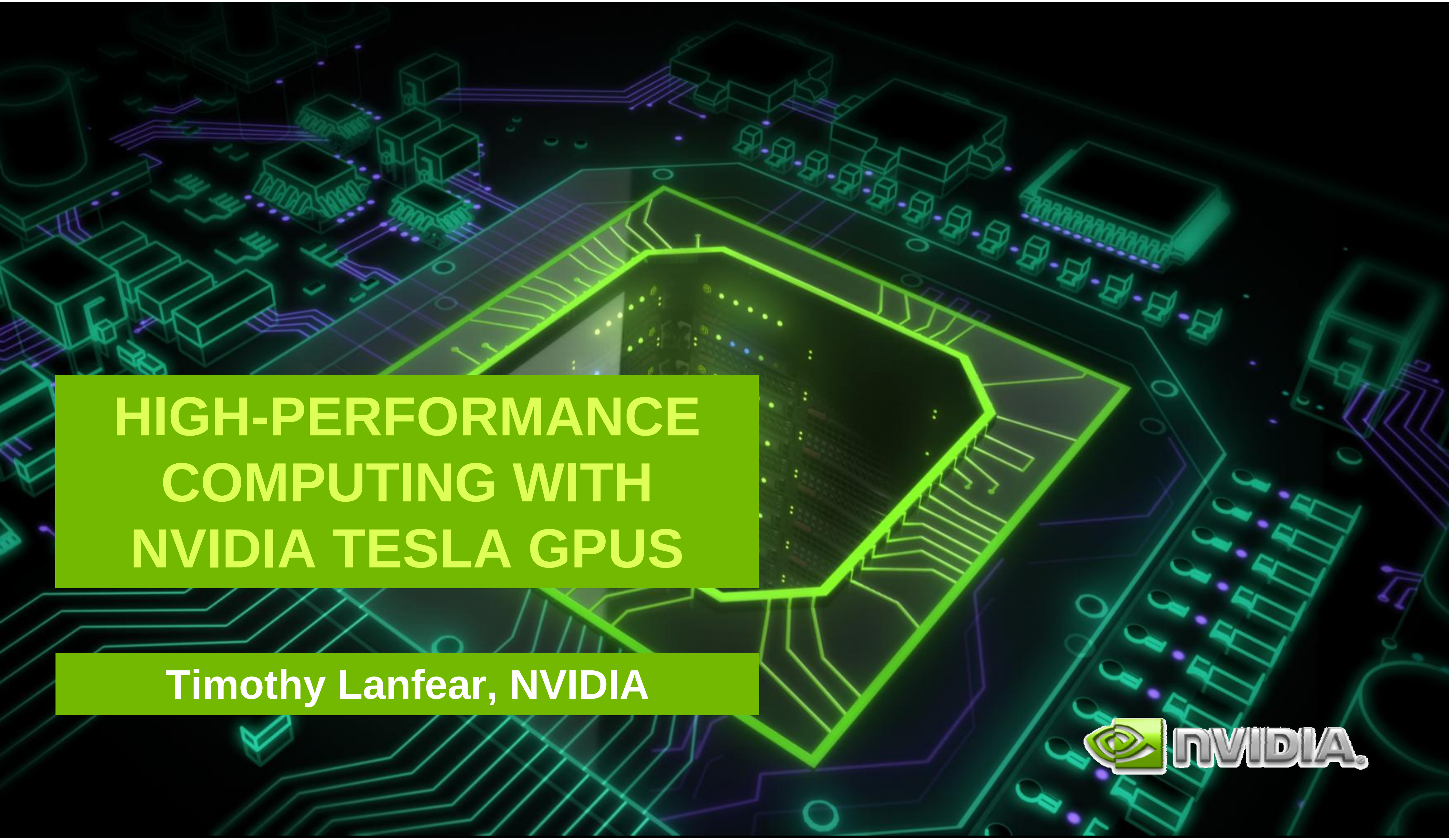
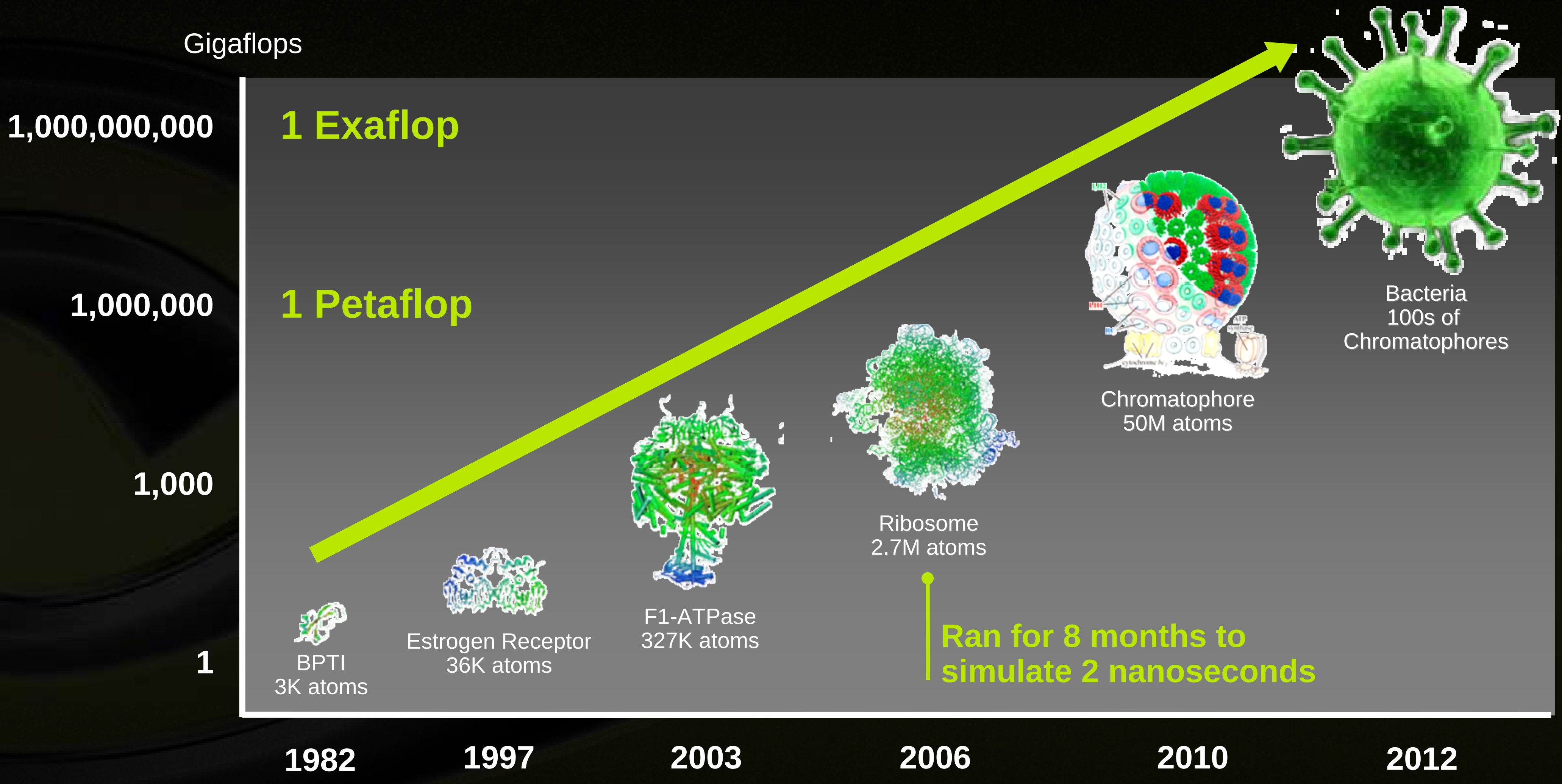


