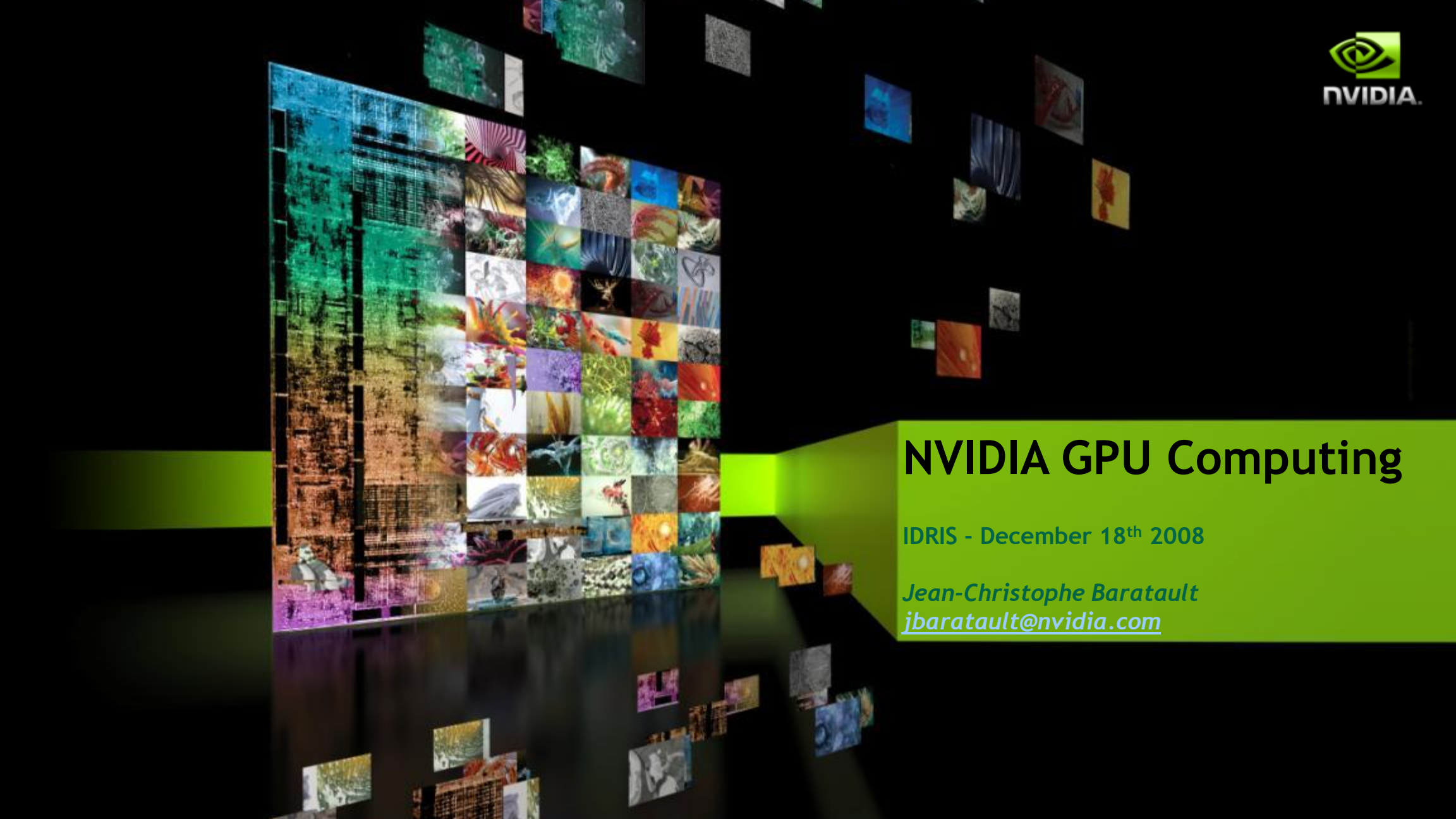


NVIDIA GPU Computing

IDRIS - December 18th 2008

Jean-Christophe Baratault
jbaratault@nvidia.com

GPU as CPU coprocessor: Not a new idea

- Hoff, et al. 1999] Hoff, K.E.I., Culver, T., Keyser, J., Lin, M. and Manocha, D. Fast Computation of Generalized Voronoi Diagrams Using Graphics Hardware.
- [Jobard, et al. 2001] Jobard, B., Erlebacher, G. and Hussaini, M.Y. Lagrangian-Eulerian Advection for Unsteady Flow Visualization.
- [Lengyel, et al. 1990] Lengyel, J., Reichert, M., Donald, B.R. and Greenberg, D.P. Real-Time Robot Motion Planning Using Rasterizing Computer Graphics Hardware.
- [Peercy, et al. 2000] Peercy, M.S., Olano, M., Airey, J. and Ungar, P.J. Interactive Multi-Pass Programmable Shading.
- [Potmesil and Hoffert 1989] Potmesil, M. and Hoffert, E.M. The Pixel Machine: A Parallel Image Computer.
- [Proudfoot, et al. 2001] Proudfoot, K., Mark, W.R., Tzvetkov, S. and Hanrahan, P. A Real-Time Procedural Shading System for Programmable Graphics Hardware.
- [Purcell, et al. 2002] Purcell, T.J., Buck, I., Mark, W.R. and Hanrahan, P. Ray Tracing on Programmable Graphics Hardware.
- [Rhoades, et al. 1992] Rhoades, J., Turk, G., Bell, A., State, A., Neumann, U. and Varshney, A. Real-Time Procedural Textures.
- [Trendall and Steward 2000] Trendall, C. and Steward, A.J. General Calculations using Graphics Hardware, with Applications to Interactive Caustics.

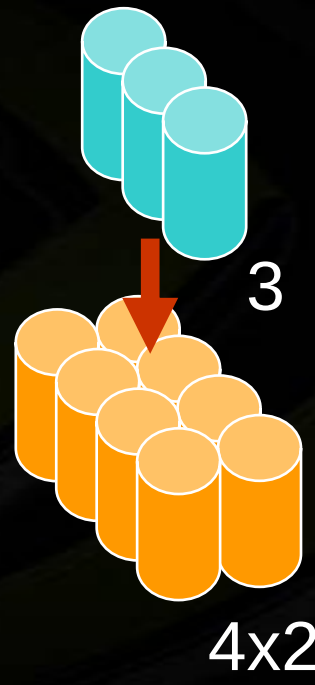
Source: www.gpgpu.org

GPU: Parallelized Architecture for years

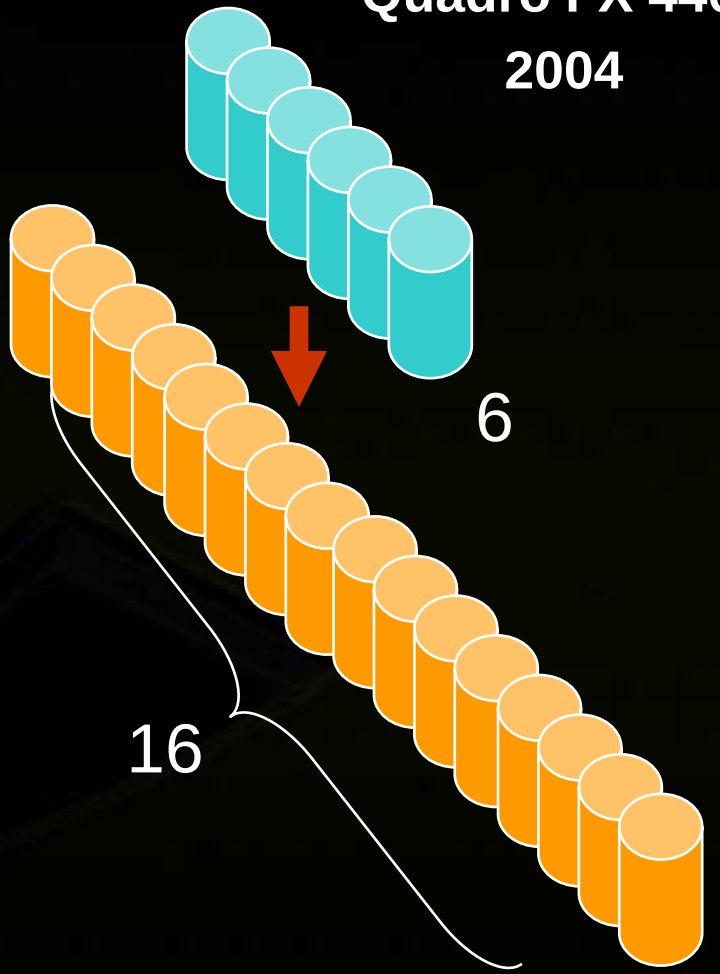
Quadro FX 3000
2003

Vertex

Pixel



Quadro FX 4400
2004



Quadro FX 4500
2005

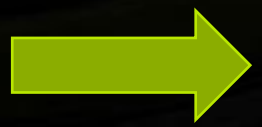
8 vertex pipes
24 pixel pipes

Why didn't GPU Computing took off sooner?

- GPU Architecture
  - Gaming oriented, process pixel for display
  - Single threaded operations
  - No shared memory
- Development Tools
  - Graphics oriented (OpenGL, GLSL)
  - University research (Brook)
  - Assembly language
- Deployment
  - Gaming solutions with limited lifetime
  - Expensive OpenGL professional graphics boards
  - No HPC compatible products

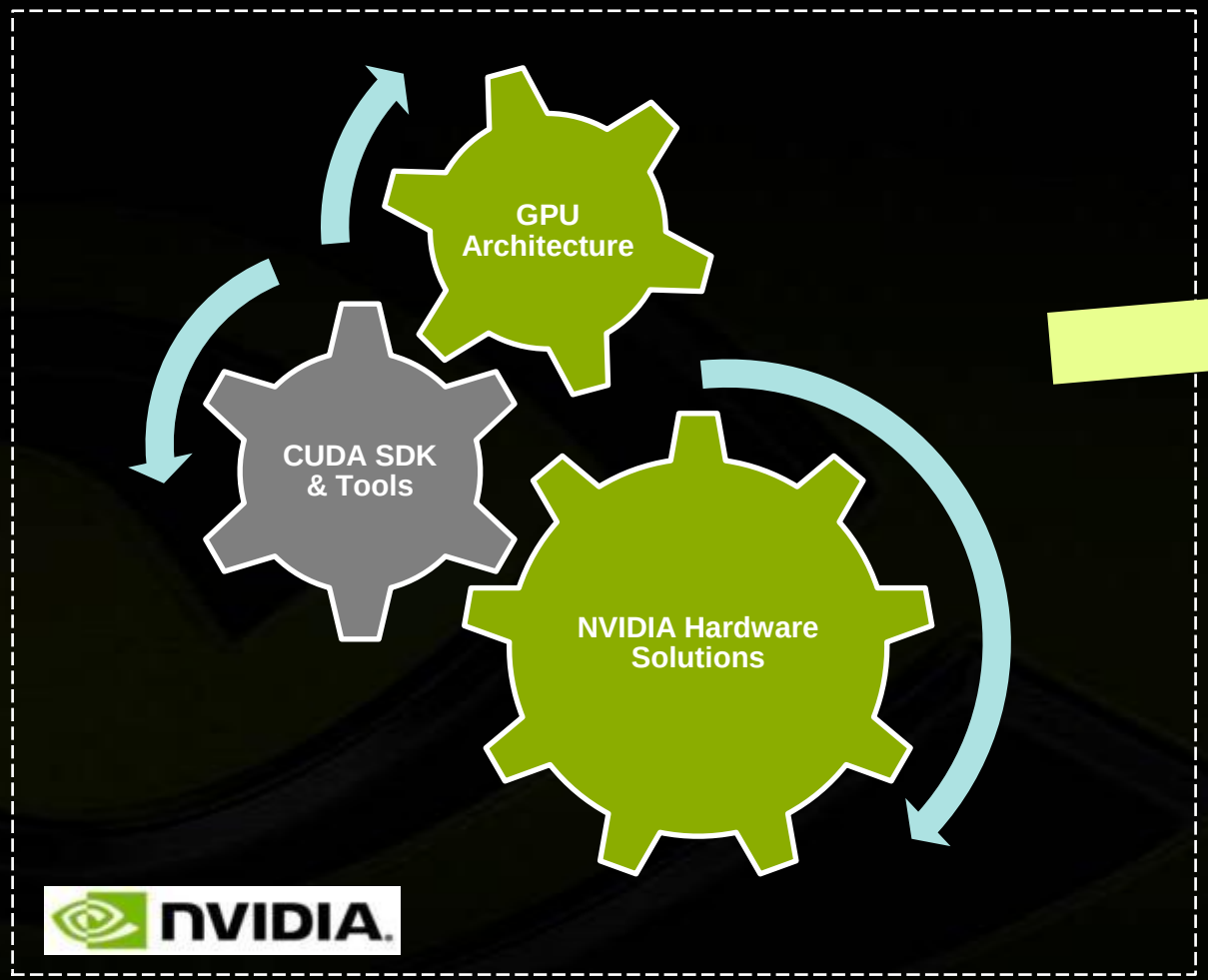
NVIDIA invested in GPU Computing back in 2004

- Strategic move for the company
  - Expand GPU architecture beyond pixel processing
  - Future platforms will be hybrid, multi/many cores based
- Hired key industry experts
  - x86 architecture
  - x86 compiler
  - HPC hardware specialist



Provide a GPU based Compute Ecosystem by 2008

NVIDIA GPU Computing Ecosystem



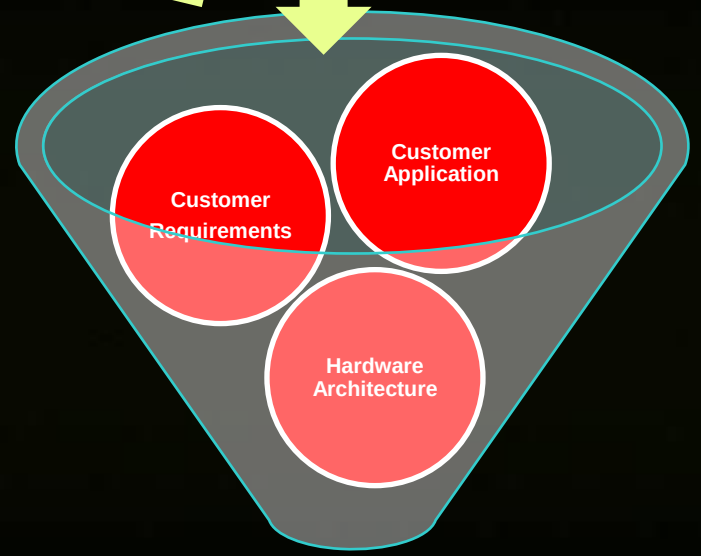
CUDA Training Company

CUDA Development Specialist

ISV

OEMs

Hardware Architect VAR



Deployment

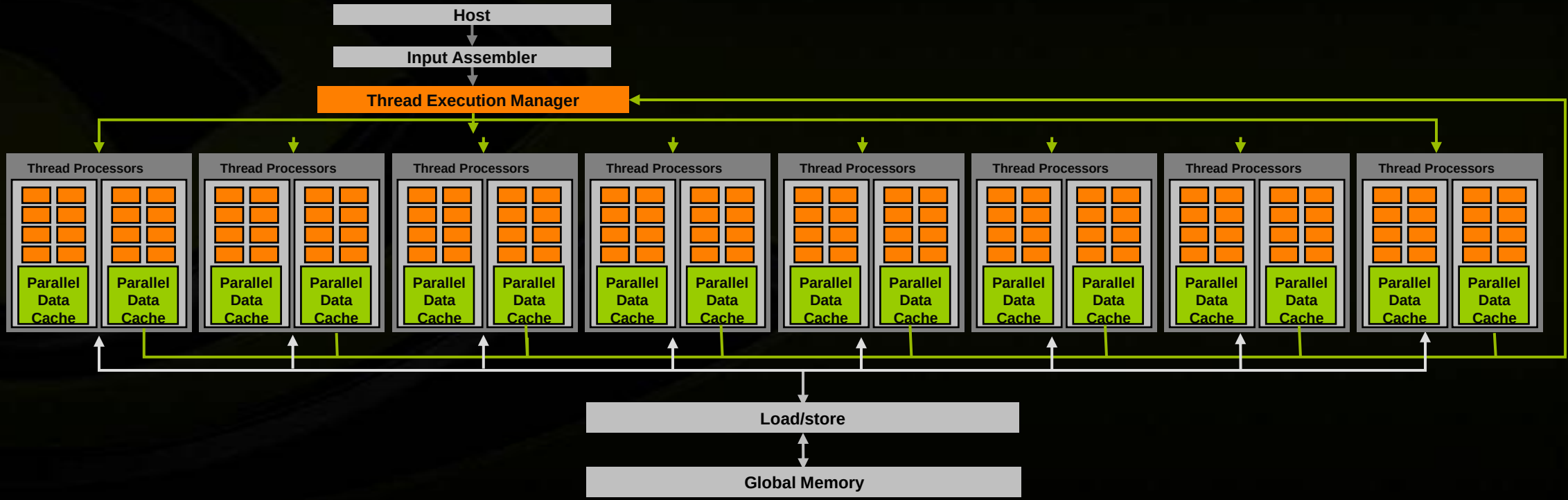
The Past 2 years

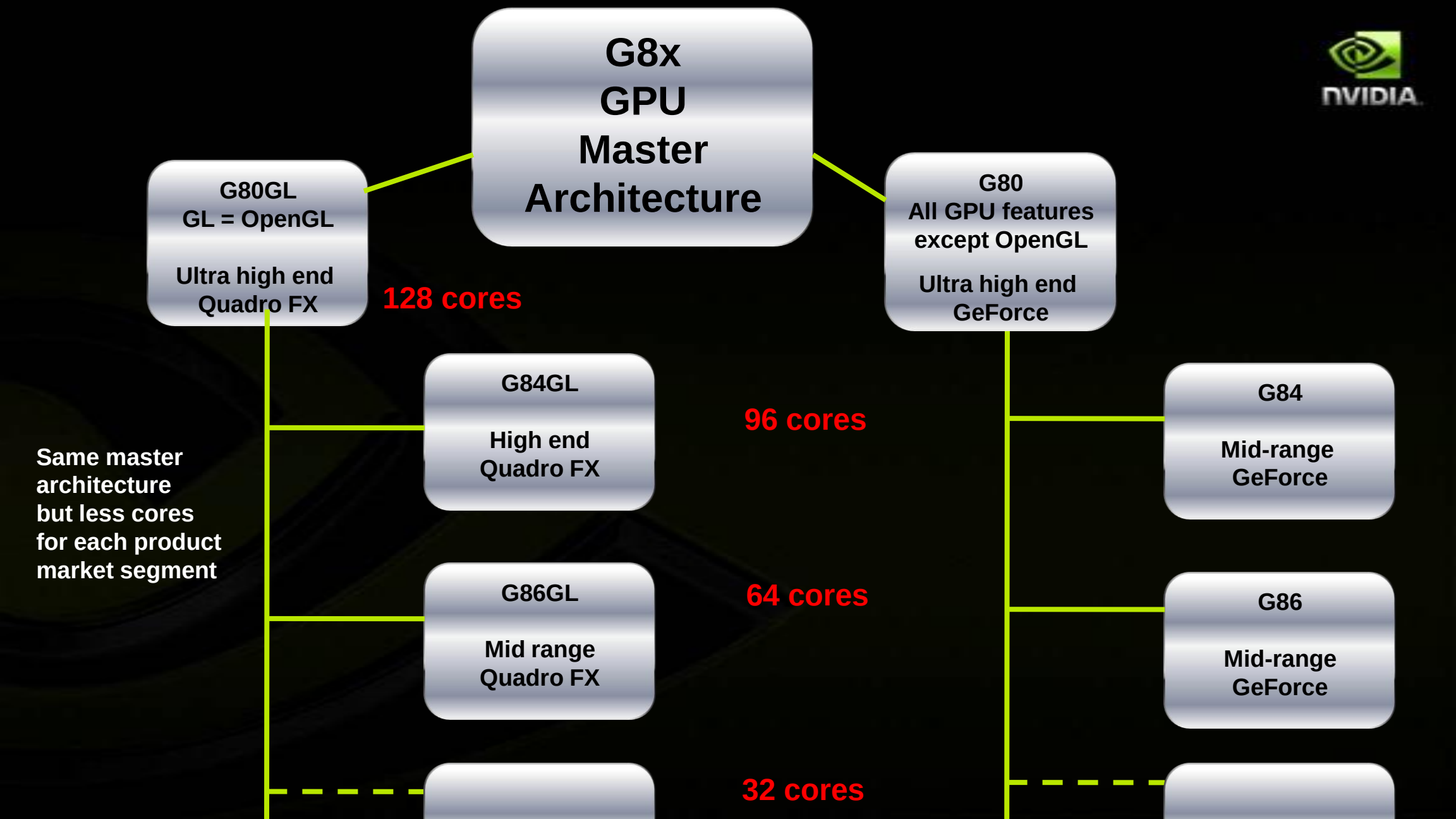
- 2006
  - G80, first GPU with built-in Compute features
    - 128 core, scalable multi-threaded architecture
  - CUDA SDK Beta
- 2007
  - Tesla HPC product line
  - CUDA SDK 1.0, 1.1
  - University trainings programs

#1 GPU Architecture

G80 – GPU Architecture Tuned for Compute

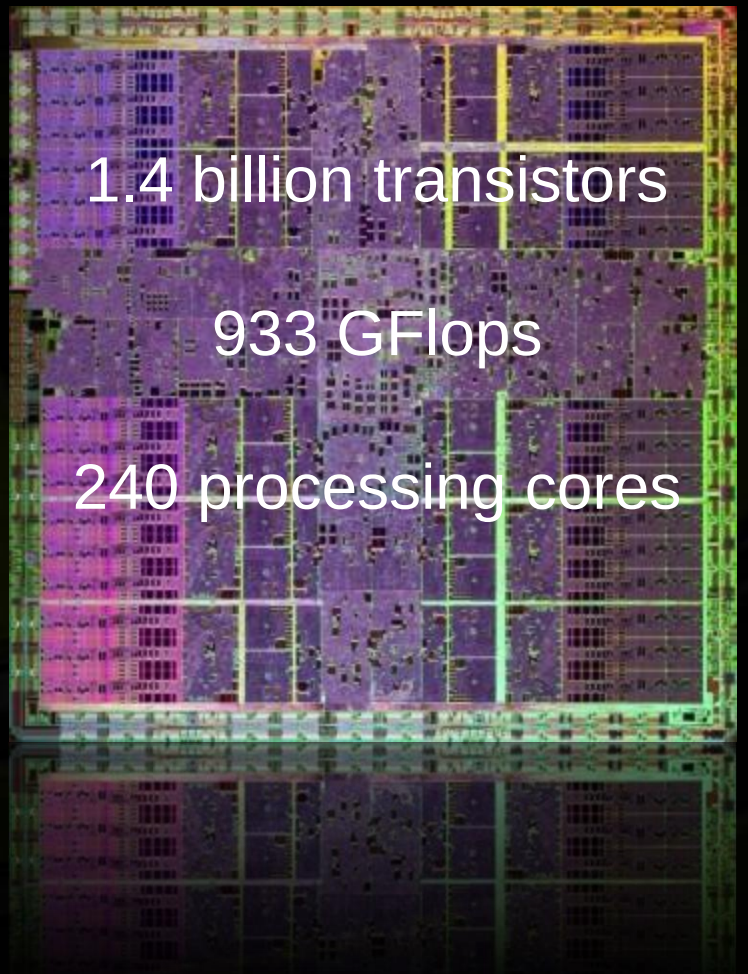
- Processors execute computing threads
- Thread Execution Manager issues threads
- 128 Thread Processors grouped into 16 Multiprocessors
- Parallel Data Cache (Shared Memory) enables thread cooperation




G8x
GPU
Master
Architecture
G80GL
GL = OpenGL
Ultra high end
Quadro FX
128 cores
G80
All GPU features
except OpenGL
Ultra high end
GeForce
G84GL
High end
Quadro FX
96 cores
G84
Mid-range
GeForce
Same master
architecture
but less cores
for each product
market segment
G86GL
Mid range
Quadro FX
64 cores
G86
Mid-range
GeForce
32 cores

June 2008: NVIDIA GT200 GPU

2nd Generation Parallel Computing Architecture



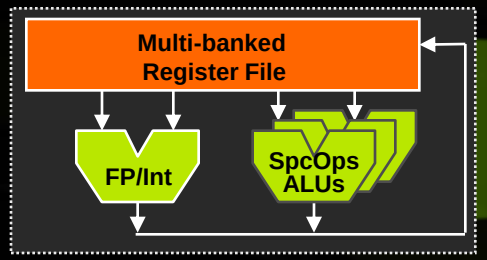
NVIDIA GPU marketing names

- **GT200** — **Consumer GeForce**
- **GT200GL** — **Professional Quadro**
- **T10** — **HPC Tesla**

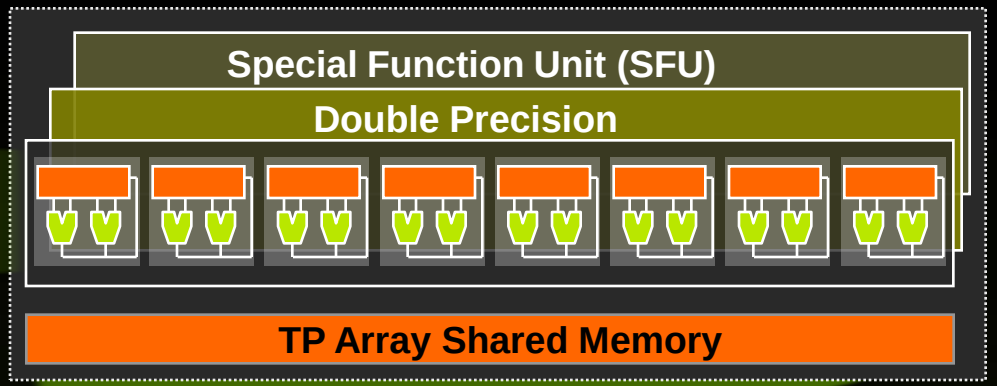
GT200, GT200GL and T10 are based on the same master architecture but different features are enabled for each target market

GT200 / GT200GL / T10

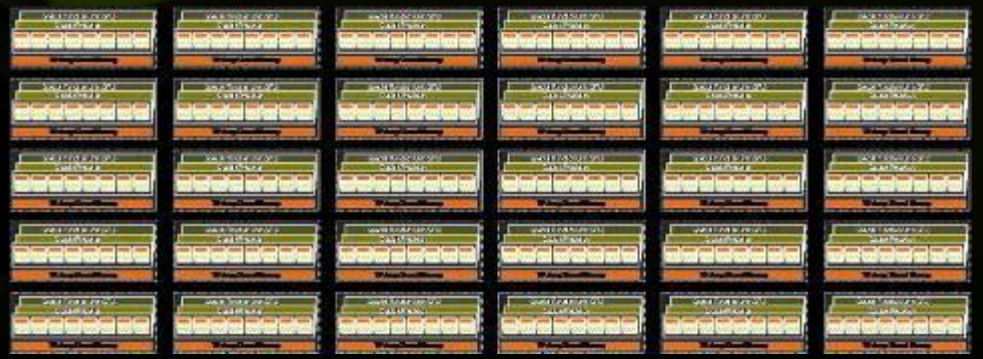
Thread Processor (TP)



Thread Processor Array (TPA)



- 240 SP thread processors
- 30 DP thread processors
- Full scalar processor
- IEEE 754 64-bit floating point



