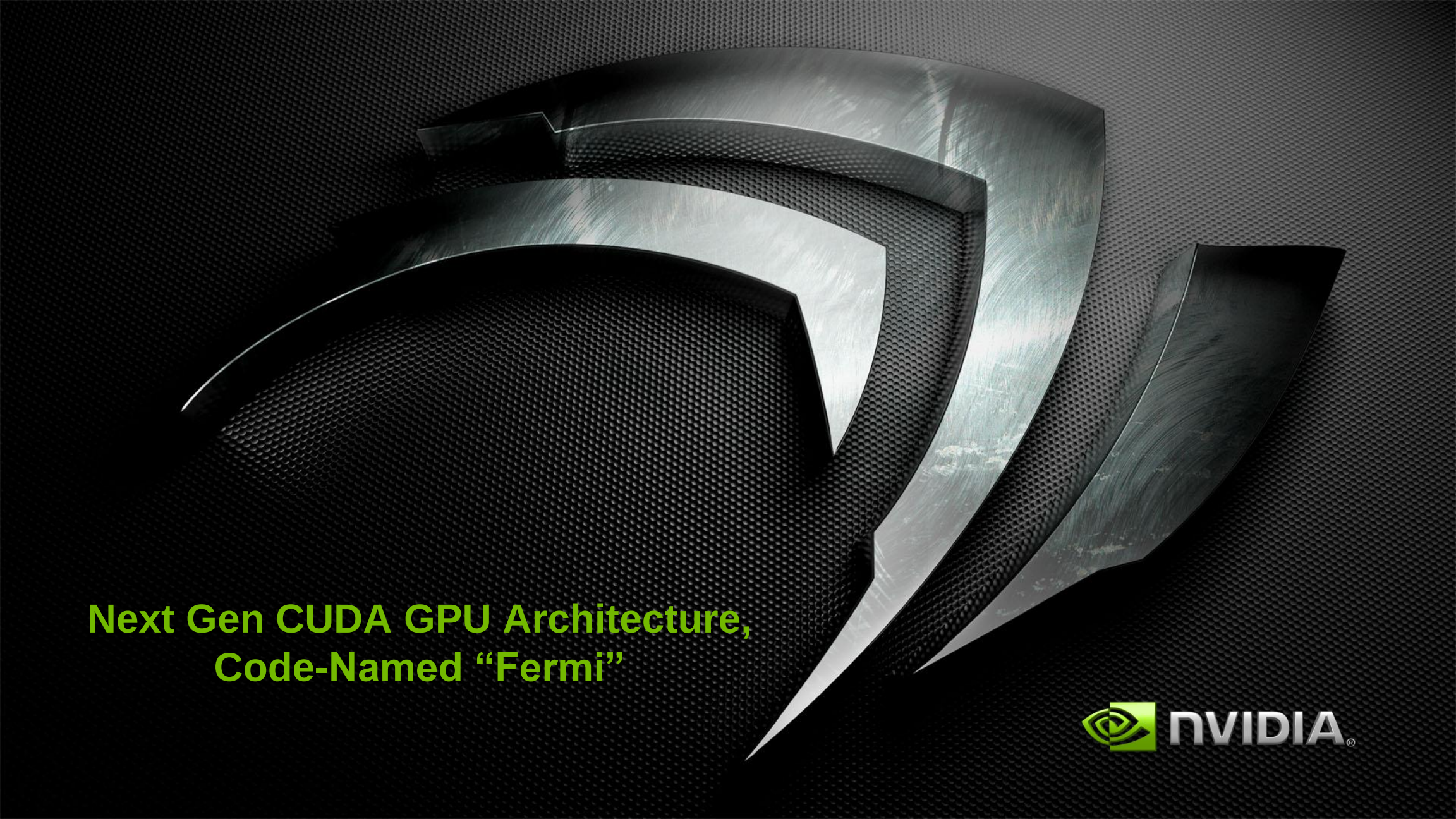
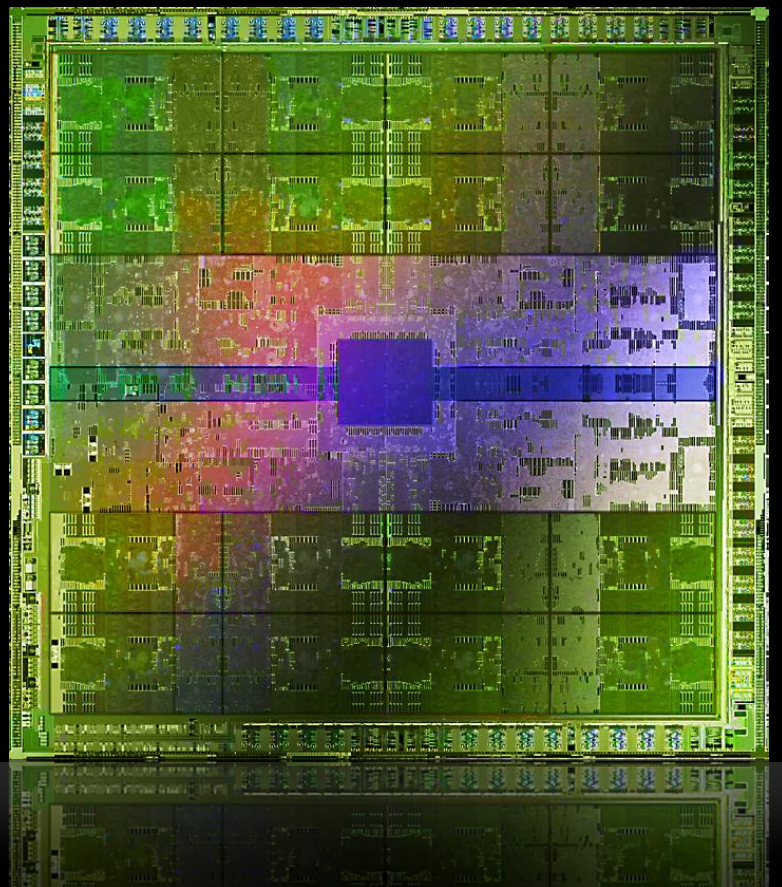

Next Gen CUDA GPU Architecture,
Code-Named “Fermi”
NVIDIA

Introducing the 'Fermi' Architecture

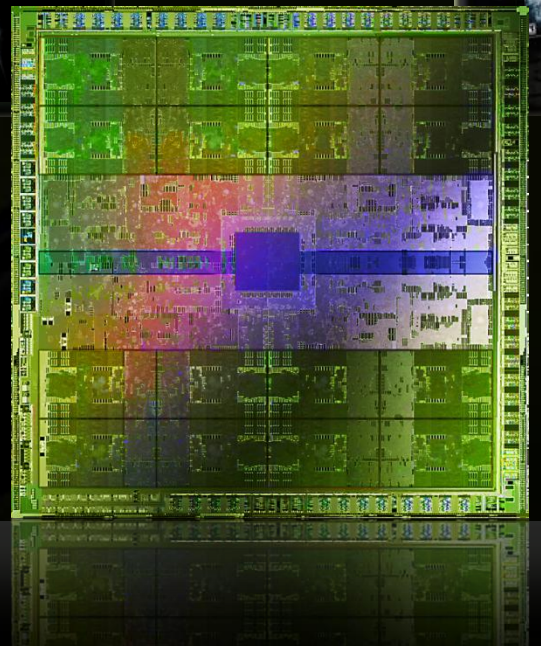
The Soul of a Supercomputer in the body of a GPU



- 3 billion transistors
- Over 2x the cores (512 total)
- 8x the peak DP performance
- ECC
- C++

HPC

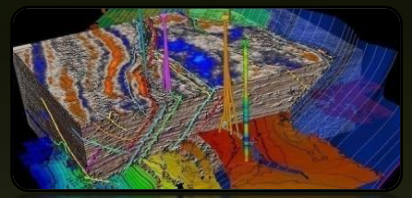
Consumer



NVIDIA Confidential

HPC Addressable TAM

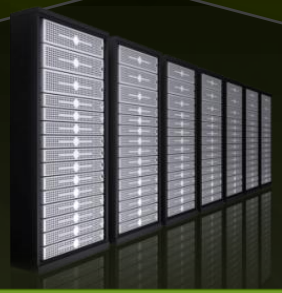
Seismic



- Energy Discovery
- Broad adoption

GPU TAM
$300M

Supercomputing



- World class science
- Top500

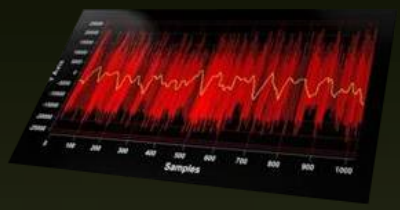
GPU TAM
$200M

Universities

- Desk supercomputing
- 1000s of customers

GPU TAM
$150M

Defense



- Signal analysis
- Very high need for compute resource

GPU TAM
$250M

Finance

- Pricing and risk
- Higher accuracy, faster

GPU TAM
$230M

NVIDIA Confidential

Source: NVIDIA, IDC

CUDA Co-Processing Ecosystem

Over 200 Universities Teaching CUDA

UIUC
MIT
Harvard
Berkeley
Cambridge
Oxford
…

IIT Delhi
Tsinghua
Dortmund
ETH Zurich
Moscow
NTU
…

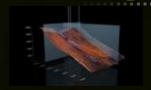
Languages

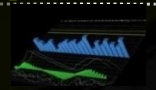
C, C++
DirectX
Fortran
Java
OpenCL
Python