HIGH-PERFORMANCE COMPUTING WITH NVIDIA TESLA GPUS

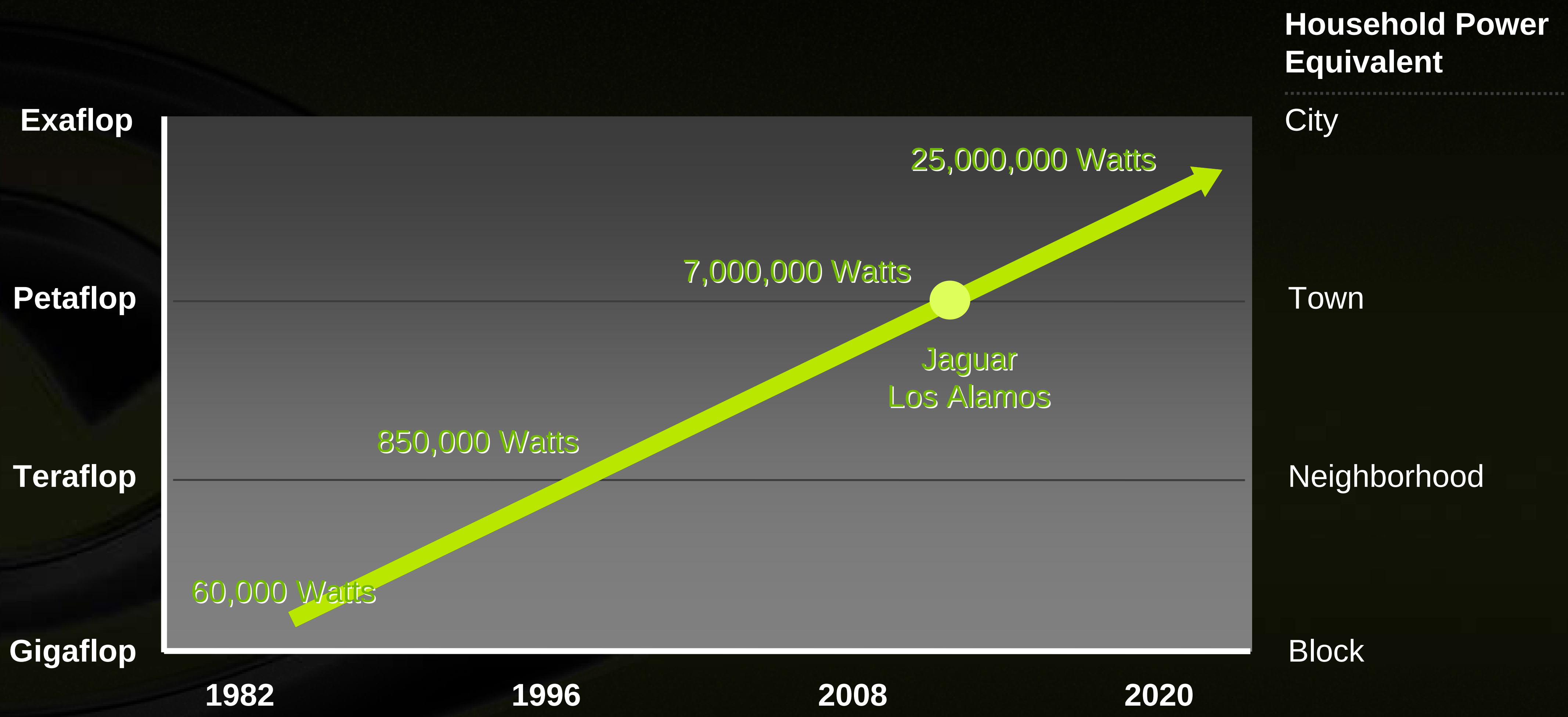
Timothy Lanfear, NVIDIA

WHY GPU COMPUTING?

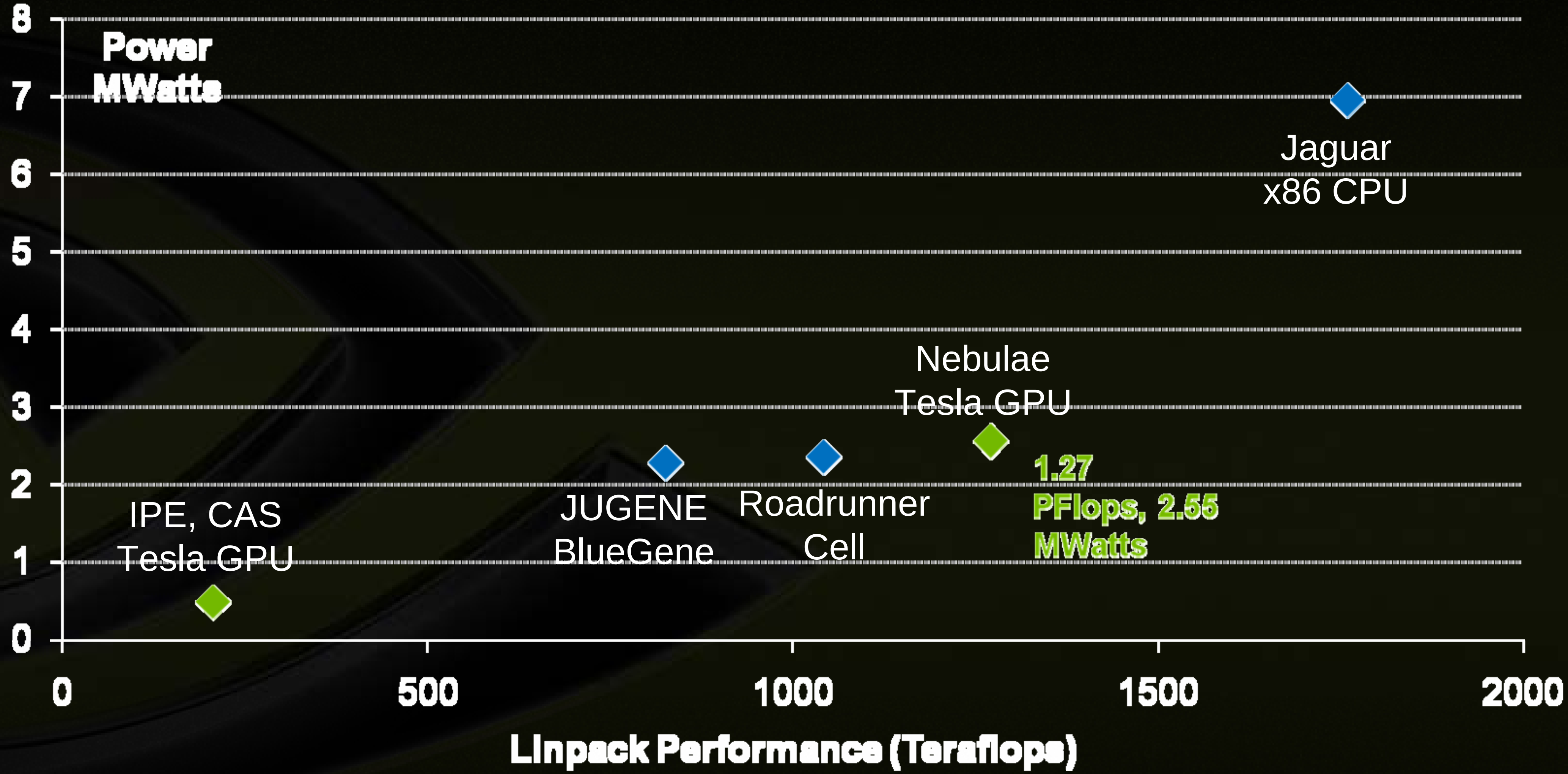
Science is Desperate for Throughput



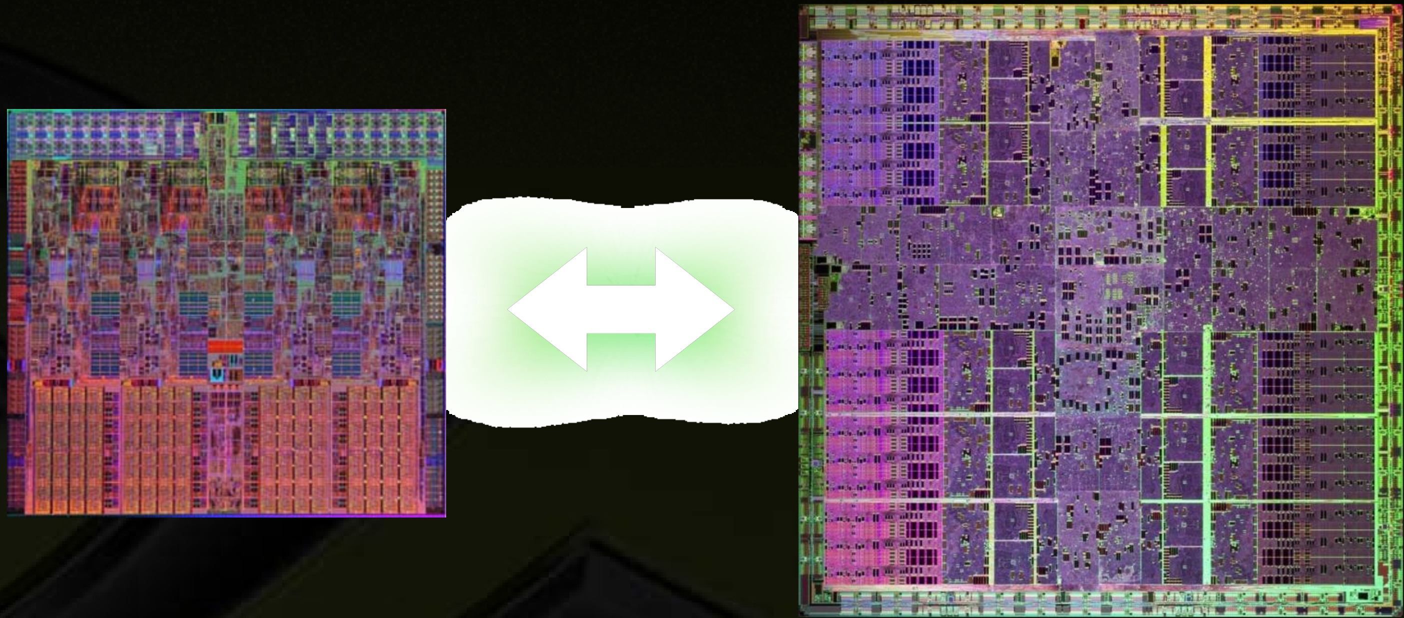
Power Crisis in Supercomputing



Double Performance per Watt



What is GPU Computing?

