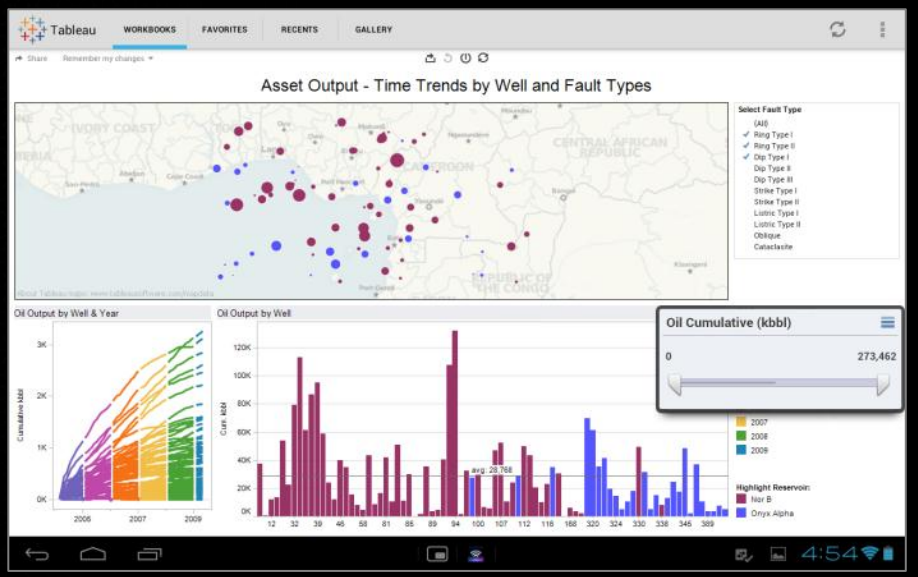
NVIDIA®

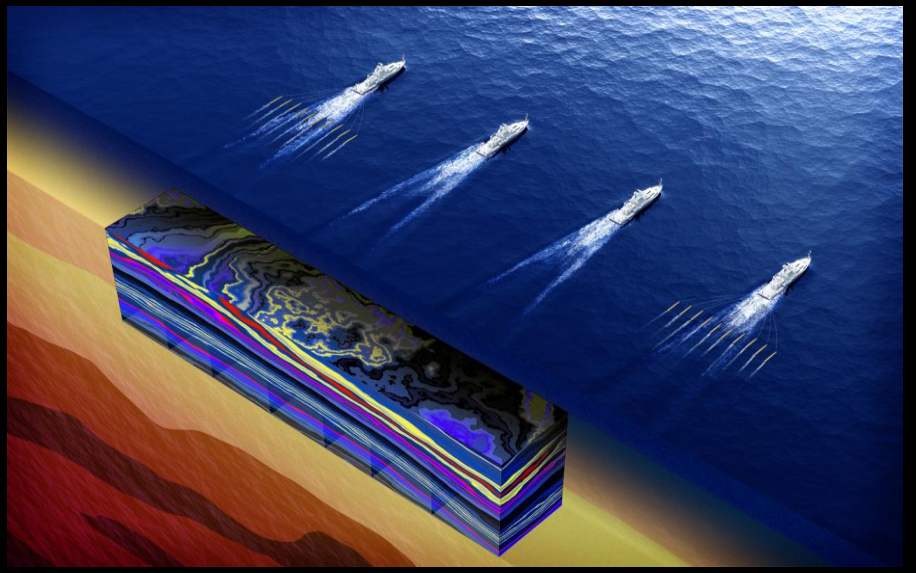
GPU – БУДУЩЕЕ ГИБРИДНЫХ ВС

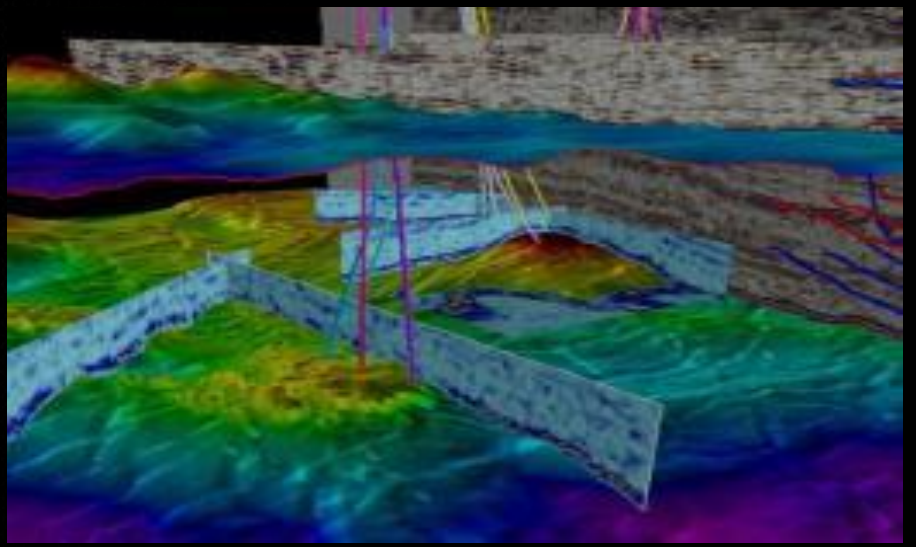
Антон Джораев

Современный мир требует все больших вычислительных ресурсов

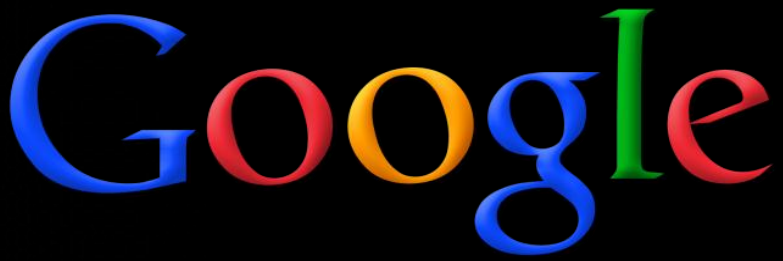


