The Future of Massively Parallel and GPU Computing

Poznaj świadectwa z dokumentacji procesu beatyfikacyjnego i kanonizacyjnego Jana Pawła II

CUDA

WYDAWNICTWO ŚW. STANISŁAWA BM

już od 7 marca

CUD

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

Simple “C” Description For Parallelism

```
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);
```

Standard C Code

```
__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);
```

Parallel C Code

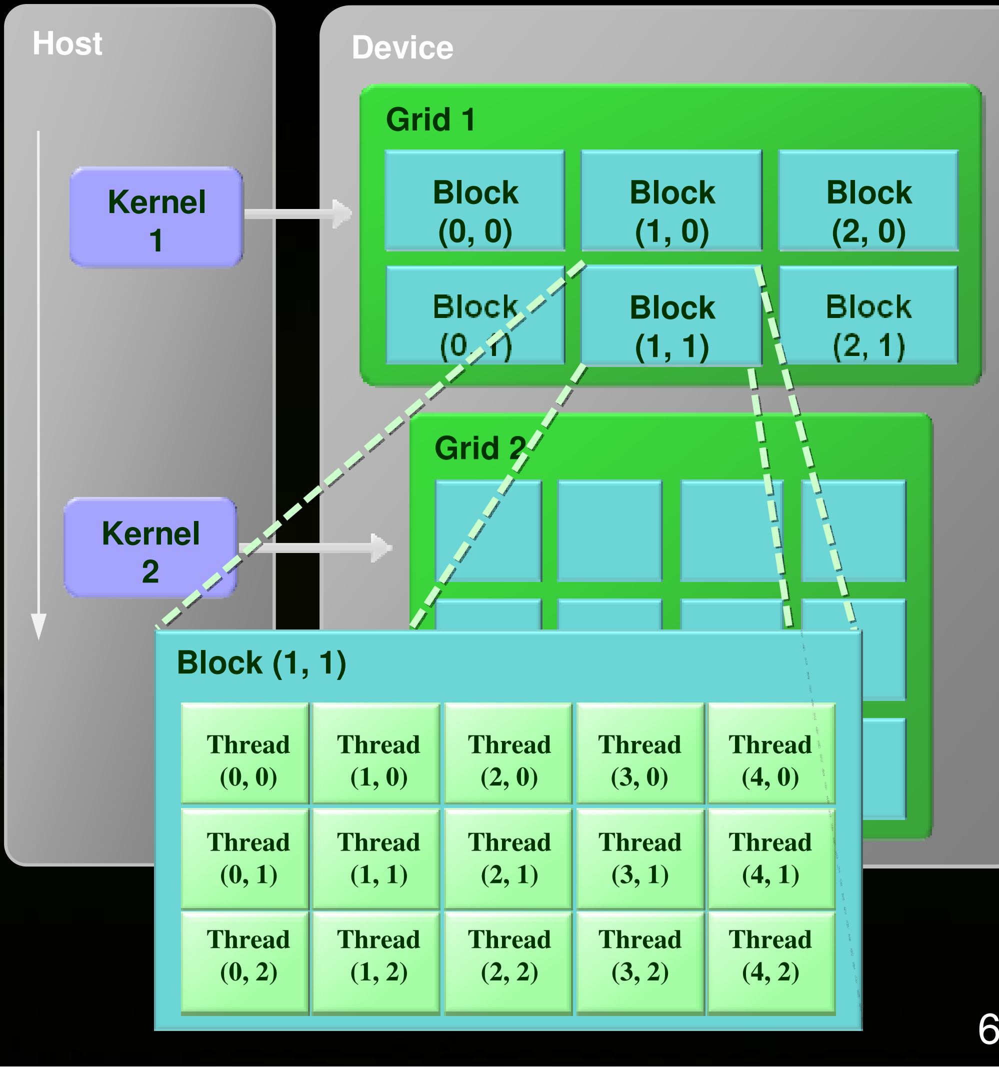
The Key to Computing on the GPU

- 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

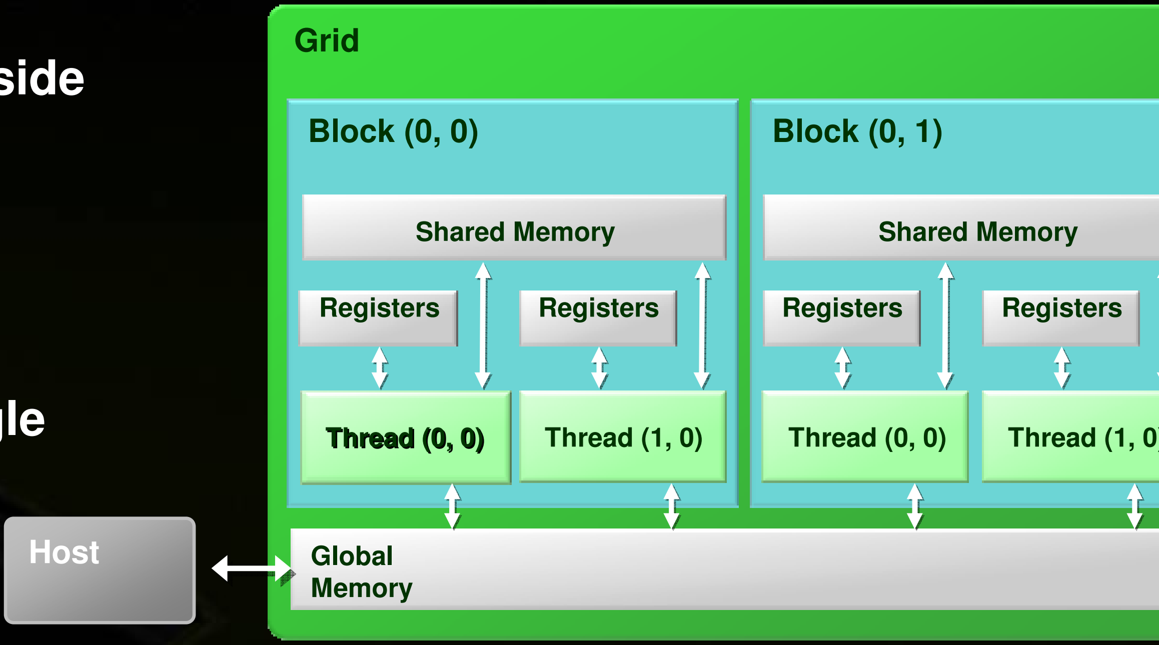
A kernel is executed by a grid of thread blocks

- 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



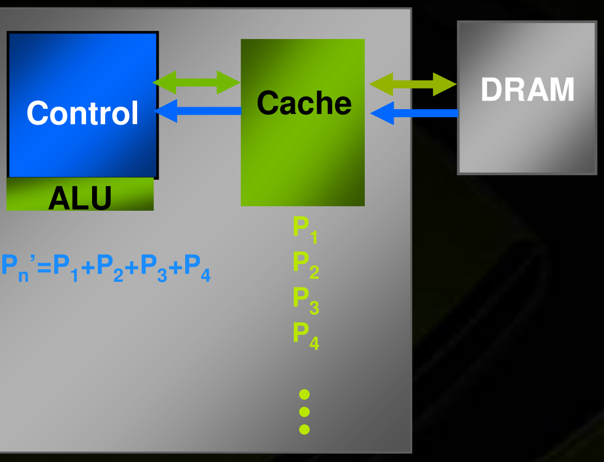
Kernel Memory Access

- 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



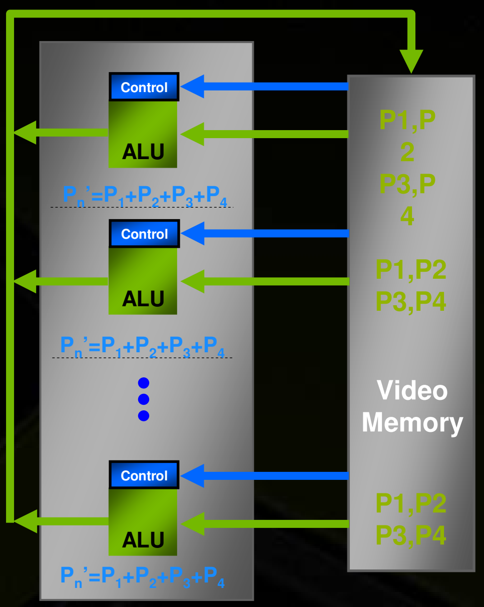
Example Fluid Algorithm

CPU



Single thread out of cache

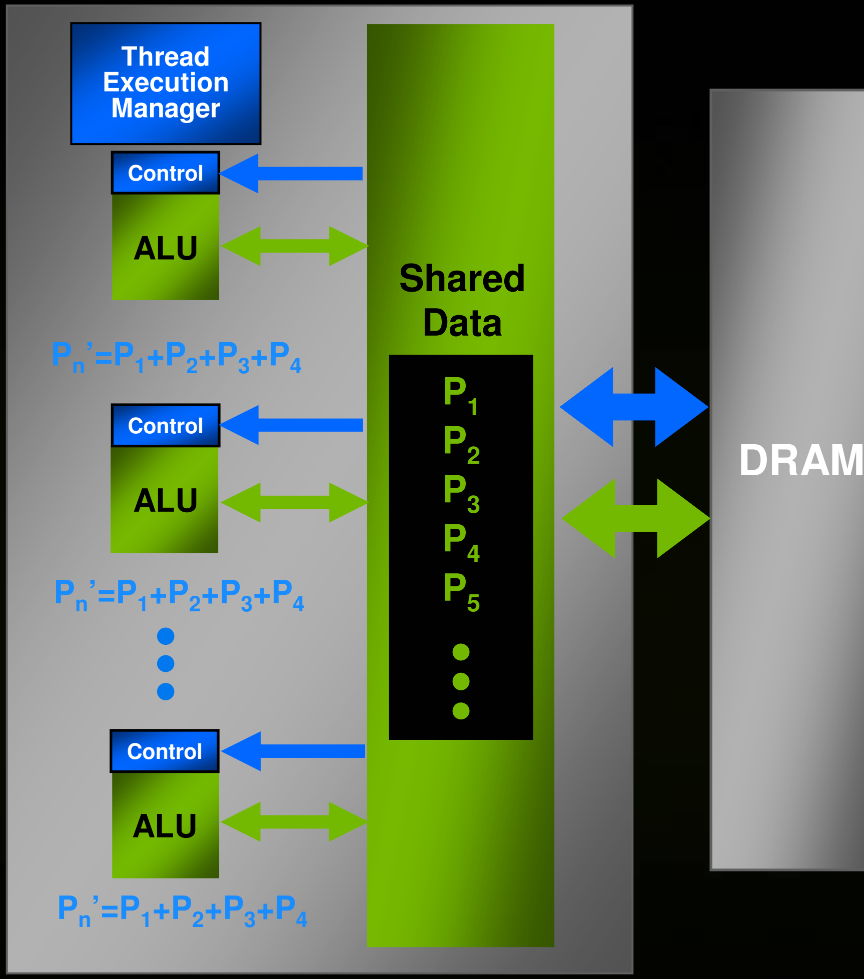
GPGPU


Control
ALU
$P_n' = P_1 + P_2 + P_3 + P_4$
Control
ALU
$P_n' = P_1 + P_2 + P_3 + P_4$
Control
ALU
$P_n' = P_1 + P_2 + P_3 + P_4$
P1,P2 P3,P4
P1,P2 P3,P4
Video Memory
P1,P2 P3,P4

