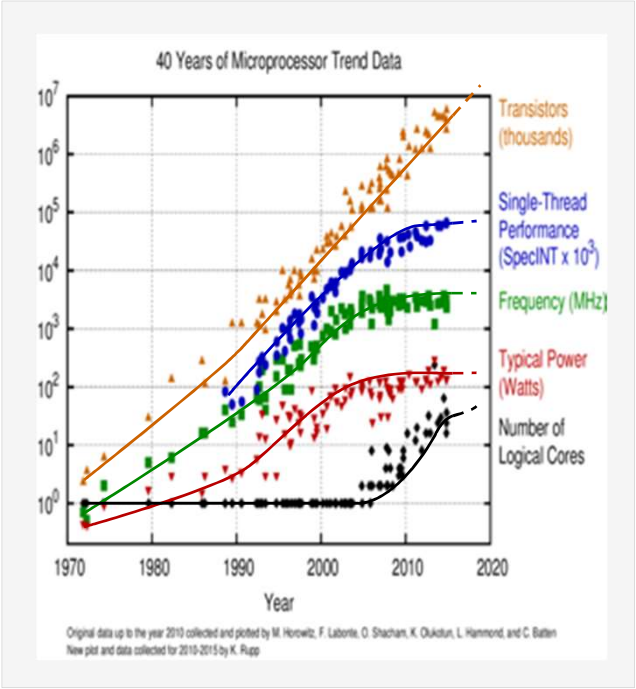
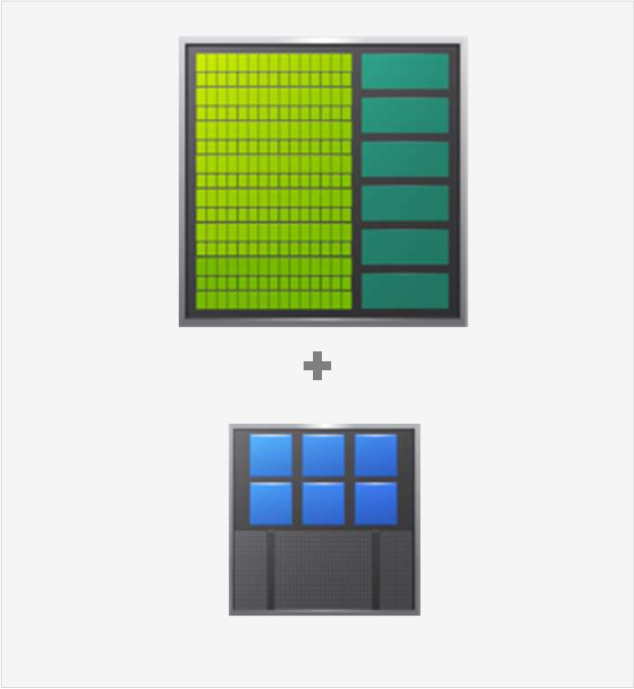
# NVIDIA CUDA PLATFORM

Paul Graham, Senior Solutions Architect, NVIDIA
ECHEP February 2020

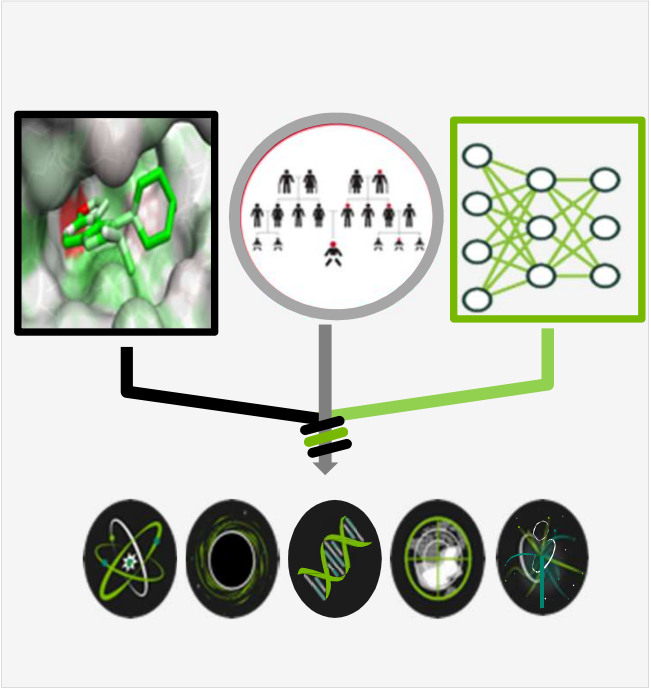
# FORCES SHAPING HIGH PERFORMANCE COMPUTING




END OF MOORES LAW

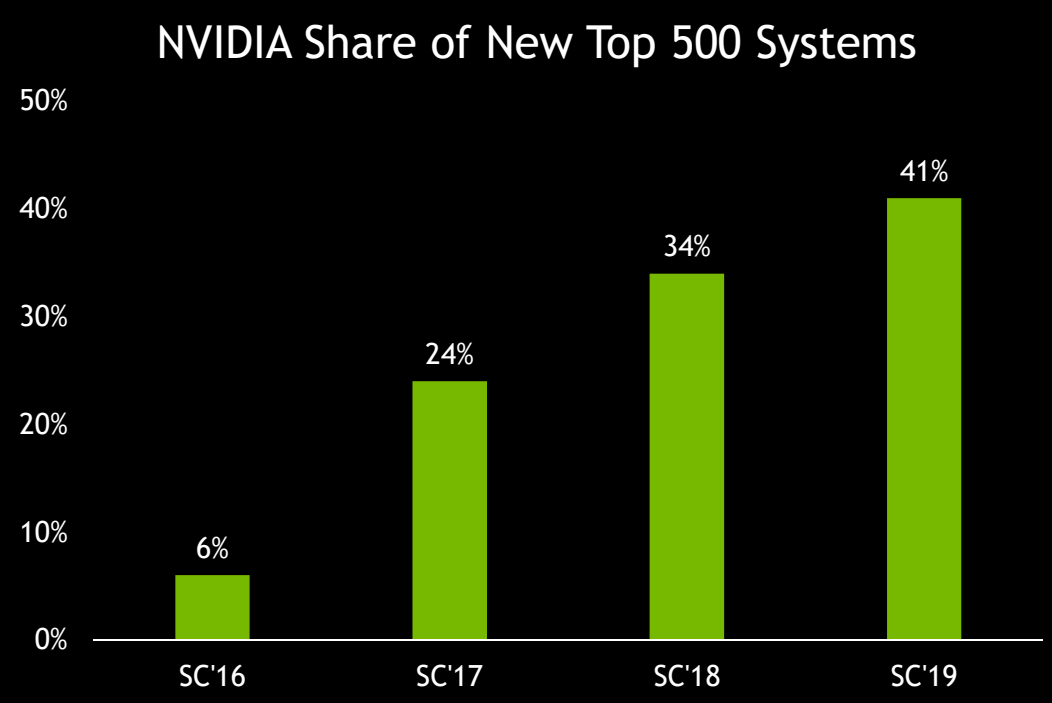


ACCELERATED COMPUTING



AI – A NEW TOOL FOR SCIENCE

# NVIDIA ACCELERATED COMPUTING IS ACCELERATING





**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

**ENI HPC5**
Fastest Industrial
7,280 GPUs| 52 PF

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

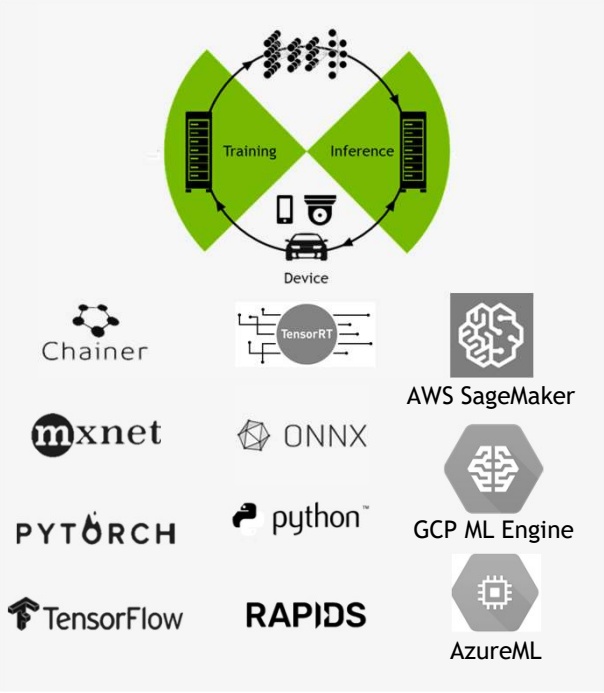
# MOST ADOPTED PLATFORM FOR ACCELERATING AI

