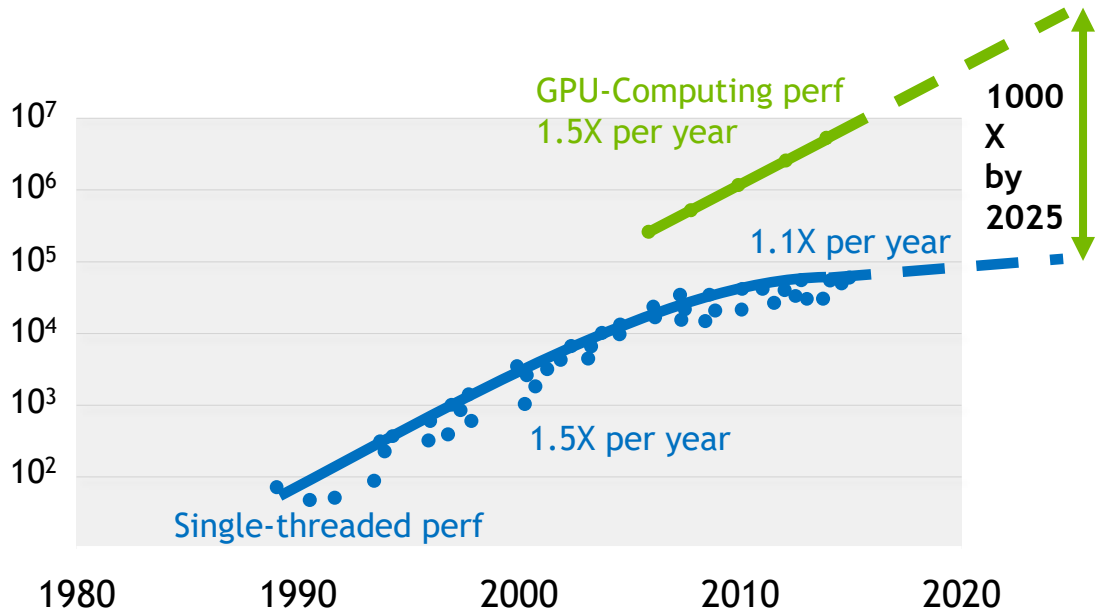
NVIDIA GPU: WHAT AND WHY

Reynaldo Gomez

reynaldog@nvidia.com

ACCELERATING HIGH-GROWTH CLOUD WORKLOADS



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

CPU No Longer Keeping Pace With Demand

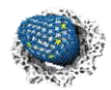
Artificial Intelligence

Sentiment

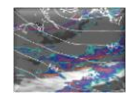
Recommender

Video | Images

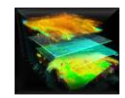
High Performance Computing



Molecular Simulations



Weather Forecasting

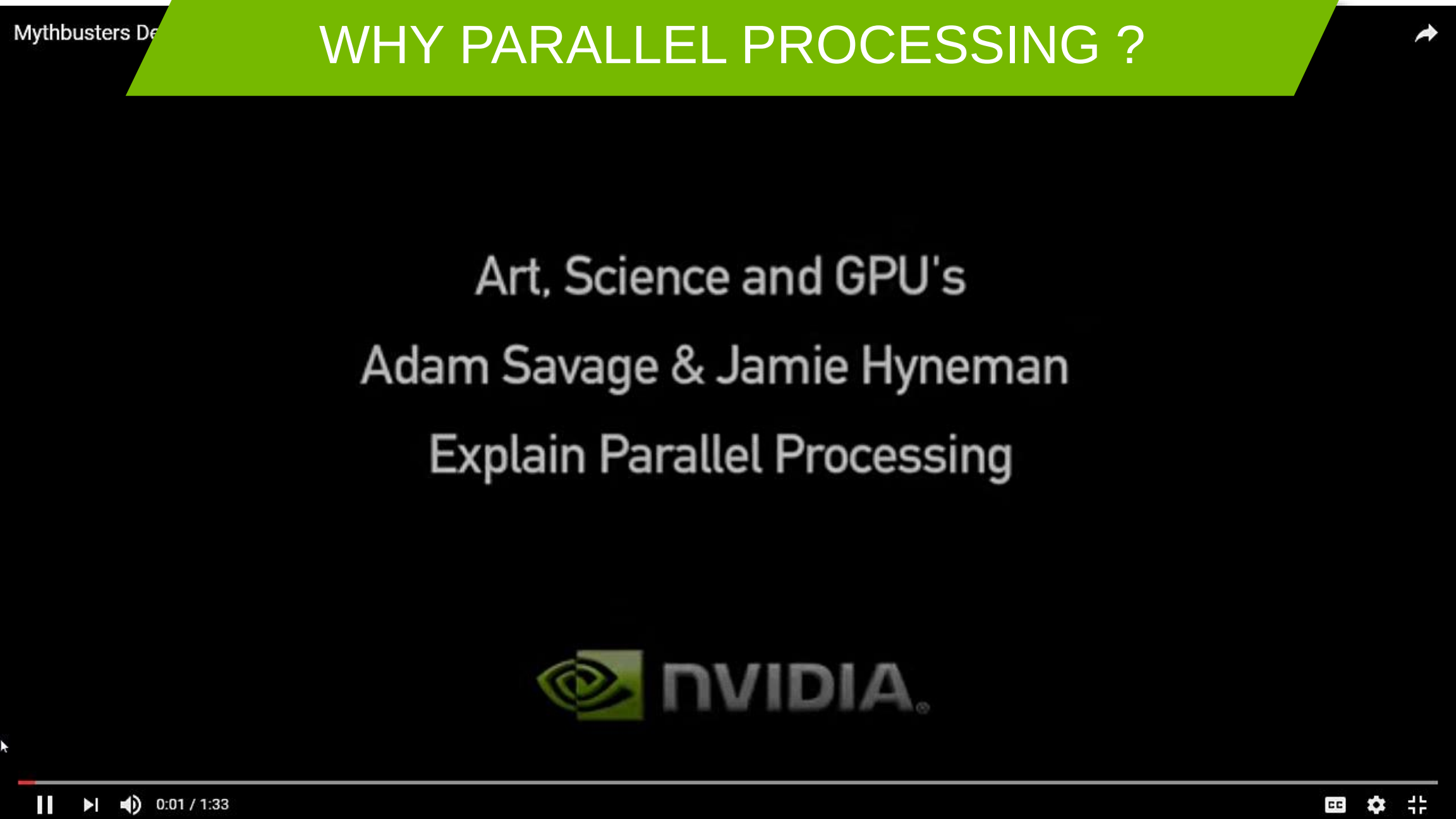


Seismic Mapping

Rendering & Graphics

Growth Workloads Require Acceleration

WHY PARALLEL PROCESSING ?



NVIDIA DATA CENTER PLATFORM

Single Platform Drives Utilization and Productivity

CUSTOMER USE CASES

Speech | Translate | Recommender | Healthcare | Manufacturing | Finance

CONSUMER INTERNET & INDUSTRY APPLICATIONS

Molecular Simulations | Weather Forecasting | Seismic Mapping

SCIENTIFIC APPLICATIONS

Creative & Technical | Knowledge Workers

VIRTUAL GRAPHICS

APPS & FRAMEWORKS

python | TensorFlow | mxnet | RAPIDS | PYTORCH | Chainer | ONNX

Amber | NAMD | +600 Applications

CATIA | Ps | AUTODESK 3DS MAX | Windows 10

CUDA-X & NVIDIA SDKs

MACHINE LEARNING: cuDF | cuML | cuGRAPH

