What is GPU Computing?

Computing with CPU + GPU

Heterogeneous Computing
GPUs: Turning Point in Supercomputing

Desktop beats Cluster

CalcUA
$5 Million

Tesla Personal Supercomputer
$10,000

Source: University of Antwerp, Belgium
Not 2x or 3x: Speedups are 20x to 150x

- Medical Imaging
  - U of Utah
  - 146X

- Molecular Dynamics
  - U of Illinois, Urbana
  - 36X

- Video Transcoding
  - Elemental Tech
  - 18X

- MATLAB Computing
  - AccelerEyes
  - 50X

- Astrophysics
  - RIKEN
  - 100X

- Financial simulation
  - Oxford
  - 149X

- Linear Algebra
  - Universidad Jaime
  - 47X

- 3D Ultrasound
  - Techniscan
  - 20X

- Quantum Chemistry
  - U of Illinois, Urbana
  - 130X

- Gene Sequencing
  - U of Maryland
  - 30X
CUDA Parallel Computing Architecture

- Parallel computing architecture and programming model
- Includes a C compiler plus support for OpenCL and DX11 Compute
- Architected to natively support all computational interfaces (standard languages and APIs)
Tesla T10 GPU: 240 Processor Cores

Thread Processor (TP)

- Processor core has
  - Floating point / Integer unit
  - Move, compare, logic, branch unit
- IEEE 754 floating point
  - Single and Double
- 102 GB/s high-speed interface to memory

Thread Processor Array (TPA)

- Special Function Unit (SFU)
- Double Precision
- TP Array Shared Memory

30 TPAs = 240 Processors

GDDR3
512 bit
102 GB/sec
NVIDIA Tesla 10-Series GPU

Massively parallel, many core architecture

240 Processor Cores

1 Teraflops - 1,000 times Cray X-MP

IEEE Compliant Double Precision Floating Point

Designed for Scientific Computing
CUDA Facts

- 750+ Research Papers
- 85+ universities teaching CUDA
- 100 Million CUDA-Enabled GPUs
- 25K Active Developers

www.NVIDIA.com/CUDA
Background

• Who is AccelerEyes?
  – AccelerEyes is a MathWorks partner
  – Simple software for visual computing
Background

• Who is AccelerEyes?
  – AccelerEyes is a MathWorks partner
  – Simple software for visual computing
• What is Jacket?
  – GPU engine for MATLAB
  – CUDA powered language extension
Background

• Who is AccelerEyes?
  – AccelerEyes is a MathWorks partner
  – Simple software for visual computing
• What is Jacket?
  – GPU engine for MATLAB
  – CUDA powered language extension
• Why Jacket?
  – Challenges in technical computing
  – Low-cost speed, high-value graphics
  – Increased productivity
MATLAB Options

• CPU Solutions (blue arrows)
  – MATLAB and the Parallel Computing toolbox enable PC and clustered MATLAB computing

• GPU Solutions (green arrows)
  – Jacket enables CUDA MATLAB computing
Jacket combines the speed of CUDA and the graphics of the GPU with the user friendliness of MATLAB.
Jacket at Work

This Neural Net only required 3 changes (in red).
Functionality

- Generators: `geye`, `gones`, `gzeros`
- Element-wise: `+`, `*`, `-`, `/`
- Reductions: `sum`, `min`, `max` ...
- Indexing: subscripted referencing / subscripted assignment
- Linear Algebra: matrix multiply, ...
- FFT: `fft`, `ifft`, `fftn`, `ifftn`
- Filtering: `filter`, `filter2`, `convn`
- Interpolation: `interp2`
- Parallel for-loops: `gfor`
Graphics Toolbox

Jacket includes the Graphics Toolbox

- True visual computing
- OpenGL API in MATLAB
- Interactive OpenGL
- Key functions: gsurf, gimage, gscatter3, gplot, ...
- Visualization scripts are open and modifiable.
Application Benchmarks

Canny Edge Detection

- Image Size: 3936x3936, Time: 3.8 minutes, Speedup: 114x, Time: 7.2 hours
- Image Size: 2048x2048, Time: 8 seconds, Speedup: 1.86 hours
- Image Size: 1024x1024, Time: 2 seconds, Speedup: 28.4 minutes

Black-Scholes Simulation

- Vector Length: 3.84, Time: 0.13 seconds, Speedup: 143x, Time: 18.1 seconds
- Vector Length: 2.24, Time: 0.12 seconds, Speedup: 10.6 seconds
- Vector Length: 1.04, Time: 0.11 seconds, Speedup: 5.0 seconds

