The Future of Massively Parallel and GPU Computing
Poznaj świadectwa z dokumentacji procesu beatyfikacyjnego i kanonizacyjnego Jana Pawła II

CUDA

już od 7 marca

Książka wraz z filmem VCD „Cud” dostępna w księgarniach, parafiach, salonach EMPiK, w księgarniach internetowych i na stronie www.stanislawbm.pl tel. 012 429 52 17
CUDA Uses Kernels and Threads for Fast Parallel Execution

Parallel portions of an application are executed on the GPU as kernels.

- One kernel is executed at a time
- Many threads execute each kernel

Differences between CUDA and CPU threads

- CUDA threads are extremely lightweight
  - Very little creation overhead
  - Instant switching
- CUDA uses 1000s of threads to achieve efficiency
  - Multi-core CPUs can use only a few
void saxpy_serial(int n, float a, float *x, float *y)
{
    for (int i = 0; i < n; ++i)
        y[i] = a*x[i] + y[i];
}
// Invoke serial SAXPY kernel
saxpy_serial(n, 2.0, x, y);

__global__ void saxpy_parallel(int n, float a, float *x, float *y)
{
    int i = blockIdx.x*blockDim.x + threadIdx.x;
    if (i < n)  y[i] = a*x[i] + y[i];
}
// Invoke parallel SAXPY kernel with 256 threads/block
int nblocks = (n + 255) / 256;
saxpy_parallel<<<nblocks, 256>>>(n, 2.0, x, y);
The Key to Computing on the GPU

- Standard high level language support
  - C, soon C++ and Fortran
  - Standard and domain specific libraries

- Hardware Thread Management
  - No switching overhead
  - Hide instruction and memory latency

- Shared memory
  - User-managed data cache
  - Thread communication / cooperation within blocks

- Runtime and tool support
  - Loader, Memory Allocation
  - C stdlib
CUDA Programming Model

A kernel is executed by a grid of thread blocks

- A thread block is a batch of threads that can cooperate:
  - Sharing data through shared memory
  - Synchronizing their execution

- Threads from different blocks operate independently
Kernel Memory Access

- **Registers**
  - Global Memory
    - Kernel input and output data reside here
    - Off-chip, large
    - Uncached
  - Shared Memory
    - Shared among threads in a single block
    - On-chip, small
    - As fast as registers

- The host can read & write global memory but not shared memory
**Example Fluid Algorithm**

**CPU**
- Control
- ALU
- Cache
- DRAM

Single thread out of cache

**GPGPU**
- Control
- ALU
- P_1
- P_2
- P_3
- P_4
- Video Memory

Multiple passes through video memory

**GPU Computing with CUDA**

- Thread Execution Manager
- ALU
- Shared Data
- P_1
- P_2
- P_3
- P_4
- P_5

Parallel execution on-chip
100M CUDA GPUs

CUDA

Heterogeneous Computing

Oil & Gas  Finance  Medical  Biophysics  Numerics  Audio  Video  Imaging
Parallel Computing on All GPUs
Almost 100 Million CUDA GPUs Deployed

GeForce®
Entertainment

Tesla™
High-Performance Computing

Quadro®
Design & Creation
Developer Categories

- Apparel design and simulation
- Dental x-ray system
- Image analysis from confocal microscope
- Image capture
- Laser scanning feature extraction
- Acoustic and electromagnetic simulation
- Acoustic ray tracing
- Adaptive radiation therapy
- Airline trimming
- Analysis of electroencephalograms
- Animation
- Astronomical adaptive optics
- Astronomical imaging
- Astronomics image scanning system
- Astrophysics simulation
- Astrophysics simulation
- Audio conferencing enhancement
- Audio processing
- Audio rendering of complex scenes
- Audio visual editing and scripting
- Automated Web page classification
- Automobile vision system
- Automotive vision system
- Biochannel simulations
- Bioinformatics
- Bioinformatics for protein structure and cellular modeling
- Bioinformatics for sequence alignment
- Biological circuits
- Biological imaging
- Biological simulation using evolutionary algorithms
- Biological simulations
- Biomedical cell imaging
- Biomedical image registration and segmentation
- Biomedical imaging

