NVIDIA GPU Computing

A Revolution in High Performance Computing

GPUs and the Future of Accelerated Computing

Emerging Technology Conference 2014
University of Manchester

John Ashley

Senior Solutions Architect

jashley@nvidia.com

Disclaimers

- These are my views, not NVIDIA's (although I do hope you all will agree by the time I'm done!) , and where I've quoted others, any misrepresentation of their views is mine and mine alone.

- There are "forward looking statements" in here. You are warned that my crystal ball is no better than anyone else's, even if I'm speaking about things like NVIDIA roadmaps, etc. It's a lawyer thing.

- Trademarks of other firms are their own, etc.

Agenda

- Who is NVIDIA?
- Why should I care about accelerated computing?
- What sort of difference can CUDA make?
- Where is NVIDIA going?

Who is NVIDIA?

Audi connect
piloted driving – the new dimension of driving pleasure
“Super” Computing
From Super Computers to Super Phones

NVIDIA products span the power-performance spectrum

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)

GPUs are becoming…

100M
CUDA -Capable GPUs

150K
CUDA Downloads

77
Supercomputing Teraflops

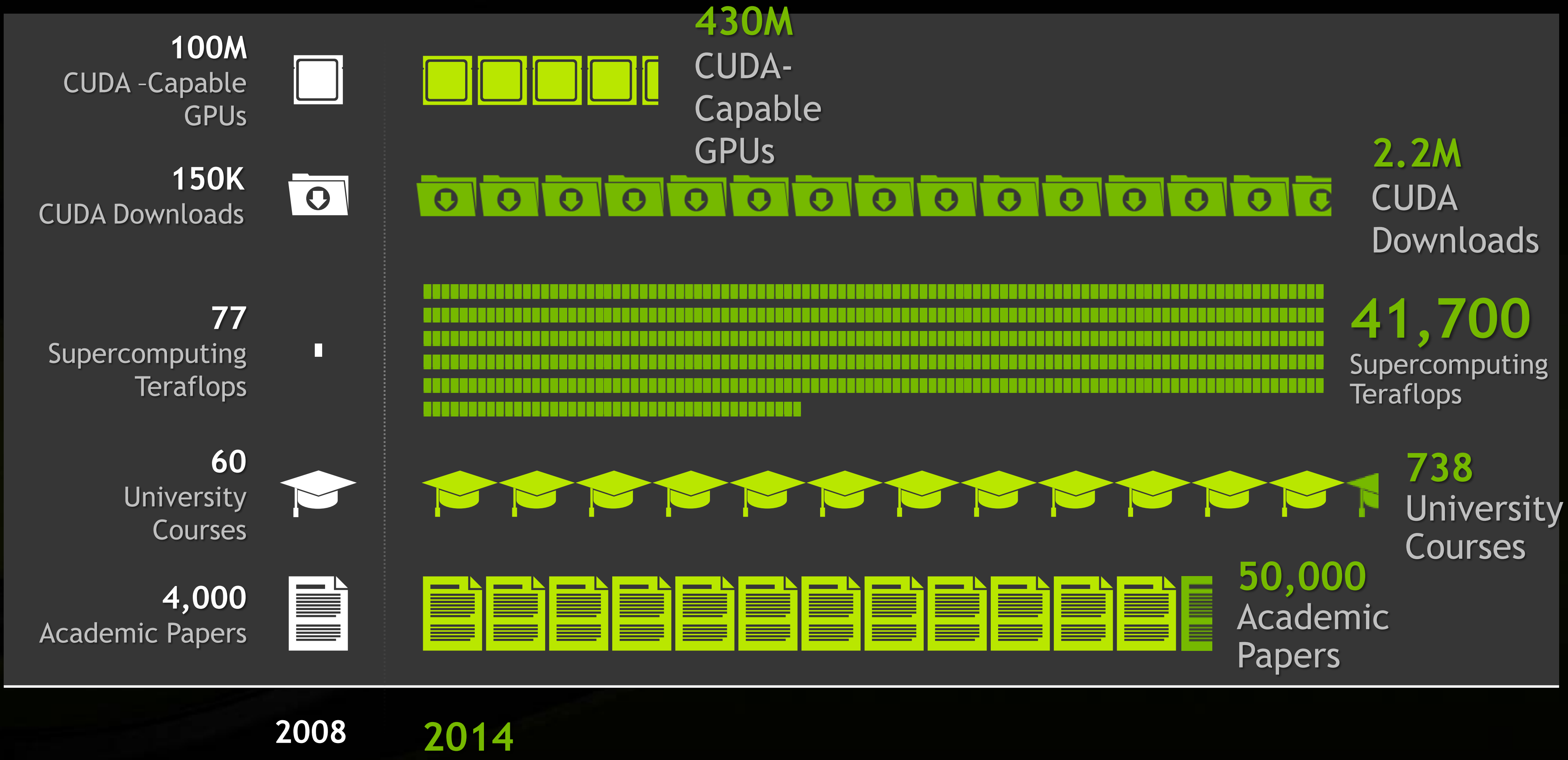
60
University Courses

4,000
Academic Papers

2008

2014

GPUs are becoming…pervasive



100M
CUDA -Capable GPUs

150K
CUDA Downloads

77
