Accelerated Computing from Mobile Devices to Supercomputers

Dale Southard, NVIDIA

Power of 600 Petaflop
CPU-only Supercomputer = Power for the city
of San Francisco

HPC’s Biggest Challenge: Power

GPUs Power World's 10 Greenest Supercomputers

Green500 Rank	MFLOPS/W	Site
1	4,503.17	GSIC Center, Tokyo Tech
2	3,631.86	Cambridge University
3	3,517.84	University of Tsukuba
4	3,185.91	Swiss National Supercomputing (CSCS)
5	3,130.95	ROMEO HPC Center
6	3,068.71	GSIC Center, Tokyo Tech
7	2,702.16	University of Arizona
8	2,629.10	Max-Planck
9	2,629.10	(Financial Institution)
10	2,358.69	CSIRO
37	1959.90	Intel Endeavor (top Xeon Phi cluster)
49	1247.57	Météo France (top CPU cluster)

Accelerated Computing

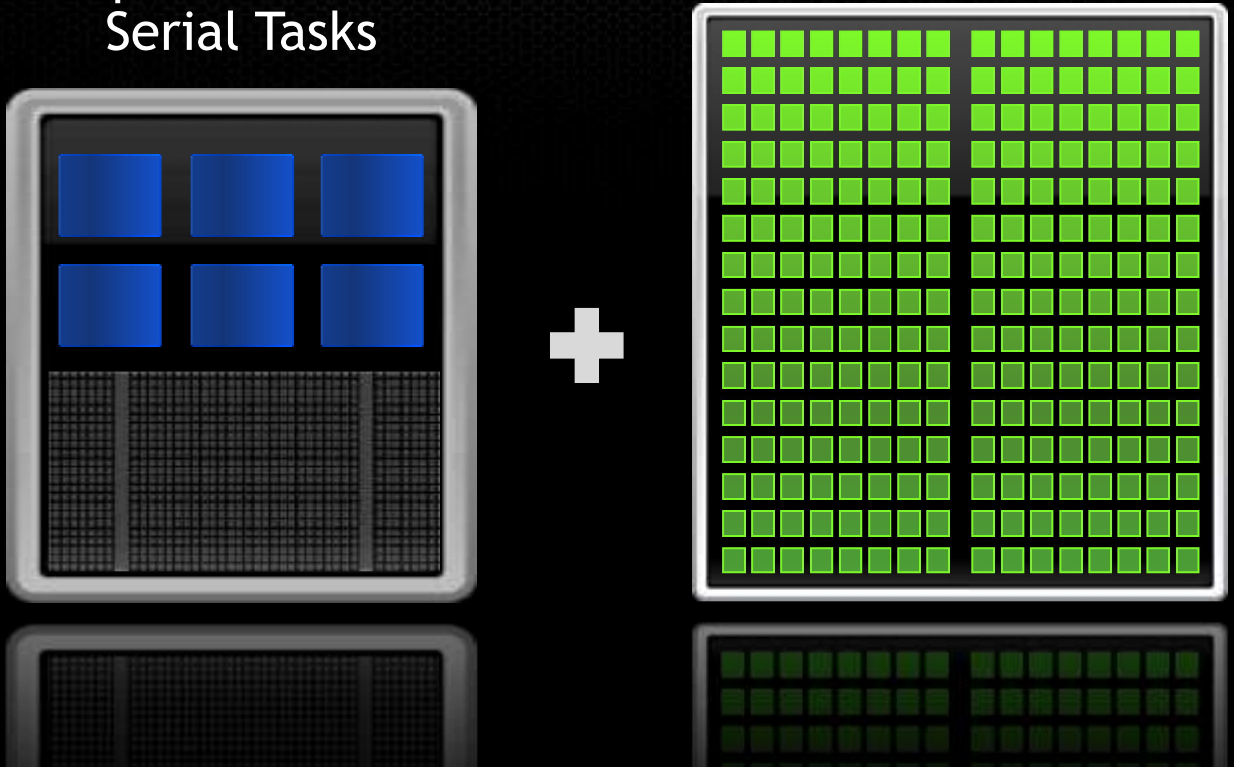
10x Performance & 5x Energy Efficiency for HPC

