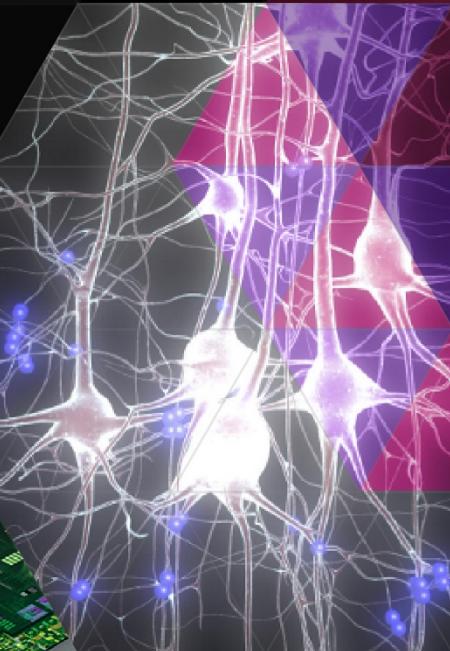
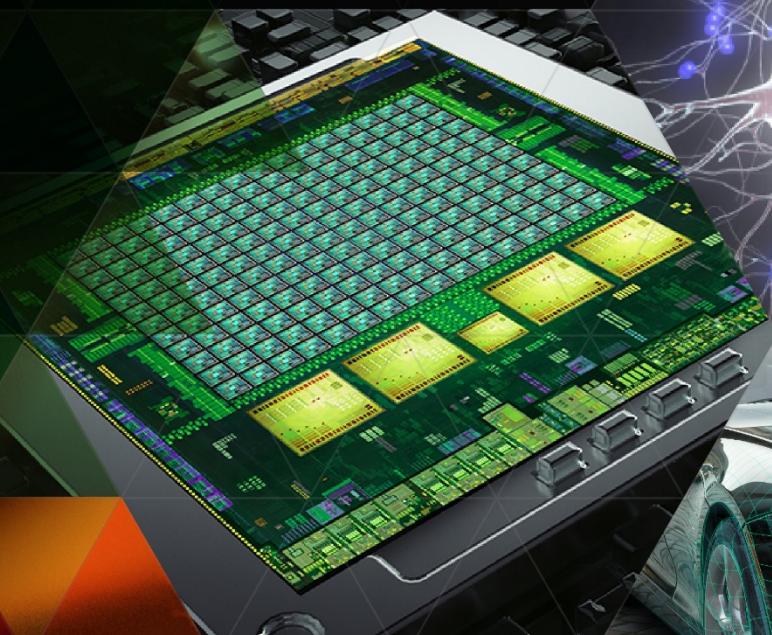




# GPU COMPUTING AND THE FUTURE OF HPC

Timothy Lanfear, NVIDIA

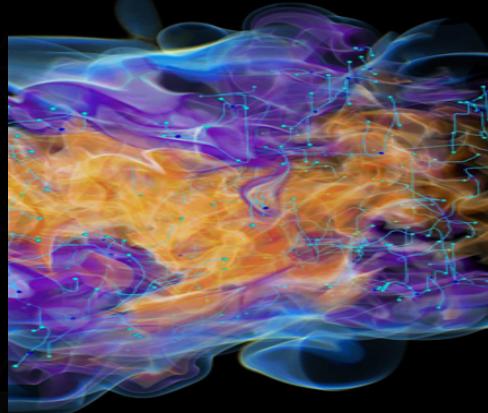




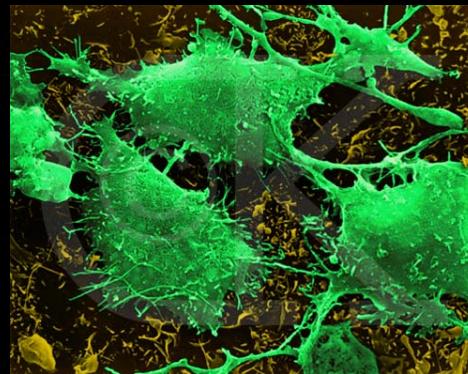
# EXASCALE COMPUTING WILL ENABLE TRANSFORMATIONAL SCIENCE RESULTS



Comprehensive Earth System Model at 1km scale, enabling modeling of cloud convection and ocean eddies.

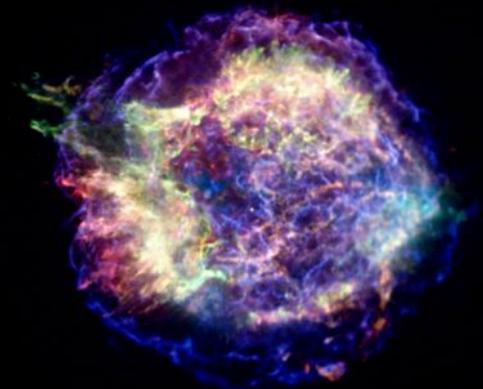


First-principles simulation of combustion for new high-efficiency, low-emission engines.

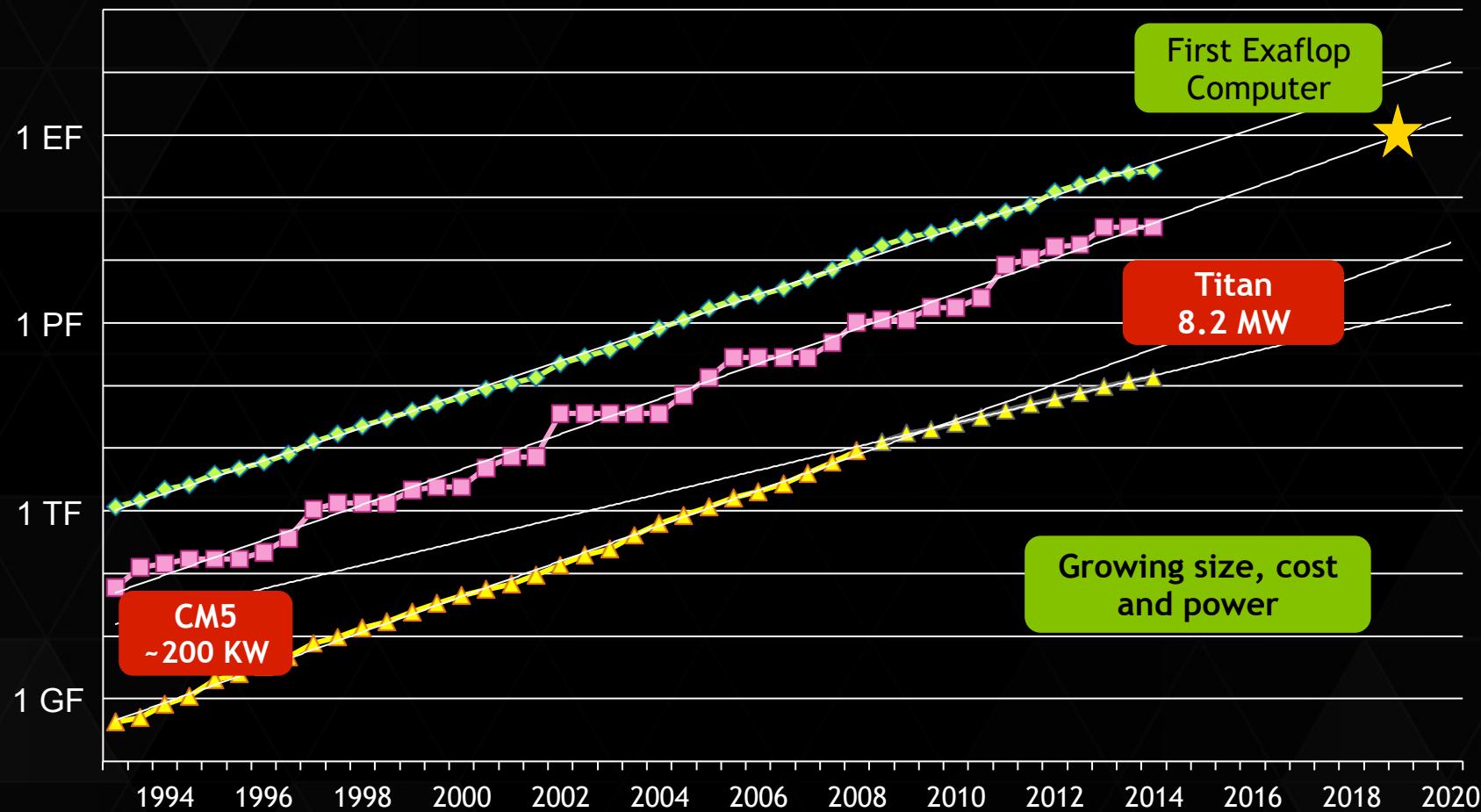


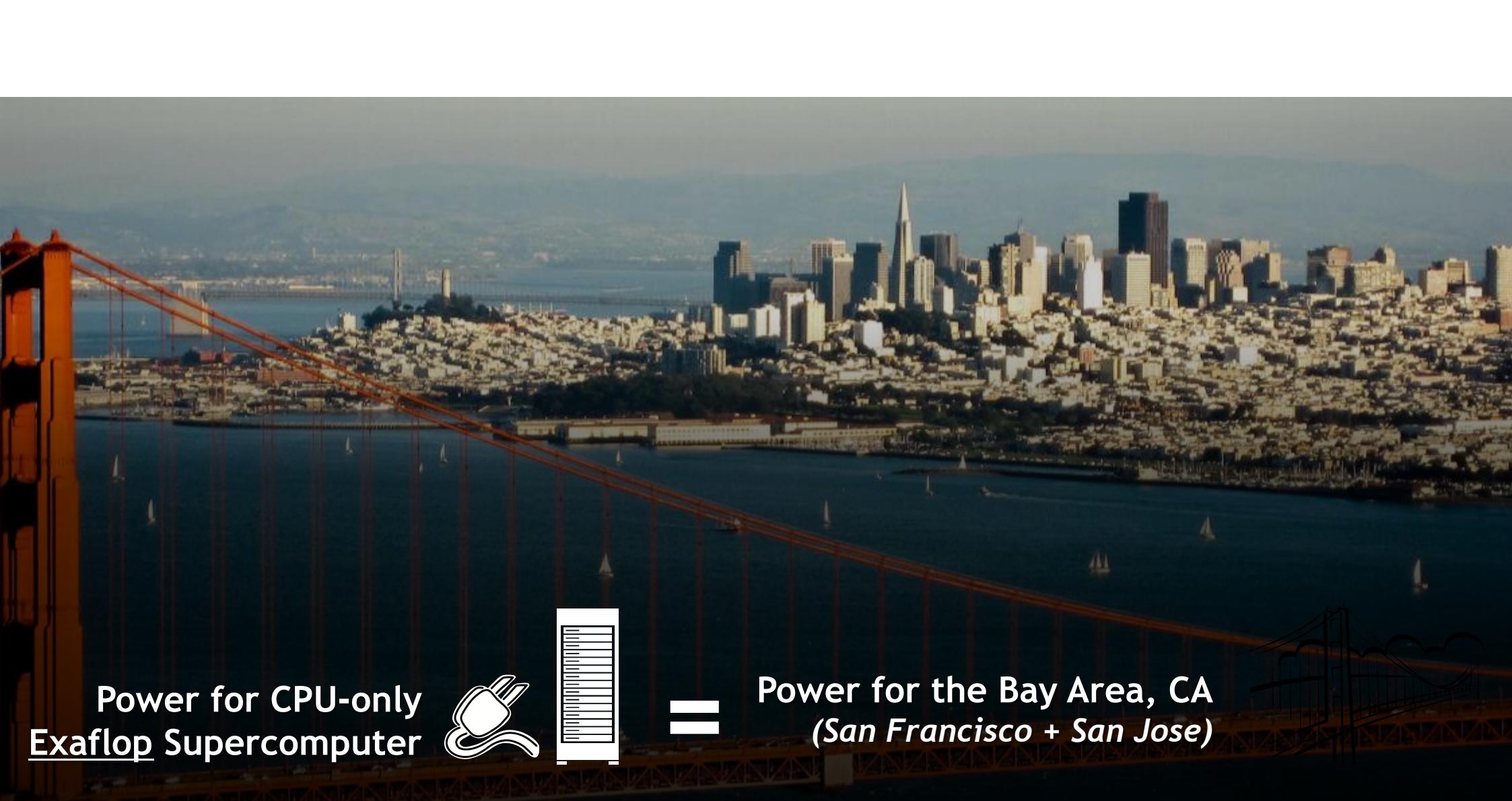
Coupled simulation of entire cells at molecular, genetic, chemical and biological levels.

