

Exascale Challenges for the Computational Science Community

**Horst Simon
Lawrence Berkeley National Laboratory
and UC Berkeley**

**Oklahoma Supercomputing Symposium 2010
October 6, 2010**



Key Message

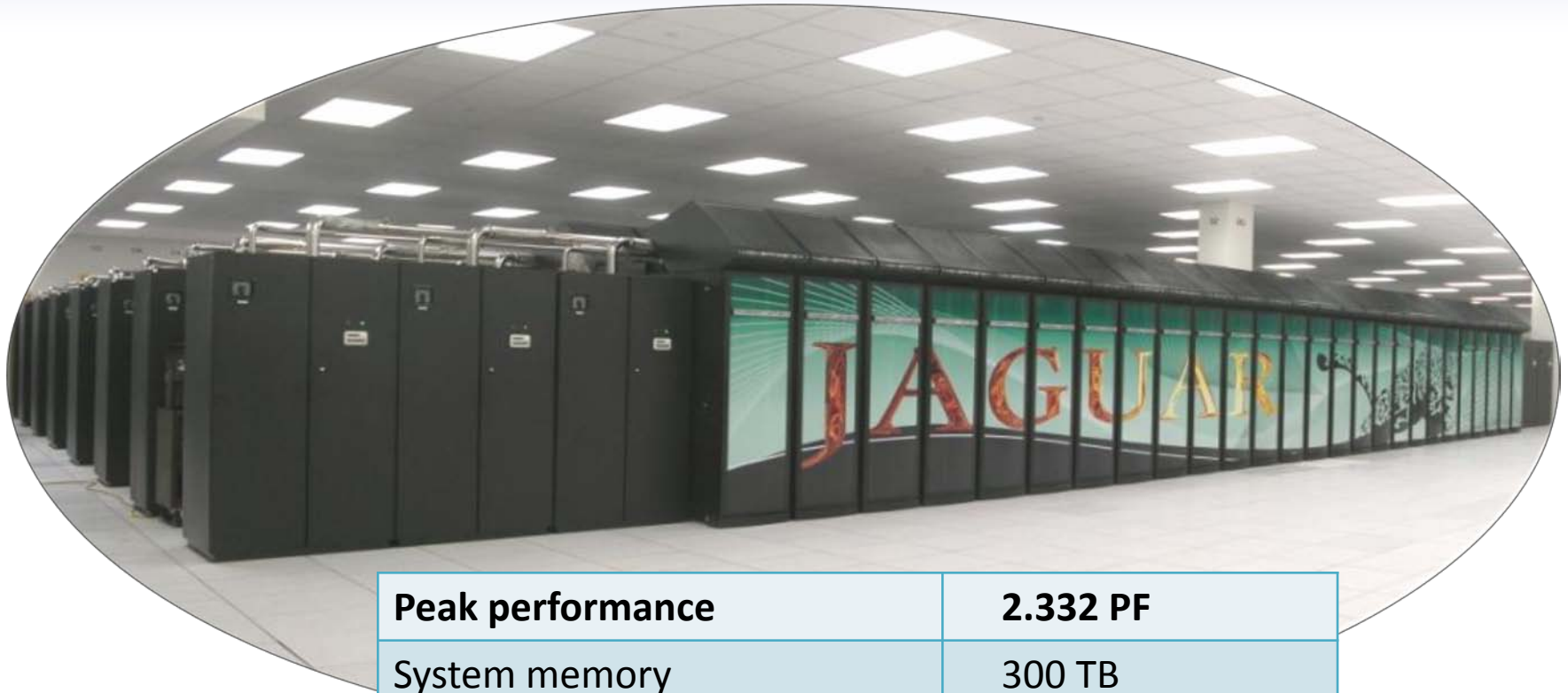
- **The transition from petascale to exascale will be characterized by significant and dramatic changes in hardware and architecture.**
- **This transition will be disruptive, but create unprecedented opportunities for computational science.**



Overview

- **From 1999 to 2009: evolution from Teraflops to Petaflops computing**
- **From 2010 to 2020: key technology changes towards Exaflops computing**
- **Impact on Computational Science**

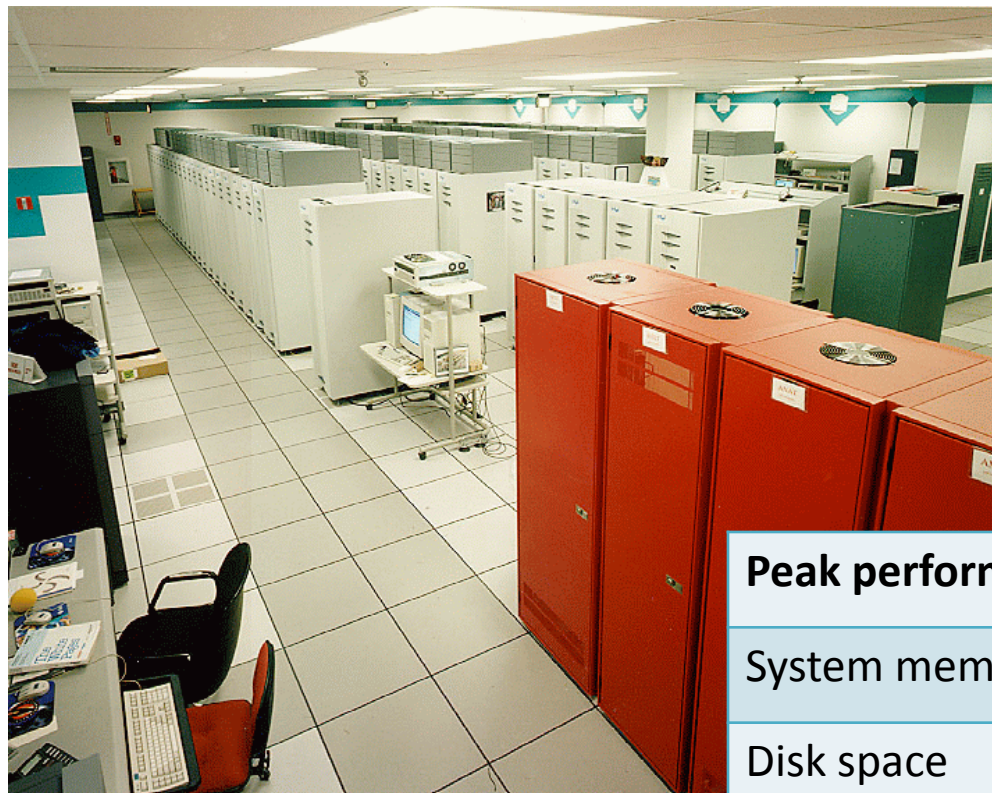
Jaguar: World's most powerful computer since 2009



Peak performance	2.332 PF
System memory	300 TB
Disk space	10 PB
Processors	224K
Power	6.95 MW



ASCI Red: World's Most Powerful Computer in 1999



#1 Nov. 1999

Peak performance	3.154 TF
System memory	1.212 TB
Disk space	12.5 TB
Processors	9298
Power	850 kW

Comparison

Jaguar (2009) vs. ASCI Red (1999)

- 739x performance (LINPACK)
- 267x memory
- 800x disk
- 24x processors/cores
- 8.2x power

Parallelism and faster processors made about equal contributions to performance increase

Significant increase in operations cost

Essentially the same architecture and software environment



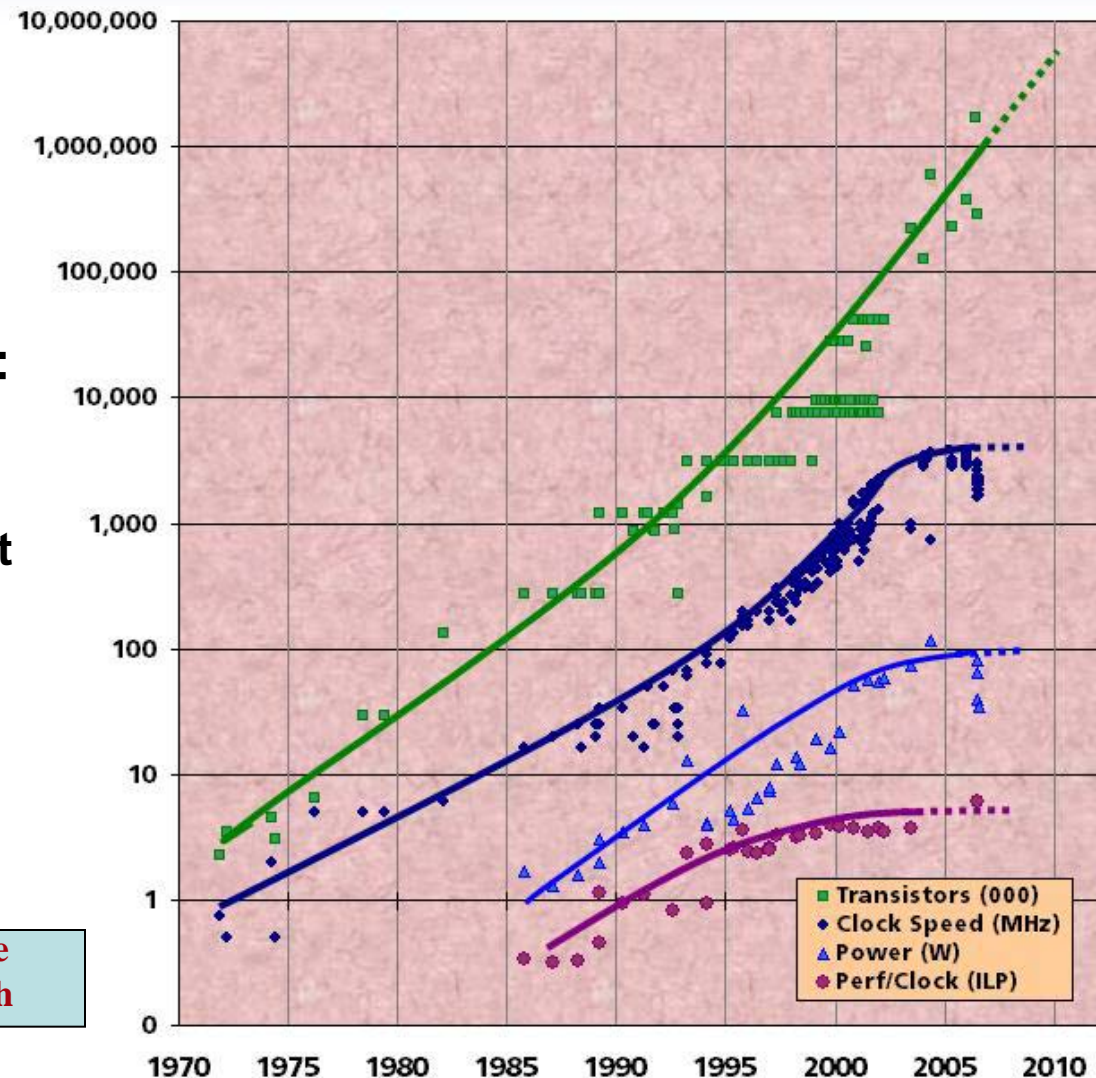
Overview

- From 1999 to 2009: evolution from Teraflops to Petaflops computing
- From 2010 to 2020: key technology changes towards Exaflops computing
- Impact on Computational Science

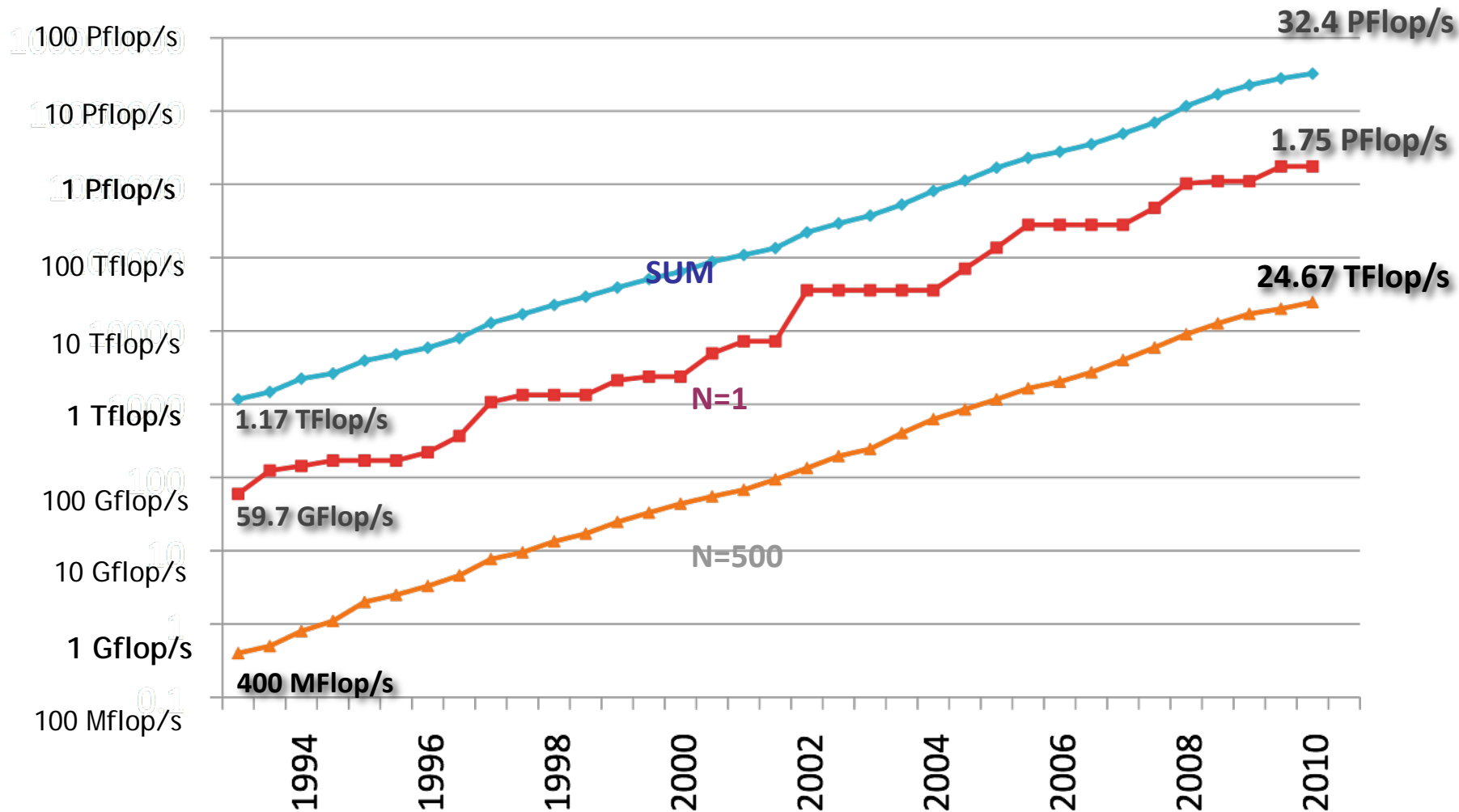
Traditional Sources of Performance Improvement are Flat-Lining (2004)

- New Constraints
 - 15 years of *exponential* clock rate growth has ended
- Moore's Law reinterpreted:
 - How do we use all of those transistors to keep performance increasing at historical rates?
 - Industry Response: #cores per chip doubles every 18 months *instead* of clock frequency!