CPU

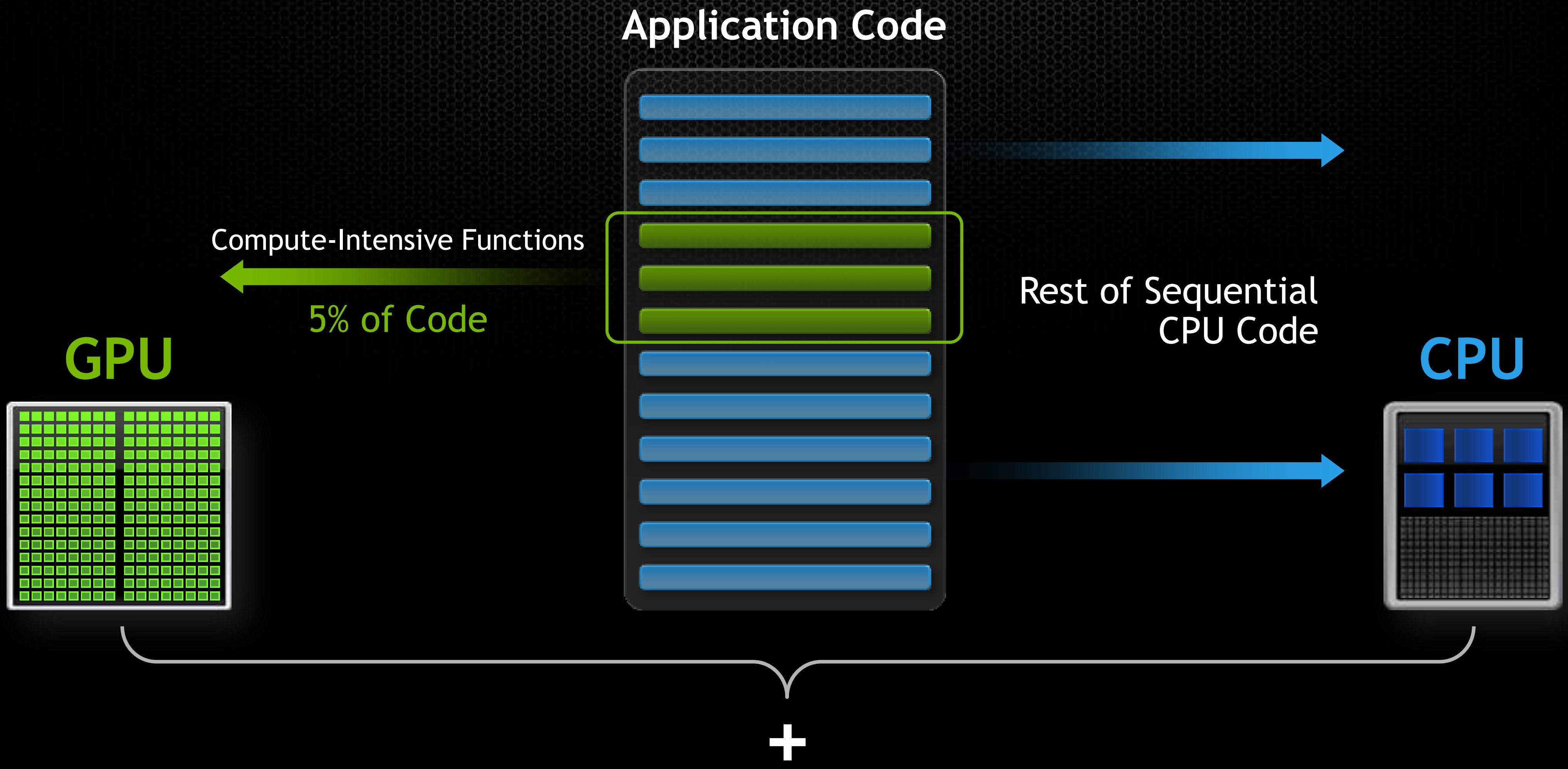
Optimized for
Serial Tasks

GPU Accelerator

Optimized for
Parallel Tasks



How GPU Acceleration Works



Application Code

Compute-Intensive Functions

5% of Code

GPU

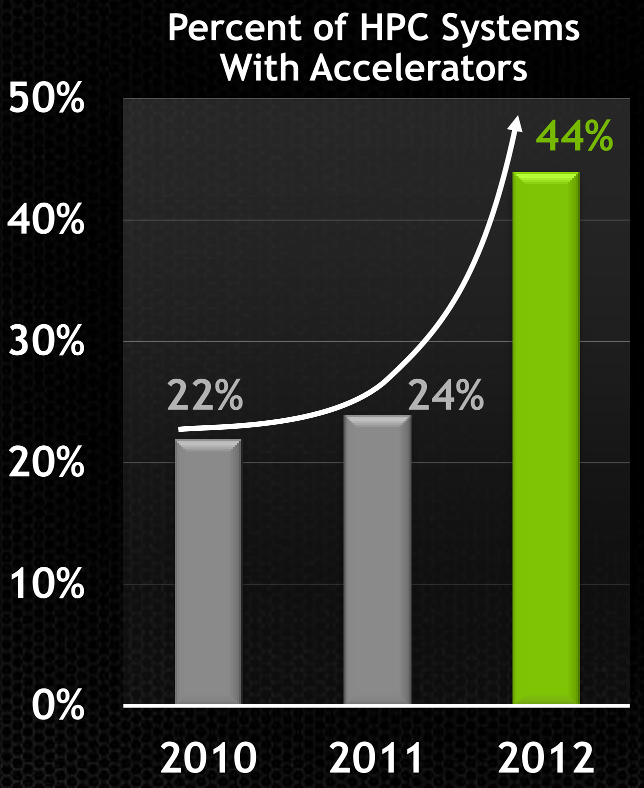
Rest of Sequential CPU Code

CPU

+

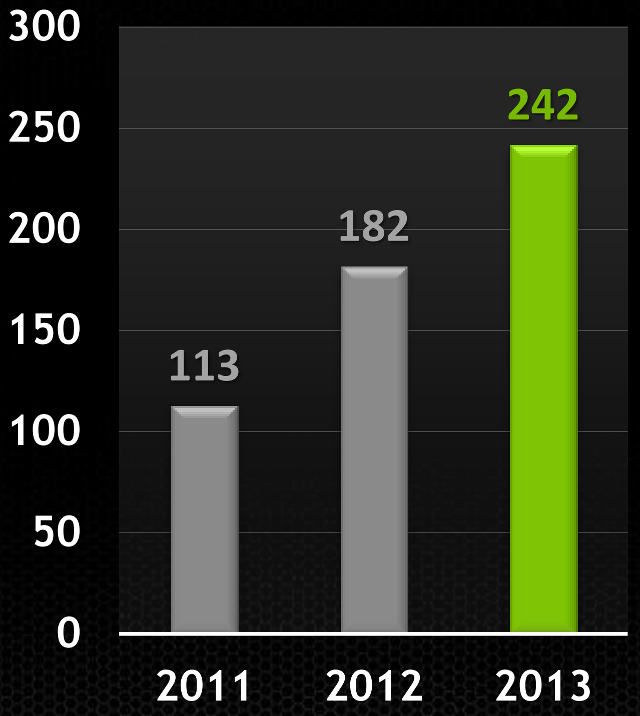
Accelerated Computing Growing Fast

2x Growth in One Year

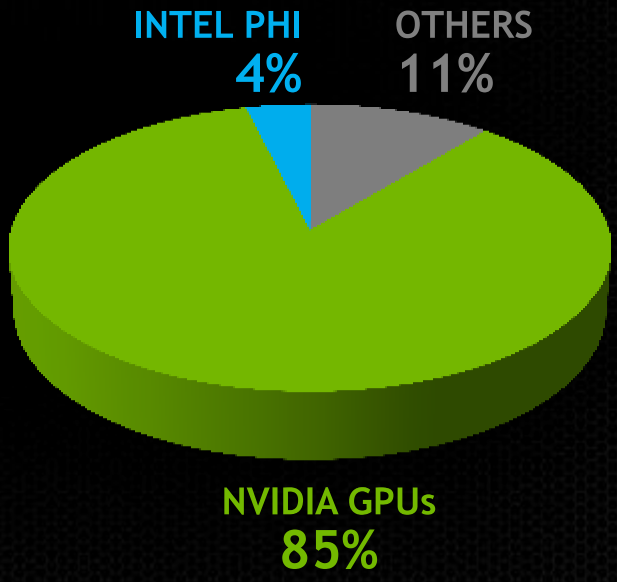


Intersect360 Research
HPC User Site Census: Systems, July 2013

Hundreds of GPU Accelerated Apps



NVIDIA GPU is Accelerator of Choice



Intersect360 Research
HPC User Site Census: Systems, July 2013

NVIDIA

POPULAR GPU-ACCELERATED APPLICATIONS

Research: Higher Education and Supercomputing

COMPUTATIONAL CHEMISTRY AND BIOLOGY

Bioinformatics

APPLICATION	DESCRIPTION	SUPPORTED FEATURES	EXPECTED SPEED UP	RECOMMENDED GPU	MULTI-GPU SUPPORT	RELEASE STATUS
BarraCUDA	Sequence mapping software	Alignment of short sequencing reads	6-10x	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 0.6.2
CUDASW++	Open source software for Smith-Waterman protein database searches on GPUs	Parallel search of Smith-Waterman database	10-50x	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 2.0.8
CUSHAW	Parallelized short read aligner	Parallel, accurate long read aligner - gapped alignments to large genomes	10x	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 1.0.40
GPU-BLAST	Local search with fast k-tuple heuristic	Protein alignment according to blastp, multi cpu threads	3-4x	T 2075, 2090, K10, K20, K20X	Single only	Available now Version 2.2.26
