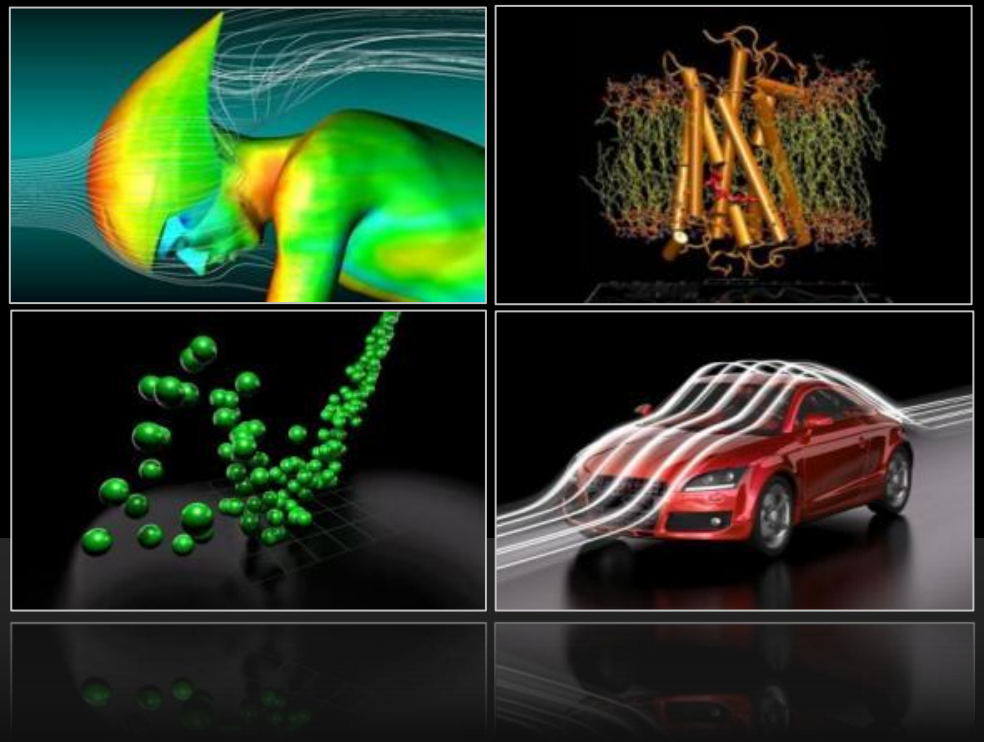
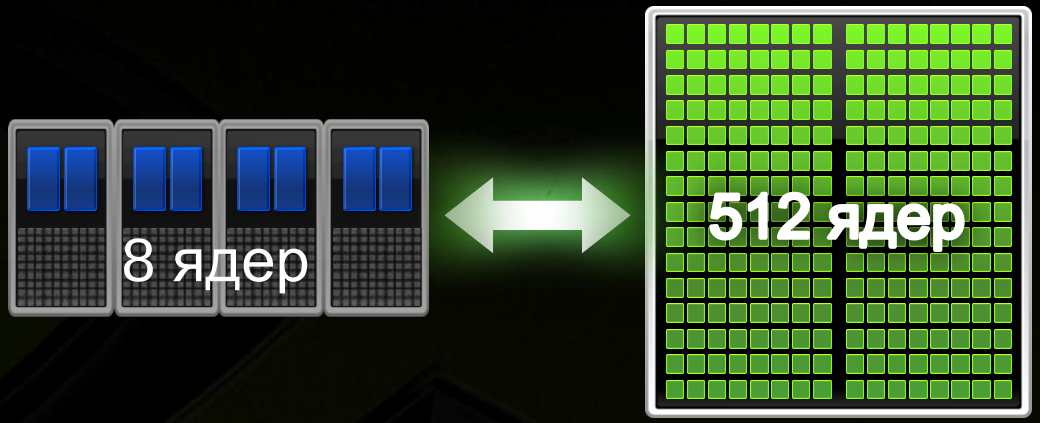
GPU NVIDIA: высокопроизводительные вычисления.



Антон Джораев, NVIDIA

Вычисления на GPU



CPU + GPU

Гибридная архитектура

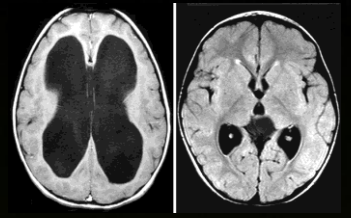
Обзор развития архитектур

Задачи для суперкомпьютеров

Разработка лекарств
МД

Проектирование
газодинамика

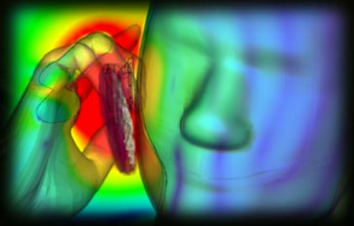
Разведка нефти и газа
Обработка сейсмических данных



