Facts and Wishful Thinking about the Future of Supercomputing

Horst D. Simon

Director, NERSC Center and Computational Research Divisions

Scicomp 6, Berkeley, California, USA

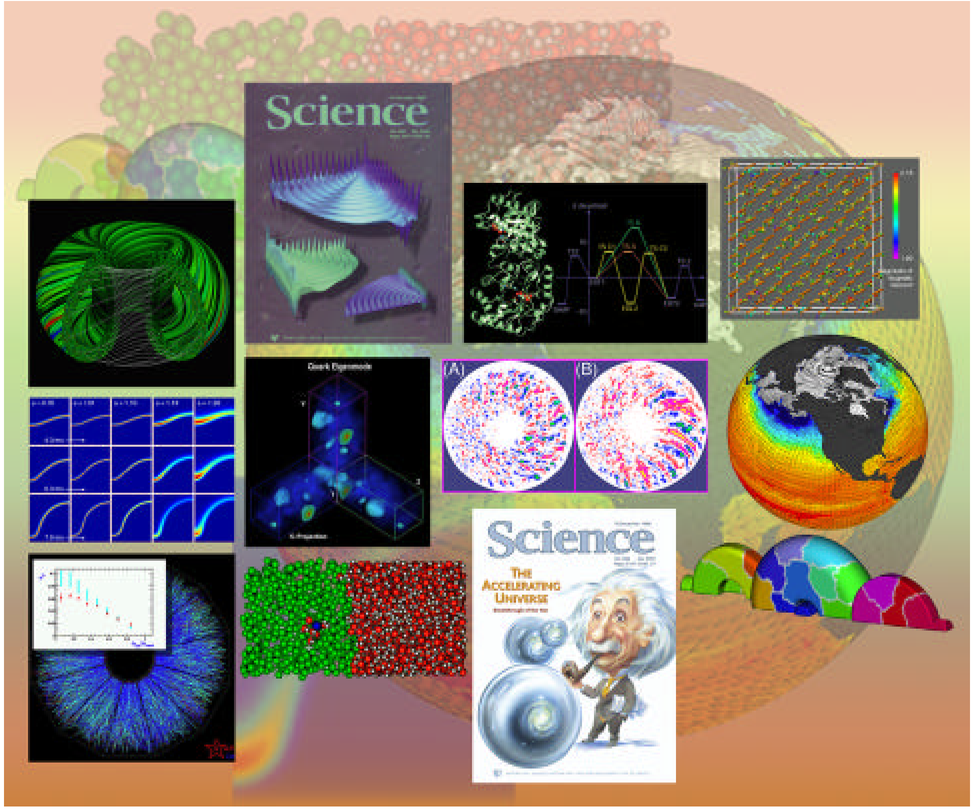
August 2002

Welcome to NERSC and LBNL

- Located in the hills next to University of California, Berkeley campus
- close collaborations between university and NERSC in computer science and computational science

NERSC - Overview

- the Department of Energy, Office of Science, supercomputer facility
- unclassified, open facility; serving >2000 users in all DOE mission relevant basic science disciplines
- 25th anniversary in 1999 (one of the oldest supercomputing centers)



NERSC at Berkeley: six years of excellence in computational science

1997: Expanding Universe is Breakthrough of the year

1998: Fernbach and Gordon Bell Award

1999: Collisional breakup of quantum system

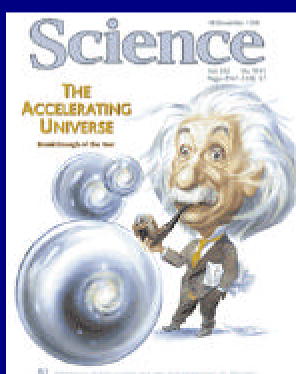
2000: BOOMERANG data analysis= flat universe

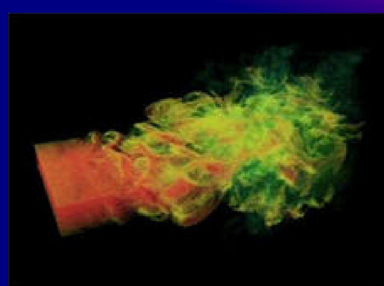
2001: Most distant supernova

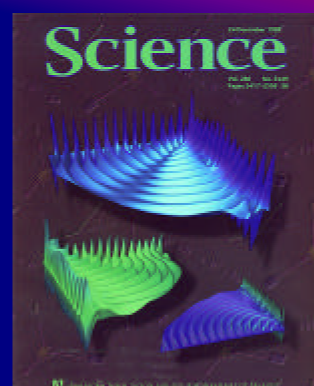
1996

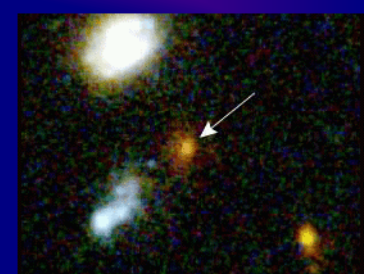
National Energy Research Scientific Computing Center

2002









Outline

- Where are we today?
  - — NERSC examples
  - — current status of supercomputing in the US
- The 40 Tflop/s Earth Simulator and the "Computenik" effect
- Business as usual won't work
- Technology Alternatives

TOP500 – June 2002

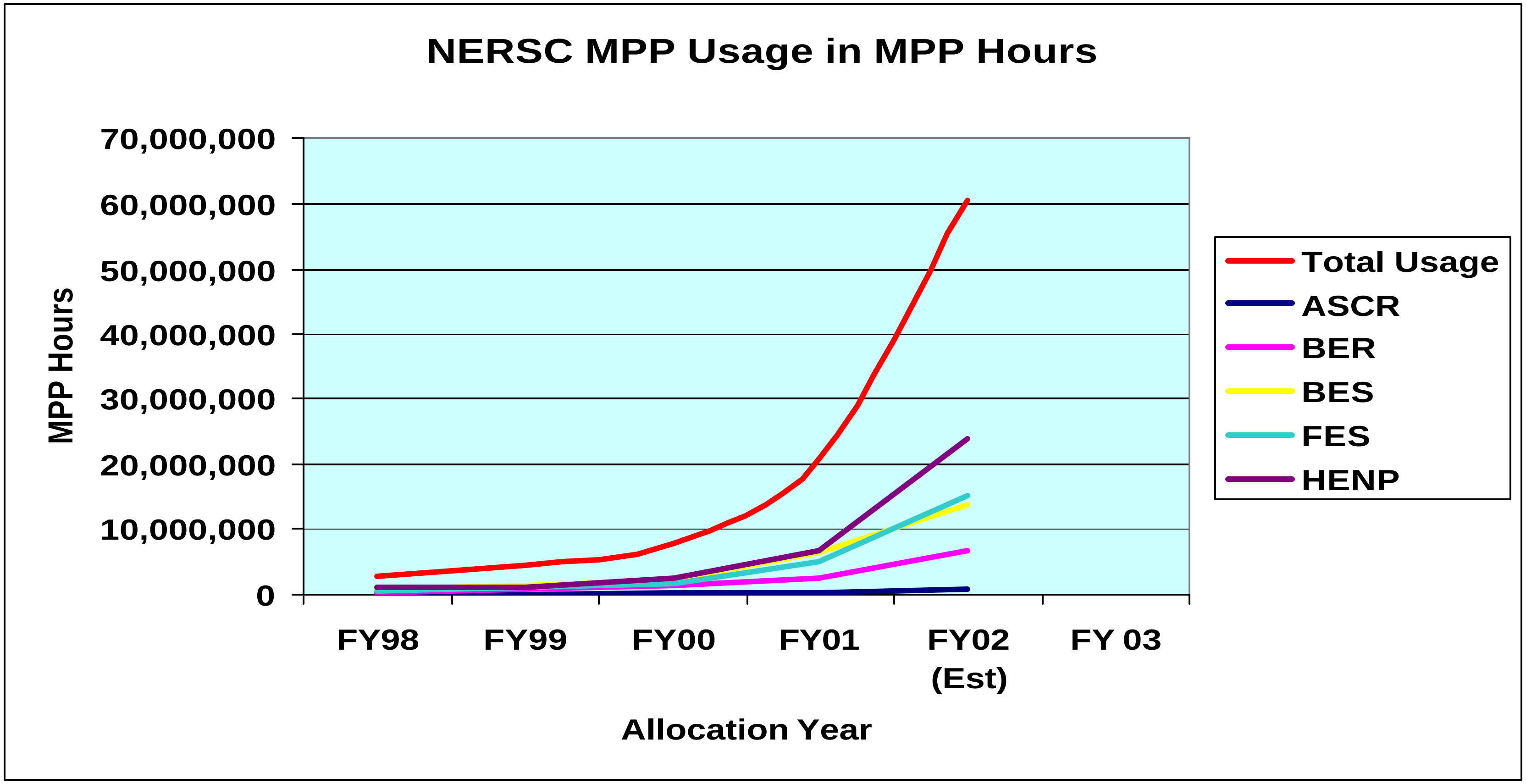
Rank	Manufacturer	Computer	Rmax	Installation Site	Country	Year	Area of Installation	# Proc	Rpeak	Nmax	N1/2
1	NEC	Earth-Simulator	**35860**	Earth Simulator Center Kanazawa	Japan	2002	Research	5120	40960	1075200	266240
2	IBM	ASCI White, SP Power3 375 MHz	**7226**	Lawrence Livermore National Laboratory Livermore	USA	2000	Research Energy	8192	12288	518096	179000
3	Hewlett-Packard	AlphaServer SC ES45/1 GHz	**4463**	Pittsburgh Supercomputing Center Pittsburgh	USA	2001	Academic	3016	6032	280000	85000
4	Hewlett-Packard	AlphaServer SC ES45/1 GHz	**3980**	Commissariat a l'Energie Atomique (CEA) Bruyeres-le-Chatel	France	2001	Research	2560	5120	360000	85000
5	IBM	SP Power3 375 MHz 16 way	**3052**	NERSC/LBNL Berkeley	USA	2001	Research	3328	4992	371712	102400
6	Hewlett-Packard	AlphaServer SC ES45/1 GHz	**2916**	Los Alamos National Laboratory Los Alamos	USA	2002	Research	2048	4096	272000	.
7	Intel	ASCI Red	**2379**	Sandia National Laboratories Albuquerque	USA	1999	Research	9632	3207	362880	75400
8	IBM	pSeries 690 Turbo 1.3GHz	**2310**	Oak Ridge National Laboratory Oak Ridge	USA	2002	Research	864	4493	275000	62000
9	IBM	ASCI Blue-Pacific SST, IBM SP 604e	**2144**	Lawrence Livermore National Laboratory Livermore	USA	1999	Research Energy	5808	3868	431344	.
10	IBM	pSeries 690 Turbo 1.3GHz	**2002**	IBM/US Army Research Laboratory (ARL) Poughkeepsie	USA	2002	Vendor	768	3994	252000	.
		SP Power3 375 MHz		Atomic Weapons							

