Five Trends in Supercomputing for the Next Five Years

Horst D. Simon
Director
National Energy Research Scientific Computing Center
(NERSC)
Berkeley, California, USA

July 2002

“Per Aspera Ad Astra”

Dedicated to Prof. Dr. Friedel Hossfeld
on the occasion of his retirement

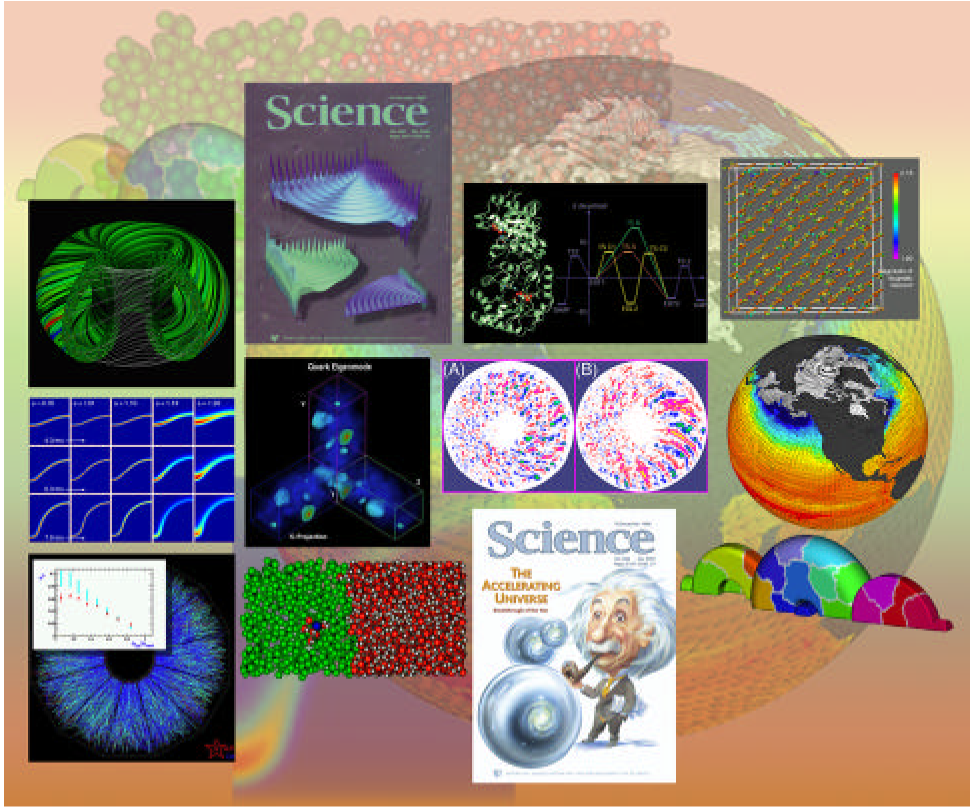
July 11, 2002

NERSC Overview

- 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-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

TOP500 – June 2002

NERSC

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 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

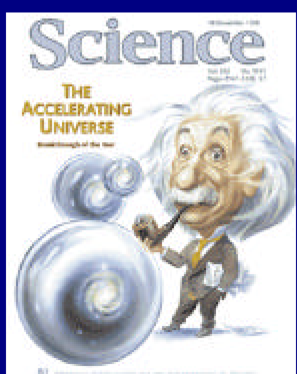
2000: BOOMERANG data analysis= flat universe

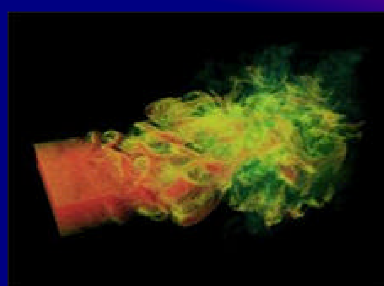
2001: Most distant supernova

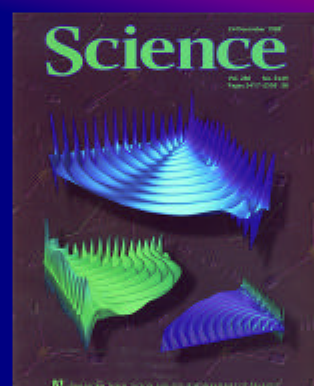
1996

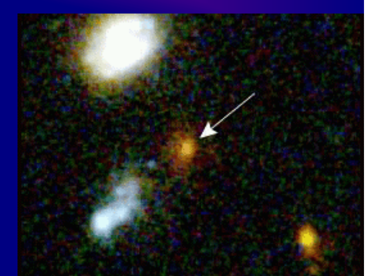
National Energy Research Scientific Computing Center

