

УСКОРЕННЫЕ ВЫЧИСЛЕНИЯ - ПУТЬ В БУДУЩЕЕ

Дмитрий Конягин

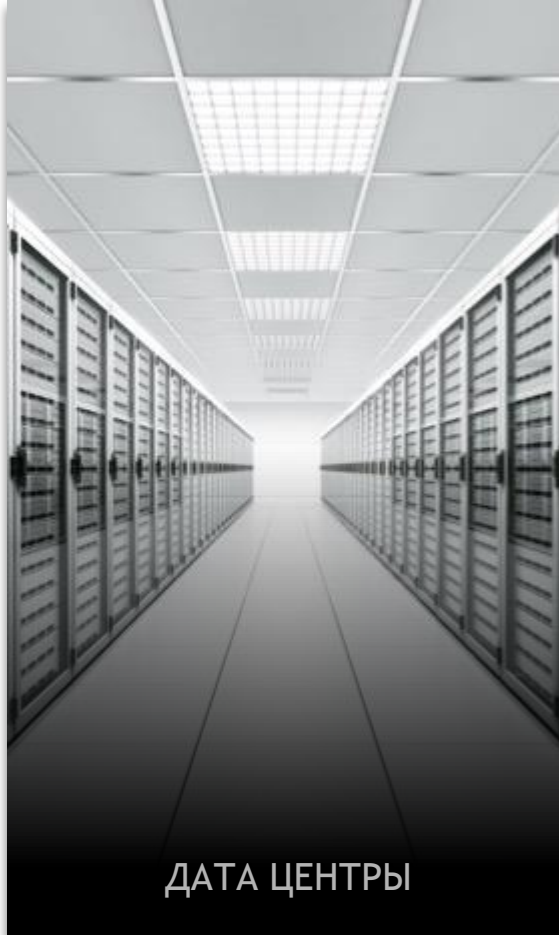




ГЕЙМИНГ



ПРОФ ВИЗУАЛИЗАЦИЯ



ДАТА ЦЕНТРЫ



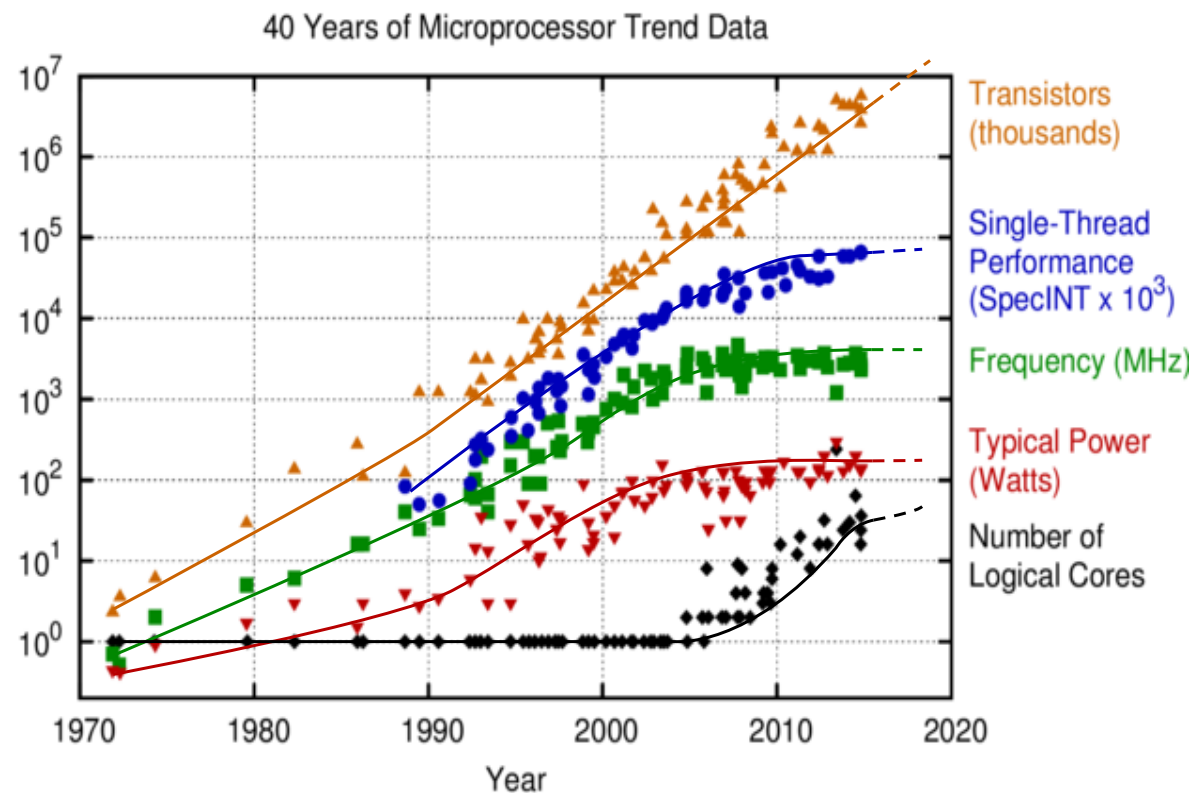
АВТО

ЛИДЕР В ОБЛАСТИ ВИЗУАЛЬНЫХ ВЫЧИСЛЕНИЙ

УСКОРЕННЫЕ ВЫЧИСЛЕНИЯ - ПУТЬ В БУДУЩЕЕ

“Пора начать думать о конце закона Мура. Вопрос уже не в том, когда это произойдет, а в том что делать после.”

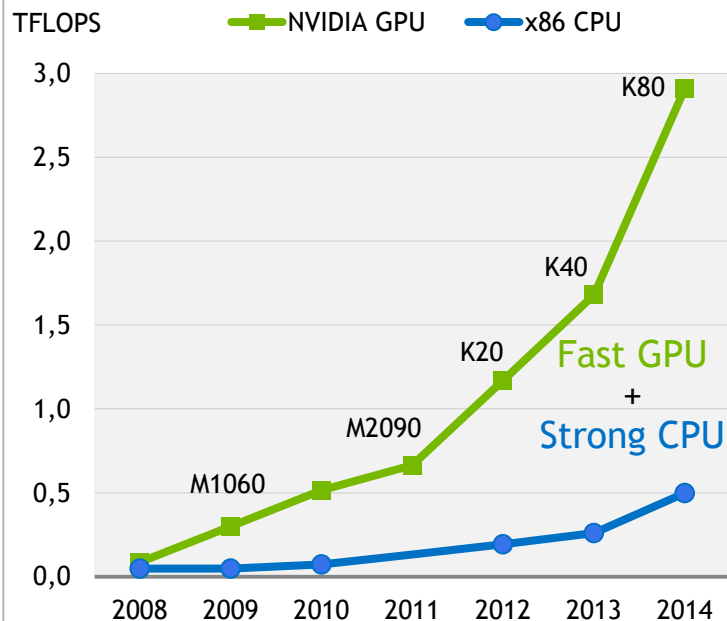
Robert Colwell
Director, Microsystems Technology Office, DARPA



Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2015 by K. Rupp

NVIDIA УСКОРЯЕТ ВЫЧИСЛЕНИЯ

Быстрые GPU для максимальной производительности



Эффективная модель программирования и инструменты



Совместные разработки

ПРИЛОЖЕНИЯ

ПРОМЕЖУТОЧНОЕ ПО

СИСТЕМНОЕ ПО

БОЛЬШИЕ СИСТЕМЫ

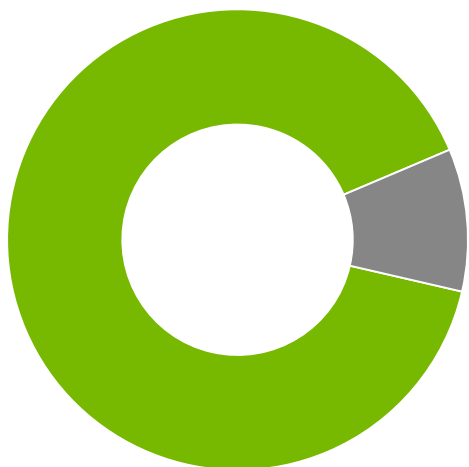
ПРОЦЕССОРЫ

Доступность



70% КЛЮЧЕВЫХ НРС ПРИЛОЖЕНИЙ УЖЕ УСКОРЯЮТСЯ НА GPU

ИССЛЕДОВАНИЕ КЛЮЧЕВЫХ ПРИЛОЖЕНИЙ ОТ INTERSECT360



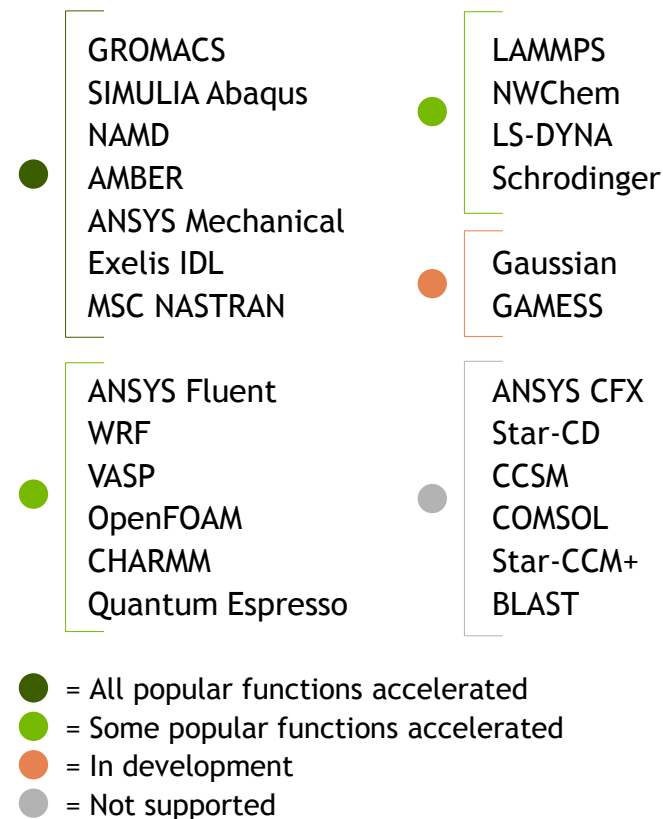
9 из топ 10 приложений ускоряются



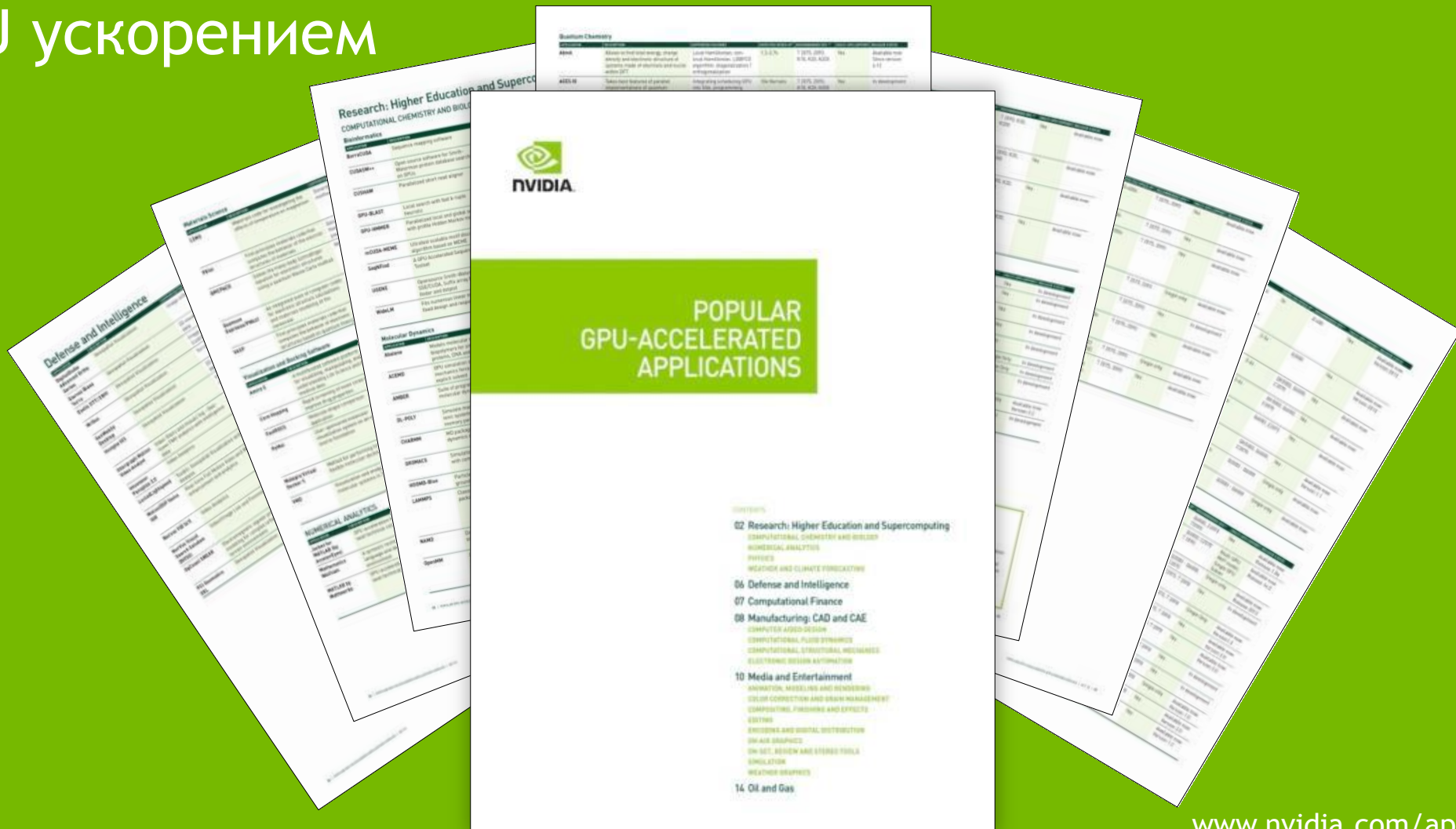
35 из топ 50 приложений ускоряются

Intersect360, Nov 2015
“HPC Application Support for GPU Computing”

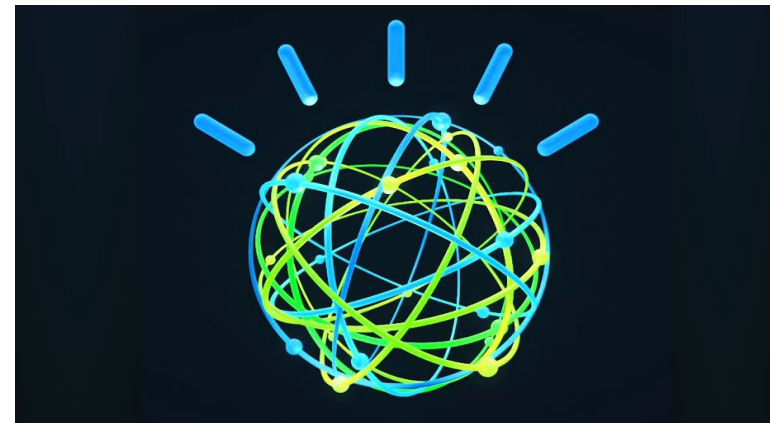
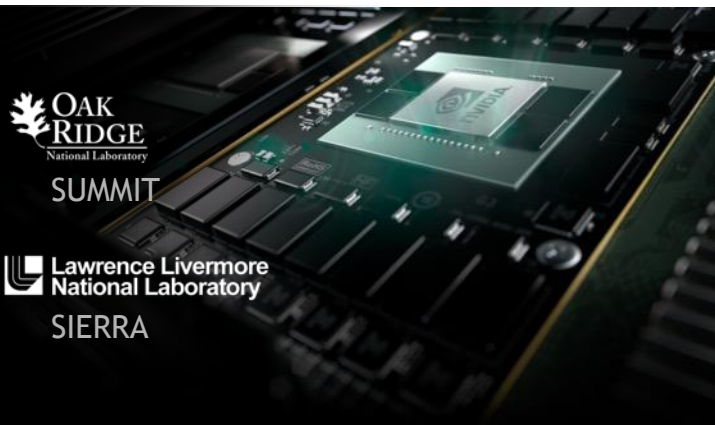
TOP 25 ПРИЛОЖЕНИЙ



400+ приложений с GPU ускорением



СУПЕРКОМПЬЮТЕРЫ СЛЕДУЮЩЕГО ПОКОЛЕНИЯ УСКОРЯЮТСЯ GPU



Минэнерго США

Пре-Exascale суперкомпьютеры
для научных исследований

NOAA

Новый суперкомпьютер для новой
прогностической модели

IBM Watson

Прорыв в обработке запросов на
естественном языке

НРС & ГЛУБОКОЕ ОБУЧЕНИЕ ВЗАИМОДОПОЛНЯЮТ ДРУГ ДРУГА

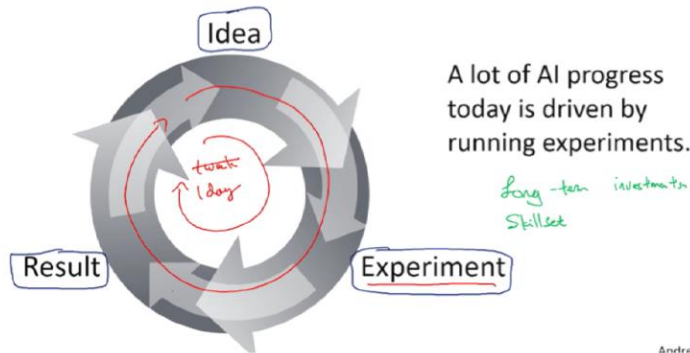
Why HPC is
speeding up
machine learning
research



“Investments in computer systems – and I think the bleeding-edge of AI, and deep learning specifically, is shifting to HPC – can cut down the time to run an experiment, and therefore go around the circle, from a week to a day and sometimes even faster.”

