NVIDIA

ВВЕДЕНИЕ

Перепёлкин Евгений Евгеньевич

Почему GPU?

ЭВОЛЮЦИЯ ЦЕНТРАЛЬНЫХ ПРОЦЕССОРОВ

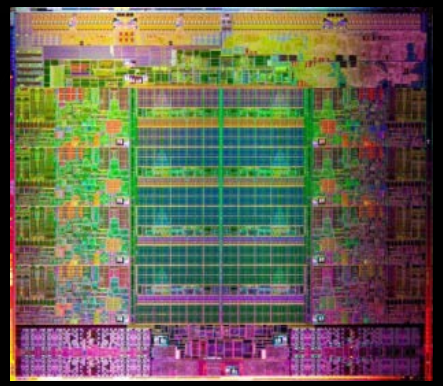
Частоты CPU

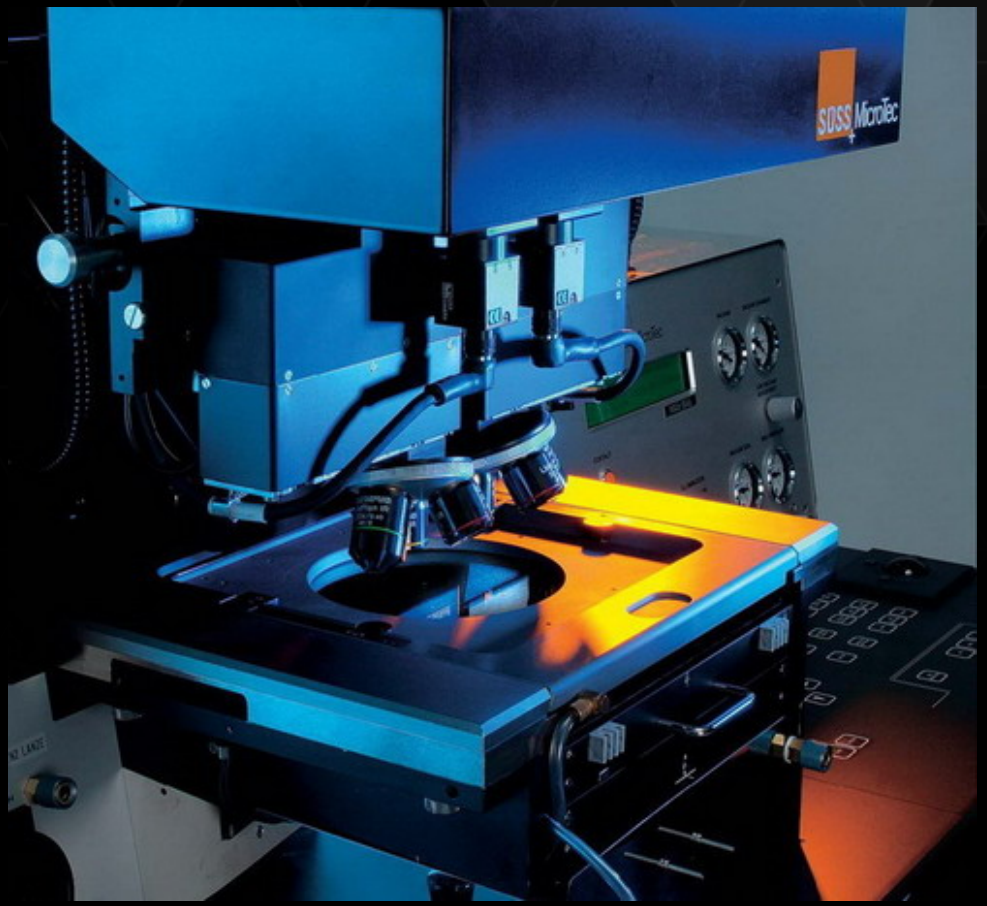
- 2004 г. Pentium 4, 3.46 GHz
- 2005 г. Pentium 4, 3.8 GHz
- 2006 г. Core Duo T2700, 2.33 GHz
- 2007 г. Core 2 Duo E6700, 2.66 GHz
- 2007 г. Core 2 Duo E6800, 3 GHz
- 2008 г. Core 2 Duo E8600, 3.33 GHz
- 2009 г. Core i7 950, 3.06 GHz (4 ядра)
- 2010 г. Core i7 9xxX, 3.2 - 3.47 GHz (6 ядер)

ПРОБЛЕМЫ

Рост частоты практически отсутствует

- Энерговыделение ~ второй степени частоты
- Ограничения техпроцесса

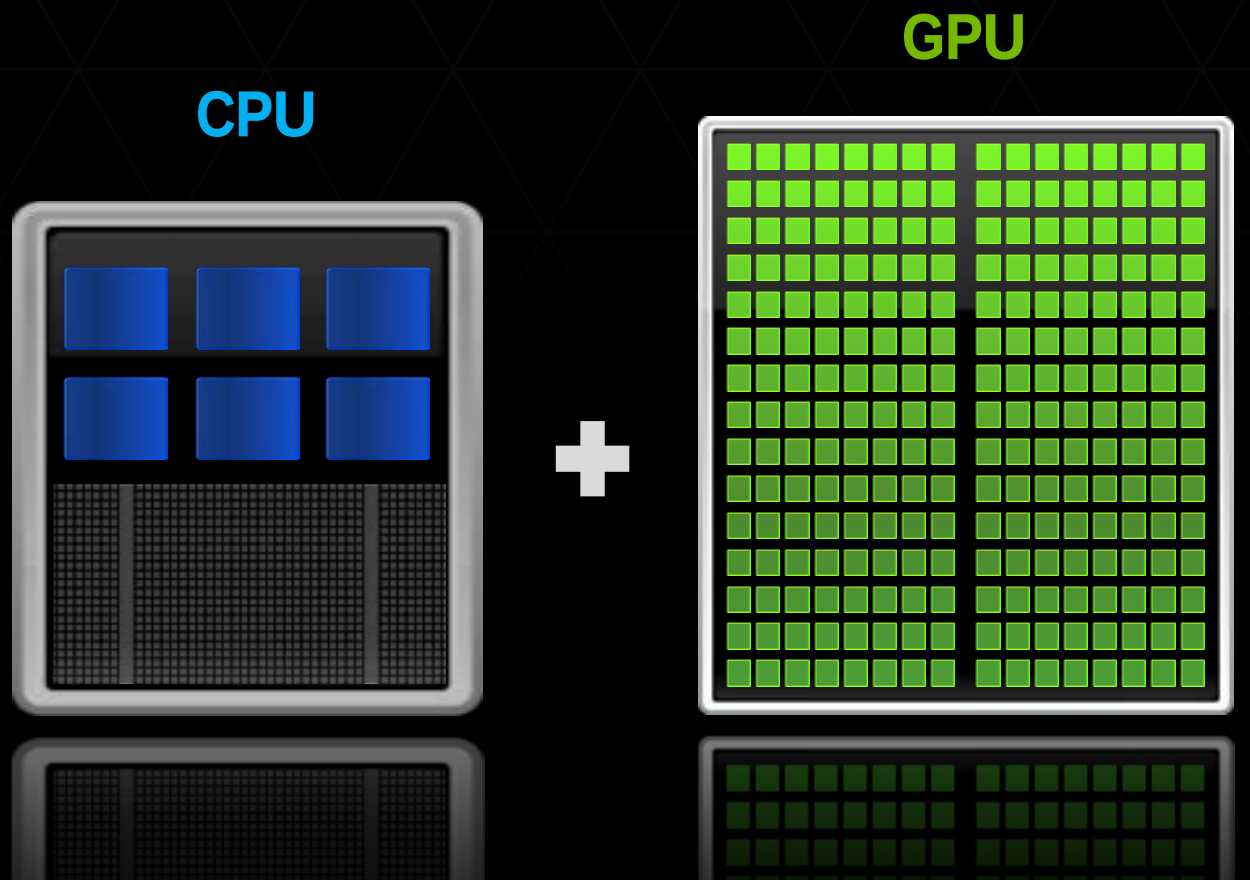




Гибридная модель вычислений

РАБОТА В ТАНДЕМЕ

Распределение вычислительной нагрузки между архитектурами



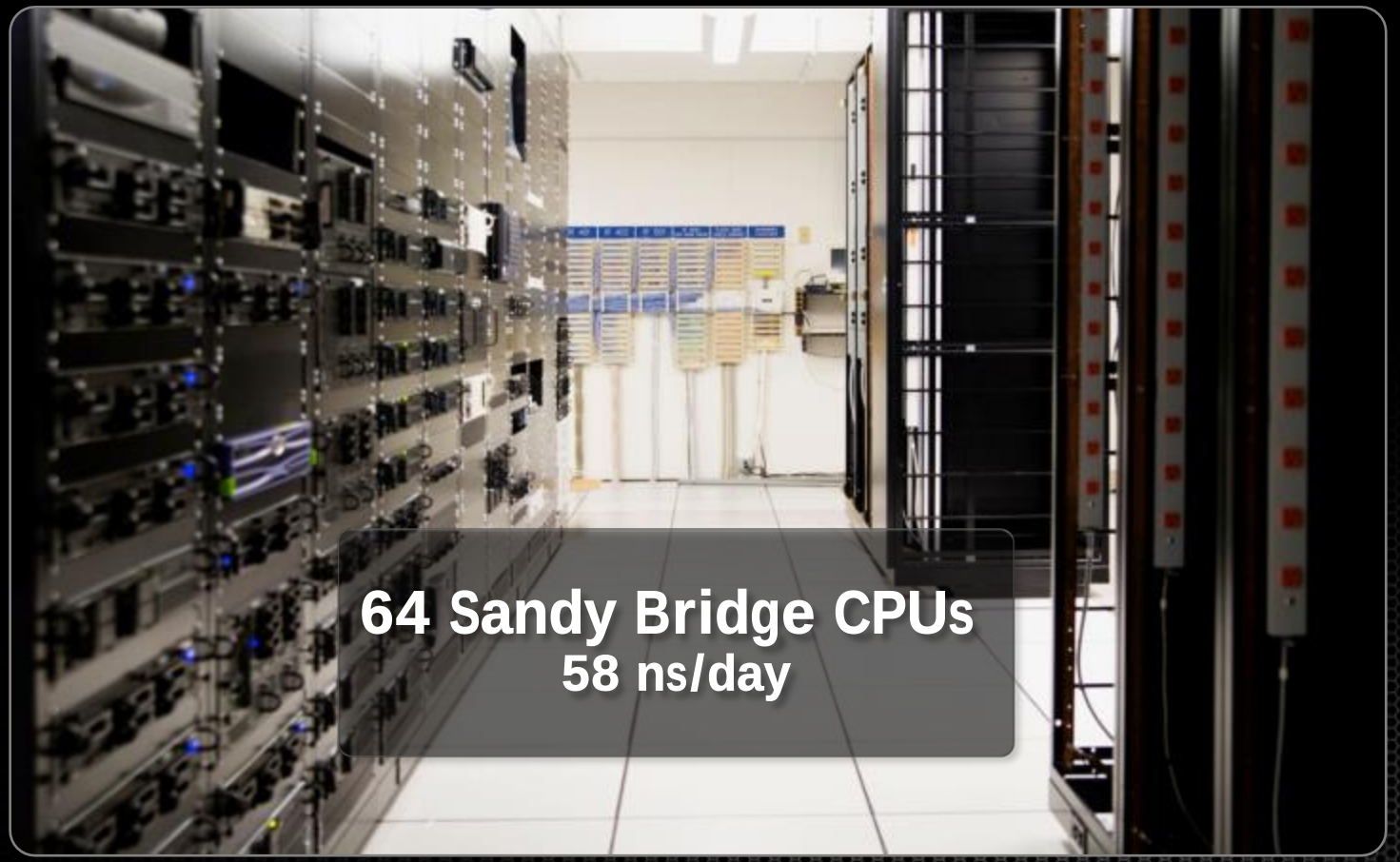
ГИБРИДНАЯ МОДЕЛЬ ВЫЧИСЛЕНИЙ

GPUs POWER WORLD'S 10 GREENEST SUPERCOMPUTERS

Green500 Rank	MFLOPS/W	Site
1	4,503.17	GSIC Center, Tokyo Tech
2	3,631.86	Cambridge University
3	3,517.84	University of Tsukuba
4	3,185.91	Swiss National Supercomputing (CSCS)
5	3,130.95	ROMEO HPC Center
6	3,068.71	GSIC Center, Tokyo Tech
7	2,702.16	University of Arizona
8	2,629.10	Max-Planck
9	2,629.10	(Financial Institution)
10	2,358.69	CSIRO
37	1959.90	Intel Endeavor (top Xeon Phi cluster)
49	1247.57	Météo France (top CPU cluster)