Современный мир требует все больших вычислительных ресурсов





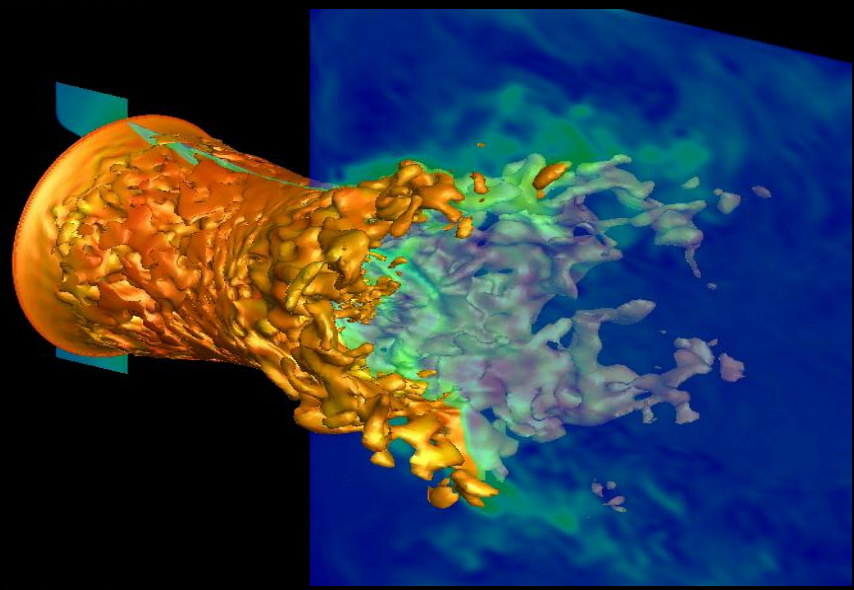
Современный мир требует все больших вычислительных ресурсов



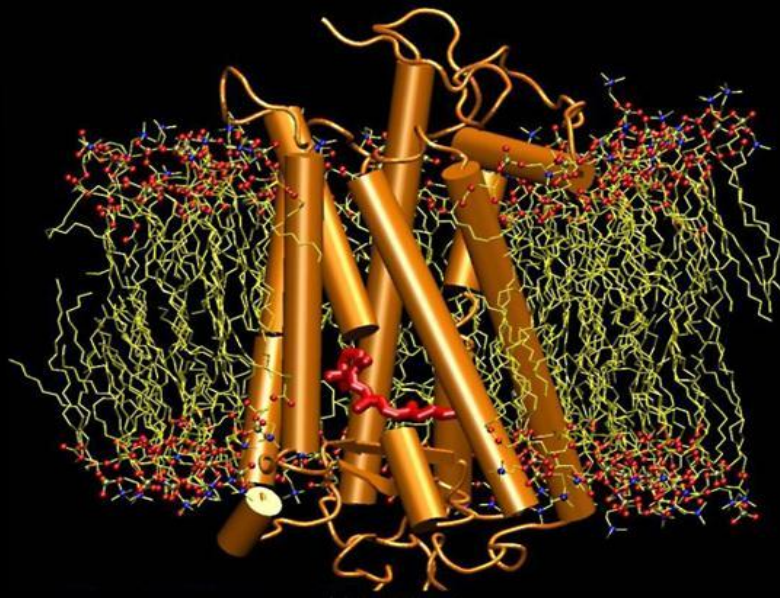
Яндекс

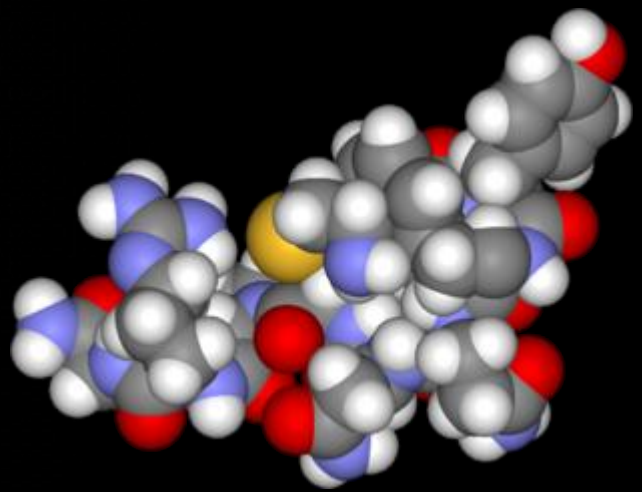
Современный мир требует все больших вычислительных ресурсов

