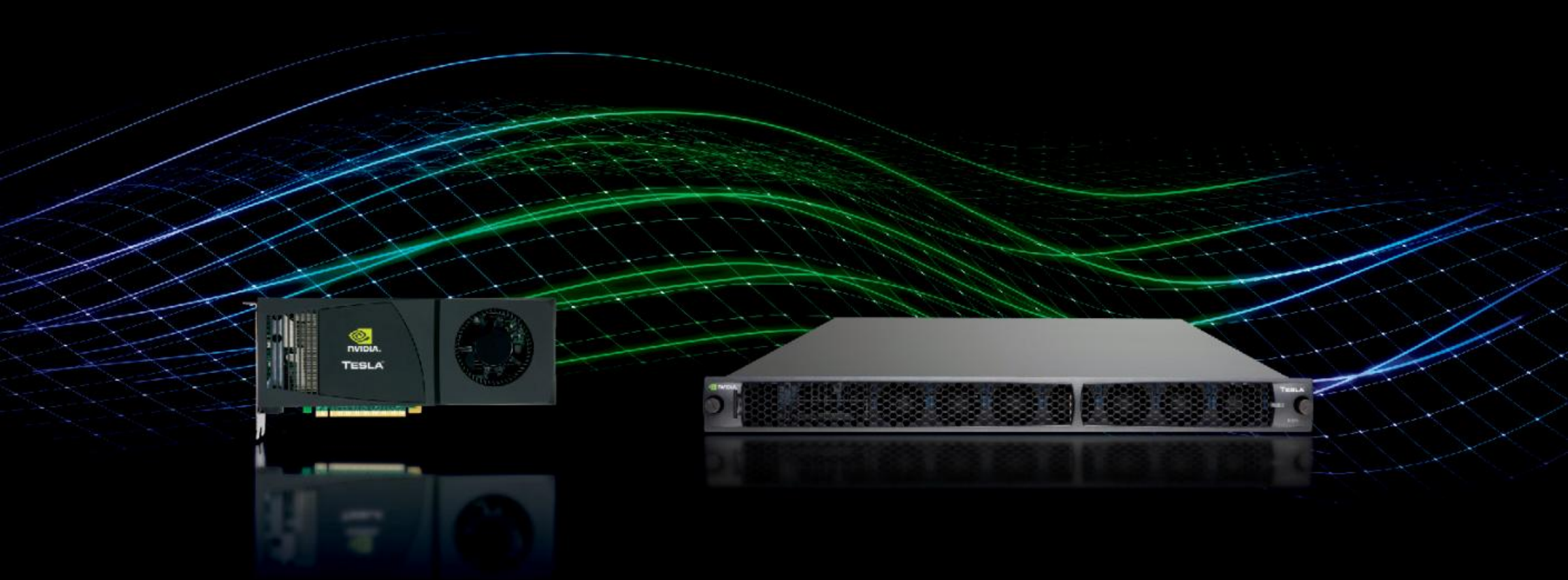
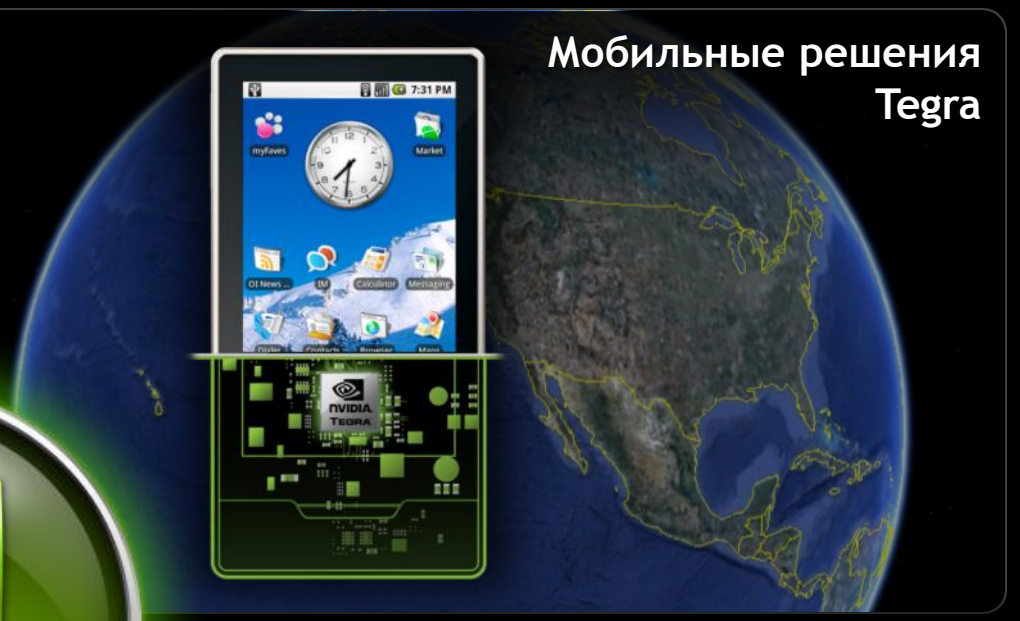


