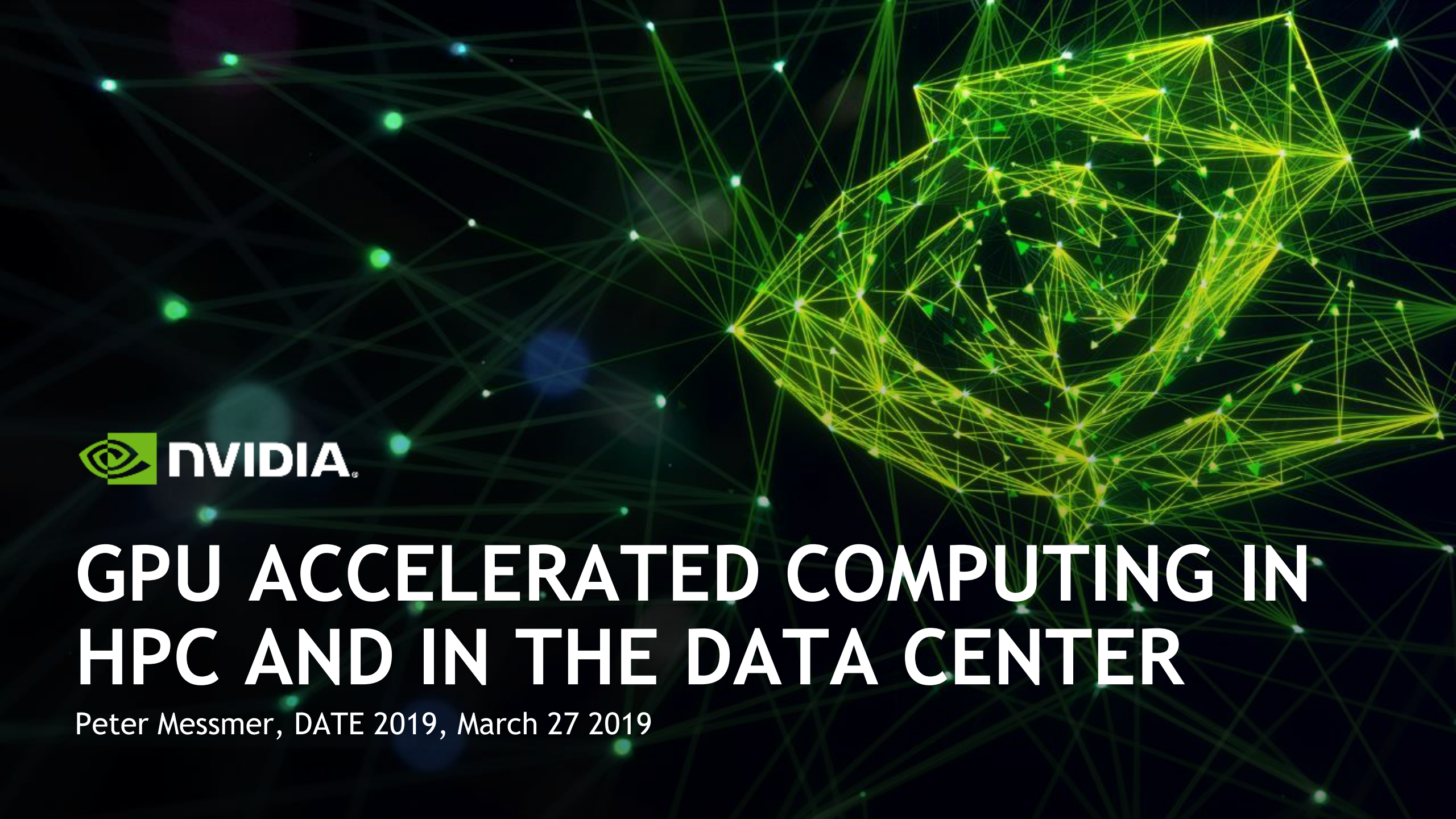
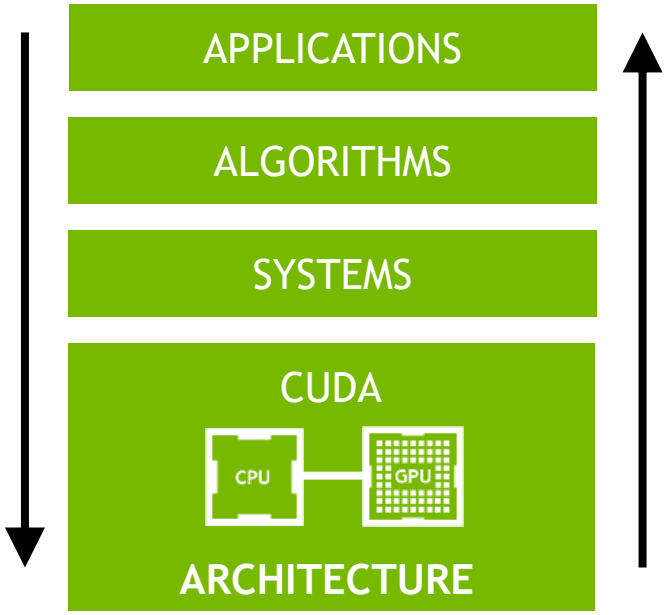
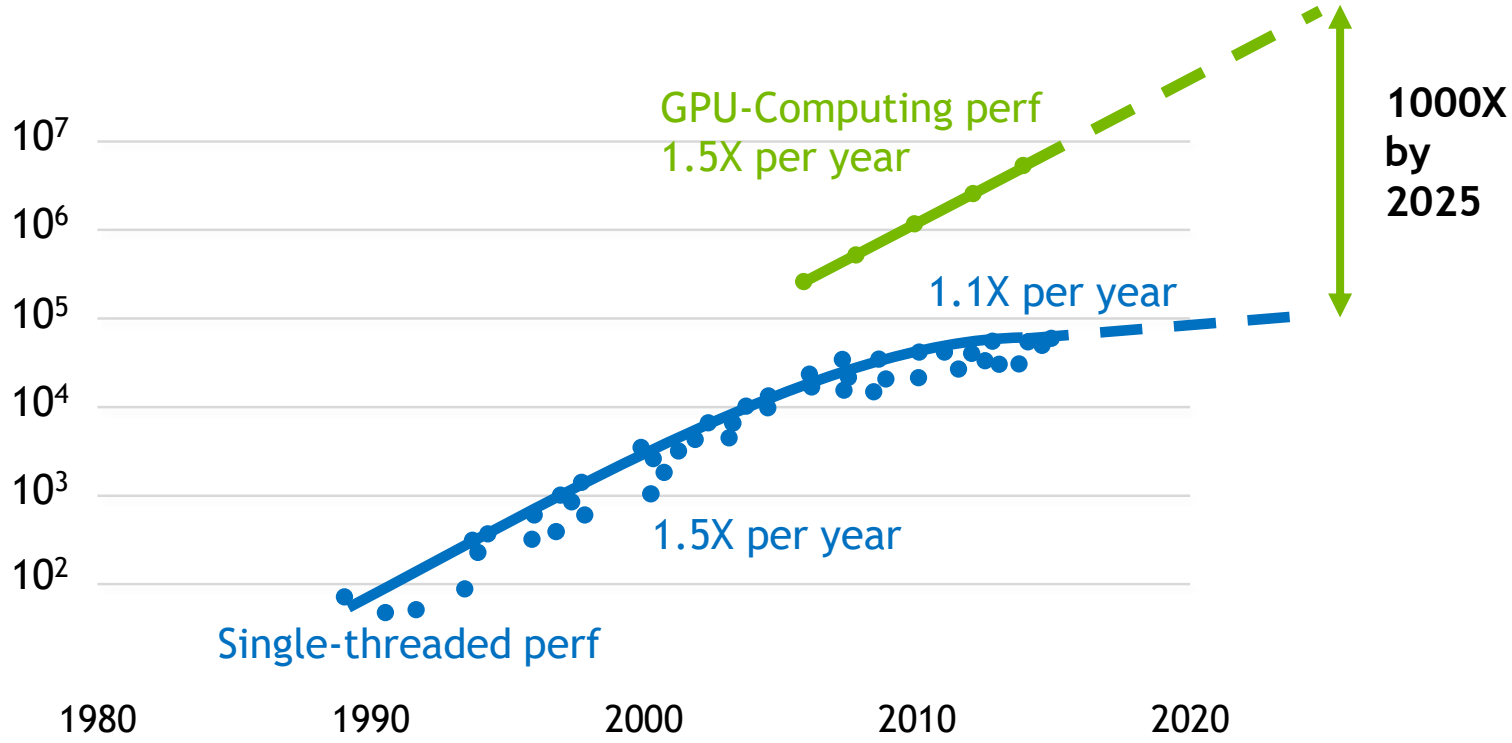


