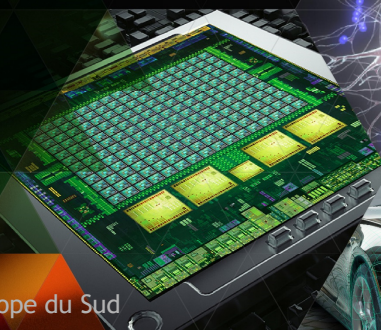




TESLA MASTER DECK

March 2014

Guillaume BARAT - gbarat@nvidia.com
Relation Education / Research & DL Europe du Sud



GAMING	DESIGN	ENTERPRISE VIRTUALIZATION	HPC & CLOUD SERVICE PROVIDERS	AUTONOMOUS MACHINES
	PC	DATA CENTER	MOBILE	

THE WORLD LEADER IN VISUAL COMPUTING



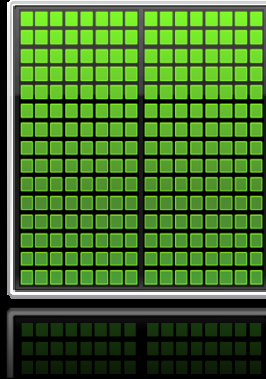
ACCELERATED COMPUTING

10X PERFORMANCE & 5X ENERGY EFFICIENCY

CPU
Optimized for
Serial Tasks



GPU Accelerator
Optimized for
Parallel Tasks



FROM HPC TO ENTERPRISE DATACENTERS

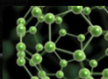


Oil & Gas

Schlumberger

BP
PETROBRAS

Eni
Chevron
Statoil

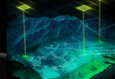


Higher Ed

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

CSGS

NCSA

Tokyo Institute
of Technology
Lawrence Livermore
National Laboratory



Finance

J.P.Morgan

BARCLAYS

STANDARD LIFE
BNP PARIBAS
MUREX



Consumer Web

Baidu 百度

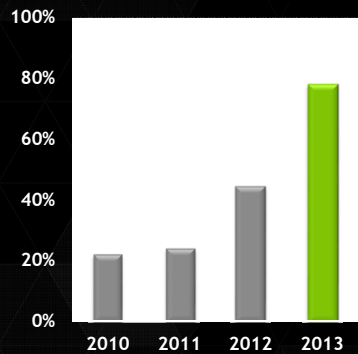
salesforce

SHAZAM
amazon.com
Yandex

RAPID ADOPTION OF ACCELERATED COMPUTING

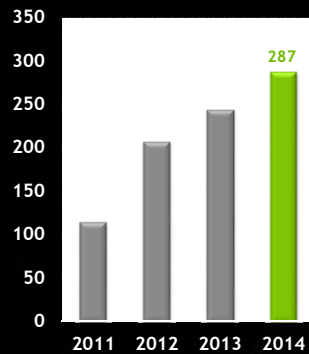
Rapid Adoption of Accelerators

% of HPC Customers with Accelerators

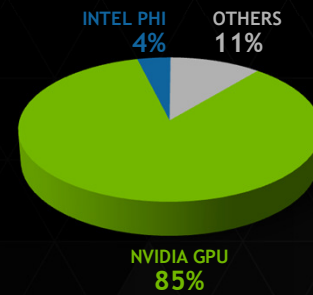


Intersect360 HPC User Site Census: Systems, July 2013
IDC HPC End-User MSC Study, 2013