Решения для высокопроизводительных вычислений от NVIDIA

Февраль 2010

Игры
GeForce

Мобильные решения
Tegra

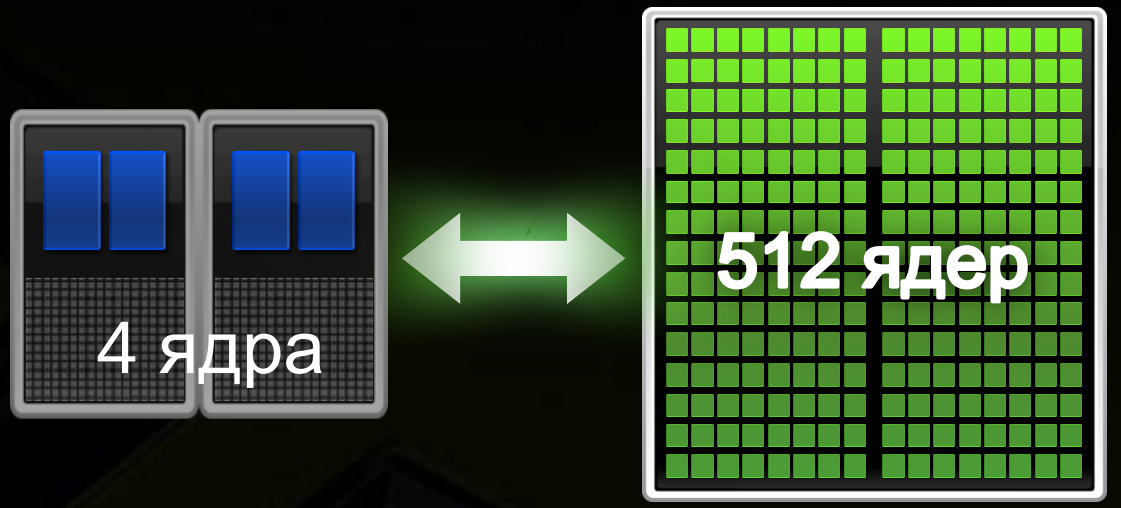


NVIDIA

Профессиональная визуализация & DCC
Quadro

HPC
Tesla

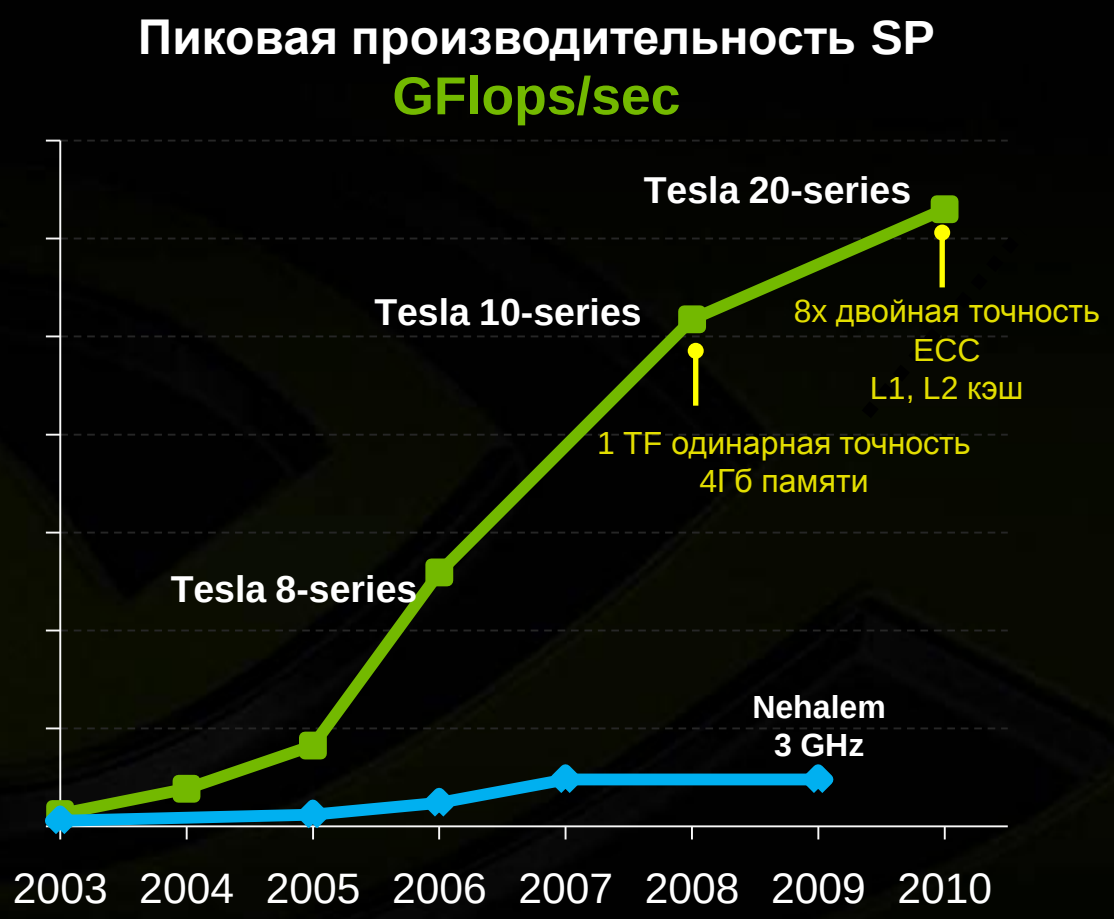
Гетерогенные вычисления

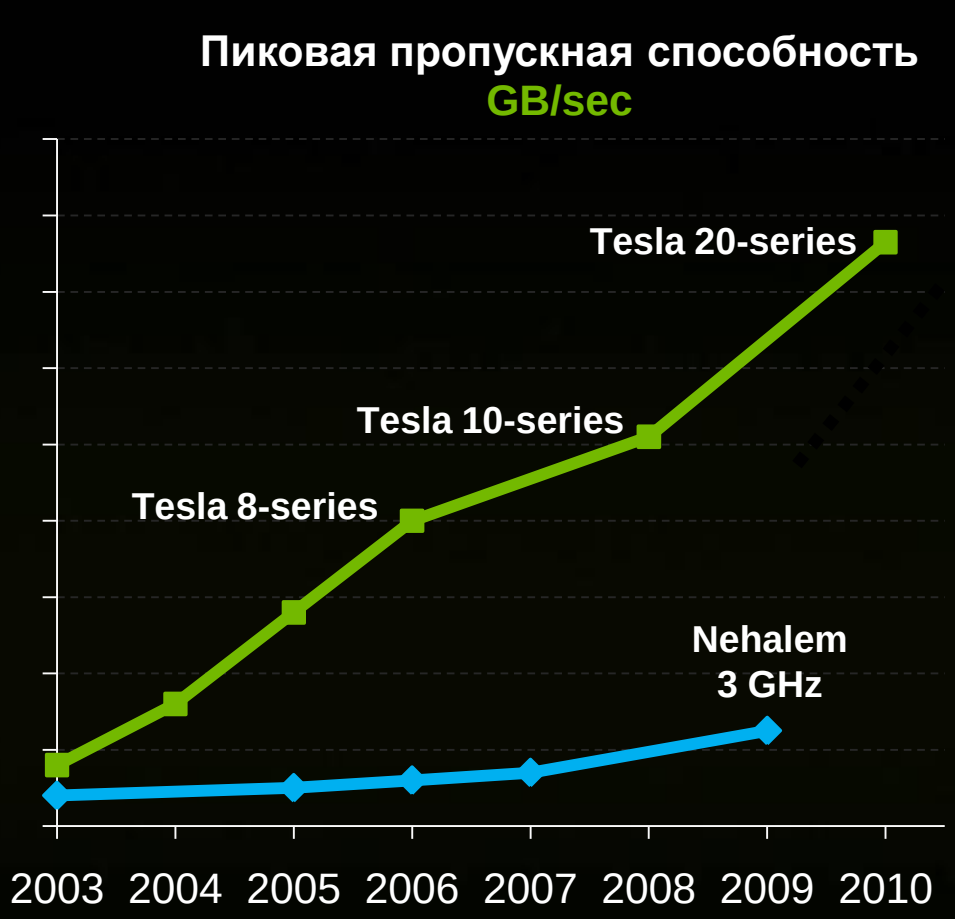


Вычисления CPU + GPU

