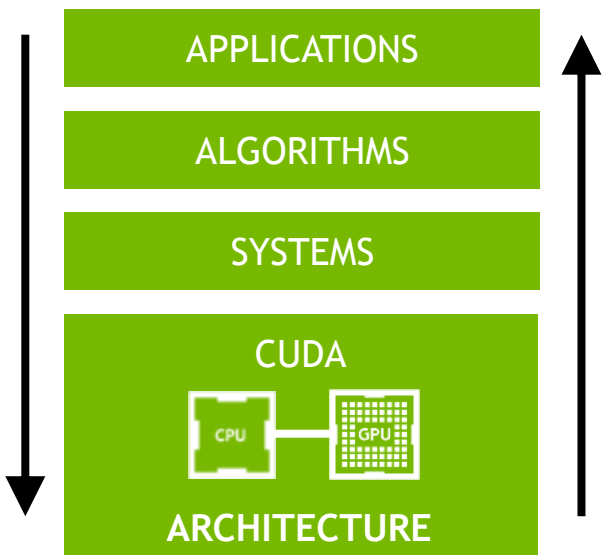
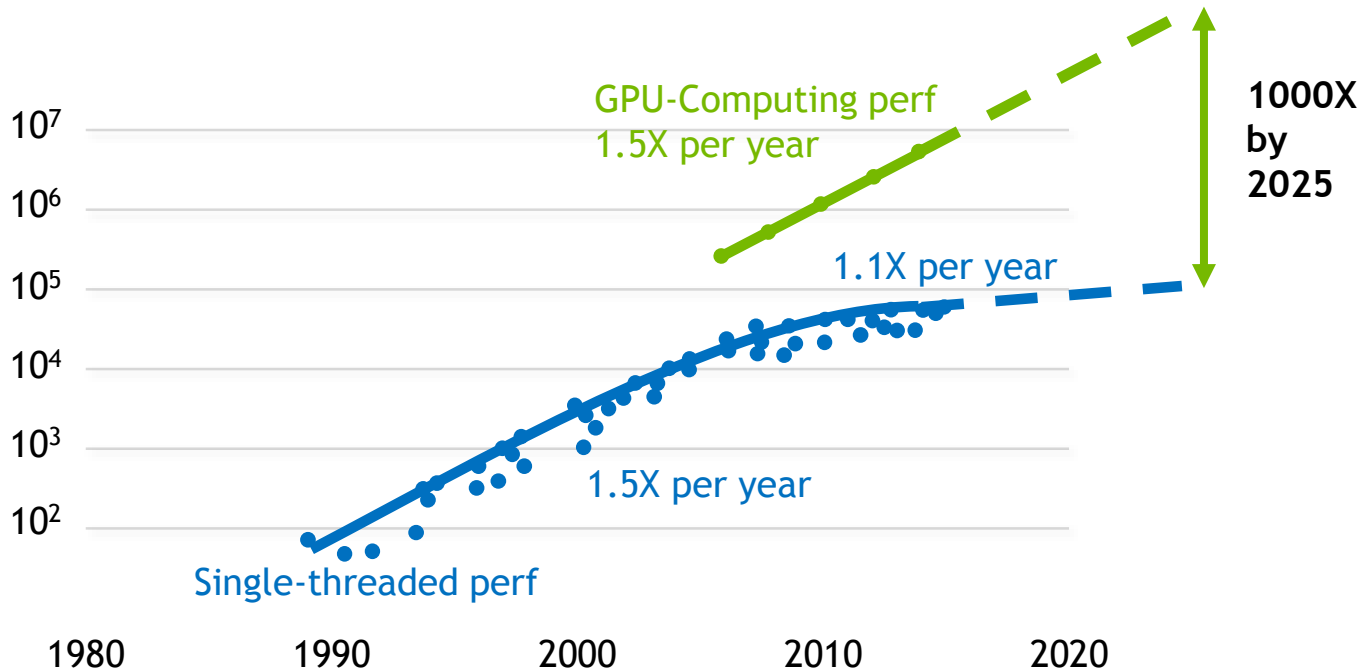
ERAD-RS'2019
TESLA PLATFORM – HPC & AI

Pedro Mario Cruz e Silva
Solutions Architect Manager, Latin América | Global Energy Team

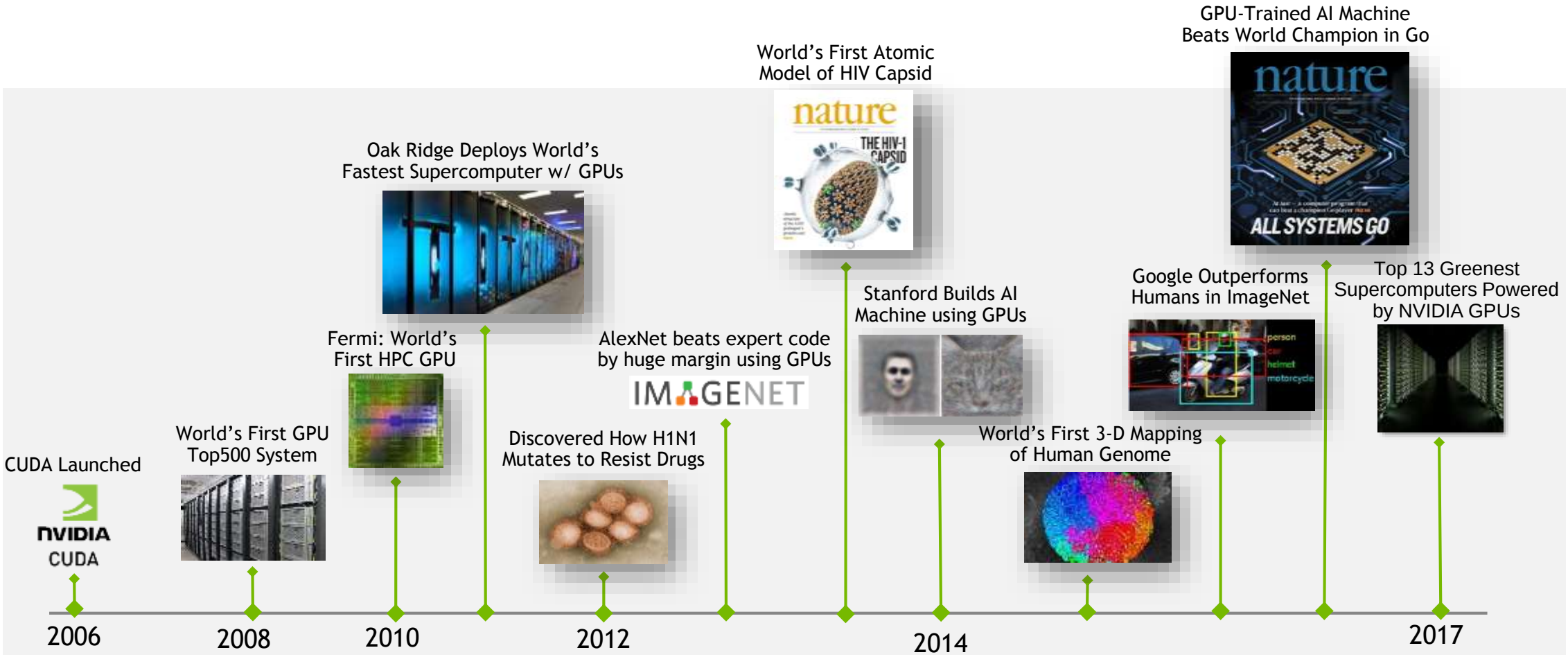
RISE OF GPU COMPUTING





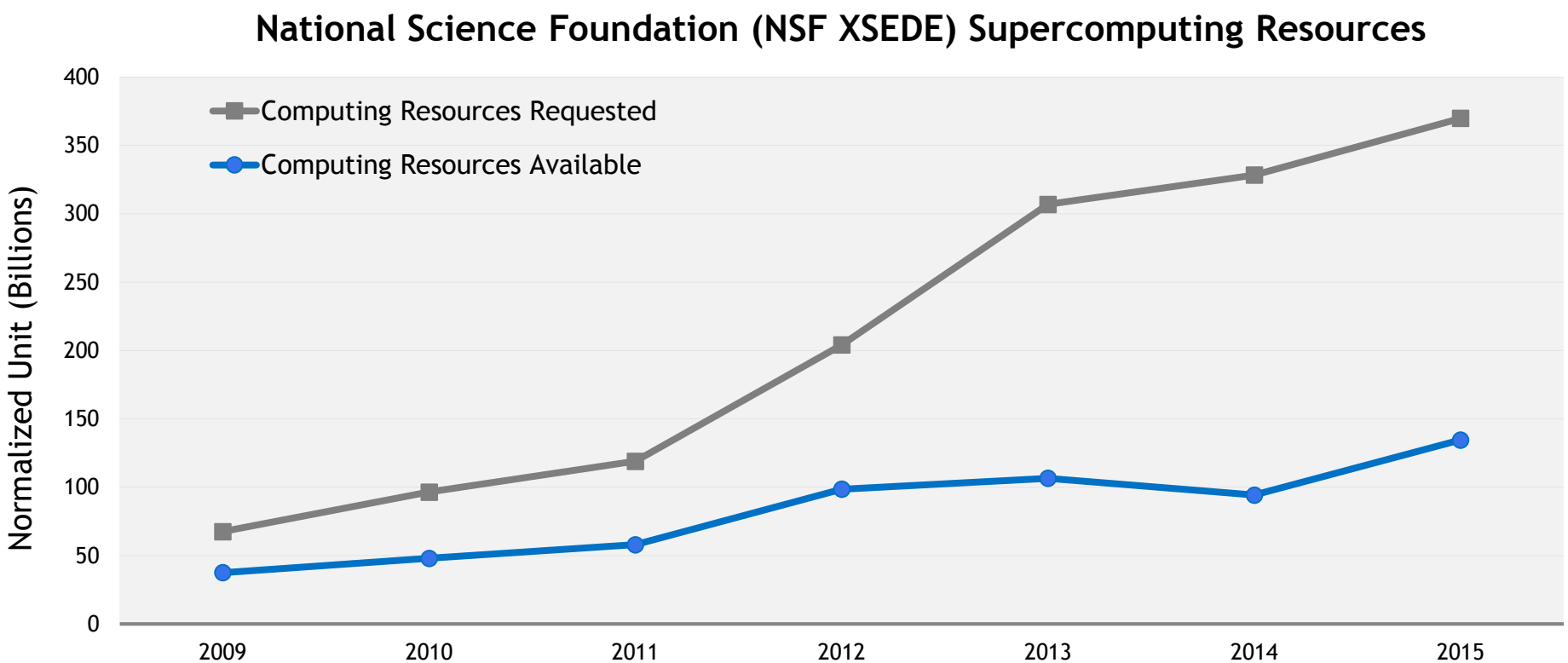
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

ELEVEN YEARS OF GPU COMPUTING



200B CORE HOURS OF LOST SCIENCE

Data Center Throughput is the Most Important Thing for HPC



Source: NSF XSEDE Data: https://portal.xsede.org/#/gallery
NU = Normalized Computing Units are used to compare compute resources across supercomputers and are based on the result of the High Performance LINPACK benchmark run on each system

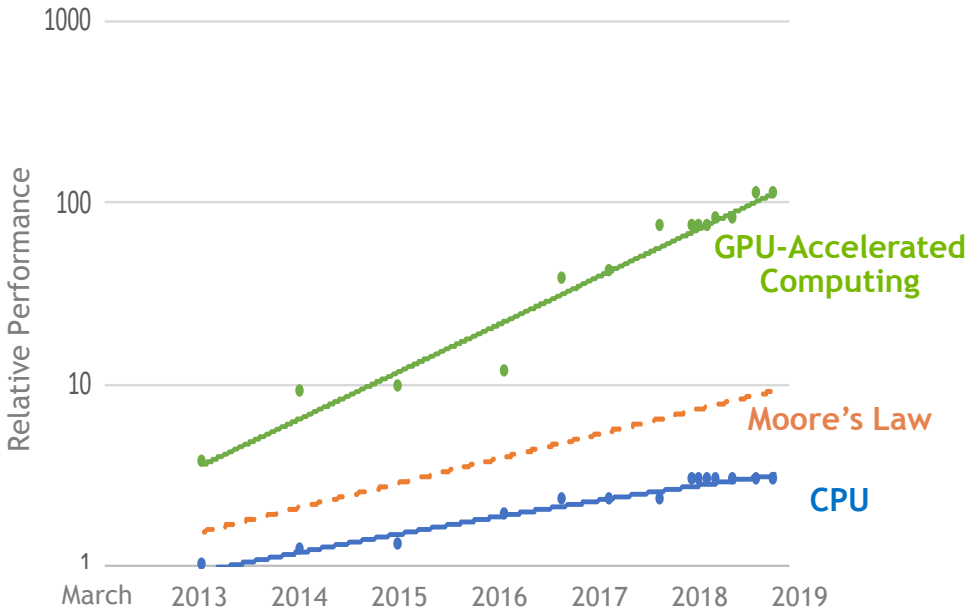
BEYOND MOORE'S LAW

Progress Of Stack In 6 Years

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

2019

- cuBLAS: 10.0
- cuFFT: 10.0
- cuRAND: 10.0
- cuSOLVER: 10.0
- cuSPARSE: 10.0
- NPP: 10.0
- Thrust: 1.9.0
- CUDA: 10.0
- Resource Mgr: r384
- Base OS: Ubuntu 16.04

Accelerated Server with Volta

NVIDIA DATA CENTER PLATFORM

Single Platform Drives Utilization and Productivity

CUSTOMER USE CASES

Speech

Translate

Recommender

Healthcare

Manufacturing

Finance

Molecular Simulations

Weather Forecasting

Seismic Mapping

Creative & Technical

Knowledge Workers

CONSUMER INTERNET & INDUSTRY APPLICATIONS

SCIENTIFIC APPLICATIONS

VIRTUAL GRAPHICS

APPS & FRAMEWORKS

python · TensorFlow · mxnet · ONNX · RAPIDS · PYTORCH · Chainer

Amber · NAMD · +600 Applications

CATIA · Ps · AUTODESK 3DS MAX

CUDA-X
NVIDIA SDK & LIBRARIES)

MACHINE LEARNING: cuDF | cuML | cuGRAPH

DEEP LEARNING: cuDNN | CUTLASS | TensorRT

HPC: OpenACC | cuFFT

VIRTUAL GPU: vDWS | vPC | vAPPS

CUDA & CORE LIBRARIES - cuBLAS | NCCL

TESLA GPUs & SYSTEMS

TESLA GPU

NVIDIA DGX FAMILY

NVIDIA HGX

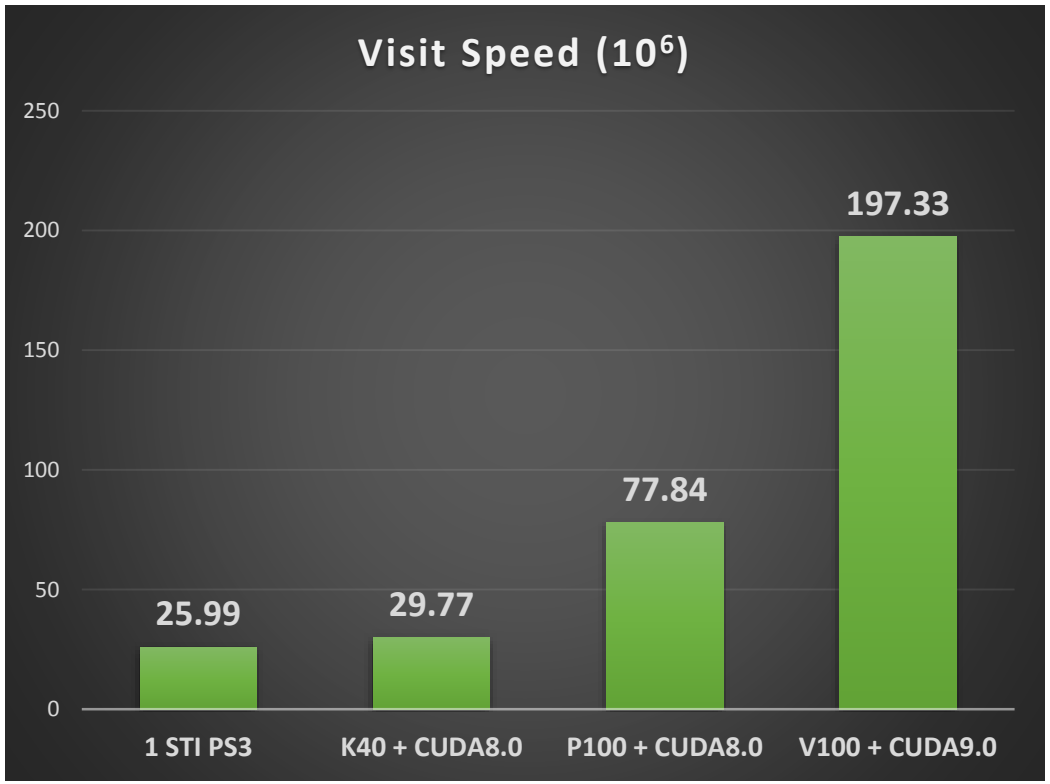
DELL · Hewlett Packard Enterprise · IBM
SYSTEM OEM

aws · Google Cloud Platform · Microsoft Azure
CLOUD

TRADITIONAL HPC

“SCALABILITY OF CPU AND GPU SOLUTIONS OF THE PRIME ELLIPTIC CURVE DISCRETE LOGARITHM PROBLEM”

