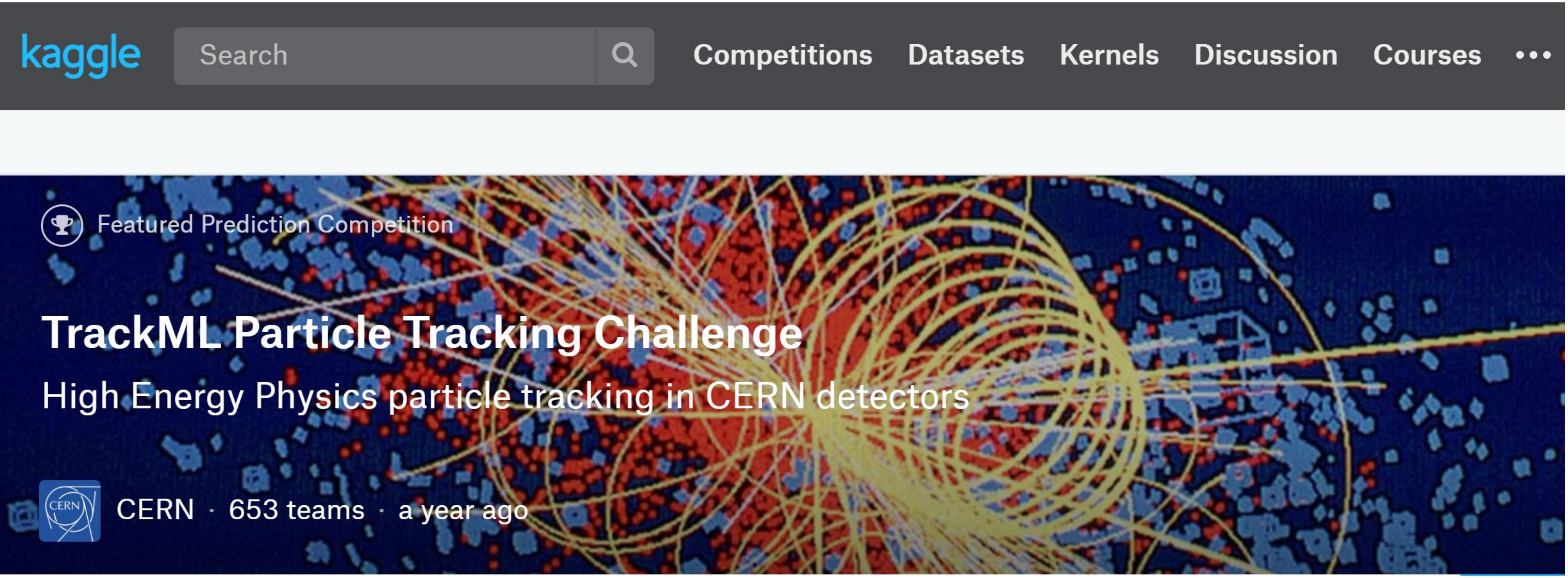
TRACKING: A PERFECT USE CASE FOR GPUS?

Peter Messmer, 7/1/2019

TrackML Grand Finale, CERN

THANK YOU FOR THE PARTICIPATION!



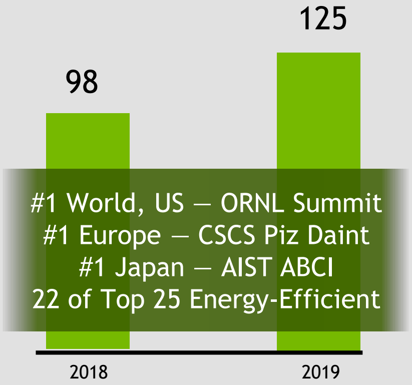
Featured Prediction Competition

TrackML Particle Tracking Challenge

High Energy Physics particle tracking in CERN detectors

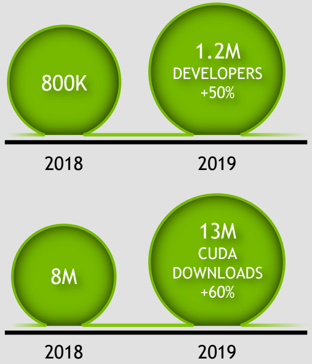
CERN · 653 teams · a year ago

A YEAR OF RAPID GROWTH

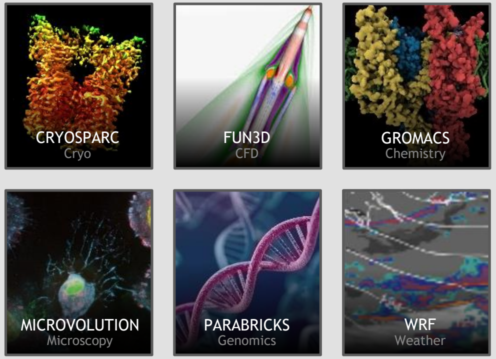


#1 World, US — ORNL Summit
#1 Europe — CSCS Piz Daint
#1 Japan — AIST ABCI
22 of Top 25 Energy-Efficient

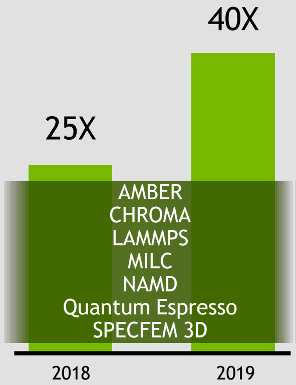
25% MORE TOP500 SUPERCOMPUTERS



50% GROWTH OF NVIDIA DEVELOPERS



600+ CUDA APPS



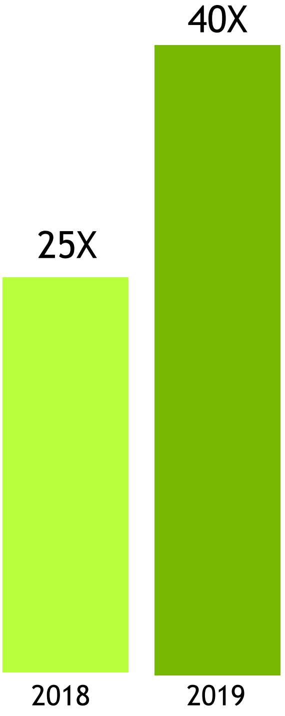
AMBER
CHROMA
LAMMPS
MILC
NAMD
Quantum Espresso
SPECFEM 3D

