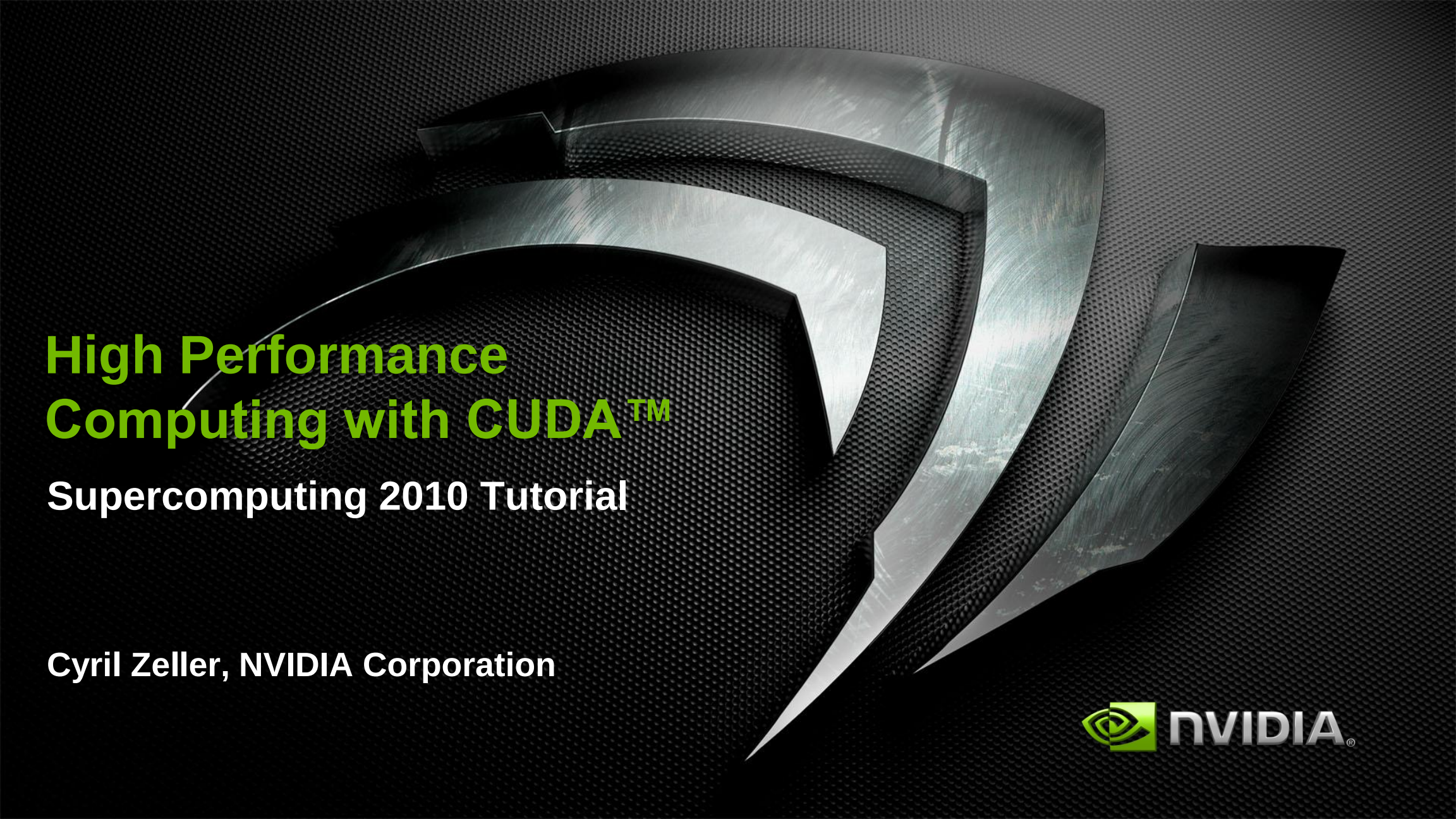




High Performance Computing with CUDA™

Supercomputing 2010 Tutorial

Cyril Zeller, NVIDIA Corporation



Welcome!