NERSC-3 Vital Statistics

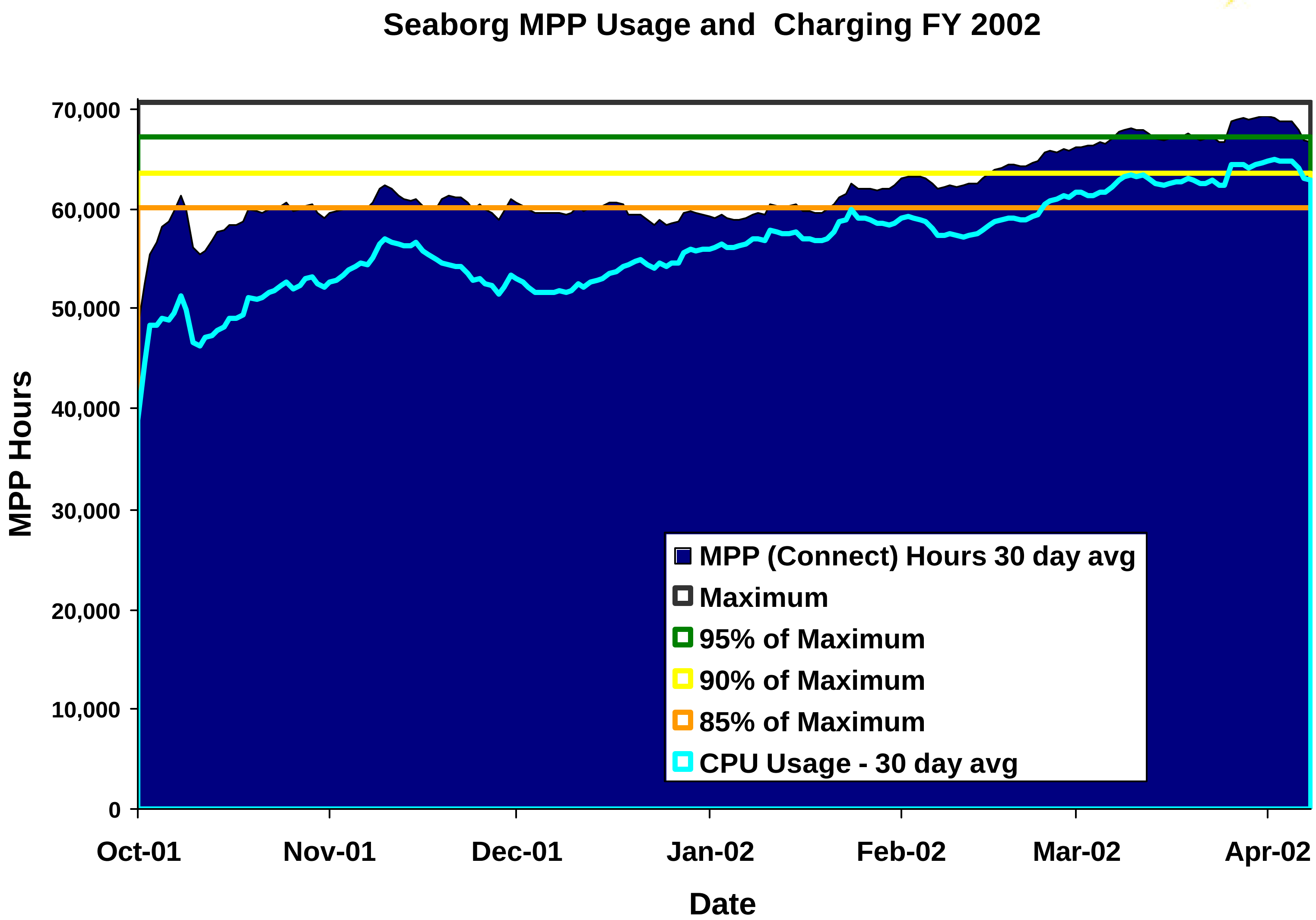
- 5 Teraflop/s Peak Performance – 3.05 Teraflop/s with Linpack
  - 208 nodes, 16 CPUs per node at 1.5 Gflop/s per CPU
  - "Worst case" Sustained System Performance measure .358 Tflop/s (7.2%)
  - "Best Case" Gordon Bell submission 2.46 on 134 nodes (77%)
- 4.5 TB of main memory
  - 140 nodes with 16 GB each, 64 nodes with 32 GBs, and 4 nodes with 64 GBs.
- 40 TB total disk space
  - 20 TB formatted shared, global, parallel, file space; 15 TB local disk for system usage
- Unique 512 way Double/Single switch configuration

The Demand for Supercomputing Cycles is Urgent and Growing

- The growth is dramatically evident at NERSC.

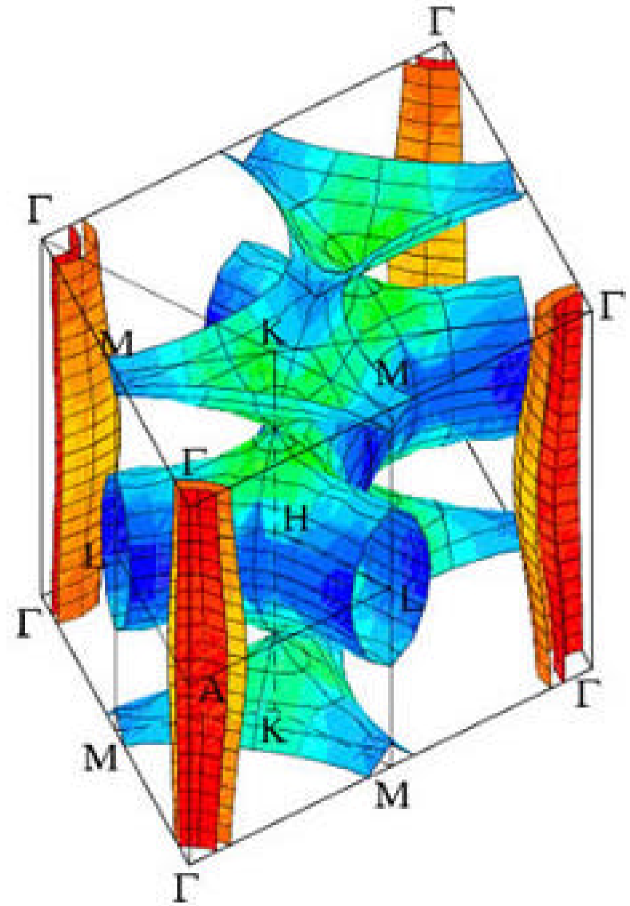


NERSC-3, Installed in 2001, is Already Fully Utilized



Computational Science at NERSC: Explaining HT Superconductivity

- Published August 15, 2002 in Nature by Marvin Cohen and Steven Louie of Berkeley Lab's Materials Sciences Division, and UC Berkeley
- Calculated the properties of the unique superconductor MgB2 from first principles, revealing the secrets of its anomalous behavior, including more than one superconducting energy gap.
- MgB2 becomes superconducting at 39 degrees Kelvin, one of the highest known transition temperatures (Tc) of any superconductor.
- The theorists report that MgB2's odd features arise from two separate populations of electrons -- nicknamed "red" and "blue" -- that form different kinds of bonds among the material's atoms.

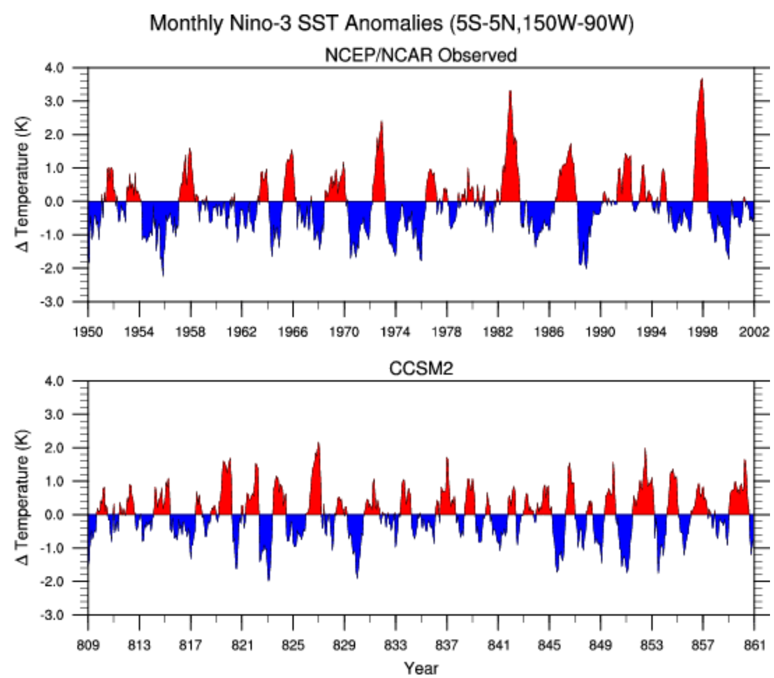


Computational Science at NERSC: A 1000 year climate simulation

- *Warren Washington and Jerry Meehl, National Center for Atmospheric Research; Bert Semtner, Naval Postgraduate School; John Weatherly, U.S. Army Cold Regions Research and Engineering Lab Laboratory.*

• A 1000-year simulation demonstrates the ability of the new Community Climate System Model (CCSM2) to produce a long-term, stable representation of the earth's climate.

•760,000 processor hours by July



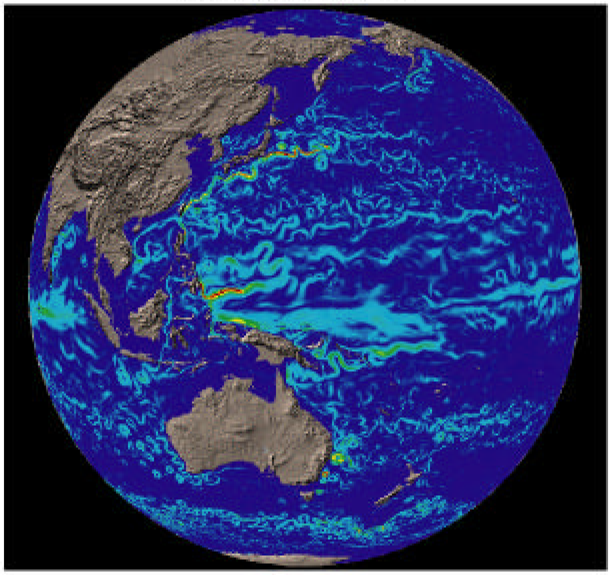
Computational Science at NERSC: High Resolution Global Coupled Ocean/Sea Ice Model

- *Mathew E. Maltrud, Los Alamos National Laboratory; Julie L. McClean, Naval Postgraduate School.*
- The objective of this project is to couple a high-resolution ocean general circulation model with a high-resolution dynamic-thermodynamic sea ice model in a global context.

•Currently, such simulations are typically performed with a horizontal grid resolution of about 1 degree. This project is running a global ocean circulation model with horizontal resolution of approximately 1/10th degree.

•Allows resolution of geographical features critical for climate studies such as Canadian Archipelago

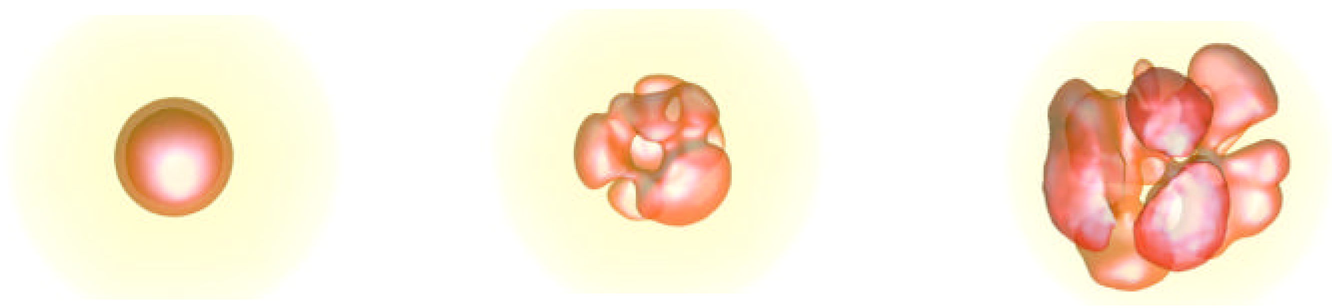
1/10 Degree Global POP Ocean Model Currents at 50m Depth
(blue = 0; red > 150 cm/s)



Computational Science at NERSC: Supernova Explosions and Cosmology

Peter Nugent and Daniel Kasen, Lawrence Berkeley National Laboratory; Peter Hauschildt, University of Georgia; Edward Baron, University of Oklahoma; Stan Woosley and Gary Glatzmaier, University of California, Santa Cruz; Tom Clune, Goddard Space Flight Center; Adam Burrows, Salim Hariri, Phil Pinto, Hessam Sarjoughian, and Bernard Ziegler, University of Arizona; Chris Fryer and Mike Warren, Los Alamos National Laboratory; Frank Dietrich and Rob Hoffman, Lawrence Livermore National Laboratory

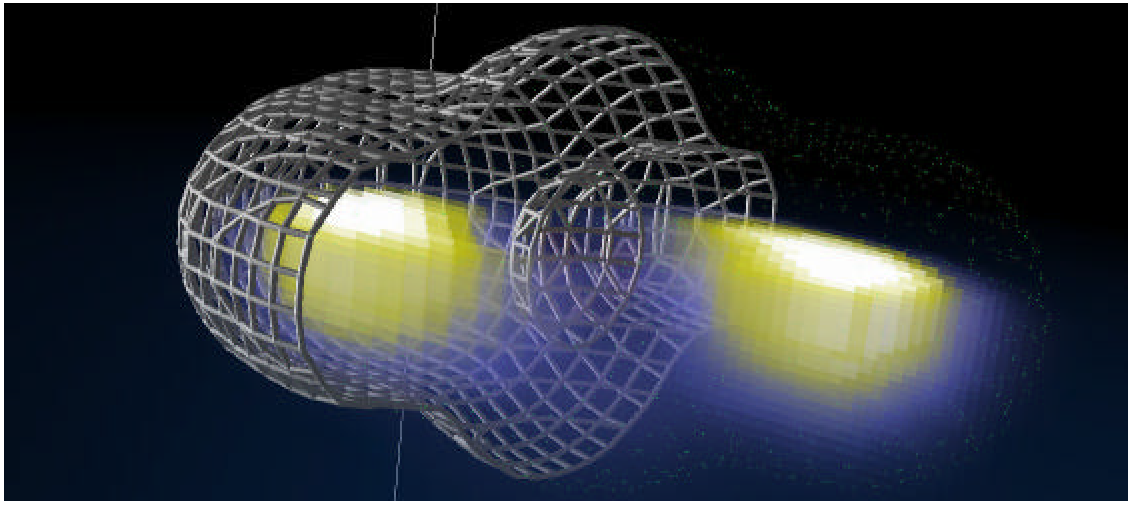
- First 3-D supernova explosion simulation, based on computation at NERSC. This research eliminates some of the doubts about earlier 2-D modeling and paves the way for rapid advances on other questions about supernovae.



