Future Trends in Computing

Horst Simon
Lawrence Berkeley National Laboratory
and UC Berkeley

BERAC Meeting
September 1, 2009

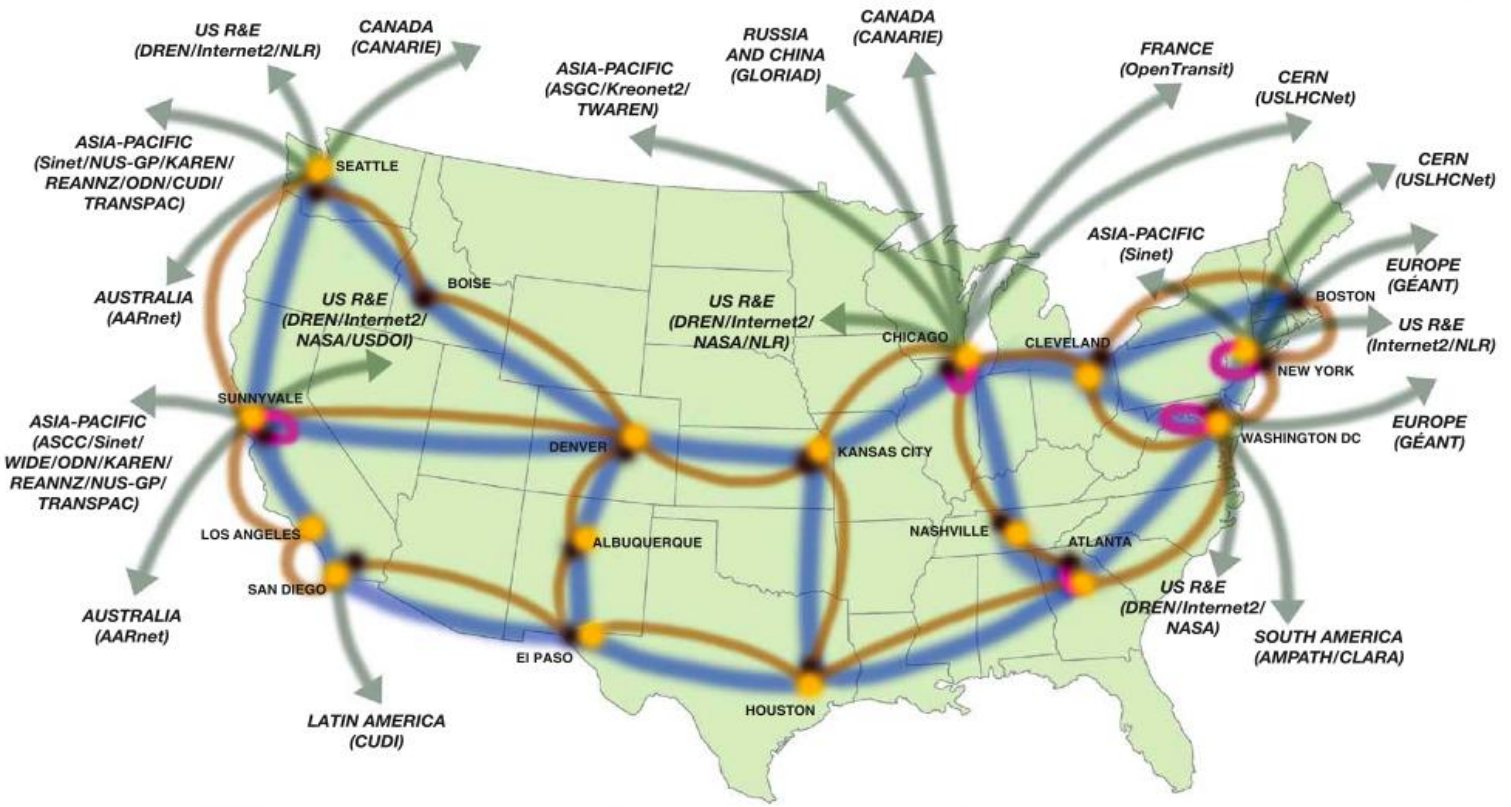
BERKELEY LAB

NERSC

U.S. DEPARTMENT OF ENERGY
Office of Science

ESnet4

Connecting the U.S. Department of Energy Laboratories to the world of science.



Core networks: scalable to 50-60 Gbps by 2009-2010, 200-600 Gbps by 2011-2012

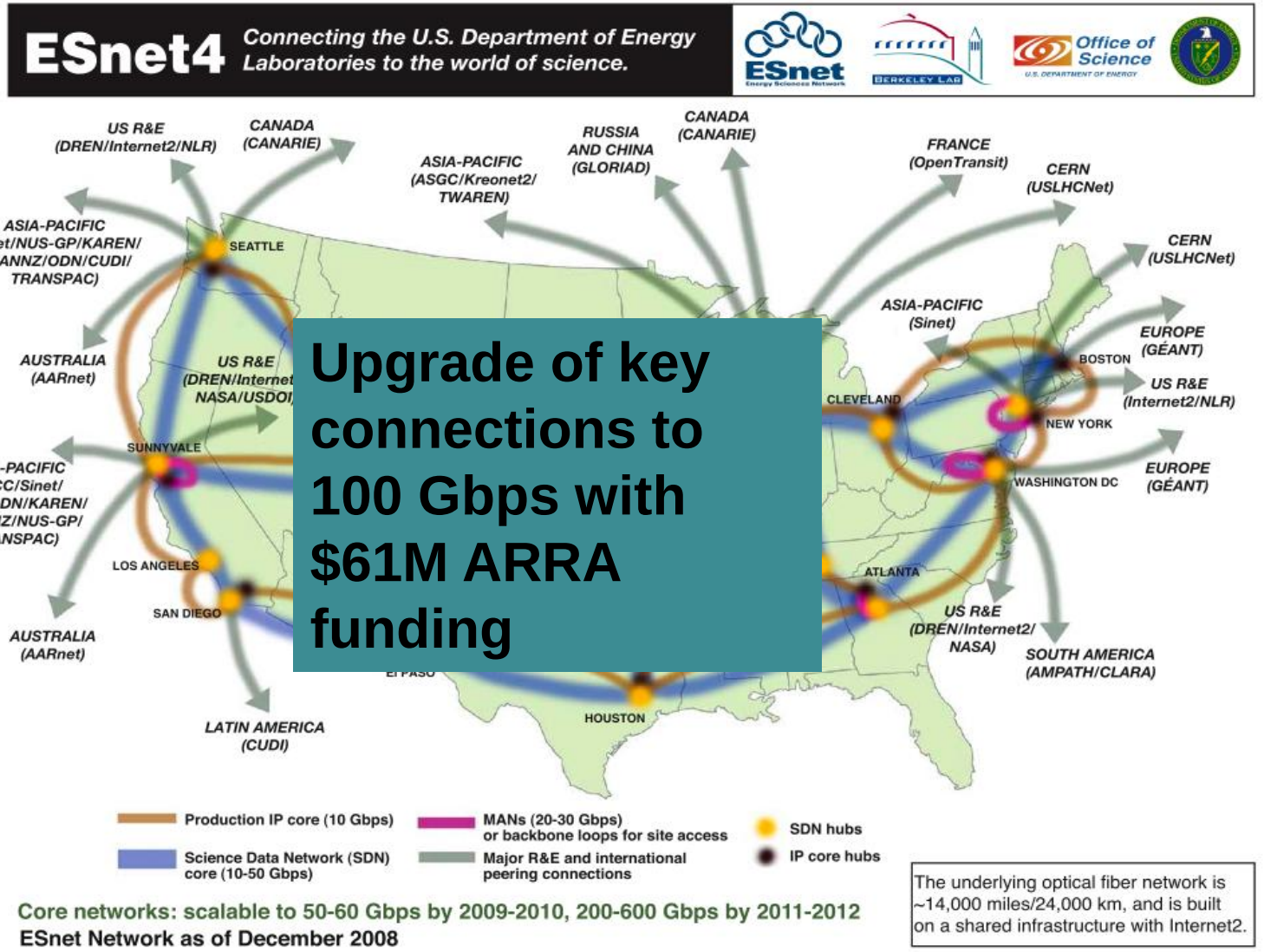
ESnet Network as of December 2008

The underlying optical fiber network is ~14,000 miles/24,000 km, and is built on a shared infrastructure with Internet2.

