BERKELEY LAB

Why we need Exascale and why we won't get there by 2020

Horst Simon
Lawrence Berkeley National Laboratory

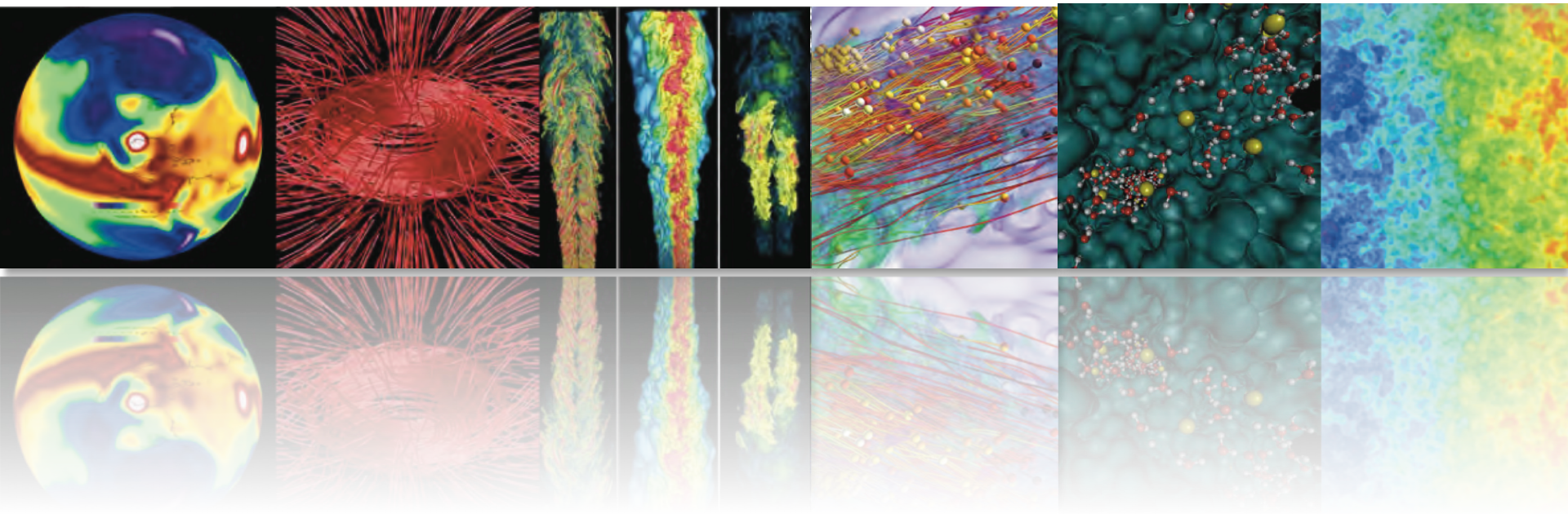
August 27, 2013

Overview

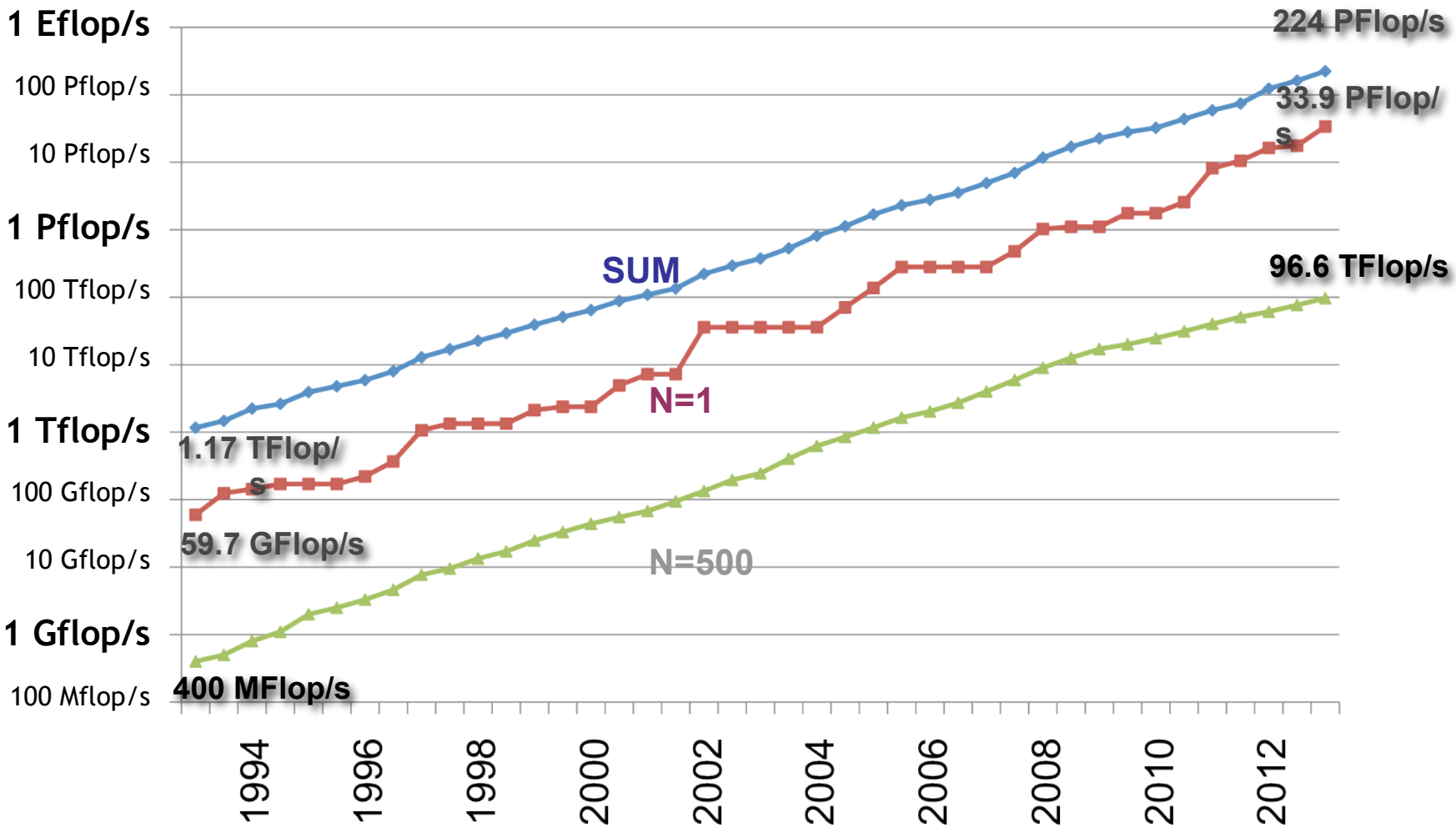
- **Current state of HPC: petaflops firmly established**
- **Why we won't get to exaflops (in 2020)**
- **The case for exascale**
  - **Science (old)**
  - **Science (new)**
  - **Technology and economic competitiveness**
  - **Discovery**

Thank you to the DOE Hexlab group, the TOP500 team, John Shalf, Kathy Yelick and many others

Current State of HPC: petaflops firmly established



Performance Development



Source: TOP500 June 2013

Top 10 Systems in June 2013

#	Site	Manufacturer	Computer	Country	Cores	Rmax [Pflops]	Power [MW]
1	National University of Defense Technology	NUDT	Tianhe-2 NUDT TH-IVB-FEP, Xeon 12C 2.2GHz, IntelXeon Phi	China	3,120,000	33.9	17.8
2	Oak Ridge National Laboratory	Cray	Titan Cray XK7, Opteron 16C 2.2GHz, Gemini, NVIDIA K20x	USA	560,640	17.6	8.21
3	Lawrence Livermore National Laboratory	IBM	Sequoia BlueGene/Q, Power BQC 16C 1.6GHz, Custom	USA	1,572,864	17.2	7.89
4	RIKEN Advanced Institute for Computational Science	Fujitsu	K Computer SPARC64 VIIIfx 2.0GHz, Tofu Interconnect	Japan	795,024	10.5	12.7
5	Argonne National Laboratory	IBM	Mira BlueGene/Q, Power BQC 16C 1.6GHz, Custom	USA	786,432	8.59	3.95
6	Texas Advanced Computing Center/UT	Dell	Stampede PowerEdge C8220, Xeon E5 8C 2.7GHz, Intel Xeon Phi	USA	462,462	5.17	4.51
7	Forschungszentrum Juelich (FZJ)	IBM	JuQUEEN BlueGene/Q, Power BQC 16C 1.6GHz, Custom	Germany	458,752	5.01	2.30
8	Lawrence Livermore National Laboratory	IBM	Vulcan BlueGene/Q, Power BQC 16C 1.6GHz, Custom	USA	393,216	4.29	1.97
9	Leibniz Rechenzentrum	IBM	SuperMUC iDataPlex DX360M4, Xeon E5 8C 2.7GHz, Infiniband FDR	Germany	147,456	2.90	3.52
10	National SuperComputer Center in Tianjin	NUDT	Tianhe-1A NUDT TH MPP, Xeon 6C, NVidia, FT-1000 8C	China	186,368	2.57	4.04

Technology Paths to Exascale

Leading Technology Paths (Swim Lanes):

Multicore	Maintain complex cores, and replicate (x86, SPARC, Power7) [#4 and 9]
Manycore / Embedded	Use many simpler, low power cores from embedded (BlueGene) [#3, 5, 7, and 8]
GPU / Accelerator	Use highly specialized processors from gaming/graphics market space (NVidia Fermi, Cell, Intel Phi (MIC)) [# 1, 2, 6, and 8]

Tianhe-2 (TH-2) at NUDT, China

Summary of the Tianhe-2 (TH-2) Milkyway 2

Model	TH-IVB-FEP
Nodes	16,000
Processor	Intel Xeon IvyBridge E5-2692
Speed	2.200 GHz
Sockets per Node:	2
Cores per Socket:	12
Coprocessors:	Intel Xeon Phi 31S1P
Coprocessors per Node:	3
Cores per Coprocessor:	57
Coprocessors total	48,000
Operating System:	Kylin Linux
Primary Interconnect:	Proprietary high-speed interconnecting network (TH Express-2)
Peak Power (MW):	17.8
Size of Power Measurement (Cores)	3,120,000
Memory per Node (GB)	64

Summary of all components

CPU Cores	384,000
Accelerators/CP	48,000
Accelerator/CP Cores	
Memory	1,024,000 GB

Summary of the Tainhe-2 (TH-2) or Milkyway-2	
Items	Configuration
Processors	32,000 Intel Xeon CPU's + 48,000 Xeon Phi's (+ 4096 FT-1500 CPU's frontend) Peak Performance 54.9 PFlop/s (just Intel parts)
Interconnect	Proprietary high-speed interconnection network, TH Express-2
Memory	1 PB
Storage	Global Shared parallel storage system, 12.4 PB
Cabinets	125 + 13 + 24 = 162 compute/communication/storage cabinets
Power	17.8 MW
Cooling	Closed air cooling system

BERKELEY LAB
Lawrence Berkeley National Laboratory

Sequoia at Lawrence Livermore National Laboratory (Manycore/embedded)

- Fully deployed in June 2012
- 16.32475 petaflops Rmax and 20.13266 petaflops Rpeak
- 7.89 MW Power consumption
- IBM Blue Gene/Q
- 98,304 compute nodes
- 1.6 million processor cores
- 1.6 PB of memory
- 8 or 16 core Power Architecture processors built on a 45 nm fabrication method
- Lustre Parallel File System

K Computer at RIKEN Advanced Institute for Computational Science (Multicore)

- #3 on the TOP500 list
- Distributed memory architecture
- Manufactured by Fujitsu
- Performance: 10.51 PFLOPS Rmax and 11.2804 PFLOPS Rpeak
- 12.65989 MW Power Consumption, 824.6 GFLOPS/Kwatt
- Annual running costs - $10 million
- 864 cabinets:
  - 88,128 8-core SPARC64 VIIIfx processors @ 2.0 GHz
  - 705,024 cores
  - 96 computing nodes + 6 I/O nodes per cabinet
- Water cooling system minimizes failure rate and power consumption

Choice of Swim Lanes

- **Swim Lane choices – currently all three are equally represented**
- **Risk in Swim Lane switch for large facilities**
  - **Select too soon:** Users cannot follow
  - **Select too late**: Fall behind performance curve
  - **Select incorrectly:** Subject users to multiple disruptive technology change

Events of 2011/2012 and the Big Questions for the Next Three Years until 2015

Leading Technology Paths (Swim Lanes):

IBM cancels "Blue Waters" contract.
Is this an indication that multicore with complex cores is nearing the end of the line?

Multicore

Blue Gene is the last of the line.
Will there be no more large scale embedded multicore machines?

Manycore / Embedded

Intel bought Cray's interconnect technology and WhamCloud and introduced Xeon-Phi.
Will Intel develop complete systems?