Compilers

PGI Fortran
CAPs HMPP
MCUDA
MPI
NOAA Fortran2C
OpenMP

Applications


Oil & Gas

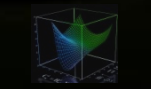

Finance

CFD

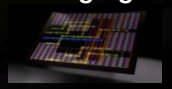
Medical

Biophysics

Imaging


Numerics

DSP


EDA

Libraries

FFT
BLAS
LAPACK
Image processing
Video processing
Signal processing
Vision

Consultants

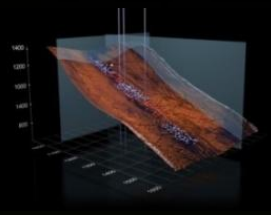
acceleware
HPC Project
ANEO
Scalable Graphics
CAPS
Stone Ridge Technology
EM Photonics
TECH
GASS
GPU Tech
WIPRO

OEMs

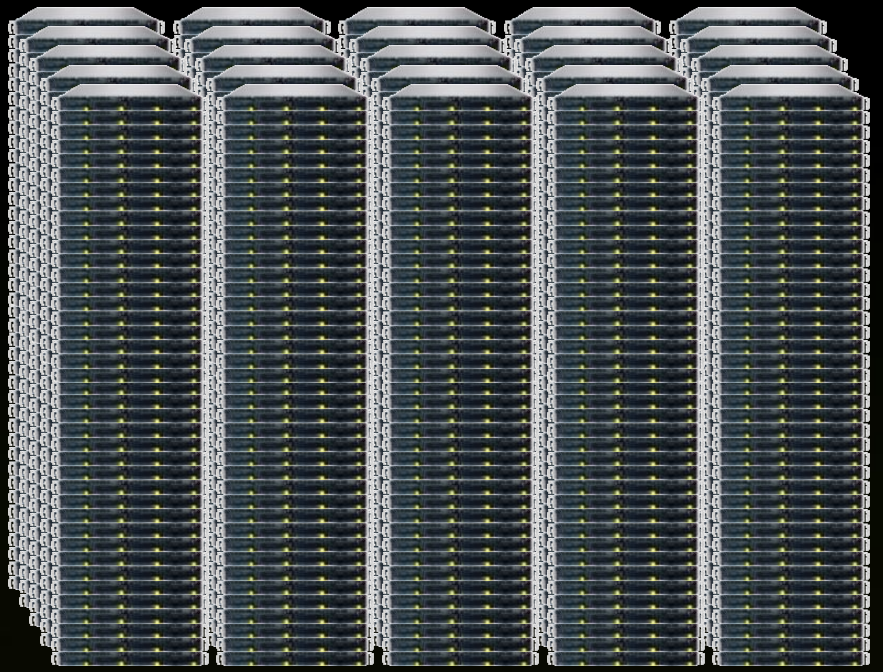
DELL
hp
Penguin Computing
CRAY
Sun
Supermicro
sgi
FUJITSU
BULL
APPRO
lenovo 联想
NEC

NVIDIA Confidential

Oil & Gas: Seismic Processing



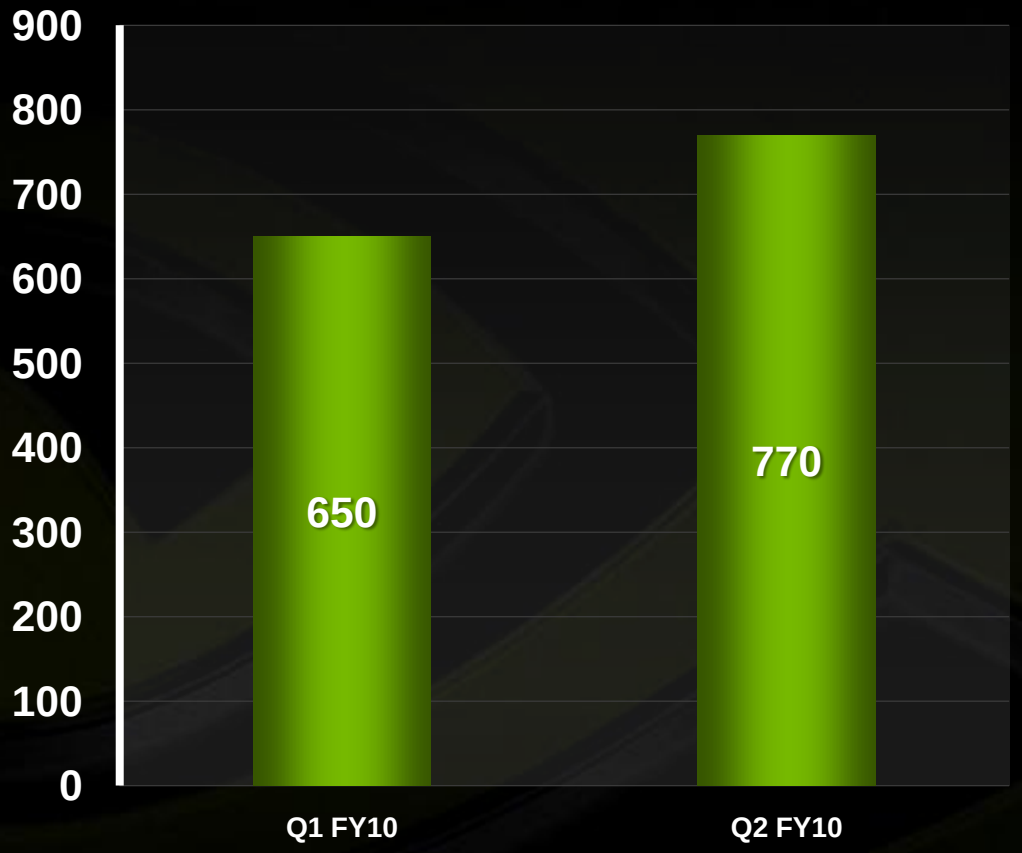
HESS



1	**Equal Performance**	1
32 Tesla S1070s	**31x Less Space**	2000 CPU Servers
~$400 K	**20x Lower Cost**	~$8 M
45 kWatts	**27x Lower Power**	1200 kWatts

NVIDIA Confidential

Tesla Server Installations



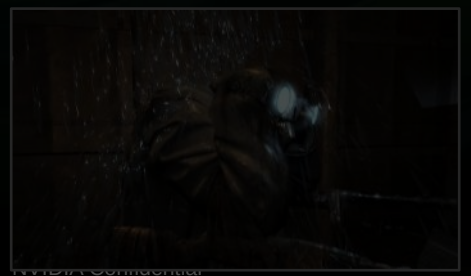
	# of GPUs
Commercial cluster	>3000
Research cluster	>2000
Chinese Academy of Sciences - Industrial Process Institute	828
Tokyo Institute of Technology Supercomputing Center	680
NCSA – National Center for Supercomputing Applications	384
Seismic processing	256
Pacific Northwest National Labs – Biomedical research	256
CSIRO – Australian National Supercomputing Center	252
Riken – Japanese Astrophysical research	220
Bloomberg – Bond pricing	200
Seismic processing	200
Chinese Academy of Sciences – Institute of Modern Physics	200

NVIDIA Confidential

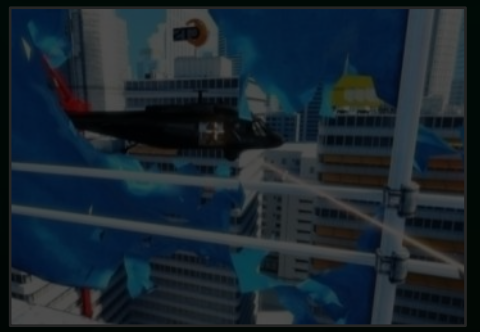
NVIDIA PhysX Creates Truly Dynamic Worlds Where Games Come Alive!