Гибридные вычислительные системы

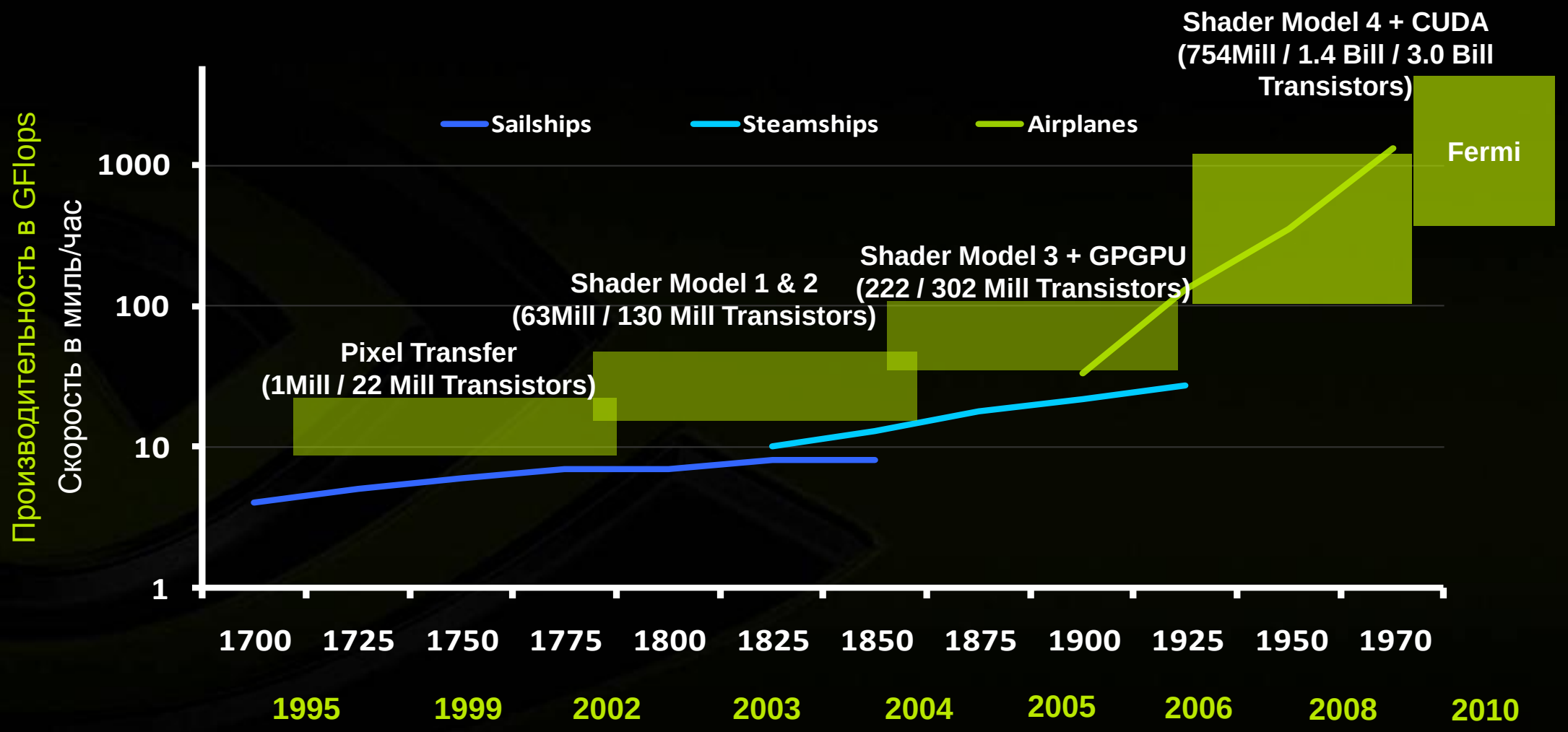
Постоянный прирост производительности





NVIDIA GPU
X86 CPU

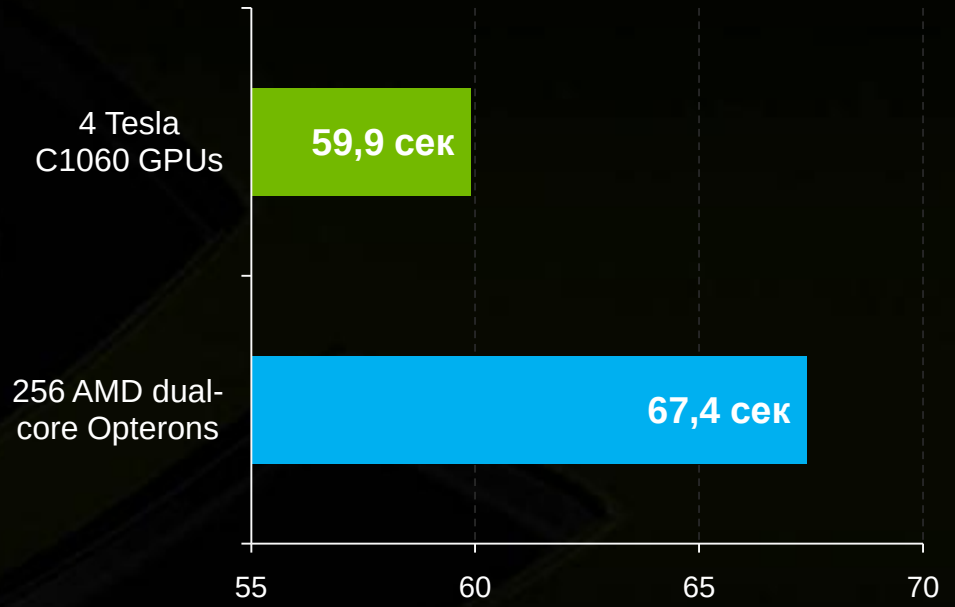
Когда прирост производительности имеет значение?



GPU : переломный момент в отрасли суперкомпьютеров

Десктоп эффективнее кластера

CalcUA
$5 млн.