REVOLUTIONIZING SCIENTIFIC COMPUTING

AMBER Molecular Dynamics Simulation
DHFR NVE Benchmark



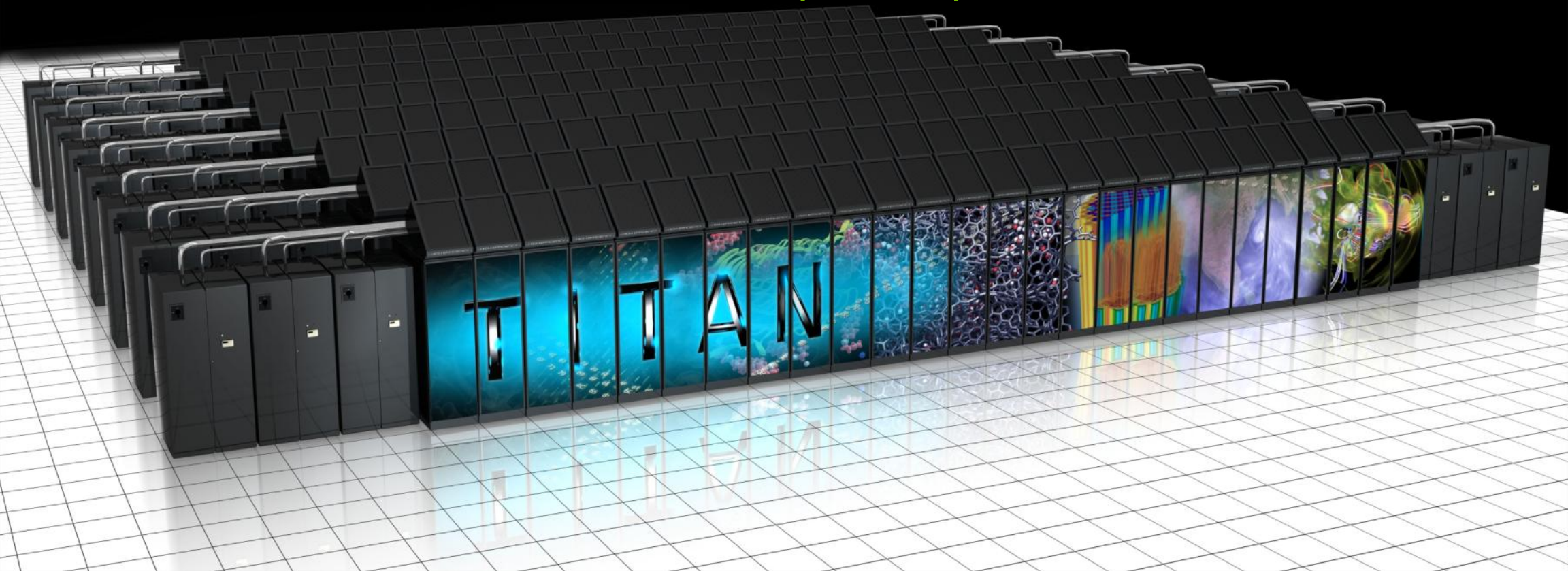
1 Tesla K40 GPU
102 ns/day

СУПЕРКОМПЬЮТЕР TITAN

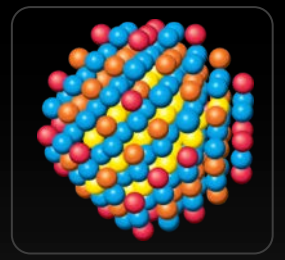
18,688 Tesla K20X GPU

27 Petaflops пик: 90% производительности обеспечено GPU

17.59 Petaflops в Linpack

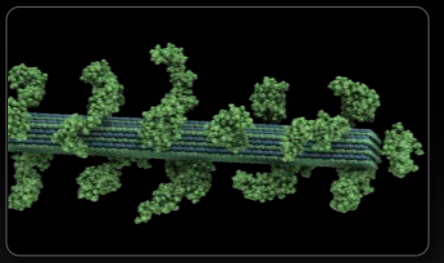


ПЕРЕДОВЫЕ ИССЛЕДОВАНИЯ НА TITAN



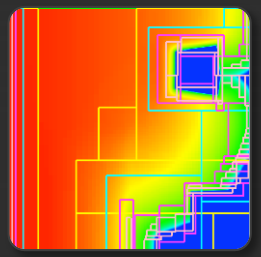
Материаловедение (WL-LSMS)
Role of material disorder, statistics, and fluctuations in nanoscale materials and systems.

Климат (CAM-SE)
Answer questions about specific climate change adaptation and mitigation scenarios; realistically represent features like precipitation patterns/statistics and tropical storms.

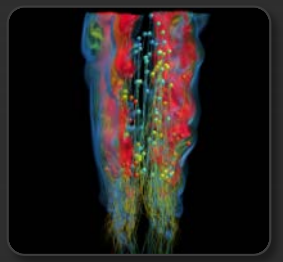


Биотопливо (LAMMPS)
A multiple capability molecular dynamics code.

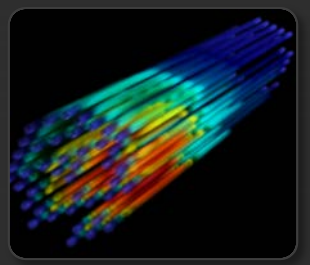
Астрофизика (NRDF)
Radiation transport – critical to astrophysics, laser fusion, combustion, atmospheric dynamics, and medical imaging.



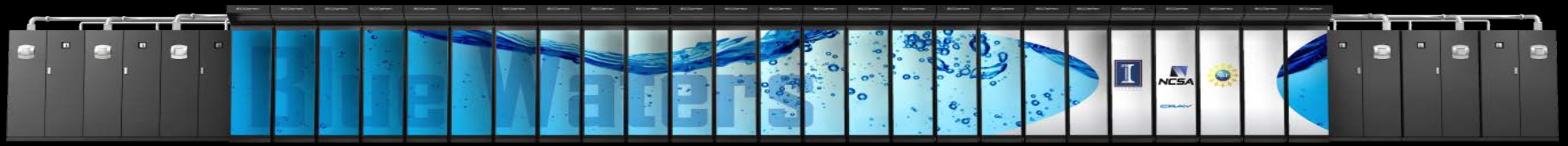
Сгорание топлива (S3D)
Combustion simulations to enable the next generation of diesel/bio-fuels to burn more efficiently.



Атомная энергия (Denovo)
Unprecedented high-fidelity radiation transport calculations that can be used in a variety of nuclear energy and technology applications.



NCSA ВЫБРАЛ GPU ДЛЯ СИСТЕМЫ BLUE WATERS

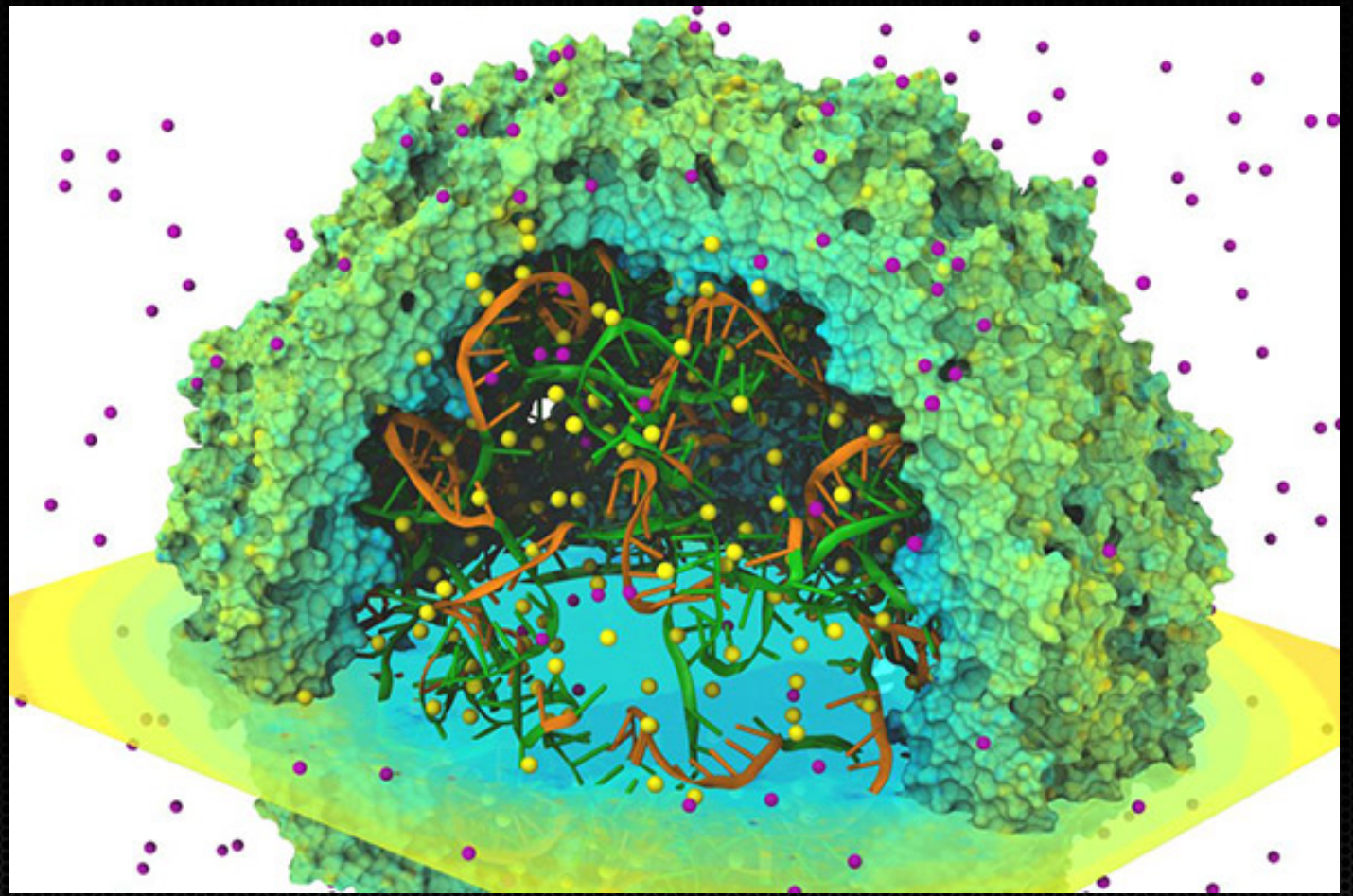


“NCSA воодушевлен включением GPU NVIDIA Tesla в Blue Waters. GPU предоставляют экстраординарные вычислительные возможности, энергоэффективность и экономическую выгоду для завтрашних систем петафлопсного уровня.”