GPU-HMMER	Parallelized local and global search with profile Hidden Markov models	Parallel local and global search of Hidden Markov Models	60-100x	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 2.3.2
mCUDA-MEME	Ultrafast scalable motif discovery algorithm based on MEME	Scalable motif discovery algorithm based on MEME	4-10x	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 3.0.12
SeqNFind	A GPU Accelerated Sequence Analysis Toolset	Reference assembly, blast, smith-waterman, hmm, de novo assembly	400x	T 2075, 2090, K10, K20, K20X	Yes	Available now
UGENE	Opensource Smith-Waterman for SSE/CUDA, Suffix array based repeats finder and dotplot	Fast short read alignment	6-8x	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 1.11
WideLM	Fits numerous linear models to a fixed design and response	Parallel linear regression on multiple similarly-shaped models	150x	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 0.1-1

Molecular Dynamics

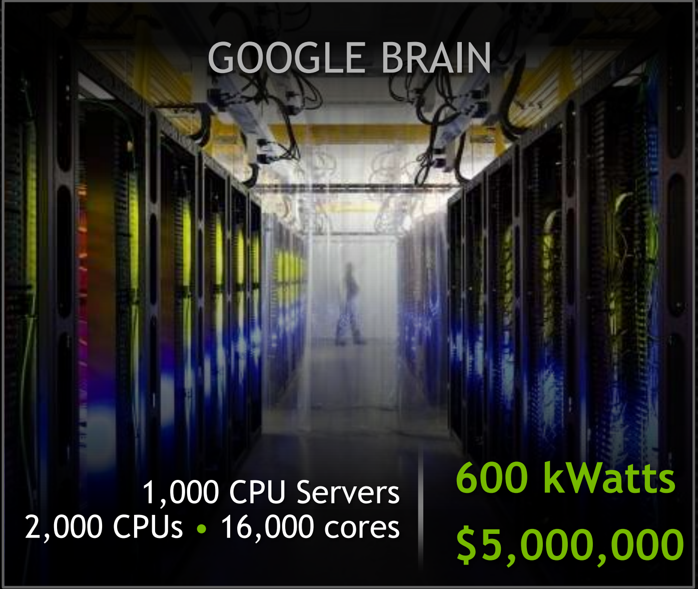
APPLICATION	DESCRIPTION	SUPPORTED FEATURES	EXPECTED SPEED UP	RECOMMENDED GPU	MULTI-GPU SUPPORT	RELEASE STATUS
Abalone	Models molecular dynamics of biopolymers for simulations of proteins, DNA and ligands	Simulations (on 1060 GPU)	4-29x	T 2075, 2090, K10, K20, K20X	Single Only	Available now Version 1.8.48
ACEMD	GPU simulation of molecular mechanics force fields, implicit and explicit solvent	Written for use on GPUs	160 ns/day GPU version only	T 2075, 2090, K10, K20, K20X	Yes	Available now