Медицинская визуализация
КТ и МРТ

Астрофизика
Происхождение вселенной

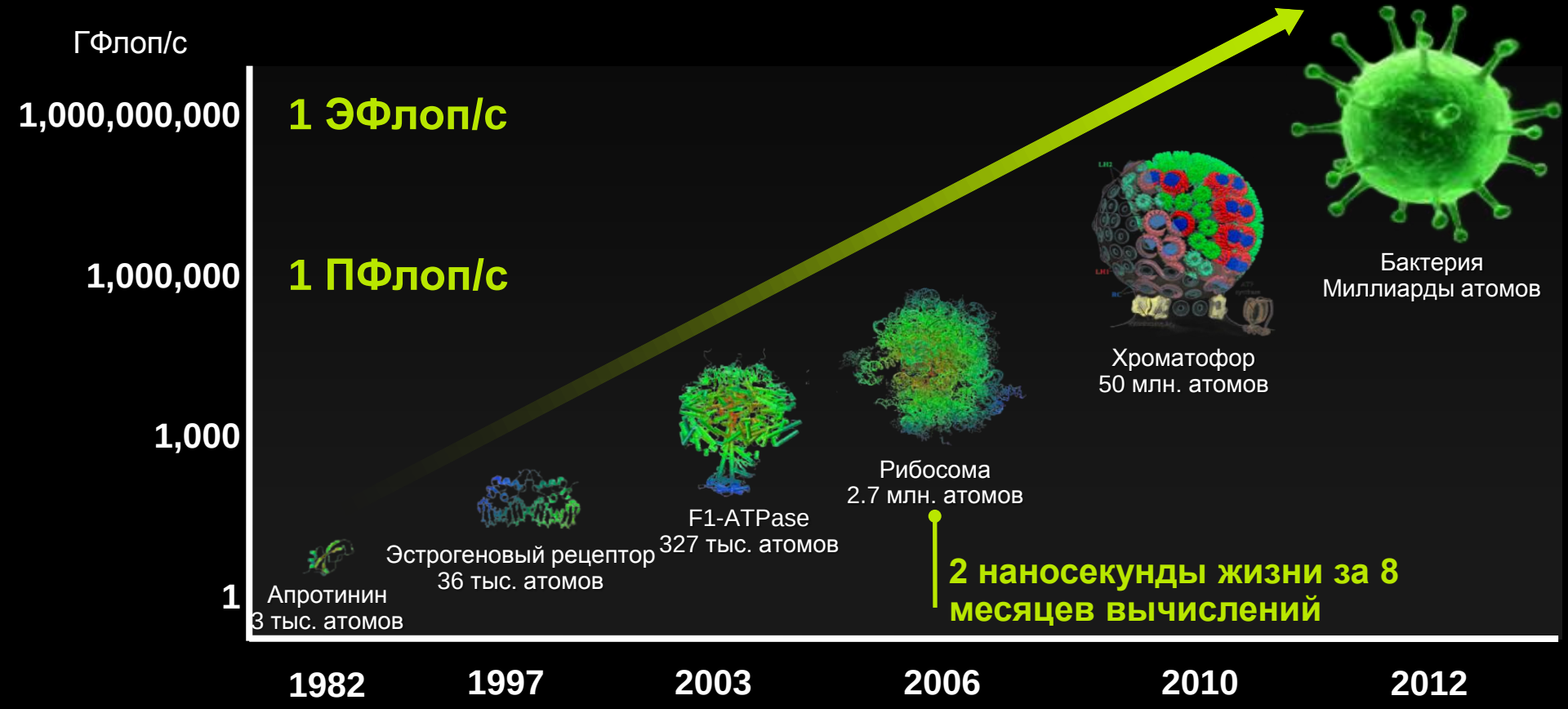
Финансы
Биржевая торговля



Разработка РЭА
Проектирование антенн

Прогнозирование погоды
Физика атмосферы

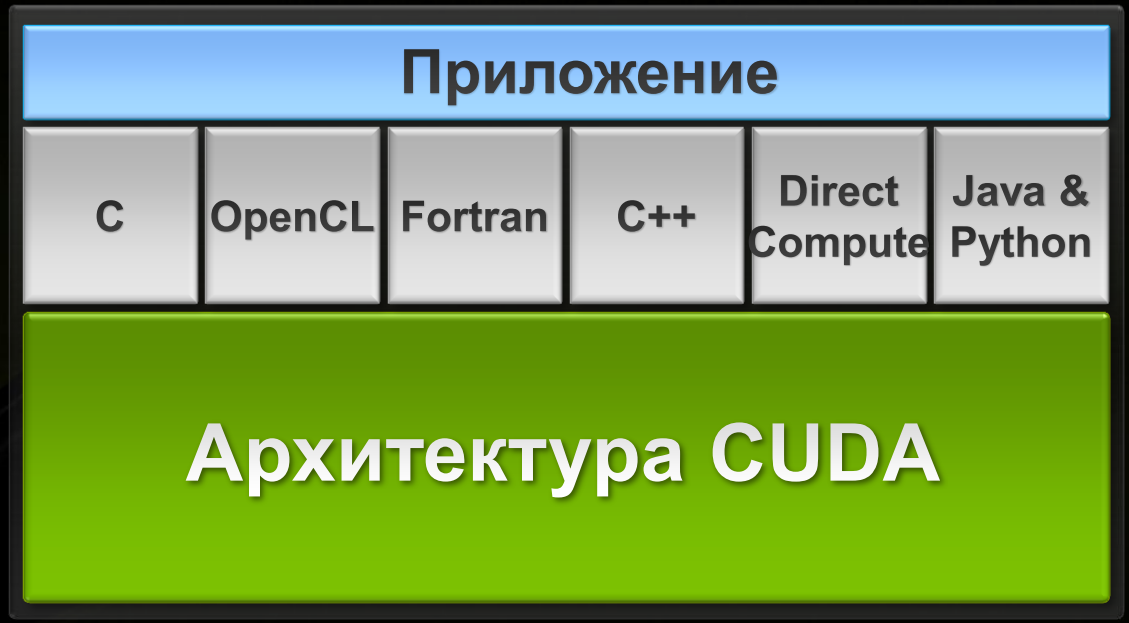
Науке нужны на порядки более мощные системы



Программно-аппаратная архитектура для параллельных вычислений

CUDA для параллельных вычислений

- Среда разработки
- Документация, сотни примеров
- Компилятор C/C++, Фортран (PGI)
- Поддержка OpenCL и DirectCompute
- Библиотеки
  - FFT
  - BLAS
  - LAPACK
  - MAGMA
  - Image processing
  - Video processing
  - Signal processing
  - Computer Vision
  - Sparse Matrix
  - Random Generator



CUDA. Мировое признание.

- 2500+ научных работ
- 1500+ приложений
- 350+ Университетов преподающих CUDA

ДОСТУПНОСТЬ

- 200+ млн. GPU с CUDA
- 100,000+ разработчиков

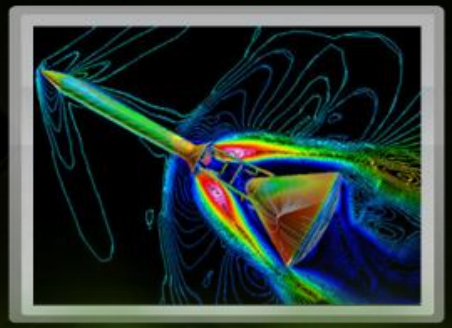

Programming Mas
Parallel Proces
基于GPU的多尺度
离散模拟并行计算
CUDA
BY EXAMPLE
An Introduction to General-Purpose GPU Programming
OpenCL
並列プログラミング
マルチコアCPU/GPUのための標準フレームワーク
CUDA
高速GPU
プログラミング入門
The OpenCL
Programming Book
Parallel Programming for MultiCore CPU and GPU
はじめての
CUDA
プログラミング
自腐のPCに
ビデオカードを
挿すだけでOK！
ОСНОВЫ РАБОТЫ
С ТЕХНОЛОГИЕЙ
CUDA
NVIDIA
GPU
高性能运算之CUDA
OpenCL入門
マルチコアCPU・GPUのための並列プログラミング

Универсальность архитектуры

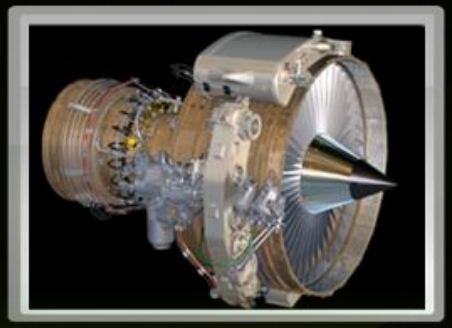
200+ млн. GPU поддерживающих CUDA по всему миру

GeForce®
Развлечения

Tesla™
Высокопроизводительные вычисления



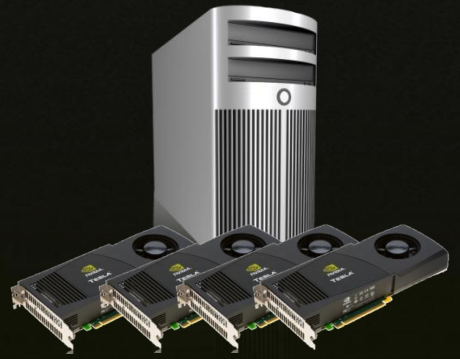
Quadro®
Конструирование и дизайн



Системы Tesla нового поколения (Fermi)

Tesla S2050 / S2070
1U System

Tesla C2050 / C2070



Tesla Personal Supercomputer
(4 Tesla C20x0)

GPUs	4 Tesla GPUs	1 Tesla GPU	4 Tesla GPUs
Произв. двойной точн.	2,06 ТФлопс	515 ГФлопс	2,06 ТФлопс
Произв. один. точн.	4,12 ТФлопс	1,03 ТФлопс	4,12 ТФлопс
Память	3 или 6 ГБ / GPU	3 или 6 ГБ	3 или 6 ГБ / GPU
Защита данных	ECC	ECC	ECC