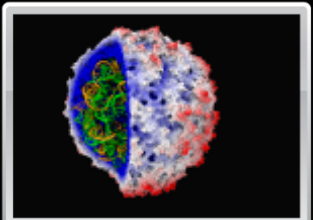
Tesla Personal Supercomputer
$10,000

Источник: University of Antwerp, Belgium

Прирост производительности до 150 раз

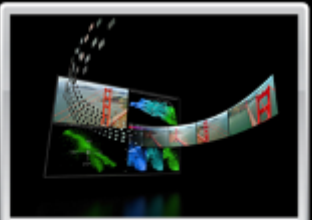
146X

Рентгенография, томография
U of Utah



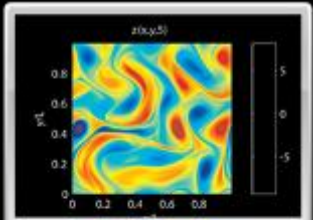
36X

Молекулярная динамика
U of Illinois, Urbana



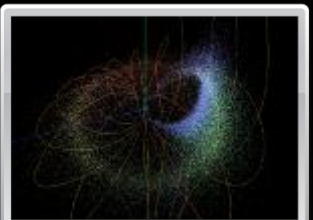
18X

Транскодирование видео
Elemental Tech



50X

Математические вычисления
AccelerEyes

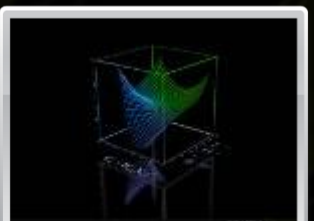


100X

Астрофизика
RIKEN

149X

Финансовые задачи
Oxford



47X

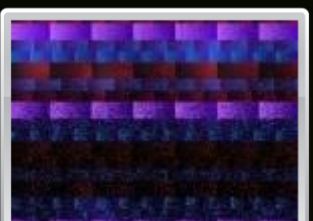
Линейная алгебра
Universidad Jaime

20X

3D ультразвук
Techniscan

130X

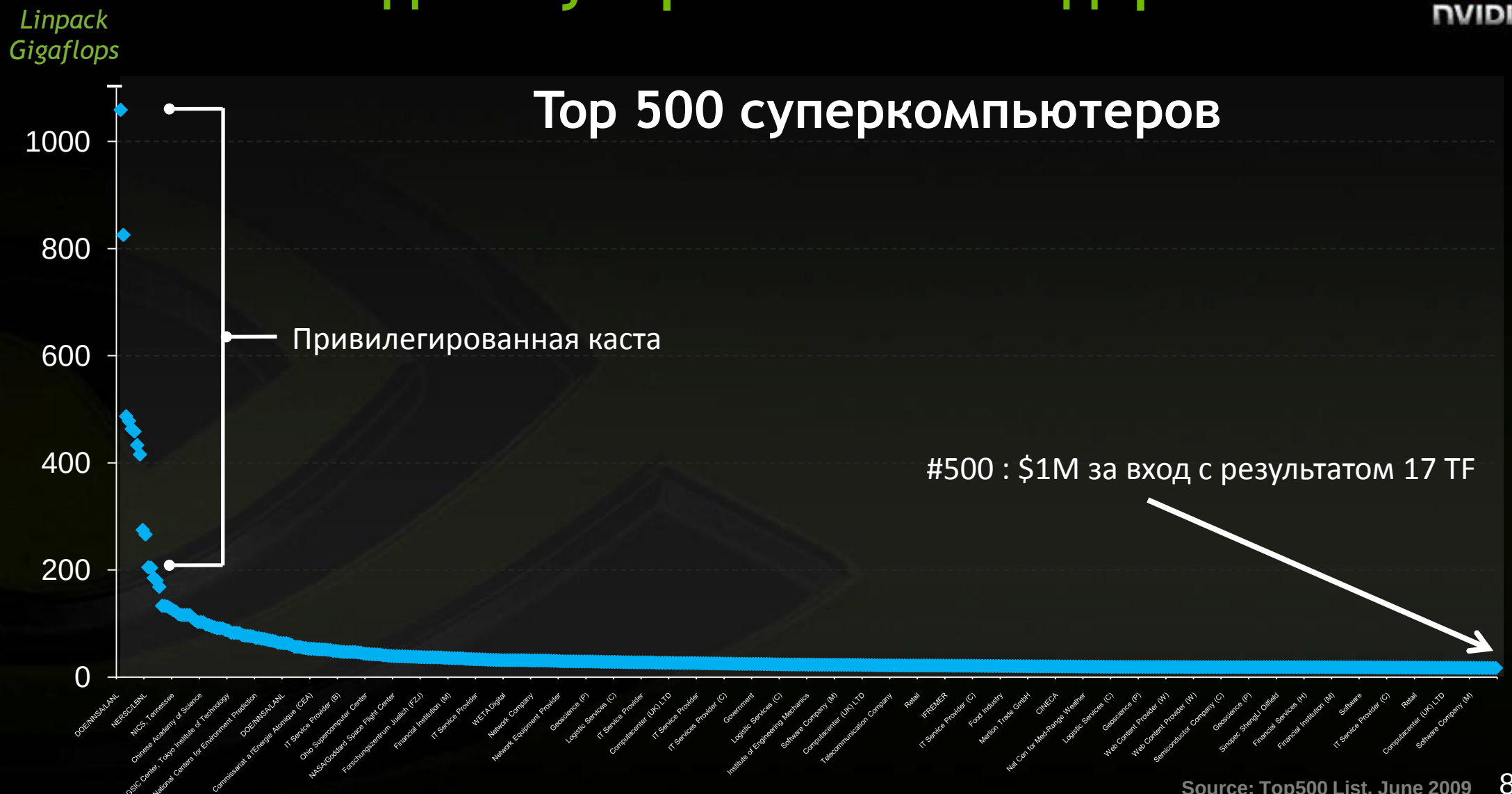
Квантовая химия
U of Illinois, Urbana



30X

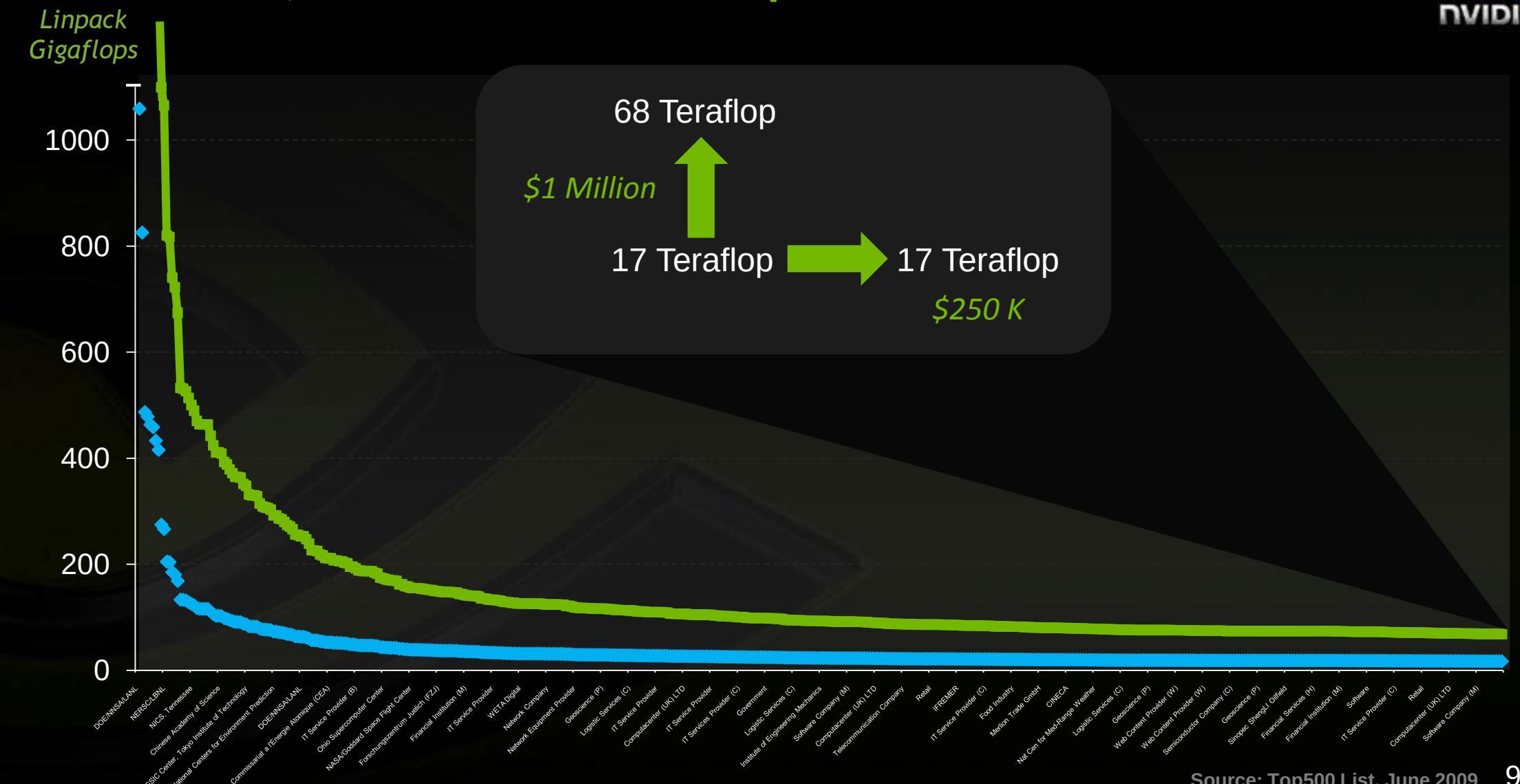
Генная инженерия
U of Maryland

Сегодня: супервычисления дороги



Source: Top500 List, June 2009

Что, если бы все из Top 500 использовали GPU



Source: Top500 List, June 2009

Решения для нефтегазового сектора

GPU vs CPU преимущества	HESS	PETROBRAS
Производительность / Ватт	18x - 27x	12x - 17x
Производительность / м^2	20x - 31x	15x - 20x
Производительность / $	15x - 20x	10x - 12x

Bloomberg: прогнозирование рынка облигаций

48 GPUs

42x экономия места

2000 CPUs

$144K

28x меньше стоимость

$4 млн

Финансы: оценка котировок

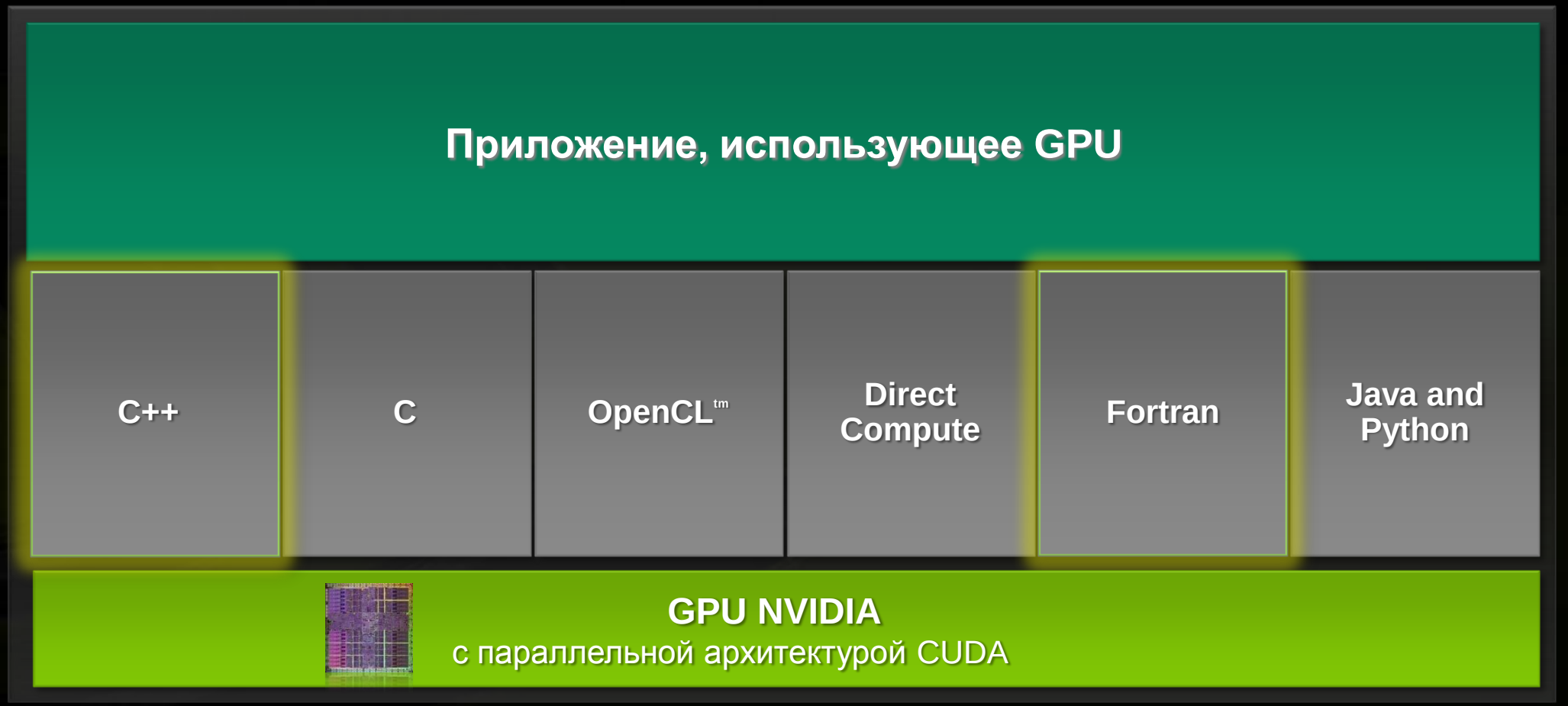
1	**Одинаковая производительность**	1
2 Tesla S1070	**16x экономия места**	500 CPU серверов
$24 K	**10x меньше стоимость**	$250 K
2.8 кВт	**13x меньше потребление**	37.5 кВт