Hundreds of GPU Accelerated Apps



NVIDIA GPU is Accelerator of Choice



Intersect360 Research
HPC User Site Census: Systems, July 2013 NVIDIA

HIGH GPU DENSITY SERVERS NOW MAINSTREAM



Cray CS-Storm
8 K80s per Node



Dell C4130
4 K80s per Node

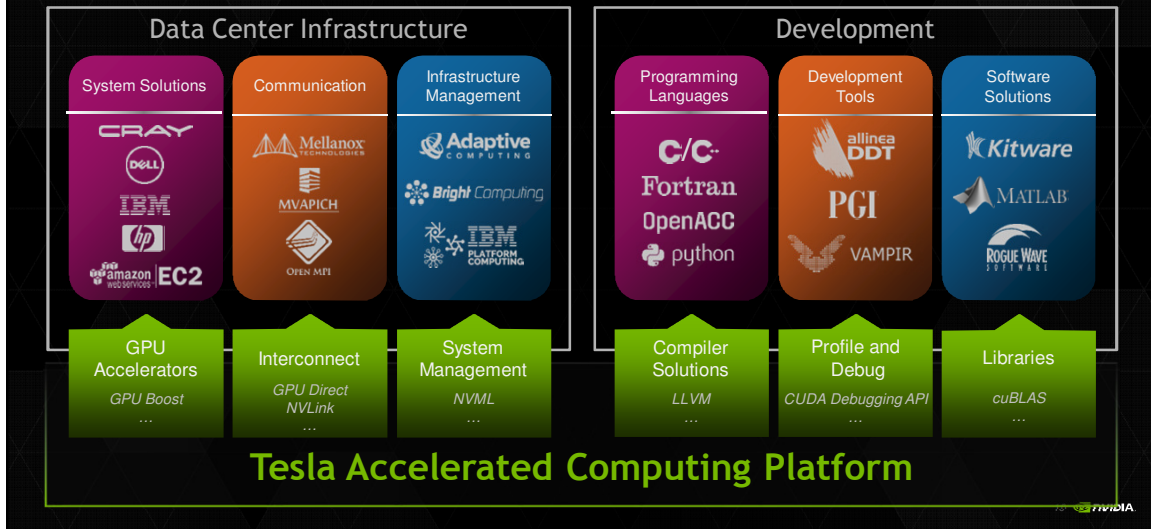


HP SL270
8 K80s per Node



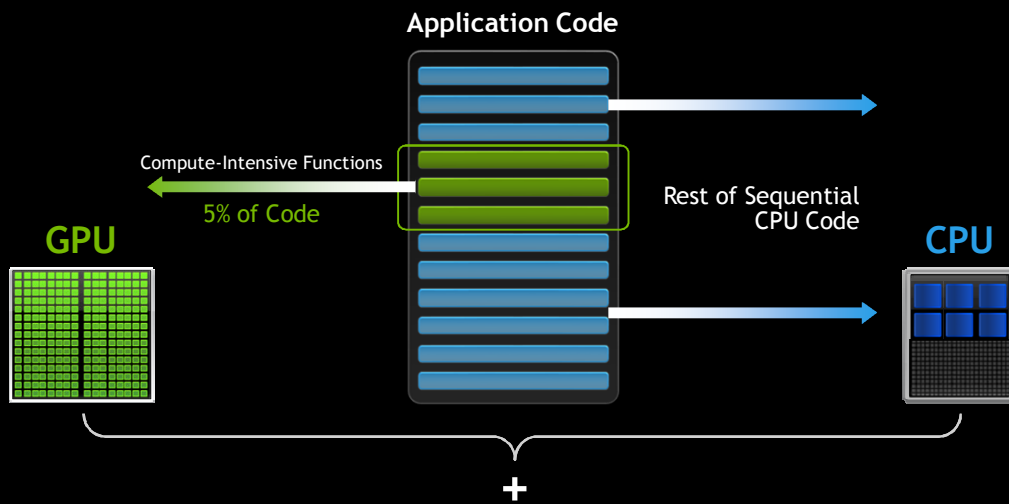
Quanta S2BV
4 K80s per Node

TESLA: PLATFORM FOR ACCELERATED DATACENTERS

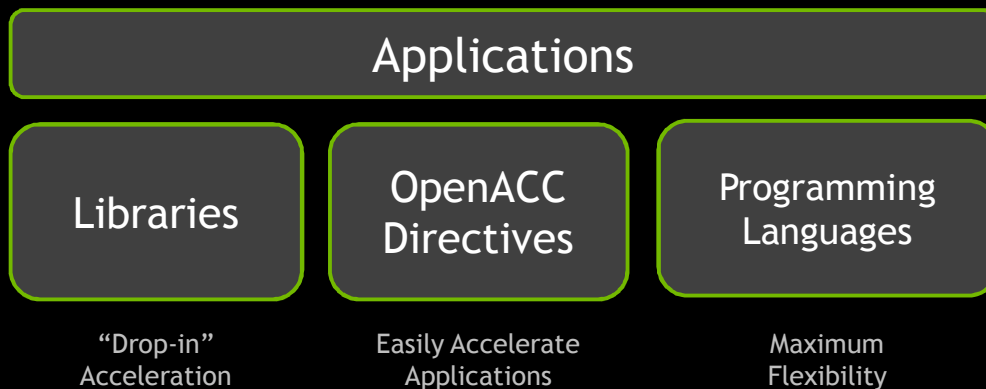


TESLA PLATFORM FOR DEVELOPERS

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



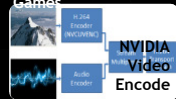
Data Struct. & AI

Sort, Scan, Zero Sum

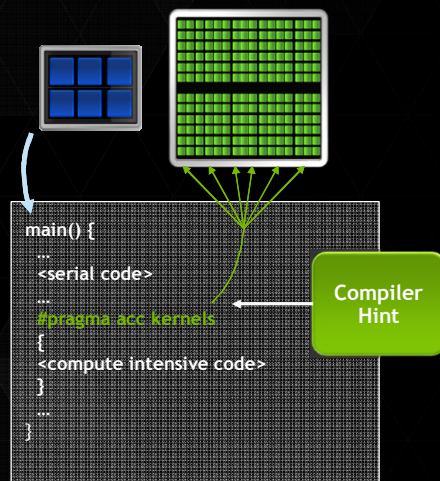


Visual Processing

Image & Video



OPENACC: OPEN, SIMPLE, PORTABLE



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

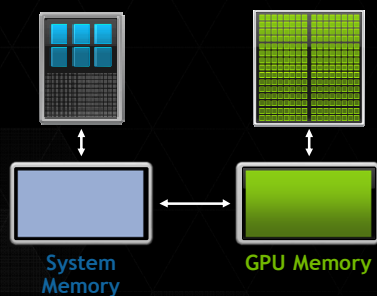
CAM-SE Climate
6x Faster on GPU
Top Kernel: 50% of Runtime



