An Overview of High Performance Computing and Future Requirements

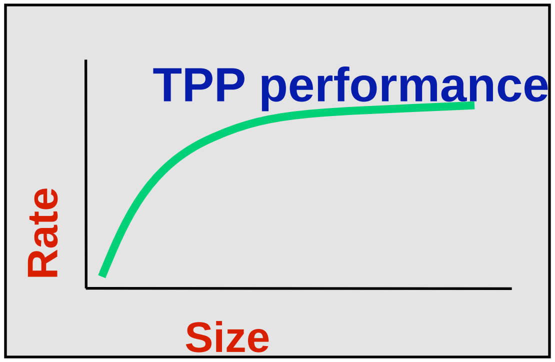
Jack Dongarra

University of Tennessee
Oak Ridge National Laboratory

H. Meuer, H. Simon, E. Strohmaier, & JD

- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

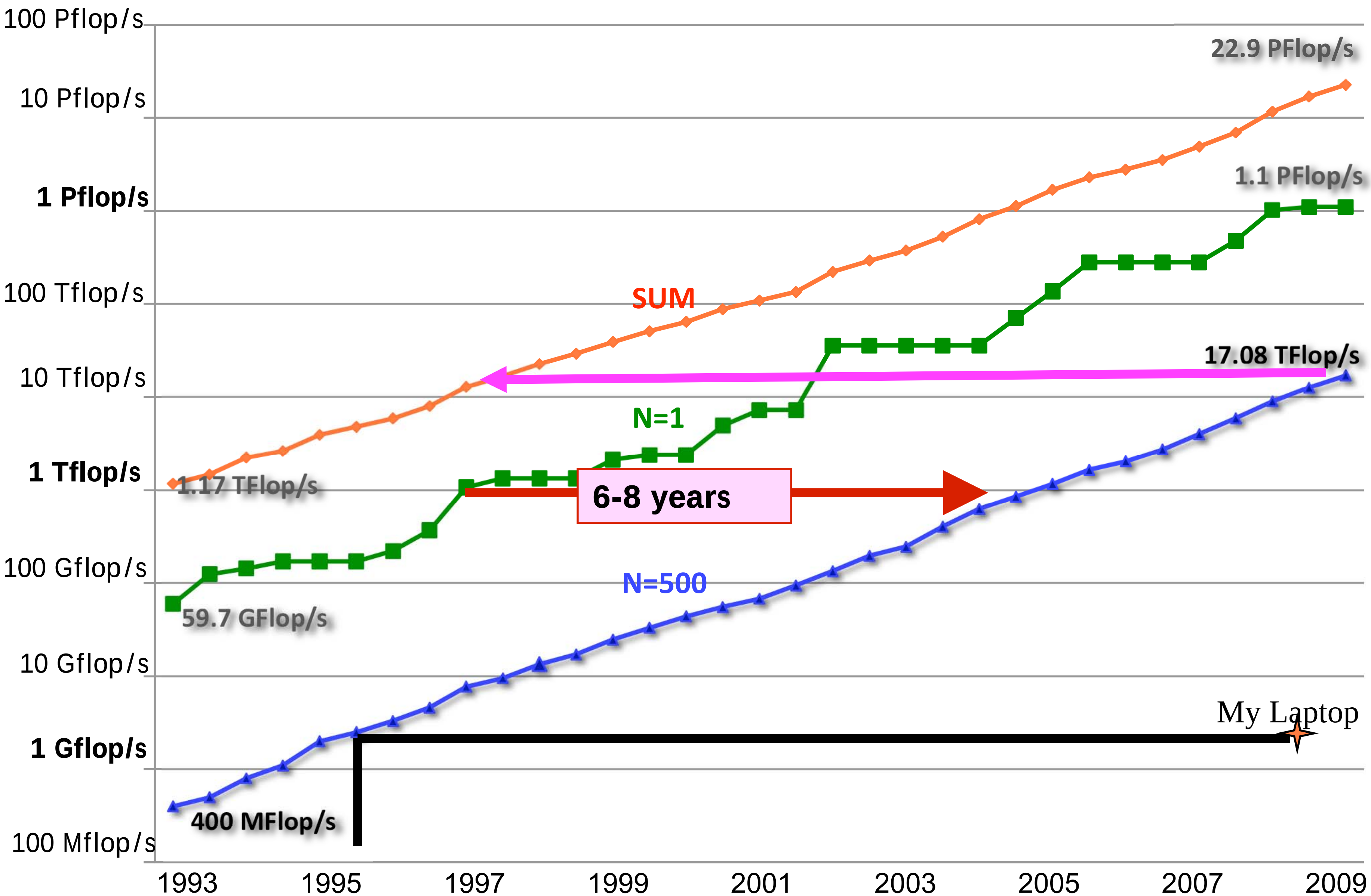
  Ax=b, dense problem



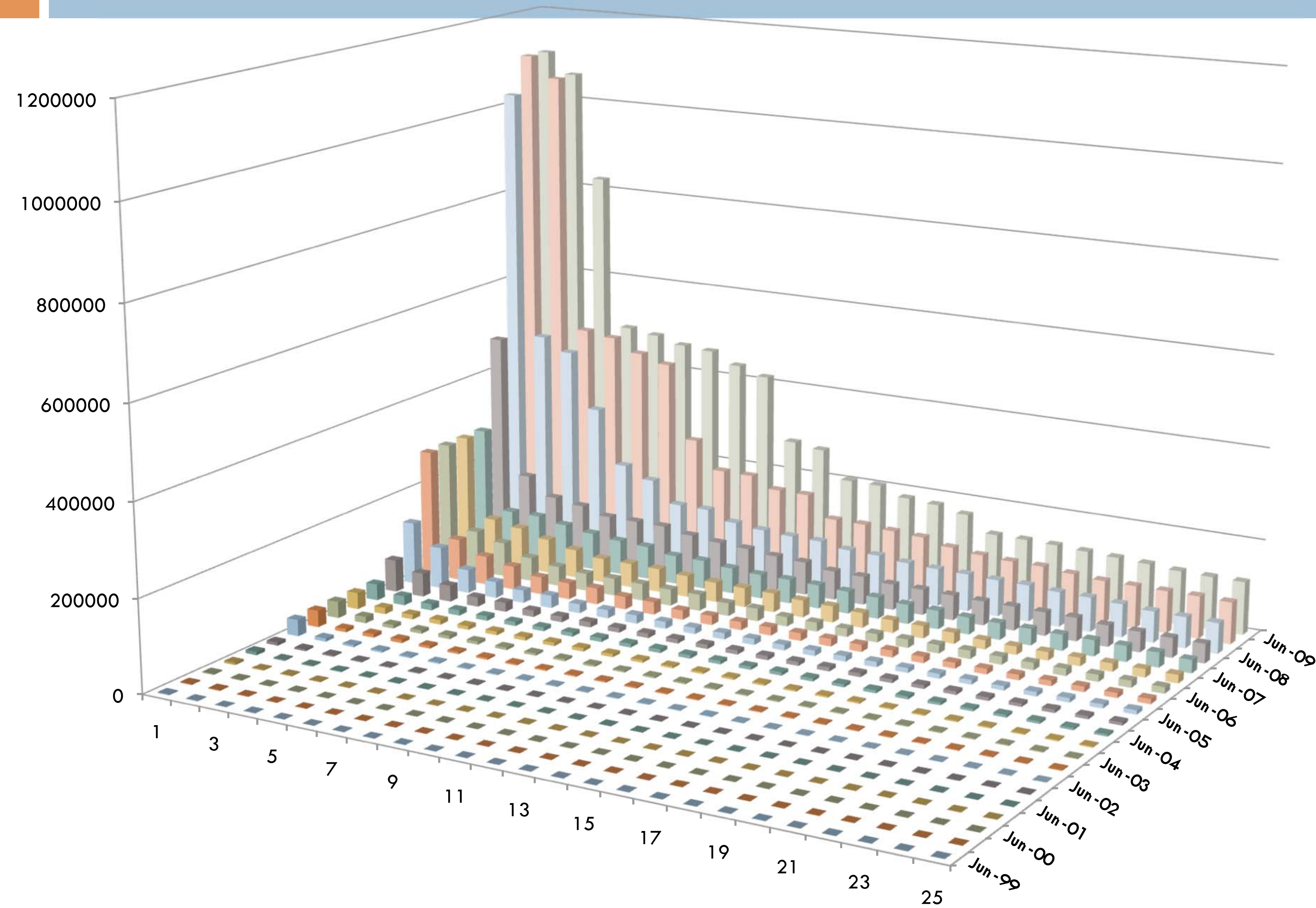
- Updated twice a year
  SC'xy in the States in November
  Meeting in Germany in June