GPU / Accelerator

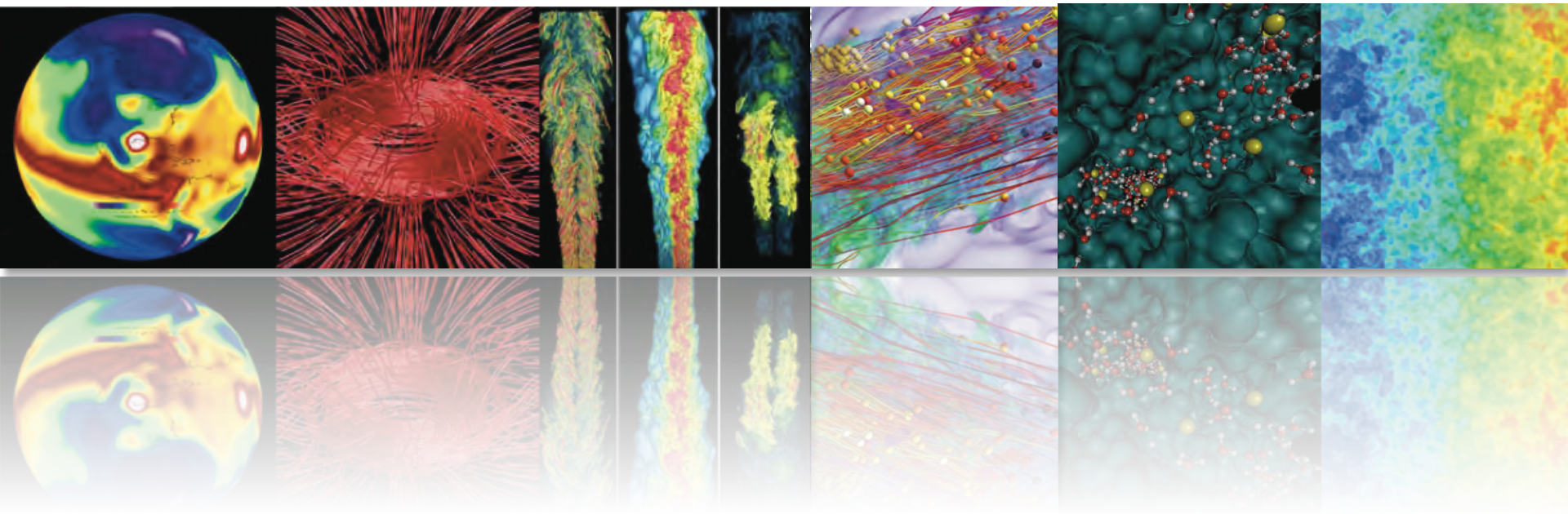
Prediction: All Top 10 systems in 2015 will be GPU/Accelerator based

State of Supercomputing in 2013

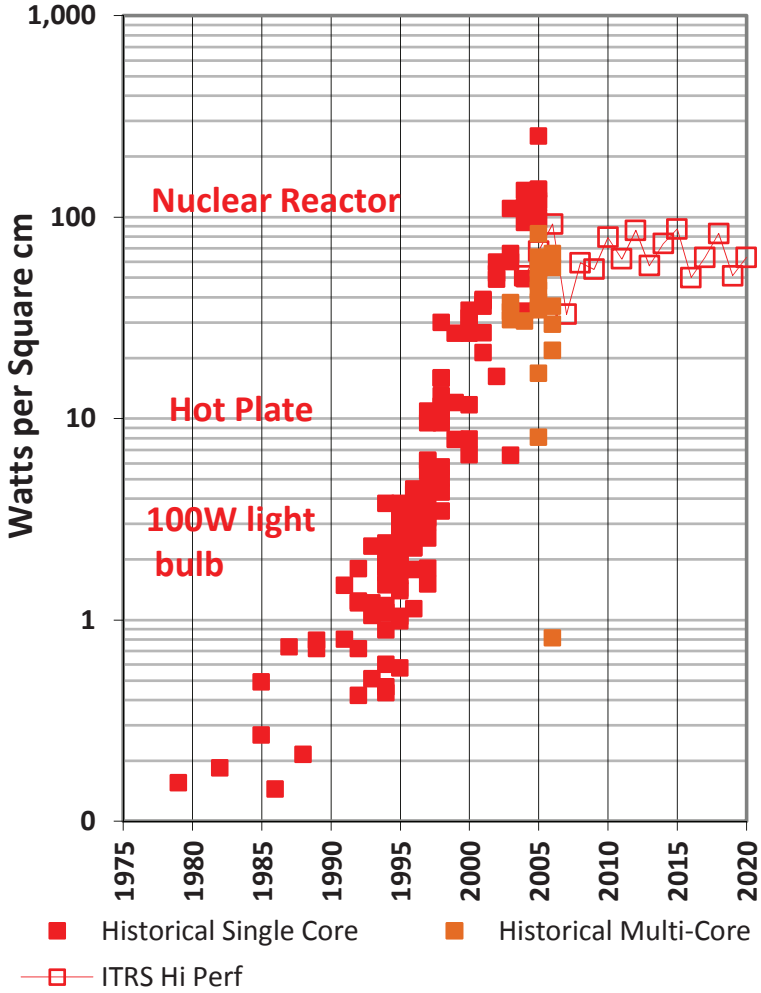
- Pflops computing fully established with more than 20 machines
- Three technology "swim lanes" are thriving
- Interest in supercomputing is now worldwide, and growing in many new markets
- Exascale projects in many countries and regions
- Rapid growth predicted by IDC for the next three years

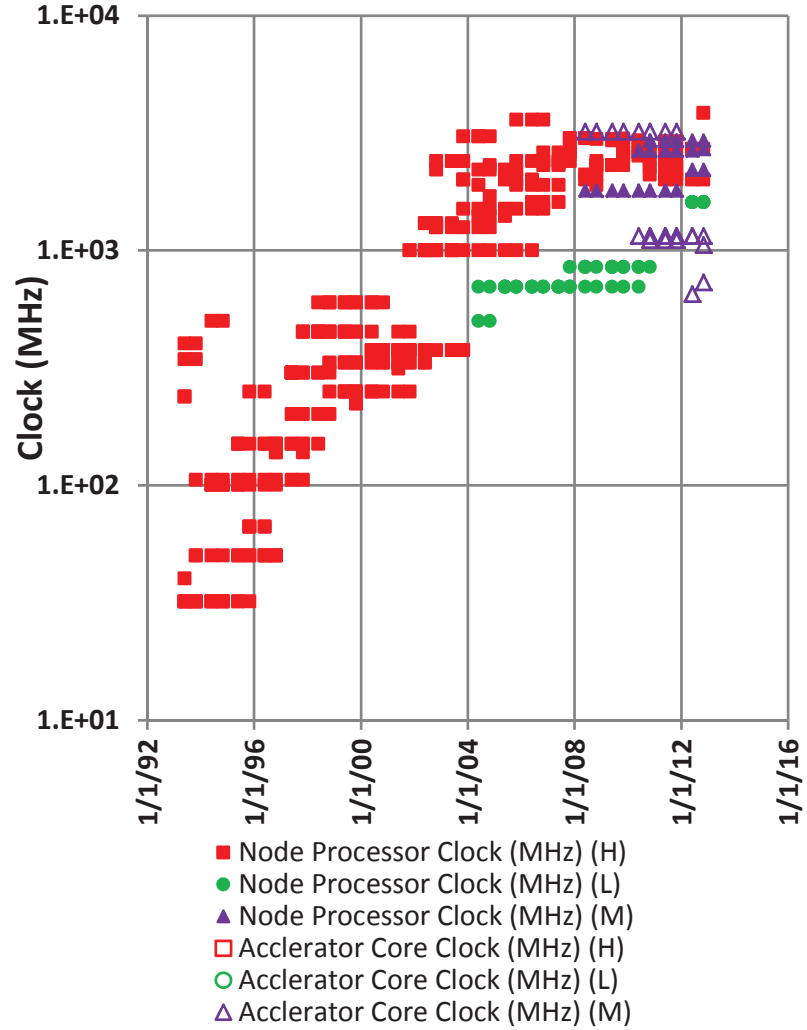
BERKELEY LAB
Lawrence Berkeley National Laboratory

Why we won't get to Exaflops by 2020



The Power and Clock Inflection Point in 2004

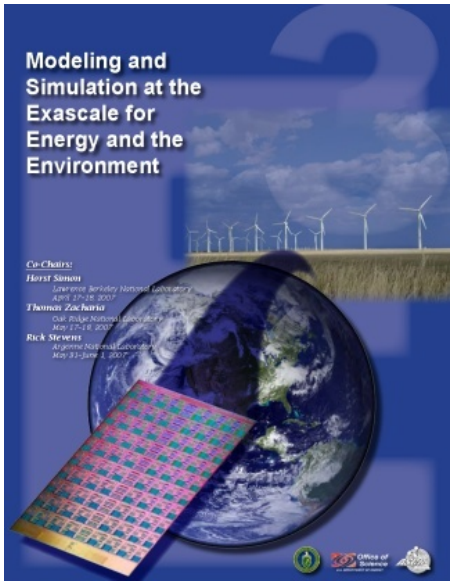


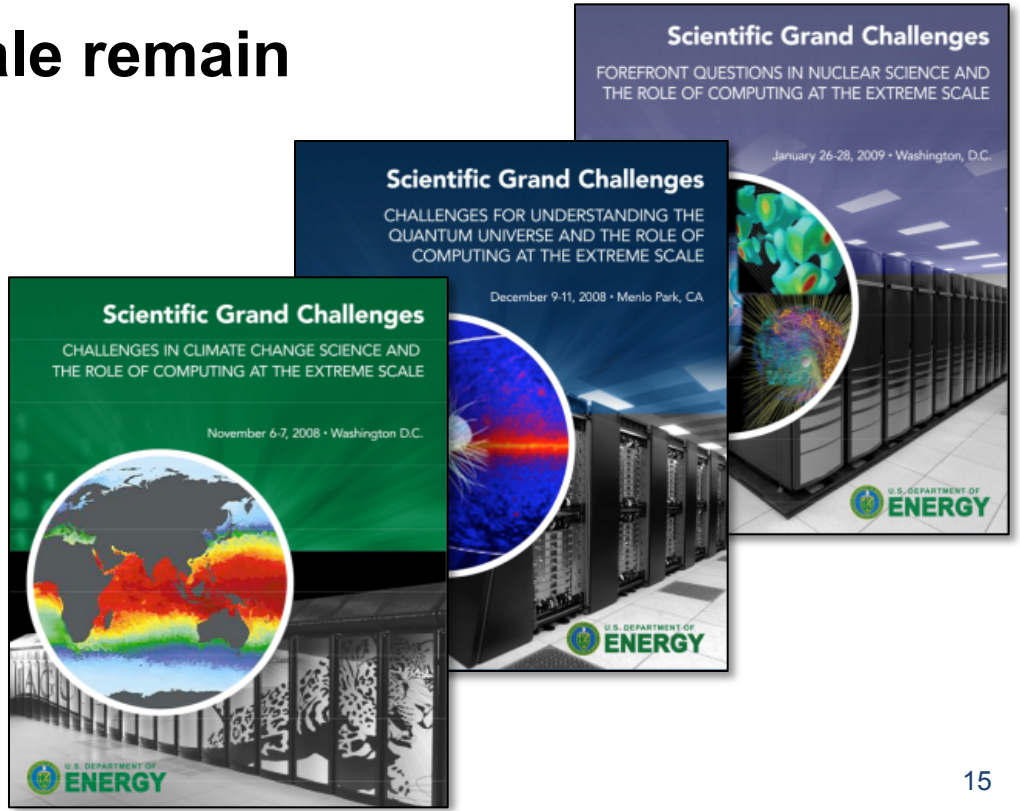


Source: Kogge and Shalf, IEEE CISE

Towards Exascale

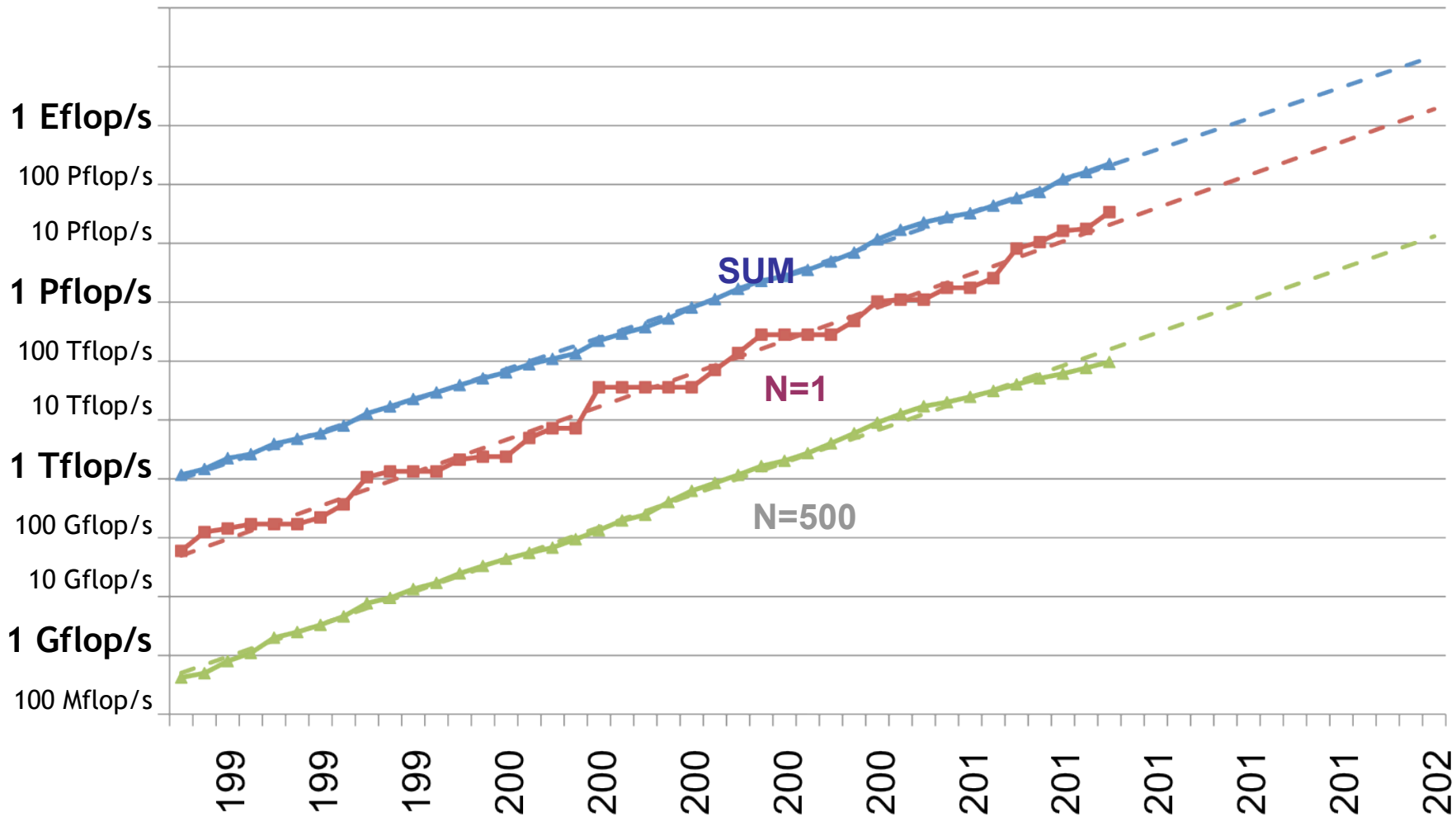
- **Exascale has been discussed in numerous workshops, conferences, planning meetings for about six years**
- **Exascale projects have been started in the U.S. and many other countries and regions**
- **Key challenges to exascale remain**





BERKELEY LAB
Lawrence Berkeley National Laboratory

Projected Performance Development



Data from: TOP500 June 2013

The Exascale Challenge

- There is a lot of vagueness in the exascale discussion and what it means to reach exascale. Let me propose a concrete and measurable goal.
- Build a system before 2020 that will be #1 on the TOP500 list with an Rmax > 1 exaflops
- Personal bet (with Thomas Lippert, Jülich, Germany) that we will not reach this goal by November 2019 (for $2,000 or €2000)
- Unfortunately, I think that I will win the bet. But I would rather see an exaflops before the end of the decade and loose my bet.

