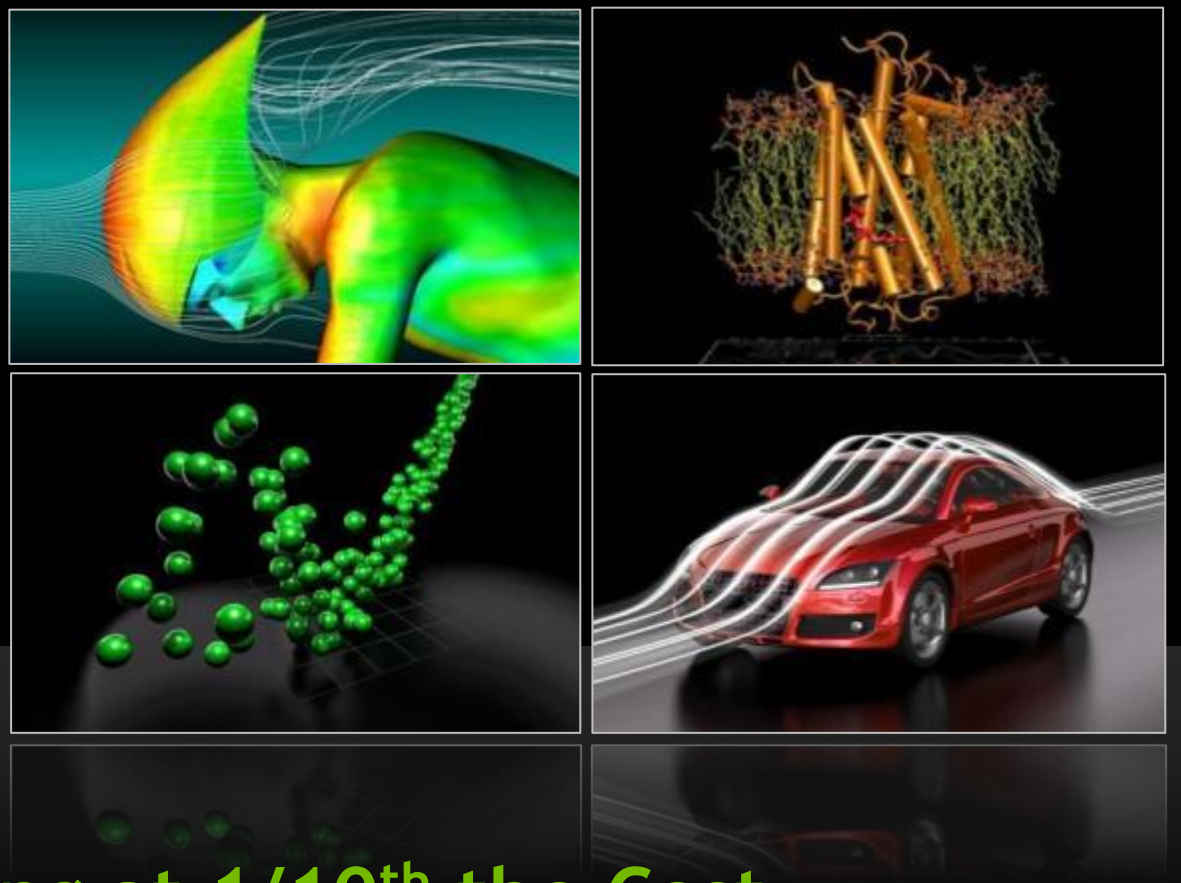
TESLA
GPU Computing



Supercomputing at 1/10th the Cost

http://www.nvidia.com/tesla

Mainstream Applications Going Parallel

CUDA Accelerates Adobe Mercury Playback Engine

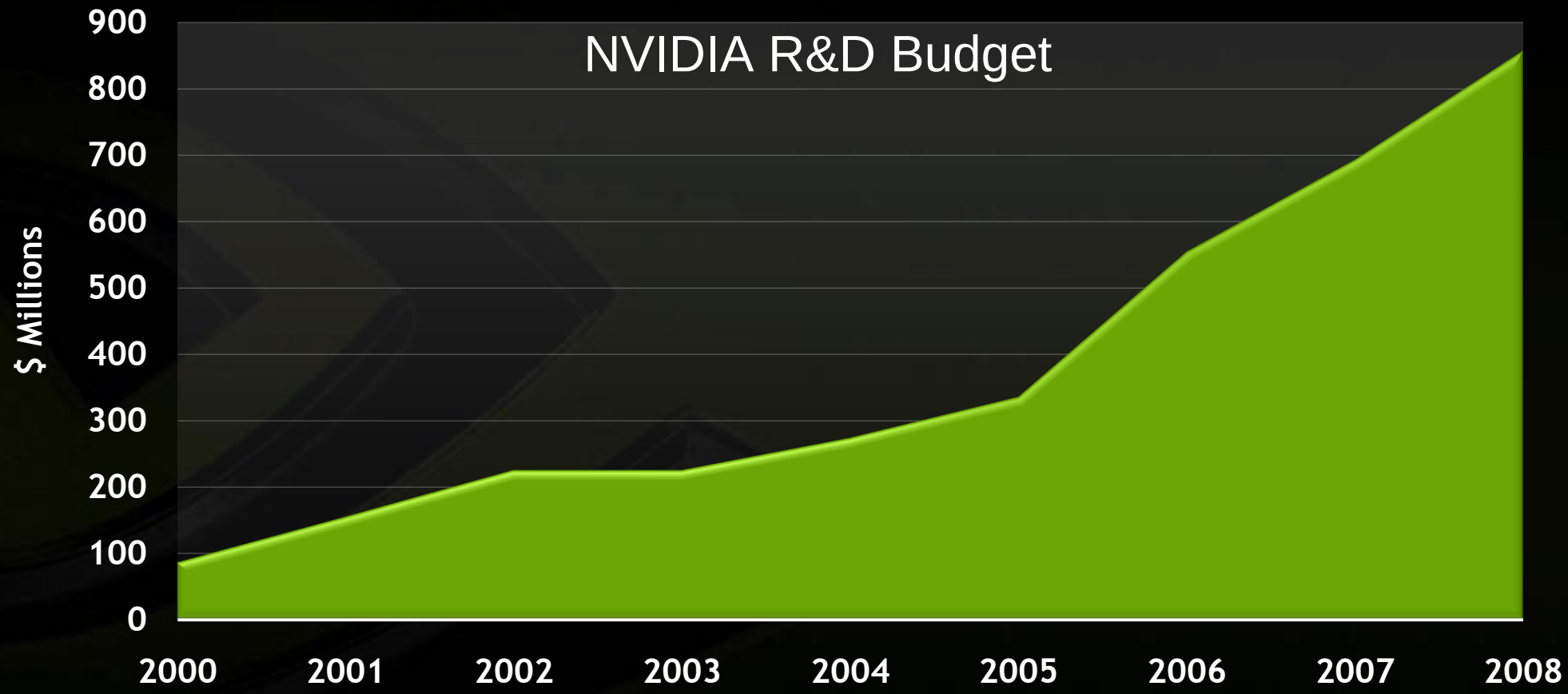
Amazingly fluid,
real-time video editing

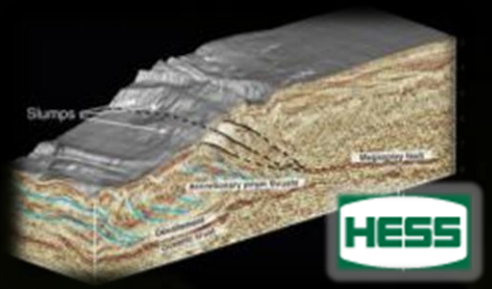
Quick preview of real time edits and effects

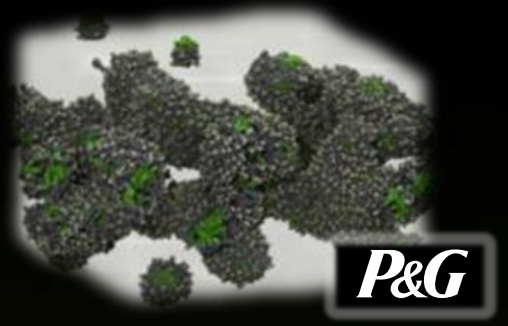
Realistic preview of final content

Faster encoding

GPU Innovation Accelerating




HESS


P&G


Bloomberg


Lowry
Digital


DIGITALGLOBE


BAE SYSTEMS


BECKMAN
COULTER

Max Planck Institute

National Center for Supercomputing Applications

Tokyo Institute of Technology

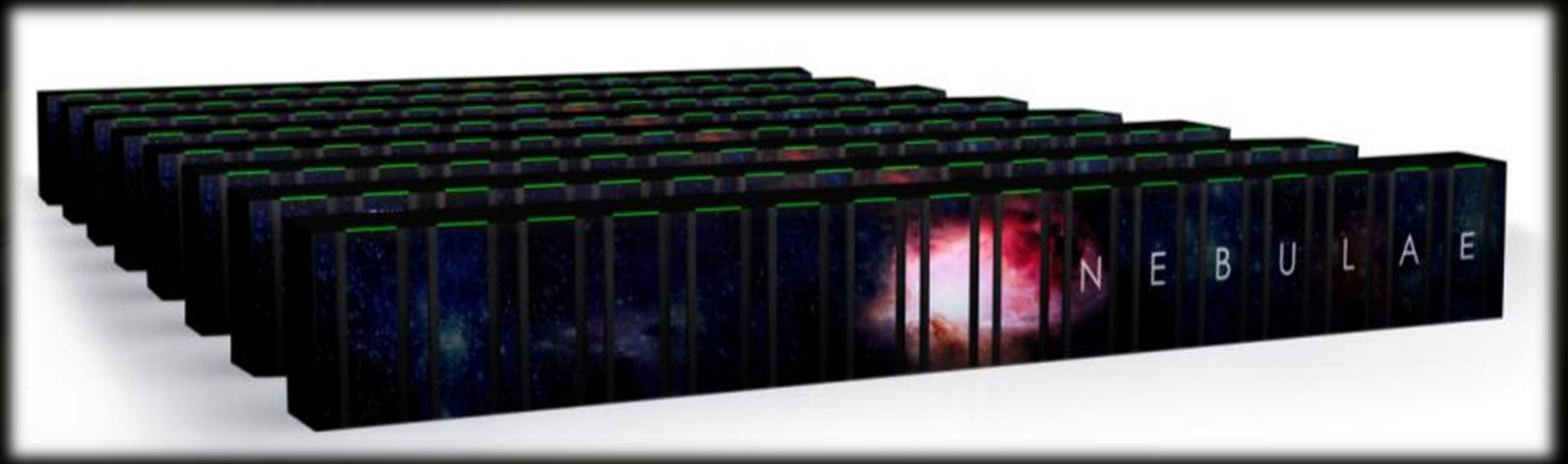
CSIRO

Dawning Nebulae

Second Fastest Supercomputer in the World

1.27 Petaflop

4640 Tesla GPUs




David B. Kirk
Wen-mei W. Hwu
Programming Massively Parallel Processors
A Hands-on Approach
基于GPU的多尺度离散模拟并行计算
CUDA
BY EXAMPLE
An Introduction to General-Purpose GPU Programming
OpenCL
並列プログラミング
マルチコアCPUとGPUのための標準フレームワーク
はじめての
CUDA
プログラミング
既存のPCに
ビデオカードを
挿すだけでOK！
ОСНОВЫ РАБОТЫ
С ТЕХНОЛОГИЕЙ
CUDA
NVIDIA
GPU
高性能运算之CUDA
OpenCL入門

1000+ GPU Clusters Around the World

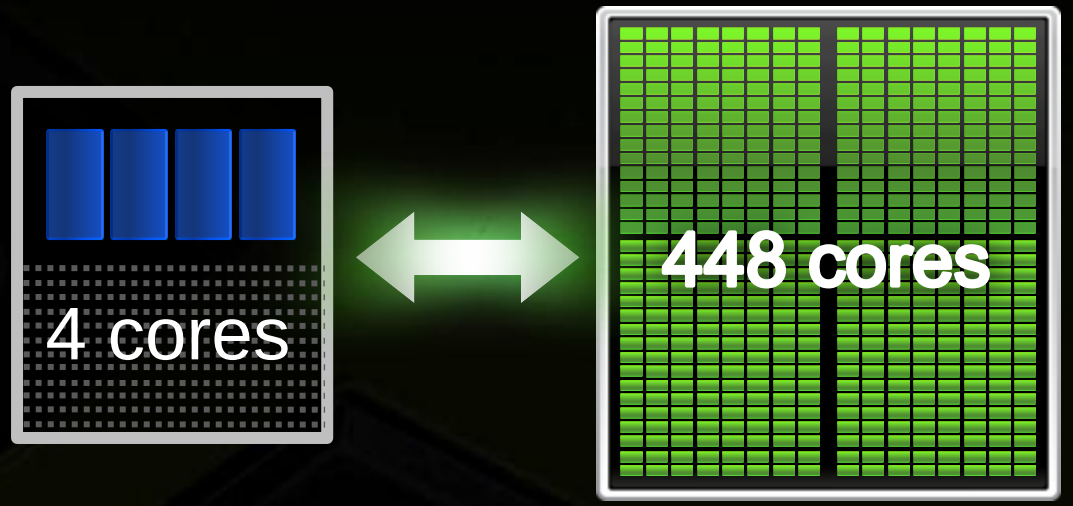
NVIDIA Tesla 20-Series Products

Data Center

Workstation

GPU Computing

CPU + GPU Co-Processing



CPU
48 GigaFlops (DP)

GPU
515 GigaFlops (DP)

NVIDIA Developer Eco-System

Numerical Packages
- MATLAB
- Mathematica
- NI LabView
- pyCUDA

Debuggers & Profilers
- cuda-gdb
- NV Visual Profiler
- Parallel Nsight
- Visual Studio
- Allinea
- TotalView