## 8 MLPerf 0.6 Training Records

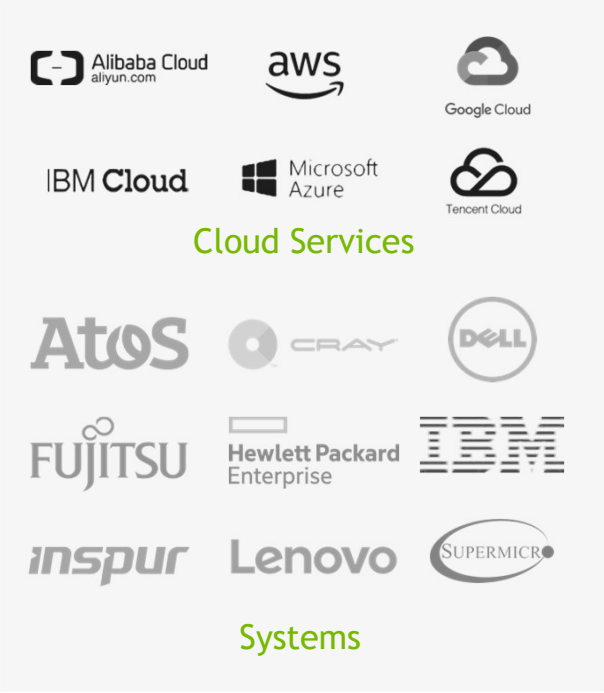
MLPerf

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

**RECORD-SETTING PERFORAMNCE**




**END-TO-END SOFTWARE STACK**




**AVAILABLE EVERYWHERE**

# NVIDIA ACCELERATED 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 | PRO VIZ & GRAPHICS

**APPS & FRAMEWORKS**

python | TensorFlow | mxnet | Chainer | ONNX | R | PYTORCH

Amber | NAMD | +600 Applications

CATIA | Ps | 3DS MAX

**CUDA-X & NVIDIA SDKs**

| MACHINE LEARNING | DEEP LEARNING | HPC | VISUALIZATION |
|---|---|---|---|
| cuDF, cuML, cuGRAPH | cuDNN, CUTLASS, TensorRT | cuBLAS, cuFFT, OpenACC | OPTIX, NVEncode, NVDecode |

CUDA & Magnum IO

**VIRTUALIZATION**

Virtual Compute Server | QvDWS | GRID vPC | GRID vApps

vGPU

**NVIDIA GPUs & SYSTEMS**

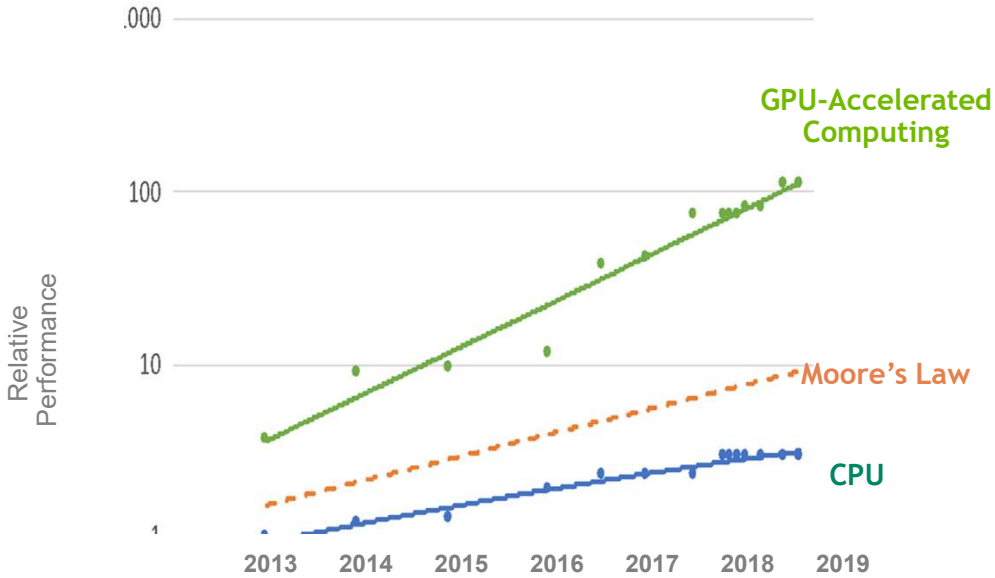
NVIDIA GPU | NVIDIA DGX FAMILY | NVIDIA HGX | NVIDIA EGX | EVERY OEM | EVERY MAJOR CLOUD

# PROGRESS OF STACK IN 6 YEARS

## CONVENTIONAL HPC BEYOND MOORE'S LAW

**2014**

- APPLICATIONS
- TOOLS
- COMPILERS
- ALGORITHMS
- LIBRARIES
- CUDA




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

**2019**

- WORKFLOW TOOLS
- CONTAINERS
- APPLICATIONS
- TOOLS
- COMPILERS
- ALGORITHMS
- LIBRARIES
- CUDA

# 3X MORE PERFORMANCE IN 2 YEARS

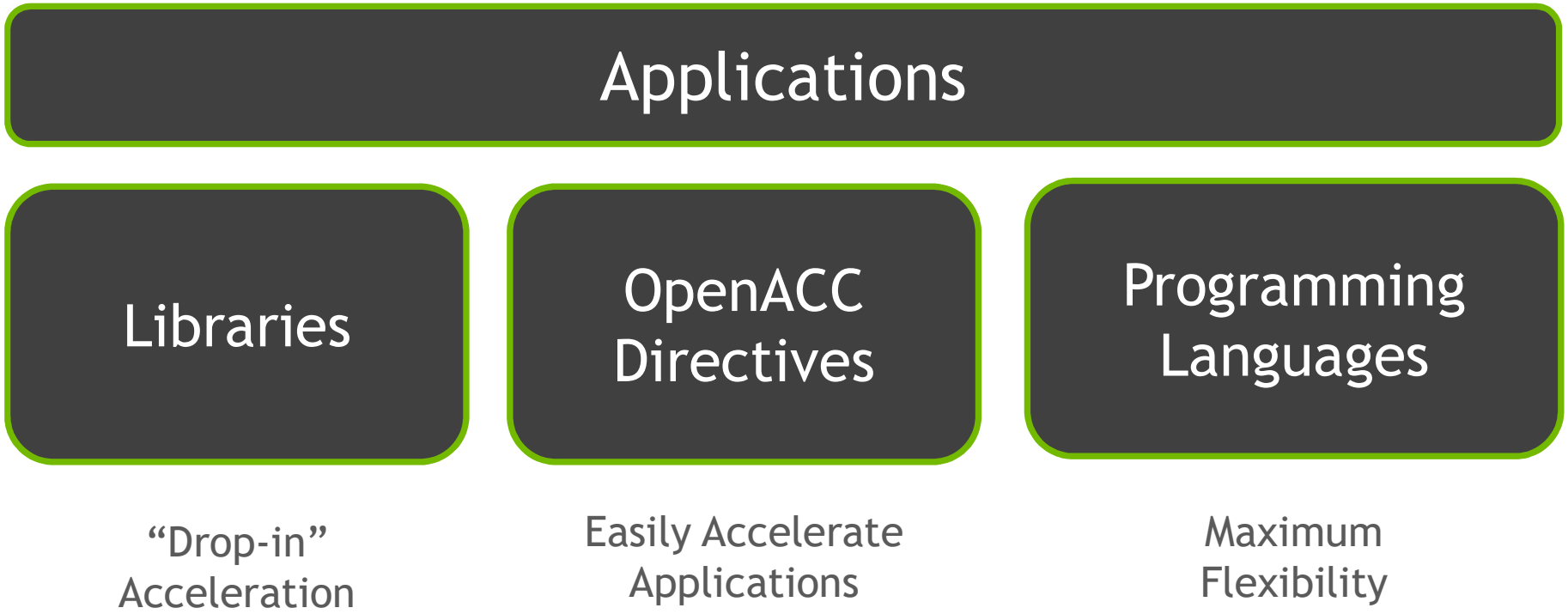
## Beyond Moore's Law


Amber Chroma
GROMACS GTC
LAMMPS
MILC
NAMD
QE
SPECFEM3D
TensorFlow
VASP

Benchmark Application: Amber [PME-Cellulose_NVE], Chroma [szscl21_24_128], GROMACS [ADH Dodec], GTC [moi#proc.in], LAMMPS [LJ 2.5], MILC [Apex Medium], NAMD [stmv_nve_cuda], Quantum Espresso [AUSURF112-jR], SPECFEM3D [four_material_simple_model]; TensorFlow [ResNet 50] VASP [Si Huge]; [GPU node: with dual-socket CPUs with 4x V100 GPU .

