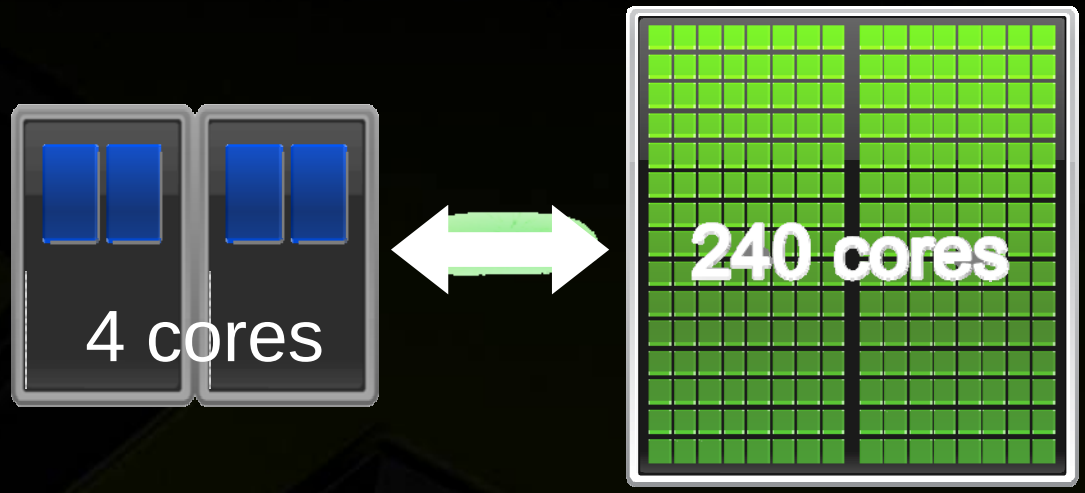
CUDA & Tesla GPU Computing

Accelerating Scientific Discovery

March 2009

What is GPU Computing?



Computing with CPU + GPU

Heterogeneous Computing

“Homemade” Supercomputer Revolution

16 GPUs

MIT, Harvard

8 GPUs

University of Antwerp Belgium

4 GPUs

TU Braunschweig, Germany

3 GPUs

University of Illinois

3 GPUs

Yonsei University, Korea

3 GPUs

Rice University

3 GPUs

University of Cambridge, UK

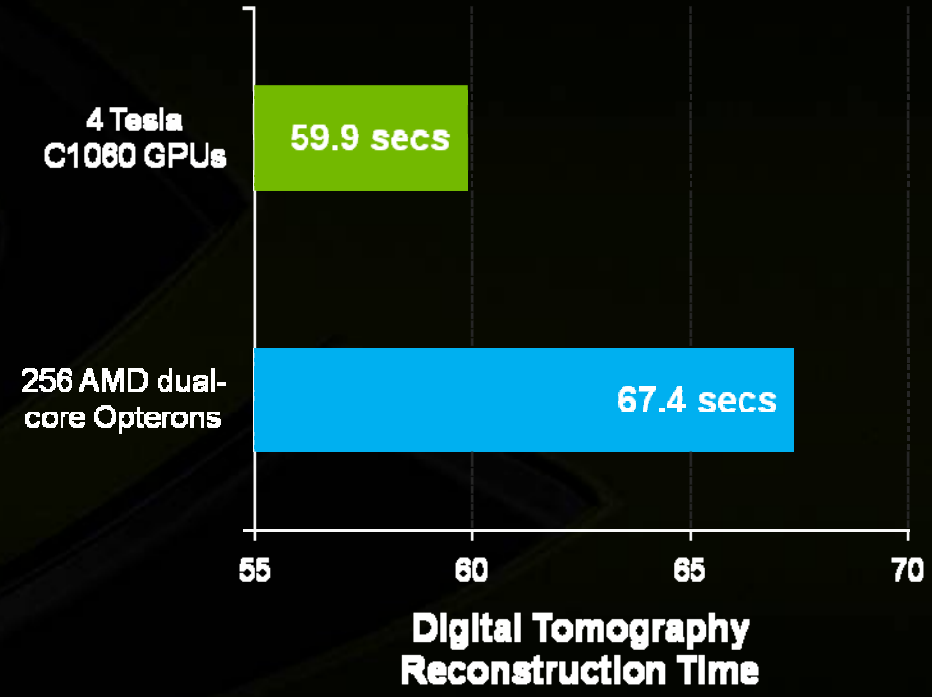
2 GPUs

Georgia Tech

GPUs: Turning Point in Supercomputing

Desktop beats Cluster

CalcUA
$5 Million



Tesla Personal Supercomputer
$10,000

Source: University of Antwerp, Belgium

Introducing the *Tesla Personal Supercomputer*

Supercomputing Performance

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

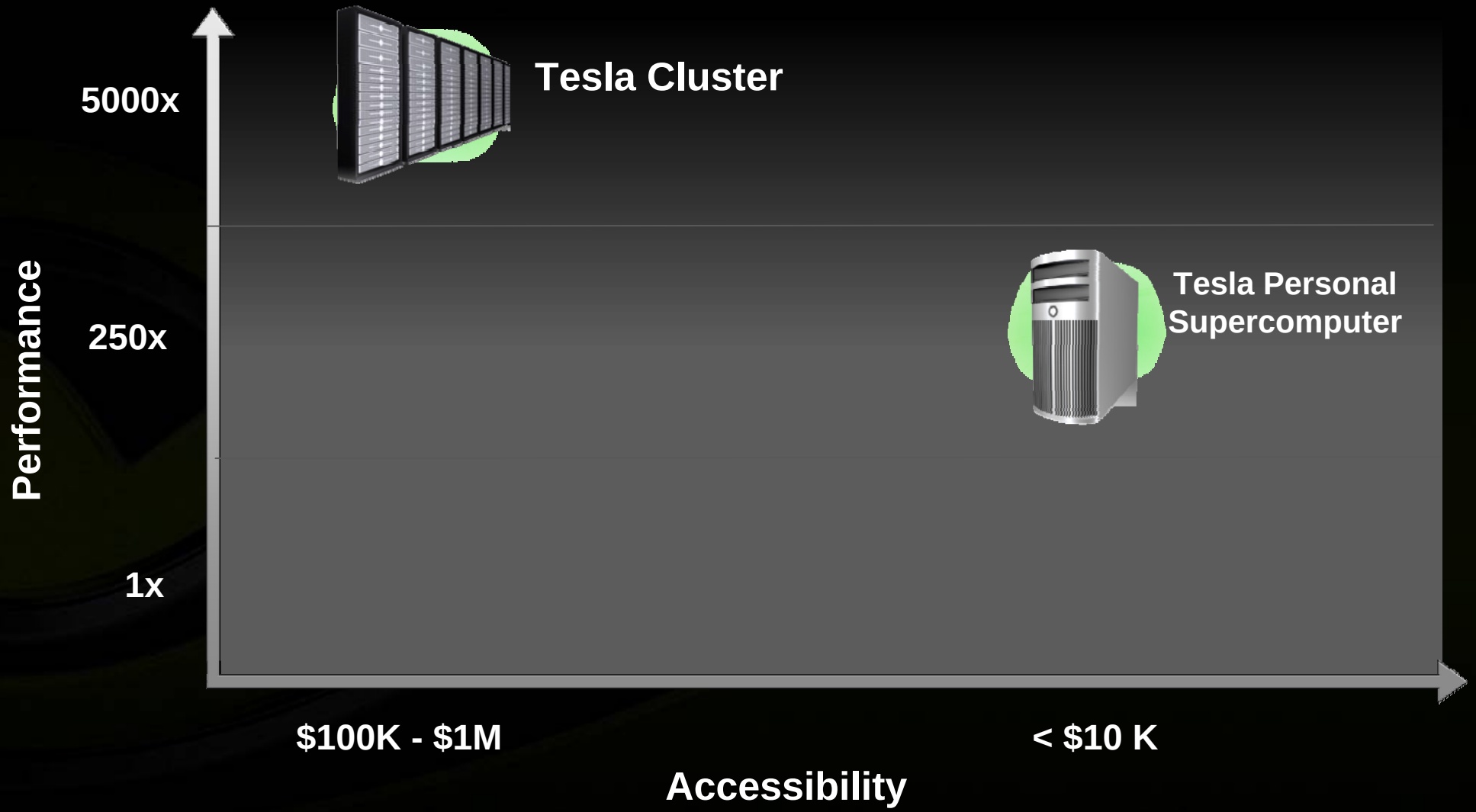
Personal

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

Accessible

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

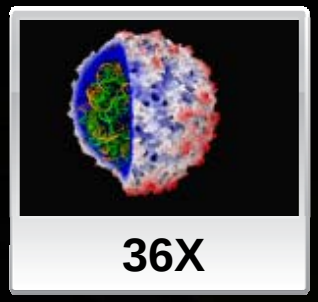
New GPU-based High-Performance Computing Landscape



