An Overview of High Performance Computing
and
Future Requirements

Jack Dongarra

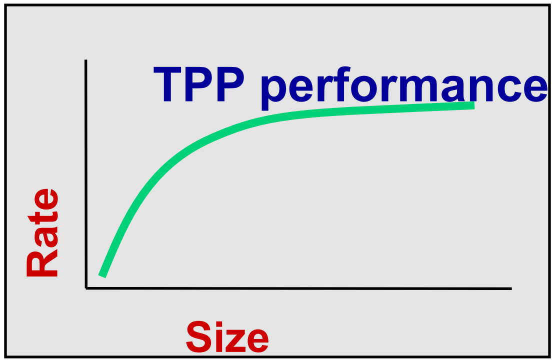
University of Tennessee
Oak Ridge National Laboratory

TOP500 superCOMPUTER

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



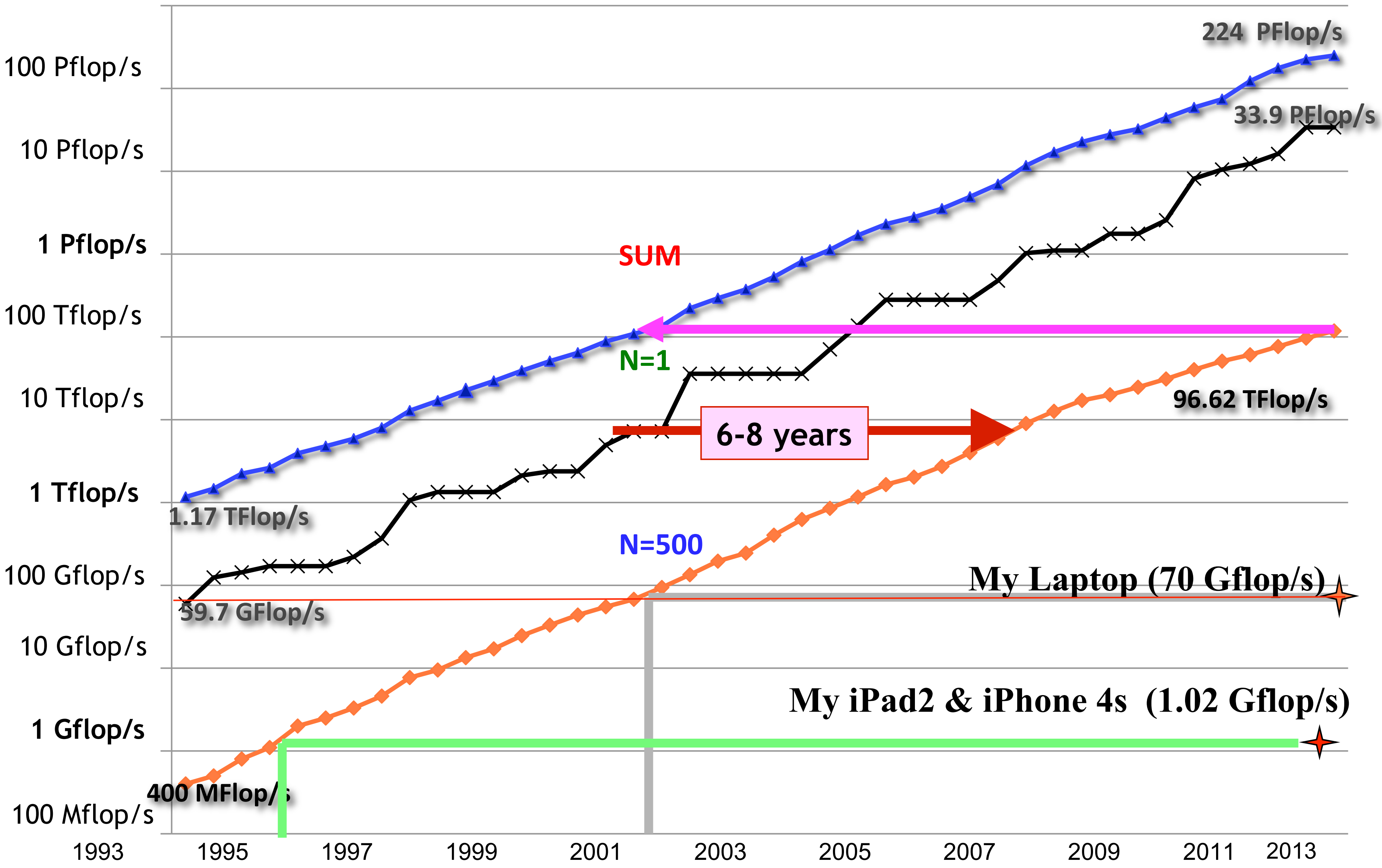
- Updated twice a year

  SC‘xy in the States in November

  Meeting in Germany in June

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

Performance Development of HPC Over the Last 20 Years



Petaflops Club 07

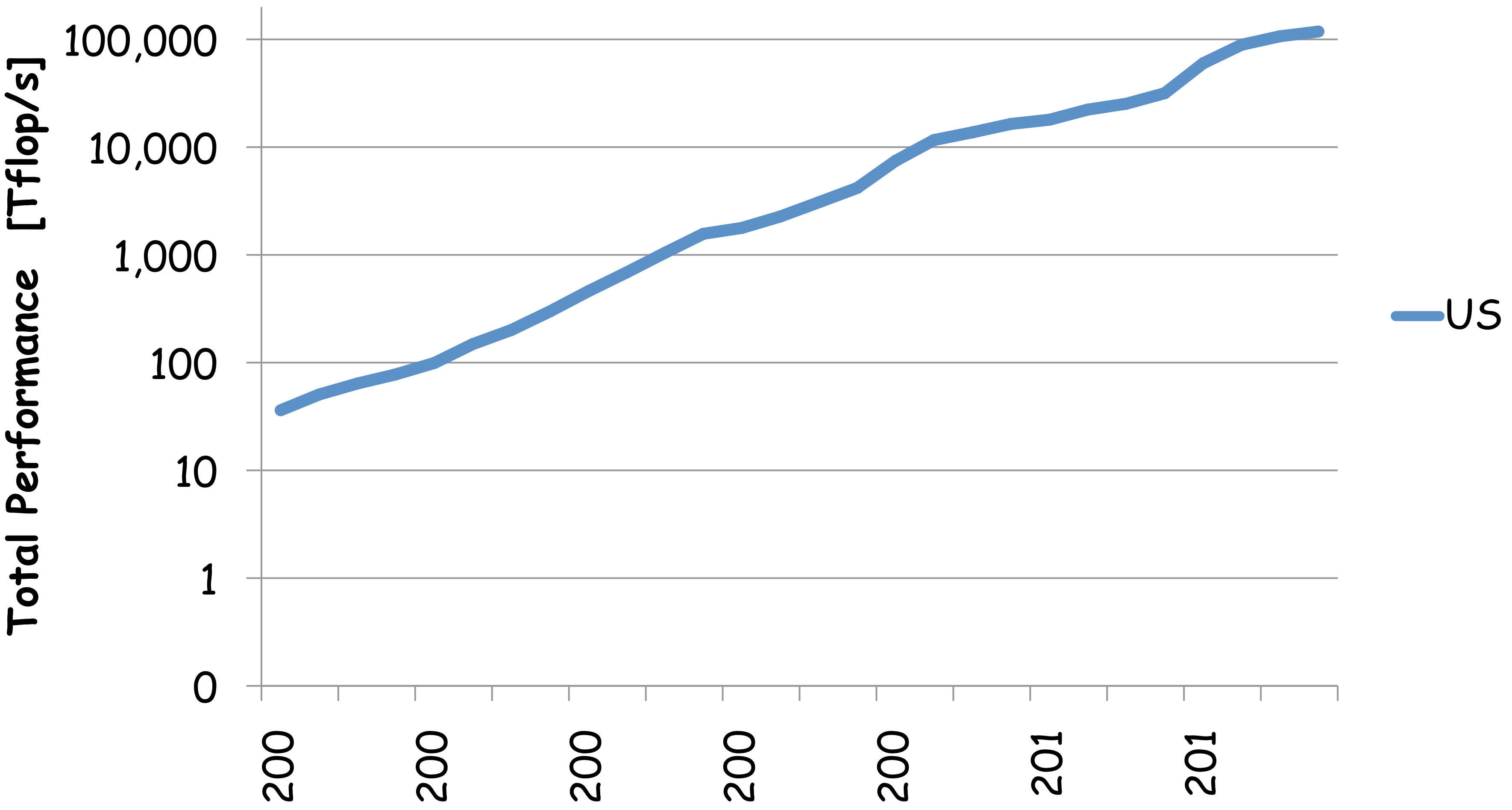
31 Systems

13 🇺🇸 4 🇯🇵 3 🇩🇪 3 🇨🇳 3 🇬🇧 3 🇫🇷 1 🇮🇹 1 🇨🇭

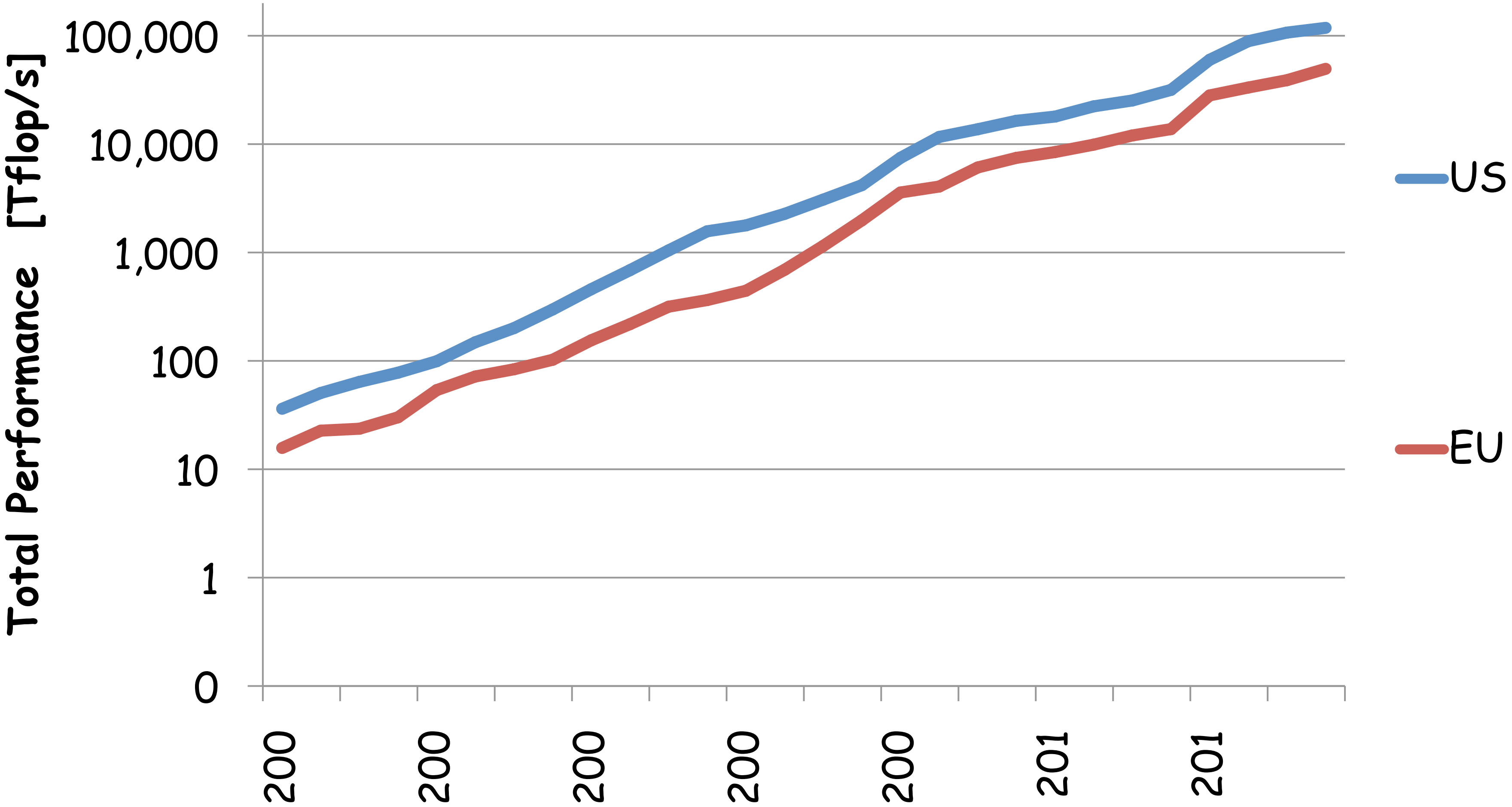
Name	Rmax Linpack# Pflops	Country	
Tianhe-2 (MilkyWay-2)	33.9	China	NUDT: Hybrid Intel/Intel/Custom
Titan	17.6	US	Cray: Hybrid AMD/Nvidia/Custom
Sequoia	17.2	US	**IBM: BG-Q/Custom**
K Computer	10.5	Japan	Fujitsu: Sparc/Custom
Mira	8.59	US	**IBM: BG-Q/Custom**
Piz Daint	6.27	Switzerland	Cray: Hybrid AMD/Nvidia/Custom
Stampede	5.17	US	Dell: Hybrid/Intel/Intel/IB
JUQUEEN	5.01	Germany	**IBM: BG-Q/Custom**
Vulcan	4.29	US	**IBM: BG-Q/Custom**
SuperMUC	2.9	Germany	IBM: Intel/IB
TSUBAME 2.5	2.84	Japan	Cluster Pltf: Hybrid Intel/Nvidia/IB
Tianhe-1A	2.57	China	NUDT: Hybrid Intel/Nvidia/Custom
cascade	2.35	US	Atipa: Hybrid Intel/Intel/IB
Pangea	2.1	France	Bull: Intel/IB
Fermi	1.79	Italy	**IBM: BG-Q/Custom**
Pleiades	1.54	US	SGI Intel/IB
DARPA Trial Subset	1.52	US	IBM: Intel/IB
Spirit	1.42	US	SGI: Intel/IB
ARCHER	1.37	UK	Cray: Intel/Custom
Curie thin nodes	1.36	France	Bull: Intel/IB
Nebulae	1.27	China	Dawning: Hybrid Intel/Nvidia/IB
Yellowstone	1.26	US	**IBM: BG-Q/Custom**
Blue Joule	1.25	UK	**IBM: BG-Q/Custom**
Helios	1.24	Japan	Bull: Intel/IB
Garnet	1.17	US	Cray: AMD/Custom
Cielo	1.11	US	Cray: AMD/Custom
DiRAC	1.07	UK	**IBM: BG-Q/Custom**
Hopper	1.05	US	Cray: AMD/Custom
Tera-100	1.05	France	Bull: Intel/IB
Oakleaf-FX	1.04	Japan	Fujitsu: Sparc/Custom
MPI	1.03	Germany	iDataFlex: Intel/IB

