



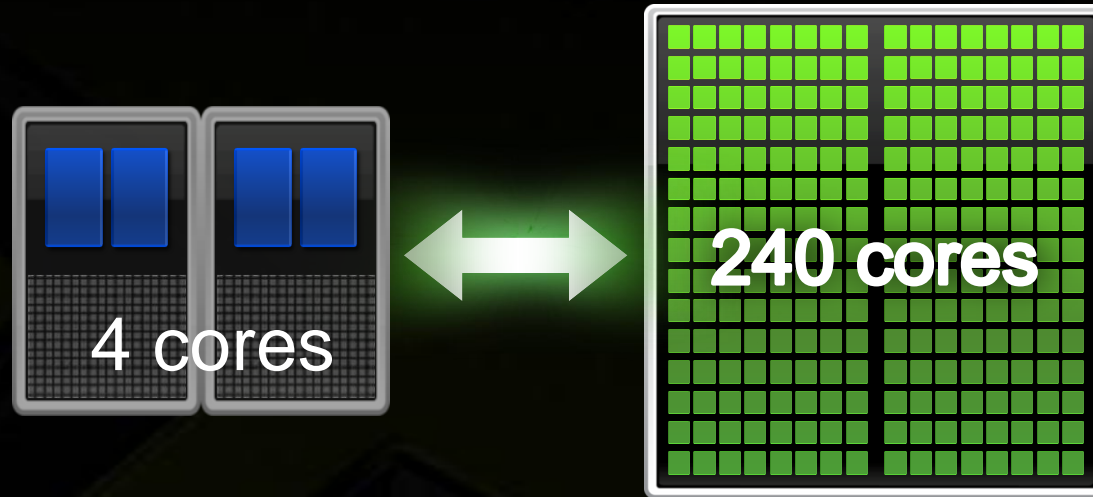
Tesla GPU Computing

A Revolution in High Performance Computing

CSIRO / IEEE EMBS Science Symposium on GPGPU Techniques
for Medical Image Processing & Simulation
Brisbane, 30 March 2009



What is GPU Computing?



Computing with CPU + GPU
Heterogeneous Computing

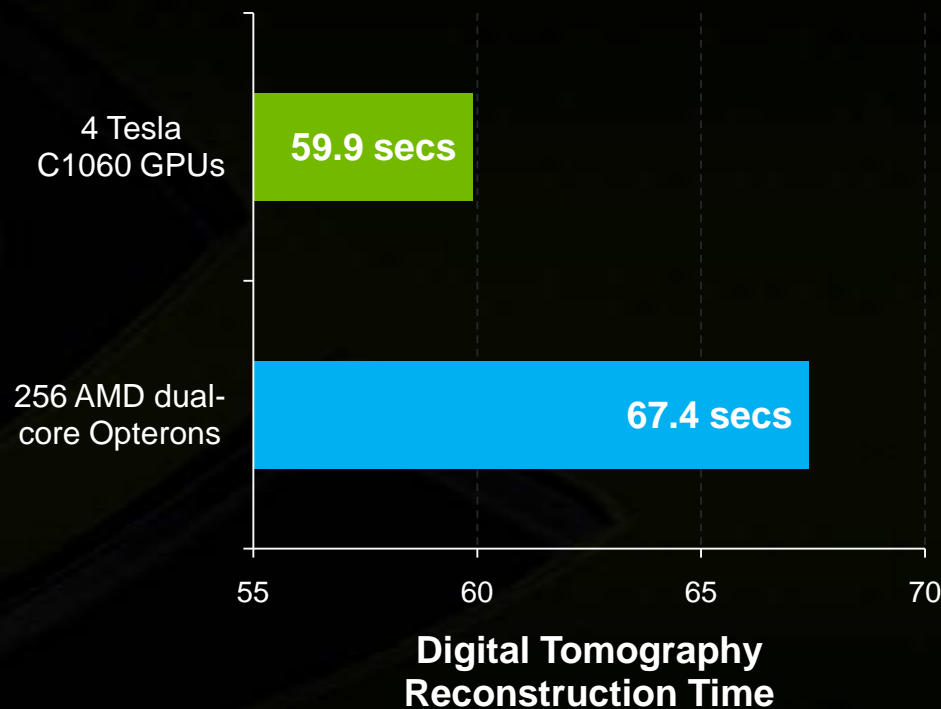
GPUs: Turning Point in Supercomputing



Desktop beats Cluster



CalcUA
\$5 Million



**Tesla Personal
Supercomputer**
\$10,000

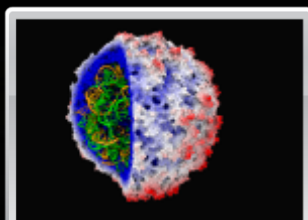
Source: University of Antwerp, Belgium

Not 2x or 3x : Speedups are 20x to 150x



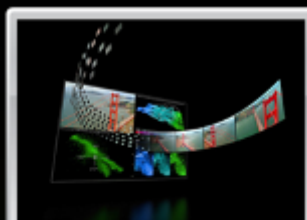
146X

Medical Imaging
U of Utah



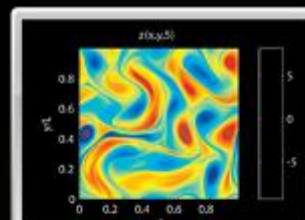
36X

Molecular Dynamics
U of Illinois, Urbana



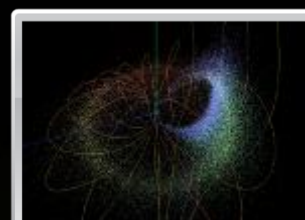
18X

Video Transcoding
Elemental Tech



50X

Matlab Computing
AccelerEyes



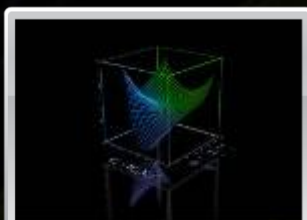
100X

Astrophysics
RIKEN



149X

Financial simulation
Oxford



47X

Linear Algebra
Universidad Jaime



20X

3D Ultrasound
Techniscan



130X

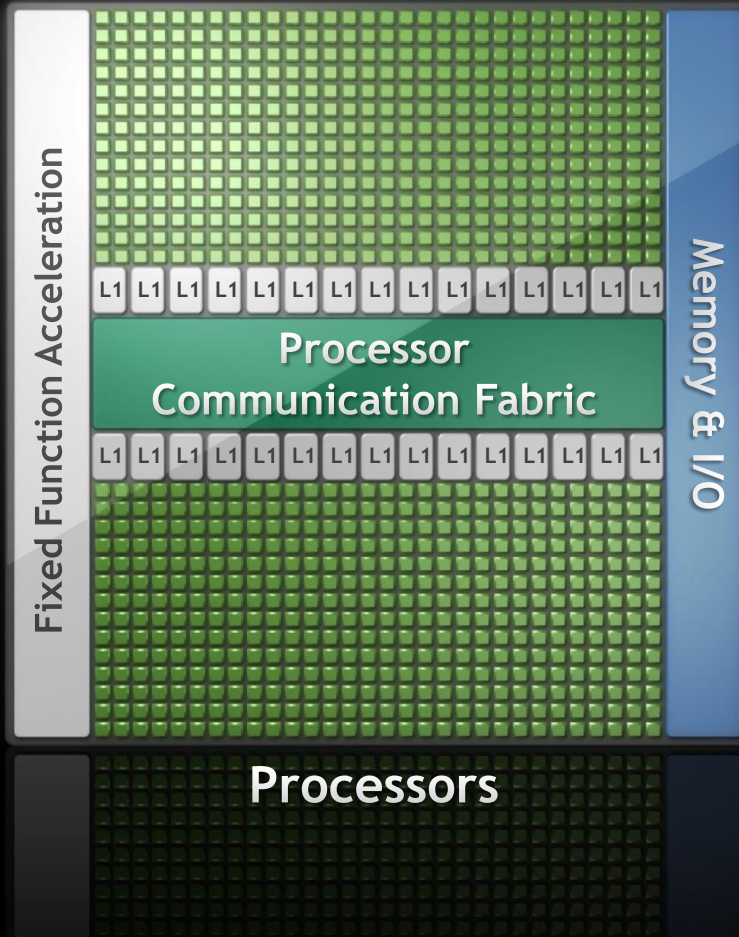
Quantum Chemistry
U of Illinois, Urbana



30X

Gene Sequencing
U of Maryland

Processors



NVIDIA Tesla 10-Series GPU

Massively parallel, many core architecture

240 Processor Cores

1 Teraflops - 1,000 times Cray X-MP

IEEE Compliant Double Precision Floating Point

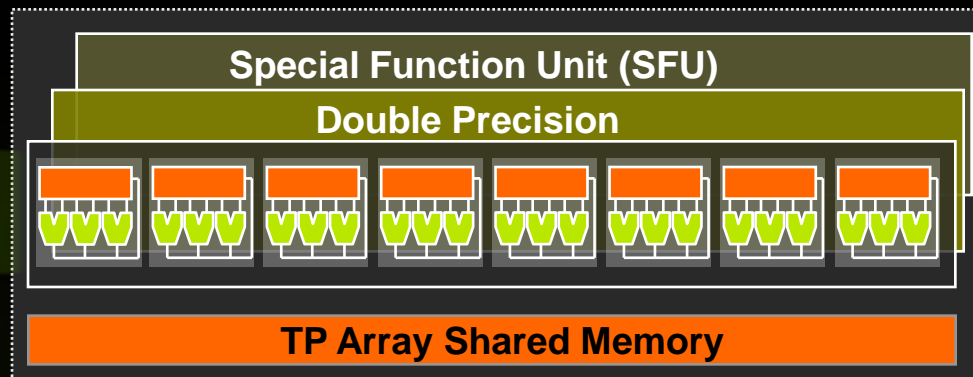
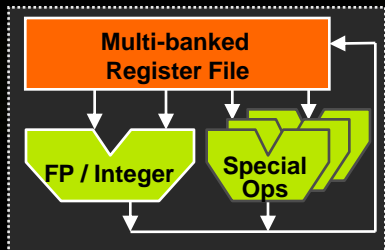
Designed for Scientific Computing

Tesla T10 GPU: 240 Processor Cores

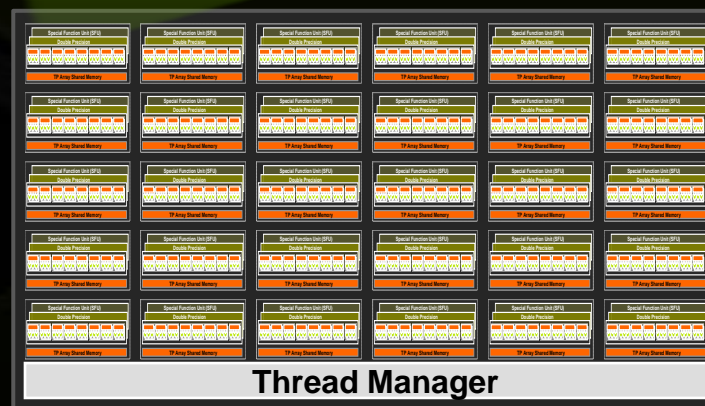


Thread Processor Array (TPA)

Thread Processor (TP)



- Processor core has
 - Floating point / Integer unit
 - Move, compare, logic, branch unit
- IEEE 754 floating point
 - Single and Double
- 102 GB/s high-speed interface to memory