2002





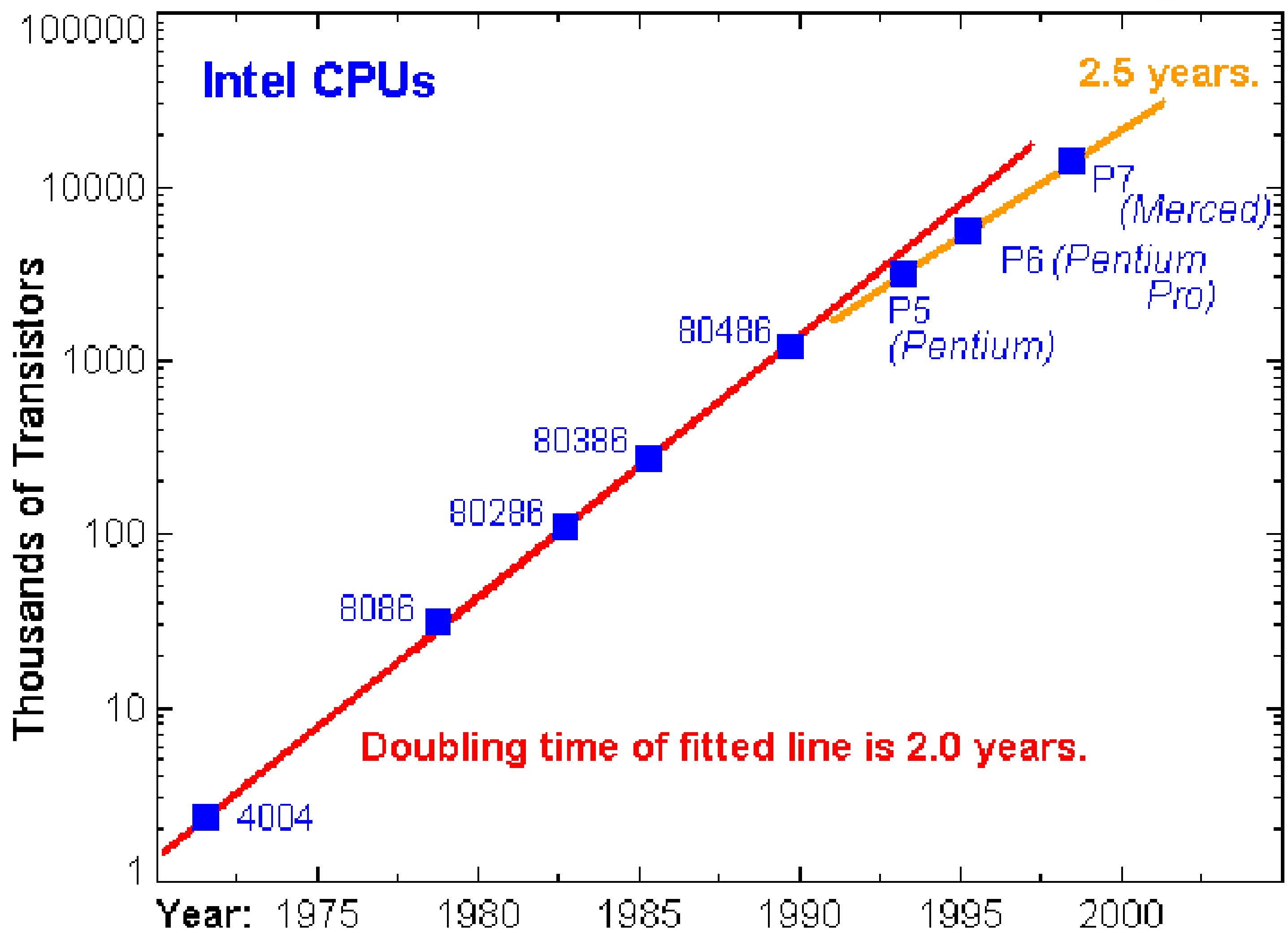




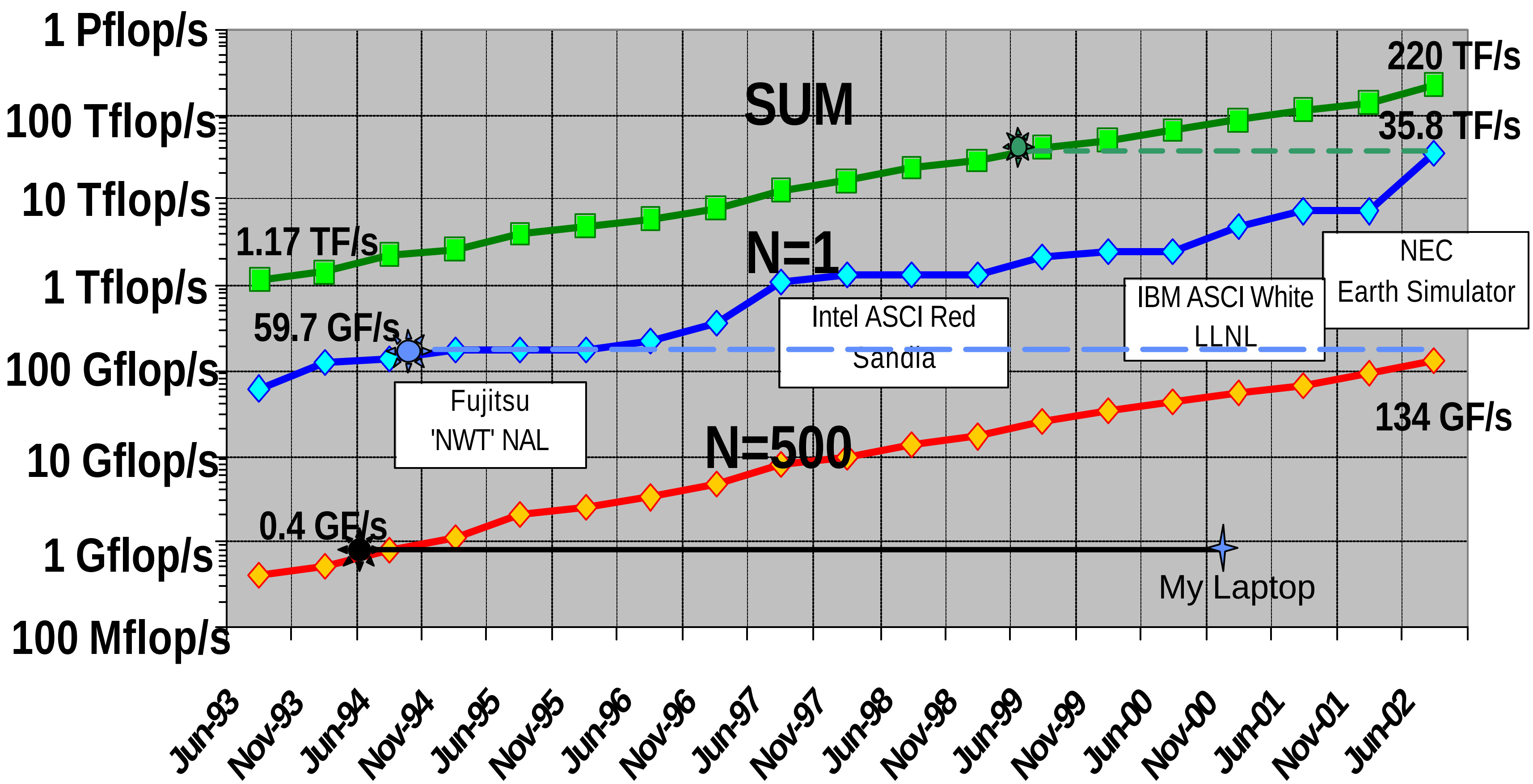
Five Computing Trends for the Next Five Years

- Continued rapid processor performance growth following Moore's law
- Open software model (Linux) will become standard
- Network bandwidth will grow at an even faster rate than Moore's Law
- Aggregation, centralization, colocation
- Commodity products everywhere