DEEP LEARNING: cuDNN | CUTLASS | TensorRT

HPC: OpenACC | cuFFT

VIRTUAL GPU: vDWS | vPC | vAPPS

CUDA & CORE LIBRARIES - cuBLAS | NCCL

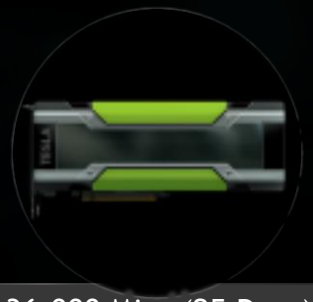
TESLA GPUs & SYSTEMS

TESLA GPU | NVIDIA DGX FAMILY | NVIDIA HGX

SYSTEM OEM: DELL | Hewlett Packard Enterprise | IBM | inspur

CLOUD: aws | Google Cloud Platform | Microsoft Azure

AI TRAINING REQUIRES FULL STACK INNOVATION



36,000 Mins (25 Days)
1xK80 | 2015
CUDA

1,200 Mins (20 Hours)
DGX-1P | 2016
NVLink

480 Mins (8 Hours)
DGX-1V | 2017
Tensor Core

70 Minutes on MLPerf
DGX-2H | 2018
NVSwitch

6.3 Minutes on MLPerf
At Scale | 2018
DGX Cluster

2015 2016 2017 2018

Workload: Time (in minutes) to train ResNet-50 network.

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

THE DGX FAMILY OF AI SUPERCOMPUTERS

CLOUD-SCALE AI

NVIDIA GPU Cloud

Simple Access to GPU Accelerated Software

AI WORKSTATION

DGX Station

with

AI Workstation for Data Science Teams

AI DATA CENTER

DGX-1

with

The Essential Instrument for AI Research

DGX-2

with

The World's Most Powerful AI System for the Most Complex AI Challenges

CONTAINERS: SIMPLIFYING WORKFLOWS

WHY CONTAINERS

Simplifies Deployments

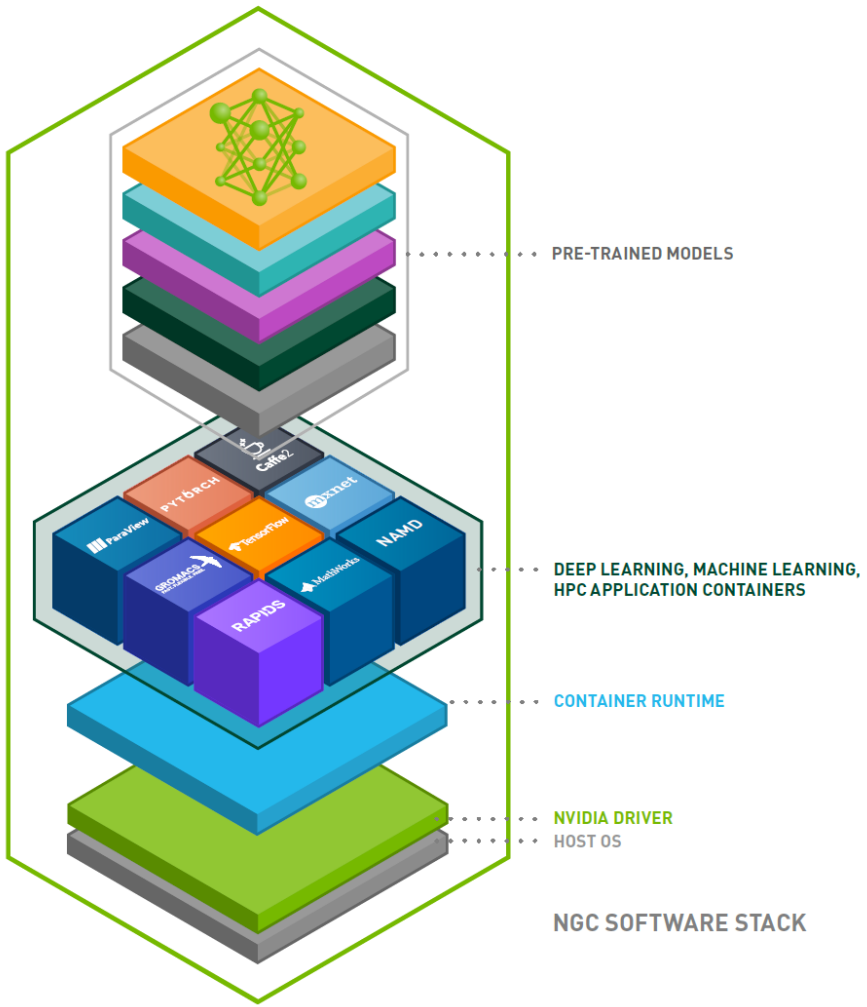
- Eliminates complex, time-consuming builds and installs