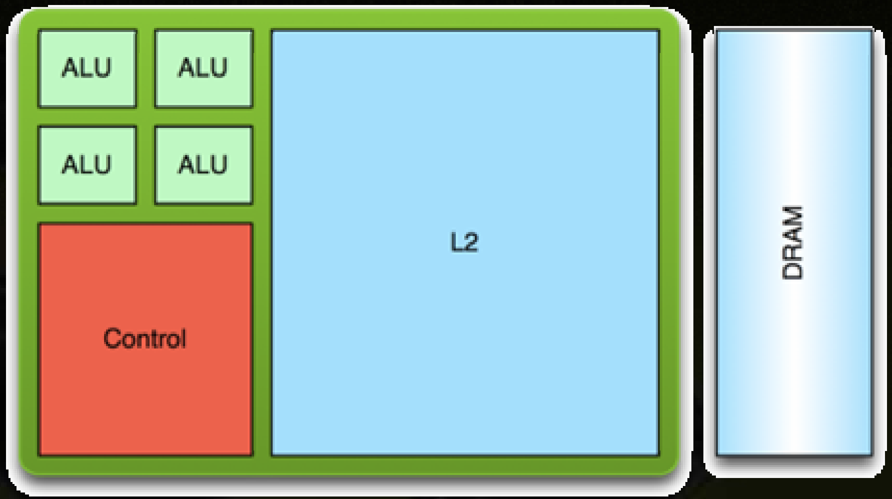


x86 PCIe bus GPU

Computing with CPU + GPU

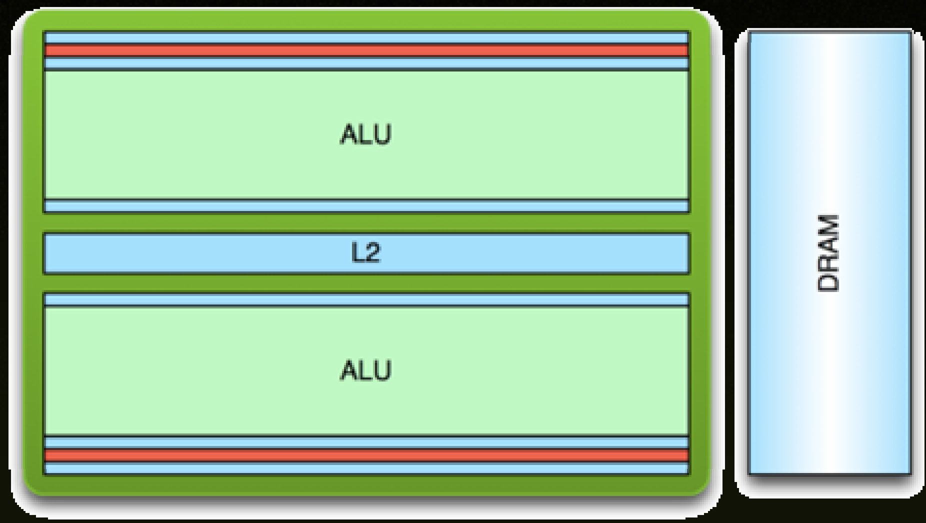
Heterogeneous Computing

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

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 : Leadership in GPU Computing

NVIDIA

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
Nexus
MCUDA
MPI
NOAA Fortran2C
OpenMP

CUDA

Applications

Oil & Gas
Finance
CFD
Medical
Biophysics
Imaging
Numerics
DSP
EDA

Libraries

FFT
BLAS
LAPACK
Image processing
Video processing
Signal processing
Vision

Consultants

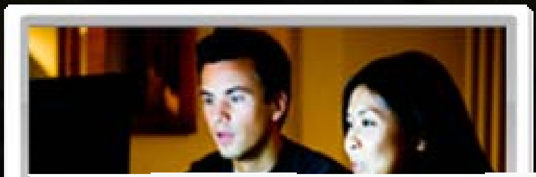
acceleware
HPC Project
ANEO
Scalable Graphics
CAPS
Stone Ridge Technology
EM Photonics
Tech-X
GASS
GPU Tech
Wipro

OEMs

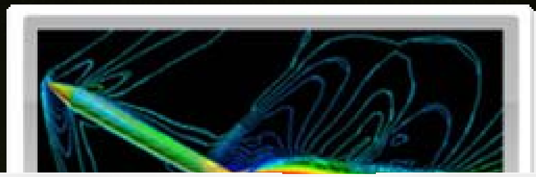
Dell
HP
Penguin Computing
Cray
IBM
Supermicro
SGI
Fujitsu
Bull
Appro
Lenovo
NEC

NVIDIA GPU Product Families

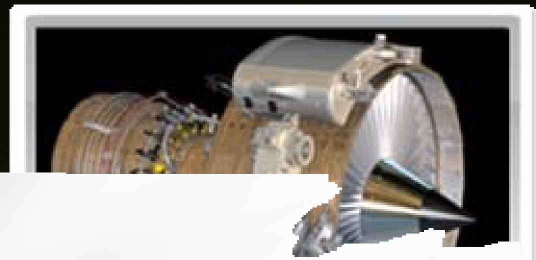
GeForce®
Entertainment

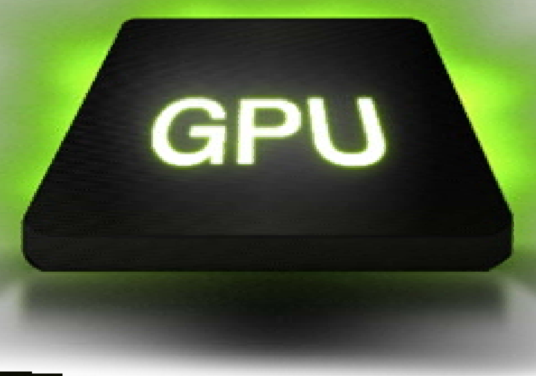


Tesla™
High-Performance Computing



Quadro®
Design & Creation





Fermi: The Computational GPU

Performance

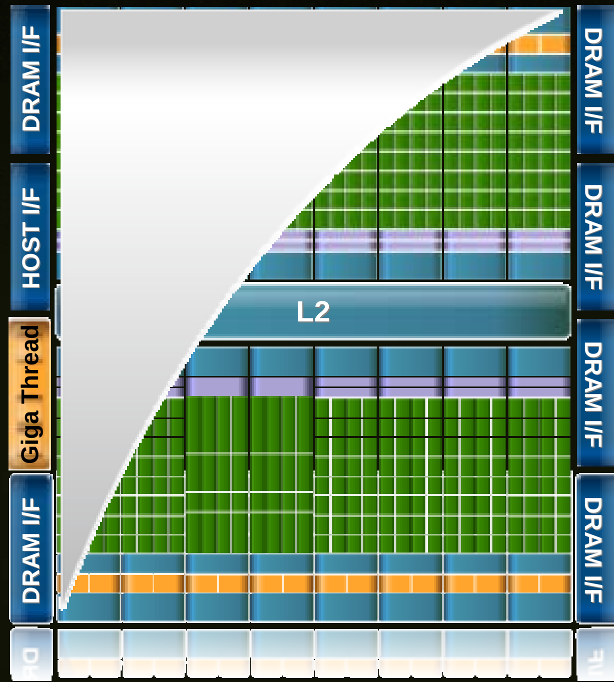
- More than ½ Teraflop 64-bit performance
- IEEE 754-2008 SP and DP Floating Point

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

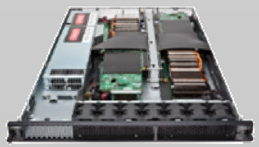
- Multiple Simultaneous Tasks on GPU
- 10× Faster Atomic Operations
- C++ Support
- System Calls, printf support



NVIDIA Tesla GPU Computing Products

Data Center Products

Workstation



	Tesla M1060	Tesla M2050	Tesla S2070	Tesla 2050	Tesla S1070	Tesla C2070	Tesla C2050	Tesla C1060
GPUs	**1 T10 GPU**	**1 T20 GPU**	**4 T20 GPUs**		**4 T10 GPUs**	**1 T20 GPU**		**1 T10 GPU**
Single Precision	933 GFlops	1030 GFlops	4120 GFlops		4140 GFlops	1030 Gflops		933 GFlops
Double Precision	78 GFlops	515 GFlops	2060 GFlops		346 GFlops	515 Gflops		78 GFlops
Memory	4 GB	3 GB	12 GB 3 GB / GPU	24 GB 6 GB / GPU	16 GB 4 GB / GPU	6 GB	3 GB	4 GB
Mem BW	102 GB/s	148.4 GB/s	148.4 GB/s		102 GB/s	144 GB/s		102 GB/s
Display	No display IO		No display IO			Single dual-link DVI		No display IO

