NVIDIA GPU コンピューティング
エクサスケールへの道

エヌビディア ジャパン
Tesla Quadro 事業部 マーケティング・マネージャー
林　憲一 (khayashi@nvidia.com)

NVIDIAについて

1993年に設立

設立以来、 半導体企業の中で最速で
10億ドルの収益を達成

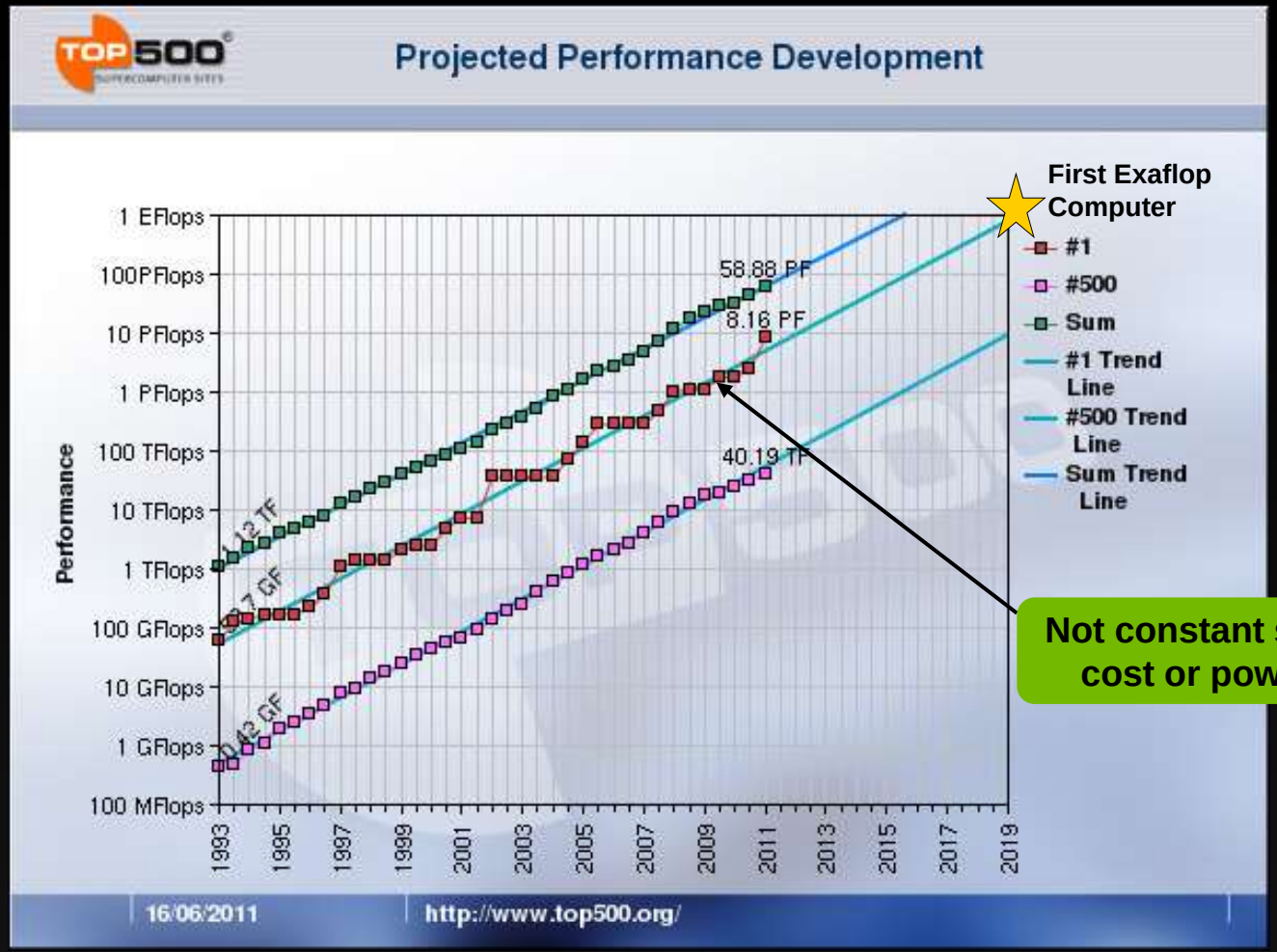
従業員：20カ国に6,800人

特許数：2,000

本社：カリフォルニア州サンタクララ


GeForce
Quadro
NVIDIA.
Tegra
Tesla

Exaflop Expectations



K Computer
~10 MW

CM5
~200 KW

Not constant size, cost or power

The Future Belongs to the *Efficient*

Chips have become power (not area) constrained

density increases *quadratically* with feature size

energy/op decreases *linearly* with feature size

Achieving Energy Efficiency

Reduce overhead

(spend more transistors doing actual work)

Minimize data motion

(data movement is much more expensive than computation)

Which Takes More Energy?

Performing a 64-bit floating-point FMA:

893,500.288914668
× 43.90230564772498
= 39,226,722.78026233027699
\+ 2.02789331400154
= 39,226,724.80815564

Or moving the three 64-bit operands 18 mm across the die:

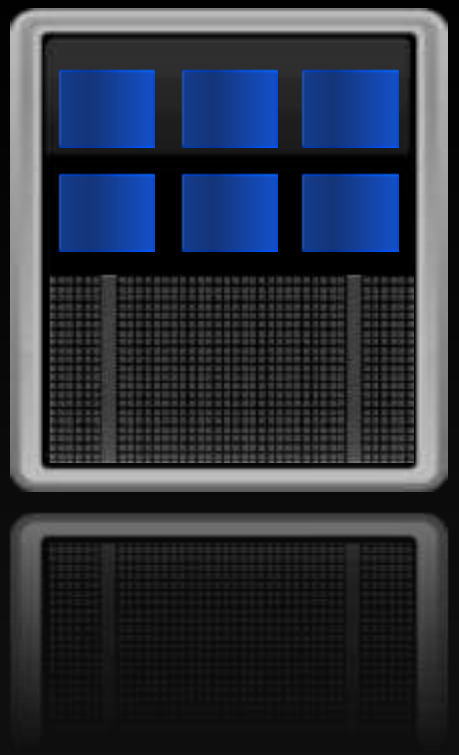
This one takes over 4.2x the energy (40nm)!

It's getting worse: in10nm, relative cost will be 15x!

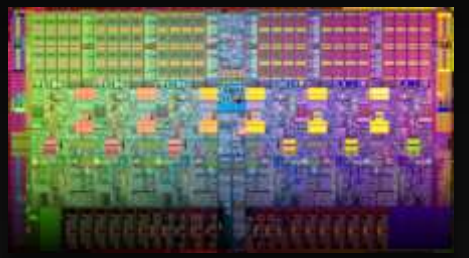
Loading the data from off chip takes >> 100x the energy.

Flops are cheap. ***Communication is expensive.***

Multi-core CPUs



- Industry has gone multi-core as a first response to power issues
  - Performance through parallelism
  - Dial back complexity and clock rate
  - Exploit locality
- But CPUs are fundamentally designed for single thread performance rather than energy efficiency
  - Fast clock rates with deep pipelines
  - Data and instruction caches optimized for latency
  - Superscalar issue with out-of-order execution
  - Dynamic conflict detection
  - Lots of predictions and speculative execution
  - Lots of instruction overhead per operation

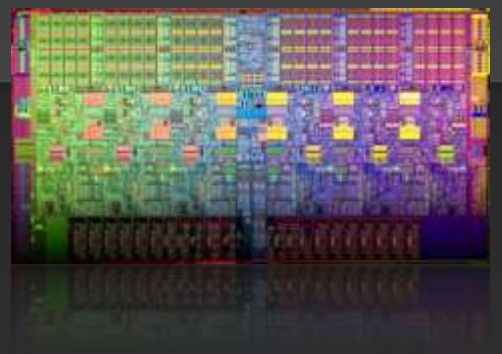


Less than 2% of chip power today goes to flops.

