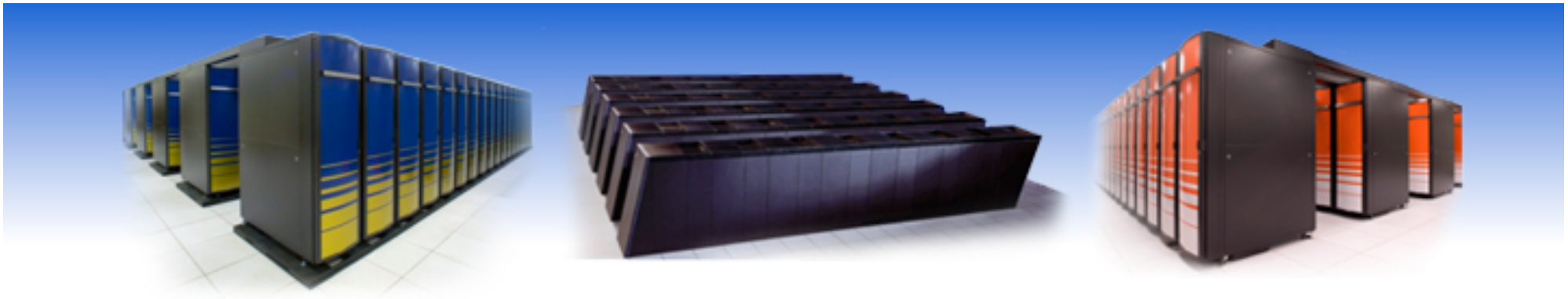


Future Trends in Computing

Horst Simon
Lawrence Berkeley National Laboratory
and UC Berkeley

CS267 – Lecture 21
April 6, 2010

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

Key Message

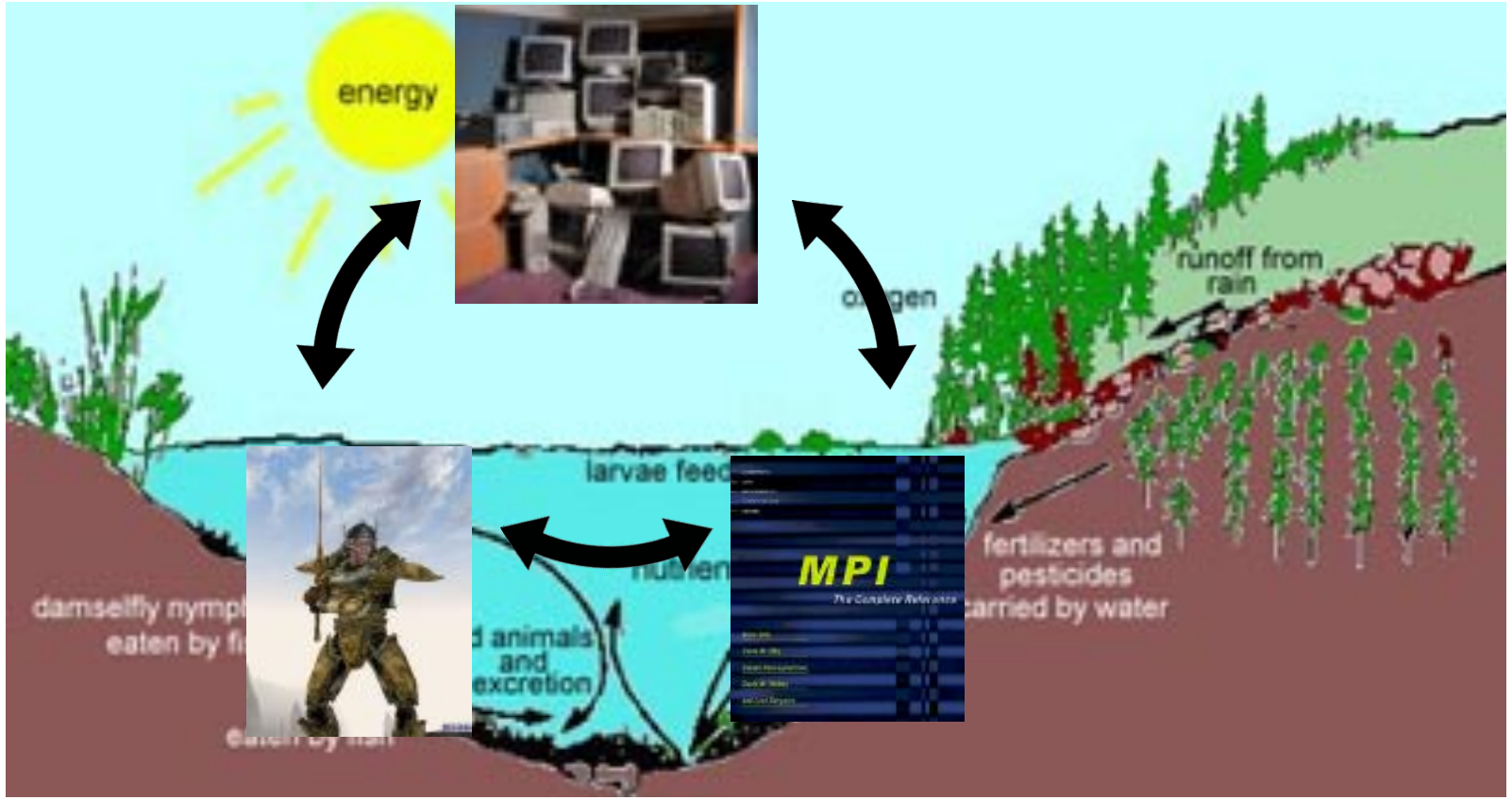
Computing is changing more rapidly than ever before, and scientists have the unprecedented opportunity to change computing directions

Overview

- **Turning point in 2004**
- **Current trends and what to expect until 2014**
- **Long term trends until 2019**

Supercomputing Ecosystem (2005)

Commercial Off The Shelf technology (COTS)



“Clusters”

12 years of legacy MPI applications base

From my presentation at ISC 2005

Supercomputing Ecosystem (2005)

Commercial Off The Shelf technology (COTS)

“Clusters”

12 years of legacy MPI applications base

From my presentation at ISC 2005

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!

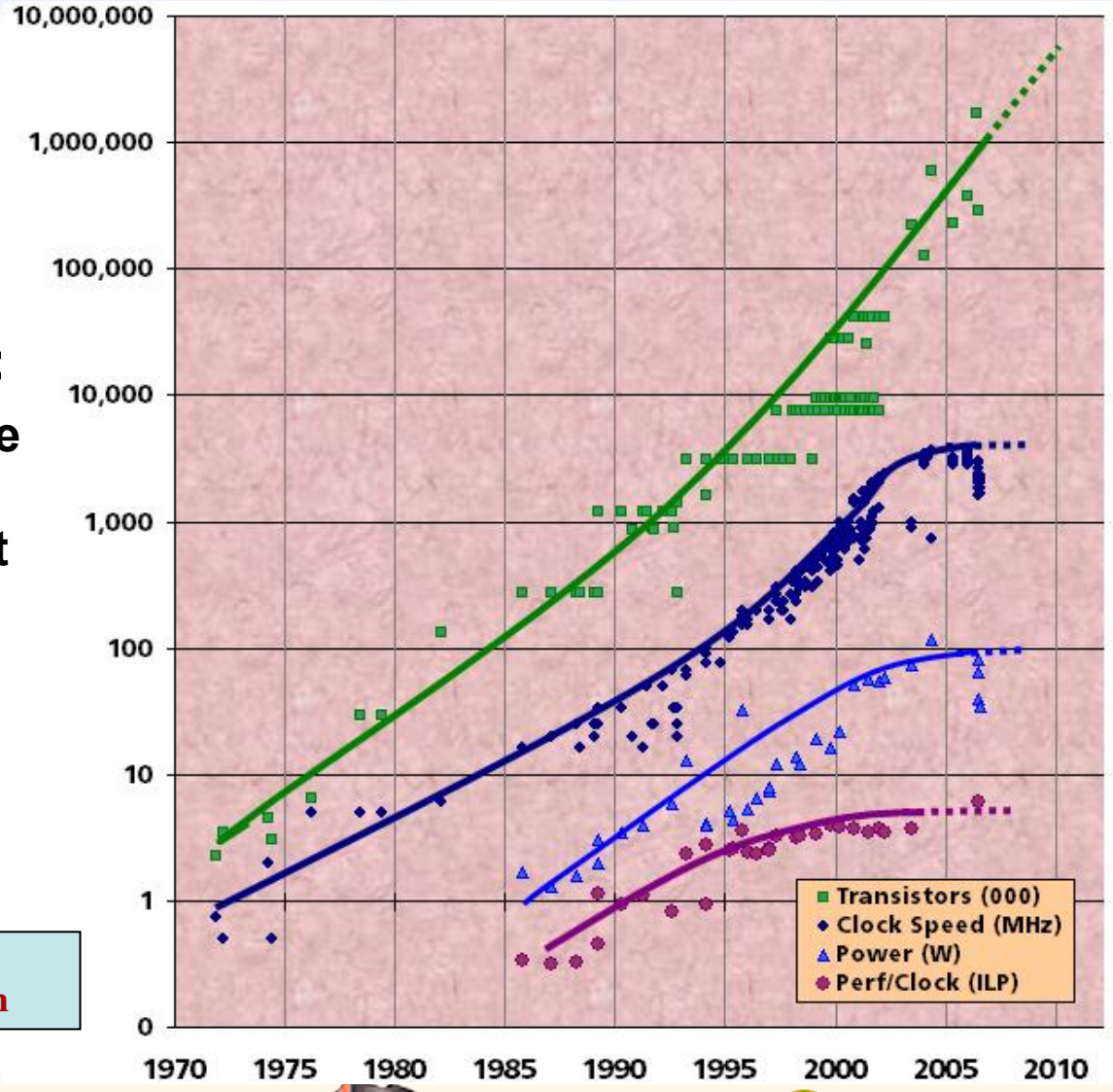


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Supercomputing Ecosystem (~~2005~~) 2010

Commercial Off The Shelf technology (COTS)

PCs and desktop systems are no longer the economic driver.

Architecture and programming model are about to change

"Clusters"

12 years of legacy MPI applications base

Overview

- Turning point in 2004
- Current trends and what to expect until 2014
- Long term trends until 2019

The TOP500 Project

- Listing the 500 most powerful computers in the world
- Yardstick: Rmax of Linpack
  - Solve Ax=b, dense problem, matrix is random
- Update twice a year:
  - ISC'xy in June in Germany • SCxy in November in the U.S.
- All information available from the TOP500 web site at: www.top500.org
- Compiled by Dongarra (UTK&ORNL), Meuer (Univ. Mannheim, Germany), Simon, and Strohmaier (LBNL)

34th List: The TOP10

				Country	Cores	Rmax [Tflops]	Power [MW]
1	Oak Ridge National Laboratory	Cray	Jaguar Cray XT5 HC 2.6 GHz	USA	224,162	1,759	6.95
2	DOE/NNSA/LANL	IBM	Roadrunner BladeCenter QS22/LS21	USA	122,400	1,042	2.34
3	University of Tennessee	Cray	Kraken Cray XT5 HC 2.36GHz	USA	98,928	831.7	
4	Forschungszentrum Juelich (FZJ)	IBM	Jugene Blue Gene/P Solution	Germany	294,912	825.5	2.26
5	National SuperComputer Center	NUDT	Tianhe-1 NUDT TH-1 Cluster, Xeon, ATI Radeon, Infiniband	China	71,680	563.1	
6	NASA/Ames Research Center/ NAS	SGI	Pleiades SGI Altix ICE 8200EX	USA	56,320	544.3	2.34
7	DOE/NNSA/LLNL	IBM	BlueGene/L eServer Blue Gene Solution	USA	212,992	478.2	2.32
8	Argonne National Laboratory	IBM	Intrepid Blue Gene/P Solution	USA	163,840	458.6	1.26
9	TACC/U. of Texas	Sun	Ranger SunBlade x6420	USA	62,976	433.2	2.0
10	Sandia National Labs	Sun	Red Sky - Sun Blade x6275, Xeon 2.93 Ghz, Infiniband	USA	41,616	423.9	

BERKELEY LAB

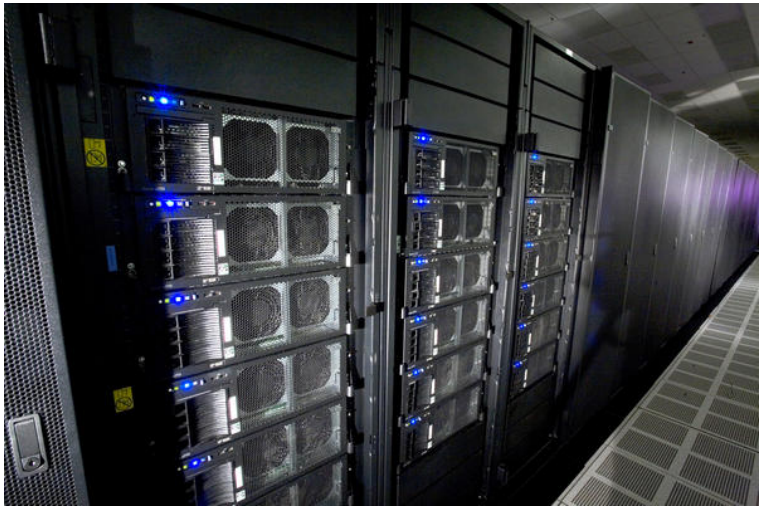
ENERGY
Office of Science

Jaguar @ ORNL: 1.75 PF/s

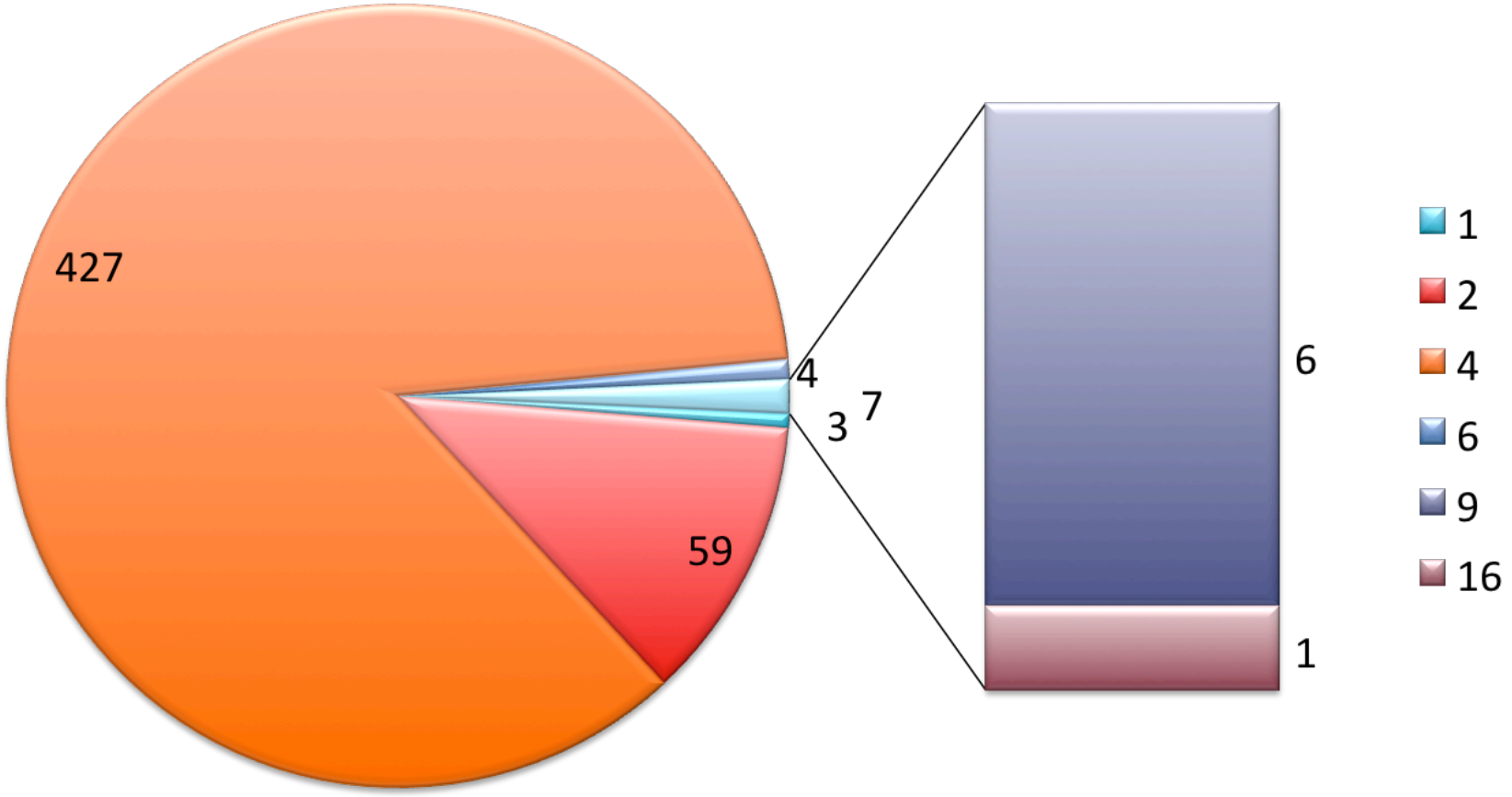
- Cray XT5-HE system
- Over 37,500 quad-core AMD Opteron processors running at 2.6 GHz, 224,162 cores.
- 300 terabytes of memory
- 10 petabytes of disk space.
- 240 gigabytes per second of disk bandwidth
- Cray's SeaStar2+ interconnect network.