OEM Servers with Tesla M2050 GPUs

Announced on May 4th, 2010

SUPERMICRO

2 Tesla M2050 GPUs

SuperServer 6016GT-TF
2 CPUs + 2 GPUs in 1U

APPRO

4 Tesla M2050 GPUs

Appro Tetra
2 CPUs + 4 GPUs in 1U

10 Tesla M2050 GPUs

Appro GreenBlade
10 CPUs + 10 GPUs in 5U

TYAN®

8 Tesla M2050 GPUs

Tyan B7015
2 CPUs + 8 GPUs in 4U

….. many more coming soon ……

OEM Servers with Tesla M1060 GPUs

NVIDIA

SUPERMICRO

CRAY THE SUPERCOMPUTER COMPANY

Bull

TYAN®

2 Tesla
M1060 GPUs

SuperServer 6016GT-TF
2 CPUs + 2 GPUs in 1U

Upto 18 Tesla
M1060 GPUs

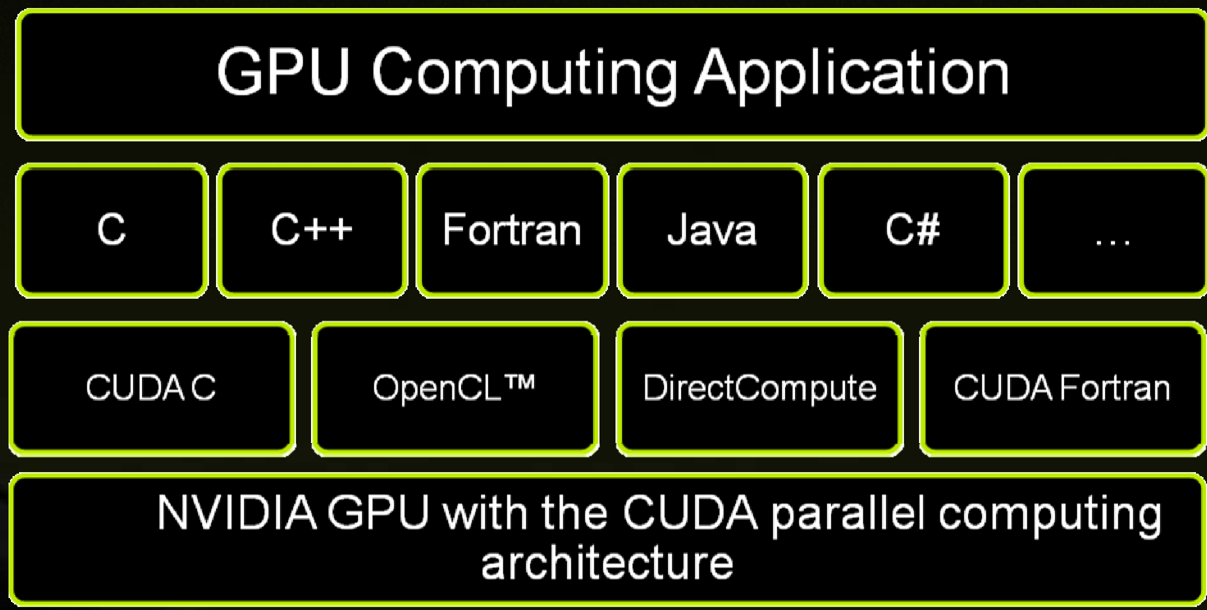
Cray CX1000 and Bull Bullx
36 CPUs + 18 GPUs in 7U

8 Tesla
M1060 GPUs

Tyan B7015
2 CPUs + 8 GPUs in 4U

CUDA Parallel Computing Architecture

- **Parallel computing architecture and programming model**
- **Includes a CUDA C compiler, support for OpenCL and DirectCompute**
- **Architected to natively support multiple computational interfaces (standard languages and APIs)**



Application Software (written in C)