Современный мир требует все больших вычислительных ресурсов

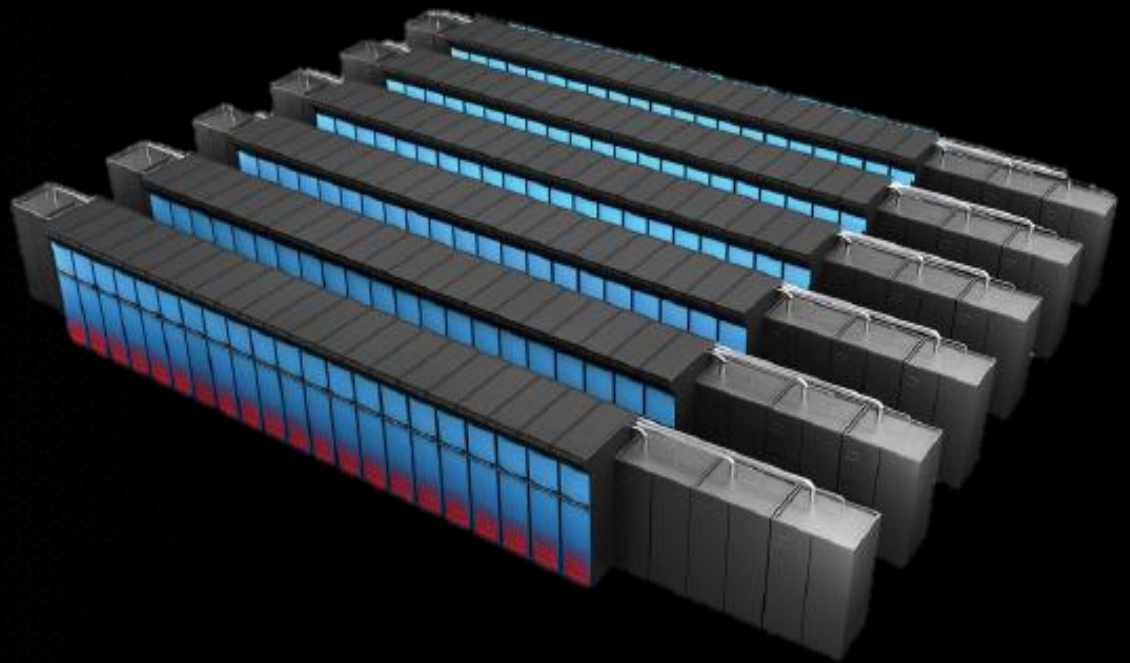


Современный мир требует все больших вычислительных ресурсов





Современный мир требует все больших вычислительных ресурсов



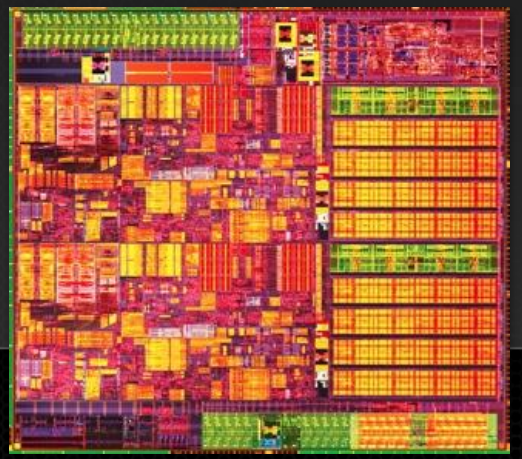
Проблемы архитектуры x86

- Оптимизация под последовательный код, низкая произв. SIMD
- Низкая эффективность - стоимость, энергетика, охлаждение
- Огромное количество ПО написанного давно и неэффективно

CPU
1690 pJ/flop

Оптимизирован для исполнения последовательного кода

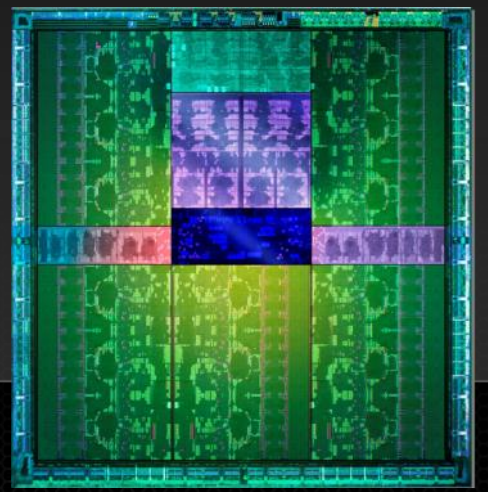
Неконтролируемый кэш



GPU
140 pJ/flop

Оптимизирован для работы с большим количеством данных

Явное управление памятью GPU

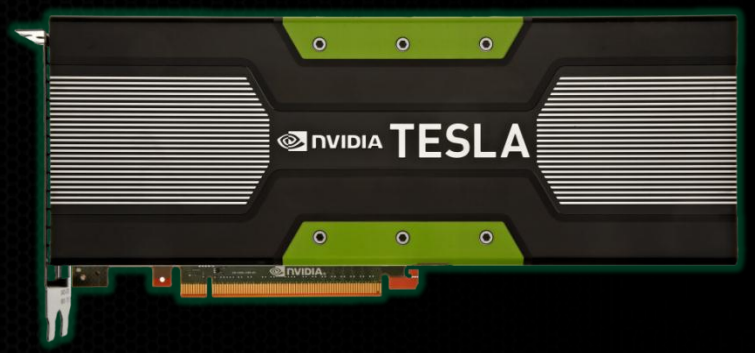


GPU - будущее HPC

- GPU – самая производительная архитектура
- GPU – наиболее энергоэффективная архитектура
- GPU – незаменимы для визуализации

GPU - самый производительный процессор

Tesla K20X



Кол-во CUDA ядер	2688
Пиковая произв-ть DP DGEMM	1.32 TF 1.22 TF
Пиковая произв-ть SP SGEMM	3.95 TF 2.90 TF
Пропускная способность памяти	250 GB/s
Объем памяти	6 GB
Потребление	235W