Thom Dunning
Директор NCSA

НОБЕЛЕВСКАЯ ПРЕМИЯ ПО ХИМИИ ЗА 2013 Г.

Моделирование молекулярной динамики на GPU



ACCELERATING MAINSTREAM DATACENTERS

Oil & Gas

Schlumberger

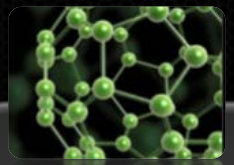
BR PETROBRAS

Chevron

Eni

Statoil

Paradigm



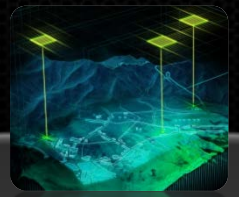
Higher Ed

Chinese Academy of Sciences

Georgia Tech

HARVARD School of Engineering and Applied Sciences

STANFORD UNIVERSITY



Government

Air Force Research Laboratory

Raytheon

NASA

Naval Research Laboratory

MITRE

Supercomputing

CSCS Swiss National Supercomputing Centre

NCSA

Tokyo Institute of Technology

OAK RIDGE National Laboratory

Finance

J.P.Morgan

BARCLAYS

STANDARD LIFE

BNP PARIBAS

MUREX

Web 2.0

Baidu 百度

salesforce SOFTWARE

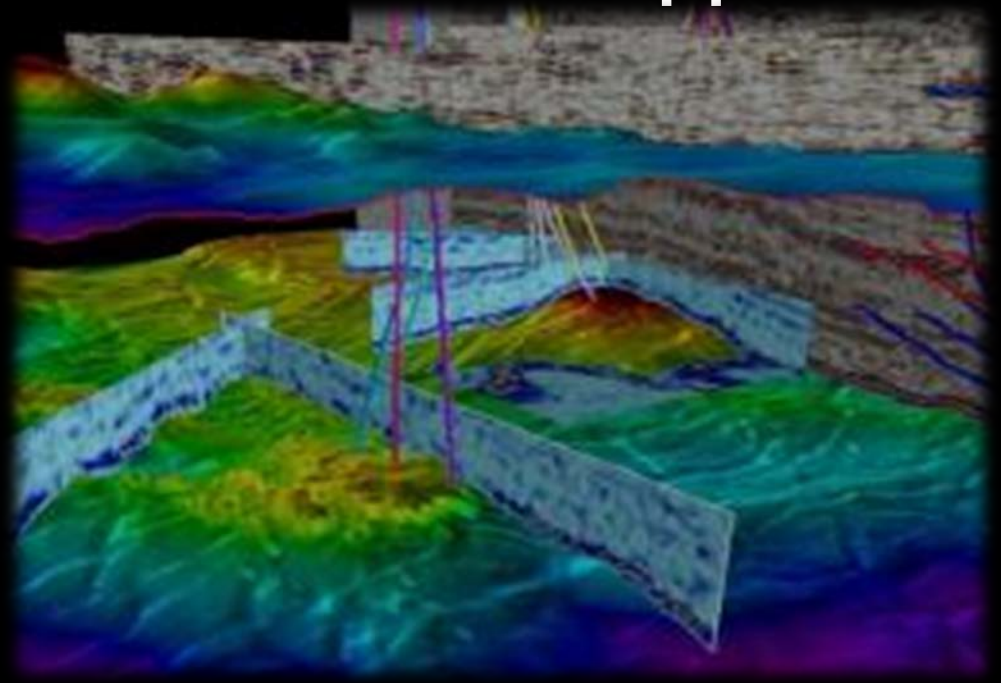
SHAZAM

amazon.com

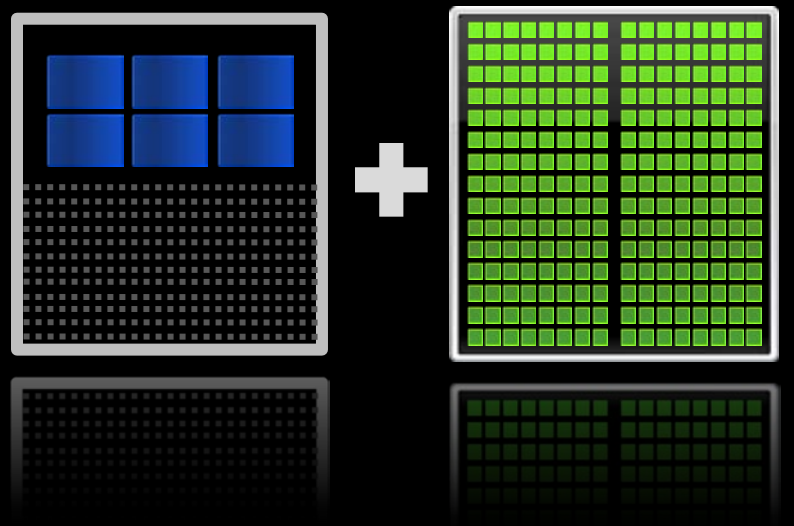
OIL & GAS: СЕЙСМОРАЗВЕДКА

1	**Идентичная производительность**	1
32 Tesla S1070s	**31x экономия места**	2000 CPU Servers
~$400 K	**20x ниже стоимость**	~$8 M
45 kWatts	**27x ниже энергопотребление**	1200 kWatts

GPU ДЛЯ СЕЙСМОАНАЛИЗА



Global Tier 1 RTM
Seismic Contractor



Время моделирования	7 дней вместо 28 дней
Расходы на обеспечение	75% экономия
Общее число Tesla GPU	тысячи

BLOOMBERG: МОДЕЛИРОВАНИЕ РЫНКА АКЦИЙ

48 GPUs	**42x экономия места**	2000 CPUs
$144K	**28x ниже стоимость**	$4 Million
$31K / year	**38x ниже энергопотребление**	$1.2 Million / year