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY

Office of Science



Upgrade of key connections to 100 Gbps with $61M ARRA funding

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

Key Message

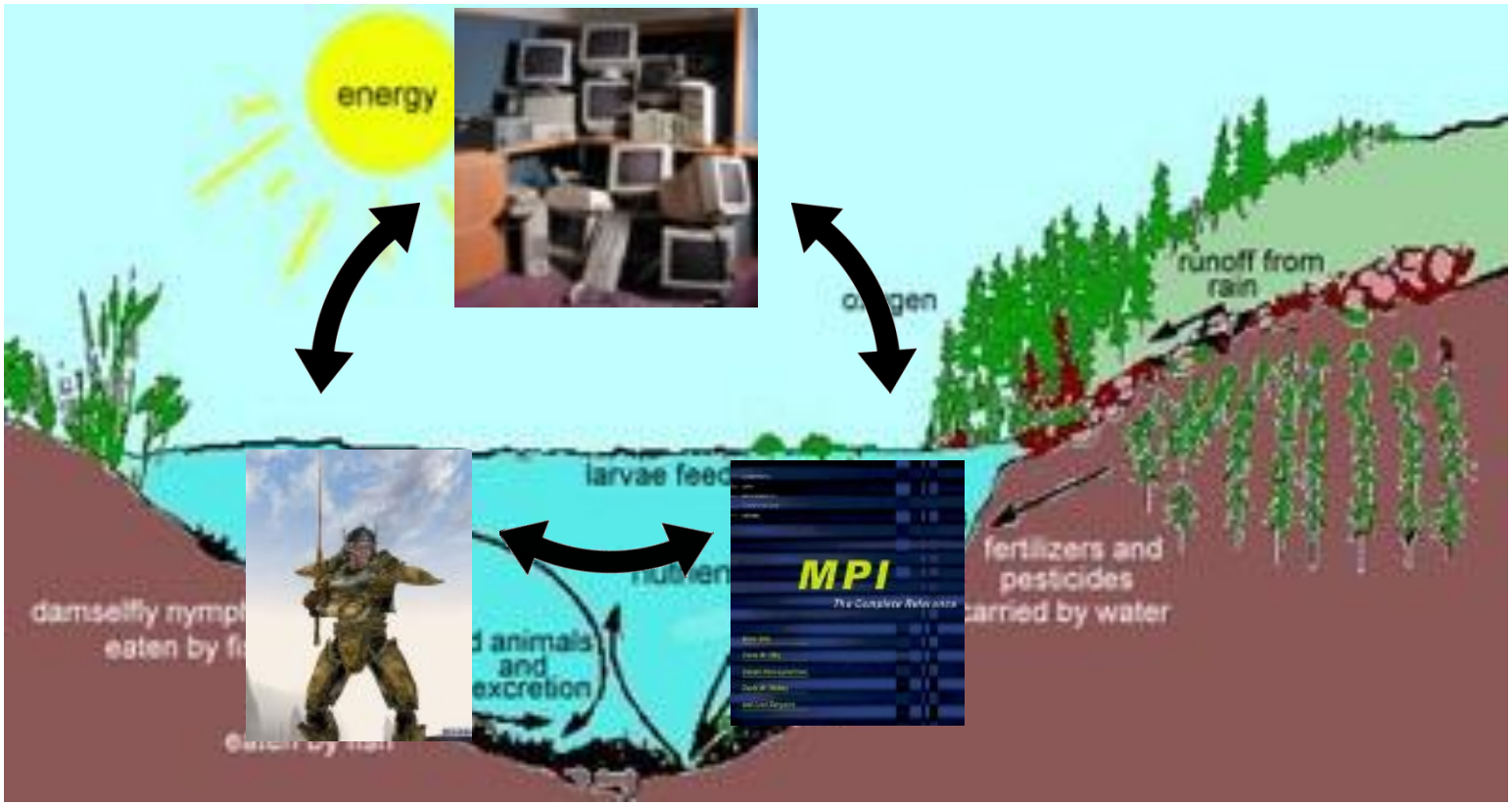
Computing is changing more rapidly than ever before, and scientists have the unprecedented opportunity to change computing directions

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!**

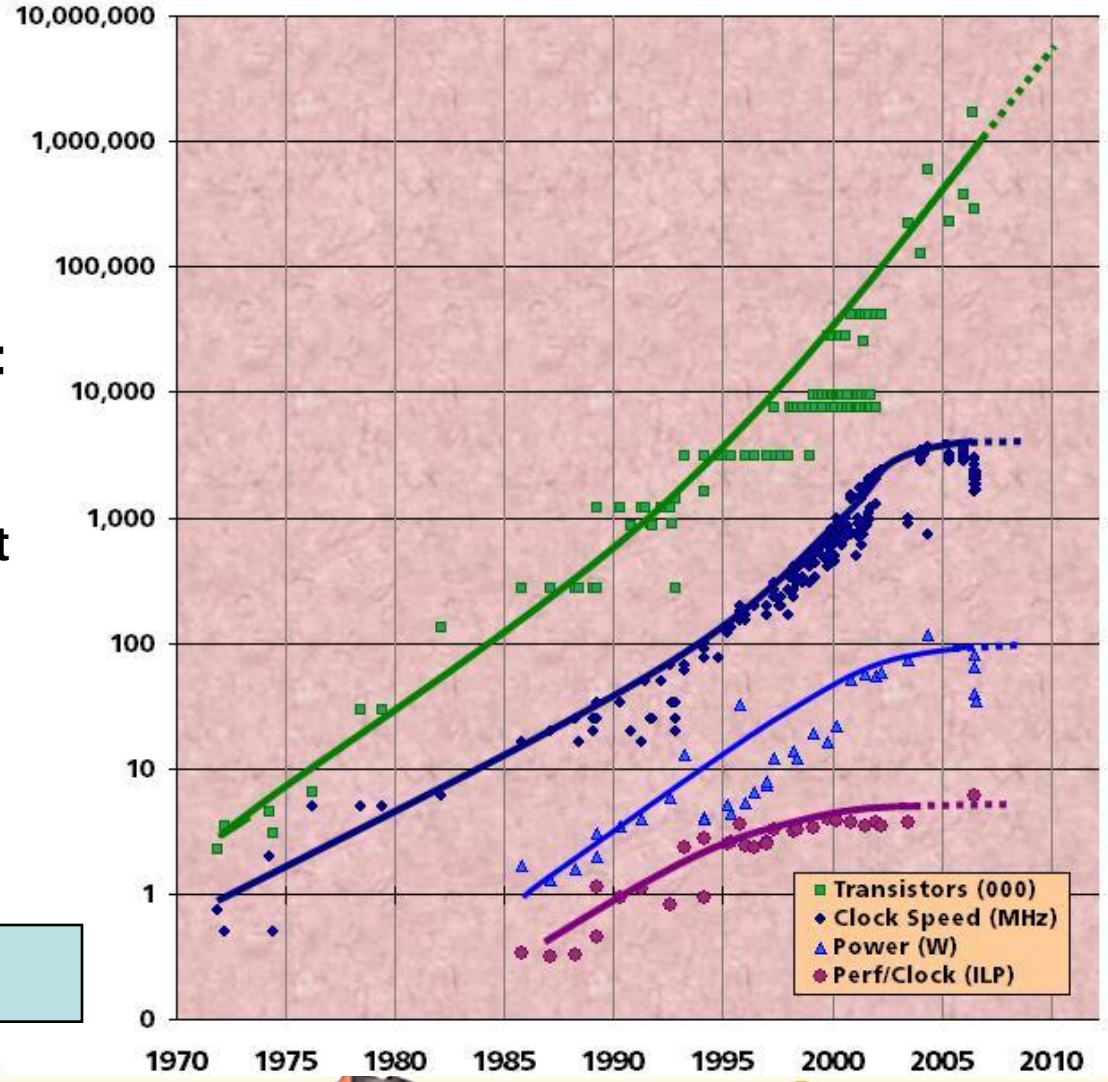


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Supercomputing Ecosystem (~~2005~~ 2009)

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**

Roadrunner Breaks the Pflop/s Barrier

- 1,026 Tflop/s on LINPACK reported on June 9, 2008
- 6,948 dual core Opteron + 12,960 cell BE
- 80 TByte of memory
- IBM built, installed at LANL

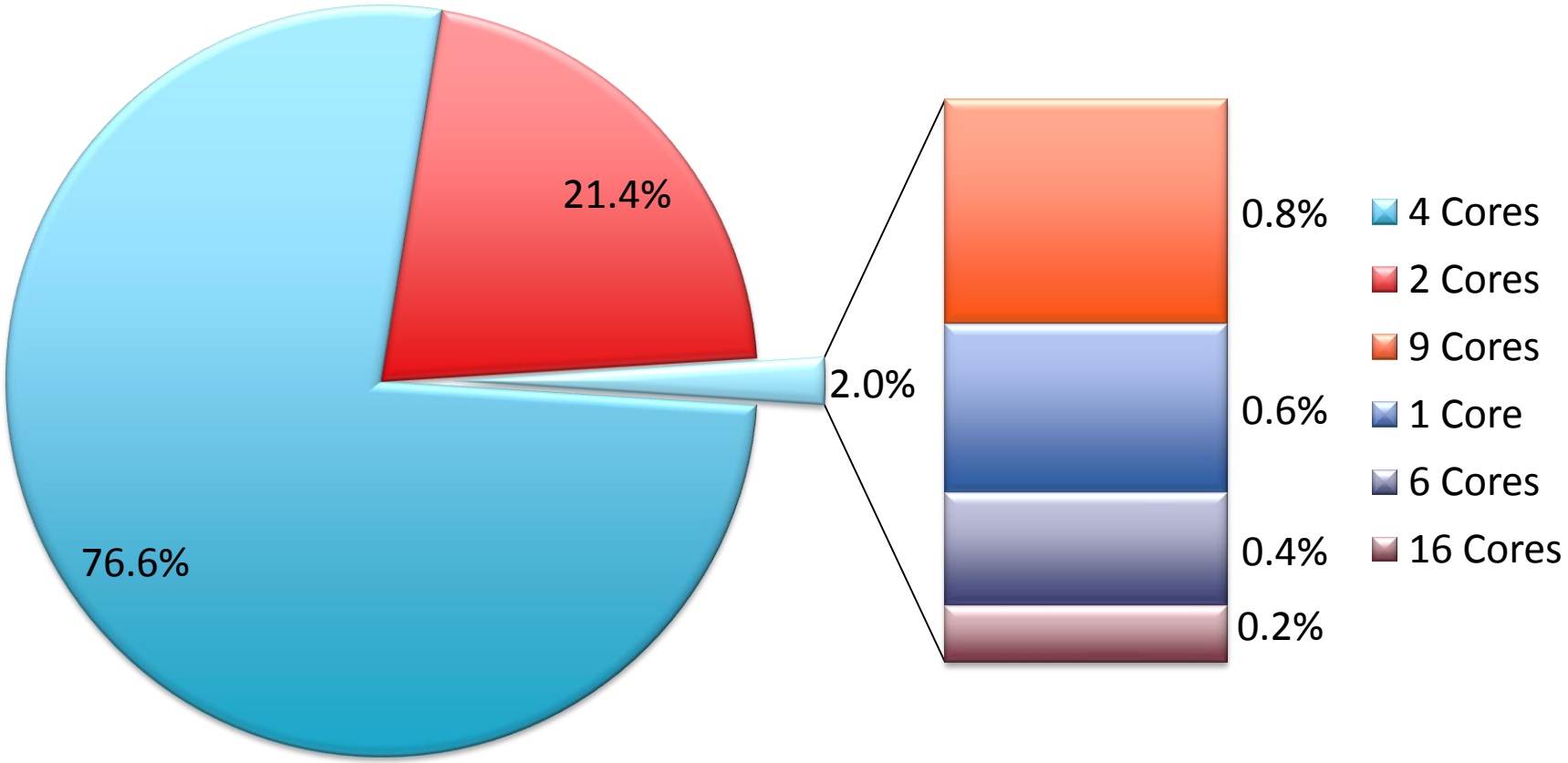
Cray XT5 at ORNL -- 1 Pflop/s in November 2008