– Andrew Ng, Baidu

Iterating quickly is important for research progress



“...deep learning and cognitively enabled applications are driving large-scale high-performance computing (HPC) projects that are heavier on GPUs. IDC expects major advances and potential large build-outs...”

– IDC



THE LEARNING MACHINES

Using massive amounts of data to recognize photos and speech, deep-learning computers are taking a big step towards true artificial intelligence.

BY NICOLA JONES

Three years ago, researchers at the secretive Google X lab in Mountain View, California, extracted some 10 million still images from YouTube videos and fed them into Google Brain — a network of 1,000 computers programmed to soak up the world much as a human toddler does. After three days looking for recurring patterns, Google Brain decided, all on its own, that there were certain repeating categories it could identify: human faces, human bodies and ... cats?

Google Brain's discovery that the Internet is full of cat videos provoked a flurry of jokes from journalists. But it was also a landmark in the resurgence of deep learning: a three-decade-old technique in which massive amounts of data and processing power

help computers to crack messy problems that humans solve almost intuitively, from recognizing faces to understanding language.

Deep learning itself is a revival of an even older idea for computing: neural networks. These systems, loosely inspired by the densely interconnected neurons of the brain, mimic human learning by changing the strength of simulated neural connections on the basis of experience. Google Brain, with about 1 million simulated neurons and 1 billion simulated connections, was ten times larger than any deep neural network before it. Project founder Andrew Ng, now director of the Artificial Intelligence Laboratory at Stanford University in California, has gone on to make deep-learning systems ten times larger again.

Such advances make for exciting times in

Tesla преобразует глубокое обучение

ПРИЛОЖЕНИЕ GOOGLE BRAIN - ГЛУБКОЕ ОБУЧЕНИЕ

	ДО TESLA	С TESLA
Цена	\$5M	\$0.2M
Серверы	1000 серверов	16 Tesla серверов
Потребление	600 KW	4 KW
Производительность	1x	6x

ПОЛНОЕ СЕМЕЙСТВО ПРОДУКТОВ TESLA

HYPERSCALE HPC

Tesla M4, M40



hyperscale решения для
глубокого обучения,
инференса, обработки
изображений и видео

MIXED-APPS HPC

Tesla K80



HPC дата-центры со широким
набором задач для CPU и GPU

STRONG-SCALING HPC

Tesla P100



hyperscale & HPC дата-
центры с задачами хорошо
масштабируемыми на GPU

FULLY INTEGRATED DL SUPERCOMPUTER

DGX-1



Полностью интегрированное
решение по принципу
включи-и-работай

ПЛАТФОРМА TESLA ДЛЯ МАСШТАБИРУЕМЫХ НРС ПРИЛОЖЕНИЙ

ЦОД СЕГОДНЯ

Хорошо подходит для транзакционных задач, использующих множество узлов



Традиционные вычислительные узлы со значительными накладными расходами на сетевое взаимодействие

ИДЕАЛ

Для важных задач с бесконечными требованиями к производительности



Несколько супер узлов с производительностью тысяч традиционных узлов

TESLA P100

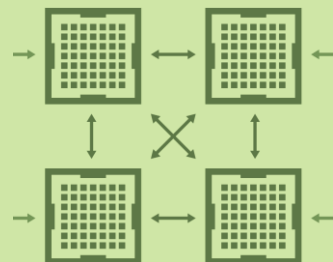
Новая архитектура GPU для самых быстрых вычислительных узлов

Архитектура Pascal



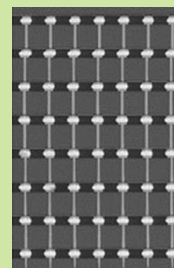
Максимальная
производительность

NVLink



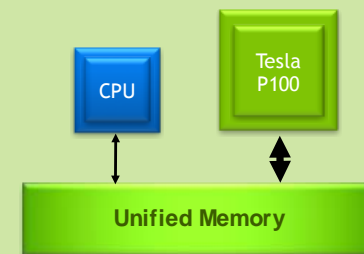
GPU интерконнект для
максимальной масштабируемости

CoWoS HBM2



Объединение вычислителя и
памяти на одной подложке

Page Migration Engine



Упрощение параллельного
программирования с
виртуально общей памятью

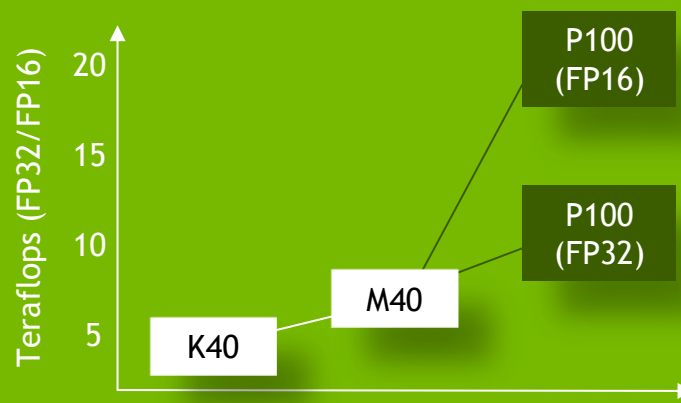


РАДИКАЛЬНЫЕ УЛУЧШЕНИЯ

ВО ВСЕМ

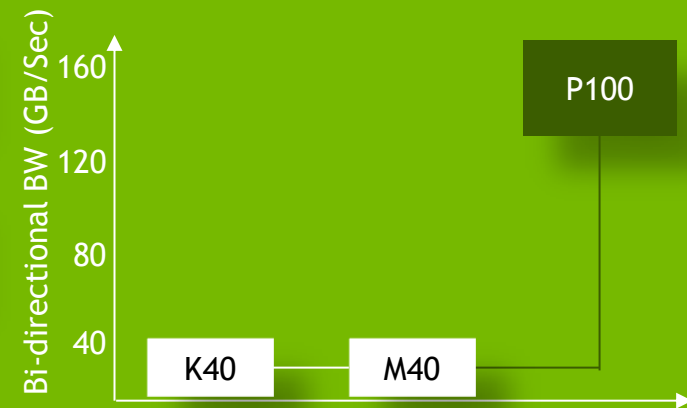
АРХИТЕКТУРА PASCAL

21 Tf в FP16 для глубокого обучения



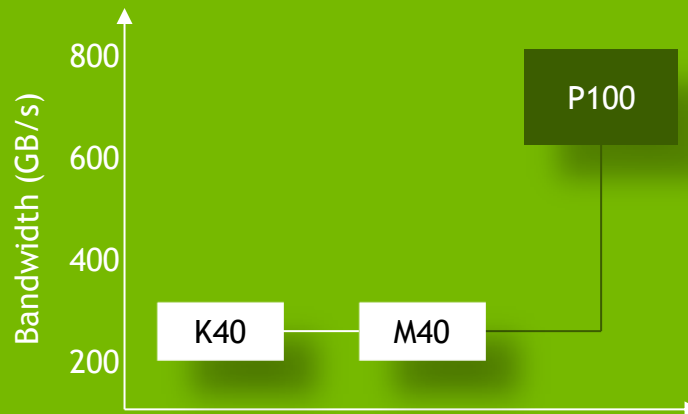
NVLINK

5x GPU-GPU пропускная способность



HBM2 стекируемая память

3x пропускная способность



PAGE MIGRATION ENGINE

Виртуально неограниченное адресное пространство памяти



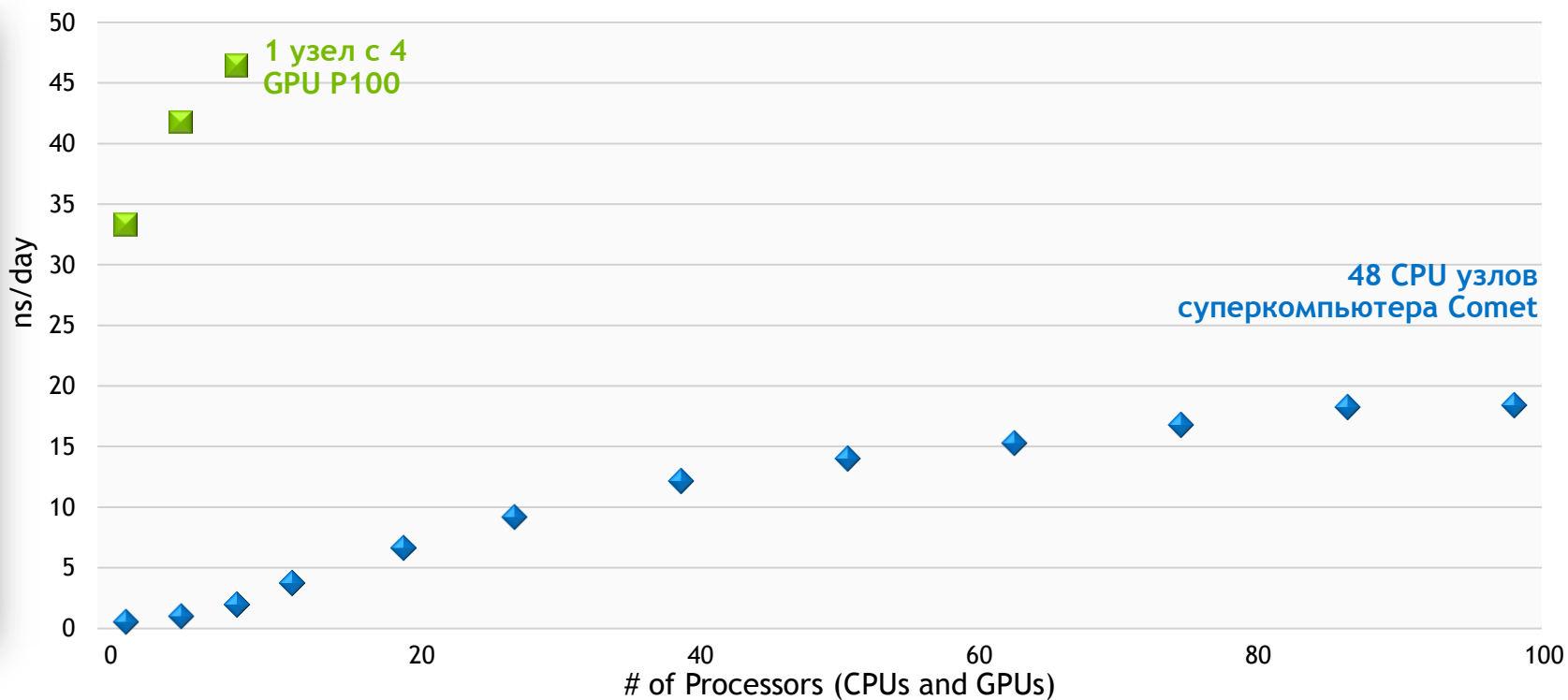
СЕРЬЕЗНЫЕ ЗАДАЧИ ТРЕБУЮТ БЫСТРЫХ КОМПЬЮТЕРОВ

В 2.5 раза быстрее, чем целый ЦОД с CPU



“Biotech discovery of the century”
-MIT Technology Review 12/2014

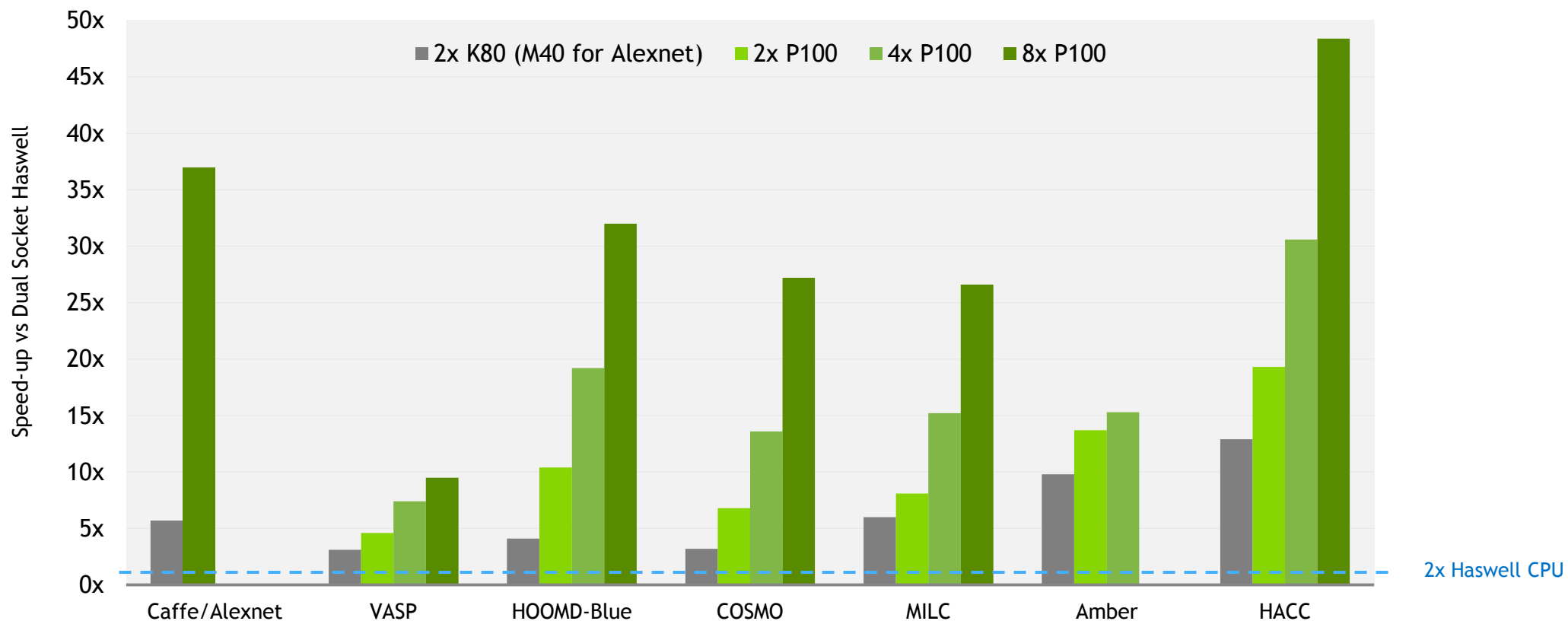
AMBER Simulation of CRISPR, Nature’s Tool for Genome Editing