Computational Science at NERSC: Black Hole Merger Simulations

- *Ed Seidel, Gabrielle Allen, Denis Pollney, and Peter Diener, Max Planck Institute for Astrophysics; John Shalf, Lawrence Berkeley National Laboratory.*
- Simulations of the spiraling coalescence of two black holes, a problem of particular importance for interpreting the gravitational wave signatures that will soon be seen by new laser interferometric detectors around the world.

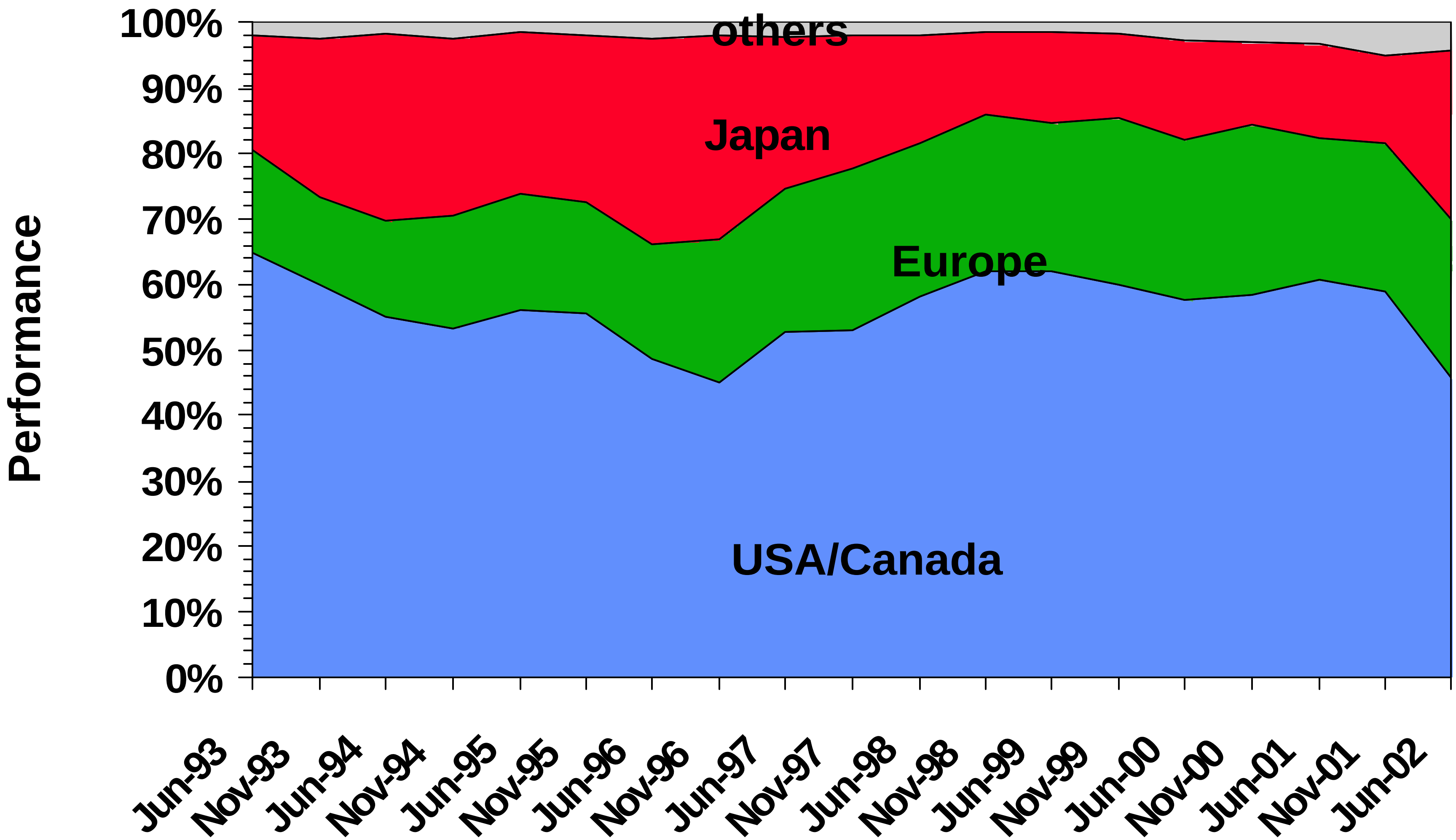
- Required 1.5 Tbytes of memory and was run on the large 64 Gbyte nodes



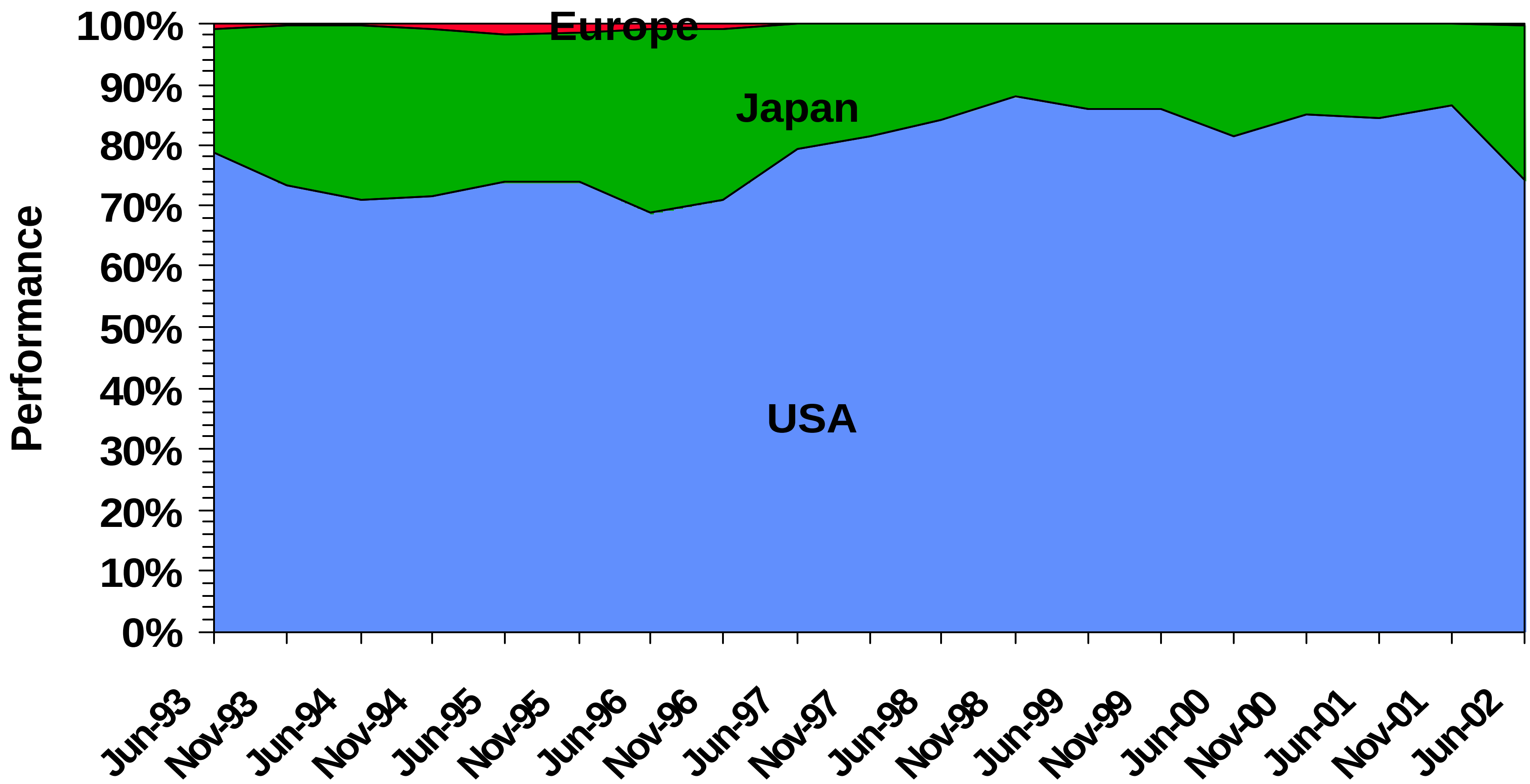
Outline

- Where are we today?
  — NERSC examples
  — current status of supercomputing in the US
- The 40 Tflop/s Earth Simulator and the "Computenik" effect
- Business as usual won't work
- Technology Alternatives

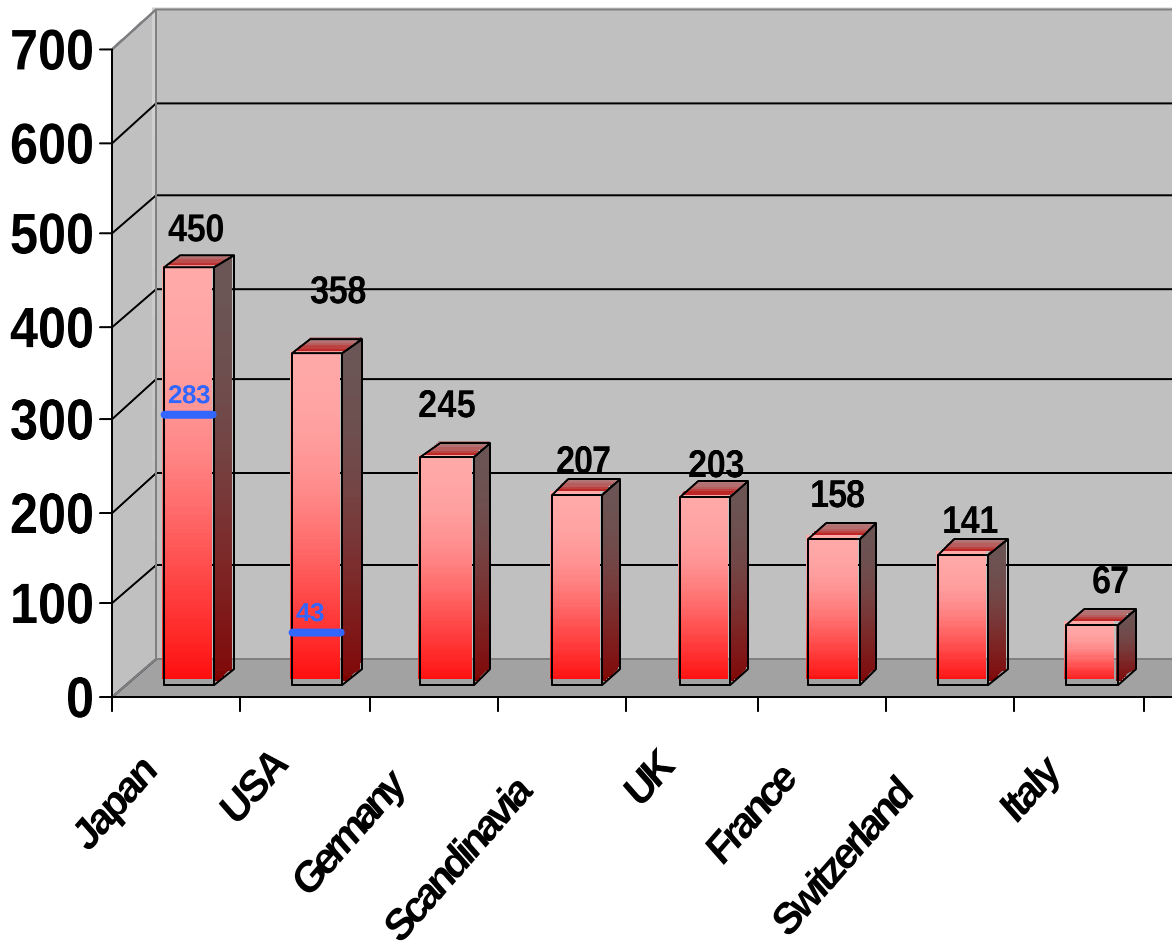
TOP 500: Continents - Performance



TOP 500: Producers - Performance



TOP 500: Kflops per Inhabitant



Current Status of Applications Software

There is a large national investment in scientific software that is dedicated to current massively parallel hardware Architectures

— Scientific Discovery Through Advanced Computing (SciDAC) initiative in DOE
— Accelerated Strategic Computing Initiative (ASCI) in DOE
— Supercomputing Centers of the National Science Foundation (NCSA, NPACI, Pittsburgh)
— Cluster computing in universities and labs

This is a strong a vibrant field. Computational Simulation is well established in the US as the third "leg" of science.