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

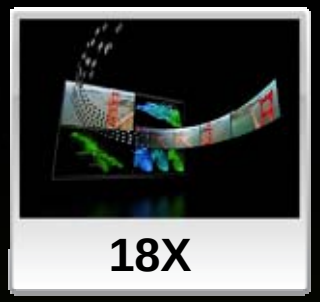
146X

Medical Imaging
U of Utah



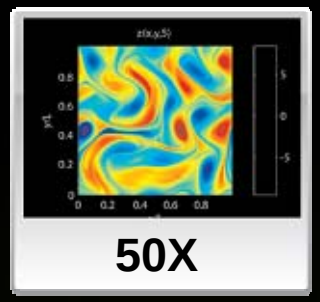
36X

Molecular Dynamics
U of Illinois, Urbana



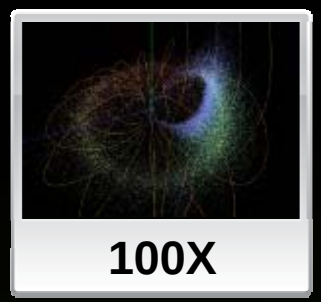
18X

Video Transcoding
Elemental Tech



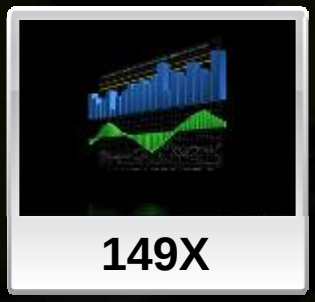
50X

Matlab Computing
AccelerEyes



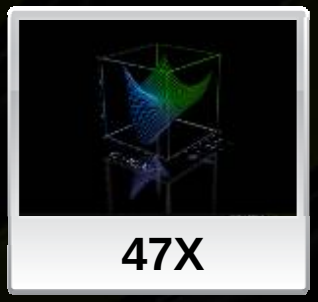
100X

Astrophysics
RIKEN



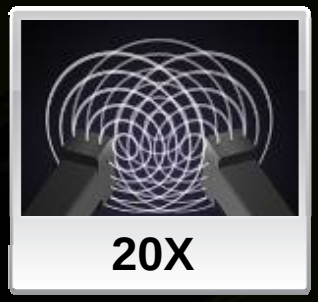
149X

Financial simulation
Oxford



47X

Linear Algebra
Universidad Jaime

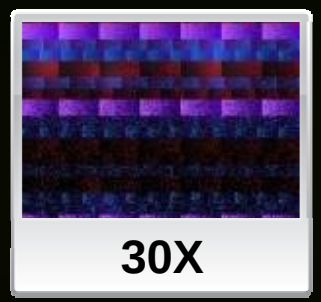


20X

3D Ultrasound
Techniscan

130X

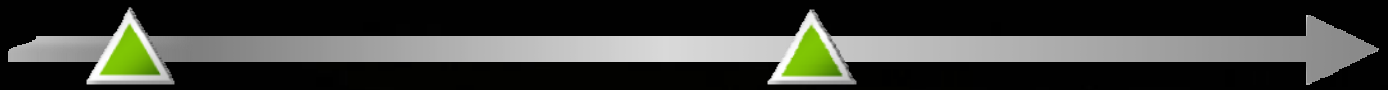
Quantum Chemistry
U of Illinois, Urbana



30X

Gene Sequencing
U of Maryland

Parallel vs Sequential Architecture Evolution



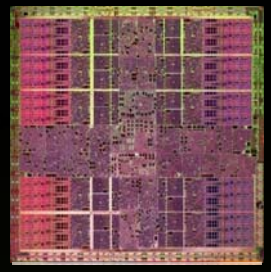
ILLIAC IV

Cray-1

Maspar

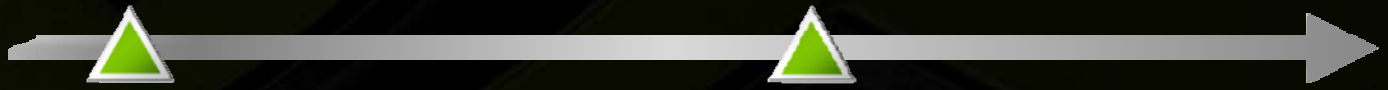
Thinking Machines

Blue Gene



Many-Core
GPUs

High Performance Computing Architectures



DEC PDP-1

Intel 4004

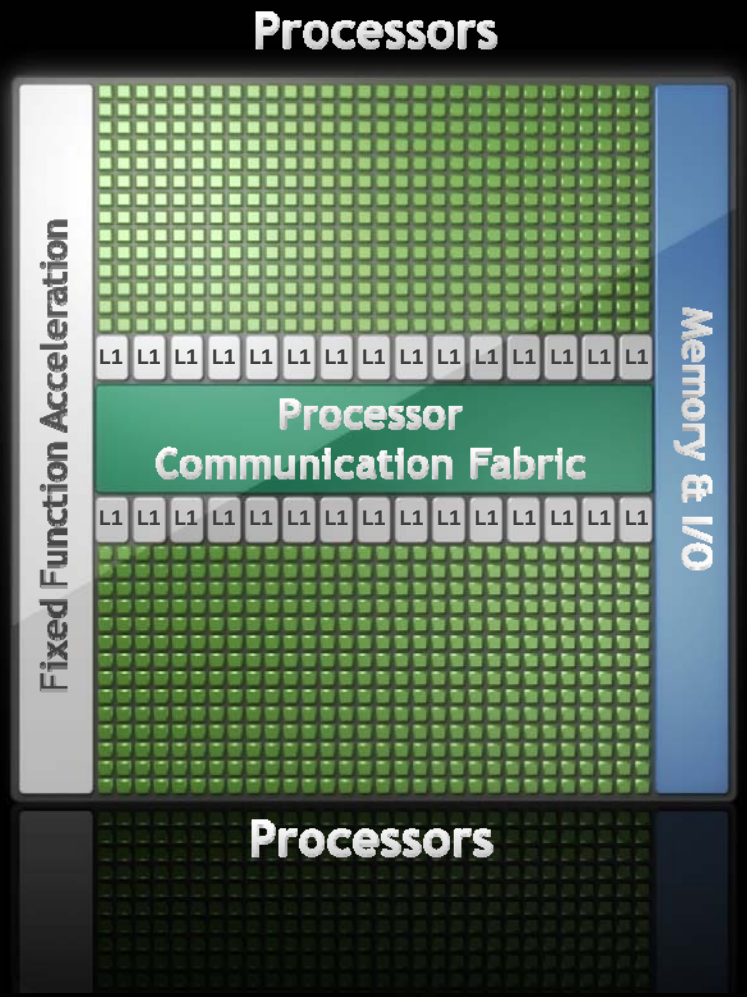
IBM System 360

VAX

IBM POWER4

Multi-Core
x86

Database, Operating System Sequential Architectures



NVIDIA Tesla 10-Series GPU

Massively parallel, many core architecture

240 Processor Cores

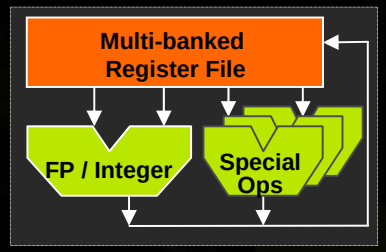
1 Teraflops - 1,000 times Cray X-MP

IEEE Compliant Double Precision Floating Point

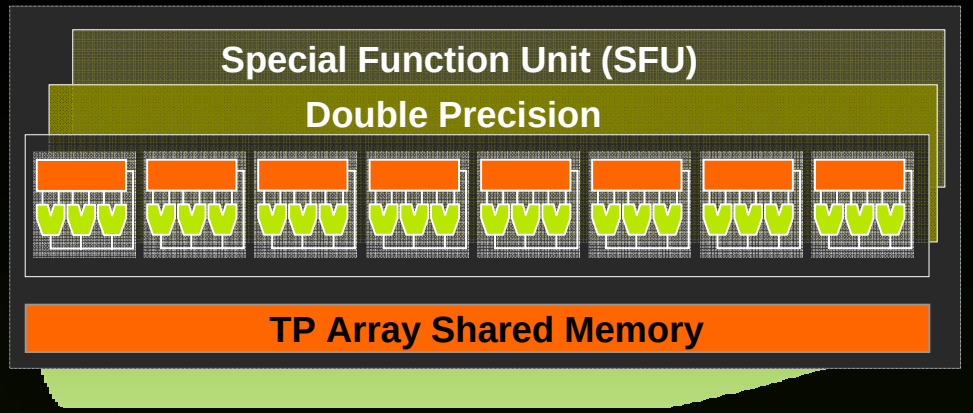
Designed for Scientific Computing

Tesla T10 GPU: 240 Processor Cores

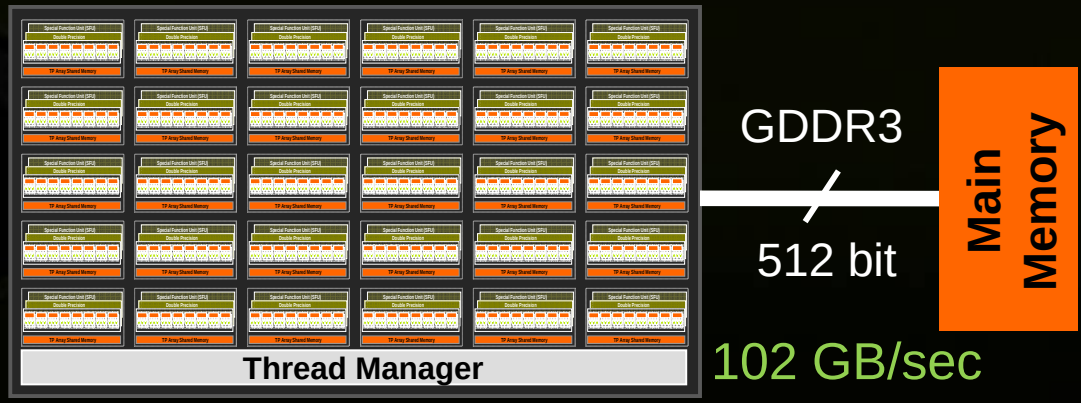
Thread Processor (TP)



Thread Processor Array (TPA)



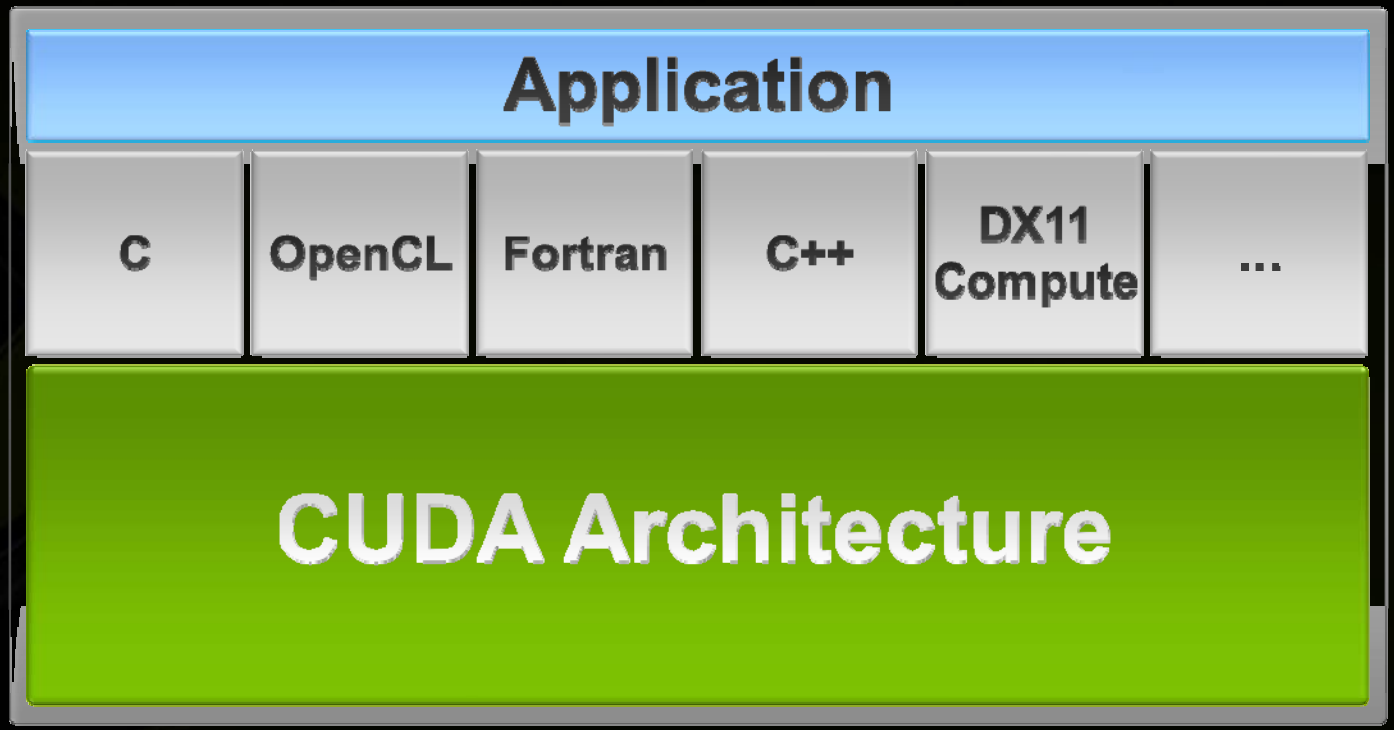
- Processor core has
  - Floating point / Integer unit
  - Move, compare, logic, branch unit
- IEEE 754 floating point
  - Single and Double
- 102 GB/s high-speed interface to memory



30 TPAs = 240 Processors

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)

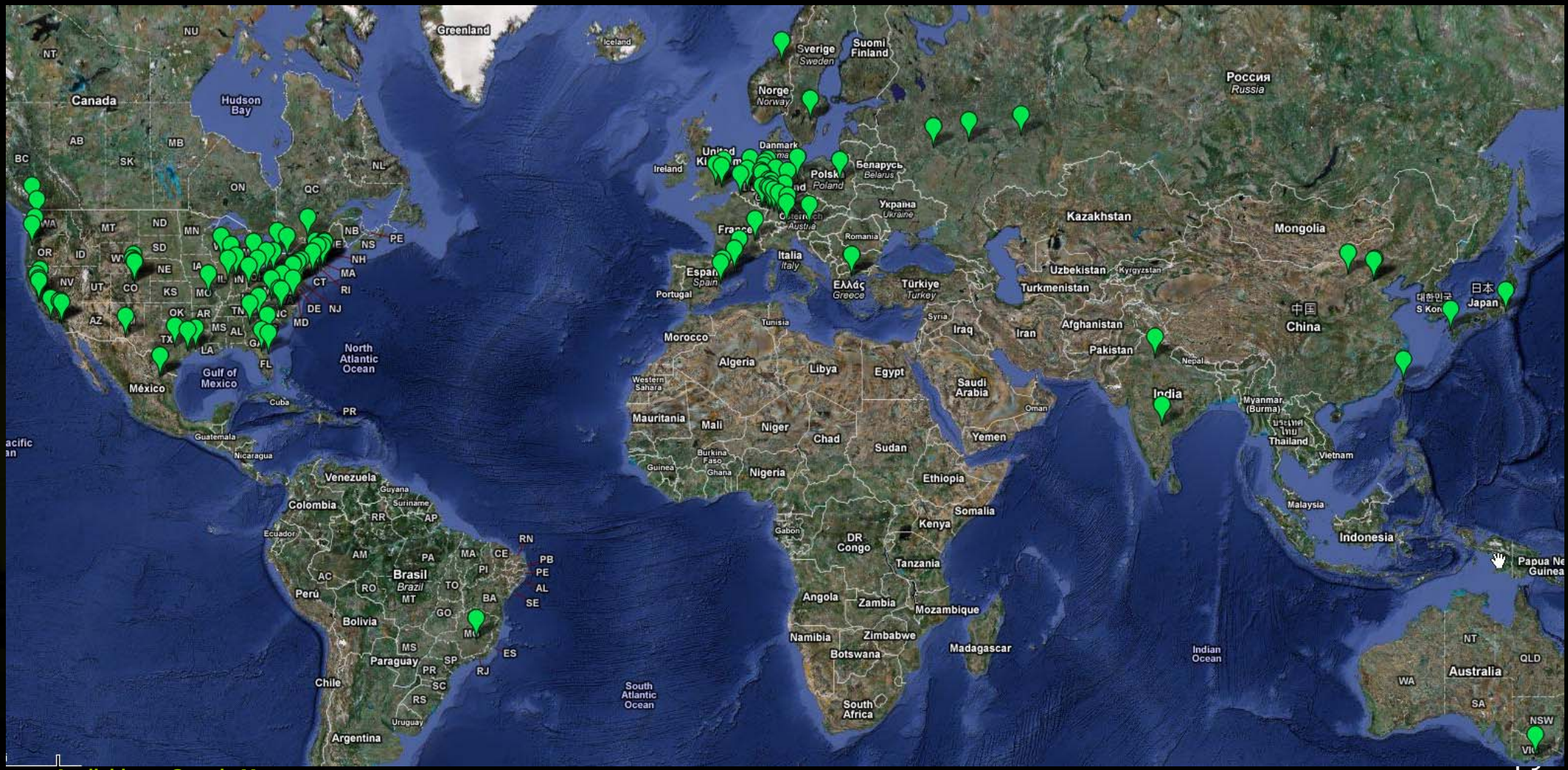


CUDA Facts

- 750+ Research Papers
  http://scholar.google.com/scholar?hl=en&q=cuda+gpu
- 100+ universities teaching CUDA
  http://www.nvidia.com/object/cuda_university_courses.html
- 100 Million CUDA-Enabled GPUs
- 25,000 Active Developers

