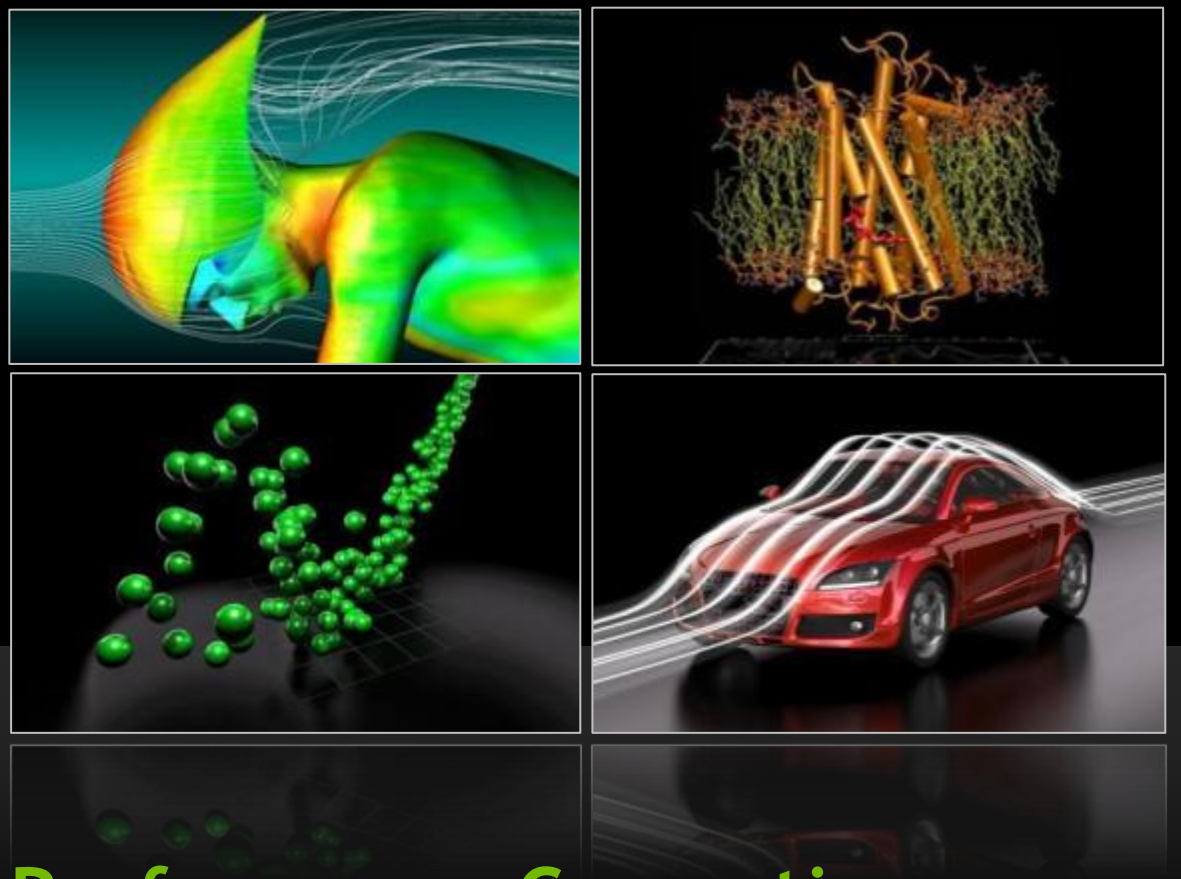
TESLA

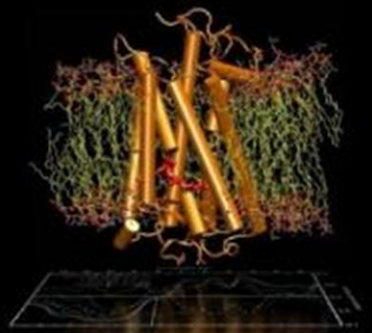
GPU Computing



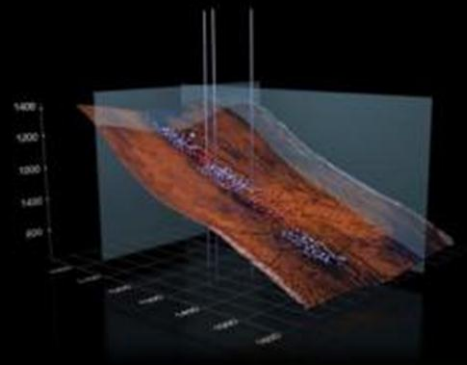
Accelerating High Performance Computing

http://www.nvidia.com/tesla

Computing – The 3rd Pillar of Science

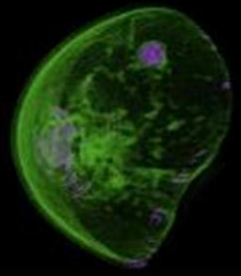


Drug Design
Molecular Dynamics

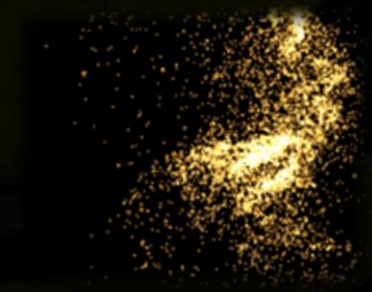


Seismic Imaging
Reverse Time Migration

Automotive Design
Computational Fluid Dynamics

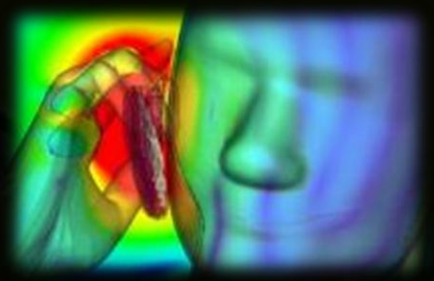


Medical Imaging
Computed Tomography