Current Trends in Computer Science Research in the US

The attention of research in computer science is not directed towards scientific supercomputing

- —Primary focus is on Grids and Information Technology
- —Only a handful of supercomputing relevant computer architecture projects currently exist at US universities; versus of the order of 50 in 1992
- —Parallel language and tools research has been almost abandoned
- —Petaflops Initiative (~1997) was not extended beyond the pilot study by any federal sponsors

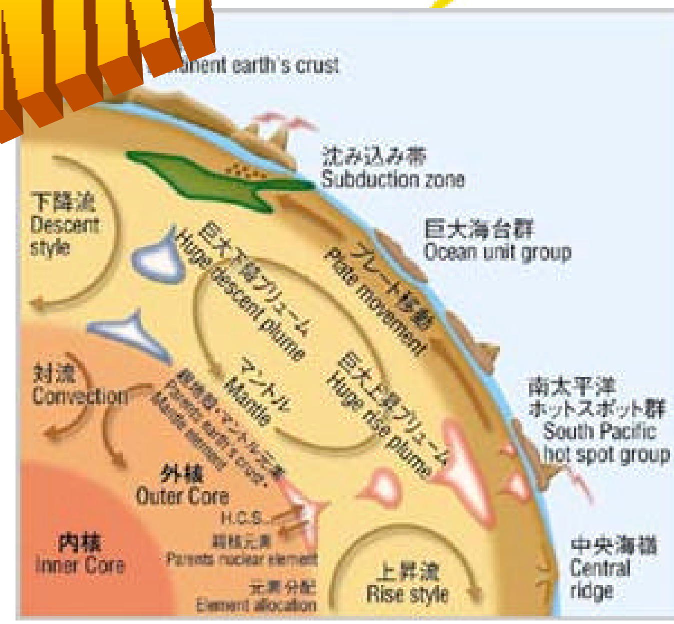
Outline

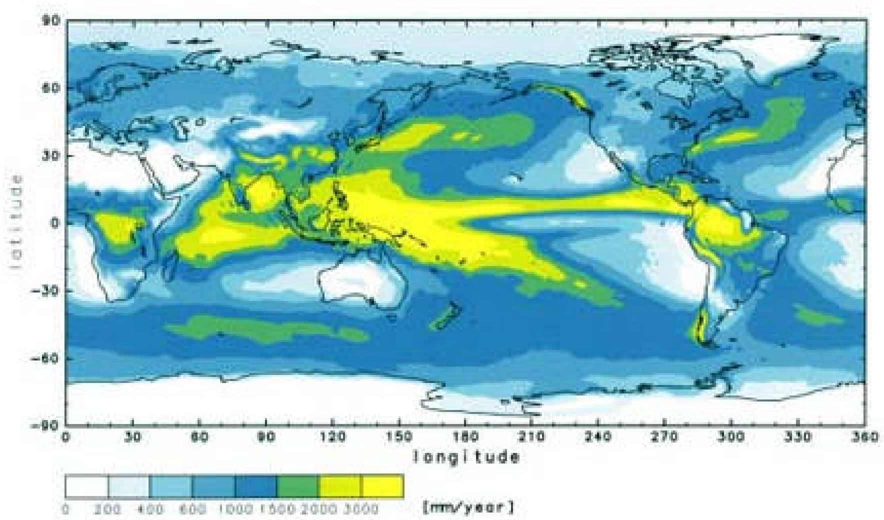
- Where are we today?
  — NERSC examples
  — current status of supercomputing in the US
- The 40 Tflop/s Earth Simulator and the "Computenik" effect
- Business as usual won't work
- Technology Alternatives

The Earth Simulator i… an

COMPUTENIK!

- **Linpack benchn… TF/s = 87% of 4(…**
- **Completed Apr…**
- **Driven by clima… d earthquake simulation**
- **Built by NEC**





http://www.es.jamstec.go.jp/esrdc/eng/menu.html

Understanding and Prediction of Global Climate Change	Understanding of Plate Tectonics
Occurrence prediction of meteorological disaster	Understanding of long-range crustal movements
Occurrence prediction of El Niño	Understanding of mechanism of seismicity
Understanding of effect of global warming	Understanding of migration of underground water and materials transfer in strata
Establishment of simulation technology with 1km resolution	

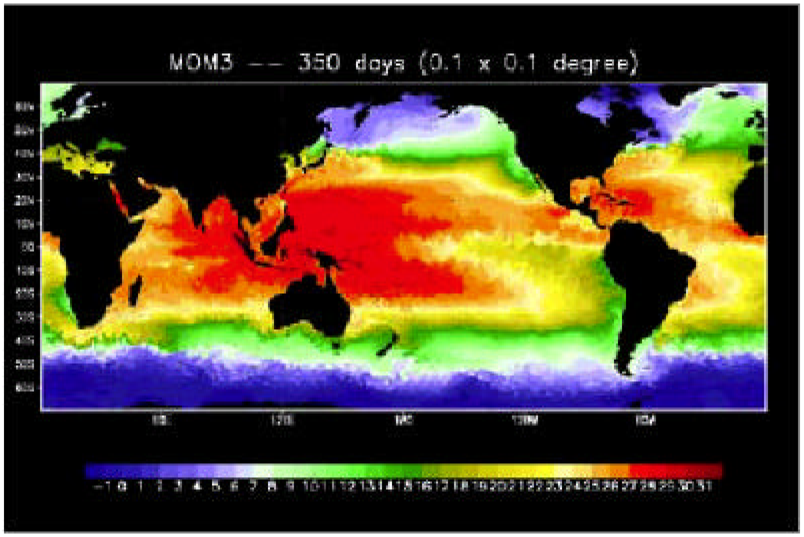
Earth Simulator – Configuration of a General Purpose Supercomputer

- 640 nodes
  - 8 vector processors of 8 GFLOPS and 16GB shared memories per node.
  - Total of 5,120 processors
  - Total 40 Tflop/s peak performance
  - Main memory 10 TB
- High bandwidth (32 GB/s), low latency network connecting nodes.
- Disk
  - 450 TB for systems operations
  - 250 TB for users.
- Mass Storage system: 12 Automatic Cartridge Systems (U.S. made STK PowderHorn9310); total storage capacity is approximately 1.6 PB.

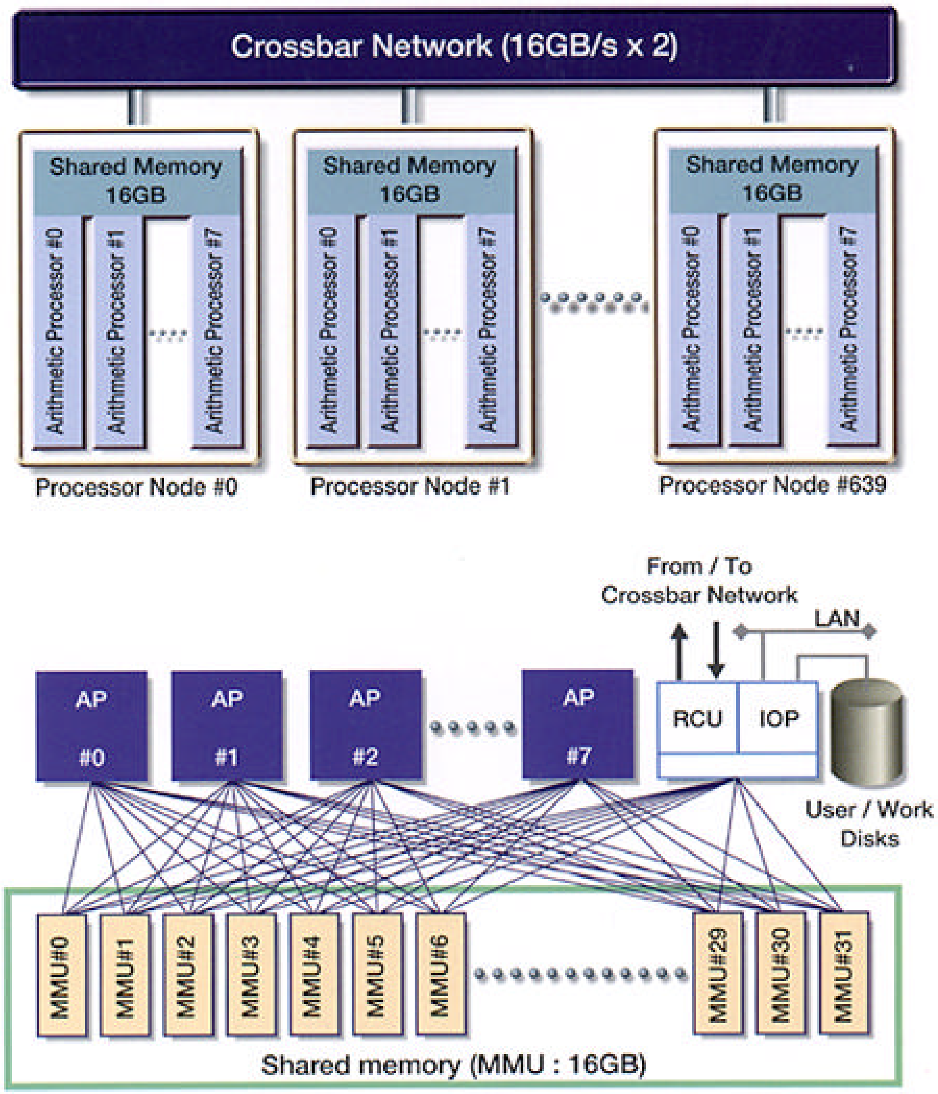
Earth Simulator Performance on Applications

- Test run on global climate model reported sustained performance of 14.5 TFLOPS on 320 nodes (*half the system*): atmospheric general circulation model (spectral code with full physics) with 10 km global grid. The next best climate result reported in the US is about 361 Gflop/s – a factor of 40 less than the Earth Simulator

- MOM3 ocean modeling (code from GFDL/Princeton). The horizontal resolution is 0.1 degrees and the number of vertical layers is 52. It took 275 seconds for a week simulation using 175 nodes. A full scale application result!



Earth Simulator Architecture: Optimizing for the full range of tasks



Parallel Vector Architecture

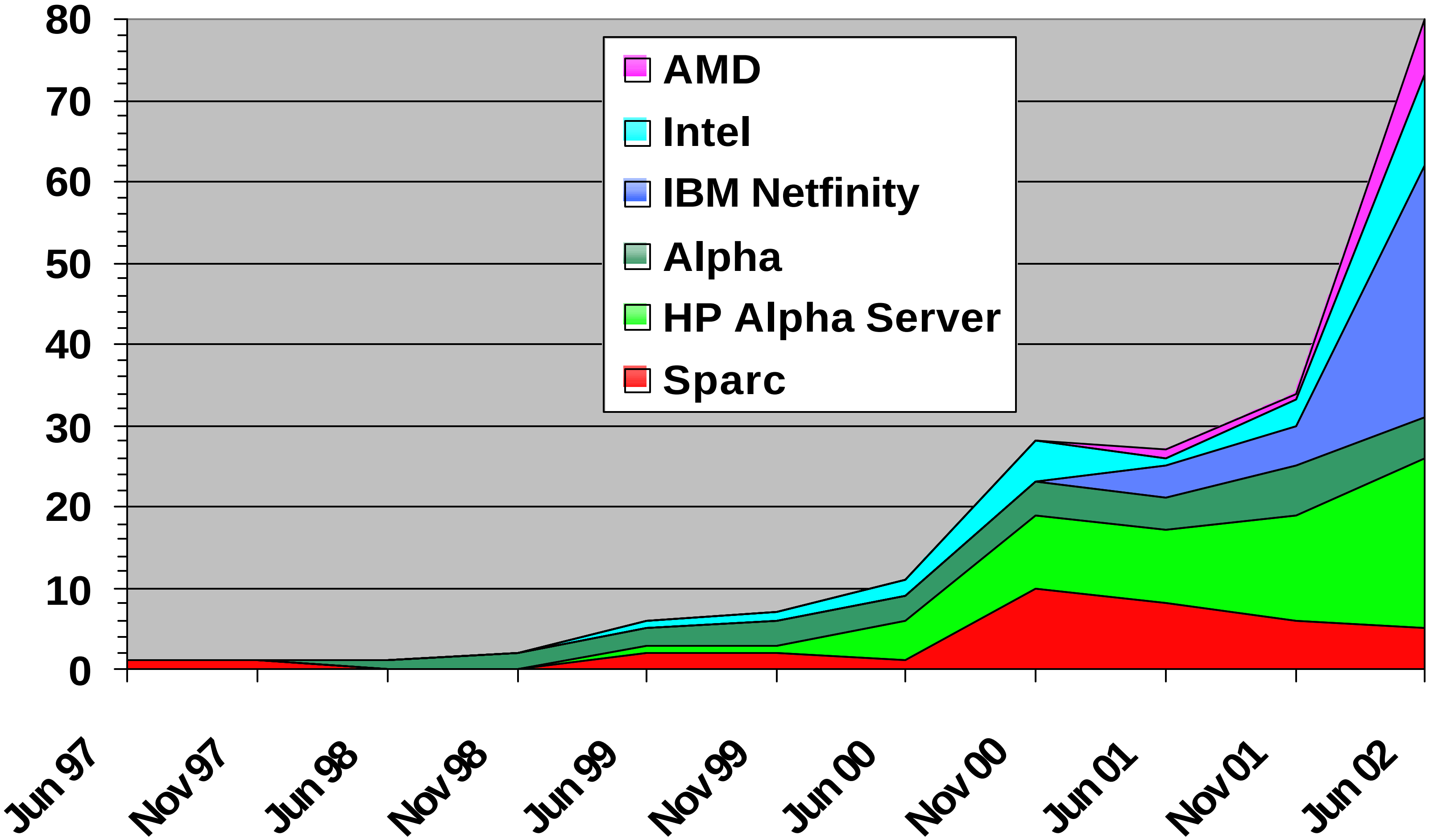
- High speed (vector) processors
- High memory bandwidth (vector architecture)
- Fast network (new crossbar switch)

