HIGH-PERFORMANCE COMPUTING WITH NVIDIA TESLA GPUs

Timothy Lanfear, NVIDIA
WHY GPU COMPUTING?
Science is Desperate for Throughput

- **1 Exaflop**
  - BPTI: 3K atoms
  - Estrogen Receptor: 36K atoms
  - F1-ATPase: 327K atoms
  - Ribosome: 2.7M atoms
  - Chromatophore: 50M atoms
  - Ran for 8 months to simulate 2 nanoseconds

- **1 Petaflop**
  - Bacteria: 100s of Chromatophores
Power Crisis in Supercomputing

- 1982
- 1996
- 2008
- 2020

- Exaflop
- Petaflop
- Teraflop
- Gigaflop

Household Power Equivalent

- City
- Town
- Neighborhood
- Block

- Jaguar Los Alamos

- 7,000,000 Watts
- 25,000,000 Watts
- 850,000 Watts
- 60,000 Watts

7,000,000 Watts
25,000,000 Watts
850,000 Watts
60,000 Watts

© NVIDIA Corporation 2009
“Oak Ridge National Lab (ORNL) has already announced it will be using Fermi technology in an upcoming super that is ‘expected to be 10-times more powerful than today’s fastest supercomputer.’

Since ORNL’s Jaguar supercomputer, for all intents and purposes, holds that title, and is in the process of being upgraded to 2.3 Petaflops ...

... we can surmise that the upcoming Fermi-equipped super is going to be in the **20 Petaflops** range.”
What is GPU Computing?

Computing with CPU + GPU

Heterogeneous Computing
Low Latency or High Throughput?

**CPU**
- Optimised for low-latency access to cached data sets
- Control logic for out-of-order and speculative execution

**GPU**
- Optimised for data-parallel, throughput computation
- Architecture tolerant of memory latency
- More transistors dedicated to computation
Why Didn’t GPU Computing Take 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 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

Create a GPU based Compute Ecosystem by 2008
NVIDIA GPU Computing Ecosystem

GPU Architecture
CUDA SDK & Tools
NVIDIA Hardware Solutions

CUDA Training Company
CUDA Development Specialist
ISV
TPP / OEM
Hardware Architect
VAR

Deployment

Customer Requirements
Customer Application
Hardware Architecture
Fermi: The Computational GPU

**Performance**
- 13 × 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
- Enable up to 1 TeraByte of GPU Memories
- High Speed GDDR5 Memory Interface

**Usability**
- Multiple Simultaneous Tasks on GPU
- 10 × Faster Atomic Operations
- C++ Support
- System Calls, printf support

Disclaimer: Specifications subject to change
# NVIDIA Tesla GPU Computing Products

## Data Center Products

<table>
<thead>
<tr>
<th>Server Module</th>
<th>1U Systems</th>
<th>Workstation Boards</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>GPUs</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tesla M1060</td>
<td>Tesla S2070</td>
<td>Tesla C2070</td>
</tr>
<tr>
<td>Tesla M2050</td>
<td>Tesla 2050</td>
<td>Tesla C2050</td>
</tr>
<tr>
<td><strong>Single Precision</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 T10 GPU</td>
<td>4 T20 GPUs</td>
<td>1 T20 GPU</td>
</tr>
<tr>
<td>1 T20 GPU</td>
<td>4 T10 GPUs</td>
<td>1 T10 GPU</td>
</tr>
<tr>
<td>933 GFlops</td>
<td>4120 GFlops</td>
<td>1030 Gflops</td>
</tr>
<tr>
<td>1030 GFlops</td>
<td>4140 GFlops</td>
<td>933 GFlops</td>
</tr>
<tr>
<td><strong>Double Precision</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>78 GFlops</td>
<td>2060 GFlops</td>
<td>515 Gflops</td>
</tr>
<tr>
<td>515 GFlops</td>
<td>346 GFlops</td>
<td>78 GFlops</td>
</tr>
<tr>
<td><strong>Memory</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 GB</td>
<td>12 GB</td>
<td>6 GB</td>
</tr>
<tr>
<td>3 GB</td>
<td>3 GB / GPU</td>
<td>4 GB / GPU</td>
</tr>
<tr>
<td><strong>Mem BW</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>102 GB/s</td>
<td>142.4 GB/s</td>
<td>144 GB/s</td>
</tr>
<tr>
<td>148.4 GB/s</td>
<td>102 GB/s</td>
<td>102 GB/s</td>
</tr>
<tr>
<td><strong>Display</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No display IO</td>
<td>No display IO</td>
<td>Single dual-link DVI</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No display IO</td>
</tr>
</tbody>
</table>