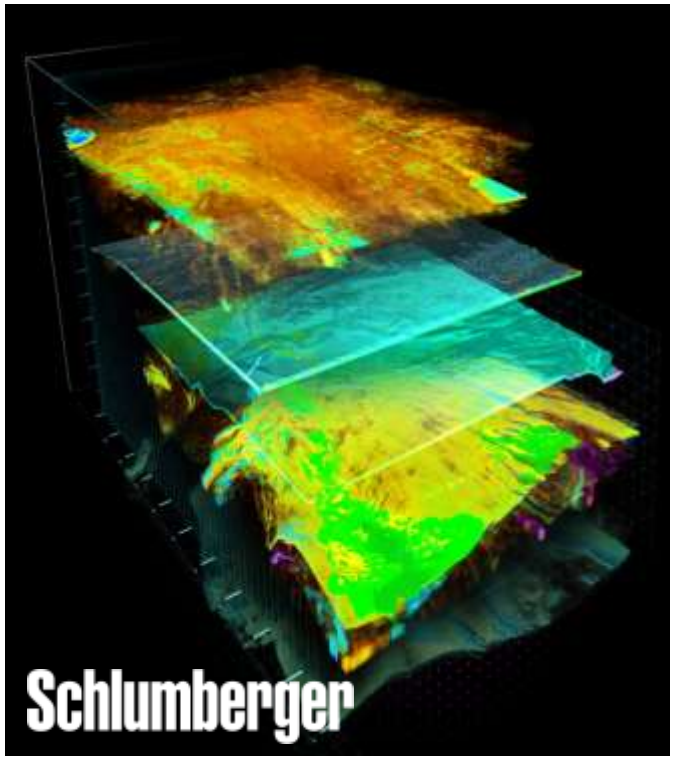
Jairo Panetta (ITA), Paulo Souza (ITA), Luiz Laranjeira (UnB), Carlos Teixeira Jr (UnB)



SBAC-PAD

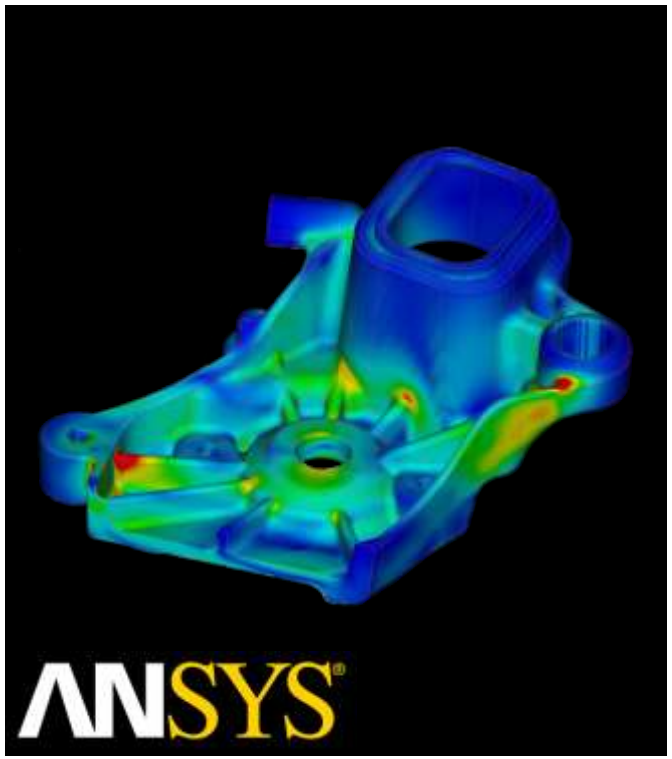
RENASIC
Rede Nacional em Segurança da Informação e Criptografia

INDUSTRY EMBRACING GPU SUPERCOMPUTING



Oil and Gas Discovery
10X increase in data processing

Realtime Fleet Analytics
Streamline routes to save >$28M



Engineering Design
Accelerate from hours to minutes

“IBM-NVIDIA SERVERS ACHIEVE HIGH-PERFORMANCE COMPUTING MILESTONE IN OIL INDUSTRY”

≡ Forbes

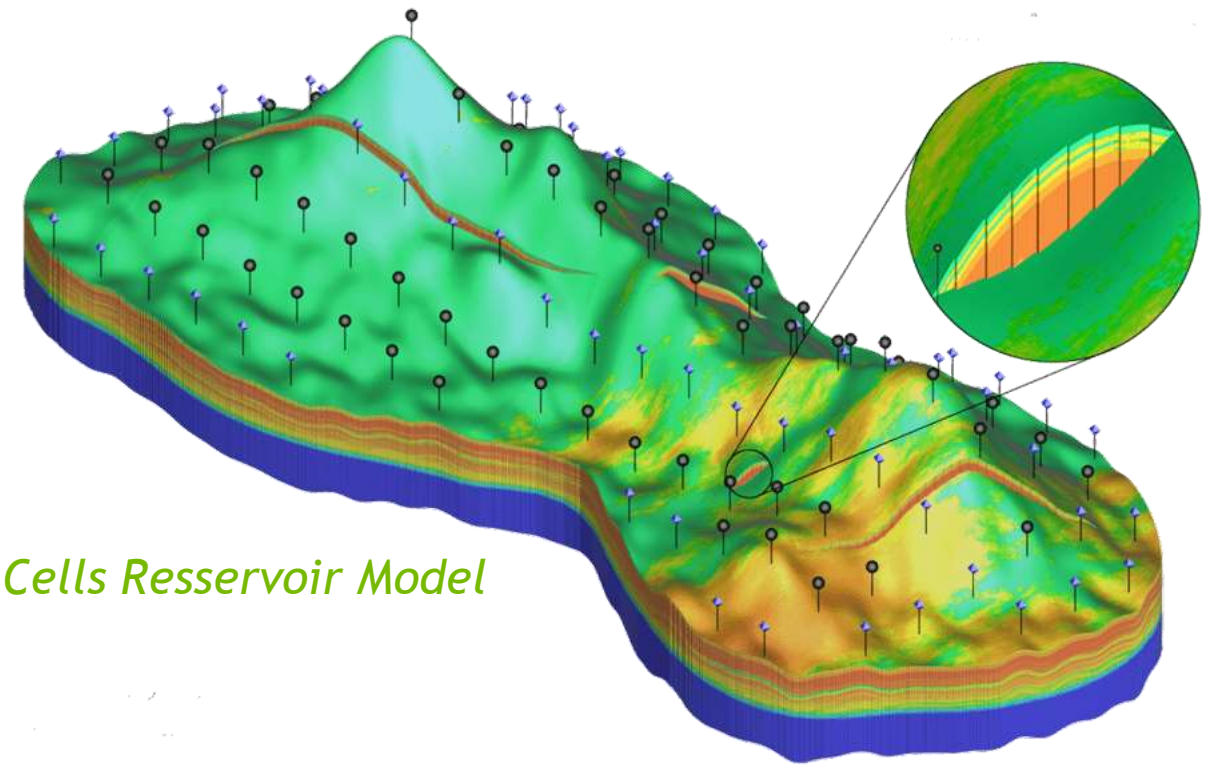
25 April 2017

Servers	22,400
Processors	24
Total CPUs	537,600

ExxonMobil using the Blue Water facility at NCSA

Servers	30
GPUs	4
Total GPUs	120

ECHELON – Simulation on GPUs Stone Ridge Technologies



1 Billion Cells Resservoir Model

https://www.forbes.com/sites/aarontilley/2017/04/25/ibm-nvidia-servers-achieve-high-performance-computing-milestone-in-oil-industry/#8e3b56626330

RESERVOIR SIMULATION

Performance Comparison

Company	Simulator/Method	Model	Production Simulation	Runtime	Reference	Cores/Servers
Saudi Aramco	GIGAPOWERS Three-phase black oil	1.03 Billion cells 3,000 wells	60 years	4 days	[1]	
Saudi Aramco	GIGAPOWERS Three-phase black oil	1.03 Billion cells 3,000 wells	60 years	21 hours	[2]	5640 Cores 470 Servers
Total/Schlumberger	INTERSECT	1.1 Billion cells 361 wells	20 years	10.5 hours	[3]	576 Cores 288 Servers
ExxonMobil	?	1 Billion cells	?	?	?	716,800 Cores 22,400 Servers
StoneRidge	Echelon Three-phase black oil	1.01 Billion cells 1,000 wells	45 years	92 minutes	?	120 GPUS 30 Servers

[1] SPE 119272 “A Next-Generation Parallel Reservoir Simulator for Giant Reservoirs”, A. Dogru et. al. 2009 SPE Reservoir Simulation Symposium.

[2] SPE 142297 “New Frontiers in Large Scale Reservoir Simulation”, A. Dogru et. al. 2011 SPE Reservoir Simulation Symposium.

[3] IPTC 17648 “Giga Cell Compositional Simulation”, E. Obi et. al., 2014 International Petroleum Technology Conference.

ENI HPC4 – GREEN DATA CENTER

The World's Most Powerful Industrial System

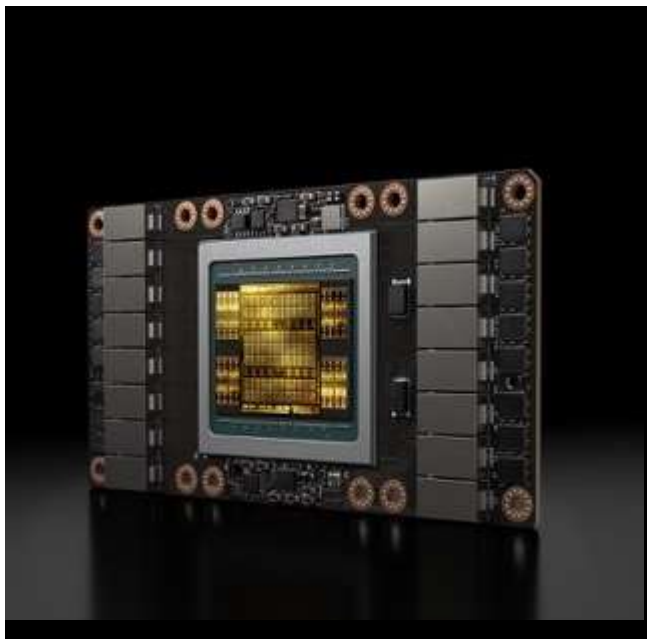
3,200 NVIDIA Tesla P100 GPU's

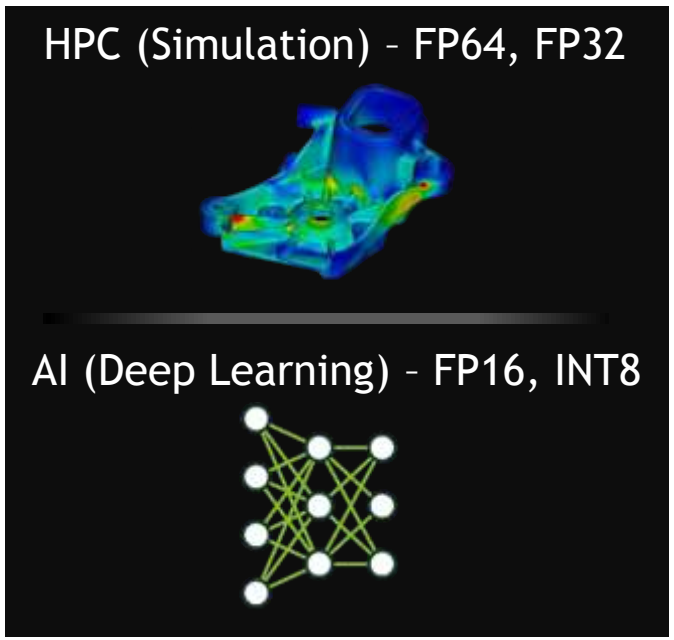
100,000 high-resolution reservoir model simulation runs, taking into account geological uncertainties, in a record time of 15 hours.

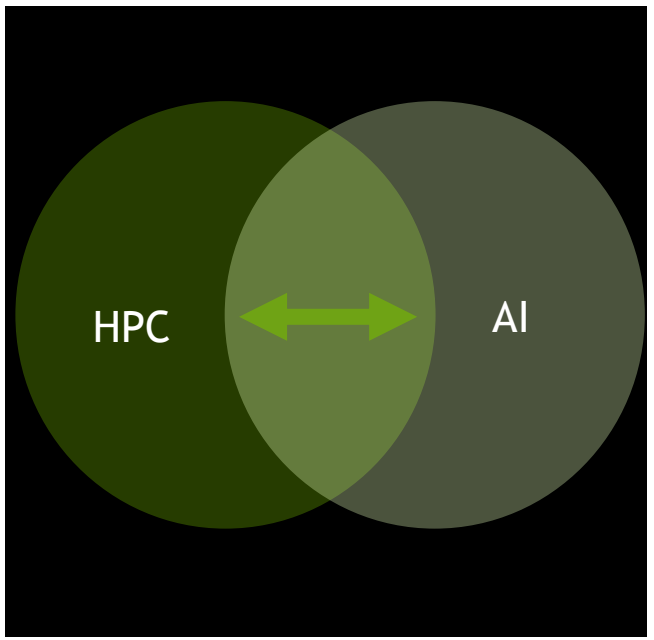
Source: https://www.eni.com/en_IT/innovation/technological-platforms/maximize-recovery/hpc.page#

DIGITAL SCIENCE
HPC + AI + DATA

GPU FUSES HPC & AI COMPUTING







AI – A NEW INSTRUMENT FOR SCIENCE

HPC

> Algorithms based on first principles theory.

> Proven models for accurate results

AI

> Neural Networks that learn patterns from large data sets

> Improve predictive accuracy and faster response time.

