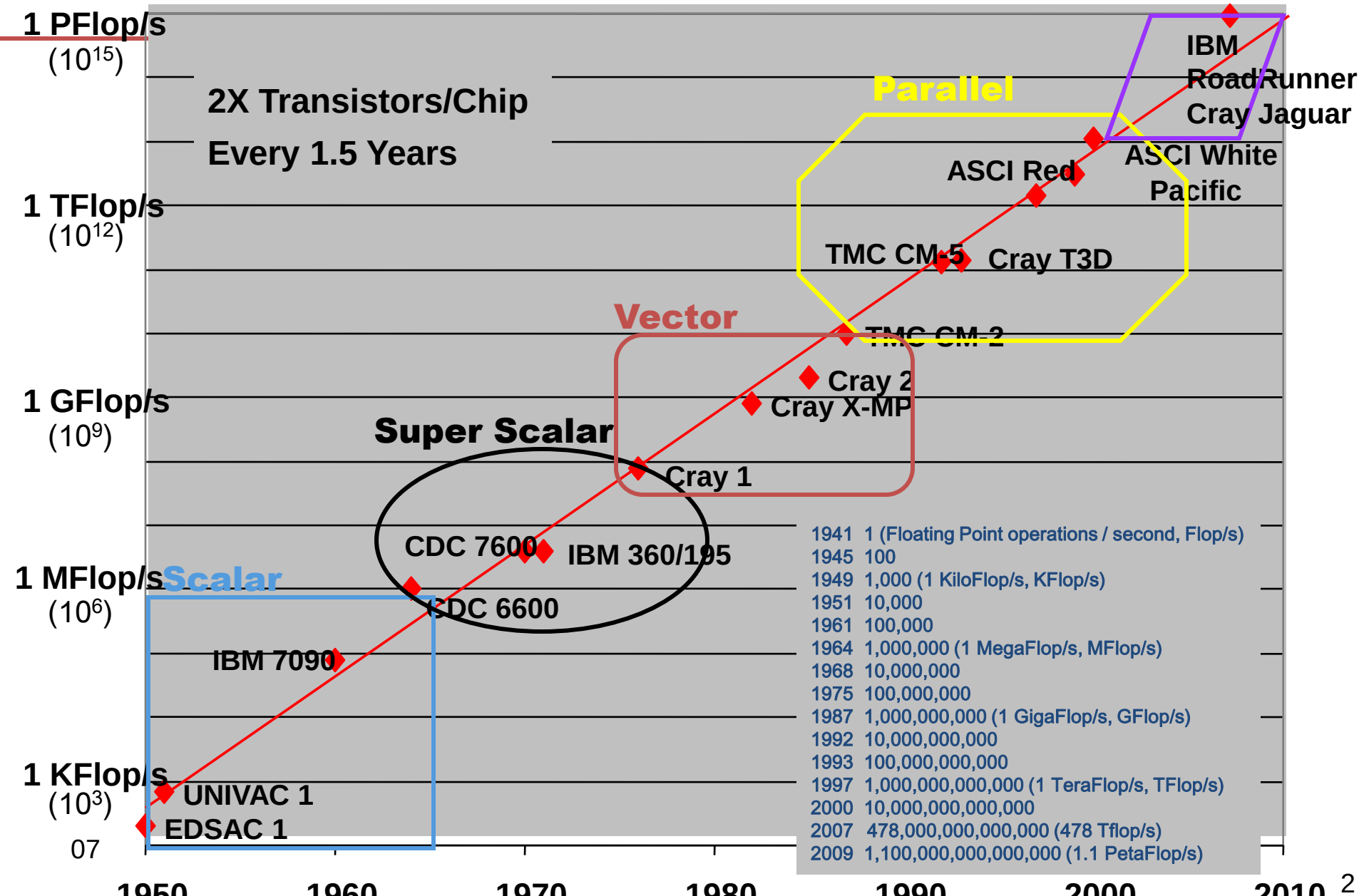
Wake Forest University

An Overview Of High Performance Computing And Challenges For The Future

Jack Dongarra
University of Tennessee
Oak Ridge National Laboratory
University of Manchester

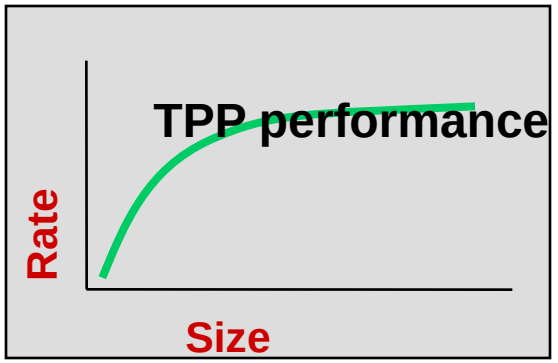
A Growth-Factor of a Billion in Performance in a Career



