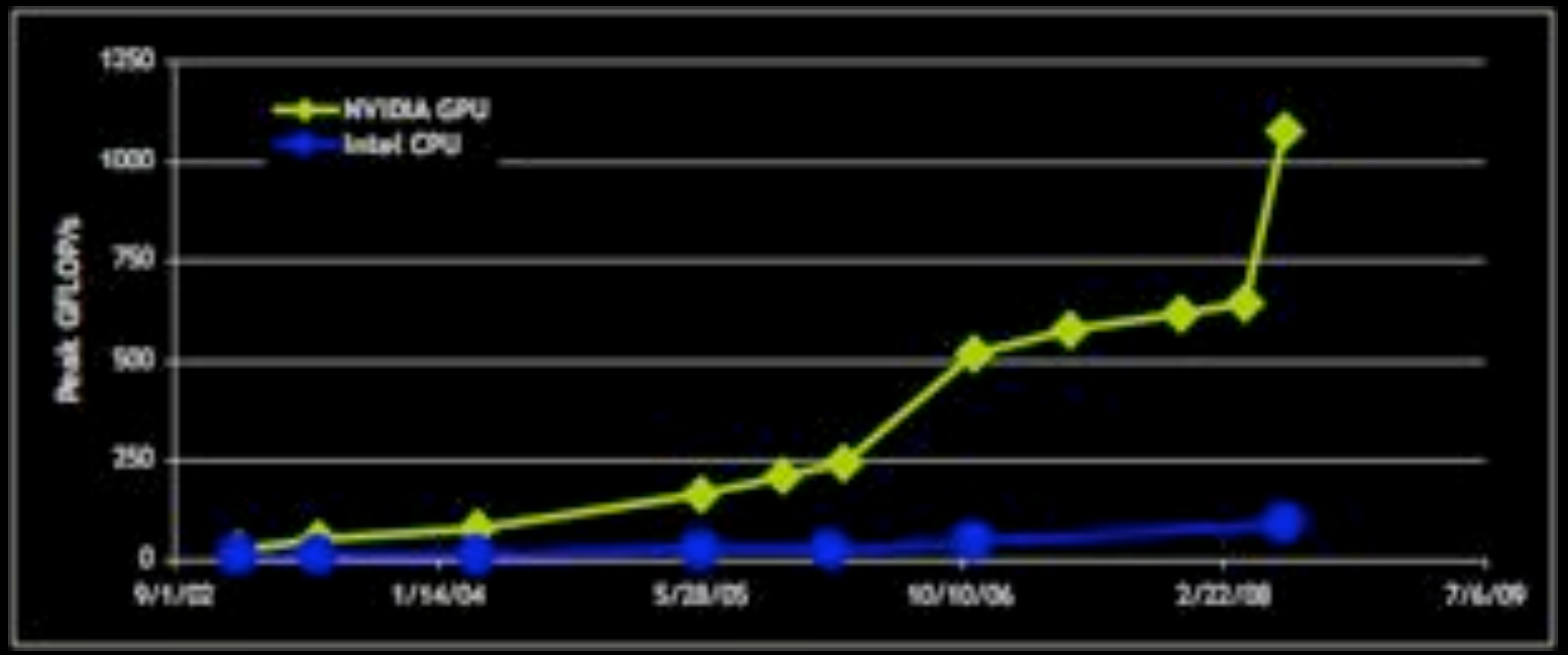
NVIDIA®

Introduction to CUDA

GPU Performance History

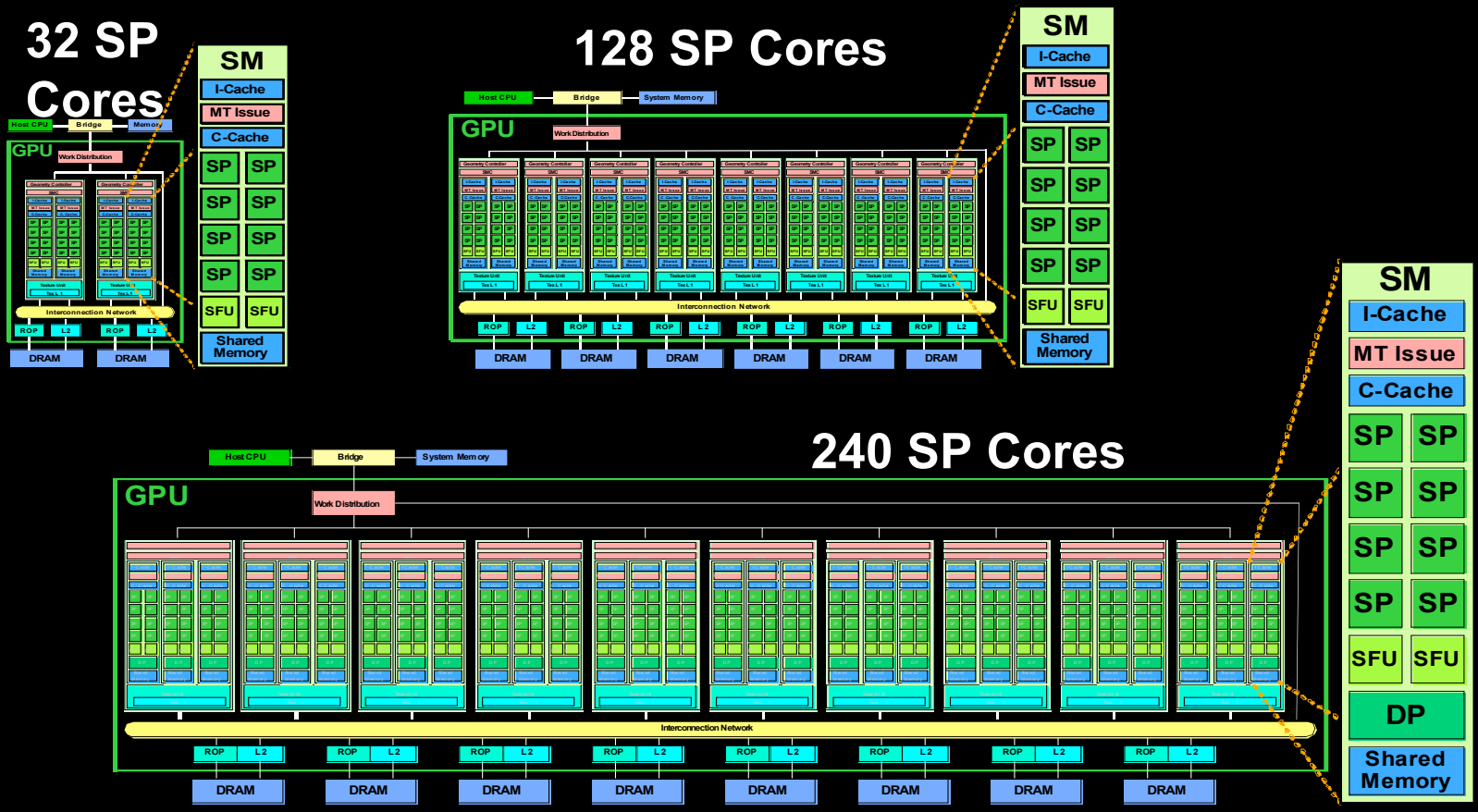
- GPUs are massively multithreaded many-core chips
  - Hundreds of cores, thousands of concurrent threads
  - Huge economies of scale
  - Still on aggressive performance growth



CUDA

- CUDA is industry-standard C
  - Write a program for one thread
  - Instantiate it on many parallel threads
  - Familiar programming model and language
- CUDA is a scalable parallel programming model
  - Program runs on any number of processors without recompiling