AMBER 16 Pre-release, CRISPR based on PDB ID 5f9r, 336,898 atoms
CPU: Dual Socket Intel E5-2680v3 12 cores, 128 GB DDR4 per node, FDR IB

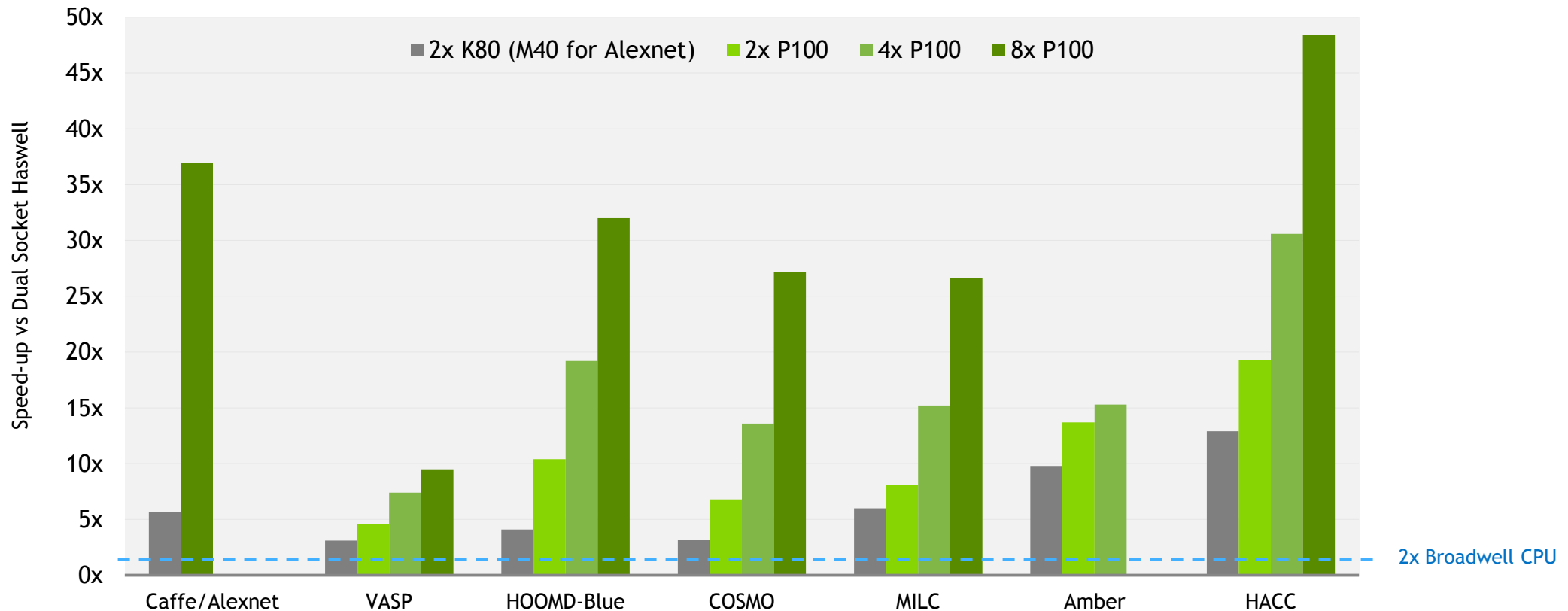
АБСОЛЮТНАЯ ПРОИЗВОДИТЕЛЬНОСТЬ

NVLink для максимальной масштабируемости, в 45 раз быстрее с 8x P100



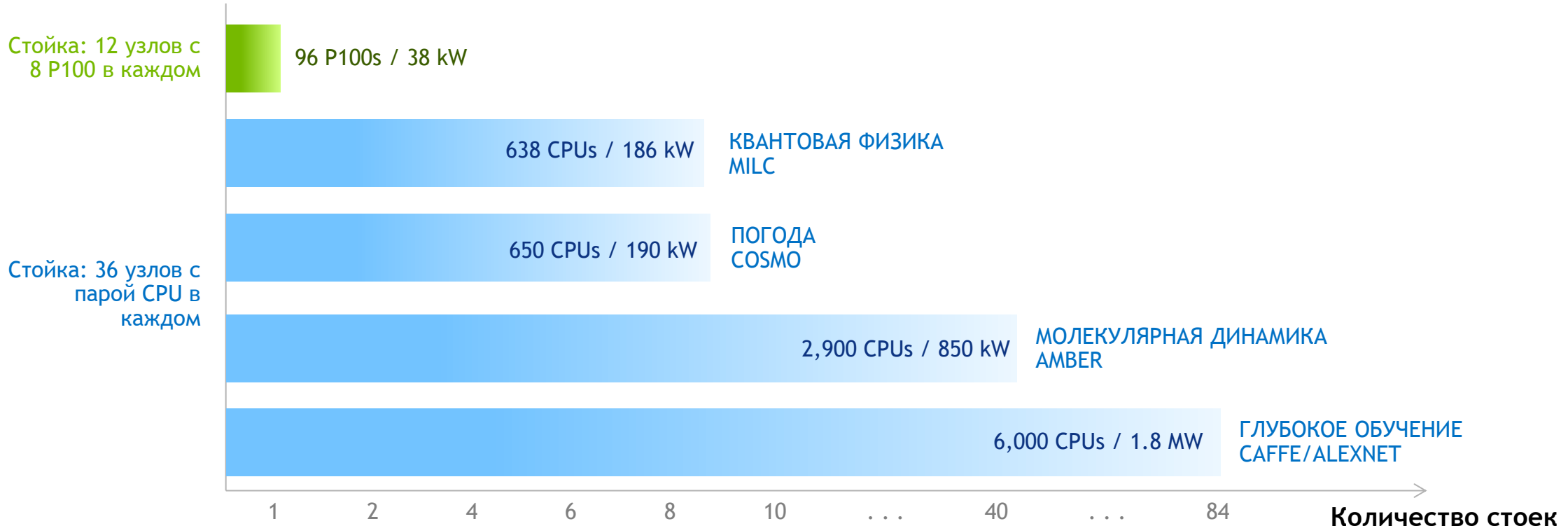
АБСОЛЮТНАЯ ПРОИЗВОДИТЕЛЬНОСТЬ

NVLink для максимальной масштабируемости, в 45 раз быстрее с 8x P100



ЦОД В ОДНОЙ СТОЙКЕ

1 стойка с Tesla P100 эквивалентна ЦОД с 6000 CPU



ПЕРВЫЕ СУПЕРКОМПЬЮТЕРЫ С P100 ДЛЯ РЕШЕНИЯ ГЛОБАЛЬНЫХ ПРОБЛЕМ



CSCS

Ведущий европейский СК для пользователей CERN, Human Brain и др



NOAA

Новый СК для новой модели прогнозирования погоды

УСКОРИТЕЛЬ TESLA P100



Производительность	5.3 TF DP · 10.6 TF SP · 21.2 TF HP
Память	HBM2: 720 GB/s · 16 GB
Интерконнект	NVLink (up to 8 way) + PCIe Gen3
Программируемость	Page Migration Engine Унифицированная память
Доступность	DGX-1: доступен для заказа Cray, Dell, HP, IBM: Q1 2017

**ПОЛНОСТЬЮ ИНТЕГРИРОВАННЫЙ
СУПЕРКОМПЬЮТЕР
ДЛЯ ГЛУБОКОГО ОБУЧЕНИЯ**

NVIDIA DGX-1

ПЕРВЫЙ В МИРЕ СУПЕРКОМПЬЮТЕР ДЛЯ ГЛУБОКОГО ОБУЧЕНИЯ



170 TFLOPS FP16

8x Tesla P100 16GB

NVLink Hybrid Cube Mesh

Ускорение основных DL-фреймворков

Dual Xeon

7 TB SSD

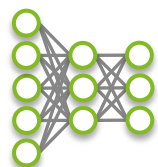
Dual 10GbE, Quad IB 100Gb

3RU - 3200W

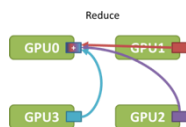
NVIDIA DGX-1 СОФТВЕРНЫЙ СТЕК

Оптимизирован для обеспечения производительности задач глубокого обучения

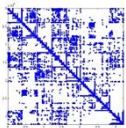
Ускорение глубокого обучения



cuDNN



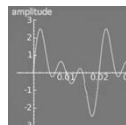
NCCL



cuSPARSE

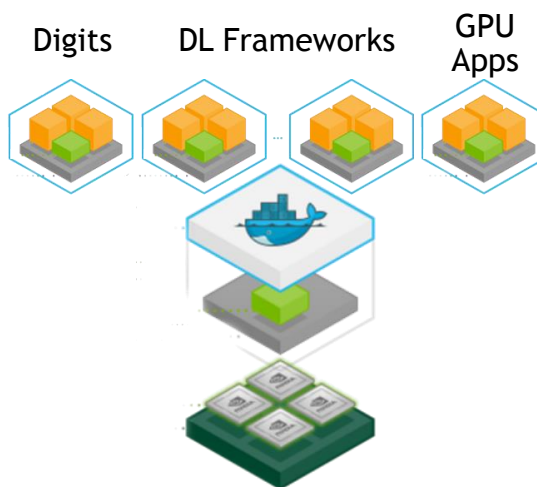


cuBLAS



cuFFT

Контейнеризация приложений



Управление через облако NVIDIA



УСКОРЕНИЕ ВСЕХ ФРЕЙМВОРКОВ

ACADEMIA

CAFFE 	TORCH 
THEANO 	MATCONVNET 
MOCHA.JL 	PURINE  
MINERVA   	MXNET*  

TENSORFLOW

Google





TORCH

facebook.

CNTK


Microsoft

START-UPS

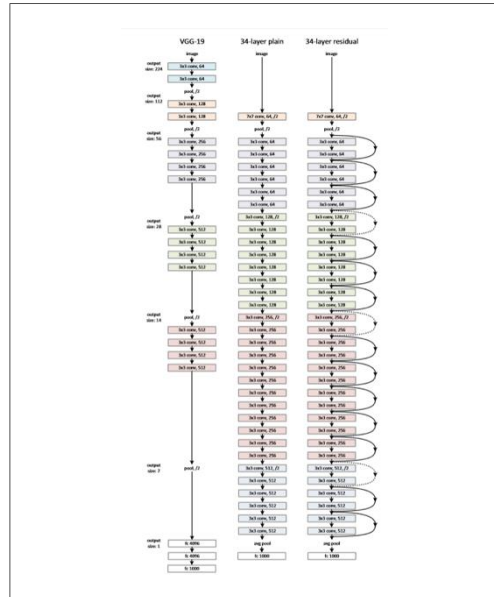
CHAINER 	DL4J 	KERAS  SCHULTS LABORATORIES	OPENDEEP  VITRUVIAN
---	--	--	---

NVIDIA DGX-1

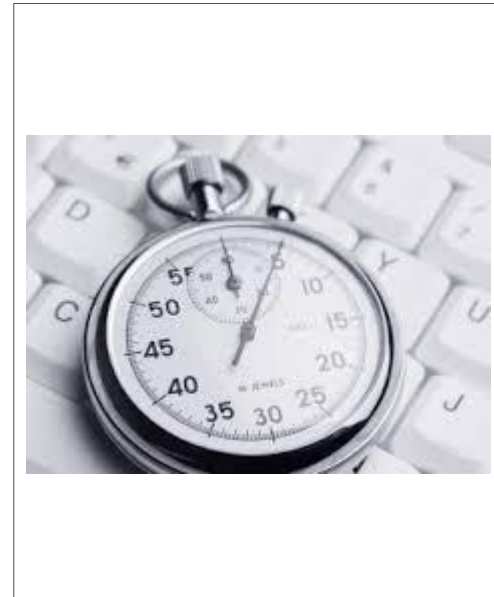
ПРЕИМУЩЕСТВА ДЛЯ ИССЛЕДОВАТЕЛЕЙ



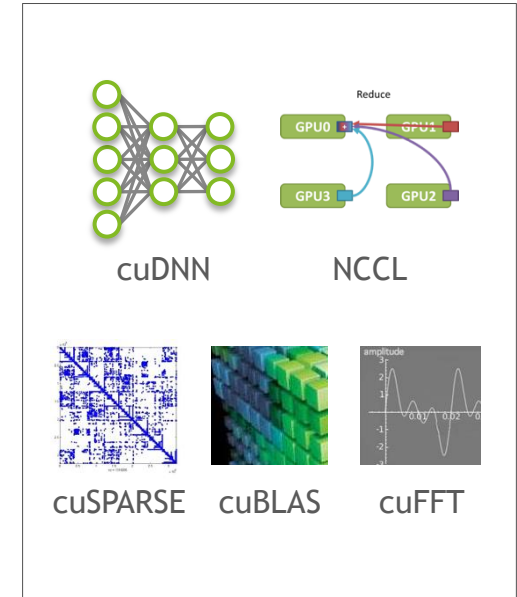
Самый быстрый СК для глубокого обучения



Создание больших сетей



Сокращение времени обучения

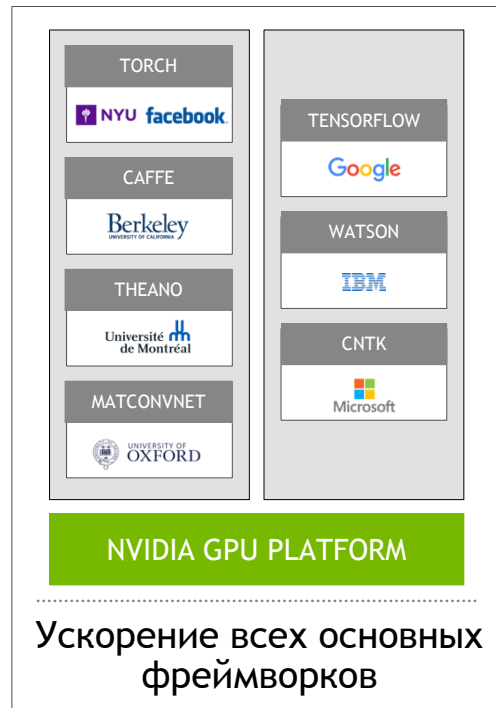


Постоянные обновления SDK для глубокого обучения

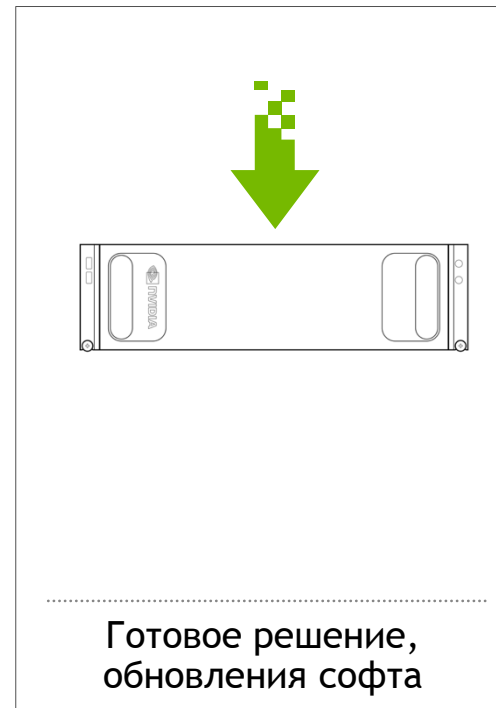
ПРЕИМУЩЕСТВА ДЛЯ ИНДУСТРИАЛЬНЫХ ПОЛЬЗОВАТЕЛЕЙ