Supercomputing Teraflops

60
University Courses

4,000
Academic Papers

430M
CUDA-Capable GPUs

2.2M
CUDA Downloads

41,700
Supercomputing Teraflops

738
University Courses

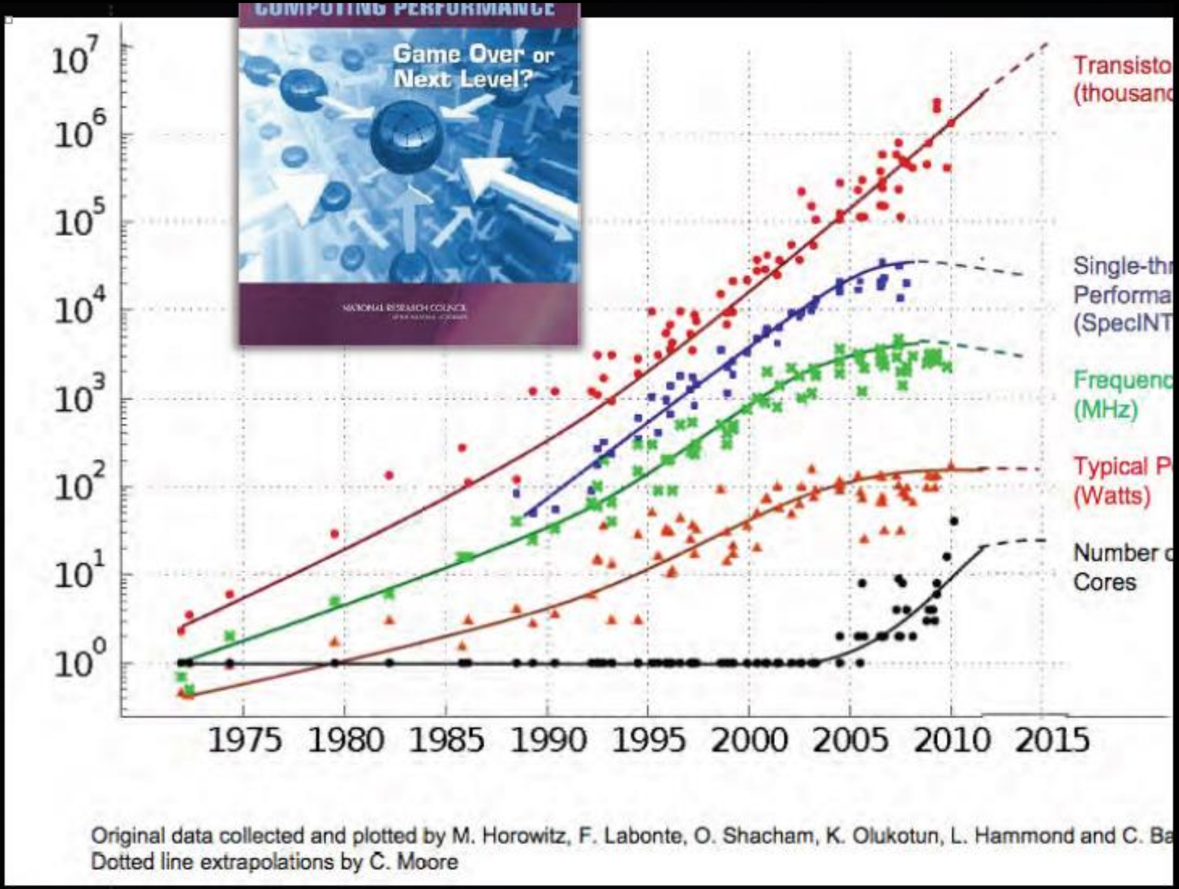
50,000
Academic Papers

2008

2014

Why should I care about accelerated computing?

Moore's Law isn't what it used to be.



Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Ba
Dotted line extrapolations by C. Moore

Moore's law is alive and well, but…

- Instruction-level parallelism (ILP) was mined out in 2001
- Voltage scaling (Dennard scaling) ended in 2005
- Most power is spent on communication

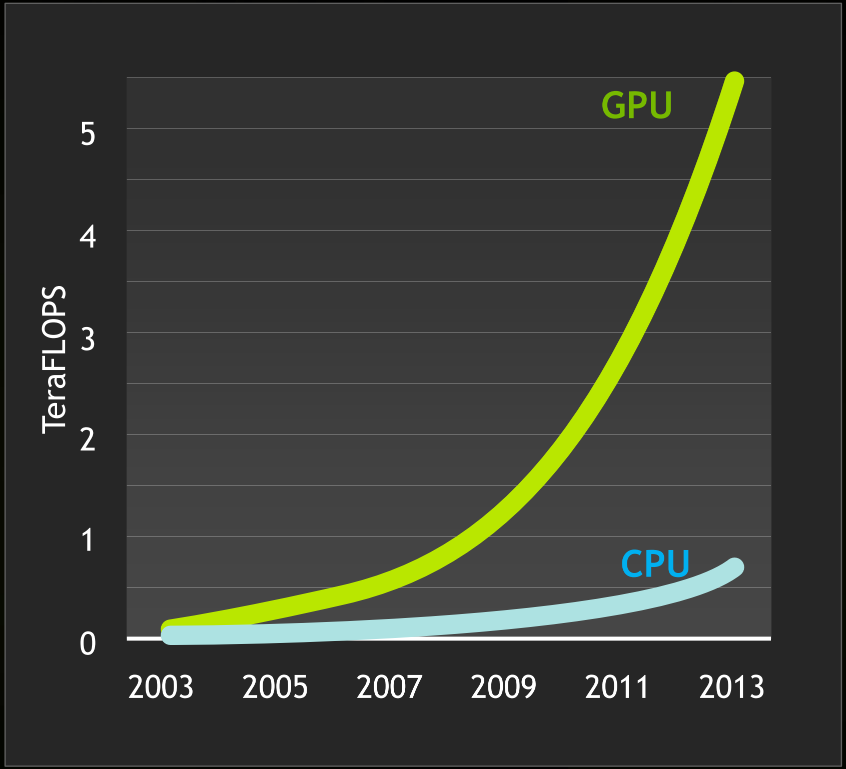
What does this mean to you?

The future is here today …

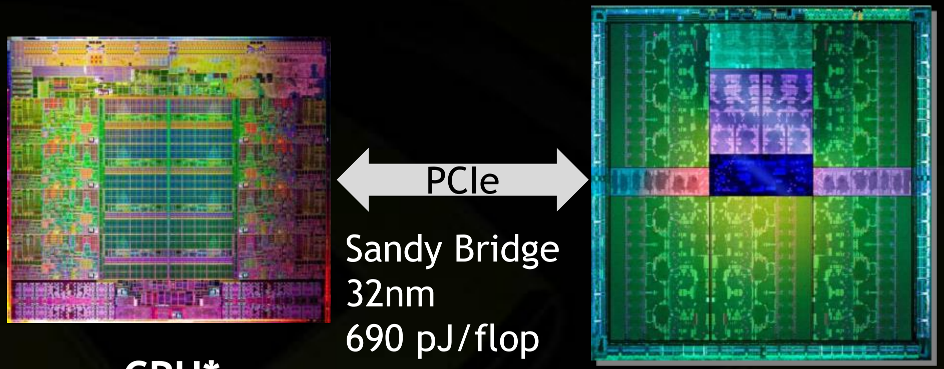
- Performance gains come from parallelism

- Systems are power limited
(efficiency IS performance)

- Systems are communication limited
(locality IS performance)



High Performance Computing is Going Hybrid



PCIe

Sandy Bridge
32nm
690 pJ/flop

Kepler
28nm
134 pJ/flop

CPU*
Fast single threads
(serial work)

GPU
Extreme power-efficiency
(throughput work)

- Do most work by many cores optimized for **extreme energy efficiency**
- Still need a few cores optimized for **fast serial work**
- **Amdahl’s law continues to apply to all architectures**

*x86, ARM, Power

What sort of difference can CUDA make?