Moore's Law —
The Traditional (Linear) View



TOP500 - Performance

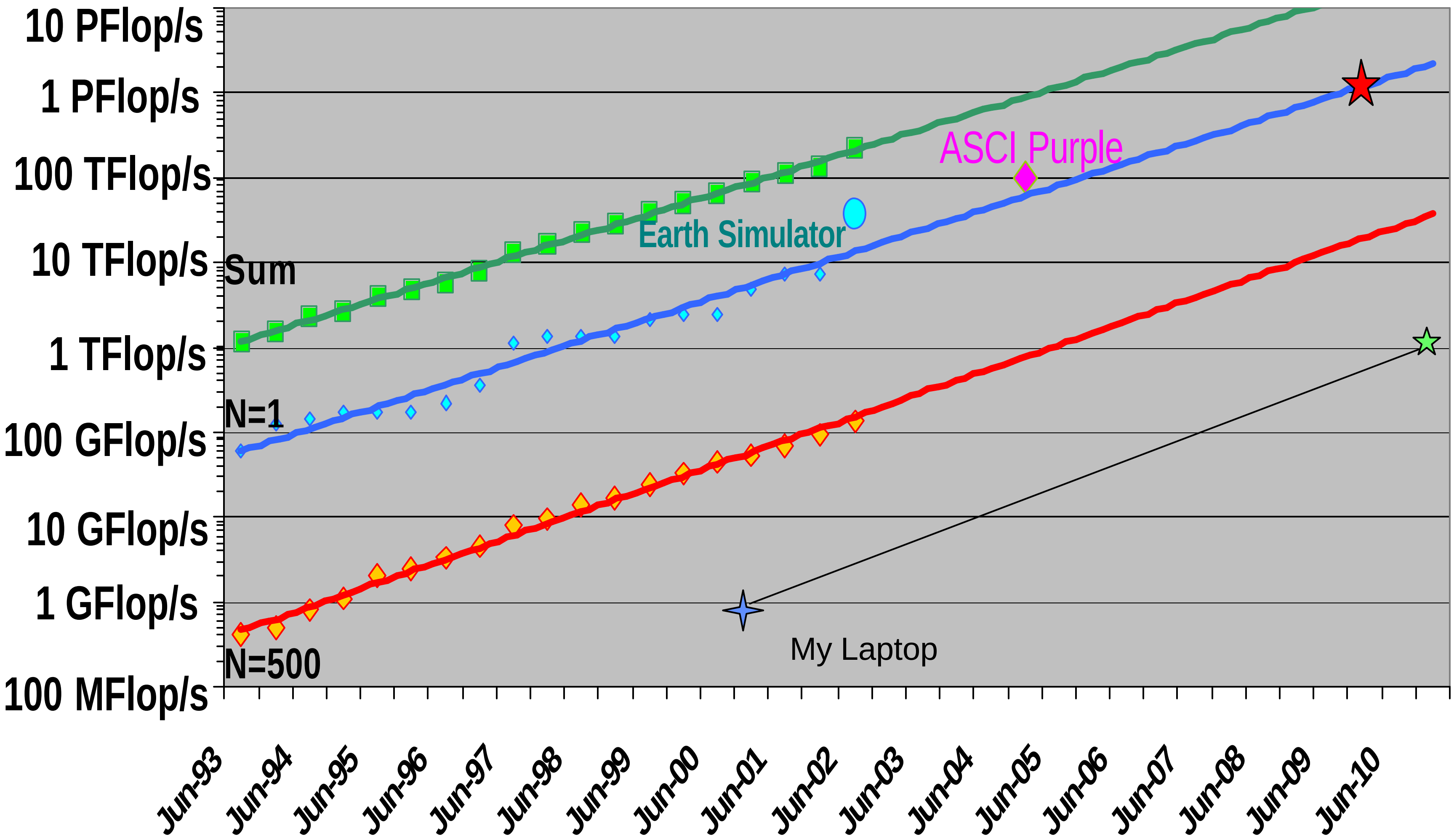


Analysis of TOP500 Data

- **Annual performance growth about a factor of 1.82**
- **Two factors contribute almost equally to the annual total performance growth**
- **Processor number grows per year on the average by a factor of 1.30 and the**
- **Processor performance grows by 1.40 compared to 1.58 of Moore's Law**

Strohmaier, Dongarra, Meuer, and Simon, Parallel Computing 25, 1999, pp 1517-1544.

Performance Extrapolation



Analysis of TOP500 Extrapolation

Based on the extrapolation from these fits we predict:

- **First 100~TFlop/s system by 2005**
- **About 1–2 years later than the ASCI path forward plans.**
- **No system smaller than 1 TFlop/s should be able to make the TOP500**
- **First Petaflop system available around 2009**
- **Rapid changes in the technologies used in HPC systems, therefore a projection for the architecture/technology is difficult**
- **Continue to expect rapid cycles of re-definition**

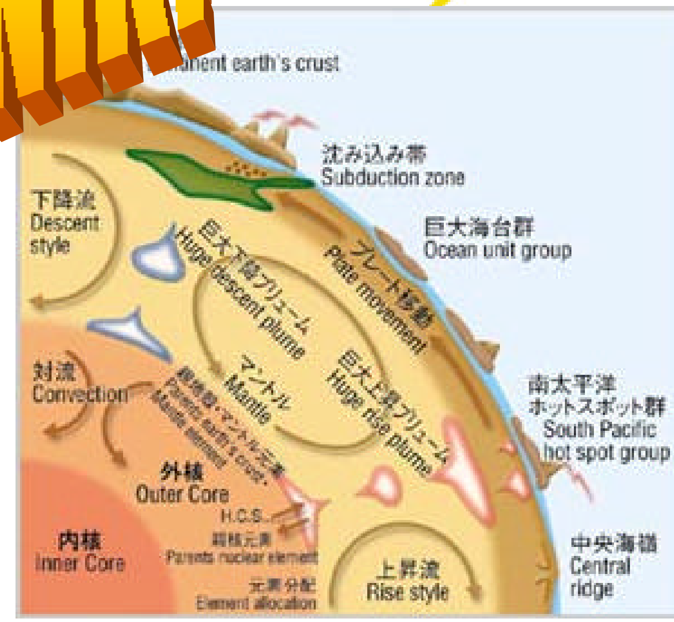
TOP500 – June 2002

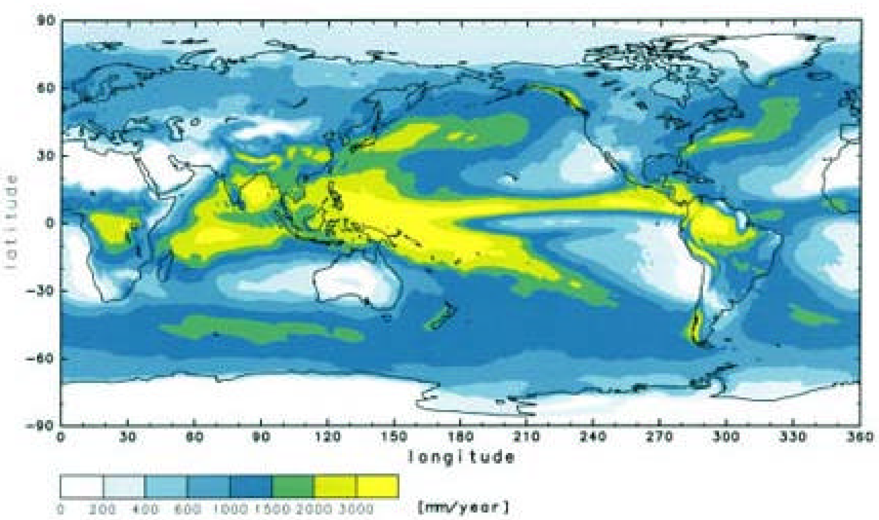
Rank	Manufacturer	Computer	Rmax	Installation Site	Country	Year	Area of Installation	# Proc	Rpeak	Nmax	N1/2
1	NEC	Earth-Simulator	**35860**	Earth Simulator Center Kanazawa	Japan	2002	Research	5120	40960	1075200	266240
2	IBM	ASCI White, SP Power3 375 MHz	**7226**	Lawrence Livermore National Laboratory 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							

The Earth Simulator i… an

COMPUTENIK!