GDDR3
512 bit

Main
Memory

102 GB/sec

30 TPAs = 240 Processors

Parallel Computing on All GPUs

100+ Million CUDA GPUs Deployed



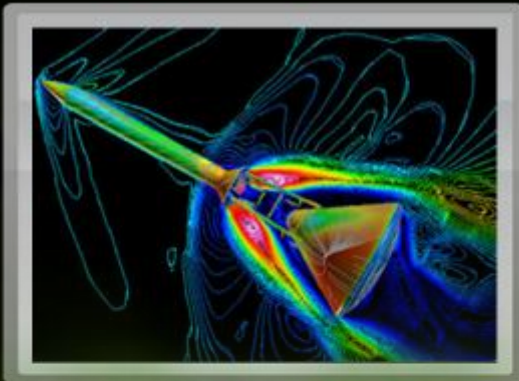
GeForce®

Entertainment



Tesla™

High-Performance Computing



Quadro®

Design & Creation



GPU

Tesla GPU Computing Products



Tesla S1070 1U System



Tesla C1060
Computing Board



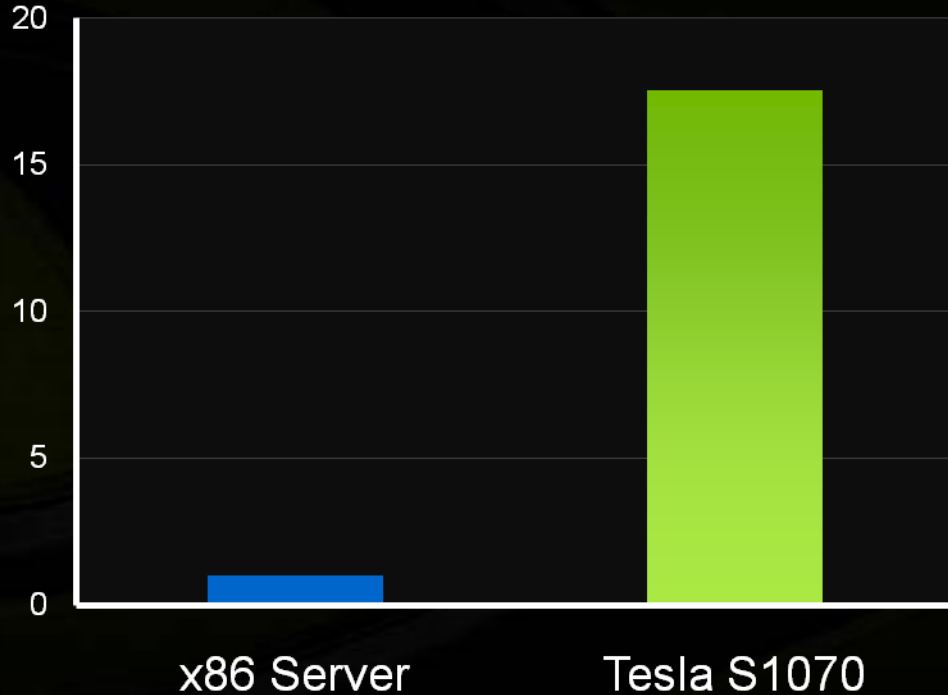
Tesla Personal
Supercomputer
(4 Tesla C1060s)

GPUs	4 Tesla GPUs	1 Tesla GPU	4 Tesla GPUs
Single Precision Perf	4.14 Teraflops	933 Gigaflops	3.7 Teraflops
Double Precision Perf	346 Gigaflops	78 Gigaflops	312 Gigaflops
Memory	4 GB / GPU	4 GB	4 GB / GPU

Tesla S1070: Green Supercomputing



**20X Better
Performance / Watt**



- Hess
- Chevron
- Petrobras
- NCSA
- CEA
- Tokyo Tech
- JFCOM
- SAIC
- Federal
- Motorola
- Kodak
- University of Heidelberg
- University of Illinois
- University of North Carolina
- Max Planck Institute
- Rice University
- University of Maryland
- GusGus
- Eotvas University
- University of Wuppertal
- Chinese Academy of Sciences
- National Taiwan University

A \$5 Million Datacenter



CPU 1U Server



2 Quad-core Xeon
CPUs: 8 cores

0.17 Teraflop (single)
0.08 Teraflop (double)

\$ 3,000

700 W

8 CPU Cores +
4 GPUs = 968 cores

4.14 Teraflops (single)
0.346 Teraflop (double)

\$ 11,000

1500 W

CPU 1U Server
Tesla 1U System



1819 CPU servers

310 Teraflops (single)

155 Teraflops (double)

Total area 16K sq feet

Total 1273 KW

455 CPU servers
455 Tesla systems

1961 Teraflops (single)

196 Teraflops (double)

Total area 9K sq feet

Total 682 KW

6x more perf

60% smaller

½ the power

Introducing the *Tesla Personal Supercomputer*



Supercomputing Performance

- Massively parallel CUDA Architecture
- 960 cores. 4 TeraFlops
- 250x the performance of a desktop

Personal

- One researcher, one supercomputer
- Plugs into standard power strip

Accessible

- Program in C for Windows, Linux
- Available now worldwide under \$10,000



More Information



- **Tesla main page**

- <http://www.nvidia.com/tesla>
- Product Specs, Marketing Literature

- **CUDA Zone**

- <http://www.nvidia.com/cuda>
- Applications, Papers, Videos

- **YouTube Videos**

- <http://www.youtube.com/nvidiatesla>
- Hear CUDA developers talk about their experiences

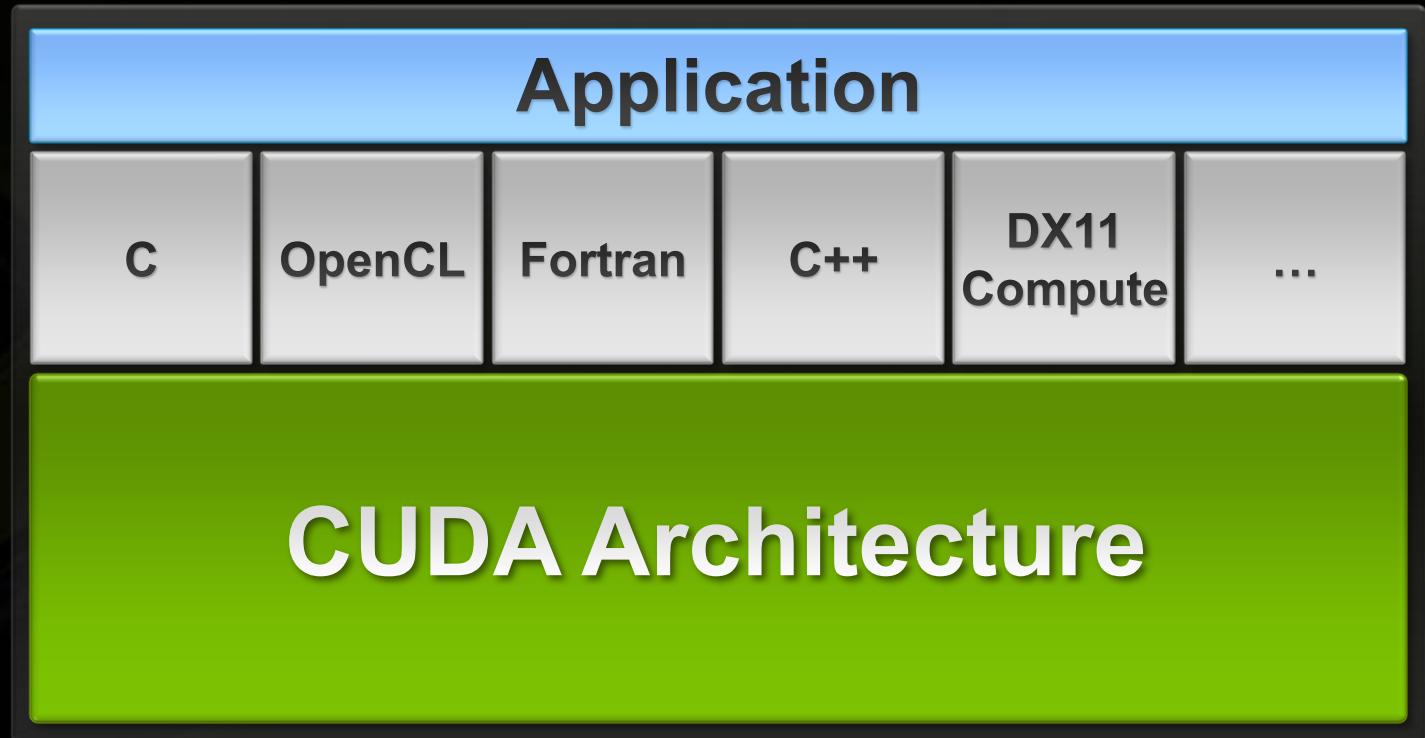
CUDA

CUDA Parallel Programming Architecture and Programming Model

CUDA Parallel Computing Architecture




- Parallel computing architecture and programming model
- Includes a C compiler plus support for OpenCL and DX11 Compute
- Architected to natively support all computational interfaces (standard languages and APIs)



CUDA Facts

- 750+ Research Papers
- 50+ universities teaching CUDA
- 100 Million CUDA-Enabled GPUs
- 25K Active Developers



The screenshot shows the NVIDIA CUDA Zone website. At the top, there's the NVIDIA logo and the "CUDA ZONE" header. A navigation bar includes links for "DOWNLOAD CUDA", "WHAT IS CUDA", "DEVELOPING WITH CUDA", "FORUMS", and "NEWS AND EVENTS". A search bar is also present. Below the navigation bar, a banner reads "LATEST CUDA NEWS Get the Next 20x Performance – Sign Up for Advanced CUDA Training Boot Camp @ NVISION 2008". The main content area displays a grid of project thumbnails, each with a title, a small image, and a download count (e.g., "35 x", "18 x", "50 x"). The projects include "Audio FIR Crossover", "GLAMElab API for Linear Algebra Operations on GPUs", "H.264 Video Encoder", "Large Vocabulary Continuous Speech Recognition", "Quantitative Risk Analysis and Algorithmic Trading Systems", "Canny Edge Detection", "GPU Acceleration Solutions", "Innovative 3D visualization solutions for Oil and Gas", "LIBOR Interest rate Model", "Ray Casting Algebraic Surfaces using the Frustum Form", "Dirac Video Codec", "GPU4Vision", "Jacket: GPU Engine for MATLAB", "Prestack Seismic Data Interaction", and "Ray Casting Deformable Models". At the bottom, there's a search bar, a "Sort by Name" dropdown, and a "Share Your Work" button.

