

GPU COMPUTING

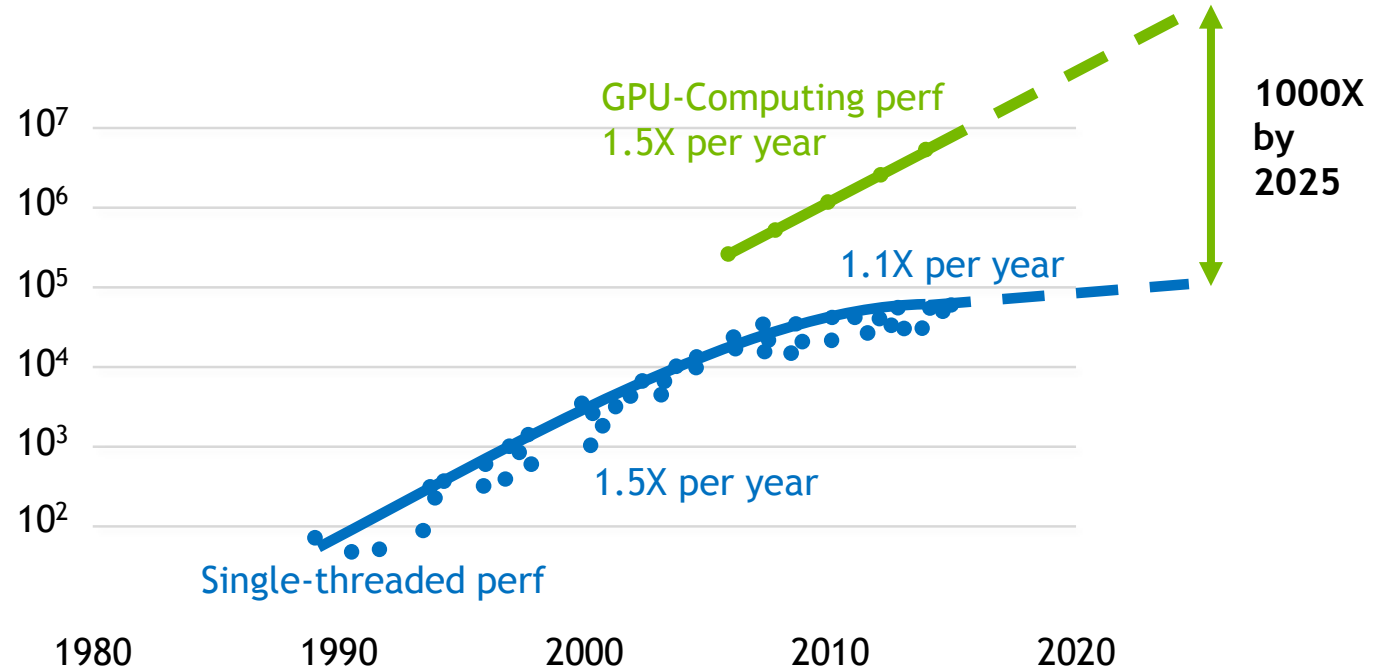
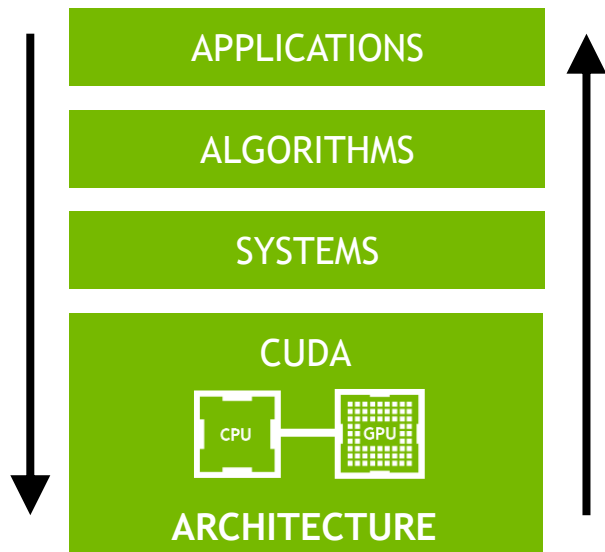
What's now ?



Guillaume BARAT
gbarat@nvidia.com
EMEA Higher-Education & Research

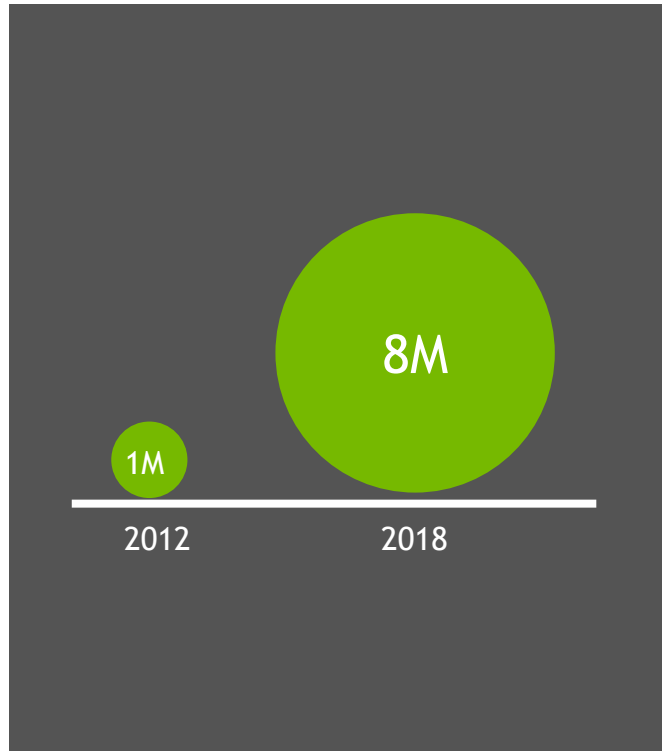


RISE OF GPU COMPUTING

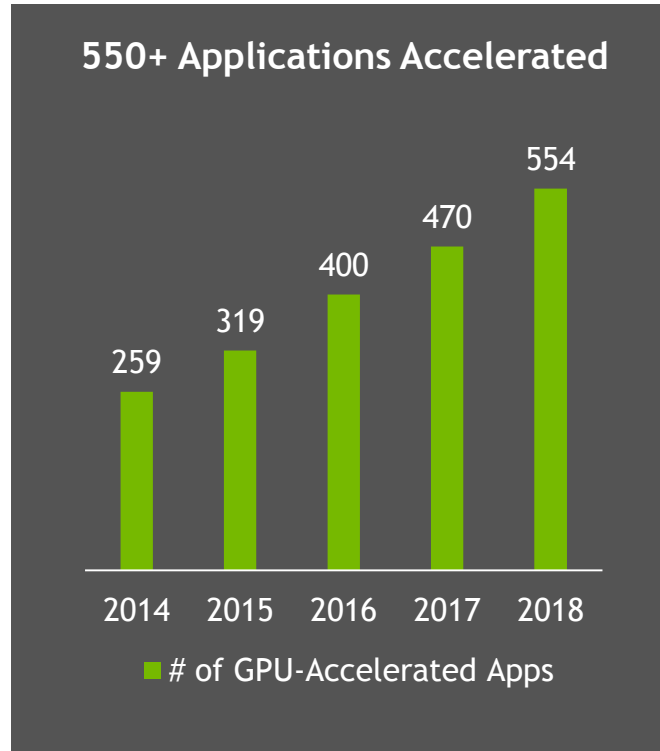


Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp

MOST ADOPTED PLATFORM FOR ACCELERATING HPC



8X CUDA DOWNLOADS



ALL TOP 15 APPLICATIONS ACCELERATED



OAK RIDGE SUMMIT

US's next fastest supercomputer
200+ Petaflop HPC; 3+ Exaflop of AI

ABCI Supercomputer (AIST)

Japan's fastest AI supercomputer

Piz Daint

Europe's fastest supercomputer

DEFINING THE NEXT GIANT WAVE IN HPC

GPU-ACCELERATED HPC APPLICATIONS

550+ APPLICATIONS

LIFE SCIENCES

50+
app

- Including:
- Gaussian
 - VASP
 - AMBER
 - HOOMD-Blue
 - GAMESS

MFG, CAD, & CAE

111
apps

- Including:
- Ansys
 - Fluent
 - Abaqus
 - SIMULIA
 - AutoCAD
 - CST Studio Suite

PHYSICS

25
apps

- Including:
- QUDA
 - MILC
 - GTC-P

OIL & GAS

18
apps

- Including:
- RTM
 - SPECFEM 3D

CLIMATE & WEATHER

3
apps

- Including:
- Cosmos
 - Gales
 - WRF

DEEP LEARNING

38
apps

- Including:
- Caffe2
 - MXNet
 - Tensorflow

MEDIA & ENT.

142
apps

- Including:
- DaVinci Resolve
 - Premiere Pro CC
 - Redshift Renderer

FEDERAL & DEFENSE

14
apps

- Including:
- ArcGIS Pro
 - EVNI
 - SocetGXP

DATA SCI. & ANALYTICS

23
apps

- Including:
- MapD
 - Kinetica
 - Graphistry

SAFETY & SECURITY

19
apps

- Including:
- Cyllance
 - FaceControl
 - Syndex Pro

COMP. FINANCE

16
apps

- Including:
- O-Quant Options Pricing
 - MUREX
 - MISYS

TOOLS & MGMT.

16
apps

- Including:
- Bright Cluster Manager
 - HPCtoolkit
 - Vampir

70% OF THE WORLD'S SUPERCOMPUTING WORKLOAD ACCELERATED

GROMACS

ANSYS Fluent

Gaussian

VASP

NAMD

Simula Abaqus

WRF

OpenFOAM

ANSYS

LS-DYNA

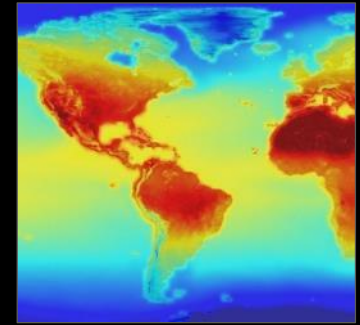
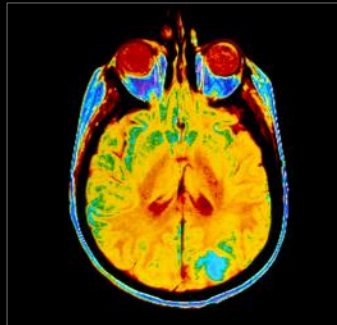
NCBI-BLAST