GPU серверы стали мейнстримом

SUPERMICRO

IBM

DELL

hp

APPRO

BULL

CRAY THE SUPERCOMPUTER COMPANY

SUPERMICRO

TYAN ®

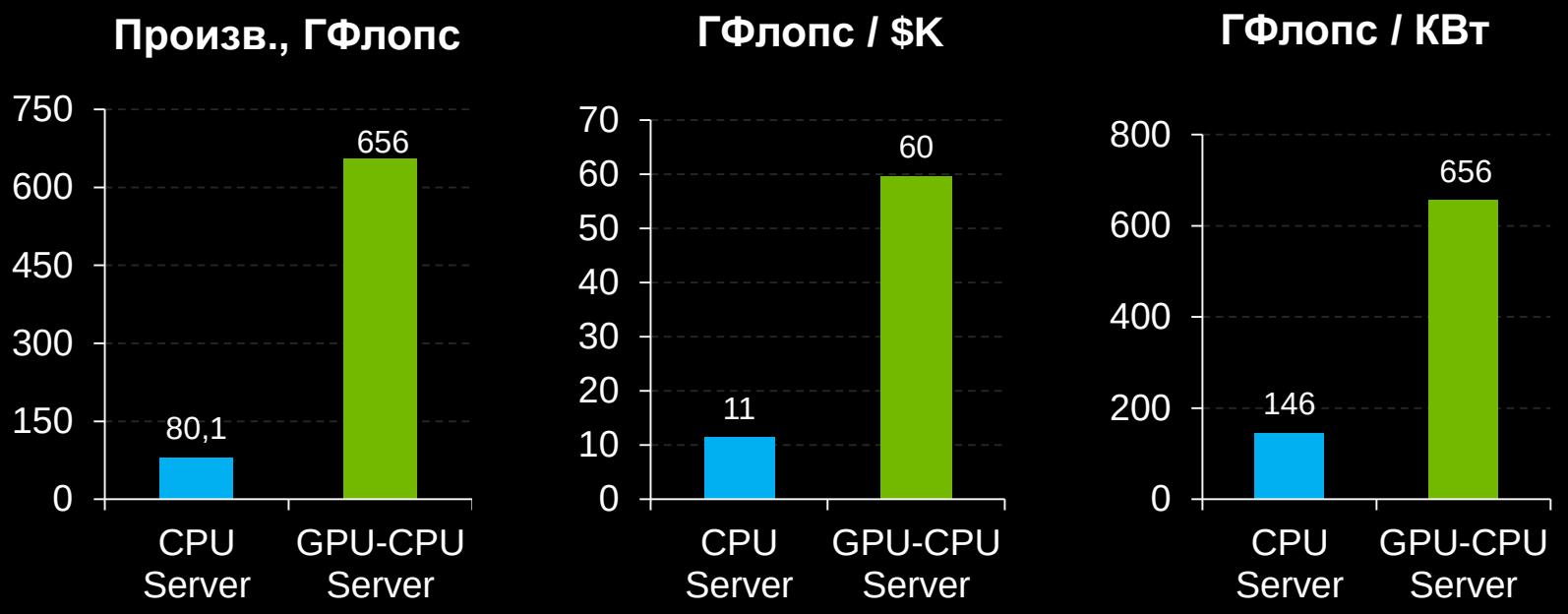
PLATFORMS

Tesla S870
Декабрь 2007

Tesla S1070 / M1060
2008-2009

Tesla M2050 / M2070
2010

В 8 раз лучше результат по 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

Мощнейшие суперкомпьютеры мира

Китай

Tianjin National Supercomputing Center

Tianhe-1A

Япония

Tokyo Institute of Technology

Tsubame 2.0

#1

Россия

Московский Государственный Университет

Ломоносов

США

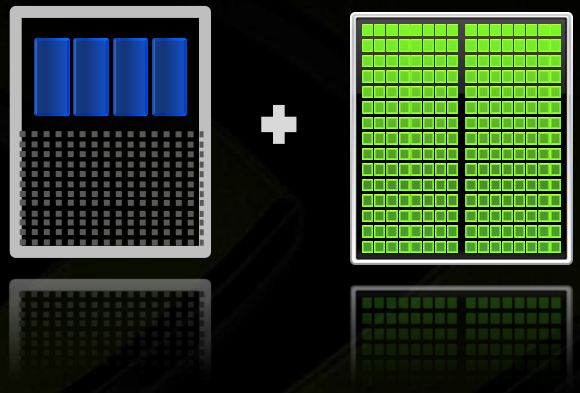
Oak Ridge National Lab

Jaguar

Официально объявлено о намерении использовать Tesla

Энергоэффективность

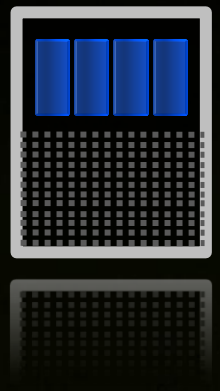
Гибридная CPU-GPU система



7,168 Tesla + 14,336 CPU

4.04 МВт

CPU-система



50,000 x86 CPUs

12.7 МВт

Китай: технологический прорыв

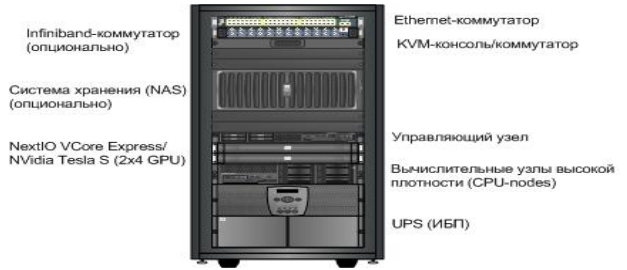
№	Вендор	Архитектура	Год	Linpack	Пик
1	国防科大	Intel X5670 + NVIDIA Tesla M2050	2010	2507	4701
2	曙光	Intel X5650 + NVIDIA Tesla C2050	2010	1271	2984
3	中科院过程所	Intel E5520 + NVIDIA Tesla C2050	2010	207	1138
4	曙光	AMD Barcelona	2008	180	233
5	联想	Intel E5450	2008	106	293
6	曙光	Intel X5650 + NVIDIA Tesla C2050	2010	76	141
7	曙光	Intel X5650 + NVIDIA Tesla C2050	2010	55	102

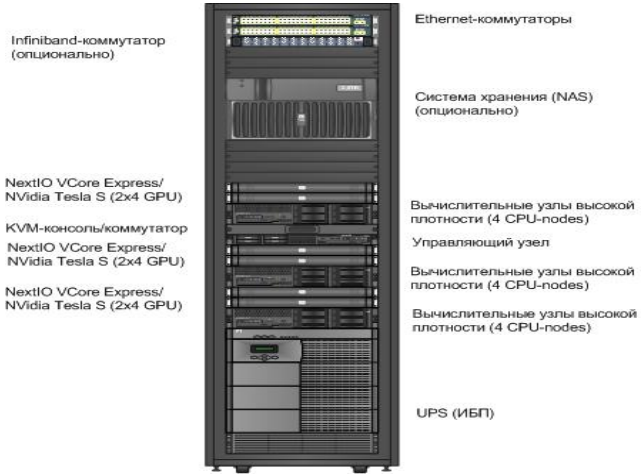
- Пять из семи самых мощных (включая три первых) – на Tesla
- Все построенные в 2010 из первой семерки – на Tesla

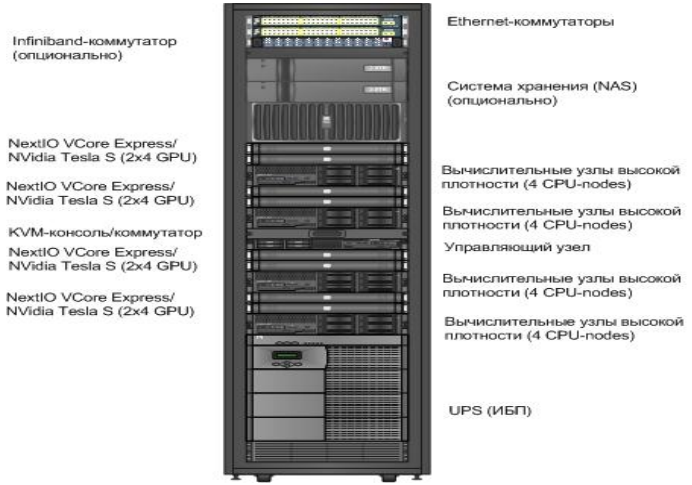
Экосистема гибридных ВС

- Проектирование и поставка высокопроизводительных вычислительных систем на базе GPU (оптимизация производительности по Linpack)
- Обучение специалистов заказчика распараллеливанию вычислений на GPU
- Аутсорсинг разработки на GPU – оптимизация и перенос существующего ПО на гибридную архитектуру