- All data available from **www.top500.org**

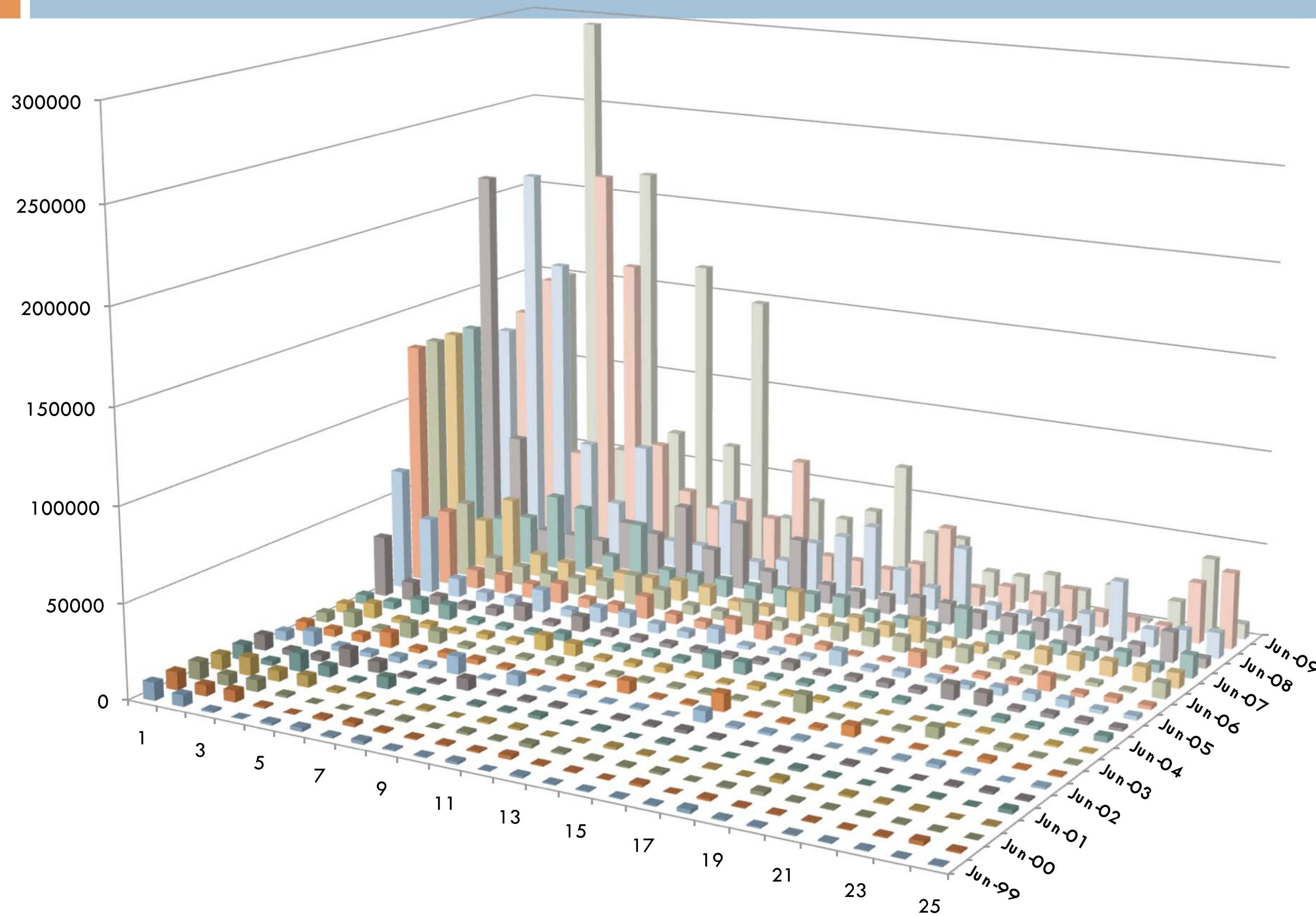
Performance Development



Performance of Top25 Over 10 Years



Cores in the Top25 Over Last 10 Years

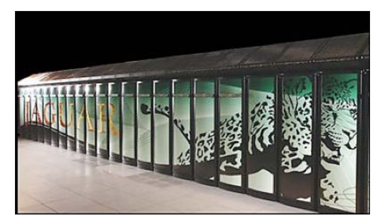


Looking at the Gordon Bell Prize

(Recognize outstanding achievement in high-performance computing applications and encourage development of parallel processing)

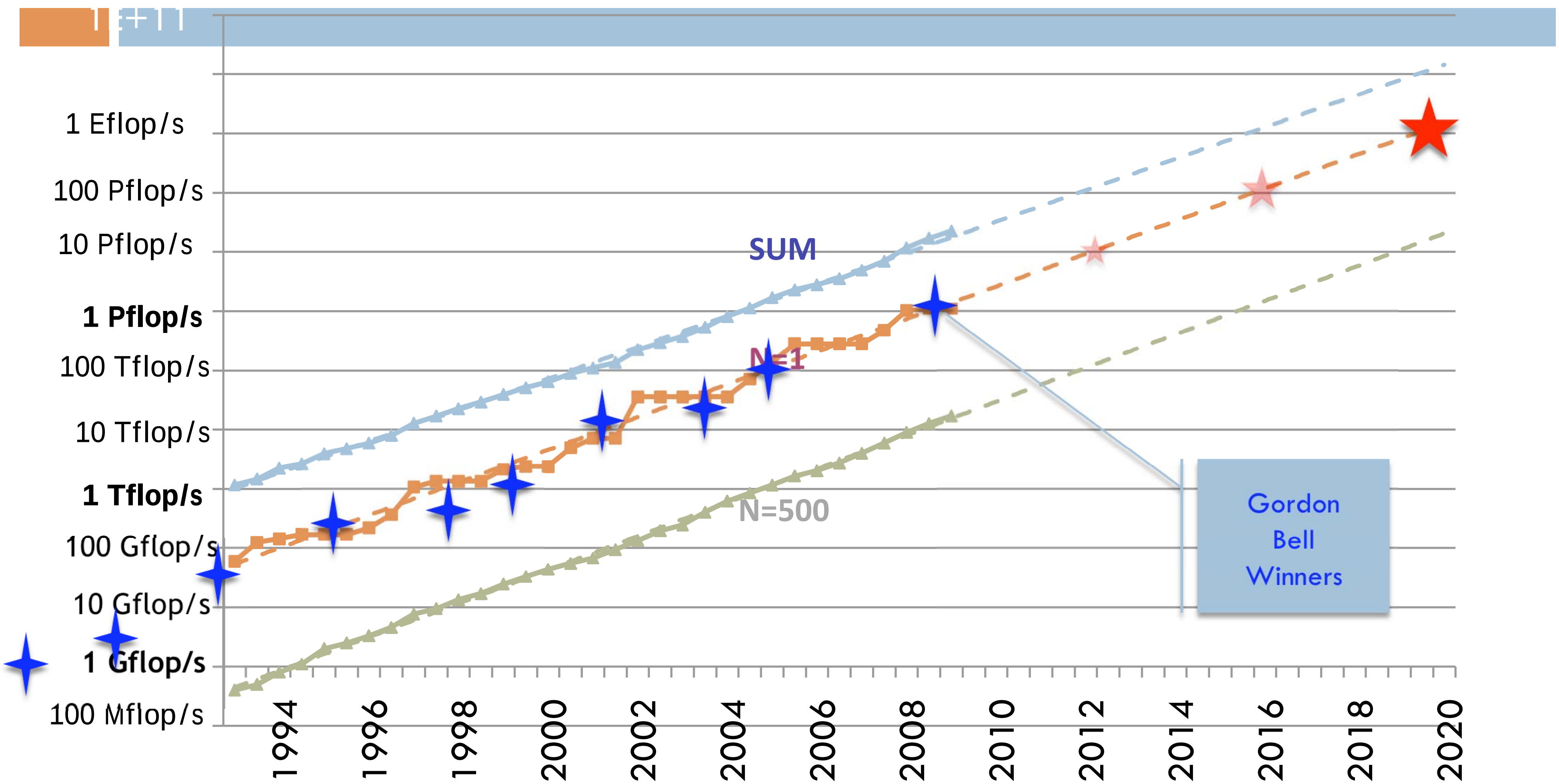
- 1 GFlop/s; 1988; Cray Y-MP; 8 Processors
  - Static finite element analysis
- 1 TFlop/s; 1998; Cray T3E; 1024 Processors
  - Modeling of metallic magnet atoms, using a variation of the locally self-consistent multiple scattering method.

- 1 PFlop/s; 2008; Cray XT5; 1.5×10^5 Processors
  - Superconductive materials

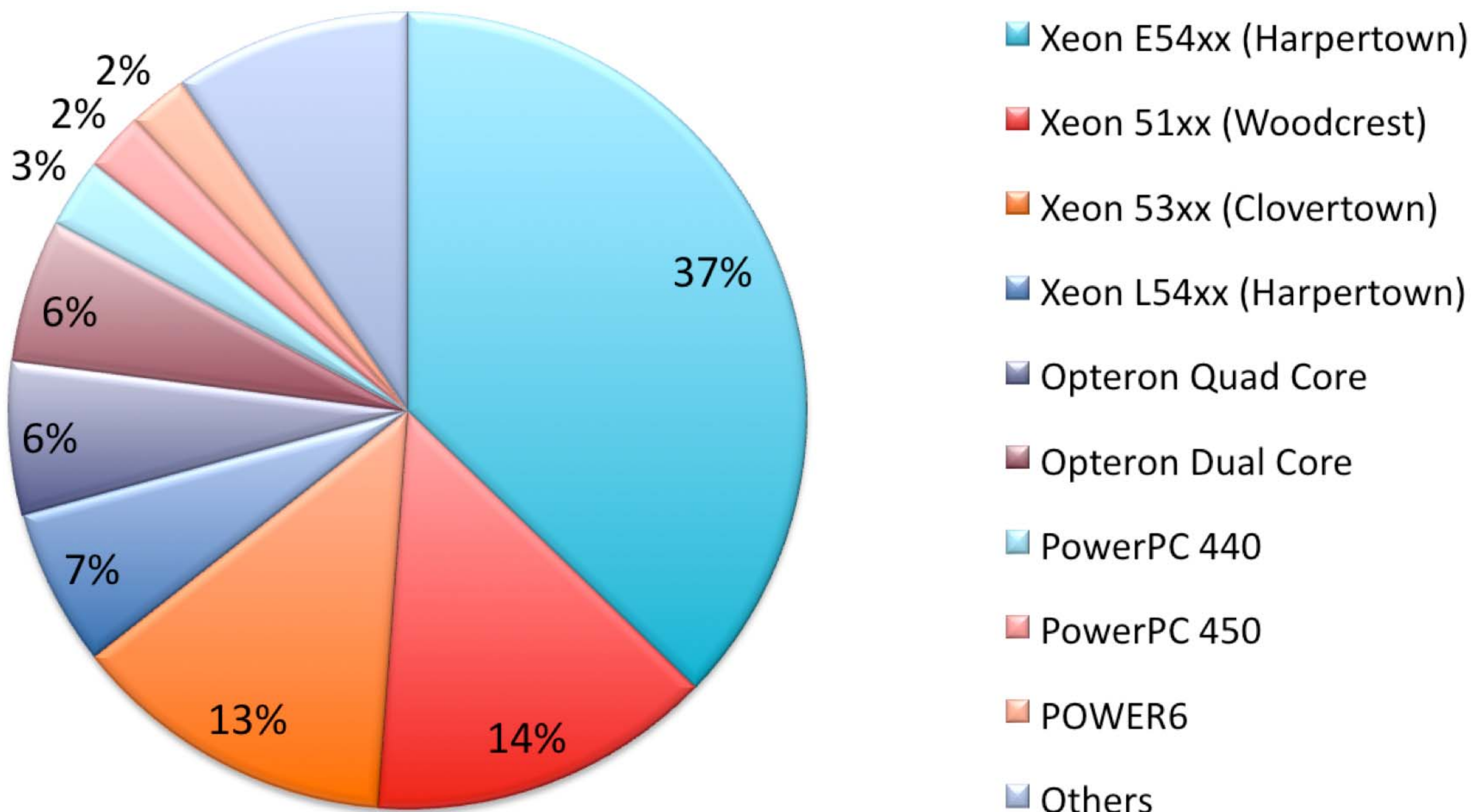


- 1 EFlop/s; ~2018; ?; 1×10^7 Processors (10^9 threads)

