Future Directions in High Performance Computing

Juan Meza and Horst Simon

High Performance Computing Research
Lawrence Berkeley National Laboratory

Society of Exploration Geophysicists, Houston, TX, October 26, 2009

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

Key Message

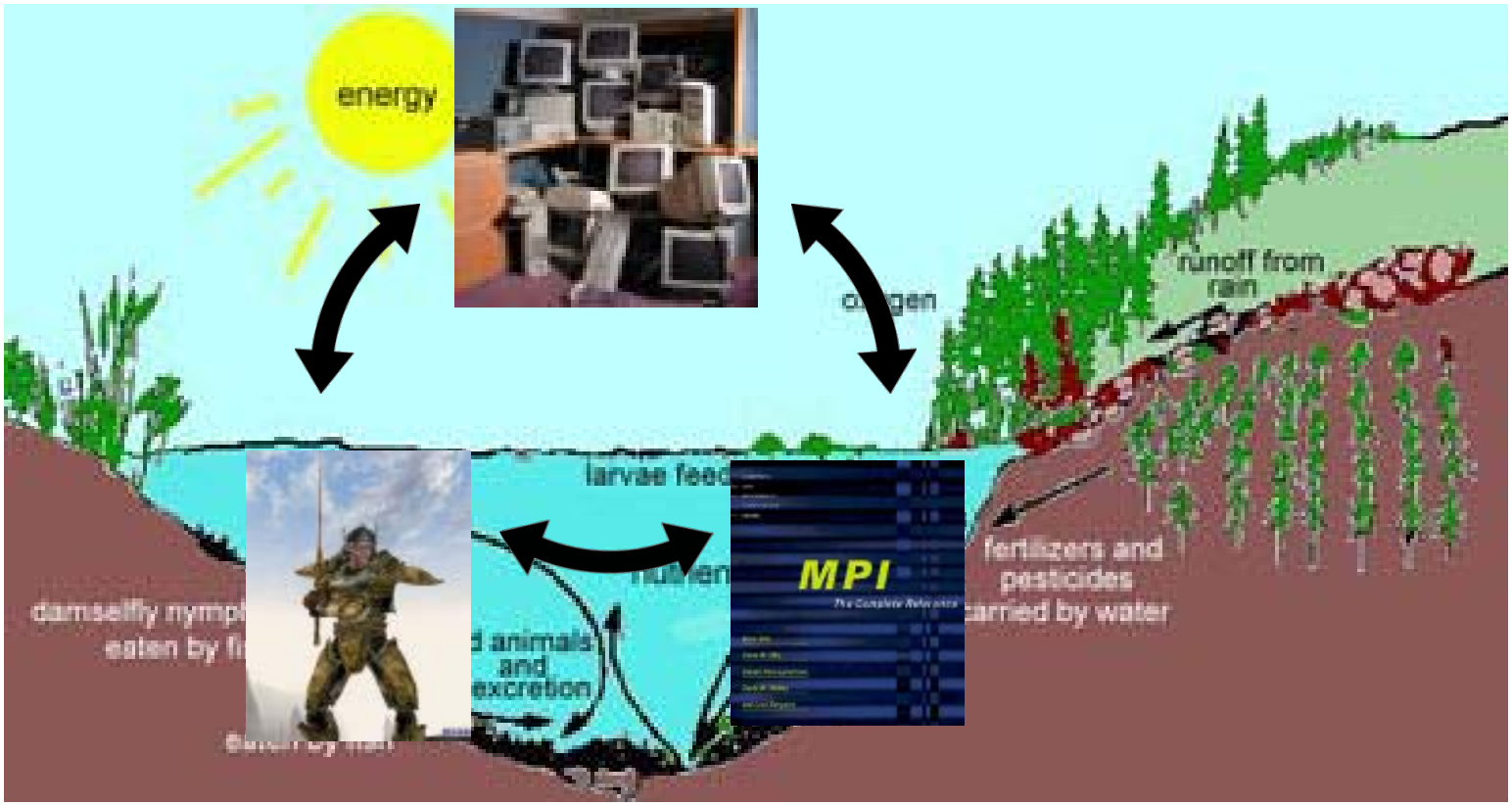
Computing is changing more rapidly than ever before, and scientists have the unprecedented opportunity to change computing directions

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)



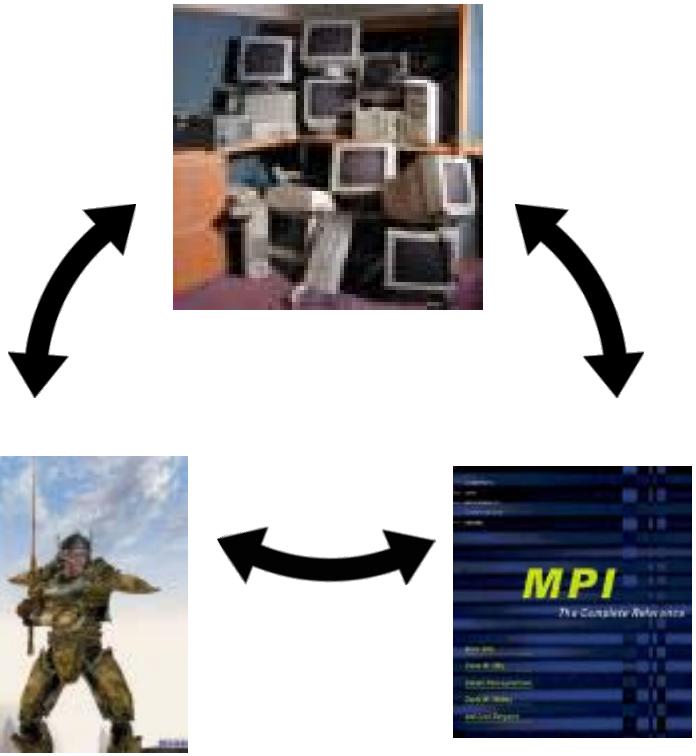
“Clusters”

12 years of legacy MPI applications base

From Horst Simon presentation at ISC 2005

Supercomputing Ecosystem (2005)

Commercial Off The Shelf technology (COTS)



“Clusters”