8 Hybrid Architectures
8 IBM BG/Q
18 Custom X
12 Infiniband X
9 Look like “clusters”

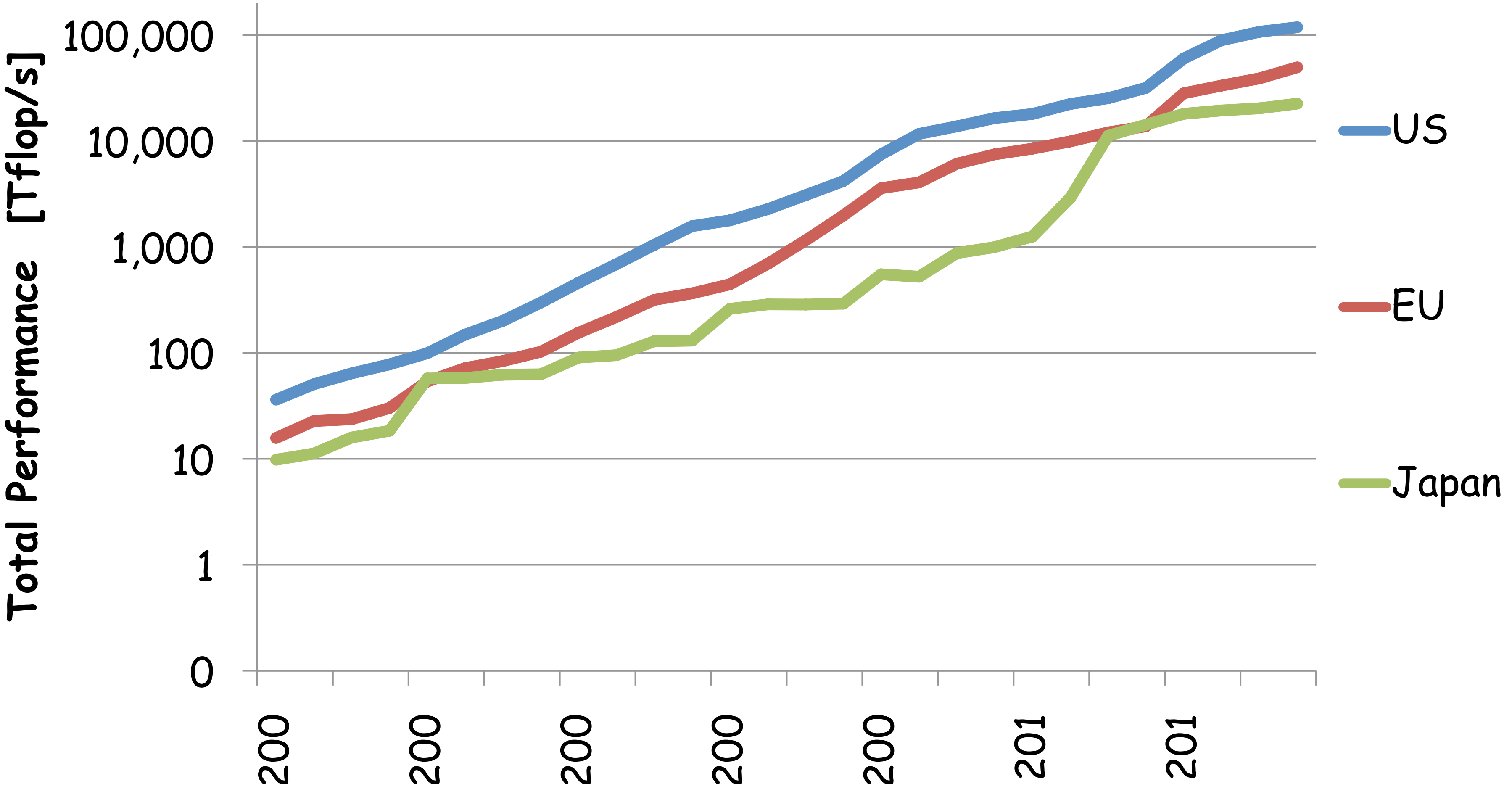
Performance of Countries



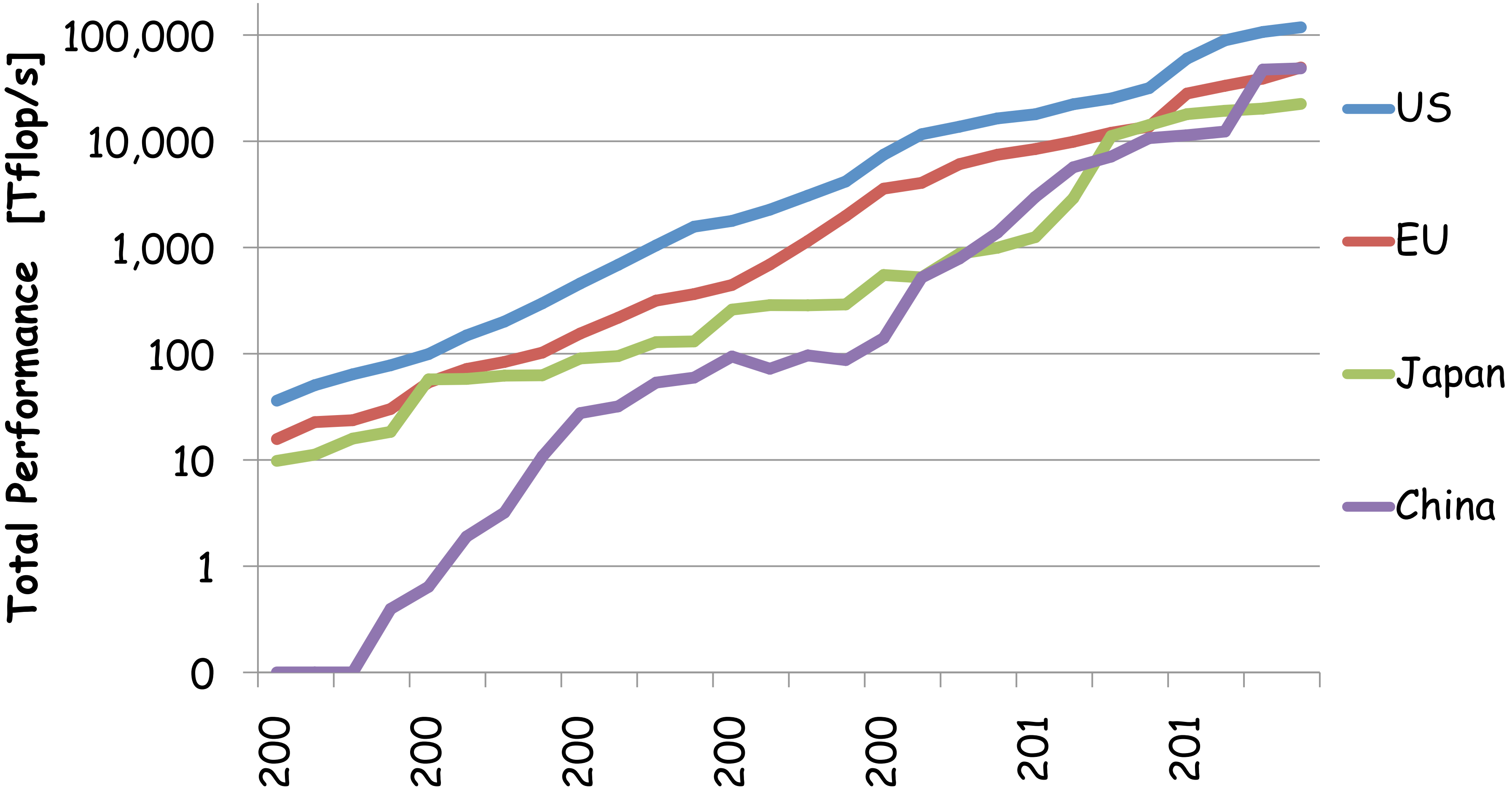
Performance of Countries



Performance of Countries



Performance of Countries



November 2013: The TOP10

Rank	Site	Computer	Country	Cores	Rmax [Pflops]	% of Peak	Power [MW]	MFlops /Watt
1	National University of Defense Technology	Tianhe-2 NUDT, Xeon 12C 2.2GHz + IntelXeon Phi (57c) + Custom	China	3,120,000	33.9	62	17.8	1905
2	DOE / OS Oak Ridge Nat Lab	Titan, Cray XK7 (16C) + Nvidia Kepler GPU (14c) + Custom	USA	560,640	17.6	65	8.3	2120
3	DOE / NNSA L Livermore Nat Lab	Sequoia, BlueGene/Q (16c) + custom	USA	1,572,864	17.2	85	7.9	2063
4	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx (8c) + Custom	Japan	705,024	10.5	93	12.7	827
5	DOE / OS Argonne Nat Lab	Mira, BlueGene/Q (16c) + Custom	USA	786,432	8.16	85	3.95	2066
6	Swiss CSCS	Piz Daint, Cray XC30, Xeon 8C + Nvidia Kepler (14c) + Custom	Swiss	115,984	6.27	81	2.3	2726
7	Texas Advanced Computing Center	Stampede, Dell Intel (8c) + Intel Xeon Phi (61c) + IB	USA	204,900	2.66	61	3.3	806
8	Forschungszentrum Juelich (FZJ)	JuQUEEN, BlueGene/Q, Power BQC 16C 1.6GHz+Custom	Germany	458,752	5.01	85	2.30	2178
9	DOE / NNSA L Livermore Nat Lab	Vulcan, BlueGene/Q, Power BQC 16C 1.6GHz+Custom	USA	393,216	4.29	85	1.97	2177
10	Leibniz Rechenzentrum	SuperMUC, Intel (8c) + IB	Germany	147,456	2.90	91*	3.42	848
500	Banking	HP	USA	22,212	.118	50		

Commodity plus Accelerator Today

Commodity

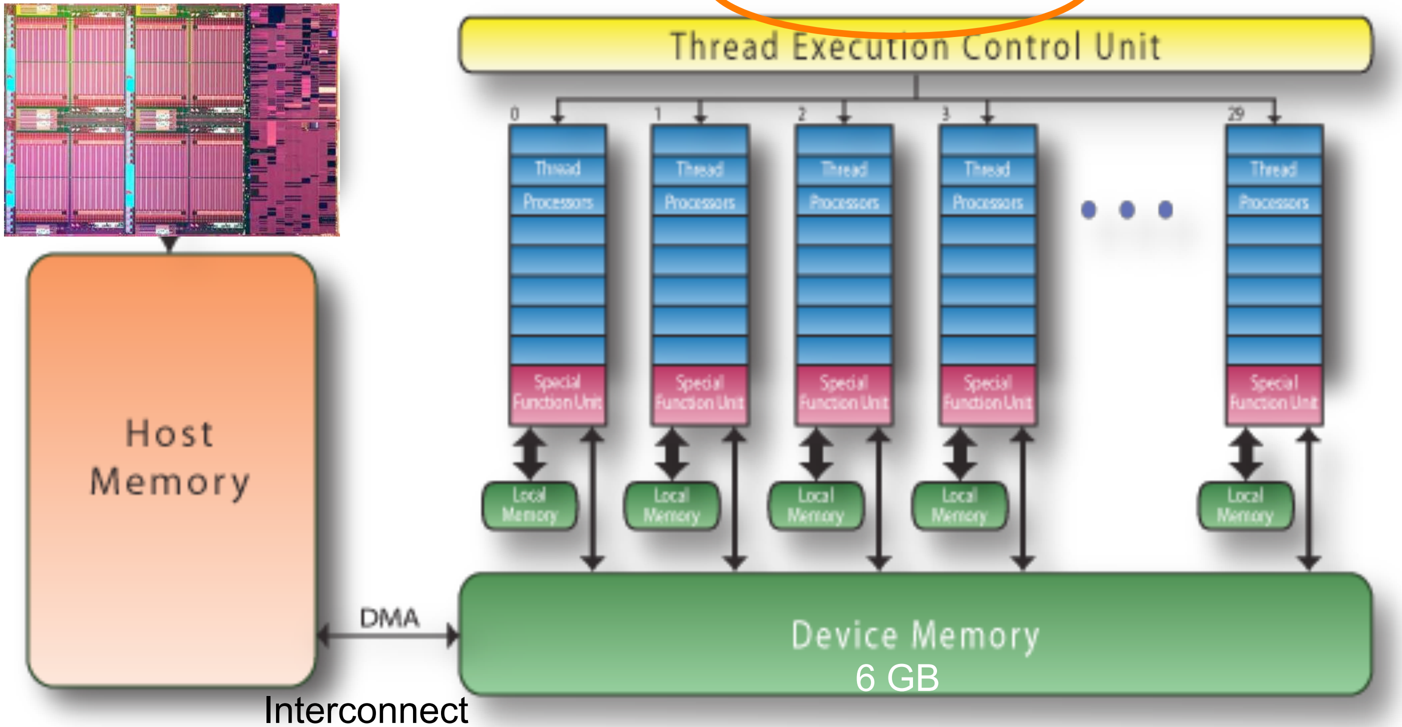
Intel Xeon
8 cores
3 GHz
8*4 ops/cycle
96 Gflop/s (DP)

Accelerator (GPU)