CUDA Parallel Computing Platform

www.nvidia.com/getcuda

Programming Approaches

Libraries	Directives	Programming Languages
"Drop-in" Acceleration	Easily Accelerate Apps	Maximum Flexibility

Development Environment

Nsight IDE
Linux, Mac and Windows
GPU Debugging and Profiling

CUDA-GDB debugger
NVIDIA Visual Profiler

Open Compiler Tool Chain

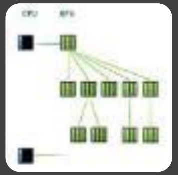
LLVM COMPILER INFRASTRUCTURE

Enables compiling new languages to CUDA platform, and CUDA languages to other architectures

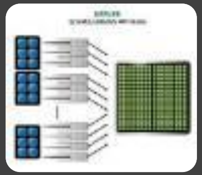
Hardware Capabilities

SMX

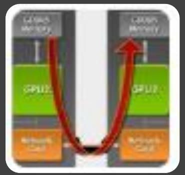
Dynamic Parallelism



HyperQ

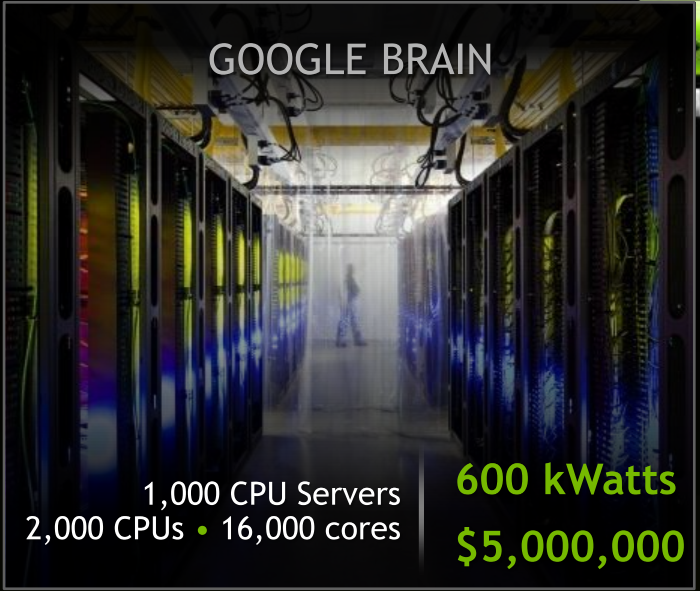


GPUDirect



Artificial Neural Network at a Fraction of the Cost with GPUs

> “*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

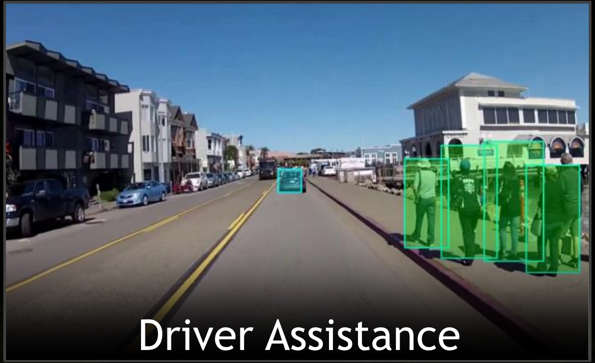
3 GPU-Accelerated Servers
12 GPUs • 18,432 cores

4 kWatts

$33,000

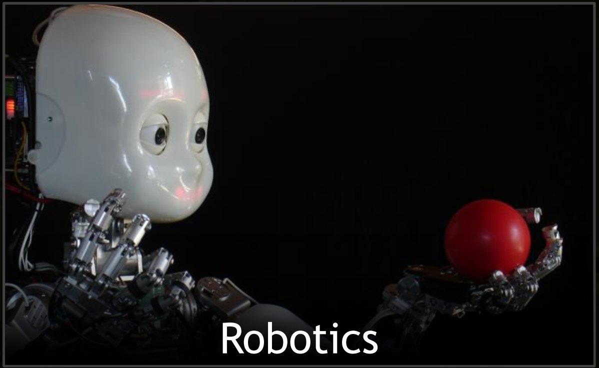
VISIONWORKS

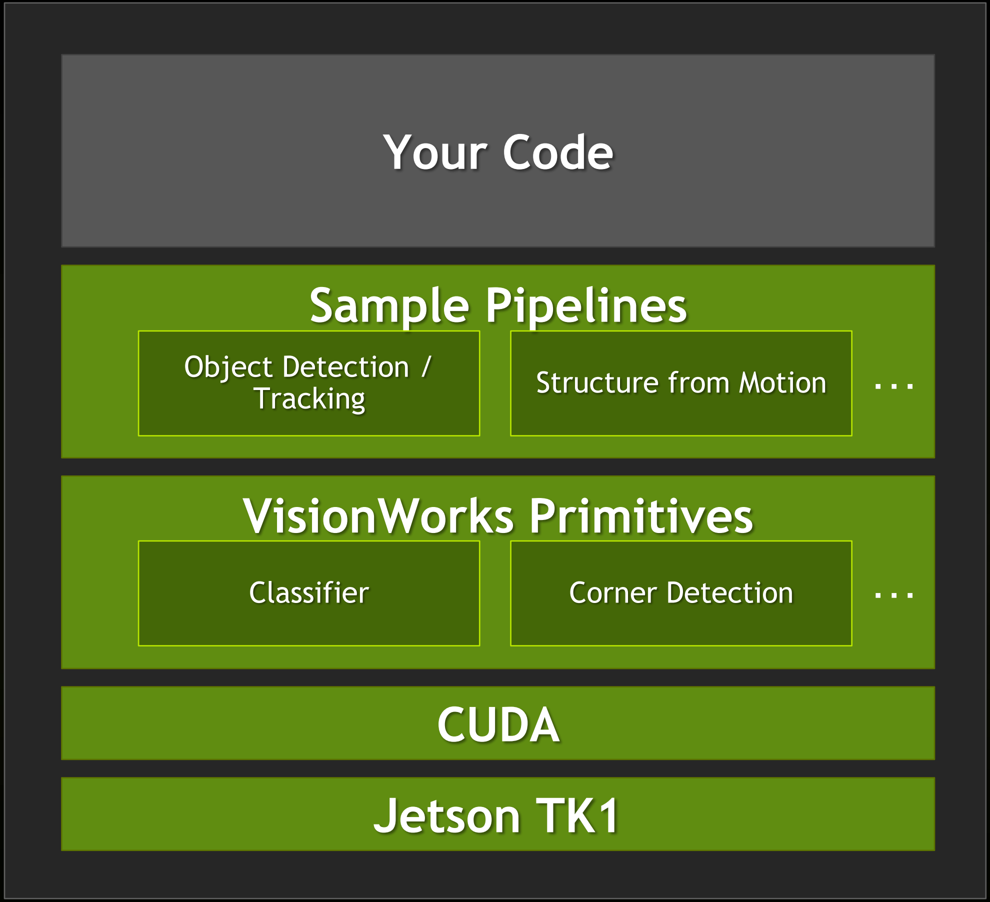
COMPUTER VISION ON CUDA


Driver Assistance

Computational Photography

Augmented Reality


Robotics


Your Code

Sample Pipelines

Object Detection / Tracking

Structure from Motion

. . .