Multiple passes through video memory

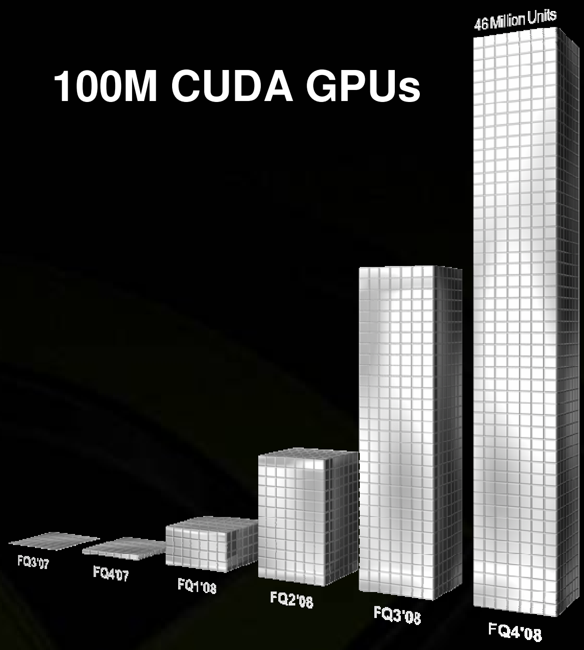
🟩 Data/Computation

🟦 Program/Control

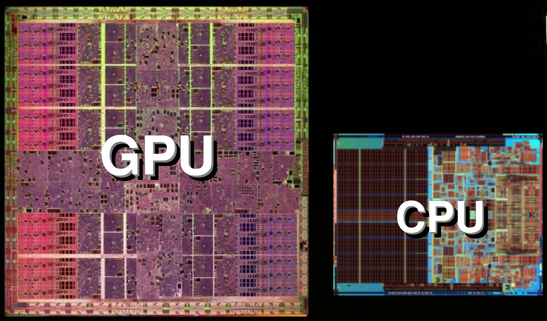


Parallel execution on-chip

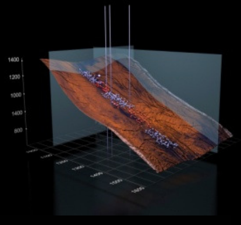
100M CUDA GPUs

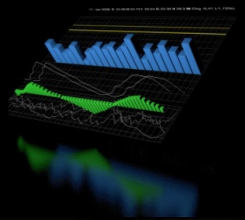


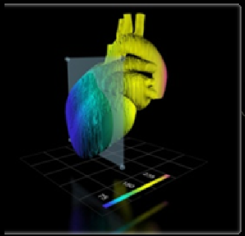
CUDA

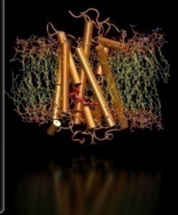


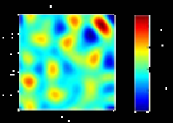
Heterogeneous Computing


Oil & Gas

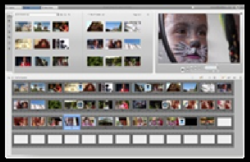

Finance

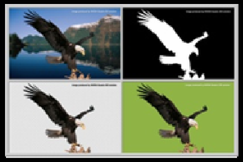

Medical


Biophysics


Numerics

Audio


Video

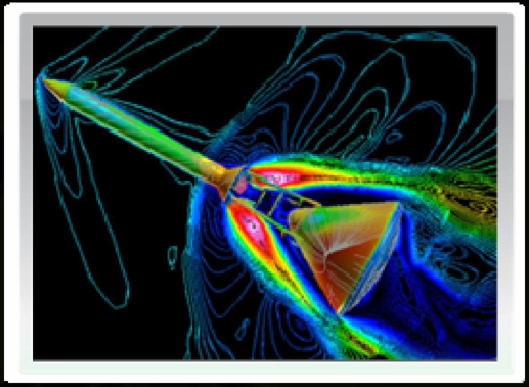

Imaging

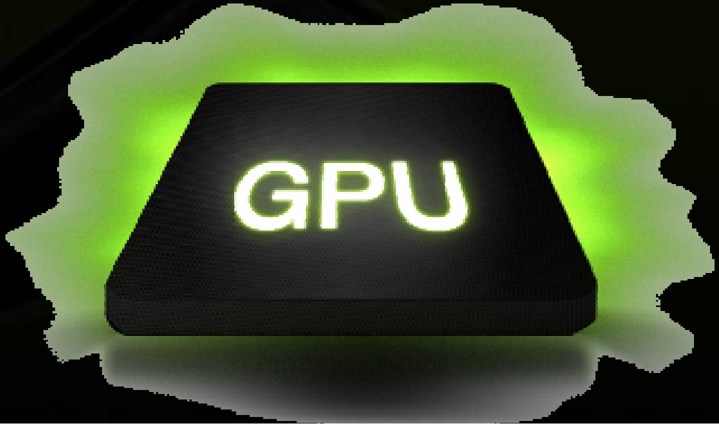
Parallel Computing on All GPUs

Almost 100 Million CUDA GPUs Deployed

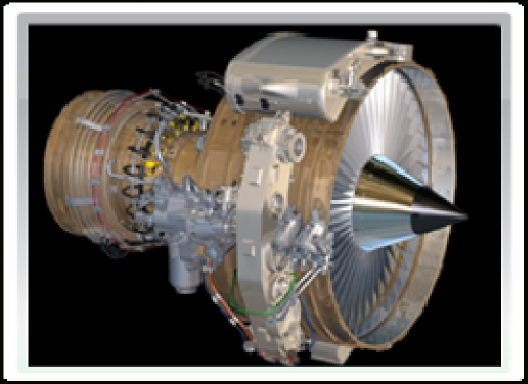
GeForce®
Entertainment

Tesla™
High-Performance Computing





Quadro®
Design & Creation



Developer Categories

D apparel design and simulation
D dental x-ray system
D image analysis from confocal microscope
D image capture
D-laser scanning feature extraction
coustic and electromagnetic simulation
coustic ray tracing
daptive radiation therapy
irline trining
LICE geometry processing
nalysis of electroencephalograms
nimation
stronomical adaptive optics
stronomical imaging
stronomics image scanning system
strophysics simulation
strophysics simulation
udio conferencing enhancement
udio processing
udio rendering of complex scenes
udio visual editing and scripting
utomated Web page classification
utomobile vision system
utomotive vision system
iochannel simulations
ioinfomatics
ioinfomatics for protein structure and cellular modelling
ioinfomatics for sequence alignment
iological circuits
iological imaging
iological simulation using evolutionary algorithms
iological simulations
iomedical cell imaging
iomedical image registration and segmentation
iomedical 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
CemSSIS space reconstruction
CFD
CFD for high speed aircraft engine design
CFD for ocean modelling
CFD with particle flow
Chess
Chromatography analysis
Climate models
Cloth simulation
Color correction for film
Color correction for projectors
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 fim 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 modelling
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
Fluorsescence Lifetime imaging
Folding at Home
Folding at home clone
Formal verification methods
Fortran, C/C++ compilers
Games
Gene sequence alignment
Genetics

Geomachanics using discrete or finite element anal
Geometric modelling
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 for security
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 extentsions
Large format imaging
Large scale neural networks
Large text database search
Linear programming
LISTSERV email list management
Machine automatioin
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
Multiphasic flow simulator
Multispectral 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
Nightime 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 engineeering
Optical processing
Optical security scanner
Optical simulation
Optronic 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

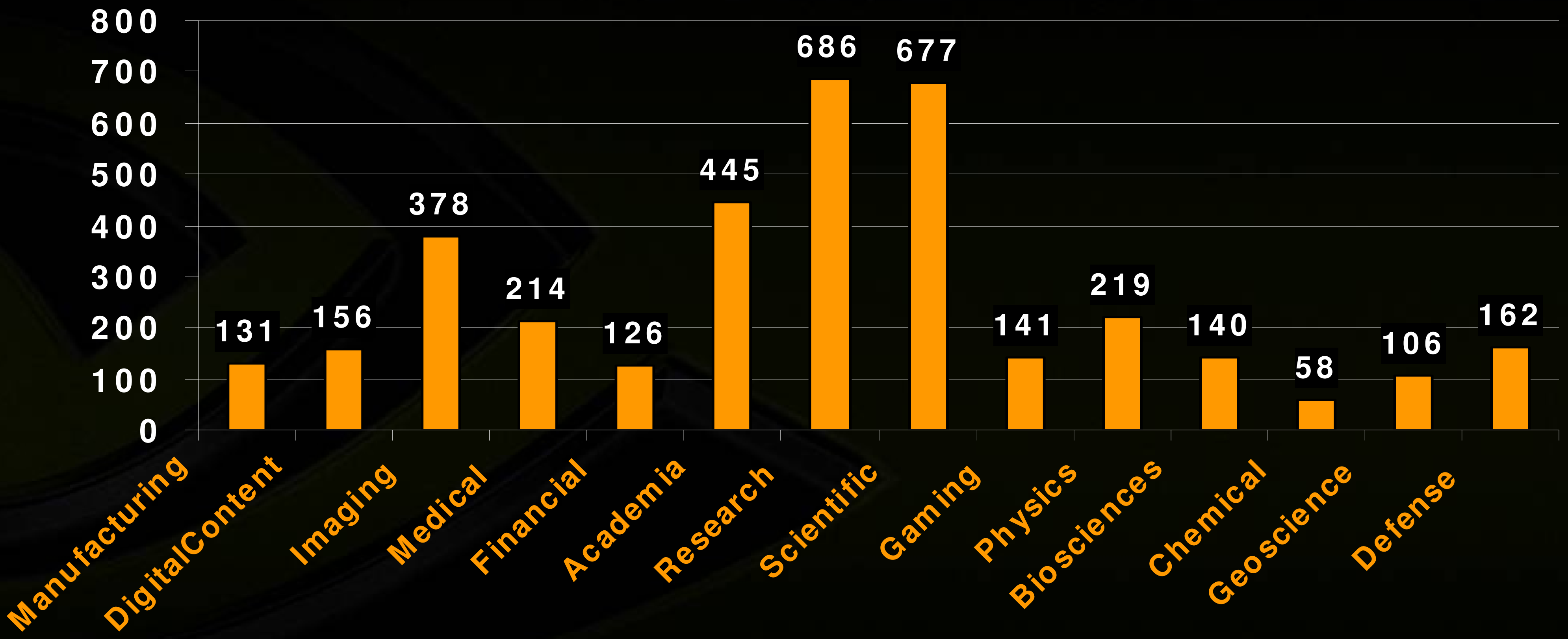
Programmable automation controllers
Protein crystallography
Protein folding
Protein structure prediction and design
Proteomics data diagnostics
Pulsar data analysis
Quantum chemistry
Quantum Chromodynamical calculations
Quantum molecular dynamics
Radar processing
Radar simulation
Radiation therapy machine
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
Speek 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 interation analysis software
Transaction query for mobile commerce
Ultrasonic inspection and testing
Ultrasound imaging
Ultrasound medical imaging
Unlimited precision mathematics
Urban 3D models from video streams
Video and audio finishinig
Video compression
Video compression with cupolet 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