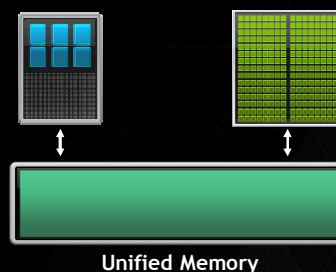
CUDA 6: UNIFIED MEMORY

Dramatically Lower Developer Effort

Developer View Today



Developer View With Unified Memory



SUPER SIMPLIFIED MEMORY MANAGEMENT CODE

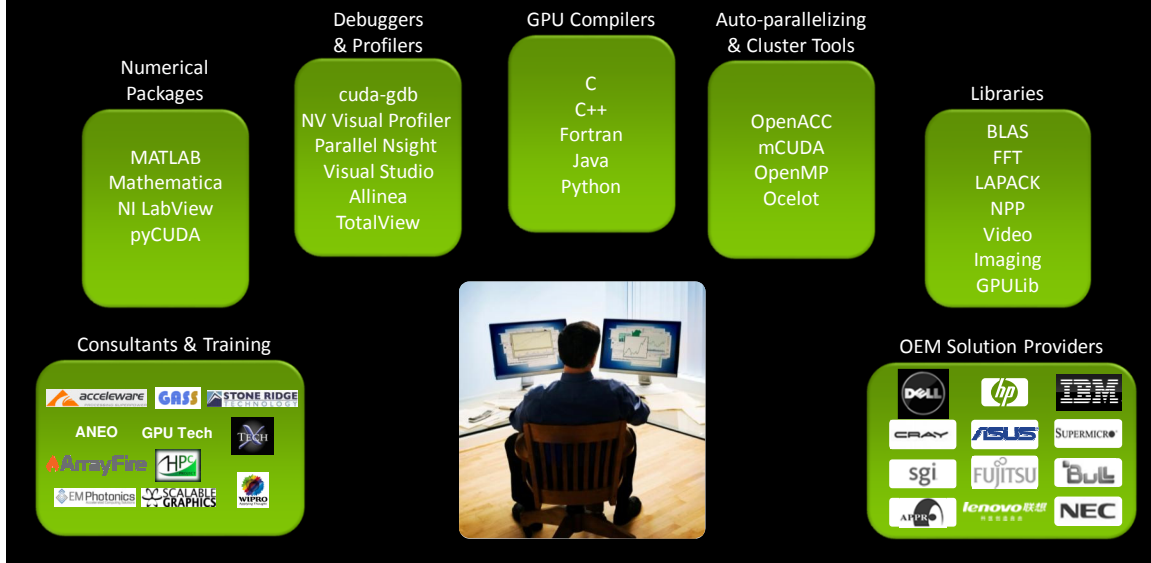
CPU Code

```
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);  
}
```

GPU DEVELOPER ECO-SYSTEM



DEVELOP ON GEFORCE, DEPLOY ON TESLA



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

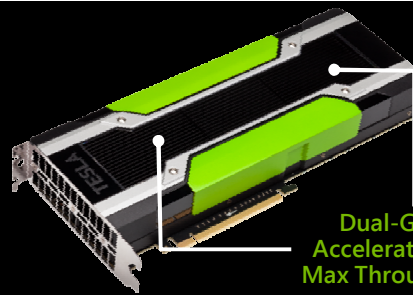


21 NVIDIA

ACCELERATED COMPUTING ROADMAP

TESLA K80

WORLD'S FASTEST ACCELERATOR
FOR DATA ANALYTICS AND
SCIENTIFIC COMPUTING

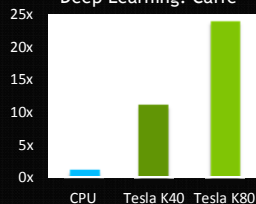


Dual-GPU
Accelerator for
Max Throughput

2x Faster

2.9 TF | 4992 Cores | 480 GB/s

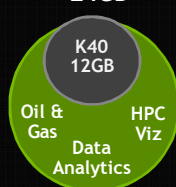
Deep Learning: Caffe



Double the Memory

Designed for Big Data Apps

24GB



Maximum Performance

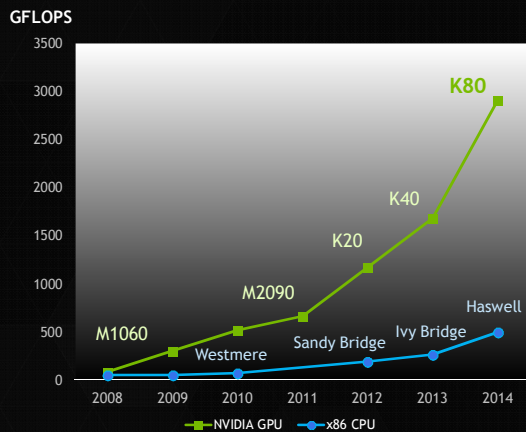
Dynamically Maximize Perf for
Every Application