Get started in minutes

- Simply Pull & Run the app

Portable

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



NGC CONTAINERS: ACCELERATING WORKFLOWS

WHY CONTAINERS

Simplifies Deployments

- Eliminates complex, time-consuming builds and installs

Get started in minutes

- Simply Pull & Run the app

Portable

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

WHY NGC CONTAINERS

Optimized for Performance

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

Scalable Performance

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

Designed for Enterprise & HPC environments

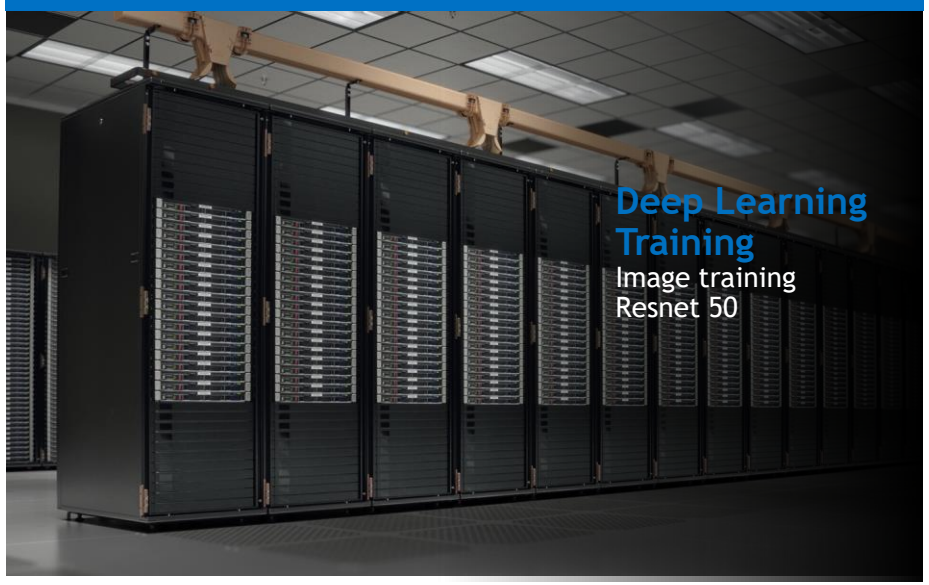
- Supports Docker & Singularity runtimes

