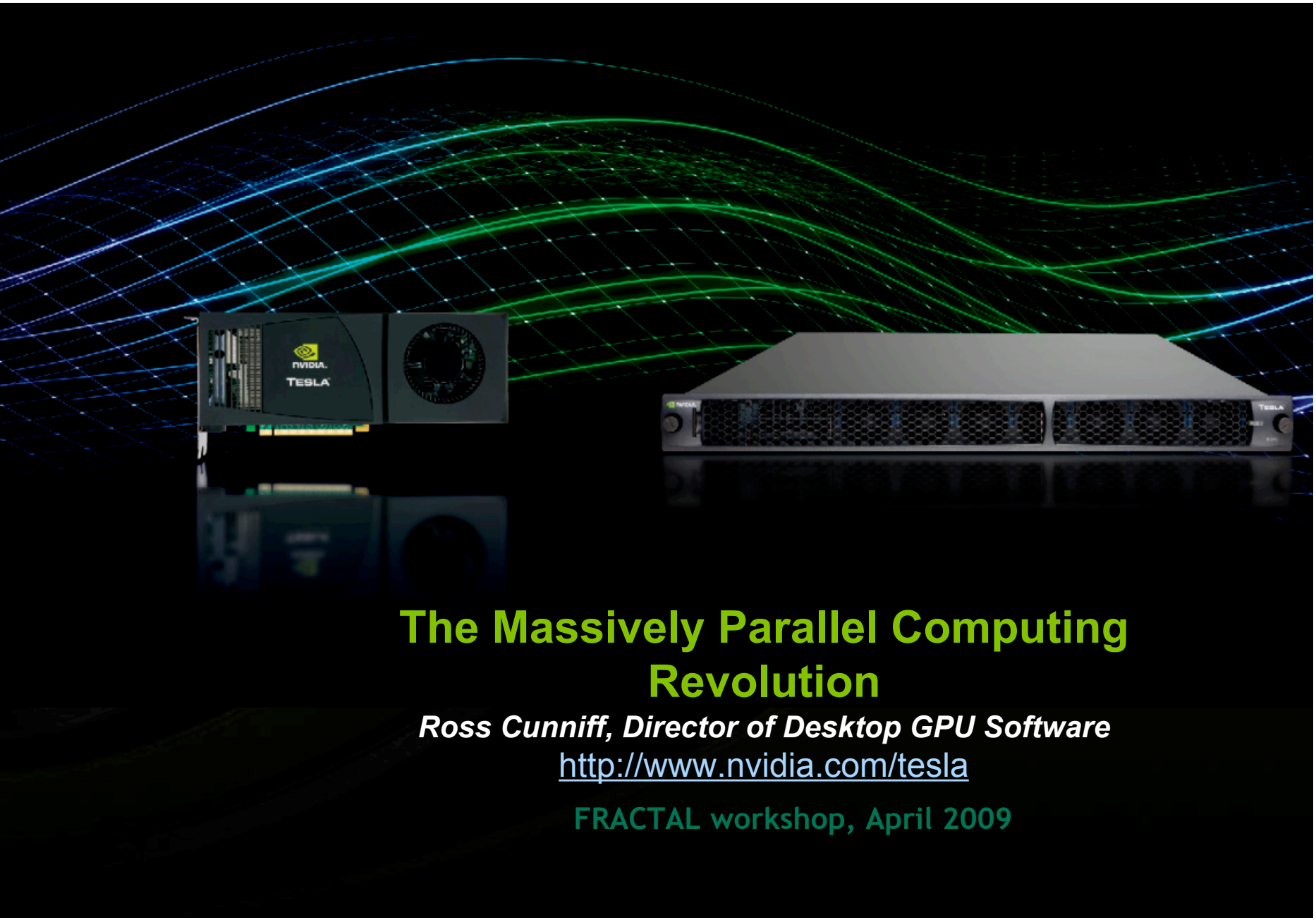


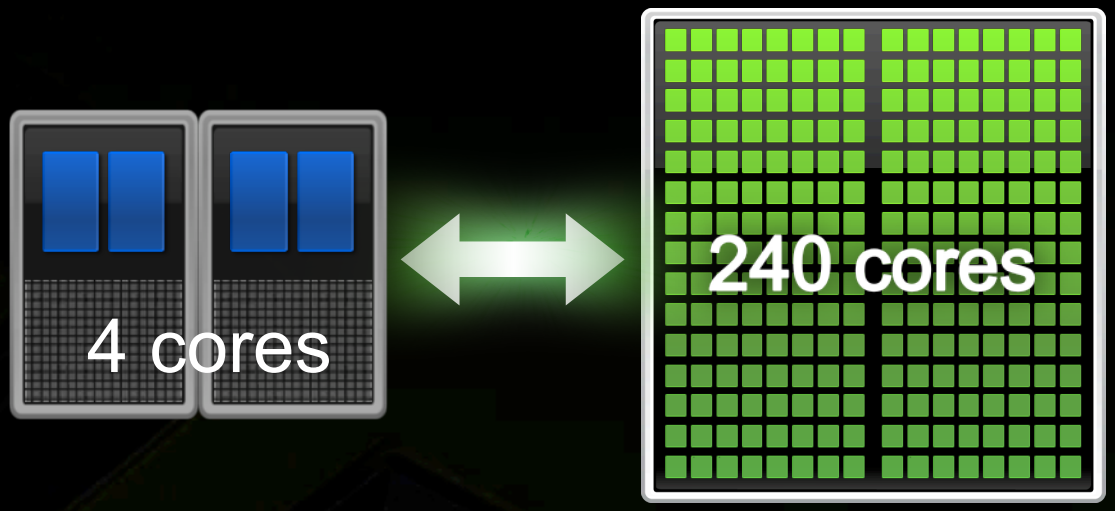
The Massively Parallel Computing Revolution