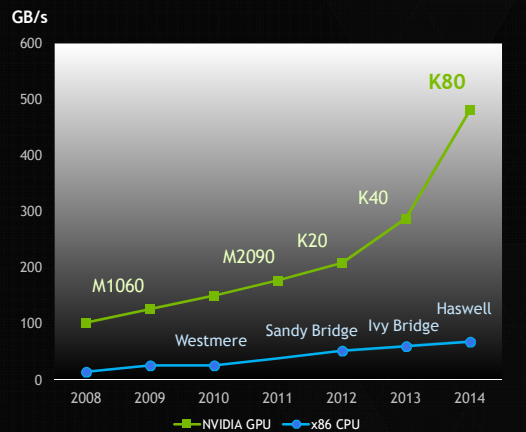
Caffe Benchmark: AlexNet training throughput based on 20 iterations, CPU: E5-2697v2 @ 2.70GHz, 64GB System Memory, CentOS 6.2

PERFORMANCE LEAD CONTINUES TO GROW

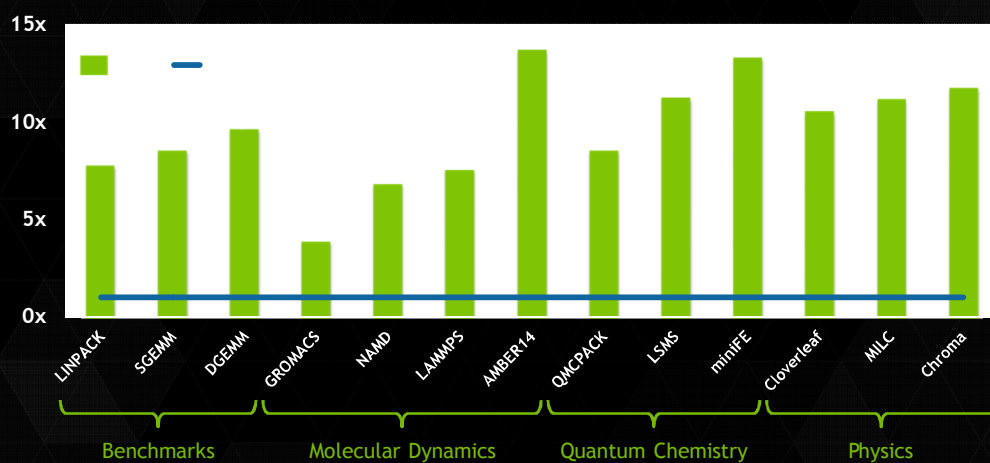
Peak Double Precision FLOPS



Peak Memory Bandwidth



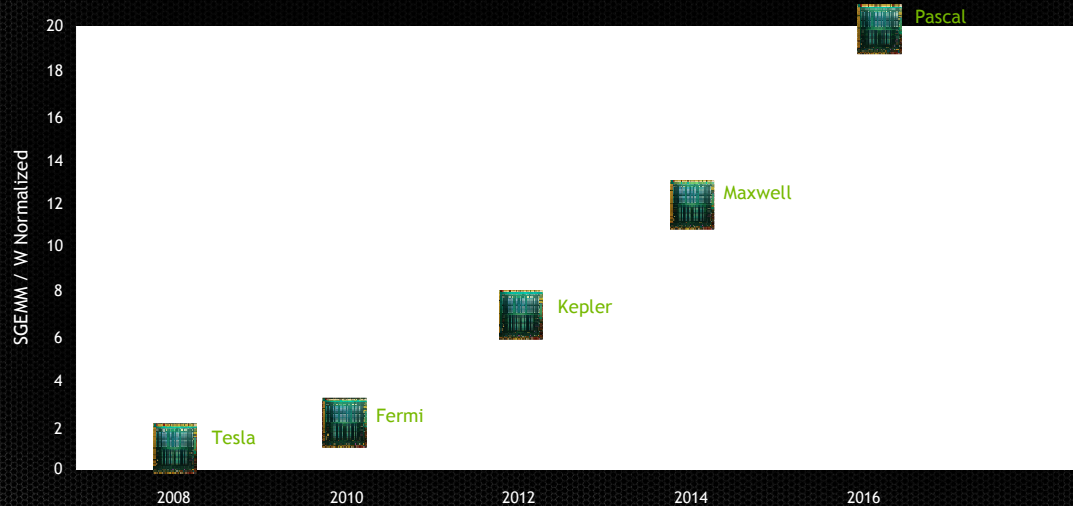
TESLA K80: 10X FASTER ON SCIENTIFIC APPS



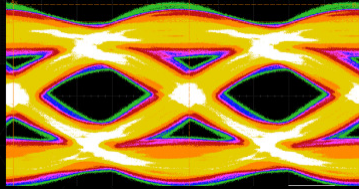
CPU: 12 cores, E5-2697v2 @ 2.70GHz. 64GB System Memory, CentOS 6.2
GPU: Single Tesla K80, Boost enabled

25 NVIDIA

GPU ROADMAP

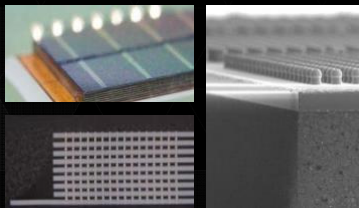


PASCAL GPU FEATURES NVLINK AND STACKED MEMORY



NVLINK

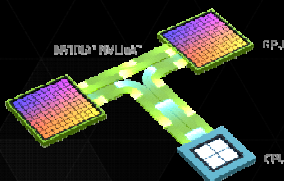
- GPU high speed interconnect
- 80-200 GB/s



