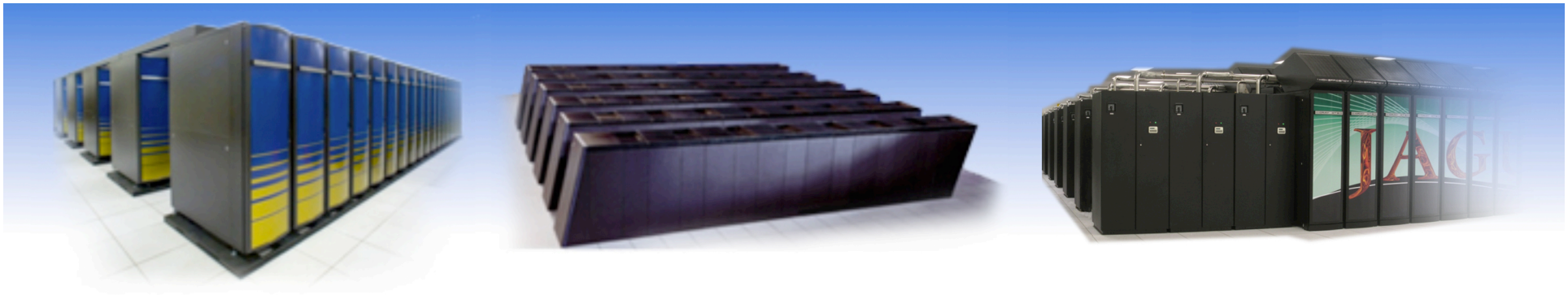


What Supercomputers Do, and What Supercomputers Still Can't Do

Horst Simon

Lawrence Berkeley National Laboratory and UC Berkeley

10 th DOE ACTS Collection Workshop

August 20, 2009

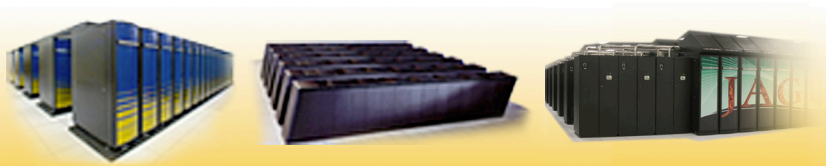
BERKELEY LAB

U.S. DEPARTMENT OF ENERGY

Office of Science

Overview

- Introducing Computing Sciences at Berkeley Lab
- Current Trends in Supercomputing (High-End Computing)
- What Supercomputers Do
- What Supercomputers Still Can't Do



Computing Sciences at Berkeley Lab (LBNL)

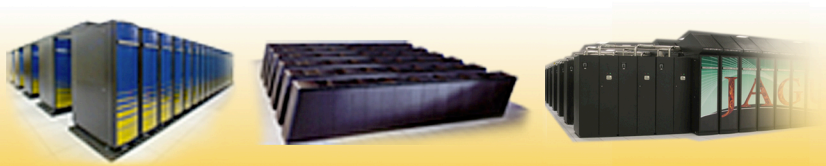
- Mostly funded by DOE, Office of Science, annual budget $100M, more than 200 staff
  - NERSC Center
  - ESnet
  - Computational Research (including SciDAC projects)
- Supports open, unclassified, basic research
- Close collaborations between UC Berkeley and LBNL in applied mathematics, computer science, and computational science

BERKELEY LAB

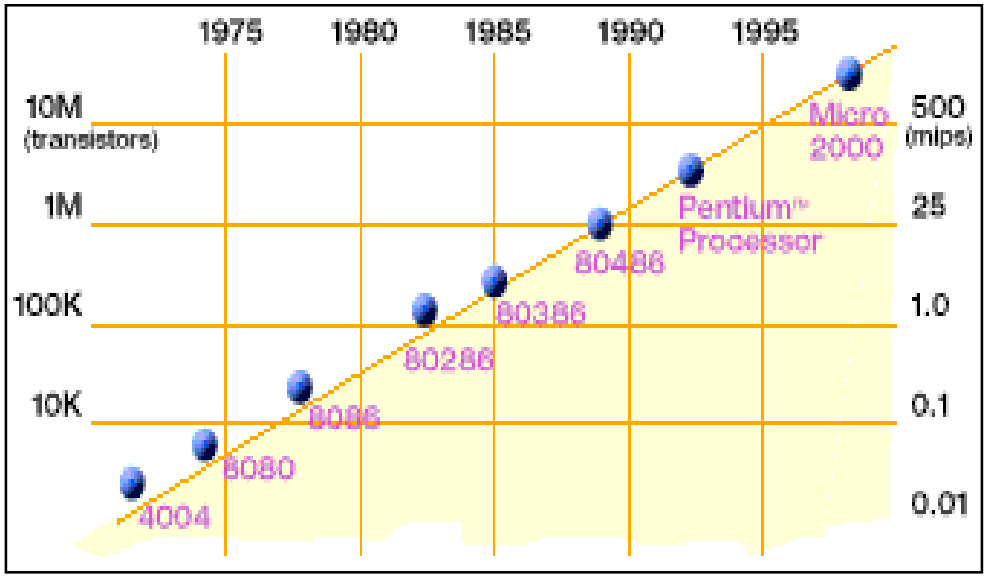
U.S. DEPARTMENT OF ENERGY
Office of Science

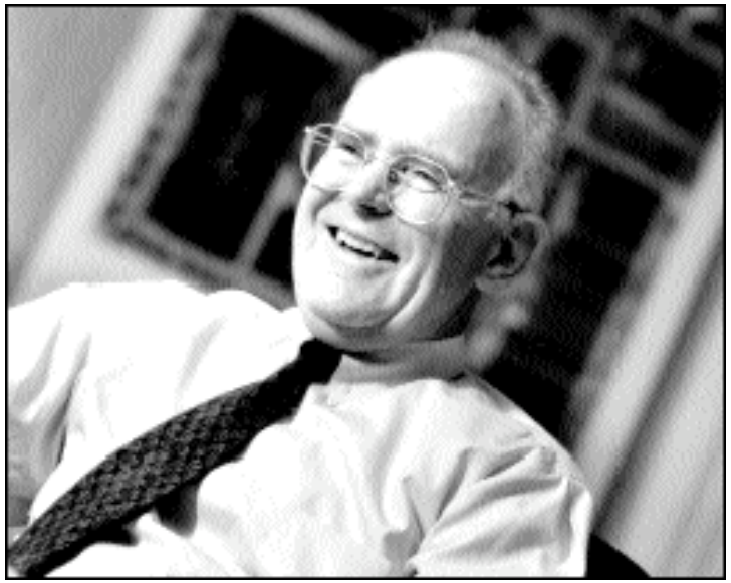
Overview

- Introducing Computing Sciences at Berkeley Lab
- Current Trends in Supercomputing (High-End Computing)
- What Supercomputers Do
- What Supercomputers Still Can't Do



Technology Trends: Microprocessor Capability





2X transistors/chip every 1.5 years
Called "Moore's Law"

Microprocessors have become smaller, denser, and more powerful.

Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.

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!**

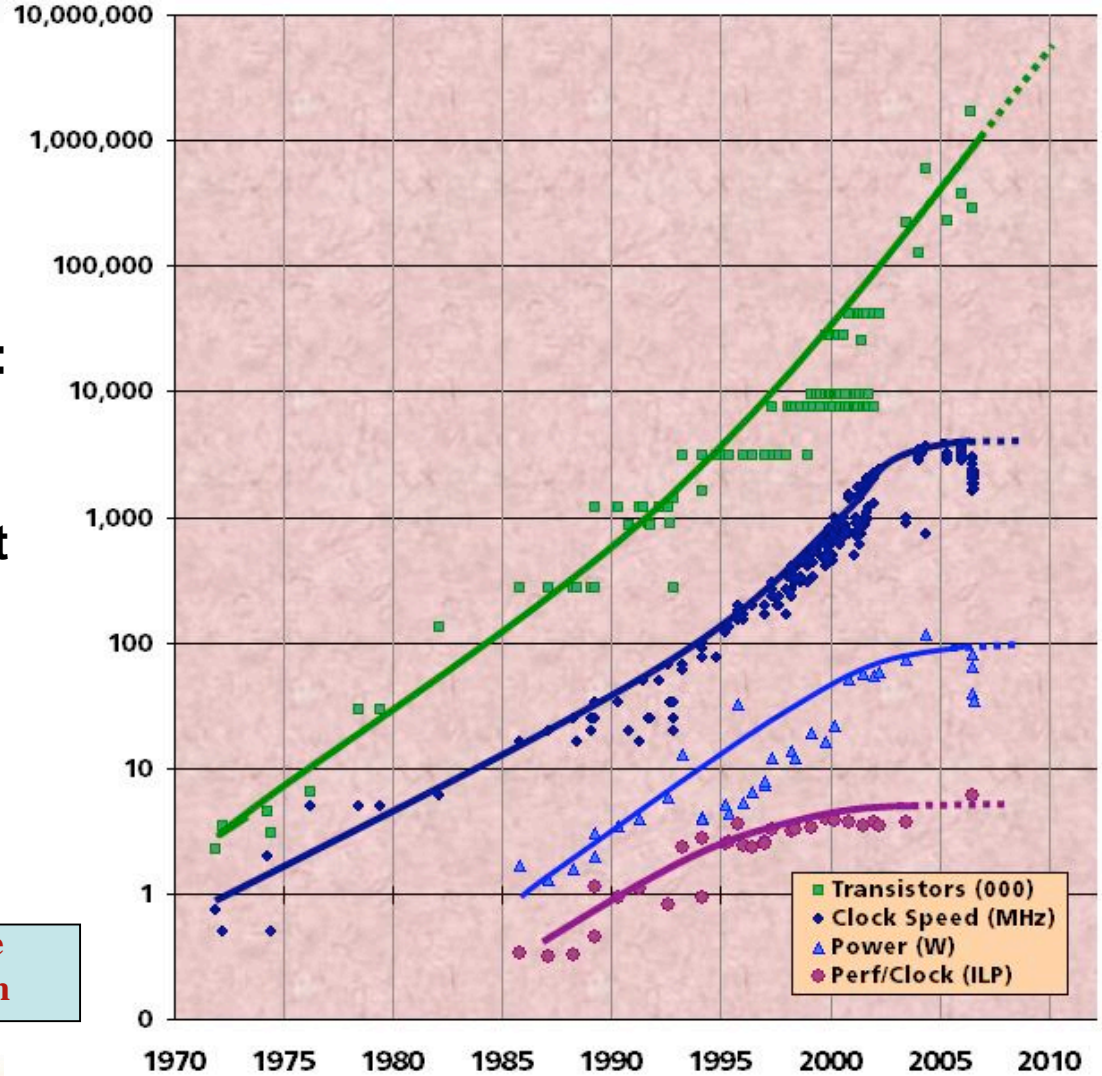
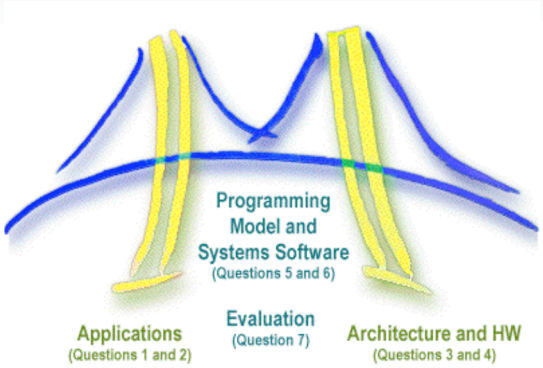
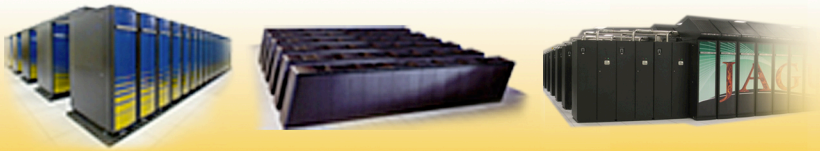


Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith

Is Multicore the Correct Response?

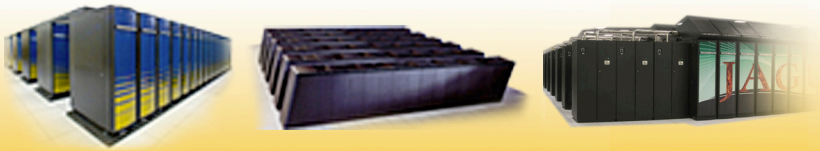


- **"The View from Berkeley",** http://view.eecs.berkeley.edu/wiki/Main_Page
- **Kurt Keutzer:** "This shift toward increasing parallelism is not a triumphant stride forward based on breakthroughs in novel software and architectures for parallelism; instead, this plunge into parallelism is actually a retreat from even greater challenges that thwart efficient silicon implementation of traditional uniprocessor architectures."
- **David Patterson:** "Industry has already thrown the hail-Mary pass. . . But nobody is running yet."



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



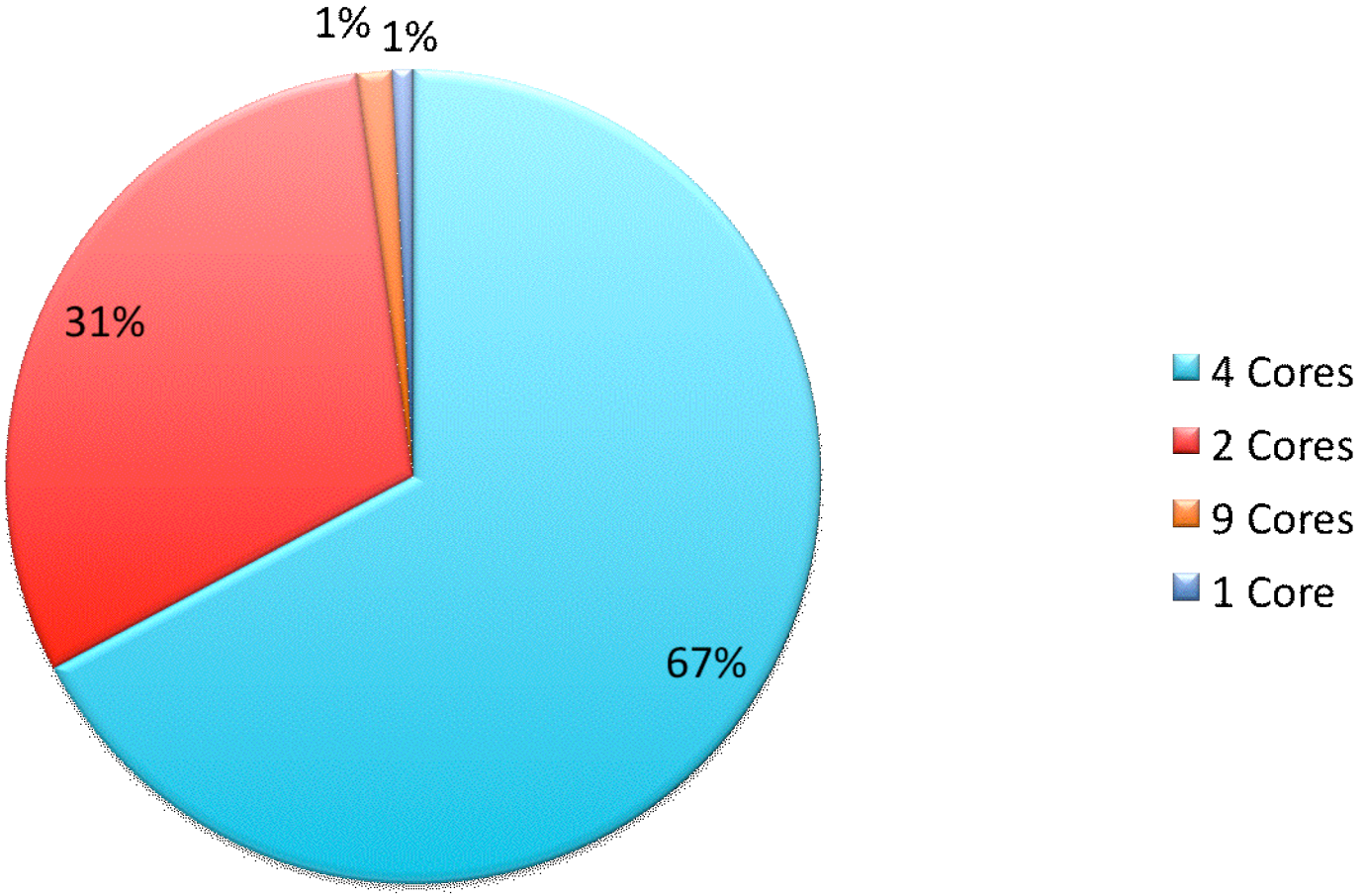
Cray XT5 at ORNL -- 1 Pflop/s in November 2008

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**

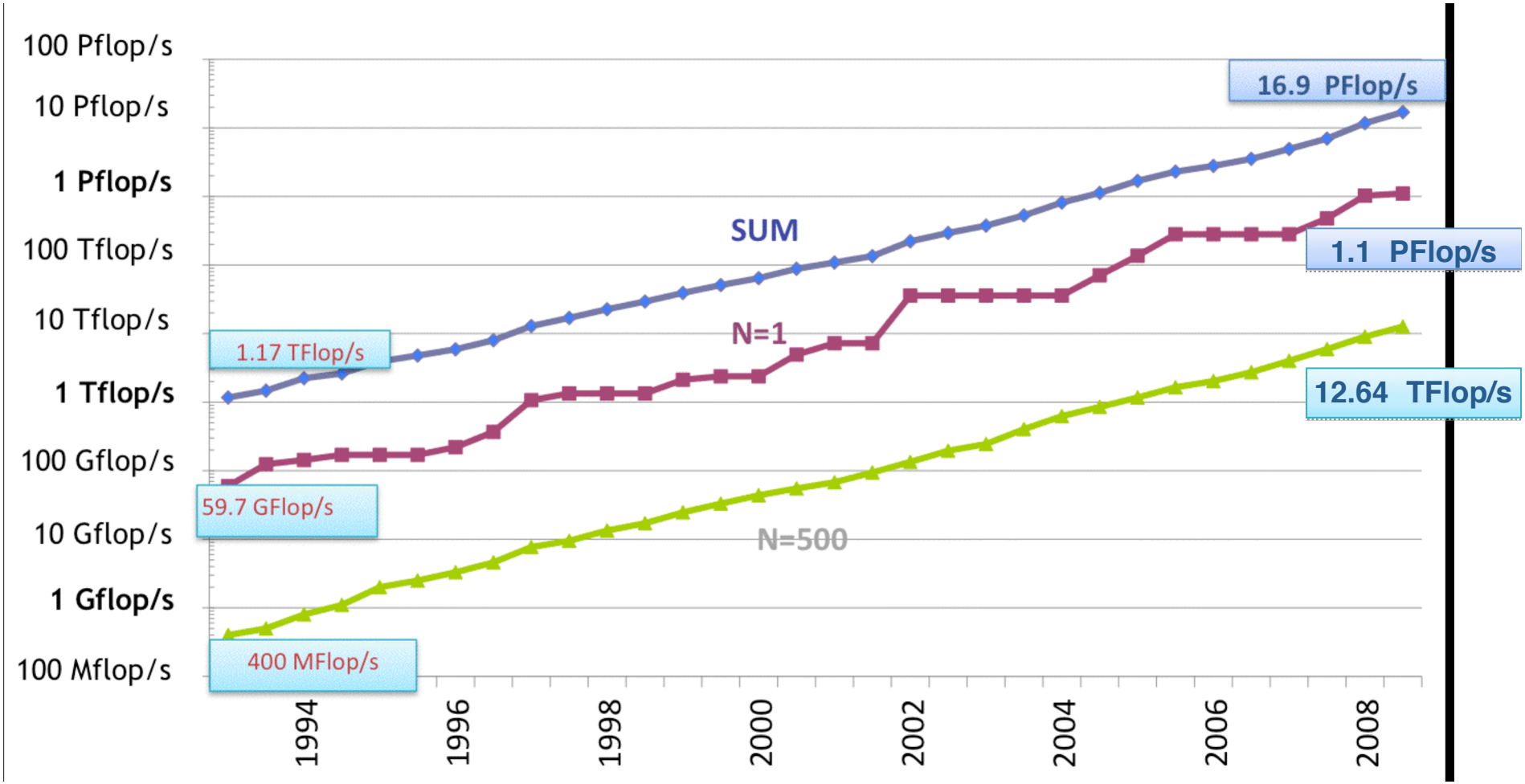
The systems will be combined after acceptance of the new XT5 upgrade. Each system will be linked to the file system through 4x-DDR Infiniband