Predictive calculations for thermonuclear and core-collapse supernovae, allowing confirmation of theoretical models.



# EXAFLOP EXPECTATIONS





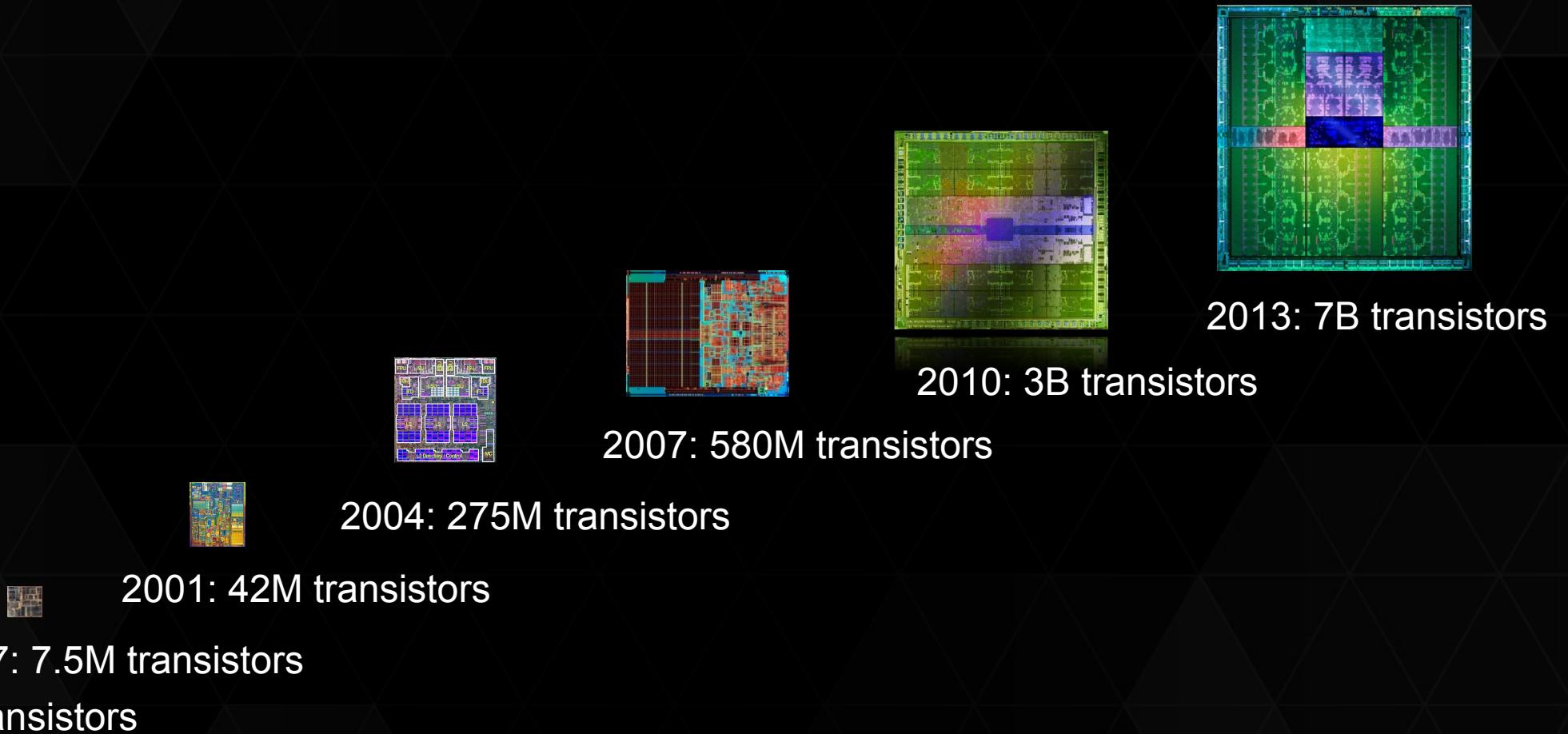
Power for CPU-only  
Exaflop Supercomputer



Power for the Bay Area, CA  
(*San Francisco + San Jose*)

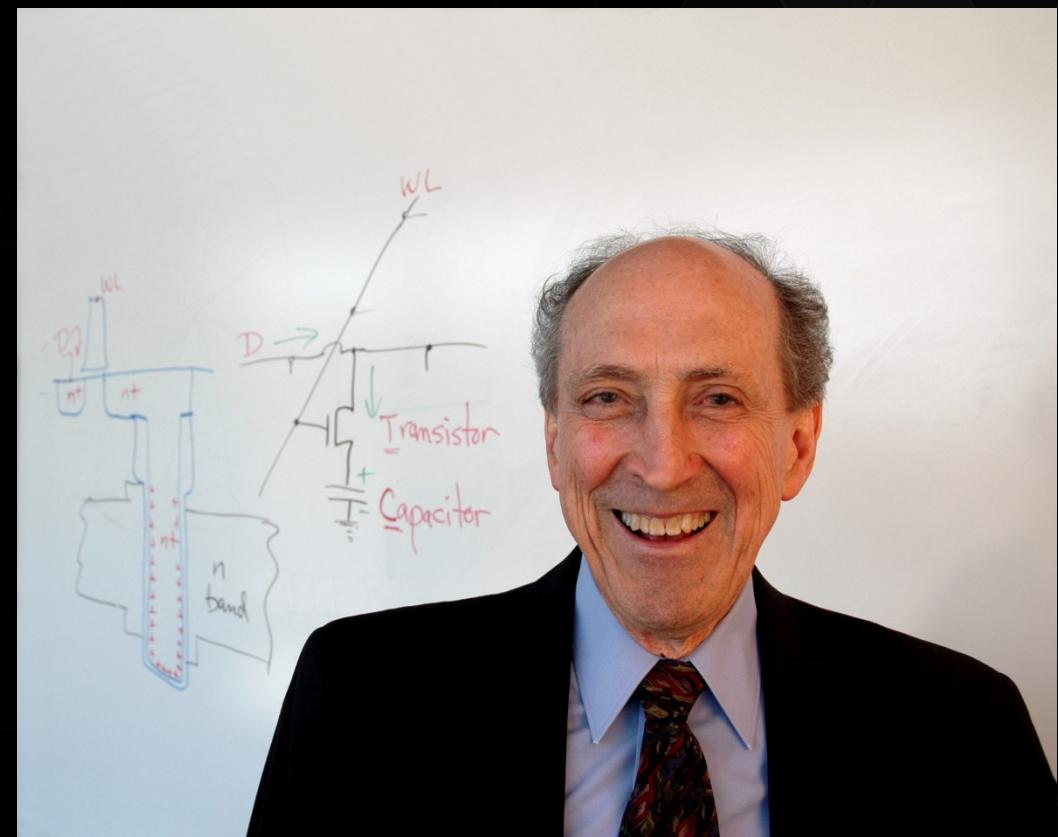
# HPC's Biggest Challenge: Power

# MOORE'S LAW IS ONLY PART OF THE STORY



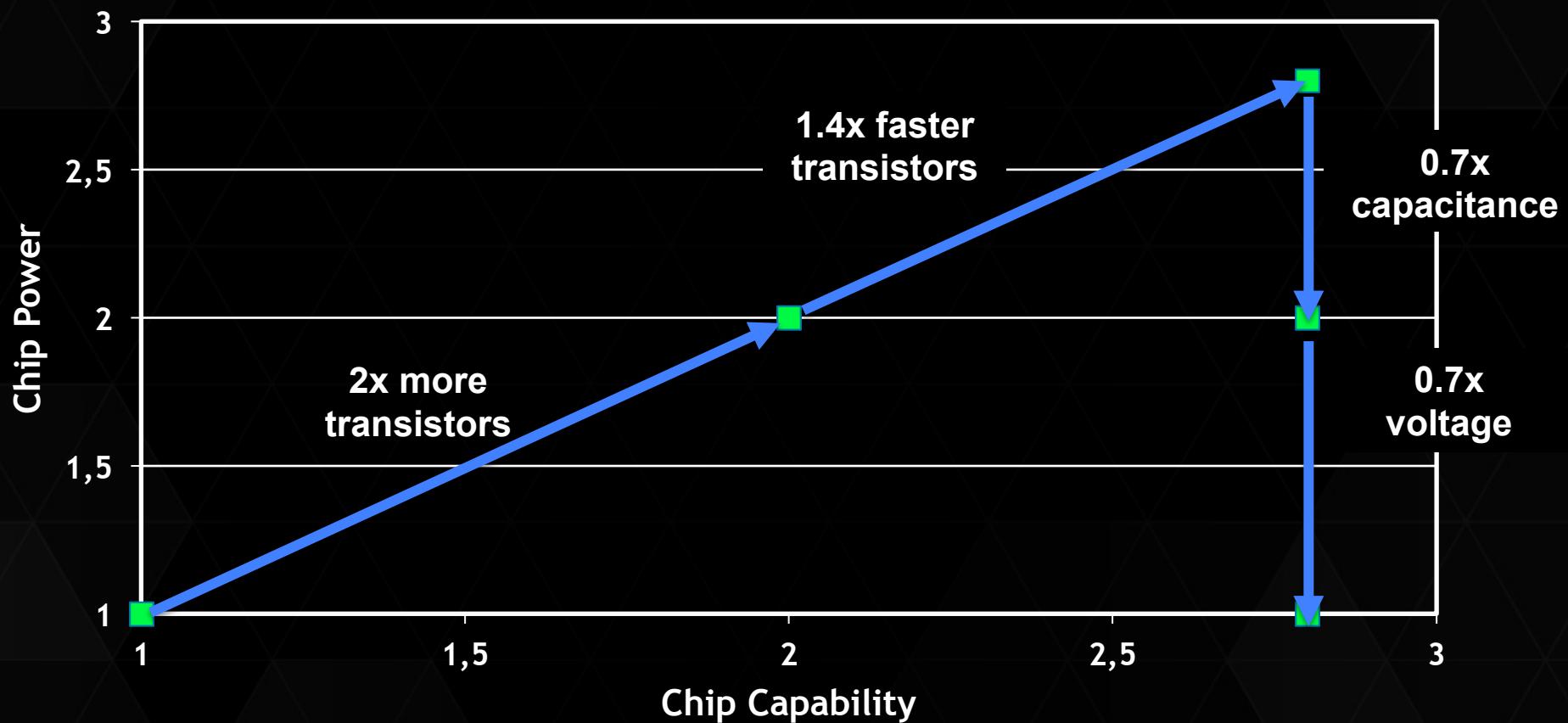
# ROBERT DENNARD, IBM

- 1968: invented DRAM
- 1974: postulated all key figures of merit of MOSFETs improve provided geometric dimensions, voltages, and doping concentrations are consistently scaled to maintain the same electric field.



