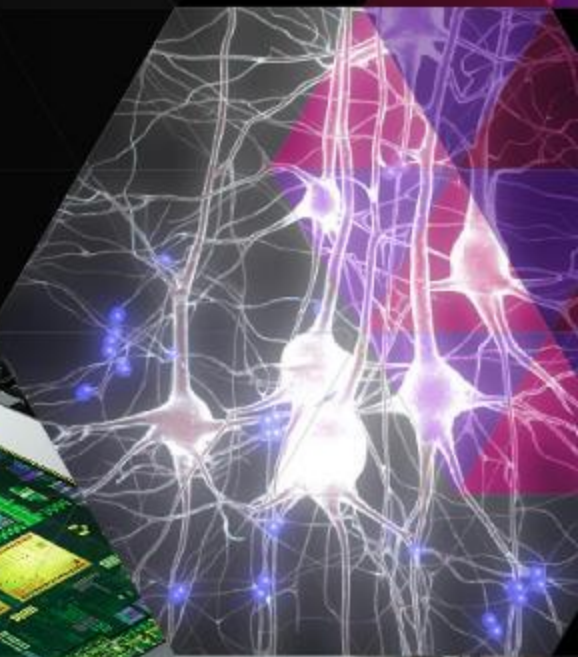




ВЫЧИСЛЕНИЯ НА GPU ОТ МОБИЛЬНЫХ УСТРОЙСТВ ДО СУПЕРКОМПЬЮТЕРОВ

Дмитрий Конягин

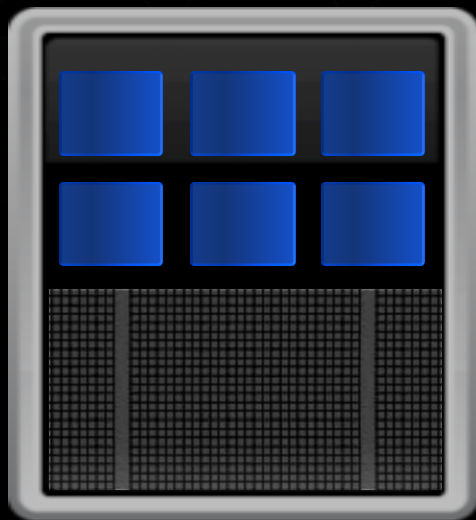


ГИБРИДНЫЕ ВЫЧИСЛЕНИЯ

ПРАВИЛЬНЫЙ ПРОЦЕССОР ДЛЯ ПРАВИЛЬНЫХ ЗАДАЧ

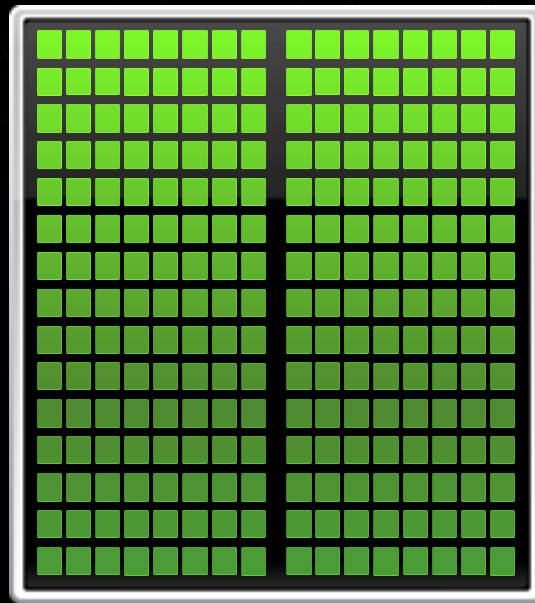
CPU

Оптимизирован для
последовательных задач



GPU ускоритель

Оптимизирован для
параллельных задач





TESLA® GPUS ЛЕЖАТ В
ОСНОВЕ 85%
ГИБРИДНЫХ HPC СИСТЕМ

ВЕДУЩИЕ МИРОВЫЕ
КОМПАНИИ
ИСПОЛЬЗУЮТ TESLA ДЛЯ
АНАЛИЗА БОЛЬШИХ
ДАННЫХ, ОБРАБОТКИ
ИЗОБРАЖЕНИЙ И
МАШИННОГО ОБУЧЕНИЯ

ПЕРВЫЙ В МИРЕ СЕРВЕР НА БАЗЕ POWER8 С GPU

IBM POWER S824L



- ▶ 2x POWER8 CPU
 - ▶ До 1TB памяти
 - ▶ Пропускная способность памяти до 384 GB/s
- ▶ 2 ускорителя Tesla K40
- ▶ Linux

Доступно с 31 октября



The Green500 List

Listed below are the June 2014 The Green500's energy-efficient supercomputers ranked from 1 to 100.

	Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
	1	4,389.82	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	34.58
	2	3,631.70	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
	3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
	4	3,459.46	SURFsara	Cartesius Accelerator Island - Bullx B515 cluster, Intel Xeon E5-2450v2 8C 2.5GHz, InfiniBand 4x FDR, Nvidia K40m	44.40
	5	3,185.91	Swiss National Supercomputing Centre (SCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect, NVIDIA K20x Level 3 measurement data available	1,753.66
	6	3,131.06	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
	7	3,019.72	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C 2GHz, Infiniband FDR, Nvidia K20m	86.20
	8	2,951.95	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.93GHz, Infiniband QDR, NVIDIA K20x	927.86
	9	2,813.14	Exploration & Production - Eni S.p.A.	HPC2 - iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, NVIDIA K20x	1,067.49
	10	2,678.41	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	54.60
	11	2,629.42	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband FDR, NVIDIA K20x	66.25
	12	2,629.42	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband FDR, NVIDIA K20x	66.25
	13	2,629.42	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband FDR, NVIDIA K20x	66.25
	14	2,629.42	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband FDR, NVIDIA K20x	66.25
	15	2,629.10	Max-Planck-Gesellschaft MPVIPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	269.94

ОСНОВНАЯ ТРУДНОСТЬ ДЛЯ СВЕРХБОЛЬШИХ HPC СИСТЕМ:

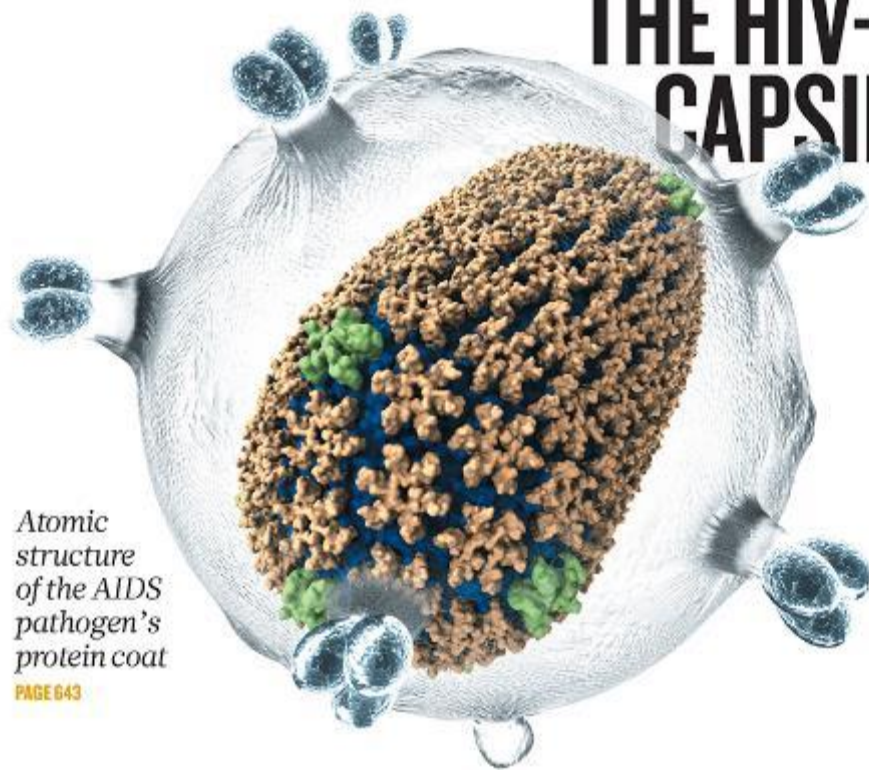
ЭНЕРГОЭФФЕКТИВНОСТЬ

GPU СОЗДАНЫ ДЛЯ ЭНЕРГОЭФФЕКТИВНЫХ ПАРАЛЛЕЛЬНЫХ ВЫЧИСЛЕНИЙ.

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

THE HIV-1 CAPSID



Atomic
structure
of the AIDS
pathogen's
protein coat

PAGE 643

**УСКОРЕНИЕ ИССЛЕДОВАНИЙ
ИСПОЛЬЗУЯ 3000 ПРОЦЕССОРОВ TESLA**

УЧЕНЫЕ ИЗ УНИВЕРСИТЕТА
ИЛЛИНОЙСА ОСУЩЕСТВИЛИ
ПЕРВУЮ ПОЛНУЮ СИМУЛЯЦИЮ
АТОМА ВИЧ И ОТКРЫЛИ СТРУКТУРУ
ЕГО ПРОТЕИНОВОЙ ОБОЛОЧКИ



