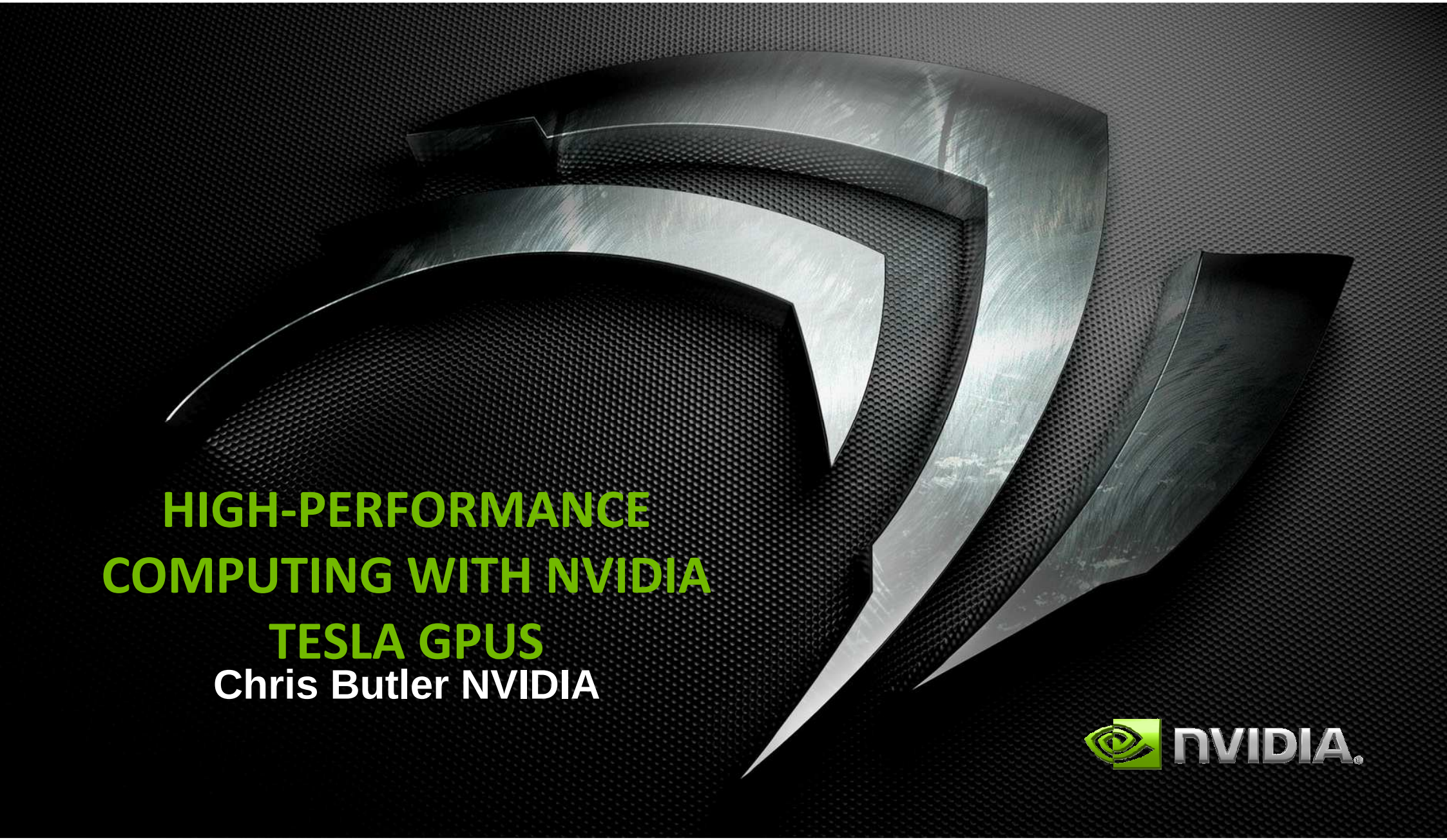
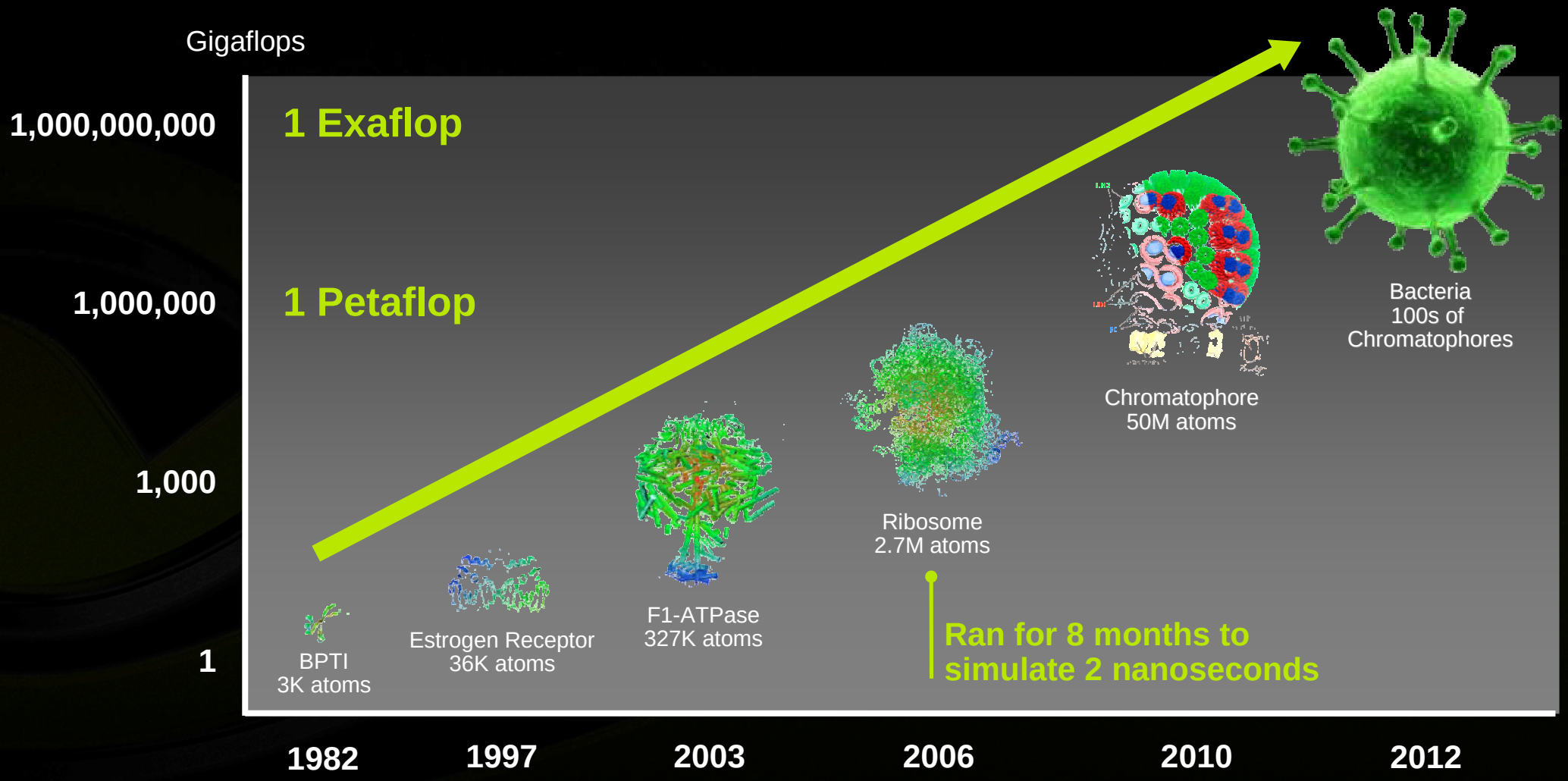
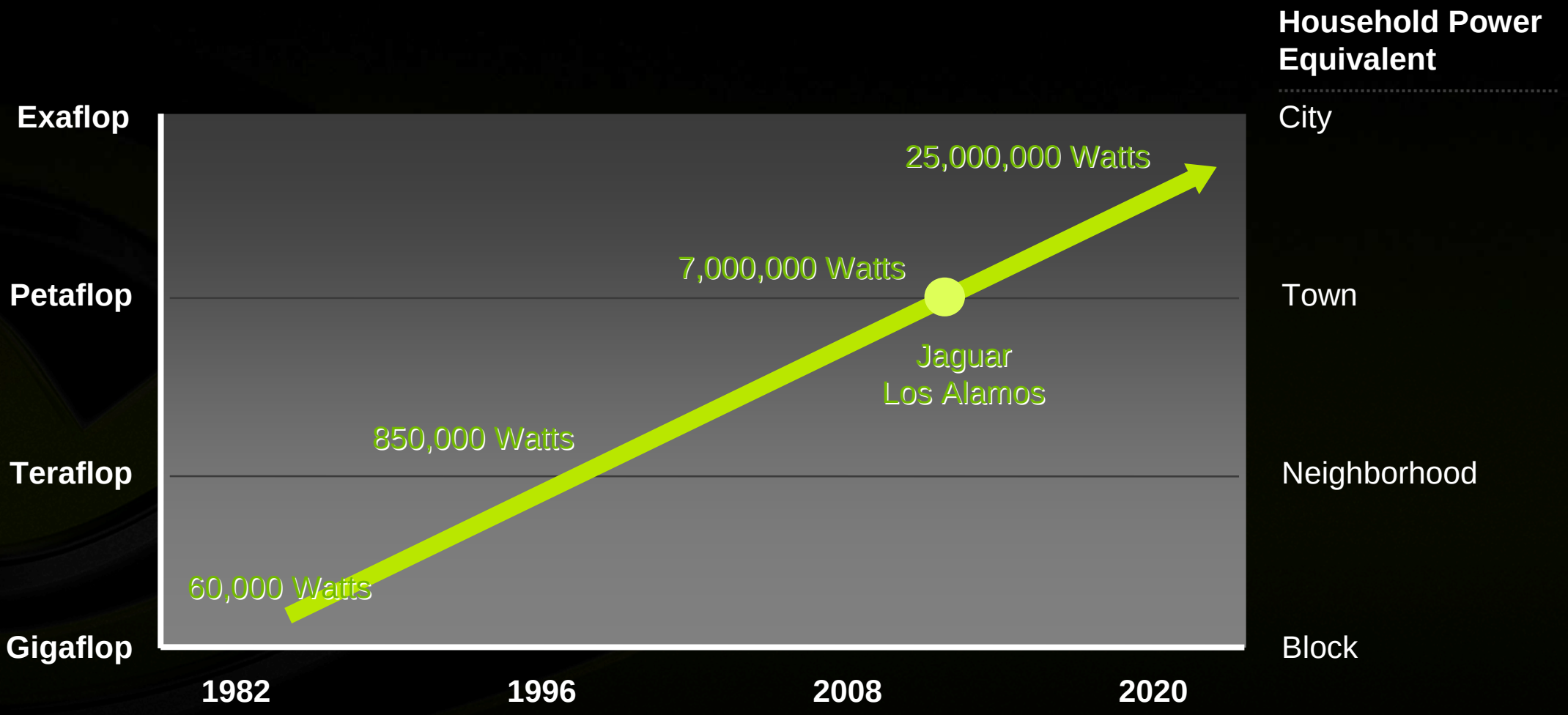