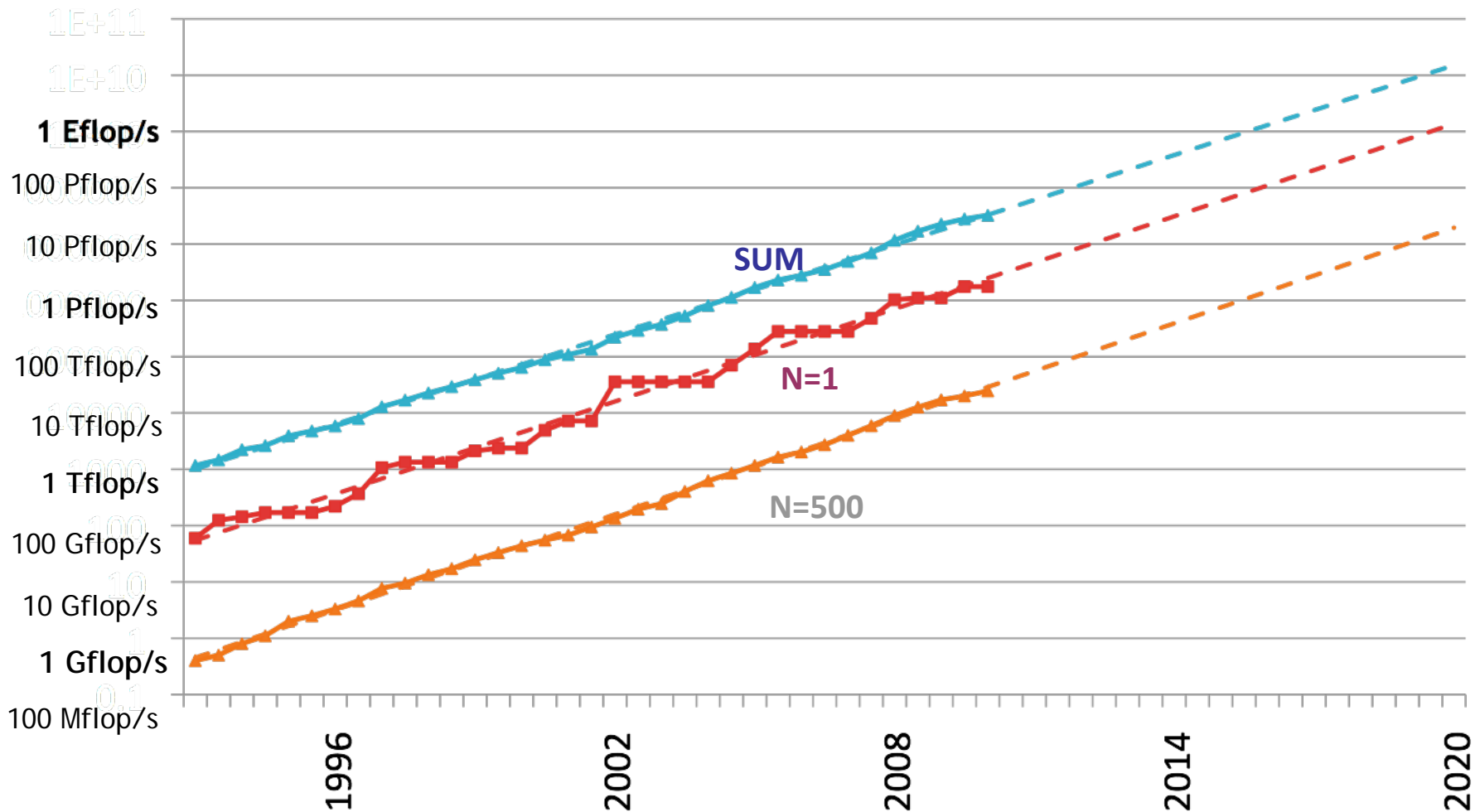
Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith



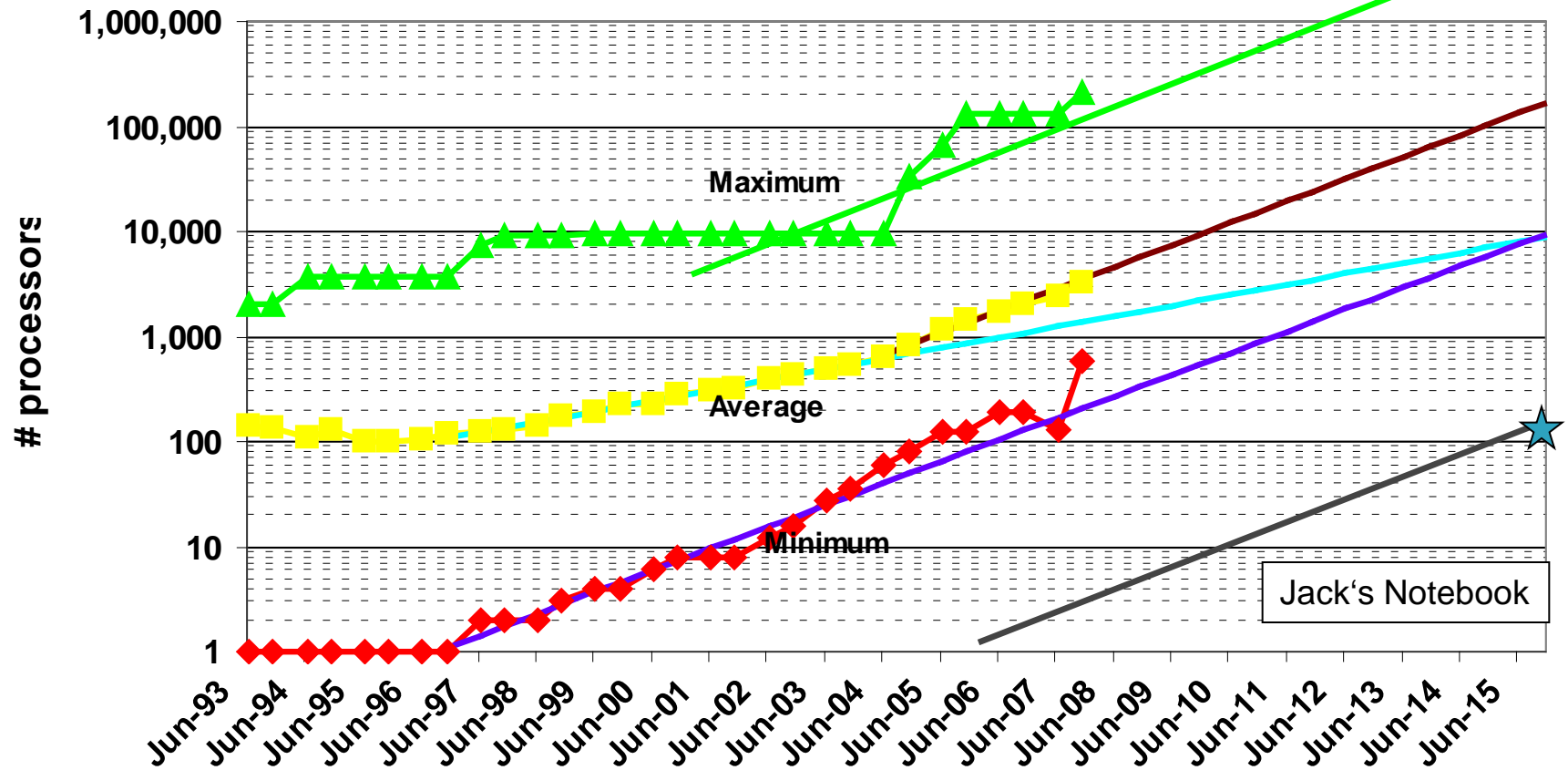
Performance Development



Projected Performance Development

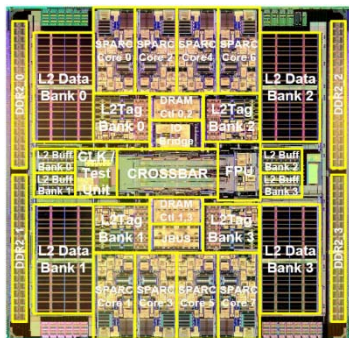


Concurrency Levels



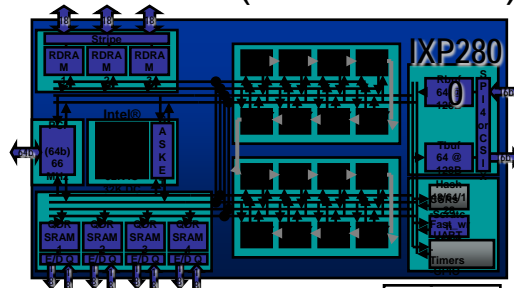
Multicore comes in a wide variety

- Multiple parallel general-purpose processors (GPPs)
- Multiple application-specific processors (ASPs)

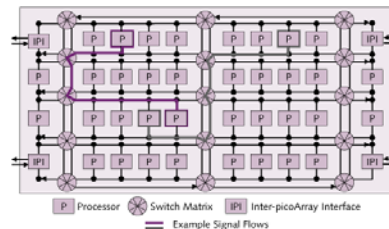


Sun Niagara
8 GPP cores (32 threads)

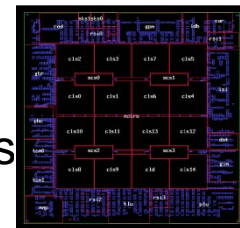
Intel Network Processor
1 GPP Core
16 ASPs (128 threads)



IBM Cell
1 GPP (2 threads)
8 ASPs

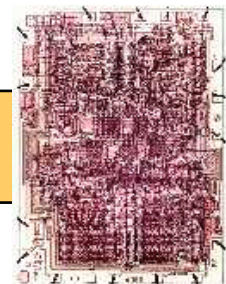


Picochip DSP
1 GPP core
248 ASPs



Cisco CRS-1
188 Tensilica GPPs

Intel 4004 (1971):
4-bit processor,
2312 transistors,
~100 KIPS,
10 micron PMOS,
11 mm² chip

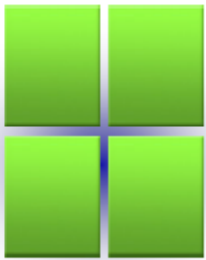


1000s of
processor
cores per
die

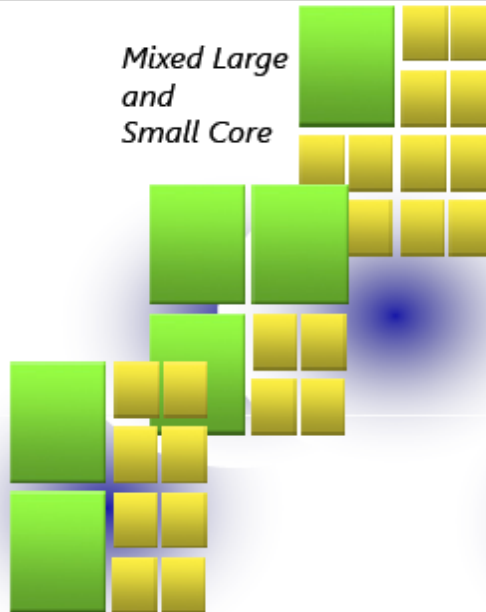
***“The Processor is
the new Transistor”
[Rowen]***

What's Next?

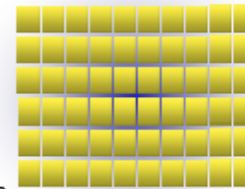
All Large Core



Mixed Large and Small Core



Many Small Cores

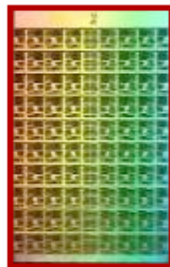


All Small Core

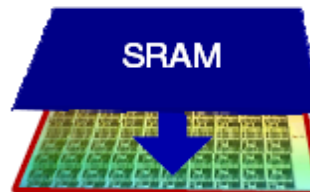


Different Classes of Chips
Home
Games / Graphics
Business
Scientific

Many Floating-Point Cores



+ 3D Stacked Memory



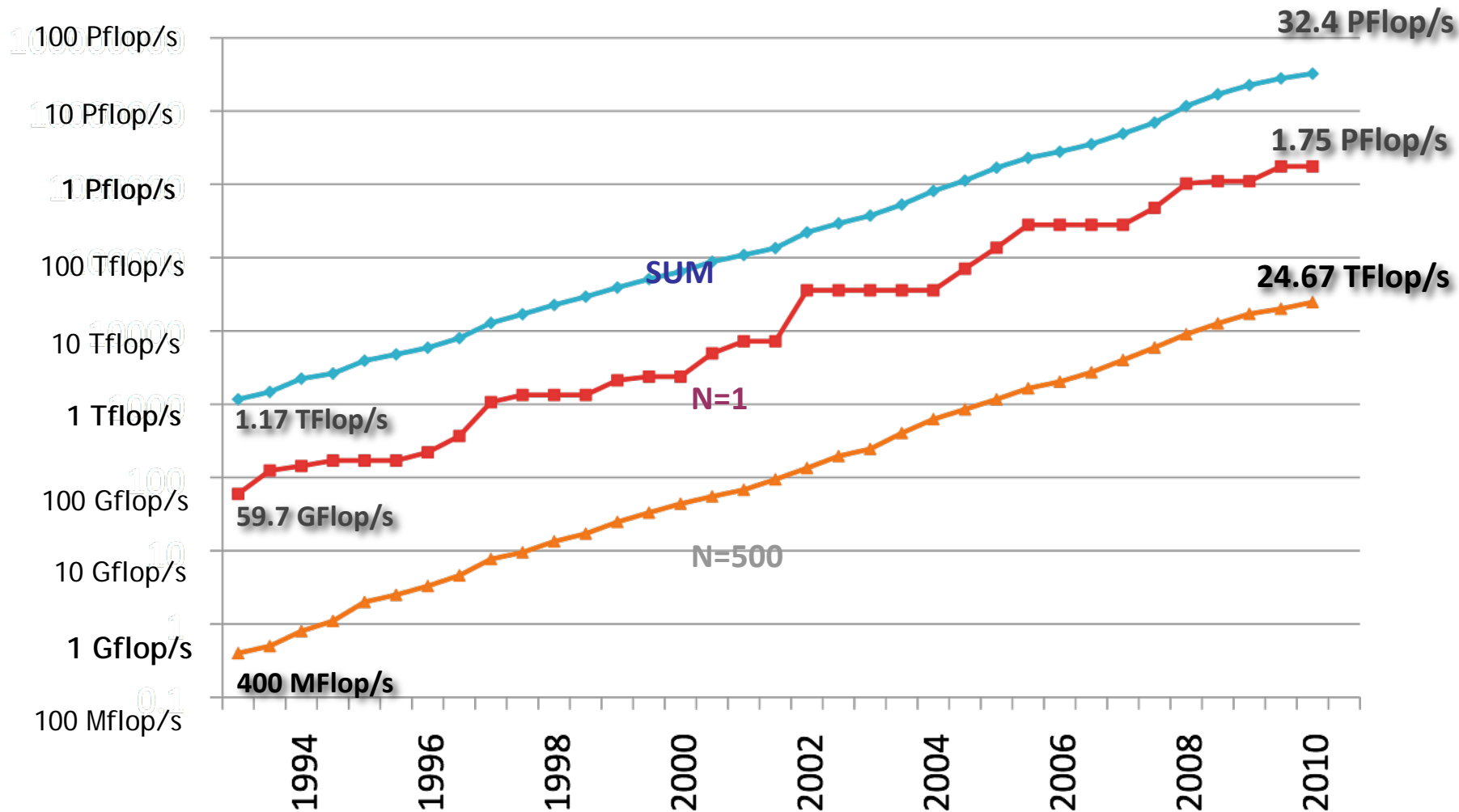
The question is not whether this will happen but whether we are ready

