Parallel Processing and Applied Mathematics
Warsaw, POLAND
52.259 °N 21.020 °E
10th International Conference on
Parallel Processing and Applied Mathematics
PPAM
2013
Warsaw, Poland
September 8-11, 2013

An Overview of High Performance Computing and Challenges for the Future

Jack Dongarra

University of Tennessee
Oak Ridge National Laboratory
University of Manchester

Key Message

- **Exascale has been discussed in numerous workshops, conferences, planning meetings for about five years**

- **Exascale projects have been started in the US and many other countries and regions**

- **Progress has been made, but key challenges to exascale remain**

Modeling and Simulation at the Exascale for Energy and the Environment

Co-Chairs:
Horst Simon
Lawrence Berkeley National Laboratory
April 17–18, 2007
Thomas Zacharia
Oak Ridge National Laboratory
May 17–18, 2007
Rick Stevens
Argonne National Laboratory
May 31–June 1, 2007

INTERNATIONAL EXASCALE SOFTWARE PROJECT

ROADMAP

Scientific Grand Challenges

CHALLENGES IN CLIMATE CHANGE SCIENCE AND THE ROLE OF COMPUTING AT THE EXTREME SCALE

November 6-7, 2008 • Washington D.C.

Scientific Grand Challenges

CHALLENGES FOR UNDERSTANDING THE QUANTUM UNIVERSE AND THE ROLE OF COMPUTING AT THE EXTREME SCALE

December 9-11, 2008 • Menlo Park, CA

U.S. DEPARTMENT OF ENERGY

Scientific Grand Challenges

FOREFRONT QUESTIONS IN NUCLEAR SCIENCE AND THE ROLE OF COMPUTING AT THE EXTREME SCALE

January 26-28, 2009 • Washington, D.C.

U.S. DEPARTMENT OF ENERGY

State of Supercomputing in 2013

- Pflops computing fully established with 26 machines
- Three technology “swim lanes” or architecture possibilities are thriving
- Interest in supercomputing is now worldwide, and growing in many new markets
- Exascale projects in many countries and regions
- Rapid growth predicted by IDC for the next three years

June 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	18	1902
2	DOE / OS Oak Ridge Nat Lab	Titan, Cray XK7 (16C) + Nvidia Kepler GPU (14c) + Custom	USA	560,640	17.6	65	8.3	2143
3	DOE / NNSA L Livermore Nat Lab	Sequoia, BlueGene/Q (16c) + custom	USA	1,572,864	17.1	85	7.9	2177
4	RIKEN Advanced Inst for Comp Sci	K computer Fujitsu SPARC64 VIIIfx (8c) + Custom	Japan	705,024	10.5	93	13	830
5	DOE / OS Argonne Nat Lab	Mira, BlueGene/Q (16c) + Custom	USA	786,432	8.58	85	3.9	2177
6	Texas Advanced Computing Center	Stampede, Dell Intel (8c) + Intel Xeon Phi (61c) + IB	USA	204,900	5.16	61	4.5	1146
7	Forschungszentrum Juelich (FZJ)	JuQUEEN, BlueGene/Q, Power BQC 16C 1.6GHz+Custom	Germany	458,752	5.01	85	2.3	2177
8	DOE / NNSA L Livermore Nat Lab	Vulcan, BlueGene/Q, Power BQC 16C 1.6GHz+Custom	USA	393,216	4.29	85	2.0	2177
9	Leibniz Rechenzentrum	SuperMUC, Intel (8c) + IB	Germany	147,456	2.90	91*	3.4	846
10	Nat. SuperComputer Center in Tianjin	Tianhe-1A, NUDT Intel (6c) + Nvidia Fermi GPU (14c) + Custom	China	186,368	2.57	55	4.0	635
500	Web Company	HP Cluster	USA	17,904	.096	50		

