

High-Performance Computing

Past, Present ==> Future

Tack till Cetac och



Synective Labs

Dr. Olaf O. Storaasli USEC & Synective Labs

Future Technologies Group, ORNL '05-'12

Sr. Research Scientist, NASA '70-'05



Chalmers University of Technology 13 May 2013

High-Performance Computing

Past, Present ==> Future



Oslo Norway 22 May 2013

Dr. Olaf O. Storaasli USEC & Synective Labs
Future Technologies Group, ORNL '05-'12
Sr. Research Scientist, NASA '70-'05

High-Performance Computing

Past, Present ==> Future



Synective Labs

18 maj 2013


Dr. Olaf O. Storaasli USEC & Synective Labs
Future Technologies Group, ORNL '05-'12
Sr. Research Scientist, NASA '70-'05

High-Performance Computing

Past, Present ==> Future


Tack till




Synective Labs

14 maj 2013

Dr. Olaf O. Storaasli USEC & Synective Labs
Future Technologies Group, ORNL '05-'12
Sr. Research Scientist, NASA '70-'05

A man with a beard and a light-colored short-sleeved shirt stands behind a large, curved stone sign. He has his arms resting on the sign and is smiling. The sign is made of large, light-colored stone blocks and is set on a base of rough-hewn stones. In the background, there is a modern building with a large, blue-tinted glass facade that is shaped like a trapezoid. To the right, there is a long, multi-story building with a red brick facade and many windows. The sky is overcast and grey.

OAK RIDGE NATIONAL LABORATORY

MANAGED BY UT BATTELLE
FOR U.S. DEPARTMENT OF ENERGY

ORNL “X-10” History

1st Graphite Plutonium Reactor => PNL



ORNL: *US #1 Energy & Science Lab, #1 Materials*

- 4K + 3K **guest** researchers
- **#1 HPC** +NSF +NOAA
- \$1.3B/yr



HPC Speedup: 1000X per decade

Evolution of the fastest sustained performance
in real simulations



~1 Exaflop/s
~ 10^7 processing units
(?)

1.35 Petaflop/s

Cray XT5

$1.5 \cdot 10^5$ processor cores

(Shultness - ORNL)

1.02 Teraflop/s

Cray T_{3D}

$1.5 \cdot 10^3$ processors
(ORNL)

1.5 Gigaflop/s

Cray YMP

$0.8 \cdot 10^1$ processors
(Storaasli - NASA)

1989

1998

2008

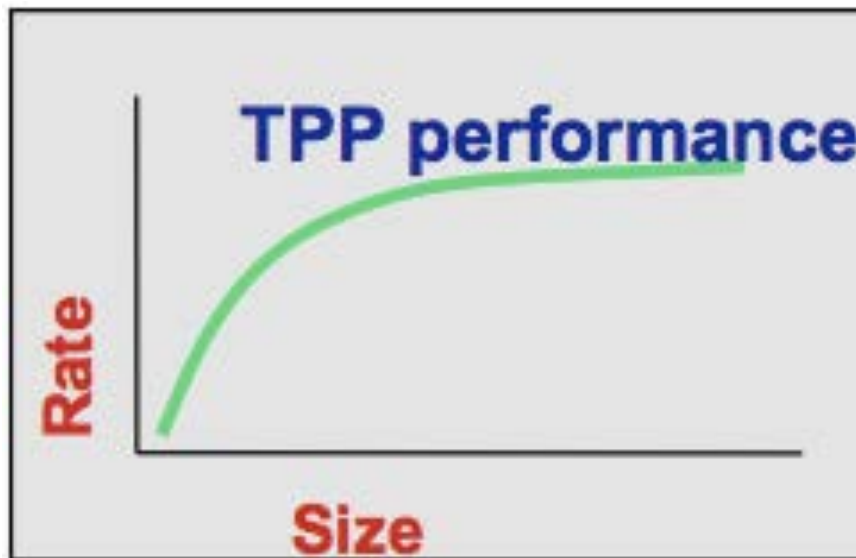
2018

1 Exaflop = 1,000 Petaflops = 1,000,000 Teraflops = 1,000,000,000 Gigaflops = 1,000,000,000,000 Mflops

World's 500 Fastest Computers

www.top500.org

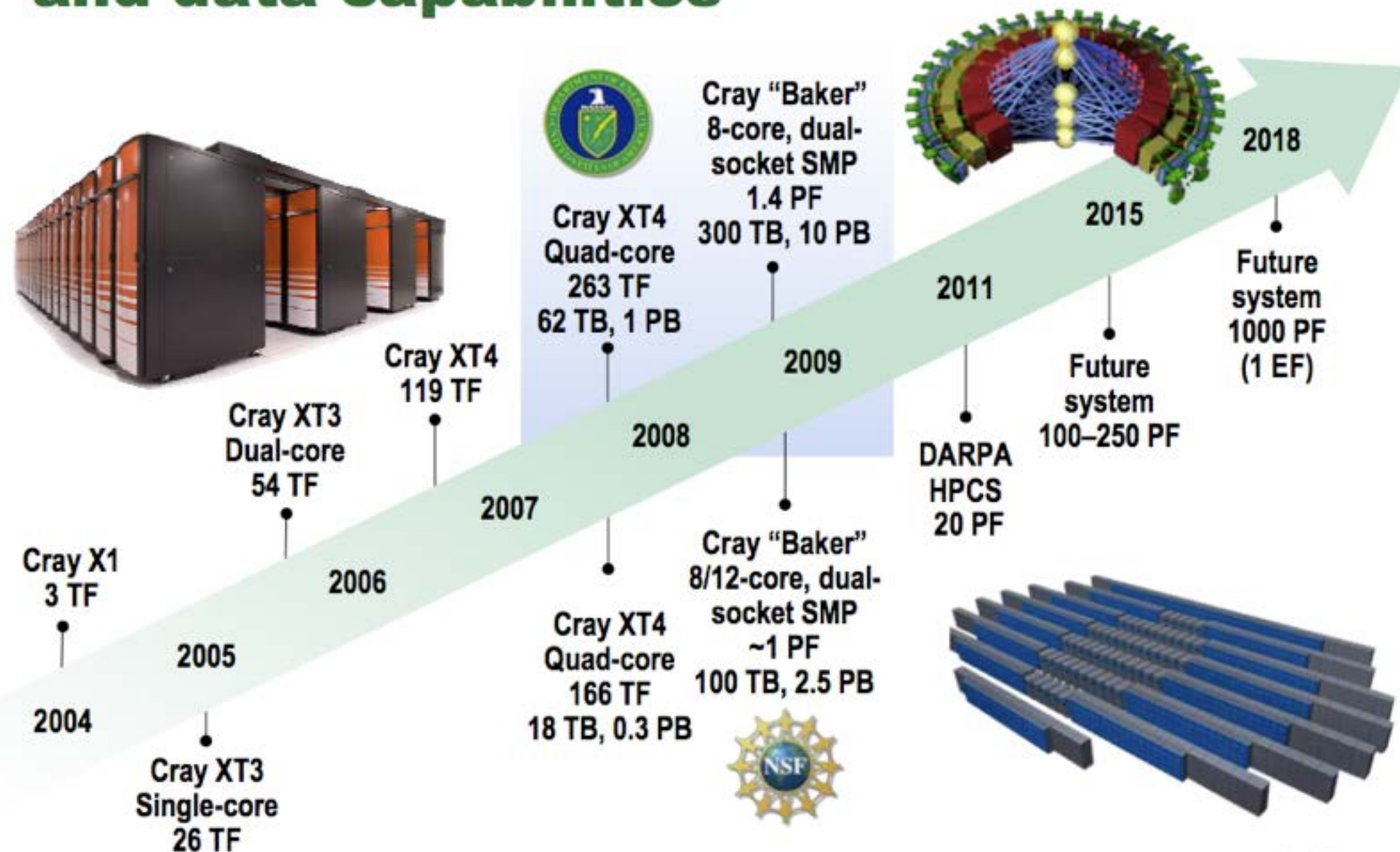
- Linpack* benchmark: $Ax=b$ (dense matrix A)



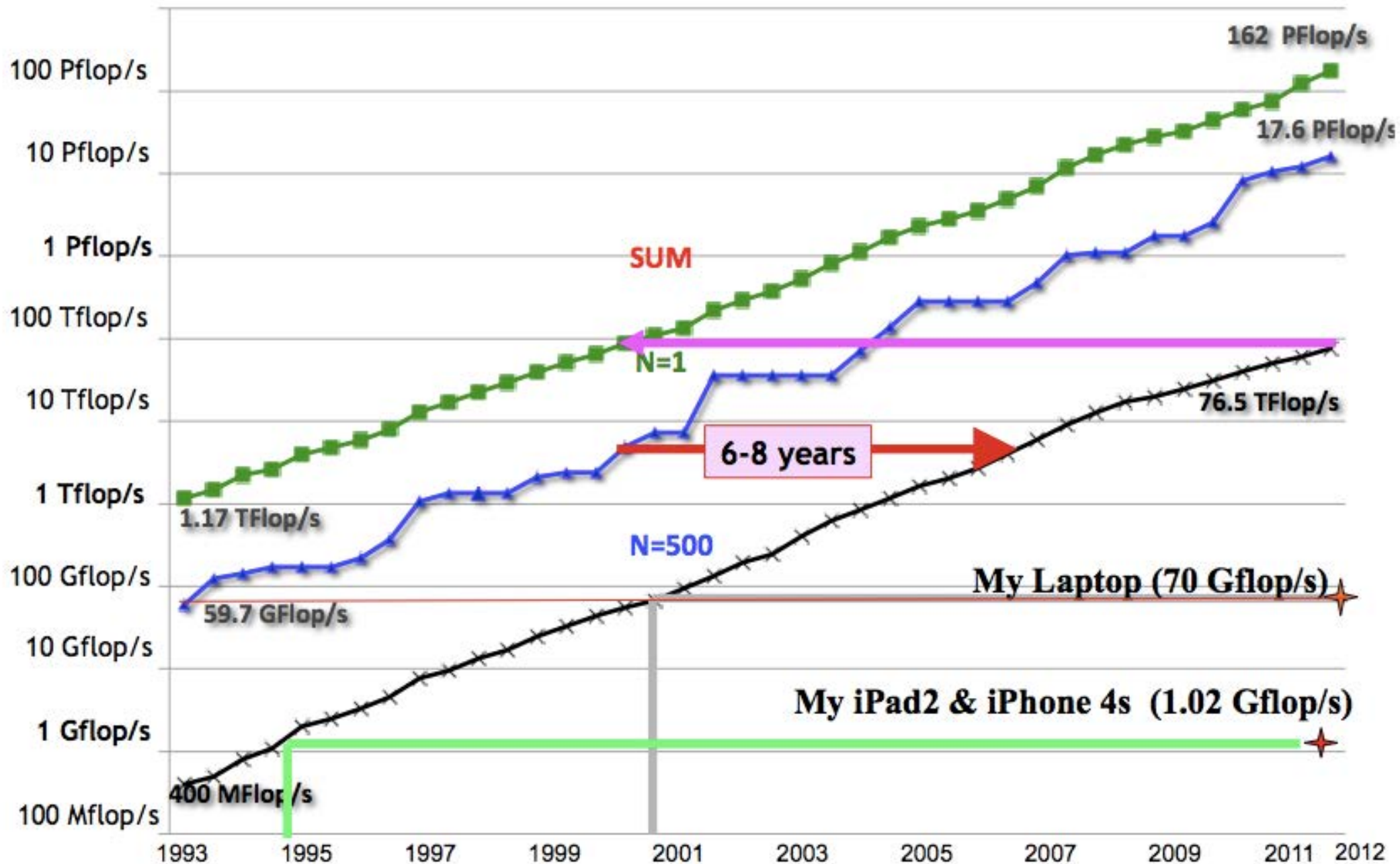
*Try App yourself (iPad, Mac etc.)