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

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

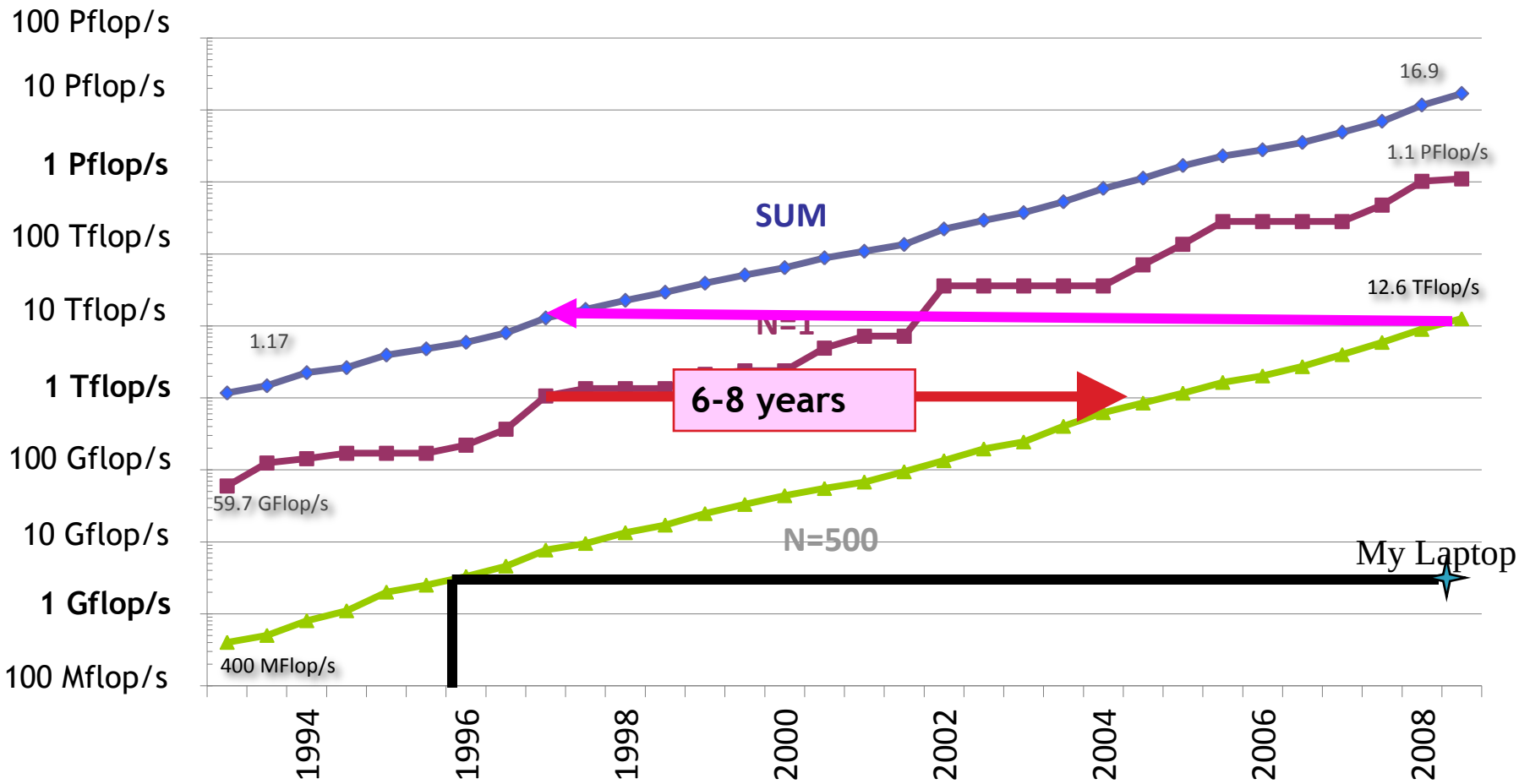
$Ax=b$, *dense problem*



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

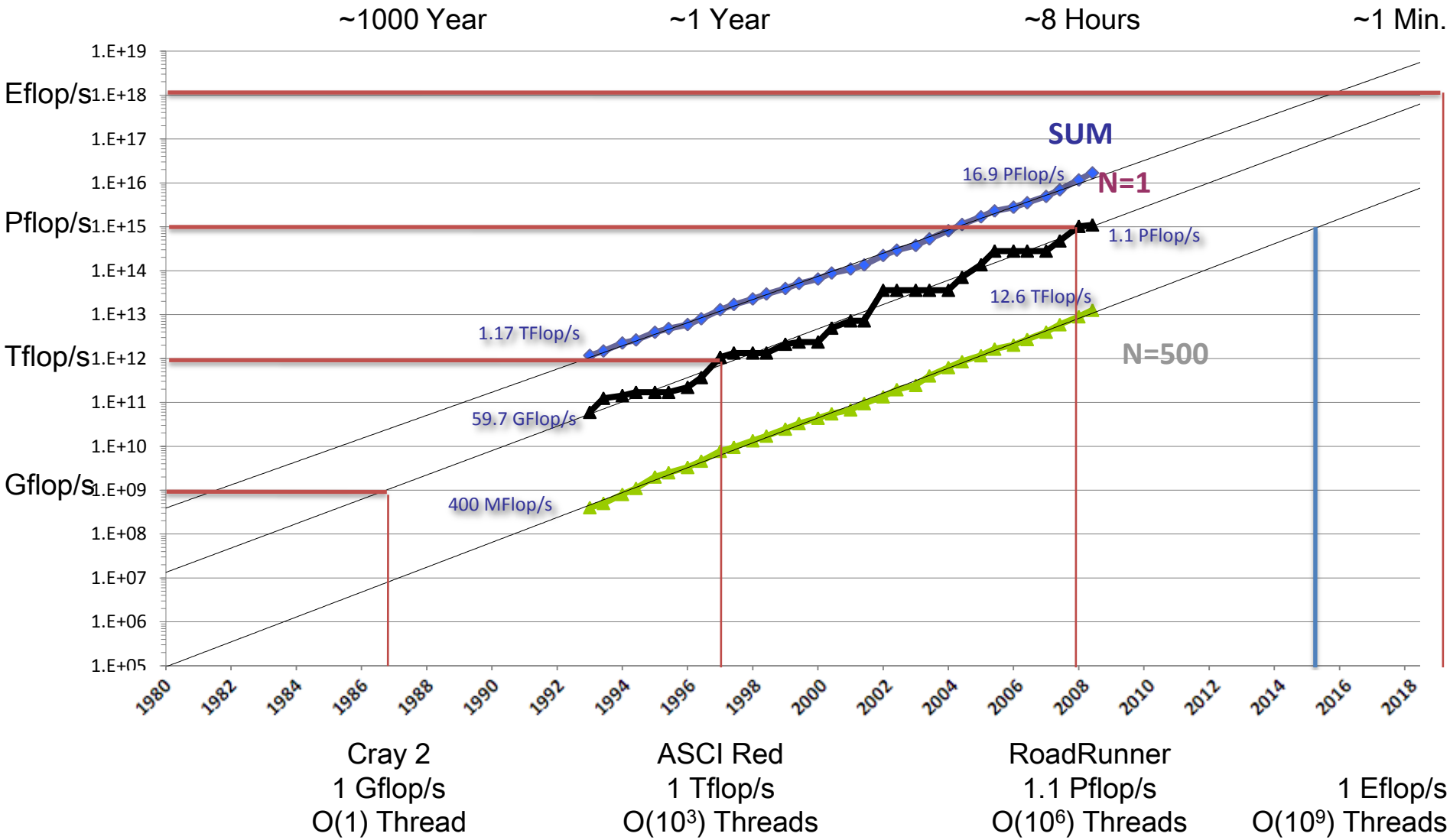
- All data available from **www.top500.org**