G8x

- Up to 128 cores
- No async. transfer*

- GeForce 8-serie
- Quadro FX 5600/4600
- Tesla C870

G9x

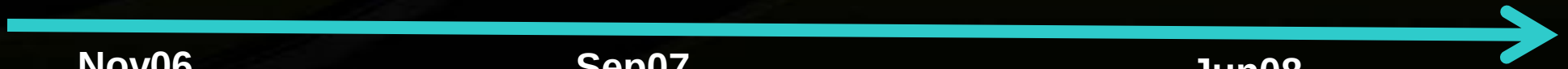
- Up to 112 cores
- Async. transfer

- GeForce 9-serie
- Quadro FX 3700

GT200

- Up to 240 cores
- Async. Transfer
- Double Precision

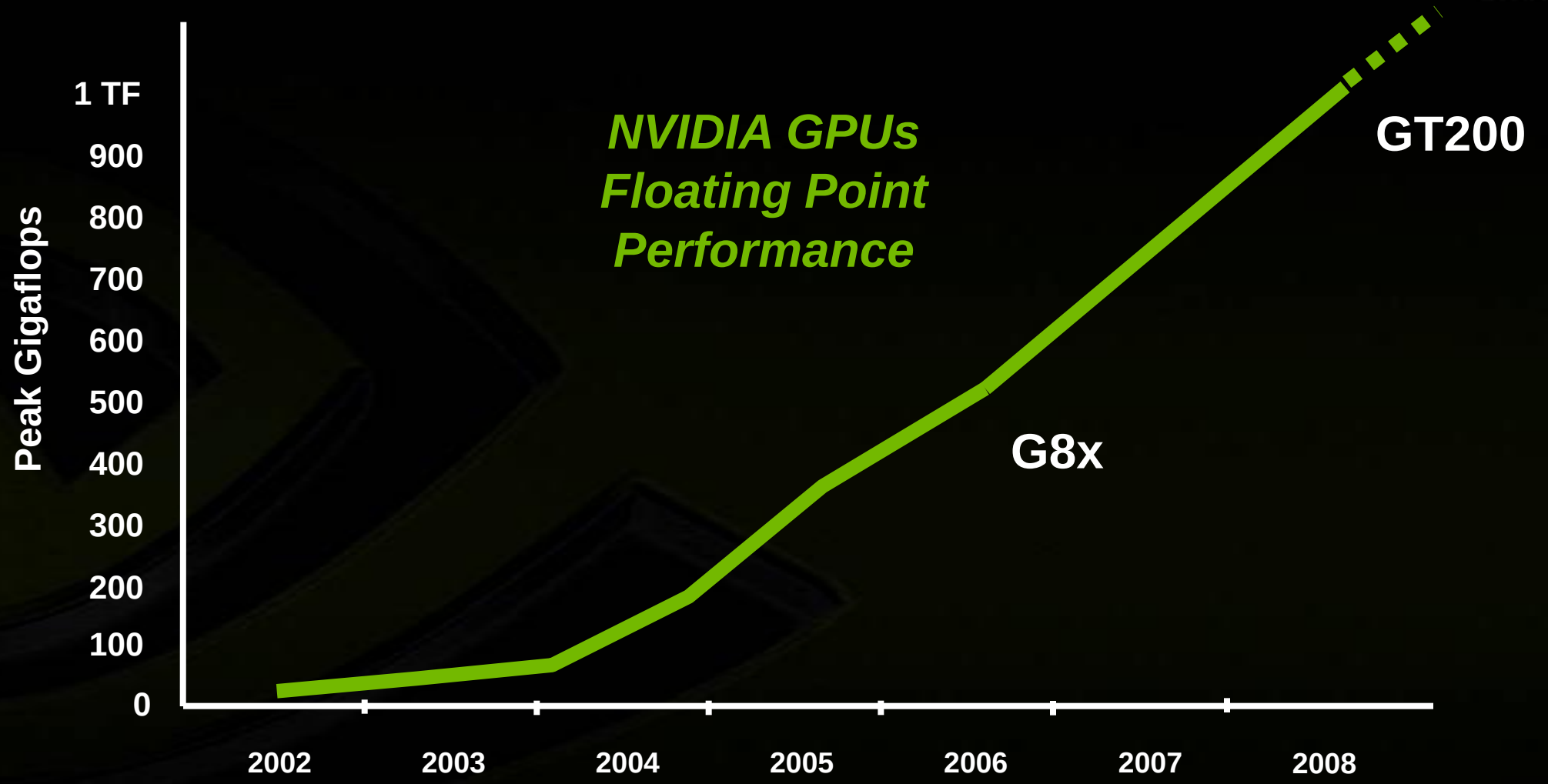
- GeForce GTX280/260
- Quadro FX 5800/4800
- Tesla C1060



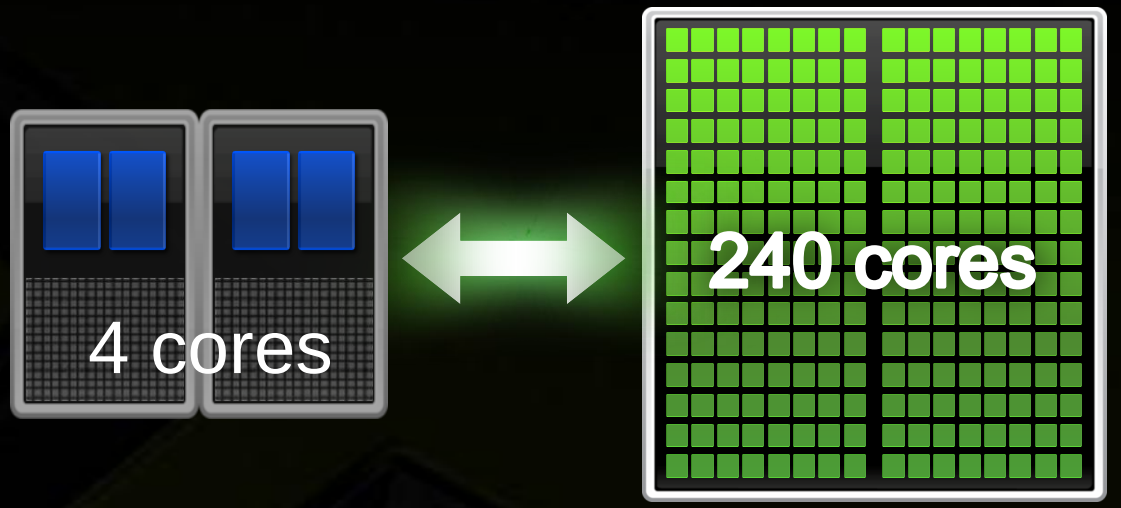
Nov06 Sep07 Jun08

Asynchronous transfer is used to hide data transfer from CPU to GPU or GPU to CPU while the GPU is processing data, thus improving application speedup

Ever Increasing Floating Point Performance



GPU Computing : Heterogeneous Computing



Computing with CPU + GPU

Heterogeneous Computing

y[i] = a*x[i] + y[i] – Computed Sequentially

y[i] = a*x[i] + y[i] – Computed In Parallel

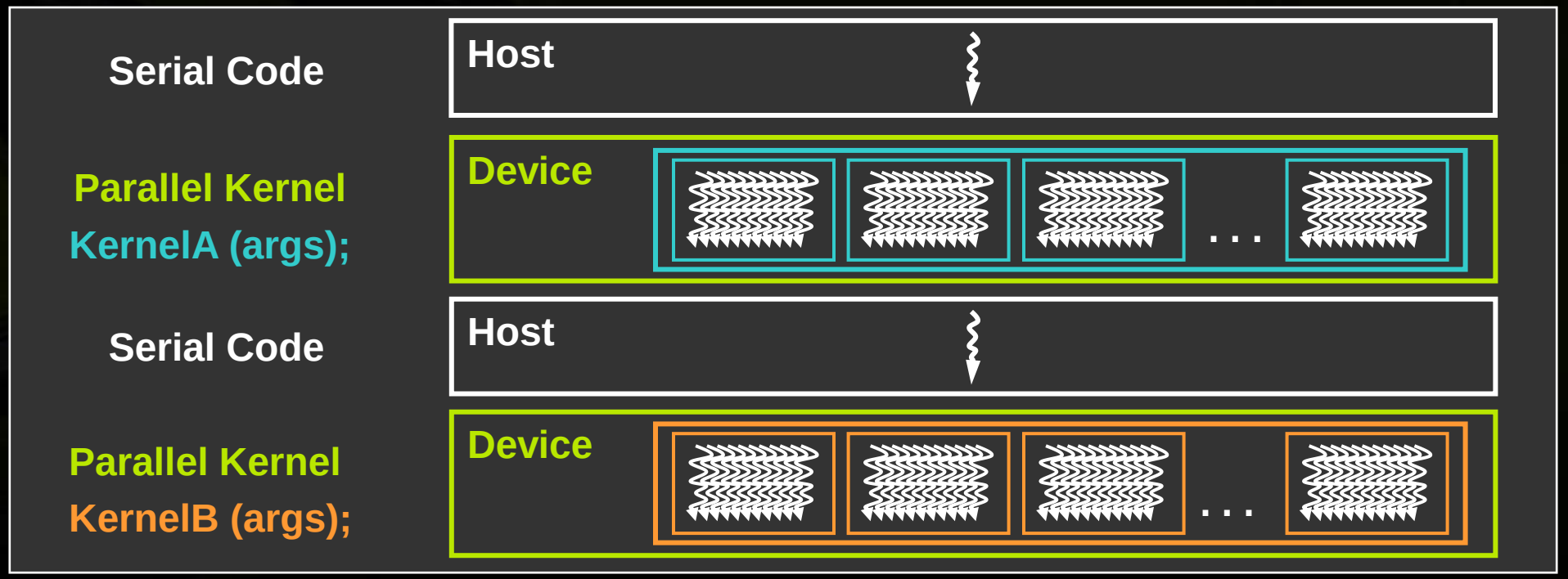
#2 CUDA SDK

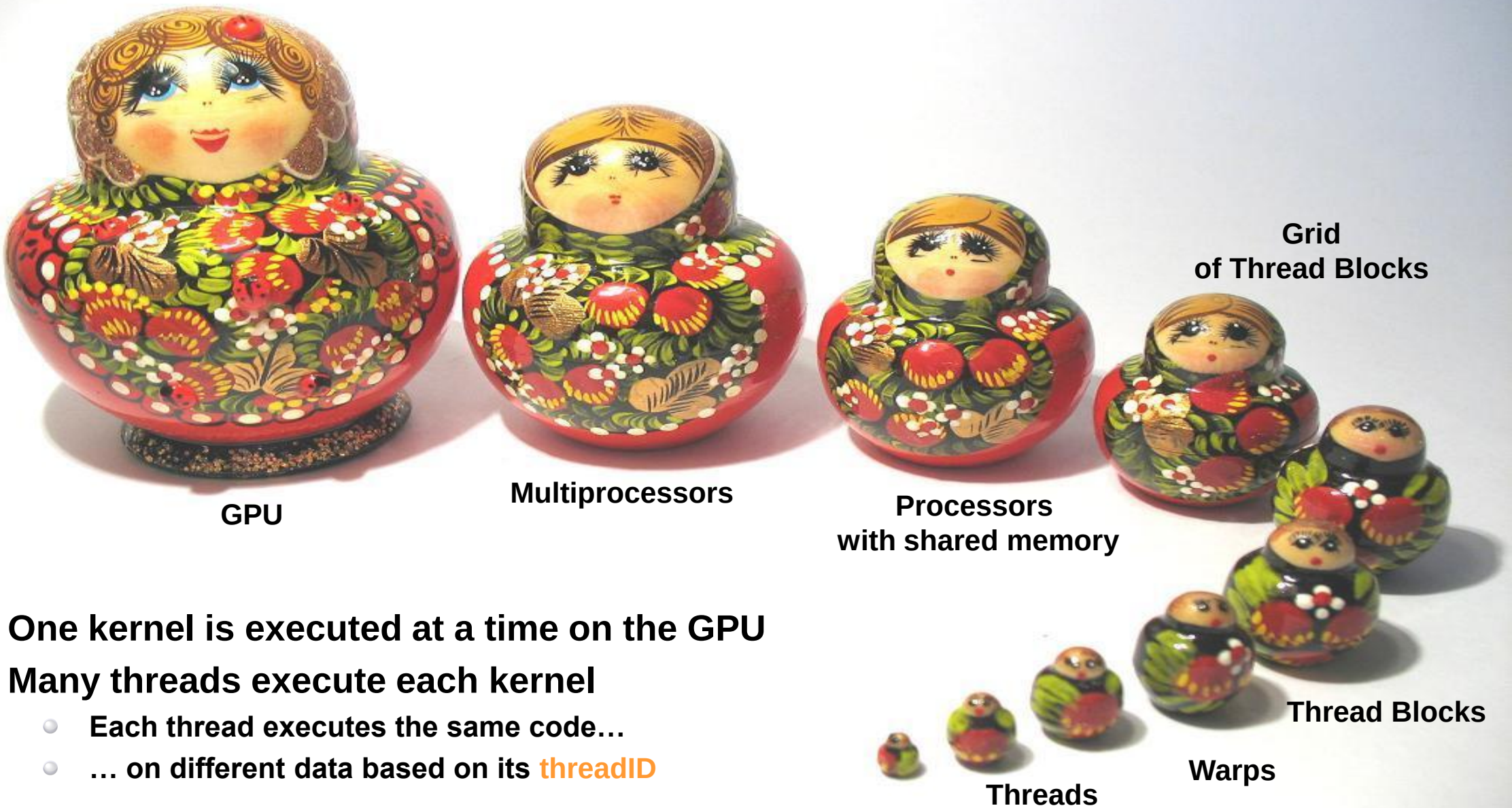
CUDA is C for Parallel Processors

- 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
- CUDA parallelism applies to both CPUs and GPUs
  - Compile the same program source to run on different platforms with widely different parallelism
  - Map to CUDA threads to GPU threads or to CPU vectors

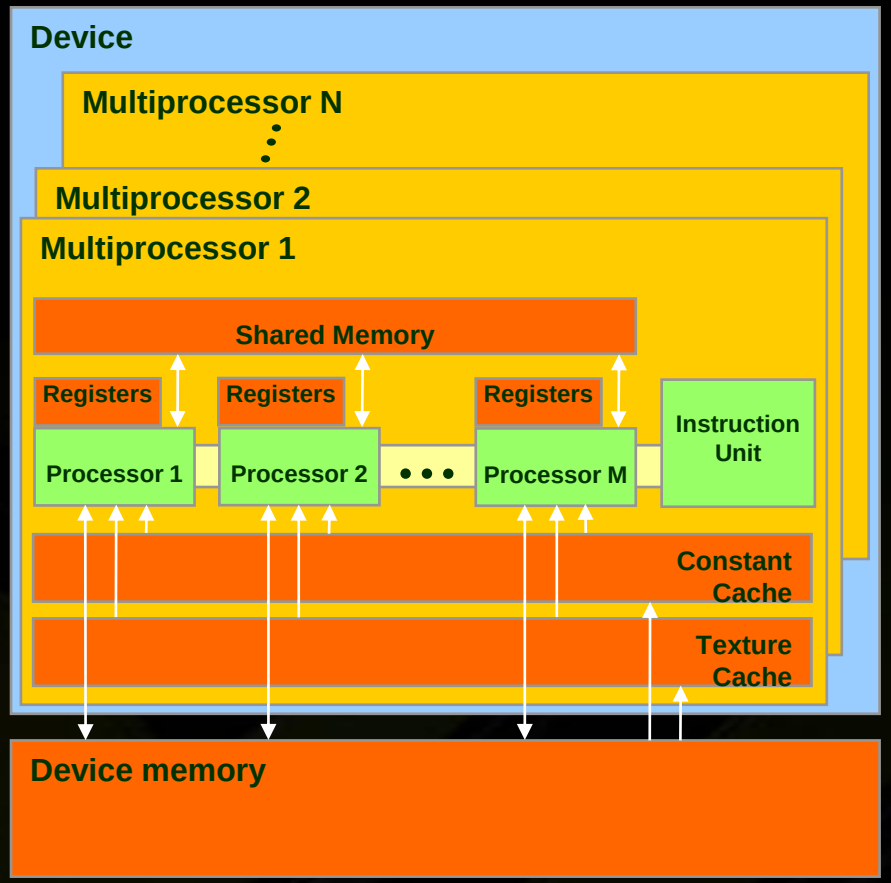
Heterogeneous Programming

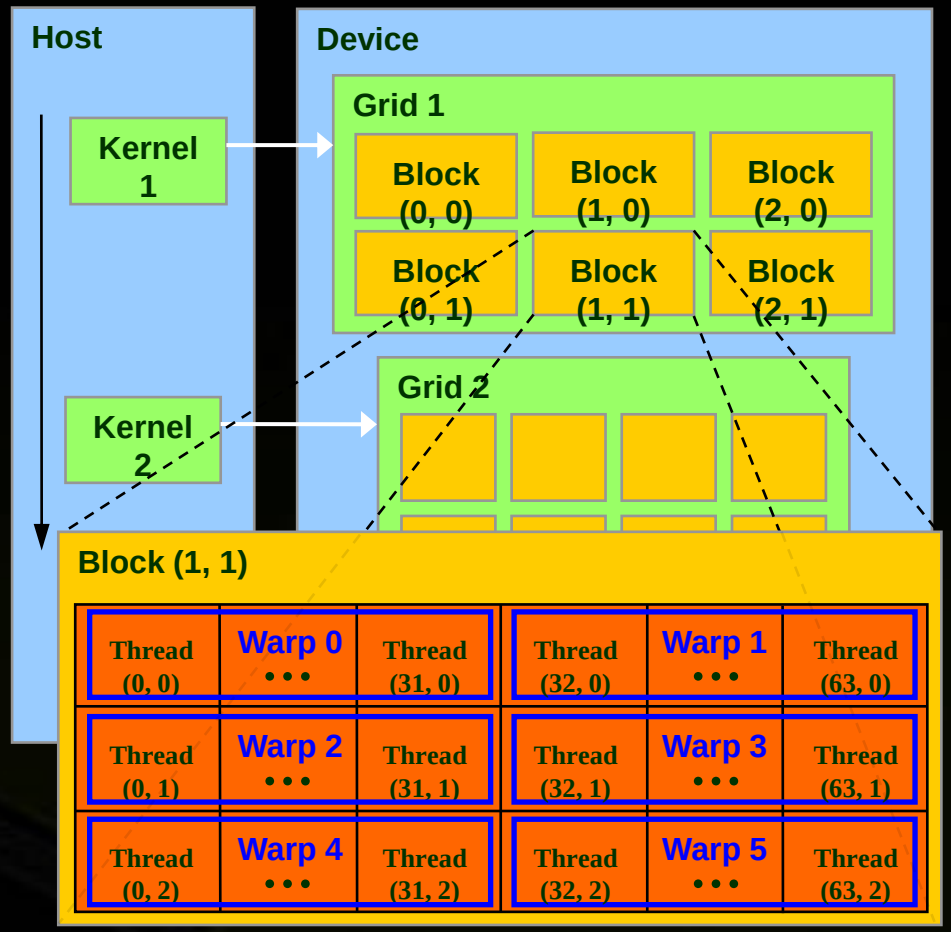
- CUDA = serial program with parallel kernels, all in C
  - Serial C code executes in a host thread (i.e. CPU thread)
  - Parallel kernel C code executes in many device threads across multiple processing elements (i.e. GPU threads)





- One kernel is executed at a time on the GPU
- Many threads execute each kernel
  - Each thread executes the same code…
  - … on different data based on its threadID



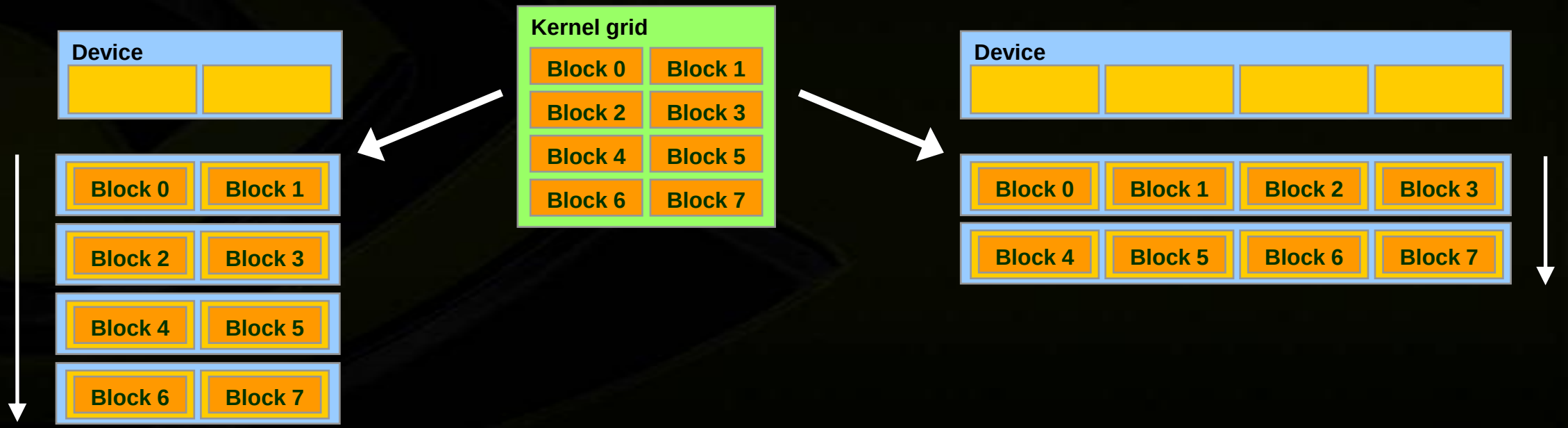


Device = **GPU** = set of multiprocessors
Multiprocessor = set of processors & shared memory

Kernel = GPU program
Grid = array of thread blocks that execute a kernel
Thread block = group of SIMD threads
Warp = group of threads

Transparent Scalability

- Hardware is free to schedule thread blocks on any processor
  - So they can run in any order, concurrently or sequentially
- This independence gives scalability
  - A kernel scales across parallel multiprocessors



CUDA Language: C with Minimal Extensions

- Philosophy: provide minimal set of extensions necessary to expose power

- Declaration specifiers to indicate where things live

```
__global__ void KernelFunc(...);   // kernel function, runs on device
__device__ int  GlobalVar;         // variable in device memory
__shared__ int  SharedVar;         // variable in per-block shared memory
```

- Extend function invocation syntax for parallel kernel launch

```
KernelFunc<<<500, 128>>>(...);     // launch 500 blocks w/ 128 threads each
```

- Special variables for thread identification in kernels

```
dim3 threadIdx;  dim3 blockIdx;  dim3 blockDim;  dim3 gridDim;
```

- Intrinsics that expose specific operations in kernel code

```
__syncthreads();                   // barrier synchronization within kernel
```

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

CUDA Toolkit

The CUDA development environment includes:

- **nvcc C compiler**
- **CUDA FFT and BLAS libraries for the GPU**
- **Profiler**
- **gdb debugger for the GPU**
- **CUDA runtime driver (also available in the standard NVIDIA GPU driver)**
- **CUDA programming manual**

NVIDIA DEVELOPER ZONE

Search Developer Zone

- Developer Site Homepage
- Developer News Homepage
- Developer Login
- Become a Registered Developer
- Developer Tools
- Developer Forums
- Documentation
- NVIDIA PhysX
- DirectX
- OpenGL
- CUDA GPU Computing
- Handheld
- Events Calendar
- Newsletter Sign-Up
- Newsletter Archive
- Drivers
- Jobs (1)
- Contact
- Legal Information
- Site Feedback

Last Updated: 05 / 05 / 2008

MATLAB plug-in for CUDA

This MATLAB plug-in for CUDA provides: acceleration of standard MATLAB 2D FFTs and CUDA/MEX example plug-in and build environment using Chris Bretherton's Fourier spectral simulation of 2D fluid flow MATLAB scripts from his course material at the University of Washington.

When MATLAB makes 2D FFT calls of any size, the NVIDIA plug-in intercepts them and handles them with a MEX file that in-turn utilizes an optimized CUDA FFT implementation on the GPU.

This is transparent to MATLAB users. Note: this current implementation uses a single-precision 2-FFT on the NVIDIA hardware, so the results are not in 64-bit precision that the native MATLAB implementation uses without the NVIDIA plug-in. Can be run with and without CUDA acceleration. Time to run the example shows a 14X speedup (from 216 seconds to 15 seconds using CUDA via the MEX file interface).

The MEX file example and build environment uses an FS_2Dflow example but the method is illustrative on how to write a custom CUDA interface via a MATLAB MEX file interface for all CUDA libraries .

[Download] MathWorks MATLAB® Plug-in for Linux

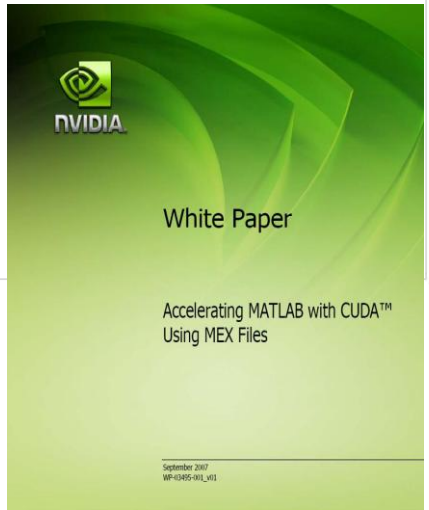
[Download] MathWorks MATLAB® Plug-in for Windows

[Download] Whitepaper: Accelerating MathWorks MATLAB® with CUDA

Previous Versions

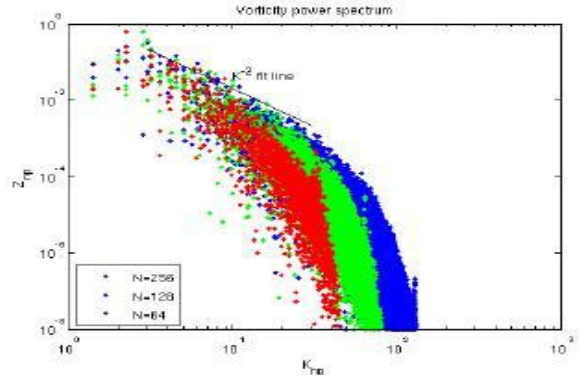
[Download] CUDA 1.0 Plug-in for Linux

[Download] CUDA 1.0 Plug-in for Windows



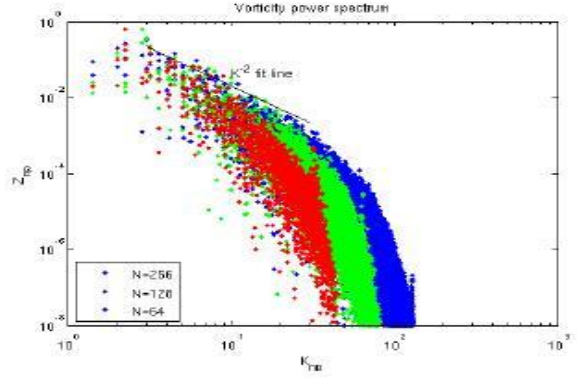
Using MATLAB on Linux, the results for the computation of the advection of an elliptic vortex on a 256×256 mesh, stream function (left) and vorticity (right) in Figure 1 required 168 seconds. By contrast, the results using MATLAB with CUDA in Figure 2 required only 14.9 seconds.

For a better comparison of the quality of the results, we ran a 2D isotropic turbulence simulation compared the vorticity power spectra of the different runs. The first used the original MATLAB code (Figure 3) and the second used MATLAB accelerated with CUDA code (Figure 4). Even for this quantity, that is very sensitive to fine scales, the results are in close agreement.



168”

Figure 3. Final Results Using MATLAB



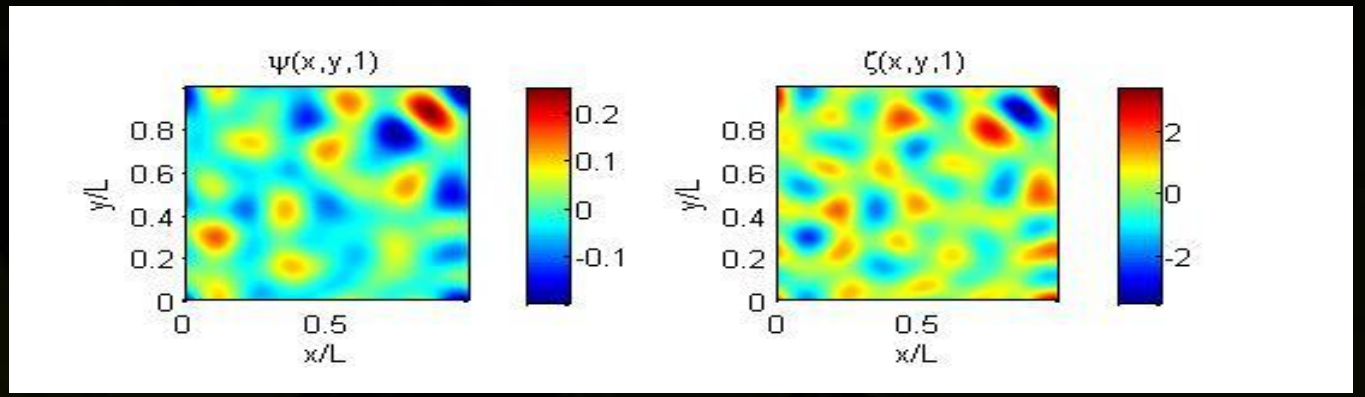
15”

x11 speedup

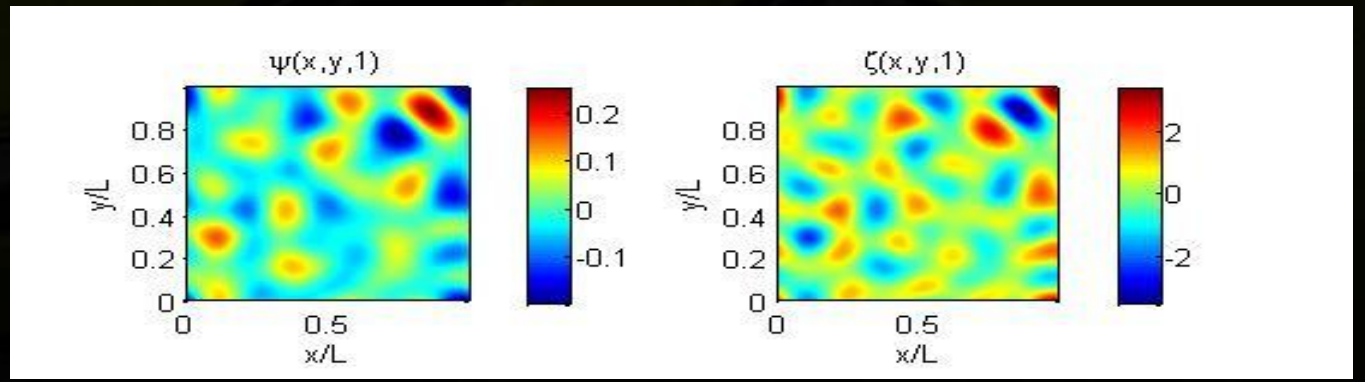
Figure 4. Final Results Using MATLAB with CUDA

Pseudo-spectral simulation of 2D Isotropic turbulence.