CUDA

Compute Unified Device Architecture

Программно-аппаратная архитектура для параллельных вычислений

Параллельная архитектура CUDA



CUDA – самая массово преподаваемая модель параллельного программирования

269 университетов преподают параллельную модель программирования CUDA

Более 250 заказчиков / разработчиков ПО

Life Sciences & Medical Equipment		Productivity / Misc	Oil and Gas	EDA	Manufacturing	Finance	CAE / Numerics	Communication
Max Planck	GE Healthcare	CEA	Hess	Synopsys	Renault	Symcor	The Mathworks	Nokia
FDA	Siemens	WRF Weather Modeling	TOTAL	Nascentric	Boeing	Level 3	Wolfram	RIM
Robarts Research	Techniscan	OptiTex	CGG/Veritas	Gauda		SciComp	National Instruments	Philips
Medtronic	Boston Scientific	Tech-X	Chevron	CST		Hanweck	Access Analytics	Samsung
AGC	Eli Lilly	Elemental Technologies	Headwave	Agilent		Quant Catalyst	Tech-x	LG
Evolved machines	Silicon Informatics	Dimensional Imaging	Acceleware			RogueWave	RIKEN	Sony Ericsson
Smith-Waterman DNA sequencing	Stockholm Research	Manifold	Seismic City			BNP Paribas	SOFA	NTT DoCoMo
AutoDock	Harvard	Digisens	P-Wave Seismic Imaging					Mitsubishi
NAMD/VMD	Delaware	General Mills	Mercury Computer					Hitachi
Folding@Home	Pittsburg	Rapidmind	ffA					Radio Research Laboratory
Howard Huges Medical	ETH Zurich	MS Visual Studio						US Air Force
CRIBI Genomics	Institute Atomic Physics	Rhythm & Hues						
		xNormal						
		Elcomsoft						
		LINZIK						

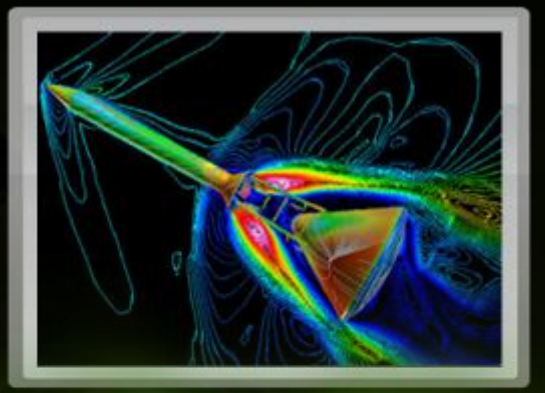
CUDA

Параллельные вычисления на GPU

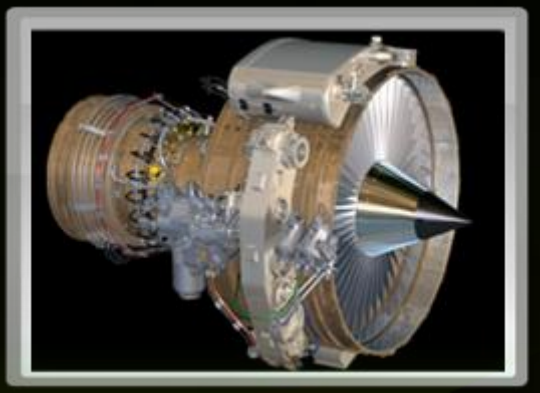
200+ млн. GPU в мире поддерживают CUDA

GeForce®
Развлечения

Tesla™
Высокопроизводительные вычисления



Quadro®
Дизайн, разработка



Выбор CUDA платформы

	Tesla	Quadro	GeForce
Стресс-тест с проверкой точности вычислений	X		
Произведено NVIDIA из высококачественных комплектующих	X	X	
3-х летняя гарантия, корпоративная поддержка	X	X	
4 Гб оперативной памяти для работы с большими объемами данных	X	X	
Единое профессиональное решение для вычислений и графики		X	
Пользовательские приложения: PhysX, Video, Imaging			X
Короткий жизненный цикл пользовательского продукта			X
Производится и сопровождается партнерами NVIDIA			X
Поддержка осуществляется через партнеров NVIDIA			X

Вычислительные решения Tesla

Созданы для профессионалов.