NVIDIA

Smoke and Water



Cloth and Clothing



Destructible Objects

Sparks and Dust

NVIDIA Confidential

DirectCompute Enables Amazing Visuals

- **Image Post-Processing**
  - Blurs
  - Depth-of-Field
  - Soft Shadows
- **Artificial Intelligence (AI)**
  - More game computation can move to the GPU
- **Faster Deferred Shading and Lighting**

GPU Accelerated Consumer Apps

DESIGNED FOR NVIDIA CUDA

MotionDSP vReveal

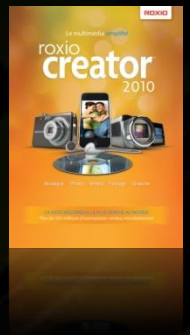
badaboom MEDIA CONVERTER

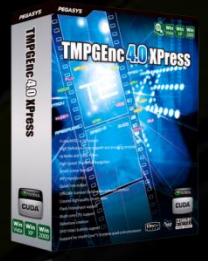
GPUGRID.net

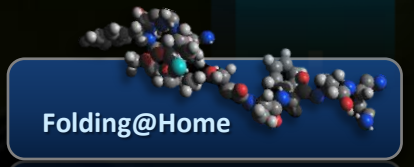
cooliris

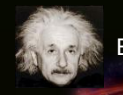
Adobe

CORECODEC COREAVC









Einstein@Home

3D COAT

SETI HOME

distributed.net

NVIDIA Confidential

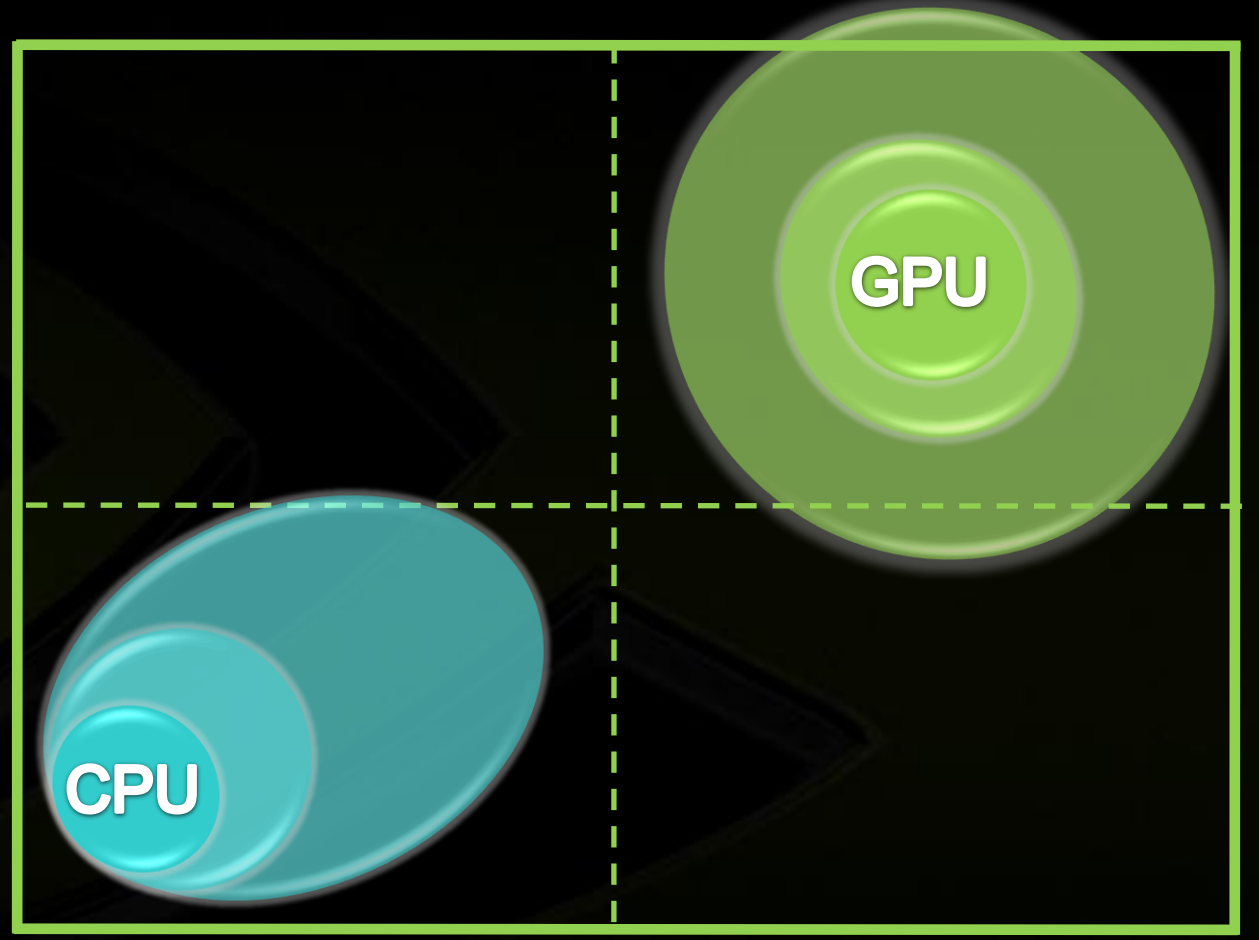
“For 50 years we’ve done things one way, and now we’re changing to a different model.

Parallelism is the only way to get there.”

Craig Mundie
Chief Research & Strategy Officer, Microsoft
As quoted in “Moore’s Law Doesn’t Matter” Article
Newsweek, August 24, 2009