Rearranging commodity parts can't match this performance

Outline

- Where are we today?
  - — NERSC examples
  - — current status of supercomputing in the US
- The 40 Tflop/s Earth Simulator and the "Computenik" effect
- Business as usual won't work
- Technology Alternatives

Number of NOW Clusters in TOP500



What about PC Clusters? Contributions of Beowulf

- An experiment in parallel computing systems
- Established vision of low cost, high end computing
- Demonstrated effectiveness of PC clusters for some (not all) classes of applications
- Provided networking software
- Conveyed findings to broad community (great PR)
- Tutorials and book
- Design standard to rally community!
- Standards beget: books, trained people, software … virtuous cycle

Adapted from Gordon Bell, presentation at Salishan

Commercially Integrated Tflop/s Clusters Are Already Happening

- Shell: largest engineering/scientific cluster
- NCSA: 1024 processor cluster (IA64)
- Univ. Heidelberg cluster
- PNNL: announced 9.1 Tflops (peak) IA64 cluster from HP with Quadrics interconnect
- DTF in US: announced 4 clusters for a total of 13 Teraflops (peak)

… **But make no mistake: Itanium and McKinley are not a commodity product**

Comparison Between Architectures (2001)

	Alvarez	Seaborg	Mcurie
Processor	Pentium III	Power 3	EV-5
Clock speed	867	375	450
# nodes	80	184	644
# processors/node	2	16	
Peak (GF/s)	139	4416	579.6
Memory (GB/node)	1	16-64	0.256
Interconnect	Myrinet 2000	Colony	T3E
Disk (TB)	1.5	20	2.5

Source: Tammy Welcome, NERSC

Performance Comparison(2)
Class C NPBs

	Alvarez		Seaborg		Mcurie	
	64	128	64	128	64	128
BT	61.0		111.9		55.7	
CG	17.1	13.9	34.0	30.9	9.3	11.8
EP	3.9	3.9	3.9	3.9	2.6	2.6
FT	31.3	20.0	61.2	54.6	30.8	30.1
IS	2.4	2.1	2.1	1.3	1.1	1.0
LU	26.9	38.7	209.0	133.7	60.4	56.0
MG	56.6	46.9	133.2	101.7	93.9	80.0
SP	40.9		100.7		41.8	
per processor	**39.0**		**108.3**		**48.7**	
SSP (Gflops/s)	**6.2**		**318.9**		**31.3**	

Source: Tammy Welcome, NERSC

Effectiveness of Commodity PC Clusters

- Dollars/performance based on peak
  - SP and Alvarez are comparable $/TF
- Get lower % of peak on Alvarez than SP
  - Based on SSP, 4.5% versus 7.2% for FP intensive applications
  - Based on sequential NPBs, 5-13.8% versus 6.3-21.6% for FP intensive applications
  - x86 known not to perform well on FP intensive applications
- $/Performance and cost of ownership need to be examined much more closely
  - Above numbers do not take into account differences in system balance or configuration
  - SP was aggressively priced
  - Alvarez was vendor-integrated, not self-integrated

Source: Tammy Welcome, NERSC

Limits to Cluster Based Systems for HPC

- Memory Bandwidth
  - Commodity memory interfaces [SDRAM, RDRAM, DDRAM]
  - Separation of memory and CPU implementations limits performance
- Communications fabric/CPU/Memory Integration
  - Current networks are attached via I/O devices
  - Limits bandwidth and latency and communication semantics
- Node and system packaging density
  - Commodity components and cooling technologies limit densities
  - Blade based servers moving in right direction but are not High Performance
- Ad Hoc Large-scale Systems Architecture
  - Little functionality for RAS
  - Lack of systems software for production environment
- … but departmental and single applications clusters will be highly successful

After Rick Stevens, Argonne

Cluster of SMP Approach

- A supercomputer is a stretched high-end server
- Parallel system is built by assembling nodes that are modest size, commercial, SMP servers – just put more of them together

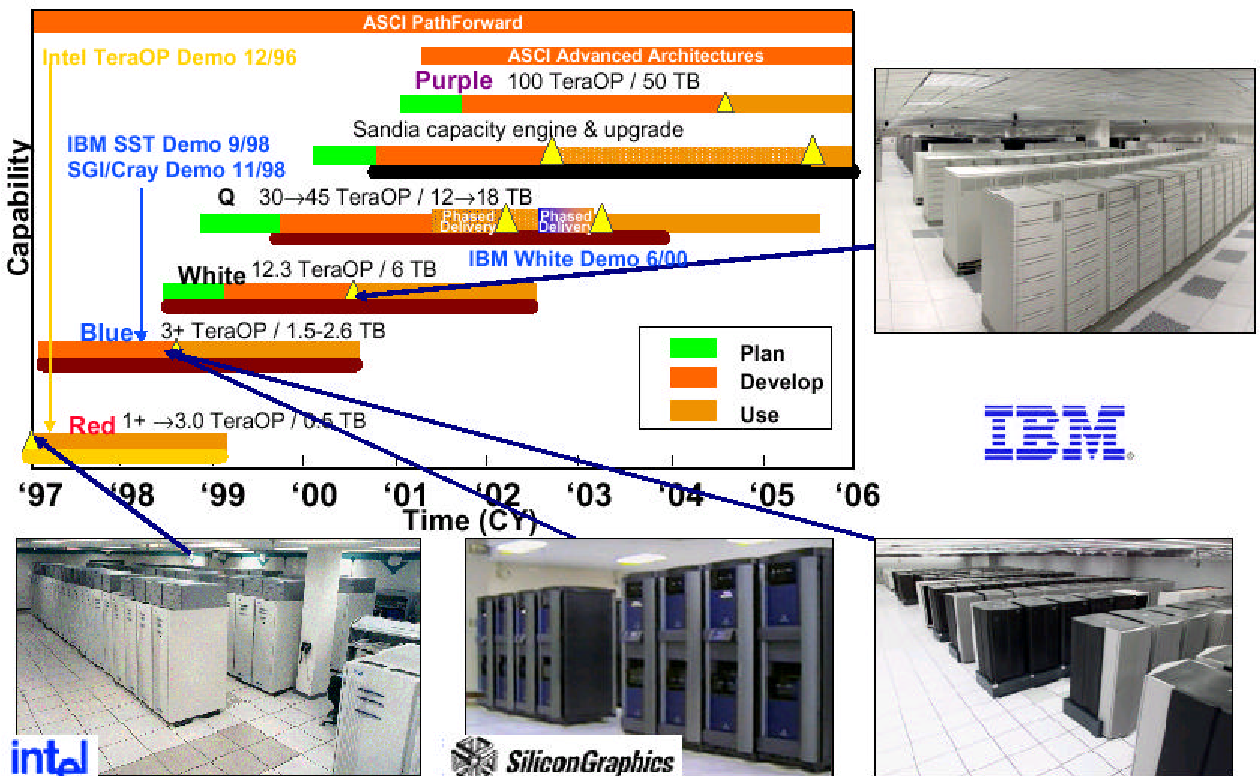


Image from LLNL

UCRL-PRES-147124-7

Comments on ASCI

- Mission focus (stockpile stewardship)
- Computing a tool to accomplish the mission
- Accomplished major milestones
- Success in creating the computing infrastructure in order to meet milestones
- Technology choice in 1995 was appropriate

IBM's Response to the Earth Simulator

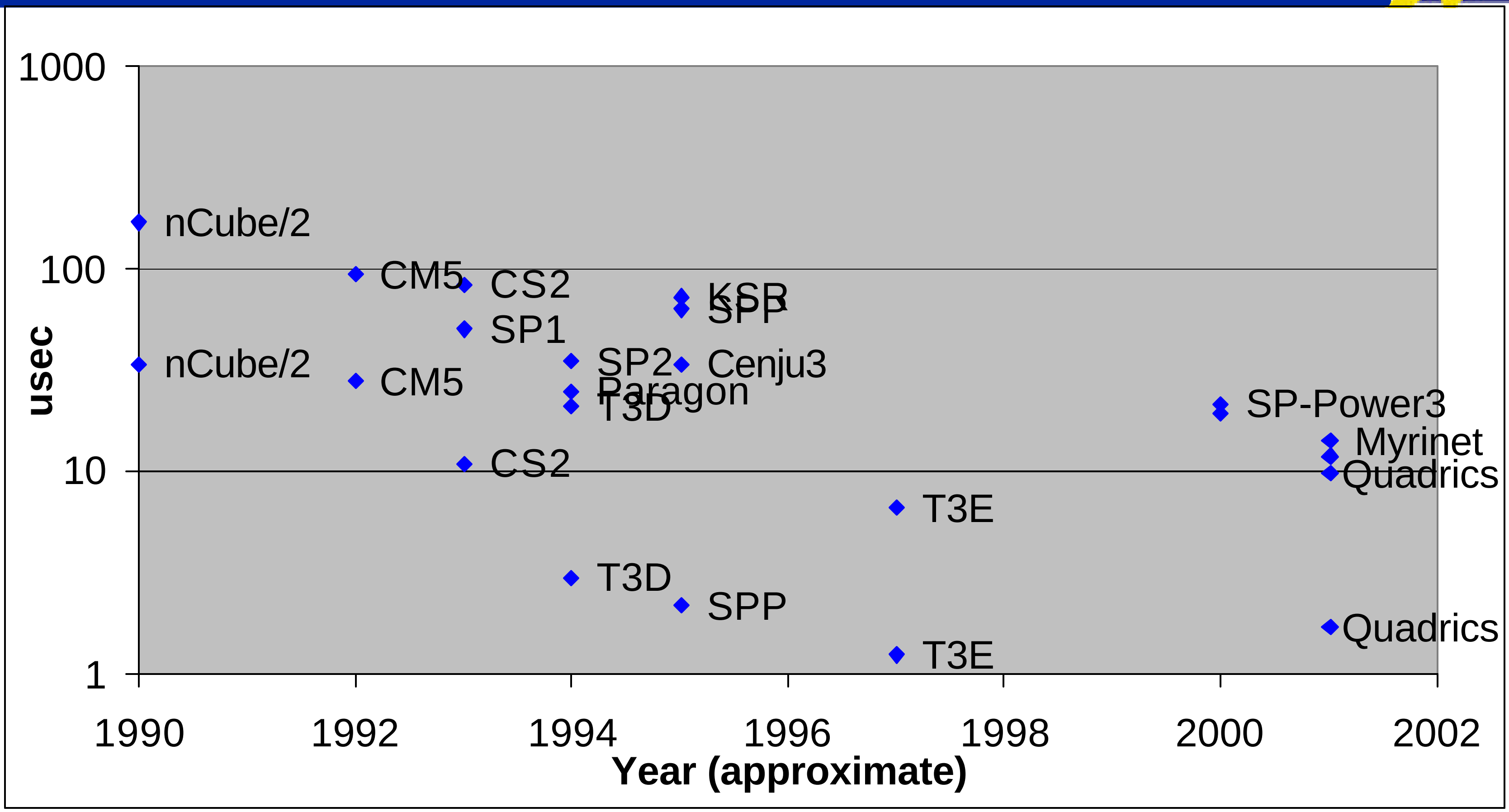
- White paper circulated in Washington in July 2002 states: "… we could construct a supercomputer in 12 to 18 months that would deliver 25 – 50 Tflops of sustainable performance on climate modeling codes … using IBM's next generation interconnect, memory subsystem and processor …"