GPU ДЛЯ ФИНАНСОВОГО СЕКТОРА

Примеры от партнеров

Бонды

2 часа вместо 16 часов

Bloomberg

Капитализация

В 10 раз меньше энергии

BNP PARIBAS

Страхование

Минуты вместо дней

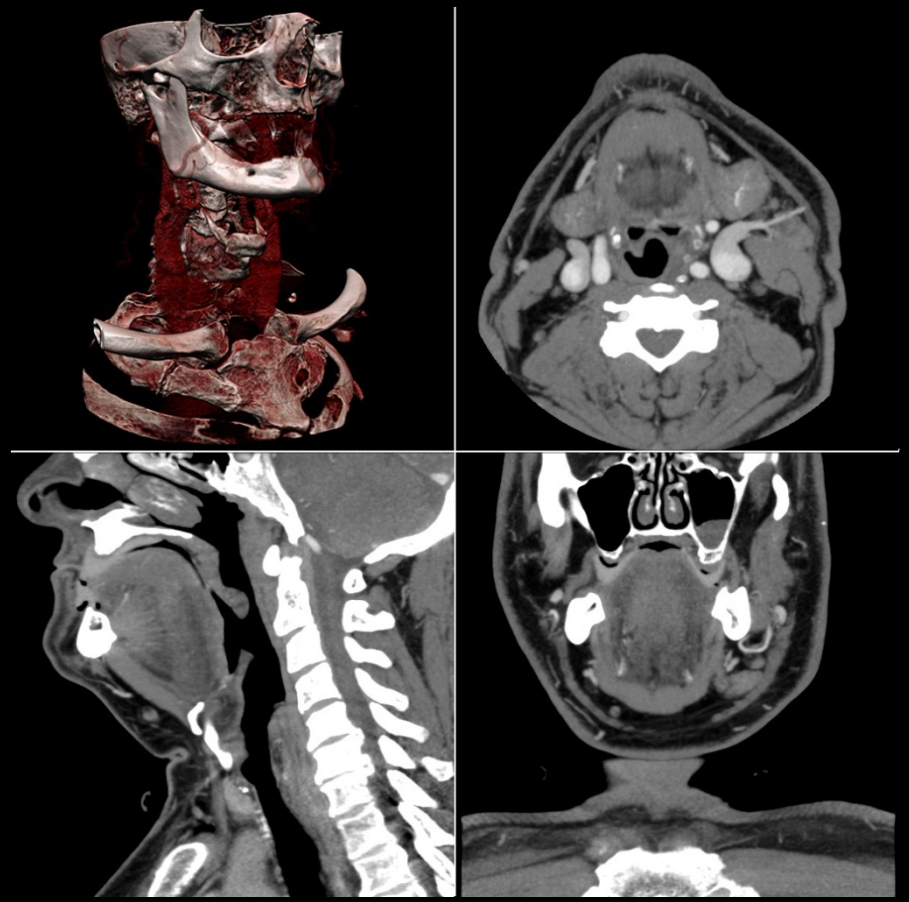
AON BENFIELD

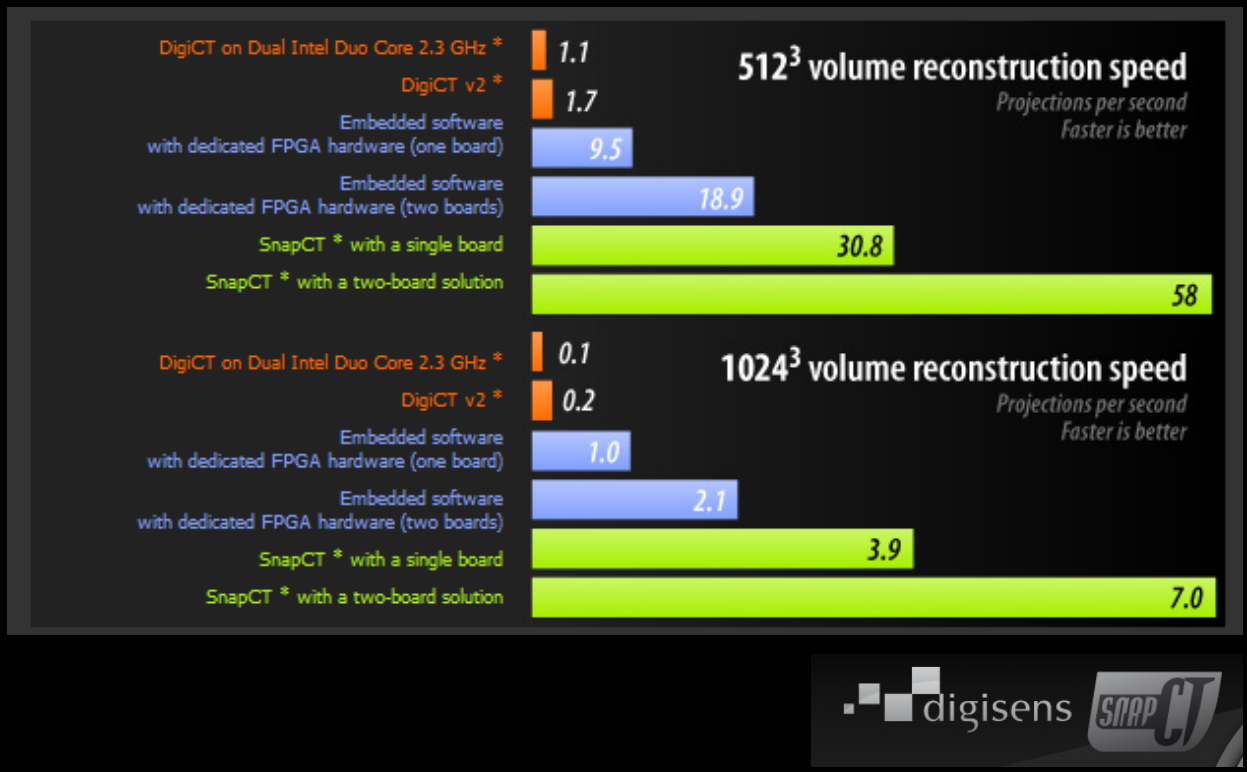
Опционы

В 70 раз быстрее

CITADEL | Technology

КОМПЬЮТЕРНАЯ ТОМОГРАФИЯ



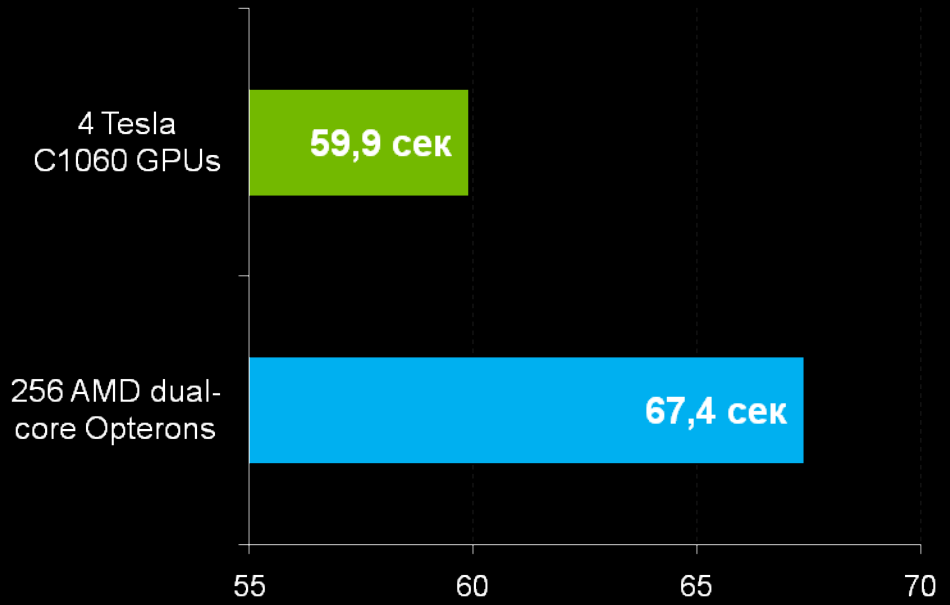


digisens SNAP CT

GPU : ПЕРЕЛОМНЫЙ МОМЕНТ В ОТРАСЛИ СУПЕРКОМПЬЮТЕРОВ

Персональный компьютер эффективнее кластера

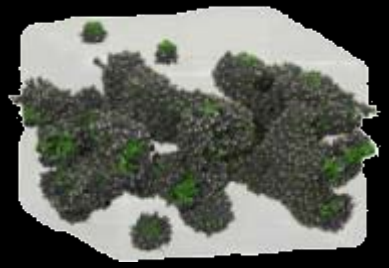
CalcUA
$5 млн.



Персональный суперкомпьютер Tesla
$10,000

Источник: University of Antwerp, Belgium

ПОИСК ЛУЧШЕГО ШАМПУНЯ



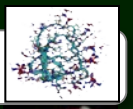
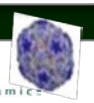
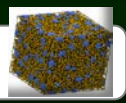
TEMPLE UNIVERSITY®

P&G

Tesla PSC	**Одинаковая производительность**	32 CPU Servers
1 kWatt	**Нет необходимости в ЦОД**	21 kWatts
$7 K	**13x ниже стоимость**	$128 K
$2 K	**19x экономия электроэнергии**	$37 K

TESLA ДЛЯ БИОМЕДИЦИНЫ

Applications

Amber 10 	GROMACS FAST. FLEXIBLE. FREE.	TeraChem 	HMMER Scalable Informatics	Hex (Docking)
NAMD Scalable Molecular Dynamics 	LAMMPS	acemd 	CUDA-BLASTP	CUDA-MEME
VMD Visual Molecular Dynamics 	GROMOS	HOOMD blue 	MUMmerGPU	CUDA-EC

Community

Download, Documentation

Technical papers

Discussion Forums

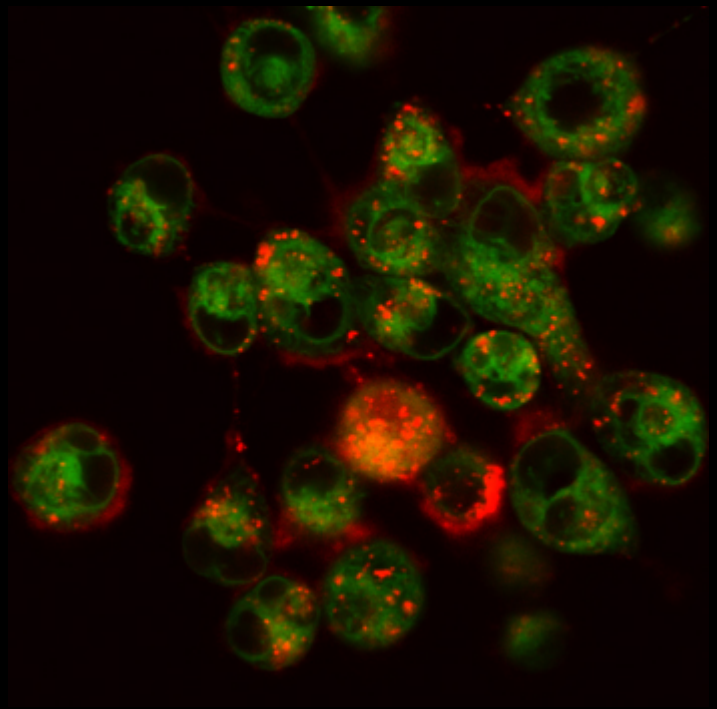
Benchmarks & Configurations

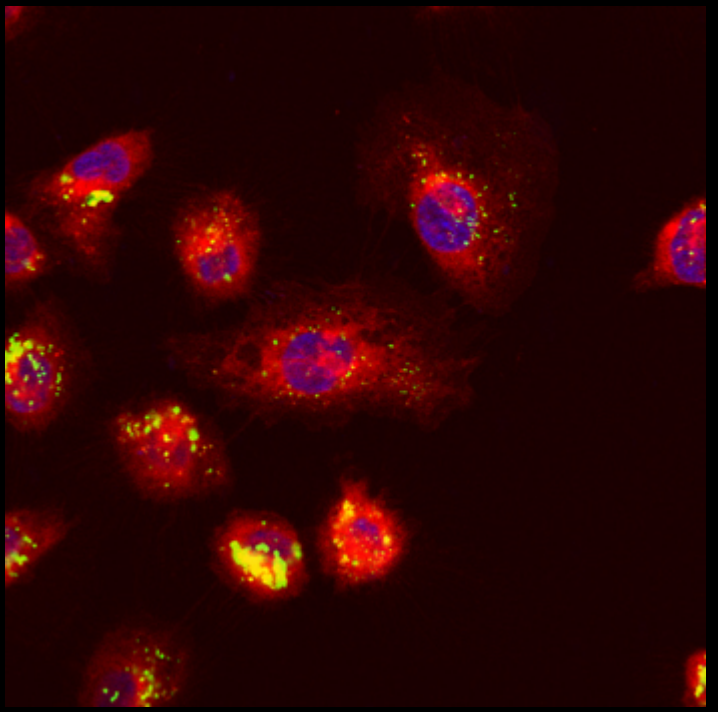
Platforms

Tesla Personal Supercomputer

Tesla GPU Clusters

МЕДИЦИНА БУДУЩЕГО УЖЕ СЕГОДНЯ



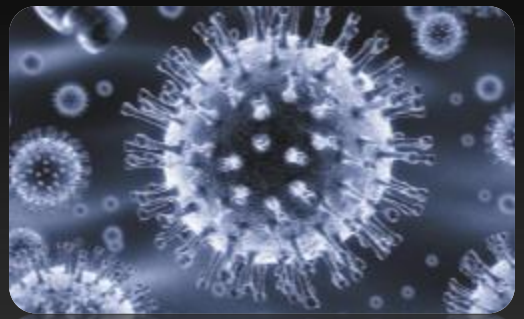


“Перенос вычислений на архитектуру графических процессоров NVIDIA CUDA дал более чем стократный прирост производительности. Среднее время получения результата уменьшилось с двух с половиной часов до 1,5 минут.”

Алексей Катичев, младший научный сотрудник института прикладной физики РАН

НАУЧНЫЕ ПРОРЫВЫ БЛАГОДАРЯ ПРИМЕНЕНИЮ GPU

Первая полная симуляция вируса H1N1



Дальнейшее понимание взаимодействия с лекарствами

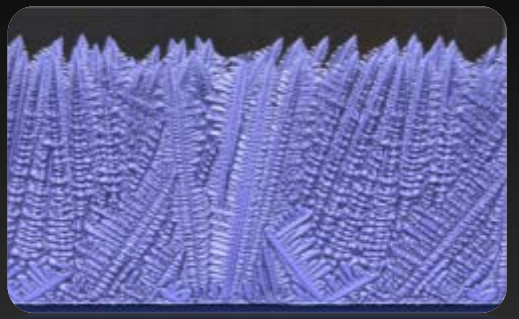
1700 GPU в Китайской академии наук

Симуляция кристаллич. кремния, солнечные батареи

Повышение эффективности элементов солнечных батарей

Производительность
1.87 Пф/сек
с 7168 GPU на Tianhe-1A

Премия Гордона Белла, материаловедение



Более прочные и легкие металлы, для снижения расхода топлива а/м

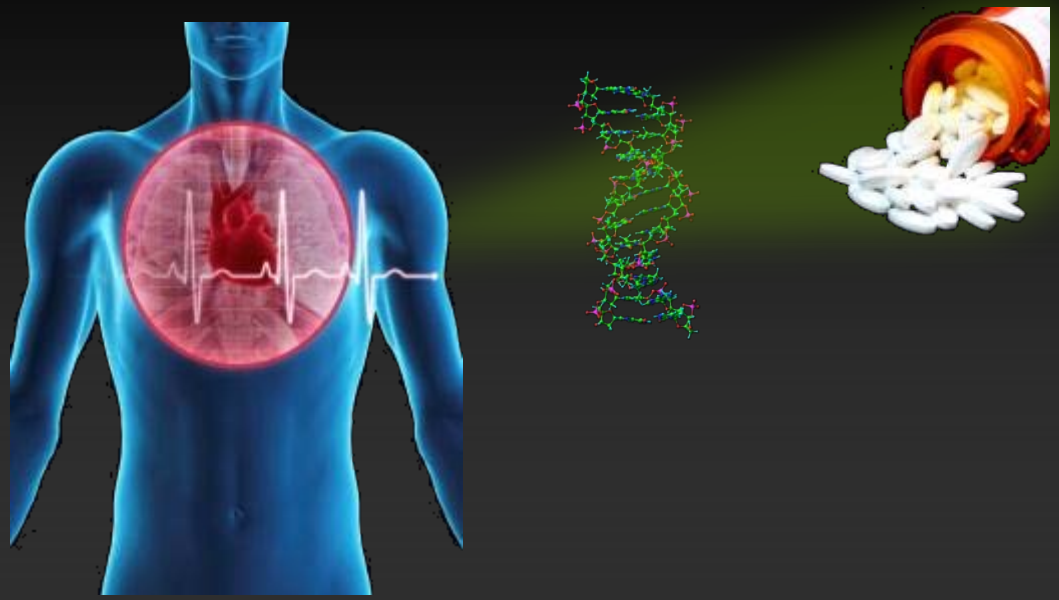
4224 GPU в Токийском технологическом институте