LAMMPS

AMBER

Quantum Espresso

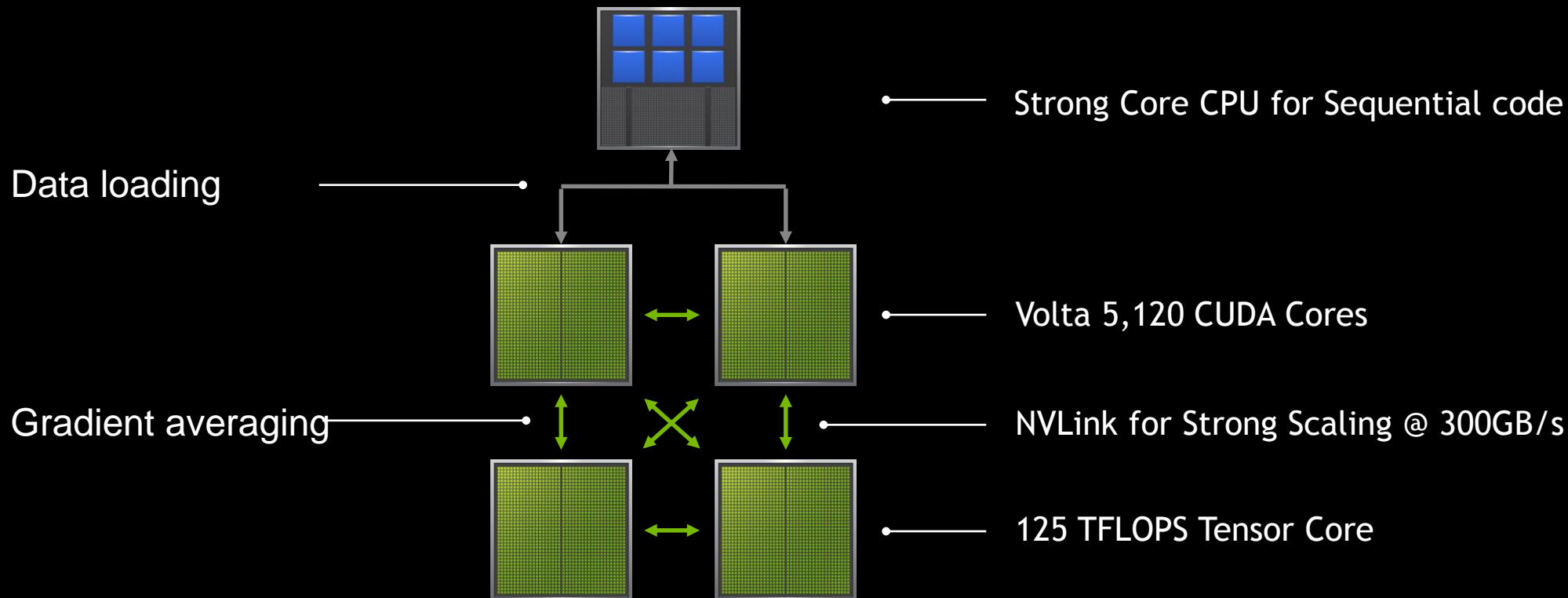
GAMSS



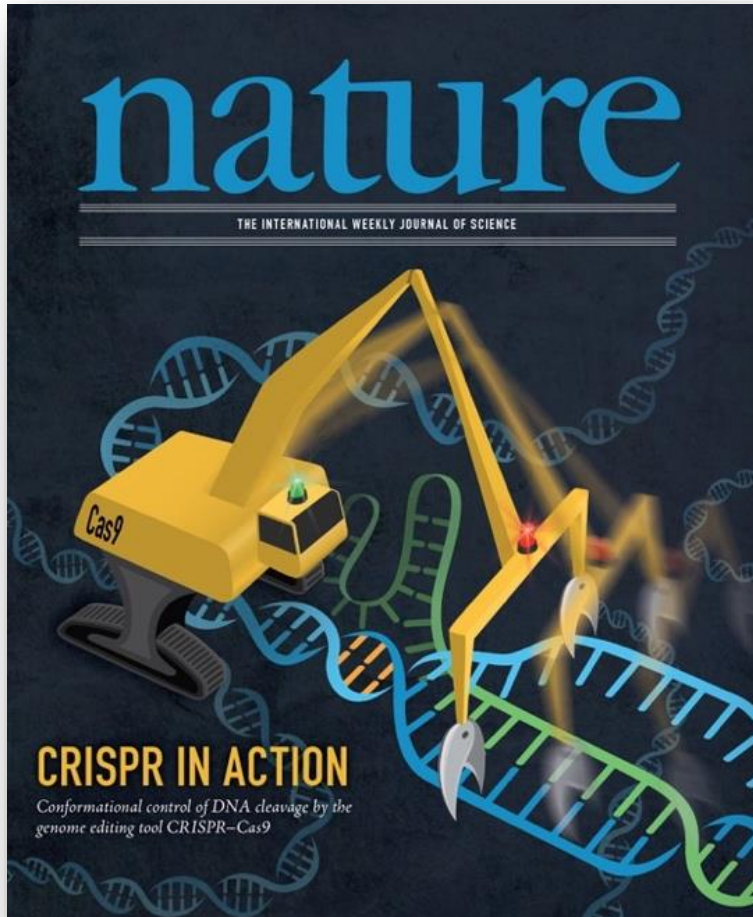
Top 15 HPC Applications

500+ Accelerated Applications

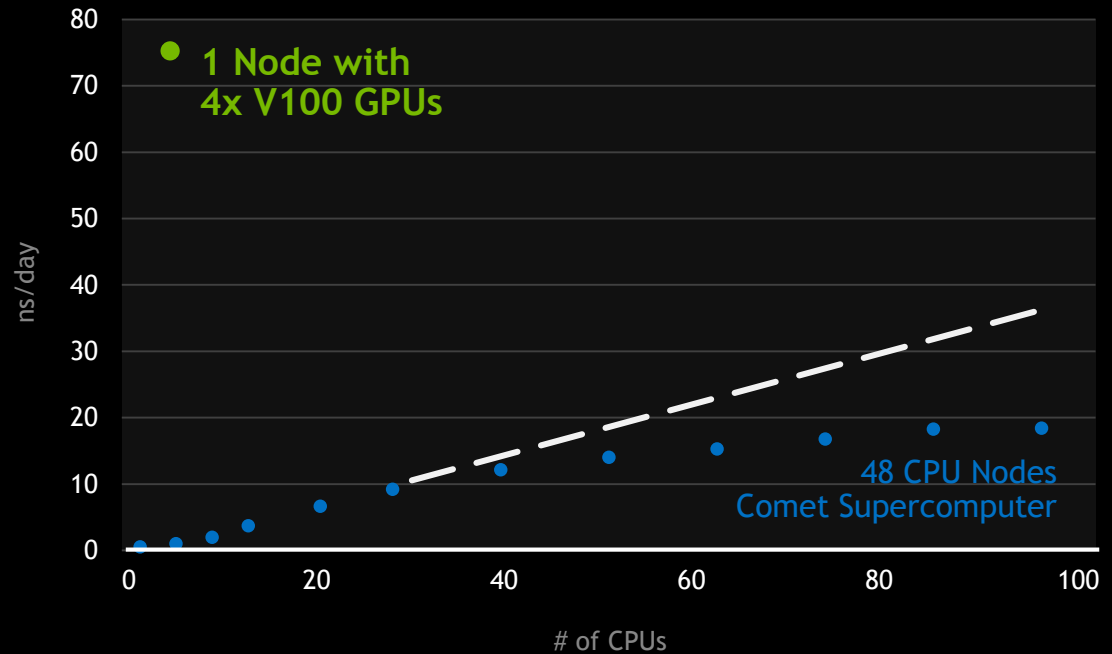
ARCHITECTING MODERN DATACENTERS



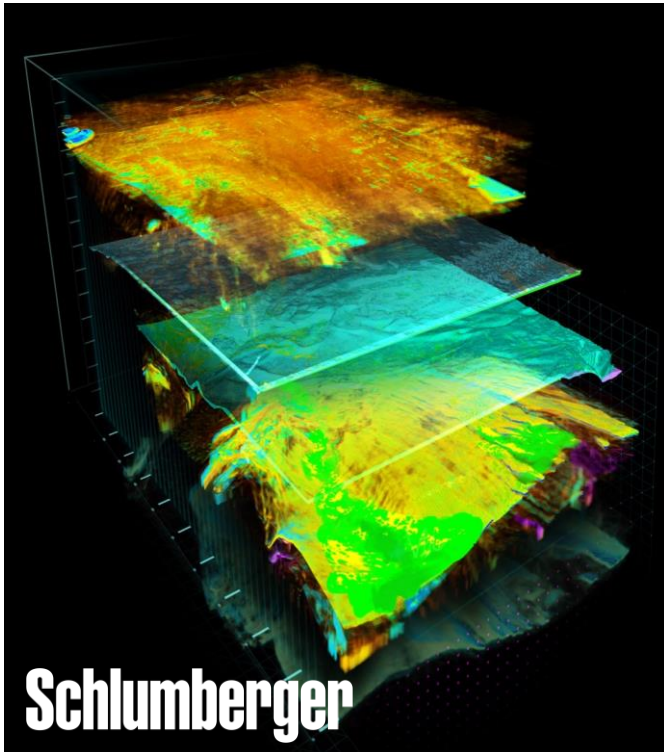
THE POWER OF ACCELERATED COMPUTING



AMBER Simulation of CRISPR



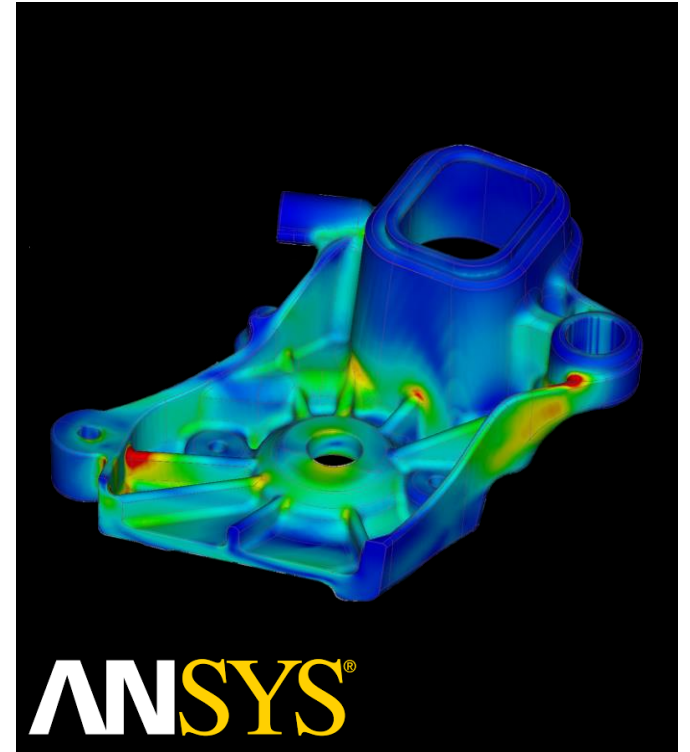
INDUSTRY EMBRACING GPU SUPERCOMPUTING



OIL AND GAS DISCOVERY
10X increase in data processing

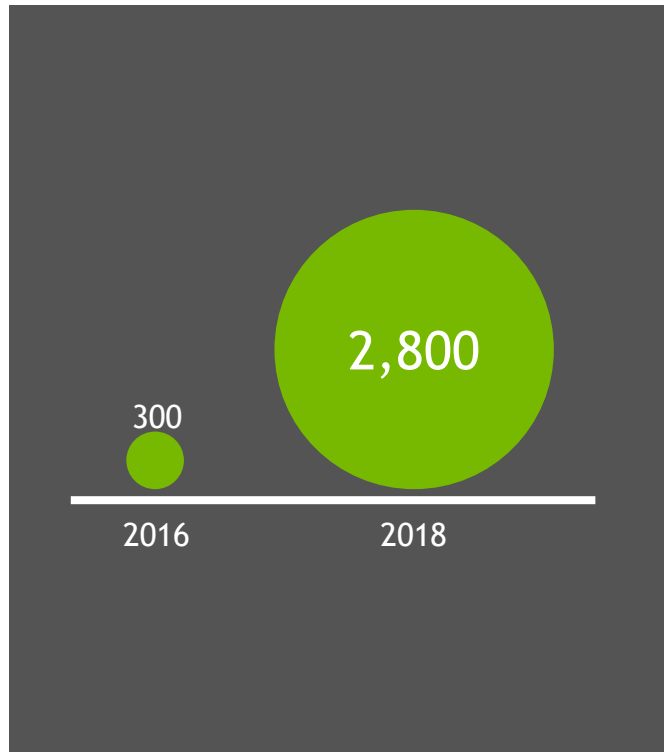


REALTIME FLEET ANALYTICS
Streamline routes to save >\$28M

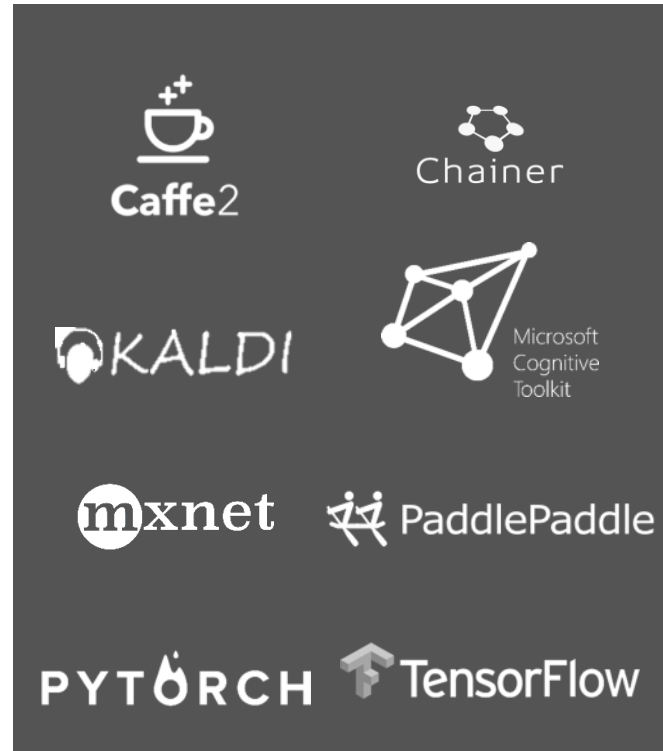


ENGINEERING DESIGN
Accelerate from hours to minutes

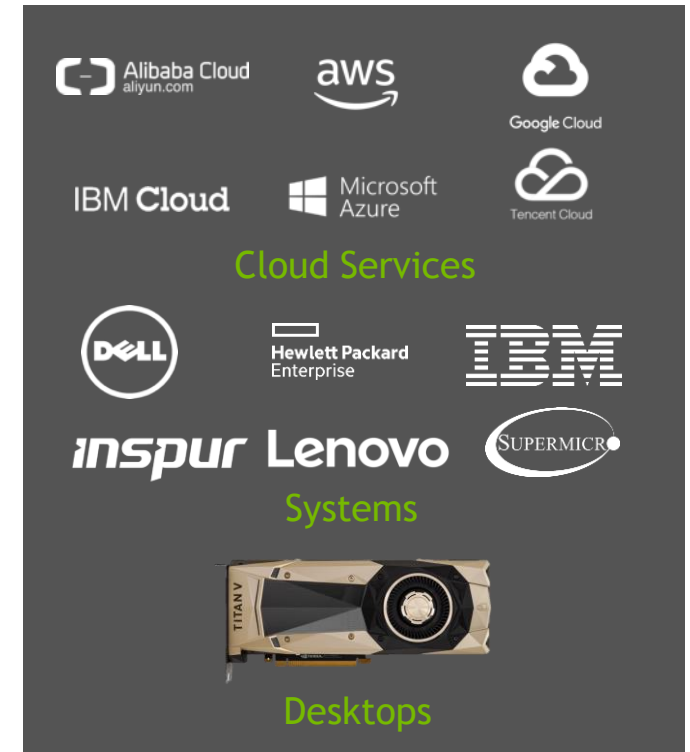
MOST ADOPTED PLATFORM FOR ACCELERATING AI



9X STARTUPS ENGAGED
VIA INCEPTION PROGRAM



EVERY DEEP LEARNING
FRAMEWORK ACCELERATED



AVAILABLE EVERYWHERE

INTELLIGENT HPC

DL Driving Future HPC Breakthroughs

- Trained networks as solvers
- Super-resolution of coarse simulations
- Low- and mixed-precision
- Simulation for training, network in production

From
calendar
time to real
time?

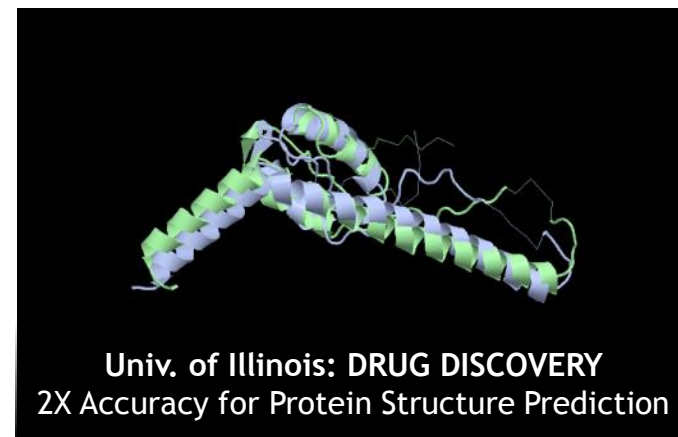
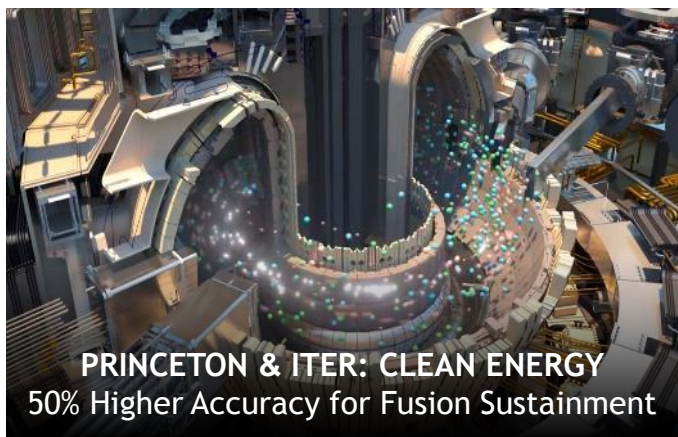
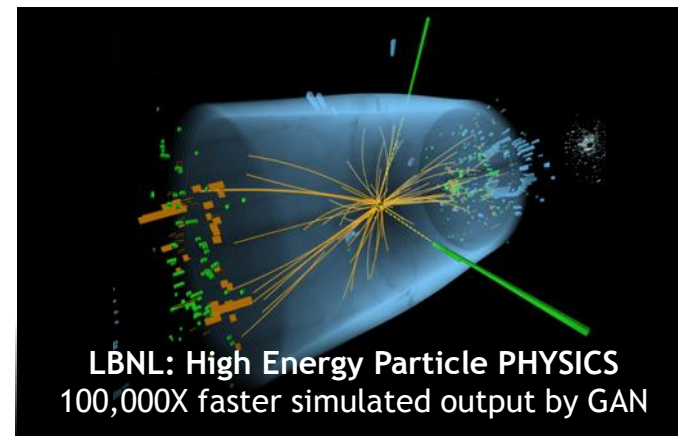
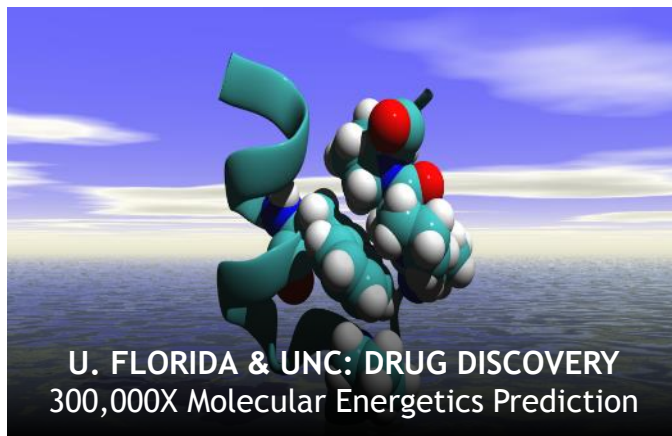
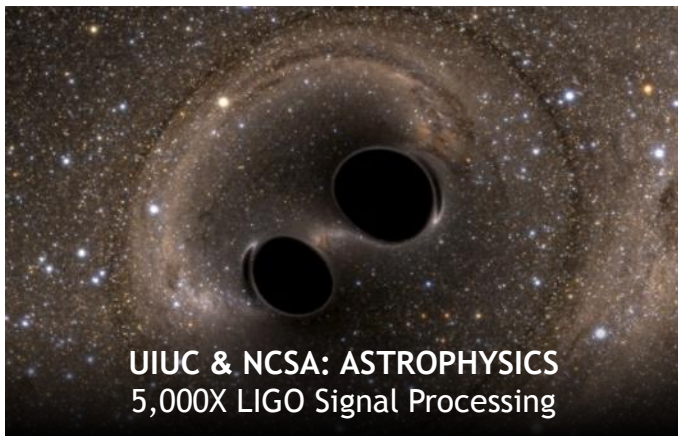


- Select/classify/augment/distribute input data
- Control job parameters

- Analyze/reduce/augment output data
- Act on output data

