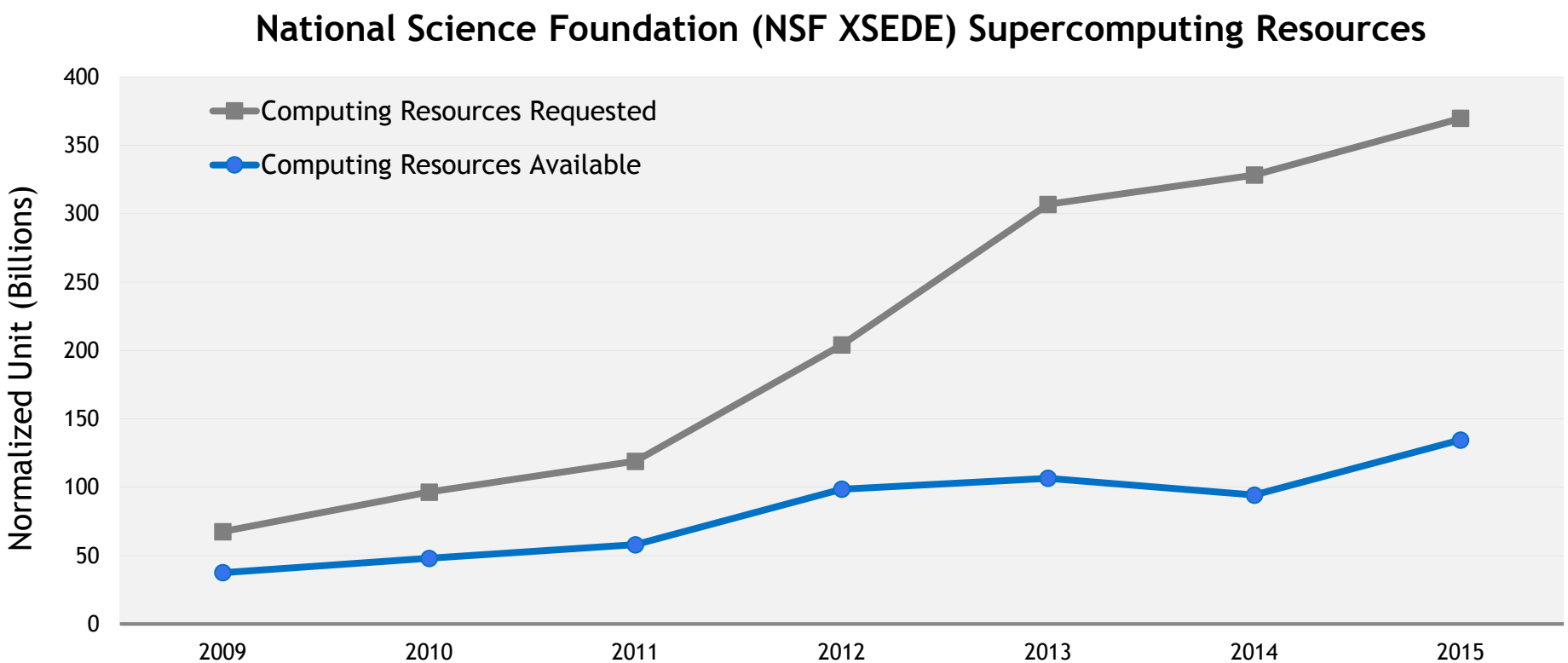
"ACCELERATING PYTHON DATA SCIENCE WITH NVIDIA RAPIDS"

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

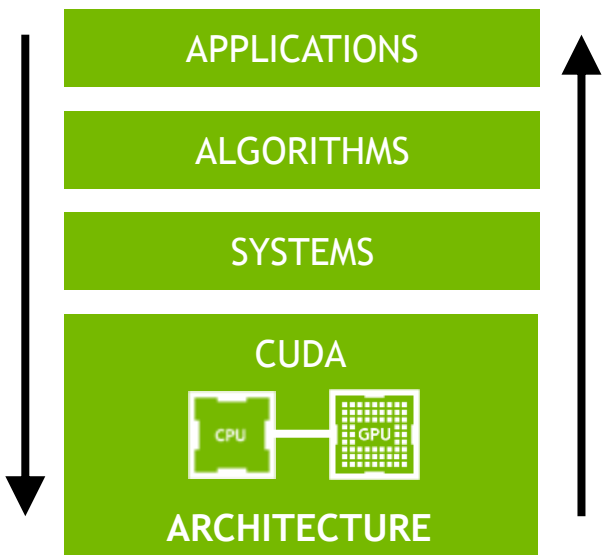
200B CORE HOURS OF LOST SCIENCE

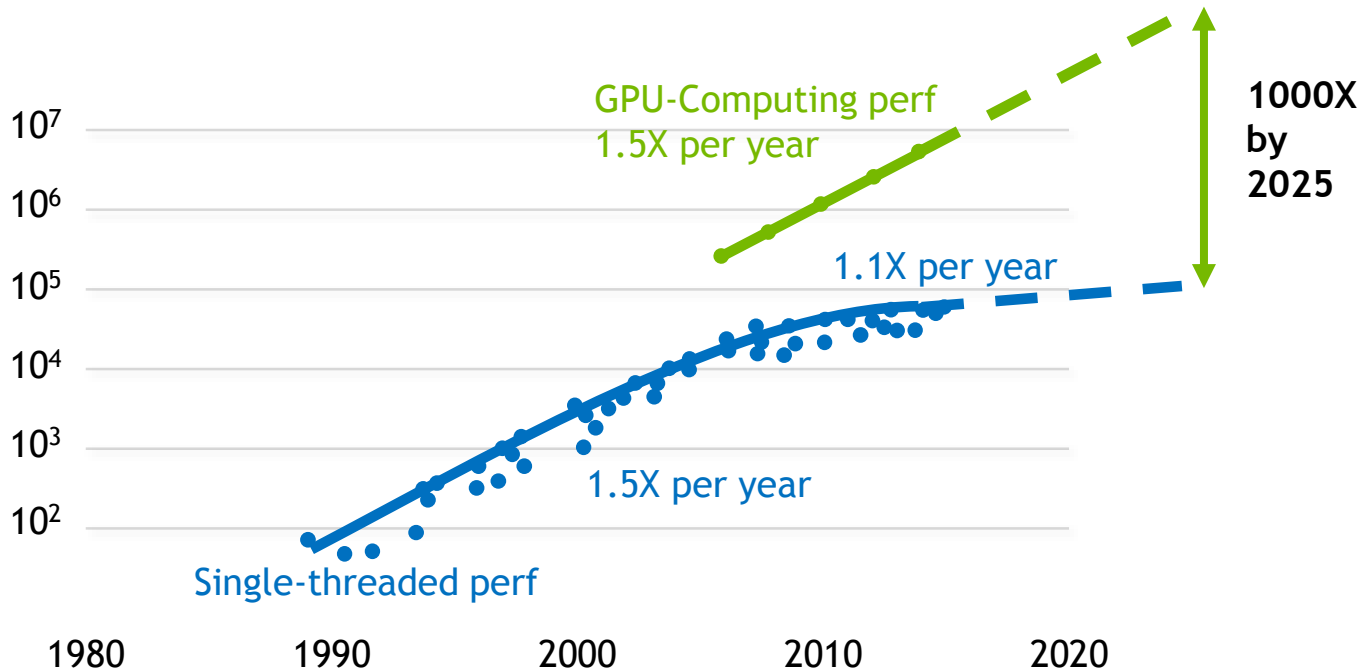
Data Center Throughput is the Most Important Thing for HPC



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

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

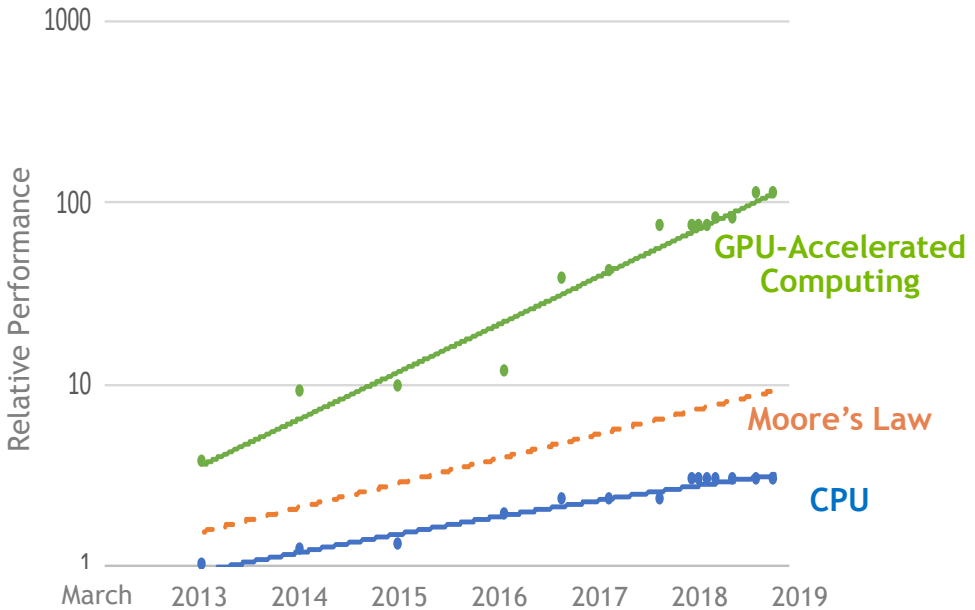
BEYOND MOORE'S LAW

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

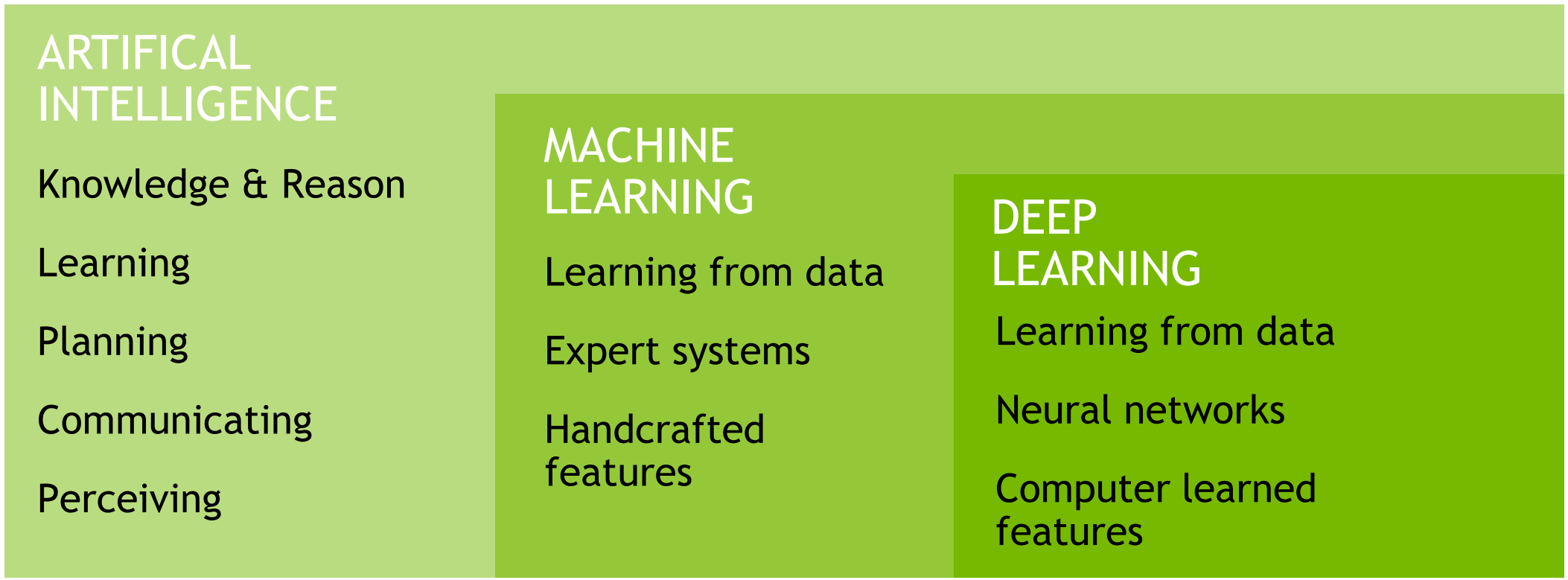
DELL · Hewlett Packard Enterprise · IBM
SYSTEM OEM

aws · Google Cloud Platform · Microsoft Azure
CLOUD


DEEP LEARNING

LEARNING FROM DATA

AND SOME BUZZ WORDS



A NEW COMPUTING MODEL

TRAINING

Training Data

Input

"Label"
Output

Trained Neural Network

INFERENCE

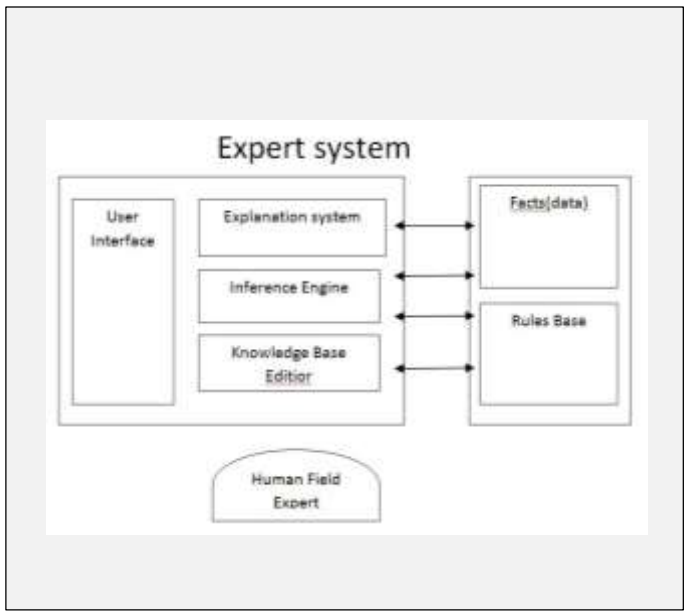
Input

Trained Neural Network

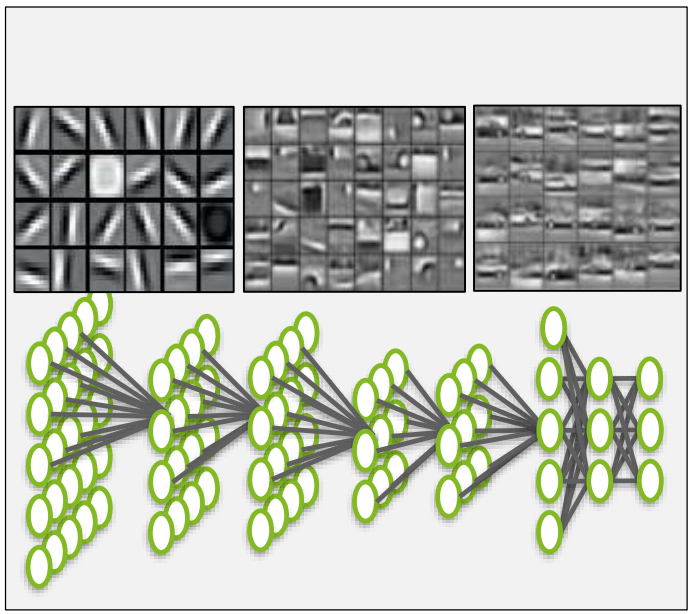
"Label"
Output

A NEW COMPUTING MODEL

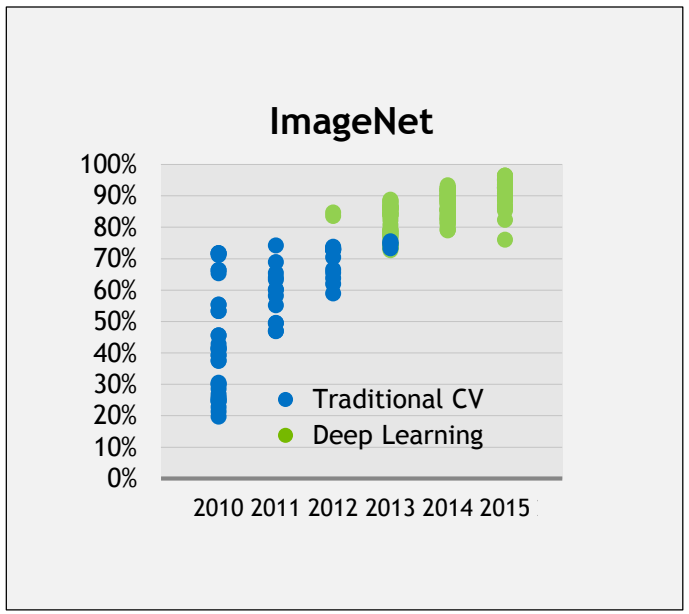
Outperform experts, facts, rules with software that writes software



Traditional Computer Vision
Experts + Time



Deep Learning Object Detection
DNN + Data + GPU



Deep Learning Achieves
“Superhuman” Results

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.