© NVIDIA Corporation 2009
## Application Software (written in C)

<table>
<thead>
<tr>
<th>CUDA Libraries</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>cuFFT</td>
<td>cuBLAS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CPU Hardware</th>
<th>CUDA Compiler</th>
<th>CUDA Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>1U PCI-E Switch</td>
<td>C Fortran</td>
<td>Debugger Profiler</td>
</tr>
</tbody>
</table>

- **CUDA Libraries**
  - cuFFT
  - cuBLAS
  - cuDPP

- **CPU Hardware**
  - 1U PCI-E Switch

- **CUDA Compiler**
  - C Fortran

- **CUDA Tools**
  - Debugger
  - Profiler

- **Application Software**
  - (written in C)

- **4 cores**
  - 240 cores
NVIDIA Parallel Nsight™

The first development environment for massively parallel applications.

**Hardware** GPU Source Debugging

**Platform-wide** Analysis

**Complete** Visual Studio integration

Register for the Beta

http://developer.nvidia.com/nsight
CUDA Zone: www.nvidia.com/CUDA

- CUDA Toolkit
  - Compiler
  - Libraries
- CUDA SDK
  - Code samples
- CUDA Profiler
- Forums
- Resources for CUDA developers
Wide Developer Acceptance and Success

146X
Interactive visualization of volumetric white matter connectivity

36X
Ion placement for molecular dynamics simulation

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
NVIDIA : Leadership in GPU Computing

Over 240 Universities Teaching CUDA

UIUC
MIT
Harvard
Berkeley
Cambridge
Oxford

IIT Delhi
Tsinghua
Dortmundt
ETH Zurich
Moscow
NTNU

Languages
C, C++
DirectX
Fortran
Java
OpenCL
Python

Tools
PGI Fortran
CAPs HMPP
Nexu
MCUDA
MPI
NOAA Fortran2C
OpenMP

Applications
Oil & Gas
Finance
CFD
Medical
Biophysics
Numerics
DSP
EDA

Consultants
ANEQ
CAPS
EM Photonics
CASS
GPU Tech

Libraries
FFT
BLAS
LAPACK
Image processing
Video processing
Signal processing
Vision

OEMs
Dell
HP
Cray
Summit
Supermicro
Sgi
Fujitsu
Lenovo
NEC
What We Did in the Past Three Years

- **2006**
  - G80, first GPU with built-in compute features, 128 core multi-threaded, scalable architecture
  - CUDA SDK Beta
- **2007**
  - Tesla HPC product line
  - CUDA SDK 1.0, 1.1
- **2008**
  - GT200, second GPU generation, 240 core, 64-bit
  - Tesla HPC second generation
  - CUDA SDK 2.0
- **2009 ...**
NEXT-GENERATION GPU ARCHITECTURE — ‘FERMI’
Introducing the ‘Fermi’ Architecture

The Soul of a Supercomputer in the body of a GPU

- 3 billion transistors
- Over 2× the cores (512 total)
- 8× the peak DP performance
- ECC
- L1 and L2 caches
- ~2× memory bandwidth (GDDR5)
- Up to 1 Terabyte of GPU memory
- Concurrent kernels
- Hardware support for C++
Design Goal of Fermi