DEEP LEARNING COMES TO HPC

Accelerates Scientific Discovery



ONE PLATFORM BUILT FOR BOTH DATA SCIENCE & COMPUTATIONAL SCIENCE



The image shows a server rack with a Tesla GPU card installed. Above the server is a close-up of the GPU card. The NVIDIA logo is visible on the server's front panel.

CUDA

Tesla Platform



The cover of Nature journal features a Go board on a glowing blue circuit board. The text on the cover includes the title 'nature', the subtitle 'THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE', and the main headline 'At last — a computer program that can beat a champion Go player PAGE 484'. Below this is the large text 'ALL SYSTEMS GO'. At the bottom, there are three article teasers: 'CONSERVATION: SONGBIRDS A LA CARTE', 'RESEARCH ETHICS: SAFEGUARD TRANSPARENCY', and 'POPULAR SCIENCE: WHEN GENES GOT "SELFISH"'. The URL 'NATURE.COM/NATURE' and the date '22 January 2015' are also present.

Accelerating AI



The cover of Nature journal features a 3D model of an HIV-1 capsid. The text on the cover includes the title 'nature', the subtitle 'THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE', and the main headline 'THE HIV-1 CAPSID'. Below this is the text 'Atomic structure of the AIDS pathogen's protein coat PAGE 643'. At the bottom, there are three article teasers: 'COSMOLOGY: THE FIRST LIGHT', 'CITATION: CROSSING THE BORDERS', and 'ANTICANCER DRUGS: A SITTING TARGET'. The URL 'NATURE.COM/NATURE' and the date '22 May 2015' are also present.