NVIDIA

# WAYS TO ACCELERATE APPLICATIONS



# 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

Thrust

# CUDA ENHANCEMENTS

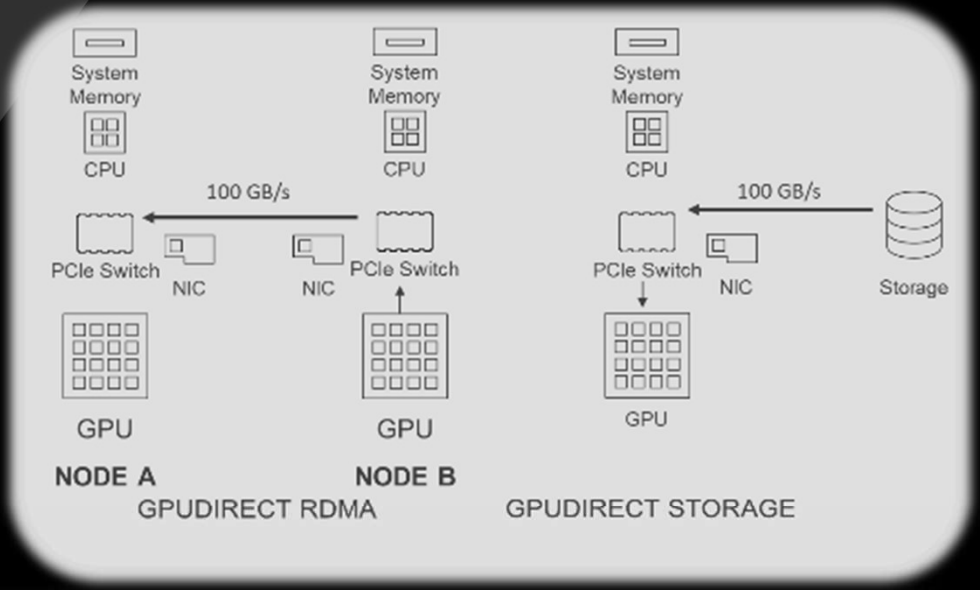
- **CUDA Graphs** allow workflows to be submitted to GPU rather than single operations, to reduce overheads and allow more holistic optimizations.
- Hierarchical parallelism is becoming increasingly important (within and across GPUs)
  - **Cooperative Groups** allow the programmer to map application-level parallelism to the hardware in a flexible and efficient manner.
  - **Multi-GPU programming** techniques are becoming more sophisticated and performant.
- Programming difficulty associated with complex hardware can be alleviated with use of **Unified Memory**. This makes it easier for users to get started with GPUs.
- There is an increasing awareness of the fact that use of **Reduced Precision** is feasible in many cases, allowing improved performance. Hardware and software support continues to evolve.

# NVIDIA MAGNUM IO

GPU-Accelerated I/O and Storage Software to Eliminate Data Transfer Bottlenecks for AI, Data Science and HPC Workloads

High-Bandwidth, Low-Latency Massive Storage Access with Lower CPU Utilization

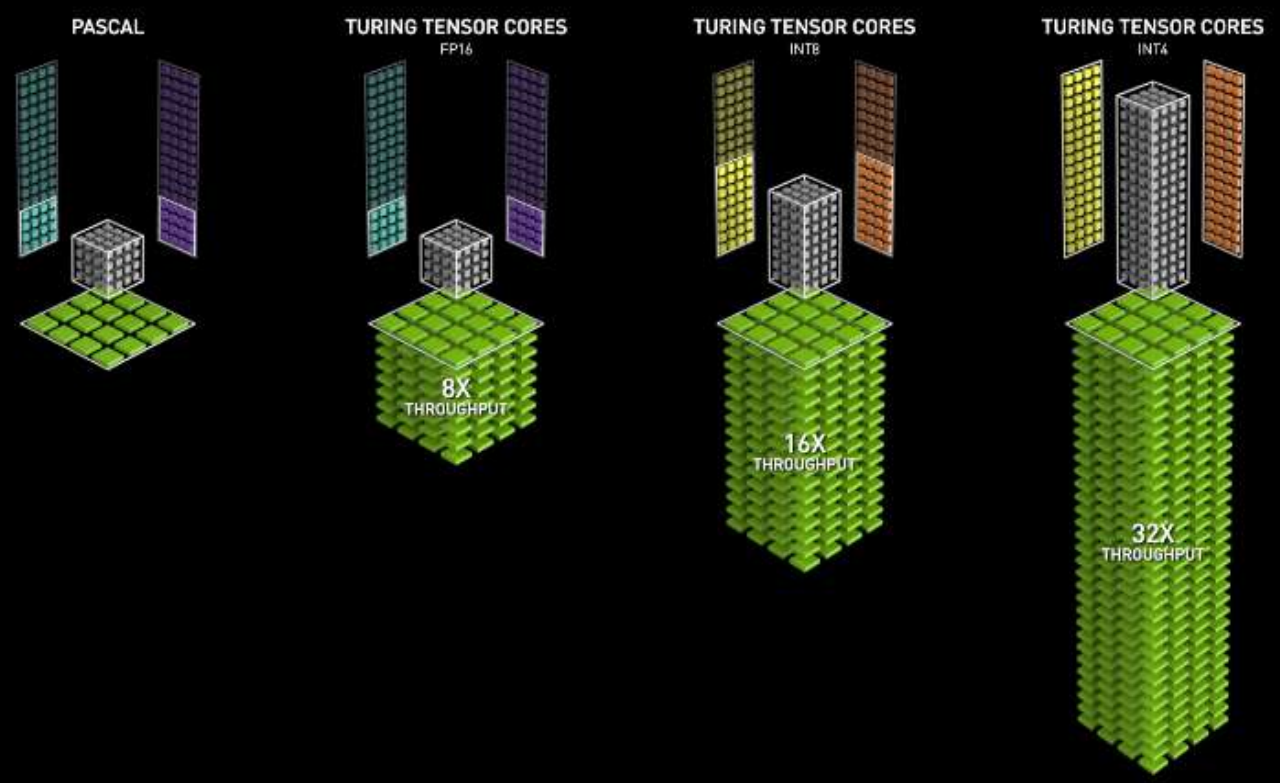
Delivers up to 20x faster data throughput on multi-server, multi-GPU computing nodes




DDN · IBM Spectrum Scale · NetApp · PURESTORAGE · WEKA.IO · Excelero · LIQID · Mellanox · MICROCHIP · Micron · CRAY · DELL EMC · Hewlett Packard Enterprise · IBM · inspur · Atos · Lenovo · PENGUIN COMPUTING · SUPERMICRO