Run Anywhere

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

DRAMATICALLY MORE FOR YOUR MONEY

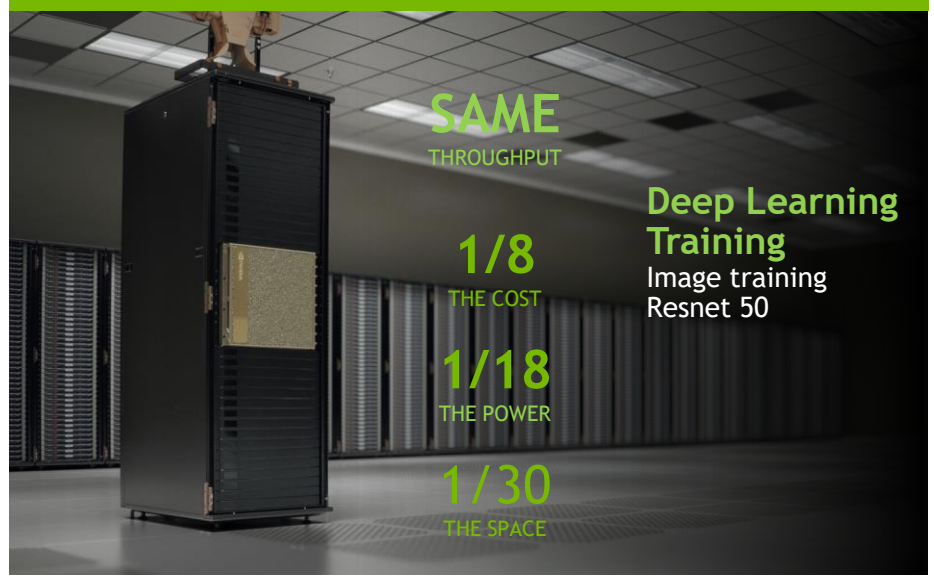
CPU-Only Cluster



300 Self-hosted Broadwell CPU Servers
180 KWatts

=

GPU-Accelerated

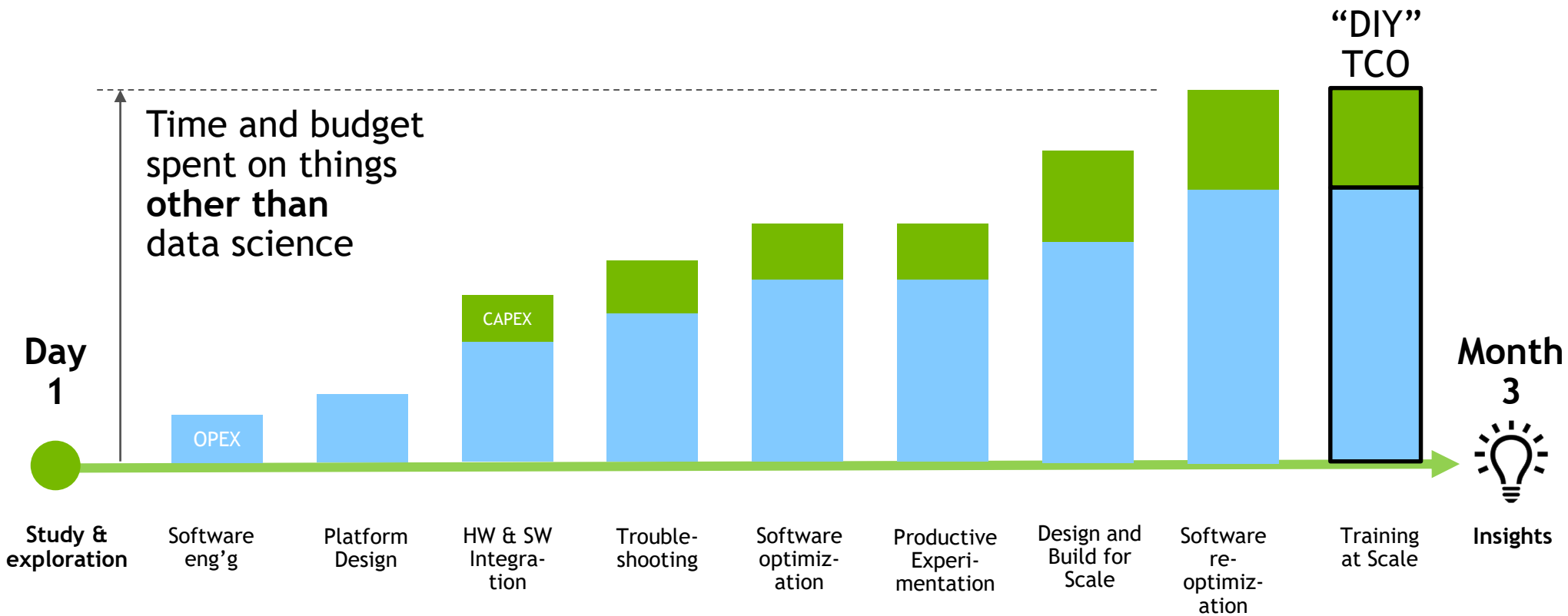


1 DGX-2
10 KWatts

HOW TESLA SAVES MONEY

	GPU Node	CPU Node	# CPU Nodes To Match 1 GPU Node	$ Spend on CPU Nodes	$ Saved with GPU Node
AMBER	$~45K	$~9K	74 CPU Nodes	$666K	$621K
GTC			26 CPU Nodes	$234K	$189K