Accelerating HPC

4X BETTER HPC SYSTEM TCO



Mixed Workload:
Materials Science (VASP)
Life Sciences (AMBER)
Physics (MILC)
Deep Learning (ResNet-50)

160 Self-hosted Servers
96 KWatts

4X BETTER HPC SYSTEM TCO



Mixed Workload:
Materials Science (VASP)
Life Sciences (AMBER)
Physics (MILC)
Deep Learning (ResNet-50)

12 Accelerated Servers w/4 V100 GPUs
20 KWatts

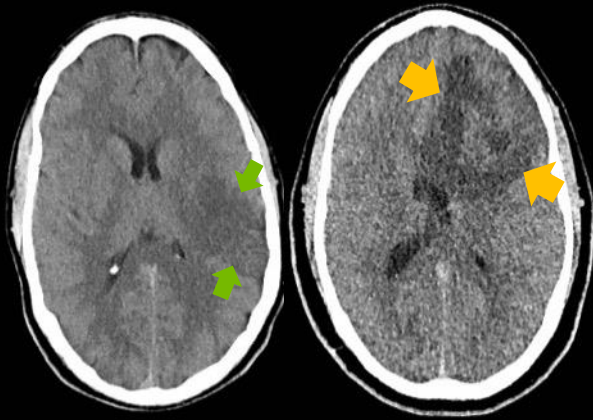
1/3 the Cost
1/4 the Space
1/5 the Power



CUSTOMERS WANT MORE

AI TO TRANSFORM EVERY INDUSTRY

HEALTHCARE



>80% Accuracy & Immediate
Alert to Radiologists

INFRASTRUCTURE



50% Reduction in Emergency
Road Repair Costs

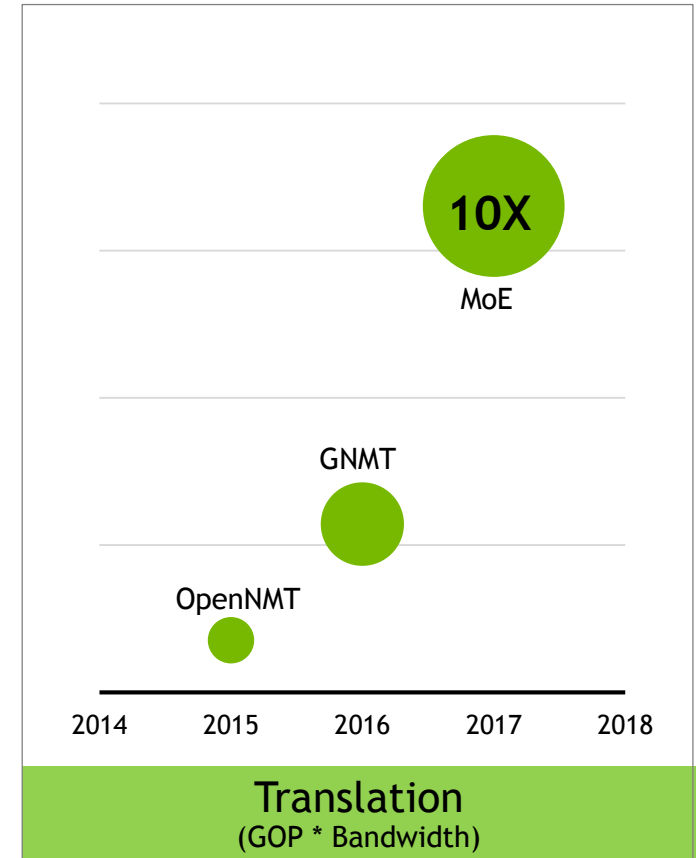
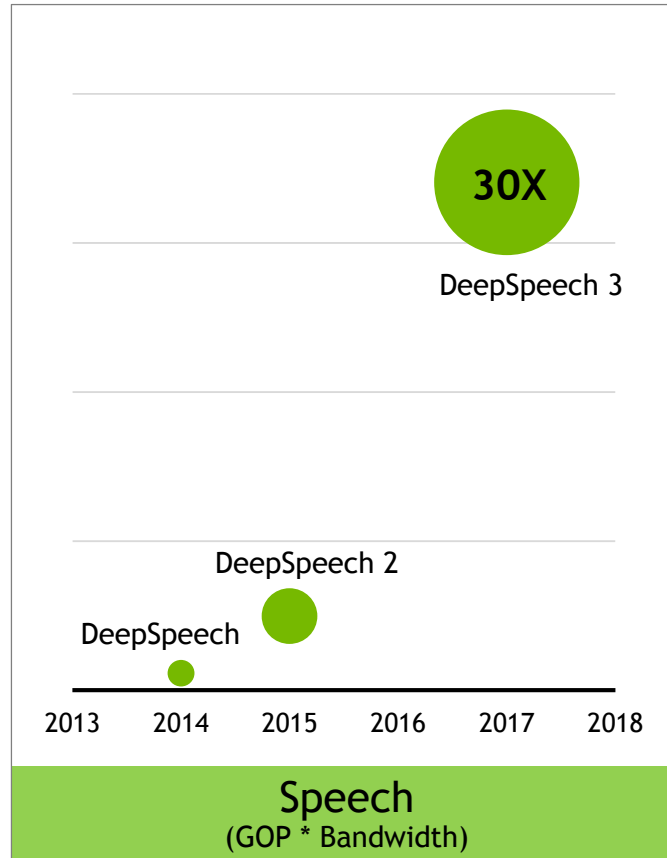
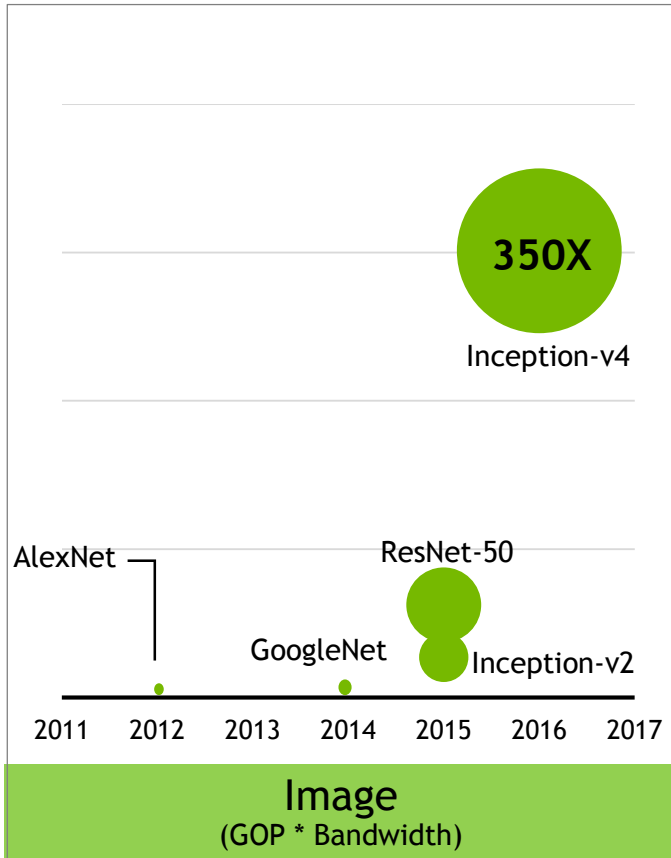
IOT



>\$6M / Year Savings and
Reduced Risk of Outage

NEURAL NETWORK COMPLEXITY IS EXPLODING

Bigger and More Compute Intensive



TESLA V100 32GB

WORLD'S MOST ADVANCED DATA CENTER GPU
NOW WITH 2X THE MEMORY

5,120 CUDA cores

640 NEW Tensor cores

7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS | 125 Tensor TFLOPS

20MB SM RF | 16MB Cache

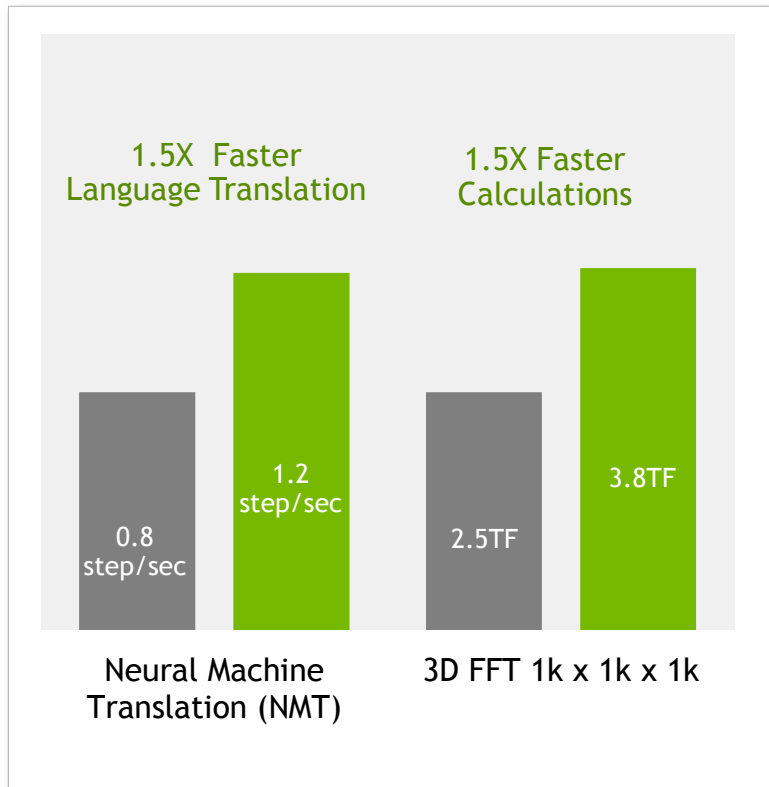
32GB HBM2 @ 900GB/s | 300GB/s NVLink



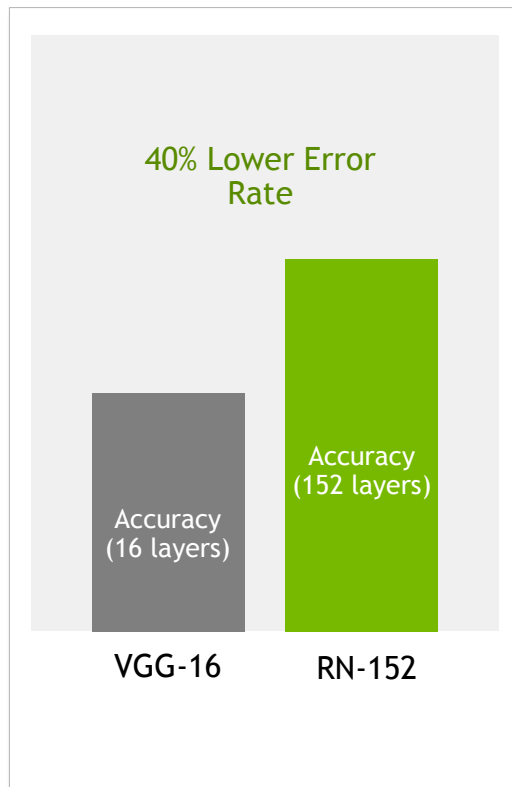
FASTER RESULTS ON COMPLEX DL AND HPC

Up to 50% Faster Results With 2x The Memory

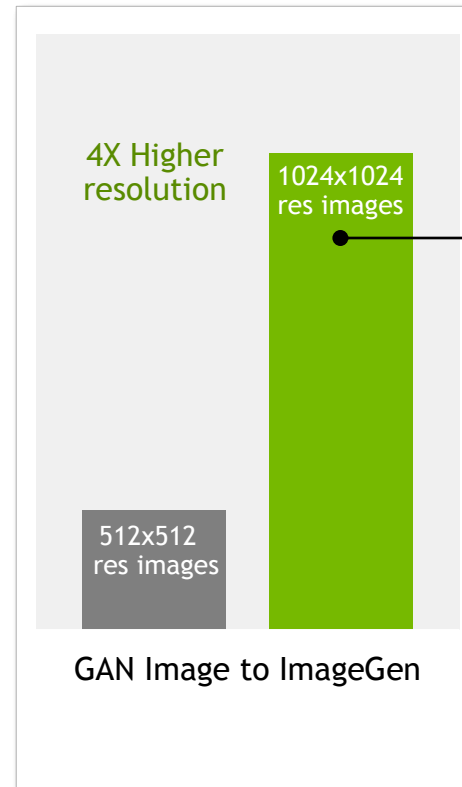
FASTER RESULTS



HIGHER ACCURACY



HIGHER RESOLUTION



Unsupervised Image Translation

Input winter photo



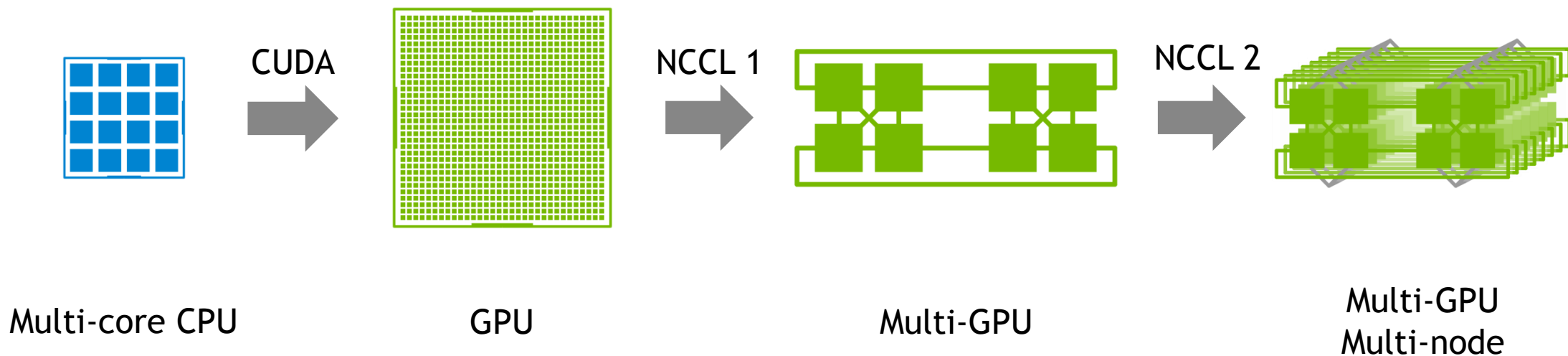
AI converts it to summer

■ V100 16GB ■ V100 32GB

DEEP LEARNING ON GPUS

Making DL training times shorter

Deeper neural networks, larger data sets ... training is a very, very long operation !