GPU ACCELERATED COMPUTING IN HPC AND IN THE DATA CENTER

Peter Messmer, DATE 2019, March 27 2019

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

NVIDIA POWERS WORLD'S FASTEST SUPERCOMPUTERS

48% More Systems | 22 of Top 25 Greenest

ORNL Summit
World's Fastest
27,648 GPUs| 144 PF

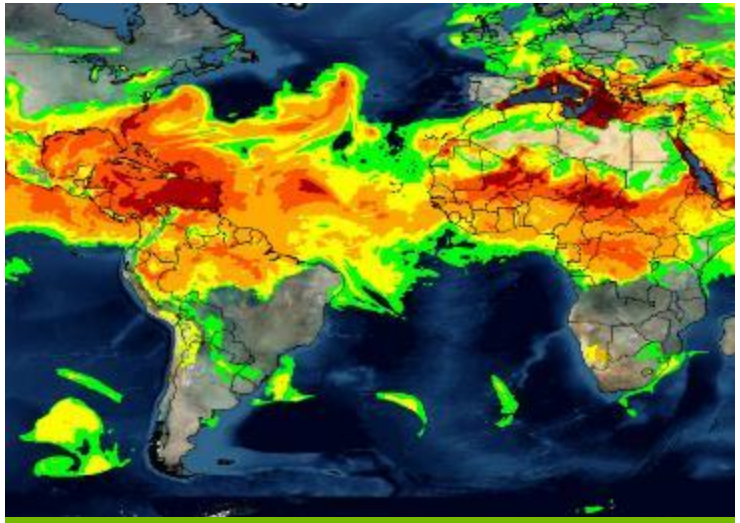
LLNL Sierra
World's 2nd Fastest
17,280 GPUs| 95 PF

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

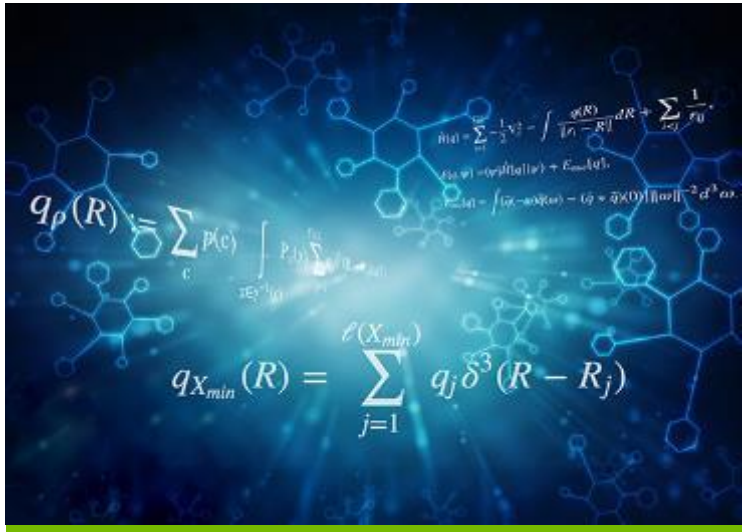
ABCI
Japan's Fastest
4,352 GPUs| 20 PF

ENI HPC4
Fastest Industrial
3,200 GPUs| 12 PF

THE NEW HPC MARKET



SIMULATION



MACHINE LEARNING

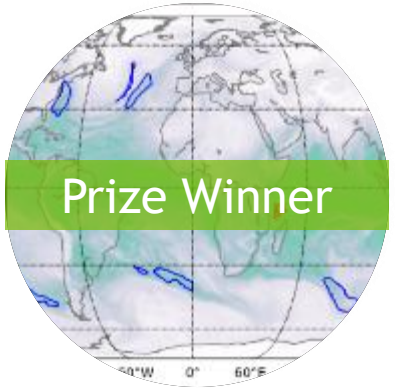
DEEP LEARNING

NVIDIA POWERS 5 OF 6 GORDON BELL NOMINATIONS

GPU Acceleration Critical To HPC At Scale Today

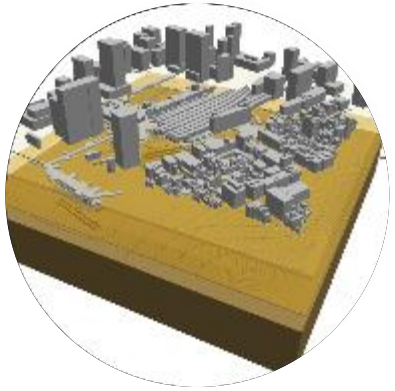
BERKELEY LAB Lawrence Berkeley National Laboratory
OAK RIDGE National Laboratory

Genomics
2.36 ExaOps



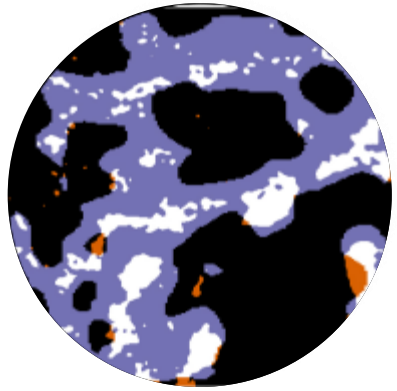
BERKELEY LAB Lawrence Berkeley National Laboratory
OAK RIDGE National Laboratory
NVIDIA

Weather
1.13 ExaOps



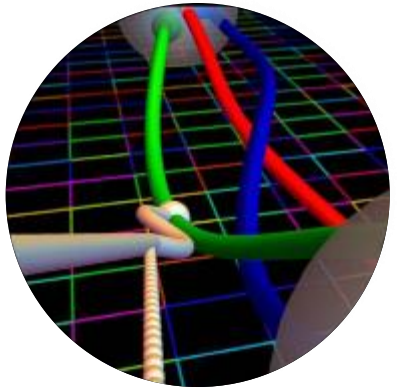
東京大学 THE UNIVERSITY OF TOKYO
OAK RIDGE National Laboratory

Seismic
1st Soil & Structure Simulation



OAK RIDGE National Laboratory

Material Science
300X Higher Performance



BERKELEY LAB Lawrence Berkeley National Laboratory
Lawrence Livermore National Laboratory
NVIDIA