3D Stacked Memory

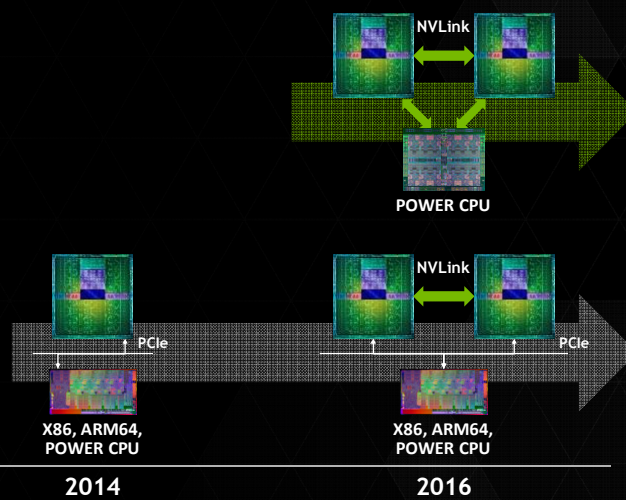
- 4x Higher Bandwidth (~1 TB/s)
- 3x Larger Capacity
- 4x More Energy Efficient per bit

NVLINK HIGH-SPEED GPU INTERCONNECT



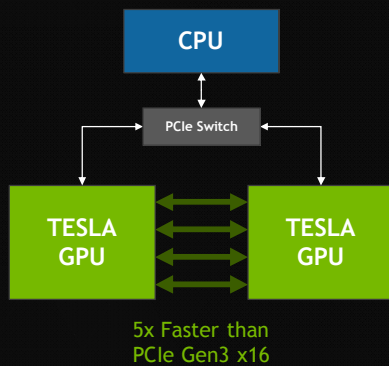
KEPLER GPU

PASCAL GPU



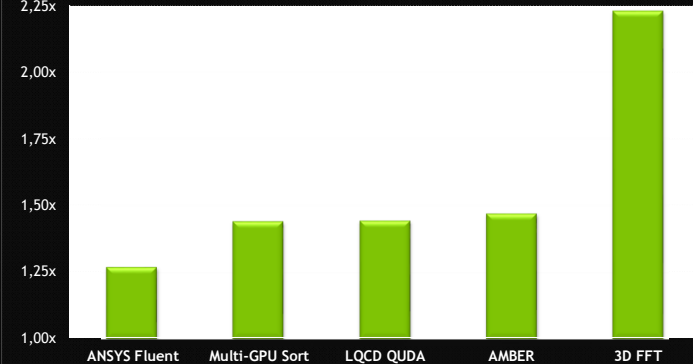
NVLINK UNLEASHES MULTI-GPU PERFORMANCE

GPUs Interconnected with NVLink



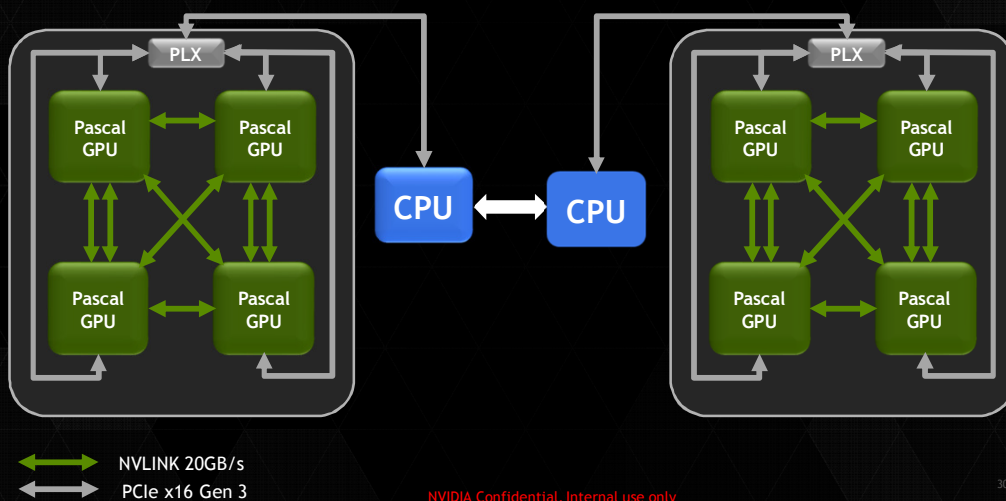
Over 2x Application Performance Speedup When Next-Gen GPUs Connect via NVLink Versus PCIe