Source: Jack Dongarra, ISC 2008

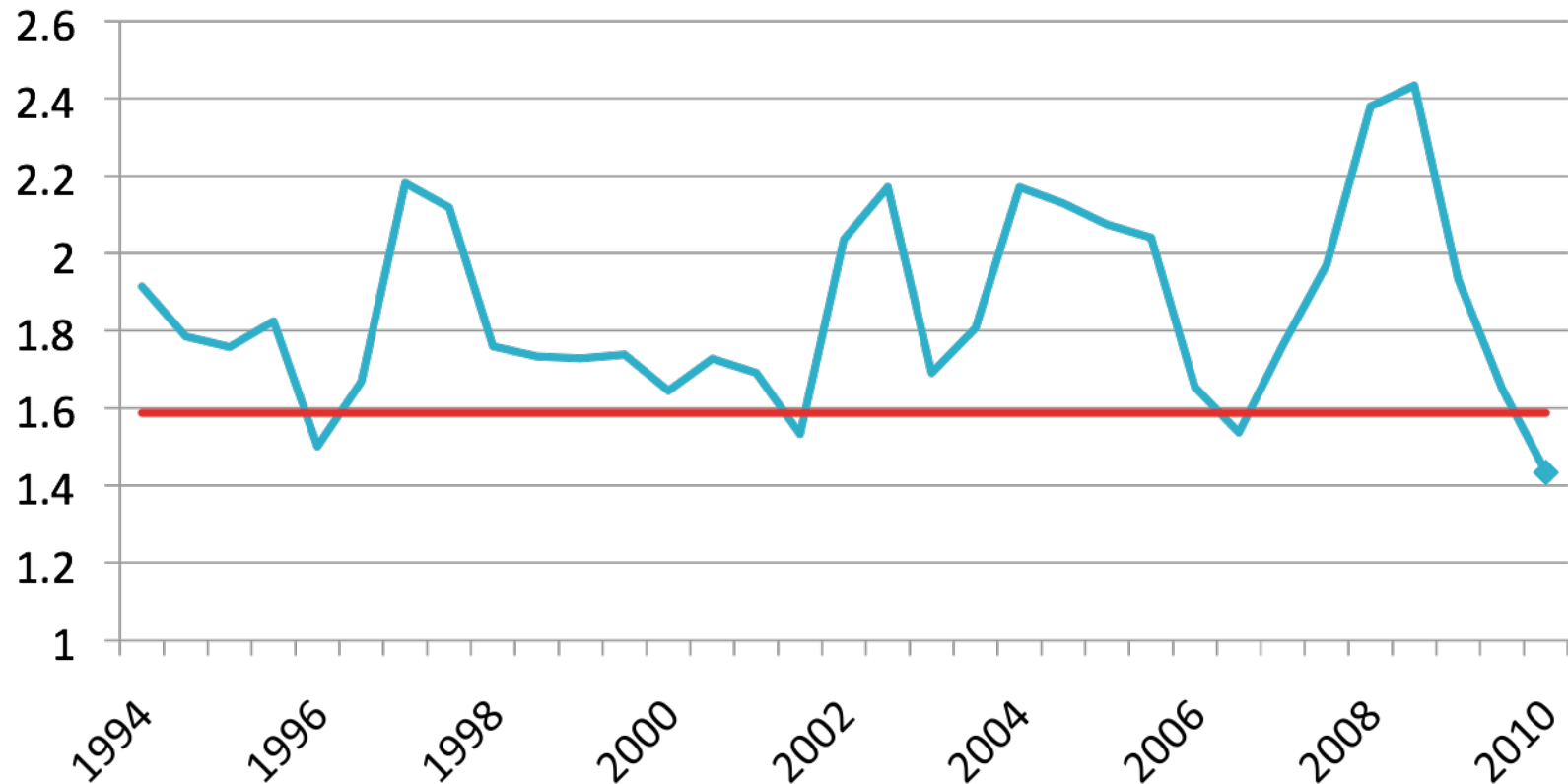
Moore's Law reinterpreted

- **Number of cores per chip will double every two years**
- **Clock speed will not increase (possibly decrease)**
- **Need to deal with systems with millions of concurrent threads**
- **Need to deal with inter-chip parallelism as well as intra-chip parallelism**

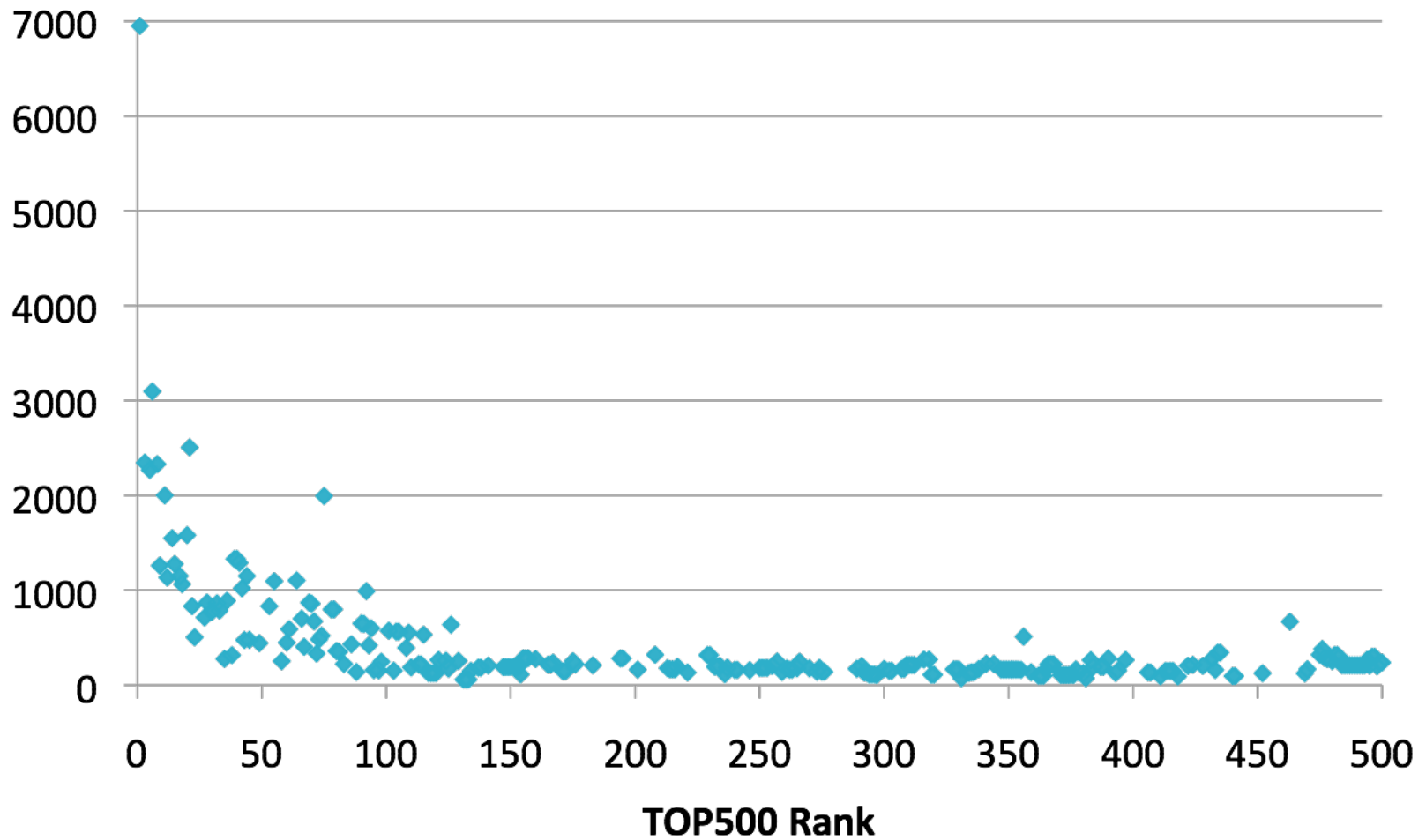
Performance Development



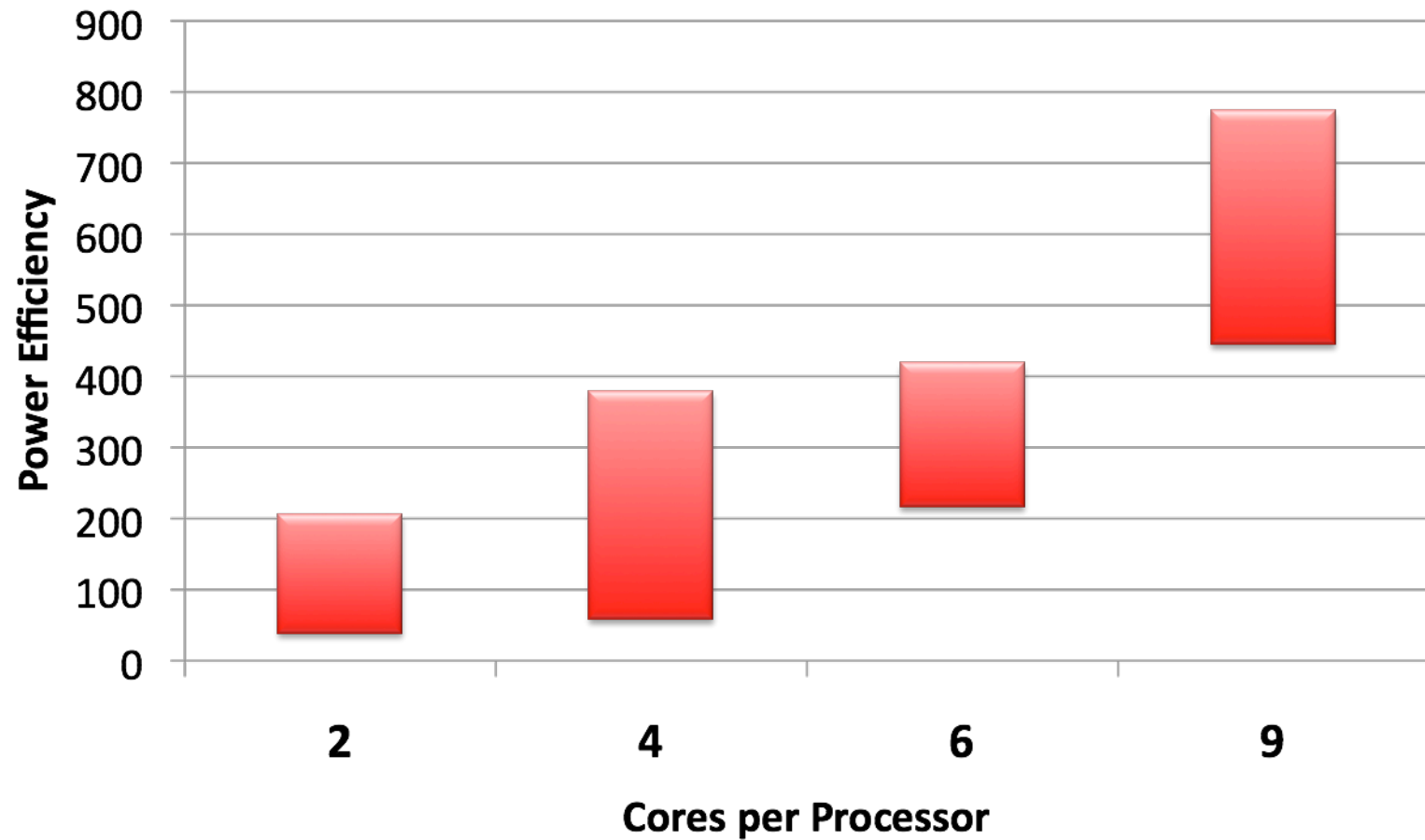
Annual Performance Increase of the TOP500



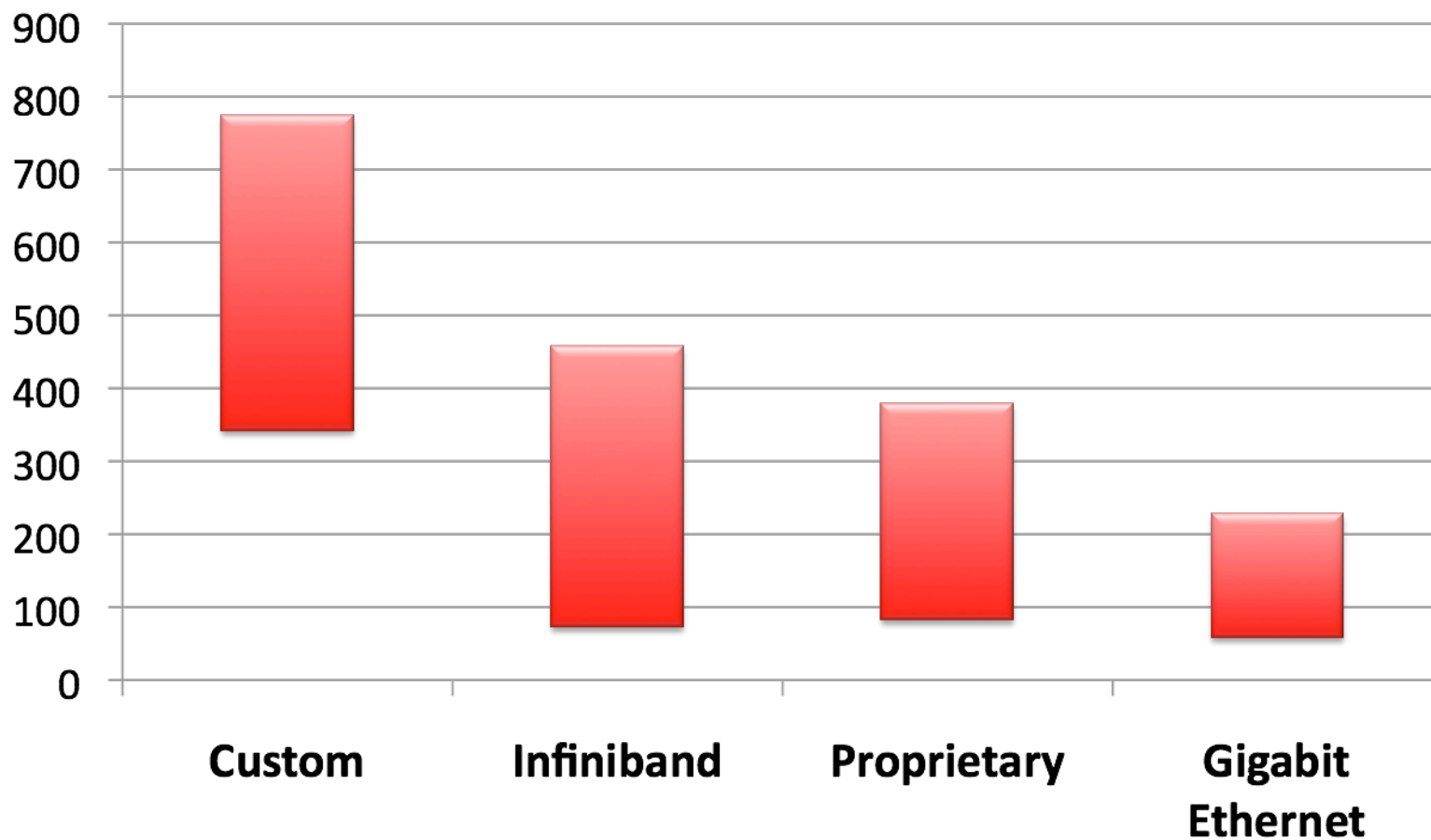
Total Power Levels (kW) for TOP500 systems



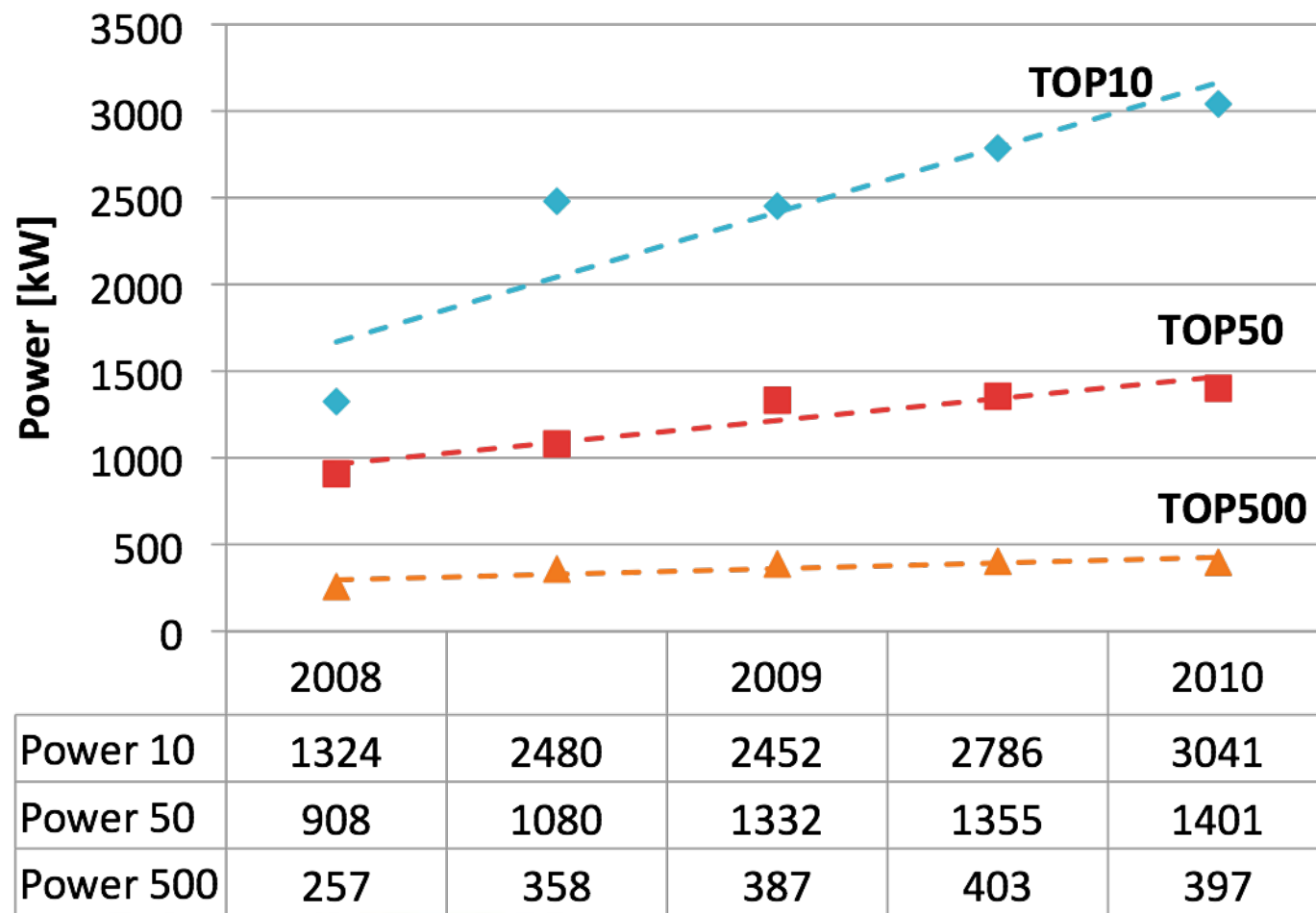
Power Efficiency (Mflops/Watt) for different Processor Generations