Project Name	Download Count
Audio FIR Crossover	35 x
GLAMElab API for Linear Algebra Operations on GPUs	35 x
H.264 Video Encoder	18 x
Large Vocabulary Continuous Speech Recognition	50 x
Quantitative Risk Analysis and Algorithmic Trading Systems	50 x
Canny Edge Detection	3 x
GPU Acceleration Solutions	35 x
Innovative 3D visualization solutions for Oil and Gas	50 x
LIBOR Interest rate Model	50 x
Ray Casting Algebraic Surfaces using the Frustum Form	16 x
Dirac Video Codec	
GPU4Vision	
Jacket: GPU Engine for MATLAB	50 x
Prestack Seismic Data Interaction	100 x
Ray Casting Deformable Models	

www.NVIDIA.com/CUDA

Simple “C” Description For Parallelism



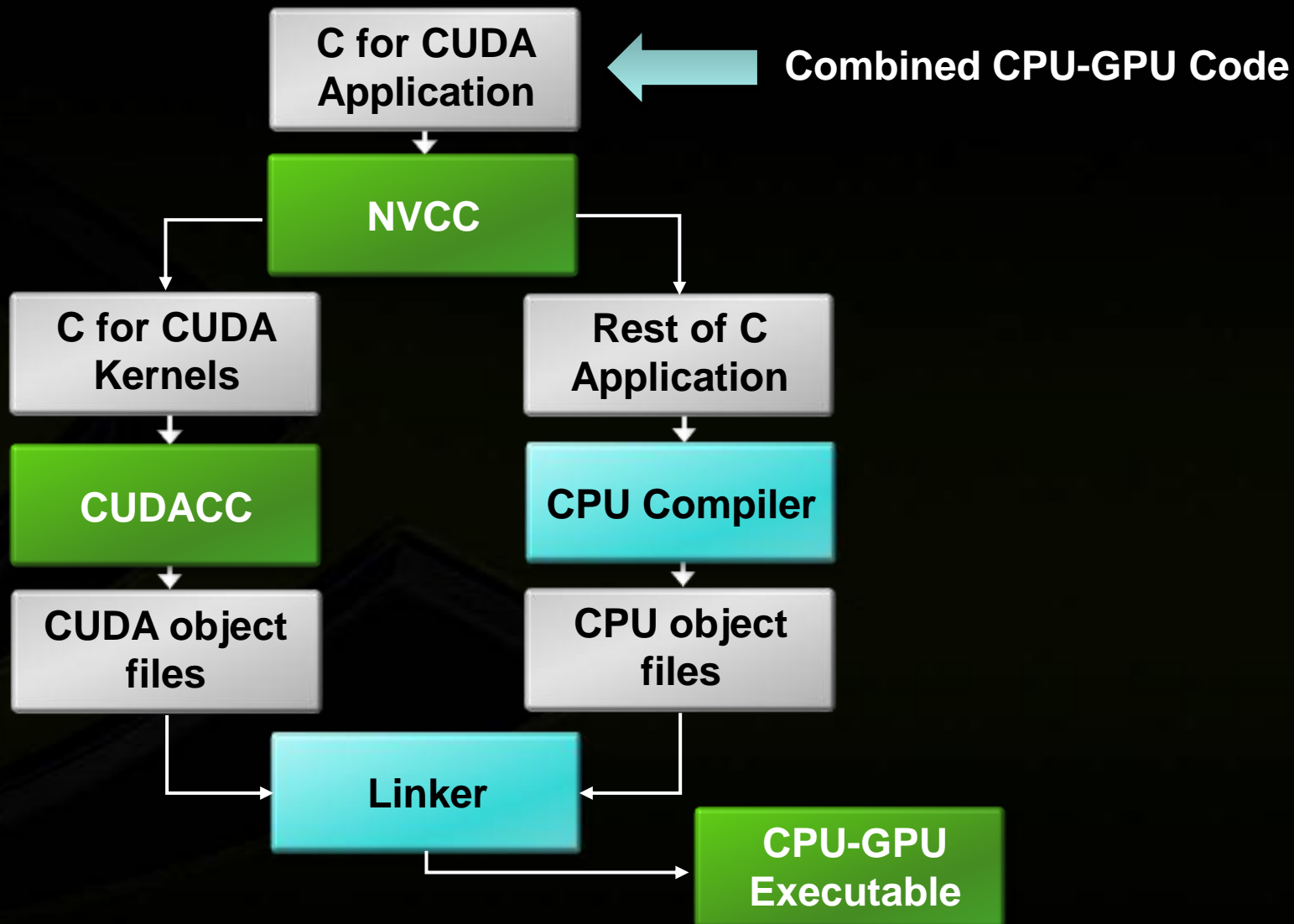
```
void saxpy_serial(int n, float a, float *x, float *y)
{
    for (int i = 0; i < n; ++i)
        y[i] = a*x[i] + y[i];
}
// Invoke serial SAXPY kernel
saxpy_serial(n, 2.0, x, y);
```

Standard C Code

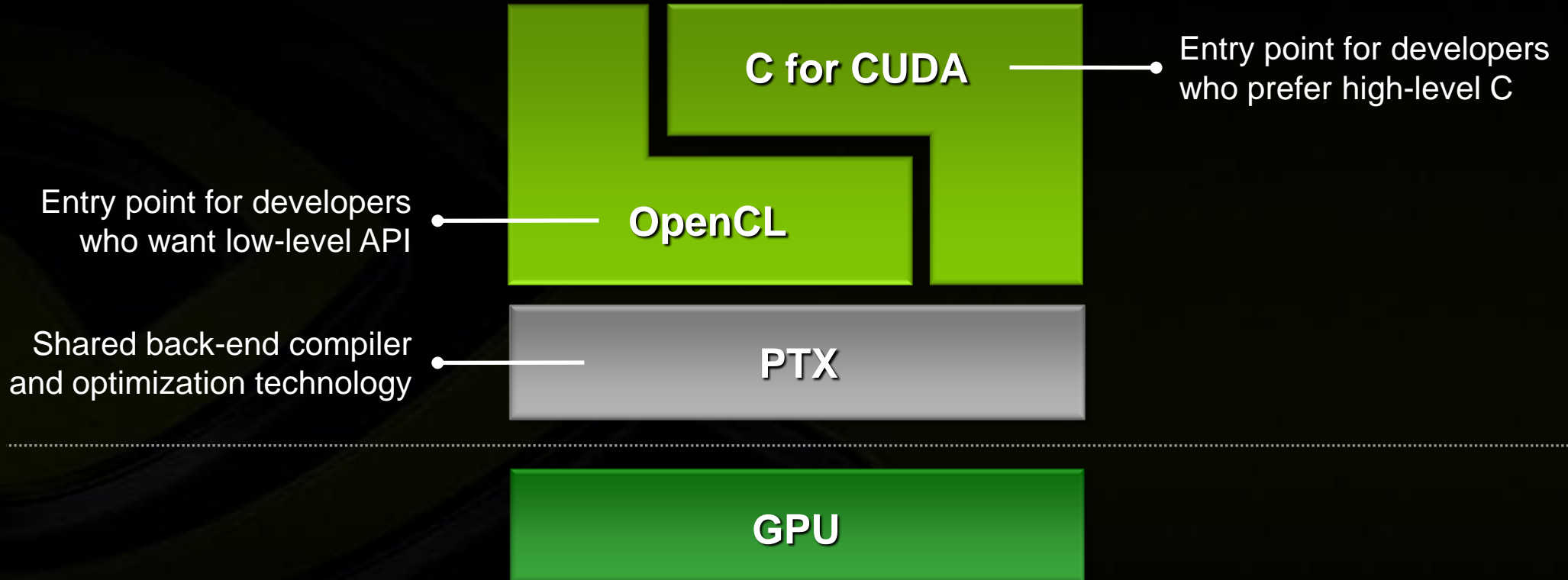
```
__global__ void saxpy_parallel(int n, float a, float *x, float *y)
{
    int i = blockIdx.x*blockDim.x + threadIdx.x;
    if (i < n) y[i] = a*x[i] + y[i];
}
// Invoke parallel SAXPY kernel with 256 threads/block
int nblocks = (n + 255) / 256;
saxpy_parallel<<<nblocks, 256>>>(n, 2.0, x, y);
```

Parallel C Code

Compiling C for CUDA Applications



NVIDIA C for CUDA and OpenCL



Different Programming Styles



- **C for CUDA**

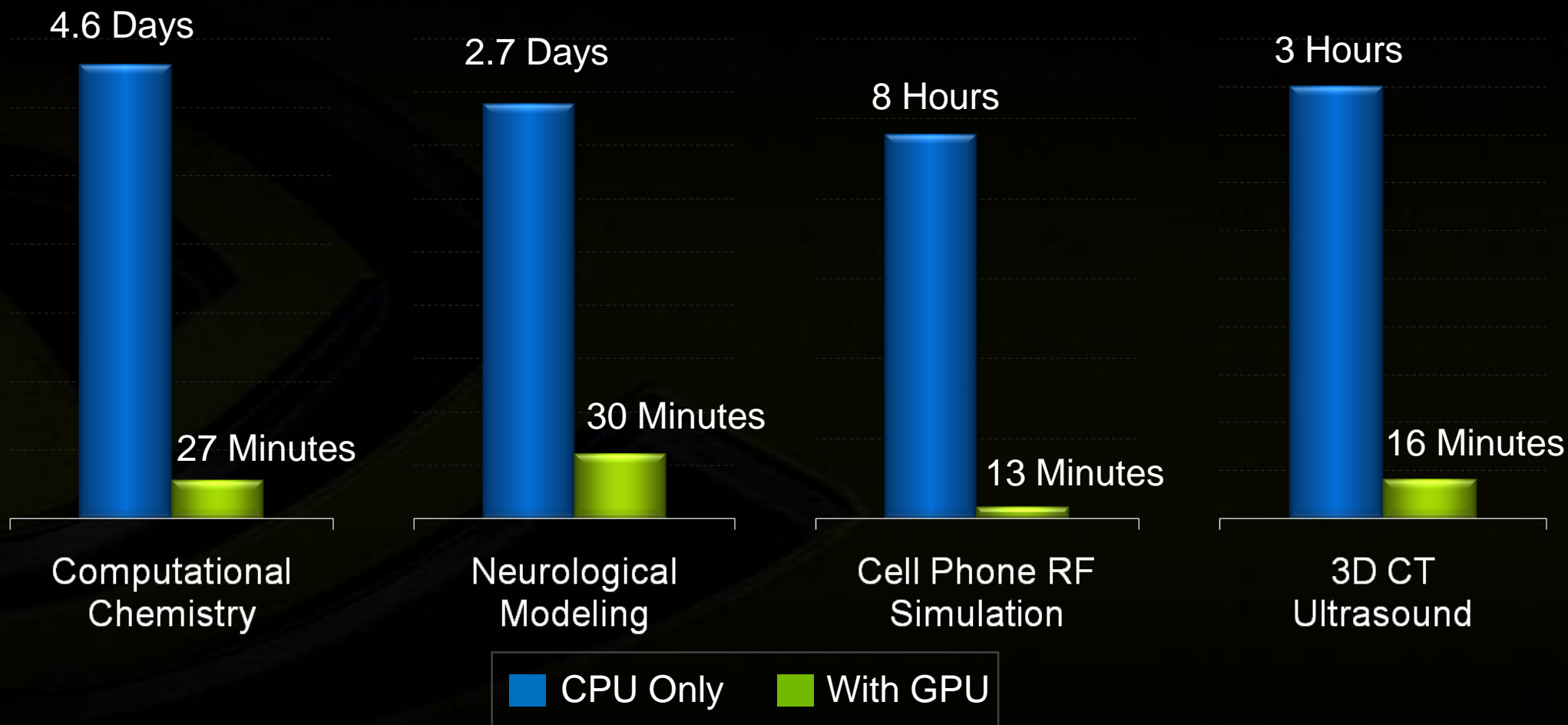
- C with parallel keywords
- C runtime that abstracts driver API
- Memory managed by C runtime (familiar malloc, free)
- Generates PTX
- Low-level “driver” API optionally available

- **OpenCL**

- Hardware API - similar to OpenGL and CUDA driver API
- Memory managed by programmer
- Generates PTX