Персональный суперкомпьютер Tesla

Производительность

Tesla C1060
933 Гфлопс SP
78 Гфлопс DP
4 Гб памяти

Tesla C2050
520-630 Гфлопс DP
3 Гб памяти
ECC

Tesla C2070
520-630 Гфлопс DP
6 Гб памяти
ECC

Решения среднего уровня

8x производительность в DP

ЦОД

Q4	Q1	Q2	Q3	Q4
2009	**2010**			

Disclaimer: performance specification may change

Tesla для ЦОД

Производительность

Tesla S1070-500
4.14 Тфлопс SP
345 Гфлопс DP
4 Гб памяти/ GPU

Tesla S2050
2.1-2.5 Тфлопс DP
3 Гб памяти/ GPU
ECC

Tesla S2070
2.1-2.5 T Тфлопс DP
6 Гб памяти/ GPU
ECC

ЦОД

8x производительность в DP

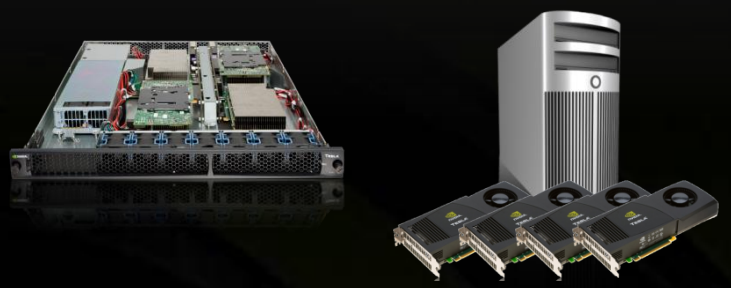
Решения среднего уровня

Q4	Q1	Q2	Q3	Q4
2009	**2010**			

Disclaimer: performance specification may change

Tesla: созданы для вычислений

Системные решения



Профессиональный уровень

- Высококачественная память
- Надежность вычислений
- 24-часовой стресс-тест

Поддержка NVIDIA

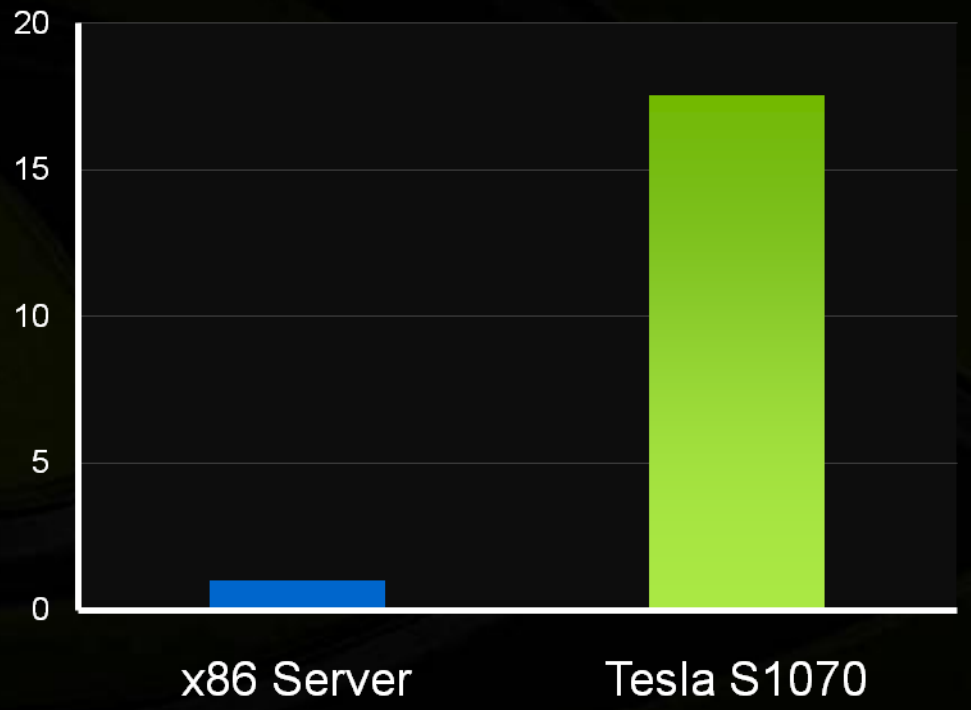
- 3-х летняя гарантия
- Длинный жизненный цикл
- Корпоративная поддержка

Вычислительные ресурсы

- Архитектура вычислений CUDA
- 4 ГБ ОЗУ
- двойная точность IEEE-754

Tesla S1070: эффективное решение

До 20 раз лучше показатель производительность/Ватт



- Hess
- Chevron
- Petrobras
- NCSA
- CEA
- TiTech
- JFCOM
- SAIC
- Federal
- Motorola
- Kodak
- University of Heidelberg
- University of Illinois
- University of North Carolina
- Max Planck Institute
- Rice University
- University of Maryland
- GusGus
- Eotvas University
- University of Wuppertal
- Chinese Academy of Sciences
- National Taiwan University

Пример: ЦОД стоимостью $5 млн.

CPU 1U Server	2 Quad-core Xeon CPUs: 8 cores	8 CPU Cores + 960 GPU Cores	CPU 1U Server Tesla 1U System
	0.17 Teraflop (single) 0.08 Teraflop (double)	4.14 Teraflops (single) 0.346 Teraflop (double)	
	$ 3,000	$ 11,000	
	700 W	1500 W	
	1819 CPU servers	455 CPU servers 455 Tesla systems	
	310 Teraflops (single)	1961 Teraflops (single)	→ *6x more perf*
	155 Teraflops (double)	196 Teraflops (double)	
	Total area 16K sq feet	Total area 9K sq feet	→ *60% smaller*
	Total 1273 KW	Total 682 KW	→ *½ the power*

Персональный суперкомпьютер Tesla

Производительность

- Массивно параллельная CUDA архитектура
- 960 ядер. 4 Терафлоп/с
- В 250 раз мощнее ПК

Удобство

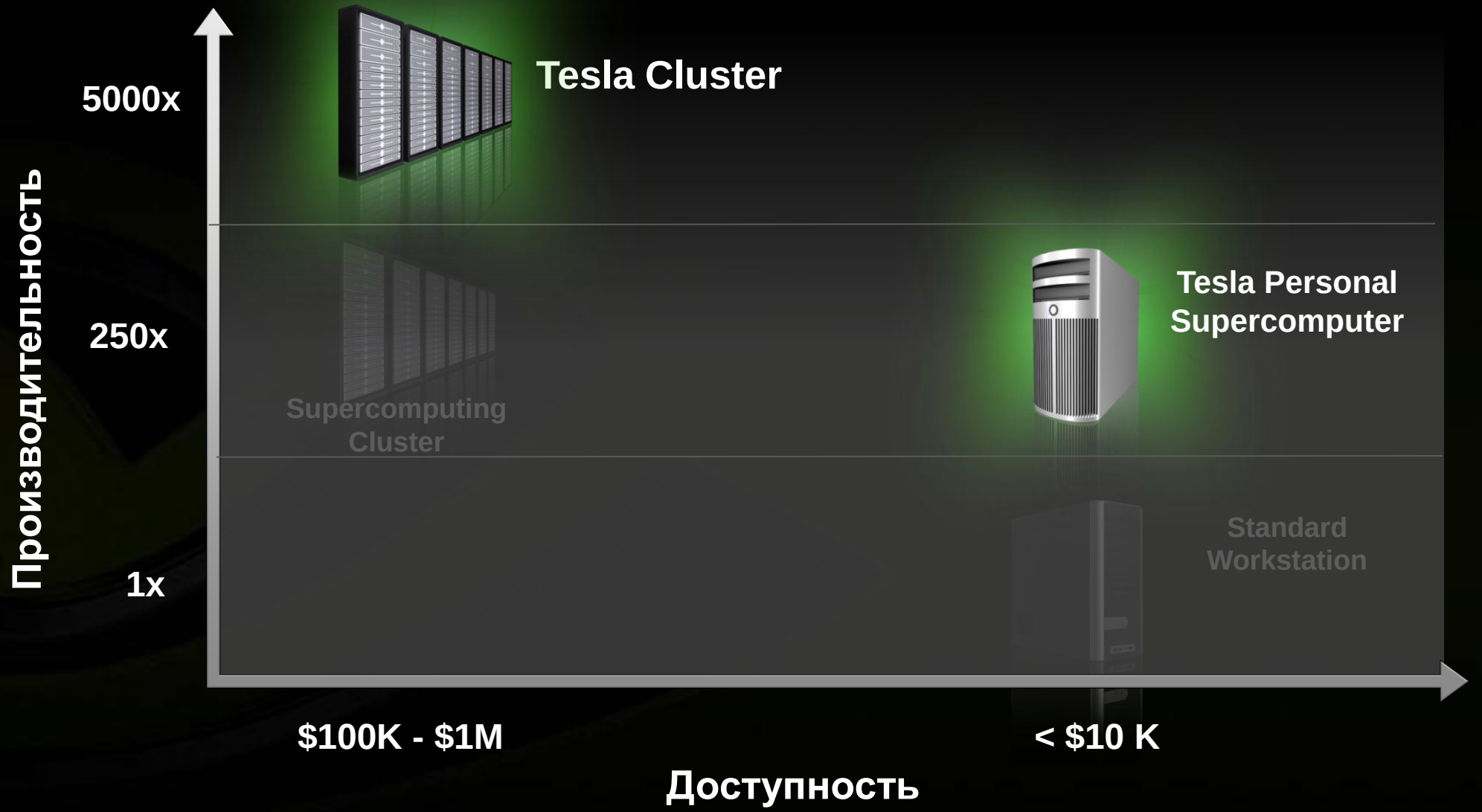
- Суперкомпьютер на рабочем столе
- Включается в обычную розетку

Доступность

- Программируется на Си под Windows и Linux
- Стоимость порядка $10,000


В 100 раз доступнее
В 20 раз экономичнее
Tesla
Personal
Supercomputer
Supercomputing
Cluster
В 250 раз
быстрее
Современный
ПК
250x
1x
Производительность
$100K - $1M
< $10 K
Доступность