Преконфигурированные GPU-кластеры



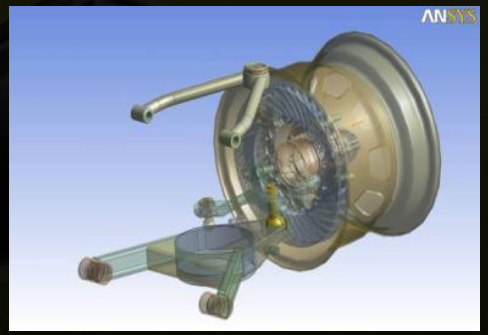




Приложения

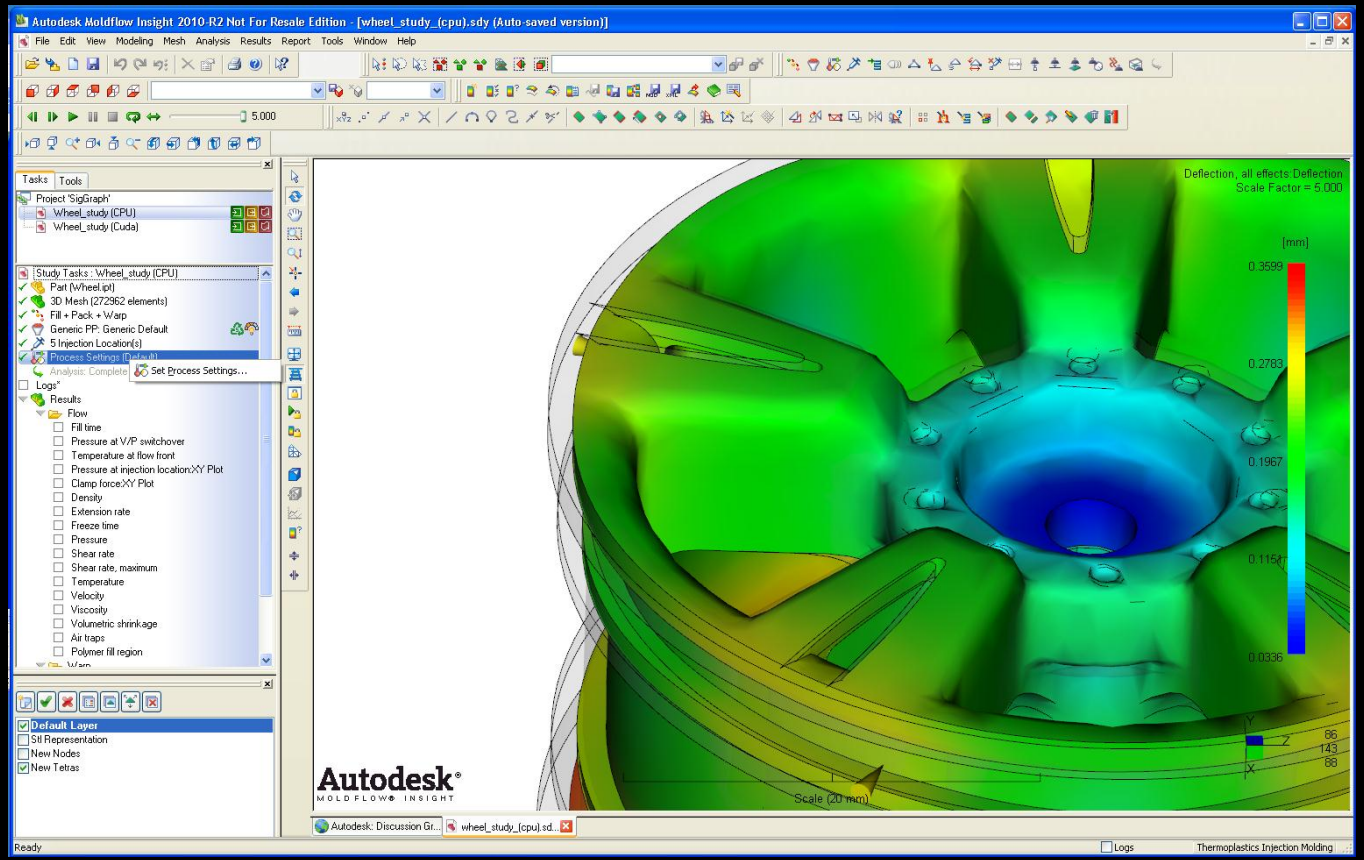
Приложения с поддержкой CUDA

Adobe

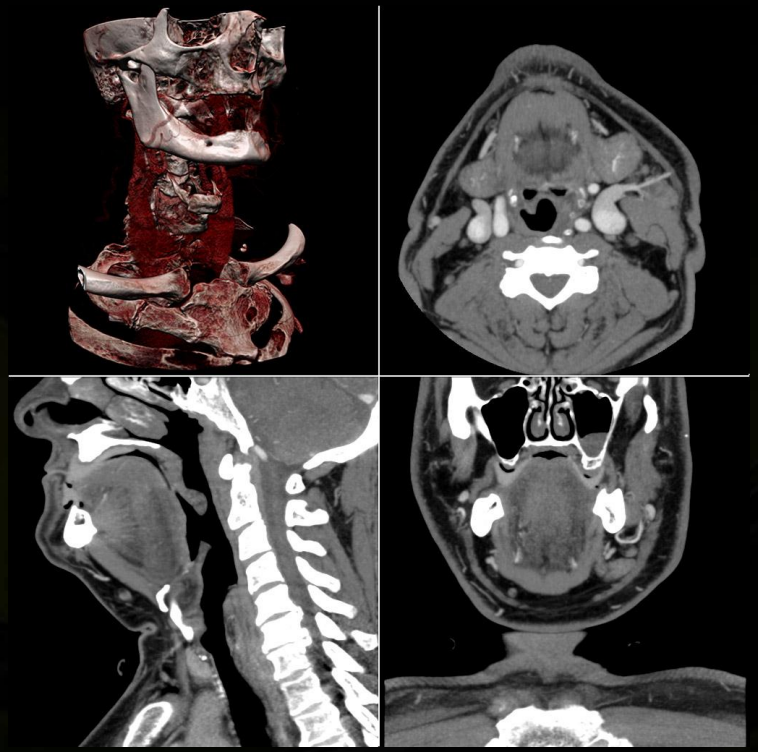


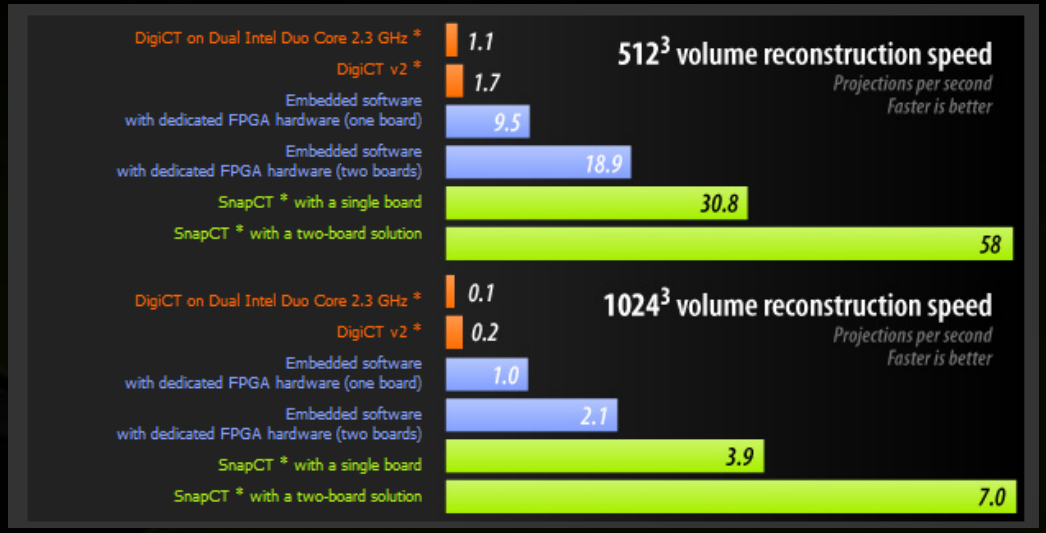
ANSYS

Moldflow – ускорение на GPU



Компьютерная томография





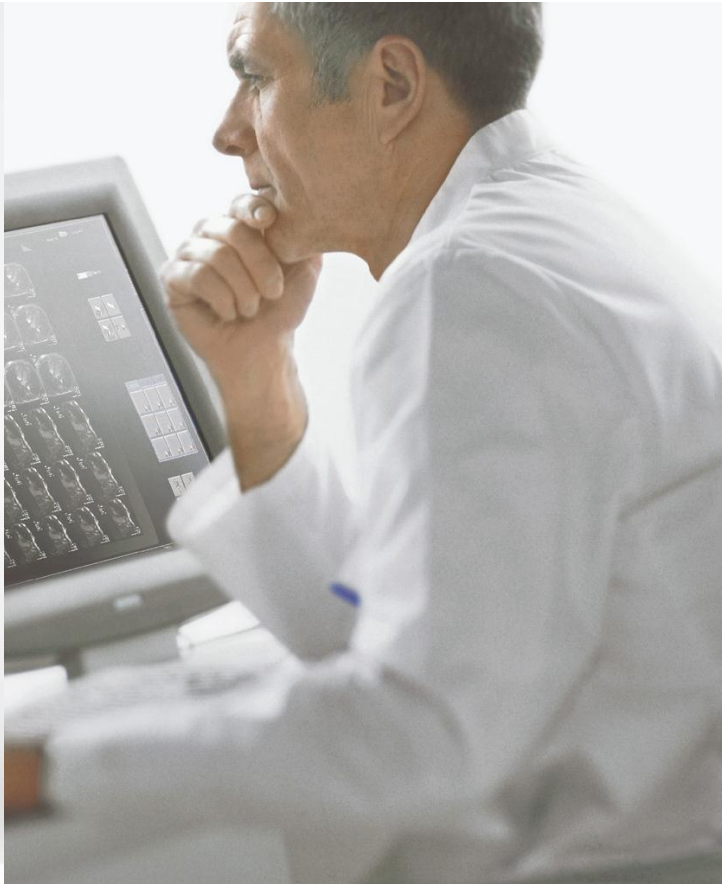
digisens SnapCT

Demo

Расчет стоимости опционов и стратегии трейдинга

- Расчеты используют симуляции Монте-Карло
- Базируются на генераторе случайных чисел
  - Ускорение на CUDA – до 100 раз
- Общее ускорение алгоритма 25-60 раз

SIEMENS