- Biomimetic neural network simulation
- BOINC cluster
- Broadcast graphics
- Broadcast production
- Business analytics
- Business intelligence
- C#
- CAD
- Call center analysis
- Casting simulation
- Cell phone
- Cellular automata for organizational behavior
- Com SSIS space reconstruction
- CFD
- CFD for high speed aircraft engine design
- CFD for ocean modeling
- CFD with particle flow
- Chess
- Chromatography analysis
- Climate models
- Cloth simulation
- Color correction for film
- Color correction for projects
- Computer design simulation
- Computer vision
- Computer vision for food inspection
- Computer vision simulation of primate vision
- Constraint fluid simulation
- Consulting
- Corporate data analysis
- Cosmological simulations
- Crash simulation
- Cryptography
- Crystallography
- CT Image reconstruction

- Cytogenetics
- Data mining
- Data reduction software for crystallography
- Database search
- Defibrillator design
- Dental CT scanner
- Design for manufacturing
- Design for manufacturing software
- Diabetic retinal analysis
- Digital audio
- Digital cinema image reconstruction
- Digital film processing
- Digital image correlation
- Digital projector
- Digital prototyping
- Digital speech processing
- Digital video management
- Digital video recorder
- DNA analysis
- DNA gene expression data analysis
- DNA research
- DNA sequence analysis
- Document data mining
- Dredging simulator
- DSP
- DVD distribution
- Earthquake engineering FEA
- Economic modeling
- EDA
- EDA
- Electromagnetic simulation
- Electron CAD flow model
- Elementary particle research
- Email and web security
- Equity trading
Developer Categories

Exact real arithmetic
Face recognition
Facial recognition
Factory design management
FEM in CFD and chemical processes
Film
Film and video production
Film animation
Film processing
Film special effects
Film visual effects
Financial option pricing
Financial pricing
Financial risk analysis
Financial trading
Fingerprint matching
Finite element simulation
Finite element solver
FLASH - adaptive mesh fluid simulation
Flight training simulations
Floodplain simulation
Flow Cytometer
Flow visualization
Fluid dynamics
Fluid dynamics
Fluid flow simulation
Fluid simulation
Fluorescence Lifetime imaging
Folding at Home
Folding at home clone
 Formal verification methods
Fortran, C/C++ compilers
Games
Gene sequence alignment
Genetics
Geomatics using discrete or finite element analysis imaging for security
Geometric modeling
Geophysical imaging
Geospatial image processing
GIS
Graphics
Graphics jpeg viewer
Grid computing
Grid computing
Grid media encoding
Harbor management - vessel navigation
HDR display
Health care sensory processing
High end imaging for professional photography
Holographic cinema
Holographic cinema
Holographic optical trapping
Human language analysis
Hydraulics simulation
Hydrodynamics
Hyperspectral image analysis
IC CAD
Image analysis
Image analysis for cancer research
Image analysis for surveillance systems
Image compression
Image data mining
Image enhancement
Image feature tracking on high speed video
Image processing
Image processing
Image registration
Image scanning
Image tracking for brain research
Imaging for defect detection
Imaging in high end digital imaging
Immersive display
In flight entertainment system
Infectious disease simulation
Infrared imaging
Infrared imaging
Injection molding CAD software
Interactive TV graphics
Interest rate risk calculation
Internet video compression for distribution
IPTV
IPTV format conversion
Language
Language - CSAIL
Language - MPI extensions
Large format imaging
Large scale neural networks
Large text database search
Linear programming
LISTSERV email list management
Machine automation
Machine learning
Machine vision
Manufacturing simulation
Mathematics - 3D framework
Mathematics - Computation geometry
Mathematics - fast multipole method
Mathematics - fractals
Mathematics - linear algebra
Mathematics - LSF-SGE
Mathematics - projective space
Mathematics library
Mathematics research - algebraic surface visualization
Mathematics research - interior point methods
Developer Categories