CPU
1690 pJ/FLOP

Optimized for Latency

Caches

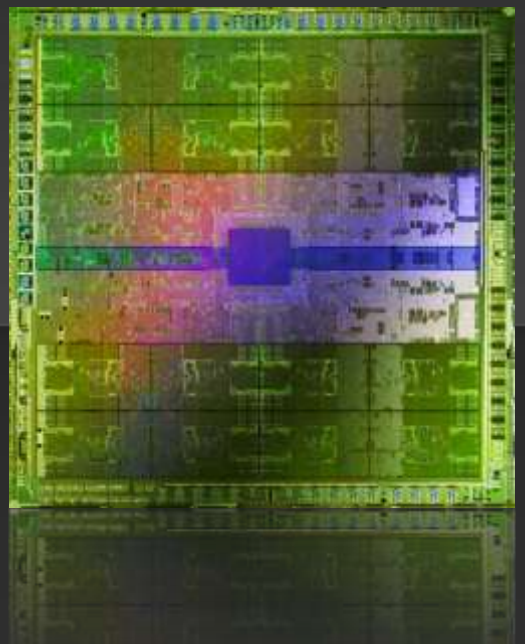


Westmere
32nm

GPU
225 pJ/FLOP

Optimized for Throughput

Explicit Management
of On-chip Memory



Fermi
40nm

Growing Momentum for GPUs in Supercomputing

Tesla Powers 3 of 5 Top Systems (November 2011)

#1 : K Computer
88K Fujitsu Sparc CPUs
10.5 PFLOPS

#2 : Tianhe-1A
7168 Tesla GPUs
2.6 PFLOPS

#3 : Jaguar
36K AMD Opteron CPUs
1.8 PFLOPS

Titan
18000 Tesla GPUs
>20 PFLOPS

#4 : Nebulae
4650 Tesla GPUs
1.3 PFLOPS

#5 : Tsubame 2.0
4224 Tesla GPUs
1.2 PFLOPS
(most efficient PF system)

ORNL Adopts GPUs for Next-Gen Supercomputer

Could not achieve goals using CPUs alone.