MORE PERF SAME GPU

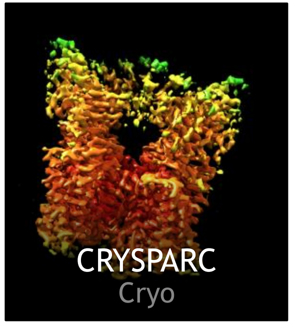
EXPANDING VALUE FOR HPC CUSTOMERS

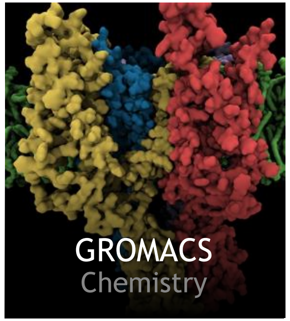
Partnering With HPC Development Community

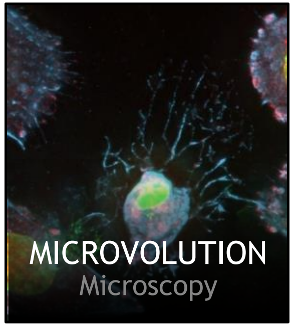
AMBER
CHROMA
GTC
LAMMPS
MILC
NAMD
QUANTUM ESP
SPECFEM3D

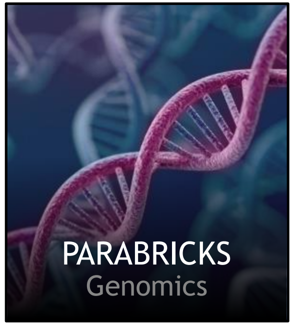


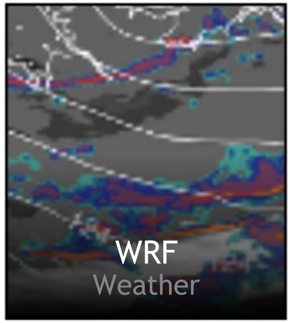
MORE PERFORMANCE WITH SAME GPU



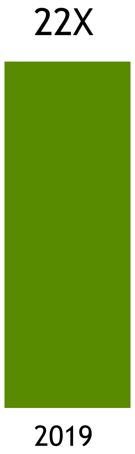








CRYOSPARC	24x
FUN3D	24x
GROMACS	7x
MICROVOLUTION	48x
PARABRICKS	22x
WRF	8x



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

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

INTERSECTION OF HPC & AI TRANSFORMING SCIENCE

HPC

> Algorithms based on first principles theory

> Proven models for accurate results

AI

> Neural networks that learn patterns from large data sets

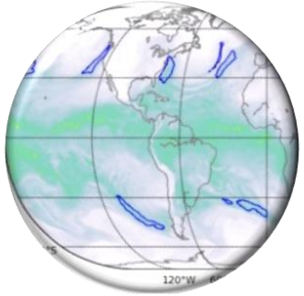
> Improve predictive accuracy and faster response time