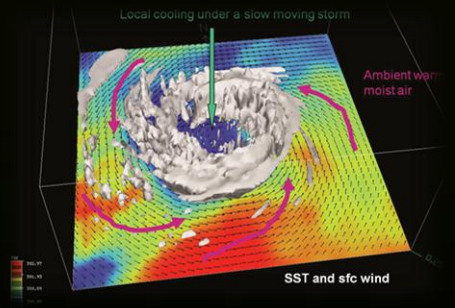


Astrophysics
n-body

Options Pricing
Monte Carlo

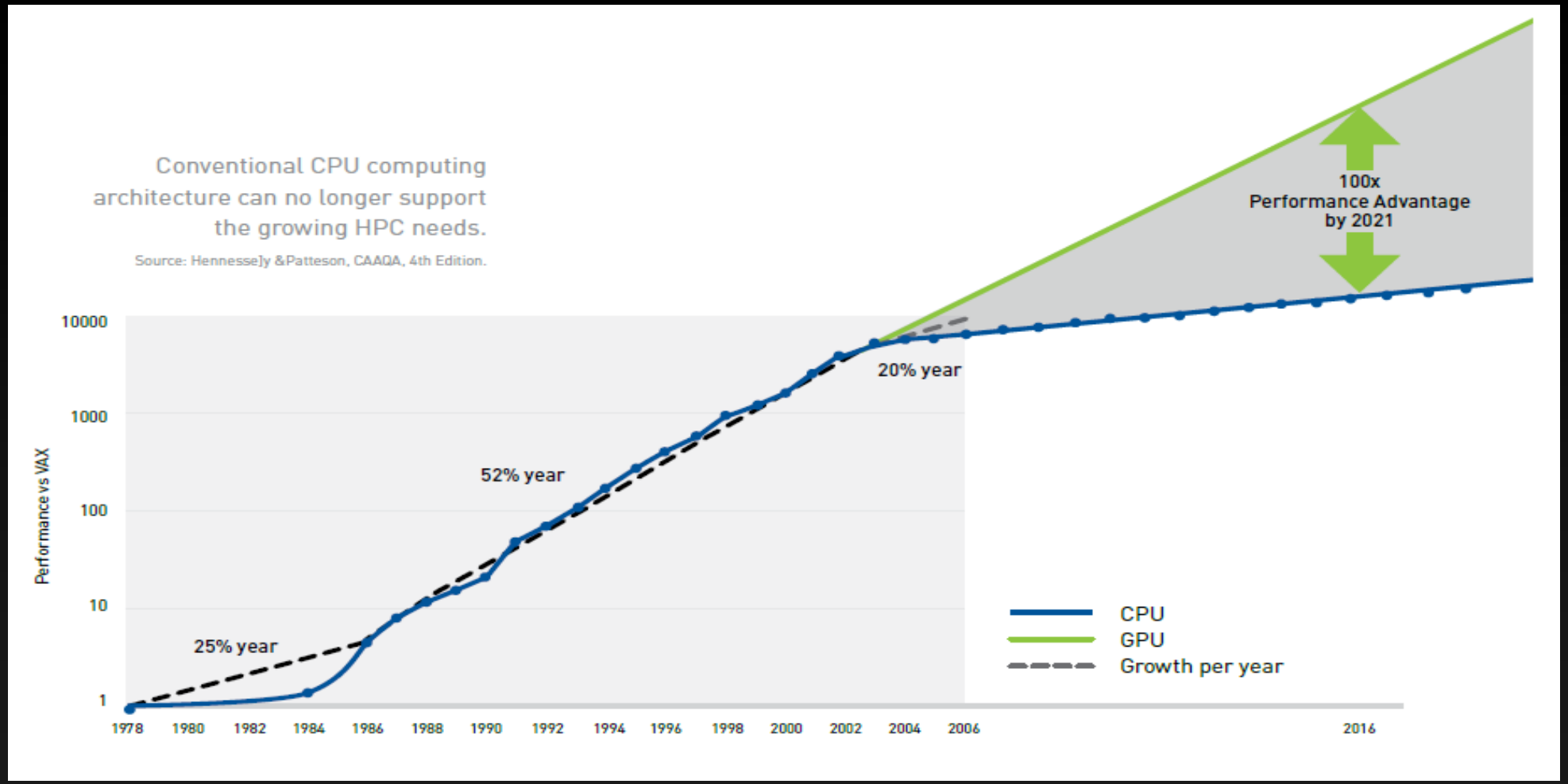


Product Development
Finite Difference Time Domain



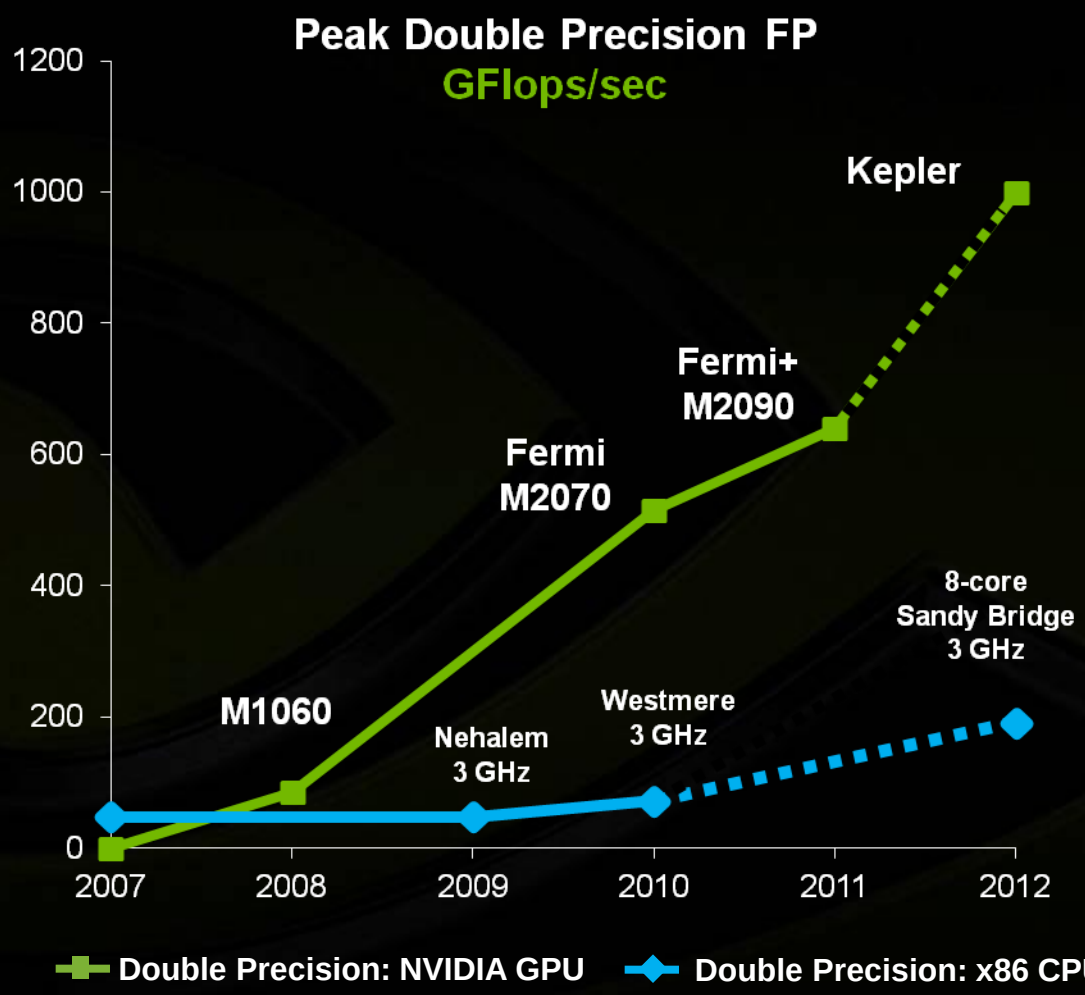
Weather Forecasting
Atmospheric Physics

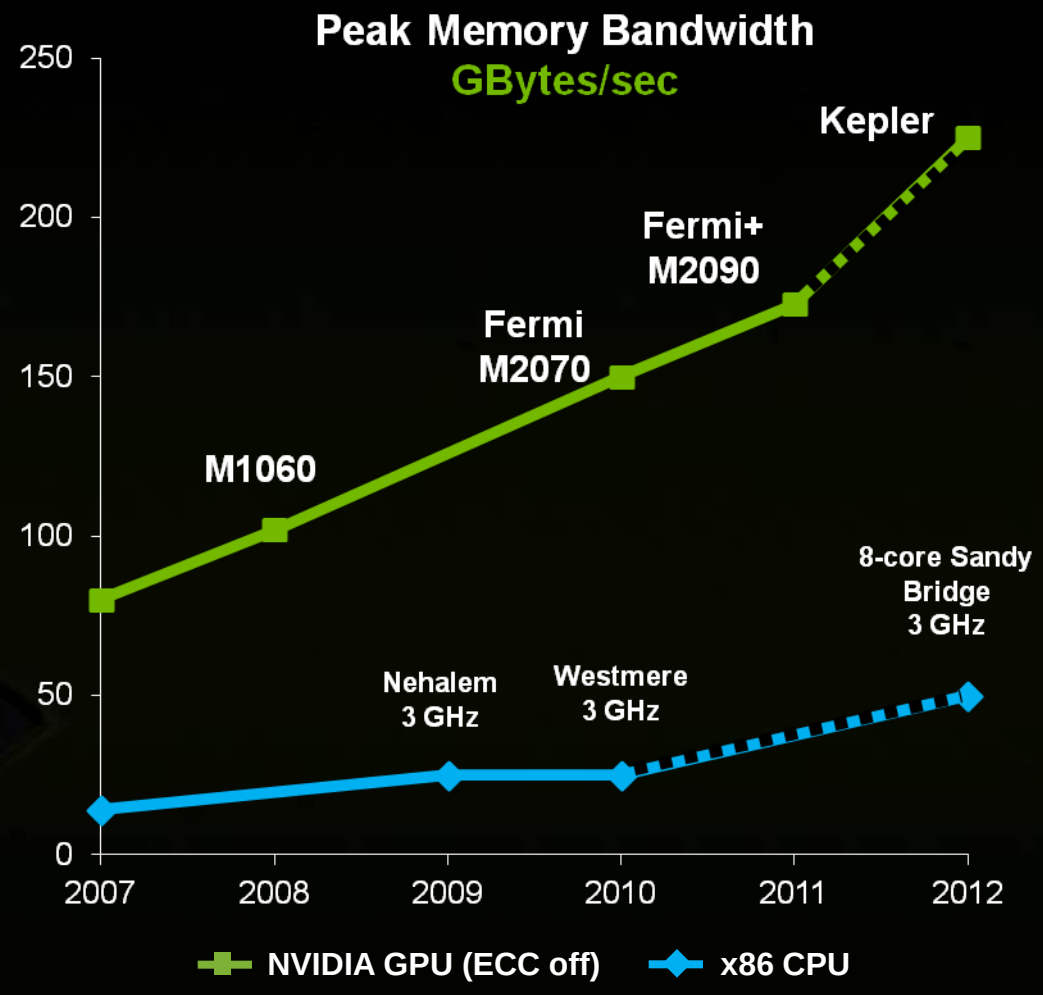
GPU Computing Bridging the CPU Wall



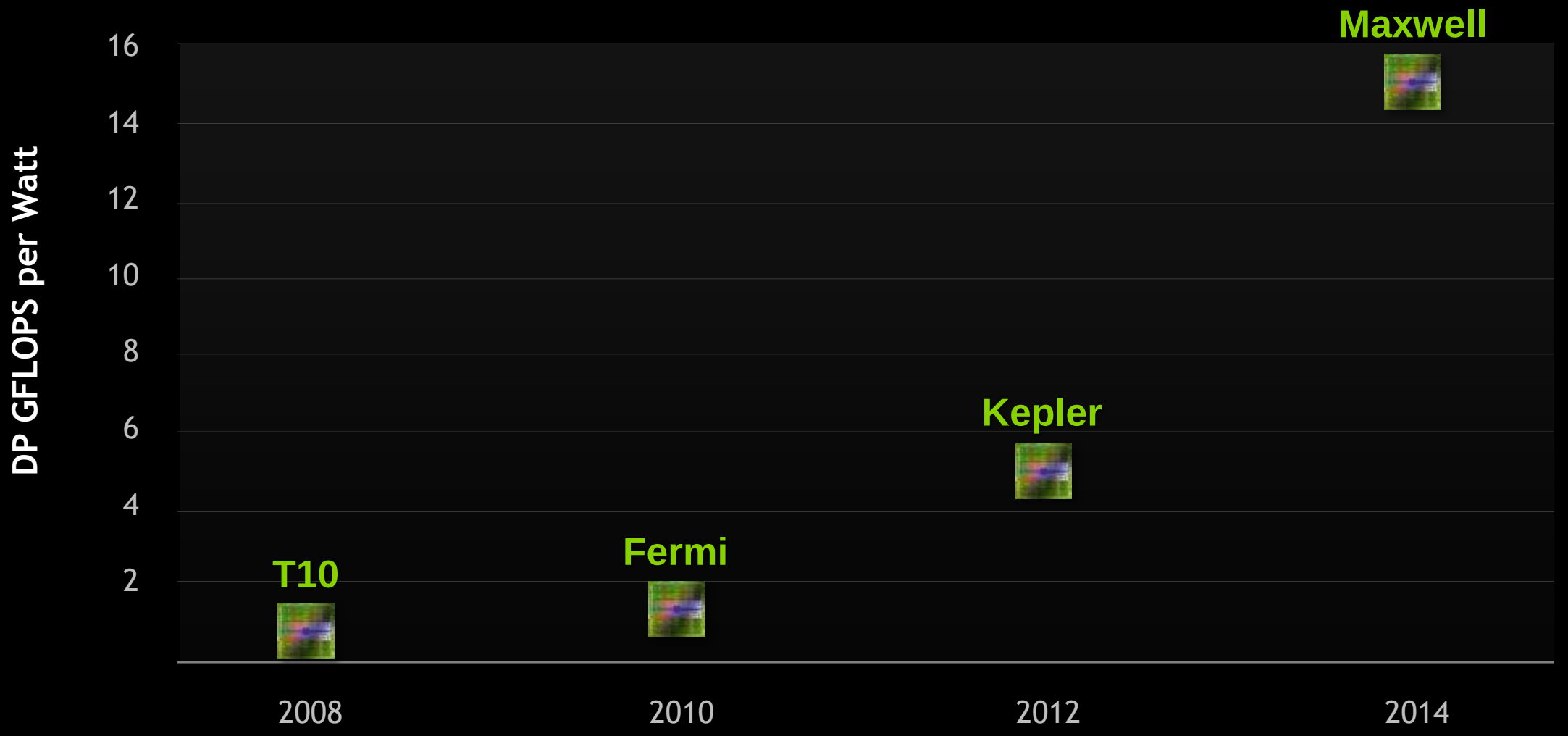
Source: David Patterson, Director Parallel Computing Research Laboratory, U.C. Berkeley

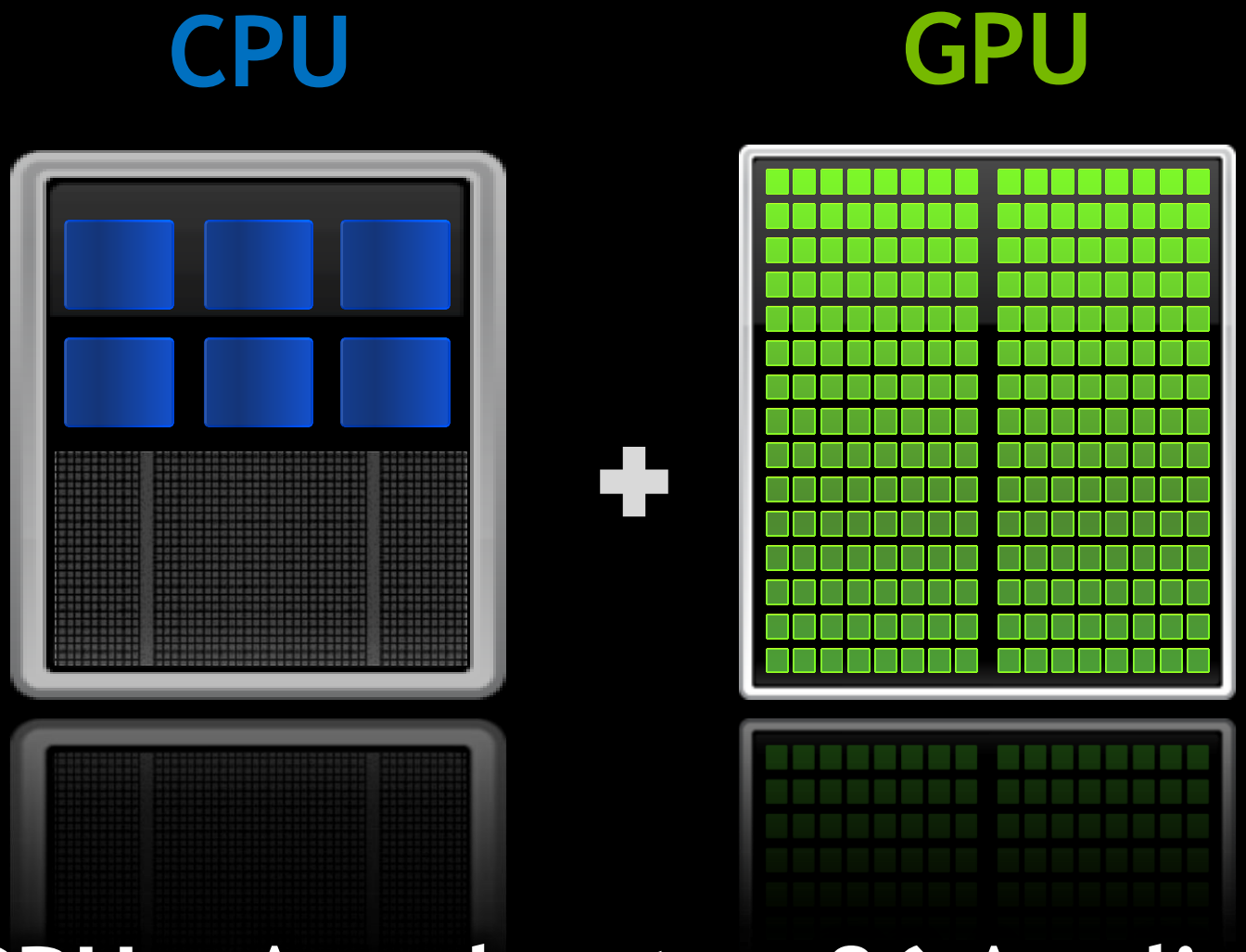
GPUs = Higher Flops and Memory Bandwidth