Quantum Chromodynamics
<1% of Uncertainty Margin

TESLA UNIVERSAL ACCELERATION PLATFORM

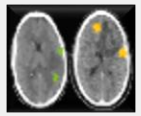
Single Platform To Drive Utilization and Productivity

CUSTOMER USECASES


Speech


Translate

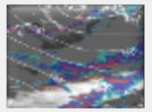

Recommender

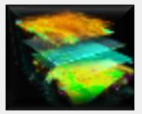

Healthcare


Manufacturing


Finance


Molecular Simulations


Weather Forecasting


Seismic Mapping

CONSUMER INTERNET | INDUSTRIAL APPLICATIONS | SUPERCOMPUTING

APPS & FRAMEWORKS

python

TensorFlow | mxnet | ONNX | PYTORCH | RAPIDS | KALDI | Chainer

Amber | ANSYS | NAMD | SIMULIA | +550 Applications

NVIDIA SDK & LIBRARIES

MACHINE LEARNING | RAPIDS: cuDF | cuML | cuGRAPH

DEEP LEARNING: cuDNN | cuBLAS | CUTLASS | NCCL | TensorRT

SUPERCOMPUTING: CuBLAS | CuFFT | OpenACC

CUDA

TESLA GPUs & SYSTEMS


TESLA GPU


VIRTUAL GPU


NVIDIA DGX FAMILY


NVIDIA HGX

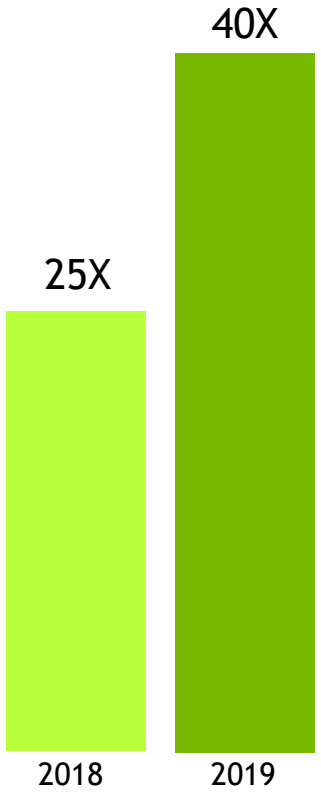
DELL | Hewlett Packard Enterprise | IBM
SYSTEM OEM

aws | Google Cloud Platform | Microsoft Azure
CLOUD

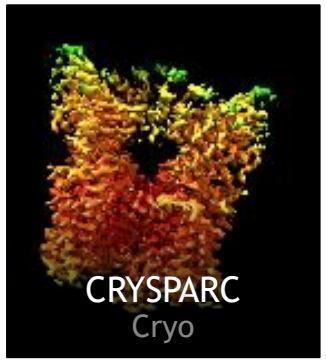
EXPANDING VALUE FOR HPC CUSTOMERS

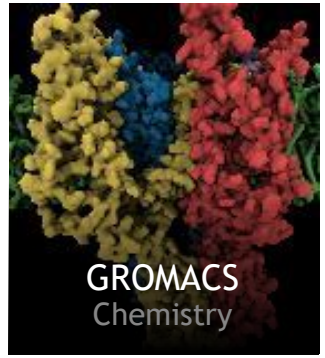
Partnering With HPC Development Community

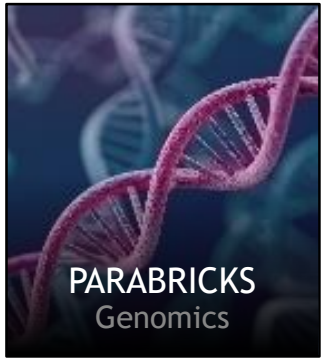
AMBER
CHROMA
GTC
LAMMPS
MILC
NAMD
QUANTUM ESP
SPECFEM3D

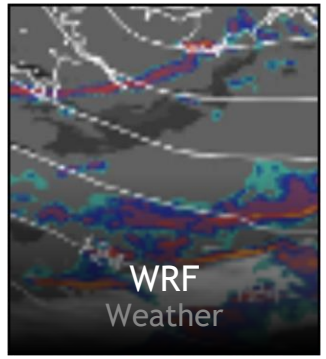


MORE PERFORMANCE WITH SAME GPU

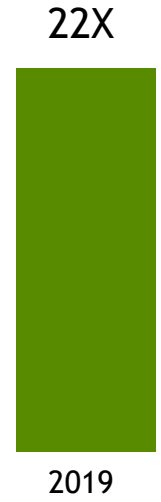








CRYOSPARC	24x
FUN3D	24x
GROMACS	7x
MICROVOLUTION	48x
PARABRICKS	22x
WRF	8x



ADDING NEW AND IMPROVED TOP APPLICATIONS

CPU Server: Dual Xeon Gold 6140@2.30GHz, GPU Servers: same CPU server w/ 4 NVIDIA V100 PCIe or SXM2 GPUs

CUDA DEVELOPMENT ECOSYSTEM

GPU Users

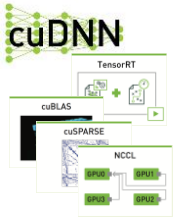
Domain Specialists

Problem Specialists

New Algorithm Developers and Optimization Experts

Applications

Frameworks



Libraries

Directives and Standard Languages

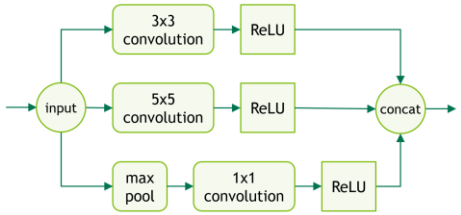
Extended Standard Languages

Ease of use ——— *Specialized Performance* →

CUDA: Programming Model, GPU Architecture, System Architecture

NEW PROGRAMMING MODEL FEATURES

Execution



Asynchronous Task Graphs

Precision

```
atomicAdd(&h, (half)1.15f);

half2 hvec(0.94f, -2.13f);
atomicAdd(&h2, hvec);
```