U.S. DEPARTMENT OF ENERGY
Office of Science

Cores per Socket

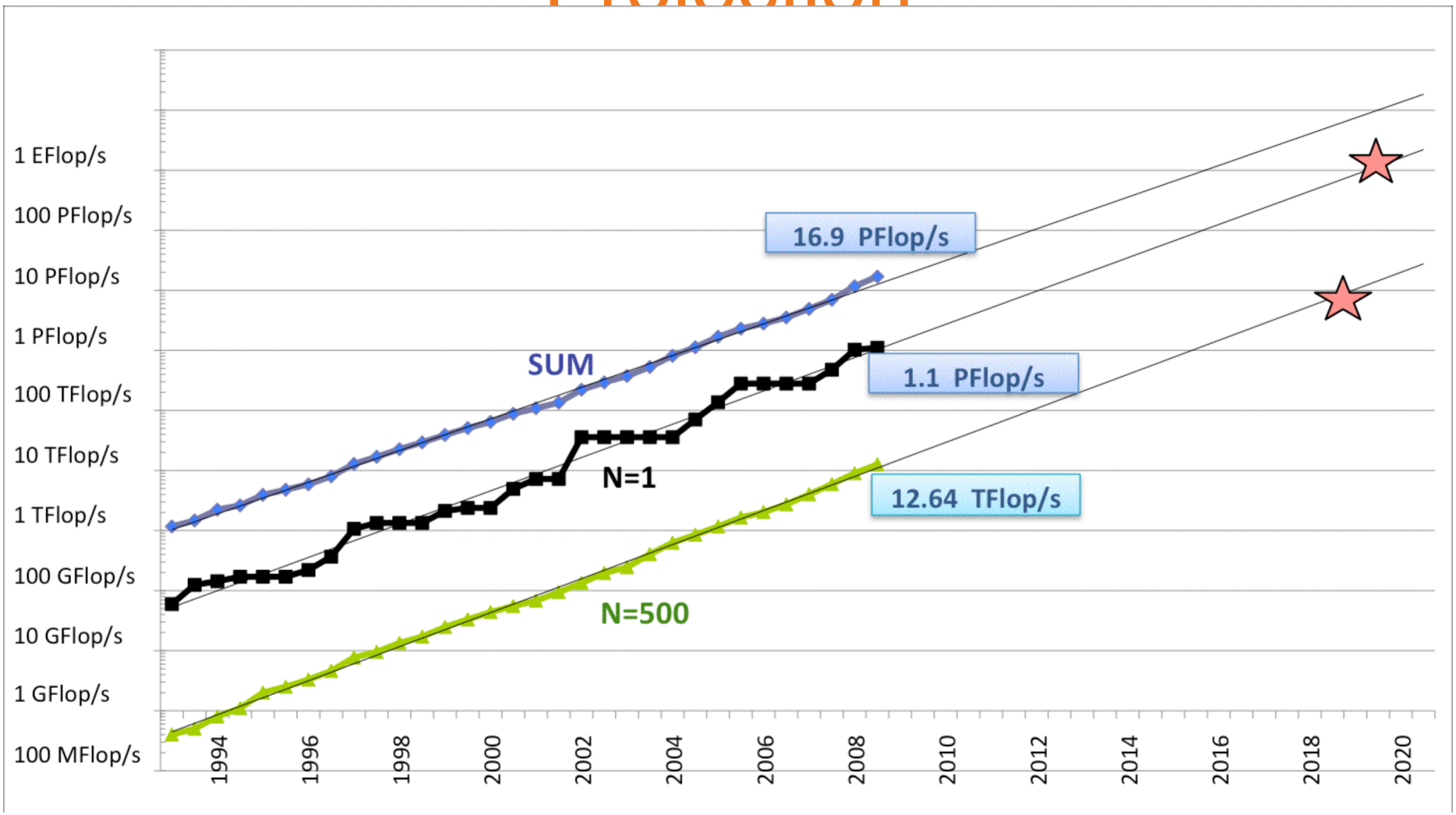


Performance Development



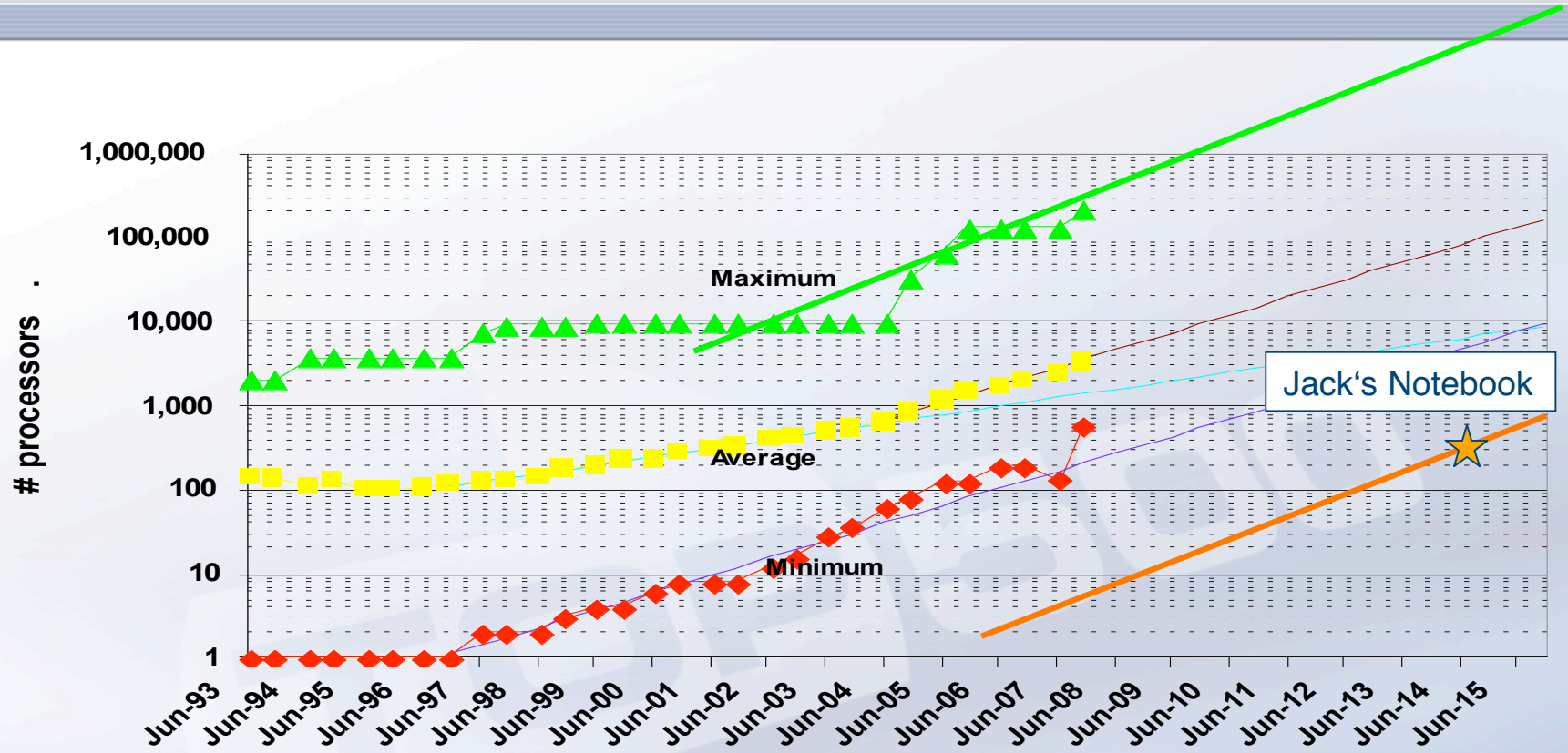
TOP500® SUPERCOMPUTER SITES

Performance Development Projection



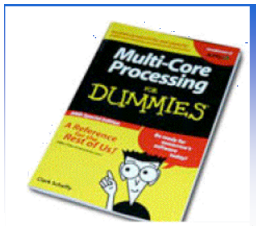
TOP500® SUPERCOMPUTER SITES

Concurrency Levels



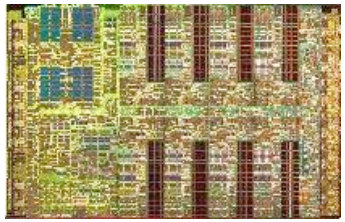
Moore's Law reinterpreted

- 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



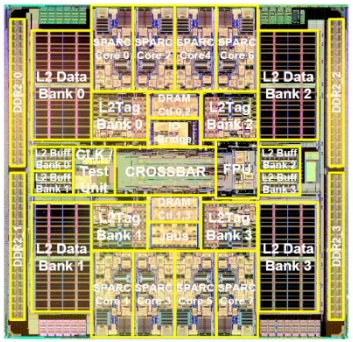
Multicore comes in a wide variety

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

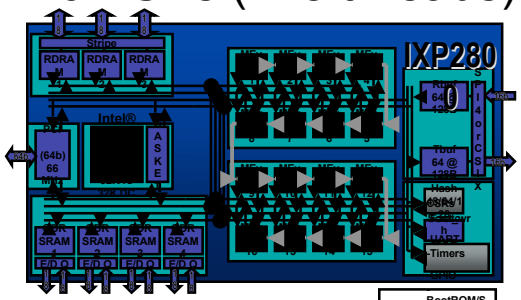


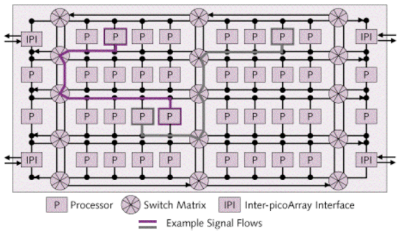
IBM Cell
1 GPP (2 threads)
8 ASPs

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



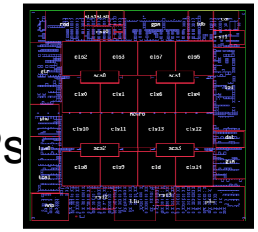
Sun Niagara
8 GPP cores (32 threads)

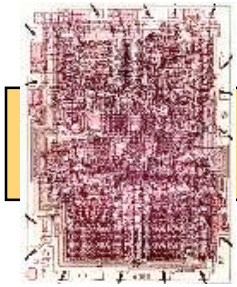




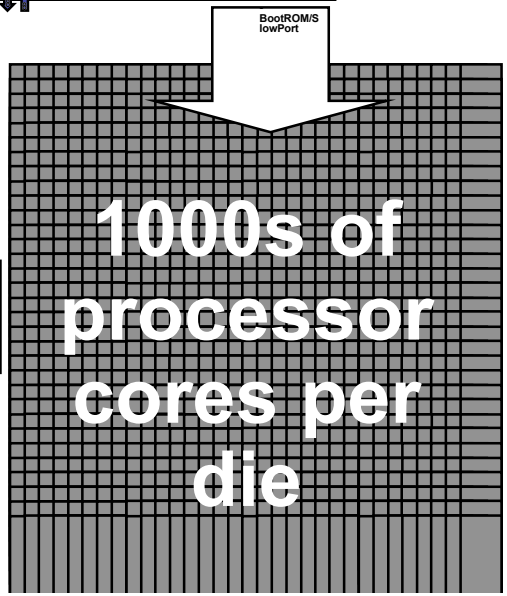
Picochip DSP
1 GPP core
248 ASPs

Cisco CRS-1
188 Tensilica GPPs





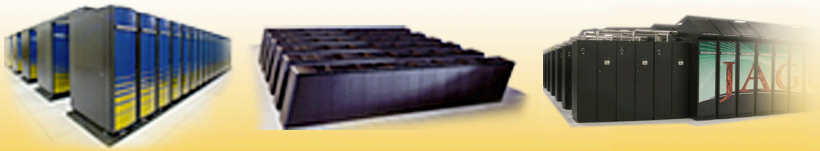
Intel 4004 (1971):
4-bit processor,
2312 transistors,
~100 KIPS,
10 micron PMOS,
11 mm^2 chip



"The Processor is the new Transistor" [Rowen]

Trends for the next five years up to 2014

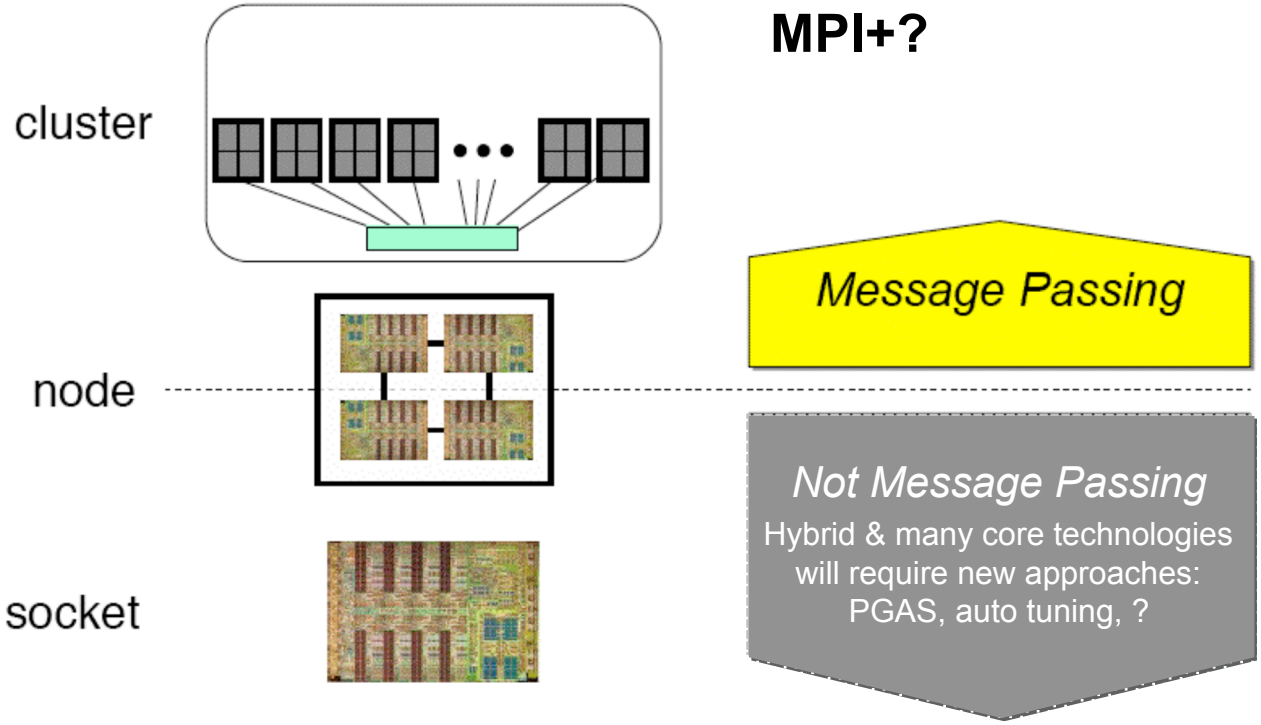
- After period of rapid architectural change we will likely settle on a future standard processor architecture
- A good bet: Intel will continue to be a market leader
- Impact of this disruptive change on software and systems architecture not clear yet



A Likely Future Scenario (2014)

System: cluster + many core node

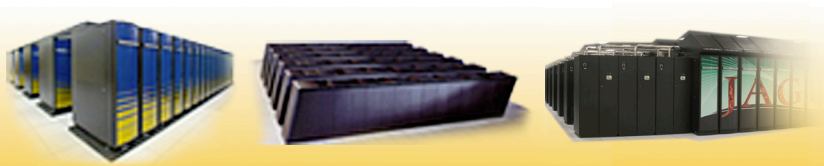
Programming model: MPI+?



after Don Grice, IBM, Roadrunner Presentation, ISC 2008

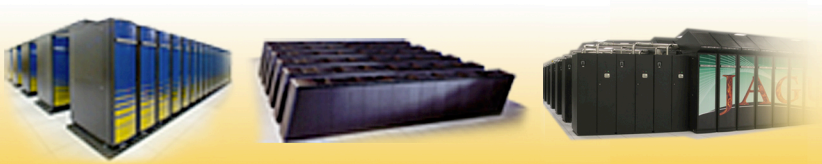
What will be the "?" in MPI+?

