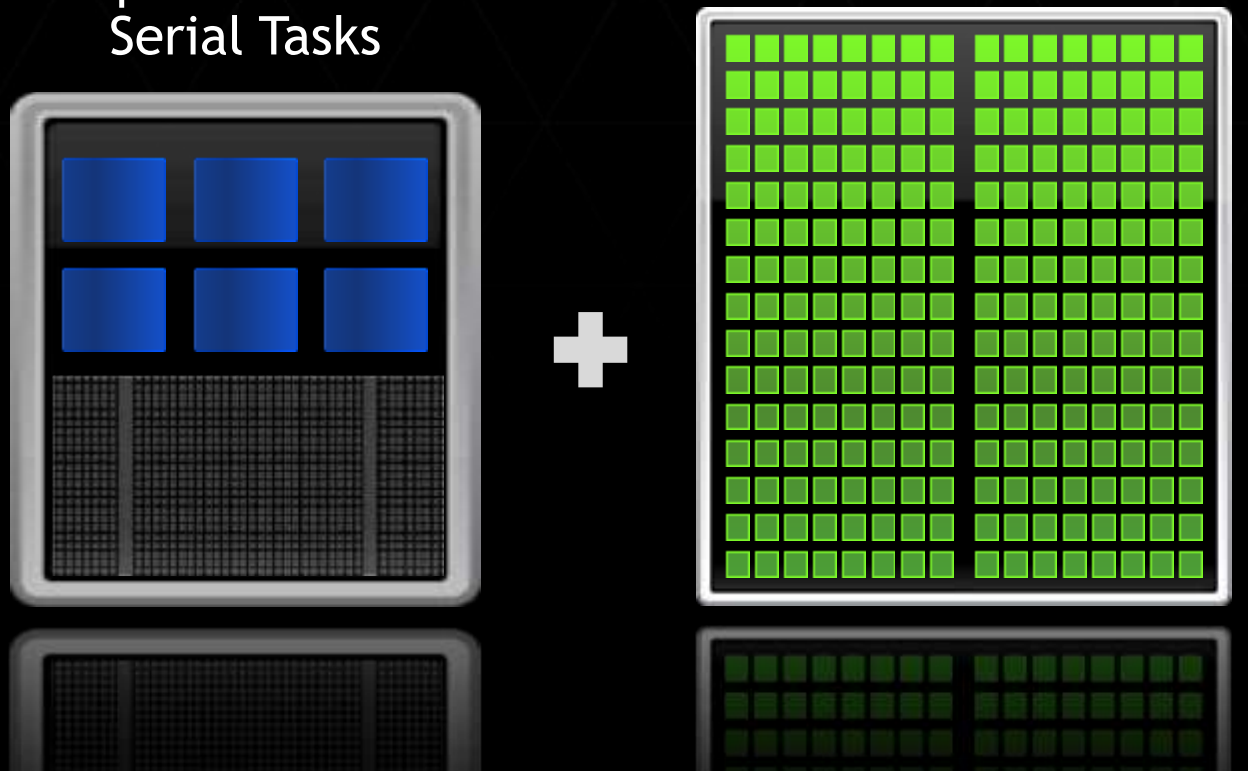
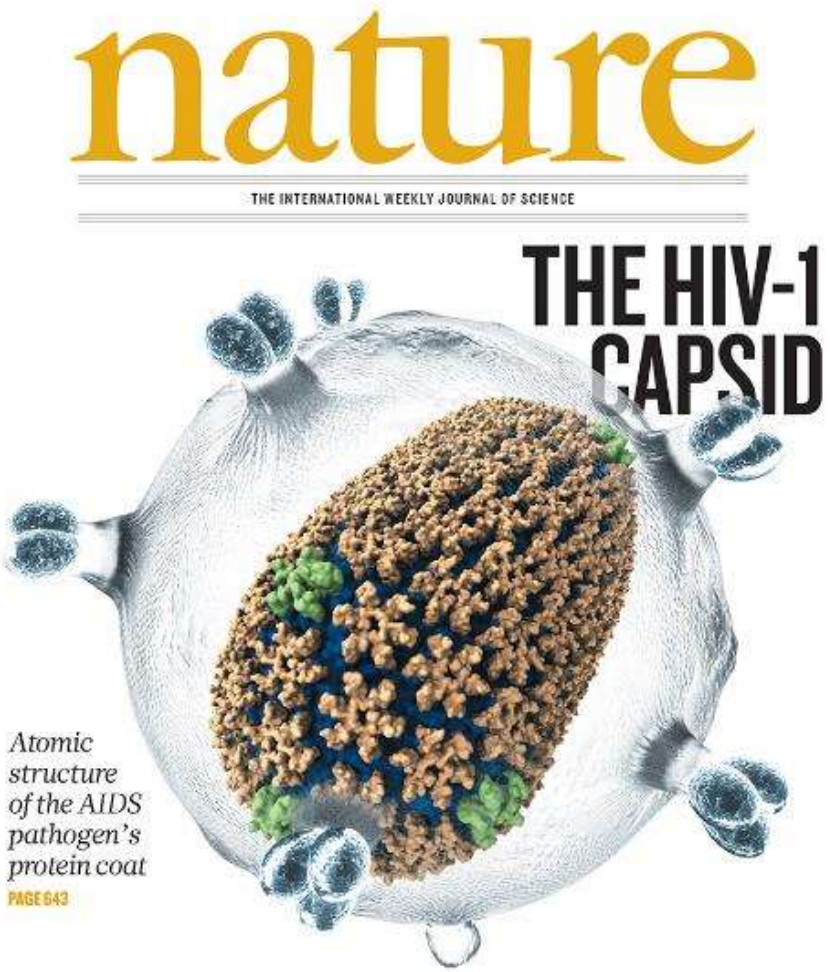
# ACCELERATED COMPUTING

## 10X PERFORMANCE & 5X ENERGY EFFICIENCY

**CPU**
Optimized for
Serial Tasks

**GPU Accelerator**
Optimized for
Parallel Tasks






# ACCELERATING DISCOVERIES

USING A SUPERCOMPUTER POWERED BY 3,000 TESLA PROCESSORS, UNIVERSITY OF ILLINOIS SCIENTISTS PERFORMED THE FIRST ALL-ATOM SIMULATION OF THE HIV VIRUS AND DISCOVERED THE CHEMICAL STRUCTURE OF ITS CAPSID – "THE PERFECT TARGET FOR FIGHTING THE INFECTION."

WITHOUT GPU, THE SUPERCOMPUTER WOULD NEED TO BE 5X LARGER FOR SIMILAR PERFORMANCE.

# ACCELERATING INSIGHTS

*“Now You Can Build Google’s $1M Artificial Brain on the Cheap”*

WIRED

## GOOGLE DATACENTER

1,000 CPU Servers
2,000 CPUs • 16,000 cores

**600 kWatts**
**$5,000,000**

## STANFORD AI LAB

3 GPU-Accelerated Servers
12 GPUs • 18,432 cores

**4 kWatts**
**$33,000**

*Deep learning with COTS HPC systems*, A. Coates, B. Huval, T. Wang, D. Wu, A. Ng, B. Catanzaro ICML 2013

# FROM HPC TO ENTERPRISE DATACENTERS

## Oil & Gas

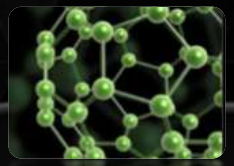
Schlumberger

BR PETROBRAS

Eni

Chevron

Statoil



## Higher Ed

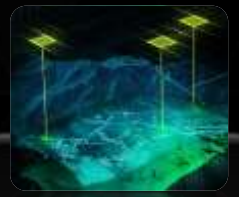
HARVARD School of Engineering and Applied Sciences

STANFORD UNIVERSITY

Georgia Tech

ETH

UNIVERSITY OF CAMBRIDGE



## Government

Air Force Research Laboratory

Raytheon

NASA

Naval Research Laboratory

## Supercomputing

CSCS

NCSA

Tokyo Institute of Technology

OAK RIDGE National Laboratory

Lawrence Livermore National Laboratory

## Finance

J.P.Morgan

BARCLAYS

STANDARD LIFE

BNP PARIBAS

MUREX

## Consumer Web

Baidu 百度

salesforce

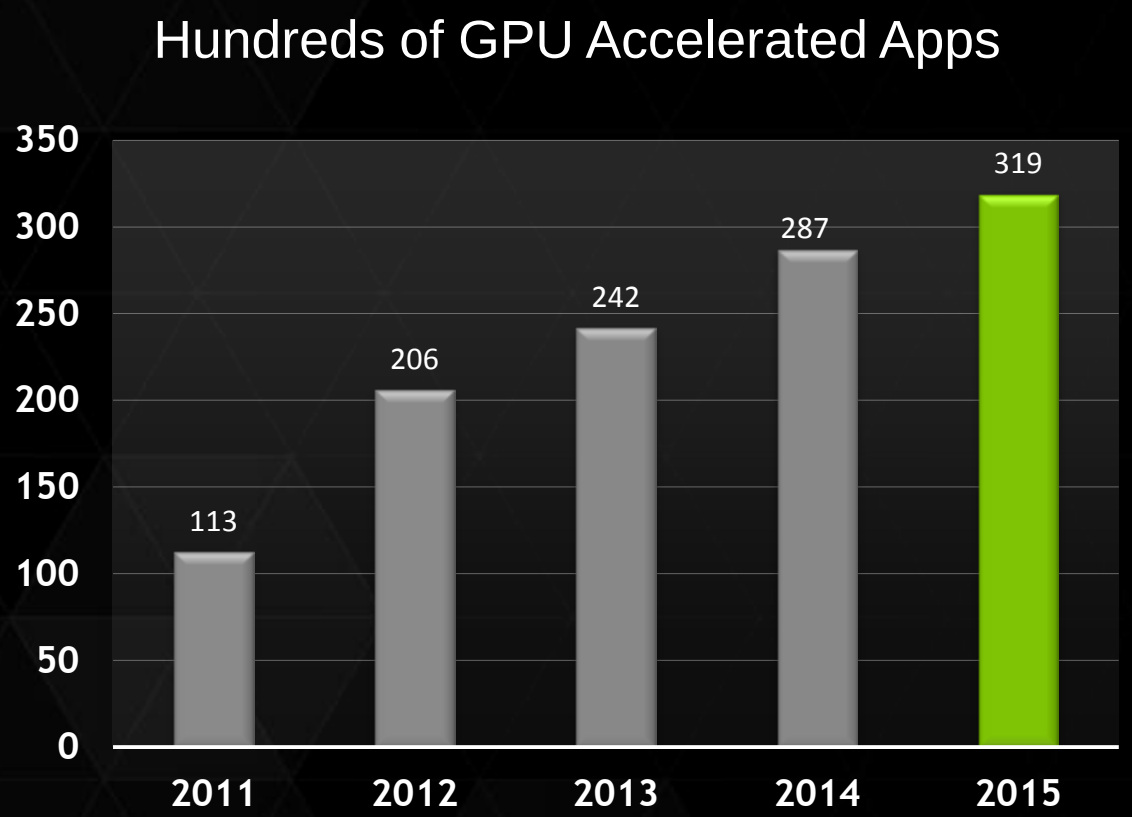
SHAZAM

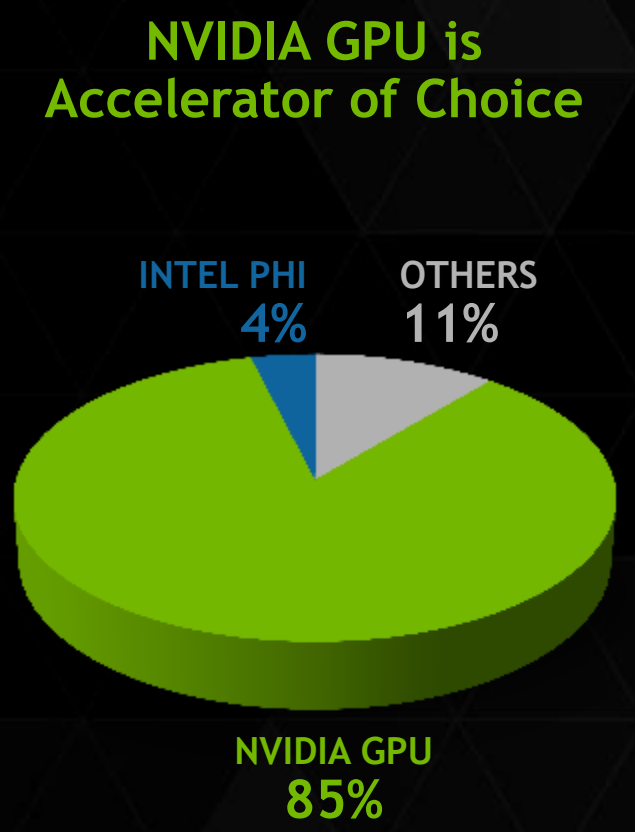
amazon.com

Yandex


NVIDIA
POPULAR GPU-ACCELERATED APPLICATIONS
02 Research: Higher Education and Supercomputing
06 Defense and Intelligence
07 Computational Finance
08 Manufacturing: CAD and CAE
10 Media and Entertainment
14 Oil and Gas
Research: Higher Education and Supercomputing
COMPUTATIONAL CHEMISTRY AND BIOLOGY
Bioinformatics
Molecular Dynamics
287 GPU-Accelerated Applications
www.nvidia.com/appscatalog