POPULAR GPU-ACCELERATED APPLICATIONS

02 Research: Higher Education and Supercomputing

COMPUTATIONAL CHEMISTRY AND BIOLOGY

NUMERICAL ANALYSIS

PHYSICS

WEATHER AND CLIMATE FORECASTING

06 Defense and Intelligence

07 Computational Finance

08 Manufacturing: CAD and CAE

COMPUTER AIDED DESIGN

COMPUTATIONAL FLUID DYNAMICS

COMPUTATIONAL STRUCTURAL MECHANICS

ELECTRONIC DESIGN AUTOMATION

10 Media and Entertainment

ANIMATION, MODELING AND RENDERING

COLOR CORRECTION AND GRAIN MANAGEMENT

COMPOSITING, FINISHING AND EFFECTS

EDITING

ENCODING AND DIGITAL DISTRIBUTION

ON-SET GRAPHICS

ON-SET REVIEW AND STEREO TOOLS

SIMULATION

WEATHER GROUPS

14 Oil and Gas

Research: Higher Education and Supercomputing

COMPUTATIONAL CHEMISTRY AND BIOLOGY

Bioinformatics

Software	Description	License/Model	Performance	Architecture	GPU Accelerated	Availability
BerryGDB	Sequence mapping software	alignment of short sequencing reads	4-18s	T.2075, 2090, 410, K20, K20K	Yes	Available now Version 5.4.2
CUSAM+ve	Open source software for Smith-Waterman protein database searches on GPUs	Parallel search of Smith-Waterman database	10-50s	T.2075, 2090, 410, K20, K20K	Yes	Available now Version 2.0.8
CUSHAM	Parallelized short read aligner	Parallel, accurate long read aligner - gapped alignments to large genomes	10s	T.2075, 2090, 410, K20, K20K	Yes	Available now Version 1.0.40
GPU-BLAST	Local search with fast k-tuple heuristic	Protein alignment according to BLAST, multi-cpu threads	3-4s	T.2075, 2090, 410, K20, K20K	Single only	Available now Version 2.2.34
GPU-HMMER	Parallelized local and global search with profile Hidden Markov models	Parallel local and global search of Hidden Markov Models	60-100s	T.2075, 2090, 410, K20, K20K	Yes	Available now Version 2.2.2
mDASA-MEME	Ultra-fast scalable motif discovery algorithm based on MEME	Scalable motif discovery algorithm based on MEME	4-11s	T.2075, 2090, 410, K20, K20K	Yes	Available now Version 3.0.12
SeqTried	A GPU Accelerated Sequence Analysis Toolkit	Reference assembly, GCAT, protein-expression, terms, de novo assembly	400s	T.2075, 2090, 410, K20, K20K	Yes	Available now
USEN	Open-source Smith-Waterman for SGE/CUDA. Suffix array based repeats filter and output	Fast short read alignment	9-3s	T.2075, 2090, 410, K20, K20K	Yes	Available now Version 1.11
WishLM	Fits thousands linear models to a fixed design and response	Parallel linear regression on multiple similarly-shaped matrices	150s	T.2075, 2090, 410, K20, K20K	Yes	Available now Version 0.1-1

Molecular Dynamics

Software	Description	License/Model	Performance	Architecture	GPU Accelerated	Availability
Alamo	Models molecular dynamics of biopolymers for simulations of proteins, DNA and ligands	Simulations on 1000 GPUs	4-7hr	T.2075, 2090, 410, K20, K20K	Single Only	Available now Version 1.0.40
ACEMD	GPU simulation of molecular mechanics force fields, implicit and explicit solvent	Written for use on GPUs	140 ns/day GPU version only	T.2075, 2090, 410, K20, K20K	Yes	Available now
AMBER	Suite of programs to simulate molecular dynamics on biomolecules	PMEMD: explicit and implicit solvent	89-44 ns/day JAC NVE	T.2075, 2090, 410, K20, K20K	Yes	Available now Version 12 + toulg8
DL-POLY	Simulate macromolecules, polymers, ions systems, etc on a distributed memory parallel computer	Two-body forces, Link-cell pairs, Ewald SPME forces, Stokes W	4s	T.2075, 2090, 410, K20, K20K	Yes	Available now Version 0.1 Tensor only
CHARMM	MD package to simulate molecular dynamics on biomolecules	Implicit Sol, Explicit Sol, Solvent via OpenMM	780s	T.2075, 2090, 410, K20, K20K	Yes	in Development 04/12
BRIMACS	Simulation of biochemical molecules with complicated bond interactions	Implicit Sol, Explicit Sol, solvent	145 ns/day D46F9	T.2075, 2090, 410, K20, K20K	Single only	Available now Version 6.5 in 04/12
HOOMD-blue	Particle dynamics package written primarily for GPUs	Written for GPUs	3s	T.2075, 2090, 410, K20, K20K	Yes	Available now
LAMMPS	Classical molecular dynamics package	Lennard-Jones, Morse, Buckingham, CHARMM, Tabulated, Coarse grain, SW, Anisotropic Dipole, RE-squared, Hybrid combination	3-18s	T.2075, 2090, 410, K20, K20K	Yes	Available now
MD	Designed for high-performance simulation of large molecular systems	100M atom-capable	4-44 ns/day 570W 98% 3000s	T.2075, 2090, 410, K20, K20K	Yes	Available now Version 2.9
OpenMM	Library and application for molecular dynamics for HPC with GPUs	Implicit and explicit solvent, custom forces	127-203 ns/day	T.2075, 2090, 410, K20, K20K	Yes	Available now Version 6.1.1