Military - SONAR
Military - swimmer detection sonar
Military - training
Military - UAV image processing
Military - Weapons systems physics
Military hyperspectral target detection
Military target modelling
Mine planning
Mixed signal data processing for testing
Molecular dynamics simulation
Molecular dynamics simulation
Molecular dynamics simulation
Molecular dynamics simulation
Molecular modelling
Molecular properties classification
Molecular simulation
Molecular simulation - GROMACS
Molecular structure simulation
Molecular visualization
Motion capture
Movie production special effects
MPEG
MPEG2 decode
MRI analysis of brain function
MRI image reconstruction
MRI imaging
Multibody simulations
Multiphase flow simulator
Multipath scene generation
Nano-carbon materials molecular dynamics
Natural language processing
nbody simulation
Netflix competition
Network analysis
Network hub line card

Network load balancer
Network packet inspection
Network processing
Network processor
Network security monitoring
Neural net AI
Neural network research
Neural networks for computer vision
Neuron modelling with XPP
Nighttime driving simulator
NMR data analysis
N-particle code for particle transport
Nuclear reactor physics simulation
Object recognition
Oceanographic research
Octopus molecular simulation
Online mapping
Open source mathematics software
Optical inspection
Optical modelling and engineering
Optical processing
Optical security scanner
Optical simulation
Optronics scene simulator
Orbital analysis
PACS medical record storage
Particle physics
Particle visualization
Pattern analysis tools for neuroimaging
PCB optical inspection
Physics engine
Plasma particle simulation
Power generation statistics
Print pre-processing
Probabilistic model checker

Programmable automation controllers
Protein crystallography
Protein folding
Protein structure prediction and design
Proteomics data diagnostics
Pulsar data analysis
Quantum chemistry
Quantum Chromodynamical calculations
Quantum molecular dynamics
Radar processing
Radar simulation
Radiation therapy machine
Radion astronomy
Ray tracing
Real time rendering
Real time signal processing
Realtime live video encoding
Realtime simulation of machining
Remote graphics
Research - astrophysics
Research - developmental biology
Research - fire simulation, cellular automata
Research - image segmentation
Research - Large particle physics simulation
Research - Mars instruments
Research - optical tracking
Research - reconfigurable computing
Research - visualization
Reservoir simulation
Reservoir simulation
Robot vision
Robot vision
Robotic AI
Robotic radiation therapy machines
Robotic surgery
Developer Categories

Robotic vision
RSA factoring
RTFSS
SAR
Satellite data analysis
Satellite data processing
Satellite development simulators
Satellite image processing
Scanning electron microscope imaging
Scientific data mining
Scientific numerical simulation
Scientific visualization
Search engine
Seismic damage simulation
Seismic imaging
Seismic processing
SIFT algorithm research
Signal processing
Simulation of micro and nano biochemical reactors
Small molecule dynamics simulation
Smoother particle hydrodynamics
Sound synthesis
South Pole Telescope data analysis
Spatial data integration
Spatial heart modelling
Spectral Imaging
Spectroscopic data optimization
Speech processing
Speak recognition
Sports broadcasting enhancement
Statistical analysis
Sterographic vision
Stock market fraud detection
Structural simulation
Surgery simulator

Surveillance research
Surveillance system
Television broadcast
Temperature simulation for architecture
Traffic analysis
Train-track interaction analysis software
Transaction query for mobile commerce
Ultrasound inspection and testing
Ultrasound imaging
Ultrasound medical imaging
Unlimited precision mathematics
Urban 3D models from video streams
Video and audio finishing
Video compression
Video compression with cupola technology
Video compression
Video conferencing
Video editing
Video effects generator
Video encoding
Video enhancement
Video processing
Virus scanning
Vision-aided navigation for robotics
Visual information system
Visual search
Visualization
Volume rendering
Wall turbulent flows
Weather forecasting
Web conferencing
Wind engineering for urban and rural environment
Wireless network simulation software
Wireless system design
X-ray tomosynthesis
Developers by Category