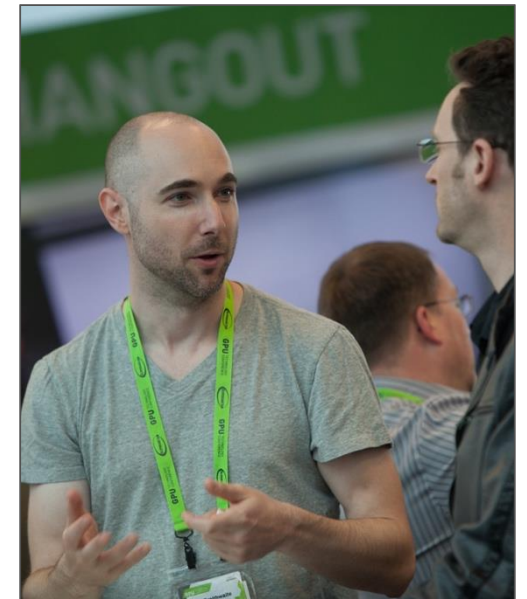
Специализированное решение



Ускорение всех основных фреймворков



Готовое решение, обновления софта

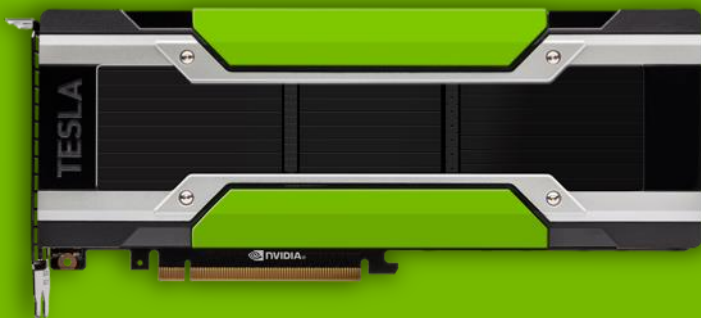


Поддержка экспертов NVIDIA

ПЛАТФОРМА TESLA ДЛЯ РАЗНОРОДНЫХ НРС ПРИЛОЖЕНИЙ

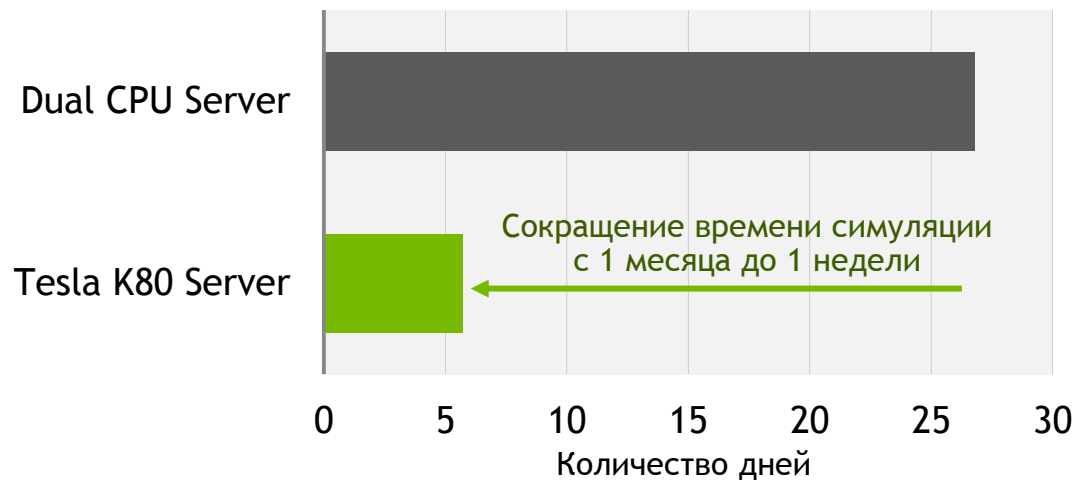
TESLA K80

Самый быстрый ускоритель
для HPC



В 5 раз быстрее

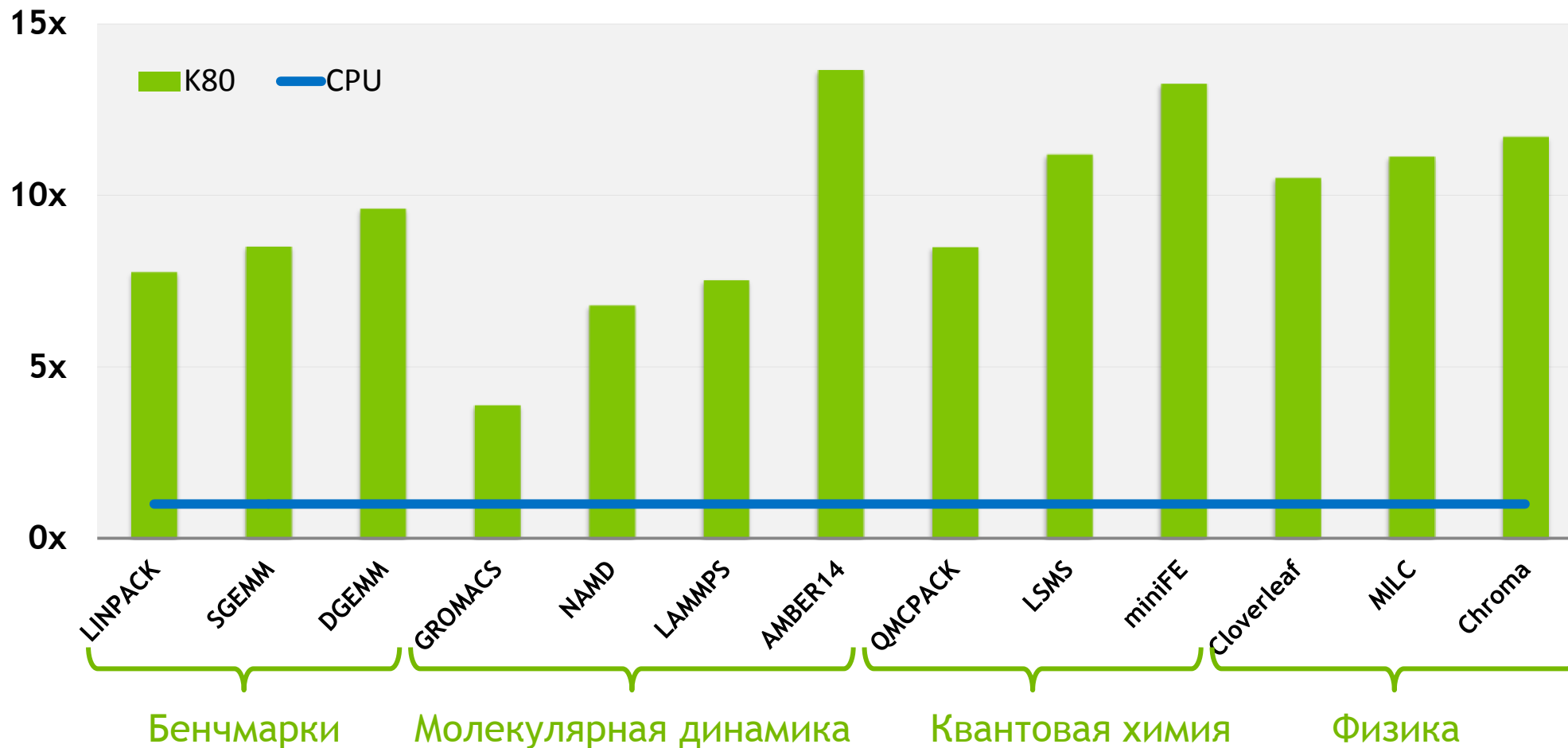
Производительность в AMBER



CUDA ядра	4992
DP пик	1.9 TFLOPS
DP пик с Boost	2.9 TFLOPS
GDDR5 память	24 GB
Пропускная способность	480 GB/s
Потребление	300 W

CPU: E5-2698v3 @ 2.3GHz. 64GB System Memory, CentOS 6.2

TESLA K80: В 10 РАЗ БЫСТРЕЕ В РЕАЛЬНЫХ ПРИЛОЖЕНИЯХ



TESLA K80 ИДЕАЛЬНО ПОДХОДИТ ДЛЯ HPC ЦЕНТРОВ С РАЗНОРОДНЫМИ ПРИЛОЖЕНИЯМИ

ПРИЛОЖЕНИЕ	УЗЕЛ С ПАРОЙ CPU	УЗЕЛ С ПАРОЙ K80	УЗЕЛ С ПАРОЙ P100
MILC (основная нагрузка на GPU)	6 часов	1 час (6x)	0.8 часа (8x)
AMBER (основная нагрузка на GPU)	6 часов	0.6 часа (10x)	0.4 часа (14x)
NAMD (использует и CPU и GPU)	6 часов	1 час(6x)	1 час (6x)
NWChem (использует только CPU)	6 часов	6 часов	6 часов
Общее время & пропускная способность	24 часа	8.6 часов (2.8x)	8.2 часа (2.9x)
Пропускная способность / \$	Низкая	Высокая	Средняя

ЭКОНОМИЯ НА ЦОД ДО 60%

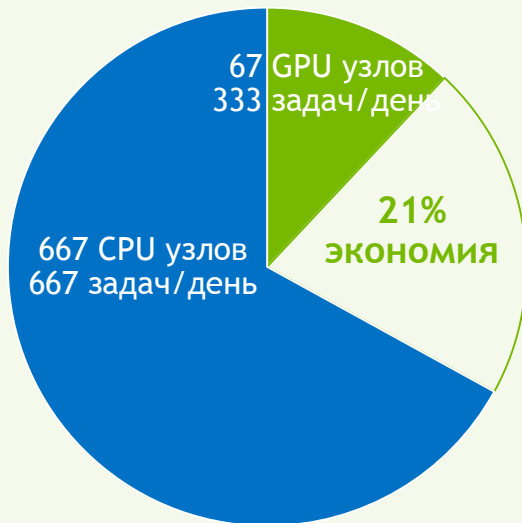
ЦОД только с CPU

1,000 CPU узлов
1,000 задач/день

Бюджет: \$10М

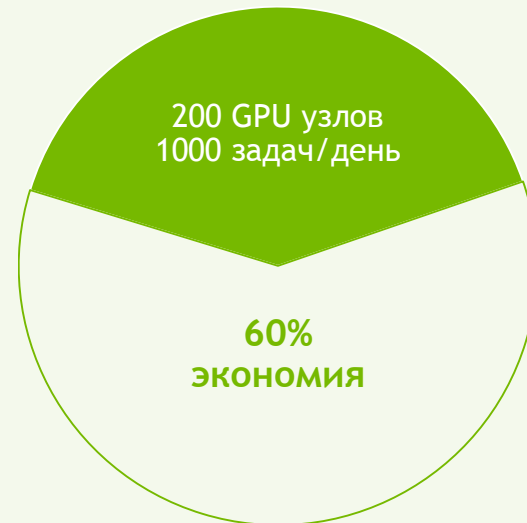
ЦОД, ускоряемый с GPU

33% приложений
в 5 раз быстрее на GPU



Бюджет: \$7.9М

100% приложений
в 5 раз GPU



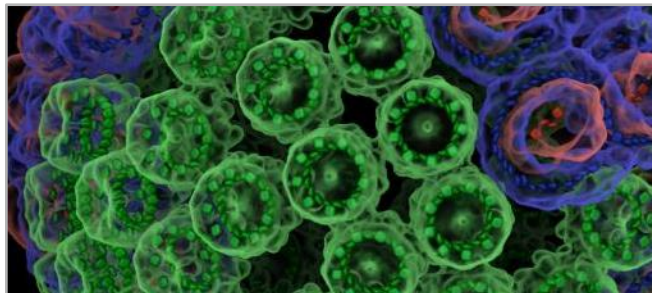
Бюджет: \$4.0М

TESLA ДЛЯ ВИЗУАЛИЗАЦИИ

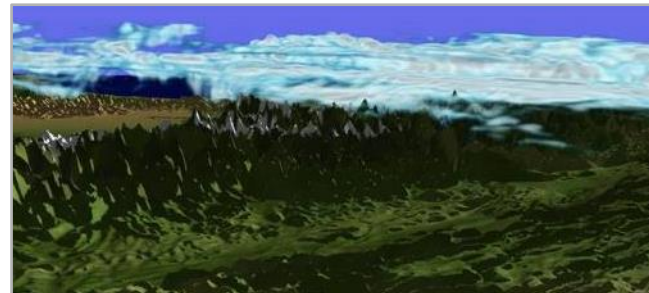
IRAY



OPTIX



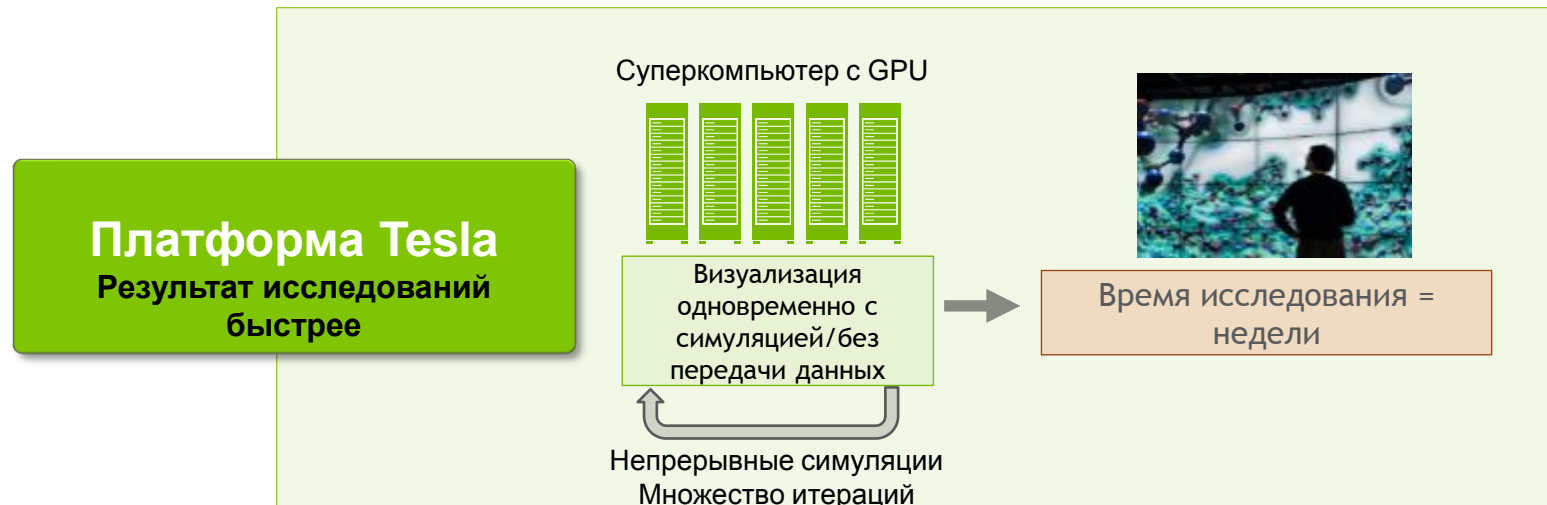
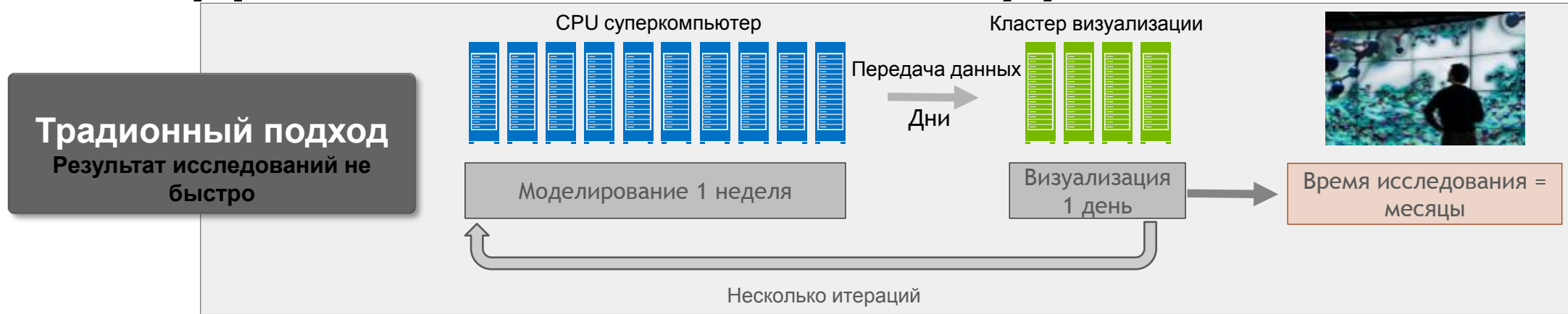
INDEX