- Updated: June@ISC & Nov@SC

Million-fold increase in computing and data capabilities



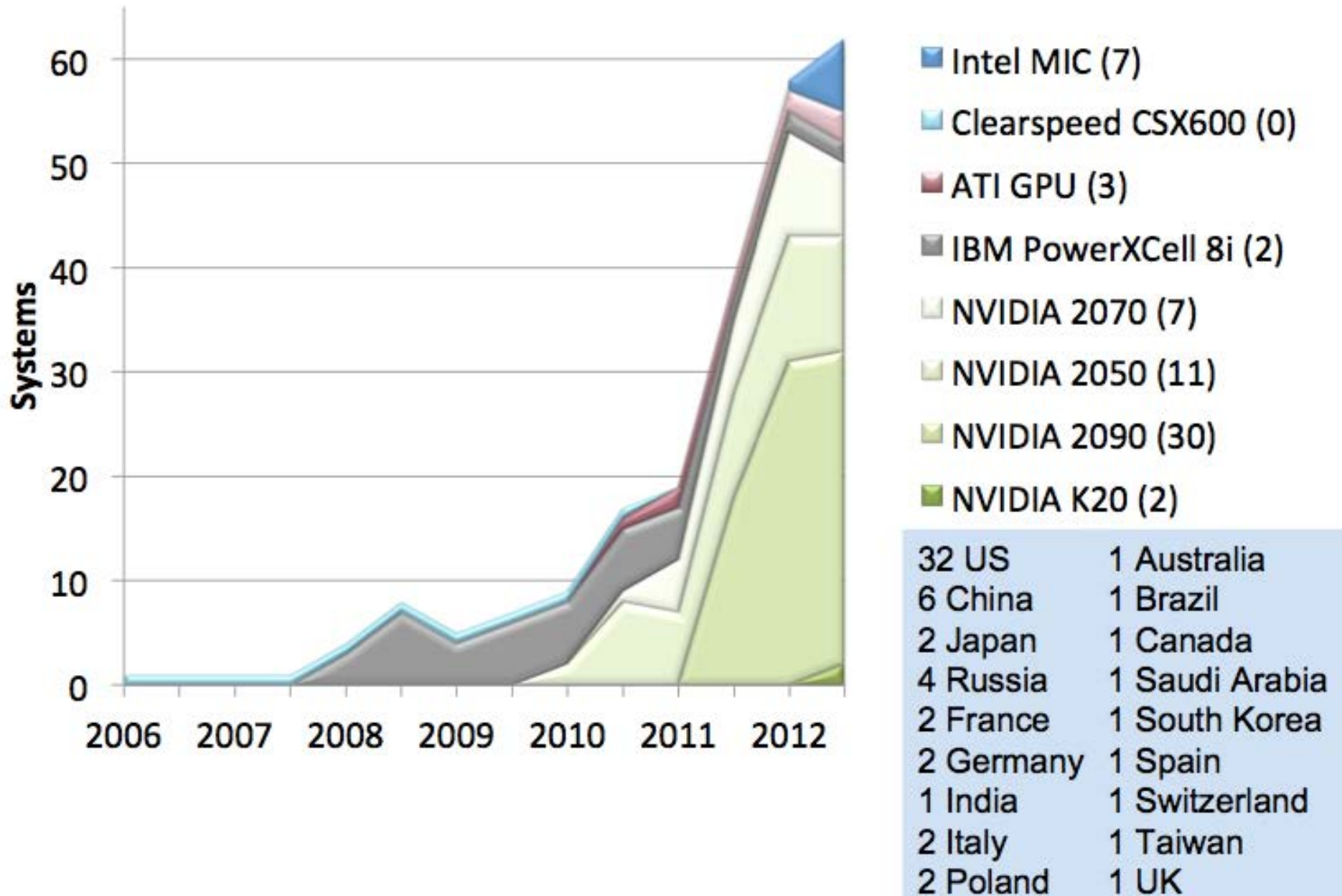
HPC Performance – 20 Years



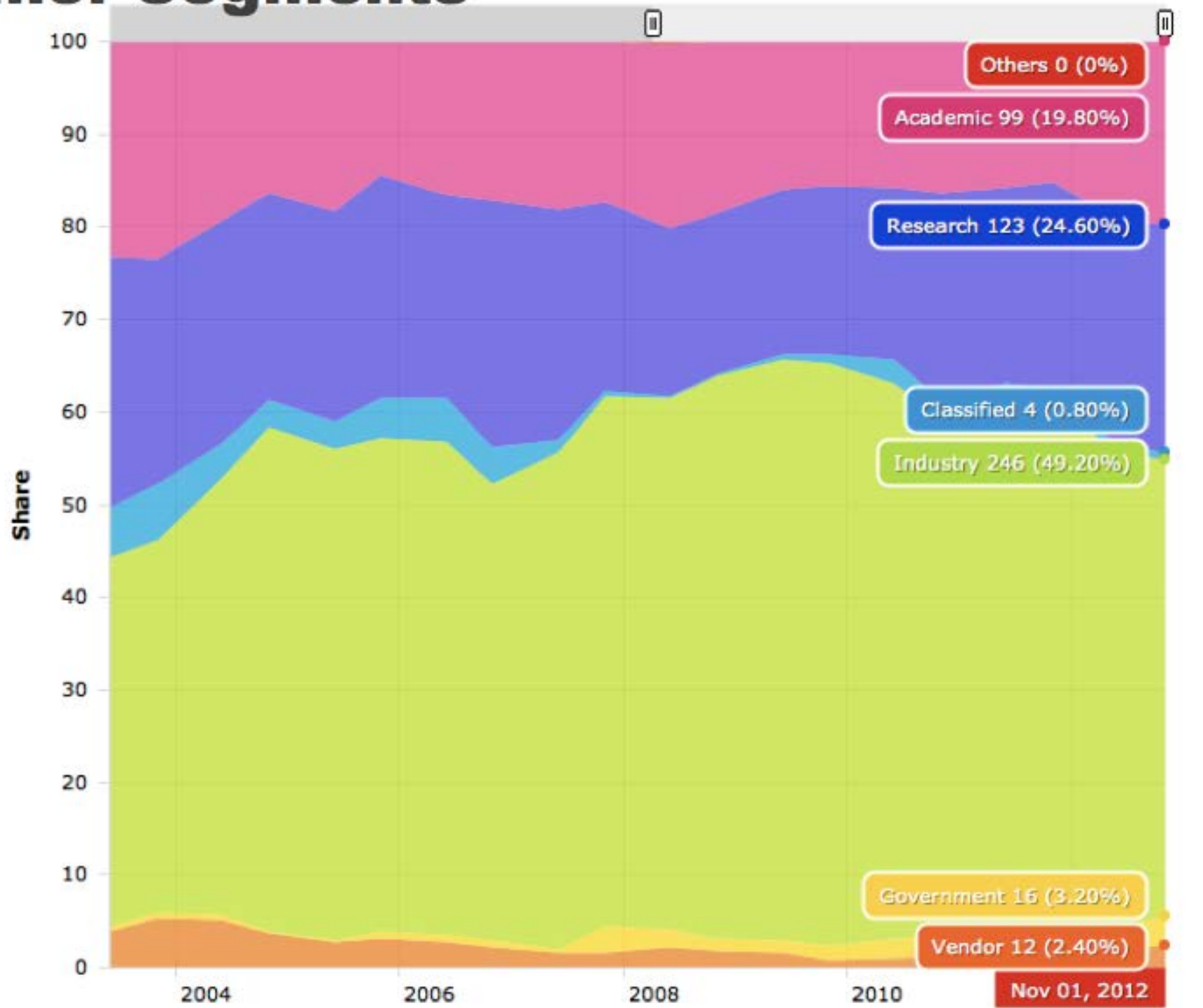
November 2012: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	MFlops /Watt
1	DOE / OS Oak Ridge Nat Lab	Titan, Cray XK7 (16C) + Nvidia Kepler GPU (14c) + custom	USA	560,640	17.6	66	8.3	2120
2	DOE / NNSA L Livermore Nat Lab	Sequoia, BlueGene/Q (16c) + custom	USA	1,572,864	16.3	81	7.9	2063
3	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx (8c) + custom	Japan	705,024	10.5	93	12.7	827
4	DOE / OS Argonne Nat Lab	Mira, BlueGene/Q (16c) + custom	USA	786,432	8.16	81	3.95	2066
5	Forschungszentrum Juelich	JuQUEEN, BlueGene/Q (16c) + custom	Germany	393,216	4.14	82	1.97	2102
6	Leibniz Rechenzentrum	SuperMUC, Intel (8c) + IB	Germany	147,456	2.90	90*	3.42	848
7	Texas Advanced Computing Center	Stampede, Dell Intel (8) + Intel Xeon Phi (61) + IB	USA	204,900	2.66	67	3.3	806
8	Nat. SuperComputer Center in Tianjin	Tianhe-1A, NUDT Intel (6c) + Nvidia Fermi GPU (14c) + custom	China	186,368	2.57	55	4.04	636
9	CINECA	Fermi, BlueGene/Q (16c) + custom	Italy	163,840	1.73	82	.822	2105
10	IBM	DARPA Trial System, Power7 (8C) + custom	USA	63,360	1.51	78	.358	422

Accelerators (62 systems)