Roadrunner @ LANL: 1.04 PF/s

- 12,240 Cell chips (8+1 cores) (on IBM Model QS22 blade servers)
- 6,562 dual-core AMD Opteron (LS21 blades)
- 98 TB main memory
- Power is approximately 2.35 MWs at load
- 278 racks grouped in 18 units
- 5,200 square feet



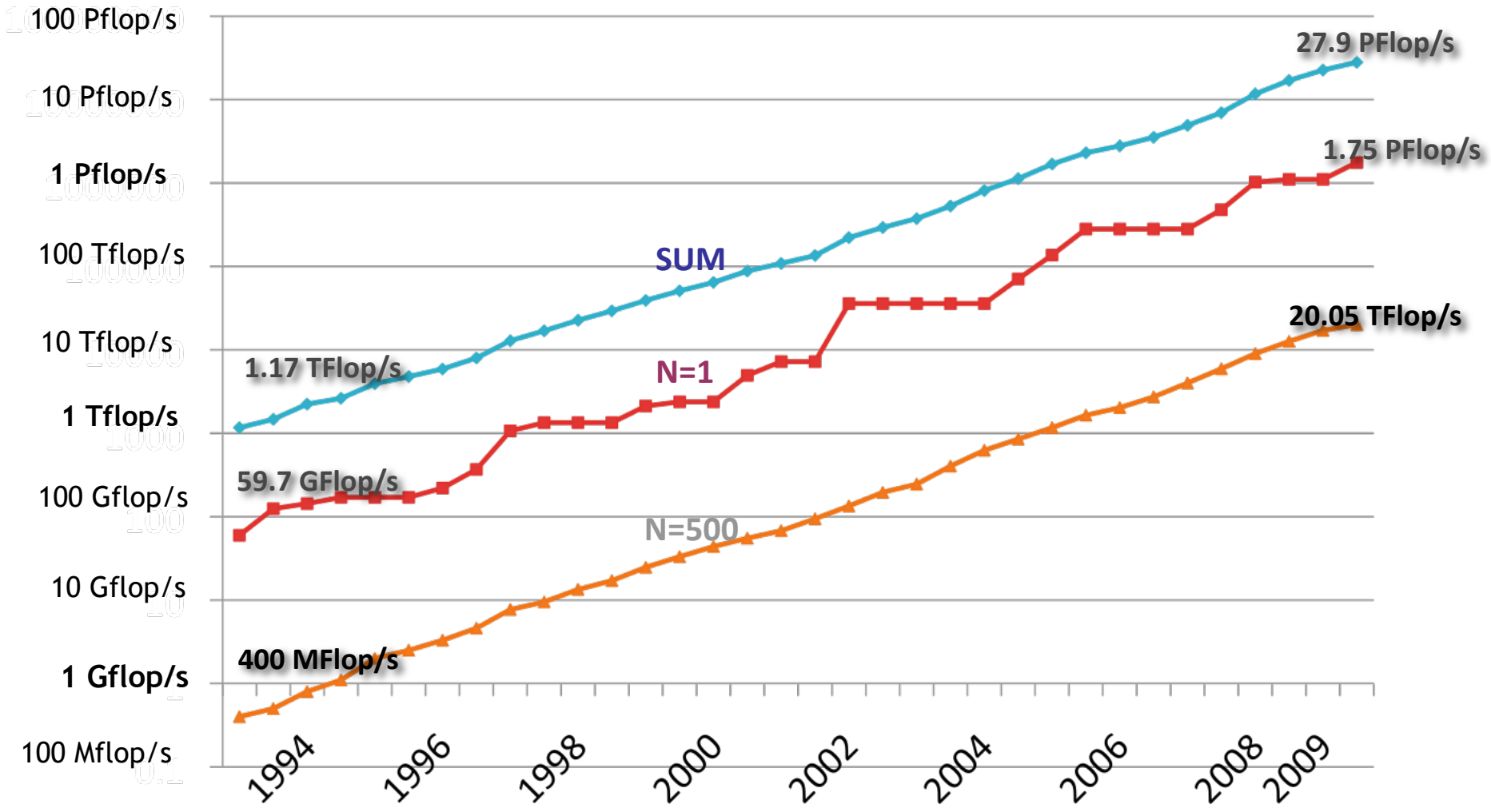
Cores per Socket (Nov. 2009)



Multi-Core and Many-Core

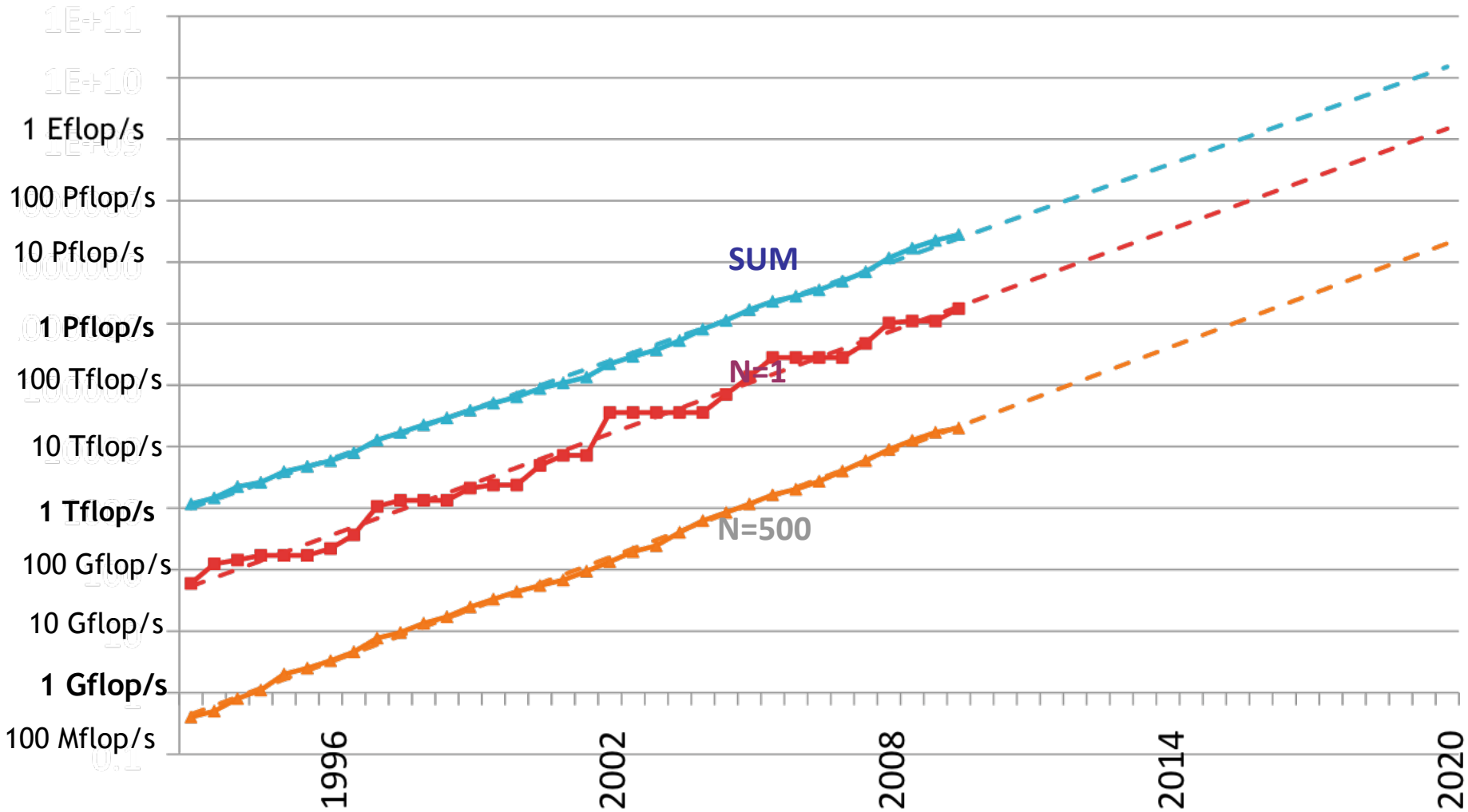
- Power consumption of chips and systems has increased tremendously, because of 'cheap' exploitation of Moore's Law.
  - Free lunch has ended
  - Stall of frequencies forces increasing concurrency levels, Multi-Cores
  - Optimal core sizes/power are smaller than current 'rich cores', which leads to Many-Cores
- Many-Cores, more (10-100x) but smaller cores:
  - Intel Polaris – 80 cores,
  - Clearspeed CSX600 – 96 cores,
  - nVidia G80 – 128 cores, or
  - CISCO Metro – 188 cores

Performance Development

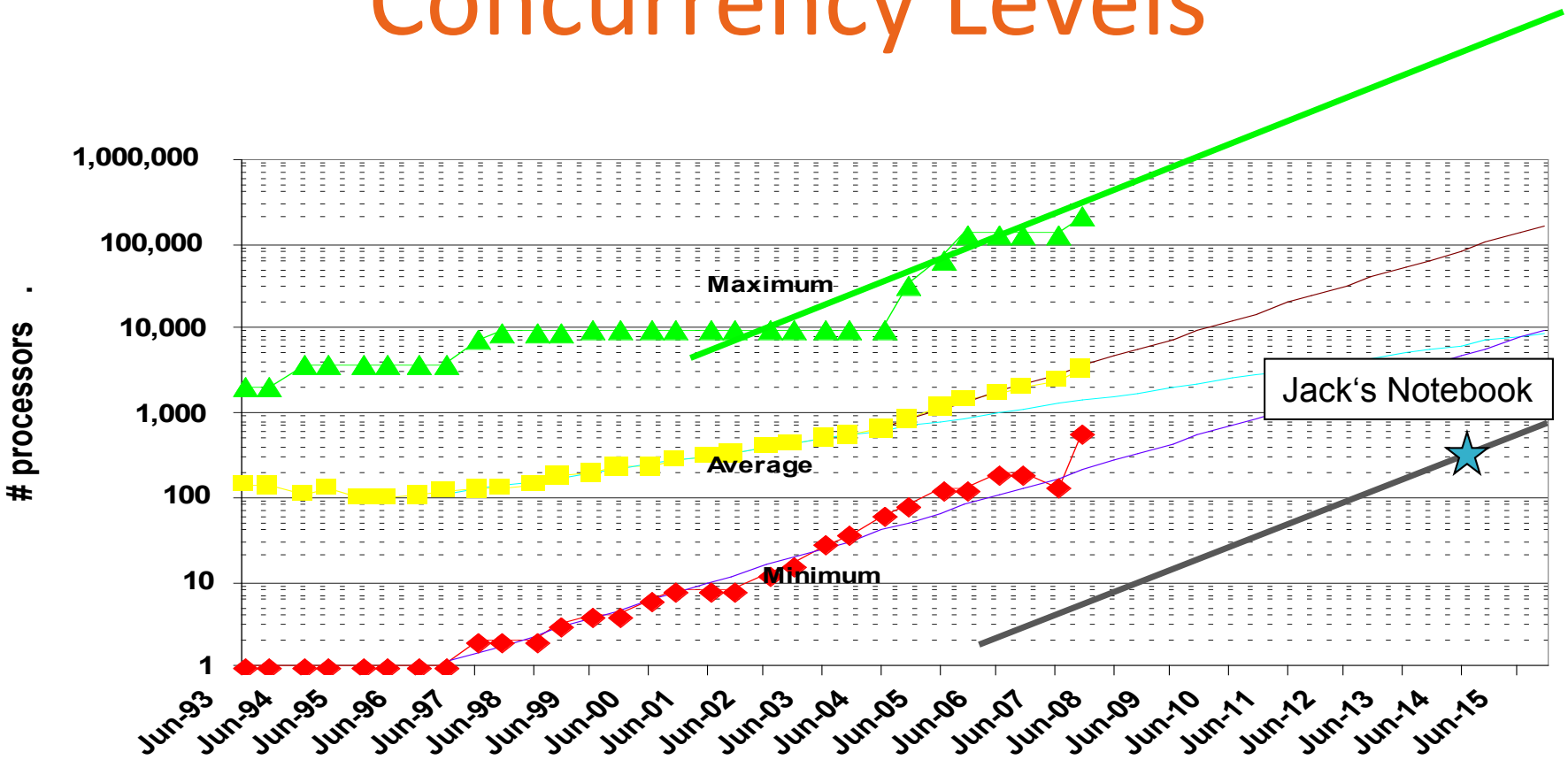


TOP500 SUPERCOMPUTER SITES

Projected Performance Development

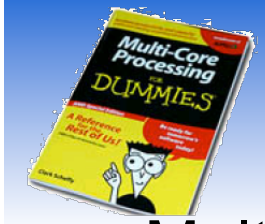


Concurrency Levels



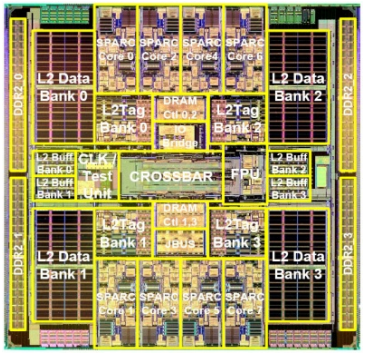
Moore's Law reinterpreted

- 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



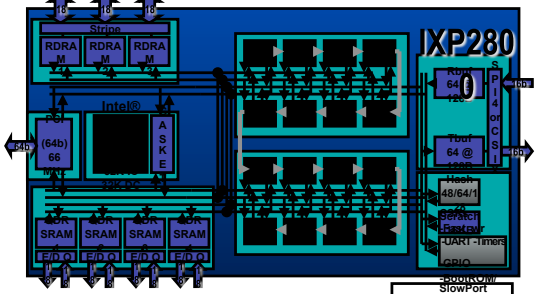
Multicore comes in a wide variety

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

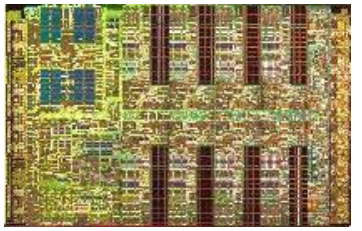


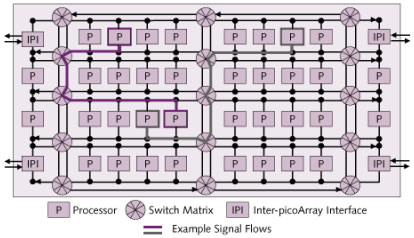
Sun Niagara
8 GPP cores (32 threads)