SPEEDING PATH TO FUSION ENERGY

PRINCETON UNIVERSITY

90% Prediction Accuracy Publish in Nature April 2019

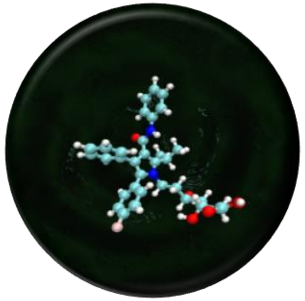
EXASCALE WEATHER MODELING



BERKELEY LAB

Tensor Cores Achieved 1.13 EF 2018 Gordon Bell Winner

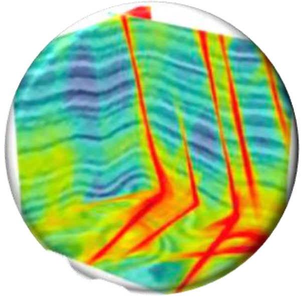
INDENTIFYING CHEMICAL COMPOUNDS



DOW

Orders Of Magnitude Speedup 3M New Compounds In 1 Day

O&G FAULT INTERPRETATION

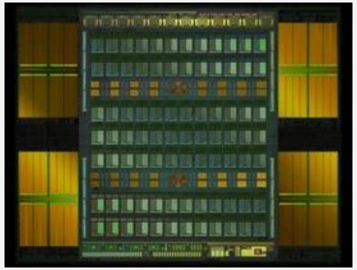


BUREAU OF ECONOMIC GEOLOGY

Time-to-solution Reduced From Weeks To 2 Hours

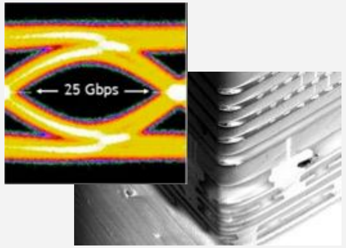
INTRODUCING TESLA V100

Volta Architecture



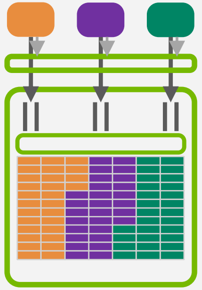
Most Productive GPU

Improved NVLink & HBM2



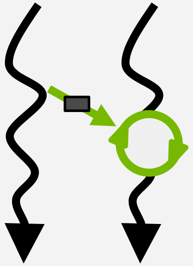
Efficient Bandwidth

Volta MPS



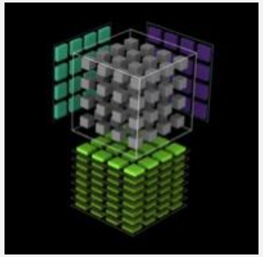
Inference Utilization

Improved SIMT Model



New Algorithms

Tensor Core



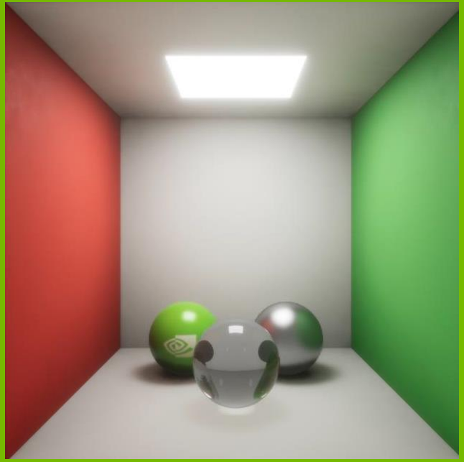
120 Programmable
TFLOPS Deep Learning

The Fastest and Most Productive GPU for Deep Learning and HPC

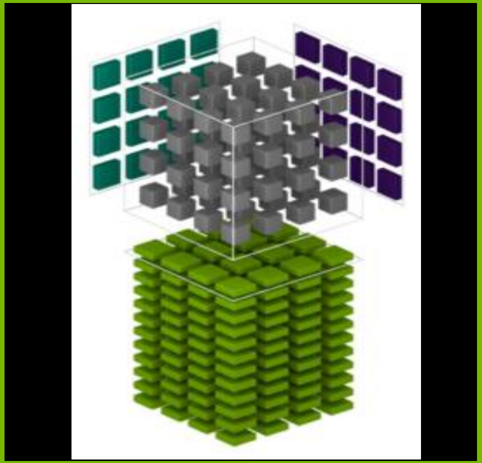
NVIDIA TURING: GRAPHICS REINVENTED

Built to Revolutionize the Work of Creative Professionals

RT Cores

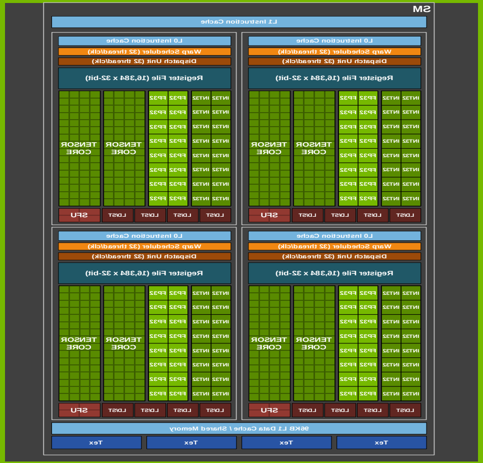


Tensor Cores



CUDA Cores

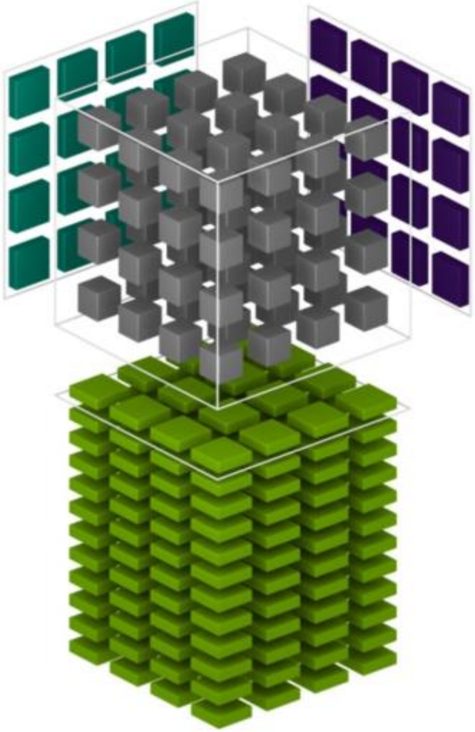
Streaming Multiprocessors


SM
L1 Instruction Cache
L0 Instruction Cache
Warp Scheduler (32 thread/clk)
Dispatch Unit (32 thread/clk)
Register File (16,384 x 32-bit)
TENSOR CORE
FP32
INT32
SFU
LD/ST
96KB L1 Data Cache / Shared Memory
Tex

TENSOR CORES IN TURING

New in Volta, Upgraded in Turing

GPU	SMs	Total	Peak Half FLOPS	PEAK INT8 OPS	PEAK INT4 OPS	PEAK B1 OPS
V100	80	640		N.A.	N.A.	N.A.
Quadro RTX 6000/8000	72	576	130.5 TFLOPS*	260 TOPS*	521 TOPS*	2087 TOPS*



Matrix Multiplication Pipeline,
half precision inputs → half / float accumulator
8bit/4bit INT inputs → 32bit INT accumulator
1bit Binary inputs → 32 bit INT accumulator (XOR + POPC)

Used in CUBLAS, CUDNN, CUTLASS
Exposed in CUDA 10 (4bit INT and 1bit binary are experimental)