Performance Development

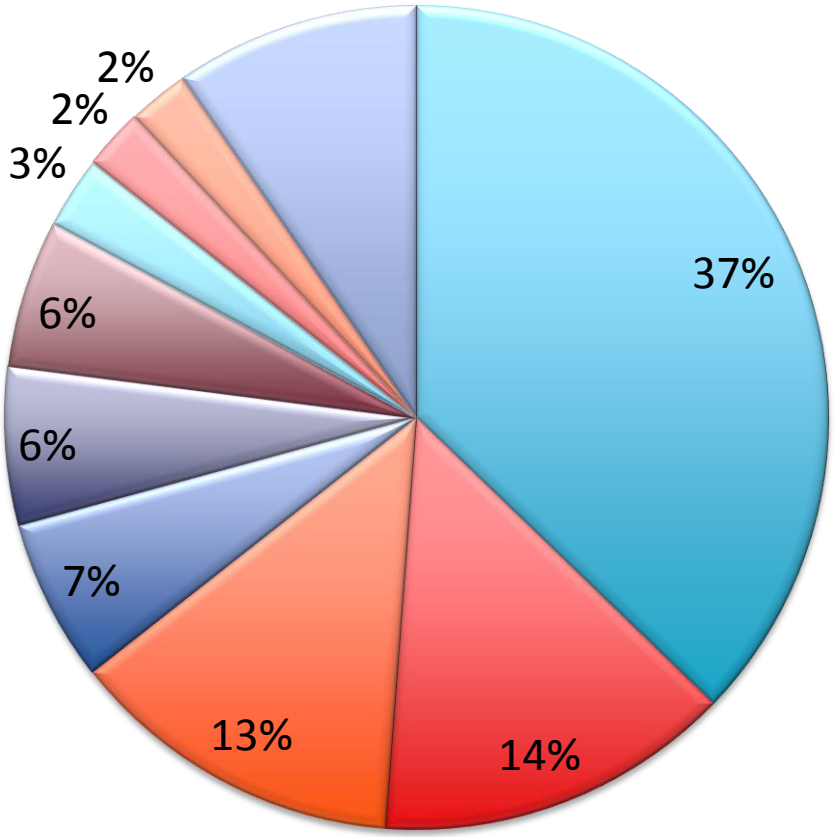


TOP500 SUPERCOMPUTER SITES

Performance Development and Projections



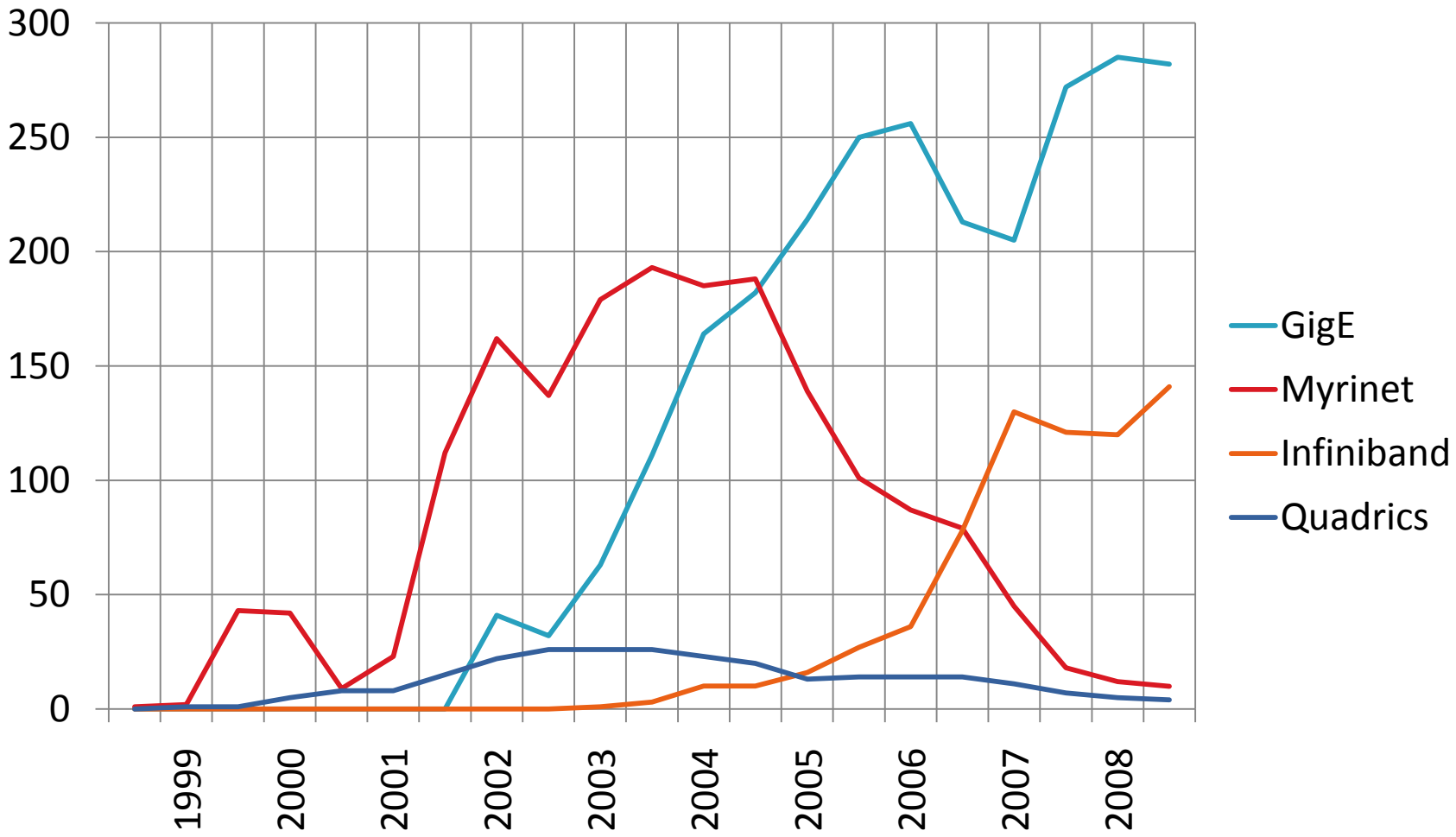
Processors / Systems



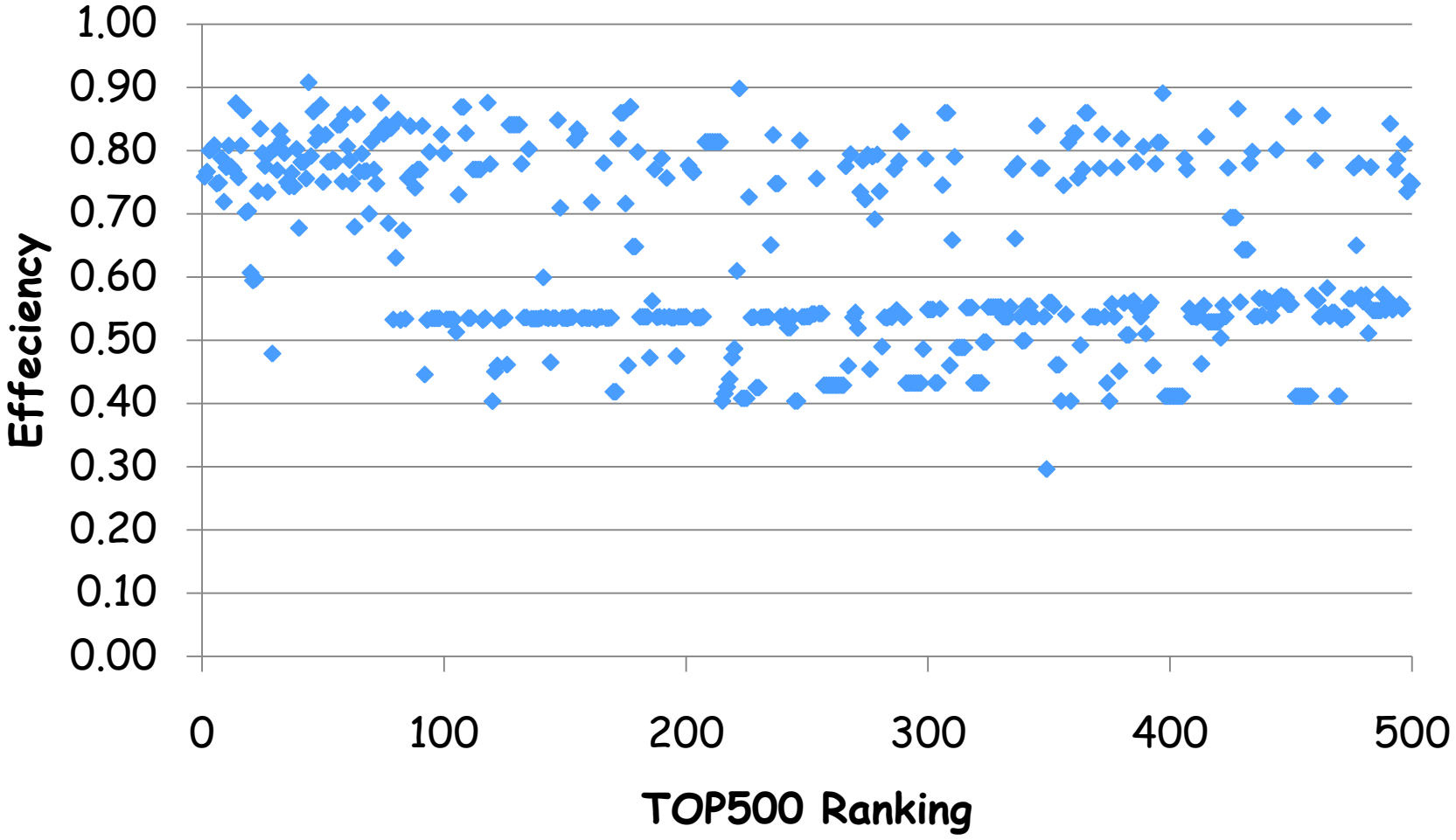
Intel 71%
AMD 13%
IBM 7%

- Xeon E54xx (Harpertown)
- Xeon 51xx (Woodcrest)
- Xeon 53xx (Clovertown)
- Xeon L54xx (Harpertown)
- Opteron Quad Core
- Opteron Dual Core
- PowerPC 440
- PowerPC 450
- POWER6
- Others

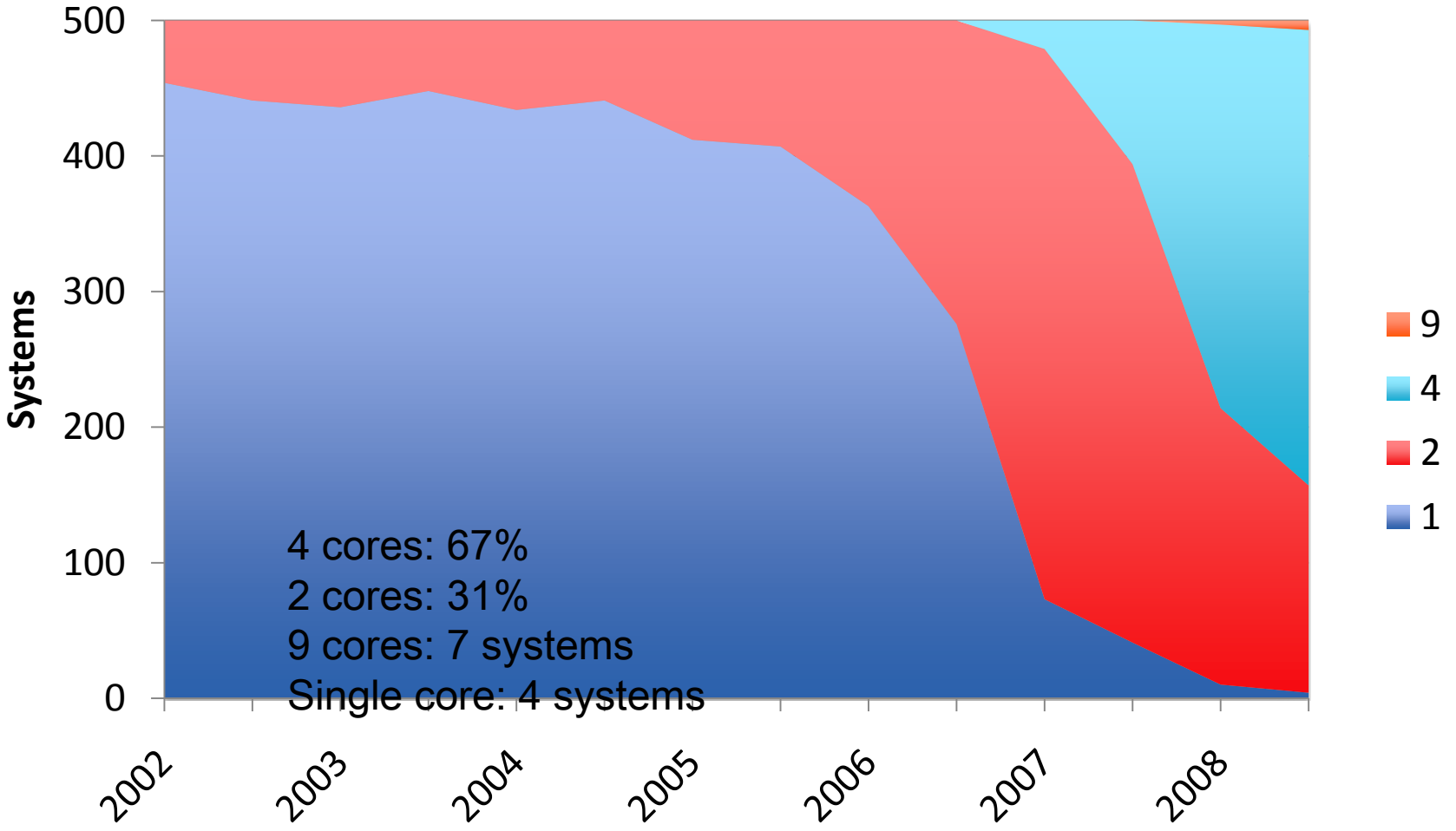
Cluster Interconnects



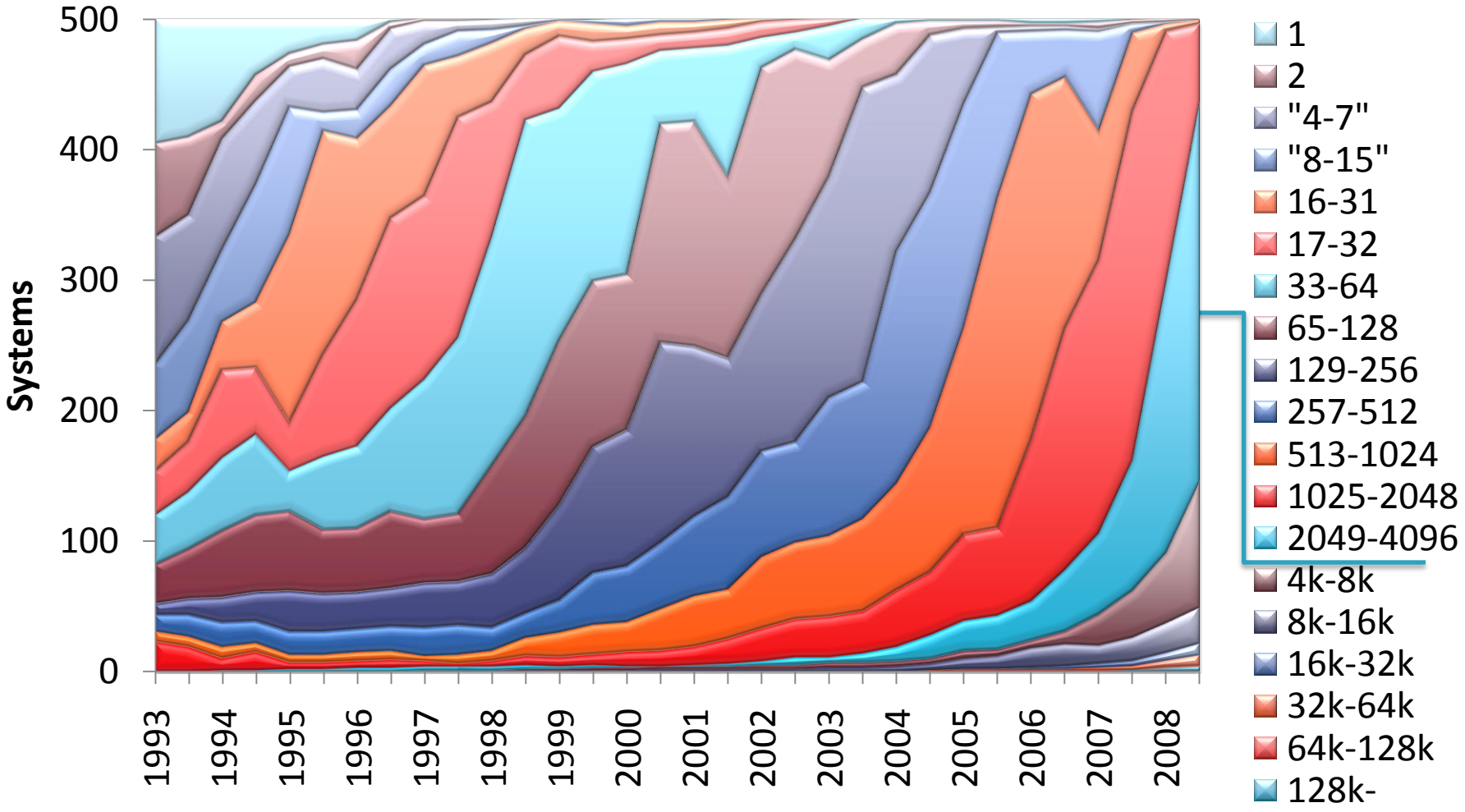
Efficiency


1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Effeciency
0
100
200
300
400
500
TOP500 Ranking