# RAPID ADOPTION OF ACCELERATED COMPUTING




Intersect360 Research
HPC User Site Census: Systems, July 2013

# HIGH GPU DENSITY SERVERS NOW MAINSTREAM

CRAY
THE SUPERCOMPUTER COMPANY

**Cray CS-Storm**
**8 K80s per Node**

DELL

**Dell C4130**
**4 K80s per Node**

hp

**HP SL270**
**8 K80s per Node**

Quanta Computer

**Quanta S2BV**
**4 K80s per Node**

# TESLA ACCELERATED COMPUTING PLATFORM

## Data Center Infrastructure

### System Solutions

CRAY
DELL
IBM
hp
lenovo

### Communication

Mellanox TECHNOLOGIES
MVAPICH
OPEN MPI

### Infrastructure Management

Adaptive COMPUTING
Bright Computing
IBM PLATFORM COMPUTING

## Development

### Programming Languages

C/C++
Fortran
OpenACC
python

### Development Tools

allinea DDT
PGI
VAMPIR

### Software Solutions

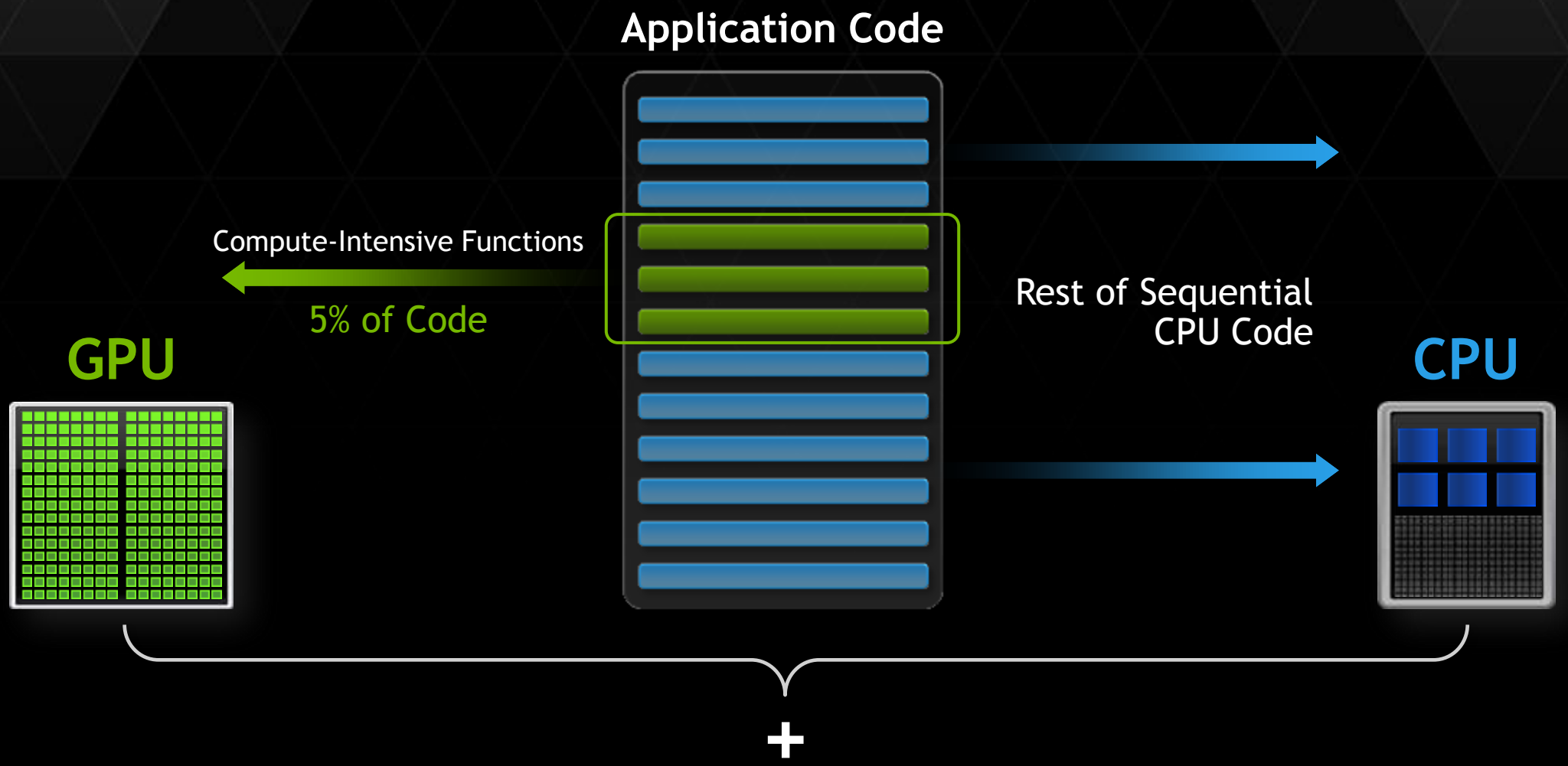
Kitware
MATLAB
ROGUE WAVE SOFTWARE

| GPU Accelerators | Interconnect | System Management | Compiler Solutions | Profile and Debug | Accelerated Libraries |
|---|---|---|---|---|---|
| *GPU Boost* | *GPU Direct NVLink* | *NVML* | *LLVM* | *CUPTI* | *cuBLAS* |

Enterprise Services Support & Maintenance

# TESLA PLATFORM FOR DEVELOPERS

# HOW GPU ACCELERATION WORKS



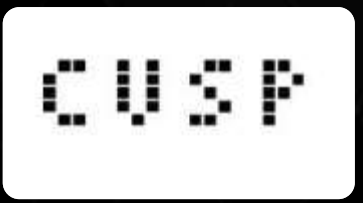
# 3 WAYS TO PROGRAM GPUS

# GPU ACCELERATED LIBRARIES

## "Drop-in" Acceleration for Your Applications

### Linear Algebra
FFT, BLAS,
SPARSE, Matrix

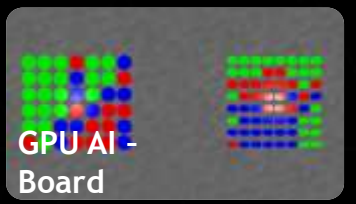



### Numerical & Math
RAND, Statistics




### Data Struct. & AI
Sort, Scan, Zero Sum

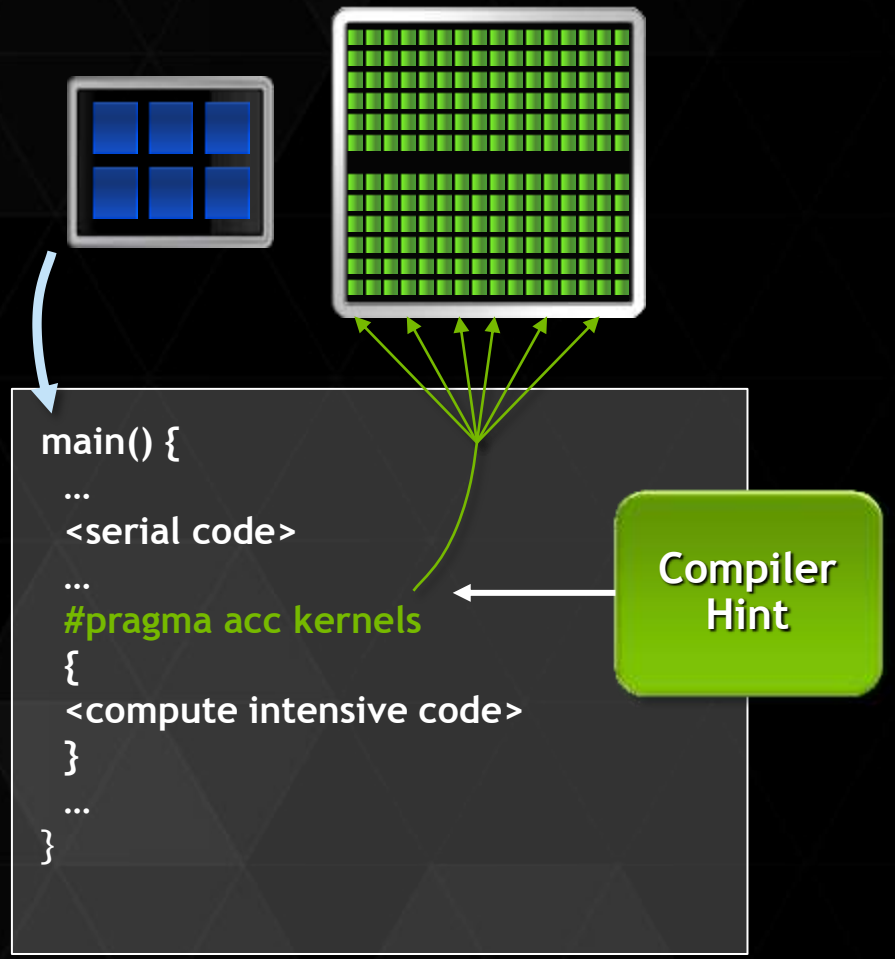





### Visual Processing
Image & Video



# OPENACC: OPEN, SIMPLE, PORTABLE

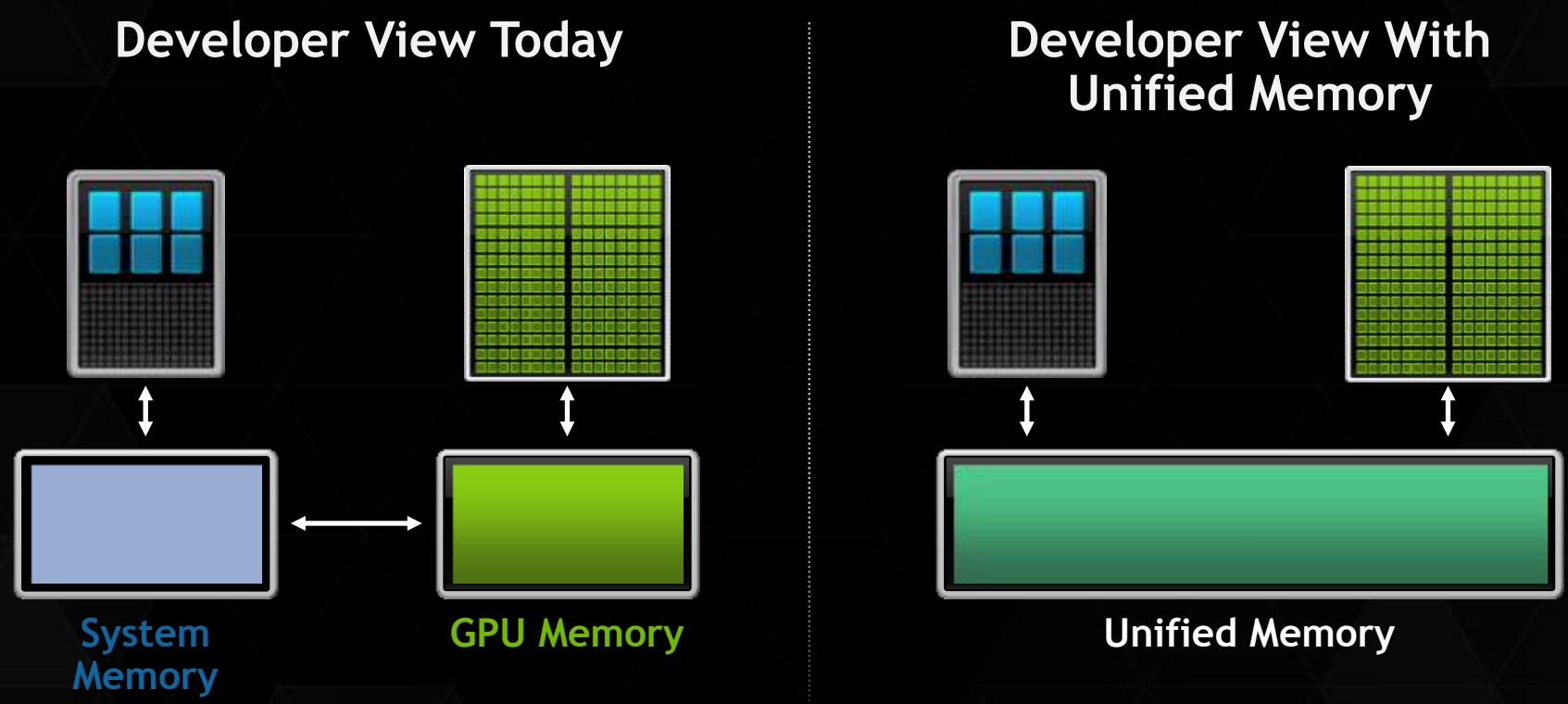



- Open Standard
- Easy, Compiler-Driven Approach
- Portable

**CAM-SE Climate**
6x Faster on GPU
Top Kernel: 50% of Runtime

# CUDA 6: UNIFIED MEMORY

## Dramatically Lower Developer Effort



# SUPER SIMPLIFIED MEMORY MANAGEMENT CODE

## CPU Code

```
void sortfile(FILE *fp, int N) {
  char *data;
  data = (char *)malloc(N);

  fread(data, 1, N, fp);

  qsort(data, N, 1, compare);


  use_data(data);

  free(data);
}
```

## CUDA 6 Code with Unified Memory

```
void sortfile(FILE *fp, int N) {
  char *data;
  cudaMallocManaged(&data, N);

  fread(data, 1, N, fp);

  qsort<<<...>>>(data,N,1,compare);
  cudaDeviceSynchronize();

  use_data(data);

  cudaFree(data);
}
```