VisionWorks Primitives

Classifier

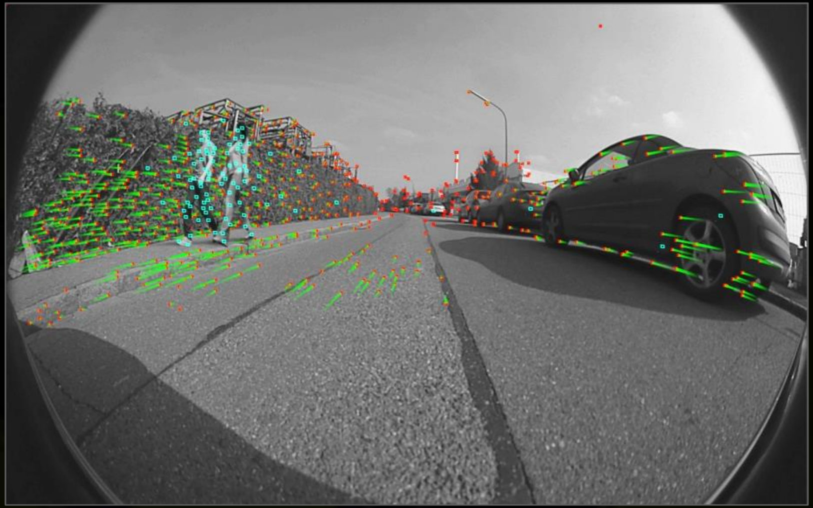
Corner Detection

. . .