# CLASSIC DENNARD SCALING

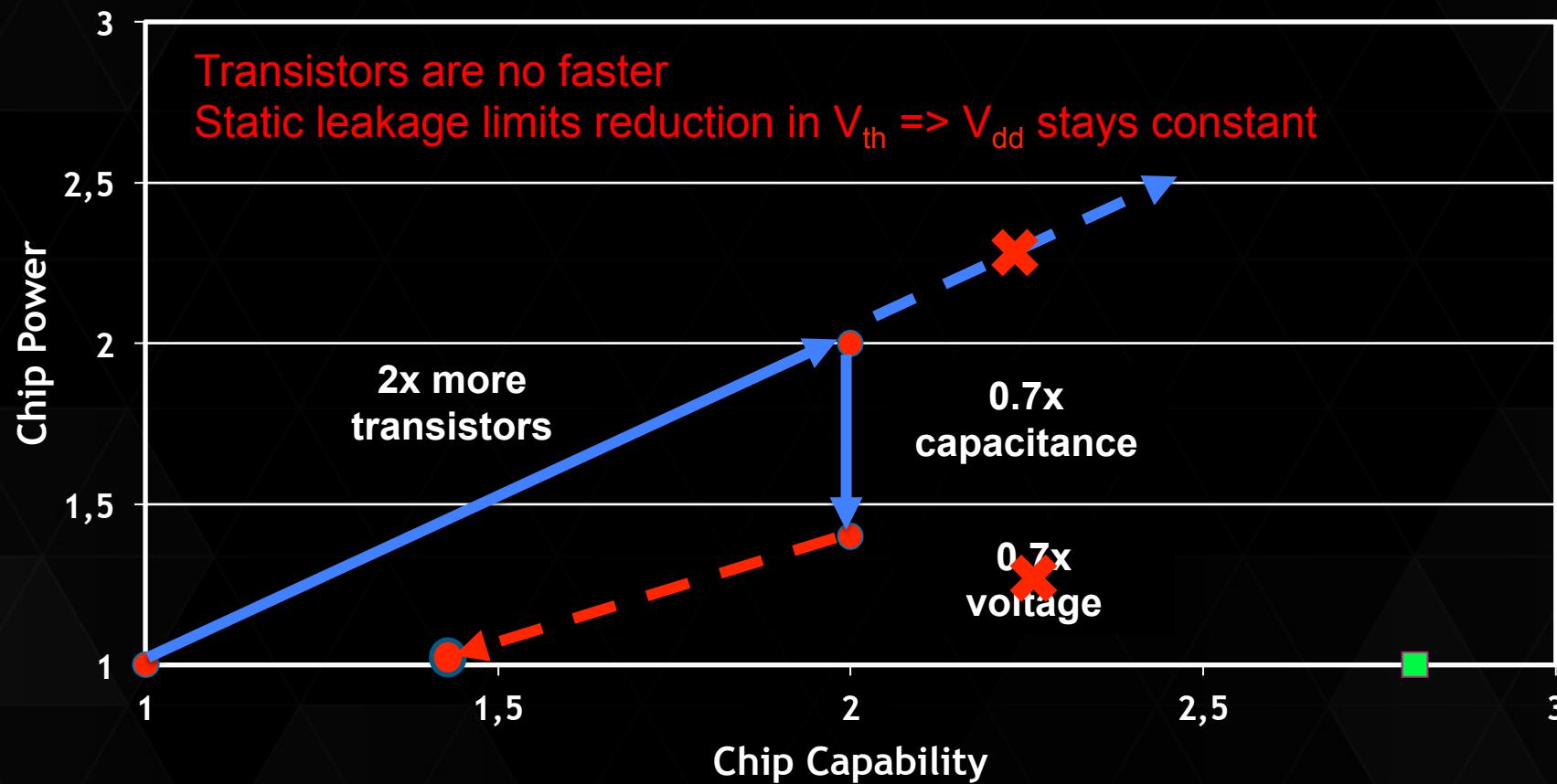
2.8x chip capability in same power



# POST DENNARD SCALING

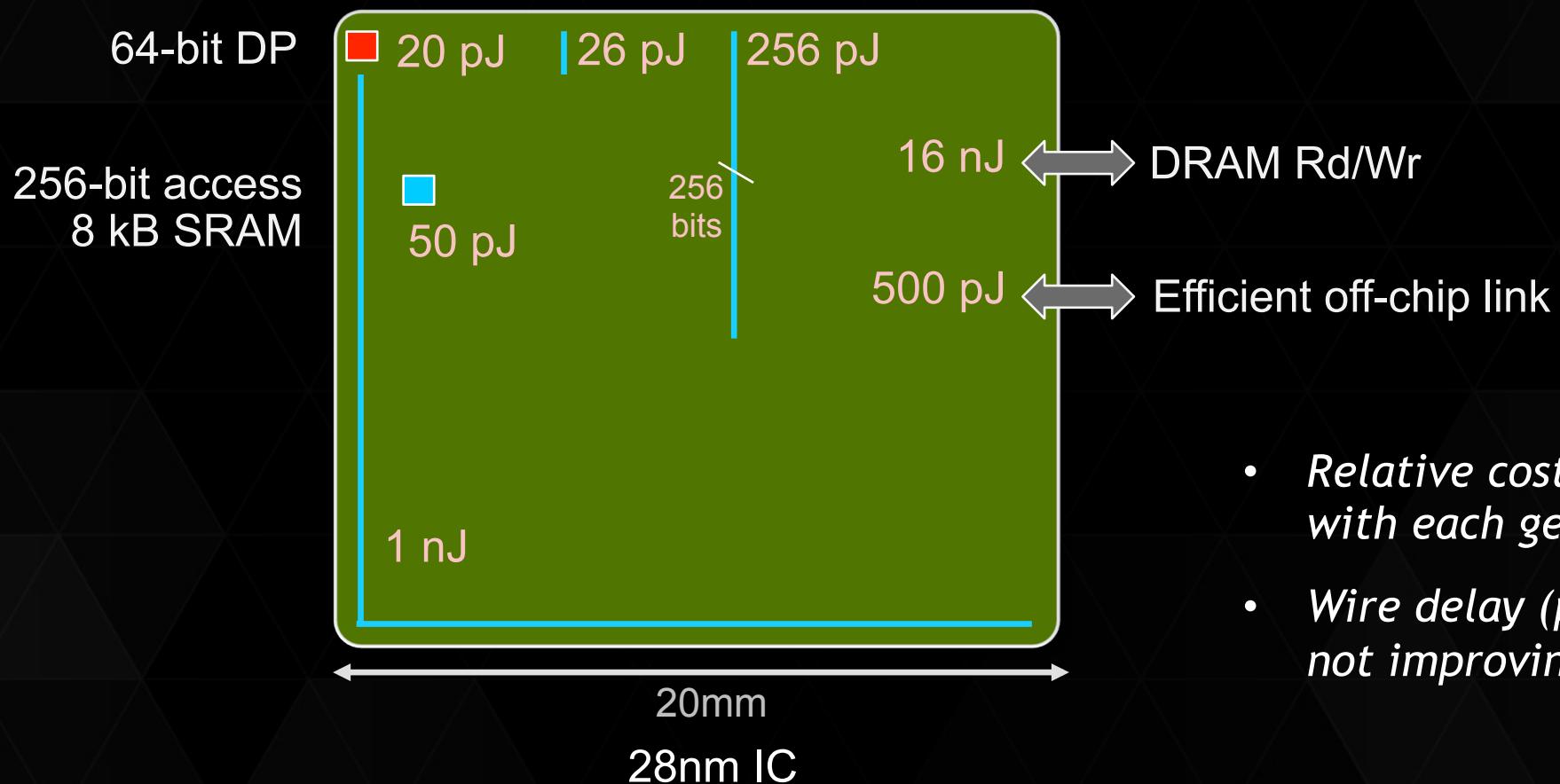
2x chip capability at 1.4x power

1.4x chip capability at same power



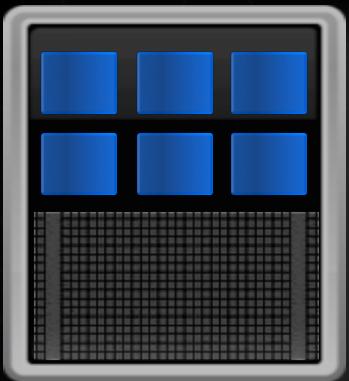
# THE HIGH COST OF DATA MOVEMENT

Fetching operands costs more than computing on them

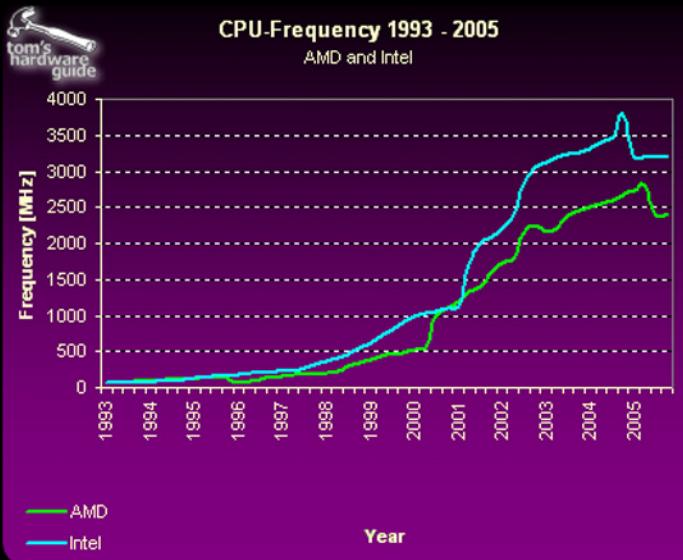


# SO, WHAT TO DO?

- 1) Stop making it worse...