# COMMON PROGRAMMING MODELS ACROSS MULTIPLE CPUS

**Libraries**


AmgX

cuDNN


cuBLAS

OpenCV

Thrust

**Compiler Directives**

OpenACC

**Programming Languages**

C/C++ Fortran python Java

# GPU DEVELOPER ECO-SYSTEM

**Numerical Packages**
MATLAB
Mathematica
NI LabView
pyCUDA

**Debuggers & Profilers**
cuda-gdb
NV Visual Profiler
Parallel Nsight
Visual Studio
Allinea
TotalView

**GPU Compilers**
C
C++
Fortran
Java
Python

**Auto-parallelizing & Cluster Tools**
OpenACC
mCUDA
OpenMP
Ocelot

**Libraries**
BLAS
FFT
LAPACK
NPP
Video
Imaging
GPULib

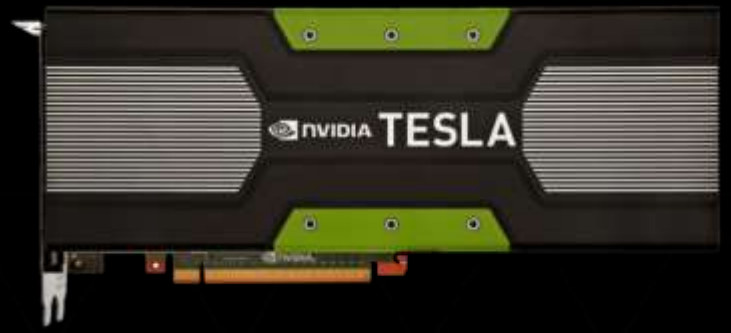
**Consultants & Training**
acceleware
GASS
STONE RIDGE TECHNOLOGY
ANEO
GPU Tech
TECH-X
ArrayFire
HPC
EM Photonics
SCALABLE GRAPHICS
WIPRO

**OEM Solution Providers**
DELL
hp
IBM
CRAY
ASUS
SUPERMICRO
sgi
FUJITSU
BULL
APPRO
lenovo 联想
NEC

# DEVELOP ON GEFORCE, DEPLOY ON TESLA

**GeForce GPUs**


**Tesla K40/K80**




## Designed for Gamers & Developers

Available Everywhere

https://developer.nvidia.com/cuda-gpus

## Designed for Cluster Deployment

ECC
24x7 Runtime
GPU Monitoring
Cluster Management
GPUDirect-RDMA
Hyper-Q for MPI
3 Year Warranty
Integrated OEM Systems, Professional Support