GPU Sizes Require CUDA Scalability



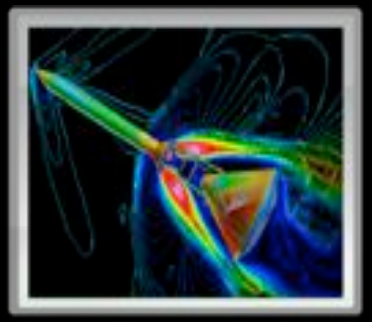
CUDA runs on NVIDIA GPUs…

Over 100 Million CUDA GPUs Deployed

GeForce®
Entertainment

Quadro®
Design & Creation

Tesla™
High-Performance Computing

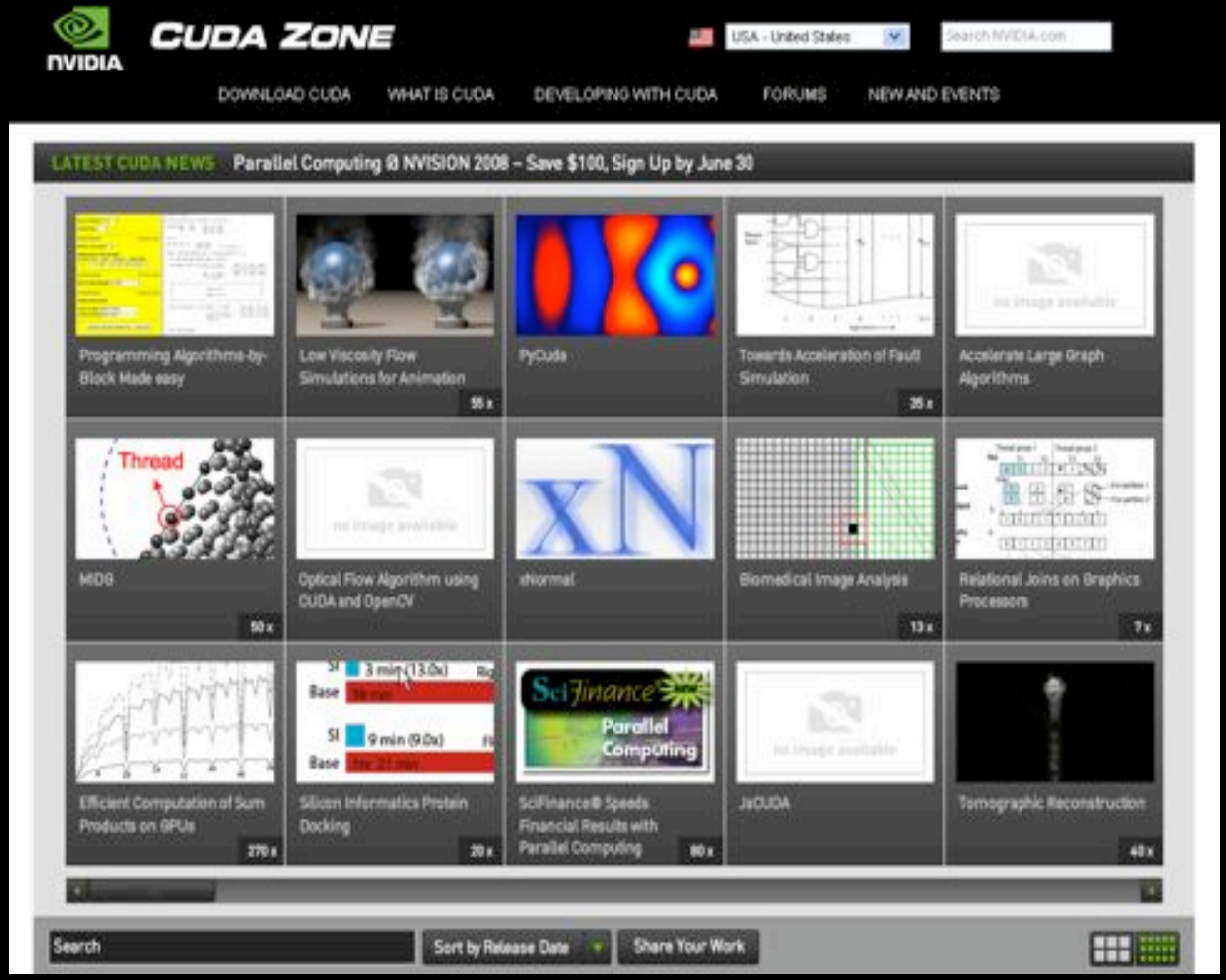


GPU

Pervasive CUDA Parallel Computing

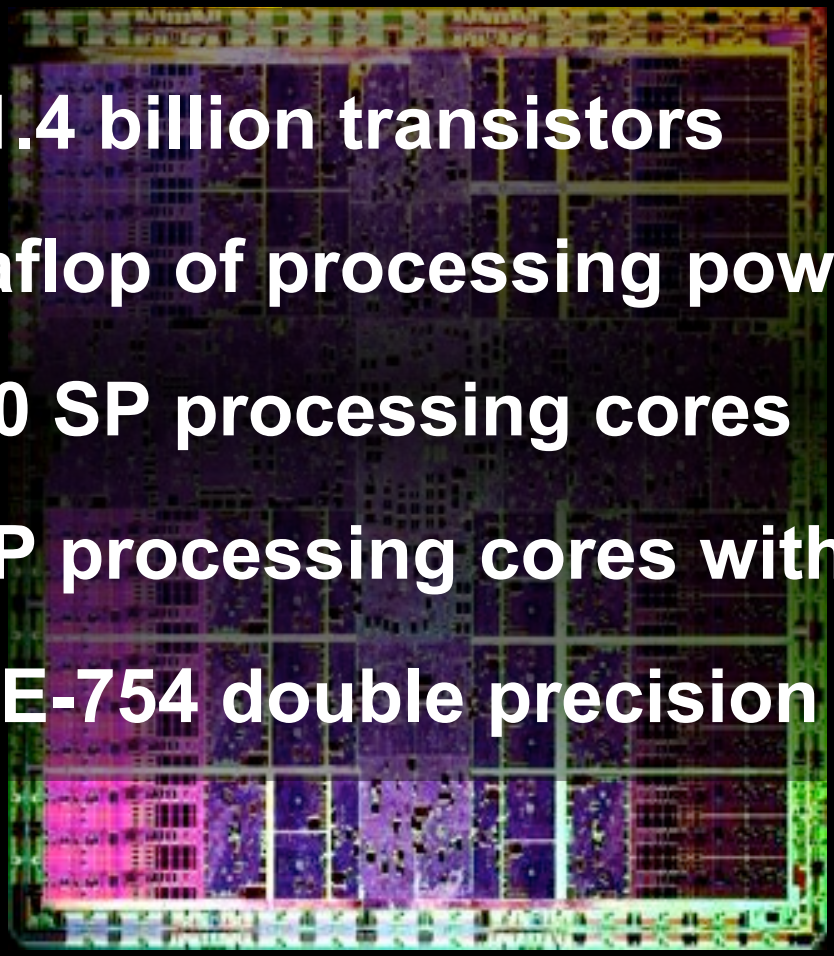
- CUDA brings data-parallel computing to the masses
  - Over 100 M CUDA-capable GPUs deployed since Nov 2006
- Wide developer acceptance
  - Download CUDA from www.nvidia.com/cuda
  - Over 50K CUDA developer downloads
  - A GPU "developer kit" costs ~$100 for several hundreds GFLOPS
- Data-parallel supercomputers are everywhere!
  - CUDA makes this power readily accessible
  - Enables rapid innovations in data-parallel computing
- Parallel computing rides the commodity technology wave