Performance Development in Top500

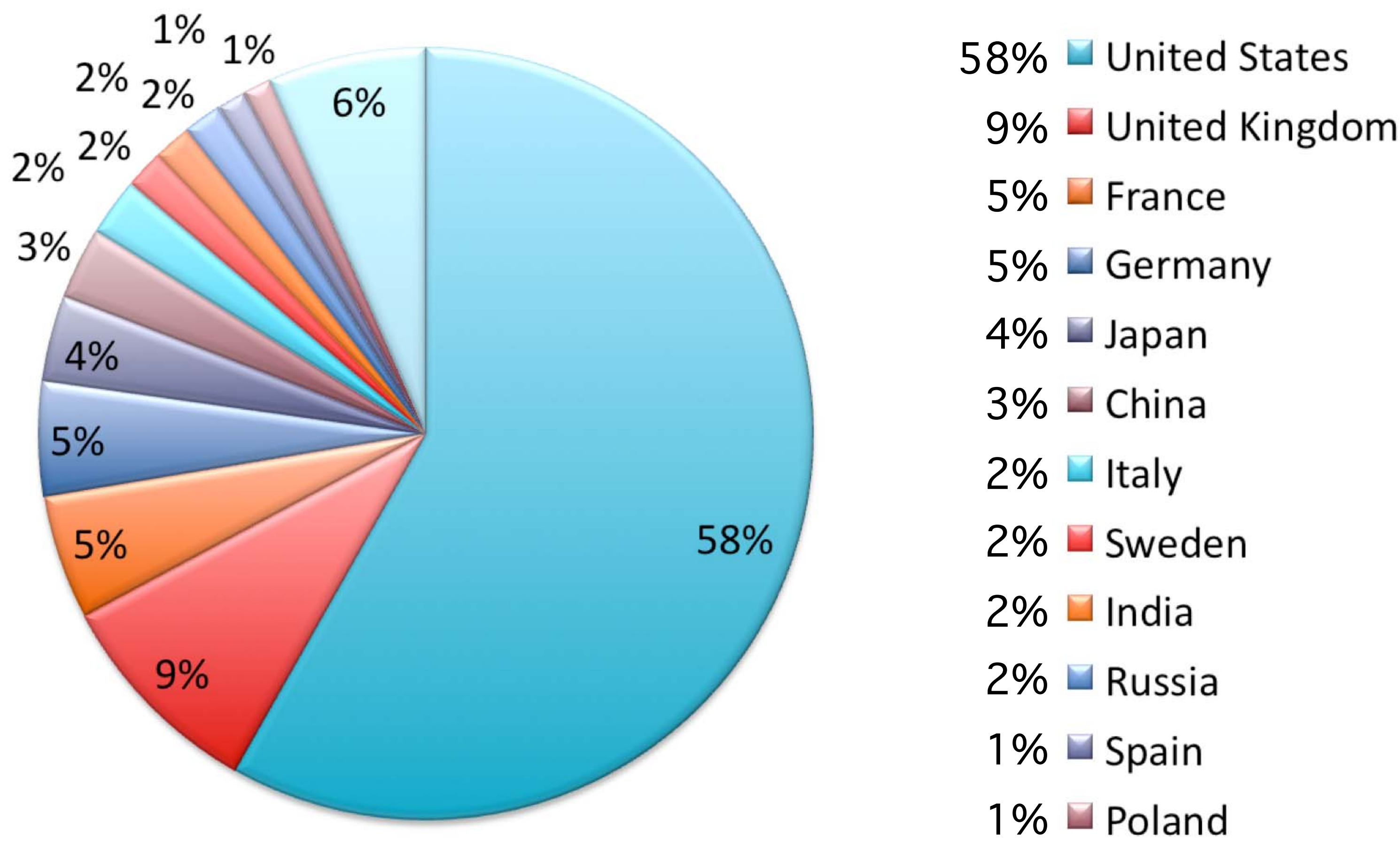


Processors Used in Supercomputers

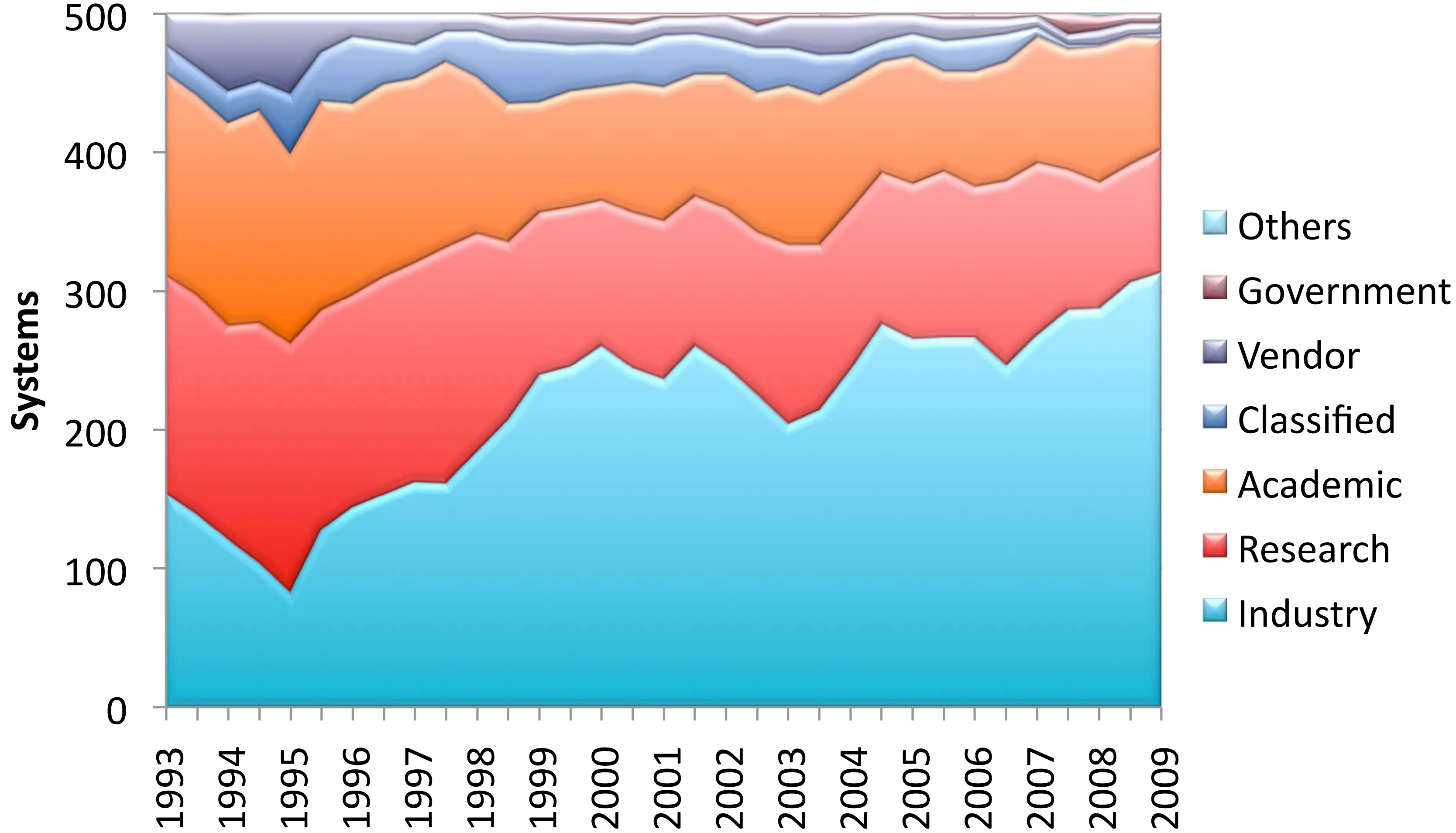


Intel 71%
AMD 13%
IBM 7%

Countries / System Share



Customer Segments



Industrial Use of Supercomputers

- **Of the 500 Fastest Supercomputer**
  - **Worldwide, Industrial Use is > 60%**

- Aerospace
- Automotive
- Biology
- CFD
- Database
- Defense
- Digital Content Creation
- Digital Media
- Electronics
- Energy
- Environment
- Finance
- Gaming
- Geophysics
- Image Proc./Rendering
- Information Processing Service
- Information Service
- Life Science
- Media
- Medicine
- Pharmaceutics
- Research
- Retail
- Semiconductor
- Telecomm
- Weather and Climate Research
- Weather Forecasting



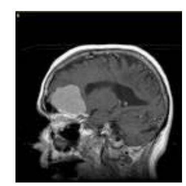




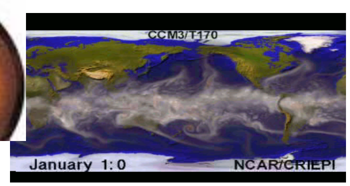












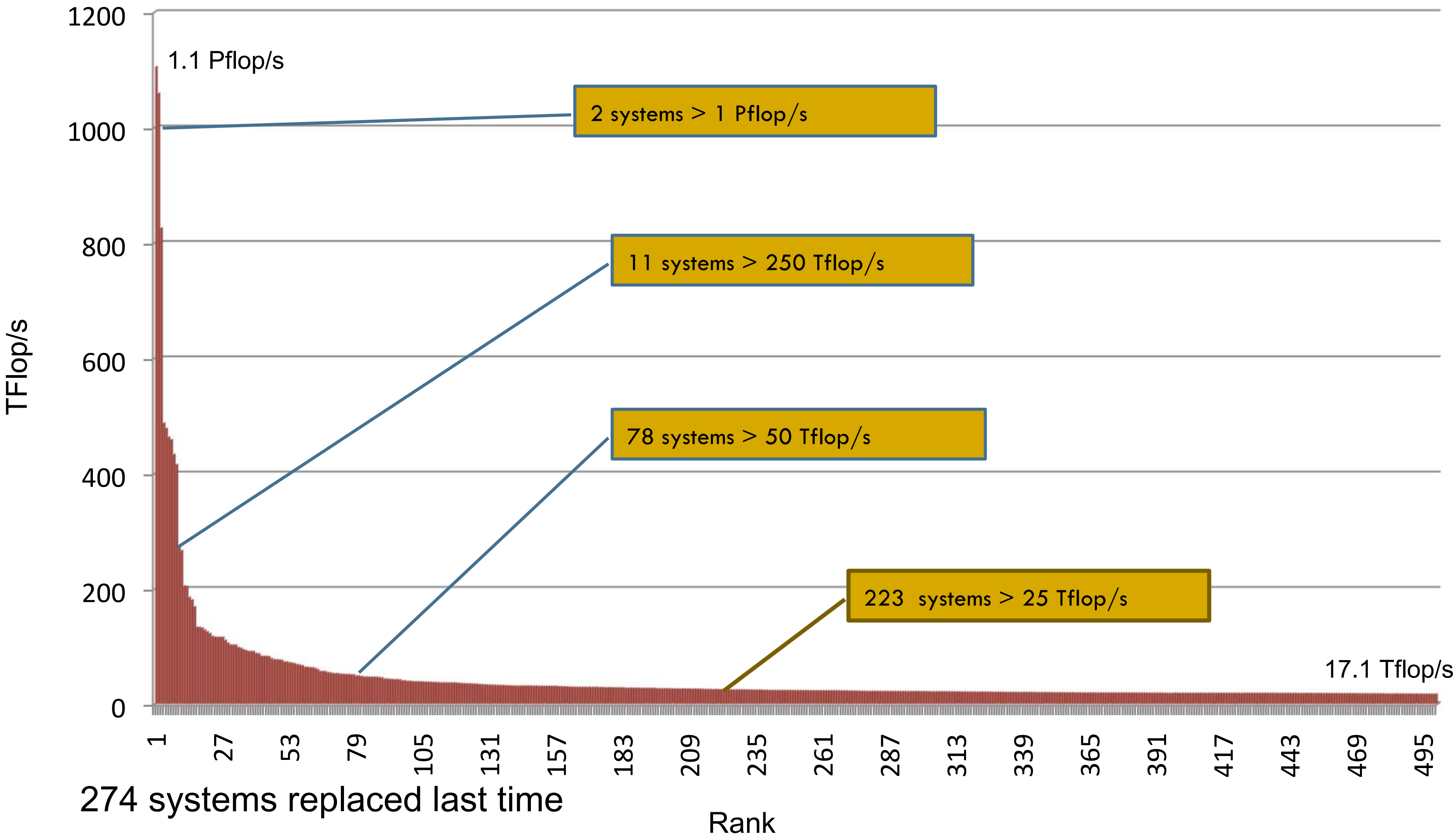

33rd List: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Tflops]	% of Peak
1	DOE / NNSA Los Alamos Nat Lab	Roadrunner / IBM BladeCenter QS22/LS21	USA	129,600	1,105	76
2	DOE / OS Oak Ridge Nat Lab	Jaguar / Cray Cray XT5 QC 2.3 GHz	USA	150,152	1,059	77
3	Forschungszentrum Juelich (FZJ)	Jugene / IBM Blue Gene/P Solution	Germany	294,912	825	82
4	NASA / Ames Research Center/NAS	Pleiades / SGI SGI Altix ICE 8200EX	USA	51,200	480	79
5	DOE / NNSA Lawrence Livermore NL	BlueGene/L IBM eServer Blue Gene Solution	USA	212,992	478	80
6	NSF NICS/U of Tennessee	Kraken / Cray Cray XT5 QC 2.3 GHz	USA	66,000	463	76
7	DOE / OS Argonne Nat Lab	Intrepid / IBM Blue Gene/P Solution	USA	163,840	458	82
8	NSF TACC/U. of Texas	Ranger / Sun SunBlade x6420	USA	62,976	433	75
9	DOE / NNSA Lawrence Livermore NL	Dawn / IBM Blue Gene/P Solution	USA	147,456	415	83
10	Forschungszentrum Juelich (FZJ)	JUROPA /Sun - Bull SA NovaScale /Sun Blade	Germany	26,304	274	89