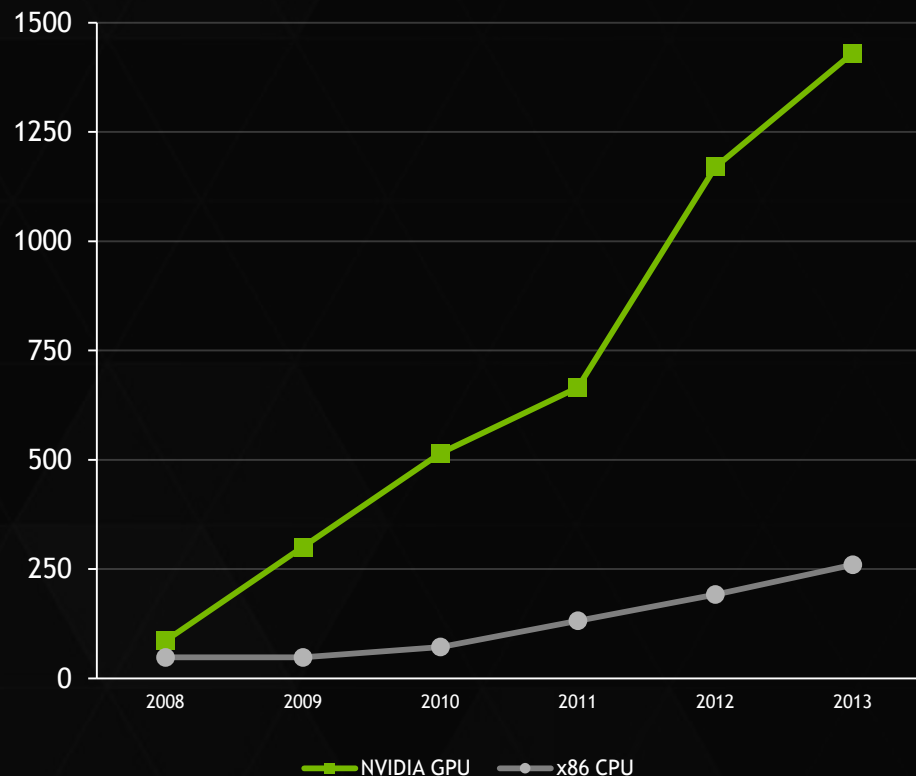
275+ приложений с GPU-ускорением
www.nvidia.com/appscatalog