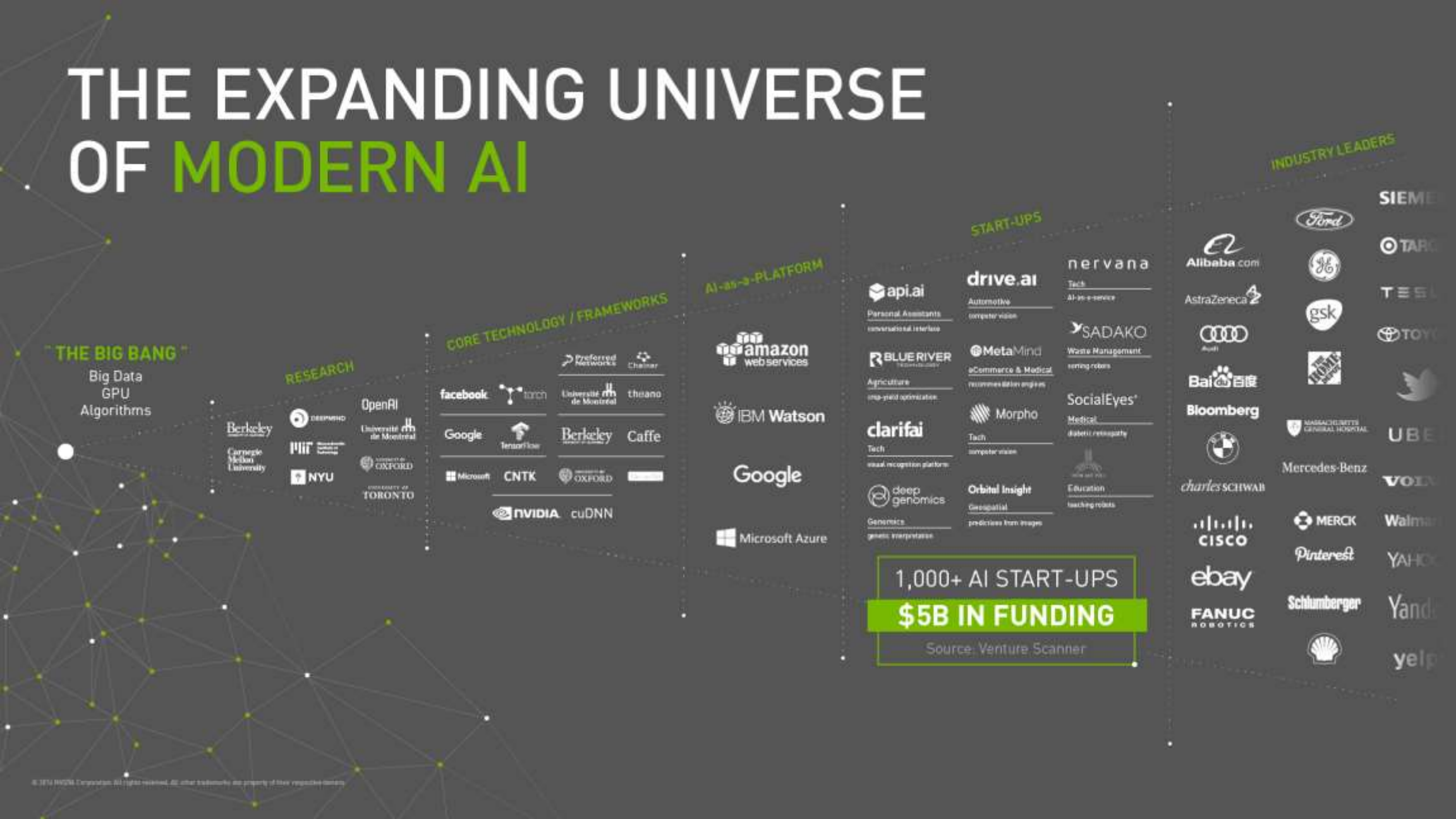
Such advances make for exciting times in

TESLA REVOLUTIONIZES DEEP LEARNING

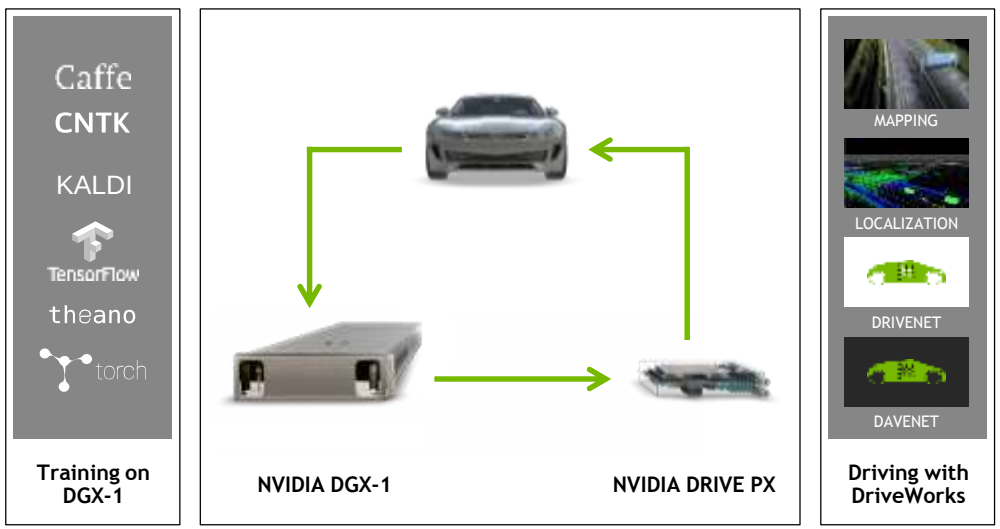
GOOGLE BRAIN APPLICATION

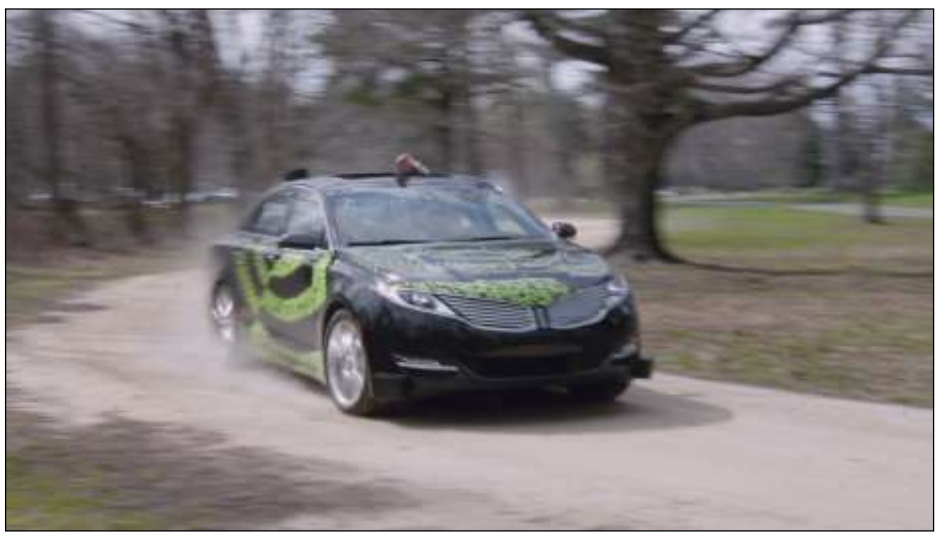
	BEFORE TESLA	AFTER TESLA
Cost	$5,000K	$200K
Servers	1,000 Servers	16 Tesla Servers
Energy	600 KW	4 KW
Performance	1x	6x

THE EXPANDING UNIVERSE OF MODERN AI



NEW AI DRIVING



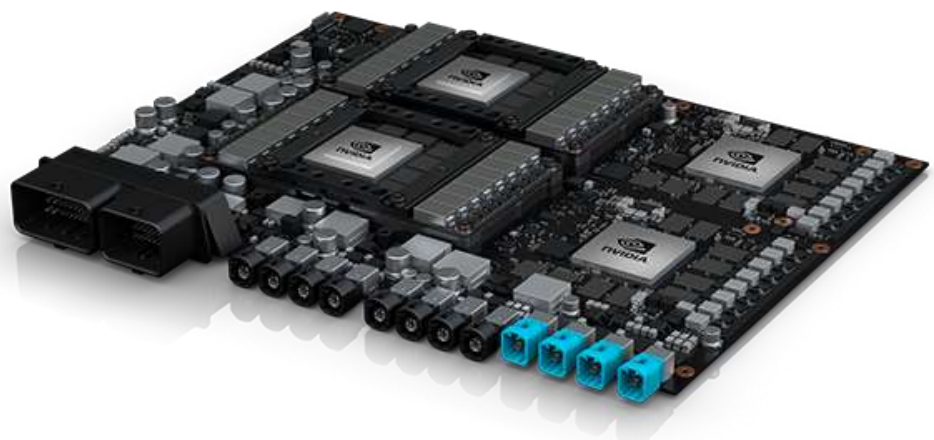


WATCH VIDEO

NVIDIA DRIVE PEGASUS

First AI Computer to Make Robotaxis a Reality

WATCH VIDEO

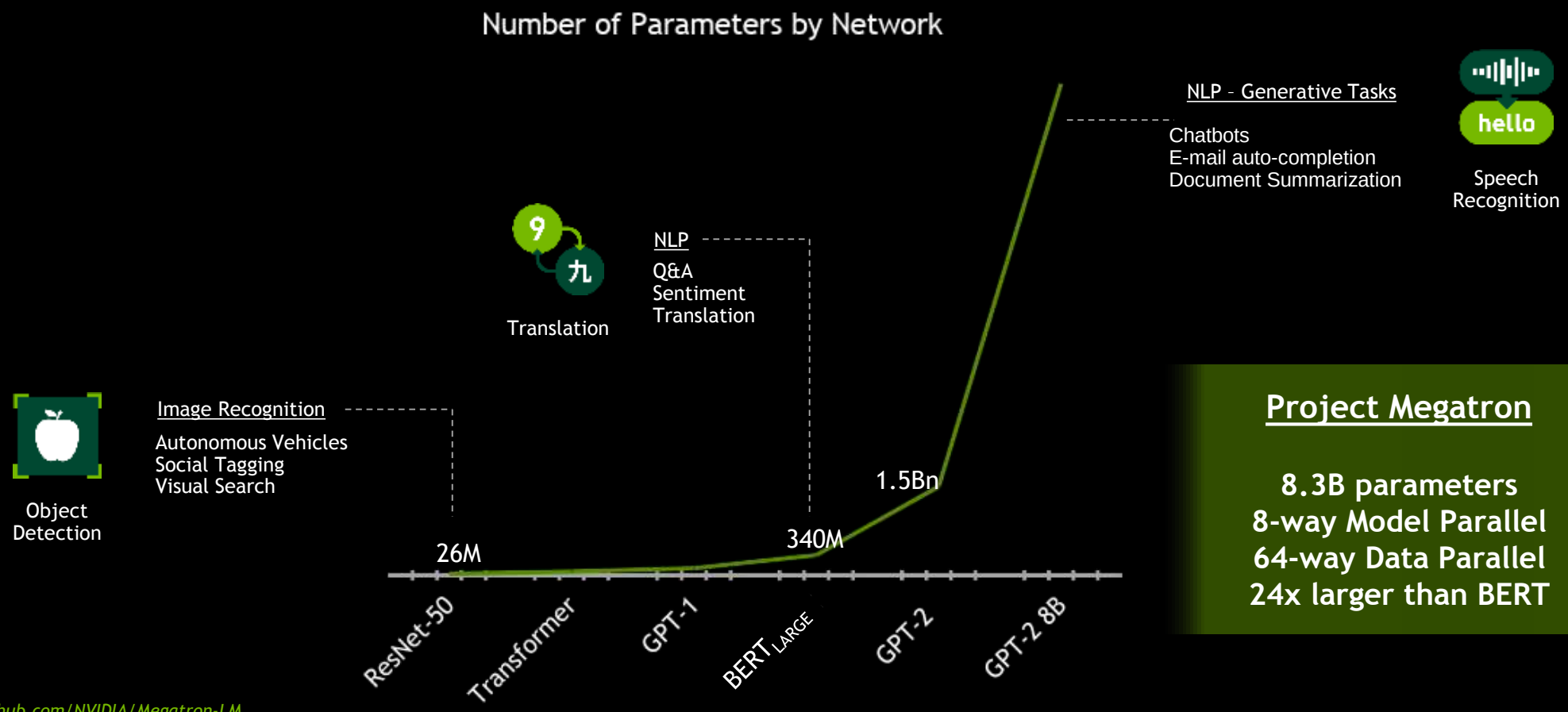


Case Study: BERT

(Bidirectional Encoder Representations from Transformer)

Deep Learning Models Increasing in Complexity

Next-Level Use-Cases Require Gigantic Models



Project Megatron

8.3B parameters
8-way Model Parallel
64-way Data Parallel
24x larger than BERT

https://github.com/NVIDIA/Megatron-LM

NVIDIA BREAKS RECORDS IN AI PERFORMANCE

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

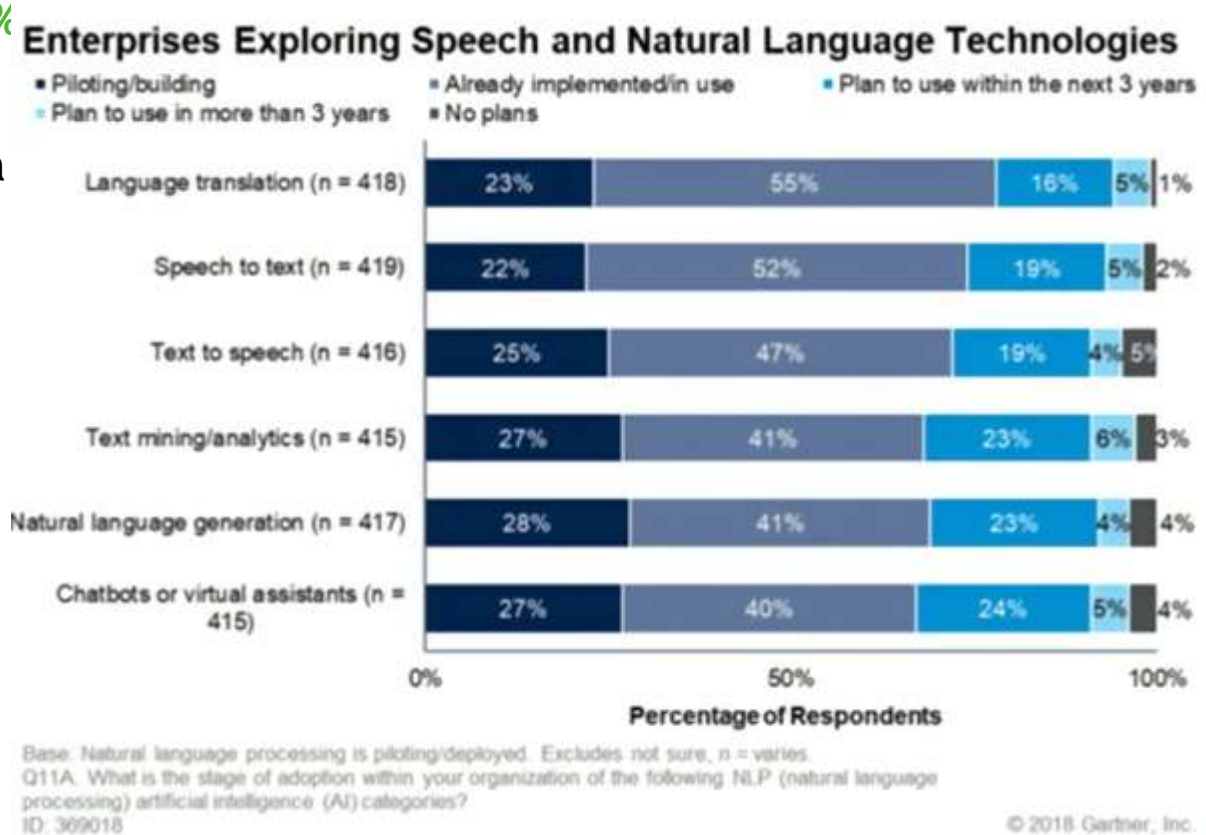
Per Accelerator comparison using reported performance for MLPerf 0.6 NVIDIA DGX-2H (16 V100s) compared to other submissions at same scale except for MiniGo where NVIDIA DGX-1 (8 V100s) submission was used| MLPerf ID Max Scale: Mask R-CNN: 0.6-23, GNMT: 0.6-26, MiniGo: 0.6-11 | MLPerf ID Per Accelerator: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

Enterprise NLP Trend

Unstructured content represents as much as **80%** of enterprise information resources.

A recent Gartner Research Circle survey on data and analytics trends shows that organizations are **actively developing text analytics** as part of their data and analytics strategies.

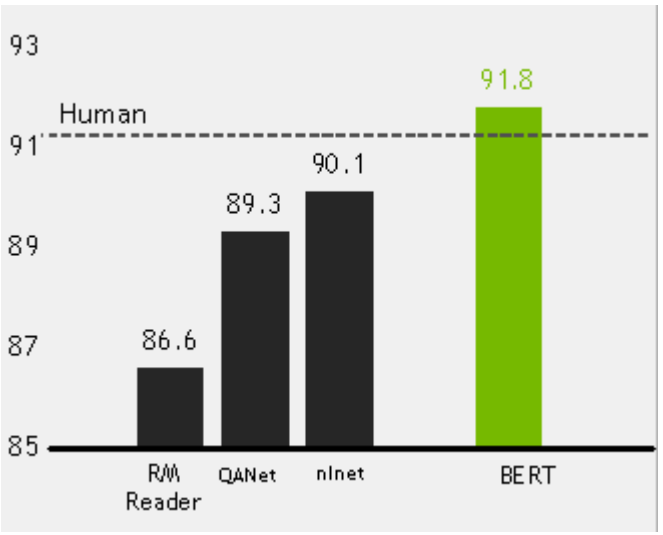
80% of survey respondents either have text analytics in use or plan to use it within the next **two years**.



BERT: Flexibility + Accuracy for NLP Tasks

“BERT is a method of pre-training language representations, meaning that we train a general-purpose "language understanding" model on a large text corpus (like Wikipedia), and then use that model for downstream NLP tasks that we care about (like question answering).

BERT outperforms previous methods because it is the first *unsupervised*, *deeply bidirectional* system for pre-training NLP.”