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



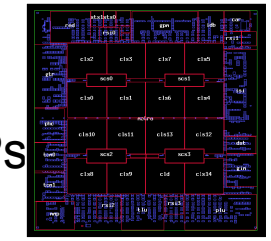
IBM Cell
1 GPP (2 threads)
8 ASPs

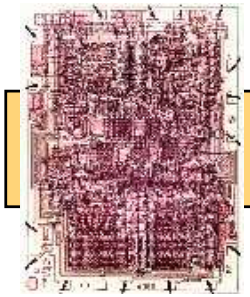




Picochip DSP
1 GPP core
248 ASPs

Cisco CRS-1
188 Tensilica GPPs



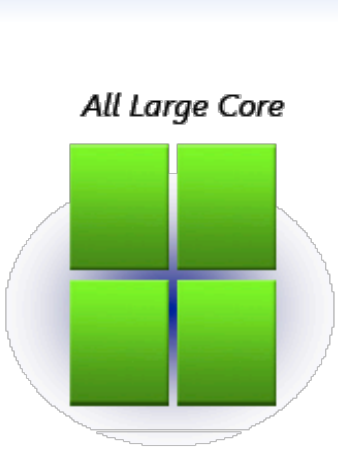


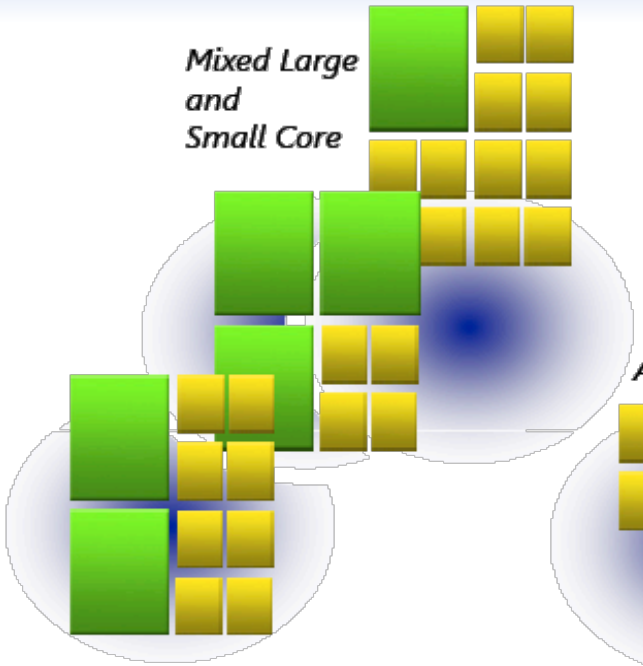
Intel 4004 (1971):
4-bit processor,
2312 transistors,
~100 KIPS,
10 micron PMOS,
11 mm^2 chip

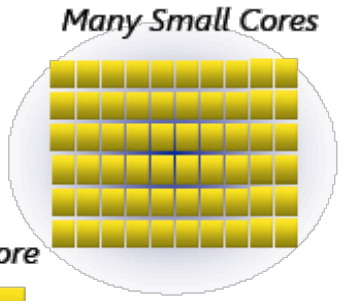
1000s of processor cores per die

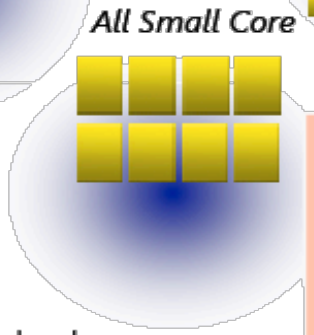
"The Processor is the new Transistor" [Rowen]

What's Next?



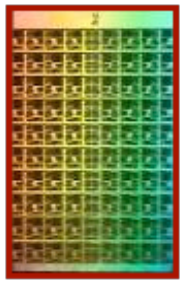




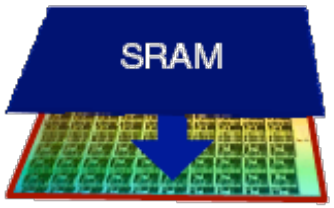


Different Classes of Chips
- Home
- Games / Graphics
- Business
- Scientific

Many Floating-Point Cores



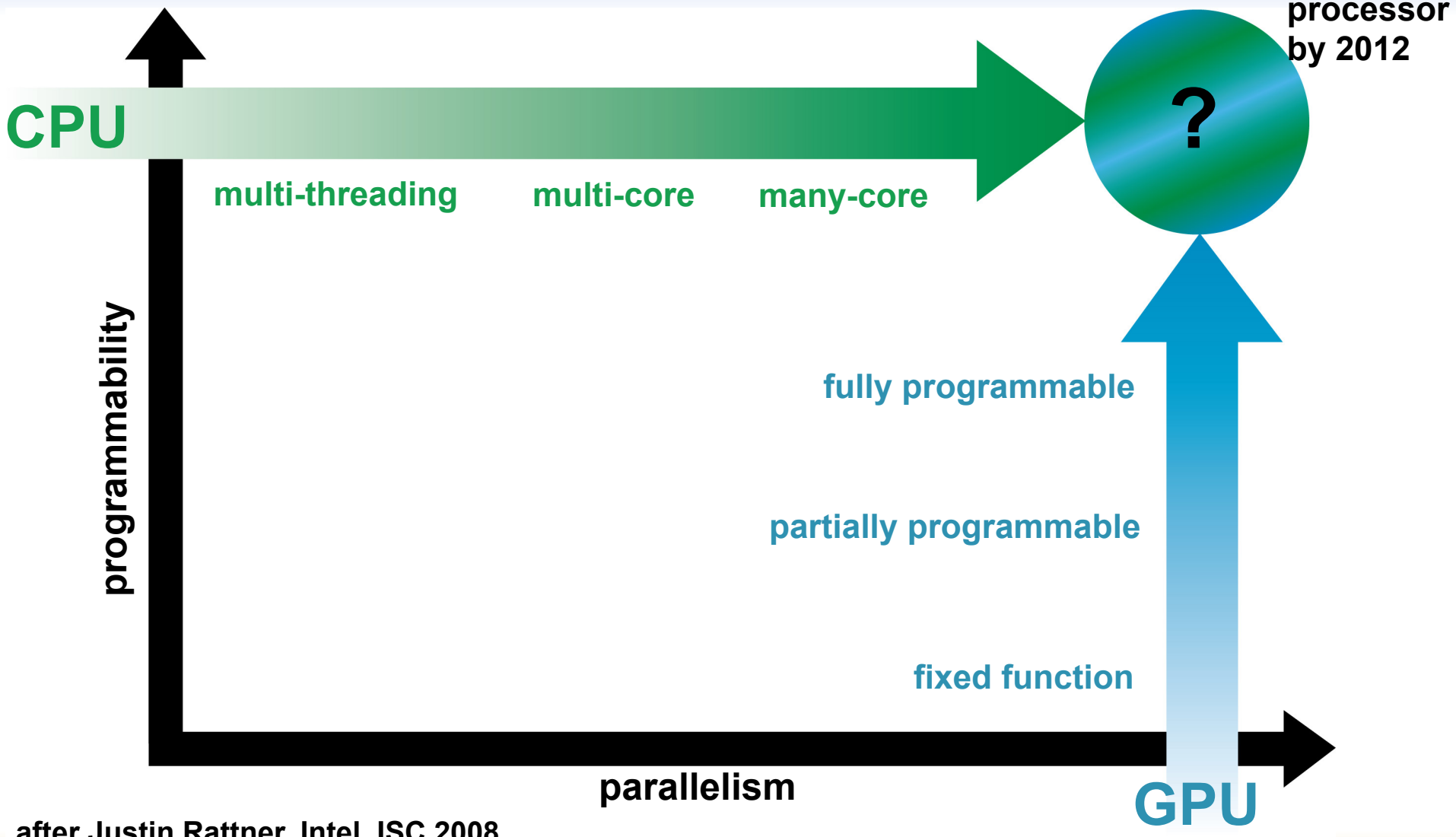
+ 3D Stacked Memory



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

Source: Jack Dongarra, ISC 2008

A Likely Trajectory - Collision or Convergence?



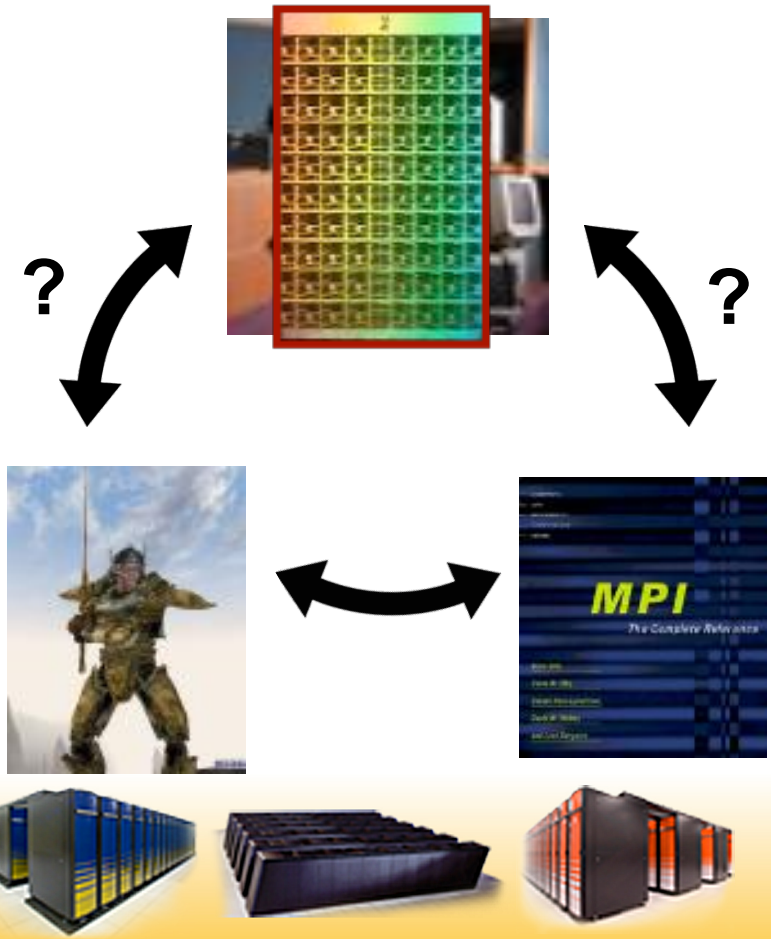
after Justin Rattner, Intel, ISC 2008

Trends for the next five years up to 2013

- After period of rapid architectural change we will likely settle on a future standard processor architecture
- A good bet: Intel will continue to be a market leader
- Impact of this disruptive change on software and systems architecture not clear yet

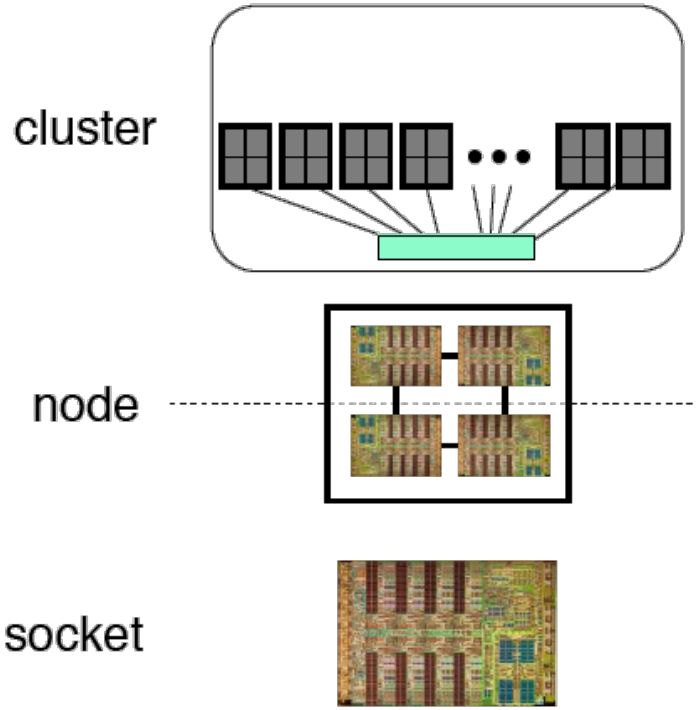
Impact on Software

- **We will need to rethink and redesign our software**
  - **Similar challenge as the 1990 to 1995 transition to clusters and MPI**



A Likely Future Scenario (2014)

System: cluster + many core node



Programming model: MPI+X

Message Passing

Not Message Passing

Hybrid & many core technologies will require new approaches: PGAS, auto tuning, ?

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

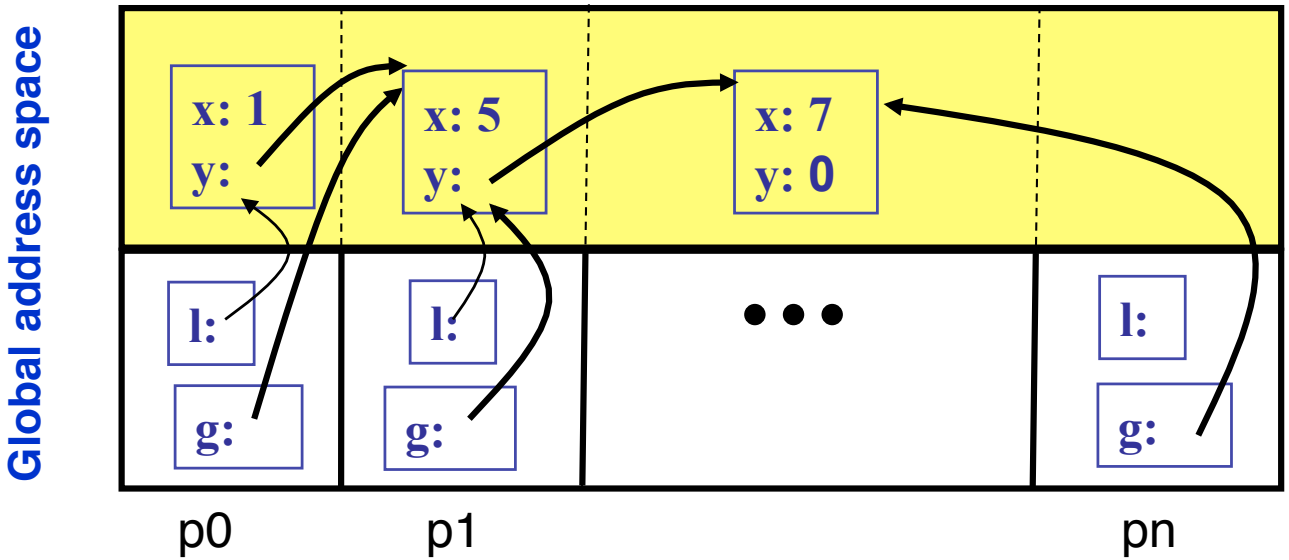
What's Wrong with MPI Everywhere?

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?