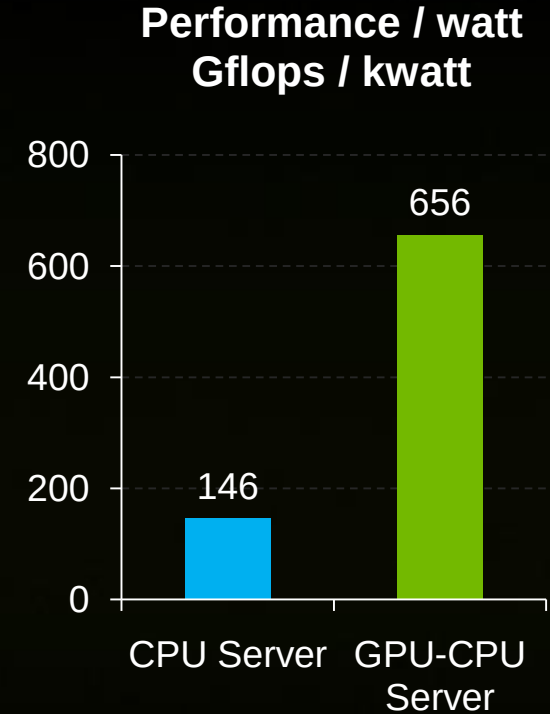
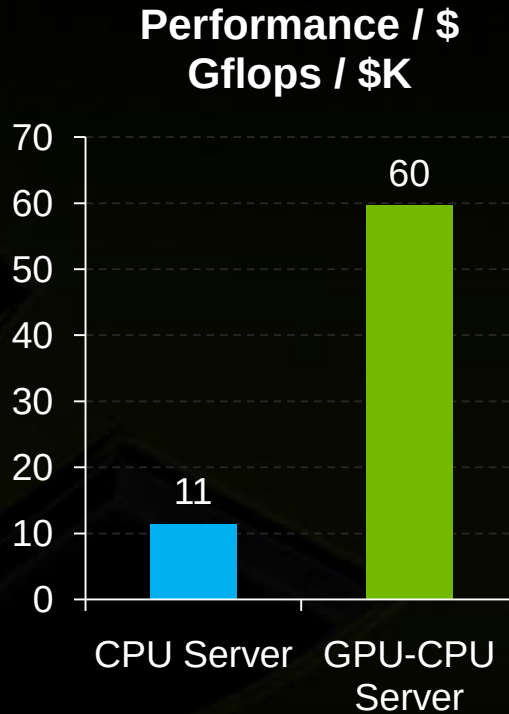
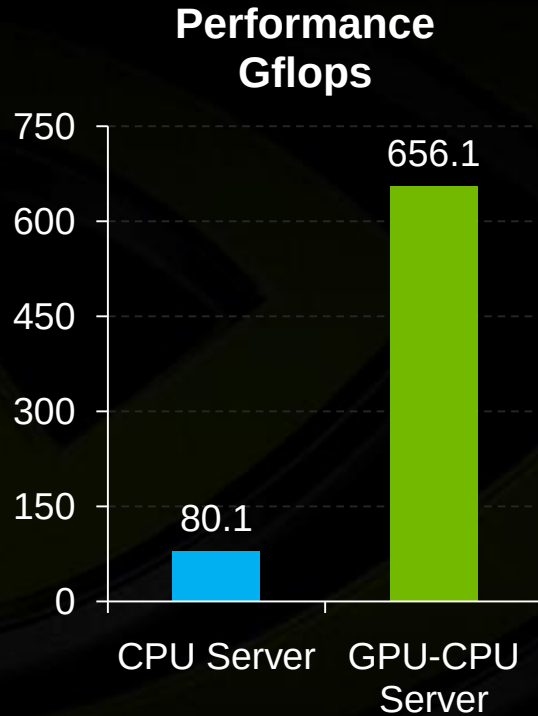


- **Goal:**
 - A detailed introduction to high performance computing with CUDA
 - CUDA = NVIDIA's architecture for GPU computing
- **Outline:**
 - Motivation and introduction
 - GPU programming basics
 - Code analysis and optimization
 - Lessons learned from production codes

GPUs are Fast!