Dramatically Improves Accuracy and Time-to-Solution

Commercially viable fusion energy

Understanding cosmological dark energy and matter

Clinically viable precision medicine

Improvement and validation of the Standard Model of Physics

Climate/weather forecasts with ultra-high fidelity

AI FOR SCIENCE

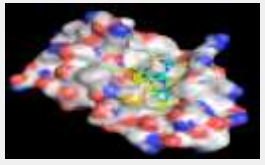
Transformative Tool To Accelerate The Pace of Scientific Innovation

90% accuracy
Fusion Sustainment
Clean Energy

33% Faster
Track Neutrinos
Particle Physics

5,000X Faster
Process LIGO Signal
Understanding Universe

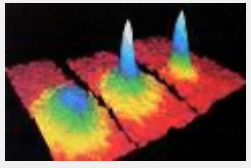
300,000X Faster
Predict Molecular Energetics
Drug Discovery



70% accuracy
Score Protein Ligand
Drug Discovery

11% higher accuracy
Monitor Earth's Vital
Climate

Weeks to 10 milliseconds
Analyze Gravitational Lensing
Astrophysics



14X Faster
Generate Bose-Einstein
Condensate (Physics)

Improves Accuracy
Enabling realization of full scientific potential

Accelerates Time to Solution
Unlocking the use of science in exciting new ways

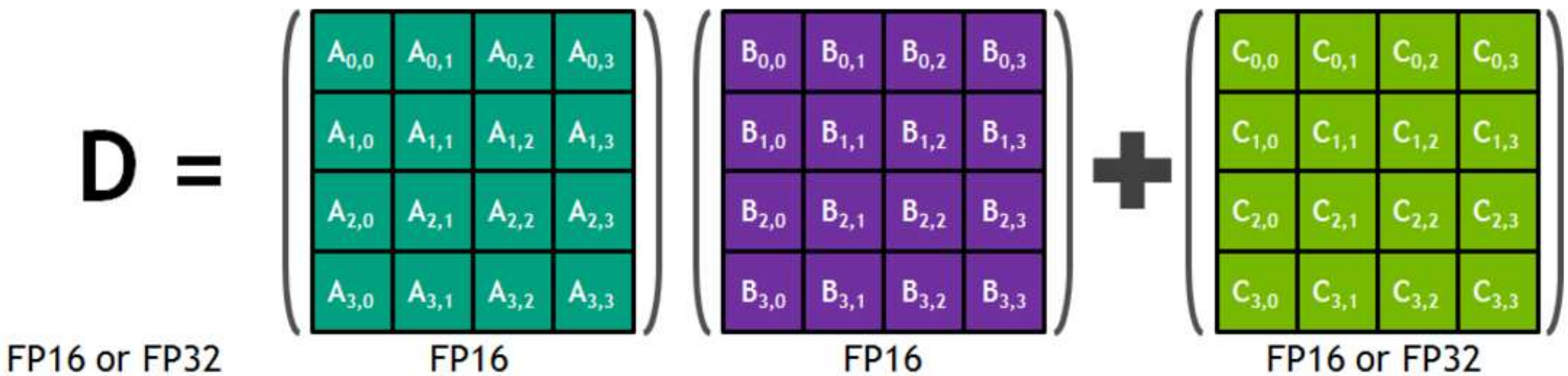
TESLA V100 TENSOR CORE GPU

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

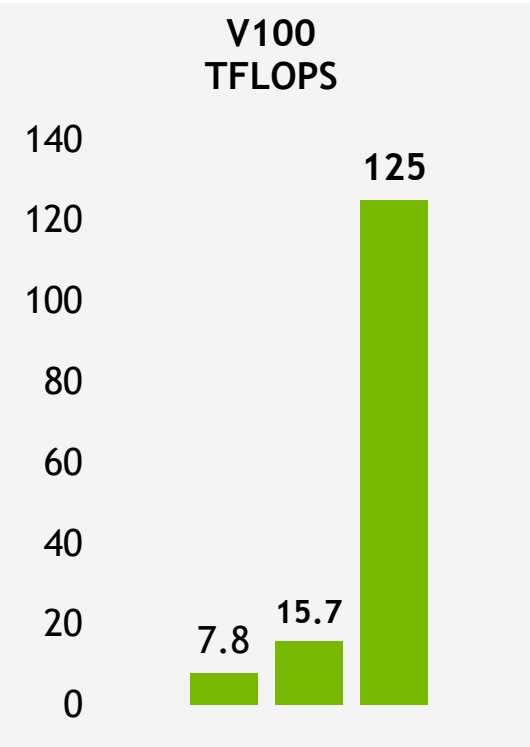
TENSOR CORE

4x4x4 matrix multiply and accumulate

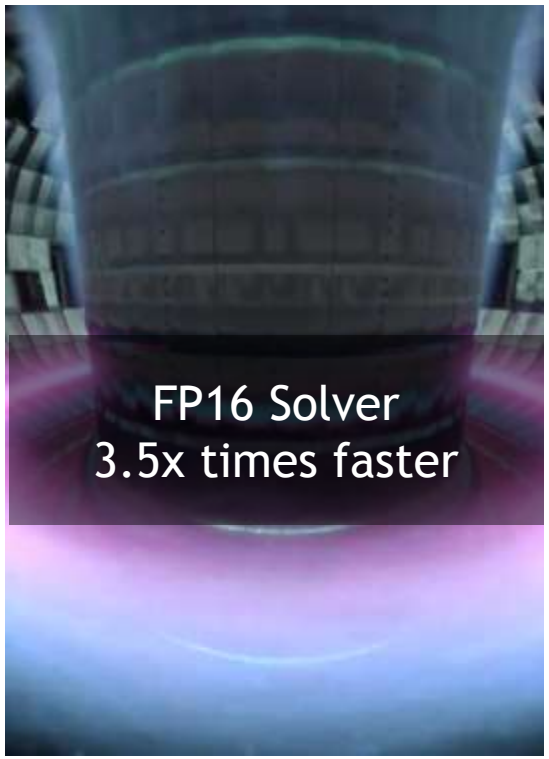


TENSOR CORES FOR SCIENCE

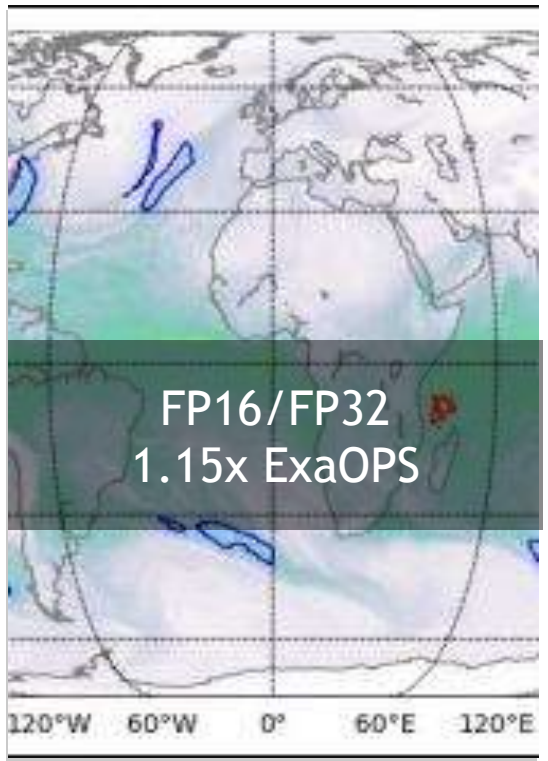
Multi-precision computing



FP64+ MULTI-PRECISION



PLASMA FUSION APPLICATION



AI-POWERED WEATHER PREDICTION

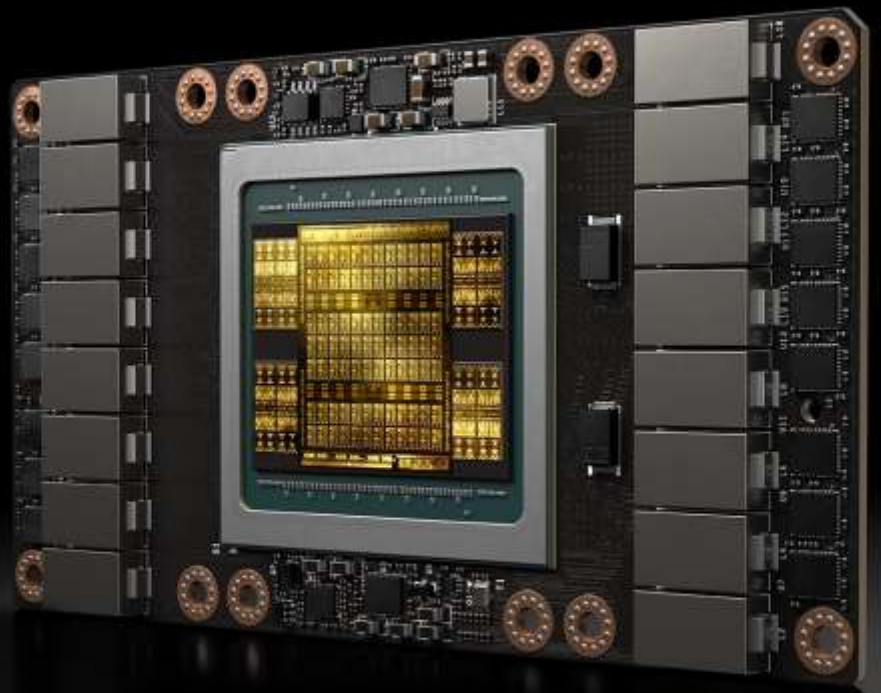
EARTHQUAKE SIMULATION

NVIDIA POWERS WORLD'S FASTEST SUPERCOMPUTER

Summit Becomes First System To Scale The 100 Petaflops Milestone

122 PF
HPC

3 EF
AI



27,648
Volta Tensor Core GPUs

NVIDIA POWERS FASTEST SUPERCOMPUTERS IN US, EUROPE, JAPAN, INDUSTRY

17 of World's 20 Most Energy-efficient Supercomputers

ORNL Summit
World's Fastest
27,648 GPUs| 122 PF

LLNL Sierra
US 2nd Fastest
17,280 GPUs| 72 PF

ABCI
Japan's Fastest
4,352 GPUs| 20 PF

Piz Daint
Europe's Fastest
5,320 GPUs| 20 PF

ENI HPC4
Fastest Industrial
3,200 GPUs| 12 PF

Twice as Nice: NVIDIA Powers Not One, But Two, Gordon Bell Prizes

November 16, 2018 by KRISTIN BRYSON

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

BUILDING A PETAFLOP(*) MACHINE

*Peak (With GPU)

How many GPUs do you need?

BUILDING A PETAFLOP(*) MACHINE

*Peak (With GPU)

How many GPUs do you need?

- 1 PFLOPS = 1000 TFLOPS
- Tesla Volta V100 32GB
  - 7.8 TFLOPS FP64
- N = 1000 / 7.8 ~= 128

BUILDING A PETAFLOP(*) MACHINE

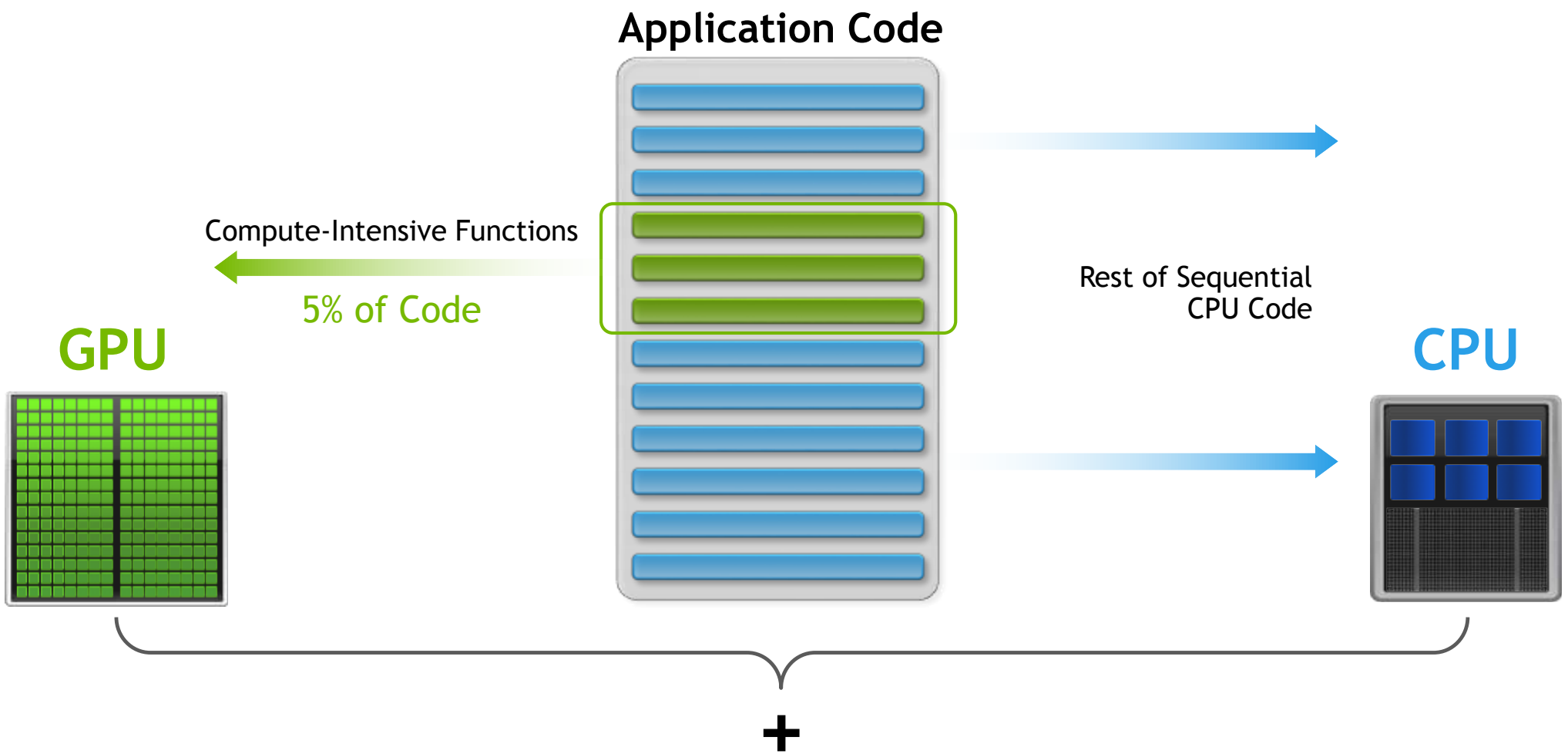
*Peak (With GPU)

How many GPUs do you need?

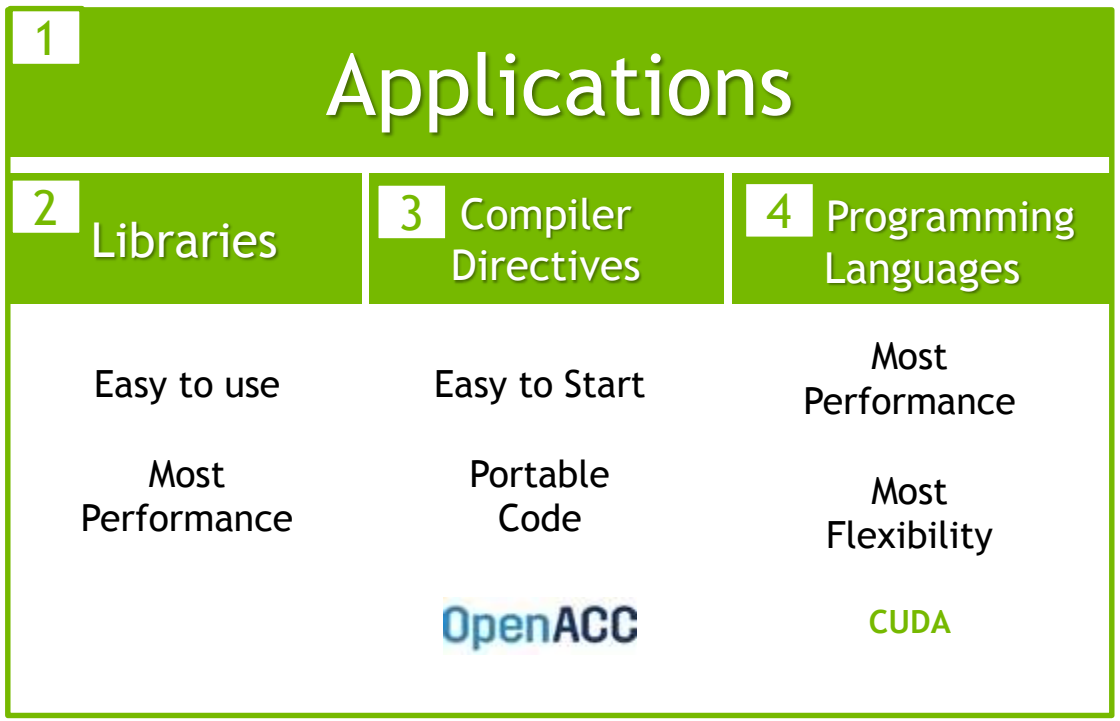
- 1 PFLOPS = 1000 TFLOPS
- Tesla Volta V100 32GB
  - 7.8 TFLOPS FP64
- N = 1000 / 7.8 ~= 128
- Server w/ 8x GPUs and 4U ~= 16 Server (Strong Node)
- 1 Rack 48U = 12x 4U Server
- 1.33 Racks!