* Using 1.77GHz Boost Clock

TURING RT CORES

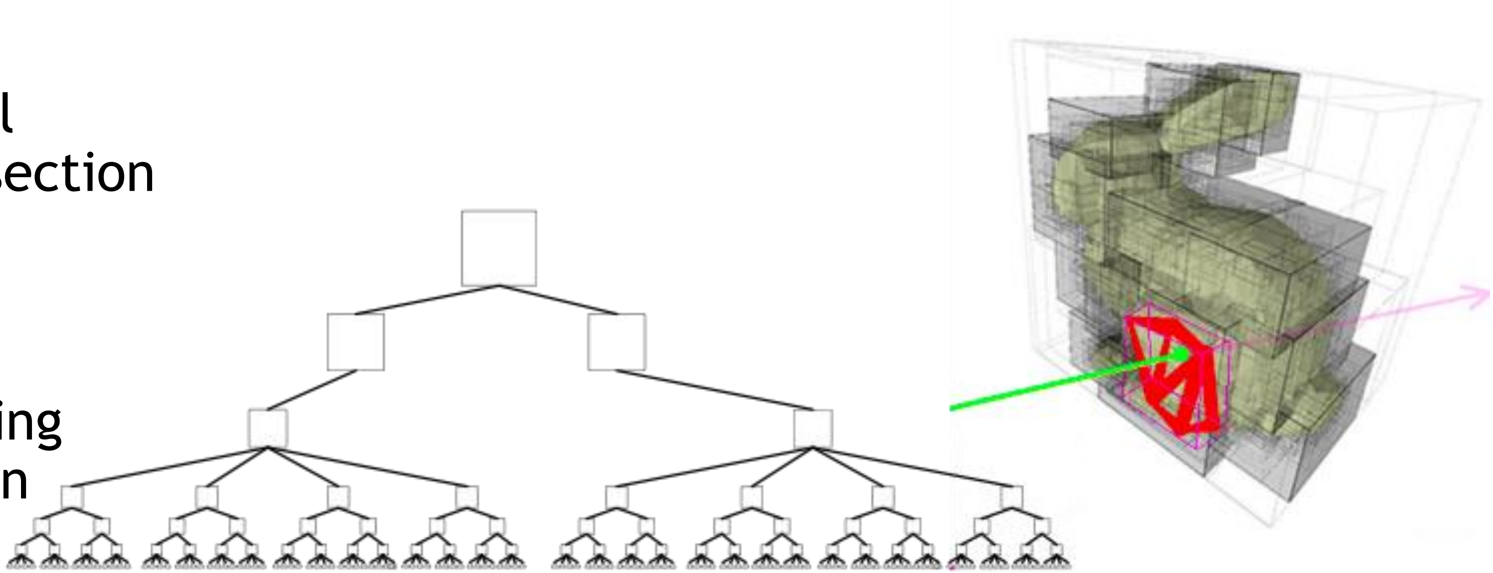
Hardware Accelerated Ray Tracing

RT Cores perform

- Ray-BVH Traversal
- Instancing: 1 Level
- Ray-Triangle Intersection

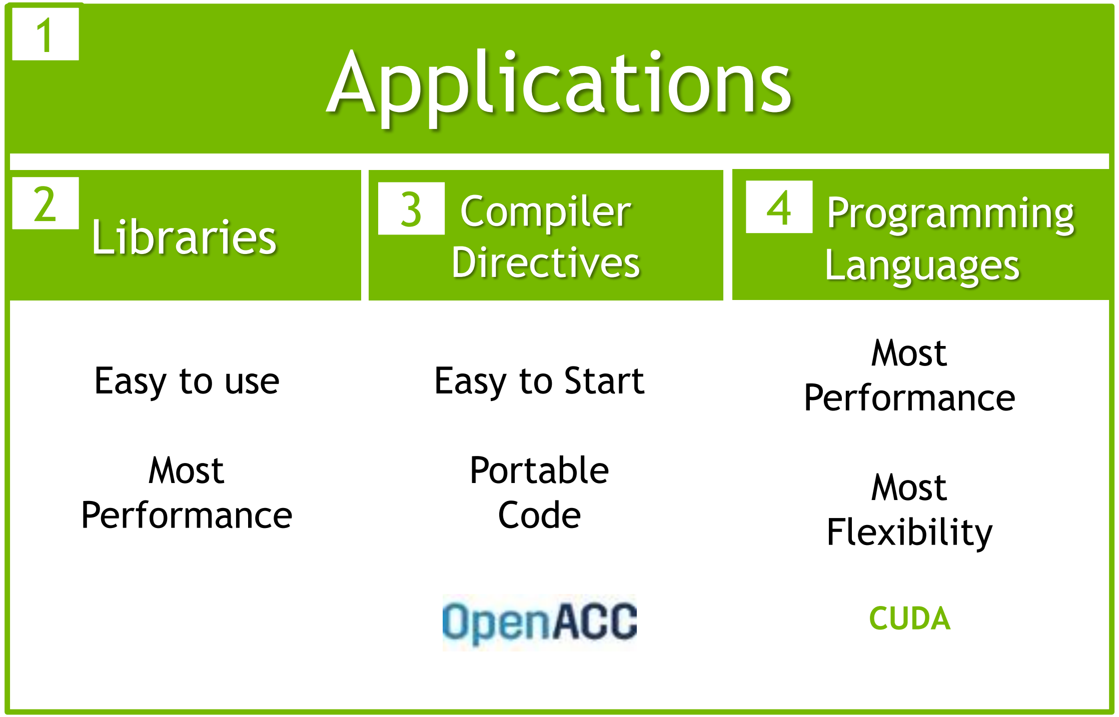
Return to SM for

- Multi-level Instancing
- Custom Intersection
- Shading



Programming via OptiX RT framework
Low overhead interop with CUDA

HOW TO START WITH GPUS



1 Applications		
2 Libraries	3 Compiler Directives	4 Programming Languages
Easy to use	Easy to Start	Most Performance
Most Performance	Portable Code	Most Flexibility
	OpenACC	CUDA