Jaguar	Total	XT5	XT4
Peak Performance	1,645	1,382	263
AMD Opteron Cores	**181,504**	**150,176**	**31,328**
System Memory (TB)	**362**	**300**	**62**
Disk Bandwidth (GB/s)	**284**	**240**	**44**
Disk Space (TB)	**10,750**	**10,000**	**750**
Interconnect Bandwidth (TB/s)	**532**	**374**	**157**

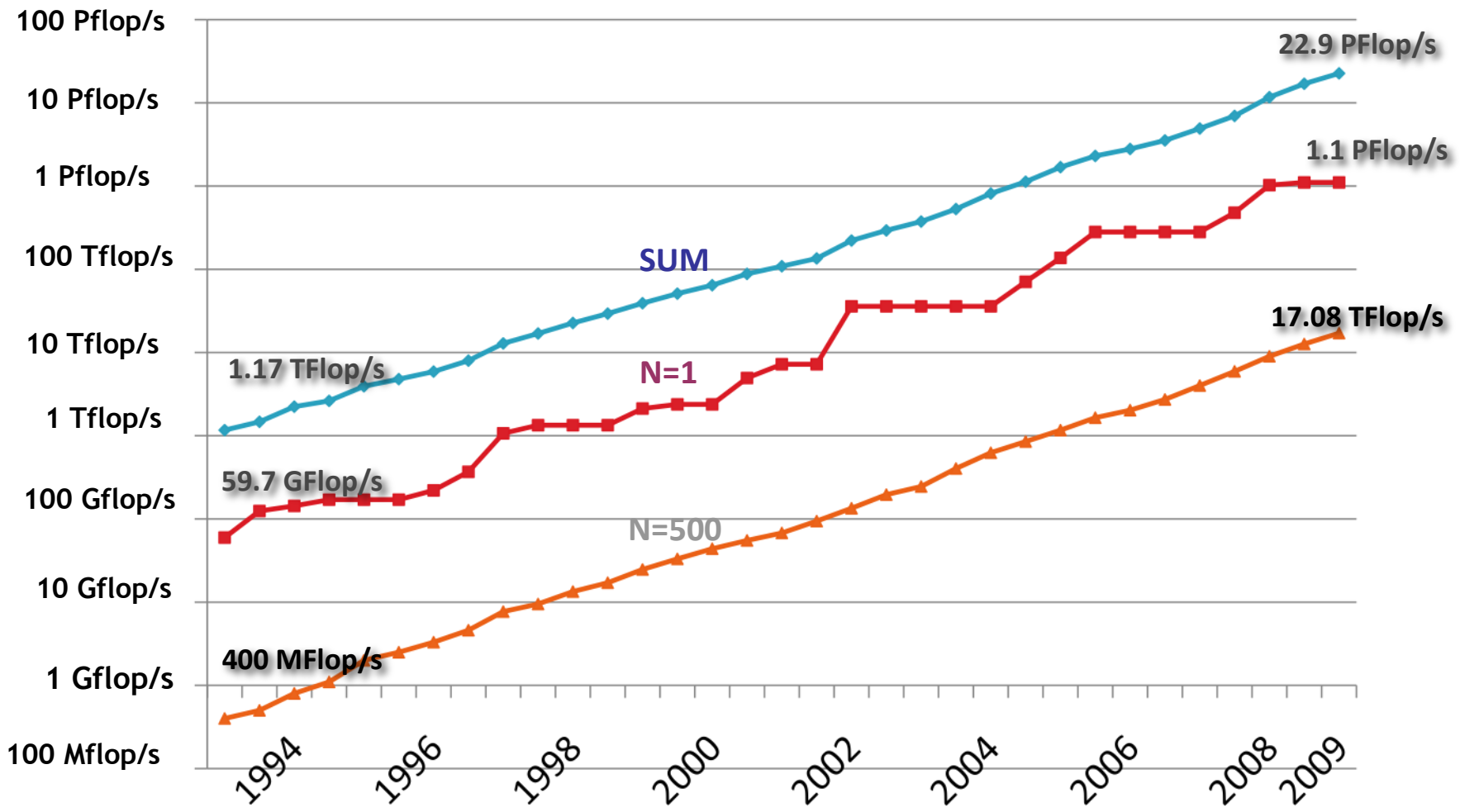
The systems will be combined after acceptance of the new XT5 upgrade. Each system will be linked to the file system through 4x-DDR Infiniband