Petaflops Club

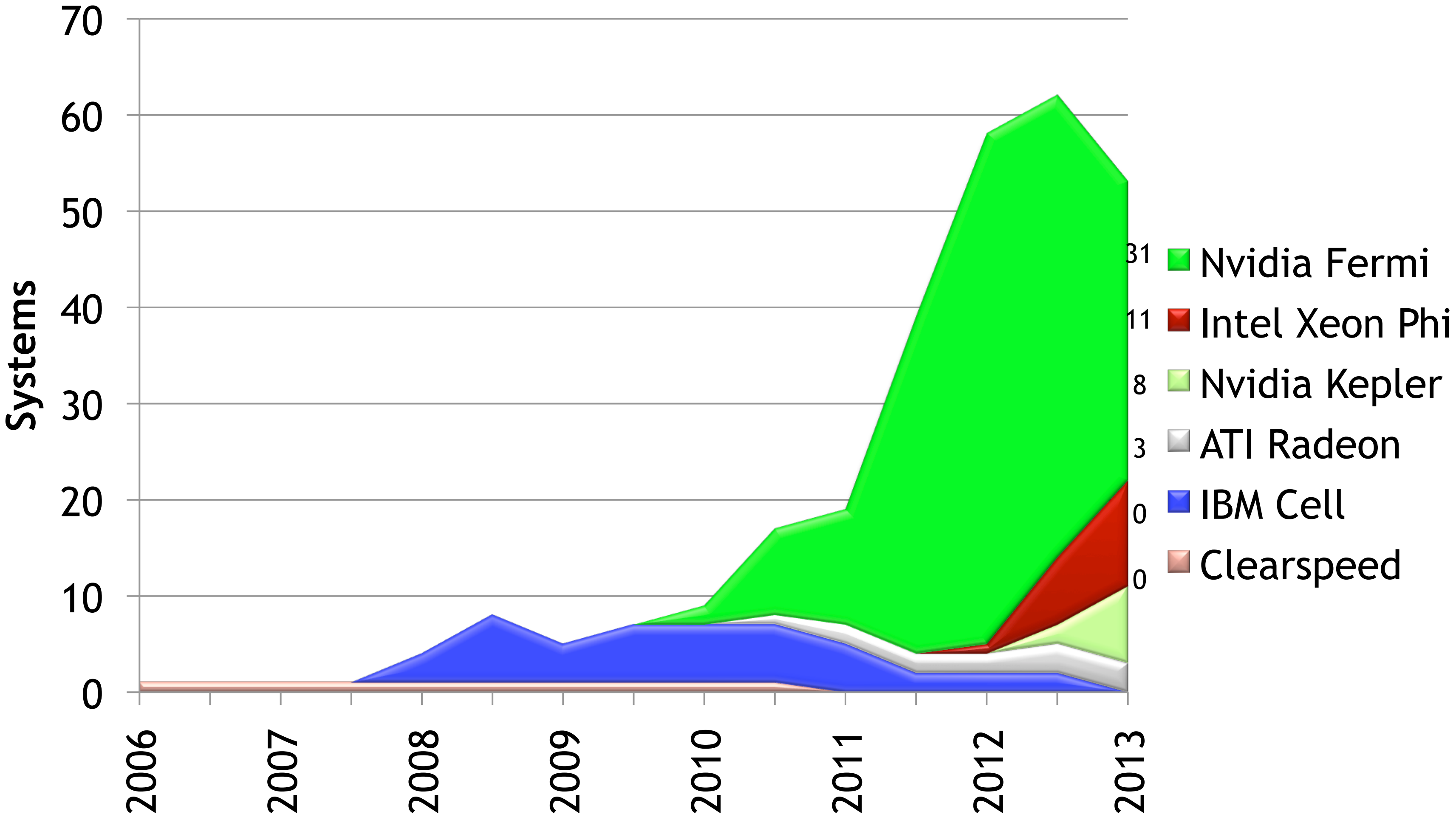
Name	Rmax Linpack# Pflops	Country	11 🇺🇸 4 🇯🇵 2 🇩🇪 3 🇨🇳 2 🇬🇧 3 🇫🇷 1 🇮🇹
Tianhe-2 (MilkyWay-2)	33.9	China	NUDT: Hybrid Intel/Intel/Custom
Titan	17.6	US	Cray: Hybrid AMD/Nvidia/Custom
Sequoia	17.2	US	IBM: BG-Q/Custom
K Computer	10.5	Japan	Fujitsu: Sparc/Custom
Mira	8.59	US	IBM: BG-Q/Custom
Stampede	5.17	US	Dell: Hybrid/Intel/Intel/IB
JUQUEEN	5.01	Germany	IBM: BG-Q/Custom
Vulcan	4.29	US	IBM: BG-Q/Custom
SuperMUC	2.90	Germany	IBM: Intel/IB
Tianhe-1A	2.57	China	NUDT: Hybrid Intel/Nvidia/Custom
Pangea	2.10	France	Bull: Intel/IB
Fermi	1.79	Italy	IBM: BG-Q/Custom
DARPA Trial Subset	1.52	US	IBM: Intel/IB
Spirit	1.42	US	SGI: Intel/IB
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
Pleiades	1.24	US	SGI Intel/IB
Helios	1.24	Japan	Bull: Intel/IB
TSUBAME 2.0	1.19	Japan	NEC/HP: Hybrid Intel/Nvidia/IB
Cielo	1.11	US	Cray: AMD/Custom
DiRAC	1.07	K	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