Tesla: 2-3x Faster GPU Every 2 Years

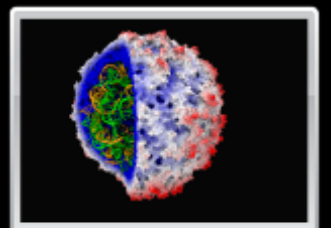




Add GPUs: Accelerate x86 Applications

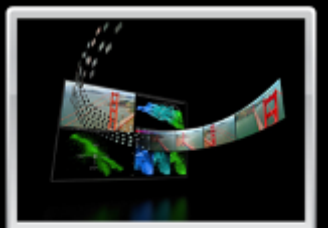
146X

Medical Imaging
U of Utah



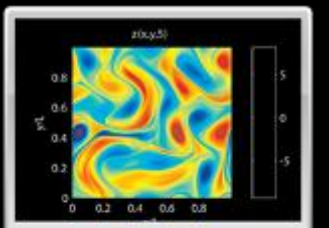
36X

Molecular Dynamics
U of Illinois, Urbana



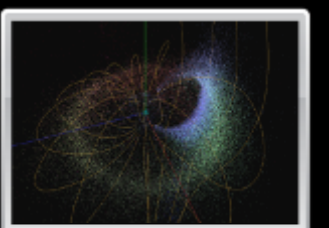
18X

Video Transcoding
Elemental Tech



50X

Matlab Computing
AccelerEyes



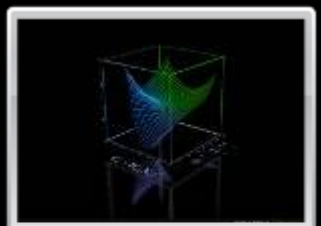
100X

Astrophysics
RIKEN

GPUs Accelerate Science

149X

Financial Simulation
Oxford



47X

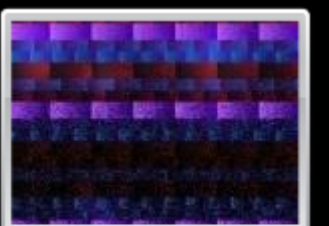
Linear Algebra
Universidad Jaime

20X

3D Ultrasound
Techniscan

130X

Quantum Chemistry
U of Illinois, Urbana

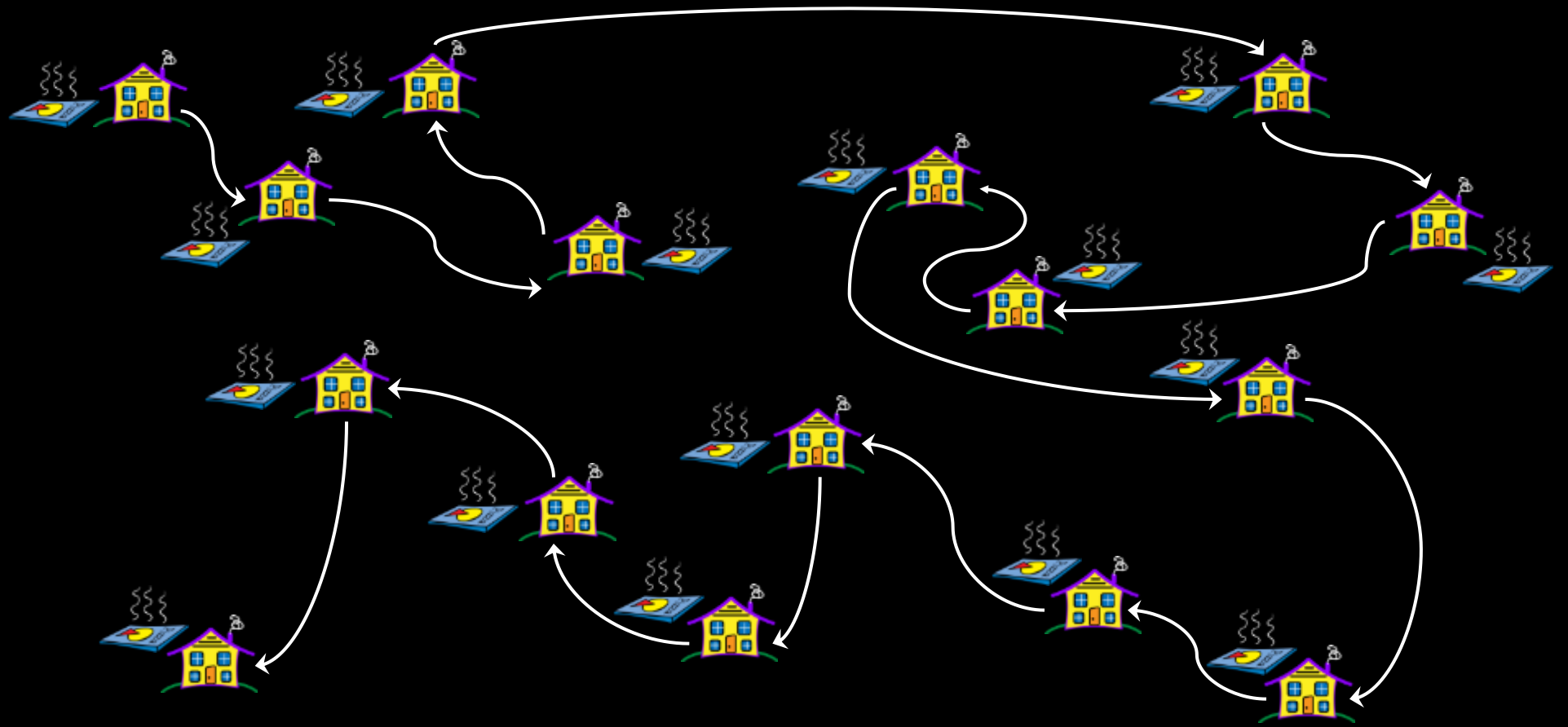


30X

Gene Sequencing
U of Maryland

CPU Pizza Delivery

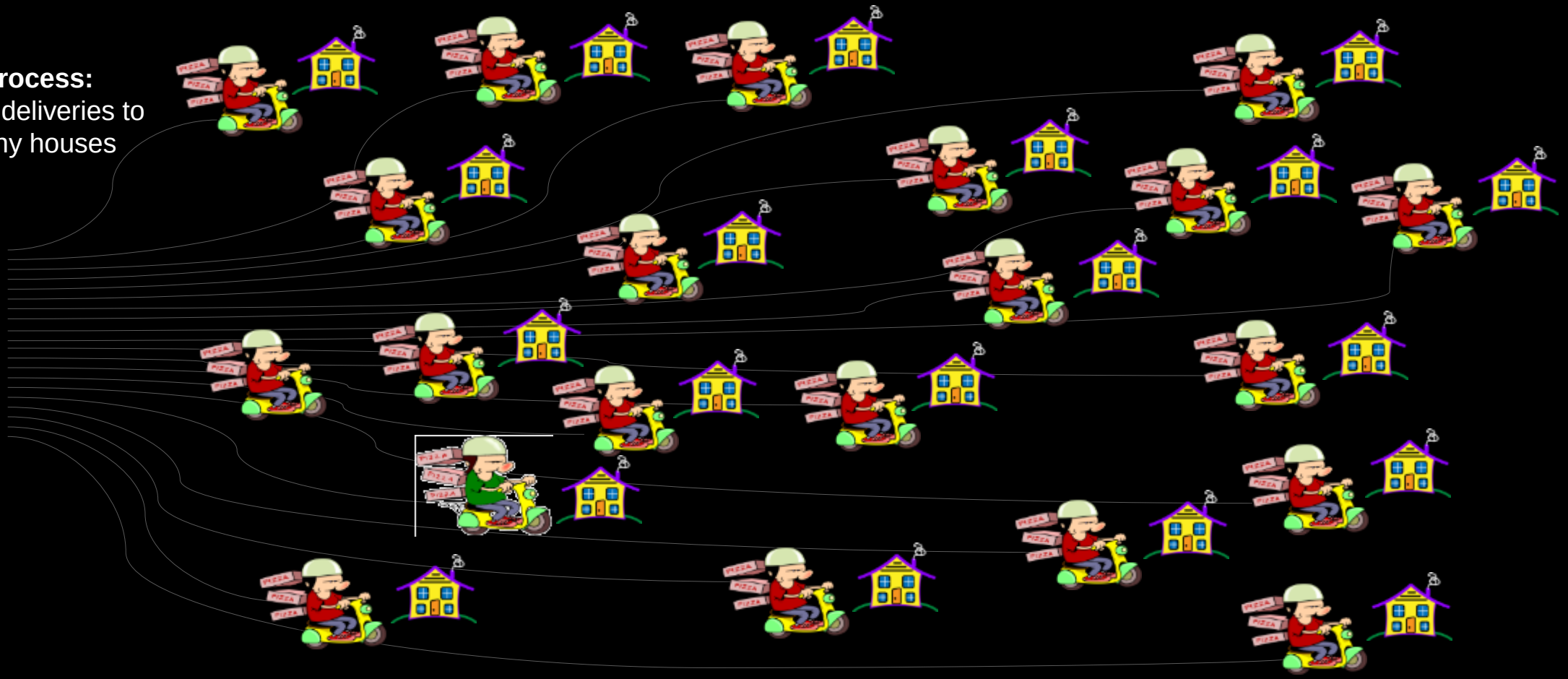
Process:
Delivery truck delivers one pizza and then moves to next house



Original Idea by Jedox
www.jedox.com

NVIDIA GPU Pizza Delivery

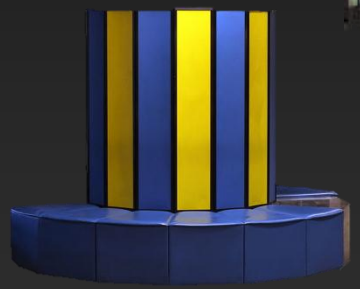
Process:
Many deliveries to many houses



Original Idea by Jedox
www.jedox.com

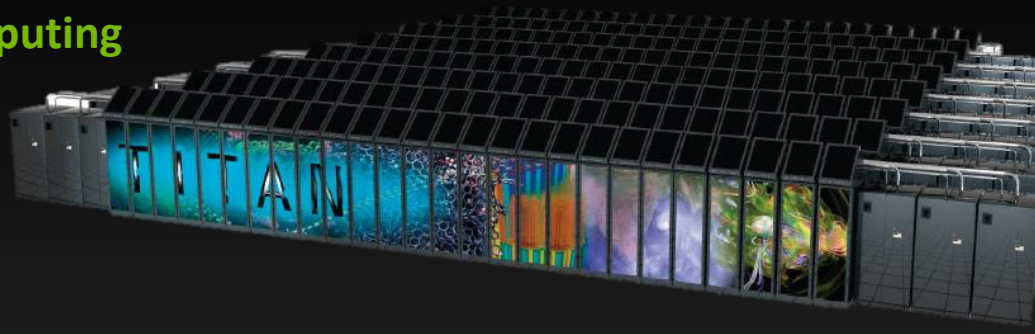
The Era of Accelerated Computing is Here

Era of
Vector Computing



Era of
Distributed Computing

Era of
Accelerated Computing



1980 1990 2000 2010 2020

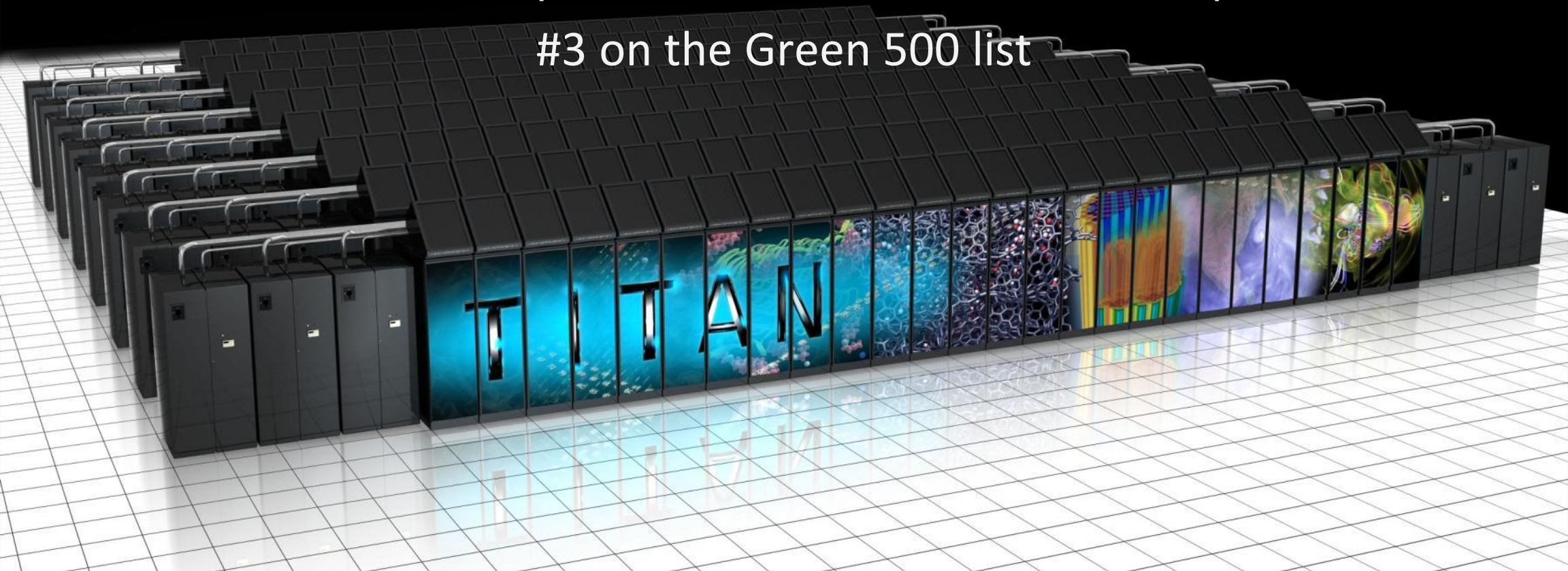
Titan: World's Fastest Supercomputer 2012

18,688 Tesla K20X GPUs

27 Petaflops Peak: 90% of Performance from GPUs

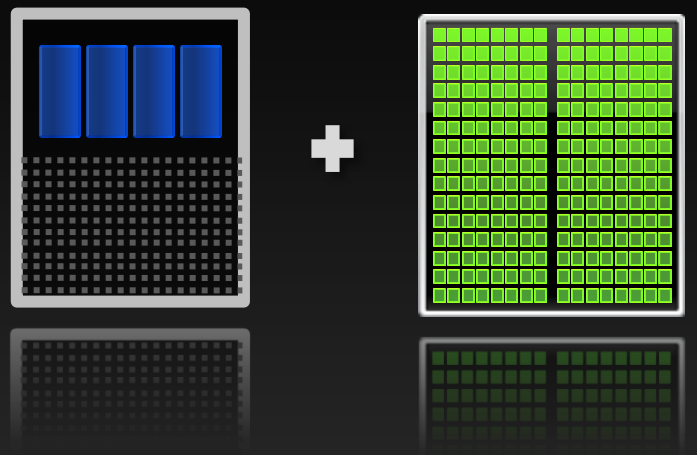
17.59 Petaflops Sustained Performance on Linpack

#3 on the Green 500 list



Two Supercomputers Built at the Same Time

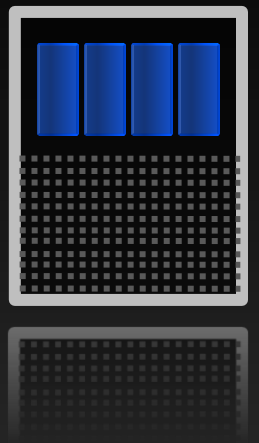
Tsubame 2.0



4,224 Tesla GPUs + 2,816 x86 CPUs

1.4 Megawatts
2060 Homes in Japan

Hopper- NERSC

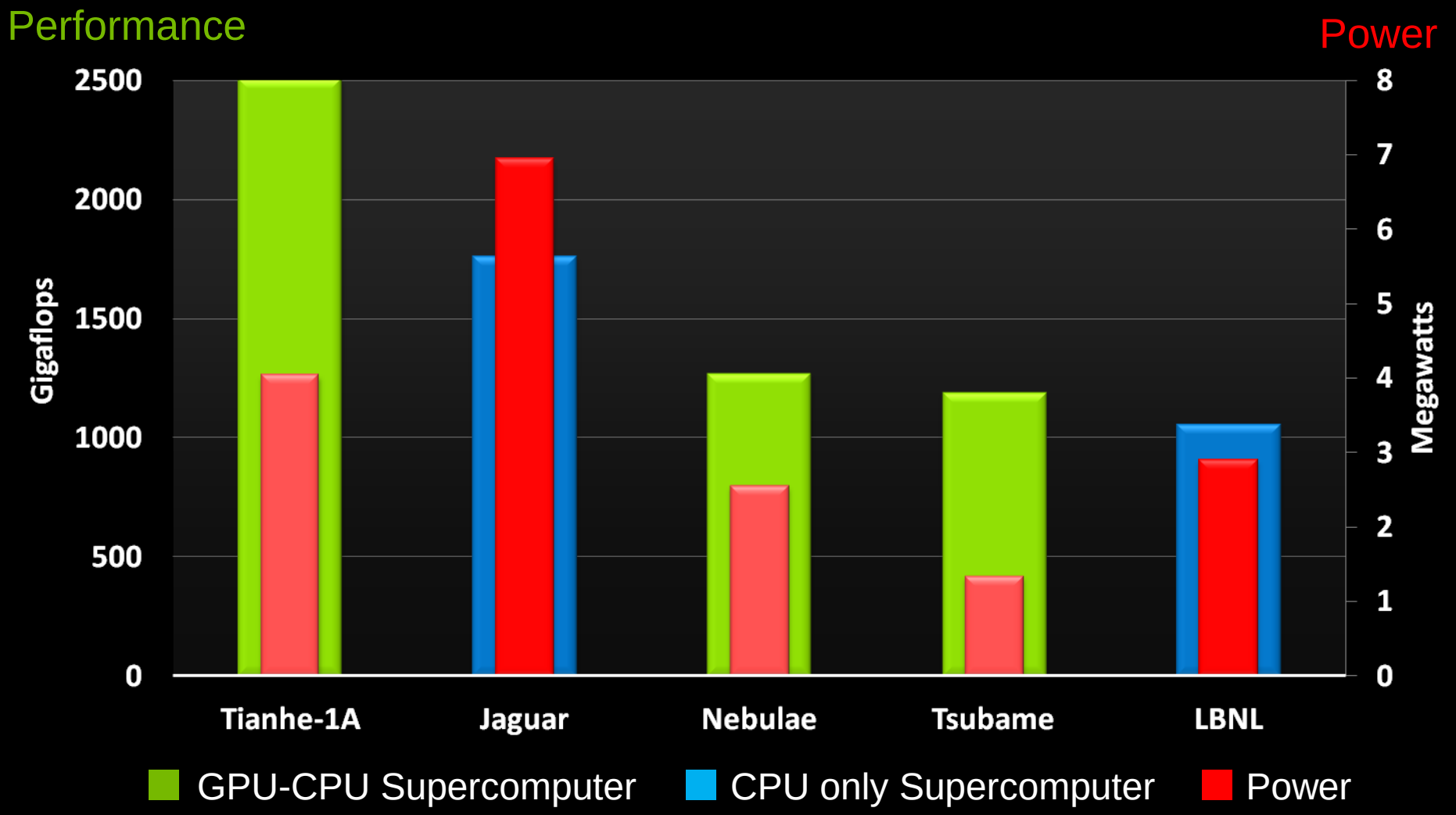


12,784 x86 CPUs

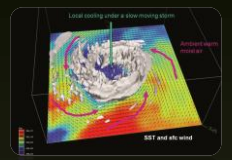
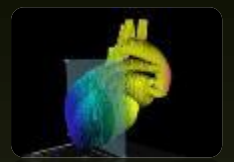
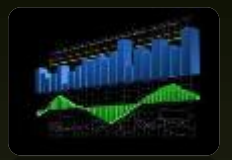
4.0 MegaWatts
5860 Homes in Japan