U.S. DEPARTMENT OF ENERGY
Office of Science

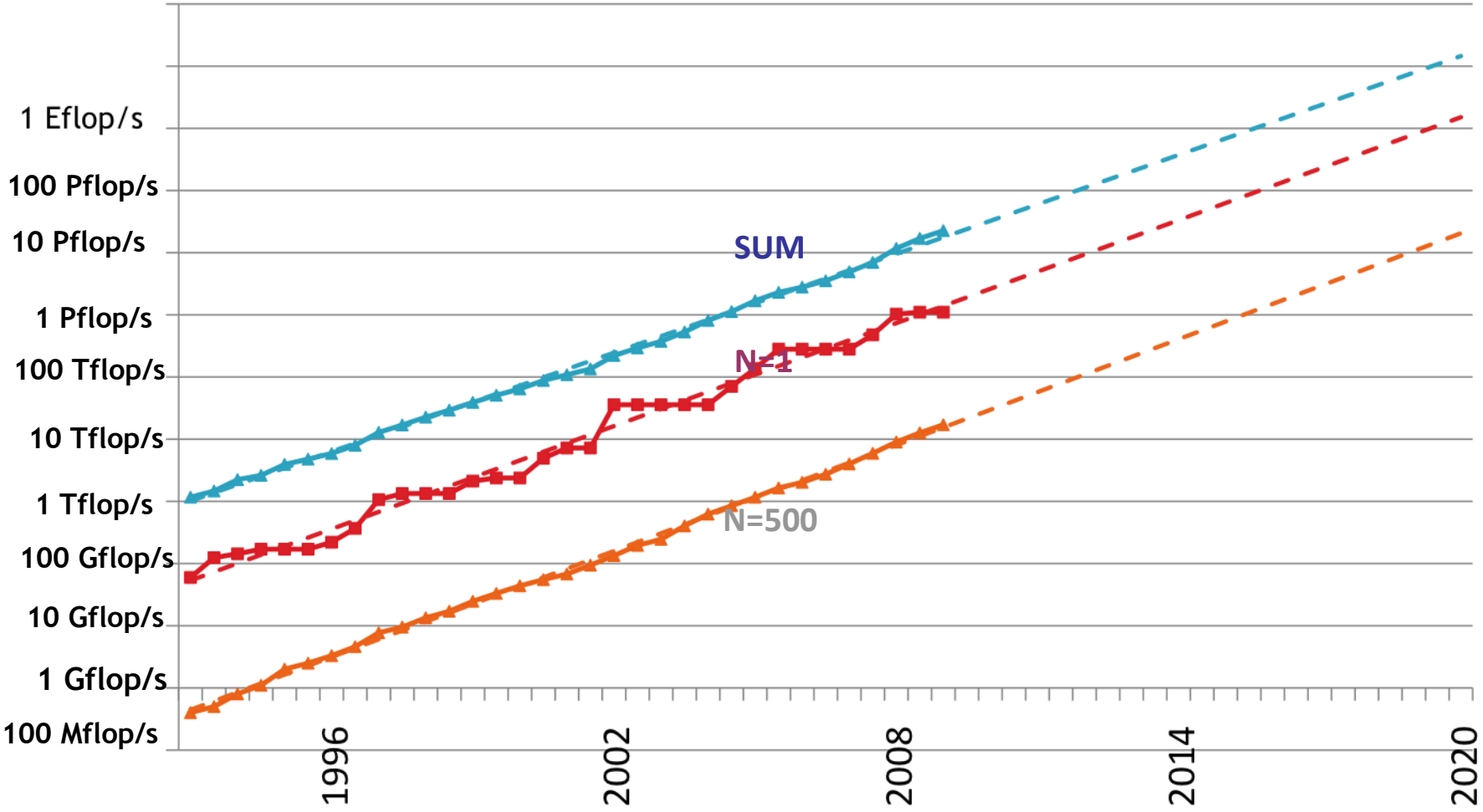
Cores per Socket



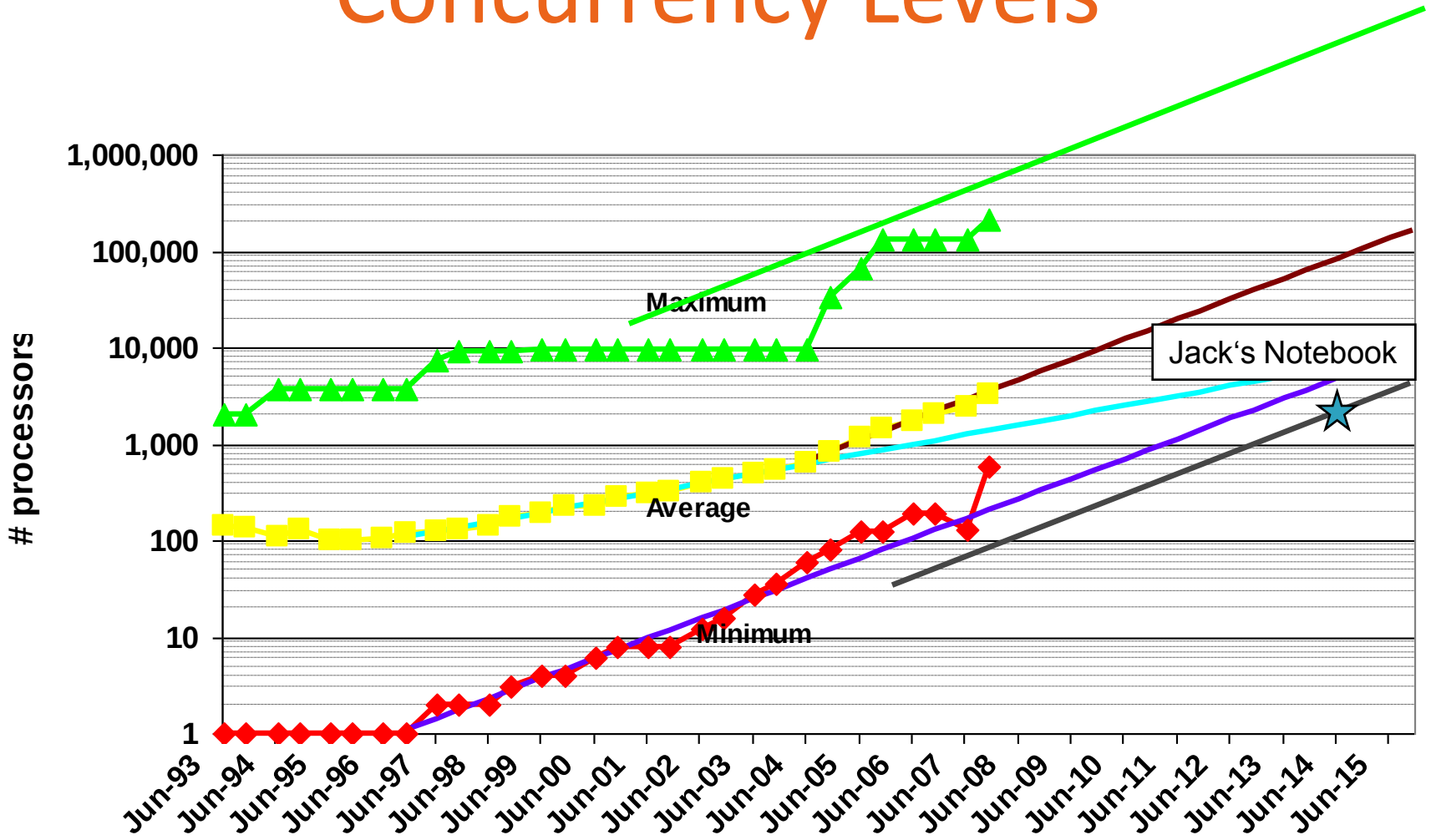
Performance Development



Performance Development Development

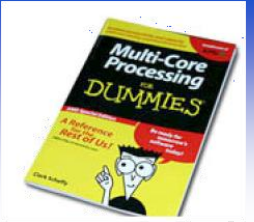


Concurrency Levels



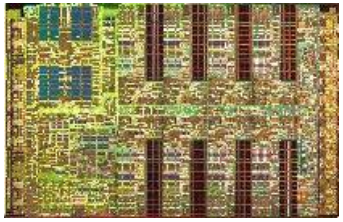
Moore's Law reinterpreted

- 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



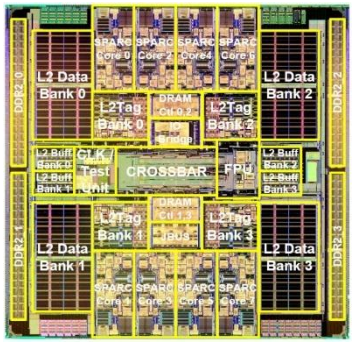
Multicore comes in a wide variety

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

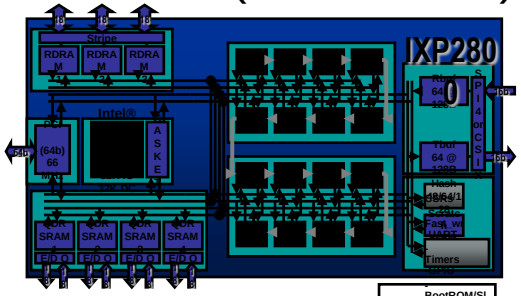


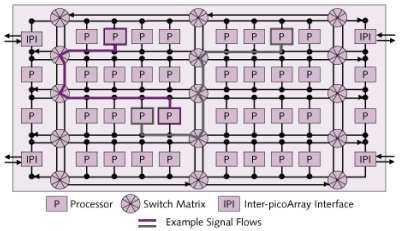
IBM Cell
1 GPP (2 threads)
8 ASPs

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

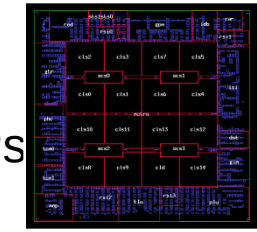


Sun Niagara
8 GPP cores (32 threads)

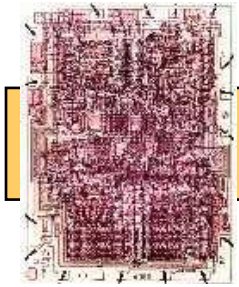




Picochip DSP
1 GPP core
248 ASPs



Cisco CRS-1
188 Tensilica GPPs



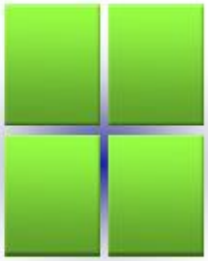
Intel 4004 (1971):
4-bit processor,
2312 transistors,
~100 KIPS,
10 micron PMOS,
11 mm^2 chip

1000s of processor cores per die

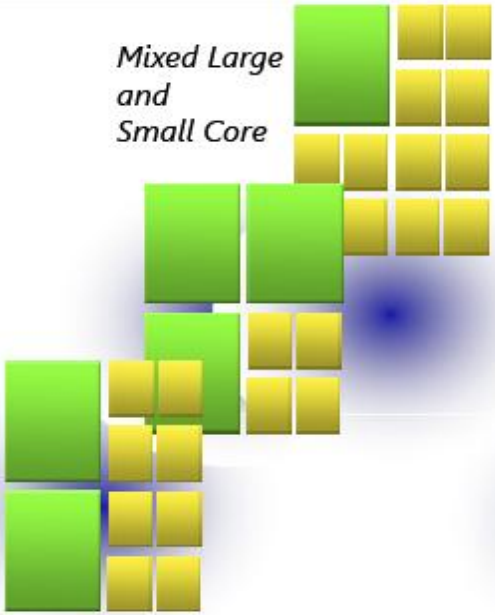
"The Processor is the new Transistor" [Rowen]

What's Next?

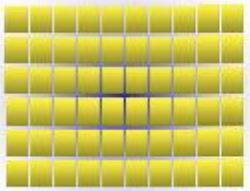
All Large Core



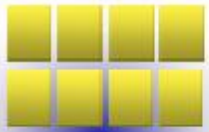
Mixed Large and Small Core



Many Small Cores

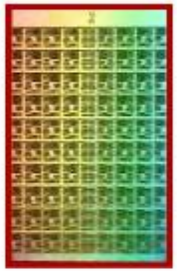


All Small Core



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

Many Floating-Point Cores

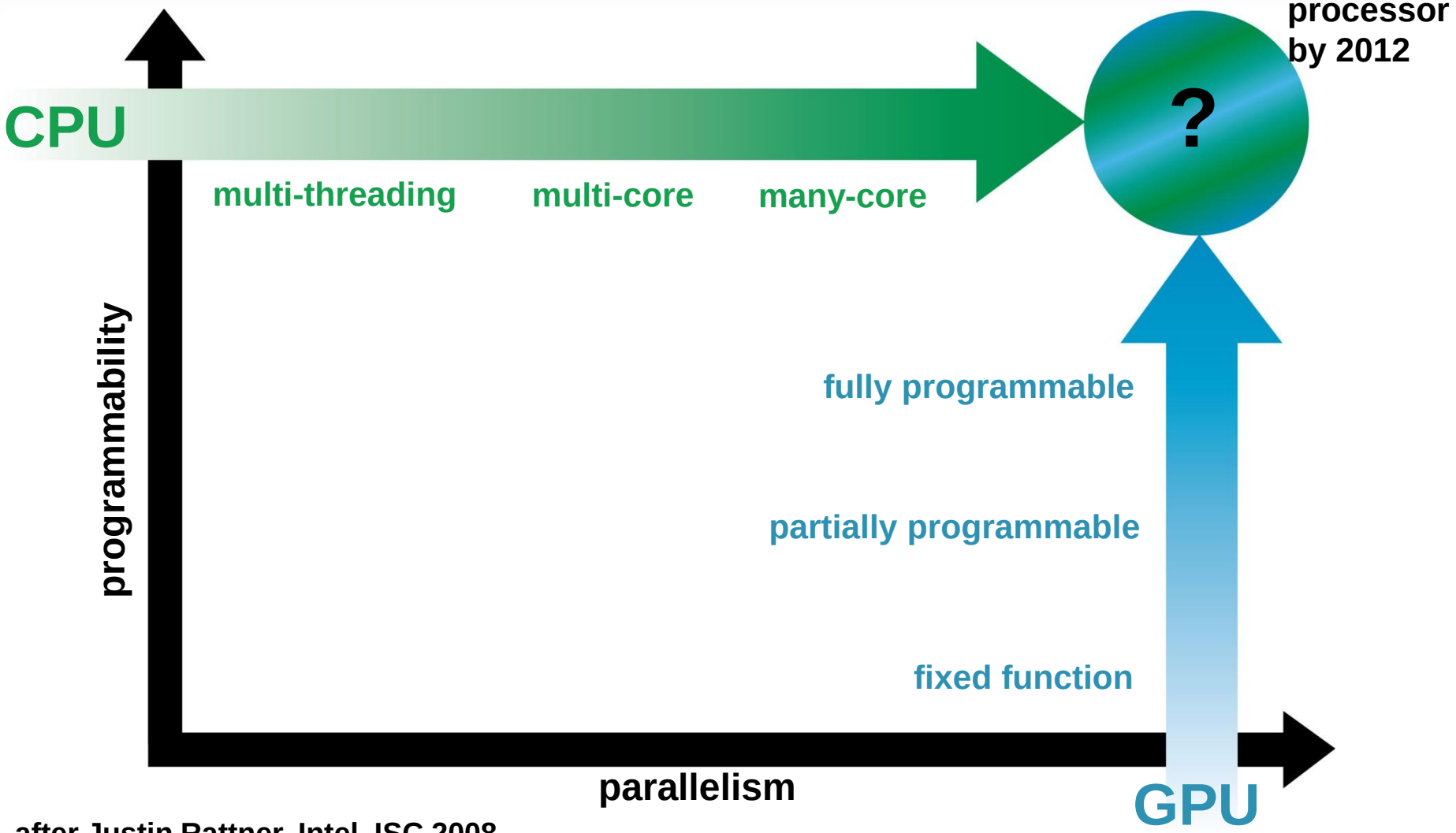


+ 3D Stacked Memory

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

Source: Jack Dongarra, ISC 2008

A Likely Trajectory - Collision or Convergence?