Super Human Question & Answering

9th October, Google submitted GLUE benchmark

- **Sentence Pair Classification:** MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG
- **Single Sentence Classification:** SST-2, CoLA
- **Question Answering:** SQuAD
- **Single Sentence Tagging:** CoNLL - 2003 NER

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$BERT_{BASE}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**


DGX REFERENCE
ARCHITECURES

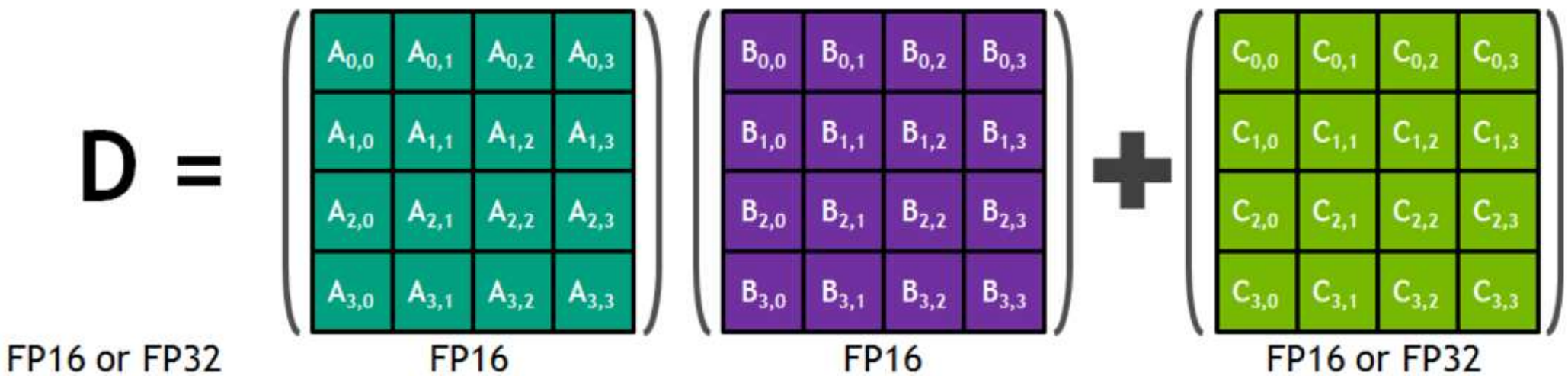
TESLA V100 TENSOR CORE GPU

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

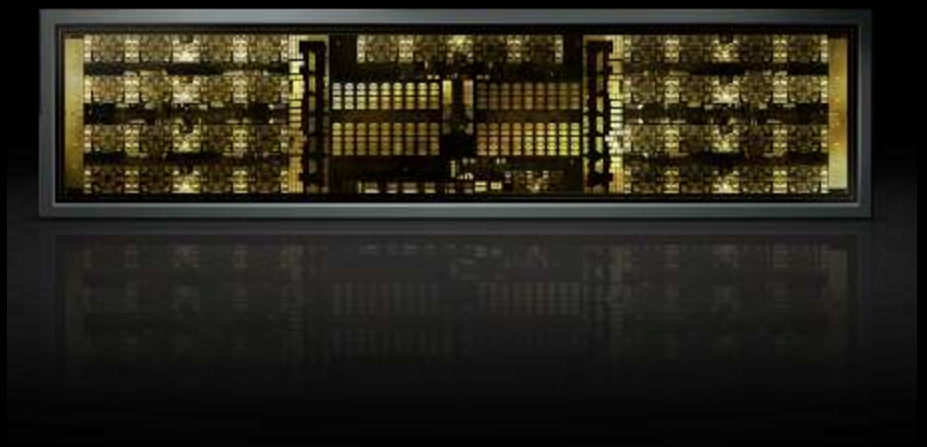


NVIDIA® DGX-1™

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

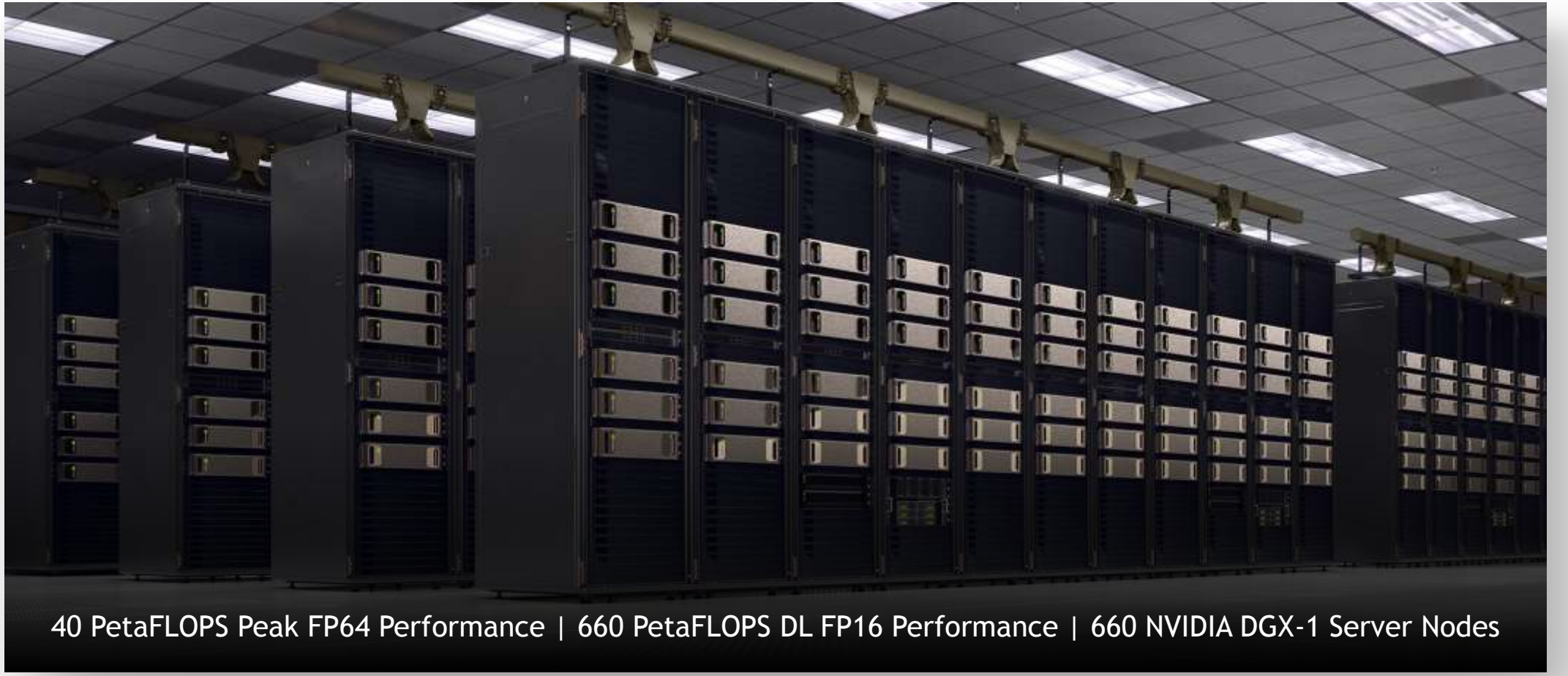


NVIDIA DGX-2

THE LARGEST GPU EVER CREATED

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

DGX-POD
NVIDIA SATURNV WITH VOLTA



NVIDIA DGX SUPERPOD

Terabit-Speed InfiniBand Networking per Node

Mellanox EDR 100G InfiniBand Network

Mellanox Smart Director Switches

In-Network Computing Acceleration Engines

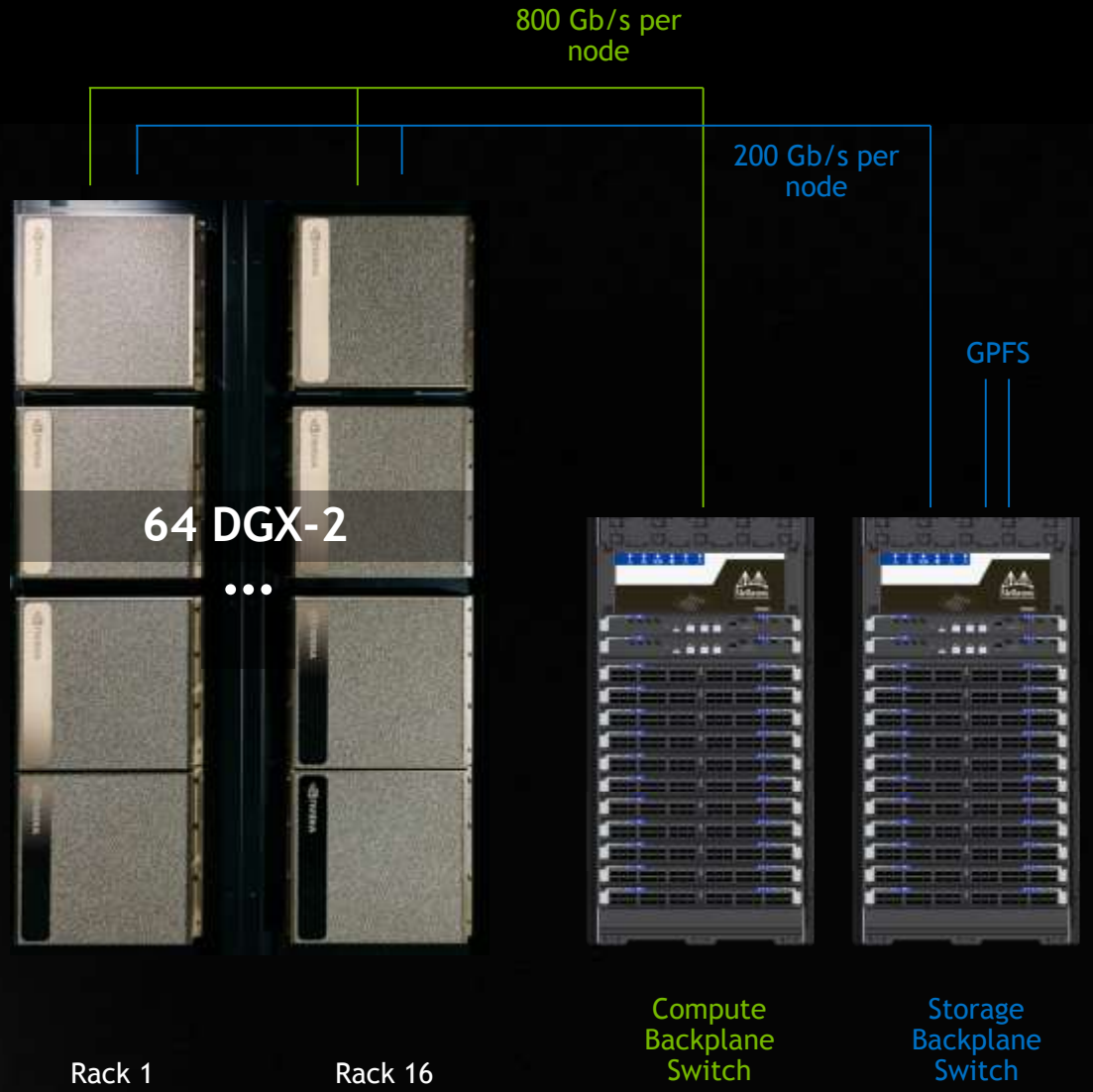
Fast and Efficient Storage Access with RDMA

Up to 130Tb/s Switching Capacity per Switch

Ultra-Low Latency of 300ns

Integrated Network Manager

White paper:
https://nvidia.highspot.com/items/5d073ad681171721086b2788




PLATFORM FOR AI
INFERENCE

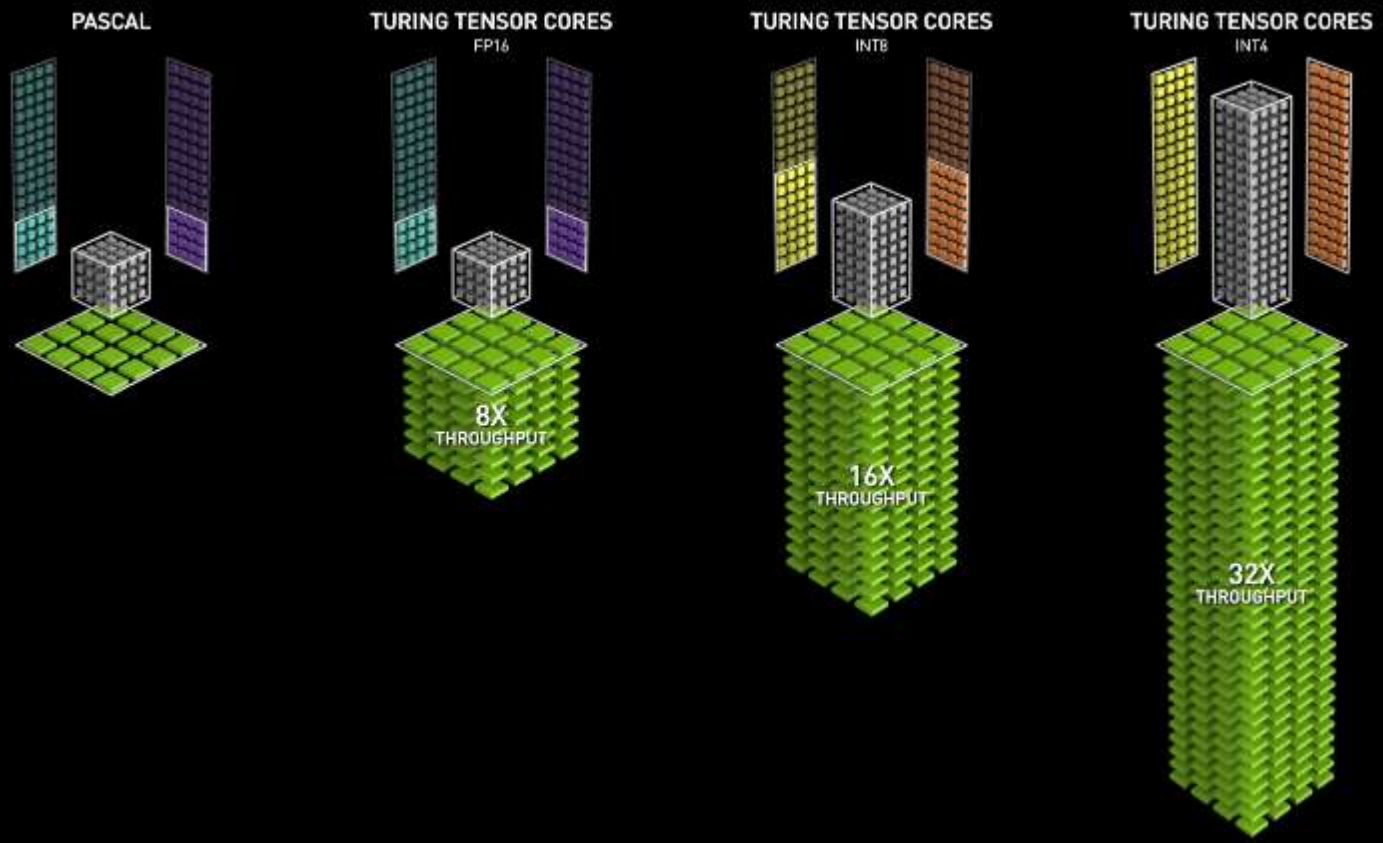
TESLA T4

WORLD'S MOST EFFICIENT GPU FOR MAINSTREAM SERVERS

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

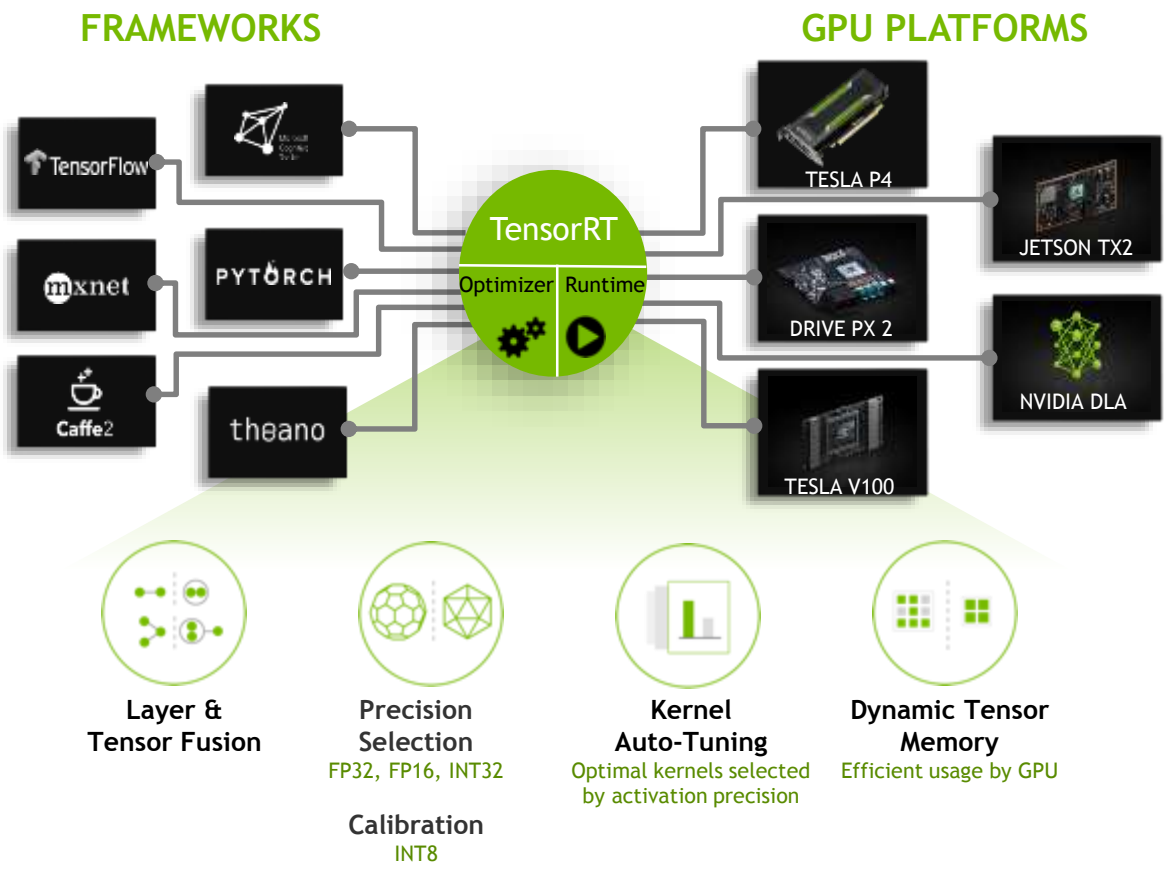
NEW TURING TENSOR CORE

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



NVIDIA TensorRT 5 INFERENCE PLATFORM

Accelerates Throughput On Leading Industry Platforms



Automotive

Drive

Embedded

Jetson

Data center

Tesla

NVIDIA GPUs IN THE CLOUD

AVAILABLE ON-DEMAND FROM THE TOP CLOUD SERVICE PROVIDERS

- Immediate access to NVIDIA GPU infrastructure for data science in the cloud
- Wide variety of deployment and management options using containers, Kubernetes, Kubeflow, support for cloud native services, and more