HIGH-PERFORMANCE
COMPUTING WITH NVIDIA
TESLA GPUS
Chris Butler NVIDIA
NVIDIA

Science is Desperate for Throughput



Power Crisis in Supercomputing



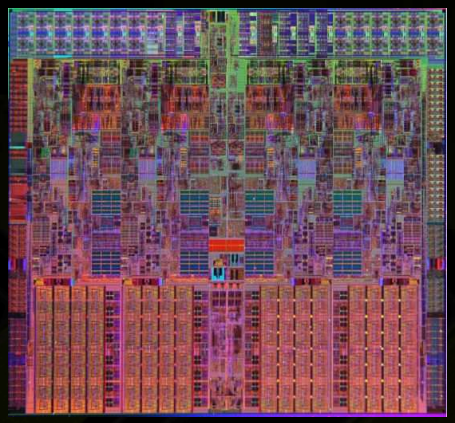
“Oak Ridge National Lab (ORNL) has already announced it will be using Fermi technology in an upcoming super that is ‘expected to be 10-times more powerful than today’s fastest supercomputer.’

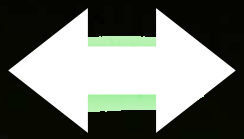
Since ORNL’s Jaguar supercomputer, for all intents and purposes, holds that title, and is in the process of being upgraded to 2.3 Petaflops …

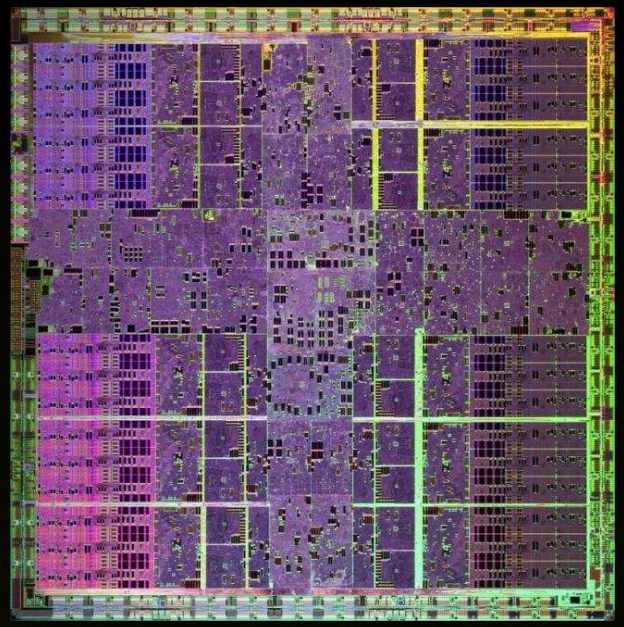
… we can surmise that the upcoming Fermi-equipped super is going to be in the 20 Petaflops range.”

September 30 2009

What is GPU Computing?