BGI УСКОРЯЕТ АНАЛИЗ ГЕНОМА БЛАГОДАРЯ GPU

Петабайты данных
эквивалент 15,000 генов в год

Понимание процесса лечения

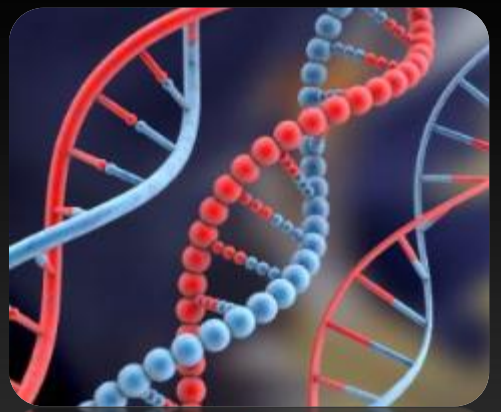
Изучение реакции индивидуумов на бактерии, вирусы, лекарства



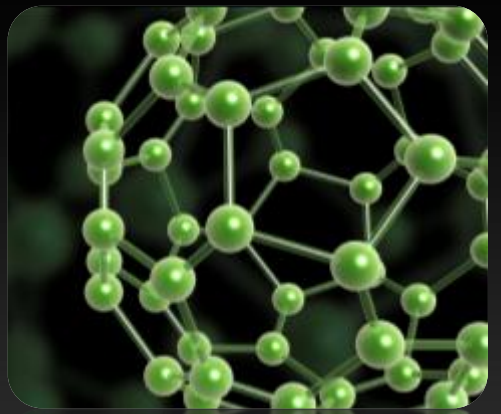
Персонализированная медицина

GPU УСКОРЯЕТ ЕСТЕСТВЕННЫЕ НАУКИ

Генетическое секвенирование



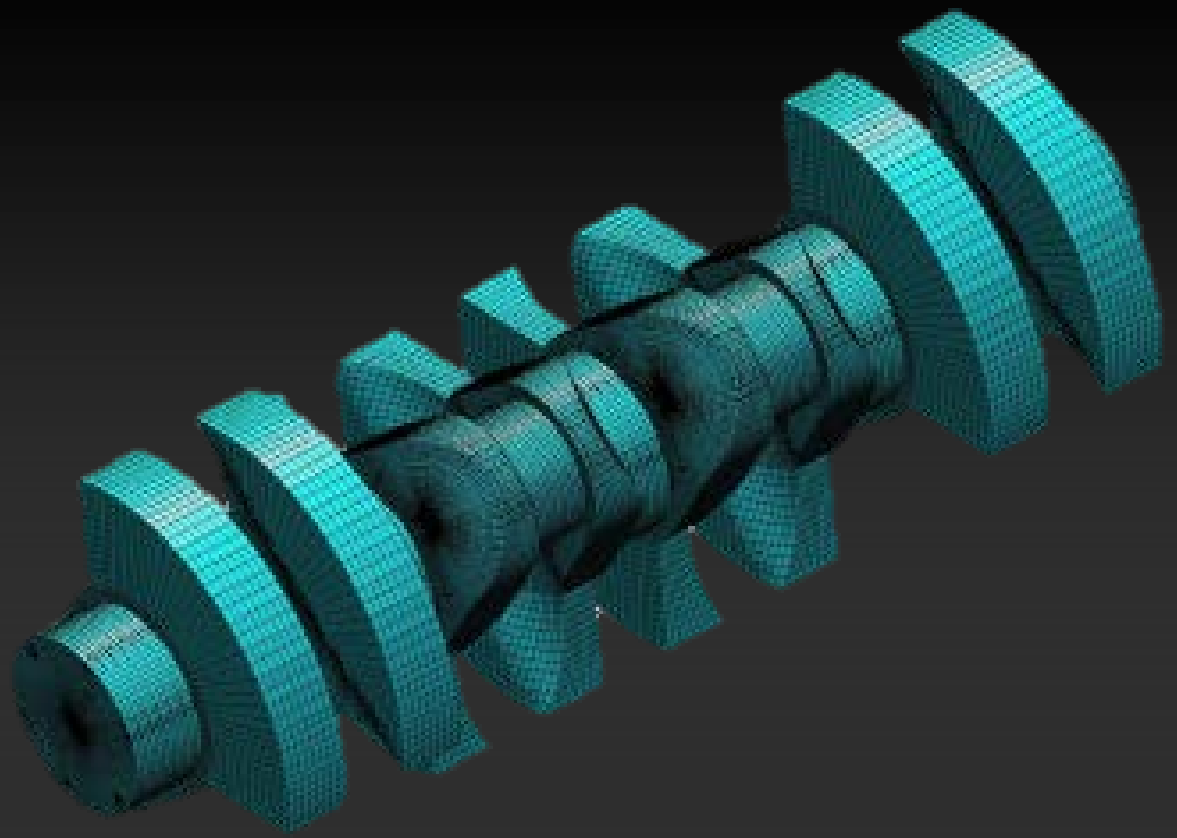
Анализ цепочек



Молекулярное моделирование

Медицинская визуализация

GPU + MSC NASTRAN = ВЫШЕ КАЧЕСТВО



Коленвал двигателя: миллионы рабочих циклов

MSC Nastran для моделирования износа и долговечности

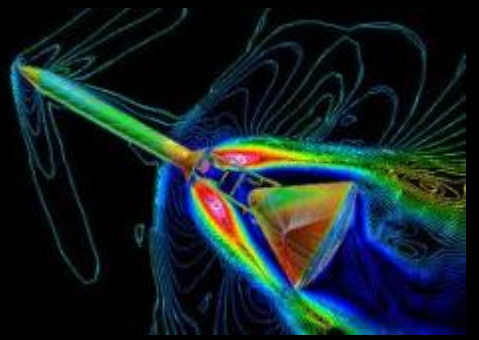
Ускорение = выше качество и надежность

Меньше отказов у клиентов/ меньше отзывов

GPU ДЛЯ ГОСУДАРСТВЕННЫХ НУЖД

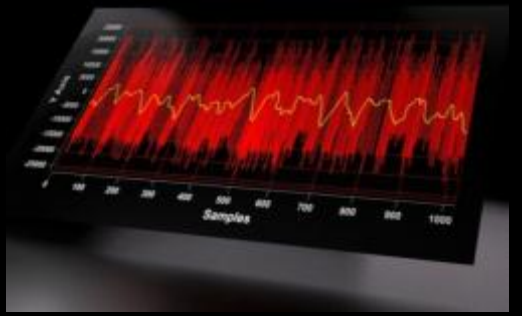
Обработка спутниковых изображений

Видео аналитика

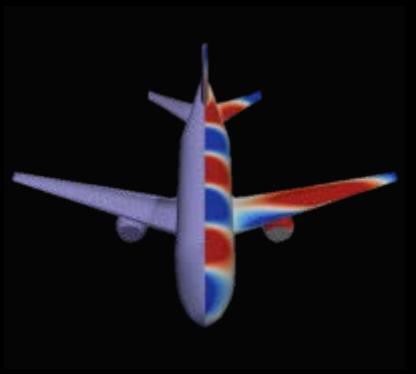


Гидро- газодинамика

Машинное зрение

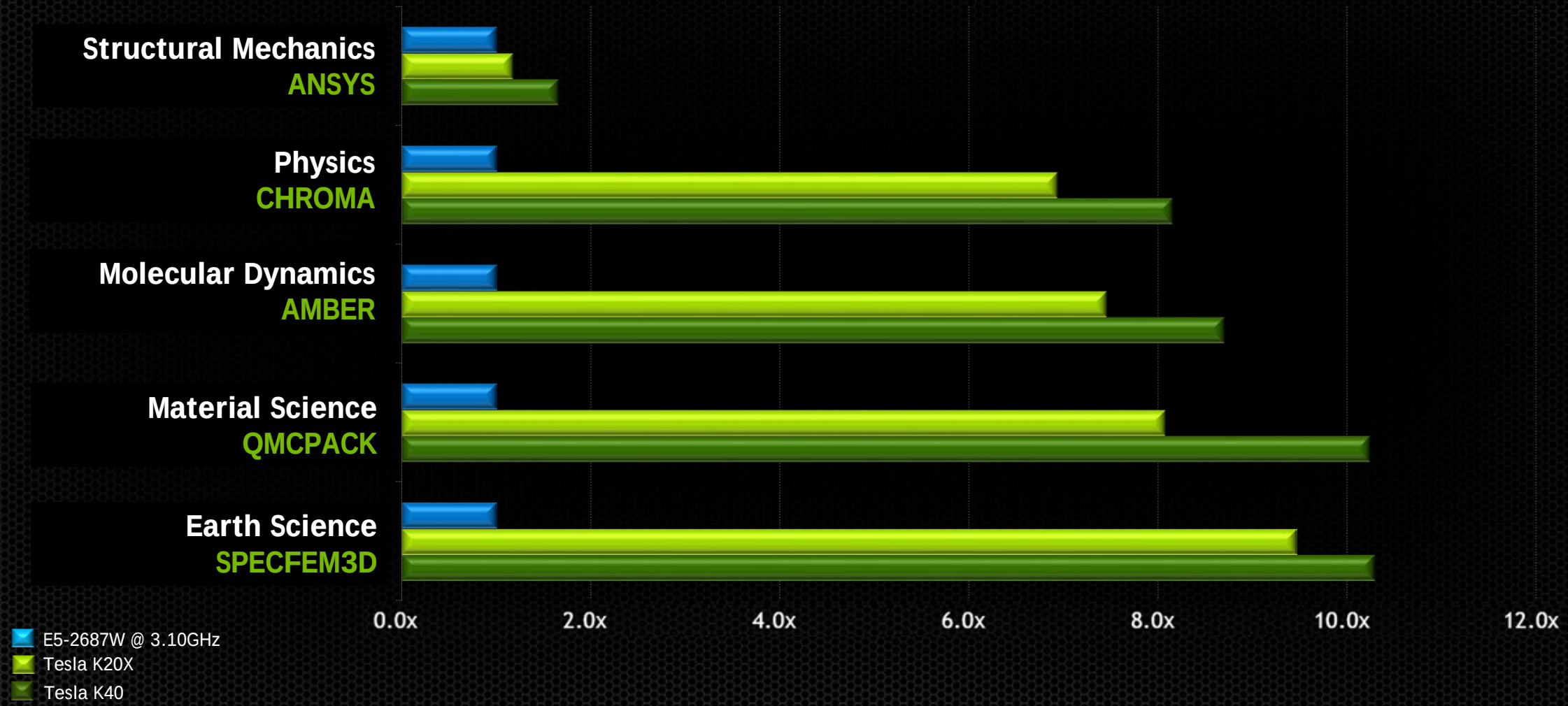


Обработка сигналов



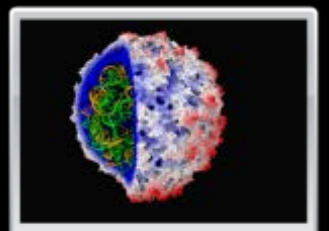
Электромагнетизм

PERFORMANCE ON LEADING SCIENTIFIC APPLICATIONS



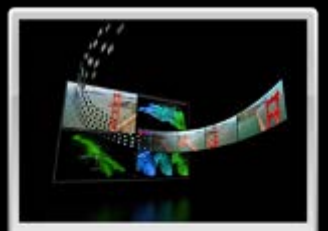
146X

Рентгенография



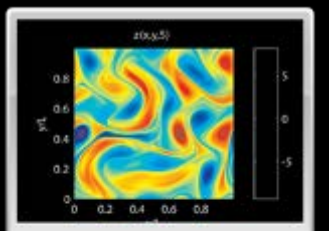
36X

Молекулярная динамика



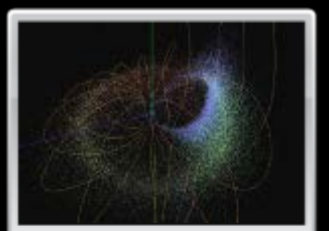
18X

Транскодирование видео



50X

Matlab



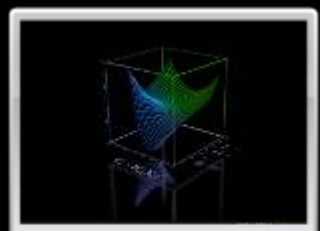
100X

Астрофизика

50x - 150x

149X

Финансы



47X

Гидро-газодинамика

20X

3D ультразвуковое сканирование

130X

Квантовая химия

30X

Биоинформатика и генетика

РОСТ ЧИСЛА ПРОФЕССИОНАЛЬНЫХ ПРИЛОЖЕНИЙ НА GPU

Tools	CUDA C/C++	PGI Accelerators	Platform LSF Cluster Mgr	TauCUDA Perf Tools	Parallel Nsight Vis Studio IDE	MATLAB	PGI CUDA x86	TotalView Debugger
	PGI CUDA Fortran	CAPS HMPP	Bright Cluster Manager	Allinea DDT Debugger	ParaTools VampirTrace	AccelerEyes Jacket MATLAB	Wolfram Mathematica	
Libraries	CUDA FFT CUDA BLAS	EMPhotonics CULAPACK	Thrust C++ Template Lib	NVIDIA NPP Perf Primitives	MAGMA (LAPACK)	NVIDIA Video Libraries	RNG & SPARSE CUDA Libraries	
Oil & Gas	Headwave Suite	OpenGeoSolutions OpenSEIS	GeoStar Seismic Suite	Acceleware RTM Solver	StoneRidge RTM		Paradigm RTM	Panorama Tech
	ffA SVI Pro	VSG Open Inventor	Seismic City RTM	Tsunami RTM			Paradigm SKUA	
Bio-Chemistry	AMBER	NAMD	HOOMD	TeraChem	BigDFT ABINT	Acellera ACEMD	DL-POLY	
	GROMACS	LAMMPS	VMD	GAMESS	CP2K			
Bio-Informatics	CUDA-BLASTP	MUMmerGPU	CUDA-MEME	PIPER Docking			OpenEye ROCS	
	CUDA SW++ SmithWaterm	GPU-HMMR	CUDA-EC	HEX Protein Docking				
CAE	ACUSIM AcuSolve 1.8	Autodesk Moldflow	Prometch Particleworks	Remcom XFdtd 7.0	ANSYS Mechanical		LSTC LS-DYNA 971	FluiDyna OpenFOAM
							Metacomp CFD++	MSC.Software Marc 2010.2