Titan Cray XK6

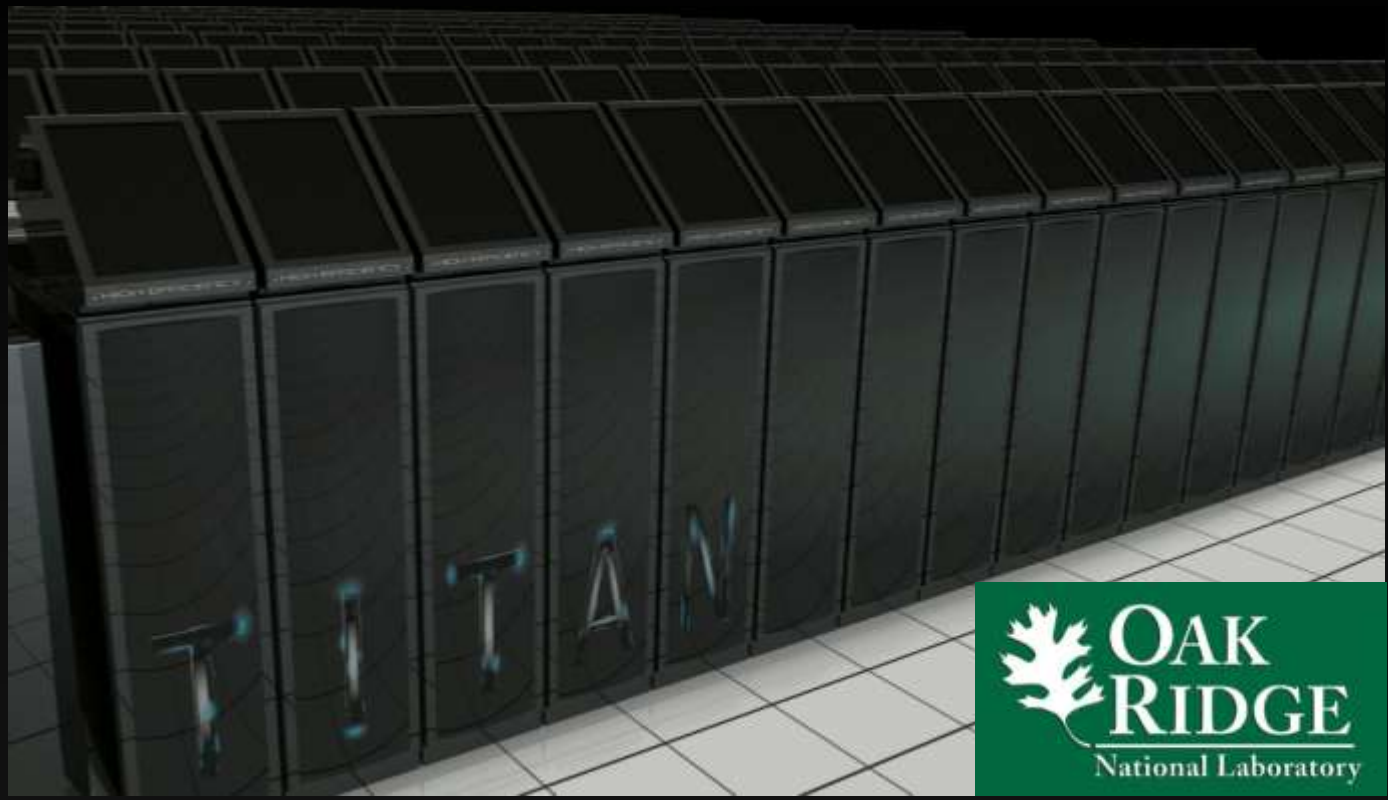
18,000 Tesla GPUs

2x Faster

3x More Energy Efficient

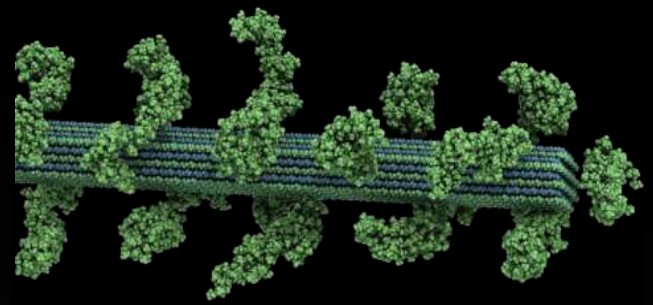
(and much smaller!)

than Current #1 (K Computer)



S3D
Model Combustion for higher efficiency fuels & engines

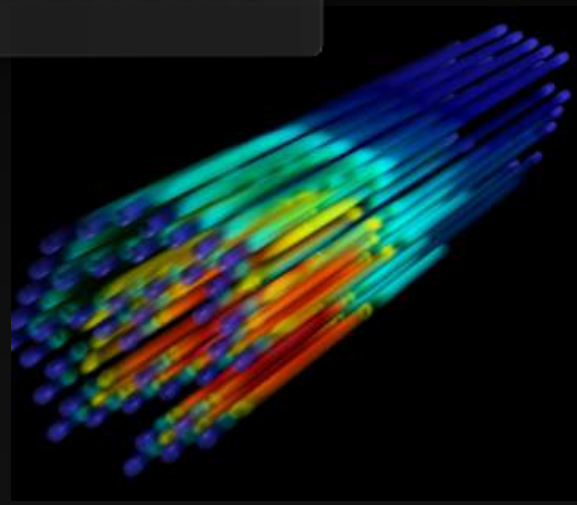
LAMMPS
Model biofuels; Reduce carbon emissions; Reduce need for petroleum