# CUDA: WORLD'S MOST PERVASIVE PARALLEL PROGRAMMING MODEL

**14,000** Institutions with CUDA Developers

**2,000,000** CUDA Downloads

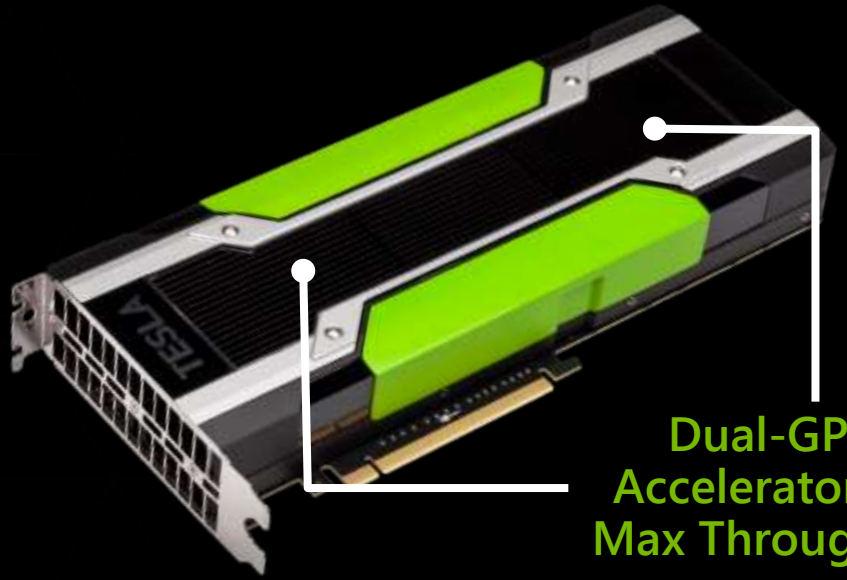
**487,000,000** CUDA GPUs Shipped

**700+ University Courses In 62 Countries**

# ACCELERATED COMPUTING ROADMAP

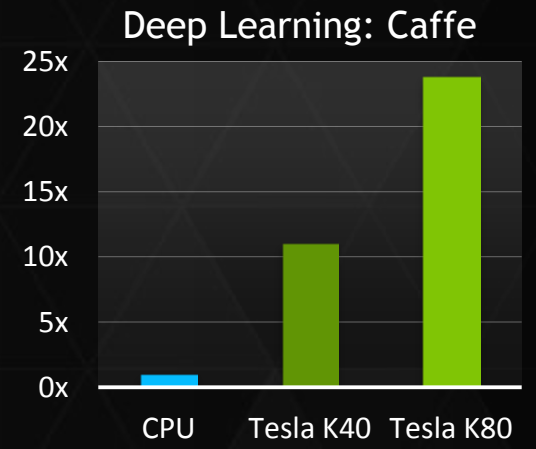
# TESLA K80

## WORLD'S FASTEST ACCELERATOR FOR DATA ANALYTICS AND SCIENTIFIC COMPUTING



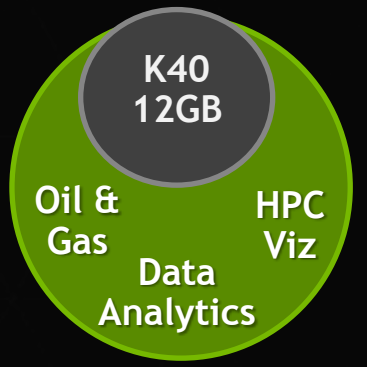

### 2x Faster

2.9 TF| 4992 Cores | 480 GB/s




### Double the Memory

Designed for Big Data Apps

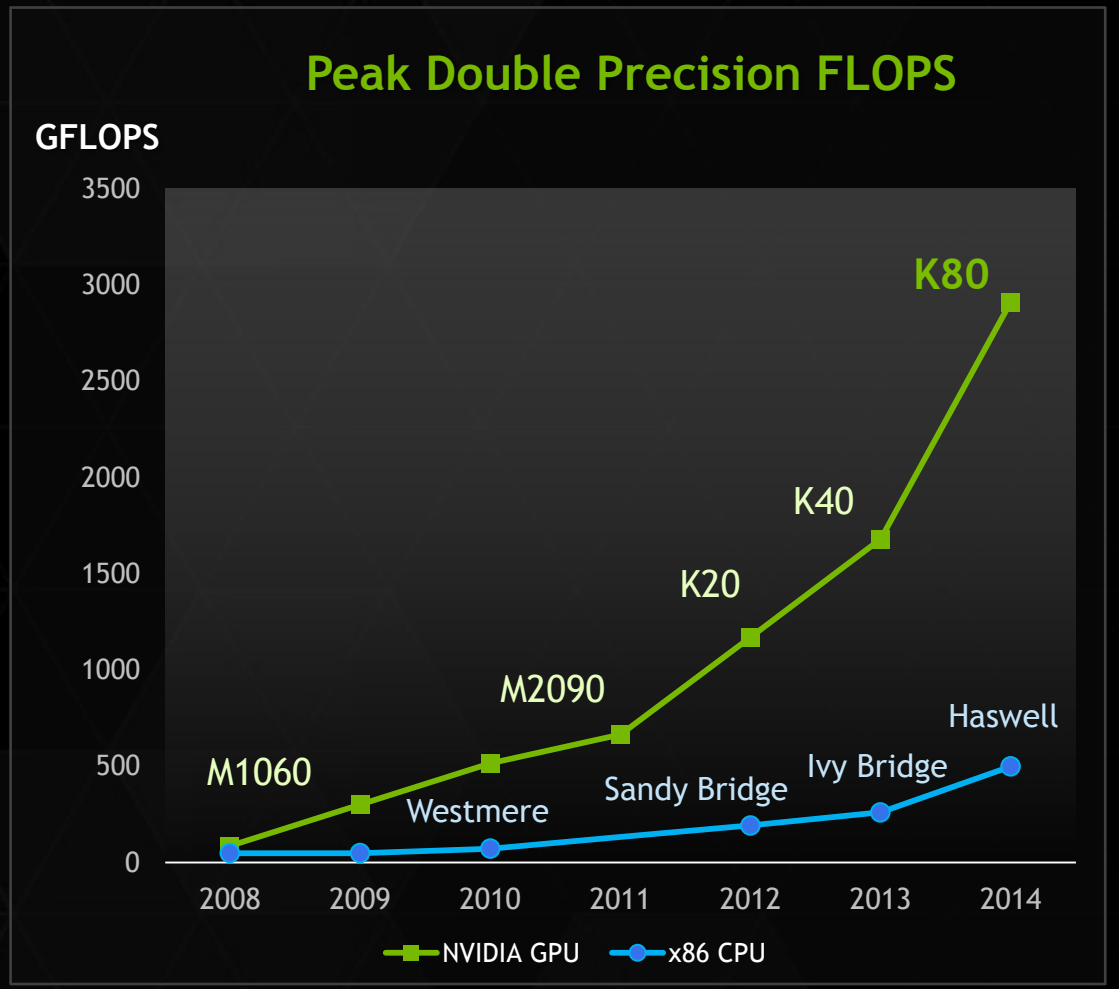


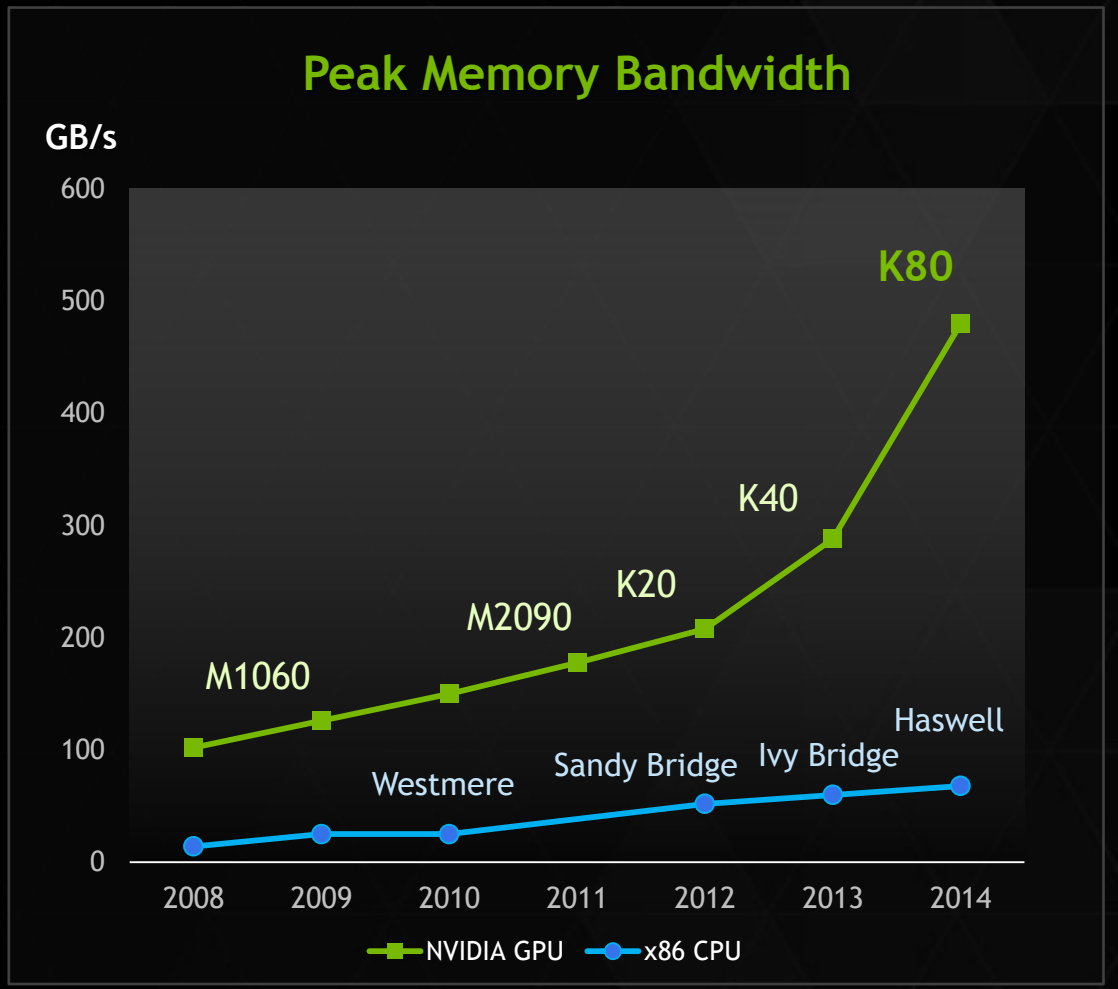

### Maximum Performance

Dynamically Maximize Perf for Every Application

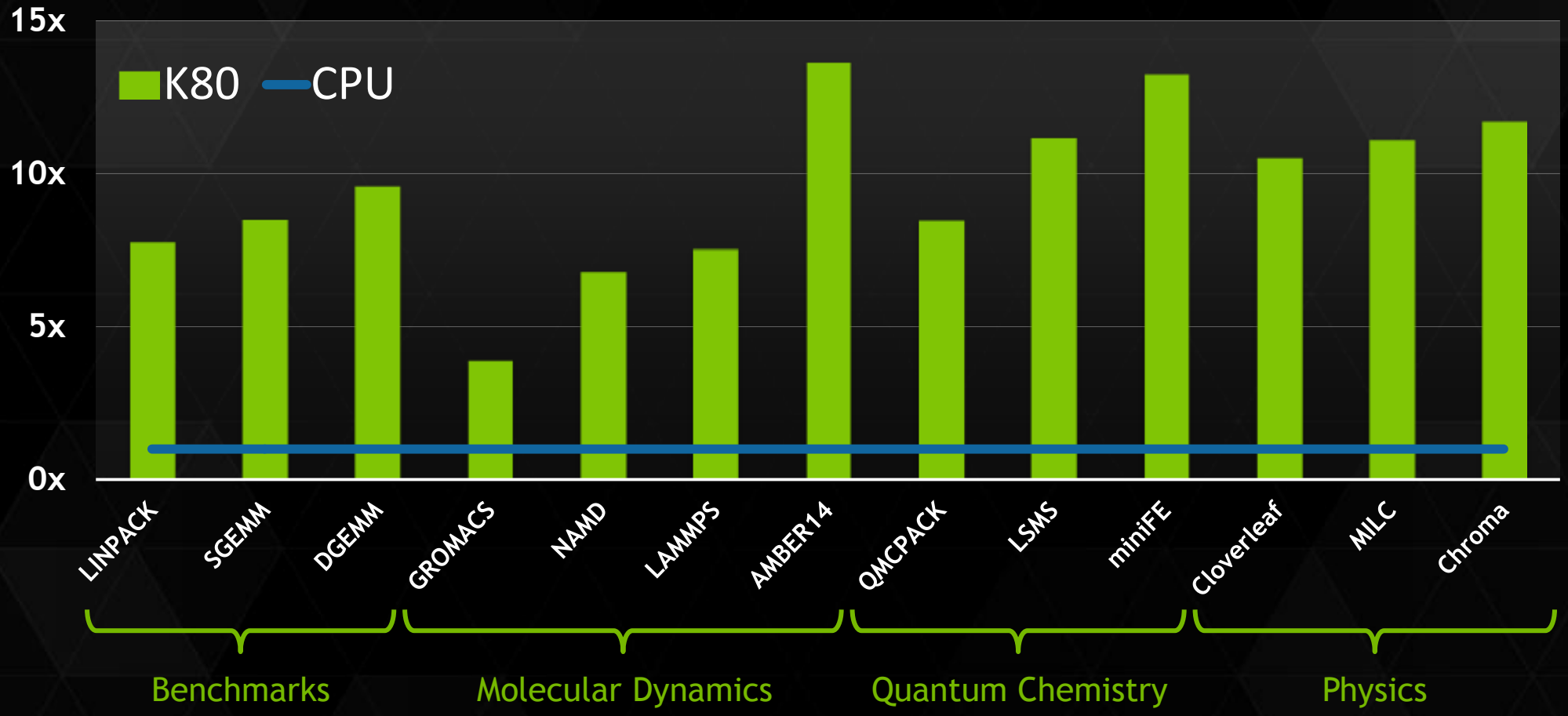

Caffe Benchmark: AlexNet training throughput based on 20 iterations, CPU: E5-2697v2 @ 2.70GHz. 64GB System Memory, CentOS 6.2

# PERFORMANCE LEAD CONTINUES TO GROW



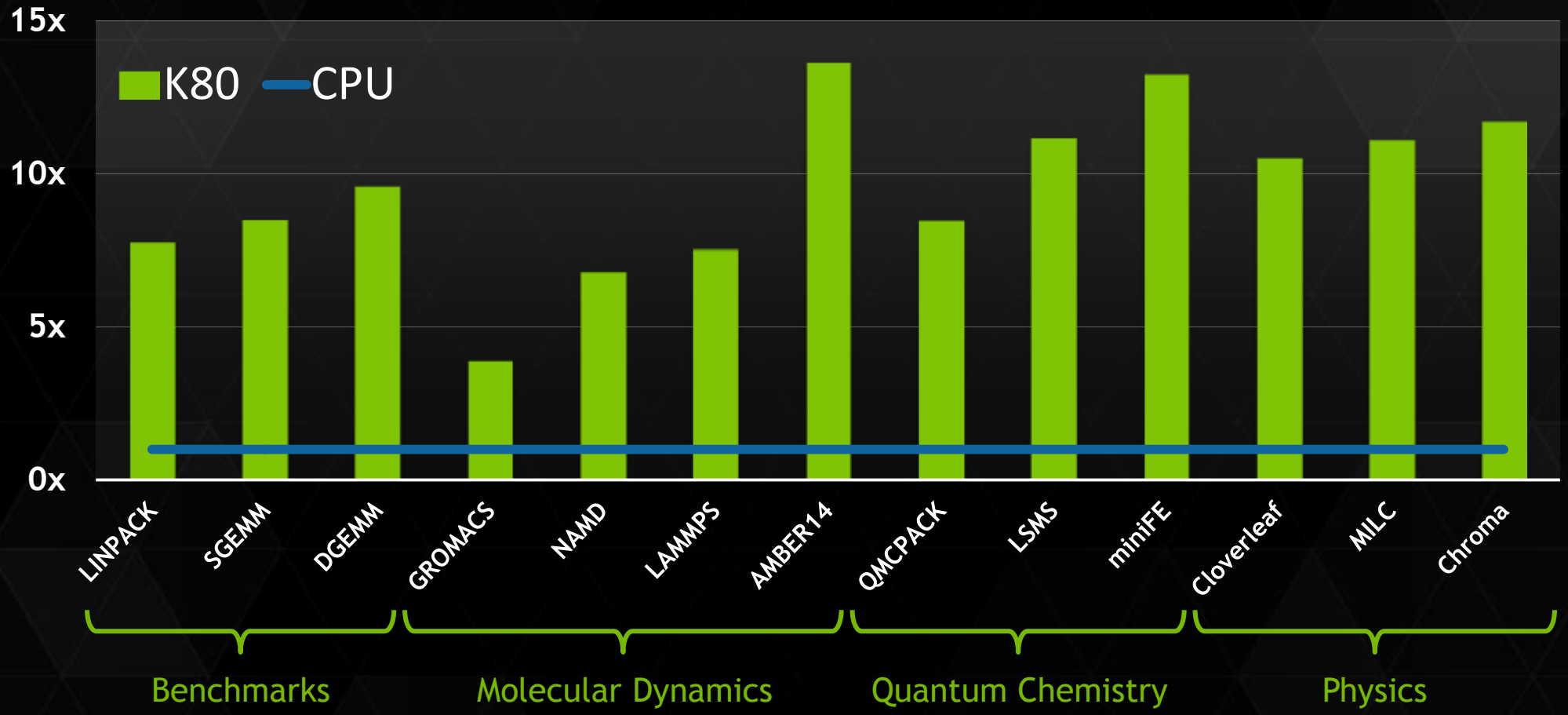



# TESLA K80: 10X FASTER ON SCIENTIFIC APPS



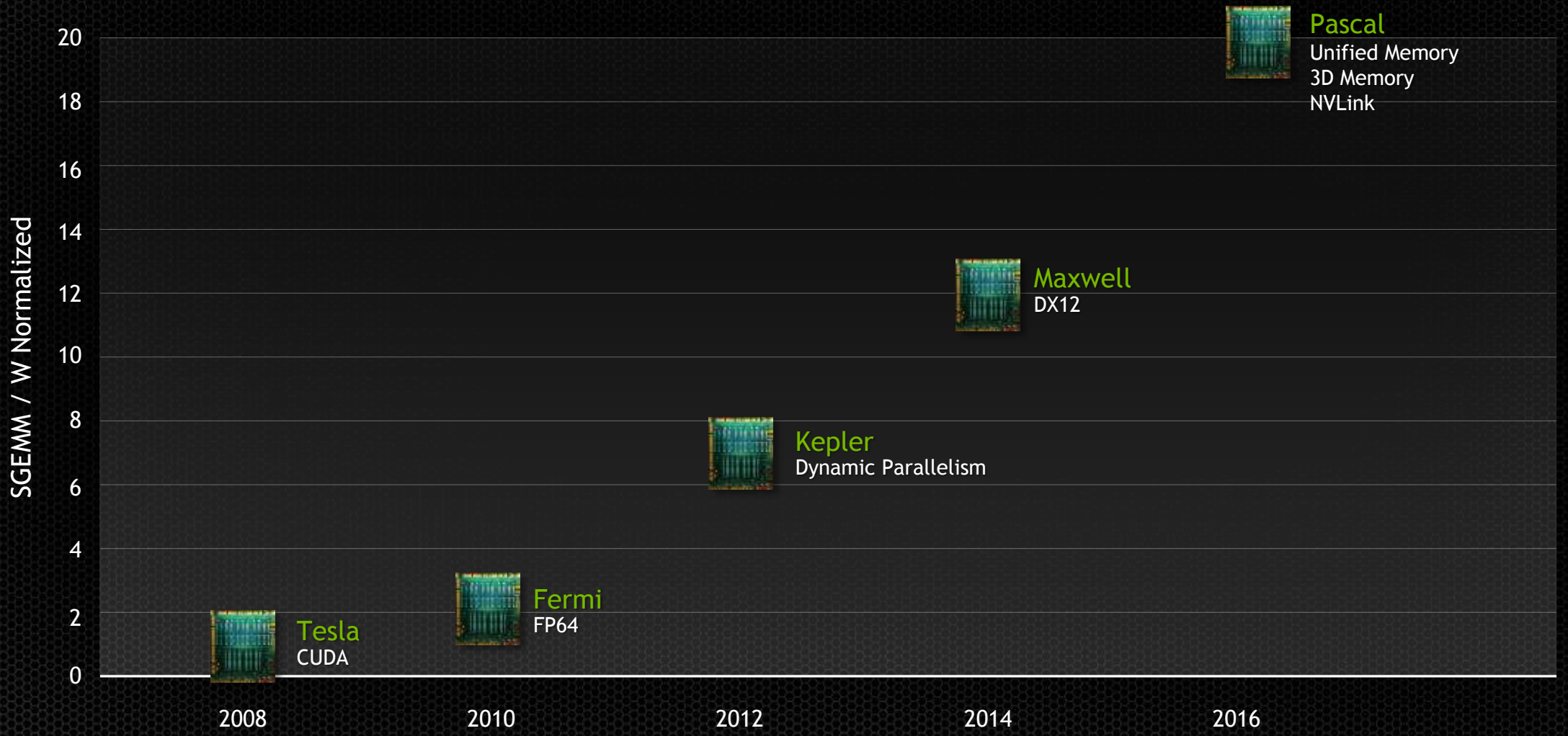

CPU: 12 cores, E5-2697v2 @ 2.70GHz. 64GB System Memory, CentOS 6.2
GPU: Single Tesla K80, Boost enabled

# TESLA K80: 10X FASTER ON REAL-WORLD APPS

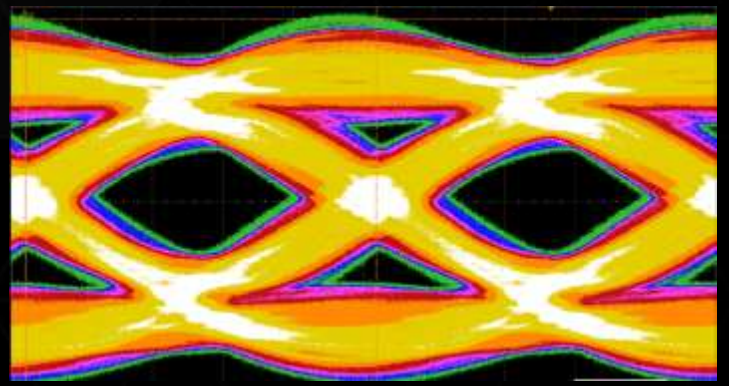



CPU: 12 cores, E5-2697v2 @ 2.70GHz. 64GB System Memory, CentOS 6.2
GPU: Single Tesla K80, Boost enabled

# GPU ROADMAP

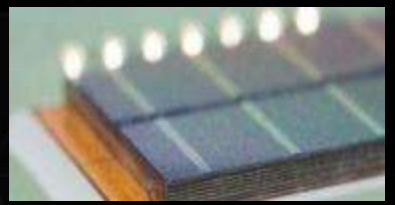

SGEMM / W Normalized
20
18
16
14
12
10
8
6
4
2
0
2008
2010
2012
2014
2016
Tesla
CUDA
Fermi
FP64
Kepler
Dynamic Parallelism
Maxwell
DX12
Pascal
Unified Memory
3D Memory
NVLink

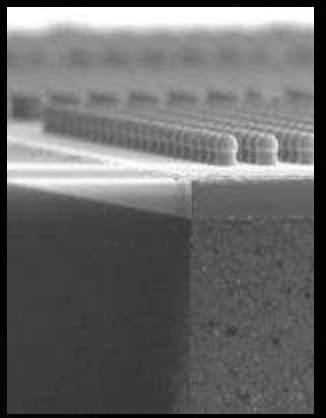
# PASCAL GPU FEATURES
# NVLINK AND STACKED MEMORY