PGAS Languages

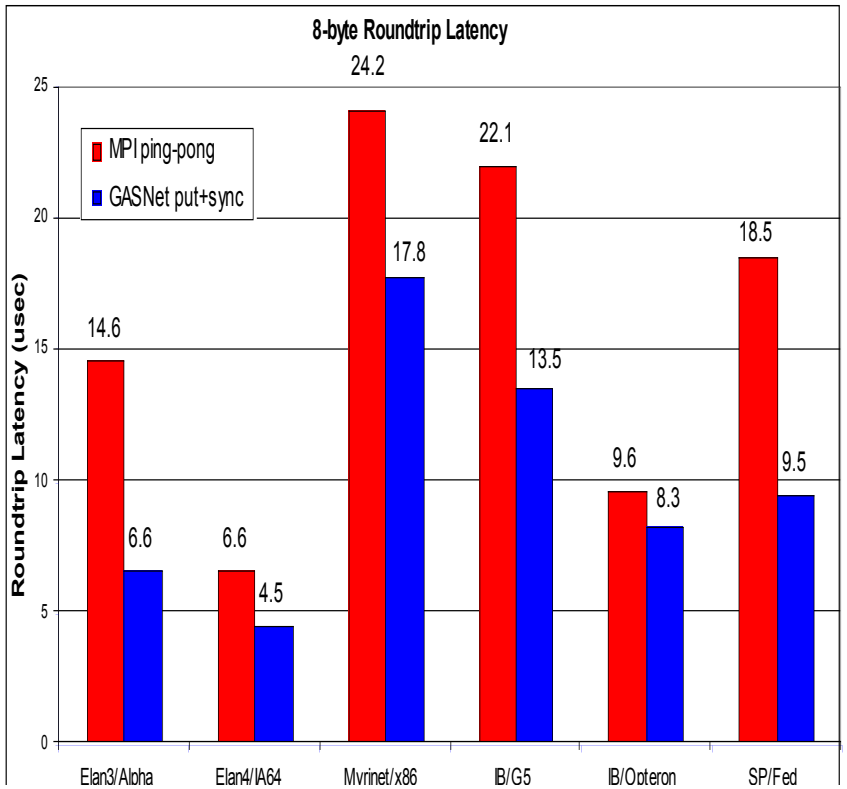
- *Global address space:* thread may directly read/write remote data
- *Partitioned:* data is designated as local or global

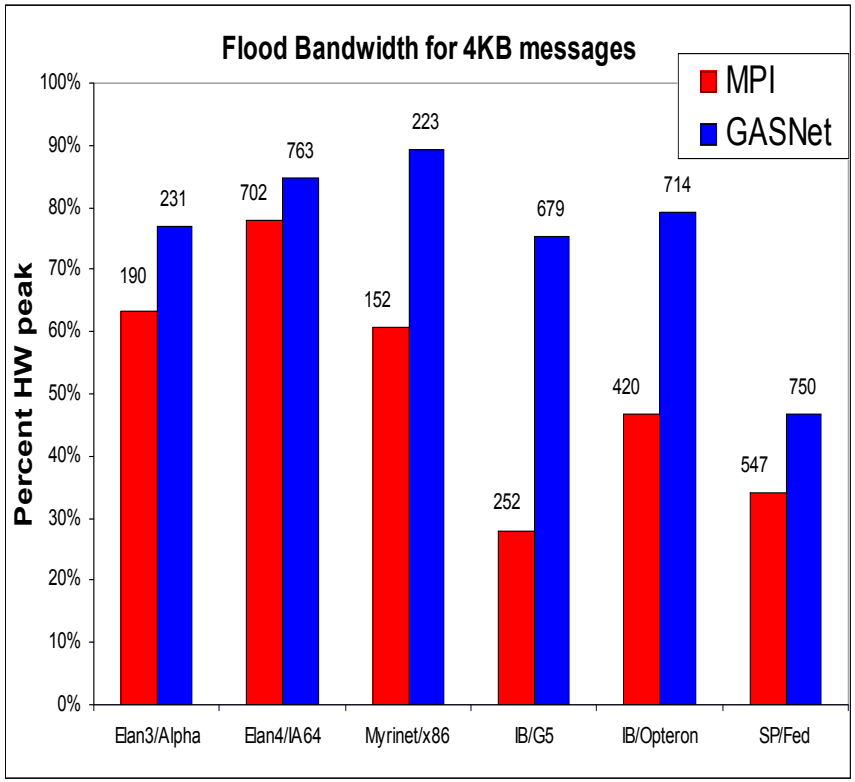


- Implementation issues:
  - Distributed memory: Reading a remote array or structure is explicit, not a cache fill
  - Shared memory: Caches are allowed, but not required
- No less scalable than MPI!
- Permits sharing, whereas MPI rules it out!

Performance Advantage of One-Sided Communication

- The put/get operations in PGAS languages (remote read/write) are one-sided (no required interaction from remote proc)
- This is faster for pure data transfers than two-sided send/receive





BERKELEY LAB

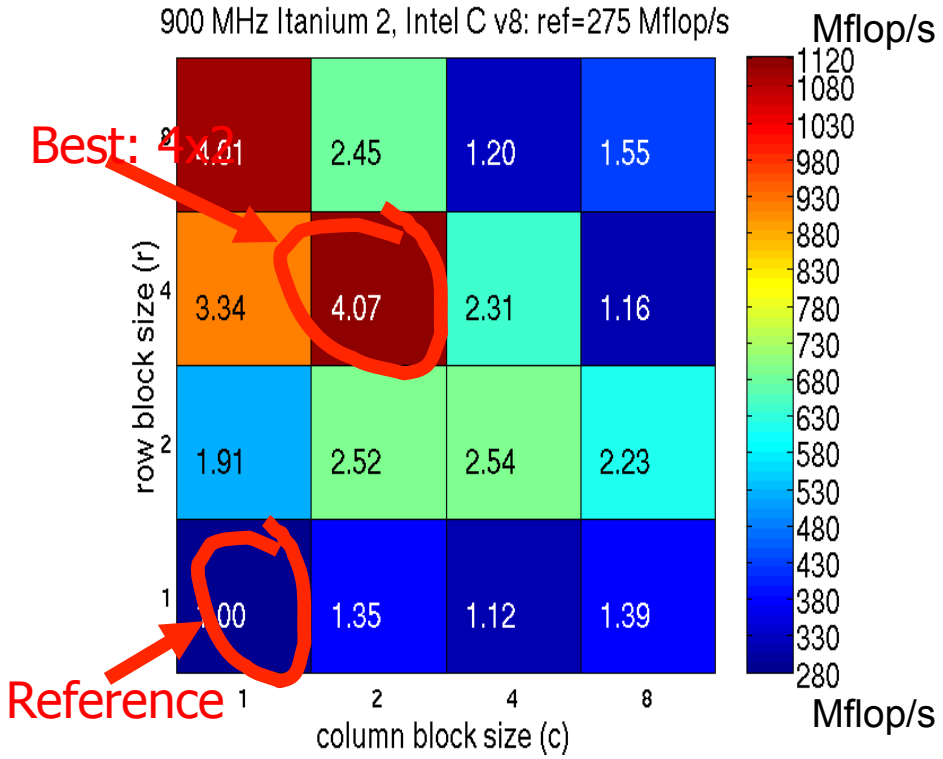
U.S. DEPARTMENT OF ENERGY
Office of Science

Autotuning

Write programs that write programs

- **Automate search across a complex optimization space**
- **Generate space of implementations, search it**
- **Performance far beyond current compilers**
- **Performance portability for diverse architectures!**
- **Past successes: PhiPAC, ATLAS, FFTW, Spiral, OSKI**

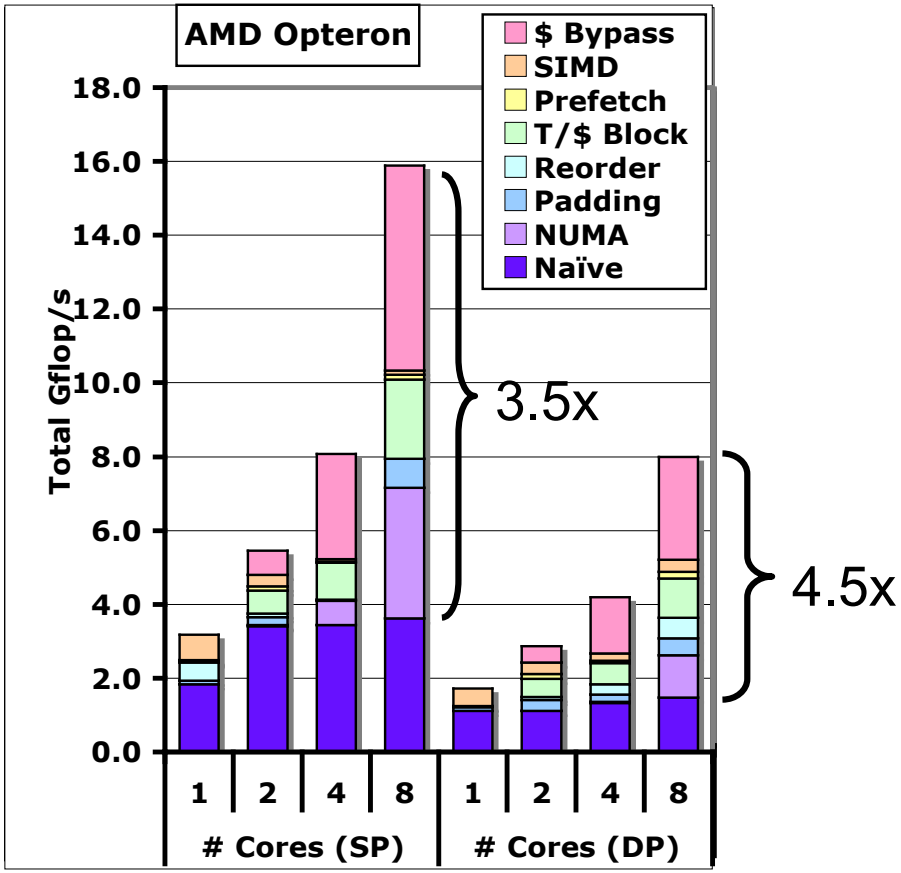
For finite element problem
[Im, Yelick, Vuduc, 2005]



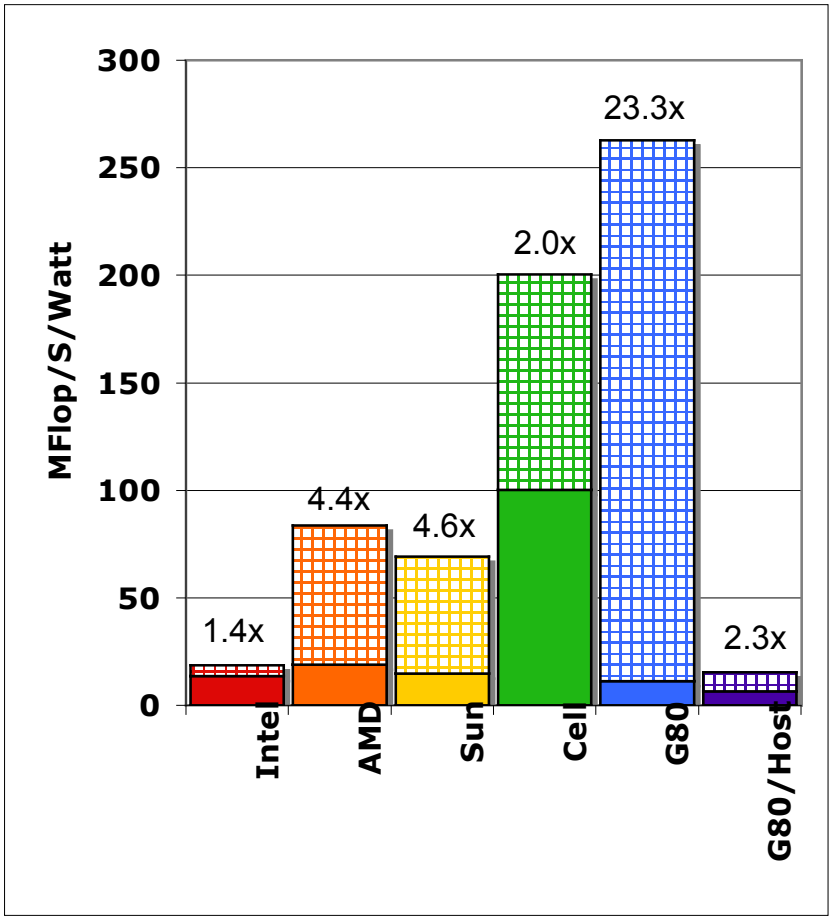
Multiprocessor Efficiency and Scaling

(auto-tuned stencil kernel; Oliker et al. , paper in IPDPS'08)