www.NVIDIA.com/CUDA

100+ Universities Teaching CUDA



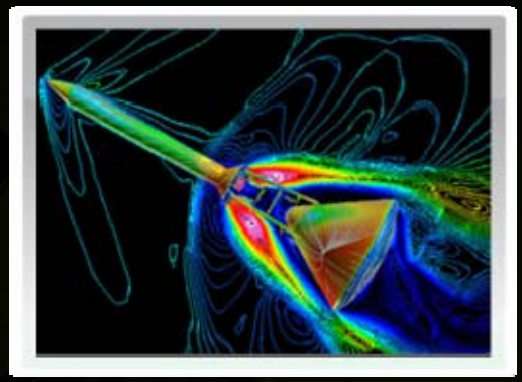
Available on Google Maps at http://maps.google.com/maps/ms?ie=UTF8&hl=en&t=h&msa=0&msid=118373856608078972372.00046298fc59576f7741a&ll=31.052934,10.546875&spn=161.641612,284.765625&z=2

Parallel Computing on All GPUs

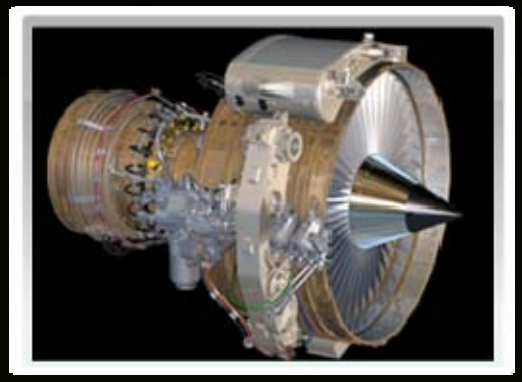
100+ Million CUDA GPUs Deployed

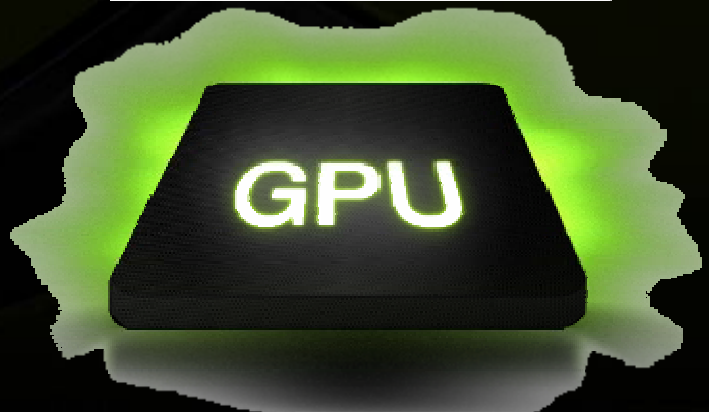
GeForce®
Entertainment

Tesla™
High-Performance Computing

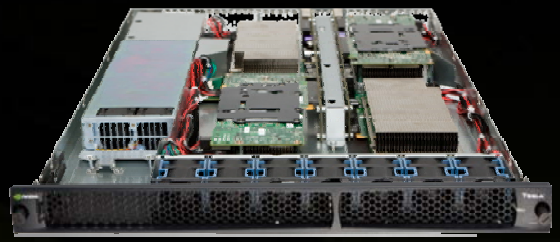


Quadro®
Design & Creation





Tesla GPU Computing Products



Tesla S1070 1U System

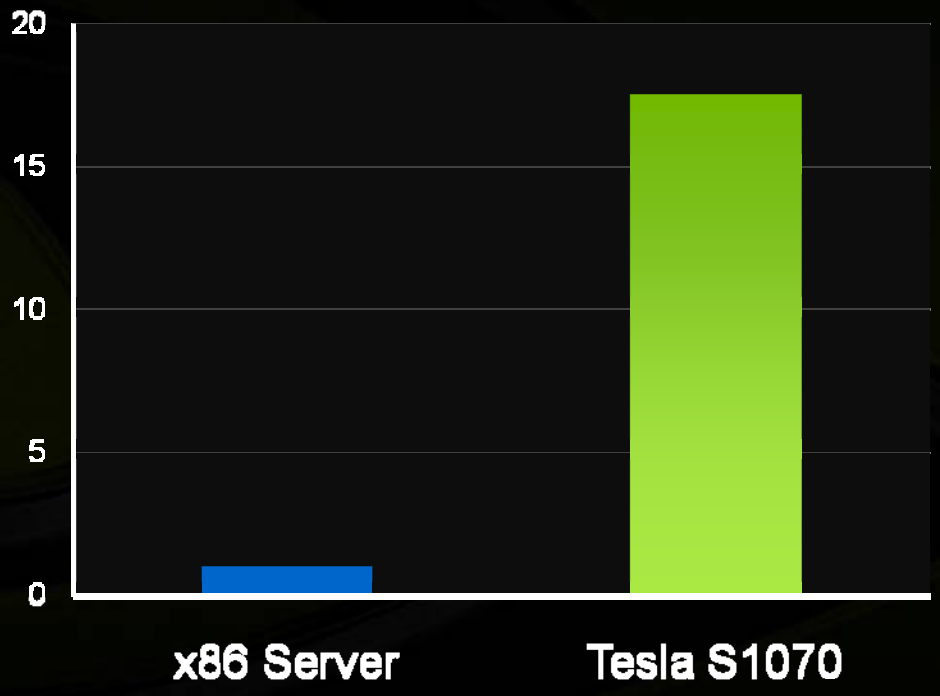
Tesla C1060
Computing Board

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
- 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
	1819 CPU servers
	310 Teraflops (single)
	155 Teraflops (double)
	Total area 16K sq feet
	Total 1273 KW

	8 CPU Cores + 4 GPUs = 968 cores	CPU 1U Server Tesla 1U System
	4.14 Teraflops (single) 0.346 Teraflop (double)	
	$ 11,000	
	1500 W	
	455 CPU servers 455 Tesla systems	
	1961 Teraflops (single)	***6x more perf***
	196 Teraflops (double)	
	Total area 9K sq feet	***40% smaller***
	Total 682 KW	***½ the power***

5000+ Customers / ISVs

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

CUDA

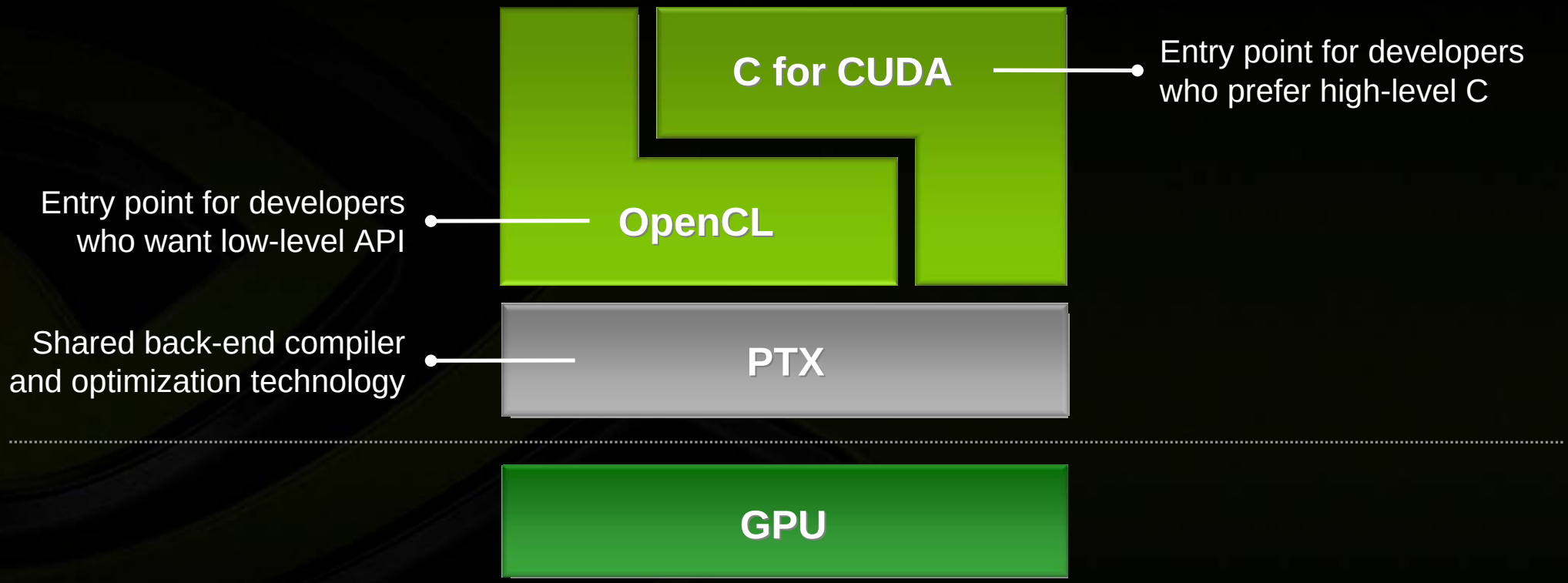
More Information

- Tesla main page
  - http://www.nvidia.com/tesla
  - Product Information
  - Industry Solutions
- CUDA Zone
  - http://www.nvidia.com/cuda
  - Applications, Papers, Videos
- Hear from Developers
  - http://www.youtube.com/nvidiatesla

CUDA

CUDA Parallel Programming Architecture and Model
Programming the GPU in High-Level Languages

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
  - Generates PTX
- OpenCL
  - Hardware API - similar to OpenGL and CUDA driver API
  - Programmer has complete access to hardware device
  - Memory managed by programmer
  - Generates PTX

Simple “C” Description For Parallelism

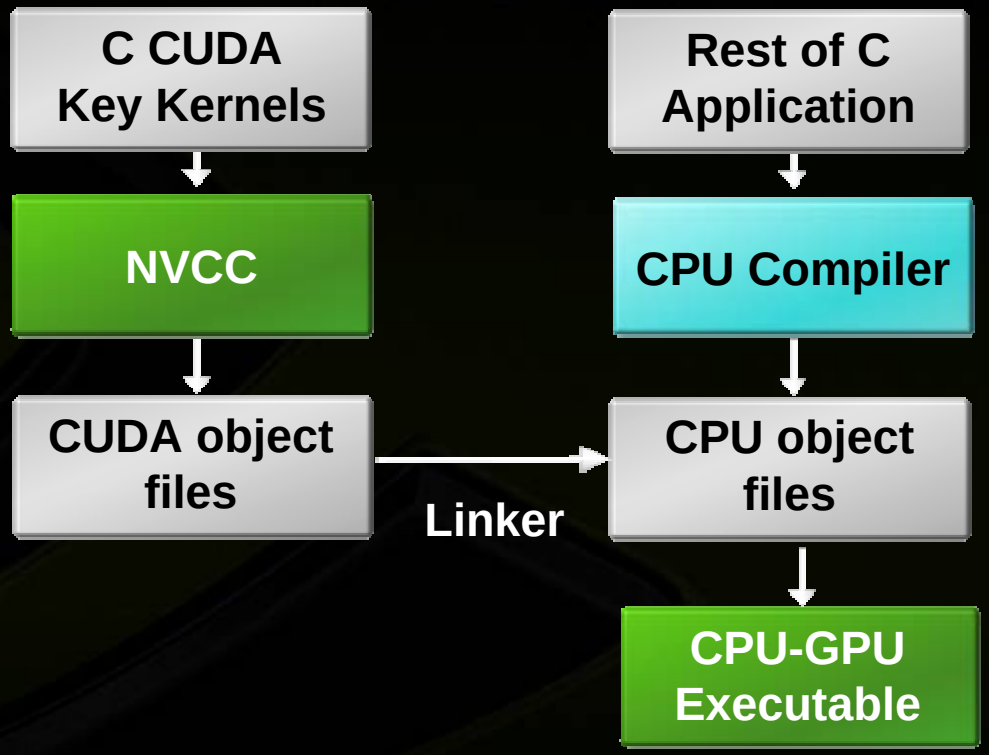
```
void saxpy_serial(int n, float a, float *x, float *y)
{
    for (int i = 0; i < n; ++i)
        y[i] = a*x[i] + y[i];
}
// Invoke serial SAXPY kernel
saxpy_serial(n, 2.0, x, y);
```