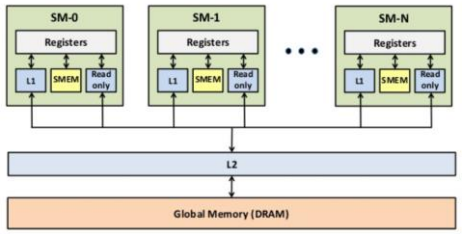
IEEE-754.2008 FP16 Specification

FP16 Operations

Interop

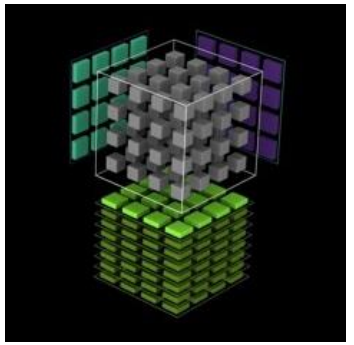
Lightweight Graphics Interop

Efficiency



NVCC Enhancements

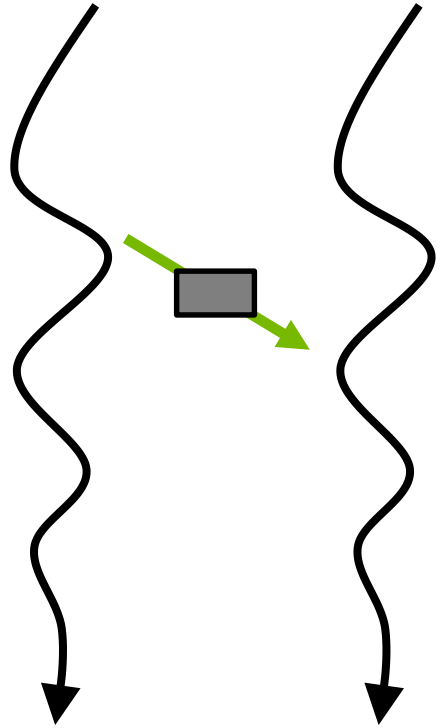
Turing



Multi-Precision Tensor Cores

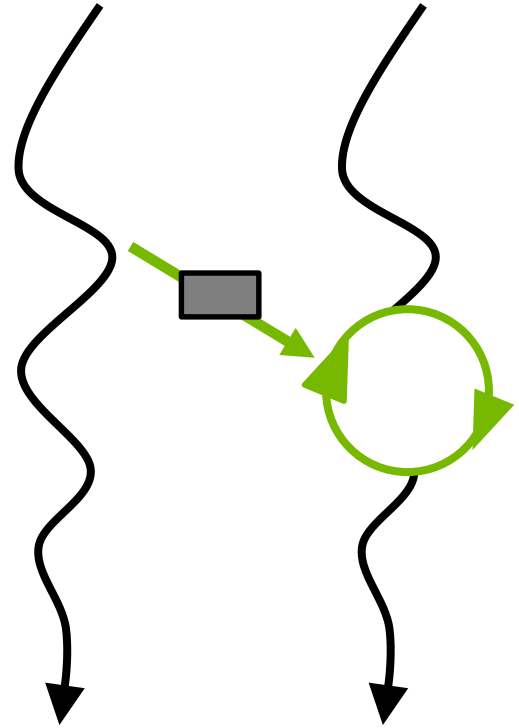
INDEPENDENT THREAD SCHEDULING

Communicating Algorithms



Pascal: Lock-Free Algorithms

Threads cannot wait for messages



Volta/Turing: Starvation Free Algorithms

Threads may wait for messages

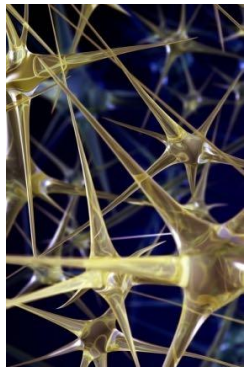
ASYNCHRONOUS TASK GRAPHS

Execution Optimization When Workflow is Known Up-Front

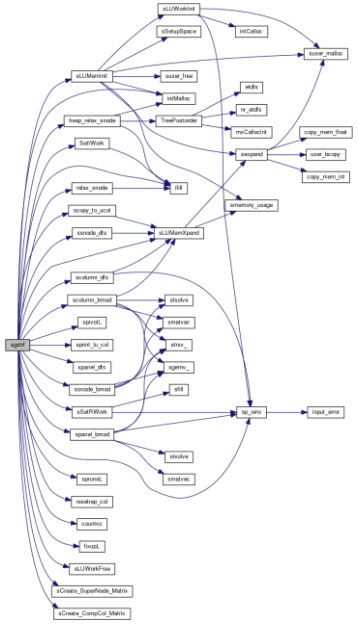
```
// Basic function to test primality.
bool IsPrime( size_t n)
{
	if (n == 2) return true;
	if ((n == 1) || ((n % 2) == 0)) return false;
	size_t iters = (unsigned int)sqrt( (double)n);
	for (size_t i = 3; i <= iters; i+=2) if (n % i == 0) return false;
	return true;
}

// Compute primes from 1 to 100,000,000.
size_t ComputePrimes()
{
	size_t Primes = 0;
	for ( size_t Start = 1; Start <= 100000000; ++Start)
	{
		if ( IsPrime( Start) )
		{
			++Primes;
		}
	}
	return Primes;
}
```

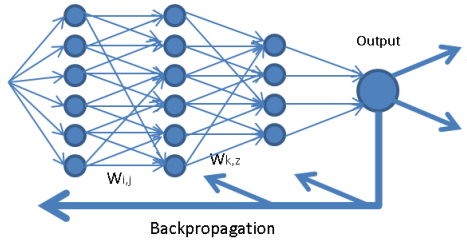
Loop & Function offload



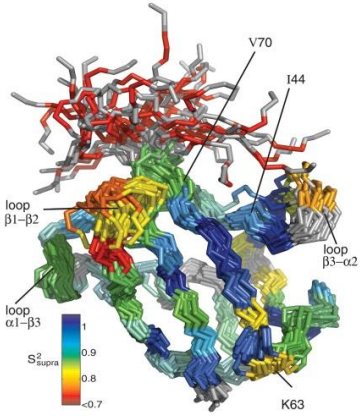
DL Inference



Linear Algebra



Deep Neural Network Training



HPC Simulation

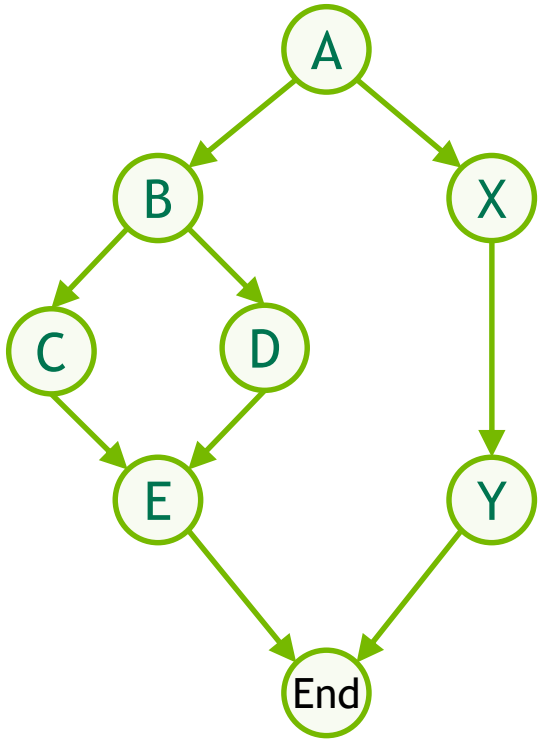
DEFINITION OF A CUDA GRAPH

Graph Nodes Are Not Just Kernel Launches

Sequence of operations, connected by dependencies.

Operations are one of:

Kernel Launch	CUDA kernel running on GPU
CPU Function Call	Callback function on CPU
Memcopy/Memset	GPU data management
Sub-Graph	Graphs are hierarchical



WHAT IS OPENACC

Open Specification Developed by OpenACC.org Consortium

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

Add Simple Compiler Directive

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

SIMPLE

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

POWERFUL & PORTABLE

Read more at www.openacc.org/about

WHO OPENACC IS FOR

The Main Focus

Domain Scientists

1. Want to do **more science & less programming**
2. Believe that GPUs are hard
3. Need help in learning how to easy start with GPUs
4. Mostly don't have a computer science degree

Application Developers

Looking for:

1. easy code maintenance,
2. better efficiency,
3. portability

Mostly computer scientists

OPENACC GROWING MOMENTUM

Wide Adoption Across Key HPC Codes

Over 100 Apps* Using OpenACC

ANSYS Fluent	GTC
Gaussian	XGC
VASP	ACME
LSDalton	FLASH
MPAS	COSMO
GAMERA	Numeca

VASP

Top Quantum Chemistry and Material Science Code

> "For VASP, OpenACC is ***the*** way forward for GPU acceleration. Performance is similar to CUDA, and OpenACC dramatically decreases GPU development and maintenance efforts. We're excited to collaborate with NVIDIA and PGI as an early adopter of Unified Memory."