- Expand performance sweet spot of the GPU
- Bring more users, more applications to the GPU
Streaming Multiprocessor Architecture

- 32 CUDA cores per SM (512 total)
- 2:1 ratio SP:DP floating-point performance
- Dual Thread Scheduler
- 64 KB of RAM for shared memory and L1 cache (configurable)
CUDA Core Architecture

- New IEEE 754-2008 floating-point standard, surpassing even the most advanced CPUs

- Fused multiply-add (FMA) instruction for both single and double precision

- Newly designed integer ALU optimized for 64-bit and extended precision operations
Cached Memory Hierarchy

First GPU architecture to support a true cache hierarchy in combination with on-chip shared memory

- L1 Cache per SM (32 cores)
  - Improves bandwidth and reduces latency

- Unified L2 Cache (768 KB)
  - Fast, coherent data sharing across all cores in the GPU

Parallel DataCache™ Memory Hierarchy
Larger, Faster Memory Interface

- GDDR5 memory interface
  - 2× speed of GDDR3
- Up to 1 Terabyte of memory attached to GPU
  - Operate on large data sets
Error Correcting Code

- ECC protection for
  - DRAM
    - ECC supported for GDDR5 memory

- All major internal memories are ECC protected
  - Register file, L1 cache, L2 cache
GigaThread™ Hardware Thread Scheduler

- Hierarchically manages thousands of simultaneously active threads
- 10× faster application context switching
- Concurrent kernel execution
GigaThread Hardware Thread Scheduler

Concurrent Kernel Execution + Faster Context Switch

Serial Kernel Execution

Parallel Kernel Execution
GigaThread Streaming Data Transfer Engine

- **Dual DMA engines**
  - Simultaneous CPU → GPU and GPU → CPU data transfer
  - Fully overlapped with CPU and GPU processing time

- **Activity Snapshot:**

```
Activity Snapshot:
```

```
Kernel 0
CPU  SDT0  GPU  SDT1

Kernel 1
CPU  SDT0  GPU  SDT1

Kernel 2
CPU  SDT0  GPU  SDT1

Kernel 3
CPU  SDT0  GPU  SDT1
```
Enhanced Software Support

- Full C++ Support
  - Virtual functions
  - Try/Catch hardware support

- System call support
  - Support for pipes, semaphores, printf, etc

- Unified 64-bit memory addressing
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

Fermi surpasses anything announced by NVIDIA's leading GPU competitor (AMD).
Fermi is the world’s first complete GPU computing architecture.

Peter Glaskowsky  
Technology Analyst  
The Envisioneering Group

The convergence of new, fast GPUs optimized for computation as well as 3-D graphics acceleration and industry-standard software development tools marks the real beginning of the GPU computing era. Gentlemen, start your GPU computing engines.

Nathan Brookwood  
Principle Analyst & Founder  
Insight 64
A 2015 GPU *
- ~20 times the performance of today’s GPU
- ~5,000 cores at ~3 GHz (50 mW each)
- ~20 TFLOPS
- ~1.2 TB/s of memory bandwidth

* This is a sketch of what a GPU in 2015 might look like, it does not reflect any actual product plans.
GPU Technology Conference 2010
Monday, September 20 – Thursday, September 23, 2010
San Jose Convention Center, San Jose, California

The most important event in the GPU ecosystem
- Learn about seismic shifts in GPU computing
- Preview disruptive technologies and emerging applications
- Get tools and techniques to impact mission critical projects
- Network with experts, colleagues, and peers across industries

“I consider the GPU Technology Conference to be the single best place to see the amazing work enabled by the GPU. It’s a great venue for meeting researchers, developers, scientists, and entrepreneurs from around the world.”

-- Professor Hanspeter Pfister, Harvard University and GTC 2009 keynote speaker