- Likely candidates are
  - PGAS languages
  - Autotuning
  - A wildcard from commercial space

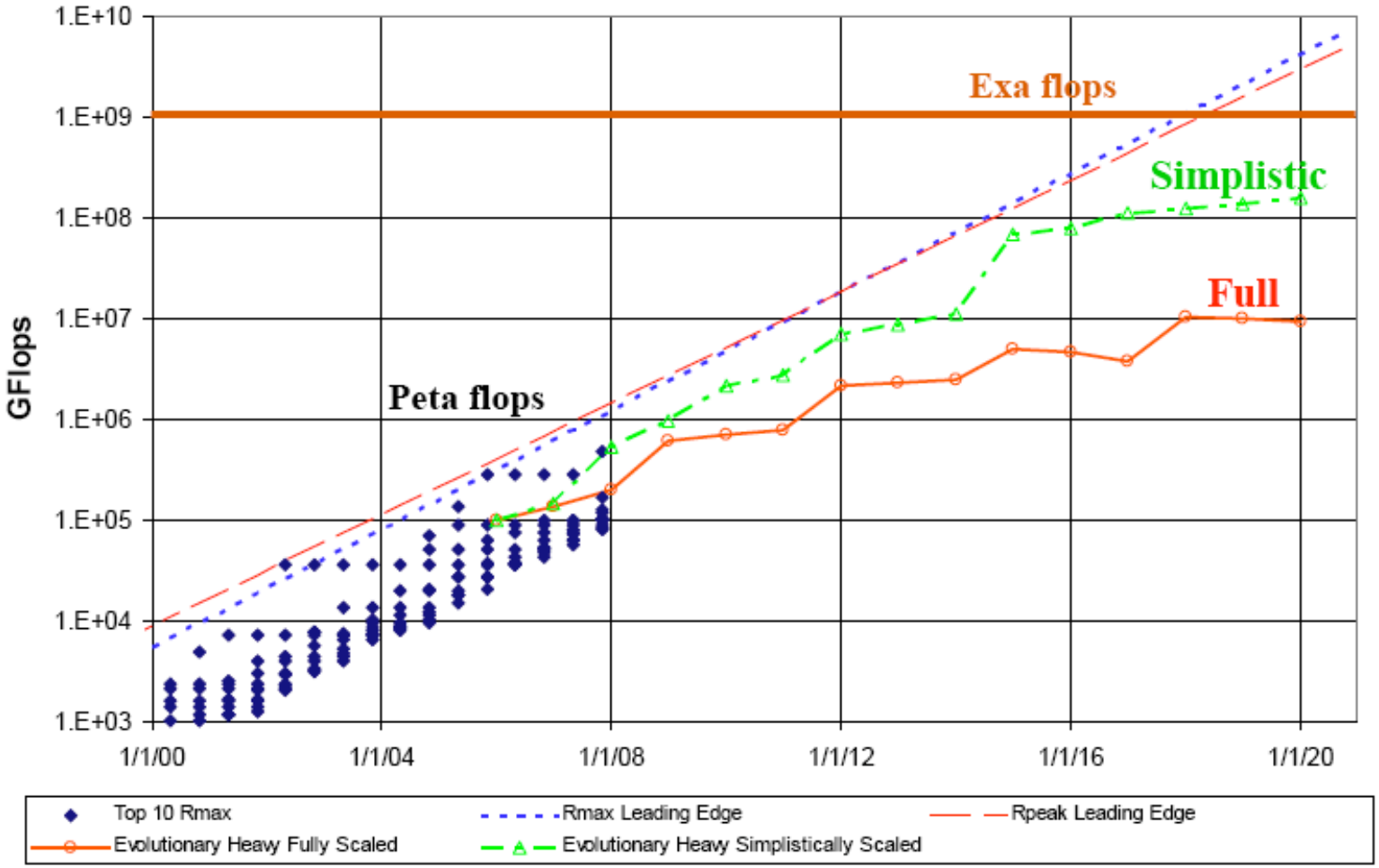


DARPA Exascale Study

- 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

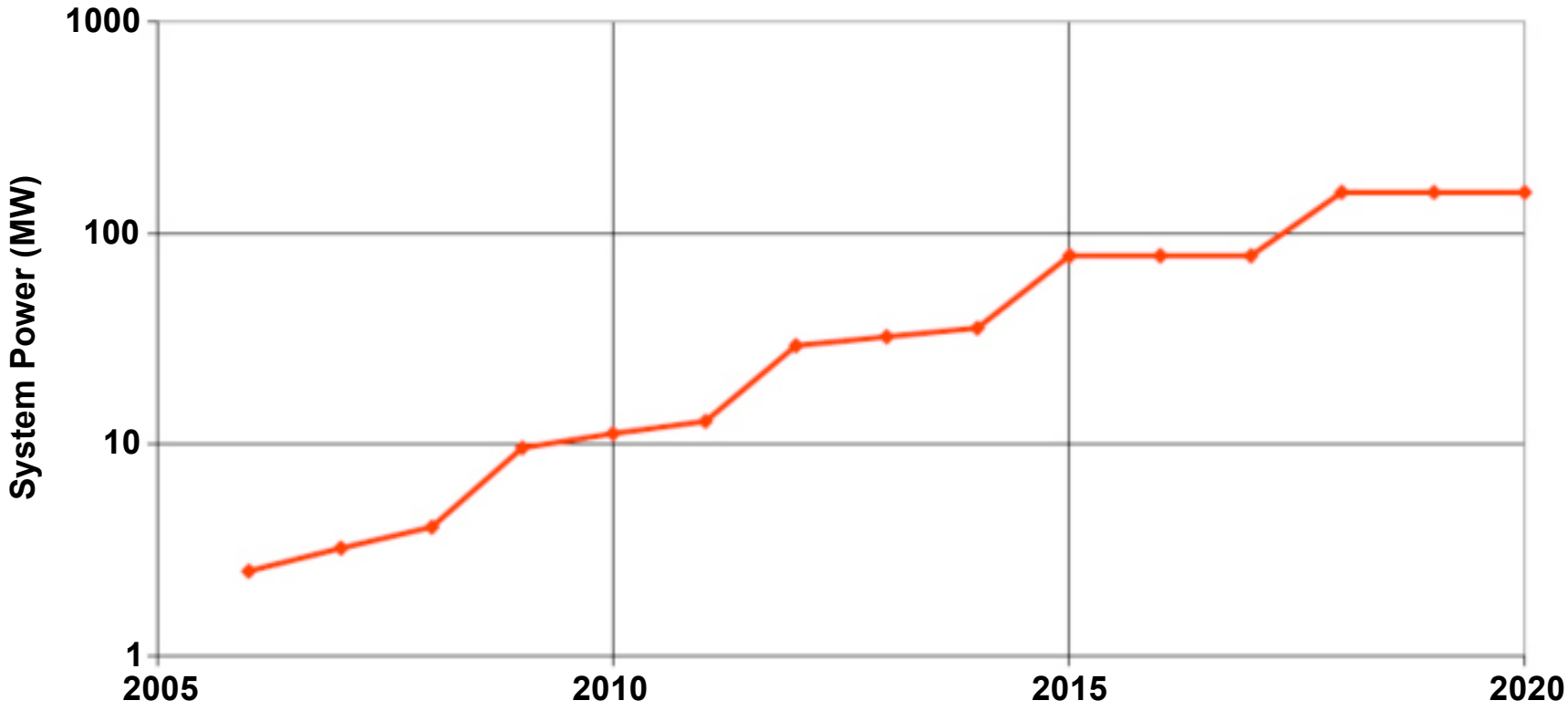


We won't reach Exaflops with this approach



From Peter Kogge, DARPA Exascale Study

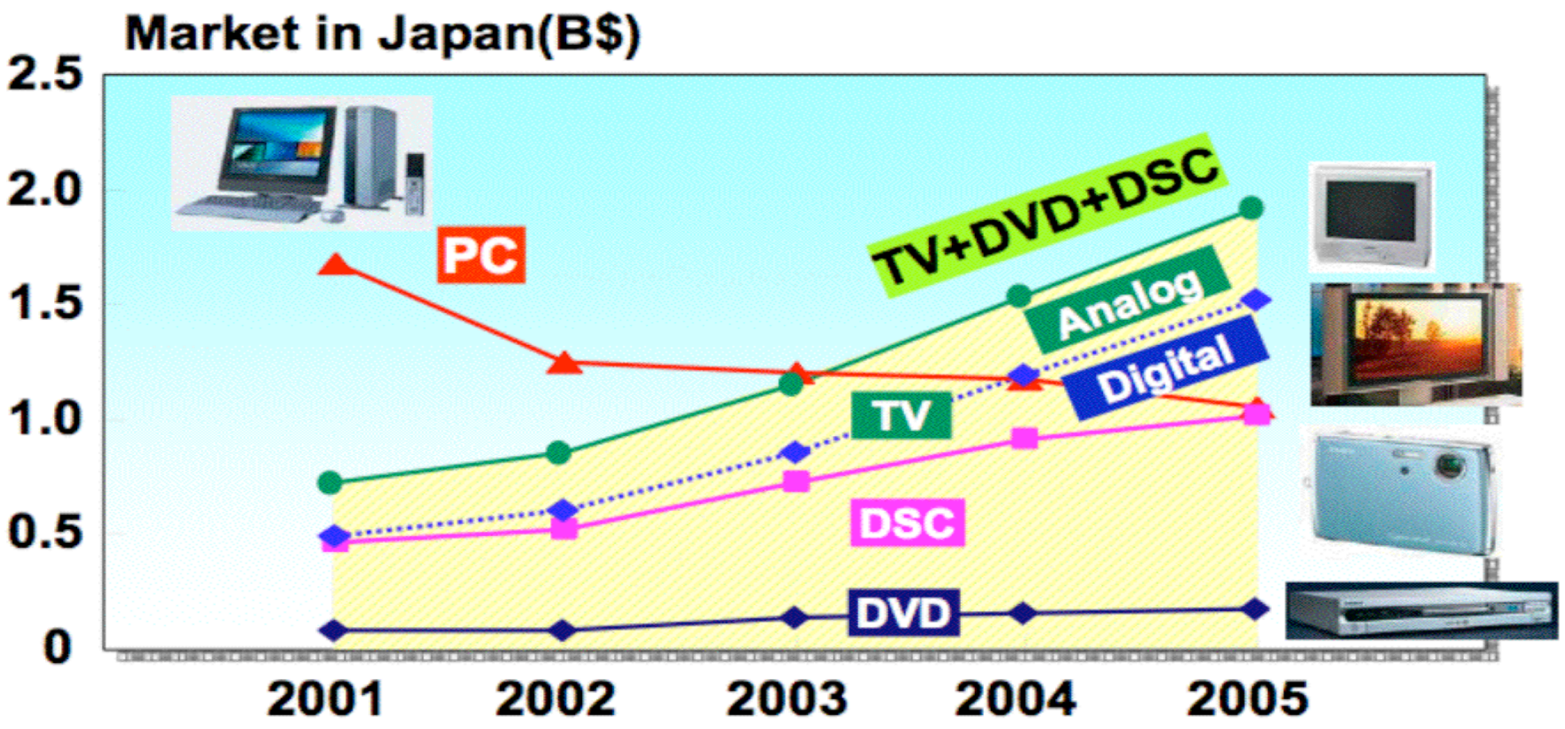
… and the power costs will still be staggering



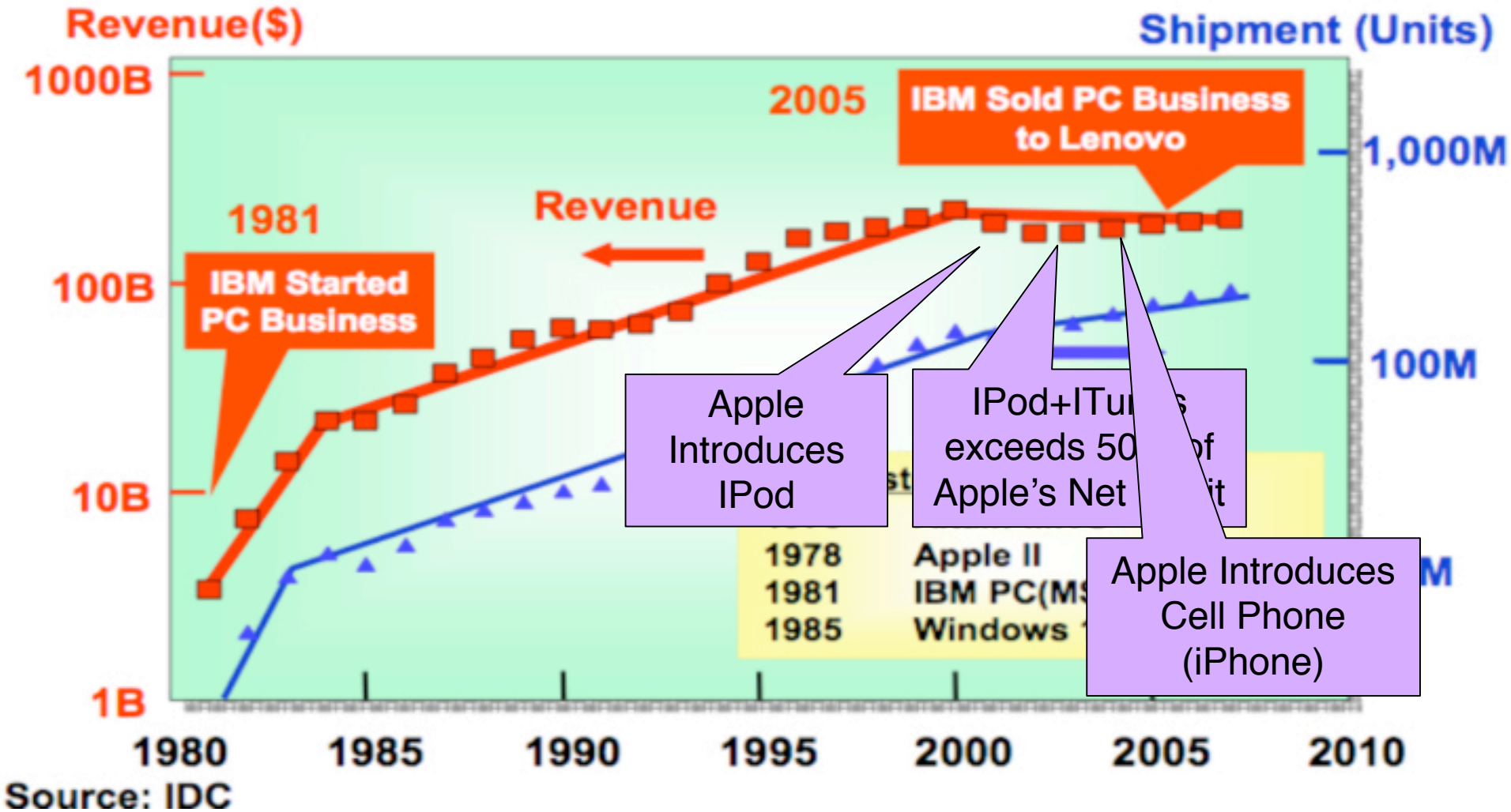
From Peter Kogge,
DARPA Exascale Study

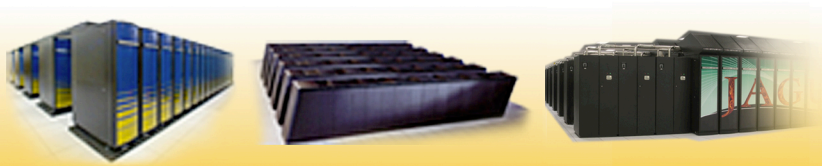
Processor Technology Trend

- 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



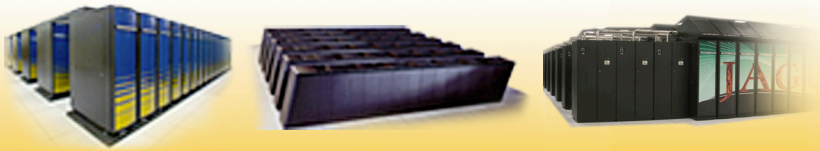
Consumer Electronics has Replaced PCs as the Dominant Market Force in CPU Design!!





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.**



Ultra-Efficient "Green Flash" Computing at NERSC: 100x over Business as Usual

Radically change HPC system development via application-driven hardware/software co-design

- ***Achieve 100x power efficiency and 100x capability of mainstream HPC approach for targeted high-impact applications***
- Accelerate development cycle for exascale HPC systems
- Approach is applicable to numerous scientific applications
- Proposed pilot application: Ultra-high resolution climate change simulation

Path to Power Efficiency

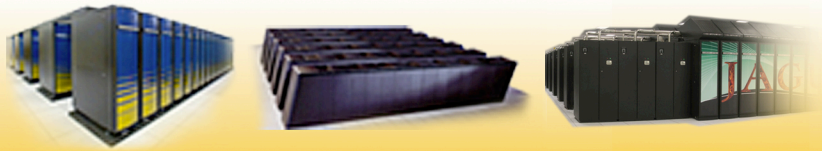
Reducing Waste in Computing

- **Examine methodology of low-power embedded computing market**
  - **optimized for low power, low cost and high computational efficiency**

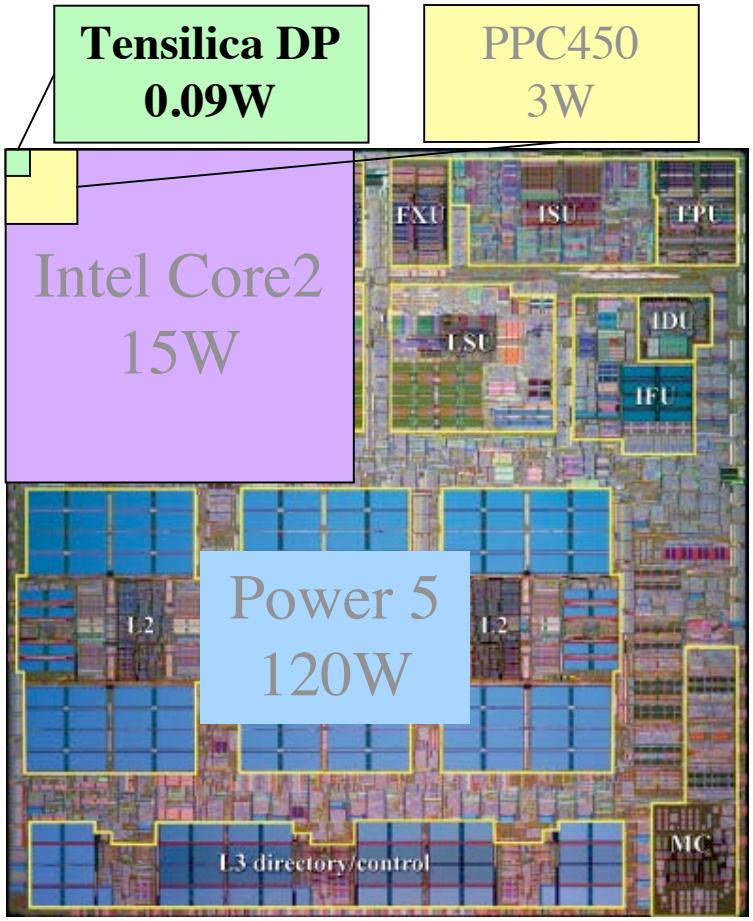
"Years of research in low-power embedded computing have shown only one design technique to reduce power: reduce waste."

— Mark Horowitz, Stanford University & Rambus Inc.