Application Domains

Accelerating Time to Discovery



Computational Finance

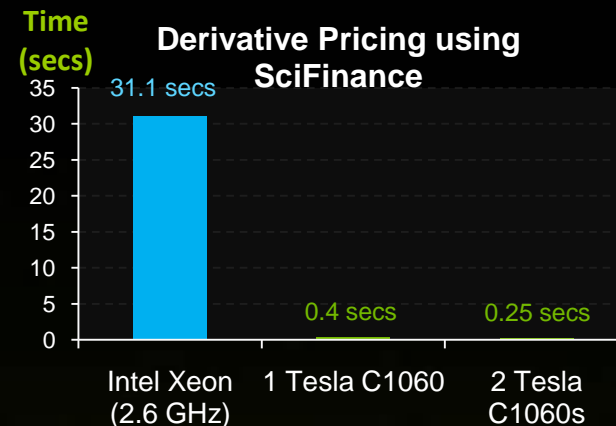


- **Financial Computing Software vendors**

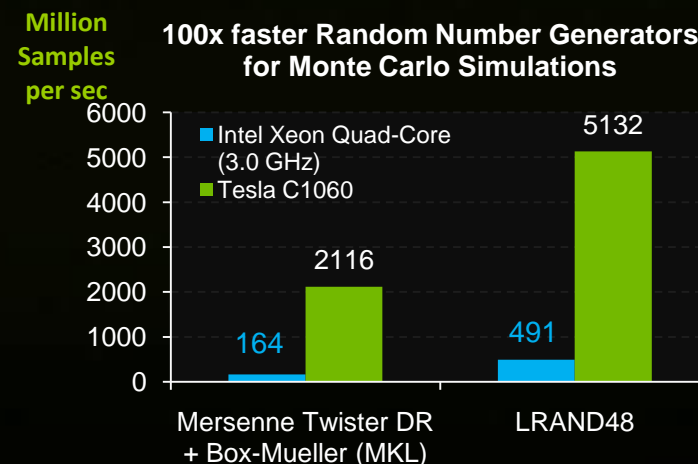
- SciComp : Derivatives pricing modeling
- Hanweck: Options pricing & risk analysis
- Aquamin: 3D visualization of market data
- Exegy: High-volume Tickers & Risk Analysis
- QuantCatalyst: Pricing & Hedging Engine
- Oneye: Algorithmic Trading
- Arbitragis Trading: Trinomial Options Pricing

- **Ongoing work**

- LIBOR Monte Carlo market model
- Callable Swaps and Continuous Time Finance



Source: SciComp



Source: CUDA SDK

Molecular Dynamics

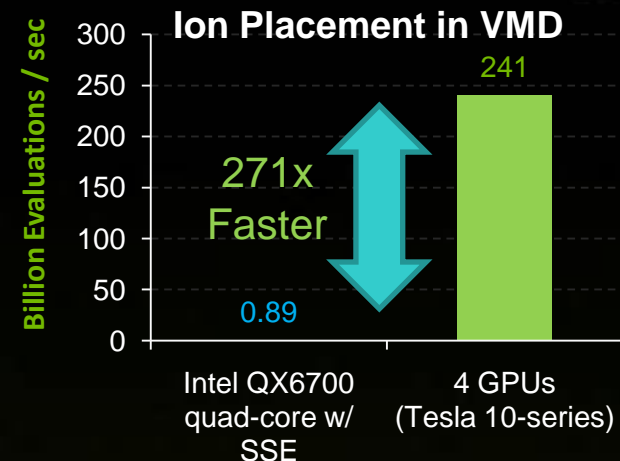


- **Available MD software**

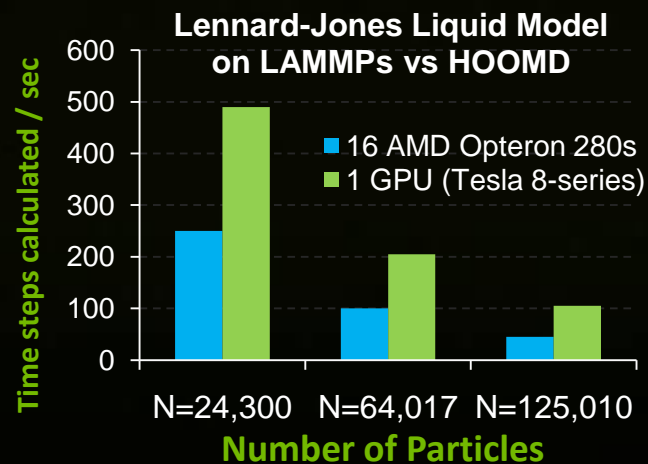
- NAMD / VMD (alpha release)
- HOOMD
- ACE-MD
- MD-GPU

- **Ongoing work**

- LAMMPS
- CHARMM
- GROMACS
- AMBER



Source: Stone, Phillips, Hardy, Schulten



Source: Anderson, Lorenz, Travesset

Quantum Chemistry

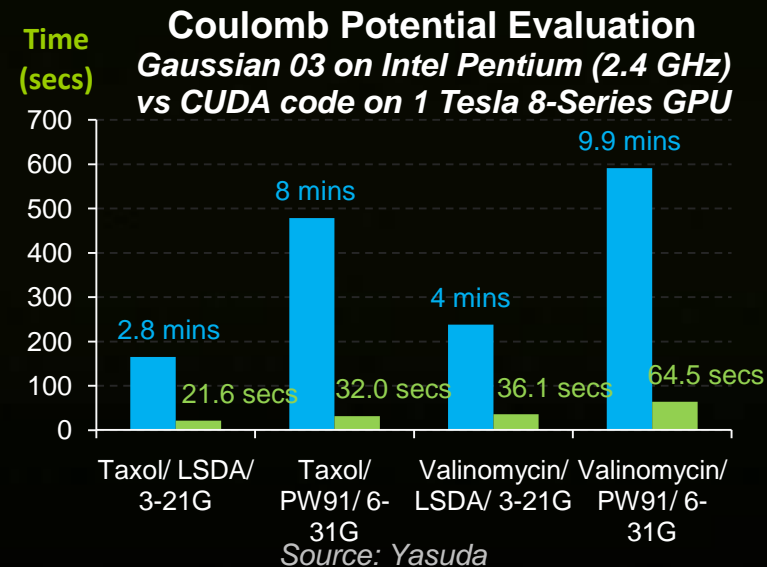
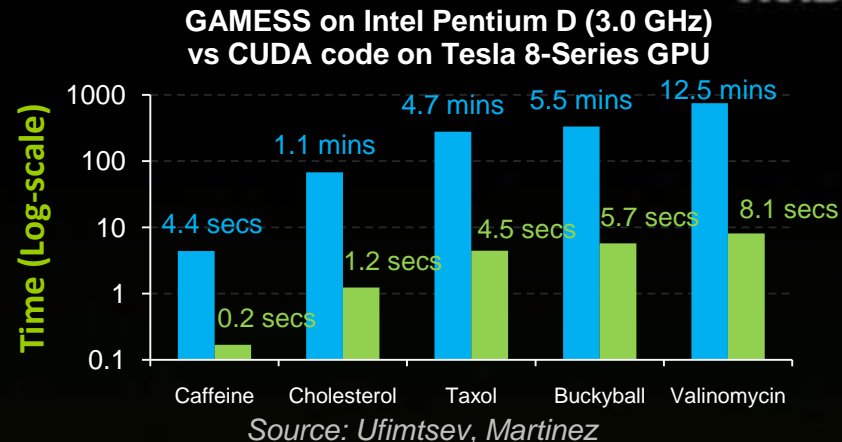


- **Available MD software**

- NAMD / VMD (alpha release)
- HOOMD
- ACE-MD
- MD-GPU

- **Ongoing work**

- LAMMPS
- CHARMM
- Q-Chem
- Gaussian

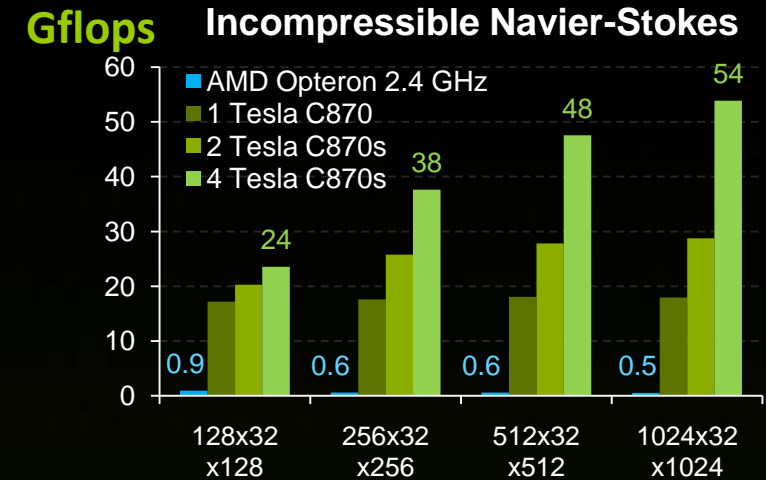


Computational Fluid Dynamics (CFD)

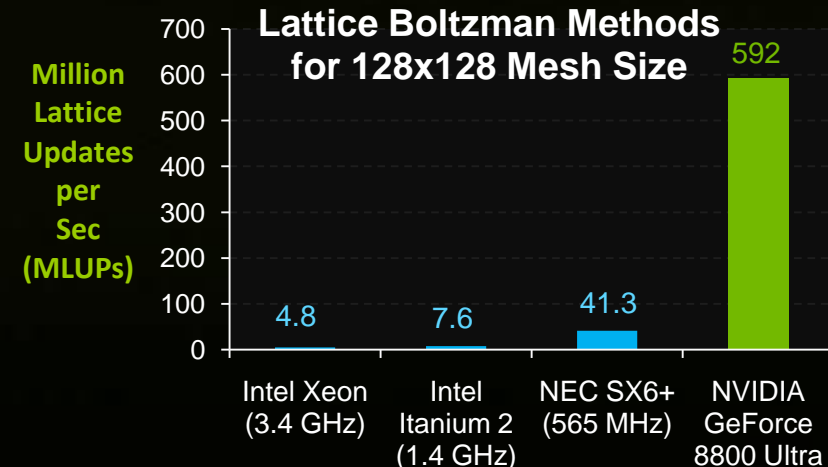


- **Ongoing work**

- Navier-Stokes
- Lattice Boltzman
- 3D Euler Solver
- Weather and ocean modeling



Source: Thibault, Senocak



Source: Tolke, Krafczyk

Electromagnetics / Electrodynamics

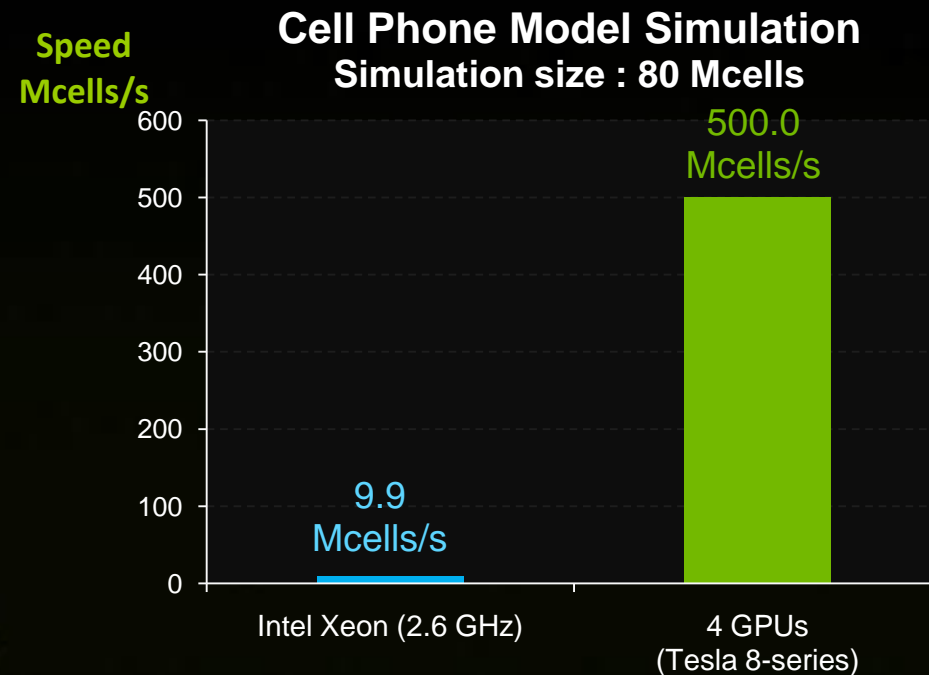


- **FDTD Solvers**

- Acceleware
- EM Photonics
- CUDA Tutorial

- **Ongoing work**

- Maxwell equation solver
- Ring Oscillator (FDTD)
- Particle beam dynamics simulator