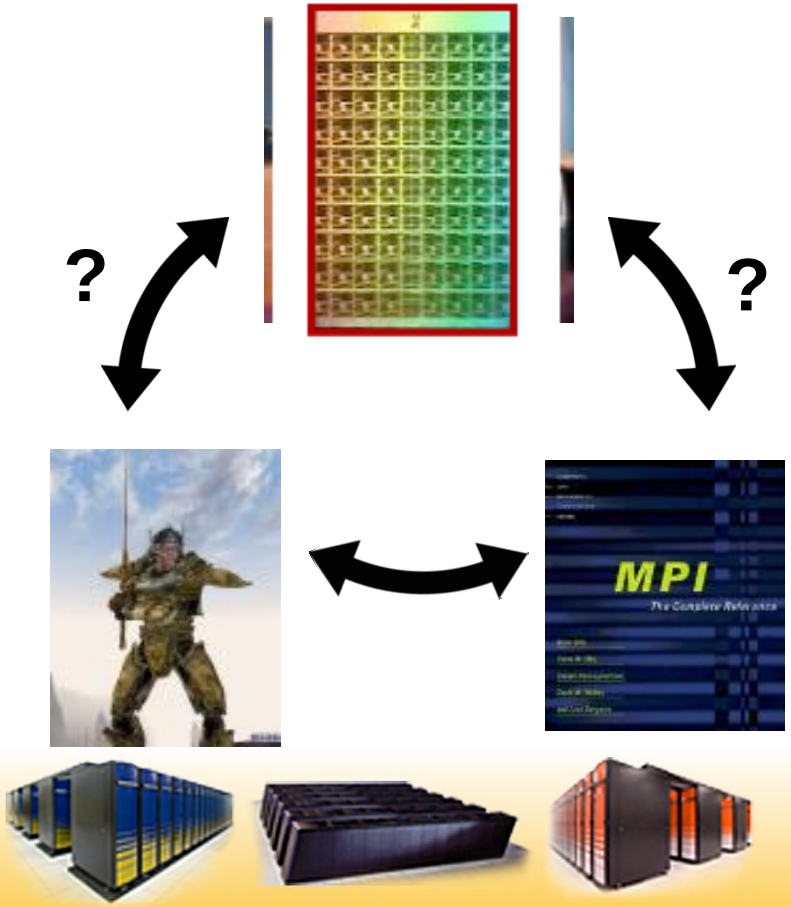
after Justin Rattner, Intel, ISC 2008

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

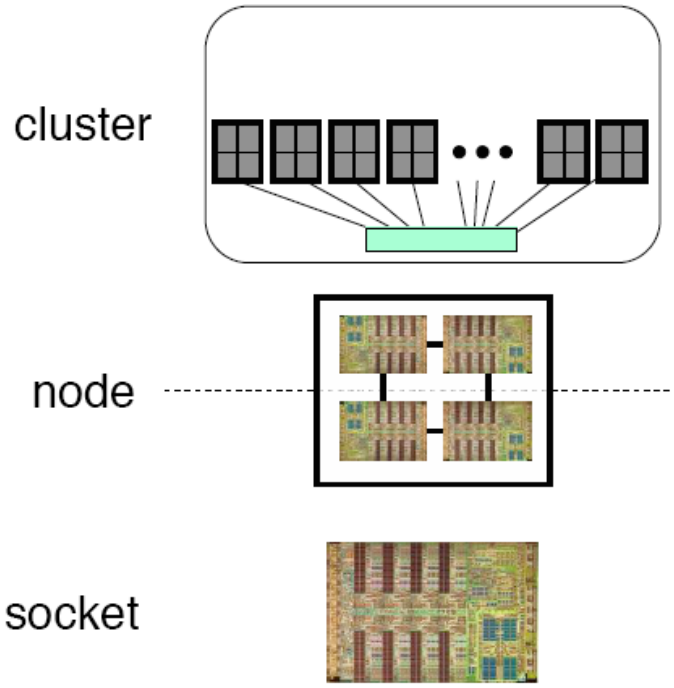
- **We will need to rethink and redesign our software**
  - **Similar challenge as the 1990 to 1995 transition to clusters and MPI**



A Likely Future Scenario (2014)

System: cluster + many core node

Programming model: MPI+?



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 “?” in MPI+?

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

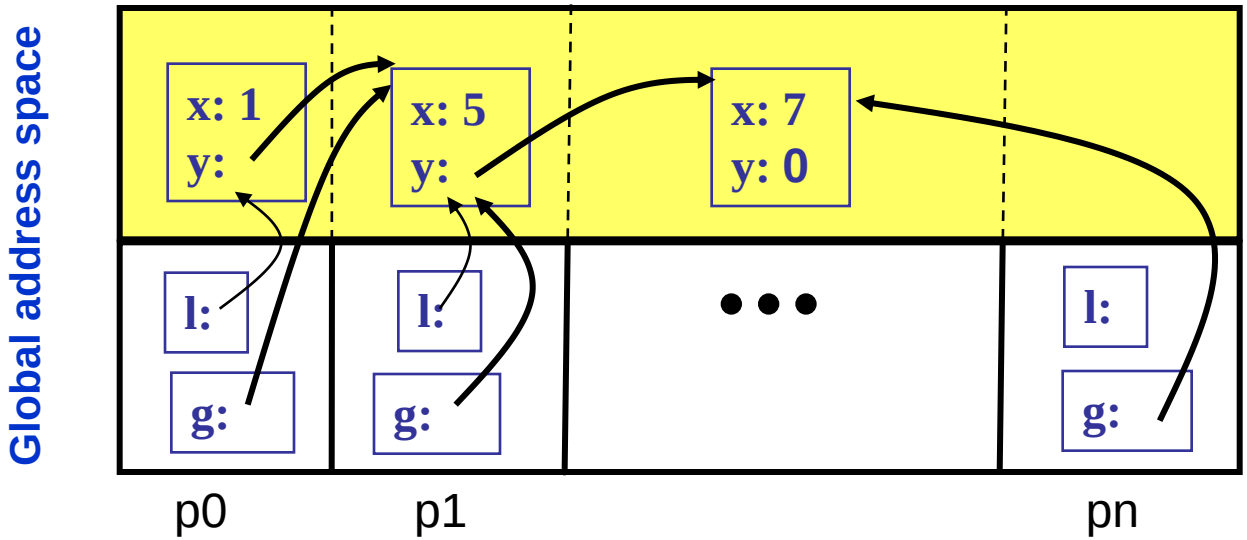
What's Wrong with MPI Everywhere?

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



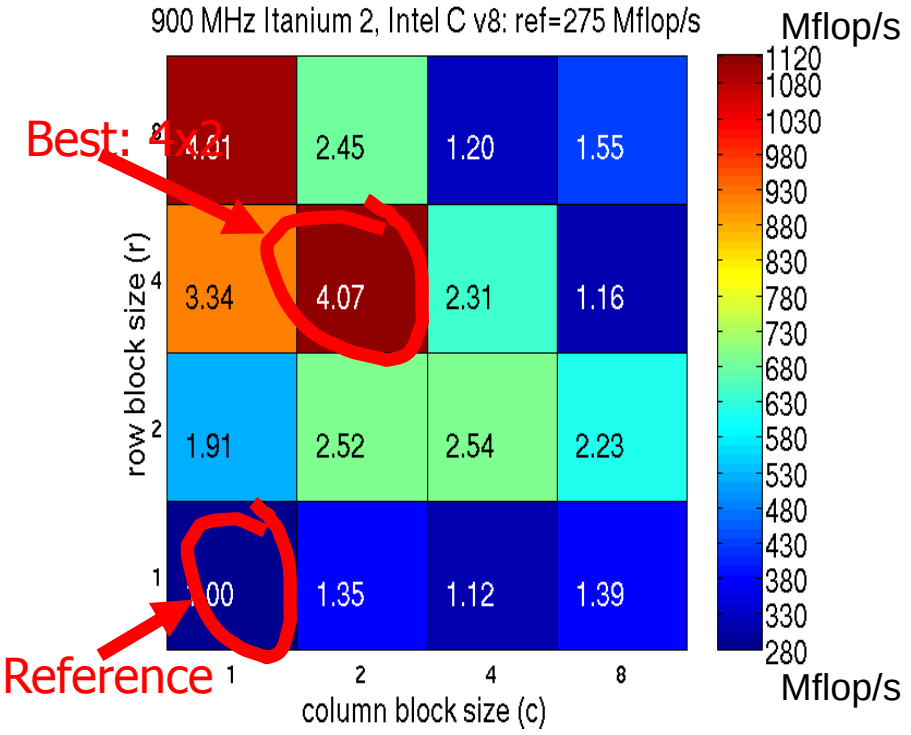
- **Implementation issues:**
  - **Distributed memory: Reading a remote array or structure is explicit, not a cache fill**
  - **Shared memory: Caches are allowed, but not required**
- **No less scalable than MPI!**
- **Permits sharing, whereas MPI rules it out!**

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**

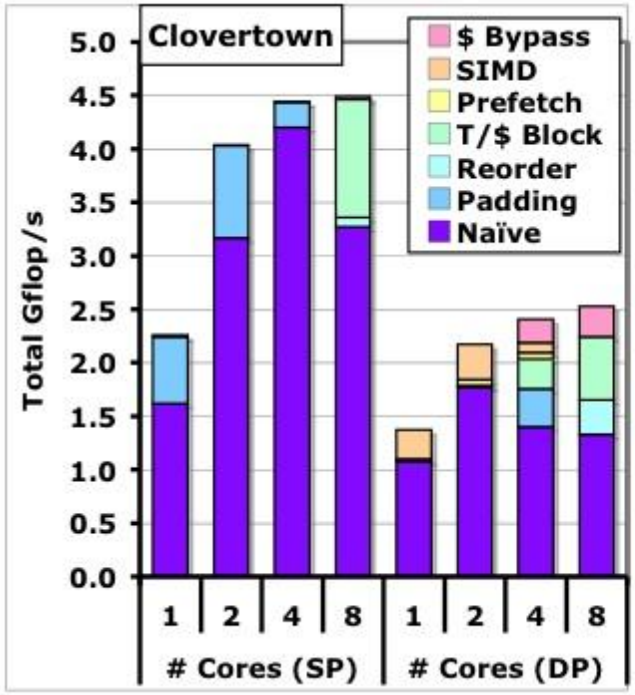
For finite element problem
[Im, Yelick, Vuduc, 2005]