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);
```

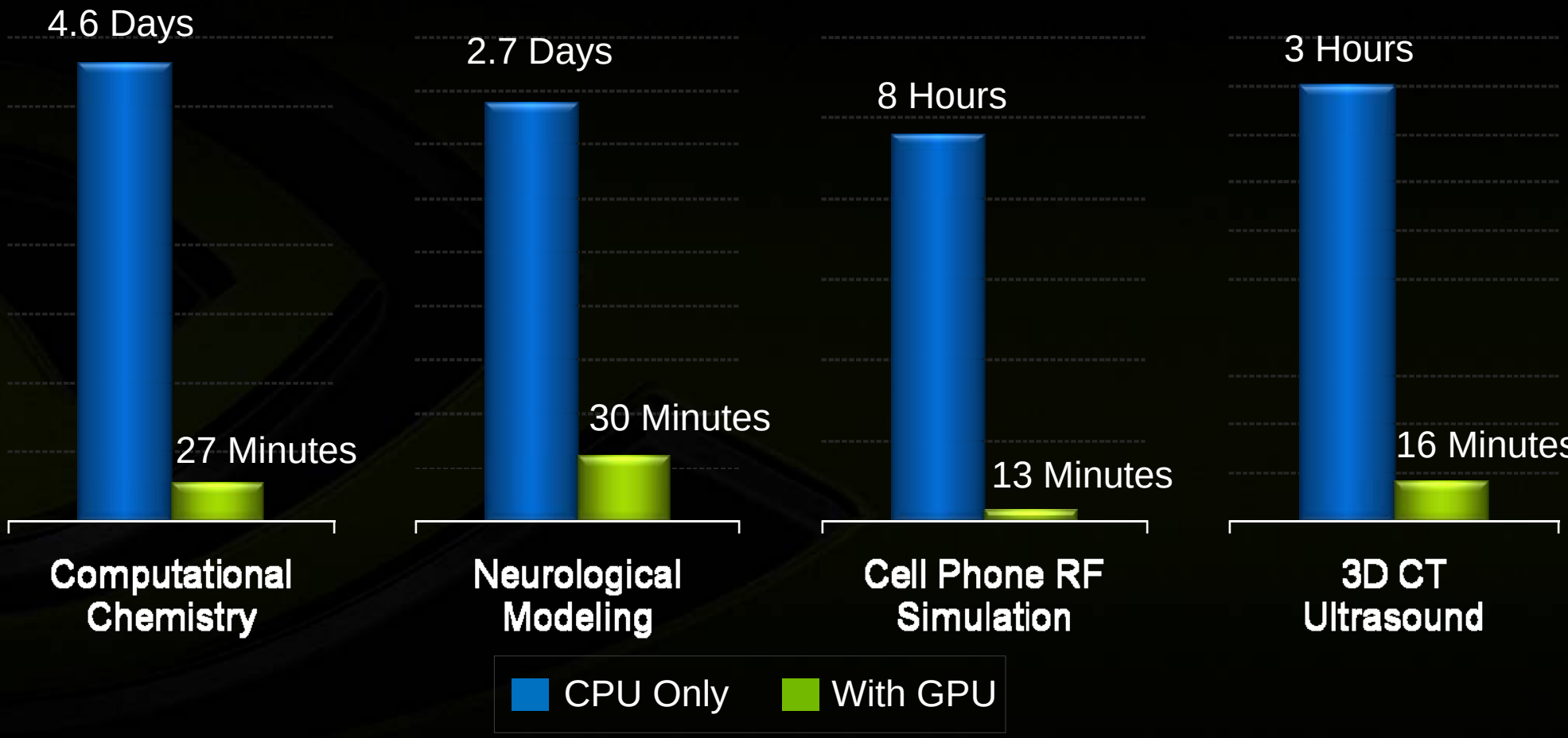
Parallel C Code

Compiling C for CUDA Applications



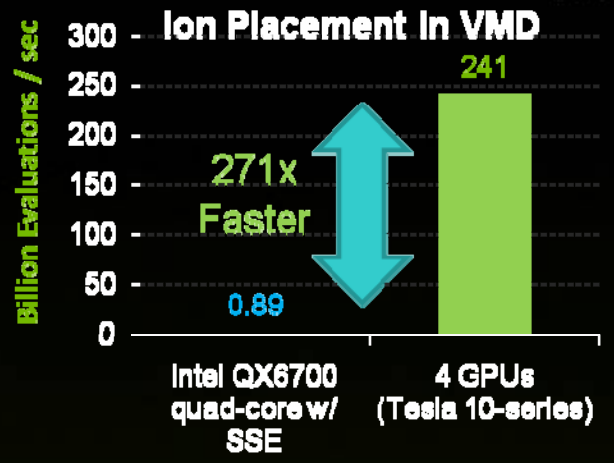
Application Domains

Accelerating Time to Discovery

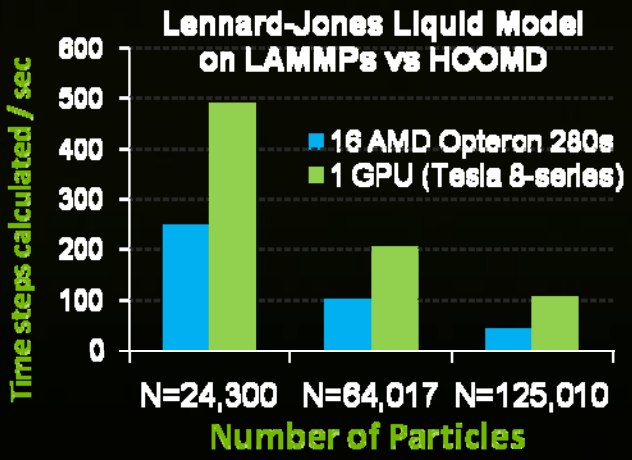


Molecular Dynamics

- Available MD software
  - NAMD / VMD (alpha release)
  - HOOMD
  - ACE-MD
  - MD-GPU
- Ongoing work
  - LAMMPS
  - CHARMM
  - GROMACS
  - AMBER



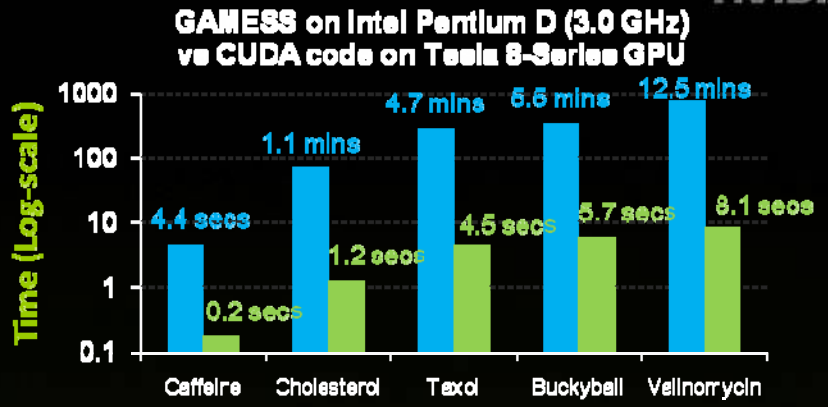
Source: Stone, Phillips, Hardy, Schulten



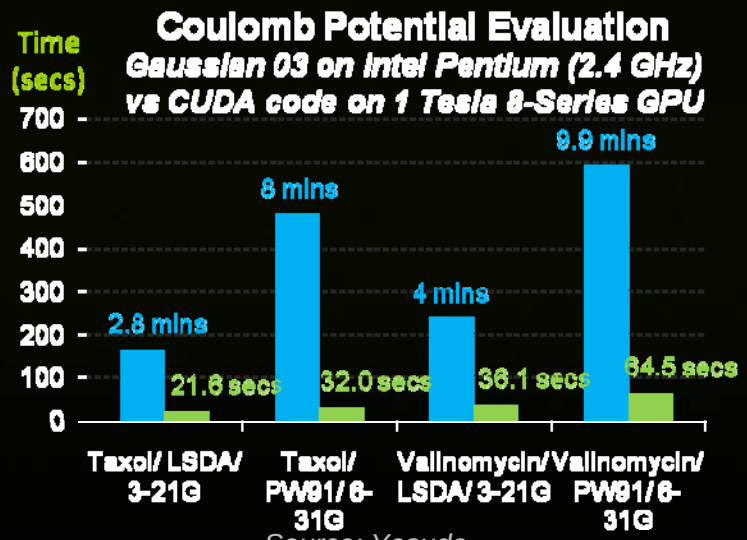
Source: Anderson, Lorenz, Travesset

Quantum Chemistry

- **Available MD software**
  - NAMD / VMD (alpha release)
  - HOOMD
  - ACE-MD
  - MD-GPU