Alibaba Cloud

aws

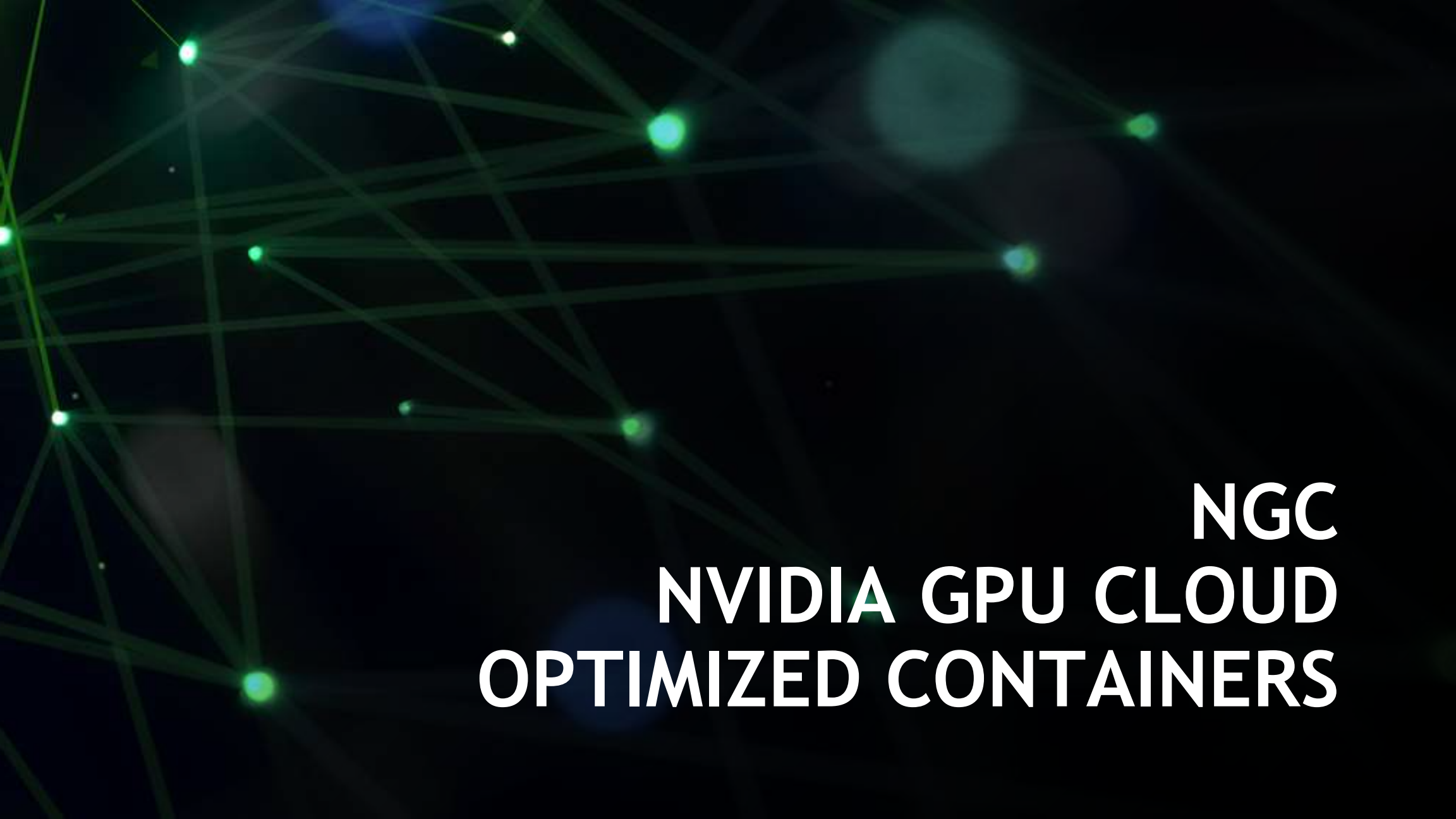
Baidu Cloud

Google Cloud

Microsoft Azure

ORACLE Cloud Infrastructure

OVH.com


NGC
NVIDIA GPU CLOUD
OPTIMIZED CONTAINERS

NGC: GPU-OPTIMIZED SOFTWARE HUB

Ready-to-run GPU Optimized Software, Anywhere

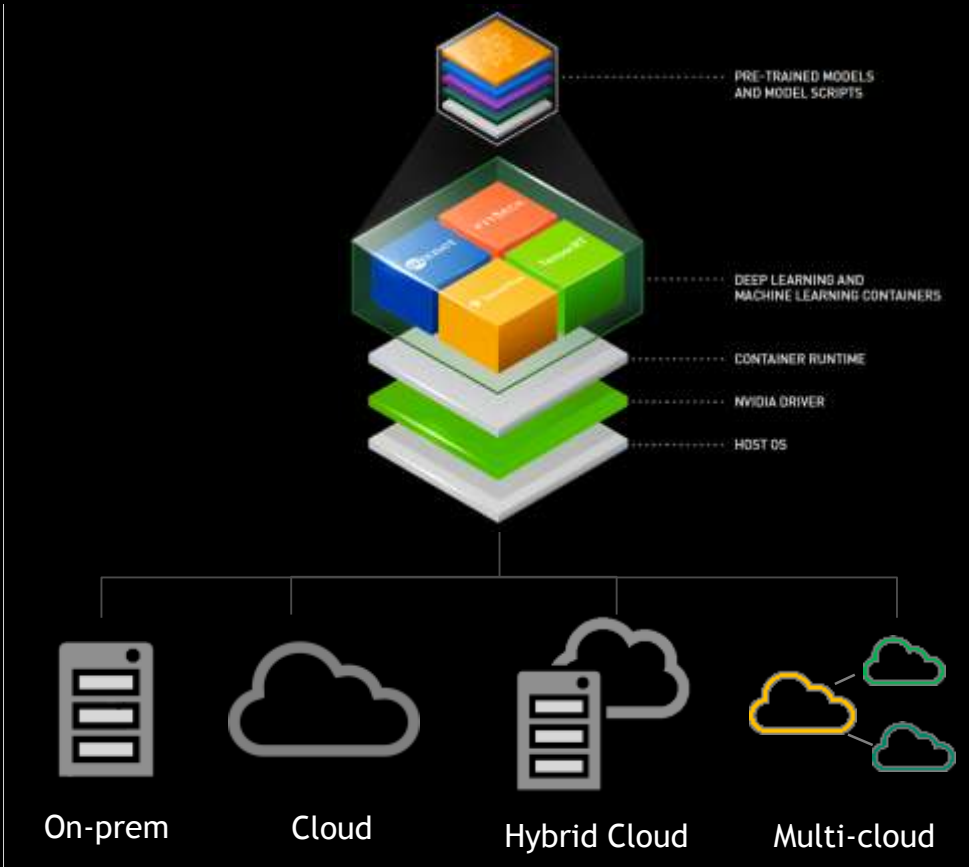
50+ Containers
DL, ML, HPC

15+ Model Training Scripts
NLP, Image Classification, Object Detection & more

NGC

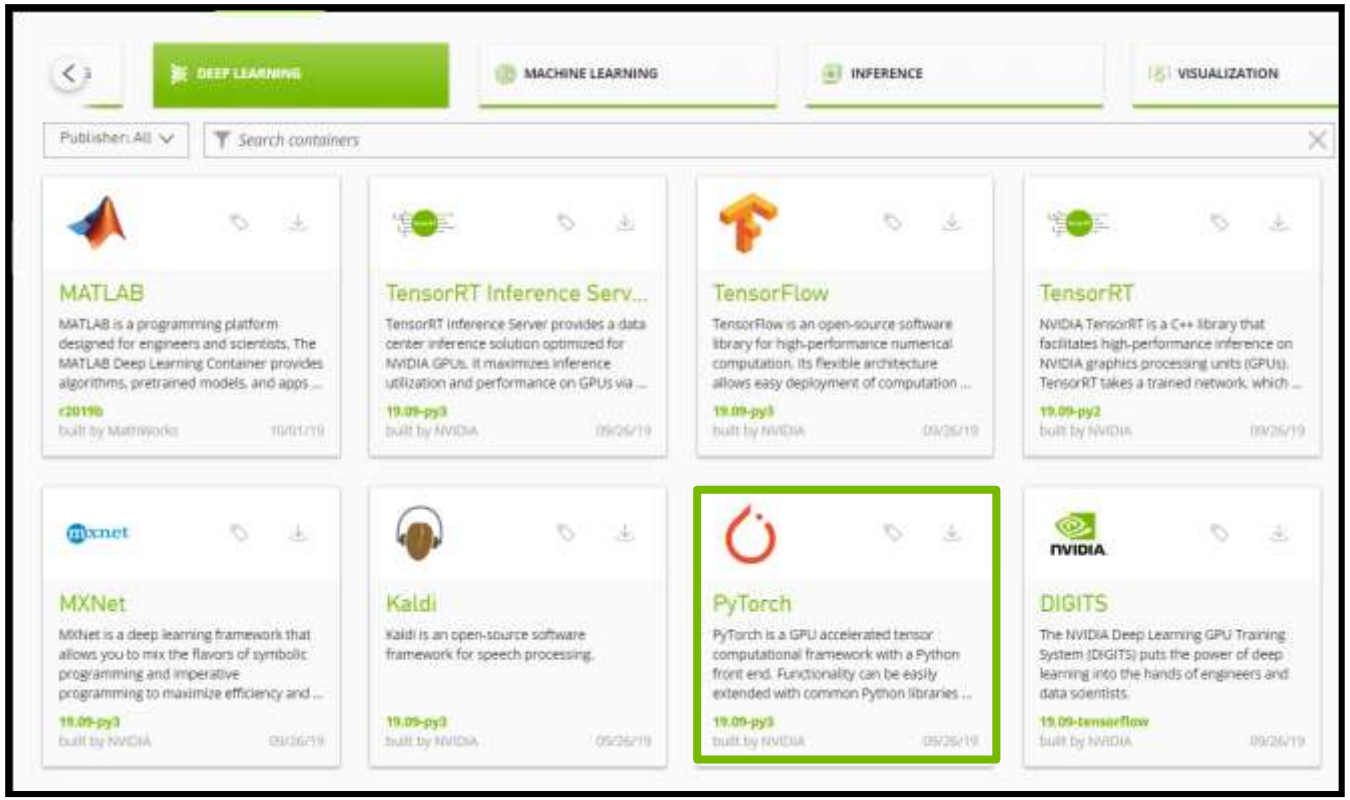
60 Pre-trained Models
NLP, Image Classification, Object Detection & more

Industry Workflows
Medical Imaging, Intelligent Video Analytics



NVIDIA GPU CLOUD

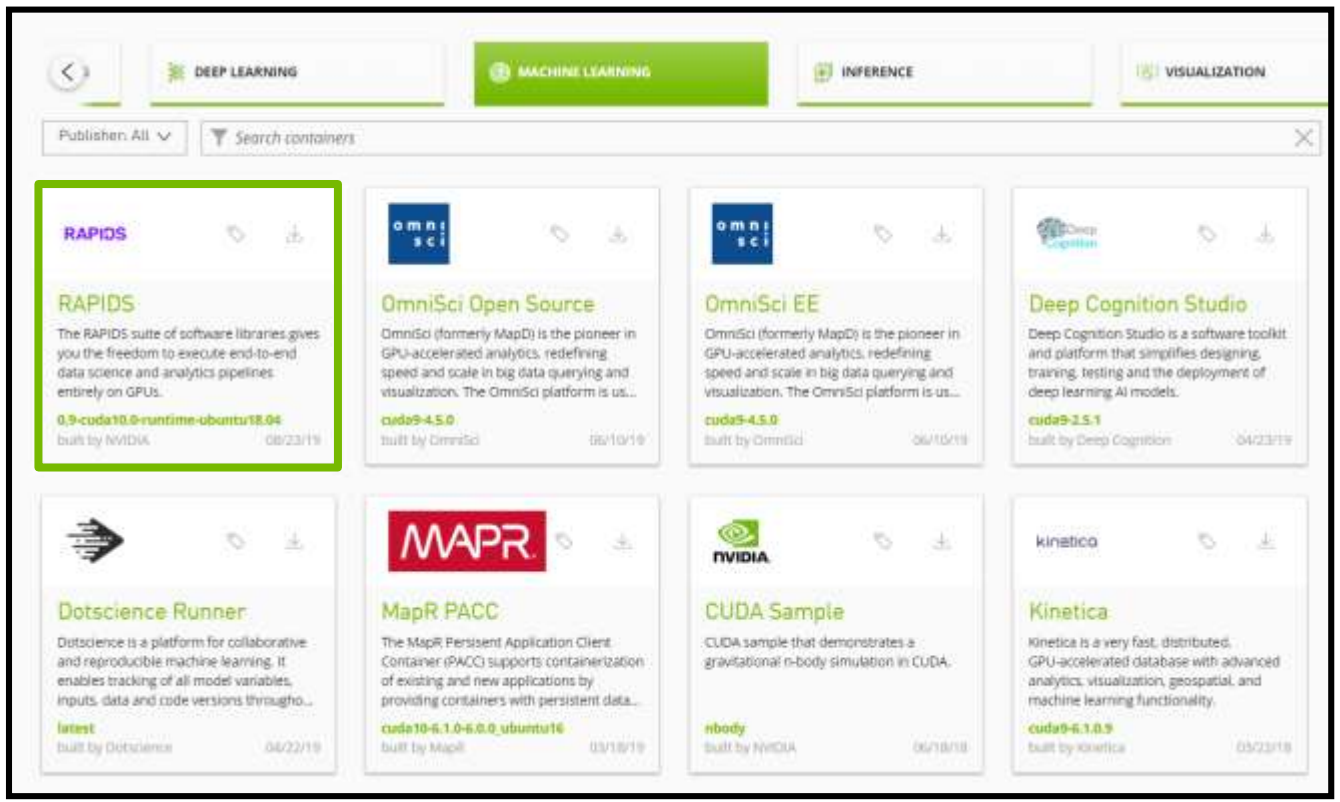
Deep Learning Containers



docker pull nvcr.io/nvidia/pytorch:19.09-py3
nvidia-docker run nvcr.io/nvidia/pytorch:19.09-py3

NVIDIA GPU CLOUD

Machine Learning Containers



docker pull nvcr.io/nvidia/rapidsai/rapidsai:0.9-cuda10.0-runtime-ubuntu18.04
nvidia-docker run nvcr.io/nvidia/rapidsai/rapidsai:0.9-cuda10.0-runtime-ubuntu18.04


EDGE COMPUTING

JETSON SUCCESS STORIES

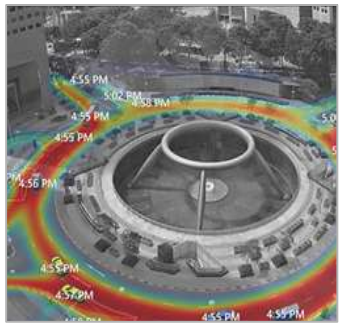
Industrial

Aerospace/Defense

Healthcare

Construction

Agriculture



Smart City

Retail

Logistics

Inventory Mgmt

Delivery

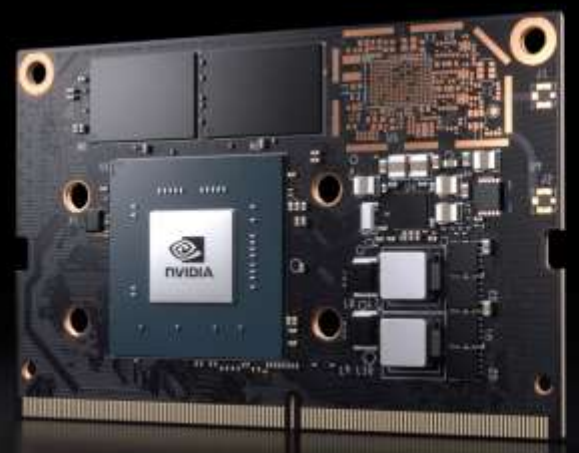
Inspection

Service

ANNOUNCING: JETSON NANO

Small, low-power AI Computer

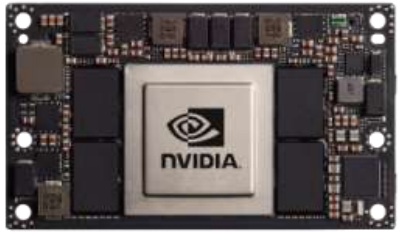
128 CUDA Cores | 4 Core CPU
4 GB Memory
472 GFLOPs
70x45mm
5W | 10W
$129



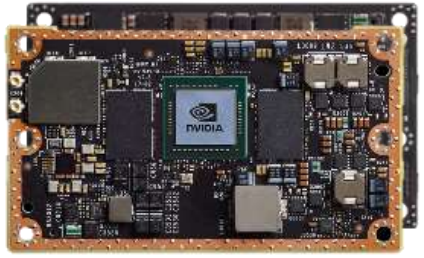
THE JETSON FAMILY

From AI at the Edge to Autonomous Machines

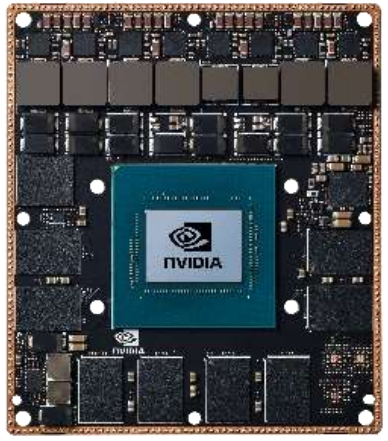
JETSON NANO
5 - 10W
0.5 TFLOPS (FP16)
45mm x 70mm
$129