Ross Cunniff, Director of Desktop GPU Software

http://www.nvidia.com/tesla

FRACTAL workshop, April 2009

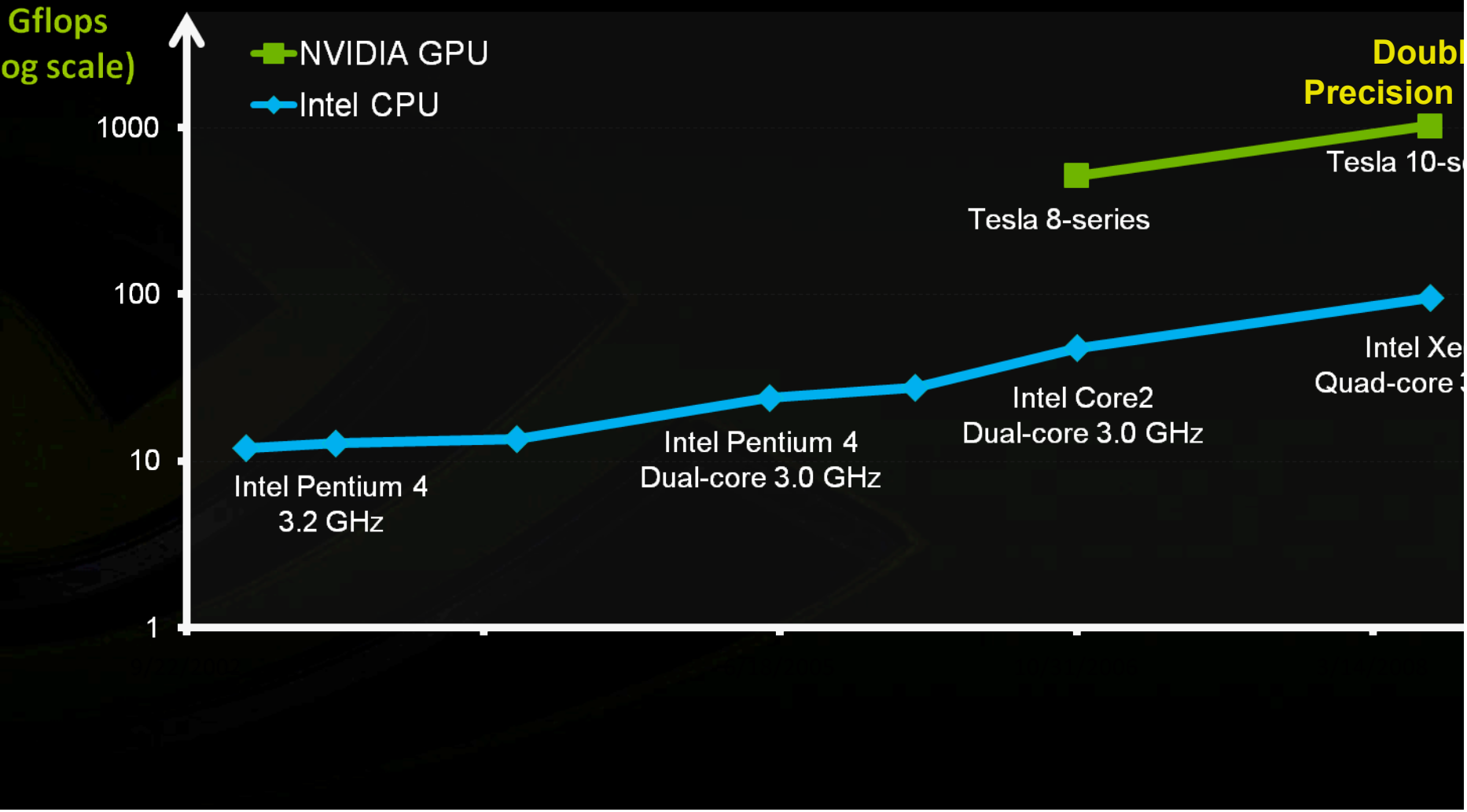
Move to Massive Parallelism



Computing with CPU + GPU

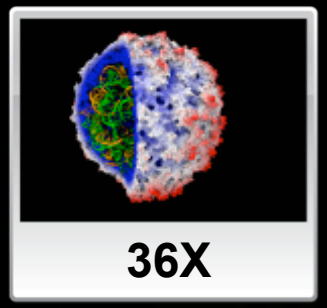
Heterogeneous Computing

Computation Discontinuity



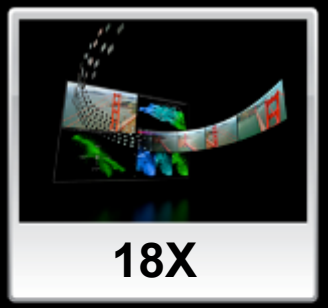
146X

Medical Imaging
U of Utah



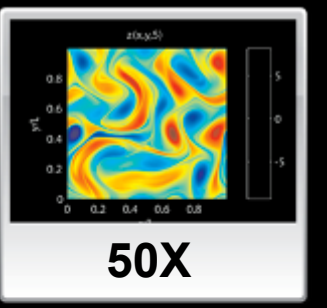
36X

Molecular Dynamics
U of Illinois, Urbana



18X

Video Transcoding
Elemental Tech



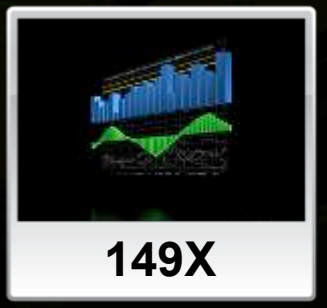
50X

Matlab Computing
AccelerEyes

1(

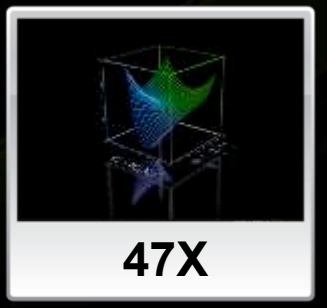
Astro
RI

50x – 150x



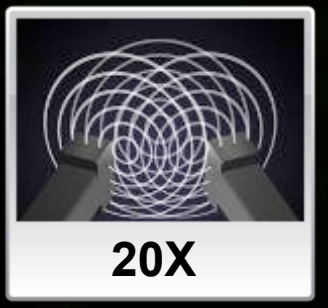
149X

Financial simulation
Oxford



47X

Linear Algebra
Universidad Jaime



20X

3D Ultrasound
Techniscan

130X

Quantum Chemistry
U of Illinois, Urbana

3

Gene S
U of M

Parallel vs Sequential Architecture Evolution

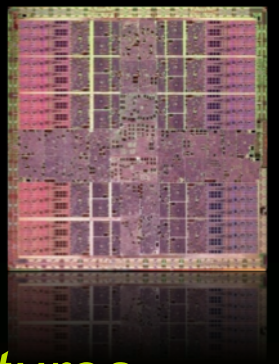
LIAC IV

Cray-1

Maspar

Thinking Machines

Blue Gene



Ma
GP

High Performance Computing Architectures

EC PDP-1

Intel 4004

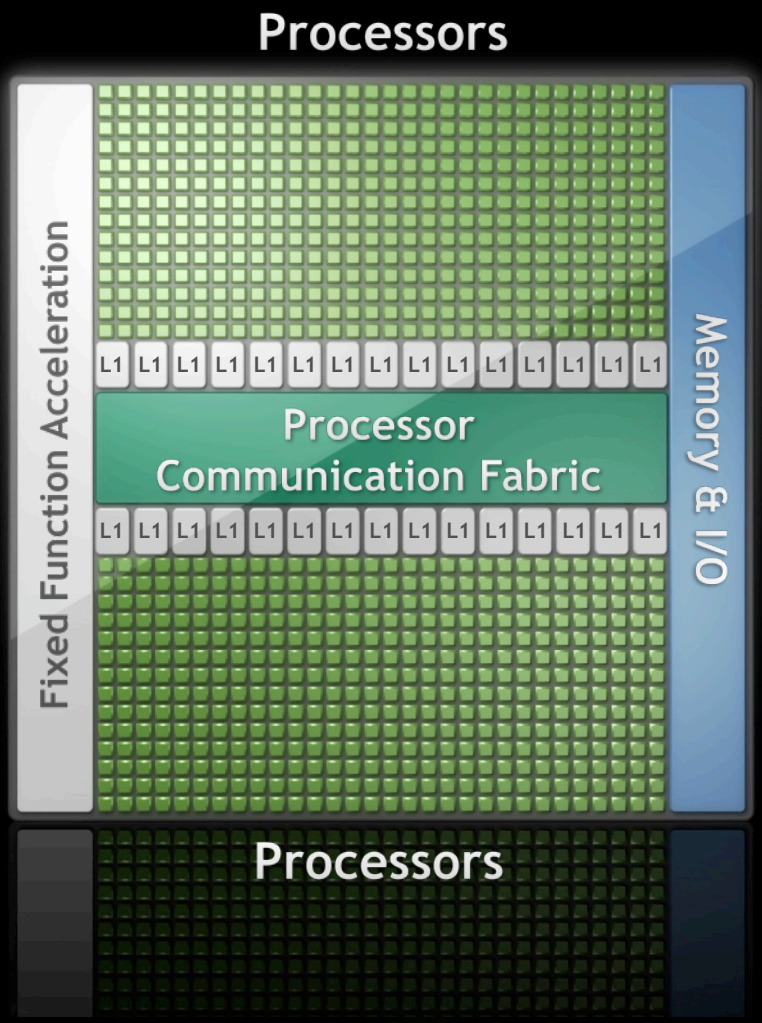
IBM System 360

VAX

IBM POWER4

Mu
x8

Database, Operating System Sequential Architectures



NVIDIA Tesla 10-Series GPU

Massively parallel, many core archite

240 Processor Cores

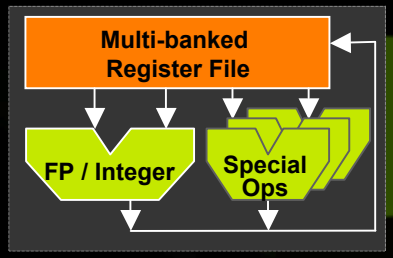
1 Teraflops - 1,000 times Cray X-MP

IEEE Compliant Double Precision Flo

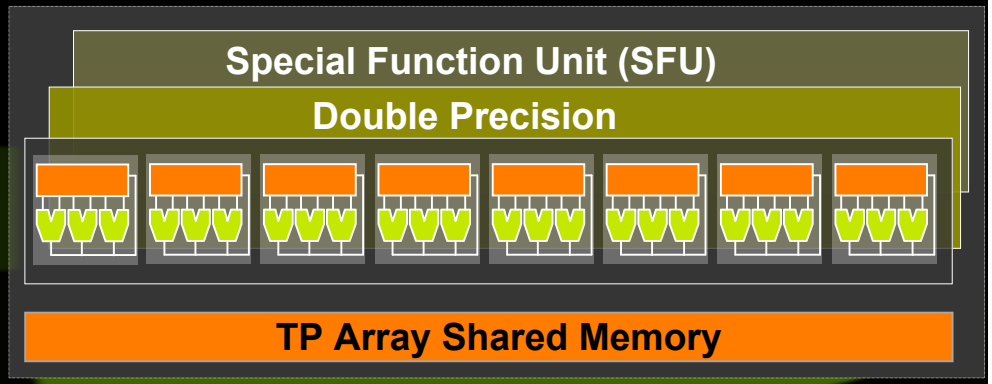
Designed for Scientific Computing

Tesla T10 GPU: 240 Processor Cores

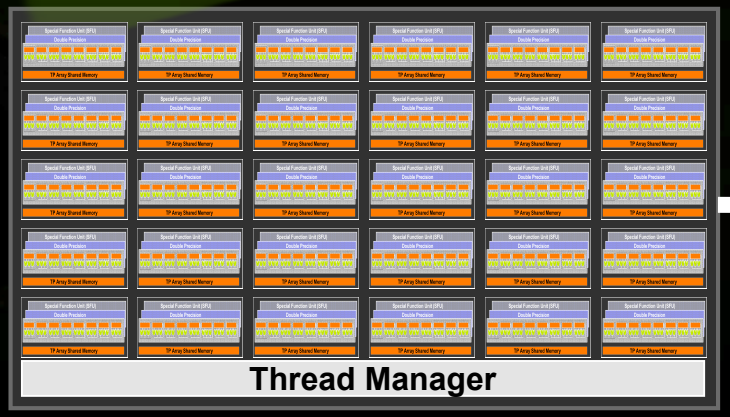
Thread Processor (TP)



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

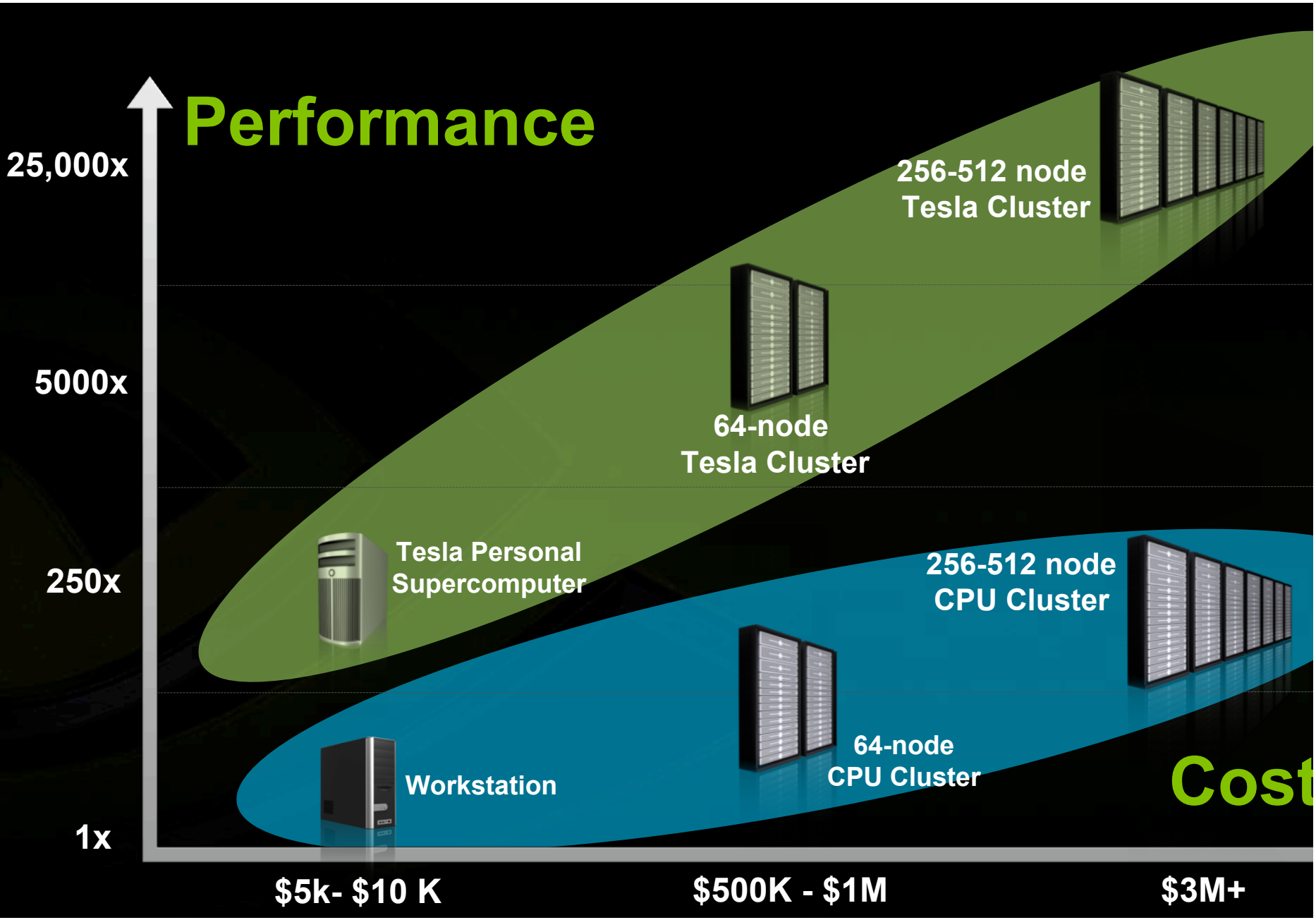


GDDR3

512 bit

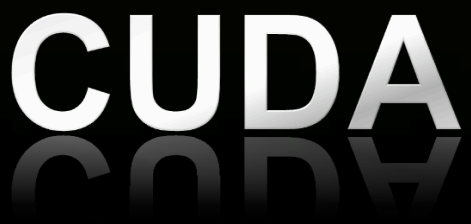
102 GB/

30 TPAs = 240 Processors


Performance
25,000x
5000x
250x
1x
256-512 node
Tesla Cluster
64-node
Tesla Cluster
Tesla Personal
Supercomputer
256-512 node
CPU Cluster
64-node
CPU Cluster
Workstation
Cost
$5k- $10 K
$500K - $1M
$3M+

5000+ Customers / ISVs

Life Sciences & Medical Equipment		Productivity / Misc	Oil and Gas	EDA	Finance	CAE / Mathematica
Max Planck	GE Healthcare	CEA	Hess	Synopsys	Symcor	AccelerEyes
FDA	Siemens	NCSA	TOTAL	Nascentric	Level 3	MathWorks
barts Research	Techniscan	WRF Weather Modeling	CGG/Veritas	Gauda	SciComp	Wolfram
Medtronic	Boston Scientific	OptiTex	Chevron	CST	Hanweck	National Instruments
AGC	Eli Lilly	Tech-X	Headwave	Agilent	Quant Catalyst	Ansys
olved machines	Silicon Informatics	Elemental Technologies	Acceleware		RogueWave	Access Analytic
mith-Waterman	Stockholm Research	Dimensional Imaging	Seismic City		BNP Paribas	Tech-x
NA sequencing	Harvard	Manifold	P-Wave Seismic Imaging			RIKEN
AutoDock	Delaware	Digisens	Mercury Computer			SOFA
NAMD/VMD	Pittsburg	General Mills	ffA			Renault
olding@Home	ETH Zurich	Rapidmind				Boeing
oward Hughes Medical	Institute Atomic Physics	Rhythm & Hues				
RIBI Genomics		xNormal				
		Elcomsoft				
		LINZIK				