33rd List: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Tflops]	% of Peak	Power [MW]	Flops/ Watt
1	DOE / NNSA Los Alamos Nat Lab	Roadrunner / IBM BladeCenter QS22/LS21	USA	129,600	1,105	76	2.48	446
2	DOE / OS Oak Ridge Nat Lab	Jaguar / Cray Cray XT5 QC 2.3 GHz	USA	150,152	1,059	77	6.95	151
3	Forschungszentrum Juelich (FZJ)	Jugene / IBM Blue Gene/P Solution	Germany	294,912	825	82	2.26	365
4	NASA / Ames Research Center/NAS	Pleiades / SGI SGI Altix ICE 8200EX	USA	51,200	480	79	2.09	230
5	DOE / NNSA Lawrence Livermore NL	BlueGene/L IBM eServer Blue Gene Solution	USA	212,992	478	80	2.32	206
6	NSF NICS/U of Tennessee	Kraken / Cray Cray XT5 QC 2.3 GHz	USA	66,000	463	76		
7	DOE / OS Argonne Nat Lab	Intrepid / IBM Blue Gene/P Solution	USA	163,840	458	82	1.26	363
8	NSF TACC/U. of Texas	Ranger / Sun SunBlade x6420	USA	62,976	433	75	2.0	217
9	DOE / NNSA Lawrence Livermore NL	Dawn / IBM Blue Gene/P Solution	USA	147,456	415	83	1.13	367
10	Forschungszentrum Juelich (FZJ)	JUROPA /Sun - Bull SA NovaScale /Sun Blade	Germany	26,304	274	89	1.54	178

Distribution of the Top500



274 systems replaced last time

15 Systems on Top 500 in Japan

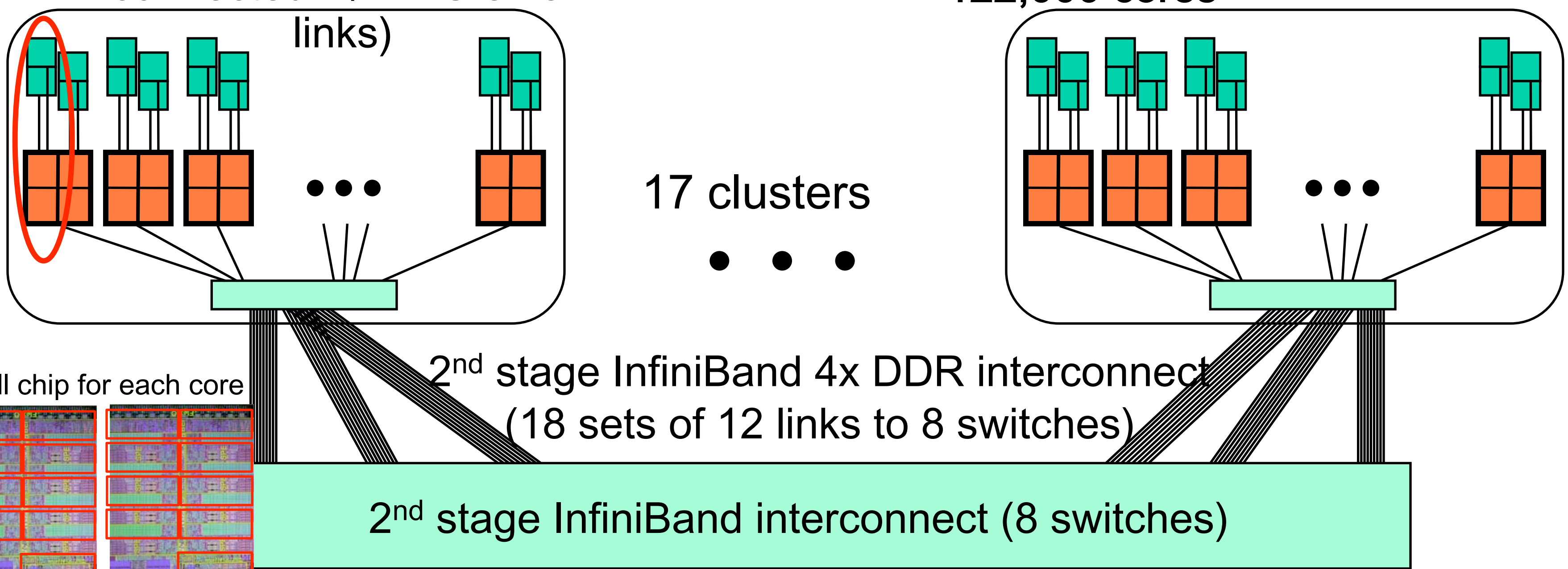
Rank	Site	Computer	Year	Cores	RMax	RPeak
22	The Earth Simulator Center	NEC Earth Simulator	2009	1280	122400	131072
28	JAXA	Fujitsu FX1, Quadcore SPARC64 VII 2.52 GHz, Infiniband DDR	2009	12032	110600	121282
40	Institute of Physical and Chemical Res. (RIKEN)	Fujitsu PRIMERGY RX200S5 Cluster, Xeon X5570 2.93GHz, Infiniband DDR	2009	8256	87890	96760
41	GSIC Center, Tokyo Institute of Technology	Sun Fire x4600/x6250, Opteron 2.4/2.6 GHz, Xeon E5440 2.833 GHz, ClearSpeed CSX600, nVidia GT200; Voltaire Infiniband	2009	31024	87010	163188
42	Information Technology Center, The University of Tokyo	Hitachi Cluster Opteron QC 2.3 GHz, Myrinet 10G	2008	12288	82984	113050
47	Center for Computational Sciences, University of Tsukuba	Appro Xtreme-X3 Server - Quad Opteron Quad Core 2.3 GHz, Infiniband	2009	10368	77280	95385
65	National Institute for Fusion Science (NIFS)	Hitachi SR16000 Model L2, Power6 4.7Ghz, Infiniband	2009	4096	56650	77004.8
69	University of Tokyo/Human Genome Center, IMS	SunBlade x6250, Xeon E5450 3GHz, Infiniband	2009	5760	54210	69120
78	Kyoto University	Fujitsu Cluster HX600, Opteron Quad Core, 2.3 GHz, Infiniband	2008	6656	50510	61235
93	National Institute for Materials Science	SGI Altix ICE 8200EX, Xeon X5560 quad core 2.8 GHz	2009	4096	42690	45875.2
259	National Astronomical Observatory of Japan	Cray XT4 QuadCore 2.2 GHz	2008	3248	22930	28582
277	National Astronomical Observatory of Japan	GRAPE-DR accelerator Cluster, Infiniband	2009	8192	21960	84480
394	Computational Biology Research Center, AIST	IBM eServer Blue Gene Solution	2005	8192	18665	22937.6
397	High Energy Accelerator Research Organization /KEK	IBM eServer Blue Gene Solution	2006	8192	18665	22937.6
398	High Energy Accelerator Research Organization /KEK	IBM eServer Blue Gene Solution	2006	8192	18665	22937.6

LANL Roadrunner
A Petascale System in 2008

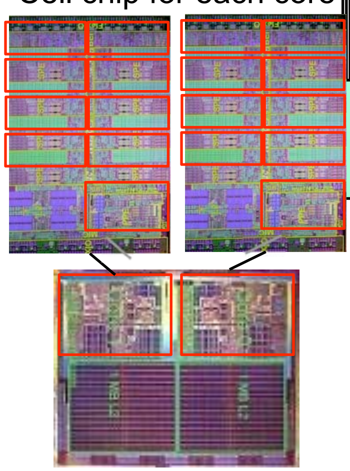
"Connected Unit" cluster
192 Opteron nodes
(180 w/ 2 dual-Cell blades
connected w/ 4 PCIe x8
links)

≈ 13,000 Cell HPC chips
- ≈ **1.33 PetaFlop/s** (from Cell)