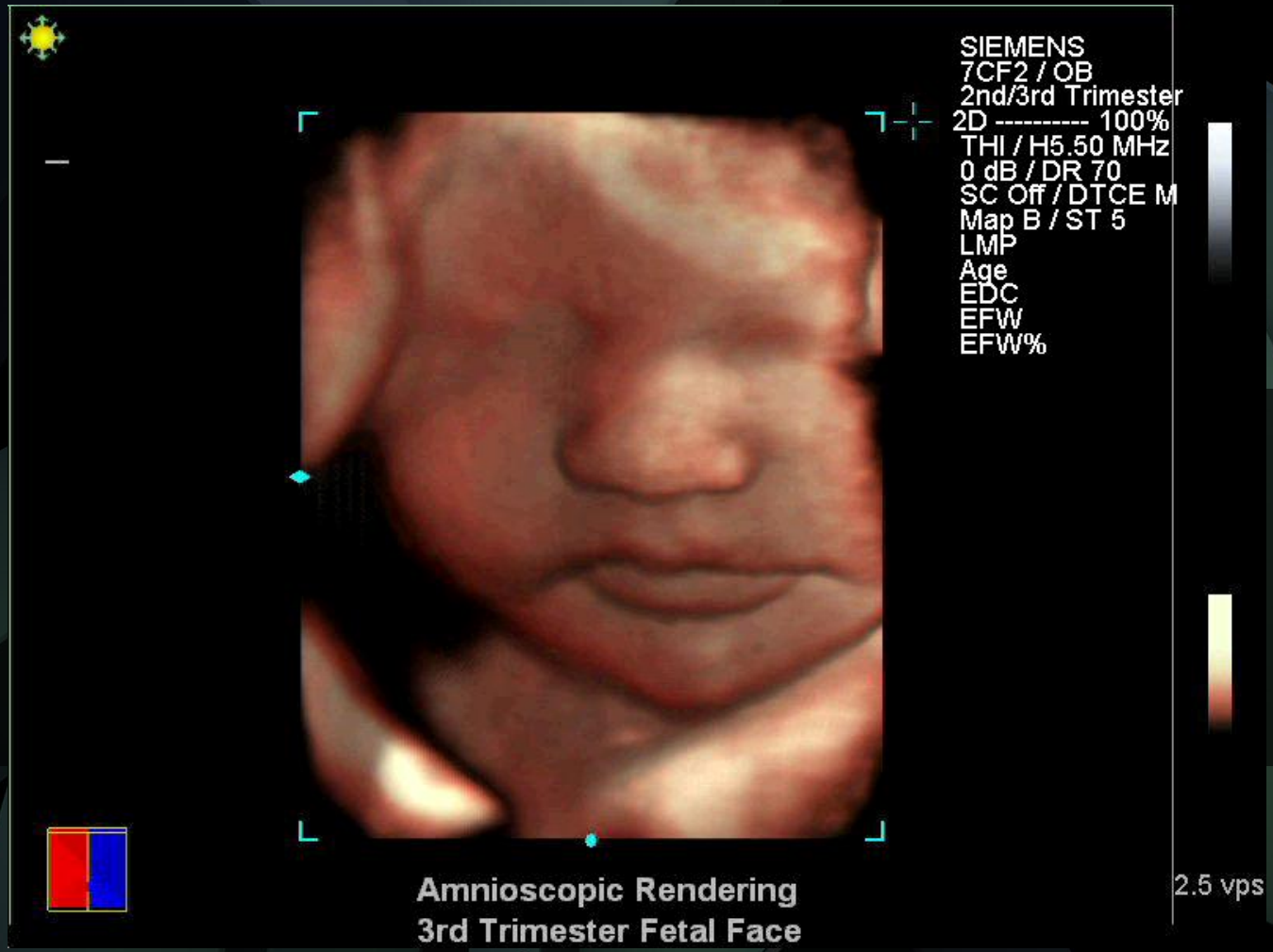


SIEMENS

Цель . Повысить наглядность и удобство диагностики плода

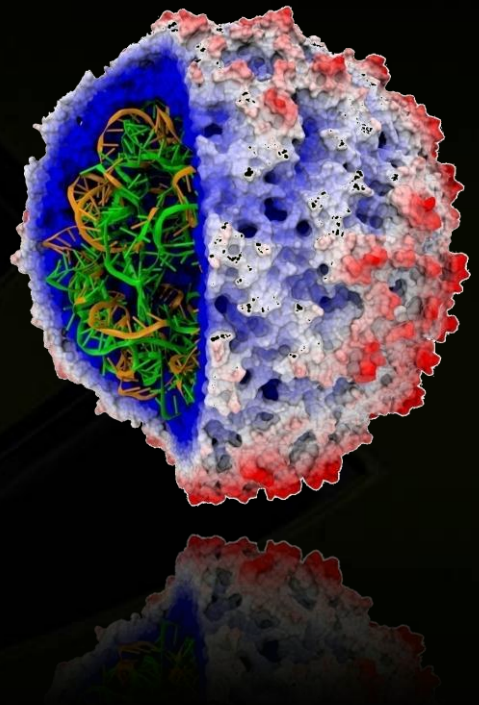
Задача . Получать результаты в высоком разрешении в режиме реального времени

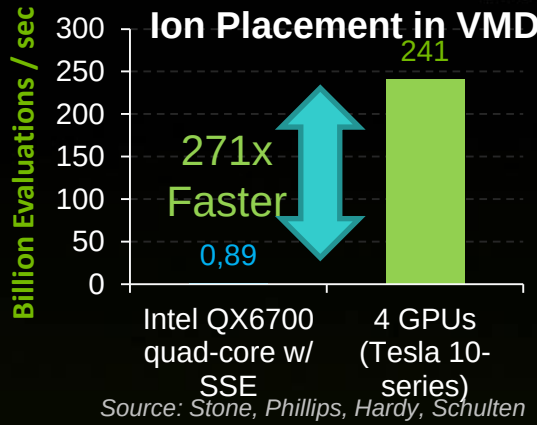
Решение . Амниоскопический рендеринг с использованием GPU NVIDIA позволяет пациентам получить трехмерное изображение младенца, выводя УЗИ на качественно новый уровень


SIEMENS
7CF2 / OB
2nd/3rd Trimester
2D ---------- 100%
THI / H5.50 MHz
0 dB / DR 70
SC Off / DTCE M
Map B / ST 5
LMP
Age
EDC
EFW
EFW%
2.5 vps
Amnioscopic Rendering
3rd Trimester Fetal Face