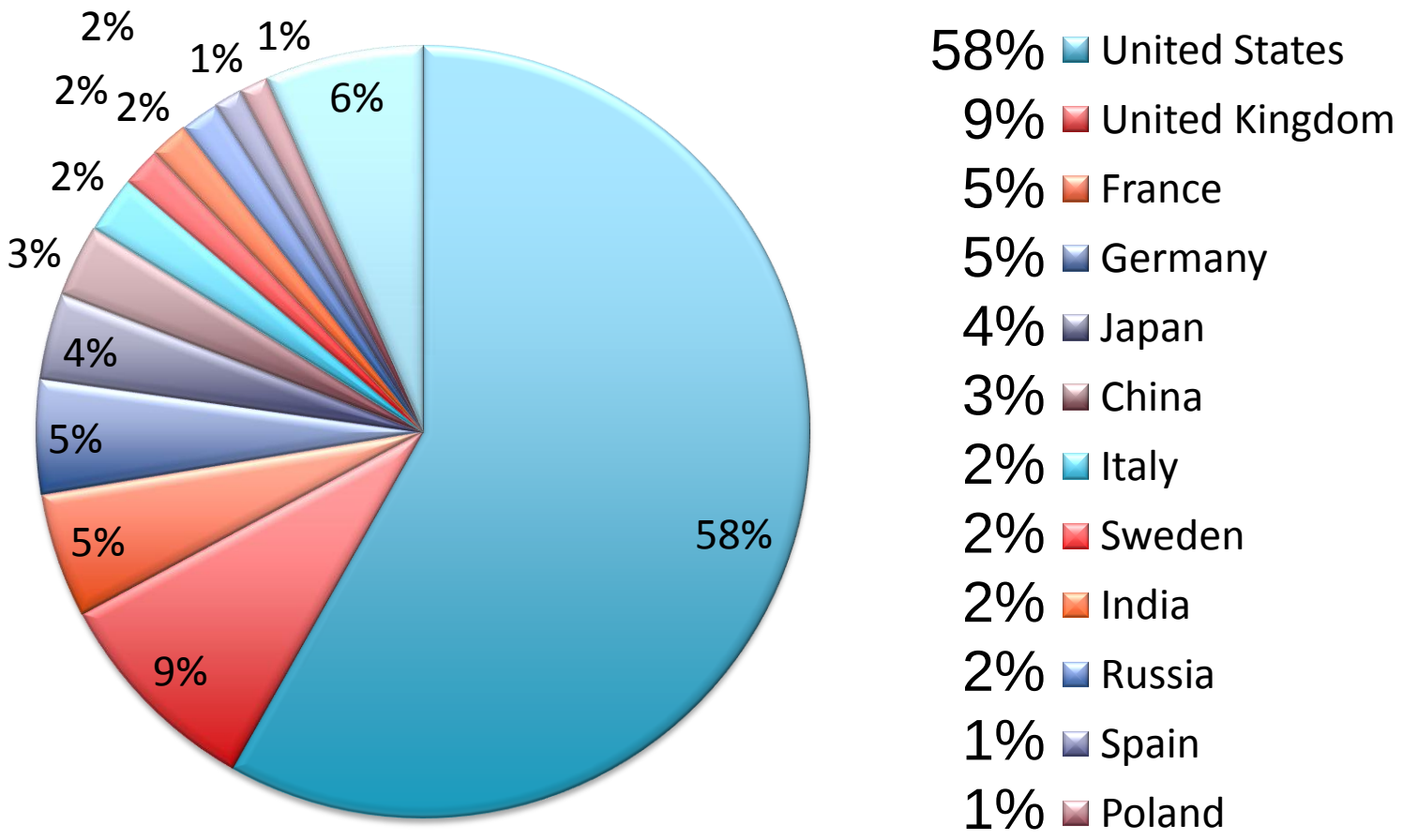
Cores Per Socket



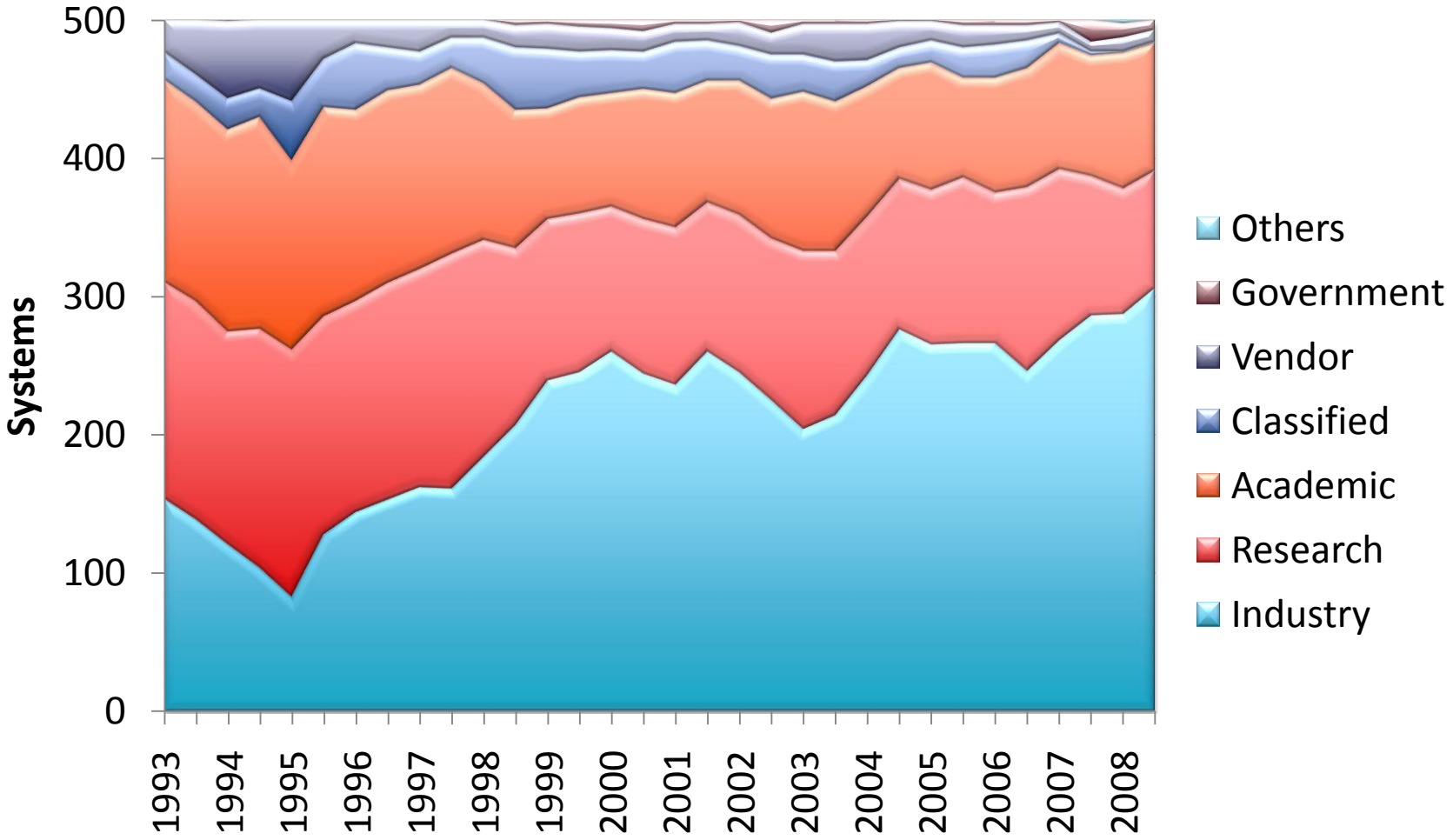
Core Count


Systems
500
400
300
200
100
0
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
1
2
"4-7"
"8-15"
16-31
17-32
33-64
65-128
129-256
257-512
513-1024
1025-2048
2049-4096
4k-8k
8k-16k
16k-32k
32k-64k
64k-128k
128k-