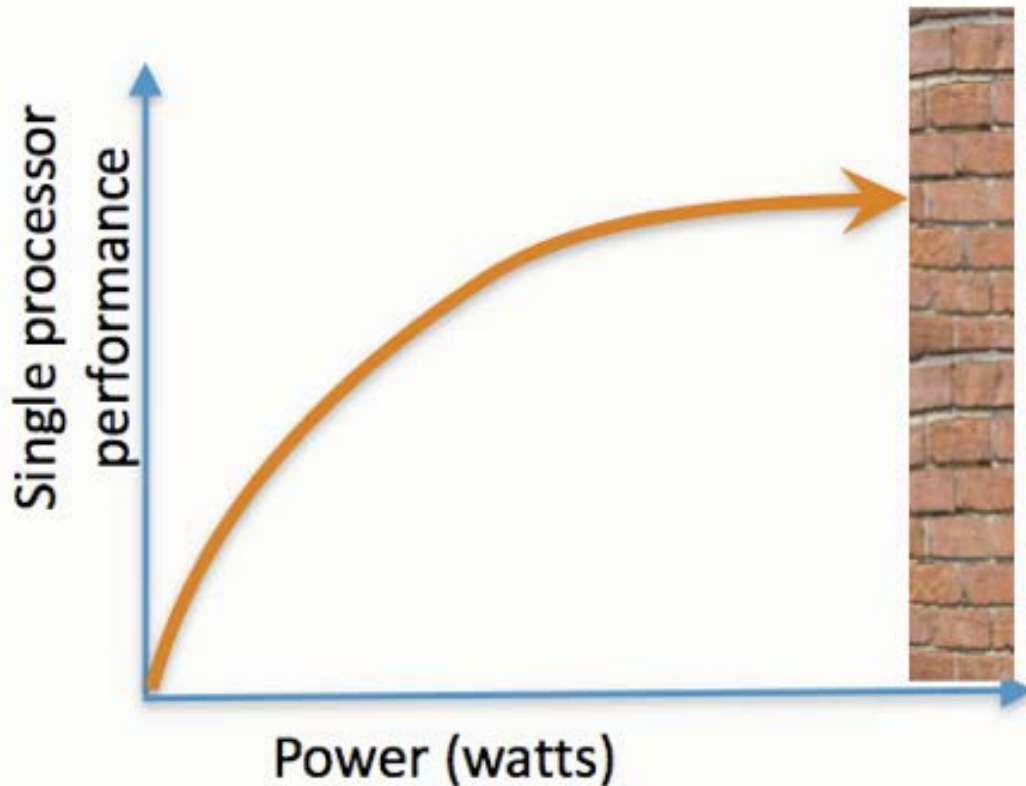


Customer Segments



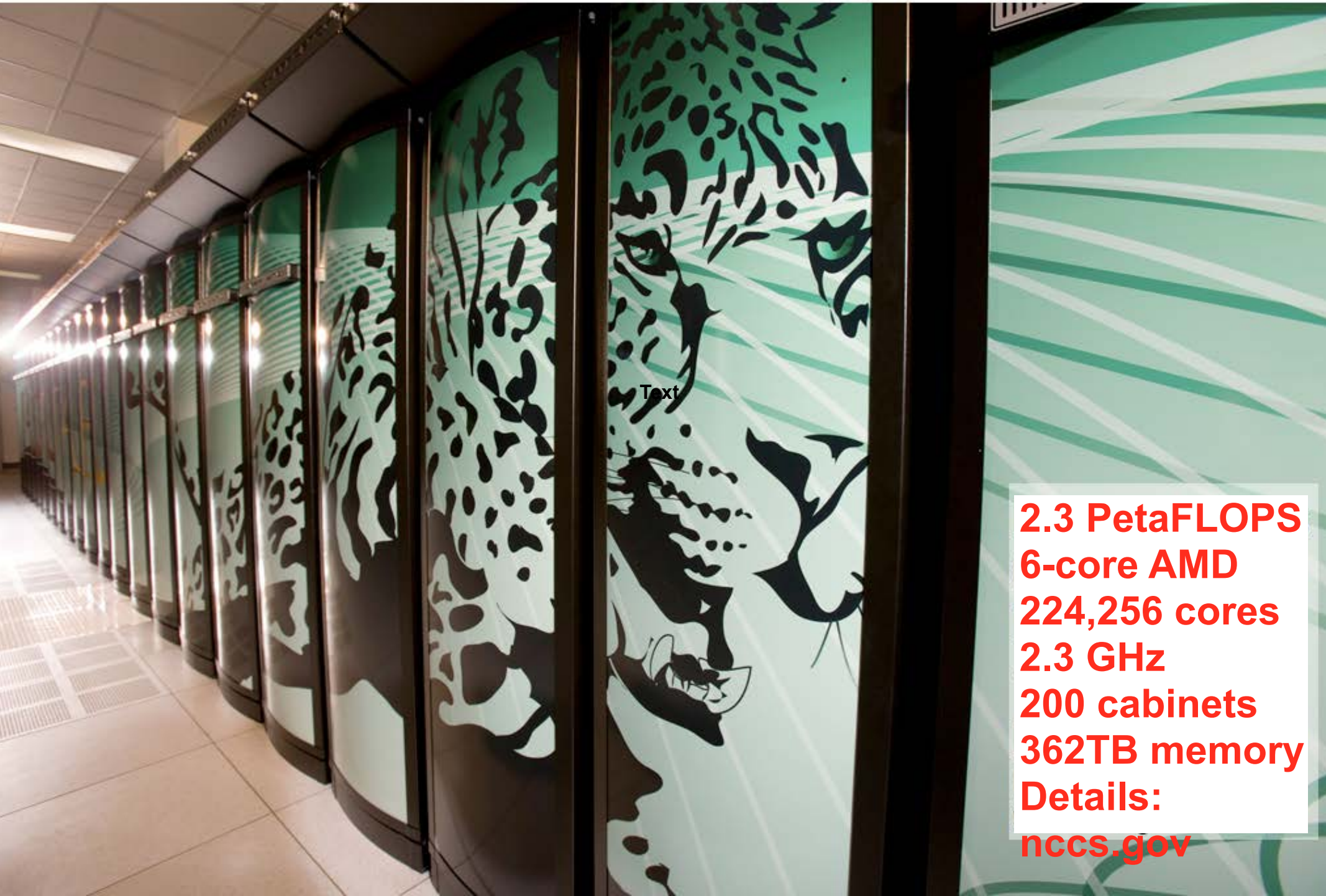
Computing has met a barrier

- In the “Good Old Days” performance doubled every 2 years
 - increased clock rate
 - architectural improvements
- But single threaded performance is increasingly limited by power & cooling



We have hit a “power wall”

Cray XT5 portion of Jaguar @ NCCS



Text

2.3 PetaFLOPS
6-core AMD
224,256 cores
2.3 GHz
200 cabinets
362TB memory
Details:
nccs.gov

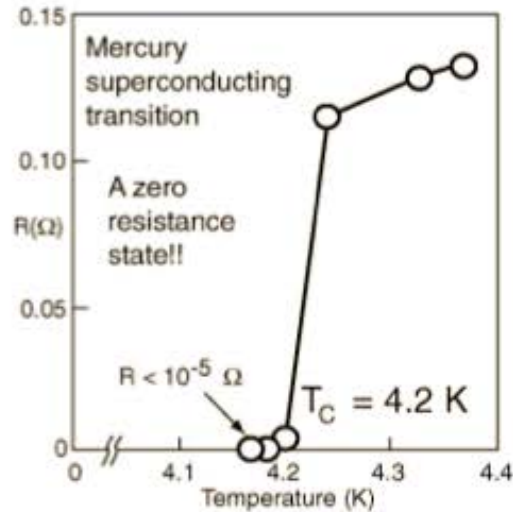
Area	Project Name	M Hrs	Institution
Astrophysics	Multidimensional Simulations of Core Collapse Supernovae	75	ORNL
Materials Sciences	Nanoscale MC Simulateton of Mott Insulators, Cuprate Superconductors	45	ORNL
Chemical Sciences	An Integrated Approach to the Rational Design of Chemical Catalysts	30	ORNL
Climate	Climate-Science Development & Grand Challenge Team	30	NCAR
Combustion	High-Fidelity Simulations for Clean, Efficient Combustion of Alternative Fuels	30	SNL
Fusion Plasma Energy	V&V off Turbulent Transport in Fusion Plasma Simulations	30	UCSD
Climate	CHIMES: Coupled High-Resolution Modeling of the Earth System-Princeton	24	NOAA/GFDL
Fusion Plasma Energy	High-fidelity tokamak edge simulation for confinement of fusion plasma	20	NYU
Fusion Plasma Energy	Validation of Plasma Microturbulence for Finite-Beta Fusion Experiments	20	LLNL
Lattice Gauge Theory	Lattice QCD	20	UCSB
Life Sciences	Gating Mechanism of Membrane Proteins	15	UChicago
Materials Sciences	Electronic, Lattice & Mechanical Properties of Nano-Structured Bulk Materials	15	GM
Nuclear Physics	Nuclear Structure	15	ORNL
Combustion	Clean and Efficient Coal Gasifier Designs using Large-Scale Simulations	13	NETL
Chemistry	Modeling Hydronium & OH- Ions in H2O & H2O/Air Interface via path Integrals	12	Catech
Geological Sciences	Modeling Reactive Flows in Porous Media	10	LLNL
Accelerator Physics	Terascale Particle Accelerator: International Linear Collider Design & Modeling	8	SLAC
Computer Science	Performance Evaluation and Analysis Consortium End Station	8	ORNL
Biophysics	Physical of Recalcitrance to Hydrolysis of Lignocellulosic Biomass	6	ORNL
Astrophysics	Intermittency and Star Formation in Turbulent Molecular Clouds	5	UCSD
Astrophysics	The Via Lactea Project: A Glimpse into the Invisible World of Dark Matter	5	UCSC
Nanoelectronics	Petascale Simulations of Nan-electronic Devices	5	Purdue
Climate	Climate Sensitivity & Abrupt Climate Change	4	UWisconsin
Astrophysics	Models of Type Ia Supernovae	3	UCSC
Biophysics	Interplay of AAA+ molecular machines, DNA repair enzymes & sliding clamps	3	UCSD
Chemistry	Dynamically tunable ferroelectric surface catalysts	2	Upa
Chemical Sciences	Molecular Simulation of Complex Chemical Systems	2	PNNL
Climate	Simulation of Global Cloudiness	2	ColoradoSU
Fusion Plasma Energy	Gyrokinetic Steady State Transport Simulations	2	Gen Atomics
Fusion Plasma Energy	High Power Electromagnetic Wave Heating in the ITER Burning Plasma	2	ORNL

Superconductivity: a state of matter with zero electrical resistivity

Discovery 1911



Heike Kamerlingh Onnes (1853-1926)



Superconductor repels magnetic field
Meissner and Ochsenfeld, **Berlin 1933**



Microscopic Theory for Superconductivity 1957

PHYSICAL REVIEW

VOLUME 108, NUMBER 2

DECEMBER 1, 1957

Theory of Superconductivity*

J. BARDEEN, L. N. COOPER,† AND J. R. SCHRIEFFER‡
Department of Physics, University of Illinois, Urbana, Illinois
(Received July 8, 1957)

A theory of superconductivity is presented, based on the fact that the interaction between electrons resulting from virtual exchange of phonons is attractive when the energy difference between the electron states involved is less than the phonon energy, $\hbar\omega$. It is favorable to form a superconducting phase when this attractive interaction dominates the repulsive screened Coulomb interaction. The normal phase is described by the Bloch individual-particle model. The ground state of a superconductor, formed from a linear combination of normal state configurations in which electrons are virtually excited in pairs of opposite spin and momentum, is lower in energy than the normal state by an amount proportional to an average $(\hbar\omega)^2$, consistent with the isotope effect. A mutually orthogonal set of excited states in

one-to-one correspondence with those of the normal phase is obtained by specifying occupation of certain Bloch states and by using the rest to form a linear combination of virtual pair configurations. The theory yields a second-order phase transition and a Meissner effect in the form suggested by Pippard. Calculated values of specific heats and penetration depths and their temperature variation are in good agreement with experiment. There is an energy gap for individual-particle excitations which decreases from about 3.5kT_c at T=0°K to zero at T_c. Tables of matrix elements of single-particle operators between the excited-state superconducting wave functions, useful for perturbation expansions and calculations of transition probabilities, are given.