NVIDIA

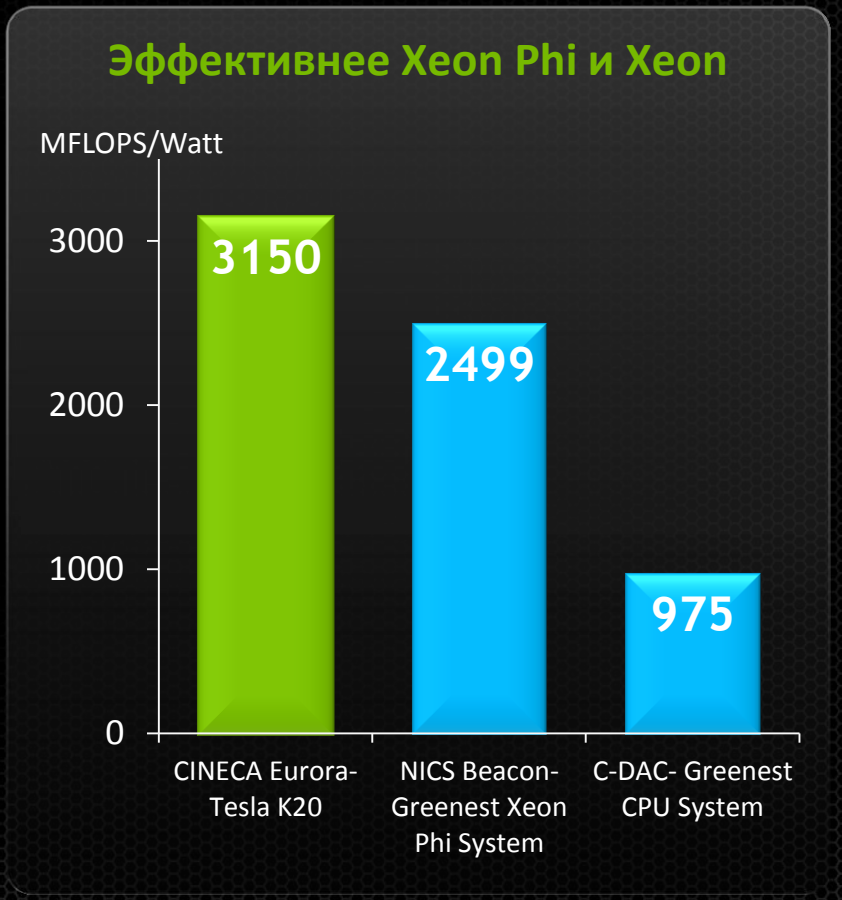
Самый энергоэффективный компьютер мира

3150 MFLOPS/Watt

128 Tesla K20

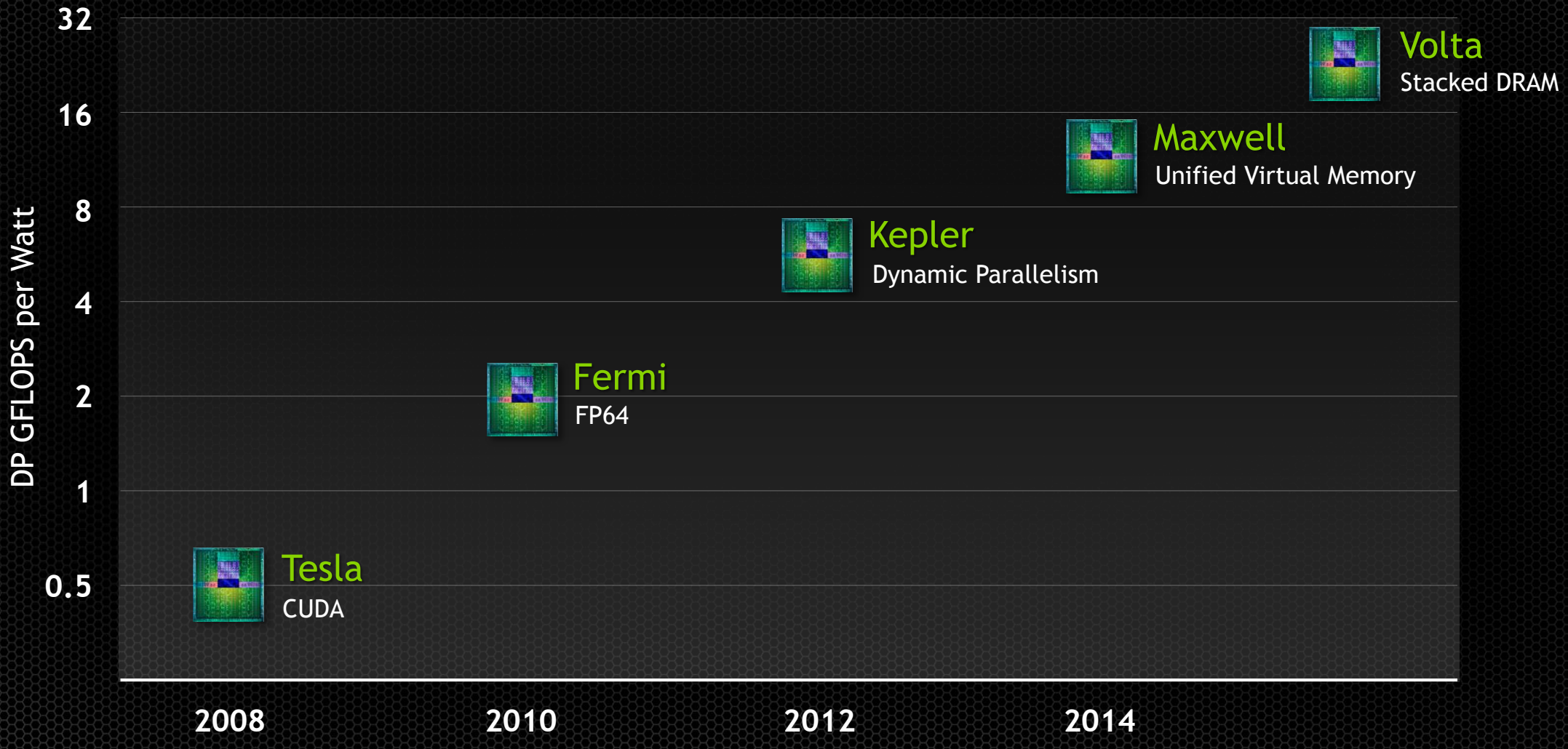
CINECA Eurora

"Liquid-Cooled" Eurotech Aurora Tigon

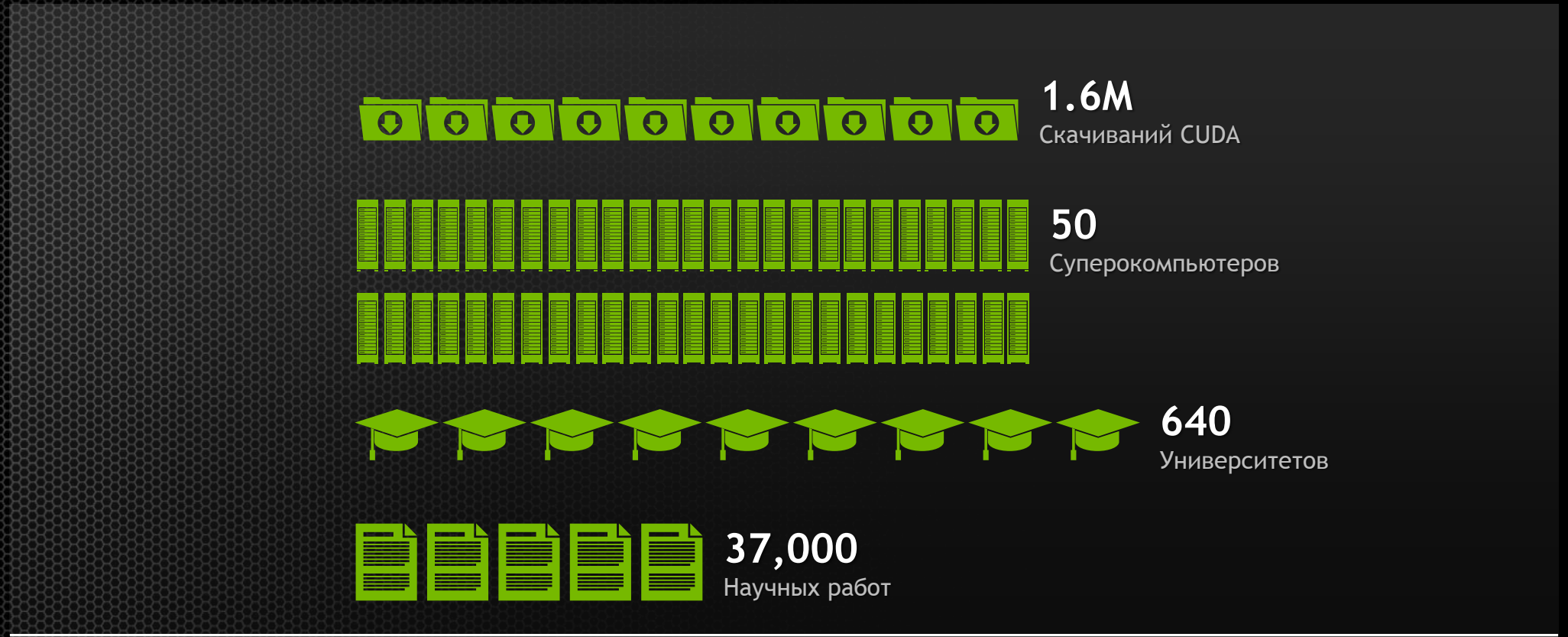


NVIDIA

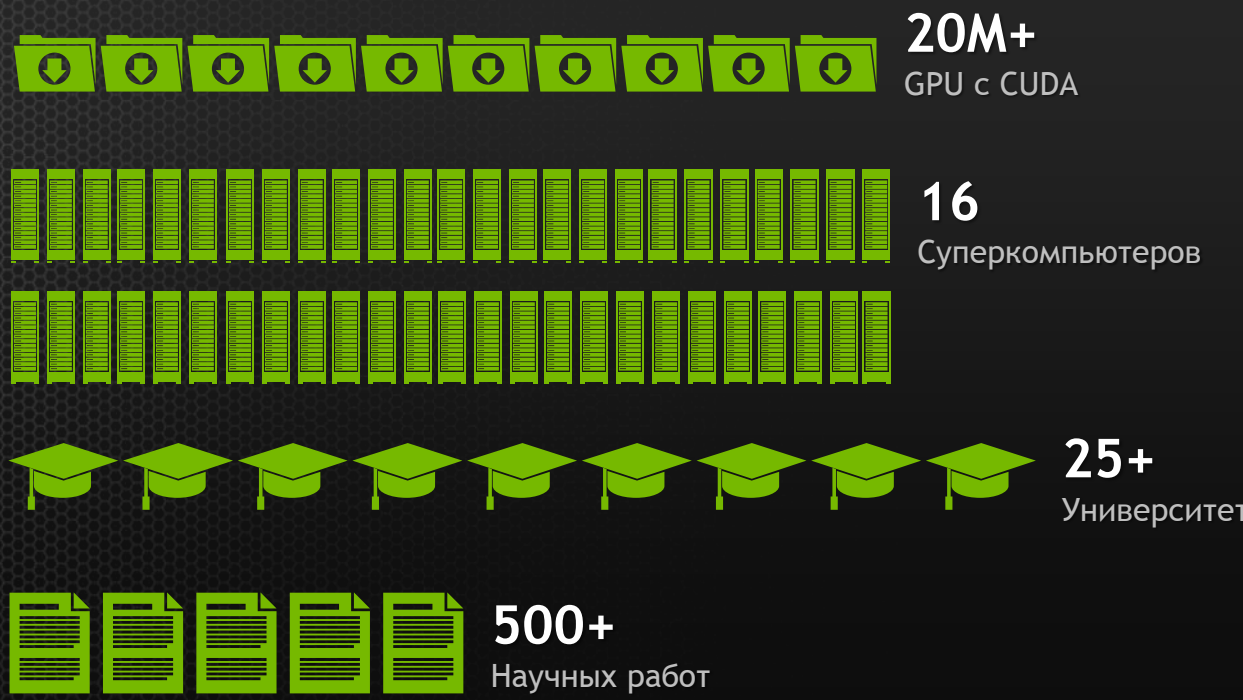
GPU Roadmap


32
16
8
4
2
1
0.5
DP GFLOPS per Watt
Tesla
CUDA
Fermi
FP64
Kepler
Dynamic Parallelism
Maxwell
Unified Virtual Memory
Volta
Stacked DRAM
2008
2010
2012
2014

Распространенность вычислений на GPU



Вычисления на GPU в России и СНГ



2013

NVIDIA

GPU доступны каждому исследователю

- 30% (120 из 400) институтов РАН решают вычислительные задачи
- 65% (78) из них применяют вычисления на GPU

GPU Accelerated Libraries

"Drop-in" Acceleration for your Applications

Linear Algebra
FFT, BLAS,
SPARSE, Matrix

NVIDIA cuFFT, cuBLAS, cuSPARSE

CULA tools

MAGMA

CUSP

Numerical & Math
RAND, Statistics

IMSL® Fortran Numerical Library

NVIDIA Math Lib

ArrayFire

NVIDIA cuRAND

Data Struct. & AI
Sort, Scan, Zero Sum

Thrust

GPU AI – Board Games

GPU AI – Path Finding

Visual Processing
Image & Video

NVIDIA NPP

H.264 Encoder (NVCUVENC)