512x512 mesh, 400 RK4 steps, Windows XP, MATLAB file
http://www.amath.washington.edu/courses/571-winter-2006/matlab/FS_2Dturb.m



MATLAB
992 seconds

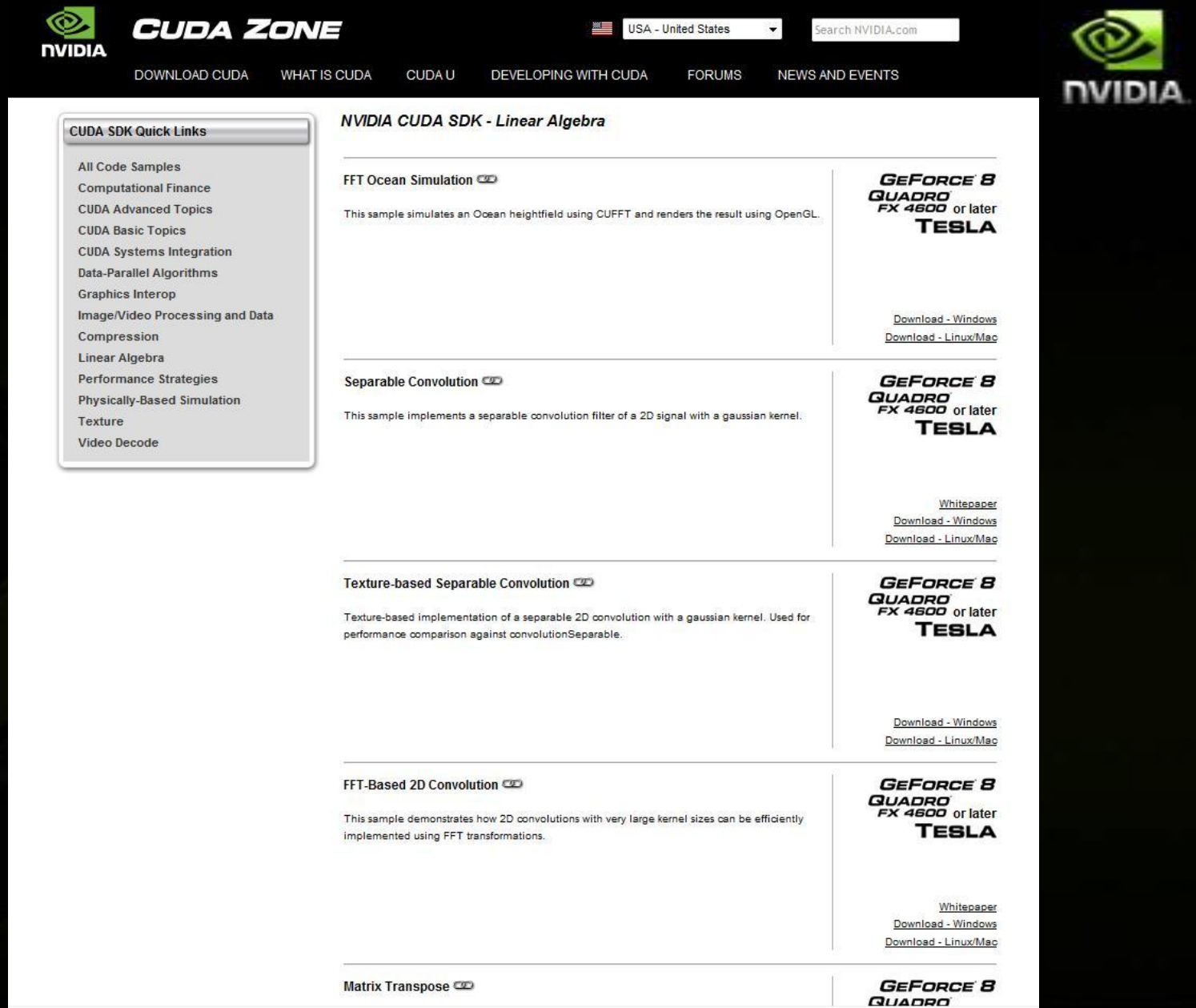


MATLAB with CUDA
(single precision FFTs)
93 seconds

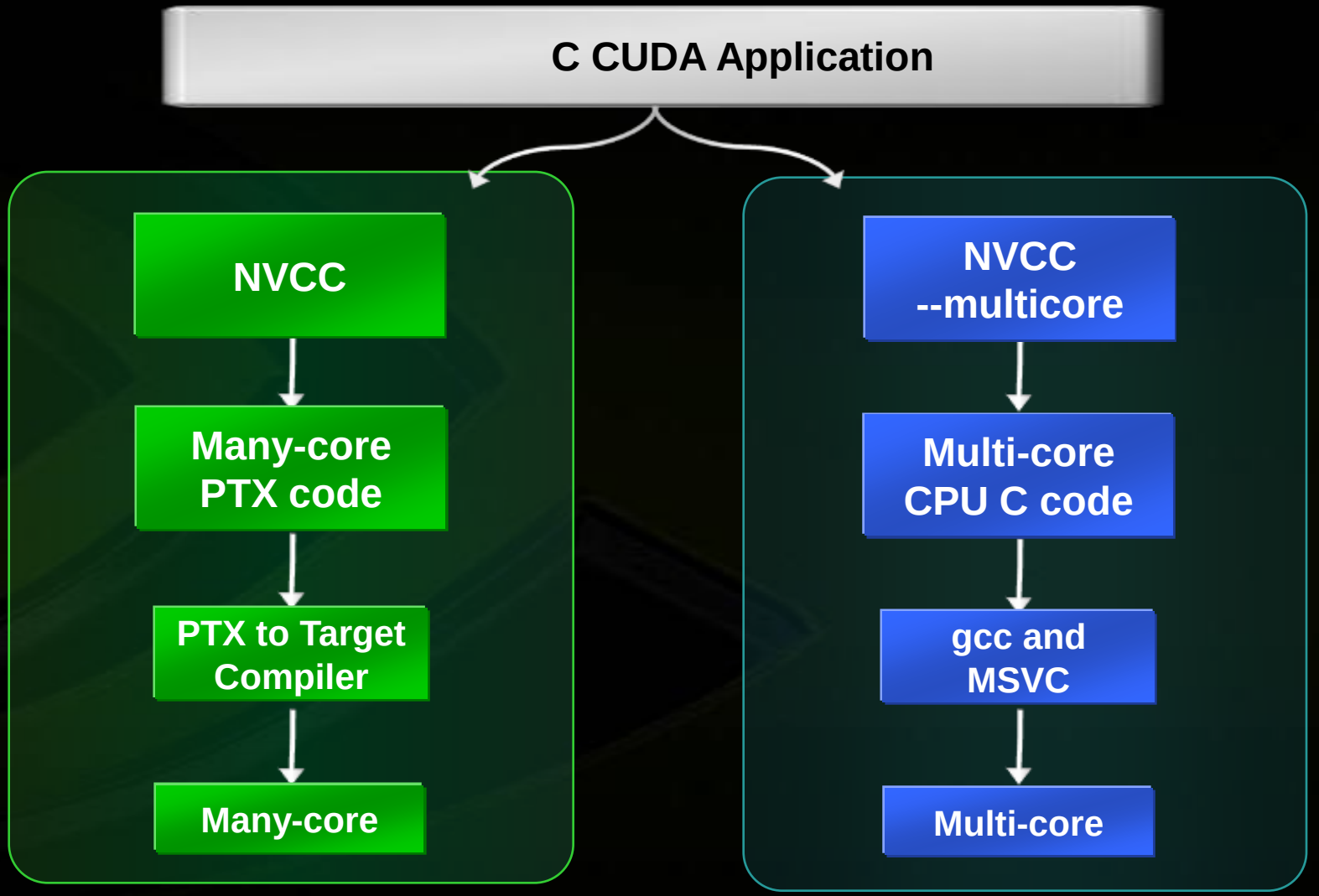
CUDA SDK Code Samples

- CUDA Basic Topics
- CUDA Advanced Topics

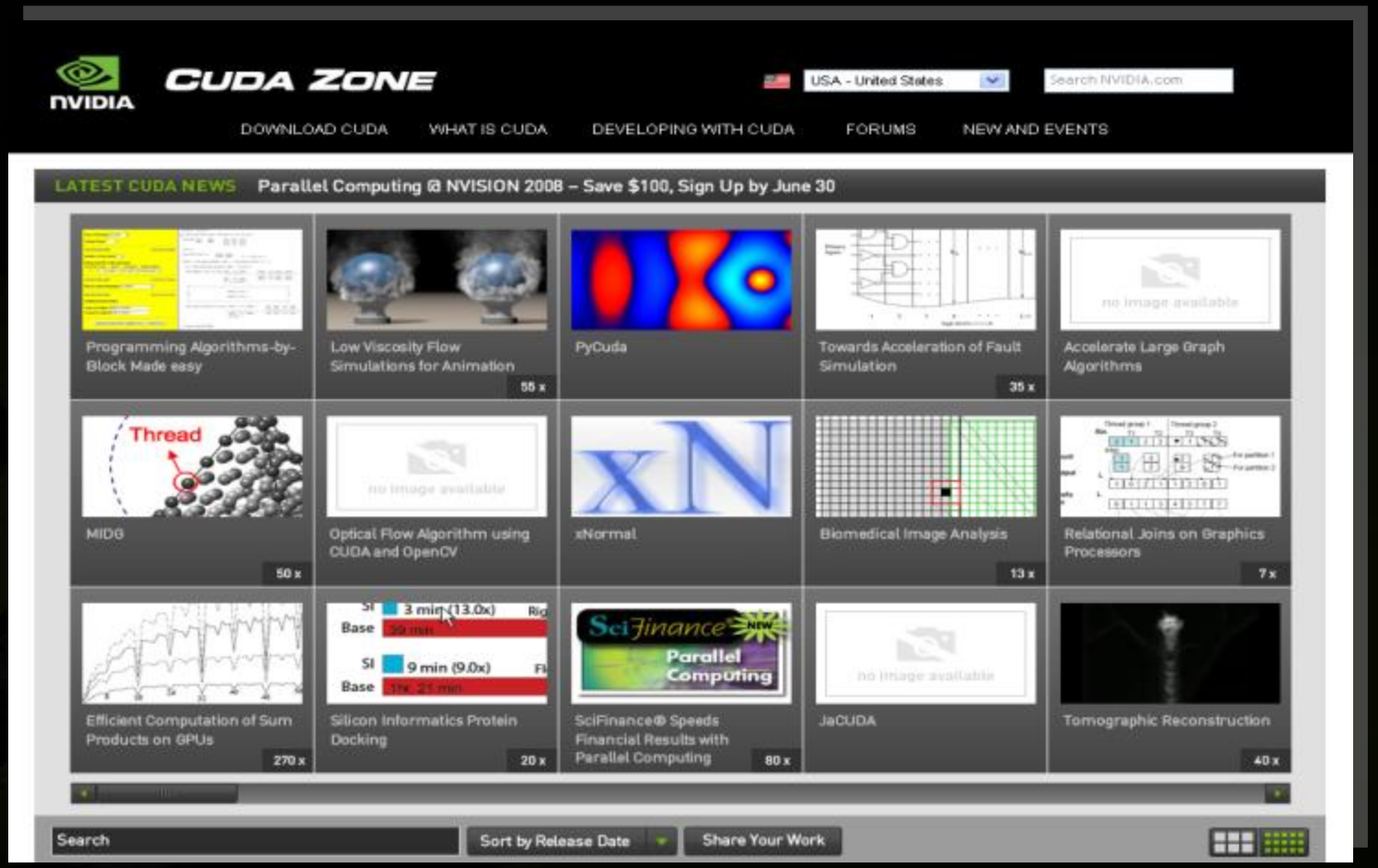
- Computational Finance
- Parallel Algorithms
- Linear Algebra
- Physically-Based Simulation
- Texture
- Video Decode
- Image/Video Processing
- Graphics Interop
- Performances Strategies



CUDA 2.0: Many-core + Multi-core support



CUDA Zone: www.nvidia.com/cuda



CUDA Tutorial

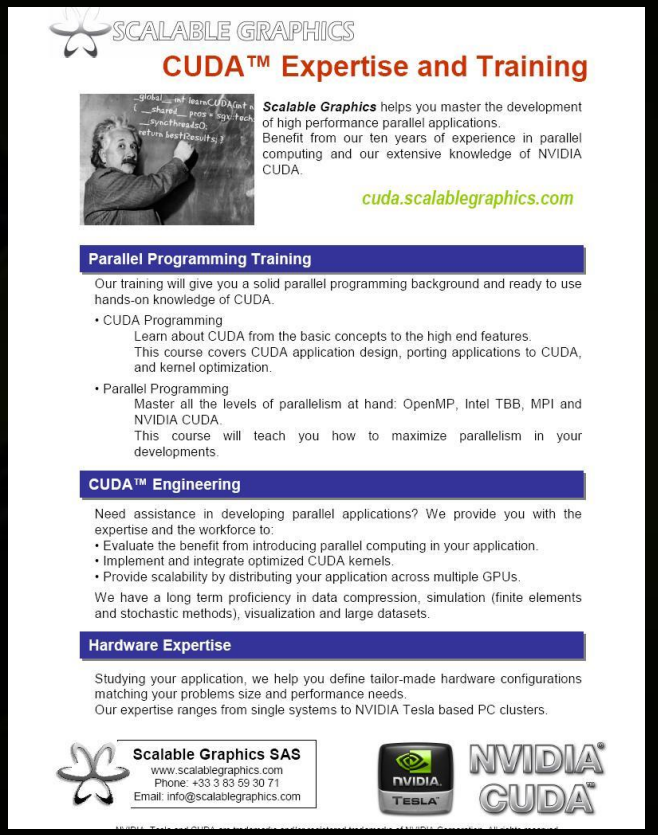
- **Latest and greatest on www.nvidia.com/object/cuda_education.html**

- **NVIDIA CUDA Tutorial, SuperComputing 2008 Austin Nov08**
  www.gpgpu.org/sc2008

  Introduction (PDF)
  Parallel Programming with CUDA (PDF)
  CUDA Toolkit (PDF)
  Optimizing CUDA (PDF)
  Seismic Imaging on NVIDIA GPUs: Algorithms and Porting & Production Experiences (PDF)
  Molecular Visualization and Analysis (PDF)
  Molecular Dynamics (PDF)
  Computational Fluid Dynamics (PDF)

NVIDIA CUDA French Partners
Training & Development

- CAPS
- ANEO
- Scalable Graphics
- GPU-Tech
- HPC Project



SCALABLE GRAPHICS

CUDA™ Expertise and Training

Scalable Graphics helps you master the development of high performance parallel applications.
Benefit from our ten years of experience in parallel computing and our extensive knowledge of NVIDIA CUDA.

cuda.scalablegraphics.com

Parallel Programming Training

Our training will give you a solid parallel programming background and ready to use hands-on knowledge of CUDA.

- CUDA Programming
  Learn about CUDA from the basic concepts to the high end features.
  This course covers CUDA application design, porting applications to CUDA, and kernel optimization.
- Parallel Programming
  Master all the levels of parallelism at hand: OpenMP, Intel TBB, MPI and NVIDIA CUDA.
  This course will teach you how to maximize parallelism in your developments.

CUDA™ Engineering

Need assistance in developing parallel applications? We provide you with the expertise and the workforce to:
- Evaluate the benefit from introducing parallel computing in your application.
- Implement and integrate optimized CUDA kernels.
- Provide scalability by distributing your application across multiple GPUs.

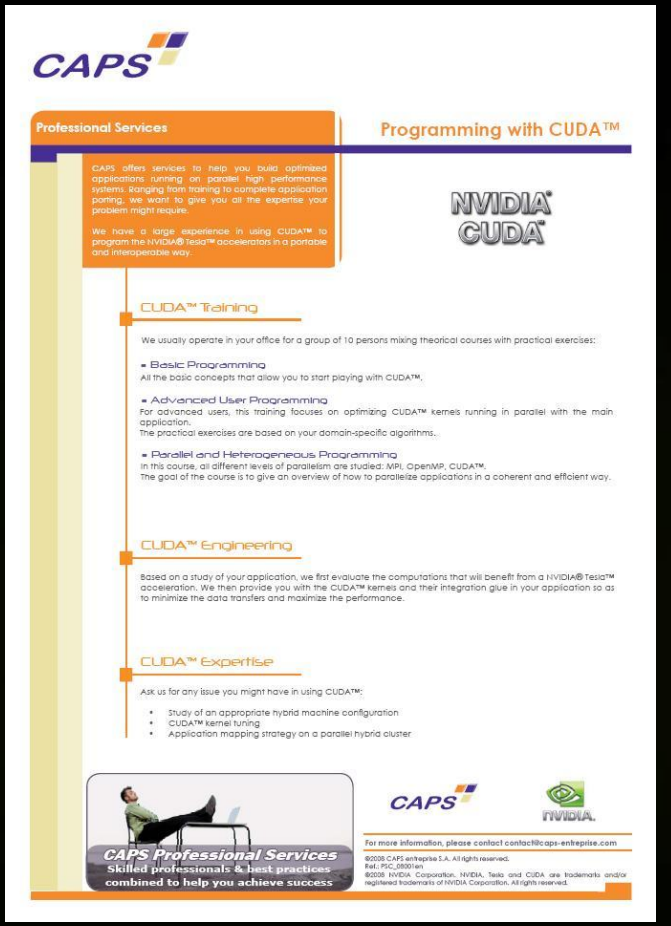
We have a long term proficiency in data compression, simulation (finite elements and stochastic methods), visualization and large datasets.

Hardware Expertise

Studying your application, we help you define tailor-made hardware configurations matching your problems size and performance needs.
Our expertise ranges from single systems to NVIDIA Tesla based PC clusters.

Scalable Graphics SAS
www.scalablegraphics.com
Phone: +33 3 83 59 30 71
Email: info@scalablegraphics.com

NVIDIA TESLA NVIDIA CUDA



CAPS

Professional Services

Programming with CUDA™

CAPS offers services to help you build optimized applications running on parallel high performance systems. Ranging from training to complete application porting, we want to give you all the expertise your problem might require.

We have a large experience in using CUDA™ to program the NVIDIA® Tesla™ accelerators in a portable and interoperable way.

NVIDIA CUDA

CUDA™ Training

We usually operate in your office for a group of 10 persons mixing theorical courses with practical exercises:

- Basic Programming
  All the basic concepts that allow you to start playing with CUDA™.
- Advanced User Programming
  For advanced users, this training focuses on optimizing CUDA™ kernels running in parallel with the main application.
  The practical exercises are based on your domain-specific algorithms.
- Parallel and Heterogeneous Programming
  In this course, all different levels of parallelism are studied: MPI, OpenMP, CUDA™.
  The goal of the course is to give an overview of how to parallelize applications in a coherent and efficient way.

CUDA™ Engineering

Based on a study of your application, we first evaluate the computations that will benefit from a NVIDIA® Tesla™ acceleration. We then provide you with the CUDA™ kernels and their integration glue in your application so as to minimize the data transfers and maximize the performance.

CUDA™ Expertise

Ask us for any issue you might have in using CUDA™:

- Study of an appropriate hybrid machine configuration
- CUDA™ kernel tuning
- Application mapping strategy on a parallel hybrid cluster

CAPS Professional Services
Skilled professionals & best practices combined to help you achieve success

CAPS NVIDIA

For more information, please contact contact@caps-entreprise.com


OpenCL

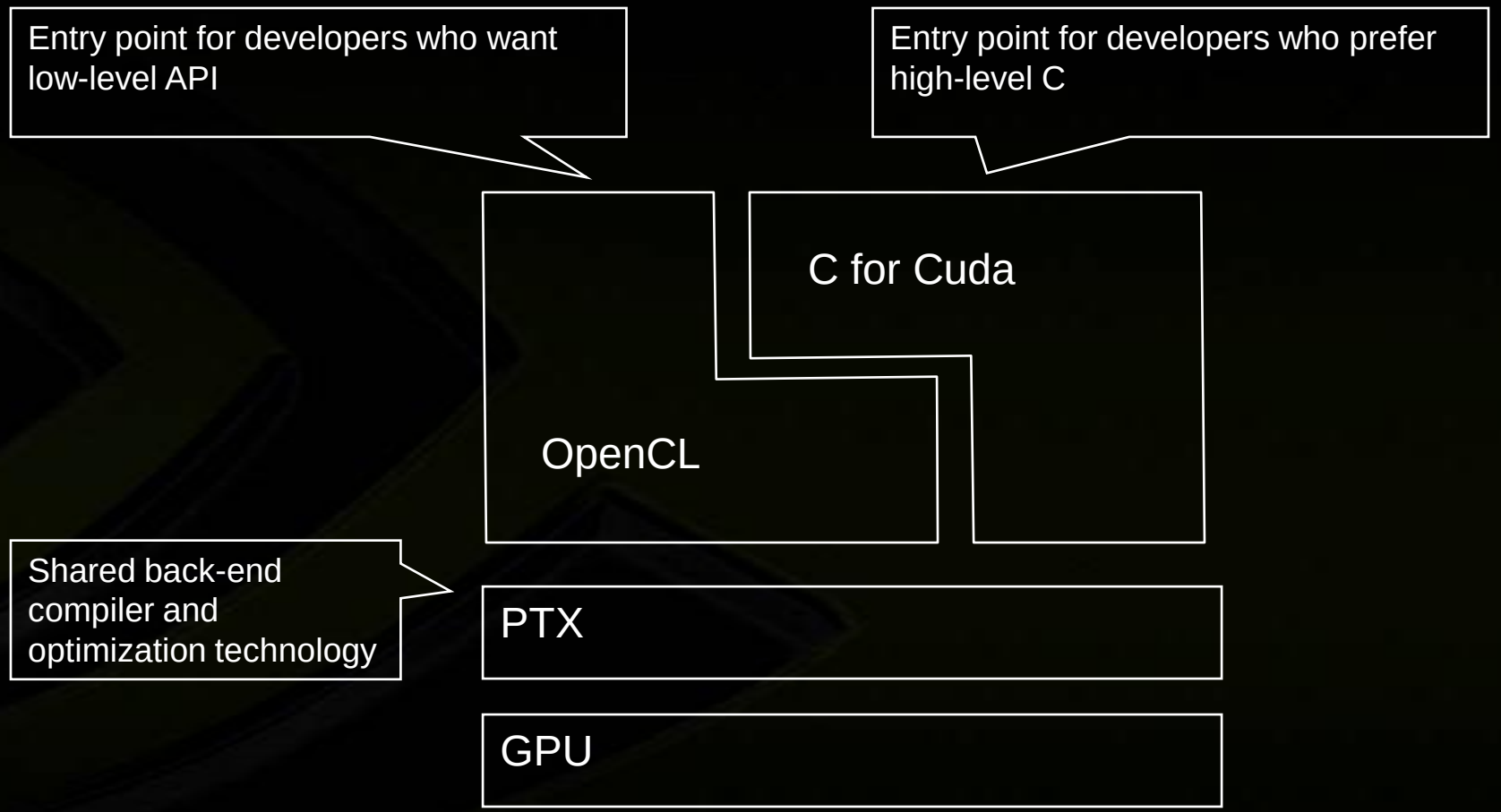
OpenCL

- A new compute API for parallel programming of heterogeneous systems
- Allows developers to harness the compute power of BOTH the GPU and the CPU
- A multi-vendor standards effort managed through the Khronos Group

NVIDIA and OpenCL

- **OpenCL is terrific**
  We support any initiative that unleashes the massive power of the GPU
- **Neil Trevett, NVIDIA VP, chairs Khronos OpenCL working group - several active NVIDIA participants**
- **NVIDIA is working closer with Apple since the inception of OpenCL**
  - OpenCL was developed on NVIDIA GPUs
  - First to show working OpenCL
  - Top to bottom supplier of GPUs for new Apple notebooks

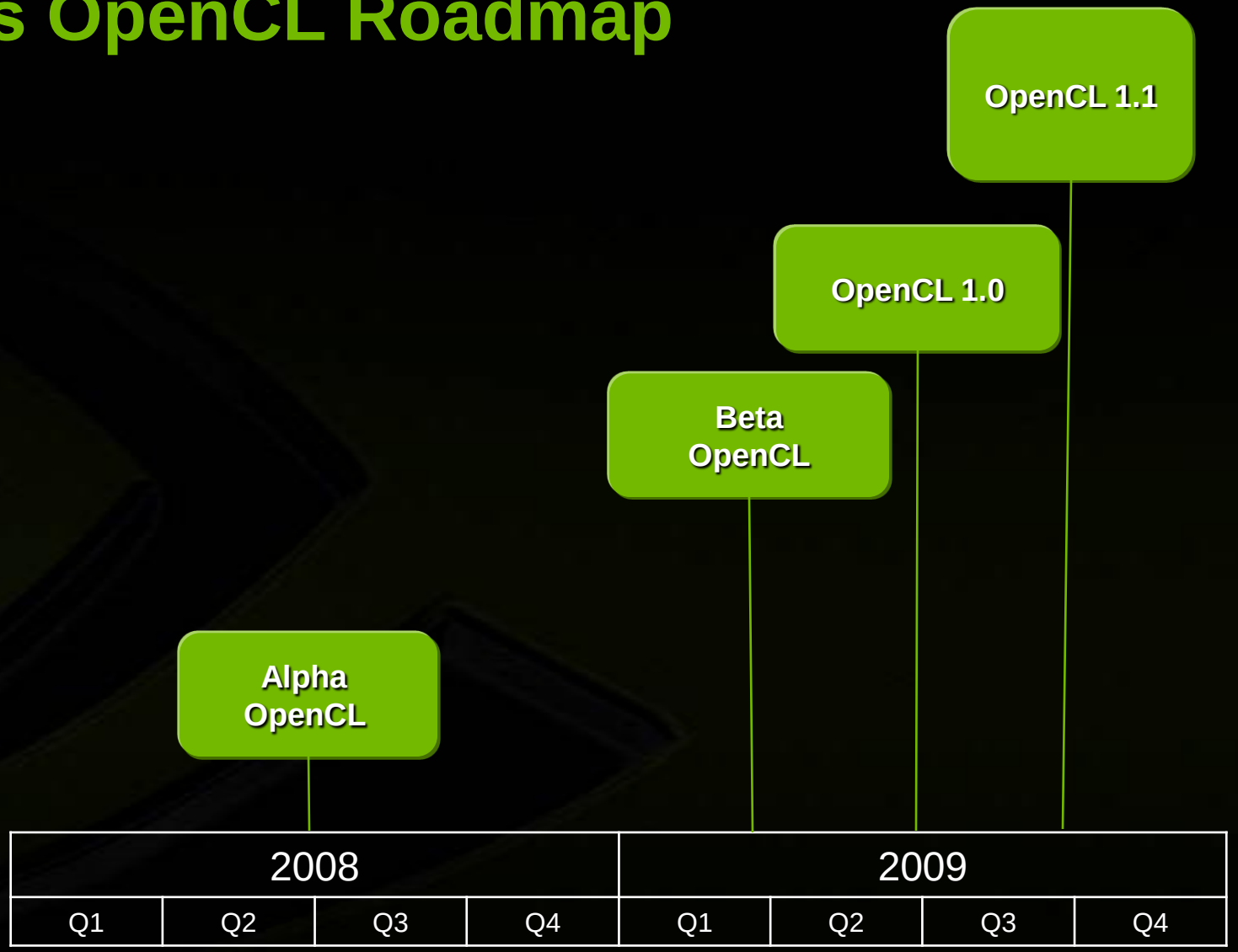
OpenCL and C for Cuda



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
  - Programmer has complete access to hardware device
  - Memory managed by programmer
  - Generates PTX

NVIDIA's OpenCL Roadmap

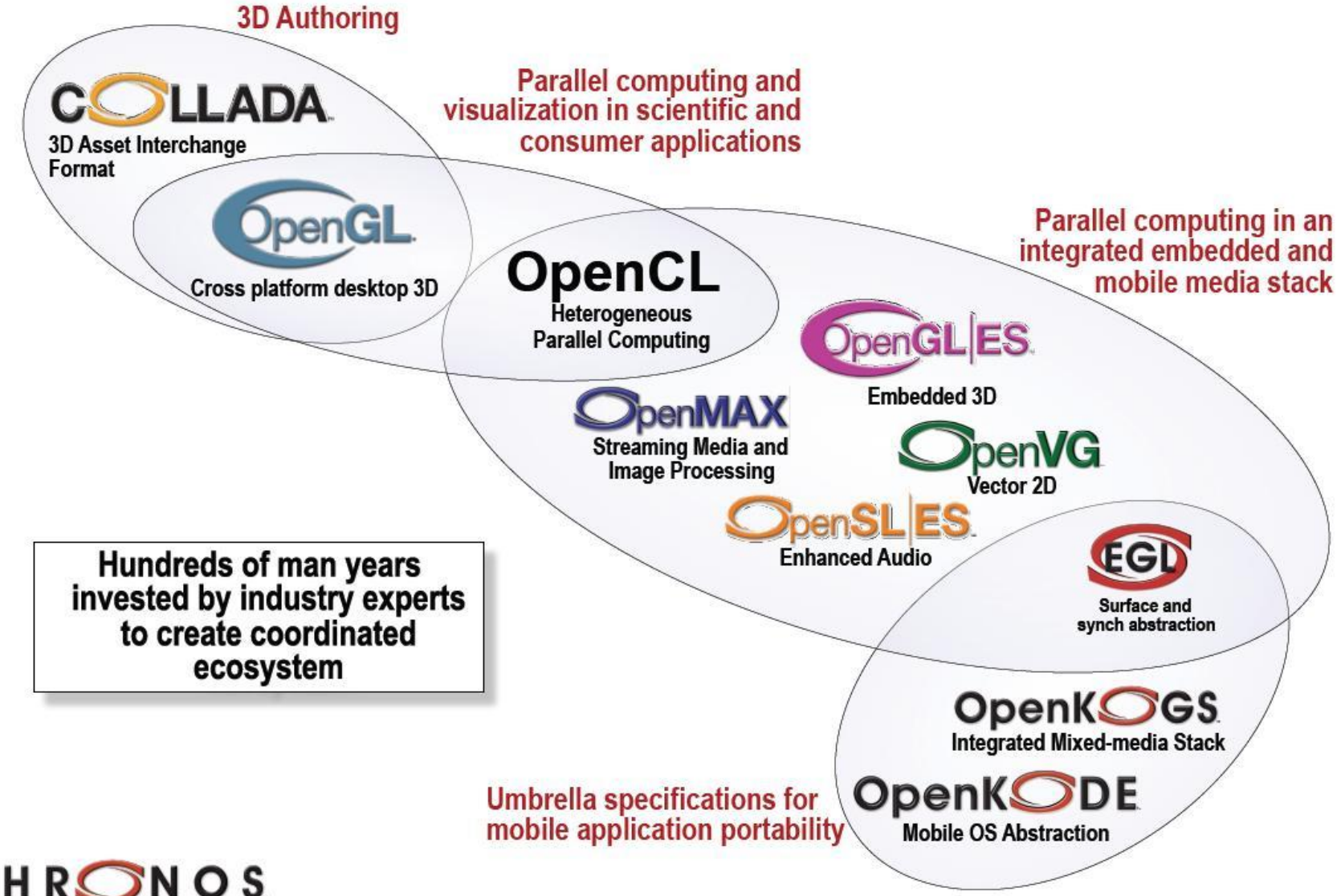


OpenCL

The Open Standard for Heterogeneous Parallel Programming

Neil Trevett
President, Khronos Group and OpenCL Chair
SIGGRAPH Asia, December 2008