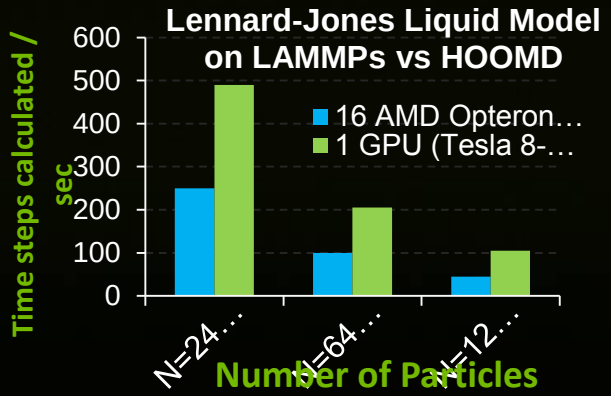
Молекулярная динамика

- **Использующие GPU**
  - NAMD / VMD
  - HOOMD
  - ACE-MD
  - MD-GPU
  - GROMACS
  - AMBER
  - LAMMPS
- **В разработке**
  - CHARMM





Source: Stone, Phillips, Hardy, Schulten



Source: Anderson, Lorenz, Travesset

KASPERSKY lab

Цель . Уменьшение времени реакции на ежесекундно появляющиеся новые угрозы

Задача . Уникальный алгоритм должен быстро сверять подозрительный файл с 50 млн. эталонов

Решение . Перенос задачи на GPU дал 360-кратный прирост производительности

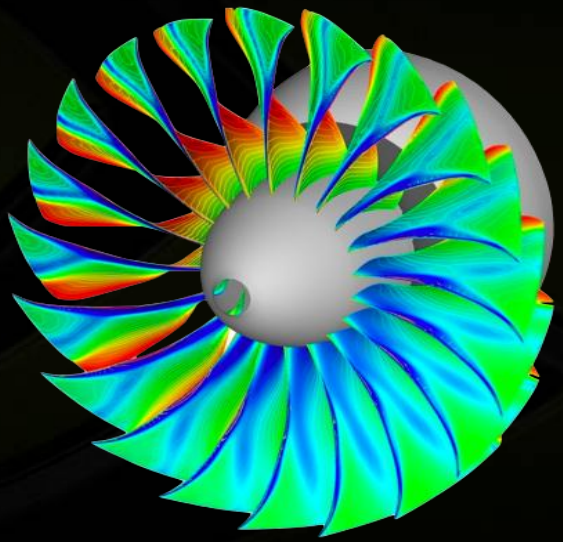
Цель . Сокращение времени обработки документов

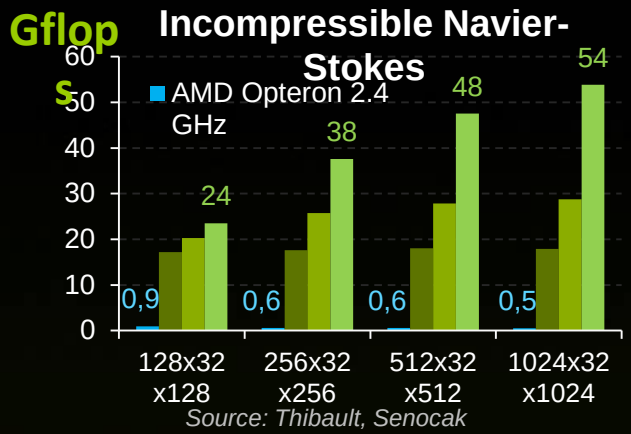
Задача . Ускорение каждого из этапов обработки изображения

Решение . Применение CUDA на одном из этапов увеличило его скорость в 30 раз, и ускорило весь процесс обработки сложных документов в 2 раза

Динамика жидкостей (CFD)

- Уравнения Навье-Стокса
- Сеточный метод Больцмана
- Трехмерные уравнения Эйлера

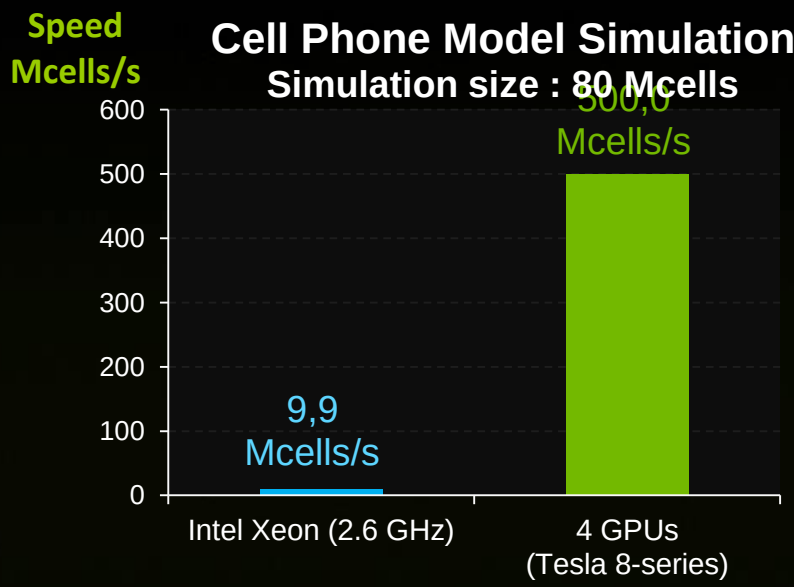




Source: Thibault, Senocak

Электромагнетизм / Электродинамика

- **Используют GPU**
  - Acceleware
  - EM Photonics
  - CST Studio
- **В разработке**
  - Maxwell equation solver
  - Ring Oscillator (FDTD)
  - Particle beam dynamics simulator



FDTD Acceleration using GPUs

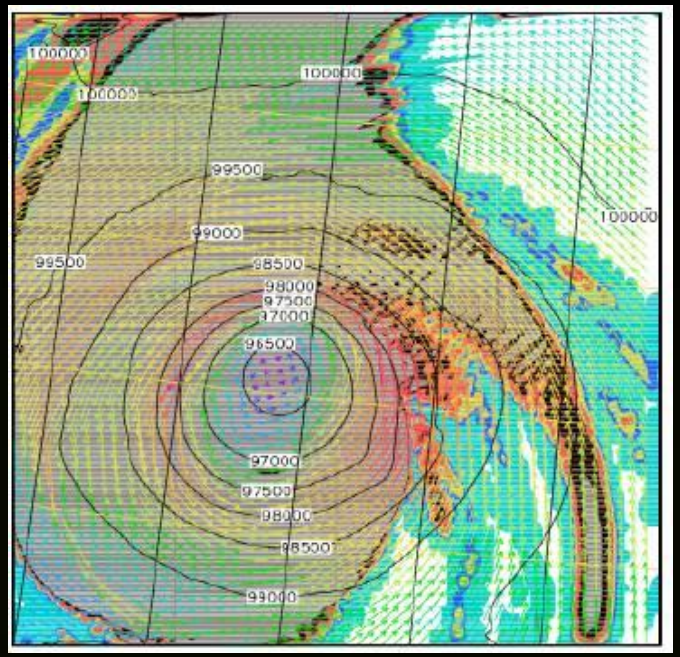
Source: Acceleware

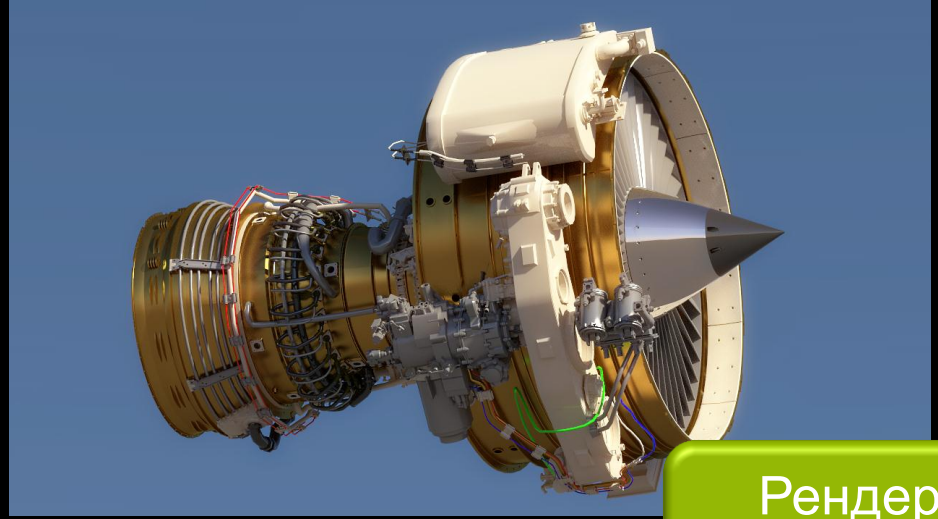
Моделирование природных явлений

Токийский Технологический Институт полностью перенес модель WRF (расчета прогнозов погоды) на GPU.