Performance Scaling



Power Efficiency

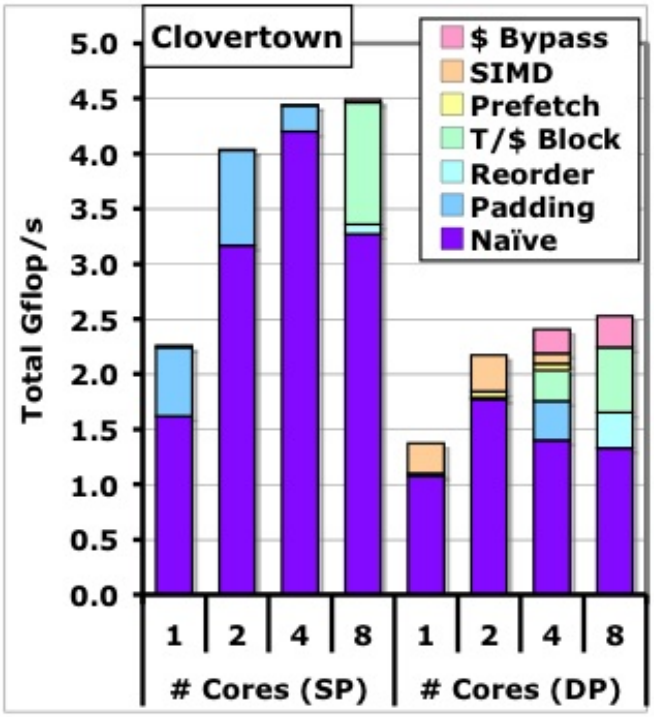


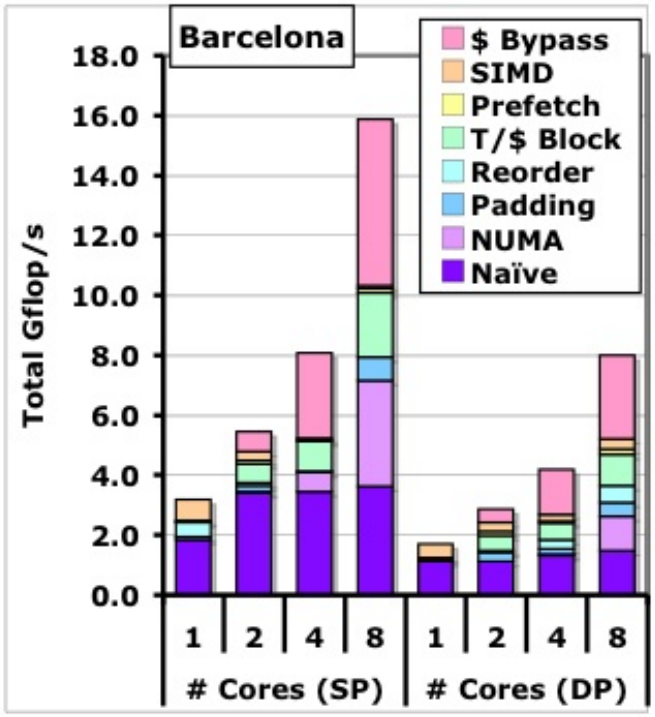
BERKELEY LAB

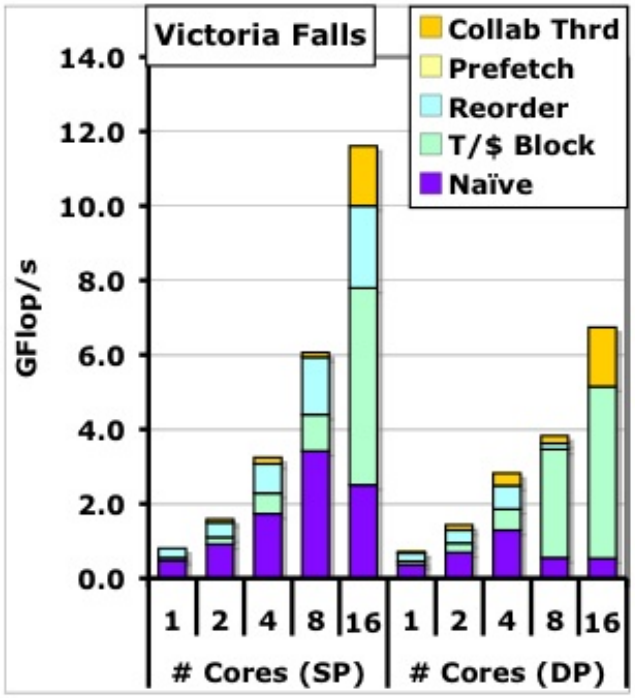
U.S. DEPARTMENT OF ENERGY

Office of Science

Autotuning for Scalability and Performance Portability





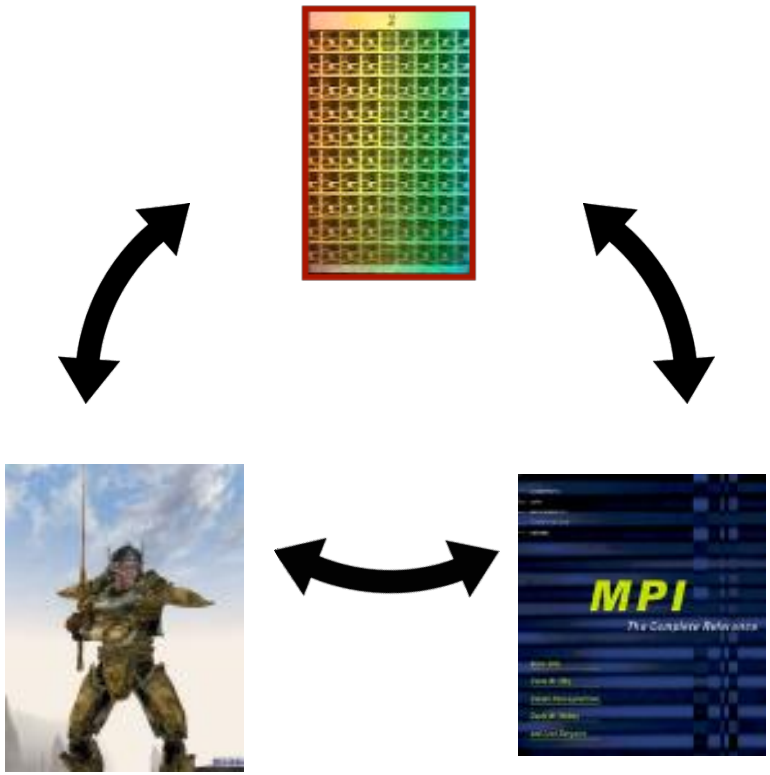


BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

The Likely HPC Ecosystem in 2014

CPU + GPU = future many-core driven by commercial applications



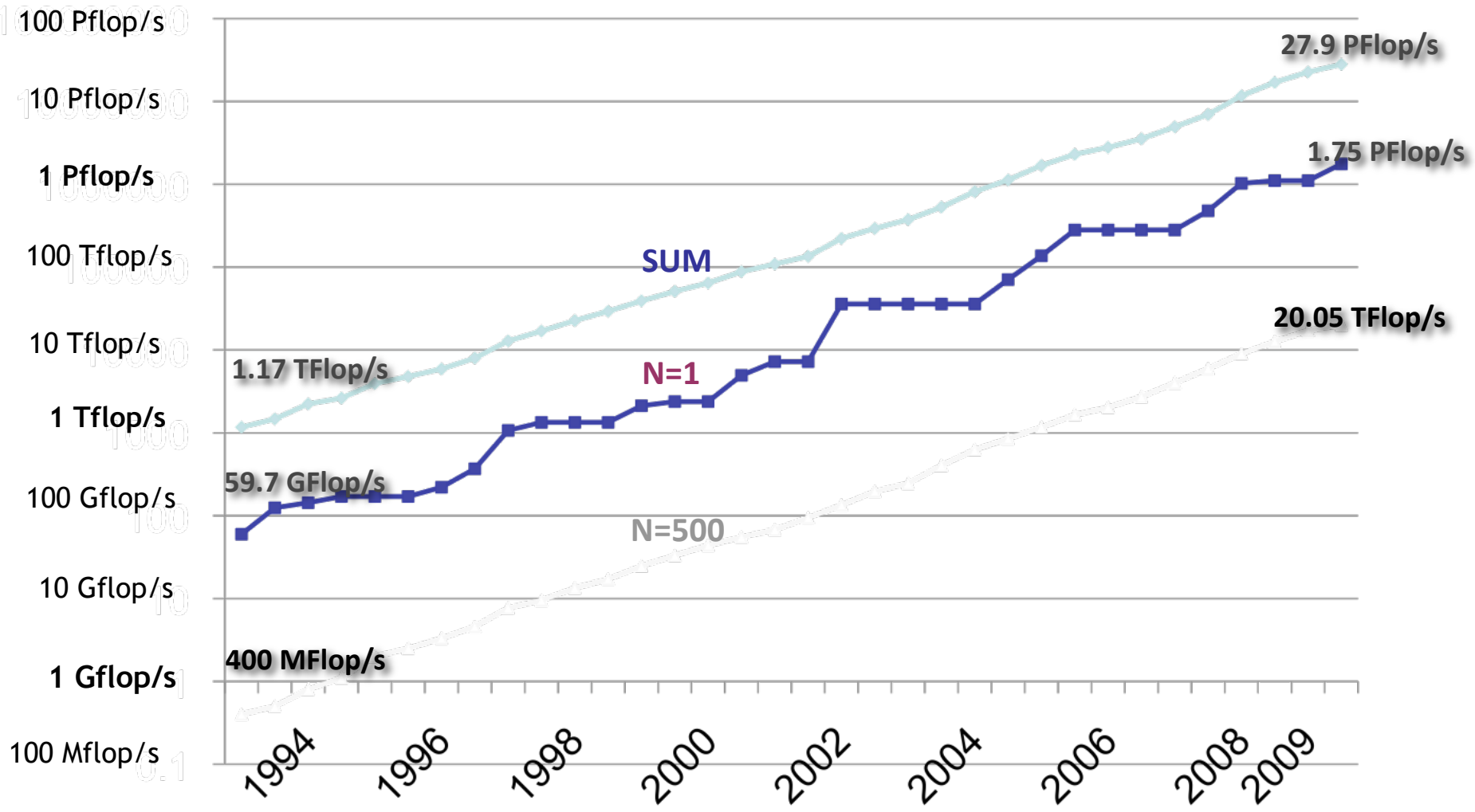
MPI+(autotuning, PGAS, ??)

Next generation “clusters” with many-core or hybrid nodes

Overview

- Turning point in 2004
- Current trends and what to expect until 2014
- Long term trends until 2019