JETSON TX1 → JETSON TX2 4 GB
7 - 15W
1 – 1.3 TFLOPS (FP16)
50mm x 87mm
$299



JETSON TX2 8GB | Industrial
7 – 15W
1.3 TFLOPS (FP16)
50mm x 87mm
$399 - $749



JETSON AGX XAVIER
10 – 30W
10 TFLOPS (FP16) | 32 TOPS (INT8)
100mm x 87mm
$1099

AI at the edge — Fully autonomous machines

Multiple devices - Same software

Listed prices are for 1000u+ | Full specs at developer.nvidia.com/jetson

JETSON NANO DEVELOPER KIT

AI Computer

128 CUDA Cores | 4 Core CPU

472 GFLOPs

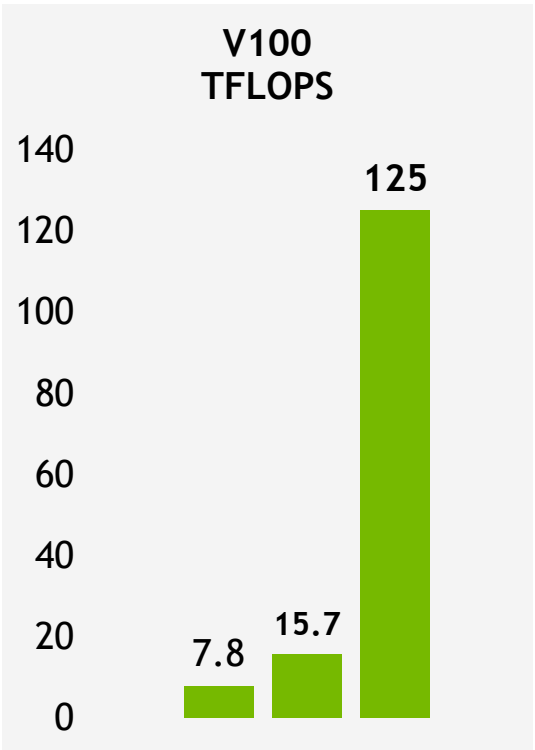
5W | 10W

Available from nvidia.com and distributors worldwide

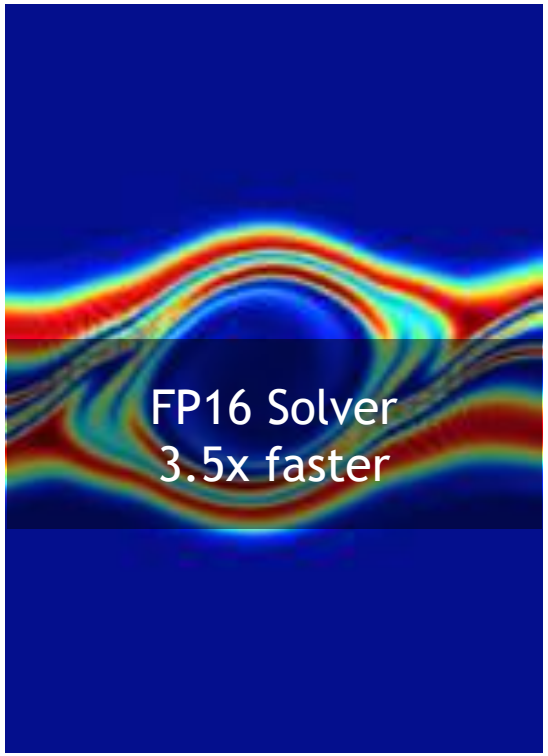

DIGITAL SCIENCE
HPC + AI + DATA

TENSOR CORES FOR SCIENCE

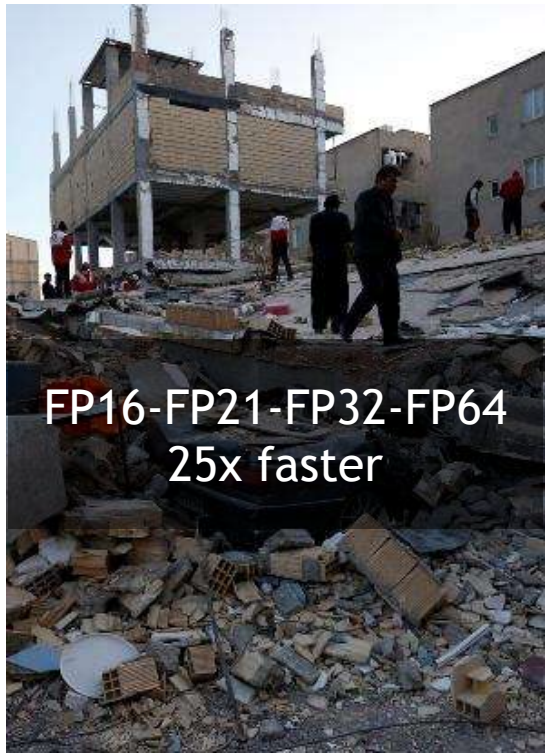
Mixed-Precision Computing



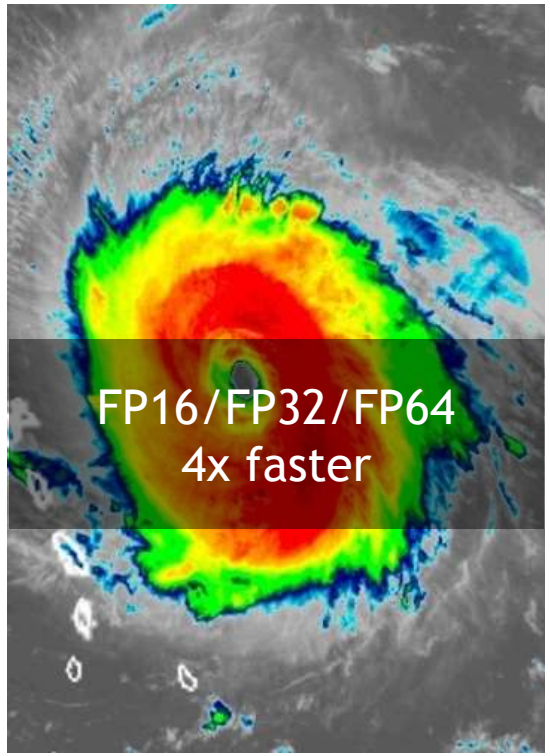
FP64+ MULTI-PRECISION



PLASMA FUSION APPLICATION



EARTHQUAKE SIMULATION



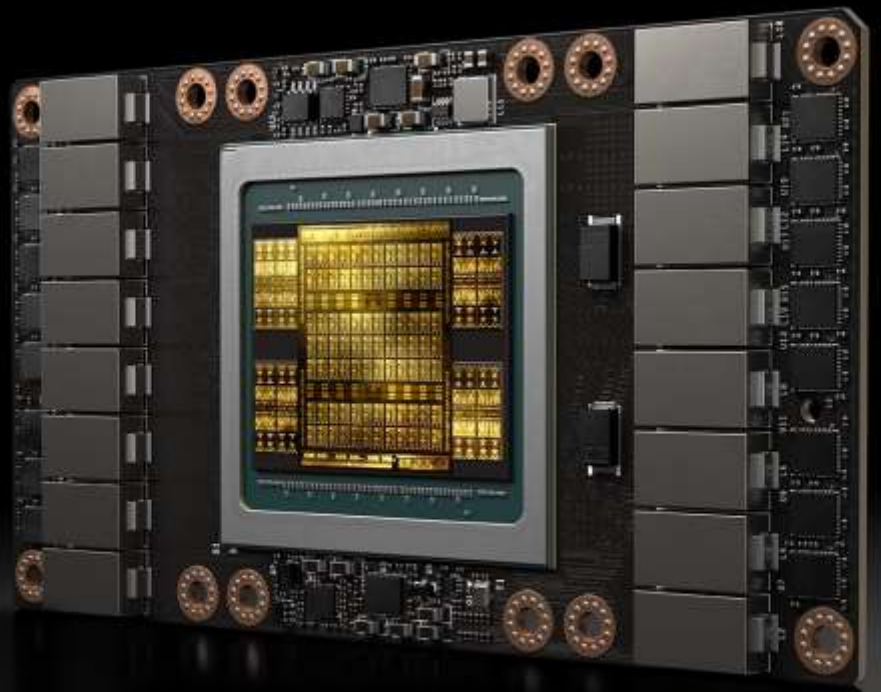
MIXED PRECISION WEATHER PREDICTION

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 TODAY'S FASTEST SUPERCOMPUTERS

22 of Top 25 Greenest

ORNL Summit
World's Fastest
27,648 GPUs| 149 PF

LLNL Sierra
World's 2nd Fastest
17,280 GPUs| 95 PF

Piz Daint
Europe's Fastest
5,704 GPUs| 21 PF

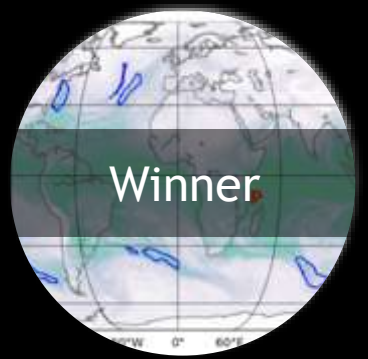
Total Pangea 3
Fastest Industrial
3,348 GPUs| 18 PF

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

NVIDIA POWER GORDON BELL WINNERS & 5 OF 6 FINALISTS

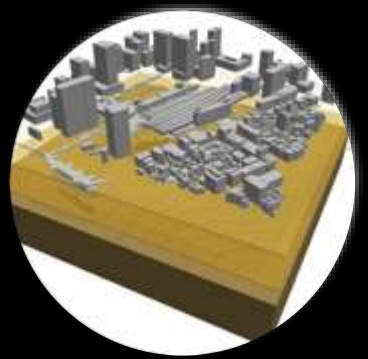
BERKELEY LAB
OAK RIDGE National Laboratory

Genomics
2.36 ExaFLOPS



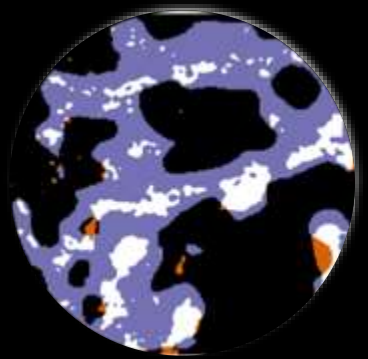
BERKELEY LAB
OAK RIDGE National Laboratory
NVIDIA

Weather
1.13 ExaFLOPS



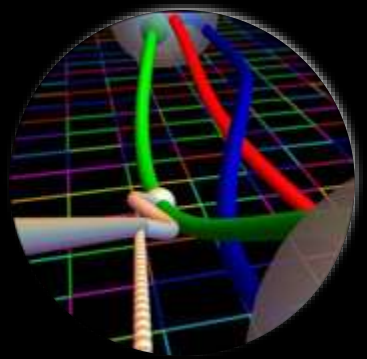
東京大学 THE UNIVERSITY OF TOKYO
OAK RIDGE National Laboratory

Seismic
1st Soil & Structure Simulation



OAK RIDGE National Laboratory

Material Science
300X Higher Performance



BERKELEY LAB
Lawrence Livermore National Laboratory
NVIDIA

Quantum Chromodynamics
<1% of Uncertainty Margin

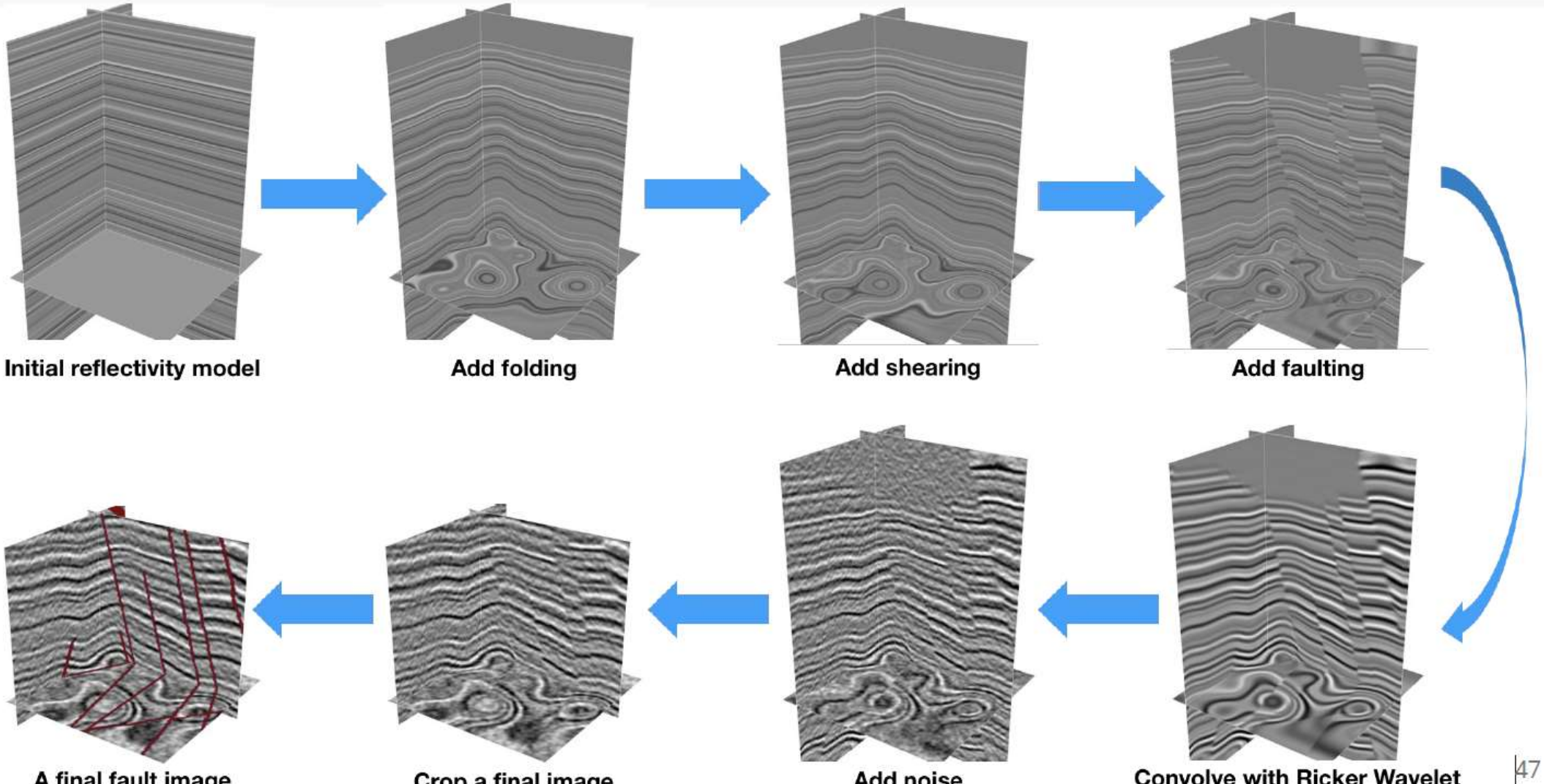
1º LATIN AMERICA SUPER-COMPUTER: PETROBRAS FENIX

Petrobras's supercomputer Fênix is among the world's 500 biggest computers and ranks first in Latin America. The list was compiled by Top500.org, based on the machines' performance in data processing, and features Fênix at the 142nd position worldwide.