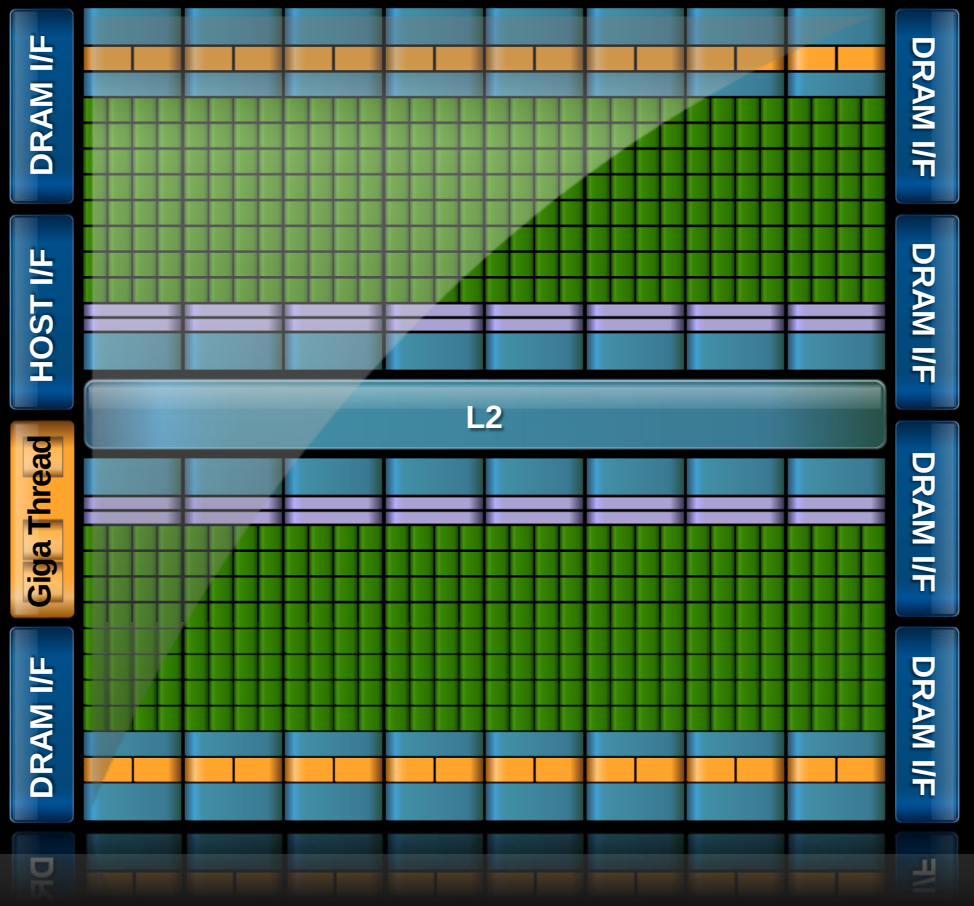
Высокопроизводительные вычисления на базе GPU



Новое поколение архитектуры GPU CUDA “Fermi”

Обзор Fermi

Суперкомпьютер в формате GPU



- *3 млрд транзисторов*
- *Вдвое больше ядер (512 ядер)*
- *8-кратный прирост DP производительности*
- *ECC*
- *L1 и L2 кэш*
- *Вдвое большая пропускная способность памяти (GDDR5)*
- *До 1Тб памяти на GPU*
- *Конкурирующее исполнение кода, C++*

“Oak Ridge National Lab (ORNL) has already announced it will be using Fermi technology in an upcoming super that is "expected to be 10-times more powerful than today's fastest supercomputer."

Since ORNL's Jaguar supercomputer, for all intents and purposes, holds that title, and is in the process of being upgraded to 2.3 PFlops….

…we can surmise that the upcoming Fermi-equipped super is going to be in the **20 Petaflops** range.”

HPC wire

September 30 2009

Tesla C – GPU для рабочих станций

	C2050	C2070
Процессоры	1 GPU 20-й серии Tesla 20	
Производительность DP	520-630 Гфлопс	
Память GPU	3 Гб	6 Гб
Интерфейс памяти	GDDR5	
Интерфес	PCIe x16 Gen2	
Энергопотребление	190 Вт (типичное) 225 Вт (максимальное)	

Disclaimer: specification subject to change
ECC

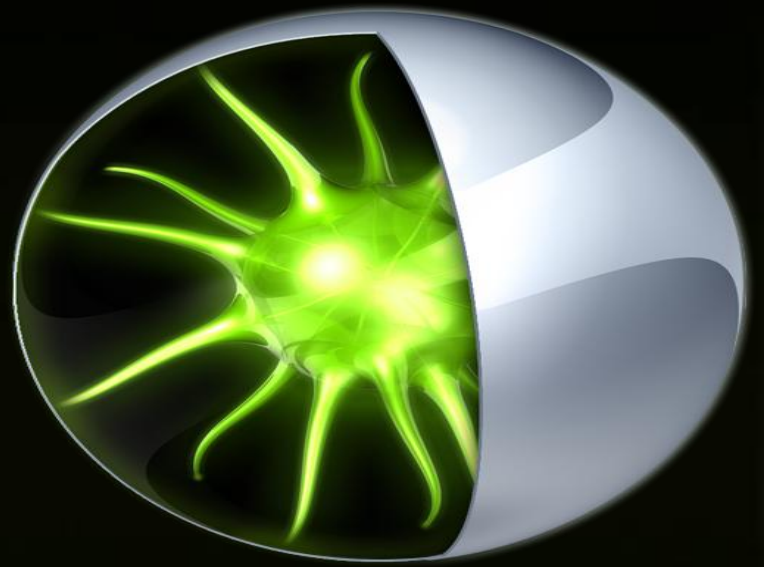
Tesla S – GPU для 1U систем

	S2050	S2070
Процессоры	4 GPU 20-й серии Tesla 20	
Производительность DP	2.1 – 2.5 Тфлопс	
Память GPU	3 Гб / GPU	6 Гб / GPU
Интерфейс памяти	GDDR5	
Интерфес	2x PCIe x16 Gen2 (опционально x8)	
Энергопотребление	900 Вт (типичное) 1200 Вт (максимальное)	

Disclaimer: specification subject to change

NVIDIA Nexus IDE

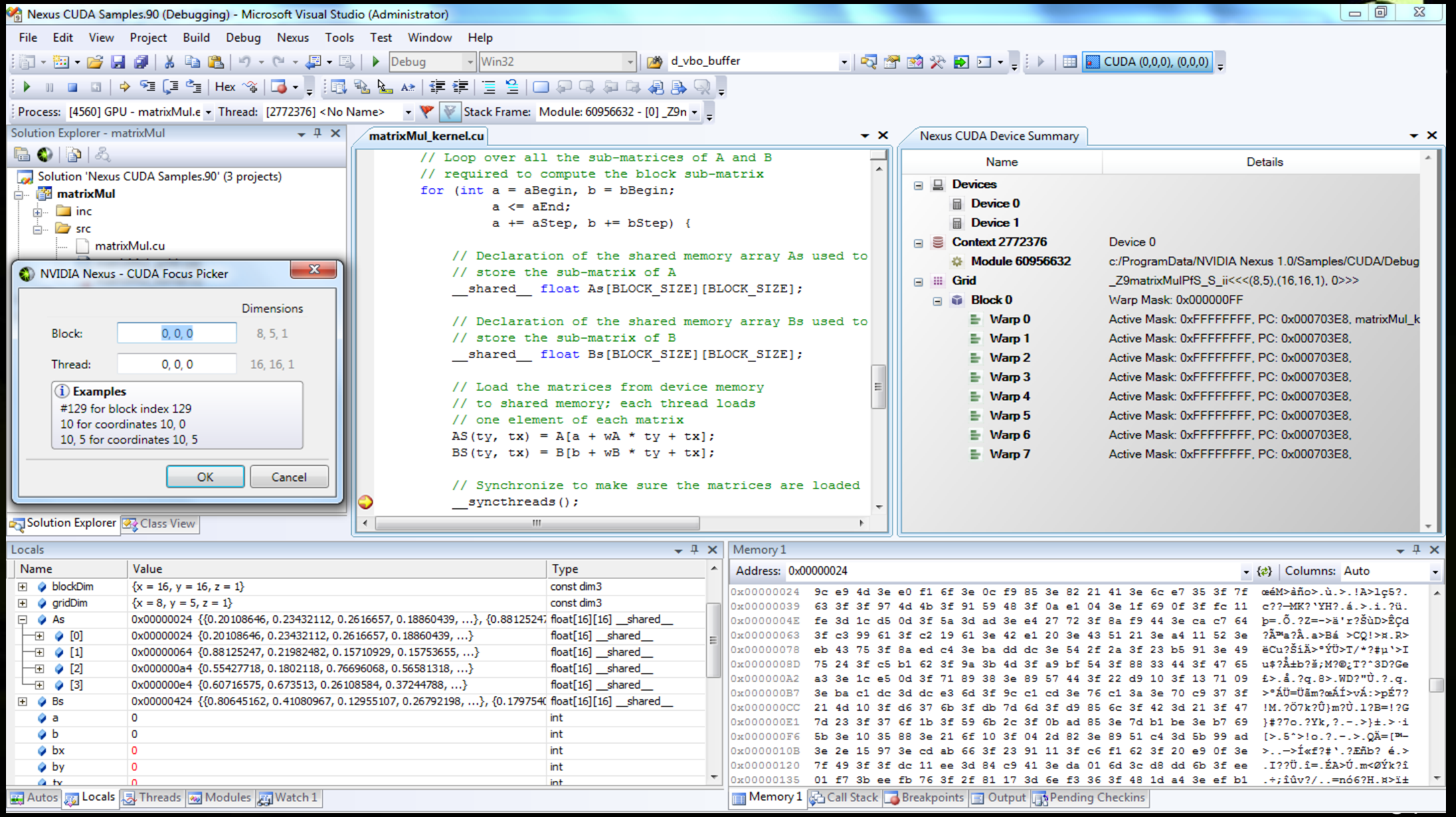
1-й в индустрии IDE (Integrated Development Environment) для массивно-параллельных приложений