CUDA Libraries

cuFFT | cuBLAS | cuDPP

CPU Hardware

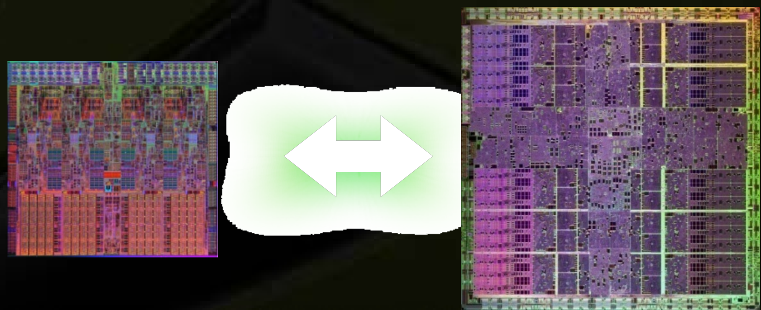
1U   PCI-E Switch

CUDA Compiler

C   Fortran

CUDA Tools

Debugger   Profiler

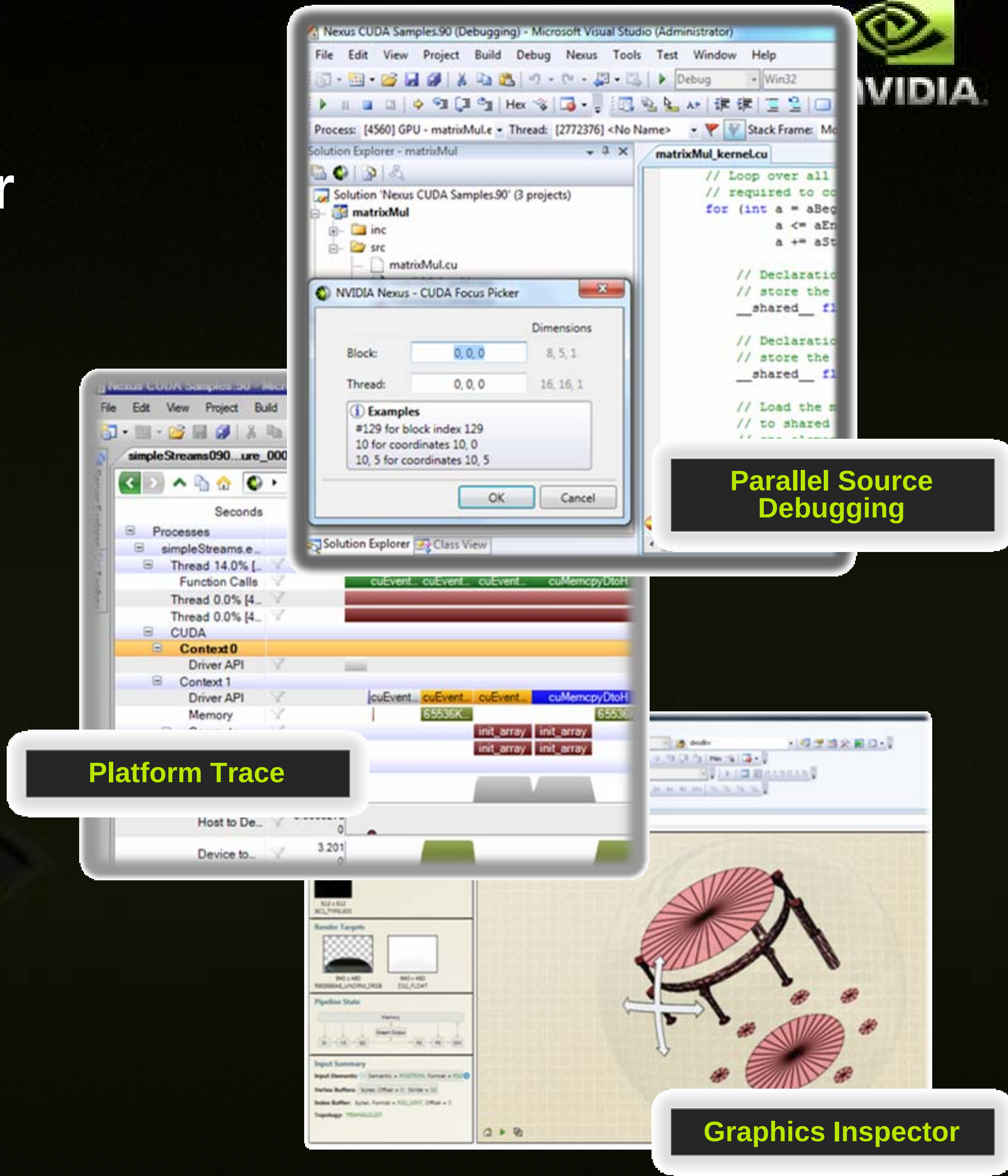


4 cores   240 cores

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

Allinea DDT

GPU Debugging

Making it easy

Allinea DDT — CUDA Enabled

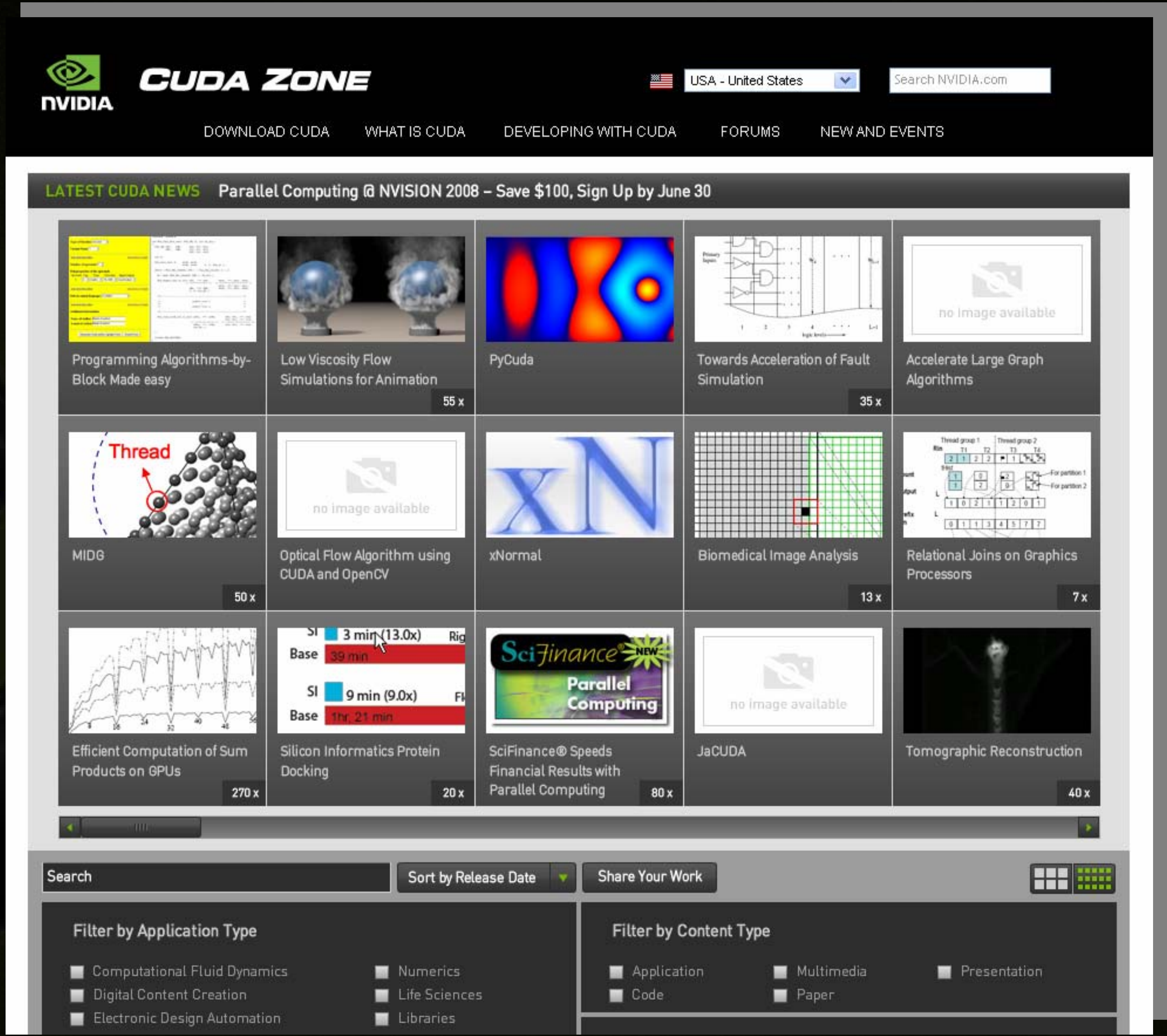
TotalView for CUDA

TotalView for CUDA

TOTALVIEW TECHNOLOGIES

CUDA Zone: www.nvidia.com/CUDA

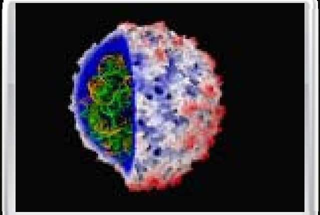
- CUDA Toolkit
  - Compiler
  - Libraries
- CUDA SDK
  - Code samples
- CUDA Profiler
- Forums
- Resources for CUDA developers



Wide Developer Acceptance and Success

146X

Interactive visualization of volumetric white matter connectivity

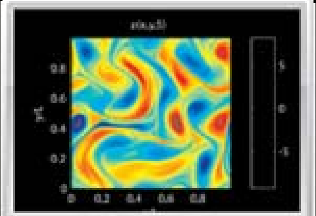


36X

Ion placement for molecular dynamics simulation

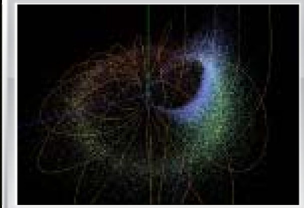
19X

Transcoding HD video stream to H.264



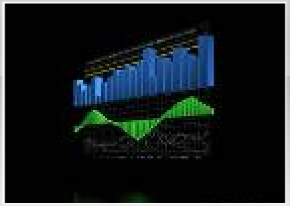
17X

Simulation in Matlab using .mex file CUDA function



100X

Astrophysics N-body simulation

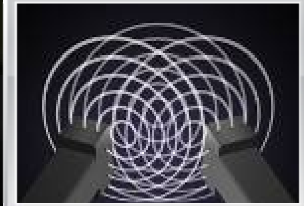


149X

Financial simulation of LIBOR model with swaptions

47X

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

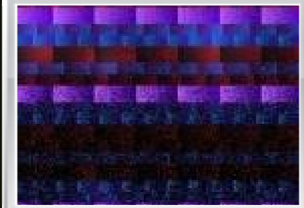


20X

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

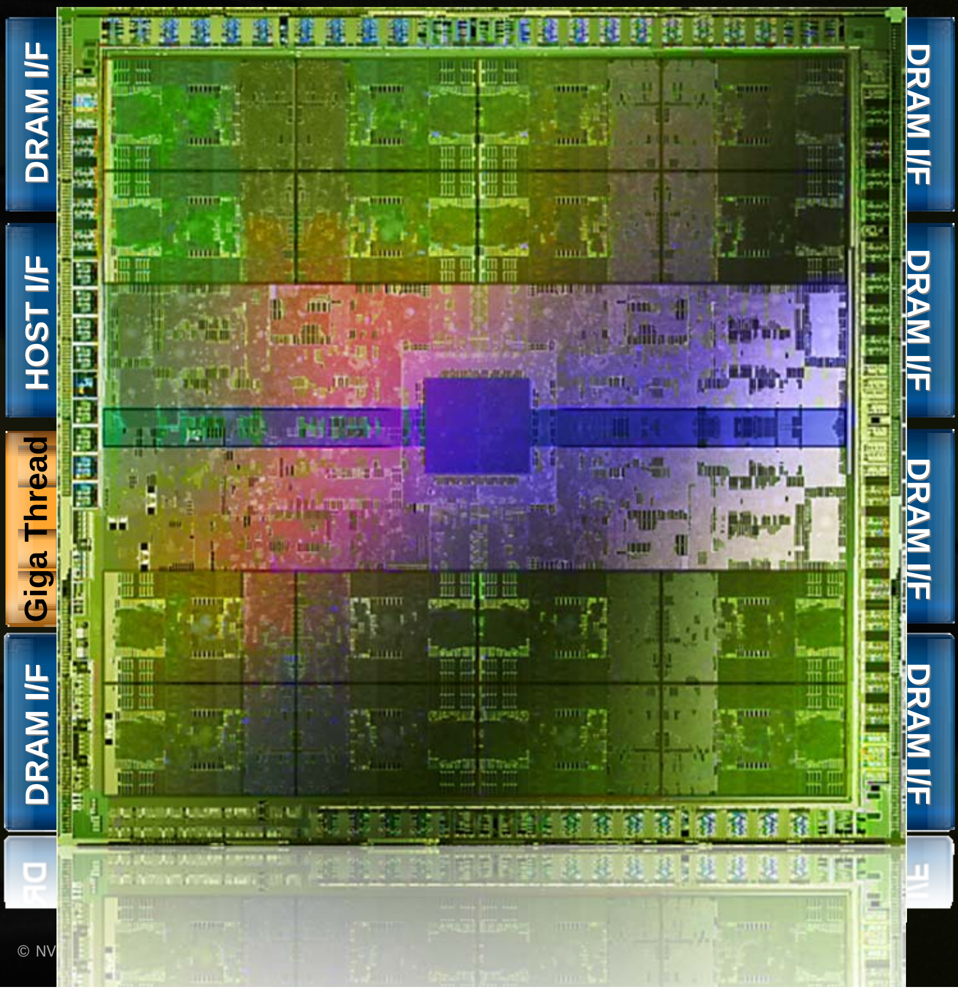
What We Did in the Past Three Years

- 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’