1. Review available GPU-accelerated applications

2. Check for GPU-Accelerated applications and libraries

3. Add OpenACC Directives for quick acceleration results and portability

4. Dive into CUDA for highest performance and flexibility

SINGLE CODE FOR MULTIPLE PLATFORMS

OpenACC - Performance Portable Programming Model for HPC

OpenPOWER

Sunway

x86 CPU

x86 Xeon Phi

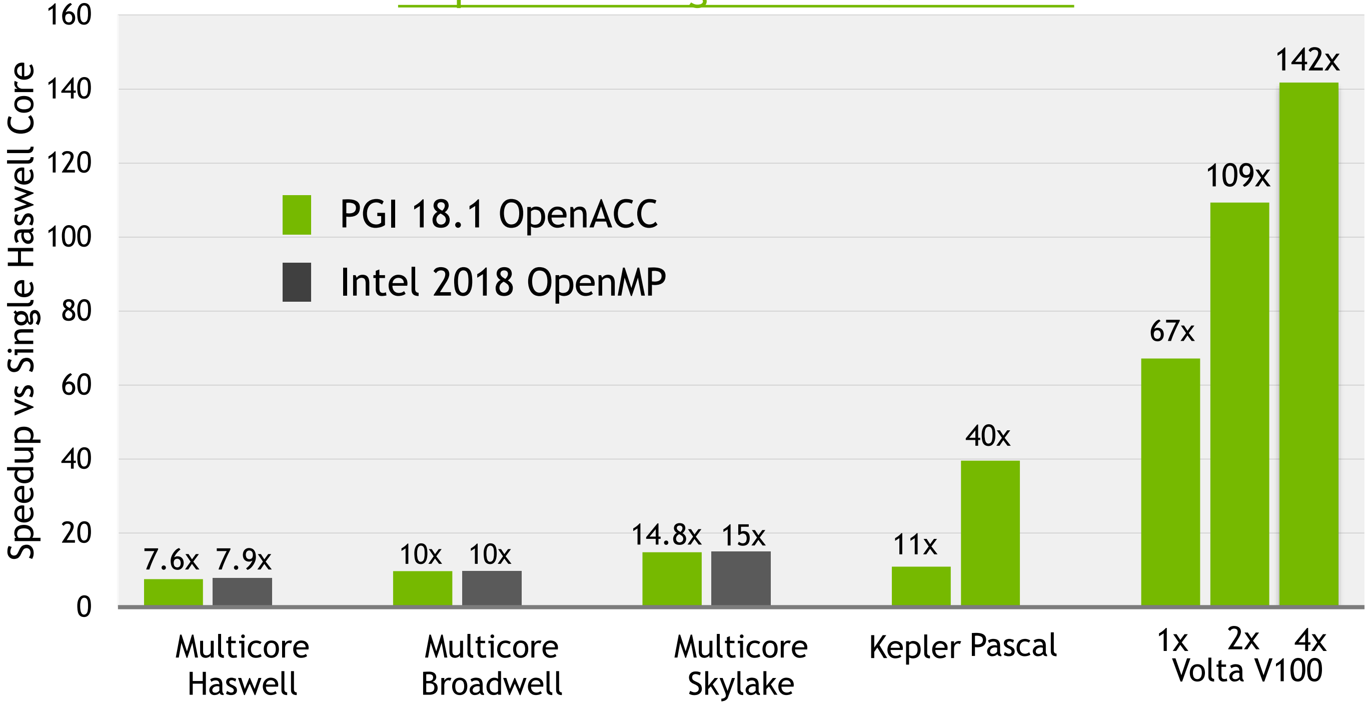
NVIDIA GPU

AMD GPU

PEZY-SC

AWE Hydrodynamics CloverLeaf mini-App, bm32 data set

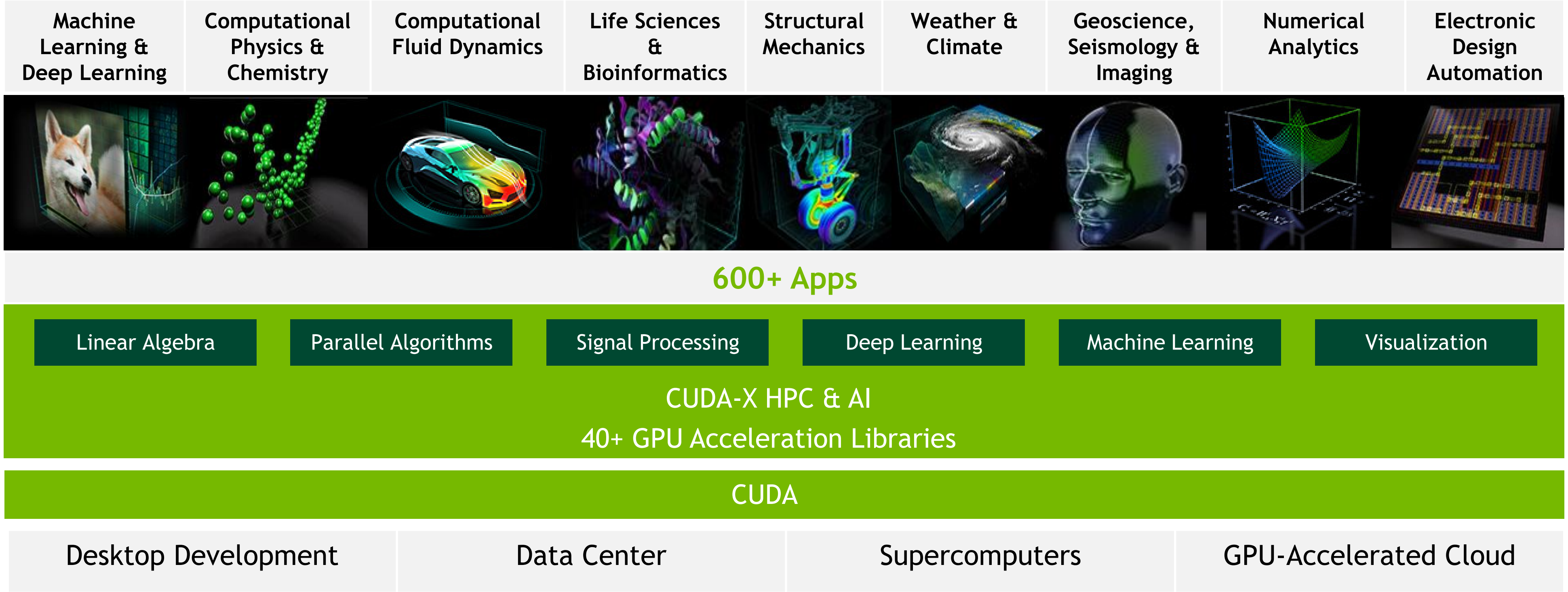
http://uk-mac.github.io/CloverLeaf



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.