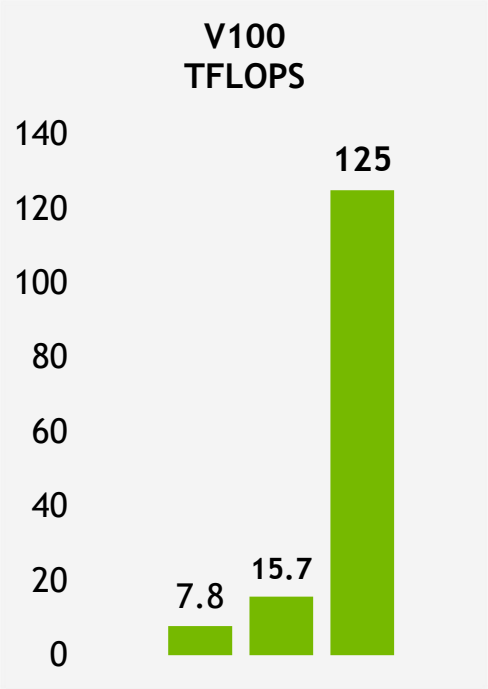
# NEW TURING TENSOR CORE

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

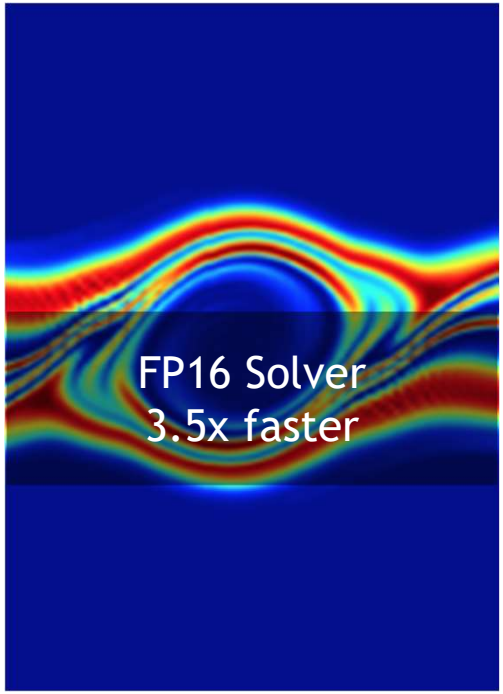


# TENSOR CORES FOR SCIENCE

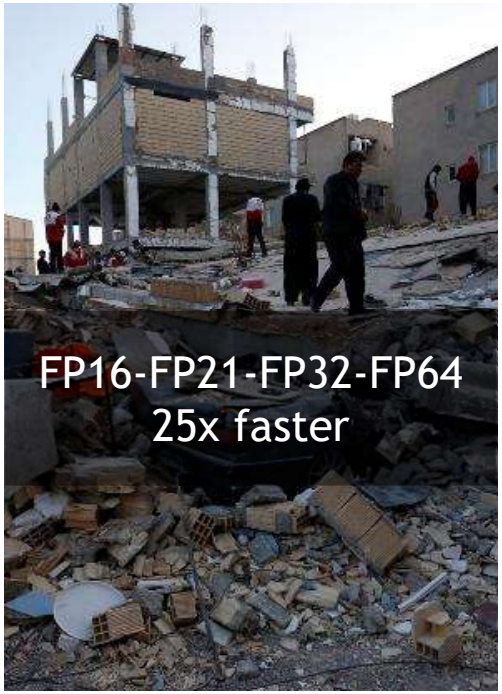
## Mixed-Precision Computing



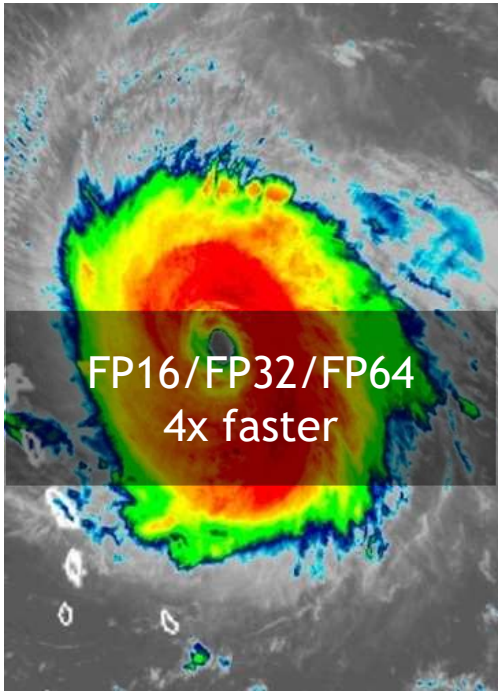

FP64+ MULTI-PRECISION




PLASMA FUSION APPLICATION




EARTHQUAKE SIMULATION

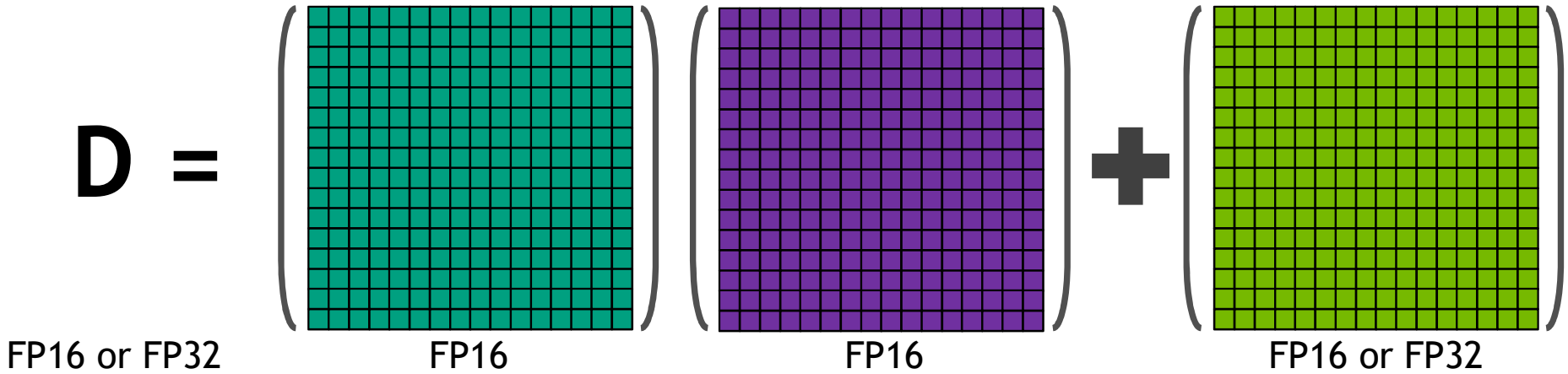



MIXED PRECISION WEATHER PREDICTION

# CUDA TENSOR CORE PROGRAMMING

## 16x16x16 Warp Matrix Multiply and Accumulate (WMMA)

```
wmma::mma_sync(Dmat, Amat, Bmat, Cmat);
```

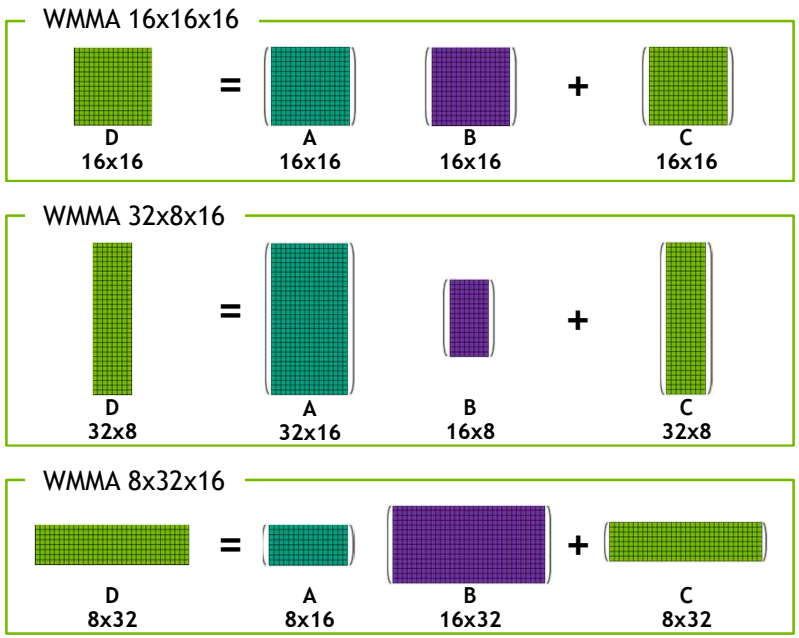



$$D = AB + C$$

# TURING TENSOR CORE

## New Warp Matrix Functions

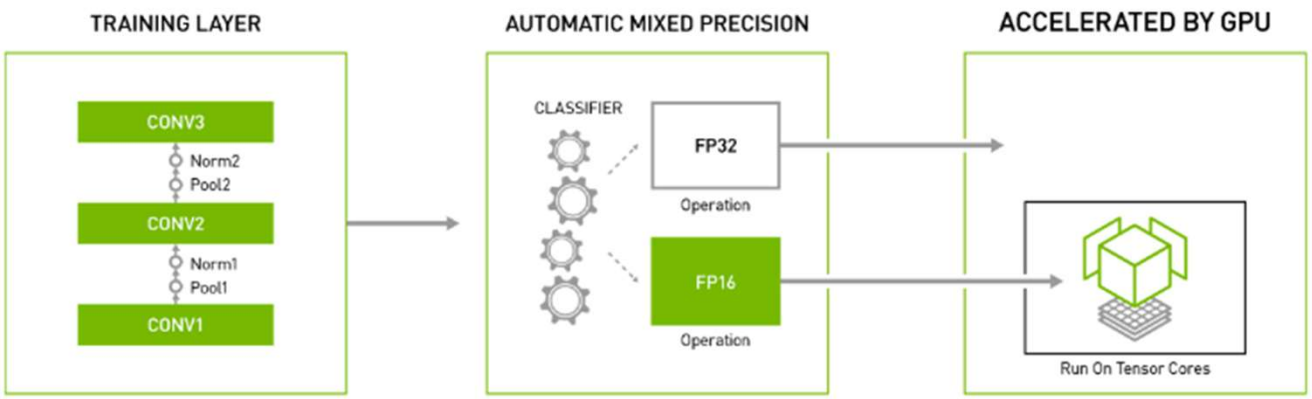
- **WMMA operations now include 8-bit integer**
- Turing (sm_75) only
- Signed & unsigned 8-bit input
- 32-bit integer accumulator
- Match input/output dimensions with *half*
- 2048 ops per cycle, per SM



# AUTOMATIC MIXED PRECISION

NEW

## Easy to Use, Greater Performance and Boost in Productivity

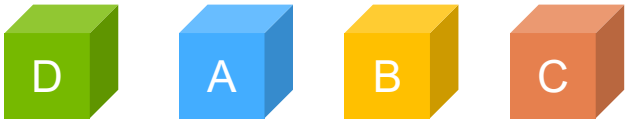



Insert ~ two lines of code to introduce Automatic Mixed-Precision and get upto 3X speedup

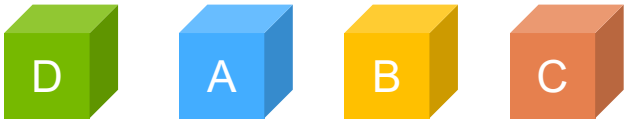
AMP uses a graph optimization technique to determine FP16 and FP32 operations

Support for TensorFlow, PyTorch and MXNet

**Unleash the next generation AI performance and get faster to the market!**




# cuTENSOR

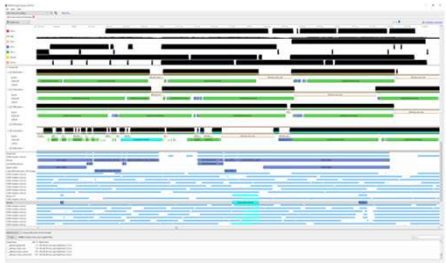



## A New High Performance CUDA Library for Tensor Primitives

- Tensor Contractions
- Elementwise Operations
- Mixed Precision
- Coming Soon
  - Tensor Reductions
  - Out-of-core Contractions
  - Tensor Decompositions
- Pre-release version available developer.nvidia.com/cuTENSOR

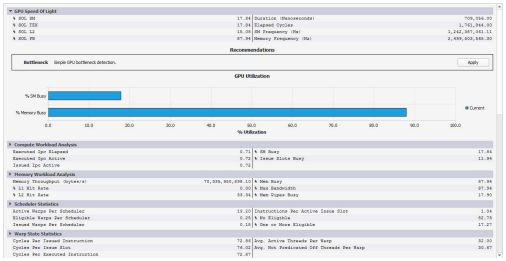
```
cutensorStatus_t cutensorCreateTensorDescriptor ( cutensorTensorDescriptor_t* desc,