OpenCL and the Khronos Ecosystem



KHRONOS GROUP

Restrictions

- Pointers to functions are not allowed
- Pointers to pointers allowed within a kernel, but not as an argument
- Bit-fields are not supported
- Variable length arrays and structures are not supported
- Recursion is not supported
- Writes to a pointer of types less than 32-bit are not supported
- Double types are not supported, but reserved
- 3D Image writes are not supported

- Some restrictions are addressed through extensions

OpenCL
Source Code Examples

Mark Harris
NVIDIA Developer Technology
December 2008

MS DirectX 11

MS DirectX11

- **Microsoft is not part of the Khronos consortium**
- **Microsoft is developing a competing technology called DirectX 11 Compute**

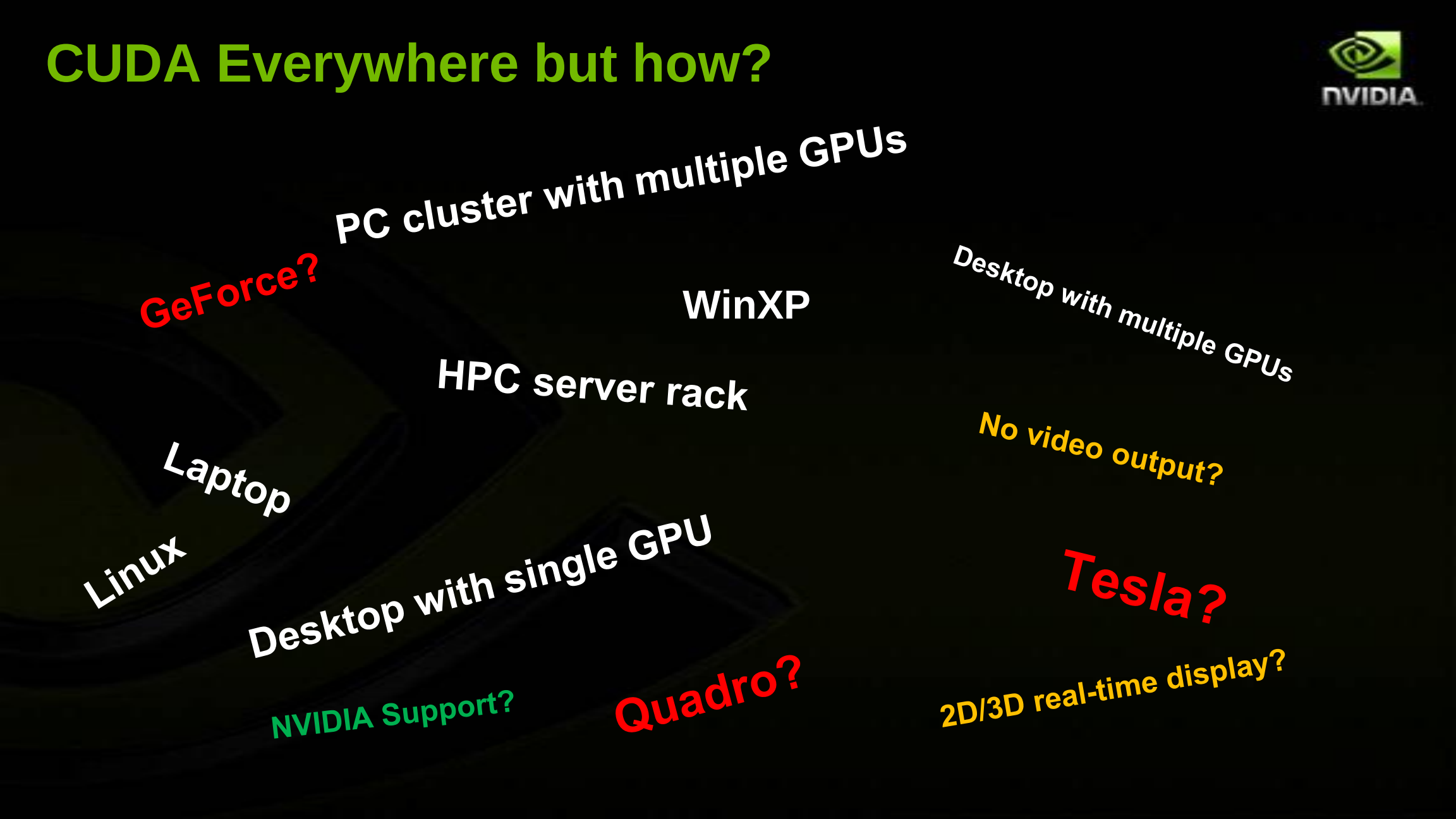
MS DirectX SDK November 2008 - Compute Shader

The Compute Shader is an additional stage independent of the Direct3D 11 pipeline that enables general purpose computing on the GPU. In addition to all shader features provided by the unified shader core, the Compute Shader also supports scattered reads and writes to resources through Unordered Access Views, a shared memory pool within a group of executing threads, synchronization primitives, atomic operators, and many other advanced data-parallel features

http://www.microsoft.com/downloads/details.aspx?familyid=5493F76A-6D37-478D-BA17-28B1CCA4865A&displaylang=en

#3 Deployment Products

CUDA Everywhere but how?

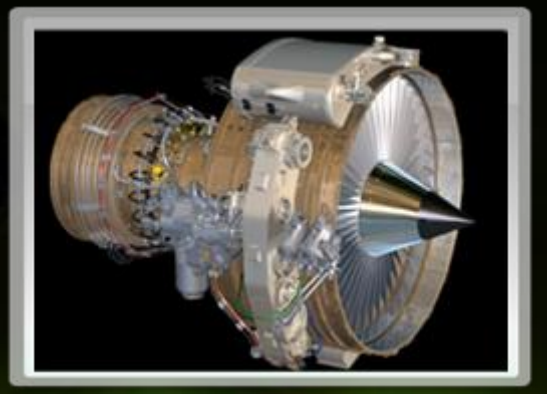


Parallel Computing on All GPUs

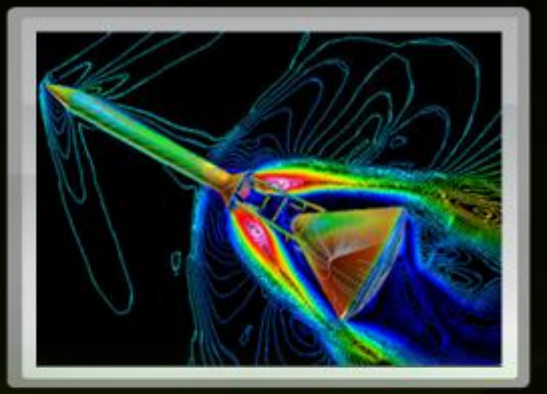
Over 90 Million CUDA Compatibles GPUs since Nov 2006

GeForce
Entertainment

Quadro
Design & Creation

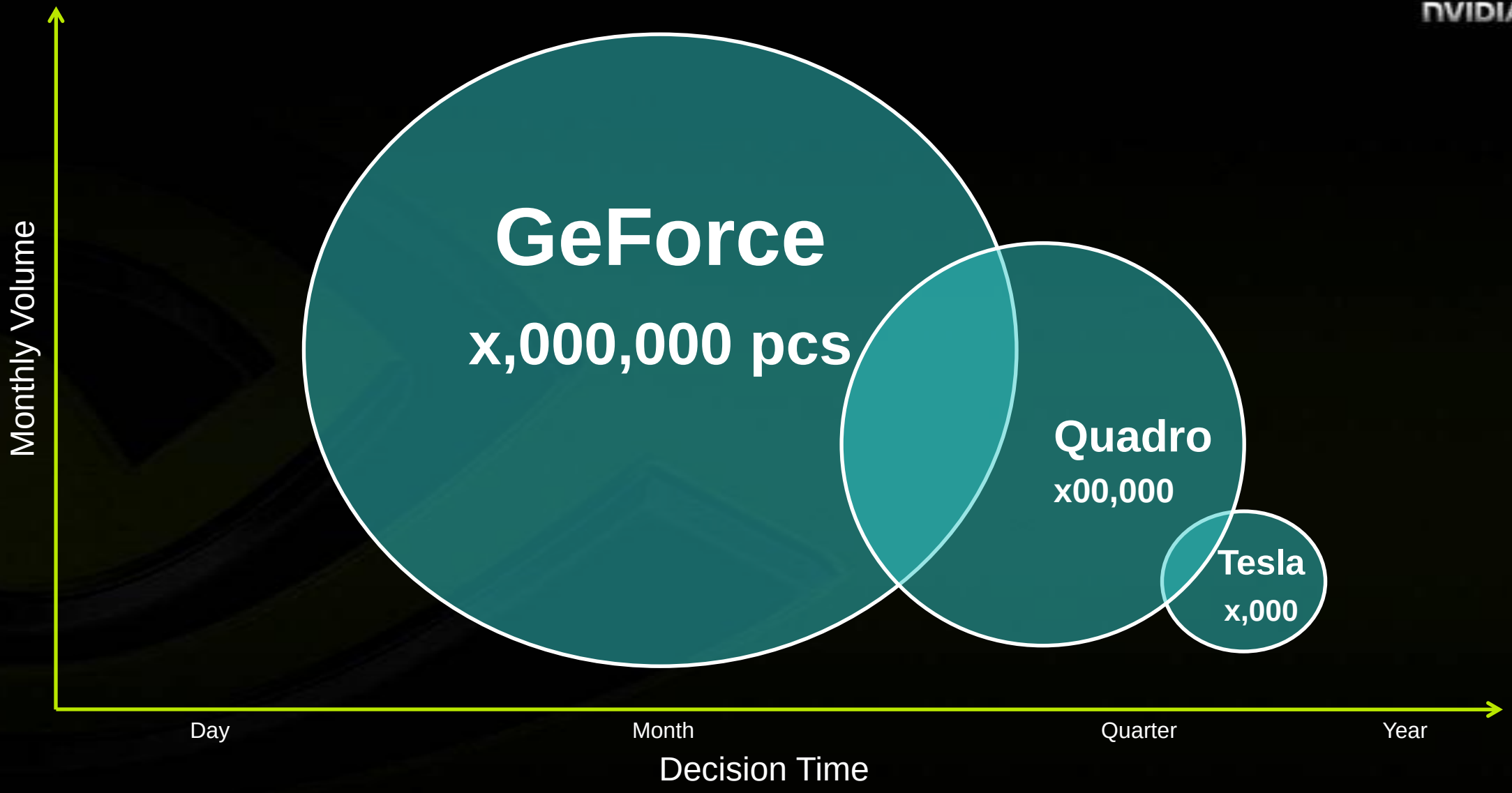


Tesla
High-Performance Computing



GPU

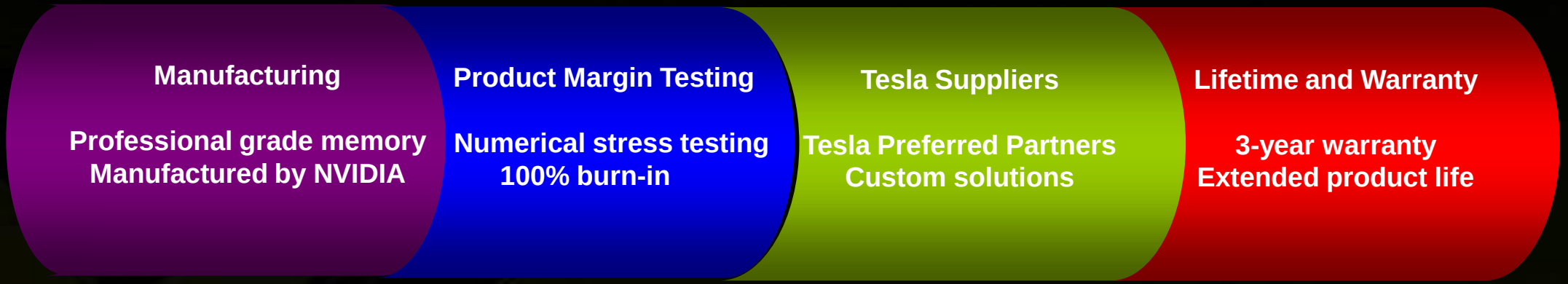
CUDA Compatible


Monthly Volume
GeForce
x,000,000 pcs
Quadro
x00,000
Tesla
x,000
Day
Month
Quarter
Year
Decision Time

Selecting a CUDA Platform

	Tesla	Quadro	GeForce
Stress tested and burned-in with added margin for numerical accuracy	X		
Manufactured by NVIDIA with professional grade memory	X	X	
NVIDIA care: 3-year warranty from NVIDIA, enterprise support	X	X	
4 Gigabyte on-board memory for large technical computing data sets	X	X	
Single card solution for professional visualization and CUDA computing		X	
Consumer middle-ware and applications: PhysX, Video, Imaging			X
Consumer product life cycle			X
Manufactured and guaranteed by NVIDIA graphics add-in card partners			X
Product support through NVIDIA graphics add-in card partners			X

NVIDIA Manufactured Computing Products



GeForce

CUDA for
consumer applications

CUDA for consumer applications

- **Over 80M GeForce** CUDA **compatible systems**
  - Widely available
  - CUDA works on desktop and laptops
  - **CUDA available for XP, Vista and MAC OS**
  - Single GPU and multi-GPU with SLI technology
- The right platform to develop consumer multimedia applications
- **Easy and fast access to CUDA programming model**
  - First step for university students to discover CUDA

badaboom !
Ultra-Fast GeForce Video Transcoding

- **Up to 19x faster than multi-core CPU**
- **Over 4 times faster than real-time**

CUDA is Taking Over Distributed Computing
Advancing Scientific Discovery

- Announcing 4 new distributed computing platforms on CUDA
  - SETI@home
  - BOINC Platform
  - GPUGRID
  - Einstein@home
- SETI@home client featured with CUDA
  - Up to 10x faster than CPU - No ATI option today
  - Available 12/17 from setiathome.berkeley.edu
- Press release on 12/17

CUDA : Pervasive in Scientific Research

BOINC

- Berkeley Open Infrastructure for Network Computing
- Foundation for many distributed compute projects
- *Now designed for CUDA*

GPUGRID

- Biomolecular simulations for scientific research
- *CUDA speeds up by average of 20x*

SETI@home

- Search for extra-terrestrial intelligence by tracking narrow-bandwidth radio signals from space
- *CUDA speeds up by as much as 10x*

Folding@home

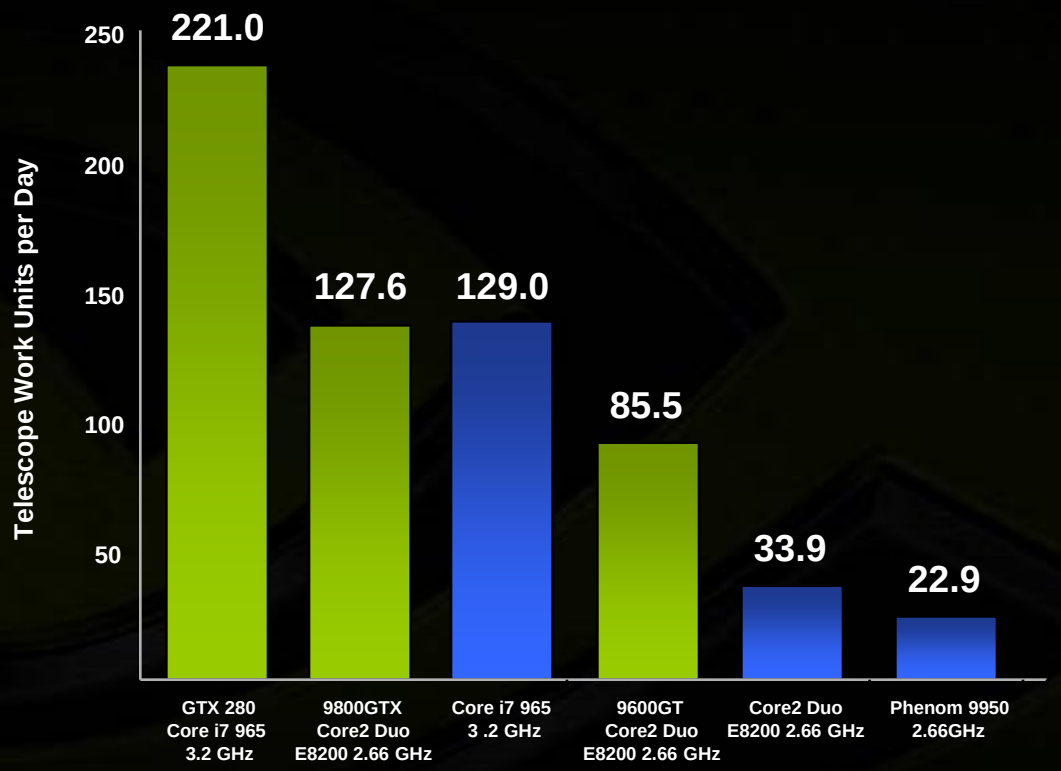
- Studying protein folding to better understand causes of diseases like Alzheimers and cancer
- *CUDA speeds up by as much as 10x*

Einstein@home

- Enhancing the search for gravitational radiation and discovery of pulsars
- *Optimizing for CUDA*

SETI@home

Powered by GeForce with CUDA



- 10x faster than an average CPU
- Nearly 2x speed of fastest home PC CPU
- Same performance at 1/10th the price *

$149

NVIDIA GeForce 9800GTX

$1,449

Intel Core i7 965 System

* Based on upgrade options for consumers who own a common Intel Core2 Duo E8200 based PC. Consumer may choose NVIDIA GeForce 9800GTX for $149 or must upgrade entire system to Intel Core i7 965, 3GB DDR3, x58 motherboard for total of $1,449. Prices based on NewEgg as of 12/15/09

Run SETI@home on your NVIDIA GPU - Windows Internet Explorer

http://setiathome.berkeley.edu/cuda.php

File Edit View Favorites Tools Help

Run SETI@home on your NVIDIA GPU

Run SETI@home on your NVIDIA GPU

HOME | PARTICIPER | A PROPOS | COMMUNAUTE | ACCOUNT | STATISTICS

Most computers are equipped with a **Graphics Processing Unit (GPU)** which handles their graphical output, including the 3-D animated graphics used in computer games. The computing power of GPUs has increased rapidly, and they are now often much faster than the computer's main processor, or CPU.

NVIDIA (a leading GPU manufacturer) has developed a system called CUDA that uses GPUs for scientific computing. With NVIDIA's assistance, we've developed a version of SETI@home that runs on NVIDIA GPUs using CUDA. **This version runs from 5X to 10X faster than the CPU-only version. We urge SETI@home participants to use it if possible.**

Just follow these instructions:

1) Check whether your computer has a CUDA-capable GPU

The CUDA version of SETI@home works on most newer NVIDIA GPUs. To find out if your GPU is compatible:

- Identify the model name of your GPU. On Windows, click on My Computer / Properties / Hardware / Device Manager, and open Display Adapters. This will show the model name.
- Check NVIDIA's list of CUDA-enabled products. If your GPU is listed here and has at least 256MB of RAM, it's compatible.

2) Download the latest NVIDIA driver

CUDA require a recent driver to work. If you alread have a recent NVIDIA driver, you can try going to step 3. If not, or if you have problems, then you should get the CUDA 2.0 driver. Download it from NVIDIA and install it (a reboot will be required). Note: you only need the driver, not the Toolkit or the SDK. Get the CUDA 2.0 version, or the latest release version. Using beta drivers is not guaranteed to work. **If BOINC (in the advanced view messages tab) reports that it has found a CUDA GPU, your current driver set may be recent enough to work properly.**

3) Install the latest BOINC software

You'll need version 6.4.4 or later of the BOINC software. Download it from here and install it.

You're done! Now start up BOINC, and before long you'll be finishing jobs in no time, and racking up big credit numbers.

Also:

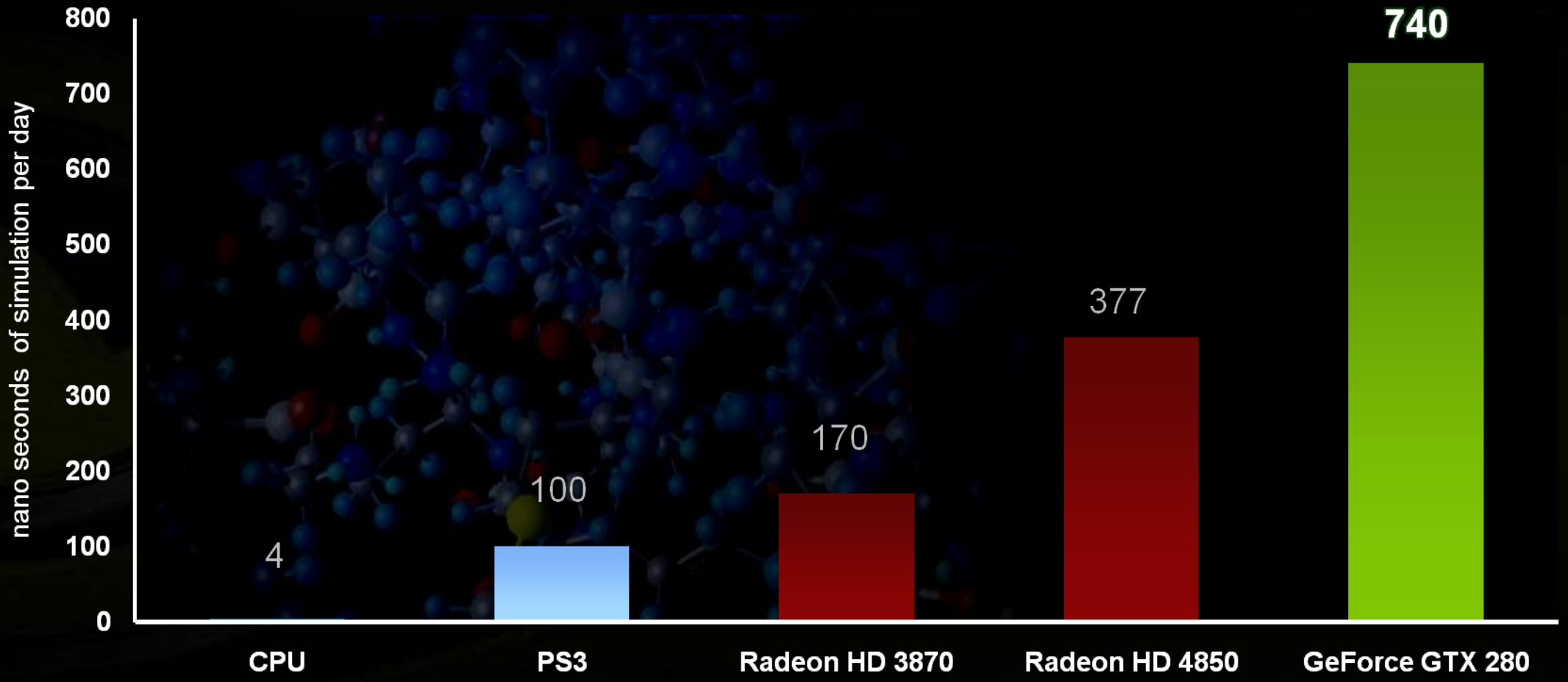
- You can use your GPU to help biomedical research as well as SETI: GPUgrid.net, also has a CUDA application (and more projects are on the way).

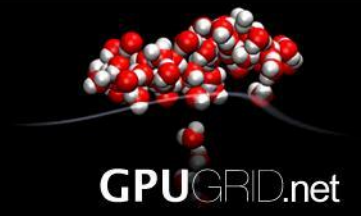
Internet | Protected Mode: Off

100%

Folding@home

Powered by GeForce with CUDA





GPUGRID.net

NVIDIA

Home | Forum | Join | Science | Gallery | Resources | News | About us | Your account | Server status | Stats

Flavour: Zen

What is it?

A volunteer distributed computing project

It is a novel distributed supercomputing infrastructure made of many **PlayStation3** and **NVIDIA graphics cards** joined together to deliver high-performance all-atom biomolecular simulations. This project gives a new **powerful computational tool** to scientists and you are an important part of it.

Be part of it

If you enjoy science, **you can participate** by donating computing time to scientific research. Simply follow the instructions below to start, gain your credits for the results you return, join a team, meet and exchange experiences with other participants in the forums.

Want to know more?

Visit our Science page and find out how PlayStation3 and NVIDIA graphics cards can help biomedical research.
And visit our Gallery to know more about us and what we do through our pictures and videos.

Spotlight

Visit this thread in the forums for full details about this machine.

Join us!