- "We could do that in a heartbeat and we could do that for a lot less money", Peter Ungarro, IBM as quoted by AP

The IBM Response Ignores Key Issues

- **Interconnect performance**
- **Memory contention on SMP nodes**
- **Processor performance**
- **System Size and Cost**

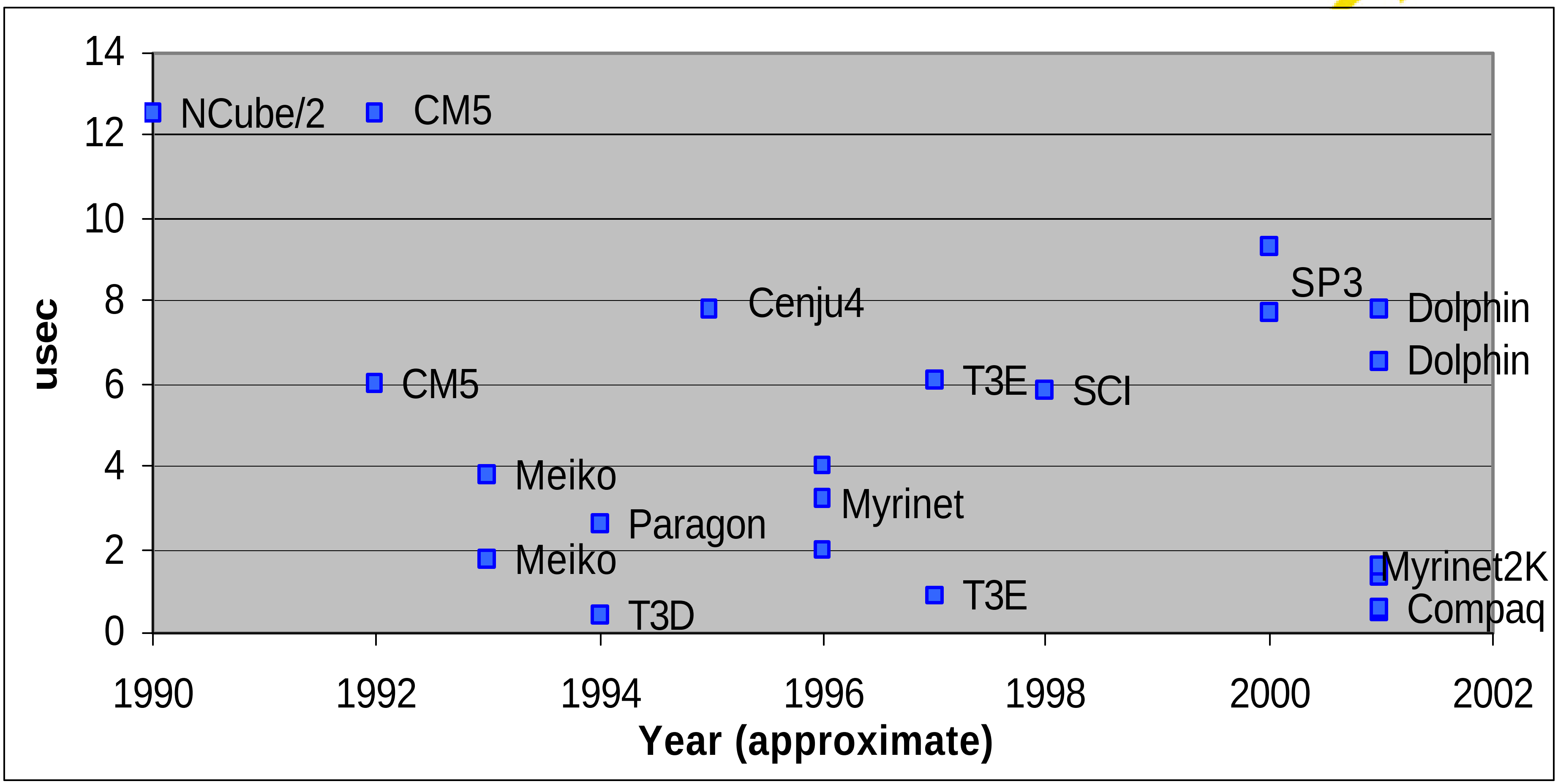
End to End Latency Over Time



- Latency has not improved significantly
  - T3E (shmem) was lowest point
  - Federation in 2003 will not reach that level – 7 years later!

Data from Kathy Yelick, UCB and NERSC

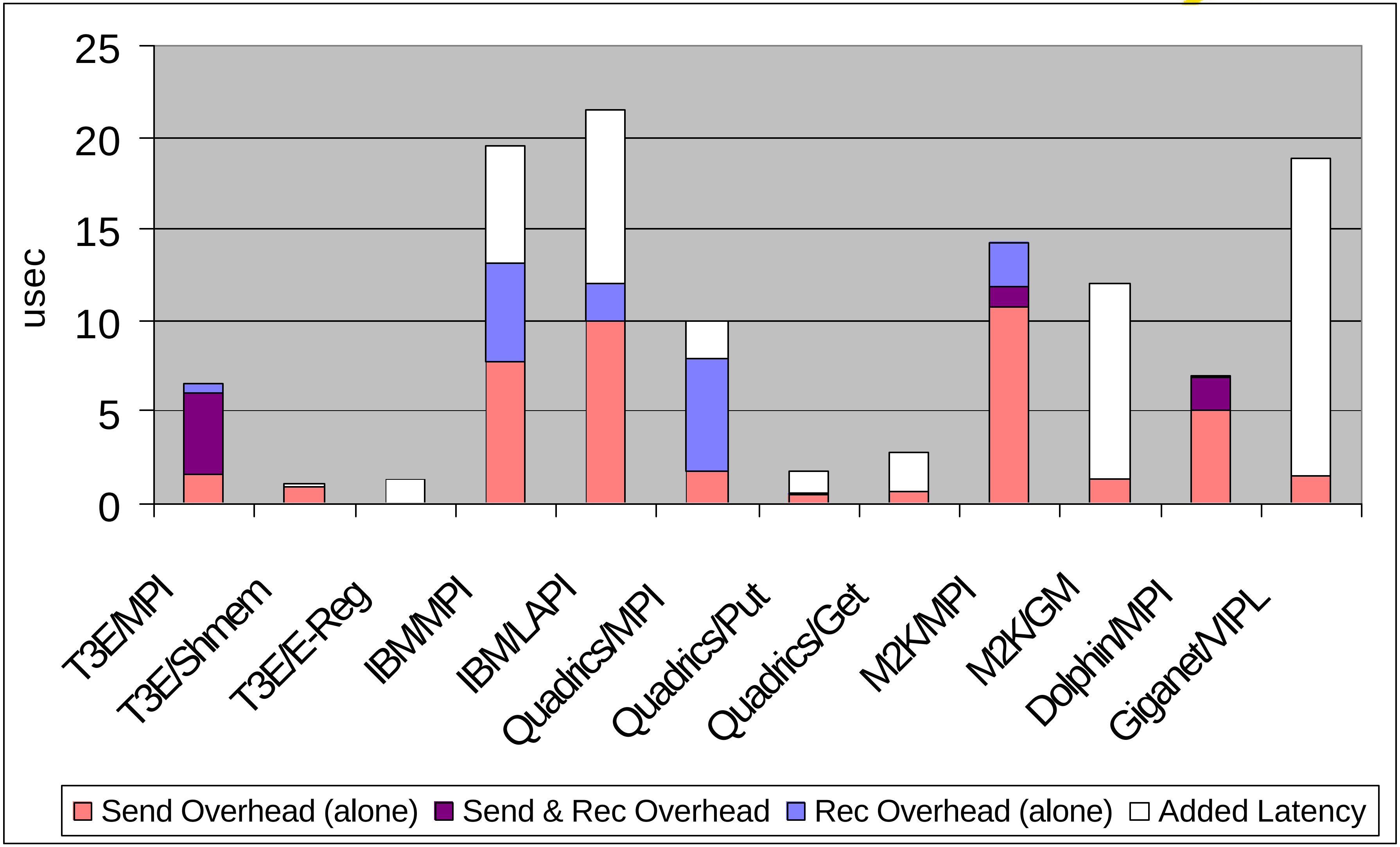
Send Overhead Over Time



- Overhead has not improved significantly; T3D was best
  — Lack of integration; lack of attention in software