12 years of legacy MPI applications base

From Horst Simon 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!**

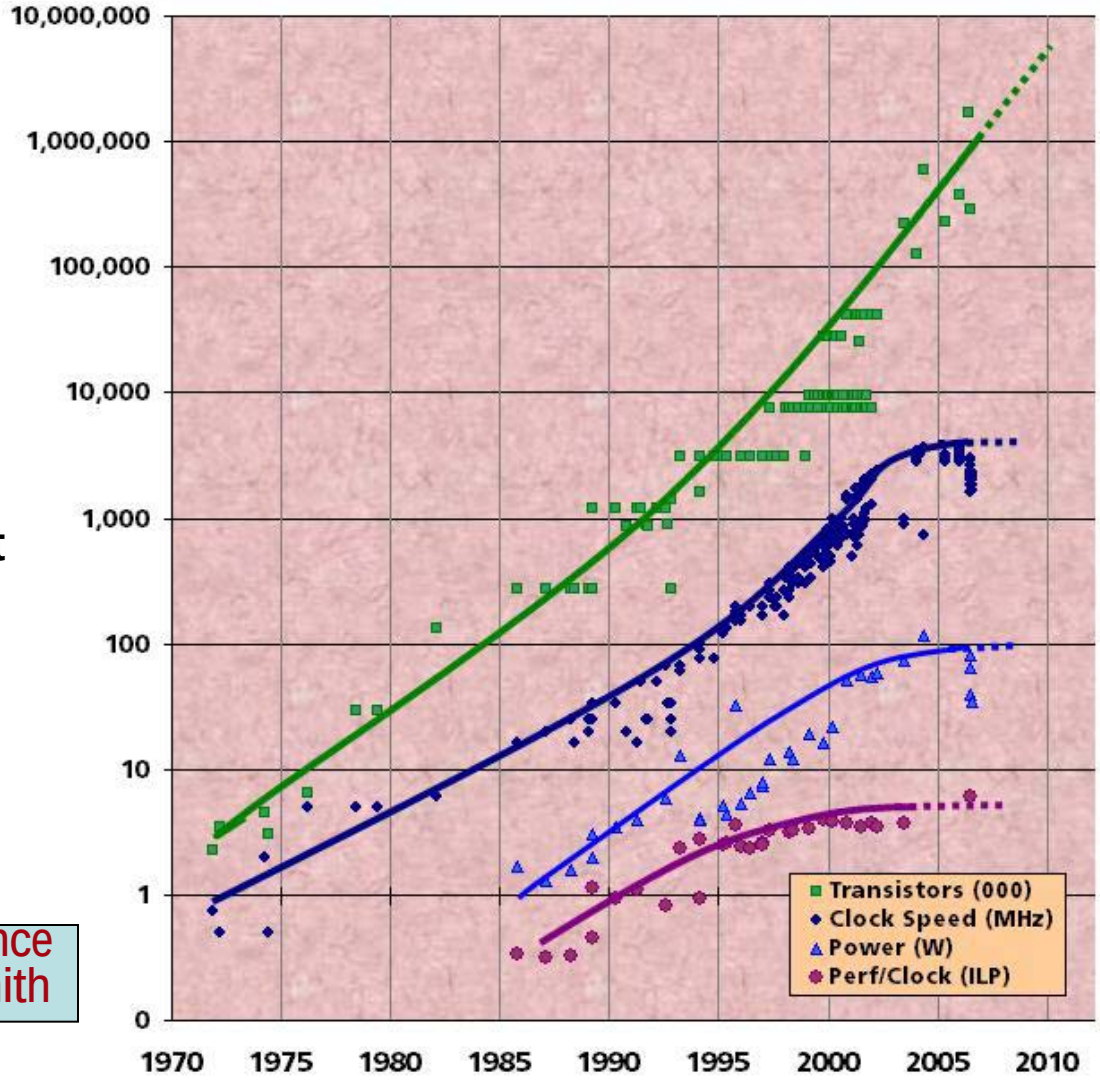


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

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

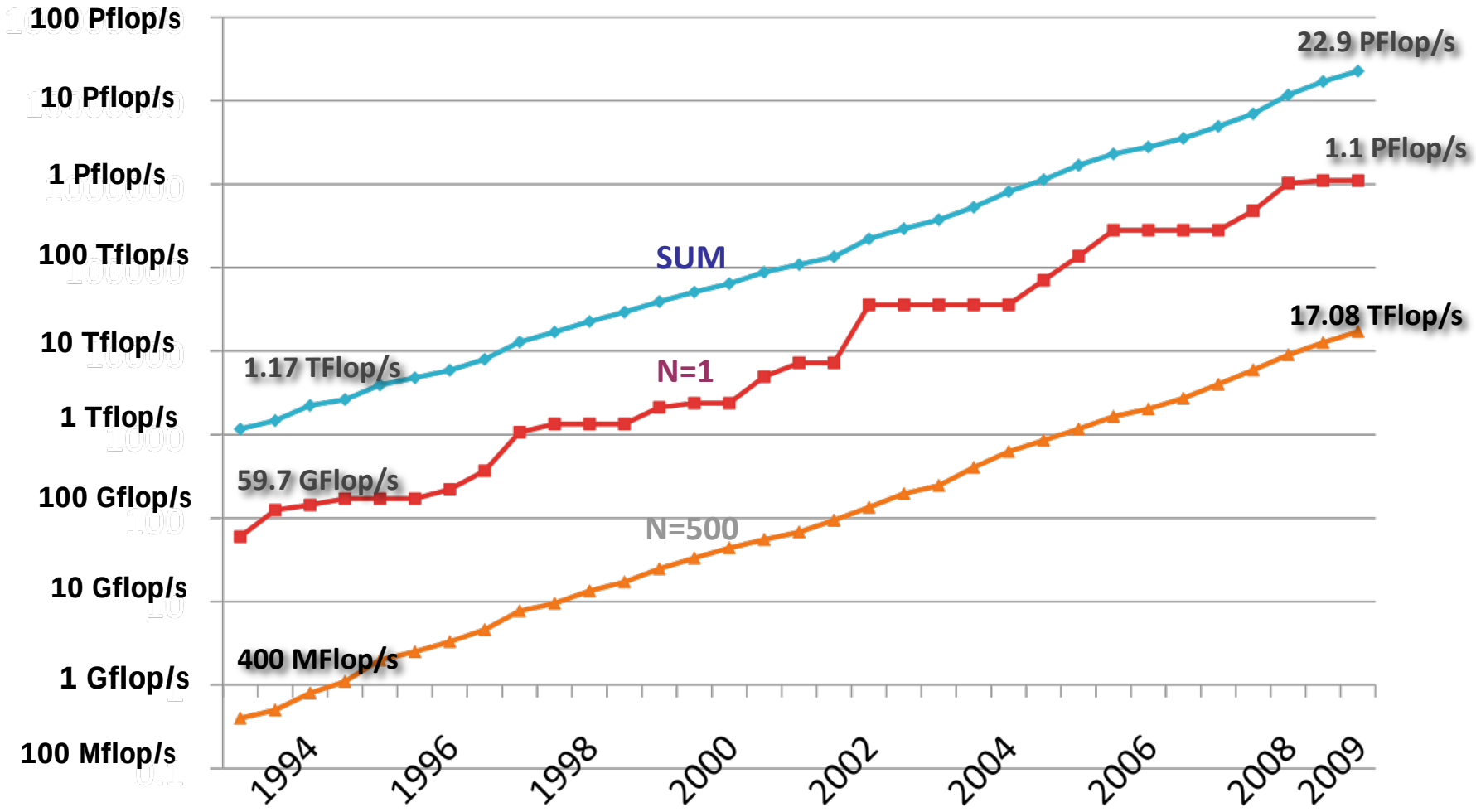
Cray XT5 at ORNL -- 1 Pflop/s in November 2008