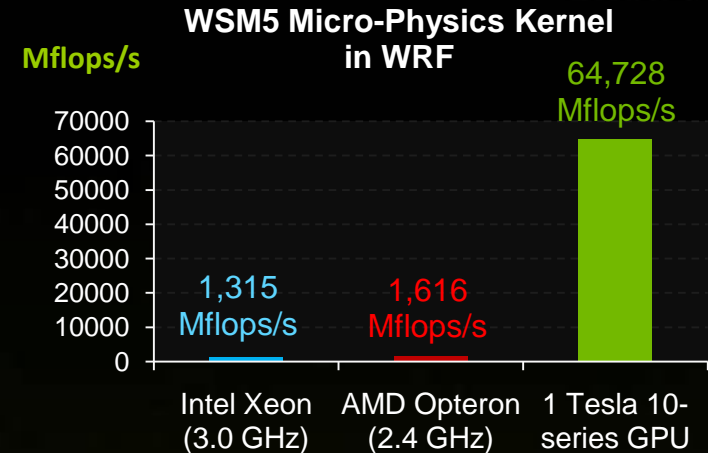
FDTD Acceleration using GPUs

Source: Acceleware

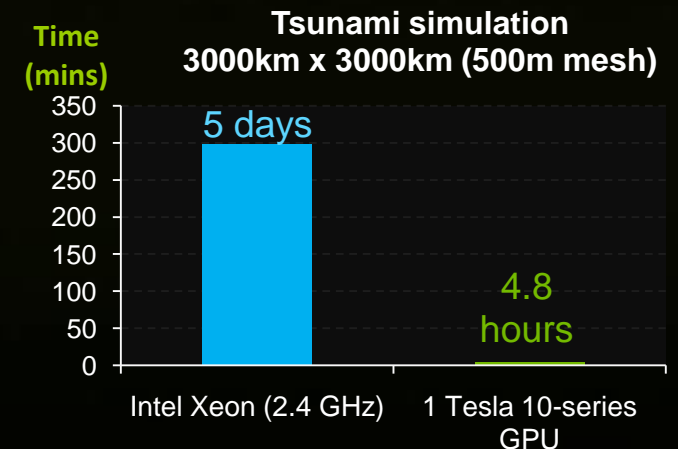
Weather, Atmospheric, & Ocean Modeling



- **CUDA-accelerated WRF available**
 - Other kernels in WRF being ported
- **Ongoing work**
 - Tsunami modeling
 - Ocean modeling
 - Several CFD codes



Source: Michalakes, Vachharajani



Source: Matsuoka, Akiyama, et al

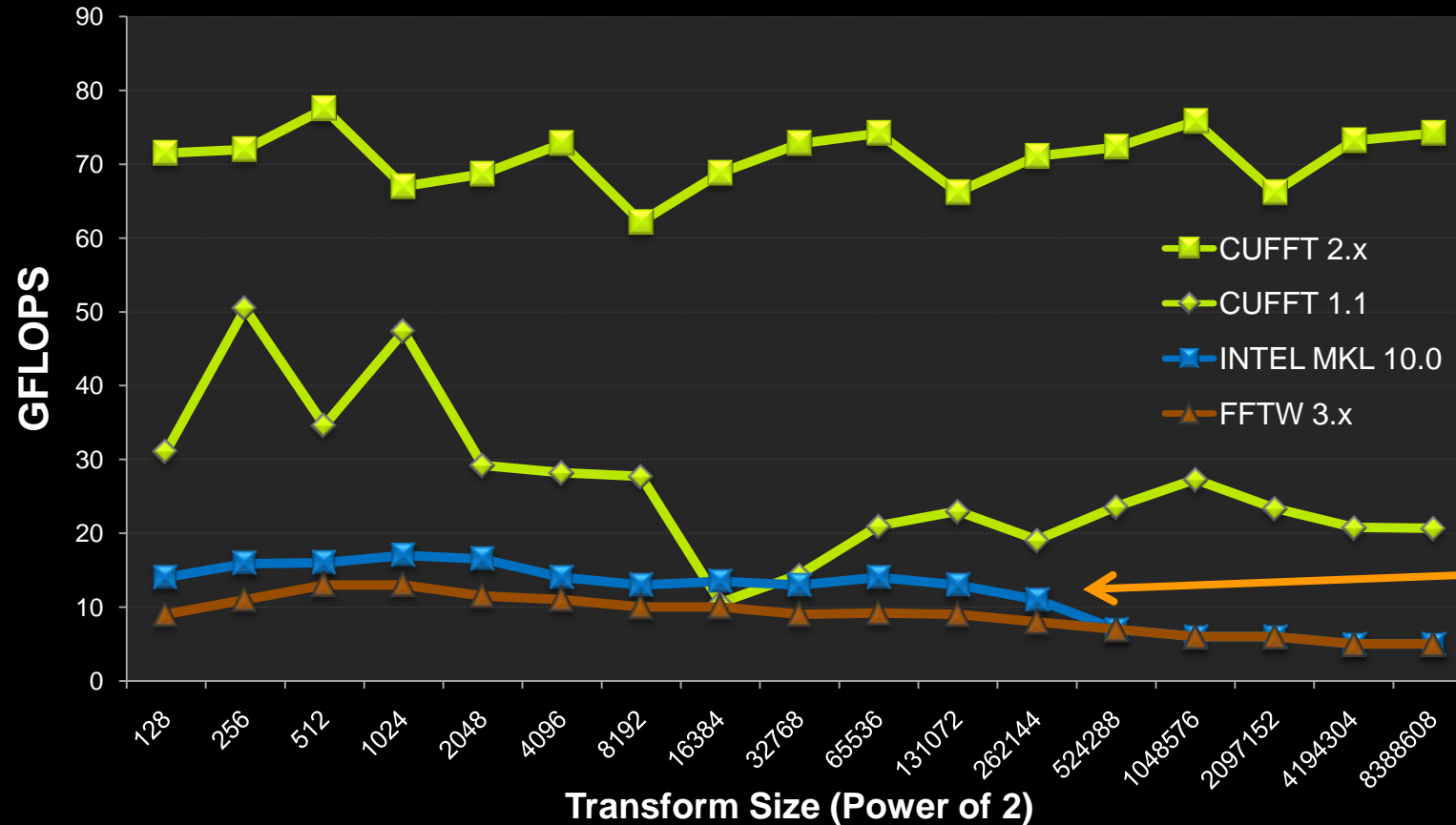
Libraries

FFT Performance: CPU vs GPU (8-Series)



1D Fast Fourier Transform On CUDA

NVIDIA Tesla C870 GPU (8-series GPU)
Quad-Core Intel Xeon CPU 5400 Series 3.0GHz,
In-place, complex, single precision

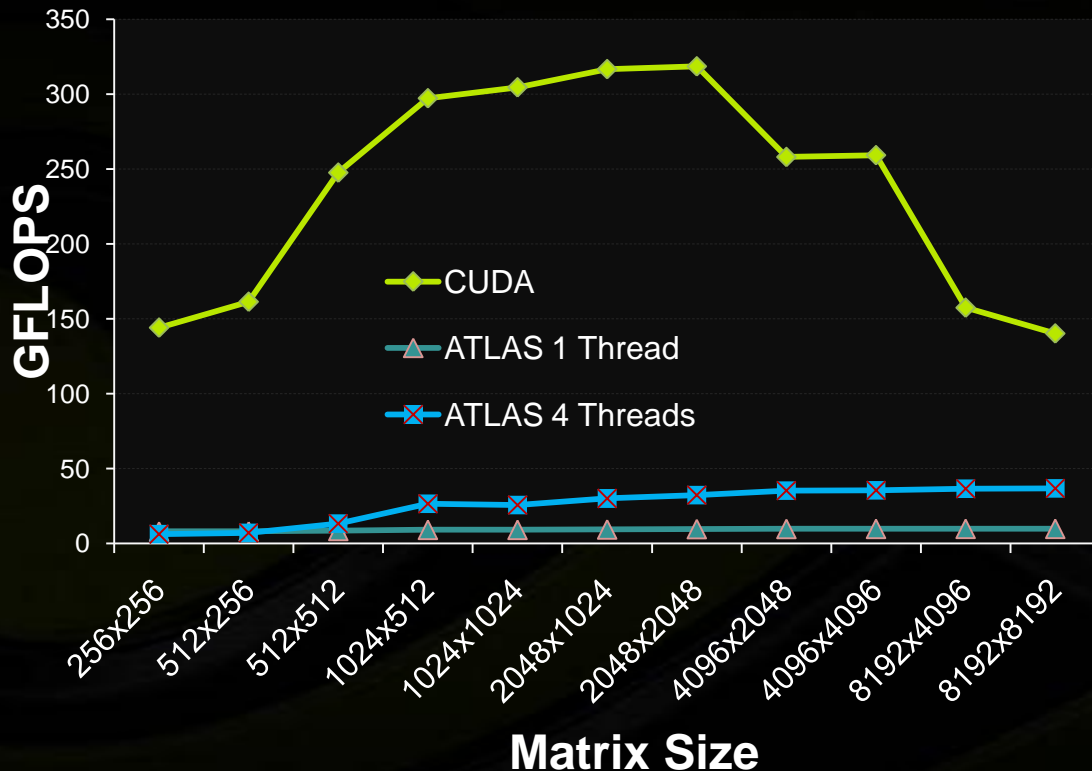


- Intel FFT numbers calculated by repeating same FFT plan
- Real FFT performance is ~10 GFlops

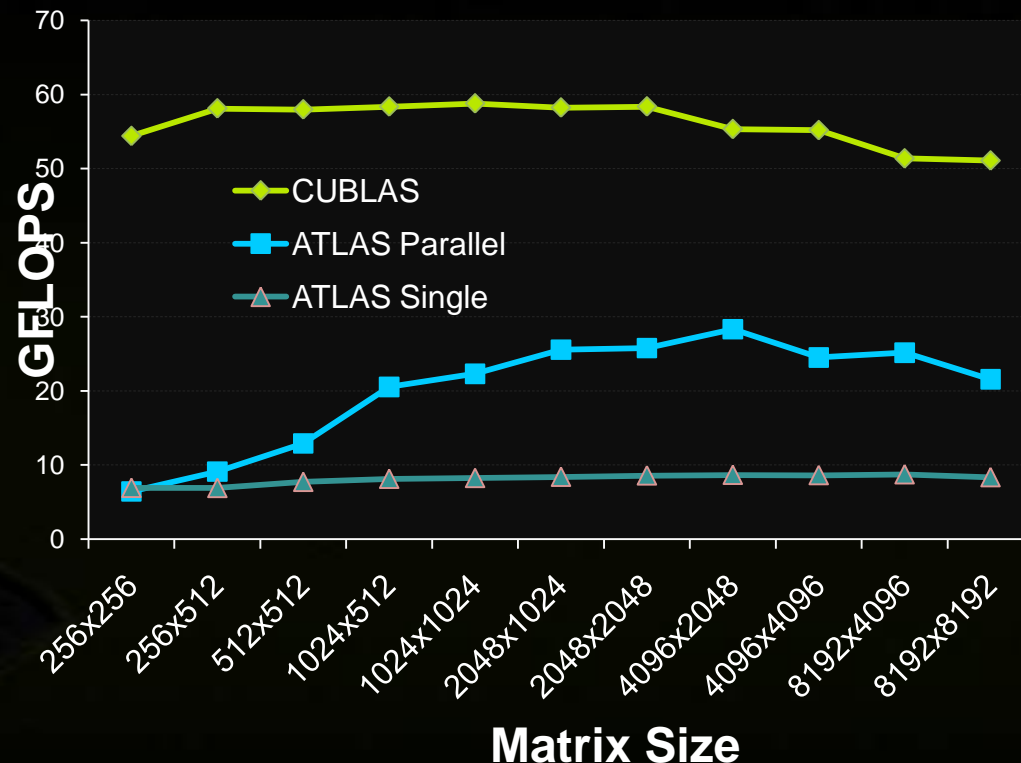
BLAS: CPU vs GPU (10-series)