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

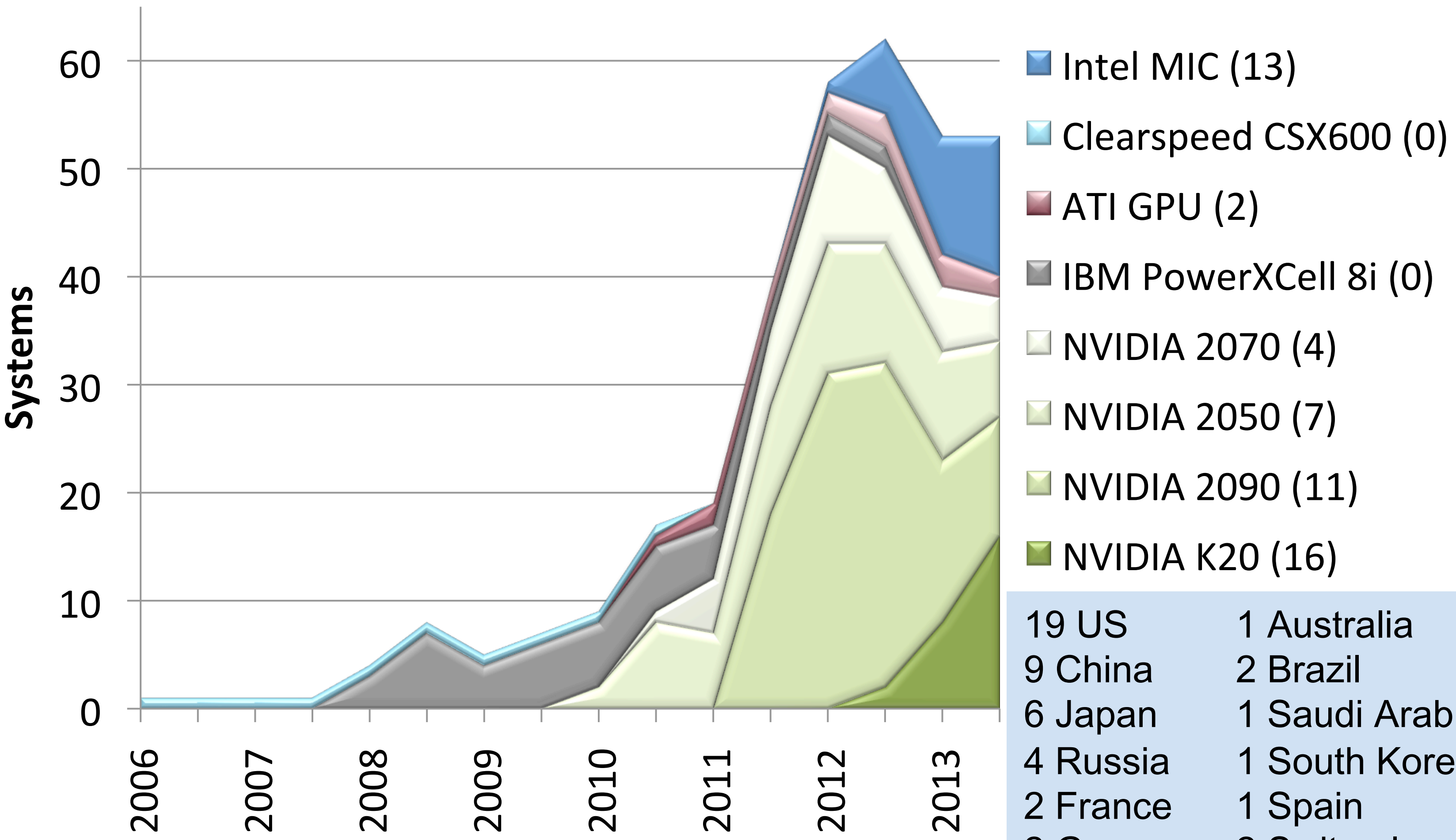
192 Cuda cores/SMX

PCI Express 3.0 Host Interface
GigaThread Engine
SMX
L2 Cache
Memory Controller



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

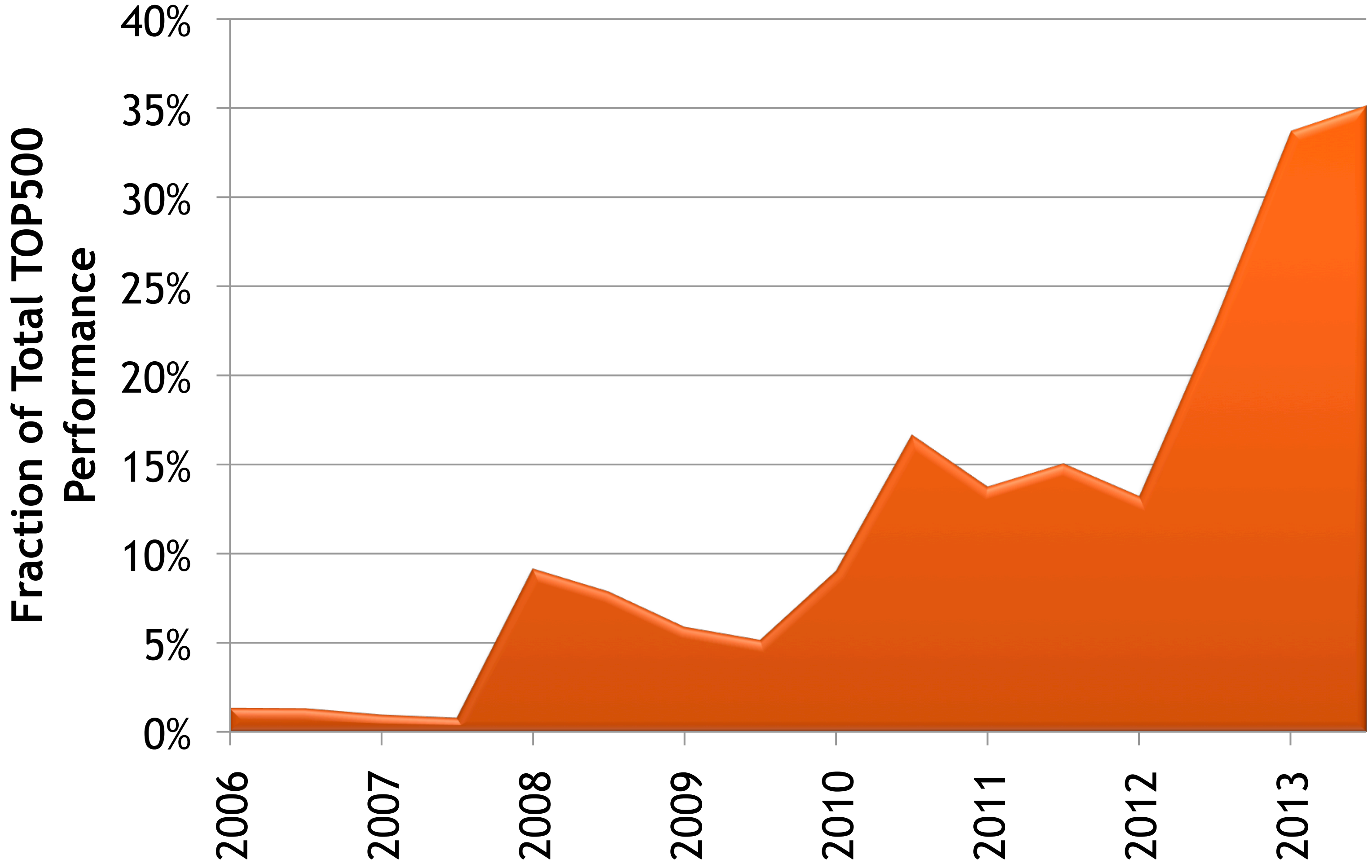
Accelerators (53 systems)



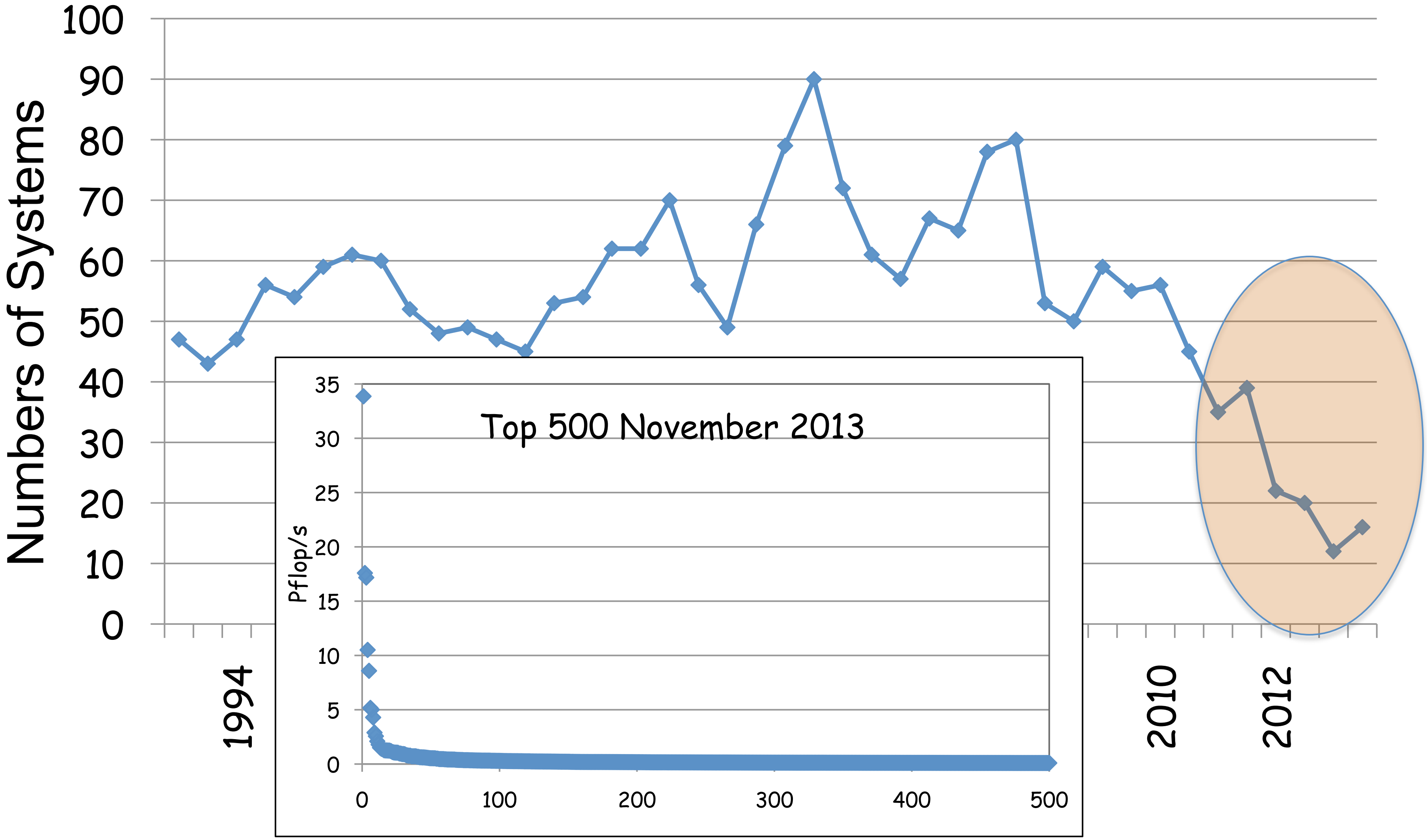
19 US
9 China
6 Japan
4 Russia
2 France
2 Germany
2 India
1 Italy
1 Poland

1 Australia
2 Brazil
1 Saudi Arabia
1 South Korea
1 Spain
2 Switzerland
1 UK

Top500 Performance Share of Accelerators



For the Top 500: Rank at which Half of Total Performance is Accumulated



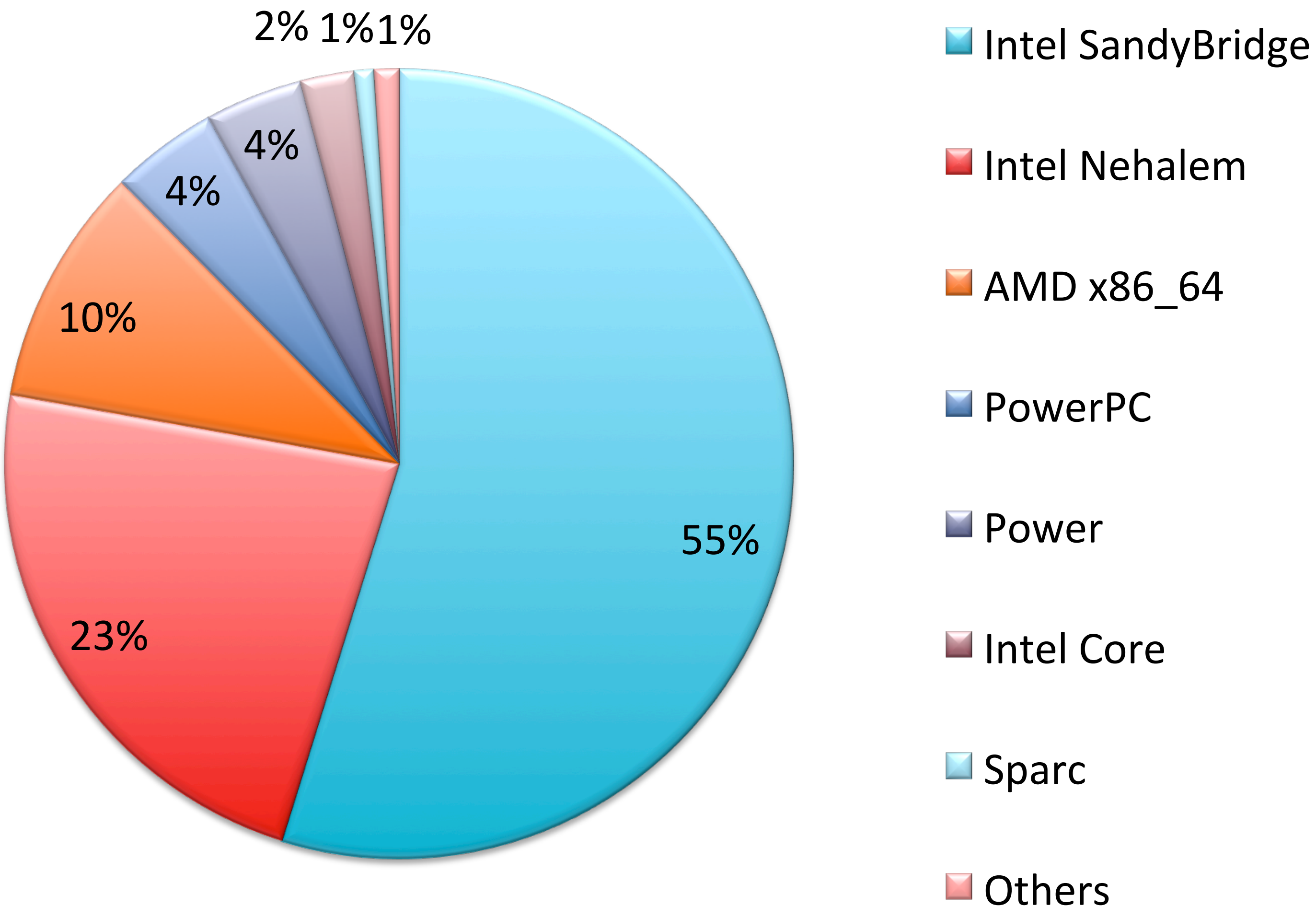
#1 System on the Top500 Over the Past 20 Years (16 machines in that club)

http://bit.ly/hpcg-benchmark

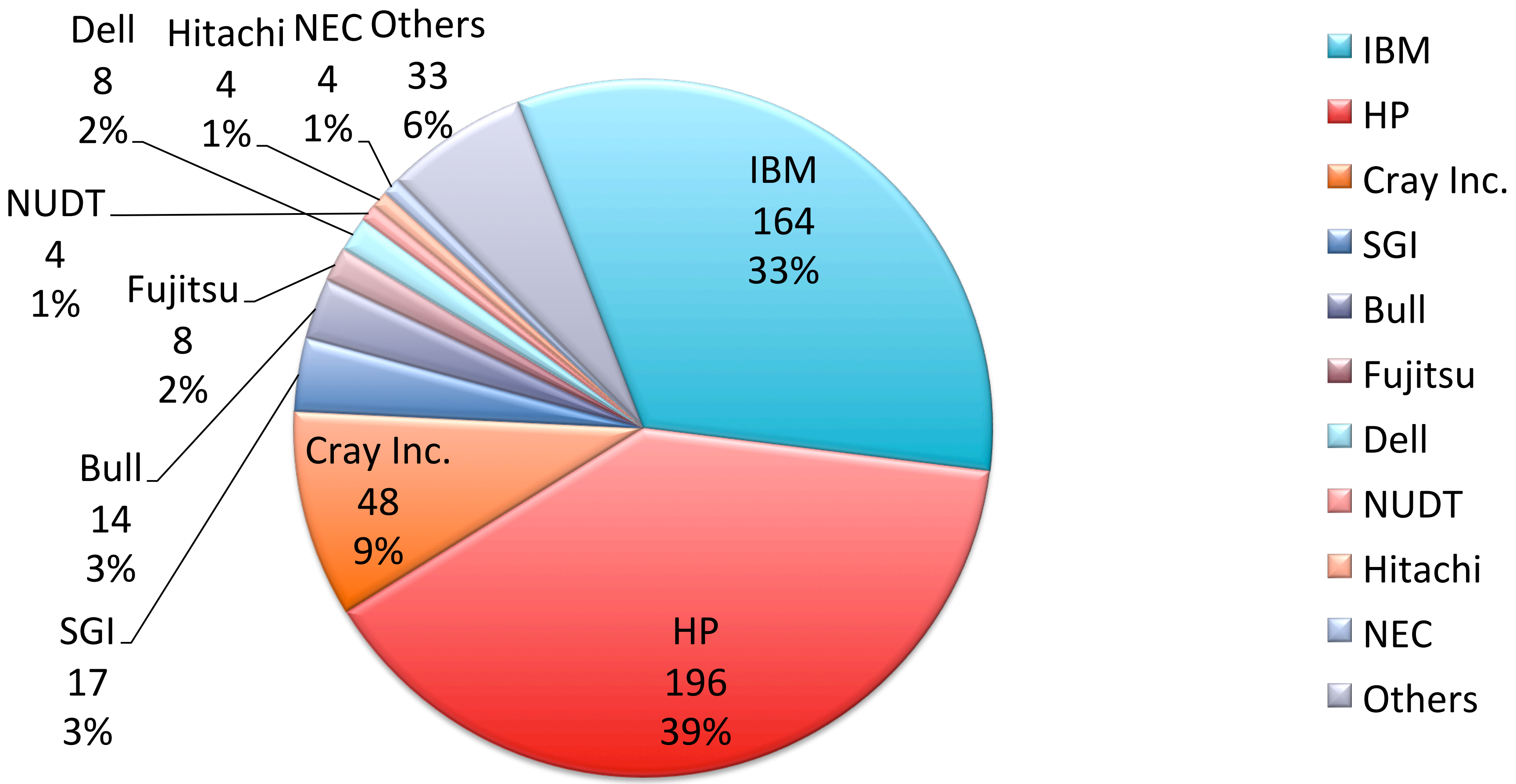
9 🇺🇸 6 🇯🇵 2 🇨🇳

Top500 List	Computer	r_max (Tflop/s)	n_max	Hours	MW
6/93 (1)	TMC CM-5/1024	.060	52224	0.4	
11/93 (1)	Fujitsu Numerical Wind Tunnel	.124	31920	0.1	1.
6/94 (1)	Intel XP/S140	.143	55700	0.2	
11/94 - 11/95 (3)	Fujitsu Numerical Wind Tunnel	.170	42000	0.1	1.
6/96 (1)	Hitachi SR2201/1024	.220	138,240	2.2	
11/96 (1)	Hitachi CP-PACS/2048	.368	103,680	0.6	
6/97 - 6/00 (7)	Intel ASCI Red	2.38	362,880	3.7	.85
11/00 - 11/01 (3)	IBM ASCI White, SP Power3 375 MHz	7.23	518,096	3.6	
6/02 - 6/04 (5)	NEC Earth-Simulator	35.9	1,000,000	5.2	6.4
11/04 - 11/07 (7)	IBM BlueGene/L	478.	1,000,000	0.4	1.4
6/08 - 6/09 (3)	IBM Roadrunner –PowerXCell 8i 3.2 Ghz	1,105.	2,329,599	2.1	2.3
11/09 - 6/10 (2)	Cray Jaguar - XT5-HE 2.6 GHz	1,759.	5,474,272	17.3	6.9
11/10 (1)	NUDT Tianhe-1A, X5670 2.93Ghz NVIDIA	2,566.	3,600,000	3.4	4.0
6/11 - 11/11 (2)	Fujitsu K computer, SPARC64 VIIIfx	10,510.	11,870,208	29.5	9.9
6/12 (1)	IBM Sequoia BlueGene/Q	16,324.	12,681,215	23.1	7.9
11/12 (1)	Cray XK7 Titan AMD + NVIDIA Kepler	17,590.	4,423,680	0.9	8.2
6/13 – 11/13(?)	NUDT Tianhe-2 Intel IvyBridge & Xeon Phi	33,862.	9,960,000	5.4	17.8