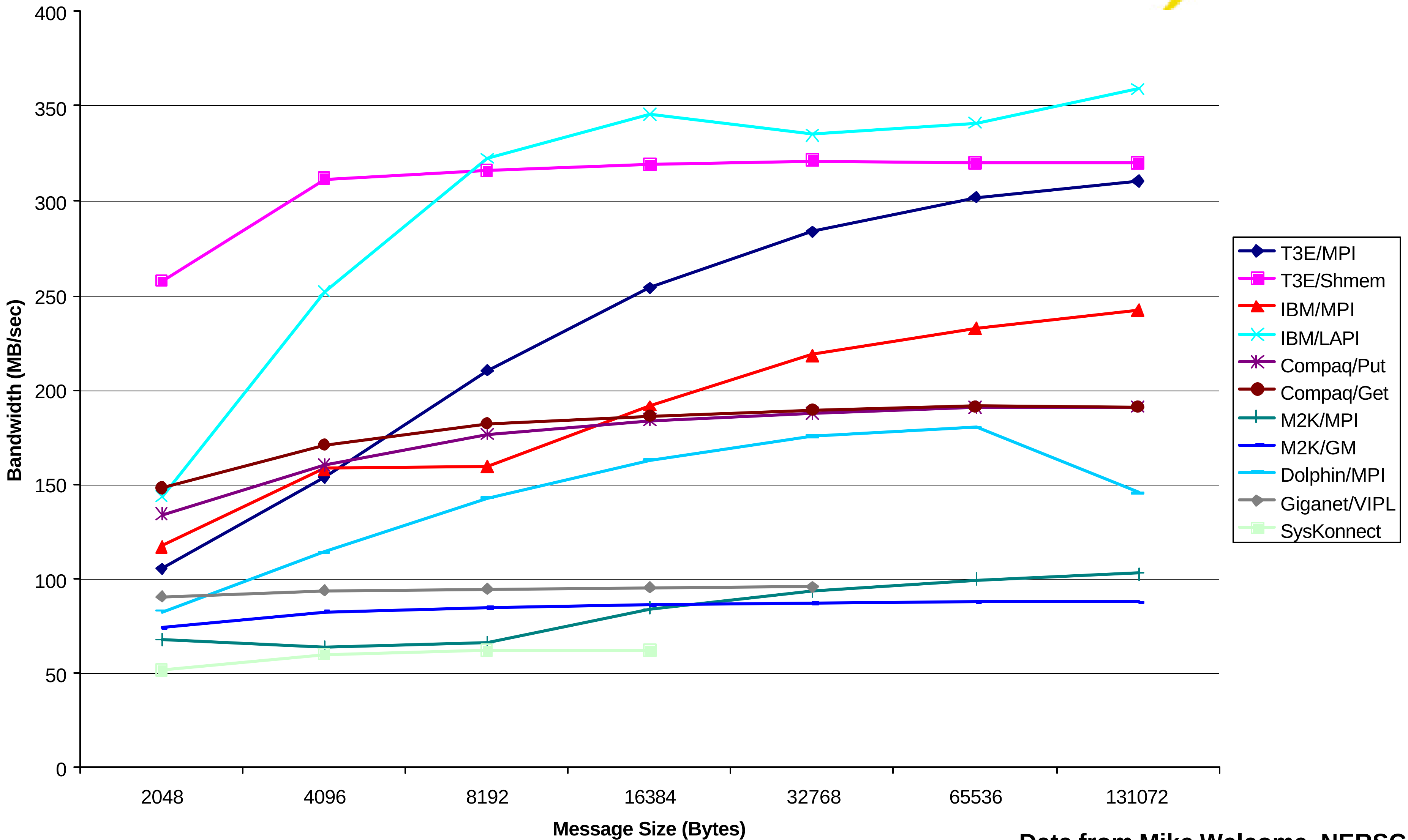
Data from Kathy Yelick, UCB and NERSC

Results: EEL and Overhead



Data from Mike Welcome, NERSC

Bandwidth Chart



Data from Mike Welcome, NERSC

PSTSWM Description

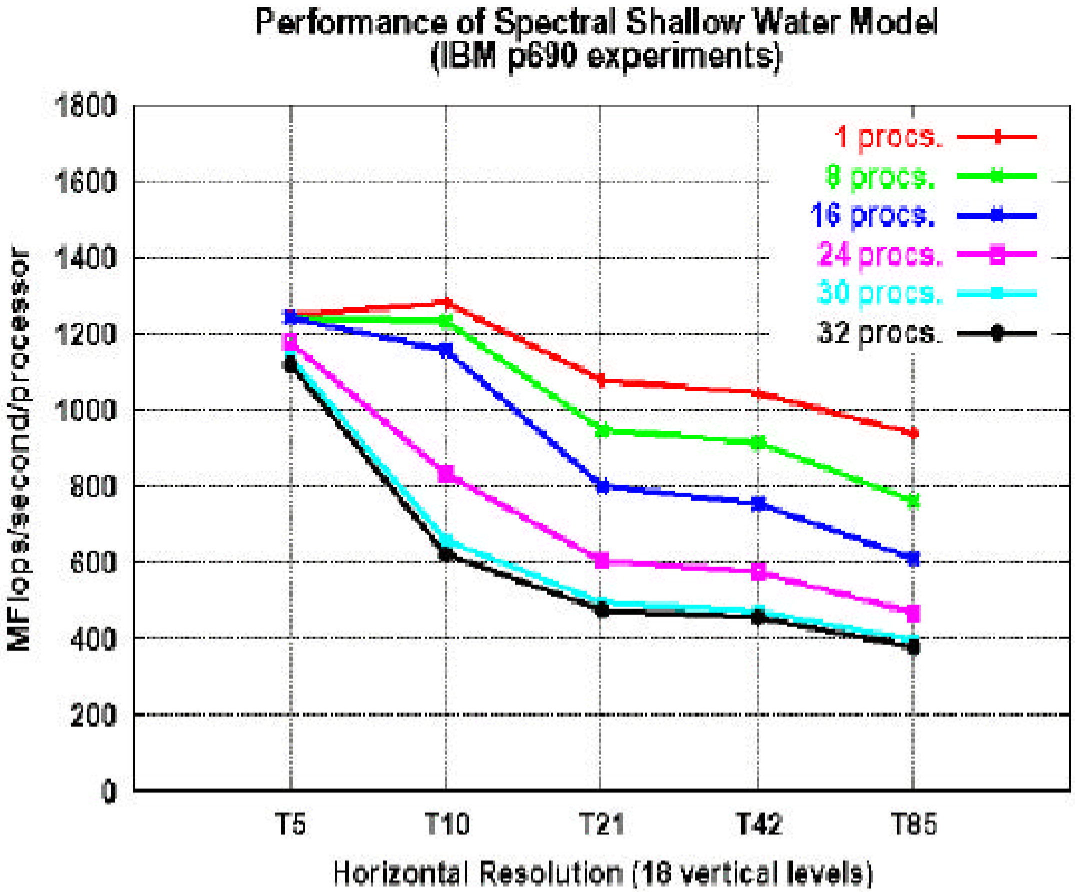
- The Parallel Spectral Transform Shallow Water Model represents an important computational kernel in spectral global atmospheric models. As 99% of the floating-point operations are multiply or add, it runs well on systems optimized for these operations. PSTSWM exhibits little reuse of operands as it sweeps through the field arrays; thus it exercises the memory subsystem as the problem size is scaled and can be used to evaluate the impact of memory contention in SMP nodes.
- These experiments examine serial performance, both using one processor and running the serial benchmark on multiple processors simultaneously. Performance is measured for a range of horizontal problems resolutions for 18 vertical levels.

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY

From Pat Worley, ORNL

PSTSWM Sensitivity to Contention

- Performance degradation is a “smooth” function of number of processes.



Process scaling on IBM p690

OAK RIDGE NATIONAL LABORATORY
U. S. DEPARTMENT OF ENERGY

From Pat Worley, ORNL

For the Next Decade, The Most Powerful Supercomputers Will Increase in Size

This

Became

And will get bigger

Power and cooling are also increasingly problematic, but there are limiting forces in those areas.

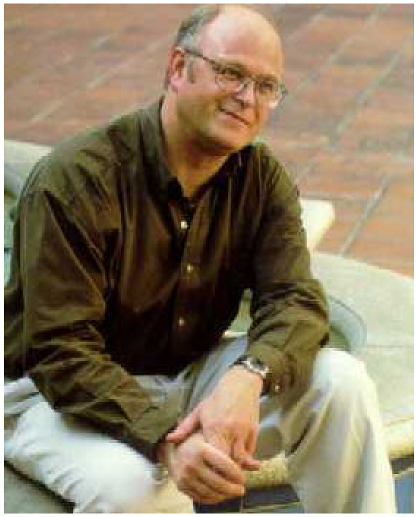
- Increased power density and RF leakage power, will limit clock frequency and amount of logic *[Shekhar Borkar, Intel]*
- So linear extrapolation of operating temperatures to Rocket Nozzle values by 2010 is likely to be wrong.

The Oakland Facility Machine Room




“I used to think computer architecture was about how to organize gates and chips – not about building computer rooms”

Thomas Sterling, Salishan, 2001



NERSC Analysis

- A system of 512 Power4 nodes, with a 1024-way Federation switch (2 adapters per node), would have a tough time achieving 5 Tflop/s without extensive optimization.
- Even assuming a 2X speedup from tuning, maybe IBM could hit 10 Tflop/s.
- 20 Tflop/s or more would require "miracles" of tuning, or, more likely, a significantly larger (and more expensive) system.
- Such a system would cost $200 million or more and require 40,000 sqft floor space or more
- … and we still would be in second place

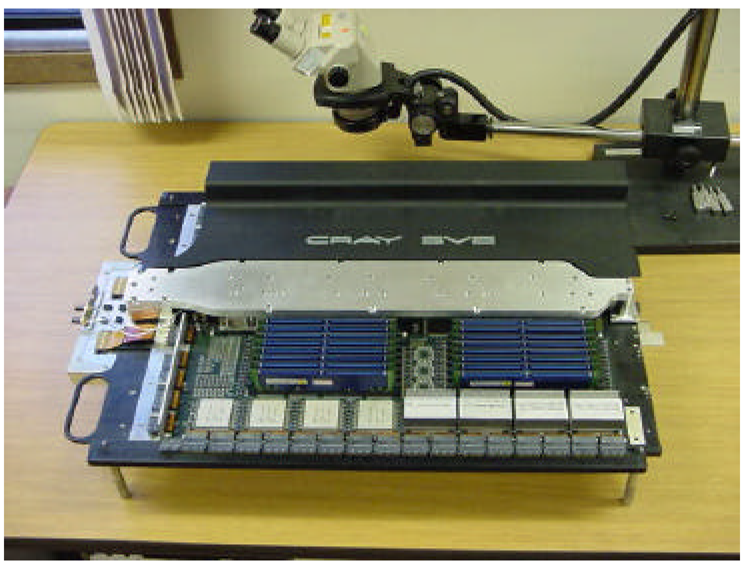
Quoted from David Bailey

Outline

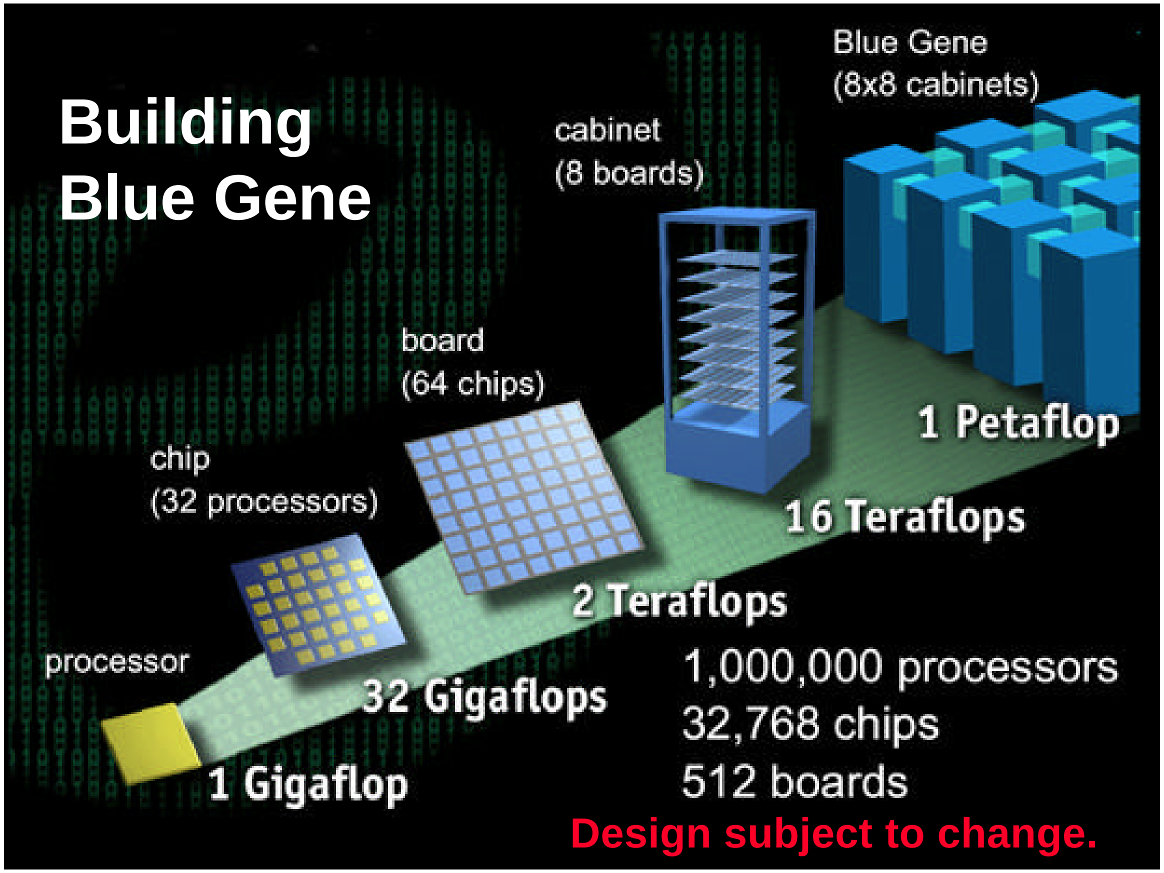
- Where are we today?
  — NERSC examples
  — current status of supercomputing in the US
- The 40 Tflop/s Earth Simulator and the “Computenik” effect
- Business as usual won’t work
- Technology Alternatives

Cray SV2: *Parallel Vector Architecture*

- 12.8 Gflop/s Vector processors
- 4 processor nodes sharing up to 64 GB of memory
- Single System Image to 4096 Processors
- 64 CPUs/800 GFLOPS in LC cabinet



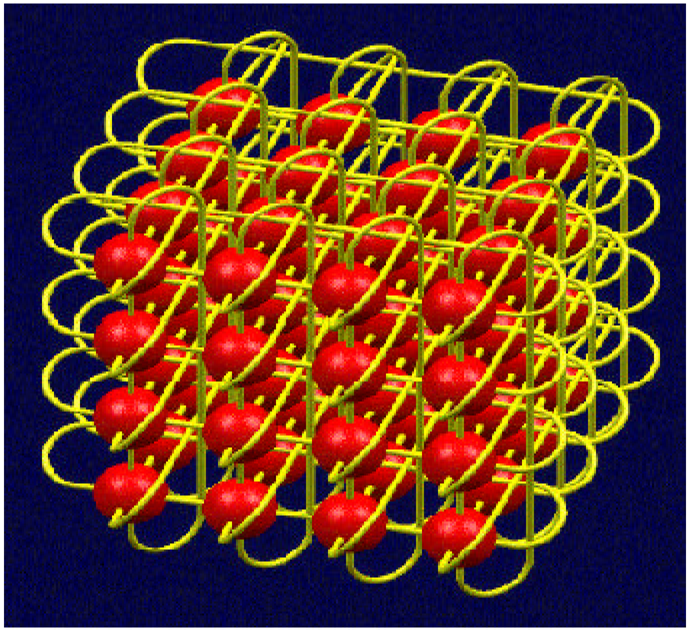
CMOS Petaflop/s Solution



- IBM's Blue Gene
- 64,000 32 Gflop/s PIM chips
- Sustain O(10^7) ops/cycle to avoid Amdahl bottleneck

Characteristics of Blue Gene/L

- **Machine Peak Speed 180 Teraflop/s**
- **Total Memory 16 Terabytes**
- **Foot Print 2500 sq. ft.**
- **Total Power 1.2 MW**
- **Number of Nodes 65,536**
- **Power Dissipation/CPU 7 W**
- **MPI Latency 5 microsec**