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

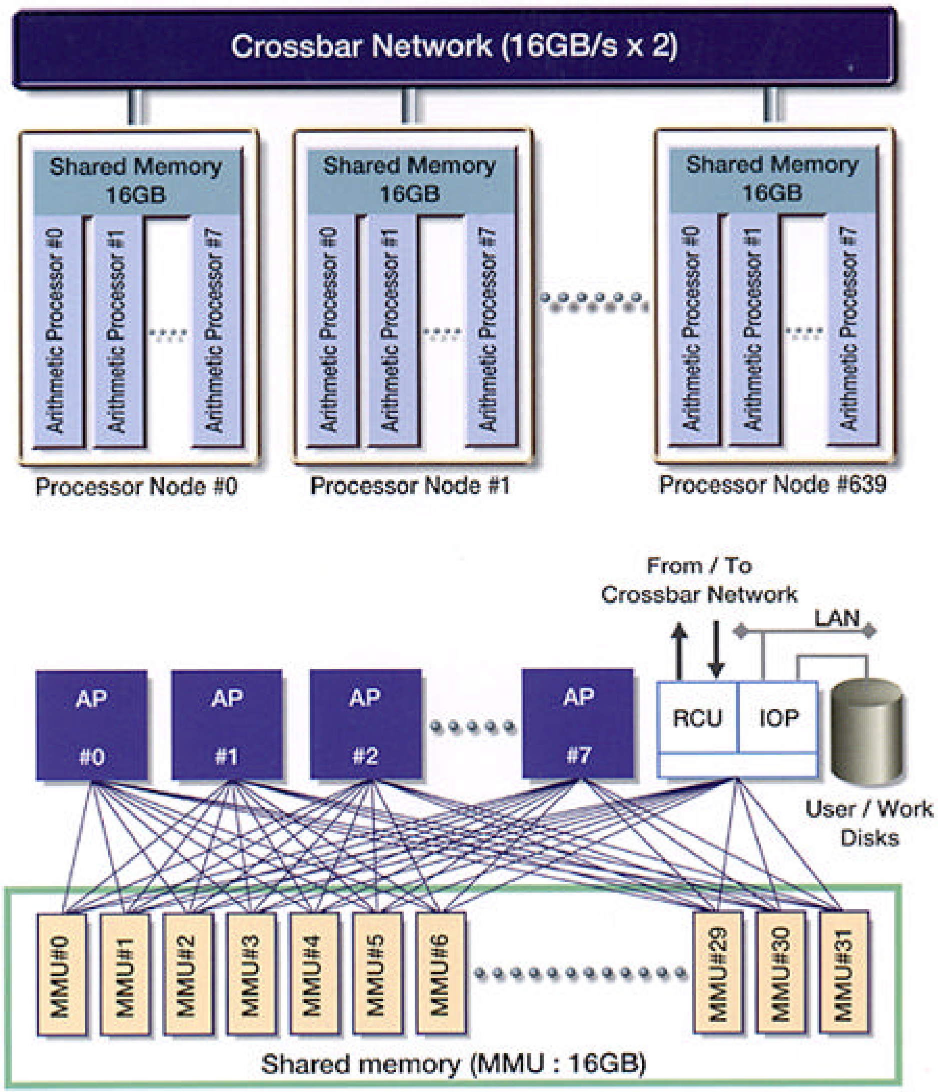




http://www.es.jamstec.go.jp/esrdc/eng/menu.html

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 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

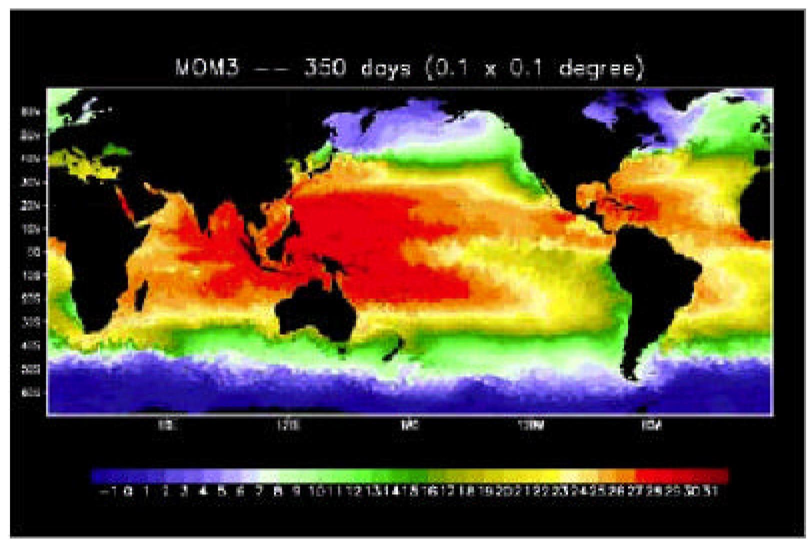
Earth Simulator – Configuration of a General Purpose Supercomputer

- 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!