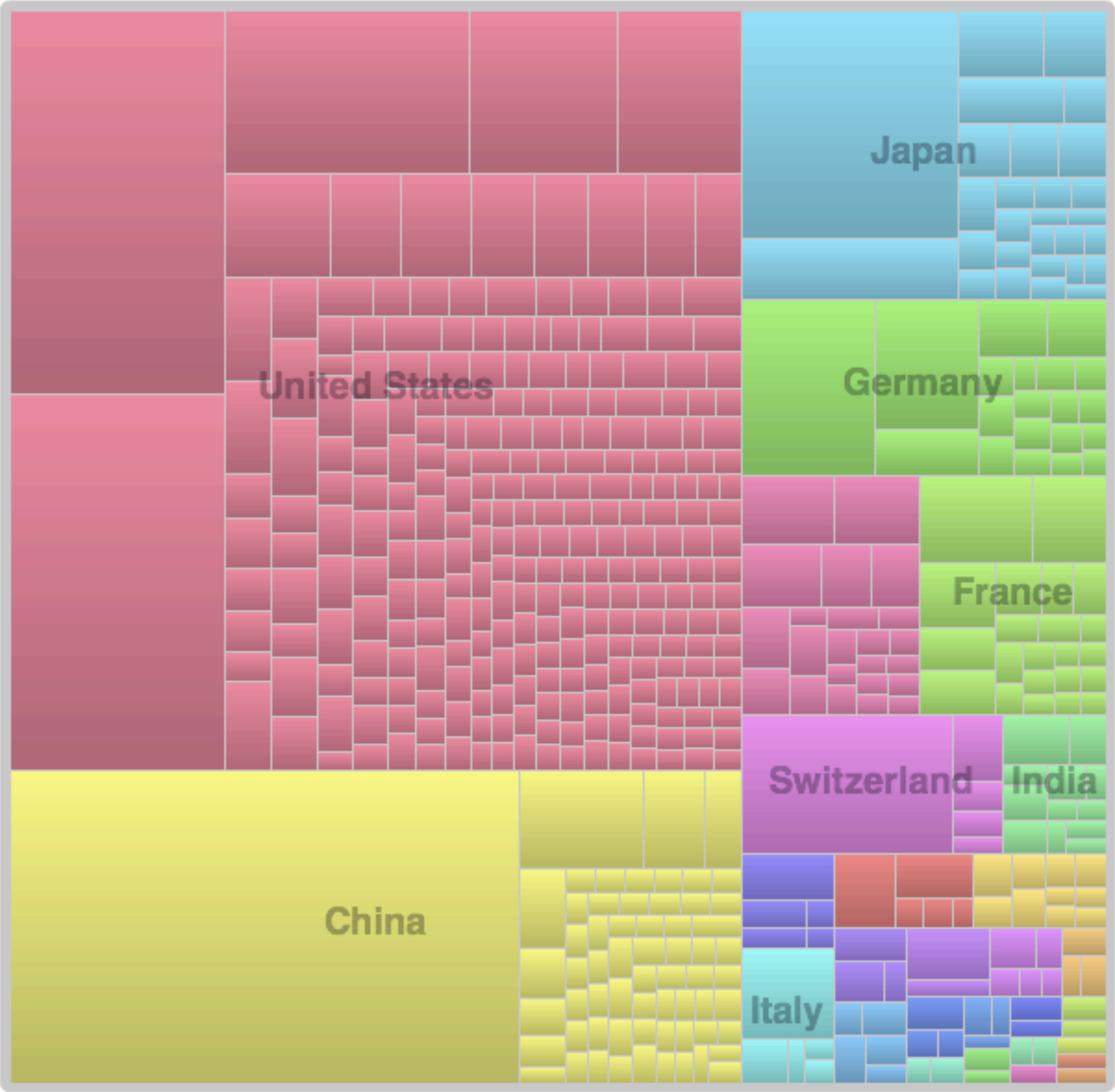
Processors / Systems



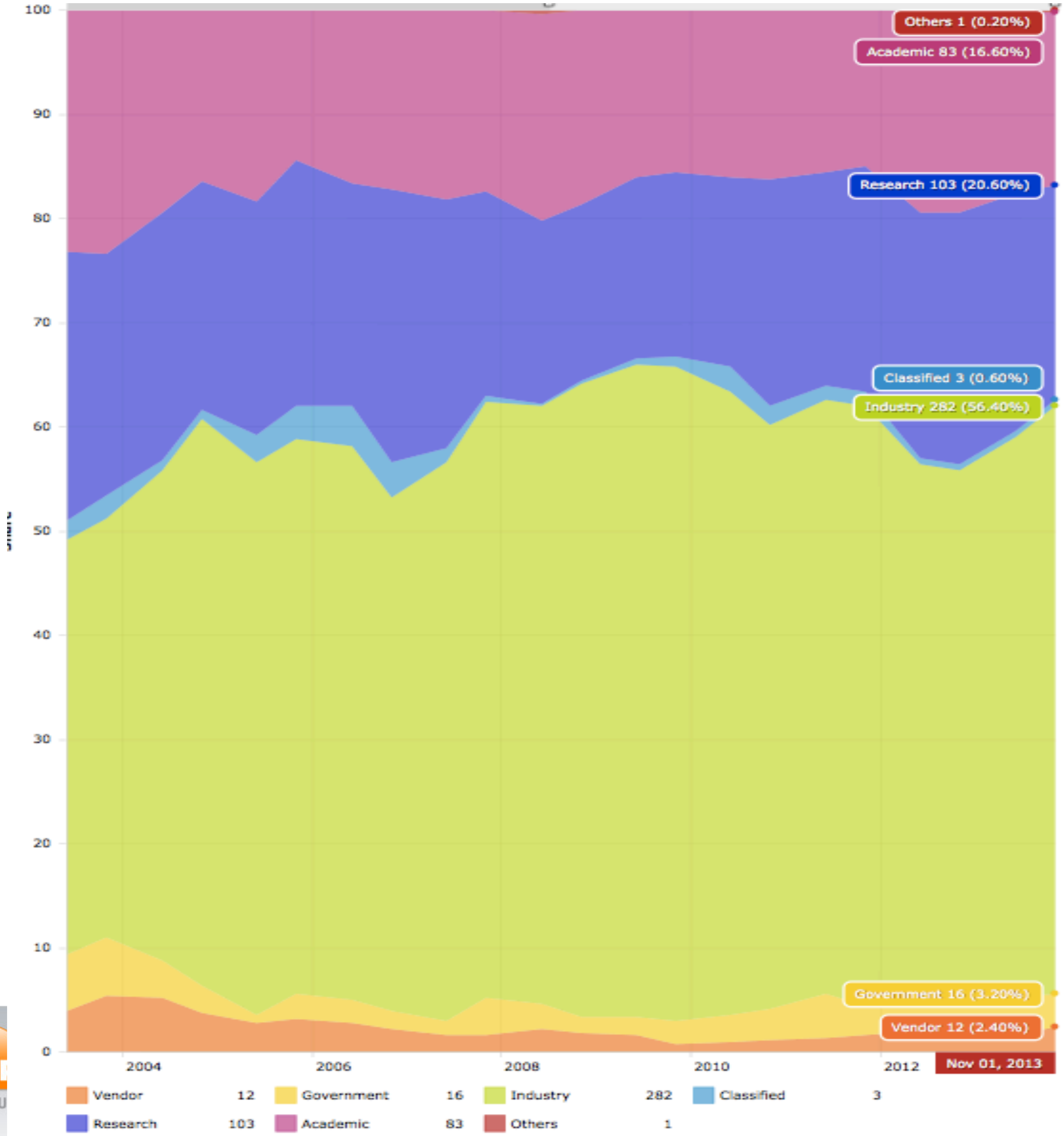
Vendors / System Share



Countries Share



Absolute Counts	
US:	267
China:	63
Japan:	28
UK:	23
France:	22
Germany:	20

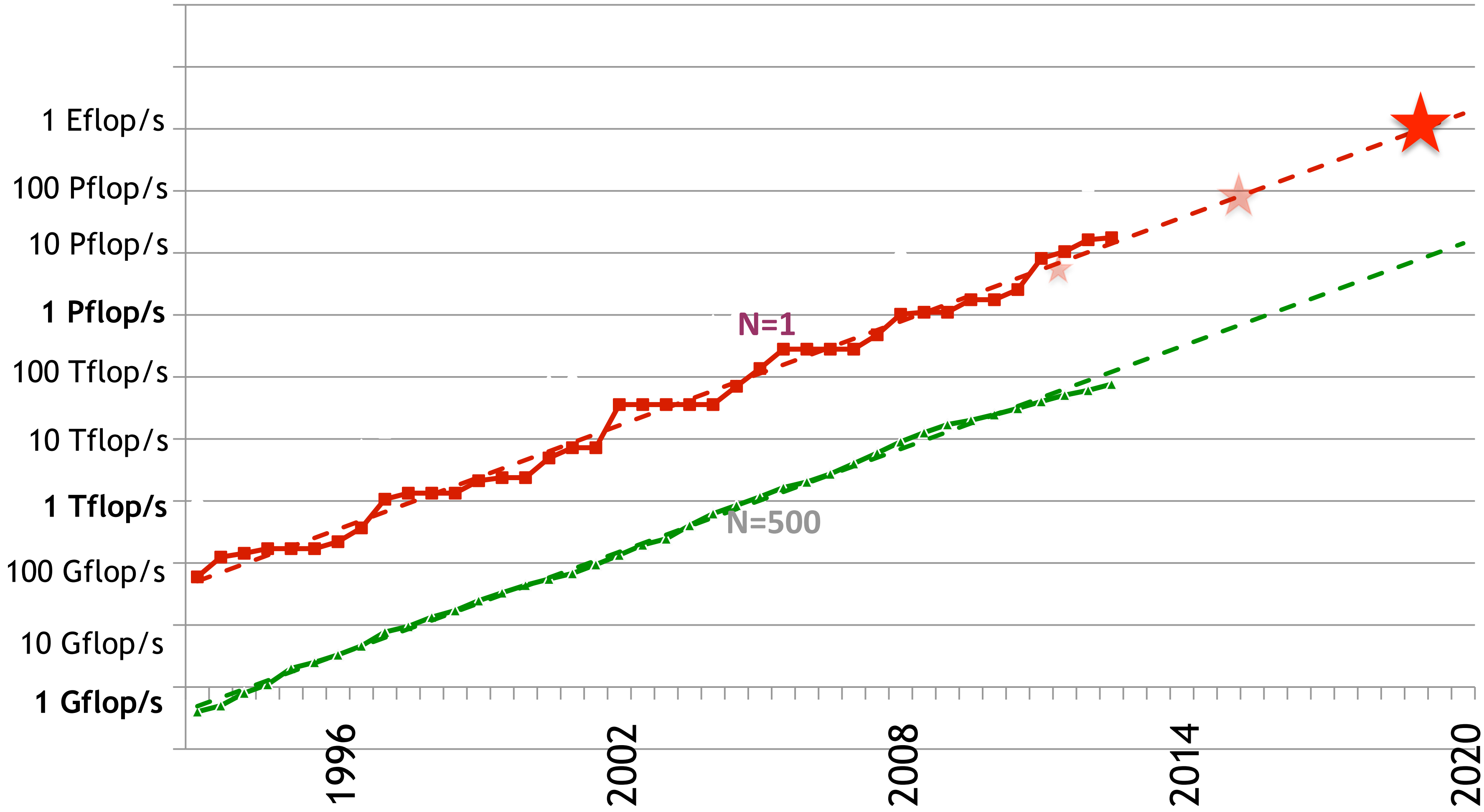

100
90
80
70
60
50
40
30
20
10
0
Others 1 (0.20%)
Academic 83 (16.60%)
Research 103 (20.60%)
Classified 3 (0.60%)
Industry 282 (56.40%)
Government 16 (3.20%)
Vendor 12 (2.40%)
2004
2006
2008
2010
2012
Nov 01, 2013
Vendor 12
Government 16
Industry 282
Classified 3
Research 103
Academic 83
Others 1

8%

20%

56%

Performance Development in Top500



Today's #1 System

Systems	2013 Tianhe-2
System peak	**55 Pflop/s**
Power	**18 MW** **(3 Gflops/W)**
System memory	1.4 PB (1.024 PB CPU + .384 PB CoP)
Node performance	3.43 TF/s (.4 CPU +3 CoP)
Node concurrency	24 cores CPU + 171 cores CoP
Node Interconnect BW	6.36 GB/s
System size (nodes)	16,000
Total concurrency	3.12 M 12.48M threads (4/core)
MTTF	Few / day

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

Systems	2013 Tianhe-2
System peak	**55 Pflop/s**
Power	**18 MW** **(3 Gflops/W)**
System memory	1.4 PB (1.024 PB CPU + .384 PB CoP)
Node performance	3.43 TF/s (.4 CPU +3 CoP)
Node concurrency	24 cores CPU + 171 cores CoP
Node Interconnect BW	6.36 GB/s
System size (nodes)	16,000
Total concurrency	3.12 M 12.48M threads (4/core)
MTTF	Few / day

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

Systems	2013 Tianhe-2	2020-2022	Difference Today & Exa
System peak	**55 Pflop/s**	**1 Eflop/s**	**~20x**
Power	**18 MW** **(3 Gflops/W)**	**~20 MW** **(50 Gflops/W)**	**O(1)** **~15x**
System memory	1.4 PB (1.024 PB CPU + .384 PB CoP)	32 - 64 PB	~50x
Node performance	3.43 TF/s (.4 CPU +3 CoP)	1.2 or 15TF/s	O(1)
Node concurrency	24 cores CPU + 171 cores CoP	O(1k) or 10k	~5x - ~50x
Node Interconnect BW	6.36 GB/s	200-400GB/s	~40x
System size (nodes)	16,000	O(100,000) or O(1M)	~6x - ~60x
Total concurrency	3.12 M 12.48M threads (4/core)	O(billion)	~100x
MTTF	Few / day	Many / day	O(?)

High Performance Linpack (HPL)

- Is a **widely recognized** and discussed metric for ranking high performance computing systems
- When HPL gained prominence as a performance metric in the early 1990s there **was a strong correlation between its predictions of system rankings and the ranking that full-scale applications would realize**.
- **Computer system vendors pursued designs that would increase their HPL performance**, which would in turn improve overall application performance.
- Today HPL remains **valuable as a measure of historical trends**, and as a stress test, especially for leadership class systems that are pushing the boundaries of current technology.

The Problem

- HPL performance of computer systems are **no longer so strongly correlated to real application performance**, especially for the broad set of HPC applications governed by partial differential equations.

- **Designing a system for good HPL performance can actually lead to design choices that are wrong** for the real application mix, or add unnecessary components or complexity to the system.