Получено 80-кратное ускорение.

Метеорологическое Агентство Японии рассматривает вариант перехода на GPU при грядущем обновлении модели расчета.





Рендеринг на GPU в CATIA v6

20+ нефтегазовых компаний используют CUDA

заказчики

Chevron, HESS, BR PETROBRAS, Schlumberger, TOTAL

вендоры ПО

acceleware, ffA, geomage, GeoStar, Headwave, SeismicCity, STONE RIDGE TECHNOLOGY

Сравнение GPU с CPU	HESS	BR PETROBRAS
производительность/ мощность	18x - 27x	12x - 17x
производительность/ площадь	20x - 31x	15x - 20x
производительность/ стоимость	15x - 20x	10x - 12x

CAE ISV Pipeline

	Available Now	Expected Release in 2011
Structural Mechanics	ANSYS Mechanical AFEA Abaqus/Standard (beta)	LS-DYNA *implicit* Marc
Fluid Dynamics	AcuSolve Moldflow Culises (OpenFOAM) Particleworks	CFD++ LS-DYNA CFD
Electromagnetics	Nexxim EMPro CST MS XFdtd SEMCAD X	Xpatch


Спасибо за внимание
Антон Джораев
adzhoraev@nvidia.com
NVIDIA

CUDA 4.0

Портирование приложений становится доступнее

Быстрый обмен между GPU
GPU Direct 2.0

Облегченное портирование приложений
Unified Virtual Addressing

Параллельное программирование в C++
Thrust

NVIDIA GPUDirect™: *Towards Eliminating the CPU Bottleneck*

Version 1.0

for applications that communicate over a network

- Direct access to GPU memory for 3rd party devices
- Eliminates unnecessary sys mem copies & CPU overhead
- Supported by Mellanox and Qlogic
- Up to 30% improvement in communication performance

Version 2.0

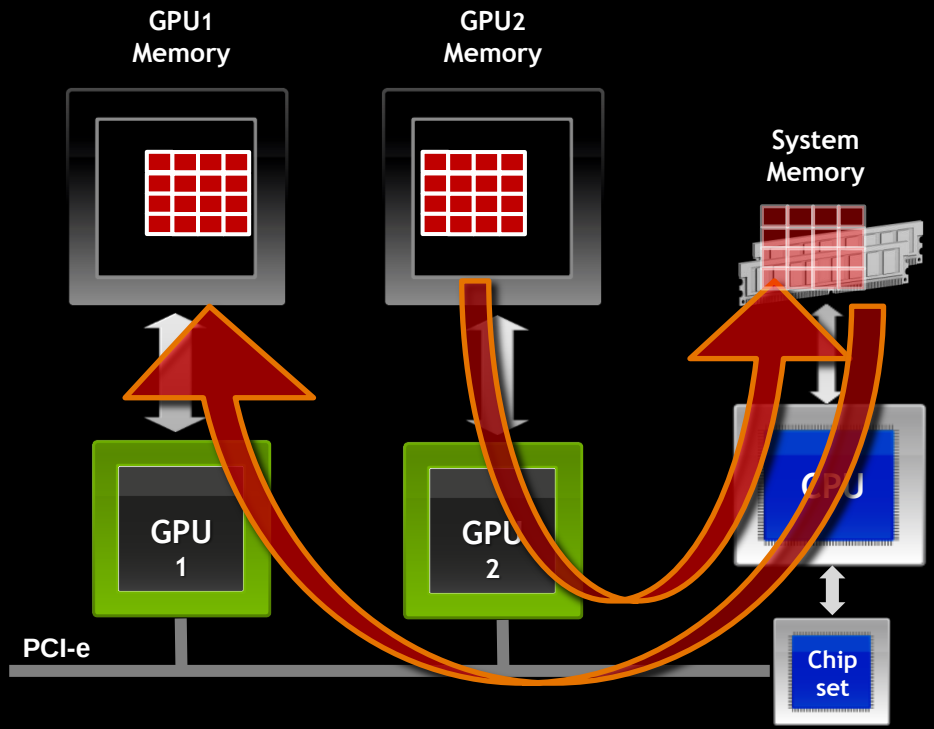
for applications that communicate within a node

- Peer-to-Peer memory access, transfers & synchronization
- Less code, higher programmer productivity