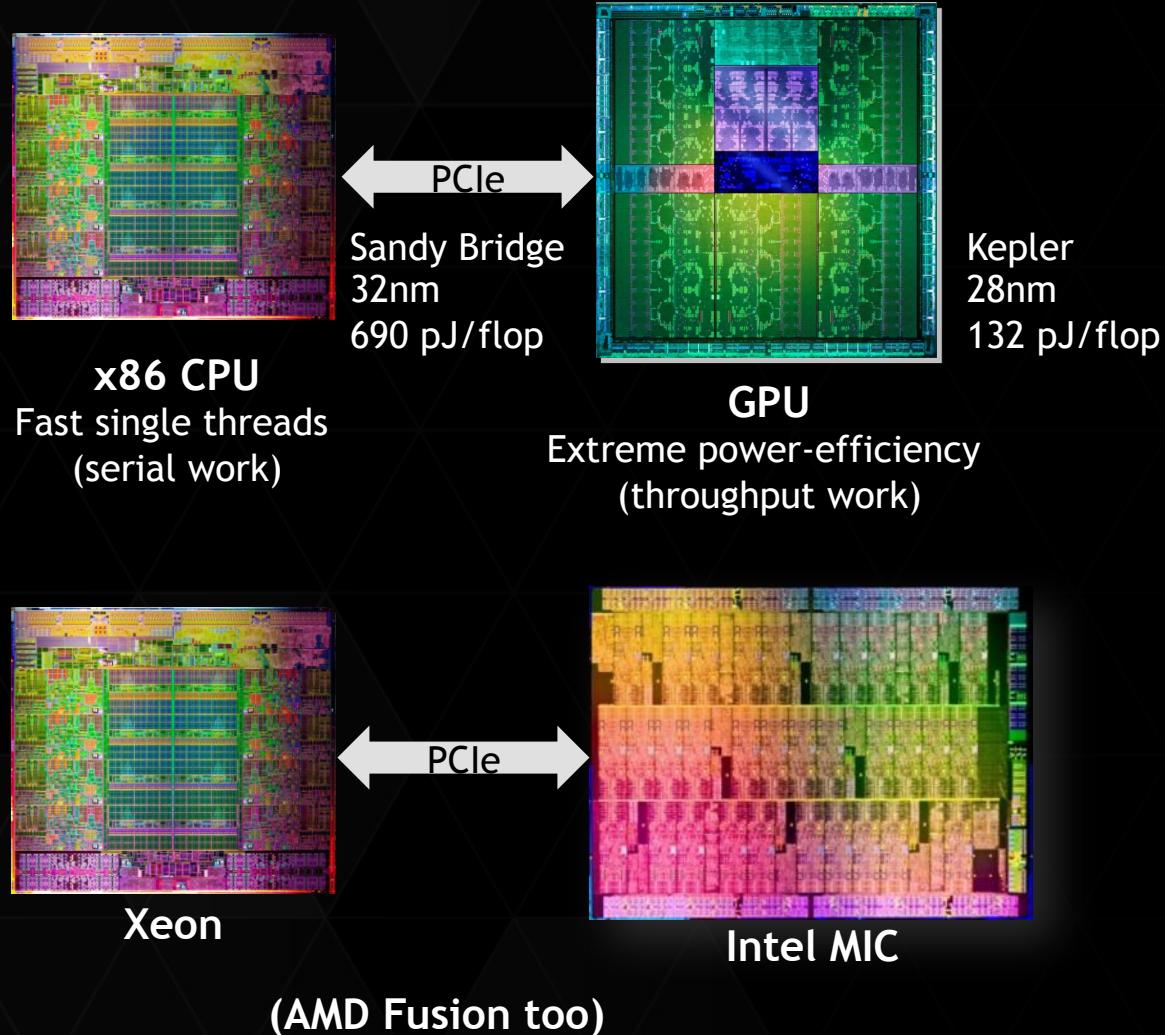
Multicore CPUs



*But still only a tiny fraction of CPU power spent on flops*

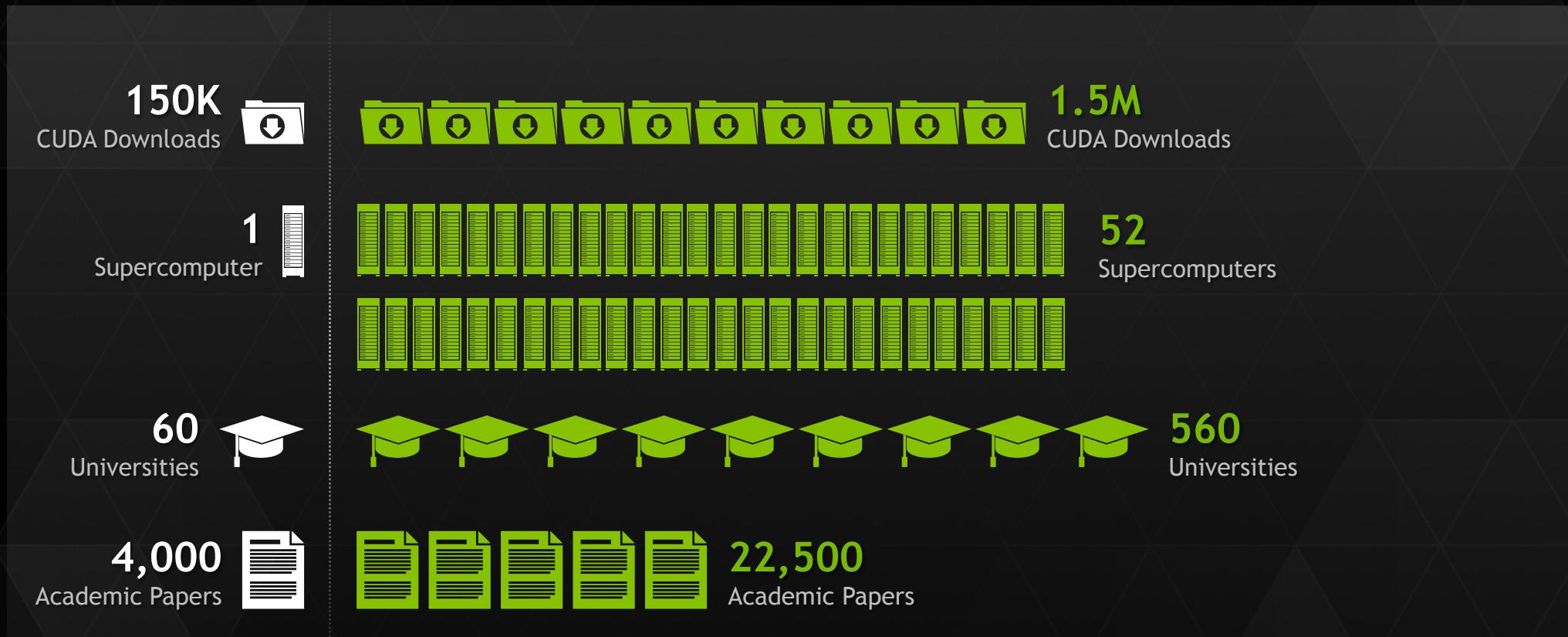
- 2) Unwind all that complexity we threw at single thread performance

# HPC IS GOING HYBRID



- Do most work by cores optimized for **extreme energy efficiency**
- Still need a few cores optimized for **fast serial work**

# EXPLOSIVE GROWTH OF GPU COMPUTING



2008

2012



## POPULAR GPU-ACCELERATED APPLICATIONS

CONTENTS	
02 Research: Higher Education and Supercomputing	COMPUTATIONAL CHEMISTRY AND BIOLOGY NUMERICAL ANALYTICS PHYSICS WEATHER AND CLIMATE FORECASTING
06 Defense and Intelligence	DEFENSE AND INTELLIGENCE
07 Computational Finance	COMPUTATIONAL FINANCE
08 Manufacturing: CAD and CAE	MANUFACTURING: CAD AND CAE
10 Media and Entertainment	MEDIA AND ENTERTAINMENT
14 Oil and Gas	OIL AND GAS

### Research: Higher Education and Supercomputing

#### COMPUTATIONAL CHEMISTRY AND BIOLOGY

##### Bioinformatics

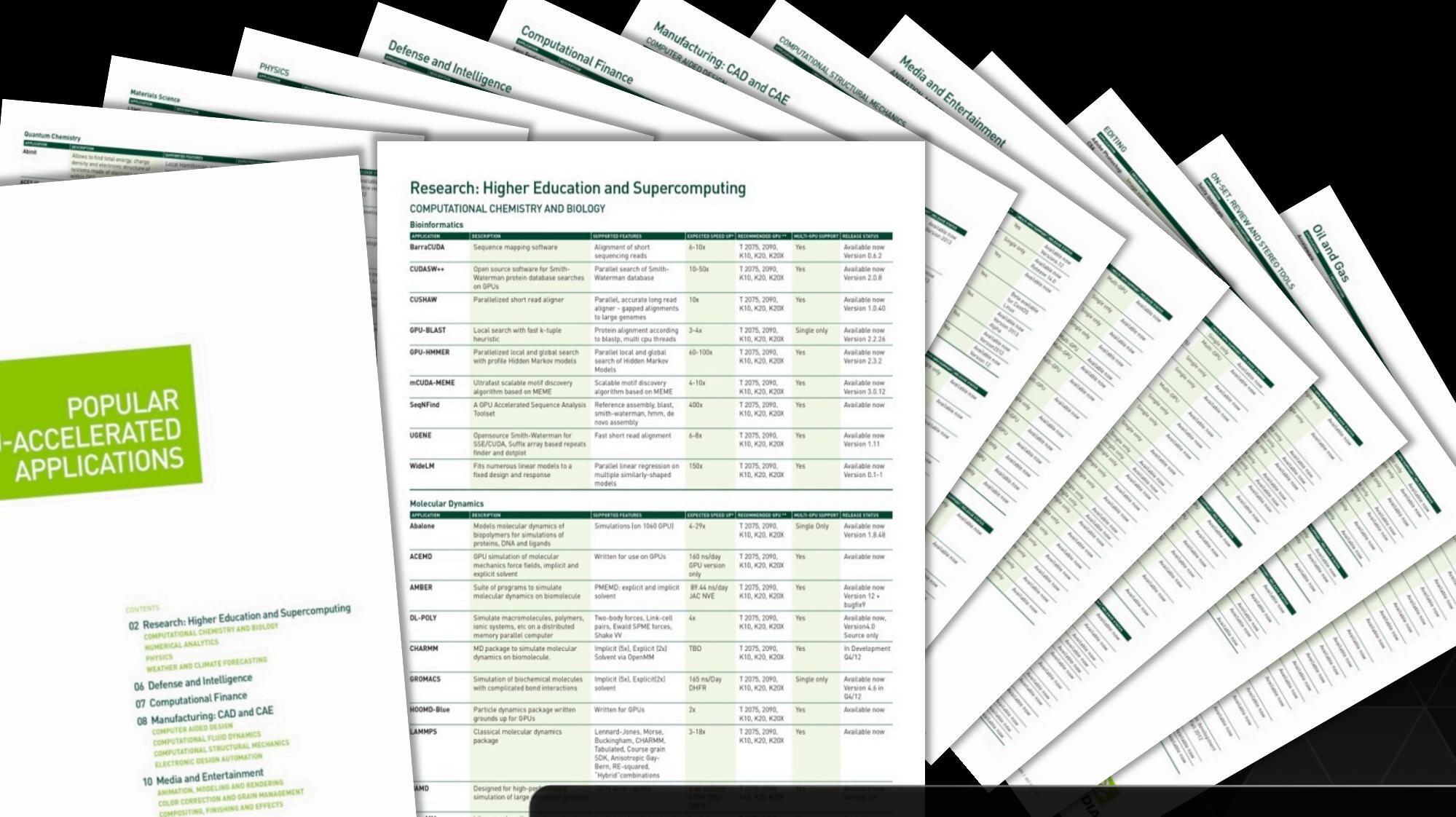
APPLICATION	DESCRIPTION	SUPPORTED FEATURES	EXPECTED SPEED UP*	RECOMMENDED GPU**	MULTI-GPU SUPPORT	RELEASE STATUS
Barracuda	Sequence mapping software	Alignment of short sequencing reads	6-10x	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 6.6.2
CUDASW++	Open source software for Smith-Waterman 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 - mapped 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
mcCUDA-MEME	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 Toolkit	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