Scanned at the American Institute of Physics



Scanned at the American Institute of Physics



Scanned at the American Institute of Physics

BCS Theory generally accepted in the early 1970s

Courtesy of Thomas Schultze

New algorithm to enable 1+ PFlop/s sustained performance in simulations of disorder effects in high- T_c superconductors

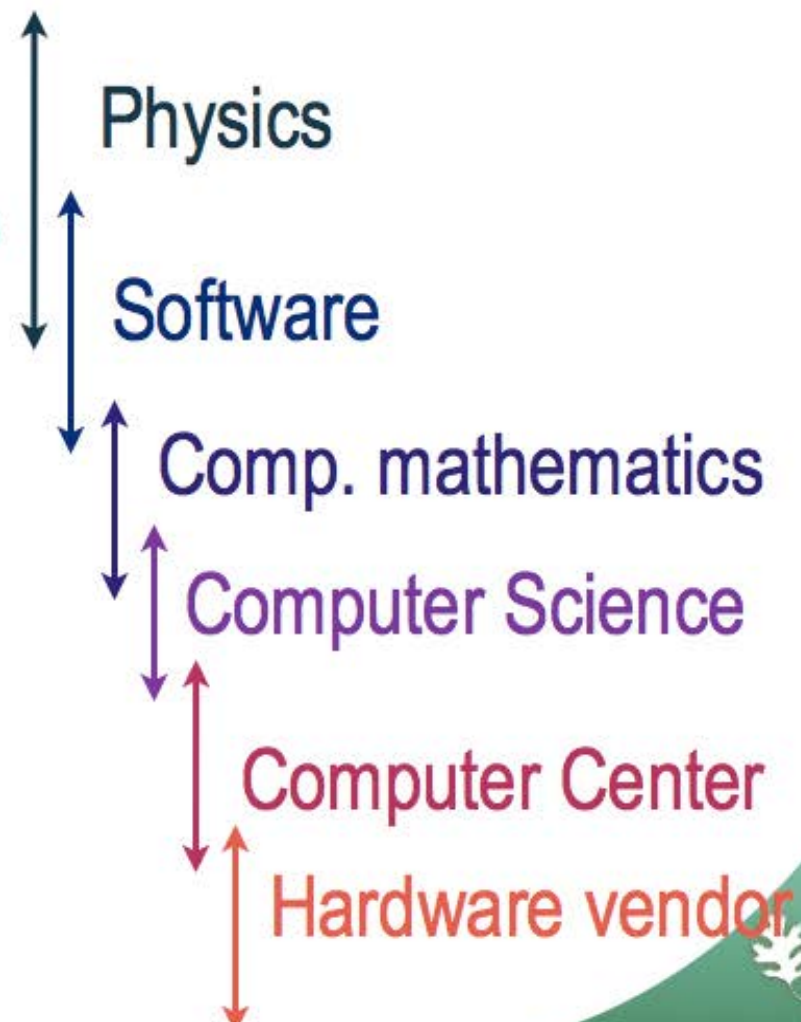
Models,
Methods,
& Implementation

Map to Hardware

Operations

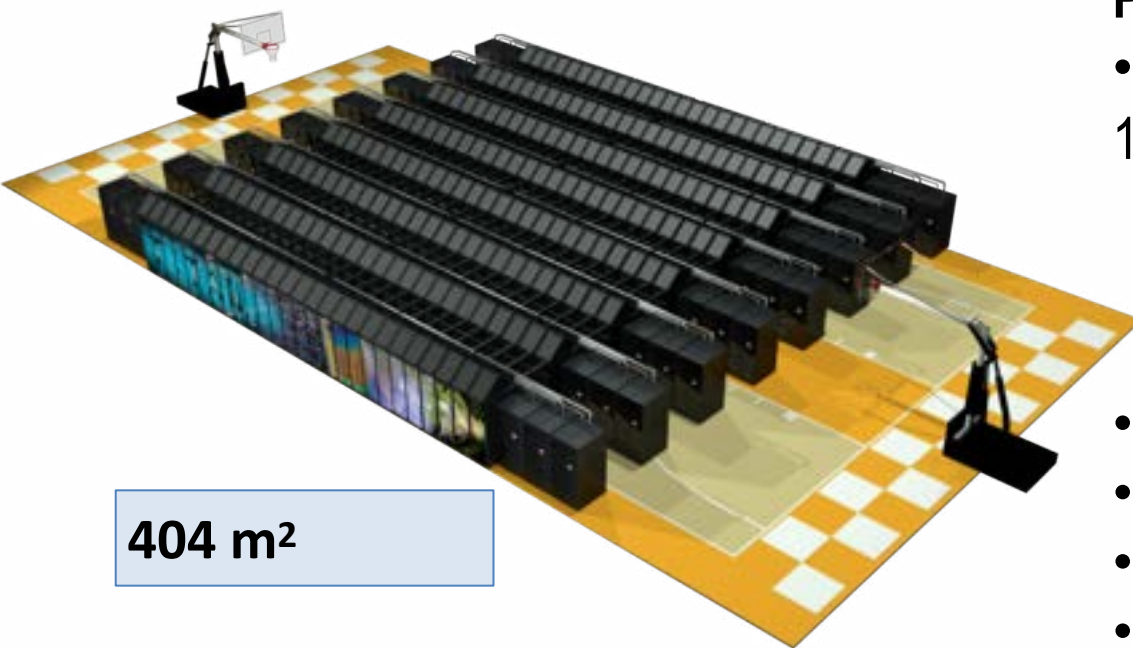
System design

T. A. Maier
P. R. C. Kent
T. C. Schulthess
G. Alvarez
M. S. Summers
E. F. D'Azevedo
J. S. Meredith
M. Eisenbach
D. E. Maxwell
J. M. Larkin
J. Levesque



Present: ORNL's *Titan* Hybrid (Cray XK7)

AMD Opteron + NVIDIA Tesla



PERFORMANCE SPECS:

- Peak: **27.1 PF/s**: 24.5 GPU + 2.6 CPU
- 18,688 Compute Nodes x 161GF:
 - 16-Core AMD Opteron CPU
 - NVIDIA Tesla "K20x" GPU
 - **Memory: 32GB CPU + 6 GB GPU**
- 710 TB total system memory
- 512 Service & I/O nodes
- 200 Cabinets
- Cray Gemini 3D Torus Interconnect
- 8.9 MW peak power

Phase 1: Jaguar Upgrade => Titan

XK5 => XT7 nodes, new fans, power supplies + 3.3MW Transformer

Replacing board



100 Shipping Crates



Adding Power Supplies



Reused Jaguar parts:

- Cabinets
- Backplanes
- RAS System
- File System
- Interconnect cables
- Liquid Cooling System

**Upgrade saved
\$25M over new
system cost!**



New Fans



**Electric
Switchboard**

Phase 2: Jaguar Upgrade => Titan

- NVIDIA K20x GPU => 18,688 compute nodes

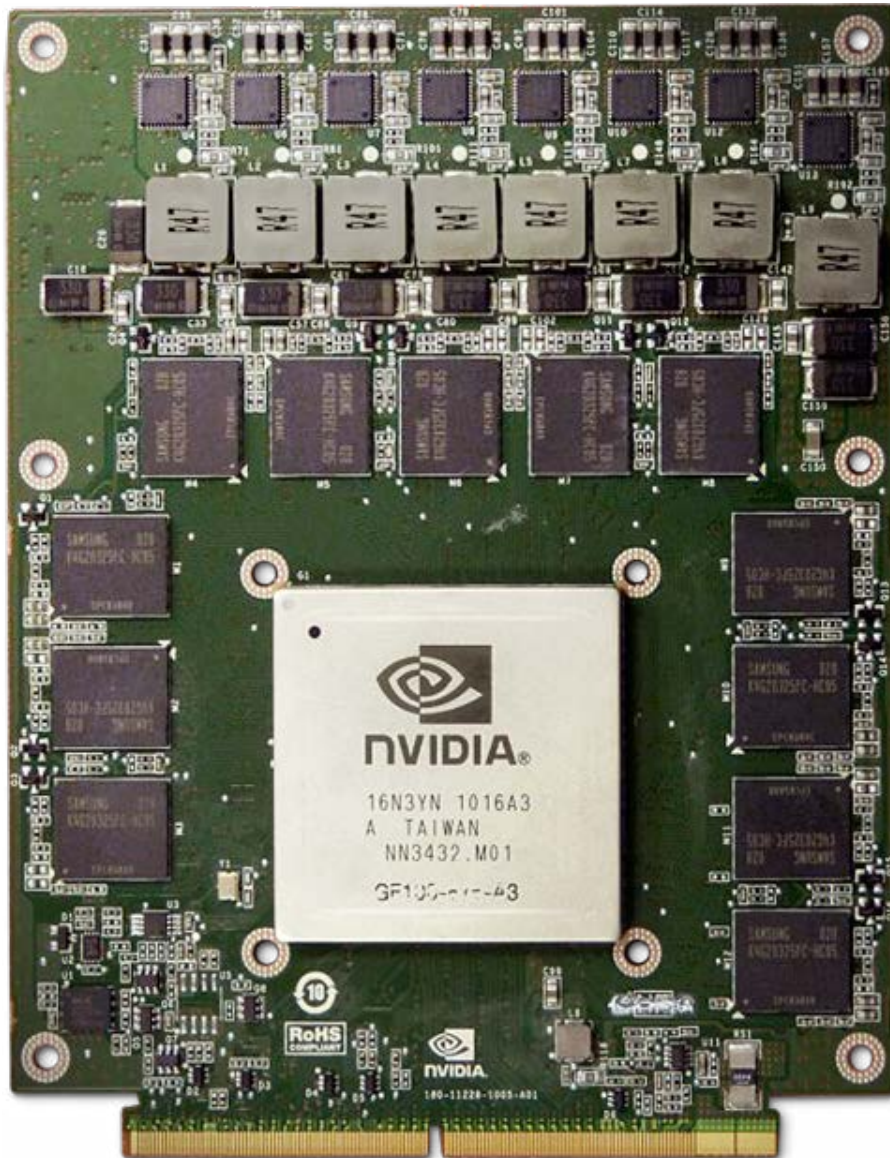


New K20x GPUs



GPU on Compute Board

GPU: Hyper-Parallel, Efficient & low power



Tesla K20x

- 14 Streaming Multiprocessors
- 2,688 CUDA cores
- 1.31 TFLOP/s peak (DP)
- 6 GB GDDR5 memory
- HPL: 2.0 GF/w (full system)



Titan Power & Cooling:

Designed for Efficiency

13,800v into building reduces transmission loss

480v to computers saved \$1M in installation & reduces losses

Liquid Cooling 1,000x more efficient than air

=> ORNL one of the world's most efficient data centers: PUE=1.25

Variable Speed Chillers save energy

Vapor barriers & positive air pressure remove humidity

UPS: highly-efficient Flywheel



Hybrid Programming Model

- *Jaguar's* 299,008 cores: at **limit** of MPI scaling
=> Hierarchical Parallelism needed for *Titan*
- Distributed memory: **MPI**, SHMEM, PGAS
Node Local: **OpenMP**, Pthreads, local MPI
In threads: GPU Vector constructs, libraries, **OpenACC**
- *Same constructs needed on **all** multi-PFLOPS computers to scale to full system size!*

How to program nodes?