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

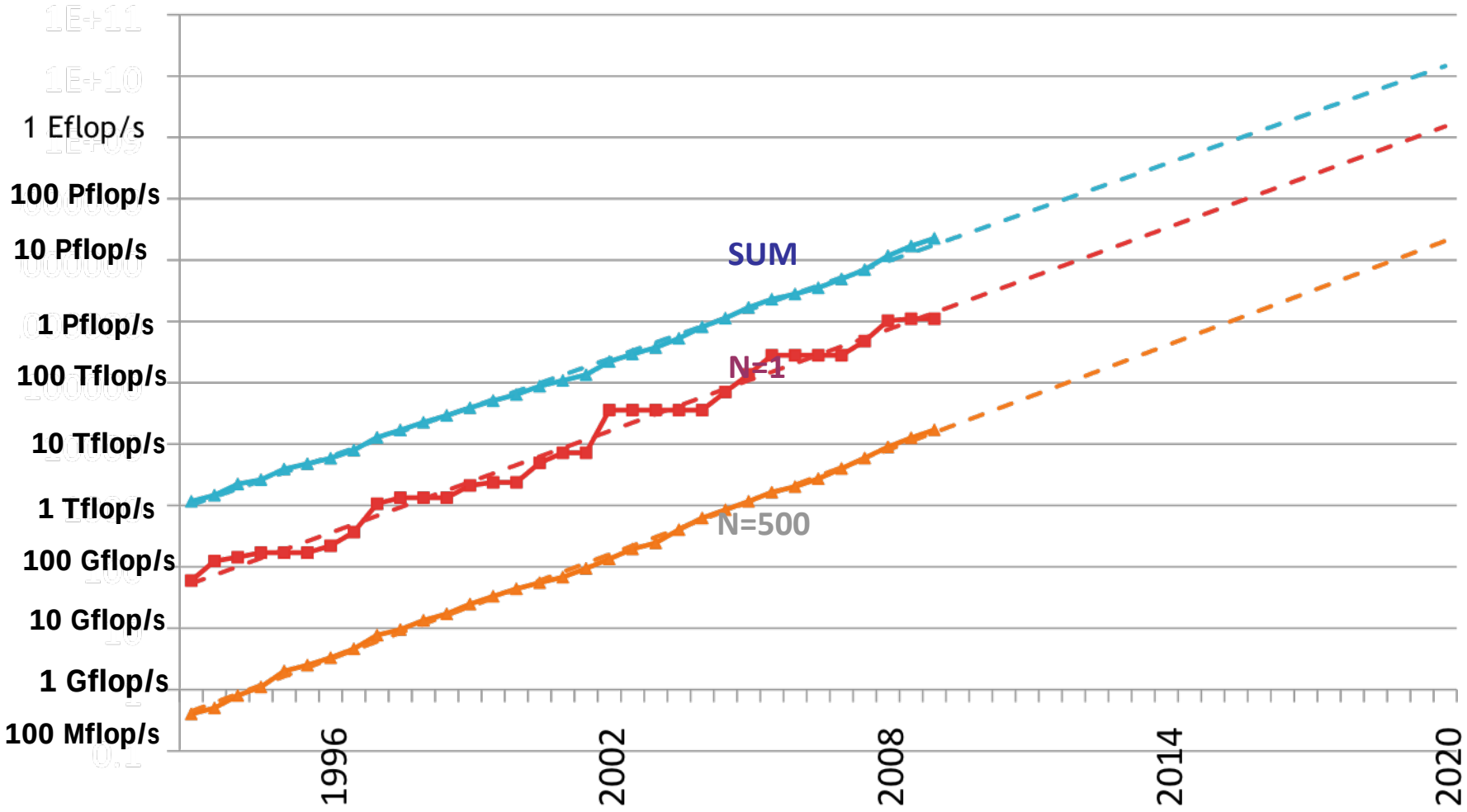
U.S. DEPARTMENT OF ENERGY
Office of Science

Performance Development

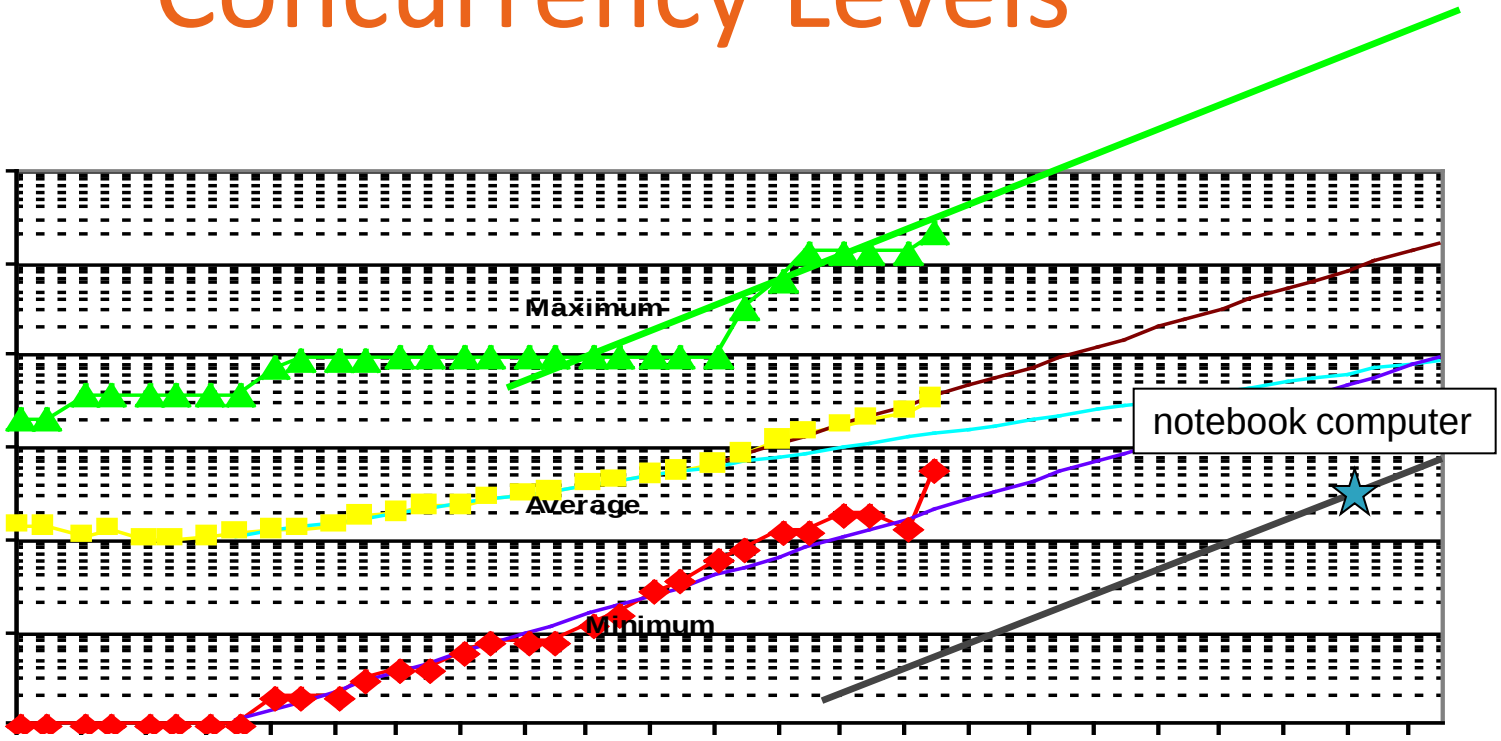


TOP500 SUPERCOMPUTER SITES

Performance Development

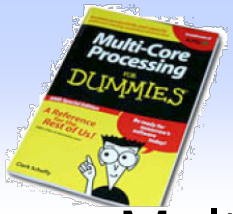


Concurrency Levels



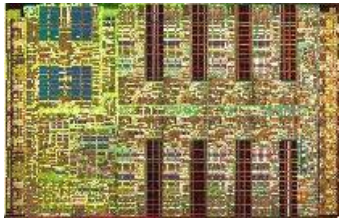
Moore's Law reinterpreted

- 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

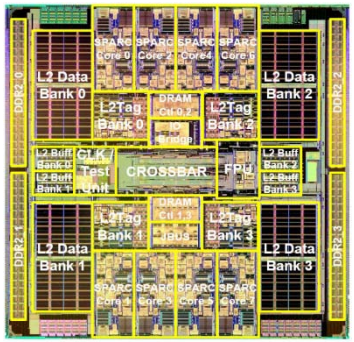


Multicore comes in a wide variety

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

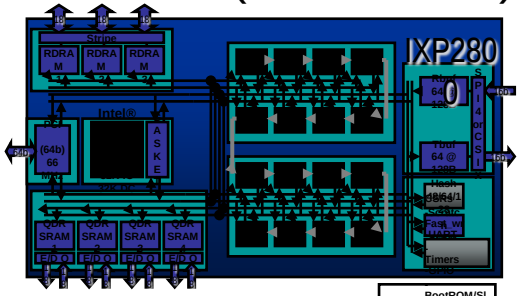


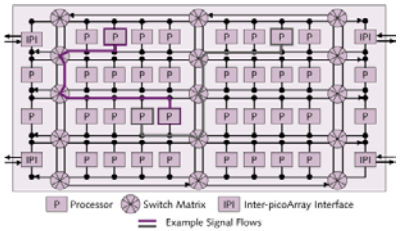
IBM Cell
1 GPP (2 threads)
8 ASPs



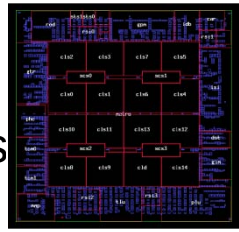
Sun Niagara
8 GPP cores (32 threads)