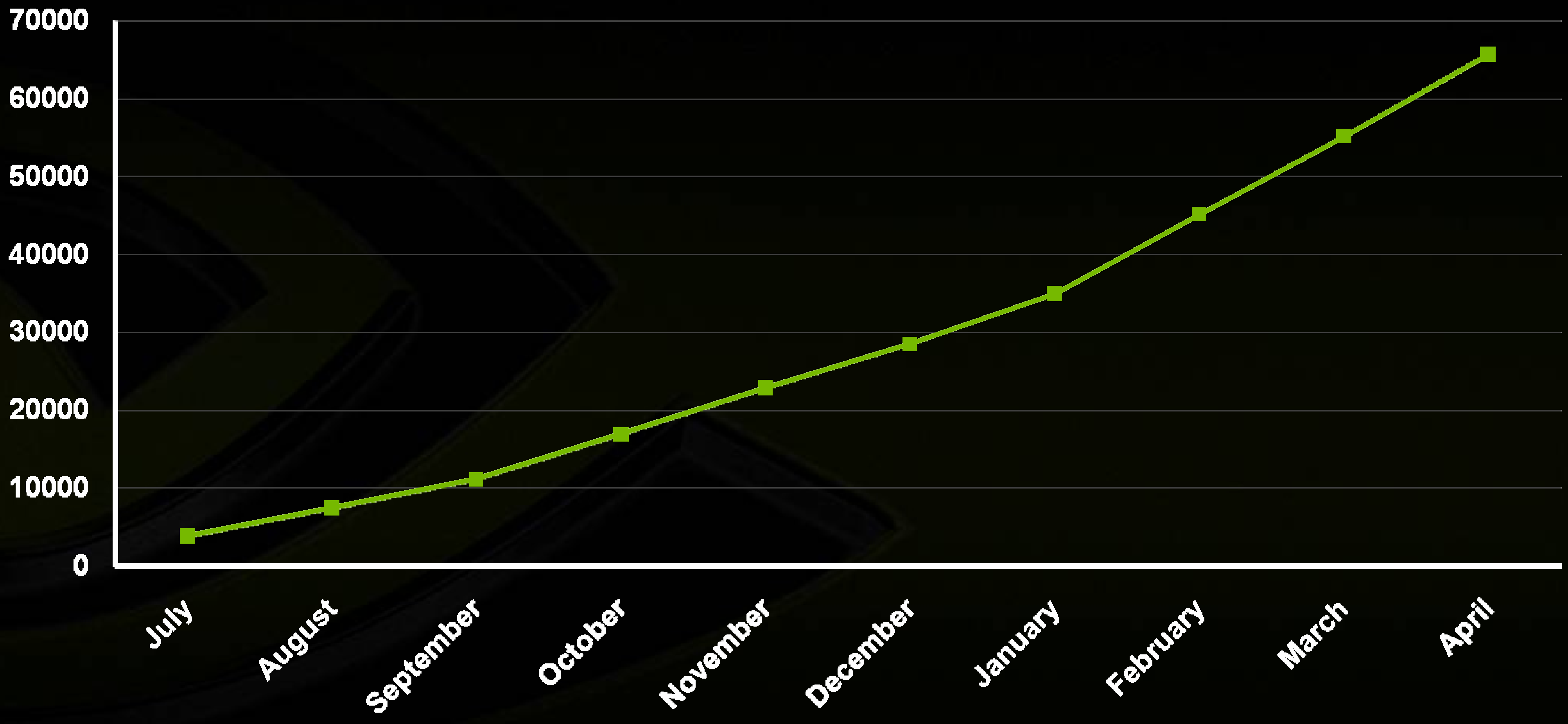


Registered developers who downloaded both CUDA 0.8 and 1.0

Parallel Computing Applications

Consumer	Business	Workstation	Technical
Imaging ideo – transcoding ames – Physics, AI Computer vision	Search Web	Oil and gas viz CAD	Seismic Finance Numerics*
Grid computing Audio Photography Virus scanning	XML parsing Database VPN/networking Backup compression RAID	Volume visualiaztion Cluster visualiztion	Medical imagi EDA CAE GIS

CUDA Compiler Downloads



niversities Teaching Parallel Programming With CUDA

Duke
Erlangen
ETH Zurich
Georgia Tech
Grove City College
Harvard
IIIT
IIT
Illinois Urbana-Champaign
INRIA
Iowa
ITESM
Johns Hopkins

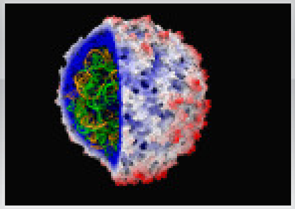
- Kent State
- Kyoto
- Lund
- Maryland
- McGill
- MIT
- North Carolina - Chapel Hill
- North Carolina State
- Northeastern
- Oregon State
- Pennsylvania
- Polimi
- Purdue

- Santa Clara
- Stanford Stuttgar
- Suny
- Tokyo
- TU-Vienna
- USC
- Utah
- Virginia
- Washington
- Waterloo
- Western Australi
- Williams College
- Wisconsin

Wide Developer Acceptance

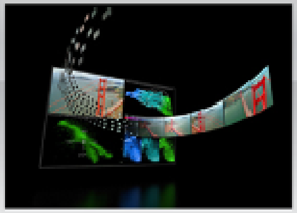
146X

Interactive visualization of volumetric white matter connectivity



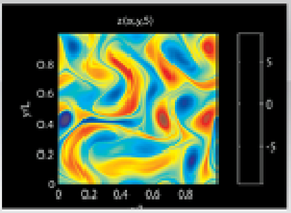
36X

Ionic placement for molecular dynamics simulation on GPU



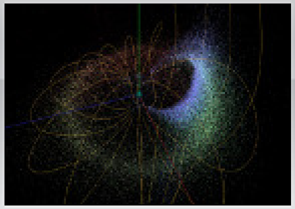
19X

Transcoding HD video stream to H.264



17X

Simulation in Matlab using .mex file CUDA function

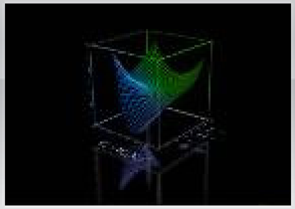


100X

Astrophysics N-body simulation

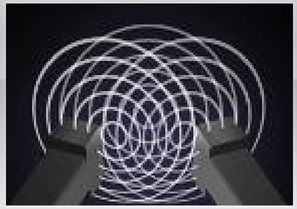
149X

Financial simulation of LIBOR model with swaptions



47X

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

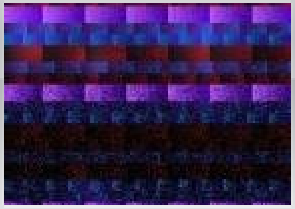


20X

Ultrasound medical imaging for cancer diagnostics

24X

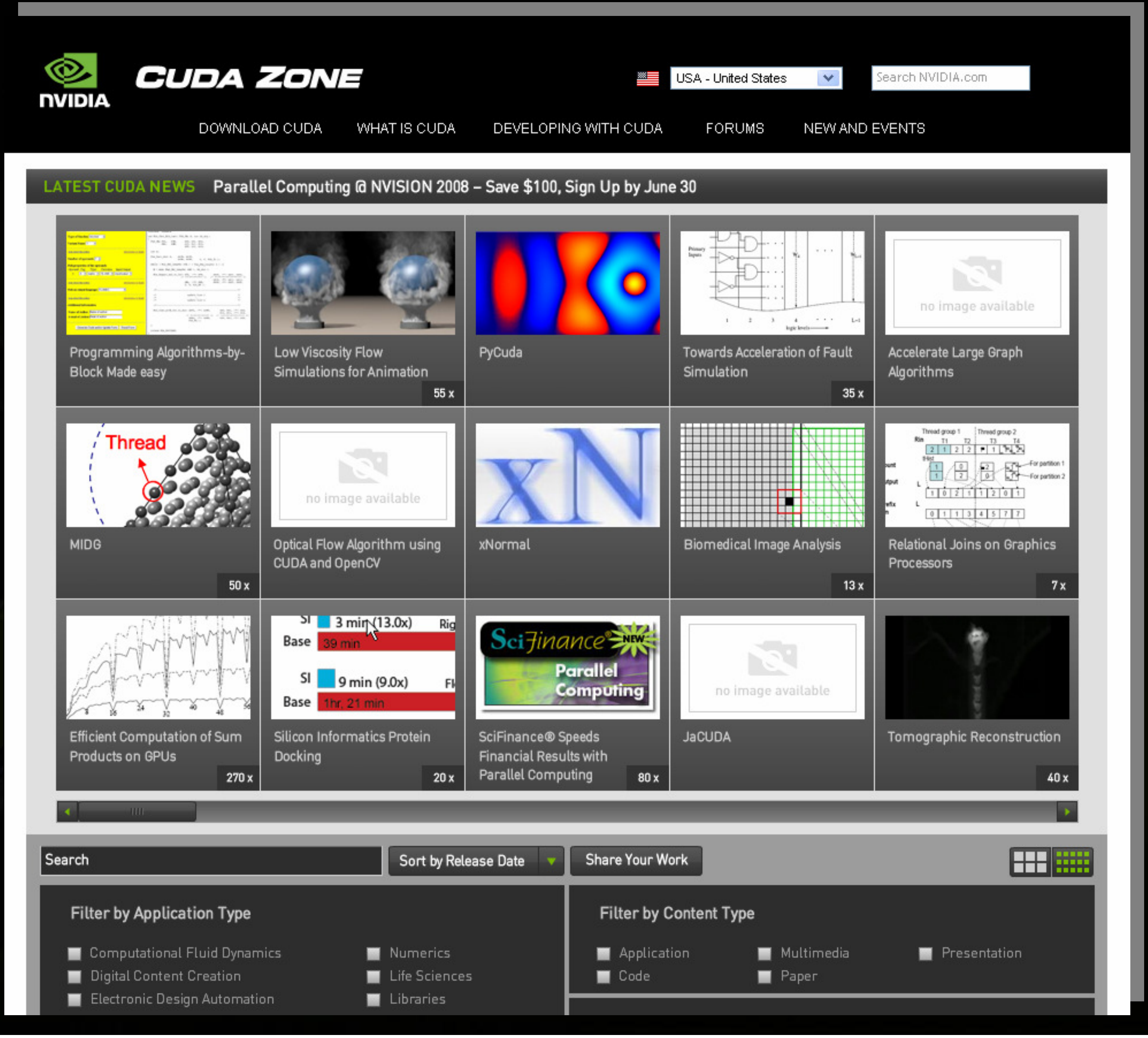
Highly optimized object oriented molecular dynamics



30X

Cmatch exact string matching to find similar proteins and gene sequences

CUDA Zone



CUDA ZONE

USA - United States

Search NVIDIA.com

DOWNLOAD CUDA | WHAT IS CUDA | DEVELOPING WITH CUDA | FORUMS | NEW AND EVENTS

LATEST CUDA NEWS Parallel Computing @ NVISION 2008 – Save $100, Sign Up by June 30

- Programming Algorithms-by-Block Made easy
- Low Viscosity Flow Simulations for Animation 55 x
- PyCuda
- Towards Acceleration of Fault Simulation 35 x
- Accelerate Large Graph Algorithms
- MIDG 50 x
- Optical Flow Algorithm using CUDA and OpenCV
- xNormal
- Biomedical Image Analysis 13 x
- Relational Joins on Graphics Processors 7 x
- Efficient Computation of Sum Products on GPUs 270 x
- Silicon Informatics Protein Docking 20 x
- SciFinance® Speeds Financial Results with Parallel Computing 80 x
- JaCUDA
- Tomographic Reconstruction 40 x

Search | Sort by Release Date | Share Your Work

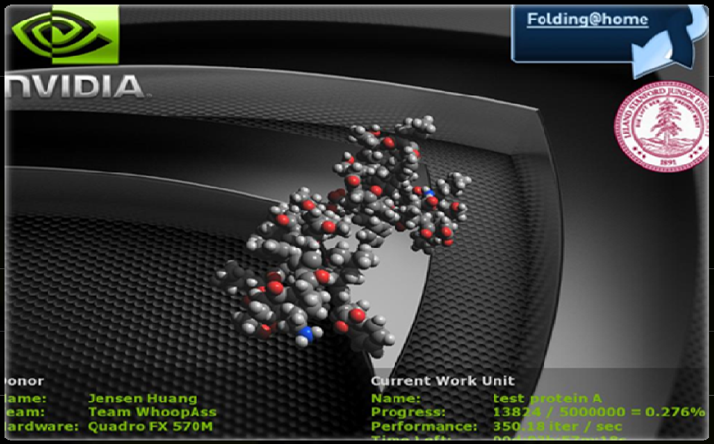
Filter by Application Type

- □ Computational Fluid Dynamics
- □ Digital Content Creation
- □ Electronic Design Automation
- □ Numerics
- □ Life Sciences
- □ Libraries