CUDA Zone: www.nvidia.com/cuda



- Resources, examples, and pointers for CUDA developers

Introducing Tesla T10P Processor



1.4 billion transistors

1 Teraflop of processing power

240 SP processing cores

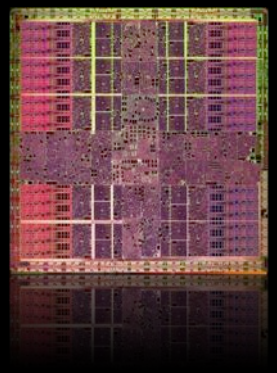
30 DP processing cores with

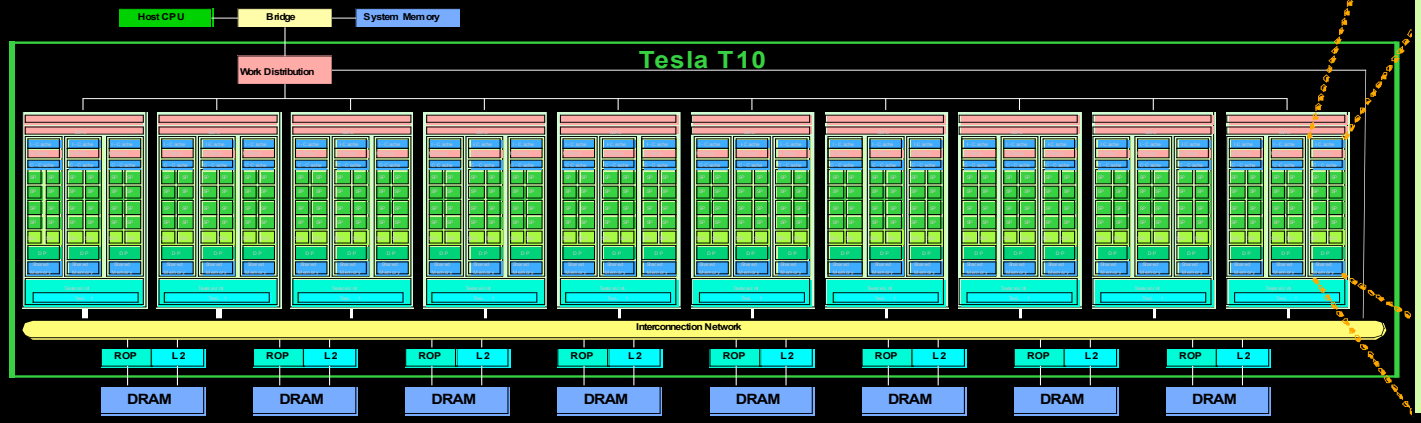
IEEE-754 double precision

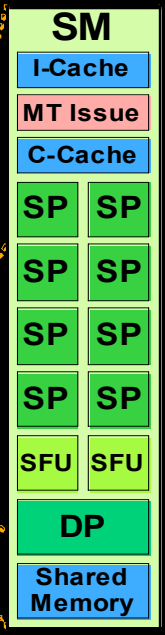
…NVIDIA's 2nd Generation CUDA Processor

CUDA Computing with Tesla T10

- 240 SP processors at 1.45 GHz: 1 TFLOPS peak
- 30 DP processors at 1.44Ghz: 86 GFLOPS peak
- 128 threads per processor: 30,720 threads total