NVIDIA Confidential

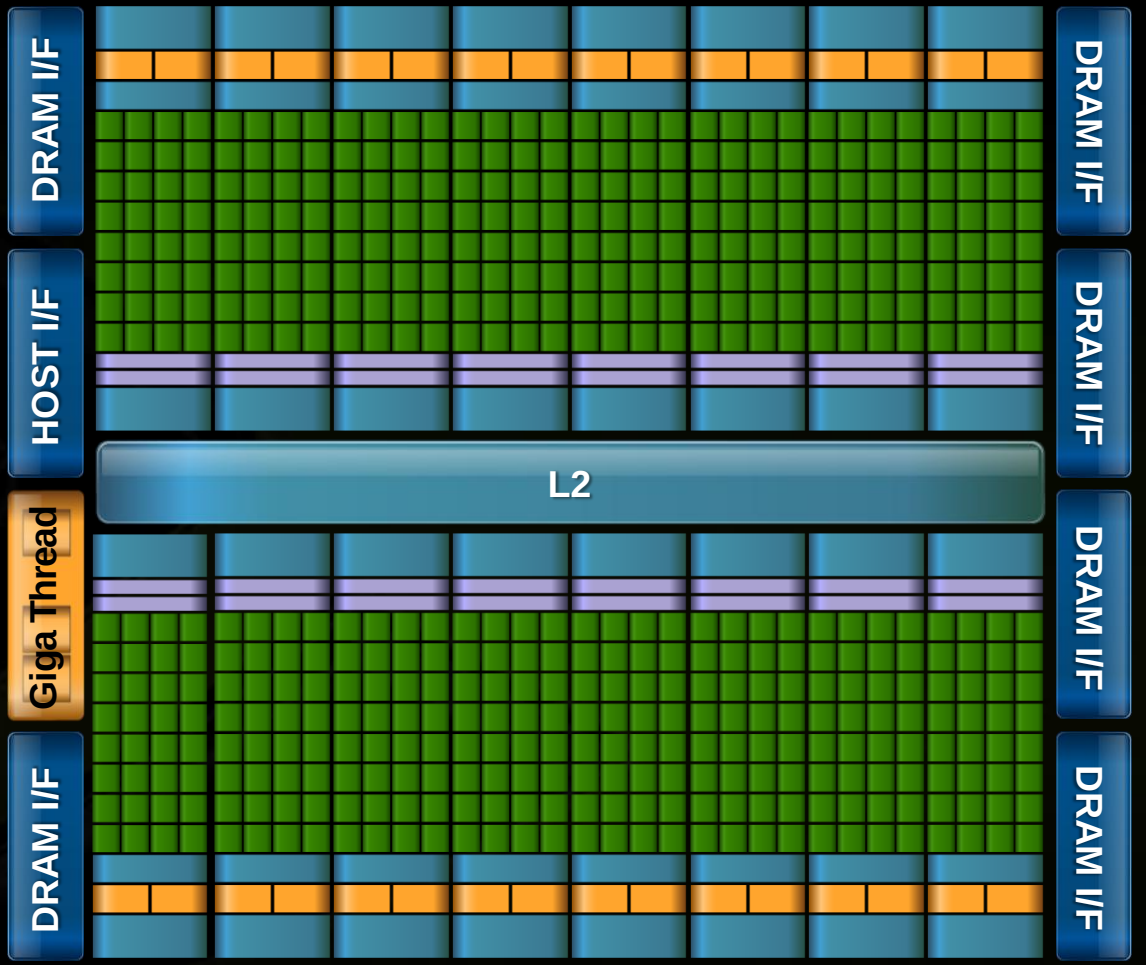
Design Goal of Fermi



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

NVIDIA Confidential

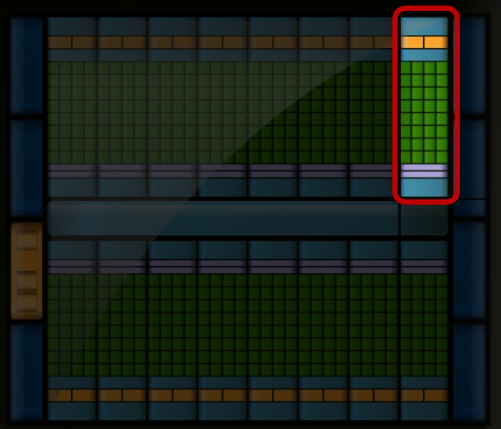
Fermi Block Diagram

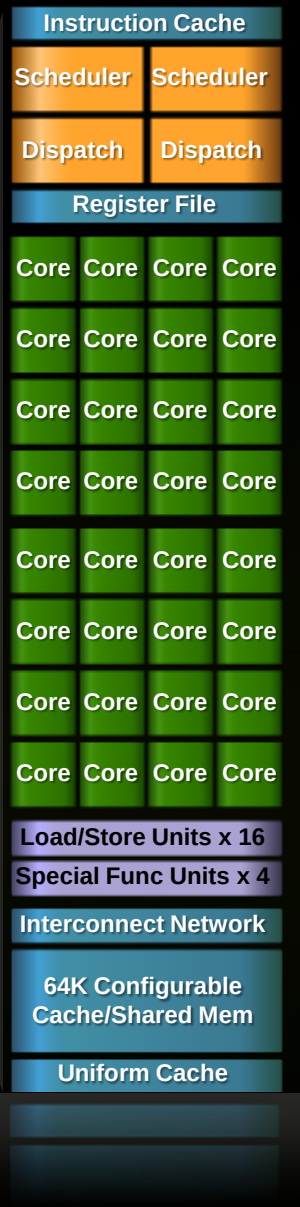


NVIDIA Confidential

SM Architecture

- 32 CUDA cores per SM (512 total)
- 8x peak double precision floating point performance
- Dual Warp Scheduler
- 64 KB of RAM