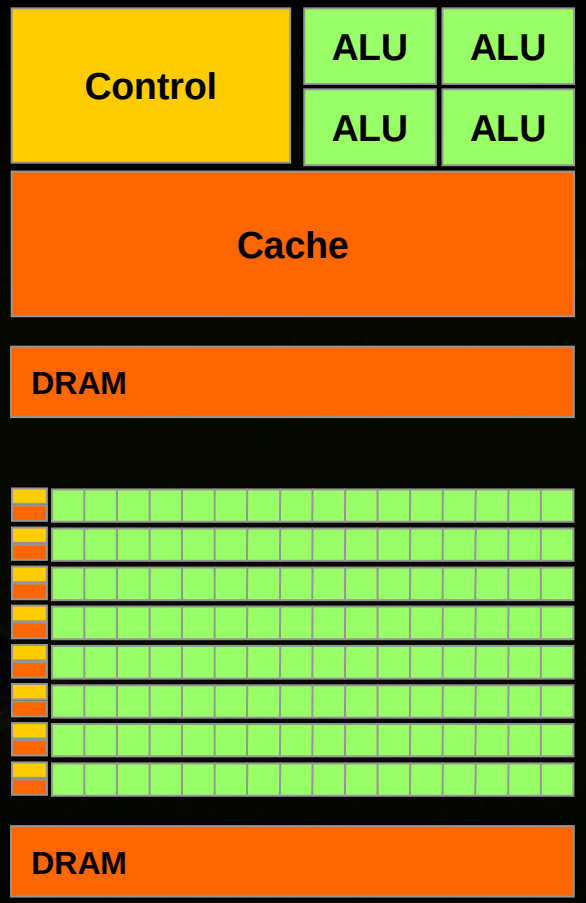
x86 PCIe bus GPU

Computing with CPU + GPU

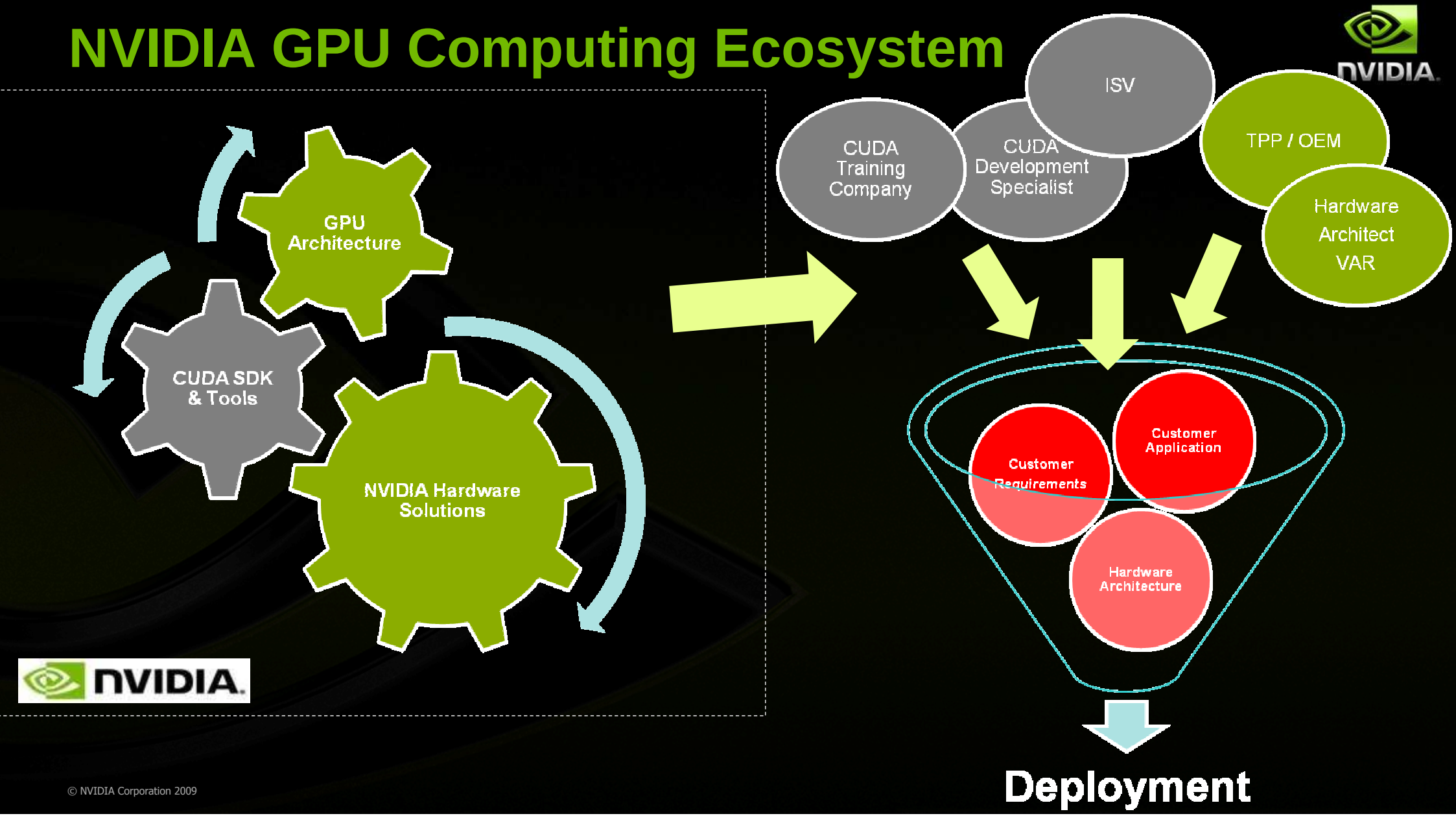
Heterogeneous Computing

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



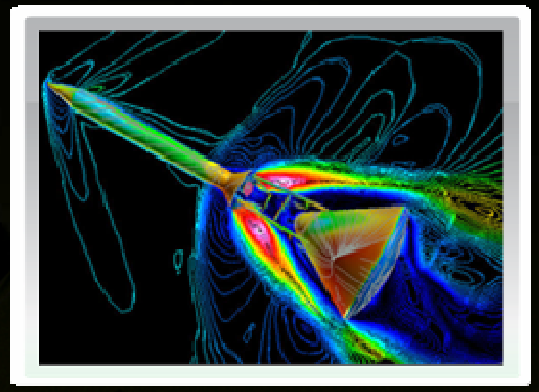
NVIDIA GPU Computing Ecosystem



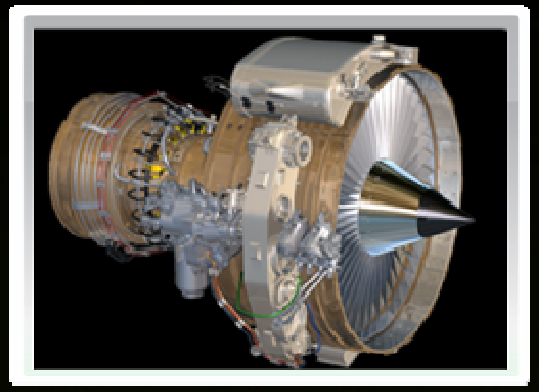
NVIDIA GPU Product Families

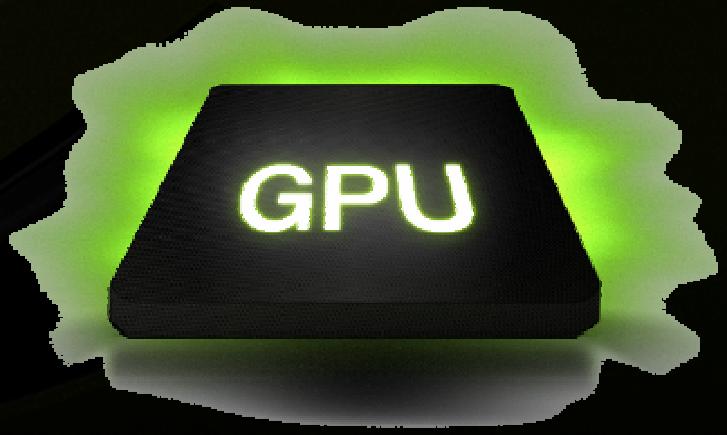
GeForce®
Entertainment

Tesla™
High-Performance Computing



Quadro®
Design & Creation

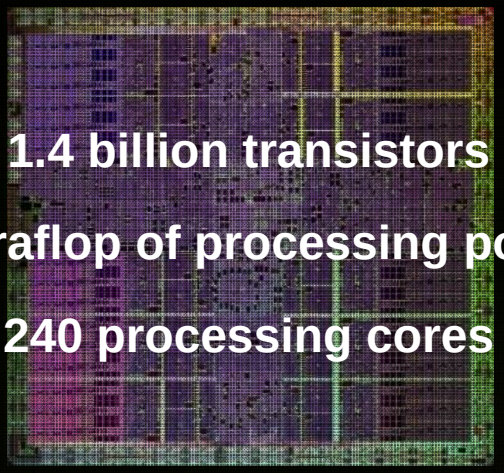




Many-Core High Performance Computing

- NVIDIA's 10-series GPU has 240 cores
- Each core has a
  - Floating point / integer unit
  - Logic unit
  - Move, compare unit
  - Branch unit
- Cores managed by thread manager
  - Thread manager can spawn and manage 30,000+ threads
  - Zero overhead thread switching

NVIDIA 10-Series GPU



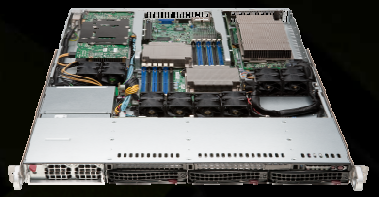
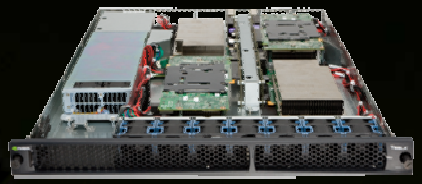
1.4 billion transistors

1 Teraflop of processing power

240 processing cores

NVIDIA's 2nd Generation CUDA Processor

Tesla GPU Computing Products

	SuperMicro 1U GPU SuperServer	Tesla S1070 1U System	Tesla C1060 Computing Board	Tesla Personal Supercomputer
				
GPUs	**2 Tesla GPUs**	**4 Tesla GPUs**	**1 Tesla GPU**	**4 Tesla GPUs**
Single Precision Performance	1.87 Teraflops	4.14 Teraflops	933 Gigaflops	3.7 Teraflops
Double Precision Performance	156 Gigaflops	346 Gigaflops	78 Gigaflops	312 Gigaflops
Memory	8 GB (4 GB / GPU)	16 GB (4 GB / GPU)	4 GB	16 GB (4 GB / GPU)

Tesla C1060 Computing Processor

Processor	1 × Tesla T10
Number of cores	240
Core Clock	1.296 GHz
Floating Point Performance	933 Gflops Single Precision
	78 Gflops Double Precision
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 ×16 Gen2
Typical power	160 W

Tesla M1060 Embedded Module

- OEM-only product
- Available as integrated product in OEM systems

Processor	1 × Tesla T10
Number of cores	240
Core Clock	1.296 GHz
Floating Point Performance	933 Gflops Single Precision
	78 Gflops Double Precision
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 ×16 Gen2
Typical power	160 W

Tesla Personal Supercomputer

Supercomputing Performance

- Massively parallel CUDA Architecture
- 960 cores. 4 Teraflops
- 250× the performance of a desktop