GPU Compilers
- C
- C++
- Fortran
- OpenCL
- DirectCompute
- Java
- Python

Parallelizing Compilers
- PGI Accelerator
- CAPS HMPP
- mCUDA
- OpenMP

Libraries
- BLAS
- FFT
- LAPACK
- NPP
- Video
- Imaging
- GPULib

GPGPU Consultants & Training

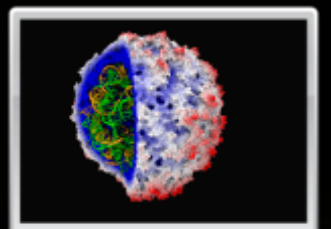
acceleware, GASS, STONE RIDGE TECHNOLOGY, ANEO, GPU Tech, TECH-X, CAPS, HPC, EM Photonics, SCALABLE GRAPHICS, WIPRO

OEM Solution Providers

DELL, hp, IBM, CRAY, ASUS, SUPERMICRO, sgi, FUJITSU, BULL, APPRO, lenovo 联想, NEC

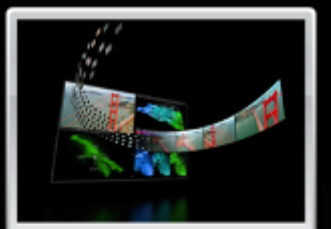
146X

Medical Imaging
U of Utah



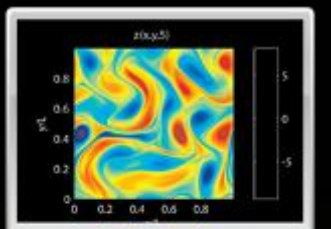
36X

Molecular Dynamics
U of Illinois, Urbana



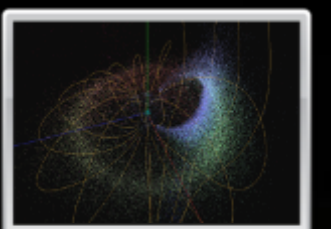
18X

Video Transcoding
Elemental Tech



50X

Matlab Computing
AccelerEyes



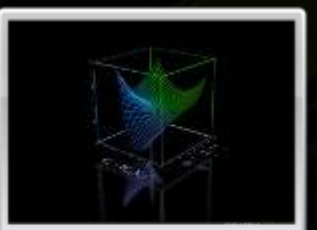
100X

Astrophysics
RIKEN

50x – 150x

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

Tesla Data Center & Workstation GPU Solutions

Tesla M-series GPUs
M2070 M2050 M1060

Tesla S-series 1U Systems
S2050 S1070

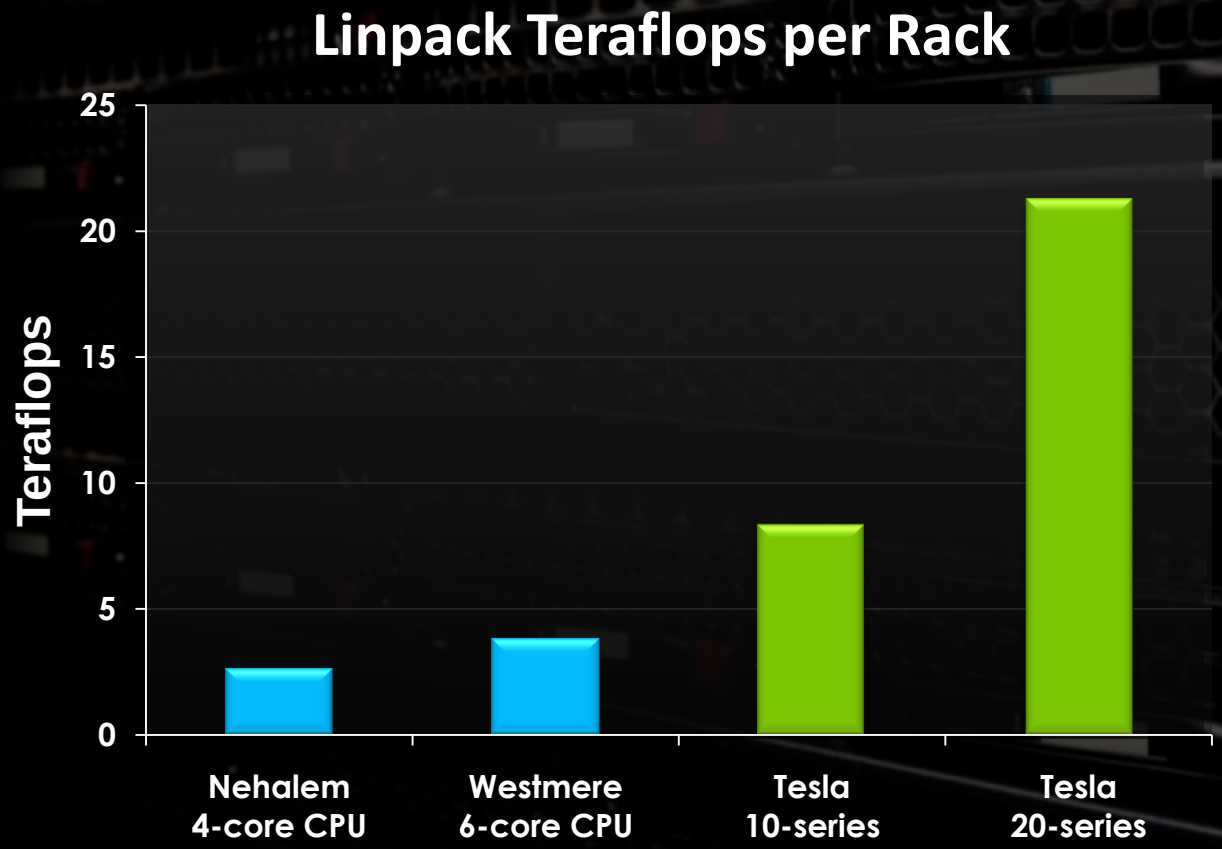
Tesla C-series GPUs
C2070 C2050 C1060

Integrated CPU-GPU
Servers & Blades

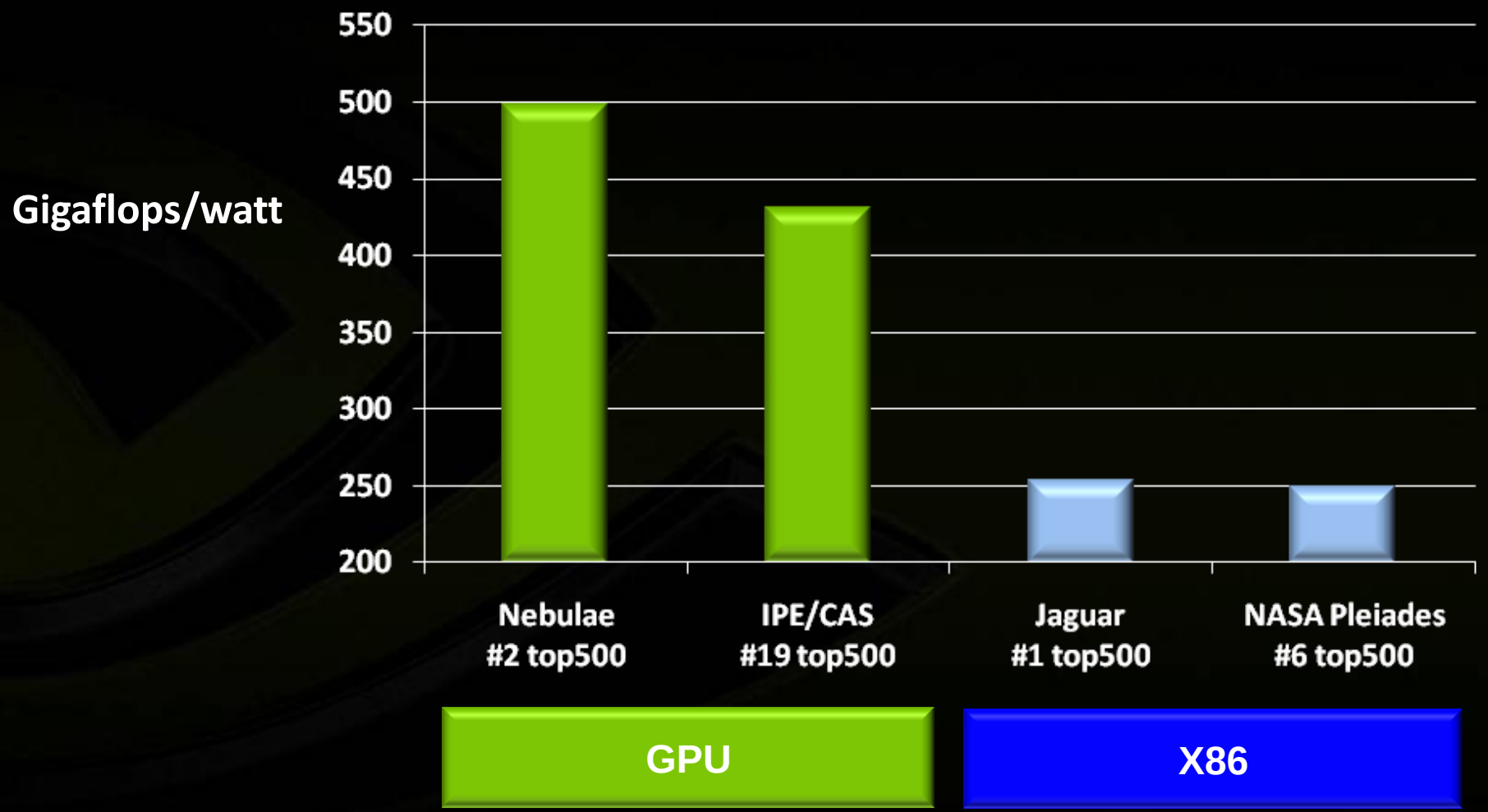
OEM CPU Server +
Tesla S-series 1U

Workstations
2 to 4 Tesla GPUs

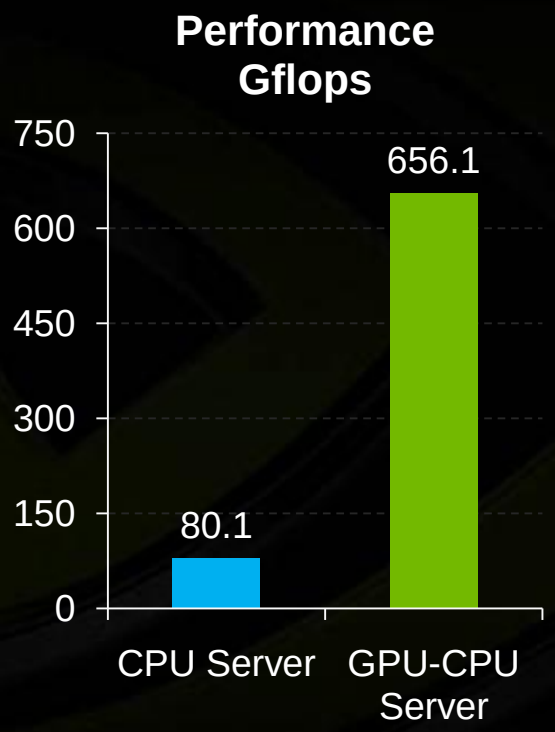
Soul of a Supercomputer

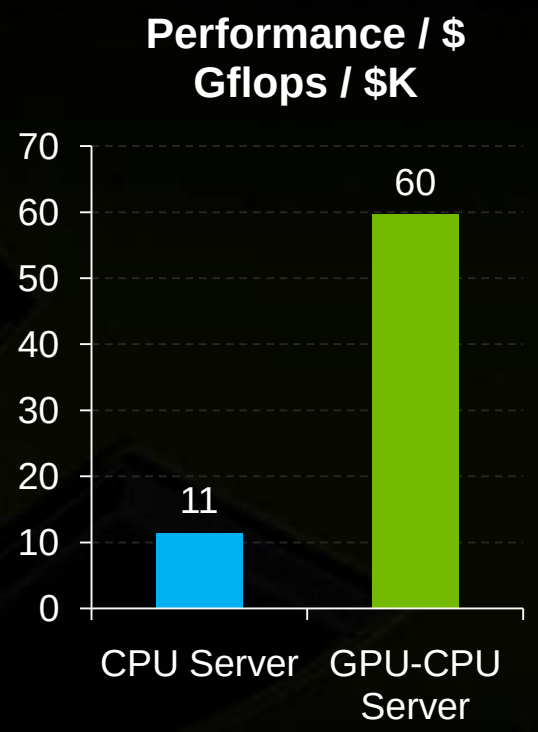


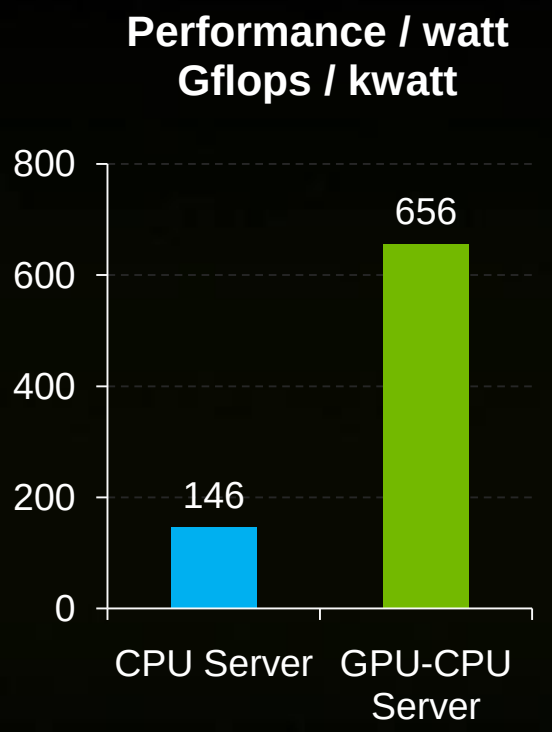
Heterogeneous = 2x Performance / Watt



8x Higher Linpack





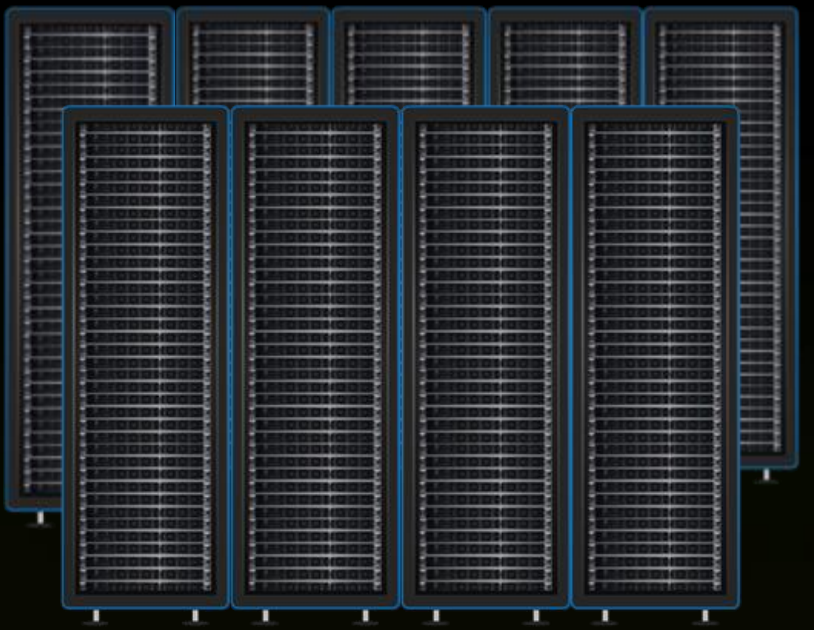


CPU 1U Server: 2x Intel Xeon X5550 (Nehalem) 2.66 GHz, 48 GB memory, $7K, 0.55 kw