РАЗРЫВ ПРОДОЛЖАЕТ УВЕЛИЧИВАТЬСЯ

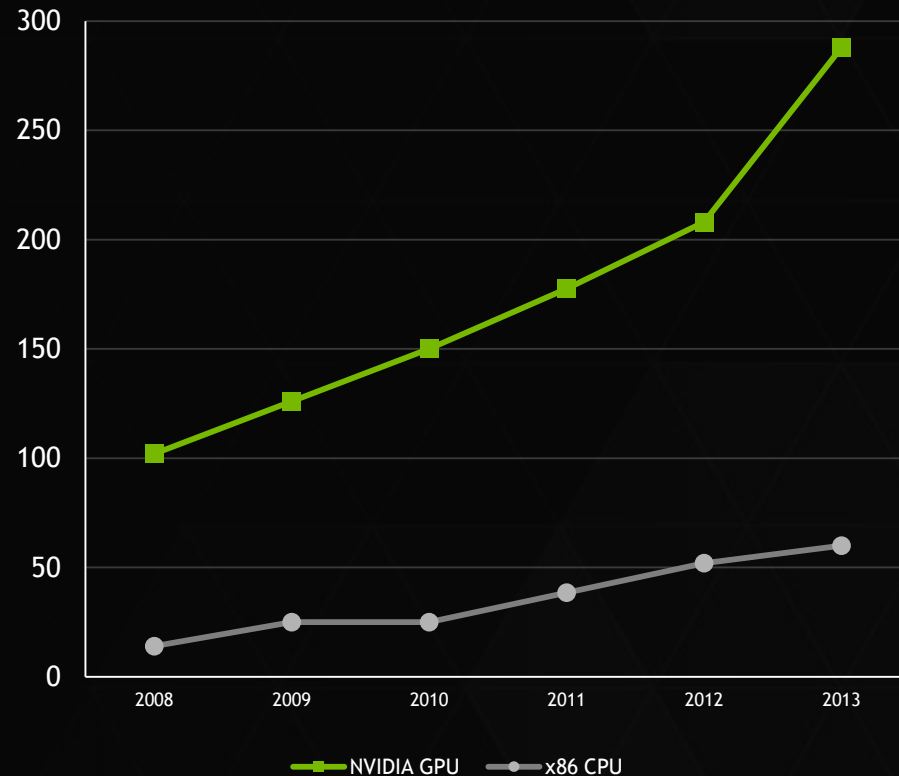
Пиковая производительность DP
FLOPS

GFLOPS



Пиковая пропускная способность
памяти

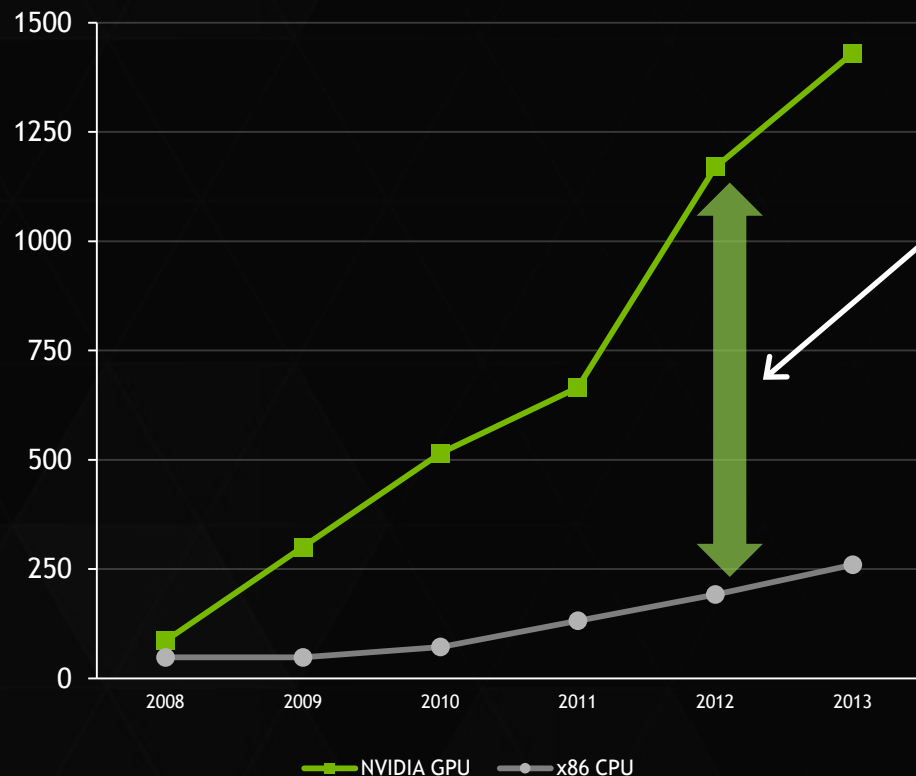
GB/Sec



РАЗРЫВ ПРОДОЛЖАЕТ УВЕЛИЧИВАТЬСЯ

Пиковая производительность DP FLOPS

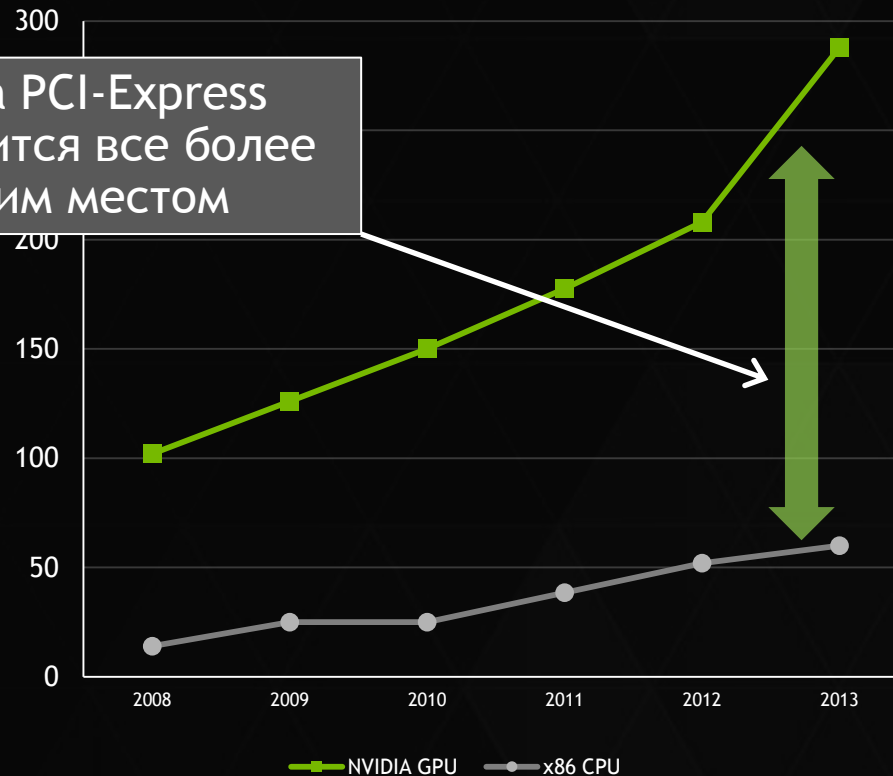
GFLOPS



Шина PCI-Express становится все более узким местом

Пиковая пропускная способность памяти

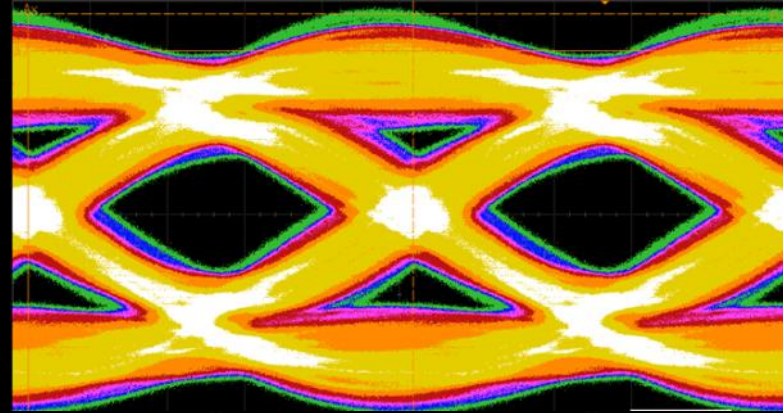
GB/Sec



NVLINK И СТЕКОВАЯ 3D ПАМЯТЬ

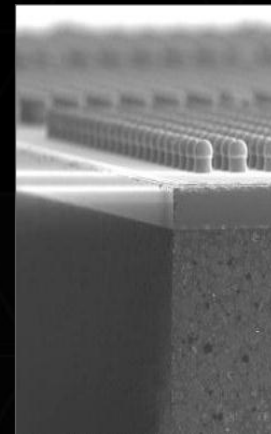
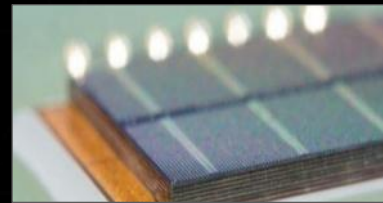
NVLINK