BERKELEY LAB
Lawrence Berkeley National Laboratory

At Exascale, even LINPACK is a Challenge

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

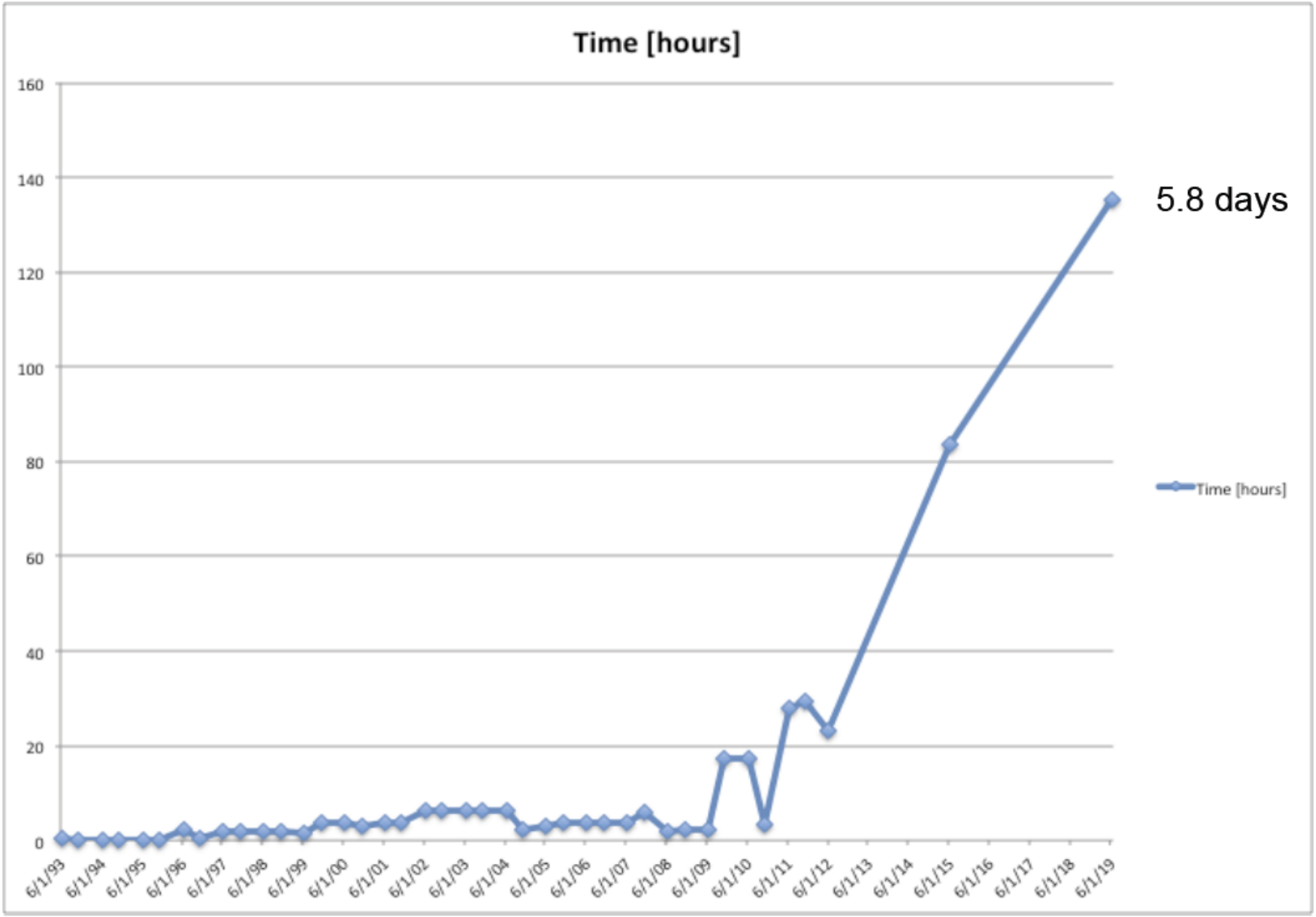
9 6 2

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	17.8

Source: Jack Dongarra **http://bit.ly/hpcg-benchmark**

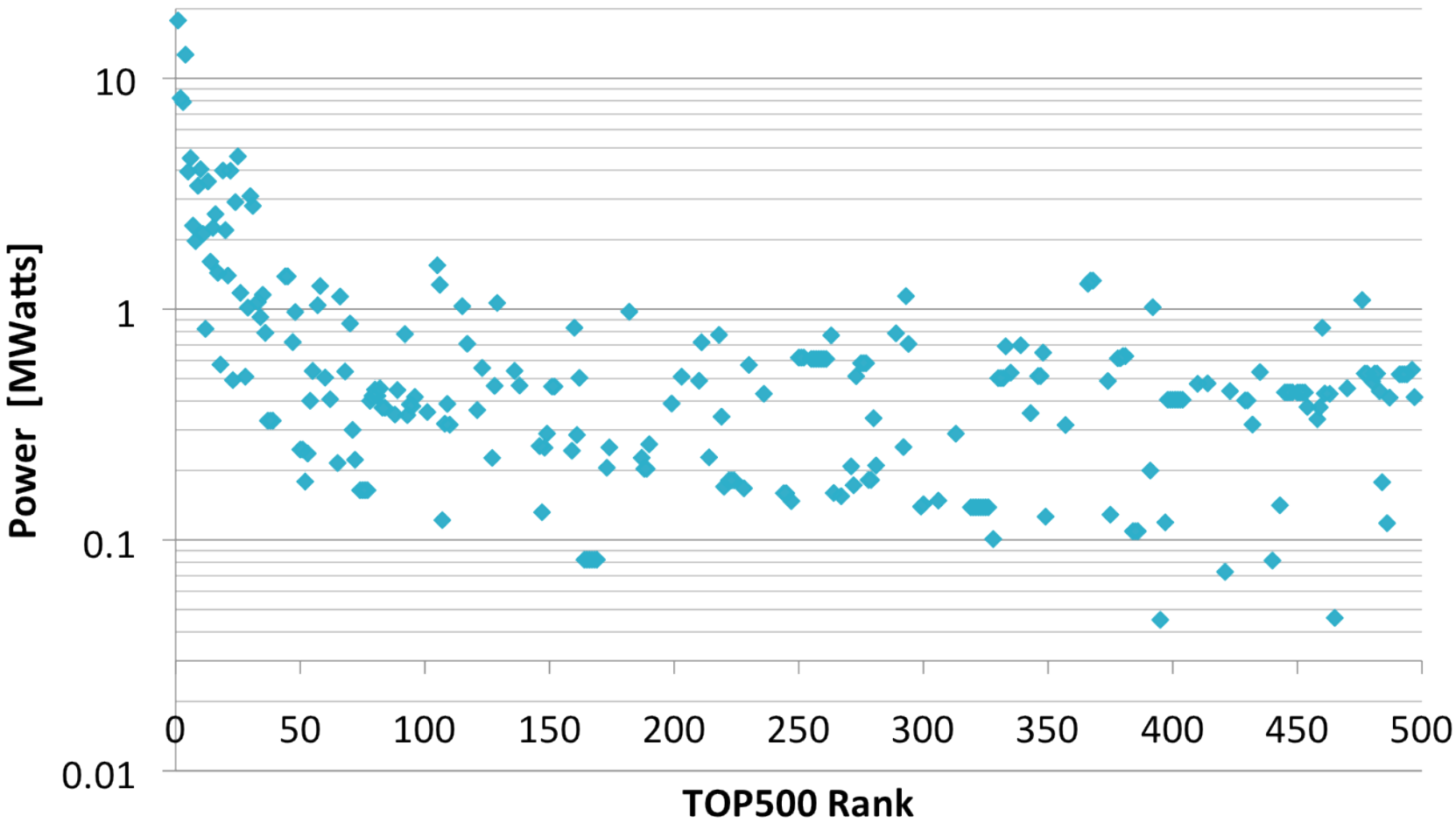
BERKELEY LAB
Lawrence Berkeley National Laboratory