AMBER	Suite of programs to simulate molecular dynamics on biomolecule	PMEMD: explicit and implicit solvent	89.44 ns/day JAC NVE	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 12 + bugfix9
DL-POLY	Simulate macromolecules, polymers, ionic systems, etc on a distributed memory parallel computer	Two-body forces, Link-cell pairs, Ewald SPME forces, Shake VV	4x	T 2075, 2090, K10, K20, K20X	Yes	Available now Version4.0 Source only
CHARMM	MD package to simulate molecular dynamics on biomolecule.	Implicit (5x), Explicit (2x) Solvent via OpenMM	TBD	T 2075, 2090, K10, K20, K20X	Yes	In Development Q4/12
GROMACS	Simulation of biochemical molecules with complicated bond interactions	Implicit (5x), Explicit(2x) solvent	165 ns/Day DHFR	T 2075, 2090, K10, K20, K20X	Single only	Available now Version 4.6 in Q4/12
HOOMD-Blue	Particle dynamics package written grounds up for GPUs	Written for GPUs	2x	T 2075, 2090, K10, K20, K20X	Yes	Available now
LAMMPS	Classical molecular dynamics package	Lennard-Jones, Morse, Buckingham, CHARMM, Tabulated, Course grain SDK, Anisotropic Gay-Bern, RE-squared, "Hybrid" combinations	3-18x	T 2075, 2090, K10, K20, K20X	Yes	Available now
AMD	Designed for high-performance simulation of large molecular systems	100M atom capable	6.44 ns/days STMV 585x 2050s	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 2.9
enMM	Library and application for molecular dynamics for HPC with GPUs	Implicit and explicit solvent, custom forces	Implicit: 127-213 ns/ day. Explic 18.55 ns/ DHFR	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 4.1.1