GPU-CPU 1U Server: 2x Tesla C2050 + 2x Intel Xeon X5550, 48 GB memory, $11K, 1.0 kw

4.5x Lower Power & Cooling Costs

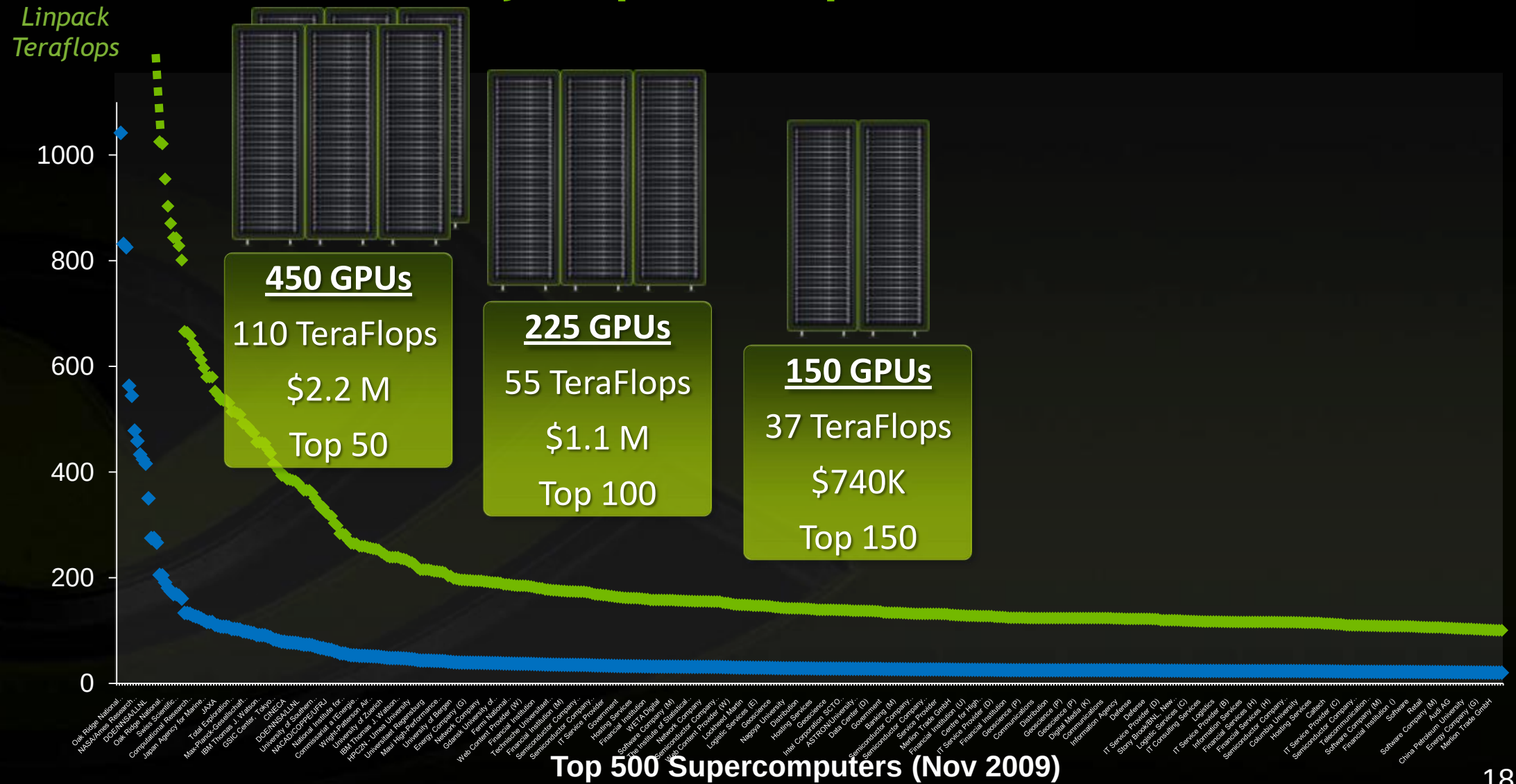
37 TeraFlop System : Top 150 System



2 Racks of GPU+CPUs	**7x Less Space Required**	15 Racks of CPUs
$740 K	**5x Lower Cost**	$3.8 M
$117 K	**4.5x Power Savings every Year**	$524 K

As per November 2009 Top 500 List

What if Every Supercomputer Had Fermi?



20+ Oil & Gas Companies Porting to CUDA

Successful Customers

Oil & Gas ISVs

GPU vs CPU Improvements	HESS	BR PETROBRAS
Performance / Watt	18x - 27x	12x - 17x
Performance / Space	20x - 31x	15x - 20x
Performance / Cost	15x - 20x	10x - 12x

Finance: 10+ Banks Porting to CUDA

Successful Customers

symcor

BNP PARIBAS

Bloomberg

Several unannounced

Finance ISVs

Hanweck Associates, LLC

nag

NumeriX

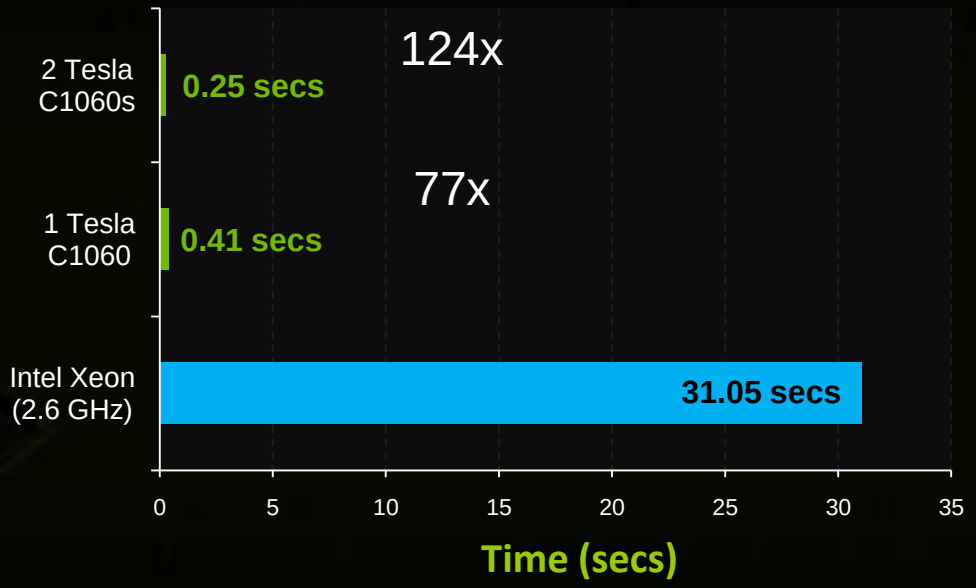
SciComp Inc.

UnRisk

Cluster Technology

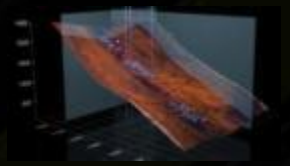
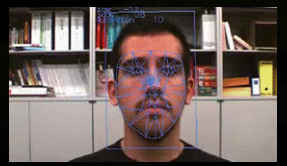
SciComp Inc.

Derivative Pricing using SciFinance
Basket Equity-Linked Structured Note



Defense / Federal Agencies

- Opportunities
  - Defense Contractors
  - Federal Agencies
  - Defense services
- Speedups 10x-50x

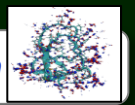
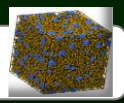





Software Available

- GIS
  - Manifold, PCI Geomatics, DigitalGlobe
- Signal Processing
  - GPU VSIPL
- MATLAB
  - GPU Plugin available
- UAV video analysis
  - MotionDSP Ikena
- Virtual Prototyping
  - RealityServer
- Surveillance, Cryptography

Tesla Bio WorkBench : Bio-Chemistry & Bio-Informatics

Applications

Amber 10 	GROMACS FAST. FLEXIBLE. FREE.	TeraChem 	HMMER Scalable Informatics University at Buffalo The State University of New York	Hex (Docking)
NAMD Scalable Molecular Dynamics 	LAMMPS	acemd GPUGRID.net 	CUDA-BLASTP	CUDA-MEME
VMD Visual Molecular Dynamics 	GROMOS	HOOMD blue 	MUMmerGPU	CUDA-EC

Community

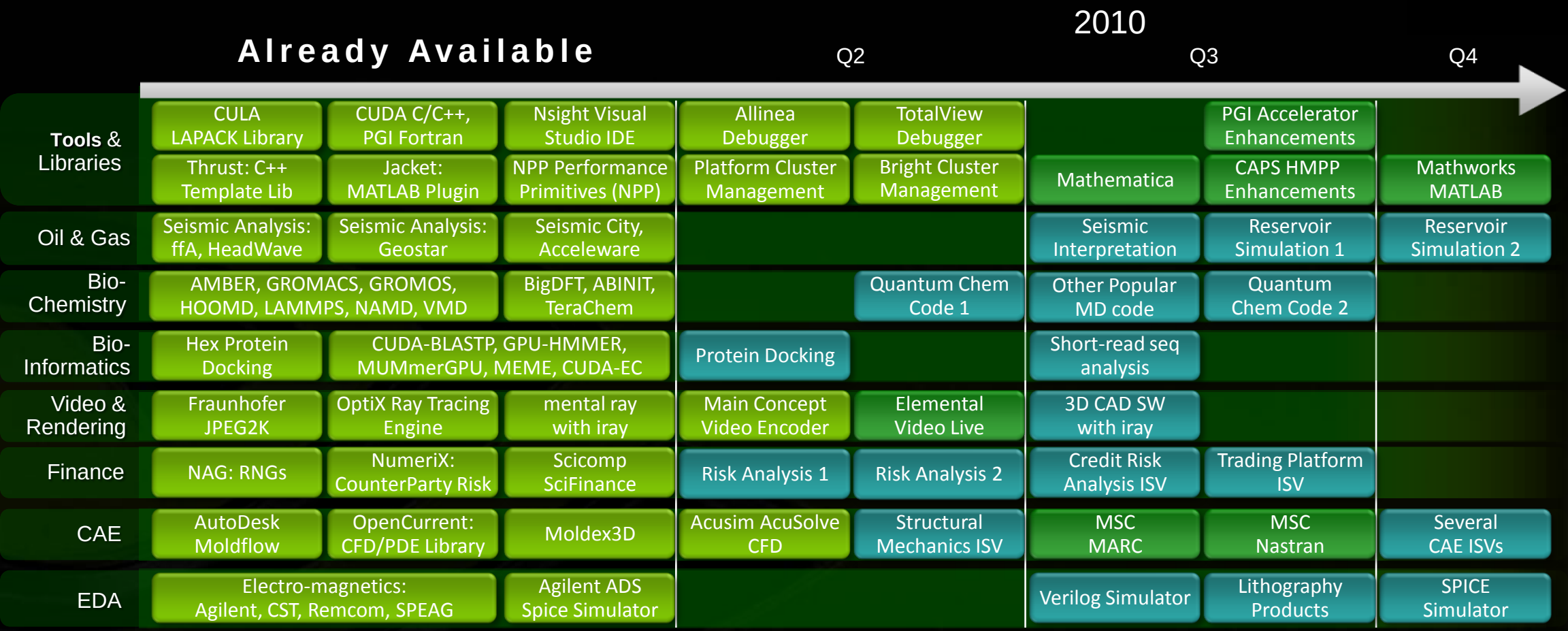
Download, Documentation | Technical papers | Discussion Forums | Benchmarks & Configurations