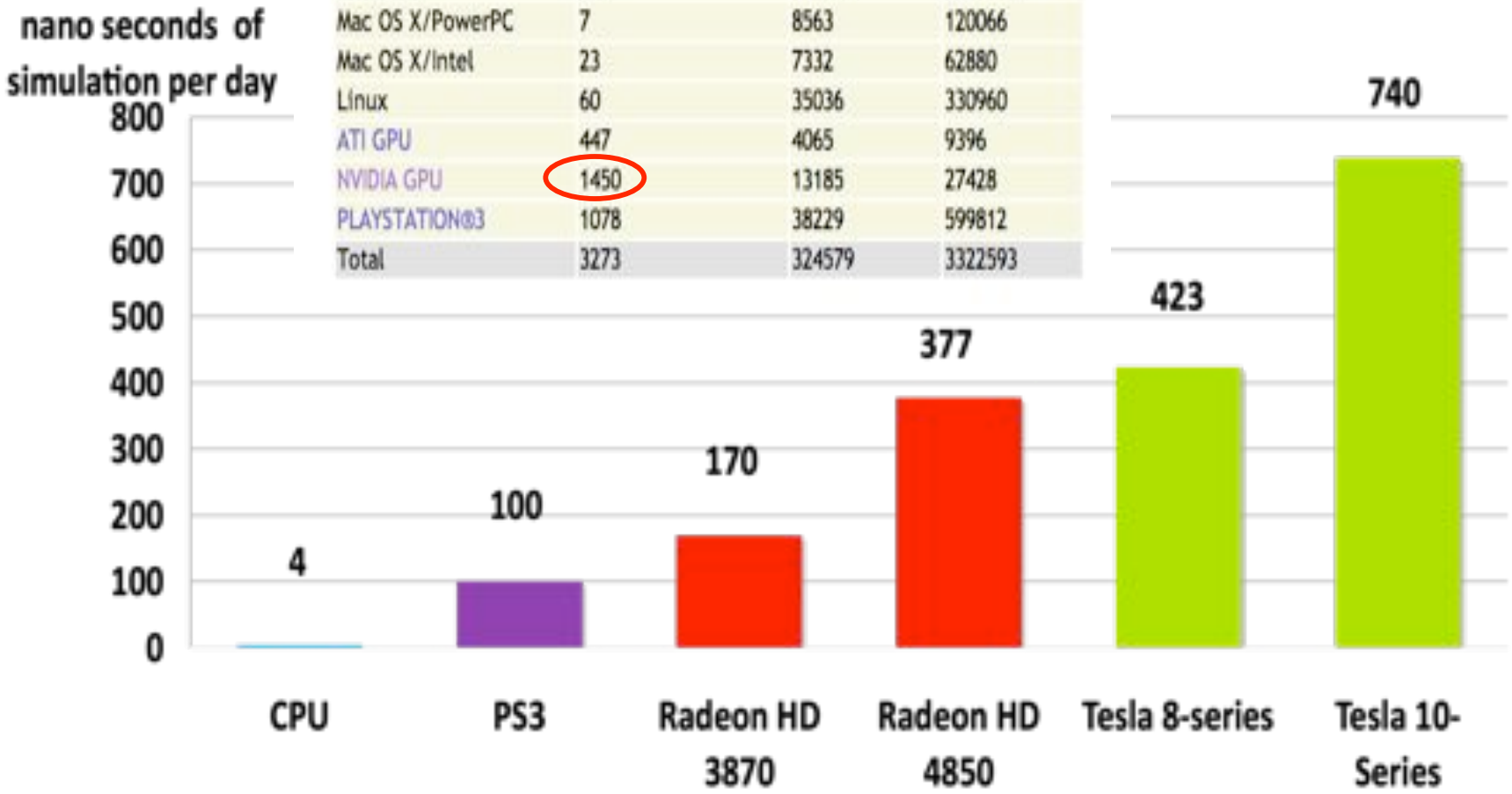
Double Precision Floating Point

	NVIDIA GPU	SSE2	Cell SPE
Precision	IEEE 754	IEEE 754	IEEE 754
Rounding modes for FADD and FMUL	All 4 IEEE, round to nearest, zero, inf, -inf	All 4 IEEE, round to nearest, zero, inf, -inf	Round to zero/truncate only
Denormal handling	Full speed	Supported, costs 1000's of cycles	Flush to zero
NaN support	Yes	Yes	No
Overflow and Infinity support	Yes	Yes	No infinity, clamps to max norm
Flags	No	Yes	Some
FMA	Yes	No	Yes
Square root	Software with low-latency FMA-based convergence	Hardware	Software only
Division	Software with low-latency FMA-based convergence	Hardware	Software only
Reciprocal estimate accuracy	24 bit	12 bit	12 bit
Reciprocal sqrt estimate accuracy	23 bit	12 bit	12 bit
log2(x) and 2^x estimates accuracy	23 bit	No	No