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

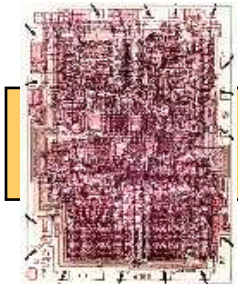




Picochip DSP
1 GPP core
248 ASPs



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

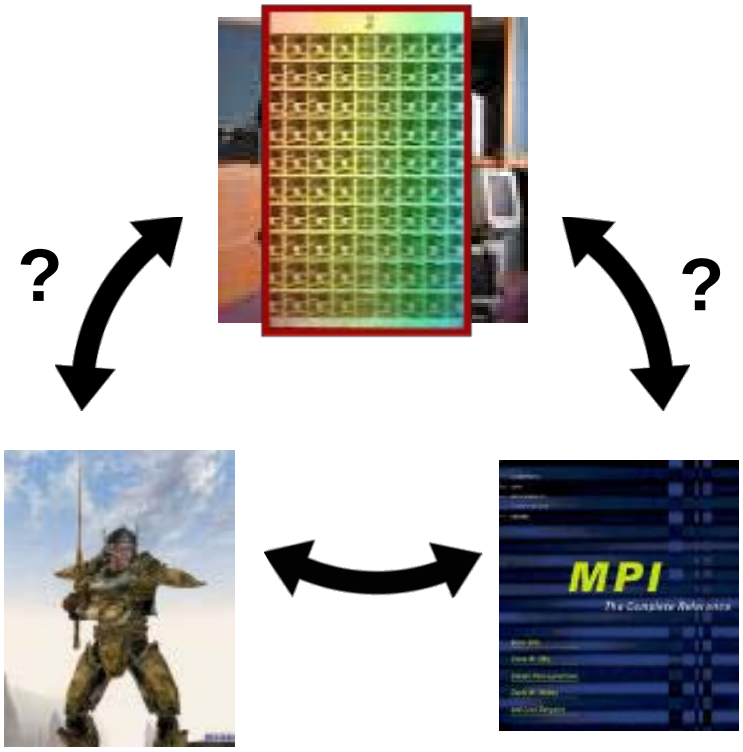
"The Processor is the new Transistor" [Rowen]

Trends for the next five years up to 2014

- 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

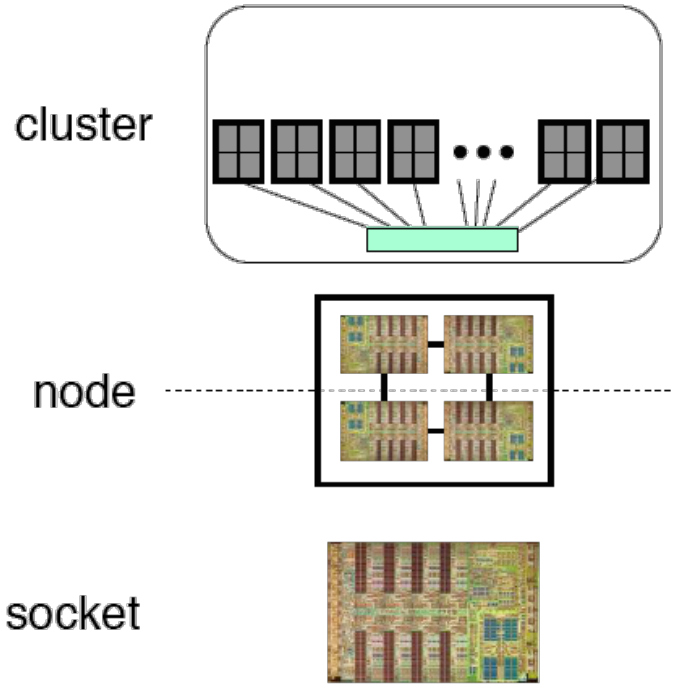
Impact on Software

- **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 will be the "?" in MPI+?

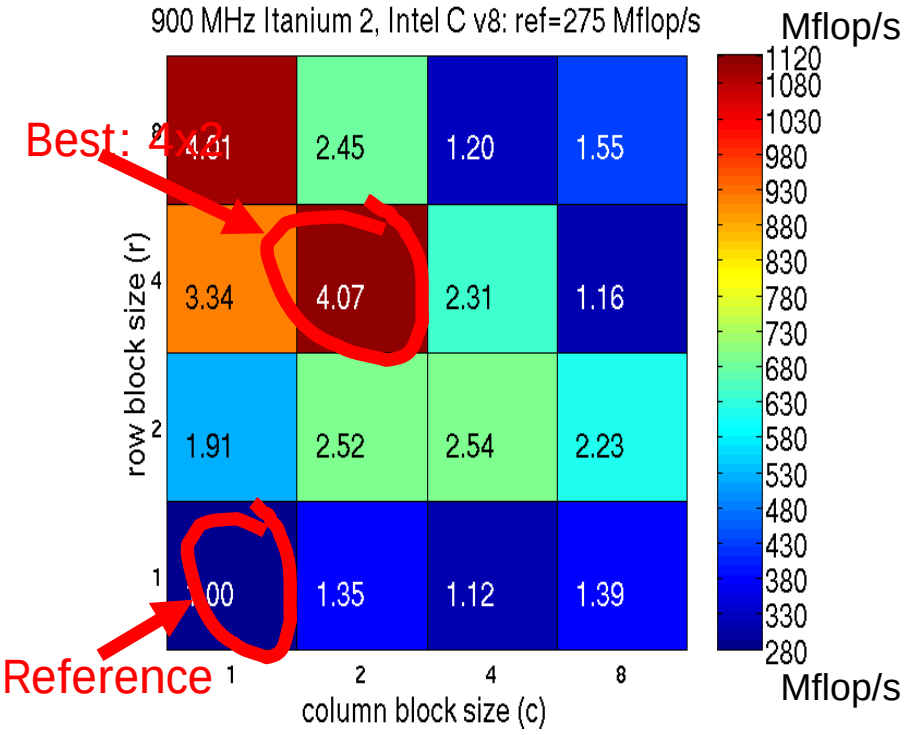
- **Likely candidates are**
  - PGAS languages
  - **Autotuning**
  - CUDA, OpenCL
  - A wildcard from commercial space