At Exascale HPL will take 5.8 days



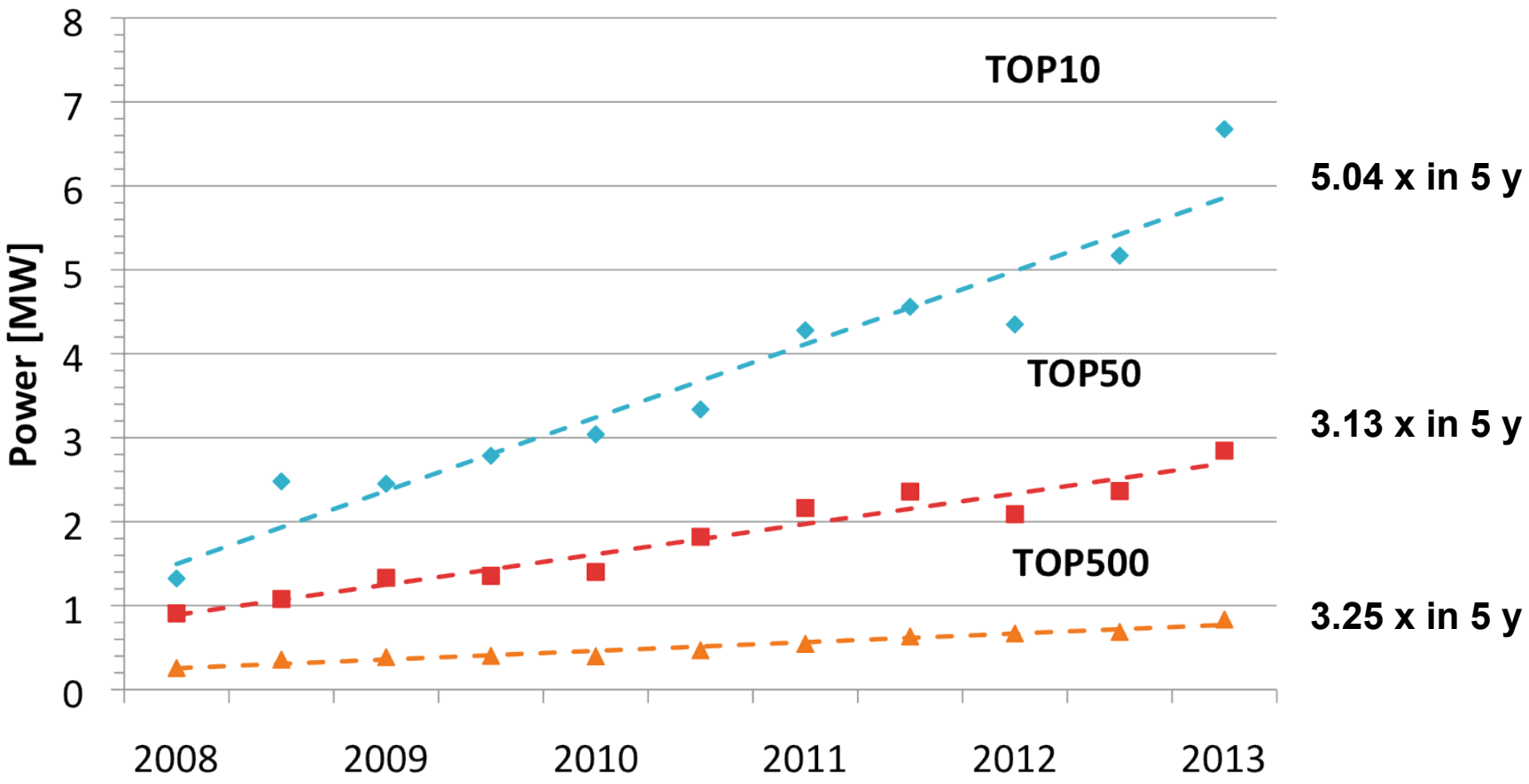
Source: Jack Dongarra, ISC'12

Total Power Levels (MW) for TOP500 Systems



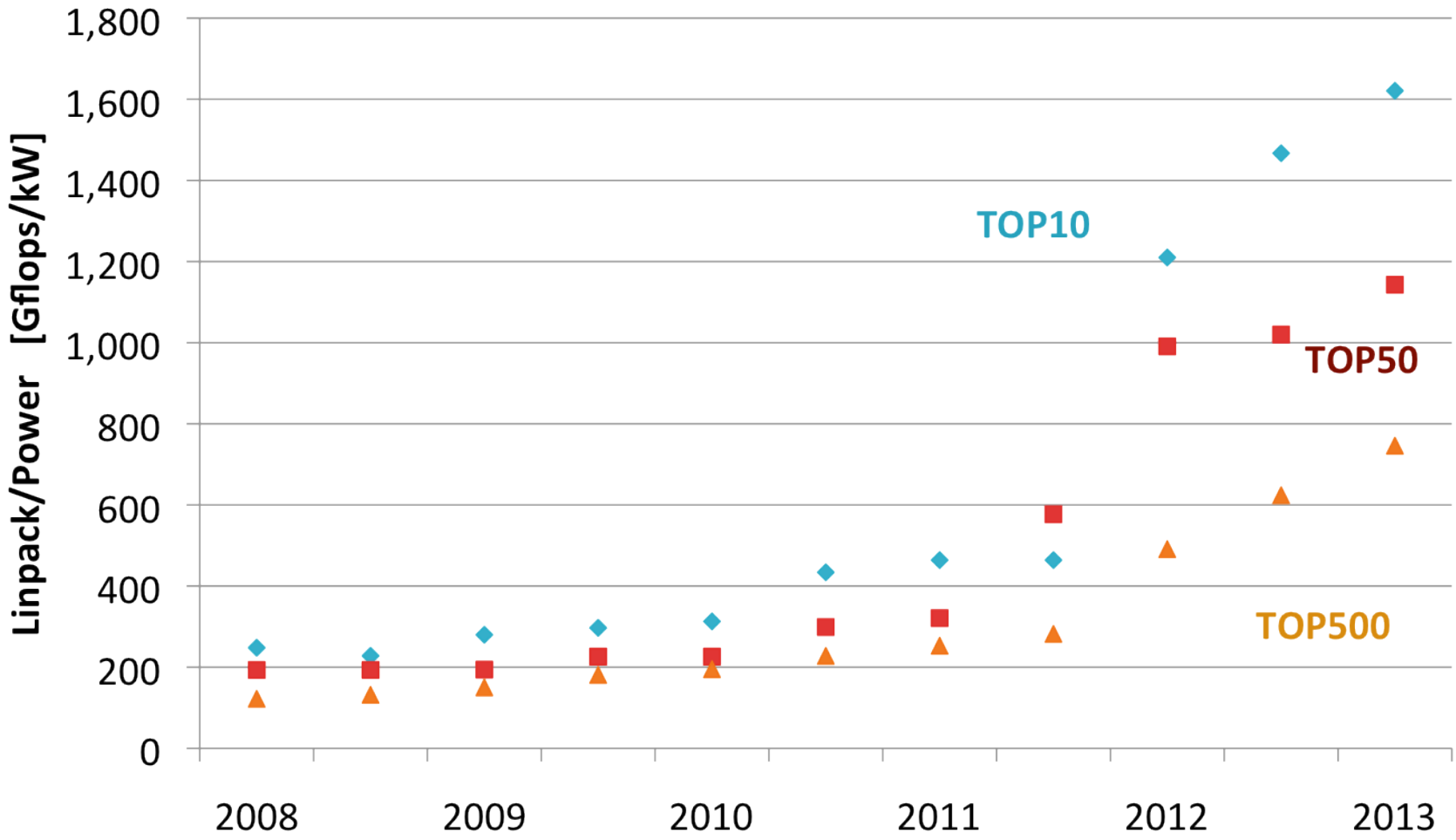
Source: TOP500 June 2013

Power Consumption



Source: TOP500 June 2013

Power Efficiency went up significantly in 2012



Data from: TOP500 June 2013

Most Power Efficient Architectures

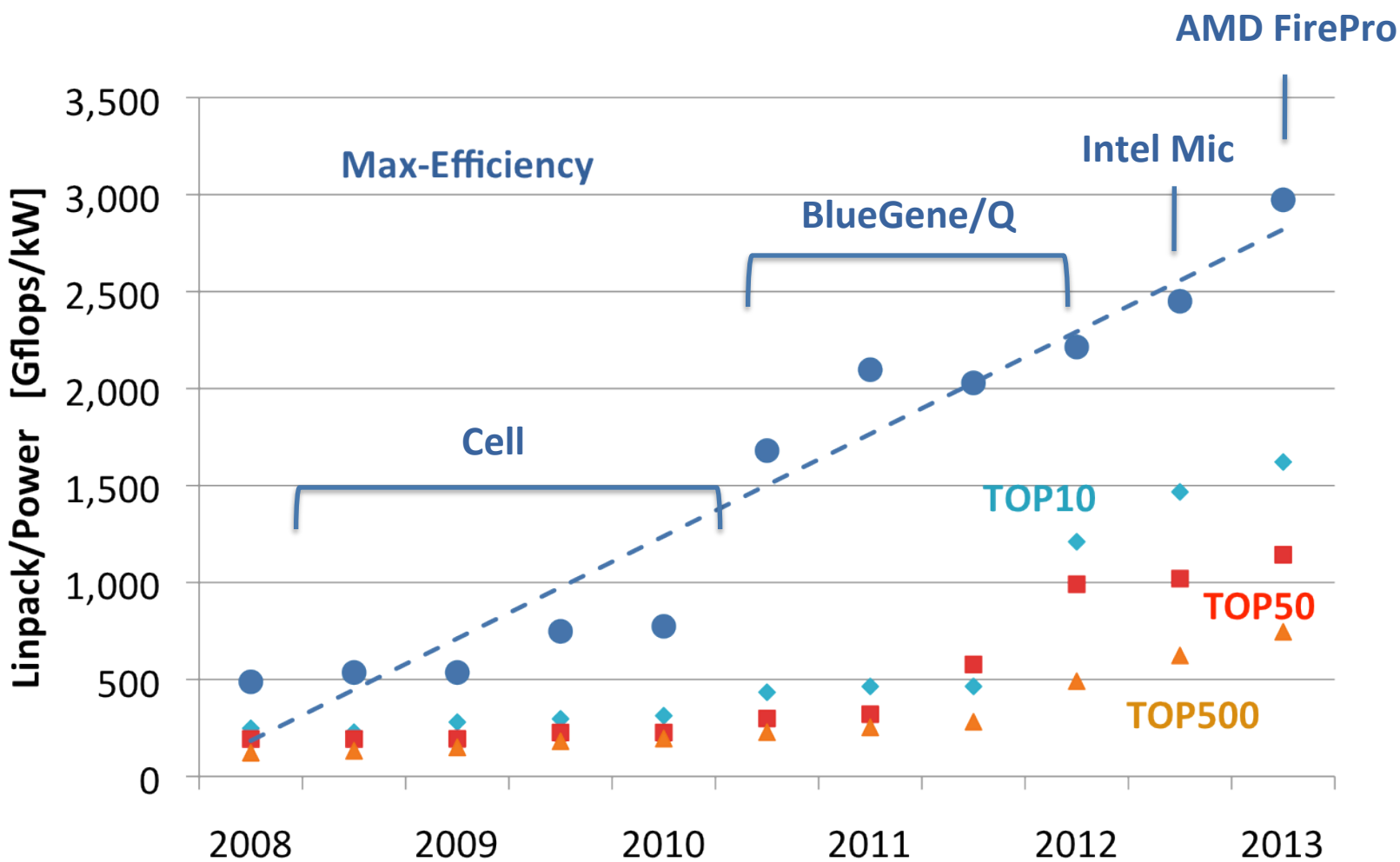
Computer	Rmax/ Power
Adtech, ASUS, Xeon 8C 2.0GHz, Infiniband FDR, **AMD FirePro**	**2,973**
Appro GreenBlade, Xeon 8C 2.6GHz, Infiniband FDR, **Intel Xeon Phi**	**2,450**
BlueGene/Q, Power BQC 16C 1.60 GHz, Custom	**2,300**
Cray XK7, Opteron 16C 2.1GHz, Gemini, **NVIDIA Kepler**	**2,243**
Eurotech Aurora HPC, Xeon 8C 3.1GHz, Infiniband QDR, **NVIDIA K20**	**2,193**
iDataPlex DX360M4, Xeon 8C 2.6GHz, Infiniband QDR, **Intel Xeon Phi**	**1,935**
Tianhe-2, NUDT, Intel Xeon 6C 2.2GHz, TH Express-2, **Intel Xeon Phi**	**1,902**
RSC Tornado, Xeon 8C 2.9GHz, Infiniband FDR, **Intel Xeon Phi**	**1,687**
SGI Rackable, Xeon 8C 2.6GHz, Infiniband FDR, **Intel Xeon Phi**	**1,613**
Chundoong Cluster, Xeon 8C 2GHz, Infiniband QDR, **AMD Radeon HD**	**1,467**

[Mflops/Watt]

BERKELEY LAB
Lawrence Berkeley National Laboratory

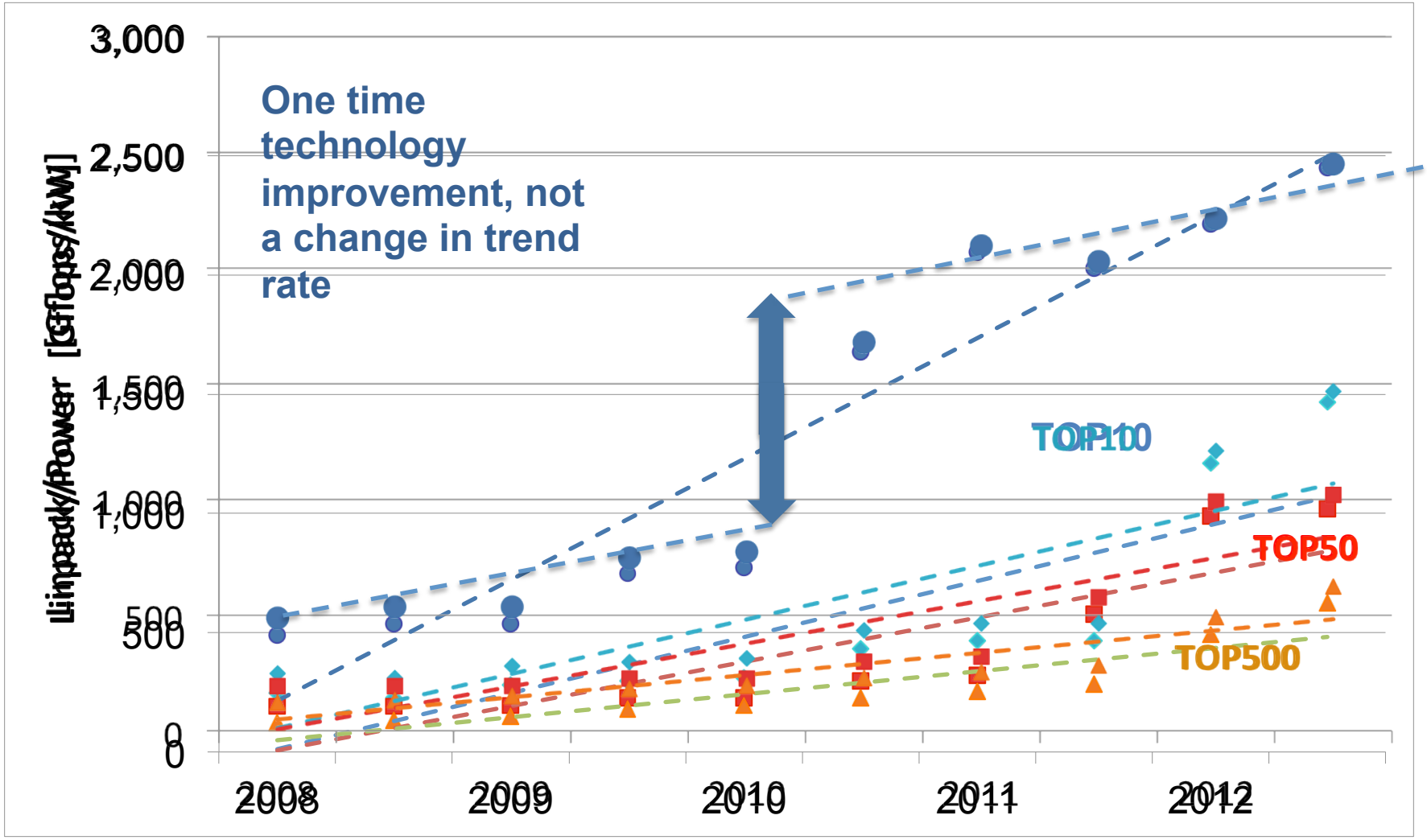
Data from: TOP500 June 2013

Power Efficiency over Time



Data from: TOP500 June 2013

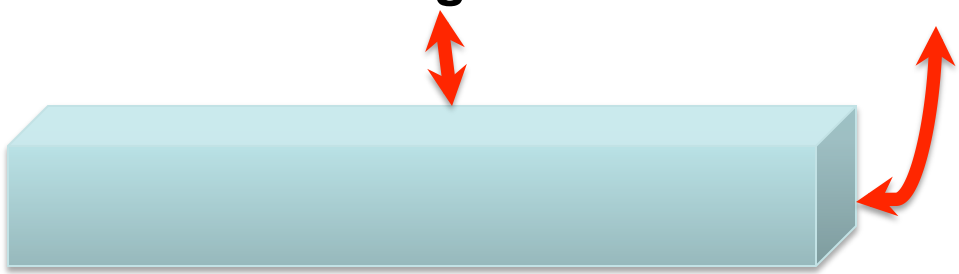
Power Efficiency over Time



Data from: TOP500 November 2012

The Problem with Wires:
Energy to move data proportional to distance

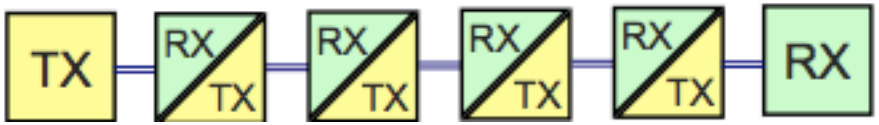
- **Cost to move a bit on copper wire:**
  - **power= bitrate * Length / cross-section area**



- **Wire data capacity constant as feature size shrinks**
- ***Cost to move bit proportional to distance***
- ***~1TByte/sec max feasible off-chip BW (10GHz/pin)***
- ***Photonics reduces distance-dependence of bandwidth***

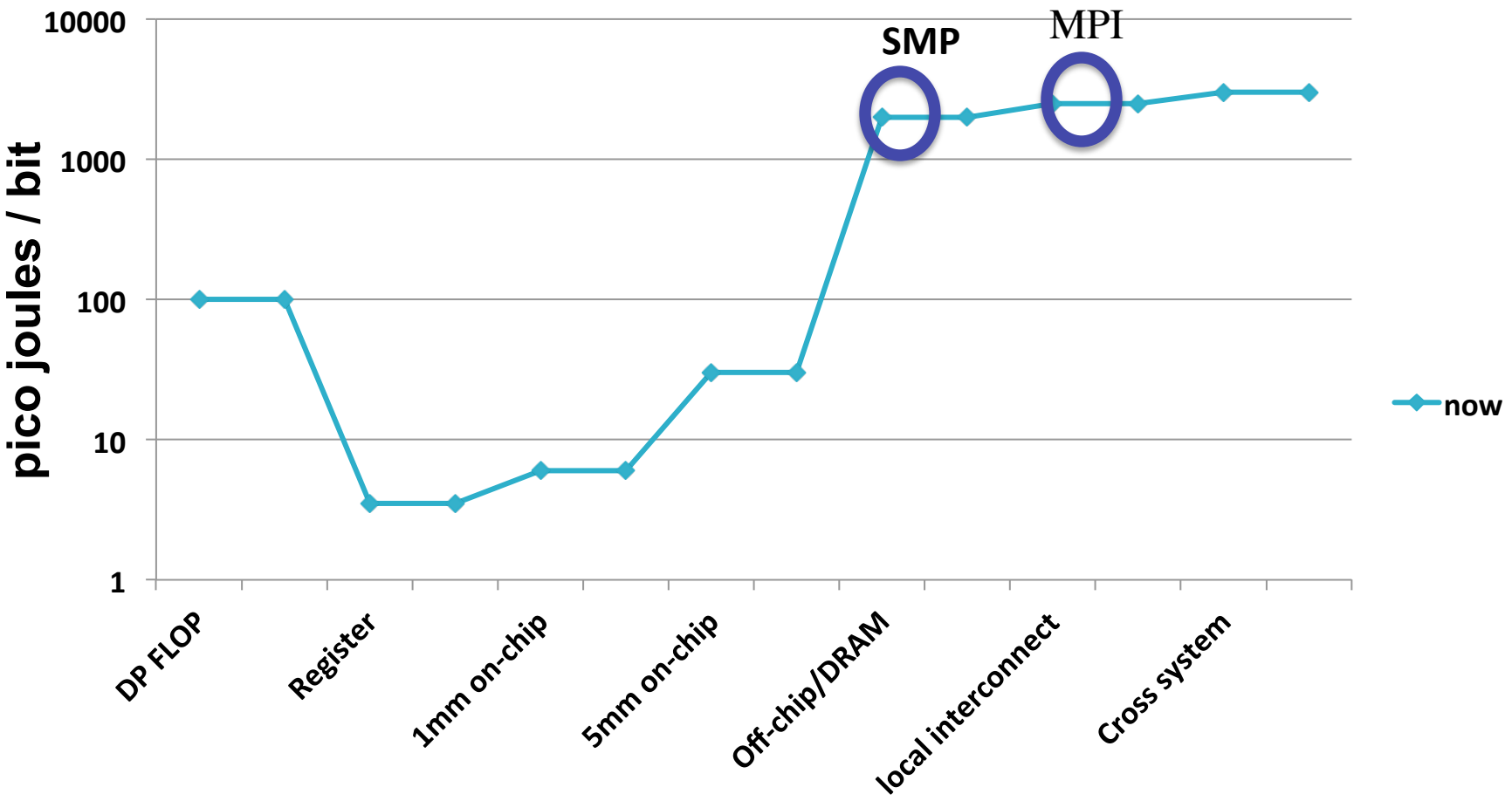
Photonics requires no redrive and passive switch little power