Petaflops Club

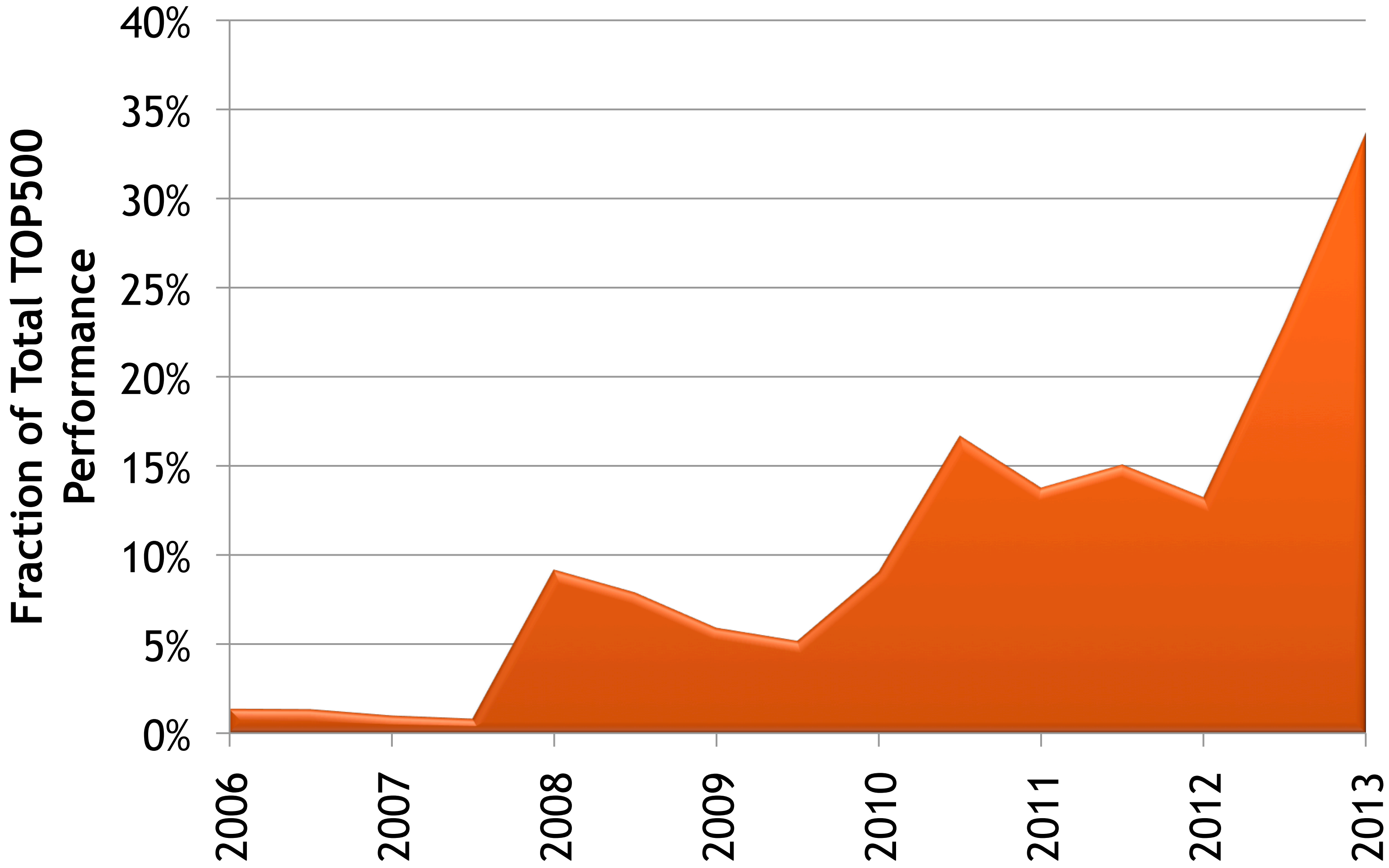
Name	Rmax Linpack# Pflops	Country	11 🇺🇸 4 🇯🇵 2 🇩🇪 3 🇨🇳 2 🇬🇧 3 🇫🇷 1 🇮🇹
Tianhe-2 (MilkyWay-2)	33.9	China	NUDT: Hybrid Intel/Intel/Custom
Titan	17.6	US	Cray: Hybrid AMD/Nvidia/Custom
Sequoia	17.2	US	**IBM: BG-Q/Custom**
K Computer	10.5	Japan	Fujitsu: Sparc/Custom
Mira	8.59	US	**IBM: BG-Q/Custom**
Stampede	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	2.90	Germany	IBM: Intel/IB
Tianhe-1A	2.57	China	NUDT: Hybrid Intel/Nvidia/Custom
Pangea	2.10	France	Bull: Intel/IB
Fermi	1.79	Italy	**IBM: BG-Q/Custom**
DARPA Trial Subset	1.52	US	IBM: Intel/IB
Spirit	1.42	US	SGI: Intel/IB
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**
Pleiades	1.24	US	SGI Intel/IB
Helios	1.24	Japan	Bull: Intel/IB
TSUBAME 2.0	1.19	Japan	NEC/HP: Hybrid Intel/Nvidia/IB
Cielo	1.11	US	Cray: AMD/Custom
DiRAC	1.07	K	**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

6 Hybrid Architectures
8 IBM BG/Q
15 Custom X
11 Infiniband X
9 Look like “clusters”

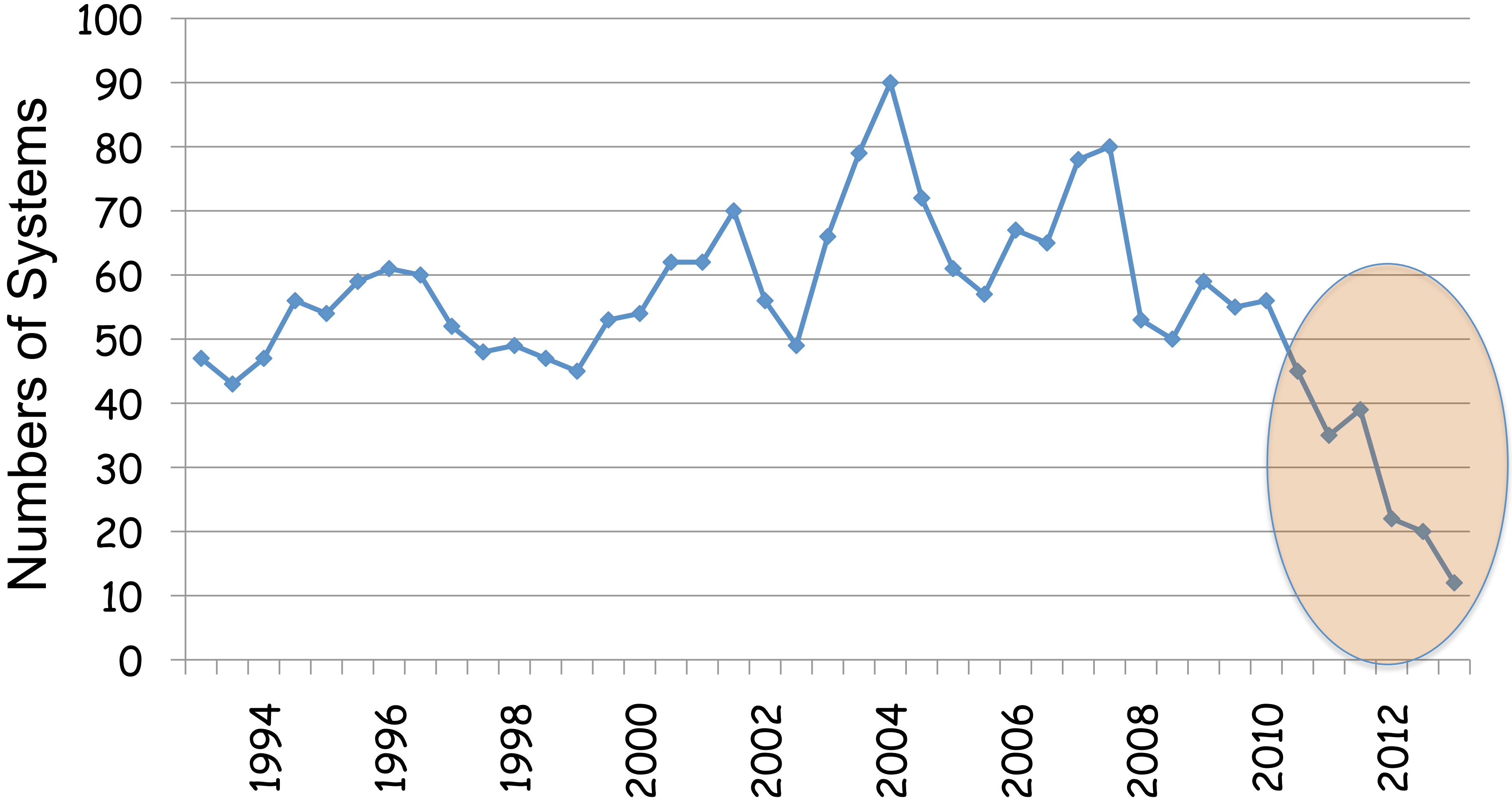
Hybrid/Accelerators (53 Systems)