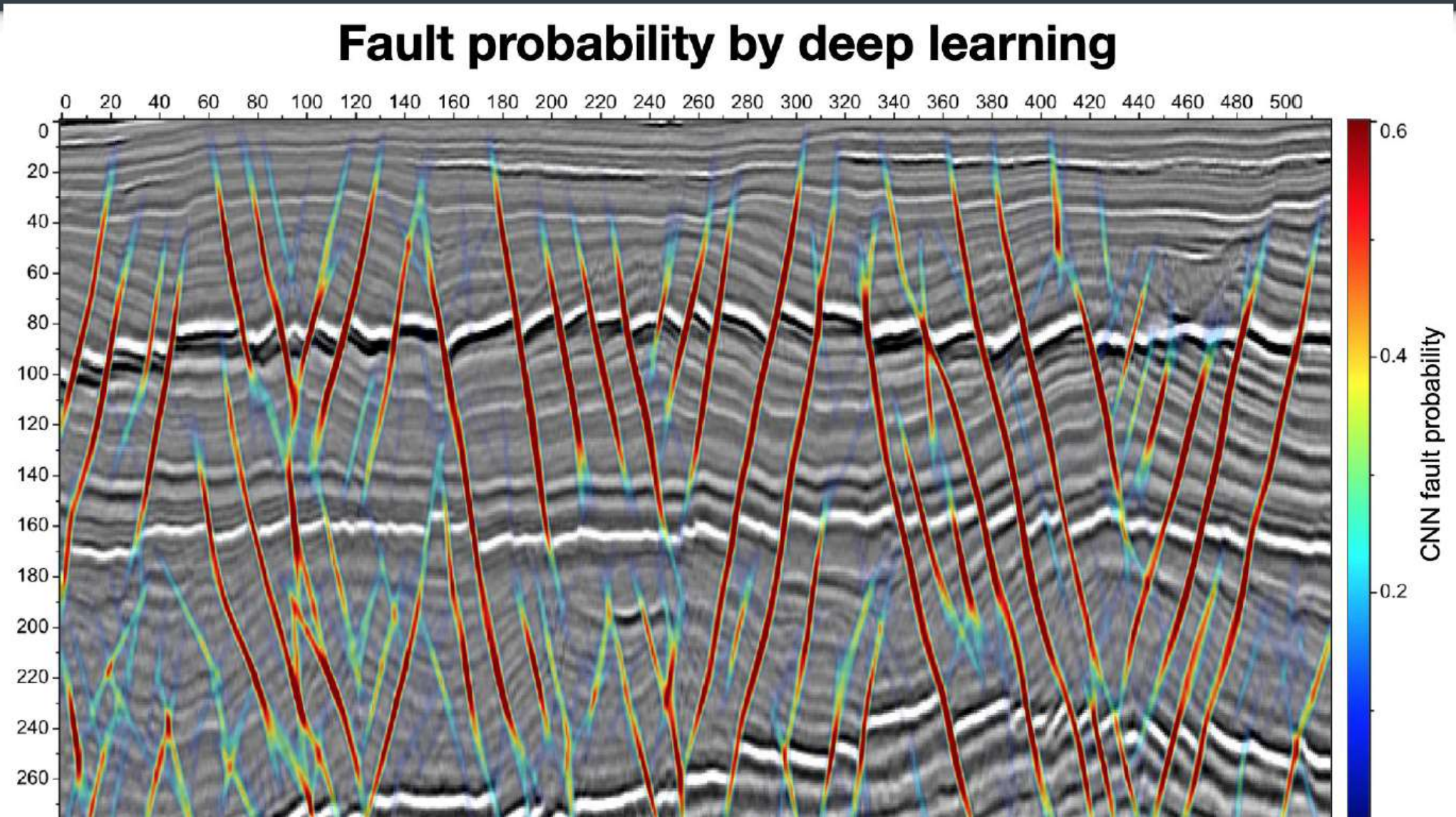
NVIDIA

BR PETROBRAS

Workflow to generate 3D fake multi-fault sample

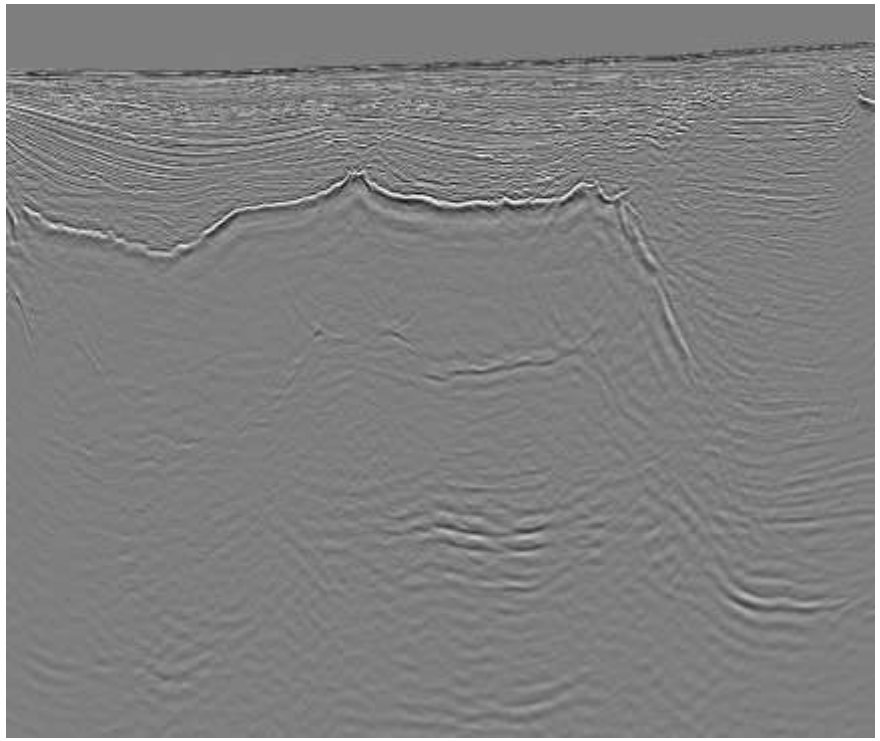


Simple fault classification

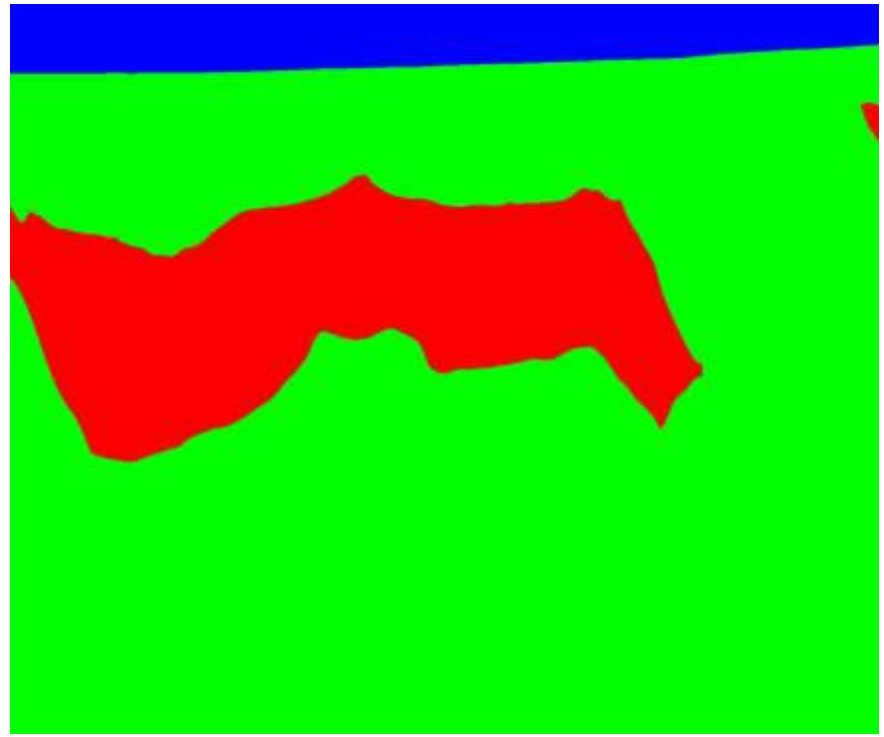


SEISMIC INTERPRETATION

New Deep Learning approaches - Salt



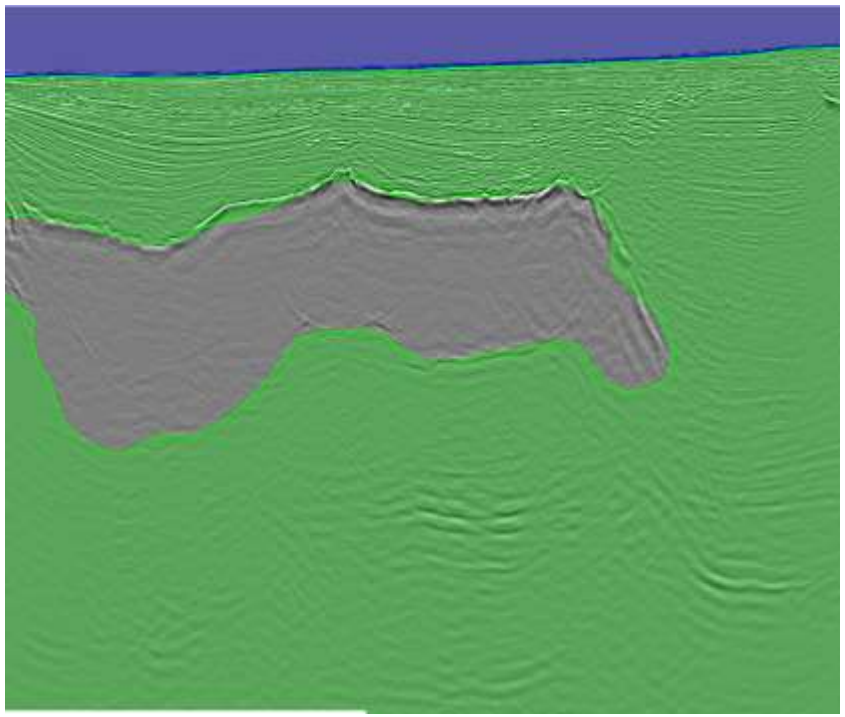
Features (Seismic)



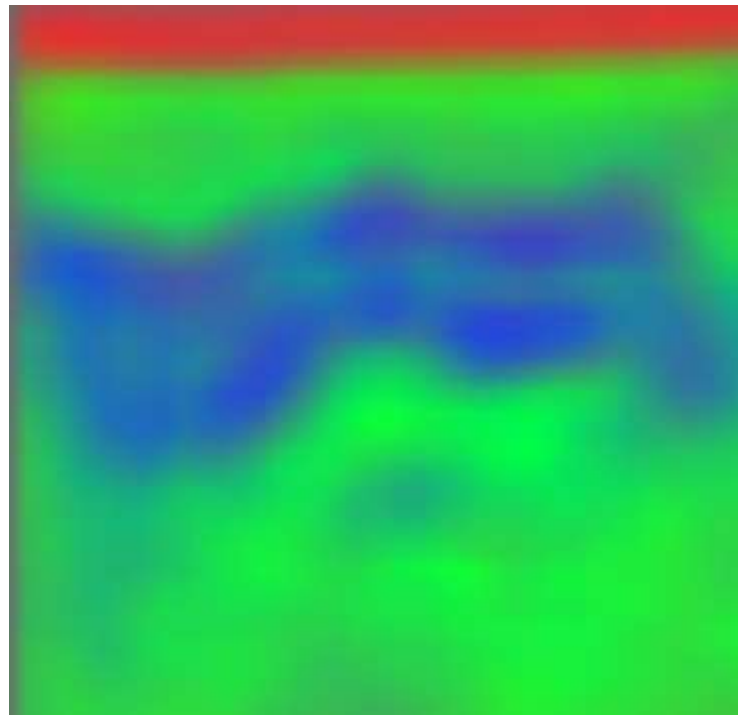
Labels

SEISMIC INTERPRETATION

New Deep Learning approaches - Salt



Detection



Prob. Map

Machine Learning in Rock Facies Classification: An Application of XGBoost

Licheng Zhang, Cheng Zhan

Summary

Big data analysis has drawn much attention across different industries. Geoscientists, meanwhile, have been doing analysis with voluminous data for many years, without even bragging how big it is. In this paper, we present an application of machine learning, to be more specific, the gradient boosting method, in Rock Facies Classification based on certain geological features and constrains. Gradient boosting is a both popular and effective approach in classification, which produces a prediction model in an ensemble of weak models, typically decision trees. The key for gradient boosting to work successfully lies in introducing a customized objective function and tuning the parameters iteratively based on cross-validation. Our model achieves a rather high F1 score in evaluating two test wells data.

Introduction and Background

Machine learning emerges to be a very promising area and should make the work of future geoscientists more fun and less tedious. Furthermore, with the maturing neural network technology, the ability for better geological interpretation

ten wells (with 4149 examples), consisting of a set of seven predictor variables and a rock facies (class) for each example vector and validation (test) data (830 examples from two wells) having the same seven predictor variables in the feature vector. Facies are based on the examination of cores from nine wells taken vertically at half-foot intervals. Predictor variables include five from the wireline log measurements and two geologic constraining variables that are derived from geologic knowledge. These are essentially continuous variables sampled at a half-foot sample rate.

The seven predictor variables are:

Five wireline log measurements	**Two geologic constrains**
• Gamma ray (GR) • Resistivity logging (ILD_log10) • Photoelectric effect (PE) • Neutron-density porosity difference (Delta PHI) • Average neutron-density porosity (PHIND)	• Nonmarine-marine indicator (NM_M) • Relative position (RELPOS)

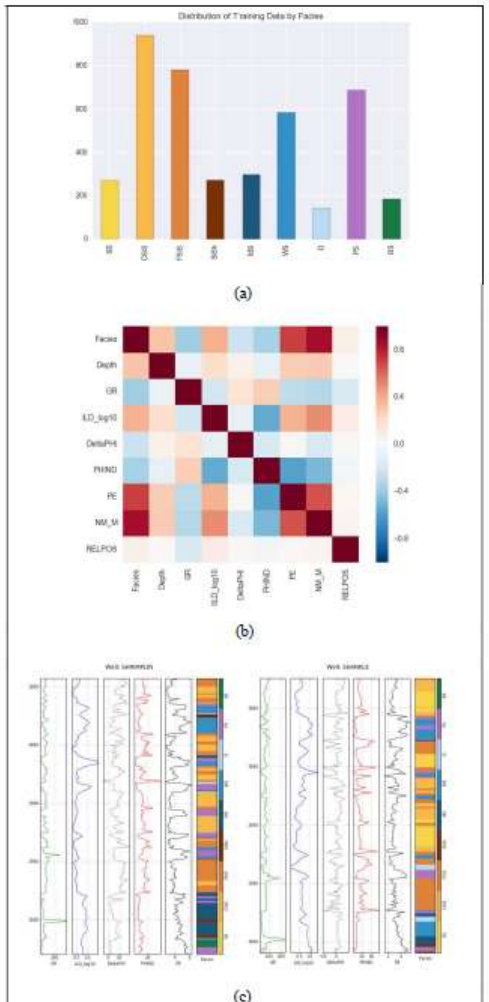


Figure 1: (a) Distribution of facies (b) Heatmap of features (c) Log plots for well SHRIMPLIN and SHANKLE

The next step is data preparation and model selection. The goal is to build a reliable model to predict the Y values (Facies) based on X values (the seven predictor variables).

To enhance the performance of XGBoost's speed over many iterations, we create a DMatrix format. Such process sorts the data initially to optimize for XGBoost in building trees, and reduces the runtime correspondingly. This is especially helpful in learning with a large number of training examples.

```
import xgboost as xgb
X_train = training_data.drop(['Facies', 'Well Name','Formation','Depth'], axis = 1 )
Y_train = training_data['Facies' ] - 1
dtrain = xgb.DMatrix(X_train, Y_train)
```

On the other hand, in order to quantify the quality of the models, certain metrics are needed. We use accuracy metrics for judging the models. A simple and easy way to learn the terminologies (e.g., accuracy, prediction, recall) can be found in the following webpage (http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/).

There are several main parameters to be tuned to get a good model for this rock facies classification problem.

```
from xgboost import XGBClassifier
xgb1 = XGBClassifier(learning_rate = 0.1,n_estimators=1000,
                     max_depth=5,min_child_weight=1,gamma = 0,subsample=0.8,
                     colsample_bytree=0.8,objective='multi:softmax',nthread =4
)
```

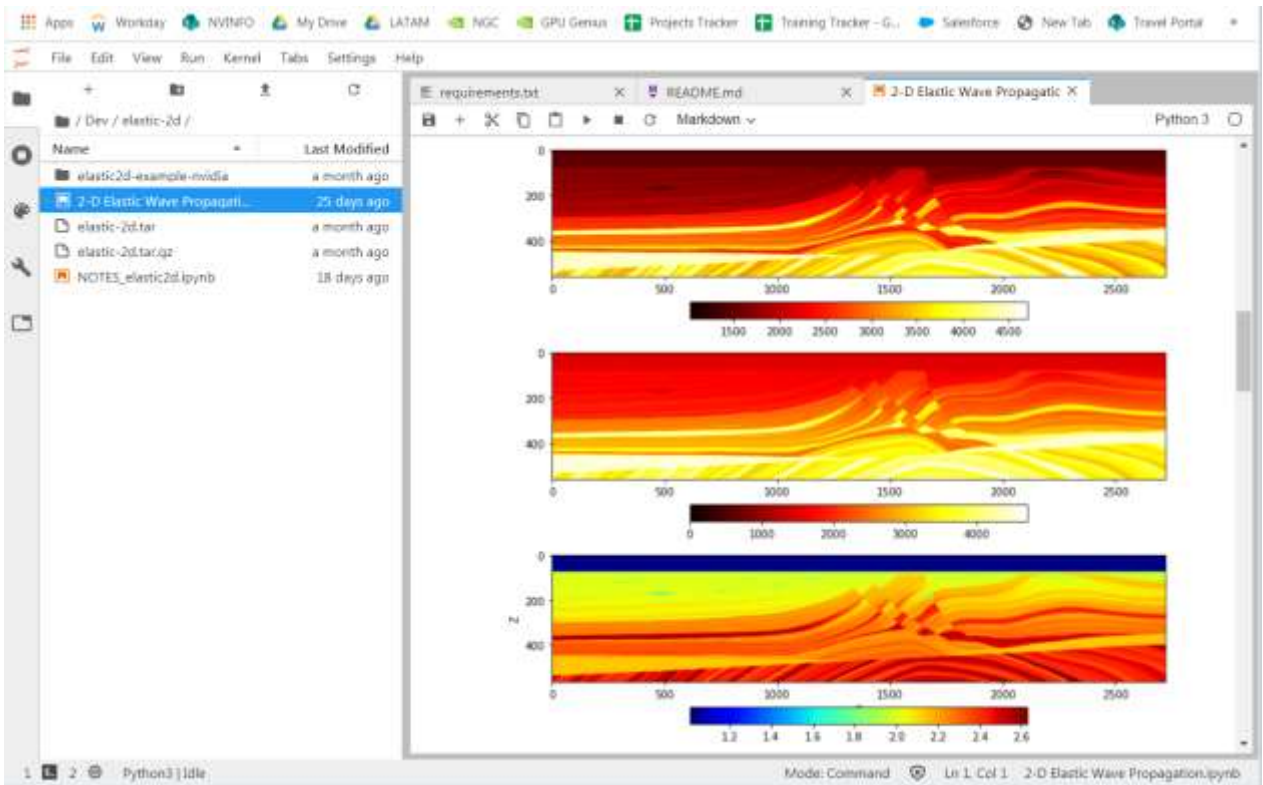
Table 2: main parameters

Learning rate	Step size shrinkage employed to prevent overfitting. We shrinks the feature weights to make the boosting process more conservative
N_estimators	The number of trees
Max_depth	Maximum depth of a tree, and increasing this value will make the model more complex(likely to be overfitting)
Min_child_weight	Minimum sum of instance weight needed in a child
Gamma	Minimum loss reduced required to make a further partition on a leaf node of the tree
Subsample	Subsample ratio of the training instance
Colsample_bytree	Subsample ratio of features when constructing each tree
Objective:'multi:softmax'	This sets XGBoost to produce multiclass classification using the softmax objective
nthread	Number of parallel threads used to run XGBoost