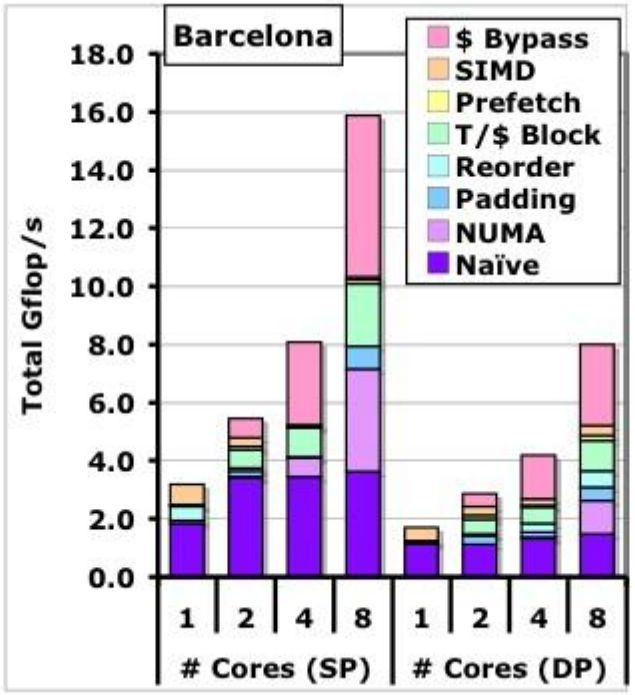


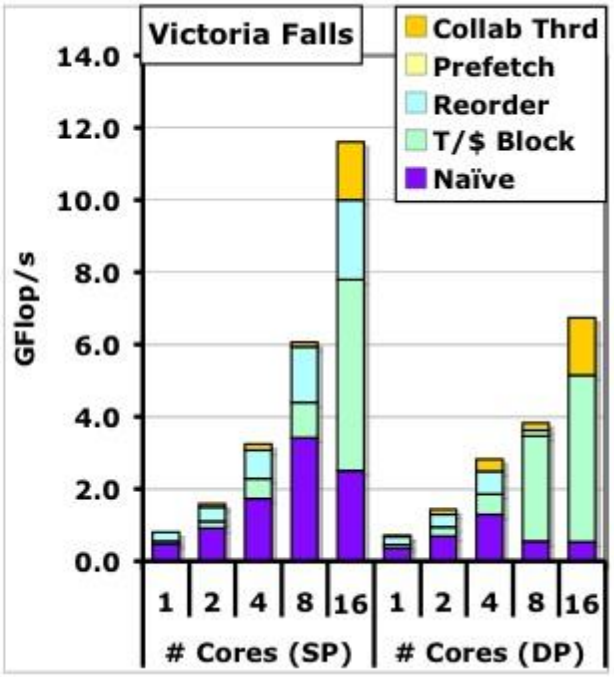
BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