Top500 Performance Share of Accelerators



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



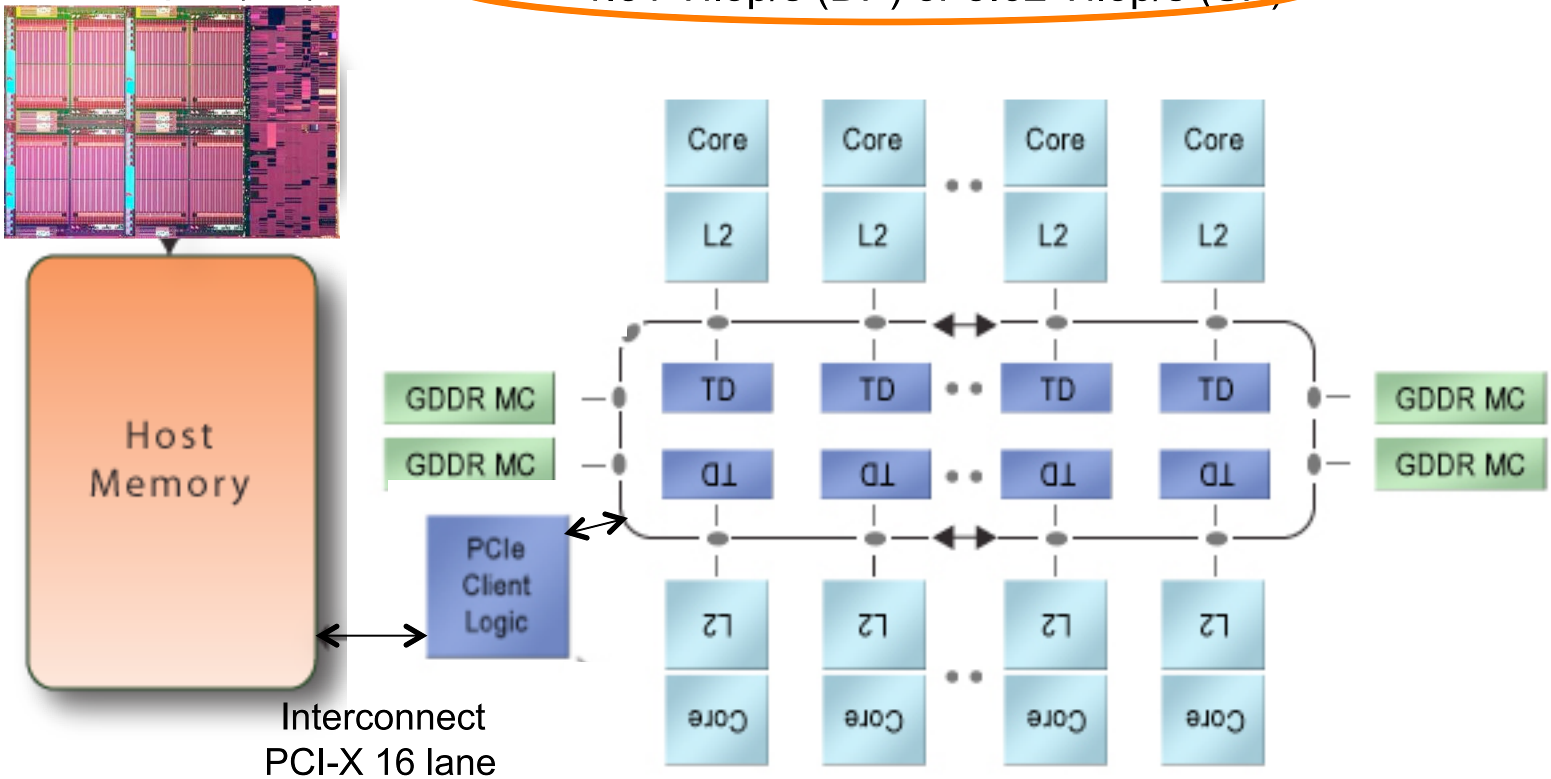
Commodity plus Accelerator Today

Commodity

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

Accelerator/Co-Processor

Intel Xeon Phi
244 “cores” (4 used by OS)
61 (60) FPU = 61 (60) cores
1.091 GHz
60*1.092*8*2 ops/cycle
1.31 Tflop/s (DP) or 3.62 Tflop/s (SP)