Ускорение разработки гетерогенных (CPU + GPU) приложений

Полная интеграция со средствами разработки Visual Studio

Microsoft Visual Studio PARTNER


Nexus CUDA Samples.90 (Debugging) - Microsoft Visual Studio (Administrator)
File Edit View Project Build Debug Nexus Tools Test Window Help
Debug
Win32
d_vbo_buffer
CUDA (0,0,0), (0,0,0)
Process: [4560] GPU - matrixMul.e
Thread: [2772376] <No Name>
Stack Frame: Module: 60956632 - [0] _Z9n
Solution Explorer - matrixMul
Solution 'Nexus CUDA Samples.90' (3 projects)
matrixMul
inc
src
matrixMul.cu
NVIDIA Nexus - CUDA Focus Picker
Dimensions
Block:
0, 0, 0
8, 5, 1
Thread:
0, 0, 0
16, 16, 1
Examples
#129 for block index 129
10 for coordinates 10, 0
10, 5 for coordinates 10, 5
OK
Cancel
Solution Explorer
Class View
matrixMul_kernel.cu
// Loop over all the sub-matrices of A and B
// required to compute the block sub-matrix
for (int a = aBegin, b = bBegin;
a <= aEnd;
a += aStep, b += bStep) {
// Declaration of the shared memory array As used to
// store the sub-matrix of A
__shared__ float As[BLOCK_SIZE][BLOCK_SIZE];
// Declaration of the shared memory array Bs used to
// store the sub-matrix of B
__shared__ float Bs[BLOCK_SIZE][BLOCK_SIZE];
// Load the matrices from device memory
// to shared memory; each thread loads
// one element of each matrix
AS(ty, tx) = A[a + wA * ty + tx];
BS(ty, tx) = B[b + wB * ty + tx];
// Synchronize to make sure the matrices are loaded
__syncthreads();
Nexus CUDA Device Summary
Name
Details
Devices
Device 0
Device 1
Context 2772376
Device 0
Module 60956632
c:/ProgramData/NVIDIA Nexus 1.0/Samples/CUDA/Debug
Grid
_Z9matrixMulPfS_S_ii<<<(8,5),(16,16,1), 0>>>
Block 0
Warp Mask: 0x000000FF
Warp 0
Active Mask: 0xFFFFFFFF, PC: 0x000703E8, matrixMul_k
Warp 1
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Warp 2
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Warp 3
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Warp 4
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Warp 5
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Warp 6
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Warp 7
Active Mask: 0xFFFFFFFF, PC: 0x000703E8,
Locals
Name
Value
Type
blockDim
{x = 16, y = 16, z = 1}
const dim3
gridDim
{x = 8, y = 5, z = 1}
const dim3
As
0x00000024 {{0.20108646, 0.23432112, 0.2616657, 0.18860439, ...}, {0.8812524
float[16][16] __shared__
[0]
0x00000024 {0.20108646, 0.23432112, 0.2616657, 0.18860439, ...}
float[16] __shared__
[1]
0x00000064 {0.88125247, 0.21982482, 0.15710929, 0.15753655, ...}
float[16] __shared__
[2]
0x000000a4 {0.55427718, 0.1802118, 0.76696068, 0.56581318, ...}
float[16] __shared__
[3]
0x000000e4 {0.60716575, 0.673513, 0.26108584, 0.37244788, ...}
float[16] __shared__
Bs
0x00000424 {{0.80645162, 0.41080967, 0.12955107, 0.26792198, ...}, {0.1797540
float[16][16] __shared__
a
0
int
b
0
int
bx
0
int
by
0
int
Autos
Locals
Threads
Modules
Watch 1
Memory 1
Address: 0x00000024
Columns: Auto
0x00000024 9c e9 4d 3e e0 f1 6f 3e 0c f9 85 3e 82 21 41 3e 6c e7 35 3f 7f
0x00000039 63 3f 3f 97 4d 4b 3f 91 59 48 3f 0a e1 04 3e 1f 69 0f 3f fc 11
0x0000004E fe 3d 1c d5 0d 3f 5a 3d ad 3e e4 27 72 3f 8a f9 44 3e ca c7 64
0x00000063 3f c3 99 61 3f c2 19 61 3e 42 e1 20 3e 43 51 21 3e a4 11 52 3e
0x00000078 eb 43 75 3f 8a ed c4 3e ba dd dc 3e 54 2f 2a 3f 23 b5 91 3e 49
0x0000008D 75 24 3f c5 b1 62 3f 9a 3b 4d 3f a9 bf 54 3f 88 33 44 3f 47 65
0x000000A2 a3 3e 1c e5 0d 3f 71 89 38 3e 89 57 44 3f 22 d9 10 3f 13 71 09
0x000000B7 3e ba c1 dc 3d dc e3 6d 3f 9c c1 cd 3e 76 c1 3a 3e 70 c9 37 3f
0x000000CC 21 4d 10 3f d6 37 6b 3f db 7d 6d 3f d9 85 6c 3f 42 3d 21 3f 47
0x000000E1 7d 23 3f 37 6f 1b 3f 59 6b 2c 3f 0b ad 85 3e 7d b1 be 3e b7 69
0x000000F6 5b 3e 10 35 88 3e 21 6f 10 3f 04 2d 82 3e 89 51 c4 3d 5b 99 ad
0x0000010B 3e 2e 15 97 3e cd ab 66 3f 23 91 11 3f c6 f1 62 3f 20 e9 0f 3e
0x00000120 7f 49 3f 3f dc 11 ee 3d 84 c9 41 3e da 01 6d 3c d8 dd 6b 3f ee
0x00000135 01 f7 3b ee fb 76 3f 2f 81 17 3d 6e f3 36 3f 48 1d a4 3e ef b1
Memory 1
Call Stack
Breakpoints
Output
Pending Checkins

Экосистема параллельных вычислений на GPU

Научные пакеты

MATLAB
Mathematica
NI LabView
pyCUDA

Дебагеры & профилировщики

cuda-gdb
NV Visual Profiler
“Nexus” VS 2008
Allinea
TotalView

GPU компиляторы

C
C++
Fortran
OpenCL
DirectCompute
Java
Python

Параллелизующие компиляторы

PGI Accelerator
CAPS HMPP
mCUDA
OpenMP

Библиотеки

BLAS
FFT
LAPACK
NPP
Video
Imaging

Консалтинг CUDA

acceleware · GASS · STONE RIDGE TECHNOLOGY · ANEO · GPU Tech · TECH-X · CAPS · HPC · EM Photonics · SCALABLE GRAPHICS · WIPRO

Интеграторы

DELL · hp · Sun microsystems · CRAY · ASUS · SUPERMICRO · sgi · FUJITSU · BULL · APPRO · lenovo 联想 · NEC

Ссылки

- Fermi
  - http://www.nvidia.ru/fermi
- CUDA Zone
  - http://www.nvidia.ru/cuda
  - Приложения, документы, видео
- Tesla
  - http://www.nvidia.ru/tesla
  - Спецификации, технические и маркетинговые материалы
- Вертикальные отраслевые решения
  - http://www.nvidia.com/object/vertical_solutions.html
- YouTube Videos
  - http://www.youtube.com/nvidiatesla