272 GPU-Accelerated Applications

www.nvidia.com/appscatalog

Artificial Neural Network at a Fraction of the Cost with GPUs

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



GOOGLE BRAIN

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

GPUs Accelerate Machine Learning & Data Analytics

Adobe

Auto Tagging in Creative Cloud

Baidu

Speech/Image Recognition

IBM

Hadoop-based Clustering

NETFLIX

Recommendation Engine

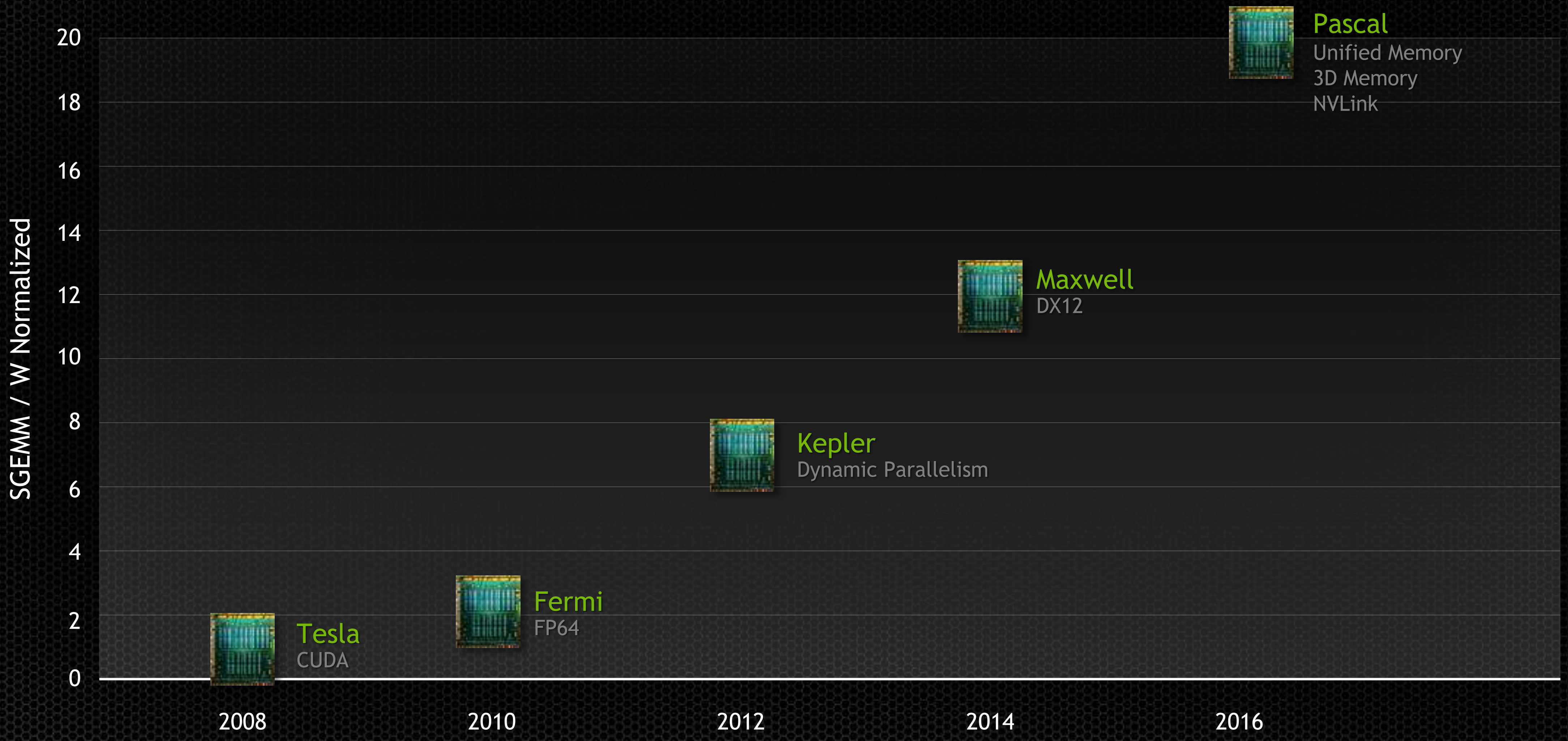
orange™

Database Queries

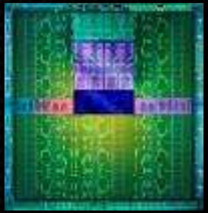
Yandex

Search Ranking

Strong CUDA GPU Roadmap