ИНСТРУМЕНТЫ ВИЗУАЛИЗАЦИИ ДЛЯ HPC

ВЫЧИСЛИТЕЛЬНАЯ ПЛАТФОРМА TESLA

МГНОВЕННАЯ ВИЗУАЛИЗАЦИЯ ДАННЫХ ДЛЯ УСКОРЕНИЯ ИССЛЕДОВАНИЙ



Интерактивность

Масштабируемость

Гибкость

СУПЕРКОМПЬЮТЕРЫ С ФУНКЦИОНАЛОМ ДЛЯ ВИЗУАЛИЗАЦИИ

Симуляция + визуализация

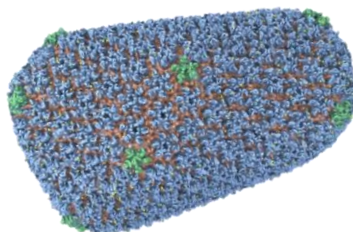
CSCS Piz Daint



Формирование
Галактик

<http://blogs.nvidia.com/blog/2014/11/19/gpu-in-situ-milky-way/>

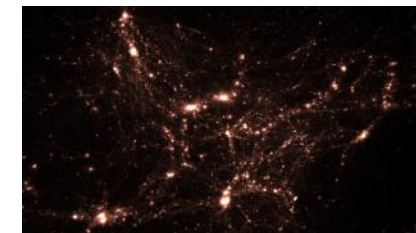
NCSA Blue Waters



Молекулярная
динамика

<http://devblogs.nvidia.com/parallelforall/hpc-visualization-nvidia-tesla-gpus/>

ORNL Titan



Космология

<http://www.sdav-scidac.org/29-highlights/visualization/66-accelerated-cosmology-data-anal.html>

РОСТ ПОПУЛЯРНОСТИ GPU ДЛЯ ПРОГНОЗИРОВАНИЯ ПОГОДЫ И КЛИМАТА



MeteoSwiss запускает первый в мире СК с GPU для прогнозирования погоды

В 2 раза выше разрешение сетки для ежедневного прогноза

В 14 раз большее количество симуляций для ансамблевого метода расчета среднесрочных прогнозов



NOAA выбирает Tesla для улучшения прогнозов

Разработка глобальной модели с разрешением 3 км, что в 5 раз больше, чем сегодня

Повышение разрешения увеличивает вычислительную сложность в 40 раз



США ПОСТРОЯТ ДВА ФЛАГМАНСКИХ СК

на базе платформы Tesla



100-300 PFLOPS пик

10x ускорение научных приложений

IBM POWER9 CPU + NVIDIA Volta GPU

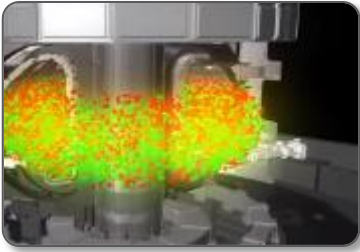
Высокоскоростной интерфейс NVLink

40 TFLOPS на узел, >3400 узлов

2017

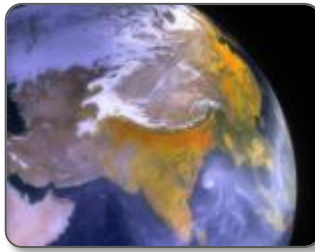
Значимый шаг на пути к экзафлопсу

CORAL: ДЛЯ РЕШЕНИЯ ВАЖНЕЙШИХ НАУЧНЫХ ПРОБЛЕМ



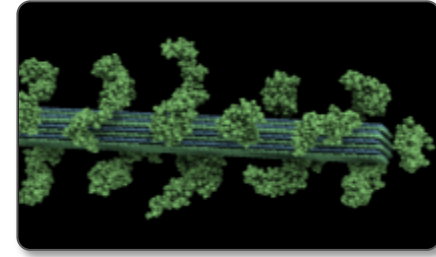
Термоядерный синтез

Роль материалов с неупорядоченной структурой, статистика и колебания в наноструктурах.



Изменения климата

Исследование изменений климата и моделирование их последствий

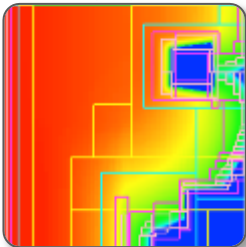


Биотопливо

Поиск возобновляемых и более эффективных источников топлива

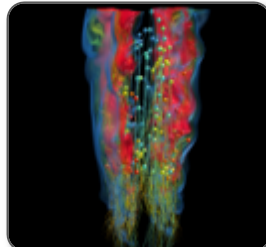
Астрофизика

Радиационный перенос – критический элемент для астрофизики, лазерного термоядерного синтеза, динамики атмосферы и медицинской визуализации



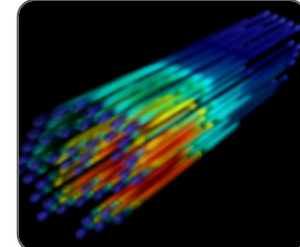
Сгорание топлива

Моделирование горения для повышения эффективности биотоплива



Атомная энергетика

Беспрецедентно точные расчеты радиационного переноса для приложений атомной энергетики



ПЛАТФОРМА TESLA ДЛЯ HYPERSCALE

ЭКЗАБАЙТЫ КОНТЕНТА СОЗДАЮТСЯ ЕЖЕДНЕВНО

Пользовательский контент доминирует в веб-сервисах

10млн пользователей
40 лет видео/день



1.7 млн. передач
Пользователи смотрят 1.5 часа/день



6 млрд запросов/день
10% голосом



270млн предметов продано/день
43% на мобильных устройствах



8 млрд просмотров видео/день
Рост 400% за 6 мес.



300 часов видео/минуту
50% на мобильных устройствах





ВЫЗОВ

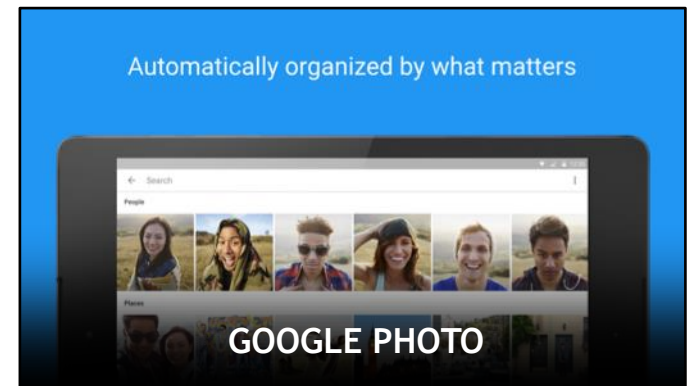
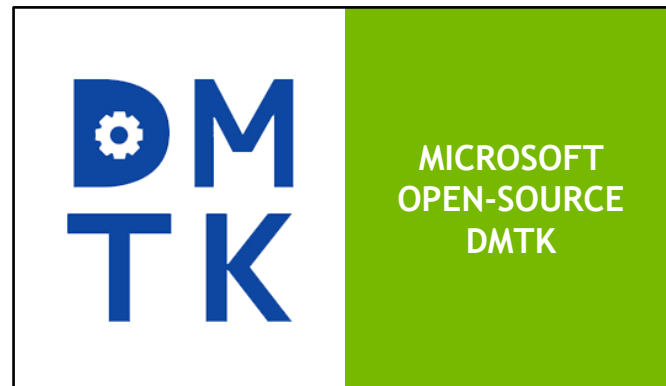
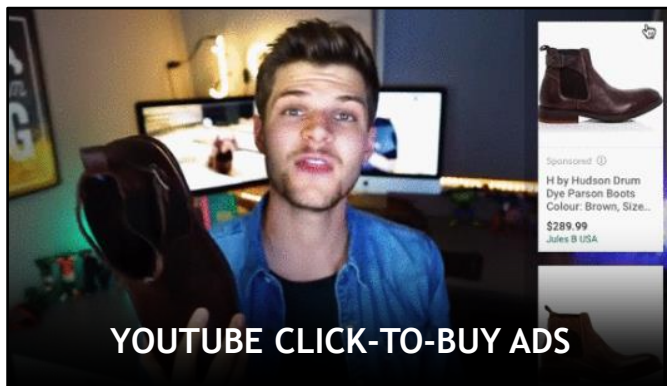
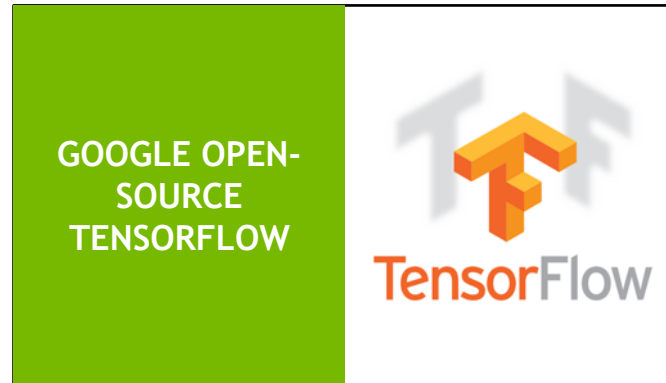
Обуздать цунами данных
в реальном времени

Миллионы пользователей делятся
друг с другом миллионами
изображений и видео

Распознавание и улучшение каждой
секунды видео

Мгновенный показ релевантной
рекламы

МАШИННОЕ ОБУЧЕНИЕ - НОВАЯ ВЕХА В ИРС



TESLA ДЛЯ HYPERSCALE

HYPERSCALE SUITE



GPU REST Engine



GPU Accelerated
FFmpeg



Image Compute
Engine

TESLA M40

ПРОИЗВОДИТЕЛЬНОСТЬ: максимальная
производительность для глубокого обучения



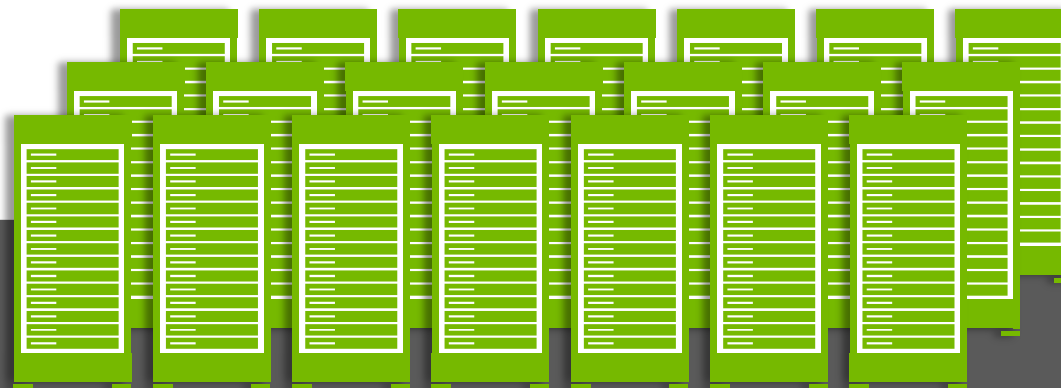
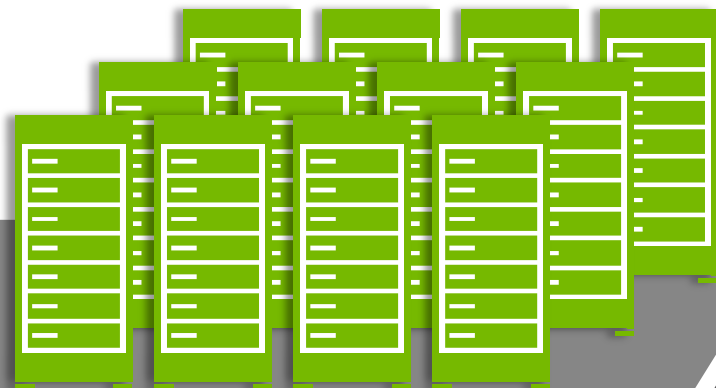
TESLA M4

НИЗКОЕ ЭНЕРГОПОТРЕБЛЕНИЕ: максимальная
пропускная способность для hyperscale задач



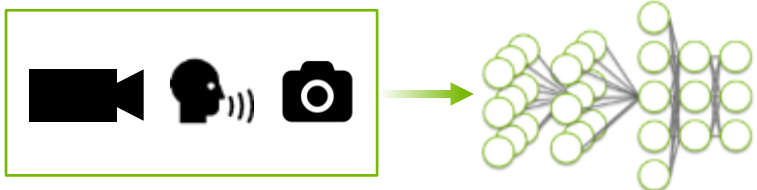
HYPERSCALE ЦОД С GPU-УСКОРИТЕЛЯМИ

Платформа Tesla



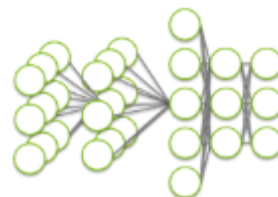
СЕРВЕРЫ ДЛЯ ОБУЧЕНИЯ
Масштабирование с ростом данных

СЕРВЕРЫ ДЛЯ ИНФЕРЕНСА, WEB СЕРВИСОВ
Масштабирование с ростом числа пользователей



Экзбайты контента / день

Обучение модели



Обученная модель на каждом сервере



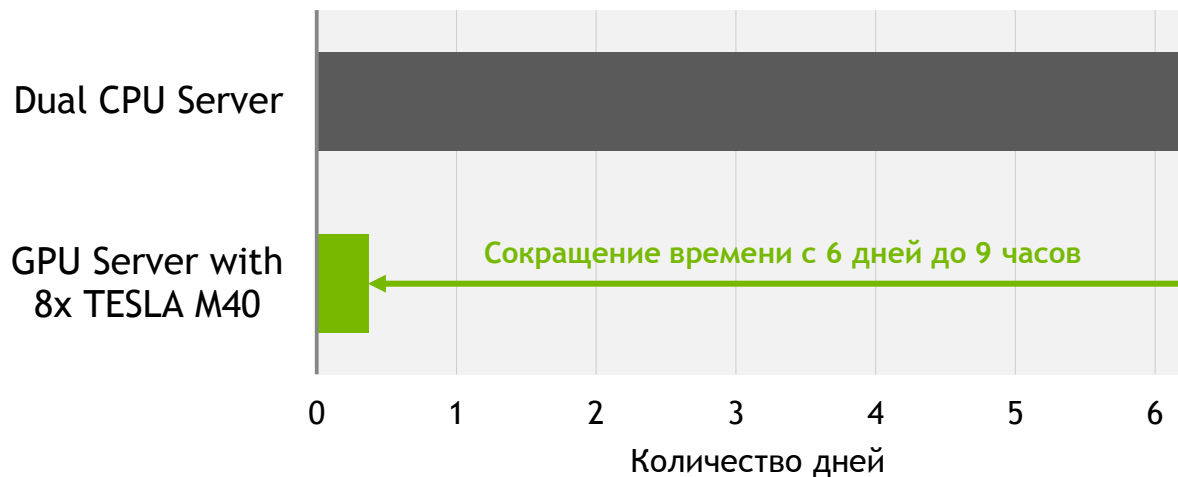
Миллиарды устройств

TESLA M40

Самый быстрый в мире ускоритель для глубокого обучения



В 17 раз быстрее обучение Alexnet

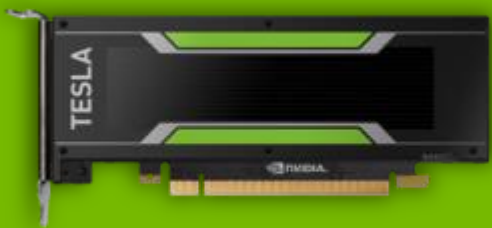


GDDR5 память	24 GB
CUDA ядра	3072
SP производительность	7 TFLOPS
Пропускная способность	288 GB/s
Потребление	250W