Copper requires to signal amplification even for on-chip connections



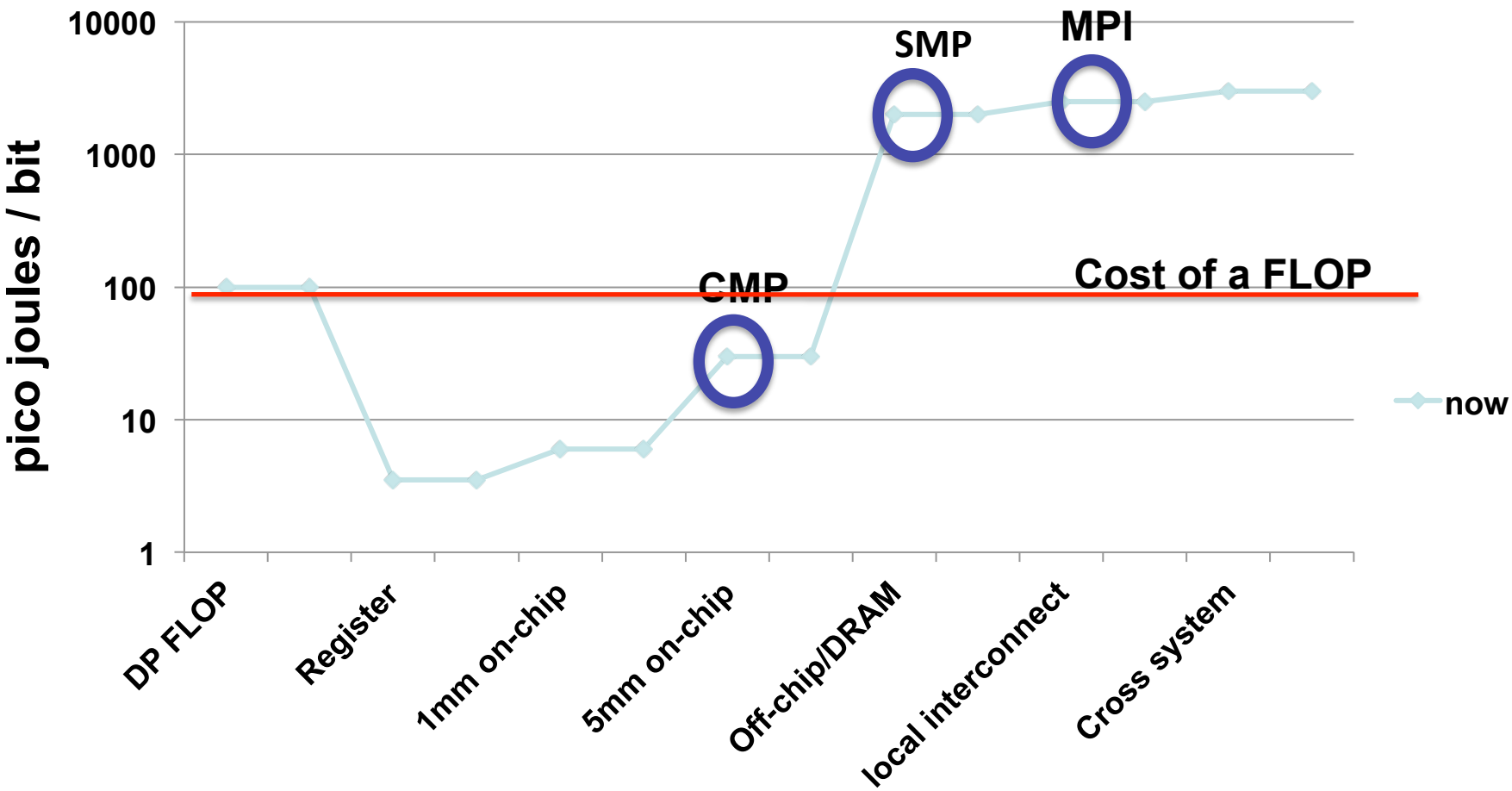
Adapted from John Shalf

The Cost of Data Movement



Adapted from John Shalf

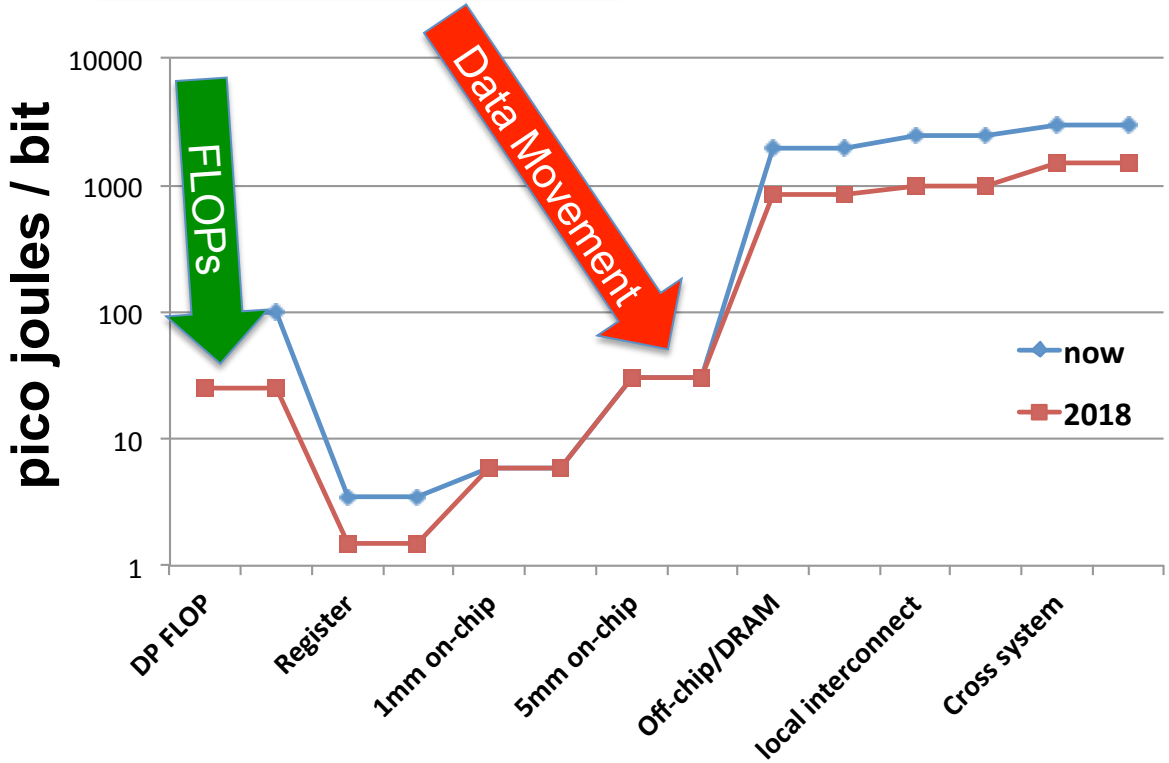
The Cost of Data Movement



Adapted from John Shalf

The Cost of Data Movement in 2018

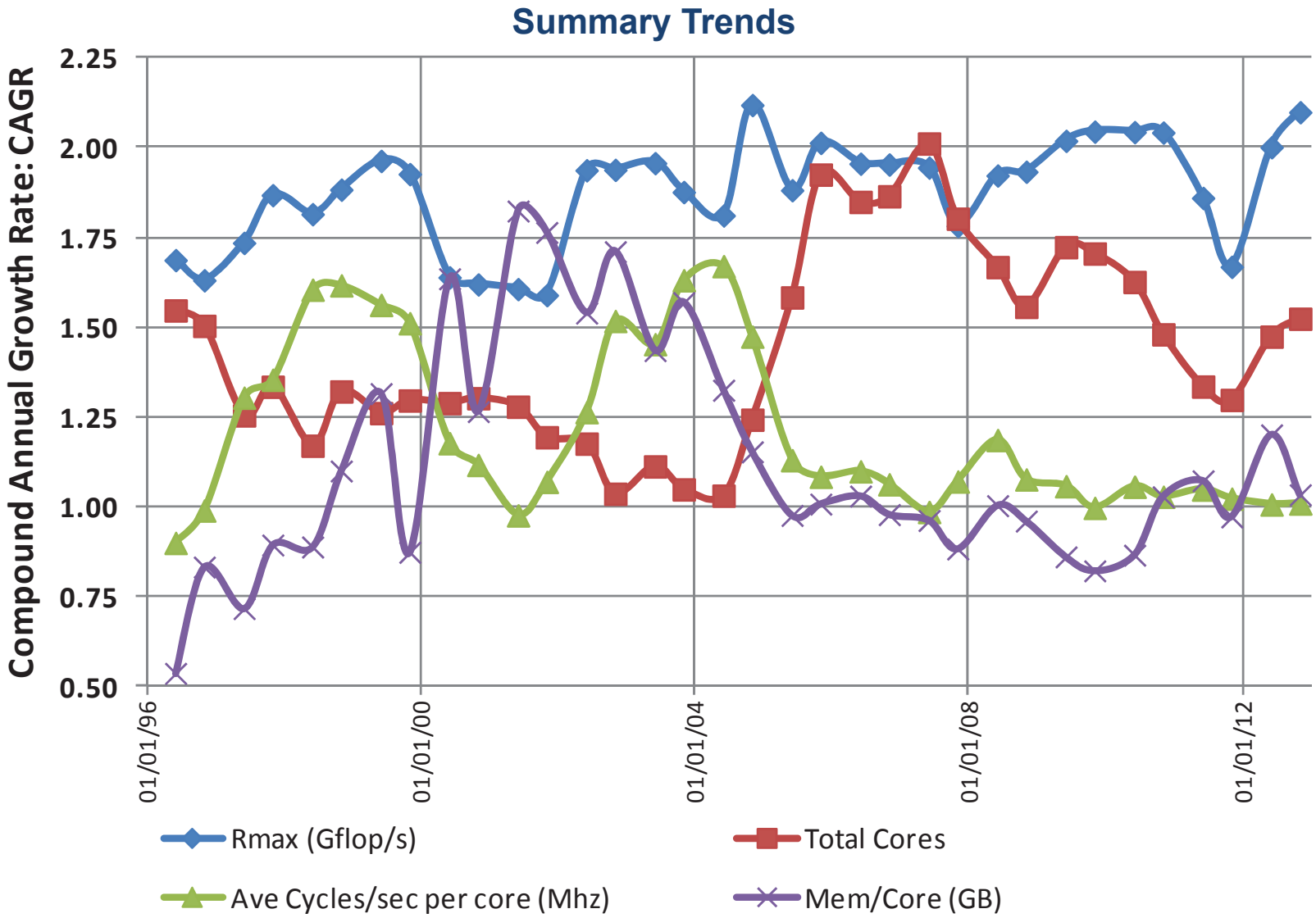
FLOPs will cost less than on-chip data movement! (NUMA)



None of the fast forward projects addresses this challenge

Adapted from John Shalf

It's the End of the World as We Know It!



Source: Kogge and Shalf, IEEE CISE

Why I believe that I will win my bet

1. HPL at Exaflop/s level will take about a week to run challenge for center directors and new systems
2. We realized a one time gain in power efficiency by switching to accelerator/manycore. This is not a sustainable trend in the absence of other new technology.
3. Data movement will cost more than flops (even on chip)
4. Limited amount of memory, low memory/flop ratios (processing is free)

BERKELEY LAB
Lawrence Berkeley National Laboratory

The Logical Conclusion

If FLOPS are free, then why do we need an "exaflops" initiative?

"Exa"-anything has become a bad brand

- **Associated with buying big machines for the labs**
- **Associated with "old" HPC**
- **Sets up the community for "failure" if "goal" can't be met**

It is not just "exaflops" – we are changing the whole computational model

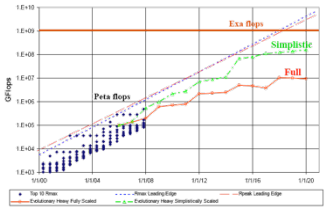
Current programming systems have WRONG optimization targets

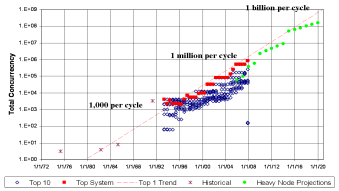
Old Constraints

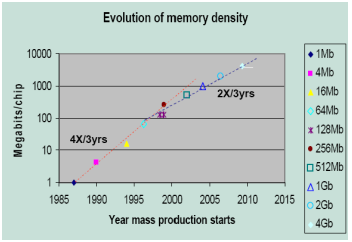
- **Peak clock frequency** *as primary limiter for performance improvement*
- **Cost:** *FLOPs* are biggest cost for system: *optimize for compute*
- **Concurrency**: *Modest growth of parallelism by adding nodes*
- **Memory scaling**: *maintain byte per flop capacity and bandwidth*
- **Locality**: *MPI+X model (uniform costs within node & between nodes)*
- **Uniformity**: *Assume uniform system performance*
- **Reliability**: *It's the hardware's problem*

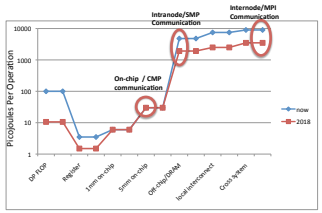
New Constraints

- **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*
- **Locality**: *must reason about data locality and possibly topology*
- **Heterogeneity**: *Architectural and performance non-uniformity increase*
- **Reliability**: *Cannot count on hardware protection alone*