Platforms

Tesla Personal Supercomputer

Tesla GPU Clusters

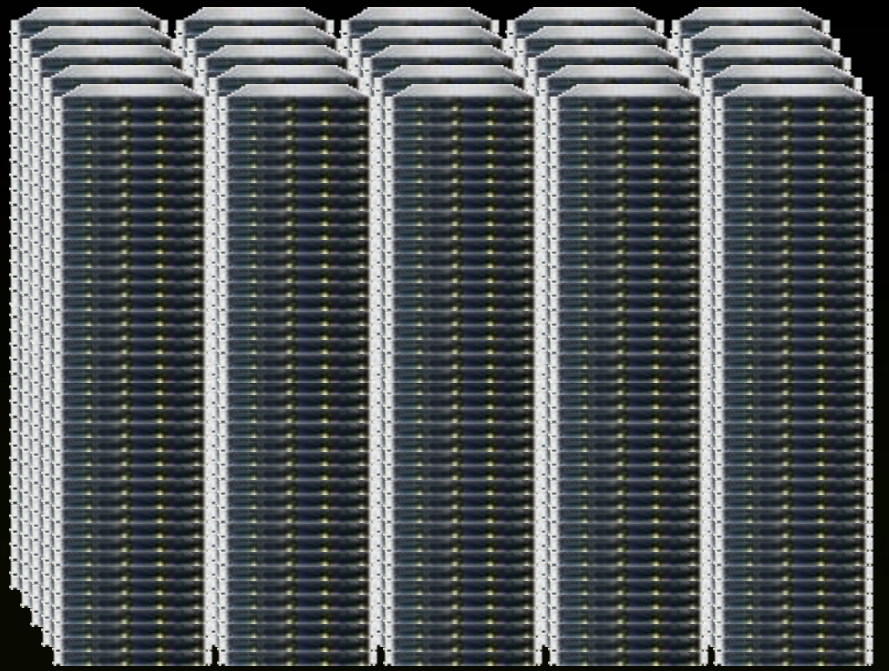
Increasing Number of CUDA Applications



Bloomberg: Bond Pricing

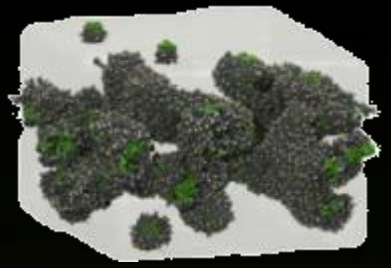
48 GPUs	**42x Lower Space**	2000 CPUs
$144K	**28x Lower Cost**	$4 Million
$31K / year	**38x Lower Power Cost**	$1.2 Million / year

Oil & Gas: Seismic Processing



1	**Equal Performance**	1
32 Tesla S1070s	**31x Less Space**	2000 CPU Servers
~$400 K	**20x Lower Cost**	~$8 M
45 kWatts	**27x Lower Power**	1200 kWatts

Finding a Better Shampoo

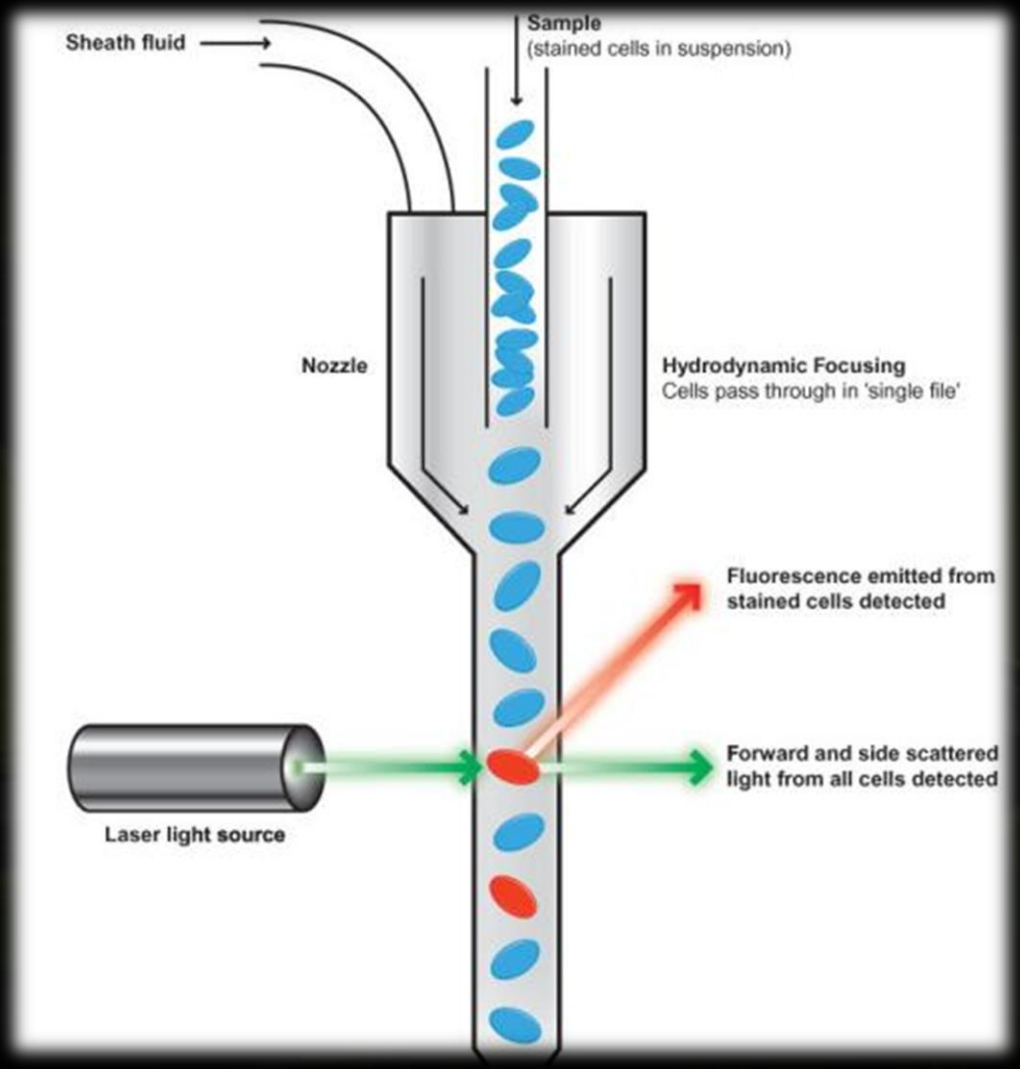


TEMPLE UNIVERSITY P&G

Tesla PSC	**Equal Performance**	32 CPU Servers
1 kWatt	**No Data Center Required**	21 kWatts
$7 K	**13x Lower System Cost**	$128 K
$2 K	**19x Lower Power & Cooling Cost**	$37 K

Reducing Radiation in CT

Flow Cytometry : Finding Cancer Cells



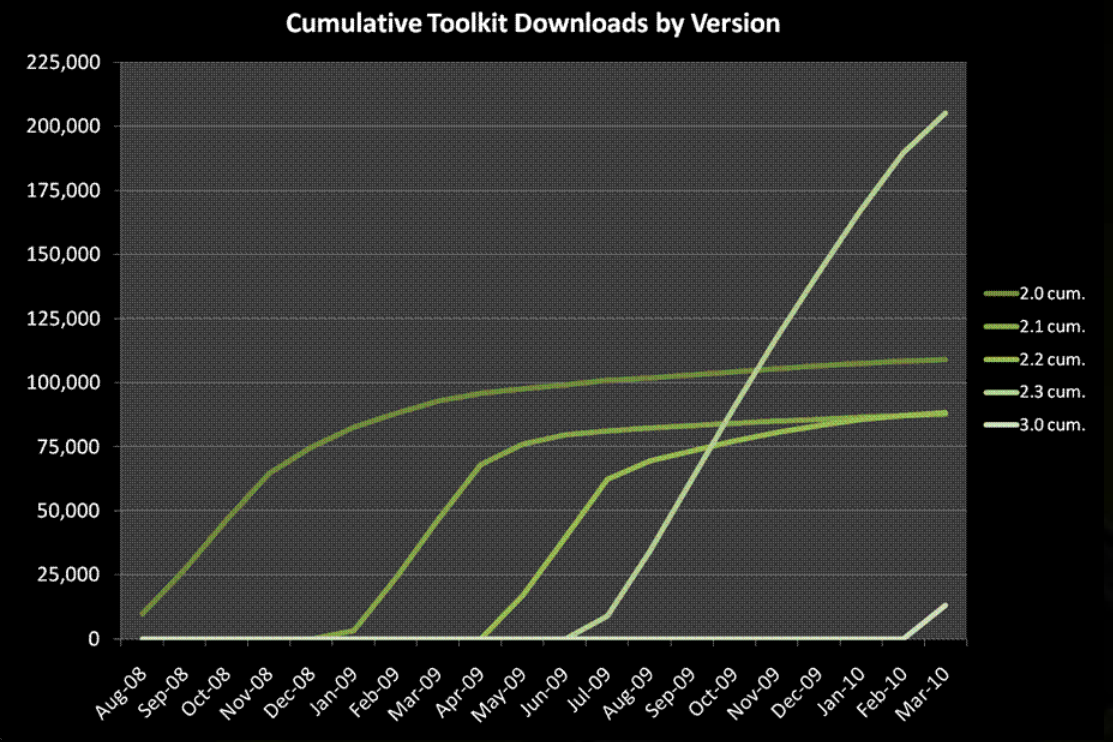
Flow Cytometer

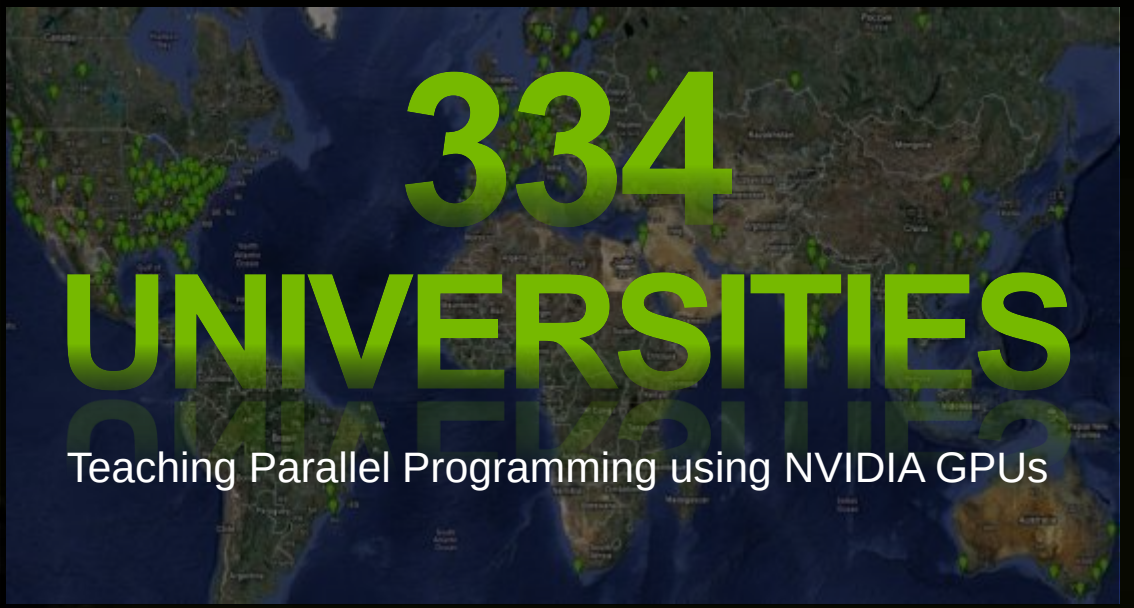
Post-Katrina Hurricane: What Google Showed