## NVLINK

- GPU high speed interconnect
- 80-200 GB/s

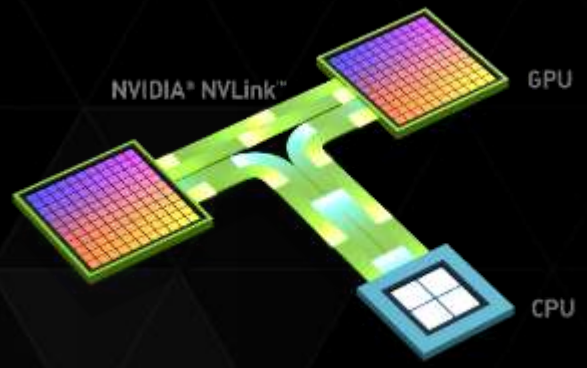


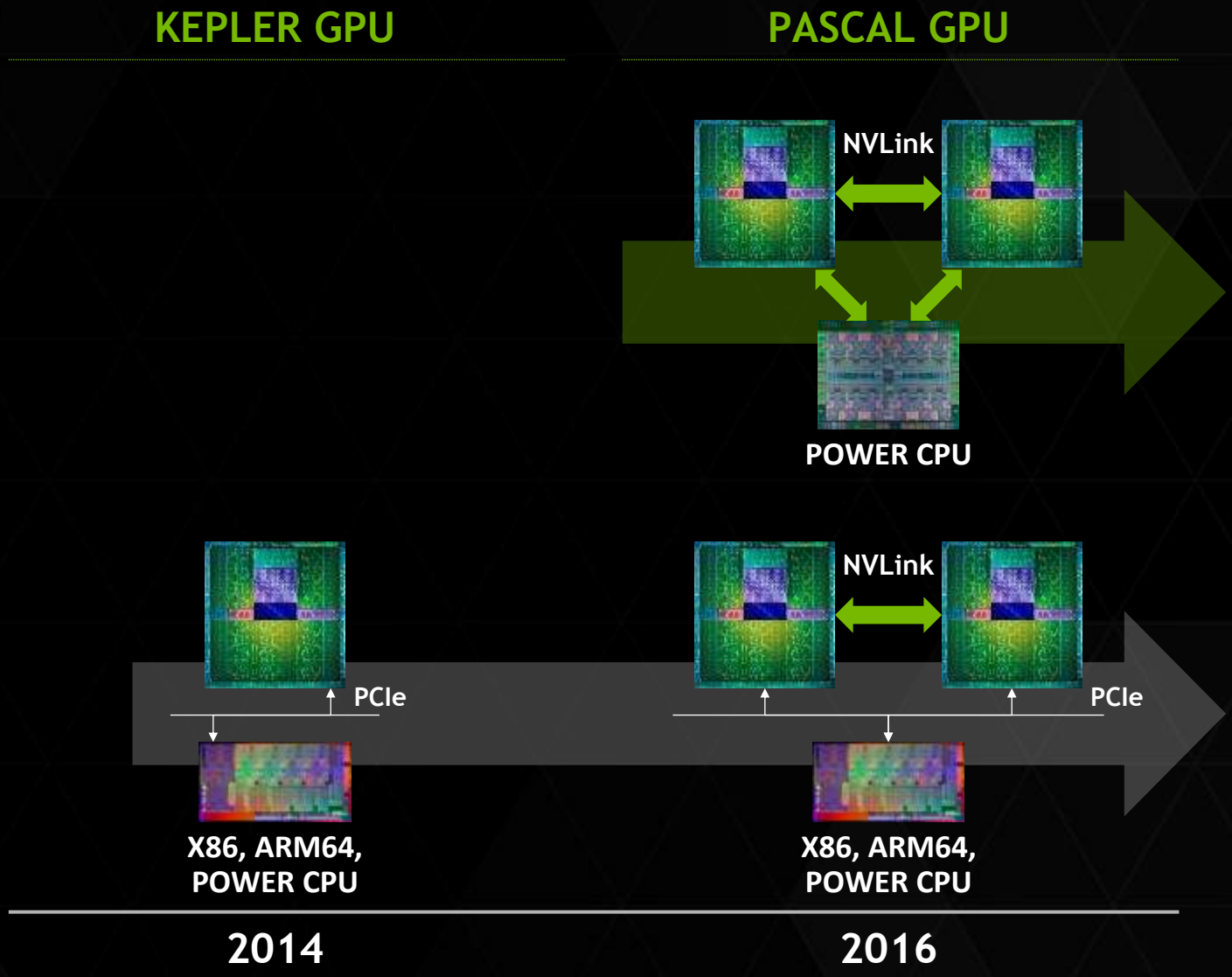


## 3D Stacked Memory

- 4x Higher Bandwidth (~1 TB/s)
- 3x Larger Capacity
- 4x More Energy Efficient per bit

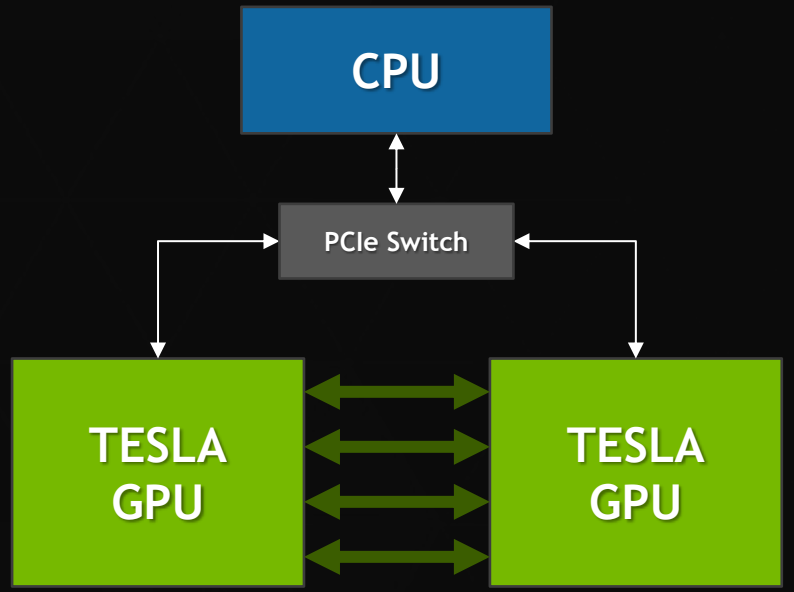
# NVLINK
## HIGH-SPEED GPU INTERCONNECT



# NVLINK UNLEASHES MULTI-GPU PERFORMANCE

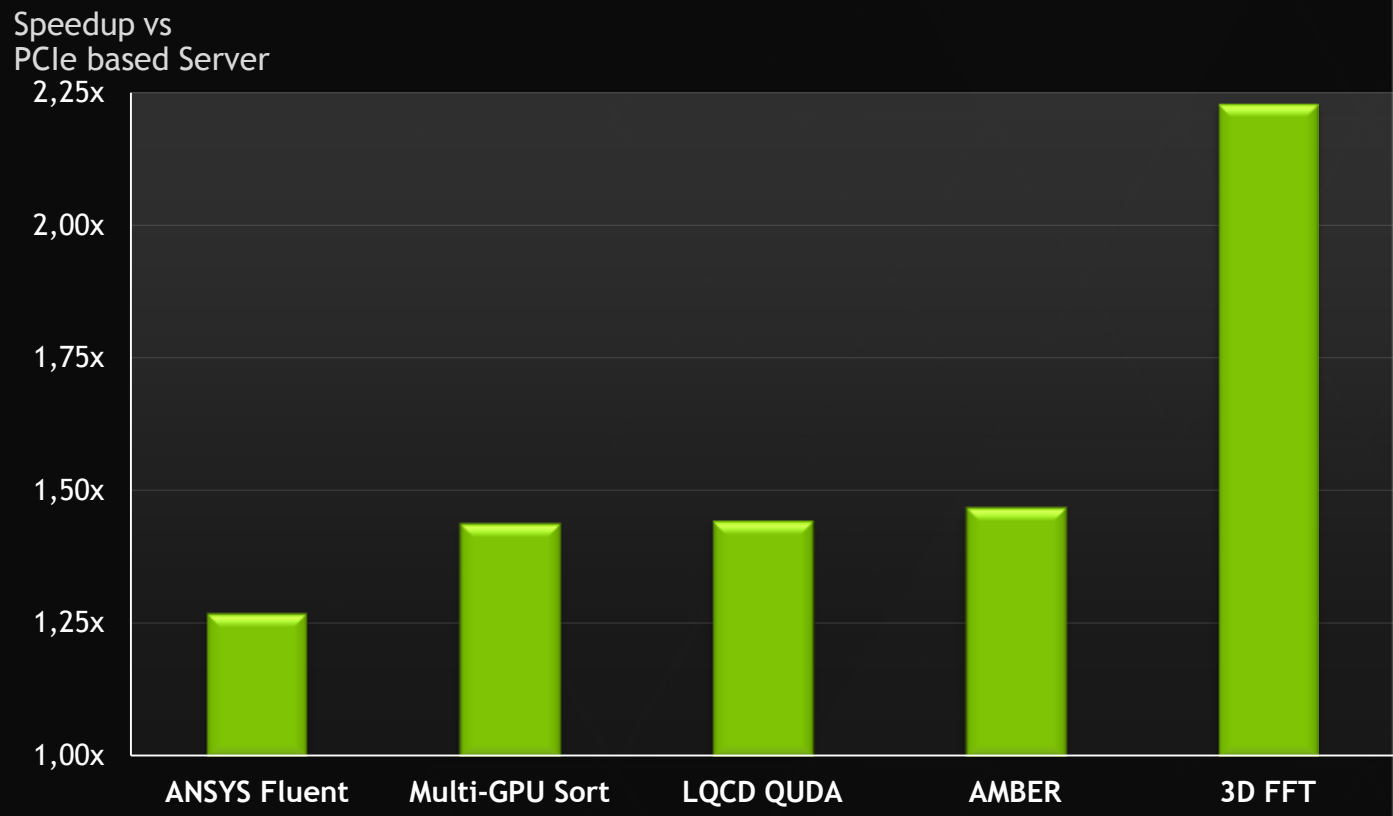
## GPUs Interconnected with NVLink




5x Faster than
PCIe Gen3 x16

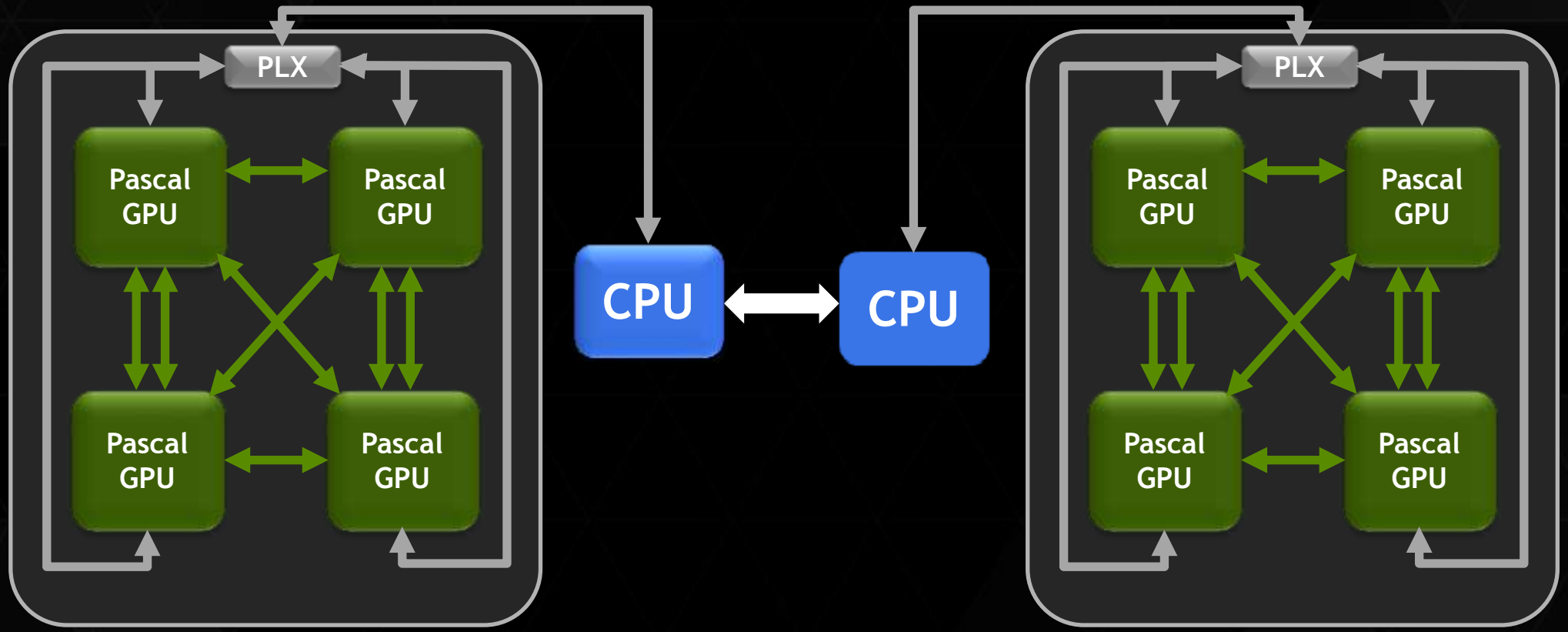
## Over 2x Application Performance Speedup