Applications

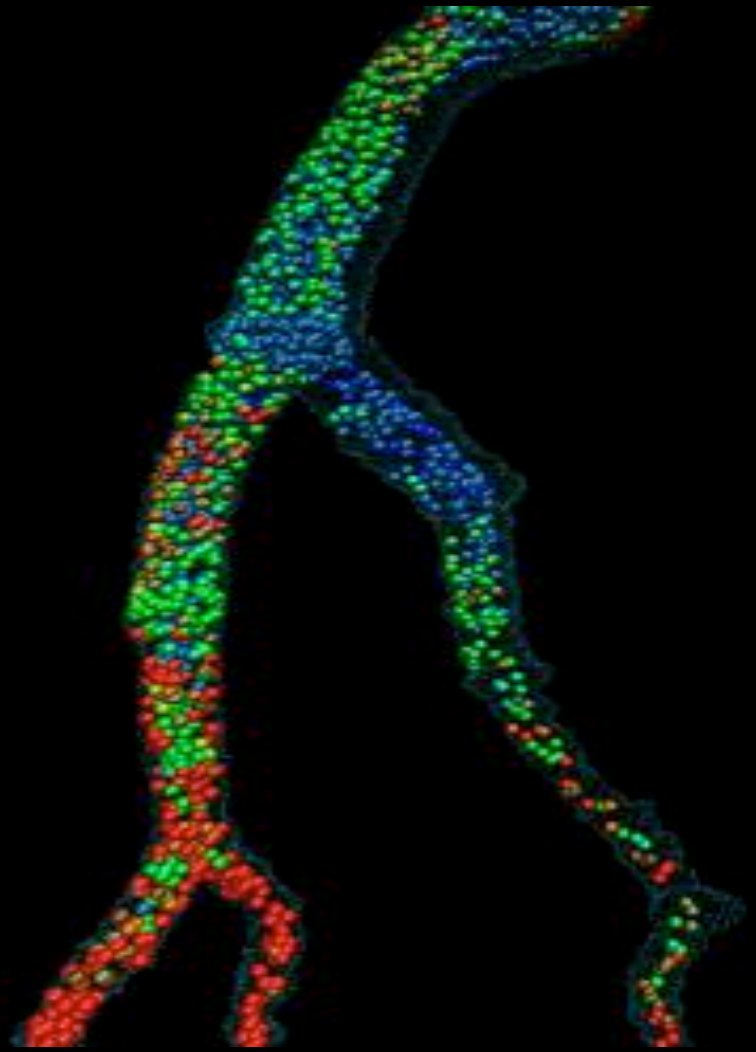
Folding@home Performance Comparison

OS Type	Current TFLOPS*	Active CPUs	Total CPUs
Windows	208	218169	2172051
Mac OS X/PowerPC	7	8563	120066
Mac OS X/Intel	23	7332	62880
Linux	60	35036	330960
ATI GPU	447	4065	9396
NVIDIA GPU	1450	13185	27428
PLAYSTATION®3	1078	38229	599812
Total	3273	324579	3322593



F@H kernel based on GROMACS code

Lattice Boltzmann



1000 iterations on a 256x128x128 domain

Cluster with 8 GPUs: 7.5 sec

Blood Gene/L 512 nodes: 21 sec

10000 iterations on irregular 1057x692x1446 domain with 4M fluid nodes

1 C870	**760 s**	**53 MLUPS**
2 C1060	**159 s**	**252 MLUPS**
8 C1060	**42 s**	**955 MLUPS**

Blood flow pattern in a human coronary artery, Bernaschi et al.

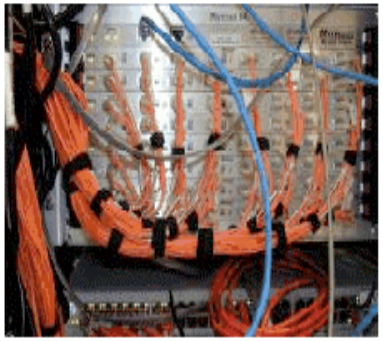
Desktop GPU Supercomputer Beats Cluster