Concerns

- The **gap between HPL predictions and real application performance will increase** in the future.
- A computer system with the potential to run **HPL at 1 Exaflops is a design that may be very unattractive for real applications.**
- Future **architectures targeted toward good HPL performance will not be a good match for most applications**.
- This leads us to a think about a different metric

Proposal: HPCG

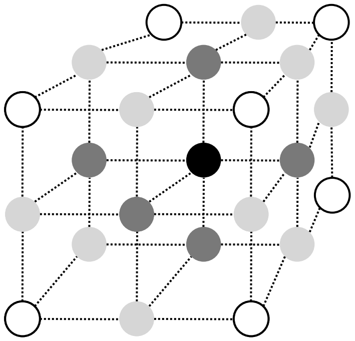
- High Performance Conjugate Gradient (HPCG).
- Solves *Ax=b*, *A* large, sparse, *b* known, *x* computed.
- An optimized implementation of PCG contains essential computational and communication patterns that are prevalent in a variety of methods for discretization and numerical solution of PDEs

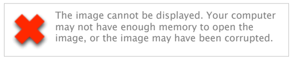
- Patterns:
  - Dense and sparse computations.
  - Dense and sparse collective.
  - Data-driven parallelism (unstructured sparse triangular solves).
- Strong verification and validation properties (via spectral properties of CG).

http://bit.ly/hpcg-benchmark

HPCG Details

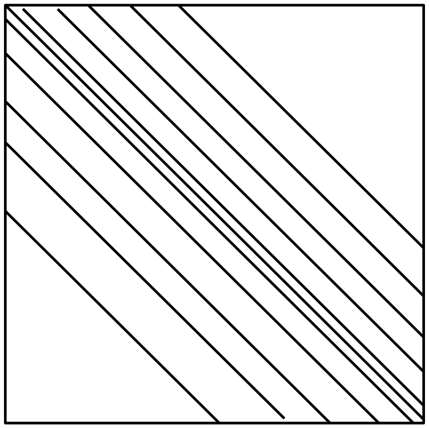
3D Laplacian discretization

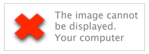




The image cannot be displayed. Your computer may not have enough memory to open the image, or the image may have been corrupted.

Sparse matrix based on 27-point stencil





The image cannot be displayed. Your computer

Preconditioned Conjugate Gradient solver

```
$p_0 := x_0, r_0 := b - A \times p_0$
Loop $i$ = 1, 2, ...
        $z_i := M^{-1} \times r_{i-1}$
        if $i = 1$
                $p_i := z_i$
                $\alpha_i :=$ dot_product($r_{i-1}, z_i$)
        else
                $\alpha_i :=$ dot_product($r_{i-1}, z_i$)
                $\beta_i := \alpha_i / \alpha_{i-1}$
                $p_i := \beta_i \times p_{i-1} + z_i$
        end if
        $\alpha_i :=$ dot_product($r_{i-1}, z_i$)/dot_product($p_i, Ap_i$)
        $x_{i+1} := x_i + \alpha_i \times p_i$
        $r_i := r_{i-1} - \alpha_i \times A \times p_i$
        if $||r_i||_2 < tolerance$ then Stop
end Loop
```

Computational Kernels

- `DotProduct()`
  - Vector dot-product
  - $y = \Sigma\ x_i \times y_i$
  - User optimization allowed: YES
- `SpMV()`
  - Sparse Matrix-Vector multiply
  - $y = A \times x$
  - User optimization allowed: YES
- `SymGS()`
  - Symmetric Gauss-Sidel
  - $z = M^{-1} \times x$
  - User optimization allowed: YES
- `WAXPBY()`
  - Scalar times vector plus scalar times vector
  - $w_i = \alpha \times x_i + \beta \times y_i$
  - User optimization allowed: YES

Verification Procedures

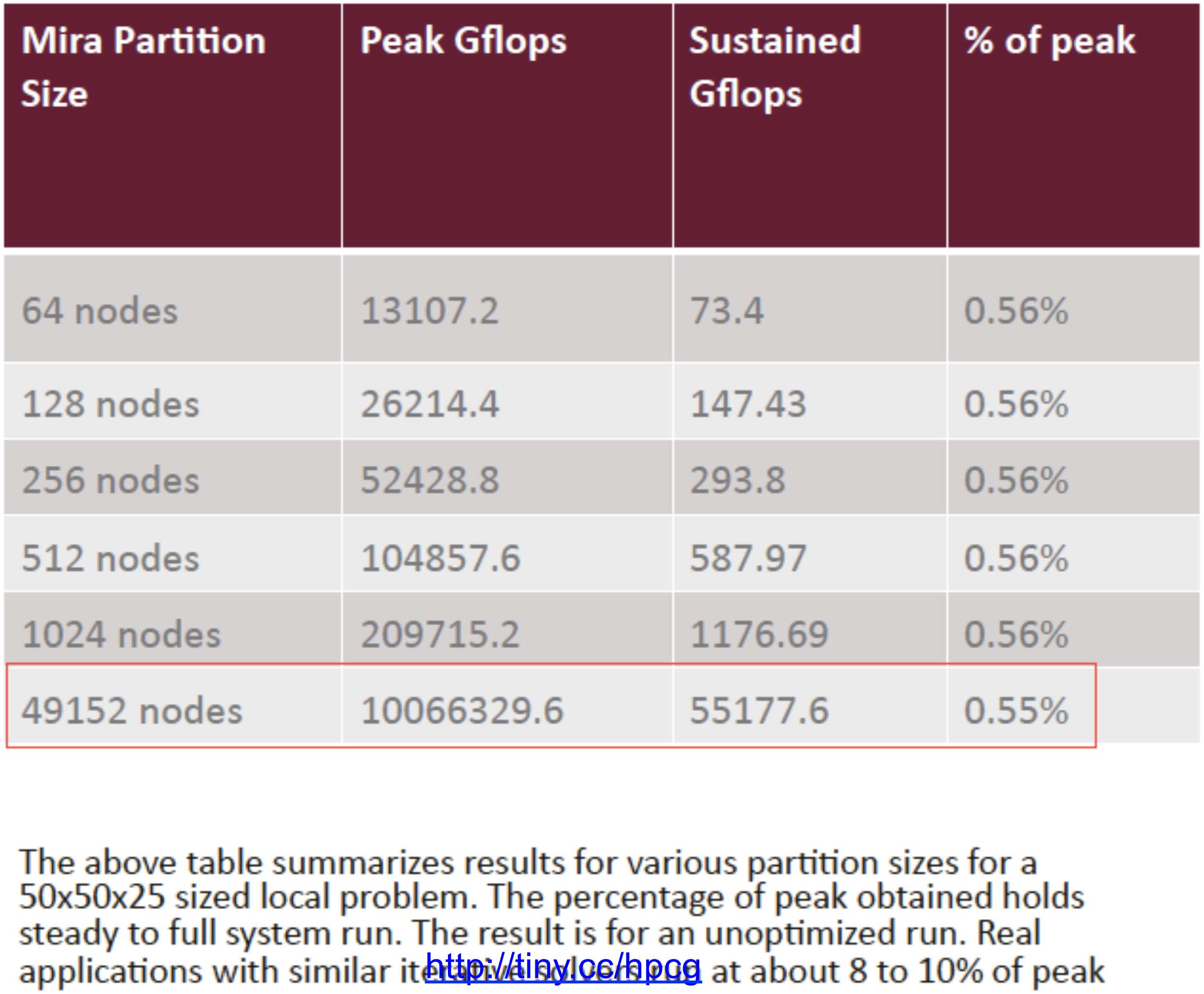
- Symmetry test
  - SpMV: $||x^{\dagger}Ay - y^{\dagger}Ax||_2$
  - SymGS: $||x^{\dagger}M^{-1}y - y^{\dagger}M^{-1}x||_2$
- CG convergence test
  - Convergence for diagonally dominant matrices should be fast
  - If $A' = A+\mathrm{diag}(A)\times 10^6$ then $x=CG(A', b, \mathrm{iterations}=12)$ and $||A'\times x-b||_2 < \varepsilon$
- Variance test
  - Repeated CG runs should yield similar residual norms despite different behavior due to runtime factors such as thread parallelism
  - $\mathrm{Variance}(||Ax^{(i)} - b||_2)$

HPCG and HPL

- We are NOT proposing to eliminate HPL as a metric.
- The historical importance and community outreach value is too important to abandon.
- HPCG will serve as an alternate ranking of the Top500.
  - ➢Similar perhaps to the Green500 listing.

http://bit.ly/hpcg-benchmark

Preliminary results

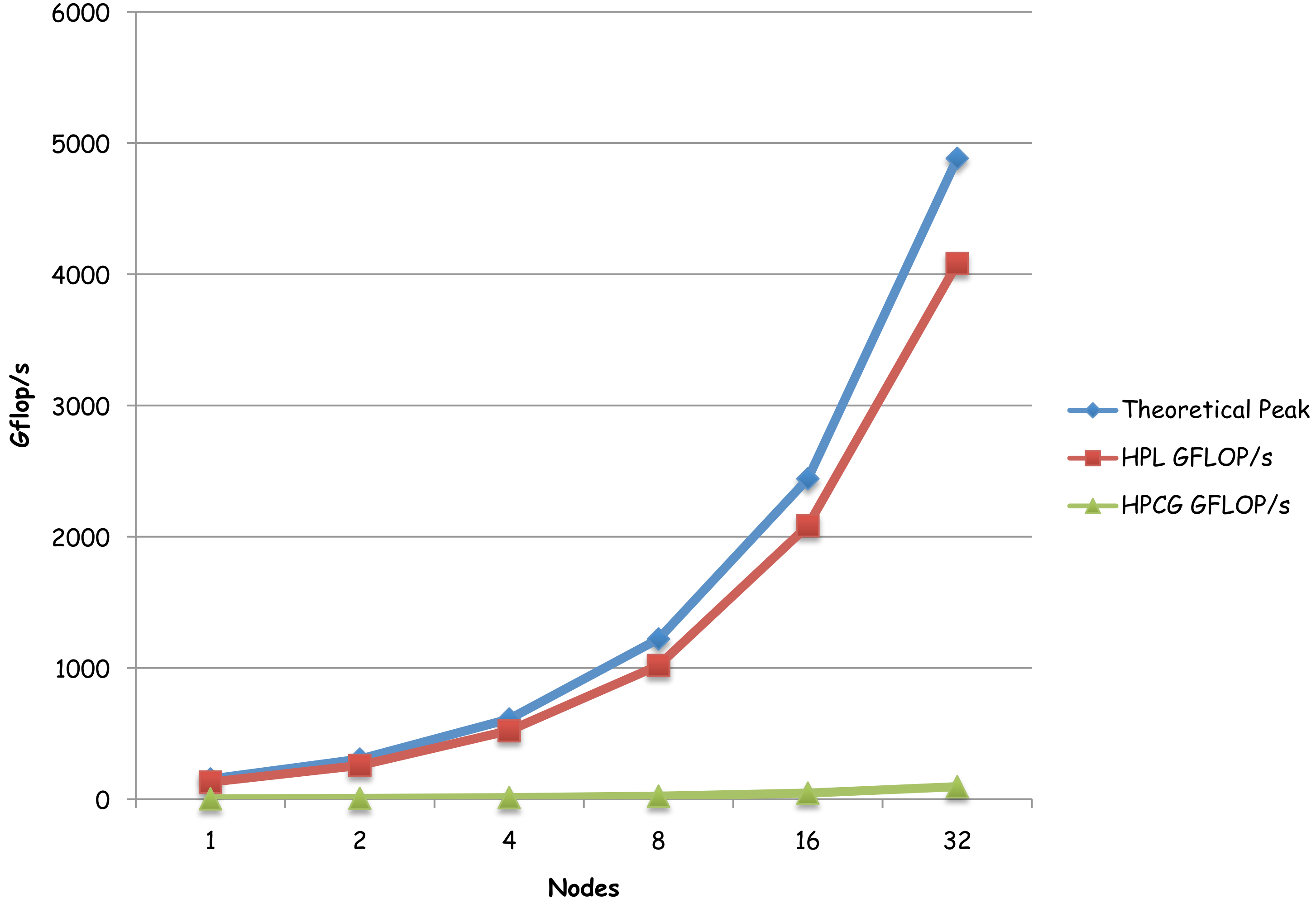


Mira Partition Size	Peak Gflops	Sustained Gflops	% of peak
64 nodes	13107.2	73.4	0.56%
128 nodes	26214.4	147.43	0.56%
256 nodes	52428.8	293.8	0.56%
512 nodes	104857.6	587.97	0.56%
1024 nodes	209715.2	1176.69	0.56%
49152 nodes	10066329.6	55177.6	0.55%

The above table summarizes results for various partition sizes for a 50x50x25 sized local problem. The percentage of peak obtained holds steady to full system run. The result is for an unoptimized run. Real applications with similar iterative solvers plug at about 8 to 10% of peak

http://tiny.cc/hpcg

Results for Cielo
Dual Socket AMD (8 core) Magny Cour
Each node is 2*8 Cores 2.4 GHz = Total 153.6 Gflops/

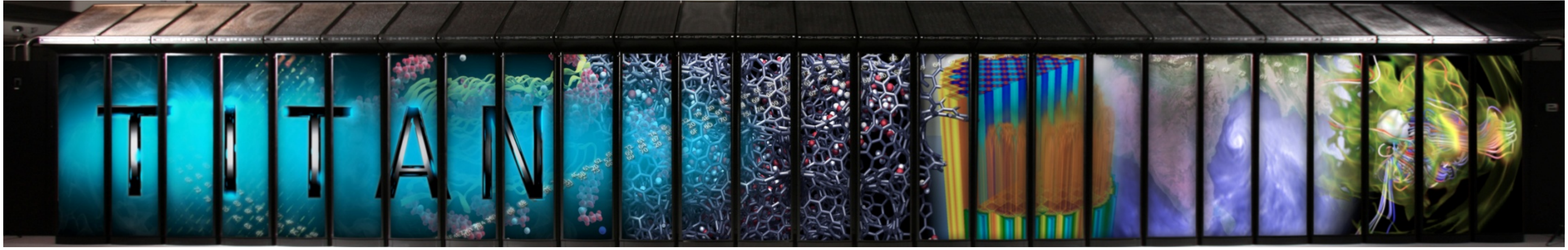


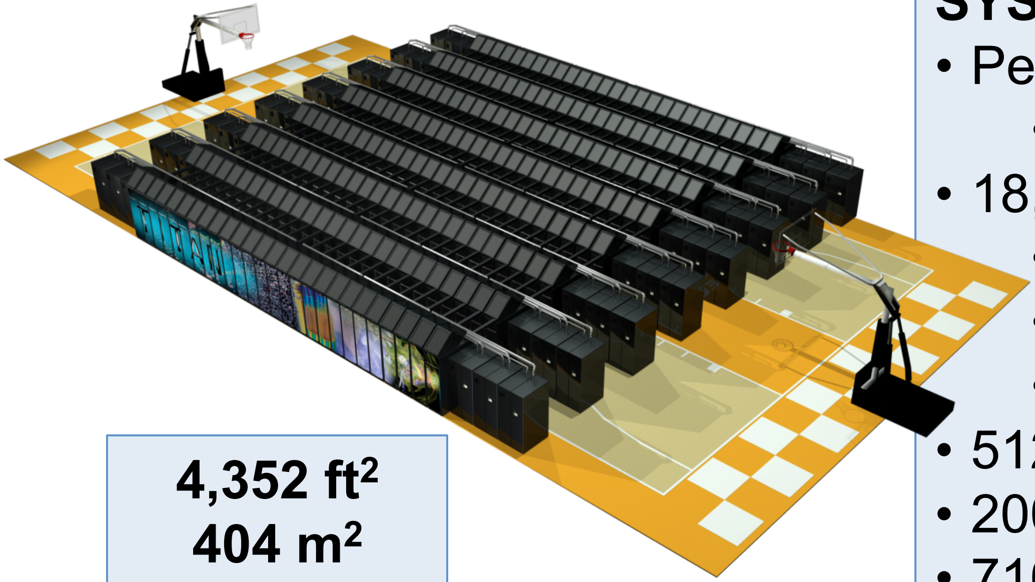
http://tiny.cc/hpcg

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.
- High Performance Ecosystem out of balance
  - Hardware, OS, Compilers, Software, Algorithms, Applications
    - No Moore's Law for software, algorithms and applications

ORNL's "Titan" Hybrid System: Cray XK7 with AMD Opteron and NVIDIA Tesla processors