World's Greenest Petaflop Supercomputer (2011)

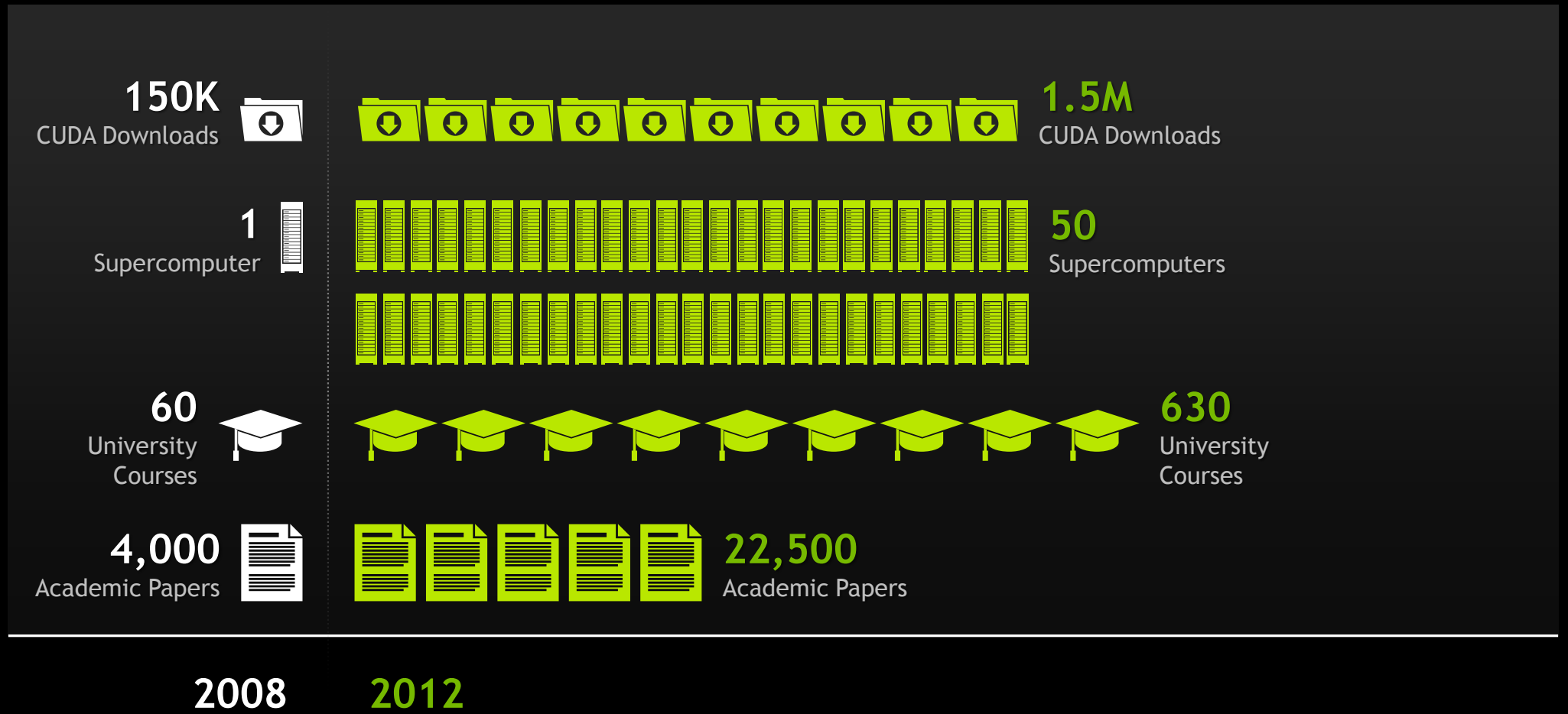
GPU Supercomputers: More Power Efficient



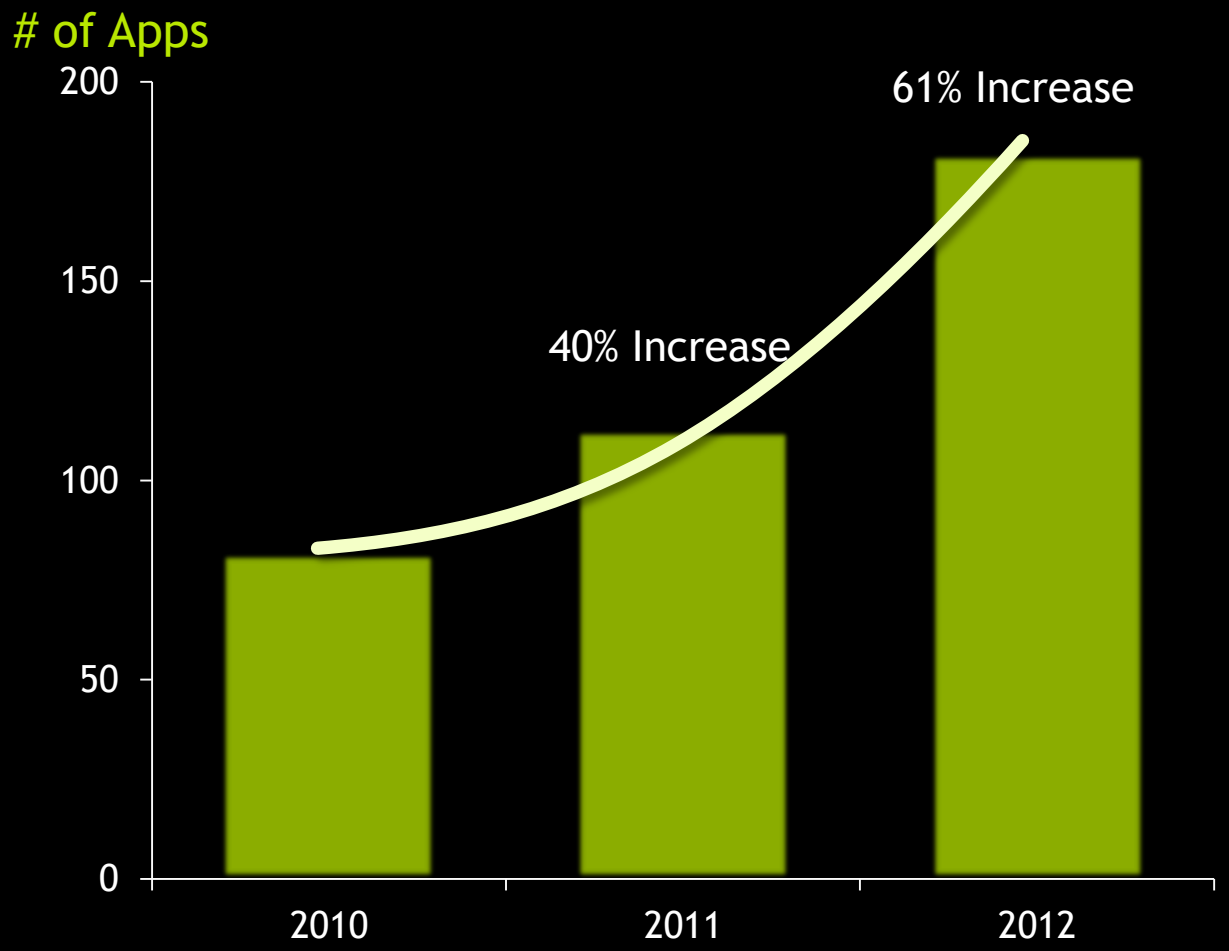
GPUs are Mainstream

Oil & Gas	Edu/Research	Government	Life Sciences	Finance	Manufacturing
					
Schlumberger	Chinese Academy of Sciences	Air Force Research Laboratory	Boston Scientific	Bloomberg	Agilent
BR PETROBRAS	Georgia Tech	NASA	MGH 1811 Mass General Hospital	BNP PARIBAS	ANSYS
TOTAL Chevron	HARVARD School of Engineering and Applied Sciences	Naval Research Laboratory	Max Planck Institute	STANDARD LIFE	Autodesk
Paradigm	OAK RIDGE National Laboratory	BAE SYSTEMS	BECKMAN COULTER	J.P.Morgan	SIMULIA
				NumeriX	ACUSIM SOFTWARE

Explosive Growth of GPU Computing



CUDA Apps Grows 60%, Accelerating Key Apps



For App Catalog: http://www.nvidia.com/teslaapps/

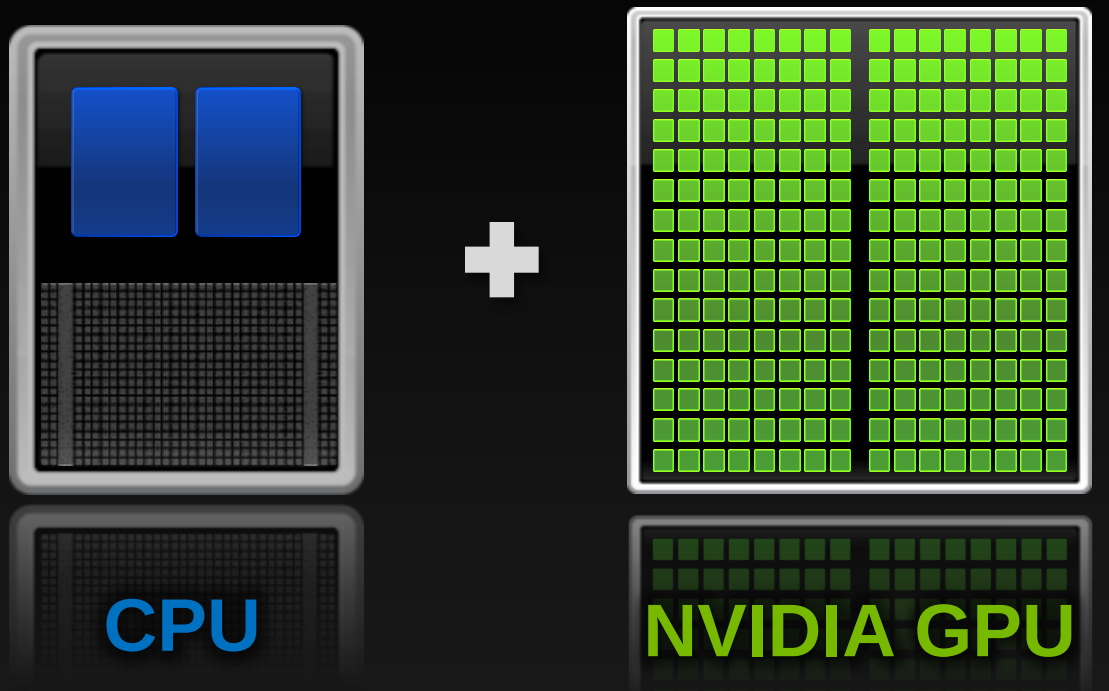
Top Supercomputing Apps		
Computational Chemistry	AMBER CHARMM GROMACS	LAMMPS NAMD DL_POLY
Material Science	QMCPACK Quantum Espresso GAMESS	Gaussian NWChem VASP
Climate & Weather	COSMO GEOS-5	CAM-SE NIM WRF
Physics	Chroma Denovo GTC	GTS ENZO MILC
CAE	ANSYS Mechanical MSC Nastran SIMULIA Abaqus	ANSYS Fluent OpenFOAM LS-DYNA

Accelerated, In Development

Accelerated computing

NVIDIA GPU Accelerates Computing

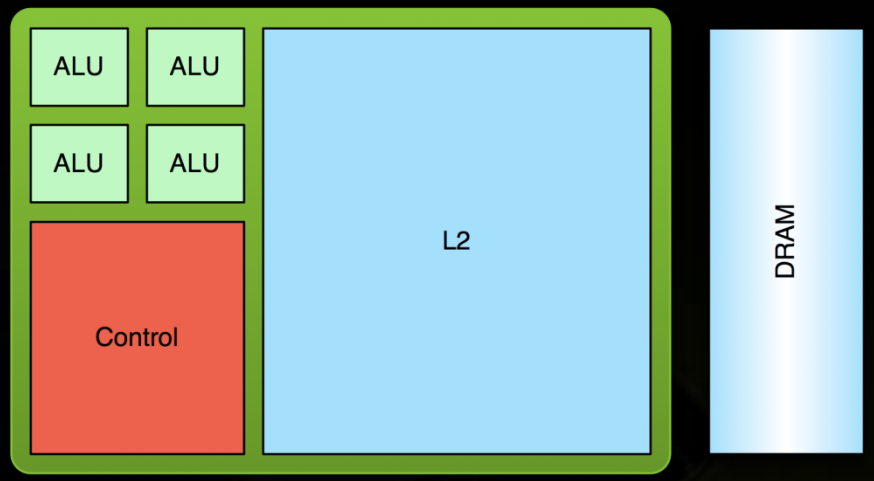
Choose the Right Processor for the Right Task



Several sequential cores

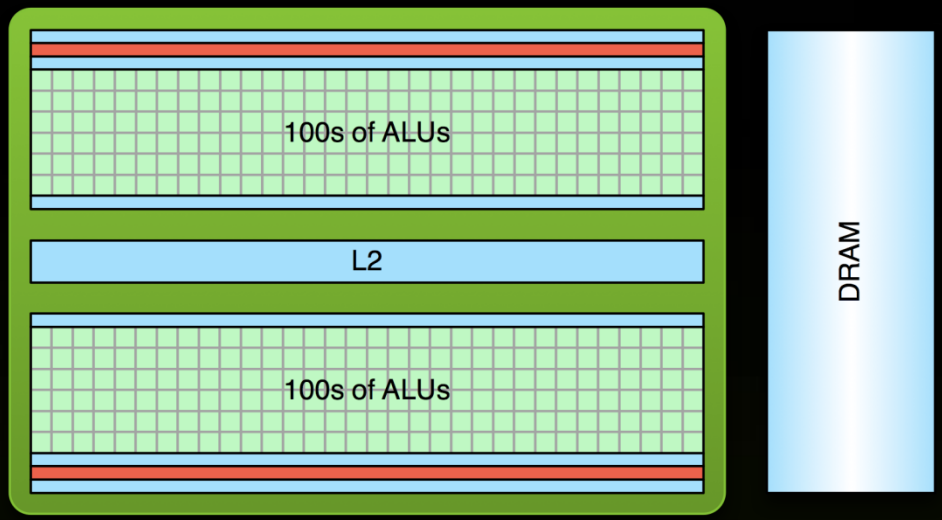
Hundreds of parallel cores

Low Latency or High Throughput?