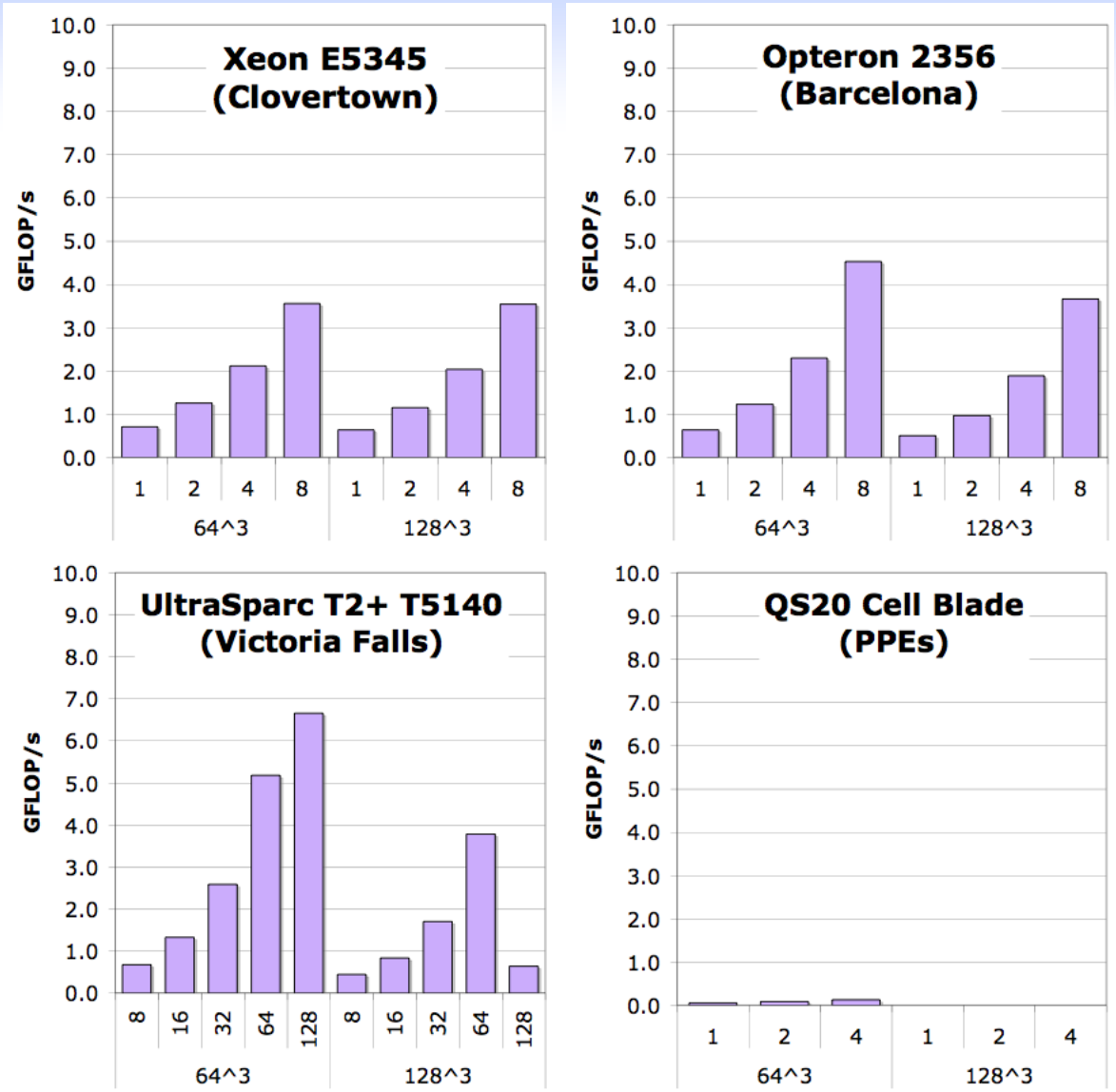
Autotuning

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]

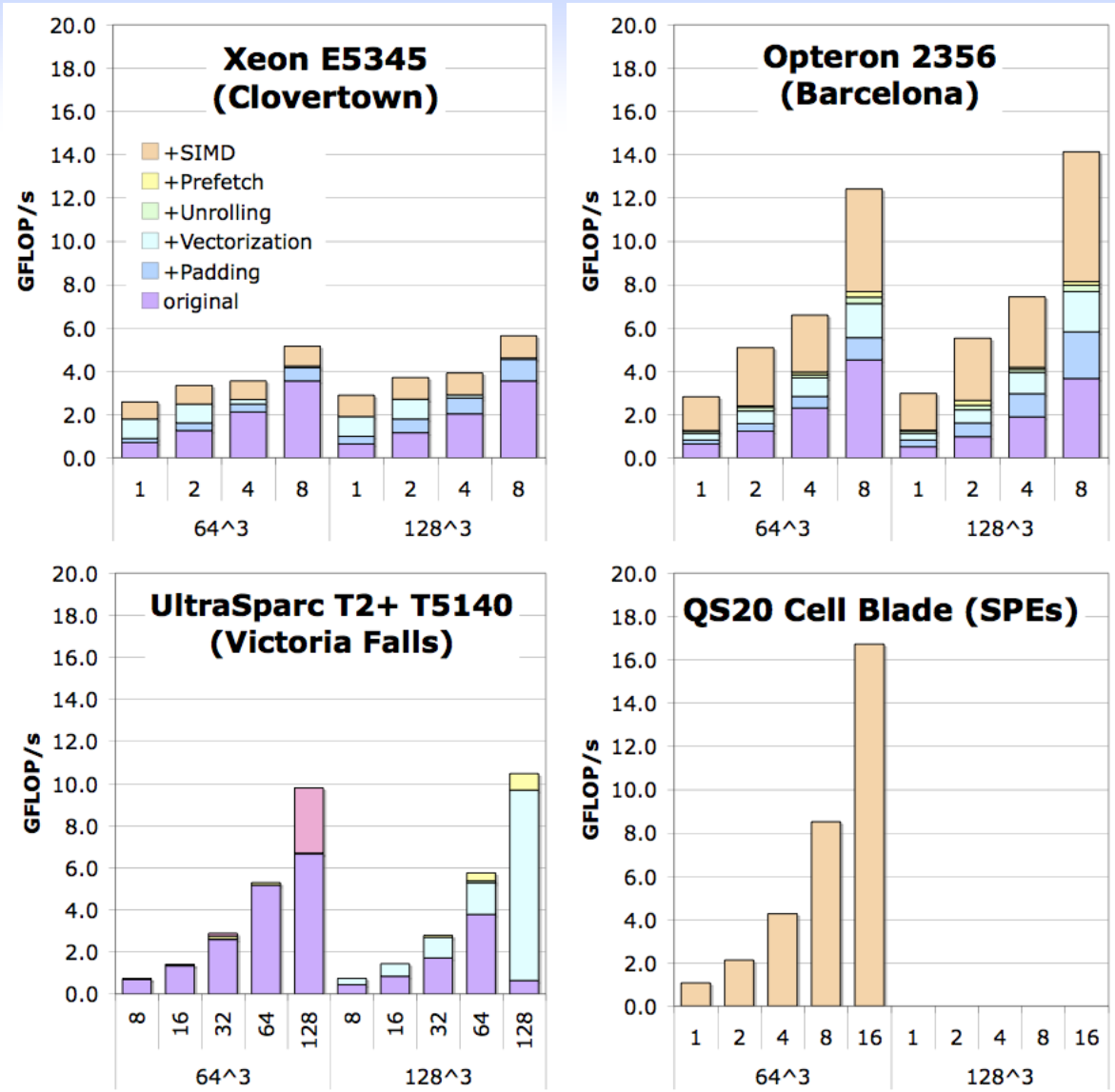




Performance

- Reference code was nominated for a Gordon Bell prize
- Used for out-of-box study on multicore performance
- Superficially, scalability looks good, but is performance good?
  - no
  - performance model

POC: Leonid Oliker, Samuel Williams (LBNL)