Click on your system and follow the instructions

PlayStation3 | NVIDIA Graphics Card

News

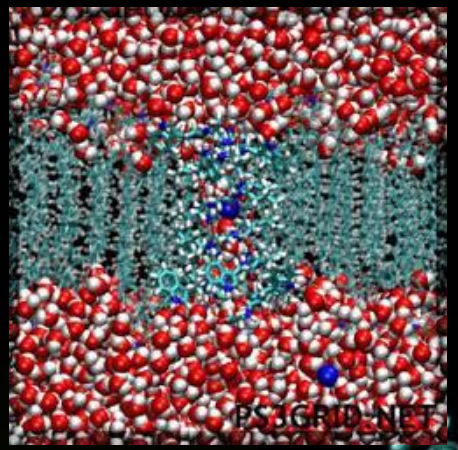
Suggested BOINC version is now 6.4.5
December 11, 2008
Please upgrade to version 6.4.5 to have a correct estimation of elapsed time.
GPUGRID t-shirt graphics and brochures available from website.
November 27, 2008
We have created a new resource section which contains project brochures and t-shirt graphics and divulgation material.
PS3GRID and GPUGRID new website in two styles selectable by users.
November 12, 2008
We have uploaded the new website with improved design and usability. Users can also choose between black (geek) and white (zen) styles. Try it out.
...more news

Subscribe to the RSS feed XML

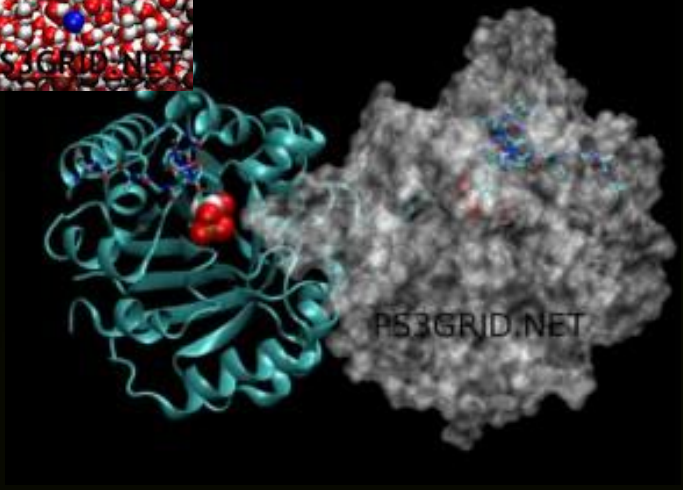
User of the day

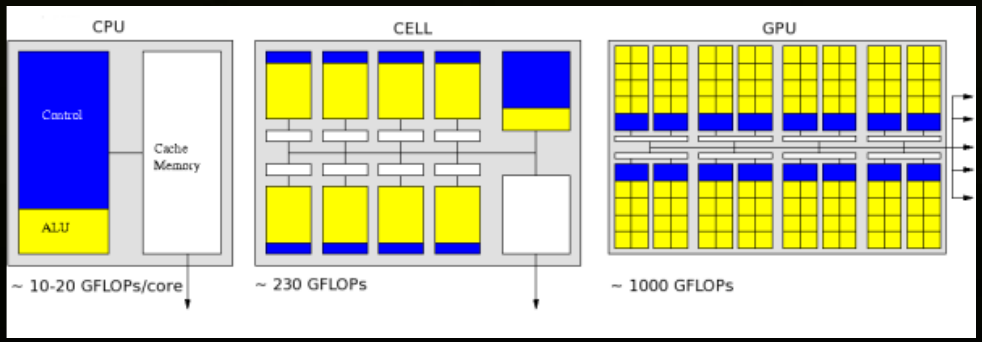
Czech Crunchers Unit

Returning participants



NVIDIA





Quadro FX

CUDA for
professional visualization applications

NVIDIA Quadro

NVIDIA Quadro
FX 5600
G80GL 1.5GB

NVIDIA Quadro
FX 5800
GT200GL 4GB

NVIDIA Quadro
FX 4600
G80GL 768MB

NVIDIA Quadro
FX 4800
GT200GL 1.5GB

NVIDIA Quadro FX 3700
G92GL 512MB

NVIDIA Quadro
FX 1700

NVIDIA Quadro
FX 570

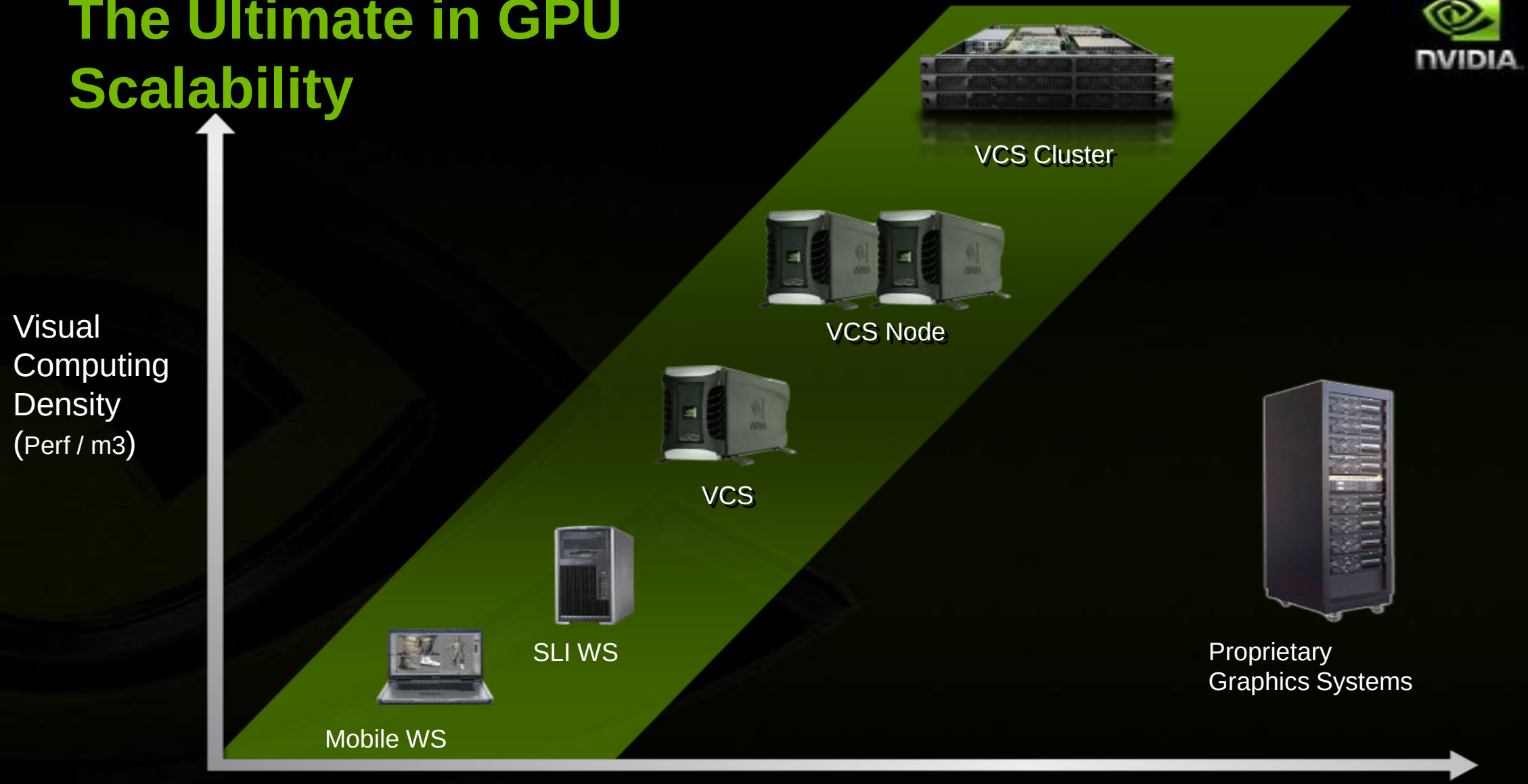
NVIDIA Quadro
FX 370

NVIDIA Quadro
NVS 290

Pro 3D

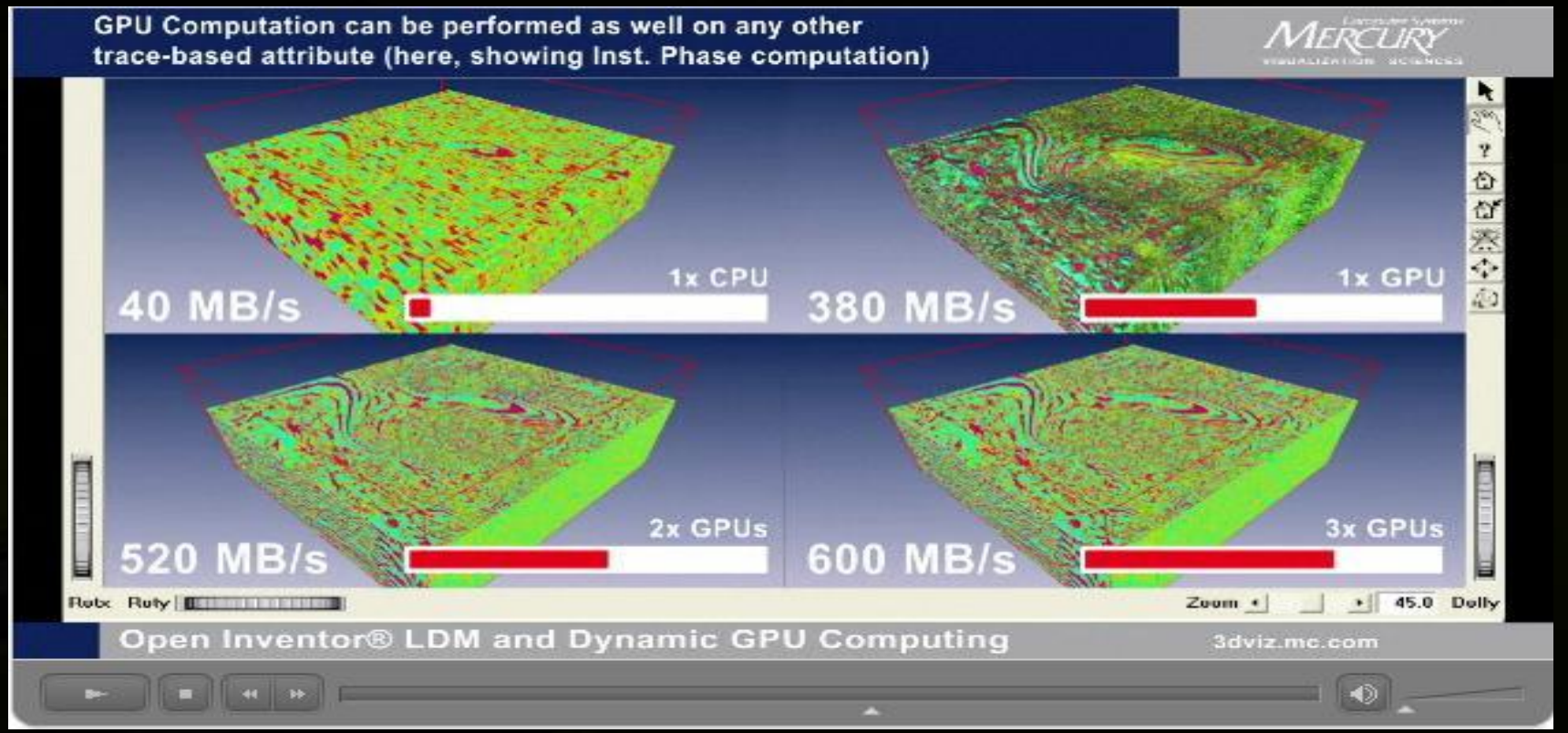
Pro 2D

The Ultimate in GPU Scalability



Mercury Computers – Oil & Gas

Quadro FX used for data and graphics processing



Tesla

CUDA for
HPC solutions

TESLA™

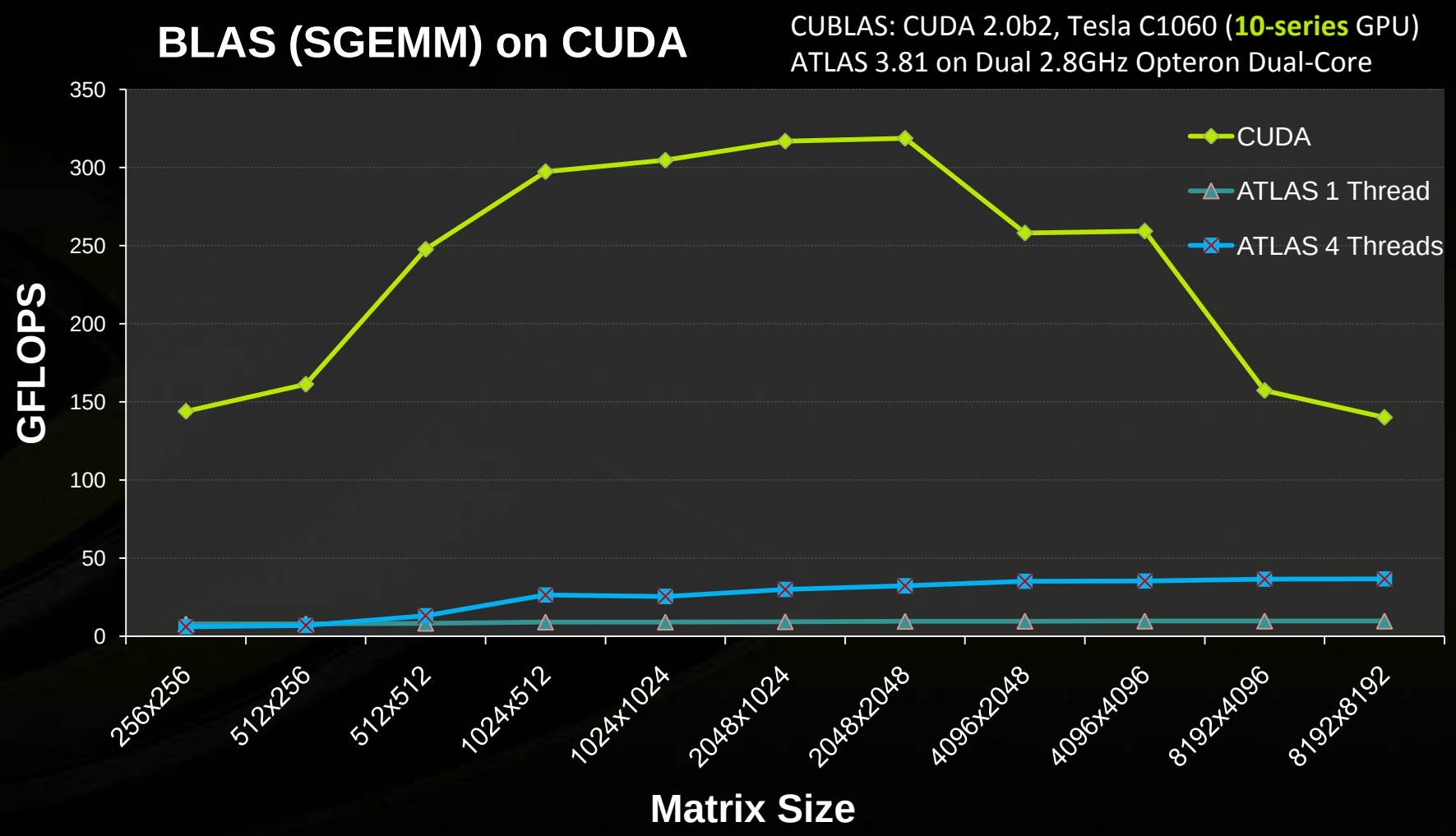
C1060
Card

S1070
1U Compute System

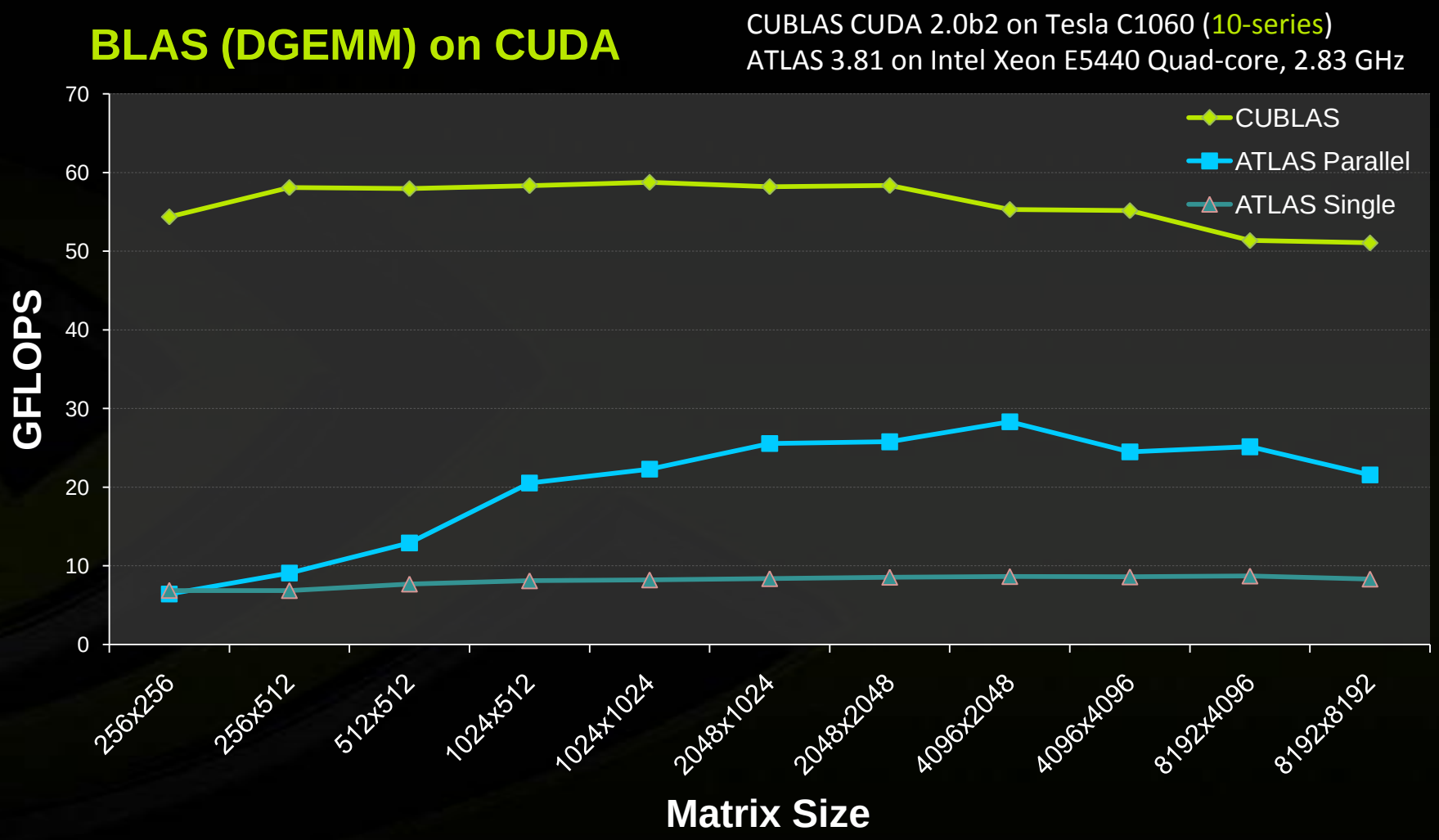
Double Precision Floating Point

	NVIDIA Tesla T10	x86 (SSE4)	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	All 4 IEEE, round to nearest, zero, inf, -inf
Denormal handling	**Full speed**	**Supported, costs 1000's of cycles**	Supported only for results, not input operands (input denormals flushed-to-zero)
NaN support	**Yes**	Yes	Yes
Overflow and Infinity support	**Yes**	Yes	Yes
Flags	**No**	Yes	Yes
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 + step**
Reciprocal sqrt estimate accuracy	**23 bit**	**12 bit**	**12 bit + step**
log2(x) and 2^x estimates accuracy	**23 bit**	**No**	**No**

Single Precision BLAS: CPU vs GPU



Double Precision BLAS: CPU vs GPU

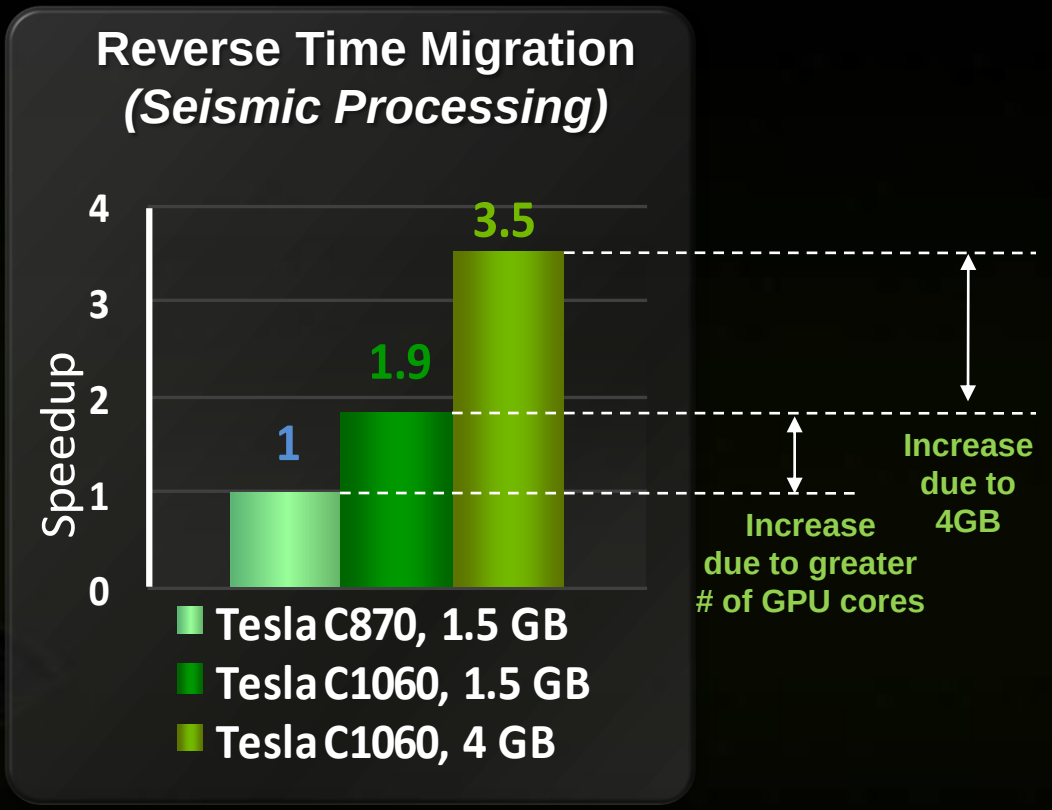


Tesla C1060 Computing Card

Processor	1 x Tesla T10
Number of cores	240
Core Clock	1.29 GHz
On-board memory	4.0 GB
Memory bandwidth	102 GB/sec peak
Memory I/O	512-bit, 800MHz GDDR3
Form factor	Full ATX: 4.736” x 10.5” Dual slot wide
System I/O	PCIe x16 Gen2
Power	200 W maximum 160 W typical (5.83 GFlops/Watt) 25 W idle

Impact of 4GB Memory on Performance

- 4GB of memory is critical for best CUDA performance
- Enables processing on larger data sets
  - Solve larger problems
- Double precision applications require more memory



Impact of 4 GB memory


NVIDIA® TESLA™
PERSONAL SUPERCOMPUTER

Globally Researchers are building GPU-based Workstations

3 GPUs

YONSEI UNIVERSITY

Korea

2 GPUs

University of Cambridge, UK

THEORETICAL and COMPUTATIONAL BIOPHYSICS GROUP

NIH RESOURCE FOR MACROMOLECULAR MODELING AND BIOINFORMATICS

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

3 GPUs

University of Illinois

MIT Graduates Build 16-GPU Monster

POST FROM UBERGIZMO ON 28 JULY 2008 01:08:38 PM. © UBERGIZMO

16 GPUs

The Visual Neuroscience Group

@ The Rowland Institute at Harvard

8 GPUs

FASTRA

GPU SuperPC

University of Antwerp, Belgium

Cluster vs Workstation : Trade-offs

Performance Gap

Cluster

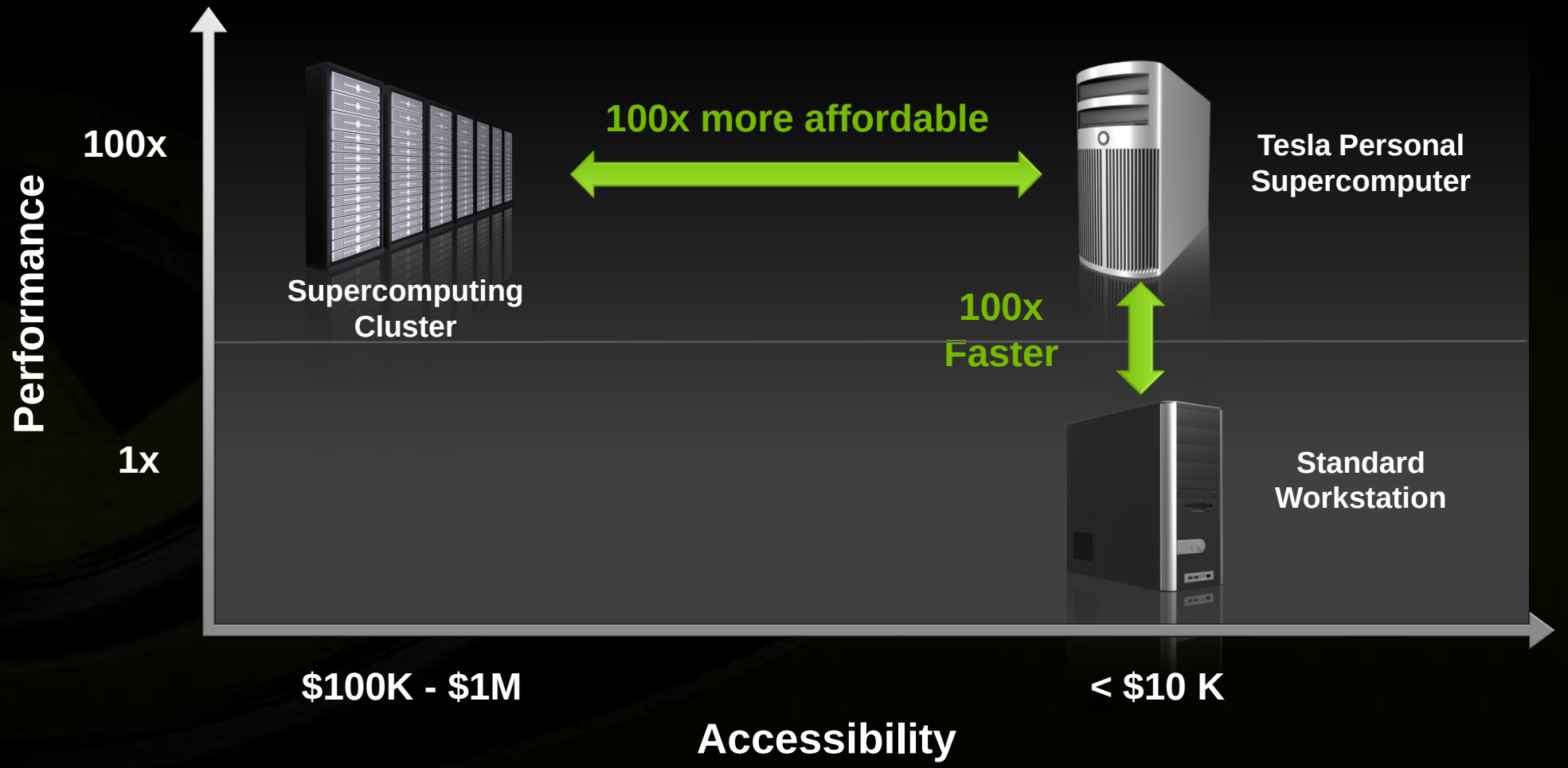
Workstation

Need server room; shared resource	→	Dedicated resource for each person
Energy hungry: to power up & cool	→	No system management/IT required
Expensive to build/maintain; need IT dept	→	Built for the office: cool, quiet, & compact

The Tesla Visual Supercomputer

Return of the Scientific Workstation

- 4 TeraFlops Workstation
  - 4 CUDA GPUs
  - 960 cores
  - 16 GB fast GPU memory
- Specs:
  - Quad-core CPU (1P or 2P)
  - 16 GB System memory
  - 4 Tesla/Quadro GPUs
- Optimized for scientific computing
- The power of a cluster in a workstation


Performance
100x
1x
Supercomputing Cluster
100x more affordable
Tesla Personal Supercomputer
100x Faster
Standard Workstation
$100K - $1M
< $10 K
Accessibility

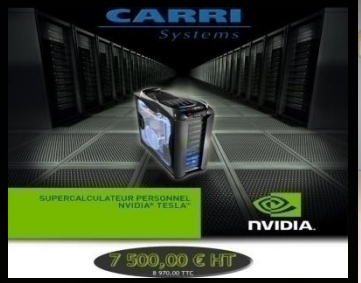
Personal Super Computers

Transtec
Germany

Comptronic
Germany

Fluidyna
Germany

CAD2
UK



Carri
France

Sprinx
Czech

Axxiv
Switzerland

NEXT
Italy

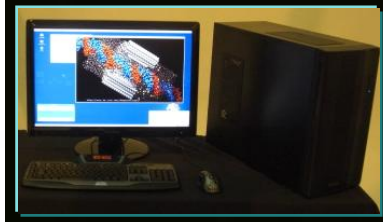
Exon
Italy

Armari
UK

Concordia Graphics
Italy

E4 Italy

Azken Muga
Spain



Viglen
UK

La puissance d'un SUPERCALCULATEUR à portée de main

Votre activité requiert des calculs mathématiques complexes et exigeants ?

Vous avez besoin de plus de puissance et de rapidité ?

Vous recherchez une solution dédiée pour développer vos propres codes de calcul ?

4340€ HT seulement !

Station HP xw8600 et carte NVIDIA® Tesla™ C1060

Offre spéciale de fin d'année

Un supercalculateur dédié à vos calculs sur votre bureau

Bénéficiez d'une solution de **supercalcul parallèle,** et intégrez dans une seule station de travail les performances traditionnellement réservées aux clusters d'entreprise :

- **Puissance brute : 1 Teraflop***
- **240 coeurs**
- **4 GB de mémoire**

Vous bénéficiez ainsi d'une ressource de bureau dédiée **bien plus rapide et écoénergétique qu'un cluster partagé** dans un centre de calcul.

* Possibilité d'ajouter une deuxième carte pour atteindre 1,9 Teraflops.

Flexibilité, puissance,

et écoénergies

pour seulement

4340€ HT !

Ecran 24" Performance

HP LP2475W en option

pour 420€ HT.

Spécifications techniques

Station de travail HP xw8600
1050W 80 Energy Efficient. **Linux** Installer Kit Software. **2 x Intel Xeon 5450 3.00 12M. QC. 8GB de mémoire.** Carte graphique NVIDIA Quadro NVS 290. **2 x 146GB SAS 3Gb/s 15K tours.** DVD /-RW SuperMulti SATA. PCI Express 16x.

Carte NVIDIA Tesla C1060 avec processeur Tesla T10P massivement parallèle et multi-coeurs, couplée au standard de programmation CUDA C afin de simplifier la programmation multi-coeur. Applicatifs supportés.