TESLA PLATFORM
FOR DEVELOPERS

HOW GPU ACCELERATION WORKS



HOW TO START WITH GPUS



1. Review available GPU-accelerated applications
2. Check for GPU-Accelerated applications and libraries
3. Add OpenACC Directives for quick acceleration results and portability
4. Dive into CUDA for highest performance and flexibility

NVIDIA SDK

The Essential Resource for GPU Developers

GPU ACCELERATED LIBRARIES

“Drop-in” Acceleration for Your Applications

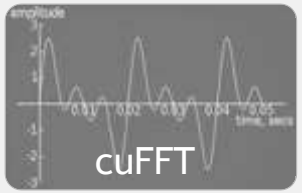
DEEP LEARNING

cuDNN

TensorRT

DeepStream SDK

SIGNAL, IMAGE & VIDEO


cuFFT

NVIDIA NPP

CODEC SDK

LINEAR ALGEBRA

cuBLAS

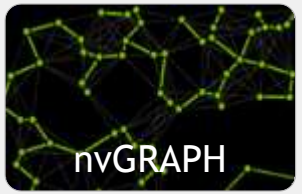
cuSPARSE

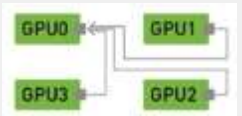
CUDA
Math library

cuSOLVER

cuRAND

PARALLEL ALGORITHMS


nvGRAPH



NCCL

WHAT IS OPENACC

Programming Model for an Easy Onramp to GPUs

Directives-based programming model for **parallel computing**

Add Simple Compiler Directive

```
main()
 {
  <serial code>
  #pragma acc kernels
  {
    <parallel code>
  }
 }
```

Designed for **performance portability** on CPUs and GPUs

Simple

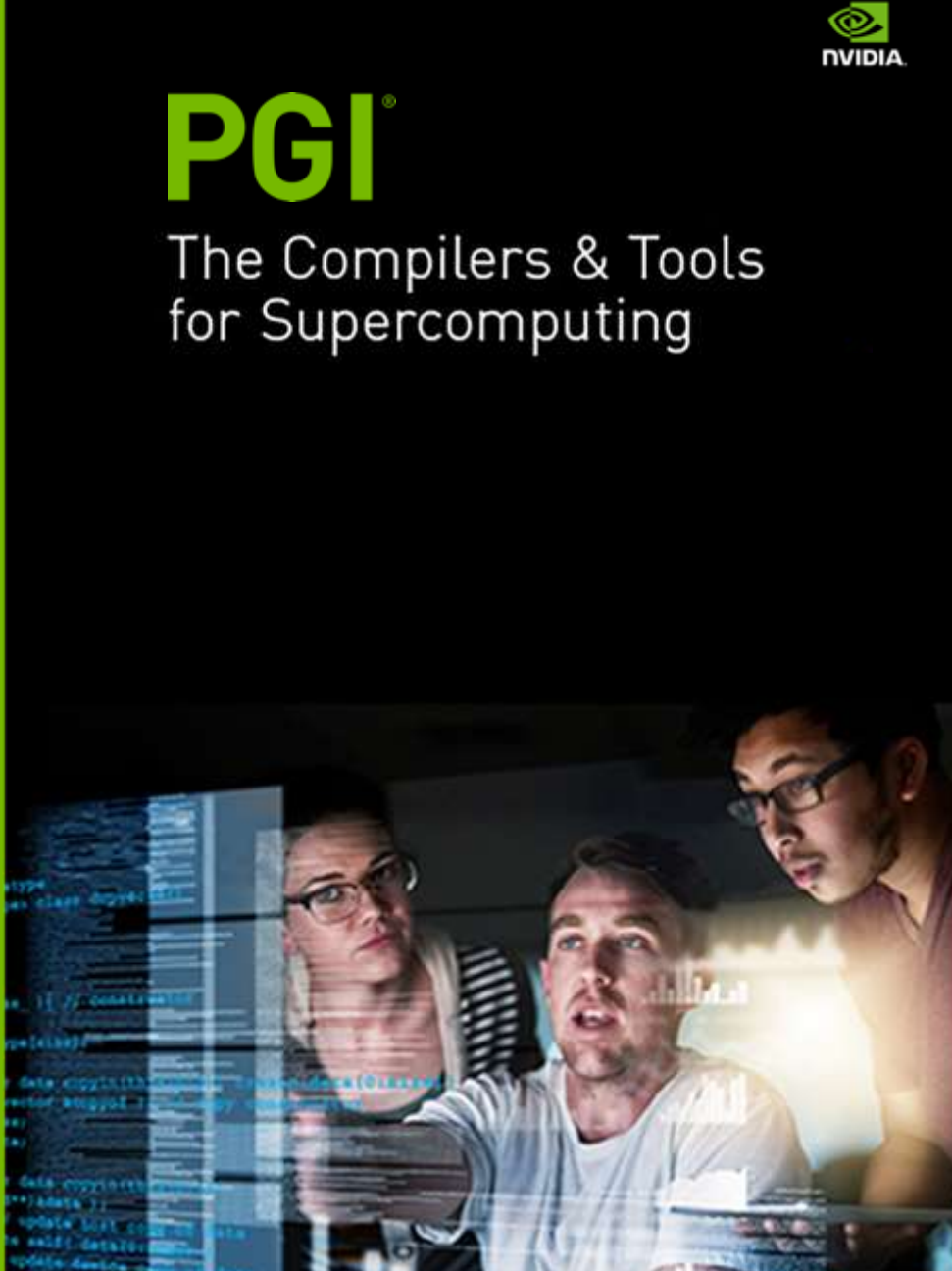
Powerful & Portable

Read more at www.openacc.org/about

OpenACC is an open specification developed by OpenACC.org consortium

PGI — THE NVIDIA HPC SDK

- Fortran, C & C++ Compilers
  - Optimizing, SIMD Vectorizing, OpenMP
- Accelerated Computing Features
  - CUDA Fortran, OpenACC Directives
- Multi-Platform Solution
  - X86-64 and OpenPOWER Multicore CPUs
  - NVIDIA Tesla GPUs
  - Supported on Linux, macOS, Windows
- MPI/OpenMP/OpenACC Tools
  - Debugger
  - Performance Profiler
  - Interoperable with DDT, TotalView



PGI | 19.1

Fortran, C and C++ for the Tesla Platform

V100 Tensor Cores

Full C++17 language

OpenACC printf()

CUDA 10.x support

OpenACC 2.6

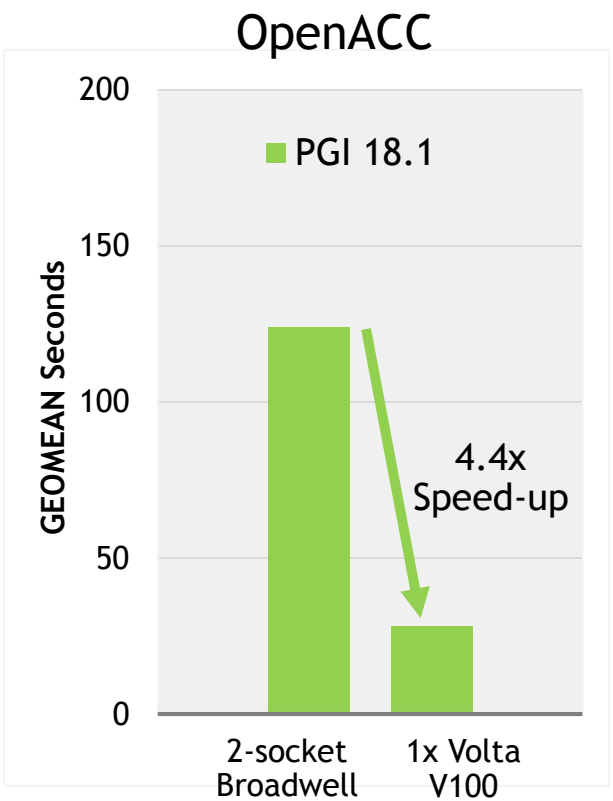
OpenMP 4.5 for multicore

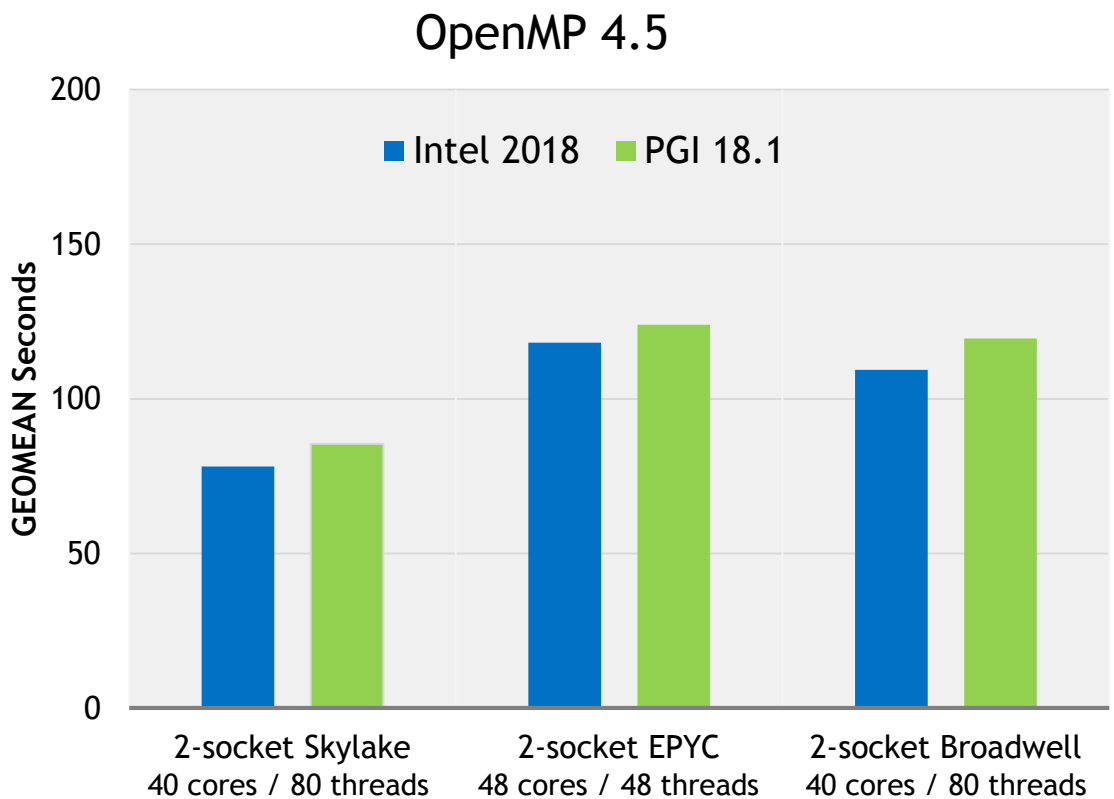
OpenACC Deep Copy

PGI in the Cloud

pgicompilers.com/whats-new

SPEC ACCEL 1.2 BENCHMARKS





Performance measured February, 2018. Skylake: Two 20 core Intel Xeon Gold 6148 CPUs @ 2.4GHz w/ 376GB memory, hyperthreading enabled. EPYC: Two 24 core AMD EPYC 7451 CPUs @ 2.3GHz w/ 256GB memory. Broadwell: Two 20 core Intel Xeon E5-2698 v4 CPUs @ 3.6GHz w/ 256GB memory, hyperthreading enabled. Volta: NVIDIA DGX1 system with two 20 core Intel Xeon E5-2698 v4 CPUs @ 2.20GHz, 256GB memory, one NVIDIA Tesla V100-SXM2 GPU @ 1.53GHz. SPEC® is a registered trademark of the Standard Performance Evaluation Corporation (www.spec.org).

PGI

SINGLE CODE FOR MULTIPLE PLATFORMS

Compiler Options

```
pgcc -fast <myCode>.c -o myApp [Serial]

pgcc -fast -ta=multicore <myCode>.c -o myApp [parallel cpu]

pgcc -fast -ta=tesla <myCode>.c -o myApp [parallel gpu]
```

OPENACC.ORG RESOURCES

Guides ● Talks ● Tutorials ● Videos ● Books ● Spec ● Code Samples ● Teaching Materials ● Events ● Success Stories ● Courses ● Slack ● Stack Overflow

Open Source Compiler

GCC 7
Includes initial support
for OpenACC 2.5

slack

https://www.openacc.org/community#slack

Resources

https://www.openacc.org/resources

Success Stories

https://www.openacc.org/success-stories

Compilers and Tools

https://www.openacc.org/tools

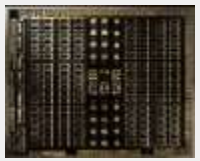
Events

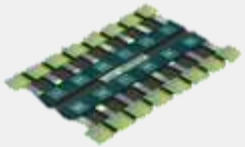
https://www.openacc.org/events

CUDA TOOLKIT 10.0

TURING AND NEW SYSTEMS

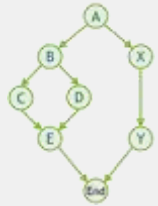
New GPU Architecture, Tensor Cores, NVSwitch Fabric





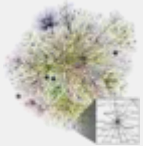
CUDA PLATFORM

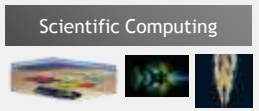
CUDA Graphs, Vulkan & DX12 Interop, Warp Matrix



LIBRARIES

GPU-accelerated hybrid JPEG decoding,
Symmetric Eigenvalue Solvers, FFT Scaling





DEVELOPER TOOLS

New Nsight Products – Nsight Systems and Nsight Compute

POWERING THE DEEP LEARNING ECOSYSTEM

NVIDIA SDK Accelerates Every Major Framework

COMPUTER VISION

OBJECT DETECTION

IMAGE CLASSIFICATION

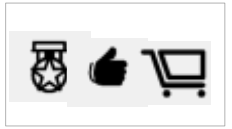
SPEECH & AUDIO

VOICE RECOGNITION

LANGUAGE TRANSLATION

NATURAL LANGUAGE PROCESSING

RECOMMENDATION ENGINES



SENTIMENT ANALYSIS

DEEP LEARNING FRAMEWORKS

Caffe

Caffe2

Chainer

Cognitive Toolkit

KALDI

mxnet

PaddlePaddle

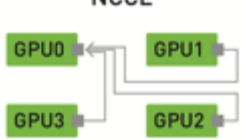
PYTORCH

TensorFlow

NVIDIA DEEP LEARNING SDK and CUDA

cuDNN

NCCL

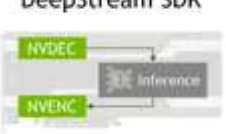


cuBLAS

TensorRT

cuSPARSE

DeepStream SDK

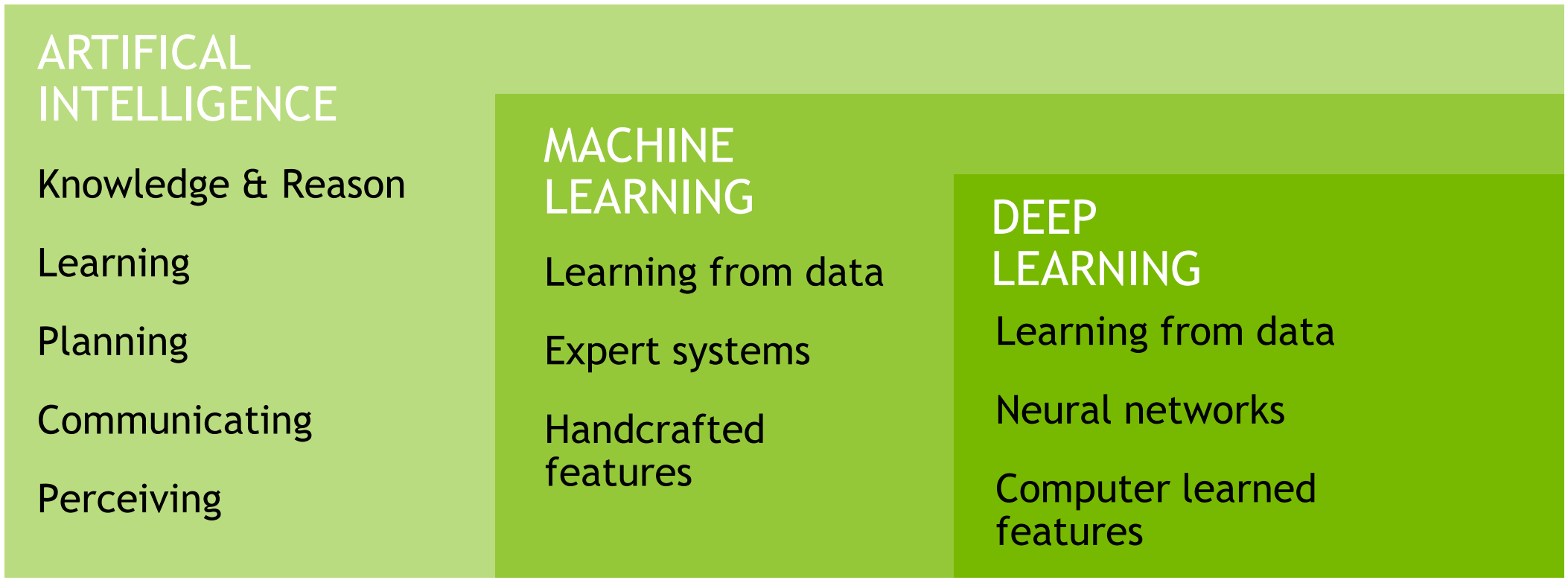


developer.nvidia.com/deep-learning-software

DEEP LEARNING

LEARNING FROM DATA

AND SOME BUZZ WORDS



NVIDIA.