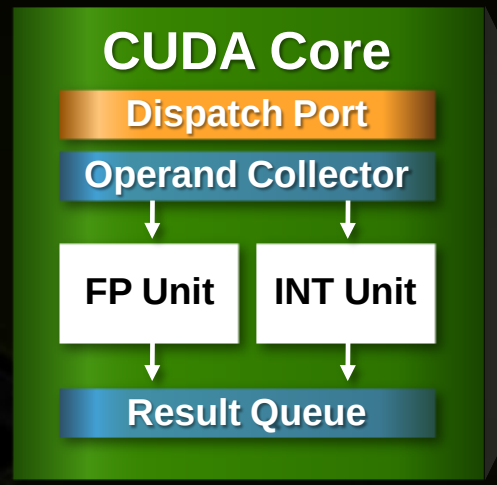




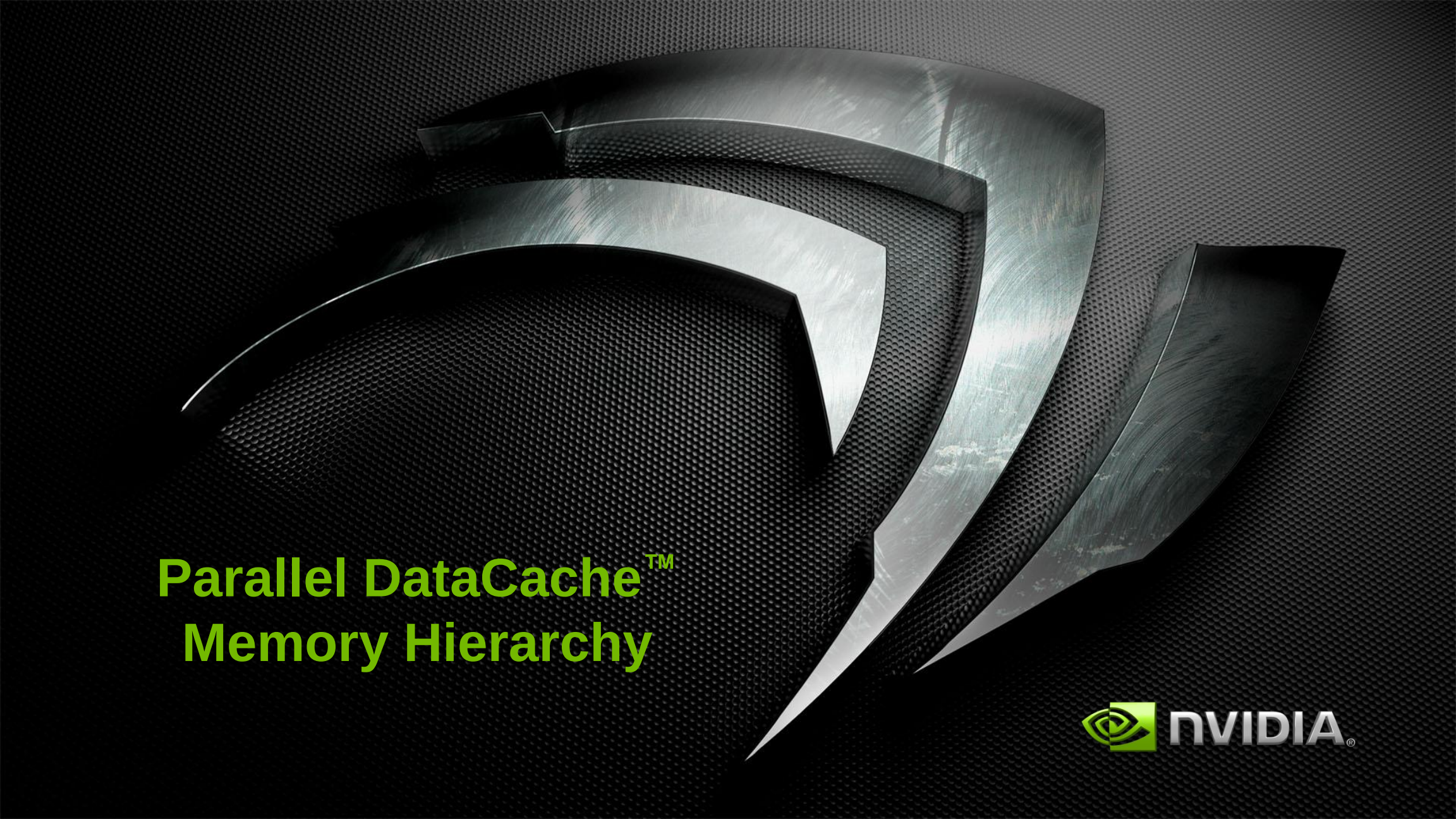
NVIDIA Confidential

CUDA Cores

- 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

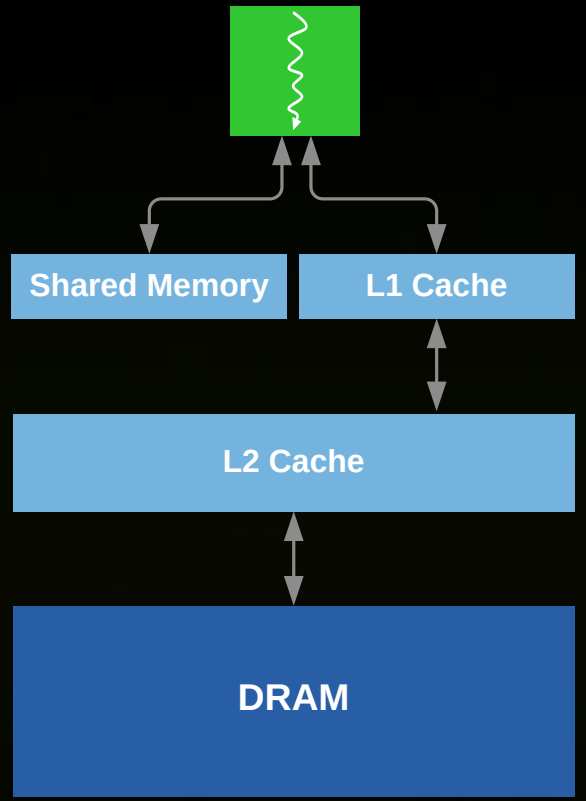


NVIDIA Confidential


Parallel DataCache™
Memory Hierarchy
NVIDIA

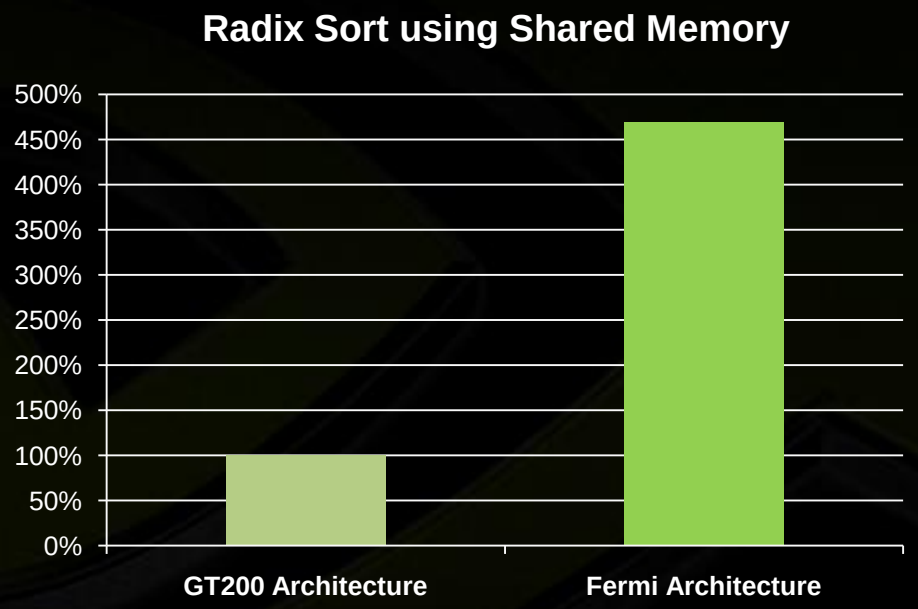
Cached Memory Hierarchy

- First GPU architecture to support a true cache hierarchy in combination with on-chip shared memory
- Improves bandwidth and reduces latency through L1 cache's configurable shared memory
- Fast, coherent data sharing across the GPU through unified L2 cache

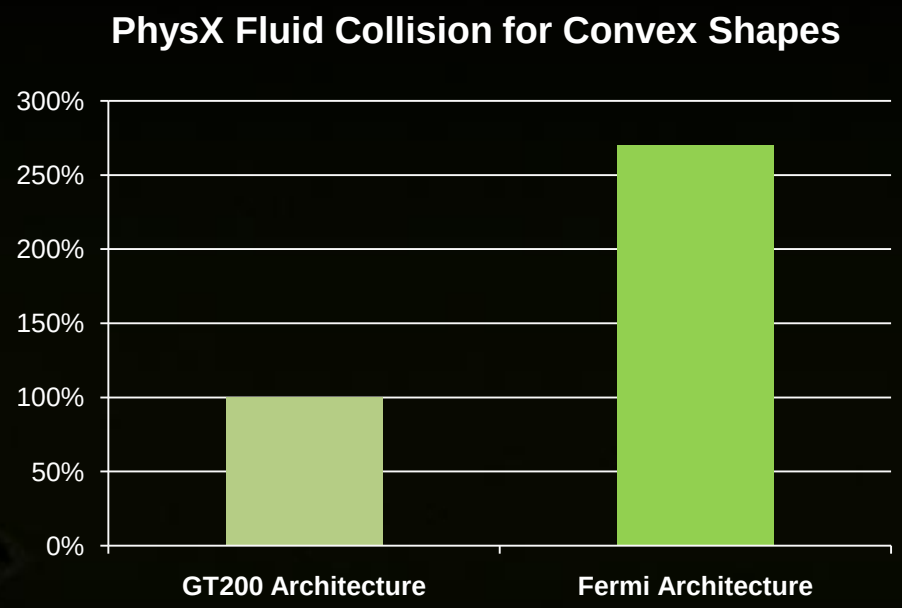


NVIDIA Confidential

Performance Benefits of Memory Architecture



4.7x faster!



2.7x faster!

NVIDIA Confidential

World's First GPUs with ECC

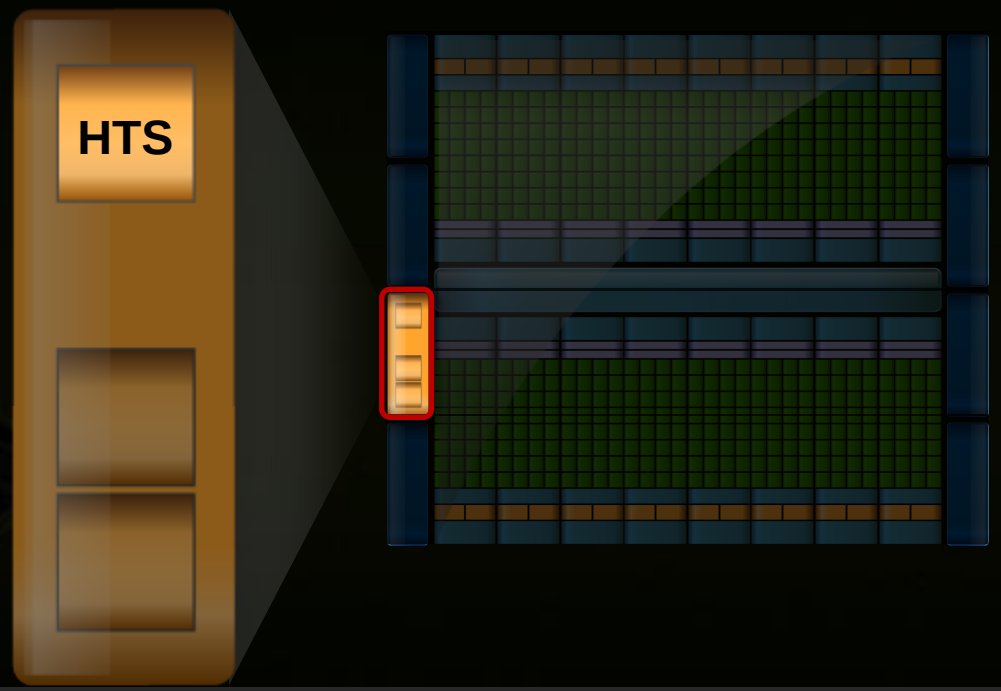
- **ECC is a must have for many computing applications**
- **So we implemented ECC**
- **DRAM protected by ECC**
  - ECC supported for GDDR5 as well as SDDR3 memory configurations
- **All major internal memories are ECC protected**
  - Register file, L1 cache, L2 cache