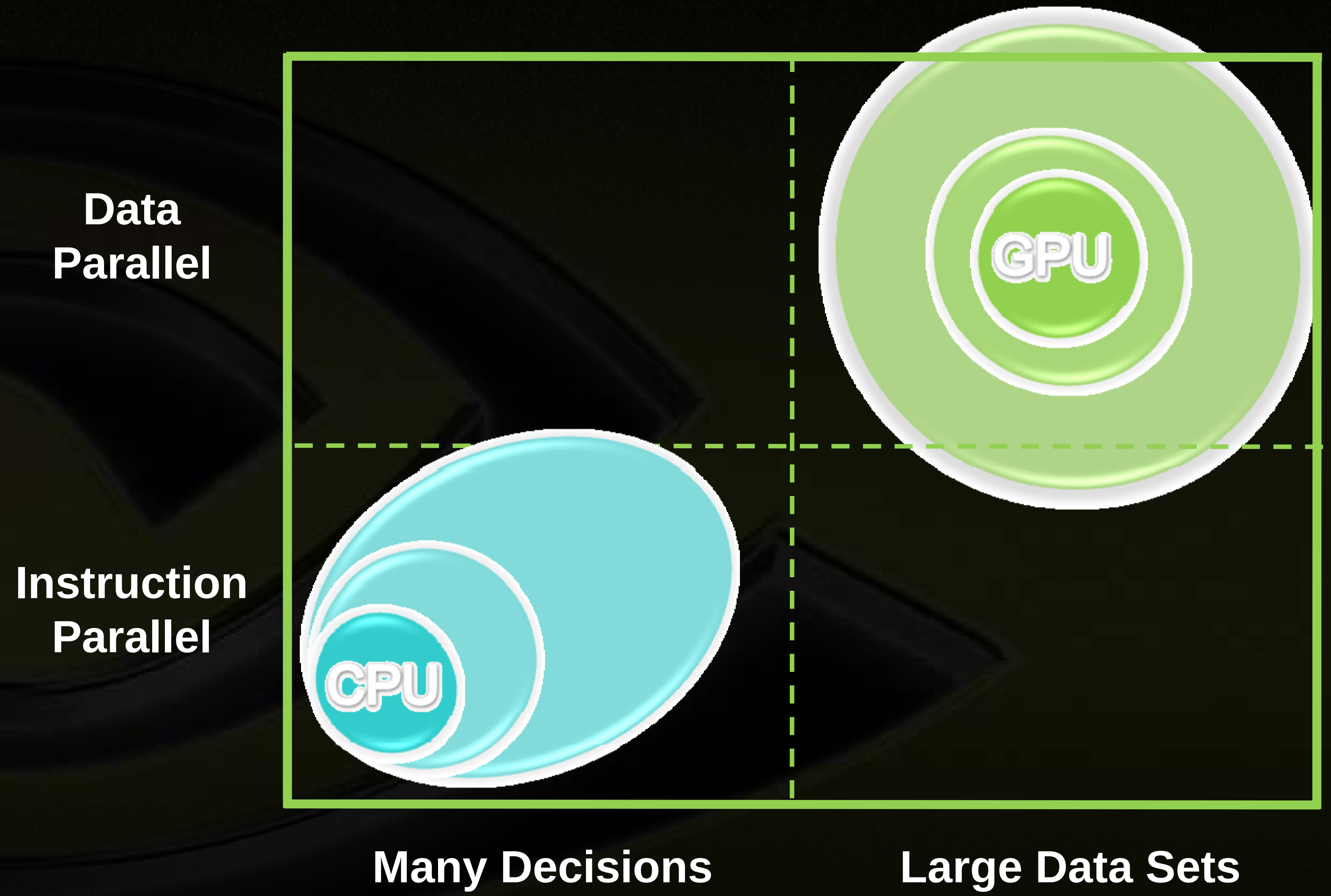
Introducing the ‘Fermi’ Architecture

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++

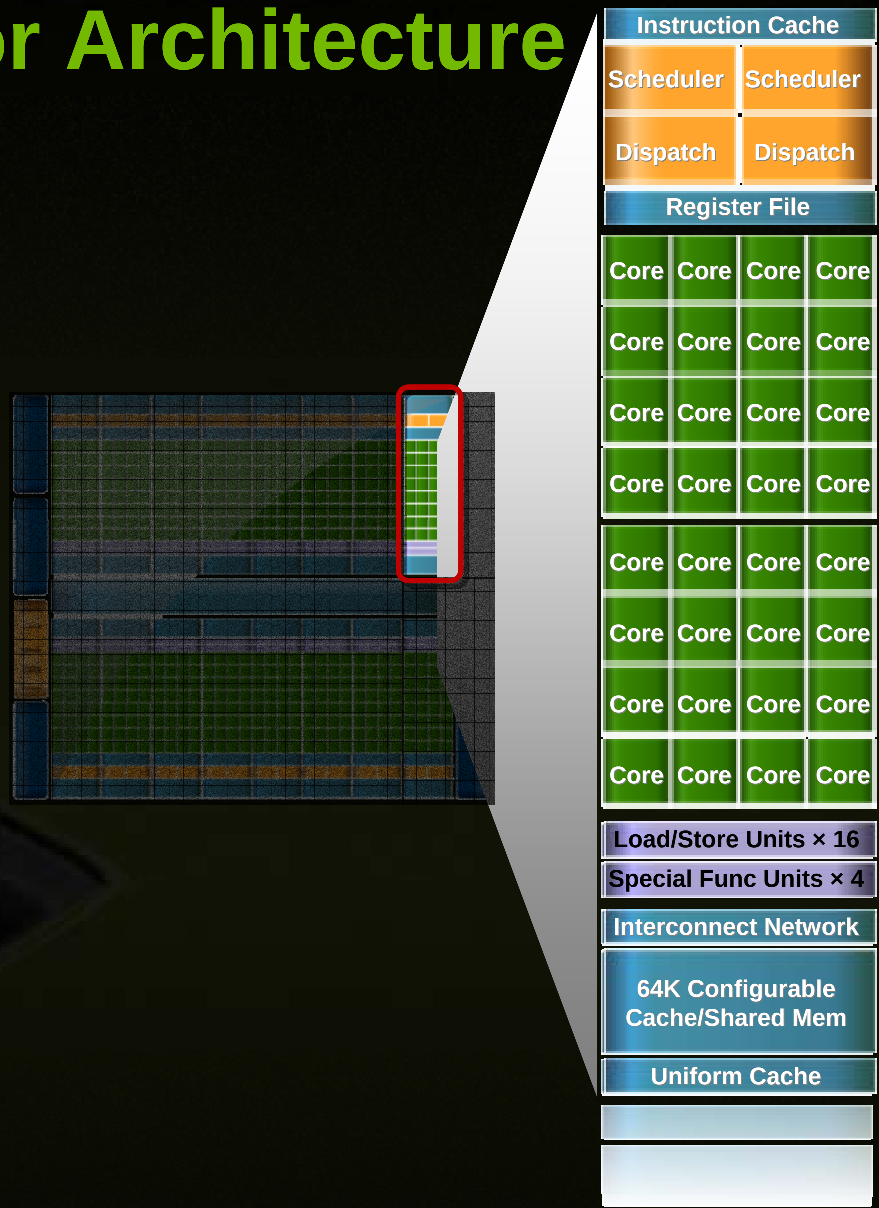
Design Goal of Fermi



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

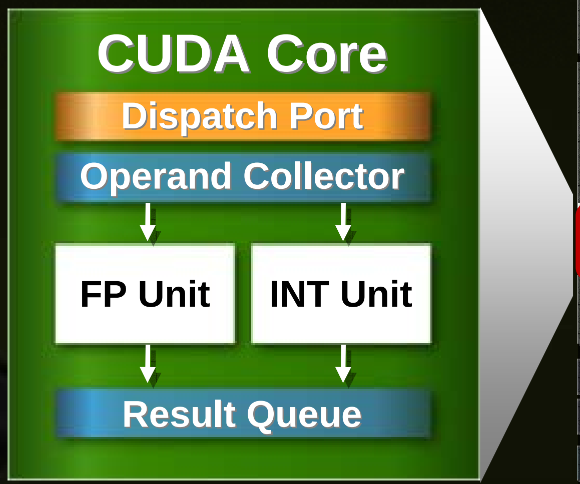
Streaming Multiprocessor Architecture

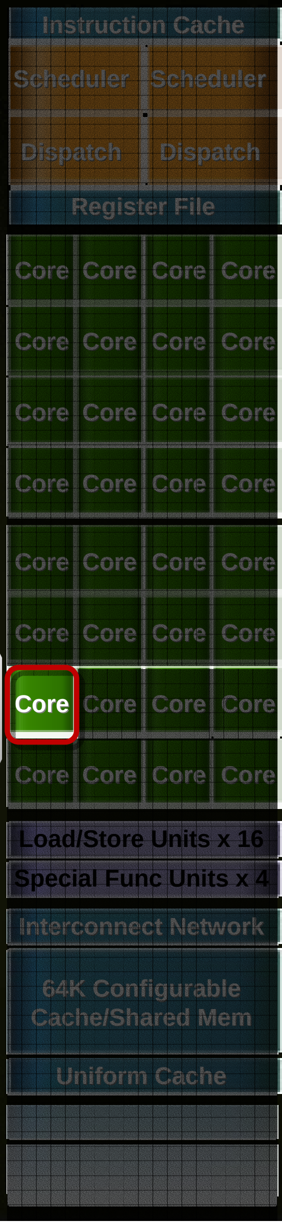
- 32 CUDA cores per SM (512 total)
- 2:1 ratio SP:DP floating-point performance
- Dual Thread Scheduler
- 64 KB of RAM for shared memory and L1 cache (configurable)



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

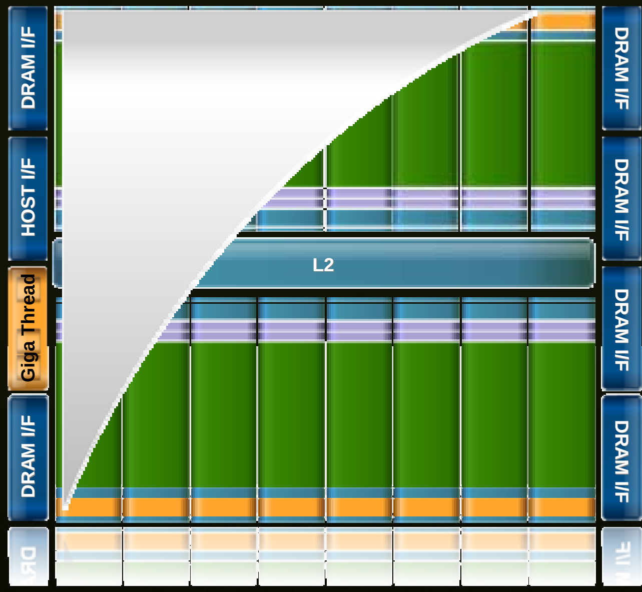




Cached Memory Hierarchy

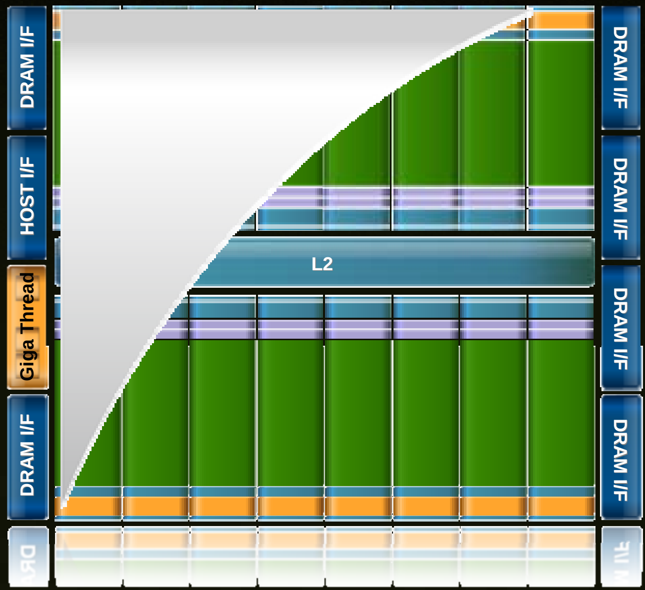
- First GPU architecture to support a true cache hierarchy in combination with on-chip shared memory
- L1 Cache per SM (32 cores)
  - Improves bandwidth and reduces latency
- Unified L2 Cache (768 KB)
  - Fast, coherent data sharing across all cores in the GPU