Registered developers who downloaded both CUDA 0.8 and 1.0

- Manufacturing: 131
- Digital Content: 156
- Imaging: 378
- Medical: 214
- Financial: 126
- Academia: 445
- Research: 686
- Scientific: 677
- Gaming: 141
- Physics: 219
- Biosciences: 140
- Chemical: 58
- Geoscience: 106
- Defense: 162
### Parallel Computing Applications

<table>
<thead>
<tr>
<th>Consumer</th>
<th>Business</th>
<th>Workstation</th>
<th>Technical</th>
</tr>
</thead>
<tbody>
<tr>
<td>Imaging Video – transcoding</td>
<td>Search Web</td>
<td>Oil and gas viz CAD</td>
<td>Seismic Finance Numerics*</td>
</tr>
<tr>
<td>Camera – Physics, AI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Computer vision</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Grid computing</td>
<td>XML parsing</td>
<td>Volume visualization Cluster visualization</td>
<td>Medical imaging</td>
</tr>
<tr>
<td>Audio</td>
<td>Database</td>
<td></td>
<td>EDA</td>
</tr>
<tr>
<td>Photography</td>
<td>VPN/networking</td>
<td></td>
<td>CAE</td>
</tr>
<tr>
<td>Virus scanning</td>
<td>Backup compression RAID</td>
<td></td>
<td>GIS</td>
</tr>
<tr>
<td>Computer vision</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Universities Teaching Parallel Programming With CUDA

<table>
<thead>
<tr>
<th>Duke</th>
<th>Kent State</th>
<th>Santa Clara</th>
</tr>
</thead>
<tbody>
<tr>
<td>Erlangen</td>
<td>Kyoto</td>
<td>Stanford Stuttgart</td>
</tr>
<tr>
<td>ETH Zurich</td>
<td>Lund</td>
<td>Suny</td>
</tr>
<tr>
<td>Georgia Tech</td>
<td>Maryland</td>
<td>Tokyo</td>
</tr>
<tr>
<td>Grove City College</td>
<td>McGill</td>
<td>TU-Vienna</td>
</tr>
<tr>
<td>Harvard</td>
<td>MIT</td>
<td>USC</td>
</tr>
<tr>
<td>IIIT</td>
<td>North Carolina - Chapel Hill</td>
<td>Utah</td>
</tr>
<tr>
<td>IIT</td>
<td>North Carolina State</td>
<td>Virginia</td>
</tr>
<tr>
<td>Illinois Urbana-Champaign</td>
<td>Northeastern</td>
<td>Washington</td>
</tr>
<tr>
<td>INRIA</td>
<td>Oregon State</td>
<td>Waterloo</td>
</tr>
<tr>
<td>Iowa</td>
<td>Pennsylvania</td>
<td>Western Australia</td>
</tr>
<tr>
<td>IITEM</td>
<td>Polimi</td>
<td>Williams College</td>
</tr>
<tr>
<td>Johns Hopkins</td>
<td>Purdue</td>
<td>Wisconsin</td>
</tr>
</tbody>
</table>
Wide Developer Acceptance

146X
Interactive visualization of volumetric white matter connectivity

36X
Ionic placement for molecular dynamics simulation on GPU

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
Folding@Home Using GROMACS

- Alzheimer’s Disease
- Huntington’s Disease
- Cancer
- Osteogenesis imperfecta
- Parkinson’s Disease
- Antibiotics
Science: National Center for Atmospheric Research