NVIDIA Confidential

HW Thread Management:
GigaThread™
NVIDIA

GigaThread Hardware Thread Scheduler

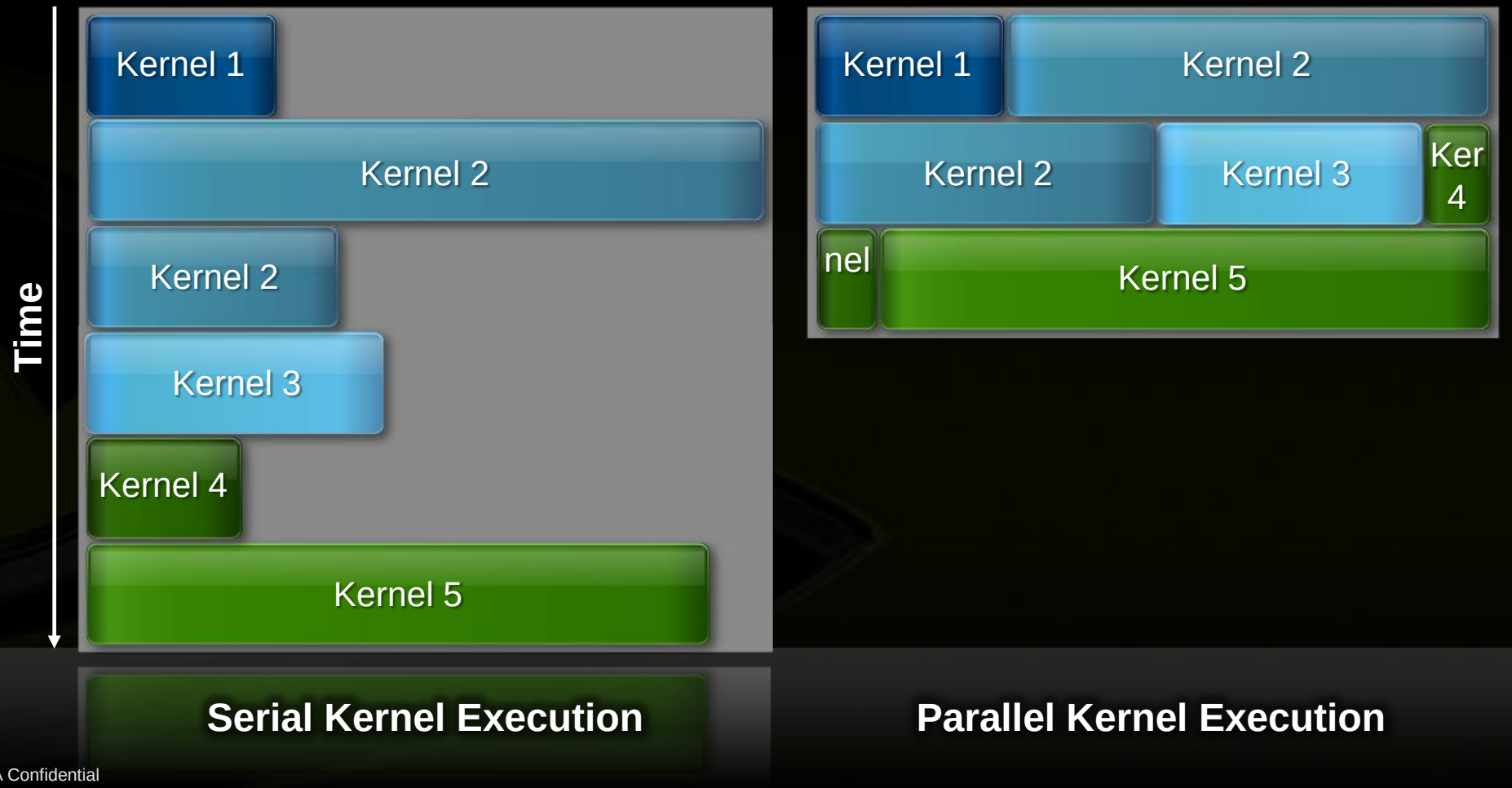
- Hierarchically manages thousands of simultaneously active threads
- 10x faster application context switching
- Concurrent kernel execution



NVIDIA Confidential

GigaThread Hardware Thread Scheduler

Concurrent Kernel Execution + Faster Context Switch



NVIDIA Confidential

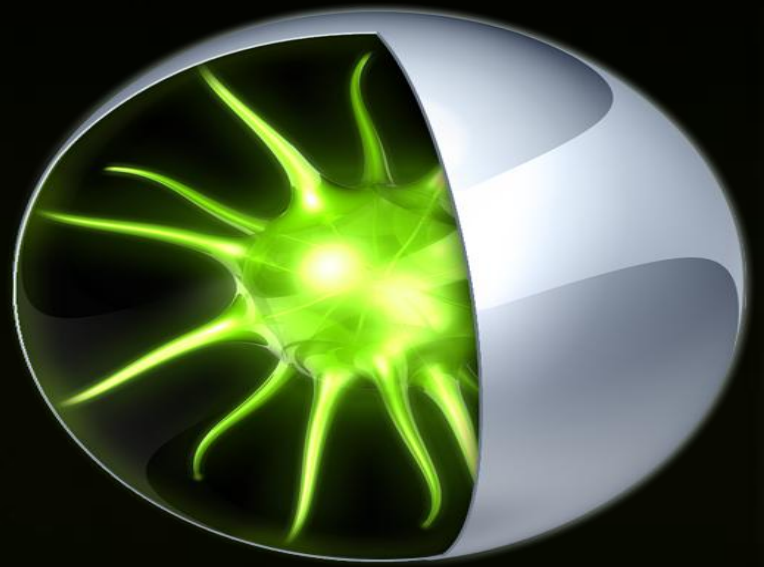
NVIDIA Nexus™
Development Environment
NVIDIA

NVIDIA Nexus IDE

The Industry's 1st IDE (Integrated Development Environment) for massively parallel applications

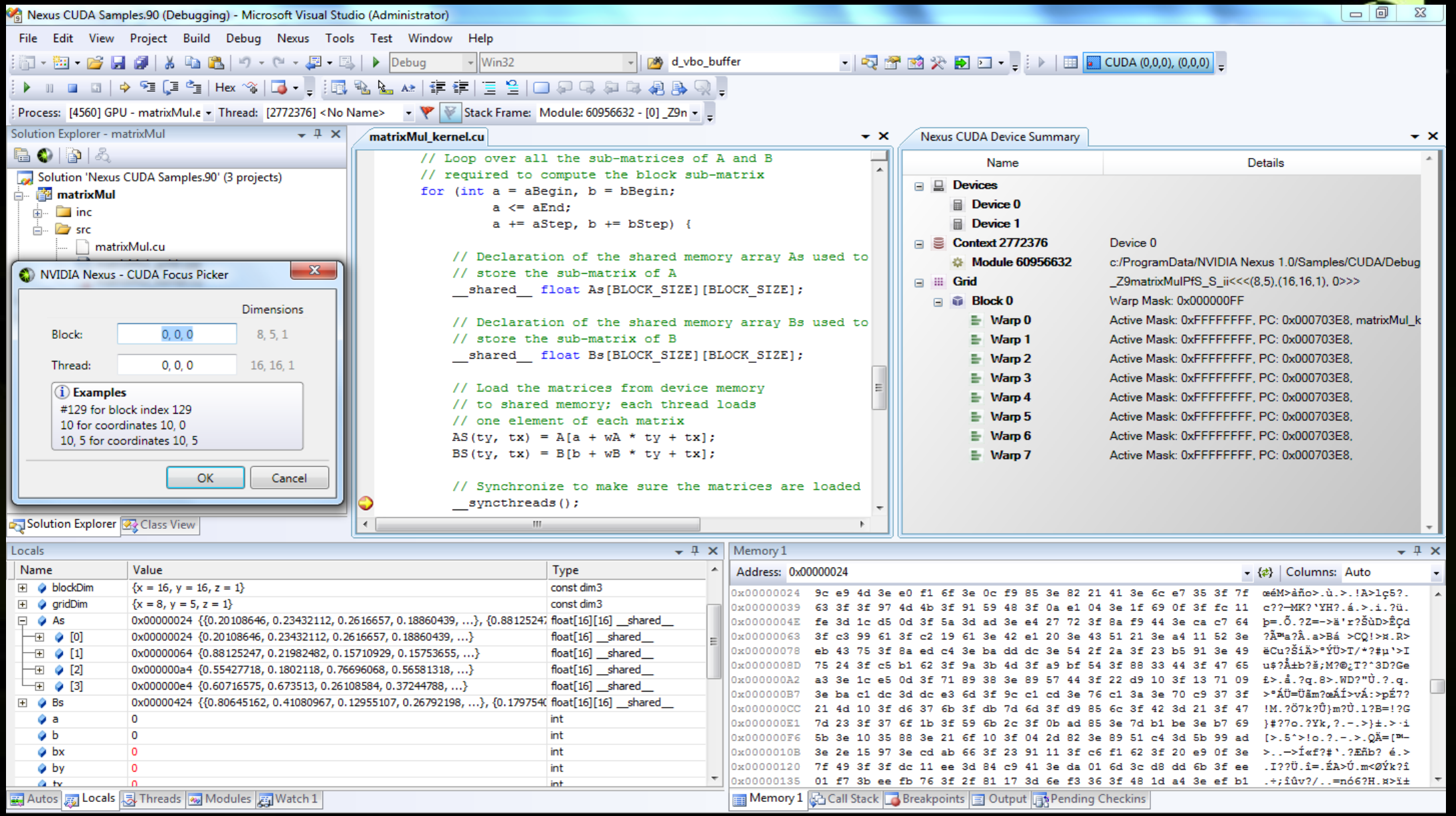
Accelerates co-processing (CPU + GPU) application development