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

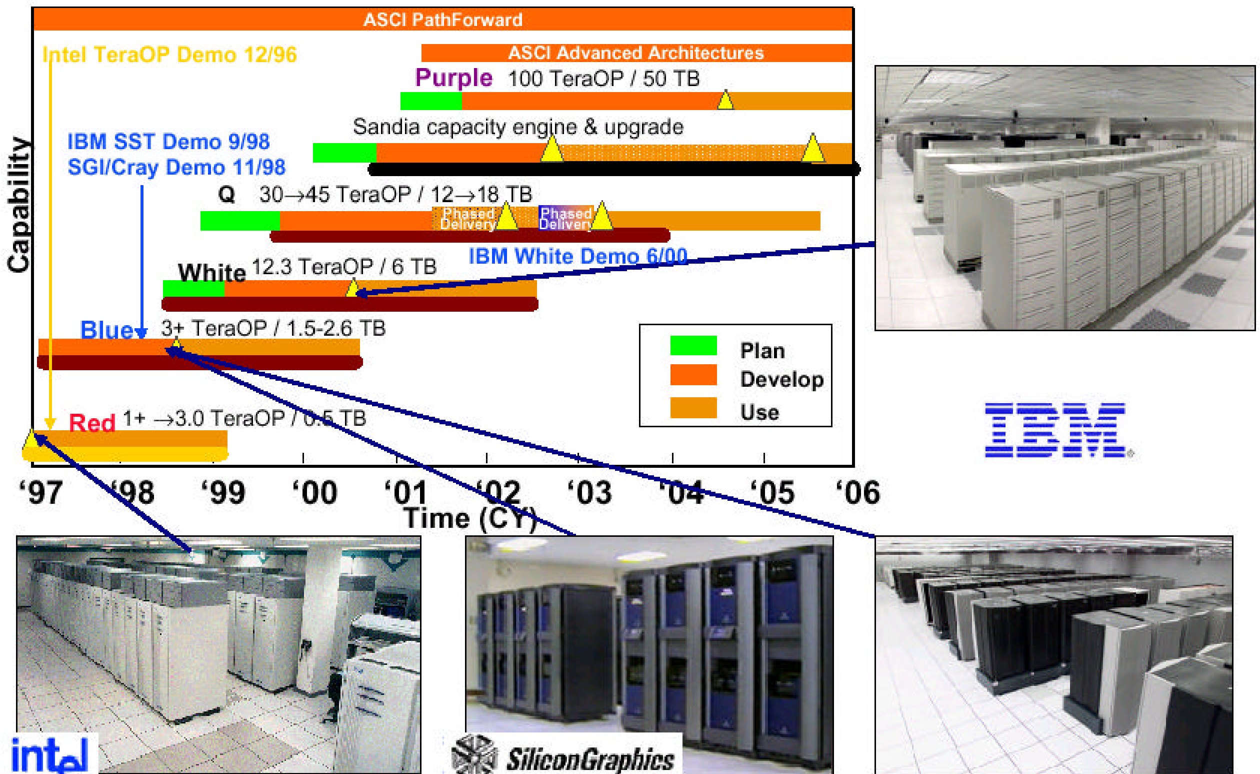


Image from LLNL

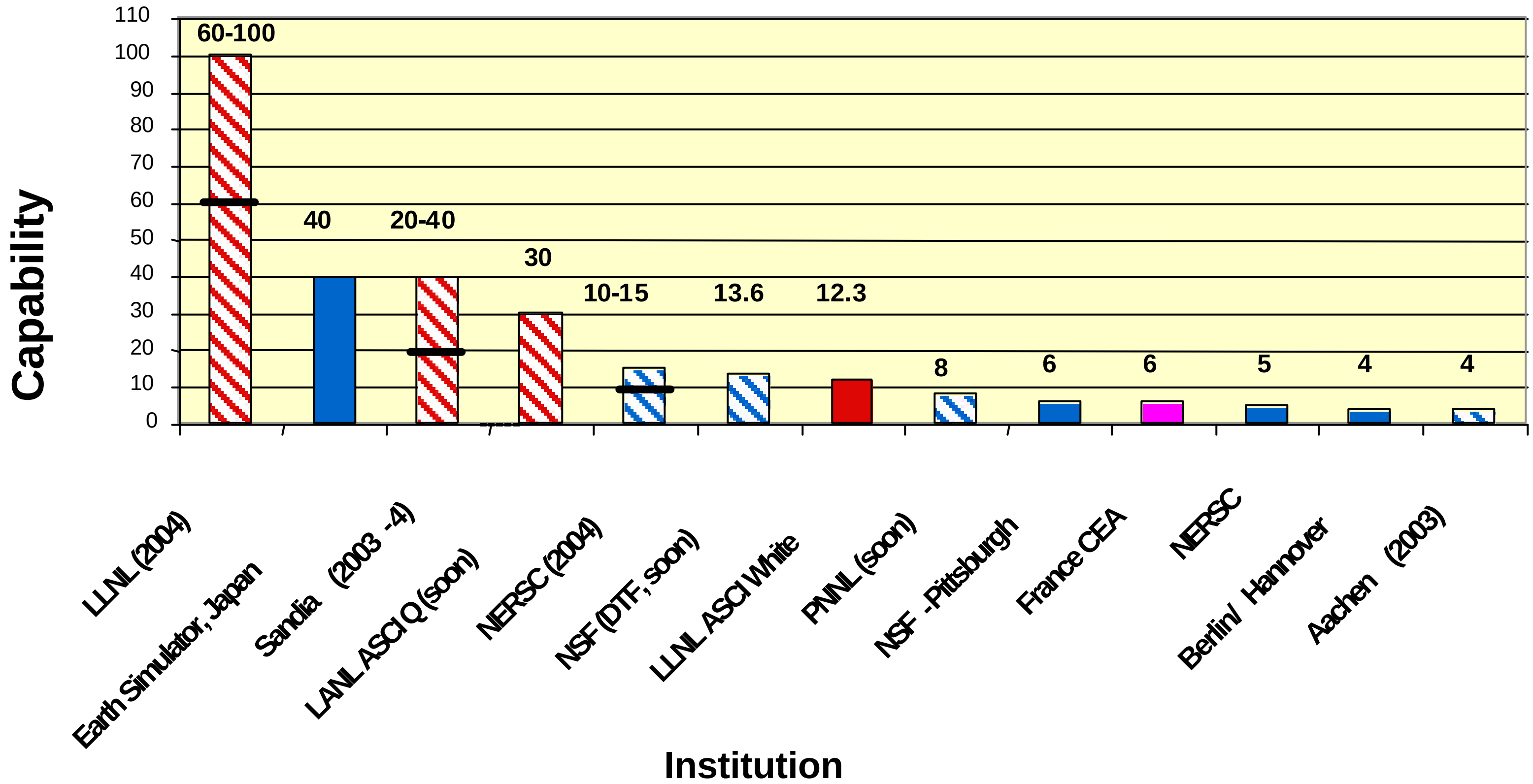
UCRL-PRES-147124-7

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
- Total hardware cost $540M
  — (Red $50M, Blue Mtn $80M, Blue Pacific $80M, White $110M, Q $220M)

The majority of terascale simulation environments continue to be based on clusters of SMPs

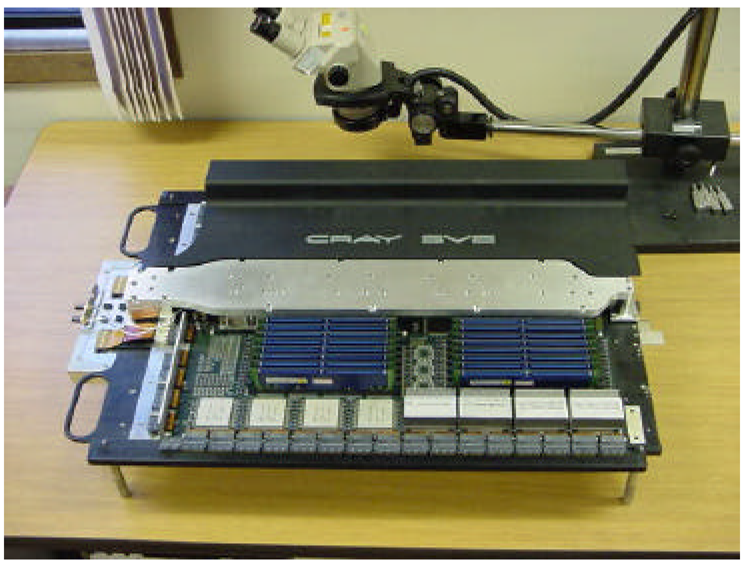
Peak Computer Capabilities



Source: Dona Crawford, LLNL

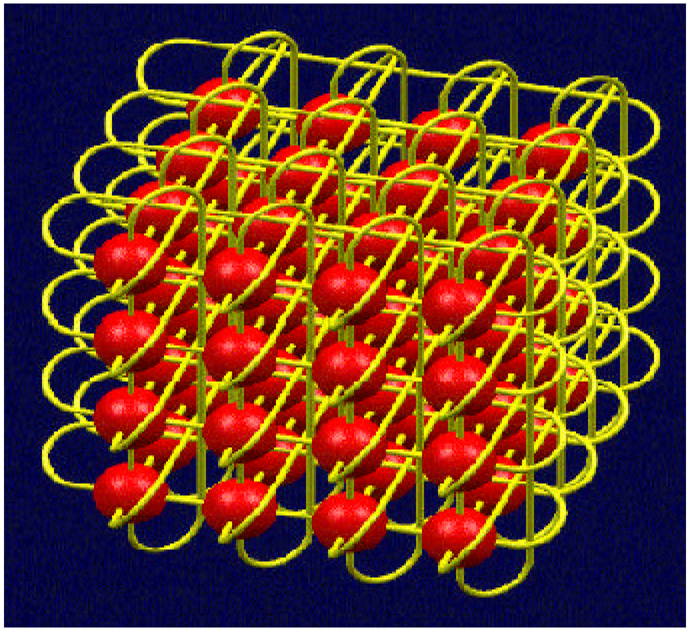
Cray SV2: *Parallel Vector Architecture*