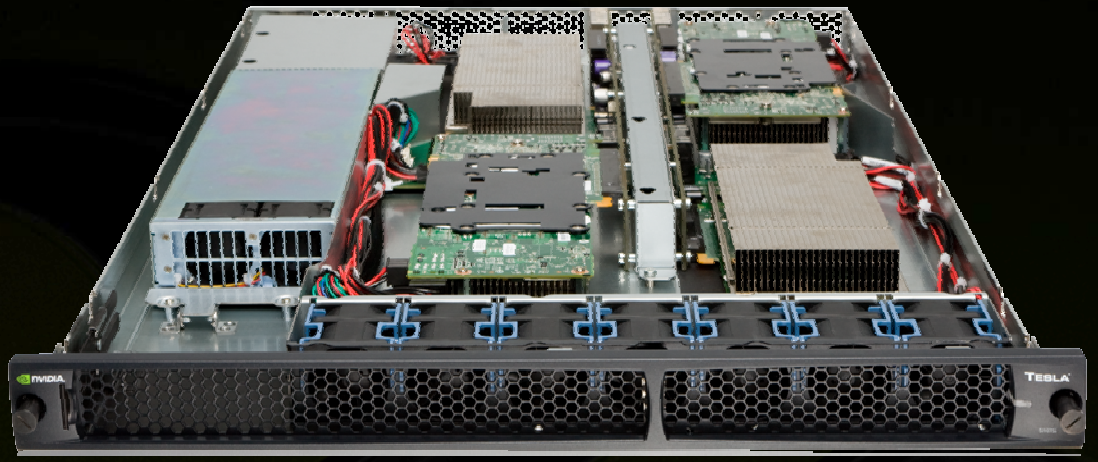
Personal

- One researcher, one supercomputer
- Plugs into standard power strip

Accessible

- Program in C for Windows, Linux
- Available now worldwide under $10,000

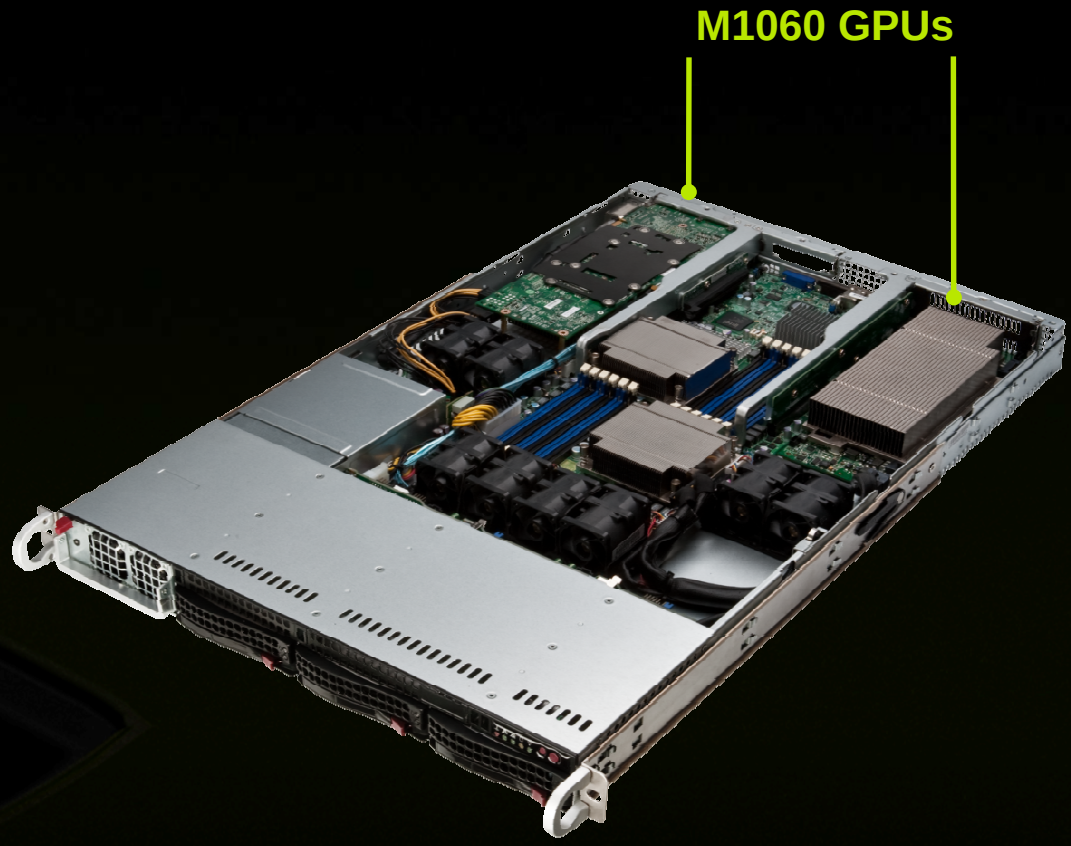
Tesla S1070 1U System



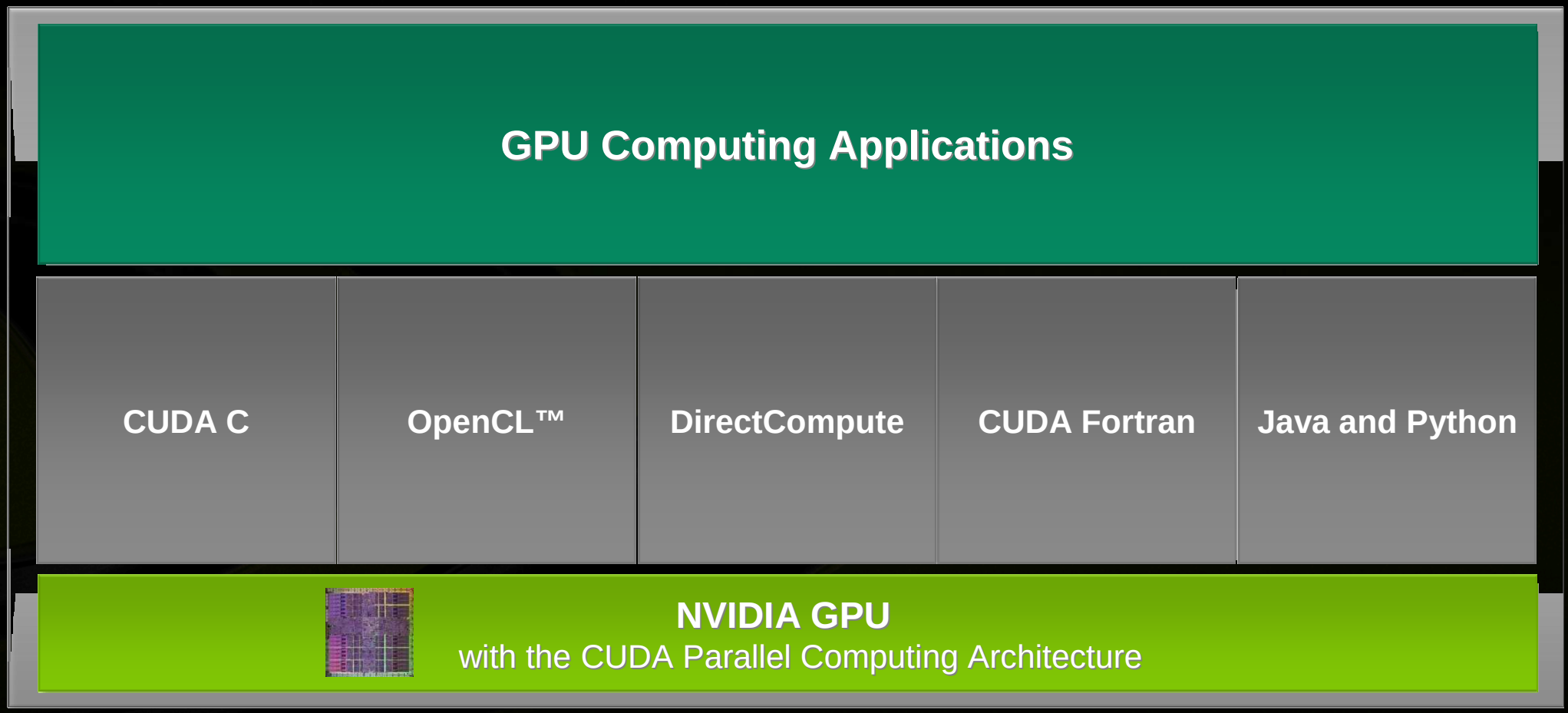
Processors	4 × Tesla T10
Number of cores	960
Core Clock	1.44 GHz
Performance	4 Teraflops
Total system memory	16.0 GB (4.0 GB per T10)
Memory bandwidth	408 GB/sec peak (102 GB/sec per T10)
Memory I/O	2048-bit, 800MHz GDDR3 (512-bit per T10)
Form factor	1U (EIA 19" rack)
System I/O	2 PCIe ×16 Gen2
Typical power	700 W

SuperMicro GPU 1U SuperServer

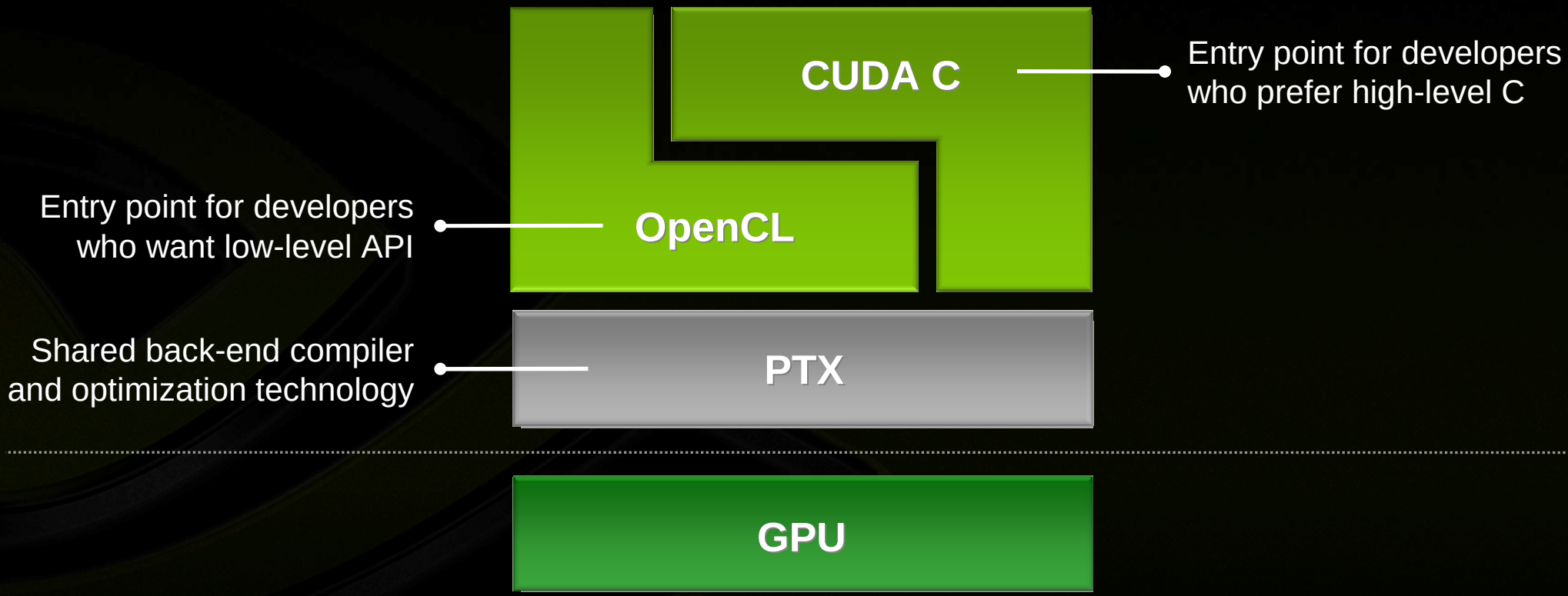
- Two M1060 GPUs in a 1U
- Dual Nehalem-EP Xeon CPUs
- Up to 96 GB DDR3 ECC
- Onboard Infiniband (QDR)
- 3× hot-swap 3.5″ SATA HDD
- 1200 W power supply