APPLICATION	DESCRIPTION	SUPPORTED FEATURES	EXPECTED SPEED UP*	RECOMMENDED GPU**	MULTI-GPU SUPPORT	RELEASE STATUS
Abalone	Models molecular dynamics of biopolymers for simulations of proteins, DNA and ligands	Simulations (on 1040 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/W	4x	T 2075, 2090, K10, K20, K20X	Yes	Available now, Version 4.0 Source only
CHARMM	MD package to simulate molecular dynamics on biomolecule.	Implicit (2x), 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 (2x), Explicit(2x) solvent	165 ns/Day DHFR	T 2075, 2090, K10, K20, K20X	Single only	Available now Version 4.6 in Q4/12
HOMO-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, Coulomb grain, SDK, Anisotropic Gay-Berne, RE-squared, "Hybrid" combinations	3-18x	T 2075, 2090, K10, K20, K20X	Yes	Available now
MD	Designed for high-performance simulation of large molecular systems	MD, MD with periodic boundary conditions, SMD, Langevin, Brownian, Langevin-Brownian, Langevin-Brownian-SMD, Langevin-SMD, Langevin-Brownian-SMD, Langevin-SMD-Brownian, Langevin-Brownian-SMD-Brownian	4-4000x	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 2.4
OpenMM	Library and application for molecular dynamics for HPC and GPUs	Implicit (2x), Explicit solvent	1x	T 2075, 2090, K10, K20, K20X	Yes	Available now Version 4.6 in Q4/12

Hundreds of GPU-Accelerated Applications

[www.nvidia.com/appscatalog](http://www.nvidia.com/appscatalog)



# BREAKTHROUGH EFFICIENCY

## The Green500 List

Listed below are the June 2014 The Green500's energy-efficient supercomputers ranked from 1 to 100.

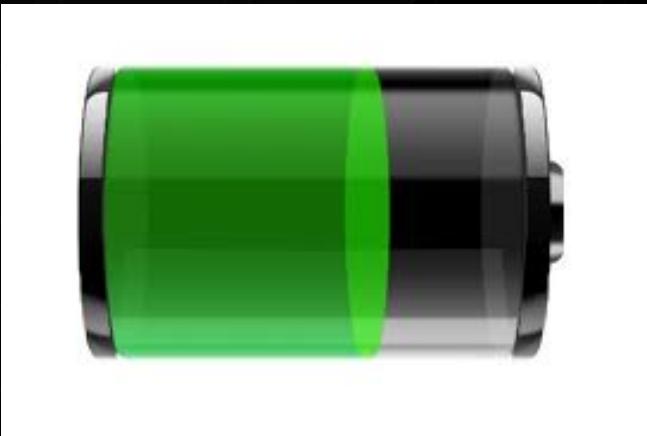
Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	4,389.82	GSIC Center, Tokyo Institute of Technology	TSUBAME-KFC - LX 1U-4GPU/104Re-1G Cluster, Intel Xeon E5-2620v2 6C 2.100GHz, Infiniband FDR, NVIDIA K20x	34.58
2	3,631.70	Cambridge University	Wilkes - Dell T620 Cluster, Intel Xeon E5-2630v2 6C 2.600GHz, Infiniband FDR, NVIDIA K20	52.62
3	3,517.84	Center for Computational Sciences, University of Tsukuba	HA-PACS TCA - Cray 3623G4-SM Cluster, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband QDR, NVIDIA K20x	78.77
4	3,459.46	SURFsara	Cartesius Accelerator Island - Bullx B515 cluster, Intel Xeon E5-2450v2 8C 2.5GHz, InfiniBand 4x FDR, Nvidia K40m	44.40
5	3,185.91	Swiss National Supercomputing Centre (CSCS)	Piz Daint - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect, NVIDIA K20x Level 3 measurement data available	1,753.66
6	3,131.06	ROMEO HPC Center - Champagne-Ardenne	romeo - Bull R421-E3 Cluster, Intel Xeon E5-2650v2 8C 2.600GHz, Infiniband FDR, NVIDIA K20x	81.41
7	3,019.72	CSIRO	CSIRO GPU Cluster - Nitro G16 3GPU, Xeon E5-2650 8C 2GHz, Infiniband FDR, Nvidia K20m	86.20
8	2,951.95	GSIC Center, Tokyo Institute of Technology	TSUBAME 2.5 - Cluster Platform SL390s G7, Xeon X5670 6C 2.93GHz, Infiniband QDR, NVIDIA K20x	927.86
9	2,813.14	Exploration & Production - Eni S.p.A.	HPC2 - iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.8GHz, Infiniband FDR, NVIDIA K20x	1,067.49
10	2,678.41	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	54.60
11	2,629.42	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband FDR, NVIDIA K20x	66.25
12	2,629.42	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband FDR, NVIDIA K20x	66.25
13	2,629.42	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband FDR, NVIDIA K20x	66.25
14	2,629.42	Financial Institution	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband FDR, NVIDIA K20x	66.25
15	2,629.10	Max-Planck-Gesellschaft MPI IPP	iDataPlex DX360M4, Intel Xeon E5-2680v2 10C 2.800GHz, Infiniband, NVIDIA K20x	269.94

June 2014 Green 500 list, Tesla powers 15 of the most energy-efficient supercomputers