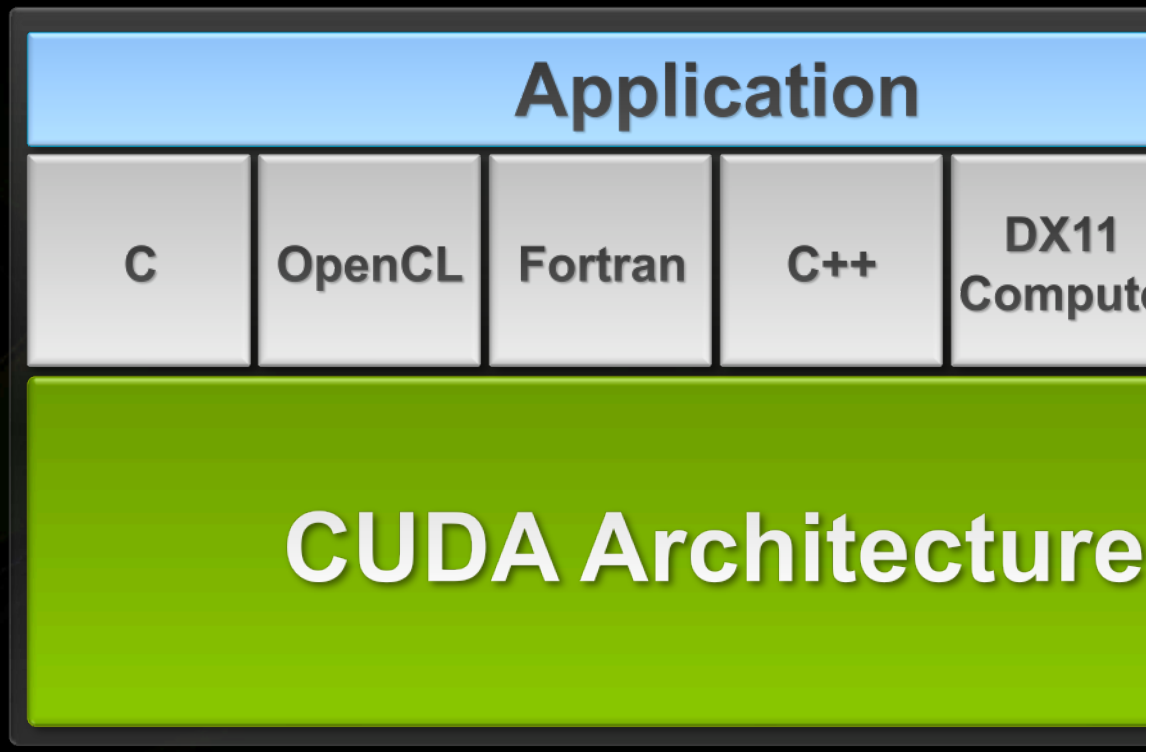


CUDA Parallel Computing Architecture

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)



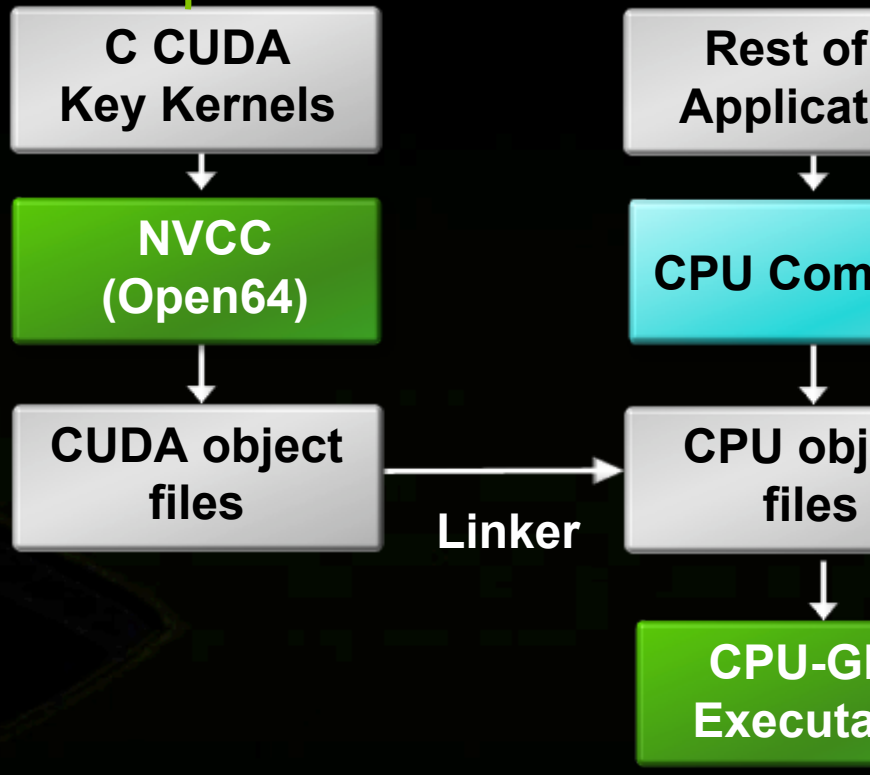
Compiling C for CUDA Applications

```
void serial_function(… ) {
  ...
}
void other_function(int ... ) {
  ...
}
```

```
void saxpy_serial(float ... ) {
   for (int i = 0; i < n; ++i)
      y[i] = a*x[i] + y[i];
}
```

```
void main( ) {
  float x;
  saxpy_serial(..);
  ...
}
```

Write Parallel CUDA code



C for CUDA : C with few keywords

```
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