- **Ongoing work**
  - LAMMPS
  - CHARMM
  - Q-Chem
  - Gaussian



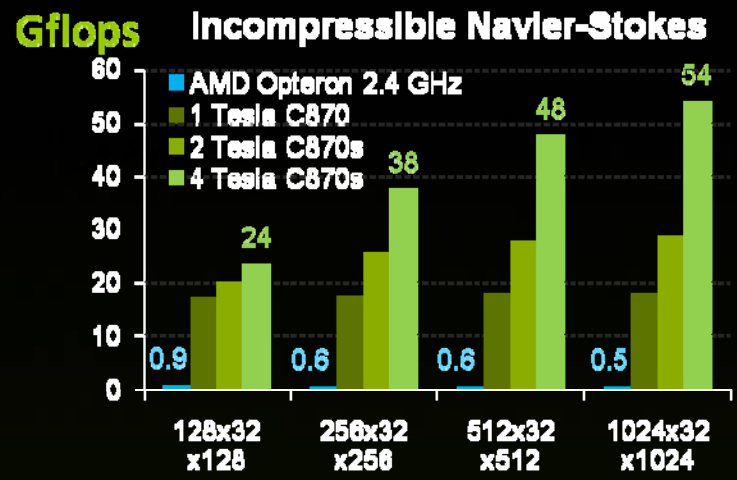
Source: Ufimtsev, Martinez



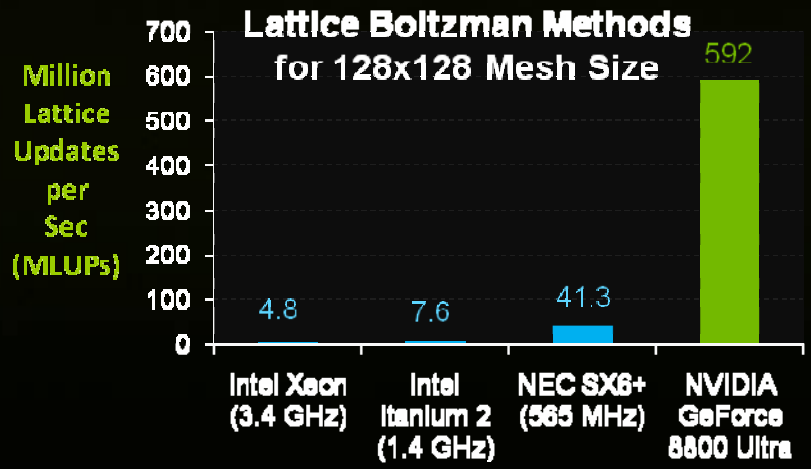
Source: Yasuda

Computational Fluid Dynamics (CFD)

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



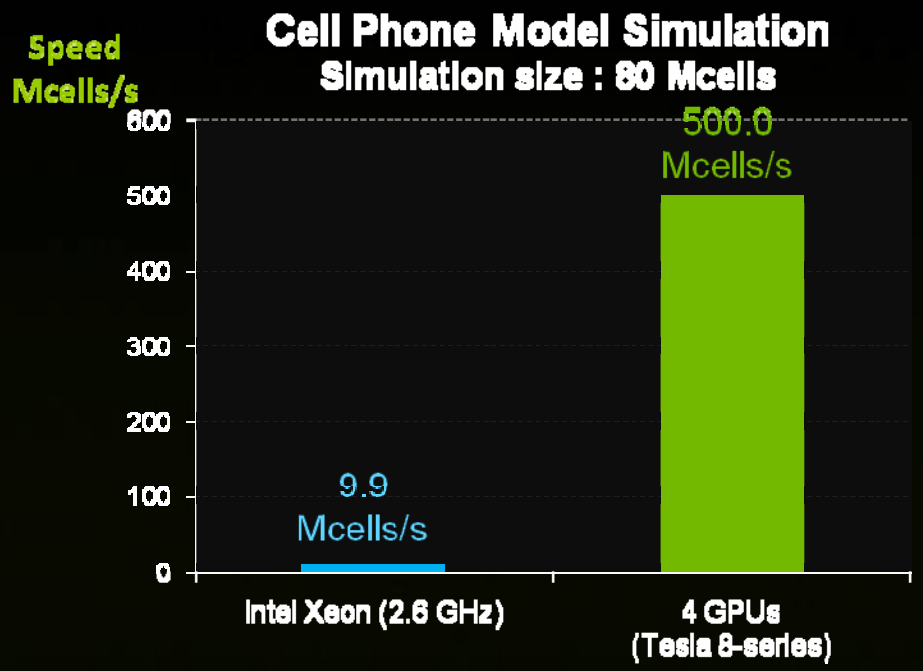
Source: Thibault, Senocak



Source: Tolke, Krafczyk

Electromagnetics / Electrodynamics

- **FDTD Solvers**
  - Acceleware
  - EM Photonics
  - CUDA Tutorial
- **Ongoing work**
  - Maxwell equation solver
  - Ring Oscillator (FDTD)
  - Particle beam dynamics simulator

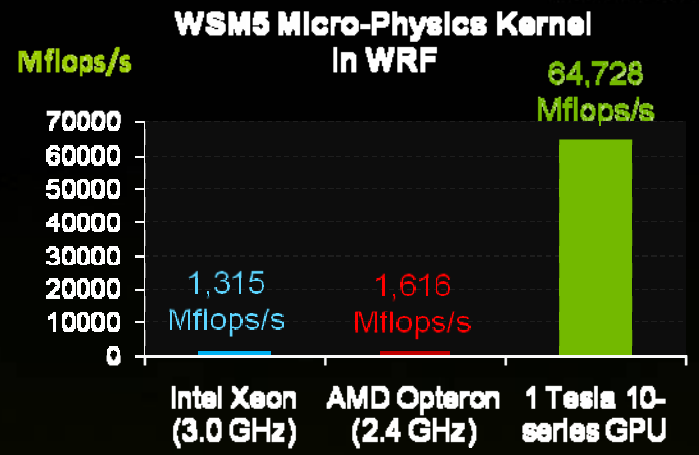


FDTD Acceleration using GPUs

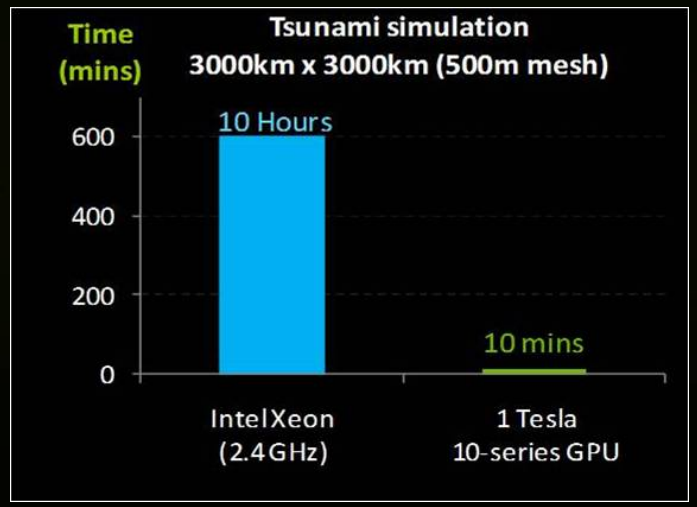
Source: Acceleware