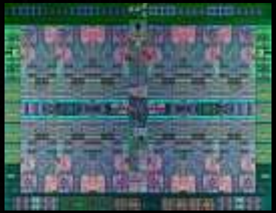


IBM Partners with NVIDIA to Build Next-Generation Supercomputers



Tesla
GPU

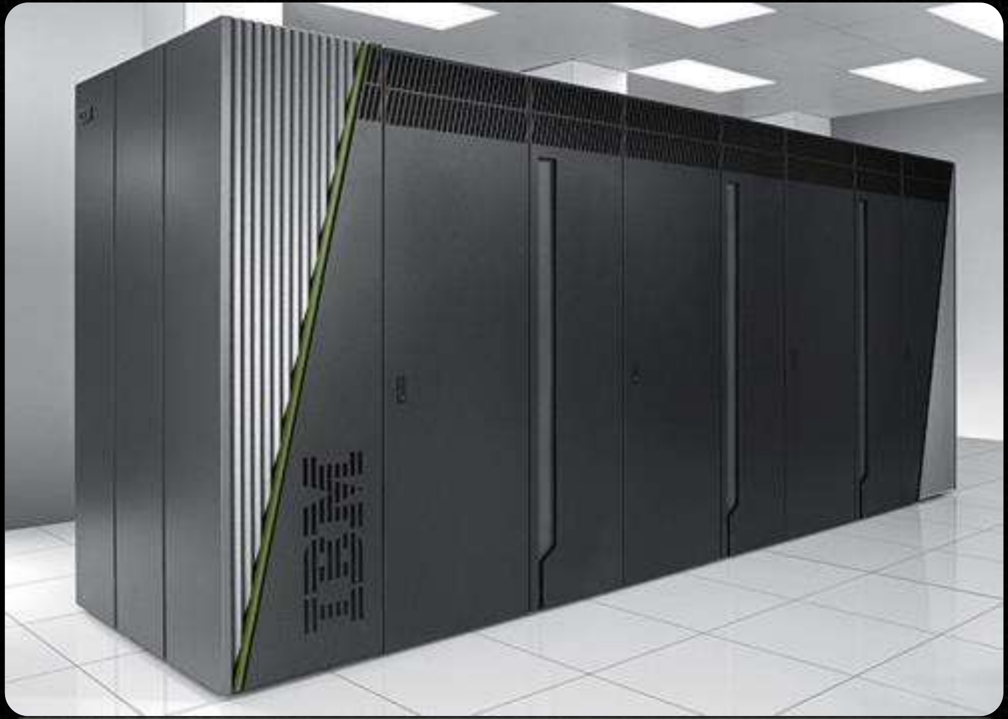
+



POWER 8
CPU

NVIDIA

IBM®



GPU-Accelerated POWER-Based Systems Available in 2014

JETSON TK1

THE WORLD'S 1st EMBEDDED SUPERCOMPUTER

Development Platform for Embedded Computer Vision, Robotics, Medical

192 Cores · 326 GFLOPS FP32

CUDA Enabled

Available Now

GPU Accelerated Libraries

“Drop-in” Acceleration for your Applications

Linear Algebra

FFT, BLAS,
SPARSE, Matrix

NVIDIA
cuFFT,
cuBLAS,
cuSPARSE