```
__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);
```

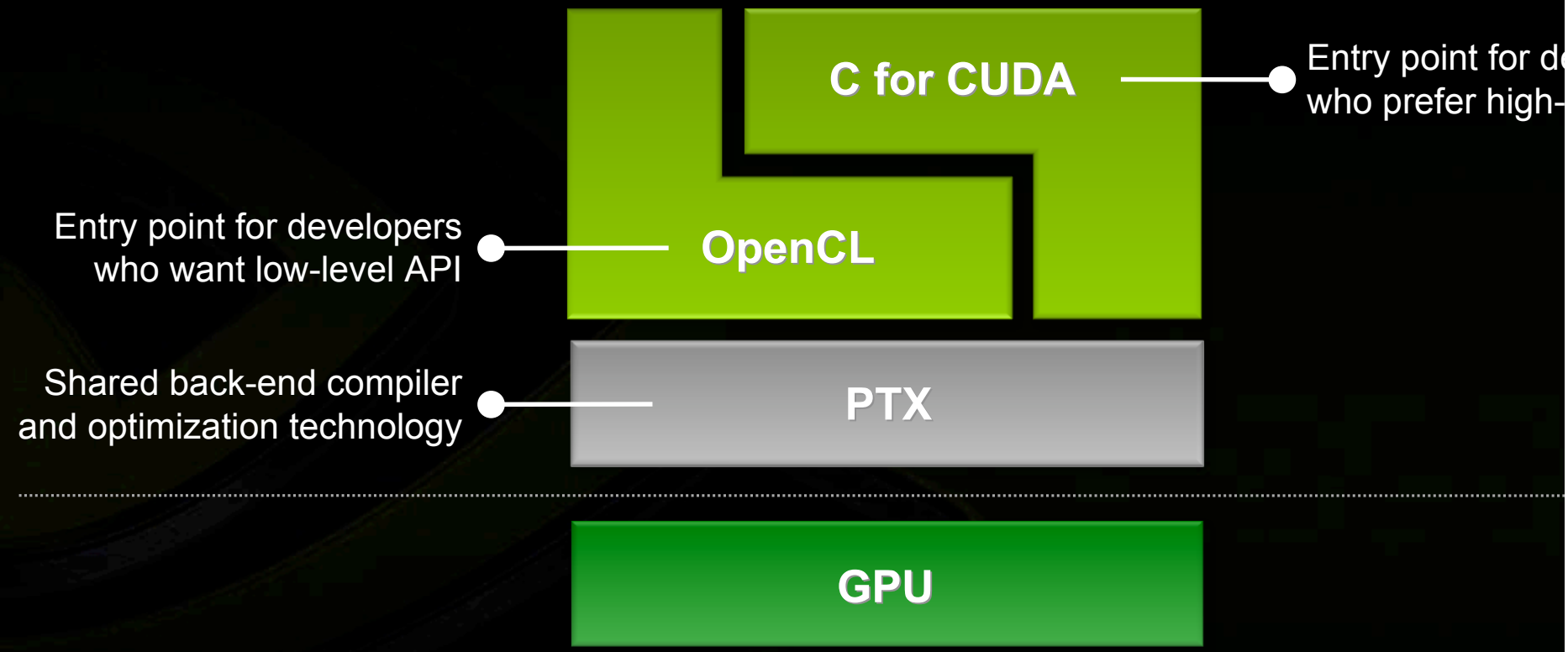
Parallel C

Support for Fortran, Python

- PGI Fortran to CUDA compiler (alpha release)
- Fortran Wrapper for CUDA
  - FLAGON : Library of numerics for Fortran 95
- Full Fortran support in Q4 2009
- Also Available
  - PyCUDA : Python Wrapper for CUDA
  - Java wrappers, .NET integration

Available from www.nvidia.com/tesla

NVIDIA C for CUDA and OpenCL

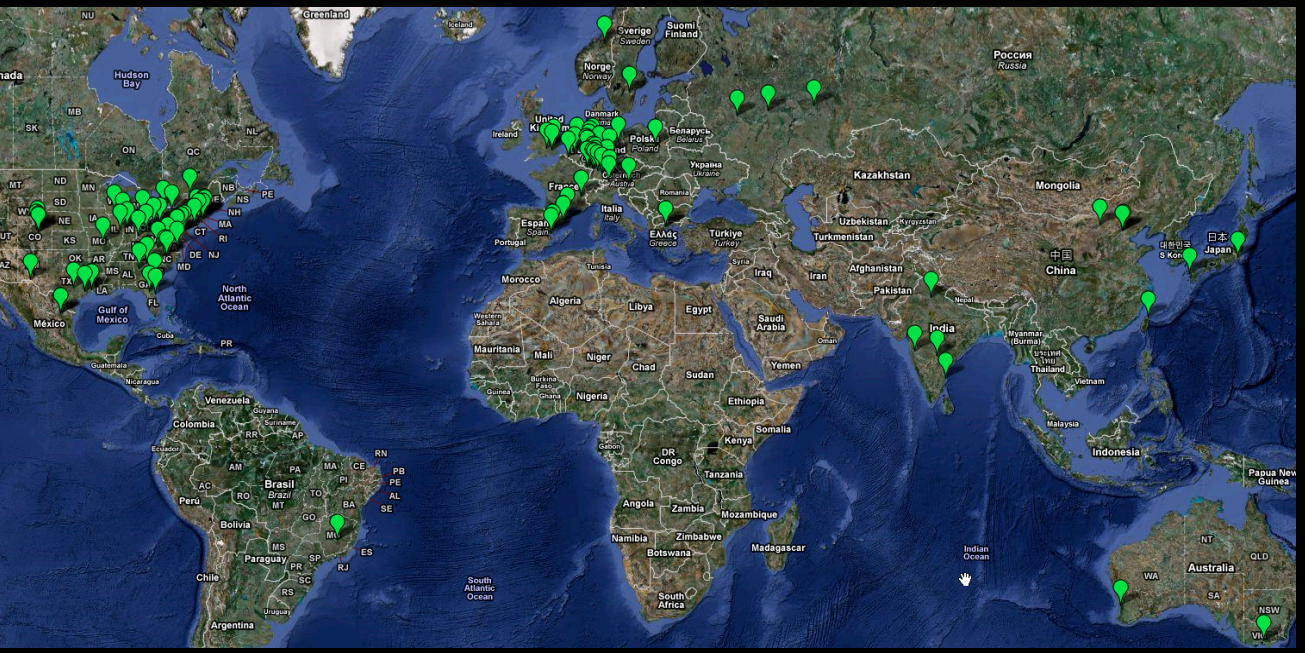


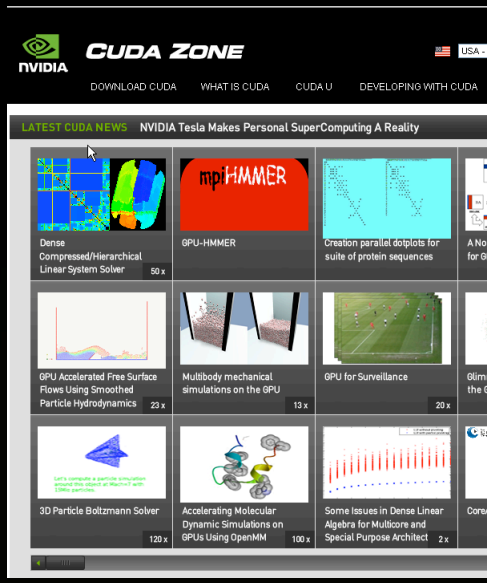
Different Programming Styles

- 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

CUDA Facts

- 900+ Research Papers
- 115+ universities teaching CUDA





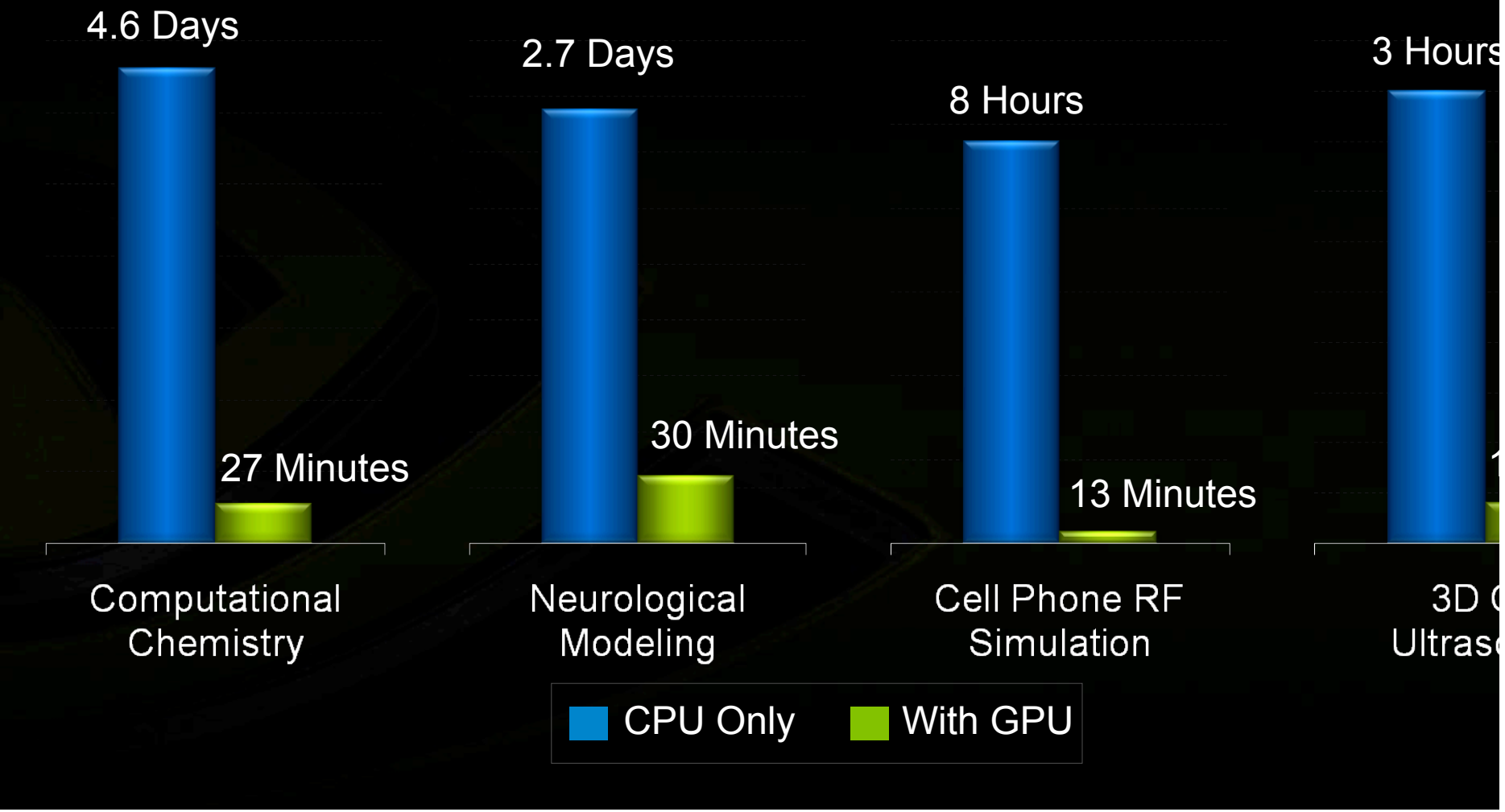
www.NVIDIA.com

- 200+ papers and a
- 110 Million CUDA-E
- 60,000+ Active D

Application Domains

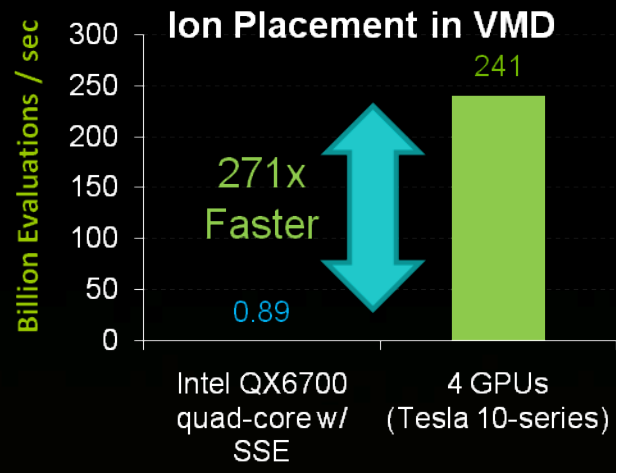
Details at
http://www.nvidia.com/object/vertical_solutions.html

Accelerating Time to Discovery

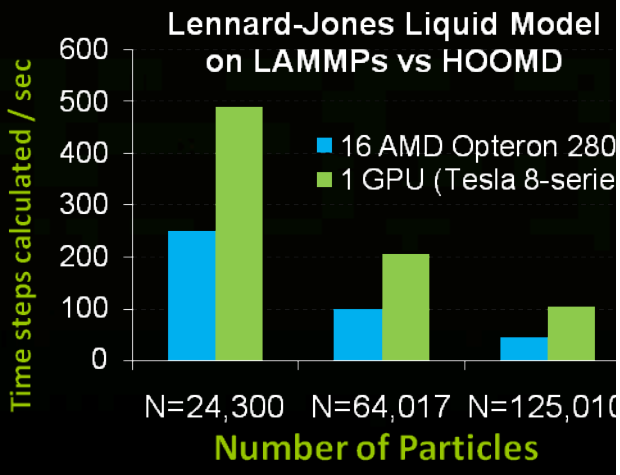


Molecular Dynamics

- Available MD software
  - NAMD / VMD (alpha release)
  - GROMACS (alpha release)
  - HOOMD