A NEW COMPUTING MODEL

TRAINING

Training Data

Input

"Label"
Output

Trained Neural Network

INFERENCE

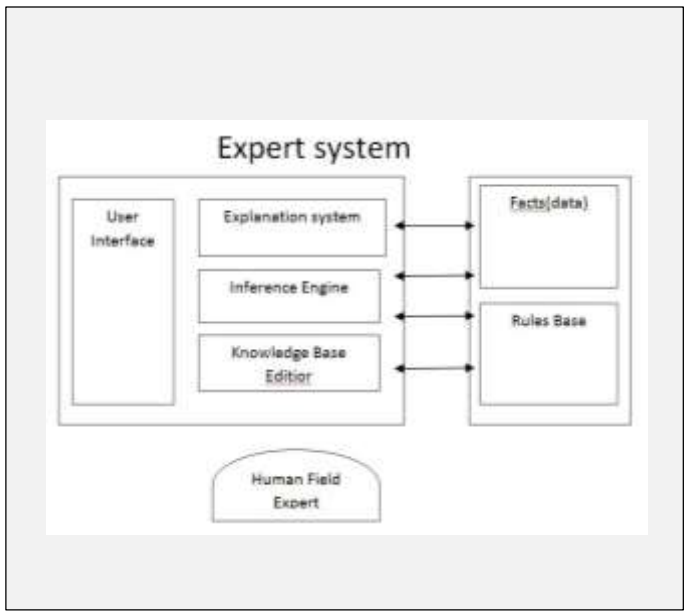
Input

Trained Neural Network

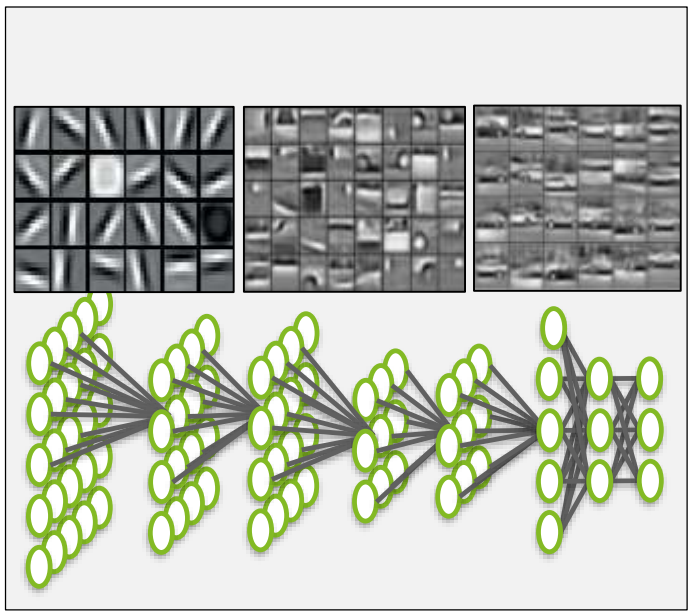
"Label"
Output

A NEW COMPUTING MODEL

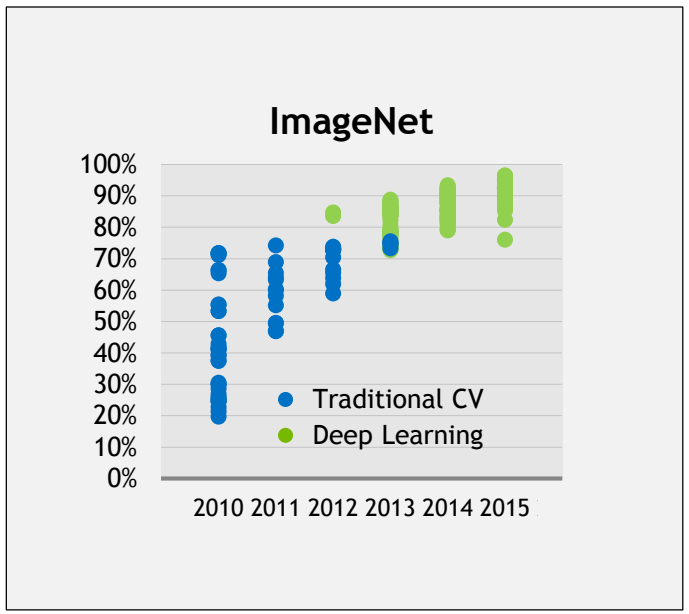
Outperform experts, facts, rules with software that writes software



Traditional Computer Vision
Experts + Time



Deep Learning Object Detection
DNN + Data + GPU



Deep Learning Achieves
"Superhuman" Results

“ACCELERATING EULERIAN FLUID SIMULATION WITH CONVOLUTIONAL NETWORKS”

Tompson, J., Schlachter, K., Sprechmann, P., & Perlin, K. (2016). Accelerating Eulerian Fluid Simulation With Convolutional Networks. *arXiv preprint arXiv:1607.03597*.

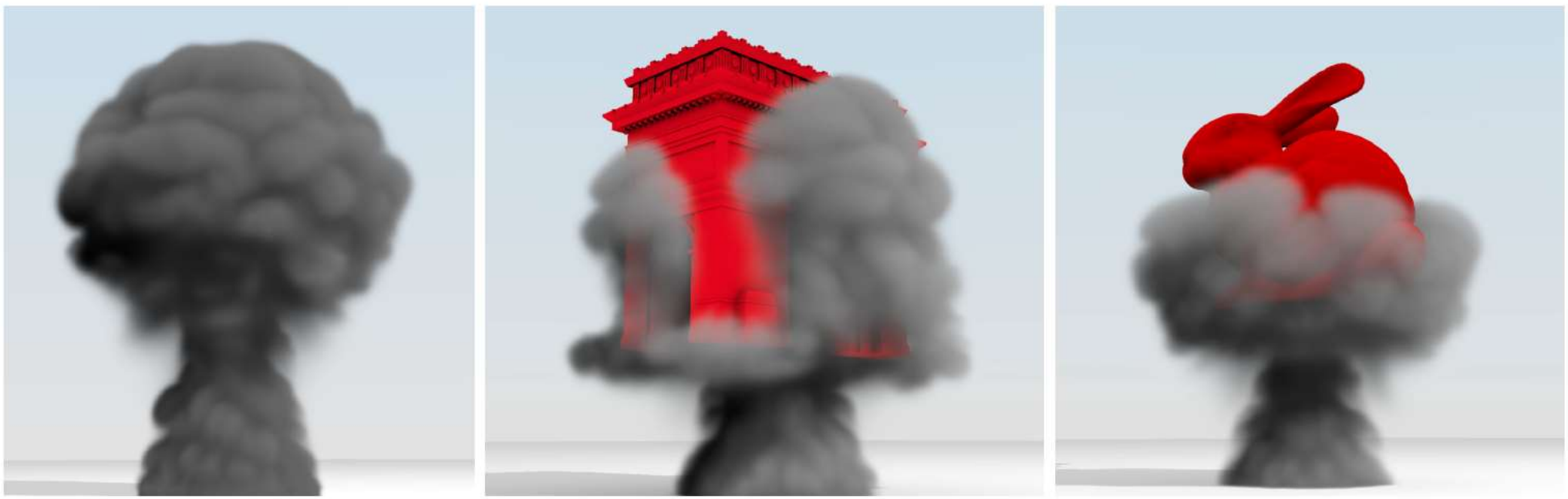
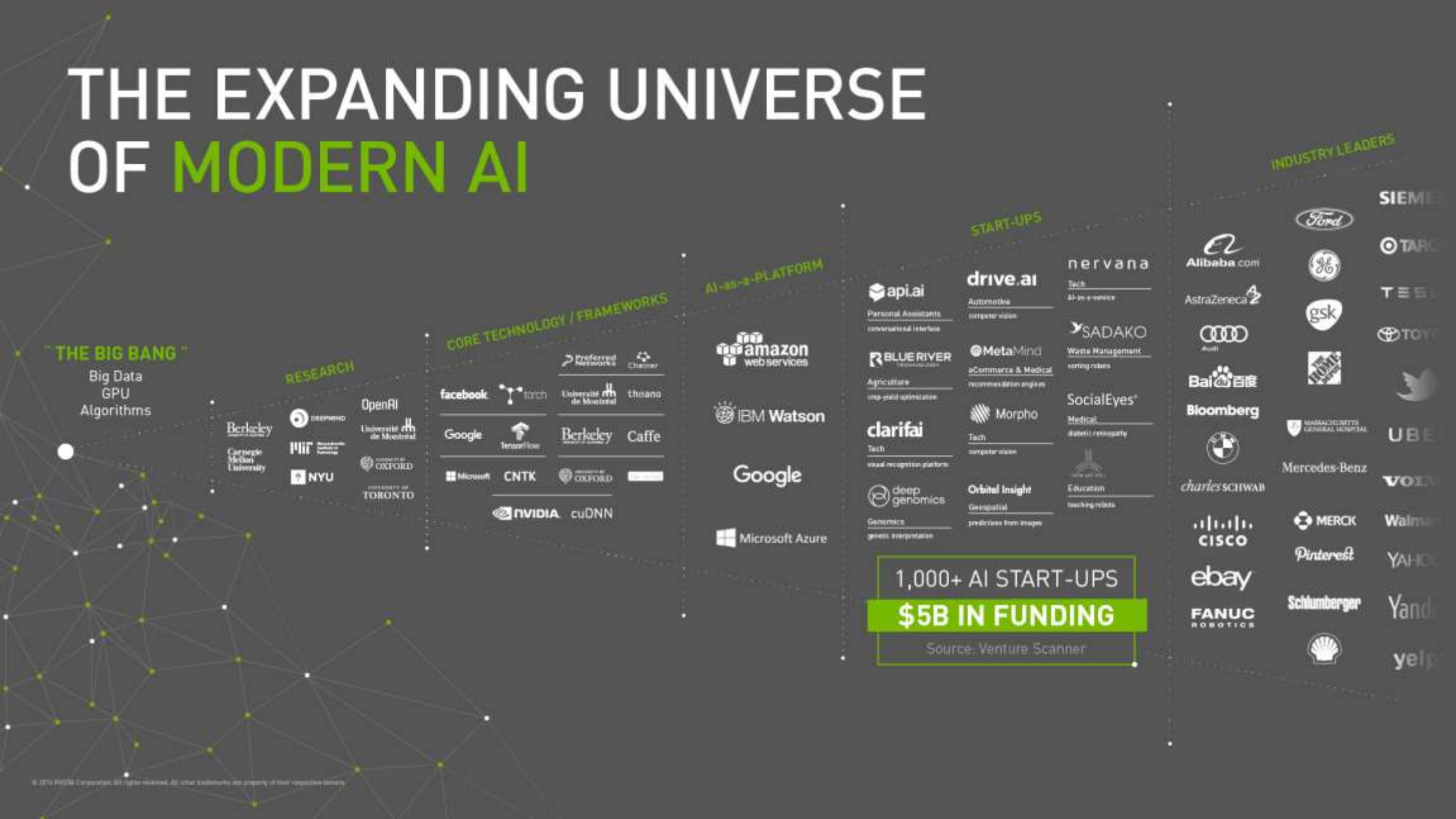


Fig. 1: Smoke simulation using our system - our method is capable of fast and accurate simulation of the Euler Equations for incompressible fluid flow at interactive frame-rates. Videos can be found at: `http://cims.nyu.edu/~schlacht/CNNFluids.htm`.

"ACCELERATING EULERIAN FLUID SIMULATION WITH CONVOLUTIONAL NETWORKS"
HTTPS://WWW.YOUTUBE.COM/WATCH?V=W71ZXKNIJFO

THE EXPANDING UNIVERSE OF MODERN AI



NEWS FEATURE

nature

THE LEARNING MACHINES

Using massive amounts of data to recognize photos and speech, deep-learning computers are taking a big step towards true artificial intelligence.

BY NICOLA JONES

Three years ago, researchers at the secretive Google X lab in Mountain View, California, extracted some 10 million still images from YouTube videos and fed them into Google Brain — a network of 1,000 computers programmed to soak up the world much as a human toddler does. After three days looking for recurring patterns, Google Brain decided, all on its own, that there were certain repeating categories it could identify: human faces, human bodies and ... cats[1].

Google Brain's discovery that the Internet is full of cat videos provoked a flurry of jokes from journalists. But it was also a landmark in the resurgence of deep learning: a three-decade-old technique in which massive amounts of data and processing power help computers to crack many problems that humans solve almost intuitively, from recognizing faces to understanding language.

Deep learning itself is a revival of an even older idea for computing: neural networks. These systems, loosely inspired by the densely interconnected neurons of the brain, mimic human learning by changing the strength of simulated neural connections on the basis of experience. Google Brain, with about 1 million simulated neurons and 1 billion simulated connections, was ten times larger than any deep neural network before it. Project founder Andrew Ng, now director of the Artificial Intelligence Laboratory at Stanford University in California, has gone on to make deep-learning systems ten times larger again.