- **Sources of waste**
  - **Wasted transistors (surface area)**
  - **Wasted computation (useless work/speculation/stalls)**
  - **Wasted bandwidth (data movement)**
  - **Designing for serial performance**

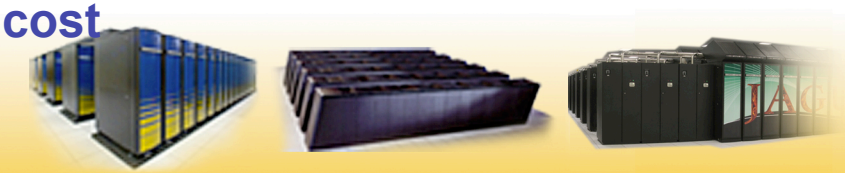


Design for Low Power: More Concurrency

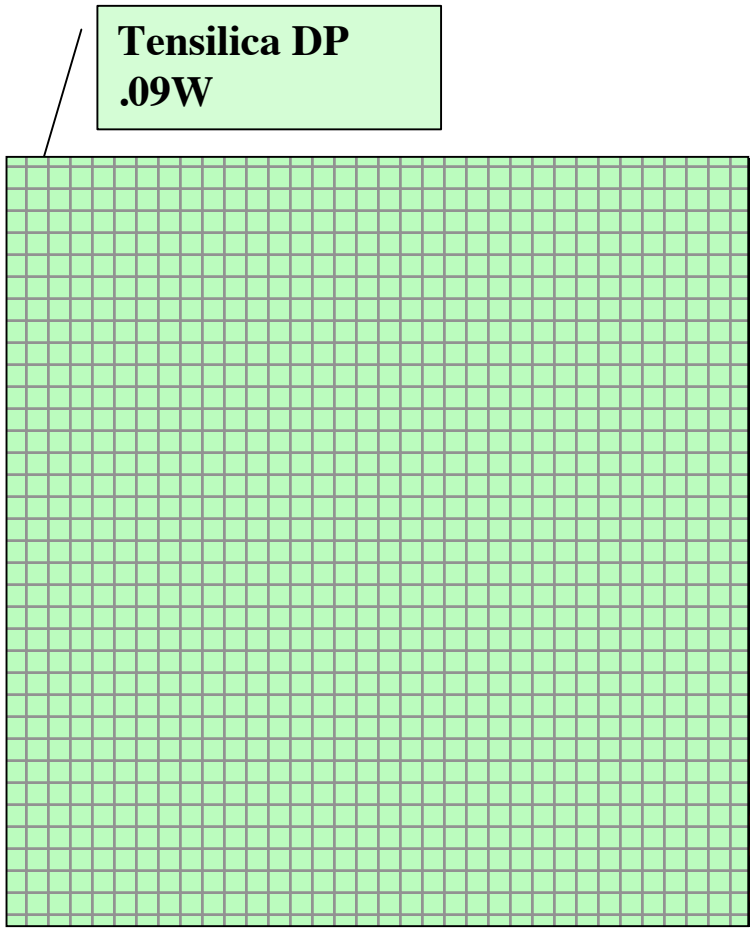


- Cubic power improvement with lower clock rate due to V^2F
- Slower clock rates enable use of simpler cores
- Simpler cores use less area (lower leakage) and reduce cost
- Tailor design to application to <u>reduce waste</u>

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



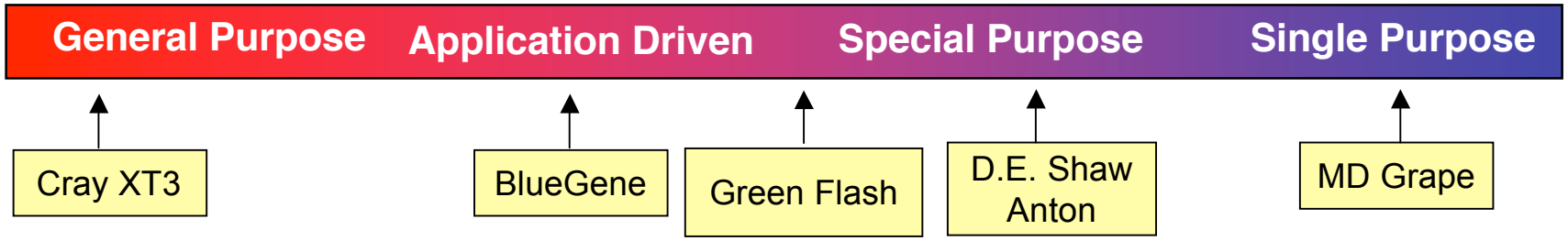
Low Power Design Principles



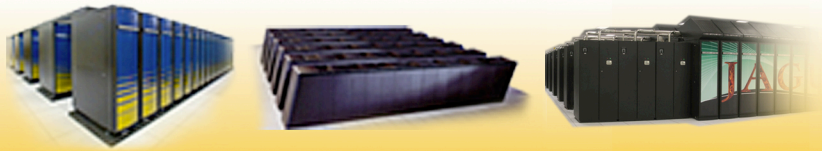
- IBM Power5 (server)
  - 120W@1900MHz
  - Baseline
- Intel Core2 sc (laptop) :
  - 15W@1000MHz
  - *4x more FLOPs/watt than baseline*
- IBM PPC 450 (BG/P - low power)
  - 0.625W@800MHz
  - 90x more
- Tensilica XTensa (Moto Razor) :
  - 0.09W@600MHz
  - 400x more

Even if each core operates at 1/3 to 1/10th efficiency of largest chip, you can pack 100s more cores onto a chip and consume 1/20 the power

Customization Continuum: Green Flash



- Application-driven does NOT necessitate a special purpose machine
- MD-Grape: Full custom ASIC design
  - 1 Petaflop performance for one application using 260 kW for $9M
- D.E. Shaw Anton System: Full and Semi-custom design
  - Simulate 100x–1000x timescales vs any existing HPC system (~200kW)
- Application-Driven Architecture (Green Flash): Semicustom design
  - Highly programmable core architecture using C/C++/Fortran
  - Goal of 100x power efficiency improvement vs general HPC approach
  - Better understand how to build/buy application-driven systems
  - Potential: 1km-scale model (~200 Petaflops peak) running in O(5 years)

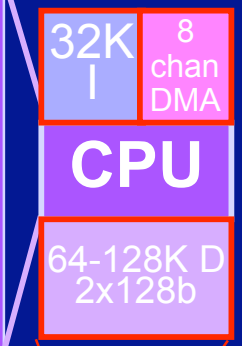


Climate System Design Concept

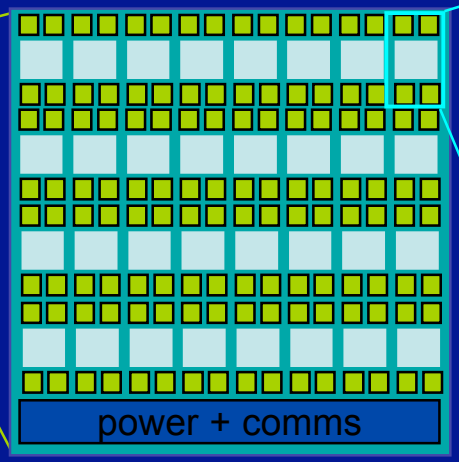
Strawman Design Study

VLIW CPU:

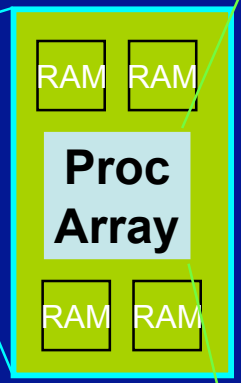
- 128b load-store + 2 DP MUL/ADD + integer op/ DMA per cycle:
- Synthesizable at 650MHz in commodity 65nm
- 1mm² core, 1.8-2.8mm² with inst cache, data cache data RAM, DMA interface, 0.25mW/MHz
- Double precision SIMD FP : 4 ops/cycle (2.7GFLOPs)
- Vectorizing compiler, cycle-accurate simulator, debugger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



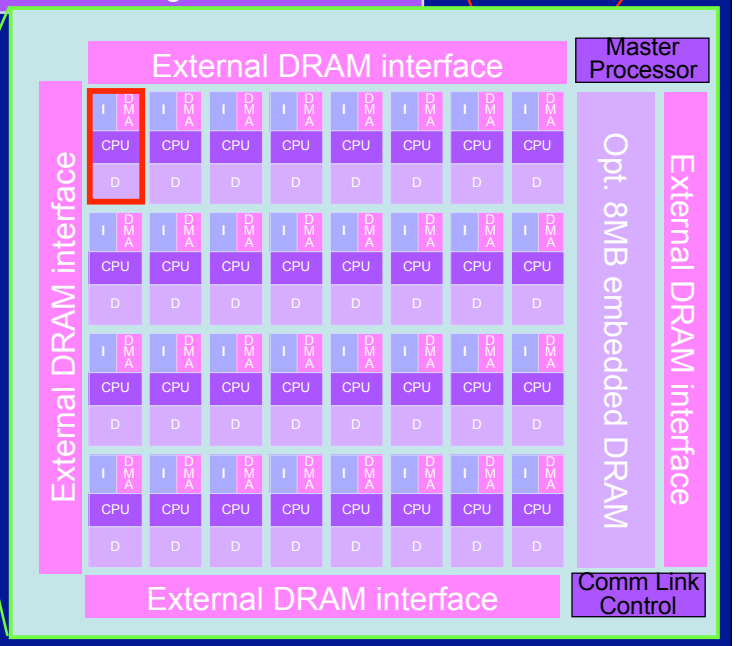
100 racks @ ~25KW



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

Green Flash Strawman System Design

We examined three different approaches (in 2008 technology)

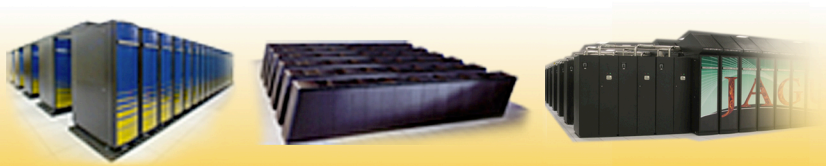
Computation .015°X.02°X100L: 10 PFlops sustained, ~200 PFlops peak

- AMD Opteron: **Commodity approach, lower efficiency for scientific applications offset by cost efficiencies of mass market**
- 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

Overview

- Introducing Computing Sciences at Berkeley Lab
- Current Trends in Supercomputing (High-End Computing)
- What Supercomputers Do
- What Supercomputers Still Can't Do

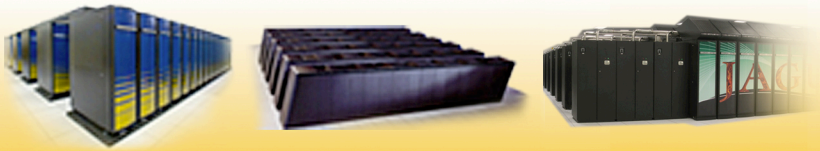


What Supercomputers Do

Introducing Computational Science and Engineering

Three Examples

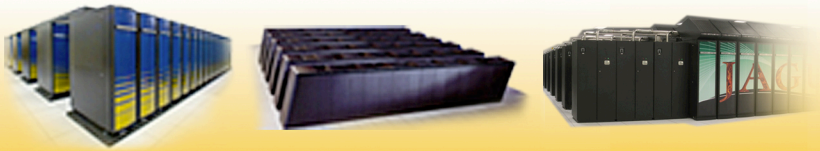
- simulation replacing experiment that is too difficult
- simulation replacing experiment that is too dangerous
- analyzing massive amounts of data with new tools



Computational Science and Engineering (CSE)

- CSE is a widely accepted label for an evolving field concerned with the science of and the engineering of systems and methodologies to solve computational problems arising throughout science and engineering
- CSE is characterized by
  - Multi - disciplinary
  - Multi - institutional
  - Requiring high-end resources
  - Large teams
  - Focus on community software
- CSE is not "just programming" (and not CS)
- Petaflop/s computing is necessary but not sufficient

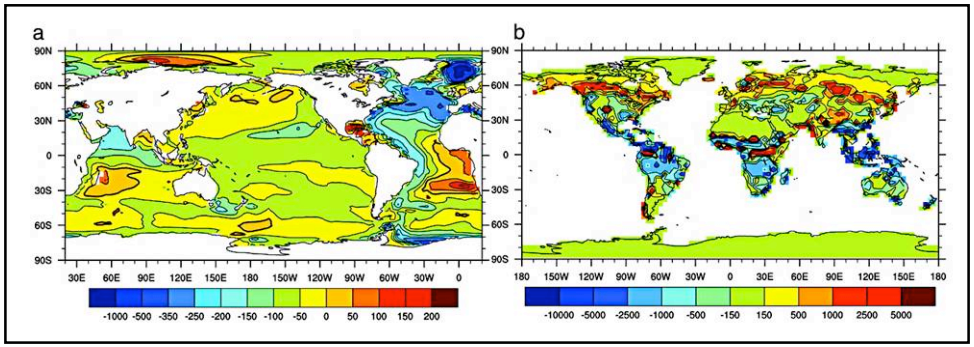
Reference: Petzold, L., *et al.*, Graduate Education in CSE, *SIAM Rev.*, 43(2001), 163-177



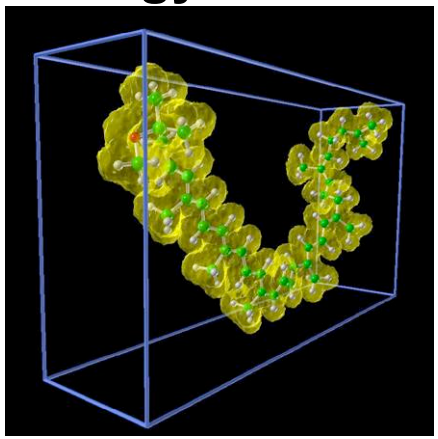
SciDAC - First Federal Program to Implement CSE

- SciDAC (Scientific Discovery through Advanced Computing) program created in 2001
  - About $50M annual funding
  - Berkeley (LBNL+UCB) largest recipient of SciDAC funding

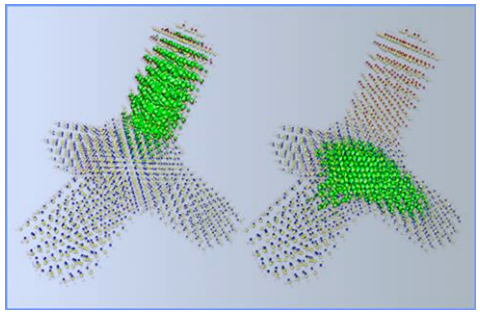
Global Climate



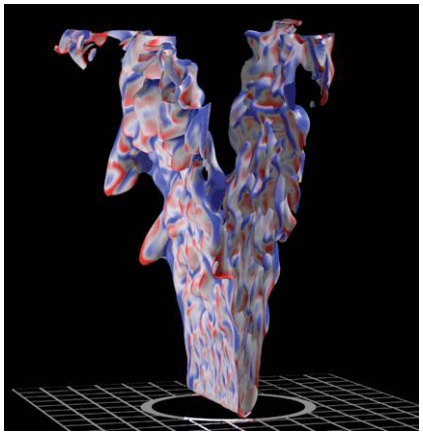
Biology



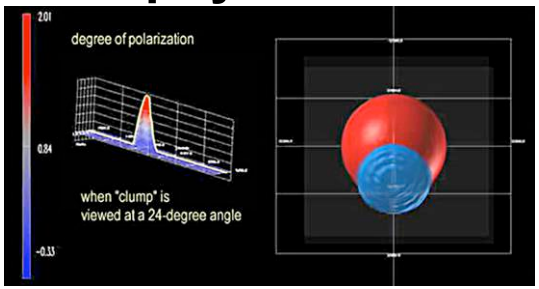
Nanoscience

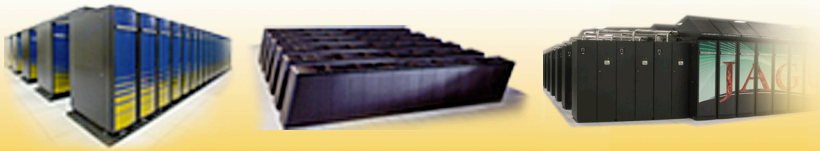


Combustion



Astrophysics



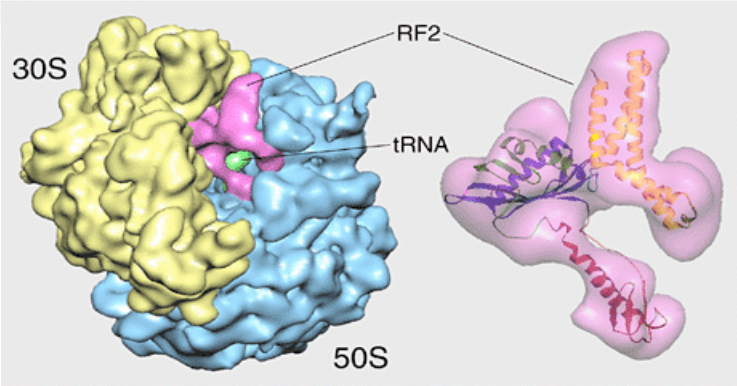


BERKELEY LAB

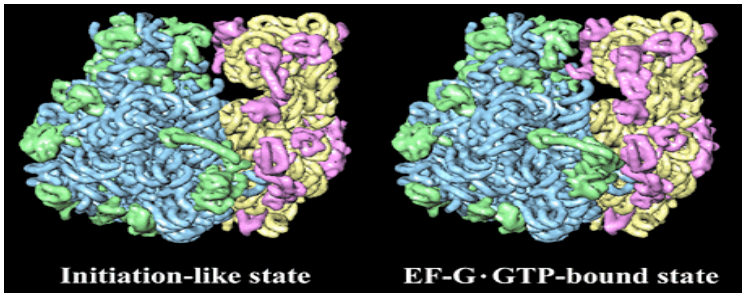
U.S. DEPARTMENT OF ENERGY
Office of Science

Cryo-EM: Significance

- **Protein structure determination is one of the building blocks for molecular biology research**
  - **Provides the mapping of subunits and active sites within a complex**
- **Standard approach is to crystallize protein**
- **However, 30% of all proteins do not crystallize or are difficult to crystallize**