- OpenMM : Library for molecular modeling https://simtk.org/home/openmm
  - Jump start work on molecular dynamics



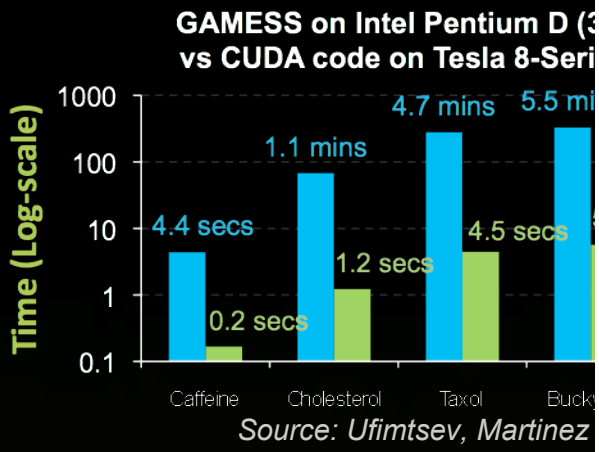
Source: Stone, Phillips, Hardy, Schulten



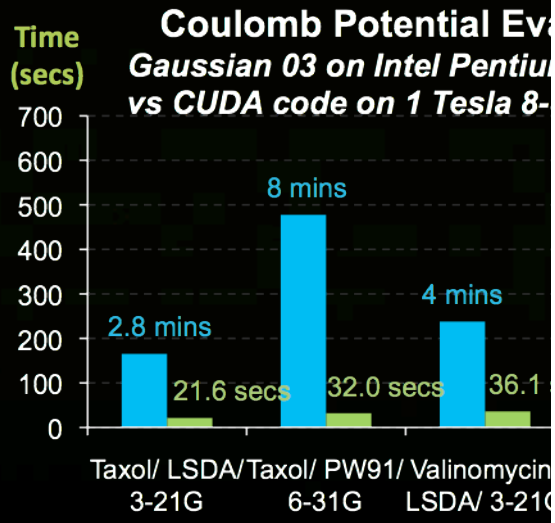
Source: Anderson, Lorenz, Travesset

Quantum Chemistry

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



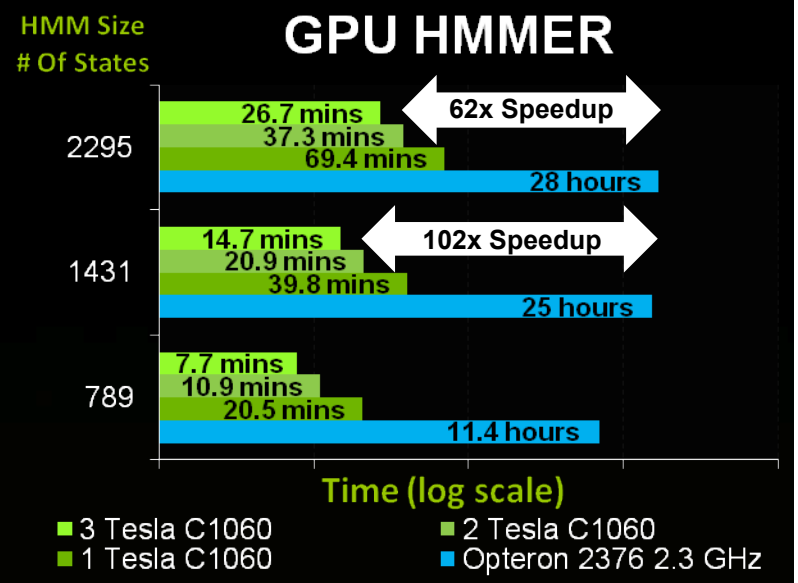
Source: Ufimtsev, Martinez

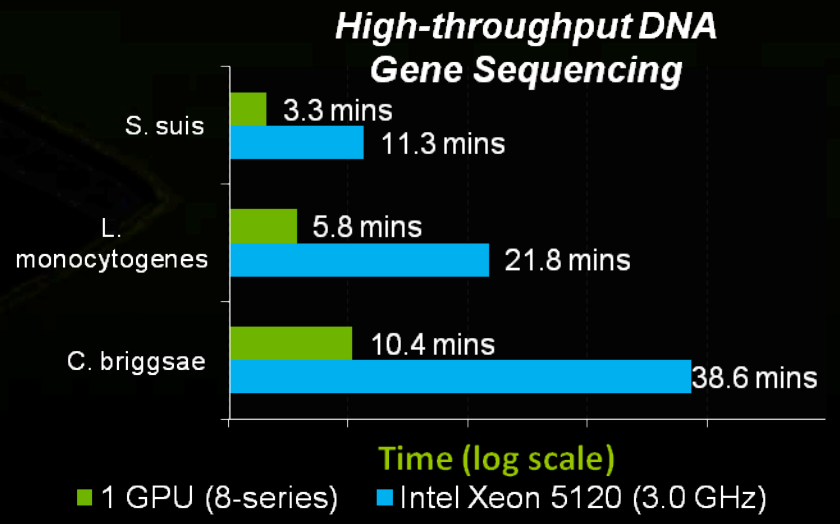


Source: Yasuda

Bio-Informatics

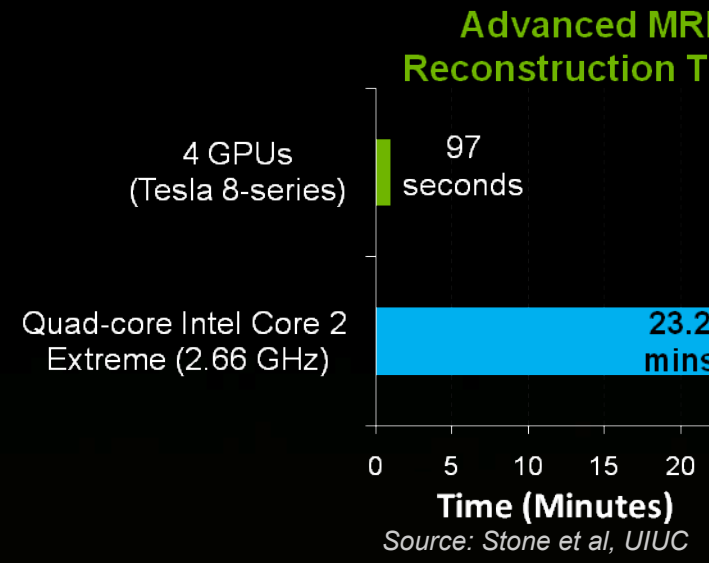
- Available applications
  - GPU HMMER
  - MUMmerGPU sequencing
  - MATLAB acceleration
  - Protein docking



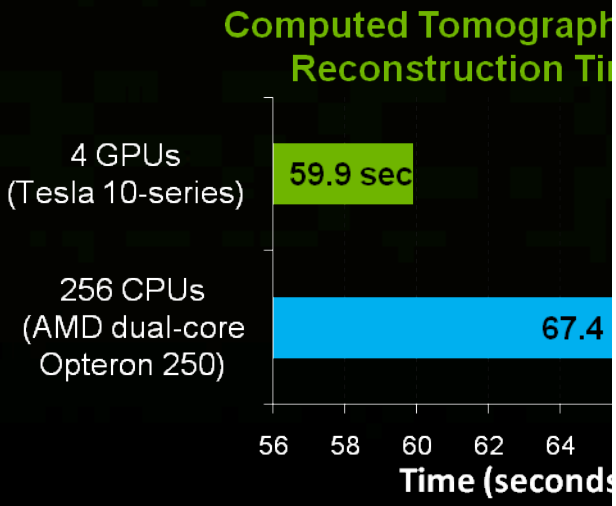


Medical Equipment

- GE Healthcare : CT
  - 40% increase in CT resolution
  - 2x increase in frame rate
- Techniscan: Ultra-sound
  - High resolution ultra-sound
  - 2x increase in acquisition
- Digisens : Tomography
  - Tomography reconstruction
- Several others on
  - X-Ray, Flow Cytometry, MRI, etc



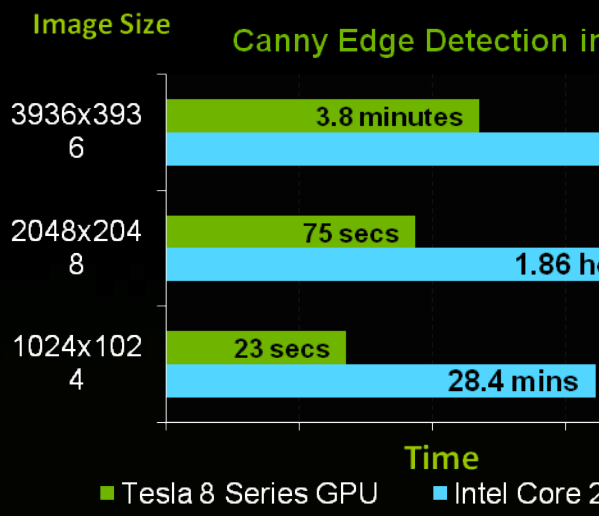
Source: Stone et al, UIUC

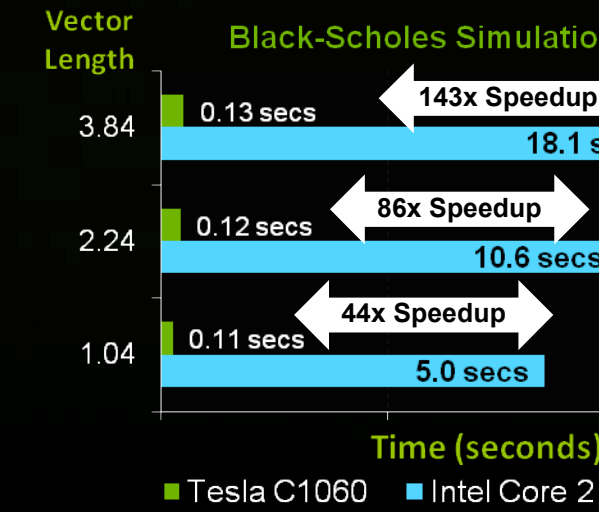


Source: Batenburg, Sijbers, e

MATLAB: Acceleration with GPUs