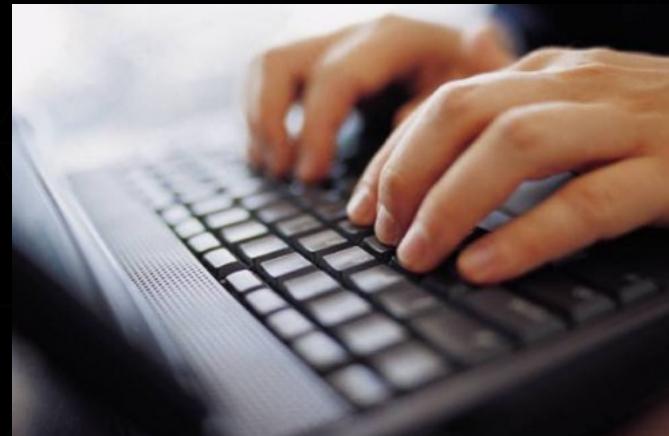
First sweep since IBM BlueGene

Tsubame-KFC: 4.3 GFLOPS / Watt

# OVERARCHING GOALS FOR TESLA



Power  
Efficiency

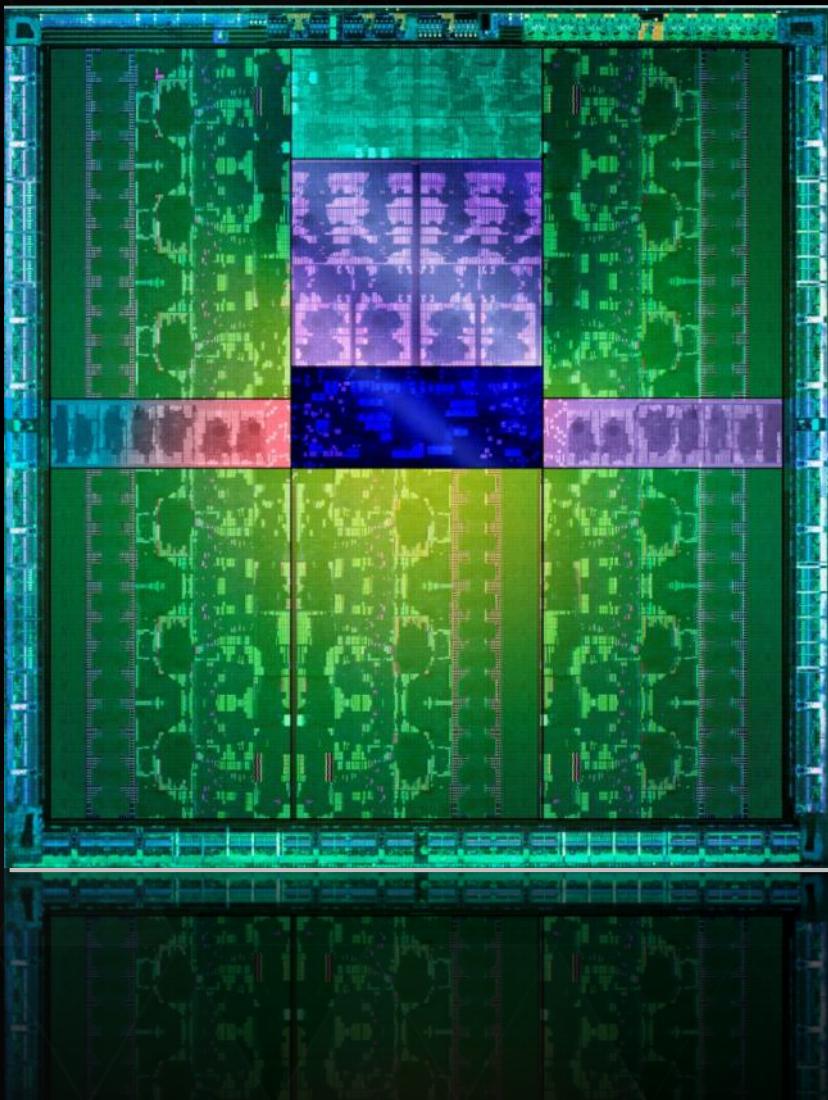


Ease of  
Programming  
And Portability



Application  
Space  
Coverage

## GK110 GPU



# KEPLER

## THE WORLD'S FASTEST, MOST EFFICIENT HPC ACCELERATOR

SMX

*(power efficiency)*

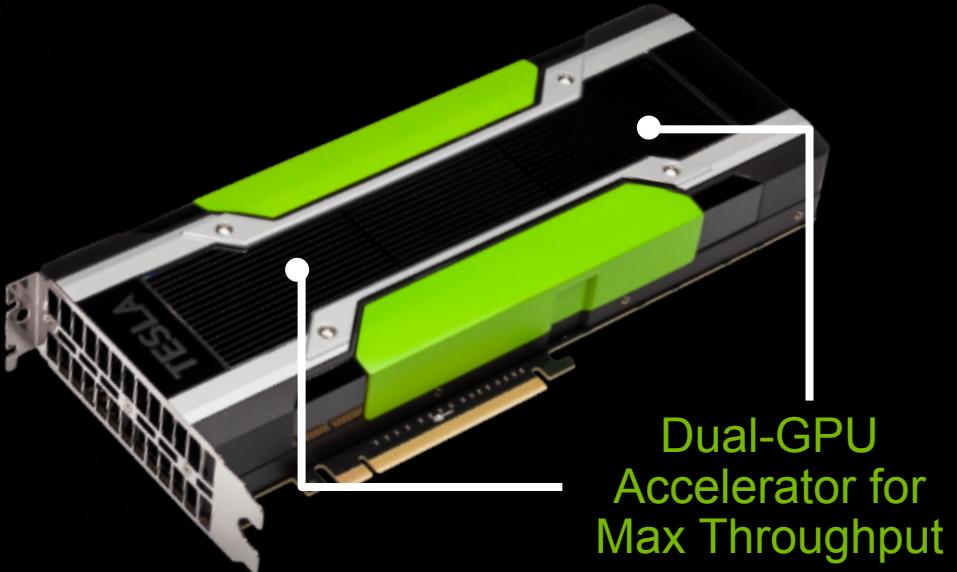
Hyper-Q

*(programmability and application coverage)*

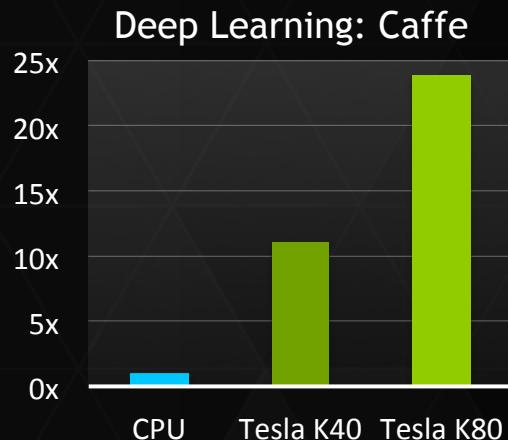
Dynamic Parallelism

# TESLA K80

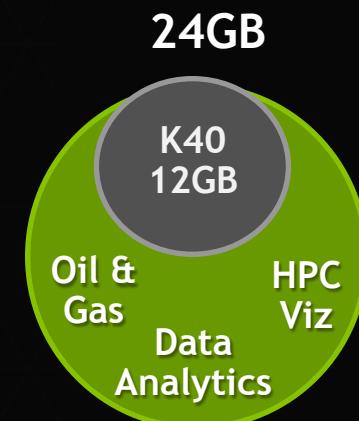
WORLD'S FASTEST ACCELERATOR  
FOR DATA ANALYTICS AND  
SCIENTIFIC COMPUTING



**2x Faster**  
2.9 TF | 4992 Cores | 480 GB/s



**Double the Memory**  
Designed for Big Data Apps



**Maximum Performance**  
Dynamically Maximize Perf for Every Application

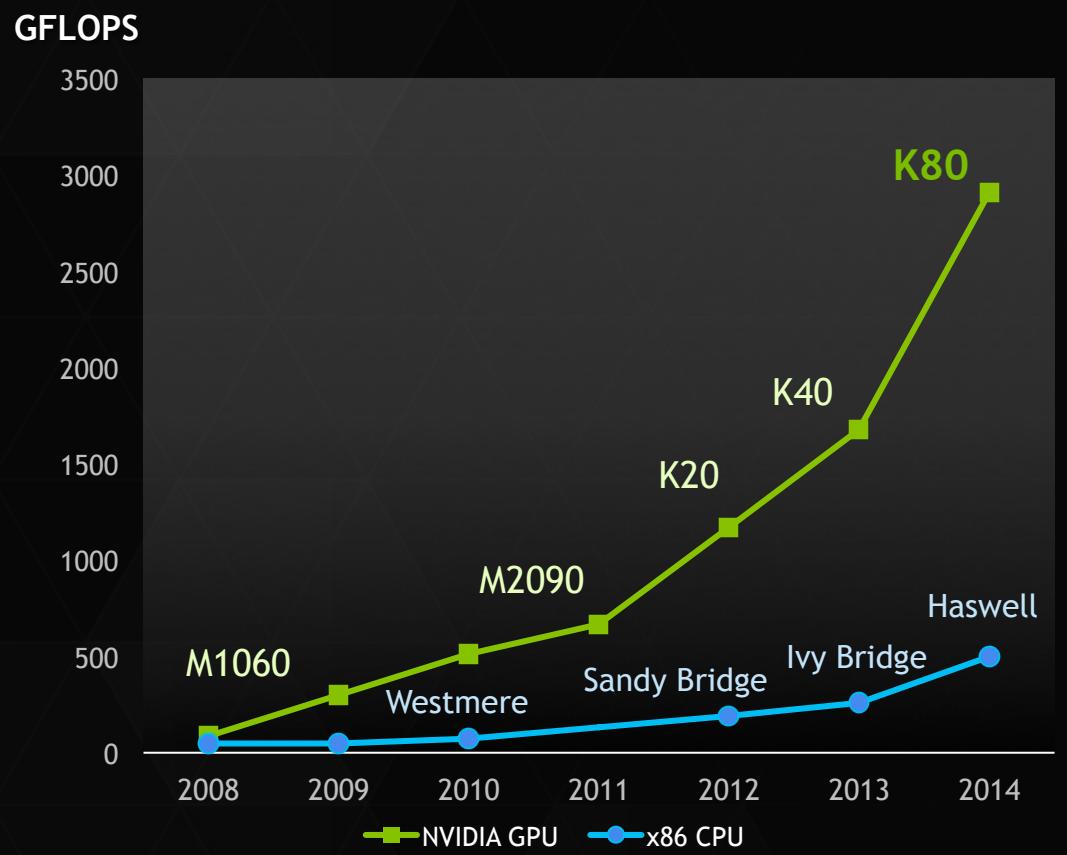


**GPU Boost**

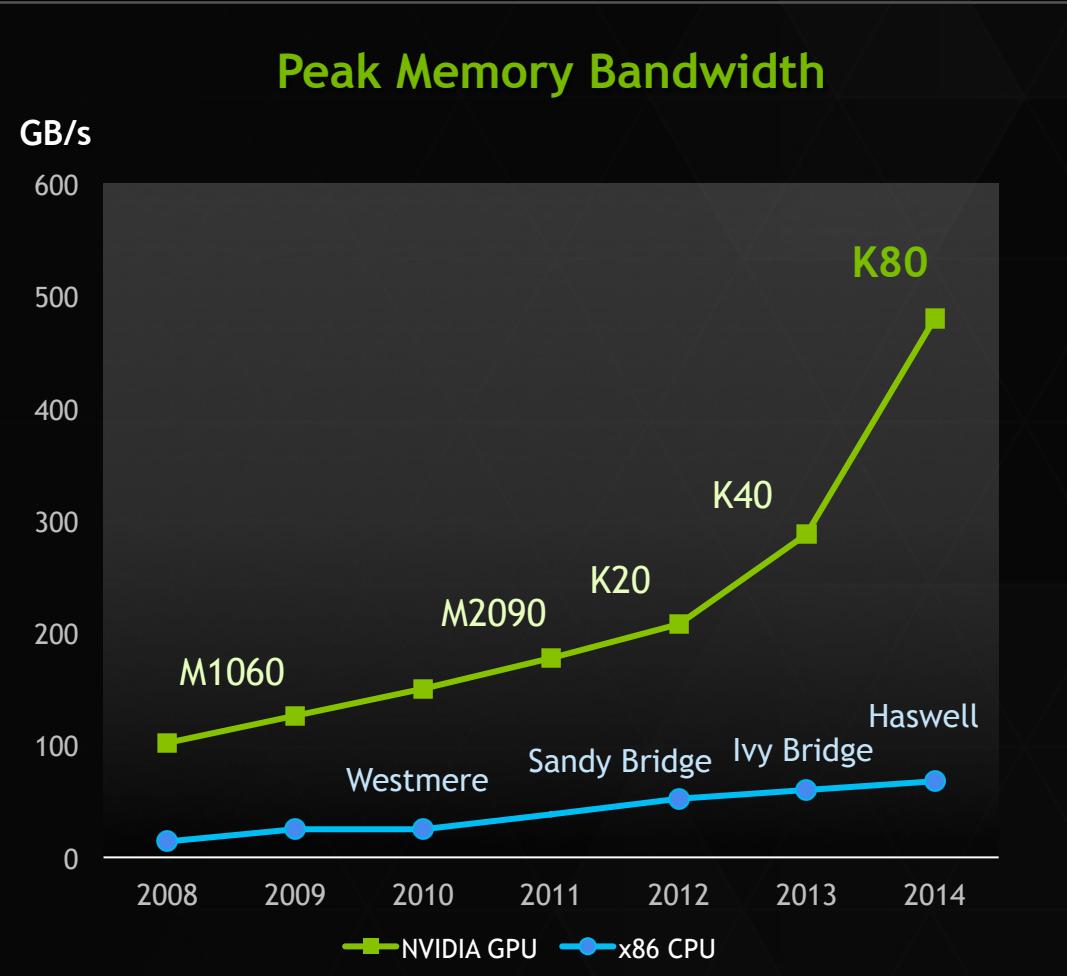
Caffe Benchmark: AlexNet training throughput based on 20 iterations, CPU: E5-2697v2 @ 2.70GHz. 64GB System Memory, CentOS 6.2