CPU

- Optimized for low-latency access to cached data sets
- Control logic for out-of-order and speculative execution

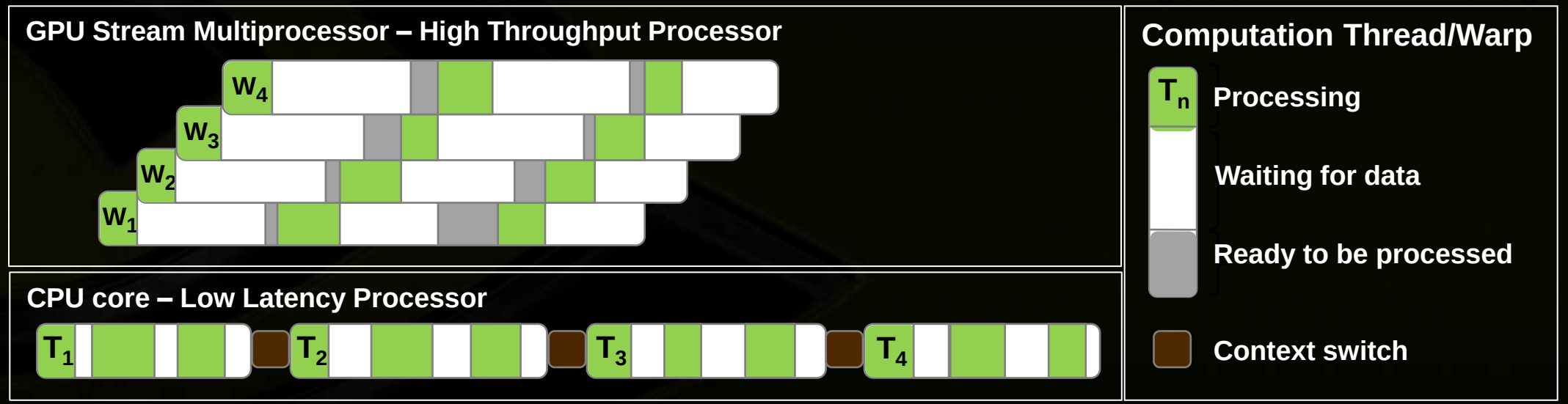


GPU

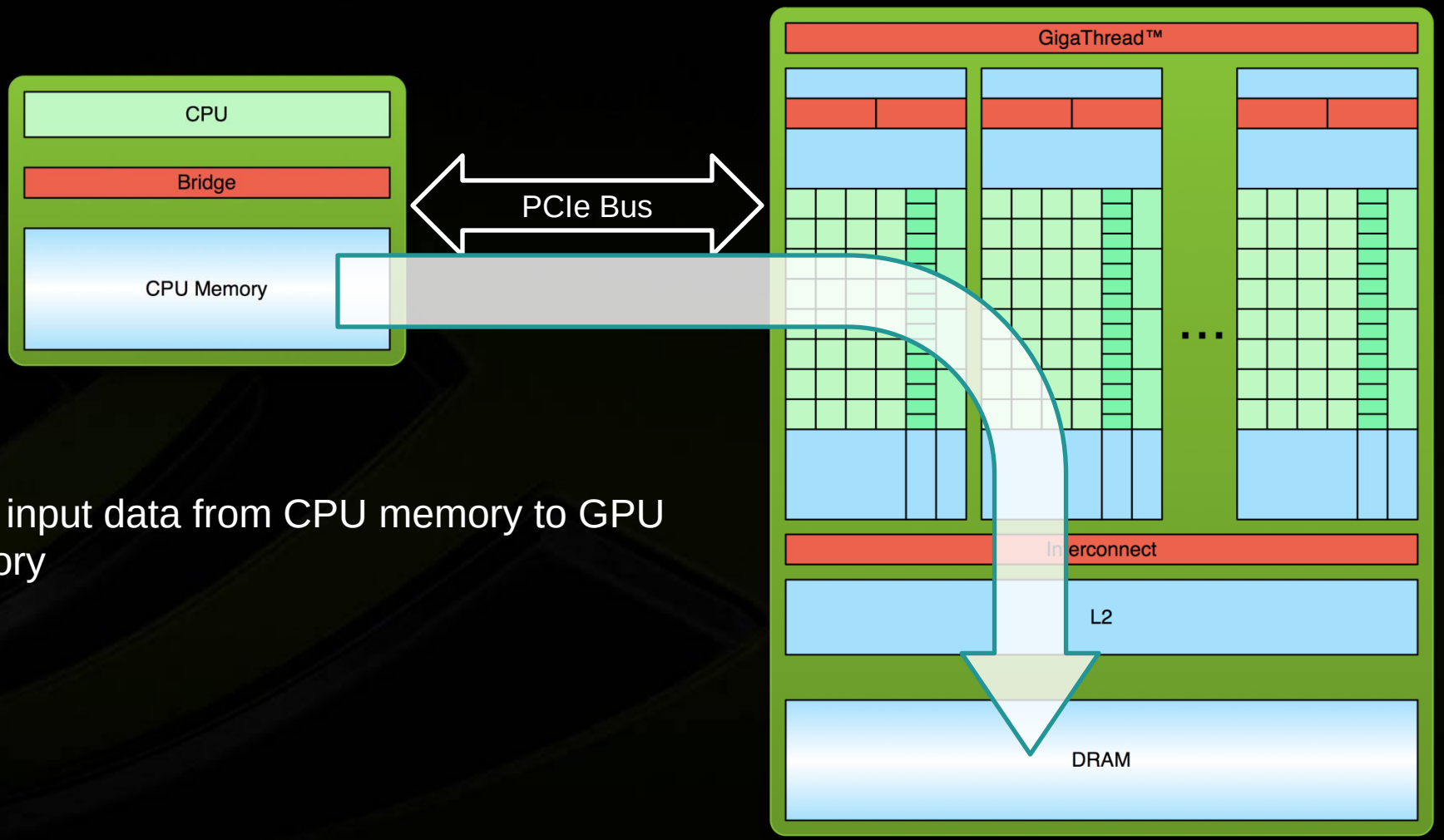
- Optimized for data-parallel, throughput computation
- Architecture tolerant of memory latency
- More transistors dedicated to computation

Low Latency or High Throughput?

- CPU architecture must minimize latency within each thread
- GPU architecture hides latency with computation from other thread warps

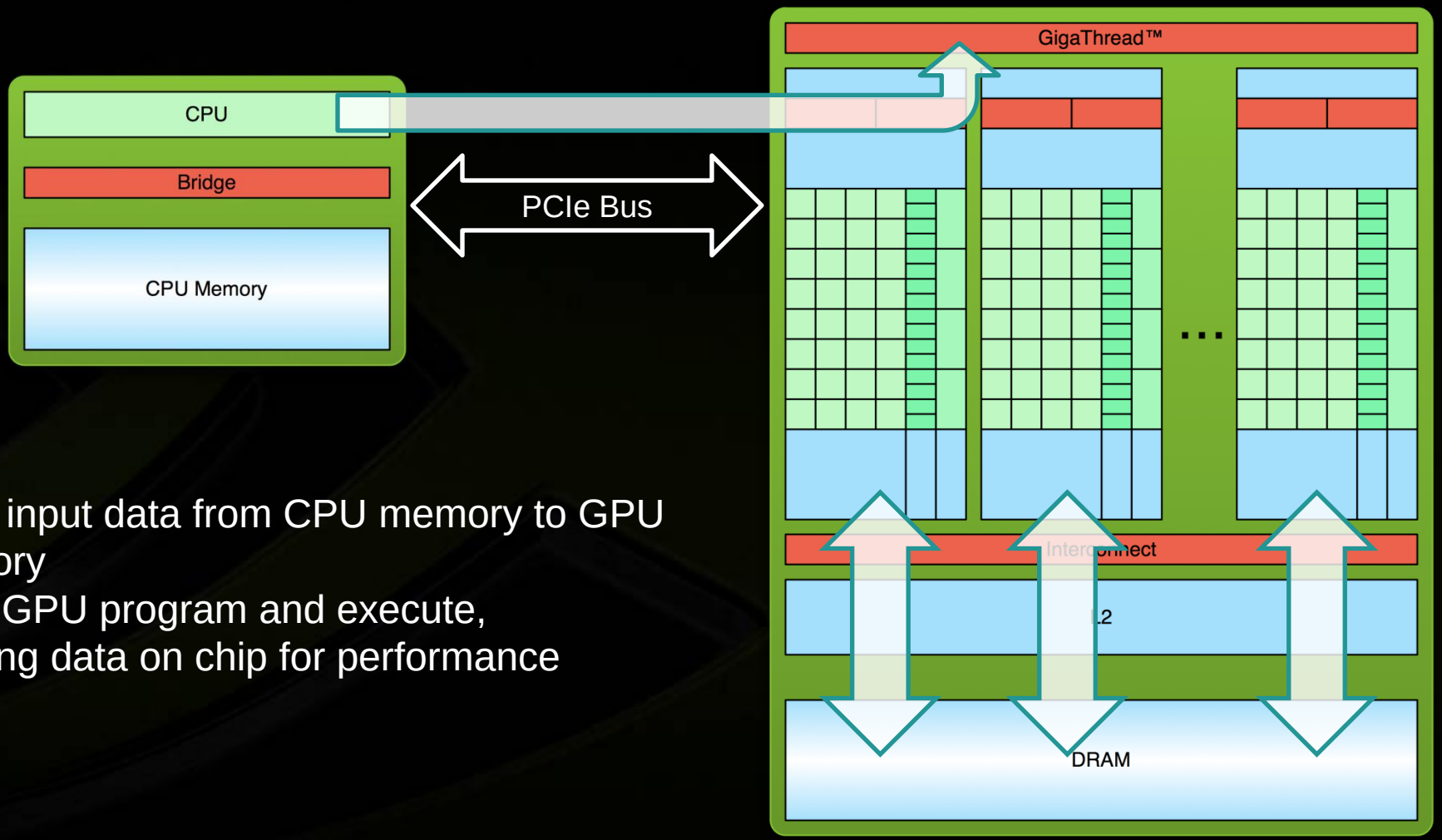


Processing Flow



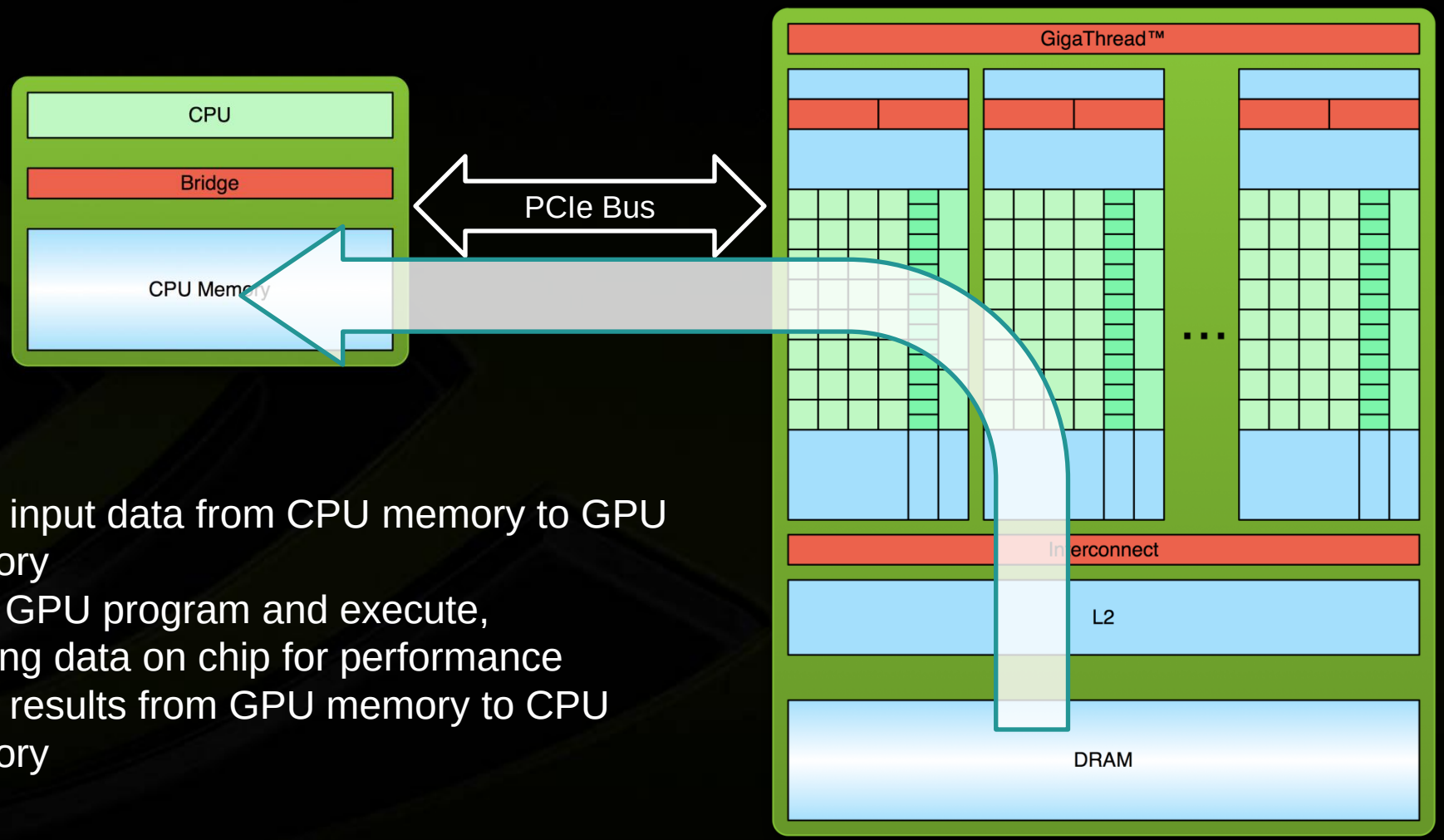
1. Copy input data from CPU memory to GPU memory