AI SUCCESS DELAYED BY DEPLOYMENT COMPLEXITY

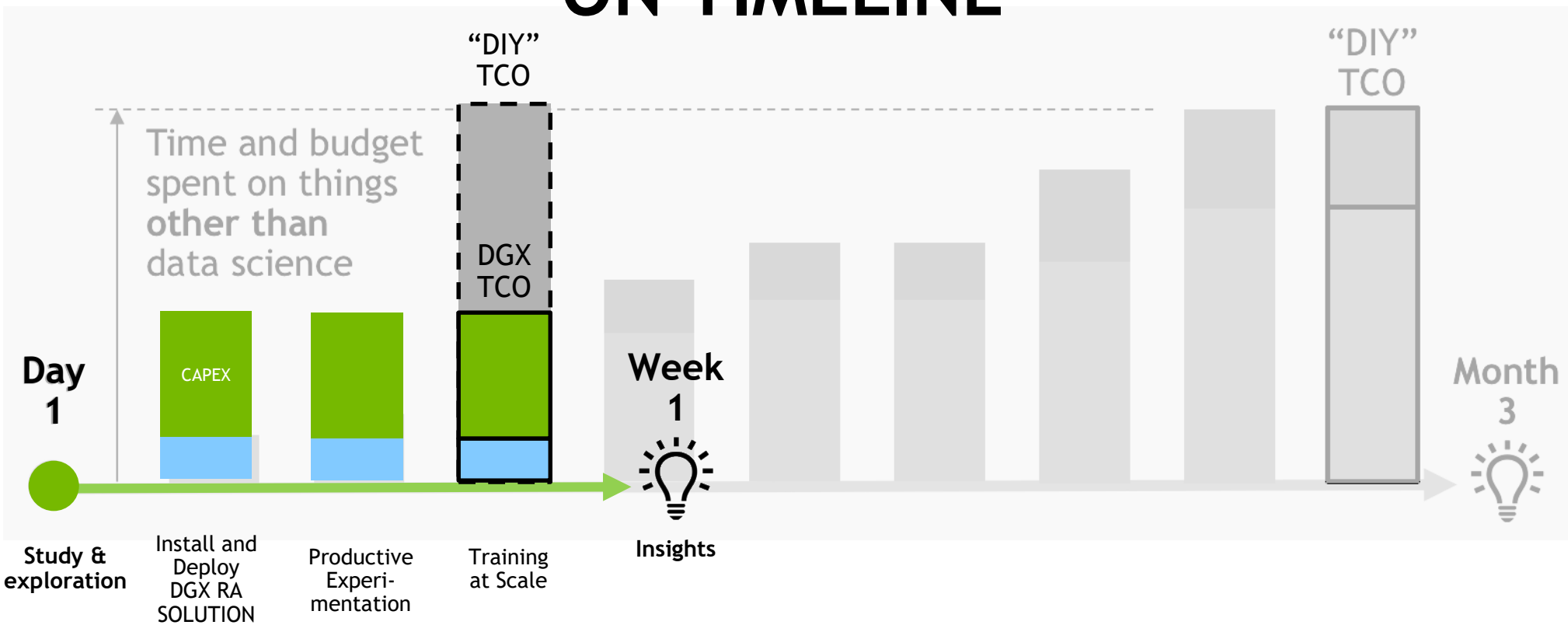


Designing, Building and Supporting an AI Infrastructure - <u>from Scratch</u>

THE IMPACT OF DGX R/A SOLUTIONS ON TIMELINE

2. Deploying an Integrated, Full-Stack AI Solution using a DGX Reference Architecture

THE IMPACT OF DGX R/A SOLUTIONS ON TIMELINE



2. Deploying an Integrated, Full-Stack AI Solution using a DGX Reference Architecture

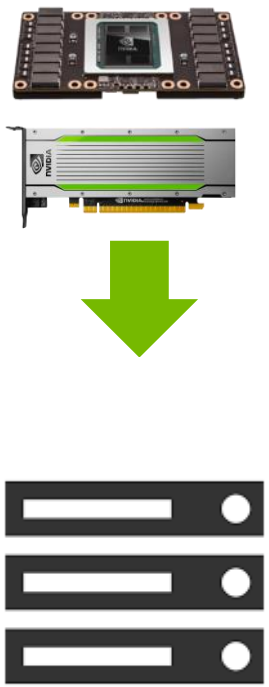
NVIDIA GPUS ARE ON EVERY CLOUD

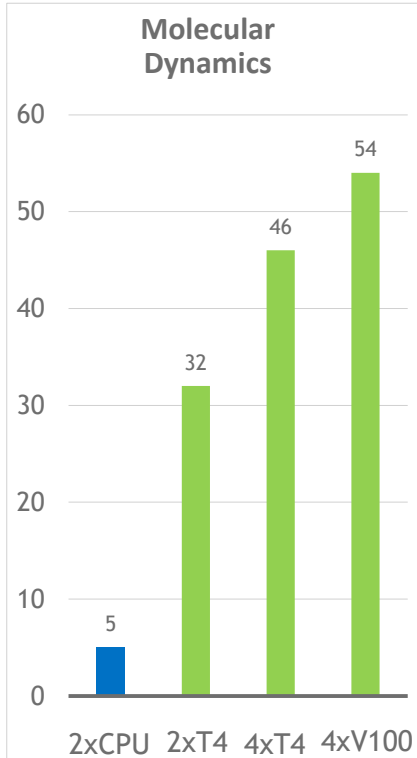
Over 30 Offerings Across USA and China

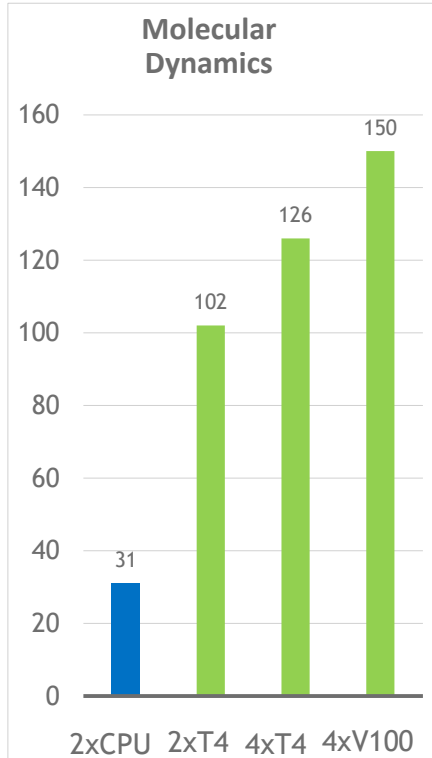
	K520	K80	P40	M60	P4	P100	T4	V100	NGC
Alibaba Cloud					●	●		●	●
AWS	●	●		●				●	●
Baidu Cloud			●		●		●		
Google Cloud		●			●	●	●	●	●
IBM Cloud		●		●		●		●	
Microsoft Azure		●	●	●		●		●	●
Oracle Cloud						●		●	●
Tencent Cloud			●		●				