- Высокоскоростной интерфейс между GPU и GPU/CPU
- В 5-12 раз выше пропускная способность, чем у PCI-E Gen3
- Поддержка в POWER CPU



СТЕКОВАЯ 3D ПАМЯТЬ

- В 4 раза выше пропускная способность (~1 TB/s)
- В 3 раза больший объем
- В 4 раза выше энергоэффективность



ЗА
ПРЕДЕЛАМИ
НРС

*GPU в области аналитики
Больших Данных,
машинного обучения и
промышленных вычислений*

ОТ НРС К АНАЛИТИКЕ БОЛЬШИХ ДАННЫХ

GTC 2009



SIEMENS

Raytheon

GTC 2014

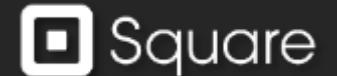


facebook

Google



NETFLIX



YAHOO!

Yandex


Список некоторых участников ежегодной конференции NVIDIA по графическим процессорам

ДОСТУПНОСТЬ НЕЙРОННЫХ СЕТЕЙ БЛАГОДАРЯ GPU

“*Now You Can Build Google’s
\$1M Artificial Brain on the Cheap*”

-Wired


GOOGLE BRAIN



1,000 CPU Servers
2,000 CPUs • 16,000 cores

600 kWatts
\$5,000,000

STANFORD AI LAB

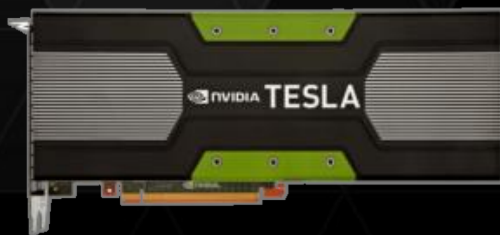


3 GPU-Accelerated Servers
12 GPUs • 18,432 cores

4 kWatts
\$33,000 NVIDIA

