CARMA
CUDA on ARM Architecture

Developing Accelerated Applications on ARM

NVIDIA

CARMA is an architectural prototype for
high performance, energy efficient hybrid computing

Schedule

- **Motivation**
- **System Overview**
- **System Details**
- **Q&A with Demonstration**

Motivation

HPC systems will be capped by power and thermal limits

- **The world's largest supercomputer systems are near their physical limits**
- **Broader market HPC installations are capped by pragmatic and site limits**

The cluster revolution was driven by

- **Cost-effective computing**
  - **Dollars per FLOP**
- **Transferable knowledge and accessibility**
  - **Skills and tools developed on personal-scale machines**
- **Long-term viable architecture**
  - **Commodity market components used at a larger scale**

We now need to incorporate power-efficient computing

The next revolution: Power Efficiency

Once again, look to commodity market for the next generation

Power-effective computing is driven by phones and tables

- ARM has an architectural and experience advantage
- System-level software complexity is high
  - Most power optimization work is being done for ARM

High performance power-efficent computing from GPGPUs

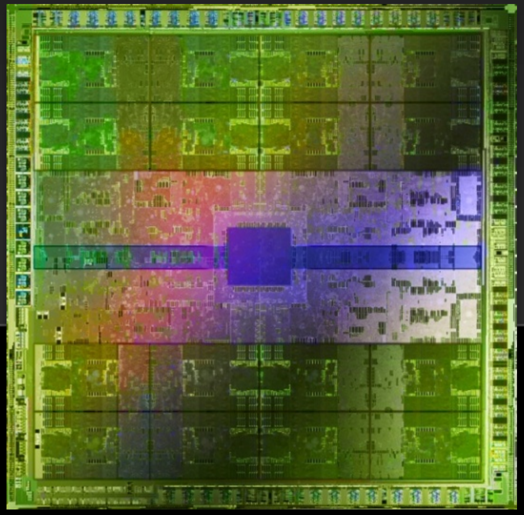
- GPUs have an architectural efficency advantage
- Many applications already effectively use GPUs

GPU

225 pJ/flop

Optimized for throughput and power efficiency

Explicit management of on-chip memory



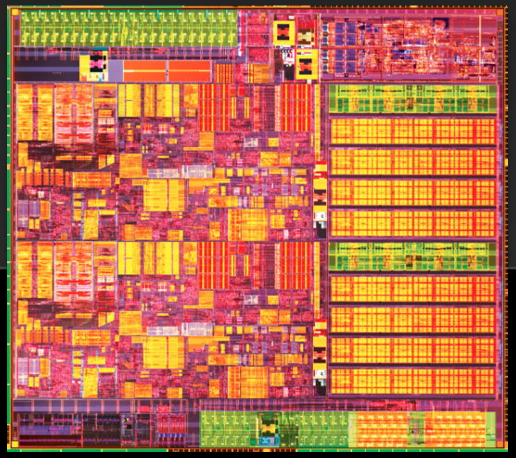
Fermi
40 nm

CPU

1700 pJ/flop

Optimized for latency

Caches



Westmere
32 nm

Why CARMA?

- Have a real prototype platform for these future HPC systems
- Explore the efficiency and performance trade-offs for existing ARM+GPU systems
- Check, tune and evaluate CUDA accelerated applications

Enabling ARM Ecosystem: CARMA DevKit

CUDA on ARM

CUDA GPU

Tegra ARM CPU

Tegra 3 Quad-core ARM A9
Quadro 1000M (96 CUDA cores)
Ubuntu

Gigabit Ethernet
SATA Connector
HDMI, DisplayPort, USB

CARMA Hardware Overview

Available from SECO

Ultra low power host CPU

- Tegra T30 “Kal-El”
- Four ARM A9 cores with NEON and VFPv3 extensions
- Q7 module

NVIDIA GPU for GPU computing

- Quadro1000m on PCIe
- 96 CUDA cores with 200GFLOPS SP peak
- MXM module

CARMA Software Overview

- ARM Linux distribution
  - Ubuntu 11.04 for ARM
  - Linux 3.1.10 kernel
  - Enhancements to support Tegra features
- CUDA 4.2 run-time and libraries
- Host x86 system support for cross development
  - CUDA cross-compiler

Developer Information

- For support and questions, register on the CUDA DevZone
  - http://www.nvidia.com/carmadevkit
  - http://www.nvidia.com/devzone
- Future enhancements
  - Native (ARM hosted) compile support
  - Updated CUDA versions e.g. CUDA 5.0
- Long term plans for the CARMA platform
  - ARMv8 64 bit platform support