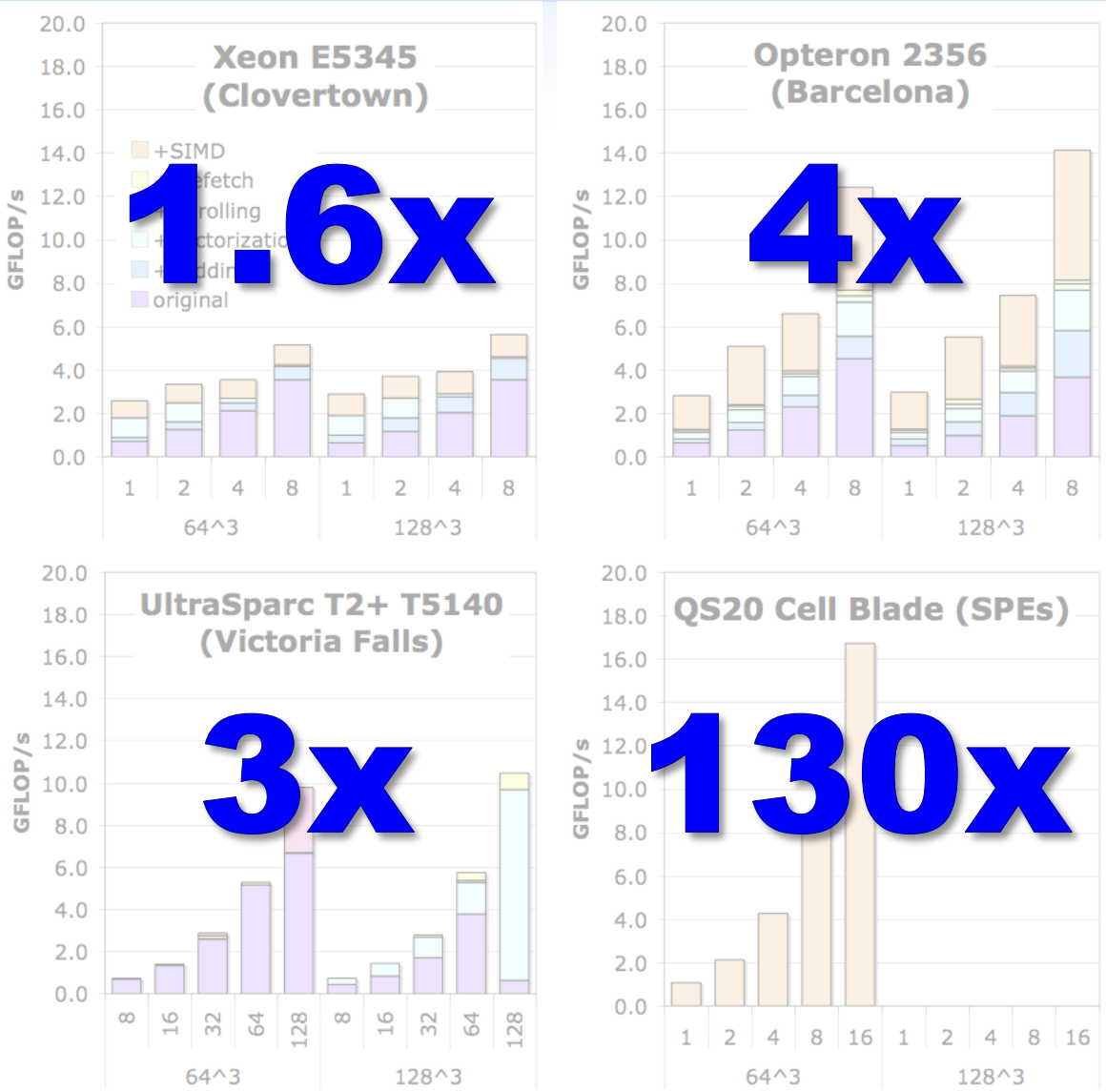


Auto-tuning

- auto-tuning dramatically improved performance
- clearly no silver bullet
- SIMD partially breaks portability premise
- Note: Cell version was optimized, not auto-tuned.

+small pages
+Explicit SIMDization
+SW Prefetching
+Unrolling
+Vectorization
+Padding
Reference+NUMA

POC: Leonid Oliker, Samuel Williams (LBNL)



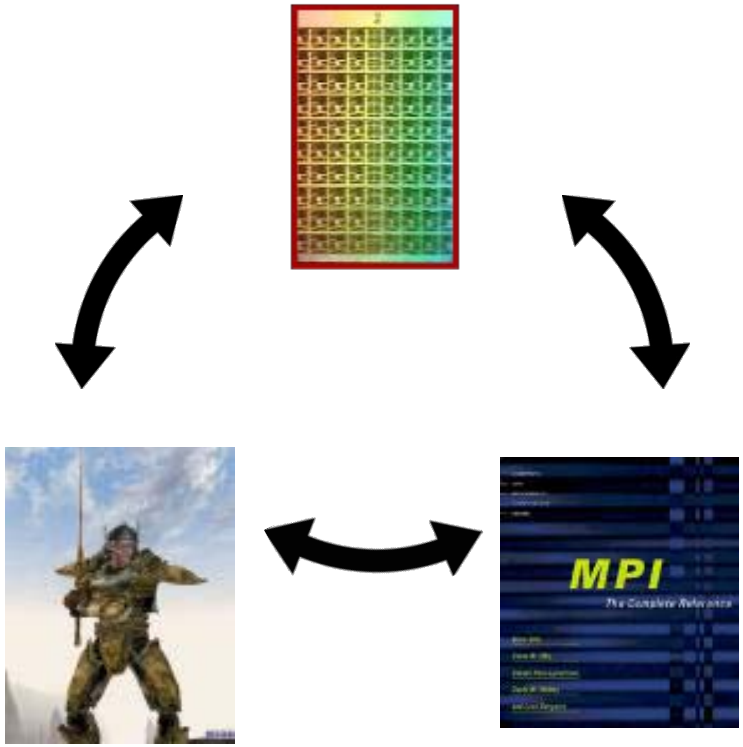
Auto-tuning

- auto-tuning dramatically improved performance
- clearly no silver bullet
- SIMD partially breaks portability premise
- Note: Cell version was optimized, not auto-tuned.

POC: Leonid Oliker, Samuel Williams (LBNL)