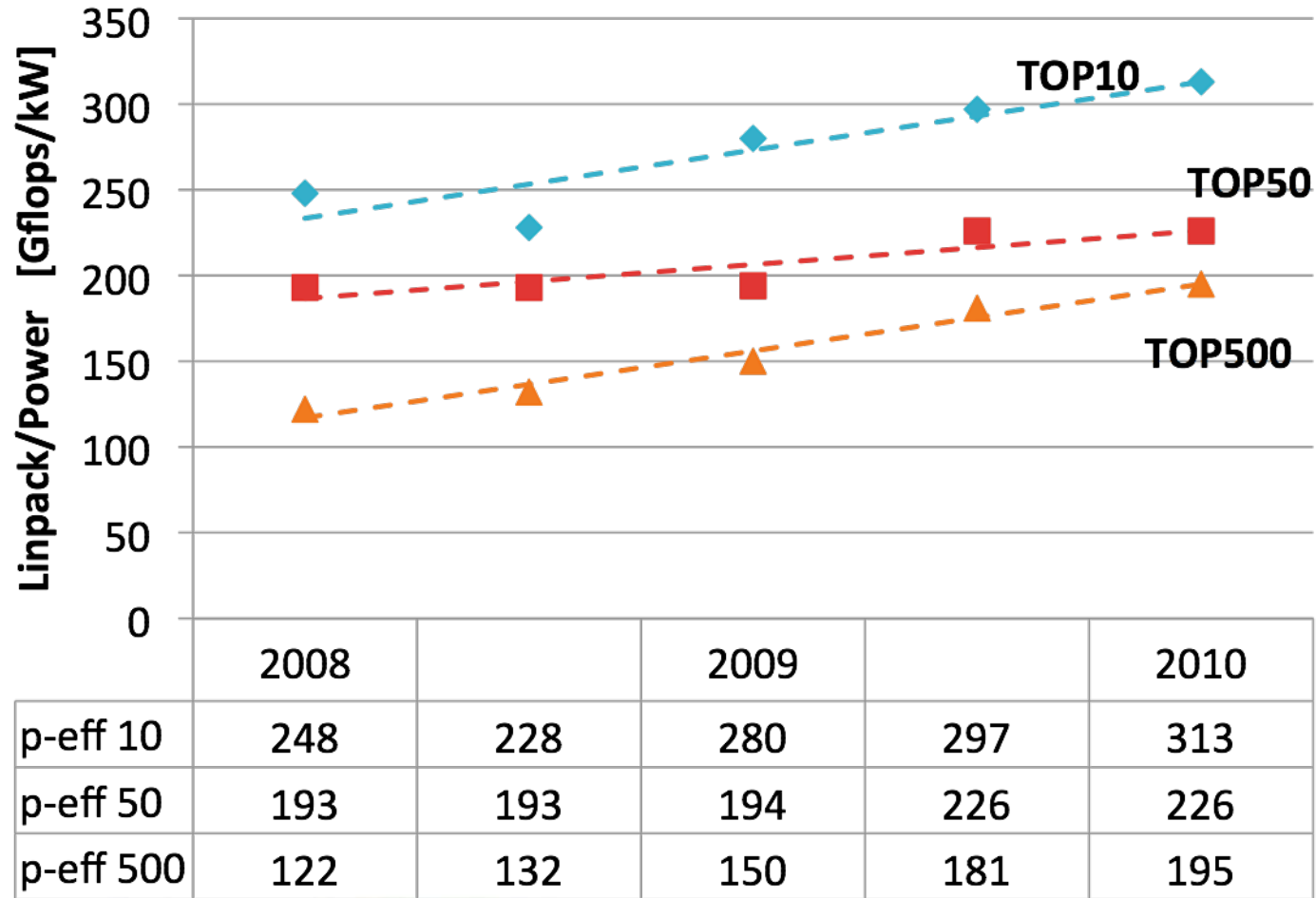
Power Efficiency (Mflops/Watt) related to Interconnects



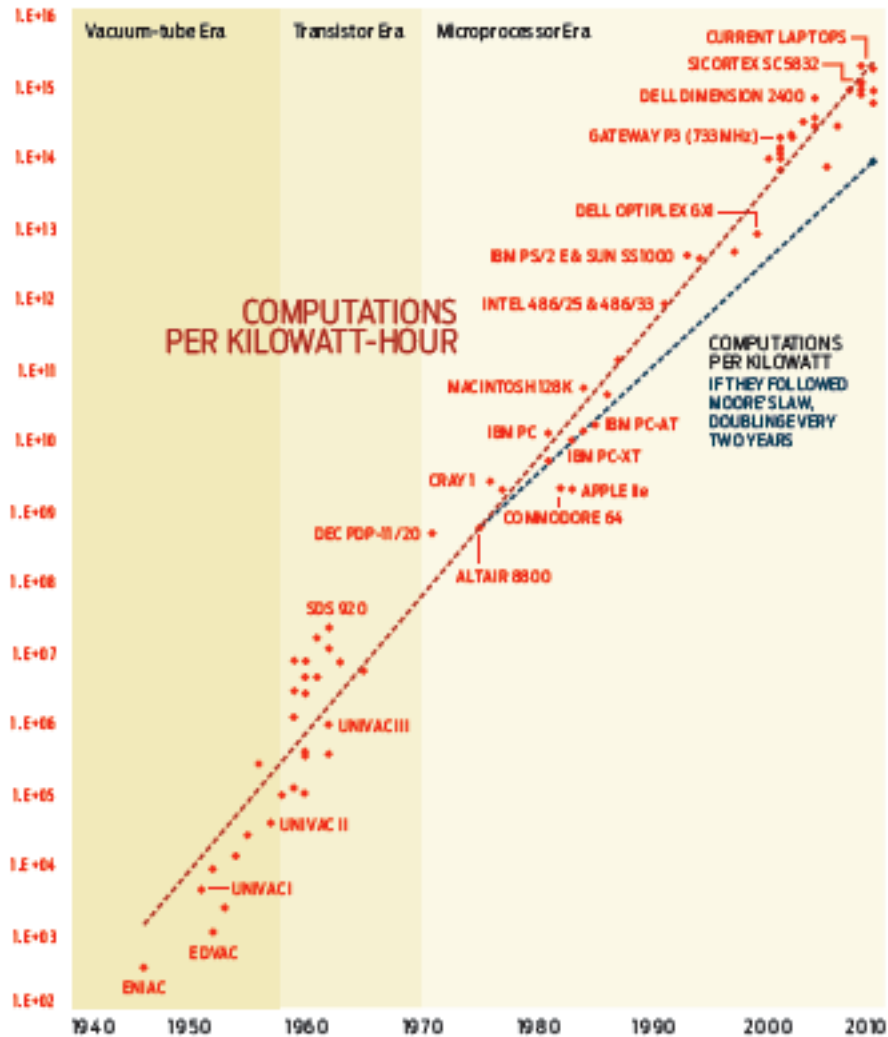
Power Consumption



Power Efficiency



Koomey's Law



- Computations per kWh have improved by a factor about 1.5 per year
- “Assessing Trends in Electrical Efficiency over Time”, see IEEE Spectrum, March 2010

The complete study is available from the Intel Web site at <http://download.intel.com/pressroom/pdf/computertrendrelease.pdf>



BERKELEY LAB

LAWRENCE BERKELEY NATIONAL LABORATORY

Managed by the University of California for the U.S. Department of Energy

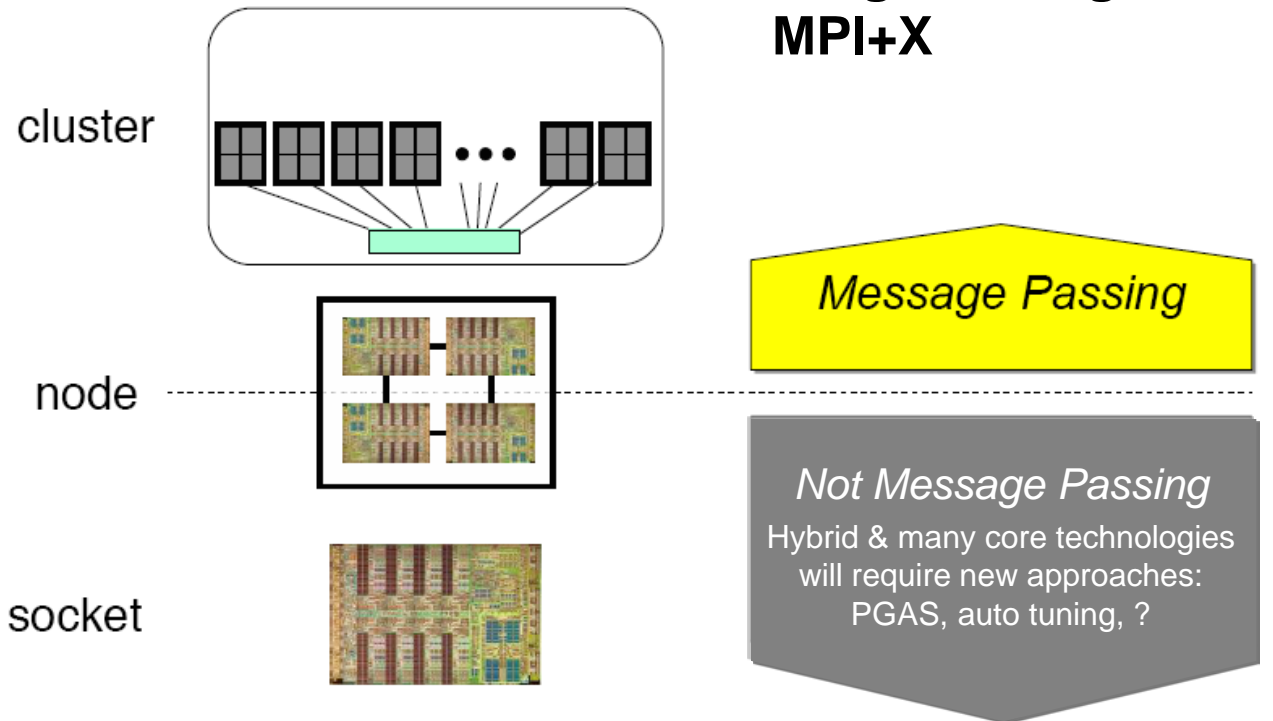
Trend Analysis

- **Processors and Systems have become more energy efficient over time**
 - **Koomey's Law shows factor of 1.5 improvement in computations/kWh**
- **Supercomputers have become more powerful over time**
 - **TOP500 data show factor of 1.86 increase of computations/sec per system**
- **Consequently power/system increases by about 1.24 per year**
- **Based on these projections: 495 Pflop/s Linpack-Rmax system with 60 MW in 2020**

Roadrunner - A Likely Future Scenario

System: cluster + many core node

**Programming model:
MPI+X**



after Don Grice, IBM, Roadrunner Presentation,
ISC 2008

Why MPI will persist

- Obviously MPI will not disappear in five years
- By 2014 there will be 20 years of legacy software in MPI
- New systems are not sufficiently different to lead to new programming model

What will be the “X” in MPI+X

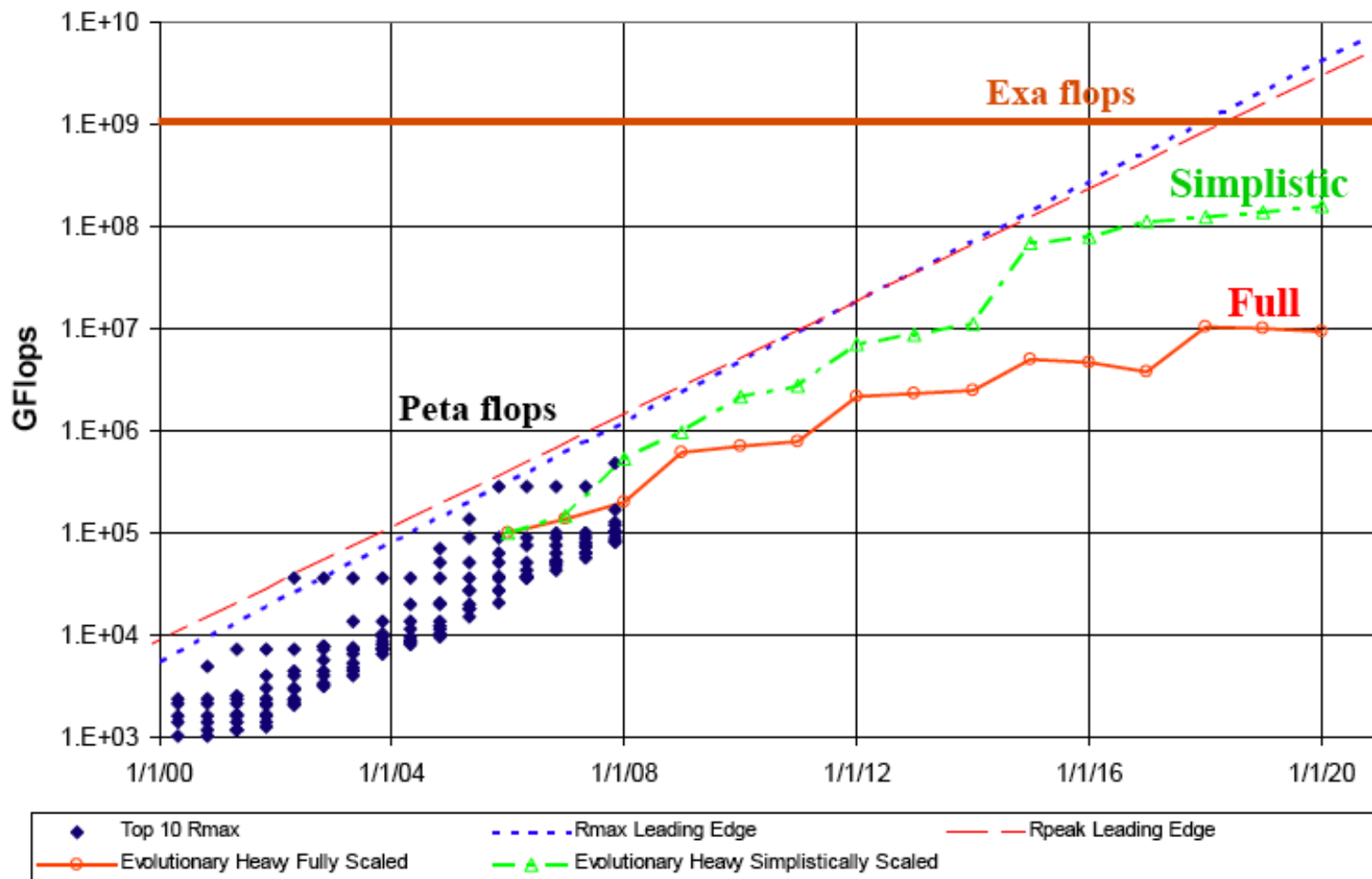
- **Likely candidates are**
 - **PGAS languages**
 - **OpenMP**
 - **Autotuning**
 - **CUDA, OpenCL**
 - **A wildcard from commercial space**

What's Wrong with MPI Everywhere?