                                                  unsigned int numModes,
                                                  const int64_t *extent,
                                                  const int64_t *stride,
                                                  cudaDataType_t dataType,
                                                  cutensorOperator_t unaryOp );


cutensorStatus_t cutensorContraction (cuTensorHandle_t handle,
       const void* alpha, const void *A, const cutensorTensorDescriptor *descA, const int modeA[],
                          const void *B, const cutensorTensorDescriptor *descB, const int modeB[],
       const void* beta,  const void *C, const cutensorTensorDescriptor *descC, const int modeC[],
                                 void *D, const cutensorTensorDescriptor *descD, const int modeD[],
       cutensorOperator_t opOut, cudaDataType_t typeCompute, cutensorAlgo_t algo,
       void* workspace, size_t workspaceSize, cudaStream_t stream );


cutensorStatus_t cutensorElementwiseTrinary ( cuTensorHandle_t handle,
       const void* alpha, const void *A, const cutensorTensorDescripto r *descA, const int modeA[],
       const void* beta,  const void *B, const cutensorTensorDescriptor *descB, const int modeB[],
       const void* beta,  const void *C, const cutensorTensorDescriptor *descC, const int modeC[],
                                 void *D, const cutensorTensorDescriptor *descD, const int modeD[],
       cutensorOperator_t opAB, cutensorOperator_t opABC, cudaDataType_t typeCompute, cudaStream_t stream );
```

# NSIGHT PRODUCT FAMILY



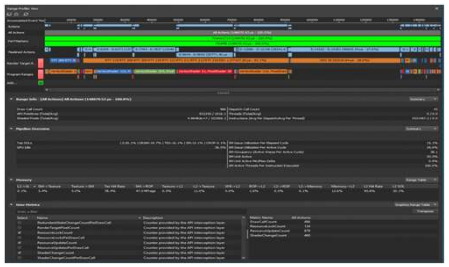
## Nsight Systems

System-wide application algorithm tuning



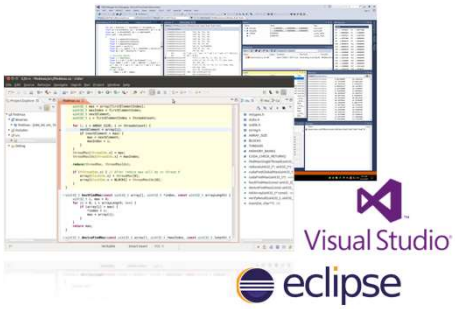
## Nsight Compute

CUDA Kernel Profiling and Debugging



## Nsight Graphics

Graphics Shader Profiling and Debugging




## IDE Plugins

Nsight Eclipse Edition/Visual Studio (Editor, Debugger)