Caffe benchmark with AlexNet
CPU server uses dual Xeon E5-2698v3, Ubuntu 14.04

TESLA M4

Высокоэффективный ускоритель hyperscale задач



20 изображений/сек/Вт

Обработка видео

4x

Стабилизация и улучшение



Обработка изображений

5x

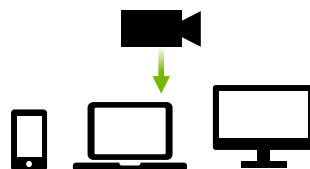
Resize, фильтр, поиск, улучшение



Транскодирование видео

2x

H.264 & H.265, SD & HD



Инференс

2x



CUDA ядра

1024

SP производительность

2.2 TFLOPS

GDDR5 память

4 GB

Пропускная способность

88 GB/s

Форм-фактор

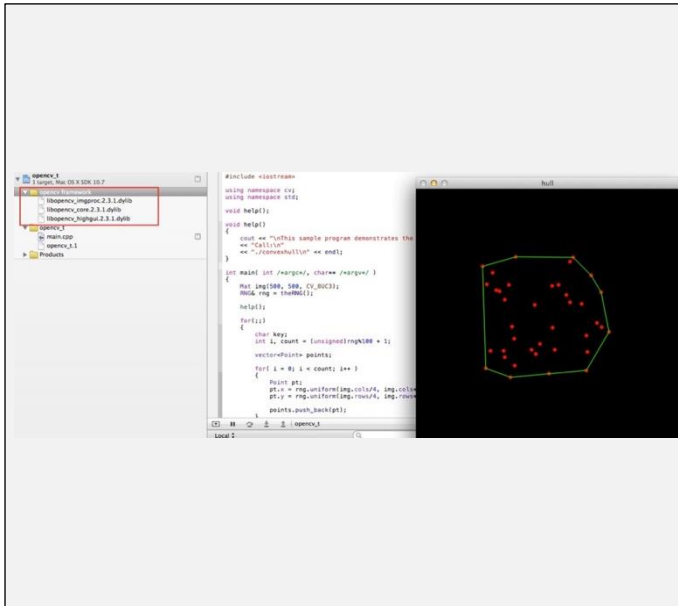
PCIe Low Profile

Потребление

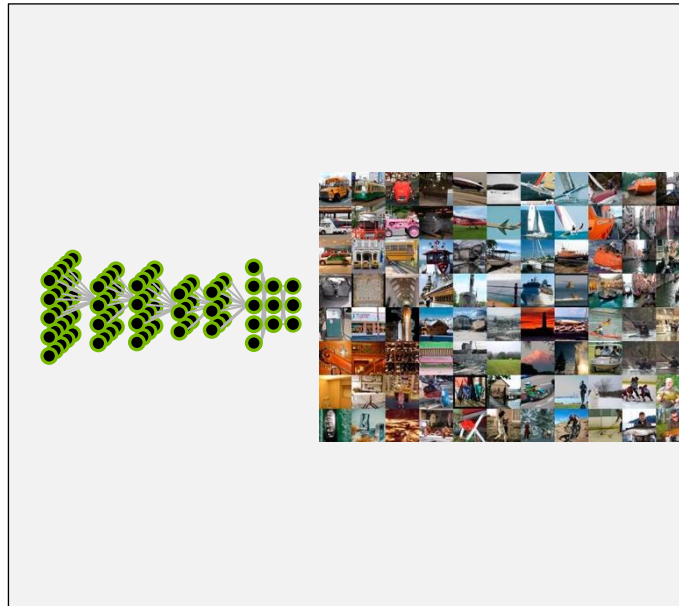
50 - 75 W

ПЛАТФОРМА TESLA ДЛЯ ГЛУБОКОГО ОБУЧЕНИЯ

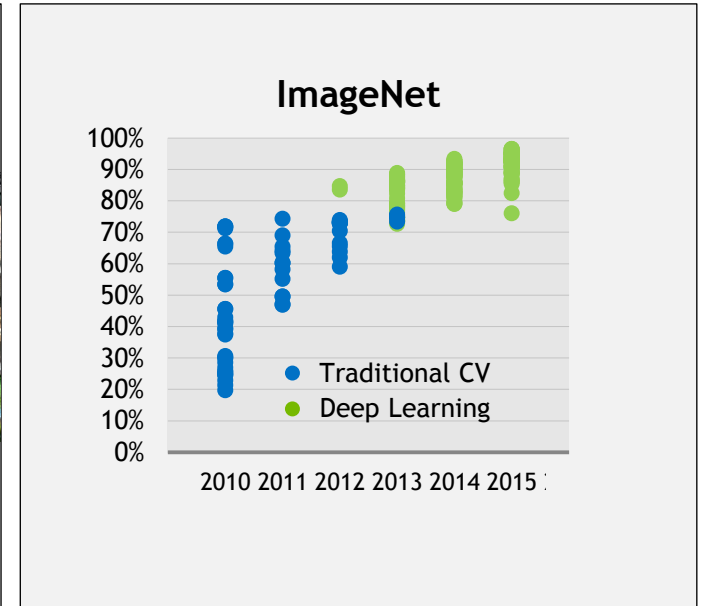
НОВАЯ ПАРАДИГМА ВЫЧИСЛЕНИЙ



Традиционное компьютерное зрение
Эксперты + Время

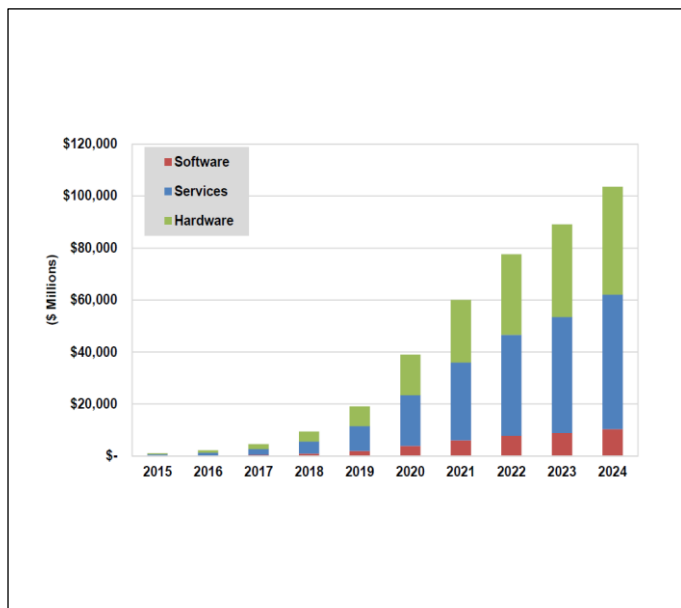


Классификация объектов с
помощью глубокого обучения
DNN + данные + HPC

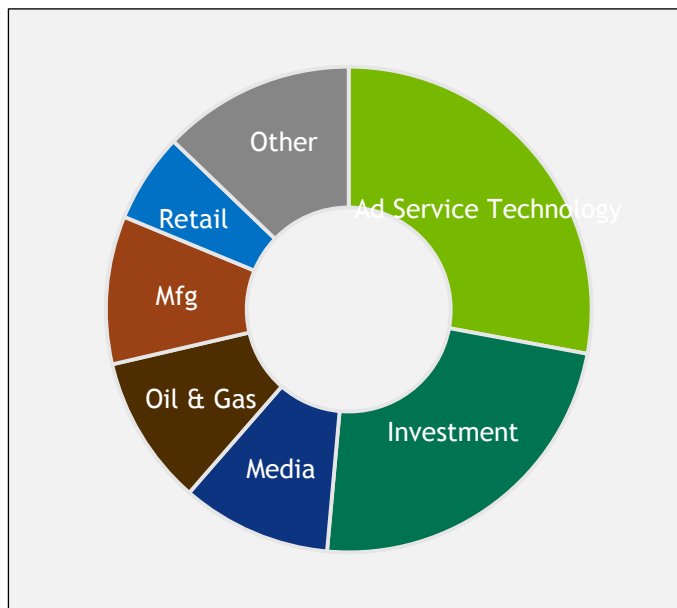


Глубокое обучение
превзошло результаты
человека

ИНДУСТРИЯ ОБЪЕМОМ В \$500 МЛРД В ТЕЧЕНИЕ 10 БЛИЖАЙШИХ ЛЕТ



Общий оборот индустрии глубокого обучения по сегментам



Оборот по софту для глубокого обучения по индустриям



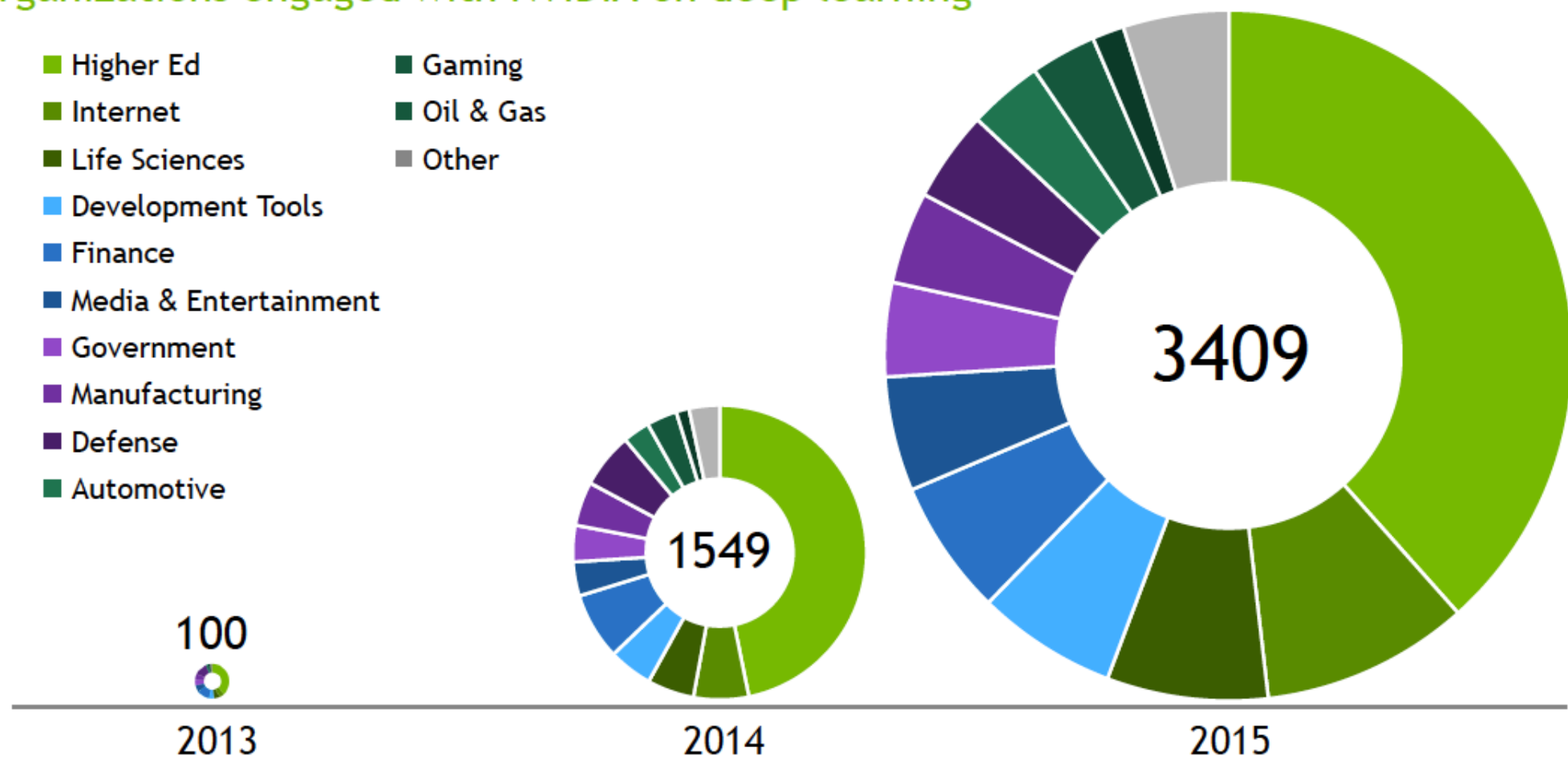
IBM: “Рынок в части когнитивных технологий размером в \$2трлн”

EVERY INDUSTRY WANTS INTELLIGENCE









Organizations engaged with NVIDIA on deep learning





- Higher Ed
- Internet
- Life Sciences
- Development Tools
- Finance
- Media & Entertainment
- Government
- Manufacturing
- Defense
- Automotive





- Gaming
- Oil & Gas
- Other



УСКОРЕНИЕ ВСЕХ ОСНОВНЫХ ФРЕЙМВОРКОВ

EDUCATION	
TORCH 	CAFFE 
THEANO 	MATCONVNET 
MOCHA.JL 	PURINE 
MINERVA 	MXNET* 

BIG SUR	TENSORFLOW	WATSON	CNTK
			

START-UPS			
CHAINER 	DL4J 	KERAS 	OPENDEEP 

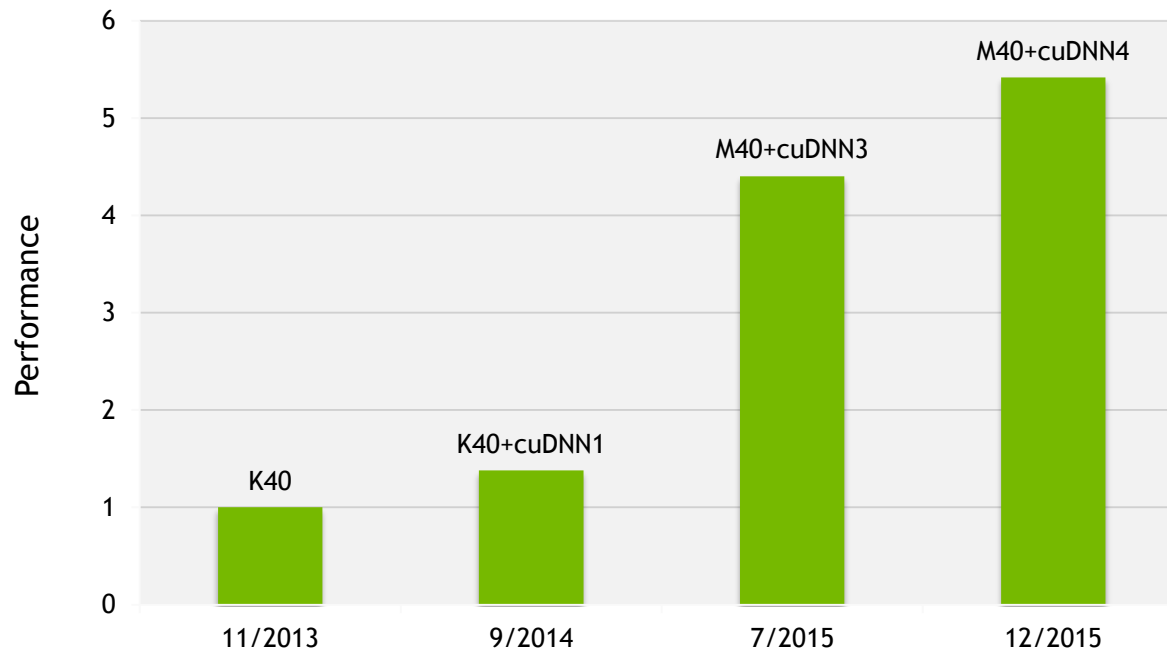
NVIDIA GPU PLATFORM

*U. Washington, CMU, Stanford, TuSimple, NYU, Microsoft, U. Alberta, NYU Shanghai