CUDA

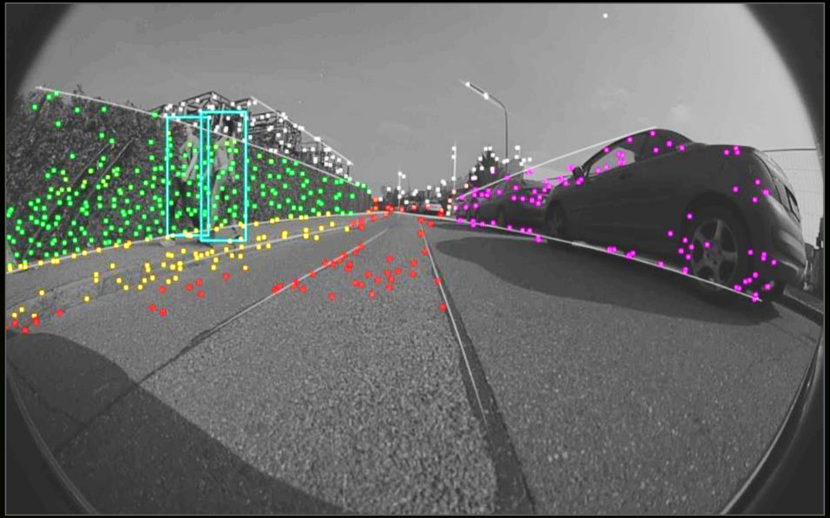
Jetson TK1

Computer Vision on CUDA



Feature Detection / Tracking

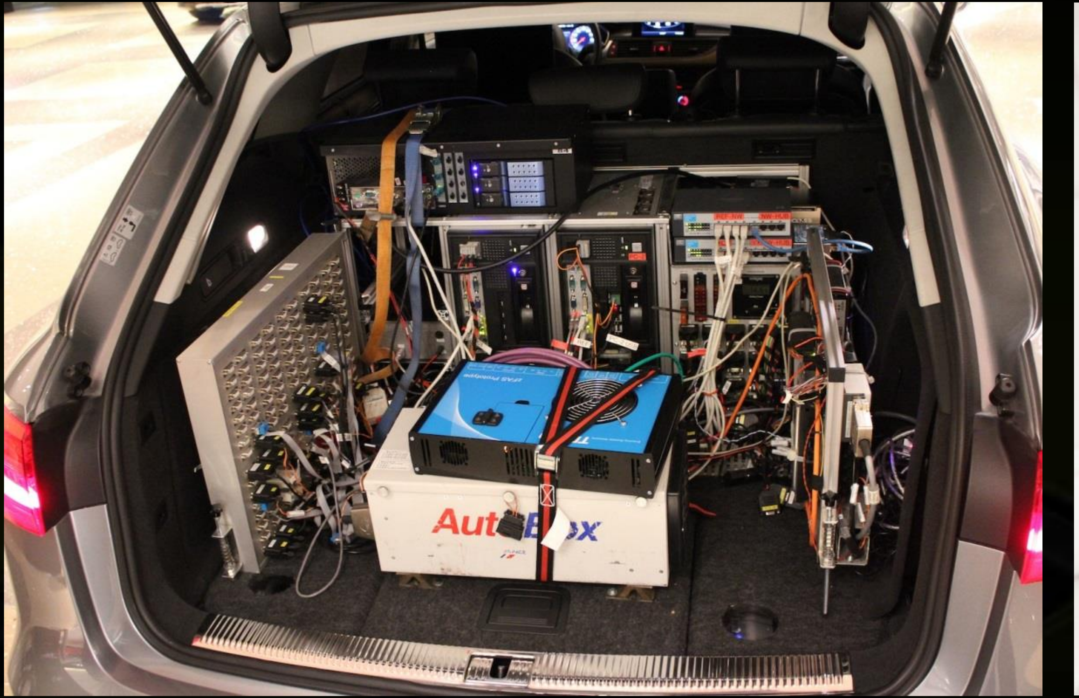
~30 GFLOPS @ 30 Hz

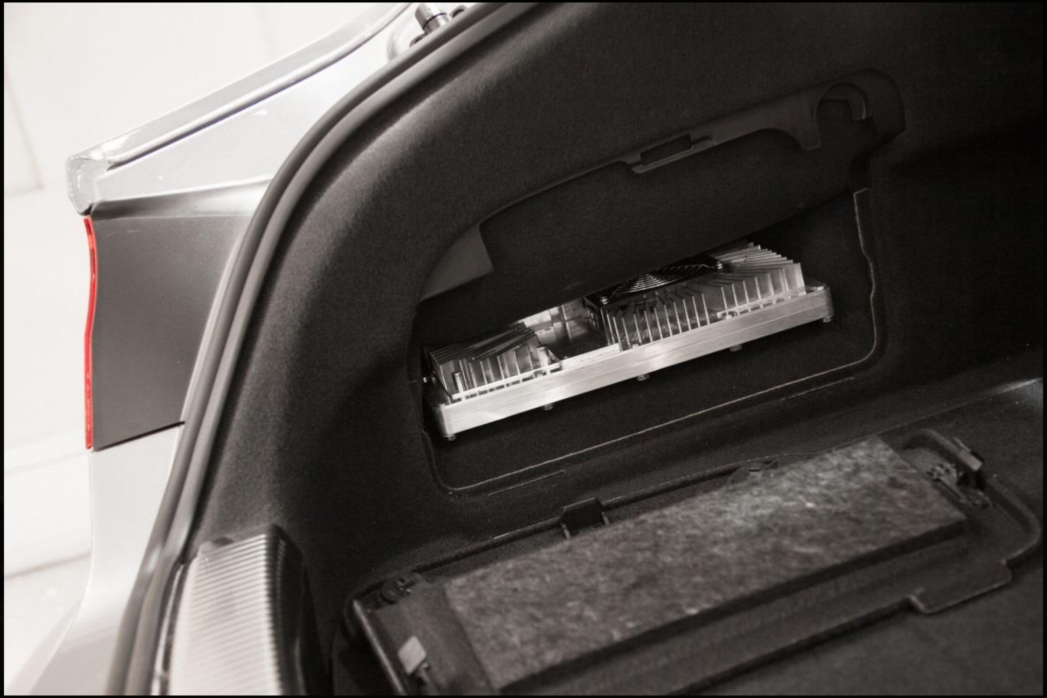


3D Scene Interpretation

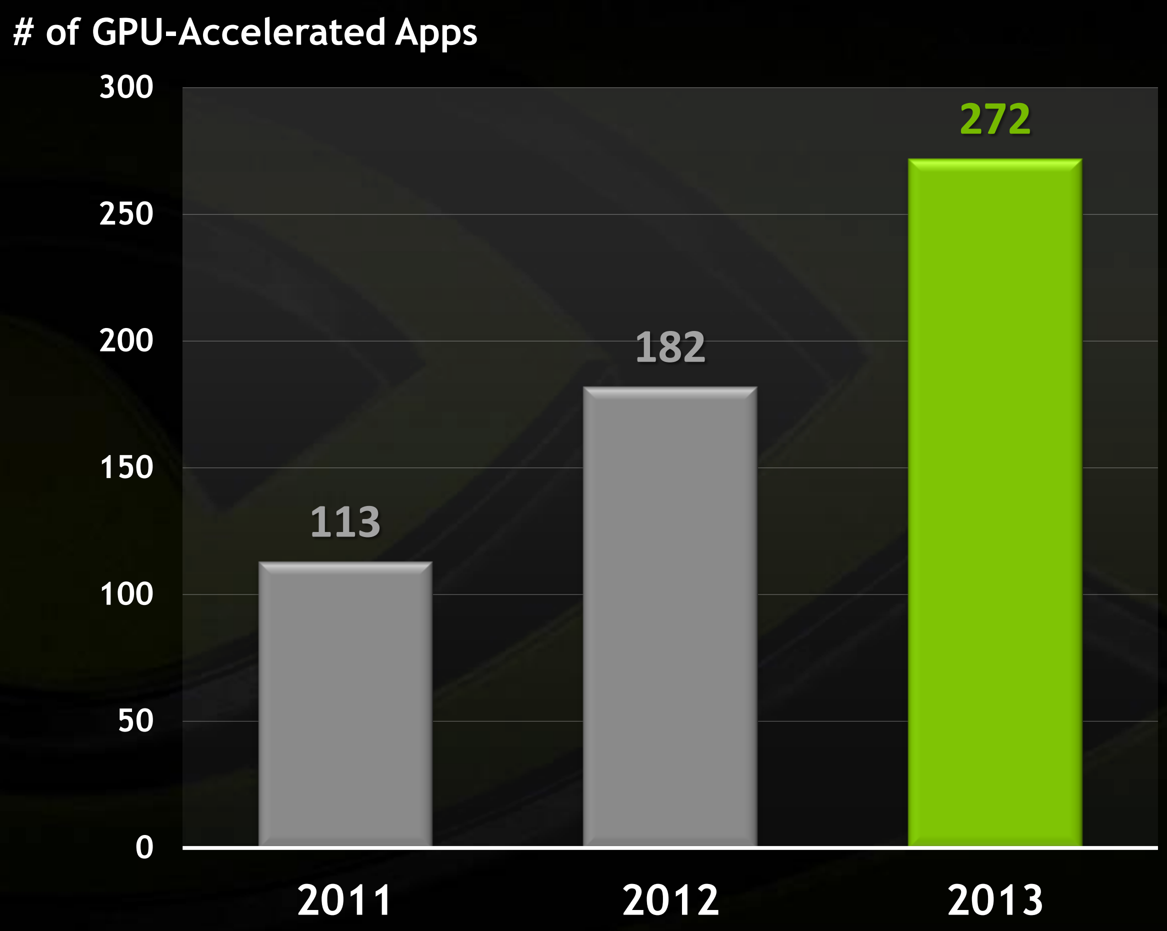
~280 GFLOPS @ 30 Hz

Audi Self Driving Car Before & After CUDA





Solid Growth of GPU Accelerated Apps



Top HPC Applications

Molecular Dynamics	AMBER CHARMM DESMOND	GROMACS LAMMPS NAMD
Quantum Chemistry	Abinit Gaussian	GAMESS NWChem
Material Science	CP2K QMCPACK	Quantum Espresso VASP
Weather & Climate	COSMO GEOS-5 HOMME	CAM-SE NEMO NIM WRF
Lattice QCD	Chroma	MILC
Plasma Physics	GTC	GTS
Structural Mechanics	ANSYS Mechanical LS-DYNA Implicit MSC Nastran	OptiStruct Abaqus/Standard
Fluid Dynamics	ANSYS Fluent	Culises (OpenFOAM)

Accelerated, In Development

NVIDIA

POPULAR GPU-ACCELERATED APPLICATIONS

Quantum Chemistry

Materials Science

PHYSICS

Defense and Intelligence

Computational Finance

Manufacturing: CAD and CAE

COMPUTER AIDED DESIGN

COMPUTATIONAL STRUCTURAL MECHANICS

Media and Entertainment

EDITING

ON-SET, REVIEW AND STEREO TOOLS

Oil and Gas

Research: Higher Education and Supercomputing

COMPUTATIONAL CHEMISTRY AND BIOLOGY

Bioinformatics

[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
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