Prof. Georg Kresse
Computational Materials Physics
University of Vienna

* Applications in production and development

SINGLE CODE FOR MULTIPLE PLATFORMS

OpenACC - Performance Portable Programming Model for HPC

OpenPOWER

Sunway

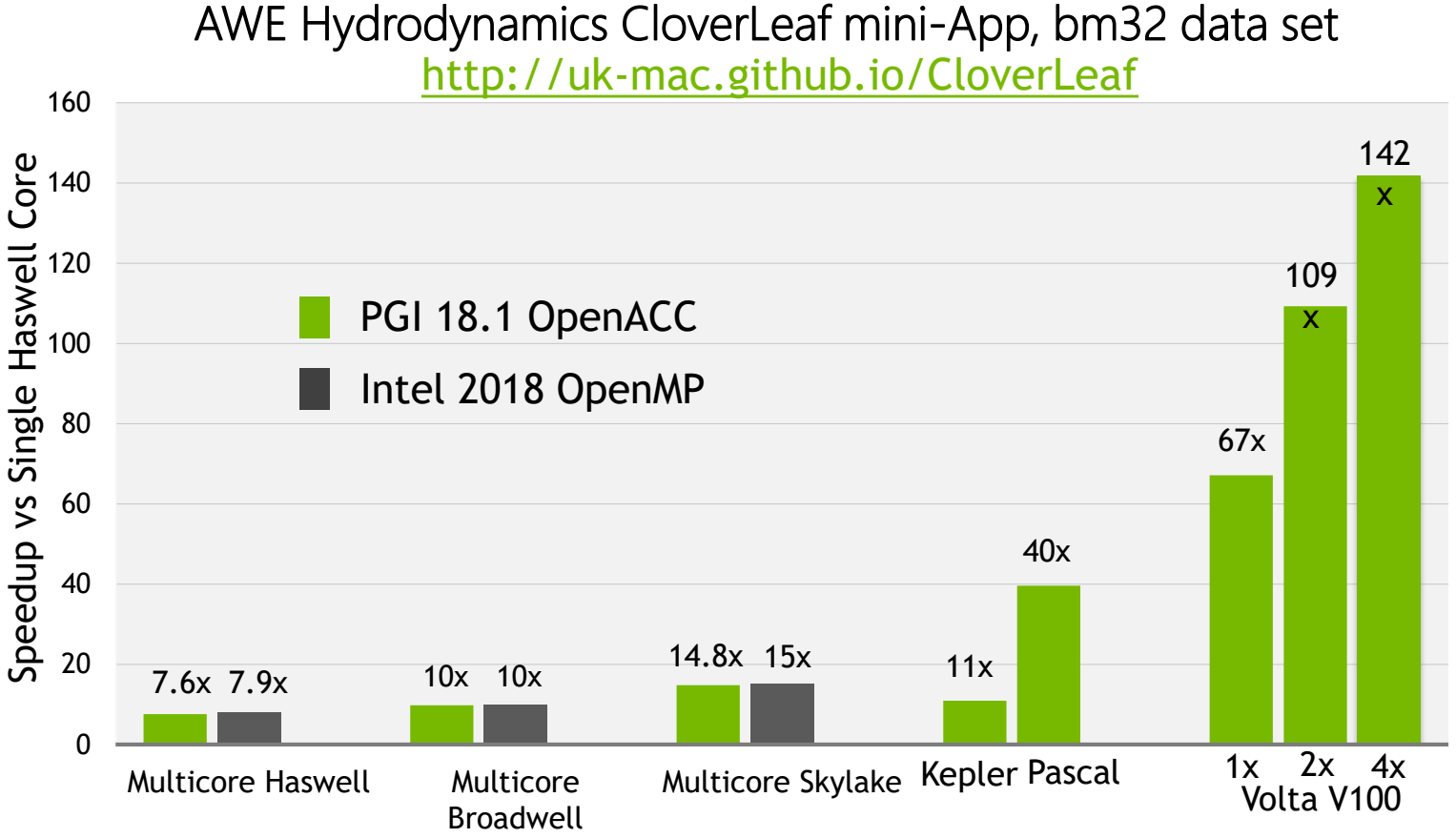
x86 CPU

x86 Xeon Phi

NVIDIA GPU

AMD GPU

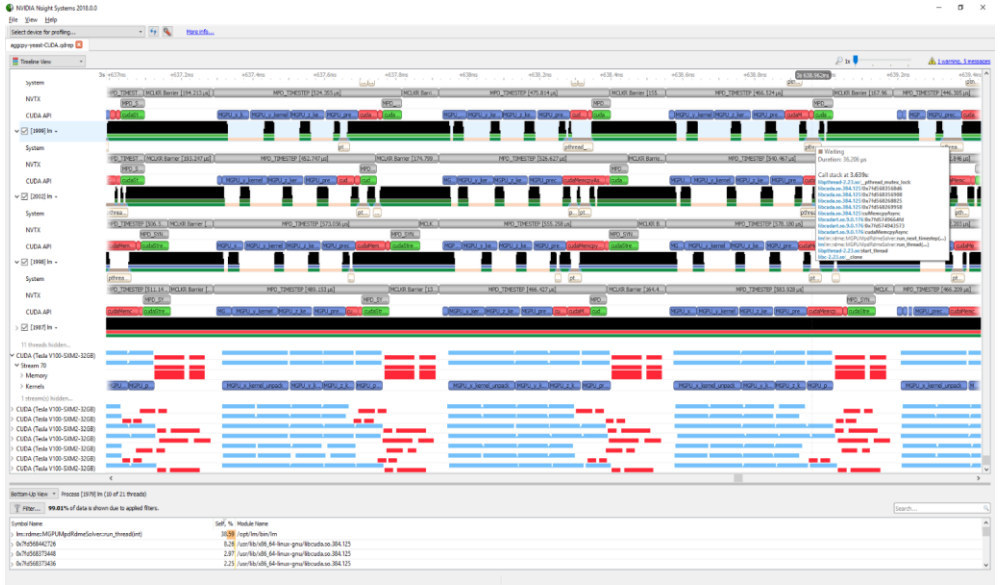
PEZY-SC



Systems: Haswell: 2x16 core Haswell server, four K80s, CentOS 7.2 (perf-hsw10), Broadwell: 2x20 core Broadwell server, eight P100s (dgx1-prd-01), Broadwell server, eight V100s (dgx07), Skylake 2x20 core Xeon Gold server (sky-4).
Compilers: Intel 2018.0.128, PGI 18.1
Benchmark: CloverLeaf v1.3 downloaded from http://uk-mac.github.io/CloverLeaf the week of November 7 2016; CloverlLeaf_Serial; CloverLeaf_ref (MPI+OpenMP); CloverLeaf_OpenACC (MPI+OpenACC)
Data compiled by PGI February 2018.