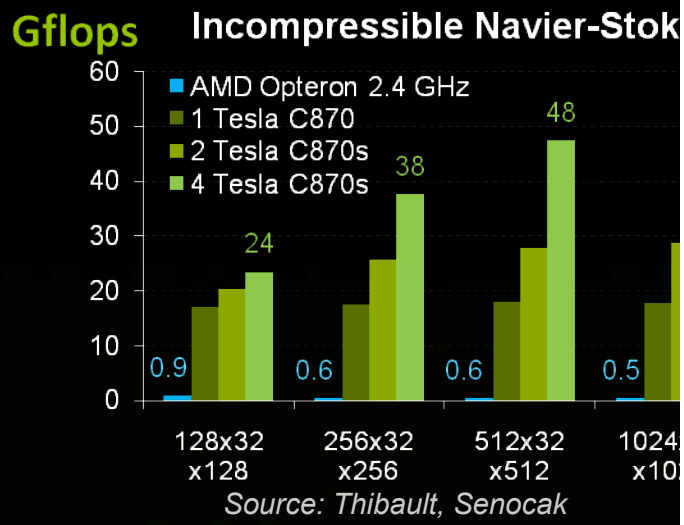
- Jacket CUDA plugin from Accelereyes
  - http://www.accelereyes.com
  - 15-day trail version avaiable
- Tesla GPU in a workstation
  - For MATLAB and research



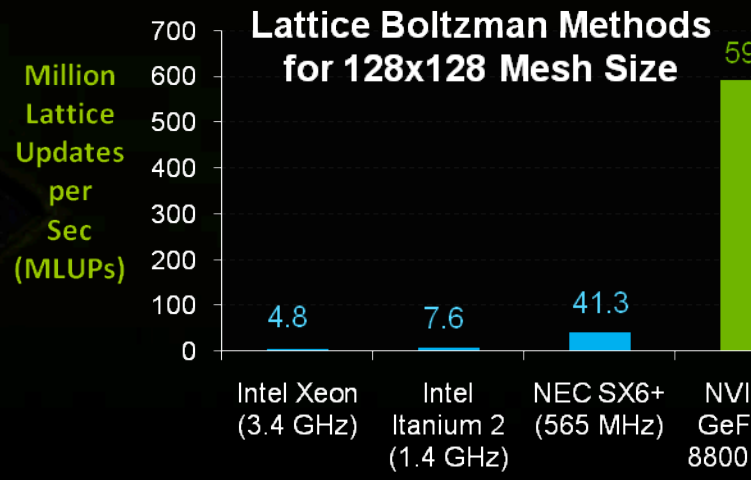


Computational Fluid Dynamics (CFD)

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



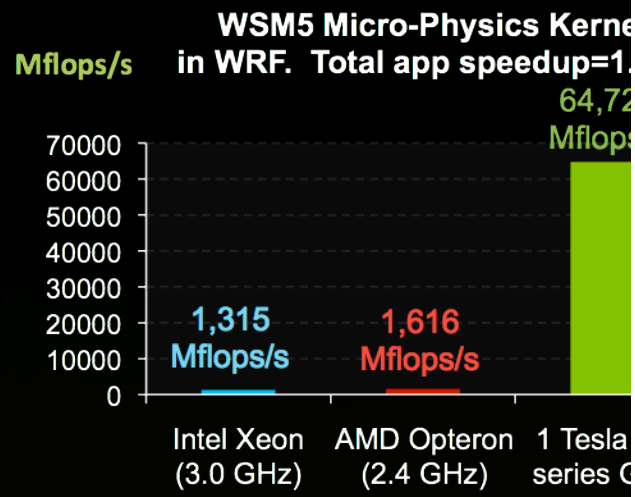
Source: Thibault, Senocak



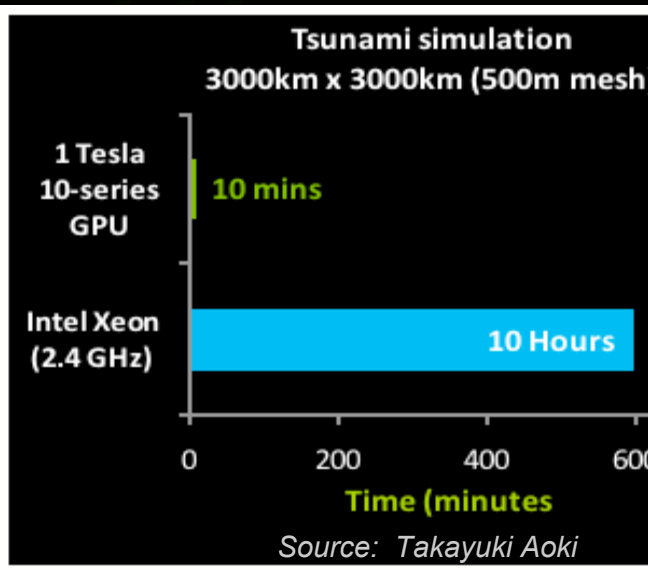
Source: Tolke, Krafczyk

Weather, Atmospheric, & Ocean Modeling

- **CUDA-accelerated WRF available**
  - 25-30% speedup in WRF so far
  - Other kernels in WRF being ported
- **Ongoing work**
  - Tsunami modeling
  - Ocean modeling
  - Several CFD codes



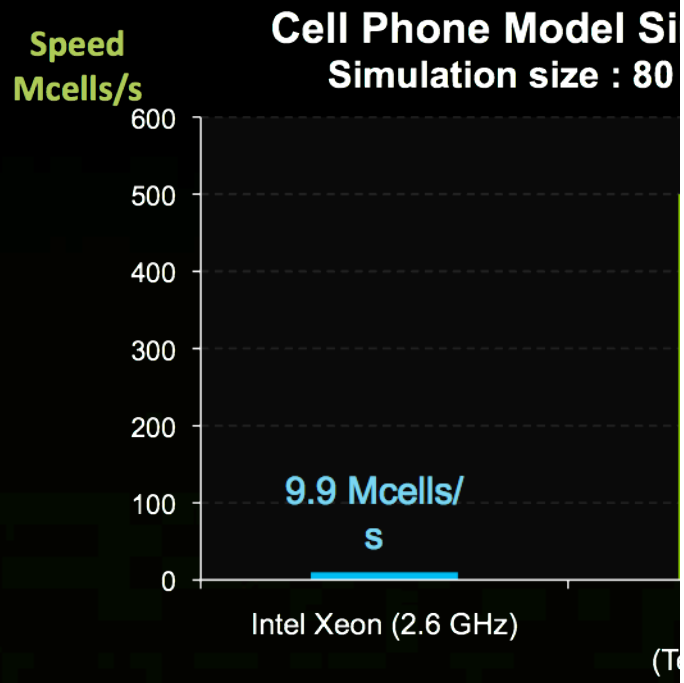
Source: Michalakes, Vachharajani



Source: Takayuki Aoki

Electromagnetics / Electrodynamics

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

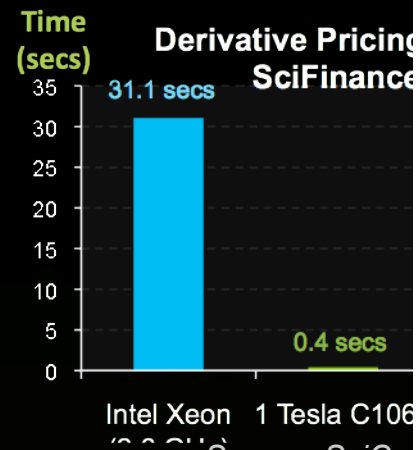


FDTD Acceleration usi

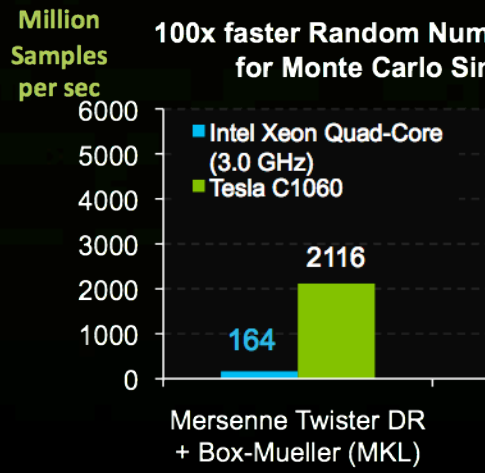
Source: Accelewa

Computational Finance

- **Financial Computing Software vendors**
  - SciComp : Derivatives pricing modeling
  - Hanweck: Options pricing & risk analysis