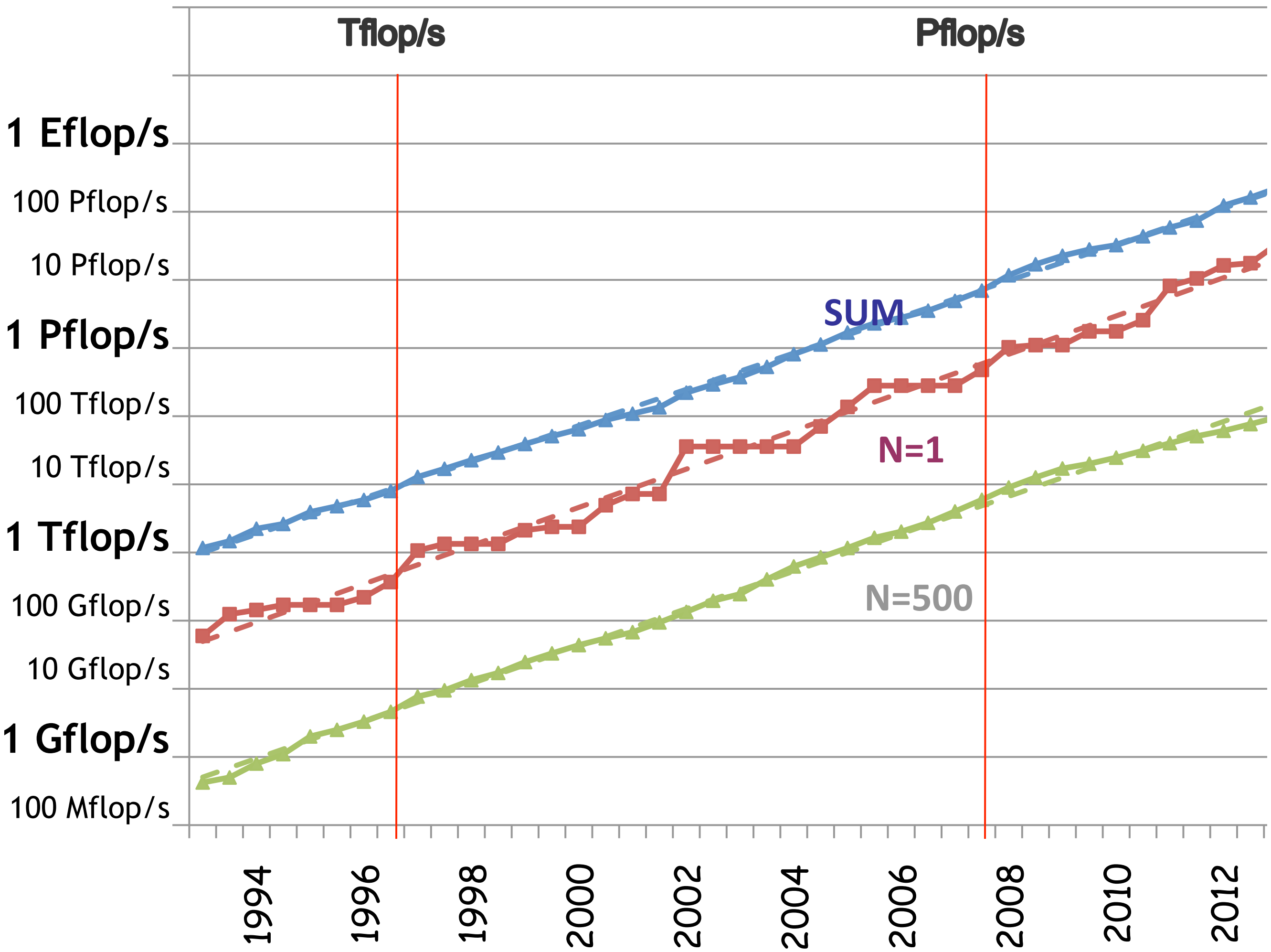
Interconnect
PCI-X 16 lane
64 Gb/s (8 GB/s)
1 GW/s