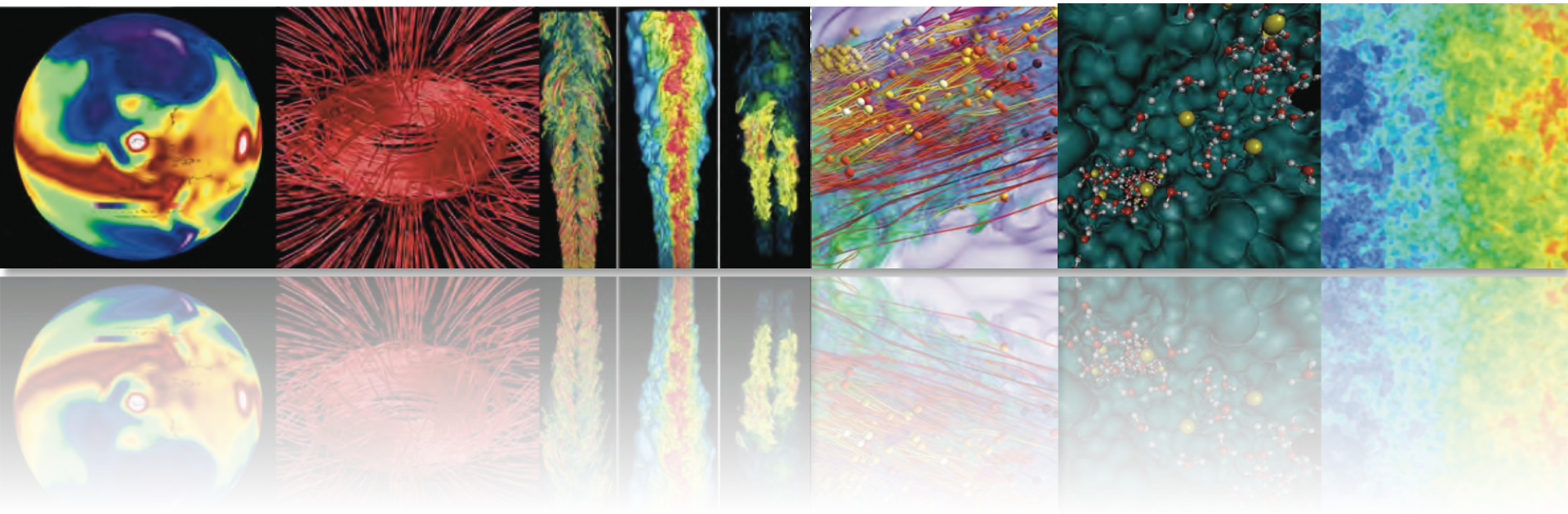




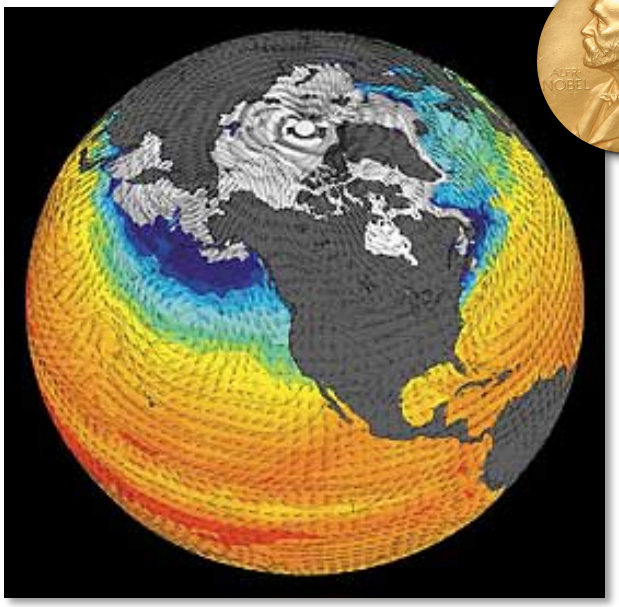
Fundamentally breaks our current programming paradigm and computing ecosystem

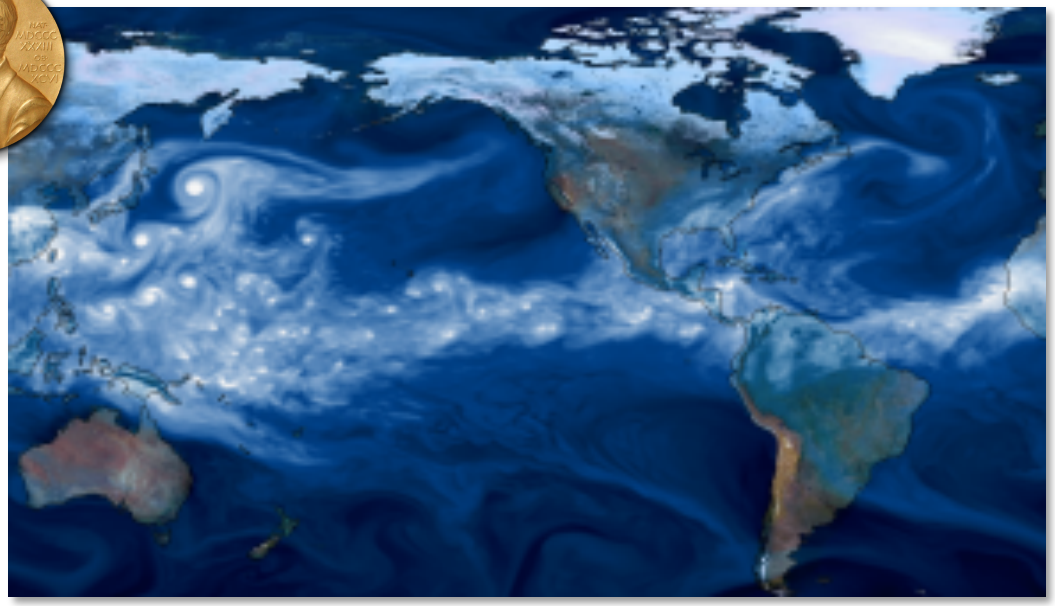
Adapted from John Shalf

The Science Case for Exascale (old)



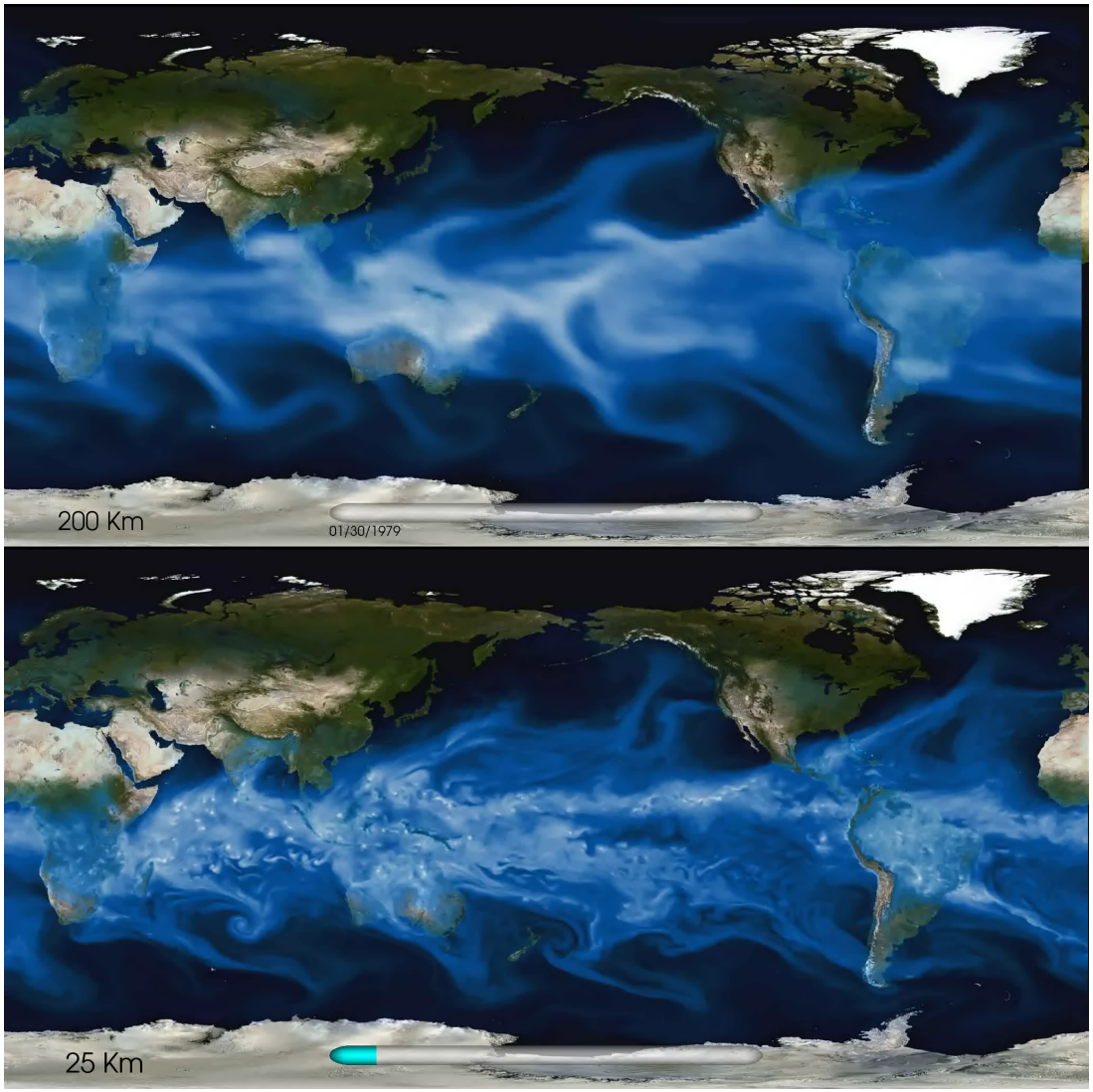
Climate change analysis





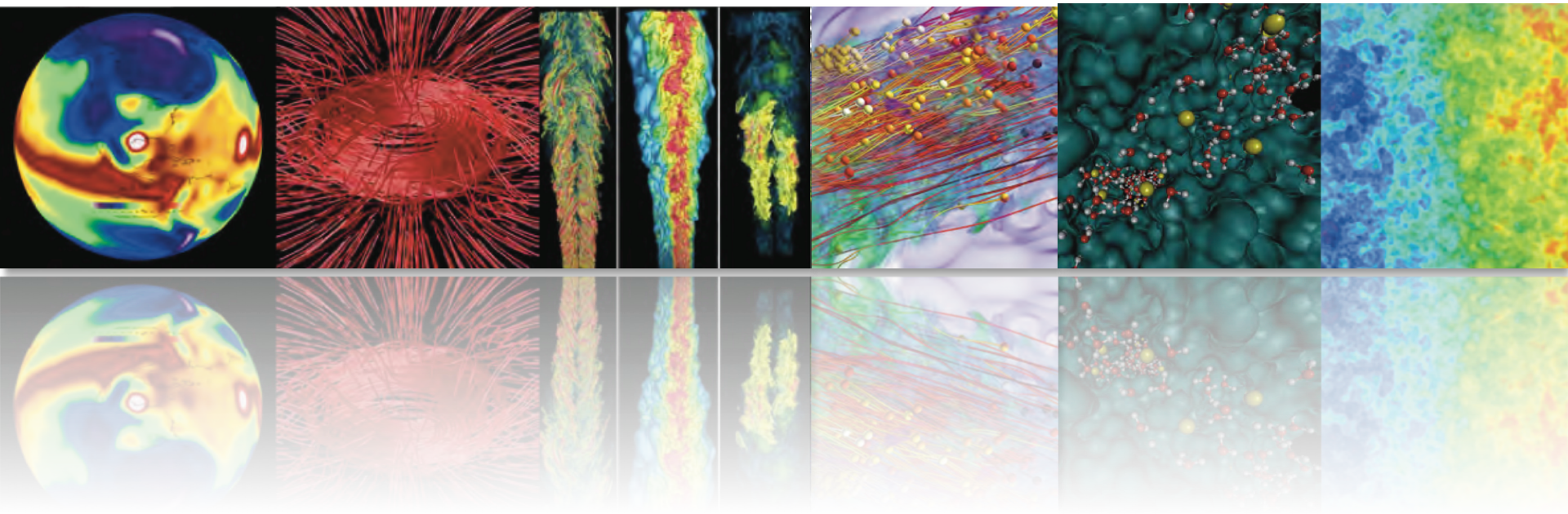
Simulations	Extreme data
• Cloud resolution, quantifying uncertainty, understanding tipping points, etc., will drive climate to exascale platforms • New math, models, and systems support will be needed	• “Reanalysis” projects need 100× more computing to analyze observations • Machine learning and other analytics are needed today for petabyte data sets • Combined simulation/observation will empower policy makers and scientists

Qualitative Improvement of Simulation with Higher Resolution (2011)



Michael Wehner, Prabhat, Chris Algieri, Fuyu Li, Bill Collins, Lawrence Berkeley National Laboratory; Kevin Reed, University of Michigan; Andrew Gettelman, Julio Bacmeister, Richard Neale, National Center for Atmospheric Research

The Science Case for Exascale (new)



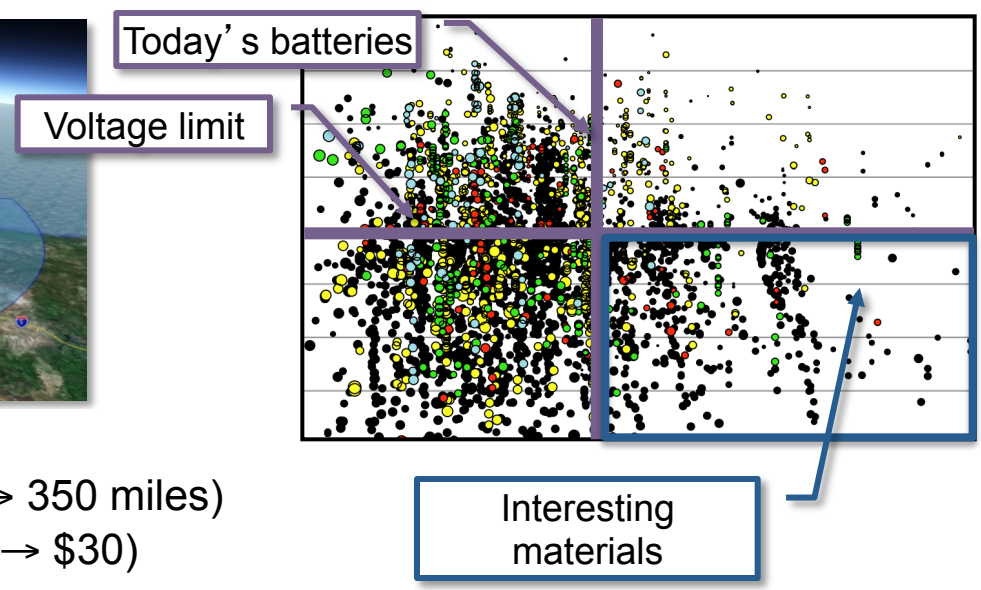
Materials Genome

Computing 1000× today

- Key to DOE's Energy Storage Hub
- Tens of thousands of simulations used to screen potential materials
- Need more simulations and fidelity for new classes of materials, studies in extreme environments, etc.