Such advances make for exciting times in

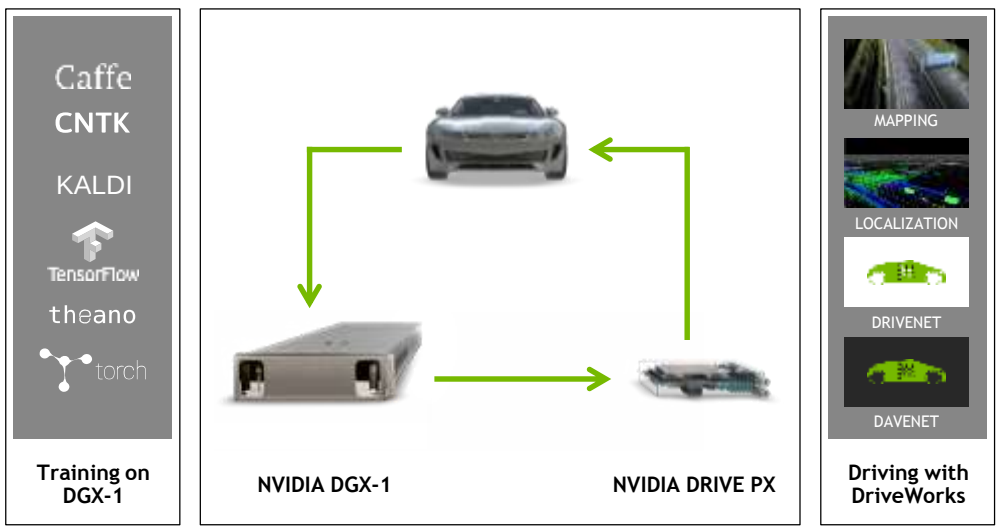
©2014 Macmillan Publishers Limited. All rights reserved

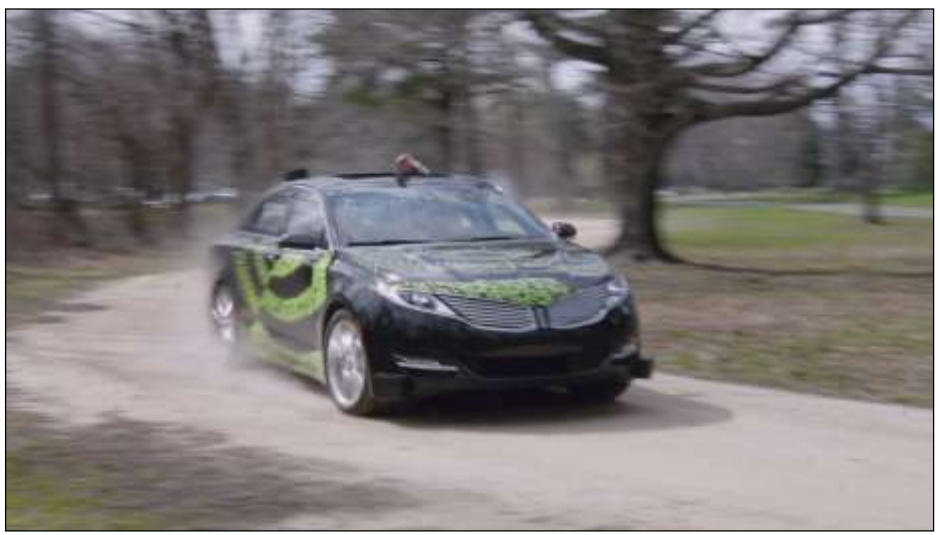
TESLA REVOLUTIONIZES DEEP LEARNING

GOOGLE BRAIN APPLICATION

	BEFORE TESLA	AFTER TESLA
Cost	$5,000K	$200K
Servers	1,000 Servers	16 Tesla Servers
Energy	600 KW	4 KW
Performance	1x	6x

NEW AI DRIVING



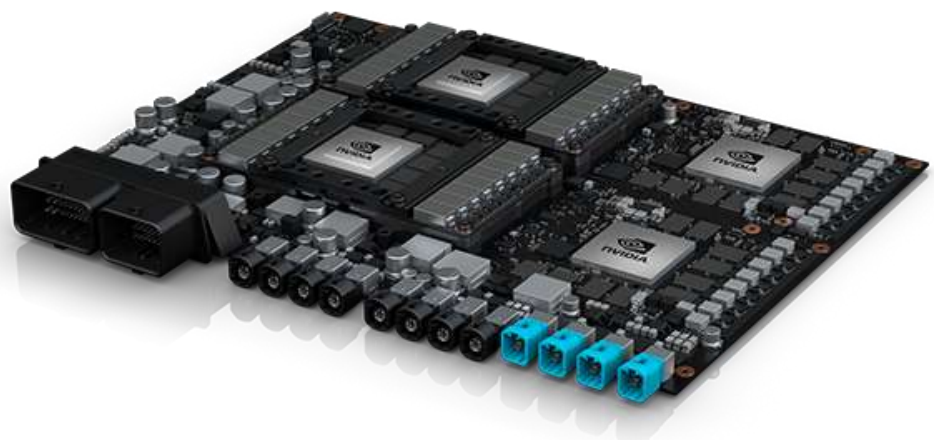


WATCH VIDEO

NVIDIA DRIVE PEGASUS

First AI Computer to Make Robotaxis a Reality

WATCH VIDEO



https://mlperf.org/

Supporting companies

Alibaba, AMD, Arm, Baidu, Cadence, Cerebras, Cisco, Cray, Dividiti, Enflame Tech, Esperanto, Google, Groq, Huawei, Intel, MediaTek, Mentor Graphics, Mythic, NetApp, NVIDIA, One Convergence, Rpa2ai, Sambanova, Samsung S.LSI, Sigopt, Synopsys, Tensyr, Wave Computing

Contributions by researchers from

Harvard University, Stanford University, University of Arkansas, Littlerock, University of California, Berkeley, University of Minnesota, University of Texas, Austin, University of Toronto

First Industry Benchmark for Measuring AI Performance

ML-PERF

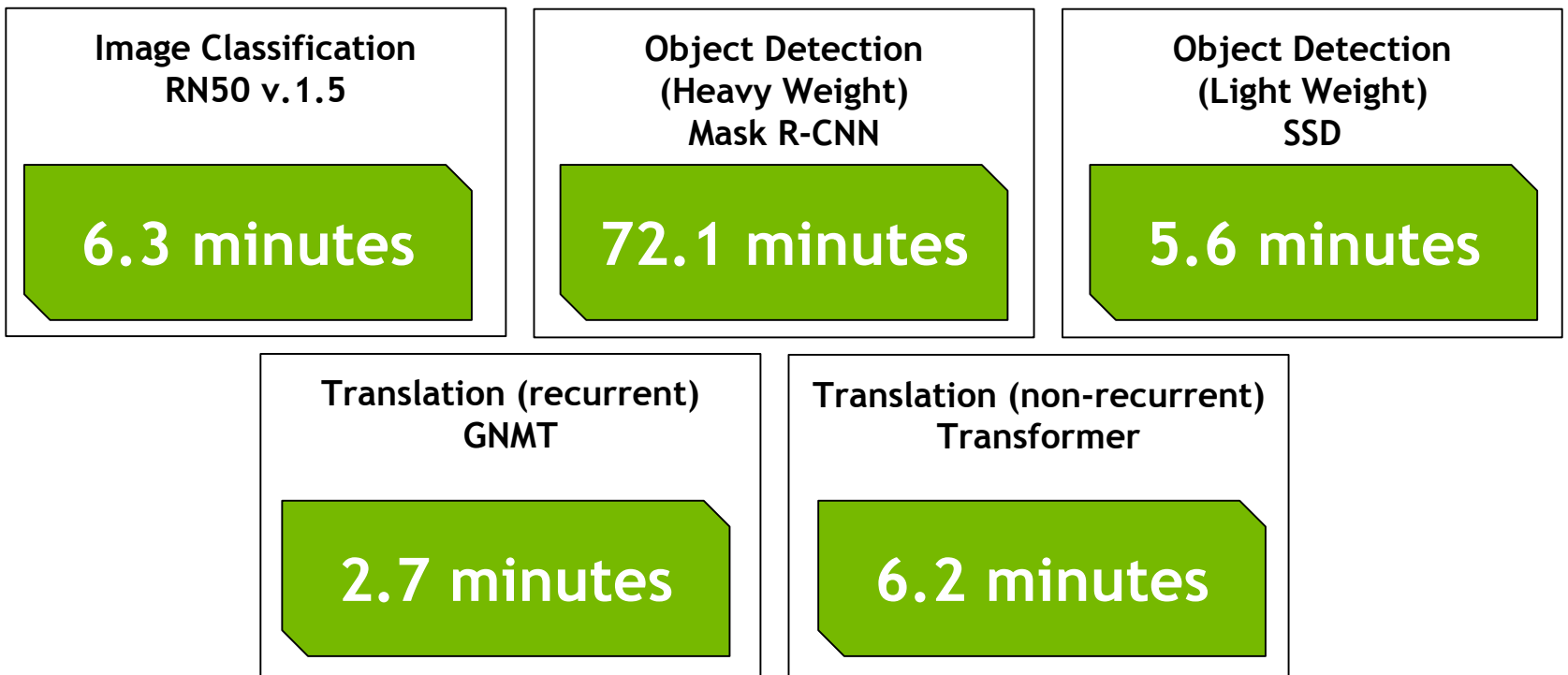
Results, December 2018

Closed Division Speedups														
					Benchmark results (speedup relative to reference implementation)									
					Image classifi-cation	Object detection, light-weight	Object detection, heavy-wt.	Translation , recurrent	Translation , non-recur.	Recom-mendation	Reinforce-ment Learning			
					ImageNet	COCO	COCO	WMT E-G	WMT E-G	MovieLens-20M	Pro games			
#	Submitter	Hardware	Chip count and type	Software	ResNet-50 v1.5	SSD w/ ResNet-34	Mask-R-CNN	NMT	Transformer	NCF	Mini Go	Cloud Scale	Power (unofficial, submitter-provided)	Details

NVIDIA

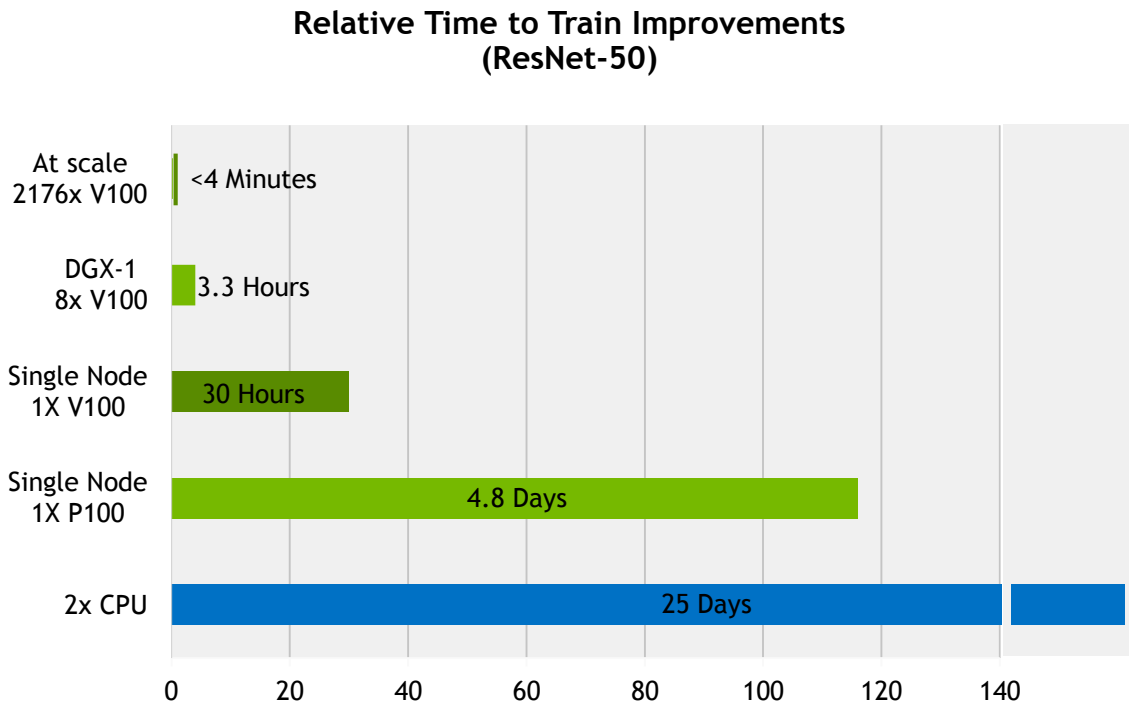
MLPERF RESULTS - AT SCALE

Results are Time to Complete Model Training



Test Platform: For Image Classification and Translation (non-recurrent), DGX-1V Cluster. For Object Detection (Heavy Weight) and Object Detection (Light Weight), Translation (recurrent) DGX-2H Cluster. Each DGX-1V, Dual-Socket Xeon E5- 2698 V4, 512GB system RAM, 8 x 16 GB Tesla V100 SXM-2 GPUs. Each DGX-2H, Dual-Socket Xeon Platinum 8174, 1.5TB system RAM, 16 x 32 GB Tesla V100 SXM-3 GPUs connected via NVSwitch

TESLA PLATFORM ENABLES DRAMATIC REDUCTION IN TIME TO TRAIN



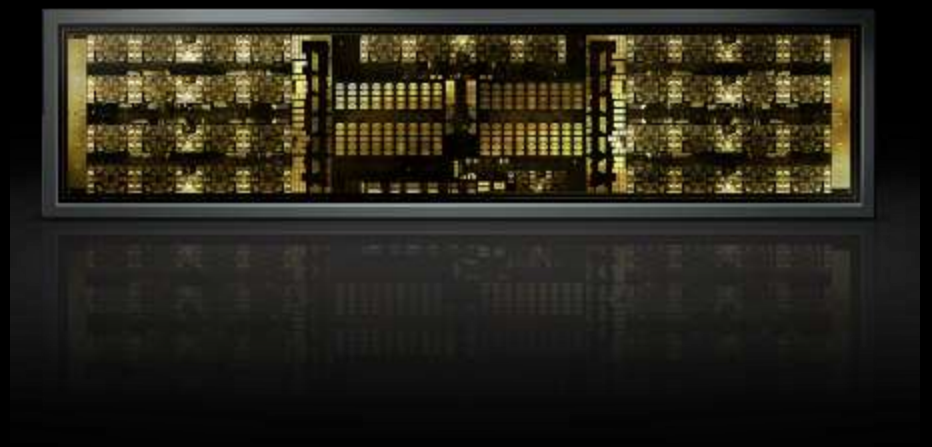
ResNet-50, 90 epochs to solution | CPU Server: dual socket Intel Xeon Gold 6140
Sony 2176x V100 record on https://nnabla.org/paper/imagenet_in_224sec.pdf

NVIDIA® DGX-1™

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



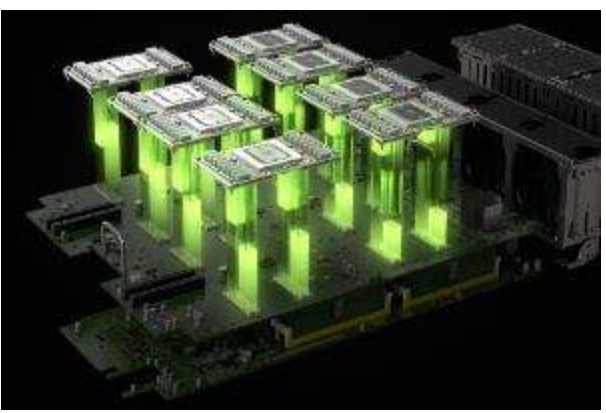
ANNOUNCING NVIDIA DGX-2

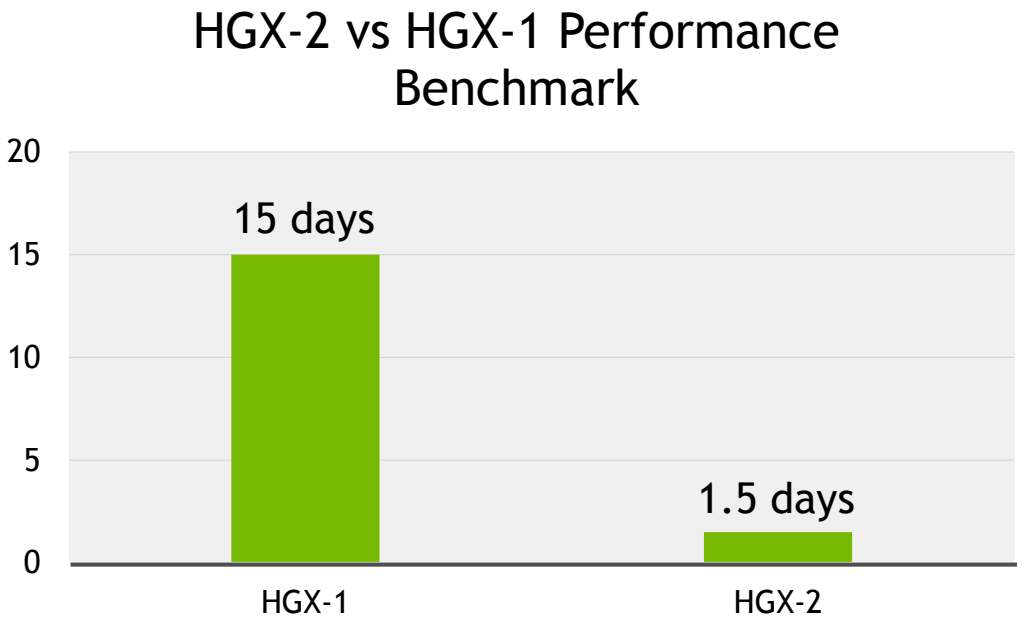
THE LARGEST GPU EVER CREATED

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

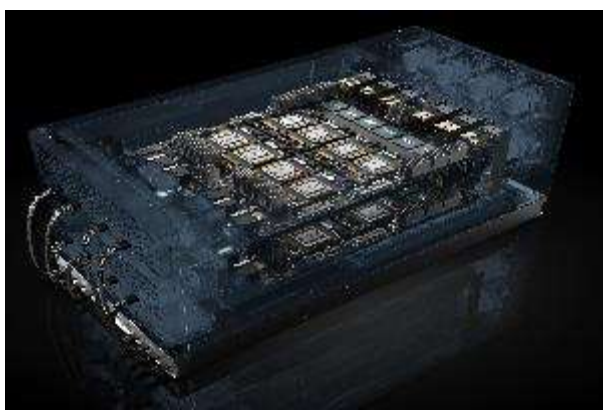
10X PERFORMANCE GAIN IN LESS THAN A YEAR