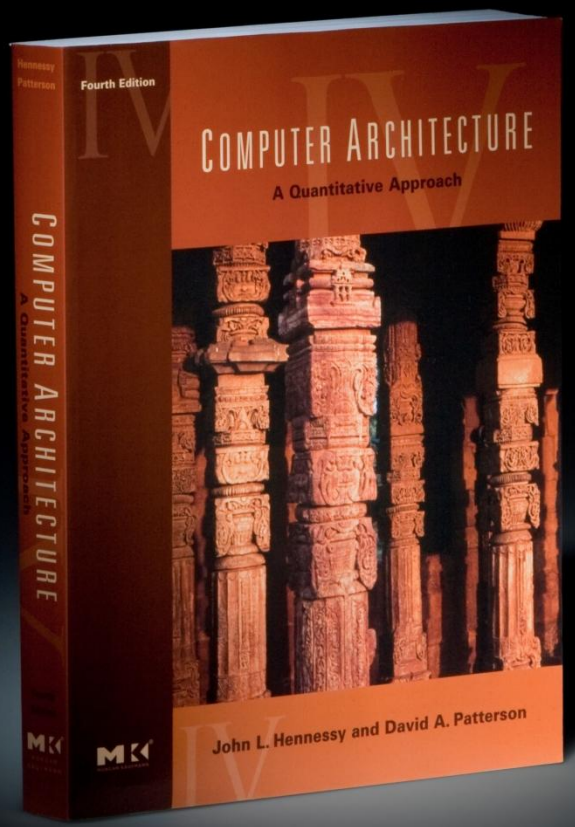
Complete Visual Studio-integrated development environment



Microsoft Visual Studio PARTNER

NVIDIA Confidential


Nexus CUDA Samples.90 (Debugging) - Microsoft Visual Studio (Administrator)
File Edit View Project Build Debug Nexus Tools Test Window Help
Debug
Win32
d_vbo_buffer
CUDA (0,0,0), (0,0,0)
Process: [4560] GPU - matrixMul.e
Thread: [2772376] <No Name>
Stack Frame: Module: 60956632 - [0] _Z9n
Solution Explorer - matrixMul
Solution 'Nexus CUDA Samples.90' (3 projects)
matrixMul
inc
src
matrixMul.cu
NVIDIA Nexus - CUDA Focus Picker
Dimensions
Block:
0, 0, 0
8, 5, 1
Thread:
0, 0, 0
16, 16, 1
Examples
#129 for block index 129
10 for coordinates 10, 0
10, 5 for coordinates 10, 5
OK
Cancel
Solution Explorer
Class View
matrixMul_kernel.cu
// Loop over all the sub-matrices of A and B
// required to compute the block sub-matrix
for (int a = aBegin, b = bBegin;
a <= aEnd;
a += aStep, b += bStep) {
// Declaration of the shared memory array As used to
// store the sub-matrix of A
__shared__ float As[BLOCK_SIZE][BLOCK_SIZE];
// Declaration of the shared memory array Bs used to
// store the sub-matrix of B
__shared__ float Bs[BLOCK_SIZE][BLOCK_SIZE];
// Load the matrices from device memory
// to shared memory; each thread loads
// one element of each matrix
AS(ty, tx) = A[a + wA * ty + tx];
BS(ty, tx) = B[b + wB * ty + tx];
// Synchronize to make sure the matrices are loaded
__syncthreads();
Nexus CUDA Device Summary
Name
Details
Devices
Device 0
Device 1
Context 2772376
Device 0
Module 60956632
c:/ProgramData/NVIDIA Nexus 1.0/Samples/CUDA/Debug
Grid
_Z9matrixMulPfS_S_ii<<<(8,5),(16,16,1), 0>>>
Block 0
Warp Mask: 0x000000FF
Warp 0
Active Mask: 0xFFFFFFFF, PC: 0x000703E8, matrixMul_k
Warp 1
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Warp 2
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Warp 3
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Warp 4
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Warp 5
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Warp 6
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Warp 7
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Locals
Name
Value
Type
blockDim
{x = 16, y = 16, z = 1}
const dim3
gridDim
{x = 8, y = 5, z = 1}
const dim3
As
0x00000024 {{0.20108646, 0.23432112, 0.2616657, 0.18860439, ...}, {0.8812524
float[16][16] __shared__
[0]
0x00000024 {0.20108646, 0.23432112, 0.2616657, 0.18860439, ...}
float[16] __shared__
[1]
0x00000064 {0.88125247, 0.21982482, 0.15710929, 0.15753655, ...}
float[16] __shared__
[2]
0x000000a4 {0.55427718, 0.1802118, 0.76696068, 0.56581318, ...}
float[16] __shared__
[3]
0x000000e4 {0.60716575, 0.673513, 0.26108584, 0.37244788, ...}
float[16] __shared__
Bs
0x00000424 {{0.80645162, 0.41080967, 0.12955107, 0.26792198, ...}, {0.1797540
float[16][16] __shared__
a
0
int
b
0
int
bx
0
int
by
0
int
Memory 1
Address: 0x00000024
Columns: Auto
0x00000024 9c e9 4d 3e e0 f1 6f 3e 0c f9 85 3e 82 21 41 3e 6c e7 35 3f 7f
0x00000039 63 3f 3f 97 4d 4b 3f 91 59 48 3f 0a e1 04 3e 1f 69 0f 3f fc 11
0x0000004E fe 3d 1c d5 0d 3f 5a 3d ad 3e e4 27 72 3f 8a f9 44 3e ca c7 64
0x00000063 3f c3 99 61 3f c2 19 61 3e 42 e1 20 3e 43 51 21 3e a4 11 52 3e
0x00000078 eb 43 75 3f 8a ed c4 3e ba dd dc 3e 54 2f 2a 3f 23 b5 91 3e 49
0x0000008D 75 24 3f c5 b1 62 3f 9a 3b 4d 3f a9 bf 54 3f 88 33 44 3f 47 65
0x000000A2 a3 3e 1c e5 0d 3f 71 89 38 3e 89 57 44 3f 22 d9 10 3f 13 71 09
0x000000B7 3e ba c1 dc 3d dc e3 6d 3f 9c c1 cd 3e 76 c1 3a 3e 70 c9 37 3f
0x000000CC 21 4d 10 3f d6 37 6b 3f db 7d 6d 3f d9 85 6c 3f 42 3d 21 3f 47
0x000000E1 7d 23 3f 37 6f 1b 3f 59 6b 2c 3f 0b ad 85 3e 7d b1 be 3e b7 69
0x000000F6 5b 3e 10 35 88 3e 21 6f 10 3f 04 2d 82 3e 89 51 c4 3d 5b 99 ad
0x0000010B 3e 2e 15 97 3e cd ab 66 3f 23 91 11 3f c6 f1 62 3f 20 e9 0f 3e
0x00000120 7f 49 3f 3f dc 11 ee 3d 84 c9 41 3e da 01 6d 3c d8 dd 6b 3f ee
0x00000135 01 f7 3b ee fb 76 3f 2f 81 17 3d 6e f3 36 3f 48 1d a4 3e ef b1
Autos
Locals
Threads
Modules
Watch 1
Memory 1
Call Stack
Breakpoints
Output
Pending Checkins



> *"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 is the world’s first complete GPU computing architecture.”

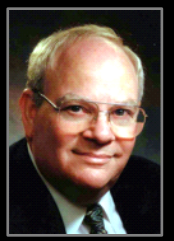
Peter Glaskowsky
Technology Analyst
The Envisioneering Group

“Fermi surpasses anything announced by NVIDIA's leading GPU competitor (AMD).”