Speedup vs
PCIe based Server



29

EXAMPLE: 8-GPU SERVER WITH NVLINK



NVIDIA Confidential. Internal use only

30 NVIDIA

US TO BUILD TWO FLAGSHIP SUPERCOMPUTERS POWERED BY THE TESLA PLATFORM



100-300 PFLOPS Peak

10x in Scientific App Performance

IBM POWER9 CPU + NVIDIA Volta GPU

NVLink High Speed Interconnect

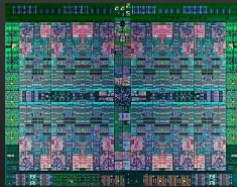
40 TFLOPS per Node, >3,400 Nodes

2017

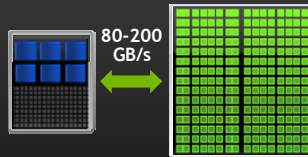
Major Step Forward on the Path to Exascale

31 NVIDIA

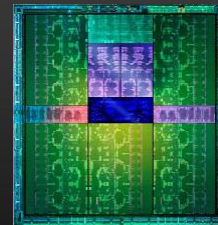
ACCELERATED COMPUTING 5X HIGHER ENERGY EFFICIENCY



IBM POWER CPU
Most Powerful Serial Processor



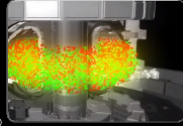
NVIDIA NVLink
Fastest CPU-GPU Interconnect



NVIDIA Volta GPU
Most Powerful Parallel Processor

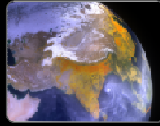
32 NVIDIA

CORAL: BUILT FOR GRAND SCIENTIFIC CHALLENGES



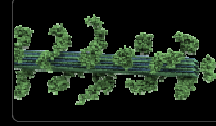
Fusion Energy

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

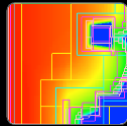


Biofuels

Search for renewable and more efficient energy sources

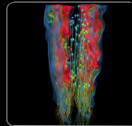
Astrophysics

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



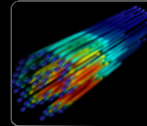
Combustion

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



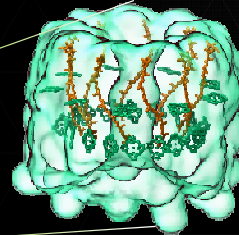
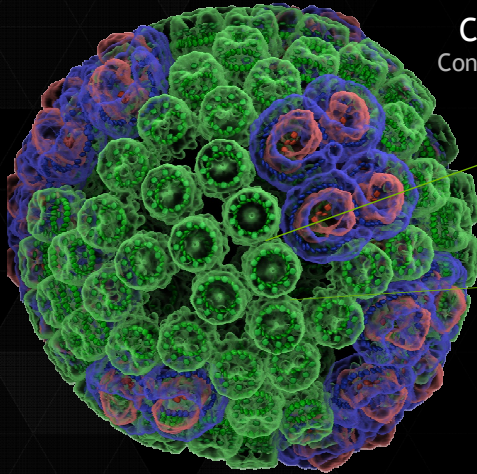
IN-SITU VISUALIZATION