HGX-1, SEP'17





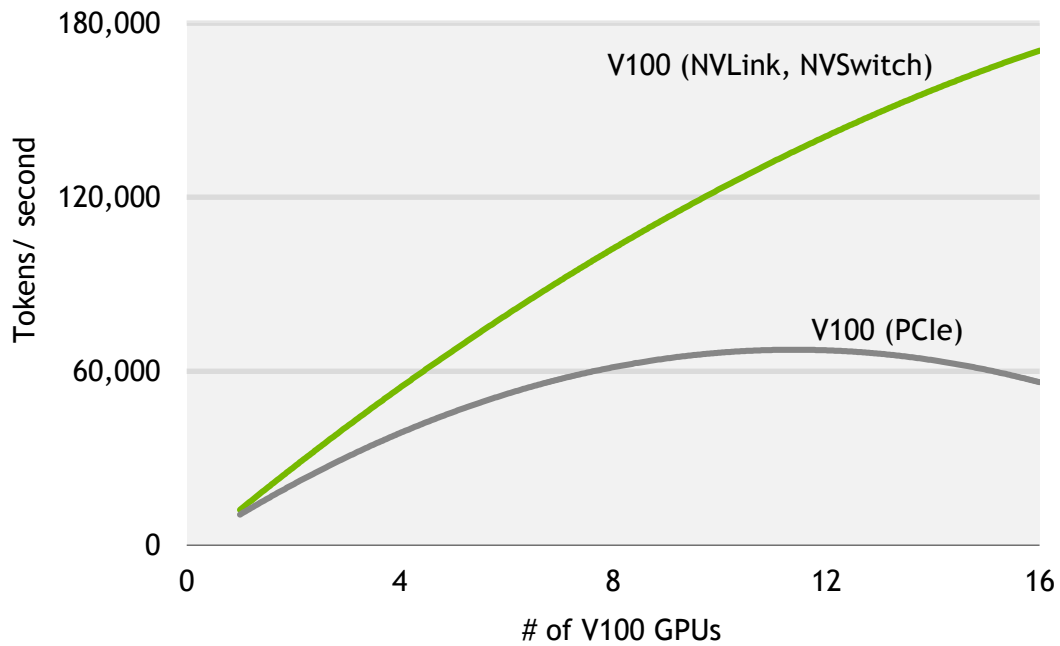
HGX-2, MAY'18



software improvements across the stack including NCCL, cuDNN, etc.

FairSeq, trained with WMT'14 English-French dataset in 55 epochs
HGX-1 9/2017 SW stack (run on NVIDIA DGX-1)
HGX-2 3/2018 SW stack (run on NVIDIA DGX-2)

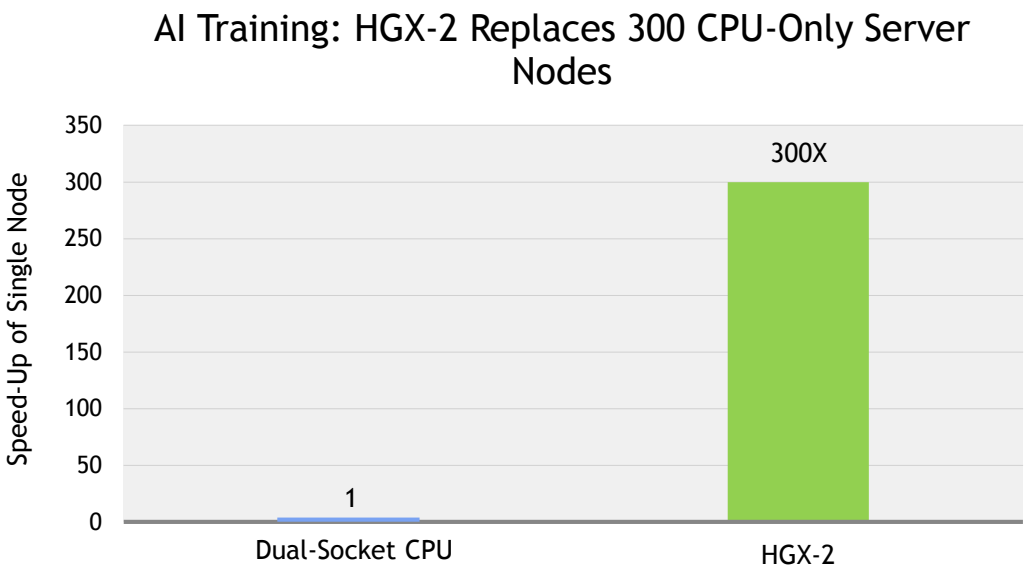
SCALING-UP PERFORMANCE WITH NVSWITCH



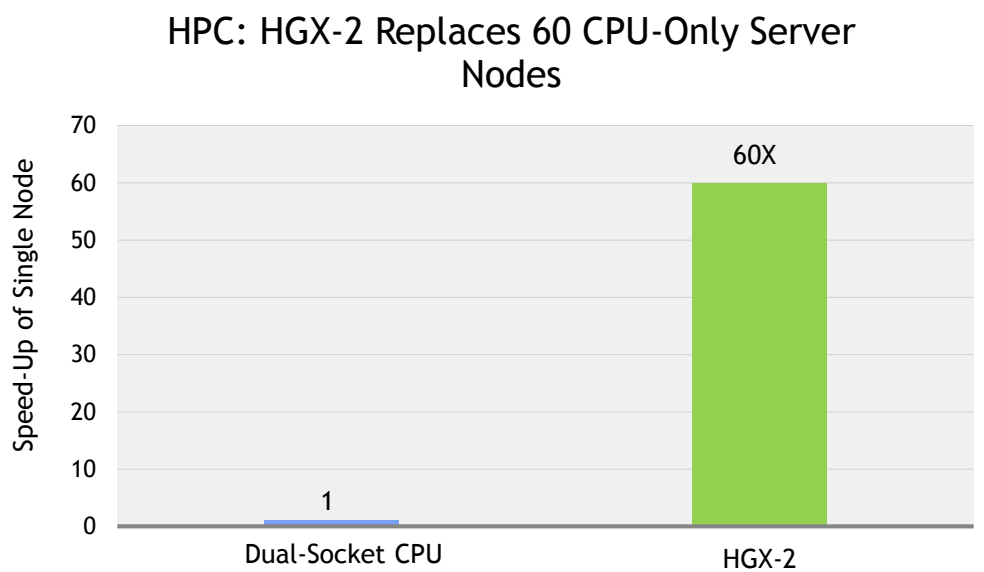
Transformer with MoE Layers | Training Dataset: 1B Word Benchmark for Language Modeling | Batch size of 8,192 per GPU

AI AND HPC BENCHMARKS: HGX-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 | Dataset: ImageNet2012 |



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

DEEP LEARNING
INFERENCE

GPU INFERENCE ADOPTION IS ACCELERATING

VISUAL SEARCH

Bing

60X Latency Improvement
Real-Time Search

VIDEO ANALYSIS

VALOSSA

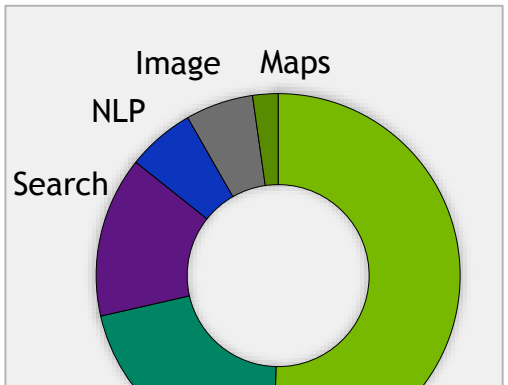
12X Faster Inference
Live Video Analysis

ADVERTISING

SAP

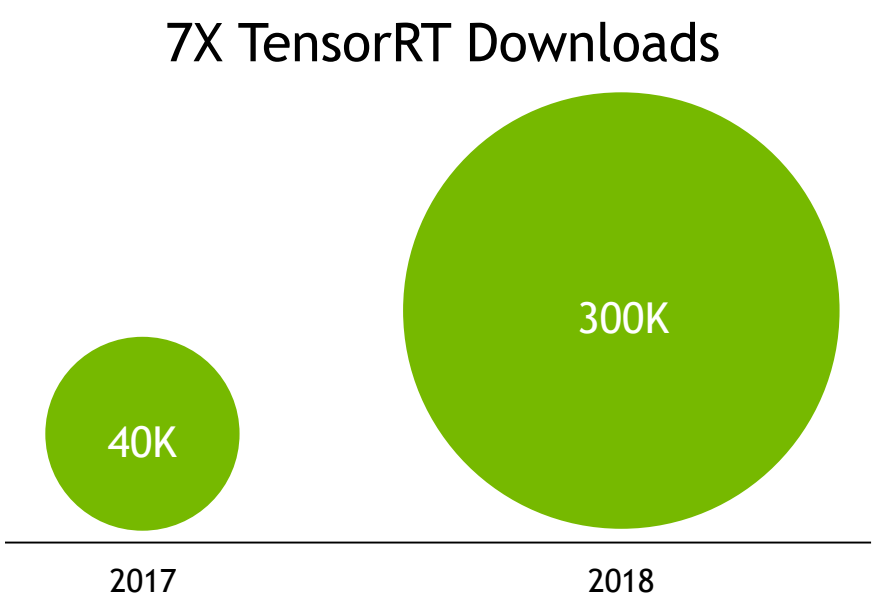
40X Higher Performance
Real-Time Brand Impact

INFERENCE USE CASES



Tesla P4, TensorRT Adoption
Use Cases

WORLD'S LEADING TECH COMPANIES ADOPT NVIDIA TO ACCELERATE AI DEPLOYMENT



Bytedance
NLP

Clarifai
Computer Vision

iFlyTek
Speech Recognition

John Deere
Smart Farming

Paypal
Fraud Detection

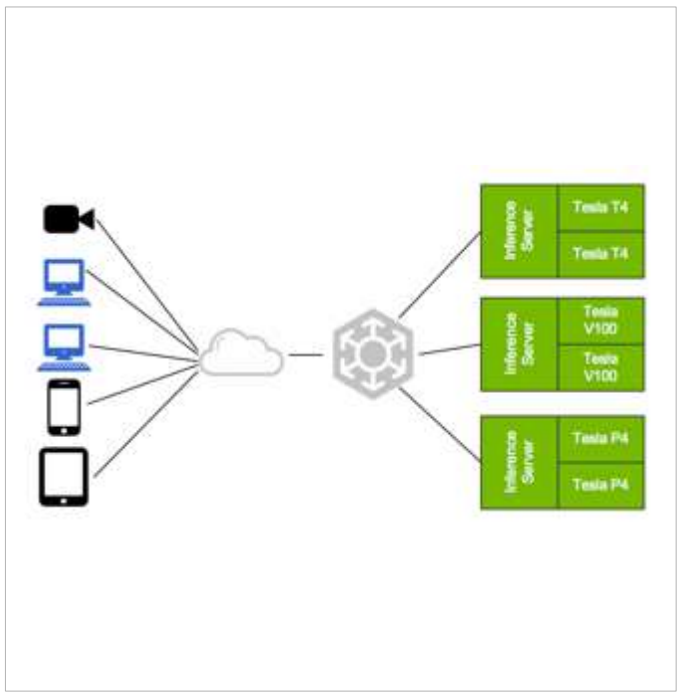
Pinterest
Visual Search

Snap
Recommendation

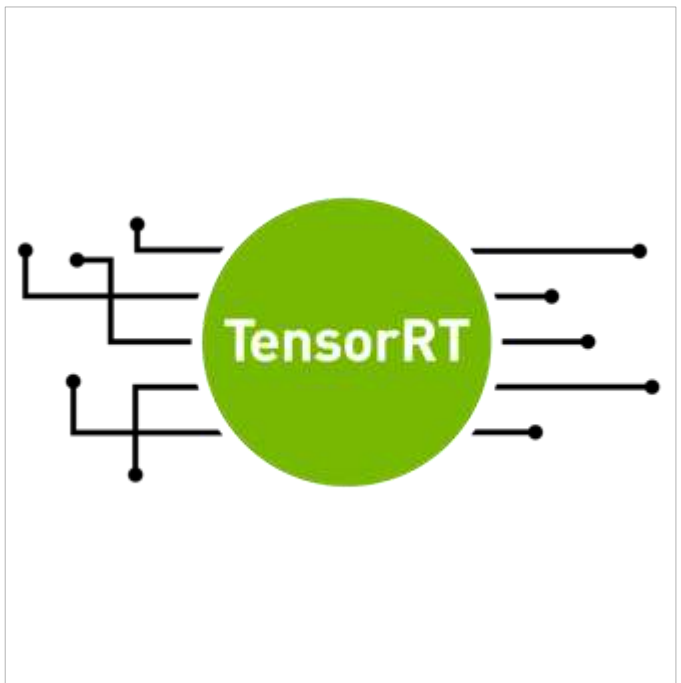
Twitter
Video Analytics

TENSORRT HYPERSCALE INFERENCE PLATFORM

WORLD'S MOST ADVANCED SCALE-OUT GPU



TENSORRT INFERENCE SERVER



INTEGRATED INTO TENSORFLOW & ONNX SUPPORT

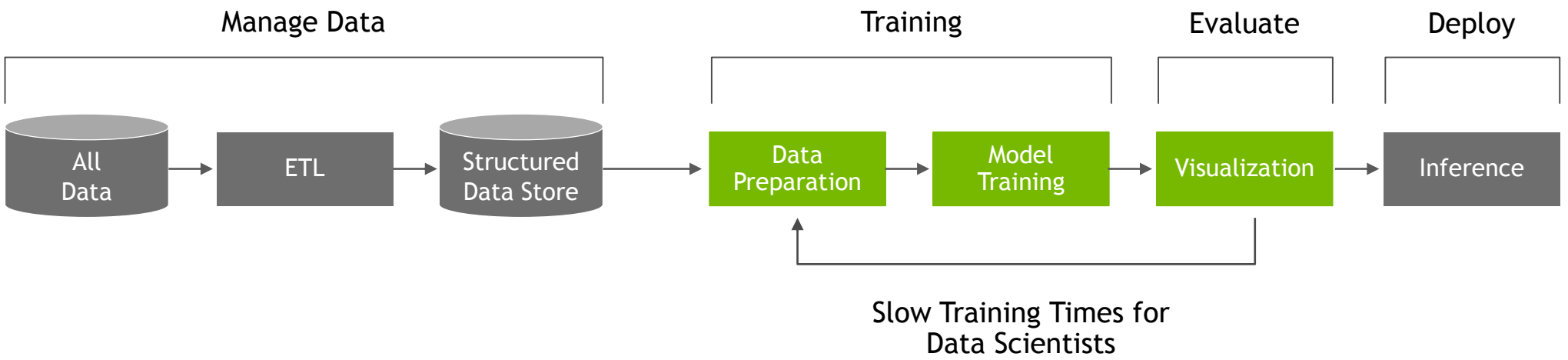
TESLA T4

WORLD'S MOST ADVANCED SCALE-OUT GPU

320 Turing Tensor Cores
2,560 CUDA Cores
65 FP16 TFLOPS | 130 INT8 TOPS | 260 INT4 TOPS
16GB | 320GB/s
70 W

MACHINE LEARNING
RAPIDS

THE BIG PROBLEM IN DATA SCIENCE



ACCELERATING MACHINE LEARNING

The RAPIDS Ecosystem

Open Source Community	Enterprise Data Science Platforms	Startups	Deep Learning Integration
ANACONDA. Apache Arrow GUNROCK PyData QUANTSIGHT learn URSA LABS	IBM ORACLE SAS SAP	BLAZINGDB FASTDATA graphistry H_2O.ai kinetica MAPR omnisci	Chainer PYTORCH