Weather, Atmospheric, & Ocean Modeling

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

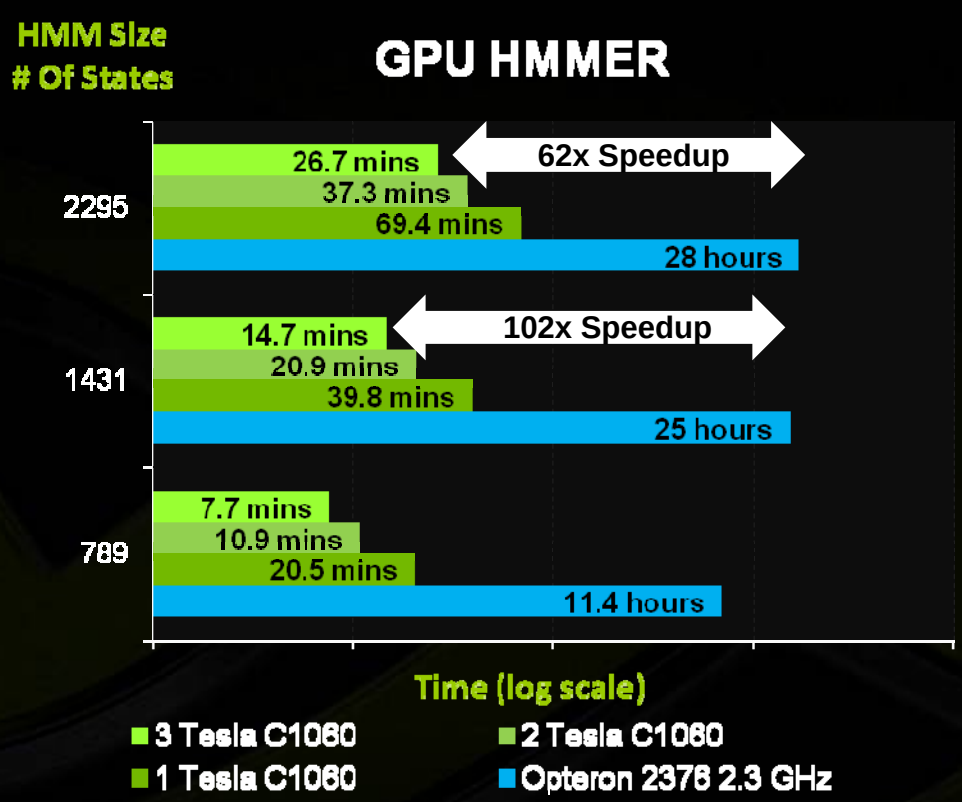


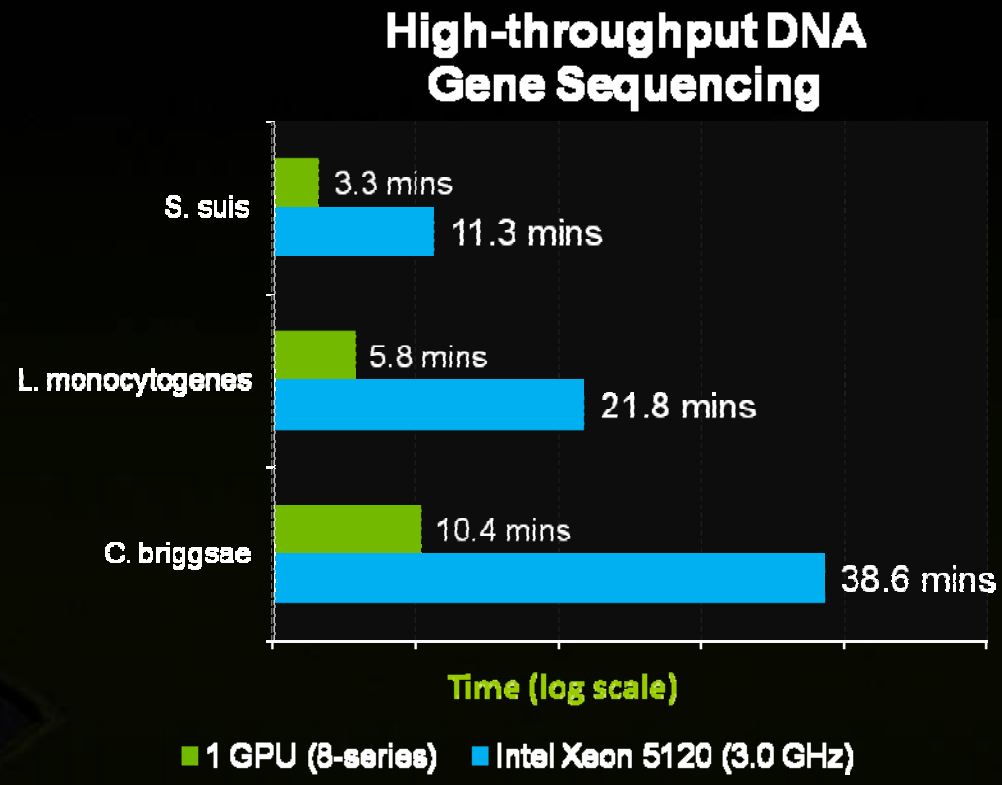
Source: Michalakes, Vachharajani



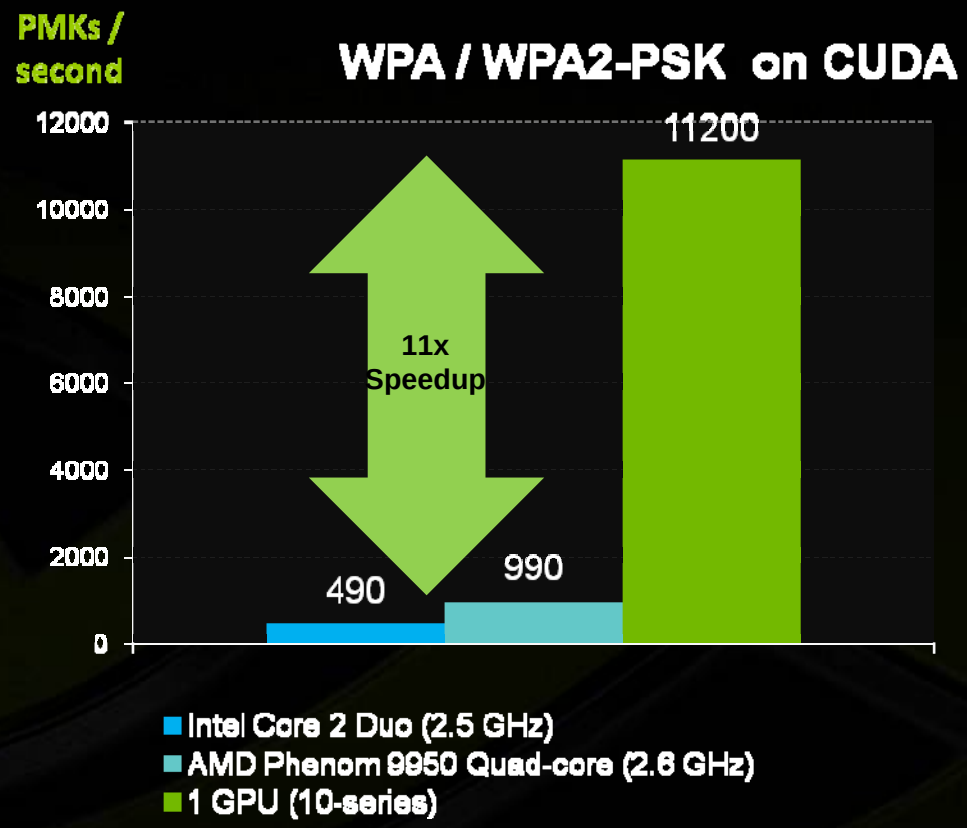
Source: Takayuki Aoki, Tokyo Tech

Bio-Informatics

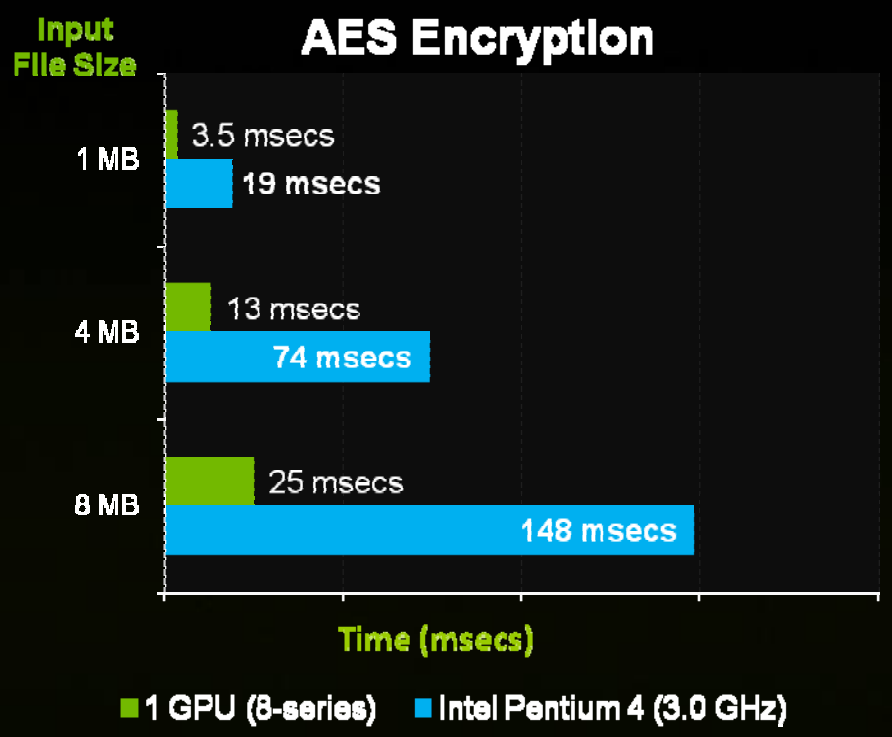




Encryption

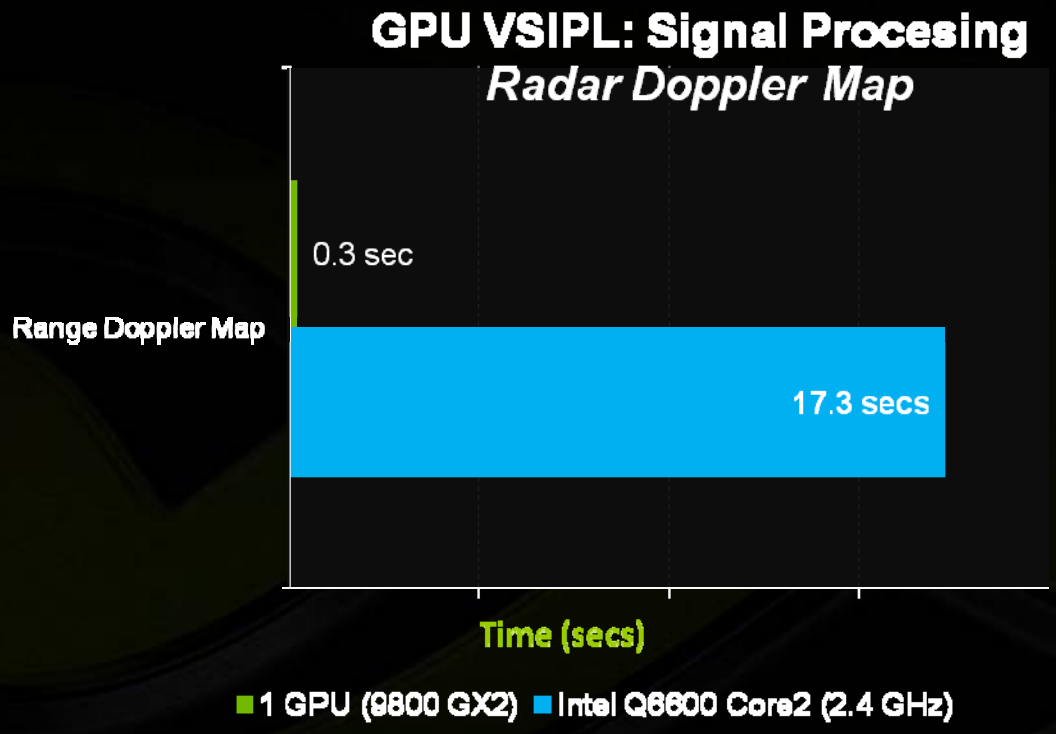


Source: http://code.google.com/p/pyrit/

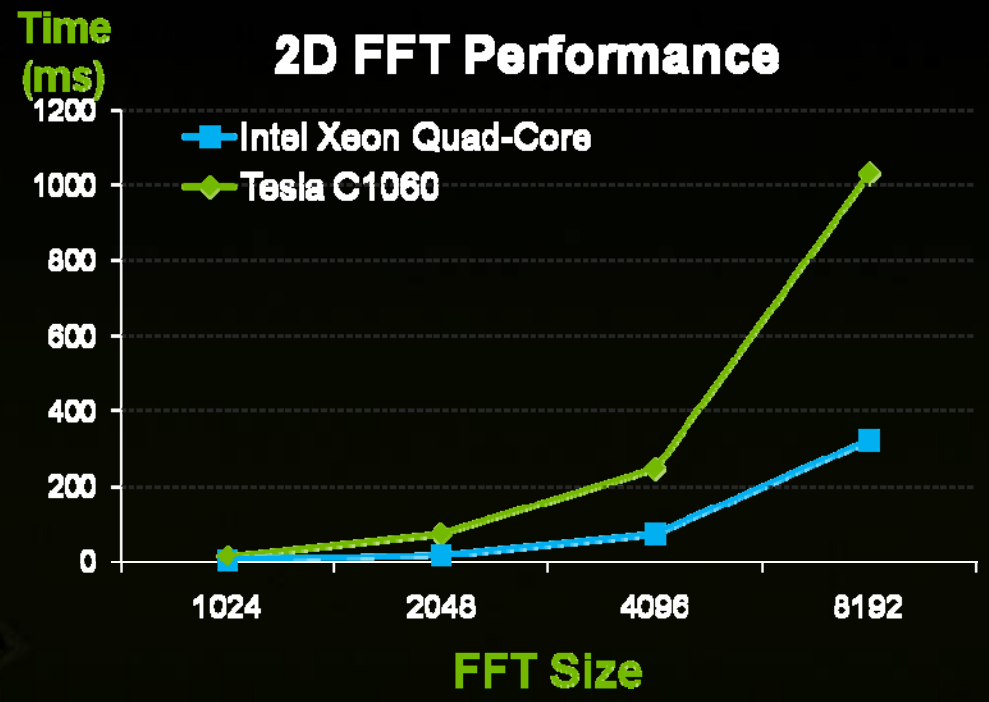


Source: S. Manavski

Signal Processing



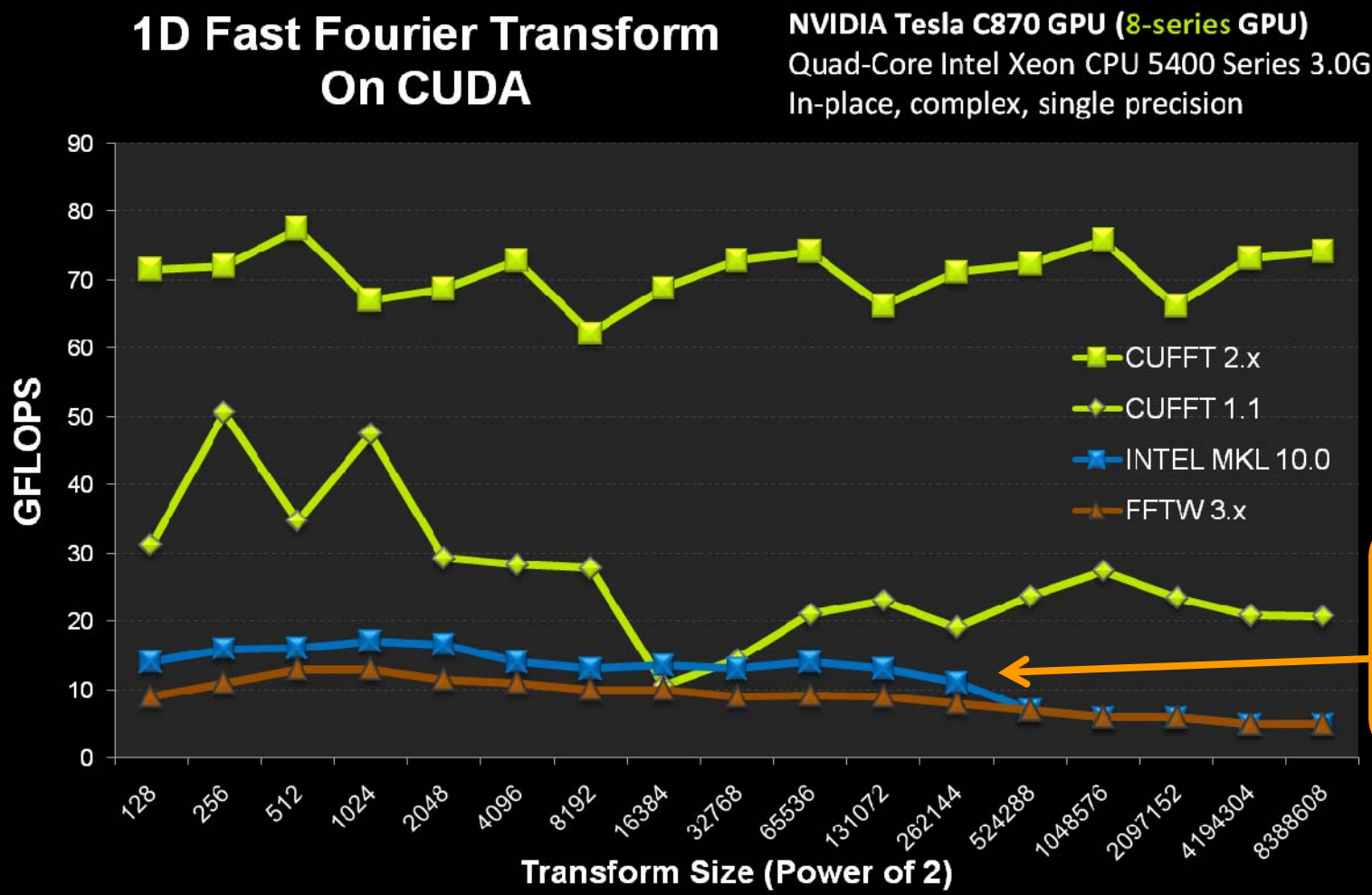
Source: http://gpu-vsipl.gtri.gatech.edu/



Source: CUDA FFT library (cuFFT)

Libraries

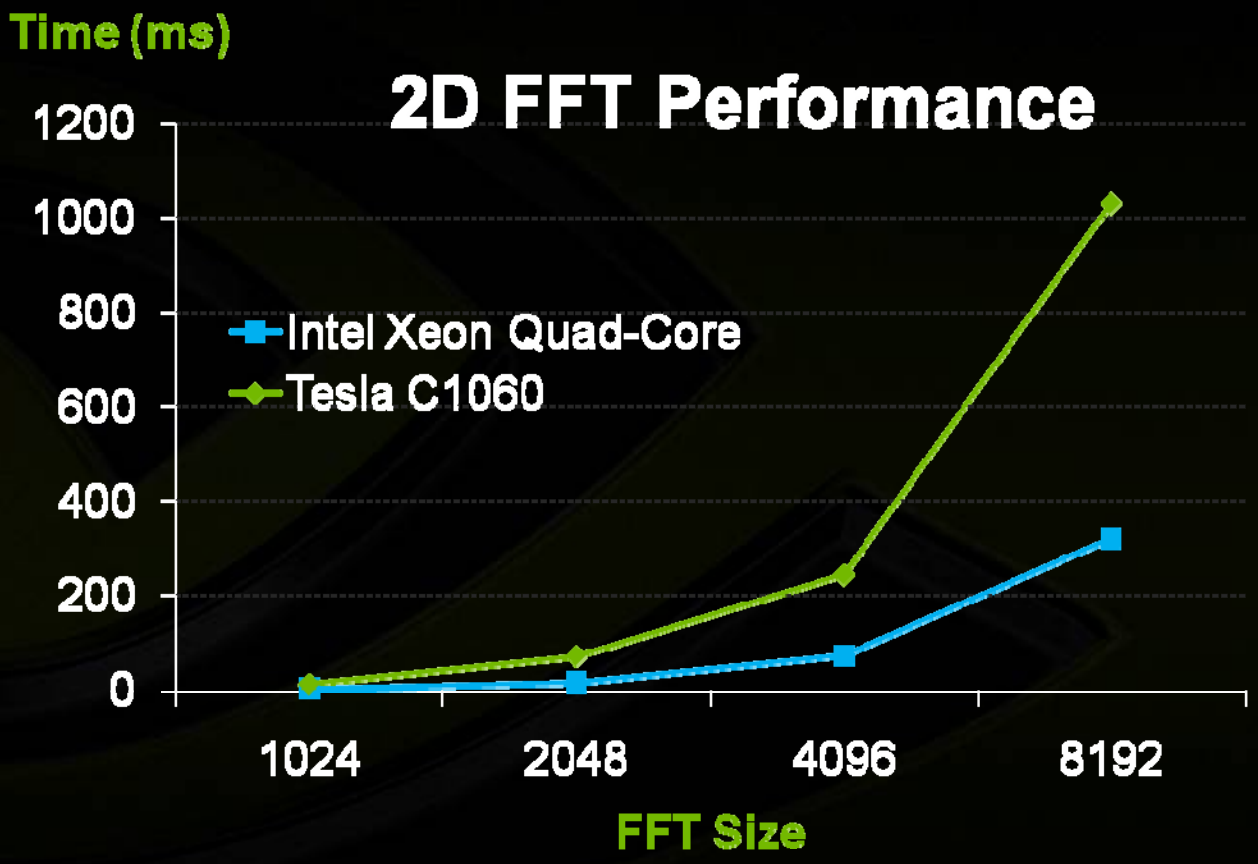
FFT Performance: CPU vs GPU (8-Series)