Autotuning for Scalability and Performance Portability





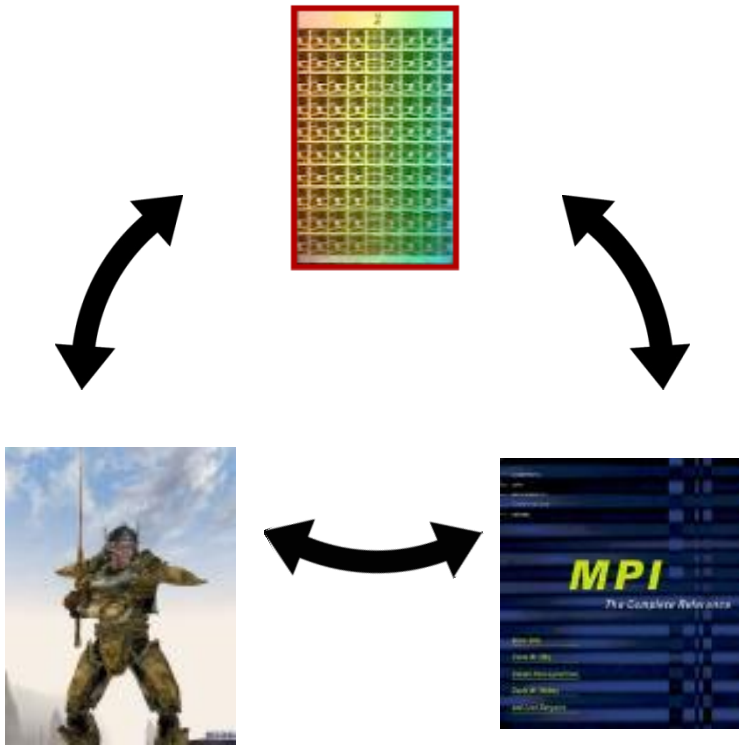


BERKELEY LAB

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

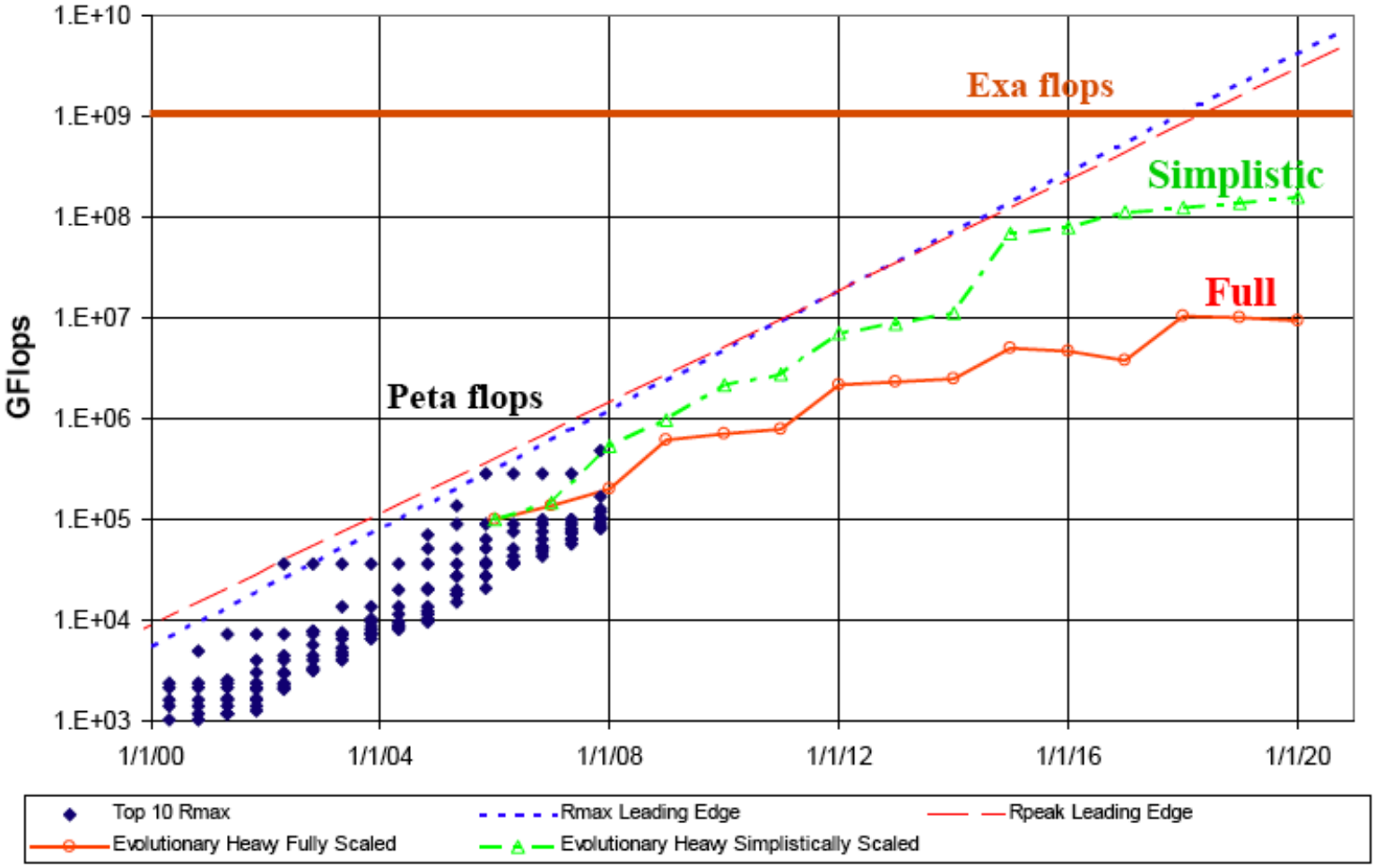
Overview

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

DARPA Exascale Study

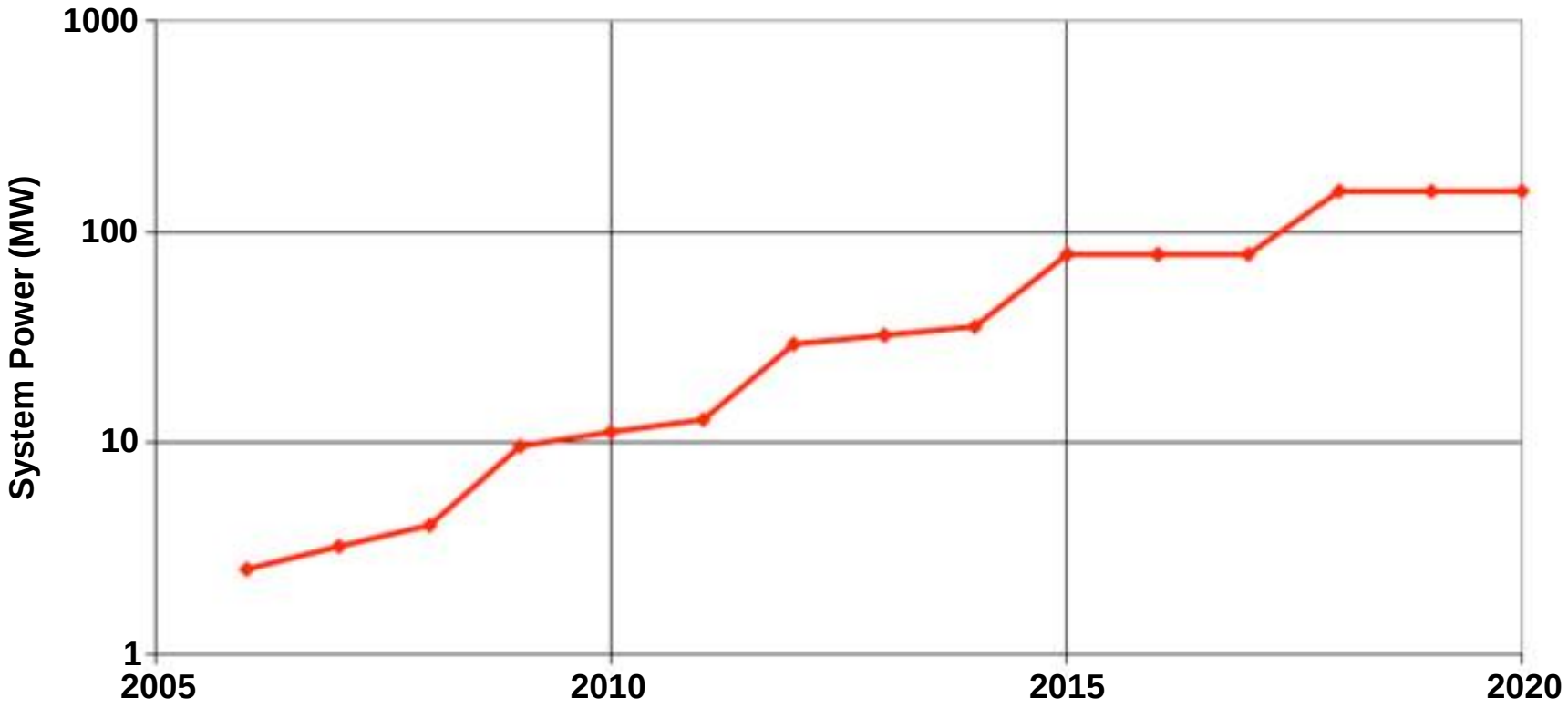
- **Commissioned by DARPA to explore the challenges for Exaflop computing (Kogge et al.)**
- **Two models for future performance growth**
  - **Simplistic: ITRS roadmap; power for memory grows linear with # of chips; power for interconnect stays constant**
  - **Fully scaled: same as simplistic, but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

We won't reach Exaflops with this approach



From Peter Kogge, DARPA Exascale Study

… and the power costs will still be staggering



From Peter Kogge,
DARPA Exascale Study

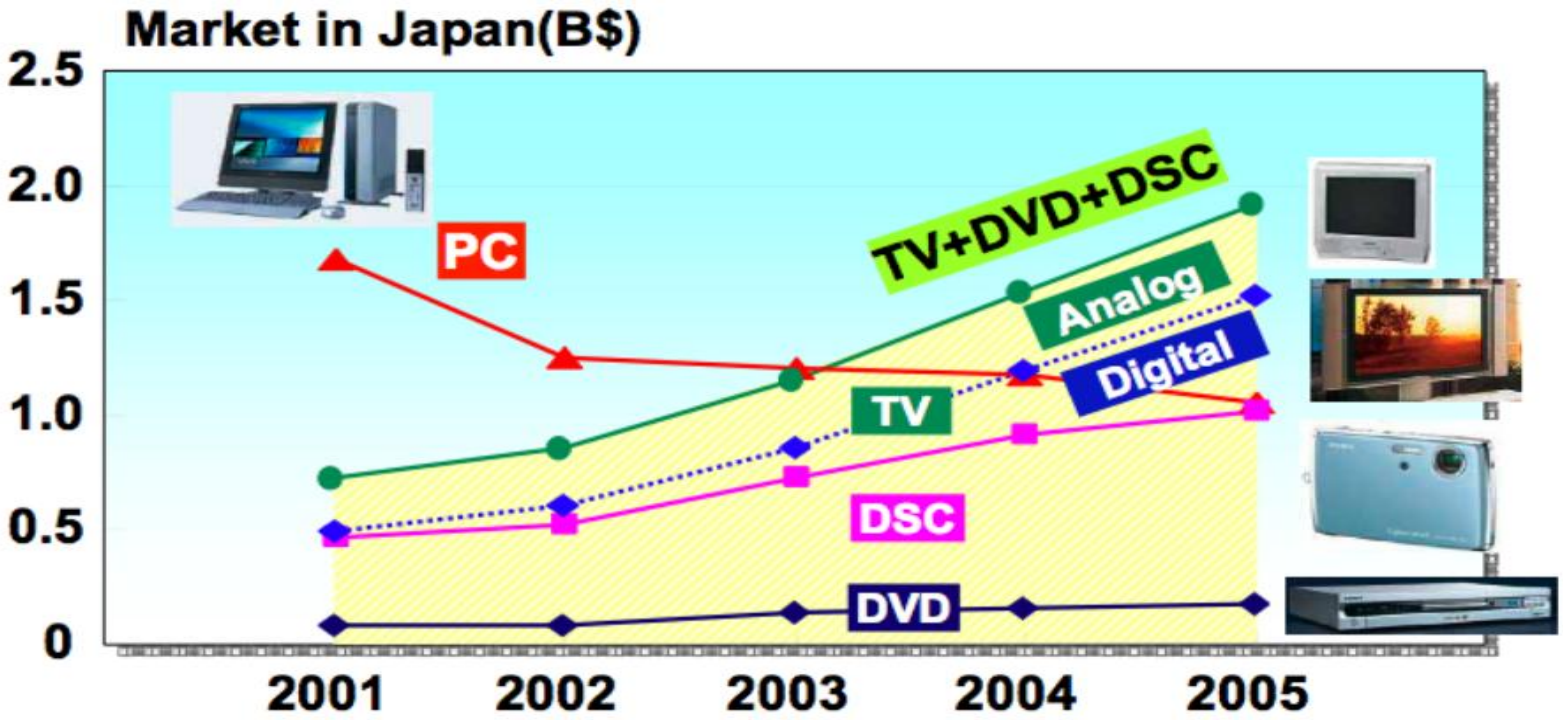
Extrapolating to Exaflop/s in 2018

	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)

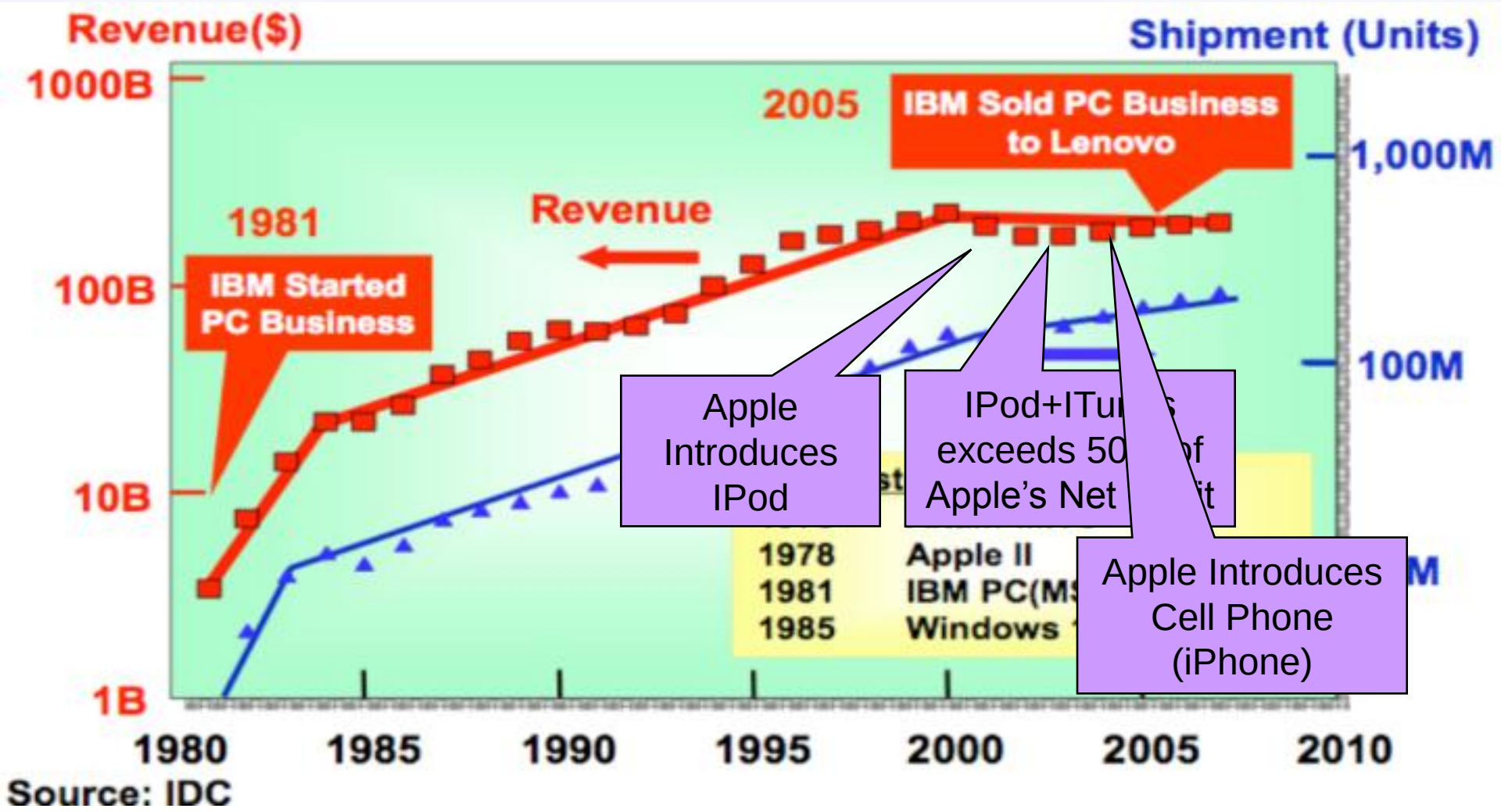
Source: David Turek, IBM

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!!