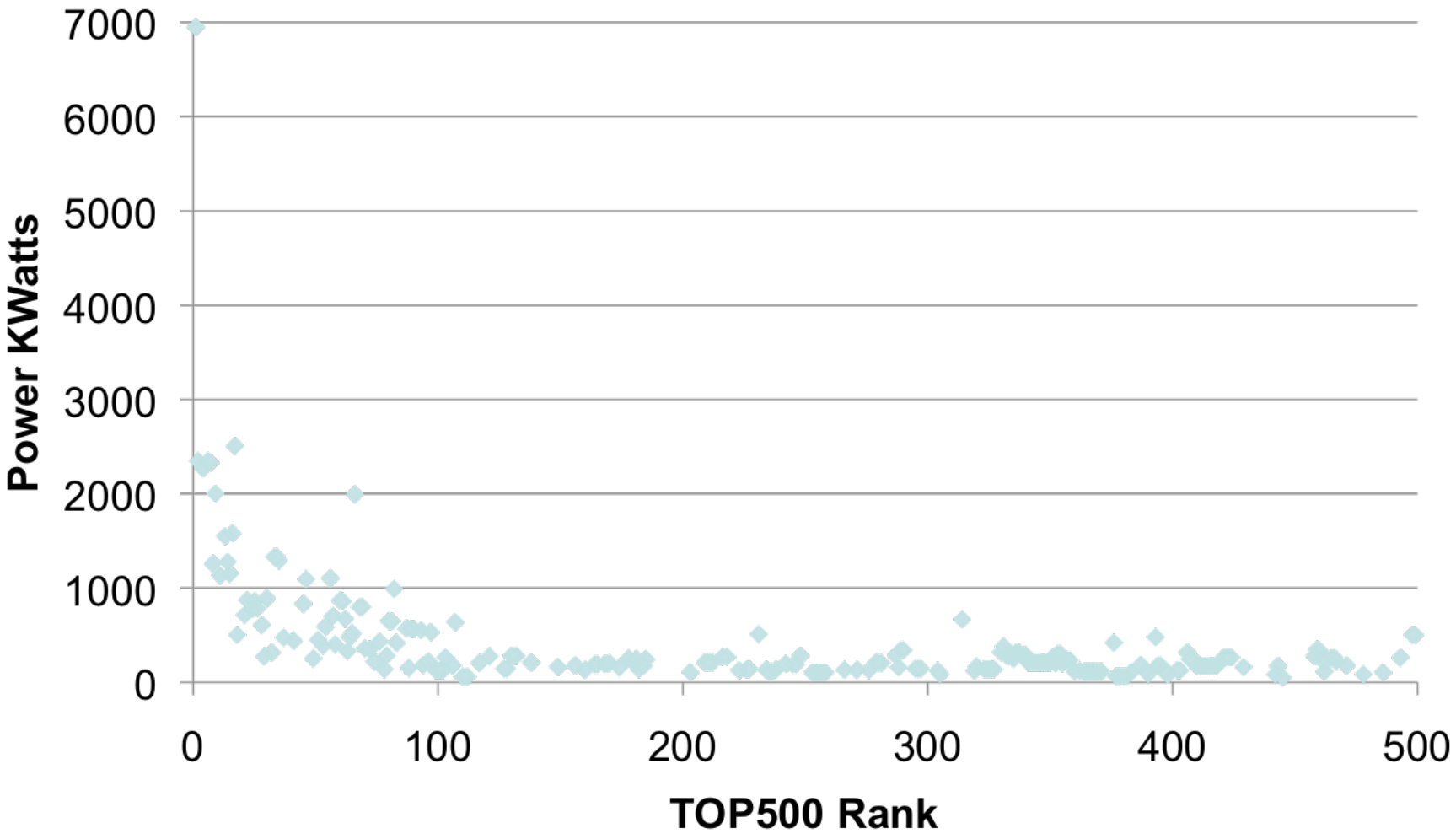
Performance Development



BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

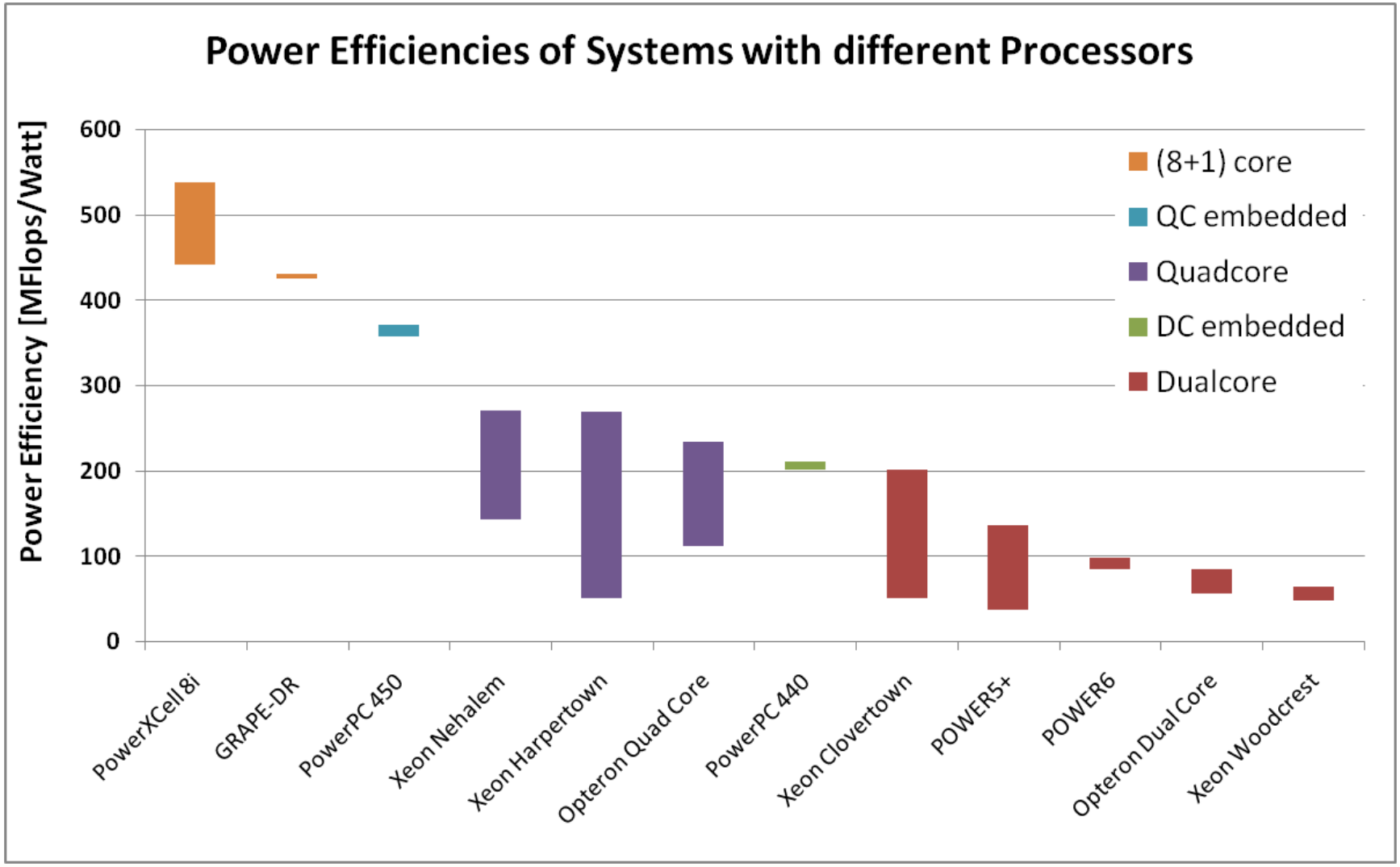
Absolute Power Levels



BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

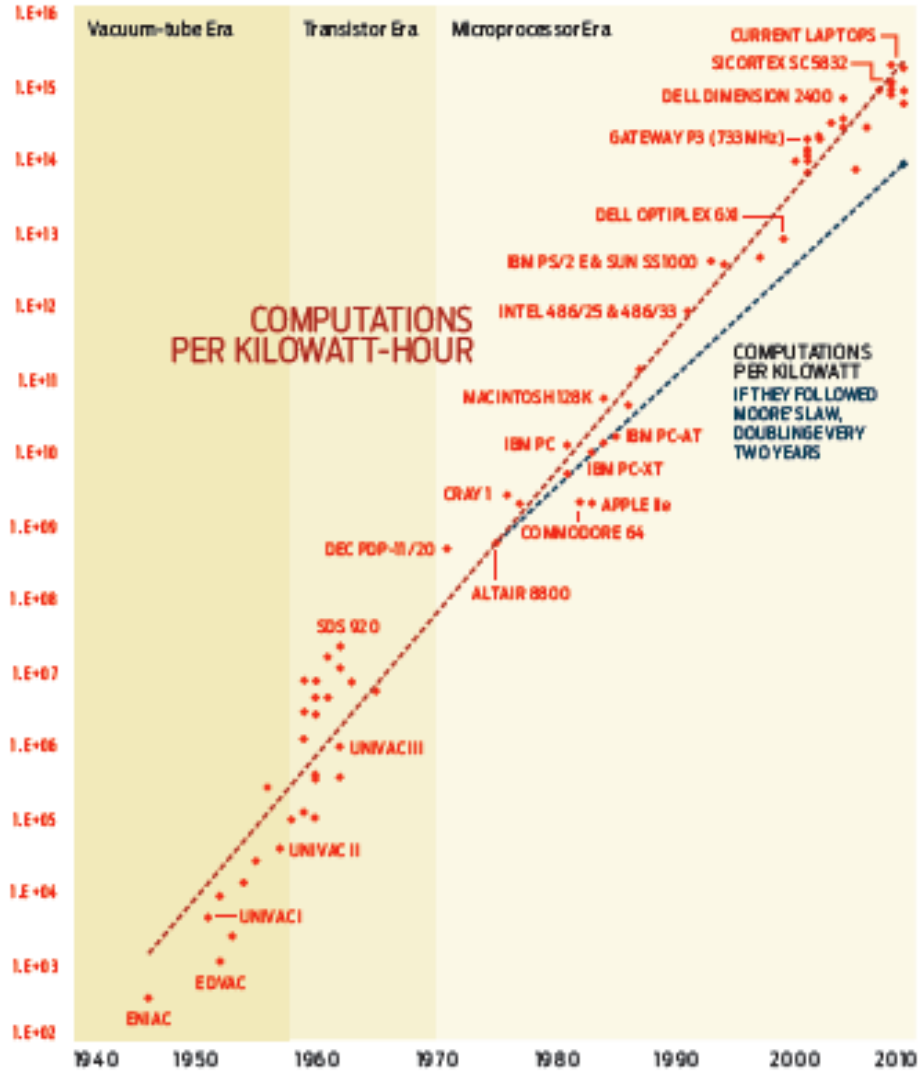
Power Efficiency related to Processors



BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

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**

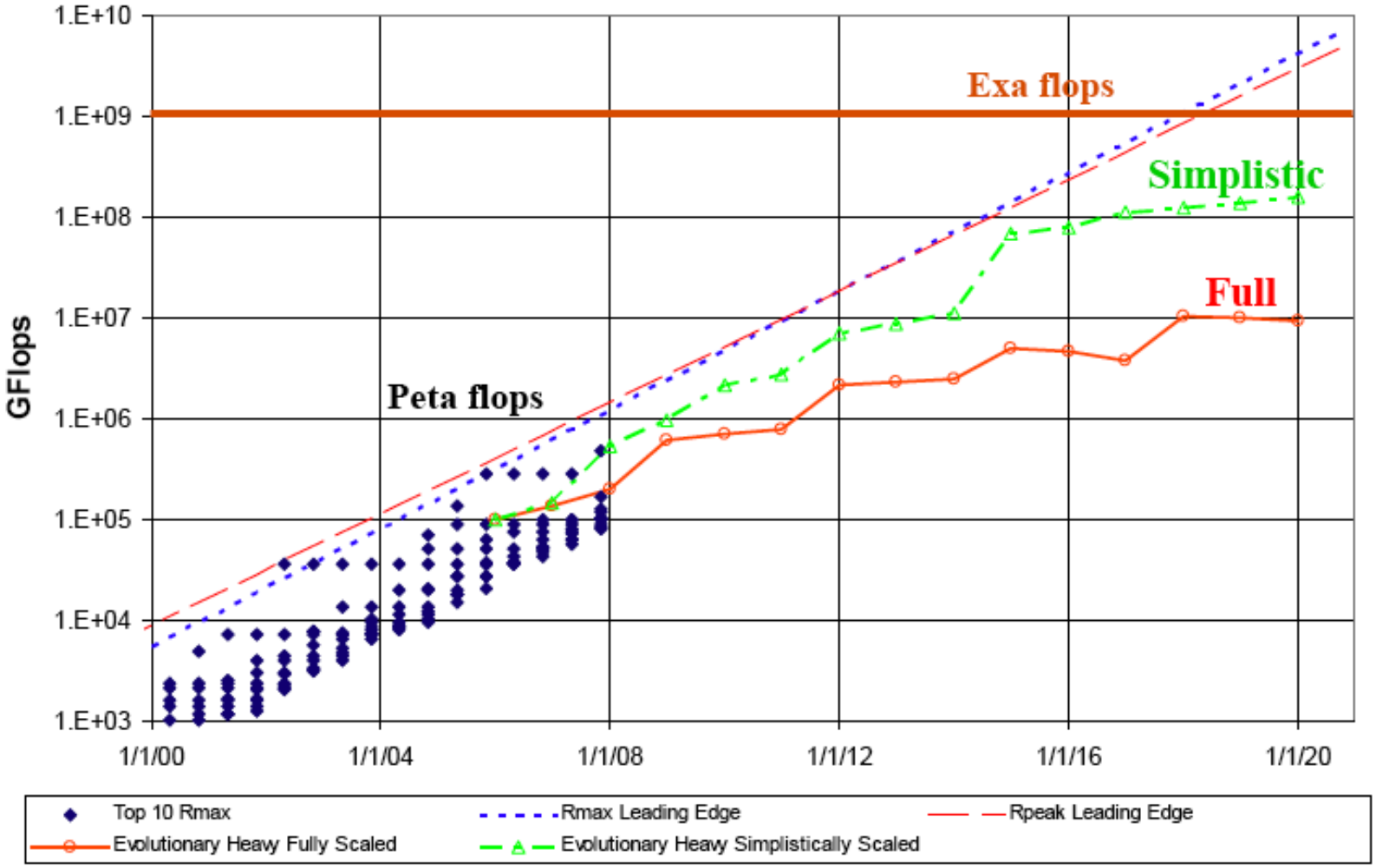
Trend Analysis

- Processors and Systems have become more energy efficient over time
  - Koomey's Law shows factor of 1.5 improvement in kWh/computations
- Supercomputers have become more powerful over time
  - TOP500 data show factor of 1.86 increase of computations/sec
- Consequently power/system increases by about 1.24 per year

DARPA Exascale Study

- Commissioned by DARPA to explore the challenges for Exaflop computing (Kogge et al.)
- Two models for future performance growth
  - Simplistic: ITRS roadmap; power for memory grows linear with # of chips; power for interconnect stays constant
  - Fully scaled: same as simplistic, but memory and router power grow with peak flops per chip

We won't reach Exaflops with this approach

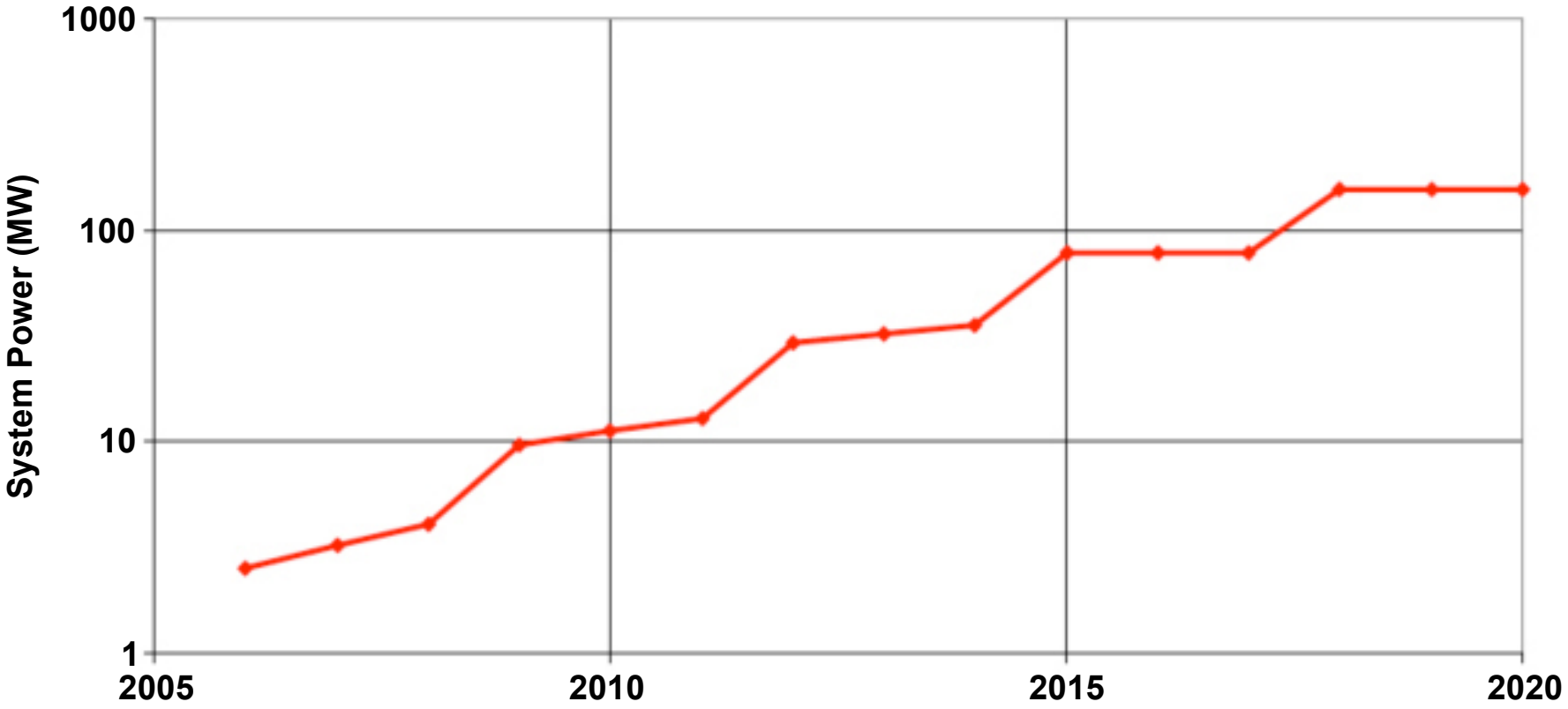


From Peter Kogge, DARPA Exascale Study

BERKELEY LAB

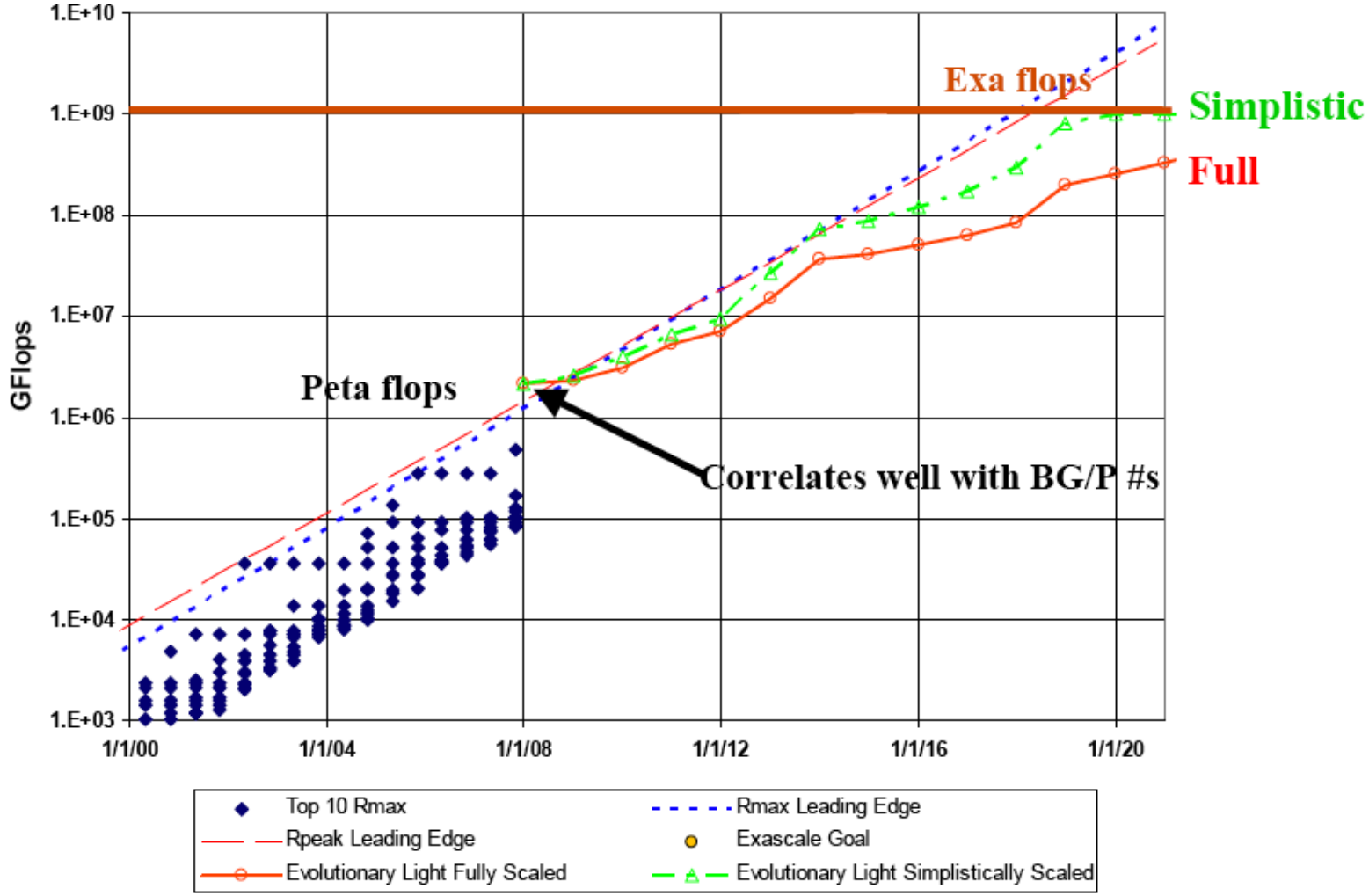
U.S. DEPARTMENT OF ENERGY
Office of Science

… and the power costs will still be staggering



From Peter Kogge,
DARPA Exascale Study

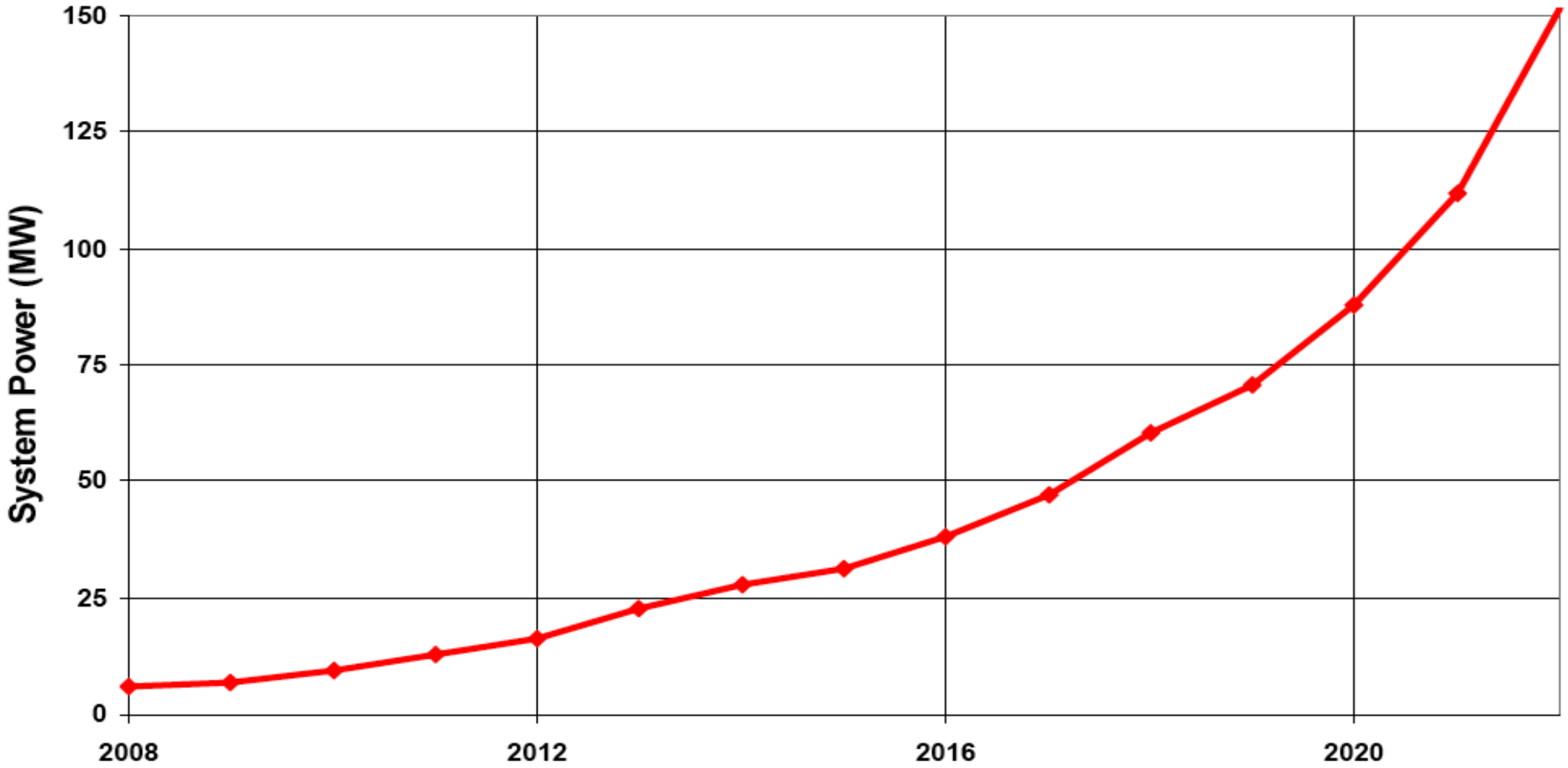
An Alternate "BG" Scenario With Similar Assumptions



BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

… and a similar, but delayed power consumption



Extrapolating to Exaflop/s in 2018

	BlueGene/L (2005)	Exaflop Directly scaled	Exaflop compromise using expected technology	Assumption for "compromise guess"
Node Peak Perf	5.6GF	20TF	20TF	Same node count (64k)
hardware concurrency/node	2	8000	1600	Assume 3.5GHz
System Power in Compute Chip	1 MW	3.5 GW	35 MW	100x improvement (very optimistic)
Link Bandwidth (Each unidirectional 3-D link)	1.4Gbps	5 Tbps	1 Tbps	Not possible to maintain bandwidth ratio.
Wires per unidirectional 3-D link	2	400 wires	80 wires	Large wire count will eliminate high density and drive links onto cables where they are 100x more expensive. Assume 20 Gbps signaling
Pins in network on node	24 pins	5,000 pins	1,000 pins	20 Gbps differential assumed. 20 Gbps over copper will be limited to 12 inches. Will need optics for in rack interconnects. 10Gbps now possible in both copper and optics.
Power in network	100 KW	20 MW	4 MW	10 mW/Gbps assumed. Now: 25 mW/Gbps for long distance (greater than 2 feet on copper) for both ends one direction. 45mW/Gbps optics both ends one direction. + 15mW/Gbps of electrical Electrical power in future: separately optimized links for power.
Memory Bandwidth/node	5.6GB/s	20TB/s	1 TB/s	Not possible to maintain external bandwidth/Flop
L2 cache/node	4 MB	16 GB	500 MB	About 6-7 technology generations
Data pins associated with memory/node	128 data pins	40,000 pins	2000 pins	3.2 Gbps per pin
Power in memory I/O (not DRAM)	12.8 KW	80 MW	4 MW	10 mW/Gbps assumed. Most current power in address bus. Future probably about 15mW/Gbps maybe get to 10mW/Gbps (2.5mW/Gbps is c*v^2*f for random data on data pins) Address power is higher.
QCD CG single iteration time	2.3 msec	11 usec	15 usec	Requires: 1) fast global sum (2 per iteration) 2) hardware offload for messaging (Driverless messaging)