What's Wrong with MPI Everywhere?

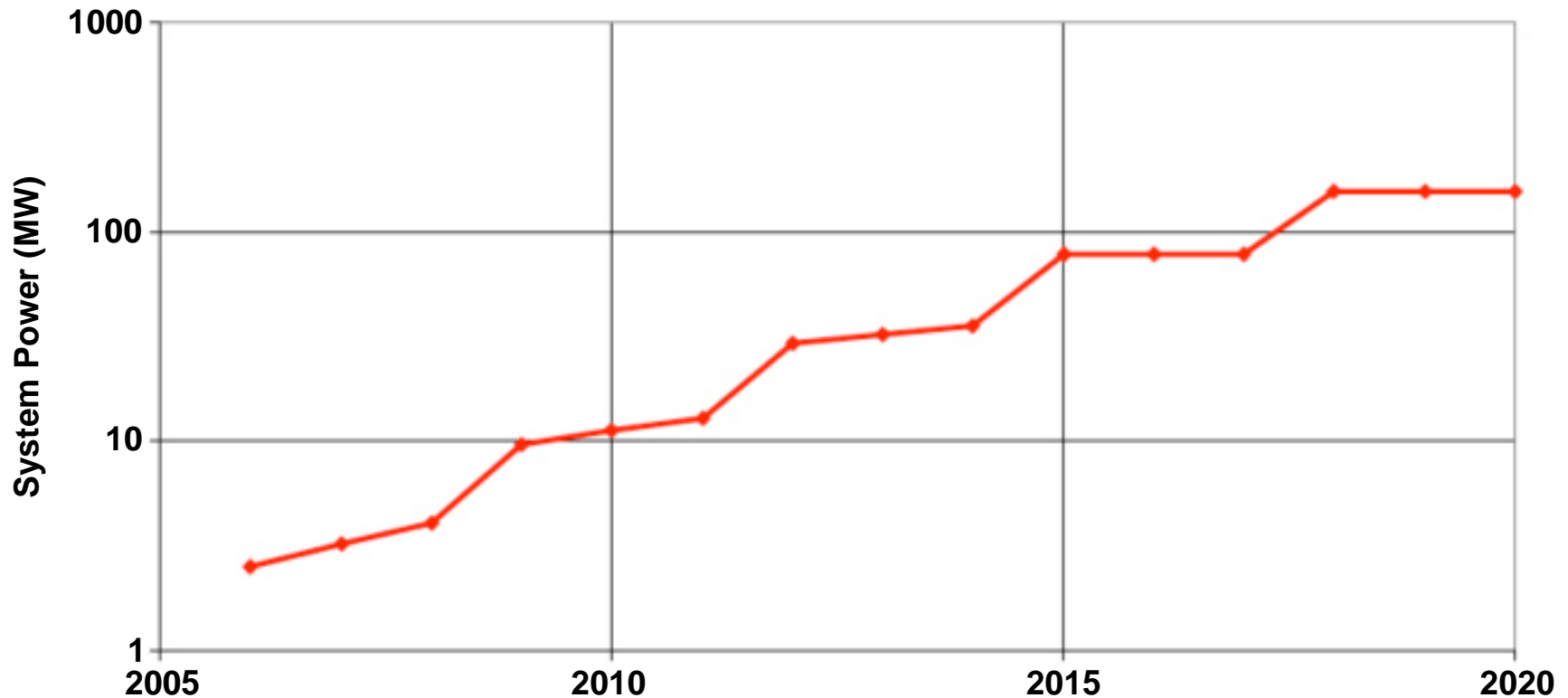
- One MPI process per core is wasteful of intra-chip latency and bandwidth
- **Weak scaling:** success model for the “cluster era”
 - not enough memory per core
- **Heterogeneity:** MPI per CUDA thread-block?

We won't reach Exaflops with the current approach



From Peter Kogge, DARPA Exascale Study

... and the power costs will still be staggering



From Peter Kogge,
DARPA Exascale Study

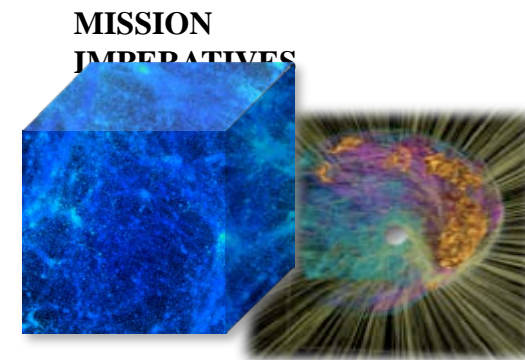
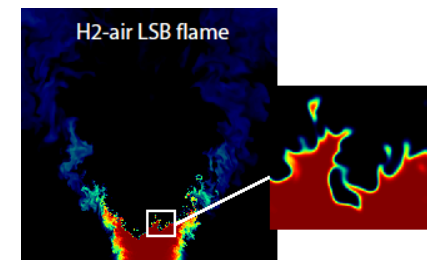
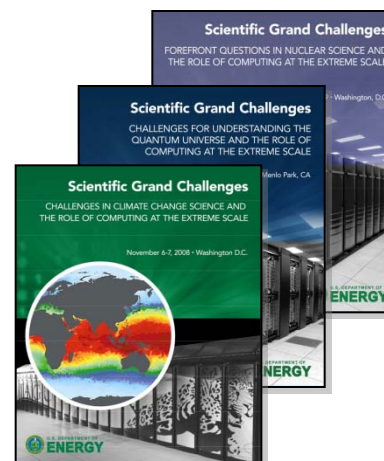
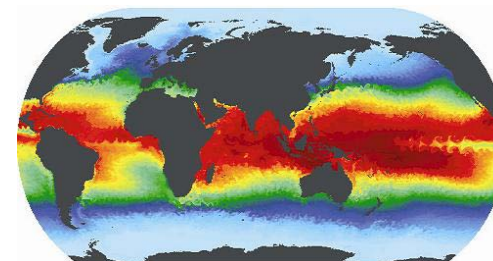
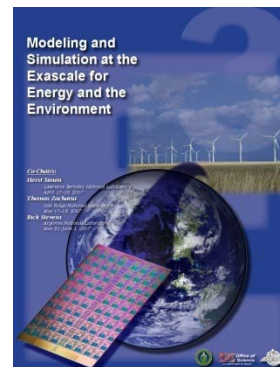
A decadal DOE plan for providing exascale applications and technologies for DOE mission needs

Rick Stevens and Andy White, co-chairs

Pete Beckman, Ray Bair-ANL; Jim Hack, Jeff Nichols, Al Geist-ORNL; Horst Simon, Kathy Yelick, John Shalf-LBNL; Steve Ashby, Moe Khaleel-PNNL; Michel McCoy, Mark Seager, Brent Gorda-LLNL; John Morrison, Cheryl Wampler-LANL; James Peery, Sudip Dosanjh, Jim Ang-SNL; Jim Davenport, Tom Schlagel, BNL; Fred Johnson, Paul Messina, ex officio

Process for identifying exascale applications and technology for DOE missions ensures broad community input

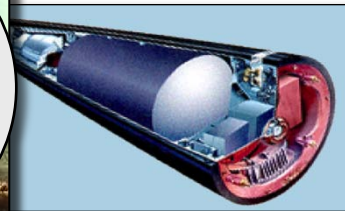
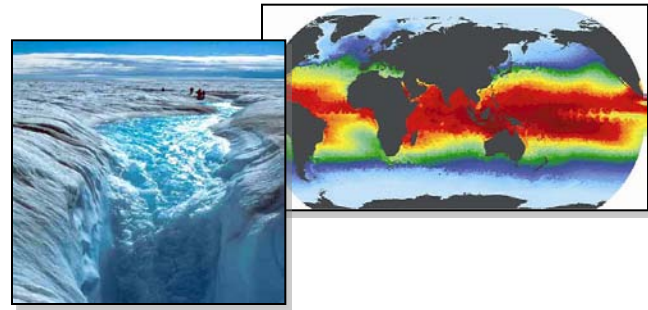
- Town Hall Meetings April-June 2007
- Scientific Grand Challenges Workshops Nov, 2008 – Oct, 2009
 - Climate Science (11/08),
 - High Energy Physics (12/08),
 - Nuclear Physics (1/09),
 - Fusion Energy (3/09),
 - Nuclear Energy (5/09),
 - Biology (8/09),
 - Material Science and Chemistry (8/09),
 - National Security (10/09)
 - Cross-cutting technologies (2/10)
- Exascale Steering Committee
 - “Denver” vendor NDA visits 8/2009
 - SC09 vendor feedback meetings
 - Extreme Architecture and Technology Workshop 12/2009
- International Exascale Software Project
 - Santa Fe, NM 4/2009; Paris, France 6/2009; Tsukuba, Japan 10/2009



FUNDAMENTAL SCIENCE

DOE mission imperatives require simulation and analysis for policy and decision making

- **Climate Change:** Understanding, mitigating and adapting to the effects of global warming
 - Sea level rise
 - Severe weather
 - Regional climate change
 - Geologic carbon sequestration
- **Energy:** Reducing U.S. reliance on foreign energy sources and reducing the carbon footprint of energy production
 - Reducing time and cost of reactor design and deployment
 - Improving the efficiency of combustion energy systems
- **National Nuclear Security:** Maintaining a safe, secure and reliable nuclear stockpile
 - Stockpile certification
 - Predictive scientific challenges
 - Real-time evaluation of urban nuclear detonation

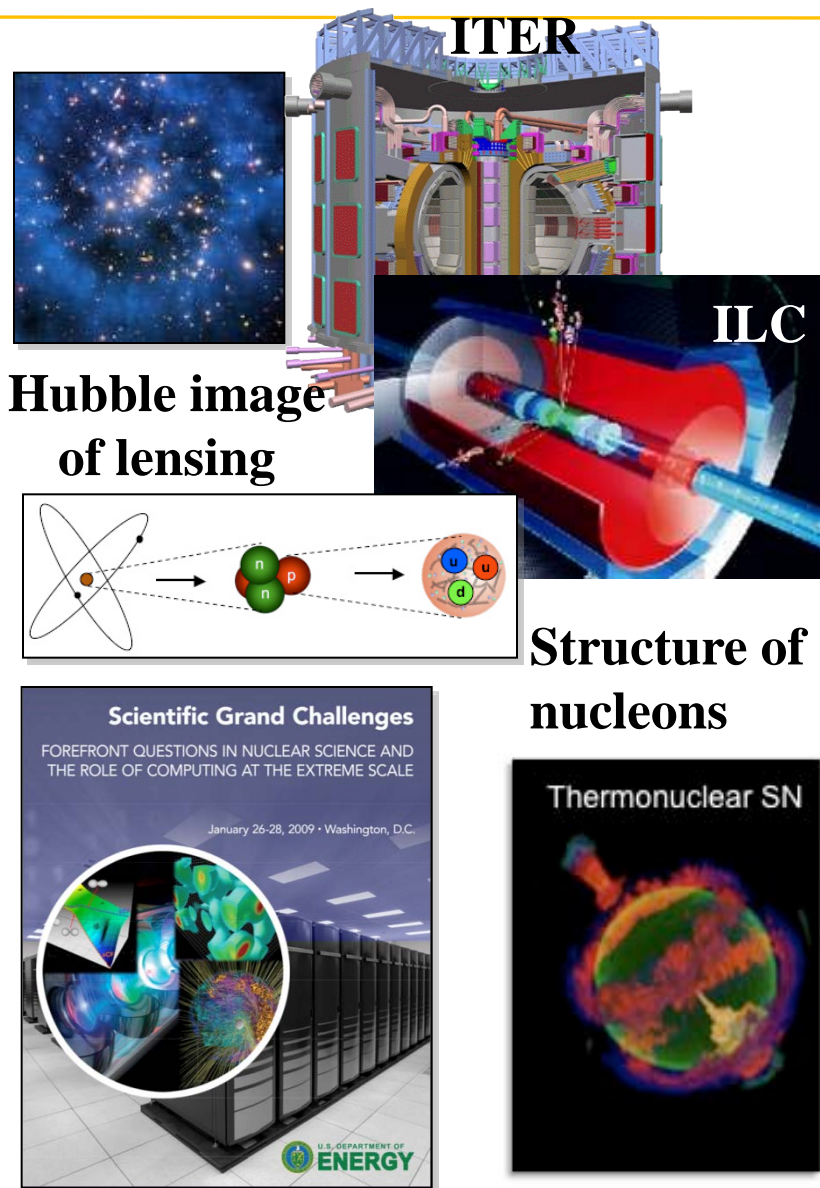


Accomplishing these missions requires exascale resources.

Exascale simulation will enable fundamental advances in basic science.

- **High Energy & Nuclear Physics**
 - Dark-energy and dark matter
 - Fundamentals of fission fusion reactions
- **Facility and experimental design**
 - Effective design of accelerators
 - Probes of dark energy and dark matter
 - ITER shot planning and device control
- **Materials / Chemistry**
 - Predictive multi-scale materials modeling: observation to control
 - Effective, commercial technologies in renewable energy, catalysts, batteries and combustion
- **Life Sciences**
 - Better biofuels
 - Sequence to structure to function

These breakthrough scientific discoveries and facilities require exascale applications and resources.





Potential System Architecture Targets

System attributes	2010	“2015”		“2018”	
System peak	2 Peta	200 Petaflop/sec		1 Exaflop/sec	
Power	6 MW	15 MW		20 MW	
System memory	0.3 PB	5 PB		32-64 PB	
Node performance	125 GF	0.5 TF	7 TF	1 TF	10 TF
Node memory BW	25 GB/s	0.1 TB/sec	1 TB/sec	0.4 TB/sec	4 TB/sec
Node concurrency	12	O(100)	O(1,000)	O(1,000)	O(10,000)
System size (nodes)	18,700	50,000	5,000	1,000,000	100,000
Total Node Interconnect BW	1.5 GB/s	20 GB/sec		200 GB/sec	
MTTI	days	O(1day)		O(1 day)	

Comparison

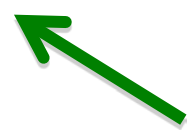
“2018” vs. Jaguar (2009)

- 500x performance (peak)
- 100x memory
- 5000x concurrency
- 3x power

All performance increase
is based on more
parallelism



Keep operating cost
about the “same”



Significantly different architecture and software environment

DOE Exascale Technology Roadmap

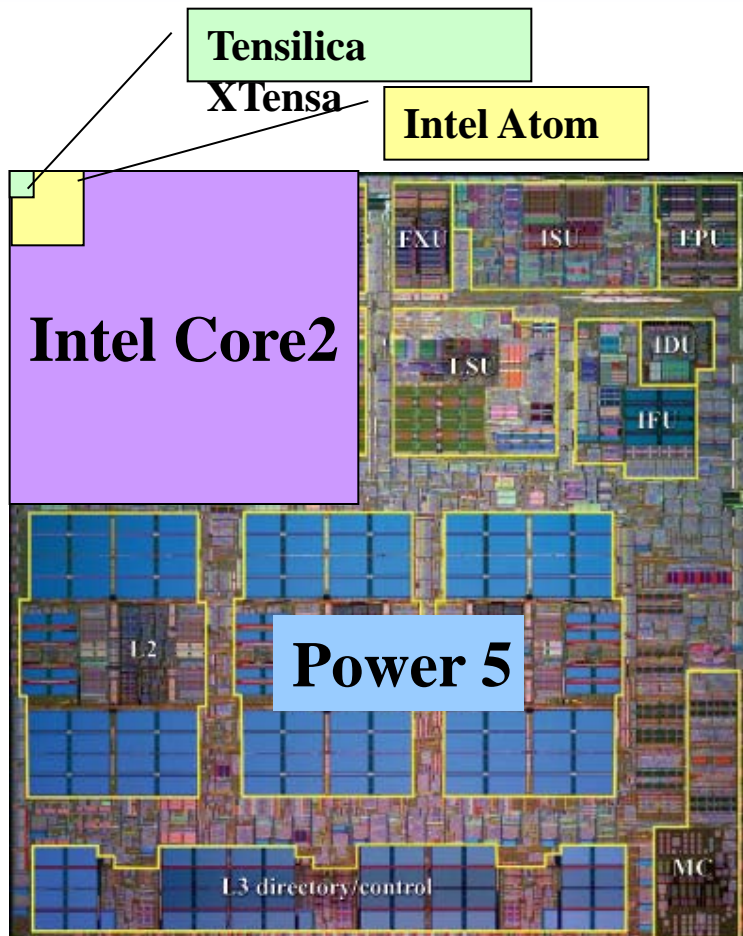
**Key Observations from DOE Exascale
Architecture and Technology Workshop,
San Diego, Dec. 2009,**

<http://extremecomputing.labworks.org/hardware/index.stm>

Where do we get 1000x performance improvement for 10x power?