Filter by Content Type

- □ Application
- □ Code
- □ Multimedia
- □ Paper
- □ Presentation

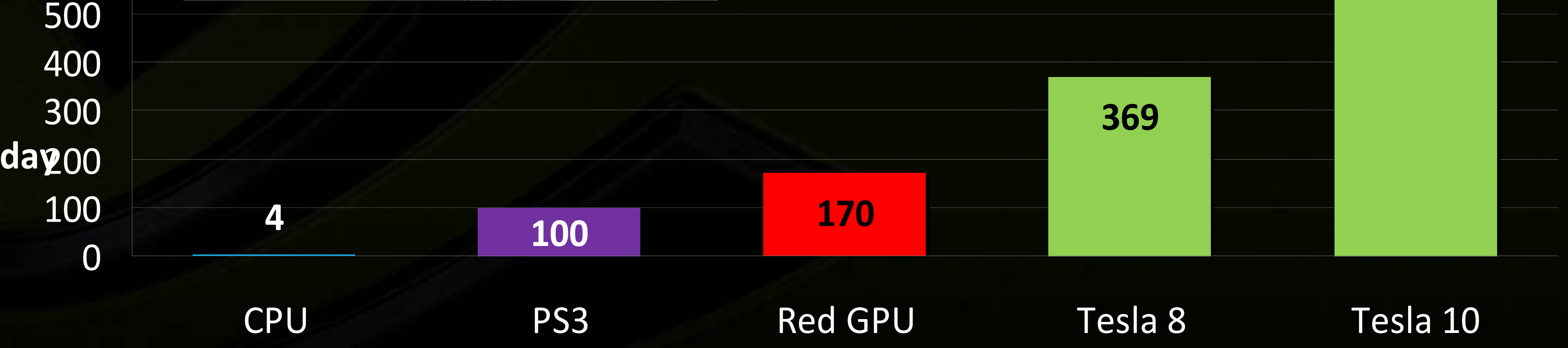
Folding@Home Using GROMACS



onor
ame: Jensen Huang
eam: Team WhoopAss
ardware: Quadro FX 570M

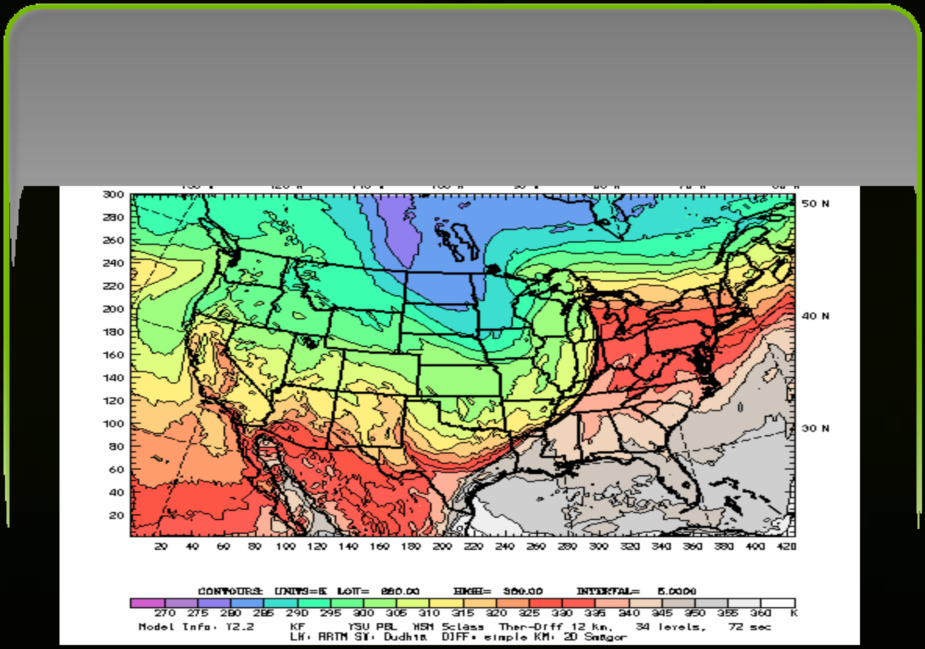
Current Work Unit
Name: test protein A
Progress: 13824 / 5000000 = 0.276%
Performance: 350.18 iter / sec

- Alzheimer’s Disease
- Huntington’s Disease
- Cancer
- Osteogensis imperfecta
- Parkinson’s Disease
- Antibiotics



Science: National Center for Atmospheric Research

- eather Research and Forecast (WRF) model
- 000+ registered users worldwide
- % speedup with 1% of WRF on CUDA
- aves 1 week analysis time


300
280
260
240
220
200
180
160
140
120
100
80
60
40
20
20 40 60 80 100 120 140 160 180 200 220 240 260 280 300 320 340 360 380 400 420
50 N
40 N
30 N
CONTOURS: UNITS=K LOW= 260.00 HIGH= 360.00 INTERVAL= 5.0000
270 275 280 285 290 295 300 305 310 315 320 325 330 335 340 345 350 355 360 K
Model Info: V2.2 KF YSU PBL WSM 5class Ther-Diff 12 km, 34 levels, 72 sec
LW: RRTM SW: Dudhia DIFF: simple KM: 2D Smagor

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

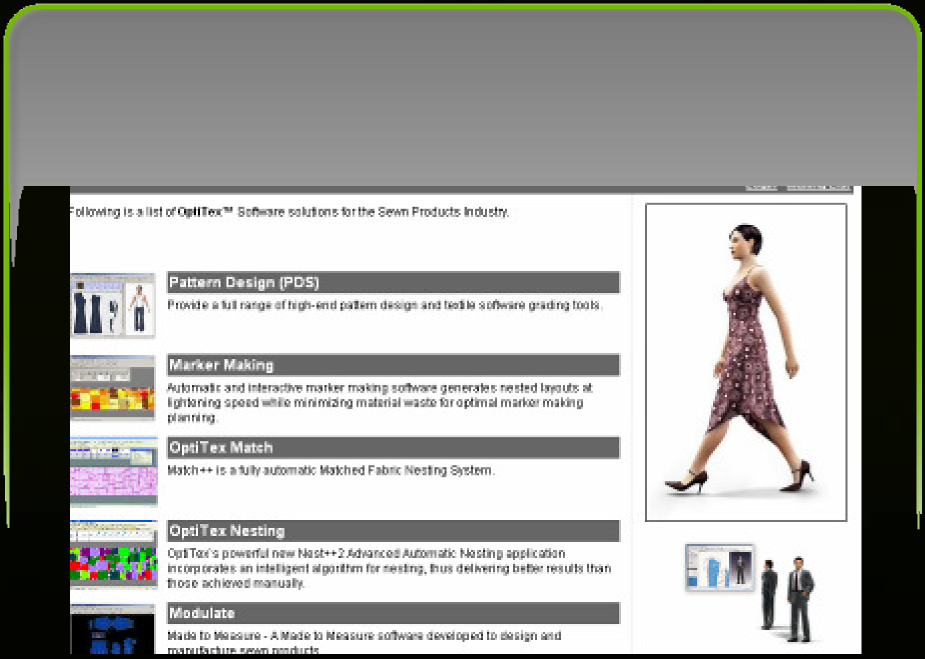
	GPUs	CPUs	Savings
Processors	12	600	
Rack Space	6U	54U	9x
Hardware Cost	$42,000	$262,000	6x
Annual Cost	$140,000	$1,200,000	9x

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

OptiTex™
Apparel | Fashion | Sewn
2D/3D CAD/CAM Professionals



Following is a list of **OptiTex™** Software solutions for the Sewn Products Industry.

Pattern Design (PDS)

Provide a full range of high-end pattern design and textile software grading tools.

Marker Making

Automatic and interactive marker making software generates nested layouts at lightening speed while minimizing material waste for optimal marker making planning.

OptiTex Match

Match++ is a fully automatic Matched Fabric Nesting System.

OptiTex Nesting

OptiTex's powerful new Nest++2 Advanced Automatic Nesting application incorporates an intelligent algorithm for nesting, thus delivering better results than those achieved manually.

Modulate

Made to Measure - A Made to Measure software developed to design and manufacture sewn products

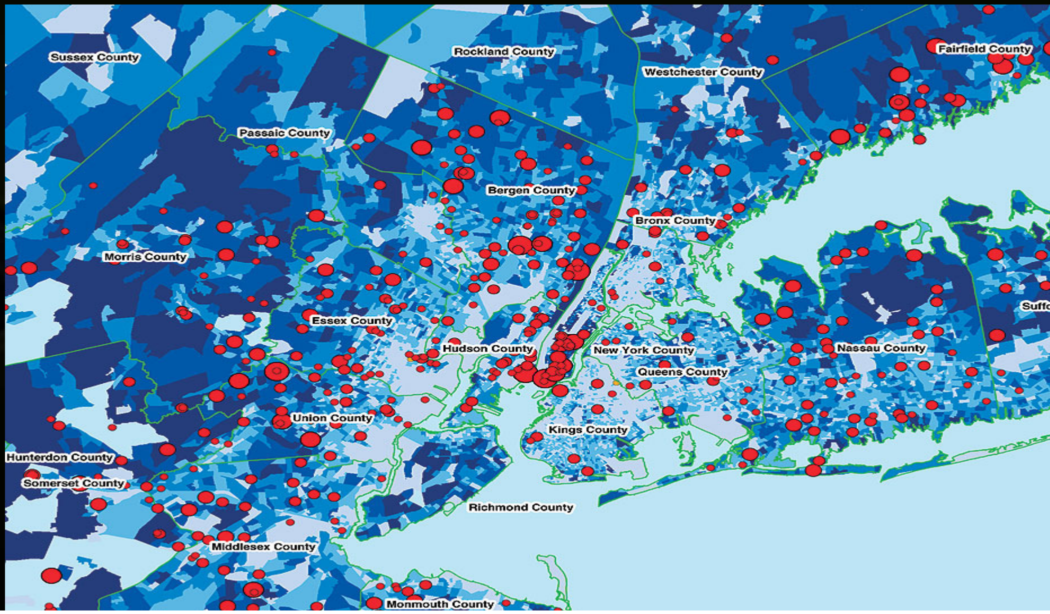
GIS Application

From the Manifold 8 feature list:

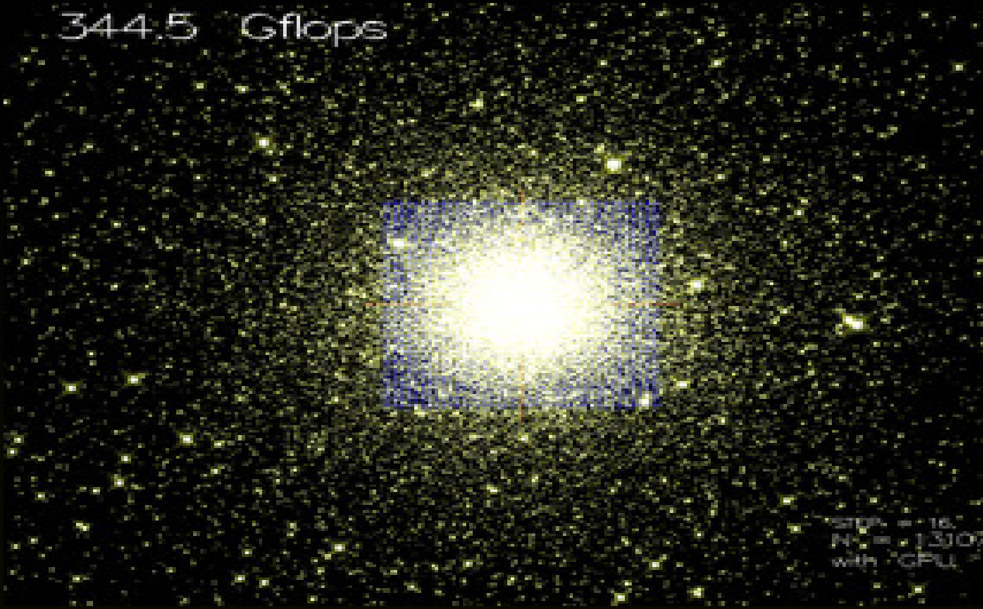
… applications fitting CUDA capabilities that might have taken tens of seconds or even minutes can be accomplished in hundredths of seconds. … CUDA will clearly emerge to be the future of almost all GIS computing