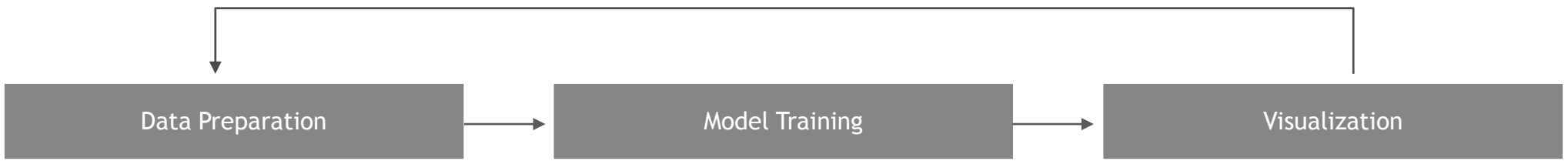
RAPIDS

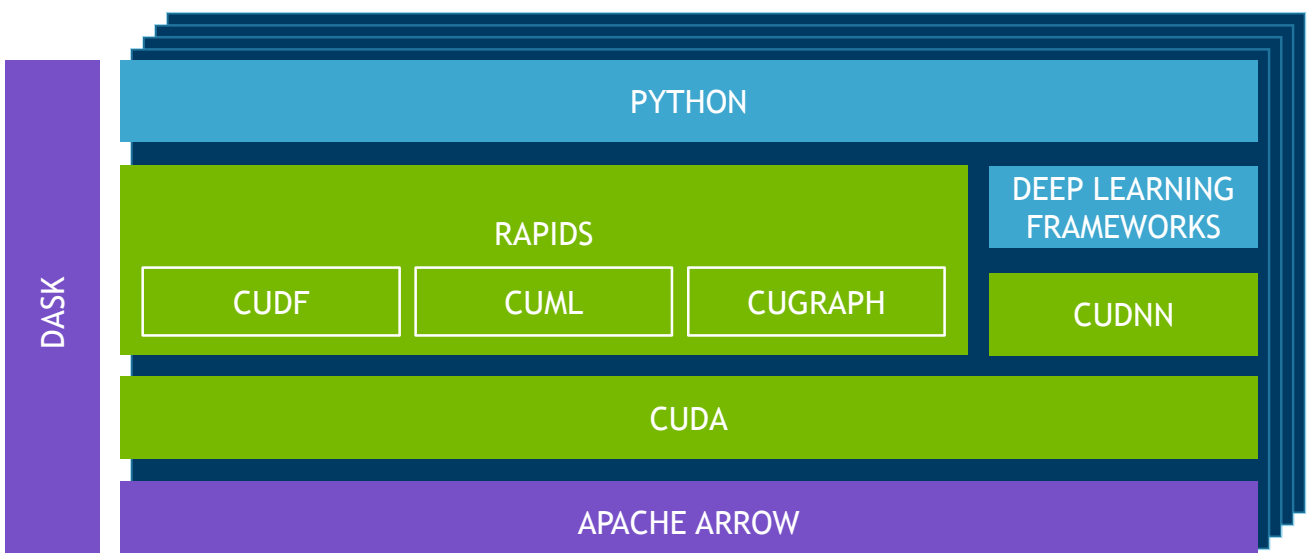
GPU Servers	Storage Partners
cisco DELL EMC Hewlett Packard Enterprise IBM Lenovo	DELL EMC IBM NetApp PURESTORAGE

NVIDIA

RAPIDS — OPEN GPU DATA SCIENCE

Software Stack





DRAMATICALLY MORE FOR YOUR MONEY

CPU-Only Cluster

300 Self-hosted Broadwell CPU Servers
180 KWatts

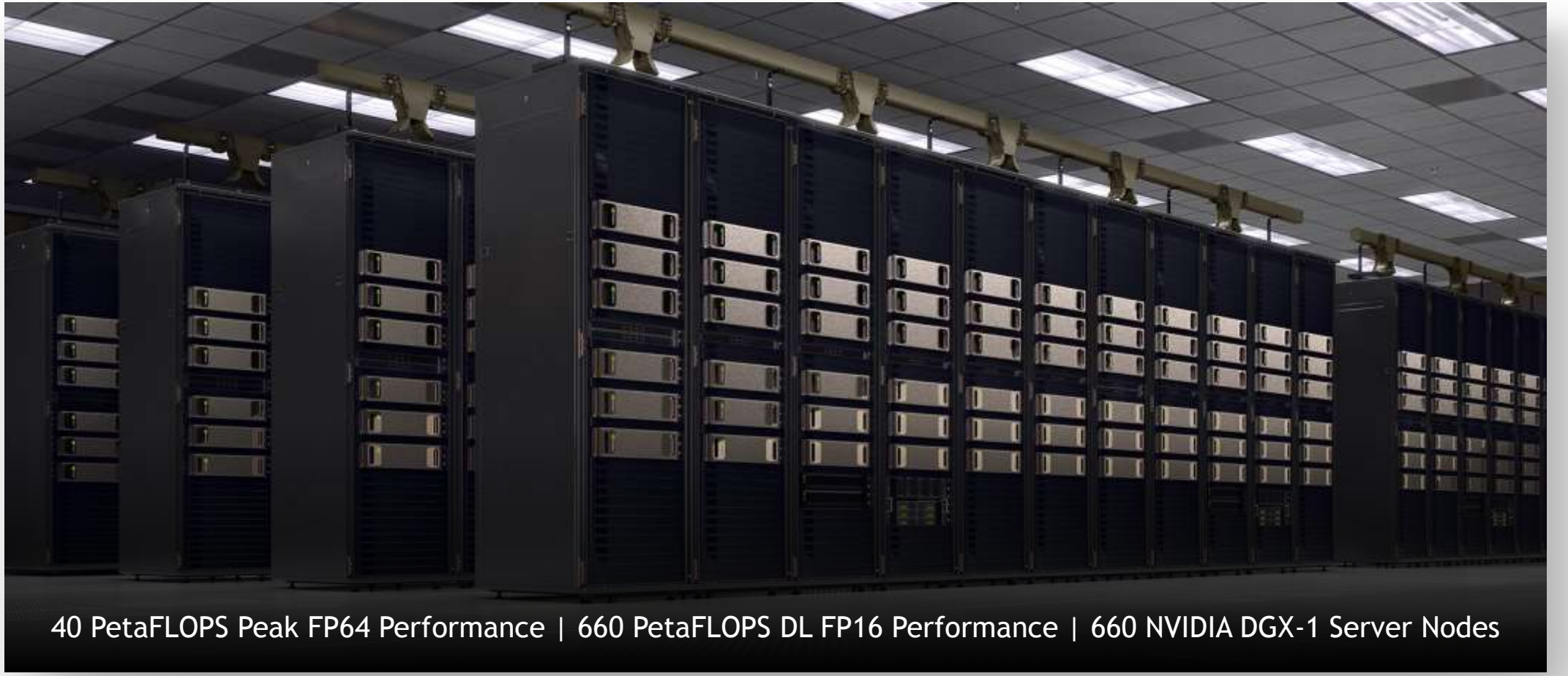
=

GPU-Accelerated

1 DGX-2
10 KWatts

DGX POD

ANNOUNCING
NVIDIA SATURNV WITH VOLTA



DGX POD — DGX-1

Reference Architecture in a Single 35 kW High-Density Rack

Fit within a standard-height 42 RU data center rack

- Nine DGX-1 servers (9 x 3 RU = 27 RU)
- Twelve storage servers (12 x 1 RU = 12 RU)
- 10 GbE (min) storage and management switch (1 RU)
- Mellanox 100 Gbps intra-rack high speed network switches (1 or 2 RU)

In real-life DL application development, one to two DGX-1 servers per developer are often required
One DGX POD supports five developers (AV workload)
Each developer works on two experiments per day
One DGX-1/developer/experiment/day*
*300,000 0.5M images * 120 epochs @ 480 images/sec Resnet-18 backbone detection network per experiment

DGX POD — DGX-2

Reference Architecture in a Single 35 kW High-Density Rack

Fit within a standard-height 48 RU data center rack

- Three DGX-2 servers (3 x 10 RU = 30 RU)
- Twelve storage servers (12 x 1 RU = 12 RU)
- 10 GbE (min) storage and management switch (1 RU)
- Mellanox 100 Gbps intra-rack high speed network switches (1 or 2 RU)

In real-life DL application development, one DGX-2 per developer minimizes model training time
One DGX POD supports at least three developers (AV workload)
Each developer works on two experiments per day
One DGX-2/developer/2 experiments/day*
*300,000 0.5M images * 120 epochs @ 480 images/sec Resnet-18 backbone detection network per experiment

NVIDIA GPU CLOUD (NGC)

DOWNLOAD AND DEPLOY

Source available on Github | Container available from NGC and Dockerhub | PIP available at a later date

GitHub NGC docker CONDA python Package Index

Source code, libraries, packages

On-premises

Cloud

50+ GPU-OPTIMIZED SOFTWARE CONTAINERS

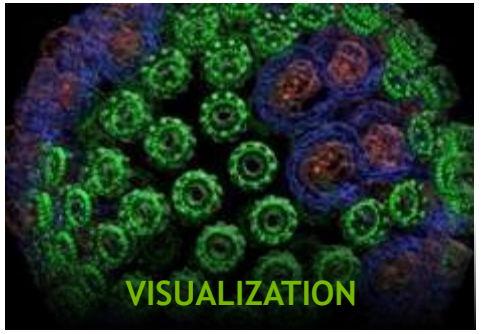
TensorFlow | PyTorch | more

RAPIDS | H2O | more

TensorRT | DeepStream | more

NAMD | GROMACS | more

Parabricks



ParaView | IndeX | more

NGC-READY SYSTEMS

VALIDATED FOR PERFORMANCE & FUNCTIONALITY OF NGC SOFTWARE

T4 & V100-ACCELERATED

*

*

*

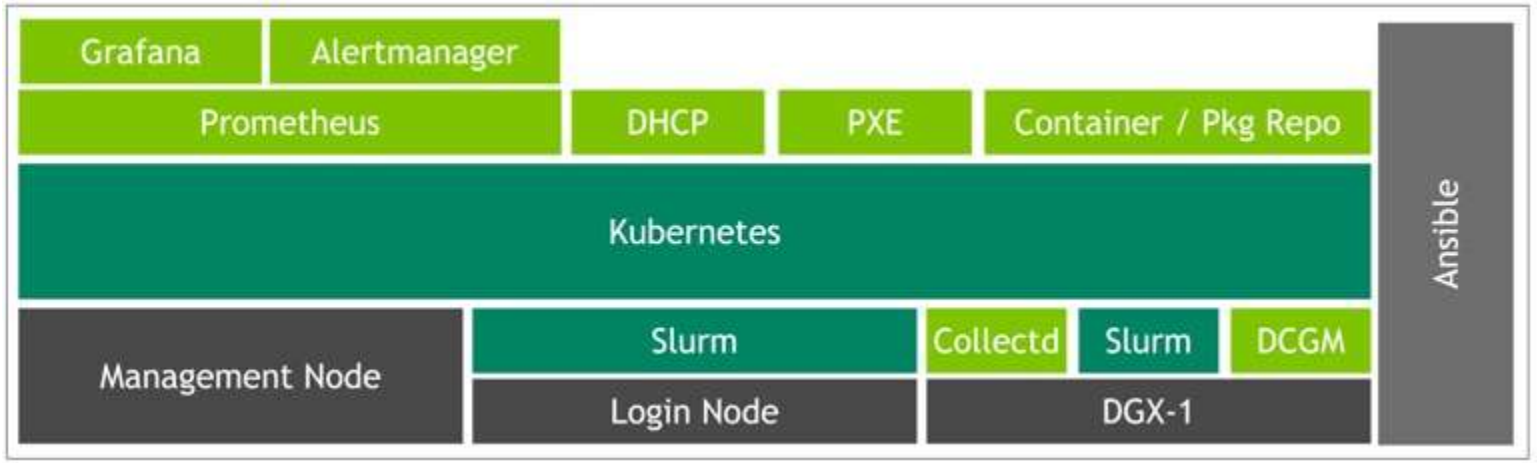
*

* Only V100 systems

DGX POD
MANAGEMENT
SOFTWARE

DGX POD MANAGEMENT SOFTWARE

For Large-Scale Multi-User AI Software Development Teams



SUPPORT PROGRAMS

IEEE – IPDPS 2019

20-24 de Maio, Rio de Janeiro

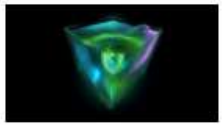
Keynote @ ScaDL Workshop
"Scalable Deep Learning over Parallel and Distributed Infrastructures"
24 de Maio

OpenACC
Hands-On Training
21 de Maio

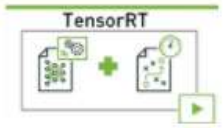
What's New

OptiX 5.0 Available OptiX 5 SDK features built-in support for motion blur, deep-learning based denoiser and more.

VRWorks 360 Video 1.1 Available This release bring many improvements listed below including new calibration technologies.



CUDA 9.1 Available New NPP functions for augmentation, multi-GPU enhancement in cuFFT, cuBLAS updates for Volta GPUs and more.



TensorRT 3 Available New TensorFlow model optimization, faster mixed-precision for CNNs and RNNs used for vision, speech & NLP.

See More

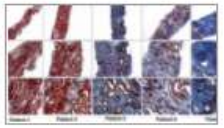
Developer News

AI Helps Farmers Distinguish Crop Data in Real Time
April 9, 2018

MIT Researchers Use AI to Capture Silent Speech
April 9, 2018

Drink up! Beer Tasting Robot Uses AI to Assess Quality
April 6, 2018

Researchers Develop AI System for License Plate Recognition
April 5, 2018



Boston University Researchers Use AI to Detect Kidney Disease
April 4, 2018

See More

Join the NVIDIA Developer Program

Access everything you need to develop with NVIDIA products.

REGISTER NOW

developer.nvidia.com

NVIDIA DEEP LEARNING INSTITUTE

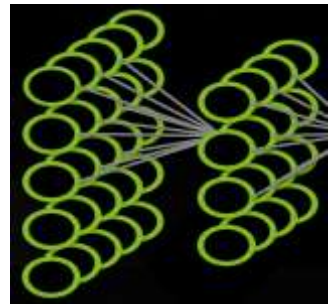
Hands-on self-paced and instructor-led training in deep learning and accelerated computing for developers

Request onsite instructor-led workshops at your organization: **www.nvidia.com/requestdli**

Take self-paced labs online: **www.nvidia.com/dlilabs**

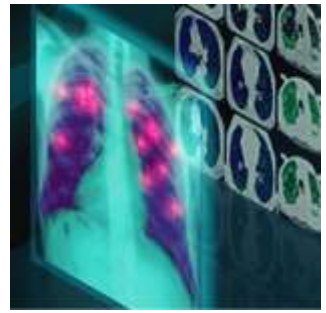
Download the course catalog, view upcoming workshops, and learn about the University Ambassador Program: **www.nvidia.com/dli**

Caffe2 · Microsoft Cognitive Toolkit · mxnet · TensorFlow · PYTORCH

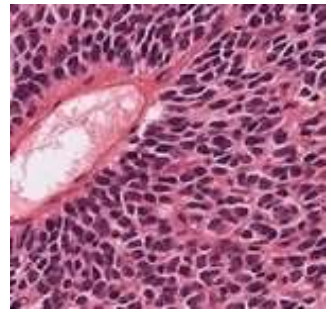


Deep Learning Fundamentals

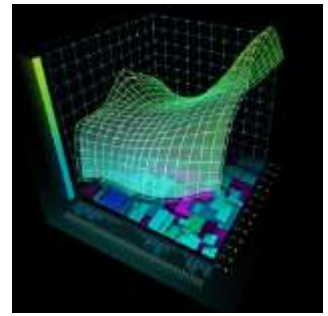
Autonomous Vehicles



Medical Image Analysis



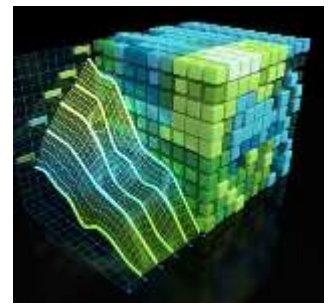
Genomics



Finance

Intelligent Video Analytics

Game Development & Digital Content



Accelerated Computing Fundamentals

More industry-specific training coming soon…

NVIDIA HW GRANT PROGRAM

Titan V Volta

- Scientific Computing
- HPC
- Deep Learning

Quadro P6000

- Scientific Visualization
- Virtual Reality

Jetson TX2 (Dev Kit)

- Robotics
- Autonomous Machines

https://developer.nvidia.com/academic_gpu_seeding

Obrigado
Gracias
Thank you
pcruzesilva@nvidia.com
NVIDIA.