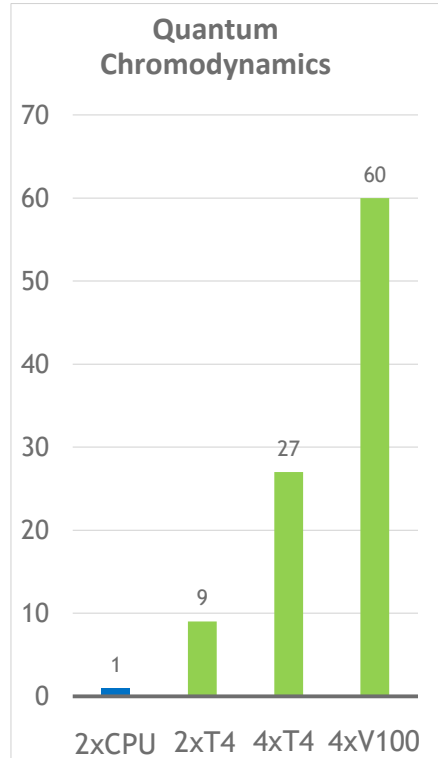
SUPERCHARGED INSTANCES FOR HPC

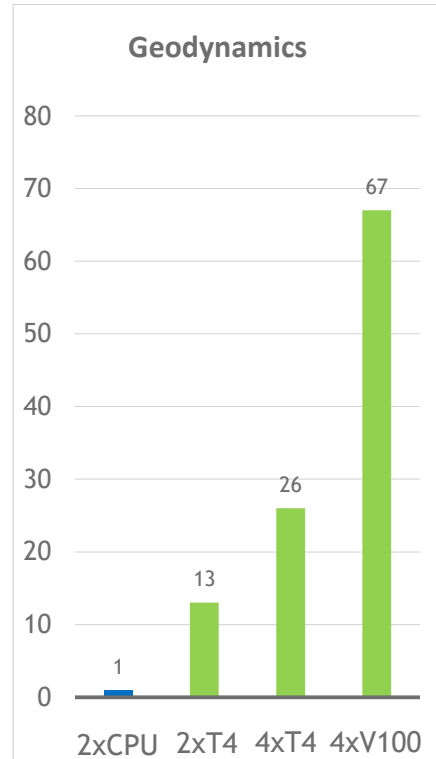
Universal Platform Drives Utilization and Productivity







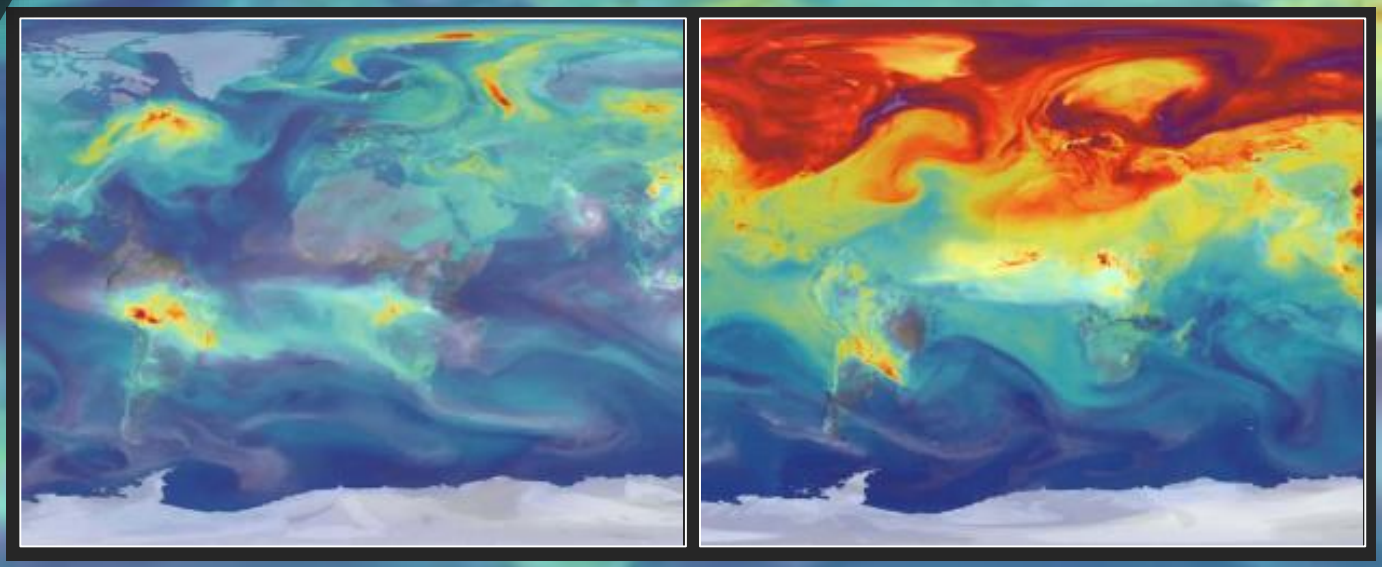




CPU: Dual Xeon Gold 6140@2.30GHz, GPU: Dual Xeon Gold 6140@2.30GHz with GPU servers as shown | CUDA 10.0.130

AN AI MONITOR OF EARTH'S VITALS

NASA Ames uses satellite imagery to measure the effects of carbon and greenhouse gas emissions on the planet. They developed DeepSat—a deep learning framework for satellite image classification trained on a GPU-powered supercomputer. The enhanced satellite imagery will help scientists plan how to protect ecosystems and improve crop production.