Titan Will Have Huge Societal Benefit

CAM-SE
Model global climate change & explore mitigation strategies

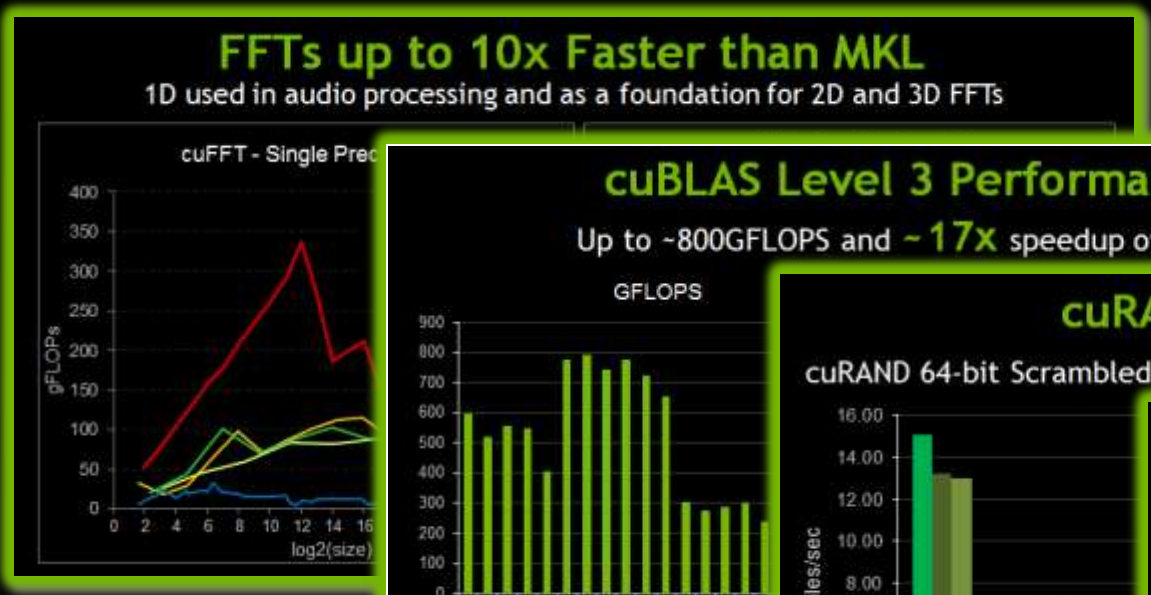
Denovo
Simulate radiation transport for safe, clean, fusion energy