8x Higher Linpack



CPU 1U Server: 2x Intel Xeon X5550 (Nehalem) 2.66 GHz, 48 GB memory, \$7K, 0.55 kw

GPU-CPU 1U Server: 2x Tesla C2050 + 2x Intel Xeon X5550, 48 GB memory, \$11K, 1.0 kw

Increasing Number of CUDA Applications



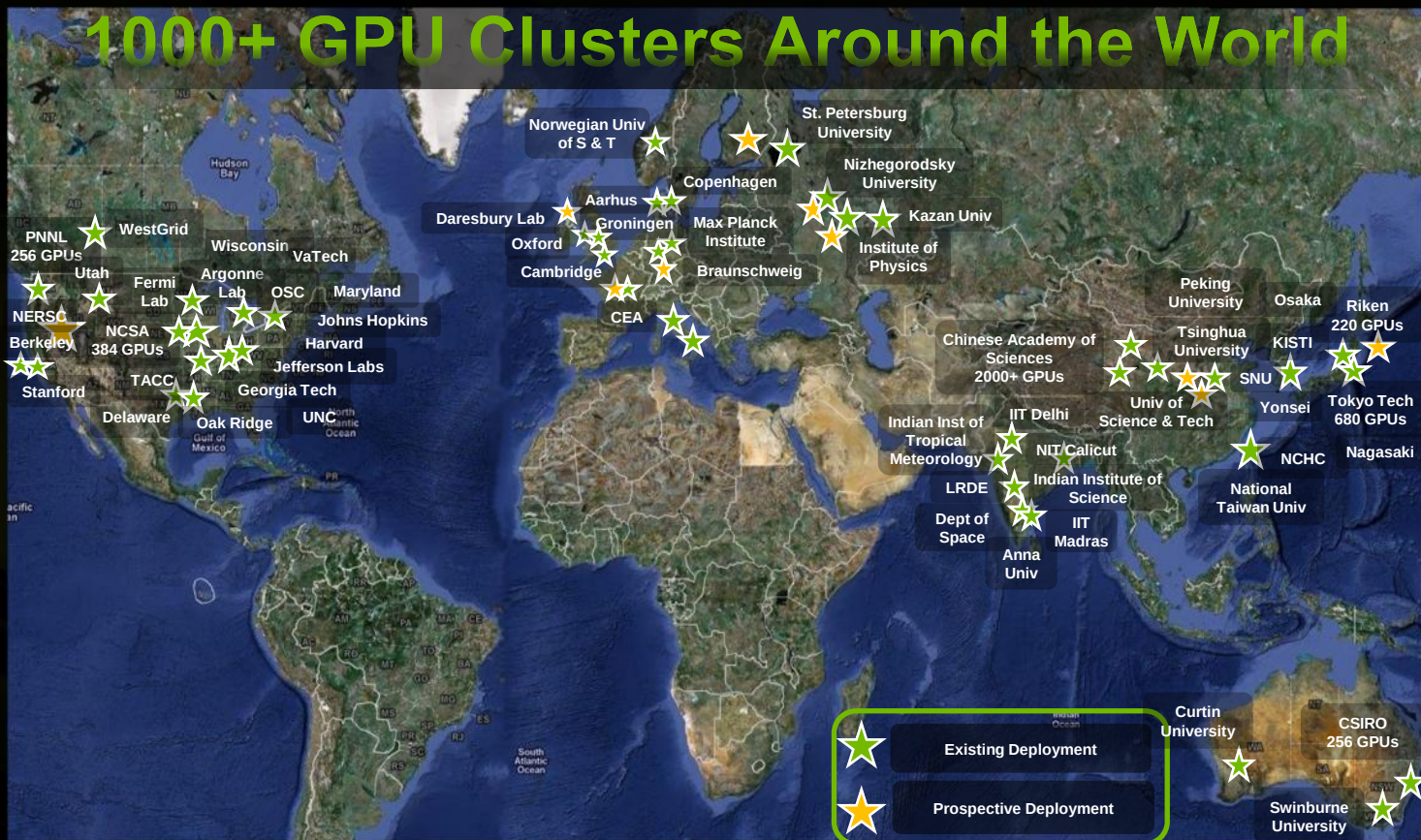
Already Available				2010				
				Q2		Q3		Q4
Tools & Libraries	CULA LAPACK Library	CUDA C/C++, PGI Fortran	Nsight Visual Studio IDE	Allinea Debugger	TotalView Debugger		PGI Accelerator Enhancements	
	Thrust: C++ Template Lib	Jacket: MATLAB Plugin	NPP Performance Primitives (NPP)	Platform Cluster Management	Bright Cluster Management	Mathematica	CAPS HMPP Enhancements	Mathworks MATLAB
Oil & Gas	Seismic Analysis: ffA, HeadWave	Seismic Analysis: Geostar	Seismic City, Acceleware			Seismic Interpretation	Reservoir Simulation 1	Reservoir Simulation 2
Bio- Chemistry	AMBER, GROMACS, GROMOS, HOOMD, LAMMPS, NAMD, VMD		BigDFT, ABINIT, TeraChem		Quantum Chem Code 1	Other Popular MD code	Quantum Chem Code 2	
Bio- Informatics	Hex Protein Docking	CUDA-BLASTP, GPU-HMMER, MUMmerGPU, MEME, CUDA-EC		Protein Docking		Short-read seq analysis		
Video & Rendering	Fraunhofer JPEG2K	OptiX Ray Tracing Engine	mental ray with iray	Main Concept Video Encoder	Elemental Video Live	3D CAD SW with iray		
Finance	NAG: RNGs	NumeriX: CounterParty Risk	Scicomp SciFinance	Risk Analysis 1	Risk Analysis 2	Credit Risk Analysis ISV	Trading Platform ISV	
CAE	AutoDesk Moldflow	OpenCurrent: CFD/PDE Library	Moldex3D	Acusim AcuSolve CFD	Structural Mechanics ISV	MSC MARC	MSC Nastran	Several CAE ISVs
EDA	Electro-magnetics: Agilent, CST, Remcom, SPEAG		Agilent ADS Spice Simulator			Verilog Simulator	Lithography Products	SPICE Simulator
100 000 active GPU computing developers						Released Product	Announced Product	Unannounced Product