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



http://progrape.jp/

Astrophysics research

1 GF on standard PC

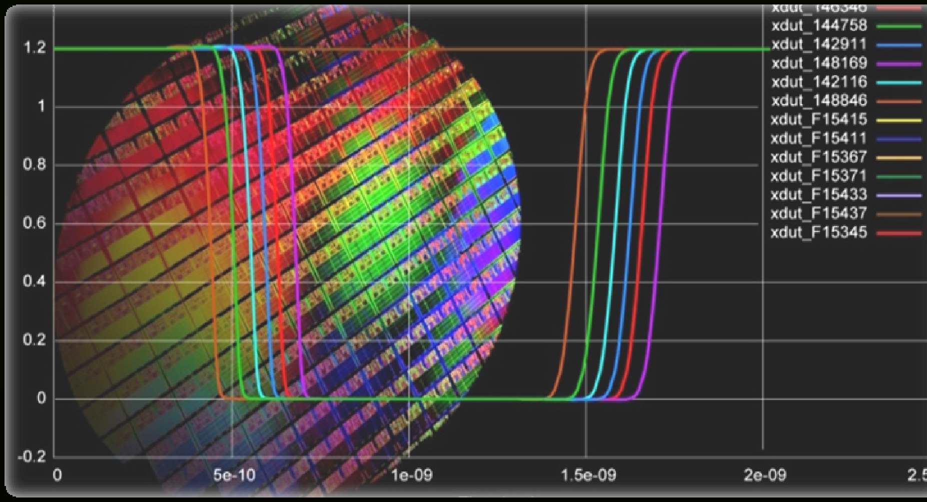
300+ GF on GeForce 8800GTX

Faster than GRAPE-6Af custom simulation computer

Video demo

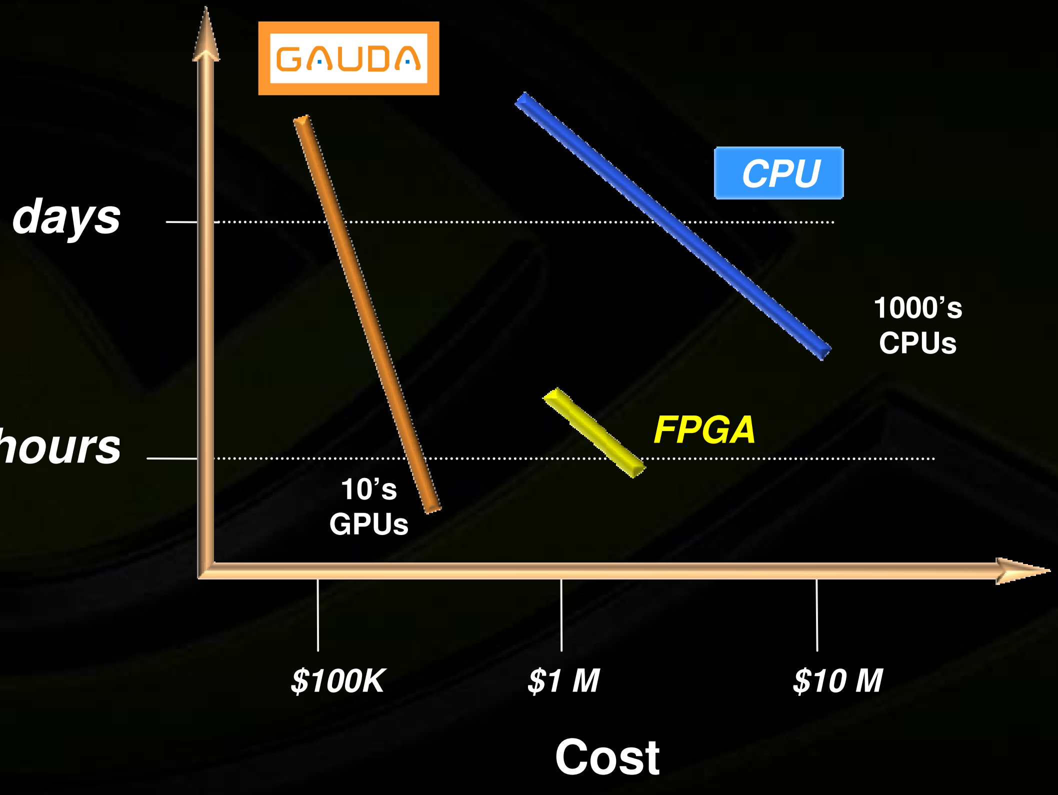
OmegaSim GX - Spice Simulation with CUDA

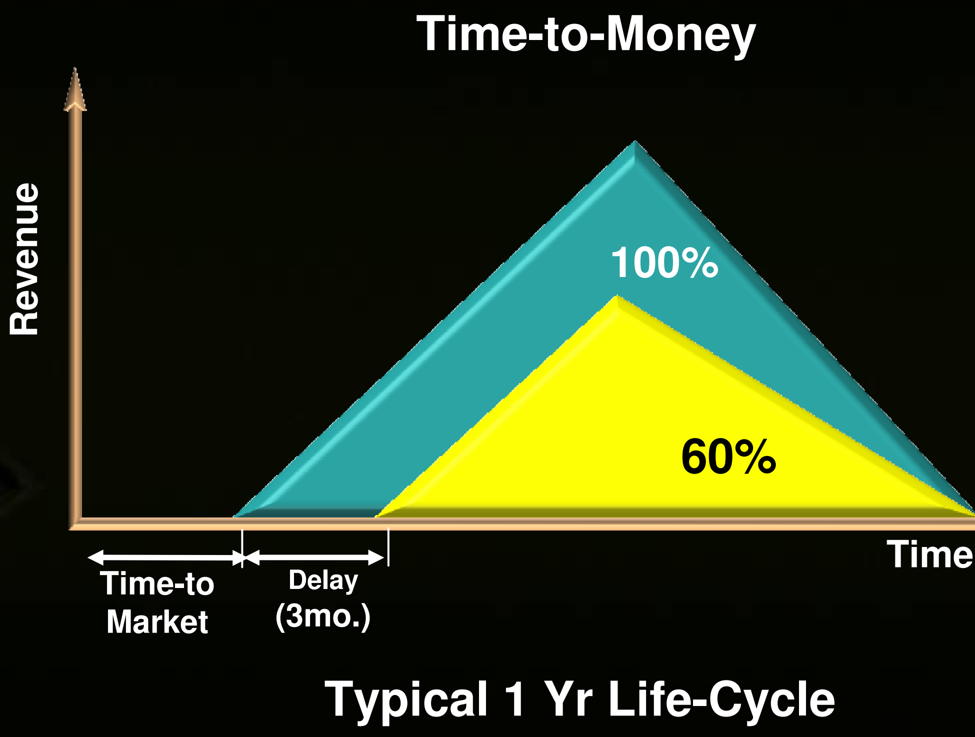
- 40x Speedup for transistor evaluation
- Up to 90% of SPICE execution time spent in transistor evaluation
- Avg. 8x overall speedup



Gauda Optical Proximity Correction (OPC)

200x Faster and Lower Cost





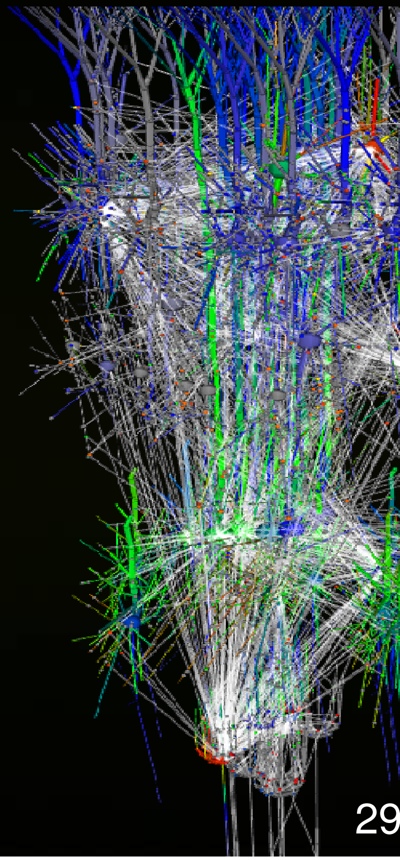
EvolvedMachines

Simulate the brain circuit

Sensory computing: vision, olfactory

130X Speed up

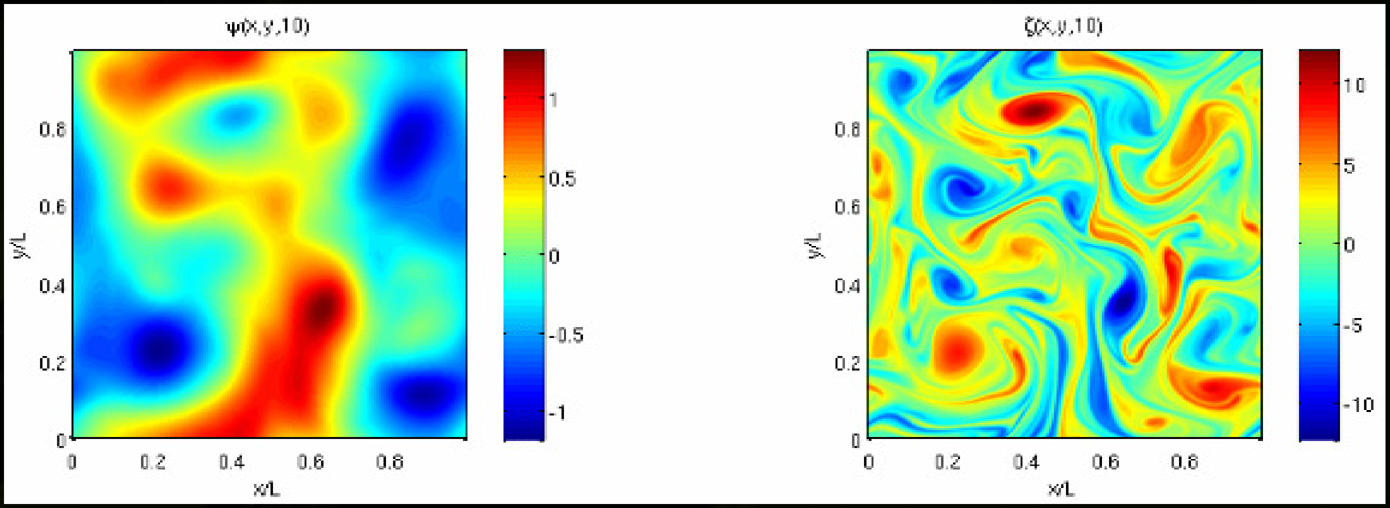
EvolvedMachines



Matlab: Language of Science

18X with MATLAB CPU+GPU

http://developer.nvidia.com/object/matlab_cuda.html



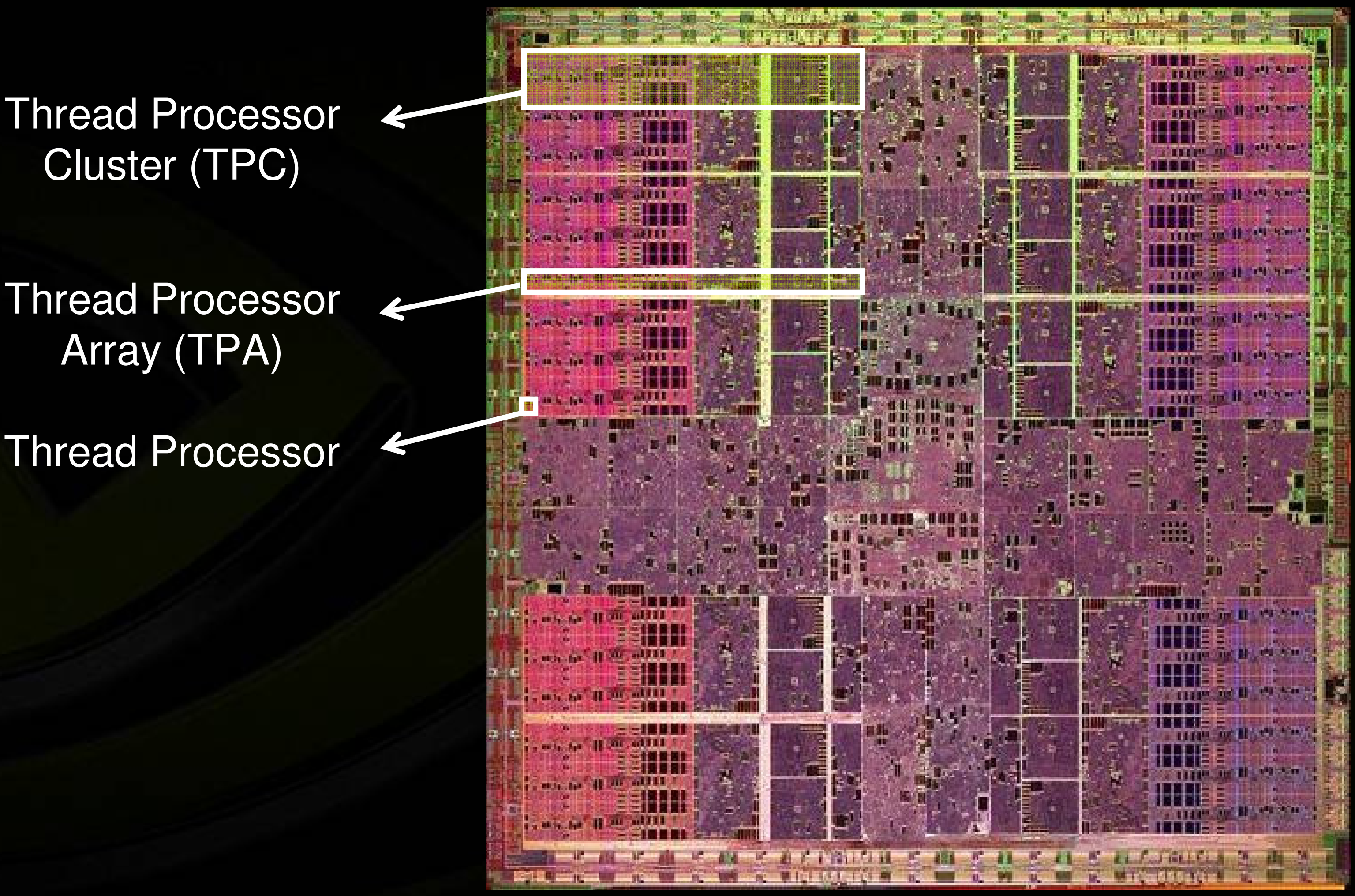
Pseudo-spectral simulation of 2D Isotropic turbulence