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

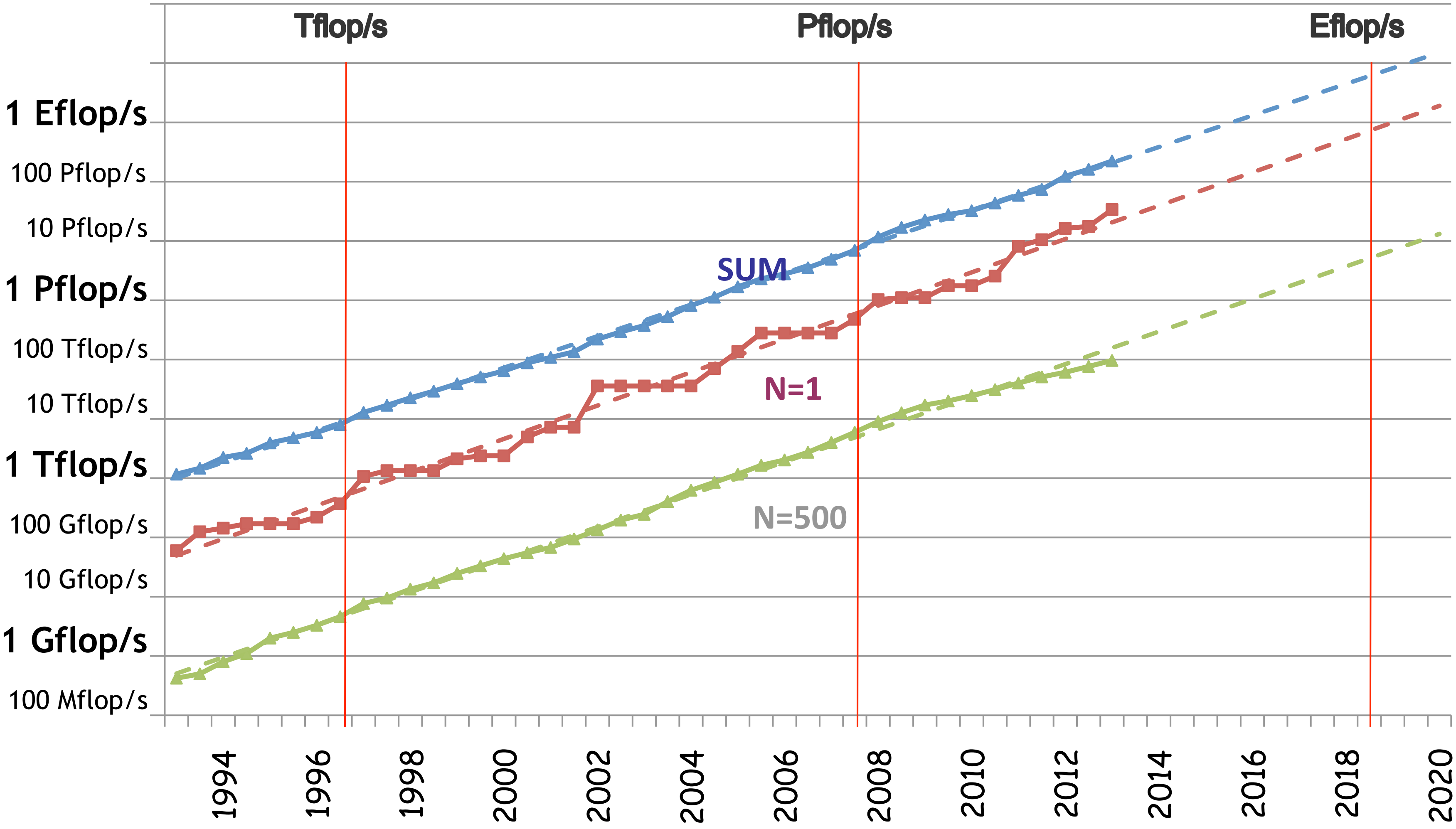
9 🇺🇸 6 🇯🇵 2 🇨🇳

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 (?)	NUDT Tianhe-2 Intel IvyBridge & Xeon Phi	33,862.	9,960,000	5.4	18.

TOP500 Editions (41 so far, 20 years)



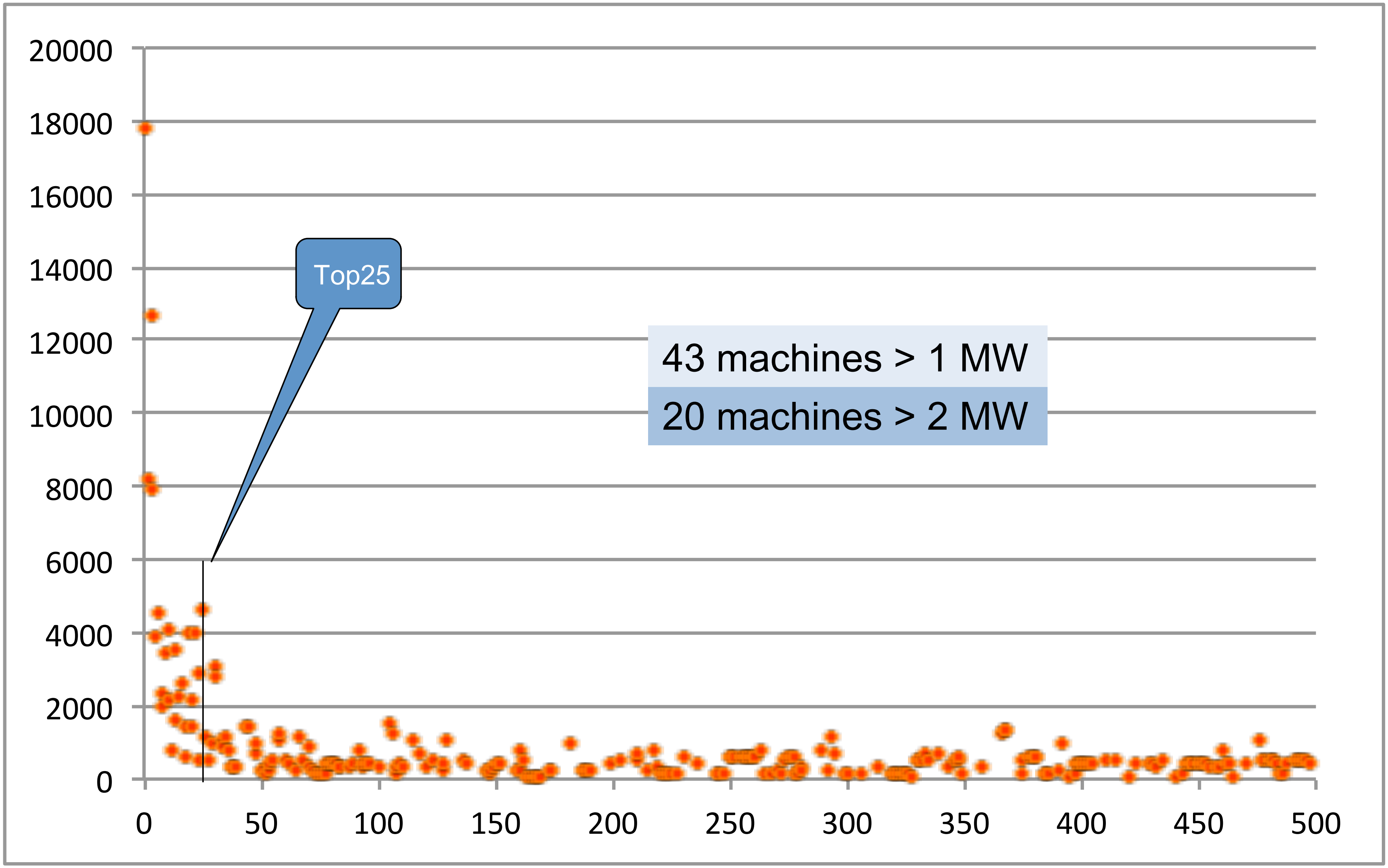
TOP500 Editions (53 edition, 26 years)



Potential 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	?? unknown	O(<1 day)	O(?)