Increasing Number of CUDA Installations

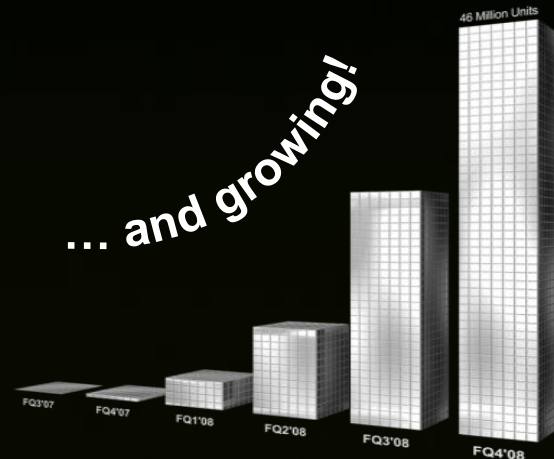


200 million
CUDA-capable GPUs
worldwide

1000+ GPU Clusters Around the World



... and growing!

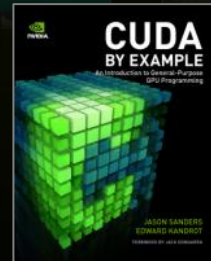
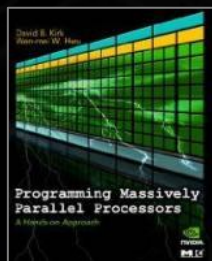


Pervasive CUDA Education



334
UNIVERSITIES

Teaching Parallel Programming using NVIDIA GPUs



GPU Computing Applications



Libraries and Middleware								
cuFFT	cuBLAS	CULA LAPACK	NPP & cuDPP	Video	PhysX Physics	OptiX Ray Tracing	mental ray iray Rendering	Reality Server 3D Web Services
C	C++	OpenCL [™]	Direct Compute	Fortran	Java and Python			



NVIDIA GPU

with the **CUDA** Parallel Computing Architecture

Fermi Architecture (Compute capabilities 2.x)	GeForce 400 Series	Quadro Fermi Series	Tesla 20 Series
Tesla Architecture (Compute capabilities 1.x)	GeForce 200 Series GeForce 9 Series GeForce 8 Series	Quadro FX Series Quadro Plex Series Quadro NVS Series	Tesla 10 Series
	 Entertainment	 Professional Graphics	 High Performance Computing

NVIDIA Tesla GPU Computing Products



Server Module



Tesla M2070 /
Tesla M2050



Tesla M1060

1U Systems



Tesla S2050



Tesla S1070

Workstation Boards



Tesla C2070 /
Tesla C2050



Tesla C1060

	Server Module		1U Systems		Workstation Boards	
GPUs	1 T20 GPU	1 T10 GPU	4 T20 GPUs	4 T10 GPUs	1 T20 GPU	1 T10 GPU
Single Precision	1030 GFlops	933 GFlops	4120 GFlops	4140 GFlops	1030 Gflops	933 GFlops
Double Precision	515 Gflops	78 GFlops	2060 GFlops	346 GFlops	515 Gflops	78 GFlops
Memory	6 GB / 3 GB	4 GB	12 GB (S2050)	16 GB 4 GB / GPU	6 GB / 3 GB	4 GB
Mem BW	148.4 GB/s	102 GB/s	148.4 GB/s	102 GB/s	144 GB/s	102 GB/s