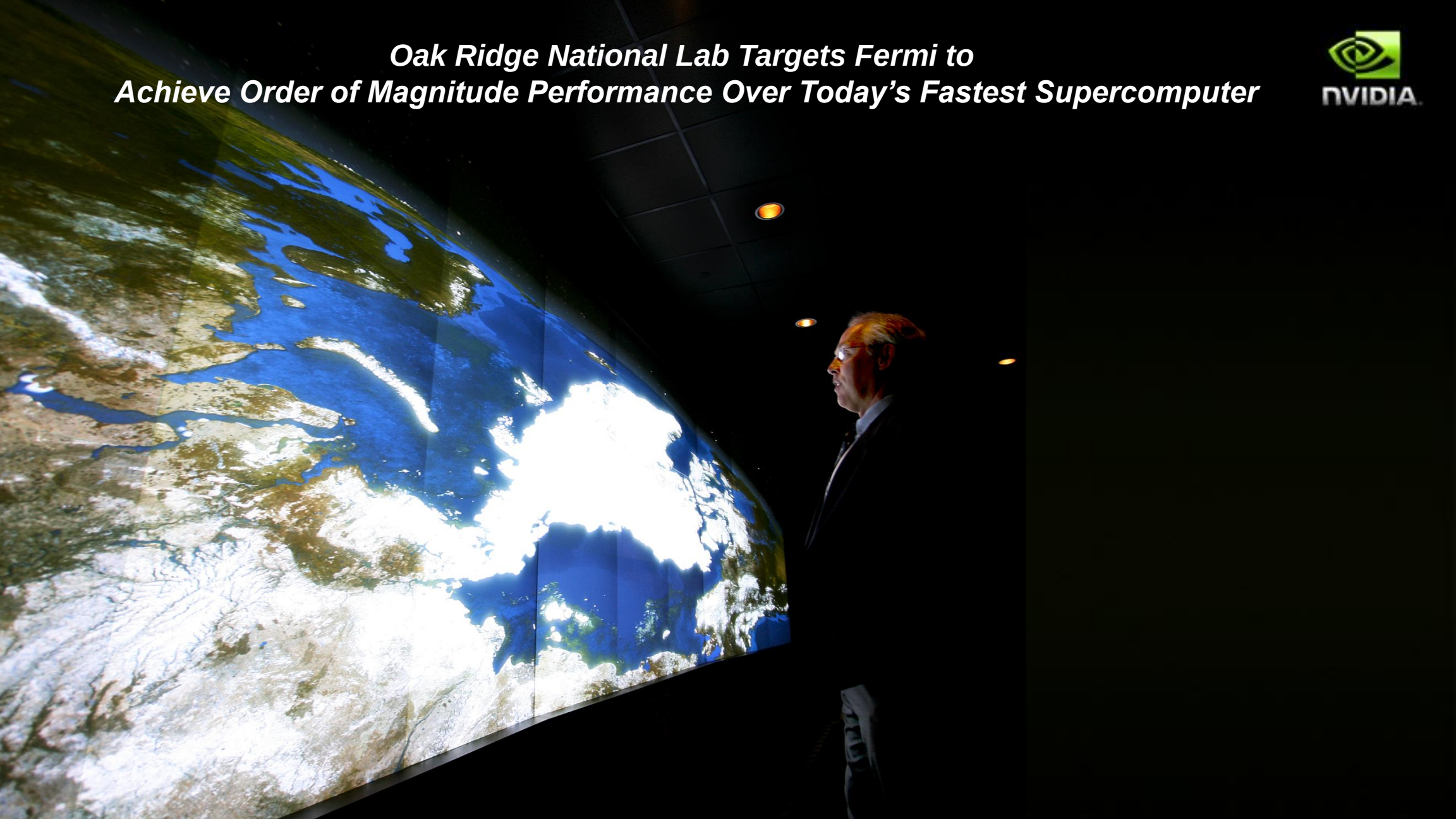
Tom Halfhill
Senior Editor
Instat’s Microprocessor Report

“ 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



Oak Ridge National Lab Targets Fermi to Achieve Order of Magnitude Performance Over Today's Fastest Supercomputer



GeForce – You CAN Have It All !

	NVIDIA Fermi Architecture
DirectX 11	✓
Advanced Physics Effects	PhysX by NVIDIA
3D Stereo Gaming	NVIDIA 3D VISION
Over 50 Consumer Compute Apps*	✓
Ray Tracing	

*Forecasted total by Q1 2010

NVIDIA Confidential

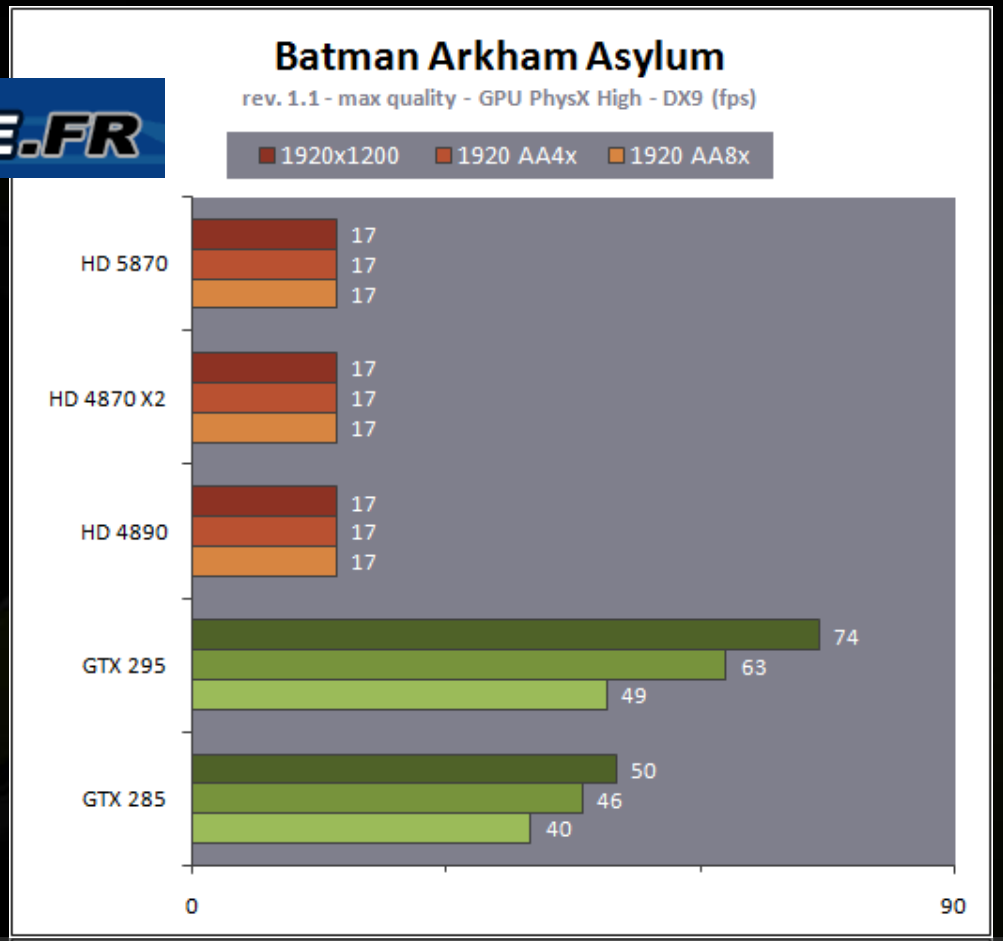
Batman: Arkham Asylum with PhysX!

- Smoke reacts with Batman, provides cover
- Walls explode
- Glass shatters
- Tattered curtains react with characters

Highest rated PC game since 2007!

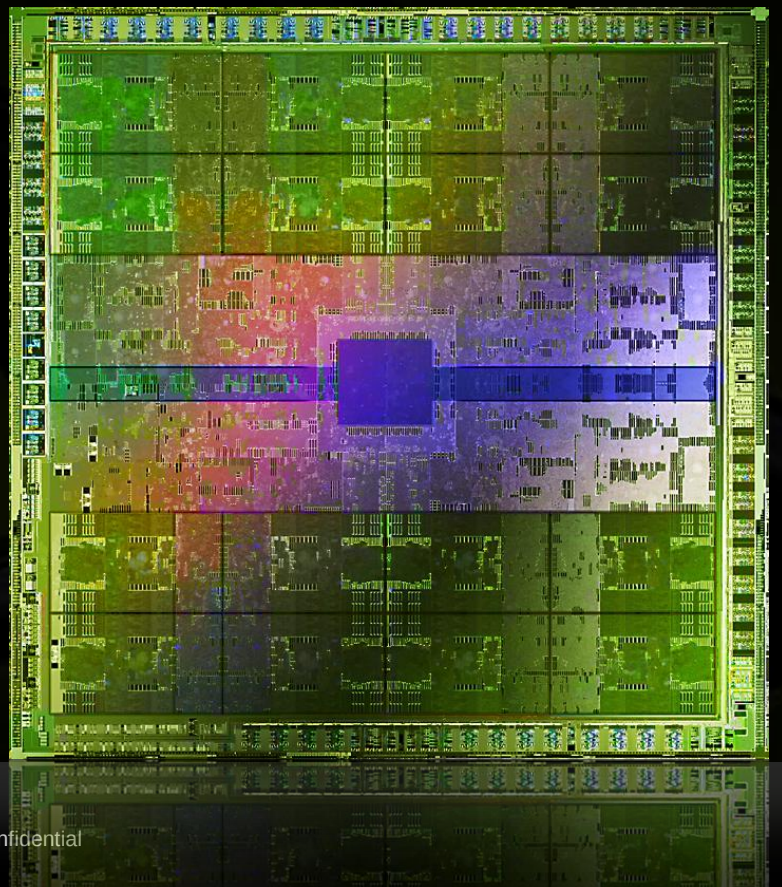
AMD Says HD 5870 is Fast….But It Doesn't Support Next Generation Games

HARDWARE.FR



The NVIDIA ‘Fermi’ Architecture

The Soul of a Supercomputer in the body of a GPU



- **Improved peak performance**
- **Improved efficiency throughput**
- **Full integration within modern development environment**

NVIDIA Confidential

Questions?
NVIDIA