CalcUA
256 Nodes (512 cores)

FASTRA
8 GPUs in a Desktop

CalcUA

- 256 Sun Fire V20z rekennodes (dual AMD Opteron 250, 2.4 GHz):
  - 192 rekennodes beschikken over 4 GB intern geheugen,
  - 64 rekennodes beschikken over 8 GB intern geheugen en zijn onderling verbonden via Myrinet
- 2 Sun Fire V440 servers
- 2 Sun StorEdge 3510 FC Array 6 TB centrale schijfcapaciteit



http://fastra.ua.ac.be/en/index.html

CUDA accelerated Linpack

Standard HPL code, with library that intercepts DGEMM and DTRSM calls and executes them simultaneously on the GPUs and CPU cores. Library is implemented with CUBLAS

Cluster with 8 nodes:

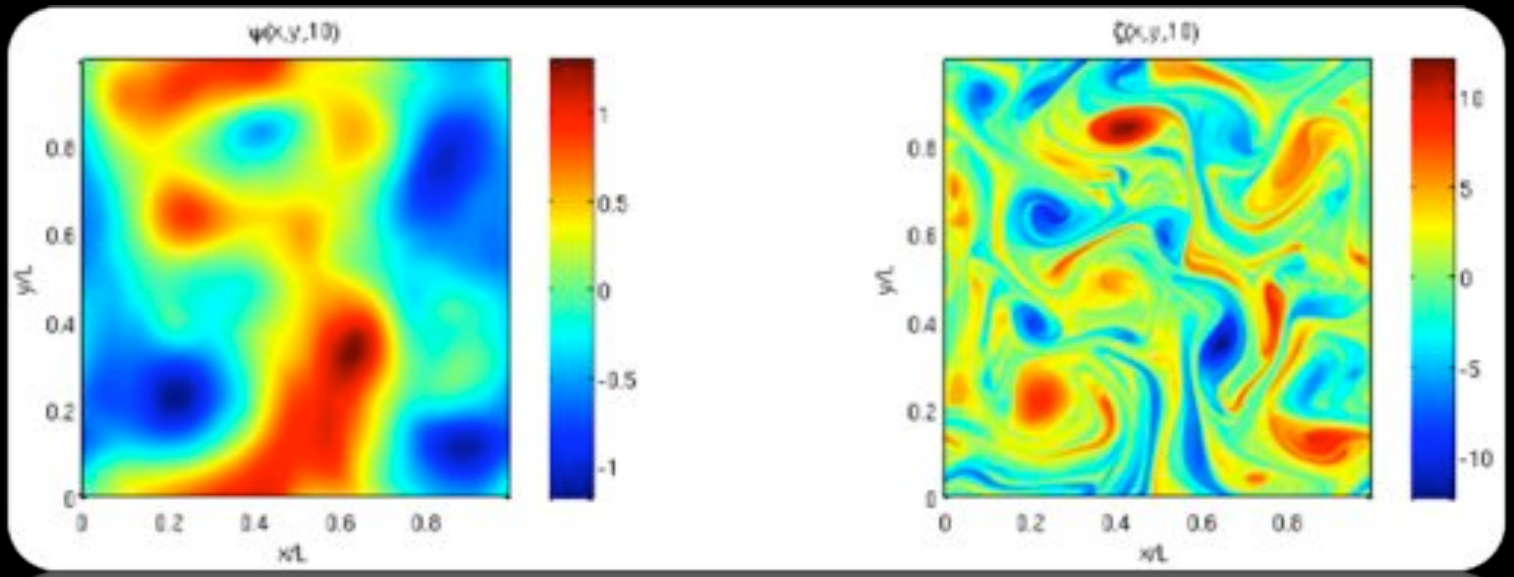
- -Each node has 2 Intel Xeon E5462 (2.8Ghz), 16GB of memory and 2 Tesla GPUs (1.44Ghz clock).
- -The nodes are connected with SDR Infiniband.