NVIDIA Developer Ecosystem



Numerical Packages

MATLAB
Mathematica
NI LabView
pyCUDA

Debuggers & Profilers

cuda-gdb
NV Visual Profiler
Parallel Nsight
Visual Studio
Allinea
TotalView

GPU Compilers

C
C++
Fortran
OpenCL
DirectCompute
Java
Python

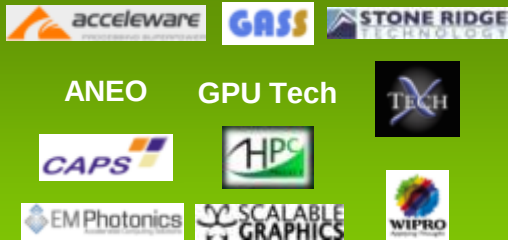
Parallelizing Compilers

PGI Accelerator
CAPS HMPP
mCUDA
OpenMP

Libraries

BLAS
FFT
LAPACK
NPP
Video
Imaging
GPULib

GPGPU Consultants & Training



OEM Solution Providers



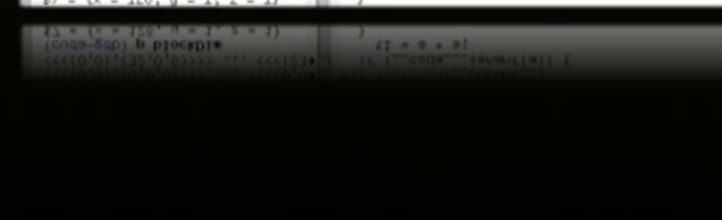
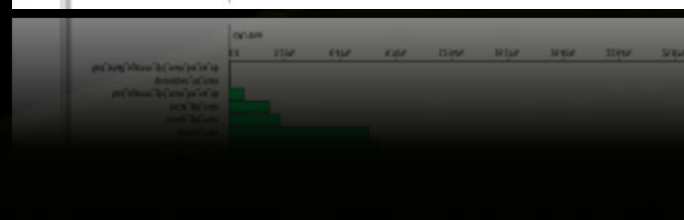
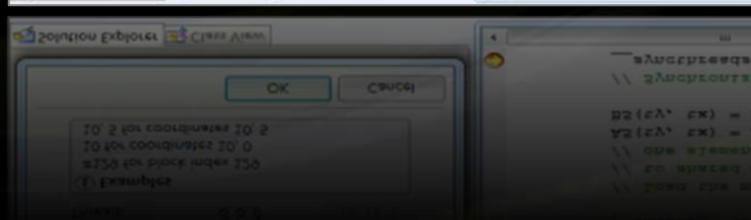
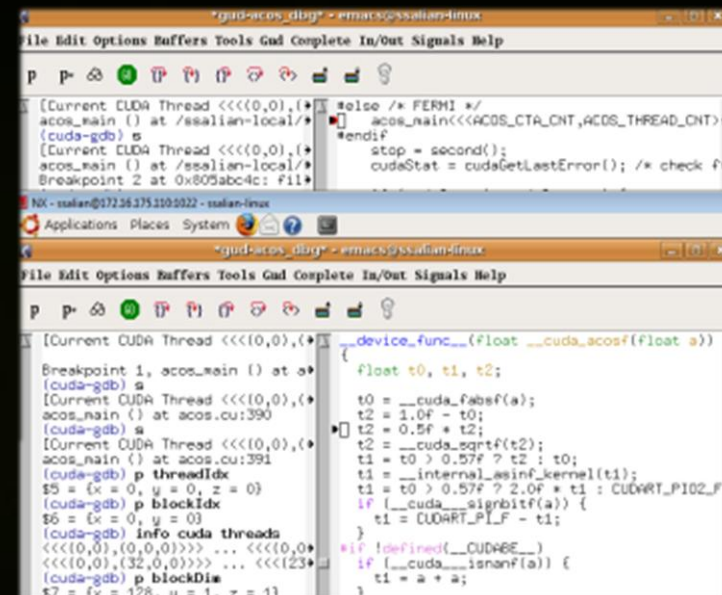
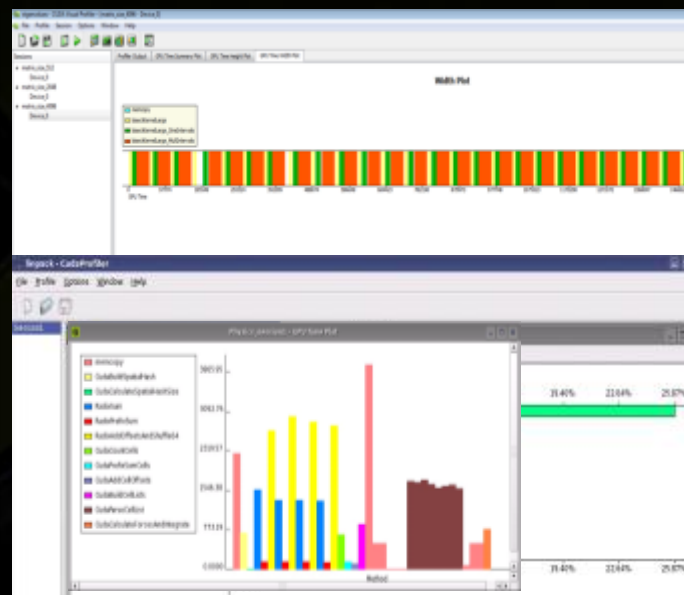
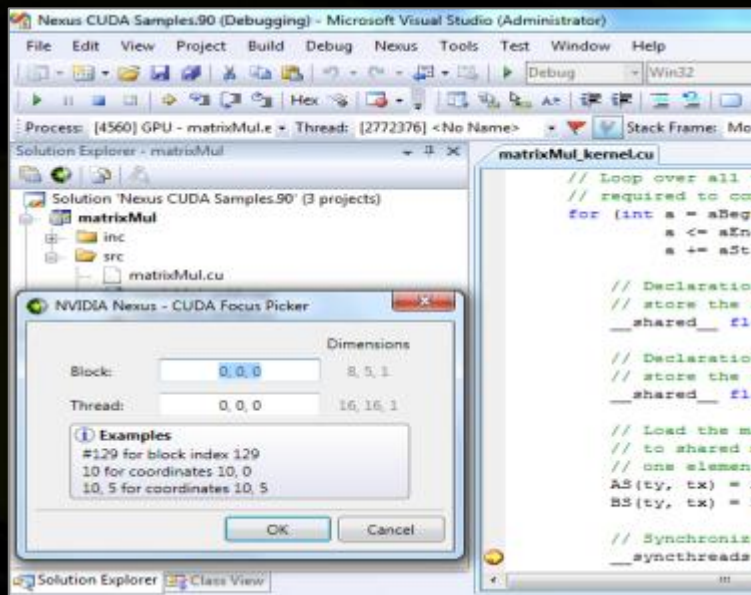