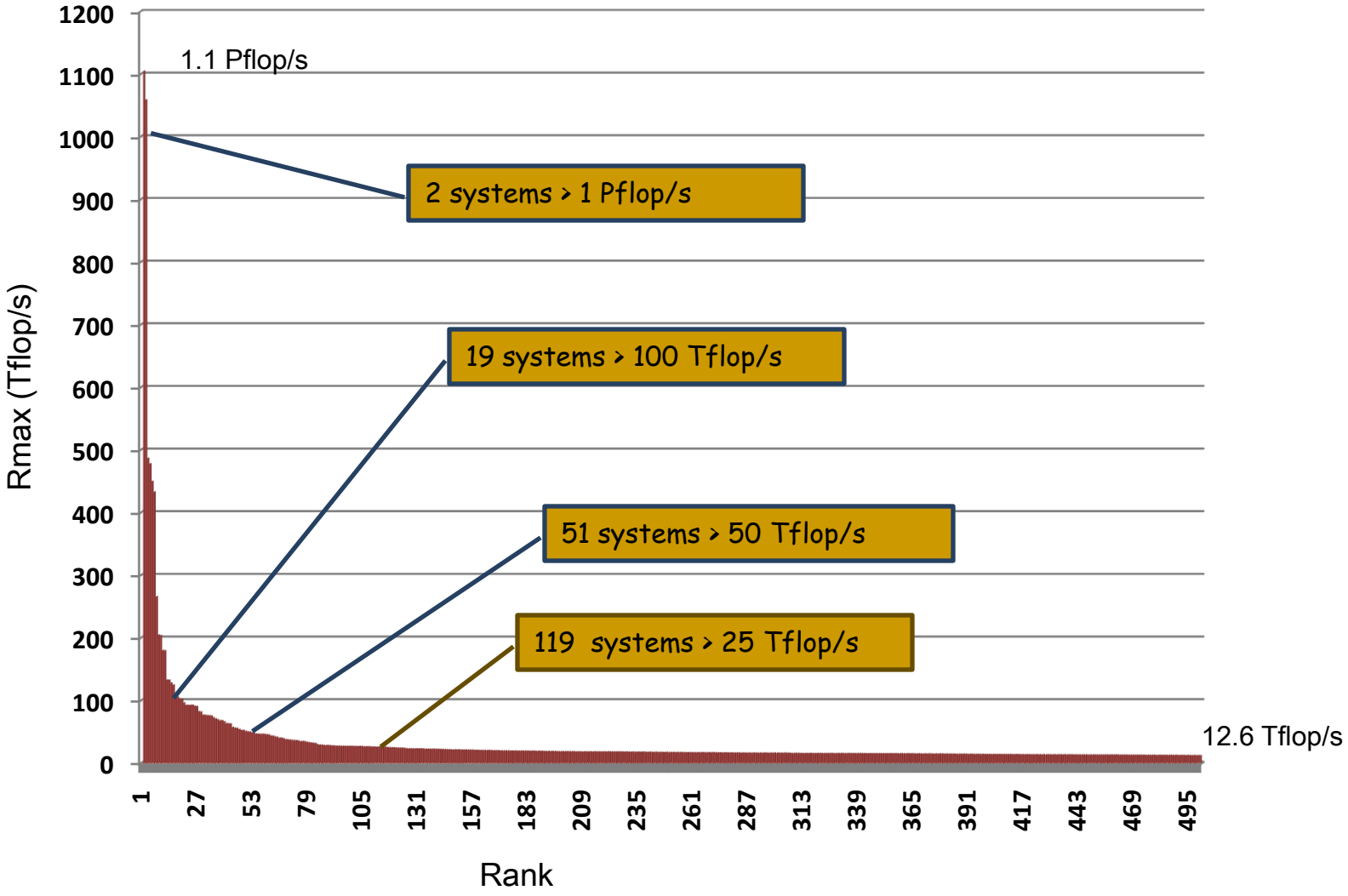
Countries / System Share



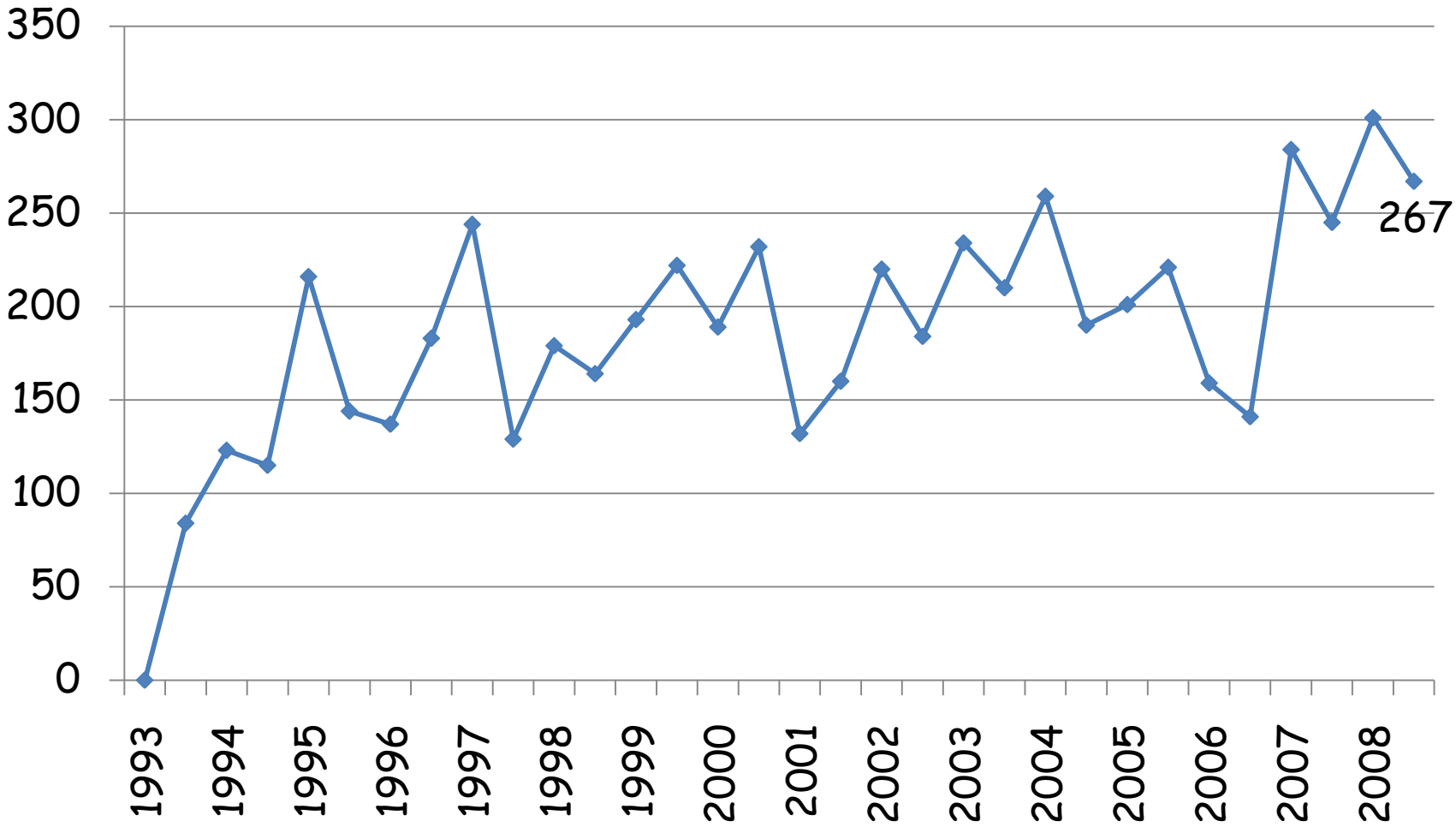
Customer Segments



Distribution of the Top500



Replacement Rate



32nd List: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Tflops]	Rmax/ Rpeak
1	DOE/NNSA/LANL	IBM / Roadrunner - BladeCenter QS22/LS21	USA	129600	1105.0	76%
2	DOE/Oak Ridge National Laboratory	Cray / Jaguar - Cray XT5 QC 2.3 GHz	USA	150152	1059.0	77%
3	NASA/Ames Research Center/NAS	SGI / Pleiades - SGI Altix ICE 8200EX	USA	51200	487.0	80%
4	DOE/NNSA/LLNL	IBM / eServer Blue Gene Solution	USA	212992	478.2	80%
5	DOE/Argonne National Laboratory	IBM / Blue Gene/P Solution	USA	163840	450.3	81%
6	NSF/Texas Advanced Computing Center/Univ. of Texas	Sun / Ranger - SunBlade x6420	USA	62976	433.2	75%
7	DOE/NERSC/LBNL	Cray / Franklin - Cray XT4	USA	38642	266.3	75%
8	DOE/Oak Ridge National Laboratory	Cray / Jaguar - Cray XT4	USA	30976	205.0	79%
9	DOE/NNSA/Sandia National Laboratories	Cray / Red Storm - XT3/4	USA	38208	204.2	72%
10	Shanghai Supercomputer Center	Dawning 5000A, Windows HPC 2008	China	30720	180.6	77%

32nd List: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Tflops]	Rmax/ Rpeak	Power [MW]	MF/W
1	DOE/NNSA/LANL	IBM / Roadrunner - BladeCenter QS22/LS21	USA	129600	1105.0	76%	2.48	**445**
2	DOE/Oak Ridge National Laboratory	Cray / Jaguar - Cray XT5 QC 2.3 GHz	USA	150152	1059.0	77%	6.95	**152**
3	NASA/Ames Research Center/NAS	SGI / Pleiades - SGI Altix ICE 8200EX	USA	51200	487.0	80%	2.09	**233**
4	DOE/NNSA/LLNL	IBM / eServer Blue Gene Solution	USA	212992	478.2	80%	2.32	**205**
5	DOE/Argonne National Laboratory	IBM / Blue Gene/P Solution	USA	163840	450.3	81%	1.26	**357**
6	NSF/Texas Advanced Computing Center/Univ. of Texas	Sun / Ranger - SunBlade x6420	USA	62976	433.2	75%	2.0	**217**
7	DOE/NERSC/LBNL	Cray / Franklin - Cray XT4	USA	38642	266.3	75%	1.15	**232**
8	DOE/Oak Ridge National Laboratory	Cray / Jaguar - Cray XT4	USA	30976	205.0	79%	1.58	**130**
9	DOE/NNSA/Sandia National Laboratories	Cray / Red Storm - XT3/4	USA	38208	204.2	72%	2.5	**81**
10	Shanghai Supercomputer Center	Dawning 5000A, Windows HPC 2008	China	30720	180.6	77%	-	**-**