- **Compilers**

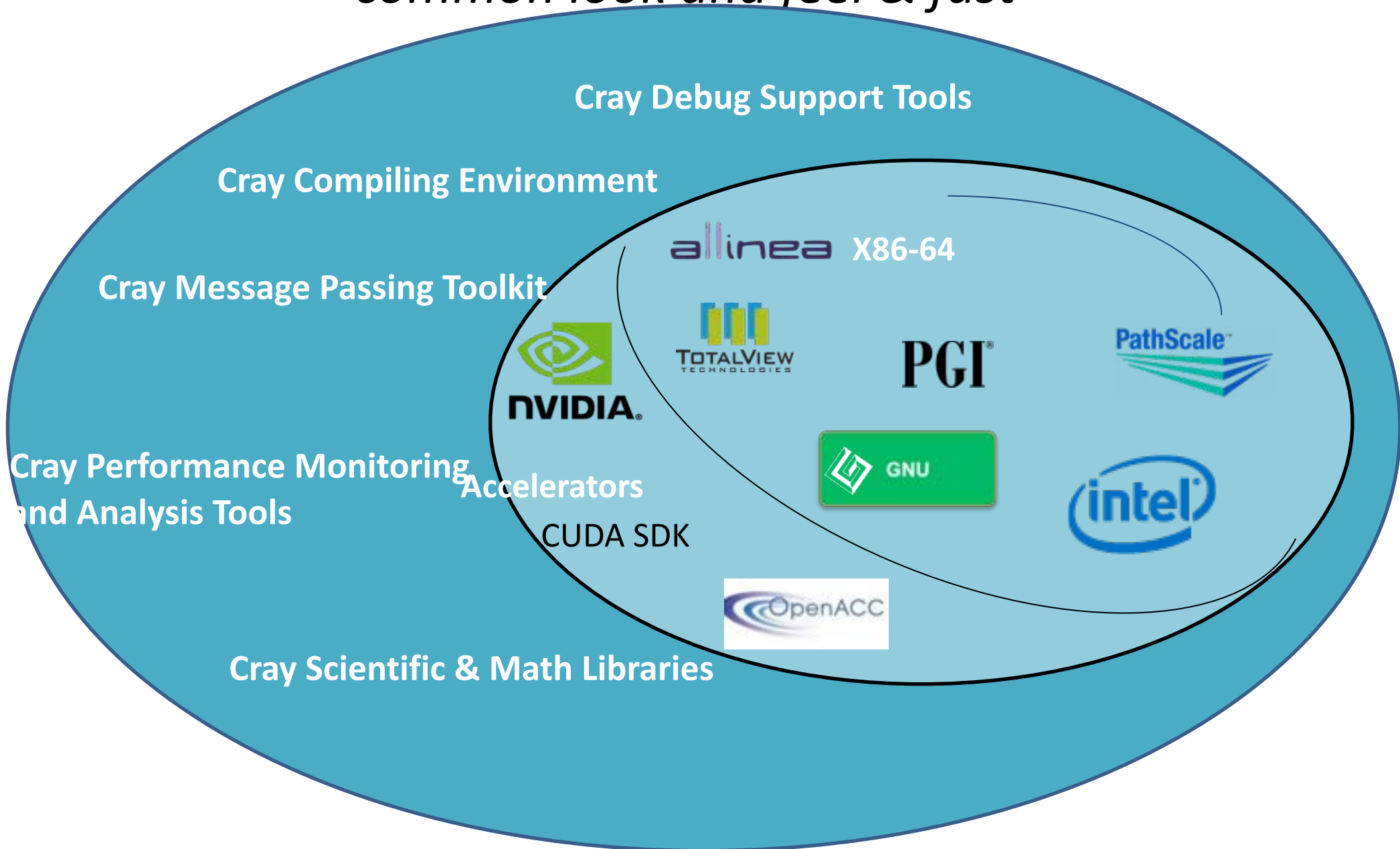
- **OpenACC**: users add `//` directives to source code => `//` code created for system: GPU, MIC, vector SIMD on CPU
- **Cray compiler** supports XK7 nodes & OpenACC
- **CAPS HMPP** compiles C, C++ & Fortran for heterogeneous nodes with OpenACC support
- **PGI** compiles OpenACC & CUDA Fortran

- **Tools**

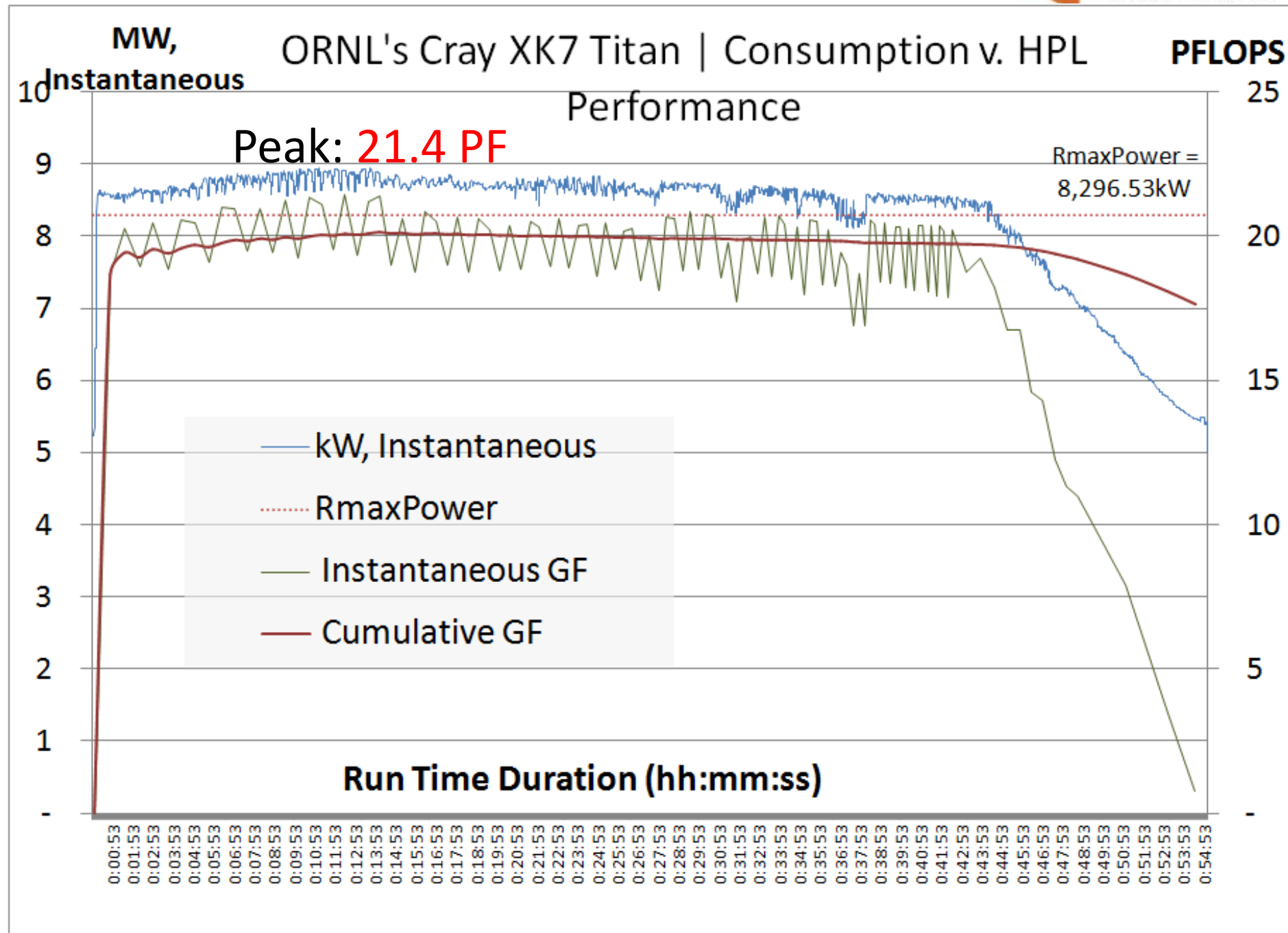
- **Allinea DDT debugger** scales to full system & debugs hybrid (x86/GPU) apps
- **TUD Vampir** profiles codes on hybrid nodes
- **CrayPAT** & Cray **Apprentice** support XK7 coding

Unified x86/Accelerator Development Environment

common look and feel & fast



ORNL's *Titan*: 17.59PF HPL Run

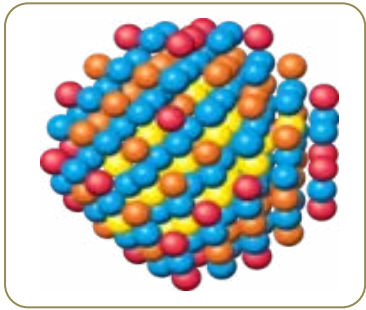


HPL Performance

- HPL took < 55 minutes: **I/O bound for 1st time**
 - Processors fast: hard to fully overlap communications with computation to keep all GPUs running.
- Used only NVIDIA K20x GPUs & its 6GB memory
 - Later extend hybrid code to use CPU & CPU's 32GB memory for better results
- Titan **10x faster** than Jaguar (when #1, Nov. 2009) & used only **19.4% more power**

Top500 List	System	Performance (PFLOPS)	Power (MW)
Nov. 2009	<i>Jaguar</i>	1.759	7.0
Nov. 2012	<i>Titan</i>	17.59	8.3

Early Science Applications on *Titan*



Material Science (WL-LSMS)

Role of material disorder, statistics & fluctuations in nanoscale materials & systems.



Climate Change (CAM-SE)

Realistic climate change adaptation & mitigation scenarios: precipitation & tropical storm patterns/statistics.

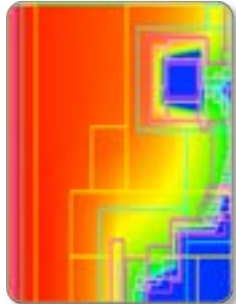


Biofuels (LAMMPS)

Multiple function molecular dynamics

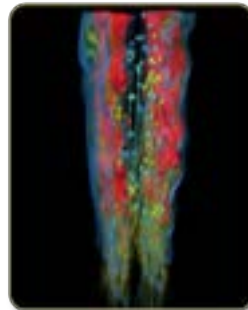
Astrophysics (NRDF)

Radiation transport for astrophysics, laser fusion, combustion, atmospheric dynamics & medical imaging.



Combustion (S3D)

Combustion simulations to enable next generation diesel/bio-fuels to burn more efficiently.



Nuclear Energy (Denovo)

Calculate Radiation transport for nuclear energy & technology apps.



GPUs Effective on Scalable Applications?