NVIDIA

Parallel Nsight Visual Studio

Visual Profiler For Linux

cuda-gdb For Linux



CUDA on the Web



- Last version of the CUDA Toolkit (currently version 3.2):
http://developer.nvidia.com/object/cuda_downloads.html
 - Includes drivers, compilers, libraries, debuggers, profilers, reference manuals, programming guides, code samples
- Online courses:
http://developer.nvidia.com/object/cuda_training.html
- CUDA books: http://www.nvidia.com/object/cuda_books.html
- CUDA forums: <http://forums.nvidia.com/index.php?showforum=62>
- And more on CUDA Zone:
http://www.nvidia.com/object/cuda_home_new.html

Schedule



8:30 **Introduction**

8:45 **CUDA C Basics**

Cyril Zeller, NVIDIA Corporation

9:45 **Break**

10:00 **Fundamental Optimizations**

11:00 **Analysis-Driver Optimization, Part 1**

Paulius Micikevicius, NVIDIA Corporation

12:00 **Lunch**

1:30 **Analysis-Driven Optimization, Part 2**

Paulius Micikevicius, NVIDIA Corporation

2:30 **Break**

Schedule



3:00 Industrial Seismic Imaging on a GPU Cluster

Scott Morton, Hess Corporation

3:30 Accelerating GPU Computation Through Mixed-Precision Methods

Mike Clark, Harvard University

4:00 Porting Large Fortran Codebases to GPUs

Andrew Corrigan, George Mason University

4:30 Heterogeneous GPU Computing for Molecular Modeling

John Stone, University of Illinois at Urbana-Champaign

5:00 End