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

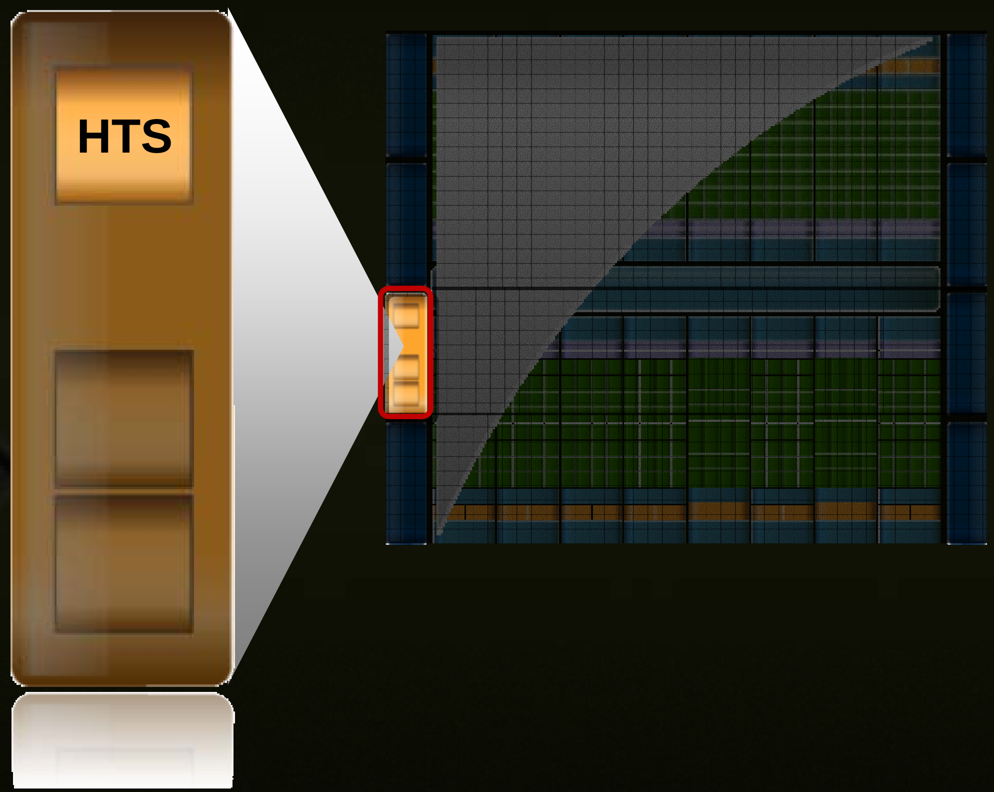


Error Correcting Code

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

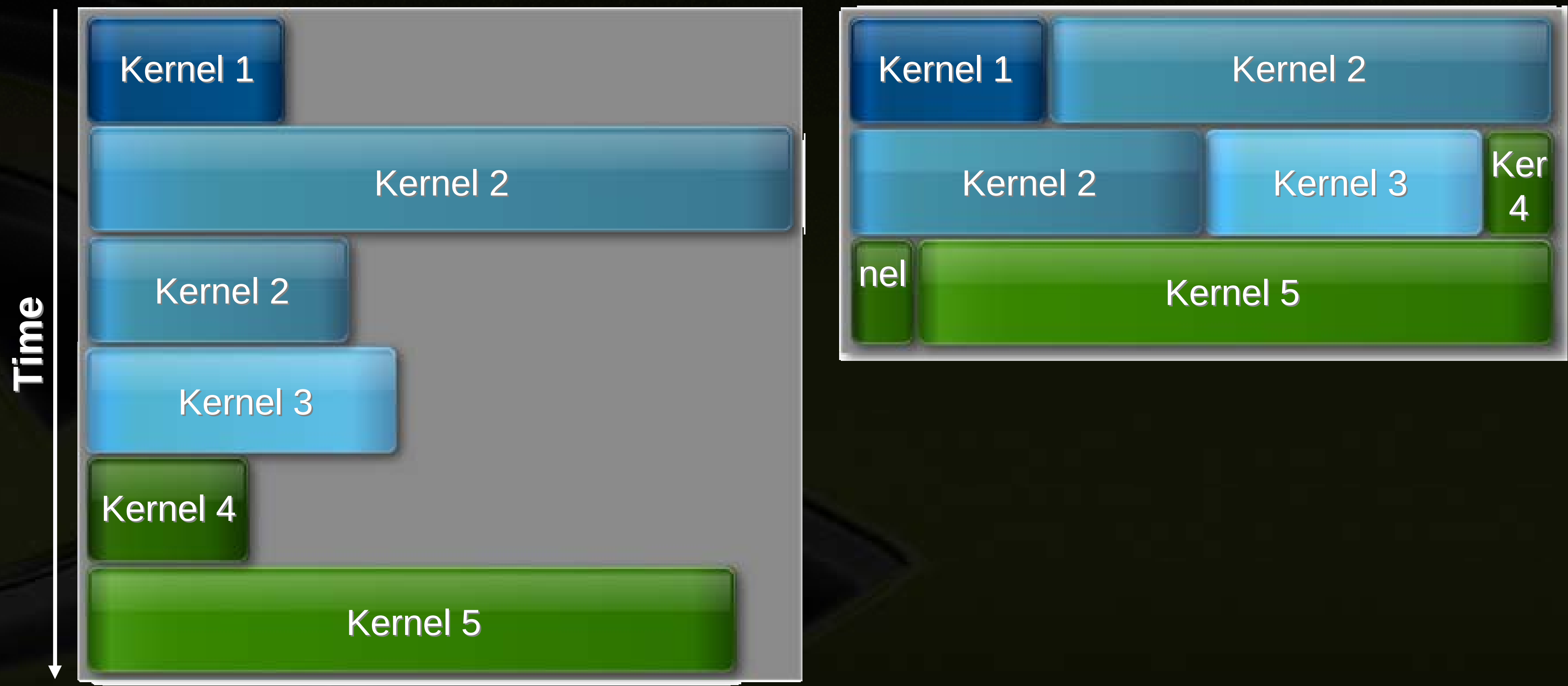
GigaThread™ Hardware Thread Scheduler

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



GigaThread Hardware Thread Scheduler

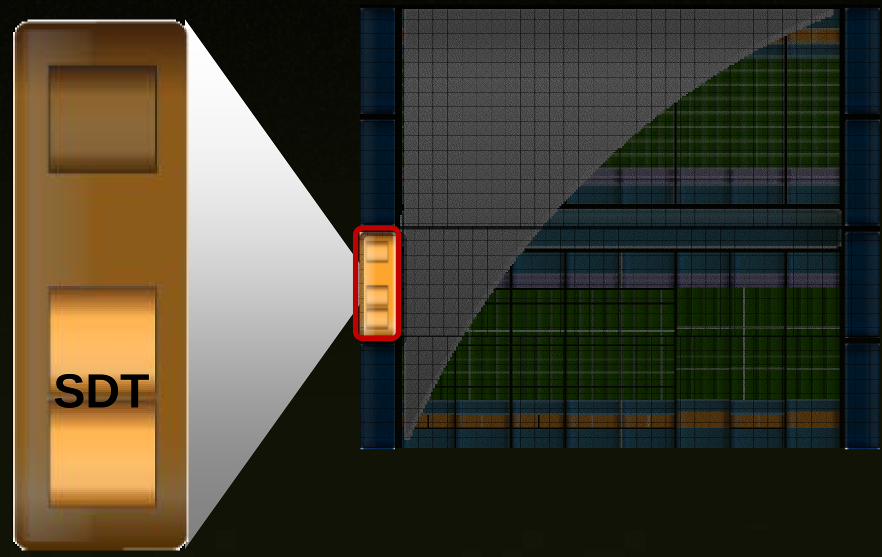
Concurrent Kernel Execution + Faster Context Switch



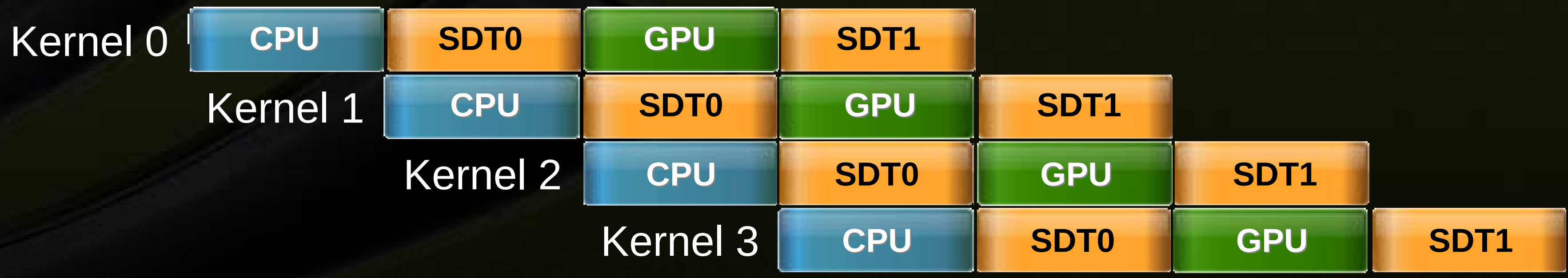
Serial Kernel Execution **Parallel Kernel Execution**

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:**



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)."