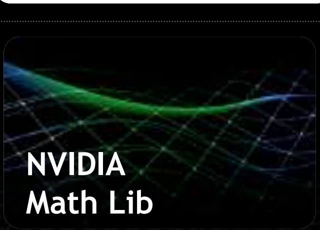
CULA|tools

MAGMA
ICL UT

CUSP

Numerical & Math

RAND, Statistics


NVIDIA
Math Lib

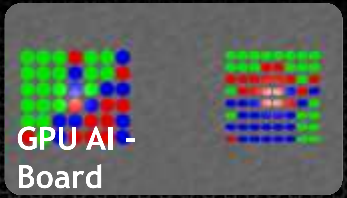
ArrayFire

NVIDIA cuRAND

Data Struct. & AI

Sort, Scan, Zero Sum

Thrust


GPU AI -
Board
Games

GPU AI -
Path Finding

Visual Processing

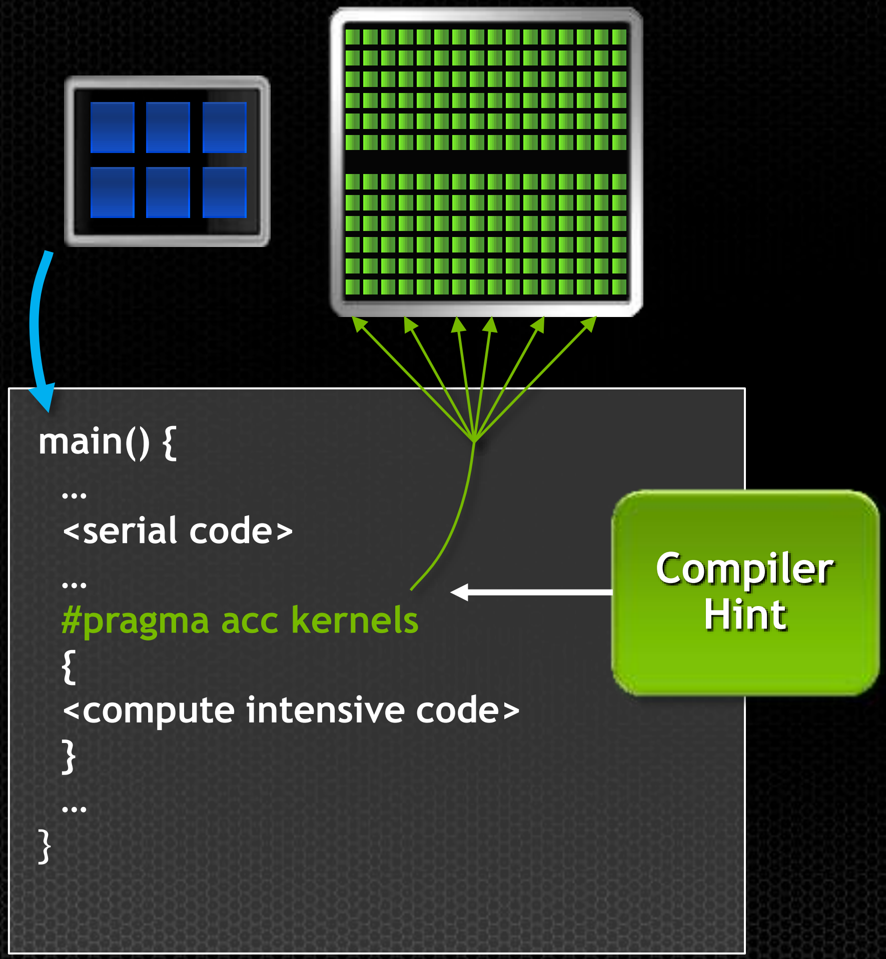
Image & Video

NVIDIA
NPP

NVIDIA
Video
Encode

Sundog™
Software

OpenACC: Open, Simple, Portable



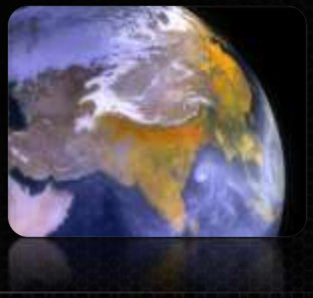
```
main() {
  …
  <serial code>
  …
  #pragma acc kernels
  {
  <compute intensive code>
  }
  …
}
```