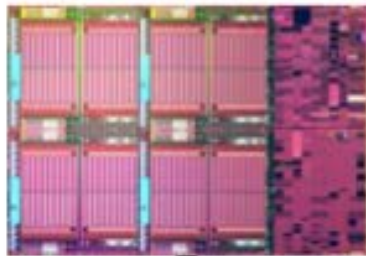
OLCF-3 Early Science Codes: **Very Early Titan performance measurements**

	XK7 (w/ K20x) vs. XE6	Cray XK7: K20x GPU plus AMD 6274 CPU Cray XE6: Dual AMD 6274 and no GPU
Application	Performance Ratio	Comments
S3D	1.8	<ul style="list-style-type: none">• Turbulent combustion• 6% of Jaguar workload
Denovo sweep	3.8	<ul style="list-style-type: none">• Sweep kernel of 3D neutron transport for nuclear reactors• 2% of Jaguar workload
LAMMPS	7.4* (mixed precision)	<ul style="list-style-type: none">• High-performance molecular dynamics• 1% of Jaguar workload
WL-LSMS	3.8	<ul style="list-style-type: none">• Statistical mechanics of magnetic materials• 2% of Jaguar workload• 2009 Gordon Bell Winner
CAM-SE	1.8* (estimate)	<ul style="list-style-type: none">• Community atmosphere model• 1% of Jaguar workload

Commodity plus Accelerator Today

Commodity

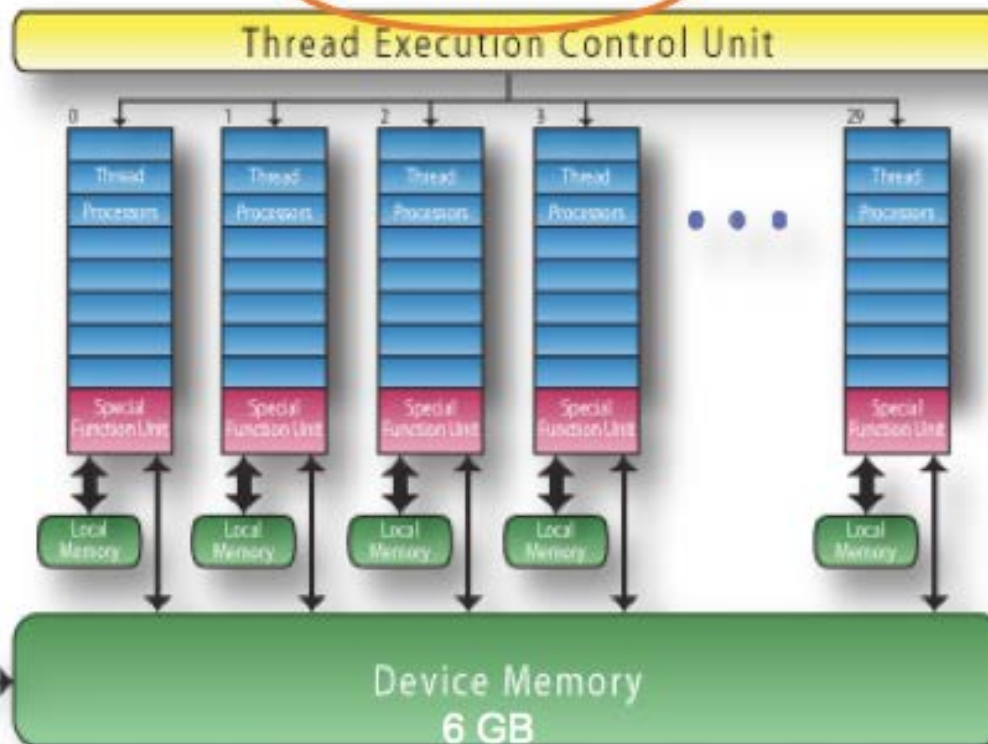
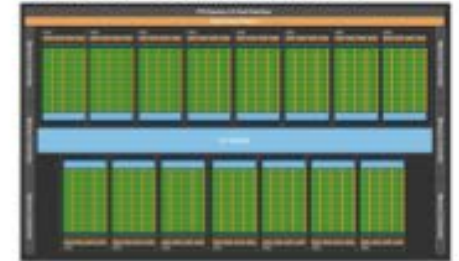
Intel Xeon
8 cores
3 GHz
8*4 ops/cycle
96 Gflop/s (DP)



Accelerator (GPU)

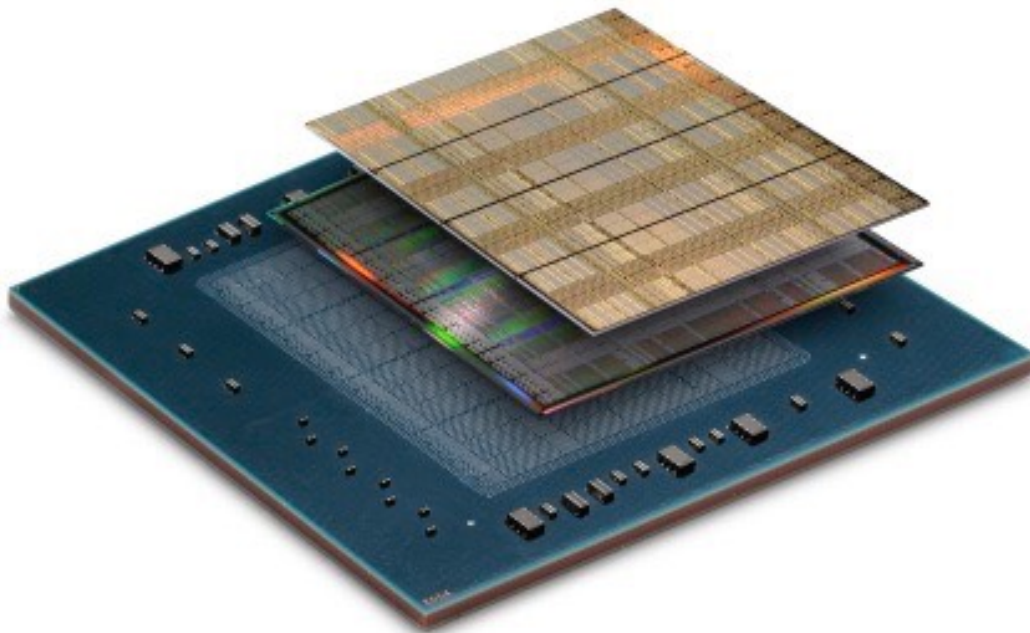
Nvidia K20X "Kepler"
2688 "Cuda cores"
.732 GHz
2688*2/3 ops/cycle
1.31 Tflop/s (DP)

192 Cuda cores/SMX
2688 "Cuda cores"



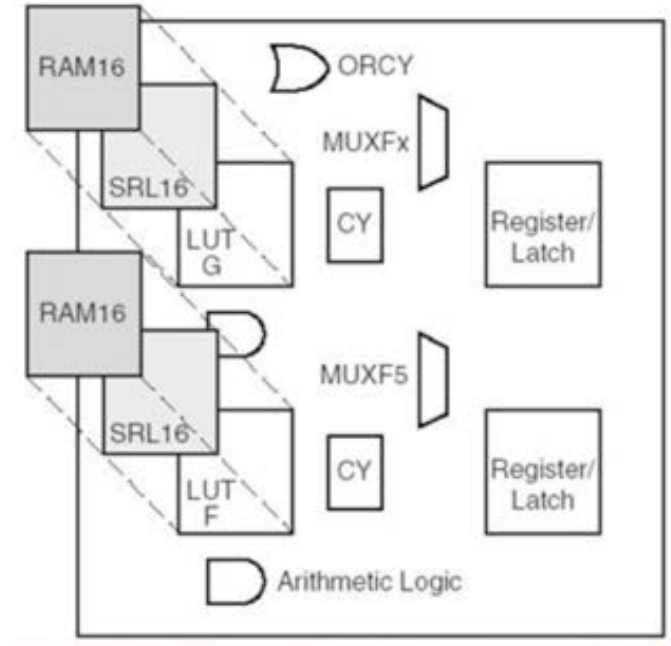
Interconnect
PCI-X 16 lane
64 Gb/s (8 GB/s)
1 GW/s

FPGA *custom chip* also advancing!



Xilinx Virtex7 FPGA:

- Logic array: user-tailored to application
- 2M logic cells. 4K slices, 2.8 Tb/s bandwidth
- On-chip RAM, multipliers & CPUs
- 100–1000 operations/clock cycle



FPGA Logic slice

FPGA Pros & Cons

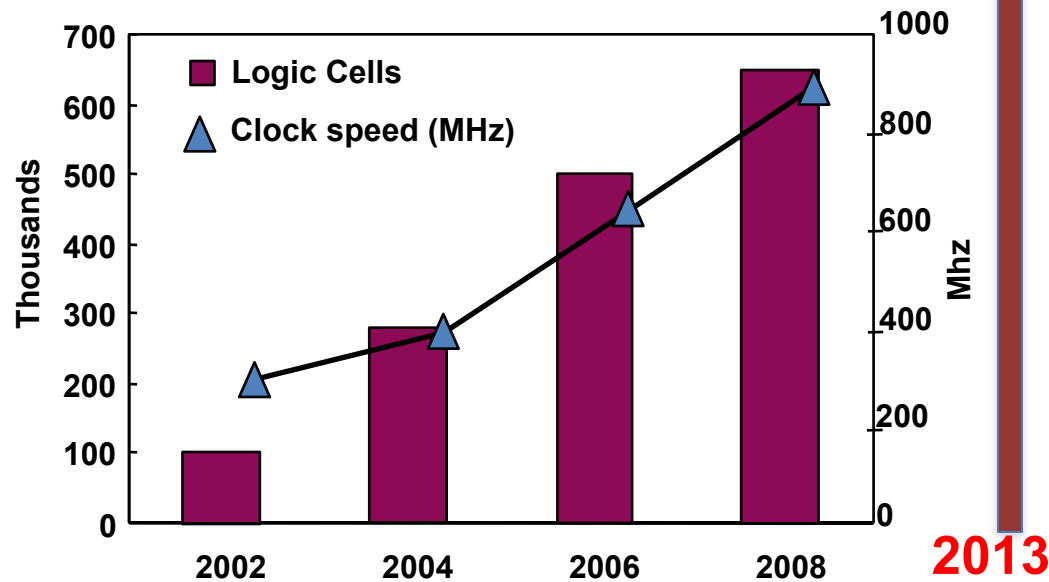
Pros

- **Performance:** optimal silicon use (most parallel ops/cycle)
- **Rapid growth:** Cells, Speed, I/O
- **Power:** < CPUs & GPUs
- **Flexible:** *tailor* to application
- **Advances:** Telecom spinoff

2M Cells

Cons

- Access:** GPUs in PCs, free Cuda
- Coding:** esoteric, \$\$ (VHDL...)
- Compile times** long