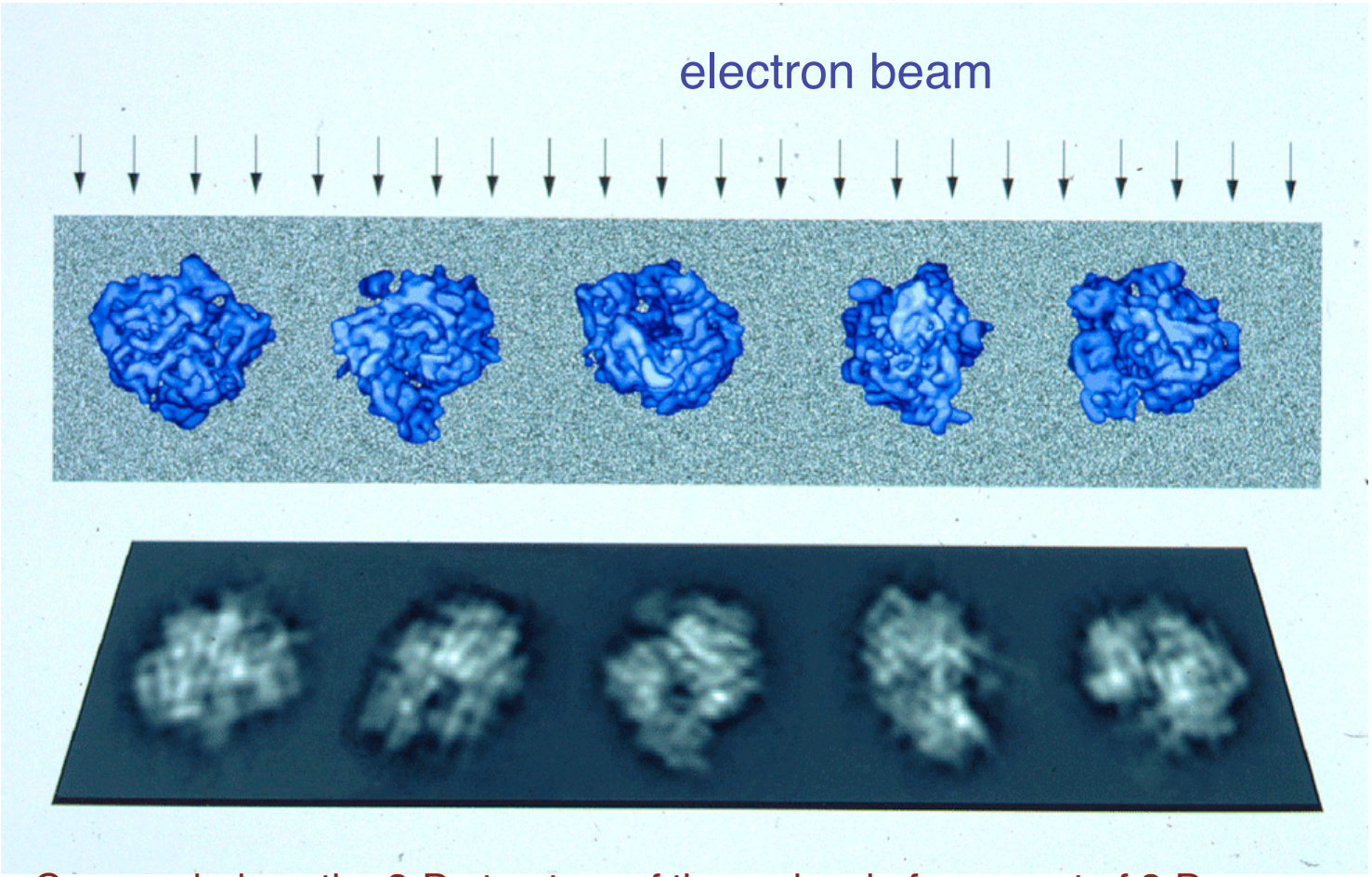


Ribosome bound with release factor RF2 in the presence of a stop codon and a P-site tRNA. (J. Frank)



Space-filling atomic models of the *E. coli* ribosome in two conformations related by the ratchet rotation. Blue and green: RNA and proteins of 50S subunit, respectively; yellow and red: RNA and proteins of the 30S subunit. While the RNA undergoes continuous elastic deformations, some of the proteins rotate and change their conformations significantly. (J. Frank)

The Reconstruction Problem

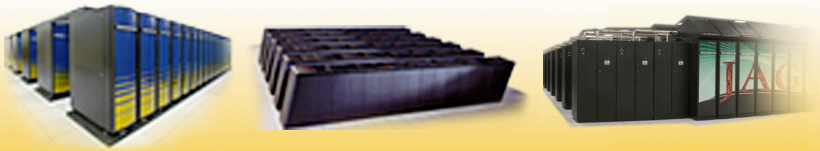


3D macromolecule

2D projections

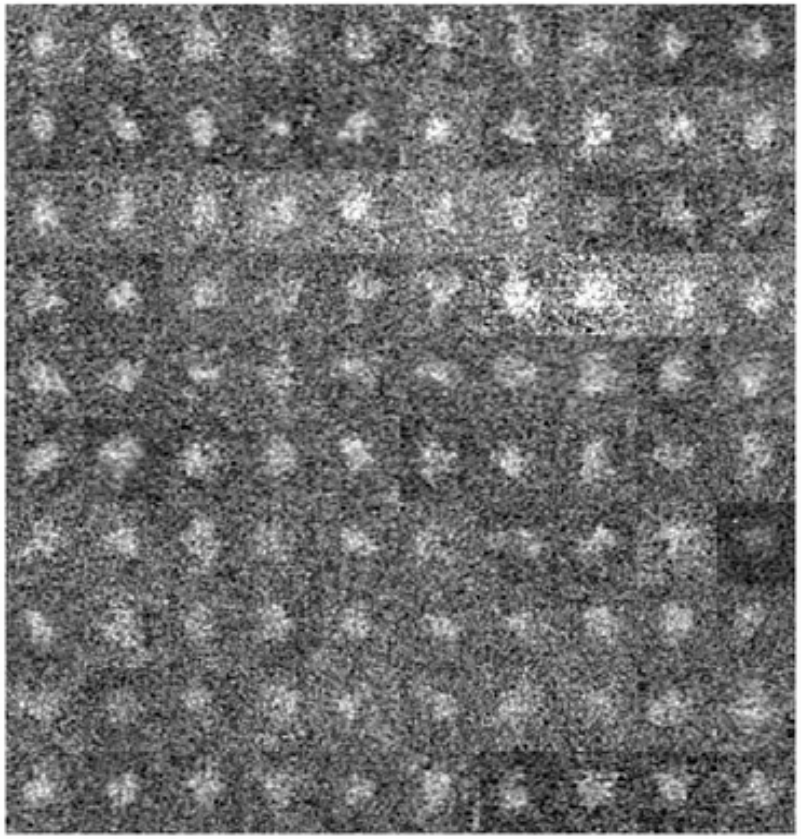
(Cryo-EM)

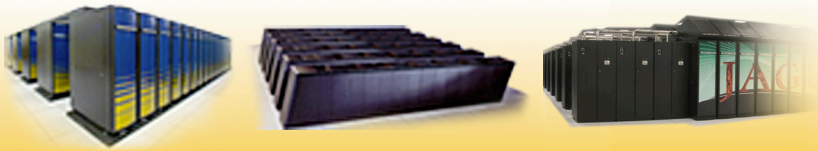
Can we deduce the 3-D structure of the molecule from a set of 2-D projection images with unknown relative orientations?



Challenge

- Nonlinear inverse problem
- Extremely noisy data
- Large volume of data
  - To make the problem well-defined (over-determined)
  - To achieve sufficient signal-to-noise ratio (SNR)
    - Higher SNR yields higher resolution
    - To reach atomic resolution requires 10^6 2-D images



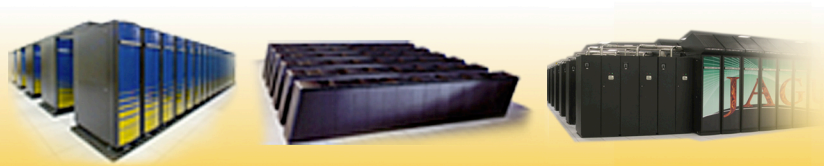


Mathematical Formulation

- Data: $g_i \in \Re^{n^2}, i = 1,2,\ldots m;$
- Unknown parameters:
  - Density: $f \in \Re^{n^3}$
  - Rotations: $(\varphi_i, \theta_i, \phi_i), i = 1,2,\ldots,m$
  - Translations: $(s_{x_i}, s_{y_i}), i = 1,2,\ldots m;$
- Objective

$$\min_{\varphi_i,\theta_i,\phi_i,f,s_{x_i},s_{y_i}} \sum_{i=1}^{m} \|r_i\|^2 = \sum_{i=1}^{m} \left\| P\left(\varphi_i, \theta_i, \phi_i, s_{x_i}, s_{y_i}\right) f - g_i \right\|^2$$

"Unified 3-D Structural and Projection Orientation Refinement Using Quasi-Newton Algorithm." C. Yang, E. Ng and P. A. Penczek. Journal of Structural Biology **149** (2005), pp. 53–64.



Computing the Search Direction

- Objective function $\rho(x) = \frac{1}{2}\sum_{i=1}^{m} \|r_i\|^2$

$$x^T = (f\ \phi_1 \cdots \phi_m\ \theta_1 \cdots \theta_m\ \psi_1 \cdots \psi_m)$$

- Gradient $\nabla\rho(x) = J^T r$

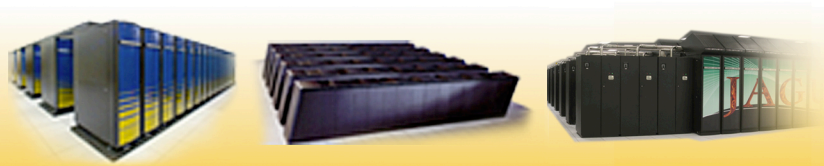
$$J = \begin{pmatrix} P_1 & g_1^{\phi} & & & & g_1^{\theta} & & & & g_1^{\psi} & & \\ P_2 & & g_2^{\phi} & & & & g_2^{\theta} & & & & g_2^{\psi} & \\ \vdots & & & \ddots & & & & \ddots & & & & \ddots \\ P_m & & & & g_m^{\phi} & & & & g_m^{\theta} & & & & g_m^{\psi} \end{pmatrix} \qquad r = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_m \end{pmatrix}$$

$$g^{\phi_1} = \frac{\partial r_1}{\partial \phi_1} \approx \frac{P(\phi_1 + \Delta\phi) f - P(\phi_1) f}{\Delta\phi}$$

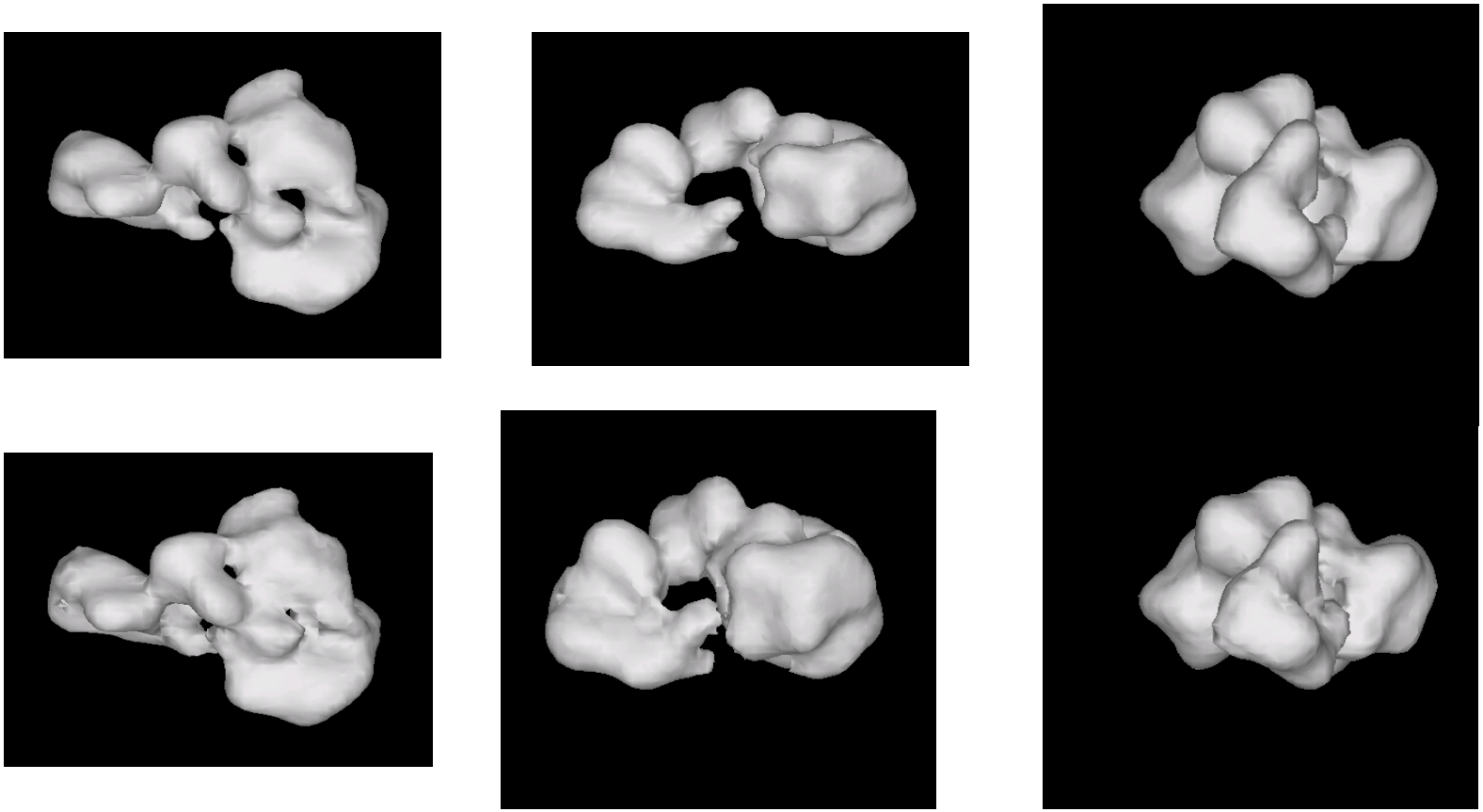
J is mn^2 by n^3+3m

m can be as large as 10^6

n can be as large as 512



Exact vs. Reconstructed Volume



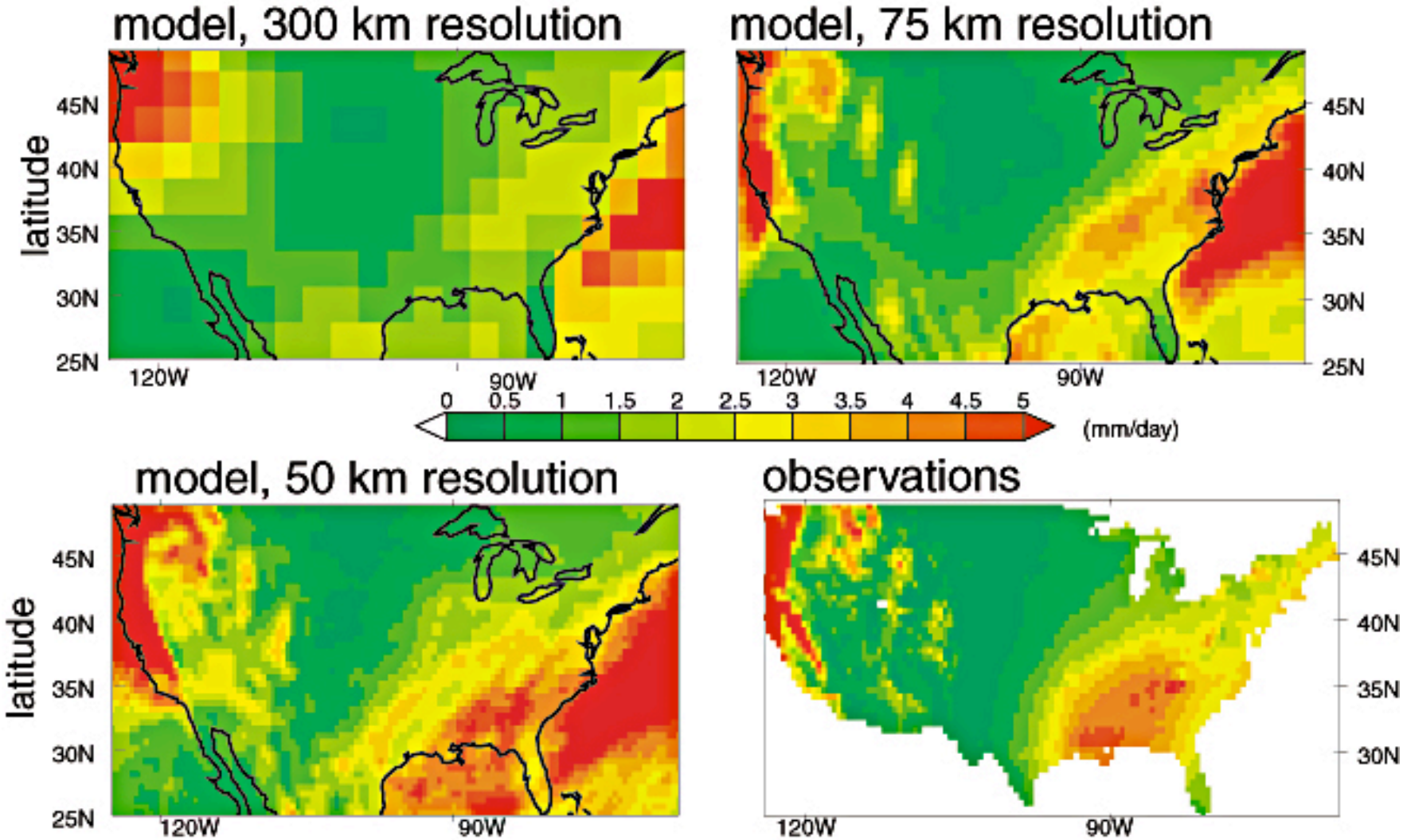
Cryo-EM - Summary

- The computer IS the microscope!
- Image resolution is directly correlated to the available compute power
- Naïve and complete ab initio calculation of a protein structure might require 10**(18) operations

High Resolution Climate Modeling – P. Duffy, et al., LLNL

Wintertime Precipitation

As model resolution becomes finer, results converge towards observations

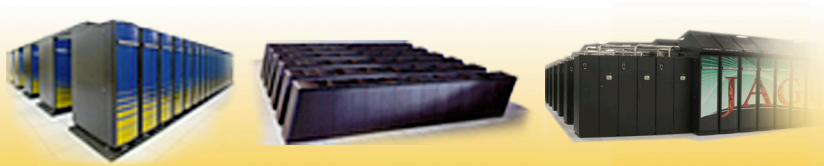


Tropical Cyclones and Hurricanes

Research by: Michael Wehner, Berkeley Lab,
Ben Santer, Phil Duffy, and G. Bala, LLNL

- Hurricanes are extreme events with large impacts on human and natural systems
- Characterized by high vorticity (winds), very low pressure centers, and upper air temperature warm anomalies
- Wind speeds on the Saffir-Simpson Hurricane Scale
  - Category one: 74-95 mph (64-82 kt or 119-153 km/hr)
  - Category two: 96-110 mph (83-95 kt or 154-177 km/hr)
  - Category three: 111-130 mph (96-113 kt or 178-209 km/hr)
  - Category four: 131-155 mph (114-135 kt or 210-249 km/hr)
  - Category five: >155 mph (135 kt or 249 km/hr).

How will the hurricane cycle change as the mean climate changes?