Weather Research and Forecast (WRF) model

400+ registered users worldwide

20% speedup with 1% of WRF on CUDA

Saves 1 week analysis time
## Finance: Real-time Options Valuation

Hanweck Associates Volera real-time option valuation engine
Value the entire U.S. listed options market in real-time using 3 NVIDIA Tesla S870’s

<table>
<thead>
<tr>
<th></th>
<th>GPUs</th>
<th>CPUs</th>
<th>Savings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processors</td>
<td>12</td>
<td>600</td>
<td></td>
</tr>
<tr>
<td>Rack Space</td>
<td>6U</td>
<td>54U</td>
<td>9x</td>
</tr>
<tr>
<td>Hardware Cost</td>
<td>$42,000</td>
<td>$262,000</td>
<td>6x</td>
</tr>
<tr>
<td>Annual Cost</td>
<td>$140,000</td>
<td>$1,200,000</td>
<td>9x</td>
</tr>
</tbody>
</table>

Figures assume:
- NVIDIA Tesla S870s with one 8-core host server per unit
- CPUs are 8-core blade servers; 10 blades per 7U
- $1,800/U/month rack and power charges
- 5-year depreciation
Design: CAD Design For Apparel Cloth Physics
GIS Application

From the Manifold 8 feature list:

... applications fitting CUDA capabilities that might have taken tens of seconds or even minutes can be accomplished in hundredths of seconds. ... CUDA will clearly emerge to be the future of almost all GIS computing

From the user manual:

"NVIDIA CUDA ... could well be the most revolutionary thing to happen in computing since the invention of the microprocessor"
nbody Astrophysics

Astrophysics research
1 GF on standard PC
300+ GF on GeForce 8800GTX
Faster than GRAPE-6Af custom simulation computer

Video demo

http://progrape.jp/c
OmegaSim GX - Spice Simulation with CUDA

- 40x Speedup for transistor evaluation
- Up to 90% of SPICE execution time spent in transistor evaluation
- Avg. 8x overall speedup
Gauda Optical Proximity Correction (OPC)

200x Faster and Lower Cost

Time-to-Money

100%

60%

Time-to Market

Delay (3mo.)

Typical 1 Yr Life-Cycle

Cost

Revenue

hours

days

$100K

$1 M

$10 M

10's CPUs

1000's CPUs

GAUDA

CPU

FPGA
Evolved Machines

Simulate the brain circuit

Sensory computing: vision, olfactory

130X Speed up
Matlab: Language of Science

18X with MATLAB CPU+GPU


Pseudo-spectral simulation of 2D Isotropic turbulence

http://www.amath.washington.edu/courses/571-winter-2006/matlab/FS_2Dturb.m
Faster is not “just Faster”

- 2-3X faster is “just faster”
  - Do a little more, wait a little less
  - Doesn’t change how you work
- 5-10x faster is “significant”
  - Worth upgrading
  - Worth re-writing (parts of) the application
- 100x+ faster is “fundamentally different”
  - Worth considering a new platform
  - Worth re-architecting the application
  - Makes new applications possible
  - Drives “time to discovery” and creates fundamental changes in Science
Tesla T10: 1.4 Billion Transistors

Thread Processor Cluster (TPC)

Thread Processor Array (TPA)