  - Aqumin: 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

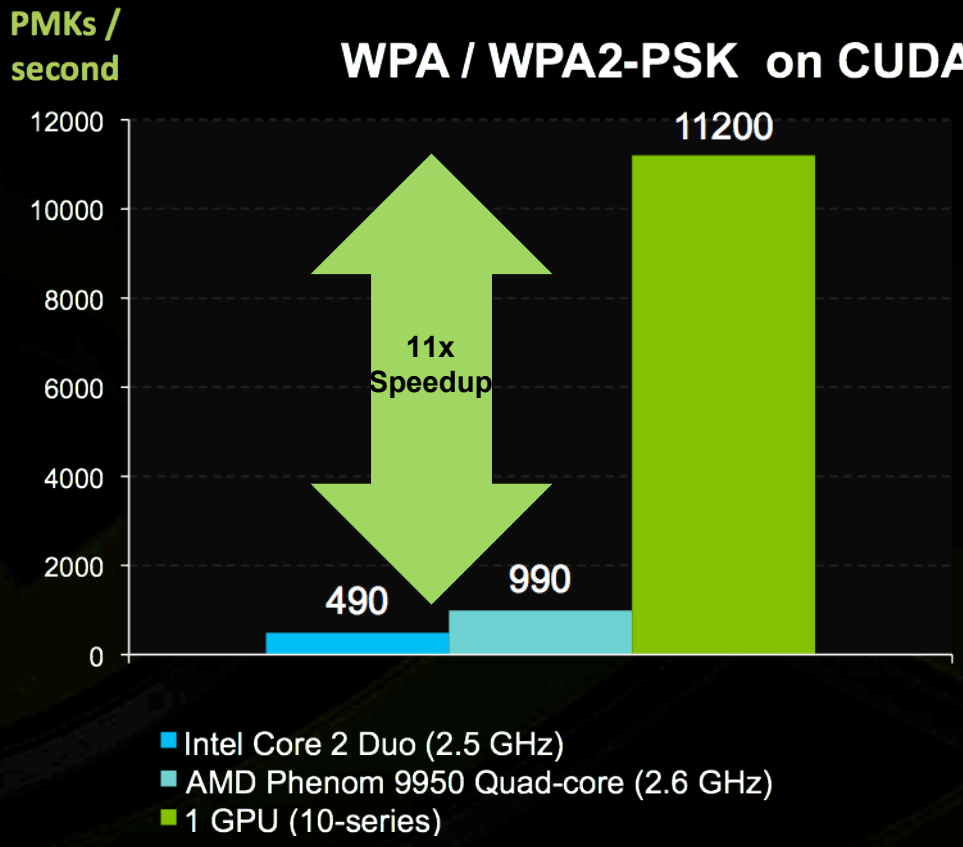


Source: SciCo

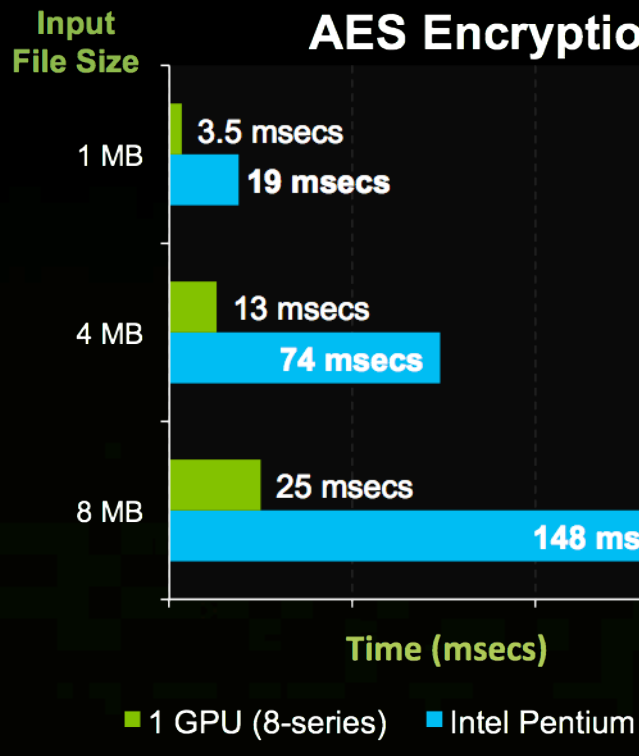


Source: CUDA

Encryption

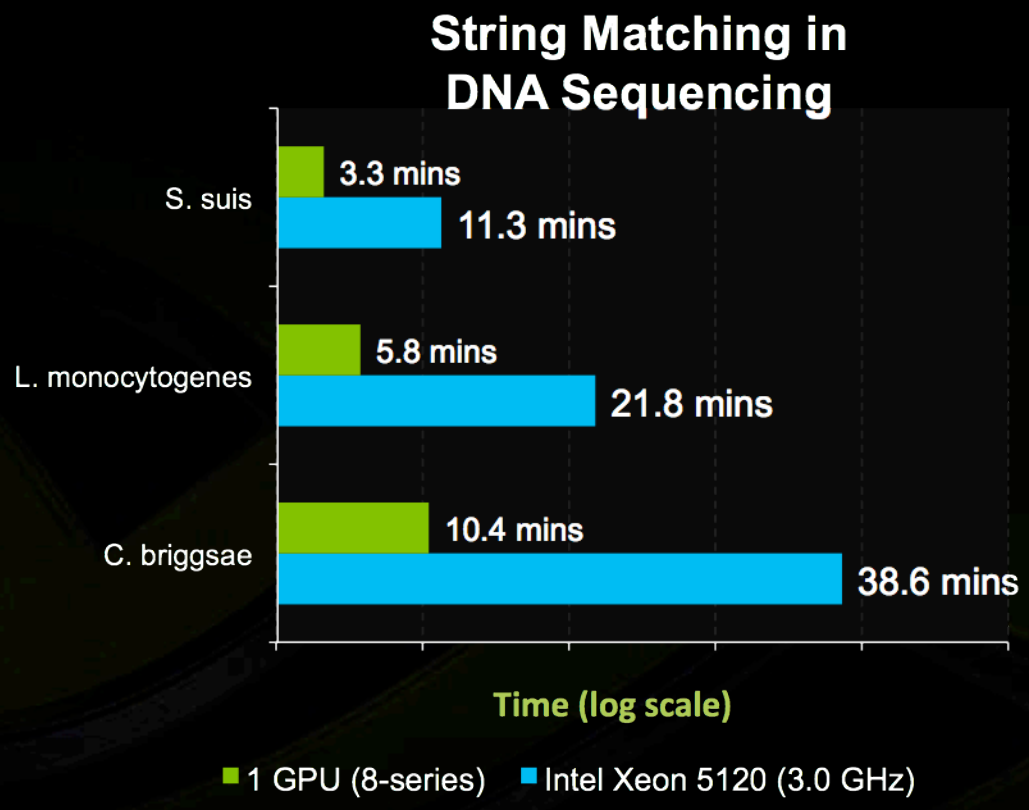


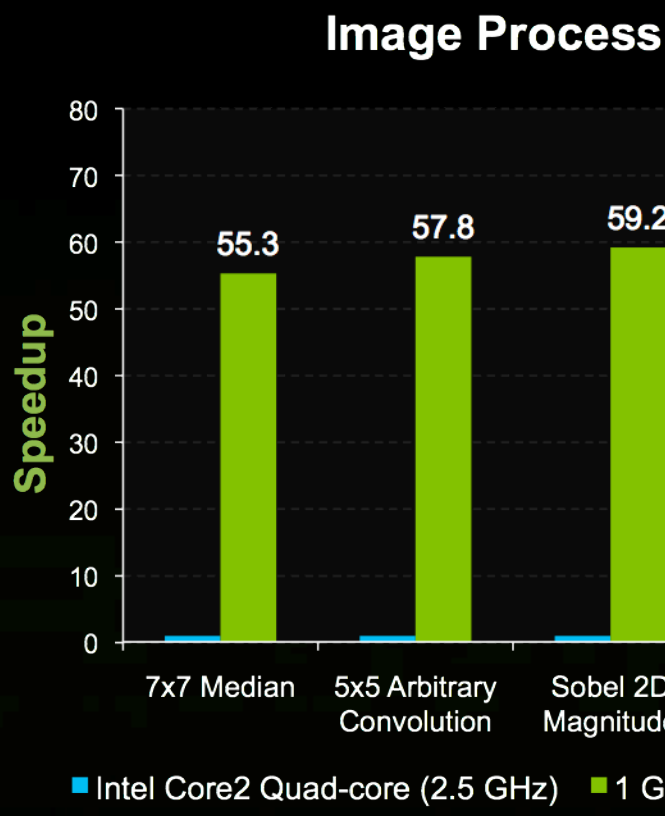
Source: http://code.google.com/p/pyrit/



Source: S. Manavski

Pattern Matching



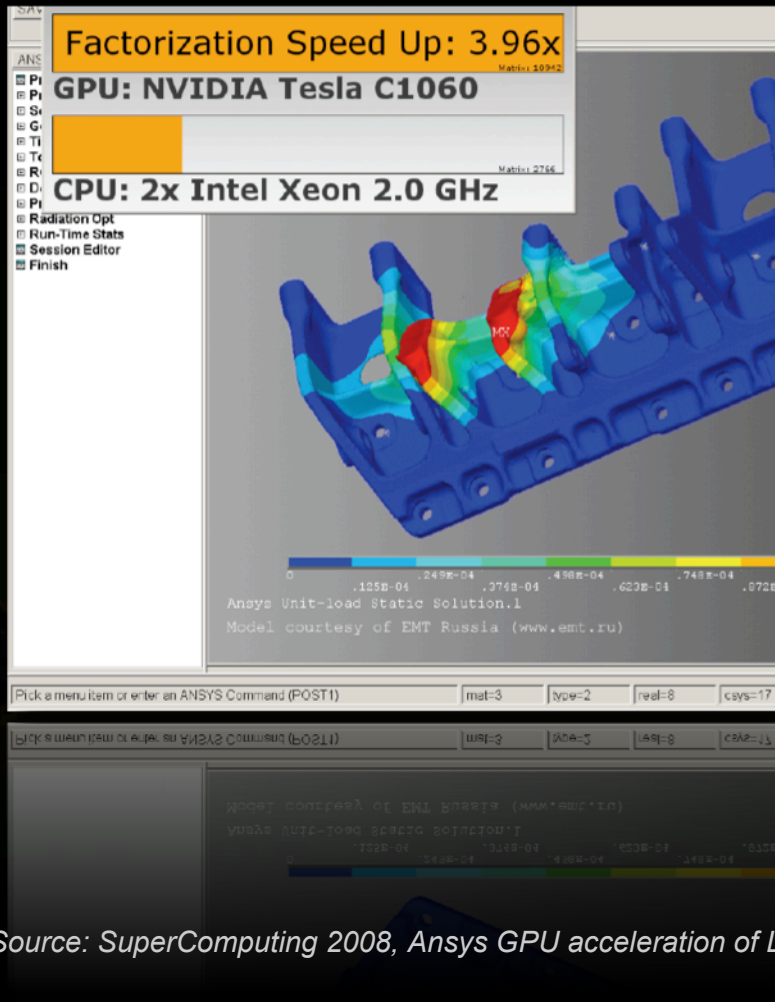


ANSYS Mechanical (Beta)

"CUDA allows us to leapfrog in a very compute intensive part of ANSYS."

Gene Poole
Principle SW Developer, ANSYS

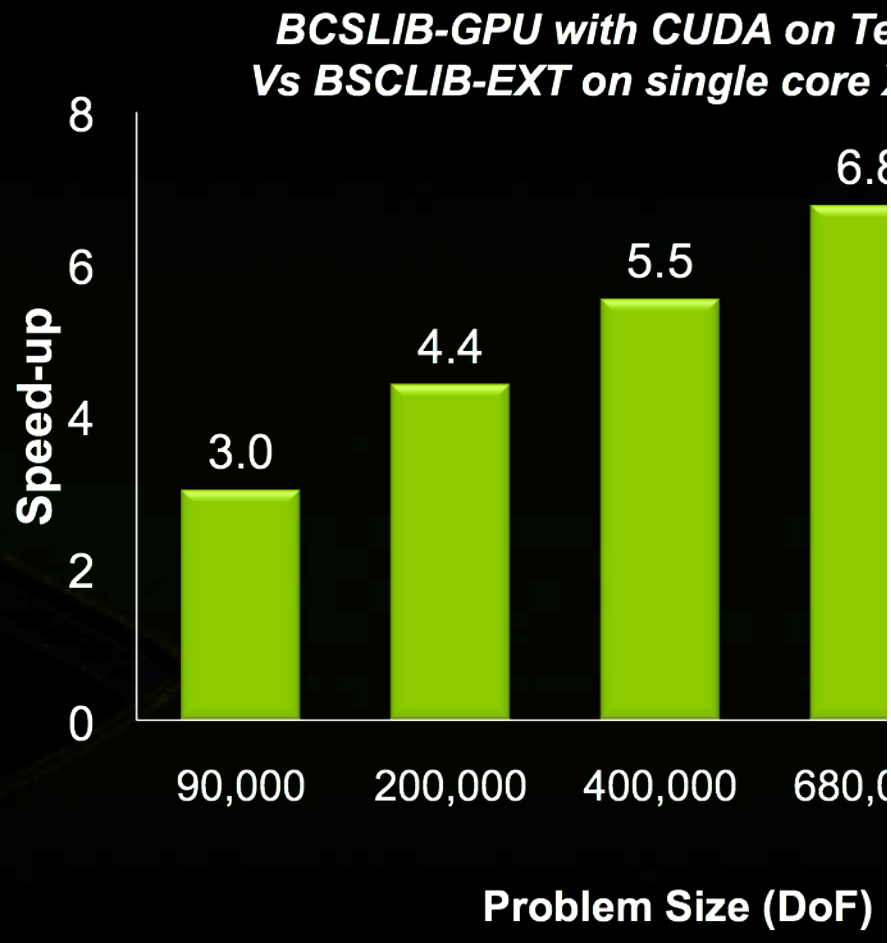
Hear Gene at http://www.youtube.com/nvidiatesla



Source: SuperComputing 2008, Ansys GPU acceleration of L

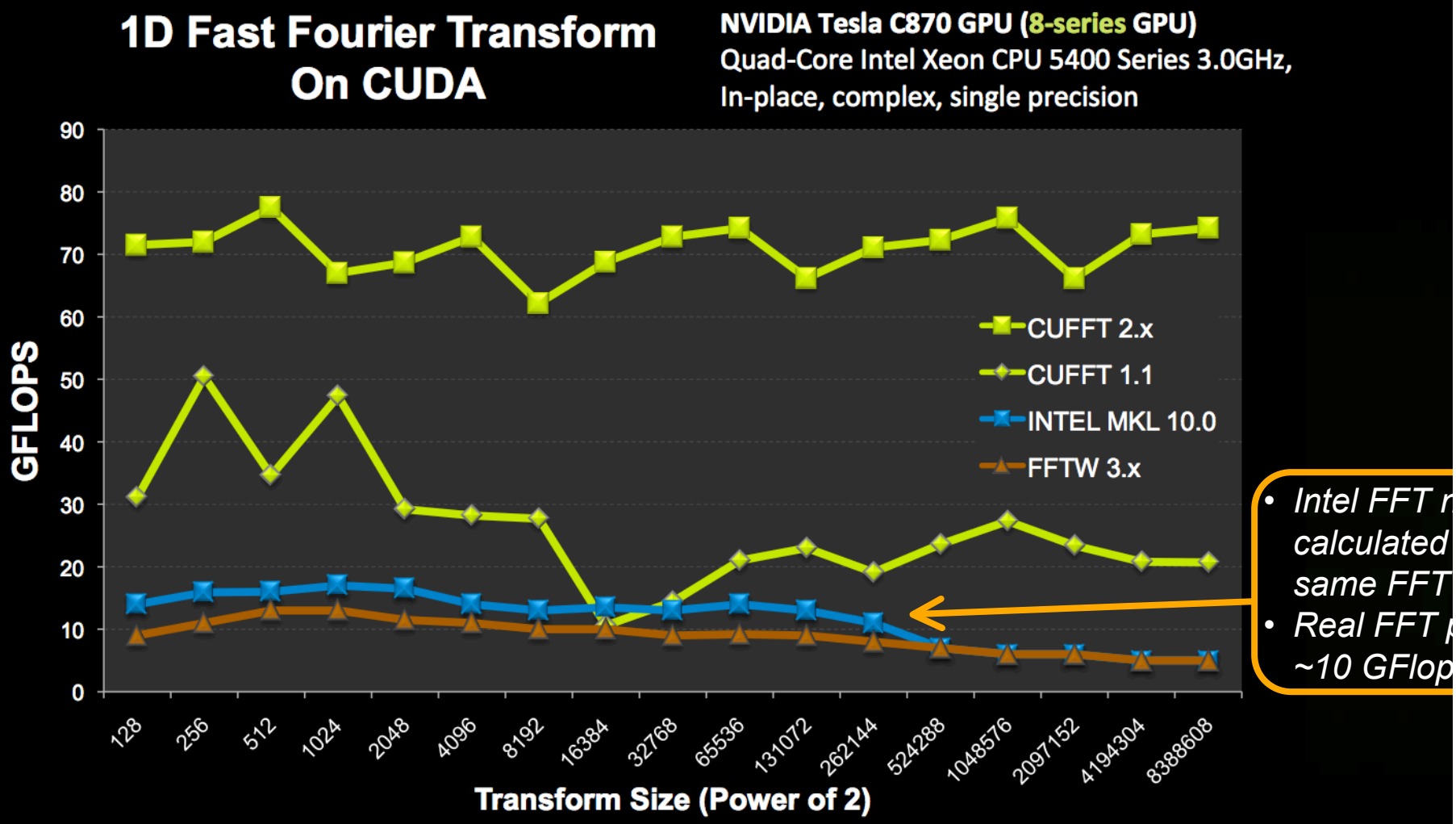
BCSLib-GPU : Sparse-Matrix Solver

- Used several solvers from Ansys, MSC
- Statics, Lanczos iterations, and Dynamics analyses
  - Not just LDL^T factorization
- Beta now, Release April 09



Libraries

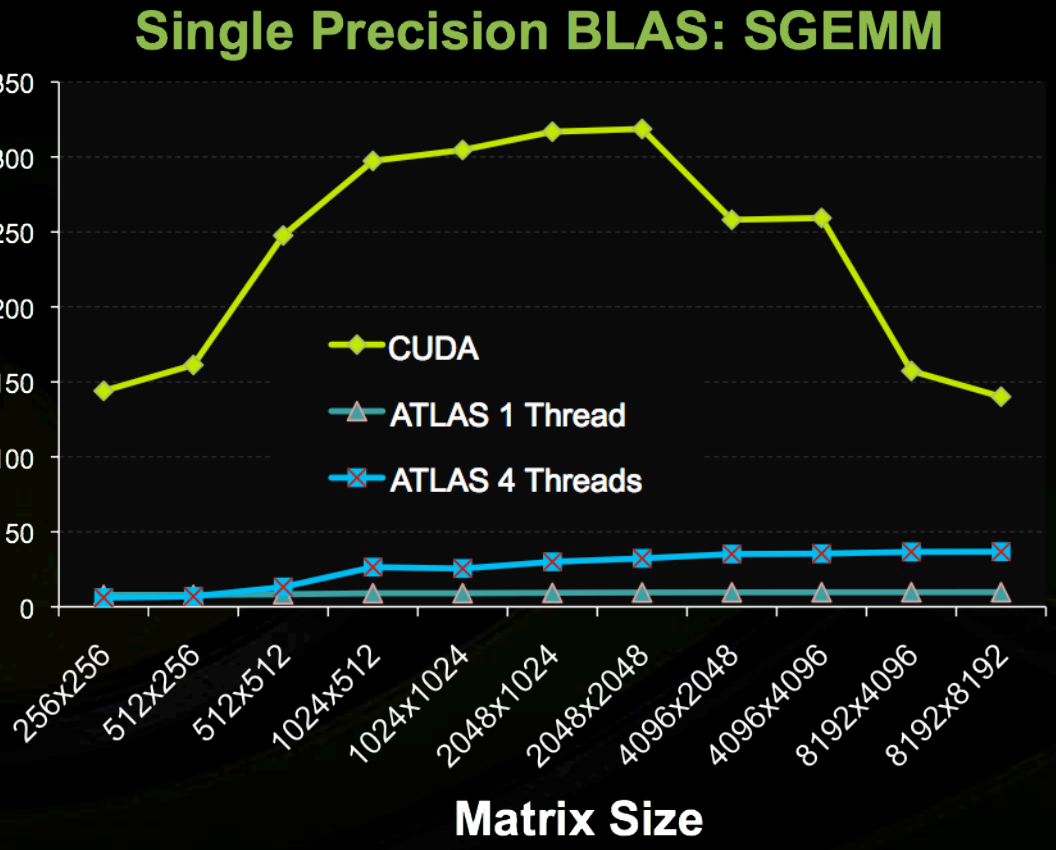
FFT Performance: CPU vs GPU (8-Series)

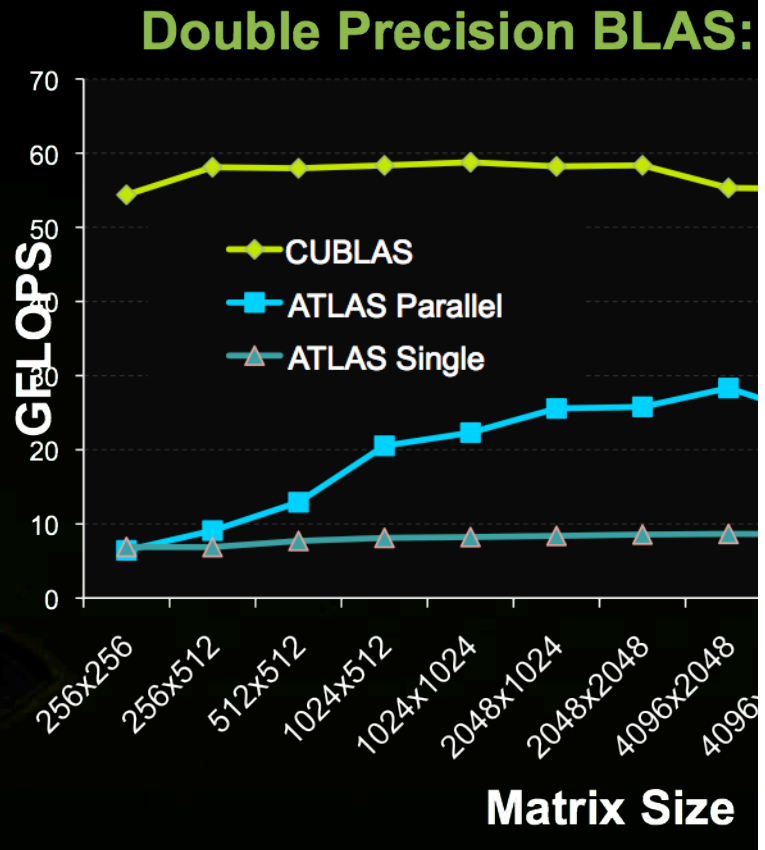


- *Intel FFT r*
  calculated
  same FFT
- *Real FFT p*
  ~10 GFlop