# PERFORMANCE LEAD CONTINUES TO GROW

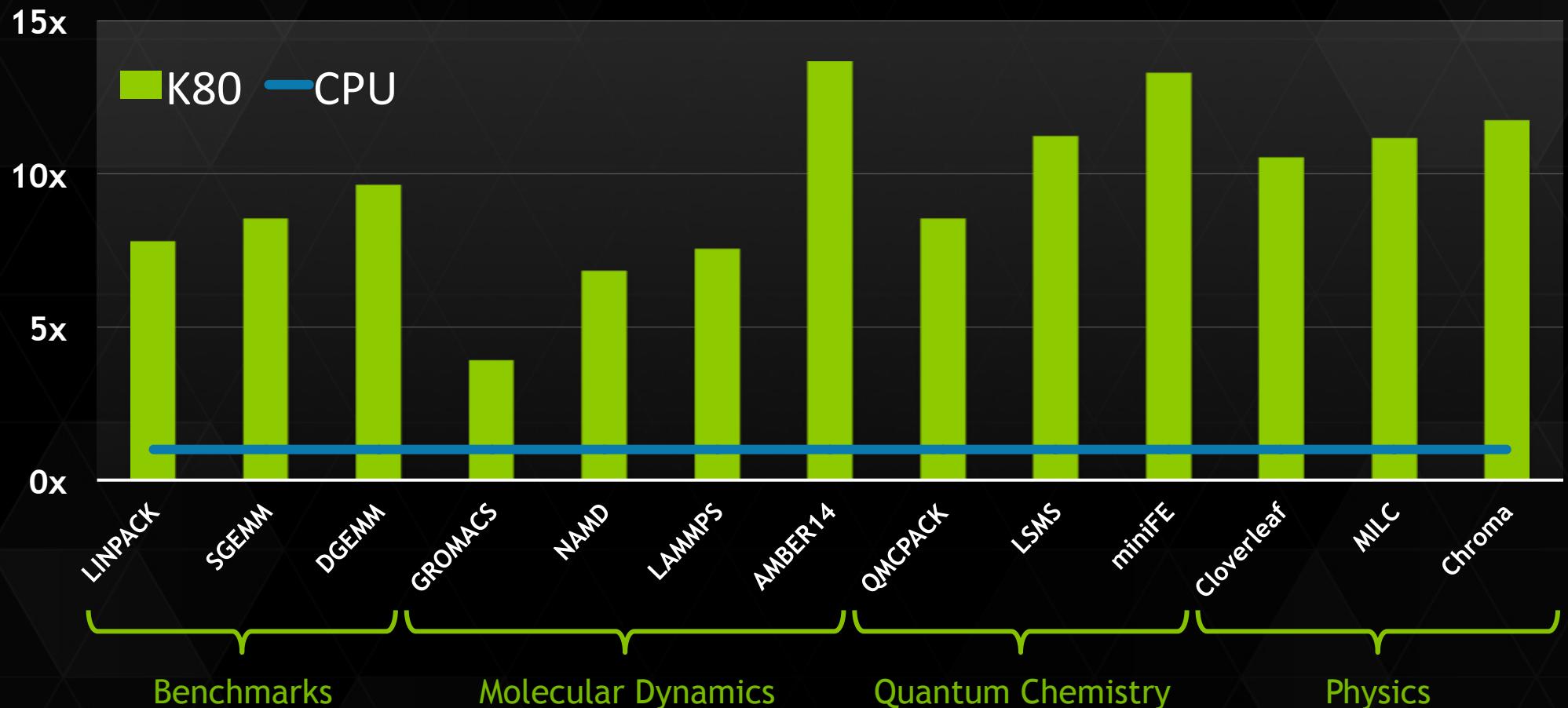
Peak Double Precision FLOPS



Peak Memory Bandwidth



# TESLA K80: 10X FASTER ON REAL-WORLD APPS

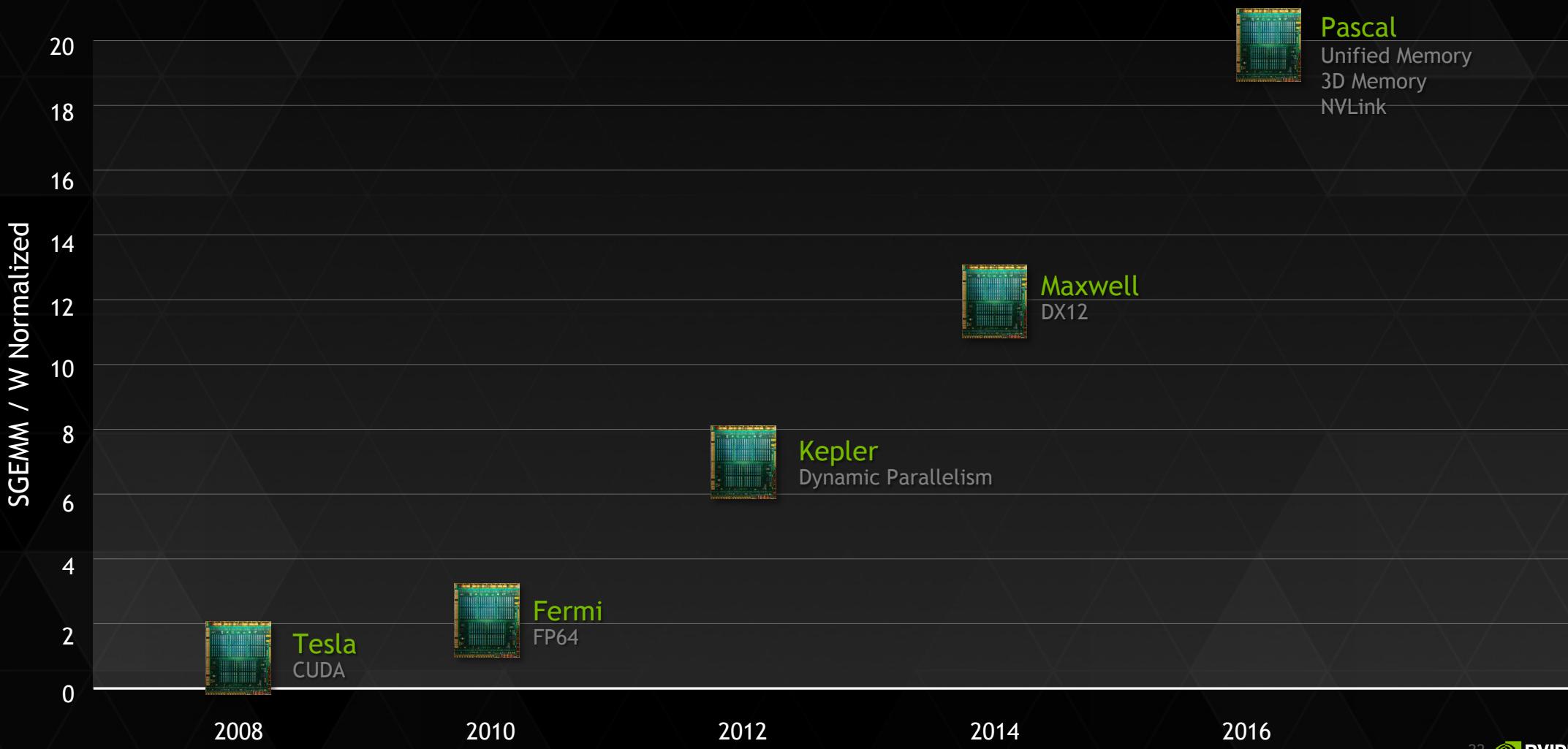


CPU: 12 cores, E5-2697v2 @ 2.70GHz. 64GB System Memory, CentOS 6.2  
GPU: Single Tesla K80, Boost enabled

# WHAT DOES THE FUTURE HOLD?

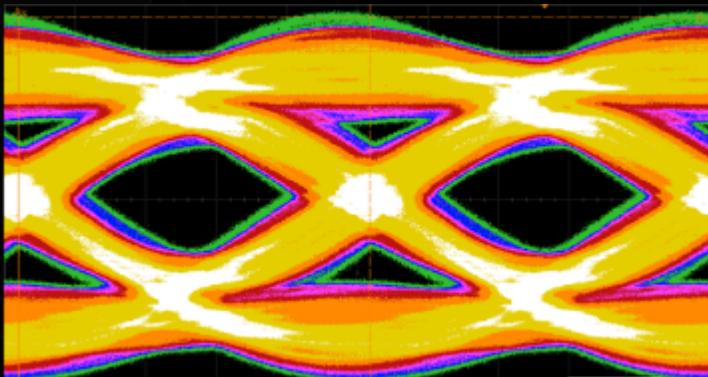


# FAST PACED CUDA GPU ROADMAP



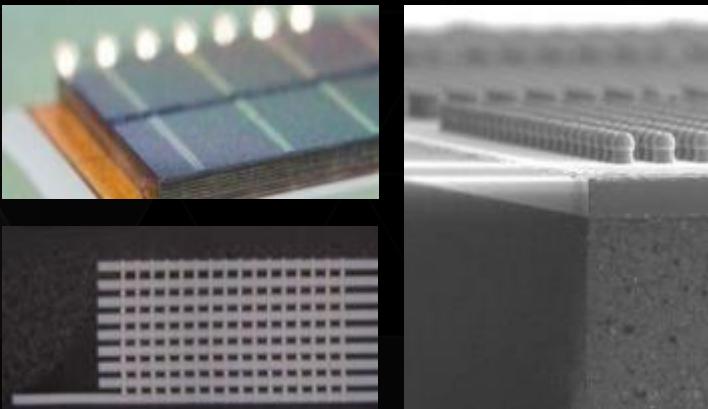
# PASCAL GPU FEATURES

## NVLINK AND STACKED MEMORY



### NVLINK

- GPU high speed interconnect
- 80-200 GB/s



### 3D Stacked Memory

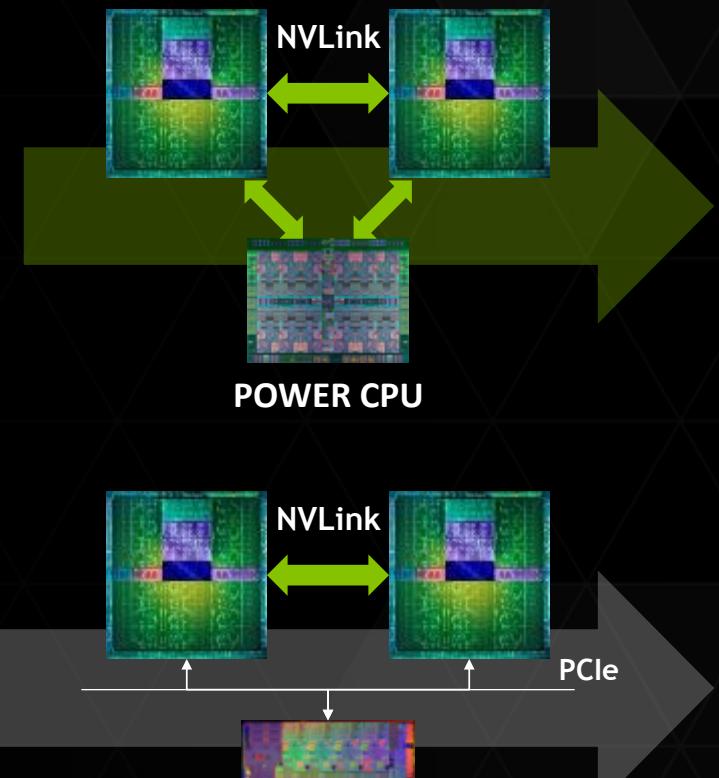
- 4x Higher Bandwidth (~1 TB/s)
- 3x Larger Capacity
- 4x More Energy Efficient per bit