1. **Processors**
2. **On-chip data movement**
3. **System-wide data movement**
4. **Memory Technology**
5. **Resilience Mechanisms**

Low-Power Design Principles



- Power5 (server)
 - 120W@1900MHz
 - **Baseline**
- Intel Core2 sc (laptop) :
 - 15W@1000MHz
 - *4x more FLOPs/watt than baseline*
- Intel Atom (handhelds)
 - 0.625W@800MHz
 - **80x more**
- Tensilica XTensa DP (Moto Razor) :
 - 0.09W@600MHz
 - **400x more** (*80x-120x sustained*)

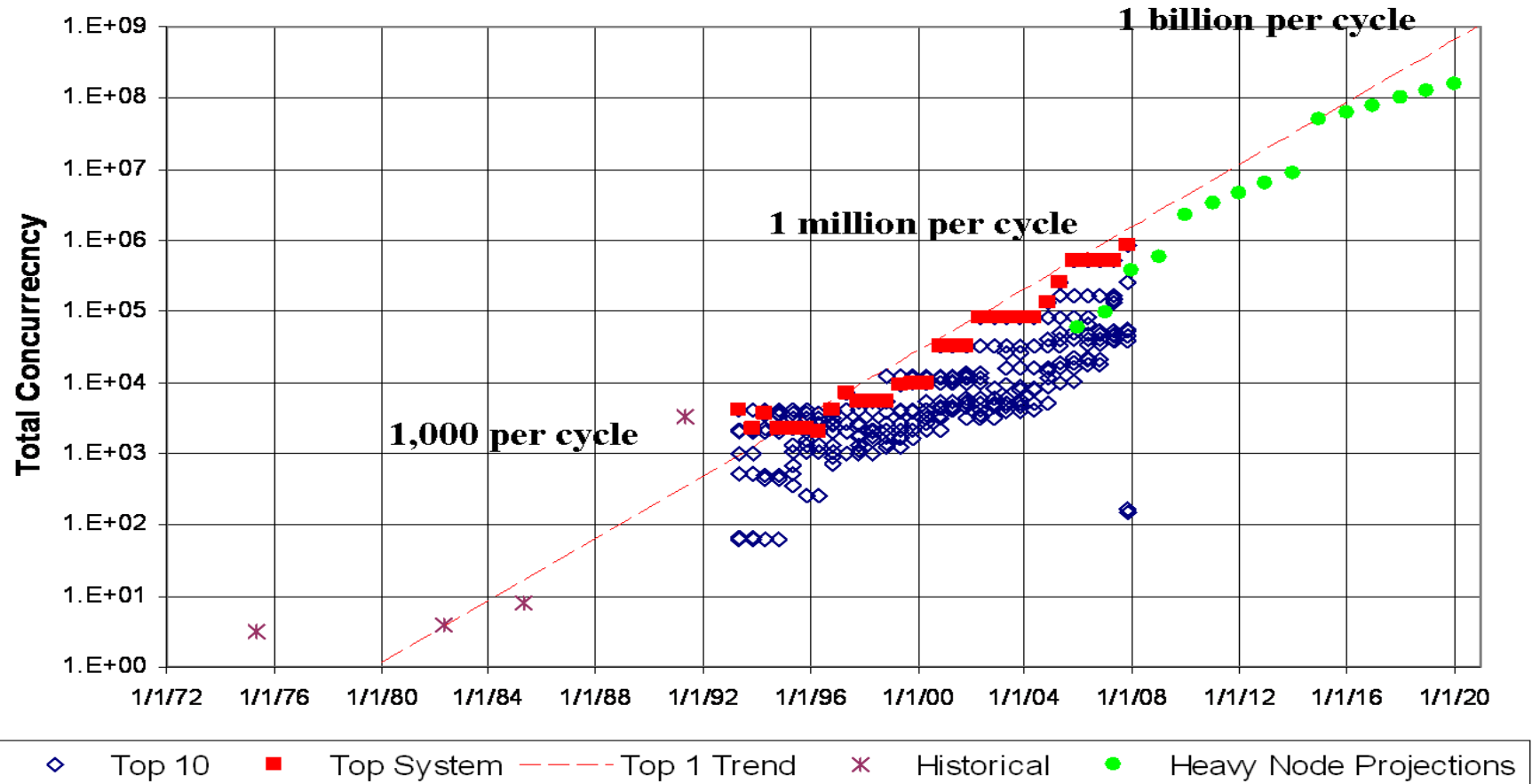
Low Power Design Principles

Tensilica
XTensa

- Power5 (server)
 - 120W@1900MHz
 - **Baseline**
- Intel Core2 sc (laptop) :
 - 15W@1000MHz
 - **4x more FLOPs/watt than baseline**
- Intel Atom (handhelds)
 - 0.625W@800MHz
 - **80x more**
- Tensilica XTensa DP (Moto Razor) :
 - 0.09W@600MHz
 - **400x more (80x-100x sustained)**

Even if each simple core is 1/4th as computationally efficient as complex core, you can fit hundreds of them on a single chip and still be 100x more power efficient.

Projected Parallelism for Exascale

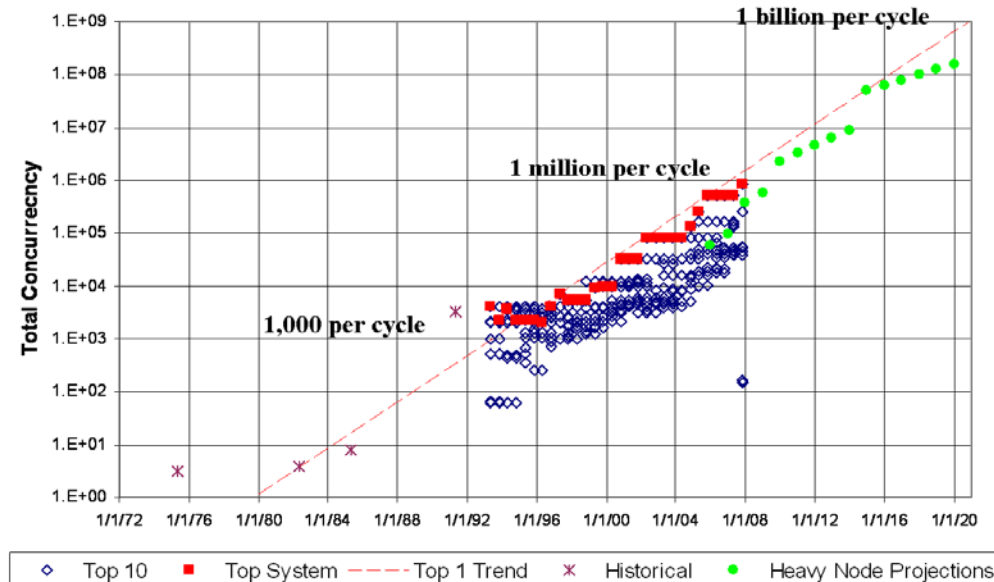


How much parallelism must be handled by the program?

From Peter Kogge (on behalf of Exascale Working Group), "Architectural Challenges at the Exascale Frontier", June 20, 2008

Conclusion: Solving Logic Power Drives Move to Massive Parallelism

- Future HPC must move to simpler power-efficient core designs
 - Embedded/consumer electronics technology is central to the future of HPC
 - Convergence inevitable because it optimizes both cost and power efficiency



How much parallelism must be handled by the program?
From Peter Kogge (on behalf of Exascale Working Group), "Architectural Challenges at the Exascale Frontier", June 20, 2008

- Consequence is massive on-chip parallelism
 - A thousand cores on a chip by 2018
 - 1 Million to 1 Billion-way System Level Parallelism
 - *Must express massive parallelism in algorithms and pmodels*
 - *Must manage massive parallelism in system software*

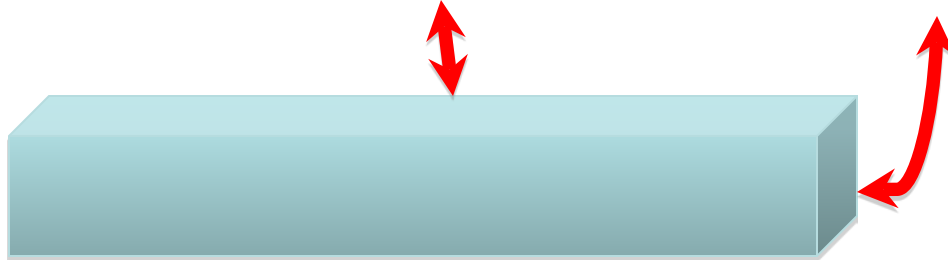
The Cost of Data Movement

How do those cores talk to each other?

The problem with Wires:

Energy to move data proportional to distance

- Cost to move a bit on copper wire:
 - $\text{energy} = \text{bitrate} * \text{Length}^2 / \text{cross-section area}$



- 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*

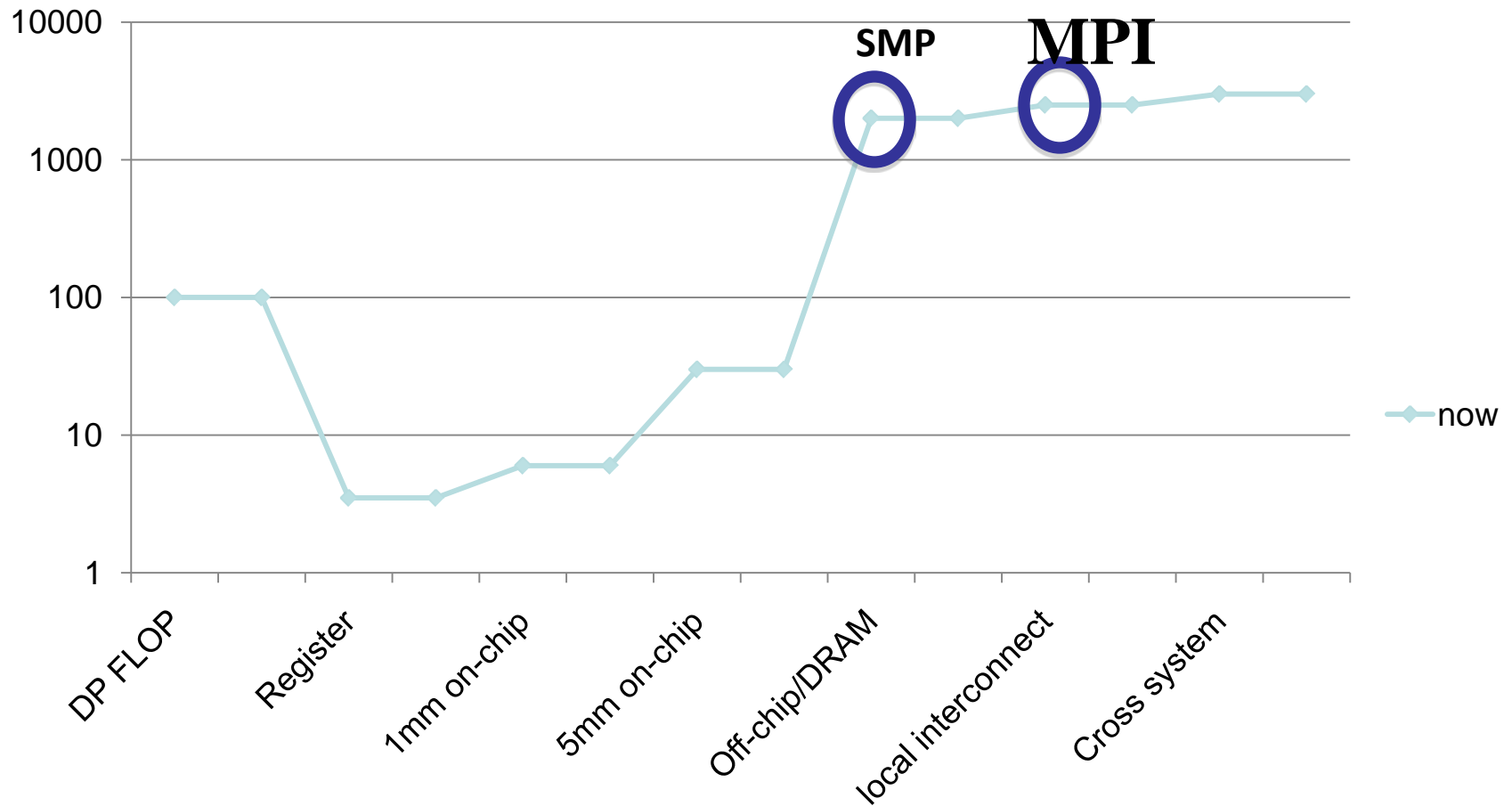
Photonics requires no redrive
and passive switch little power



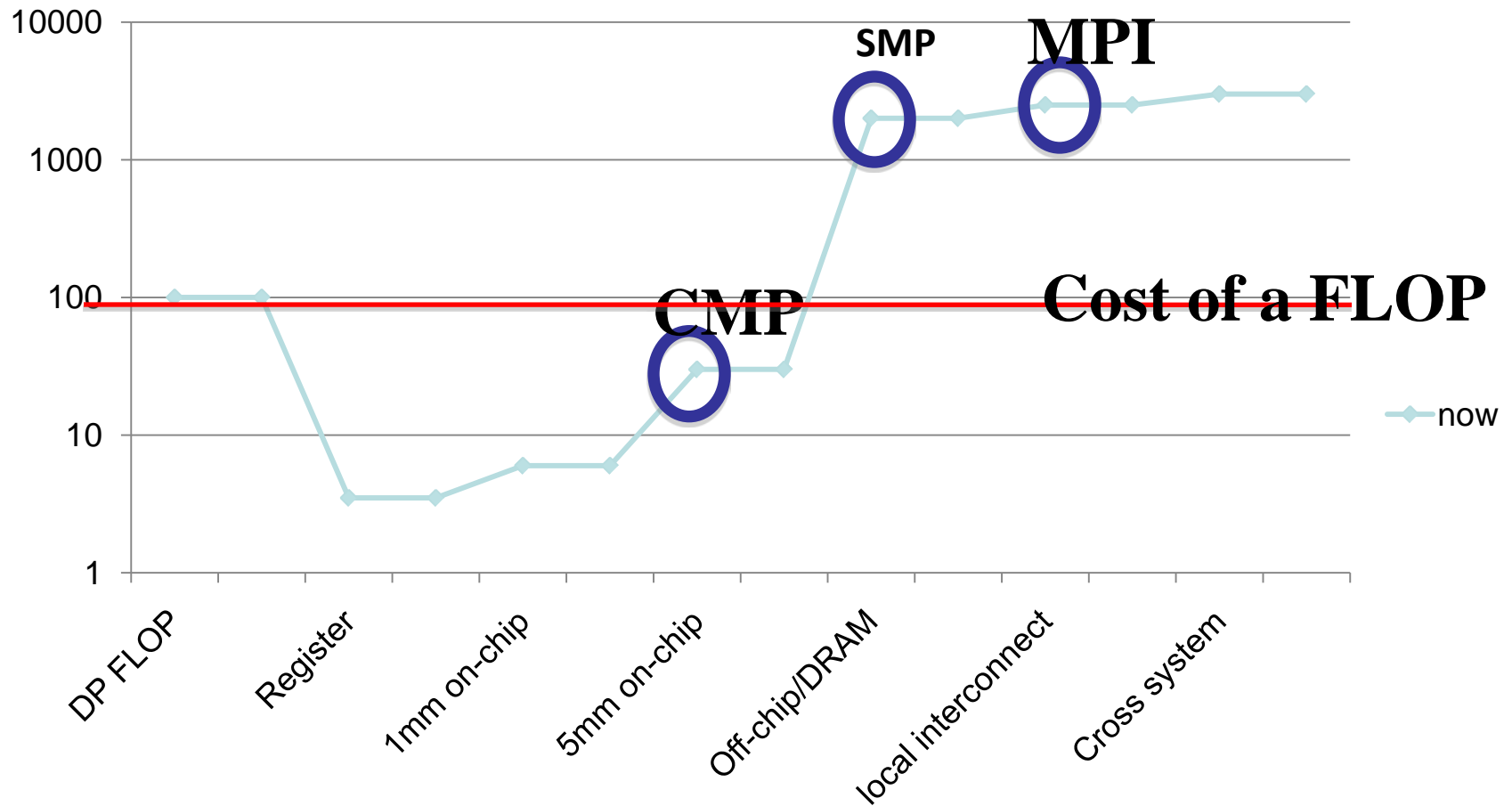
Copper requires to signal amplification
even for on-chip connections