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)

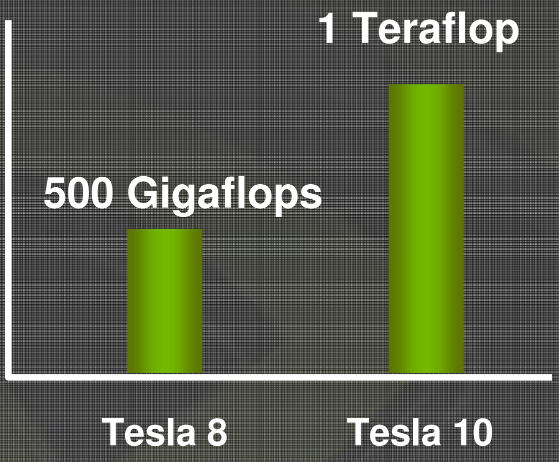
Thread Processor Array (TPA)

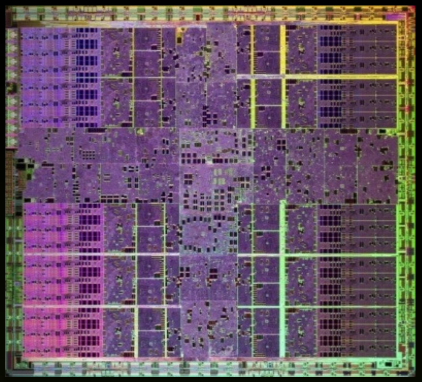
Thread Processor

Die Photo of Tesla T10

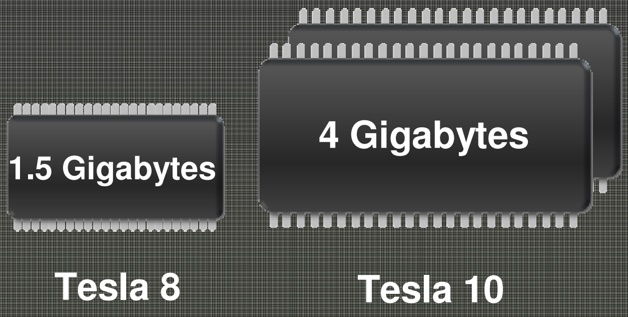
Tesla 10-Series

Double the Performance

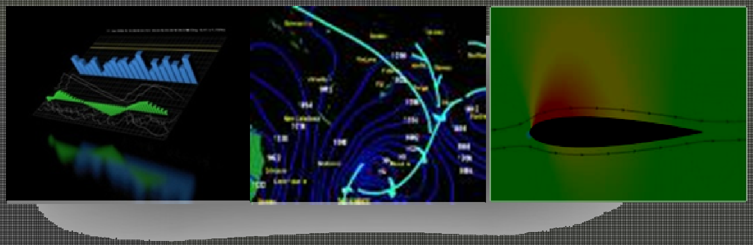




Double the Memory



Double the Precision

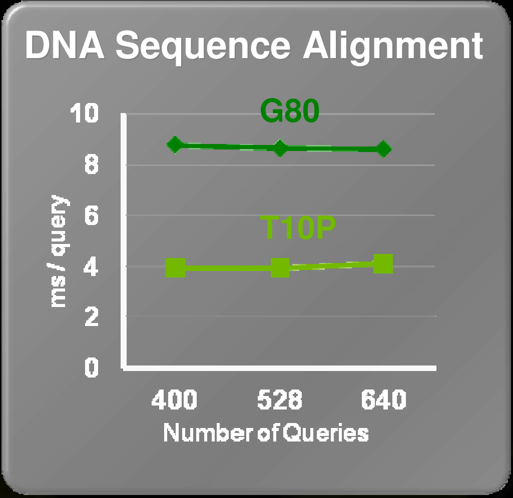


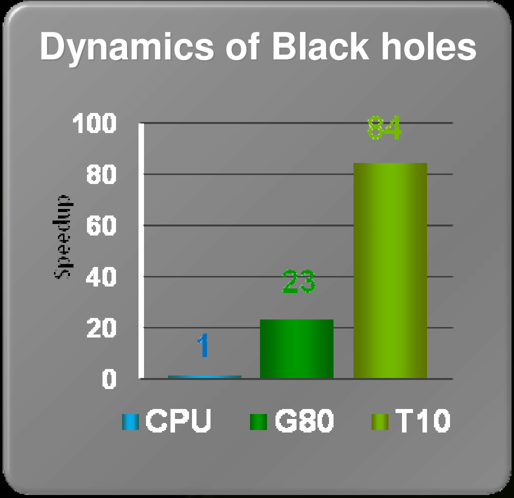
Finance Science Design

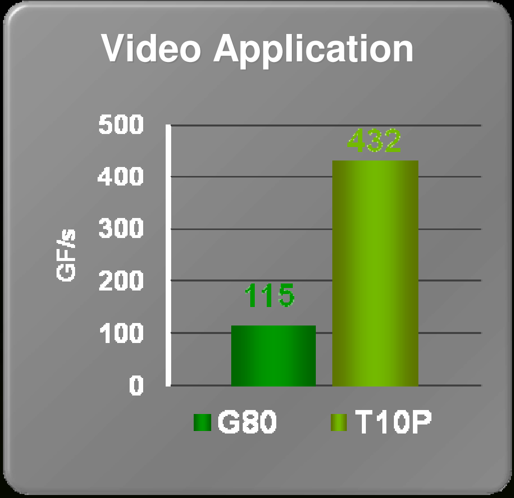
Tesla T10 Double Precision Floating Point

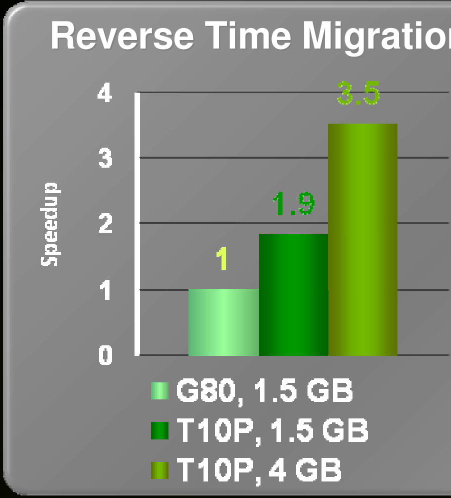
Precision	IEEE 754
Rounding modes for FADD and FMUL	All 4 IEEE, round to nearest, zero, inf, -inf
Denormal handling	Full speed
NaN support	Yes
Overflow and Infinity support	Yes
Flags	No
FMA	Yes
Square root	Software with low-latency FMA-based convergence
Division	Software with low-latency FMA-based convergence
Reciprocal estimate accuracy	24 bit
Reciprocal sqrt estimate accuracy	23 bit
log2(x) and 2^x estimates accuracy	23 bit

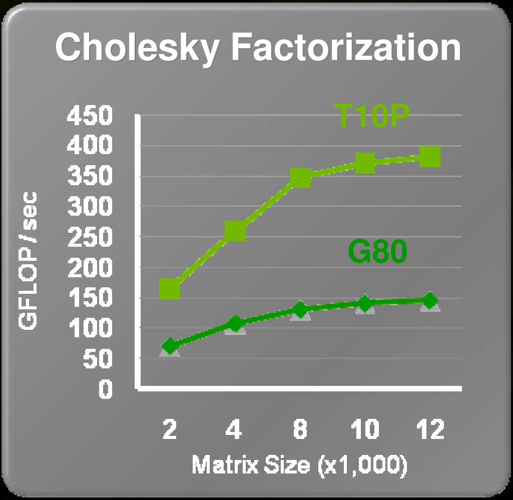
Double the Performance Using T10

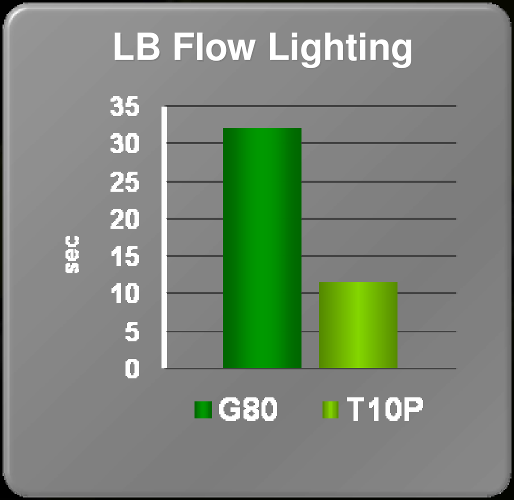

DNA Sequence Alignment

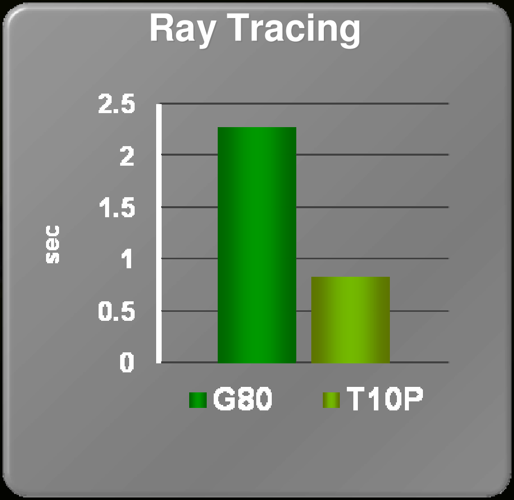

Dynamics of Black holes


Video Application


Reverse Time Migratio

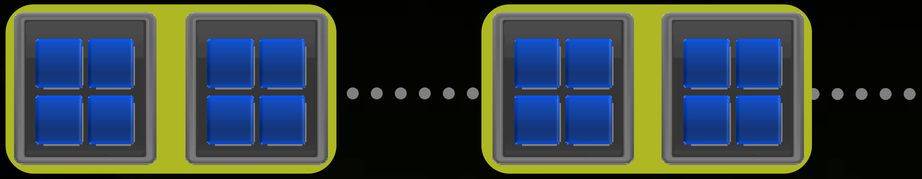

Cholesky Factorization


LB Flow Lighting


Ray Tracing

How to Get to 100X?