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.

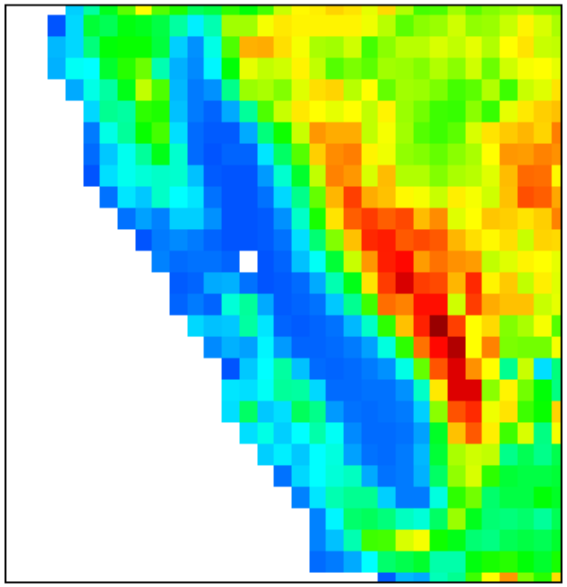
BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

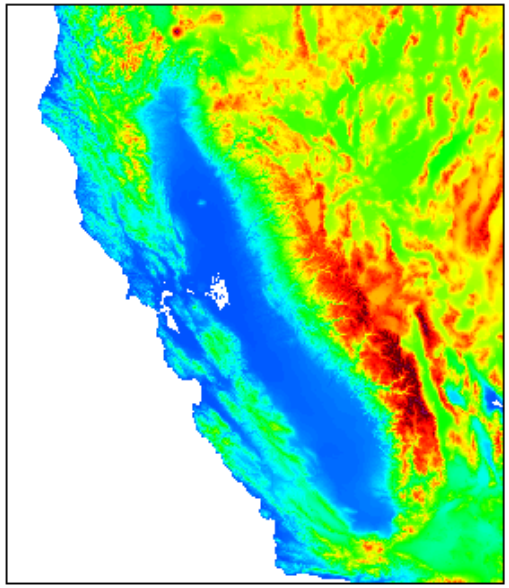
Global Cloud System Resolving Models

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

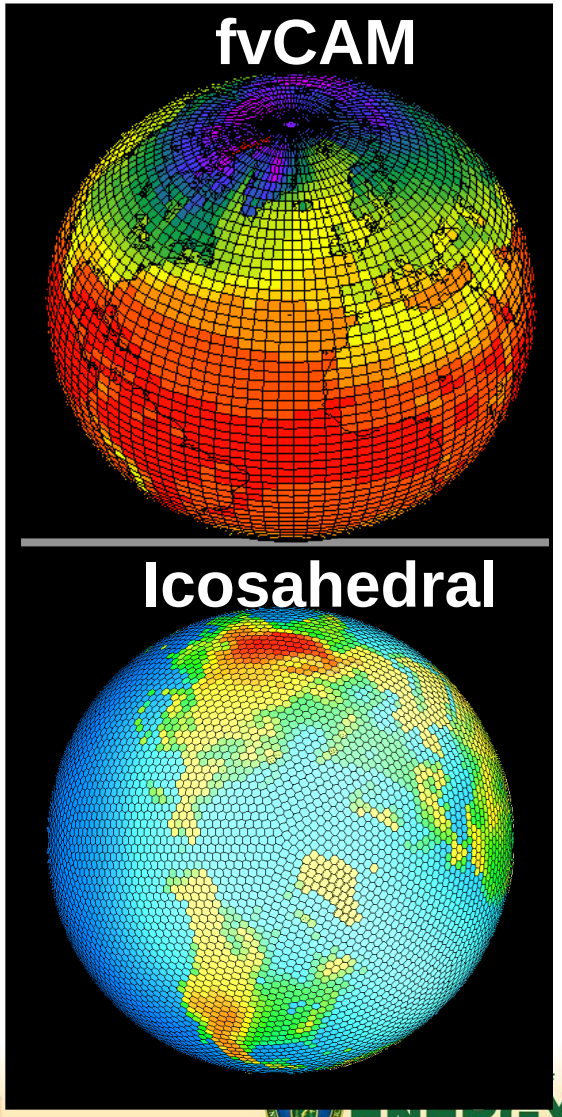
BERKELEY LAB

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

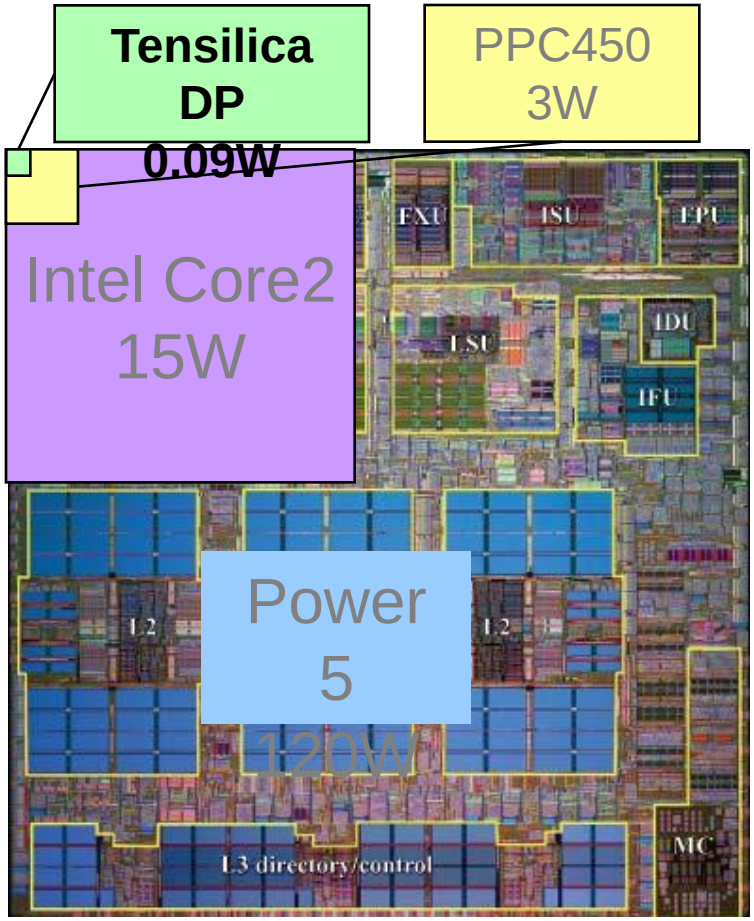
We examined three different approaches (in 2008 technology)

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

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

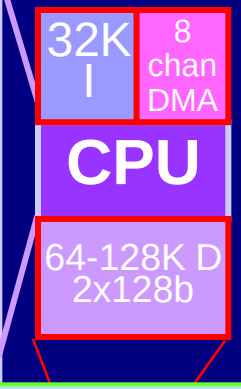
This is how iPhones and MP3 players are designed to maximize battery life and minimize cost

Climate System Design Concept

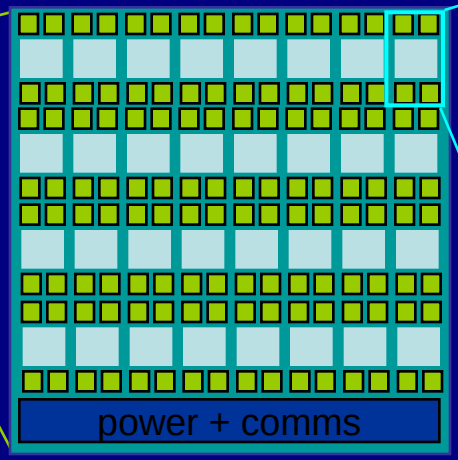
Strawman Design Study

VLIW CPU:

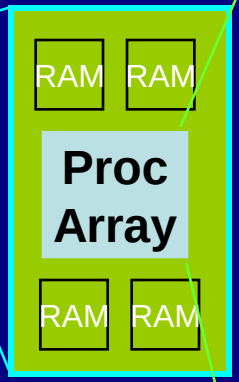
- 128b load-store + 2 DP MUL/ADD + integer op/ DMA per cycle:
- Synthesizable at 650MHz in commodity 65nm
- 1mm² core, 1.8-2.8mm² with inst cache, data cache data RAM, DMA interface, 0.25mW/MHz
- Double precision SIMD FP : 4 ops/cycle (2.7GFLOPs)
- Vectorizing compiler, cycle-accurate simulator, debugger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



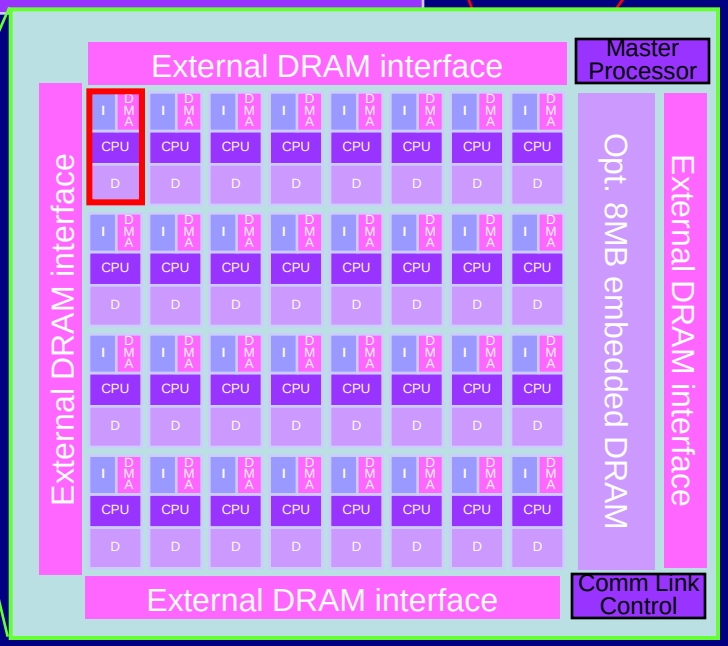
100 racks @ ~25KW



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



8 DRAM per processor chip: ~50 GB/s



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**