Available Announced

РОСТ ЧИСЛА ПРОФЕССИОНАЛЬНЫХ ПРИЛОЖЕНИЙ НА GPU

Архитектура графического процессора GPU

WORLD LEADER IN VISUAL COMPUTING

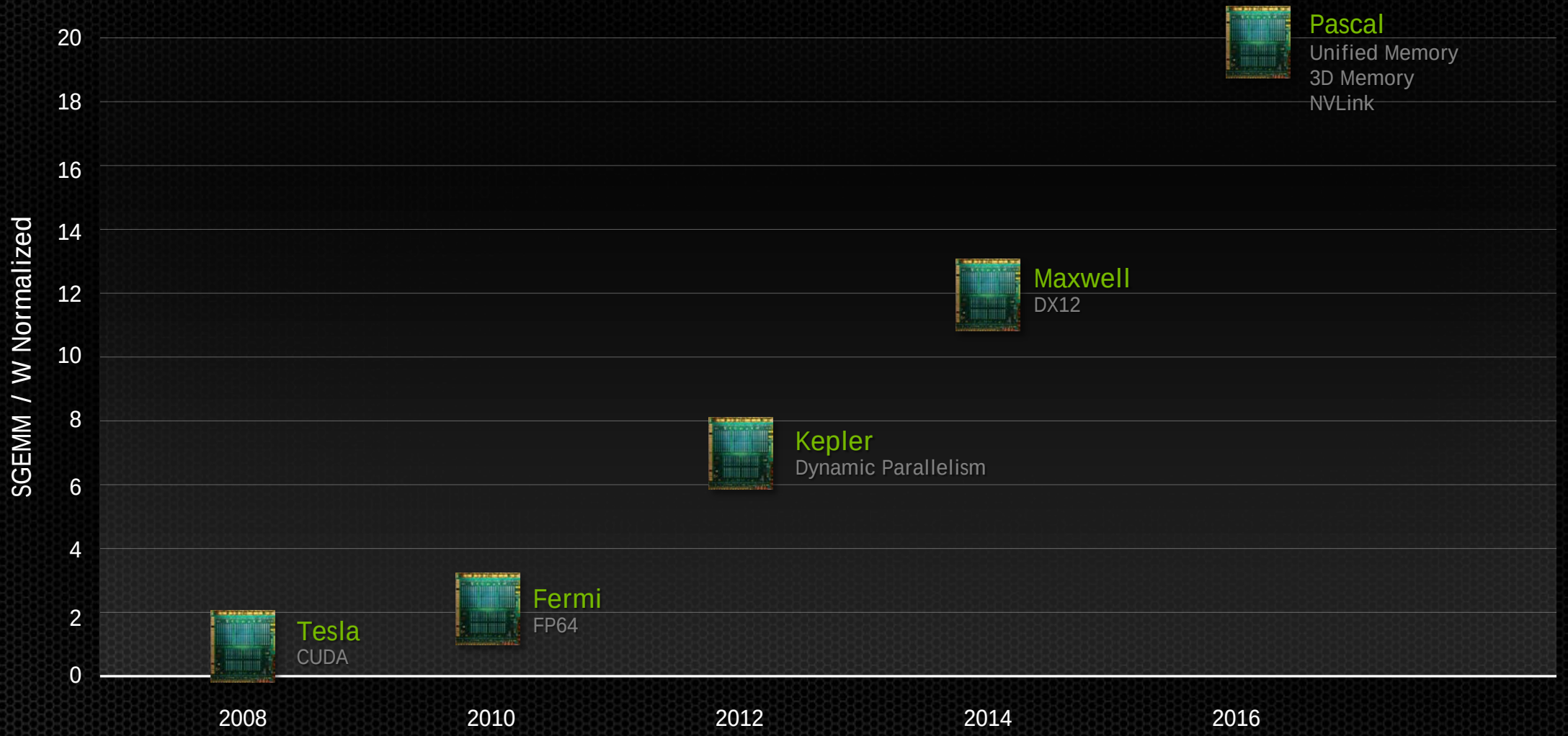
GAMING

PRO VISUALIZATION

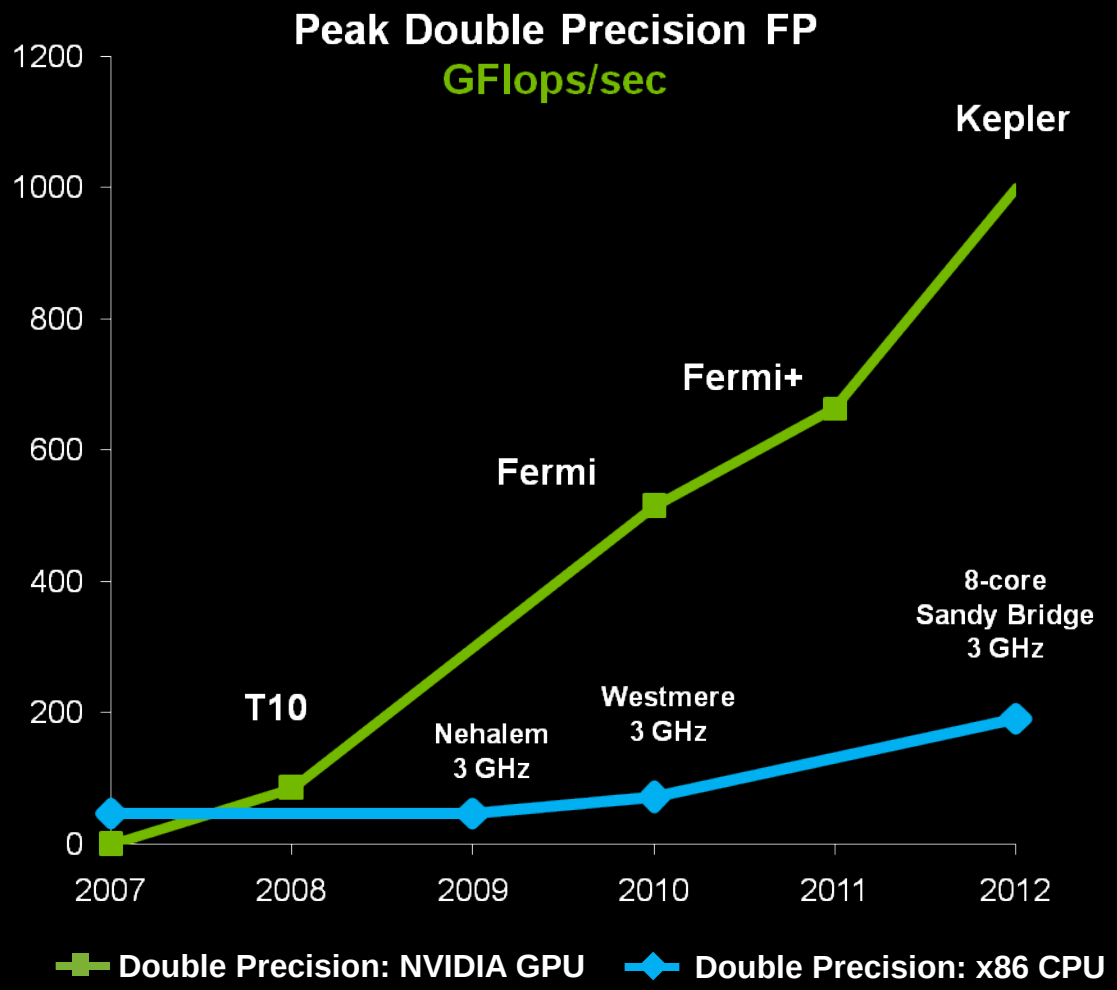
HPC & BIG DATA

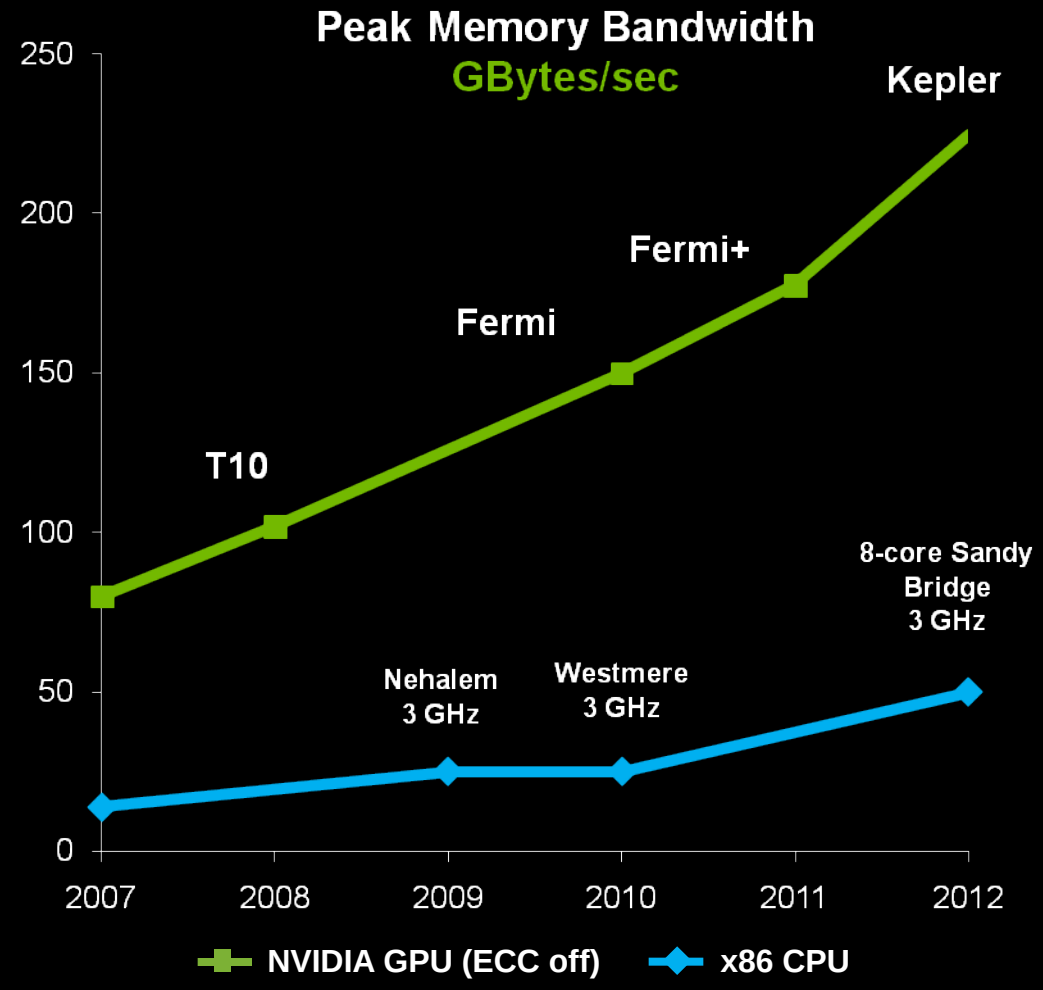
MOBILE COMPUTING

STRONG CUDA GPU ROADMAP



ЭВОЛЮЦИЯ ВЫЧИСЛИТЕЛЬНЫХ АРХИТЕКТУР





NVIDIA TESLA

Tesla C

Tesla M

Tesla S

ЭКЗАСКЕЙЛ СЕГОДНЯ С CPU

АНАЛИЗ АРХИТЕКТУР ДЛЯ ЭКЗАСКЕЙЛА

Проект Mont Blanc

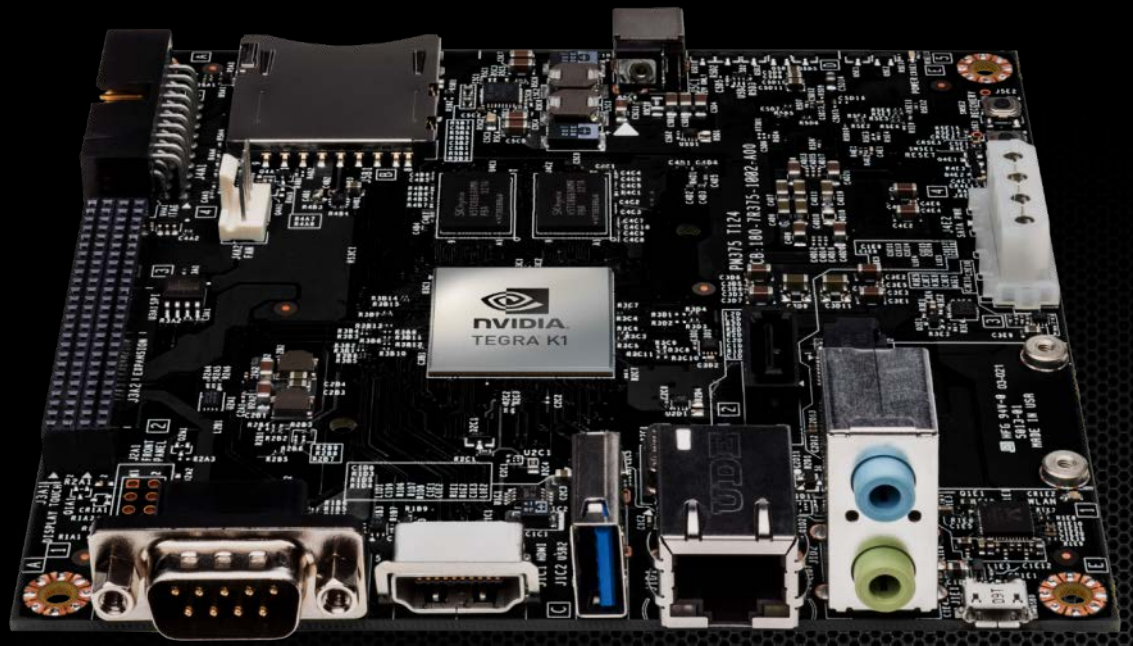
Исследование энергоэффективных архитектур для экзаскейл систем

http://www.montblanc-project.eu