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

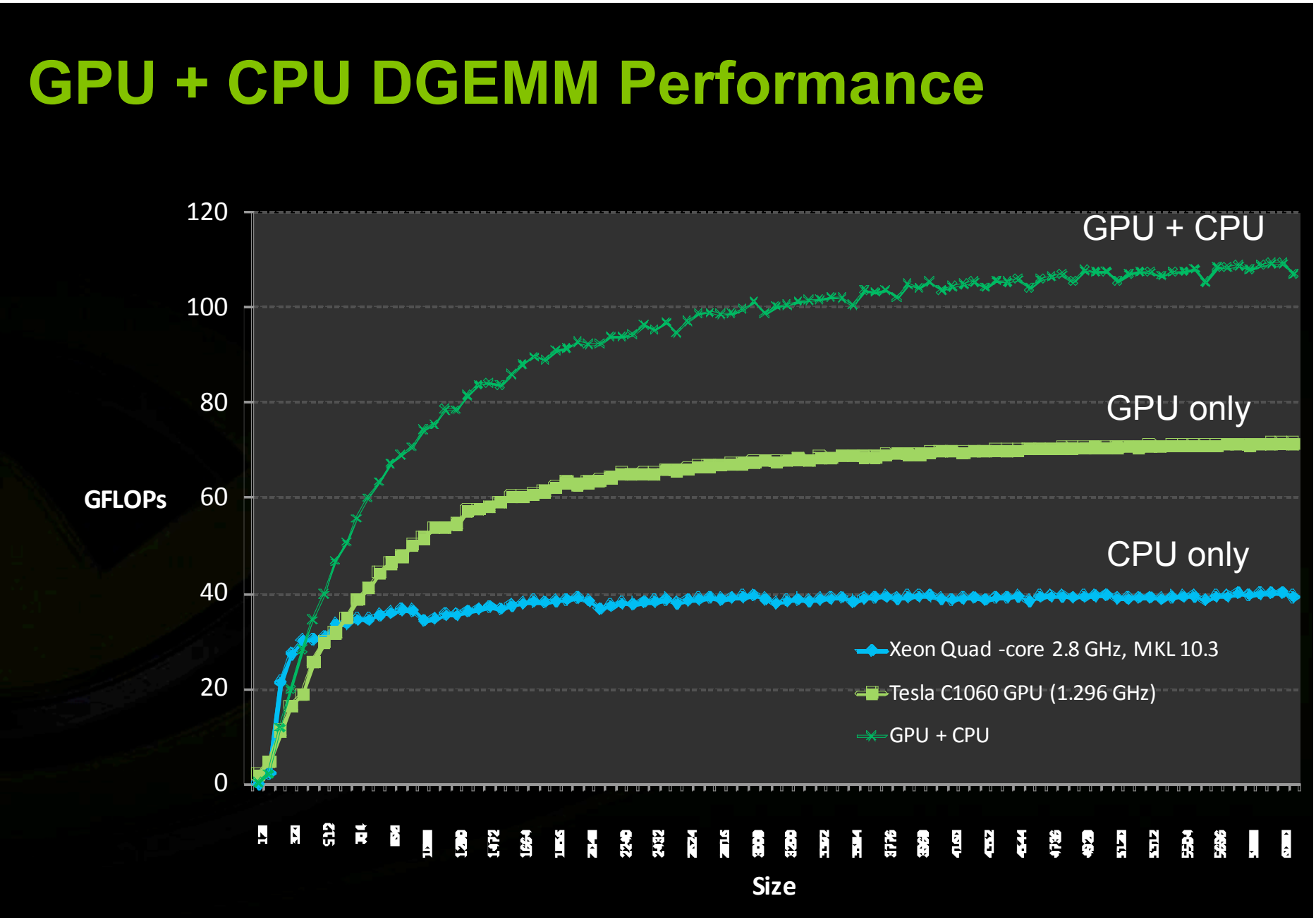
BLAS: CPU vs GPU (10-series)



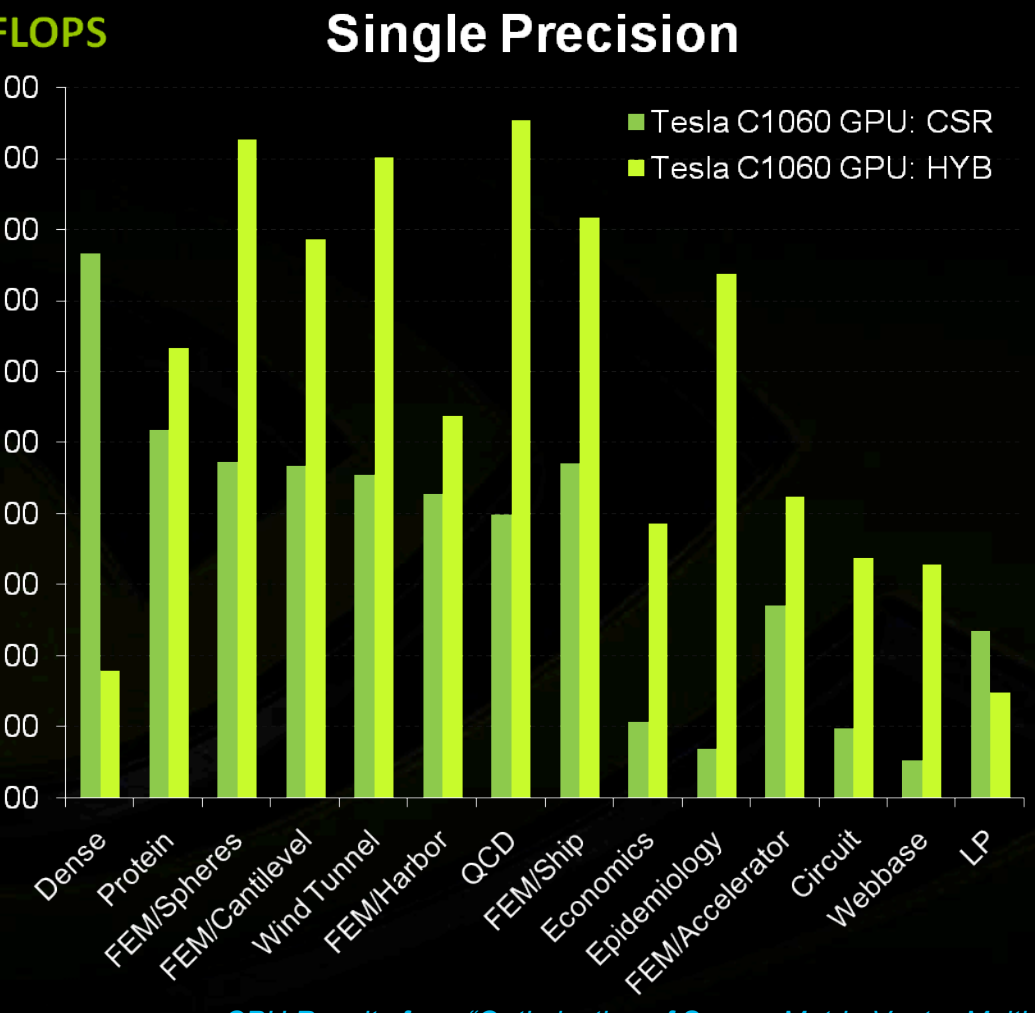


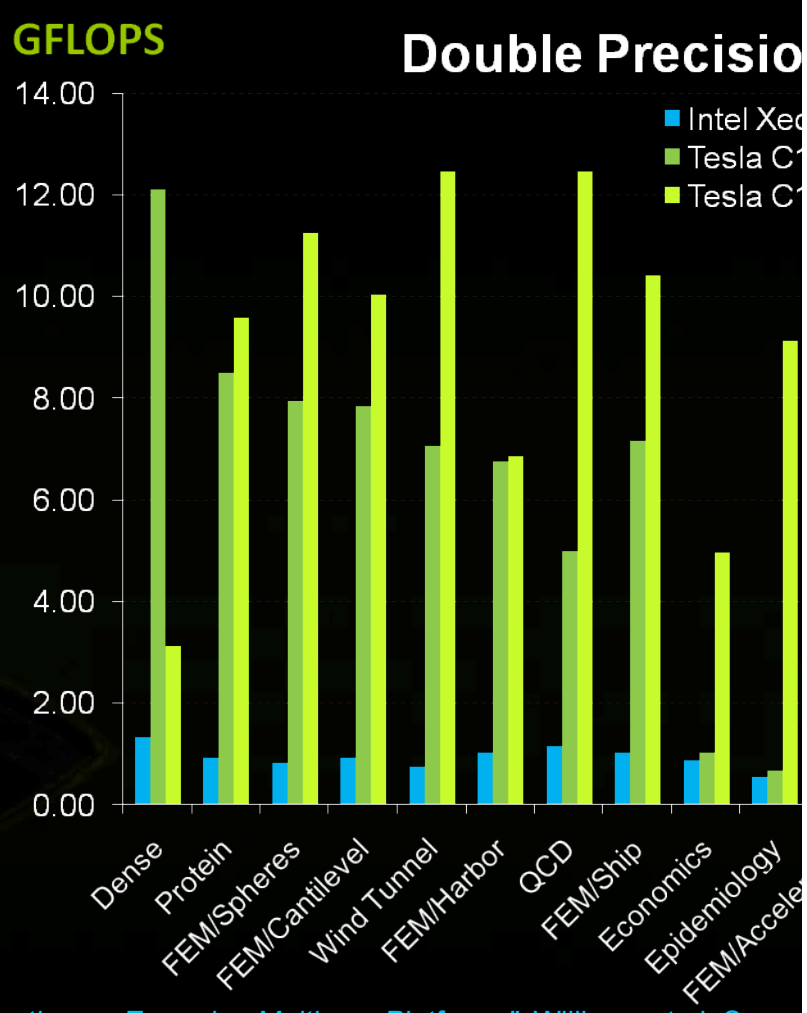
CUBLAS: CUDA 2.0, Tesla C1060 (**10-series** GPU)
ATLAS 3.81 on Dual 2.8GHz Opteron Dual-Core

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, Super

More Information

- Tesla main page
  - http://www.nvidia.com/tesla
  - Product Information
  - Software tools & libraries
- Vertical Solutions
  - http://www.nvidia.com/object/vertical_solutions.html
- CUDA Zone
  - http://www.nvidia.com
  - Applications, Papers,
  - Learn CUDA: self-pace tutorials
- Hear from Developers
  - http://www.youtube.com/nv

NVIDIA: Leadership in GPU computing

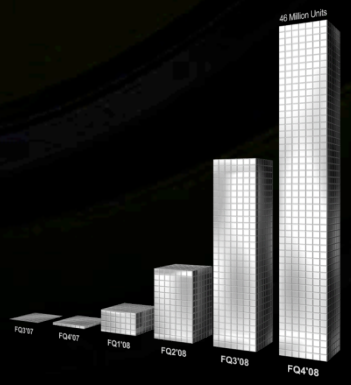
200+ Apps on CUDA Zone

30+ CUDA GPU clusters

115+ Universities Tea
900+ research p

Duke
Erlangen
ETH Zurich
Georgia Tech
Grove City College
Harvard
IISc Bangalore
IIIT Hyderabad
IIT
Illinois
INRIA
Iowa
ITESM
Johns Hopkins
Kent State
Kyoto
Lund
Maryland
McGill
MIT
North Carolina

110 M CUDA enabled GPUs
60,000+ active developers



150K CUDA compiler downloads