NVIDIA CUDA-X UPDATES

Software To Deliver Acceleration For HPC & AI Apps; 500+ New Updates



Machine Learning & Deep Learning

Computational Physics & Chemistry

Computational Fluid Dynamics

Life Sciences & Bioinformatics

Structural Mechanics

Weather & Climate

Geoscience, Seismology & Imaging

Numerical Analytics

Electronic Design Automation

600+ Apps

Linear Algebra

Parallel Algorithms

Signal Processing

Deep Learning

Machine Learning

Visualization

CUDA-X HPC & AI

40+ GPU Acceleration Libraries

CUDA

Desktop Development

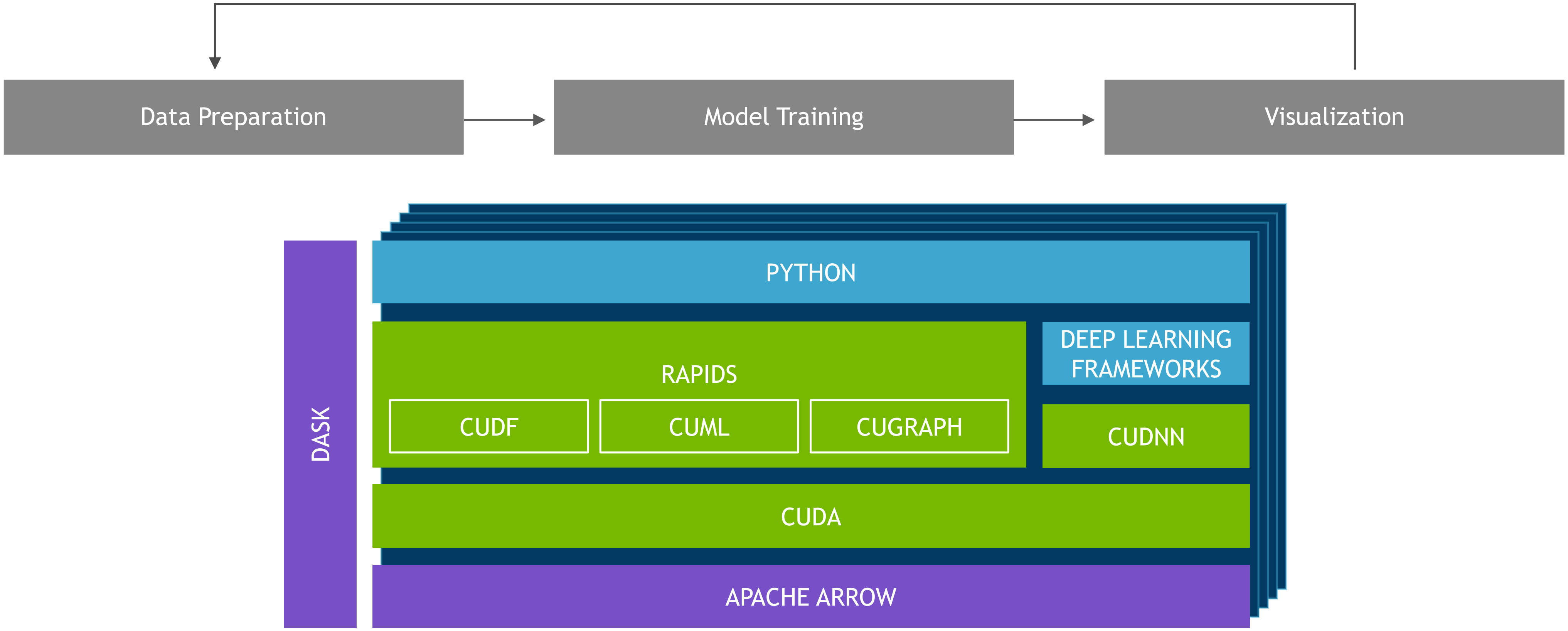
Data Center

Supercomputers

GPU-Accelerated Cloud

RAPIDS — OPEN GPU DATA SCIENCE

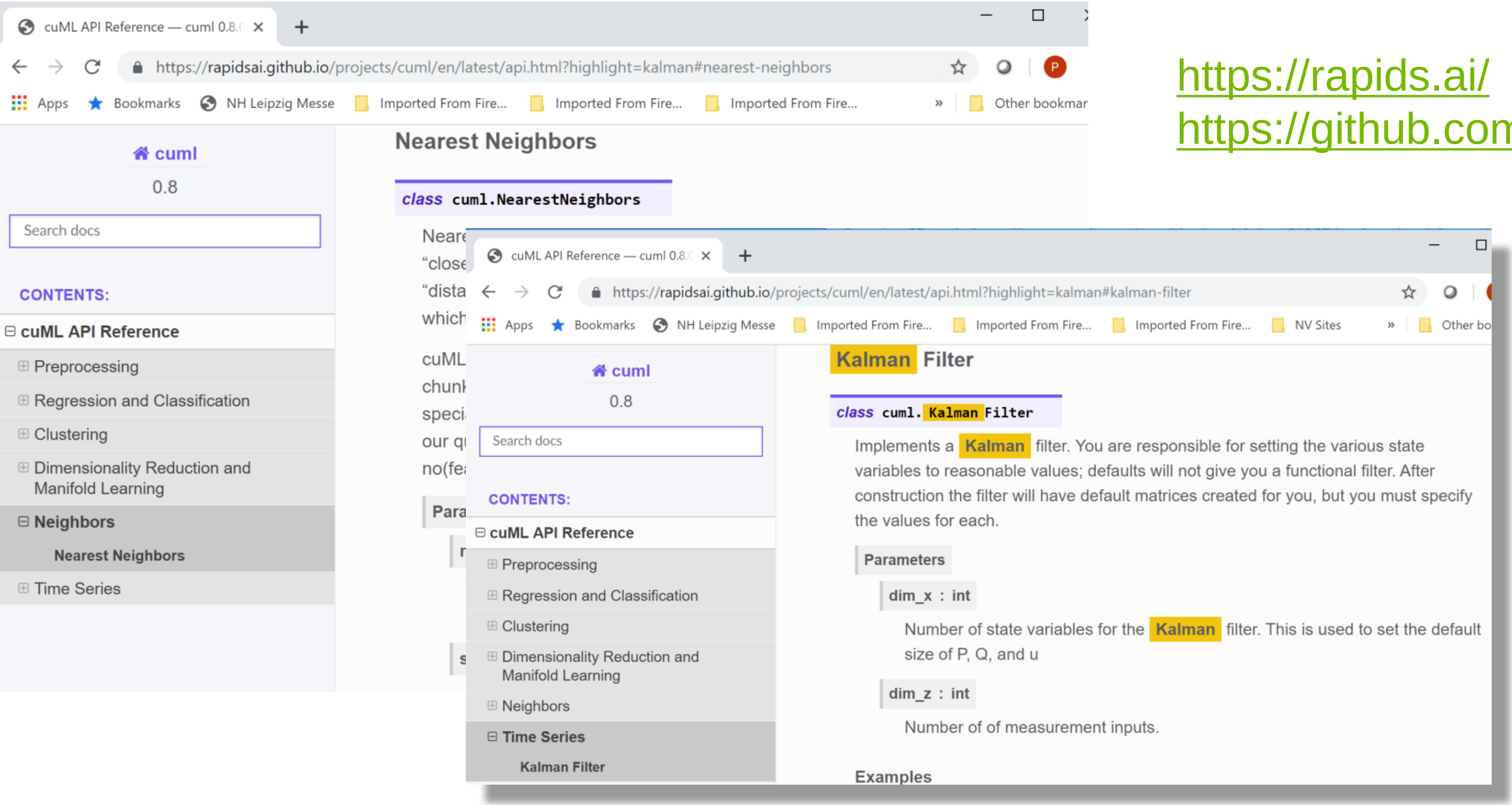
Software Stack



CUML — OPEN GPU DATA SCIENCE

Broad range of GPU accelerated algorithms

https://rapids.ai/
https://github.com/rapidsai/cuml


cuML API Reference — cuml 0.8.(
https://rapidsai.github.io/projects/cuml/en/latest/api.html?highlight=kalman#nearest-neighbors
Apps
Bookmarks
NH Leipzig Messe
Imported From Fire...
Imported From Fire...
Imported From Fire...
Other bookmar
cuML API Reference — cuml 0.8.(
https://rapidsai.github.io/projects/cuml/en/latest/api.html?highlight=kalman#kalman-filter
Apps
Bookmarks
NH Leipzig Messe
Imported From Fire...
Imported From Fire...
Imported From Fire...
NV Sites
Other bo

cuml
0.8
Search docs

CONTENTS:

- cuML API Reference
  - Preprocessing
  - Regression and Classification
  - Clustering
  - Dimensionality Reduction and Manifold Learning
  - Neighbors
    - Nearest Neighbors
  - Time Series

Nearest Neighbors

class `cuml.NearestNeighbors`

Neare
“close
“dista
which

cuML
chunk
speci
our qu
no(fea

Para

cuml
0.8
Search docs

CONTENTS:

- cuML API Reference
  - Preprocessing
  - Regression and Classification
  - Clustering
  - Dimensionality Reduction and Manifold Learning
  - Neighbors
  - Time Series
    - Kalman Filter

Kalman Filter

class `cuml.KalmanFilter`

Implements a Kalman filter. You are responsible for setting the various state variables to reasonable values; defaults will not give you a functional filter. After construction the filter will have default matrices created for you, but you must specify the values for each.

Parameters

dim_x : int

Number of state variables for the Kalman filter. This is used to set the default size of P, Q, and u

dim_z : int

Number of of measurement inputs.

Examples

GPU-ACCELERATED XGBOOST

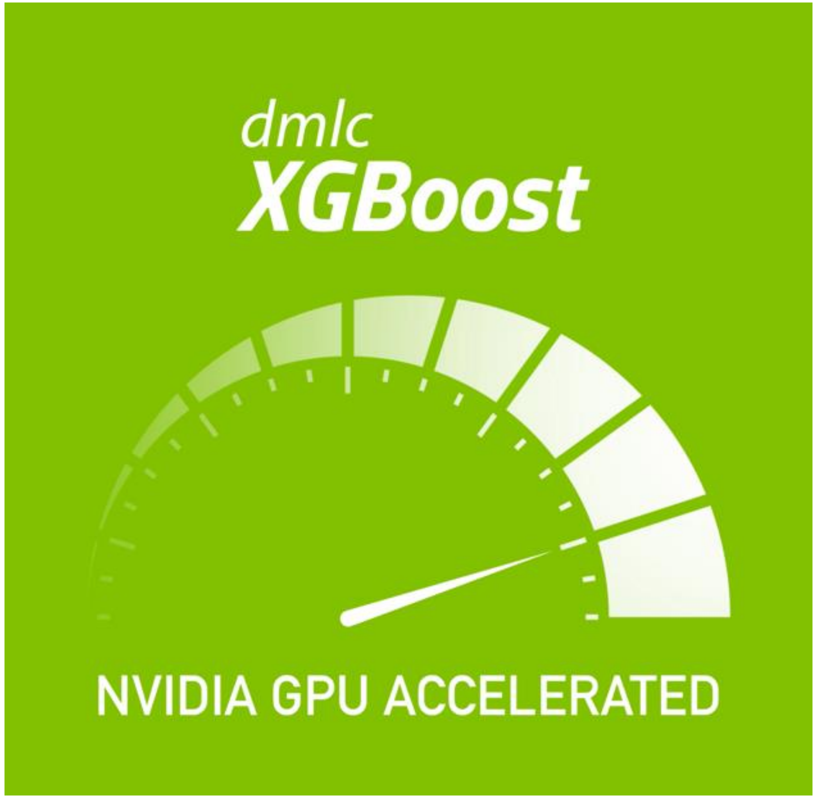
Unleashing the Power of NVIDIA GPUs for Users of XGBoost

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

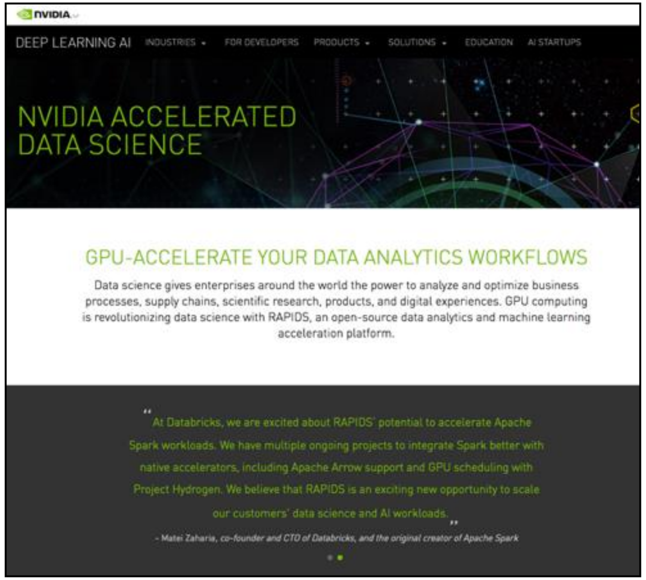
Better Predictions, Sooner
Work with larger datasets and perform more model iterations without spending valuable time waiting.