Processing Flow



1. Copy input data from CPU memory to GPU memory
2. Load GPU program and execute, caching data on chip for performance

Processing Flow

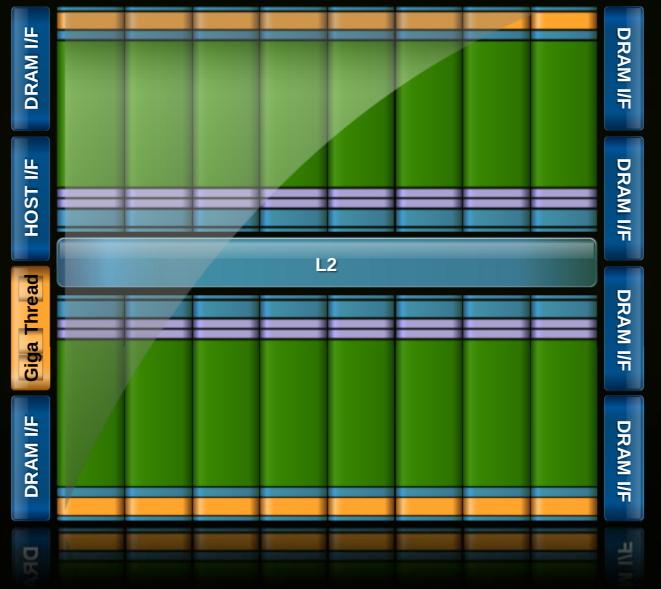


1. Copy input data from CPU memory to GPU memory
2. Load GPU program and execute, caching data on chip for performance
3. Copy results from GPU memory to CPU memory

GPU Architecture

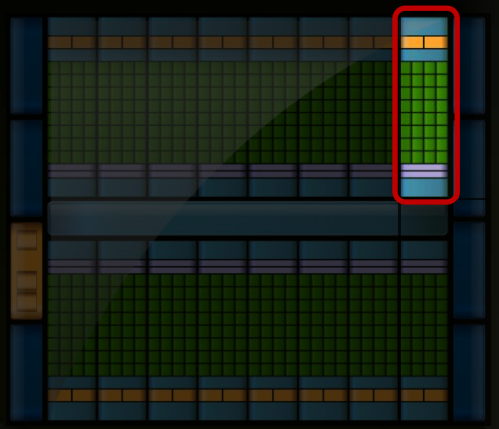
GPU Architecture: Two Main Components

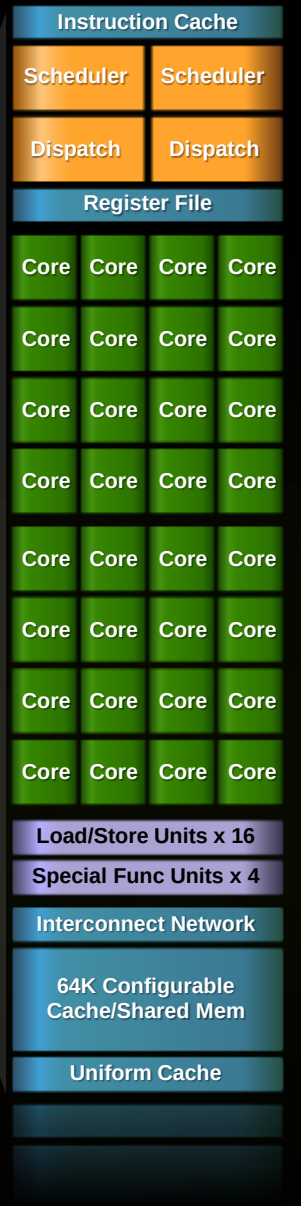
- **Global memory**
  - Analogous to RAM in a CPU server
  - Accessible by both GPU and CPU
  - Currently up to 6 GB
  - Bandwidth currently up to 150 GB/s for Quadro and Tesla products
  - ECC on/off option for Quadro and Tesla products
- **Streaming Multiprocessors (SMs)**
  - Perform the actual computations
  - Each SM has its own:
    - Control units, registers, execution pipelines, caches



GPU Architecture – Fermi: Streaming Multiprocessor (SM)

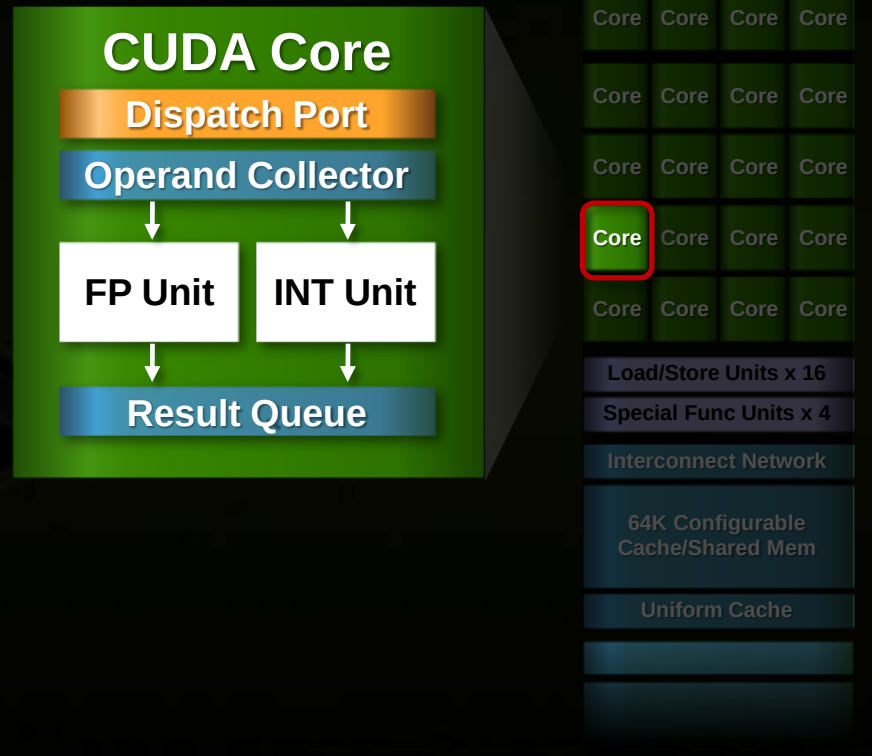
- 32 CUDA Cores per SM
  - 32 fp32 ops/clock
  - 16 fp64 ops/clock
  - 32 int32 ops/clock
- 2 warp schedulers
  - Up to 1536 threads concurrently
- 4 special-function units
- 64KB shared mem + L1 cache
- 32K 32-bit registers





GPU Architecture – Fermi: CUDA Core

- Floating point & Integer unit
  - IEEE 754-2008 floating-point standard
  - Fused multiply-add (FMA) instruction for both single and double precision
- Logic unit
- Move, compare unit
- Branch unit



GPU Architecture – Fermi: Memory System

- **L1**
  - 16 or 48KB / SM, can be chosen by the program
  - Hardware-managed
  - Aggregate bandwidth per GPU: 1.03 TB/s
- **Shared memory**
  - User-managed scratch-pad
    - Hardware will not evict until threads overwrite
  - 16 or 48KB / SM (64KB total is split between Shared and L1)
  - Aggregate bandwidth per GPU: 1.03 TB/s

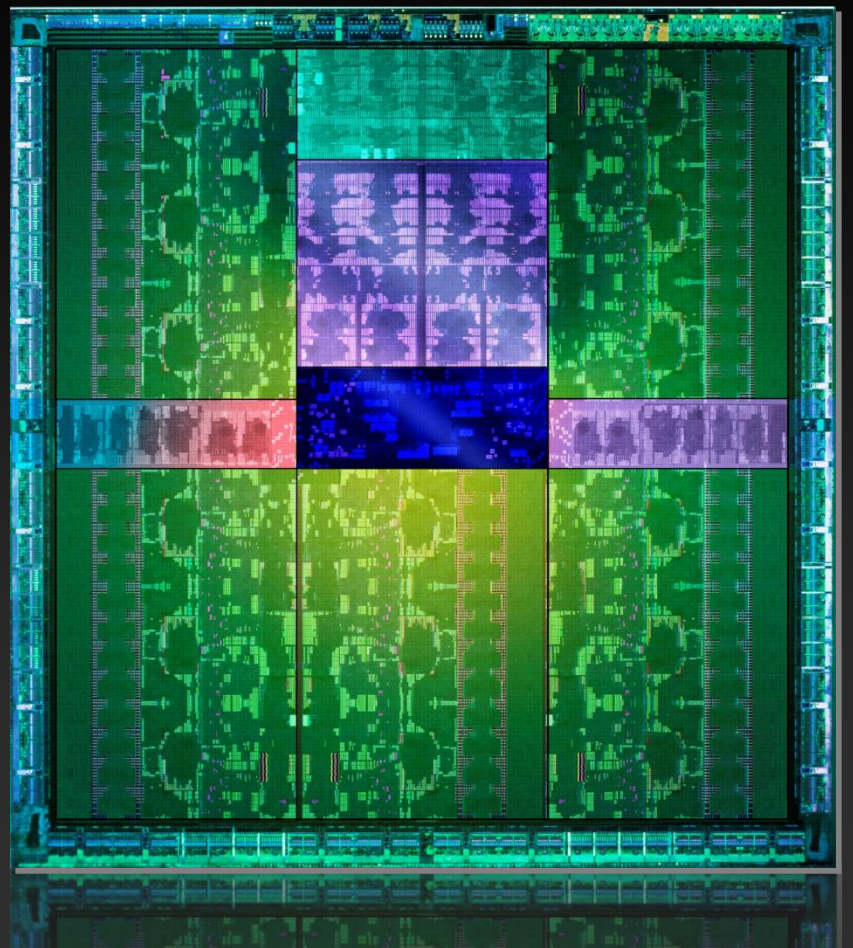
GPU Architecture – Fermi: Memory System

- **Unified L2 cache (768k)**
  - **Fast, coherent data sharing across all cores in the GPU**
- **ECC protection**
  - **DRAM**
    - ECC supported for GDDR5 memory
  - **All major internal memories are ECC protected**
    - Register file, L1 cache, L2 cache

Kepler

Kepler

Fastest, Most Efficient HPC Architecture Ever



SMX

Hyper-Q

Dynamic Parallelism

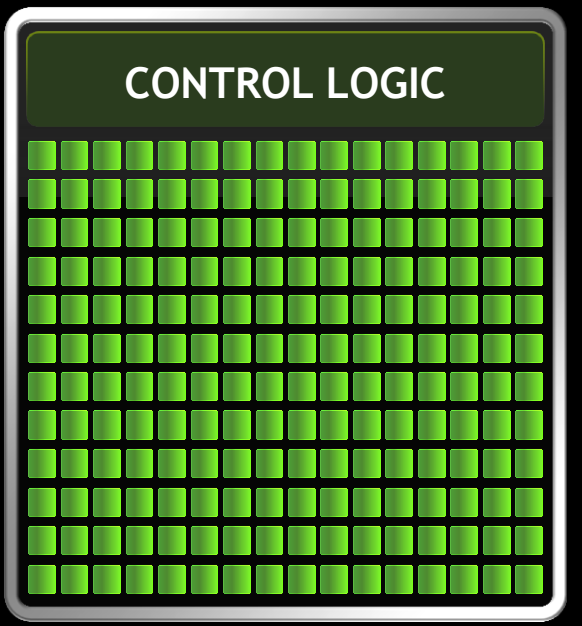
Kepler: Fast & Efficient

SM
M2090

32 cores

3x
Perf / Watt

SMX
K20

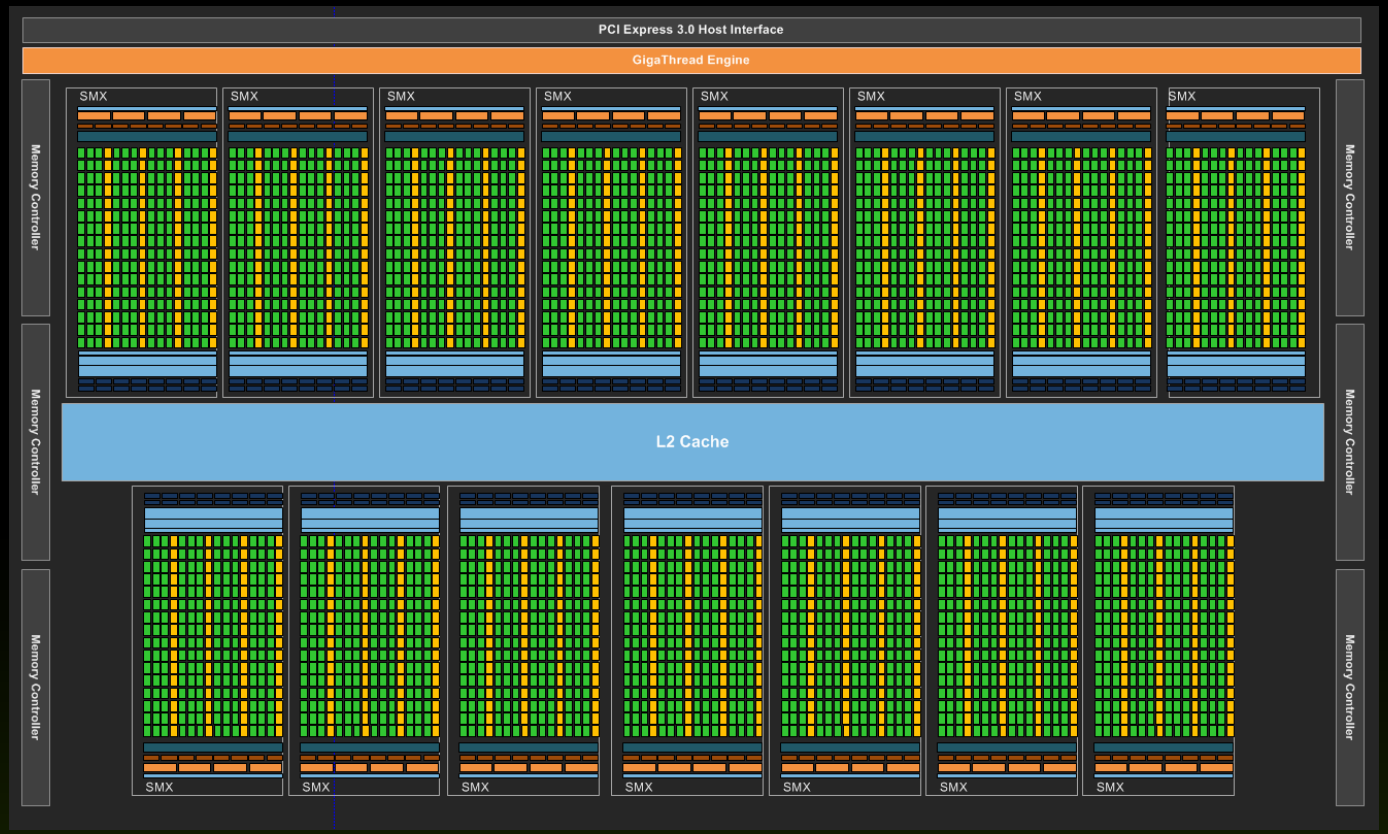


192 cores

Kepler GK110 Block Diagram

Architecture

- 7.1B Transistors
- 15 SMX units
- > 1 TFLOP FP64
- 1.5 MB L2 Cache
- 384-bit GDDR5