Ref: 74032246

NVIDIA

CV54 – Dual Socket Visualization Node

Integrates Tesla or Quadro FX boards

Tesla S1070 1U System

Processors	**4 x Tesla T10**
Number of cores	**960**
Core Clock	**1.5 GHz**
Performance	**4 Teraflops**
Total system memory	**16.0 GB** (4.0 GB per T10P)
Memory bandwidth	**408 GB/sec peak** (102 GB/sec per T10P)
Memory I/O	**2048-bit 800MHz GDDR3** (512-bit per T10P)
Form factor	**1U** (EIA 19" rack)
System I/O	**2 PCIe x16 Gen2**
Typical power	**700 W**

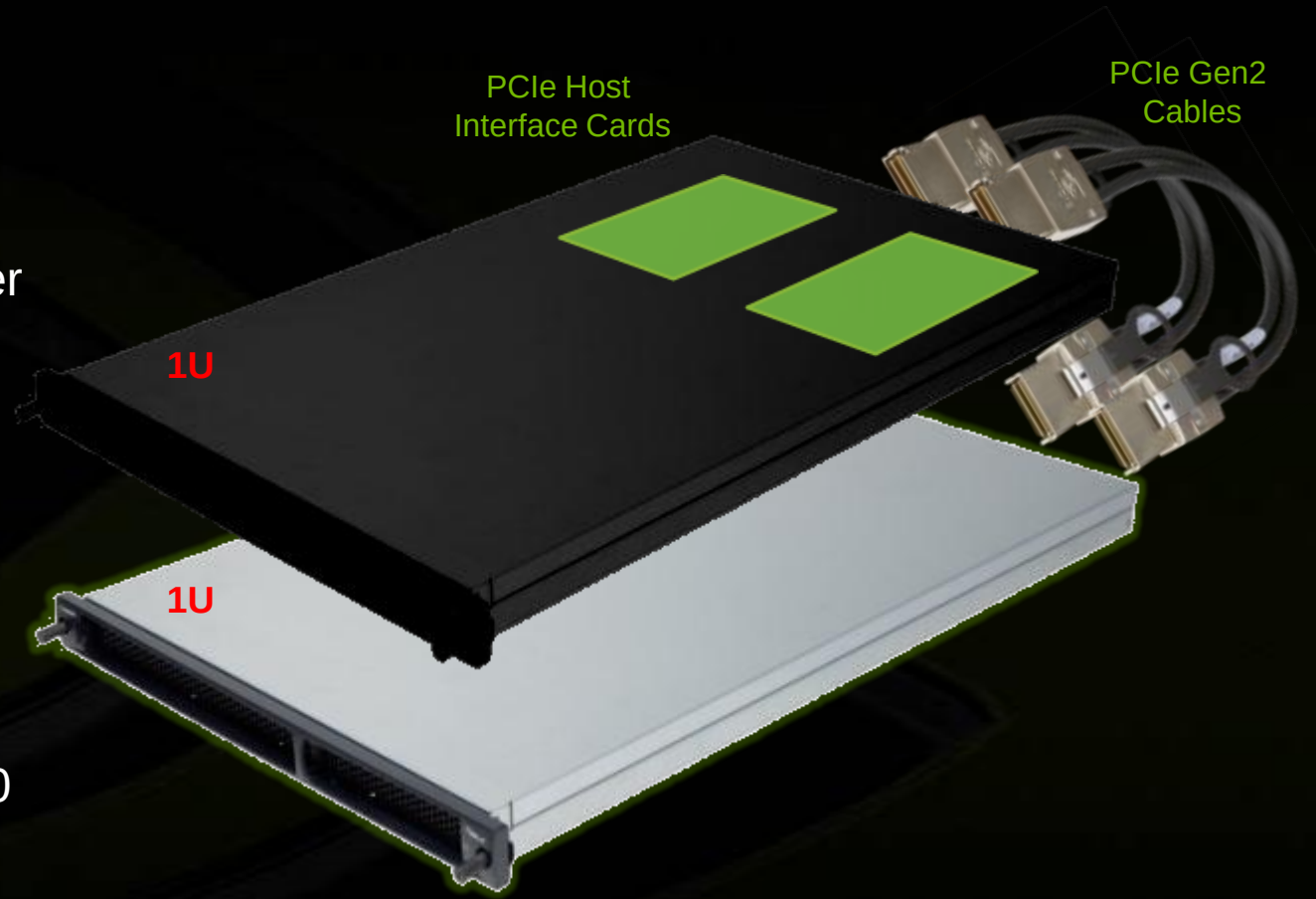
Connection to host system(s) using two PCIe interface cards

NVIDIA Tesla S1070 SKUs

- **Tesla S1070-400** **<u>* Standard SKU *</u>**
  - 1.296 GHz GPU clock
  - Peak performance
    - 32-bitt 3.73 TF (933 GF per GPU)
    - 64-bit 310 GF (77.7 GF per GPU)

- **Tesla S1070-500**
  - 1.44 GHz GPU clock
  - Limited availability
  - Peak performance
    - 32-bit 4.14 TF (1.03 TF per GPU)
    - 64-bit 345 GF (86.4 GF per GPU)

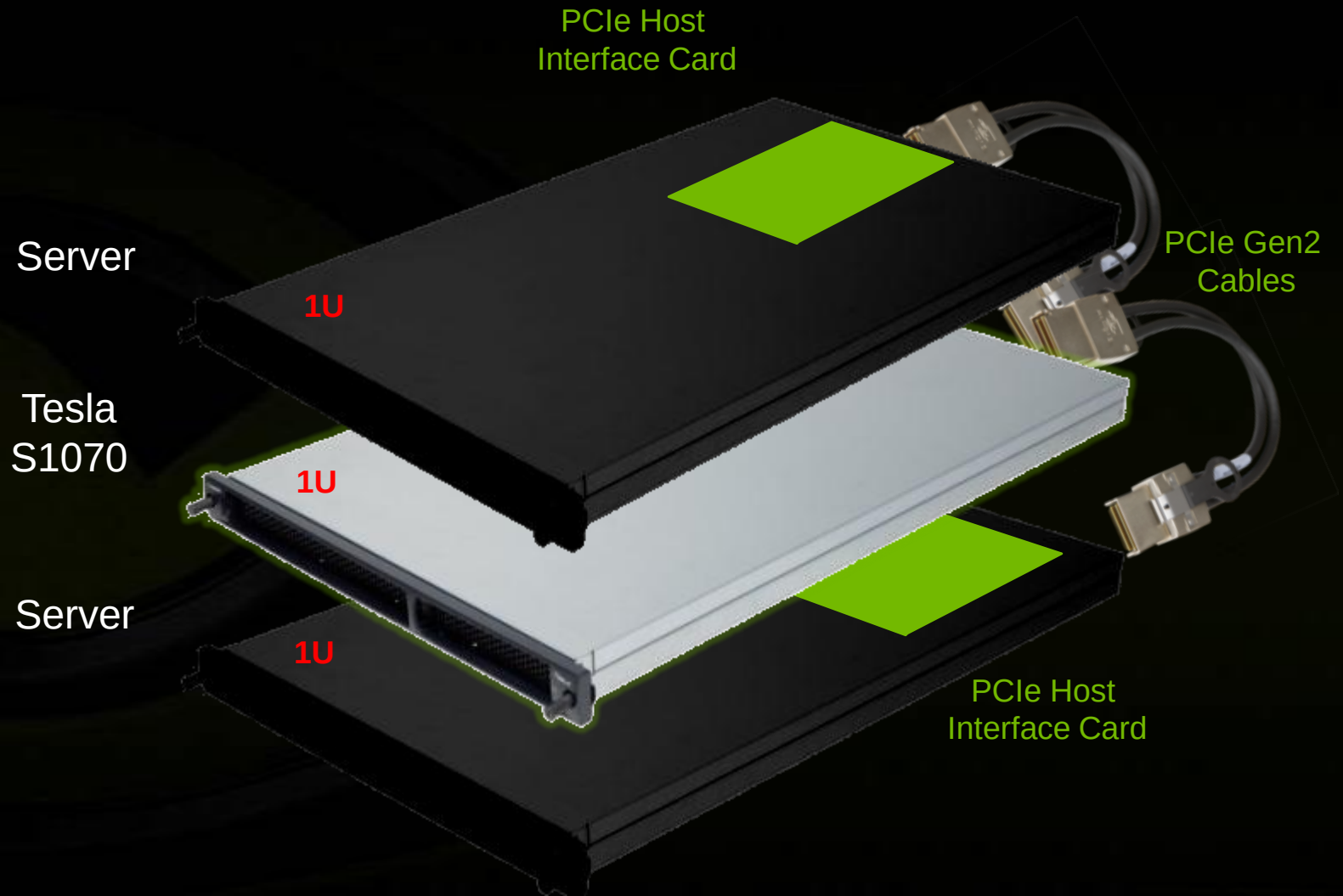
Tesla S1070: 2U Sample Configuration



Two PCIe Gen2 Cables
(50 cm or 2 m length)

Two PCIe Gen2 Host
Interface Cards

Tesla S1070: 3U Sample Configuration



Two PCIe Gen2 Cables
(50 cm or 2 m length)

Two PCIe Gen2 Host
Interface Cards

Tesla S1070 OEM Partners

- HP
- Dell
- SUN
- SGI
- Bull
- Lenovo
- IBM
- FSC
- Supermicro

Scalable Professional Development Platforms

Laptop / Desktop
single GPU

Single User

Discover GPU Computing

- Easy entry path but limited performance and memory size
- Limited dataset size

1 to 3K €

Supercomputing PC
multiple Tesla C1060

Single User

Development & Prototyping

- Multi-GPU performance scaling
- 1TFlops and 4GB per GPU
- Larger dataset

3 to 8K €

Hybrid Cluster
Tesla S1070

Multiple Users

Dev., Prototyping & Production

- 4 TFlops per 1U Tesla
- 16GB per 1U Tesla
- Up to TB datasets

8K € per 2U (CPU+GPU)

CUDA Compatible

French Atomic Energy Commission

295 TFlops Hybrid Cluster

- The new Bull NovaScale supercomputer consists of a cluster of 1,068 Intel Nehalem nodes, delivering some 103 TFlops, and 192 NVIDIA Tesla GPU nodes, providing additional power of up to 192 TFlops

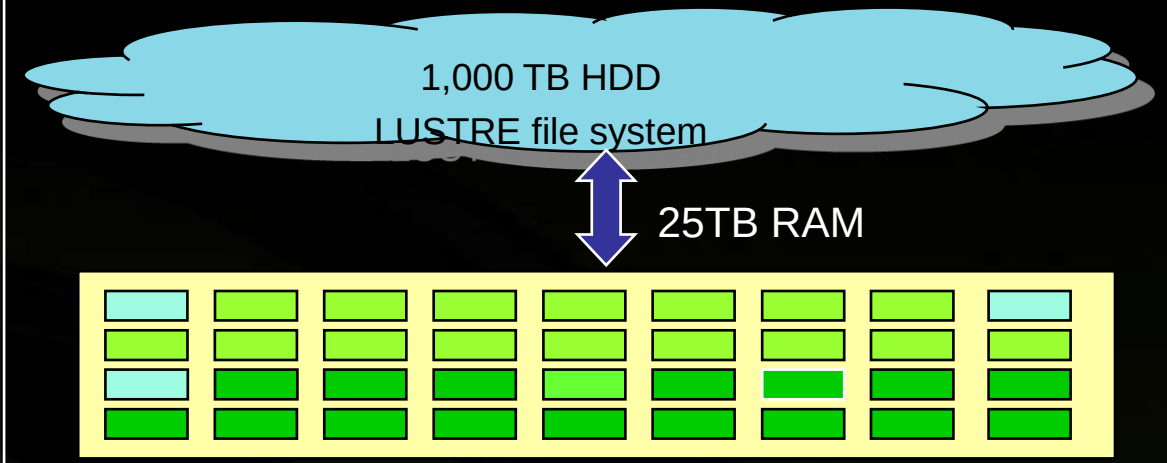
48 Tesla S1070 1U servers

= 192 GPUs

= 768GB

http://www.cea.fr/english_portal/news_list/bull_novascale_supercomputer_genci_and_the_cea

CEA Hybrid Cluster



3 x 42U rack for GPU
192 TFlops Peak

17 x 42U Bull NovaScale
103 TFlops Peak

Innovative GPU Platform

Allow specific applications to use multiple CPU-GPU cores and the entire system RAM

SMP Production Platform

- 3GB RAM per CPU core
- For all industrial and research applications

Over 295 TFlops
#1 in Europe

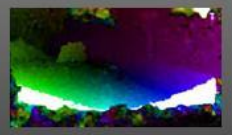
- 42U rack with 16 pcs Tesla S1070
- 42U rack with 60 pcs Intel Nehalem CPU. 3GB RAM/core
- HDD, network, service nodes
- Infiniband DDR

Open Source system software

Some Tesla S1070 Customers

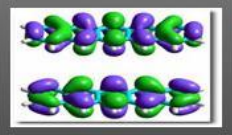
- Tokyo Technical Institute — 170 pcs S1070, #27 in TOP 500 Nov08
- Max Plank Institute
- Univ. Francfort, Cardiff, Reims…
- CEA CCRT
- CINES
- EADS
- TOTAL
- BNPParisbas


NVIDIA.
Applications



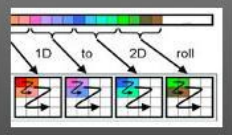
Lucas and Kanade optical flow algorithm using CUDA (LKCUDA)

55 x



Computational Chemistry Using GPUs

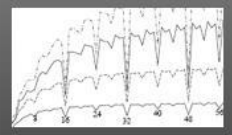
4.3 x



Concurrent Number Cruncher

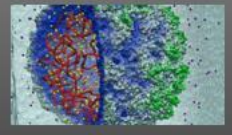
10 x

GPU4Vision

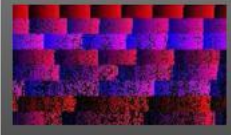


Efficient Computation of Sum Products on GPUs

270 x

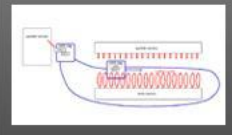


Accelerating Molecular Modeling with GPUs

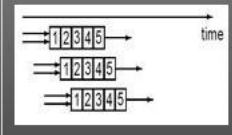


Cmatch: Fast Exact String Matching on the GPU

35 x

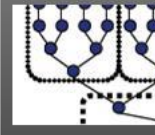


The Chamomile Scheme: N-body Simulations



LIBOR Interest rate Model

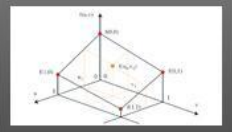
50 x



Fast Support Vector Training and Classifi

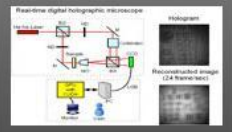
Ray tracing with CUDA (CUDART-sp)

25 x

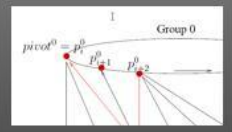


Teraflops for Games and Derivatives Pricing

50 x



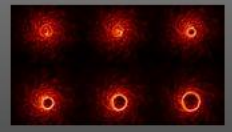
Real-time Digital Holographic Microscopy



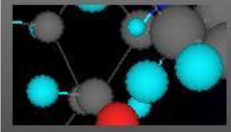
Wait-free Programming for Computations on Graphics Processors

Real-time Visual Tracker by Stream Processing

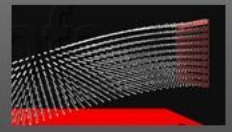
10 x



Visualization of Meshless Simulations Using Fourier Volume Rendering

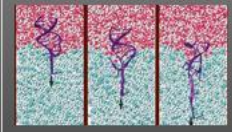


Two-electron Integral Evaluation

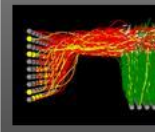


Simulation Open Framework Architecture (SOFA)

55 x



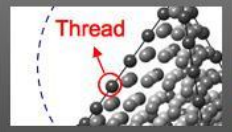
Scalable Molecular Dynamics: NAMD



Synthesis of Artificial Circuitry

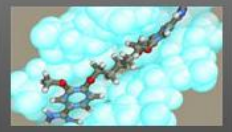
Computational Fluid Dynamics (CFD) using GPUs

17 x



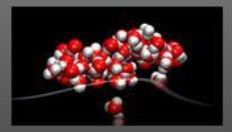
MIDG

50 x

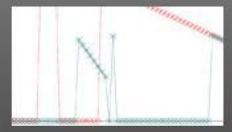


Molecular Dynamics of DNA and Liquids

18 x

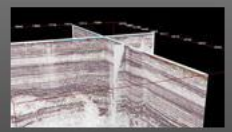


GPUGRID.NET



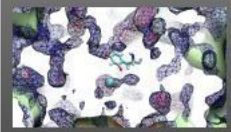
Mixed Precision Linear Solvers

27 x



Prestack Seismic Data Interaction

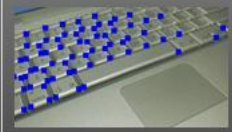
100 x



Visual Molecular Dynamics: VMD

100 x

N-body Simulations in CUDA

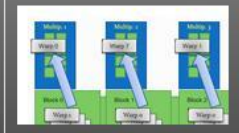


OpenVIDIA: Parallel GPU Computer Vision

SnapCT: Tomograph Reconstruction

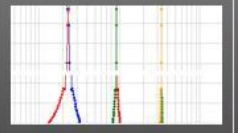
Tomographic Reconstruction

40 x



Obsidian: GPU Programming in Haskell

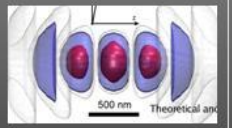
Volume Ray Casting With CUDA



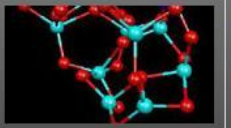
Audio FIR Crossover

35 x

CUDA vs Wizard

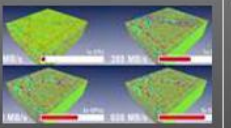


Fluorescent Microscopy



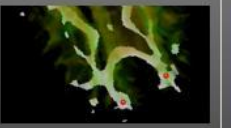
Quantum Mechanical Calculations of Molecular Properties

4 x



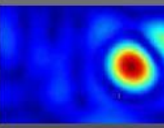
Innovative 3D visualization solutions for Oil and Gas

10 x



Interactive Visualization of Volumetric White Matter Connectivity

100 x



Real Time Capture of Au Images and Use with Vid

LINZIK: The compact optical CAD

10 x

Fast MRI Gridding on GPUs via CUDA

Accelerating Statistical Static Timing Analysis

260 x

Histogram Computation with CUDA

Dense Matrix-Vector Multiplication

32 x

High Performance Pattern Recognition on GPU

100 x

AstroGPU 2007 Workshop

NaminamiFX for Fluid Simulation

4 x

Improved Magnetic Resonance Imaging (MRI) Quality

Speeding Up Mutual Information Computatio Hardware

Cost-effective Medical Image Reconstruction

Solving Dense Linear Systems on Multi-Accelerator

Canny Edge Detection

Multiple Relatively Robust Representations (MRRR)

Fast k Nearest Neighbor Search using GPU

Fast GPU-Based CT Reconstruction

AES Crytography Acceleration

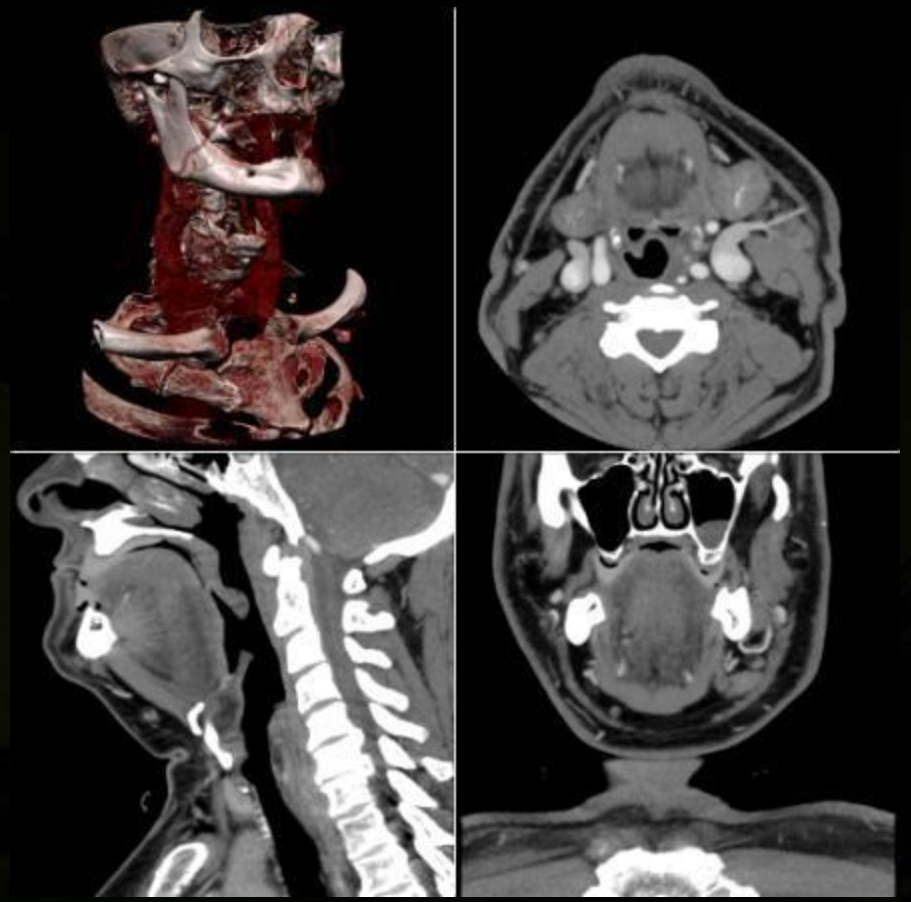
Flocking-based Document Clustering on the GPU

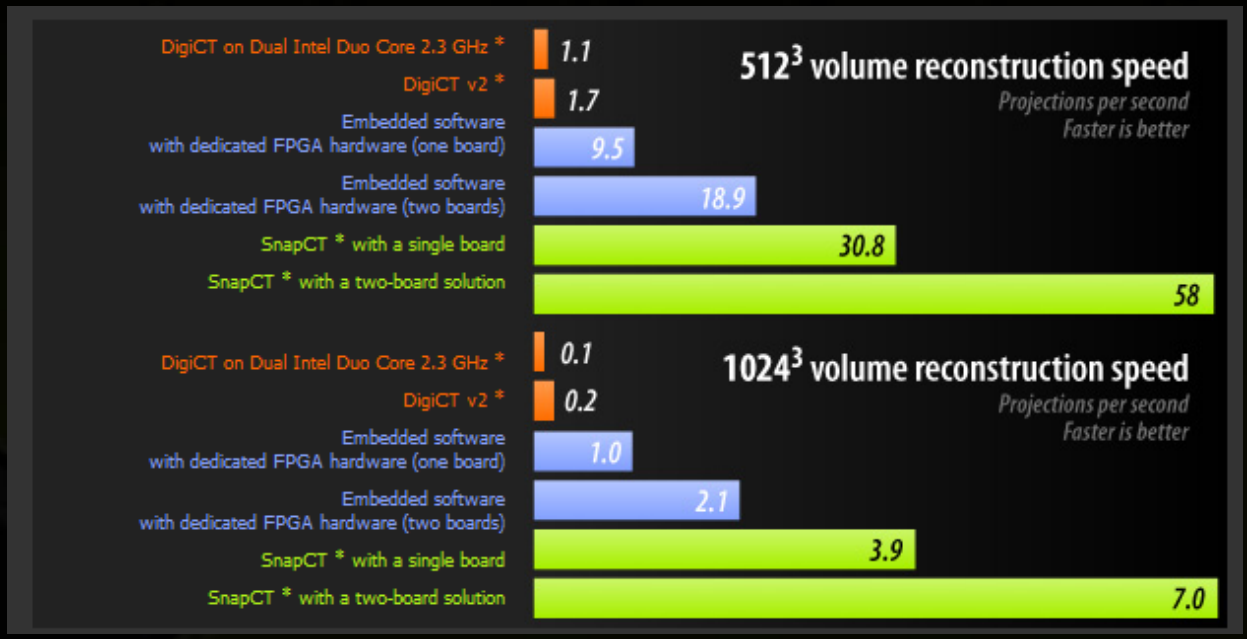
Fast Multipole Methods on Graphics Processors

Quantitative Risk Analysi Algorithmic Trading Syst

Accelerating Time to Discovery

CT Image Reconstruction



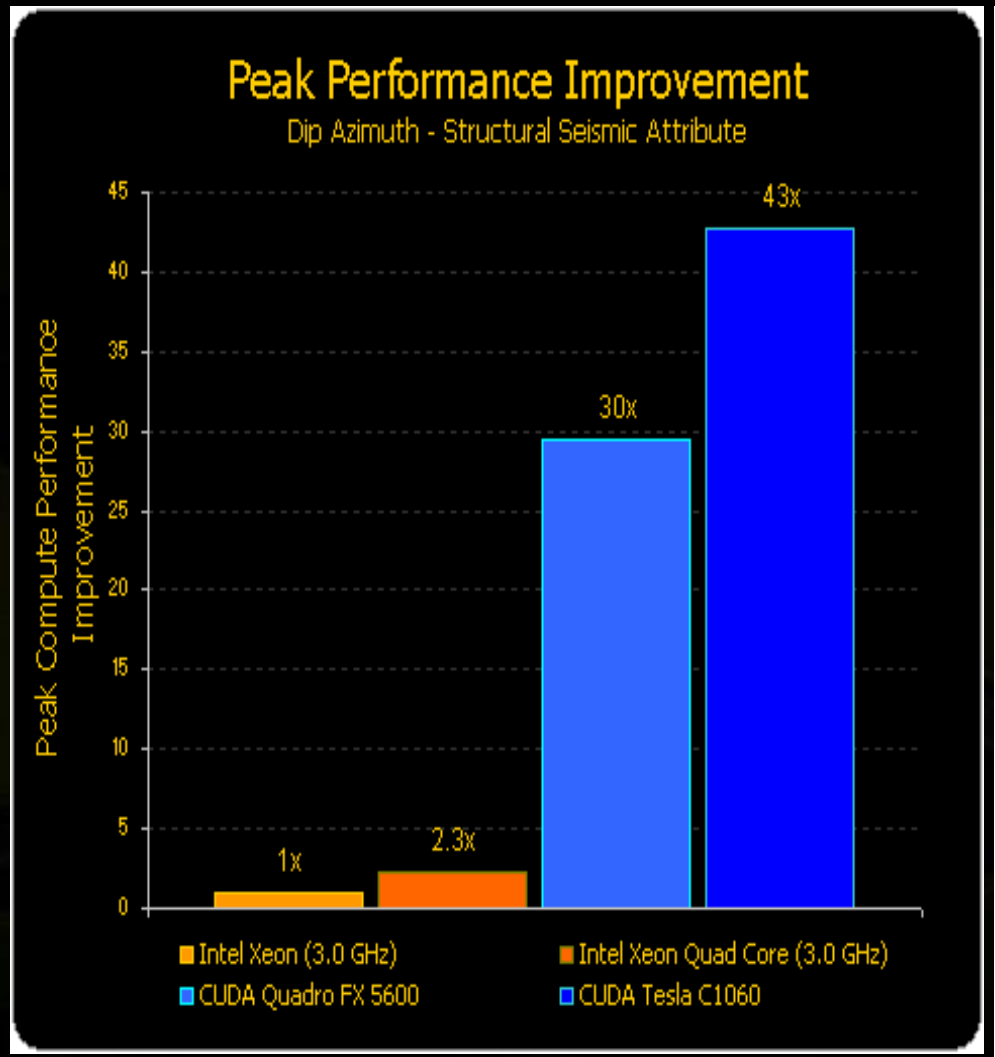


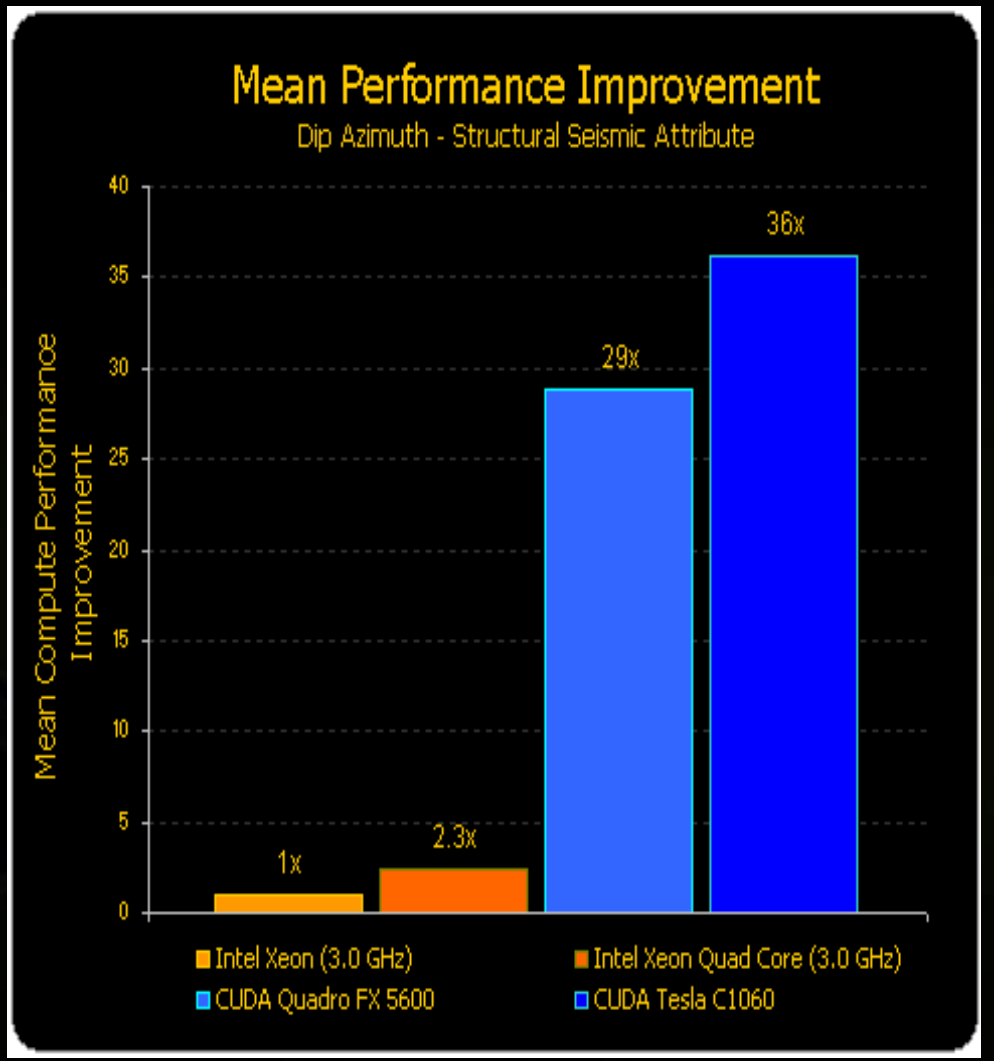
www.digisens.fr

Demo

ffA – Initial Performance Metrics

www.ffa.co.uk





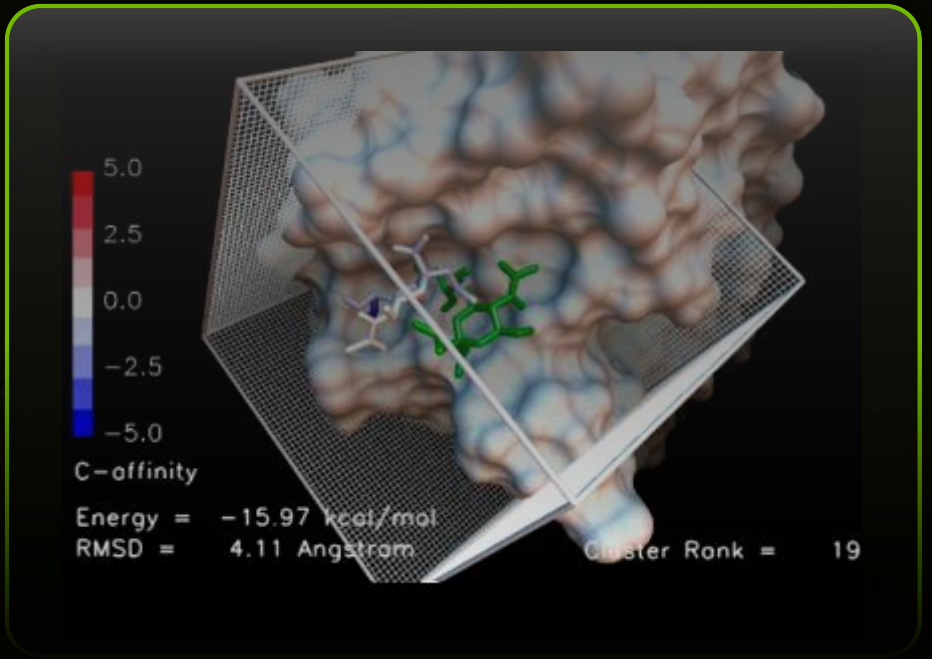
Autodock for Cancer Research

National Cancer Institute reports 12x speedup

Wait for results reduced from 2 hours to 10 minutes

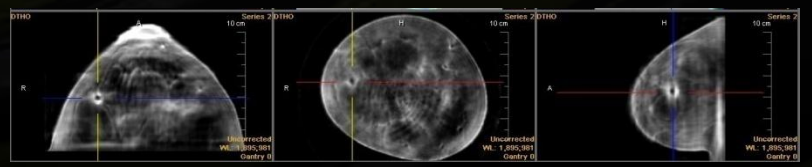
"We can only hope that in the long run, Silicon Informatics' efforts will accelerate the discovery of new drugs to treat a wide range of diseases, from cancer to Alzheimer's, HIV to malaria."