# NVLINK HIGH-SPEED GPU INTERCONNECT



KEPLER GPU

PASCAL GPU

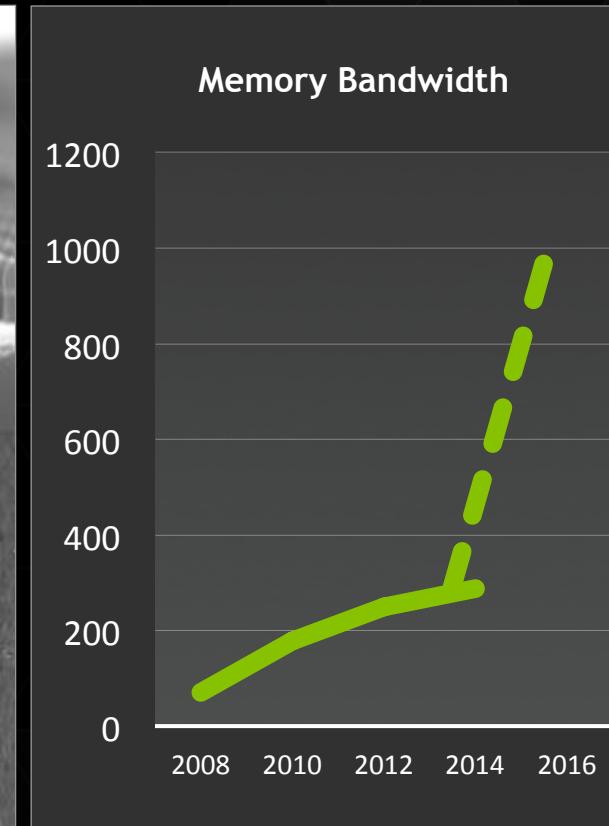
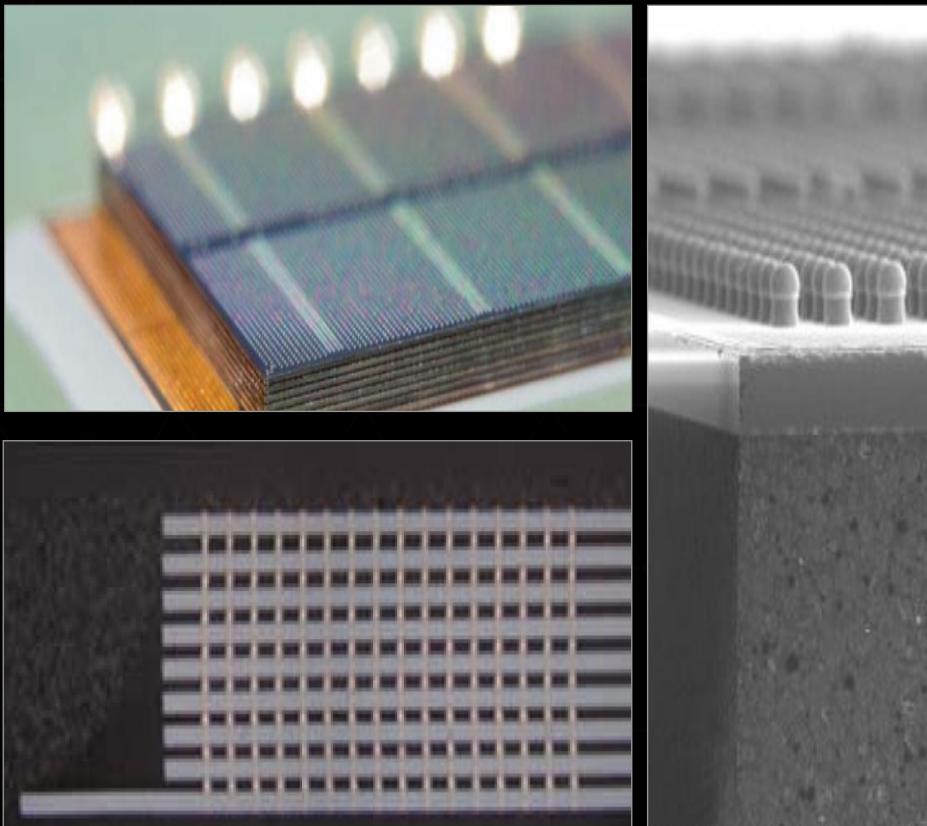


2014

2016

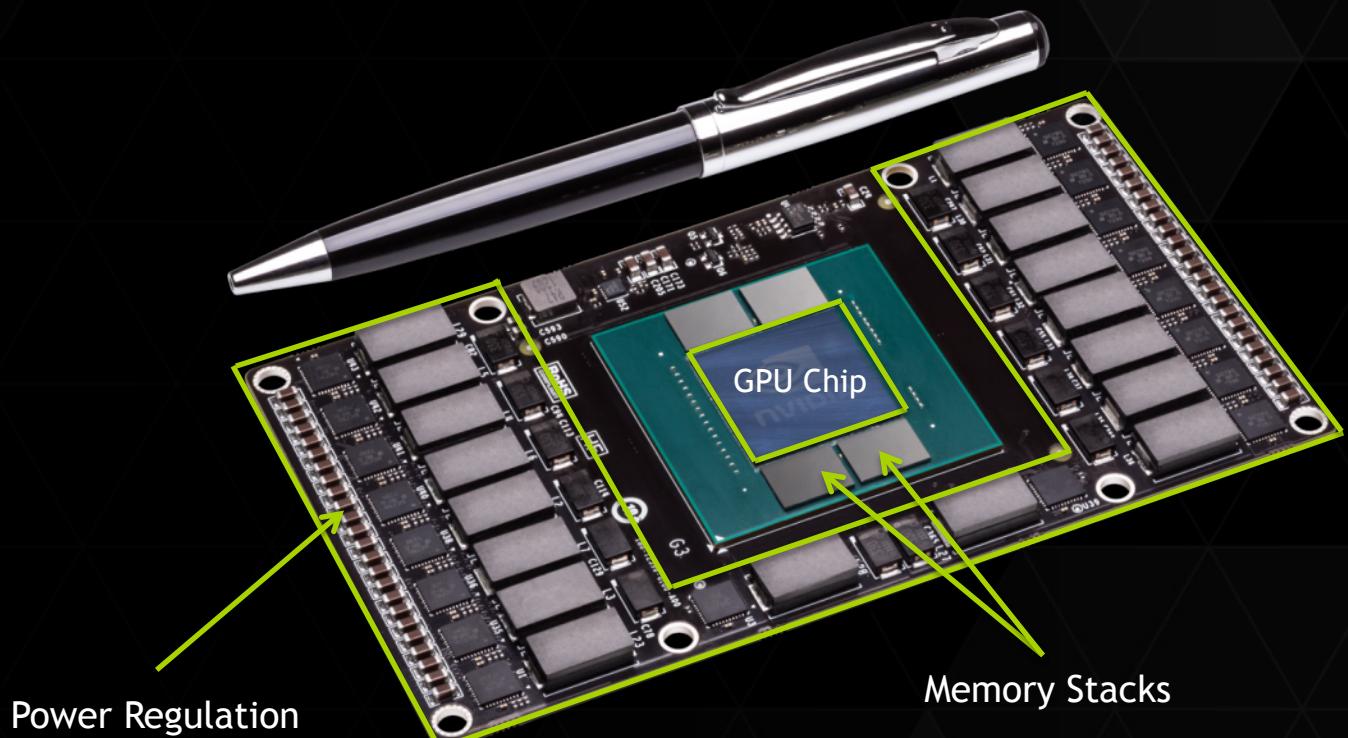
# 3D MEMORY

3D Chip-on-Wafer integration  
Many X bandwidth  
2.5X capacity  
4X energy efficiency



# PASCAL

NVLink 5 to 12X PCIe 3.0  
3D Memory 2 to 4X memory BW & size  
Module 1/3 size of PCIe card



# PARALLELISM IN MAINSTREAM LANGUAGES

- ▶ Enable more programmers to write parallel software
- ▶ Give programmers the choice of language to use
- ▶ GPU support in key languages



# C++ PARALLEL ALGORITHMS LIBRARY

```
std::vector<int> vec = ...  
  
// previous standard sequential loop  
std::for_each(vec.begin(), vec.end(), f);  
  
// explicitly sequential loop  
std::for_each(std::seq, vec.begin(), vec.end(), f);  
  
// permitting parallel execution  
std::for_each(std::par, vec.begin(), vec.end(), f);
```

- Complete set of parallel primitives:  
for\_each, sort, reduce, scan, etc.
- ISO C++ committee voted unanimously to accept as official tech. specification working draft

A Parallel Algorithms Library | N3724

Jared Hoberock     Jaydeep Marathe     Michael Garland     Olivier Giroux  
Vinod Grover     {jhoberock, jmarathe, mgarland, ogiroux, vgrover}@nvidia.com  
Artur Laksberg    Herb Sutter     {arturl, hsutter}@microsoft.com    Arch Robison

Document Number: N3960  
Date: 2014-02-28  
Reply to: Jared Hoberock  
NVIDIA Corporation  
jhoberock@nvidia.com

Working Draft, Technical Specification for C++ Extensions for Parallelism, Revision 1

N3960 Technical Specification Working Draft:  
<http://www.open-std.org/jtc1/sc22/wg21/docs/papers/2014/n3960.pdf>  
Prototype:  
<https://github.com/n3554/n3554>

# LINUX GCC COMPILER TO SUPPORT GPU

Open Source  
Free to all Linux users  
Most Widely Used HPC Compiler

**OpenACC**  
Directives for Accelerators



**“ Incorporating OpenACC into GCC is an excellent example of open source and open standards working together to make accelerated computing broadly accessible to all Linux developers. ”**

Oscar Hernandez  
Oak Ridge National Laboratories



# NUMBA PYTHON COMPILER

- ▶ Free and open source compiler for array-oriented Python
- ▶ NEW numba.cuda module integrates CUDA directly into Python

```
@cuda.jit("void(float32[:], float32, float32[:], float32[:])")
def saxpy(out, a, x, y):
    i = cuda.grid(1)
    out[i] = a * x[i] + y[i]

# Launch saxpy kernel
saxpy[griddim, blockdim](out, a, x, y)
```

▶ <http://numba.pydata.org/>



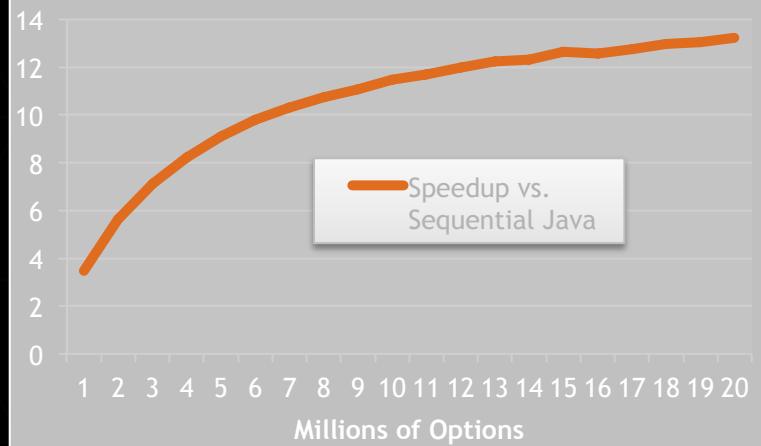
# COMPILE JAVA FOR GPUS



- Approach: apply a closure to a set of arrays

```
// vector addition
float[] X = {1.0, 2.0, 3.0, 4.0, ... };
float[] Y = {9.0, 8.1, 7.2, 6.3, ... };
float[] Z = {0.0, 0.0, 0.0, 0.0, ... };
jog.foreach(X, Y, Z, new jogContext(),
    new jogClosureRet<jogContext>() {
        public float execute(float x, float y) {
            return x + y;
        }
    }
);
```

Java Black-Scholes Options Pricing Speedup



Learn More: \$4939:  
Vinod Grover: "Accelerating JAVA on GPUs"  
Wednesday, 17:30 - 17:55  
Room LL20C

- foreach iterations parallelized over GPU threads

# THE FUTURE OF HPC IS GREEN

- Power is the constraint
  - Vast majority of work *must* be done by cores designed for efficiency
- GPU computing has a sustainable model
  - Aligned with technology trends, supported by consumer markets
- Future evolution will focus on:
  - Integration (CPU, network, memory)
  - Increased generality - efficient on *any* code with high parallelism
- This is simply how computers will be built