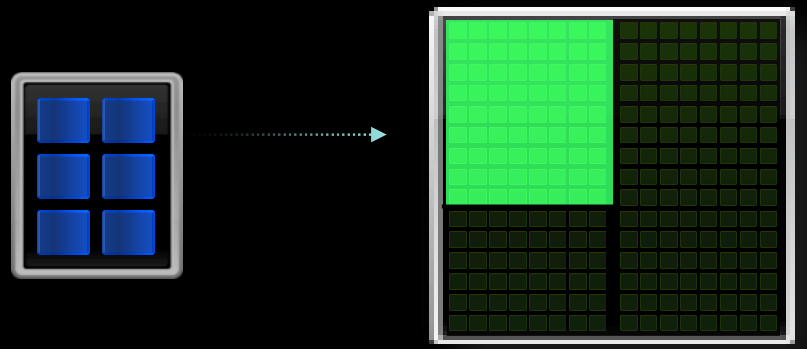


Hyper-Q

CPU Cores Simultaneously Run Tasks on Kepler

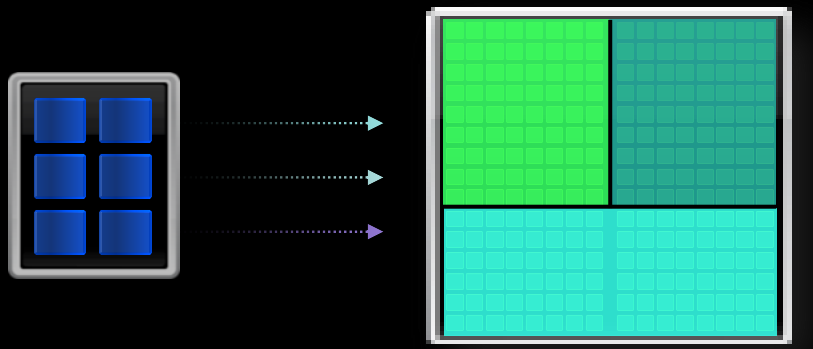
FERMI

1 MPI Task at a Time



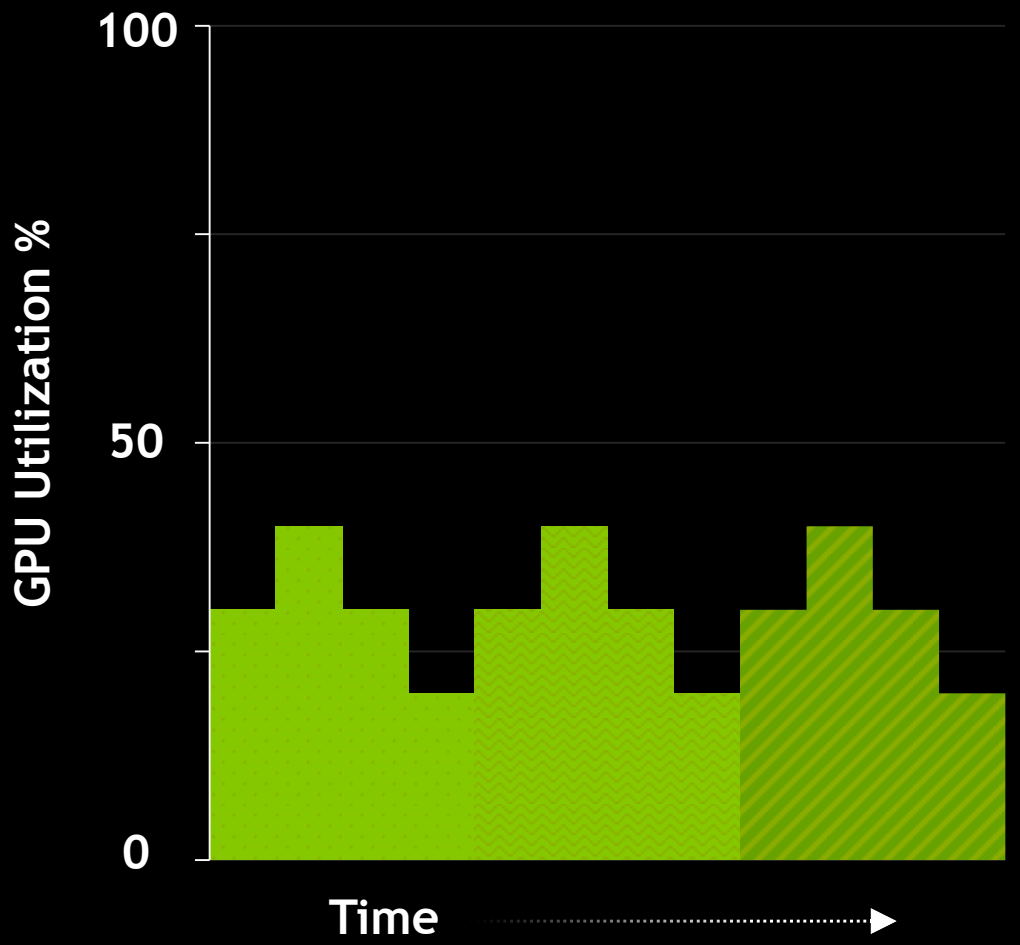
KEPLER

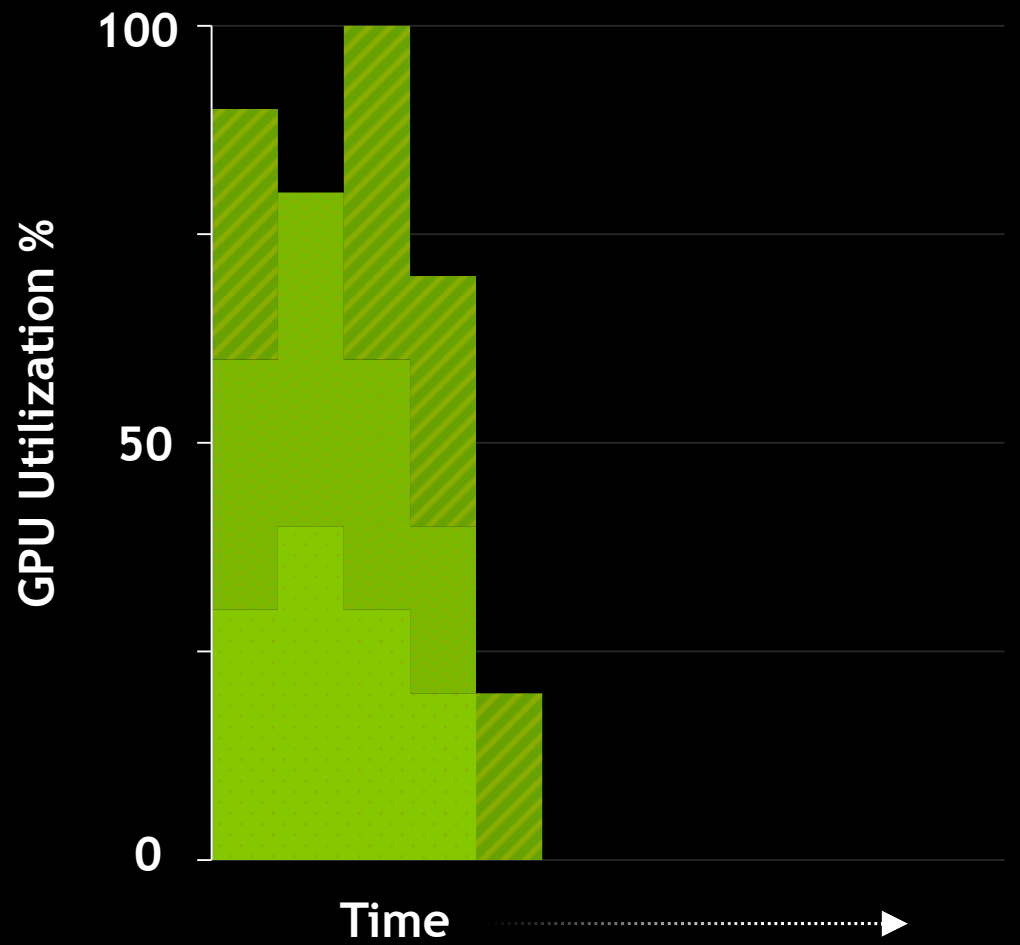
32 Simultaneous MPI Tasks



Hyper-Q

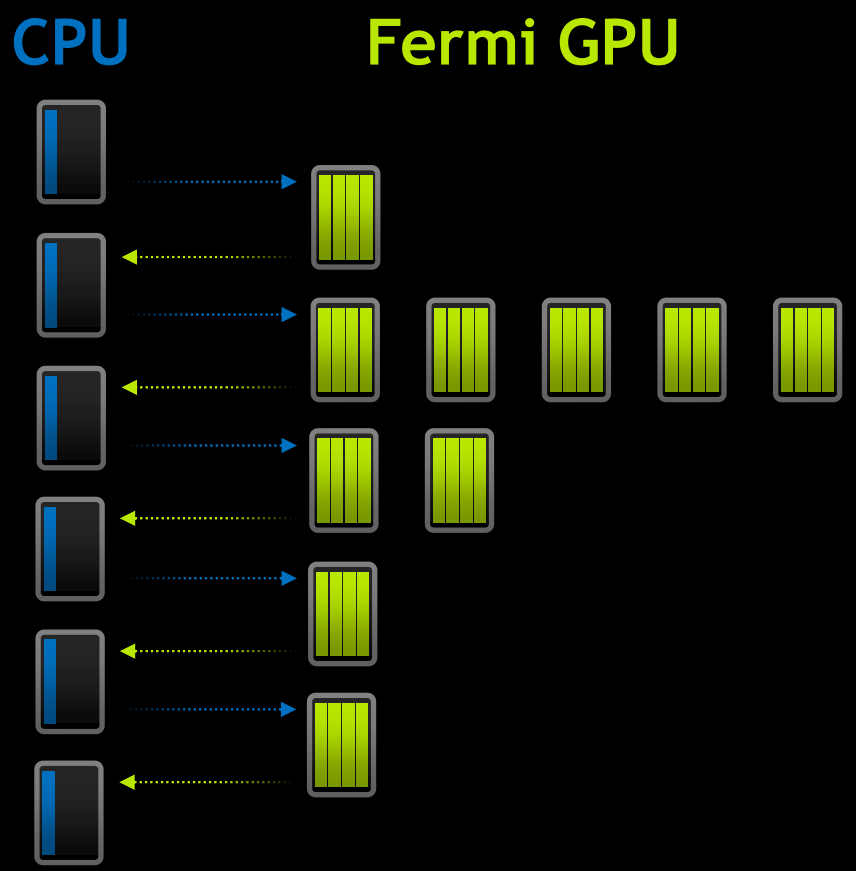
Max GPU Utilization, Slashes CPU Idle Time