# ANNOUNCING CUDA TO ARM

## ENERGY-EFFICIENT SUPERCOMPUTING

NVIDIA GPU Accelerated Computing Platform On ARM

Optimized CUDA-X HPC & AI Software Stack

CUDA, Development Tools and Compilers

Available End of 2019

NVIDIA

&

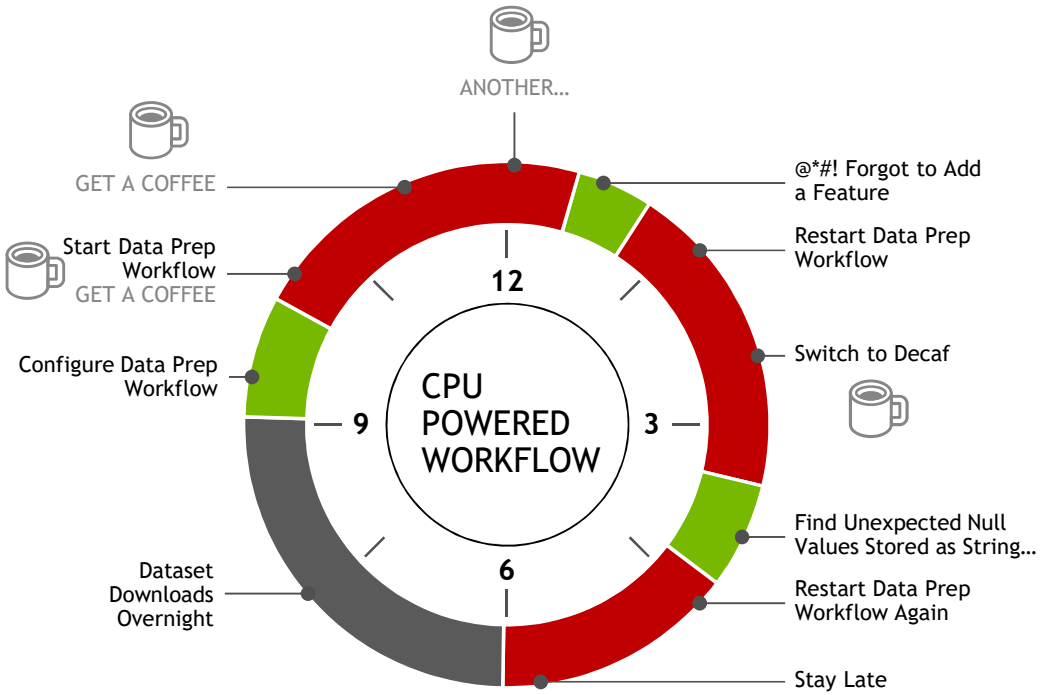
arm

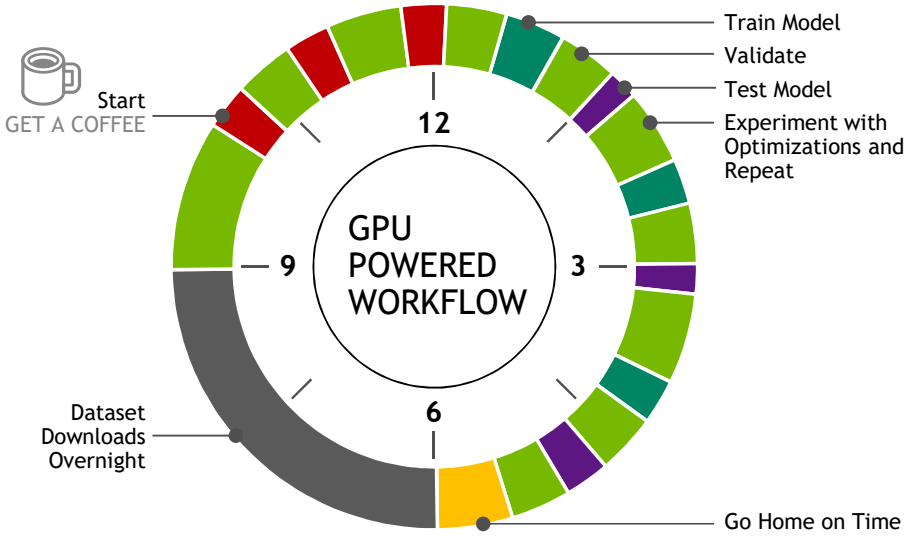
Atos

CRAY

Hewlett Packard Enterprise

# DAY IN THE LIFE OF A DATA SCIENTIST

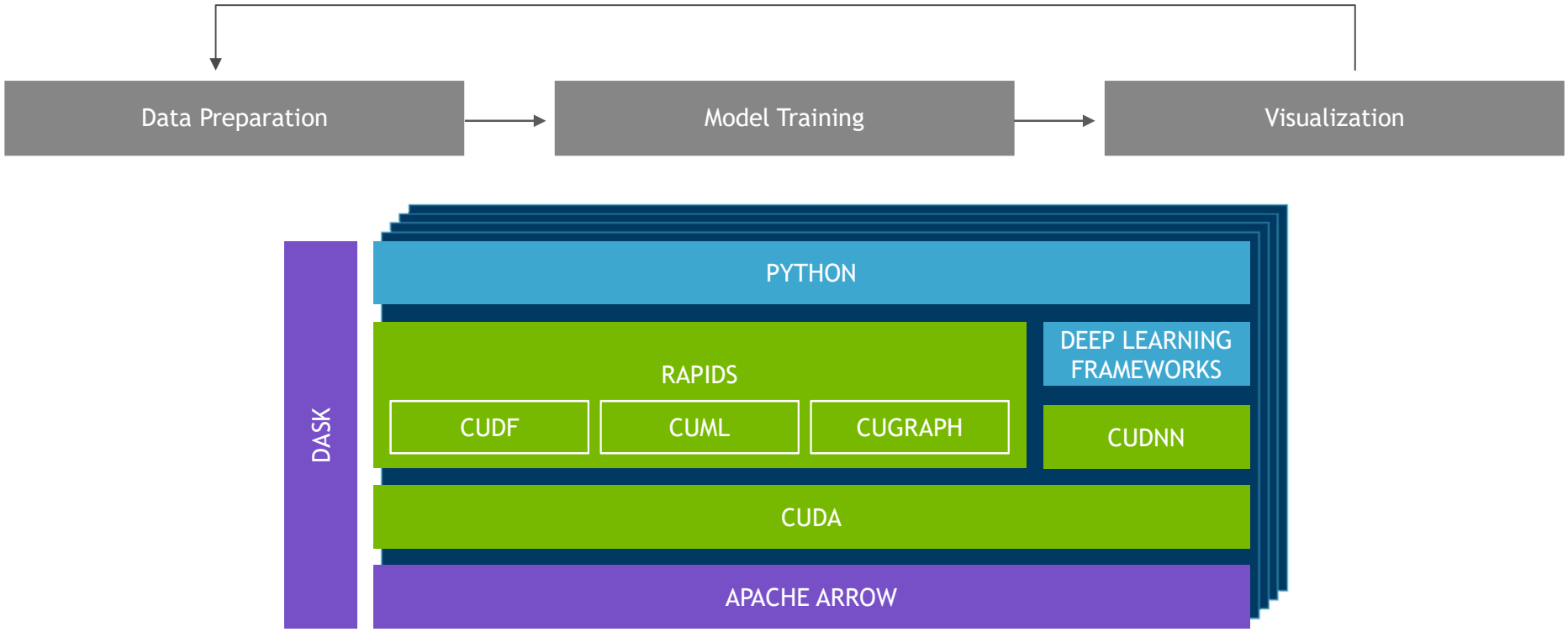



Dataset Collection | Analysis | Data Prep | Train | Inference

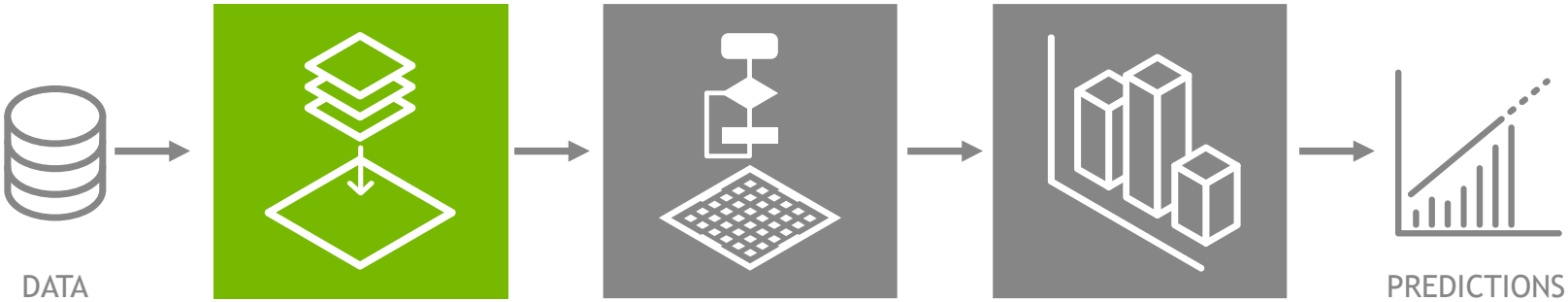
# RAPIDS — OPEN GPU DATA SCIENCE

## Software Stack



# GPU-ACCELERATED DATA SCIENCE WORKFLOW

## NVIDIA Accelerated Data Science Solution, Built on CUDA-X AI




### DATA PREPARATION

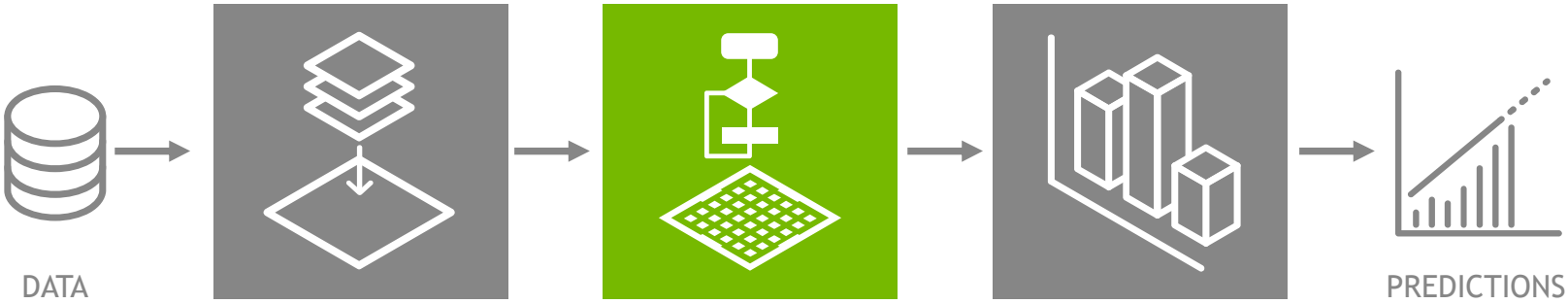
GPUs accelerated compute for in-memory data preparation

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