≈ 7,000 dual-core Opterons
≈ 122,000 cores



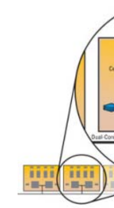
Cell chip for each core



Dual Core Opteron Chip

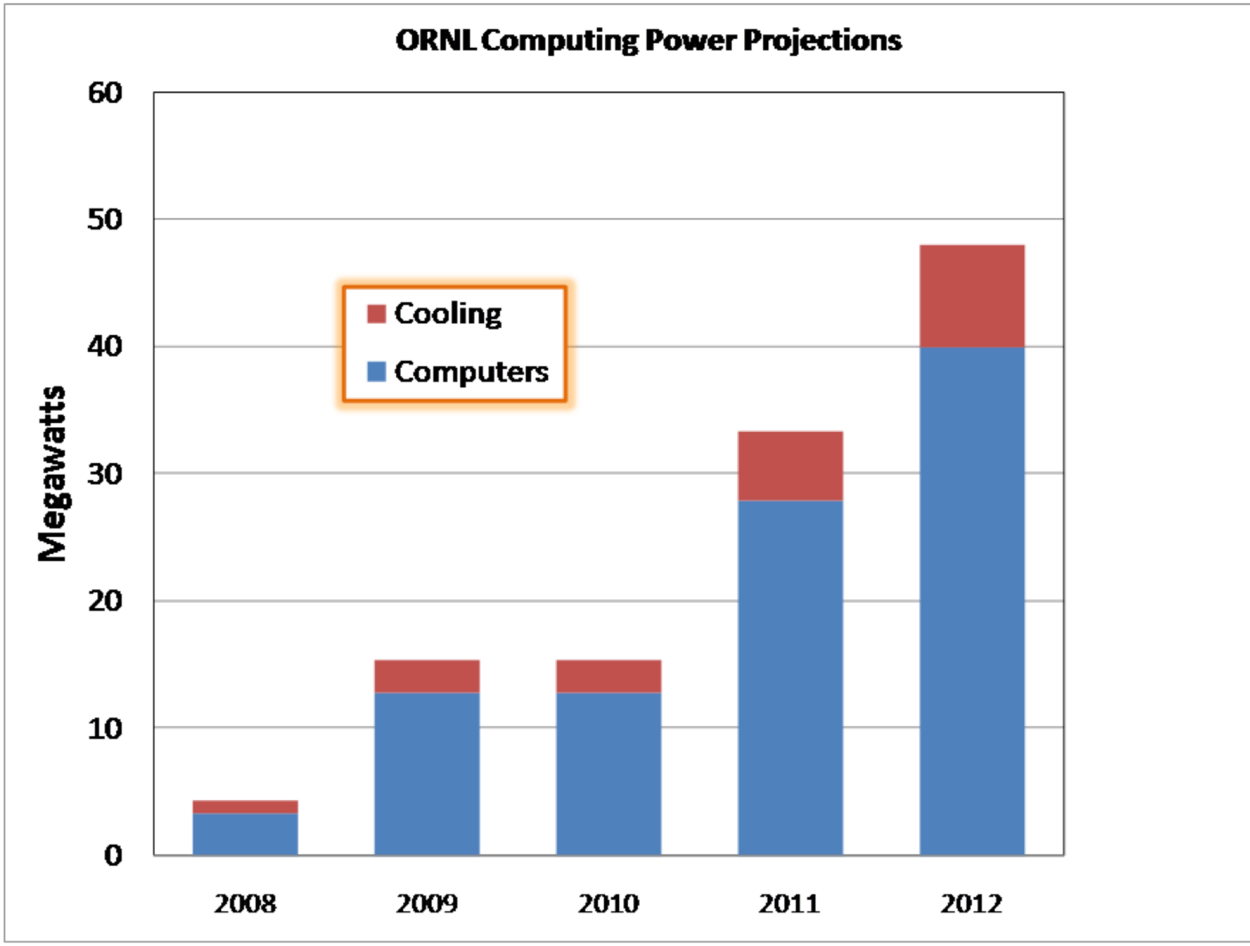
Based on the 100 Gflop/s (DP) Cell chip

Hybrid Design (2 kinds of chips & 3 kinds of cores)
Programming required at 3 levels.



ORNL/UTK Computer Power Cost Projections 2008-2012

- Over the next 5 years ORNL/UTK will deploy 2 large Petascale systems
- Using 15 MW today
- By 2012 close to 50MW!!
- Power costs close to $10M today.
- Cost estimates based on $0.07 per KwH

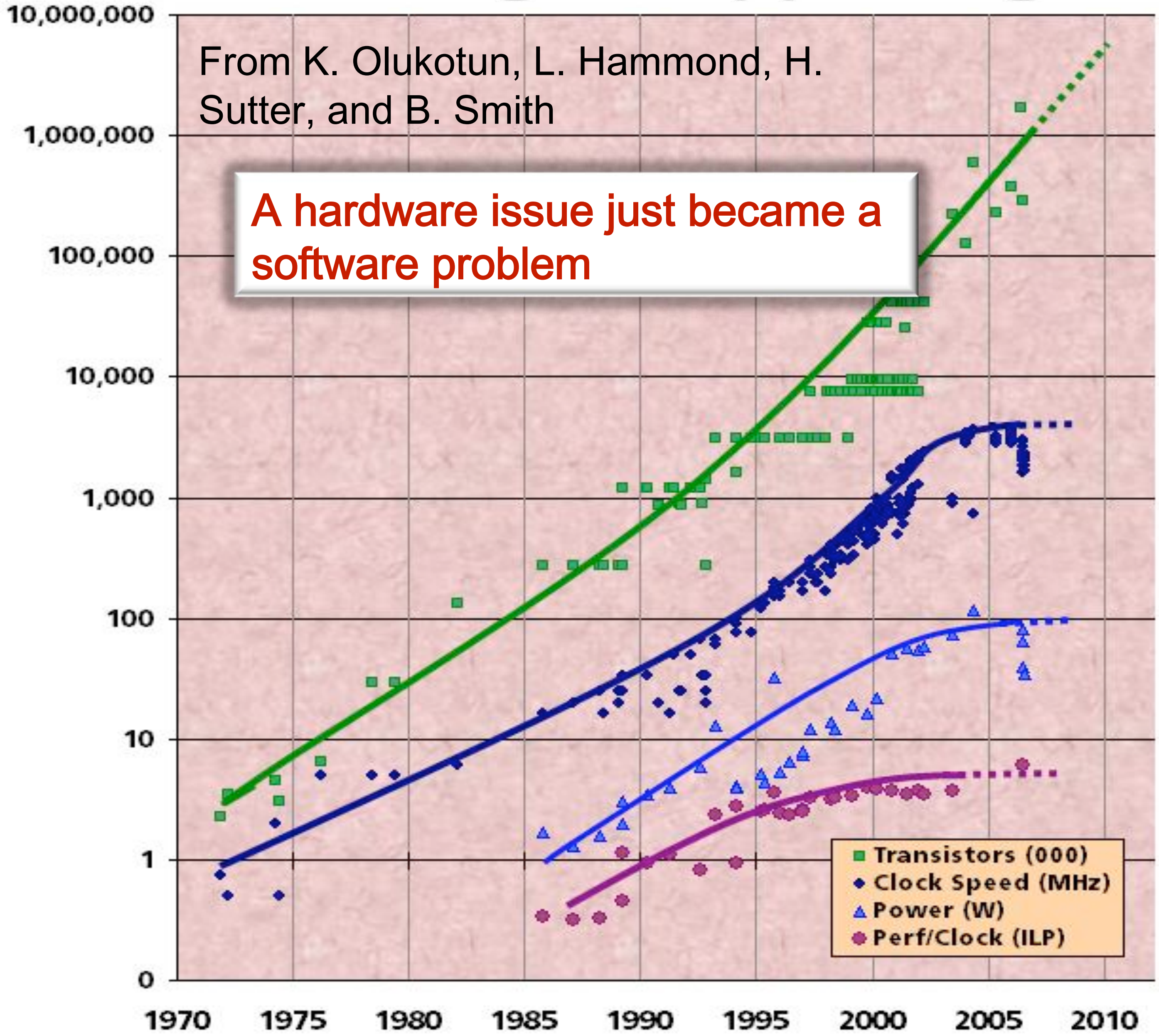


> $10M > $20M > $30M

Cost Per Year

Power becomes the architectural driver for future large systems

Something's Happening Here…



- **In the "old days" it was: each year processors would become faster**
- **Today the clock speed is fixed or getting slower**
- **Things are still doubling every 18 -24 months**
- **Moore's Law reinterpretated.**
  - **Number of cores double every 18-24 months**

Moore's Law Reinterpreted

- **Number of cores per chip doubles every 2 year, while clock speed decreases (not increases).**
  - **Need to deal with systems with millions of concurrent threads**
    - **Future generation will have billions of threads!**
  - **Need to be able to easily replace inter-chip parallelism with intro-chip parallelism**
- **Number of threads of execution doubles every 2 year**

Power Cost of Frequency

- Power ∝ Voltage^2 x Frequency (V^{2}F)
- Frequency ∝ Voltage
- Power ∝ Frequency^3

	Cores	V	Freq	Perf	Power	PE (Bops/watt)
Superscalar	1	1	1	1	1	1
"New" Superscalar	1X	1.5X	1.5X	1.5X	3.3X	0.45X

Power Cost of Frequency

- **Power ∝ Voltage^2 x Frequency (V^{2}F)**
- **Frequency ∝ Voltage**
- **Power ∝ Frequency^3**

	Cores	V	Freq	Perf	Power	PE (Bops/watt)
Superscalar	1	1	1	1	1	1
"New" Superscalar	1X	1.5X	1.5X	1.5X	3.3X	0.45X
Multicore	2X	0.75X	0.75X	1.5X	0.8X	1.88X

(Bigger # is better)

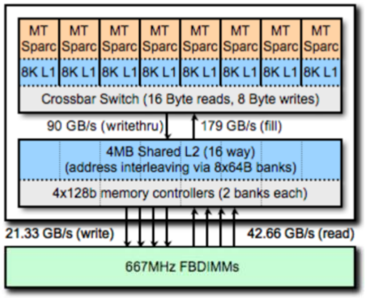
50% more performance with 20% less power

Preferable to use multiple slower devices, than one superfast device

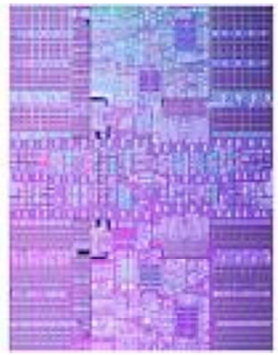
Today's Multicores

99% of Top500 Systems Are Based on Multicore

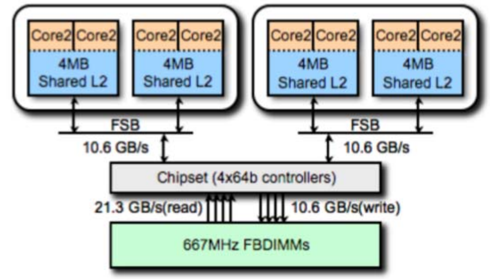
282 use Quad-Core
204 use Dual-Core
3 use Nona-core



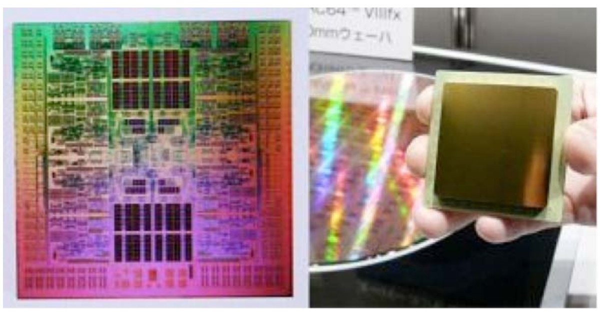
Sun Niagra2 (8 cores)



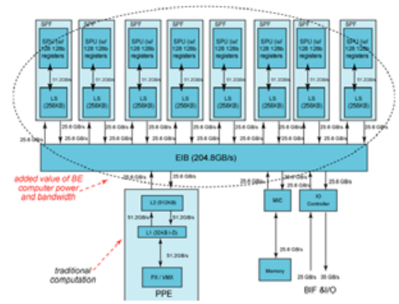
IBM Power 7 (8 cores)



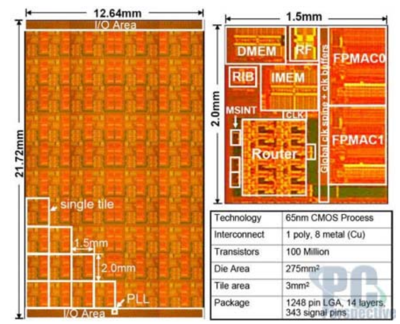
Intel Clovertown (4 cores)



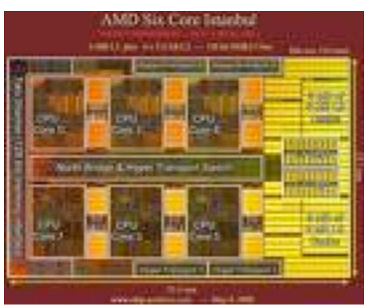
Fujitsu Venus (8 cores)