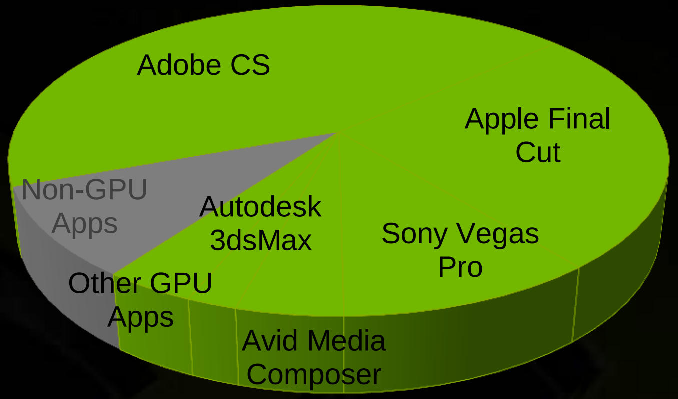
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Abalone	Models molecular dynamics of biopolymers for simulations of proteins, DNA and ligands	Simulations (on 1060 GPU)	4-29x	T 2075, 2090, K10, K20, K20X	Single Only	Available now Version 1.8.48
ACEMD	GPU simulation 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 Q4/12
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 Q4/12
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
NAMD	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
OpenMM	Library and application for molecular dynamics for HPC with GPUs	Implicit and explicit solvent, custom forces	Implicit: 127-213 ns/day; Explicit: 18-55 ns/day DHFR	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 4.1.1