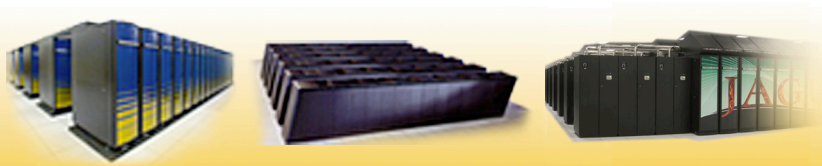


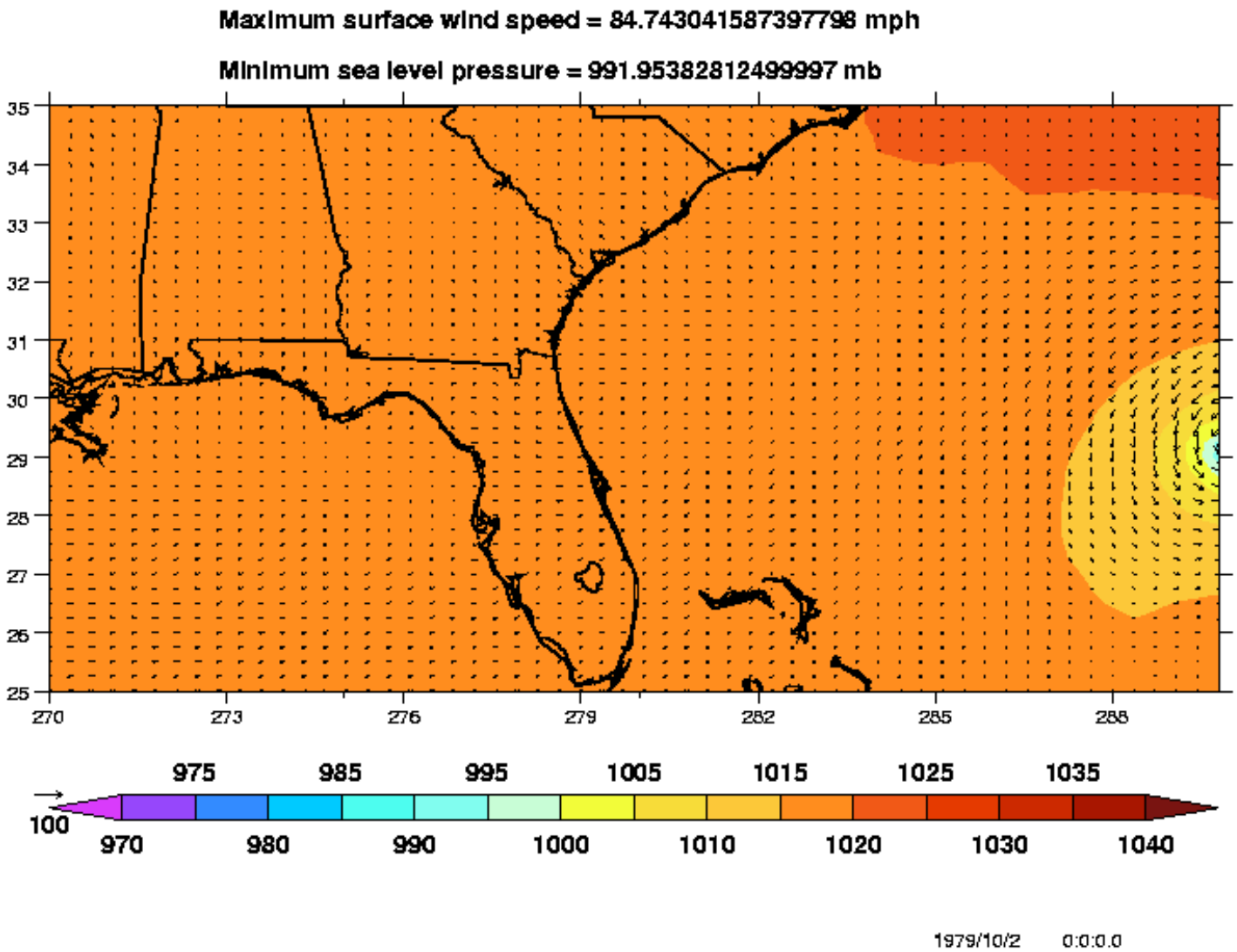
Tropical Cyclones in Climate Models

- Tropical cyclones are not generally seen in integrations of global atmospheric general circulation models at climate model resolutions (T42 ~ 300 km).
- In fact, in CCM3 at T239 (50 km), the lowest pressure attained is 995 mb. No realistic cyclones are simulated.
- However, in high resolution simulations of the finite volume dynamics version of CAM2, strong tropical cyclones are common.

Finite Volume Dynamics CAM

- Run in an 'AMIP' Mode
  - Specified sea surface temperature and sea ice extent
  - Integrated from 1979 to 2000
- We are studying four resolutions
  - B: 2^ox2.5^o
  - C: 1^ox1.25^o
  - D: 0.5^ox0.625^o
  - E: 0.25^ox0.375^o
- Processor Configuration and Cost (IBM SP3)
  - B: 64 processors, 10 wall clock hours / simulated year
  - C: 160 processors, 22 wall clock hours / simulated year
  - D: 640 processors, 33 wall clock hours / simulated year
  - E: 640 processors, 135 wall clock hours / simulated year




Maximum surface wind speed = 84.743041587397798 mph
Minimum sea level pressure = 991.95382812499997 mb
35
34
33
32
31
30
29
28
27
26
25
270
273
276
279
282
285
288
975
985
995
1005
1015
1025
1035
100
970
980
990
1000
1010
1020
1030
1040
1979/10/2 0:0:0.0

New Science Question: Hurricane Statistics

What is the effect of different climate scenarios on number and severity of tropical storms?

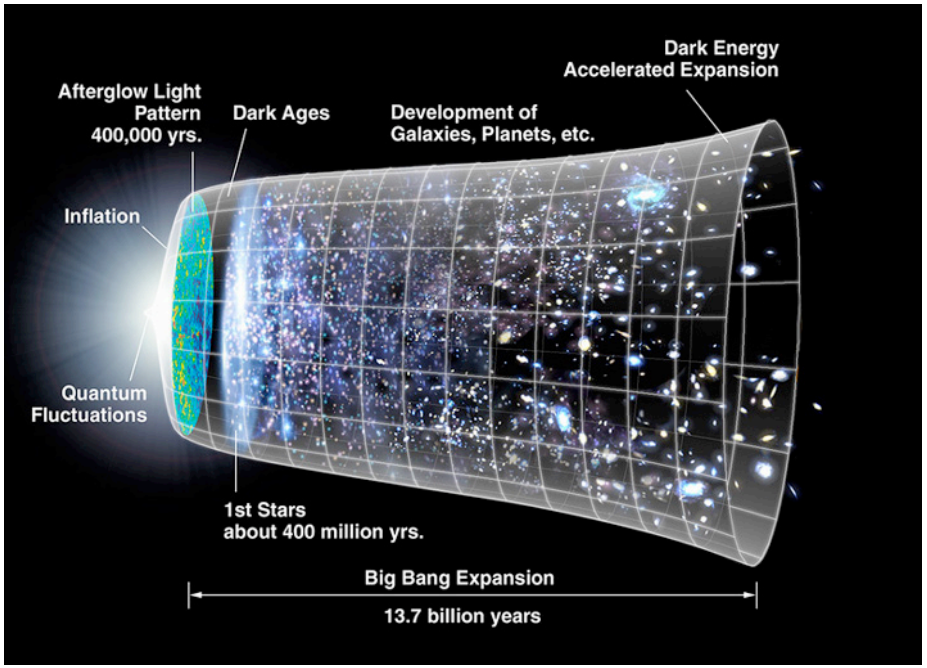
	1979	1980	1981	1982	Obs
Northwest Pacific Basin	>25	~30			40
Atlantic Basin	~6	~12			?

Work in progress—computer power insufficient!

Extreme Weather - Summary

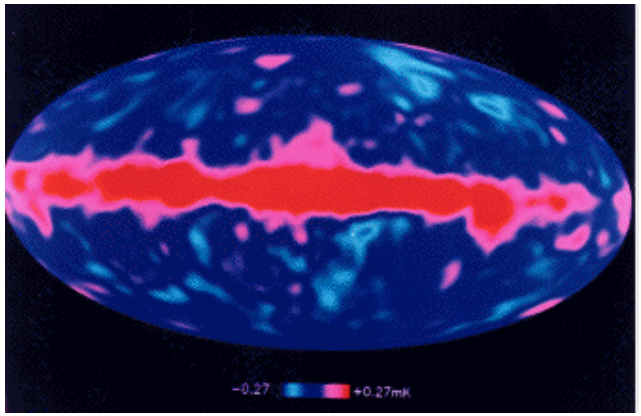
- Computer Simulation permits us to perform experiments that are too dangerous
- We can ask new scientific questions that we could not even think of before
- Current computer power still insufficient to get statistically meaningful results on possible correlation of extreme weather and climate change

NERSC User George Smoot wins 2006 Nobel Prize in Physics

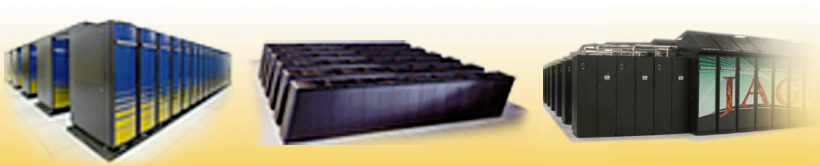


Smoot and Mather 1992

COBE Experiment showed anisotropy of CMB



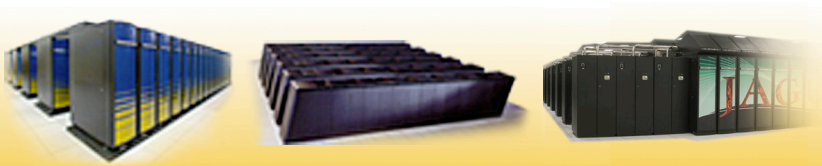
Cosmic Microwave Background Radiation (CMB): an image of the universe at 400,000 years



CMB Computing at NERSC

- CMB data analysis presents a significant and growing computational challenge, requiring
  - well-controlled approximate algorithms
  - efficient massively parallel implementations
  - long-term access to the best HPC resources
- DOE/NERSC has become the leading HPC facility in the world for CMB data analysis
  - O(1,000,000) CPU-hours/year
  - O(10) Tb project disk space
  - O(10) experiments & O(100) users (rolling)

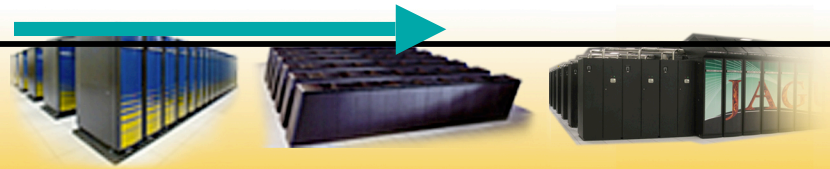
source J. Borrill, LBNL



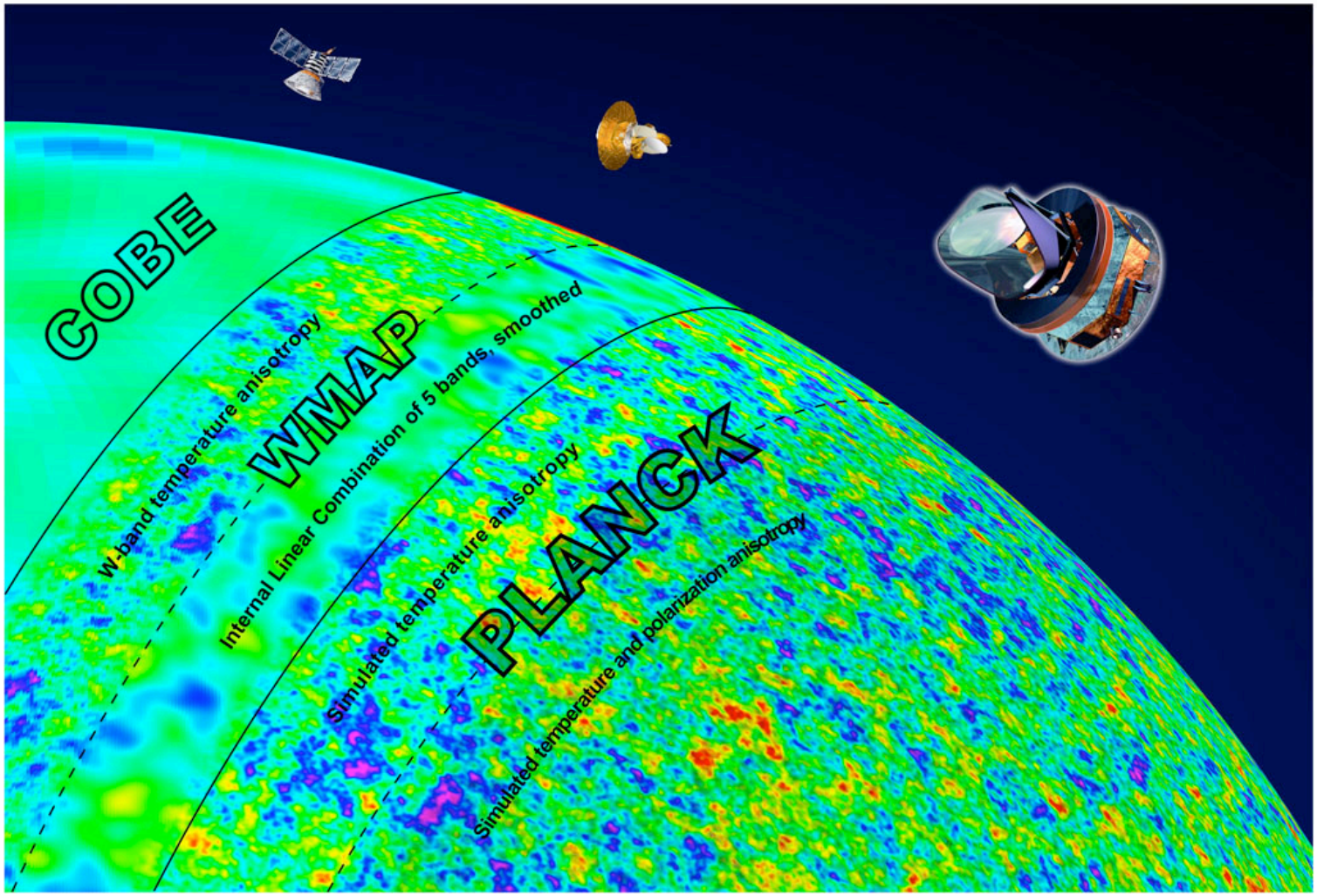
Evolution Of CMB Data Sets

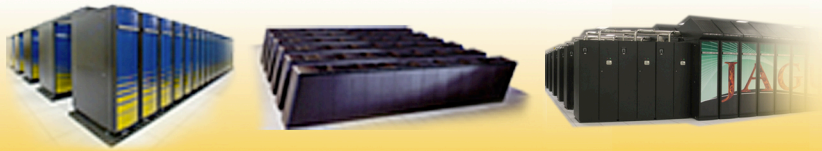
Experiment	N_t	N_p	N_b	Limiting Data	Notes
COBE (1989)	$2x10^9$	$6x10^3$	$3x10^1$	Time	Satellite, Workstation
BOOMERanG (1998)	$3x10^8$	$5x10^5$	$3x10^1$	Pixel	Balloon, 1st HPC/NERSC
(4yr) WMAP (2001)	$7x10^{10}$	$4x10^7$	$1x10^3$	?	Satellite, Analysis-bound
Planck (2007)	$5x10^{11}$	$6x10^8$	$6x10^3$	Time/ Pixel	Satellite, Major HPC/DA effort
POLARBEAR (2007)	$8x10^{12}$	$6x10^6$	$1x10^3$	Time	Ground, NG-multiplexing
CMBPol (~2020)	10^{14}	10^9	10^4	Time/ Pixel	Satellite, Early planning/design

data compression



Evolution Of CMB Satellite Maps





Algorithms & Flop-Scaling

- Map-making
  - Exact maximum likelihood : $O(N_p^3)$
  - PCG maximum likelihood : $O(N_i\ N_t \log N_t)$
  - Scan-specific, e.g.. destriping : $O(N_t \log N_t)$
  - Naïve : $O(N_t)$

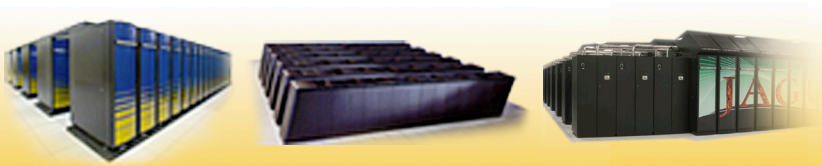
Speed ↓ Accuracy ↑

- Power Spectrum estimation
  - Iterative maximum likelihood : $O(N_i\ N_b\ N_p^3)$
  - Monte Carlo pseudo-spectral :
    - Time domain : $O(N_r\ N_i\ N_t \log N_t)$, $O(N_r\ l_{max}^3)$
    - Pixel domain : $O(N_r\ N_t)$
    - Simulations
      - exact simulation > approximate analysis !