Power Level (kW)



43 machines > 1 MW

20 machines > 2 MW

Most Power Efficient Architectures

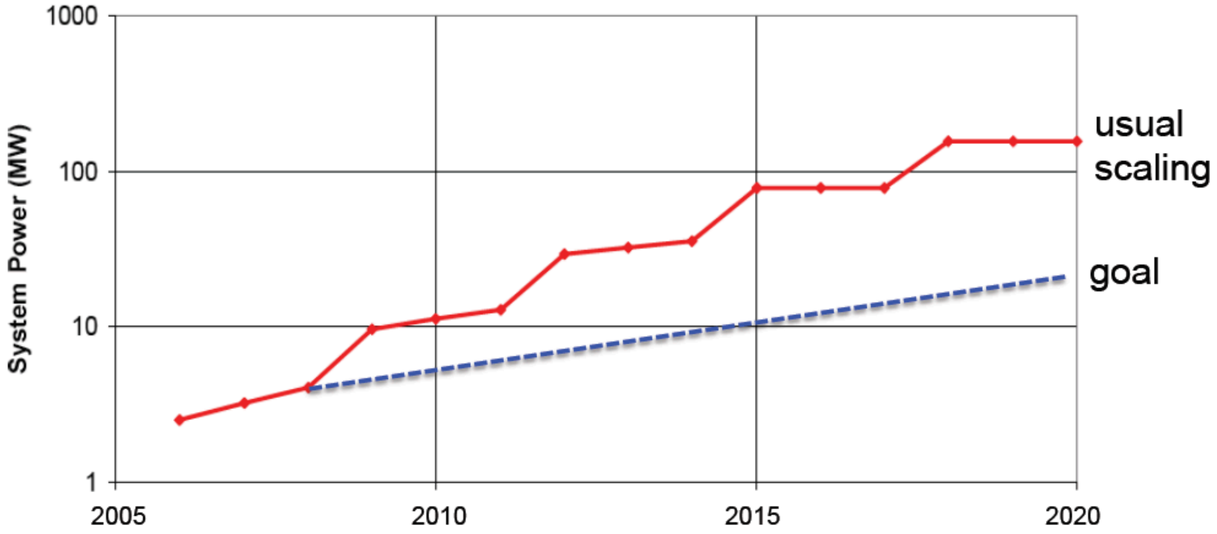
Computer	Rmax/ Power
Adtech, ASUS, Xeon 8C 2.0GHz, Infiniband FDR, **AMD FirePro**	**2.97**
Appro GreenBlade, Xeon 8C 2.6GHz, Infiniband FDR, **Intel Xeon Phi**	**2.45**
BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	**2.30**
Cray XK7, Opteron 16C 2.1GHz, Gemini, **NVIDIA Kepler**	**2.24**
Eurotech Aurora HPC, Xeon 8C 3.1GHz, Infiniband QDR, **NVIDIA K20**	**2.19**
iDataPlex DX360M4, Xeon 8C 2.6GHz, Infiniband QDR, **Intel Xeon Phi**	**1.94**
Tianhe-2, NUDT, Intel Xeon 6C 2.2GHz, TH Express-2, **Intel Xeon Phi**	**1.90**
RSC Tornado, Xeon 8C 2.9GHz, Infiniband FDR, **Intel Xeon Phi**	**1.69**
SGI Rackable, Xeon 8C 2.6GHz, Infiniband FDR, **Intel Xeon Phi**	**1.61**
Chundoong Cluster, Xeon 8C 2GHz, Infiniband QDR, **AMD Radeon HD**	**1.47**

Accelerators
&
IBM BG/Q

[Gflops/Watt]

Energy Cost Challenge

- At ~$1M per MW energy costs are substantial
  - 10 Pflop/s in 2011 uses ~10 MWs
  - 1 Eflop/s in 2020 > 100 MWs



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

The High Cost of Data Movement

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

Approximate power costs (in picoJoules)

	2012
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.

Conventional Wisdom is Changing

Old Conventional Wisdom

- **Peak clock frequency** as primary limiter for performance improvement
- **Cost:** FLOPs are biggest cost for system: optimize for compute
- **Concurrency:** Modest growth of parallelism by adding nodes
- **Memory scaling:** maintain byte per flop capacity and bandwidth
- **Uniformity:** Assume uniform system performance
- **Reliability:** It's the hardware's problem

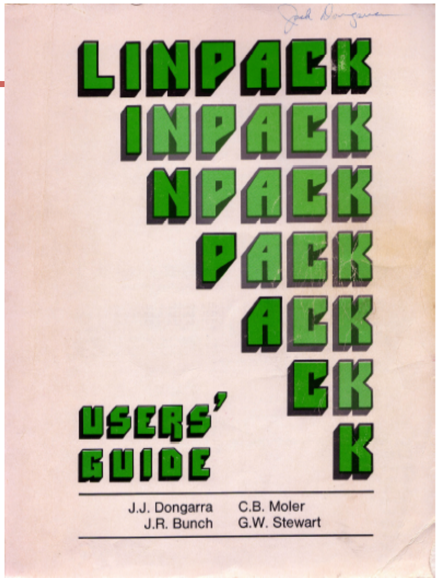
New Conventional Wisdom

- **Power** is primary design constraint for future HPC system design
- **Cost:** Data movement dominates optimize to minimize data movement
- **Concurrency:** Exponential growth of parallelism within chips
- **Memory Scaling:** Compute growing 2x faster than capacity or bandwidth
- **Heterogeneity:** Architectural and performance non-uniformity increase
- **Reliability:** Cannot count on hardware protection alone

Adapted from John Shalf, LBNL

Confessions of an Accidental Benchmarker

- Appendix B of the Linpack Users' Guide
  - Designed to help users extrapolate execution time for Linpack software package
- First benchmark report from 1977;
  - Cray 1 to DEC PDP-10



LINPACK
INPACK
NPACK
PACK
ACK
CK
K

USERS' GUIDE

J.J. Dongarra C.B. Moler
J.R. Bunch G.W. Stewart

The Problem: Linpack Benchmark

- 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

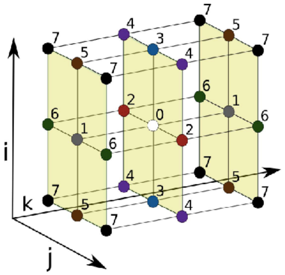
- The gap between HPL predictions and real application performance will increase in the future.
- A computer system with the potential to run HPL at 1 Exaflops is a design that may be very unattractive for real applications.
- Future architectures targeted toward good HPL performance will not be a good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

http://bit.ly/hpcg-benchmark

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).