Enabling Visualization with the Tesla Platform

VMD

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

Chromatophore
Converts Light to Energy



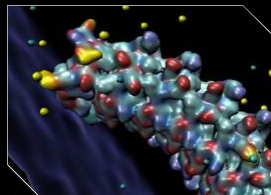
Kitware



ParaView
Parallel Visualization Application

35  NVIDIA

WORLD'S LARGEST IN-SITU HPC VISUALIZATION



2048 GPU Nodes on CSCS Piz Daint

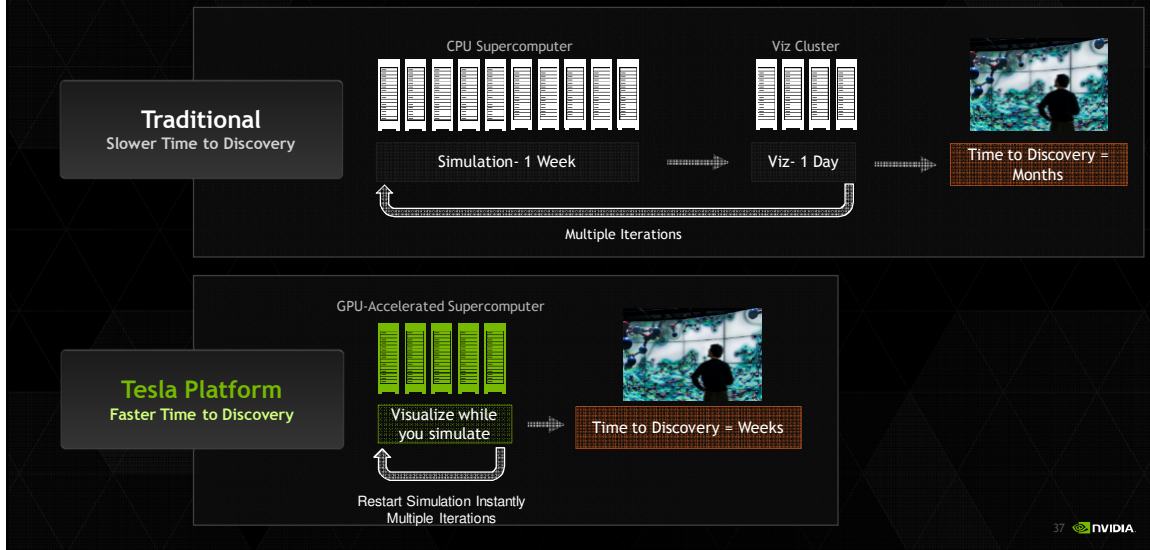
Galaxy formation and
Molecular Dynamics

Simulation + Visualization

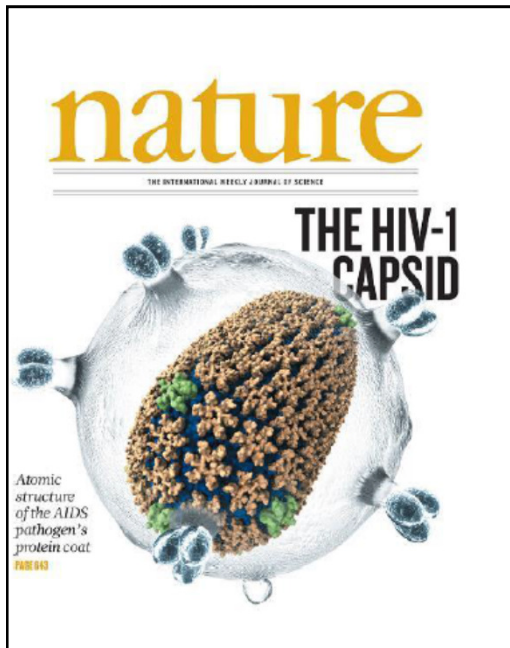


36  NVIDIA

VISUALIZE DATA INSTANTLY FOR FASTER SCIENCE



APPLICATIONS & CUSTOMER SUCCESSES



WHY HPC CUSTOMERS BUY TESLA

TO ACCELERATE SCIENTIFIC DISCOVERY

AMBER MOLECULAR DYNAMICS SIMULATION,
DHFR NVE BENCHMARK

	BEFORE TESLA	AFTER TESLA
Cost	\$200K	\$14K
Servers	32 CPU Servers	1 Dual K80 Server
Energy	21 KW	1 KW
Performance	1X (58ns/day)	4X (220ns/day)



WHY CLOUD COMPUTING CUSTOMERS BUY TESLA

DEEP LEARNING

GOOGLE BRAIN APPLICATION

	BEFORE TESLA	AFTER TESLA
Cost	\$5000K	\$200K
Servers	1000 Servers	16 Tesla Servers
Energy	600 KW	4KW
Performance	1X	6X