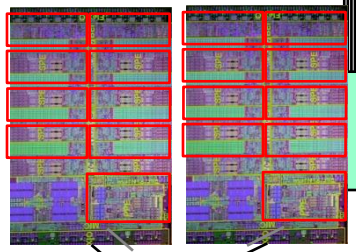
LANL Roadrunner
A Petascale System in 2008

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

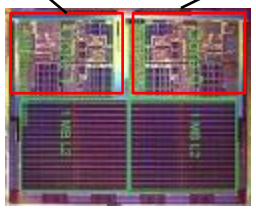
≈ 13,000 Cell HPC chips
- ≈ **1.33 PetaFlop/s** (from Cell)

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

17 clusters



Cell chip for each core



Dual Core Opteron Chip

2nd stage InfiniBand 4x DDR interconnect
(18 sets of 12 links to 8 switches)

2nd stage InfiniBand interconnect (8 switches)

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

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

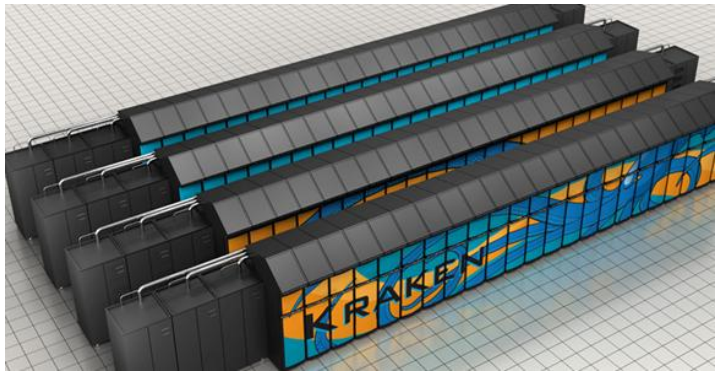
ORNL's Newest System Jaguar XT5

Jaguar	Total	XT5	XT4
Peak Performance	**1,645**	**1,382**	**263**
AMD Opteron Cores	181,504	150,176	31,328
System Memory (TB)	362	300	62
Disk Bandwidth (GB/s)	284	240	44
Disk Space (TB)	10,750	10,000	750
Interconnect Bandwidth (TB/s)	532	374	157

The systems will be combined after acceptance of the new XT5 upgrade. Each system will be linked to the file system through 4x-DDR Infiniband

NICS UT's HPC System

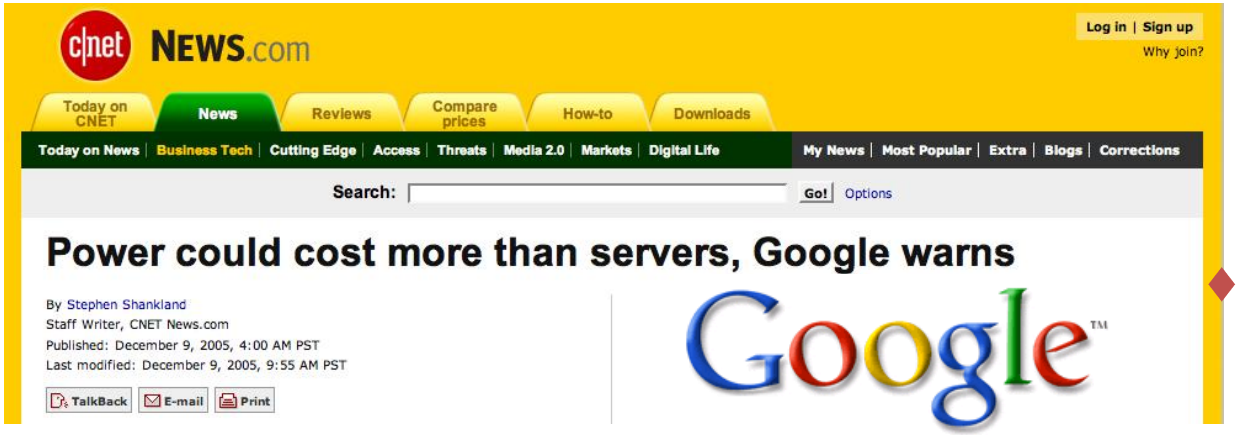
- University of Tennessee's
- National Institute for Computational Sciences
- Housed at ORNL
- Operated for the NSF
- Named Kraken



Today:

- Cray XT5 (608 TF) + Cray XT4 (167 TF)
  - XT5: 16,512 sockets, 66,048 cores
  - XT4: 4,512 sockets, 18,048 cores
- Number 15 on the Top500

Power is an Industry Wide Problem



- Google facilities
  - leveraging hydroelectric power
    - old aluminum plants

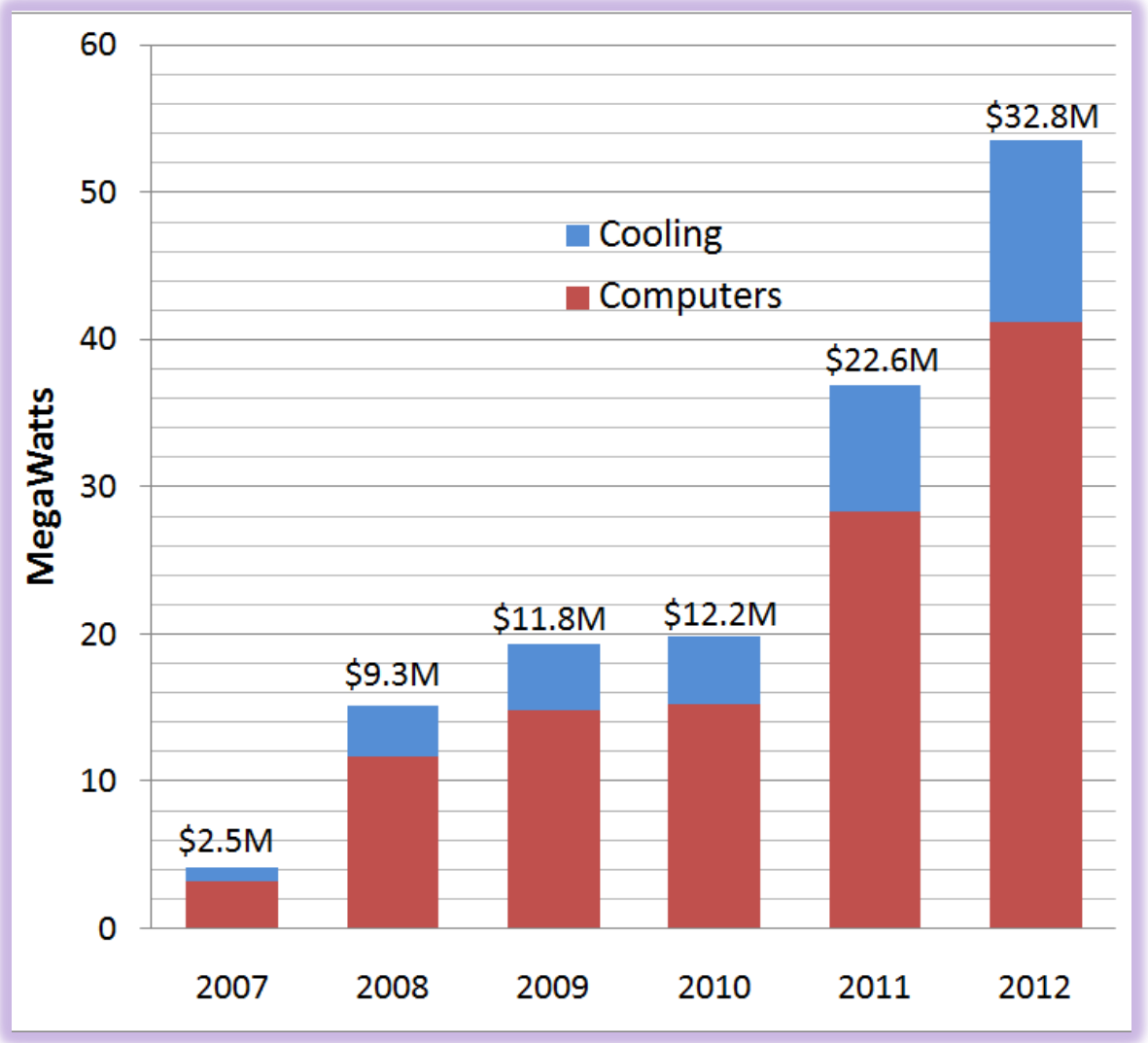
The New York Times **"Hiding in Plain Sight, Google Seeks More Power", by John Markoff, June 14, 2006**

Microsoft and Yahoo are building big data centers upstream in Wenatchee and Quincy, Wash.
– To keep up with Google, which means they need cheap electricity and readily accessible data networking

Microsoft Quincy, Wash.
470,000 Sq Ft, 47MW!