## NVIDIA Accelerated Data Science Solution, Built on CUDA-X AI



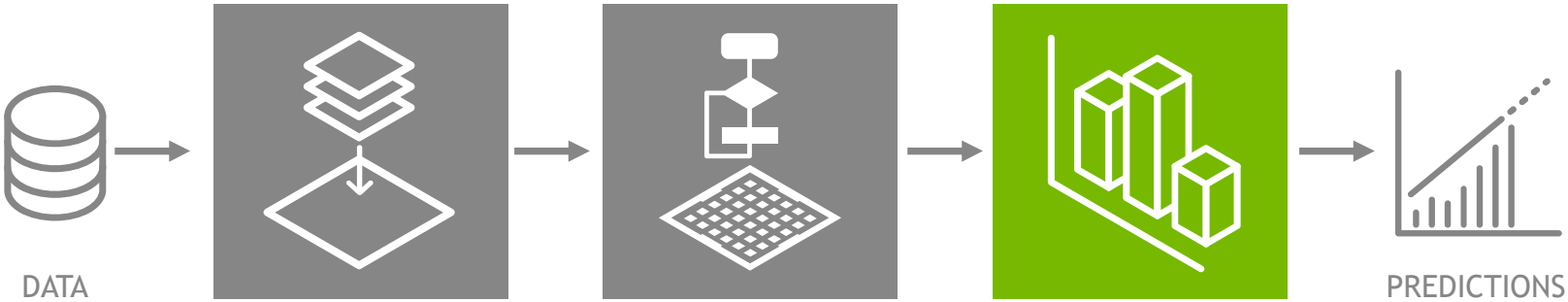

### MODEL TRAINING

GPU-acceleration of today's most popular ML algorithms

XGBoost, PCA, K-means, k-NN, DBScan, tSVD …

# GPU-ACCELERATED DATA SCIENCE WORKFLOW

## NVIDIA Accelerated Data Science Solution, Built on CUDA-X AI



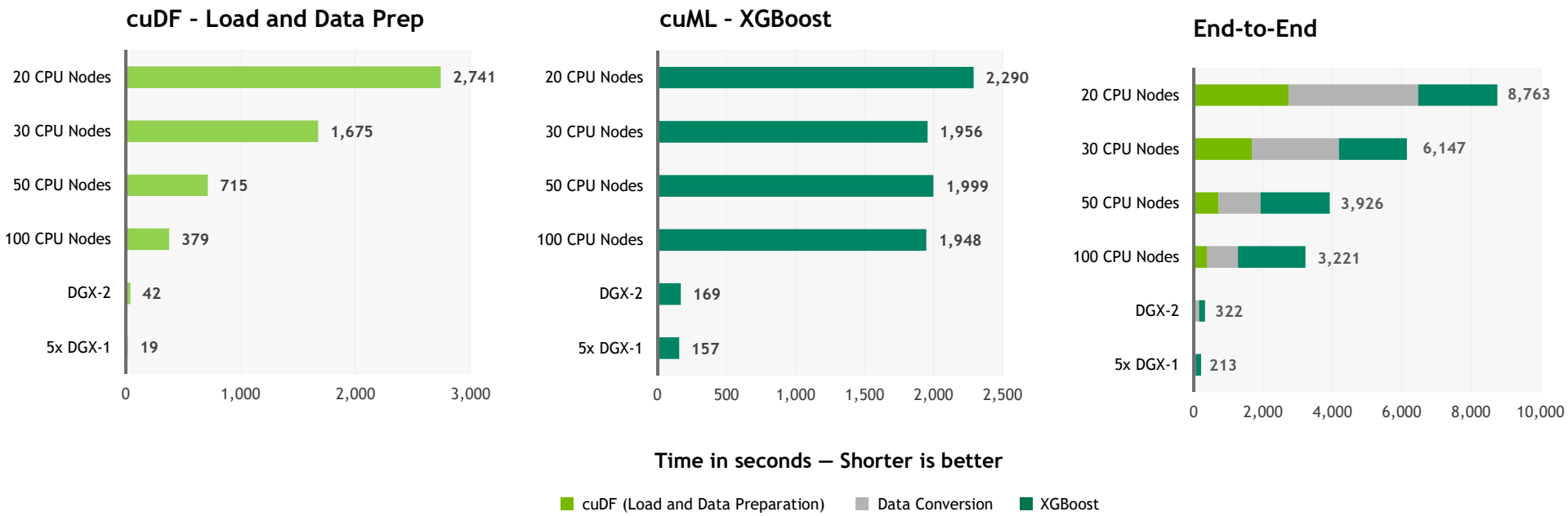

### VISUALIZATION

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

# BENCHMARKS




### 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

# DRAMATICALLY MORE FOR YOUR MONEY

## CPU-Only Cluster


300 Self-hosted Broadwell CPU Servers
180 KWatts

=

## GPU-Accelerated

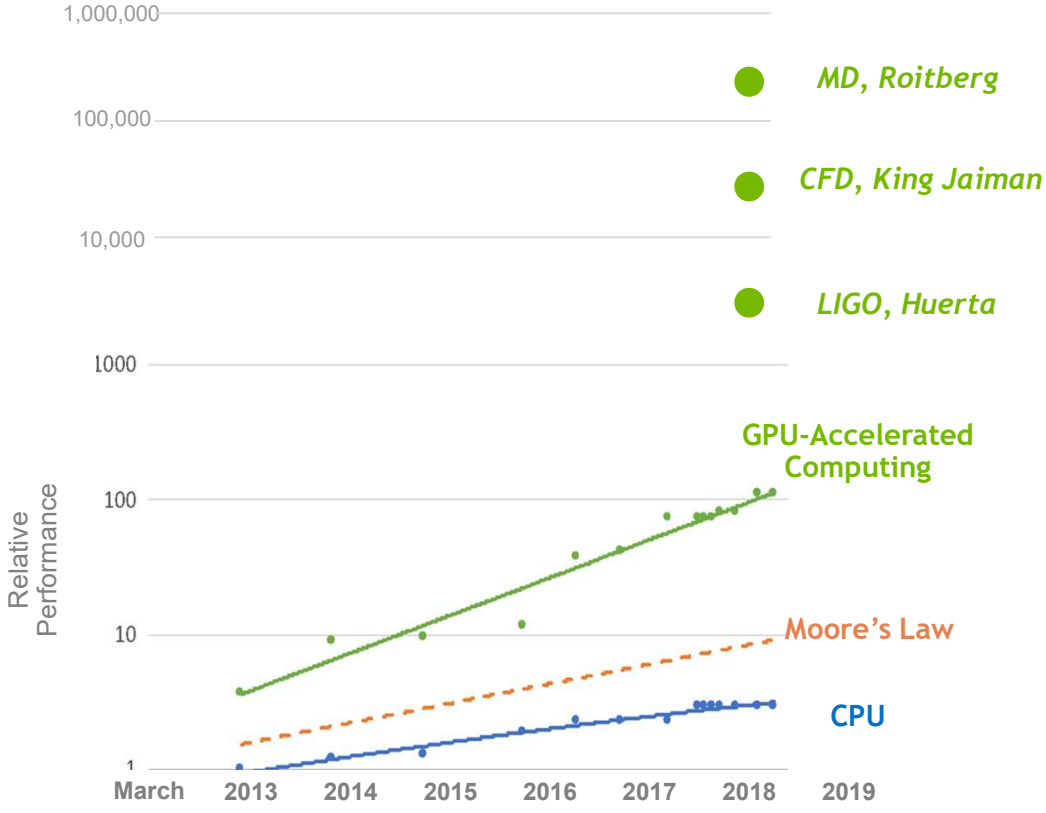

1 DGX-2
10 KWatts


THE #1 DATA SCIENTIST EXCUSE FOR LEGITIMATELY SLACKING OFF:
"MY MODEL'S TRAINING."
HEY! GET BACK TO WORK!
TRAINING!
OH. CARRY ON.