NSIGHT SYSTEMS

System-wide Performance Analysis

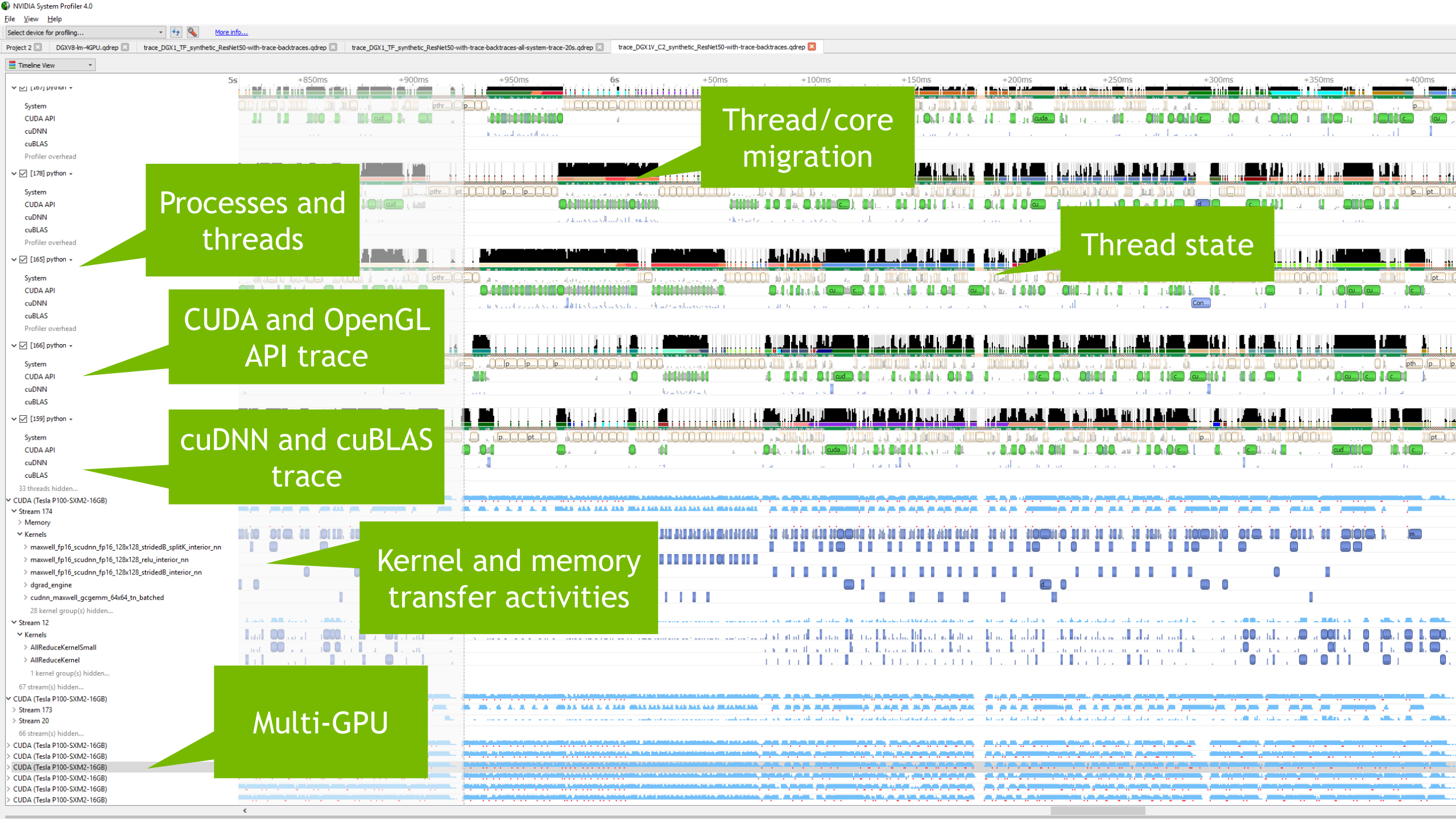


Observe Application Behavior: CPU threads, GPU traces, Memory Bandwidth and more

Locate Optimization Opportunities: CUDA & OpenGL APIs, Unified Memory transfers, User Annotations using NVTX

Ready for Big Data: Fast GUI capable of visualizing in excess of 10 million events on laptops, Container support, Minimum user privileges

https://developer.nvidia.com/nsight-systems


NVIDIA System Profiler 4.0
File View Help
Select device for profiling...
More info...
Project 2
DGX8-lm-4GPU.qdrep
trace_DGX1_TF_synthetic_ResNet50-with-trace-backtraces.qdrep
trace_DGX1_TF_synthetic_ResNet50-with-trace-backtraces-all-system-trace-20s.qdrep
trace_DGX1V_C2_synthetic_ResNet50-with-trace-backtraces.qdrep
Timeline View
5s
+850ms
+900ms
+950ms
6s
+50ms
+100ms
+150ms
+200ms
+250ms
+300ms
+350ms
+400ms
System
CUDA API
cuDNN
cuBLAS
Profiler overhead
[178] python
[165] python
[166] python
[159] python
33 threads hidden...
CUDA (Tesla P100-SXM2-16GB)
Stream 174
Memory
Kernels
maxwell_fp16_scudnn_fp16_128x128_stridedB_splitK_interior_nn
maxwell_fp16_scudnn_fp16_128x128_relu_interior_nn
maxwell_fp16_scudnn_fp16_128x128_stridedB_interior_nn
dgrad_engine
cudnn_maxwell_gcgemm_64x64_tn_batched
28 kernel group(s) hidden...
Stream 12
AllReduceKernelSmall
AllReduceKernel
1 kernel group(s) hidden...
67 stream(s) hidden...
Stream 173
Stream 20
66 stream(s) hidden...
Processes and threads
Thread/core migration
Thread state
CUDA and OpenGL API trace
cuDNN and cuBLAS trace
Kernel and memory transfer activities
Multi-GPU

CONTAINERS: SIMPLIFYING WORKFLOWS

WHY CONTAINERS

Simplifies Deployments

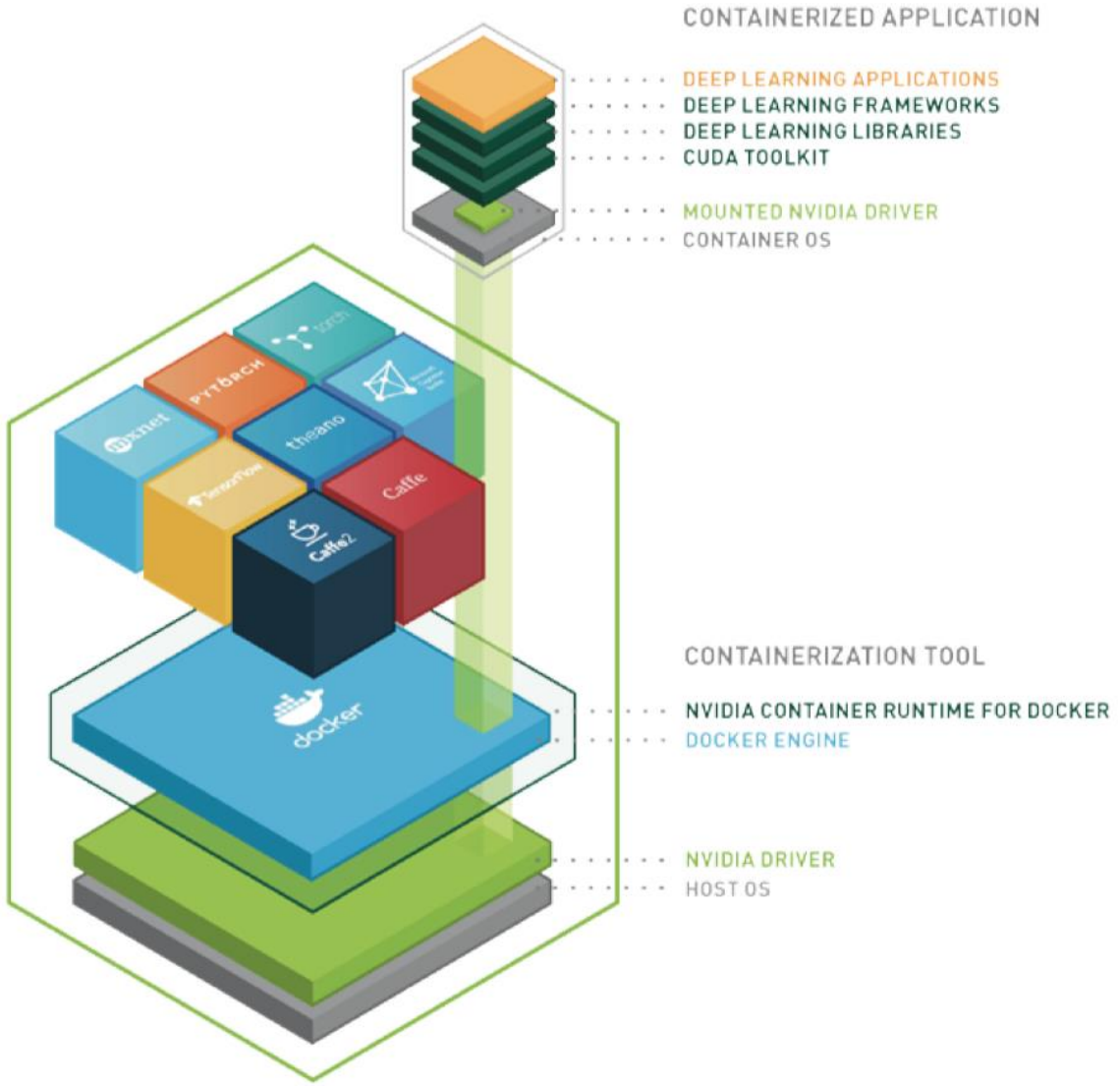
- Eliminates complex, time-consuming builds and installs