What was really there : Digital Globe

Explosive Growth in CUDA Developers

200,000 CUDA Developer Downloads





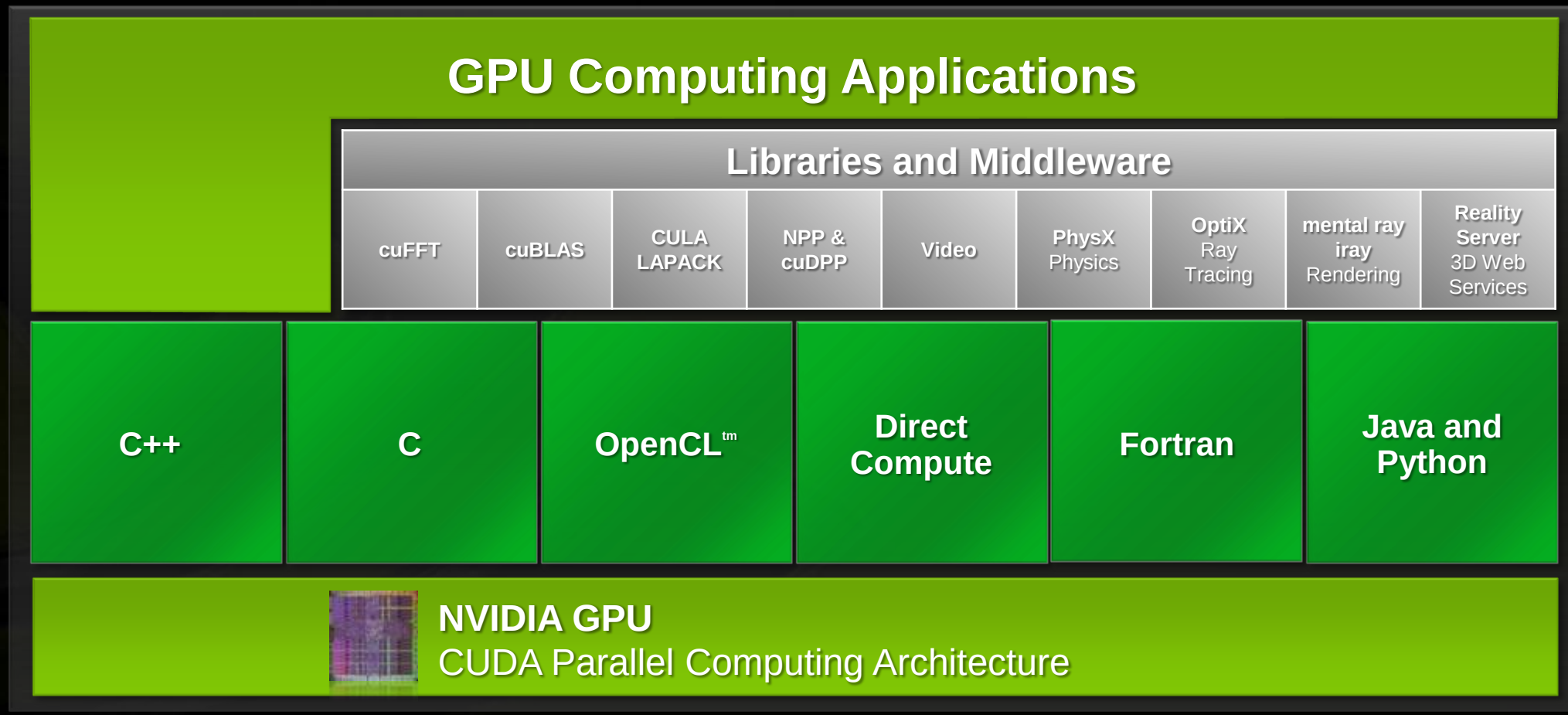
NVIDIA's Commitment to GPU Computing

1000+ Applications and Papers using NVIDIA GPUs

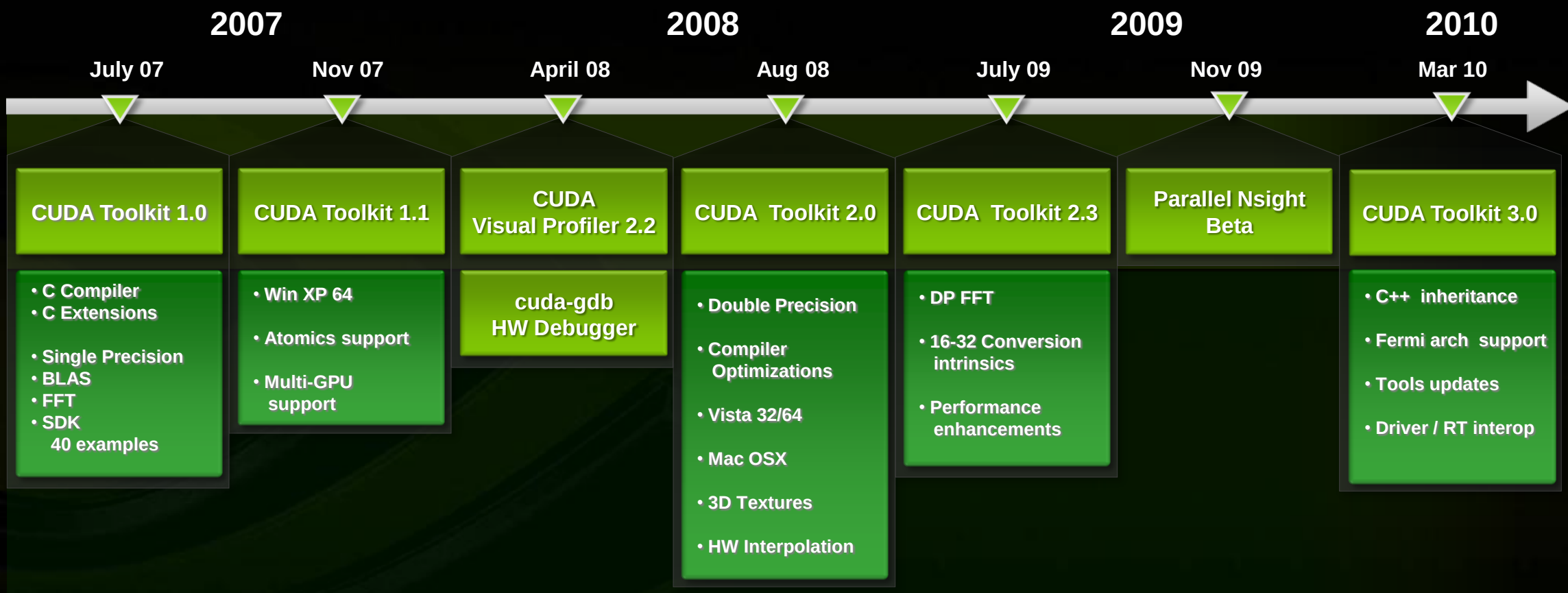
10 GPGPU Books using NVIDIA GPUs

Programming GPUs

Doing GPGPU Right: Combination of Hardware and Software



CUDA C/C++ Continuous Innovation



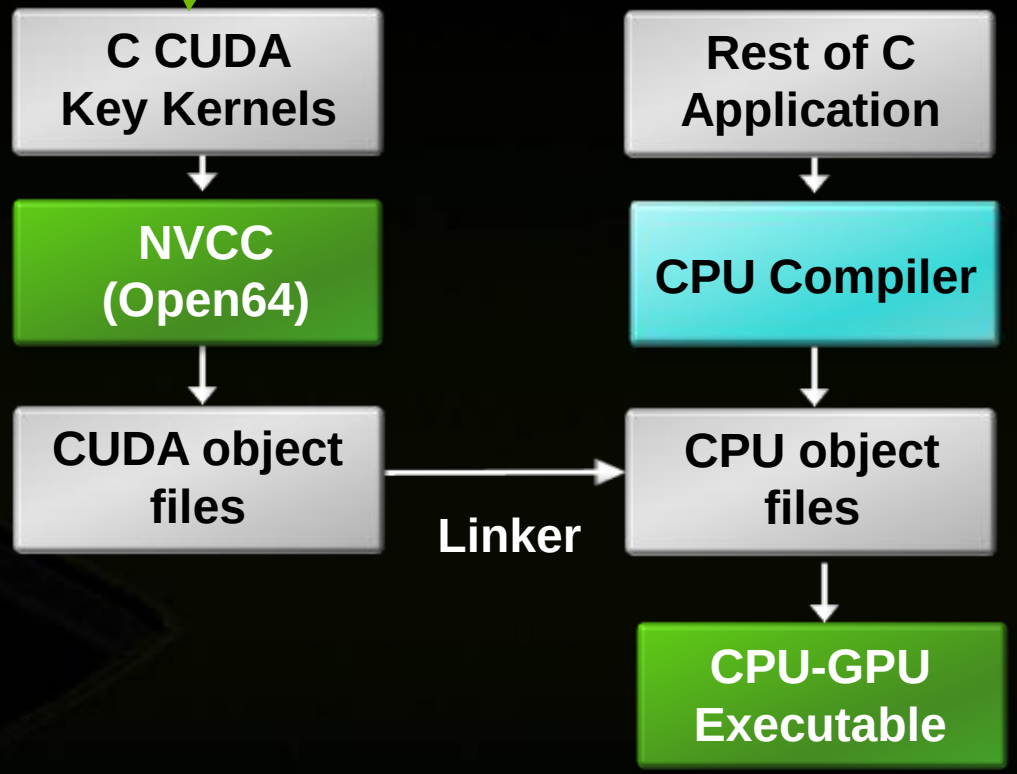
Compiling C for CUDA Applications

```
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(..);
  ...
}
```

Modify into Parallel CUDA code



C for CUDA : C with a few keywords

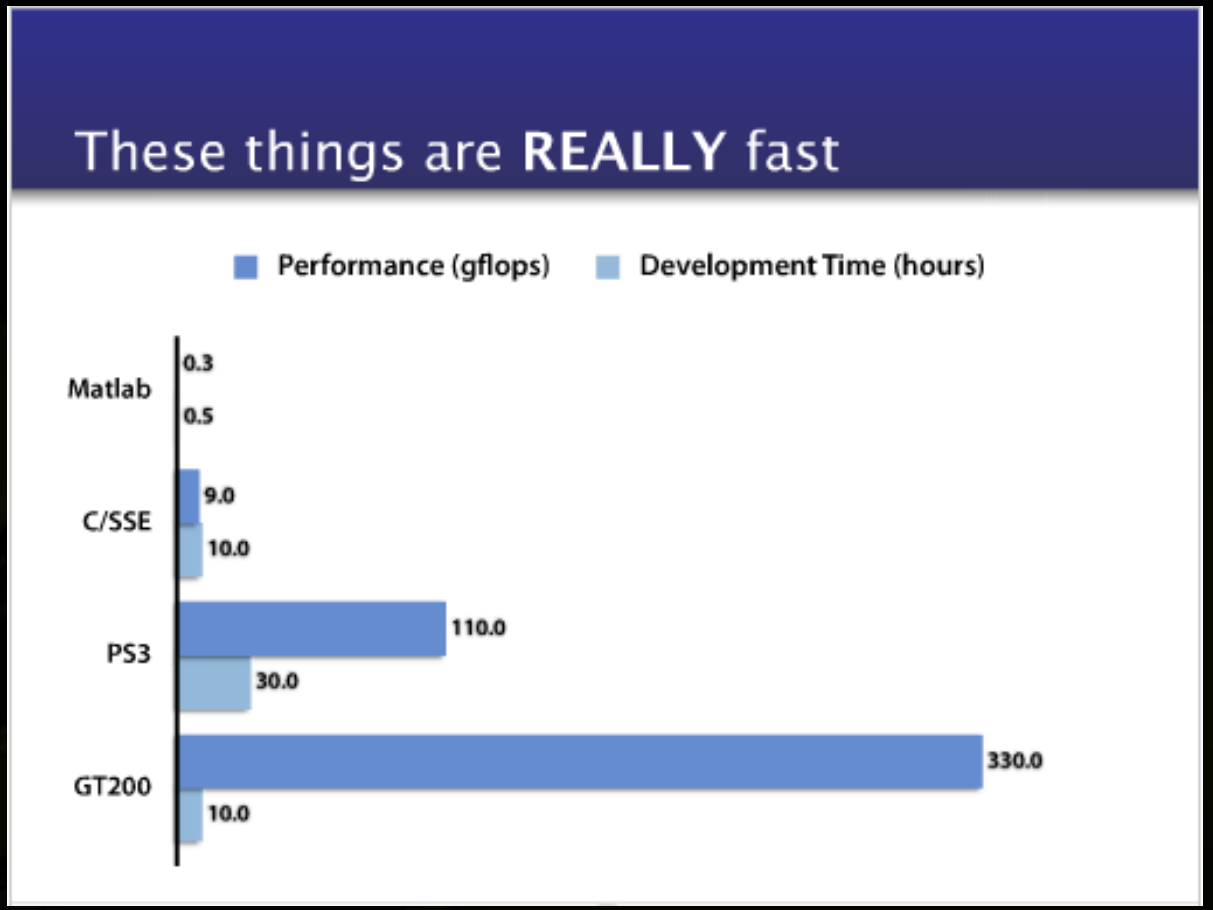
```
void saxpy_serial(int n, float a, float *x, float *y)
{
    for (int i = 0; i < n; ++i)
        y[i] = a*x[i] + y[i];
}
// Invoke serial SAXPY kernel
saxpy_serial(n, 2.0, x, y);
```