27-point stencil operator

http://bit.ly/hpcg-benchmark

Preconditioner Setup

- Symmetric Gauss-Seidel preconditioner
  - (Non-additive Schwarz)
- In Matlab that might look like:

```
LA = tril(A); UA = triu(A); DA = diag(diag(A));

x = LA\y;
x1 = y - LA*x + DA*x; % Subtract off extra diagonal
contribution
x = UA\x1;
```

What about the NAS Parallel CG Benchmark?

- NAS CG is flawed from the perspective of modeling the design choices of real science and engineering codes.
- The matrix truly random and make the placement of entries random means that, for distributed memory machines, a 2-dimensional matrix decomposition is most effective, which is fundamentally different that the 1D processor decomposition that spatial locality in PDEs needs.
- Random also meant that the natural spatial and temporal locality properties of real sparse matrices were not present, so caches were much less useful in the benchmark than in real life.
- Finally, NAS CG has no preconditioner, so it is essentially a fast sparse MV benchmark for an atypical sparse matrix.

http://bit.ly/hpcg-benchmark

Merits of HPCG

- Provides coverage for major communication and computational patterns.
    - Represents a minimal collection of the major patterns.
- Rewards investment in high-performance collective ops.
- Rewards investment in local memory system performance.
- Detects and measures variances from bitwise identical computations.

http://bit.ly/hpcg-benchmark

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

Factors that Necessitate Redesign

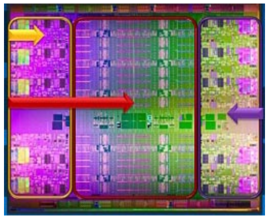
- Steepness of the ascent from terascale to petascale to exascale
- Extreme parallelism and hybrid design
  - Preparing for million/billion way parallelism
- Tightening memory/bandwidth bottleneck
  - Limits on power/clock speed implication on multicore
  - Reducing communication will become much more intense
  - Memory per core changes, byte-to-flop ratio will change
- Necessary Fault Tolerance
  - MTTF will drop
  - Checkpoint/restart has limitations
- Software infrastructure does not exist today

Key Challenges at Exascale

- Levels of parallelism
  - O(100M and beyond)
- Hybrid architectures
  - Node composed of multiple multicore sockets + accelerators
- Bandwidth vs Arithmetic rate
  - Most approaches assume flops expensive
- Storage Capacity
  - Issue of weak scalability in future systems
- Fault occurrence; shared responsibility
  - Process failure recovery
- Power Management
  - API for fine grain management
- Language constraints
  - Fortran, C & MPI, Open-MP
- Autotuning
  - Systems complex and changing
- Bulk Sync Processing
  - Break fork join parallelism
- Lack of reproducibility; unnecessarily expensive (most of the time)
  - Can't guarantee bitwise results
- Need for effective scheduling of tasks

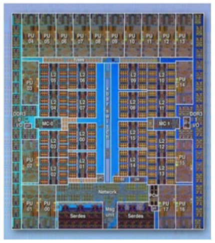
The winning architecture for building exascale systems, heterogeneous or homogeneous, and why?

- **Multicore:** Maintain complex cores, and replicate (x86, SPARC, Power7) [#3, 6, and 10]



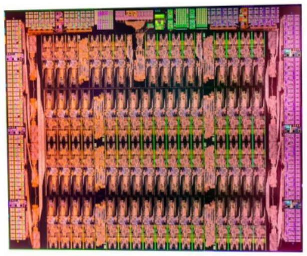
Intel Xeon E7
(10 cores)

- **Manycore/Embedded:** Use many simpler, low power cores from embedded (BlueGene, future ARM) [#2, 4, 5, and 9]



IBM BlueGene/Q
(16 +2 cores)

- **GPU/Coprocessor/Accelerator**: Use highly specialized processors from graphics market space (NVidia Fermi, Intel Xeon Phi, AMD) [# 1, 7, and 8]



Intel Xeon Phi
(60 cores)

From Horst Simon, LBNL

Critical Issues at Peta & Exascale for Algorithm and Software Design

- Synchronization-reducing algorithms
  - Break Fork-Join model
- Communication-reducing algorithms
  - Use methods which have lower bound on communication
  - Cache aware
- Mixed precision methods
  - 2x speed of ops and 2x speed for data movement
- Autotuning
  - Today's machines are too complicated, build "smarts" into software to adapt to the hardware
- Fault resilient algorithms
  - Implement algorithms that can recover from failures/bit flips
- Reproducibility of results
  - Today we can't guarantee this. We understand the issues, but some of our "colleagues" have a hard time with this.

Summary

- Major Challenges are ahead for extreme computing
  - Parallelism O(10^9)
    - Programming issues
  - Hybrid
    - Peak and HPL may be very misleading
    - No where near close to peak for most apps
  - Fault Tolerance
    - Today Sequoia BG/Q node failure rate is 1.25 failures/day
  - Power
    - 50 Gflops/w (today at 2 Gflops/w)
- We will need completely new approaches and technologies to reach the Exascale level