Data services for industry and science

- Results from tens of thousands of simulations web-searchable
- Materials Project launched in October 2012, now has >3,000 registered users
- Increase U.S. competitiveness; cut in half 18 year time from discovery to market



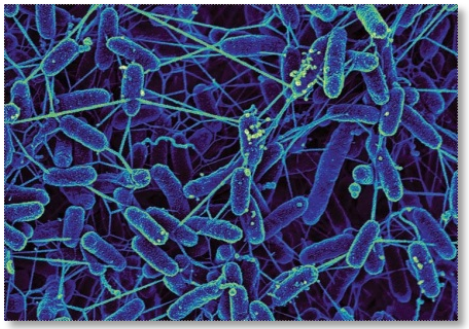
By 2018:

•Increase energy density (70 miles → 350 miles)

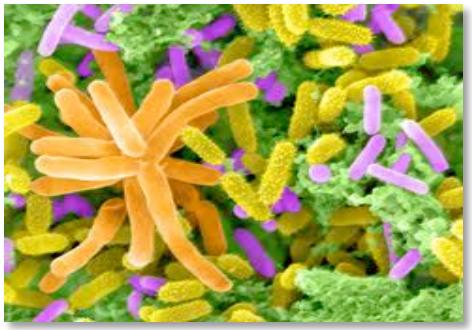
•Reduce battery cost per mile ($150 → $30)

DOE Systems Biology Knowledgebase: KBase

- **Integration and modeling for predictive biology**
- **Knowledgebase enabling predictive systems biology**
  - Powerful modeling framework
  - Community-driven, extensible and scalable open-source software and application system
  - Infrastructure for integration and reconciliation of algorithms and data sources
  - Framework for standardization, search, and association of data
  - Resource to enable experimental design and interpretation of results



Microbes



Communities

Plants

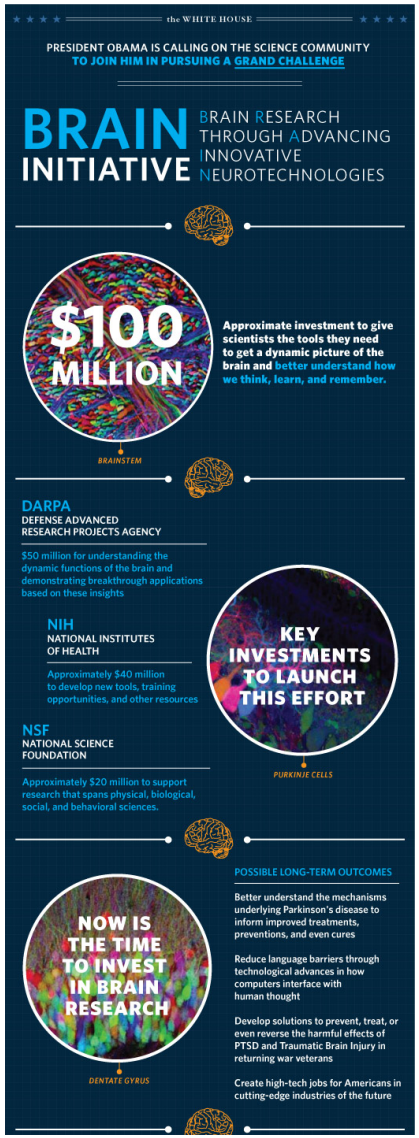
KBase
PREDICTIVE BIOLOGY

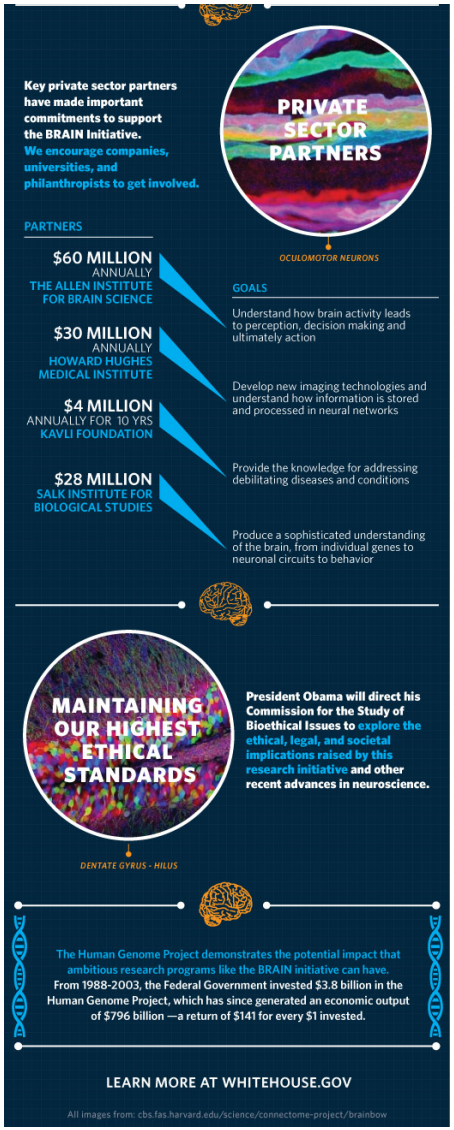
BERKELEY LAB
Lawrence Berkeley National Laboratory

Obama Announces BRAIN Initiative – Proposes $100 Million in his FY2014 Budget

- Brain Research through Advancing Innovative Neurotechnologies
- Create real-time traffic maps to provide new insights into brain disorders
- "There is this enormous mystery waiting to be unlocked, and the BRAIN Initiative will change that by giving scientists the tools they need to get a dynamic picture of the brain in action and better understand how we think and how we learn and how we remember,"

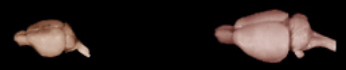
Barack Obama





See Alivisatos et al., Neuron 74, June 2012, pg 970

Modha Group at IBM Almaden



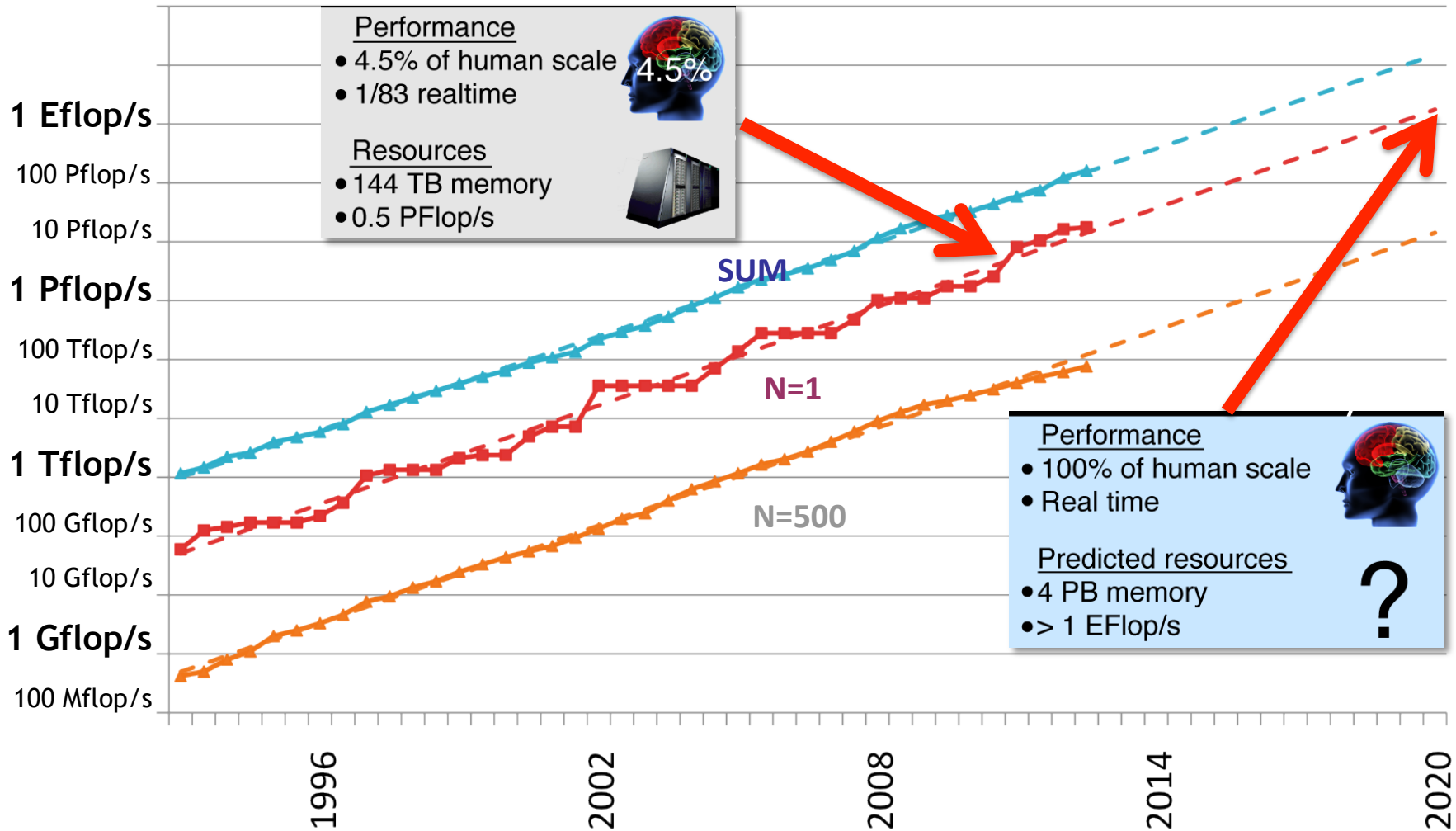
	Mouse	Rat	Cat	Monkey	Human
N:	16×10^6	56×10^6	763×10^6	2×10^9	22×10^9
S:	128×10^9	448×10^9	6.1×10^{12}	20×10^{12}	220×10^{12}

New results for SC12

Almaden	Watson	WatsonShaheen	LLNL Dawn	LLNL Sequoia
BG/L	BG/L	BG/P	BG/P	BG/Q
December, 2006	April, 2007	March, 2009	May, 2009	June, 2012

Latest simulations in 2012 achieve unprecedented scale of 65×10^9 neurons and 16×10^{12} synapses

Towards Exascale



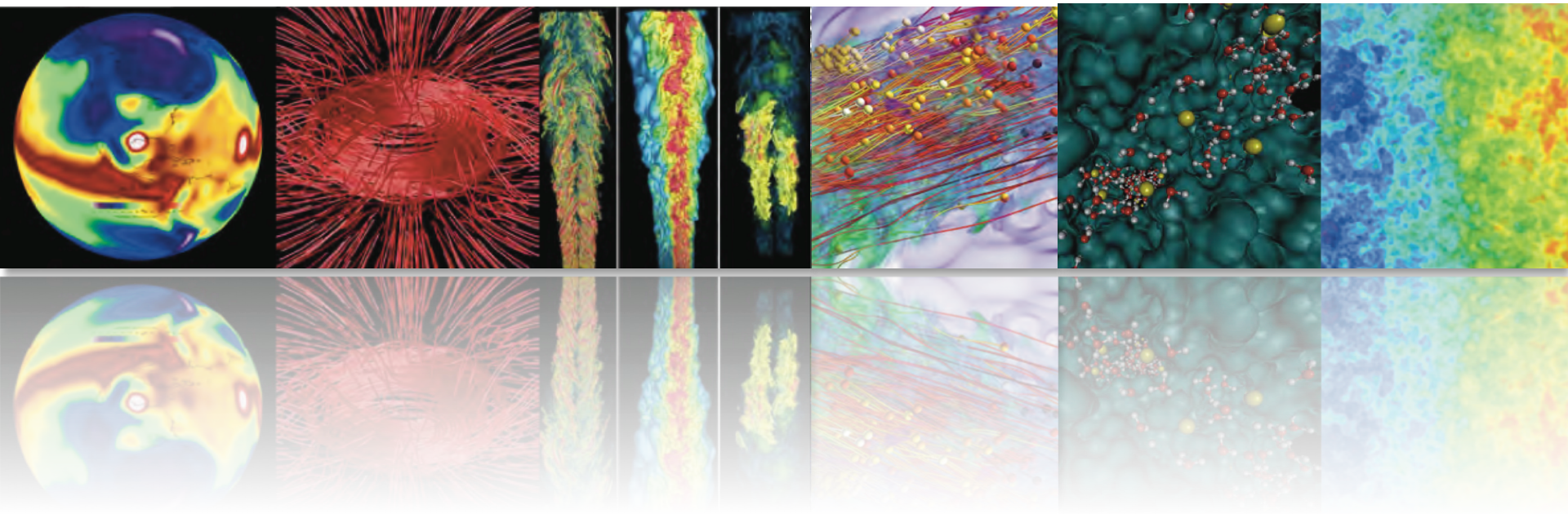
Data: TOP500 November 2012

Towards Exascale – the Power Conundrum

- **Straight forward extrapolation results in a real time human brain scale simulation at about 1 - 10 Exaflop/s with 4 PB of memory**
- **Current predictions envision Exascale computers in 2020 with a power consumption of at best 20 - 30 MW**
- **The human brain takes 20W**
- **Even under best assumptions in 2020 our brain will still be a million times more power efficient**

BERKELEY LAB
Lawrence Berkeley National Laboratory

Exascale is Critical to the Nation to Develop Future Technologies and Maintain Economic Competitiveness



U.S. competitive advantage demands exascale resources

- Digital design and prototyping at exascale enable rapid delivery of new products to market by minimizing the need for expensive, dangerous, and/or inaccessible testing
- Potential key differentiator for American competitiveness
- Strategic partnerships between DOE labs and industrial partners to develop and scale applications to exascale levels

Exascale computing is key for national security

"I think that what we've seen is that they may be somewhat further ahead in the development of that aircraft than our intelligence had earlier predicted."

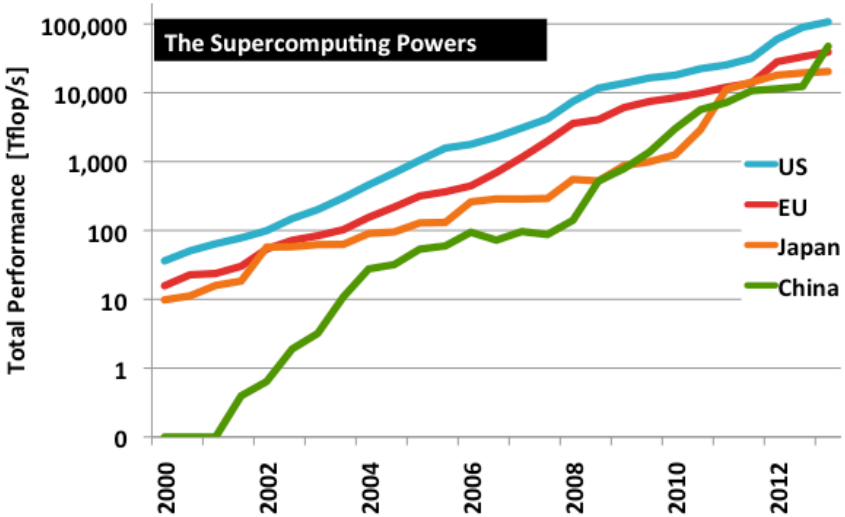
—Defense Secretary Robert M. Gates,
(*The New York Times*, Jan. 9, 2011)

Potential adversaries are not unaware of this

Exascale technologies are the foundation for future leadership in computing

The U.S. is not the only country capable of achieving exascale computing. The country that is first to exascale will have significant competitive intellectual, technological, and economic advantages.