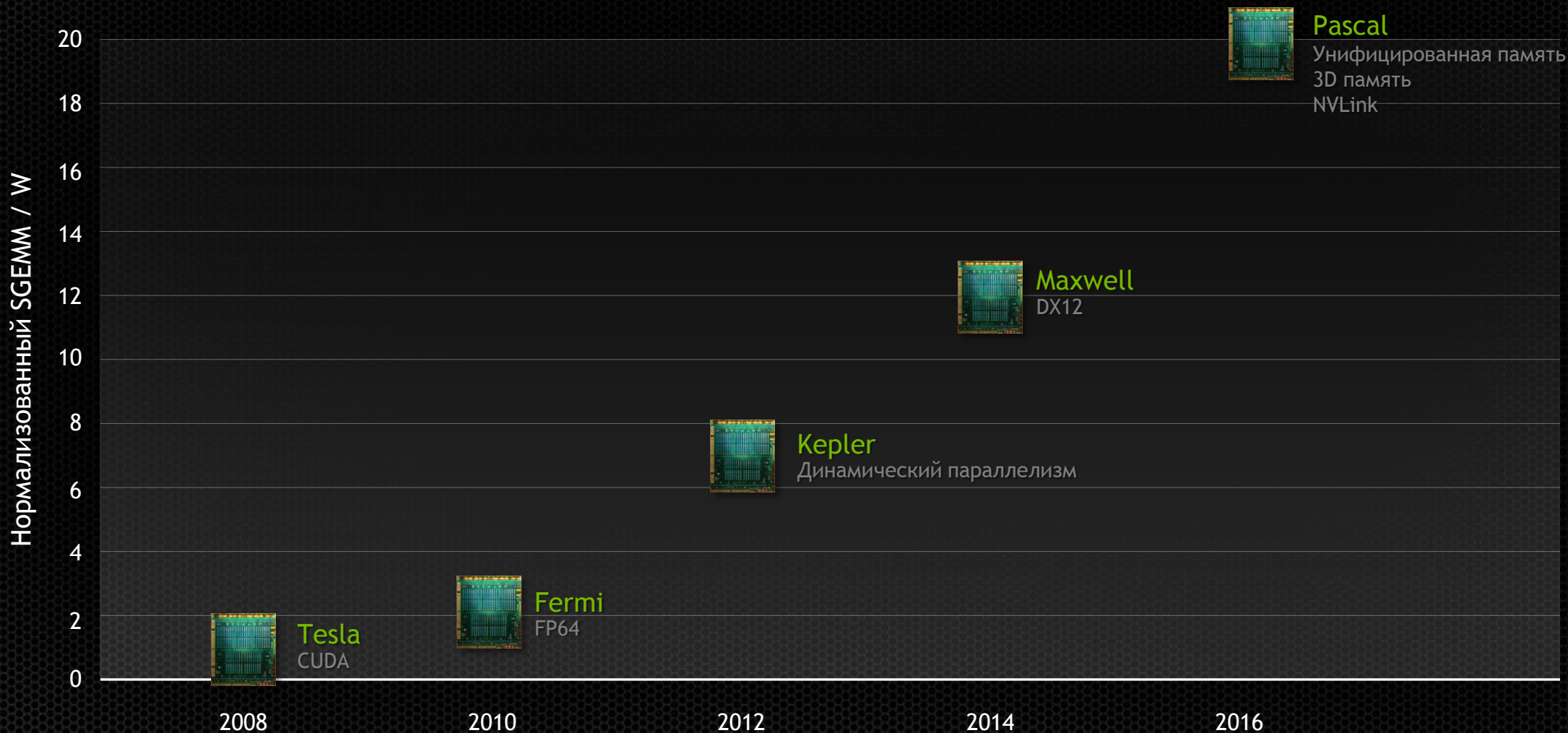
ПРОДУКТОВАЯ ЛИНЕЙКА TESLA

САМЫЕ БЫСТРЫЕ И ЭНЕРГОЭФФЕКТИВНЫЕ УСКОРИТЕЛИ

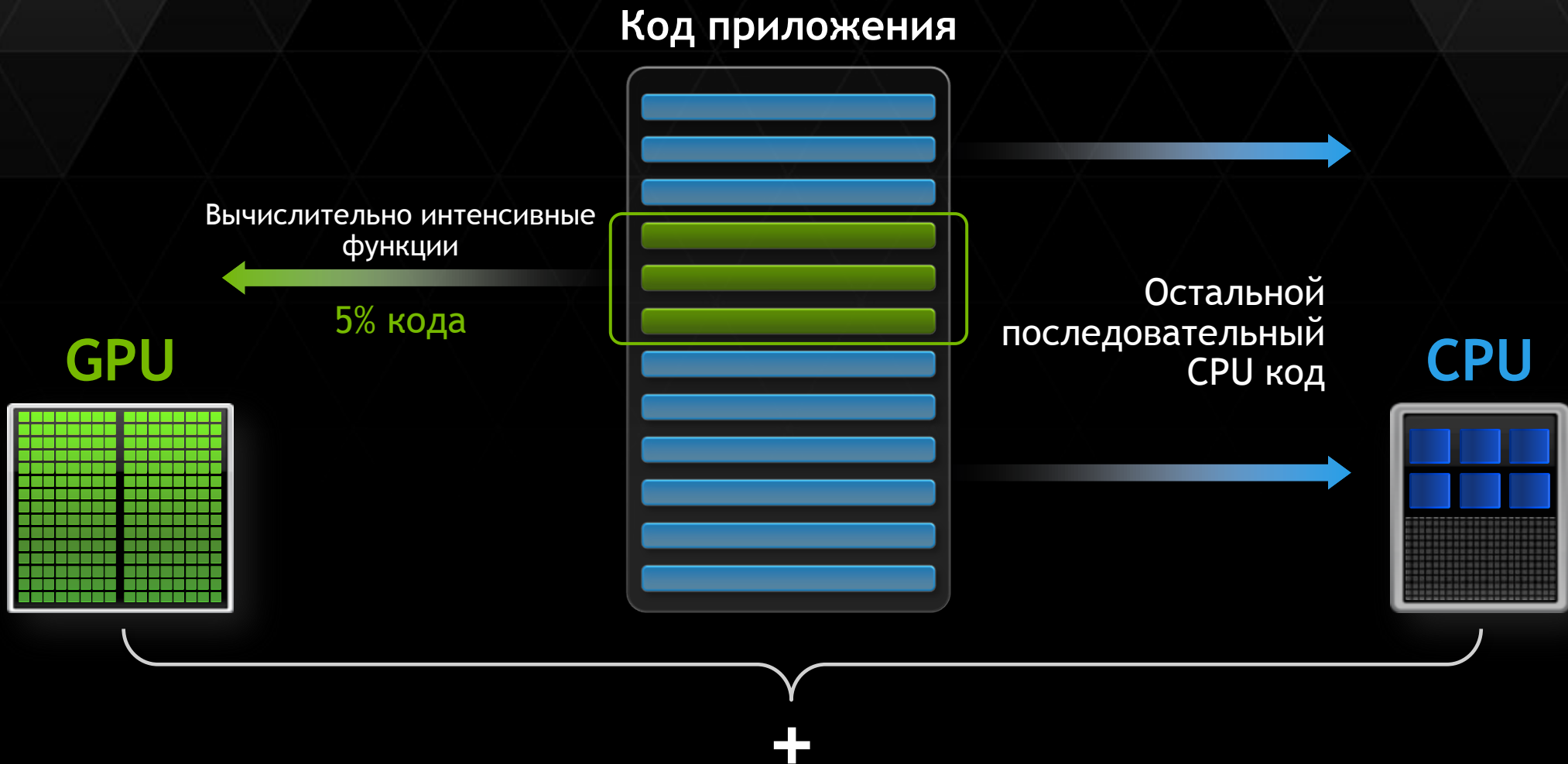


	GPUs	Single Precision Peak (<i>SGEMM</i>)	Double Precision Peak (<i>DGEMM</i>)	Memory Size	Memory Bandwidth (ECC off)	PCIe Gen	System Solution
CFD, BioChemistry, Neural Networks, High Energy Physics, Graph analytics, Material Science, BioInformatics, M&E	K40	4.29 TF (3.22TF)	1.43 TF (1.33 TF)	12 GB	288 GB/s	Gen 3	Server + Workstation
Weather & Climate, Physics, BioChemistry, CAE, Material Science	K20X	3.95 TF (2.90 TF)	1.32 TF (1.22 TF)	6 GB	250 GB/s	Gen 2	Server only
	K20	3.52 TF (2.61 TF)	1.17 TF (1.10 TF)	5 GB	208 GB/s	Gen 2	Server + Workstation
Image, Signal, Video, Seismic	K10	4.58 TF	0.19 TF	8 GB	320 GB/s	Gen 3	Server only

УВЕРЕННОЕ РАЗВИТИЕ CUDA GPU



КАК РАБОТАЮТ ВЫЧИСЛЕНИЯ НА GPU



3 СПОСОБА ПРОГРАММИРОВАТЬ НА GPU

Приложения

Библиотеки

BLAS, FFT, MAGMA & CULA
LAPACK, ...