Single Precision BLAS: SGEMM

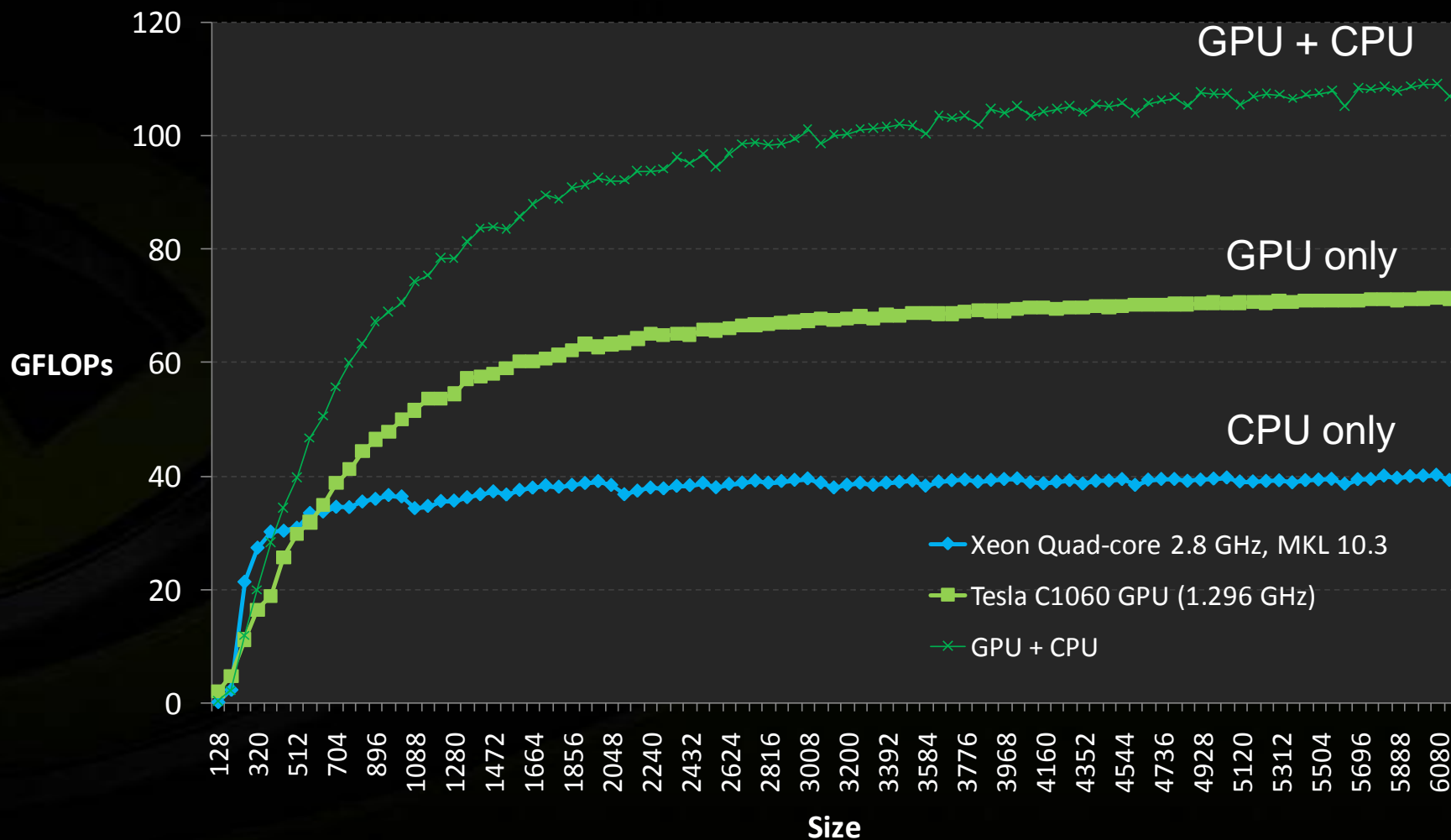


Double Precision BLAS: DGEMM



CUBLAS: CUDA 2.0, Tesla C1060 (10-series GPU)
ATLAS 3.81 on Dual 2.8GHz Opteron Dual-Core

GPU + CPU DGEMM Performance

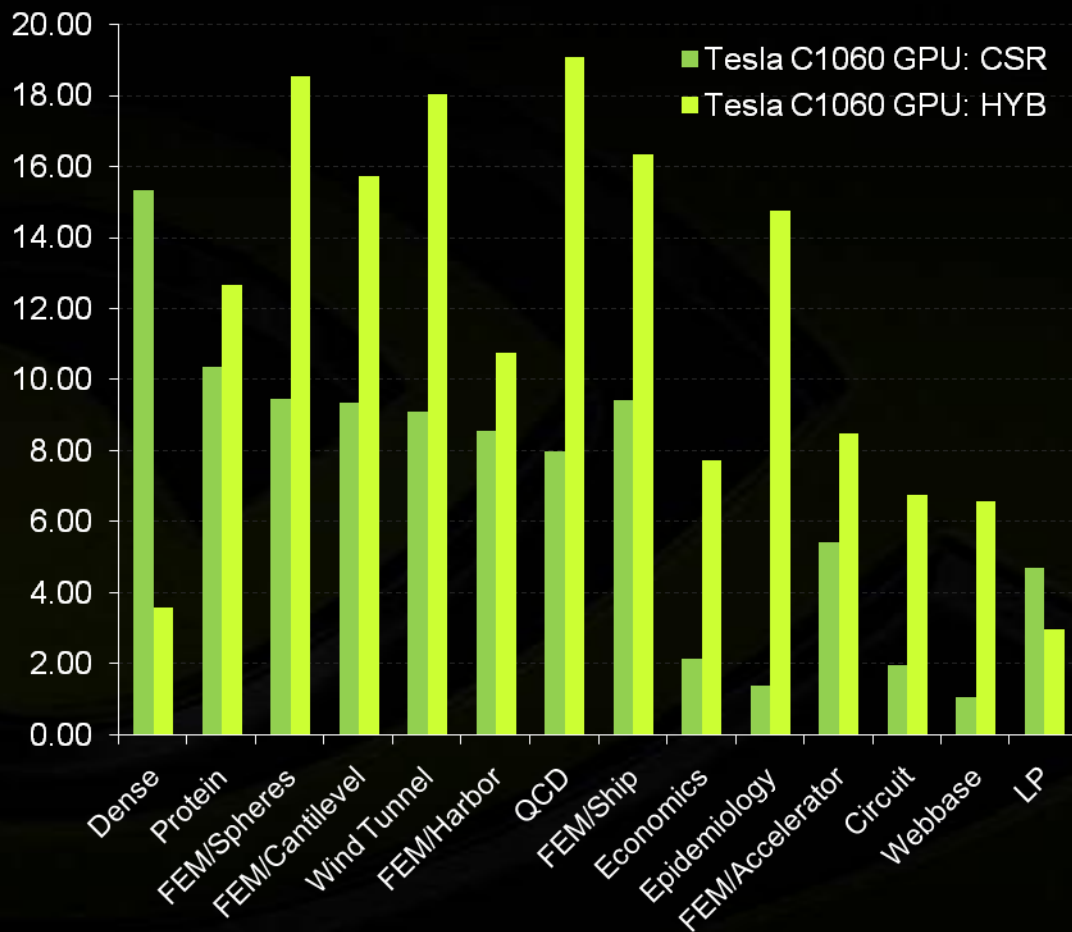


Results: Sparse Matrix-Vector Multiplication (SpMV) on CUDA



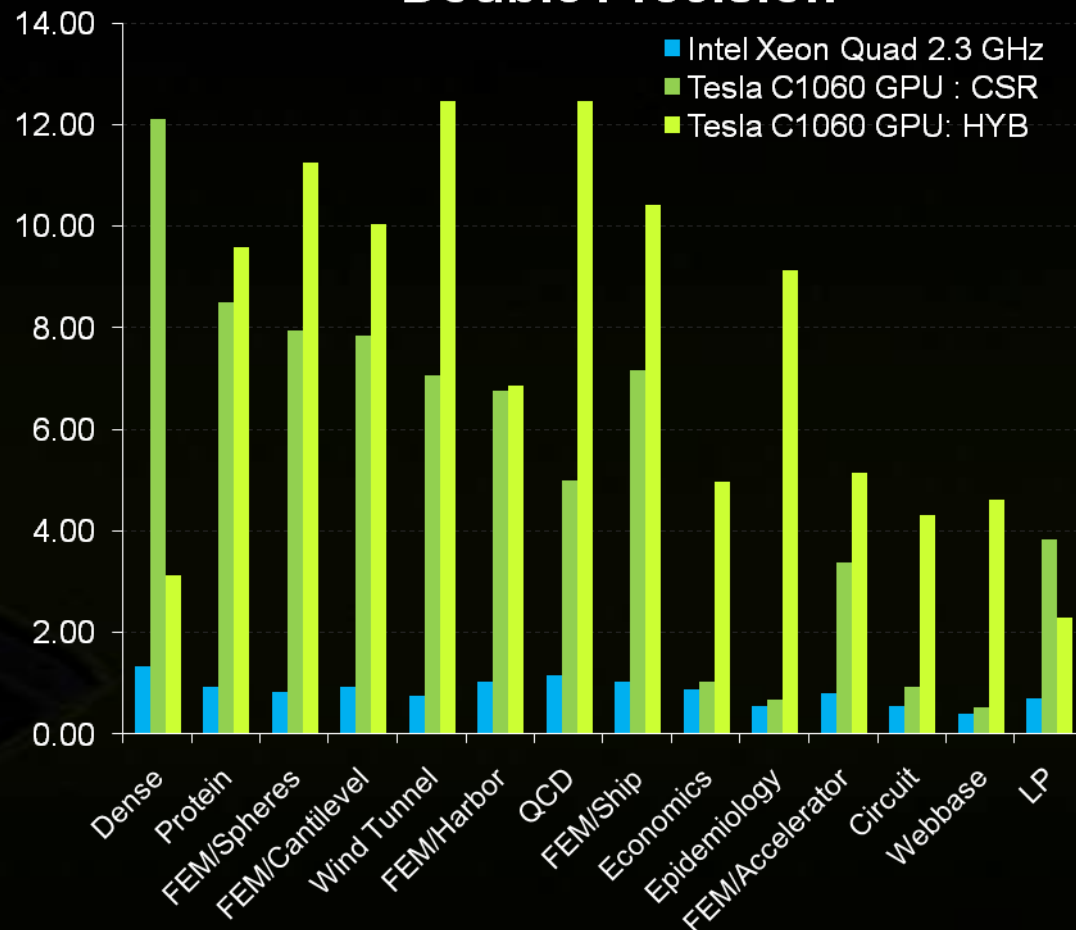
GFLOPS

Single Precision



GFLOPS

Double Precision



CPU Results from "Optimization of Sparse Matrix-Vector Multiplication on Emerging Multicore Platforms", Williams et al, Supercomputing 2007

Questions?