CARMA

CUDA on ARM Architecture

QUESTIONS & ANSWERS

NVIDIA

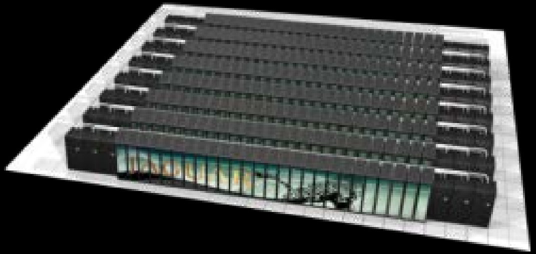
Back-up material

Growing Momentum for GPUs in Supercomputing

Tesla Powers 3 of 5 Top Systems

#1 : K Computer
68K Fujitsu Sparc CPUs
8.2 PFLOPS

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



#3 : Jaguar
36K AMD Opteron CPUs
1.8 PFLOPS

→

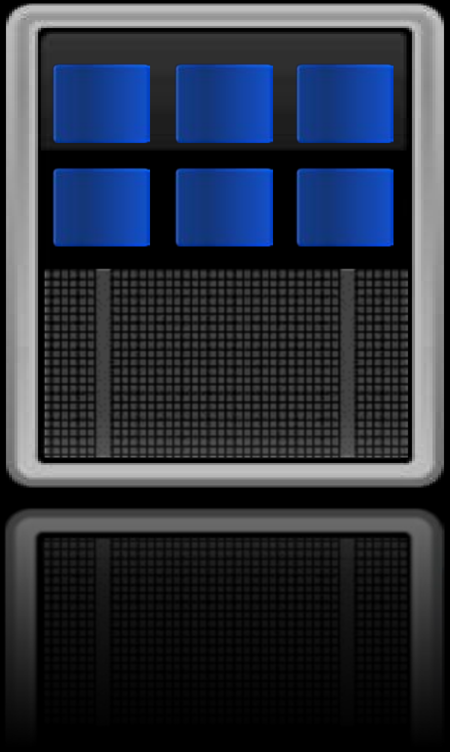
Titan
18000 Tesla GPUs
>25 PFLOPS

#4 : Nebulae
4650 Tesla GPUs
1.3 PFLOPS

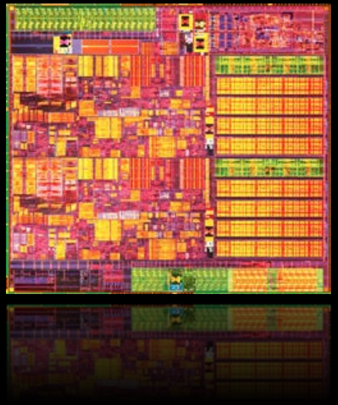
#5 : Tsubame 2.0
4224 Tesla GPUs
1.2 PFLOPS
(most efficient PF system)