Speed ↓ Accuracy ↑

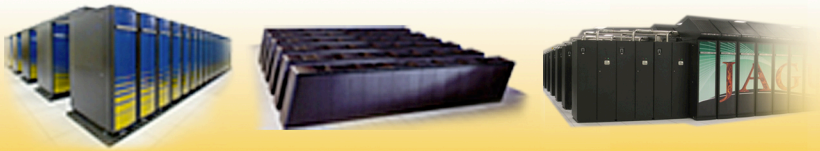
CMB is Characteristic for CSE Projects

- Petaflop/s and beyond computing requirements
- Algorithm and software requirements
- Use of new technology, e.g. NGF
- Service to a large international community
- Exciting science

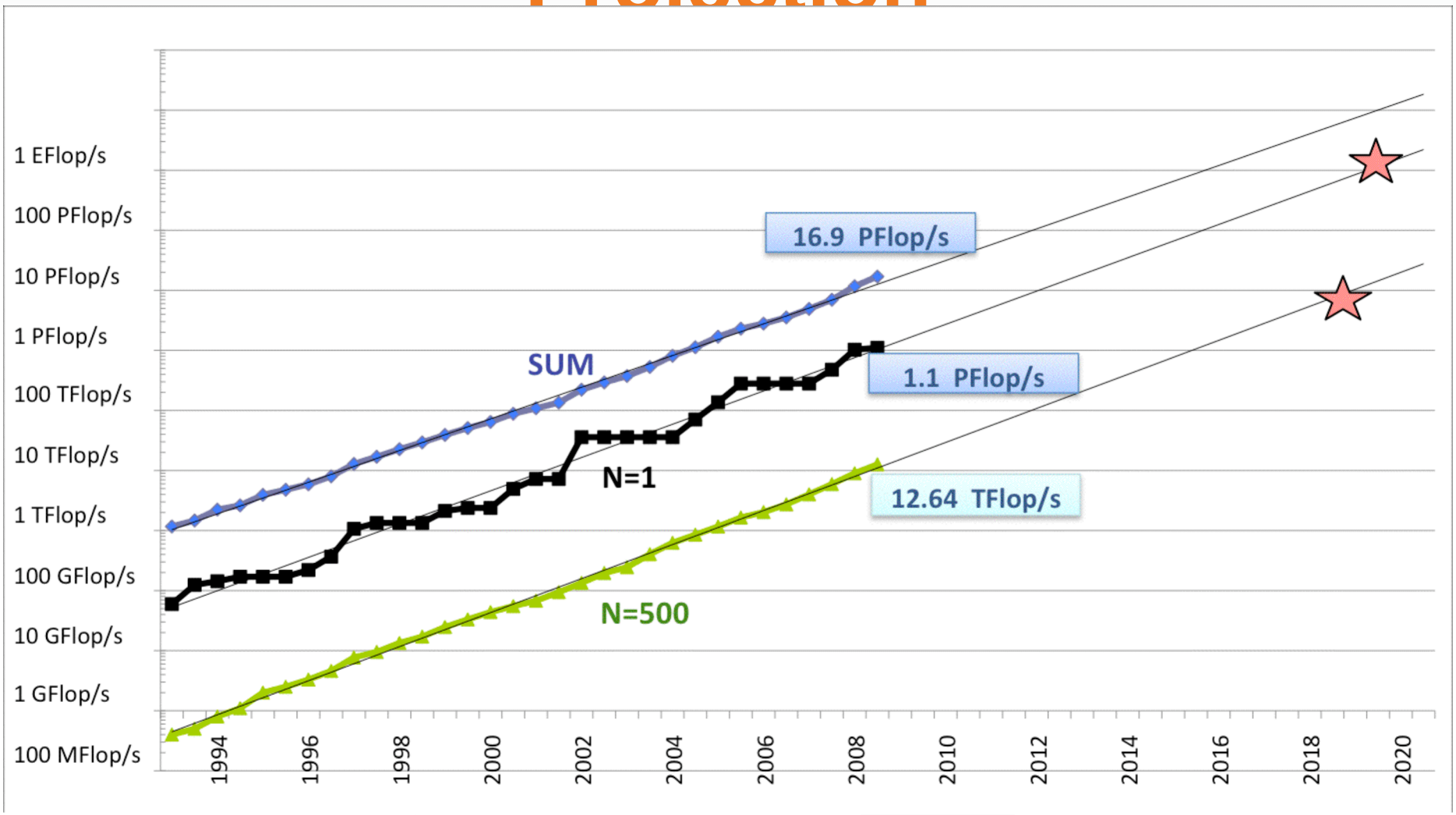


Overview

- Introducing NERSC and Computing Sciences at Berkeley Lab
- Current Trends in Supercomputing (High-End Computing)
- What Supercomputers Do
- What Supercomputers Still Can't Do



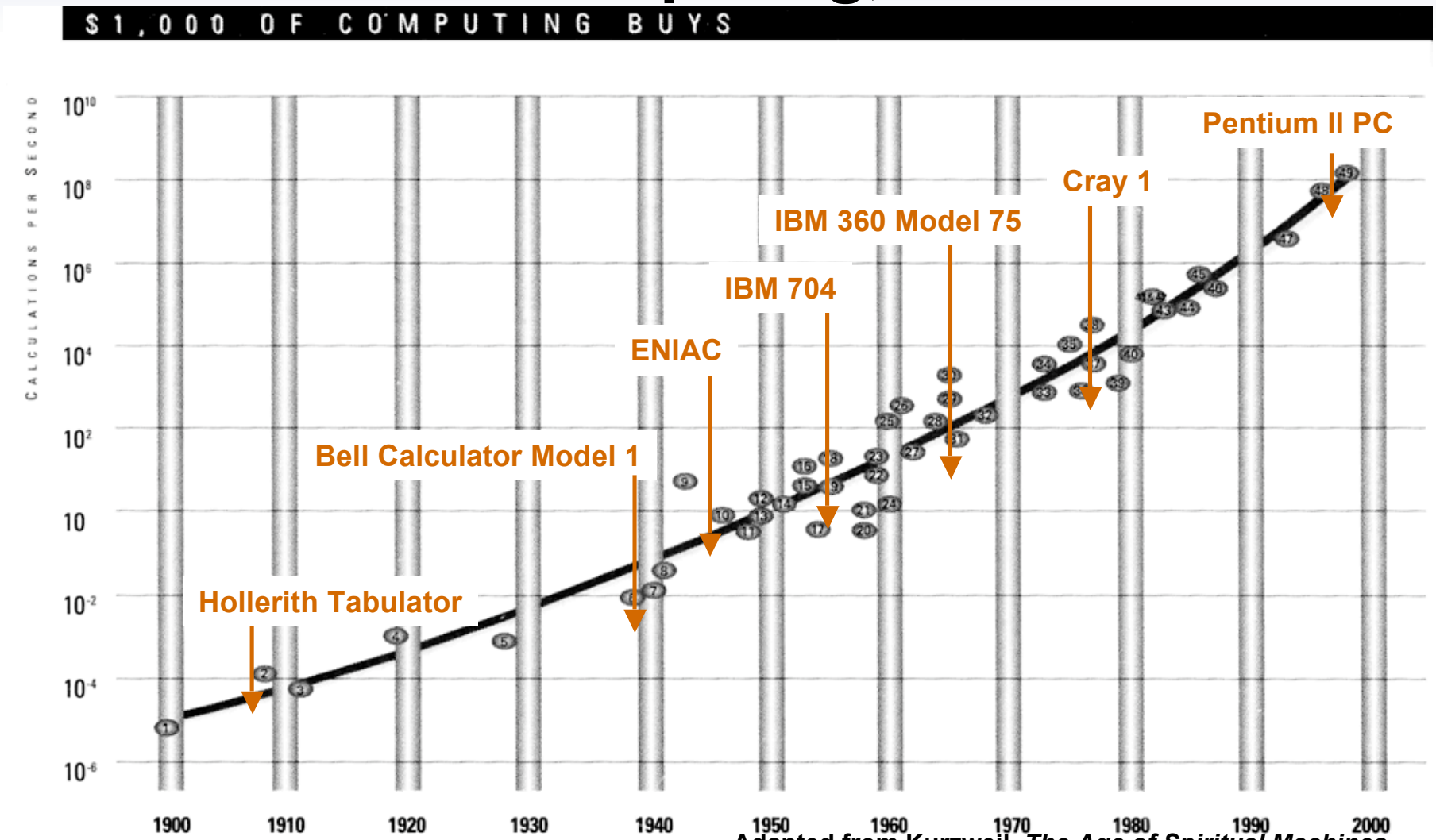
Performance Development Projection



BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

The Exponential Growth of Computing, 1900-1998

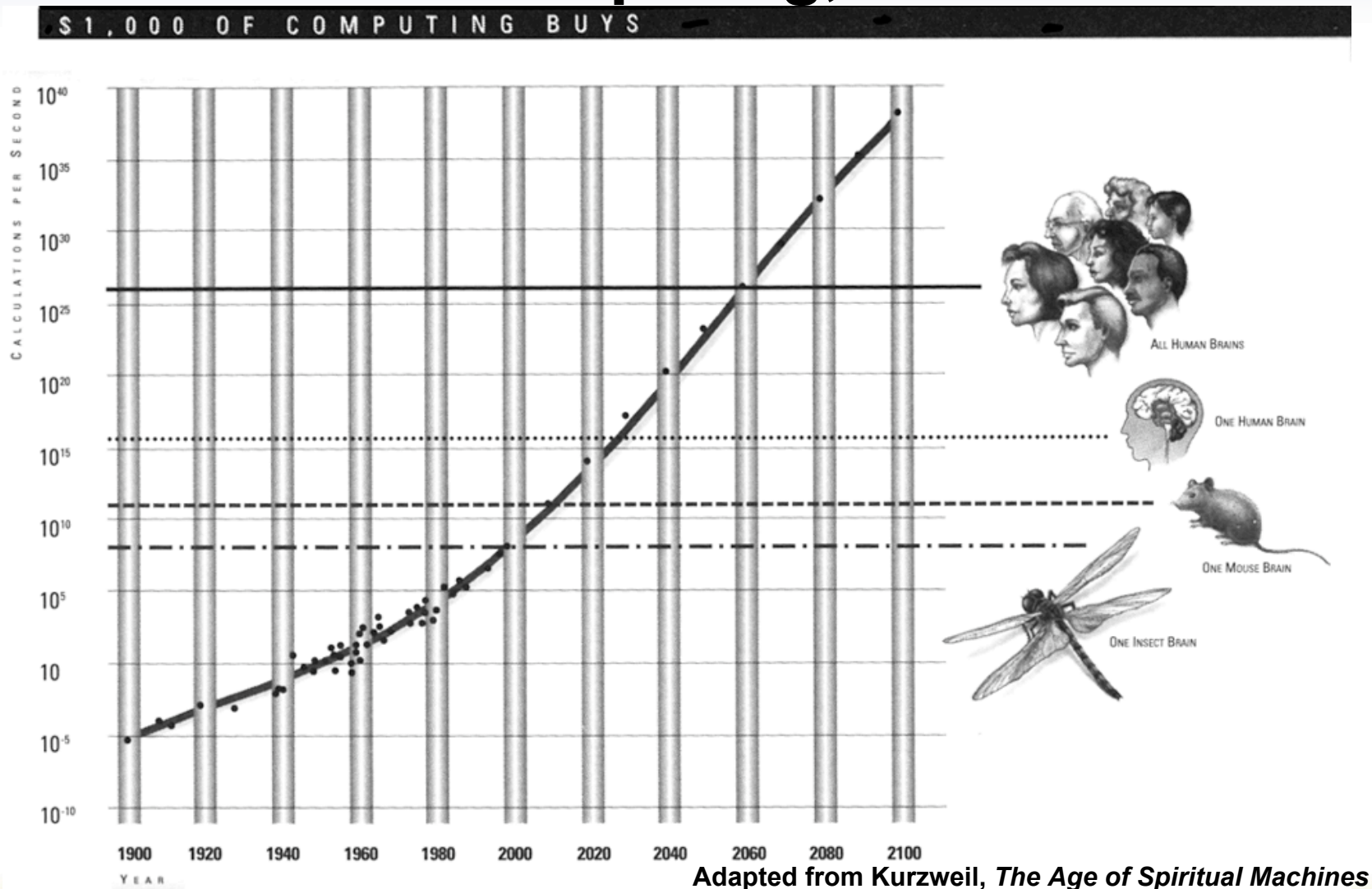


Adapted from Kurzweil, *The Age of Spiritual Machines*

BERKELEY LAB

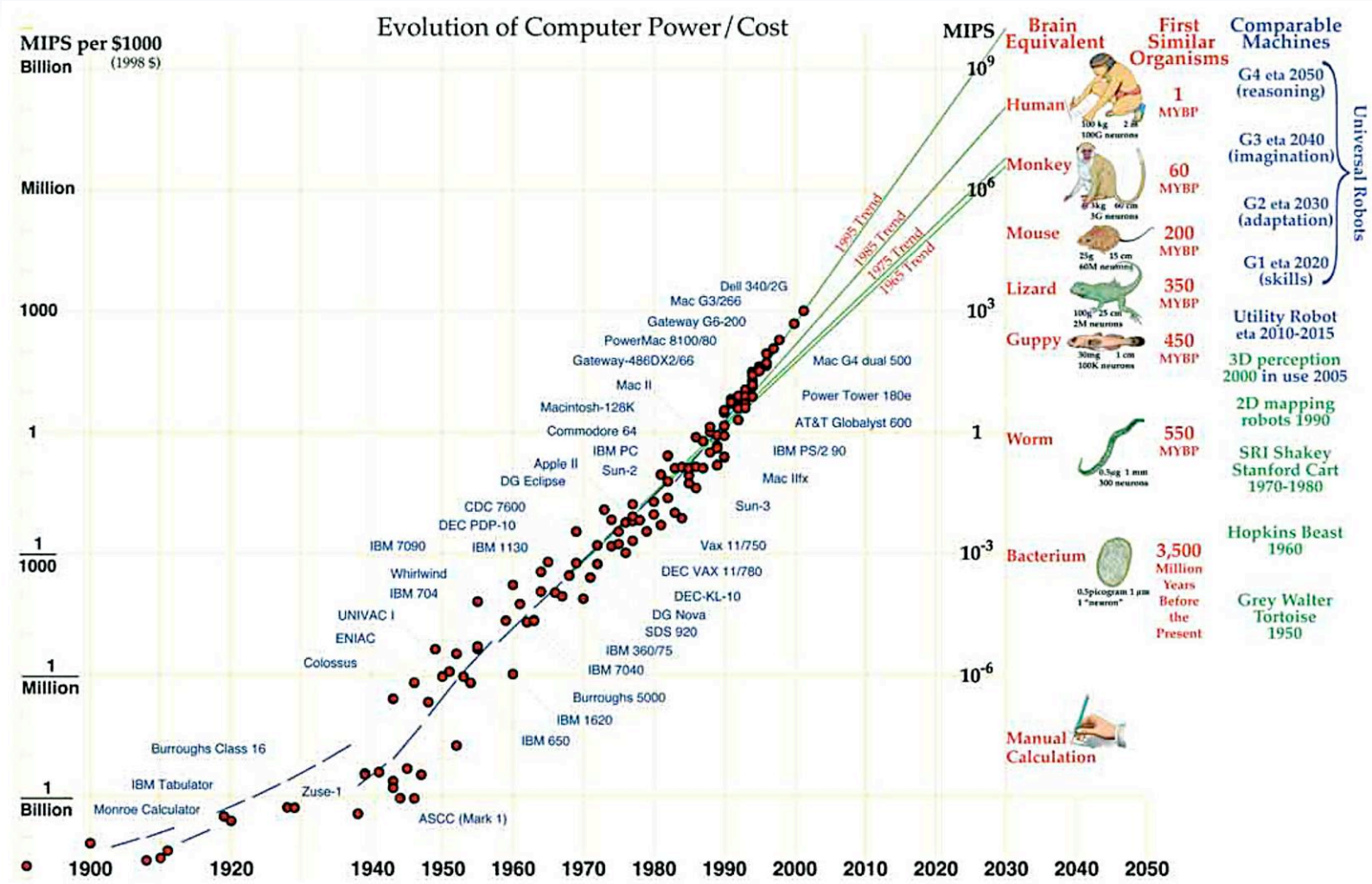
U.S. DEPARTMENT OF ENERGY
Office of Science

The Exponential Growth of Computing, 1900-2100



Adapted from Kurzweil, *The Age of Spiritual Machines*

Growth of Computing Power and "Mental Power"



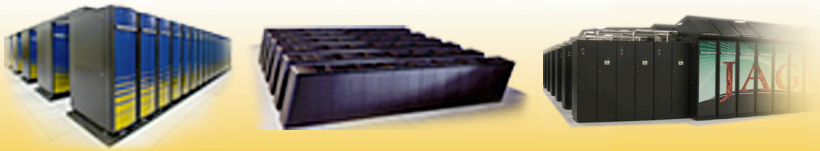
Hans Moravec, *CACM* 10, 2003, pp 90-97

BERKELEY LAB

U.S. DEPARTMENT OF ENERGY
Office of Science

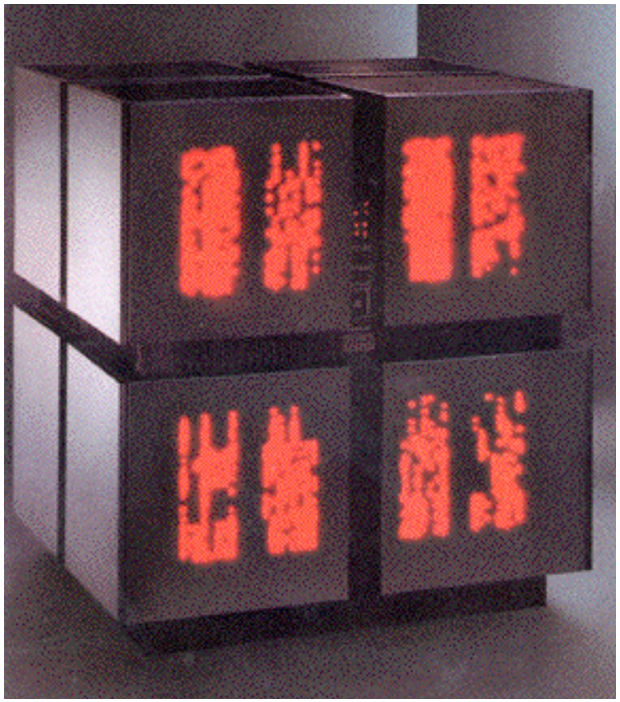
Why This Simplistic View is Wrong

- Unsuitability of Current Architectures
  - Teraflop systems are focused on excelling in computing; only one of the six (or eight) dimensions of human intelligence
- Fundamental lack of mathematical models for cognitive processes
  - That's why we are not using the most powerful computers today for cognitive tasks
- Complexity limits
  - We don't even know yet how to model turbulence, how then do we model thought?



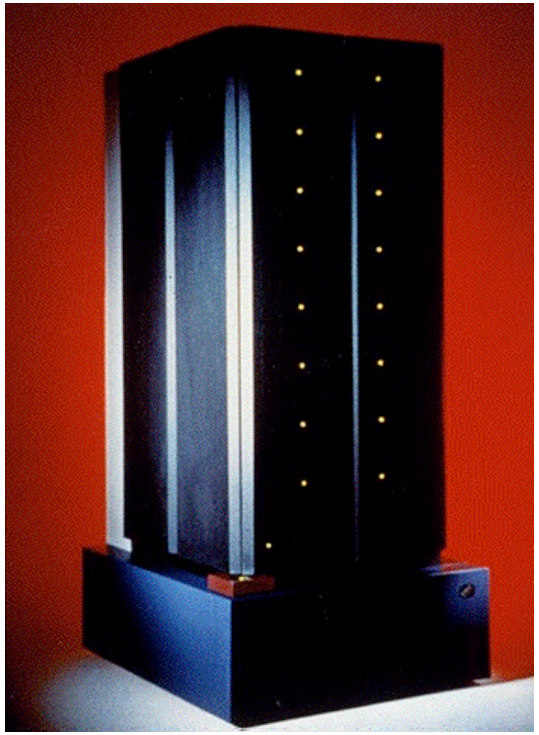
History Lesson: 1987

- "Legendary" CM-2 by Thinking Machines
- Architecture evolved into CM-5 (1992) built as MPP for scientific applications
- Early history of AI applications on parallel platforms has been lost



History Lesson: 1997

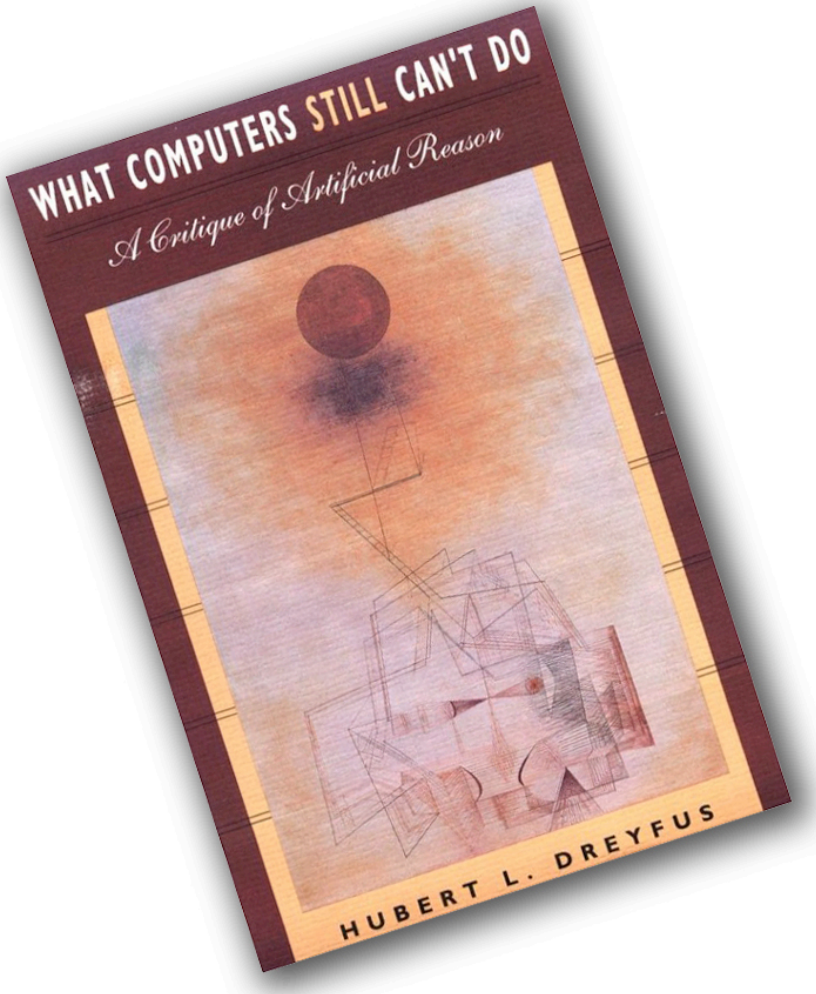
- IBM Deep Blue beats Gary Kasparov (May 1997)
- one of the biggest success stories of machine intelligence,
- however, the chess computer "Deep Blue", did not teach us anything about how a chess grandmaster thinks
- no further analysis or further developments

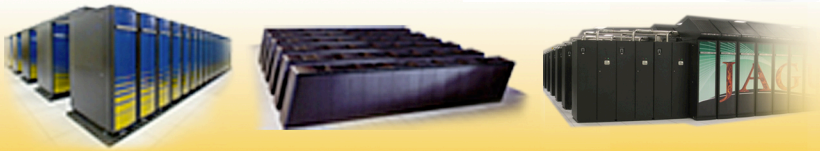


Motivation for the Title of my Talk

"The computer model turns out not to be helpful in explaining what people actually do when they think and perceive."
Hubert Dreyfus, pg.189

Example: one of the biggest success stories of machine intelligence, the chess computer "Deep Blue", did not teach us anything about how a chess grandmaster thinks.