When Next-Gen GPUs Connect via NVLink Versus PCIe




3D FFT, ANSYS: 2 GPU configuration, [illegible] comparing 4 GPU configuration
AMBER Cellulose [illegible] 3D FFT problem size 1283

# EXAMPLE: 8-GPU SERVER WITH NVLINK



# US TO BUILD TWO FLAGSHIP SUPERCOMPUTERS
## POWERED BY THE TESLA PLATFORM


OAK RIDGE National Laboratory

Lawrence Livermore National Laboratory

IBM

NVIDIA

100-300 PFLOPS Peak

10x in Scientific App Performance

IBM POWER9 CPU + NVIDIA Volta GPU

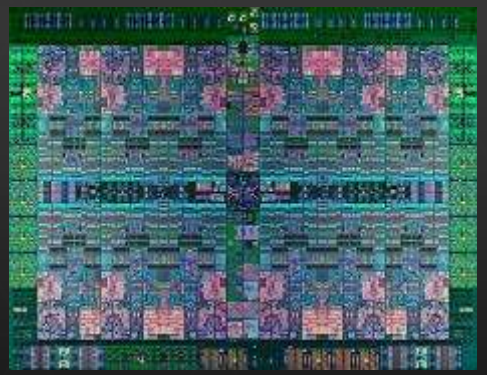
NVLink High Speed Interconnect

40 TFLOPS per Node, >3,400 Nodes

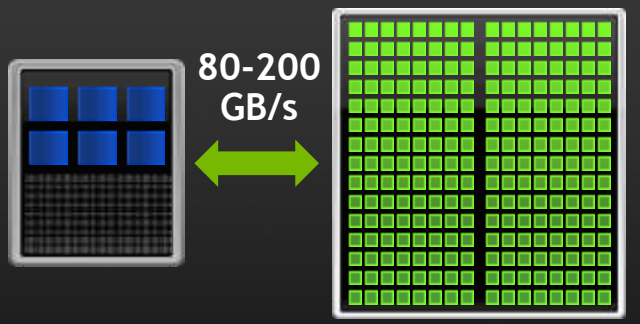
2017

Major Step Forward on the Path to Exascale

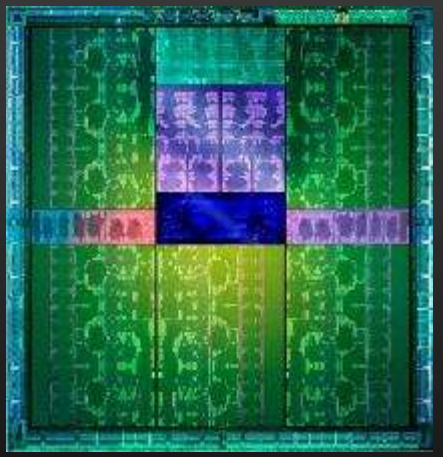
# ACCELERATED COMPUTING
# 5X HIGHER ENERGY EFFICIENCY



**IBM POWER CPU**
Most Powerful Serial Processor

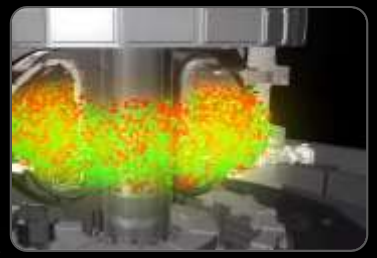



**NVIDIA NVLink**
Fastest CPU-GPU Interconnect



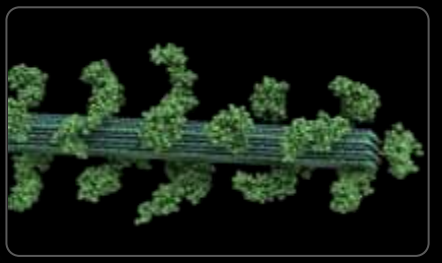
**NVIDIA Volta GPU**
Most Powerful Parallel Processor

# CORAL: BUILT FOR GRAND SCIENTIFIC CHALLENGES



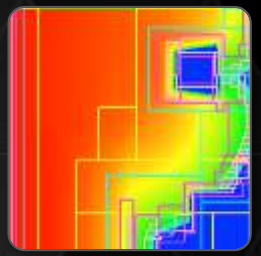
**Fusion Energy**
Role of material disorder, statistics, and fluctuations in nanoscale materials and systems.

**Climate Change**
Study climate change adaptation and mitigation scenarios; realistically represent detailed features

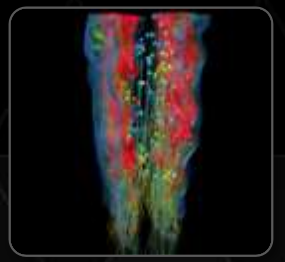


**Biofuels**
Search for renewable and more efficient energy sources

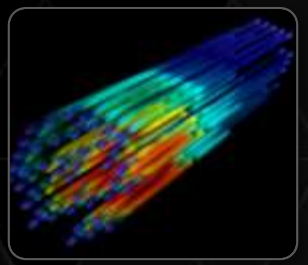
**Astrophysics**
Radiation transport – critical to astrophysics, laser fusion, atmospheric dynamics, and medical imaging



**Combustion**
Combustion simulations to enable the next gen diesel/bio-fuels to burn more efficiently



**Nuclear Energy**
Unprecedented high-fidelity radiation transport calculations for nuclear energy applications

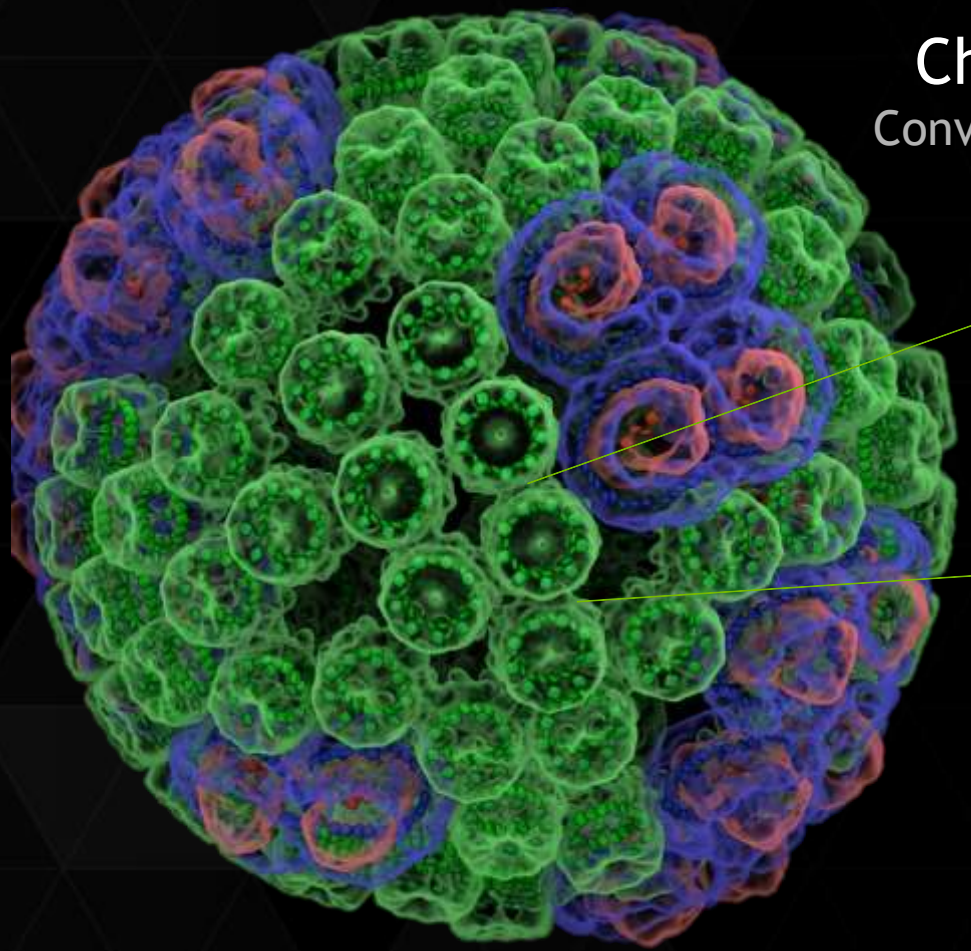


# IN-SITU VISUALIZATION

# Enabling Visualization with the Tesla Platform

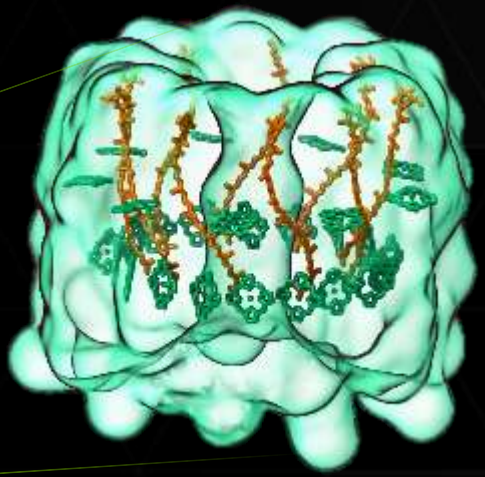
VMD

Theoretical and Computational Biophysics Group
University of Illinois at Urbana-Champaign



## Chromatophore

Converts Light to Energy

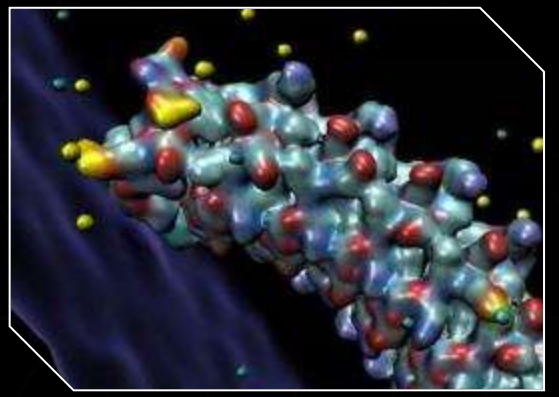


Kitware

VTK

ParaView
Parallel Visualization Application

# WORLD'S LARGEST IN-SITU HPC VISUALIZATION




2048 GPU Nodes on CSCS Piz Daint

Galaxy formation and
Molecular Dynamics

Simulation + Visualization

# VISUALIZE DATA INSTANTLY FOR FASTER SCIENCE

**Traditional**
Slower Time to Discovery

CPU Supercomputer

Simulation- 1 Week

Viz Cluster

Viz- 1 Day

Time to Discovery = Months

Multiple Iterations

**Tesla Platform**
Faster Time to Discovery

GPU-Accelerated Supercomputer

Visualize while you simulate

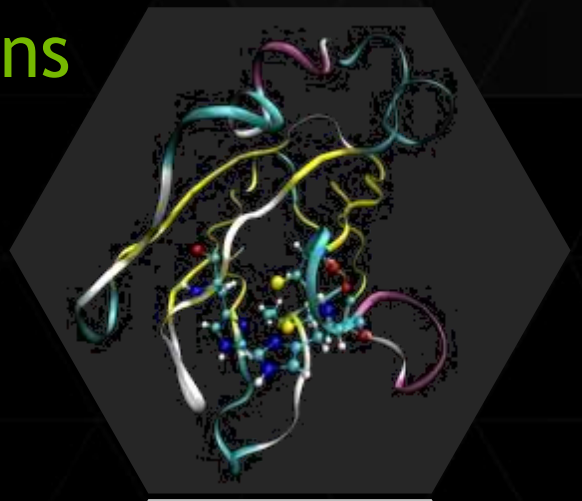
Time to Discovery = Weeks

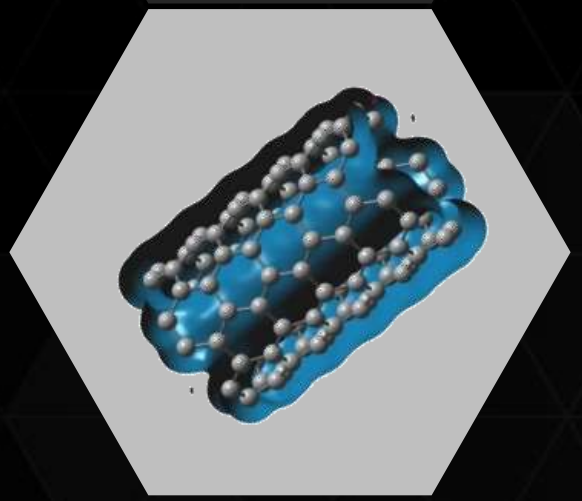
Restart Simulation Instantly
Multiple Iterations