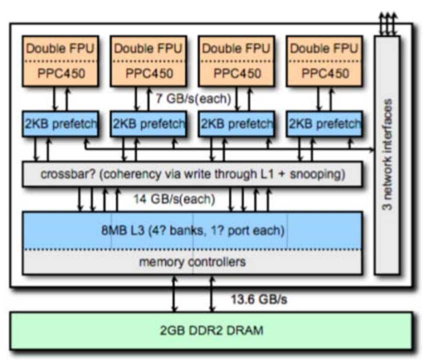
IBM Cell (9 cores)



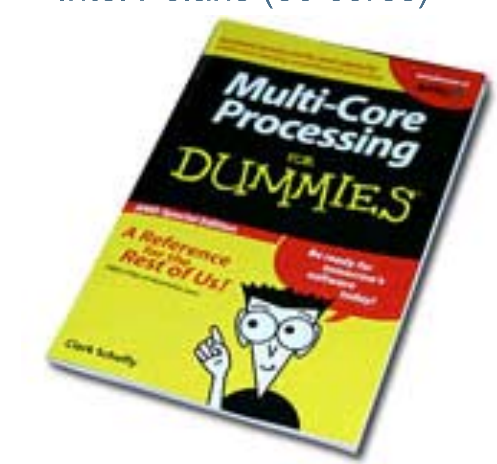
Intel Polaris (80 cores)



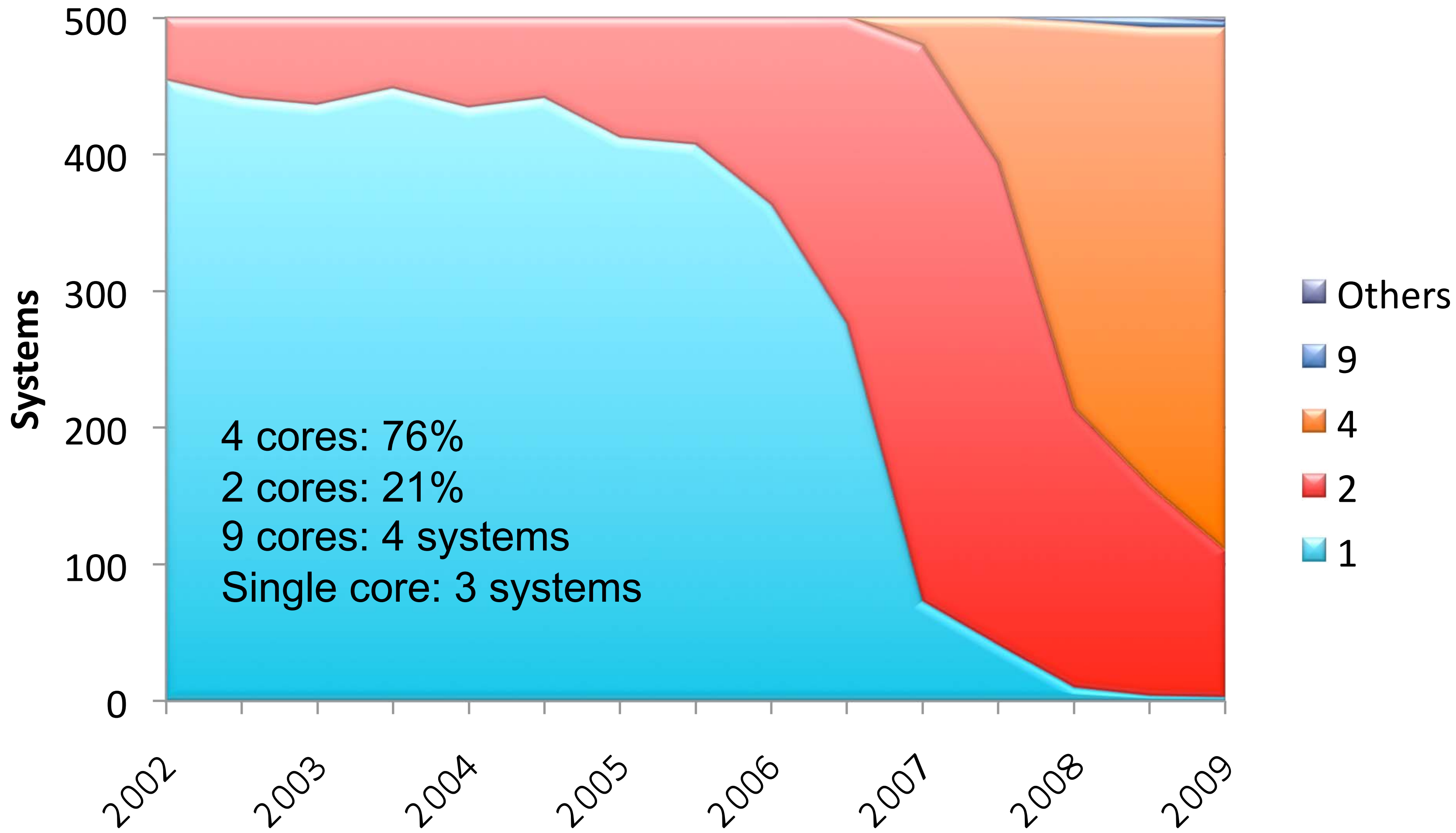
AMD Istambul (6 cores)



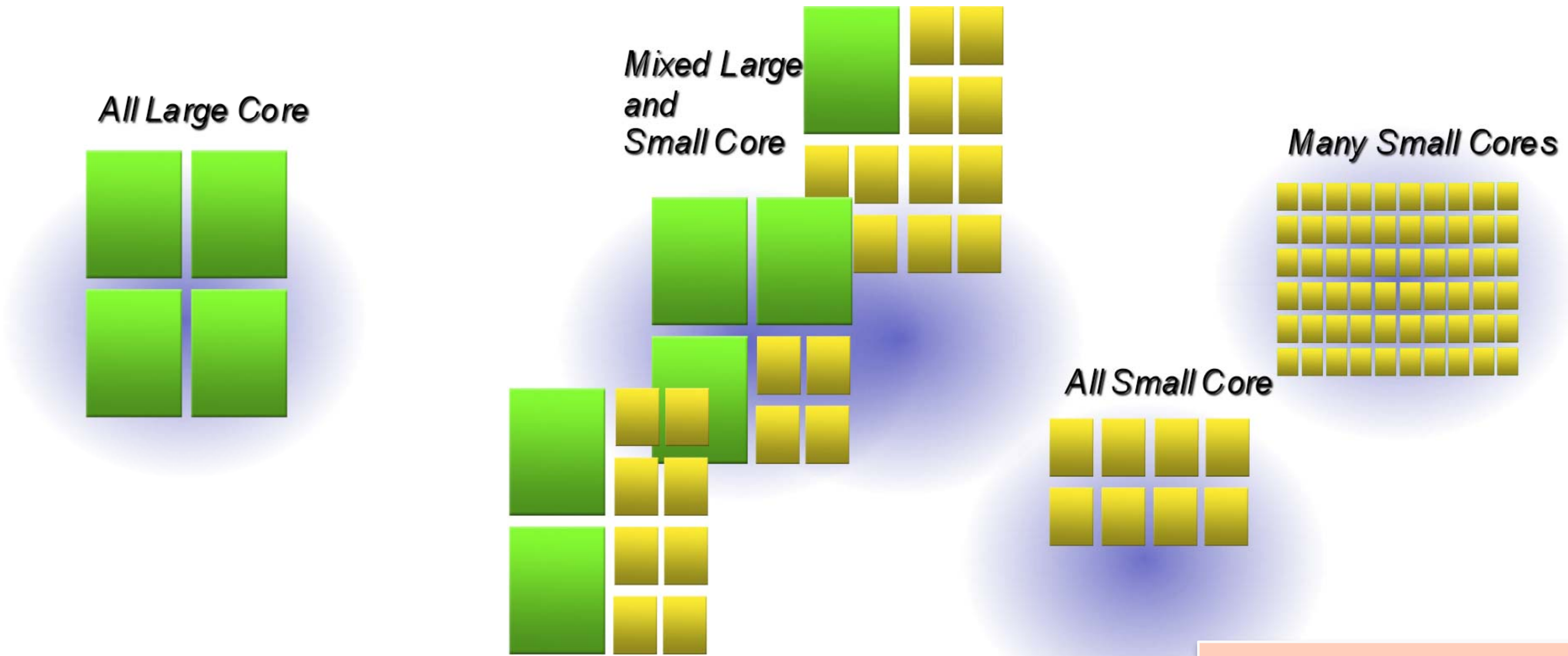
IBM BG/P (4 cores)



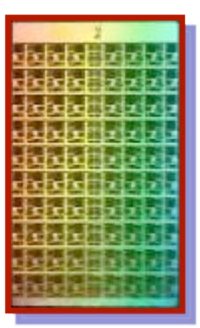
Cores per Socket

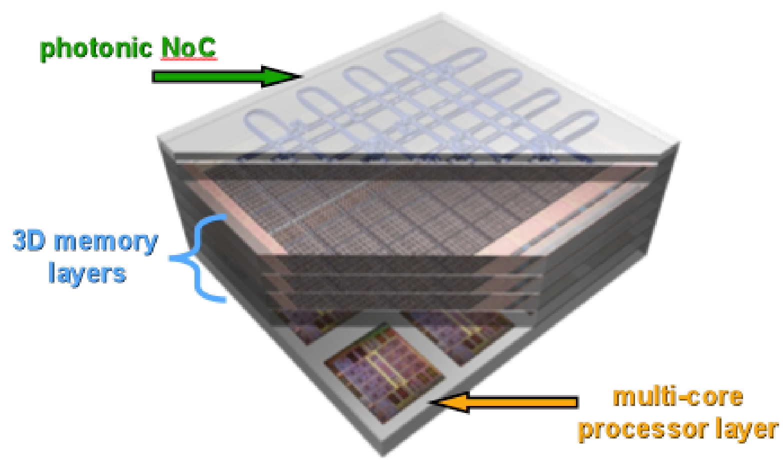


What's Next?