Source: ORNL

Ease of Programming GPUs

GPU Libraries: Plug In & Play



Parallel Algorithms

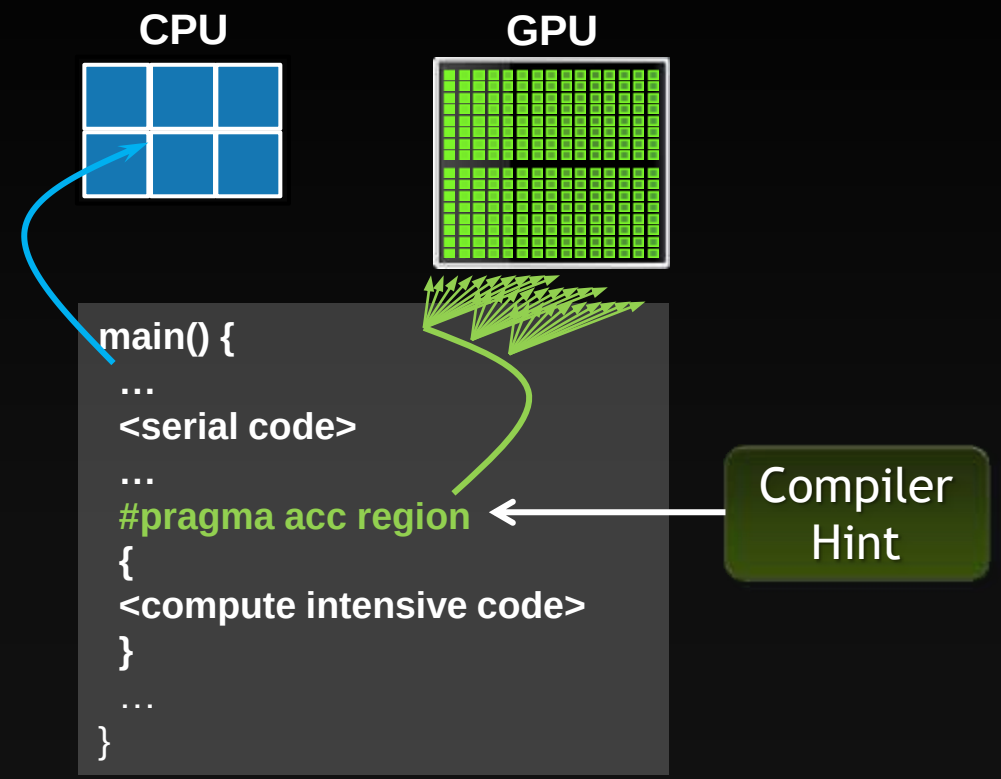
QUDA

Lattice QCD

Dense Linear Algebra

Directives: Ease of Programming and Portability

Available from PGI, CAPS, and Cray



Add hints to code

User focuses on identifying parallelism

Compiler does heavy lifting of parallelizing code

Code works on multicore CPUs & many core GPUs

2x in 4 Weeks. Guaranteed.

Free 30 day trial license
to PGI Accelerator*

Tools for quick ramp

www.nvidia.com/2xin4weeks

*Limit 1000 developers

Small Effort, Huge Speedups

7 Days 3x	2 Days 20x	2 Days 60x	4 Weeks 7x	4 Weeks 10x
	UNIVERSITY of HOUSTON CULLEN COLLEGE of ENGINEERING	THE UNIVERSITY OF MELBOURNE	УГАТУ	COMPUTING FOR SCIENCE fs
Large Oil Company Dr. Jorge Pita Oil exploration at world's largest petroleum reservoirs	**Univ. of Houston** Prof. Kayali Analyzing magneto-static interaction for better storage, memories, and biosensing	**Uni. Of Melbourne** Prof. Black Better understand lifecycles of snapper fish in Port Phillip Bay	**Ufa State Aviation** Prof. Arthur Yuldashev Generating stochastic geological models of oilfield reservoirs with borehole data	**GAMESS-UK** Prof. Karl Wilkinson Used for investigating biofuel production and molecular sensors