CUDA Parallel Computing Architecture

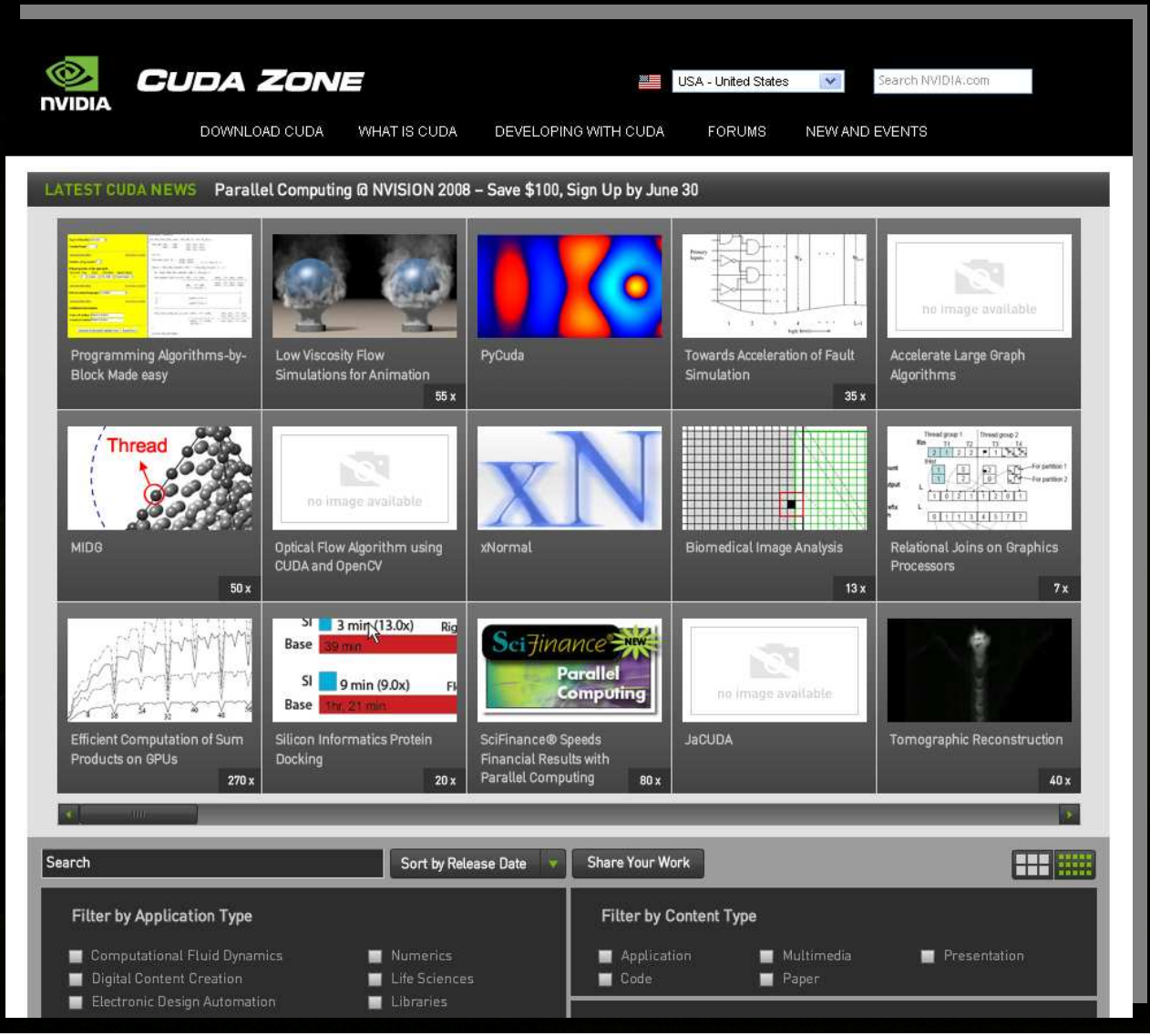


NVIDIA CUDA C and OpenCL



CUDA Zone: www.nvidia.com/CUDA

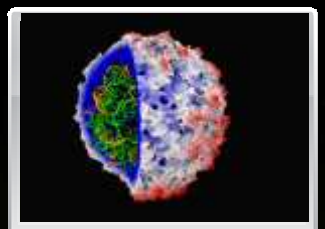
- CUDA Toolkit
  - Compiler
  - Libraries
- CUDA SDK
  - Code samples
- CUDA Profiler
- Forums
- Resources for CUDA developers



Wide Developer Acceptance and Success

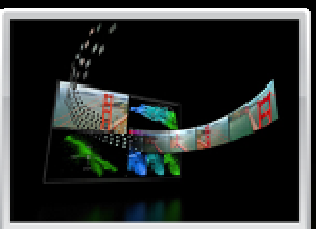
146X

Interactive visualization of volumetric white matter connectivity



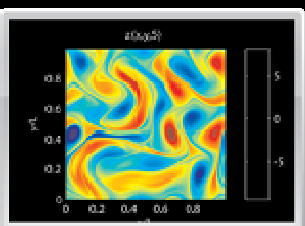
36X

Ion placement for molecular dynamics simulation



19X

Transcoding HD video stream to H.264

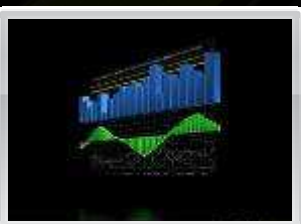


17X

Simulation in Matlab using .mex file CUDA function

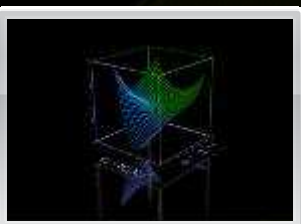
100X

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

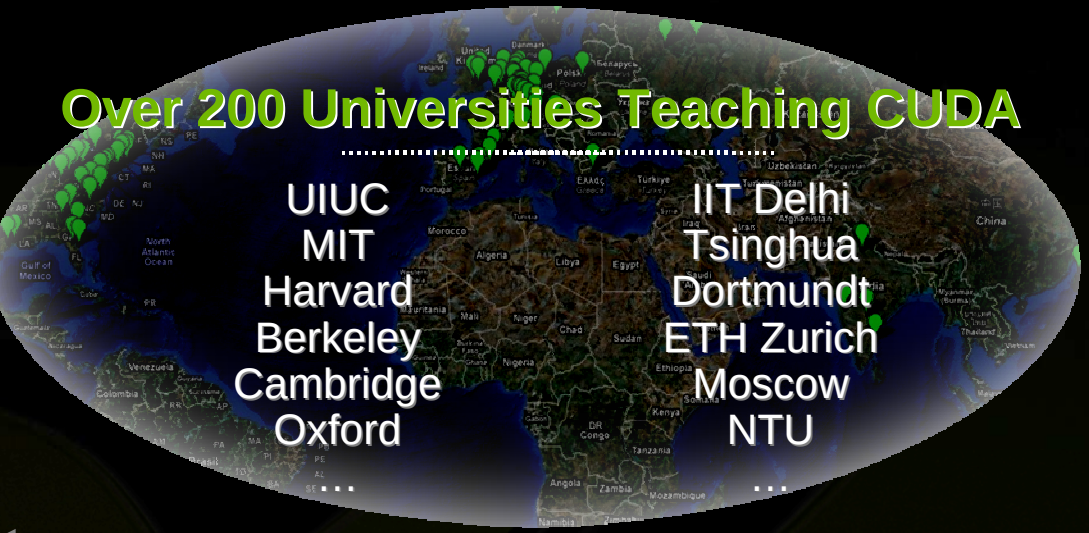
Highly optimized object oriented molecular dynamics

30X

Cmatch exact string matching to find similar proteins and gene sequences

CUDA Co-Processing Ecosystem

Over 200 Universities Teaching CUDA



UIUC
MIT
Harvard
Berkeley
Cambridge
Oxford
...

IIT Delhi
Tsinghua
Dortmundt
ETH Zurich
Moscow
NTU
...