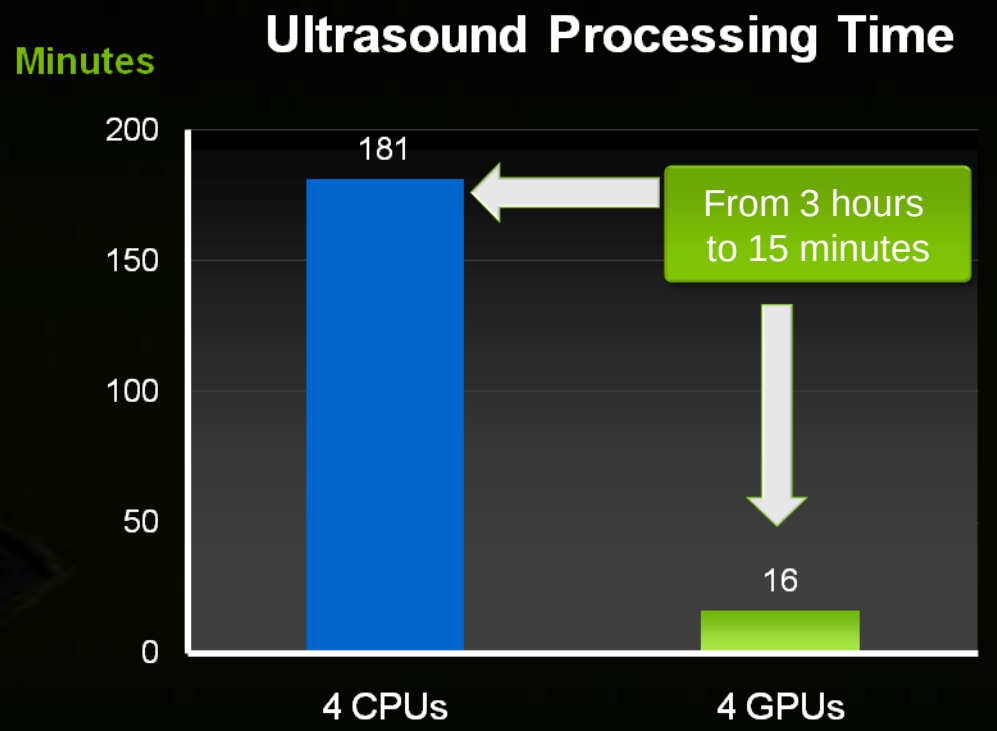
Dr. Garrett Morris, Scripps, Author of AutoDock



Tesla Revolutionizes Breast Cancer Detection

3D Ultrasound
Image Reconstruction





Techniscan Medical Systems

Bioinformatics

- **Molecular Biology**
  - **Searches for similarities in protein and DNA databases**
  - **Smith-Waterman Algorithm is the most accurate but most time consuming, CUDA enables faster results**

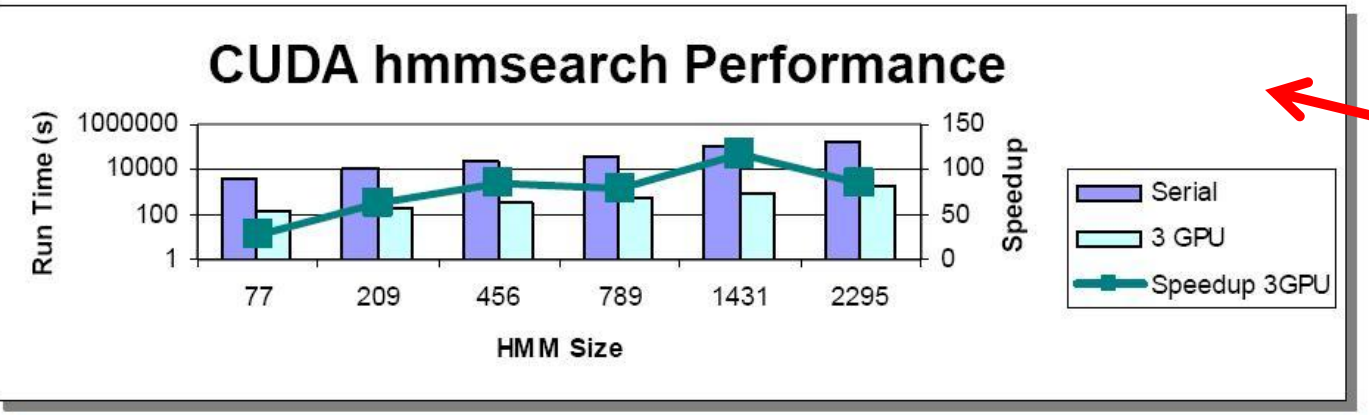
Algorithm	Scenario 1		Scenario 2		Scenario 3	
	Runtime	Speedup	Runtime	Speedup	Runtime	Speedup
Seq (1)	1.47s	1.00	113.26s	1.00	401.06s	1.00
Seq (2)	2.73s	0.54	211.92s	0.53	-	-
Par (1)	21.22s	0.07	-	-	-	-
Par (2)[1]	1.86s	0.79	96.35s	1.18	316.14s	1.27
CUDA	0.66s	2.23	22.75s	4.98	57.14s	7.02

Reference: http://www.kruber.eu/files/SE-HWAccel-presentation.pdf

MPI-HMMR

Open source MPI implementation of the HMMER protein sequence analysis suite

3 GPUs, 117x Speedup



Use as many GPUs as your system supports

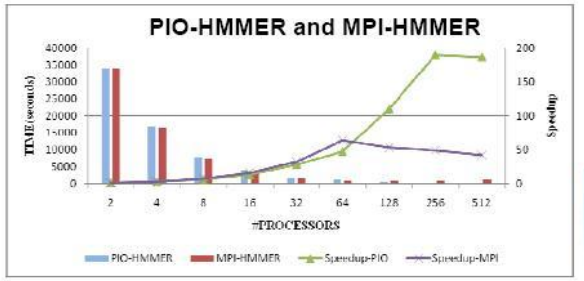
*mpi*HMMER

Available Now

***mpi*HMMER** is a multi-layer performance-enhanced version of Sean Eddy's HMMER.

Performance enhancing optimizations implemented include

- MPI Support
- Parallel I/O Support (New)
- Multi-GPU Support (New)



Currently Supports

- GPP Clusters
- NVIDIA CUDA

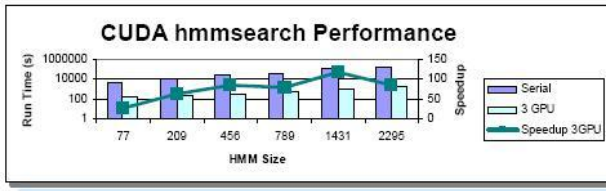
Coming Soon

- Integrated Accelerator Platform

Team

John Paul Walters
Rohan Darole
Joseph Landman
Vipin Chaudhary

3 GPUs, 117x Speedup



Use as many GPUs as your system supports

Visit www.mpihmmer.org for more information and for software downloads.

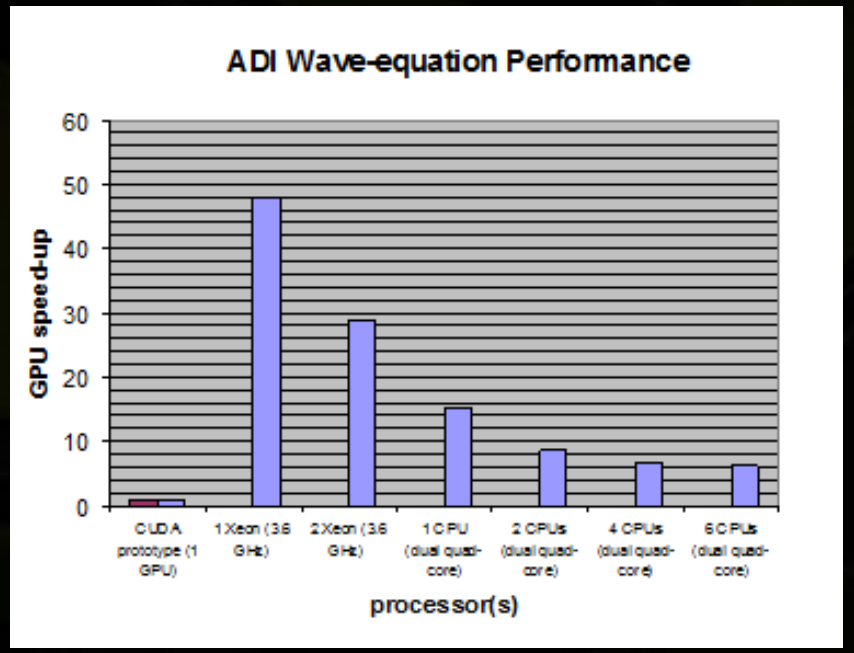
Scalable Informatics

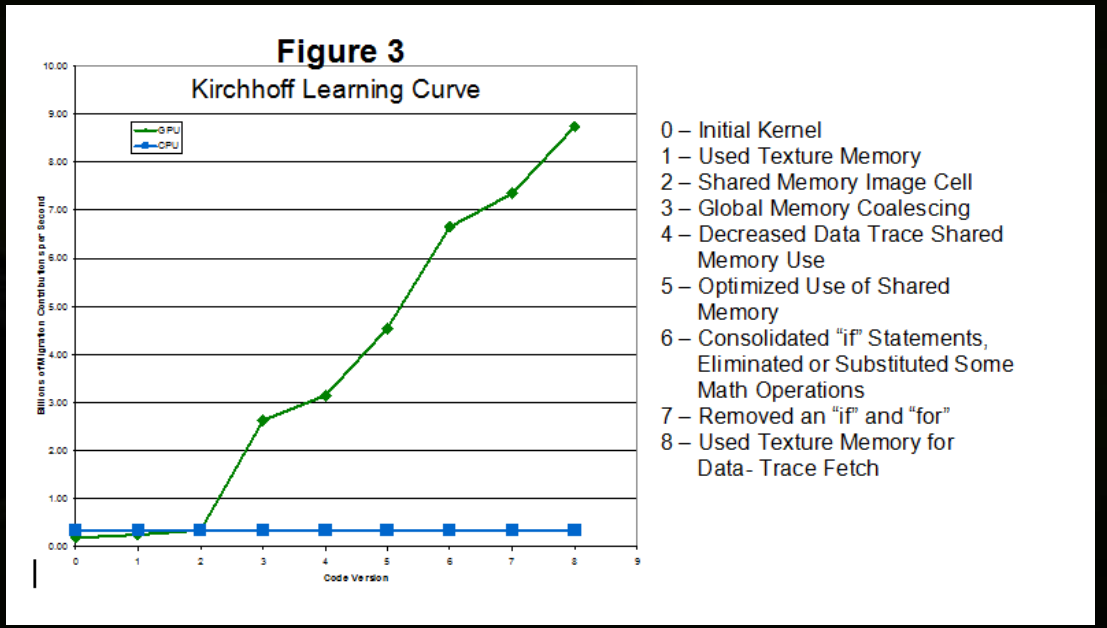
University at Buffalo
The State University of New York

Oil and Gas: Migration Codes

"Based on the benchmarks of the current prototype [128 node GPU cluster], this code should outperform our current 4000-CPU cluster"

Leading global independent energy company





nvidiatesla

S'abonner

nvidiatesla
Inscription : **16 octobre 2008**
Dernière connexion : **il y a 1 semaine**
Vidéos visionnées : **98**
Abonnés : **90**
Vues (chaîne) : **18304**

GURU

NVIDIA® Tesla™ computing solutions enable the necessary transition to energy efficient parallel computing power. With 240 cores per processor and based on the revoluationary NVIDIA® CUDA™ parallel computing architecture, Tesla scales to solve the worlds most important computing challenges—more quickly and accurately.

To learn more about Tesla computing solutions, visit http://www.nvidia.com/tesla

To learn more about the CUDA parallel computing architecture, visit http://www.nvidia.com/cuda

Âge : **46**

Pays : **États-Unis**

Signaler l'infraction concernant l'image de profil

Contacter nvidiatesla

Envoyer un message
Ajouter un commentaire
Partager la chaîne
Google Ajouter à iGoogle

http://fr.youtube.com/nvidiatesla

Activités récentes

nvidiatesla a envoyé une nouvelle vidéo.
(il y a 1 semaine)

Ian Buck talks about CUDA a...
Ian Buck of NVIDIA talks about CUDA and way he exposes to the developer ... suite

nvidiatesla a envoyé une nouvelle vidéo.
(il y a 1 semaine)

Tesla Personal Supercomputer
De : nvidiatesla
Vues : 37848
Commentaires écrits : 45

Vidéos (23)

S'abonner aux vidéos de nvidiatesla

Vidéos | Les plus regardées | Les plus commentées

Rechercher

Ian Buck talks about CUDA and
il y a 1 semaine
384 vues
nvidiatesla

Hideyuki Torii CEO of Numerical ...
il y a 1 semaine
271 vues
nvidiatesla

Matthew Walker of University of ...
il y a 1 semaine
233 vues
nvidiatesla
aucun avis


NVIDIA

For more information

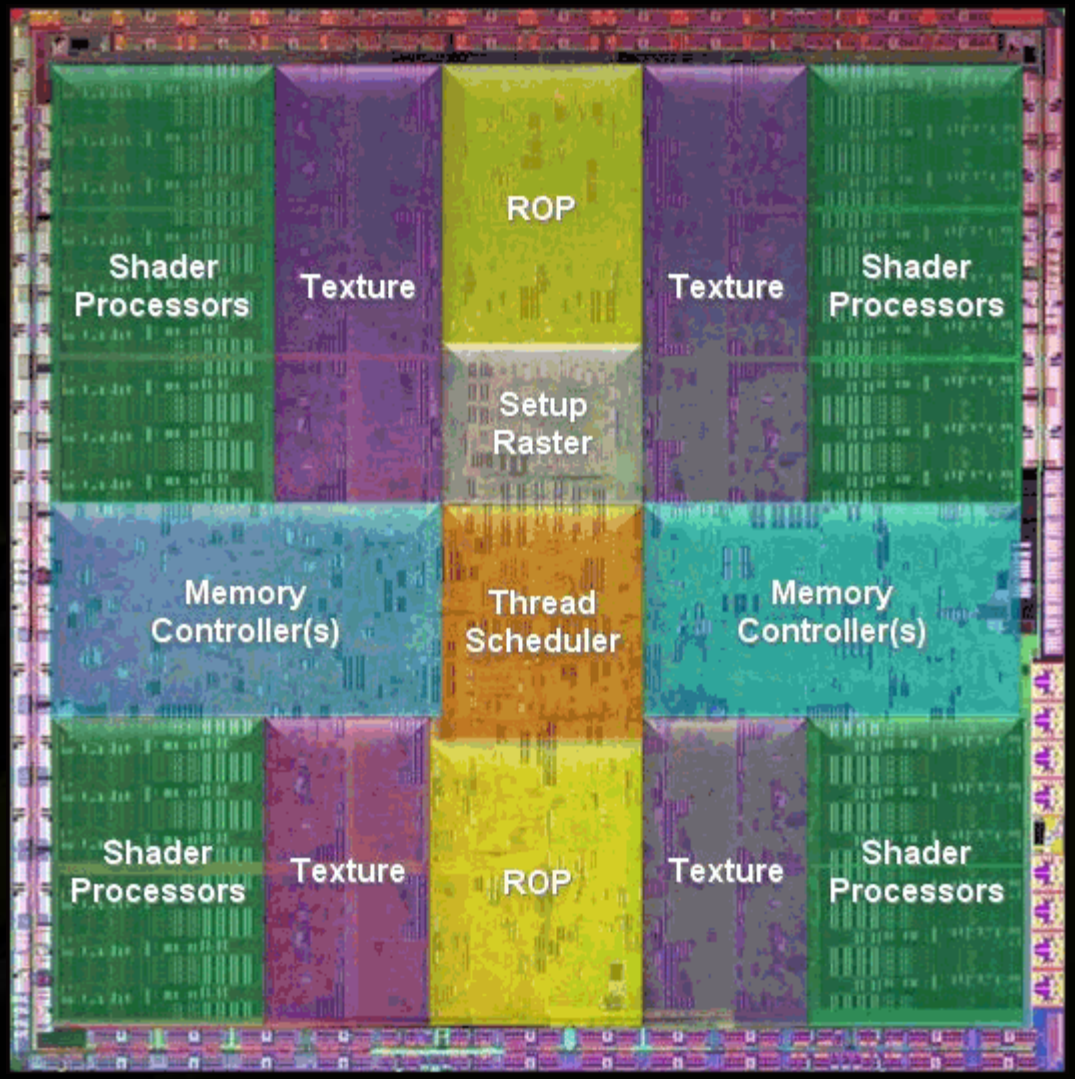
Jean-Christophe Baratault

jbaratault@nvidia.com

www.nvidia.com/Tesla

Backup Slides

GT200 Die

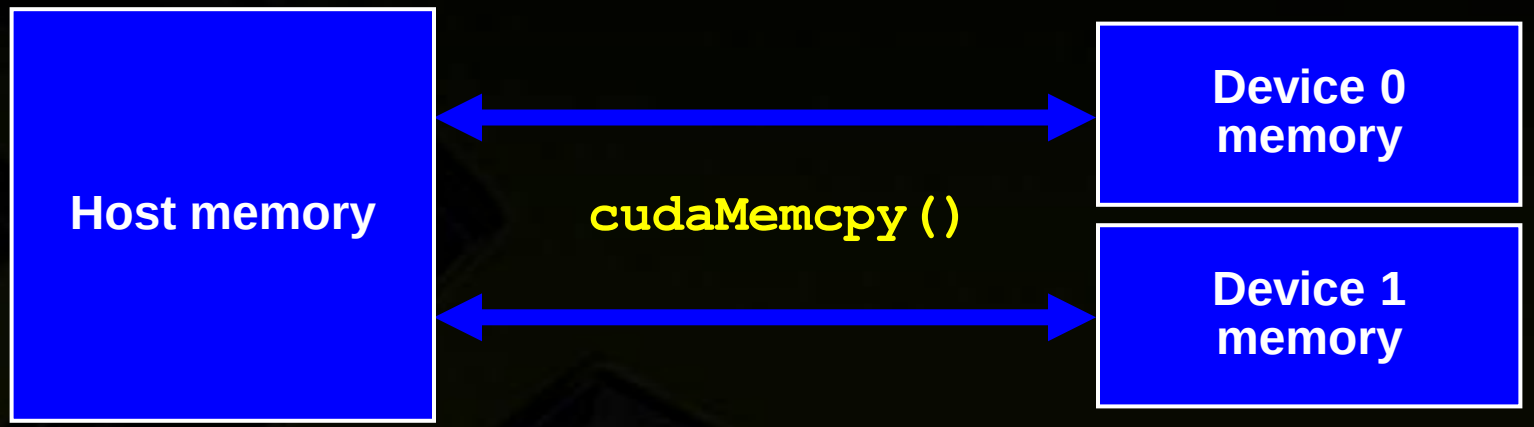

Shader Processors
Texture
ROP
Texture
Shader Processors
Setup Raster
Memory Controller(s)
Thread Scheduler
Memory Controller(s)
Shader Processors
Texture
ROP
Texture
Shader Processors

Definitions

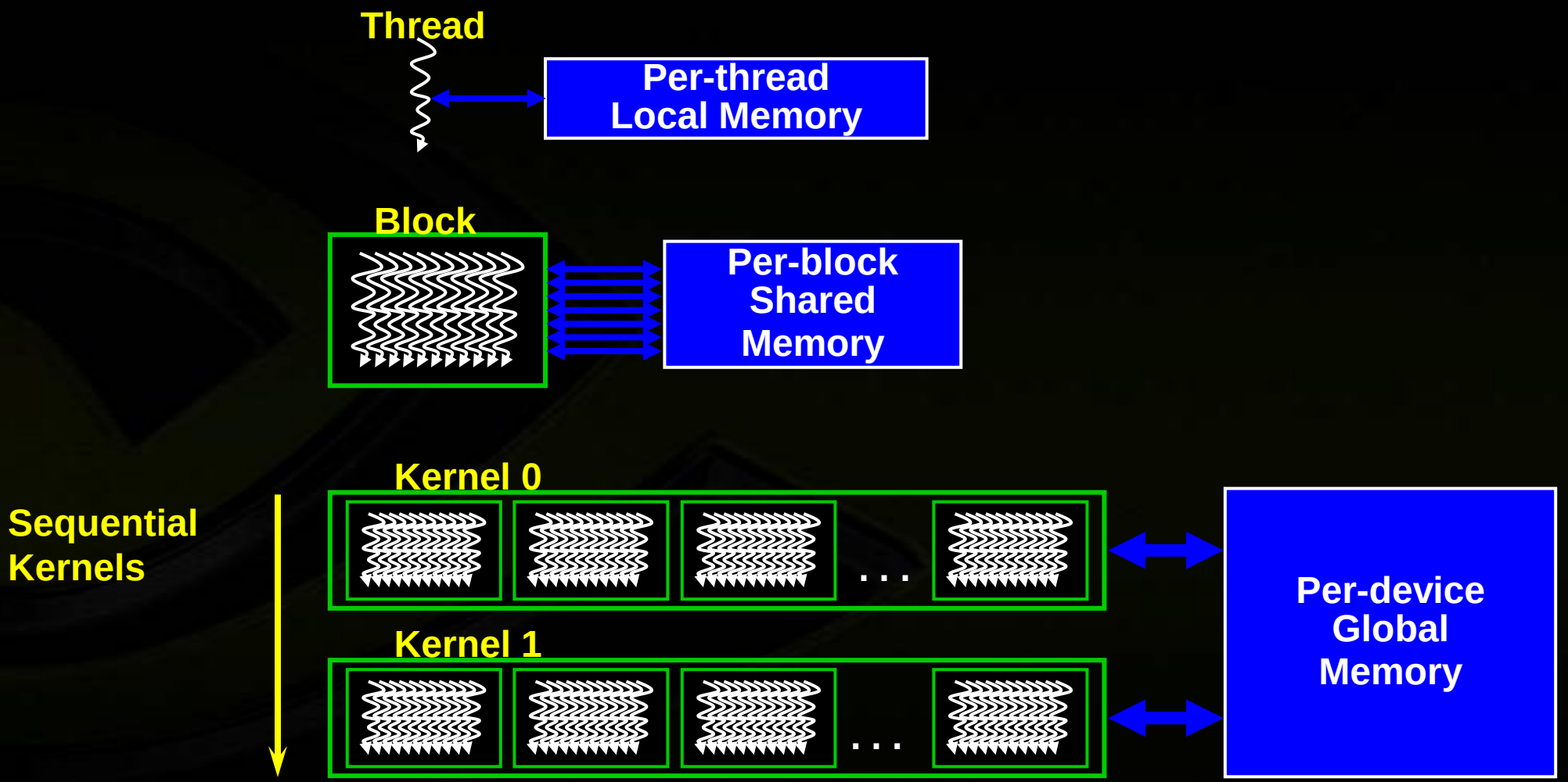
- Device = GPU = set of multiprocessors
- Multiprocessor = set of processors & shared memory
- Kernel = GPU program
- Grid = array of thread blocks that execute a kernel
- Thread block = group of SIMD threads that execute a kernel and can communicate via shared memory

Memory	Location	Cached	Access	Who
Local	Off-chip	No	Read/write	One thread
Shared	On-chip	N/A	Read/write	All threads in a block
Global	Off-chip	No	Read/write	All threads + host
Constant	Off-chip	Yes	Read	All threads + host
Texture	Off-chip	Yes	Read	All threads + host

Heterogeneous Memory Model

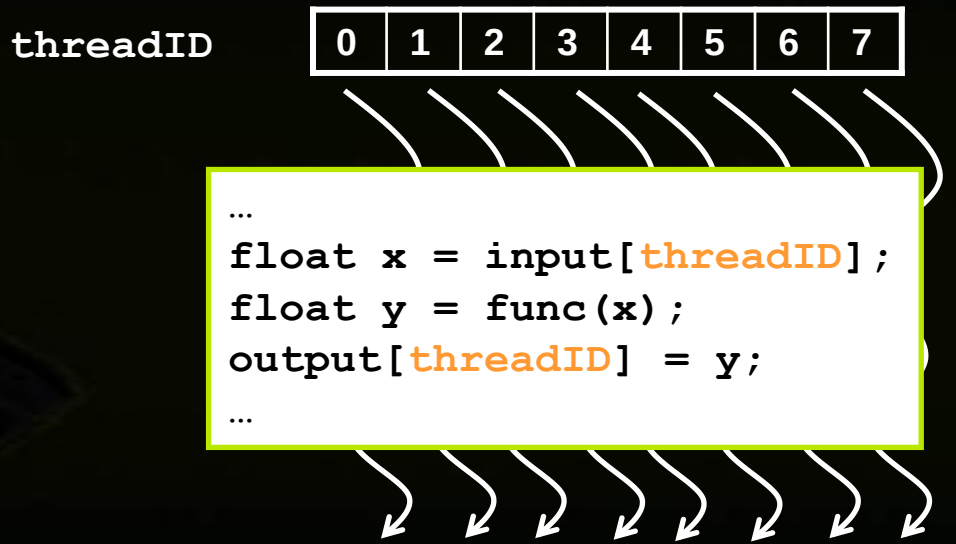


Memory Hierarchy



Kernel = Many Concurrent Threads

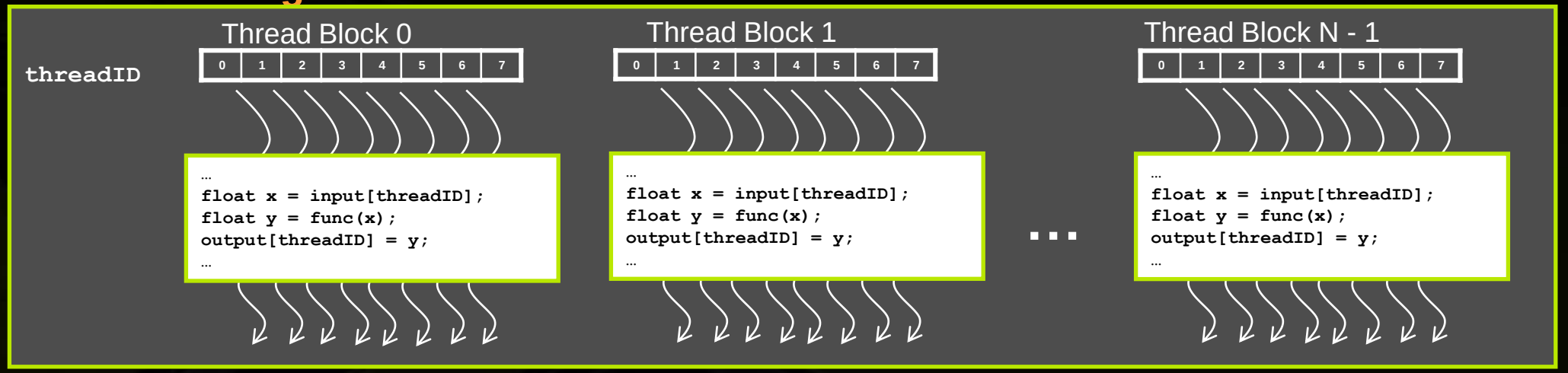
- One kernel is executed at a time on the device
- Many threads execute each kernel
  - Each thread executes the same code…
  - … on different data based on its threadID



- CUDA threads might be
  - Physical threads
    - As on NVIDIA GPUs
    - GPU thread creation and context switching are essentially free
  - Or virtual threads
    - E.g. 1 CPU core might execute multiple CUDA threads