The Cost of Data Movement

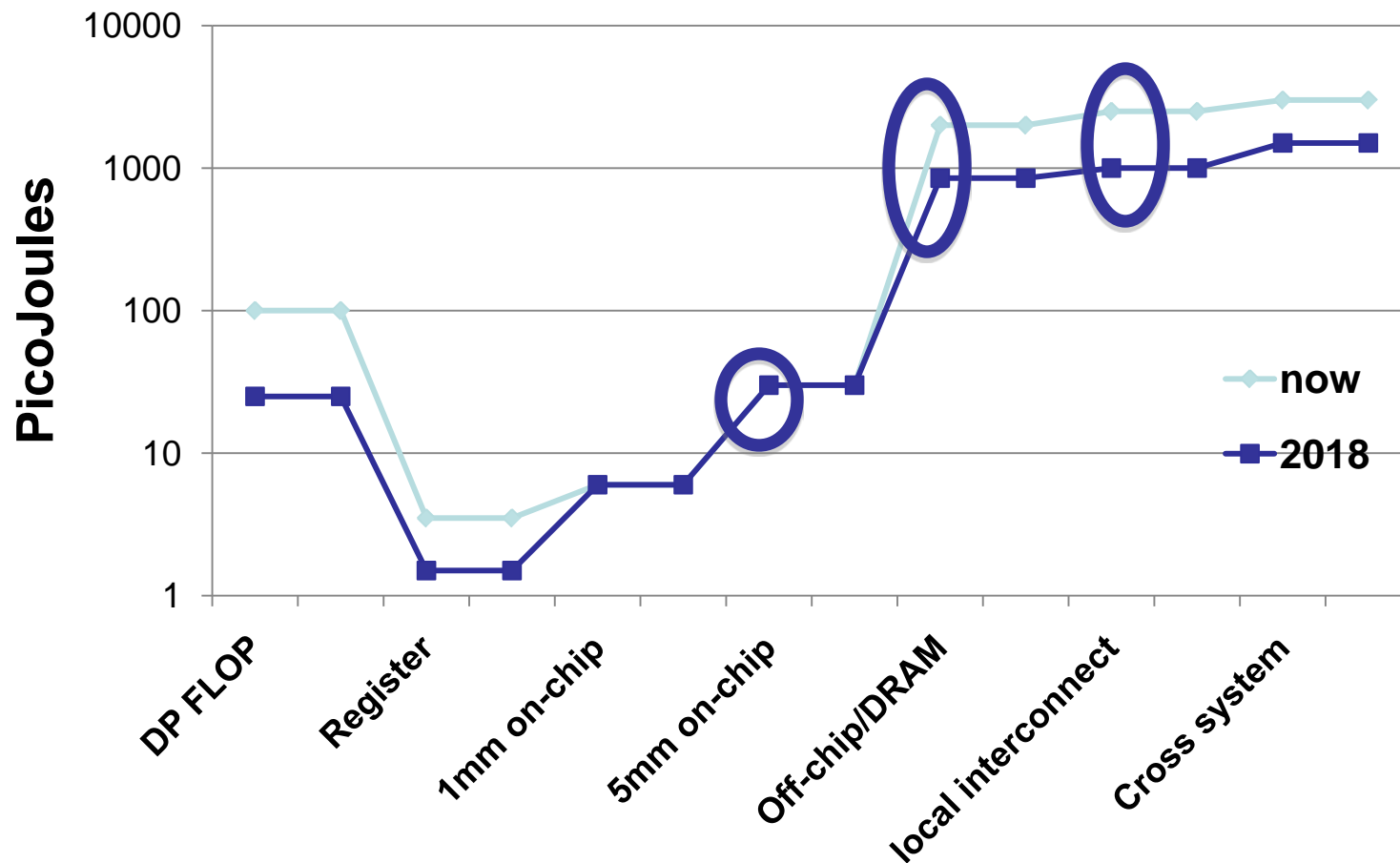


The Cost of Data Movement



The situation will not improve in 2018

Energy Efficiency will require careful management of data locality



Important to know when you are on-chip and when data is off-chip!

Memory



U.S. DEPARTMENT OF
ENERGY

Office of
Science



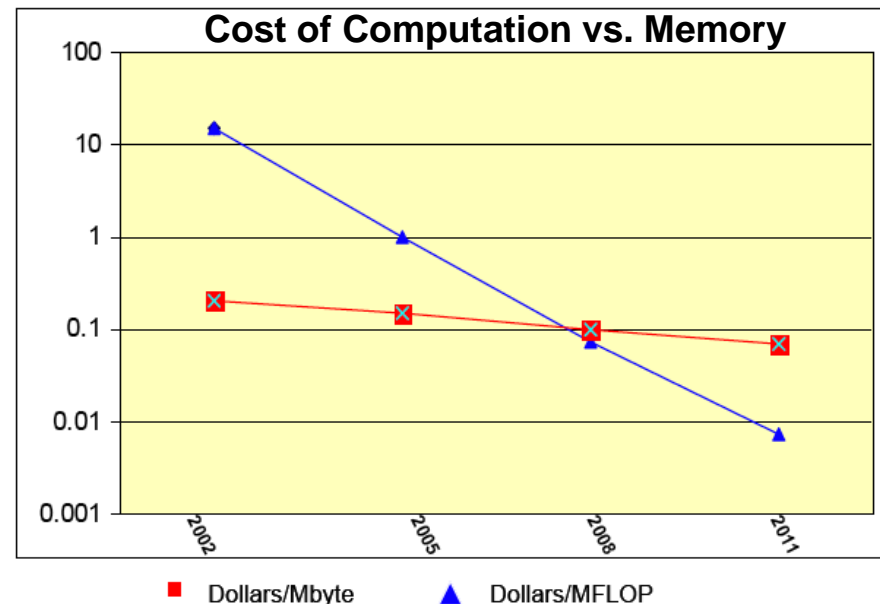
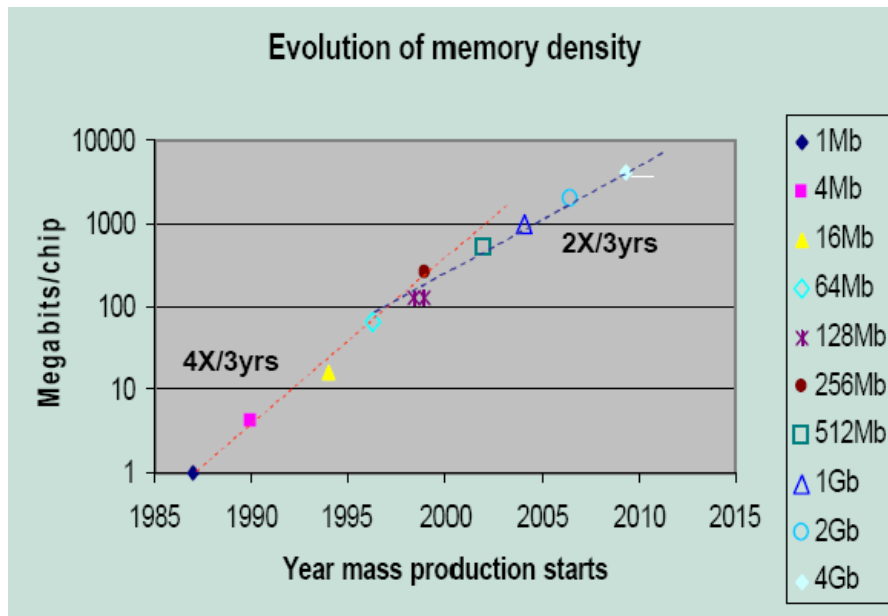
BERKELEY LAB

LAWRENCE BERKELEY NATIONAL LABORATORY

Managed by the University of California for the U.S. Department of Energy

Projections of Memory Density Improvements

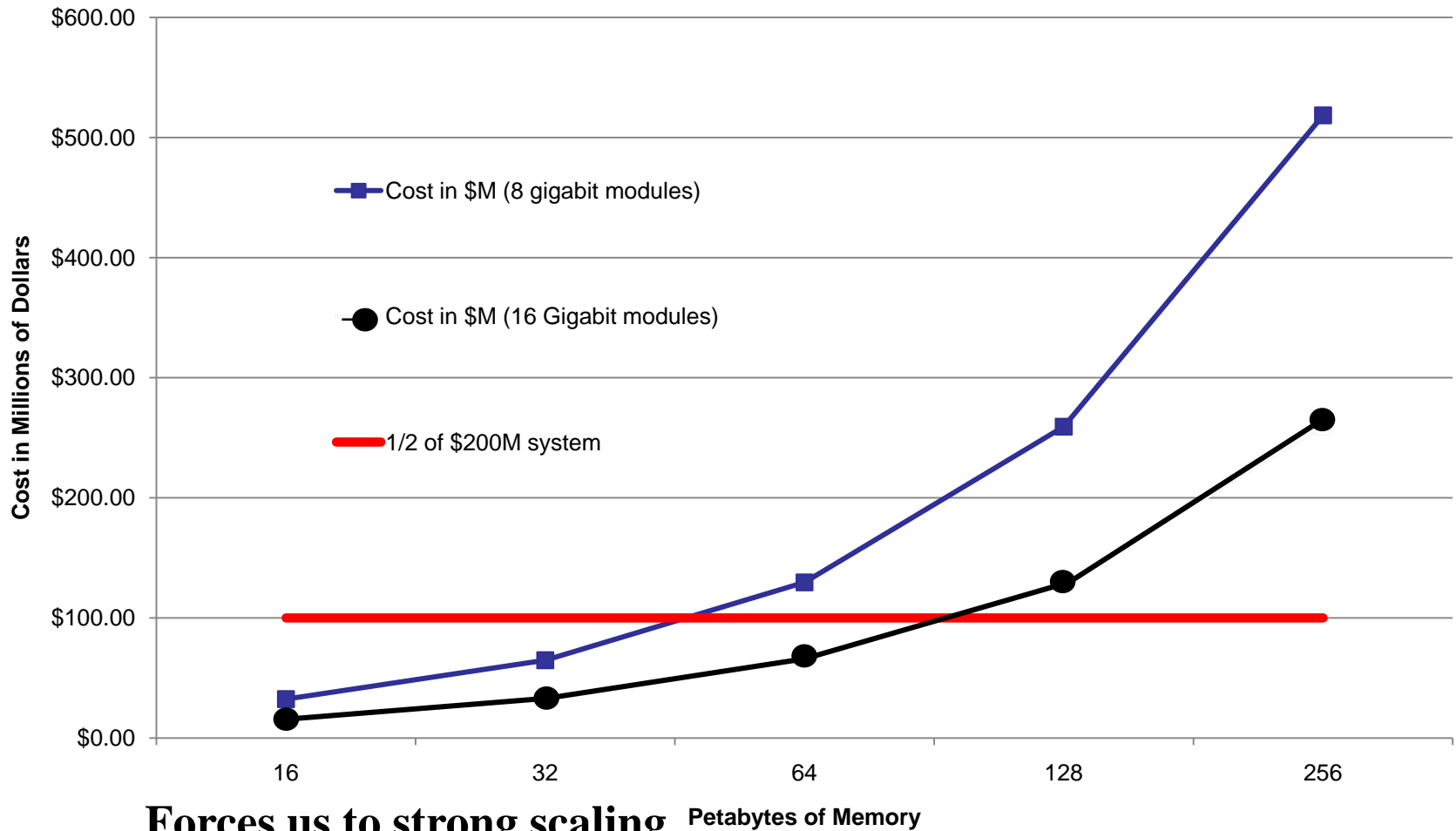
- Memory density is doubling every three years; processor logic is every two
 - Project 8Gigabit DIMMs in 2018
 - 16Gigabit if technology acceleration (or higher cost for early release)
- Storage costs (dollars/Mbyte) are dropping gradually compared to logic costs
 - Industry assumption: \$1.80/memory chip is median commodity cost



The cost to sense, collect, generate and calculate data is declining much faster than the cost to access, manage and store it

Cost of Memory Capacity

2 different potential Memory Densities

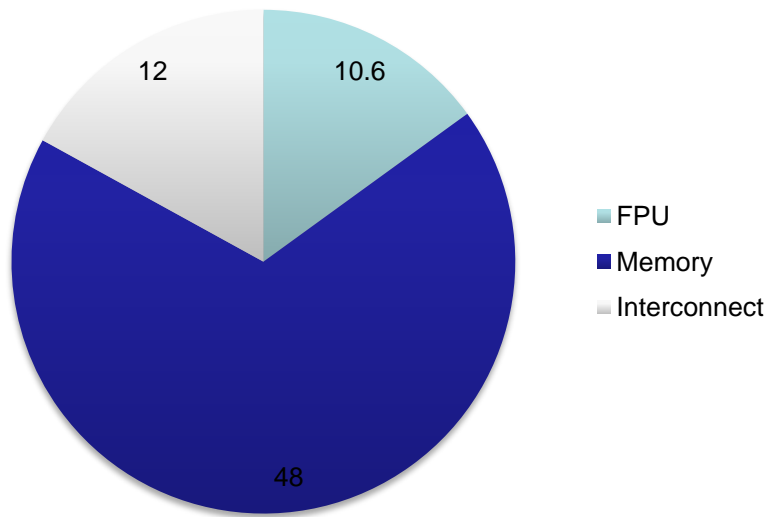


Forces us to strong scaling

Forces us to memory conservative communication (GAS)

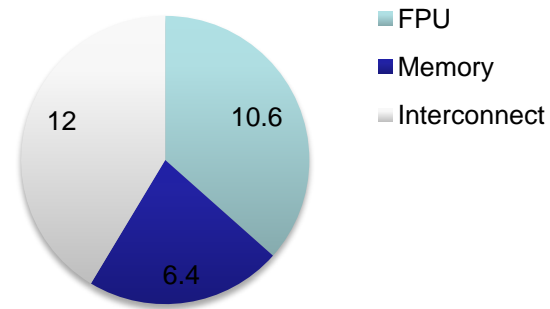
Exascale Memory Power Consumption (San Diego Meeting)