- 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



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

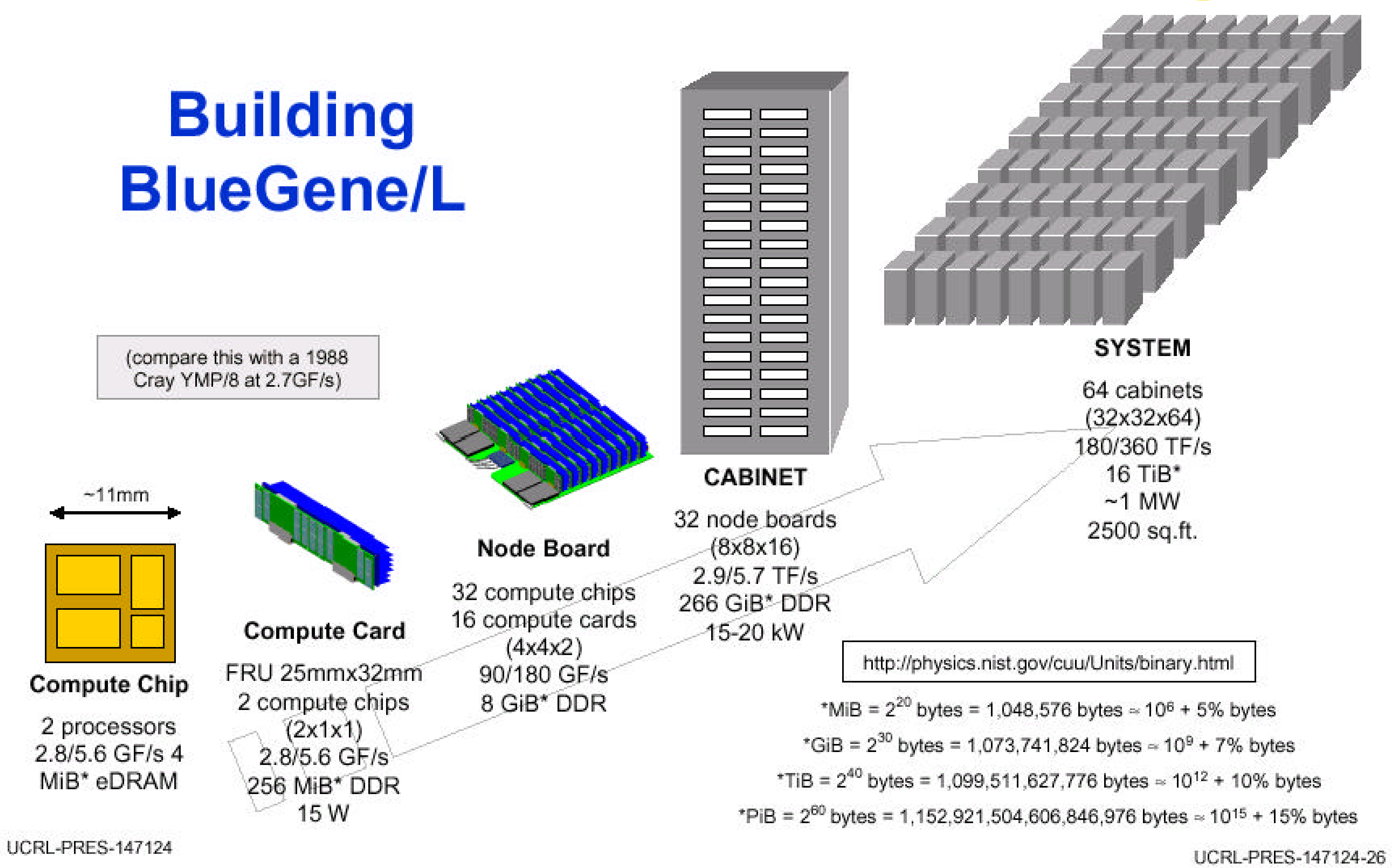


Image from LLNL

Choosing the Right Option

- Good hardware options are available
- 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 Iniative (ASCI) in DOE
  - Supercomputing Centers of the National Science Foundation (NCSA, NPACI, Pittsburgh)
  - Cluster computing in universities and labs

There is a software cost for each hardware option but,

The problem can be solved

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

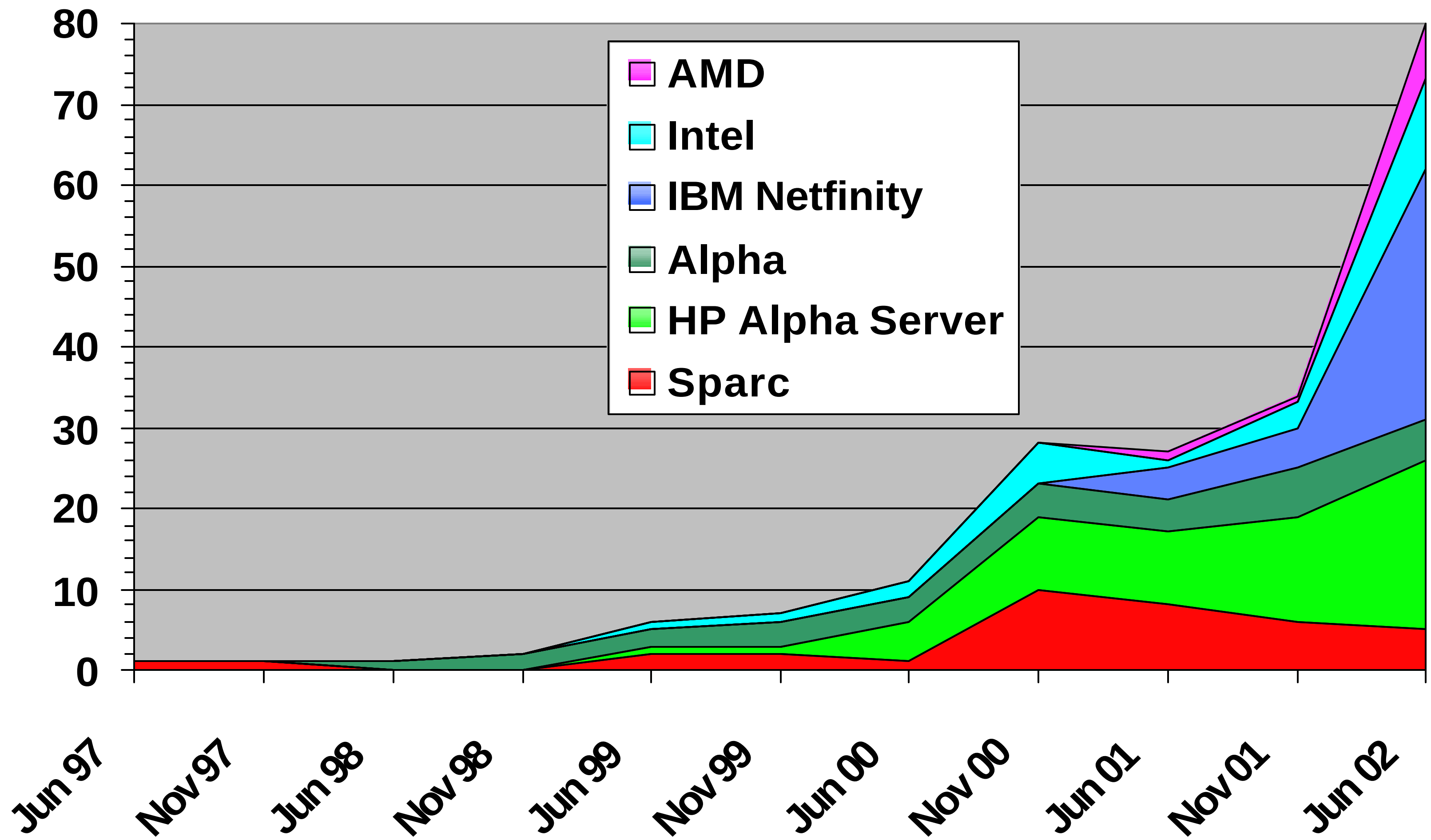
Processor Trends (summary)