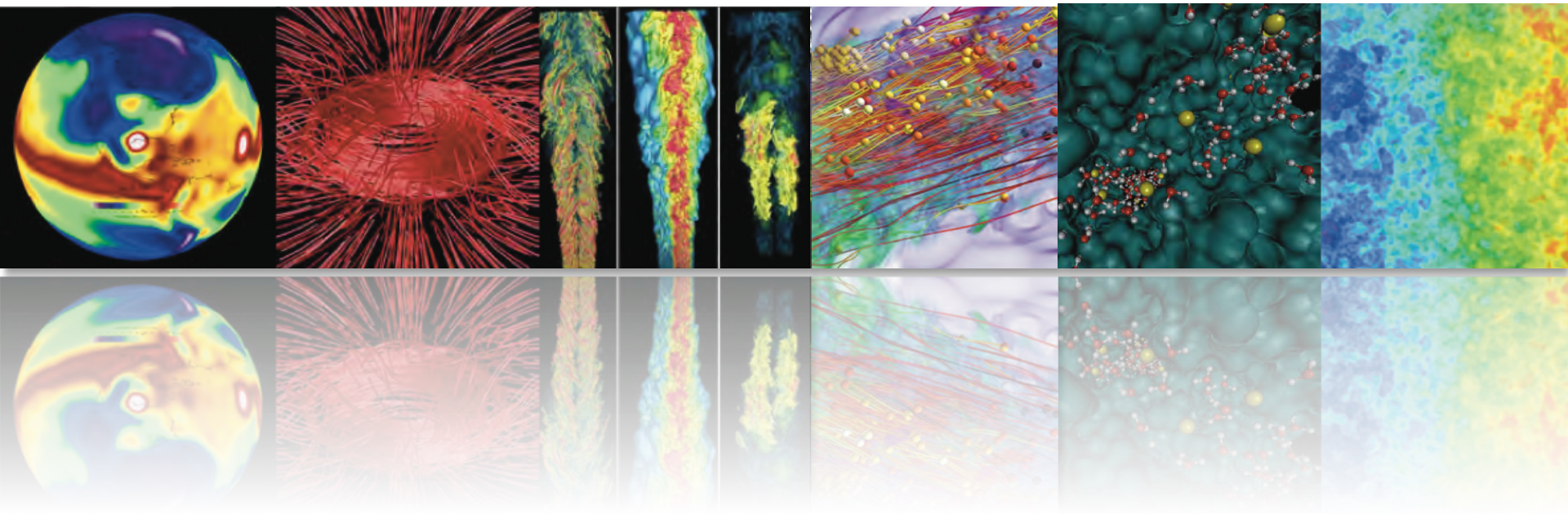
Achievement of the power efficiency and reliability goals needed for exascale will have enormous positive impacts on consumer electronics and business information technologies and facilities.



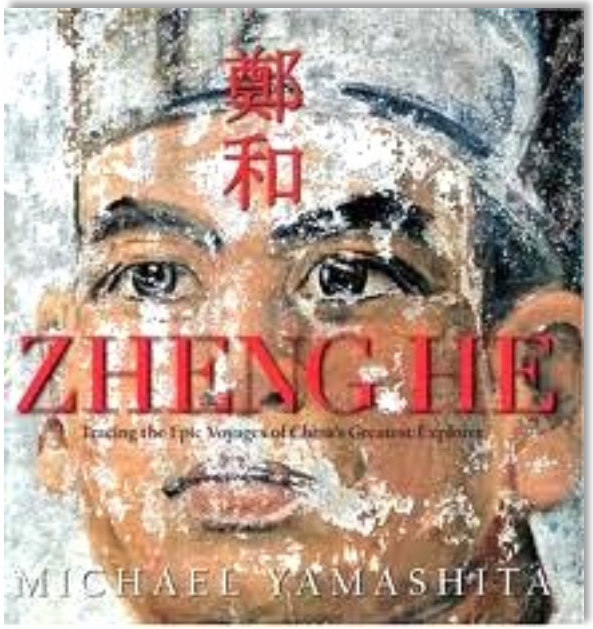
The gap in available supercomputing capacity between the United States and the rest of the world has narrowed, with China gaining the most ground.

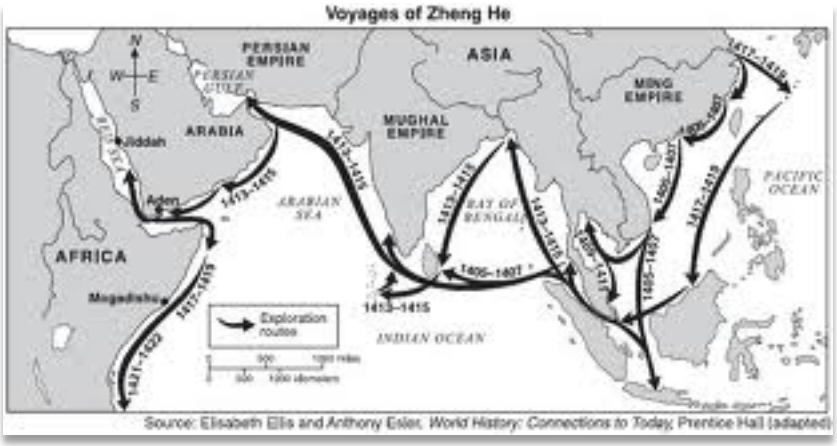
Data from TOP500.org; Figure from Science; Vol 335; 27 January 2012

Exascale is Discovery



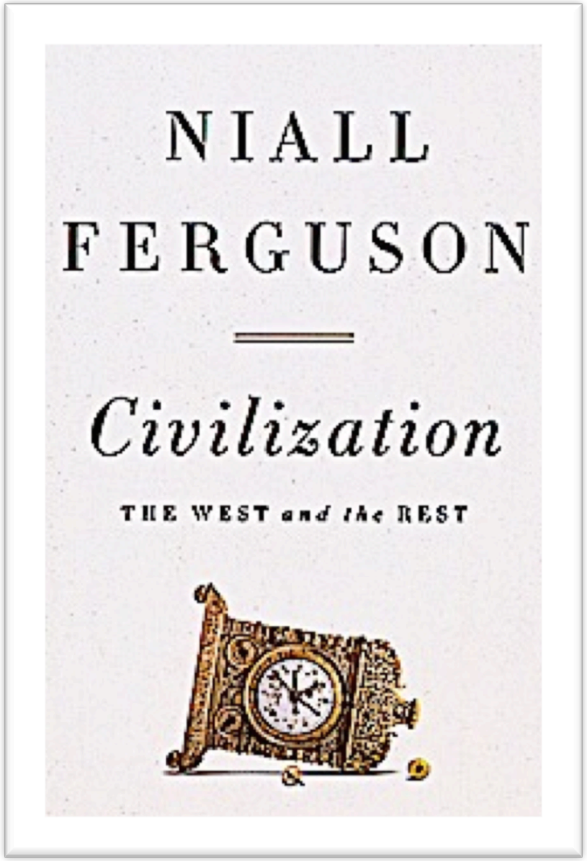
Admiral Zheng He's Voyages in 1416





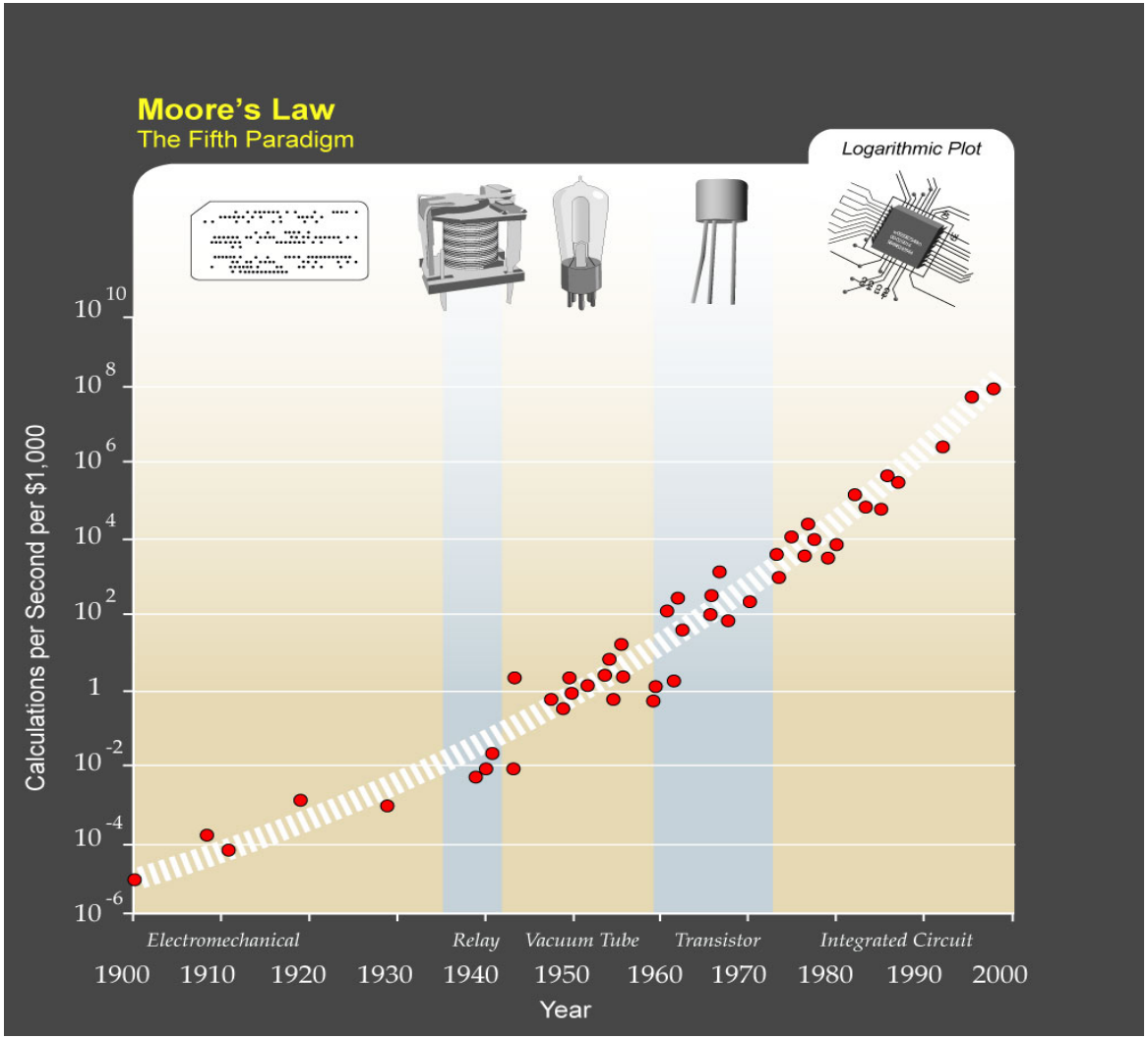
http://en.wikipedia.org/wiki/Zheng_He

The Rise of the West

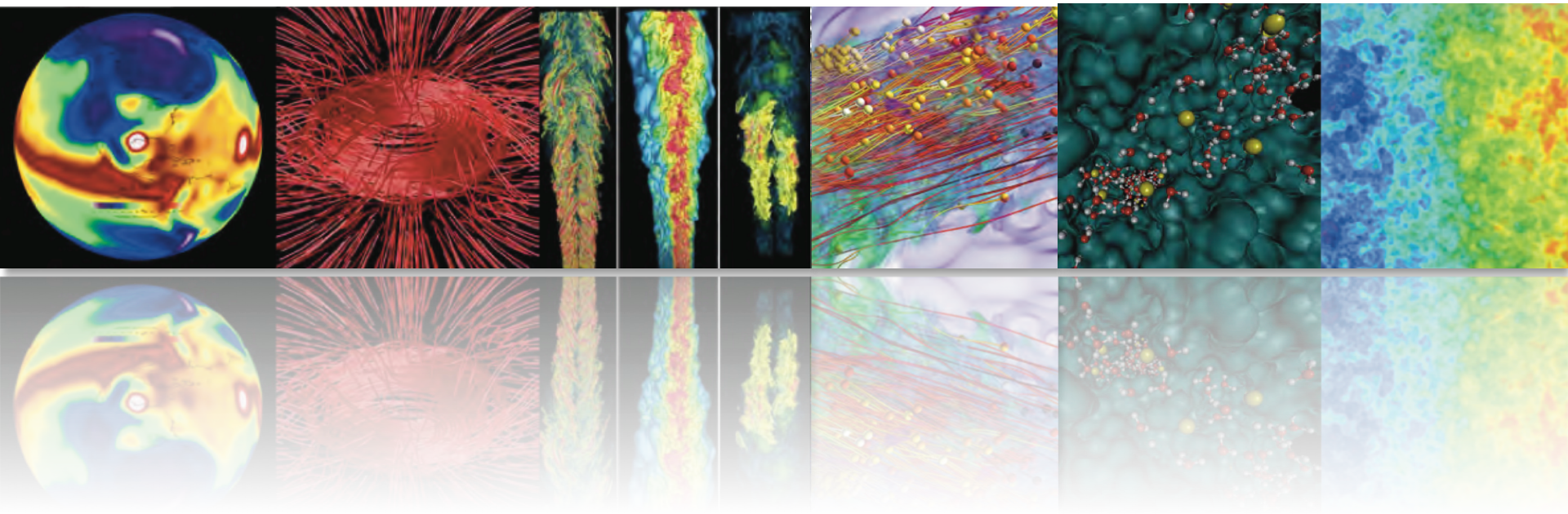


"Like the Apollo moon missions, Zheng He's voyages had been a formidable demonstration of wealth and technological sophistication. Landing … on the East African coast in 1416 was in many ways … comparable to landing an American astronaut on the moon in 1969. By abruptly canceling oceanic exploration, Yongle's successors ensured that the economic benefits remained negligible."

Exascale is the next step in a long voyage of discovery, we should not give up now and retreat



Discussion



Exascale rationale

- Exascale computing means performing computer simulations on a machine capable of 10 to the 18th power floating point operations per second – ~1,000× the current state-of-the-art.
- The U.S. is capable of reaching exascale simulation capability within ten years with cost-effective, efficient, computers. An accelerated program to achieve exascale is required to meet this goal, and doing so will enable the U.S. to lead in key areas of science and engineering.
- The U.S. is not the only country capable of achieving exascale computing. The country that is first to exascale will have significant competitive intellectual, technological, and economic advantages.
- Exascale systems and applications will improve predictive understanding in complex systems (e.g. climate, energy, materials, nuclear stockpile, emerging national security threats, biology, …)
- Tightly coupled exascale systems make possible fundamentally new approaches to uncertainty quantification and data analytics – uncertainty quantification is the essence of predictive simulation.
- American industry is increasingly dependent on HPC for design and testing in a race to get products to market. Exascale computing and partnerships with DOE laboratories could be THE key differentiator, a winning U.S. strategy for advanced manufacturing and job creation.
- The scale of computing developed for modeling and simulation will have a revolutionary impact on data analysis and data modeling. Much of the R&D required is common to big data and exascale computing.
- Exascale provides a ten year goal to drive modeling and simulation developments forward and to attract young scientific talent to DOE in a variety of scientific and technological fields of critical importance to the future of the U.S.
- Achievement of the power efficiency and reliability goals needed for exascale will have enormous positive impacts on consumer electronics and business information technologies and facilities.

Summary

- There is progress in Exascale with many projects now focused and on their way, e.g. FastForward, Xstack, and Co-Design Centers in the U.S.
- HPC has moved to low power processing, and the processor growth curves in energy-efficiency could get us in the range of exascale feasibility
- Memory and data movement are still more open challenges
- Programming model needs to address heterogeneous, massive parallel environment, as well as data locality
- Exascale applications will be a challenge just because of their sheer size and the memory limitations

BERKELEY LAB
Lawrence Berkeley National Laboratory