NVLINK MULTI-GPU SCALING

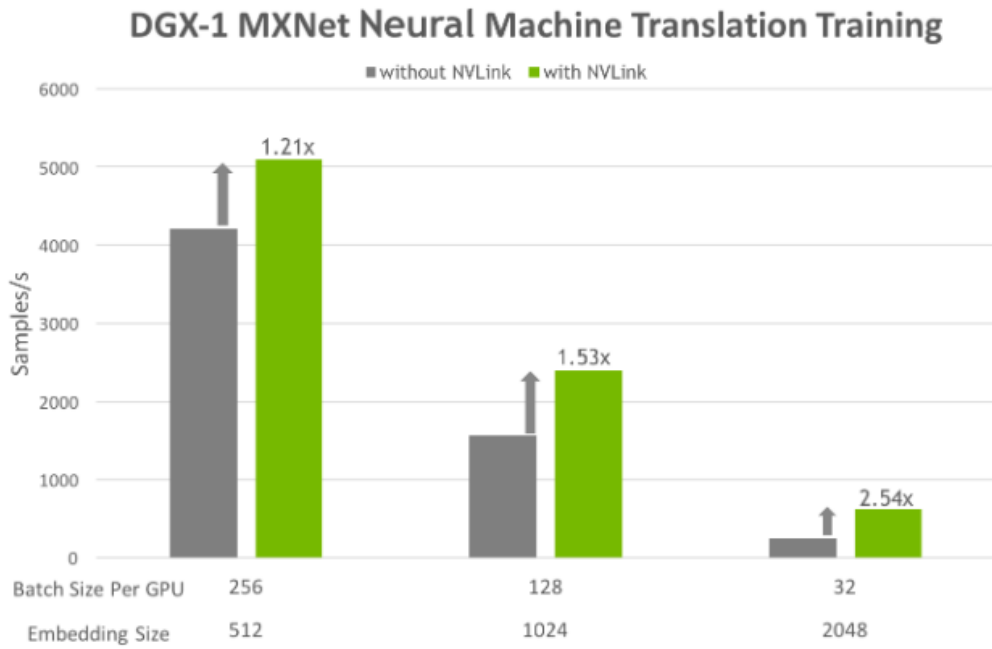


Figure 7 Sockeye neural machine translation single-precision training with MXNet using MLP attention on DGX-1, demonstrating significant NVLink performance benefits. The bars present performance on eight Tesla V100 GPUs in a DGX-1 when using NVLink for communication (green), and when using PCIe for communication (gray). Performance benefits increase with the encoder/decoder embedding size. Results are the average number of samples per second processed during a single epoch of training with the German to English dataset. Tests used NVIDIA DGX MXNet container version 17.11, processing real data with cuDNN 7.0.4, NCCL 2.1.2.

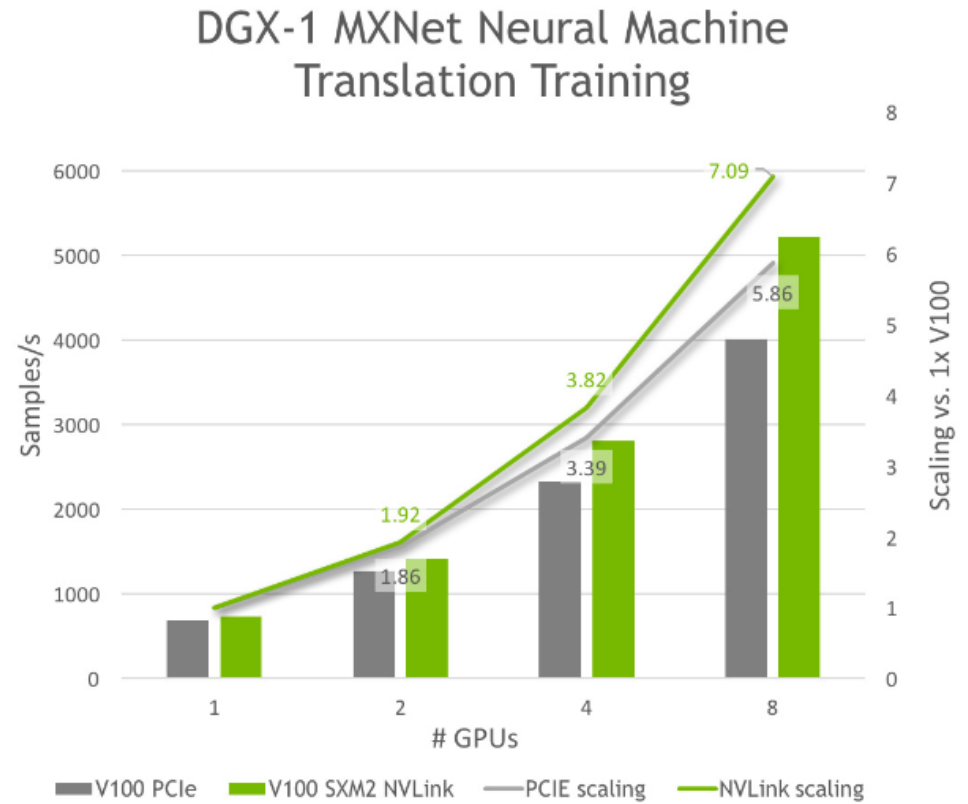
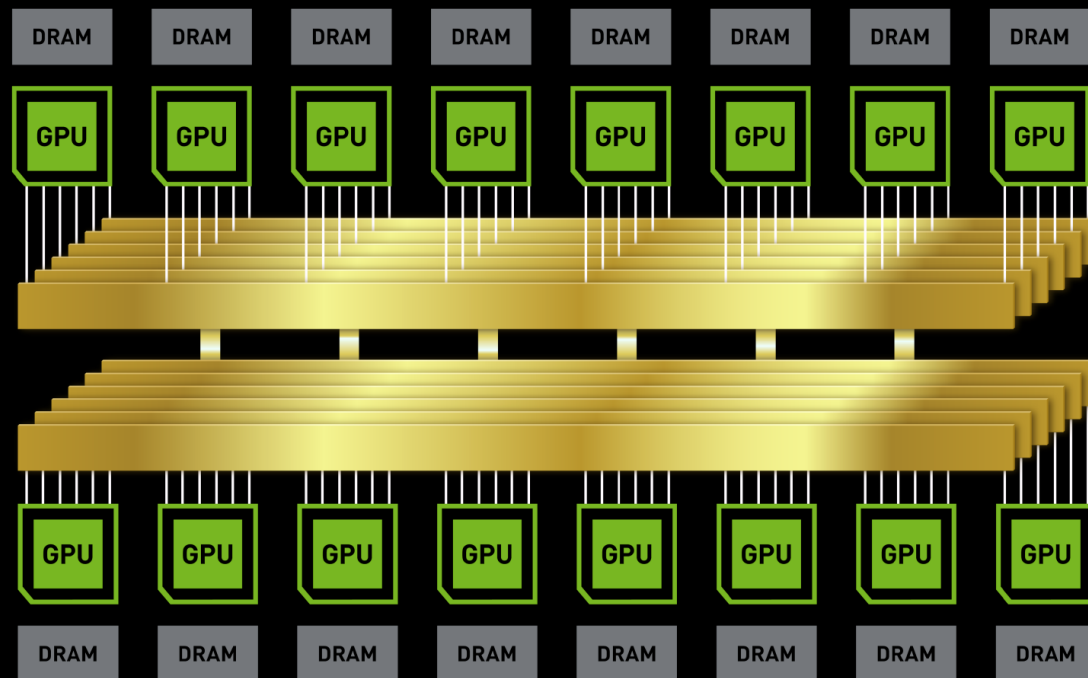


Figure 6 DGX-1 and V100 PCIe performance and scaling for single-precision training of a neural machine translation model with MLP attention and encoder/decoder embedding size of 512 and a batch size of 256 per GPU. The bars show performance on one, two, four, and eight GPUs, comparing an off-the-shelf system of eight Tesla V100 GPUs using PCIe for communication (gray) with eight Tesla V100 GPUs in a DGX-1 using NVLink for communication (green). The lines show the speedup compared to a single GPU. Tests used NVIDIA DGX containers version 17.11, processing real data with cuDNN 7.0.4, NCCL 2.1.2.

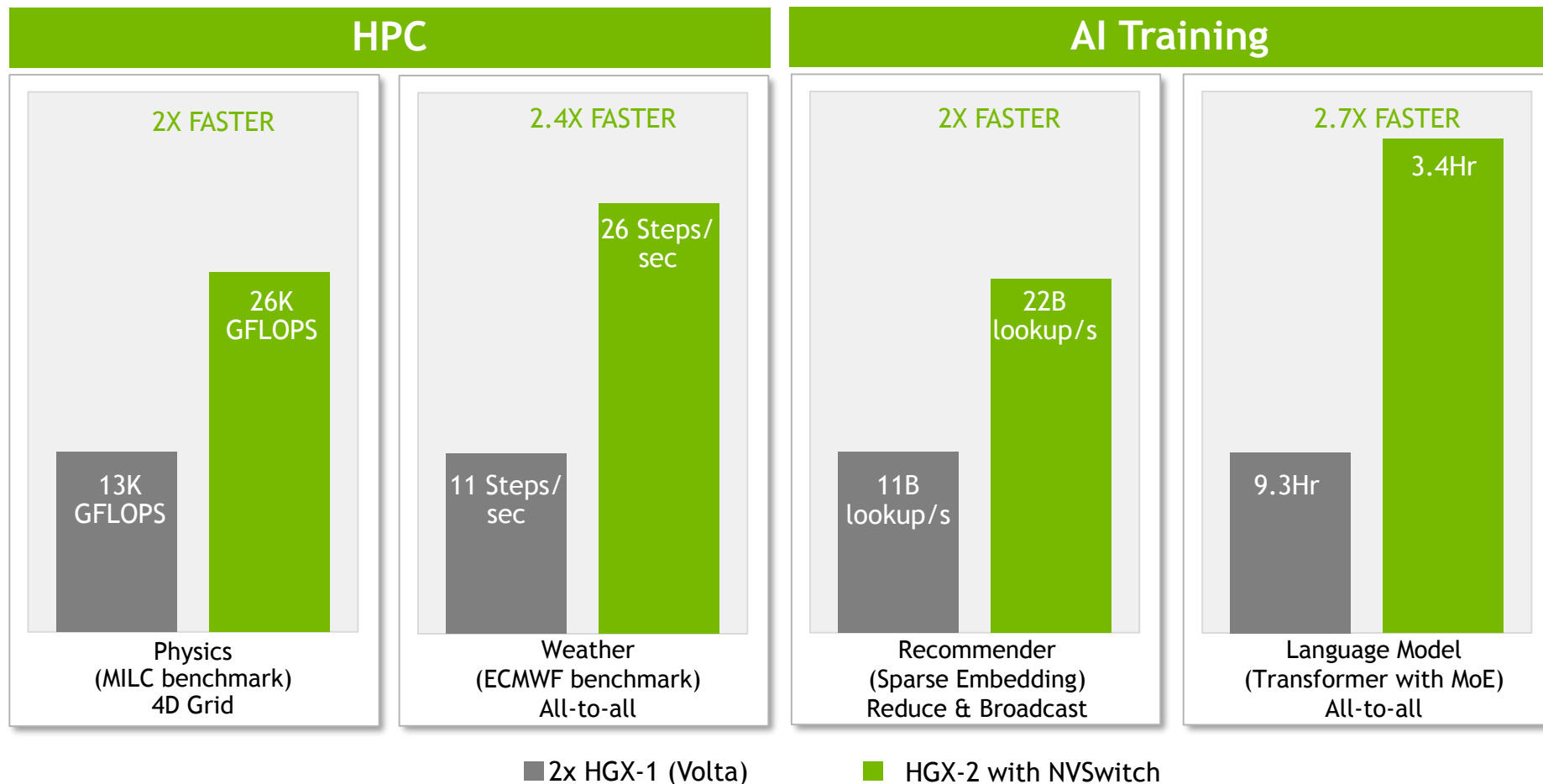
NVSWITCH

ENABLES THE WORLD'S LARGEST GPU

16 Tesla V100 32GB Connected by New NVSwitch
2 petaFLOPS of DL Compute
Unified 512GB HBM2 GPU Memory Space
300GB/sec Every GPU-to-GPU
2.4TB/sec of Total Cross-section Bandwidth