OpenACC: Open Parallel Programming Standard

Easy, Fast, Portable

NVIDIA CRAY THE SUPERCOMPUTER COMPANY PGI CAPS

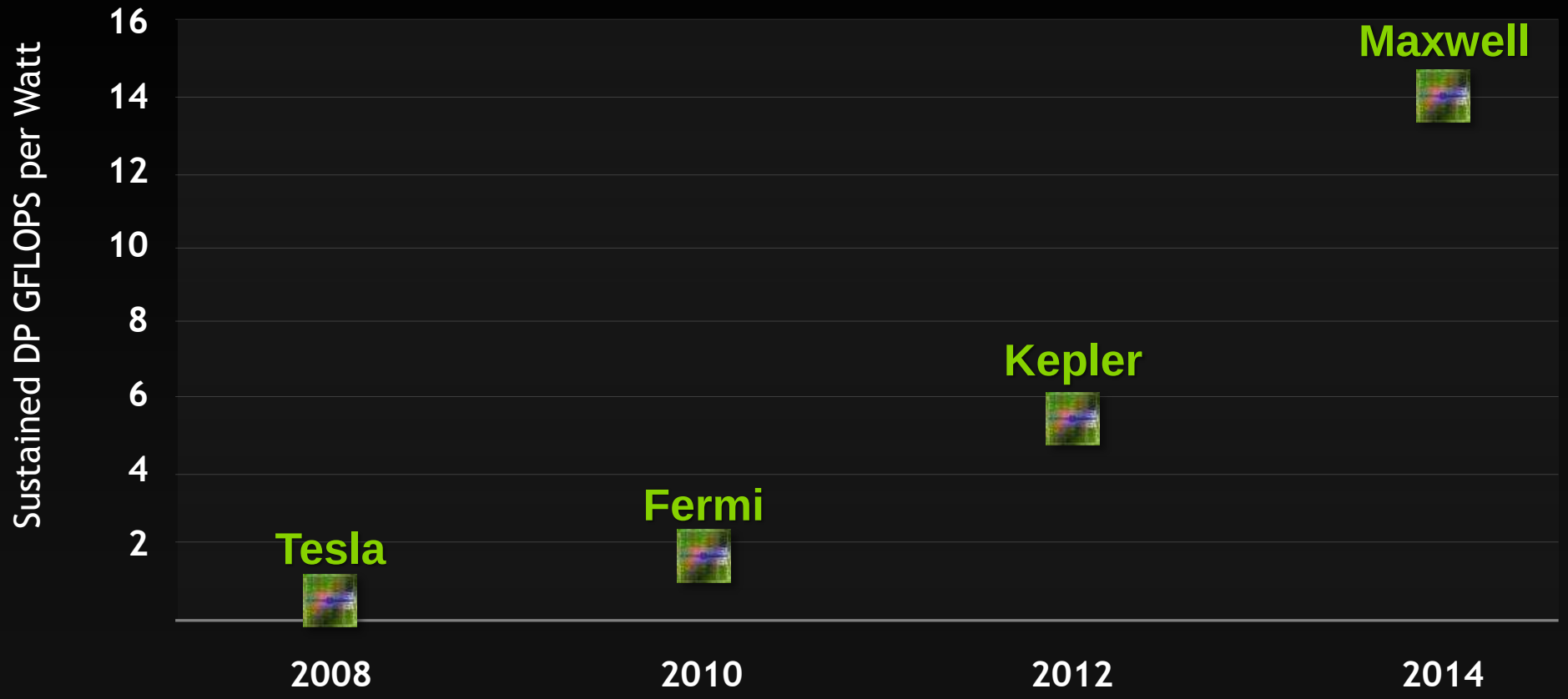
"OpenACC will enable programmers to easily develop portable applications that maximize the performance and power efficiency benefits of the hybrid CPU/GPU architecture of Titan."

Buddy Bland
Titan Project Director
Oak Ridge National Lab

"We look forward to releasing a version of this proposal in the next release of OpenMP."

Michael Wong
CEO, OpenMP
Directives Board

NVIDIA GPU Roadmap: Increasing Performance/Watt

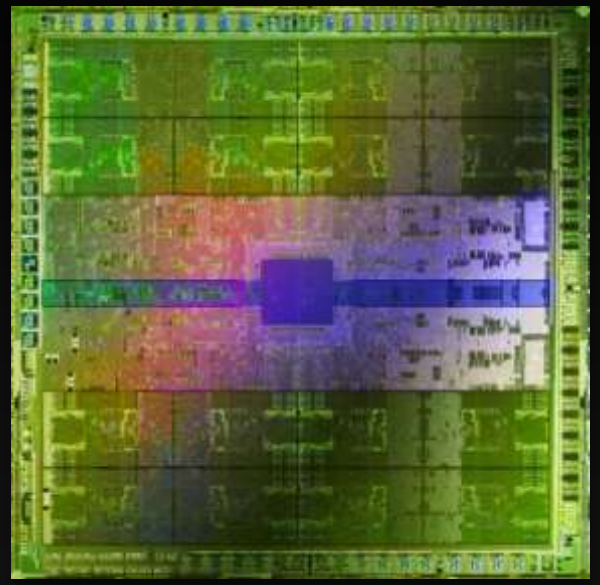


The Future of HPC is Green

- We’re constrained by power
- You can’t simultaneously optimize for single thread performance *and* power efficiency

- The future is heterogeneous
  - A few fast cores for serial work
  - Most cores optimized for power efficiency

- GPUs are the right path to this future
  - Designed for power efficiency
  - Leverage high-volume graphics business




Thank You!