mharris@nvidia.com

<http://www.nvidia.com/CUDA>

More on Tesla Hardware

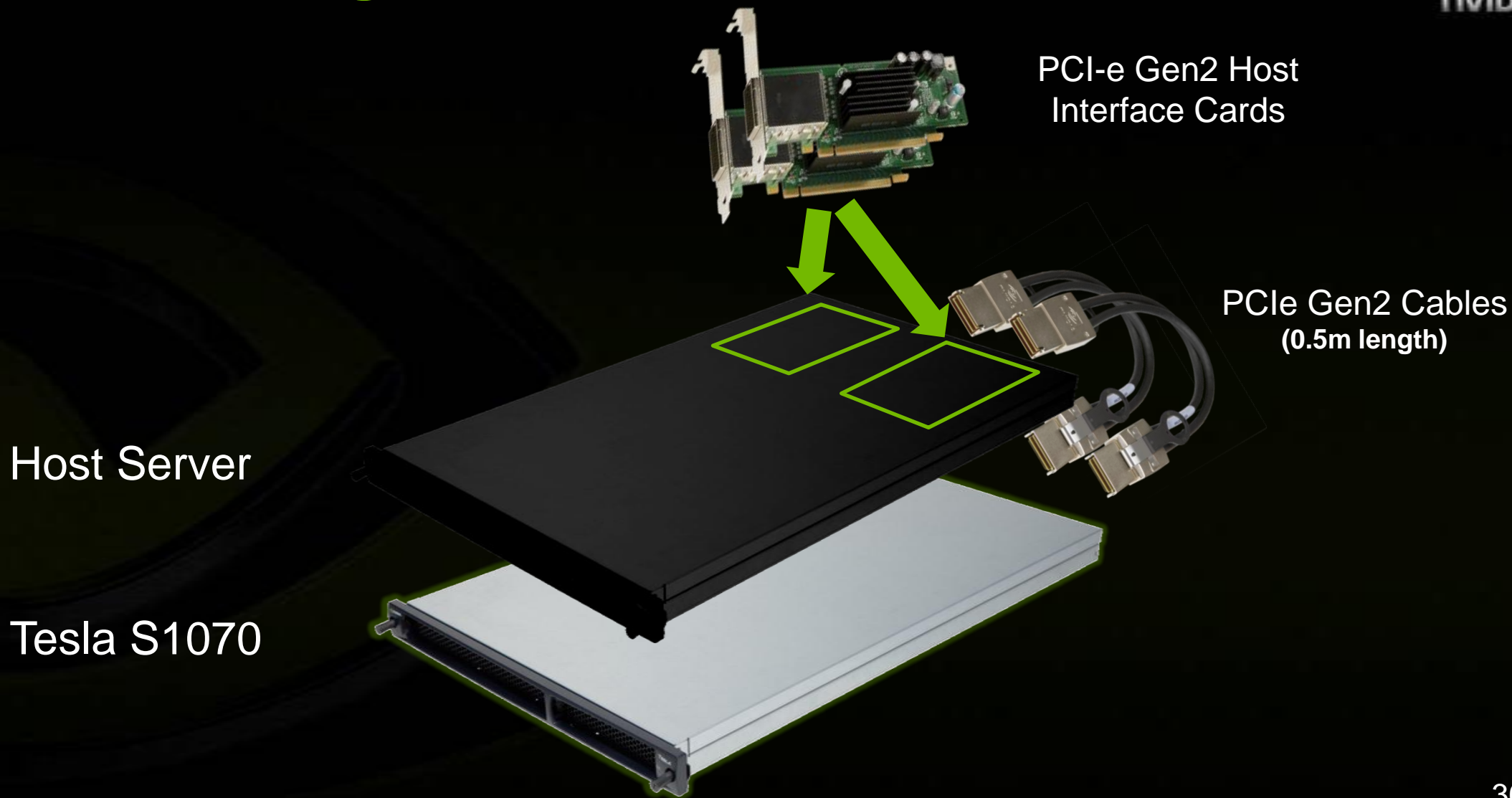
IEEE Compliant Double Precision Floating Point



NVIDIA Tesla T10

Precision	IEEE 754
Rounding modes for FADD and FMUL	All 4 IEEE, round to nearest, zero, inf, -inf
Denormal handling	Full speed
NaN support	Yes
Overflow and Infinity support	Yes
FMA	Yes
Square root	Software with low-latency FMA-based convergence
Division	Software with low-latency FMA-based convergence
Reciprocal estimate accuracy	24 bit
Reciprocal sqrt estimate accuracy	23 bit
$\log_2(x)$ and 2^x estimates accuracy	23 bit

Connecting Tesla S1070 to Host Servers



NVIDIA: Leadership in GPU computing



100s of Apps on CUDA Zone



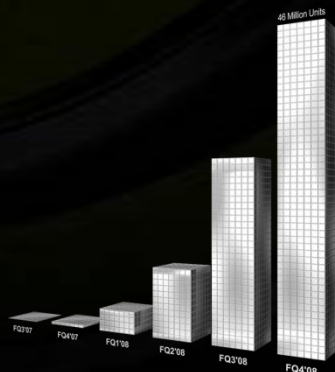
30+ CUDA GPU clusters



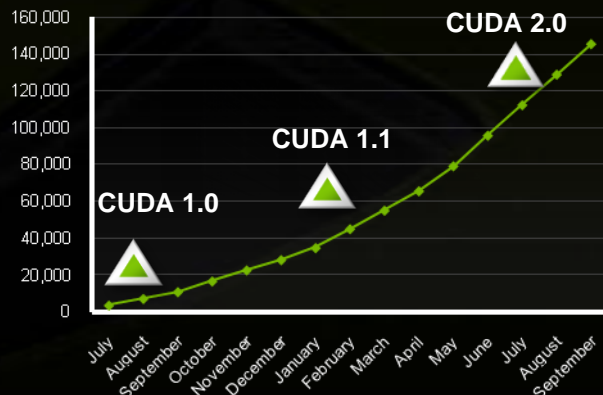
50+ Universities Teaching CUDA
750+ research papers

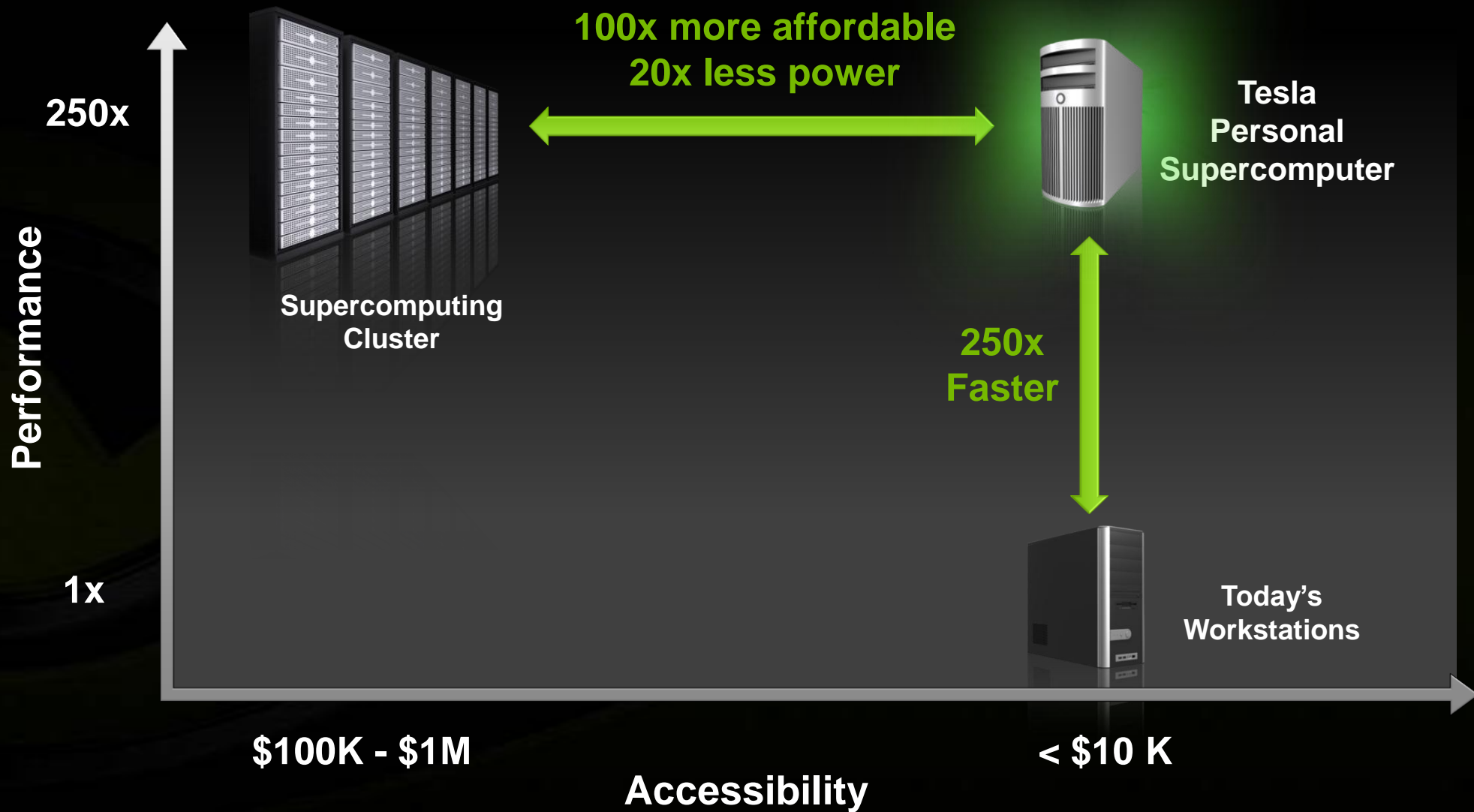
Duke	Northeastern
Erlangen	Oregon State
ETH Zurich	Pennsylvania
Georgia Tech	Polimi
Grove City College	Purdue
Harvard	Santa Clara
IISc Bangalore	Stanford
IIT Hyderabad	Stuttgart
IIT	Suny
Illinois	Tokyo
INRIA	TU-Vienna
Iowa	USC
ITESM	Utah
Johns Hopkins	Virginia
Kent State	Washington
Kyoto	Waterloo
Lund	Western Australia
Maryland	Williams College
McGill	Wisconsin
MIT	Yonsei
North Carolina	