UP TO 3X HIGHER PERFORMANCE WITH NVSWITCH



**THE WORLD'S FIRST 2
PETAFLUPS SYSTEM**



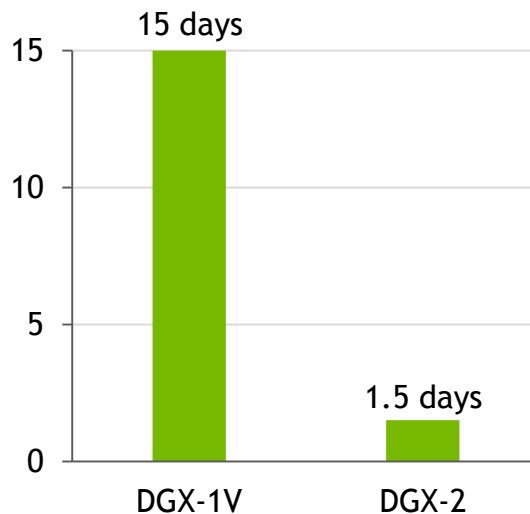
INTRODUCING NVIDIA DGX-2

THE WORLD'S MOST POWERFUL
AI SYSTEM FOR THE MOST COMPLEX
AI CHALLENGES

- DGX-2 is the newest addition to the DGX family, powered by DGX software
- Deliver accelerated AI-at-scale deployment and simplified operations
- Step up to DGX-2 for unrestricted model parallelism and faster time-to-solution

10X PERFORMANCE GAIN LESS THAN A YEAR

DGX-1, SEP'17



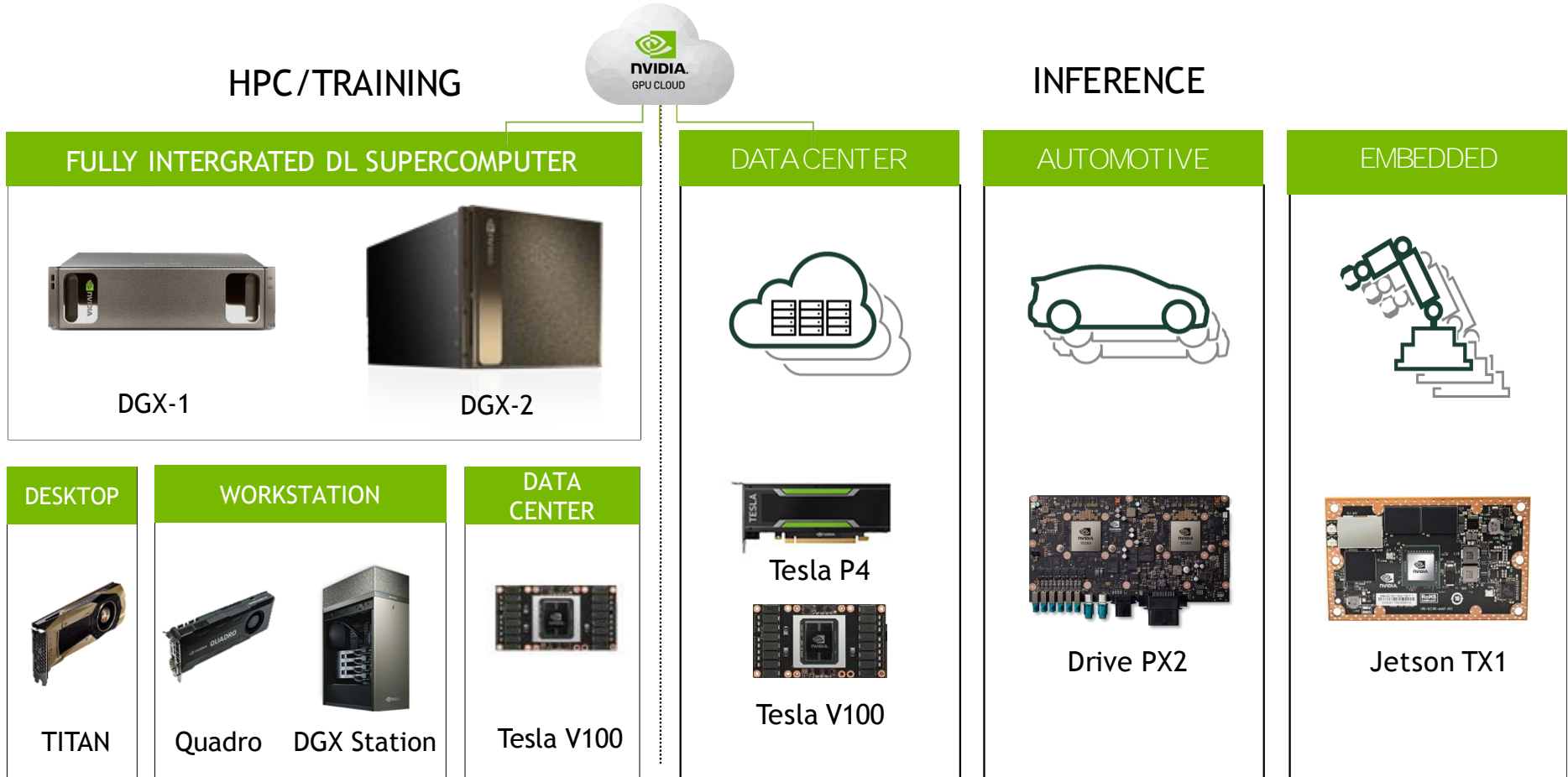
DGX-2, Q3'18



PyTorch Stack: Time to Train FAIRSEQ

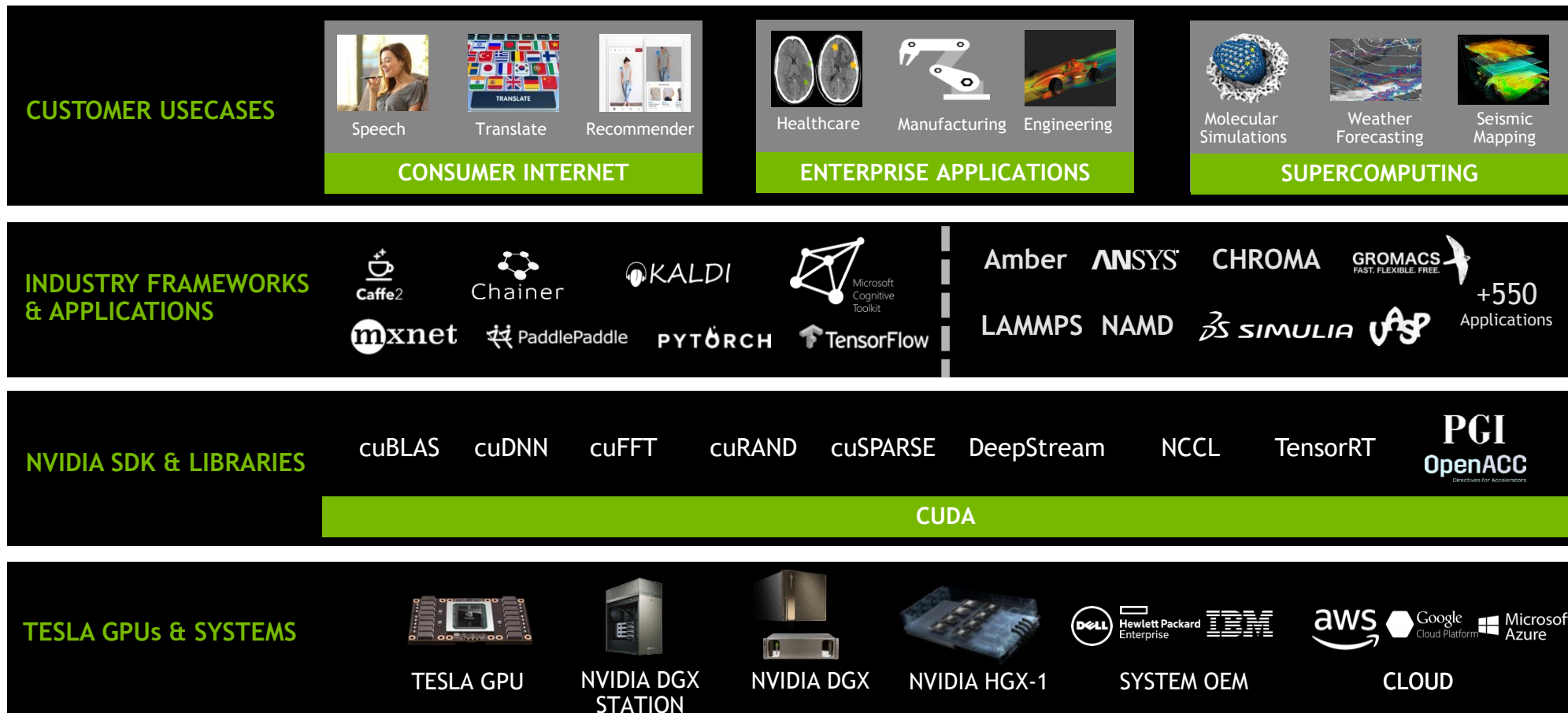
software improvements across the stack including NCCL, cuDNN, etc.