Tesla 8 Series GPU vs. Intel Core 2 Duo (2.4 GHz) for Canny Edge Detection

Tesla C1060 vs. Intel Core 2 Duo (2.0 GHz) for Black-Scholes Simulation
Custom CUDA Functions
Integration using MEX

#include <cudah.h>
#include "mex.h"

/* each element gets its index */
static __global__ void kernel(float *d_out, float *d_in)
{
    int x = blockIdx.x * blockDim.x + threadIdx.x;
    d_out[x] = d_in[x] + x;
}

void mexFunction(int nlhs, mxArray *plhs[], int nrhs, const mxArray *prhs[])
{
    /* attach to Jacket */
    C2context *ctx = C2context *(unsigned int)mxGetScalar(prhs[0]);
    cuCtxAttach(ctx, 0);

    /* get device pointer of gsingle */
    float *d_in = (float *)mxGetScalar(prhs[1]);

    /* run kernel to initialize 10 elements */
    float *d_out;
    cudaMalloc(void **d_out, 10*sizeof(float));
    kernel<<<1,2>>>(d_out, d_in);

    /* pull back to CPU and print */
    float h_out[10];
    cudaMemcpy(h_out, d_out, 10*sizeof(float), cudaMemcpyDeviceToHost);
    for (int i = 0; i < 10; i++)
        printf("%f\n", h_out[i]);

    /* detach from Jacket */
    cuCtxDetach(ctx);

    /* return device pointer */
    plhs[0] = mxCreateDoubleScalar((unsigned int)d_out);
}

---(Unix)--- mymex.cu Bot (53,0) (C/L)---(mexFunction)---17:54 [Wed]---
Some Jacket Customers
Roadmap for New Features

- more gfor
- gdouble
- multi-GPU support (for clusters of GPUs)
- LAPACK (eig, inv, etc.)
- signal processing
- image processing (and computer vision)
- Simulink® on the GPU
- statistical functions
- handle graphics
- lots of other MATLAB functions (finance, biology, etc.)
Tesla GPU Computing Products

Tesla S1070 1U System

- GPUs: 4 Tesla GPUs
- Single Precision Perf: 4.14 Teraflops
- Double Precision Perf: 346 Gigaflops
- Memory: 4 GB / GPU

Tesla C1060 Computing Board

- GPUs: 1 Tesla GPU
- Single Precision Perf: 933 Gigaflops
- Double Precision Perf: 78 Gigaflops
- Memory: 4 GB

Tesla Personal Supercomputer (4 Tesla C1060s)

- GPUs: 4 Tesla GPUs
- Single Precision Perf: 3.7 Teraflops
- Double Precision Perf: 312 Gigaflops
- Memory: 4 GB / GPU
Introducing the Tesla Personal Supercomputer

Supercomputing Performance
- Massively parallel CUDA Architecture
- 960 cores. 4 TeraFlops
- 250x 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
### A $5 Million Datacenter

#### CPU 1U Server
- **2 Quad-core Xeon CPUs:** 8 cores
  - 0.17 Teraflop (single)
  - 0.08 Teraflop (double)
  - $3,000
  - 700 W

#### Tesla 1U System
- **8 CPU Cores + 4 GPUs = 968 cores**
  - 4.14 Teraflops (single)
  - 0.346 Teraflop (double)
  - $11,000
  - 1500 W

#### CPU 1U Server
- **1819 CPU servers**
  - 310 Teraflops (single)
  - 155 Teraflops (double)
  - Total area 16K sq feet
  - Total 1273 KW

#### Tesla 1U System
- **455 CPU servers**
  - 1961 Teraflops (single)
  - 196 Teraflops (double)
  - Total area 9K sq feet
  - Total 682 KW

#### Comparisons
- **6x more perf**
- **40% smaller**
- **1/2 the power**
Tesla S1070: Green Supercomputing

20X Better Performance / Watt

- Hess
- Chevron
- Petrobras
- NCSA
- CEA
- Tokyo Tech
- JFCOM
- SAIC
- Federal
- Motorola
- Kodak
- University of Heidelberg
- University of Illinois
- University of North Carolina
- Max Planck Institute
- Rice University
- University of Maryland
- Eotvas University
- University of Wuppertal
- Chinese Academy of Sciences
- National Taiwan University
More Than 250 Customers / ISVs