Audio Encoder

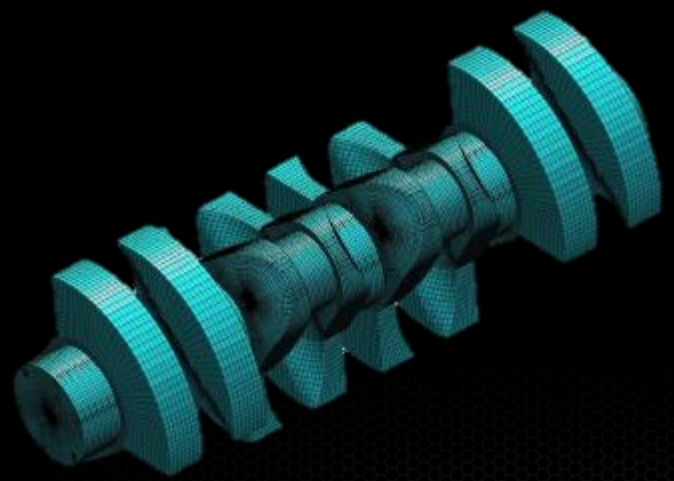
NVIDIA Video Encode

Sundog™ Software

NVIDIA


NVIDIA
POPULAR GPU-ACCELERATED APPLICATIONS
CONTENTS
02 Research: Higher Education and Supercomputing
COMPUTATIONAL CHEMISTRY AND BIOLOGY
NUMERICAL ANALYTICS
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-AIR GRAPHICS
ON-SET, REVIEW AND STEREO TOOLS
SIMULATION
WEATHER GRAPHICS
14 Oil and Gas
Research: Higher Education and Supercomputing
COMPUTATIONAL CHEMISTRY AND BIOLOGY
Bioinformatics
BarraCUDA Sequence mapping software Alignment of short sequencing reads 6-10x T 2075, 2090, K10, K20, K20X Yes Available now Version 0.6.2
CUDASW++ Open source software for Smith-Waterman protein database searches on GPUs Parallel search of Smith-Waterman database 10-50x T 2075, 2090, K10, K20, K20X Yes Available now Version 2.0.8
CUSHAW Parallelized short read aligner Parallel, accurate long read aligner - gapped alignments to large genomes 10x T 2075, 2090, K10, K20, K20X Yes Available now Version 1.0.40
GPU-BLAST Local search with fast k-tuple heuristic Protein alignment according to blastp, multi cpu threads 3-4x T 2075, 2090, K10, K20, K20X Single only Available now Version 2.2.26
GPU-HMMER Parallelized local and global search with profile Hidden Markov models Parallel local and global search of Hidden Markov Models 60-100x T 2075, 2090, K10, K20, K20X Yes Available now Version 2.3.2
mCUDA-MEME Ultrafast scalable motif discovery algorithm based on MEME Scalable motif discovery algorithm based on MEME 4-10x T 2075, 2090, K10, K20, K20X Yes Available now Version 3.0.12
SeqNFind A GPU Accelerated Sequence Analysis Toolset Reference assembly, blast, smith-waterman, hmm, de novo assembly 400x T 2075, 2090, K10, K20, K20X Yes Available now
UGENE Opensource Smith-Waterman for SSE/CUDA, Suffix array based repeats finder and dotplot Fast short read alignment 6-8x T 2075, 2090, K10, K20, K20X Yes Available now Version 1.11
WideLM Fits numerous linear models to a fixed design and response Parallel linear regression on multiple similarly-shaped models 150x T 2075, 2090, K10, K20, K20X Yes Available now Version 0.1-1
Molecular Dynamics
Abalone Models molecular dynamics of biopolymers for simulations of proteins, DNA and ligands Simulations (on 1060 GPU) 4-29x T 2075, 2090, K10, K20, K20X Single Only Available now Version 1.8.48
ACEMD GPU simulation of molecular mechanics force fields, implicit and explicit solvent Written for use on GPUs 160 ns/day GPU version only T 2075, 2090, K10, K20, K20X Yes Available now
AMBER Suite of programs to simulate molecular dynamics on biomolecule PMEMD: explicit and implicit solvent 89.44 ns/day JAC NVE T 2075, 2090, K10, K20, K20X Yes Available now Version 12 + bugfix9
DL-POLY Simulate macromolecules, polymers, ionic systems, etc on a distributed memory parallel computer Two-body forces, Link-cell pairs, Ewald SPME forces, Shake VV 4x T 2075, 2090, K10, K20, K20X Yes Available now, Version4.0 Source only
CHARMM MD package to simulate molecular dynamics on biomolecule. Implicit (5x), Explicit (2x) Solvent via OpenMM TBD T 2075, 2090, K10, K20, K20X Yes In Development 04/12
GROMACS Simulation of biochemical molecules with complicated bond interactions Implicit (5x), Explicit(2x) solvent 165 ns/Day DHFR T 2075, 2090, K10, K20, K20X Single only Available now Version 4.6 in 04/12
HOOMD-Blue Particle dynamics package written grounds up for GPUs Written for GPUs 2x T 2075, 2090, K10, K20, K20X Yes Available now
LAMMPS Classical molecular dynamics package Lennard-Jones, Morse, Buckingham, CHARMM, Tabulated, Course grain SDK, Anisotropic Gay-Bern, RE-squared, "Hybrid"combinations 3-18x T 2075, 2090, K10, K20, K20X Yes Available now
AMD Designed for high-performance simulation of large molecular systems 100M atom capable 6.44 ns/days STMV 585x 2050s T 2075, 2090, K10, K20, K20X Yes Available now, Version 2.9
enMM Library and application for molecular dynamics for HPC with GPUs Implicit and explicit solvent, custom forces Implicit: 127-213 ns/ day; Explic 18-55 ns/ DHFR T 2075, 2090, K10, K20, K20X Yes Available now Version 4.1.1
242 GPU-Accelerated Applications
www.nvidia.com/appscatalog

Инженерные расчеты на GPU



- ANSYS Mechanical
- Ansys Fluent
- Abaqus/Standard (Silmulia)
- MSC Nastran, MSC Marc
- Matlab
- CST Microwave Studio

Ускорение моделирования = больше итераций = выше качество и надежность
Меньше отказов у клиентов / меньше отзывов