END-TO-END PRODUCT FAMILY



TESLA STACK

World's Leading Data Center Platform for Accelerating HPC and AI



NVIDIA GPU CLOUD SIMPLIFYING AI & HPC

Cloud container registry for GPU accelerated apps

Containerized in NVDocker

Optimized for GPU-accelerated Systems

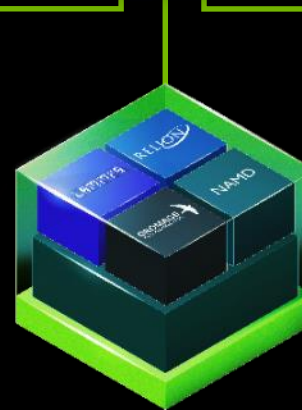
Up-to-Date Containers

Available NOW

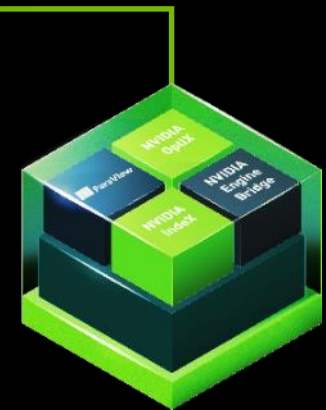
Sign up at nvidia.com/gpu-cloud



DEEP LEARNING

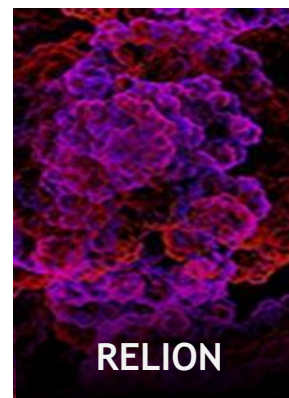
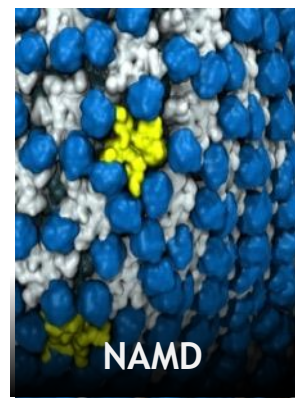
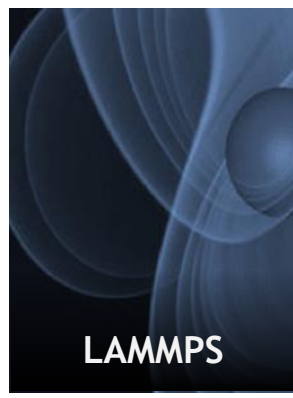
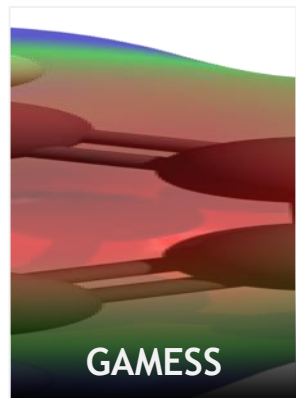
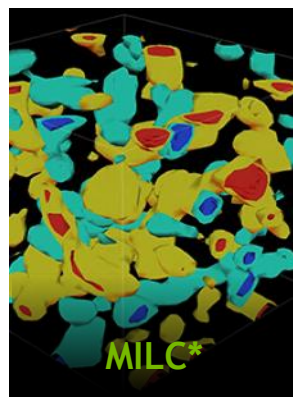
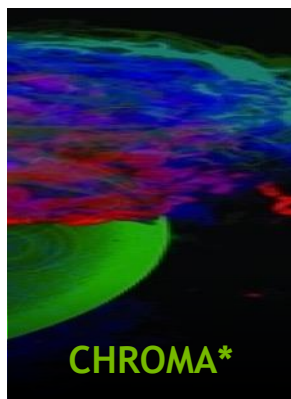


HPC APPS



HPC VIZ

HPC APPS CONTAINERS ON NVIDIA GPU CLOUD

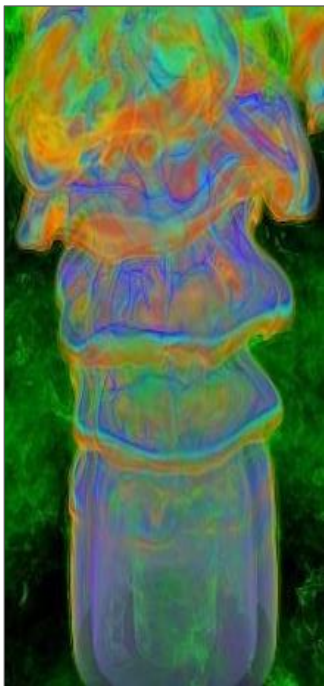


RAPID CONTAINER ADDITION

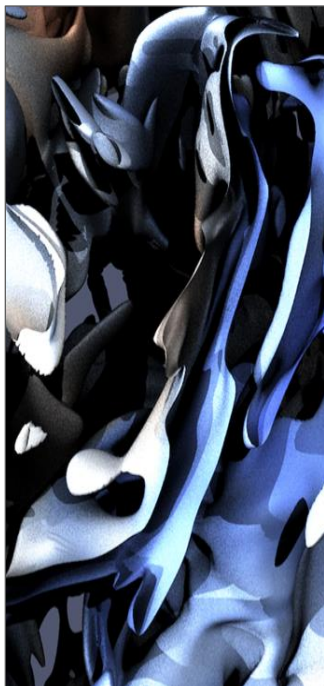


RAPID USER ADOPTION

NVIDIA GPU CLOUD FOR HPC VISUALIZATION



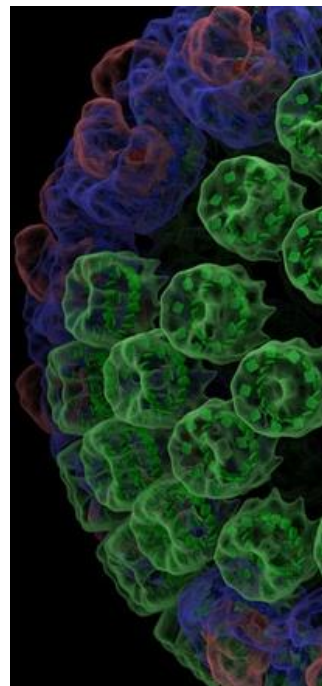
ParaView with
NVIDIA IndeX



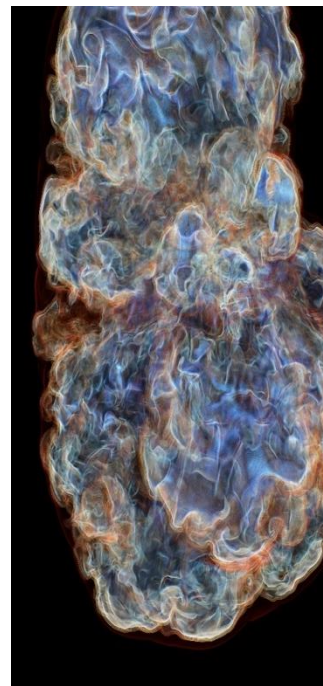
ParaView with
NVIDIA OptiX



ParaView with
NVIDIA Holodeck



VMD



IndeX



NEW CONTAINERS

TESLA PLATFORM FOR DEVELOPERS

NVIDIA SDK

The Essential Resource for GPU Developers

DEEP LEARNING

Deep Learning SDK

High-performance tools and libraries for deep learning

AUTONOMOUS VEHICLES

NVIDIA DRIVE Platform

Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous

VIRTUAL REALITY

NVIDIA VRWorks™

A comprehensive SDK for VR headsets, games and professional applications

GAME DEVELOPMENT

NVIDIA GameWorks™

Advanced simulation and rendering technology for game development

ACCELERATED COMPUTING

NVIDIA ComputeWorks™

Everything scientists and engineers need to build GPU-accelerated applications

DESIGN & VISUALIZATION

NVIDIA DesignWorks™

Tools and technologies to create professional graphics and advanced rendering applications

AUTONOMOUS MACHINES

NVIDIA JetPack™

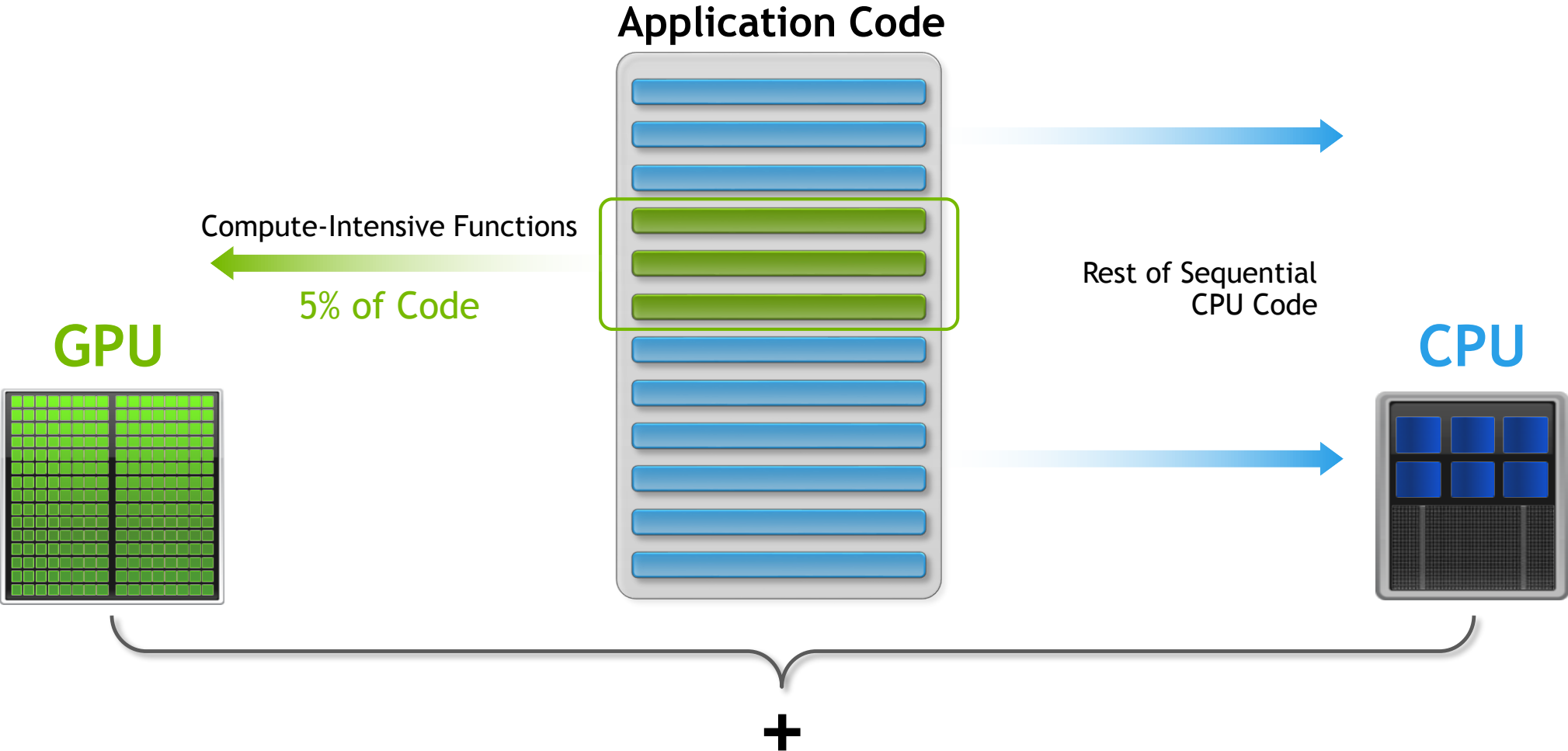
Powering breakthroughs in autonomous machines, robotics and embedded computing

SMART CITIES

NVIDIA Metropolis

Edge-to-cloud development platform for smart cities

HOW GPU ACCELERATION WORKS



HOW TO START WITH GPUS

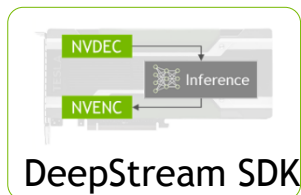
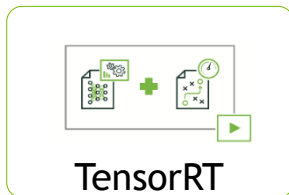
1 Applications		
2 Libraries	3 Compiler Directives	4 Programming Languages
Easy to use	Easy to Start	Most Performance
Most Performance	Portable Code	Most Flexibility
	OpenACC	CUDA