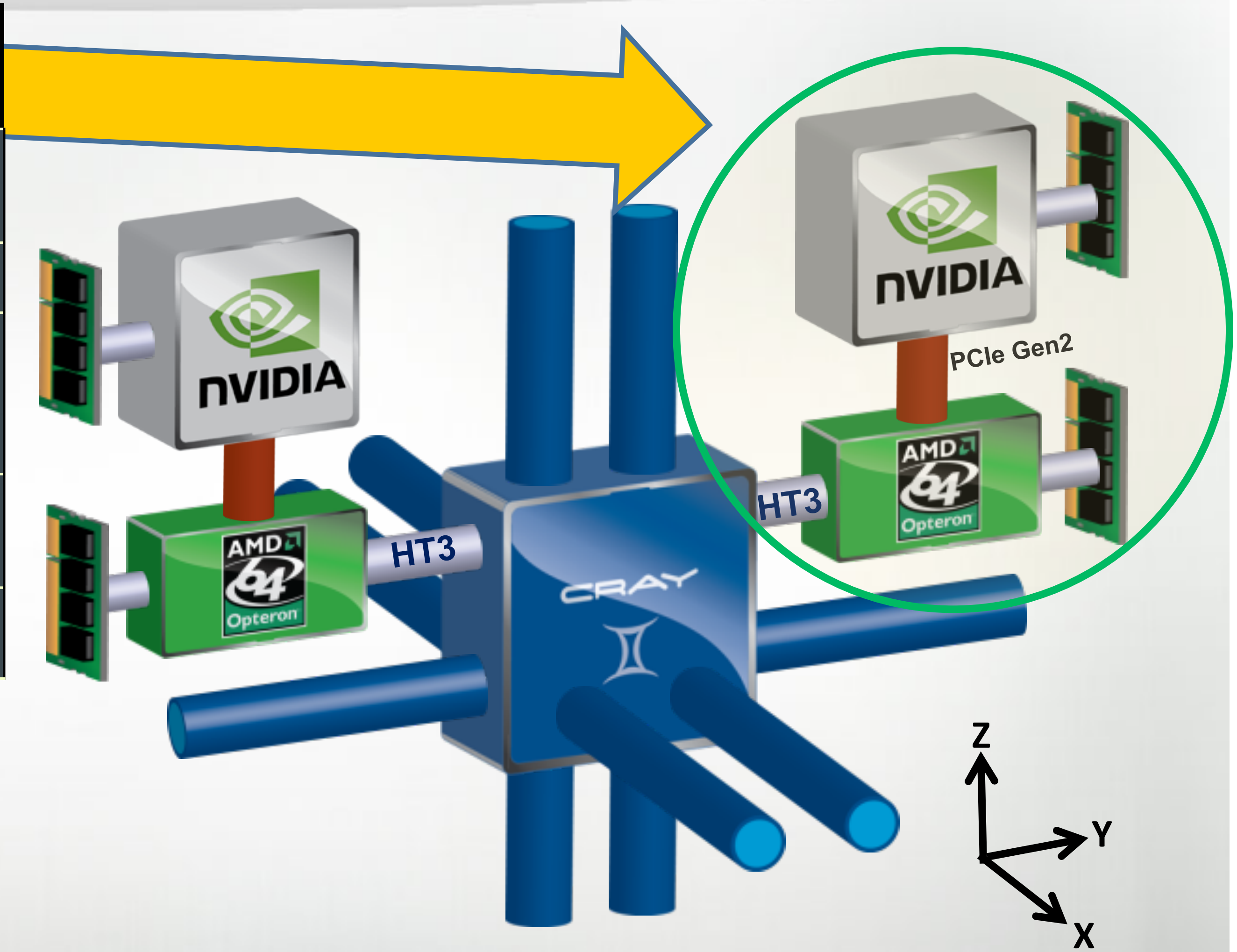
4,352 ft^2
404 m^2

SYSTEM SPECIFICATIONS:

- Peak performance of 27 PF
  - 24.5 Pflop/s GPU + 2.6 Pflop/s AMD
- 18,688 Compute Nodes each with:
  - 16-Core AMD Opteron CPU
  - NVIDIA Tesla "K20x" GPU
  - 32 + 6 GB memory
- 512 Service and I/O nodes
- 200 Cabinets
- 710 TB total system memory
- Cray Gemini 3D Torus Interconnect
- 9 MW peak power

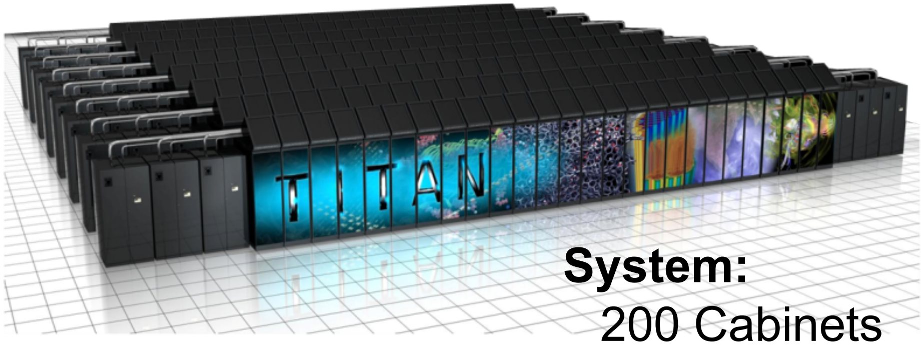
Cray XK7 Compute Node

XK7 Compute Node Characteristics
AMD Opteron 6274 Interlagos 16 core processor
Tesla K20x @ 1311 GF
Host Memory 32GB 1600 MHz DDR3
Tesla K20x Memory 6GB GDDR5
Gemini High Speed Interconnect



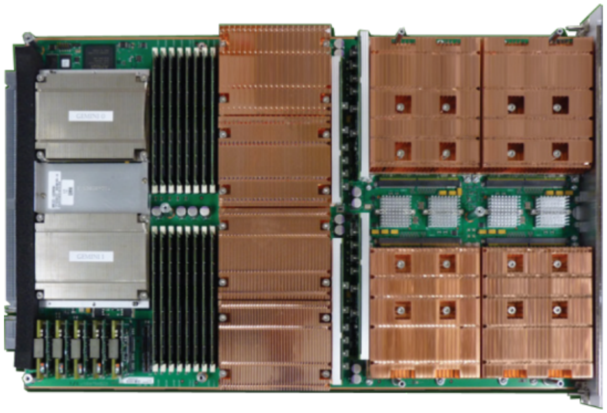
Slide courtesy of Cray, Inc.

Titan: Cray XK7 System

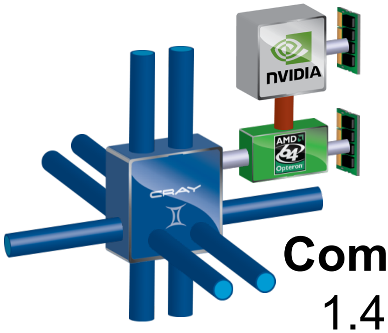


System:
200 Cabinets
18,688 Nodes
27 PF
710 TB

Cabinet:
24 Boards
96 Nodes
139 TF
3.6 TB



Board:
4 Compute Nodes
5.8 TF
152 GB



Compute Node:
1.45 TF
38 GB

Summary

- Major Challenges are ahead for extreme computing
  - Parallelism
  - Hybrid
  - Fault Tolerance
  - Power
  - … and many others not discussed here
- We will need completely new approaches and technologies to reach the Exascale level

The High Cost of Data Movement

- Flop/s or percentage of peak flop/s become much less relevant

Approximate power costs (in picoJoules)

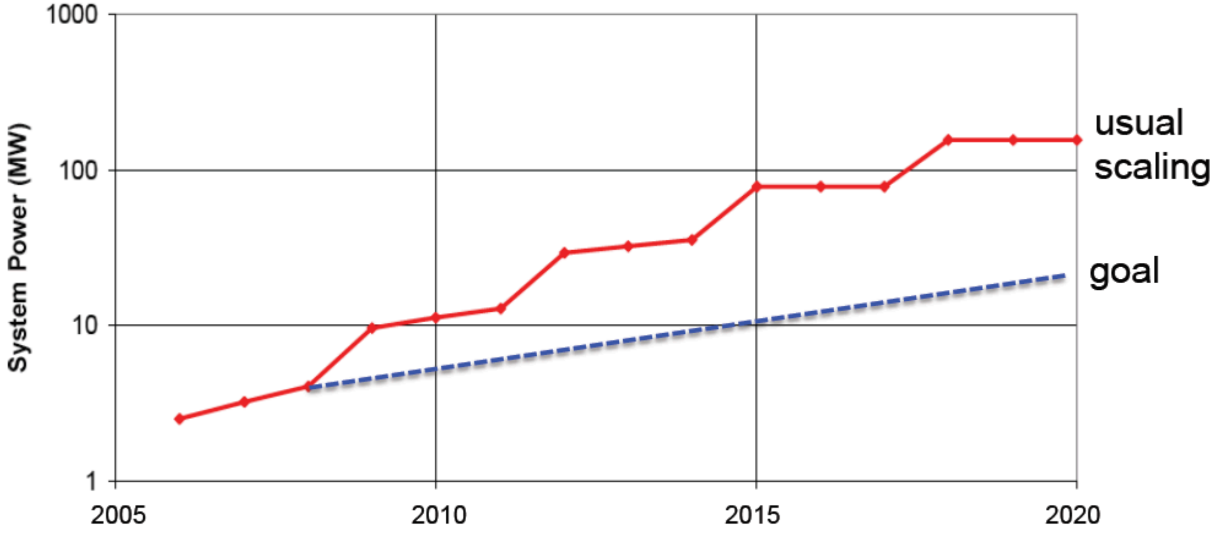
	2011
DP FMADD flop	100 pJ
DP DRAM read	4800 pJ
Local Interconnect	7500 pJ
Cross System	9000 pJ

Source: John Shalf, LBNL

- Algorithms & Software: minimize data movement; perform more work per unit data movement.

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 in 2018 at 20 MWs

A Call to Action

- Hardware has changed dramatically while software ecosystem has remained stagnant
- Need to exploit new hardware trends (e.g., manycore, heterogeneity) that cannot be handled by existing software stack, memory per socket trends
- Emerging software technologies exist, but have not been fully integrated with system software, e.g., UPC, Cilk, CUDA, HPCS
- Community codes unprepared for sea change in architectures
- No global evaluation of key missing components

Exascale is a Global Challenge

- Formed in 2008
- Goal to engage international computer science community to address common software challenges for Exascale
- Focus on open source systems software that would enable multiple platforms
- Shared risk and investment
- Leverage international talent base

International Exascale Software Program

Improve the world's simulation and modeling capability by improving the coordination and development of the HPC software environment

Workshops:

Build an international plan for coordinating research for the next generation open source software for scientific high-performance computing

www.exascale.org

Roadmap Components

www.exascale.org

4.1 Systems Software..
- 4.1.1 Operating systems ....
- 4.1.2 Runtime Systems..
- 4.1.2 I/O systems ....
- 4.1.3 External Environments..
- 4.1.4 Systems Management..

4.2 Development Environments..
- 4.2.1 Programming Models ....
- 4.2.2 Frameworks ....
- 4.2.3 Compilers..
- 4.2.4 Numerical Libraries..
- 4.2.5 Debugging tools..

4.3 Applications..
- 4.3.1 Application Element: Algorithms..
- 4.3.2 Application Support: Data Analysis and Visualization...................
- 4.3.3 Application Support: Scientific Data Management ....

4.4 Crosscutting Dimensions ....
- 4.4.1 Resilience..
- 4.4.2 Power Management ....
- 4.4.3 Performance Optimization ....
- 4.4.4 Programmability..

Where We Are Today:

- Ken Kennedy - Petascale Software Project (2006)
- SC08 (Austin TX) meeting to generate interest
- Funding from DOE's Office of Science & NSF Office of Cyberinfratructure and sponsorship by Europeans and Asians
- US meeting (Santa Fe, NM) April 6-8, 2009
    - 65 people
- European meeting (Paris, France) June 28-29, 2009
    - Outline Report
- Asian meeting (Tsukuba Japan) October 18-20, 2009
    - Draft roadmap and refine report
- SC09 (Portland OR) BOF to inform others
    - Public Comment; Draft Report presented
- European meeting (Oxford, UK) April 13-14, 2010
    - Refine and prioritize roadmap; look at management models
- Maui Meeting October 18-19, 2010
- SC10 (New Orleans) BOF to inform others (Wed 5:30, Room 389)
- Kyoto Meeting - April 6-7, 2011

Nov 2008

Apr 2009

Jun 2009

Oct 2009

Nov 2009

Apr 2010

Oct 2010

Nov 2010

Apr 2011

www.exascale.org

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

INTERNATIONAL EXASCALE SOFTWARE PROJECT ROADMAP

Jack Dongarra
Pete Beckman
Terry Moore
Jean-Claude Andre
Jean-Yves Berthou
Taisuke Boku
Franck Cappello
Barbara Chapman
Xuebin Chi
Alok Choudhary
Sudip Dosanjh
Al Geist
Bill Gropp
Robert Harrison
Mark Hereld
Michael Heroux
Adolfy Hoisie
Koh Hotta
Yutaka Ishikawa
Fred Johnson
Sanjay Kale
Richard Kenway
Bill Kramer
Jesus Labarta
Bob Lucas
Barney Maccabe
Satoshi Matsuoka
Paul Messina
Bernd Mohr
Matthias Mueller
Wolfgang Nagel
Hiroshi Nakashima
Michael E. Papka
Dan Reed
Mitsuhisa Sato
Ed Seidel
John Shalf
David Skinner
Thomas Sterling
Rick Stevens
William Tang
John Taylor
Rajeev Thakur
Anne Trefethen
Marc Snir
Aad van der Steen
Fred Streitz
Bob Sugar
Shinji Sumimoto
Jeffrey Vetter
Robert Wisniewski
Kathy Yelick

SPONSORS

Office of Science U.S. Department of Energy
NSF

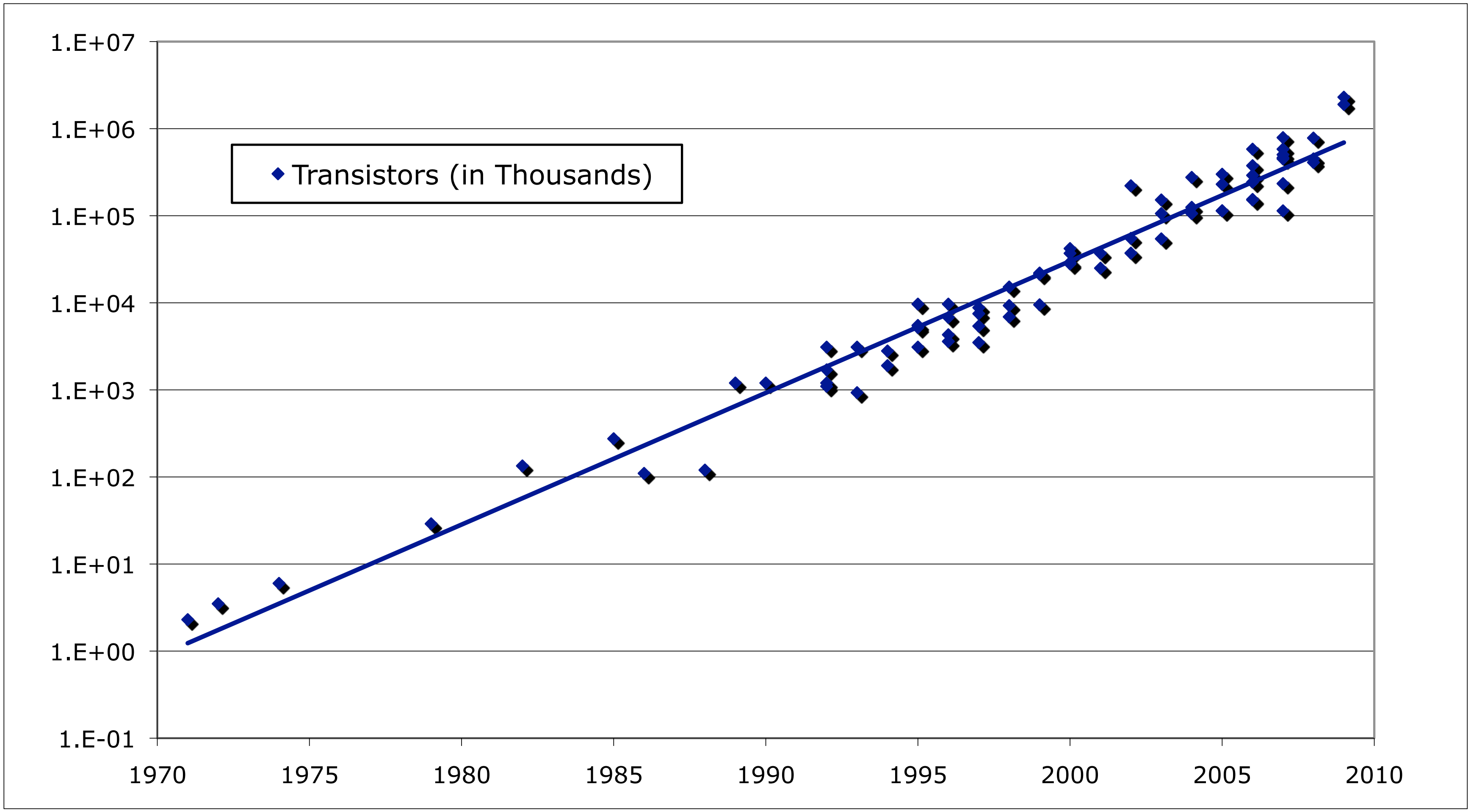
ANR
cea
CERFACS
CRAY THE SUPERCOMPUTER COMPANY
eDF
FUJITSU
INRIA
GENCI
NVIDIA
RIKEN
東京大学 THE UNIVERSITY OF TOKYO
筑波大学