http://www.eesi-project.eu/media/BarcelonaConference/Day2/13-Mont-Blanc_Overview.pdf

JETSON TK1

THE WORLD'S 1st EMBEDDED SUPERCOMPUTER



Development Platform for Embedded
Computer Vision, Robotics, Medical

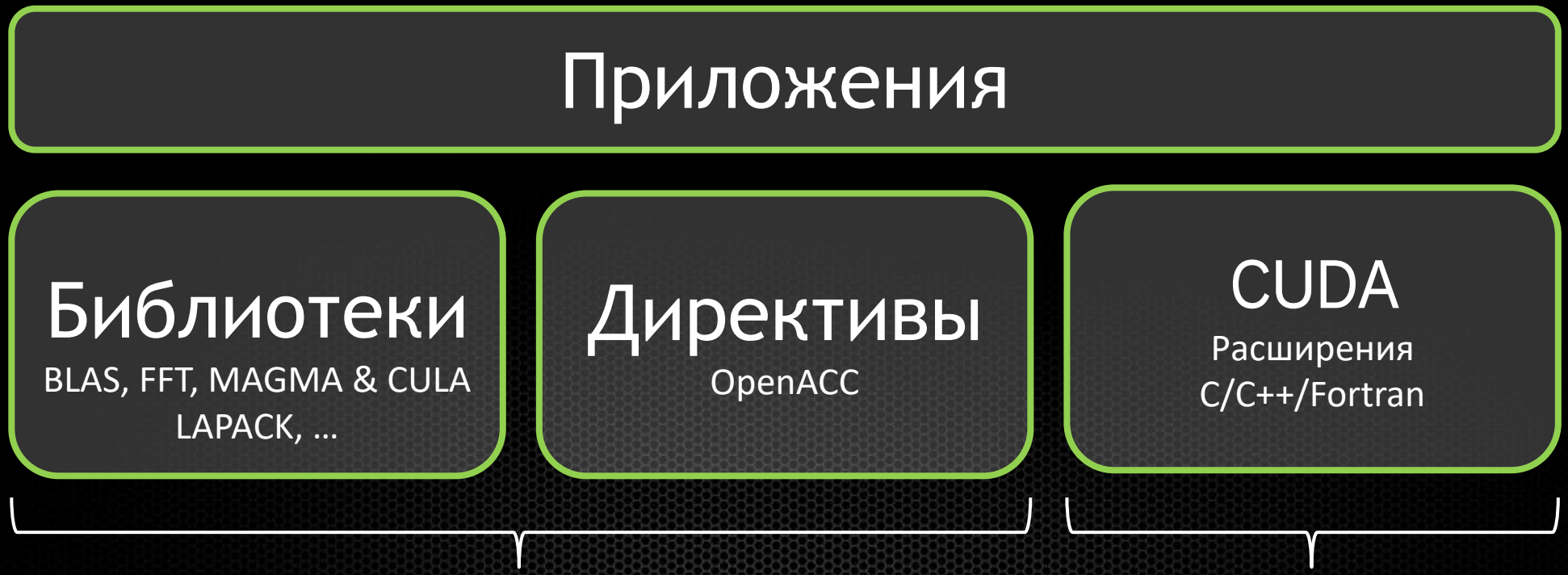
192 Cores · 326 GFLOPS

CUDA Enabled

Available Now

Программная модель CUDA

ПРОГРАММИРОВАНИЕ НА GPU



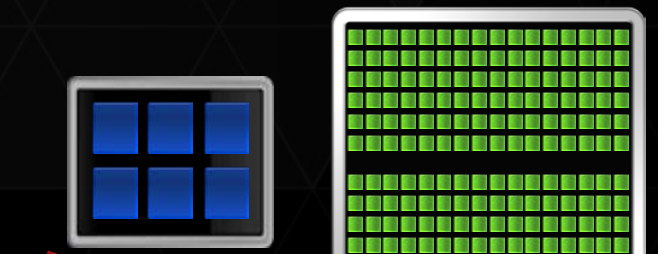
ПЕРВЫЙ ШАГ

Библиотеки

- cuFFT (Быстрое Преобразование Фурье)
- cuBLAS (библиотека линейной алгебры)
- cuRAND (генератор случайных чисел)
- cuSPARSE (работа с разряженными матрицами)
- cuDNN (нейросети, Deep Learning)
- cuSolver (поиск собственных значений)
- NPP (библиотека примитивов)
- Plugin - MATLAB, Mathematica