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

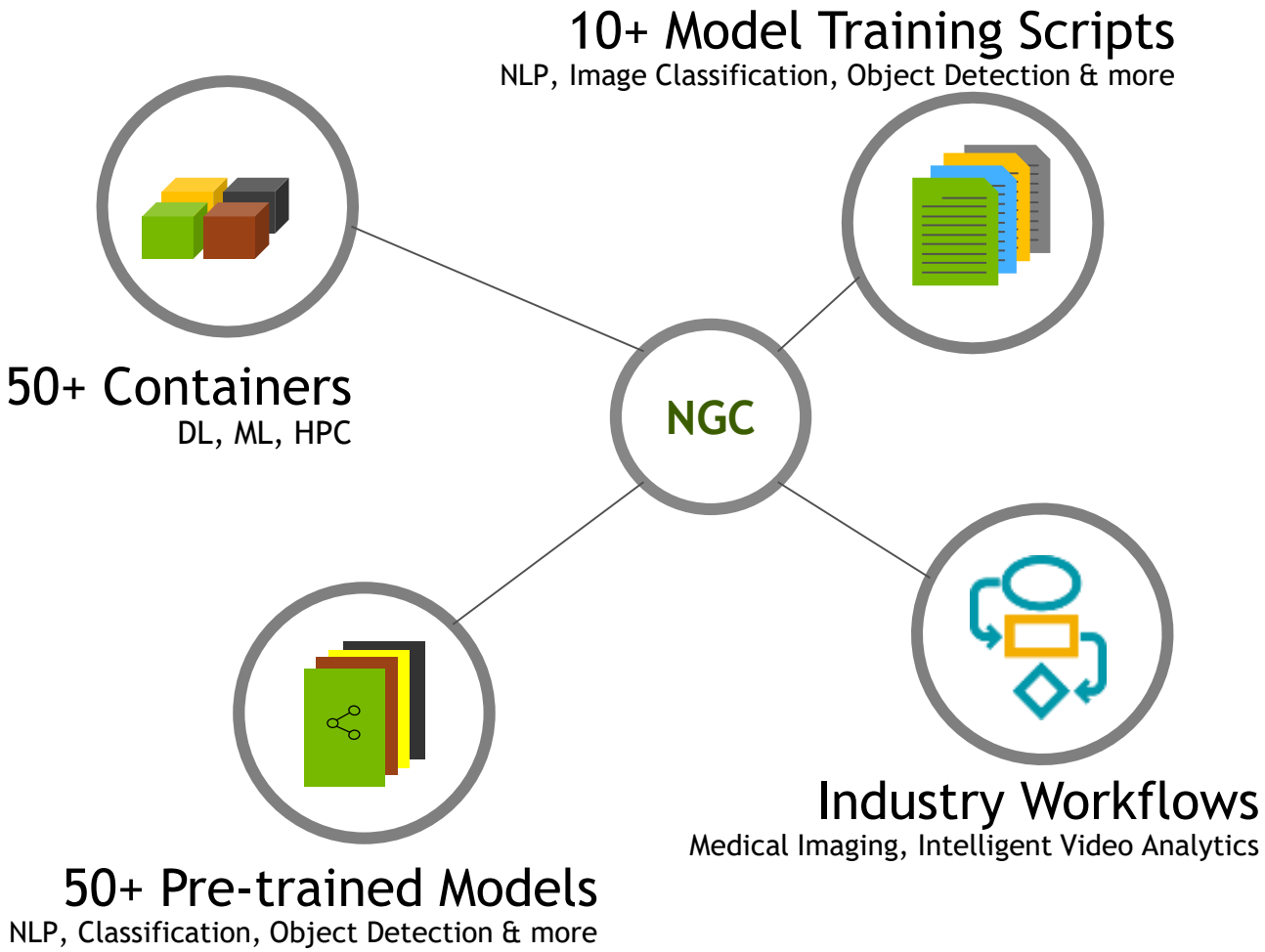
- Supports Docker & Singularity runtimes

Run Anywhere

- Pascal/Volta/Turing-powered NVIDIA DGX, PCs, workstations, servers and top cloud platforms

THE NEW NGC

GPU-optimized Software Hub. Simplifying DL, ML and HPC Workflows



Simplify Deployments

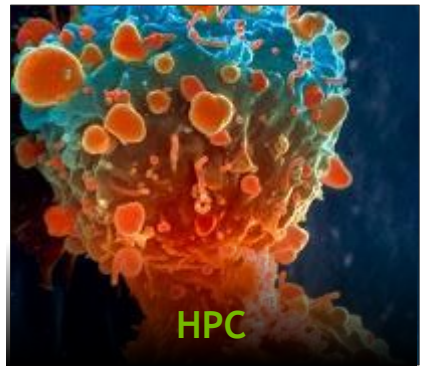
Innovate Faster

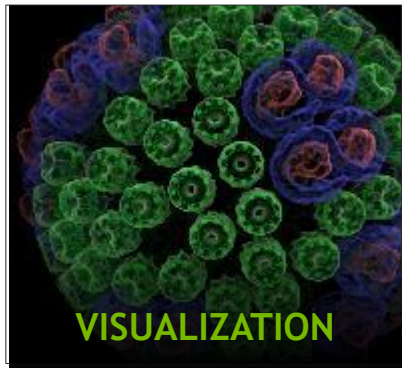
Deploy Anywhere

ngc.nvidia.com

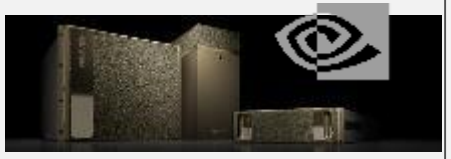
NGC-READY ECOSYSTEM

Now Over 50 GPU-Optimized Containers





docker



Atos · cisco · CRAY · DELL EMC · DELL · aws · Azure · Hewlett Packard Enterprise · hp · lenovo · SUPERMICRO · Google Cloud · ORACLE CLOUD PLATFORM