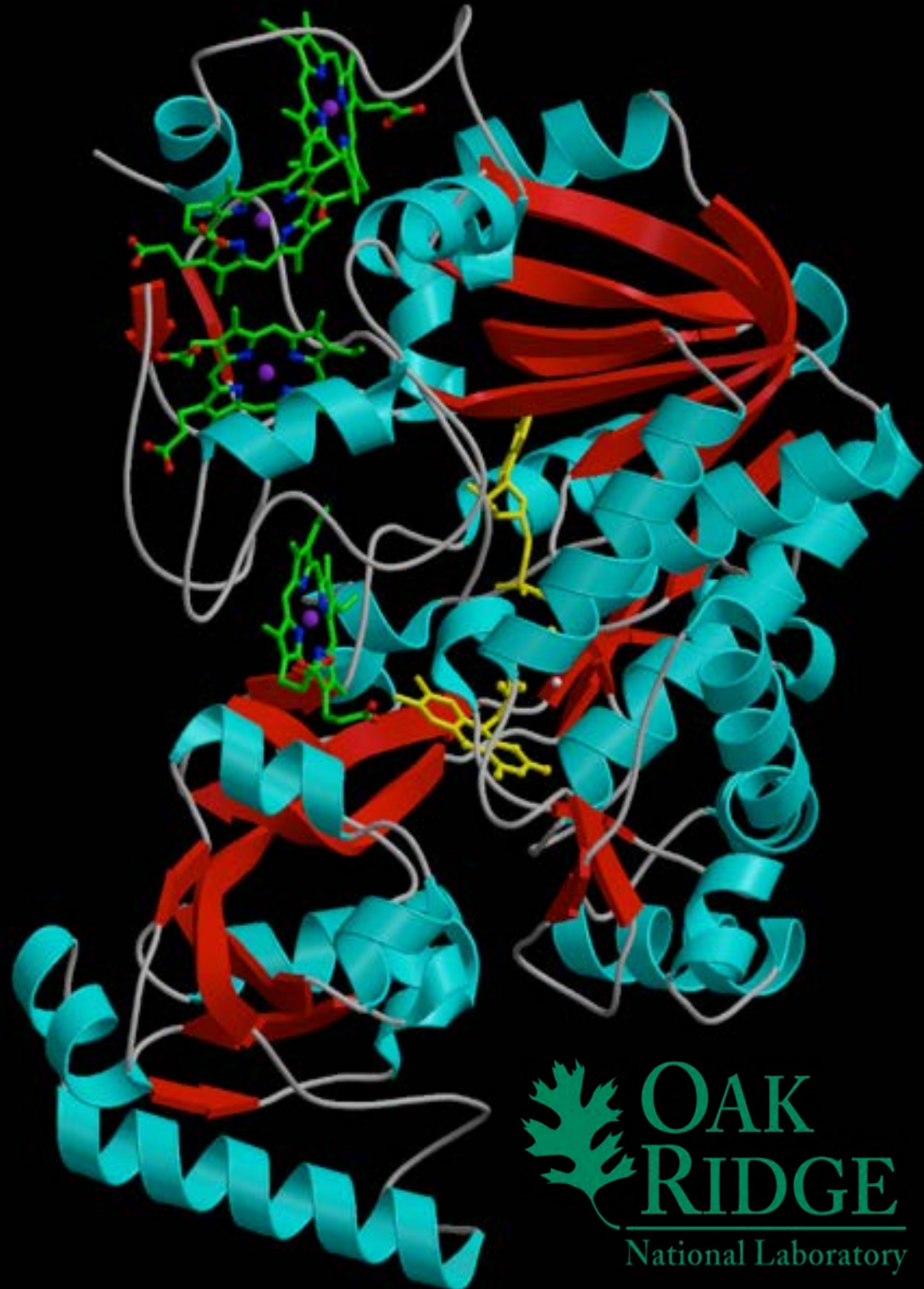
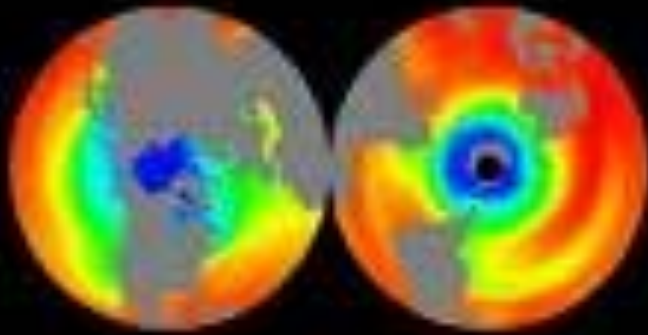
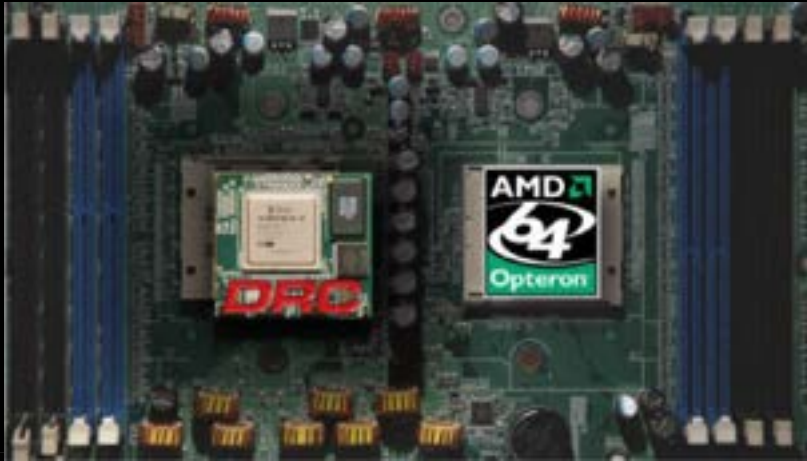


Applications

- Genomics
- Matrix Equation Solution
- Molecular Dynamics, Weather/Climate



100x Genomics Speedup/FPGA for up to 150 FPGAs

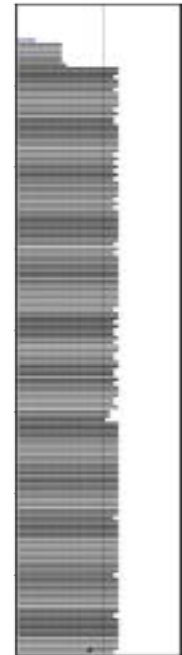
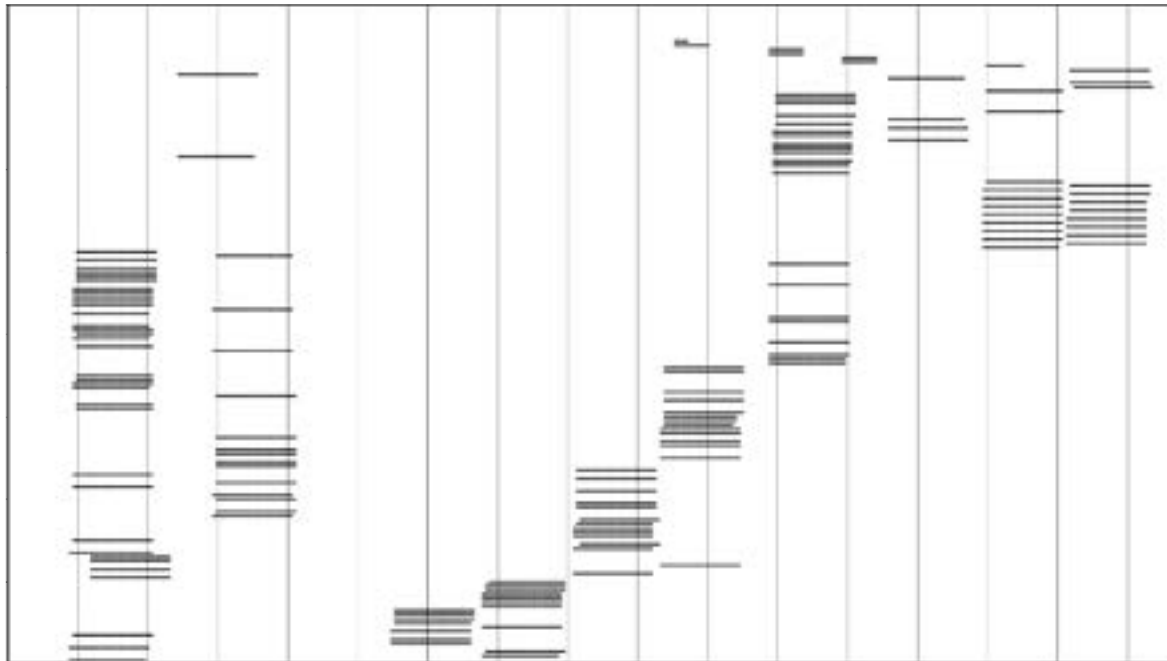


DNA Sequencing* Time on 150 FPGAs

* Human-Mouse DNA Compare (FASTA)

“Non-dedicated” FPGAs

Dedicated FPGAs



Search Time for 150 FPGAs (days)

***FPGA
Jobs***

Speedup on 150 FPGAs*

1 Opteron ==> **20 years** (240 mos)

1 FPGA ==> **5 months**

150 Opterons ==> **6 weeks**

150 FPGAs ==> **1 day** ==> 49X speedup (VirtexII)

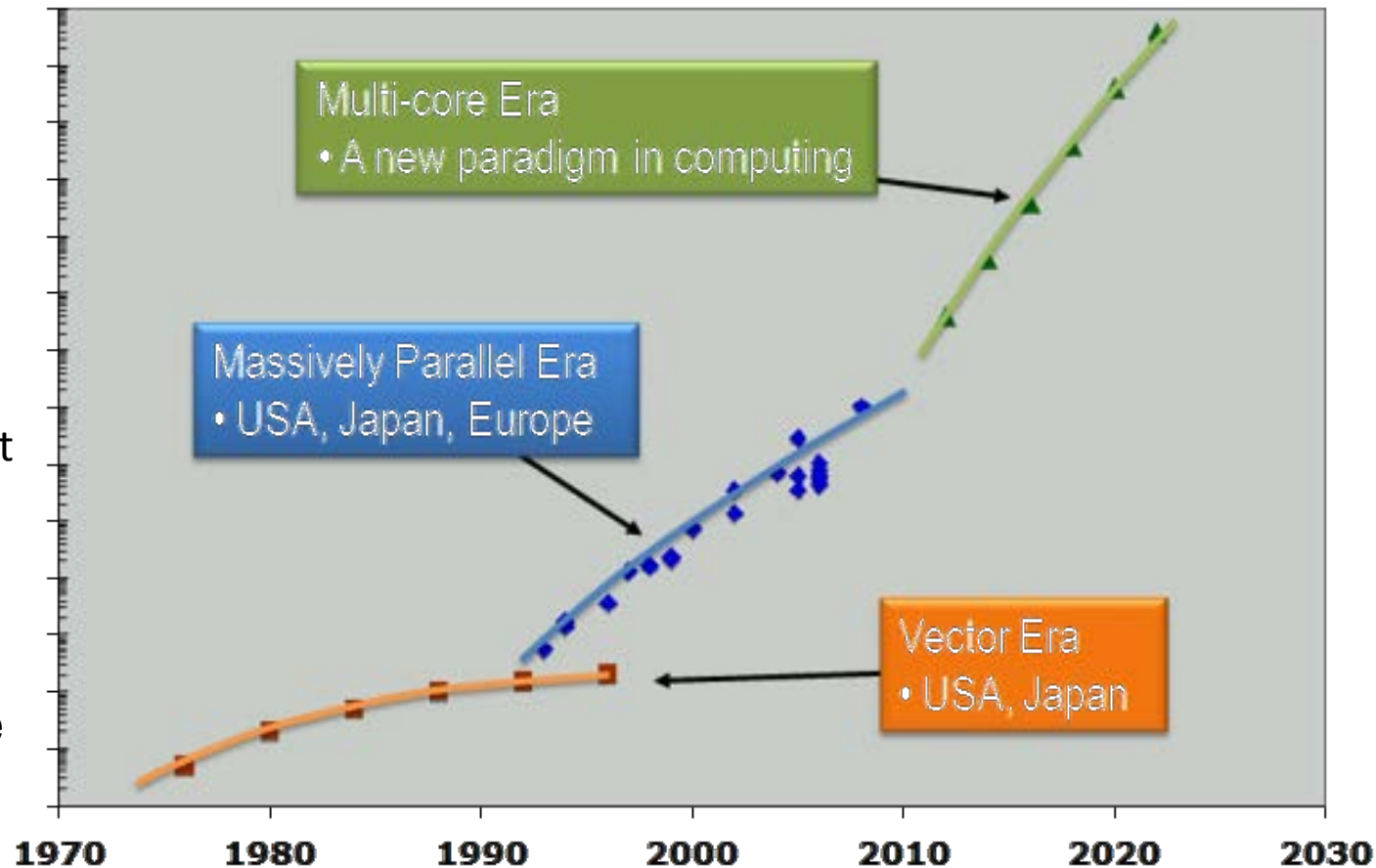
==> 7,350X faster than 1 Opteron (VirtexIIs)

==> 14,700X faster than 1 Opteron (Virtex4s)

***Compared to one 2.2 GHz Opteron**

Future HPC: Application challenges

- Huge increase in // **nodes**
 - 10 to 100× by 2015
 - 100 to 1000× by 2018
- Node increase => **lower MTTI**
- More **memory levels**
 - Algorithms must prioritize data movement over cycles
- **// Apps** to optimize data flow
- Programming **models & tools** are immature: in a state of flux