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

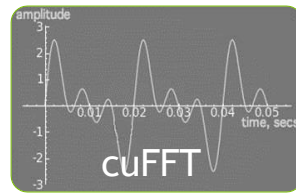
GPU ACCELERATED LIBRARIES

“Drop-in” Acceleration for Your Applications

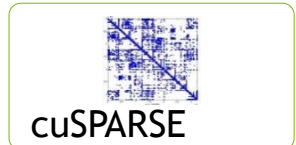
DEEP LEARNING



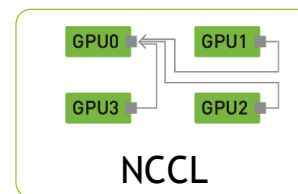
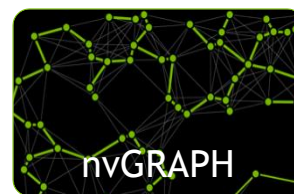
SIGNAL, IMAGE & VIDEO



LINEAR ALGEBRA



PARALLEL ALGORITHMS



WHAT IS OPENACC

Programming model for an easy onramp to GPUs

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

Add Simple Compiler Directive

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

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

SIMPLE

POWERFUL & PORTABLE

Read more at www.openacc.org/about

OpenACC is an open specification developed by OpenACC.org consortium

SINGLE CODE FOR MULTIPLE PLATFORMS

OpenACC - Performance Portable Programming Model for HPC

OpenPOWER

Sunway

x86 CPU

x86 Xeon Phi

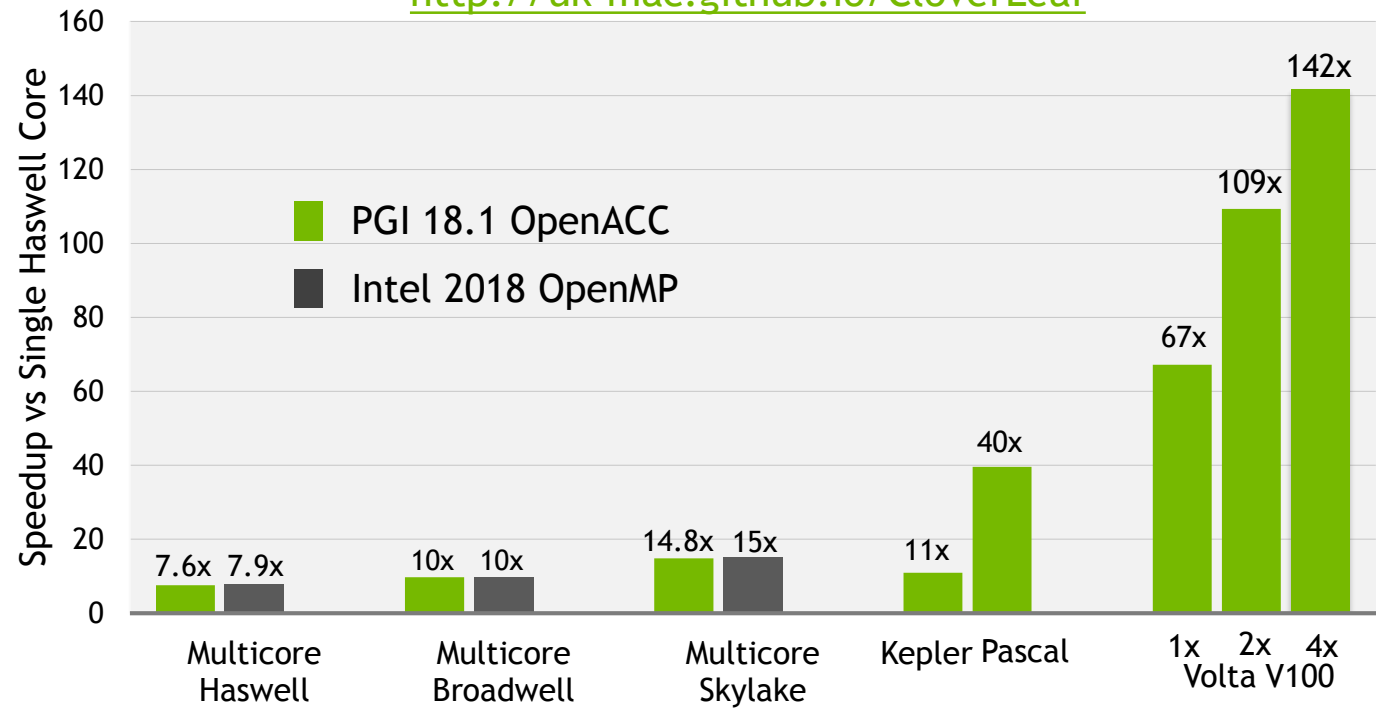
NVIDIA GPU

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; CloverLeaf_Serial; CloverLeaf_ref (MPI+OpenMP); CloverLeaf_OpenACC (MPI+OpenACC)
Data compiled by PGI February 2018.

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

OPENACC.ORG RESOURCES

Guides • Talks • Tutorials • Videos • Books • Spec • Code Samples • Teaching Materials • Events • Success Stories • Courses • Slack • Stack Overflow

OpenACC Now in GCC



<https://www.openacc.org/community#slack>

Resources

<https://www.openacc.org/resources>

A screenshot of the OpenACC website's "Resources" page. The page features a navigation bar with "About", "Tools", "News", "Events", "Resources", "Spec", and "Community". Below the navigation bar, the "Resources" section is introduced as a complete library of materials. It is divided into three main categories: "Guides" (with links to "Introduction to OpenACC Quick Guides", "OpenACC Programming and Best Practices Guide", and "OpenACC 2.3 API Reference Card"), "Books" (with "Parallel Programming with OpenACC" and "Programming Massively Parallel Processors, Third Edition: A Hands-on Approach"), and "Tutorials" (described as video tutorials to help start with OpenACC).

Success Stories

<https://www.openacc.org/success-stories>

A screenshot of the OpenACC website's "Success Stories" page. It features a navigation bar and a main heading "Success Stories". Below the heading, a paragraph states that applications across multiple domains have been accelerated with OpenACC. A carousel of three video thumbnails is displayed, each with a play button icon. The first video is titled "Researchers are using GPUs and OpenACC to accelerate the codes for their data stream simulations". The second is "Learn how OpenACC can simplify parallel programming and deliver high performance results". The third is "Alexe Sebest shows how she is using OpenACC to accelerate smaller programs to understand more science". A link at the bottom right says "Watch more OpenACC Videos on YouTube".

Compilers and Tools

<https://www.openacc.org/tools>

A screenshot of the OpenACC website's "Downloads & Tools" page. It features a navigation bar and a main heading "Downloads & Tools". Below the heading, a paragraph states that OpenACC compilers, profilers and debuggers are designed and available to download from multiple vendors and academic organizations. The page is divided into two sections: "Commercial Compilers" and "Open Source Compilers". Under "Commercial Compilers", there are logos for "CRAY THE SUPERCOMPUTER COMPANY", "PGI", and "中国科学院计算技术研究所" (Institute of Computing Technology, Chinese Academy of Sciences). Under "Open Source Compilers", there is a logo for "GCC" and a note that it "Includes initial support for OpenACC 2.3".

Events

<https://www.openacc.org/events>

A screenshot of the OpenACC website's "Events" page. It features a navigation bar and a main heading "Events". Below the heading, a paragraph states that the OpenACC Community organizes a variety of events throughout the year. A large image shows a group of people at a workshop or meeting. Below the image, there is a "2017 Calendar" section with a table listing events. The first event is "WORKSHOP Parallel Programming with OpenACC on CPUs and GPUs" on August 15, 2017, at Stanford University, Palo Alto, CA. A link at the bottom right says "Hackathons".

PGI COMPILERS FOR EVERYONE

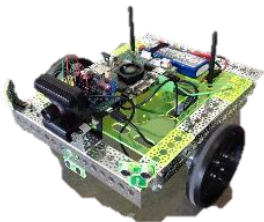
The PGI 17.10 Community Edition

	FREE PGI® Community EDITION	PGI® Professional EDITION	PGI® Enterprise EDITION
PROGRAMMING MODELS OpenACC, CUDA Fortran, OpenMP, C/C++/Fortran Compilers and Tools	✓	✓	✓
PLATFORMS X86, OpenPOWER, NVIDIA GPU	✓	✓	✓
UPDATES	1-2 times a year	6-9 times a year	6-9 times a year
SUPPORT	User Forums	PGI Support	PGI Premier Services
LICENSE	Annual	Perpetual	Volume/Site

NVIDIA ACADEMIC PROGRAMS

EDUCATION

NVIDIA Teaching kits

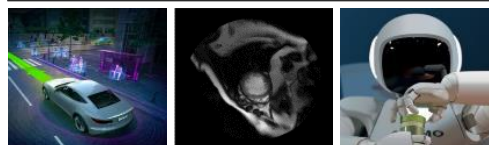
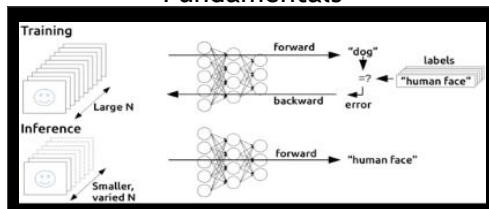


CAL POLY
SAN LUIS OBISPO



DLI Ambassador
Hands-on training

Fundamentals



Autonomous Vehicles Healthcare Robotics

GPU ACCESS

To start / teach
GeForce, Jetson



To scale
DGX, Tesla, GRID, Cloud



- Complete course solutions:
1. Deep Learning
 2. Accelerated Computing
 3. Robotics

Both Fundamentals and
Industry-Specific Labs
Available

Free for [DLI Ambassador](#)

Perfect platform to
prototype

Accelerating your ideas
Enabling
new scientific publication

Get an academic discount
on Tesla, DGX, [Jetson TX2](#) or
GRID from our [NPN partners](#)

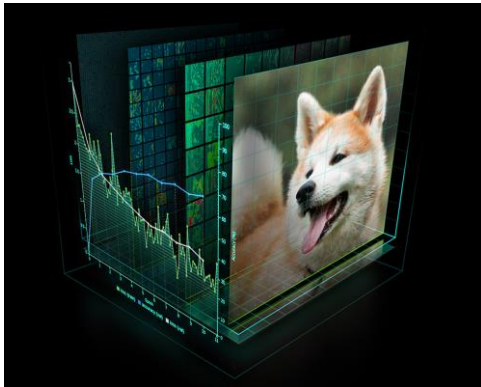
<https://developer.nvidia.com/academia>

GPU TECHNOLOGY CONFERENCE

Oct 9 -11, 2018 | Munich

www.gputechconf.eu #GTC18

EUROPE'S BRIGHTEST MINDS & BEST IDEAS



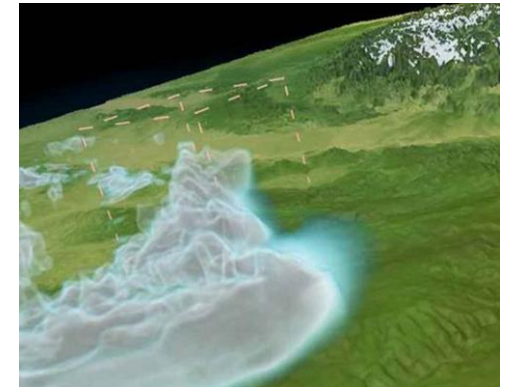
DEEP LEARNING &
ARTIFICIAL INTELLIGENCE



SELF-DRIVING CARS



VIRTUAL REALITY &
AUGMENTED REALITY



SUPERCOMPUTING & HPC

Join us in Munich and discover the latest breakthroughs in autonomous vehicles, HPC, smart cities, healthcare, big data, virtual reality, and more.

3 Days | 3000 Attendees | 80+ Exhibitors | 100+ Speakers | 10+ Tracks | 10+ Hands-on labs | 1-to-1 Meetings

-25% with the promo code NVGBARAT