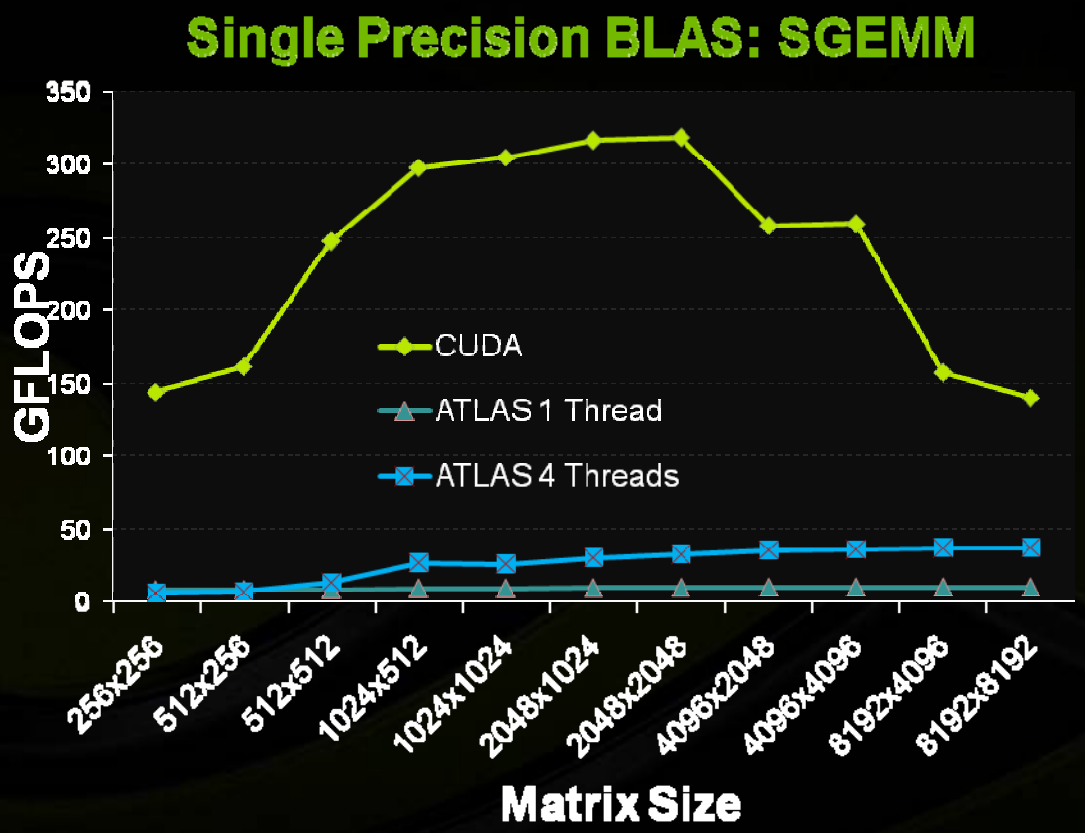
- *Intel FFT numbers calculated by repeating same FFT plan*
- *Real FFT performance is ~10 GFlops*

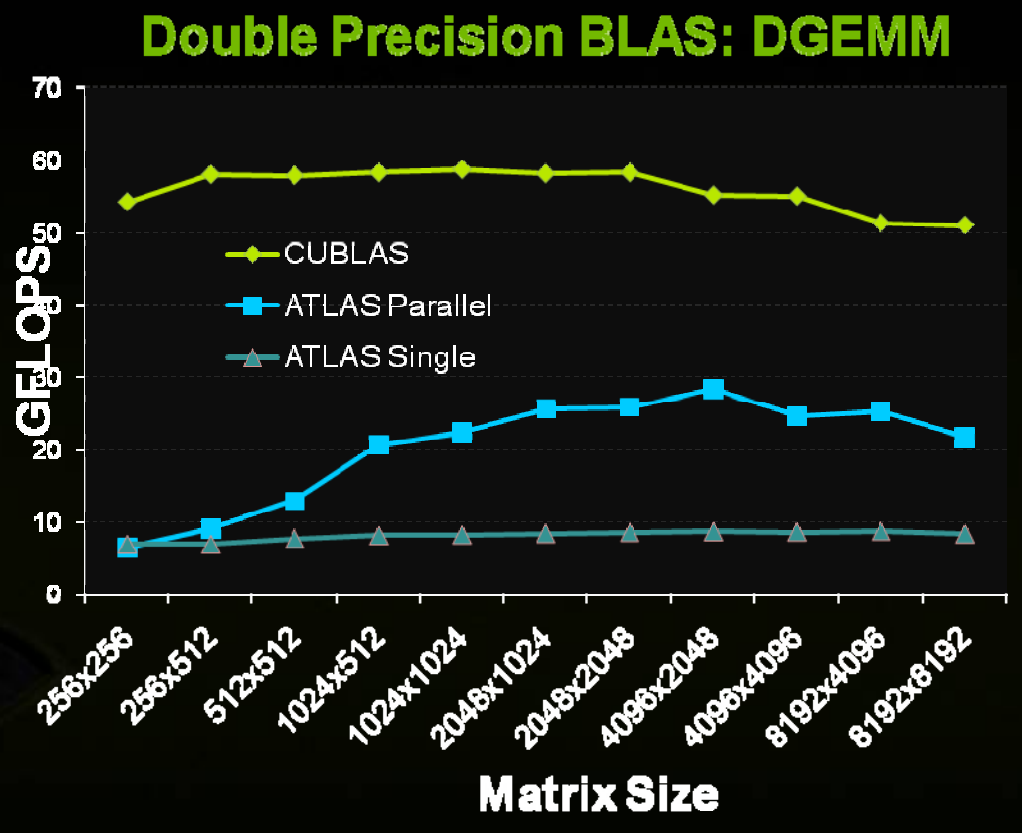
Source for Intel data : http://www.intel.com/cd/software/products/asmo-na/eng/266852.htm

2D FFT : GPU Performance



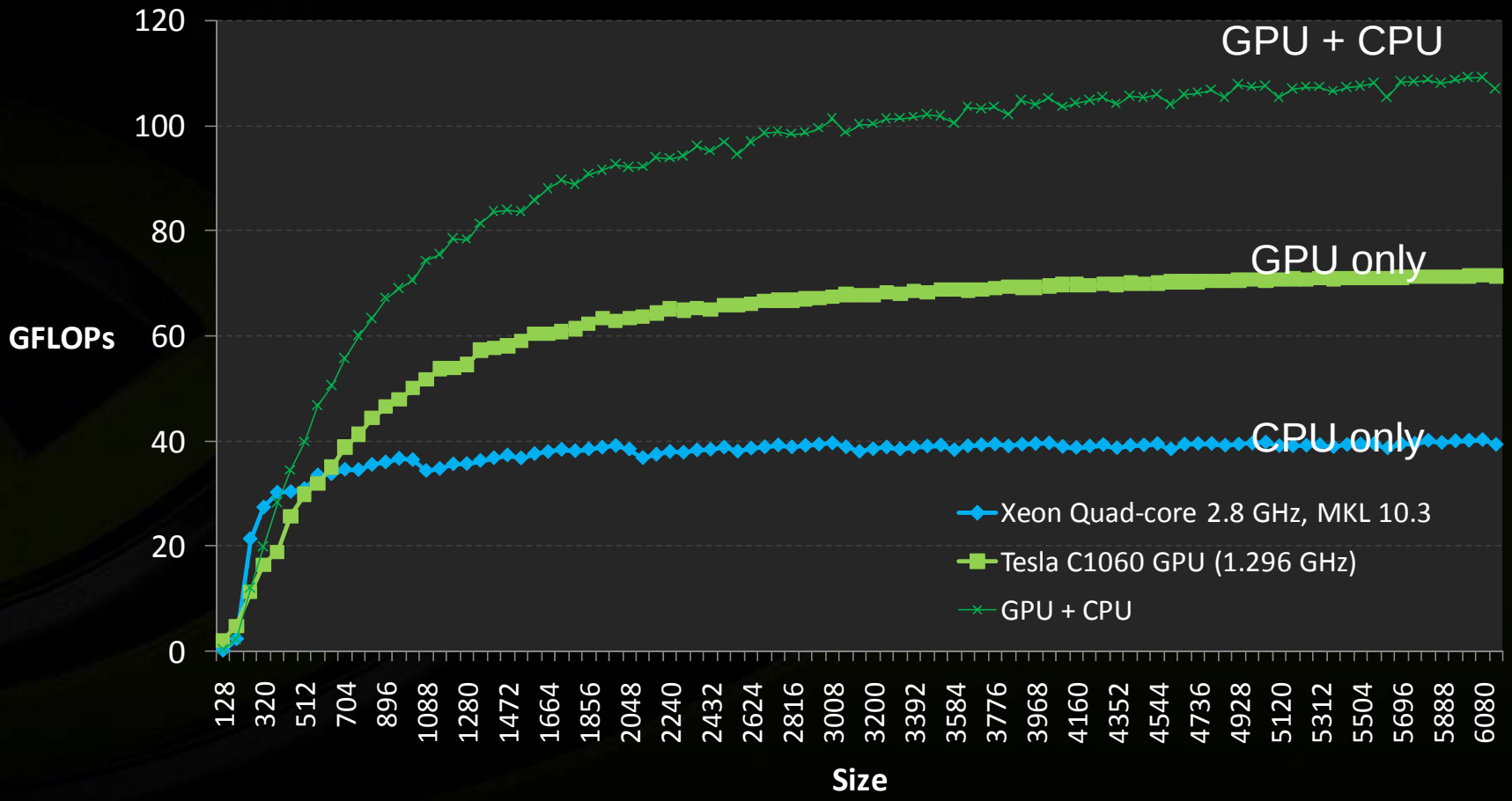
BLAS: CPU vs GPU (10-series)



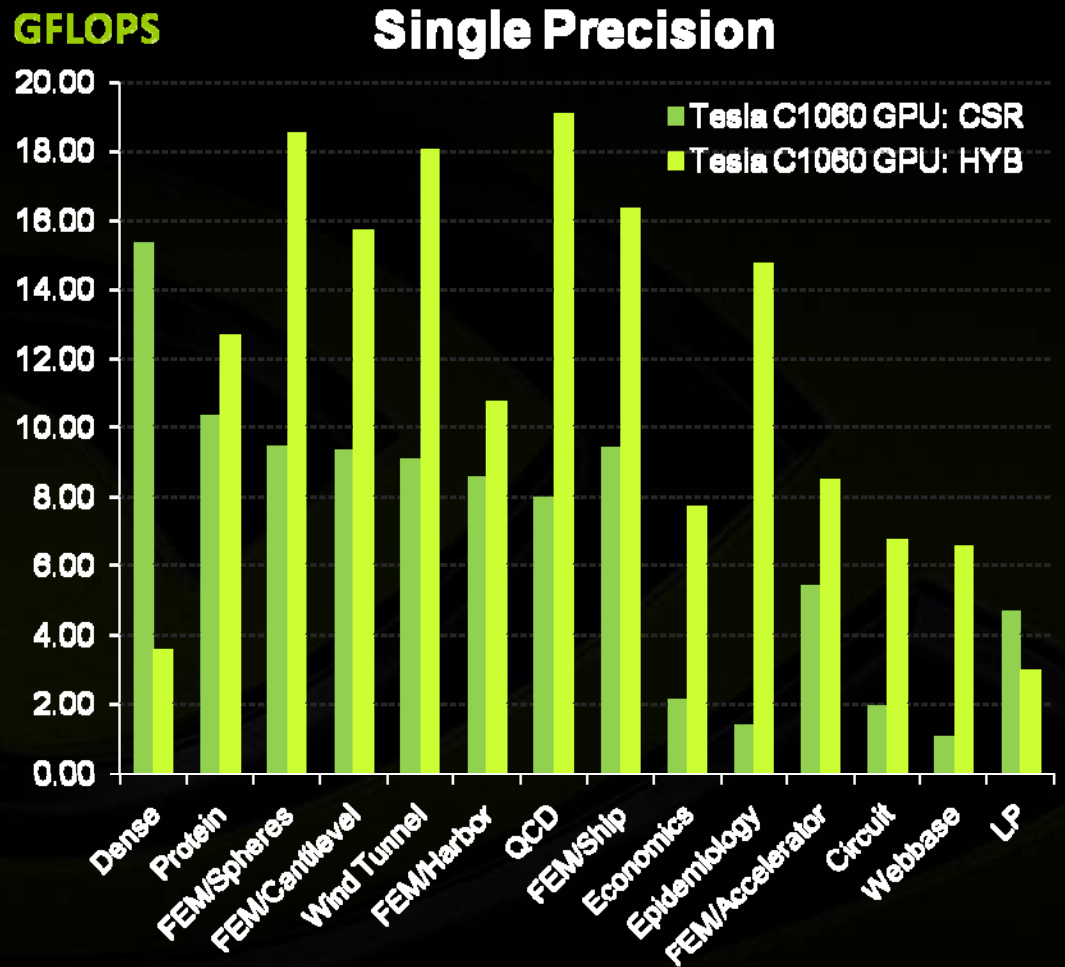


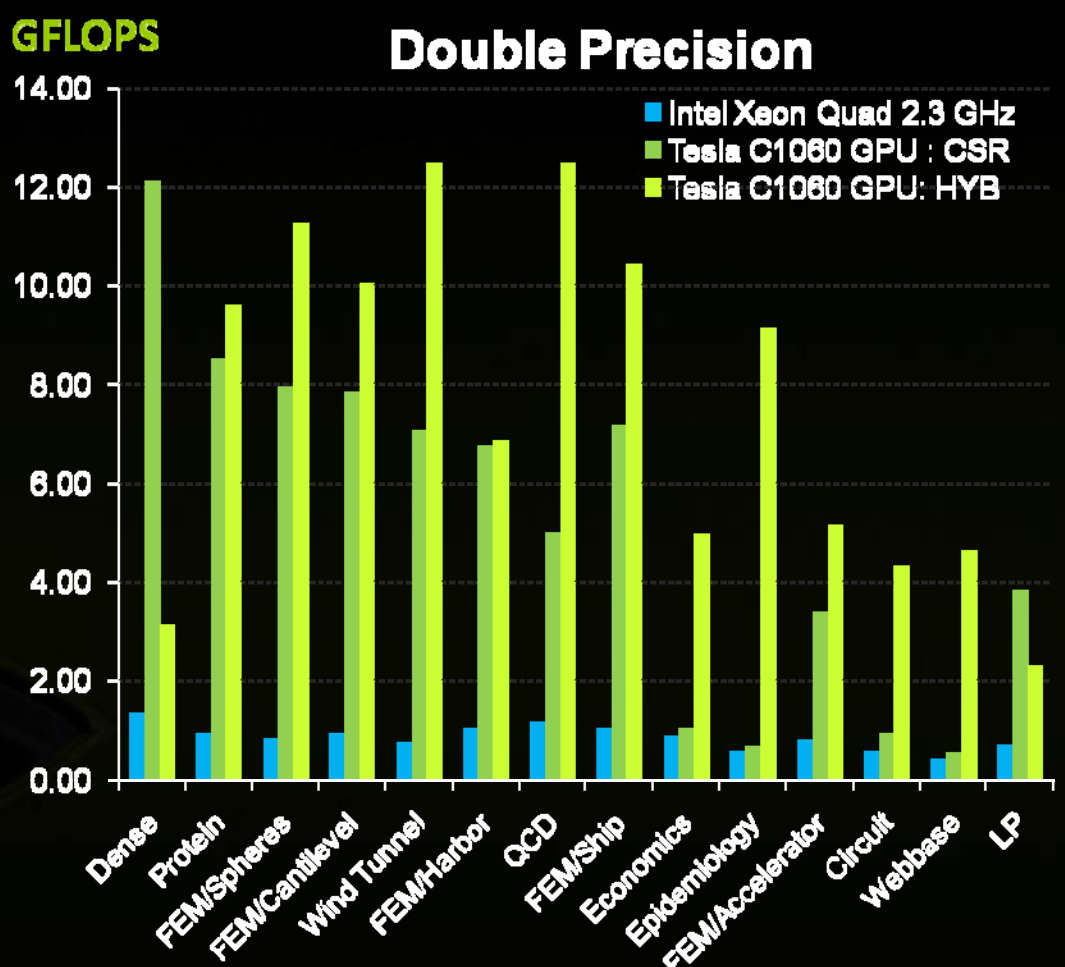
CUBLAS: CUDA 2.0, Tesla C1060 (**10-series** GPU)
ATLAS 3.81 on Dual 2.8GHz Opteron Dual-Core

GPU + CPU DGEMM Performance



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

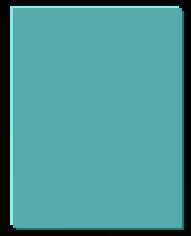


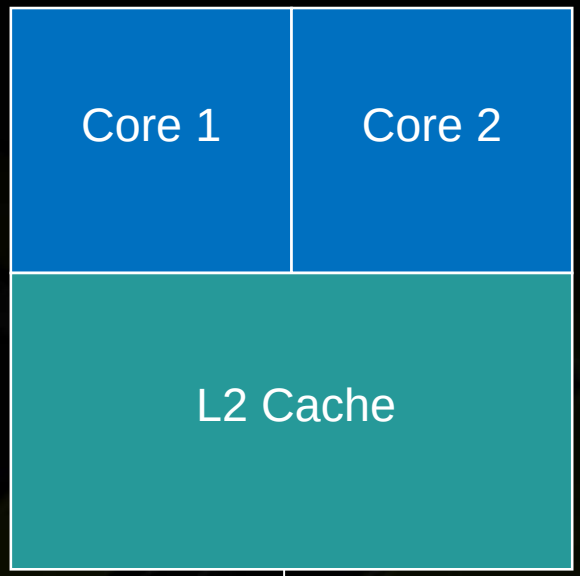