“copy-paste”
ускорение

Директивы

OpenACC

Простое ускорение
приложение

CUDA

Расширения
C/C++/Fortran

Максимальная
производительность

БИБЛИОТЕКИ С GPU УСКОРЕНИЕМ

“COPY-PASTE” УСКОРЕНИЕ ДЛЯ ВАШИХ ПРИЛОЖЕНИЙ

CUDART	CUDA Runtime Library
cuFFT	Быстрые преобразования Фурье
cuBLAS	Полноценная библиотека основных операции линейной алгебры
cuSPARSE	Библиотека для работы с разреженными матрицами
cuRAND	Библиотека генерации случайных чисел
NPP	Примитивы для обработки изображений и видео
Thrust	Шаблоны алгоритмов и структур данных для C++
math.h	C99 библиотека для работы с числами с плавающей точкой
cuDNN	Библиотека для реализации глубоких нейронных сетей

Входит в состав CUDA Toolkit (бесплатно): developer.nvidia.com/cuda-toolkit

Информация по библиотекам на CUDA : developer.nvidia.com/gpu-accelerated-libraries

ПРОГРАММА ДЛЯ CUDA РАЗРАБОТЧИКОВ

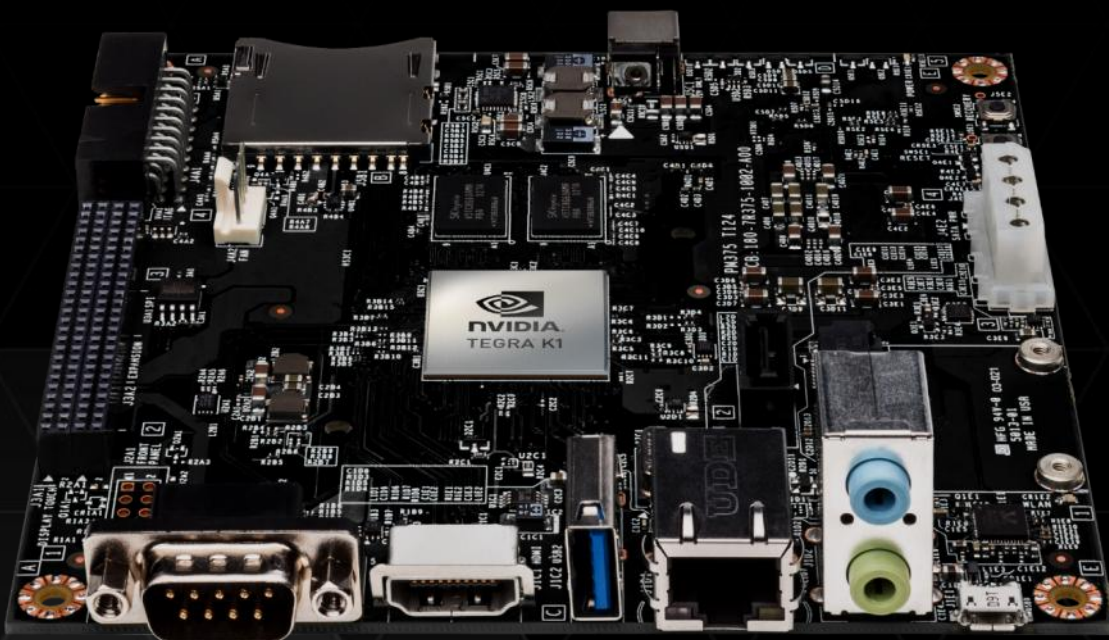
Бесплатная регистрация: www.nvidia.com/paralleldeveloper

- ▶ Exclusive access to pre-release CUDA Installers
- ▶ Submit bugs and features requests to NVIDIA
- ▶ Keep informed about latest releases and training opportunities
- ▶ Access to exclusive downloads
- ▶ Exclusive activities and special offers



JETSON TK1

ПЕРВЫЙ В МИРЕ ВСТРАЕВАЕМЫЙ СУПЕРКОМПЬЮТЕР



Платформа для разработчиков в
области компьютерного зрения,
робототехники и медицины

192 ядра · 326 GFLOPS
CUDA Enabled

Доступен сейчас

РОСТ ВЫЧИСЛЕНИЙ НА GPU

100M
GPU на базе CUDA



522M
CUDA-Capable GPUs



150K
Скачиваний CUDA



2.5M
Скачиваний CUDA



1
Суперкомпьютер



44
Суперкомпьютера



60
Университетских
курсов



770
Университетских
курсов



4,000
Научных работ



57,000
Научных работ



2008

2014



**Вычисления на GPU
от мобильных устройств до суперкомпьютеров**