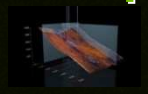
Languages

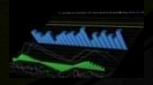
C, C++
DirectX
Fortran
Java
OpenCL
Python

Compilers

PGI Fortran
CAPs HMPP
MCUDA
MPI
NOAA Fortran2C
OpenMP

Applications


Oil & Gas

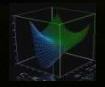

Finance


CFD

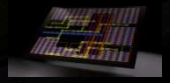

Medical


Biophysics


Imaging


Numerics


DSP


EDA

Libraries

FFT
BLAS
LAPACK
Image processing
Video processing
Signal processing
Vision

Consultants

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

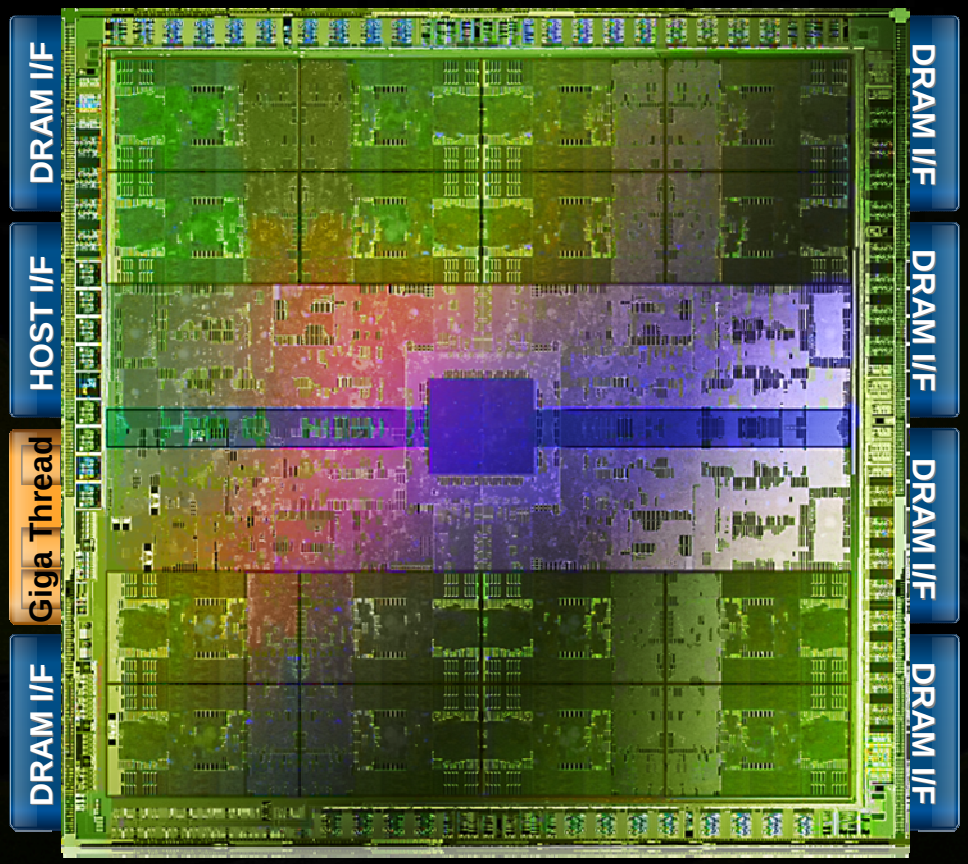
OEMs

DELL
hp
Penguin Computing
CRAY
Sun
Supermicro
sgi
FUJITSU
BULL
APPRO
lenovo
NEC

NEXT-GENERATION GPU ARCHITECTURE — ‘FERMI’

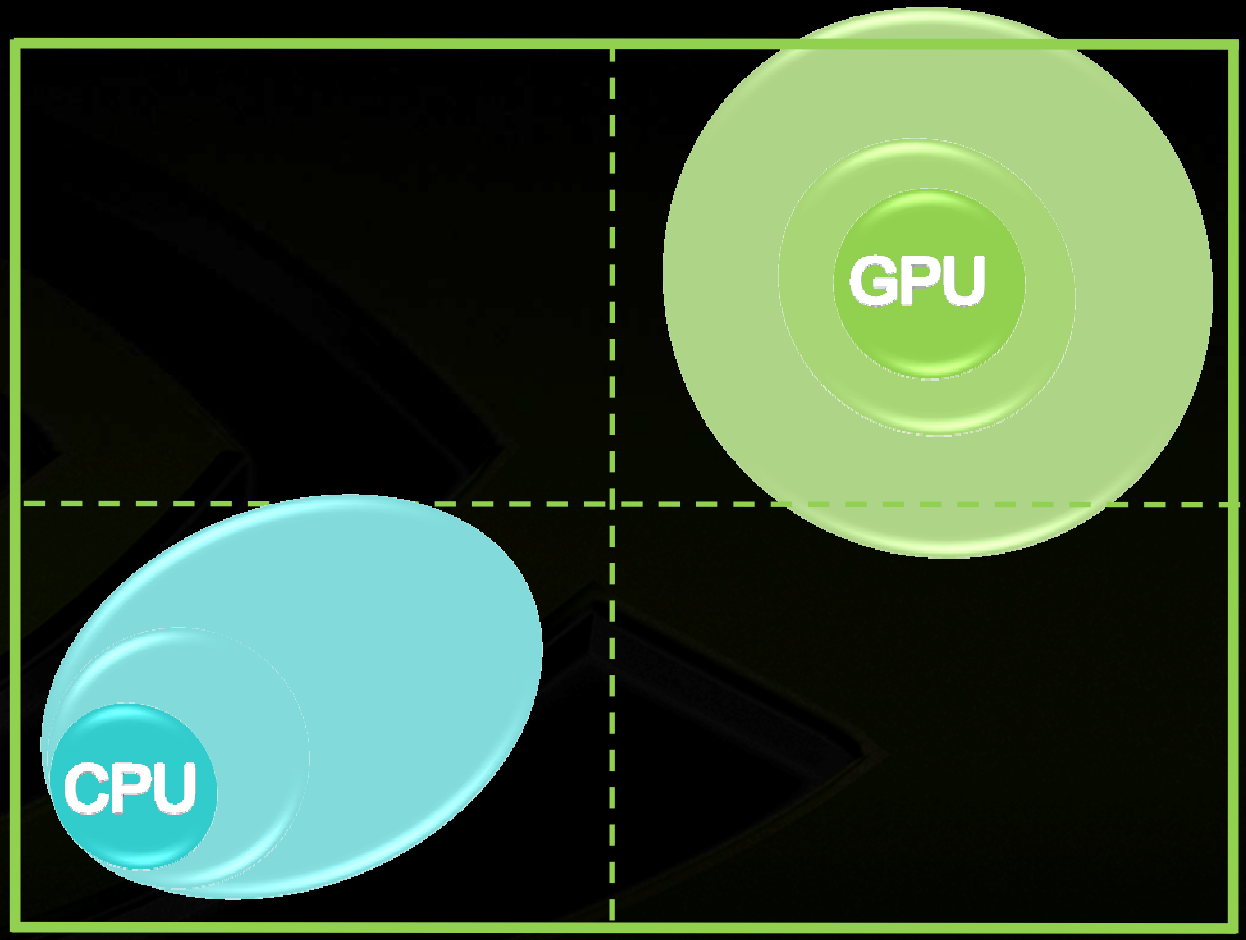
Introducing the ‘Fermi’ Architecture

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++

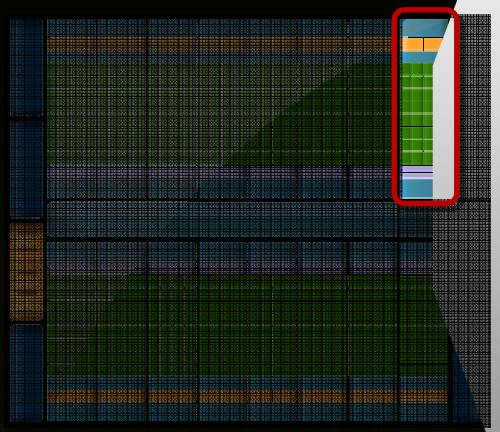
Design Goal of Fermi

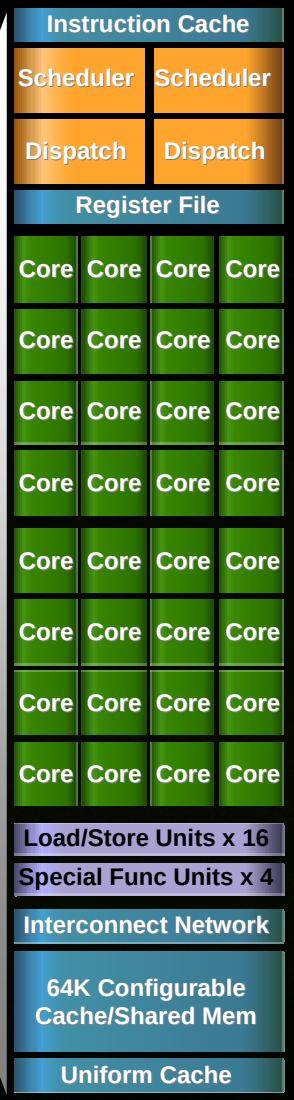


- Expand performance sweet spot of the GPU
- Bring more users, more applications to the GPU