Building Blue Gene/L

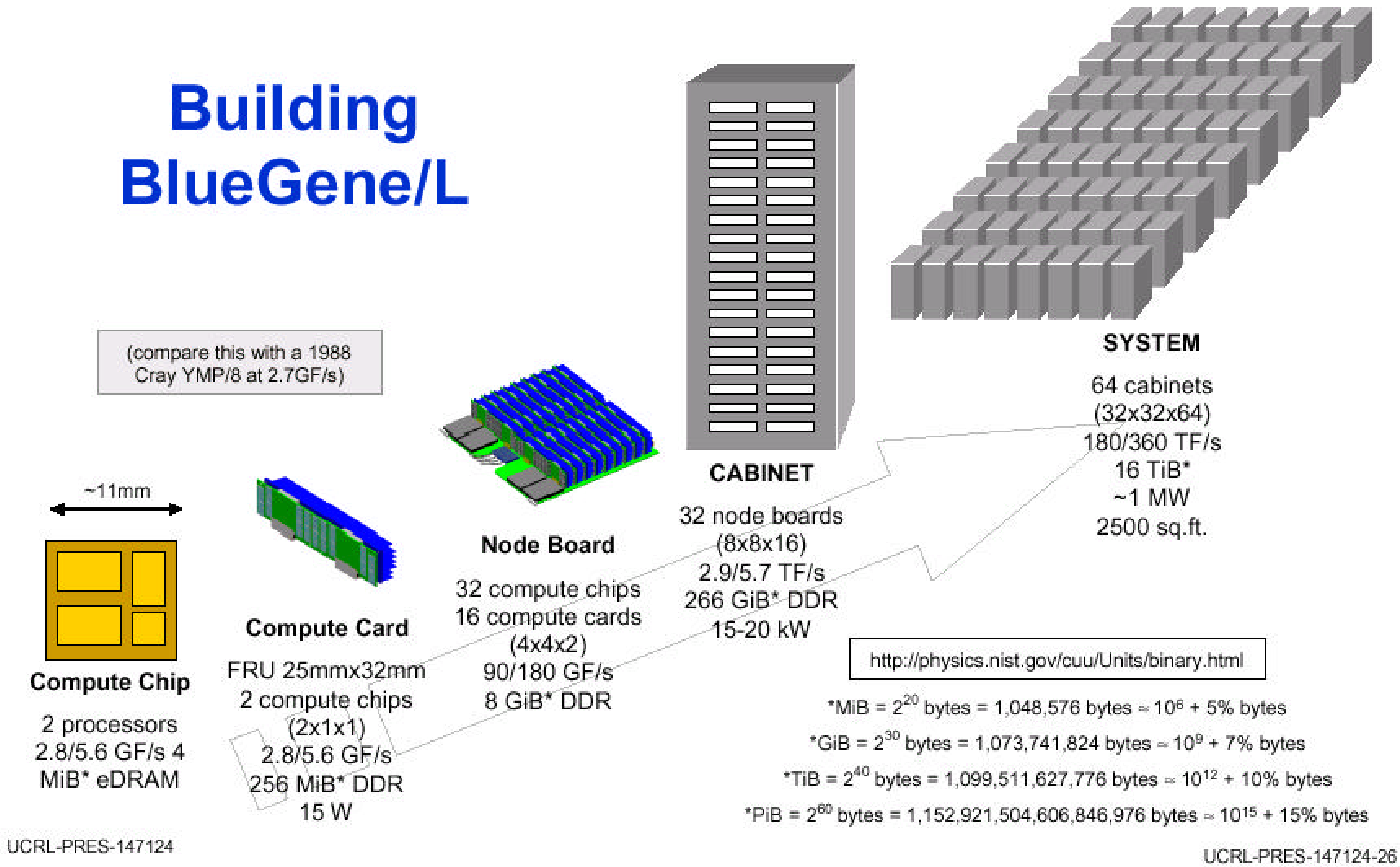
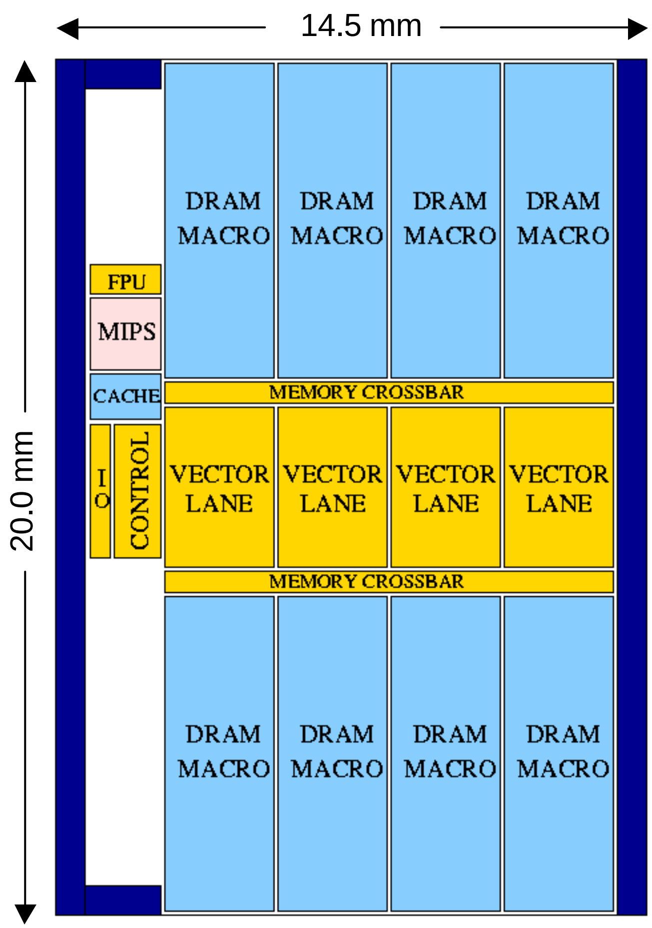


Image from LLNL

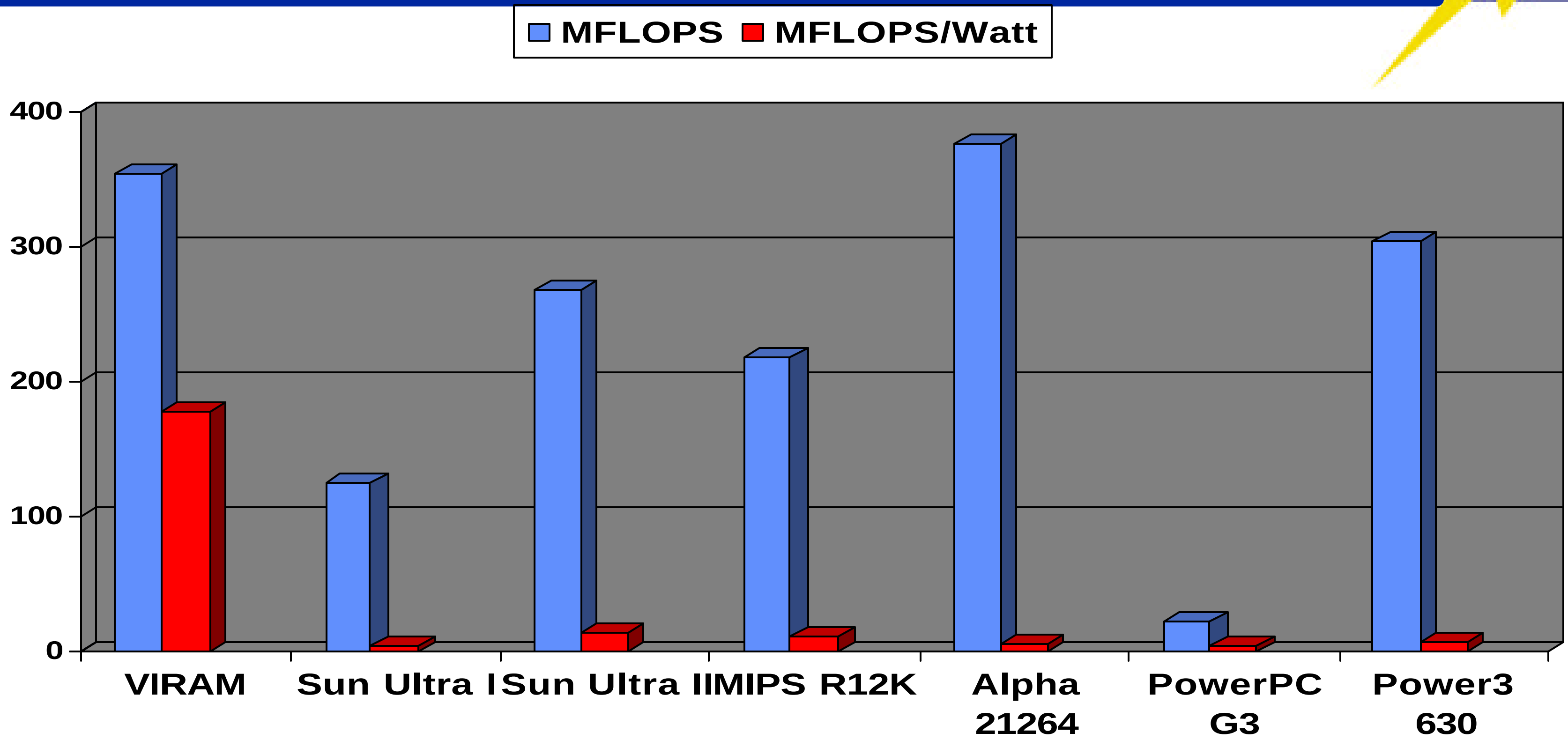
VIRAM Overview (UCB)



- MIPS core (200 MHz)
  - Single-issue, 8 Kbyte I&D caches
- Vector unit (200 MHz)
  - 32 64b elements per register
  - 256b datapaths, (16b, 32b, 64b ops)
  - 4 address generation units
- Main memory system
  - 12 MB of on-chip DRAM in 8 banks
  - 12.8 GBytes/s peak bandwidth
- Typical power consumption: 2.0 W
- Peak vector performance
  - 1.6/3.2/6.4 Gops wo. multiply-add
  - 1.6 Gflops (single-precision)
- Same process technology as Blue Gene
  - But for single chip for multi-media

Source: Kathy Yelick, UCB and NERSC

Power Advantage of PIM+Vectors



- 100x100 matrix vector multiplication (column layout)
  - —Results from the LAPACK manual (vendor optimized assembly)
  - —VIRAM performance improves with larger matrices!
  - —VIRAM power includes on-chip main memory!

Source: Kathy Yelick, UCB and NERSC, paper at IPDPS 2002

Explore the Design Space

- In between a fully custom HPC system and a COTS cluster, there are many design points where standard components are mixed with custom technology to enhance performance.
- This spectrum ranges from "pick what Dell/IBM/HP/... would do anyhow" to "build the dream machine -- cost is not an issue".
- They key is to decide what are the main performance enhancers: a custom switch, a custom package, etc.
- We seem to always veer toward the two extremes: we have Cray/Tera/... at one extreme (custom microprocessor, custom interconnect, custom package...), and Beowulfs/SPs/... at the other extreme (all commercial server technology).
- No interesting design point is explored in between and no commercial model exists to support something in between.
- We need to explore the design space in collaboration with vendors such as IBM, who have all technology easily available, or can be the integrator

After Marc Snir, UIUC

Options for New Architectures

Option	Software Impact	Cost	Timeliness	Risk Factors
Modification of commodity processors	Minimal	2 or 3 times commodity?	Can be achieved in three years	Partnership with vendors not yet established
U.S. made vector architecture	Moderate	2 or 3 times commodity at present	Deliverable in 2003 and beyond	One small vendor
Processor-in-memory (Blue Gene/L)	Extensive	Unknown, 2 to 5 time commodity?	Only prototypes available now	General purpose applicability unknown
Japanese made vector architecture	Moderate	2.5 to 3 times commodity at present	Available now	Political risk, unknown future availability and growth path
Research Architectures (Streams, VIRAM …)	Extensive or unknown	Unknown	Academic research prototypes only available now	Not practical in five years

Two Options are Preferred

1. Adapt commercial microprocessors with modifications geared towards scientific applications
   - Goal: Address the memory bandwidth and communications problem
   - Precedent: One of the most successful massively parallel computers was the Cray T3E based on commodity processors with additional special purpose components
   - Practicality: Major vendors now have robust “embedded processor” businesses which can make the parts
   - Method: Partnership and investment with U.S. vendor(s).
2. U.S. made Vector-based Massively Parallel architectures
   - Practicality: Design is extension of existing architecture
   - Method: Partnership and investment with the U.S. vendor