Thread Processor
Double the Performance

1.5 Gigabytes 4 Gigabytes
Tesla 8 Tesla 10

Double the Memory

500 Gigaflops 1 Teraflop
Tesla 8 Tesla 10

Double the Precision

Finance Science Design
### Tesla T10 Double Precision Floating Point

<table>
<thead>
<tr>
<th>Feature</th>
<th>Specification</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Precision</strong></td>
<td>IEEE 754</td>
</tr>
<tr>
<td><strong>Rounding modes for FADD and FMUL</strong></td>
<td>All 4 IEEE, round to nearest, zero, inf, -inf</td>
</tr>
<tr>
<td><strong>Denormal handling</strong></td>
<td>Full speed</td>
</tr>
<tr>
<td><strong>NaN support</strong></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Overflow and Infinity support</strong></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Flags</strong></td>
<td>No</td>
</tr>
<tr>
<td><strong>FMA</strong></td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Square root</strong></td>
<td>Software with low-latency FMA-based convergence</td>
</tr>
<tr>
<td><strong>Division</strong></td>
<td>Software with low-latency FMA-based convergence</td>
</tr>
<tr>
<td><strong>Reciprocal estimate accuracy</strong></td>
<td>24 bit</td>
</tr>
<tr>
<td><strong>Reciprocal sqrt estimate accuracy</strong></td>
<td>23 bit</td>
</tr>
<tr>
<td><strong>log2(x) and 2^x estimates accuracy</strong></td>
<td>23 bit</td>
</tr>
</tbody>
</table>
Double the Performance Using T10

DNA Sequence Alignment

Dynamics of Black holes

Video Application

Cholesky Factorization

LB Flow Lighting

Ray Tracing

Reverse Time Migration
How to Get to 100X?

*Traditional Data Center Cluster*

Quad-core CPU

8 cores per server

1000’s of cores
1000’s of servers

*More Servers To Get More Performance*
Linear Scaling with Multiple GPUs

Oil and Gas Computing: Reverse Time Migration
Hand Optimized SSE Versus CUDA

GFLOPS

Number of Cores

X86 CPU

NVIDIA GPU
Heterogeneous Computing Cluster

10,000’s processors per cluster

- Hess
- NCSA / UIUC
- JFCOM
- SAIC
- University of North Carolina
- Max Plank Institute
- Rice University
- University of Maryland
- GusGus
- Eotvas University
- University of Wuppertal
- IPE/Chinese Academy of Sciences
- Cell phone manufacturers
Building a 100TF datacenter

CPU 1U Server

- 4 CPU cores
- 0.07 Teraflop
- $2000
- 400 W
- 1429 CPU servers
- $3.1 M
- 571 KW

Tesla 1U System

- 4 GPUs: 960 cores
- 4 Teraflops
- $8000
- 700 W
- 25 CPU servers
- 25 Tesla systems
- $0.31 M
- 27 KW

10x lower cost
21x lower power
Tesla S1070 1U System

4 Teraflops\(^1\)

700 watts\(^2\)

\(^1\) single precision
\(^2\) typical power
Tesla C1060 Computing Processor

957 Gigaflops\(^1\)
160 watts\(^2\)

\(^1\) single precision
\(^2\) typical power
What’s Next for CUDA

- Fortran
- C++
- Multiple GPUs
- Debugger
- Profiler
- GPU Cluster
### Application Software

**Industry Standard C Language**

<table>
<thead>
<tr>
<th>Libraries</th>
</tr>
</thead>
<tbody>
<tr>
<td>cuFFT</td>
</tr>
<tr>
<td>cuBLAS</td>
</tr>
<tr>
<td>cuDPP</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>System</th>
<th>CUDA Compiler</th>
<th>CUDA Tools</th>
</tr>
</thead>
<tbody>
<tr>
<td>1U PCI-E Switch</td>
<td>C Fortran Multi-core</td>
<td>Debugger Profiler</td>
</tr>
</tbody>
</table>

**Diagram:**
- 4 cores
- 240 cores
Compiling CUDA

C CUDA Application

NVCC

PTX Code

CPU Code

Virtual

PTX to Target Compiler

G80

...  

GTX

Physical

Target code
CUDA Source Code
Industry Standard C Language

Industry Standard Libraries

CUDA Compiler
C Fortran

Standard
Debugger Profiler

Multi-core
CUDA 2.0: Many-core + Multi-core support

C CUDA Application

NVCC

Many-core

PTX to Target Compiler

Many-core PTX code

NVCC --multicore

Multi-core

CPU C code

gcc and MSVC

Multi-core
CUDA Everywhere!
Questions?