NASA: Late summer 2016, forest fires in Africa produce plumes of CO_2
Left: CO_2 - 10/14/2016 / Right: CO_2 - 12/24/2016

Source: https://climate.nasa.gov/climate_resources/142/

EXAMPLE: OILFIELD PREDICTIVE MAINTENANCE

Baker Hughes GE

60K oil wells WW equipped with Electric Submersible Pumps (ESP)

Average Non Producing Time (NPT) due to ESP Failure costs > $150K per day per well

- ML techniques used historically (Rule based, Fuzzy logic, traditional ML), but they don't scale
- >50% False Alarms + detect failures too late

Deep Learning: 93% detection accuracy with 2 months lead time at 5% False Alarm

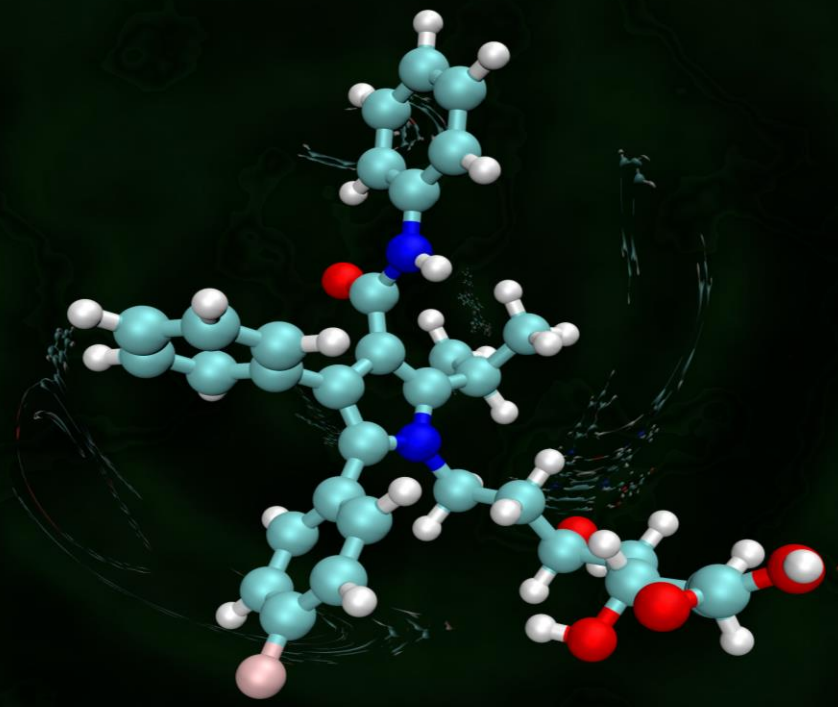
At 10% DL based anomaly detection yields $300K per well of lost productivity annually

Image source: Arcadis.com

CHASING 10^{60} CHEMICAL COMPOUNDS

Identifying molecules with desirable chemical properties is central to many industries. In the chemical space of 10^{60} conceivable compounds, only 10^{8} have been synthesized. Screening even a small fraction of the remaining compounds with legacy methods would take 100 node-seconds per compound.

Researchers at Dow are using GPU-powered deep learning to deliver completely novel molecular structures with specific properties. The AI produced 3M promising chemical leads in 1 day on an NVIDIA DGX-1.



HUNTING GHOST PARTICLES WITH DEEP LEARNING

Understanding the properties of Neutrinos is the focus of a world-wide campaign. Observing these ‘ghost particles’ requires instruments of incredible size and scale. Fermilab’s NOvA experiment applies two detectors snapping two million photos/second and analyzing them for neutrino activity.

Fermilab developed deep neural networks trained on NVIDIA GPUs and improved detection rate by 33% — increasing the discovery potential of NOvA and other experiments probing fundamental questions of the universe.

NVIDIA.

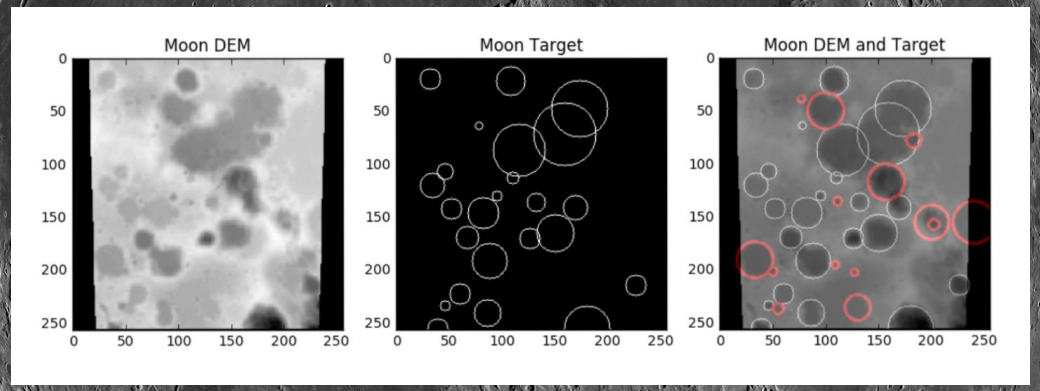
SMARTER INSPECTION SERVICES

In business, ensuring equipment uptime and meeting safety and regulatory compliance is non-negotiable. Using deep neural networks developed on NVIDIA DGX-1 in the data center that can easily extend to NVIDIA DGX Station in the field, Avitas Systems delivers inspection services using robotic-based autonomous inspection and advanced data analytics. In addition to safeguarding workers, Avitas Systems AI solutions can reduce inspection costs by 25% and reduce maintenance downtime by 15%.