2-D ELASTIC WAVE PROPAGATION

Model Properties Vp, Vs, Rho

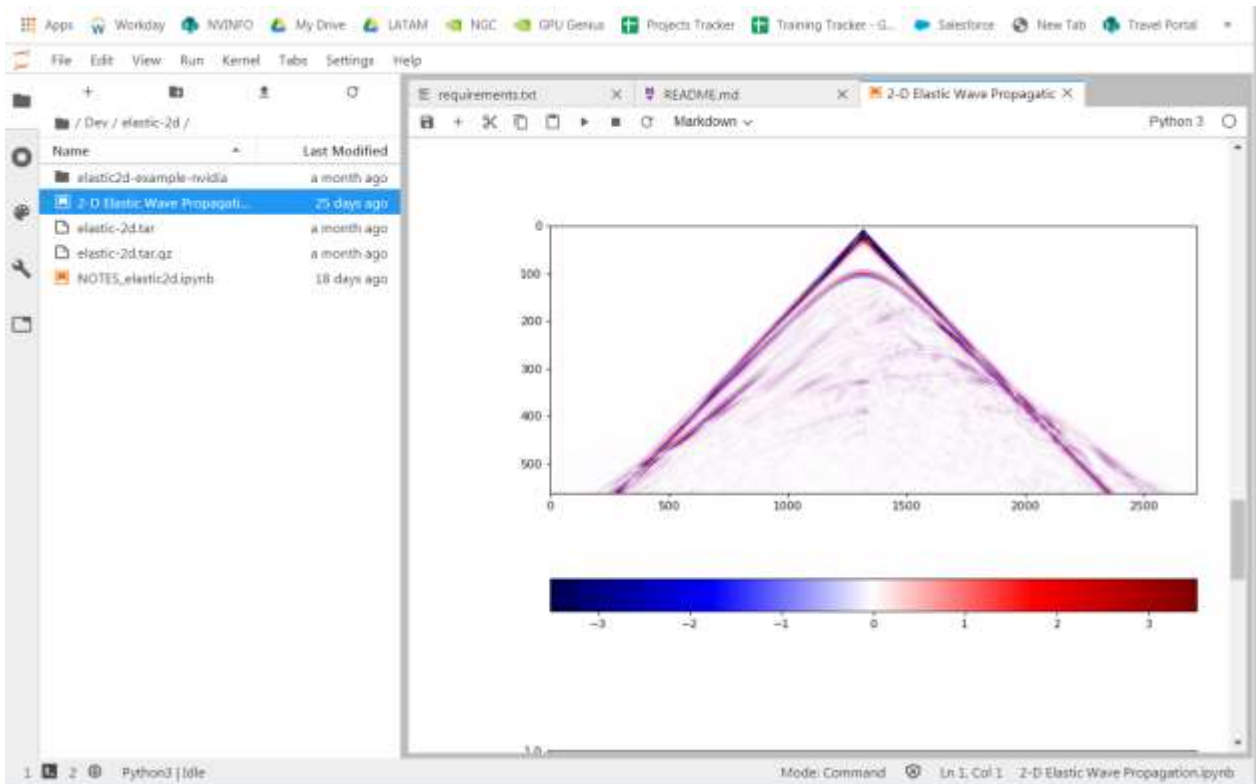
- Jupyter
- Matplotlib
- OpenACC
- GPU



2-D ELASTIC WAVE PROPAGATION

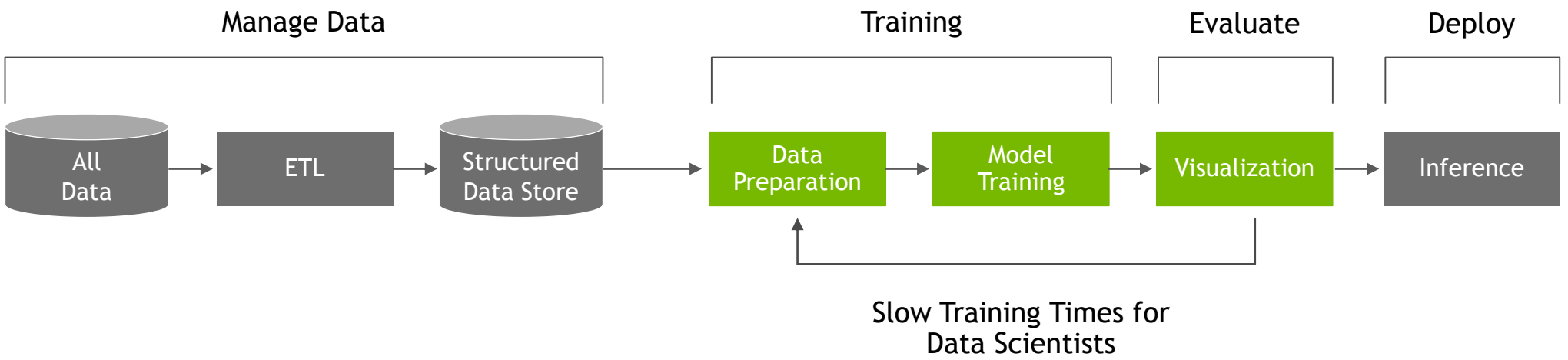
Recorded Wavefield

- Jupyter
- Matplotlib
- OpenACC
- GPU




NVIDIA POWERS
ACCELERATED DATA
SCIENCE

THE BIG PROBLEM IN DATA SCIENCE



ACCELERATING MACHINE LEARNING

The RAPIDS Ecosystem

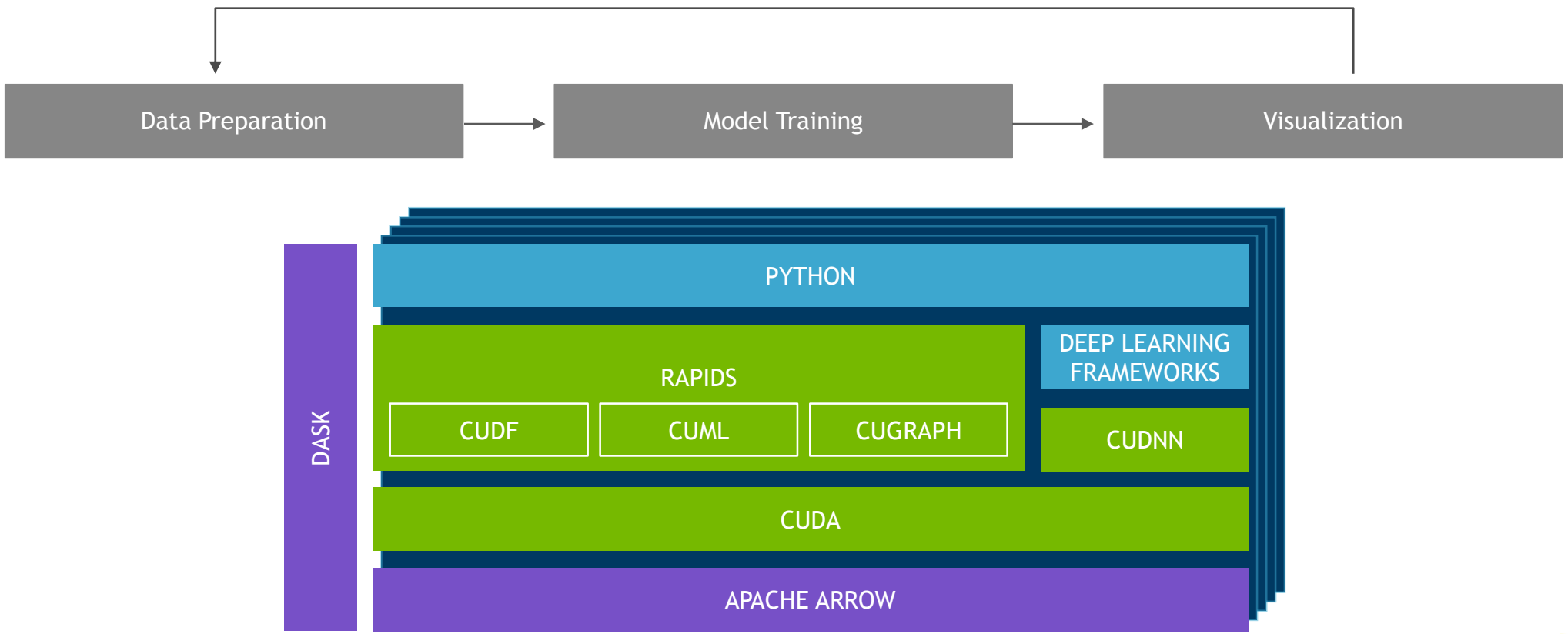
Open Source Community	Enterprise Data Science Platforms	Startups	Deep Learning Integration
ANACONDA. Apache Arrow GUNROCK PyData QUANTSIGHT learn URSA LABS	IBM ORACLE SAS SAP	BLAZINGDB FASTDATA graphistry H_2O.ai kinetica MAPR omnisci	Chainer PYTORCH

RAPIDS

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

RAPIDS — OPEN GPU DATA SCIENCE

Software Stack

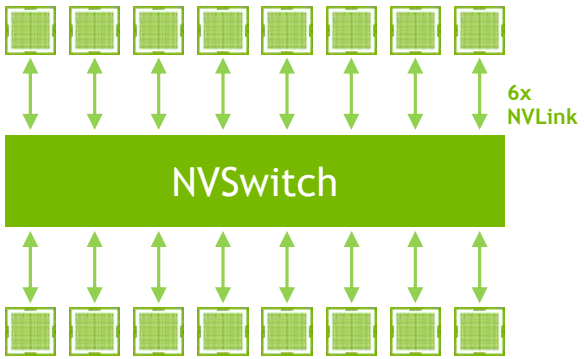


PILLARS OF DATA SCIENCE PERFORMANCE

CUDA Architecture

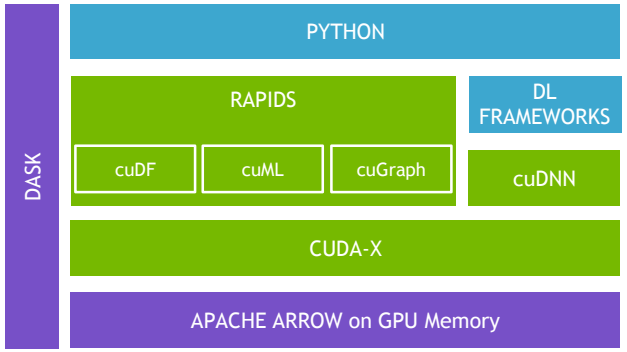
Massively Parallel Processing

NVLink/NVSwitch



High Speed Connecting between GPUs for Distributed Algorithms

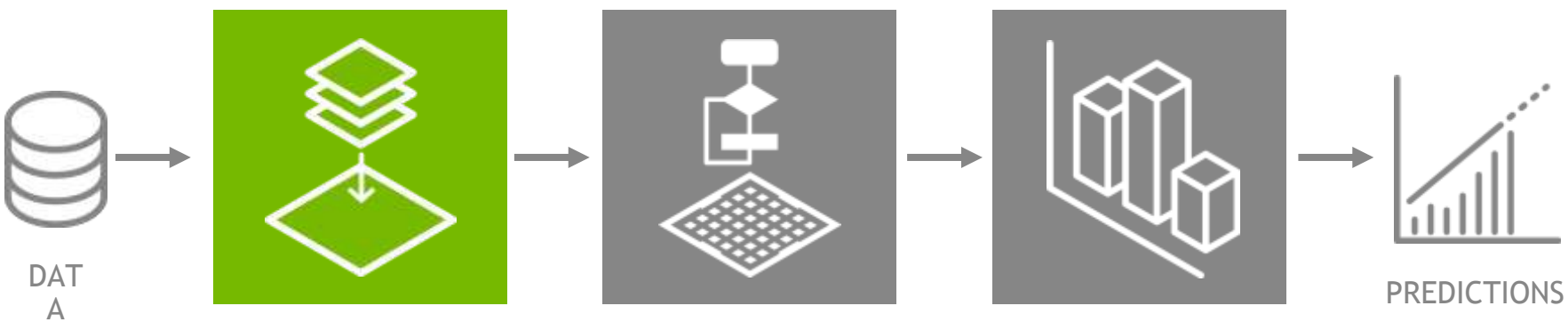
CUDA-X AI



NVIDIA GPU Acceleration Libraries for Data Science and AI

GPU-ACCELERATED DATA SCIENCE WORKFLOW WITH RAPIDS

Built on CUDA-X AI



DATA PREPARATION

GPUs accelerated compute for in-memory data preparation

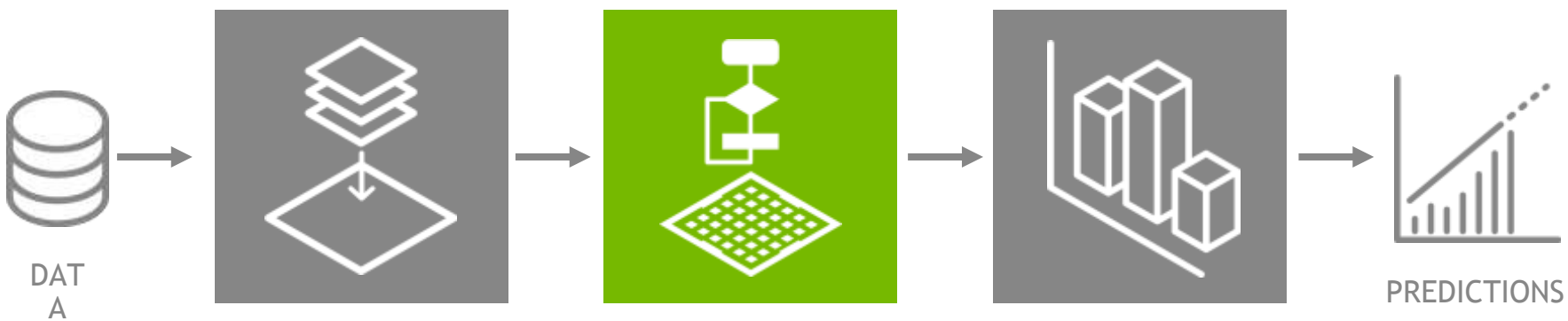
Simplified implementation using familiar data science tools

Python drop-in **pandas** replacement built on CUDA C++.

GPU-accelerated Spark (in development)

GPU-ACCELERATED DATA SCIENCE WORKFLOW WITH RAPIDS

Built on CUDA-X AI



MODEL TRAINING

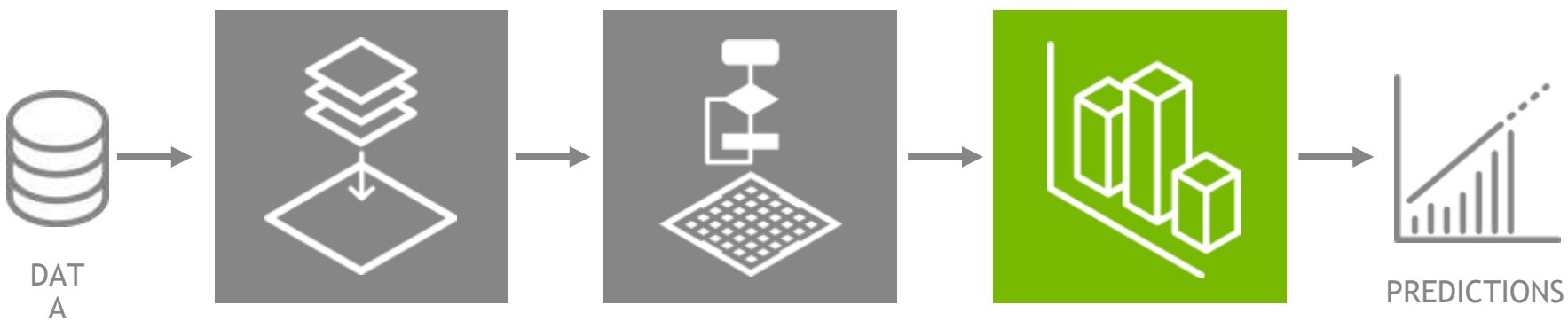
GPU-acceleration of today's most popular ML algorithms such as **XGBoost**

Also available are PCA, K-means, k-NN, DBScan, tSVD, and many more

Easy-to-adopt, scikit-learn like interface

GPU-ACCELERATED DATA SCIENCE WORKFLOW WITH RAPIDS

Built on CUDA-X AI



VISUALIZATION

Effortless exploration of datasets, billions of records in milliseconds

Dynamic interaction with data = faster ML model development

Data visualization ecosystem (Graphistry & OmniSci), integrated with RAPIDS

XGBOOST: THE WORLD'S MOST POPULAR MACHINE LEARNING ALGORITHM

Versatile and High Performance

The leading algorithm for tabular data

Outperforms most ML algorithms on regression, classification and ranking

Winner of many data science Kaggle competitions

InfoWorld Technology of the Year Award, 2019

Well known in data science community and widely used for **forecasting**, **fraud detection**, **recommender engines**, and much more

HOW CAN XGBOOST BE IMPROVED?

XGBoost Performance is Constrained by CPU Limitations

CPU processing is slow, creating issues for large data sets or when timeliness is crucial (e.g. intraday requirements for financial services)

Hyperparameter search is very slow, making search not feasible

Prediction speed limits the depth and number of trees in time sensitive applications

dmlc XGBoost

GPU-ACCELERATED XGBOOST

Unleashing the Power of NVIDIA GPUs for Users of XGBoost

Faster Time To Insight
XGBoost training on GPUs is significantly faster than CPUs, completely transforming the timescales of machine learning workflows.

Better Predictions, Sooner
Work with larger datasets and perform more model iterations without spending valuable time waiting.