272 GPU-Accelerated Applications

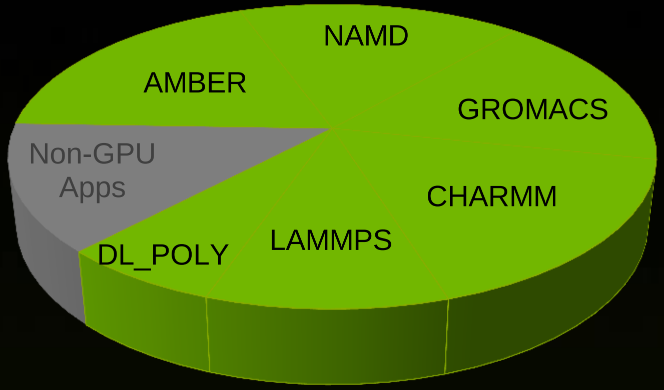
www.nvidia.com/appscatalog

Top Applications Now with Built-in GPU Support

Digital Content Creation

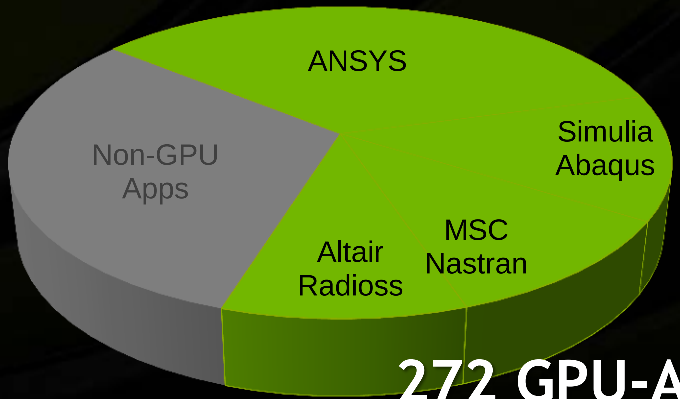


Molecular Dynamics

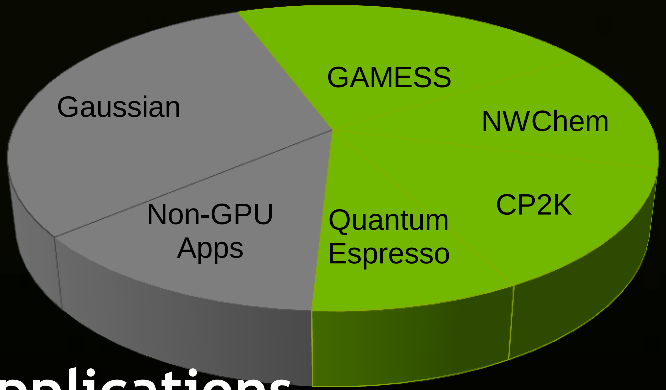


Application Market Share by Segment

Computer-Aided Engineering



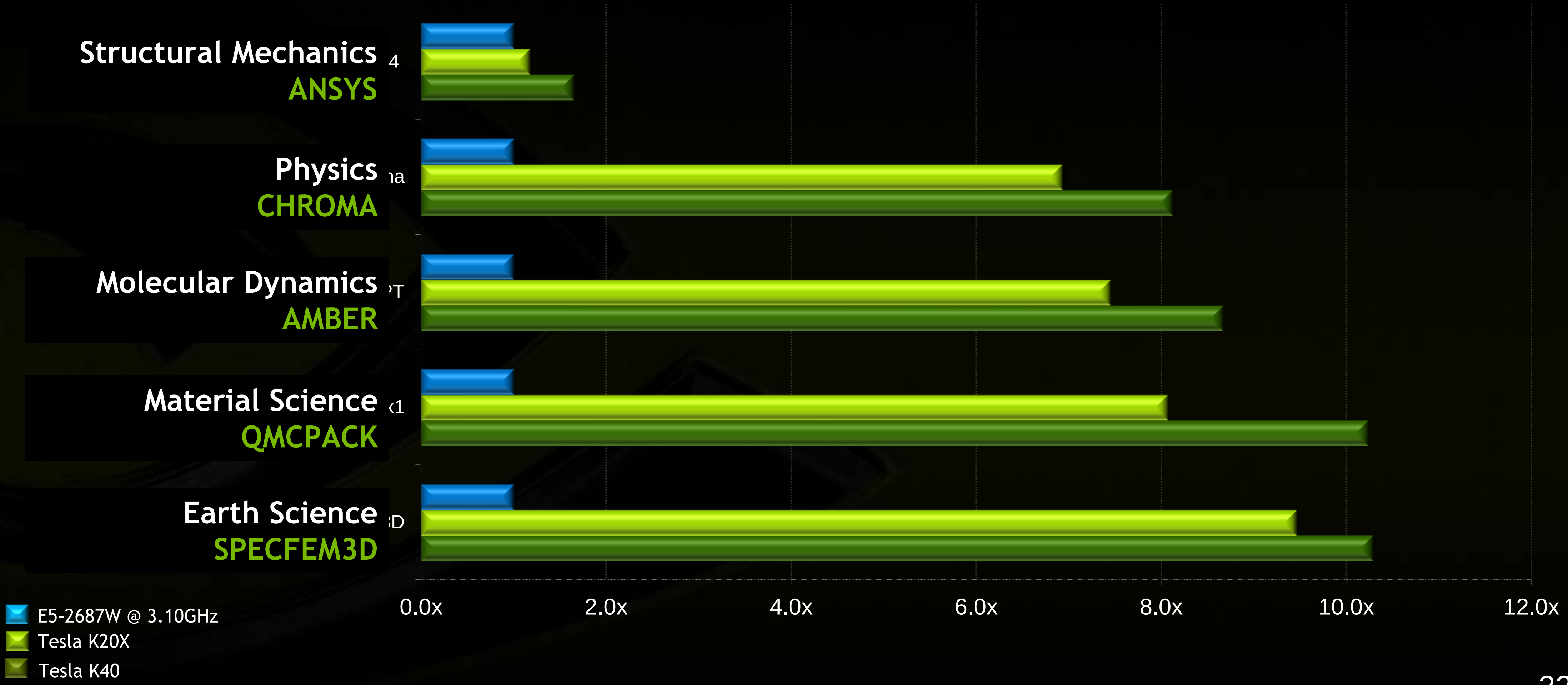
Quantum Chemistry



272 GPU-Accelerated Applications

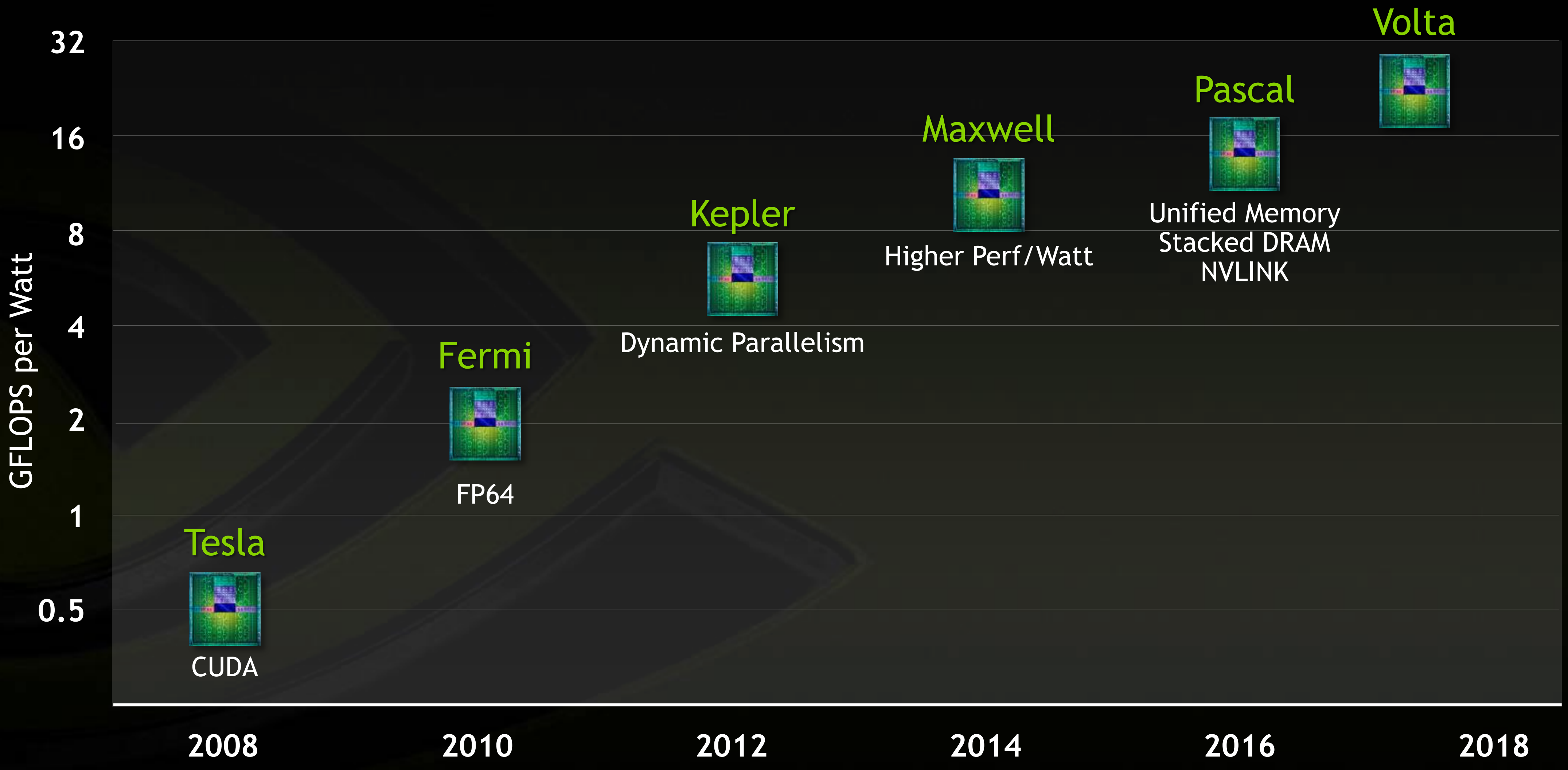
www.nvidia.com/appscatalog

Performance on Leading Scientific Applications



Where is NVIDIA going?