# APPLICATIONS & CUSTOMER SUCCESSES

# LIFE & MATERIAL SCIENCES

## Overview of Accelerated Applications





- MD: All key codes are GPU-accelerated
  - ACEMD*, AMBER (PMEMD)*, BAND, CHARMM, DESMOND, ESPResso, Folding@Home, GPUgrid.net, GROMACS, HALMD, HOOMD-Blue*, LAMMPS, Lattice Microbes, mdcore, NAMD, OpenMM, SOP-GPU
  - Great multi-GPU performance!
  - Focus: on dense (up to 16) GPU nodes & large # of GPU nodes
- QC: All key codes are ported or optimizing:
  - GPU-accelerated and available today:
    - ABINIT, ACES III, ADF, BigDFT, CP2K, GAMESS, Quantum Espresso/PWscf, MOLCAS, MOPAC2012, NWChem, QUICK, Q-Chem, TeraChem*
  - Active GPU acceleration projects:
    - CASTEP, CPMD, GAMESS, Gaussian, NWChem, ONETEP, Quantum Supercharger Library, VASP & more
  - Focus: on using GPU-accelerated math libraries, OpenACC directives

green* = application where all the workload is on GPU

# REVOLUTIONIZING SCIENTIFIC COMPUTING

AMBER Molecular Dynamics Simulation
DHFR NVE Benchmark

**64 Sandy Bridge CPUs**
**58 ns/day**

**Server with 2 Tesla K80**
**220 ns/day**

# ACCELERATING SIGNAL & VIDEO ANALYTICS

Real-time HD video enhancements and analytics

Made possible only with GPUs

MotionDSP

Video surveillance with faster than real time analytics

12x faster with GPUs

intuVision
intuitive vision solutions

Unmanned submarine with accelerated sonar processing

50-100x speed up over CPU

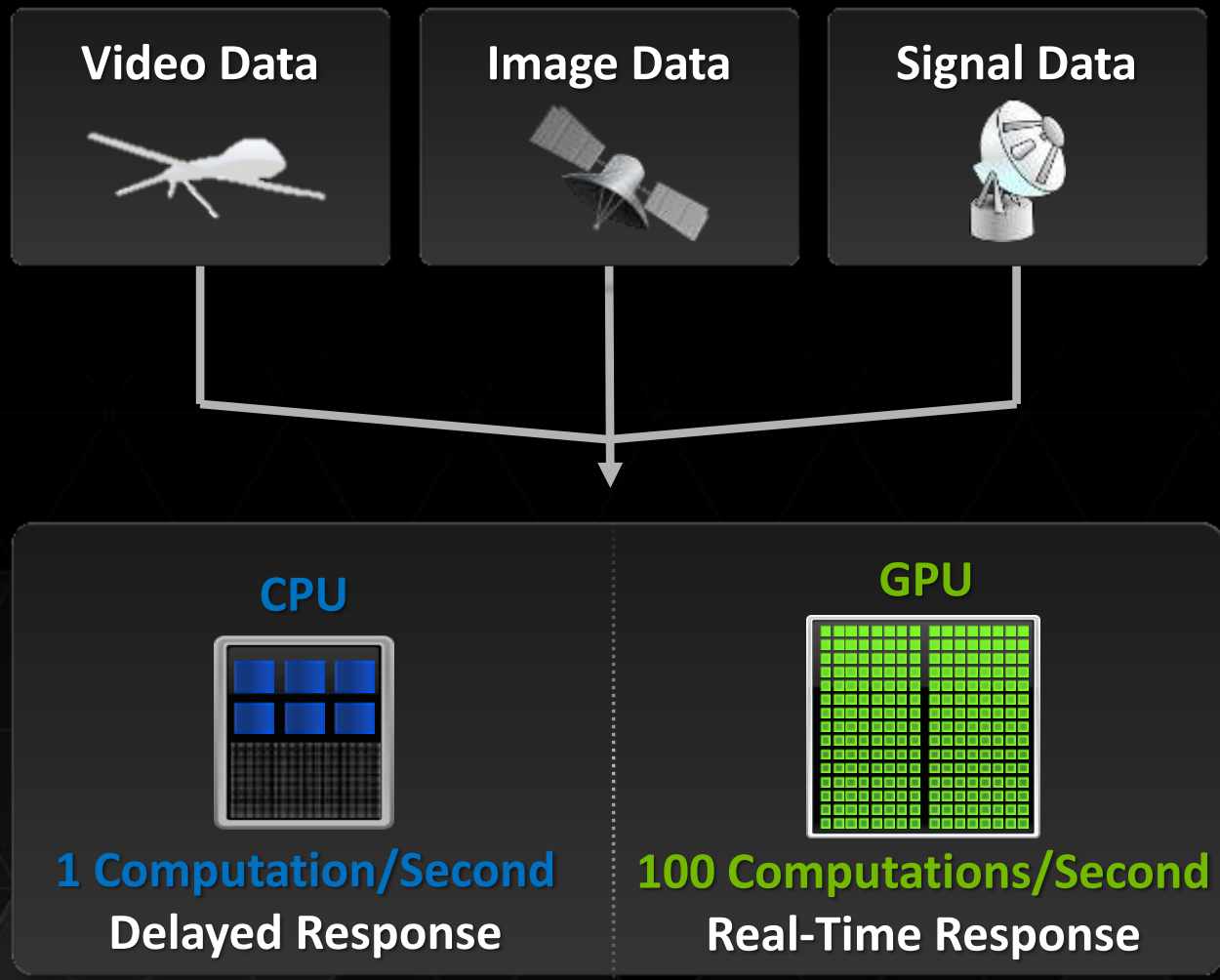
NATO OTAN

Faster satellite image processing for actionable intelligence

12x faster using GPUs

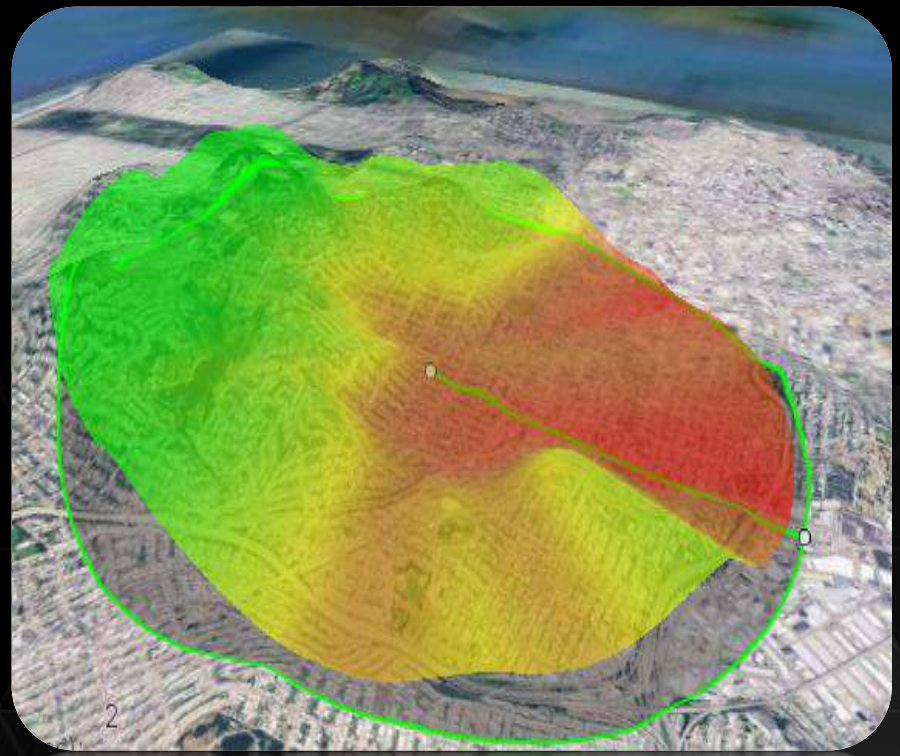
DIGITALGLOBE

# MISSION PLANNING WITH REAL-TIME LINE OF SIGHT




LUCIAD

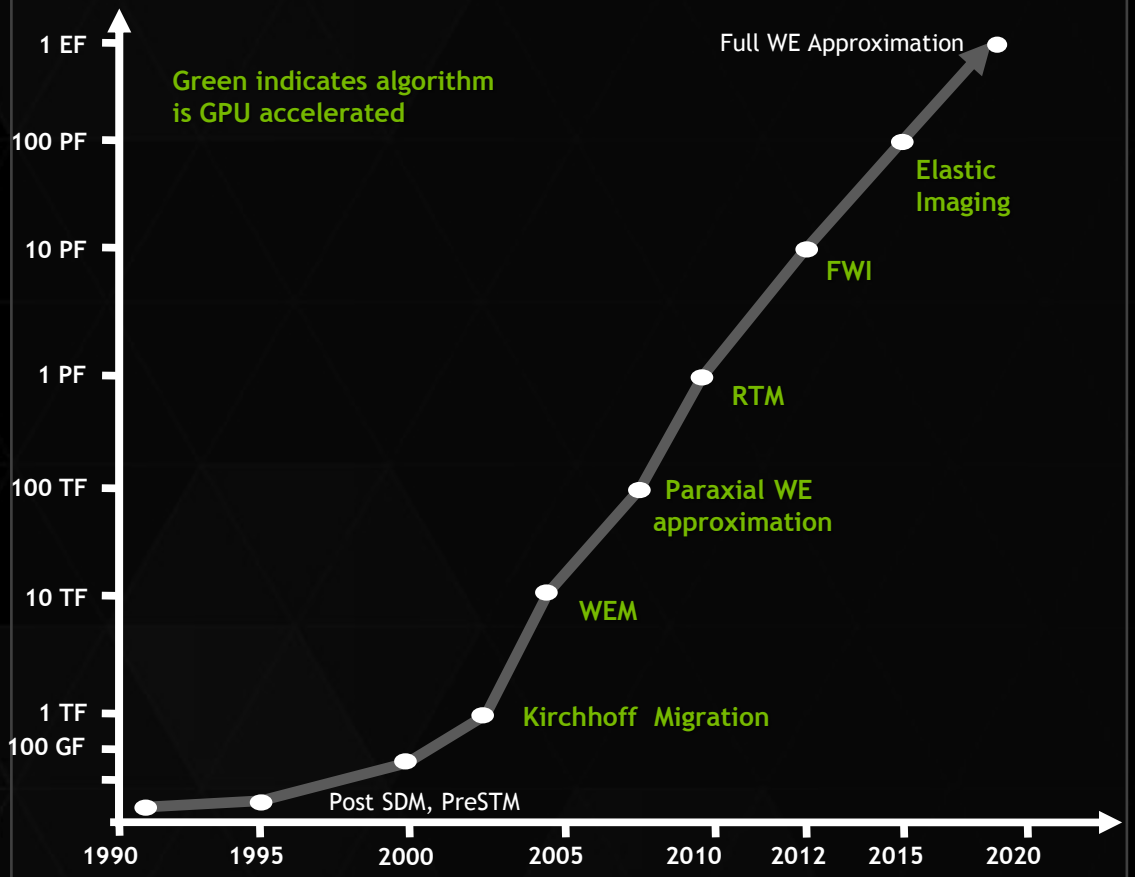
World Leader in Geospatial Situational Awareness