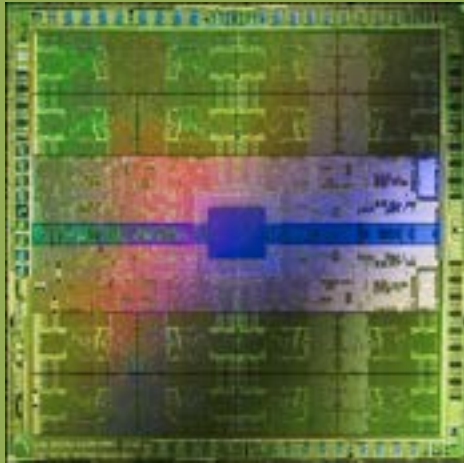


~~Desktop~~ Future HPC application challenges

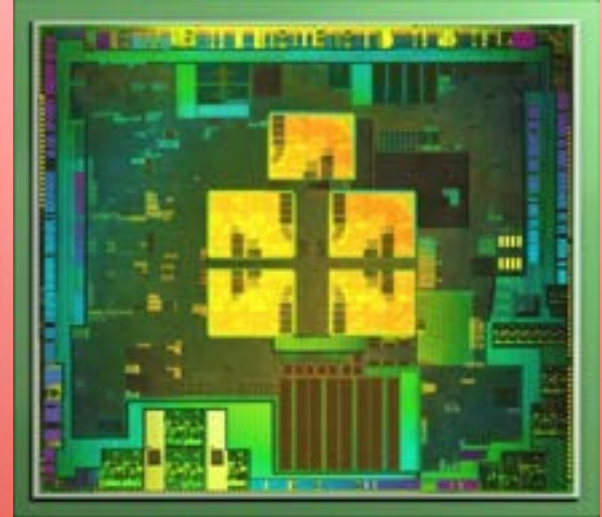
- Huge **node increase**
'15: 10-100x '18: 100-1000x
Increase => lower MTTF
- More **memory levels** => apps to prioritize data movement/cycle
- Current Programming **models & tools** immature



2010: Intel exp 48-core chip shipped



NVIDIA 512-"core"
Fermi GPU

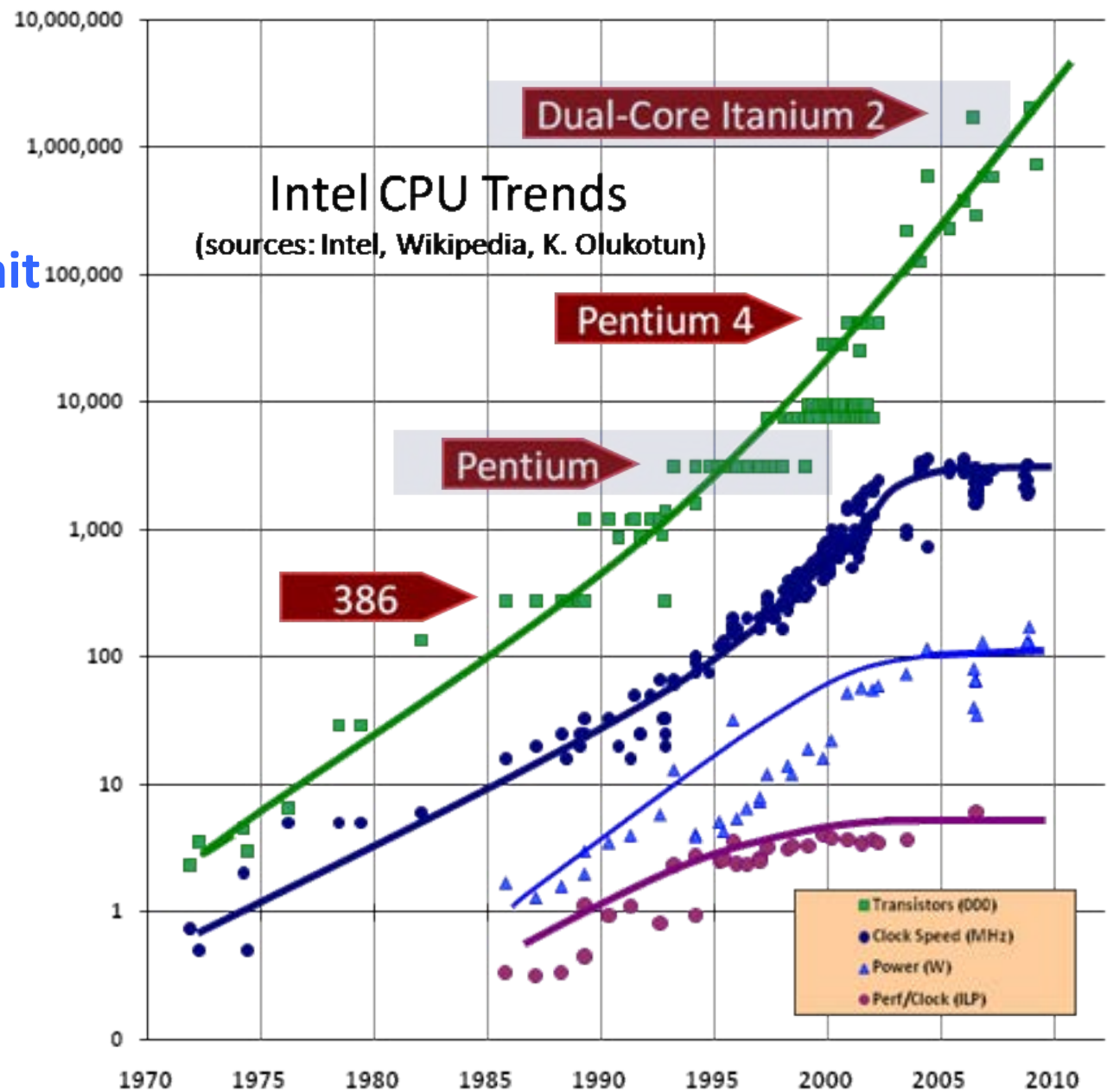


NVIDIA Tegra 3

mobile devices in next HPC system at Barcelona Supercomputing Center

Architectural Trends – No more free lunch

- **Moore's Law continues**
- **CPU clock speed increase ends in 2003 \leq power limit** (cooling & \$).
- Performance via //ism

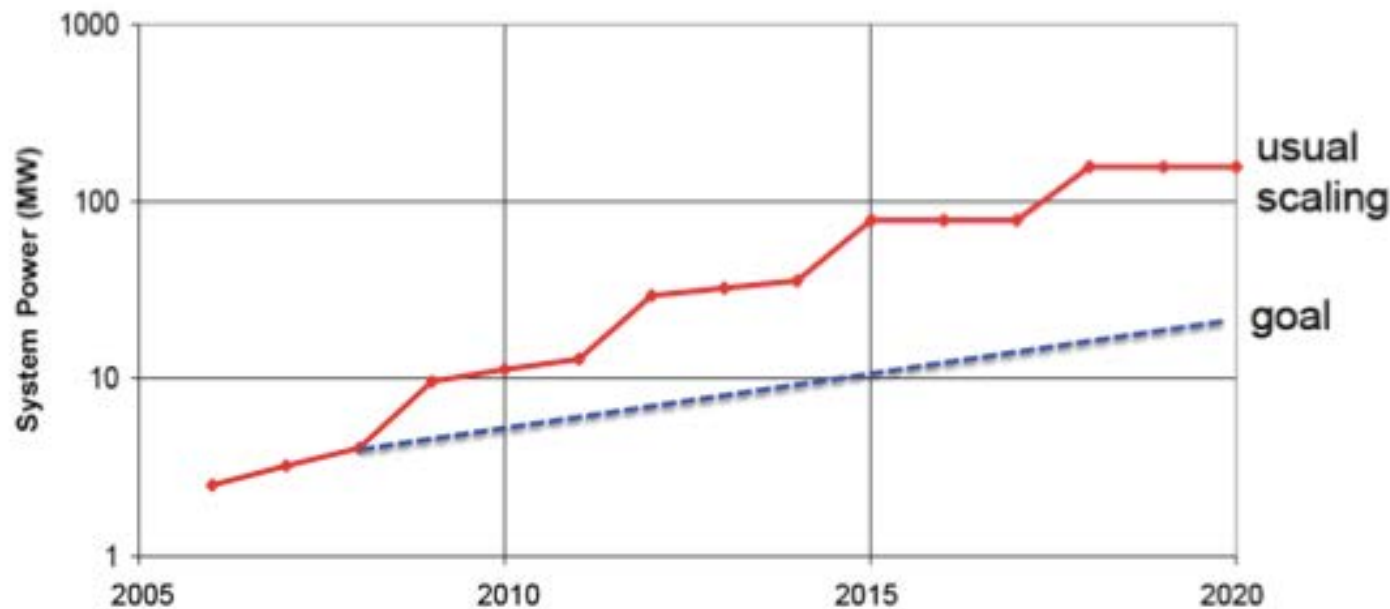


Herb Sutter: Dr. Dobb's Journal:

<http://www.gotw.ca/publications/concurrency-ddj.htm>

Energy Cost Challenge

- **At ~\$1M per MW energy costs are substantial**
 - **10 Pflop/s in 2011 uses ~10 MWs**
 - **1 Eflop/s in 2018 > 100 MWs**



- **DOE Target: 1 Eflop/s around 2020-2022 at 20 MWs**



Potential System Architecture with a cap of \$200M and 20MW

Systems	2013 Titan Computer	2020	Difference Today & 2020
System peak	27 Pflop/s	1 Eflop/s	O(100)
Power	8.3 MW (2 Gflops/W)	~20 MW (50 Gflops/W)	O(10)
System memory	710 TB (38*18688)	32 - 64 PB	O(100)
Node performance	1,452 GF/s (1311*141)	1.2 or 15TF/s	O(10)
Node memory BW	232 GB/s (52*180)	2 - 4TB/s	O(10)
Node concurrency	16 cores CPU 2688 CUDA cores	O(1k) or 10k	O(100) - O(10)
Total Node Interconnect BW	8 GB/s	200-400GB/s	O(100)
System size (nodes)	18,688	O(100,000) or O(1M)	O(10) - O(100)
Total concurrency	50 M	O(billion)	O(100)
MTTF	?? unknown	O(<1 day)	O(?)

Summary

- **Huge HPC advances:** Vector \Rightarrow MP \Rightarrow Multi-core
 - Past: Vector \Rightarrow Multi-processors (MP) 1Mx/decade
 - Present: MP \Rightarrow Multi-core + emerging accelerators
 - Future: Accelerators: GPUs & FPGAs (special Apps)
- **Major Exascale Challenges & great promise**
 - Programming Parallel Nodes $O(10^9)$
 - Hybrid: most Apps perform worse than HPL?
 - Power: 2 GF/w (today) \Rightarrow 50 GF/w (stretch)
 - Fault Tolerance: BG/Q has 1.25 MTTI
 - Success Likely: Strong support, innovative teams

Contact

Google

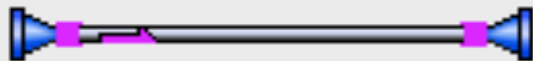


Olaf Storaasli

Email: Olaf@cox.net
Olaf@synective.se

Trevligt att vara här!
Tack allesammans!

Question



Answer