# CONVERGED HPC*AI CHANGES THE GAME

**2014**

- APPLICATIONS
- TOOLS
- COMPILERS
- ALGORITHMS
- LIBRARIES
- CUDA




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

**2019**

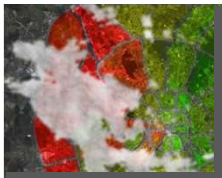
- WORKFLOW TOOLS
- CONTAINERS
- APPLICATIONS
- TOOLS
- COMPILERS
- ALGORITHMS
- LIBRARIES
- CUDA

# CONVERGED HPC*AI TAXONOMY

## How AI Algorithms are Being Applied in the HPC Workflow

### Modelling and Simulation

**Ab Initio Algorithm Enhancement**

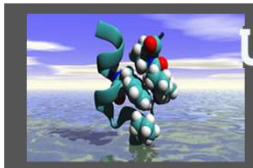


UCI University of California, Irvine

Columbia University

20X Faster Time to Solution

**Reduced Order Model Replacement**

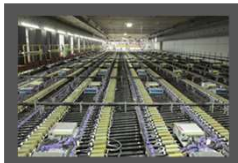


UF University of Florida

300,000X Faster Time to solution

### Experiment Data Processing

**Detection Pipeline**



U.S. Department of Energy

49% More Accurate Detection

**Real-Time Control**

Princeton University

iter

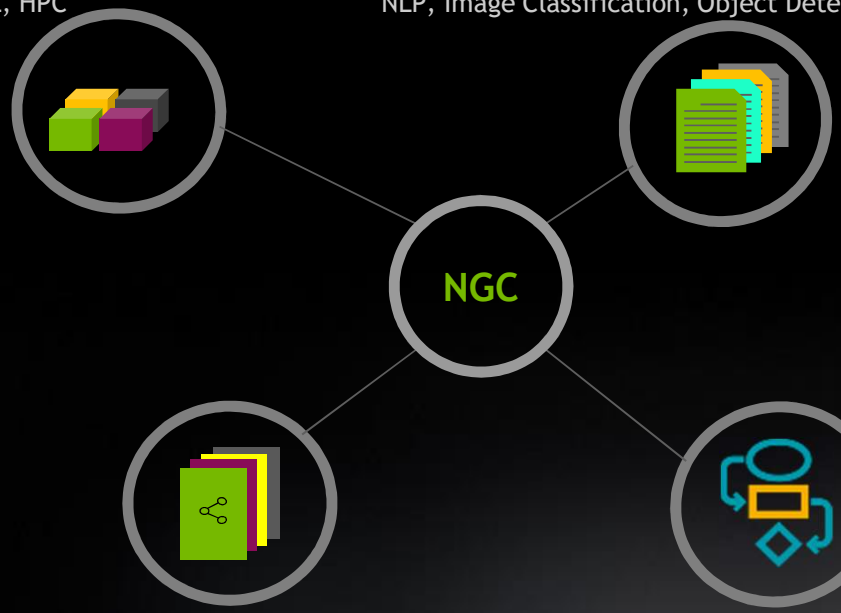
95% prediction accuracy with 5% False positives

# NGC: GPU-OPTIMIZED SOFTWARE HUB

Simplifying DL, ML and HPC Workflows

**50+ Containers**
DL, ML, HPC

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



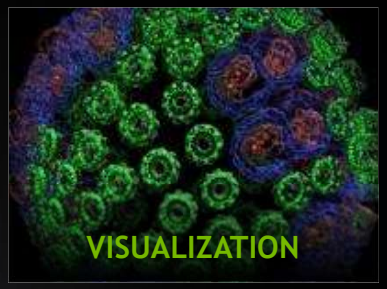

**60 Pre-trained Models**
NLP, Image Classification, Object Detection & more

**Industry Workflows**
Medical Imaging, Intelligent Video Analytics


TensorFlow | PyTorch | more


RAPIDS | H2O | more


NAMD | GROMACS | more




ParaView | IndeX | more

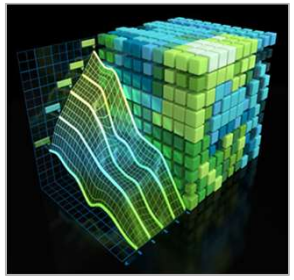
# DEEP LEARNING INSTITUTE (DLI)

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

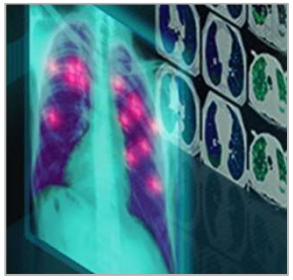
Request onsite instructor-led workshops at your organization:
www.nvidia.com/requestdli

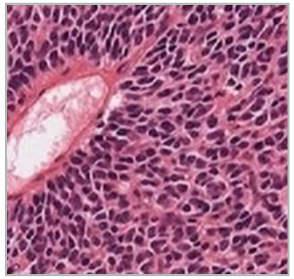
Take self-paced courses online:
www.nvidia.com/dlilabs

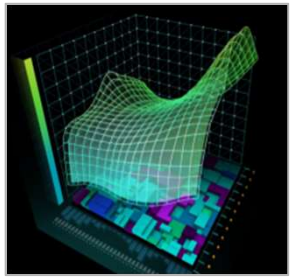
Download the course catalog, view upcoming workshops, and learn about the University Ambassador Program:
www.nvidia.com/dli


Accel. Computing Fundamentals


Autonomous Vehicles

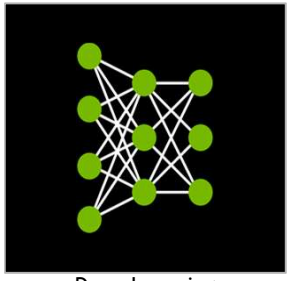

Medical Image Analysis


Genomics


Finance


Digital Content Creation


Game Development


Deep Learning Fundamentals

More industry-specific training coming soon…

# DEVELOPER ENGAGEMENT PLATFORMS

| | |
|---|---|
| Information, downloads, special programs, code samples, and bug submission | developer.nvidia.com |
| Containers for cloud and workstation environments | ngc.nvidia.com |
| Insights & help from other developers and NVIDIA technical staff | devtalk.nvidia.com |
| Technical documentation | docs.nvidia.com |
| Deep Learning Institute: workshops & self-paced courses | courses.nvidia.com |
| In depth technical how to blogs | devblogs.nvidia.com |
| Developer focused news and articles | news.developer.nvidia.com |
| Webinars | nvidia.com/webinar-portal |
| GTC on-demand content | gputechconf.com |

# RESOURCES AVAILABLE TO ACADEMICS

## TO FURTHER EDUCATION

**Developer Teaching Kits:** https://developer.nvidia.com/teaching-kits which include free access to online training for students but they have to be requested by a lecturer/professor.

**Academic Workshops:**
The NVIDIA website lists free academic workshops that our Ambassadors are giving around the world that you can go and attend: www.nvidia.co.uk/dli

**Bootcamps:**
~ 2 day tailored training events, typically for a target group

**Hackathons:**
In-depth 5-day events with access to NV *devtech* team – next UK:

**Sheffield Jul 27- Aug 2nd 2020**

Thank you

Paul Graham
pgraham@nvidia.com