Tom Halfhill
Senior Editor
Microprocessor Report

"Fermi is the world's first complete GPU computing architecture."

Peter Glaskowsky
Technology Analyst
The Envisioneering Group

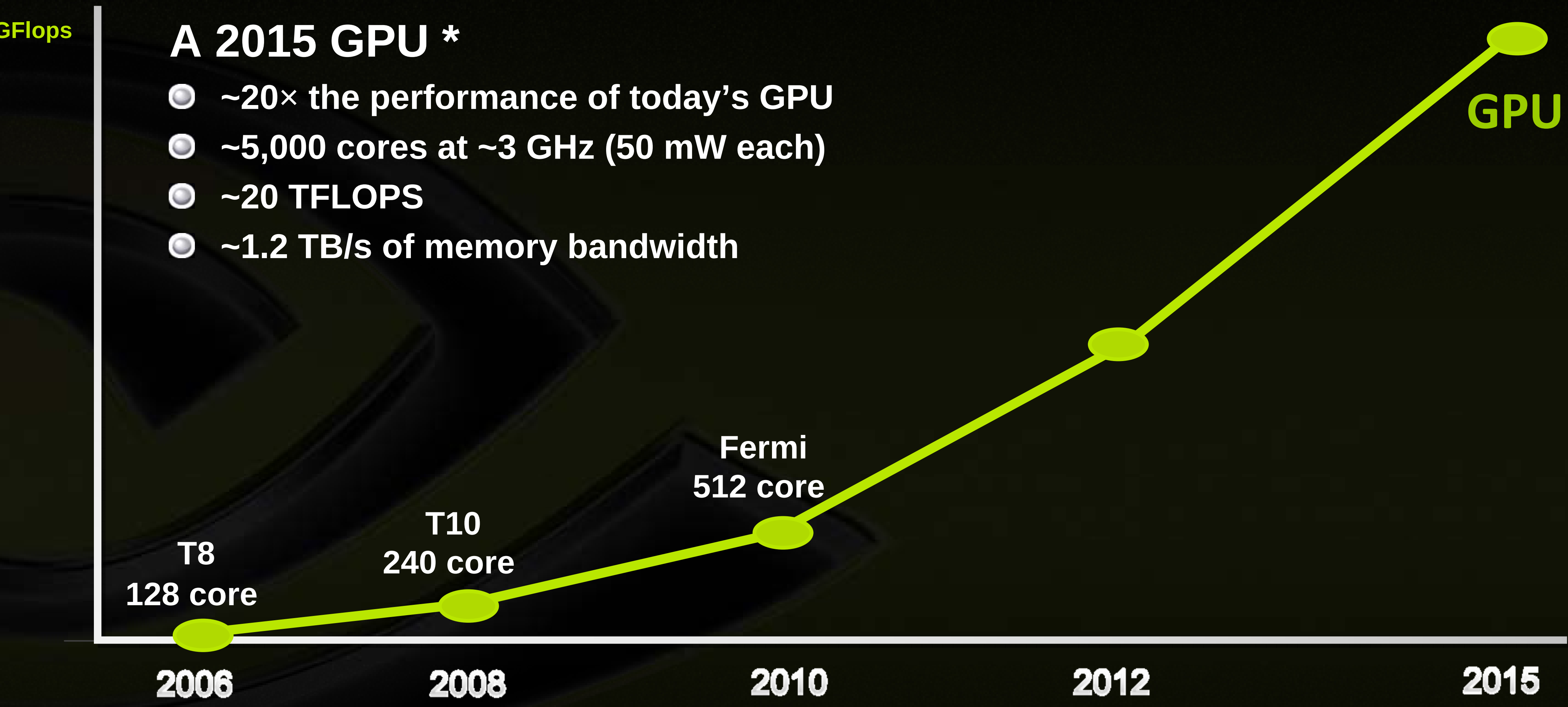
"The convergence of new, fast GPUs optimized for computation as well as 3-D graphics acceleration and industry-standard software development tools marks the real beginning of the GPU computing era. Gentlemen, start your GPU computing engines."

Nathan Brookwood
Principle Analyst & Founder
Insight 64

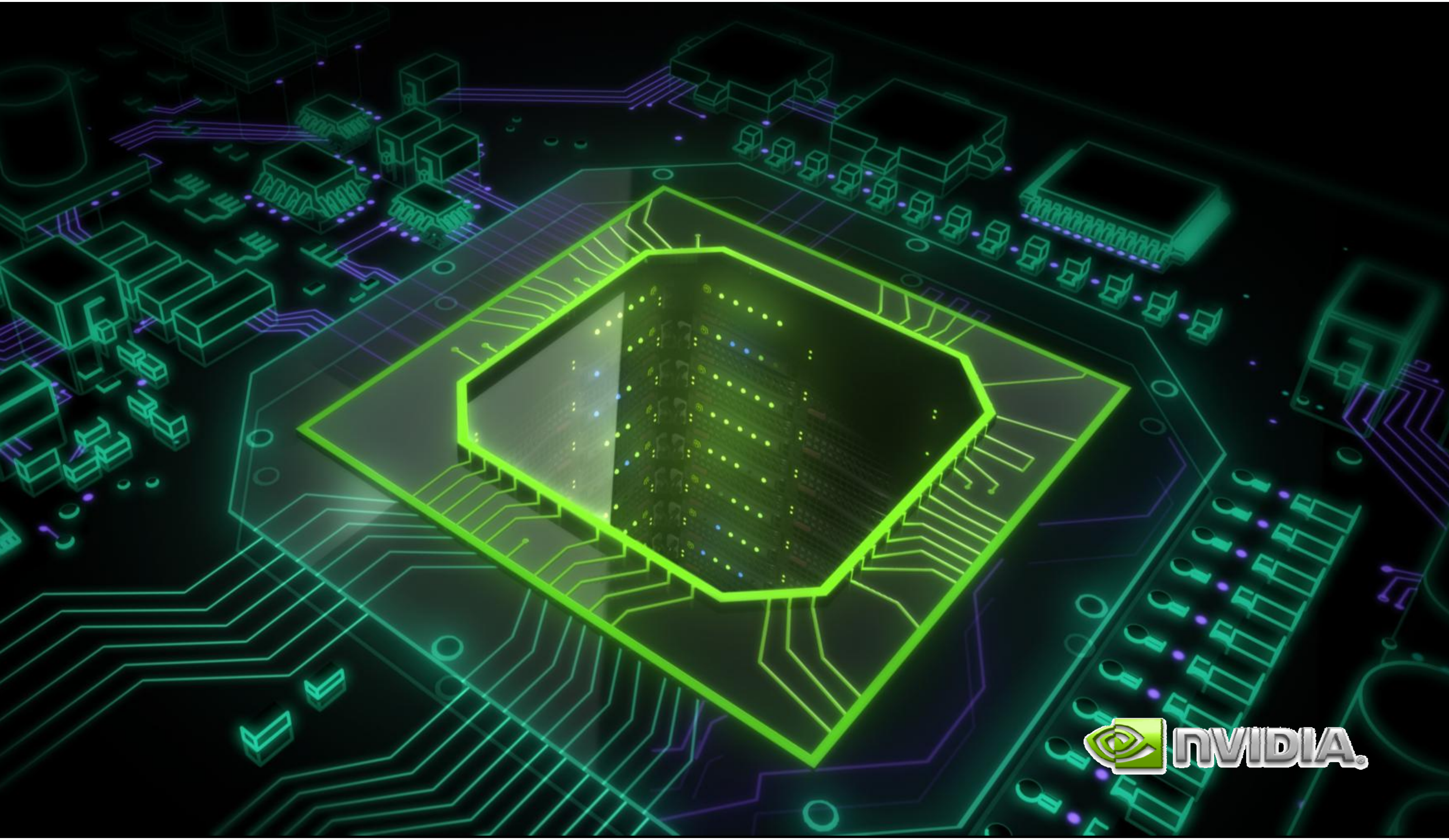
GPU Revolutionizing Computing

A 2015 GPU *

- ~20× 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 a what a GPU in 2015 might look like; it does not reflect any actual product plans


NVIDIA