Traditional Data Center Cluster



Quad-core CPU

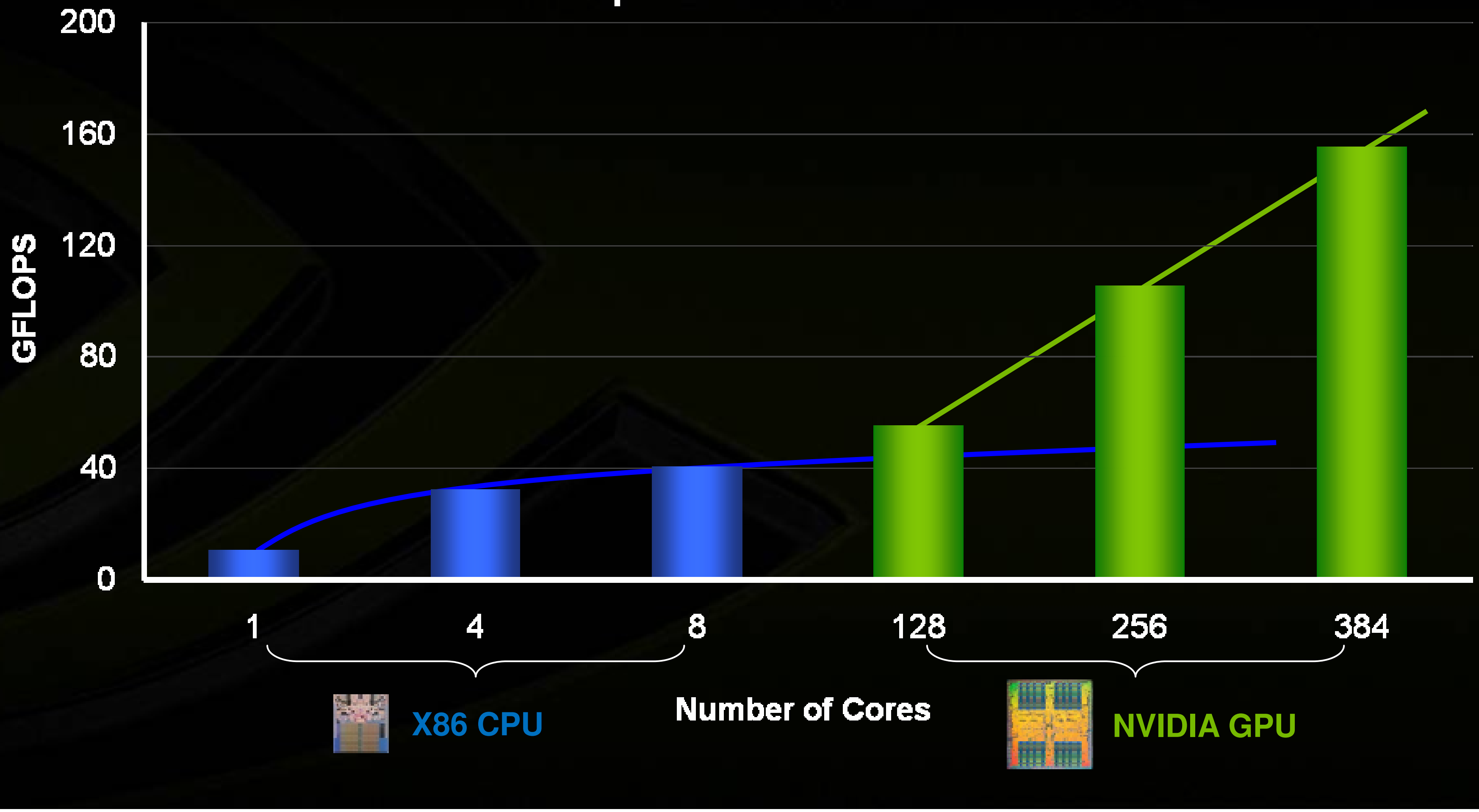
1000’s of cores
1000’s of servers

8 cores per server

More Servers To Get More Performance

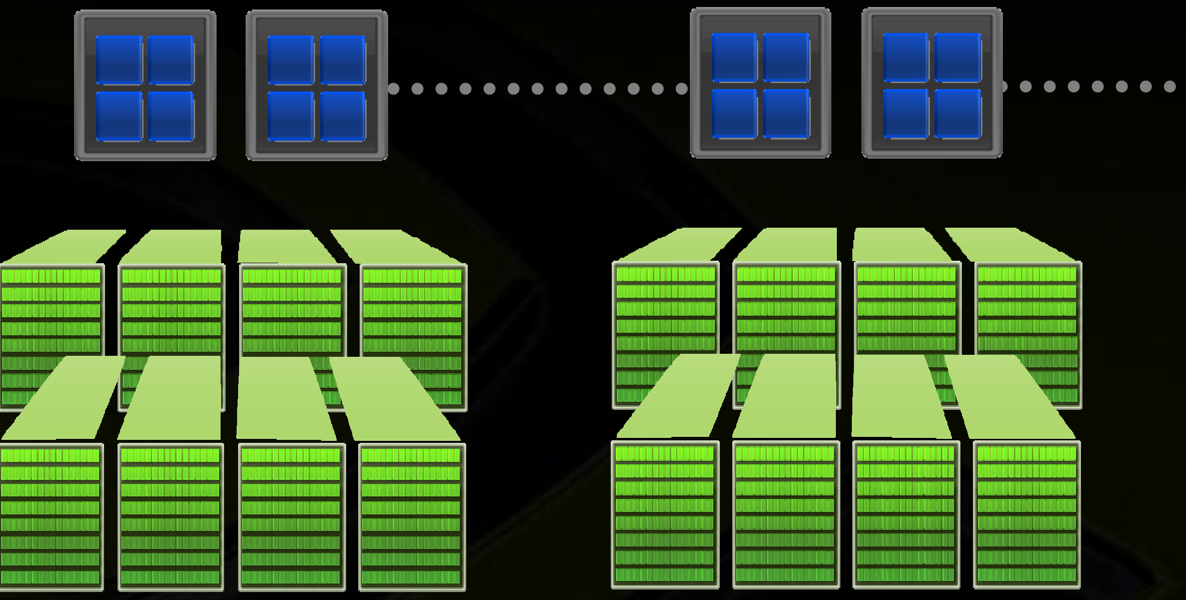
Linear Scaling with Multiple GPUs

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



Heterogeneous Computing Cluster

10,000's processors per cluster



1928 processors

1928 processors

- 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

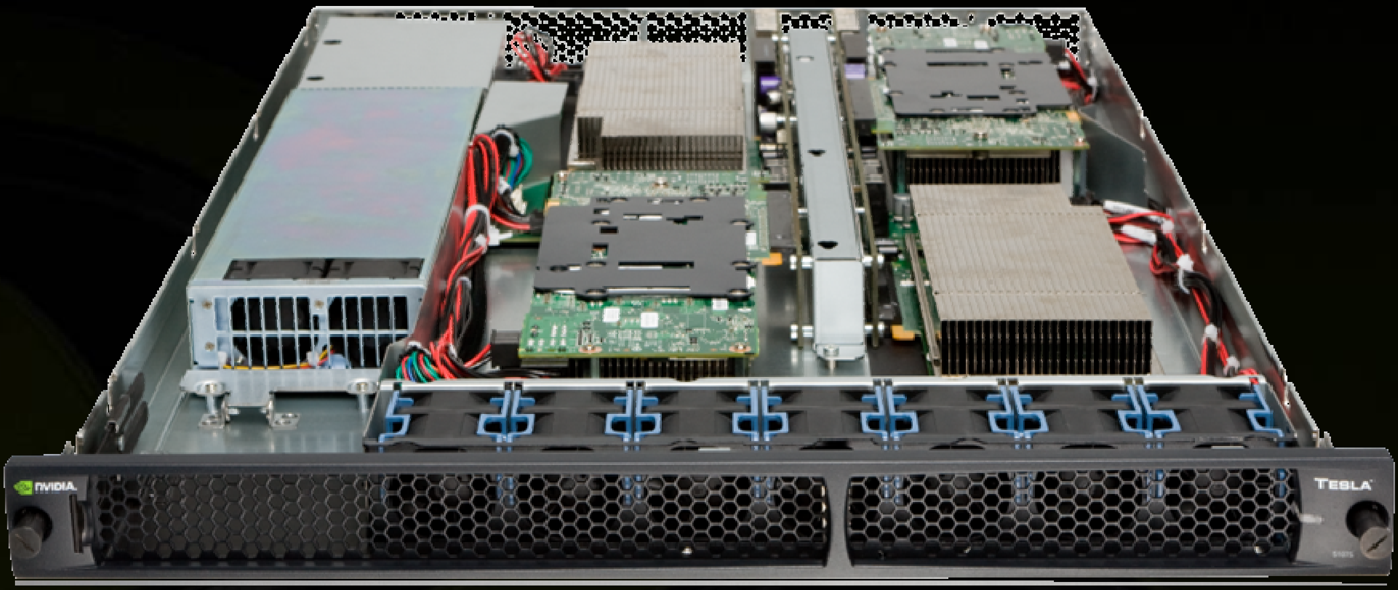
Building a 100TF datacenter

CPU 1U Server	Tesla 1U System
4 CPU cores	4 GPUs: 960 cores
0.07 Teraflop	4 Teraflops
$ 2000	$ 8000
400 W	700 W
1429 CPU servers	25 CPU servers 25 Tesla systems
$ 3.1 M	$ 0.31 M
571 KW	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