100 M CUDA enabled GPUs
25,000+ active developers

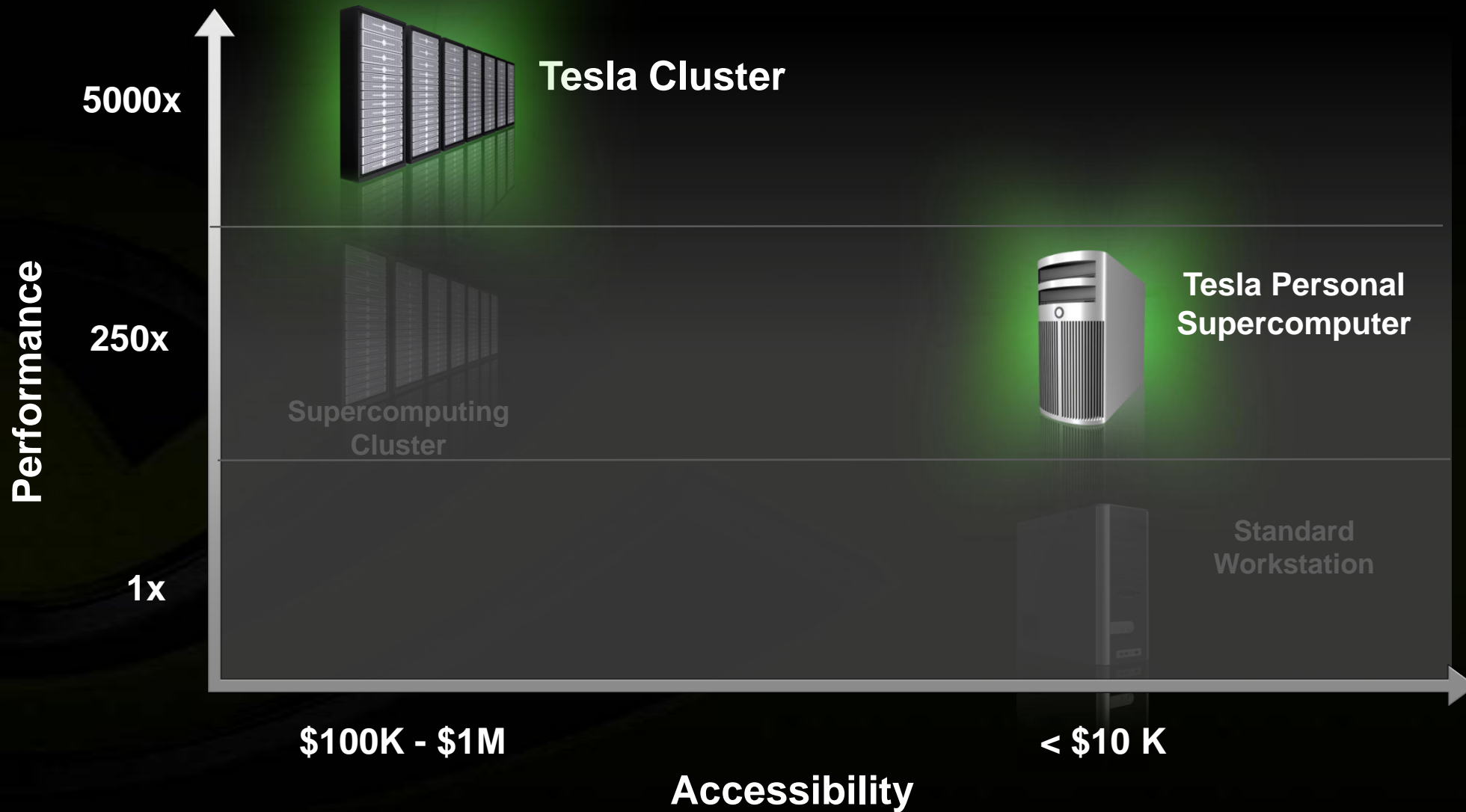


150K CUDA compiler downloads





New GPU-based High-Performance Computing Landscape



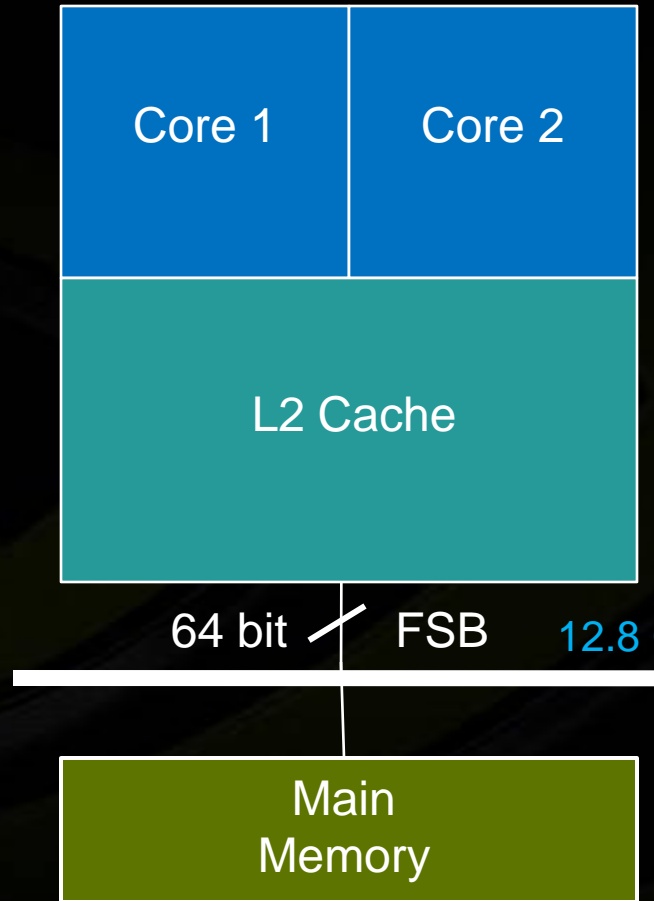
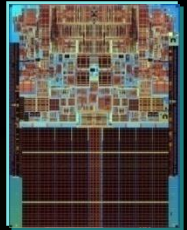
More Than 250 Customers / ISVs



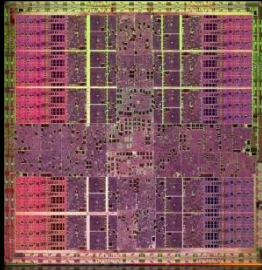
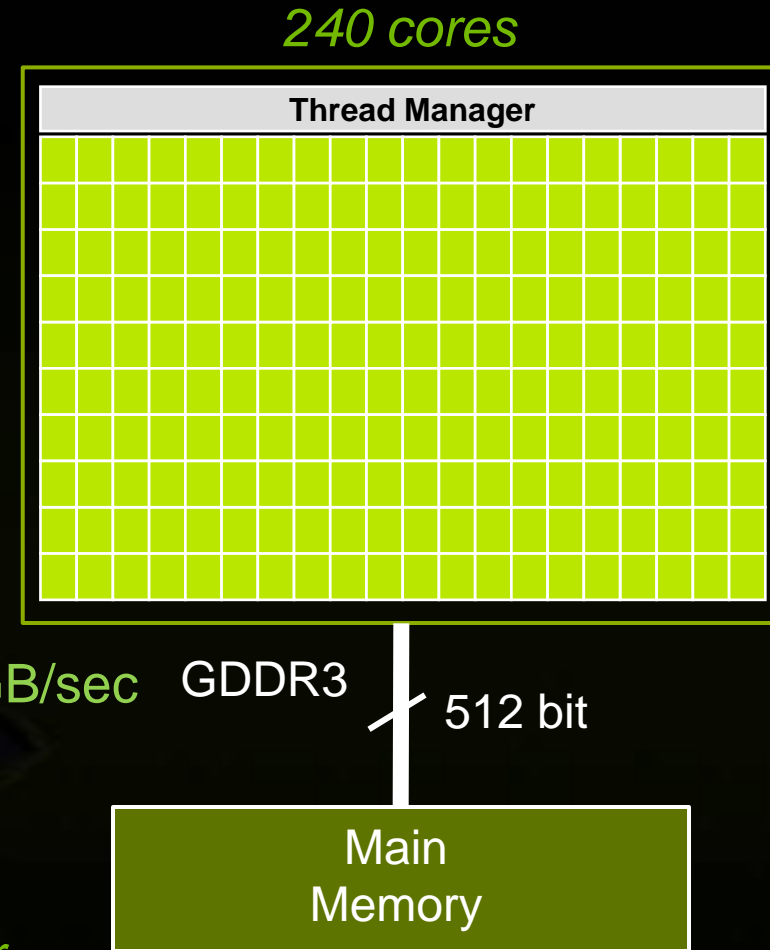
Life Sciences & Medical Equipment		Productivity / Misc	Oil and Gas	EDA	Manufacturing	Finance	CAE / Numerics	Communication
Max Planck FDA Robarts Research Medtronic AGC Evolved machines Smith-Waterman DNA sequencing AutoDock NAMD/VMD Folding@Home Howard Hughes Medical CRIBI Genomics	GE Healthcare Siemens Techniscan Boston Scientific Eli Lilly Silicon Informatics Stockholm Research Harvard Delaware Pittsburg ETH Zurich Institute Atomic Physics	CEA WRF Weather Modeling OptiTEx Tech-X Elemental Technologies Dimensional Imaging Manifold Digisens General Mills Rapidmind MS Visual Studio Rhythm & Hues xNormal Elcomsoft LINZIK	Hess TOTAL CGG/Veritas Chevron Headwave Acceleware Seismic City P-Wave Seismic Imaging Mercury Computer ffA	Synopsys Nascentric Gauda CST Agilent	Renault Boeing	Symcor Level 3 SciComp Hanweck Quant Catalyst RogueWave BNP Paribas	The Mathworks Wolfram National Instruments Access Analytics Tech-x RIKEN SOFA	Nokia RIM Philips Samsung LG Sony Ericsson NTT DoCoMo Mitsubishi Hitachi Radio Research Laboratory US Air Force

GPUs: Better Architecture for Computing

CPU: Memory Bandwidth Bottleneck



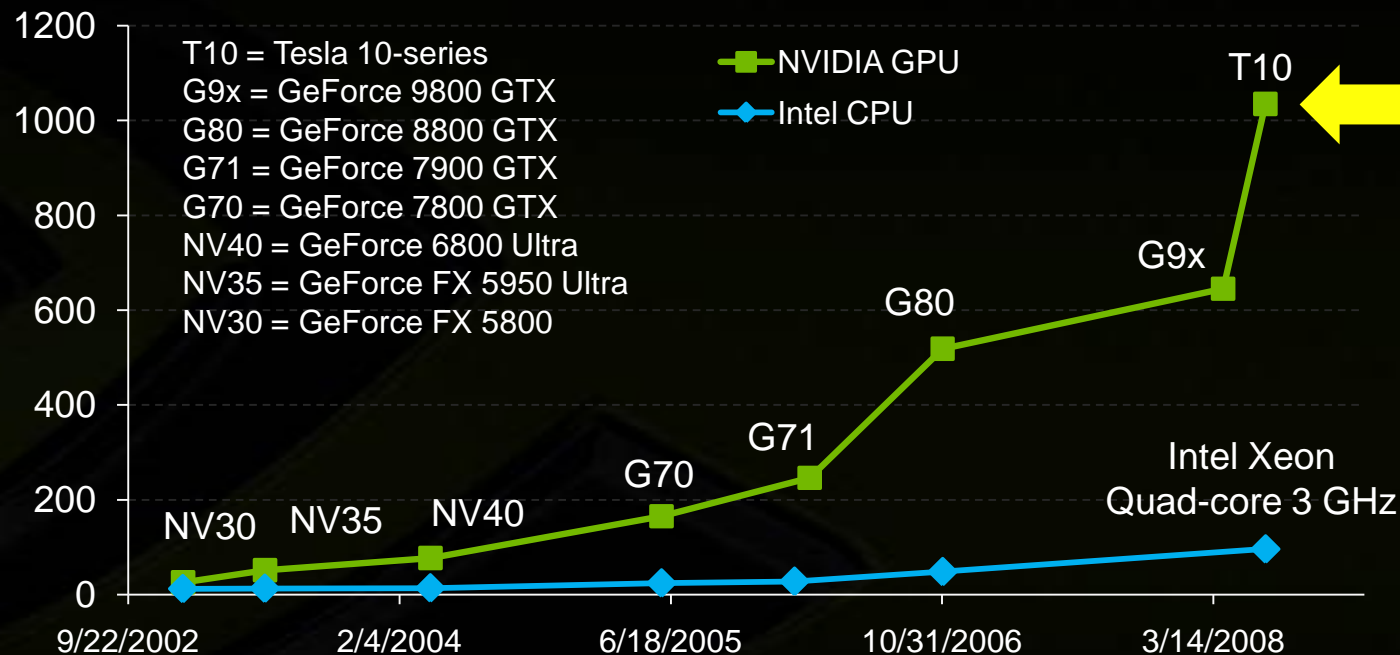
*8x faster
interface*



NVIDIA's GPUs : Ever Increasing Performance





GFlops



**Double
Precision
debut**

CPU vs GPU 1U Comparison



	CPU 1U Server	Tesla 1U System
Product		
	2x Quad Xeon: 3 GHz	Quad-GPU Tesla 1U
# of Cores	8 CPU cores	4 GPUs: 960 cores
Single Precision Flops	0.192 Teraflop	4.14 Teraflops → 21x higher Gflop
Double Precision Flops	96 GFlops	346 GFlops → 3.6x higher Gflop
Typical 1U System Power	670 W	700 W

Note: Current top Intel GFlops: Xeon Harpertown X5482 @ 3.2 GHz is 102.4 Gflops (> \$1000 CPU)