Standard C 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);
```

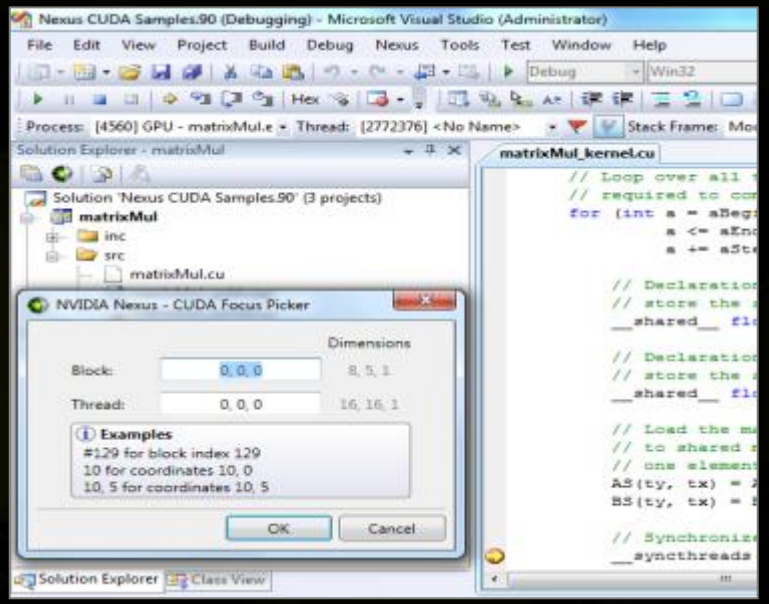
Parallel C Code

CUDA Programming Effort / Performance

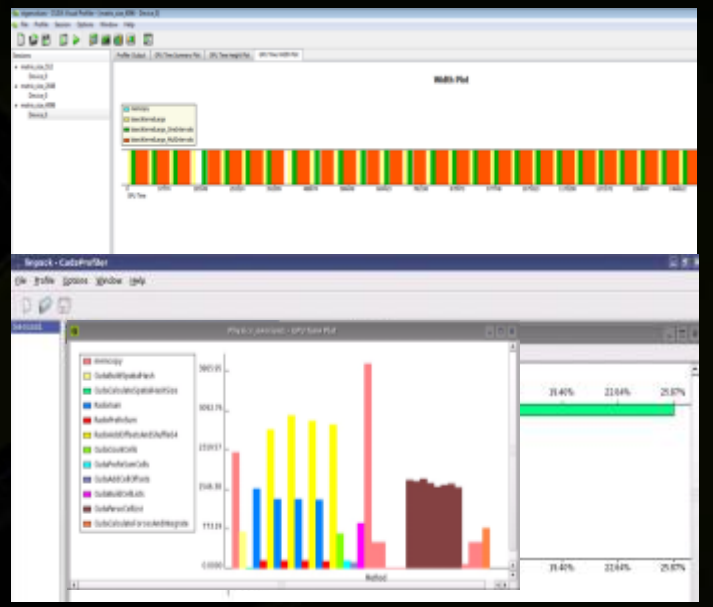


Source : MIT CUDA Course

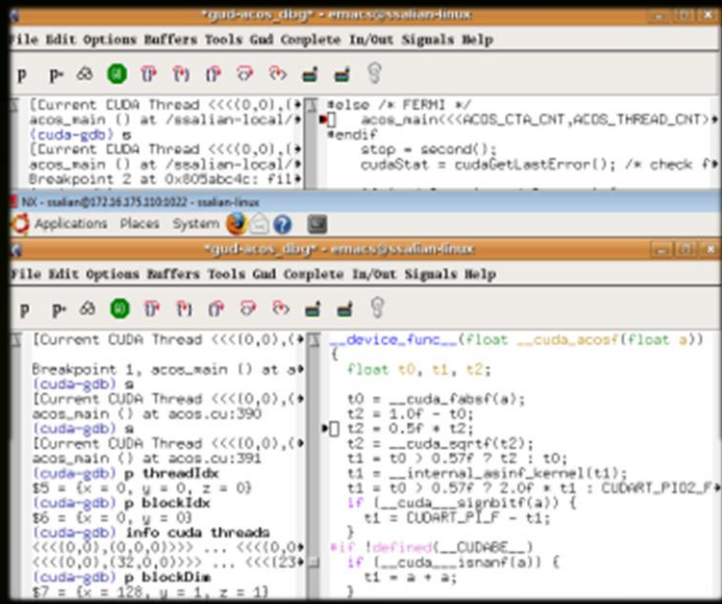
Parallel Nsight Visual Studio



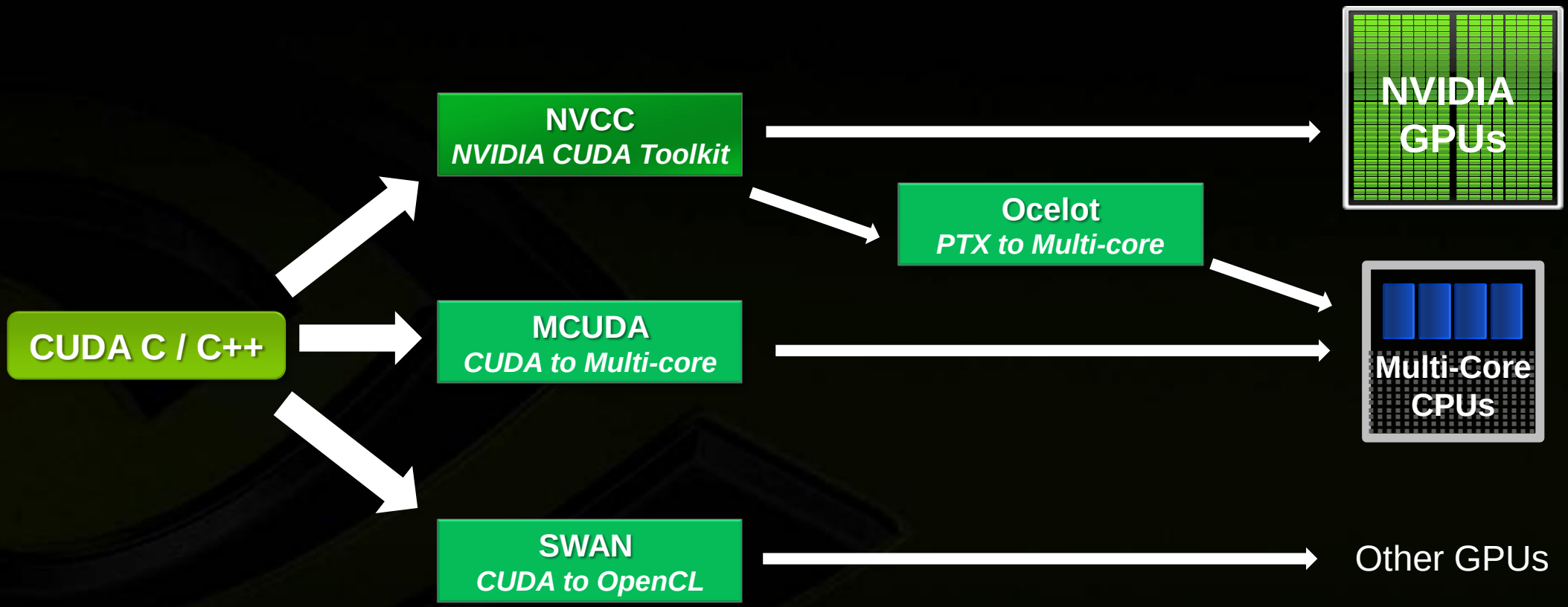
Visual Profiler For Linux



cuda-gdb For Linux



Targeting Multiple Platforms with CUDA

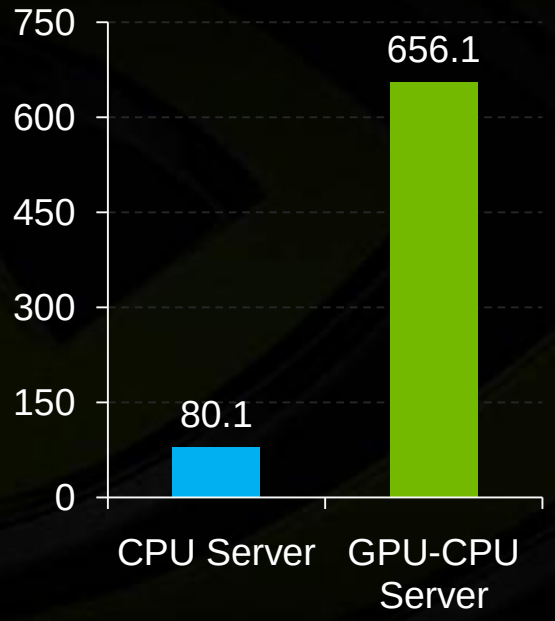


MCUDA: http://impact.crhc.illinois.edu/mcuda.php
Ocelot: http://code.google.com/p/gpuocelot/
Swan: http://www.multiscalelab.org/swan

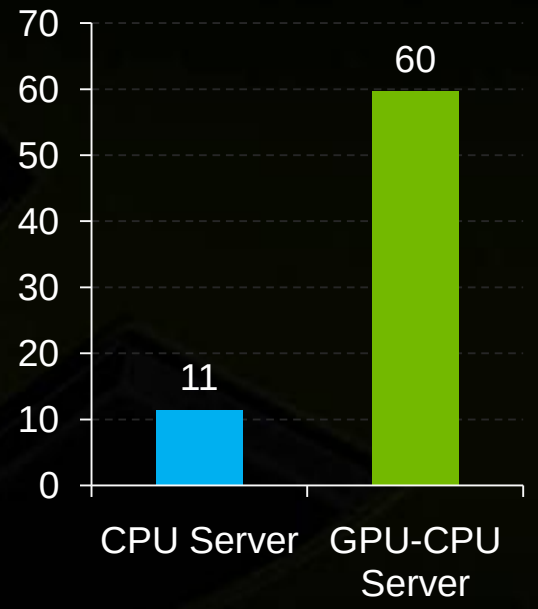
Performance Benchmarks

8x Higher Linpack

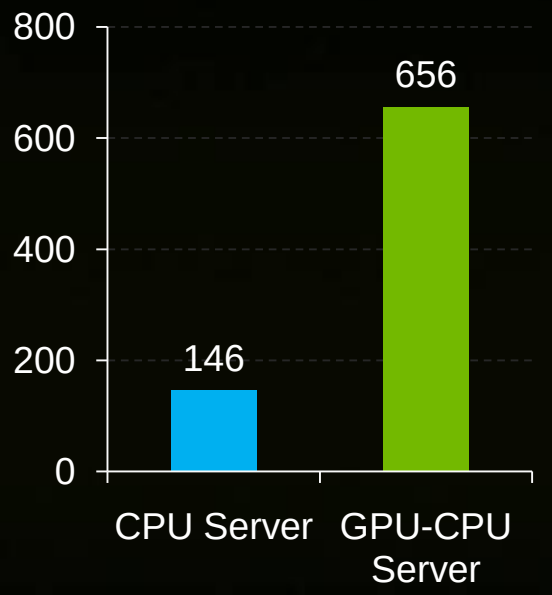
8x



5x



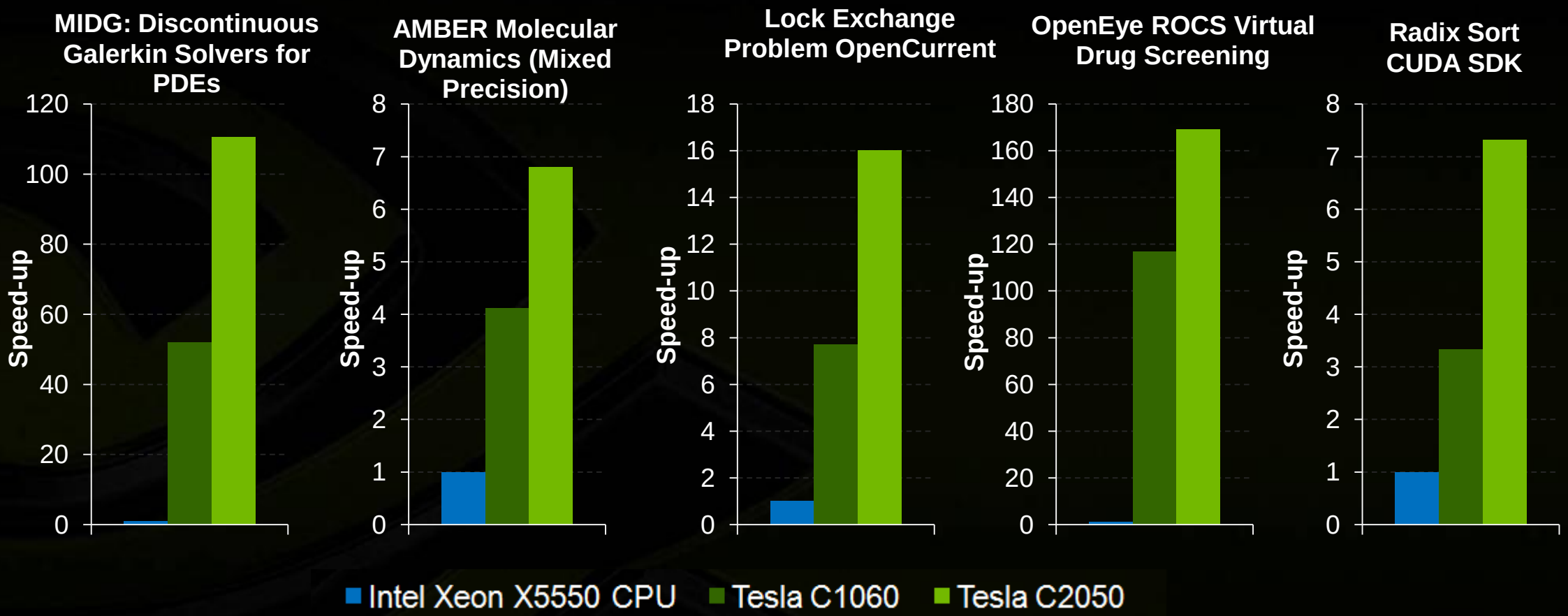
4.5x



CPU 1U Server: 2x Intel Xeon X5550 (Nehalem) 2.66 GHz, 48 GB memory, $7K, 0.55 kw

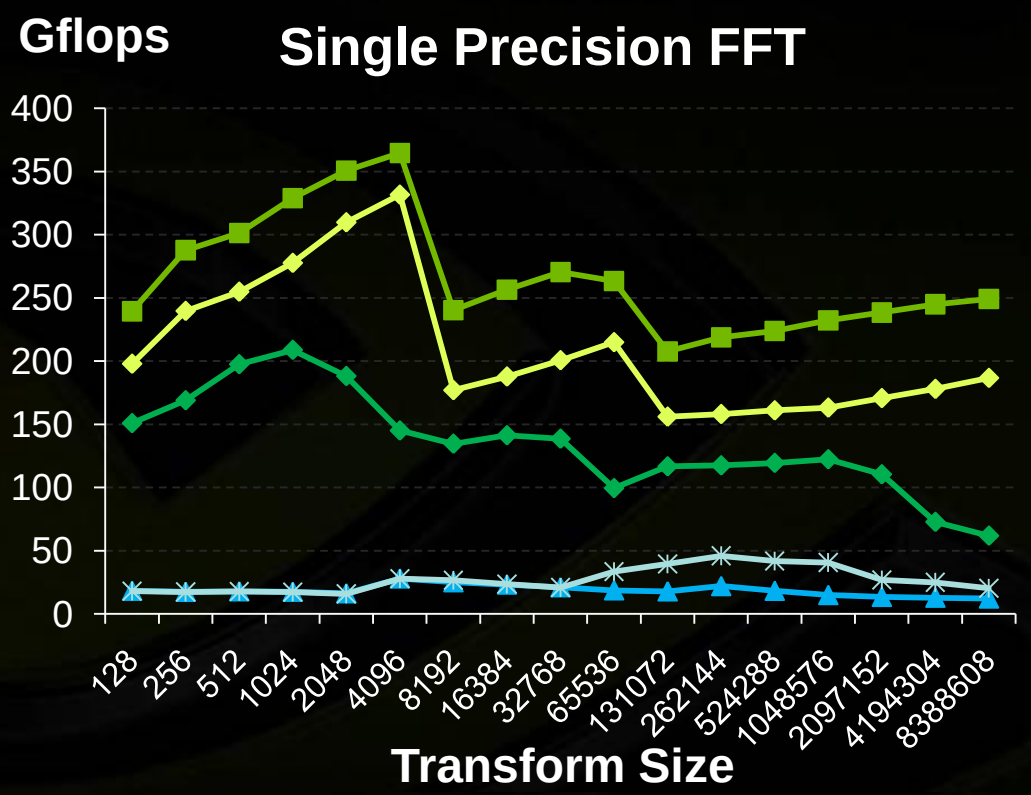
GPU-CPU 1U Server: 2x Tesla C2050 + 2x Intel Xeon X5550, 48 GB memory, $11K, 1.0 kw

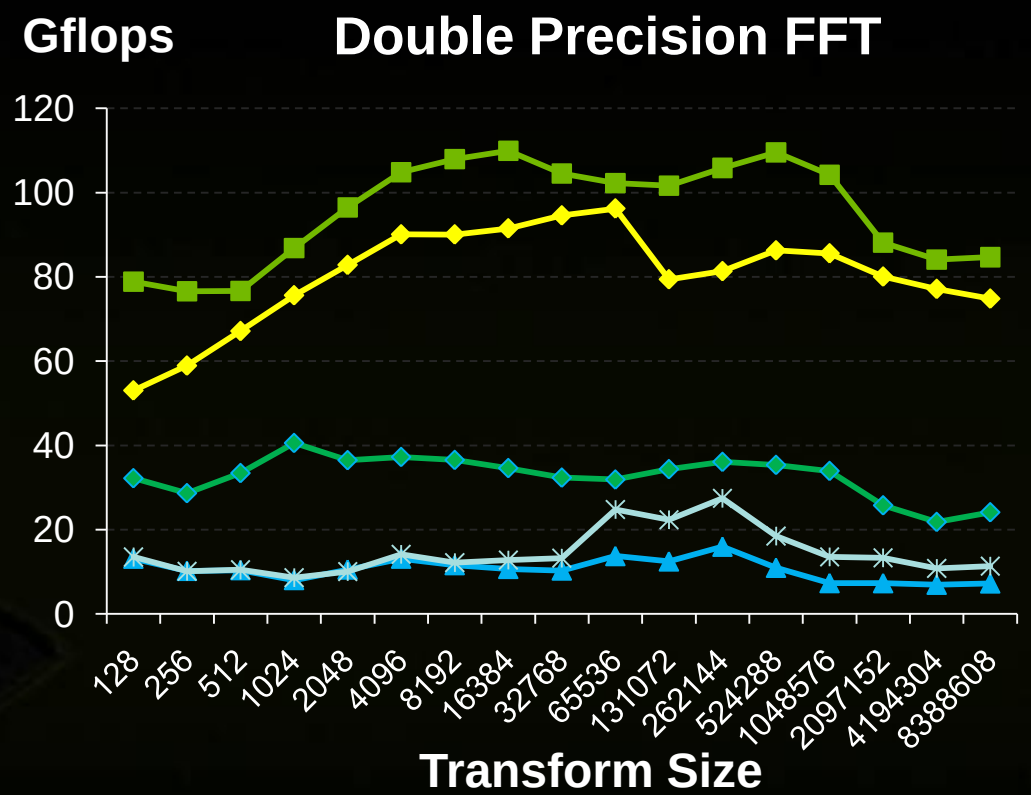
Performance Summary



Preliminary data

Standard FFT Library: cuFFT 3.1



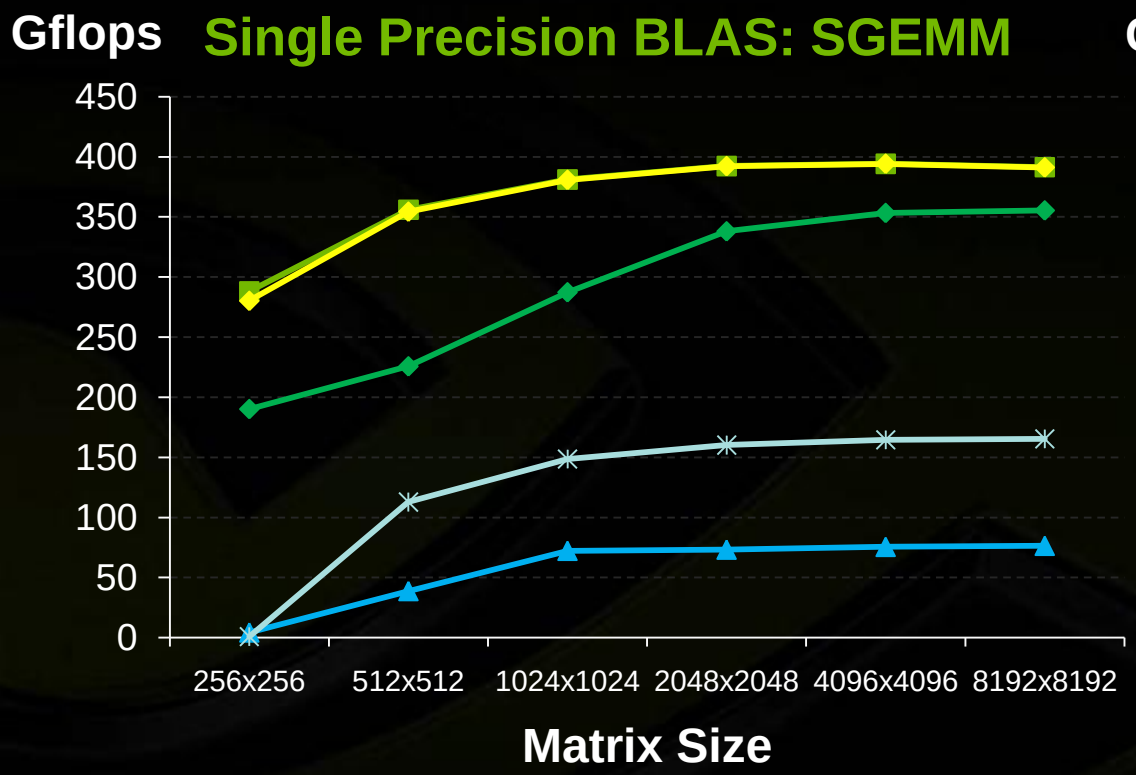


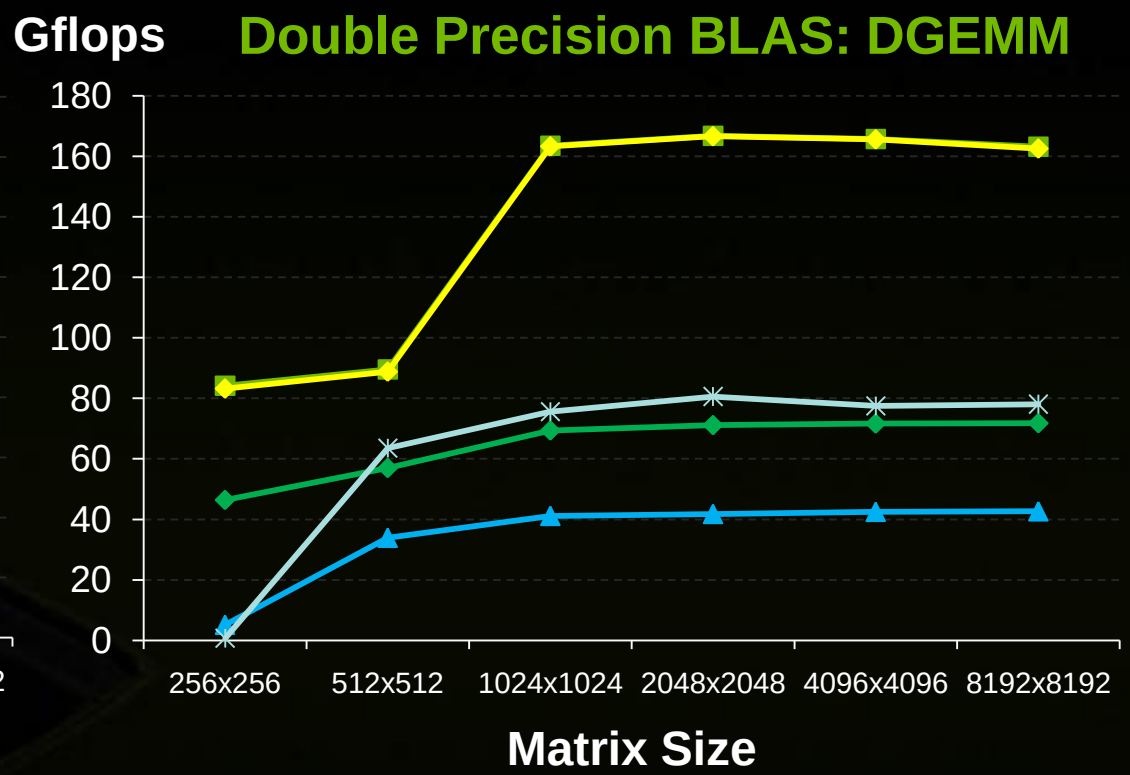
Tesla C2050 (ECC off) Tesla C2050 (ECC on) Tesla C1060 MKL 4 Threads MKL 8 Threads