Streaming Multiprocessor Architecture

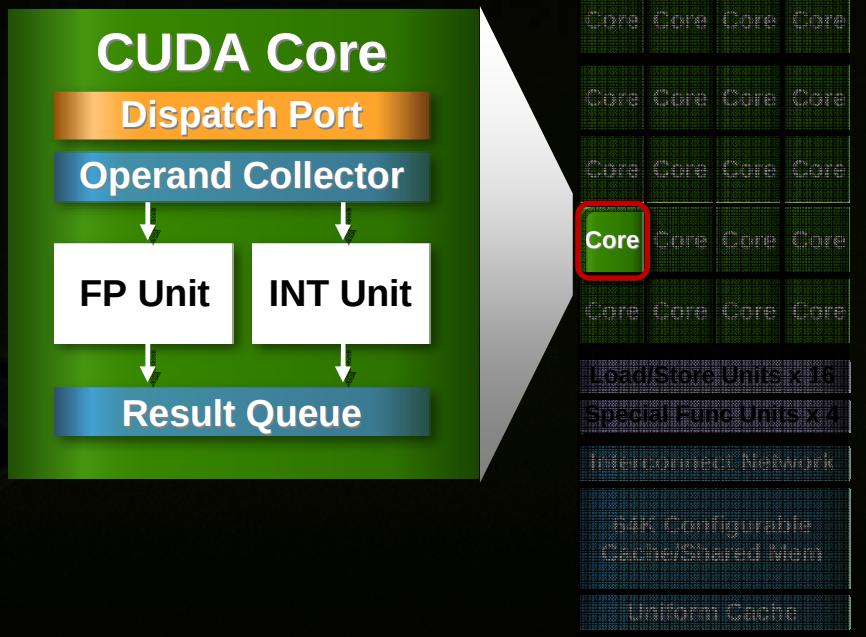
- 32 CUDA cores per SM (512 total)
- 8× peak double precision floating point performance
  - 50% of peak single precision
- Dual Thread Scheduler
- 64 KB of RAM for shared memory and L1 cache (configurable)





CUDA Core Architecture

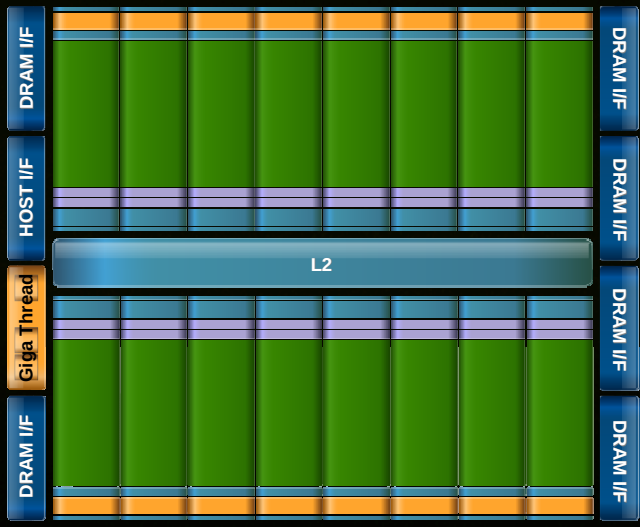
- New IEEE 754-2008 floating-point standard, surpassing even the most advanced CPUs
- Fused multiply-add (FMA) instruction for both single and double precision
- Newly designed integer ALU optimized for 64-bit and extended precision operations



Cached Memory Hierarchy

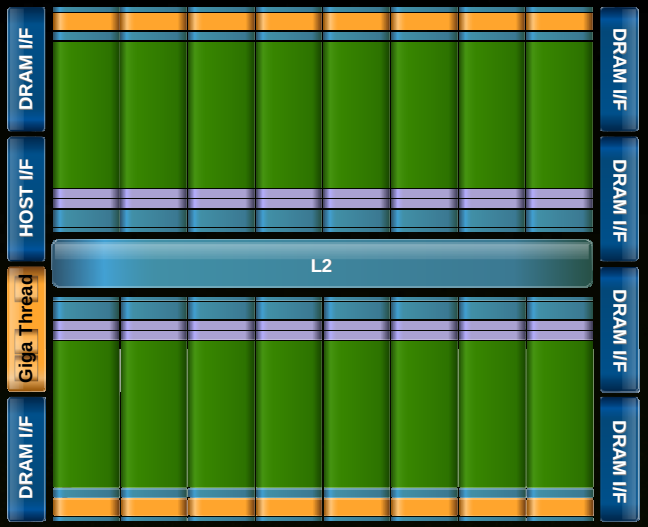
- First GPU architecture to support a true cache hierarchy in combination with on-chip shared memory
- L1 Cache per SM (32 cores)
  - Improves bandwidth and reduces latency
- Unified L2 Cache (768 KB)
  - Fast, coherent data sharing across all cores in the GPU

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

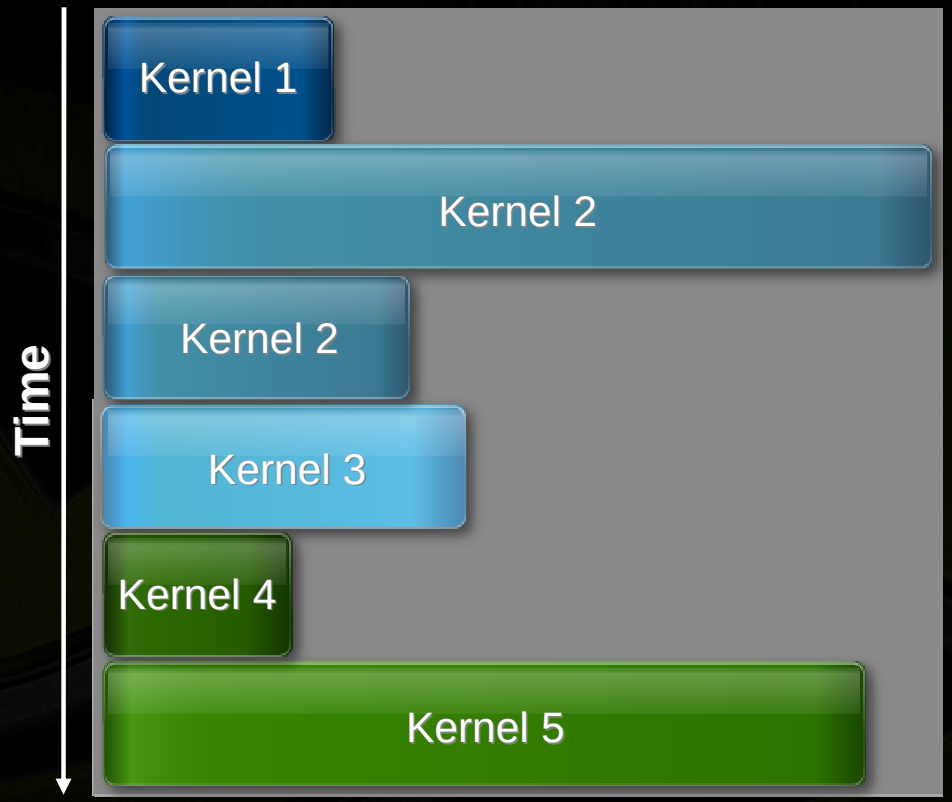


Error Correcting Code

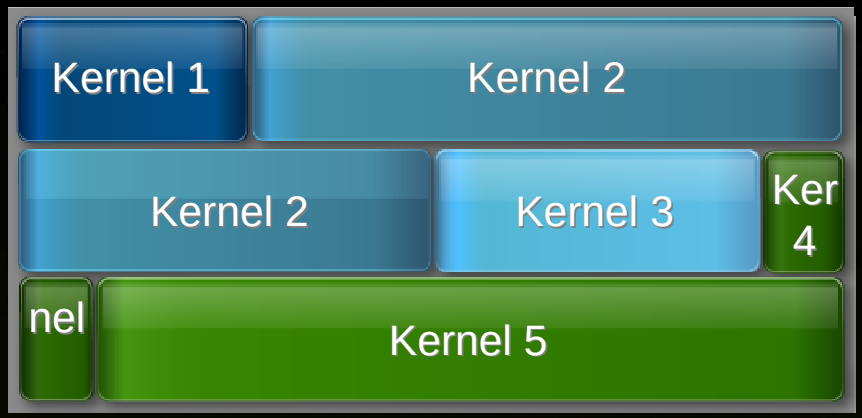
- **ECC protection for**
  - **DRAM**
    - ECC supported for GDDR5 memory
  - **All major internal memories are ECC protected**
    - Register file, L1 cache, L2 cache

GigaThread Hardware Thread Scheduler

Concurrent Kernel Execution + Faster Context Switch



Serial Kernel Execution



Parallel Kernel Execution

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

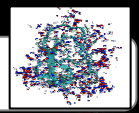
Introducing Tesla Bio WorkBench

NVIDIA

Applications

Amber 10 

GROMACS FAST. FLEXIBLE. FREE.