- Power consumption with standard technology roadmap



70 MW total

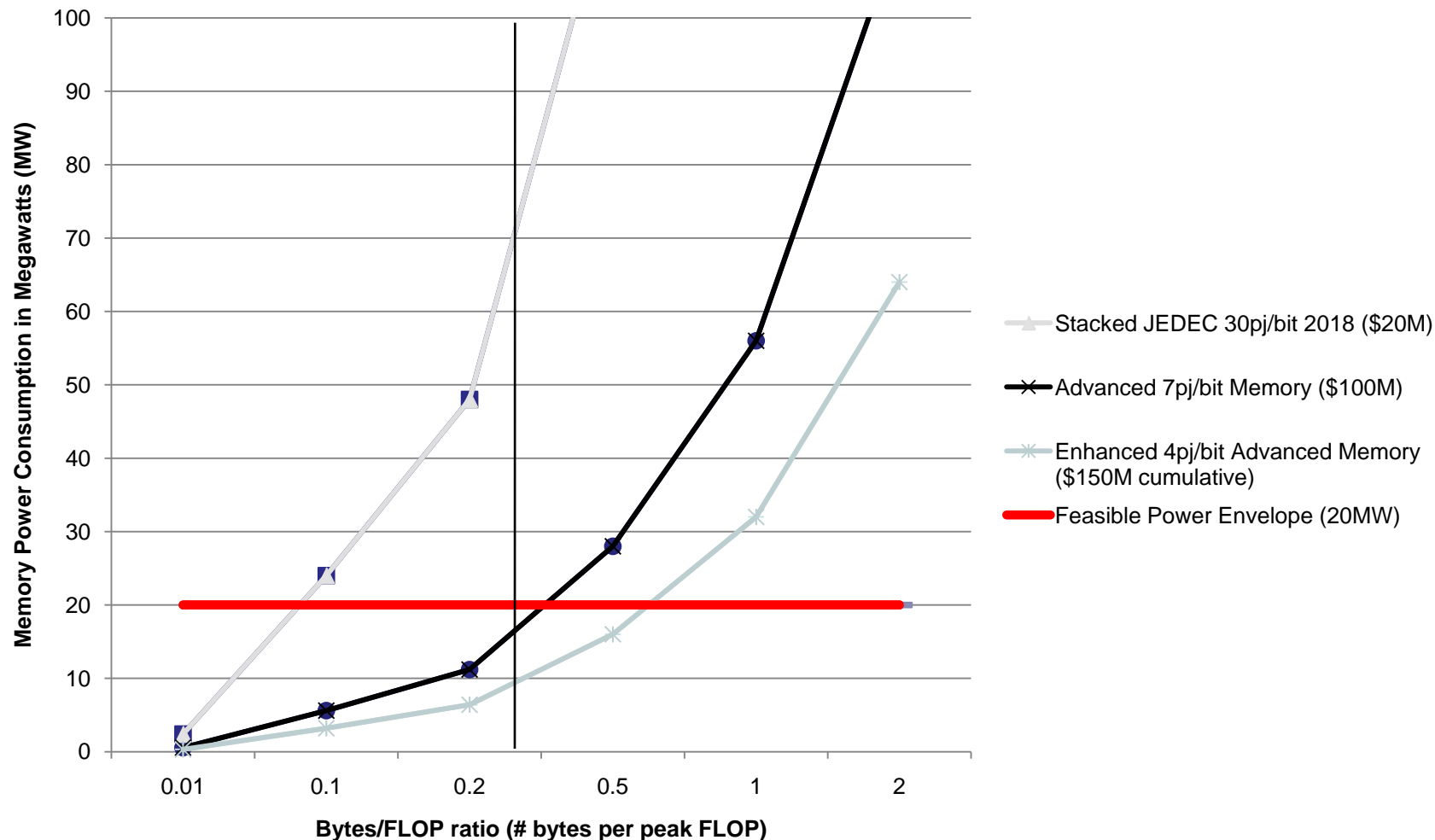
- Power consumption with investment in advanced memory technology



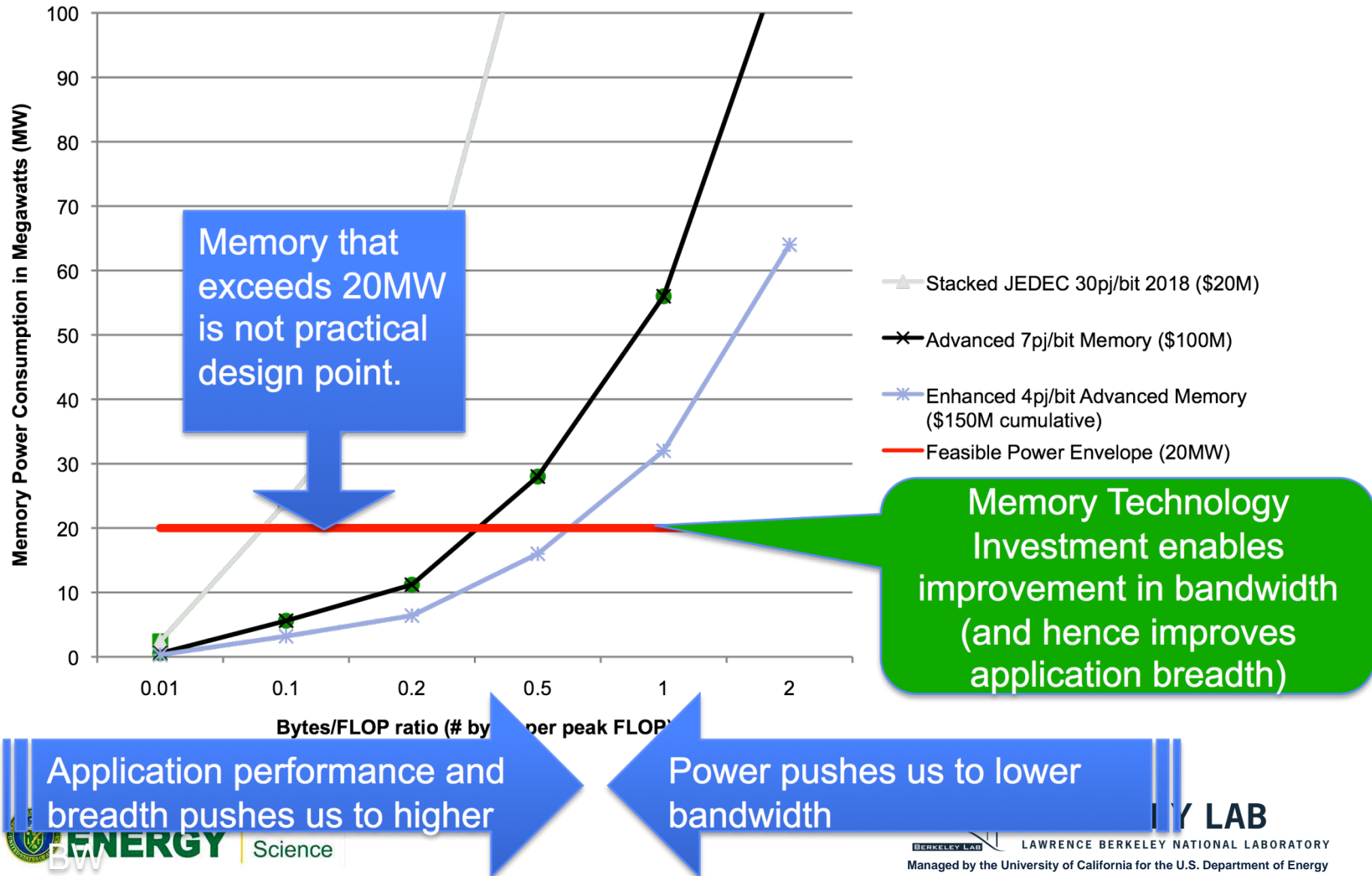
20 MW total

Memory Technology

Bandwidth costs power



Limiting Memory Bandwidth Limits System Scope



Power Considerations Drive Future Architectures in the Exascale Era

- **Massive parallelism with low power processors**
- **Limited amount of memory, low memory/flop ratios (processing is free)**
- **Cost of data movement, locality is becoming more important**

What are critical exascale technology investments?

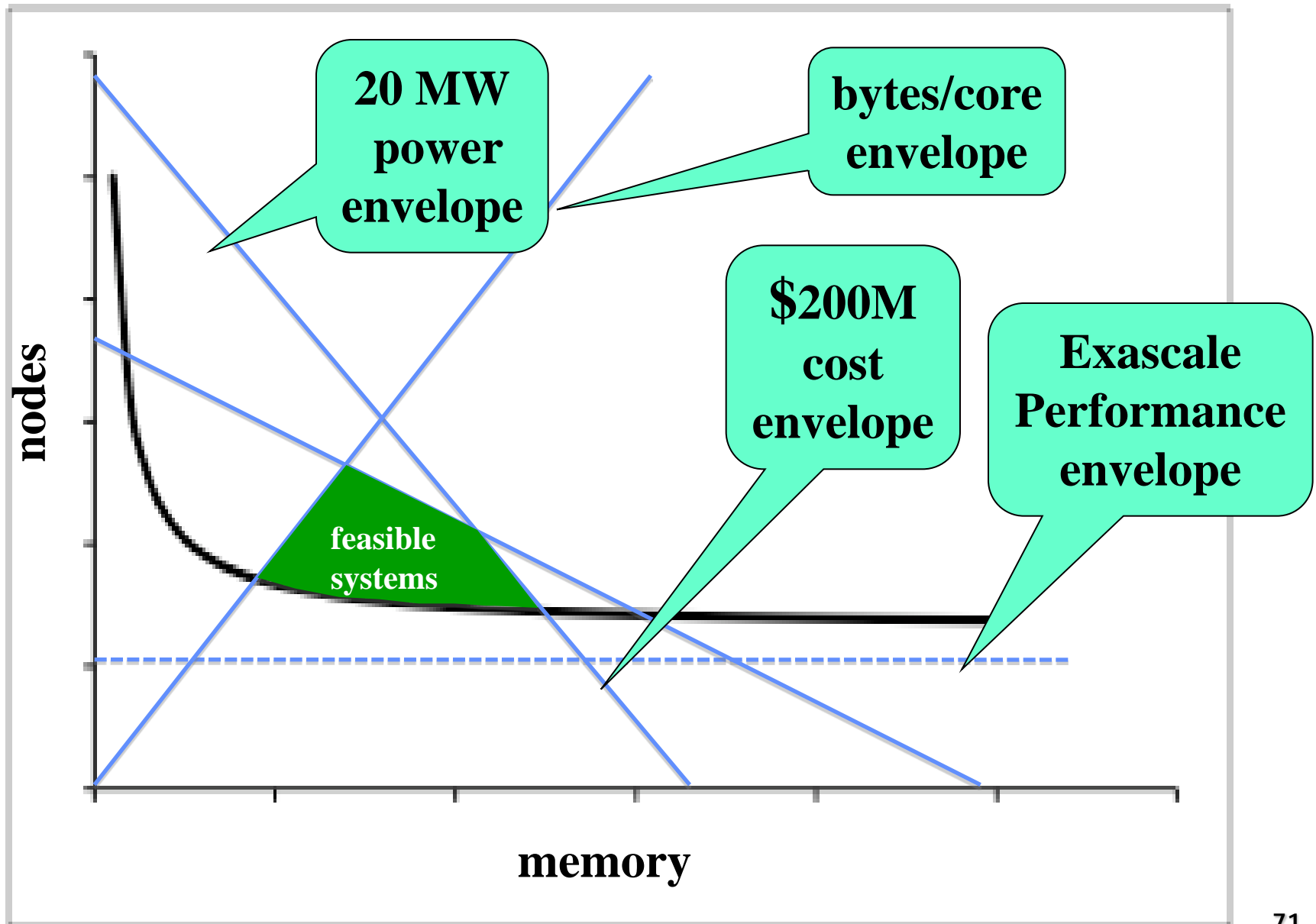
- **System power** is a first class constraint on exascale system performance and effectiveness.
- **Memory** is an important component of meeting exascale power and applications goals.
- **Programming model.** Early investment in several efforts to decide in 2013 on exascale programming model, allowing exemplar applications effective access to 2015 system for both mission and science.
- **Investment in exascale processor design** to achieve an exascale-like system in 2015.
- **Operating System strategy for exascale** is critical for node performance at scale and for efficient support of new programming models and run time systems.
- **Reliability and resiliency are critical at this** scale and require applications neutral movement of the file system (for check pointing, in particular) closer to the running apps.
- ***HPC co-design strategy and implementation*** requires a set of a hierarchical performance models and simulators as well as commitment from apps, software and architecture communities.

Overview

- **From 1999 to 2009: evolution from Teraflops to Petaflops computing**
- **From 2010 to 2020: key technology changes towards Exaflops computing**
- **Impact on Computational Science**
 - **Co-design**



The trade space for exascale is very complex.



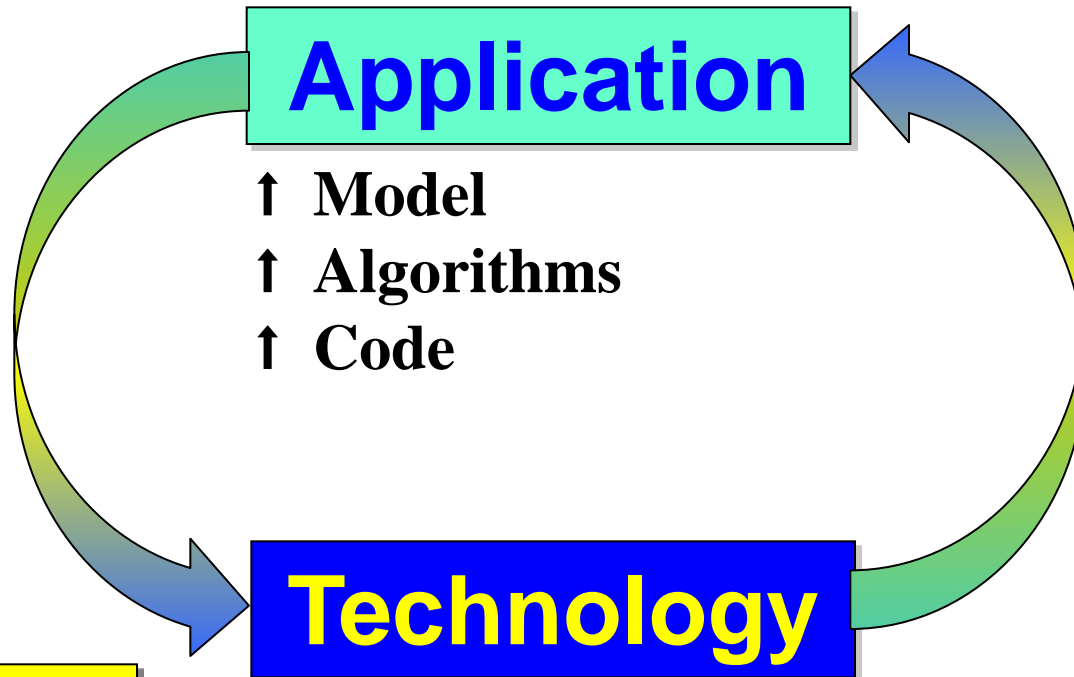


Co-design expands the feasible solution space to allow better solutions.

Application driven:

Find the best technology to run this code.

Sub-optimal



Now, we must expand the co-design space to find better solutions:

- *new applications & algorithms,*
- *better technology and performance.*

Technology driven:

Fit your application to this technology.

Sub-optimal.

A first step toward co-design was the last exascale workshop.

- The approach will be to engage experts in computational science, applied mathematics and CS with the goal of

Cross-cutting Technologies for Computing at the Exascale

February 2-5, 2010 · Washington, D.C.



- Producing a first cut at the characteristics of systems that (a) could be fielded by 2018 and (b) would meet applications' needs
- Outlining the R&D needed for "co-design" of system architecture, system software and tools, programming frameworks, mathematical models and algorithms, and scientific application codes at the exascale, and
- Exploring whether this anticipated phase change in technology (like parallel computing in 1990s) provides any opportunities for applications. That is, whether a requirement for revolutionary application design allows new methods, algorithms, and mathematical models to be brought to bear on mission and science questions.

Summary of some priority research directions (PRD)

Black – Crosscutting workshop report

Green – HDS interpretation

- *Investigate and develop new exascale programming paradigms to support ‘billion-way’ concurrency*
 - *Think 10,000 times more parallel*
 - *Expect MPI+X programming model*
 - *Think of algorithms that can easily exploit the intra node parallelism, especially if CS researchers develop automatics tools for X*

Summary of some priority research directions (PRD) -- cont.

- *Re-cast critical applied mathematics algorithms to reflect impact of anticipated macro architecture evolution, such as memory and communication constraints*
 - *Live with less memory/thread and less bandwidth*
- *Develop new mathematical models and formulations that effectively exploit anticipated exascale hardware architectures*
 - *Add more physics and not just more refinement*
- *Address numerical analysis questions associated with moving away from bulk-synchronous programs to multi-task approaches*
 - *No more SPMD; think of mapping coarse grain data flow in frameworks*

Summary of some priority research directions (PRD) – cont.

- *Adapt data analysis algorithms to exascale environments*
- *Extract essential elements of critical science applications as “mini-applications” that hardware and system software designers can use to understand computational requirements*
- *Develop tools to simulate emerging architectures for use in co-design*
 - *Applied mathematicians/computer scientists should be ready to lead co-design teams*

Summary

- **Major Challenges are ahead for extreme computing**
 - Power
 - Parallelism
 - ... and many others not discussed here
- **We will need completely new approaches and technologies to reach the Exascale level**
- **This opens up many new opportunities for computer scientists, applied mathematicians, and computational scientists**

Shackleton's Quote on Exascale



Ernest Shackleton's 1907 ad in London's Times, recruiting a crew to sail with him on his exploration of the South Pole

“Wanted. Men/women for hazardous architectures. Low wages. Bitter cold. Long hours of software development. Safe return doubtful. Honor and recognition in the event of success.”