T/V	N	NB	P	Q	Time	Gflops
WR11R2L2	118144	960	4	4	874.26	1.258e+03
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=					0.0031157 ..... PASSED	

Accelerating MATLAB®

Pseudo-spectral simulation of 2D Isotropic turbulence

Use MEX files to call CUDA from MATLAB, 17x speed-up



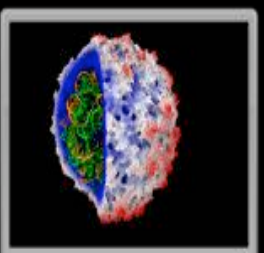
1024x1024 mesh, 400 RK4 steps, Windows XP, Core2 Duo 2.4Ghz vs GeForce 8800GTX

http://developer.nvidia.com/object/matlab_cuda.htmll

Applications in several fields

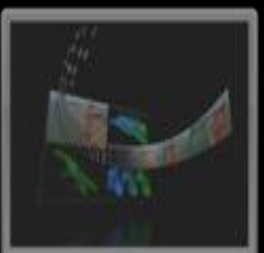
146X

Interactive visualization of volumetric white matter connectivity



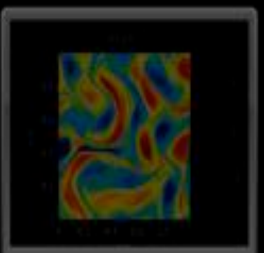
36X

Ionic placement for molecular dynamics simulation on GPU



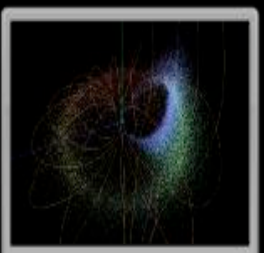
19X

Transcoding HD video stream to H.264



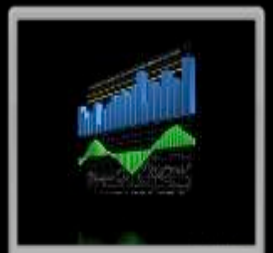
17X

Simulation in Matlab using .mex file CUDA function



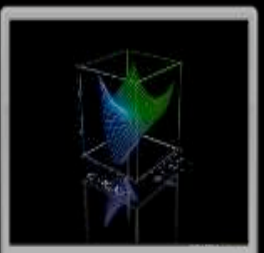
100X

Astrophysics N-body simulation



149X

Financial simulation of LIBOR model with swaptions



47X

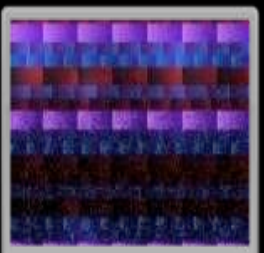
GLAME@lab: An M-script API for linear Algebra operations on GPU

20X

Ultrasound medical imaging for cancer diagnostics

24X

Highly optimized object oriented molecular dynamics



30X

Cmatch exact string matching to find similar proteins and gene sequences