ACCELERATING SIGNAL & VIDEO ANALYTICS

Real-time HD video enhancements and analytics

Made possible only with GPUs



Video surveillance with faster than real time analytics

12x faster with GPUs



Unmanned submarine with accelerated sonar processing

50-100x speed up over CPU

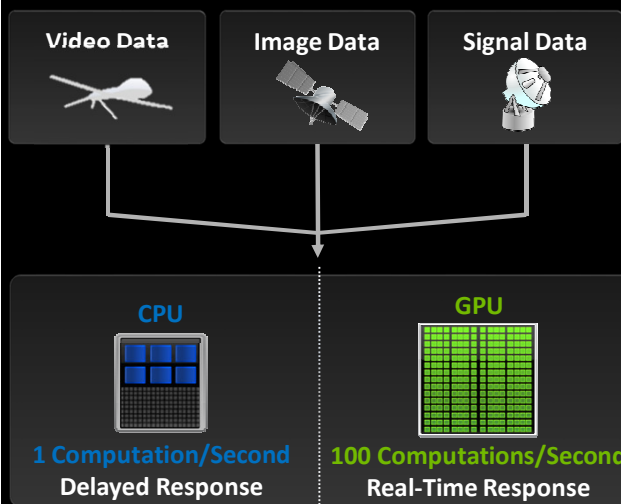


Faster satellite image processing for actionable intelligence

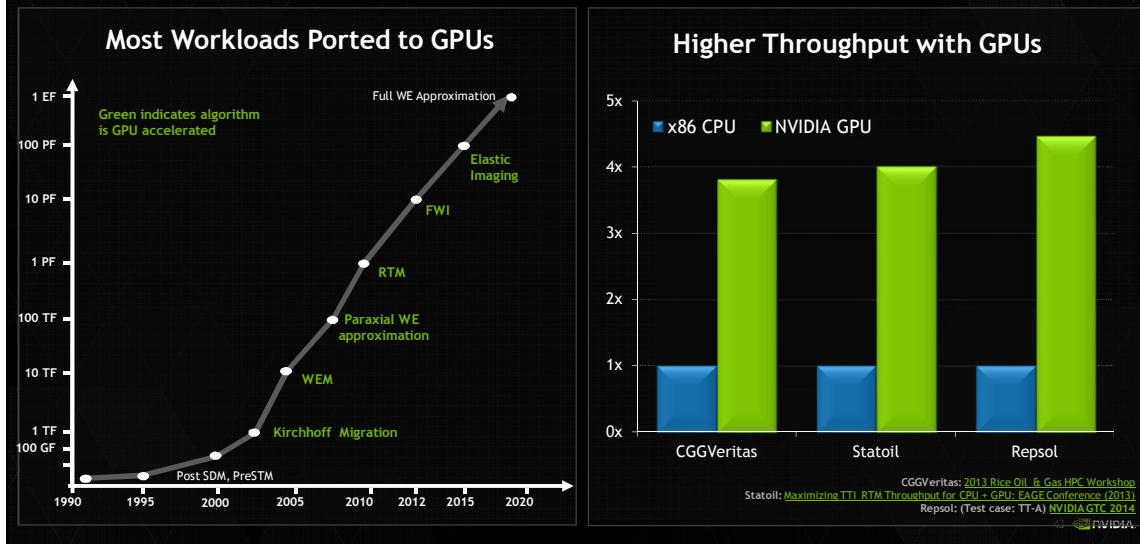
12x faster using GPUs



MISSION PLANNING WITH REAL-TIME LINE OF SIGHT



KEY O&G APPS ACCELERATED ON GPUS



Maximizing Opportunity for Oil Discovery with GPU-powered Supercomputer

WORLD'S FASTEST ENTERPRISE SUPERCOMPUTER

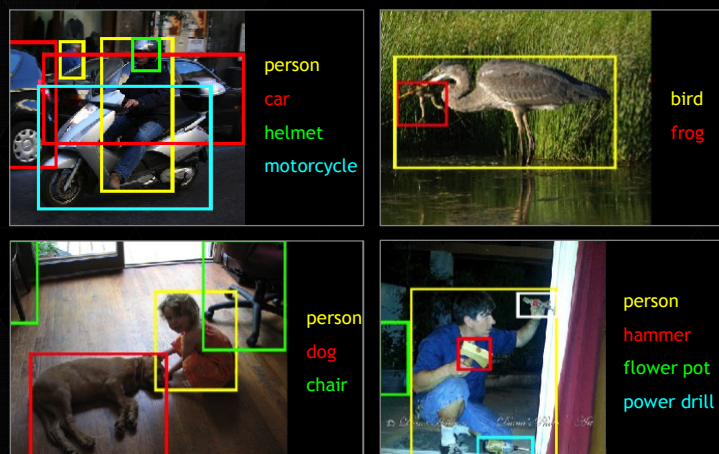
3 Petaflops Linpack Performance

Most Energy-Efficient Petascale
System in the World

3,000 NVIDIA Tesla K20X GPU Accelerators

DEEP LEARNING

DEEP LEARNING FOR IMAGE ANALYTICS



IMAGENET

46 NVIDIA

MACHINE LEARNING USING DEEP NEURAL NETWORKS

The diagram illustrates a Deep Convolutional Neural Network (DCNN) architecture for face recognition. It shows the flow from an input image to a final result. The input is a grayscale face image. This is followed by three stages of feature maps: the first stage shows a 4x4 grid of 16 small, distorted face patches; the second stage shows a 4x4 grid of 16 patches, each containing a single eye; the third stage shows a 4x4 grid of 16 patches, each containing a single nose. Below these visualizations is a schematic of the neural network layers. The input layer is labeled 'Input' and consists of 16 green circles. This is followed by three hidden layers, each with 16 green circles. The final layer is labeled 'Result' and consists of 16 green circles. The layers are connected by lines representing the network's architecture.


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>

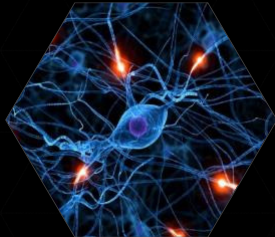
47 NVIDIA

3 DRIVERS FOR DEEP LEARNING


More Data




Better Models



Powerful GPU Accelerators



48  NVIDIA

BROAD USE OF GPUS IN DEEP LEARNING

Early Adopters



Use Cases

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

Talks @ GTC



BROAD BENEFITS OF DEEP LEARNING



Content-based music recommendation

Summer Intern implements recommendation system



Mitosis Detection in Breast Cancer Histology Images with Deep Neural Networks

Dan C. Cireşcan et al.

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



Merck Molecular Activity Challenge

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

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

51  NVIDIA

GPU TECHNOLOGY
CONFERENCE

MARCH 2015

www.gputechconf.com



4,000 guests • 550 talks • 175 posters

"At the NVIDIA GPU Developer's conference this week I'll be thinking about the future and wondering if I'm not already in it." —TechZone

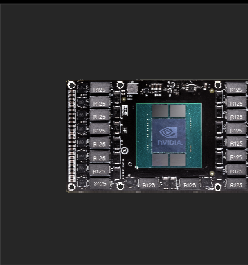
TITAN X
The World's Fastest GPU



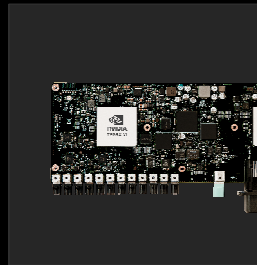
DIGITS DevBox
GPU Deep Learning Platform



Pascal — 10x Maxwell
For Deep Learning



NVIDIA DRIVE PX
Deep Learning Platform
for Self-Driving Cars



GTC 2015 focused on the promising field of deep learning.
And we made four major announcements that will fuel its advance.

gbarat@nvidia.com

www.nvidia.com

www.gputechconf.com