Hierarchy of Concurrent Threads

- Threads are grouped into thread blocks
  - Kernel = grid of thread blocks



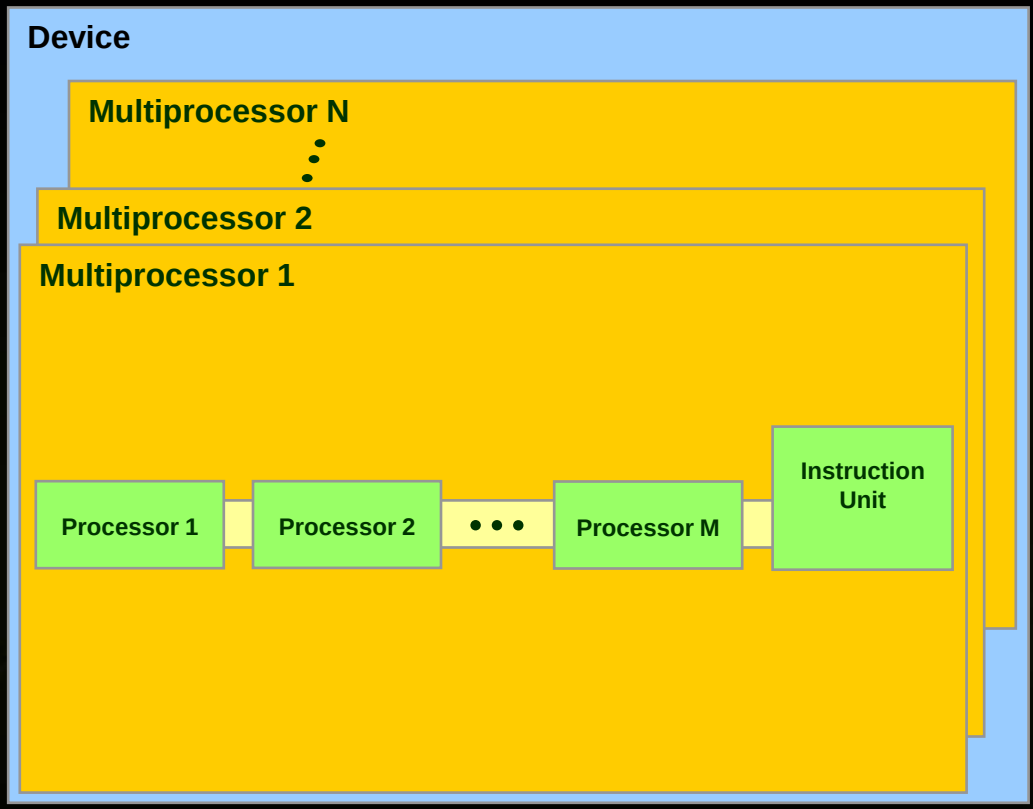
- By definition, threads in the same block may synchronize with barriers

```
__syncthreads();
int left = scratch[threadID - 1];
```

Threads wait at the barrier until all threads in the same block reach the barrier

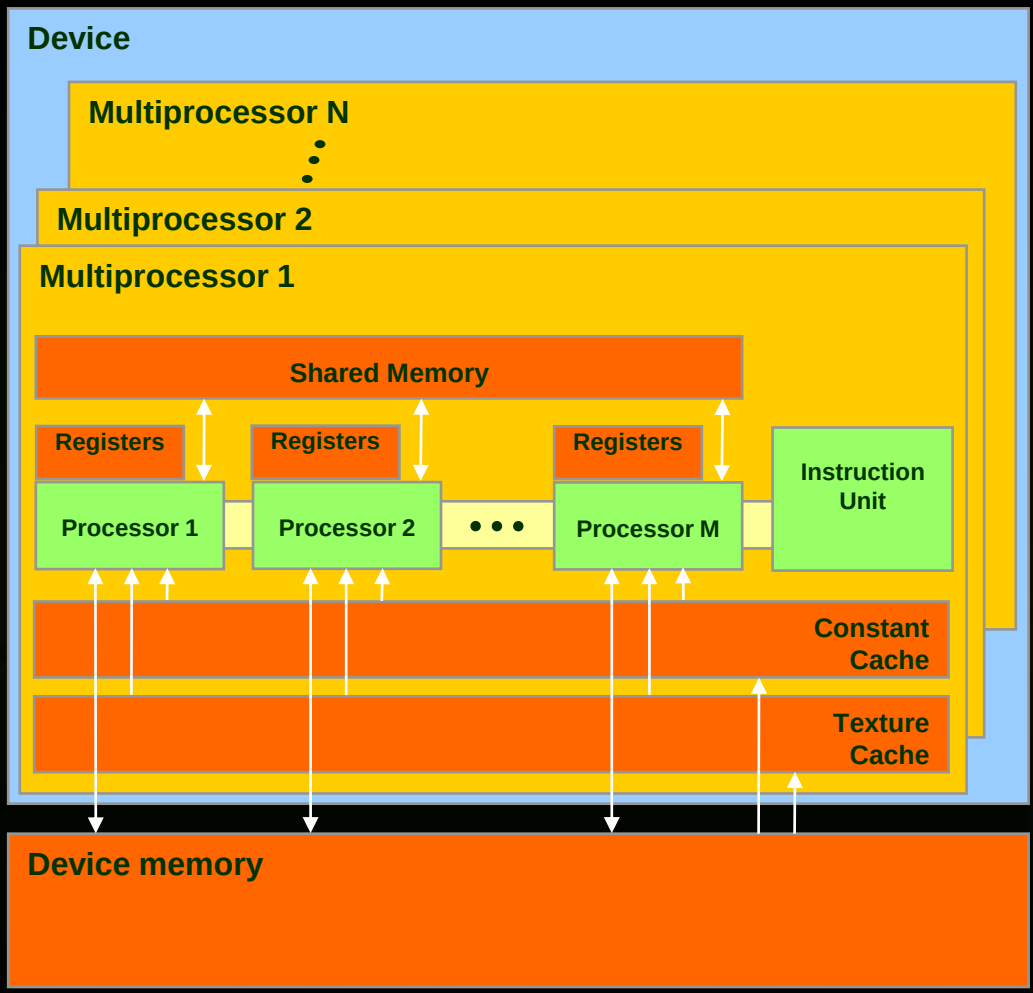
Hardware Implementation: A Set of SIMT Multiprocessors

- Each multiprocessor is a set of 32-bit processors with a Single-Instruction Multi-Thread architecture
  - 30 multiprocessors on GT200
  - 8 processors per multiprocessors
- At each clock cycle, a multiprocessor executes the same instruction on a group of threads called a warp
  - The number of threads in a warp is the warp size (= 32 threads on GT200)
  - A half-warp is the first or second half of a warp



Hardware Implementation: Memory Architecture

- The global, constant, and texture spaces are regions of device memory
- Each multiprocessor has:
  - A set of 32-bit registers per processor (16,384 on GT200)
  - On-chip shared memory (16KB on GT200)
    - Where the shared memory space resides
  - A read-only constant cache
    - To speed up access to the constant memory space
  - A read-only texture cache
    - To speed up access to the texture memory space

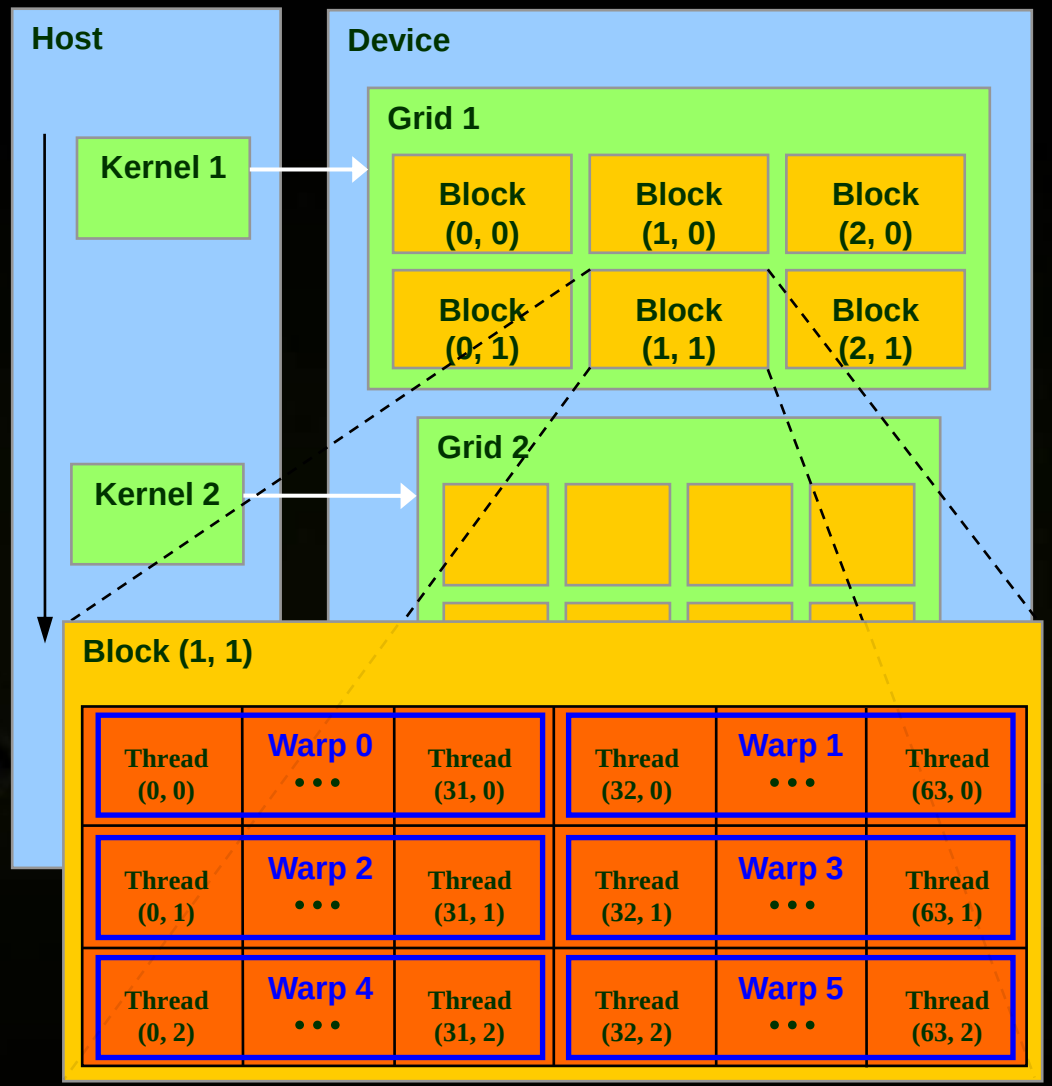


Hardware Implementation: Execution Model

- Each multiprocessor processes batches of blocks one batch after the other
  - Active blocks = the blocks processed by one multiprocessor in one batch
  - Active threads = all the threads from the active blocks
- The multiprocessor's registers and shared memory are split among the active threads
- Therefore, for a given kernel, the number of active blocks depends on:
  - The number of registers the kernel compiles to
  - How much shared memory the kernel requires
- If there cannot be at least one active block, the kernel fails to launch

Hardware Implementation: Execution Model

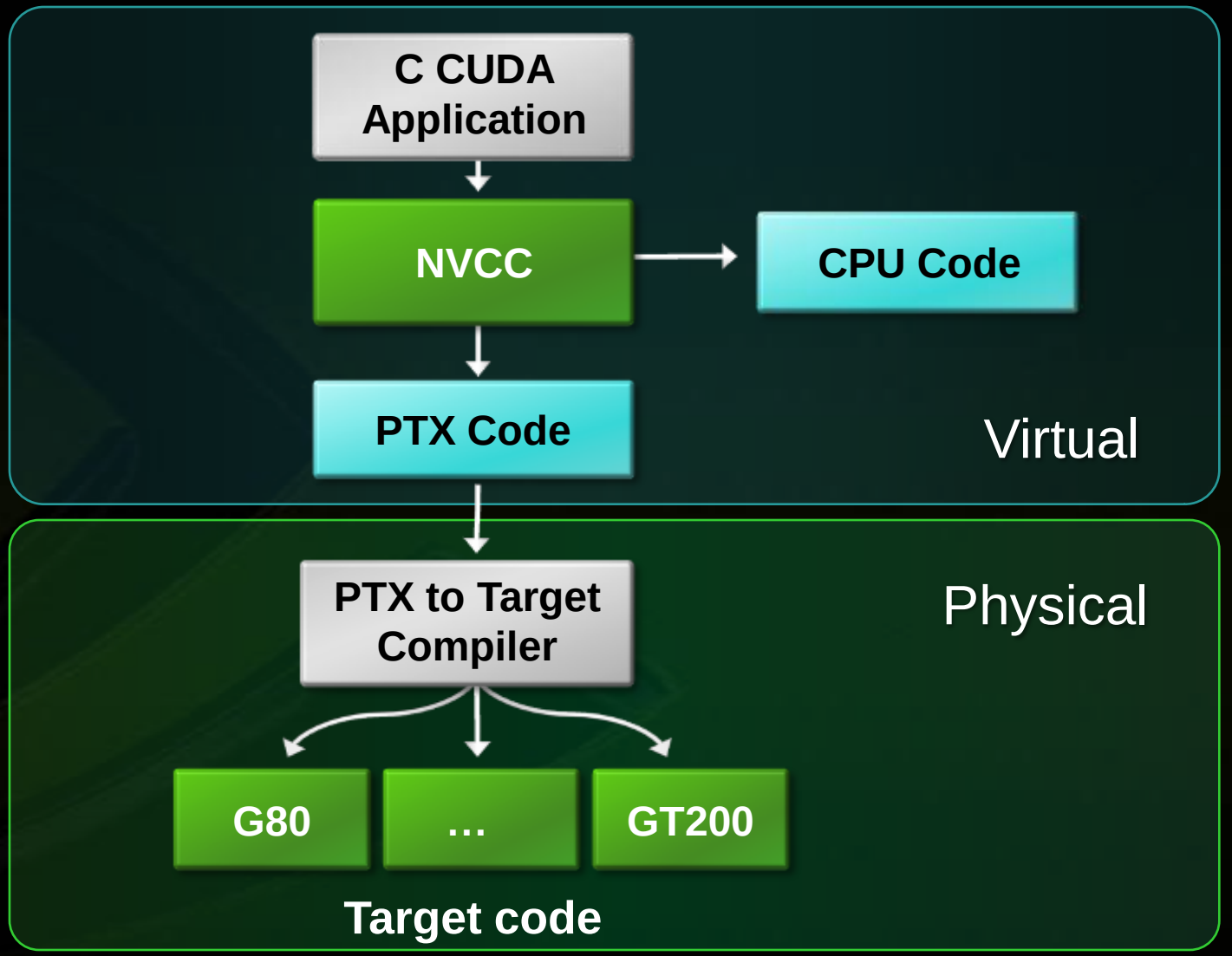
- Each active block is split into warps in a well-defined way
- Warps are time-sliced
- In other words:
  - Threads within a warp are executed *physically* in parallel
  - Warps and blocks are executed *logically* in parallel



Scalability Solution

- **Programmer uses multi-level data parallel decomposition**
  - Decomposes problem into a sequence of steps (**Grids**)
  - Decomposes Grid into independent parallel Blocks (**thread blocks**)
  - Decomposes Block into cooperating parallel elements (**threads**)
- **GPU hardware distributes thread blocks to available multiprocessors**
  - GPU balances work load across any number of multiprocessors cores
  - Core executes program that computes Block
- **Each thread block computes independently of others**
  - Enables parallel computing of Blocks of a Grid
  - No communication among Blocks of same Grid
  - Scales one program across any number of parallel cores
- **Programmer writes one program for all GPU sizes**
- **Program does not know how many cores it uses**
- **Program executes on GPU with any number of cores**

Compiling CUDA



Role of Open64

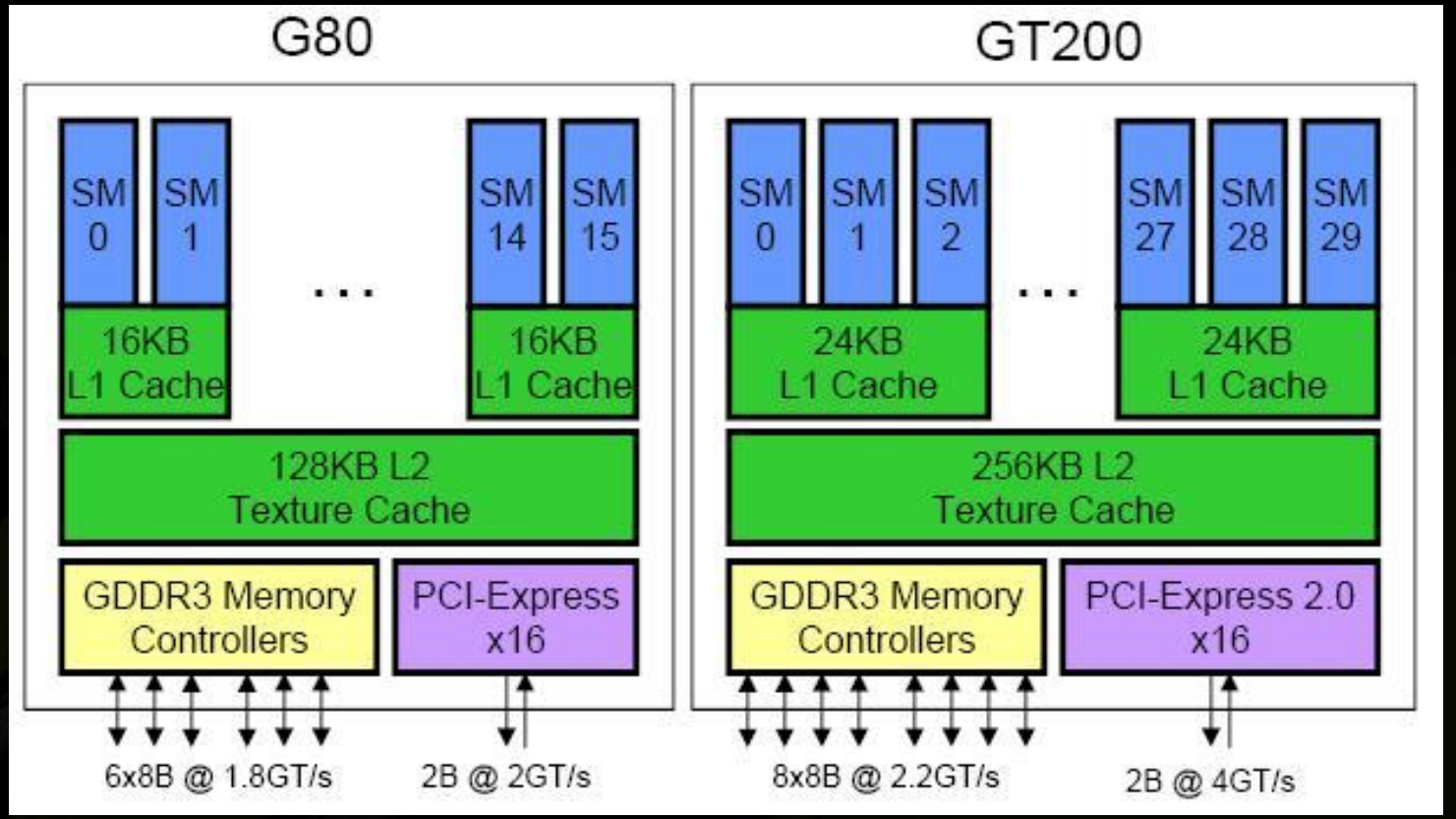
Open64 compiler gives us

- A complete C/C++ compiler framework. Forward looking. We do not need to add infrastructure framework as our hardware arch advances over time.
- A good collection of high level architecture independent optimizations. All GPU code is in the inner loop.
- Compiler infrastructure that interacts well with other related standardized tools.

CUDA Advantages over Legacy GPGPU

- Random access byte-addressable memory
  - Thread can access any memory location
- Unlimited access to memory
  - Thread can read/write as many locations as needed
- Shared memory (per block) and thread synchronization
  - Threads can cooperatively load data into shared memory
  - Any thread can then access any shared memory location
- Low learning curve
  - Just a few extensions to C
  - No knowledge of graphics is required
- No graphics API overhead

GPU Comparison



Nov06	**G80**	128 SP	384-bit mem i/f	PCIe Gen1
Jun08	**GT200**	240 SP	512-bit mem i/f	PCIe Gen2

Tesla	C870	C1060
GPU	**G80**	**T10**
Device memory	1.5GB	**4GB**
Multiprocessor	16	**30**
Cores	128	**240**

Per multiprocessor

Shared memory	16KB	16KB
Cache for constant memory	8KB	8KB
Cache for texture memory	8KB	8KB
Active block	8	8
Active warps	24	**32**
Active threads	768	**1 024**
Registers	8 192	**16 384**

Threads per block	512
x, y, z dimension	512, 512, 64
Grid thread block	65 535
Warp size	32
Constant memory	64KB

GT200 New features

- ✓Atomic functions operating on 32-bit words in global memory
- ✓Atomic functions operating in shared memory
- ✓Atomic functions operating on 64-bit words in global memory
- ✓Warp vote functions
- ✓Double-precision floating-point numbers

Selecting a CUDA Platform

	Tesla	Quadro	GeForce
Stress tested and burned-in with added margin for numerical accuracy	X		
Manufactured by NVIDIA with professional grade memory	X	X	
NVIDIA care: 3-year warranty from NVIDIA, enterprise support	X	X	
4 Gigabyte on-board memory for large technical computing data sets	X	X	
Single card solution for professional visualization and CUDA computing		X	
Consumer middle-ware and applications: PhysX, Video, Imaging			X
Consumer product life cycle			X
Manufactured and guaranteed by NVIDIA graphics add-in card partners			X
Product support through NVIDIA graphics add-in card partners			X

	Tesla 8-series C870 card	Tesla 10-series C1060 card
Number of Cores	**128**	**240**
32-bit FP Performance	**0.5 Teraflop**	**1 Teraflop**
On-board Memory	**1.5 GB**	**4.0 GB**
Memory interface	**384-bit GDDR3**	**512-bit GDDR3**
Memory I/O bandwidth	**77 GB/sec**	**102 GB/sec**
System interface	**PCIe x16 Gen1**	**PCIe x16 Gen2**

Intel PCIe bus

- PCIe x16 Gen2
  - x16 physical & electrical — 5.5GB/s
  - x16 physical / x8 electrical — 2.7GB/s
  - x16 physical / x4 electrical — 1.4GB/s
- PCIe x16 Gen1
  - x16 physical & electrical — 2.5GB/s
  - x16 physical / x8 electrical — 1.4GB/s
  - X16 physical / x4 electrical — 700MB/s

NVIDIA GPU Brand Feature Comparison

	NVIDIA Tesla	NVIDIA Quadro	NVIDIA GeForce
GPU Designed and Mfg by	NVIDIA	NVIDIA	NVIDIA
Product Engineered By	NVIDIA	NVIDIA	Add In Card maker (AIC)
Components Selected and Sourced by	NVIDIA	NVIDIA	AIC
ECO Control	NVIDIA	NVIDIA	AIC
Quality Testing	Compute and Memory	Professional Graphics	Consumer Graphics
Form Factors	Card and 1U	Card, Deskside and 1U	Card
Roadmap	High Performance Computing	Professional Graphics (Open GL & DirectX Applications)	Consumer (Gaming) (DirectX Games)
Operating Specifications	Corporate Compute Environment	Professional Workstation, Thin Client (passive)	Consumer (Gaming)
Supported Provided By	NVIDIA	NVIDIA	AIC
Max Data Readback	3 GB/s (CUDA)	3 GB/s (OGL, DX)	1 GB/s
Max Frame Buffer/GPU (On Board Memory)	4 GB	4 GB	1 GB
Lifecycle	36 months Managed by NVIDIA	24-36 months Managed by NVIDIA	9-12 month Varies by AIC manufacturer

Tesla – Quadro Positioning

		
High Level Positioning	Optimized for ***Computing***	Optimized for ***Professional Visualization***
Application Testing	**Compute validation of memories** (additional testing for data access)	**Testing for graphics image rendering** (frame buffer)
Graphics Capabilities	**Standard OpenGL** (compatible with mGPU/GeForce)	**Quadro OpenGL & Direct X** (certified for Pro WS Apps)
Products	**HPC boards & 1U systems**	**Full graphics product line** (mGPU, 2D, 3D, Vertical, Systems)
Roadmap	**Computing** > Double Precision (FP64) > ECC > Computing developer program > Tesla cluster promotion	**Professional Visualization** > More shader, geometry, fill rate > Increases in image quality > Pro App Scaling > Quadro specific features > Virtualization & Remoting

Google | CEA | Recherche Actualités | Rechercher sur le Web

Résultats dans l'actualité

Parcourir les articles à la une

depuis une heure
Hier
depuis une semaine
depuis un mois

Alertes Actualités

RSS | Atom

« Afficher tous les résultats Web pour CEA

Le **CEA** opte pour un supercalculateur Bull
Neteco - Il y a 2 heures
Comme les poids lourds américains, IBM en tête, le groupe français Bull s'active sur le marché des supercalculateurs, le **CEA** apprécie. ...
Bull : commande de Genci et du **CEA** pour un supercalculateur Boursier.com
La France se relance dans le calcul intensif Les Echos
Le plus important des supercalculateurs à base de GPU est en France PCWorld France
Le Monde - LeMondeInformatique
21 autres articles >>

e CEA et le CRNS préparer
ournal du Net - 8 avr 2008
e centre, qui cumule 500 téraflops
aut débit les sites du **CEA** et du C

e CNRS et le **CEA** ouvrent u
nunet.fr - 8 avr 2008
. du Développement et des Ressc
entre de calcul "recherche et tech

NVIDIA

PCWorld

Rechercher sur PC World.fr | Recherche

Accueil | Brèves | Tests | Vidéos | Prix

LES CATÉGORIES

Types de document

- Audio
- Business
- Composants PC
- HDTV
- Imprimantes
- Internet
- Jeux Vidéo
- Logiciels
- Mac & iPods
- Moniteurs
- Ordinateurs de bureau
- Ordinateurs portables
- Photo
- Sécurité
- Stockage
- Téléphonie
- Vidéo
- Windows Vista & XP

Tags : nvidia

Le plus important des supercalculateurs à base de GPU est en France

Un supercalculateur à base de processeurs graphiques a été commandé à Bull par l'administration française. C'est la plus puissante architecture à base de processeurs graphiques connue à ce jour et elle est en France.

Tanguy ANDRILLON
mardi, 22 avril à 15:33:36

PRINT COMMENT RSS

Recommander cet article ? Oui 9 Vote(s) Non

Créé l'année dernière par le ministère de l'Education, le Genci (Grand Equipement National de Calcul Intesif) est une société civile détenue par l'Etat à 50 %, le CEA, le CNRS et les universités. Cette société vient de commander un serveur Novascale à la société Bull. Ce supercalculateur embarque 1 068 processeurs Intel Xeon à huit cœurs et 96 processeurs graphiques NVIDIA de prochaine génération regroupant 24 000 streams proccessors. Le tout est associé à 1 To de mémoire vive et à 1 Po (pétaoctet = 1 024 To) d'espace de stockage.

Le supercalculateur Novascale délivre une puissance brute de 295 téraflops dont 192 sont obtenus à l'aide des processeurs graphiques. Evidemment, les scientifiques devront développer un code spécifique à grand niveau de

LeMondeInformatique.fr

Toute l'info et les tendances du monde IT

- Accueil
- Rubriques
- Technologie
- Economie IT
- Développement
- Solutions PME
- SSII
- Emploi/Formation
- Micro
- Numérique
- Agenda
- Thèmes
  - Décisionnel
  - Mobilité
  - Sécurité
  - Réseaux
  - Architecture logicielle
  - PC et portables
  - Applications transversales
  - Infrastructure Serveur
  - Middleware
  - Open Source
- Conférences
  - Accueil
  - Forum SOA
  - Forum Convergence
- Vidéos
- LMI Blogs
- Téléchargements
- Newsletters
- Flux RSS

lenovo

Infrastructure serveur
Inscrivez-vous XML
Consulter le centre de compétences

Version imprimable | Envoyer à un ami | Recevez les news

Le CEA et le CNRS donnent des ambitions européennes à leur supercalculateur

Edition du 18/04/2008 - par Emmanuelle Delsol

Les deux organisations installent un cluster Bull couplant CPU et GPU pour atteindre 300 Tflops en 2009. Cette puissance servira aux chercheurs pour des travaux sur la simulation numérique. Un enjeu industriel fort.

Le CEA et le CNRS réunissent leurs deux supercalculateurs de l'Essonne au sein du Centre National Jacques Louis Lions de Calcul Haute Performance de l'Essonne et en augmentent la puissance. L'ensemble devrait proposer plus de 500 Tflops (200 au CNRS et 300 au CEA) dès l'an prochain aux chercheurs français pour étudier en particulier le domaine de la simulation numérique. Cerise sur le supercalculateur, le centre de calcul est candidat pour accueillir un noeud du réseau européen Prace (Partnership for Advanced Computing in Europe) destiné aux machines de capacité pétaflopique. A la clé une subvention à hauteur de 10% pour l'infrastructure. Le projet d'un montant de 10M€ est financé par le Genci (Grand Equipement National de Calcul Intensif), société civile créée pour coordonner les politiques françaises d'équipement en supercalculateurs.

un mélange de processeurs classiques et graphiques

Le Centre Jacques Louis Lions est adossé d'une part à l'Institut du Développement et des Ressources en Informatique Scientifique du CNRS (Idris) et d'autre part au Centre de calcul « recherche et technologie » du CEA (CCRT). Le premier,