- The Earth Simulator is a singular event
- It may become a turning point for supercomputing technology in the US
- Return to vectors is unlikely, but more vigorous investment in alternate technology is likely
- Independent of architecture choice we will stay on Moore’s Law curve

Five Computing Trends for the Next Five Years

- Continued rapid processor performance growth following Moore's law
- Open software model (Linux) will become standard
- Network bandwidth will grow at an even faster rate than Moore's Law
- Aggregation, centralization, colocation
- Commodity products everywhere

Number of NOW Clusters in TOP500



PC Clusters: Contributions of Beowulf

- **An experiment in parallel computing systems**
- **Established <u>vision of</u> 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

Linus's Law: Linux Everywhere

- Software is or should be free (Stallman)
- All source code is "open"
- Everyone is a tester
- Everything proceeds a lot faster when everyone works on one code (HPC: nothing gets done if resources are scattered)
- Anyone can support and market the code for any price
- Zero cost software attracts users!
- All the developers write lots of code
- Prevents community from losing HPC software (CM5, T3E)

Commercially Integrated Tflop/s Clusters Are Happening

- **Shell: largest engineering/scientific cluster**
- **NCSA: 1024 processor cluster (IA64)**
- **Univ. Heidelberg cluster**
- **PNNL: announced 8 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**

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

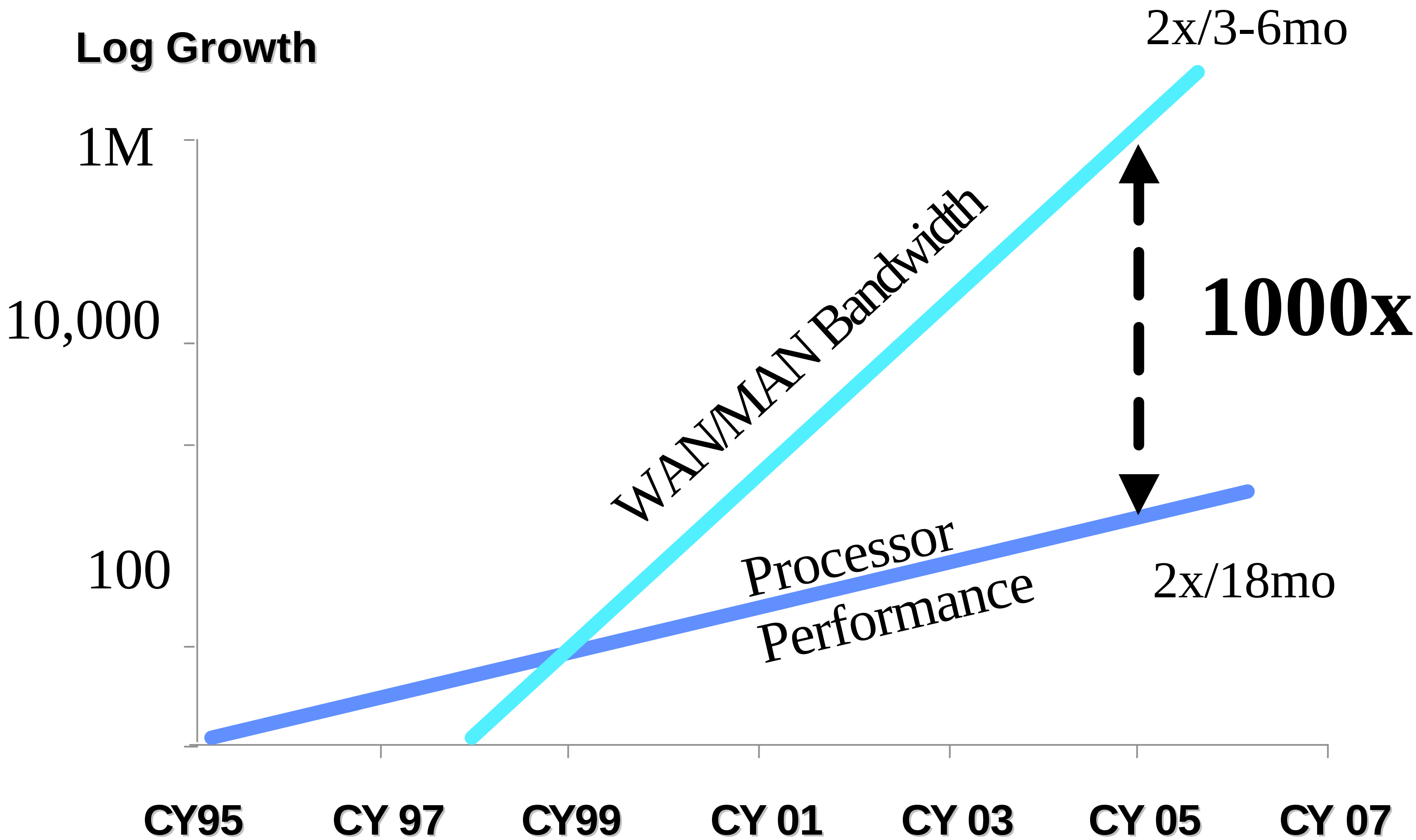
After Rick Stevens, Argonne

Five Computing Trends for the Next Five Years

- Continued rapid processor performance growth following Moore's law
- Open software model (Linux) will become standard
- Network bandwidth will grow at an even faster rate than Moore's Law
- Aggregation, centralization, colocation
- Commodity products everywhere