Lower Costs
Reduce infrastructure investment and save money with improved business forecasting.

Easy to Use
Works seamlessly with the RAPIDS open source data processing and machine learning libraries and ecosystem for end-to-end GPU-accelerated workflows.

USE WITH MINIMAL CODE CHANGES

GPU-Acceleration with the same XGBoost Usage

BEFORE

```
import xgboost as xgb

params = {'max_depth': 3,
          'learning_rate': 0.1}


dtrain = xgb.DMatrix(X, y)
bst = xgb.train(params, dtrain)
```

AFTER

```
import xgboost as xgb

params = {'tree_method': 'gpu_hist',
          'max_depth': 3,
          'learning_rate': 0.1}

dtrain = xgb.DMatrix(X, y)
bst = xgb.train(params, dtrain)
```

XGBOOST: GPU VS. CPU

Tremendous Performance Improvements and Better Accuracy

Take advantage of parallel processing with multiple GPUs

Scale to multiple nodes

GPU implementation is more memory efficient (half of CPU)

Improved accuracy by allowing time for more iterations, ability to leverage hyperparameter search, and reduced scale out needs

A single DGX-2 with GPU-accelerated XGBoost is 10x Faster than 100 CPU nodes

TRADITIONAL DATA SCIENCE CLUSTER

Workload Profile:

Fannie Mae Mortgage Data:

- 192GB data set
- 16 years, 68 quarters
- 34.7 Million single family mortgage loans
- 1.85 Billion performance records
- XGBoost training set: 50 features

300 Servers | $3M | 180 kW

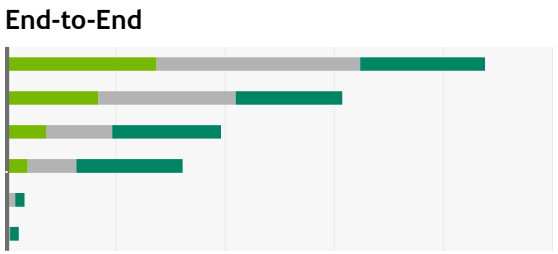
GPU-ACCELERATED DATA SCIENCE CLUSTER

GPU-accelerated XGBoost with DGX-2

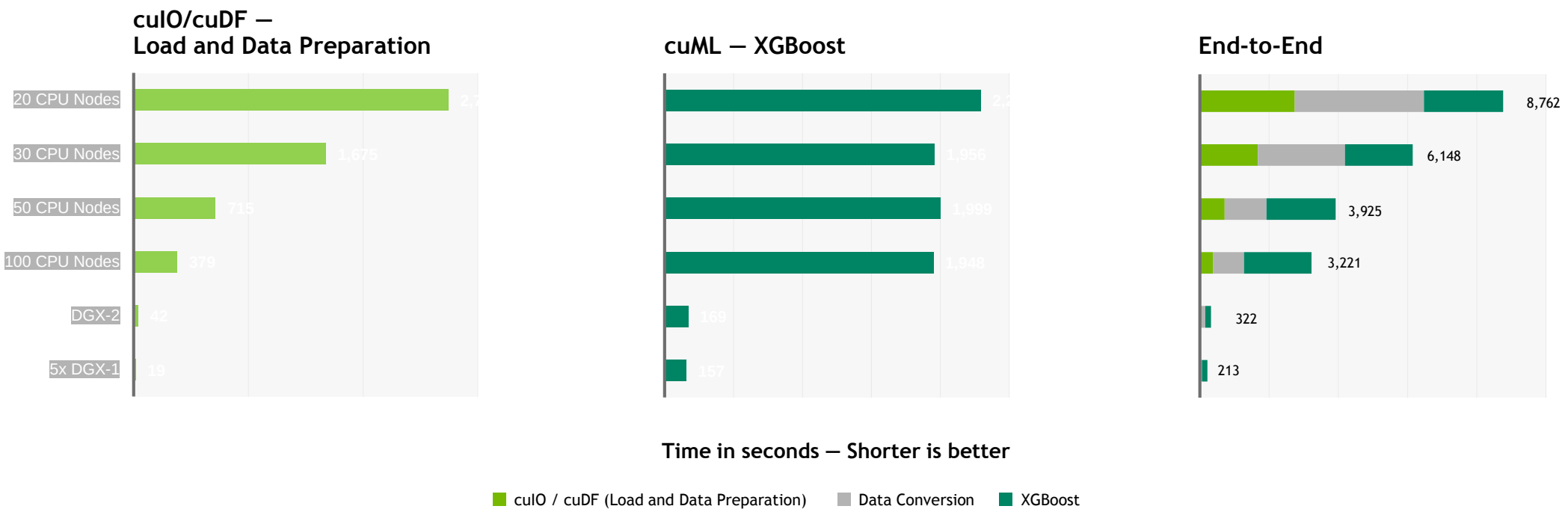
1 DGX-2 | 10 kW

1/8 the Cost | 1/15 the Space

1/18 the Power



NVIDIA GPUS ARE PROVEN FASTER FOR DATA SCIENCE



Benchmark

200GB CSV dataset; Data preparation includes joins, variable transformations.

CPU Cluster Configuration

CPU nodes (61 GiB of memory, 8 vCPUs, 64-bit platform), Apache Spark

DGX Cluster Configuration

5x DGX-1 on InfiniBand network

DISTRIBUTED XGBOOST

GPU-Accelerated XGBoost for Large Scale Workloads

GPU-acceleration for **XGBoost** with **Apache Spark** and **Dask**

Multiple nodes and multiple GPUs per node

Explore and prototype models on a PC, workstation, server, or cloud instance and scale to two or more nodes for production training

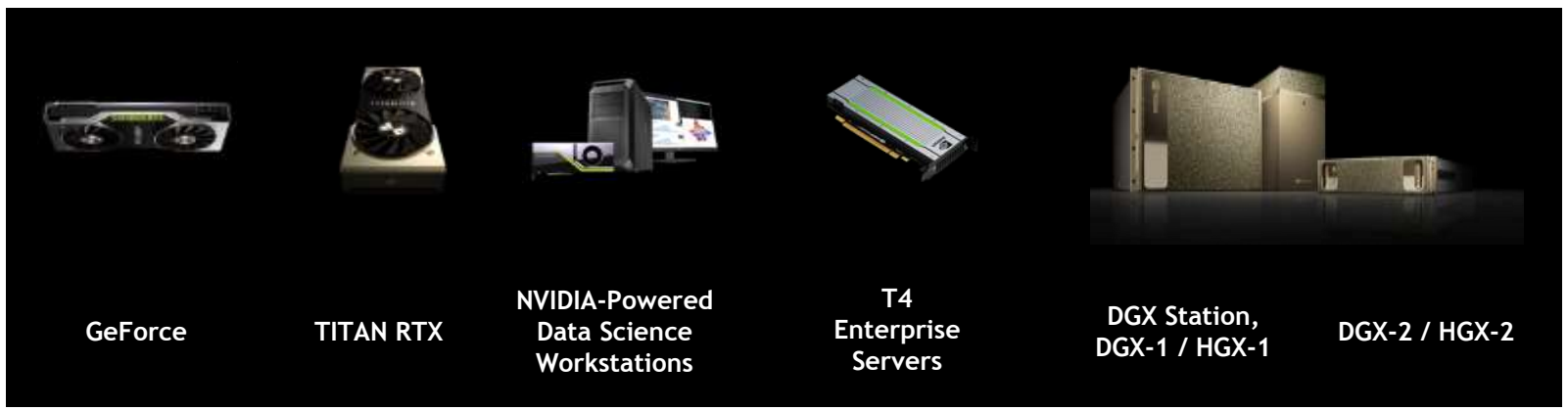
An ideal solution for GPU-accelerated clusters and enterprise scale workloads

Try out Dask support immediately using Google Cloud Dataproc

Download for on-prem and cloud deployments

GPU-ACCELERATED DATA SCIENCE

A Solution for Every User and Every Organization



ML EXPERIMENTATION

PRODUCTION DATA CENTER

GPU-ACCELERATED DATA SCIENCE PLATFORMS

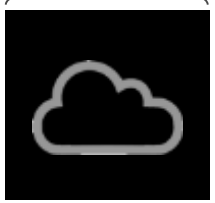
Unparalleled Performance and Productivity

ML in the Cloud
All the top CSPs

ML Enthusiast
High-end PCs

Enterprise Desktop
Individual Workstations

Enterprise Data Center
Shared Infrastructure for Data Science Teams



Max Flexibility | **Max Performance**

	NVIDIA GPUs in the Cloud	GeForce	TITAN RTX	NVIDIA-Powered Data Science Workstations	T4 Enterprise Servers	DGX Station, DGX-1 / HGX-1	DGX-2 / HGX-2
Benefit	Ease of getting started, low/no barrier to entry, elasticity of resources	Enthusiast PC solution, easy to acquire, low cost, great performance	The ultimate PC GPU for data scientists. Easy to acquire, deploy and get started experimenting.	Enterprise workstation for experienced data scientists	Standard GPU-accelerated data center infrastructures with the world's leading servers	Enterprise server, proven 4 or 8-way configuration, modular approach for scale-up, fastest multi-GPU & multi-node training	Largest compute and memory capacity in a single node, fastest training solution
Typical GPU Memory (system dependent)	varies depending on offering	22GB	48GB	96GB	64 GB (4 x 16 GB)	128GB-256GB	512GB
GPU Fabric	varies depending on offering	2-way NVLink	2-way NVLink	2-way NVLink	PCIe 3.0	4- and 8-way NVLink	16-way NVSwitch

NVIDIA ACCELERATED DATA SCIENCE KEY USE CASES

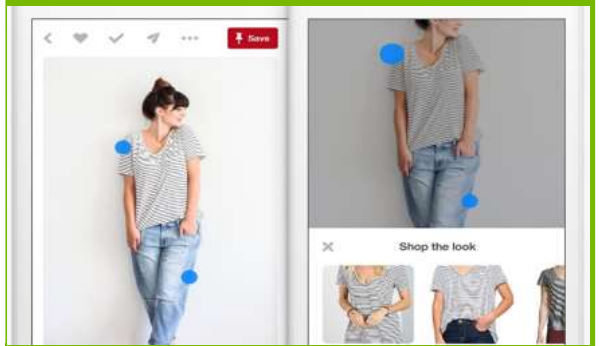
FORECASTING

Data science significantly increases the efficiency and accuracy of forecasting, directly contributing to bottom line growth.

FRAUD DETECTION

Data science dramatically improves fraud detection, saving money for organizations and customers alike.

RECOMMENDER SYSTEMS



Data science has revolutionized the connected experience, delighting users with personalized experiences.

IMPROVING DEMAND FORECASTS

With >100,000 different products in its 4,700 U.S. stores, the Walmart Labs data science team predicts demand for 500 million item-by-store combinations every week.

By performing forecasting with the open-source RAPIDS data processing and machine learning libraries built on CUDA-X AI on NVIDIA GPUs, Walmart speeds up feature engineering 100x and trains machine learning algorithms 20x faster, resulting in faster delivery of products, real-time reaction to shopper trends, and inventory cost savings at scale.

NVIDIA Walmart

AI & DATA SCIENCE FOR NETWORK OPERATIONS

A wireless network operator had access to terabytes of data daily but no efficient way to gain insights from it. That is until a deep learning solution powered by NVIDIA DGX POD, RAPIDS, and software from Datalogue and OmniSci, changed the way they collect, process, visualize and understand data.

Data prep improved from 8 days to 4 minutes and the company's new AI models predict high-surge Wi-Fi usage and detect anomalies with 99% accuracy.

NVIDIA

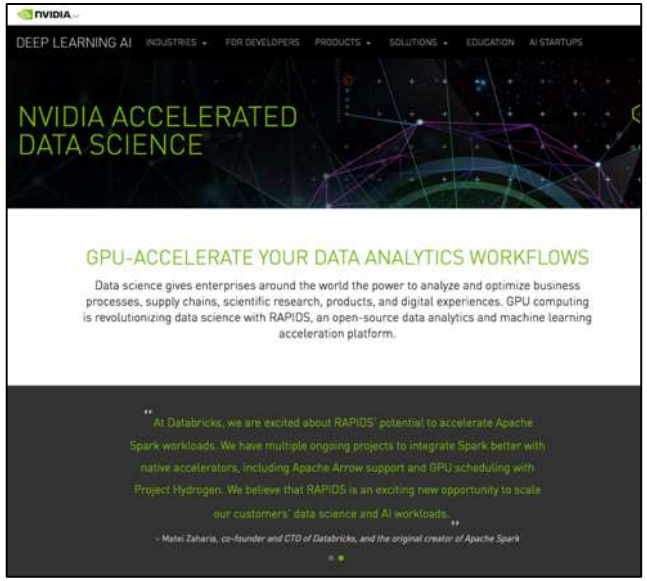
SUPERCHARGING GENOMIC ANALYTICS

China's healthcare industry is turning to AI to address the needs of its elderly population. Genetics giant BGI—which has over 1PB of data—is classifying targetable peptides for personalized immunotherapy for cancer patients.

By running the open source RAPIDS data processing and machine learning libraries built on CUDA-X AI on an NVIDIA DGX-1 AI supercomputer, BGI sped up analysis 18x using cuDF, and 10x using XGBoost. The company is now expanding analysis to millions of peptide candidates.

NVIDIA 华大基因 BGI

FOR MORE INFORMATION



nvidia.com/datascience

rapids.ai

nvidia.com/en-us/technologies/cuda-x/


LEARN & SHARE
MORE

NVIDIA GTC GPU TECHNOLOGY CONFERENCE

Don’t miss the premier AI conference.
www.nvidia.com/gtc

March 22—26, 2020 | Silicon Valley

CONNECT

Connect with hundreds of experts from top industry, academic, startup, and government organizations

LEARN

Gain insight and valuable hands-on training through over 500+ sessions

DISCOVER

See how GPU technology is creating breakthroughs in deep learning, cybersecurity, data science, healthcare and more

INNOVATE

Explore disruptive innovations that can transform your work

JOIN US AT GTC 2020 | USE VIP CODE XXXXX FOR 25% OFF

NVIDIA

GTC GPU TECHNOLOGY CONFERENCE

THE LATEST DEEP LEARNING DEVELOPER TOOLS

March 22 | Full-Day Workshops
March 23 - 26 | Conference & Training

Get the hands-on experience you need to transform the future of AI, high-performance computing and more with NVIDIA's Deep Learning Institute (DLI).

Register for GTC 2020 to earn certification in full-day workshops, join instructor-led sessions, and start self-paced training.

www.nvidia.com/en-us/gtc/sessions/training/

JOINT MACHINE LEARNING WORKSHOP

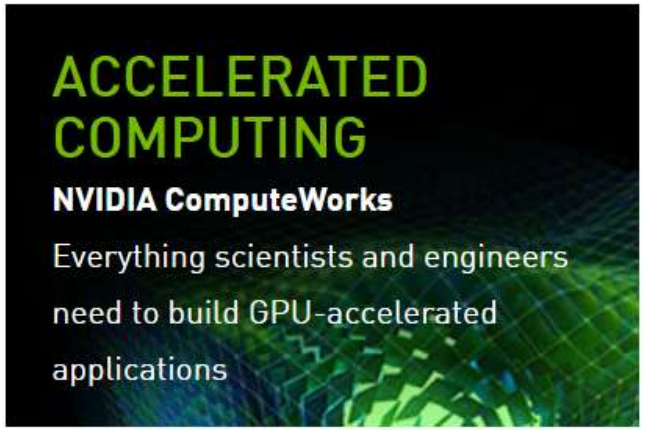
SBGf & SEG, 12-13 May 2020, Rio de Janeiro, Brazil

Join the NVIDIA Developer Program

Access everything you need to develop with NVIDIA products.

Register Now

developer.nvidia.com



NVIDIA DEEP LEARNING INSTITUTE

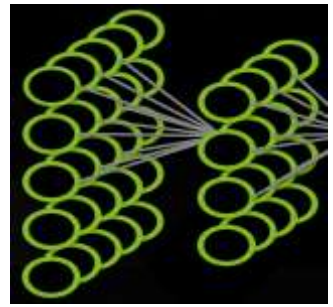
Hands-on self-paced and instructor-led training in deep learning and accelerated computing for developers

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

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

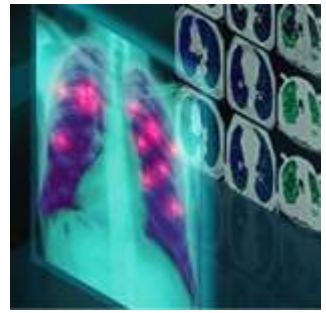
Download the course catalog, view upcoming workshops, and learn about the University Ambassador Program: **www.nvidia.com/dli**

Caffe2 · Microsoft Cognitive Toolkit · mxnet · TensorFlow · PYTORCH

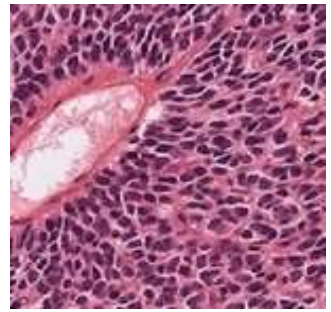


Deep Learning Fundamentals

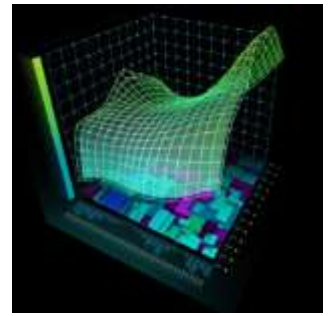
Autonomous Vehicles



Medical Image Analysis



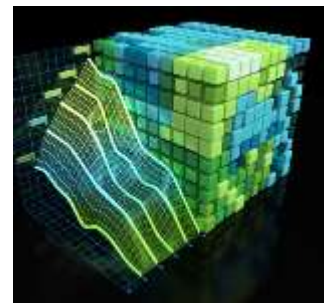
Genomics



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.