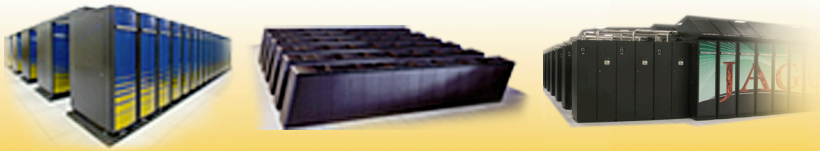




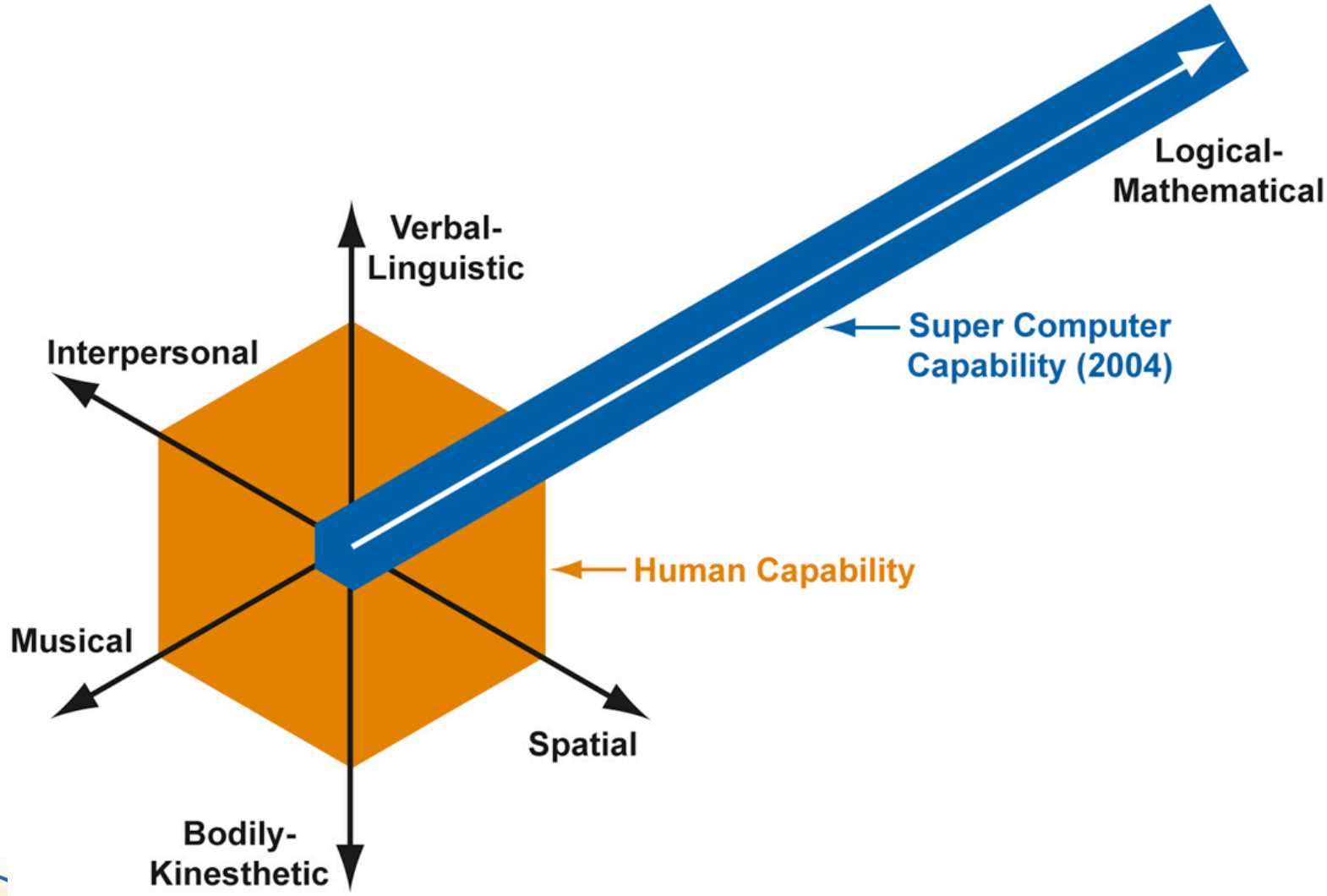
Six Dimensions of Intelligence

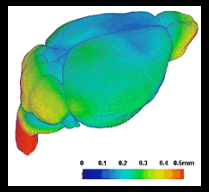
1. **Verbal-Linguistic**
   - ability to think in words and to use language to express and appreciate complex concepts
2. **Logical-Mathematical**
   - makes it possible to calculate, quantify, consider propositions and hypotheses, and carry out complex mathematical operations
3. **Spatial**
   - capacity to think and orientate in physical three-dimensional environment
4. **Bodily-Kinesthetic**
   - ability to manipulate objects and fine-tune physical skills
5. **Musical**
   - sensitivity to pitch, melody, rhythm, and tone
6. **Interpersonal**
   - capacity to understand and interact effectively with others

Howard Gardner. *Frames of Mind: The Theory of Multiple Intelligences*. New York: Basic Books, 1983, 1993.



Current State of Supercomputers



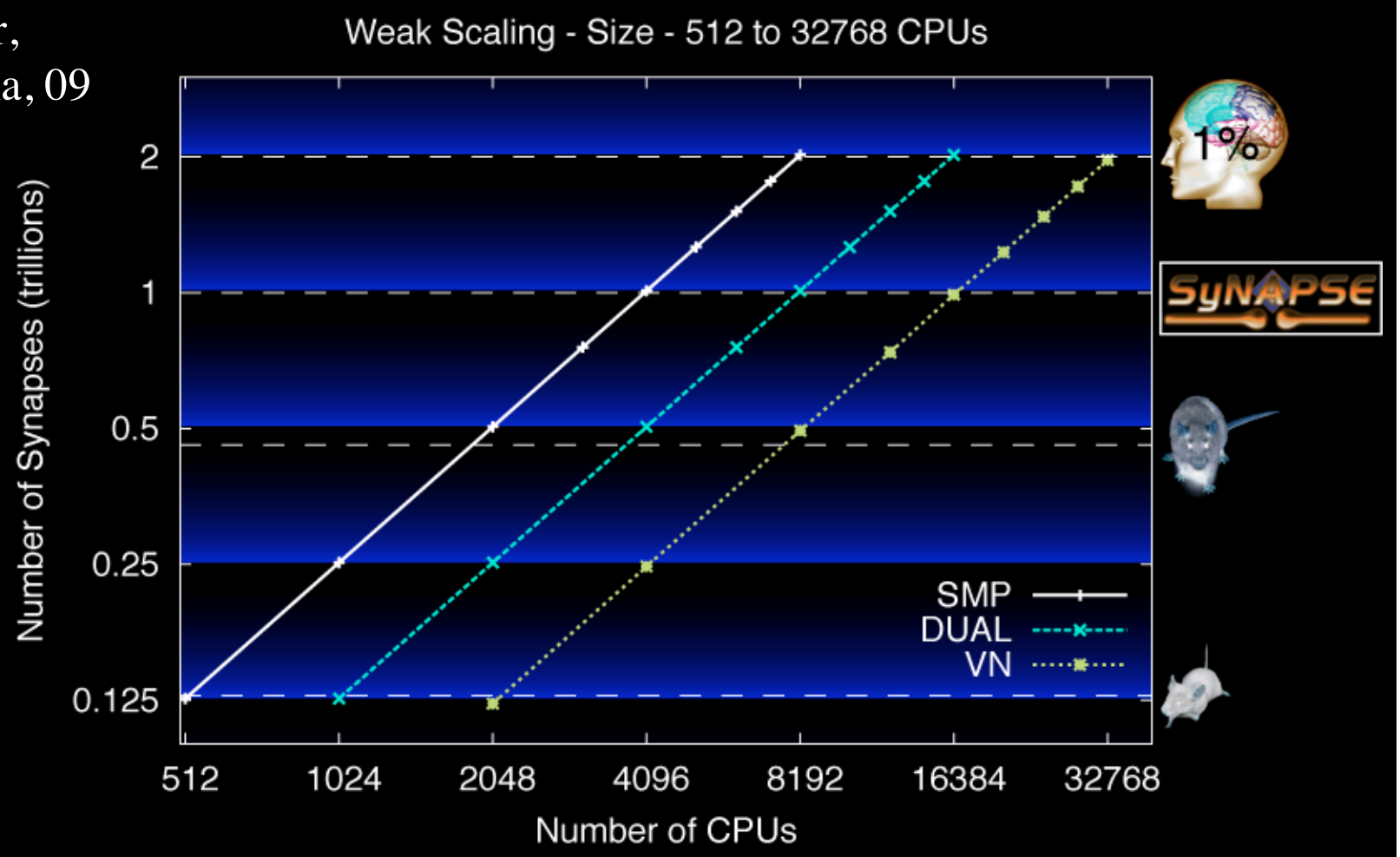
	Mouse 	Rat	Human	BlueGene/P
Neurons	$2x8x10^6$	$2x28x10^6$	$2x10x10^9$	
Synapses	128×10^9	448×10^9	200×10^{12}	
Communication 66 B/spike	128×10^9 Spikes/sec	448×10^9 Spikes/sec	200×10^{12} Spikes/sec	
Computation 175F/synapse/sec	22.4 TF	78.4 TF	~40 PF	111 TF 32,768 CPUs
Memory 16B/synapse	2 TB	7.2 TB	3.2 PB	32 TB

Assume: neurons fire at 1 Hz

© 2008 IBM Corporation

The First Percent

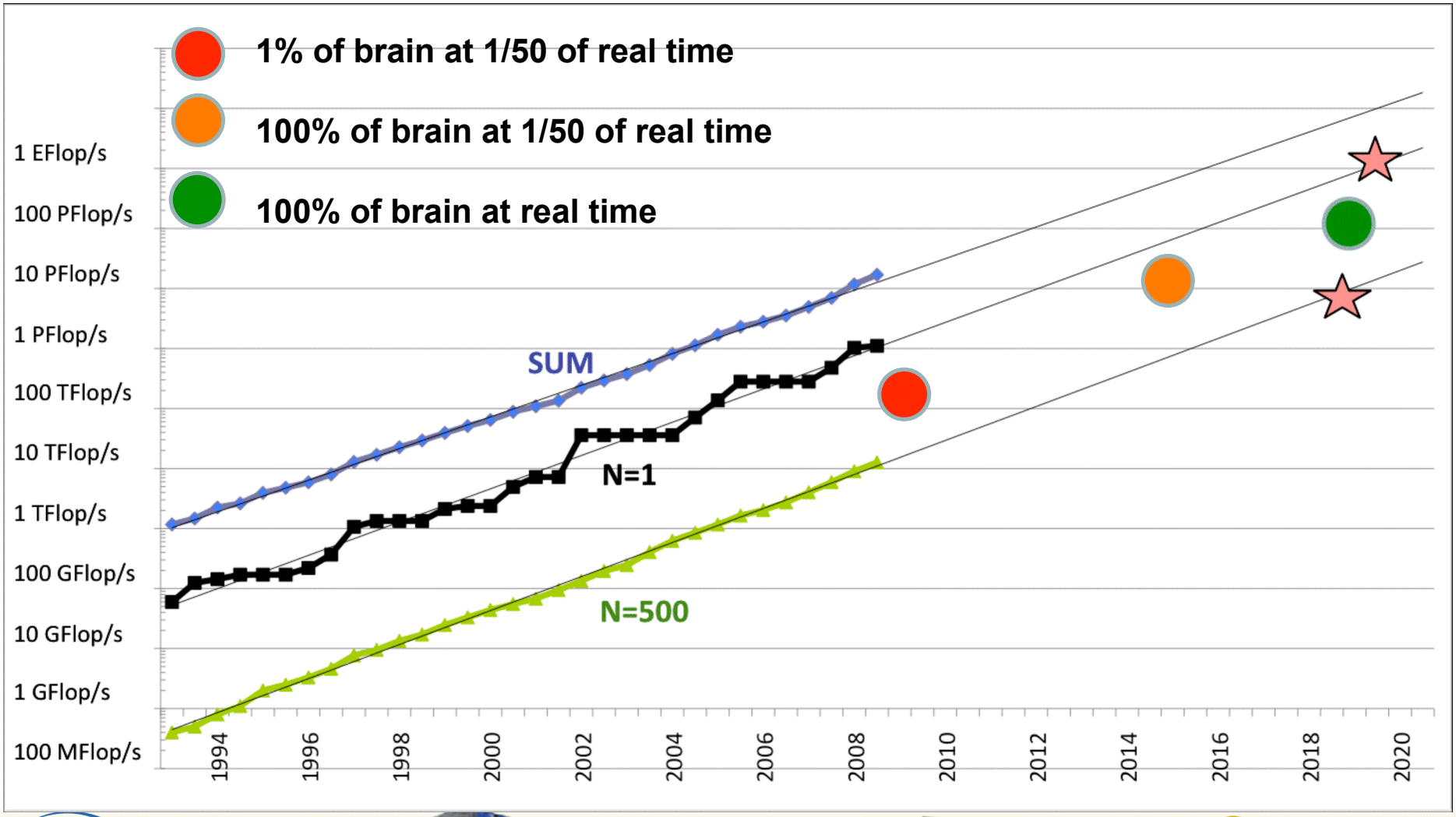
Ananth, Esser, Simon, Modha, 09

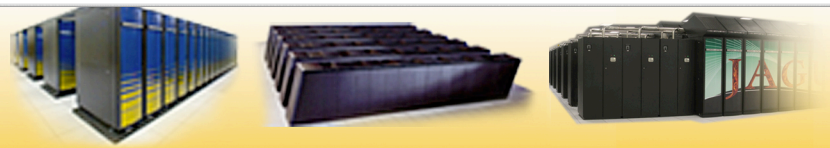


Using a 32,768 processor WatsonShaheen BG/P with 32TB, simulated 200 million spiking neurons, 2 trillion synapses, 50x slower than real-time

© 2008 IBM Corporation

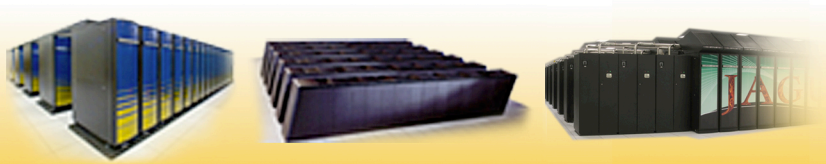
Performance Projection for Brain Simulations





The Power Conundrum

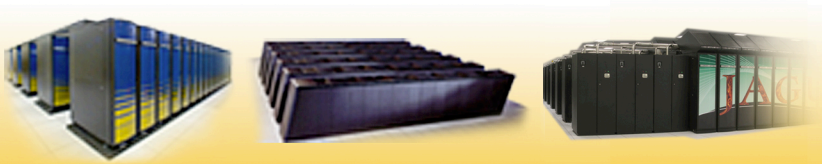
- Unless there are new breakthroughs an Exaflops computer in 2019 will consume 120 MW
- The human brain operates at 20 – 40 W
- We will need another factor of 1 M to match the energy efficiency of the human brain



Summary – Computing Technology

The next decade will see an unprecedented growth in parallelism in computing. In order to reach Exaflop/s computing it is necessary

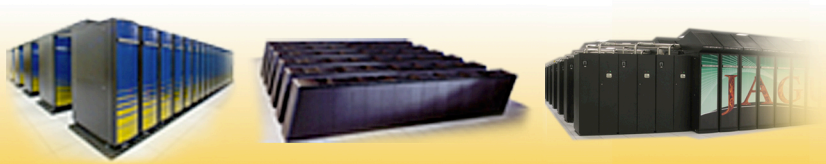
- to deal with 100M to 1B parallelism in application and systems software
- to develop new programming models
- to develop technologies that reduce power consumption by a factor of 1M
- solving any of these problems could also lead to significant commercial success



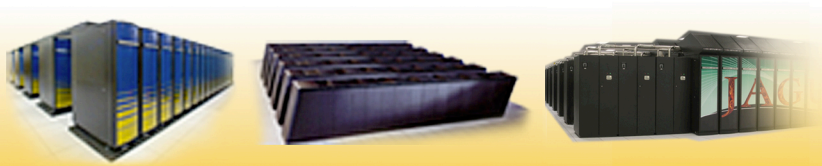
Summary – Computational Modeling

There are vast areas of science and engineering where CSE has not even begun to make an impact

- current list of CSE applications is almost the same as fifteen years ago
- in many scientific areas there is still an almost complete absence of computational models
- even in established areas many researchers do not know how to use leading-edge computational resources



Summary – Cognitive Computing

Huge research opportunities for computer scientists and applied mathematicians

- **the current set of architectures is capturing only a small cognitive abilities subset of human**
- **the use of supercomputing for modeling the brain is immature and at its very begining**
- **our tools for analyzing vast amounts of data are still primitive and not based human cognitive models**