To be published in the January 2011 issue of The International Journal of High Performance Computing Applications

> “We can only see a short distance ahead, but we can see plenty there that needs to be done.”
>
> - *Alan Turing (1912 – 1954)*

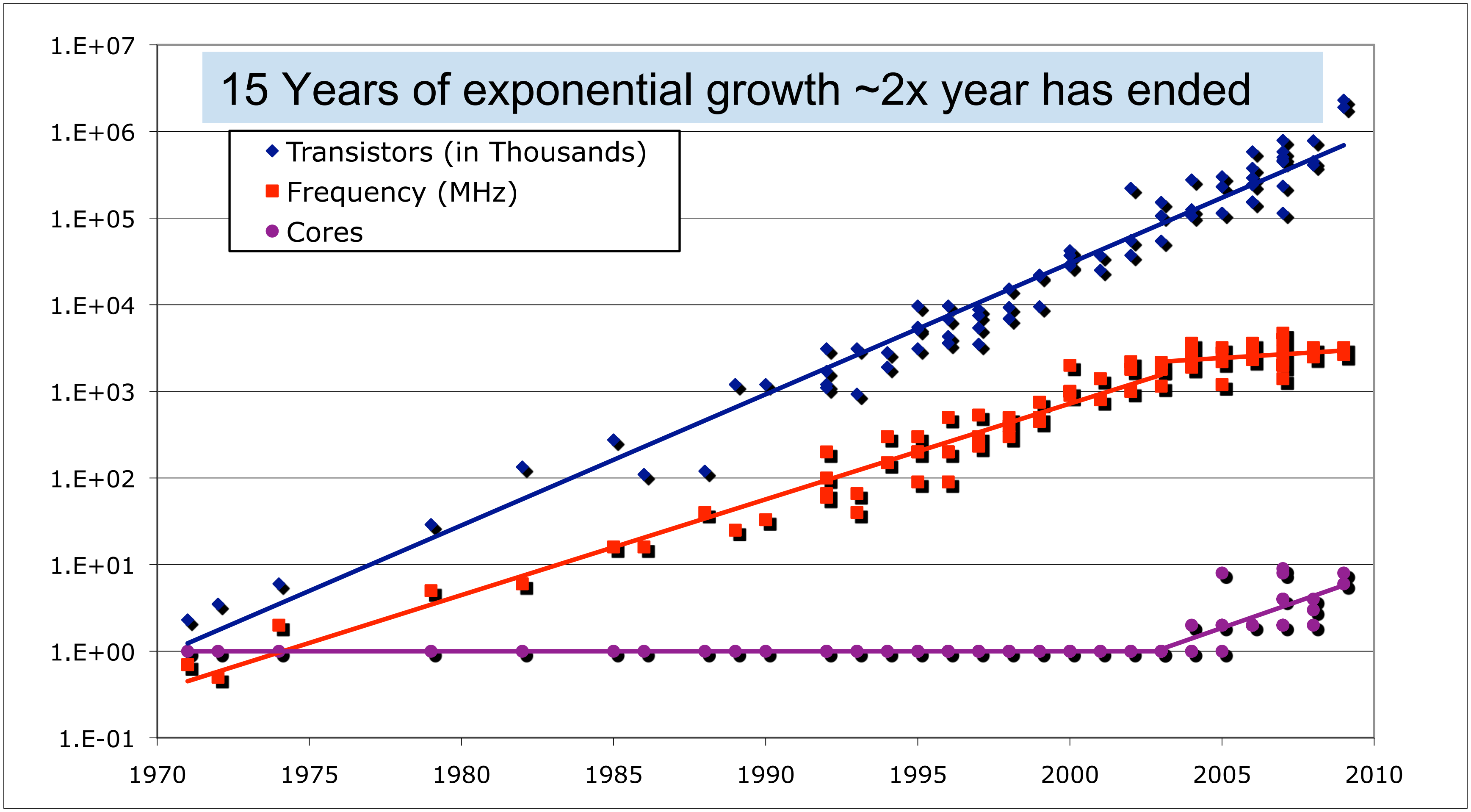
- www.exascale.org

Moore’s Law is Alive and Well



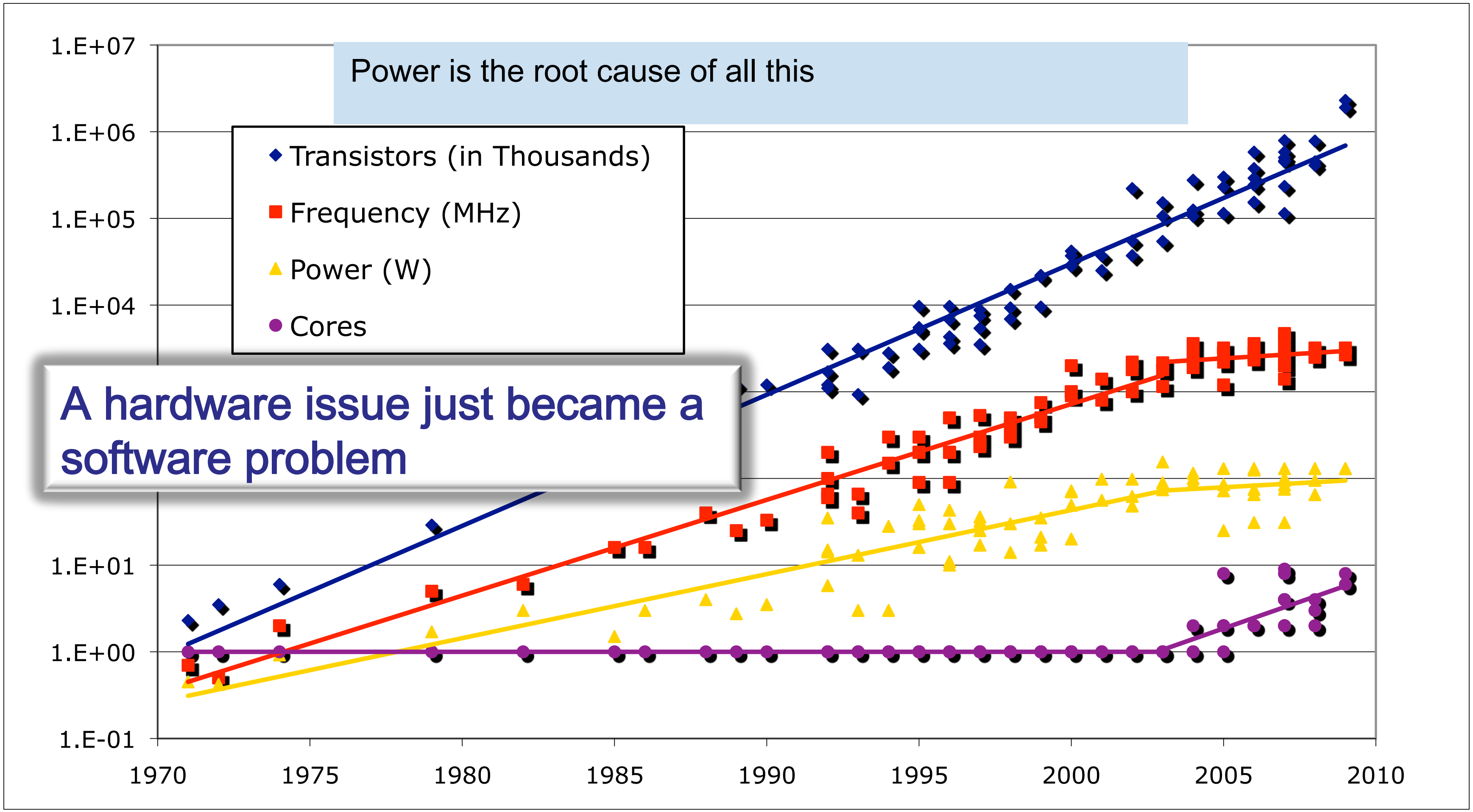
Data from Kunle Olukotun, Lance Hammond, Herb Sutter,
Burton Smith, Chris Batten, and Krste Asanoviç
Slide from Kathy Yelick

But Clock Frequency Scaling Replaced by Scaling Cores / Chip



Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç
Slide from Kathy Yelick

Performance Has Also Slowed, Along with Power



Power is the root cause of all this

A hardware issue just became a software problem

Data from Kunle Olukotun, Lance Hammond, Herb Sutter, Burton Smith, Chris Batten, and Krste Asanoviç
Slide from Kathy Yelick

Power Cost of Frequency

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

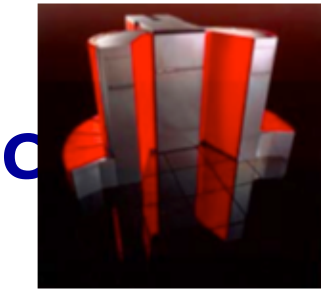
(Bigger # is better)

50% more performance with 20% less power

Preferable to use multiple slower devices, than one superfast device

Looking at the Gordon Bell Prize

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

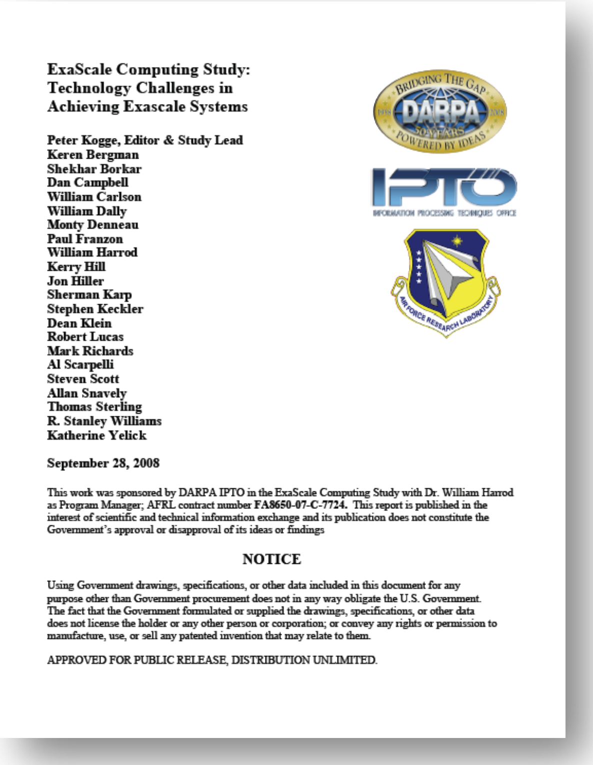


- 1 GFlop/s; 1988; Cray Y-MP; 8 Proc
  - Static finite element analysis

- 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; $1.5\text{x}10^5$ Processors
  - Superconductive materials

Exascale Computing



**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.

BRIDGING THE GAP
DARPA
1958
2008
50 YEARS
POWERED BY IDEAS
IPTO
INFORMATION PROCESSING TECHNIQUES OFFICE
AIR FORCE RESEARCH LABORATORY

- 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

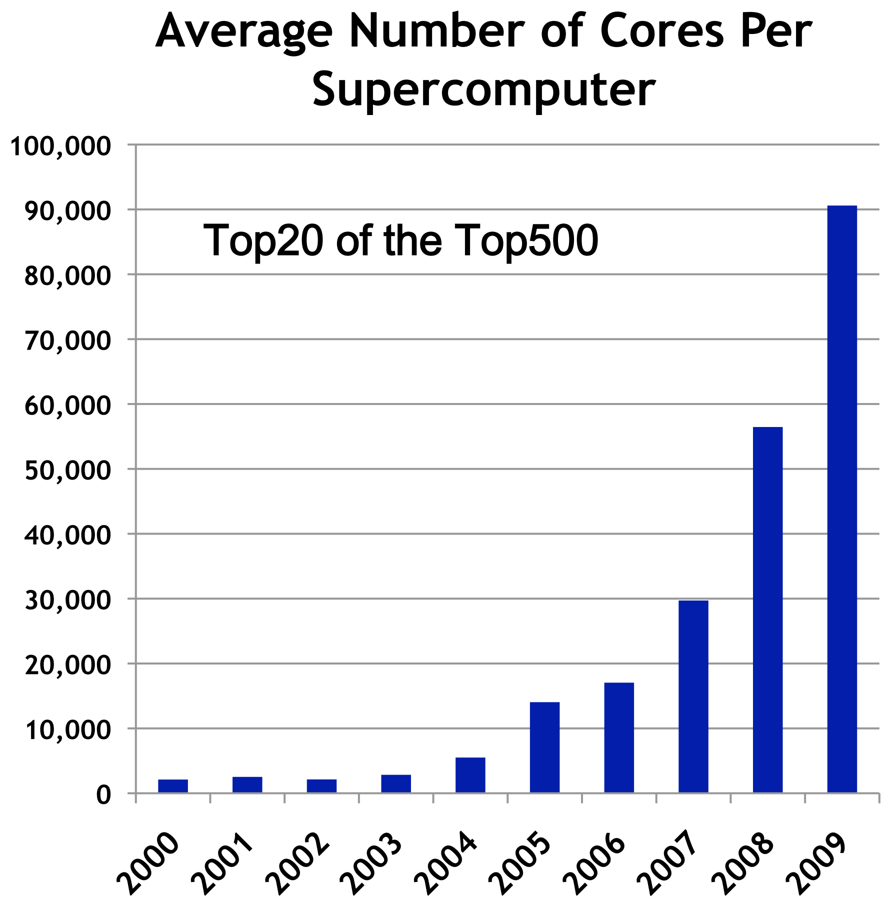
Google: exascale computing study

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**

Hardware and System Software Scalability

- **Barriers**
  - Fundamental assumptions of system software architecture did not anticipate exponential growth in parallelism
  - Number of components and MTBF changes the game
- **Technical Focus Areas**
  - System Hardware Scalability
  - System Software Scalability
  - Applications Scalability
- **Technical Gap**
  - 1000x improvement in system software scaling
  - 100x improvement in system software reliability



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

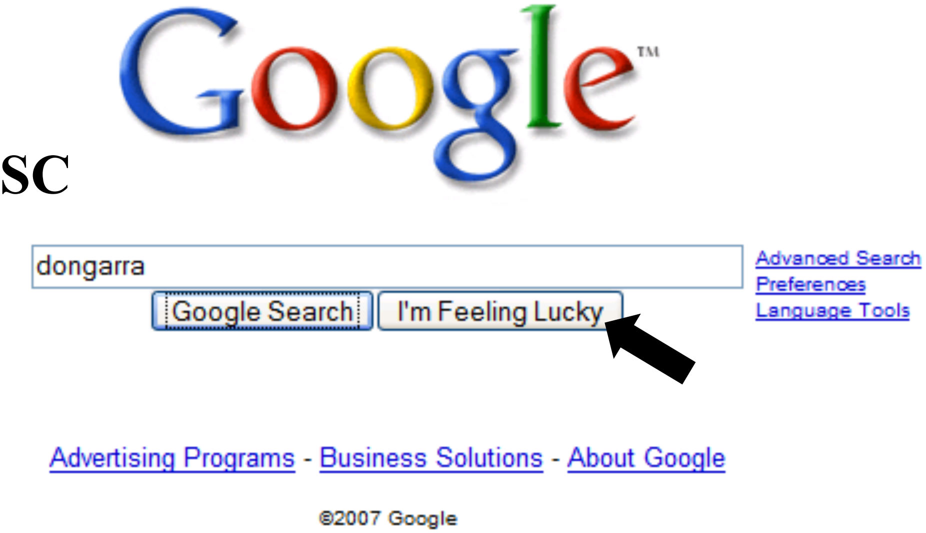
NVIDIA

Microsoft

The MathWorks

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

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



dongarra

NSF University of Illinois; Blue Waters

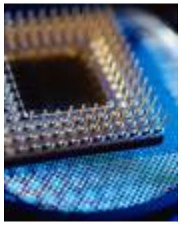
Blue Waters will be the powerhouse of the National Science Foundation’s strategy to support supercomputers for scientists nationwide