cuFFT 3.1: NVIDIA Tesla C1060, Tesla C2050 (Fermi)

MKL 10.2.4.32: Quad-Core Intel Xeon 5550, 2.67 GHz

Standard BLAS Library: cuBLAS 3.1



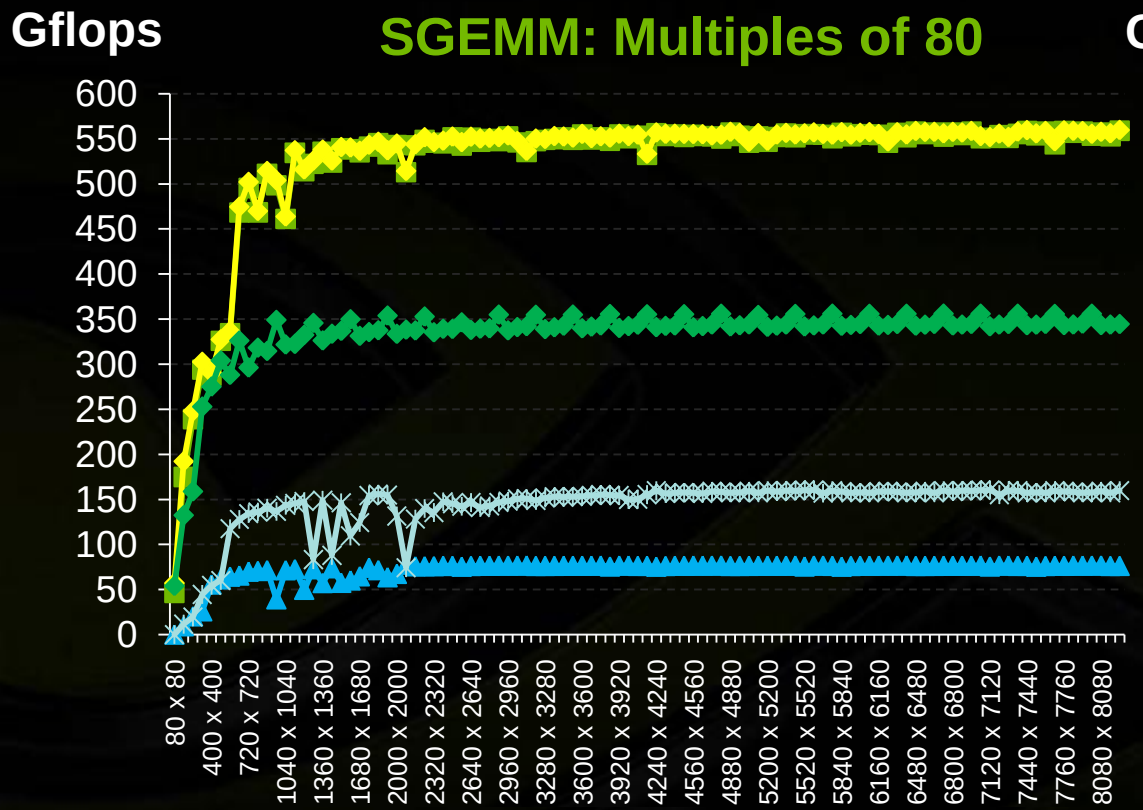


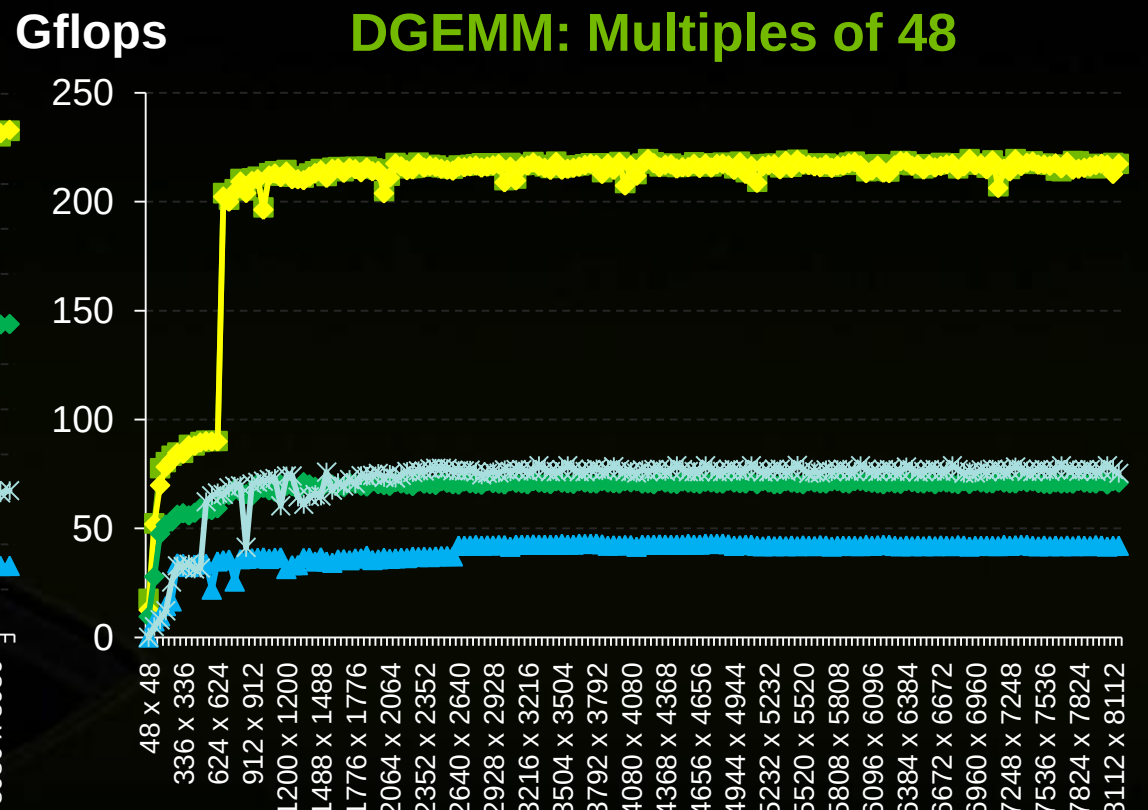
Tesla C2050 (ECC off) Tesla C2050 (ECC on) Tesla C1060 MKL 4 Threads MKL 8 Threads

cuBLAS 3.1: NVIDIA Tesla C1060, Tesla C2050 (Fermi)

MKL 10.2.4.32: Quad-Core Intel Xeon 5550, 2.67 GHz

Matrix Size for Best cuBLAS Performance



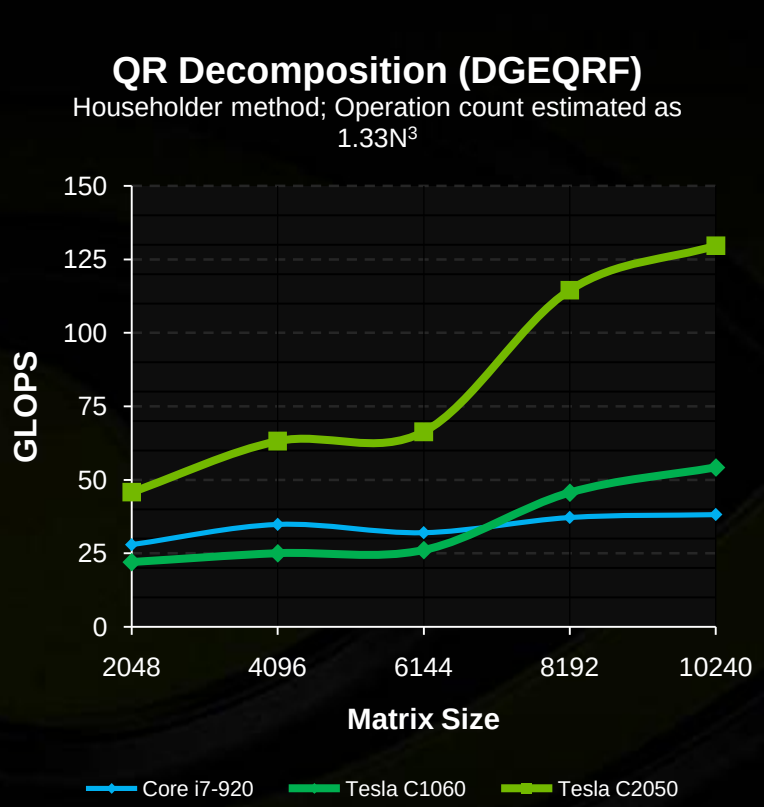


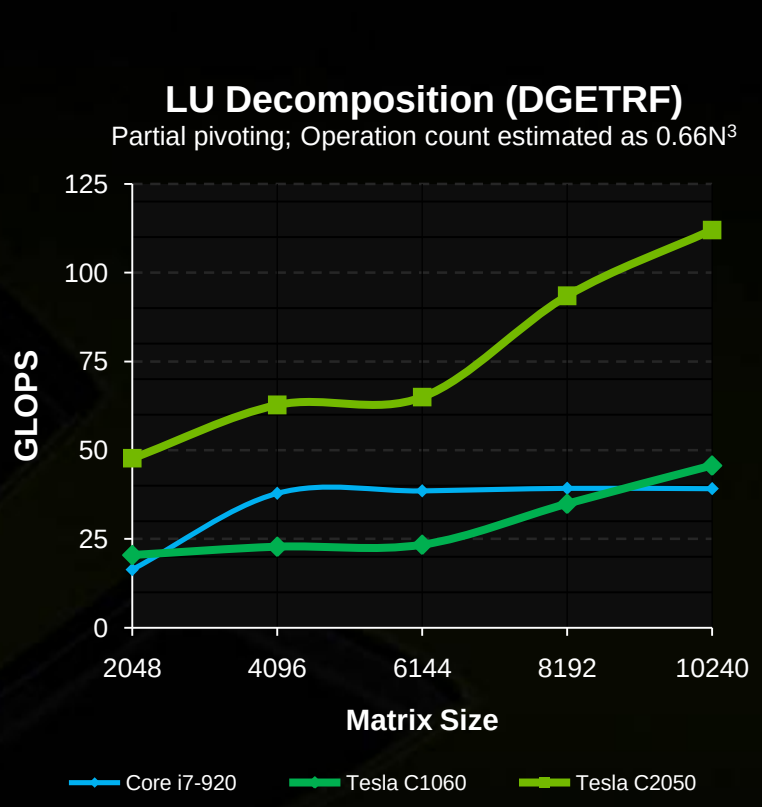
Tesla C2050 (ECC off) Tesla C2050 (ECC on) Tesla C1060 MKL 4 Threads MKL 8 Threads

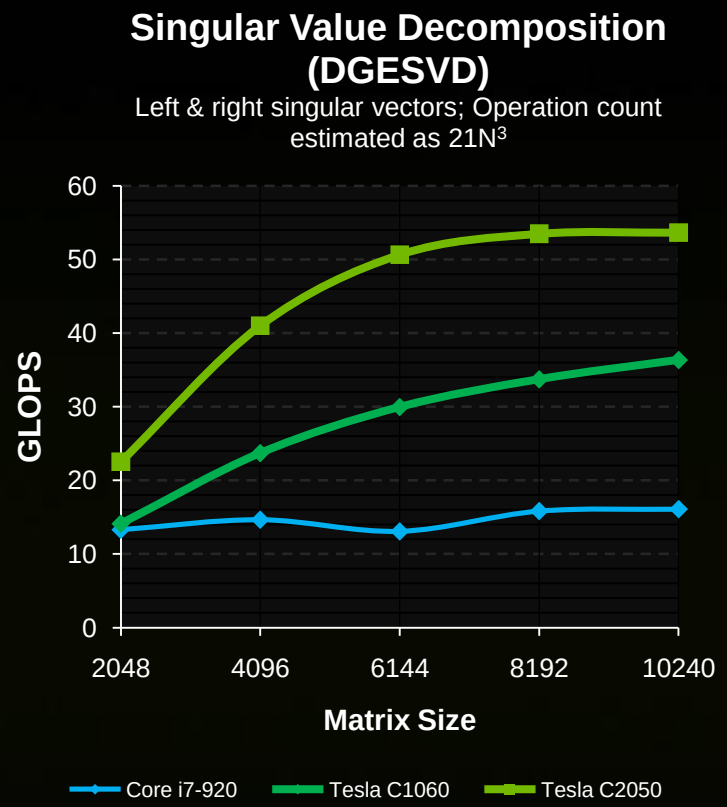
cuBLAS 3.1: NVIDIA Tesla C1060, Tesla C2050 (Fermi)

MKL 10.2.4.32: Quad-Core Intel Xeon 5550, 2.67 GHz

CULA 1.3 LAPACK Library from EM Photonics





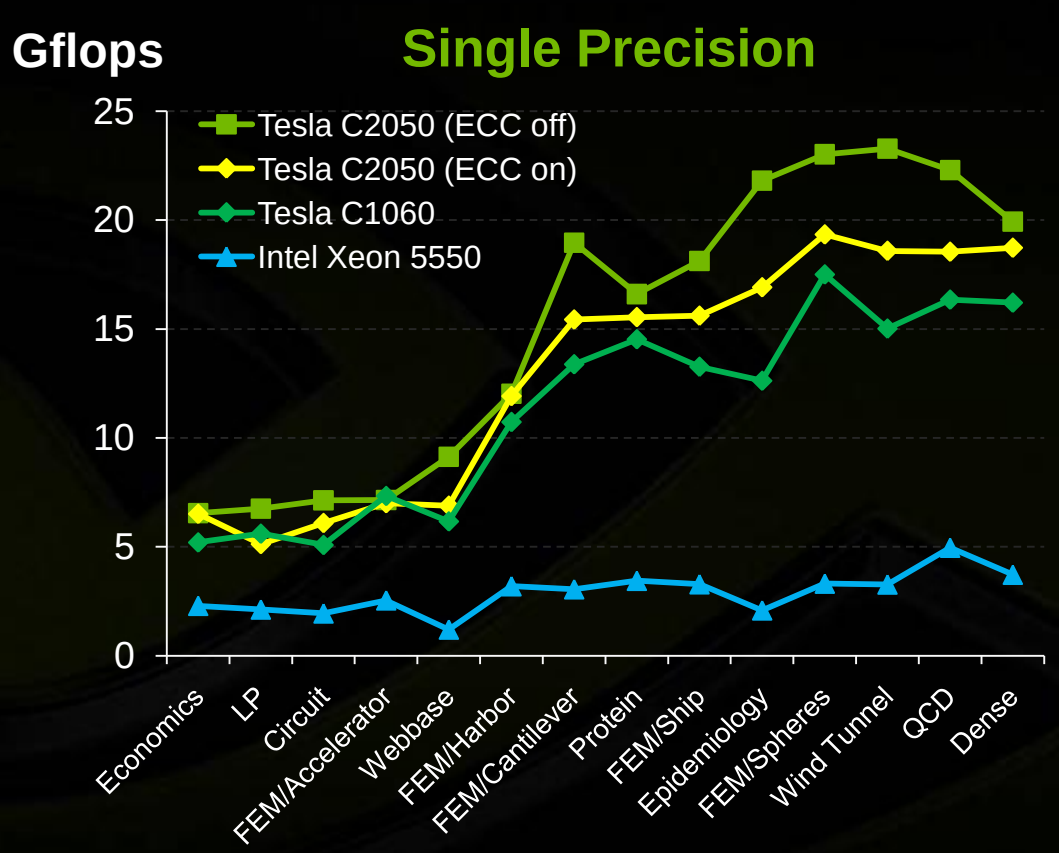


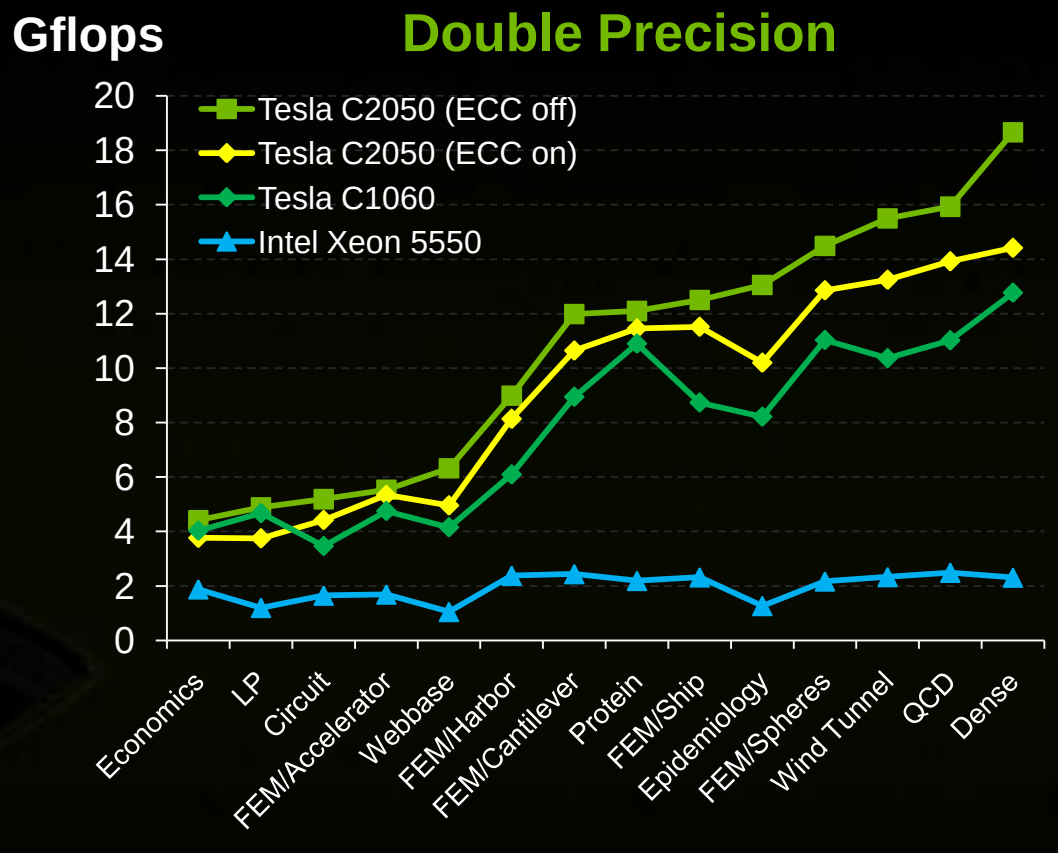
Double Precision Results

Data Courtesy: EM Photonics

EM Photonics

Sparse Matrix-Vector Multiplication (SpMV)





SpMv: CUDA 3.0, Tesla C1060 and Tesla C2050
MKL 10.2: Intel Xeon 5550, 2.67 GHz

Preliminary data

Tesla GPU Computing Products

Fermi: The Computational GPU

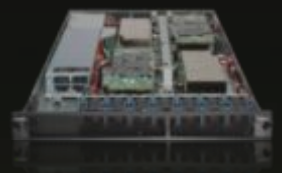
Performance

- 7x Double Precision of CPUs
- IEEE 754-2008 SP & DP Floating Point

Flexibility