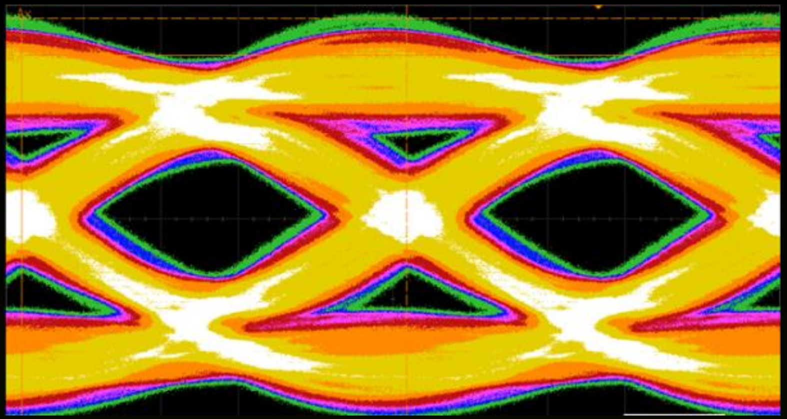
Fast Paced CUDA GPU Roadmap



Introducing NVLINK and Stacked Memory

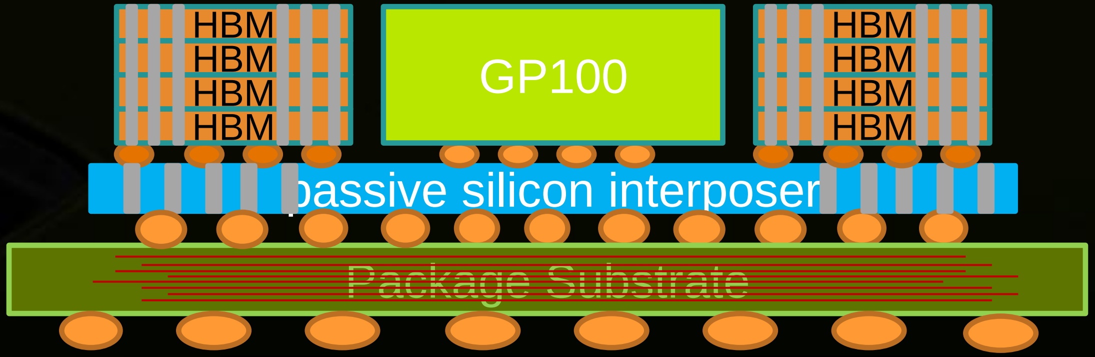
NVLINK

- GPU high speed interconnect
- 5-12x PCIe Gen 3 Bandwidth
- Drastically reduced energy/bit

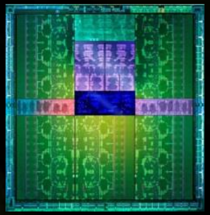


Stacked Memory

- 2-4x Capacity & Bandwidth
- 3-4x More Energy Efficient per bit
- Leaves more power for compute

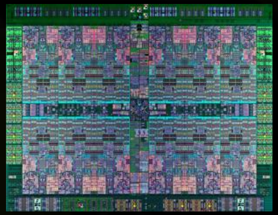


IBM Partners with NVIDIA to Build Next-Generation Supercomputers



Tesla
GPU

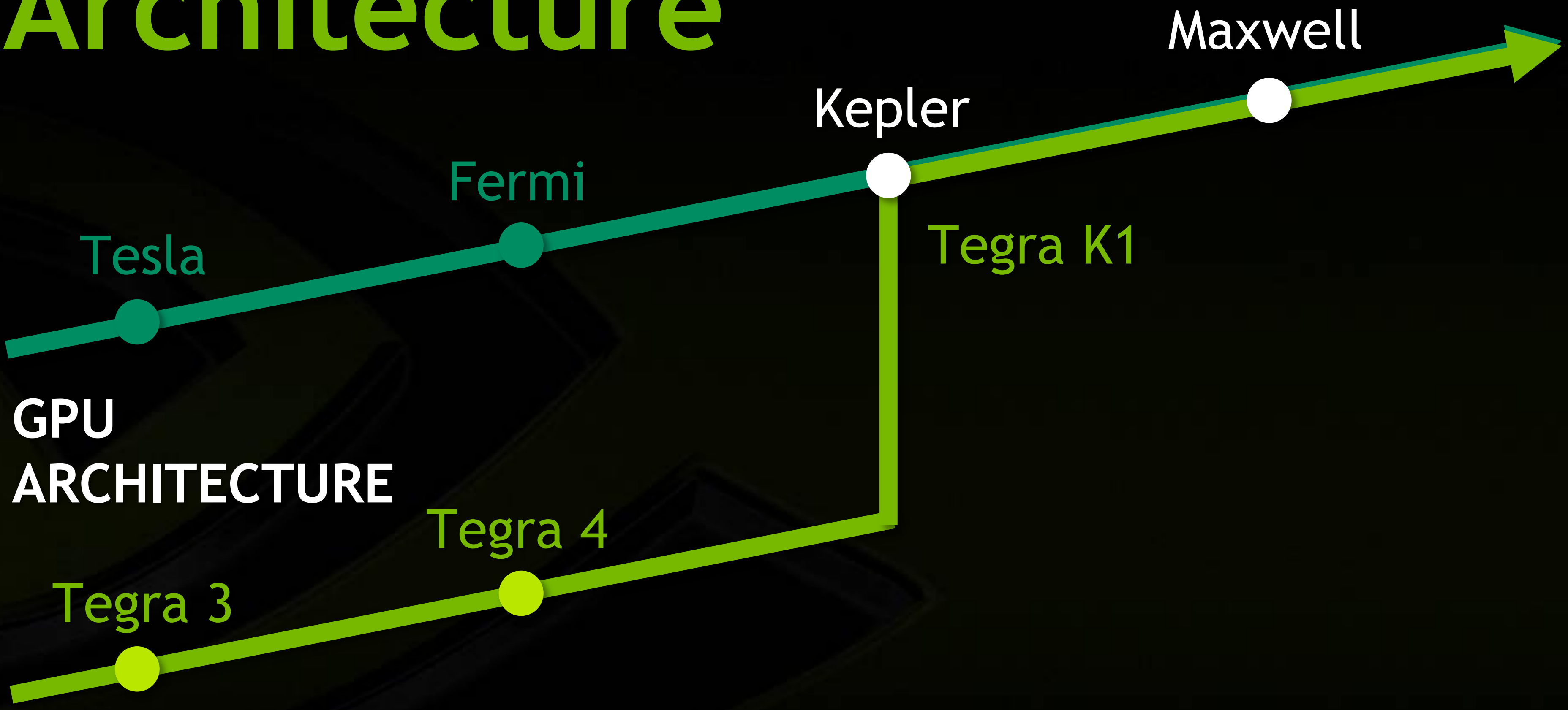
+



POWER 8
CPU

GPU-Accelerated POWER-Based Systems Available in 2014

Unify GPU and Tegra Architecture



GPU ARCHITECTURE

MOBILE ARCHITECTURE

Jetson TK1 Development Kit

- 32 Bit ARM+Kepler SMX SOC
- CUDA capable
- $192 from US retailers, cost and availability will vary elsewhere
- https://developer.nvidia.com/jetson-tk1
- http://devblogs.nvidia.com/parallelforall/jetson-tk1-mobile-embedded-supercomputer-cuda-everywhere/

Summary

- Who is NVIDIA?
  A: The world's leading Visual Computing company, from consumer devices through to world class supercomputers

- Why should I care about accelerated computing?
  A: Because parallelism and heterogeneous computing is the future of big compute and big data

- What sort of difference can CUDA make?
  A: Order of magnitude improvements in performance and efficiency are possible with CUDA

- Where is NVIDIA going?
  A: Relentlessly forward to a CUDA enabled parallel & energy efficient computing future

Questions ?