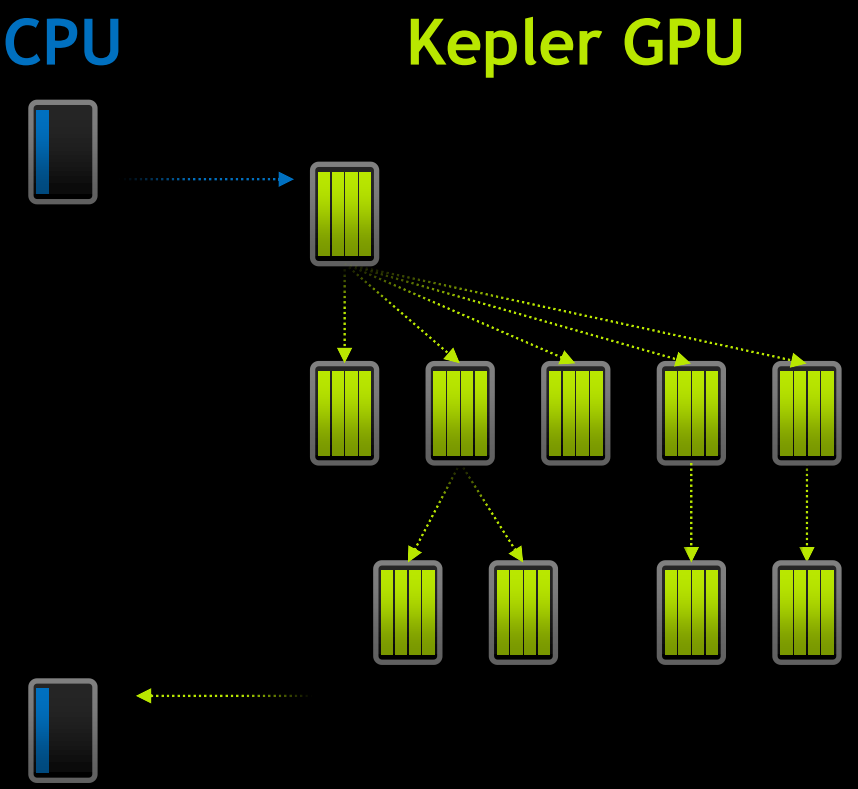




Dynamic Parallelism

GPU Adapts to Data, Dynamically Launches New Threads

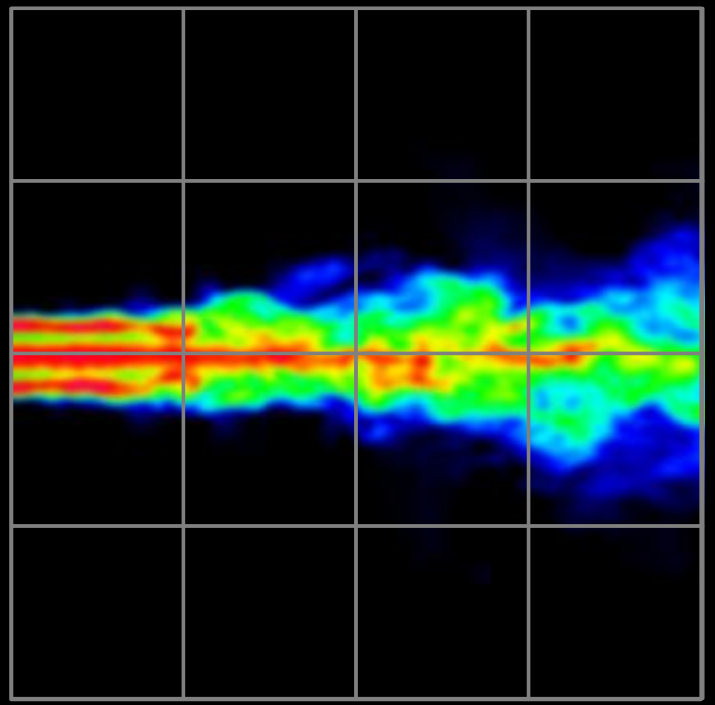




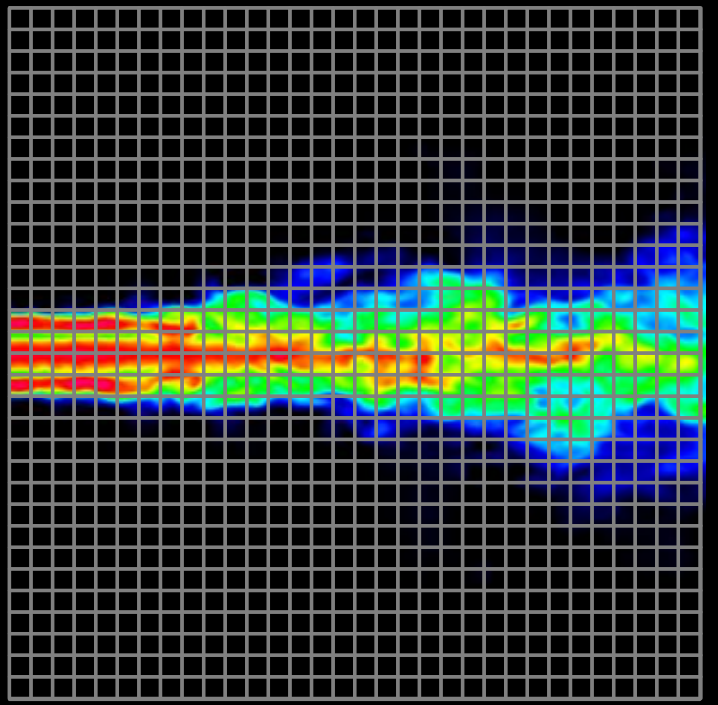
Dynamic Parallelism

Makes GPU Computing Easier & Broadens Reach

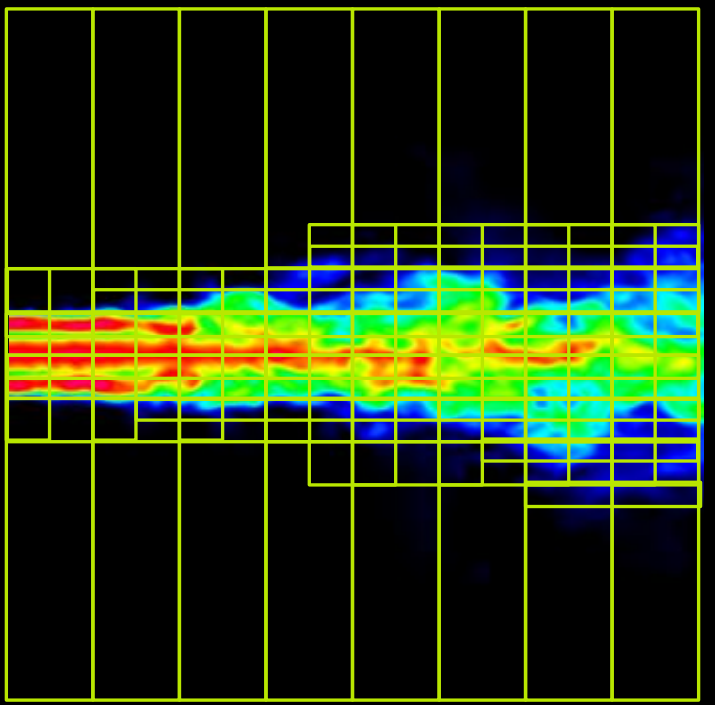
Too coarse



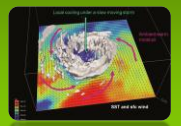
Too fine



Just right



Supercomputing



Weather / Climate Modeling
Molecular Dynamics
Computational Physics

Life Sciences

Biochemistry
Bioinformatics
Material Science

Manufacturing

Structural Mechanics
Comp Fluid Dynamics (CFD)
Electromagnetics

Defense / Govt

Signal Processing
Image Processing
Video Analytics

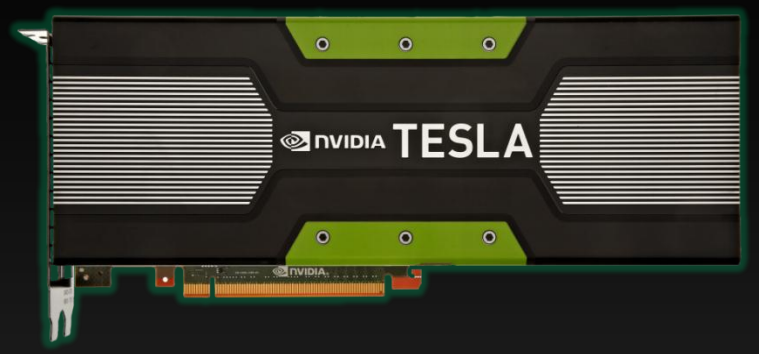
Oil and Gas

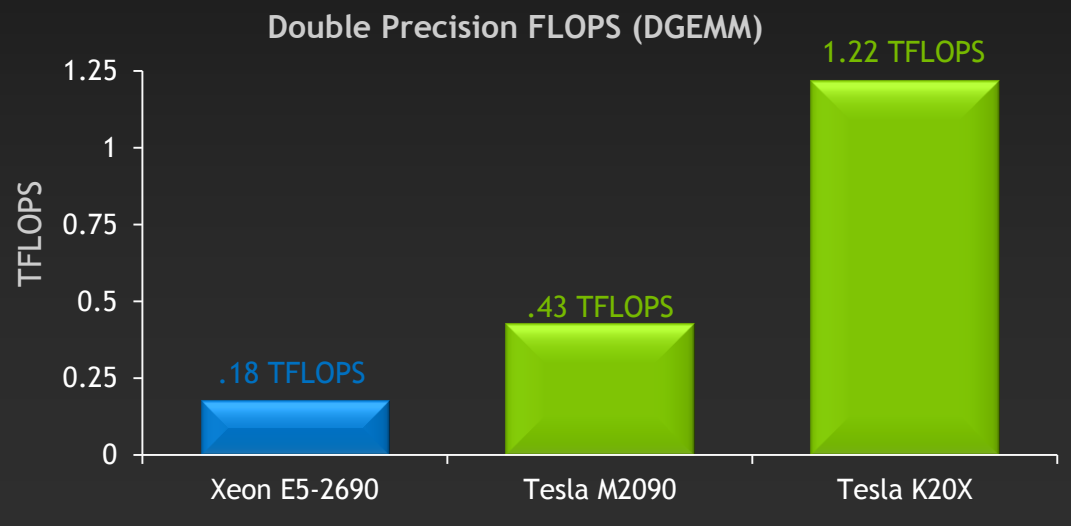
Reverse Time Migration
Kirchoff Time Migration

Q2	Q3	Q4
Tesla M2090	Fermi	Tesla K20
Tesla M2075		Kepler GK110
Tesla K10	Kepler GK104	

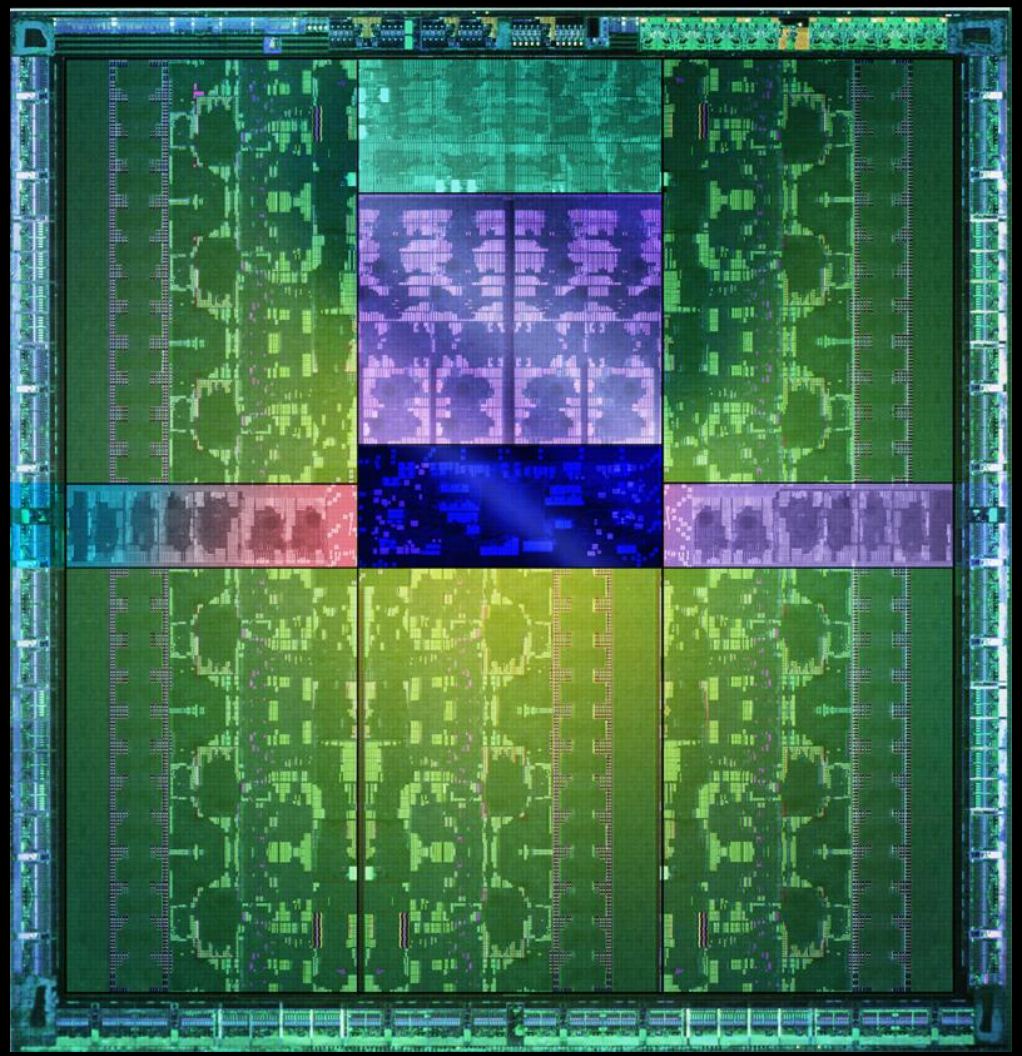
Tesla K20 Family: 3x Faster Than Fermi

Tesla K20X



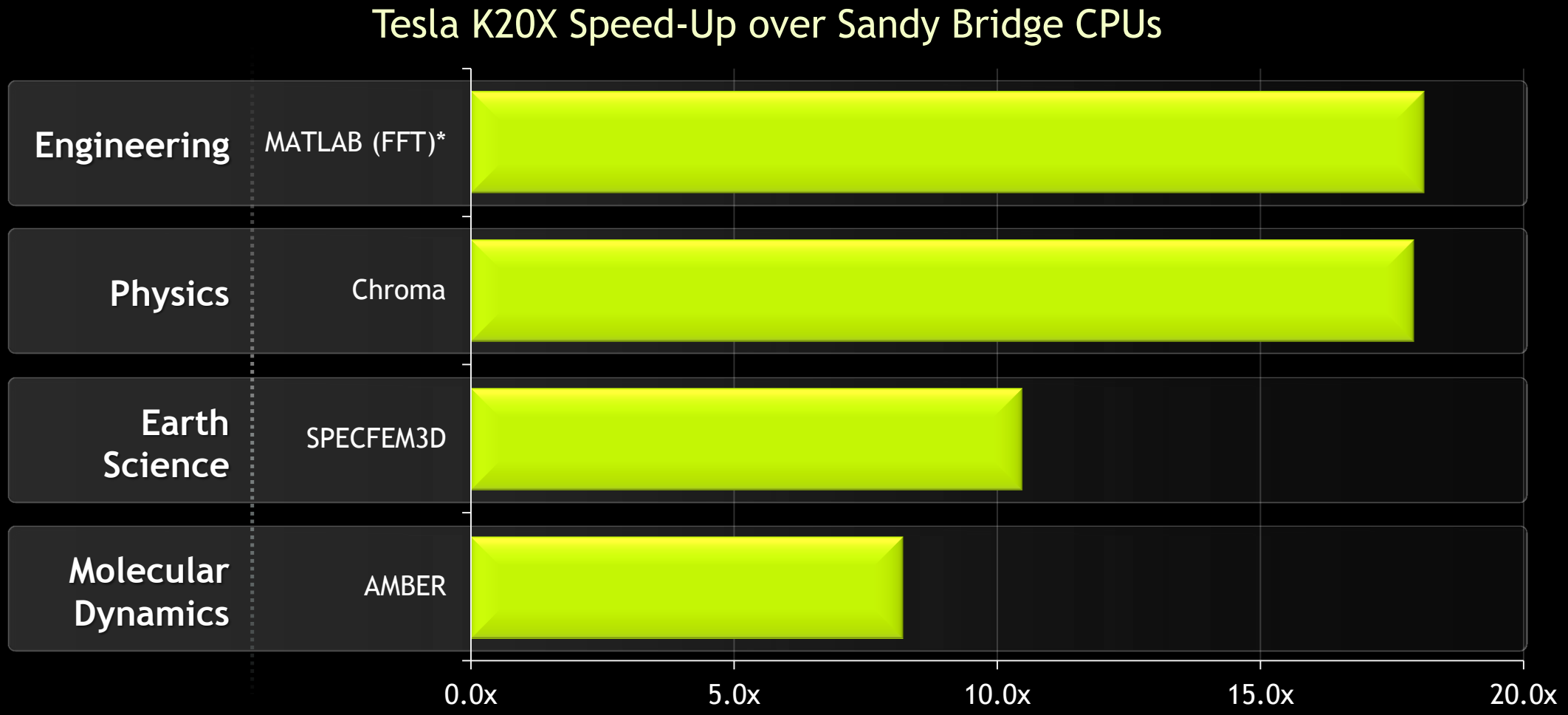


	Tesla K20X	Tesla K20
# CUDA Cores	2688	2496
Peak Double Precision	1.32 TF	1.17 TF
Peak DGEMM	1.22 TF	1.10 TF
Peak Single Precision	3.95 TF	3.52 TF
Peak SGEMM	2.90 TF	2.61 TF
Memory Bandwidth	250 GB/s	208 GB/s
Memory size	6 GB	5 GB
Total Board Power	235W	225W



Whitepaper: http://www.nvidia.com/object/nvidia-kepler.html

Fastest Performance on Scientific Applications



CPU results: Dual socket E5-2687w, 3.10 GHz, GPU results: Dual socket E5-2687w + 2 Tesla K20X GPUs
*MATLAB results comparing one i7-2600K CPU vs with Tesla K20 GPU
Disclaimer: Non-NVIDIA implementations may not have been fully optimized

Applications Scale to 1000s of GPUs