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

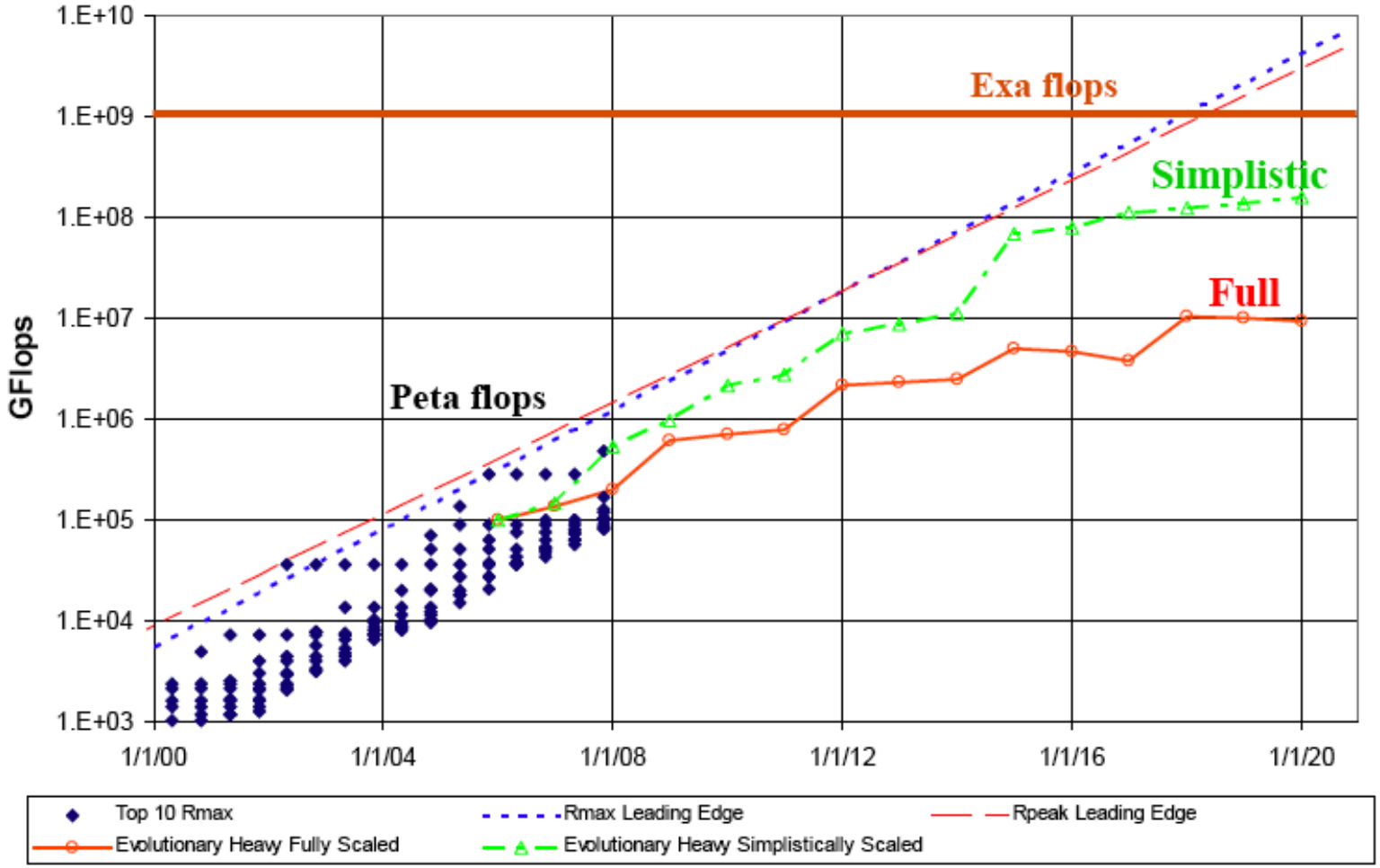
Overview

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

DARPA Exascale Study

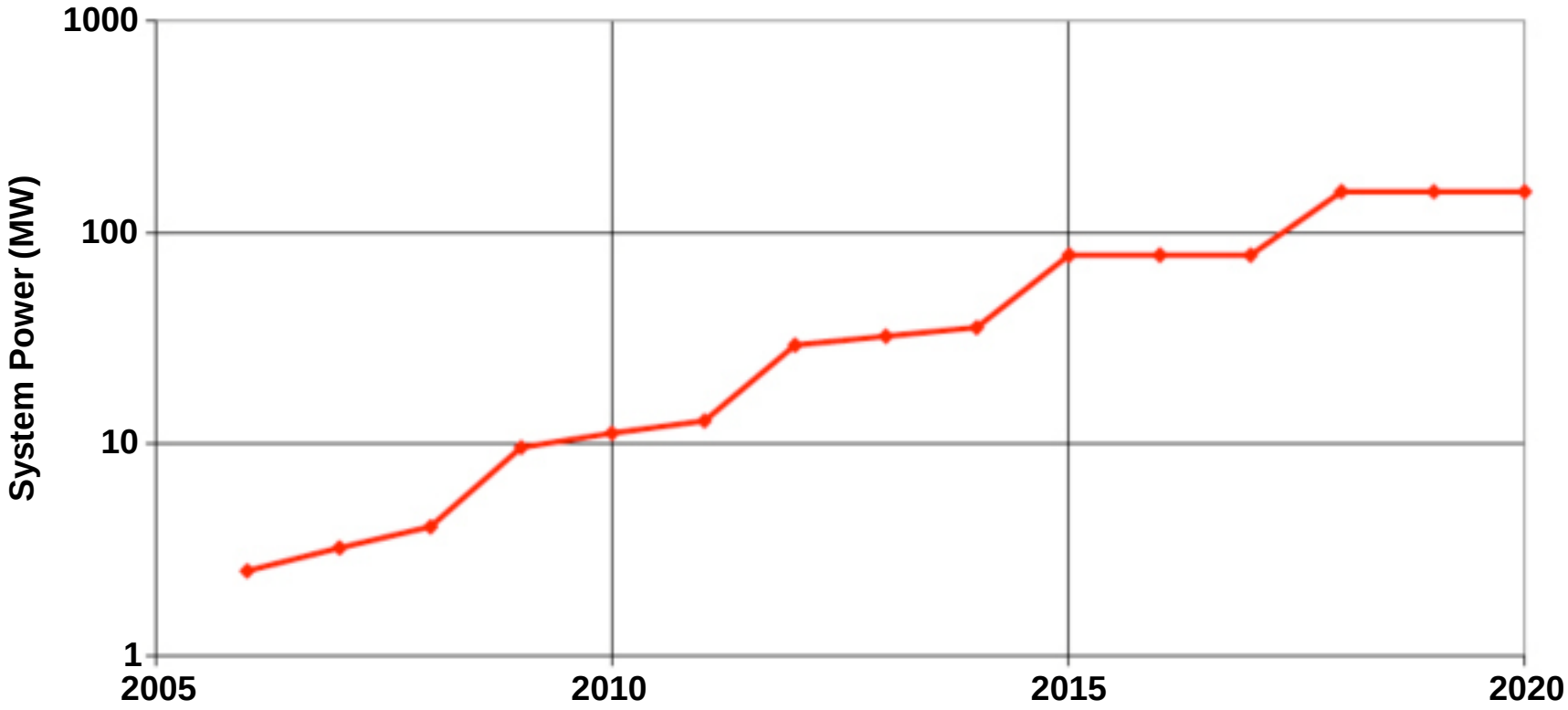
- **Commissioned by DARPA to explore the challenges for Exaflop computing (Kogge et al.)**
- **Two models for future performance growth**
  - **Simplistic: ITRS roadmap; power for memory grows linear with # of chips; power for interconnect stays constant**
  - **Fully scaled: same as simplistic, but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