- Open Standard
- Easy, Compiler-Driven Approach
- Portable on GPUs and Xeon Phi

CAM-SE Climate

6x Faster on GPU

Top Kernel: 50% of Runtime



Linux GCC Compiler to Support GPU Accelerators

Open Source

GCC Efforts by Samsung & Mentor Graphics

Pervasive Impact

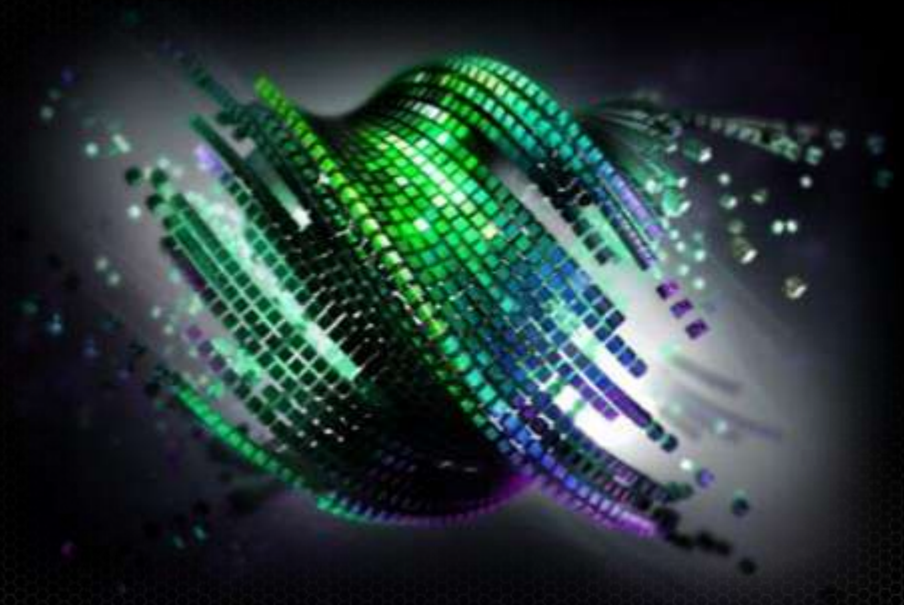
Free to all Linux users

Mainstream

Most Widely Used HPC Compiler

> *“Incorporating OpenACC into GCC is an excellent example of open source and open standards working together to make accelerated computing broadly accessible to all Linux developers.”*
>
> Oscar Hernandez
> Oak Ridge National Laboratories

ANNOUNCING
CUDA 6



Dramatically Simplifies

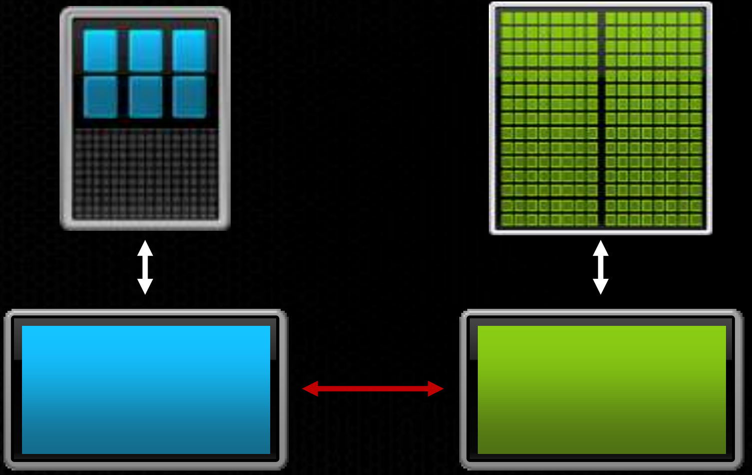
Parallel Programming with

Unified Memory

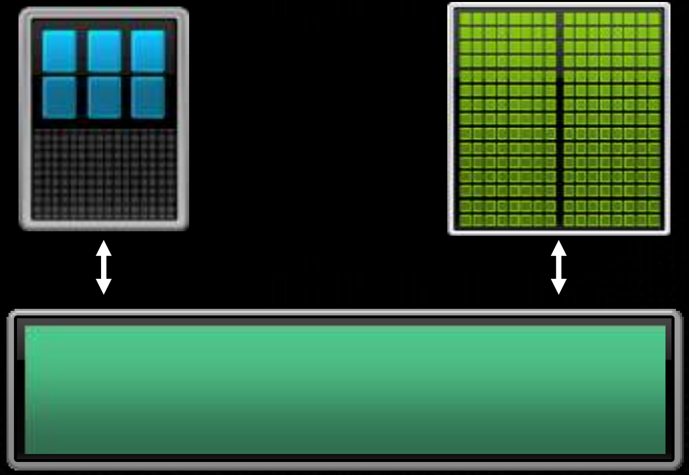
Unified Memory

Dramatically Lower Developer Effort

Developer View Today



Developer View With Unified Memory



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

Dale Southard, dsouthard@nvidia.com

Principal System Architect, Office of the CTO