Many Floating-Point Cores





+ 3D Stacked Memory

Different Classes of Chips
- Home
- Games / Graphics
- Business
- Scientific

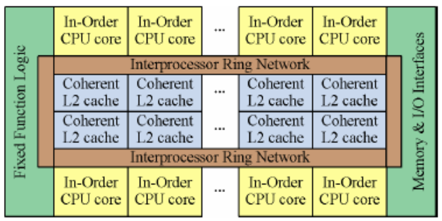
Commodity

- **Moore's "Law" favored consumer commodities**
  - **Economics drove enormous improvements**
  - **Specialized processors and mainframes faltered**
  - **Custom HPC hardware largely disappeared**
  - **Hard to compete against 50%/year improvement**
- **Implications**
  - **Consumer product space defines outcomes**
  - **It does not always go where we hope or expect**
  - **Research environments track commercial trends**
  - **Driven by market economics**
  - **Think about processors, clusters, commodity storage**

Future Computer Systems

- **Most likely be a hybrid design**
- **Think standard multicore chips and accelerator (GPUs)**

- **Today accelerators are attached**
- **Next generation more integrated**
- **Intel's Larrabee in 2010**
  - **8,16,32,or 64 x86 cores**

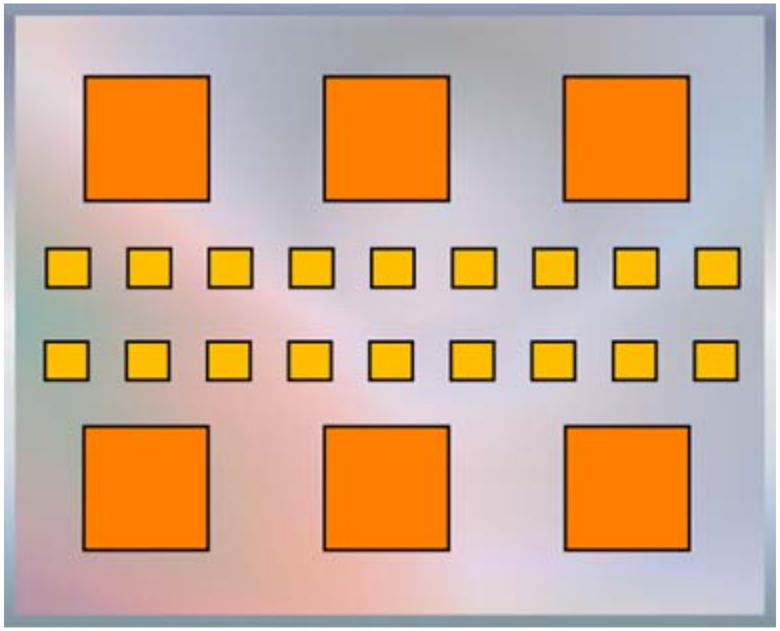


Intel Larrabee

- **AMD's Fusion in 2011**
  - **Multicore with embedded graphics ATI**
- **Nvidia's plans?**

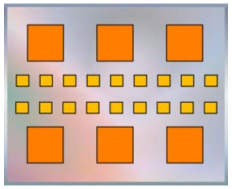
Architecture of Interest

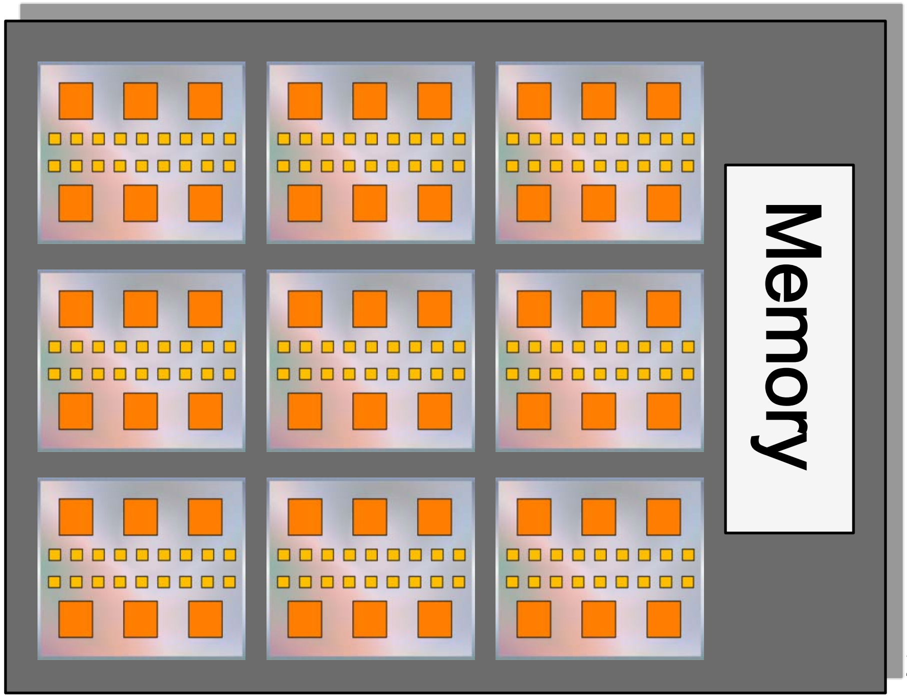
- **Manycore chip**
- **Composed of hybrid cores**
  - **Some general purpose**
  - **Some graphics**
  - **Some floating point**



Architecture of Interest

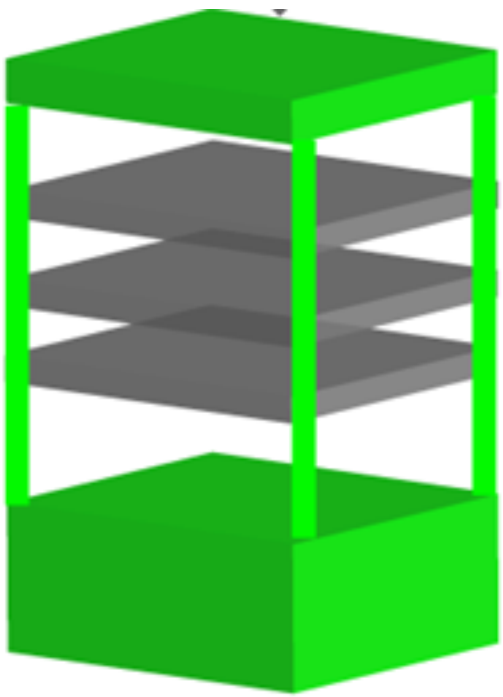
- **Board composed of multiple chips sharing memory**

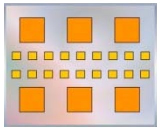


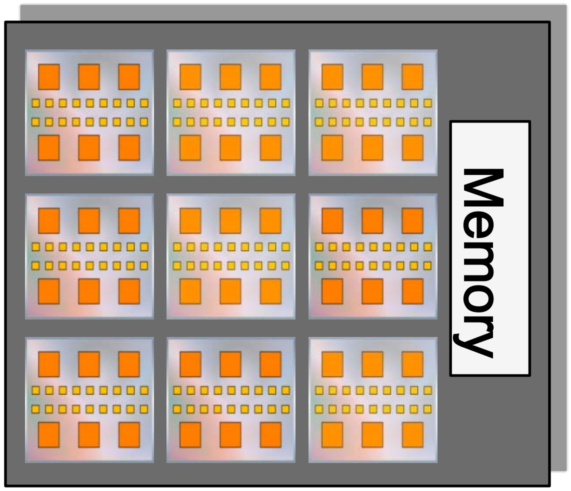


Architecture of Interest

- **Rack composed of multiple boards**

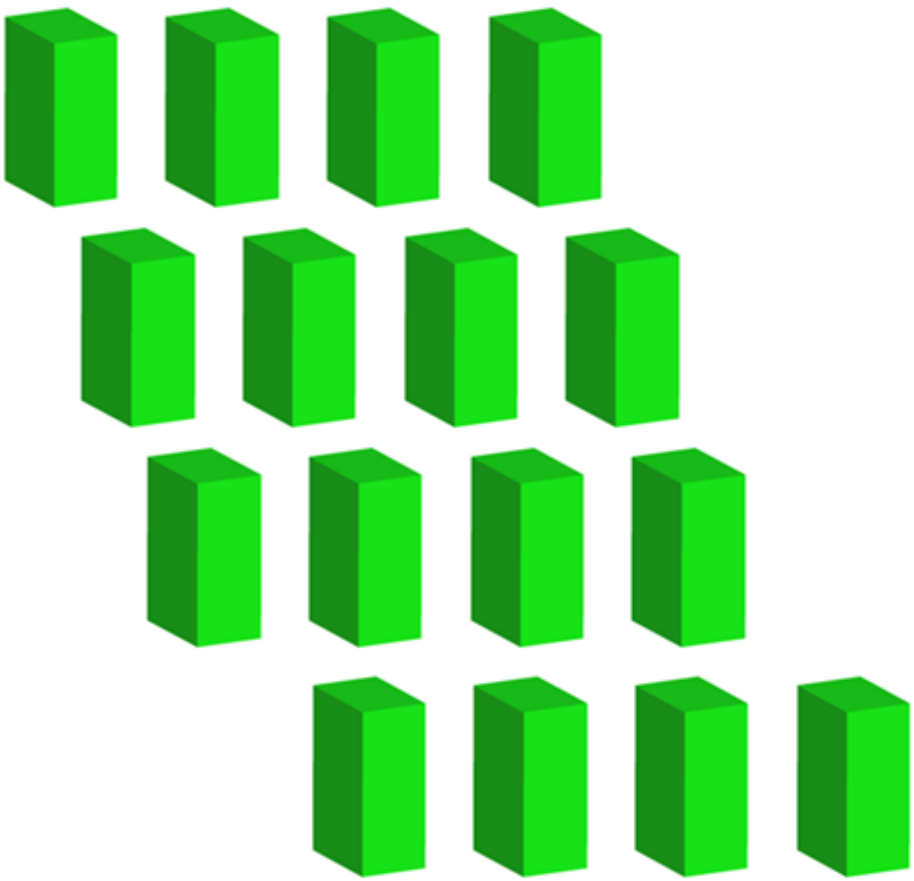


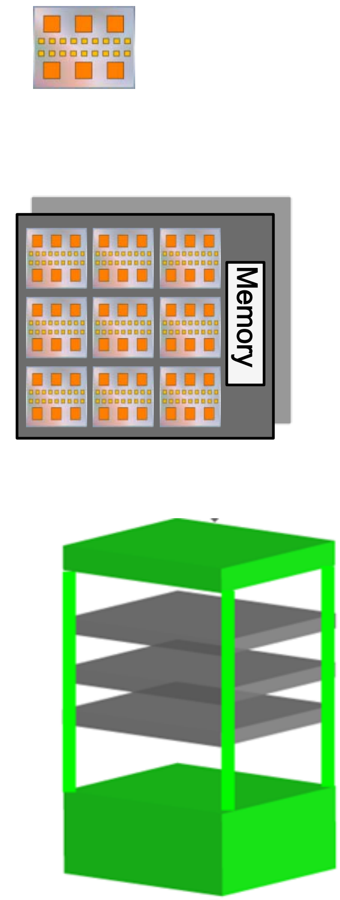




Architecture of Interest

- **A room full of these racks**





- **Think millions of cores**

Moore's Law Reinterpreted

- **Number of cores per chip doubles every 2 year, while clock speed decreases (not increases).**
  - **Need to deal with systems with millions of concurrent threads**
    - **Future generation will have billions of threads!**
  - **Need to rethink the design of our software**
  - **Very disruptive technology**
- **Number of threads of execution doubles every 2 year**

Major Changes to Software

- **Must rethink the design of our software**
  - **Another disruptive technology**
    - **Similar to what happened with cluster computing and message passing**
  - **Rethink and rewrite the applications, algorithms, and software**
- **Numerical libraries for example will change**
  - **For example, both LAPACK and ScaLAPACK will undergo major changes to accommodate this**