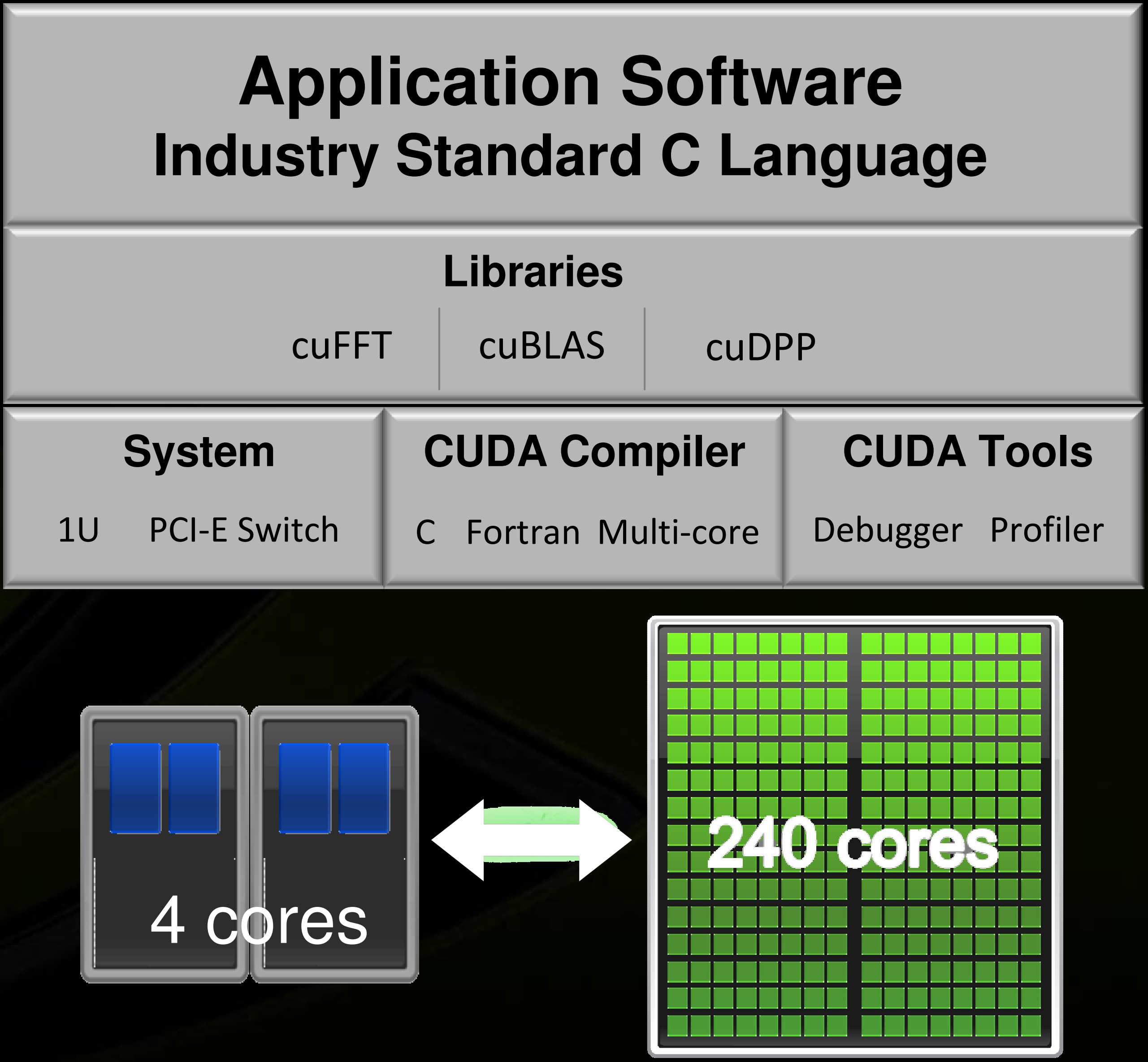
C++

Multiple GPUs

Debugger

Profiler

GPU Cluster



Application Software
Industry Standard C Language

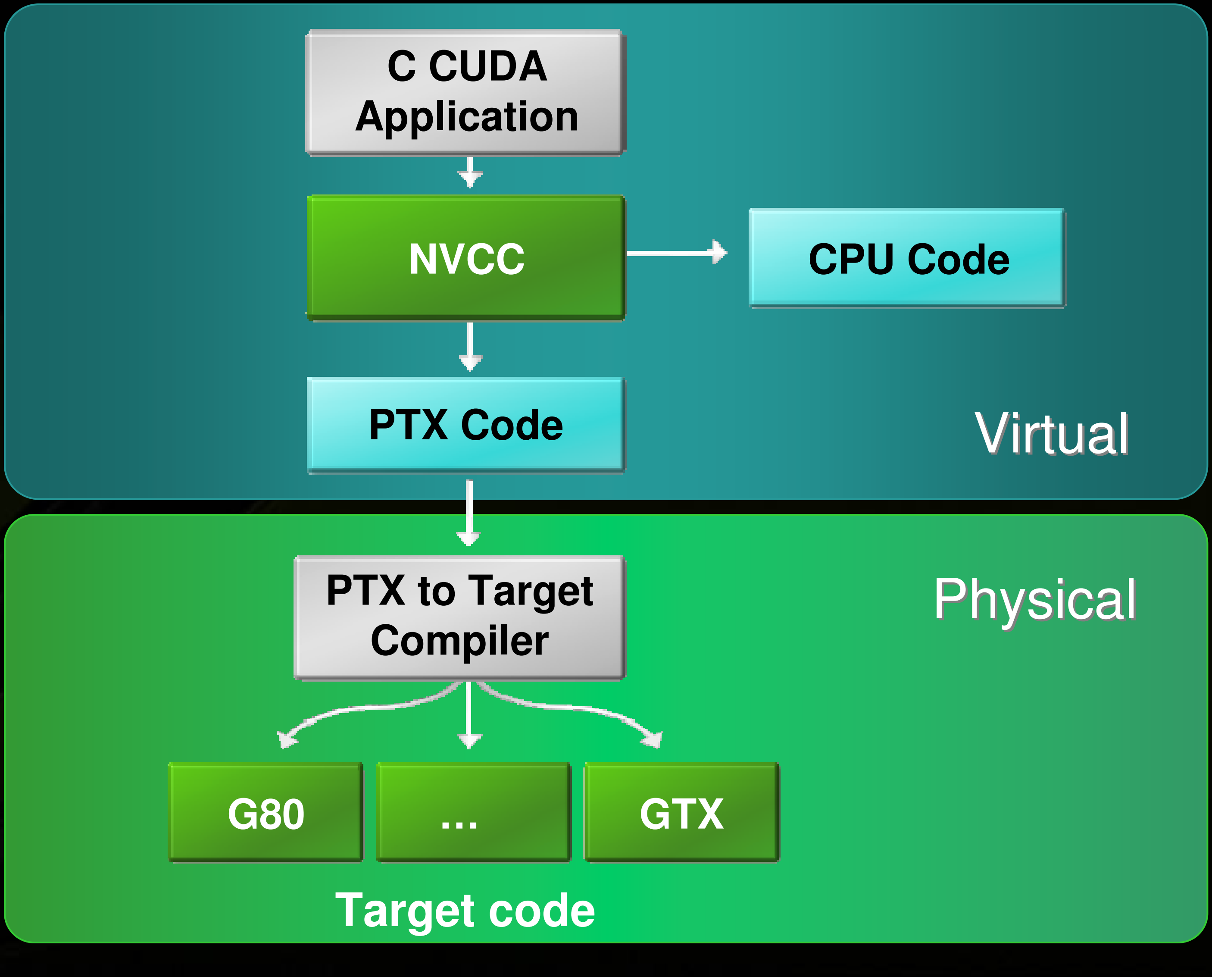
Libraries
cuFFT | cuBLAS | cuDPP

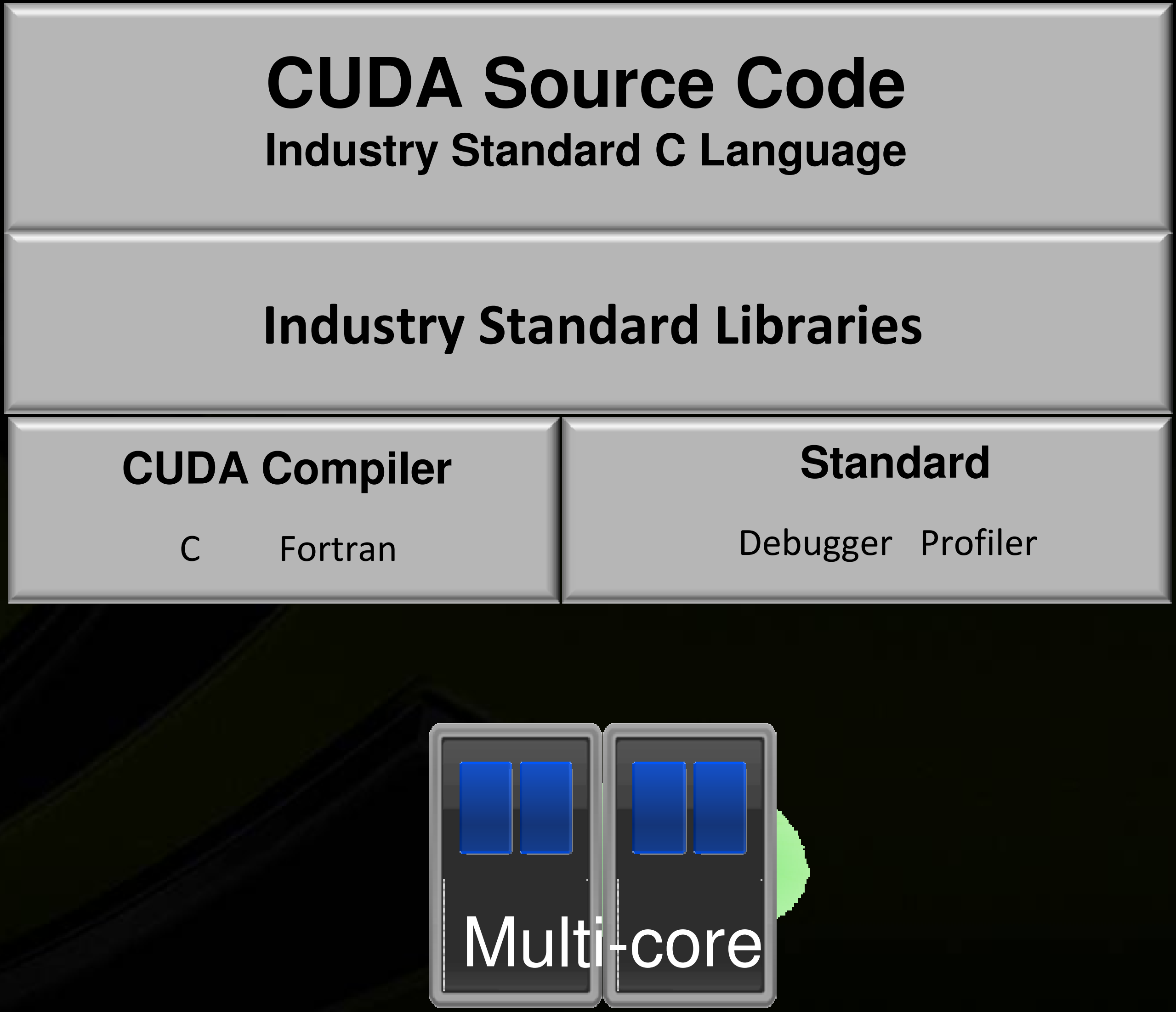
System
1U PCI-E Switch

CUDA Compiler
C Fortran Multi-core

CUDA Tools
Debugger Profiler

Compiling CUDA





CUDA Source Code

Industry Standard C Language

Industry Standard Libraries

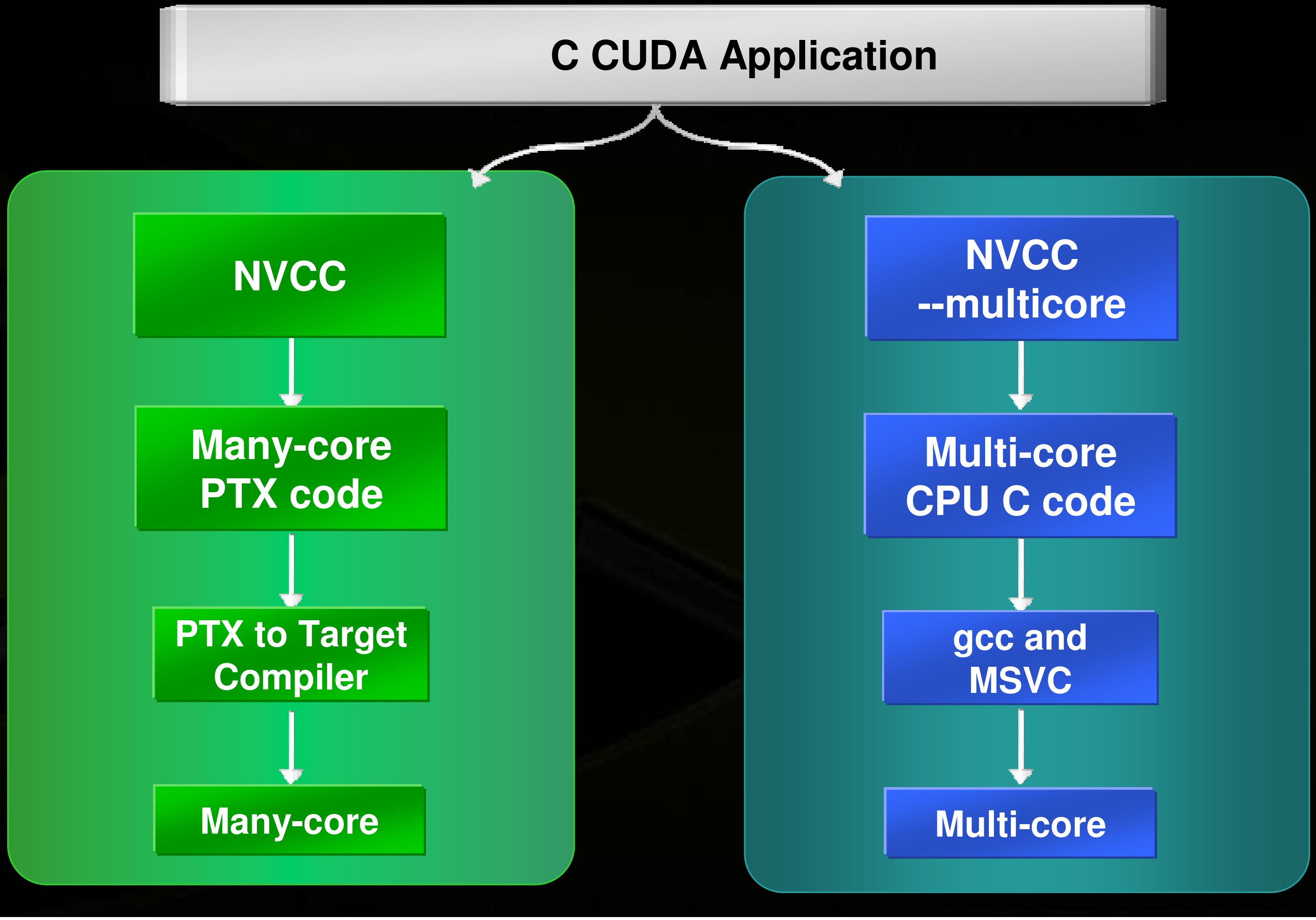
CUDA Compiler

C Fortran

Standard

Debugger Profiler

CUDA 2.0: Many-core + Multi-core support



C CUDA Application

NVCC

Many-core PTX code

PTX to Target Compiler

Many-core

NVCC --multicore

Multi-core CPU C code

gcc and MSVC

Multi-core

CUDA Everywhere!

Questions?