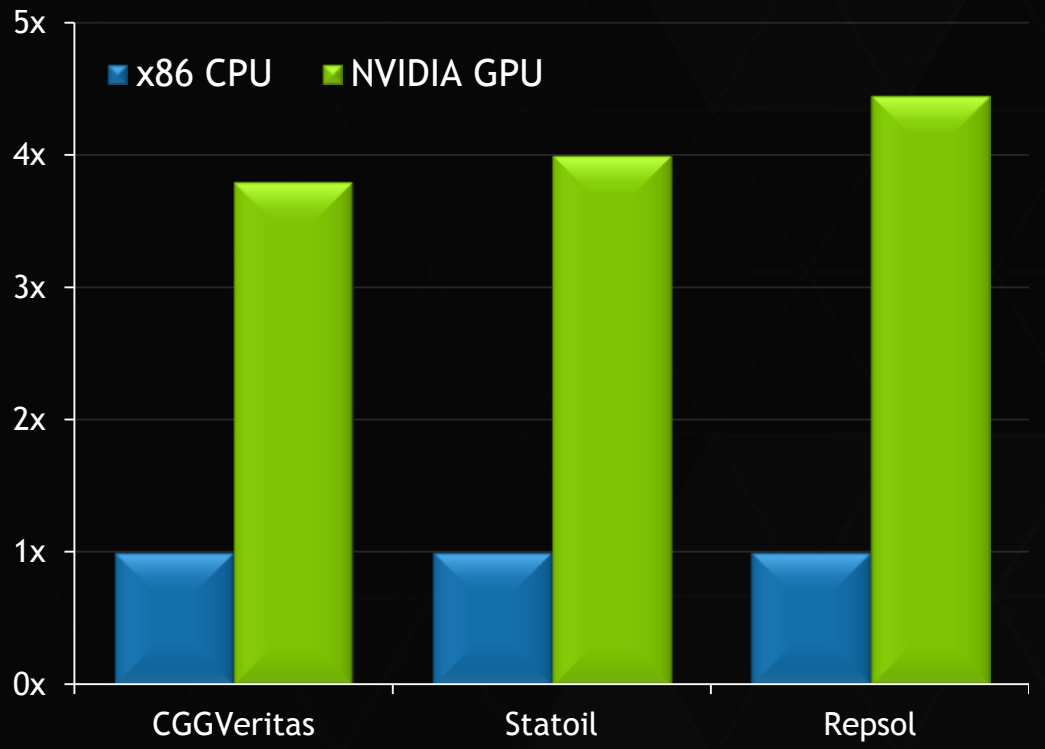
http://www.luciad.com/

# KEY O&G APPS ACCELERATED ON GPUS

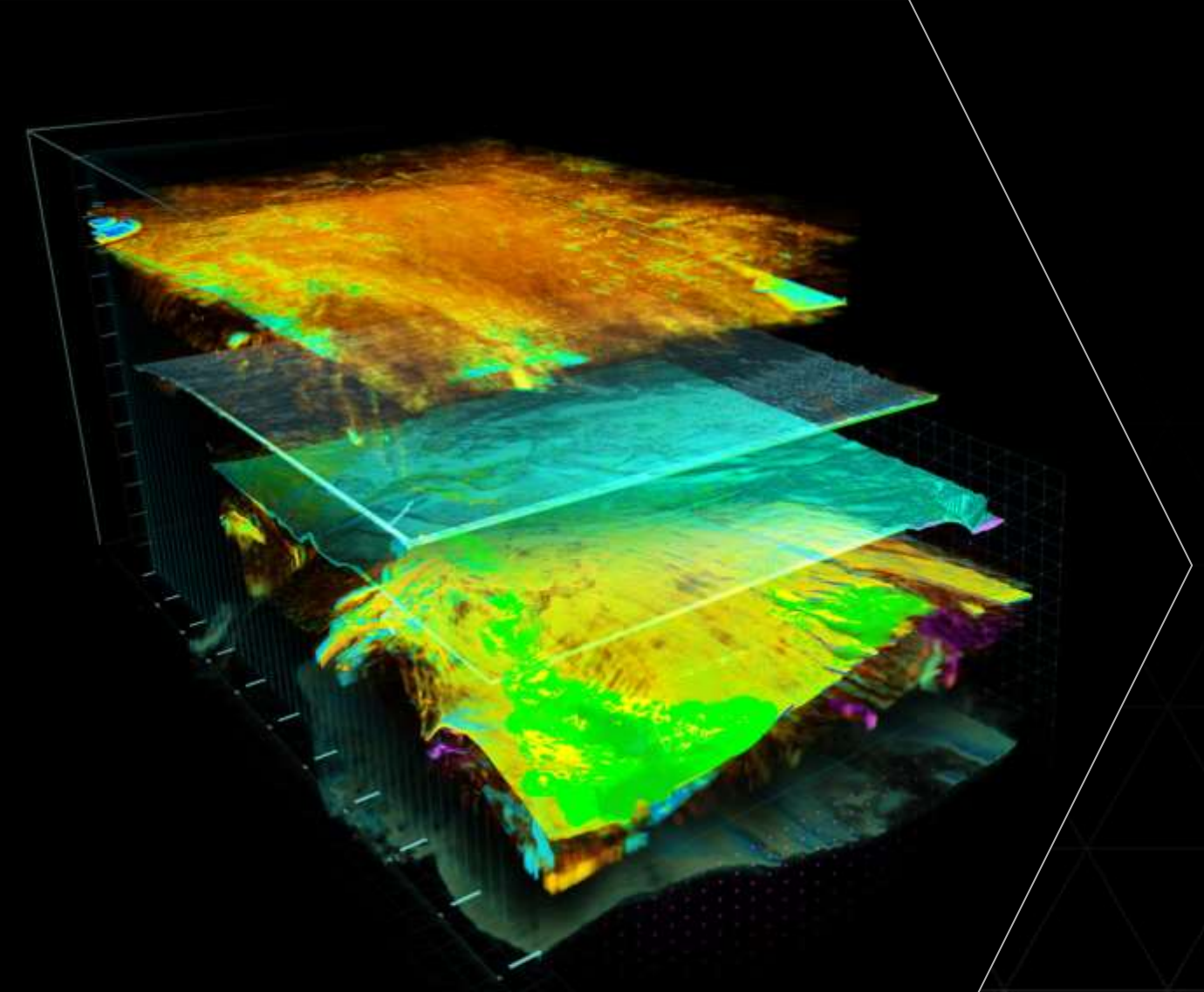
## Most Workloads Ported to GPUs




## Higher Throughput with GPUs




CGGVeritas: 2013 Rice Oil & Gas HPC Workshop
Statoil: Maximizing TTI RTM Throughput for CPU + GPU: EAGE Conference (2013)
Repsol: (Test case: TT-A) NVIDIA GTC 2014



Maximizing Opportunity for Oil Discovery with GPU-powered Supercomputer

# WORLD'S FASTEST ENTERPRISE SUPERCOMPUTER

eni

IBM

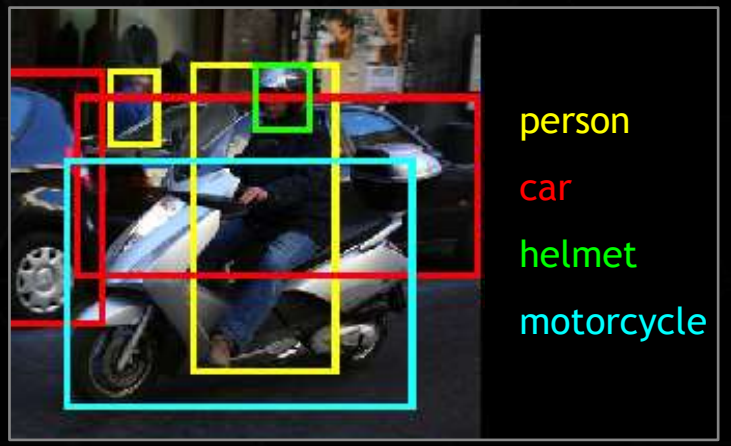
3 Petaflops Linpack Performance

Most Energy-Efficient Petascale System in the World

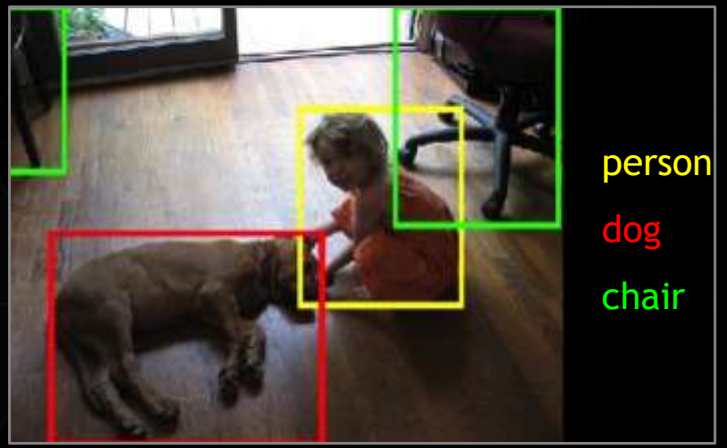
3,000 NVIDIA Tesla K20X GPU Accelerators

# DEEP LEARNING

# DEEP LEARNING FOR IMAGE ANALYTICS



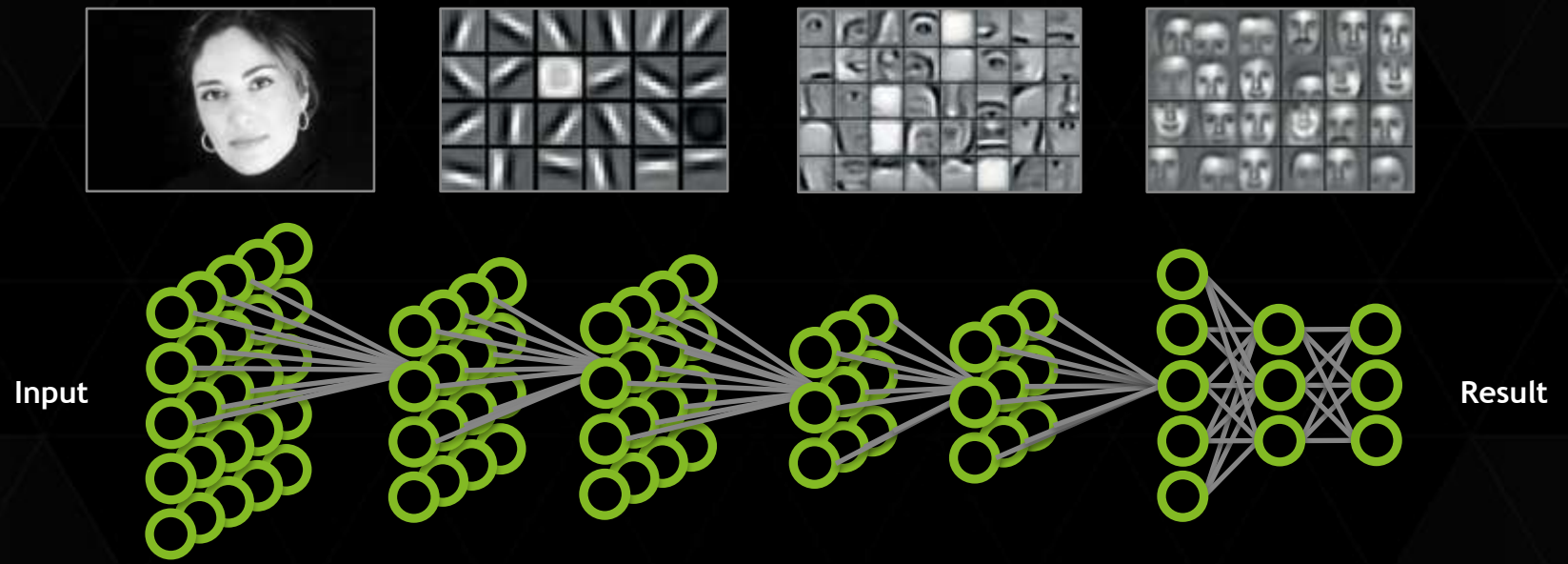






IMAGENET

# MACHINE LEARNING USING DEEP NEURAL NETWORKS



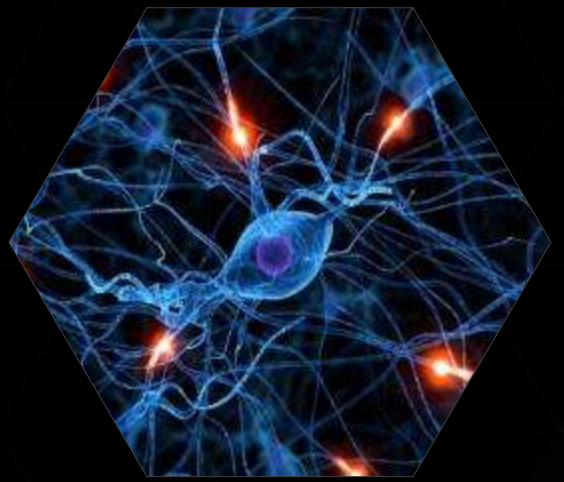

Hinton et al., 2006; Bengio et al., 2007; Bengio & LeCun, 2007; Lee et al., 2008; 2009

Visual Object Recognition Using Deep Convolutional Neural Networks
Rob Fergus (New York University / Facebook) http://on-demand-gtc.gputechconf.com/gtcnew/on-demand-gtc.php#2985

# 3 DRIVERS FOR DEEP LEARNING

## More Data


## Better Models



## Powerful GPU Accelerators

# BROAD BENEFITS OF DEEP LEARNING

Spotify

Content-based music recommendation

*Summer Intern implements recommendation system*

Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks

*Dan C. Cirescan et al.*

*Research team takes first step to bring automated mitosis detection into clinical practic*

MERCK
Be well

Merck Molecular Activity Challenge

*Winning team dominates the competition by using deep learning algorithms running on GPUs*

# BROAD USE OF GPUS IN DEEP LEARNING

## Early Adopters

Adobe
Image Analytics for Creative Cloud

Baidu
Speech/Image Recognition

flickr
Image Classification

IBM
Hadoop

NETFLIX
Recommendation

Yandex
Search Rankings

## Use Cases

- Image Detection
- Face Recognition
- Gesture Recognition
- Video Search & Analytics
- Speech Recognition & Translation
- Recommendation Engines
- Indexing & Search

## Talks @ GTC

facebook
NYU
STANFORD UNIVERSITY
DARPA
DENSO
Carnegie Mellon University
MIT Massachusetts Institute of Technology
Berkeley UNIVERSITY OF CALIFORNIA

# WHAT IS NEXT?

## Deep Learning Will Be Everywhere

**Pattern Analysis**

*Anomaly Detection*

*Behavior Prediction*

*Diagnostic Support*

*Sentiment Analysis*

....

*"Mark Zuckerberg calls it the theory of the mind. How do we model — in machines — what human users are interested in and are going to do?"*

*Yann Lecun, Director AI Research at Facebook*

*"Any product that excites you over the next five years and makes you think: 'That is magical, how did they do that?', is probably based on this [deep learning]."*

*Steve Jurvetson, Partner DFJ Venture*

# GPU TECHNOLOGY CONFERENCE

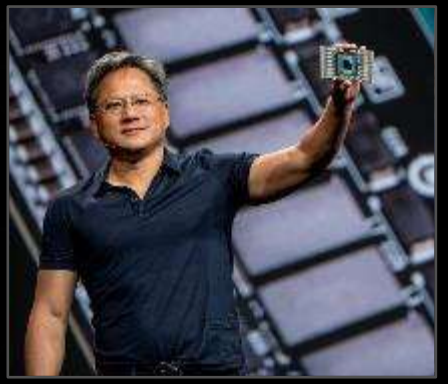
March 17-20, 2015 | Silicon Valley
www.gputechconf.com #GTC15

## CONNECT

Connect with technology experts from NVIDIA and other leading organizations


## LEARN

Gain insight and hands-on training through the hundreds of sessions and research posters



## DISCOVER

See how GPU technologies are creating amazing breakthroughs in important fields such as deep learning


## INNOVATE

Hear about disruptive innovations as early-stage startups present their work

## 2015 Theme: Deep Learning


50+
Deep Learning Sessions

www.gputechconf.com

| | |
|---|---|
| Adobe | Google |
| Alibaba | iFlytek, Ltd |
| Baidu | NUANCE |
| Carnegie Mellon | Stanford Univ |
| Facebook | UC Berkeley |
| Flickr / Yahoo | Univ of Toronto |

Developer Labs

Caffe
Torch
Theano