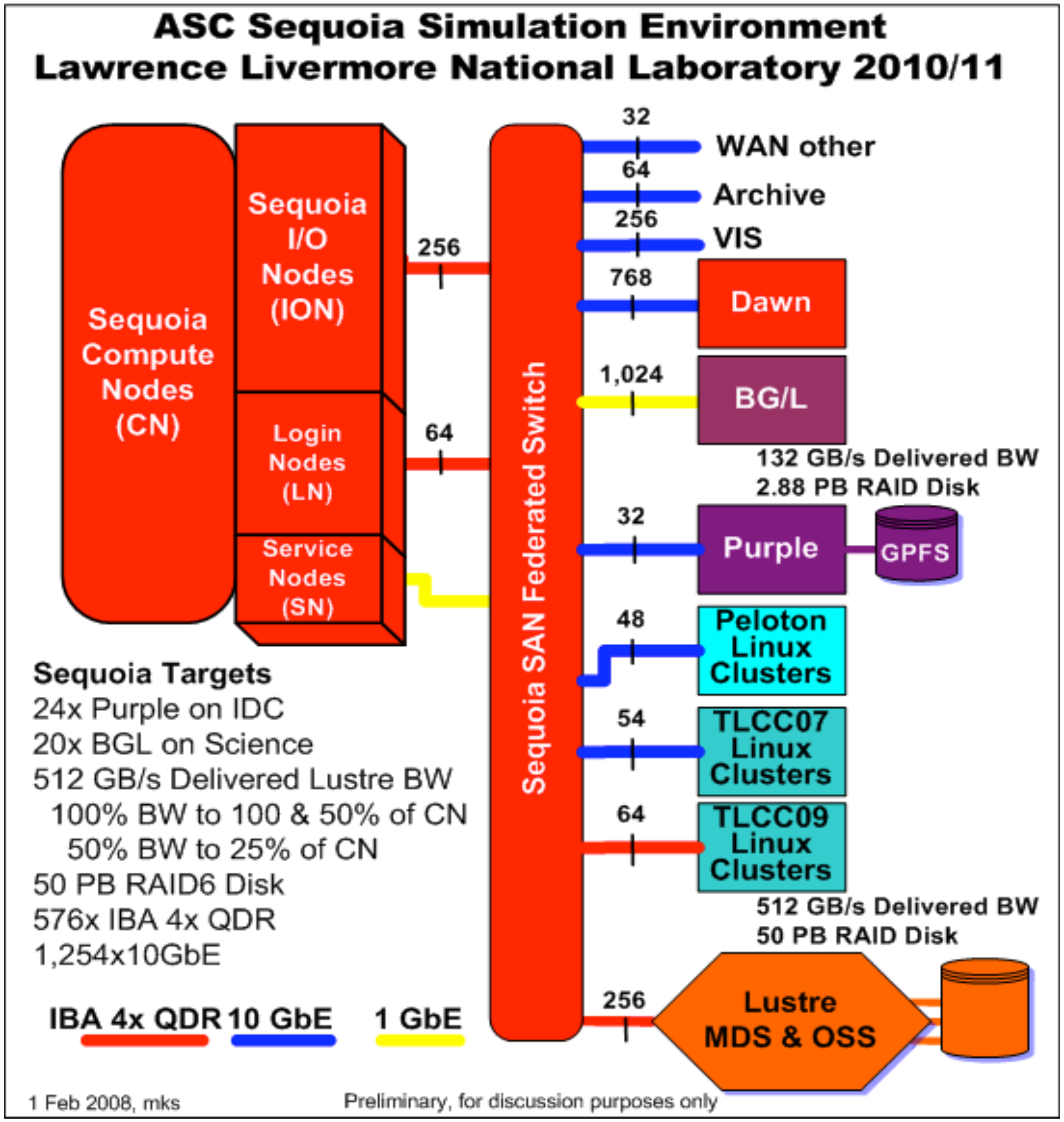
Quasi Mainstream Programming Models

- C, Fortran, C++ and MPI
- OpenMP, pthreads
- (CUDA, RapidMind, Cn) → OpenCL
- PGAS (UPC, CAF, Titanium)
- HPCS Languages (Chapel, Fortress, X10)
- HPC Research Languages and Runtime
- HLL (Parallel Matlab, Grid Mathematica, etc.)

DOE Office of Science

- **ORNL has proposed a system to meet DOE's requirement for 20-40 PF of compute capability split between the Oak Ridge and Argonne LCF centers**
- **ORNL's proposed system will be based on accelerator technology includes software development environment**
- **Plans are to deploy the system in late 2011 with users getting access in 2012**

Sequoia LLNL



- **Diverse usage models drive platform and simulation environment requirements**
  - **Will be 2D ultra-res and 3D high-res Quantification of Uncertainty engine**
  - **3D Science capability for known unknowns and unknown unknowns**
- **Peak 20 petaFLOP/s**
- **IBM BG/Q**
- **Target production 2011-2016**
- **Sequoia Component Scaling**
  - **Memory B:F = 0.08**
  - **Mem BW B:F = 0.2**
  - **Link BW B:F = 0.1**
  - **Min Bisect B:F = 0.03**
  - **SAN BW GB/:PF/s = 25.6**
  - **F is peak FLOP/s**

Blue Waters - The lay of the land

Blue Waters is the powerhouse of the National Science Foundation's strategy to support supercomputers for scientists nationwide

T1	**Blue Waters**	**NCSA/Illinois**	**1 petaflop *sustained* per second**
T2	**Roadrunner**	**DOE/Los Alamos**	**1.3 petaflops peak per second**
T2	**Ranger**	**TACC/Texas**	**504 teraflops peak per second**
T2	**Kraken**	**NICS/Tennessee**	**1 petaflops peak per second**
T3	**Campuses across the U.S.**	**Several sites**	**50-100 teraflops peak per second**

Blue Waters - Main Characteristics

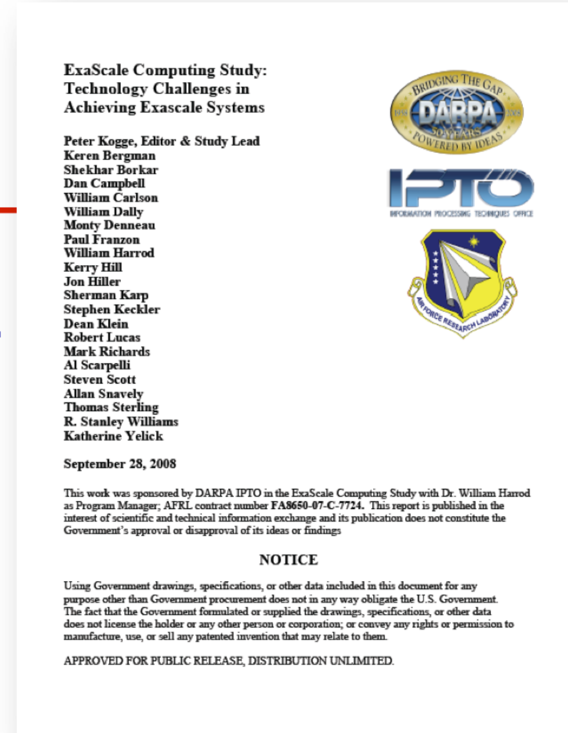
- **Hardware:**
  - **Processor: IBM Power7 multicore architecture**
  - **More than 200,000 cores will be available**
  - **Capable of simultaneous multithreading (SMT)**
  - **Vector multimedia extension capability (VMX)**
  - **Four or more floating-point operations per cycle**
  - **Multiple levels of cache – L1, L2, shared L3**
  - **32 GB+ memory per SMP, 2 GB+ per core**
  - **16+ cores per SMP**
  - **10+ Petabytes of disk storage**
  - **Network interconnect with RDMA technology**

DARPA Ubiquitous High Performance Computing Goals

- *one PFLOPS, air-cooled, single 19-inch cabinet ExtremeScale system. The power budget for the cabinet is 57 kW, including cooling.*
- *achieve 50 GFLOPS/W for the High-Performance Linpack (HPL) benchmark.*
- *The system design should provide high performance for scientific and engineering applications.*
- *The processor node should be capable of being used within terascale embedded and multiple cabinet systems.*
- *The system should be a highly programmable system that does not require the application developer to directly manage the complexity of the system to achieve high performance.*
- *The system must explicitly show a high degree of innovation and software and hardware co-design throughout the life of the program.*

Exascale Computing

- **Exascale systems are likely feasible by 2017±2**
- **10-100 Million processing elements (cores or mini-cores) with chips perhaps as dense as 1,000 cores per socket, clock rates will grow more slowly**
- **3D packaging likely**
- **Large-scale optics based interconnects**
- **10-100 PB of aggregate memory**
- **Hardware and software based fault management**
- **Heterogeneous cores**
- **Performance per watt — stretch goal 100 GF/watt of sustained performance ⇒ >> 10 - 100 MW Exascale system**
- **Power, area and capital costs will be significantly higher than for today's fastest systems**


ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems

Peter Kogge, Editor & Study Lead
Keren Bergman
Shekhar Borkar
Dan Campbell
William Carlson
William Dally
Monty Denneau
Paul Franzon
William Harrod
Kerry Hill
Jon Hiller
Sherman Karp
Stephen Keckler
Dean Klein
Robert Lucas
Mark Richards
Al Scarpelli
Steven Scott
Allan Snavely
Thomas Sterling
R. Stanley Williams
Katherine Yelick

September 28, 2008

This work was sponsored by DARPA IPTO in the ExaScale Computing Study with Dr. William Harrod as Program Manager; AFRL contract number **FA8650-07-C-7724.** This report is published in the interest of scientific and technical information exchange and its publication does not constitute the Government's approval or disapproval of its ideas or findings

NOTICE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

APPROVED FOR PUBLIC RELEASE, DISTRIBUTION UNLIMITED.

Google: exascale computing study

Conclusions

- Moore's Law Reinterpreted
  - Number of cores per chip doubles every two year, while clock speed roughly stable
  - Threads of execution double every 2 years
  - 100 M cores
- Need to deal with systems with millions of concurrent threads
  - Future generation will have billions of threads!
  - MPI and programming languages from the 60's will not make it
- Power limiting clock rate growth
  - Power becomes the architectural driver for Exescale systems.

Conclusions

- **For the last decade or more, the research investment strategy has been overwhelmingly biased in favor of hardware.**
- **This strategy needs to be rebalanced - barriers to progress are increasingly on the software side.**
- **Moreover, the return on investment is more favorable to software.**
  - **Hardware has a half-life measured in years, while software has a half-life measured in decades.**
- **High Performance Ecosystem out of balance**
  - **Hardware, OS, Compilers, Software, Algorithms, Applications**
    - **No Moore's Law for software, algorithms and applications**

Collaborators / Support

Employment opportunities for post-docs in the ICL group at Tennessee

- **Top500**
  - **Hans Meuer, Prometeus**
  - **Erich Strohmaier, LBNL/NERSC**
  - **Horst Simon, LBNL/NERSC**

If you are wondering what's beyond ExaFlops

Mega, Giga, Tera, Peta, Exa, Zetta …

10^{3}	kilo
10^{6}	mega
10^{9}	giga
10^{12}	tera
10^{15}	peta
10^{18}	exa
10^{21}	zetta

10^{24}	yotta
10^{27}	xona
10^{30}	weka
10^{33}	vunda
10^{36}	uda
10^{39}	treda
10^{42}	sorta
10^{45}	rinta
10^{48}	quexa
10^{51}	pepta
10^{54}	ocha
10^{57}	nena
10^{60}	minga
10^{63}	luma