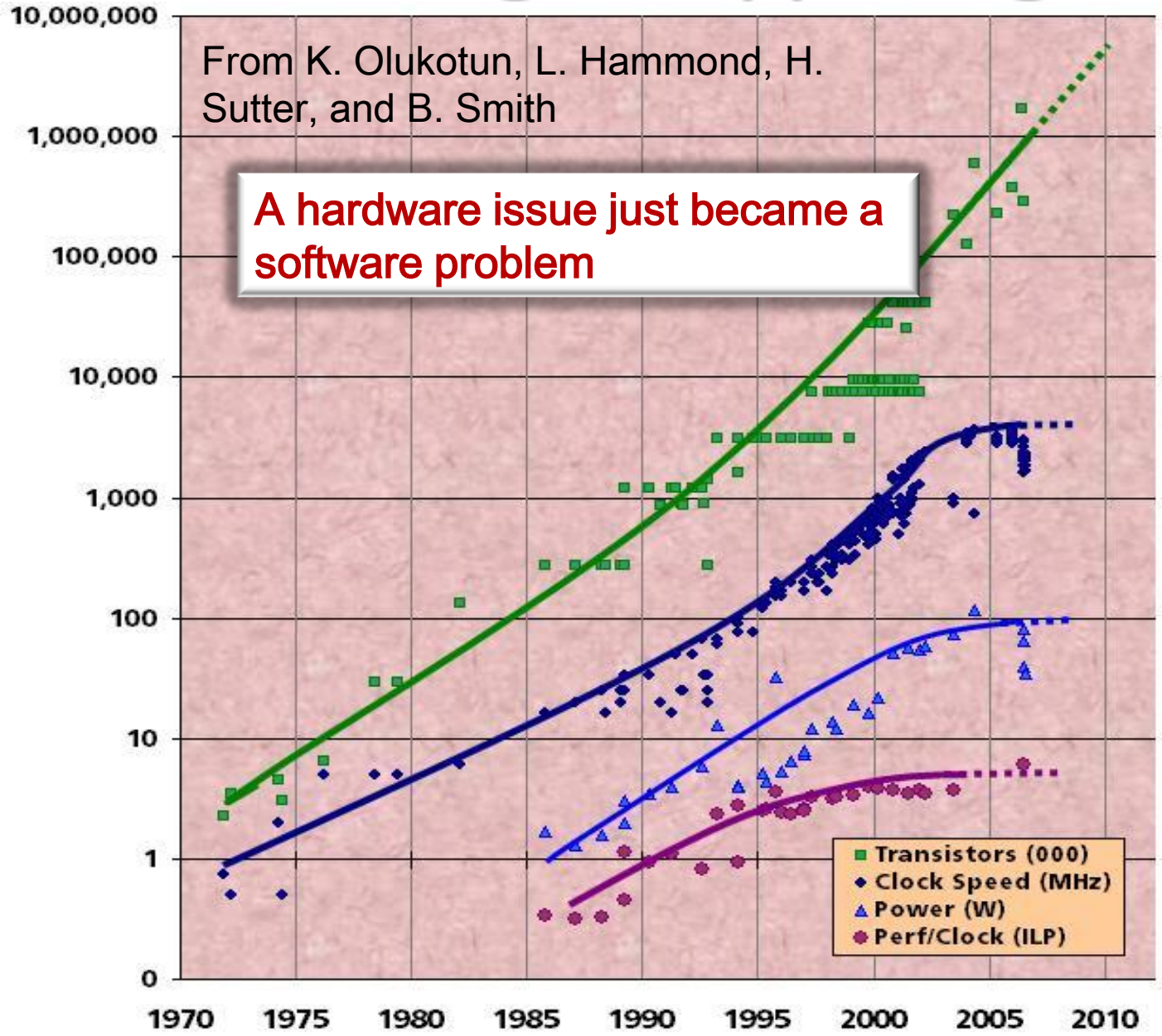
ORNL/UTK Power Cost Projections 2007-2011

- Over the next 5 years ORNL/UTK will deploy 2 large Petascale systems
- Using 4 MW today, going to 15MW before year end
- By 2012 could be using more than 50MW!!
- Cost estimates based on $0.07 per KwH



Includes both DOE and NSF 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

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

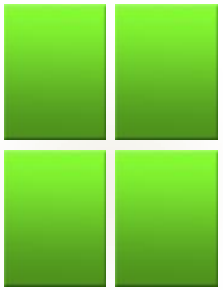
(Bigger # is better)

50% more performance with 20% less power

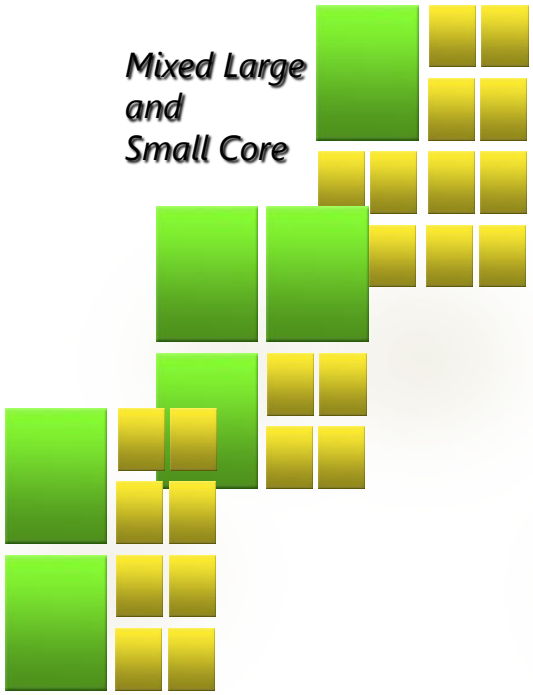
Preferable to use multiple slower devices, than one superfast device

What's Next?

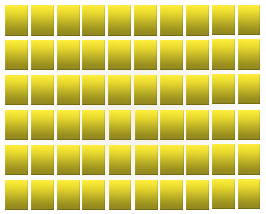
All Large Core



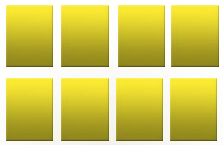
Mixed Large and Small Core



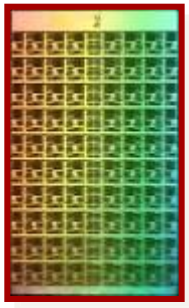
Many Small Cores



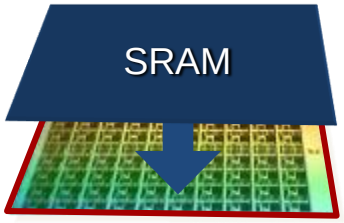
All Small Core



Many Floating-Point Cores



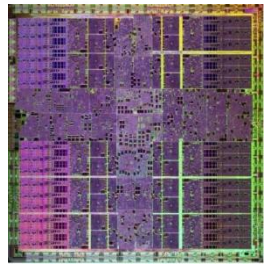
+ 3D Stacked Memory



Different Classes of Chips

- Home
- Games / Graphics
- Business
- Scientific

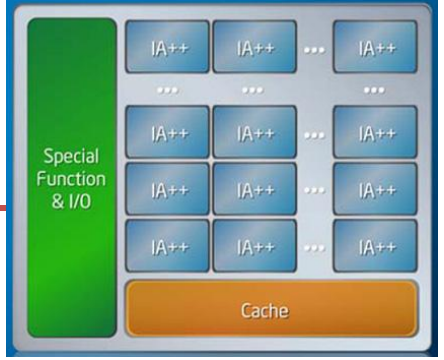
And then there's the GPGPU's NVIDIA's Tesla T10P



- T10P chip
  - 240 cores; 1.5 GHz
  - Tpeak 1 Tflop/s - 32 bit floating point
  - Tpeak 100 Gflop/s - 64 bit floating point
- S1070 board
  - 4 - T10P devices;
  - 700 Watts

- GTX 280
  - 1 – T10P; 1.3 GHz
  - Tpeak 864 Gflop/s - 32 bit floating point
  - Tpeak 86.4 Gflop/s - 64 bit floating point

Intel's Larrabee Chip



THE NEW YORK TIMES, MONDAY, AUGUST 4, 2008

Intel's Line of Graphics Chips Could Have Broader Uses

By JOHN MARKOFF

SAN FRANCISCO — **Intel** is planning to release on Monday the first technical details of a new family of chips intended to soup up computer graphics and, eventually, a broad range of computing tasks.

The new microprocessor family, code-named Larrabee, will be available in late 2009 or early 2010. Intel is releasing the details of its plans ahead of the Siggraph industry conference in Los Angeles, which starts Aug. 11.

The company said it would initially aim Larrabee at the personal-computer graphics market, where its "many-core" design, x86 instruction set, which will allow the chips to take advantage of a huge library of existing software.

In 2004, after finding that it could not make its chips faster because they were overheating, Intel adopted a strategy it referred to as a "right-hand turn." It switched to improving performance by increasing the number of processing elements, or cores, on each chip. That led first to dual-core and now quad-core chips.

Analysts said the first generation of Larrabee may have 16 to 48 cores, depending on the performance goal.

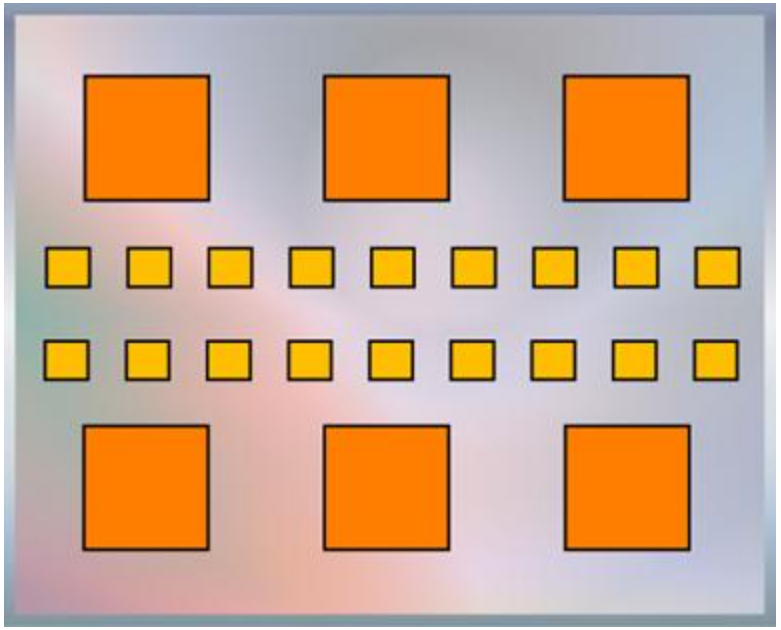
Intel has tried several approaches to chip design, but none of them have had the impact of its x86 family, which was originally introduced three decades ago. Architectures that have been less successful include the Itanium and the 432, neither of which was adopted in mainstream computing.