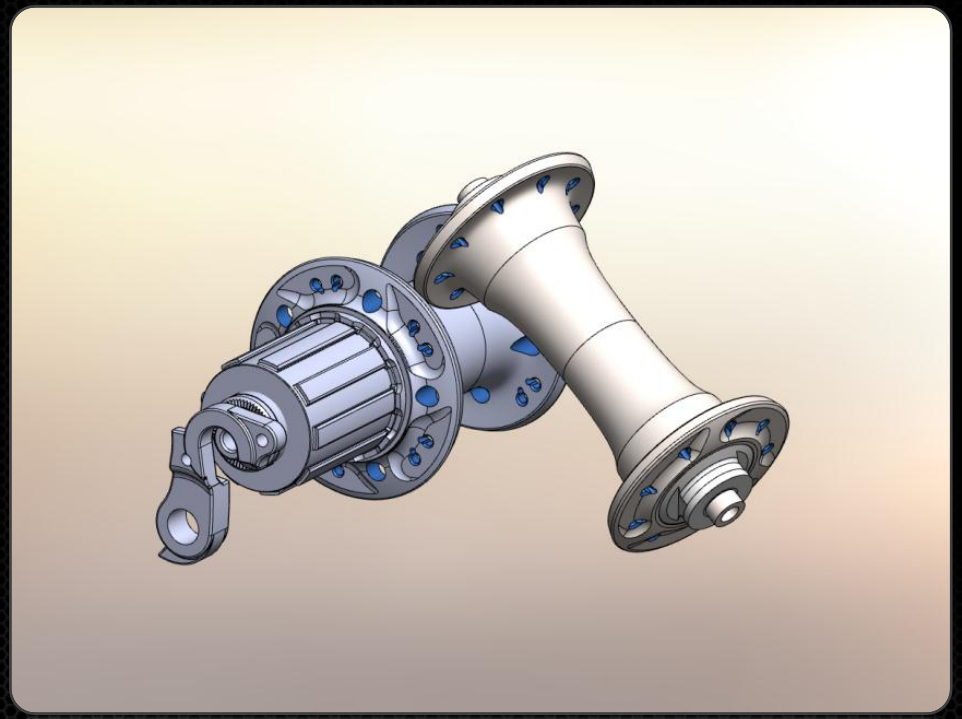
NVIDIA GRID

Производительность полноценного ПК

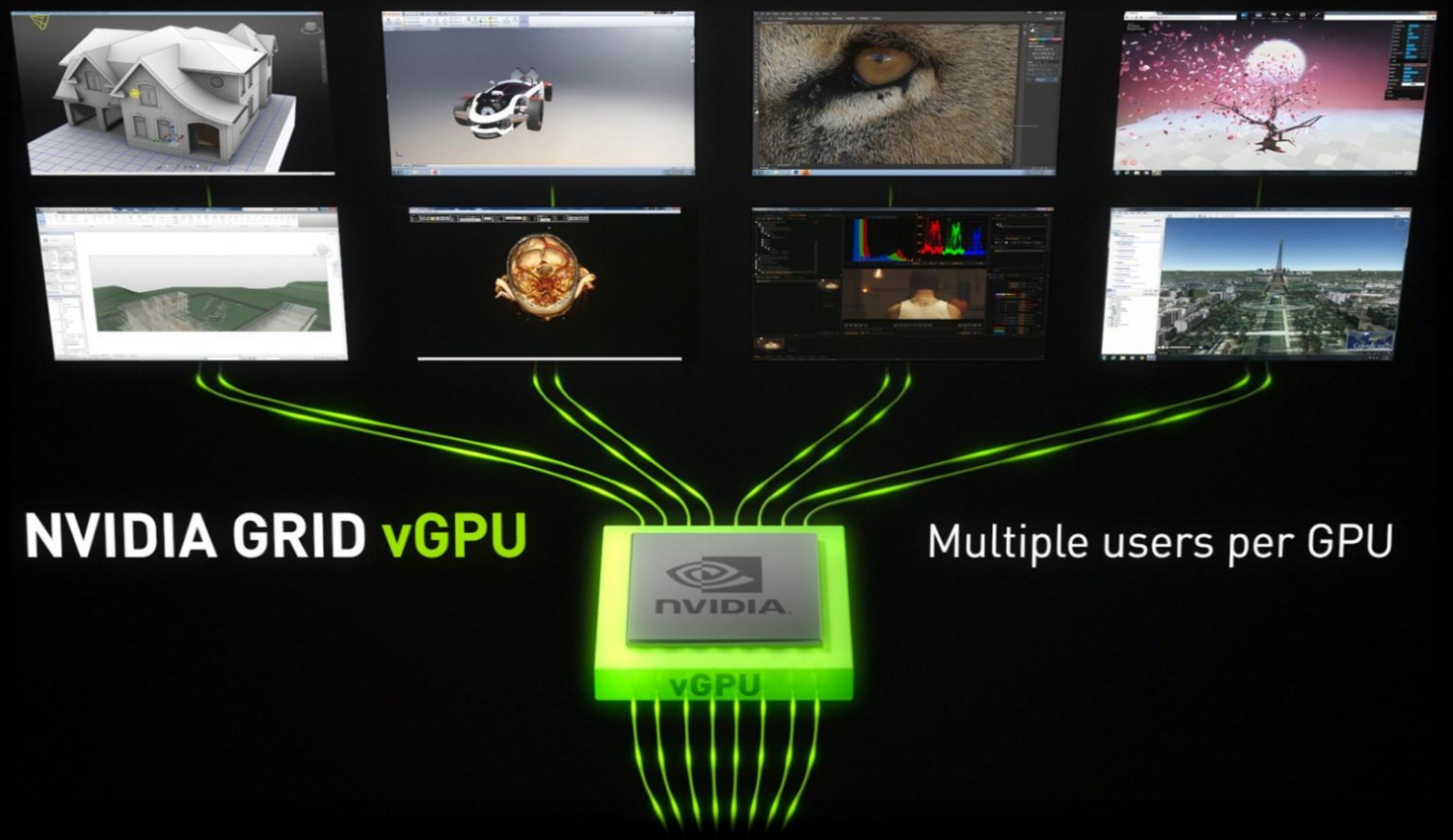
на любом персональном устройстве, с которого пользователь может подключиться к системе.

Разница для пользователя

без GPU

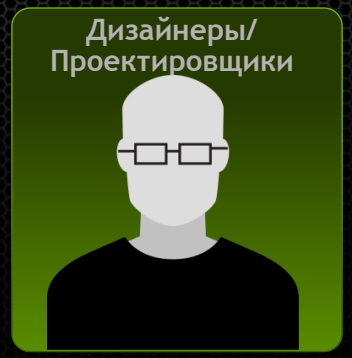


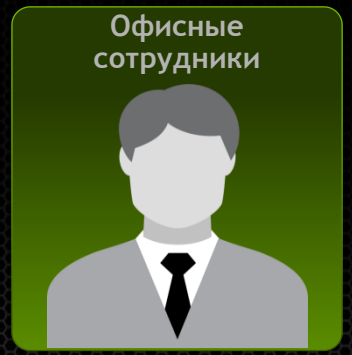
с GPU



NVIDIA GRID™ vGPU™ это технология аппаратной виртуализации графического процессора позволяющая множественным виртуальным машинам взаимодействовать с ним (с GPU) напрямую

NVIDIA





NVIDIA GRID K2

NVIDIA GRID K1

GPU	4 Kepler GPUs	2 топовых Kepler GPUs
CUDA ядра	768 (192/GPU)	3072 (1536/GPU)
Фреймбуфер	16GB DDR3 (4GB/GPU)	8GB GDDR5 (4GB/GPU)
Питание	130 W	225 W
Quadro аналог	Quadro K600	Quadro K5000

NVIDIA

GPU - ответ на вызовы будущего

- Максимальная производительность и энерго-эффектинвость
- Большое количество ПО и библиотек
- Графический потенциал

NVIDIA®

АНТОН ДЖОРАЕВ
adzhoraev@nvidia.com

Recompile and Run..?

Если это так просто – почему до сих пор нет тысяч приложений поддерживающих Phi?