Details @ http://www.nvidia.com/object/software-for-tesla-products.html

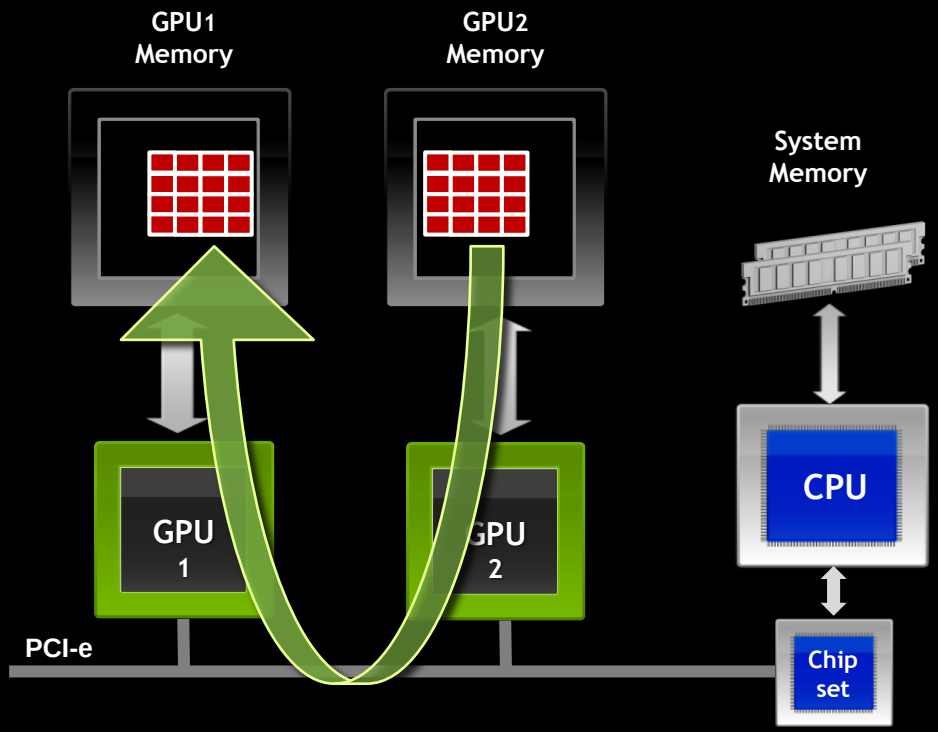
Before GPUDirect v2.0

Required Copy into Main Memory



GPUDirect v2.0: Peer-to-Peer Communication

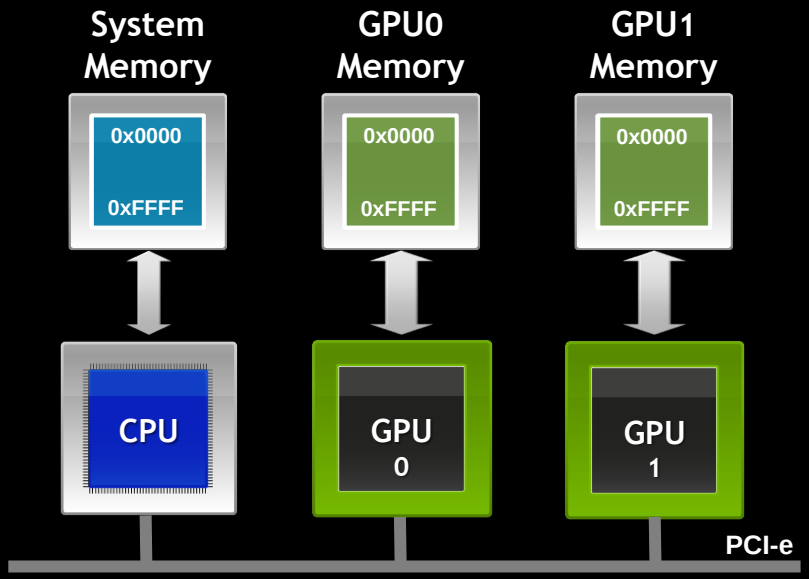
Direct Transfers b/w GPUs



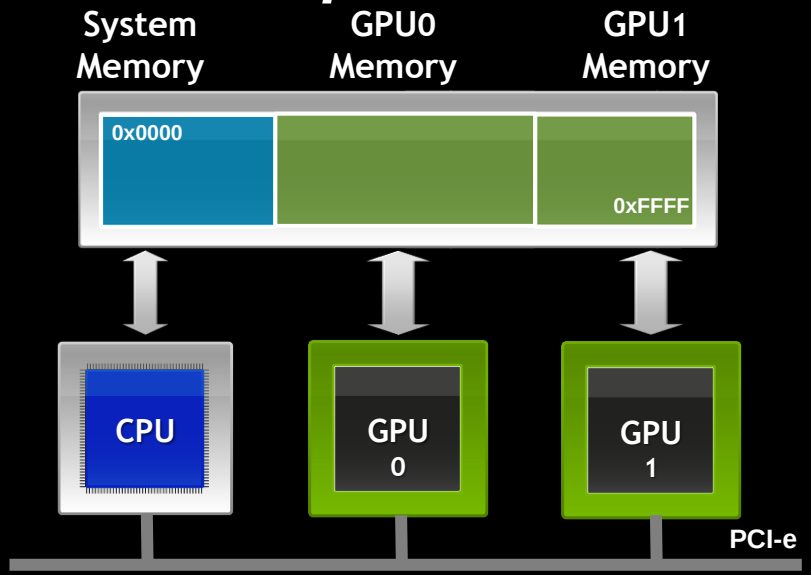
Unified Virtual Addressing

Easier to Program with Single Address Space

No UVA: Multiple Memory Spaces



UVA : Single Address Space



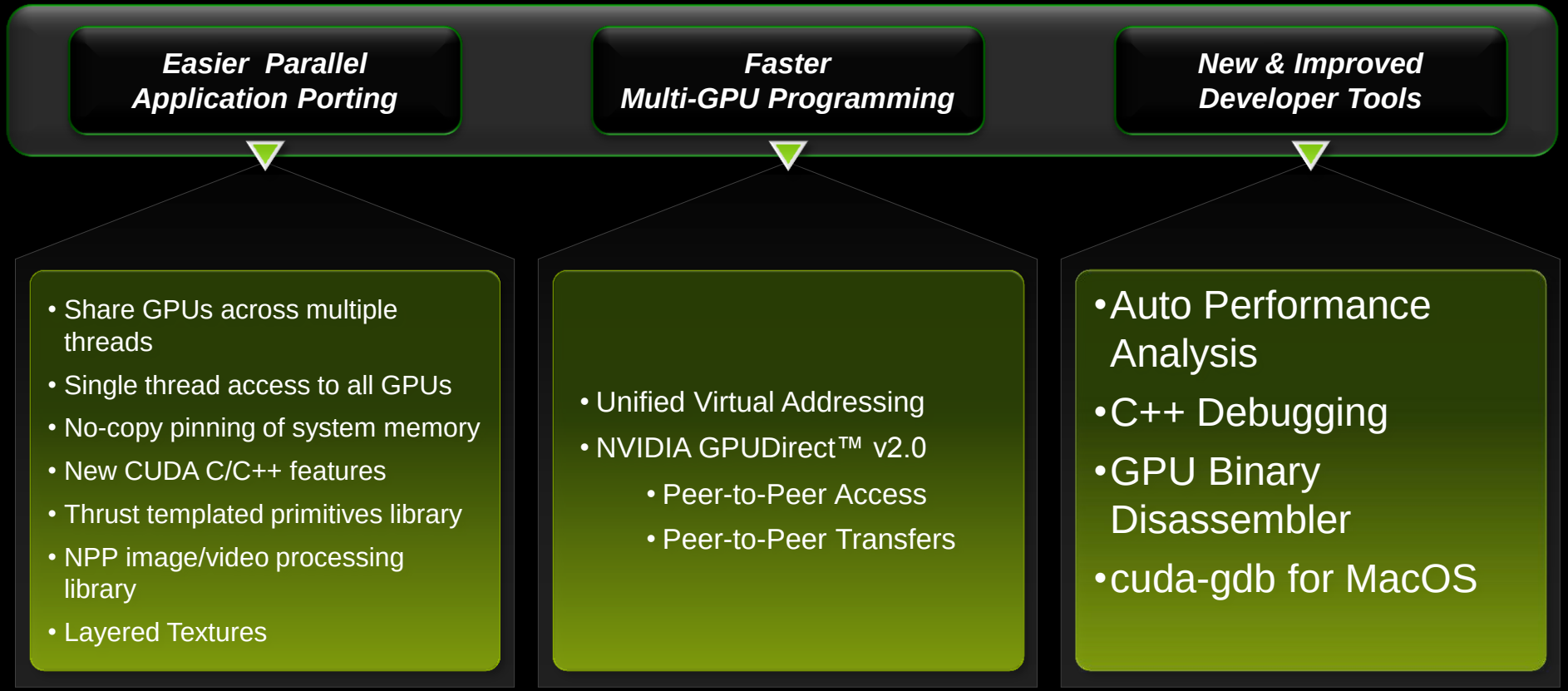
C++ Templatized Algorithms & Data Structures (Thrust)

- **Powerful open source C++ parallel algorithms & data structures**
  - **Similar to C++ Standard Template Library (STL)**
- **Automatically chooses the fastest code path at compile time**
  - **Divides work between GPUs and multi-core CPUs**
  - **Parallel sorting @ 5x to 100x faster than STL and TBB**

Data Structures
• `thrust::device_vector`
• `thrust::host_vector`
• `thrust::device_ptr`
• Etc.

Algorithms
• `thrust::sort`
• `thrust::reduce`
• `thrust::exclusive_scan`
• Etc.

CUDA 4.0: Highlights



NVIDIA CUDA Summary

New in CUDA 4.0

Platform

GPUDirect 2.0
Fast Path to Data

Hardware Support
ECC Memory
Double Precision
Native 64-bit Architecture
Concurrent Kernel Execution
Dual Copy Engines
Multi-GPU support
6GB per GPU supported

Operating System Support
MS Windows 32/64
Linux 32/64 support
Mac OSX support

Cluster Management
GPUDirect
Tesla Compute Cluster (TCC)
Graphics Interoperability

Programming Model

Unified Virtual Addressing

C++ new/delete

C++ Virtual Functions

C support
- NVIDIA C Compiler
- CUDA C Parallel Extensions
- Function Pointers
- Recursion
- Atomics
- malloc/free

C++ support
- Classes/Objects
- Class Inheritance
- Polymorphism
- Operator Overloading
- Class Templates
- Function Templates
- Virtual Base Classes
- Namespaces

Fortran, OpenCL

Libraries

Thrust C++ Library
Templated Performance Primitives

NVIDIA Library Support
Complete math.h
Complete BLAS Library (1, 2 and 3)
Sparse Matrix Math Library
RNG Library
FFT Library (1D, 2D and 3D)
Image Processing Library (NPP)
Video Processing Library (NPP)

3rd Party Math Libraries
- CULA Tools
- MAGMA
- IMSL
- VSIPL

Tools

Parallel Nsight Pro

NVIDIA Tools Support
Parallel Nsight 1.0 IDE
cuda-gdb Debugger with multi-GPU
CUDA/OpenCL Visual Profiler
CUDA Memory Checker
CUDA C SDK
CUDA Disassembler

CUDA Partner Tools
Allinea DDT
RogueWave /Totalview
Vampir
Tau
CAPS HMPP