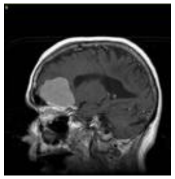
T1	**Blue Waters**	**NCSA/Illinois**	**1 Pflop *sustained* per second**
T2	**Kraken**	**NICS/U of Tennessee**	**1 Pflops peak per second**
	Ranger	**TACC/U of Texas**	**504 Tflop/s peak per second**
T3	**Campuses across the U.S.**	**Several sites**	**50-100 Tflops peak per second**

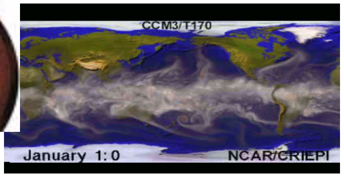
Industrial Use of Supercomputers

- Of the 500 Fastest Supercomputer
  - Worldwide, Industrial Use is > 56%

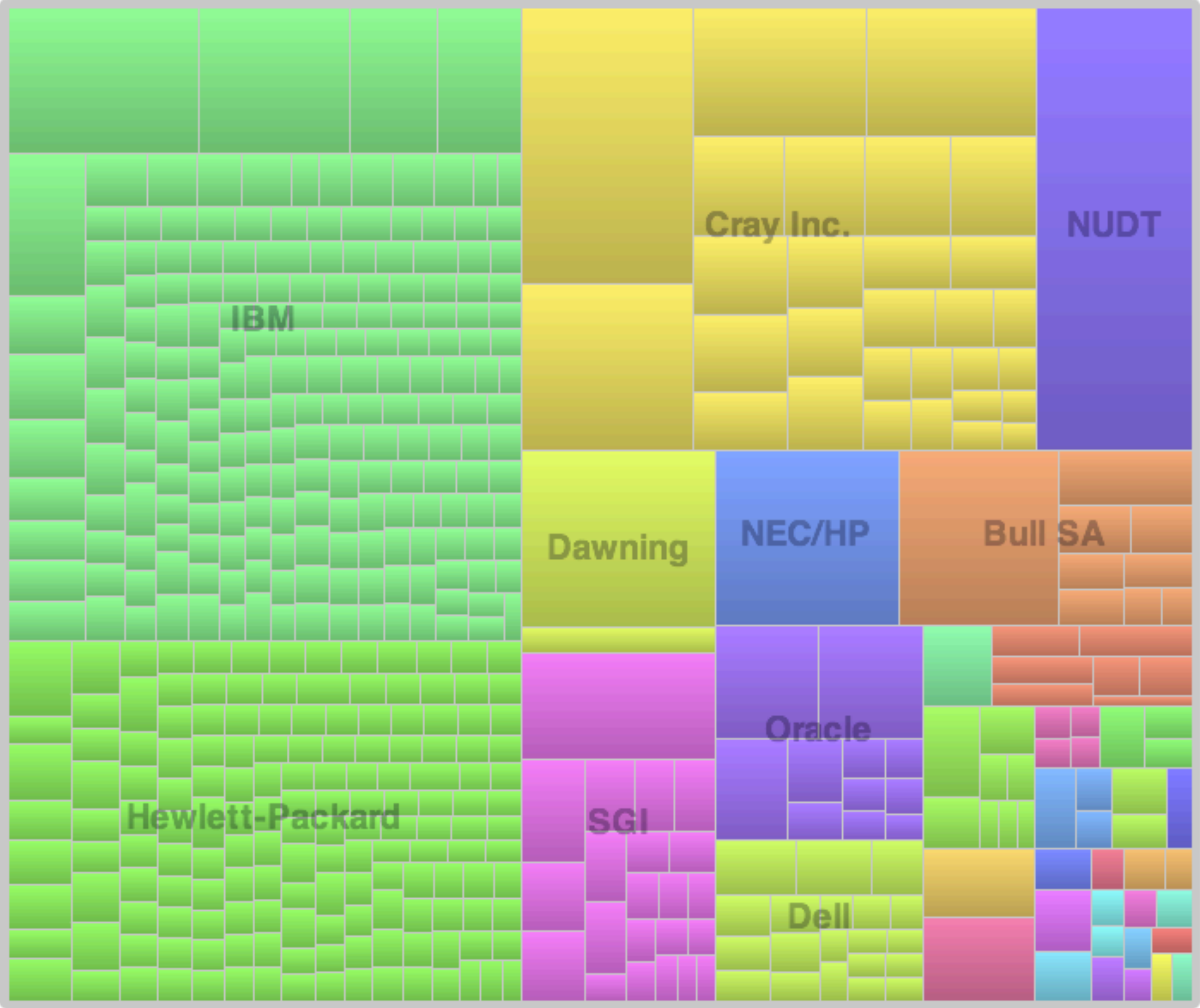
- Aerospace
- Automotive
- Biology
- CFD
- Database
- Defense
- Digital Content Creation
- Digital Media
- Electronics
- Energy
- Environment
- Finance
- Gaming
- Geophysics
- Image Proc./Rendering
- Information Processing Service
- Information Service
- Life Science
- Media
- Medicine
- Pharmaceutics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







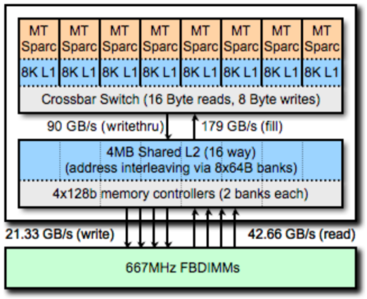
Manufacturer's Share



Today's Multicores

99% of Top500 Systems Are Based on Multicore

Of the Top500,
499 are multicore.



MT Sparc	MT Sparc	MT Sparc	MT Sparc	MT Sparc	MT Sparc	MT Sparc	MT Sparc
8K L1	8K L1	8K L1	8K L1	8K L1	8K L1	8K L1	8K L1

Crossbar Switch (16 Byte reads, 8 Byte writes)

90 GB/s (writethru)

179 GB/s (fill)

4MB Shared L2 (16 way)
(address interleaving via 8x64B banks)

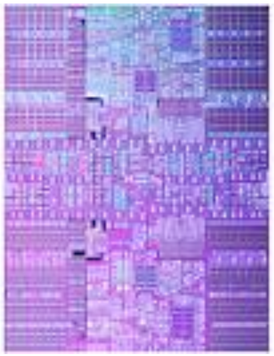
4x128b memory controllers (2 banks each)

21.33 GB/s (write)

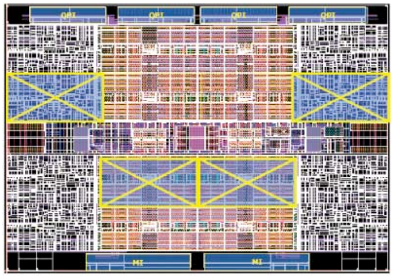
42.66 GB/s (read)

667MHz FBDIMMs

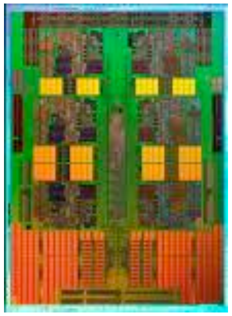
Sun Niagra2 (8 cores)



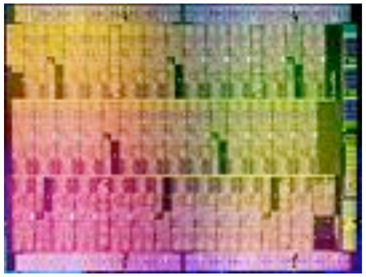
IBM Power 7 (8 cores)



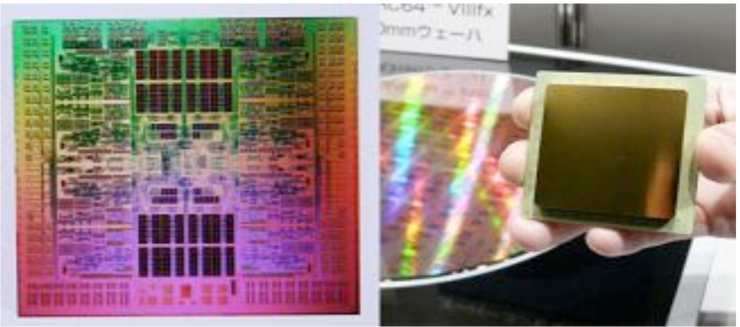
Intel Xeon(8 cores)



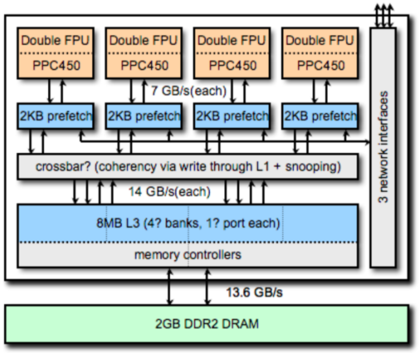
AMD Magny Cours
(12 cores)



Intel Knight's Corner
(40 cores)



Fujitsu Venus (8 cores)



Double FPU PPC450	Double FPU PPC450	Double FPU PPC450	Double FPU PPC450
2KB prefetch	2KB prefetch	2KB prefetch	2KB prefetch

7 GB/s(each)

crossbar? (coherency via write through L1 + snooping)

14 GB/s(each)

8MB L3 (4? banks, 1? port each)

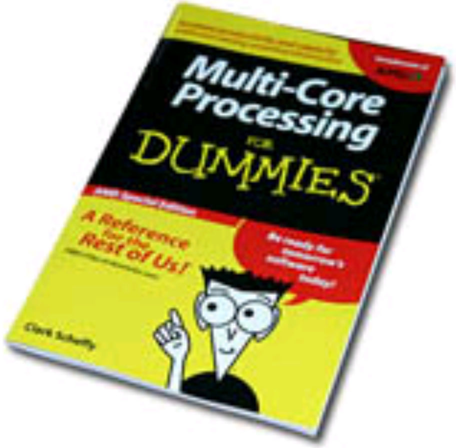
memory controllers

3 network interfaces

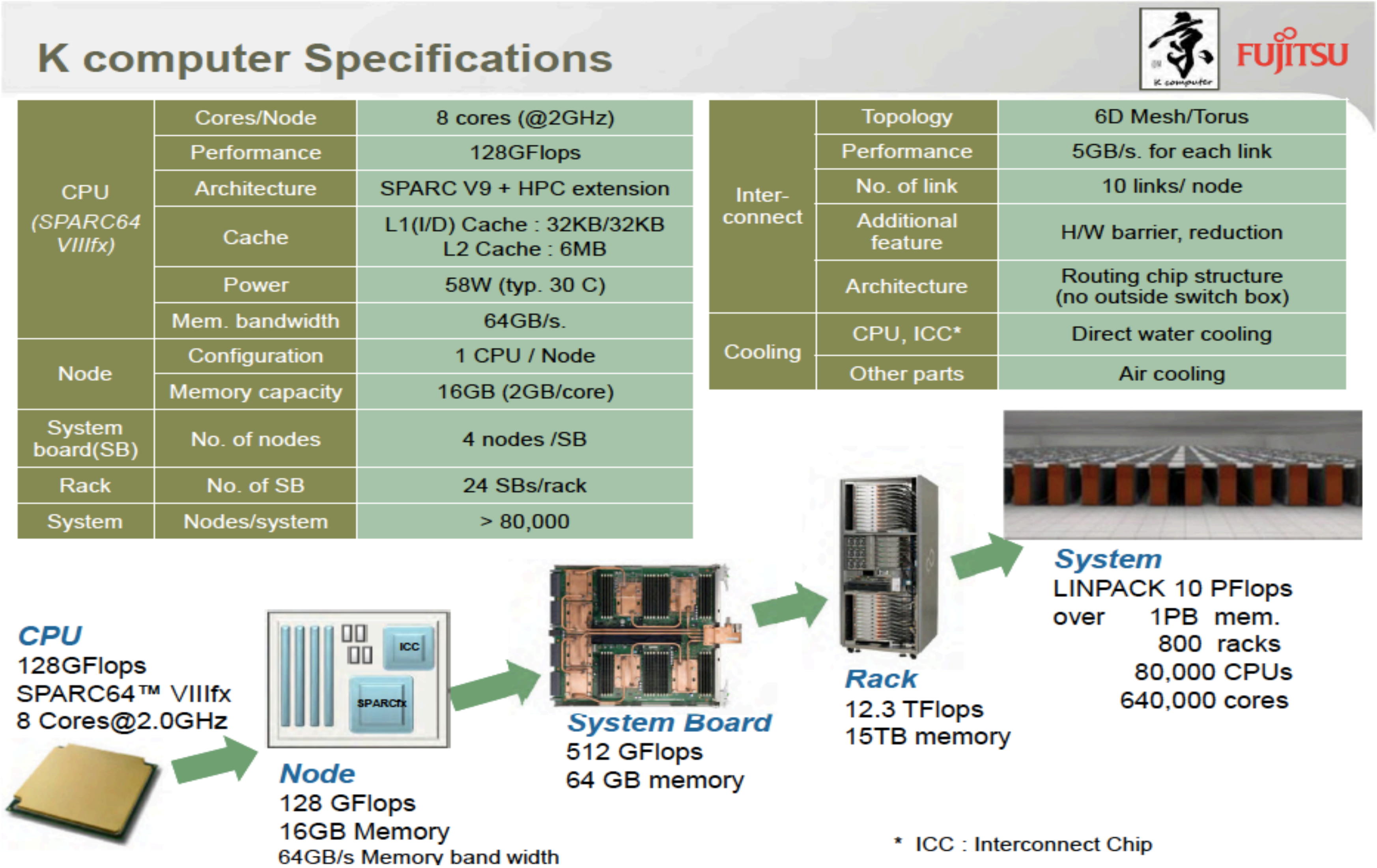
13.6 GB/s

2GB DDR2 DRAM

IBM BG/P (4 cores)



Japanese K Computer



K computer Specifications

CPU *(SPARC64 VIIIfx)*	Cores/Node	8 cores (@2GHz)
	Performance	128GFlops
	Architecture	SPARC V9 + HPC extension
	Cache	L1(I/D) Cache : 32KB/32KB L2 Cache : 6MB
	Power	58W (typ. 30 C)
	Mem. bandwidth	64GB/s.
Node	Configuration	1 CPU / Node
	Memory capacity	16GB (2GB/core)
System board(SB)	No. of nodes	4 nodes /SB
Rack	No. of SB	24 SBs/rack
System	Nodes/system	> 80,000

Inter-connect	Topology	6D Mesh/Torus
	Performance	5GB/s. for each link
	No. of link	10 links/ node
	Additional feature	H/W barrier, reduction
	Architecture	Routing chip structure (no outside switch box)
Cooling	CPU, ICC*	Direct water cooling
	Other parts	Air cooling

CPU
128GFlops
SPARC64™ VIIIfx
8 Cores@2.0GHz

Node
128 GFlops
16GB Memory
64GB/s Memory band width

System Board
512 GFlops
64 GB memory

Rack
12.3 TFlops
15TB memory

System
LINPACK 10 PFlops
over 1PB mem.
800 racks
80,000 CPUs
640,000 cores

* ICC : Interconnect Chip

New Linpack run with 705,024 cores at 10.51 Pflop/s (88,128 CPUs), 12.7 MW; 29.5 hours
Fujitsu to have a 100 Pflop/s system in 2014