<table>
<thead>
<tr>
<th>Life Sciences &amp; Medical Equipment</th>
<th>Productivity / Misc</th>
<th>Oil and Gas</th>
<th>EDA</th>
<th>Finance</th>
<th>CAE / Mathematical</th>
<th>Communication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Max Planck</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Robarts Research</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medtronic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AGC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Evolved machines</td>
<td>GE Healthcare</td>
<td>CEA</td>
<td>Hess TOTAL</td>
<td>Symcor</td>
<td>AccelerEyes</td>
<td></td>
</tr>
<tr>
<td>Smith-Waterman</td>
<td>Siemens</td>
<td>NCSA</td>
<td>CGG/Veritas</td>
<td>Level 3</td>
<td>MathWorks</td>
<td>Nokia</td>
</tr>
<tr>
<td>DNA sequencing</td>
<td>Techniscan</td>
<td>WRF Weather Modeling</td>
<td>Chevron</td>
<td>SciComp</td>
<td>Wolfram</td>
<td>RIM</td>
</tr>
<tr>
<td>AutoDock</td>
<td>Boston Scientific</td>
<td>OptiTex</td>
<td>Headwave</td>
<td>Hanweck</td>
<td>National Instruments</td>
<td>Philips</td>
</tr>
<tr>
<td>Folding@Home</td>
<td>Eli Lilly</td>
<td>Tech-X</td>
<td>Acceleware</td>
<td>Quant</td>
<td>Access Analytics</td>
<td>Samsung</td>
</tr>
<tr>
<td>Howvard Hughes</td>
<td>Silicon Informatics</td>
<td>Elemental Technologies</td>
<td>Seismic City</td>
<td>Catalyst</td>
<td>Ansys</td>
<td>LG</td>
</tr>
<tr>
<td>Medical</td>
<td>Stockholm Research</td>
<td>Dimensional Imaging</td>
<td>P-Wave</td>
<td>RogueWave</td>
<td>CST</td>
<td>Sony</td>
</tr>
<tr>
<td>CRIBI Genomics</td>
<td>Harvard</td>
<td>Manifold</td>
<td>Seismic Imaging</td>
<td>Boeing</td>
<td>Synthetron</td>
<td>Ericsson</td>
</tr>
<tr>
<td></td>
<td>Delaware</td>
<td>Digisens</td>
<td>Mercury</td>
<td>Riken</td>
<td>SOFA</td>
<td>NTT DoCoMo</td>
</tr>
<tr>
<td></td>
<td>Pittsburg</td>
<td>General Mills</td>
<td>Computer</td>
<td>Radio</td>
<td>Seismic City</td>
<td>Mitsubishi</td>
</tr>
<tr>
<td></td>
<td>ETH Zurich</td>
<td>Rapidmind</td>
<td>ffA</td>
<td>Research</td>
<td>Imaging</td>
<td>Hitachi</td>
</tr>
<tr>
<td></td>
<td>Institute Atomic</td>
<td>Rhythm &amp; Hues</td>
<td></td>
<td>Laboratory</td>
<td>Modeling</td>
<td>Radio</td>
</tr>
<tr>
<td></td>
<td>Physics</td>
<td>xNormal</td>
<td></td>
<td></td>
<td></td>
<td>Research</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Elcomsoft</td>
<td></td>
<td></td>
<td></td>
<td>Laboratory</td>
</tr>
<tr>
<td></td>
<td></td>
<td>LINZIK</td>
<td></td>
<td></td>
<td></td>
<td>US Air Force</td>
</tr>
</tbody>
</table>
More Information

- **Tesla main page**
  - [http://www.nvidia.com/tesla](http://www.nvidia.com/tesla)
  - Industry Solutions

- **CUDA Zone**
  - CUDA Tutorials, Applications

- **Hear from Developers**
  - [http://www.youtube.com/nvidiatesla](http://www.youtube.com/nvidiatesla)

- **Download Jacket Now**

- **Further Jacket Questions**
  - [http://www.accelereyes.com/forums](http://www.accelereyes.com/forums)
  - [http://www.accelereyes.com/blog](http://www.accelereyes.com/blog)

Sumit Gupta
sumitg@nvidia.com

John Melonakos
john.melonakos@accelereyes.com
NVIDIA: Leadership in GPU computing

100s of Apps on CUDA Zone

100 M CUDA enabled GPUs
25,000+ active developers

30+ CUDA GPU clusters

50+ Universities Teaching CUDA
750+ research papers

Duke
Northeastern
Erlangen
Oregon State
ETH Zurich
Pennsylvania
Georgia Tech
Polimi
Grove City College
Purdue
Harvard
Santa Clara
IISc Bangalore
Stanford
IIT
Stuttgart
IIIT Hyderabad
Suny
Illinois
Tokyo
INRIA
TU-Vienna
Iowa
USC
ITESM
Utah
Johns Hopkins
Virginia
Kent State
Washington
Kyoto
Waterloo
Lund
Western Australia
Maryland
Williams College
McGill
Wisconsin
MIT
Yonsei
North Carolina

CUDA 1.0
CUDA 1.1
CUDA 2.0

150K CUDA compiler downloads