Source: David Turek, IBM

U.S. DEPARTMENT OF ENERGY

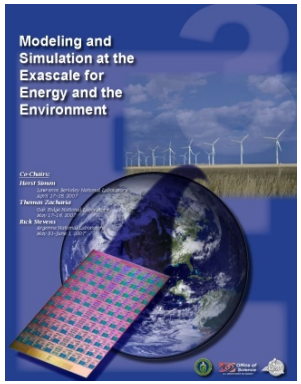
A decadal DOE plan for providing exascale applications and technologies for DOE mission needs

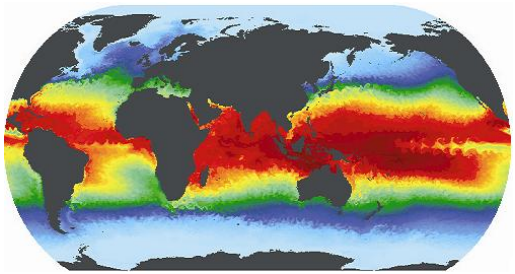
Rick Stevens and Andy White, co-chairs

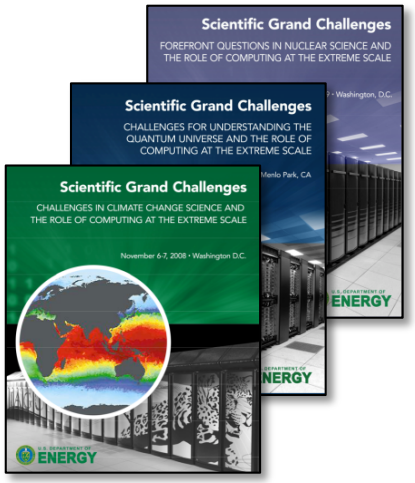
Pete Beckman, Ray Bair-ANL; Jim Hack, Jeff Nichols, Al Geist-ORNL; Horst Simon, Kathy Yelick, John Shalf-LBNL; Steve Ashby, Moe Khaleel-PNNL; Michel McCoy, Mark Seager, Brent Gorda-LLNL; John Morrison, Cheryl Wampler-LANL; James Peery, Sudip Dosanjh, Jim Ang-SNL; Jim Davenport, Tom Schlagel, BNL; Fred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

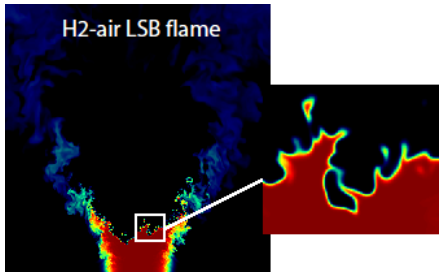
Process for identifying exascale applications and technology for DOE missions ensures broad community input

- 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

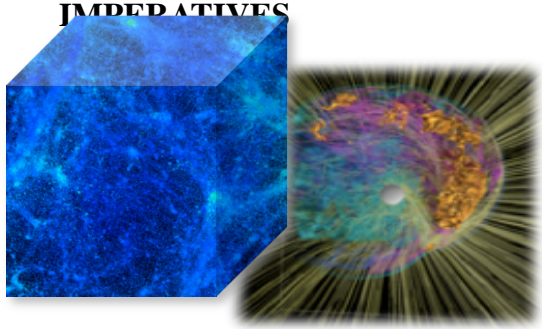








MISSION IMPERATIVES



FUNDAMENTAL SCIENCE

U.S. DEPARTMENT OF ENERGY

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)	

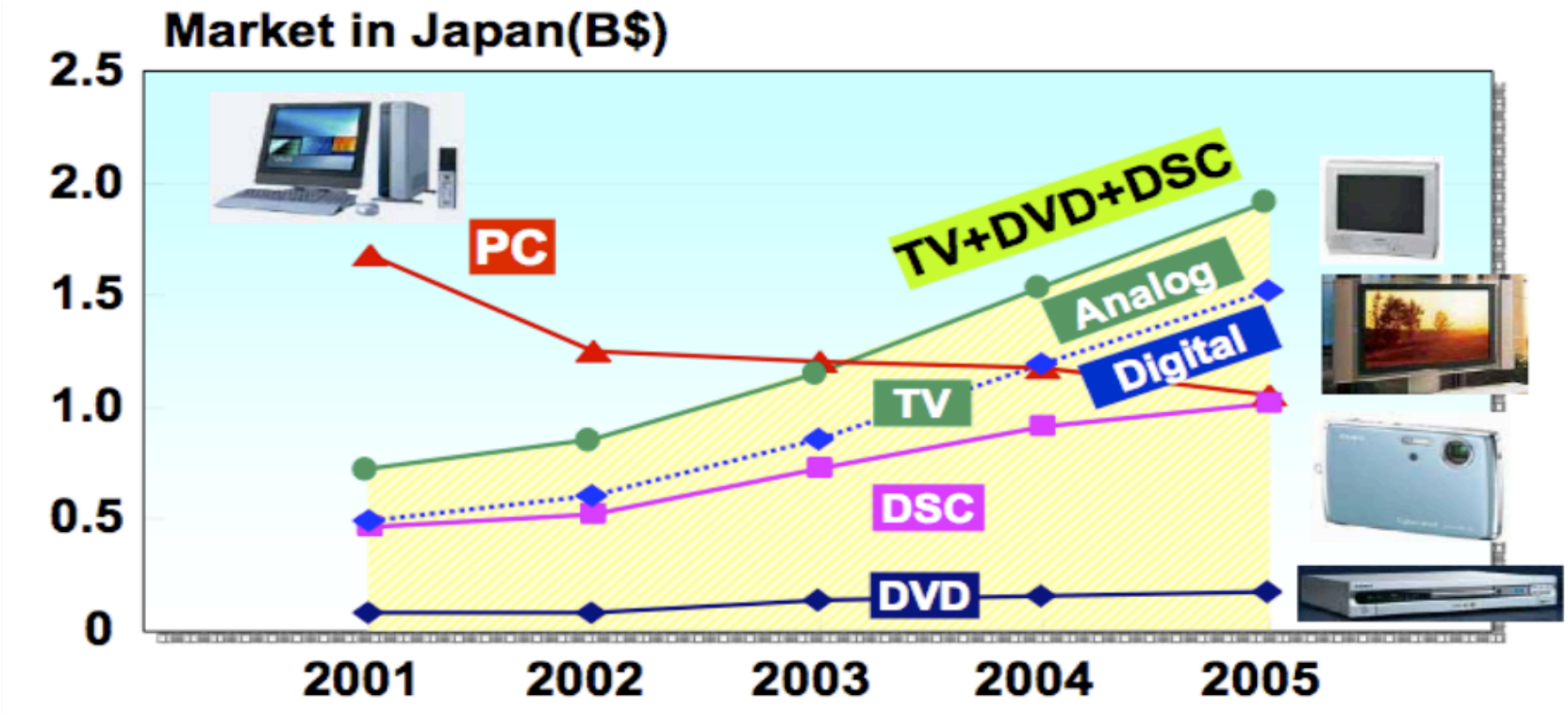
U.S. DEPARTMENT OF ENERGY

What are critical exascale technology investments?

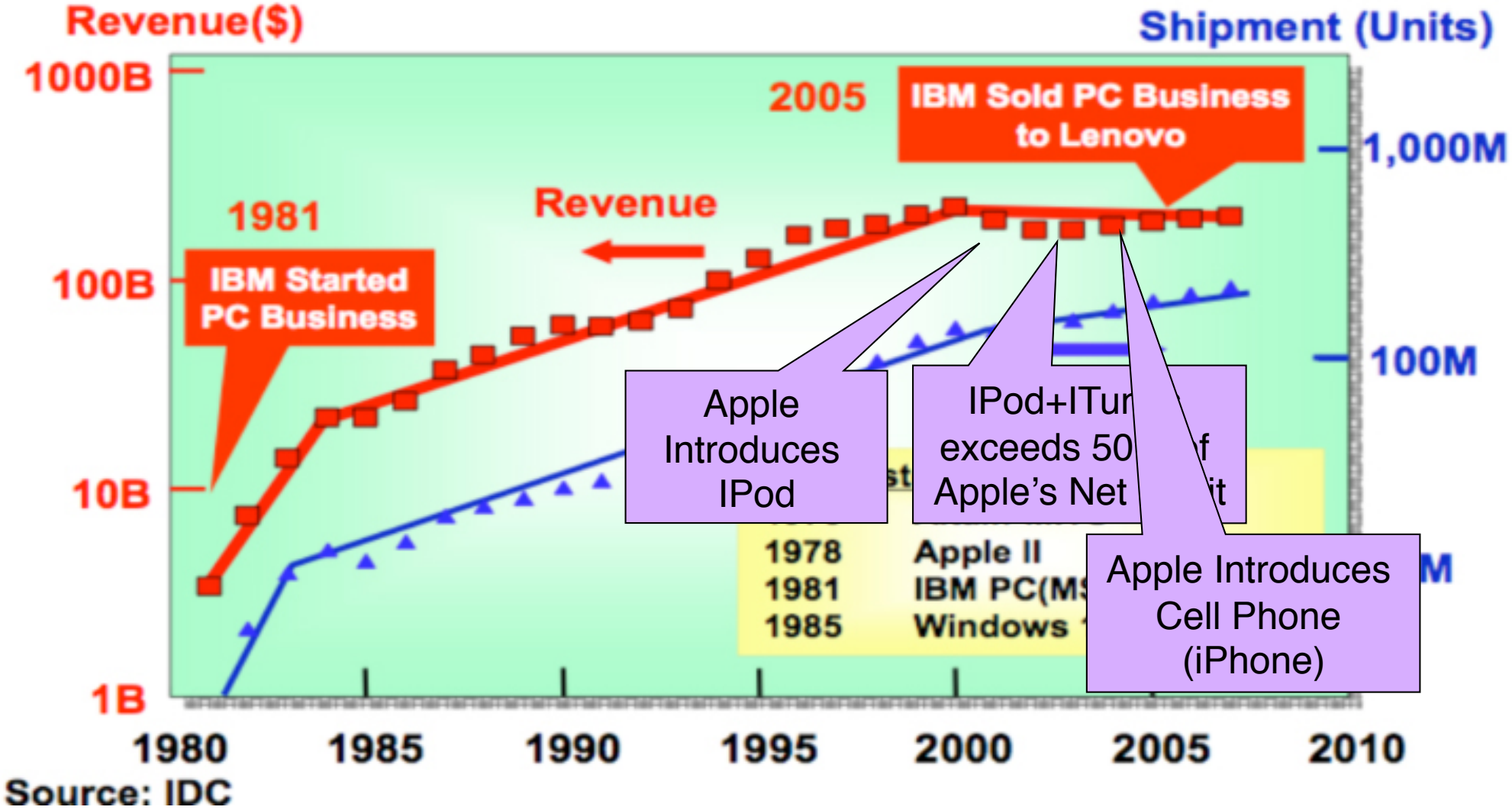
- **System power** is a first class constraint on exascale system performance and effectiveness.
- M**emory** is an important component of meeting exascale power and applications goals.
- **Programming model**. Early investment in several efforts to decide in 2013 on exascale programming model, allowing 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.*

Processor Technology Trend

- 1990s - R&D computing hardware dominated by desktop/COTS
  - Had to learn how to use COTS technology for HPC
- 2010 - R&D investments moving rapidly to consumer electronics/ embedded processing
  - Must learn how to leverage embedded processor technology for future HPC systems



Consumer Electronics has Replaced PCs as the Dominant Market Force in CPU Design!!



Berkeley Lab

U.S. Department of Energy
Office of Science

Green Flash:
Ultra-Efficient Climate Modeling

- **Project by Shalf, Oliker, Wehner and others at LBNL**
- **An alternative route to exascale computing**
  - **Target specific machine designs to answer a scientific question**
  - **Use of new technologies driven by the consumer market.**

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

Impact of Cloud Simulation

The effect of clouds in current global climate models are parameterized, not directly simulated.

Currently cloud systems are much smaller than model grid cells (unresolved).

Clouds affect both solar and terrestrial radiation, control precipitation.

Poor simulated cloud distribution impacts global moisture budget.

Several important climate features are poorly simulated including:

- Inter-tropical convergence zone (ITCZ)
- Madden-Julian Oscillation (MJO)
- Underestimation of low marine stratus clouds
- Errors in precipitation patterns, especially monsoons.

Global Cloud System Resolving Climate Modeling

Individual cloud physics fairly well understood

Parameterization of mesoscale cloud statistics performs poorly.

Direct simulation of cloud systems in global models requires exascale

Direct simulation of cloud systems replacing statistical parameterization.

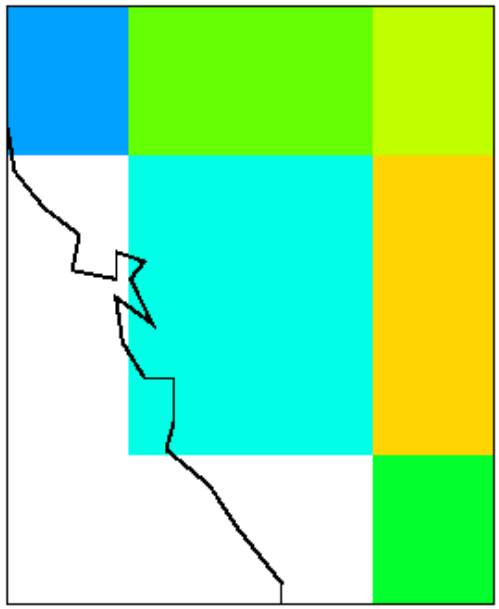
Approach recently was called a top priority by the 1st UN WMO Modeling Summit.