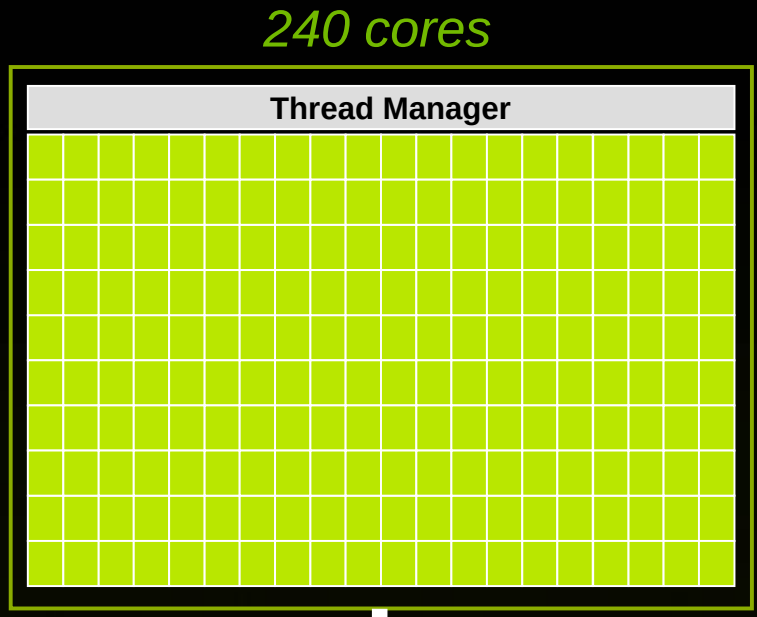
CPU Results from "Optimization of Sparse Matrix-Vector Multiplication on Emerging Multicore Platforms", Williams et al, Supercomputing 2007

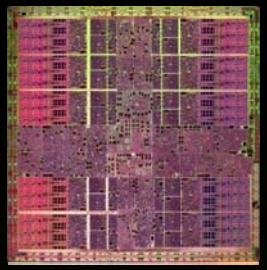
GPUs: Better Architecture for Computing

CPUs: Memory Bandwidth Bottleneck



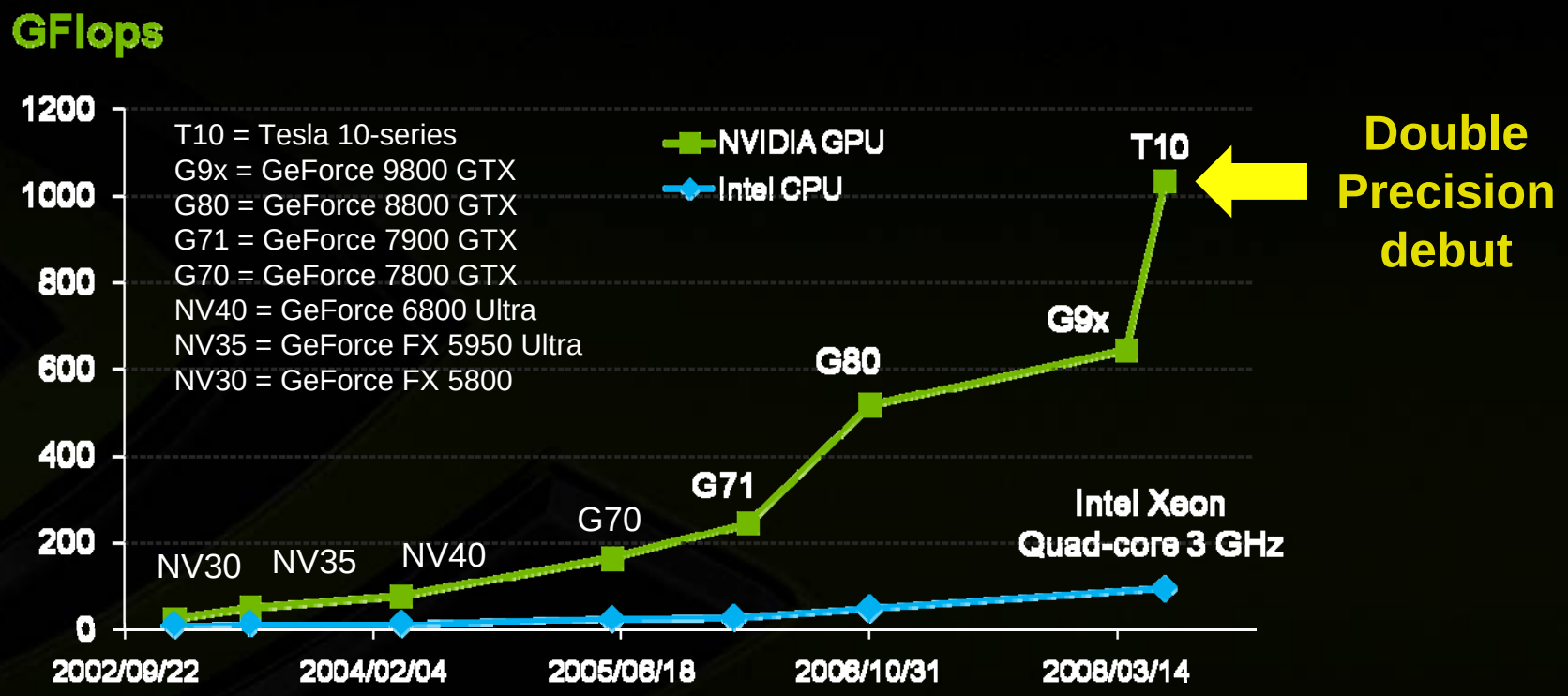






8x faster interface

NVIDIA's GPUs : Ever Increasing Performance



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)

Final Thoughts

- GPU and heterogeneous parallel architecture will revolutionize computing
- Parallel computing key to solve some of the most interesting and important human challenges ahead
- Learning parallel programming is an imperative for students in computing and sciences