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.



FOR MORE INFORMATION



NVIDIA ACCELERATED DATA SCIENCE

GPU-ACCELERATE YOUR DATA ANALYTICS WORKFLOWS

Data science gives enterprises around the world the power to analyze and optimize business processes, supply chains, scientific research, products, and digital experiences. GPU computing is revolutionizing data science with RAPIDS, an open-source data analytics and machine learning acceleration platform.

"At Databricks, we are excited about RAPIDS' potential to accelerate Apache Spark workloads. We have multiple ongoing projects to integrate Spark better with native accelerators, including Apache Arrow support and GPU scheduling with Project Hydrogen. We believe that RAPIDS is an exciting new opportunity to scale our customers' data science and AI workloads."

- Matei Zaharia, co-founder and CTO of Databricks, and the original creator of Apache Spark

nvidia.com/datascience

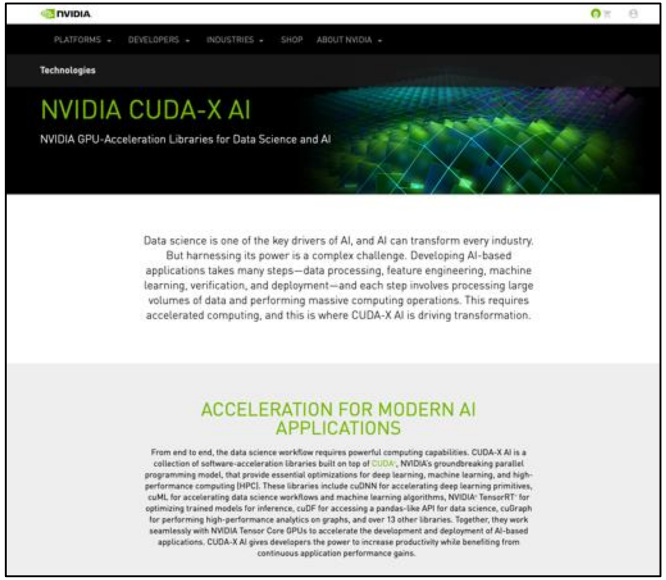
RAPIDS

Open GPU Data Science

About RAPIDS

The RAPIDS suite of open source software libraries gives you the freedom to execute end-to-end data science and analytics pipelines entirely on GPUs. It relies on NVIDIA® CUDA® primitives for low-level compute optimization, but exposes that GPU parallelism and high-bandwidth memory speed through user-friendly Python interfaces.

rapids.ai



NVIDIA CUDA-X AI

NVIDIA GPU-Acceleration Libraries for Data Science and AI

Data science is one of the key drivers of AI, and AI can transform every industry. But harnessing its power is a complex challenge. Developing AI-based applications takes many steps—data processing, feature engineering, machine learning, verification, and deployment—and each step involves processing large volumes of data and performing massive computing operations. This requires accelerated computing, and this is where CUDA-X AI is driving transformation.

ACCELERATION FOR MODERN AI APPLICATIONS

From end to end, the data science workflow requires powerful computing capabilities. CUDA-X AI is a collection of software-acceleration libraries built on top of CUDA®, NVIDIA's groundbreaking parallel programming model, that provide essential optimizations for deep learning, machine learning, and high-performance computing (HPC). These libraries include cuDNN for accelerating deep learning primitives, cuML for accelerating data science workflows and machine learning algorithms, NVIDIA® TensorRT™ for optimizing trained models for inference, cuDF for accessing a pandas-like API for data science, cuGraph for performing high-performance analytics on graphs, and over 13 other libraries. Together, they work seamlessly with NVIDIA Tensor Core GPUs to accelerate the development and deployment of AI-based applications. CUDA-X AI gives developers the power to increase productivity while benefiting from continuous application performance gains.

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

GPUS ARE SPOT ON FOR TRACKING WITH ML!

Convergence of scientific computing and ML/AI is happening now

GPUs: Accelerator for both HPC & AI/ML

Pick the right programming model: CUDA, OpenACC, libraries, frameworks

Frameworks increasingly Python driven

DNN frameworks fully supported on GPU

RAPIDS: broad framework for machine learning applications

What’s your story with GPUs?

NVIDIA®

TESLA PRODUCTS DECODER

	P100 (SXM2)	P100 (PCIE)	P40	P4	T4	V100 (PCIE)	V100 (SXM2)	V100 (FHHL)
GPU CHIP	GP100	GP100	GP102	GP104	TU104	GV100	GV100	GV100
PEAK FP64 (TFLOPs)	5.3	4.7	NA	NA	NA	7	7.8	6.5
PEAK FP32 (TFLOPs)	10.6	9.3	12	5.5	8.1	14	15.7	13
PEAK FP16 (TFLOPs)	21.2	18.7	NA	NA	65	112	125	105
PEAK TOPs	NA	NA	47	22	260	NA	NA	NA
Memory Size	16 GB HBM2	16/12 GB HBM2	24 GB GDDR5	8 GB GDDR5	16 GB GDDR6	32 GB HBM2	32 GB HBM2	16GB HBM2
Memory BW	732 GB/s	732/549 GB/s	346 GB/s	192 GB/s	320GB/s	900 GB/s	900 GB/s	900 GB/s
Interconnect	NVLINK + PCIe Gen3	PCIe Gen3	PCIe Gen3	PCIe Gen3	PCIe Gen3	PCIe Gen3	NVLINK + PCIe Gen3	PCIe Gen3
ECC	Internal + HBM2	Internal + HBM2	GDDR5	GDDR5	GDDR6	Internal + HBM2	Internal + HBM2	Internal + HBM2
Form Factor	SXM2	PCIE Dual Slot	PCIE Dual Slot	PCIE LP	PCIE LP	PCIE Dual Slot	SXM2	PCIE Single Slot Full Height Half Length
Power	300 W	250 W	250 W	50-75 W	70 W	250W	300W	150W