CUDA УСКОРЯЕТ
ГЛУБОКОЕ
ОБУЧЕНИЕ

В 5 РАЗ ЗА 2 ГОДА

Caffe Performance

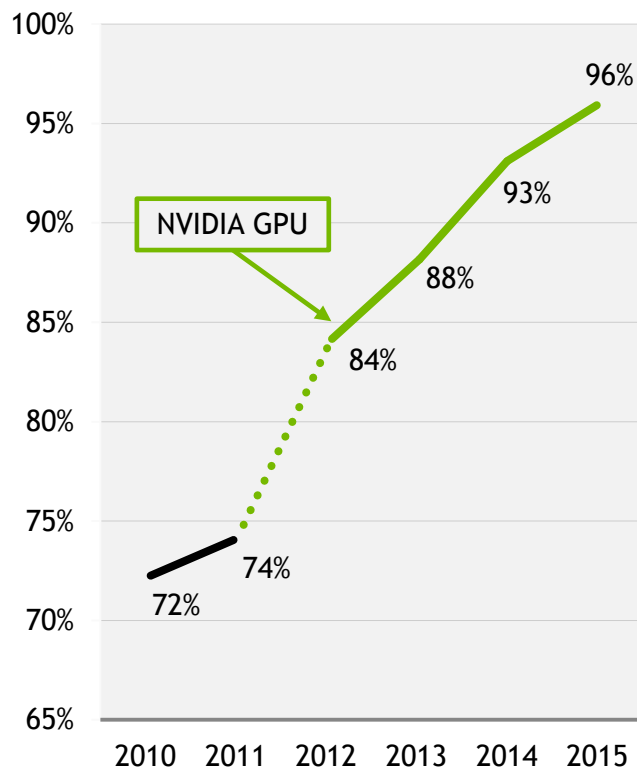


*AlexNet training throughput based on 20 iterations,
CPU: 1x E5-2680v3 12 Core 2.5GHz. 128GB System Memory, Ubuntu 14.04*

ПОТРЯСАЮЩАЯ ДИНАМИКА РАЗВИТИЯ

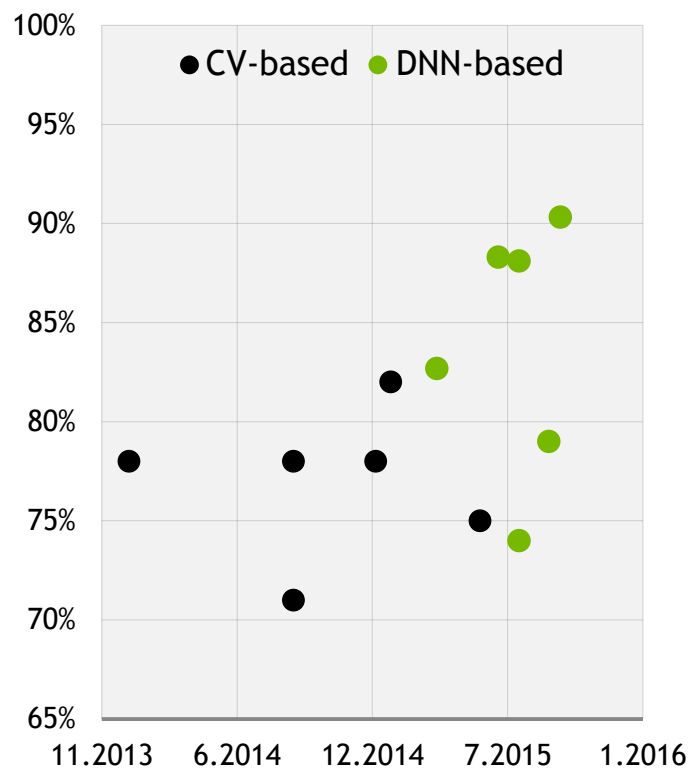
Распознавание изображений

IMAGENET



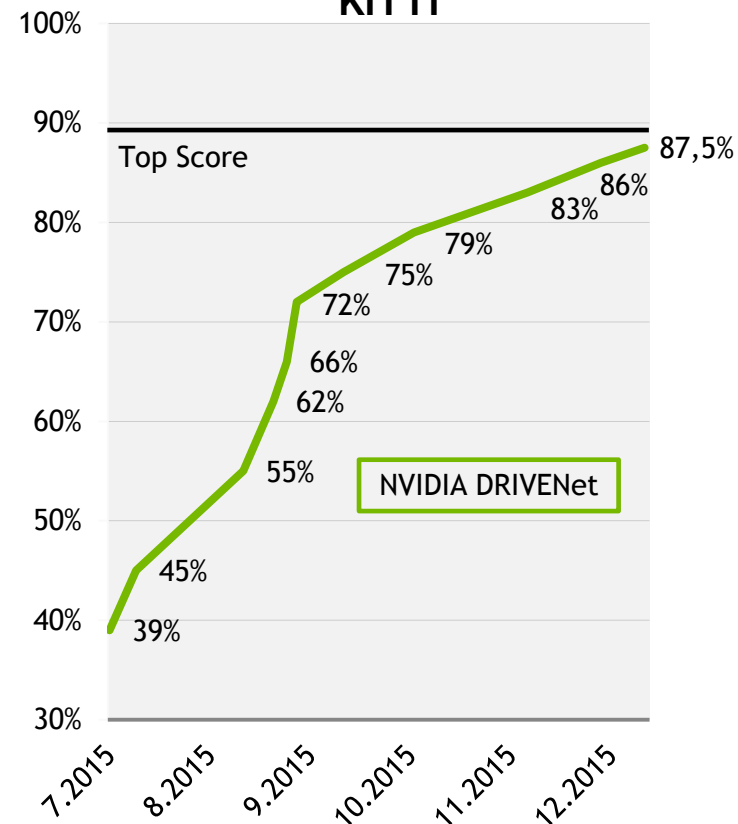
Обнаружение пешеходов

CALTECH



Распознавание объектов дорожной инфраструктуры

KITTI



СОВРЕМЕННОЕ AI-РЕШЕНИЕ ОТ FACEBOOK

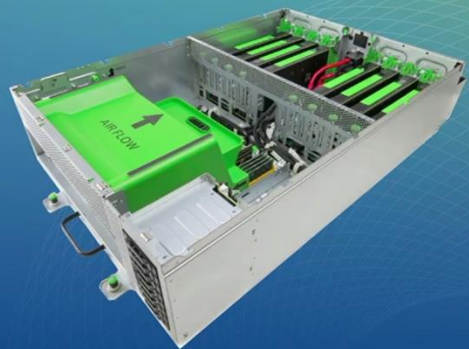
Специально созданный сервер на базе платформы NVIDIA Tesla

Решение

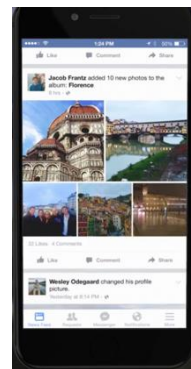
6 Million+
predictions per second

more than 25%
of engineers using AI/ML

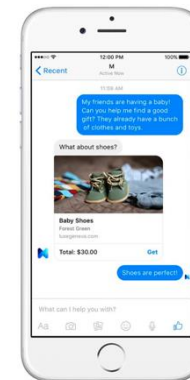
40+ PFLOP/s
available via GPU server cluster



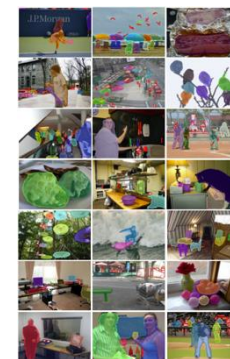
Приложения



Персонализициро
ванная лента
новостей



Персональный
цифровой
ассистент



Распознавание
фотографий

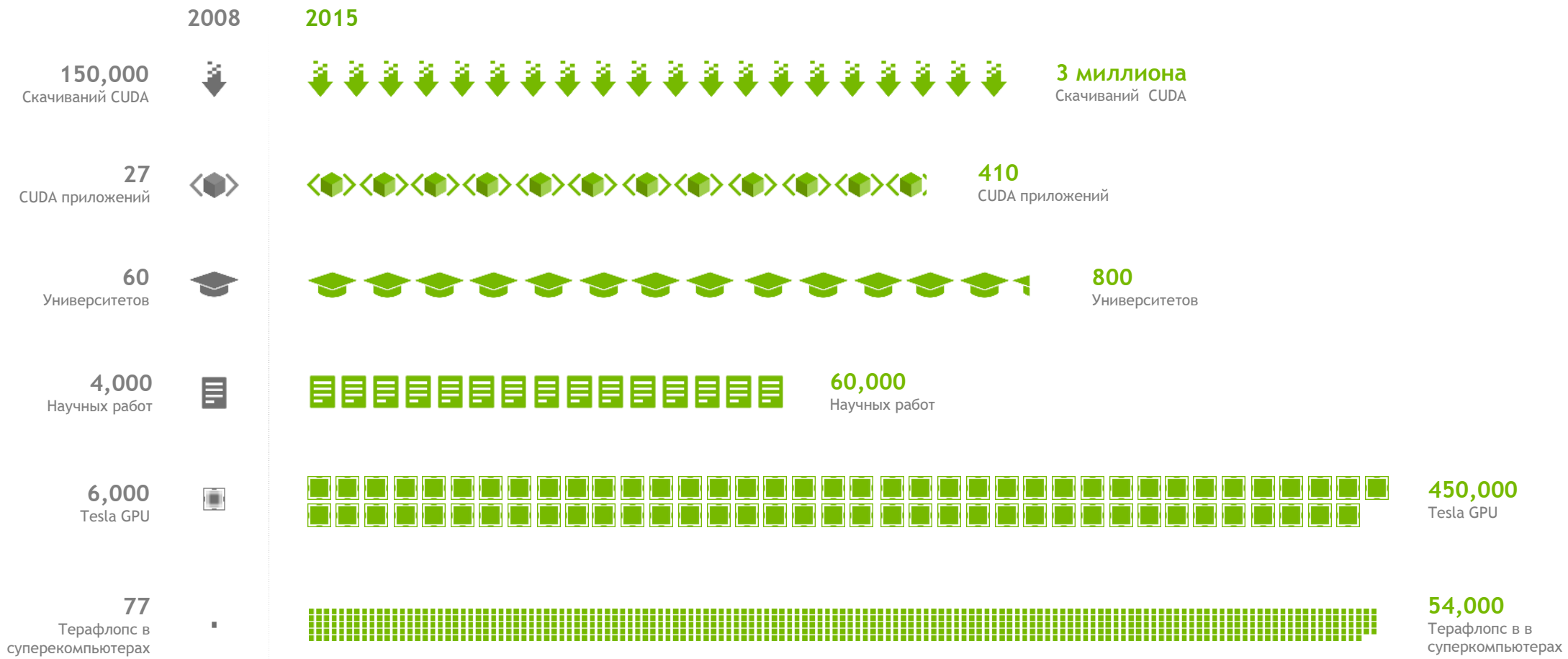
“Most of the major advances in machine learning and AI in the past few years have been contingent on **tapping into powerful GPUs** and huge data sets to build and train advanced models”



Serkan Piantino
Engineering Director of Facebook AI Research

ПЛАТФОРМА TESLA ДЛЯ РАЗРАБОТЧИКОВ

10-КРАТНЫЙ РОСТ УСКОРЕННЫХ ВЫЧИСЛЕНИЙ



NVIDIA SDK

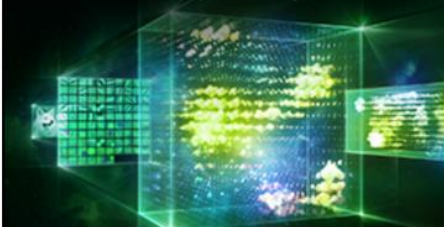
The Essential Resource for GPU Developers

NVIDIA SDK

DEEP LEARNING

Deep Learning SDK

High-performance tools and libraries for deep learning



SELF-DRIVING CARS

NVIDIA DriveWorks™

Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous



VIRTUAL REALITY

NVIDIA VRWorks™

A comprehensive SDK for VR headsets, games and professional applications



GAME DEVELOPMENT

NVIDIA GameWorks™

Advanced simulation and rendering technology for game development



ACCELERATED COMPUTING

NVIDIA ComputeWorks™

Everything scientists and engineers need to build GPU-accelerated applications



DESIGN & VISUALIZATION

NVIDIA DesignWorks™

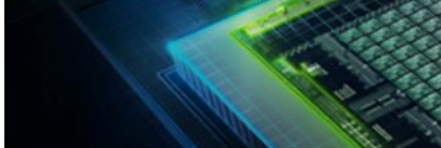
Tools and technologies to create professional graphics and advanced rendering applications