BERKELEY LAB

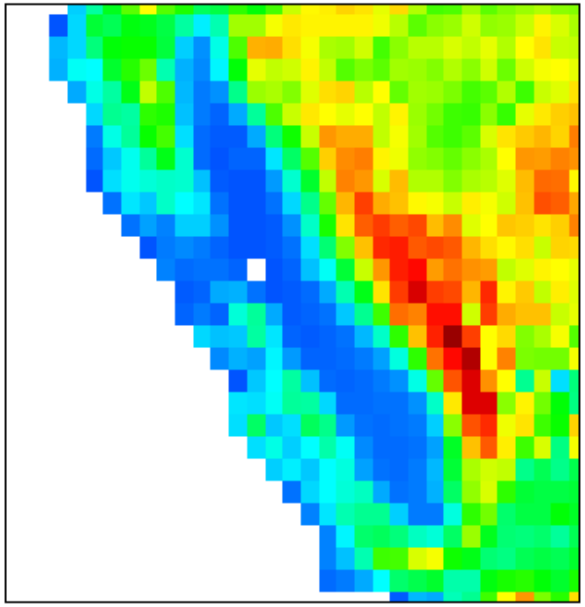
U.S. DEPARTMENT OF ENERGY
Office of Science

Global Cloud System Resolving Models

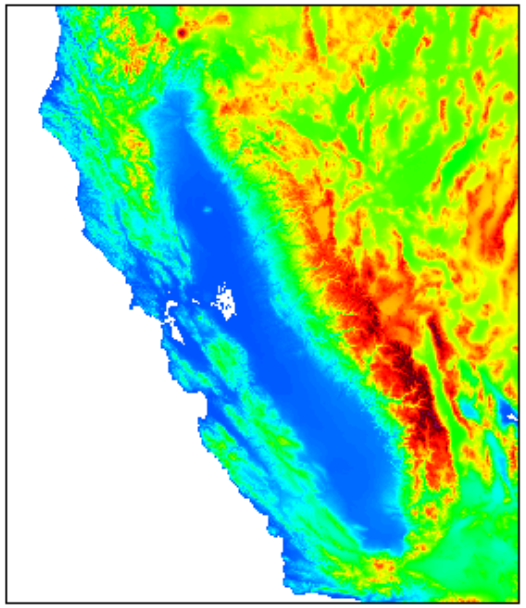
Surface Altitude (feet)



200km
Typical resolution of IPCC AR4 models



25km
Upper limit of climate models with cloud parameterizations



1km
Cloud system resolving models enable *transformational change* in quality of simulation results

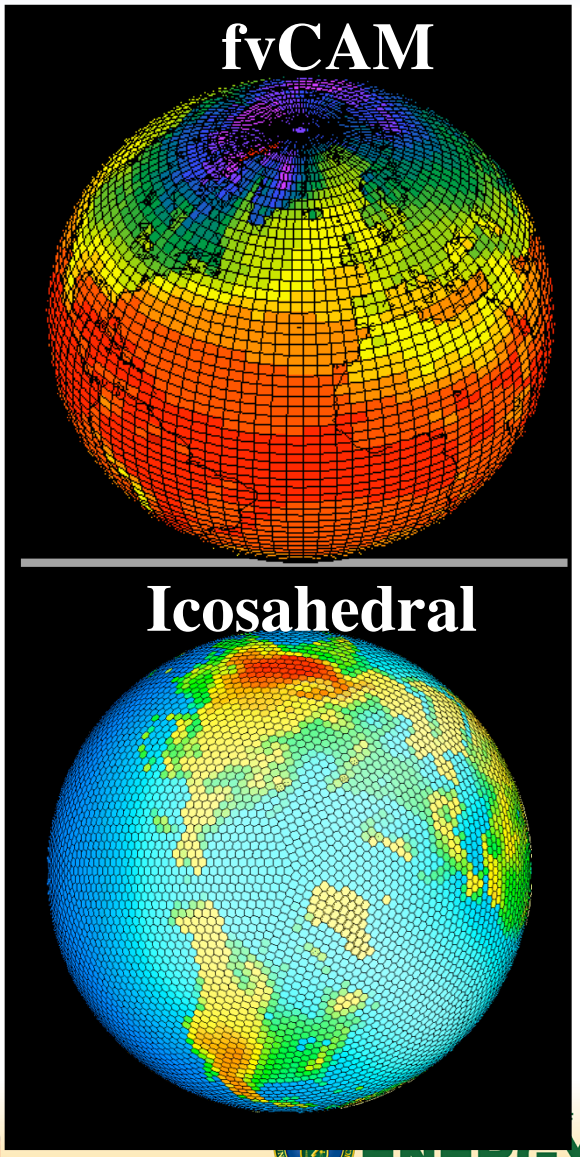
BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

Computational Requirements

Computational Requirements for 1km Global Cloud System Resolving Model, based on David Randall's (CSU) icosahedral code:

- Approximately 1,000,000x more computation than current production models
- Must achieve 1000x faster than realtime to be useful for climate studies
- 10 PetaFlops sustained, ~200PF peak
- ExaFlop(s) for required ensemble runs
- 20-billion subdomains
- *Minimum* 20-million way parallelism
- Only 5MB memory requirement per core
- 200 MB/s in 4 nearest neighbor directions
- Dominated by eqn of motion due to CFL



Green Flash Strawman System Design

We examined three different approaches (in 2008 technology)

Computation .015°X.02°X100L: 10 PFlops sustained, ~200 PFlops peak

- AMD Opteron: Commodity approach, lower efficiency for scientific codes offset by cost efficiencies of mass market. Constrained by legacy/binary compatibility.
- BlueGene: Generic embedded processor core and customize system-on-chip (SoC) to improve power efficiency for scientific applications
- Tensilica XTensa: Customized embedded CPU w/SoC provides further power efficiency benefits but maintains programmability

Processor	Clock	Peak/ Core (Gflops)	Cores/ Socket	Sockets	Cores	Power	Cost 2008
AMD Opteron	2.8GHz	5.6	2	890K	1.7M	179 MW	$1B+
IBM BG/P	850MHz	3.4	4	740K	3.0M	20 MW	$1B+
Green Flash / Tensilica XTensa	650MHz	2.7	32	120K	4.0M	3 MW	$75M

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

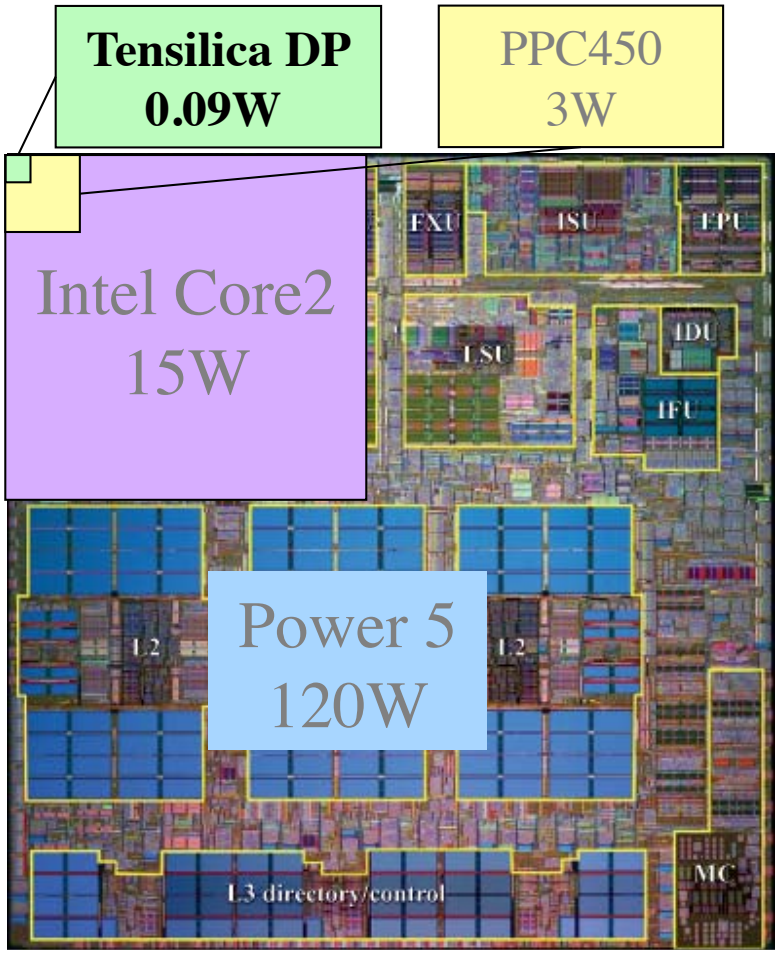
Green Flash: Ultra-Efficient Climate Modeling

- We present an alternative route to exascale computing
  - Exascale science questions are already identified.
  - Our idea is to target specific machine designs to each of these questions.
    - This is possible because of new technologies driven by the consumer market.
- We want to turn the process around.
  - Ask “What machine do we need to answer a question?”
  - Not “What can we answer with that machine?”
- Caveat:
  - We present here a feasibility design study.
  - Goal is to influence the HPC industry by evaluating a prototype design.

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

Design for Low Power: More Concurrency



- Cubic power improvement with lower clock rate due to V^2F
- Slower clock rates enable use of simpler cores
- Simpler cores use less area (lower leakage) and reduce cost
- Tailor design to application to reduce waste

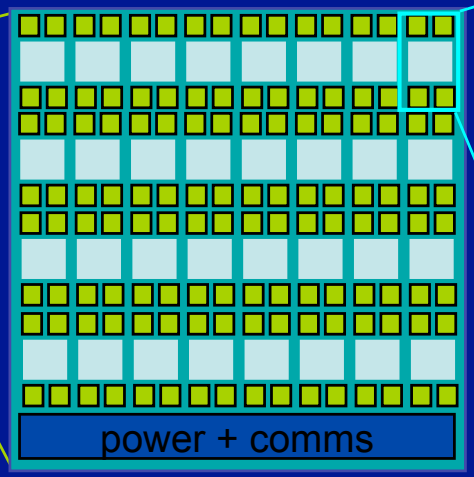
This is how iPhones and MP3 players are designed to maximize battery life and minimize cost

Climate System Design Concept

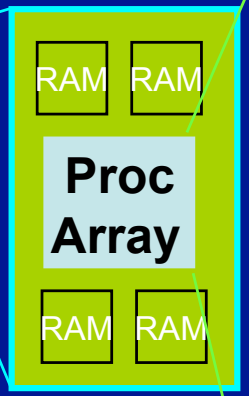
Strawman Design Study

32 boards per rack

100 racks @ ~25KW



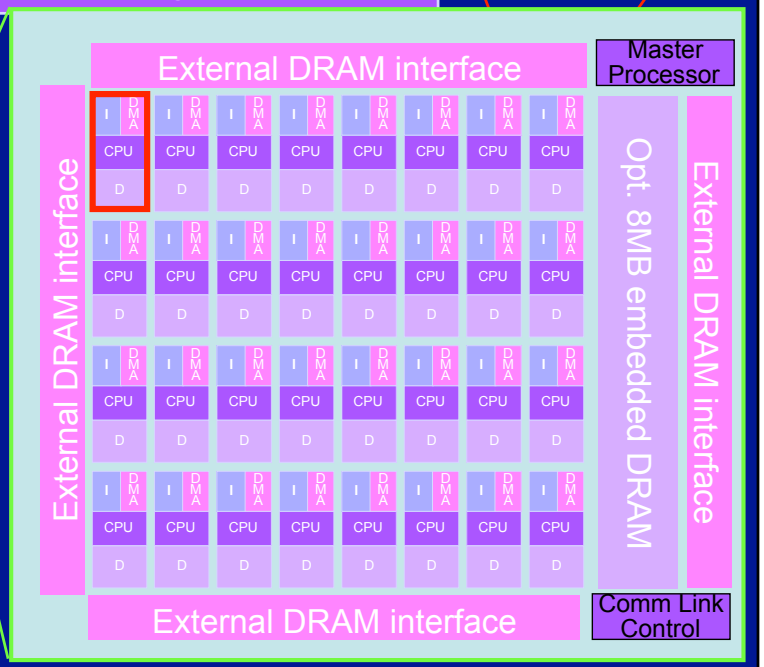
32 chip + memory clusters per board (2.7 TFLOPS @ 700W



8 DRAM per processor chip: ~50 GB/s

VLIW CPU:

- 128b load-store + 2 DP MUL/ADD + integer op/ DMA per cycle:
- Synthesizable at 650MHz in commodity 65nm
- 1mm² core, 1.8-2.8mm² with inst cache, data cache data RAM, DMA interface, 0.25mW/MHz
- Double precision SIMD FP : 4 ops/cycle (2.7GFLOPs)
- Vectorizing compiler, cycle-accurate simulator, debugger GUI (Existing part of Tensilica Tool Set)
- 8 channel DMA for streaming from on/off chip DRAM
- Nearest neighbor 2D communications grid



32 processors per 65nm chip
83 GFLOPS @ 7W

Summary on Green Flash

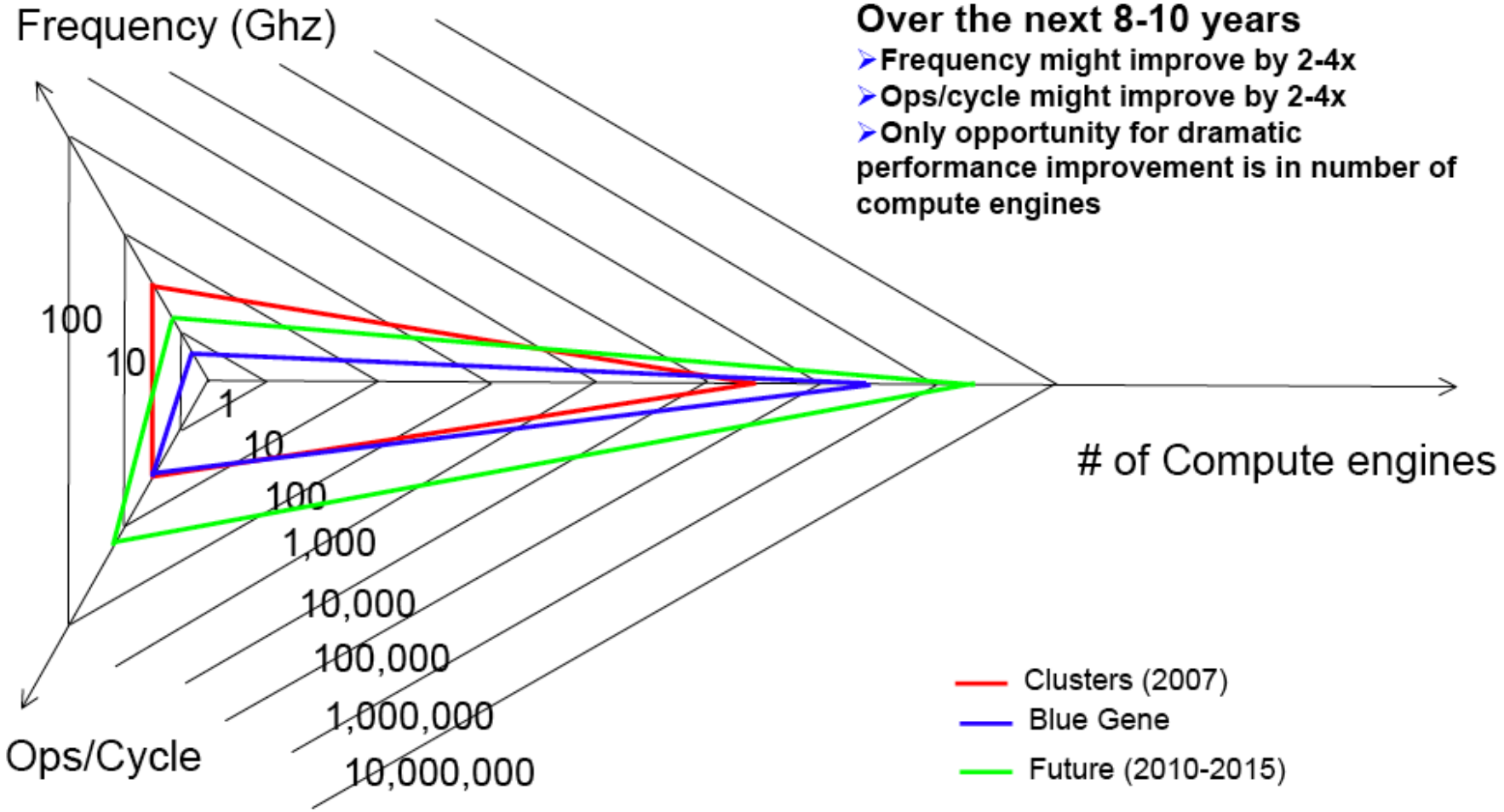
- Exascale computing is vital for numerous key scientific areas
- We propose a new approach to high-end computing that enables transformational changes for science
- Research effort: study feasibility and share insight w/ community
- This effort will augment high-end general purpose HPC systems
  - Choose the science target first *(climate in this case)*
  - Design systems for applications *(rather than the reverse)*
  - Leverage power efficient embedded technology
  - Design hardware, software, scientific algorithms together using hardware emulation and auto-tuning
  - Achieve exascale computing sooner and more efficiently

Applicable to broad range of exascale-class applications

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 a unique opportunity for science applications to lead extreme scale systems development**

Performance Improvement Trend



Over the next 8-10 years

- **Frequency might improve by 2-4x**
- **Ops/cycle might improve by 2-4x**
- **Only opportunity for dramatic performance improvement is in number of compute engines**

Source: David Turek, IBM

1 million cores ?

- What are applications developers concerned about?
- … but before we answer this question, the more interesting question is …

1000 cores on the laptop ?

- What are commercial applications developers going to do with it?

More Info

- The Berkeley View/Parlab
  - http://view.eecs.berkeley.edu
- NERSC Science Driven System Architecture Group
  - http://www.nersc.gov/projects/SDSA
- Green Flash Climate Computer
  - http://www.lbl.gov/cs/html/greenflash.html
- LS3DF
  - https://hpcrdm.lbl.gov/mailman/listinfo/ls3df