- Increased Shared Memory from 16 KB to 64 KB
- Added L1 and L2 Caches
- ECC on all Internal and External Memories

NVIDIA Tesla GPU Computing Products

	Server Module		1U Systems		Workstation Boards	
	 Tesla M2070 / Tesla M2050	 Tesla M1060	 Tesla S2050	 Tesla S1070	 Tesla C2070 / Tesla C2050	 Tesla C1060
GPUs	1 T20 GPU	1 T10 GPU	4 T20 GPUs	4 T10 GPUs	1 T20 GPU	1 T10 GPU
Single Precision	1030 GFlops	933 GFlops	4120 GFlops	4140 GFlops	1030 Gflops	933 GFlops
Double Precision	515 Gflops	78 GFlops	2060 GFlops	346 GFlops	515 Gflops	78 GFlops
Memory	6 GB / 3 GB	4 GB	12 GB (S2050)	16 GB 4 GB / GPU	6 GB / 3 GB	4 GB
Mem BW	148.4 GB/s	102 GB/s	148.4 GB/s	102 GB/s	144 GB/s	102 GB/s

"In testing our key applications, the Tesla GPUs delivered speed-ups that we had never seen before, sometimes even orders of magnitude"

Satoshi Matsuoka
Professor
Tokyo Institute of Technology

"I believe history will record Fermi as a significant milestone."

Dave Patterson
Director Parallel Computing Research Laboratory, U.C. Berkeley
Co-Author of Computer Architecture: A Quantitative Approach

"Future computing architectures will be hybrid systems with parallel-core GPUs working in tandem with multi-core CPUs"

Jack Dongarra
Professor, University of Tennessee
Author of Linpack

Mass market adoption

Applications

Scaling