AUTONOMOUS MACHINES

NVIDIA JetPack™

Powering breakthroughs in autonomous machines, robotics and embedded computing



ADDITIONAL RESOURCES

More resources for GPU Developers



NVIDIA COMPUTEWORKS

CUDA 8 – июнь | cuDNN 5 | nvGRAPH – июнь
Плагин IndeX для ParaView – май

COMPUTEWORKS

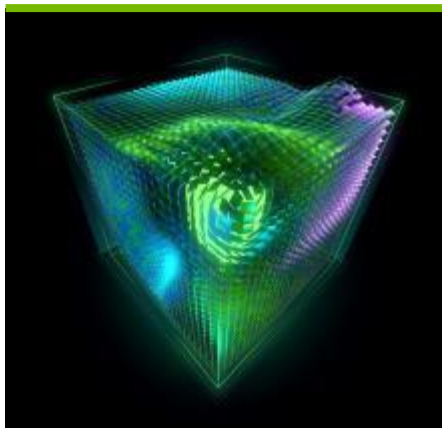
GAMEWORKS

VRWORKS

DESIGNWORKS

DRIVEWORKS

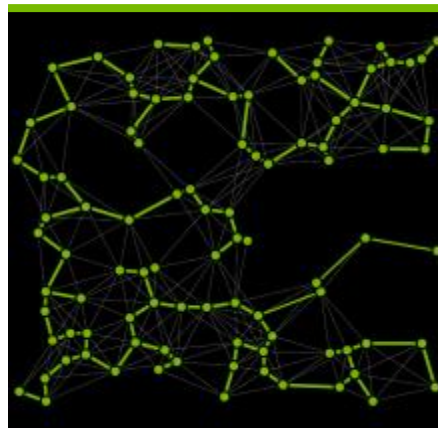
JETPACK



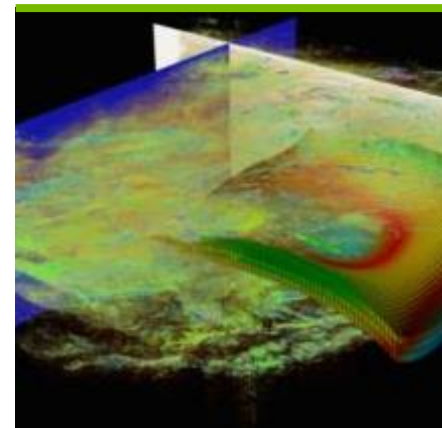
CUDA



cuDNN



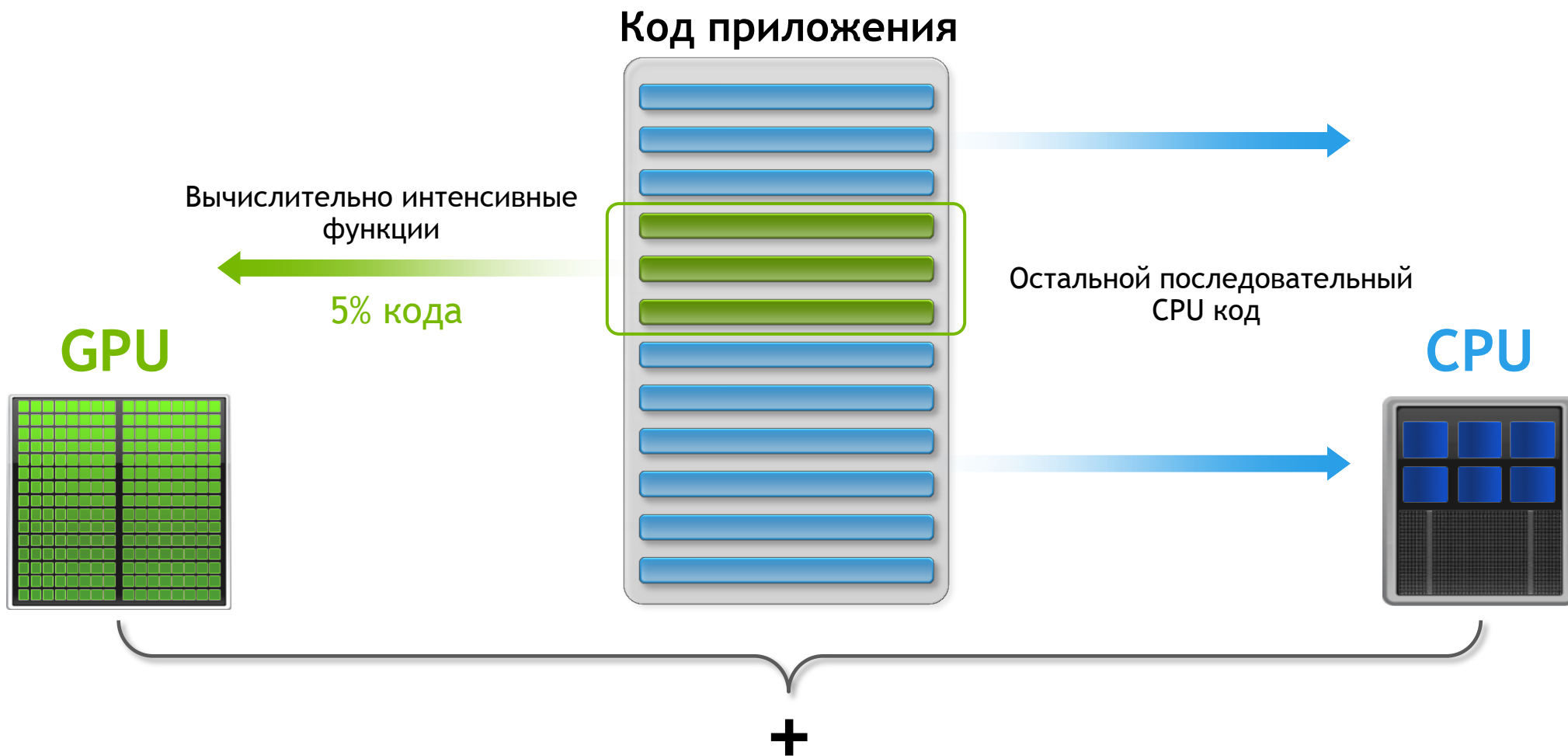
nvGRAPH



IndeX

а так же другие технологии, такие как:
AMGx, cuSOLVER, cuSPARSE, OpenACC, NSIGHT, THRUST

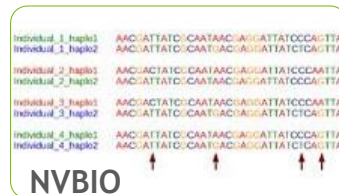
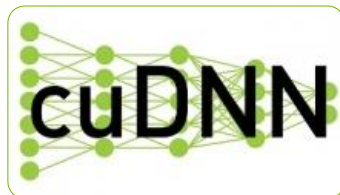
КАК РАБОТАЕТ УСКОРЕНИЕ НА GPU



БИБЛИОТЕКИ С GPU УСКОРЕНИЕМ

“copy-paste” ускорение для ваших приложений

Ориентированные
на предметные
области



Обработка
изображений



Линейная
алгебра



Мат. алгоритмы



OpenACC

Простота

Производительность

Портируемость

```
main()
{
  <serial code>
  #pragma acc kernels
  //automatically runs on GPU
  {
    <parallel code>
  }
}
```

Университет Иллинойса
Реконструкция по МРТ снимкам



70x ускорение
2 дня затрат

RIKEN Япония
NICAM- моделирование климата



7-8x ускорение
5% кода модифицировано

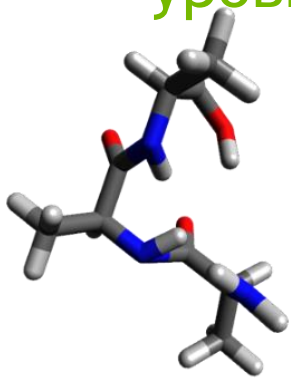
8000+

разработчиков

используют OpenACC

LS-DALTON

Масштабное приложение для
высокоточного расчета
молекулярных энергетических
уровней



“OpenACC делает вычисления на GPU доступными для специалистов в своих предметных областях. Первоначальное использование OpenACC потребовало минимум усилий и, что главное, без необходимости модифицировать существующую CPU реализацию.”

Janus Juul Eriksen, PhD Fellow
qLEAP Center for Theoretical Chemistry, Aarhus University

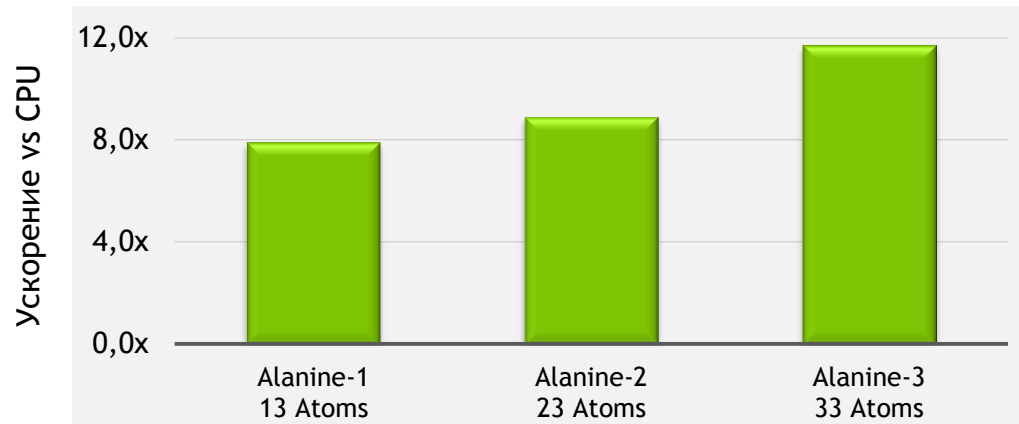


Минимум усилий

Модификация кода	Затраченное время	Версионность кода
<100 строк	1 неделя	1 версия

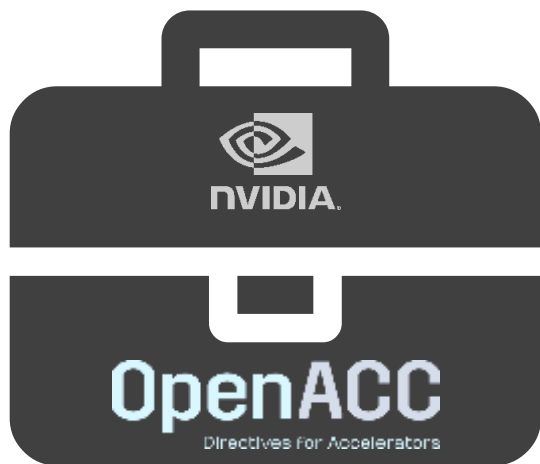
Значительное ускорение

LS-DALTON CCSD(T) модуль
Измерено на CK Titan (AMD CPU vs Tesla K20X)



NVIDIA OPENACC TOOLKIT

Бесплатный Toolkit - простой и эффективный путь для ускорения приложений



PGI компилятор

Бесплатный OpenACC компилятор для академической среды



NVProf профилировщик

Простота поиска участков кода для добавления директив



Примеры кода

Примеры реализации распространенных алгоритмов



Документация

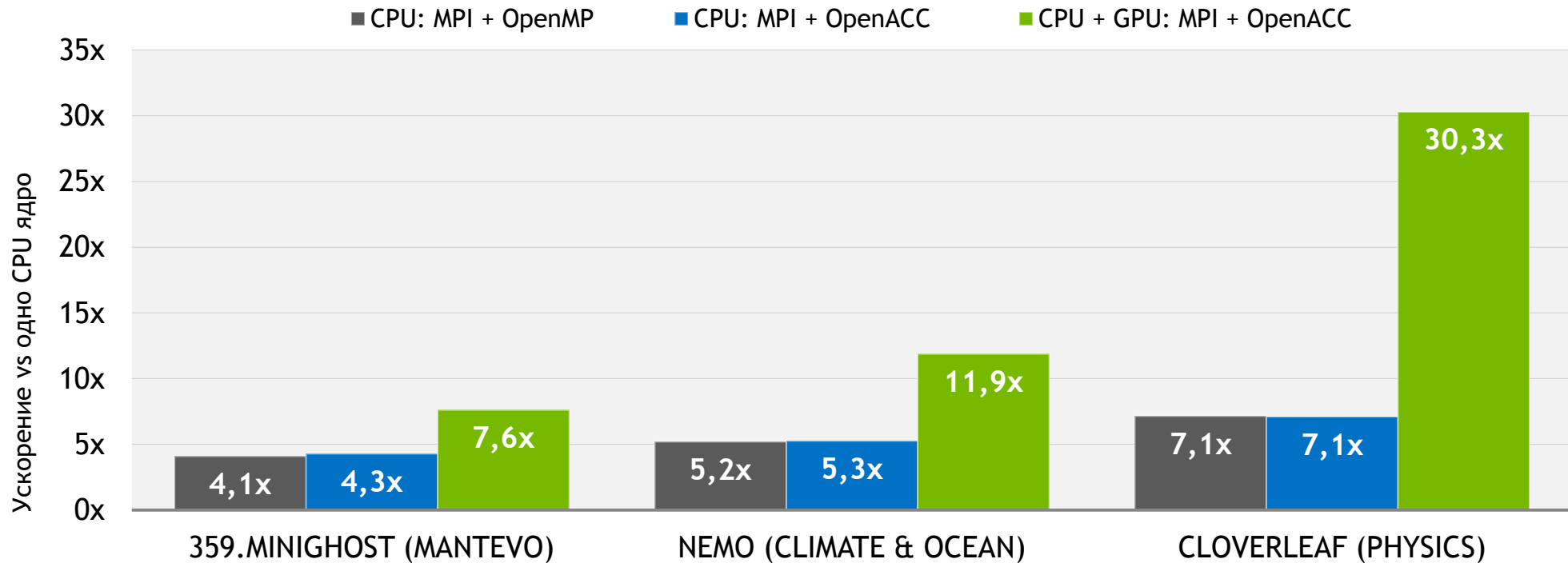
Руководства по использованию, методики, форумы

Доступно на <http://www.nvidia.com/openacc>

OPENASS ОБЕСПЕЧИВАЕТ ПОРТИРУЕМОСТЬ ПРОИЗВОДИТЕЛЬНОСТИ

Прокладывая дорогу в будущее: один код для всех HPC процессоров

Производительность приложений



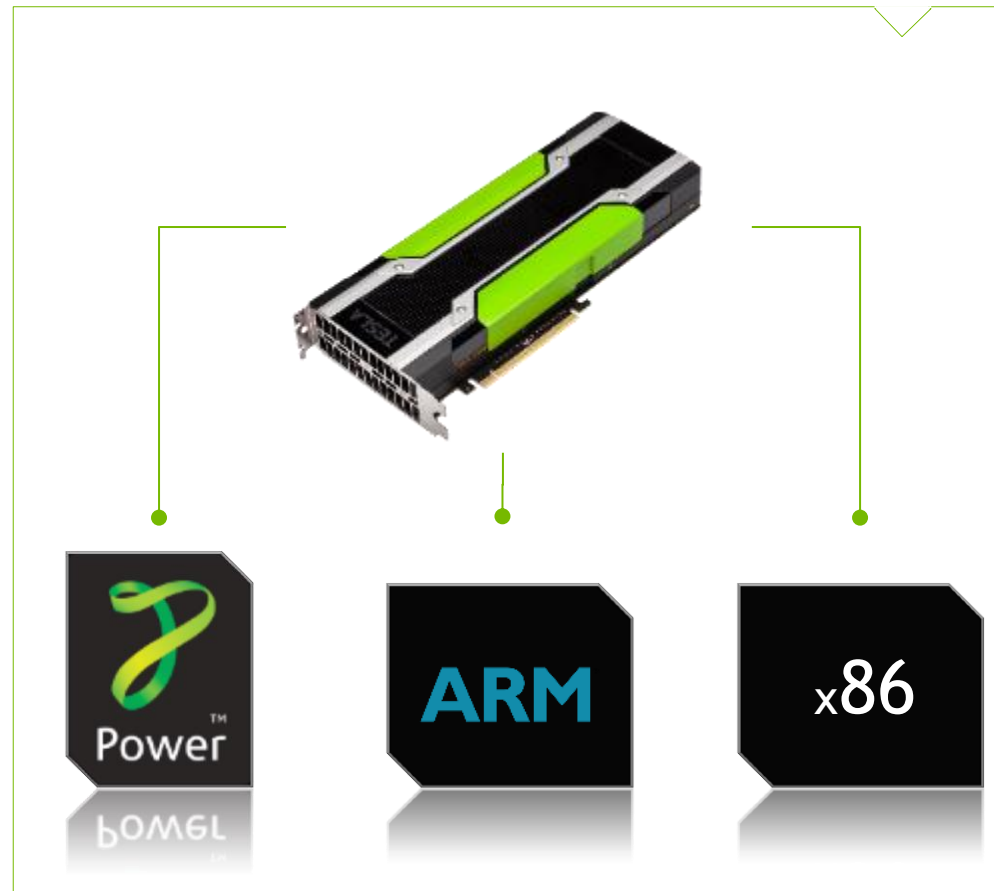
359.miniGhost: CPU: Intel Xeon E5-2698 v3, 2 sockets, 32-cores total, GPU: Tesla K80- single GPU
NEMO: Each socket CPU: Intel Xeon E5--2698 v3, 16 cores; GPU: NVIDIA K80 both GPUs
CLOVERLEAF: CPU: Dual socket Intel Xeon CPU E5-2690 v2, 20 cores total, GPU: Tesla K80 both GPUs

ОБЩАЯ ПРОГРАММНАЯ МОДЕЛЬ ДЛЯ РАЗНЫХ ПЛАТФОРМ

Библиотеки AmgX cuDNN cuBLAS
OpenCV Thrust

Директивы OpenACC

Языки программирования C/C++ Fortran python Java



ЭКОСИСТЕМА ДЛЯ РАЗРАБОТЧИКОВ НА GPU

Прикладные пакеты

MATLAB
Mathematica
NI LabView
pyCUDA

Отладчики и профилировщики

cuda-gdb
NV Visual Profiler
Parallel Nsight
Visual Studio
Allinea
TotalView

GPU компиляторы

C
C++
Fortran
Java
Python

Директивы и кластерные инструменты

OpenACC
mCUDA
OpenMP
Ocelot

Библиотеки

BLAS
FFT
LAPACK
NPP
Video
Imaging
GPULib

Консалтинг и обучение



Аппаратные OEM решения



РАЗРАБОТКА НА GEFORCE, ВНЕДРЕНИЕ НА TESLA



Создано для разработчиков и геймеров

Доступно для всех

<https://developer.nvidia.com/cuda-gpus>



Создано для дата-центров

ECC

Работа 24x7

Мониторинг GPU

Управление кластером

GPUDirect-RDMA

Hyper-Q для MPI

3 года гарантии

Интегрированные OEM системы, профессиональная поддержка