The robots can handle the heat and use infrared cameras and chemical and other sensing technologies to inspect assets under dangerous conditions and keep production running. Image credit: Avitas Systems

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

QUANTUM CHROMODYNAMICS

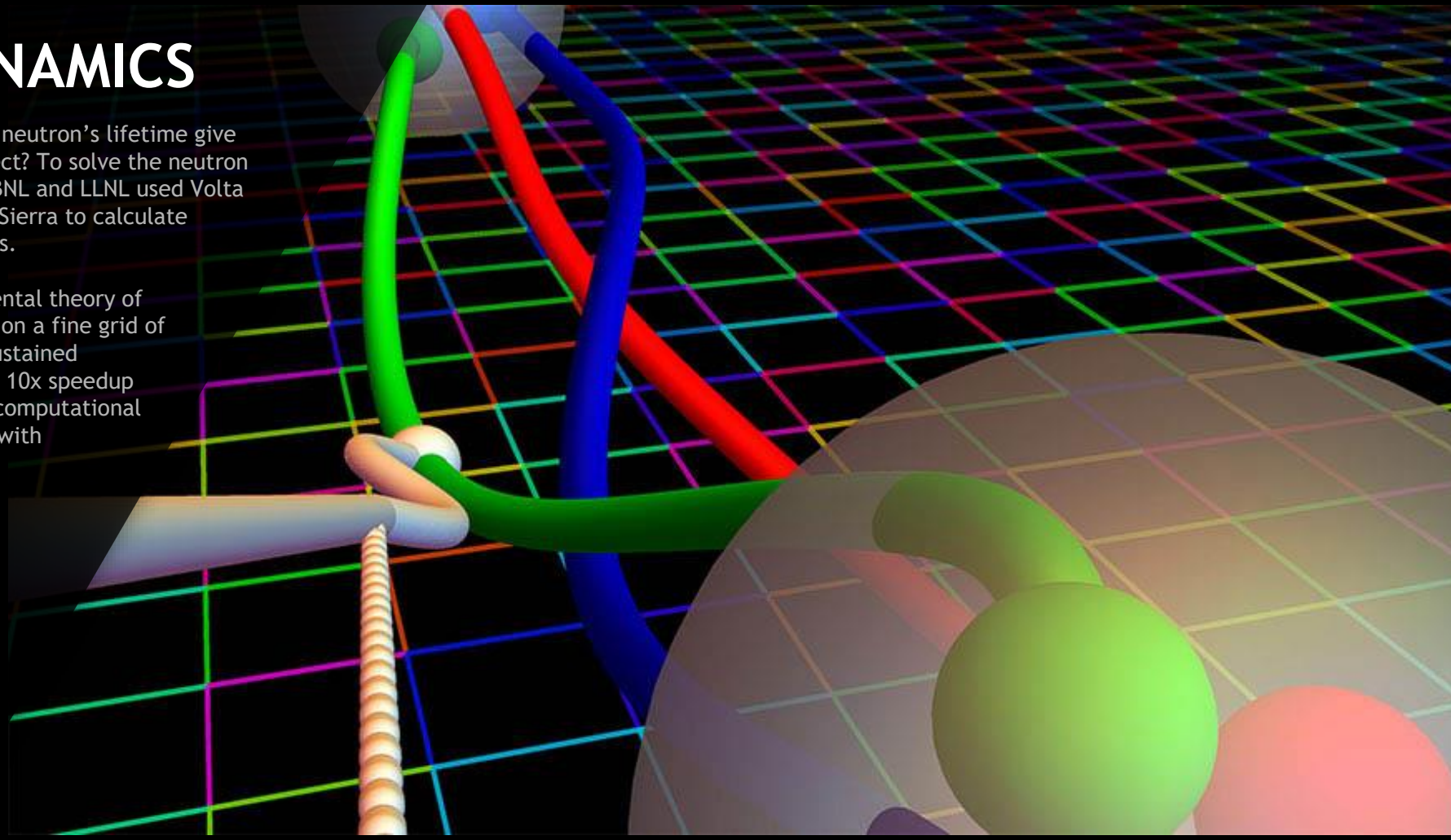
The two methods of measuring a neutron's lifetime give different answers. Which is correct? To solve the neutron lifetime puzzle scientists from LBNL and LLNL used Volta GPU-based nodes of Summit and Sierra to calculate the physics of subatomic particles.

The team simulated the fundamental theory of quantum chromodynamics (QCD) on a fine grid of space-time points. Achieving a sustained performance of ~20 petaflops —a 10x speedup over previous-gen systems— the computational breakthrough supplies physicists with the compute power to search for new physics.

NVIDIA

BERKELEY LAB

Lawrence Livermore National Laboratory



Pictured: beta decay, the decay of a neutron (n) to a proton (p) with the emission of an electron (e) and an electron-anti-neutrino (v). In the figure gA is depicted as the white node on the red line. The square grid indicates the lattice. Image credit: Evan Berkowitz/Forschungszentrum Jülich/Institut für Kernphysik /Institute for Advanced Simulation

“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

OPTIMIZING QUALITY INSPECTIONS WITH AI

Due to the increasingly sophisticated design of its cars and the high-quality standards at Audi, the company inspects all components —doors, engine hoods, fenders, etc.— in its press shop.

In addition to visual inspection by Audi employees, several small deep learning-based cameras trained on NVIDIA GPUs installed directly in the presses detect the finest cracks in sheet metal with the utmost precision in a matter of seconds.

NVIDIA.

AN EYE FOR CRISIS MANAGEMENT

Natural disasters are increasingly causing major destruction to life, property and economies. DFKI is using the NVIDIA DGX-2 to evolve DeepEye —which uses satellite images enriched with social media content to identify the effects of natural disasters— into a crisis management solution. With the increased GPU memory and fully connected GPUs based on the NVSwitch architecture, DFKI can build bigger models and process more data to aid rescuers in their decision-making for faster, more efficient dispatching of resources.

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

THANK YOU

Reynaldo Gomez
reynaldog@nvidia.com