TeraChem 

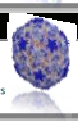
HMMER Scalable Informatics University at Buffalo The State University of New York

NAMD Scalable Molecular Dynamics 

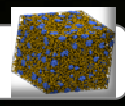
LAMMPS

ACEMD GPUGRID.net 

GPU-AutoDock

VMD Visual Molecular Dynamics 

GROMOS

HOOMD blue 

MUMmerGPU

Community

Download, Documentation

Technical papers

Discussion Forums

Benchmarks & Configurations

Platforms

Tesla Personal Supercomputer

Tesla GPU Clusters

Tesla Bio Workbench Applications

- AMBER (MD)
- ACEMD (MD)
- GROMACS (MD)
- GROMOS (MD)
- LAMMPS (MD)
- NAMD (MD)
- TeraChem (QC)
- VMD (Visualization MD & QC)

- Docking
  - GPU AutoDock
- Sequence analysis
  - CUDASW++ (SmithWaterman)
  - MUMmerGPU
  - GPU-HMMER
  - CUDA-MEME Motif Discovery

AMBER Molecular Dynamics

Alpha now
- **Generalized Born**

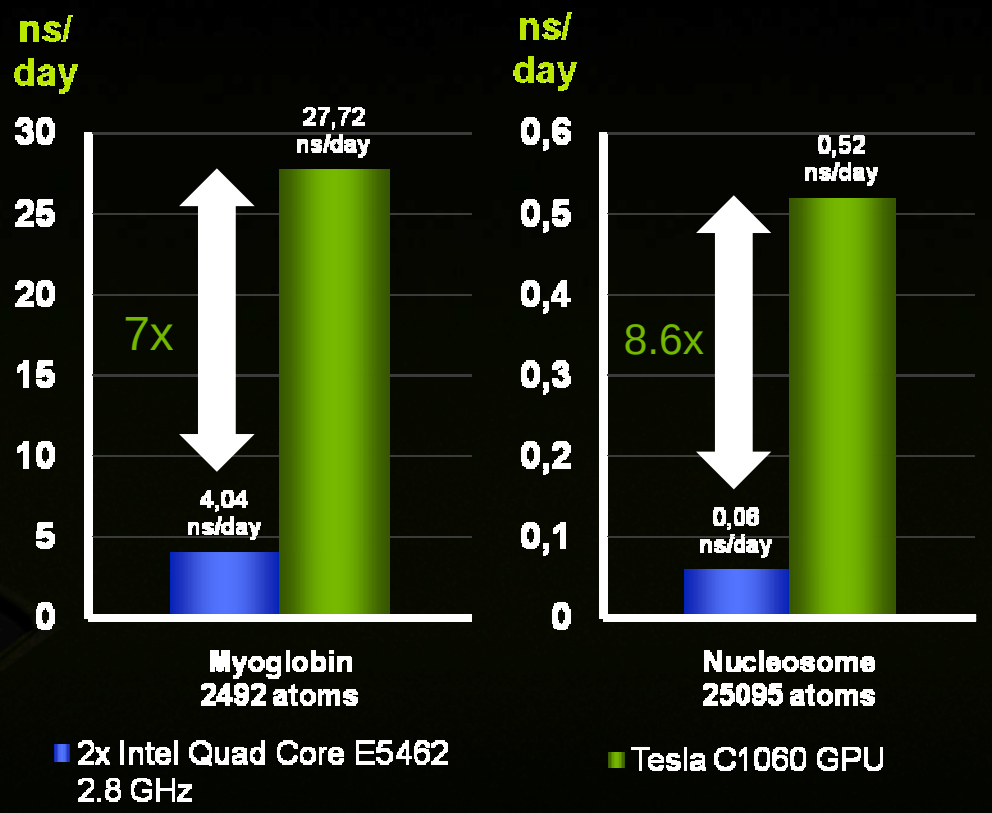
Q1 2010
- **PME: Particle Mesh Ewald**
- **Beta release**

Q2 2010
- **Multi-GPU + MPI support**
- **Beta 2 release**

- Implicit solvent GB results
  - 1 Tesla GPU 8x faster than 2 quad-core CPUs

More Info
http://www.nvidia.com/object/amber_on_tesla.html

Generalized Born Simulations



Data courtesy of San Diego Supercomputing Center

ISV status

Developer	Application	Description	Category	STATUS
ORNL	HOMME	High Order Method Modeling Environment	Government	Work in progress
Ames Lab	HOOMD	Molecular dynamics	Life Science	Available
NCI	AutoDock	Molecular dynamics	Life Science	Selected support from third parties, also open source project, need to reach out to them more
ORNL	MADNESS	Computational Chemistry	Life Science	Work in progress
	Smith-Waterman	DNA sequencing	Life Science	Various versions available such as http://gpu.epfl.ch/sw.html
Stanford	OpenMM	Molecular Library	Life Science	Available with PME in beta
Gaussian	Gaussian	Quant Chem	Life Science	Taking names, no date given

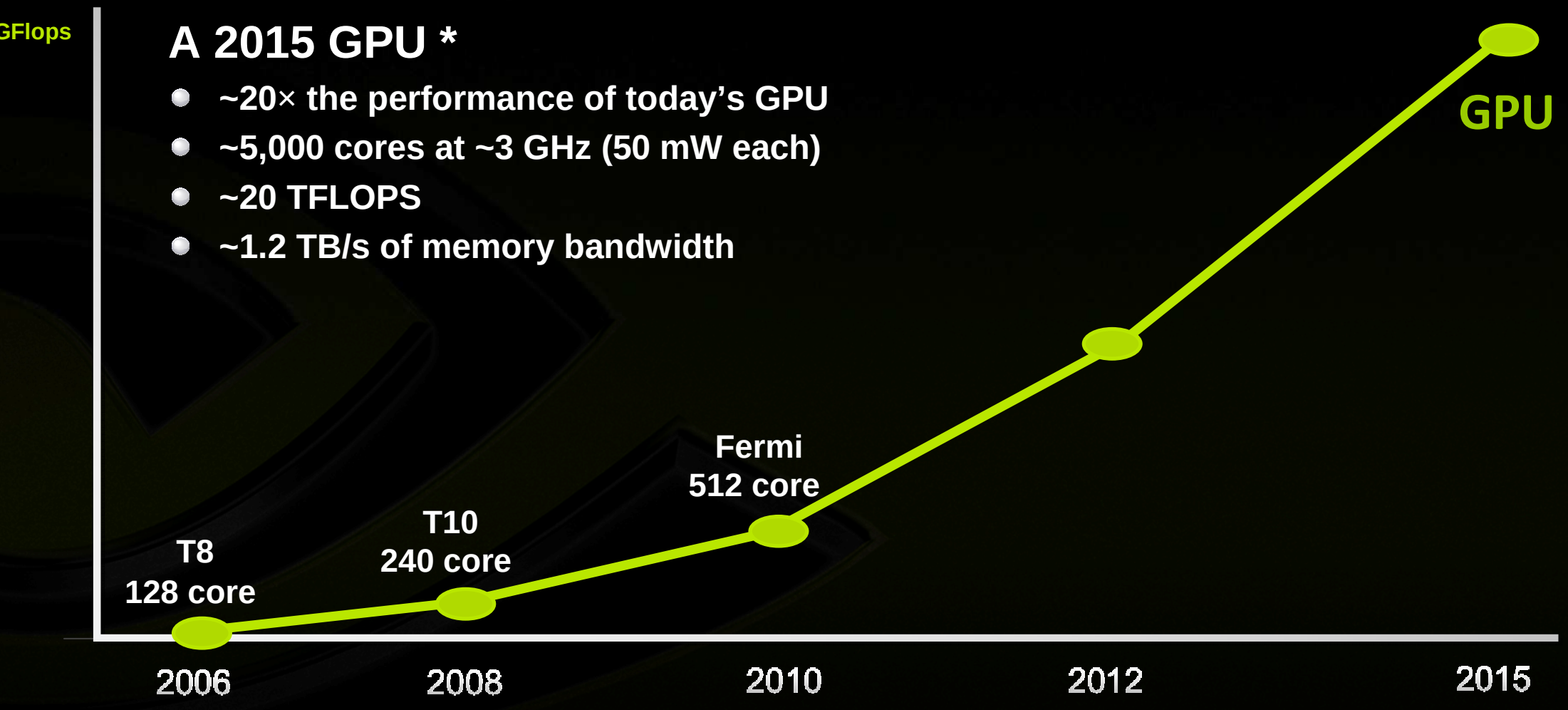
ISV Status:

Developer	Application	Description	Category	STATUS
Harvard and Univ of Delaware	CHARMM	Molecular dynamics	Life Science	Cuda support in library as part of alpha build, working with them to get a beta in Feb
Howard Hughes Med	HMMER	Hidden Markov models for bio	Life Science	Available
IA State	GAMESS	Quant Chem	Life Science	Integration in progress, hopeful of beta in Q1
Scripps	LAMMPS	Molecular dynamics	Life Science	Selected algorithms supported, project ongoing – download availalable
Scripps	AMBER	Molecular dynamics	Life Science	Amber 10 patch for GB today, PME by Dec/Jan
Scripps	AutoDock	Protien Docking	Life Science	Available via 3rd party

ISV Status:

Developer	Application	Description	Category	STATUS
Stockholm Center	GROMACS	Molecular dynamics	Life Science	Available with PME via OpenMM
UIUC	NAMD	Molecular dynamics	Life Science	Namd 2.7 B2 available
UIUC	VMD	Viz of MD	Life Science	Available
Univ of Delaware	Dockig@home	Protein docking	Life Science	Not sure.
Univ of Maryland	MUMmerGPU	DNA sequence alignment	Life Science	Available
Allinea	Allinea DDT	Linux Debugger	Tools - Debug/Profile	Beta this month
TotalView	TotalView Debugger	Linux Debugger	Tools - Debug/Profile	Beta in Q1

GPU Revolutionizing Computing



* This is a sketch of a what a GPU in 2015 might look like; it does not reflect any actual product plans.

TESLA™