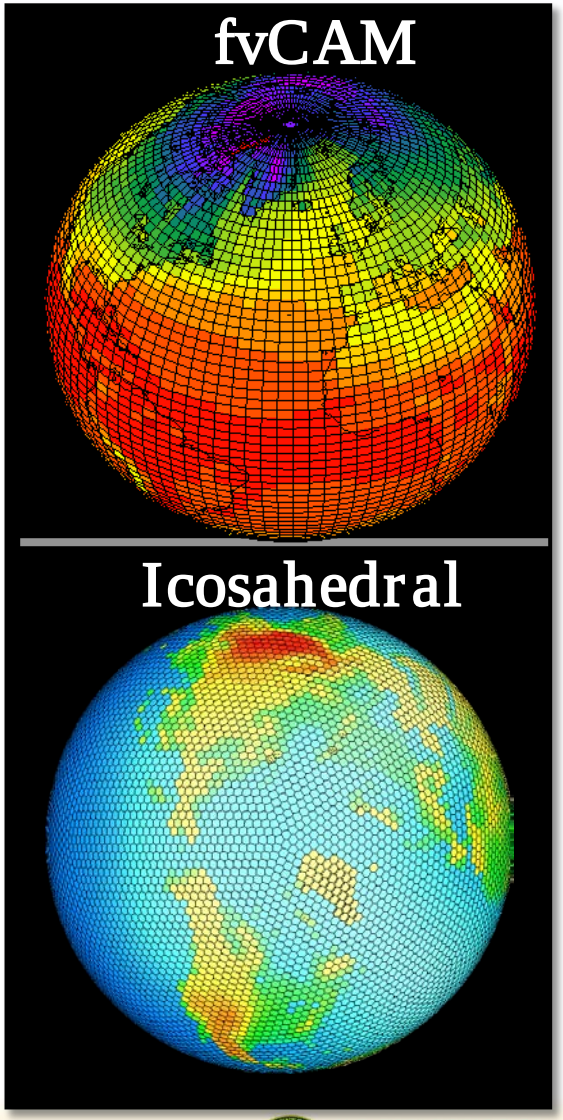
From Peter Kogge, DARPA Exascale Study

Exascale Technology Challenges

- 1B + parallelism
- Programming model
- Limit system power consumption to about 20 MW
- New memory technologies to reduce power consumption and improve bandwidth (e.g stacked memory)
- New interconnects (e.g. silicon-photonics)
- RAS

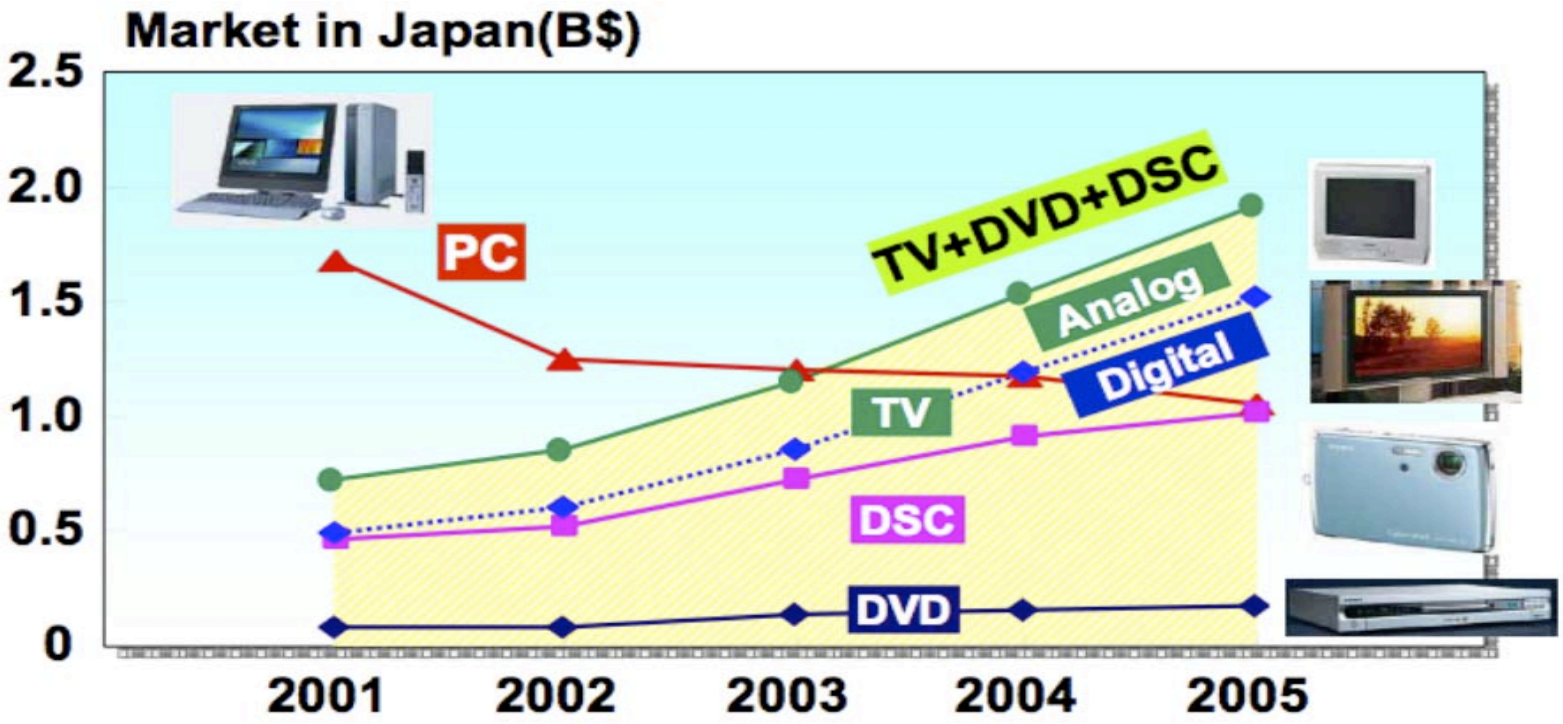
Green Flash: Ultra-Efficient Climate Modeling

- 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?"
- Goal is to influence the HPC industry by evaluating a prototype design.

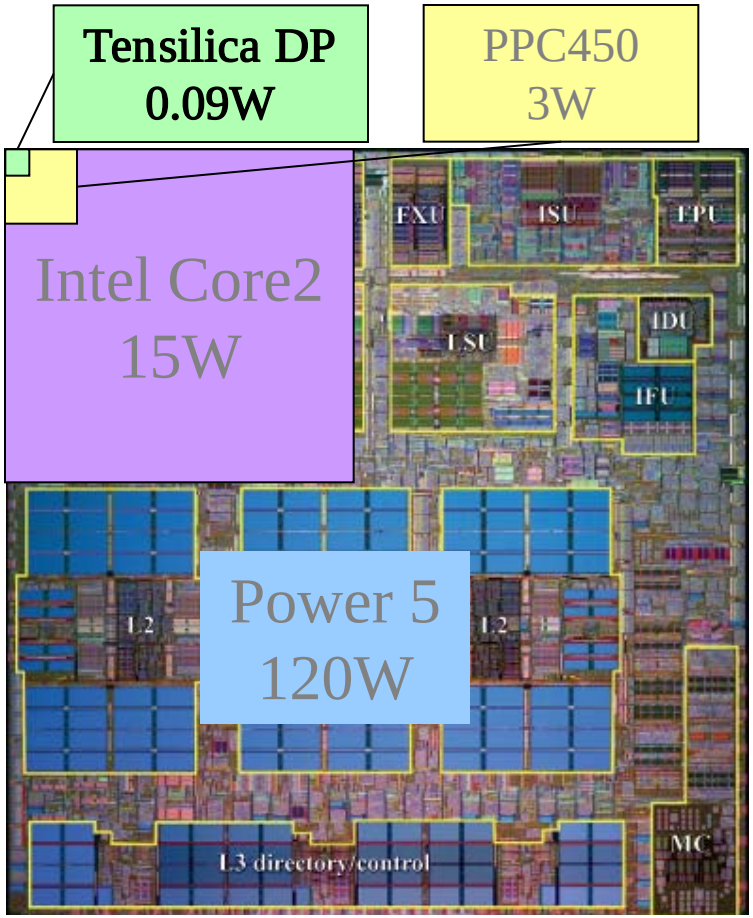


Processor Technology Trend

- 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



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

Summary on Green Flash

- **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**

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
  - http://parlab.eecs.berkeley.edu/
- NERSC System Architecture Group
  - http://www.nersc.gov/projects/SDSA
- LBNL Future Technologies Group
  http://crd.lbl.gov/ftg