Instead of speed, Intel turns to improving performance.

- Many X 86 IA cores
  - Scalable to Tflop/s
- New cache architecture
- New vector instructions set
  - Vector memory operations
  - Conditionals
  - Integer and floating point arithmetic
- New vector processing unit / wide SIMD

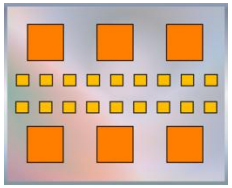
Architecture of Interest

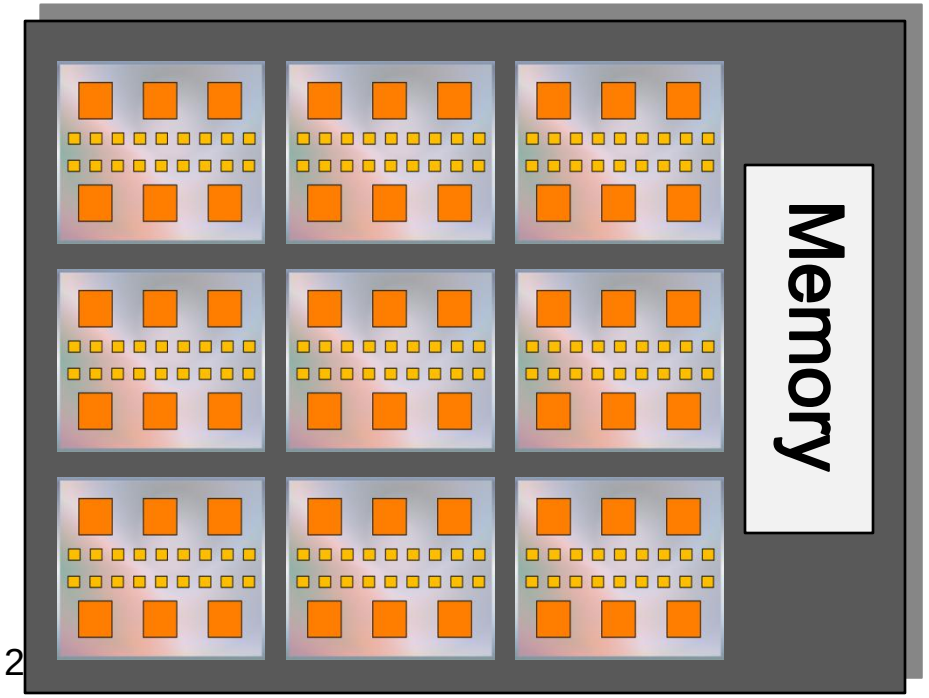
- Manycore chip
- Composed of hybrid cores
  - Some general purpose
  - Some graphics
  - Some floating point



Architecture of Interest

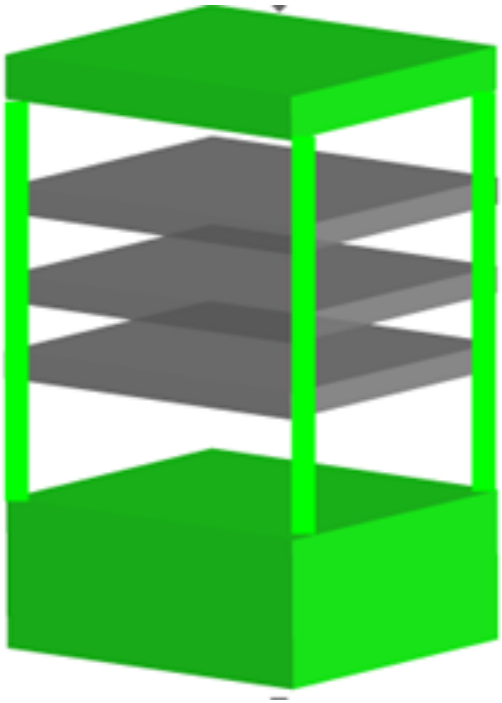
- Board composed of multiple chips sharing memory

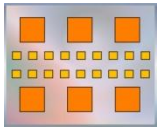


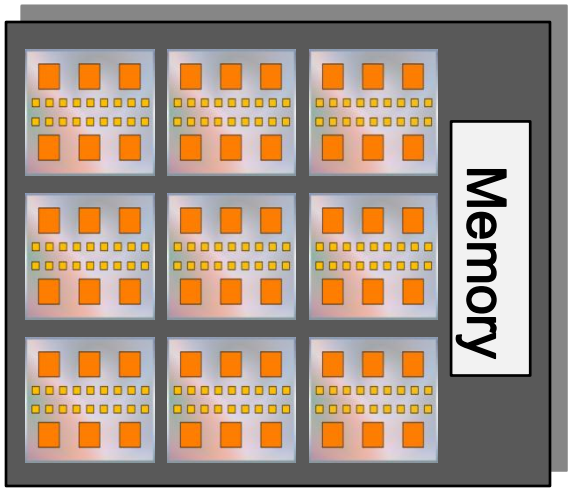


Architecture of Interest

- Rack composed of multiple boards

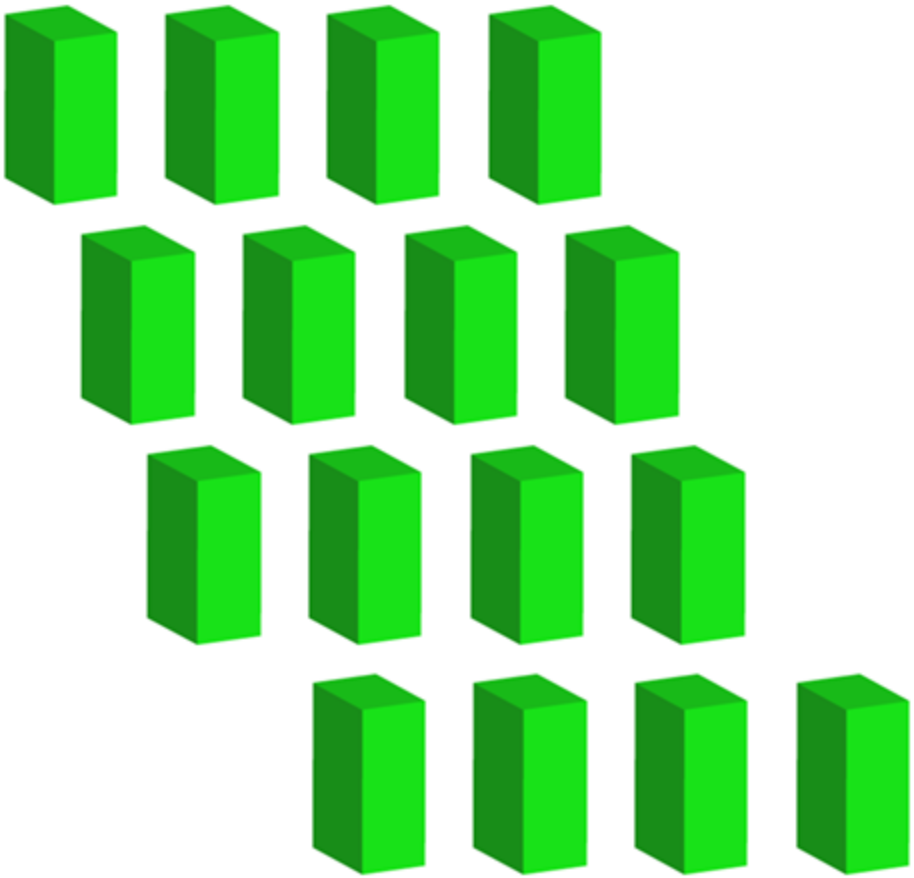


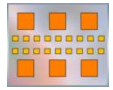


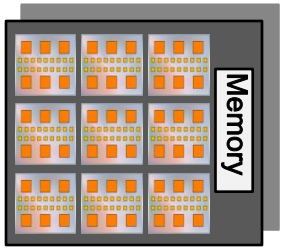


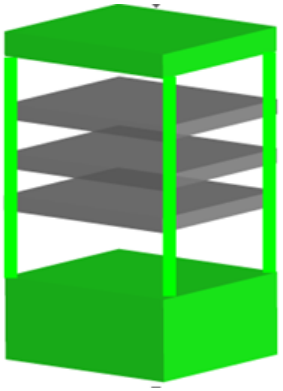
Architecture of Interest

- A room full of these racks









- Think millions of cores

Near Term Situation

- Million core systems and beyond are on the horizon
- By 2012 there will be more systems deployed in the 200K – 1M core range
- By 2020 there will be systems with perhaps 100M cores
- Personal systems with > 1000 cores within 5 years (I have over 100 cores in my office now)
- Personal systems with requirements for 1M threads is not too far fetched (think GPUs)

Exascale Computing

- Exascale systems (10^{18} Flop/s) 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
- > 10,000's of I/O channels to 10-100 Exabytes of secondary storage, disk bandwidth to storage ratios not optimal for HPC use
- Hardware and software based fault management
- Achievable performance per watt will likely be the primary measure of progress

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 coming
- 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.

Collaborators

- Top500 Team
  - Erich Strohmaier, NERSC
  - Hans Meuer, Mannheim
  - Horst Simon, NERSC

http://www.top500