THE UNIVERSITY OF ARIZONA

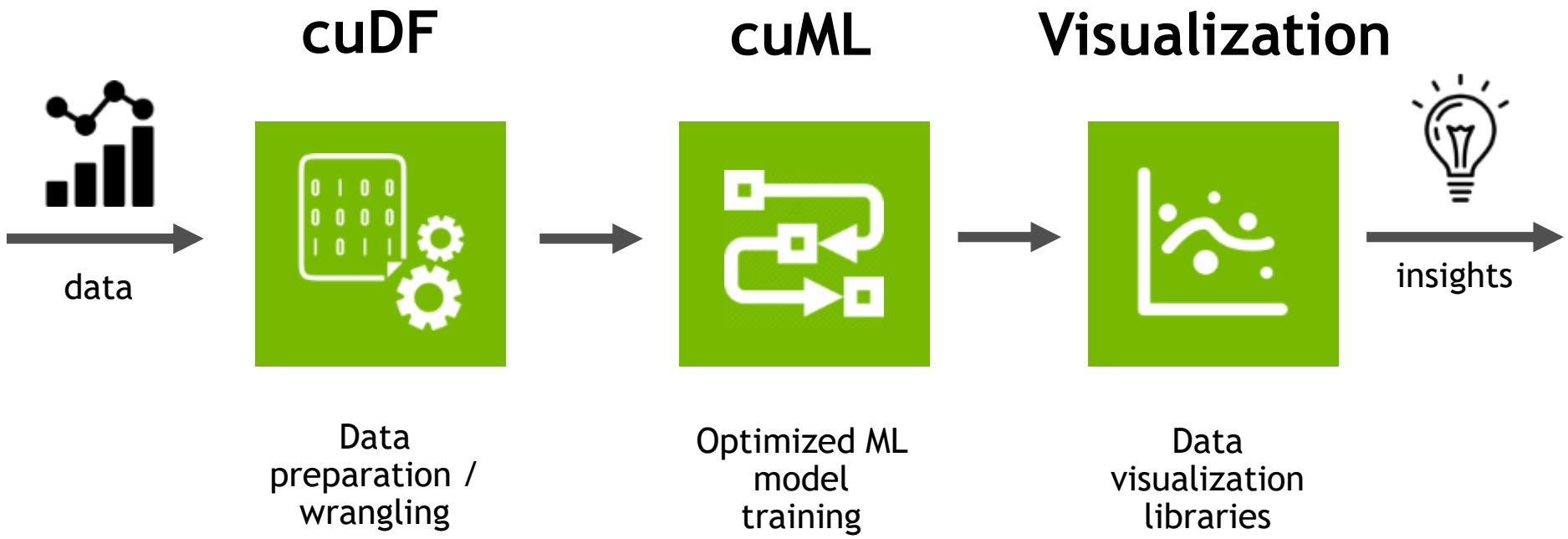
CLEMSON UNIVERSITY

PSC PITTSBURGH SUPERCOMPUTING CENTER

Stanford University

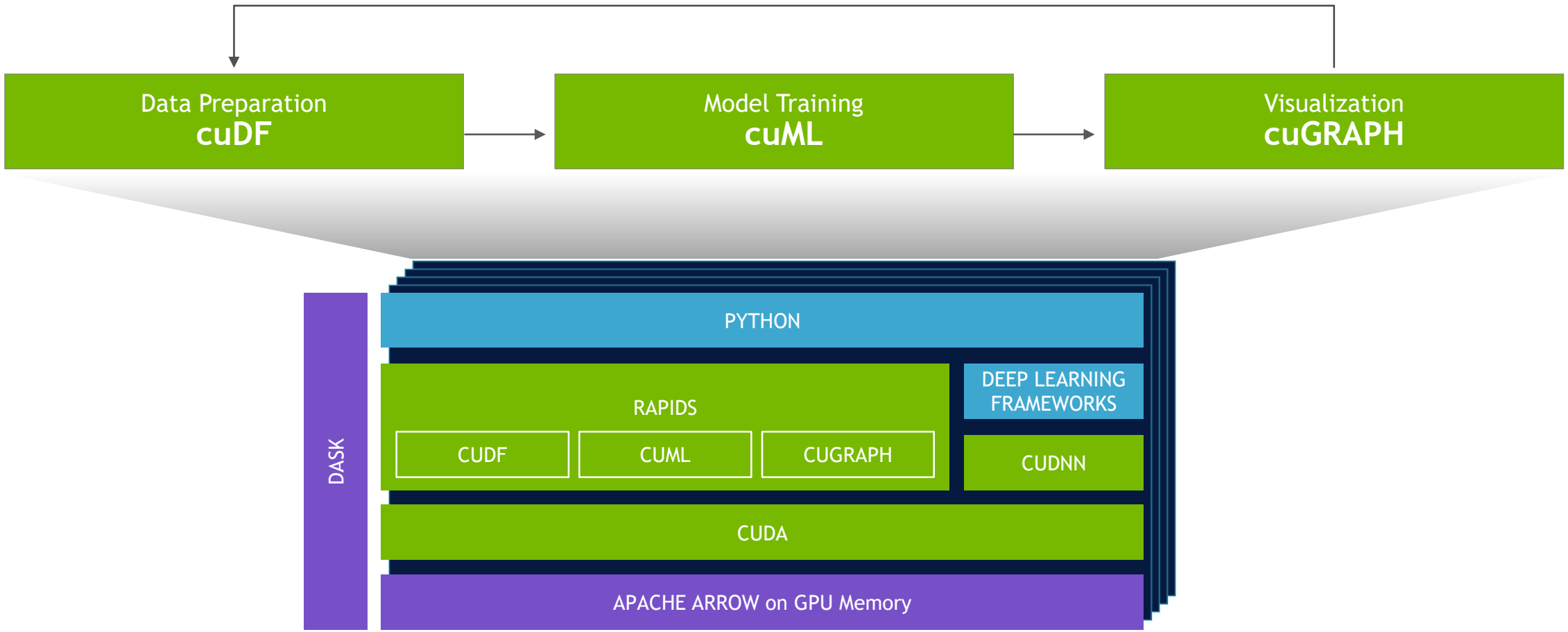
RE-IMAGINING DATA SCIENCE WORKFLOW

Open Source, End-to-end GPU-accelerated Workflow Built On CUDA



RAPIDS — OPEN GPU DATA SCIENCE

Software Stack Python



ACCELERATING MACHINE LEARNING

The RAPIDS Ecosystem

Open Source Community

ANACONDA. · Apache Arrow · GUNROCK · PyData · QUANTSIGHT · scikit-learn · URSA LABS

Enterprise Data Science Platforms

IBM · ORACLE · SAS · SAP

Startups

BLAZINGDB · FASTDATA · graphistry · $H_2O.ai$ · kinetica · MAPR · omnisci

Deep Learning Integration

Chainer · PYTORCH

RAPIDS

GPU Servers

CISCO · DELL EMC · Hewlett Packard Enterprise · IBM · Lenovo

Storage Partners

DELL EMC · IBM · NetApp · PURESTORAGE

SUMMARY

GPUs are established in HPC and Datacenter

Full stack optimization, not just selling silicon

Improvements and simplification on multiple fronts

- HW: chip, node and system level
- SW: low- and high-level languages, libraries, frameworks, apps

Convergence of HPC and accelerated machine learning in the data center


BACKUP
NVIDIA