ВТОРОЙ ШАГ

OpenACC



```
main() {
  ...
  <serial code>
  ...
  #pragma acc kernels
  {
  <compute intensive code>
  }
  ...
}
```

Выделение части кода, выполняемого на GPU

- Открытый стандарт
- Простота
- Использование на GPUs

ТРЕТИЙ ШАГ

Compute Unified Device Architecture

GPU Computing Applications

Libraries and Middleware

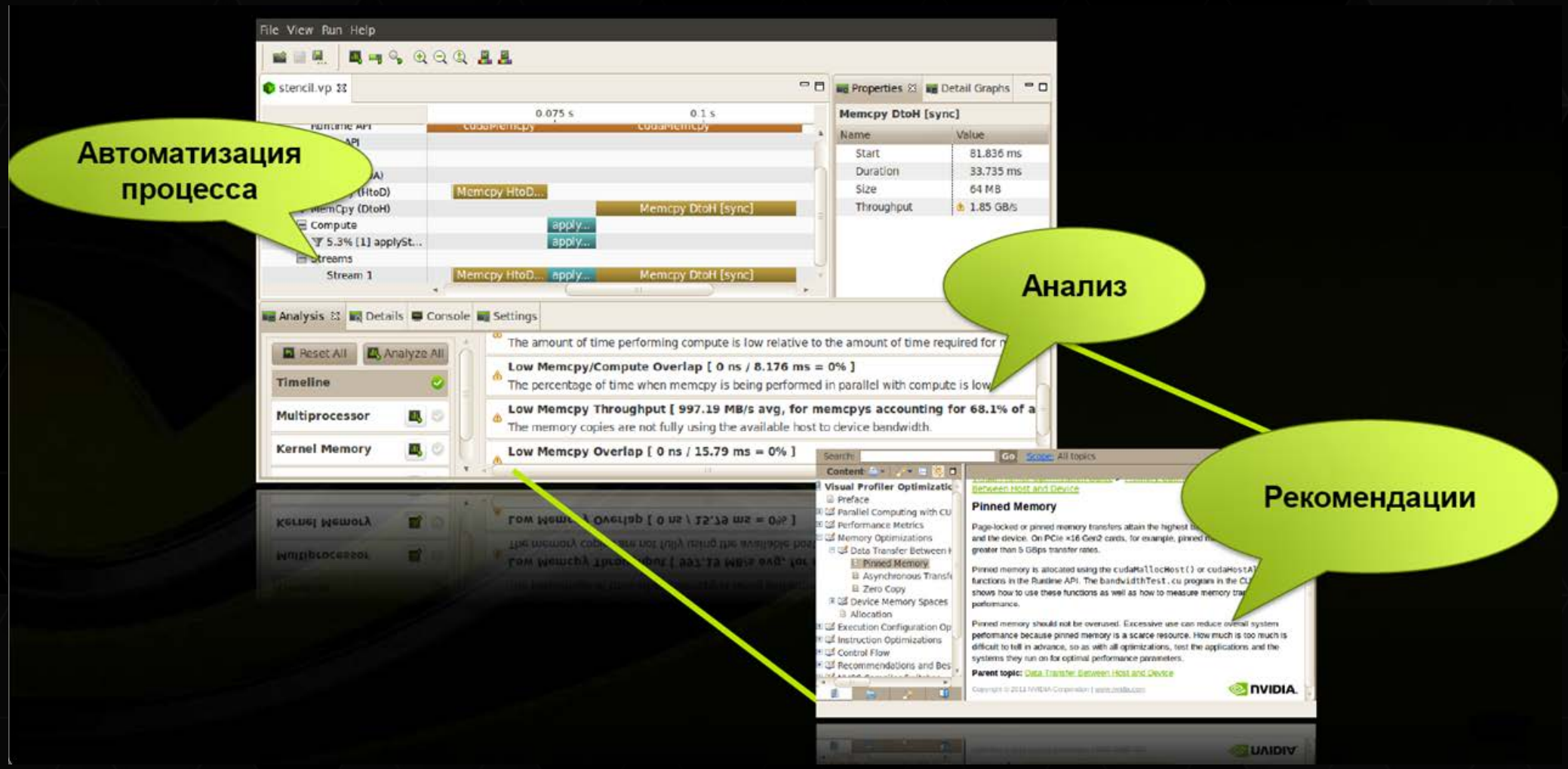
cuFFT cuBLAS cuRAND cuSPARSE	CULA MAGMA	Thrust NPP	VSIPL SVM OpenCurrent	PhysX OptiX iRay	cuDNN TensorRT	MATLAB Mathematica

Programming Languages

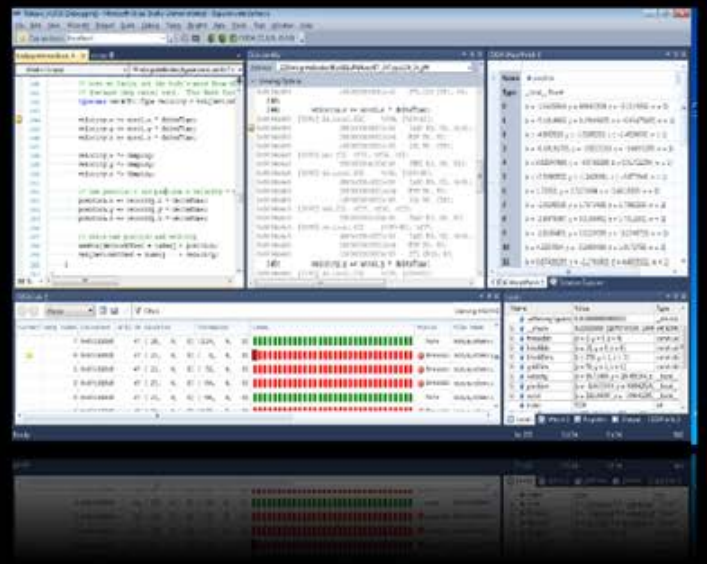
C	C++	Fortran	DirectCompute	Java Python Wrappers	Directives (e.g. OpenACC)

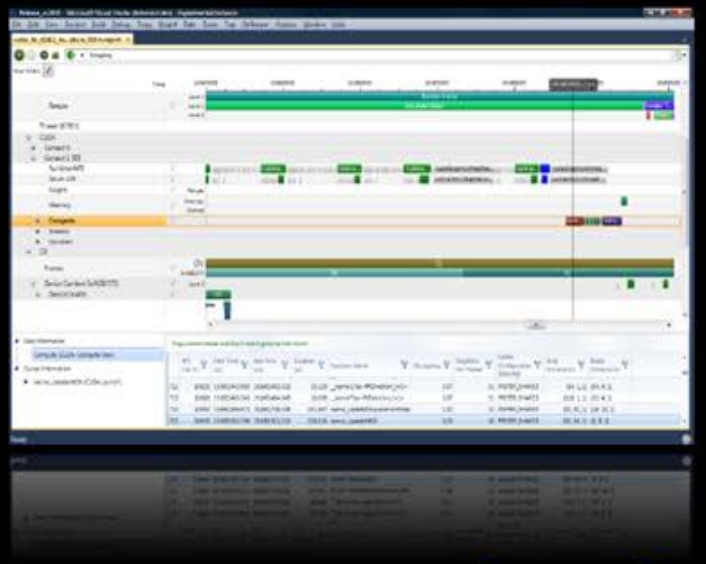
NVIDIA GPU
CUDA Parallel Computing Architecture

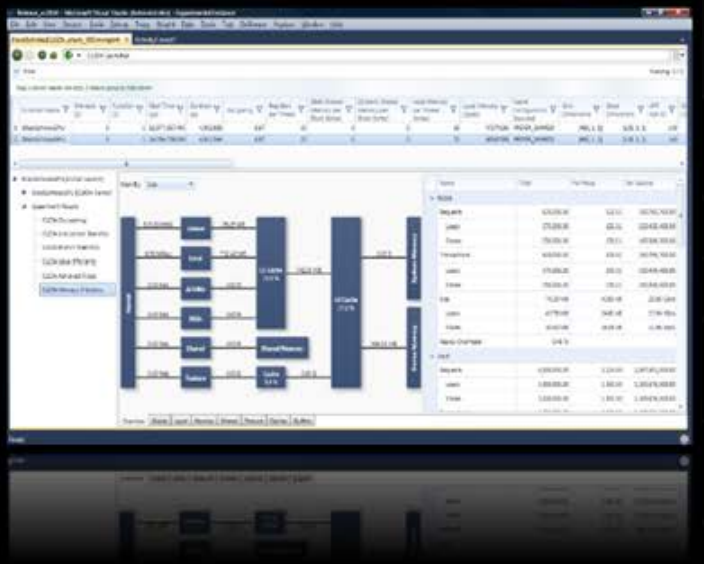
VISUAL PROFILER



NVIDIA PARALLEL NSIGHT







CUDA Debugger

- Debug CUDA kernels directly on GPU hardware
- Examine thousands of threads executing in parallel
- Use on-target conditional breakpoints to locate errors

CUDA Memory Checker

- Enables precise error detection

System Trace

- Review CUDA activities across CPU and GPU
- Perform deep kernel analysis to detect factors limiting maximum performance

CUDA Profiler

- Advanced experiments to measure memory utilization, instruction throughput and stalls

Обучение программированию на GPU

CUDA: WORLD'S MOST PERVASIVE PARALLEL PROGRAMMING MODEL

14,000 Institutions with CUDA Developers

2,000,000 CUDA Downloads

487,000,000 CUDA GPUs Shipped

700+ University Courses
In 62 Countries

КУРС ПО CUDA В 200 УНИВЕРСИТЕТАХ КИТАЯ

“Модель параллельного программирования CUDA позволяет нам учить будущих инженеров и исследователей создавать инновации за счет использования мощи современных параллельных процессоров.”

Профессор Steve Deng
Университет Цинхуа

CUDA CENTER OF EXCELLENCE

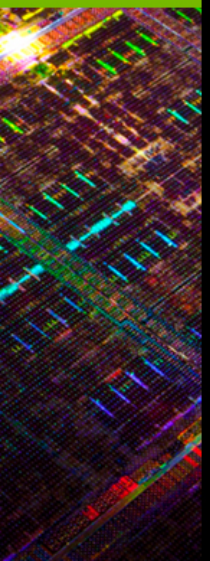
Геномика
Медицинская визуализация
Радарные технологии

ГИС
Гидро- газодинамика
Дизайн аналоговых схем

ЛИТЕРАТУРА ПО GPU

GPU TECHNOLOGY CONFERENCE

ТЕХНОЛОГИЧЕСКАЯ КОНФЕРЕНЦИЯ ПО GPU



Одно из мероприятий, которое нельзя пропустить

- Передовые достижения в области вычислений на GPU
- Новые научные и коммерческие приложения
- Лучшие умы в области параллельных вычислений
- Самые инновационные продукты и решения

Способы участия

- Докладчик – презентация полученных результатов
- Посетитель – общение с экспертами и коллегами из вашей предметной области
- Участник/стенд – реклама вашей организации, как ключевого игрока в экосистеме GPU

www.gputechconf.com