14

Multi-core CPUs

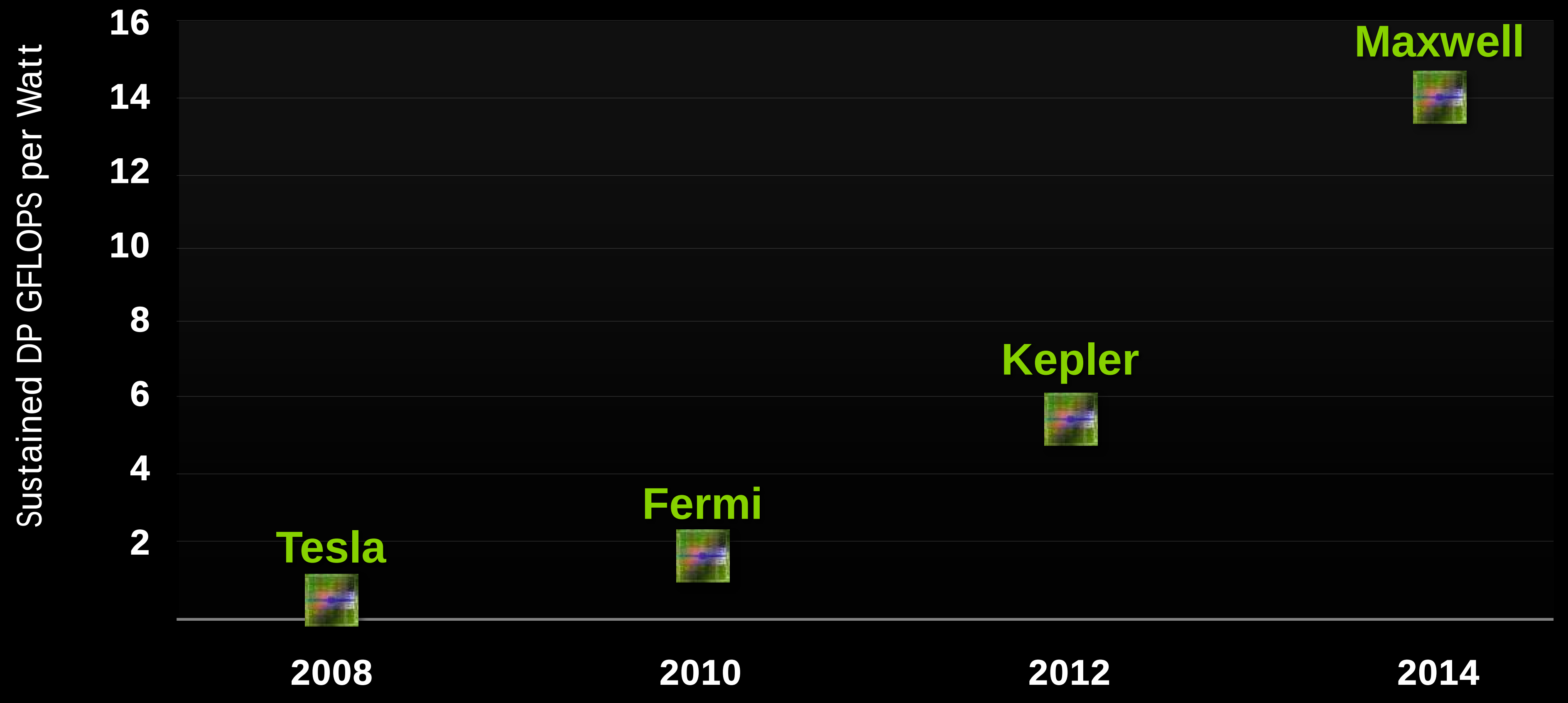


- Multi-core as a first response to power issues
  - Performance through parallelism, not frequency increases
  - Slow the complexity spiral
  - Better locality in many cases
- But CPUs have evolved for single thread performance rather than energy efficiency
  - Fast clock rates with deep pipelines
  - 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

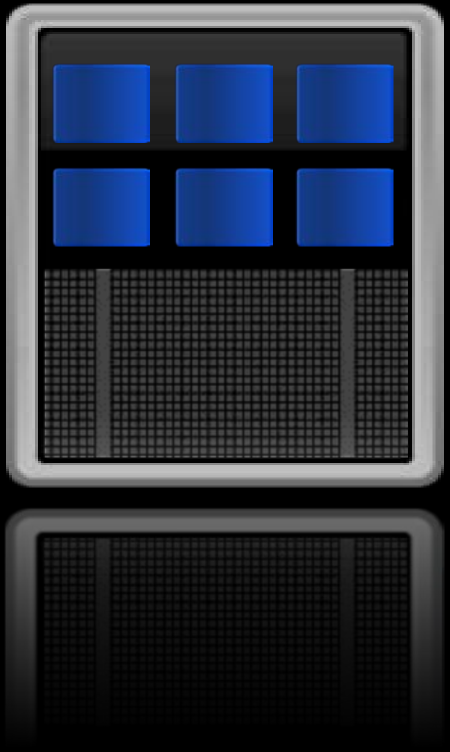


Less than 2% of chip power today goes to flops.

NVIDIA GPU Roadmap: Increasing Performance/Watt

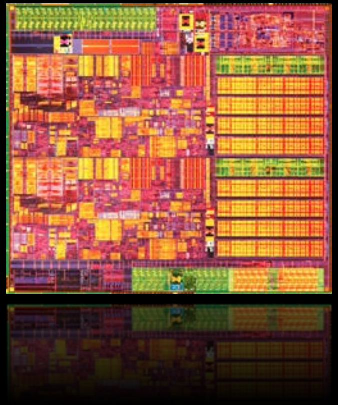


Possible Power-efficient Future



Power-efficient general core combined with GPU

- Power control shared with mobile products
  - Ultra-focused on power efficiency
  - Aggressive market forces innovation
- Technology evolution driven by commodity market
- Bulk of compute power provided by inherently efficient GPUs



Increase to over 50% of chip power for flops.

World’s First ARM CPU / CUDA GPU Supercomputer

MONT BLANC

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

http://www.montblanc-project.eu

- Mont Blanc Research project
- Exploring energy efficient supercomputer architectures
- Working towards exascale

World's Greenest Petaflop Supercomputer

Tsubame 2.0

Tokyo Institute of Technology

- 1.19 Petaflops
- 4,224 Tesla M2050 GPUs
- 0.85 sustained GF/W