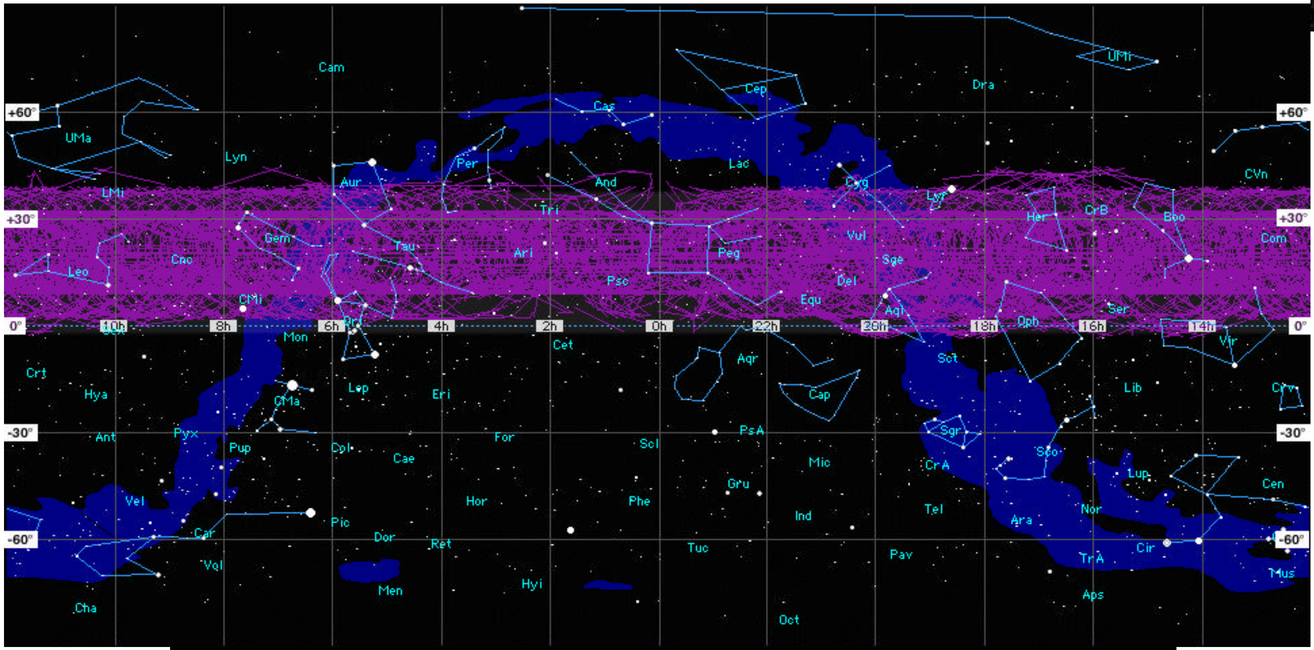
Bandwidth vs. Moore's Law

Adapted from G. Papadopoulos, Sun



Internet Computing- SETI@home

- Running on 500,000 PCs, ~1000 CPU Years per Day
  — 485,821 CPU Years so far
- Sophisticated Data & Signal Processing Analysis
- Distributes Datasets from Arecibo Radio Telescope

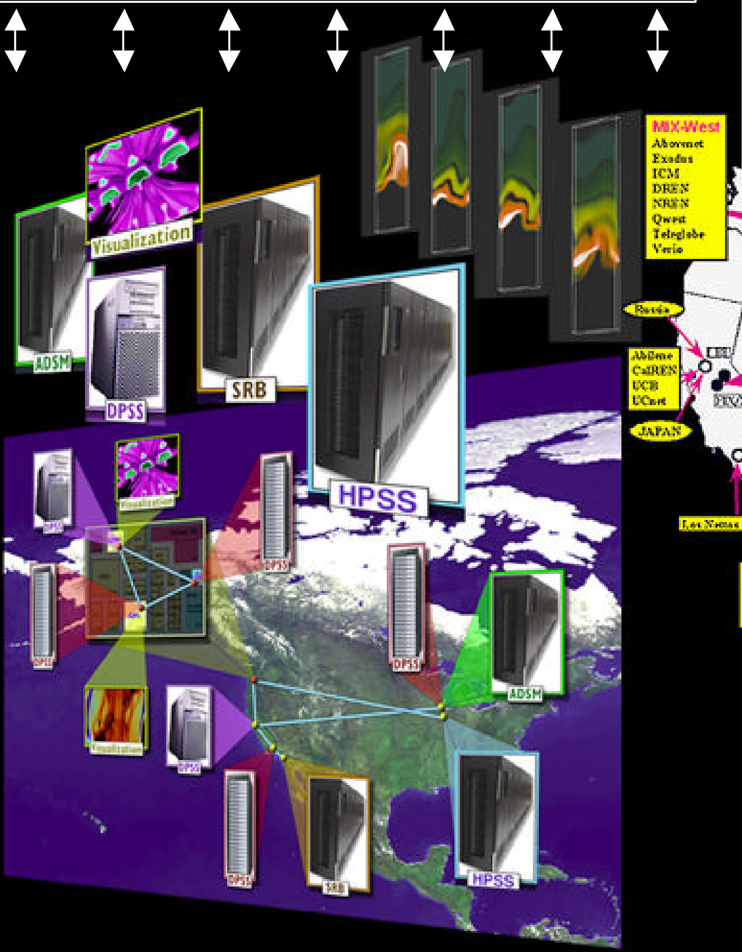


Next Step-
Allen Telescope Array

The Vision for a DOE Science Grid

Scientific applications use workflow frameworks to coordinate resources and solve complex, multi-disciplinary problems

Grid services provide a uniform view of many diverse resources

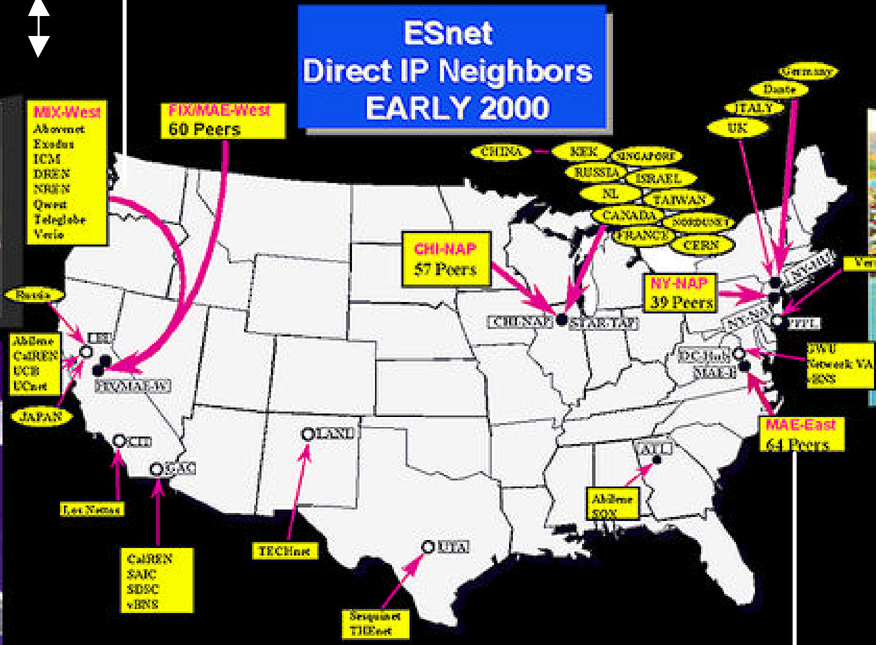


Large-scale science and engineering is typically done through the interaction of

- People,
- Heterogeneous computing resources,
- Multiple information systems, and
- Instruments

All of which are geographically and organizationally dispersed.

The overall motivation for “Grids” is to enable the routine interactions of these resources to facilitate this type of large-scale science and engineering.

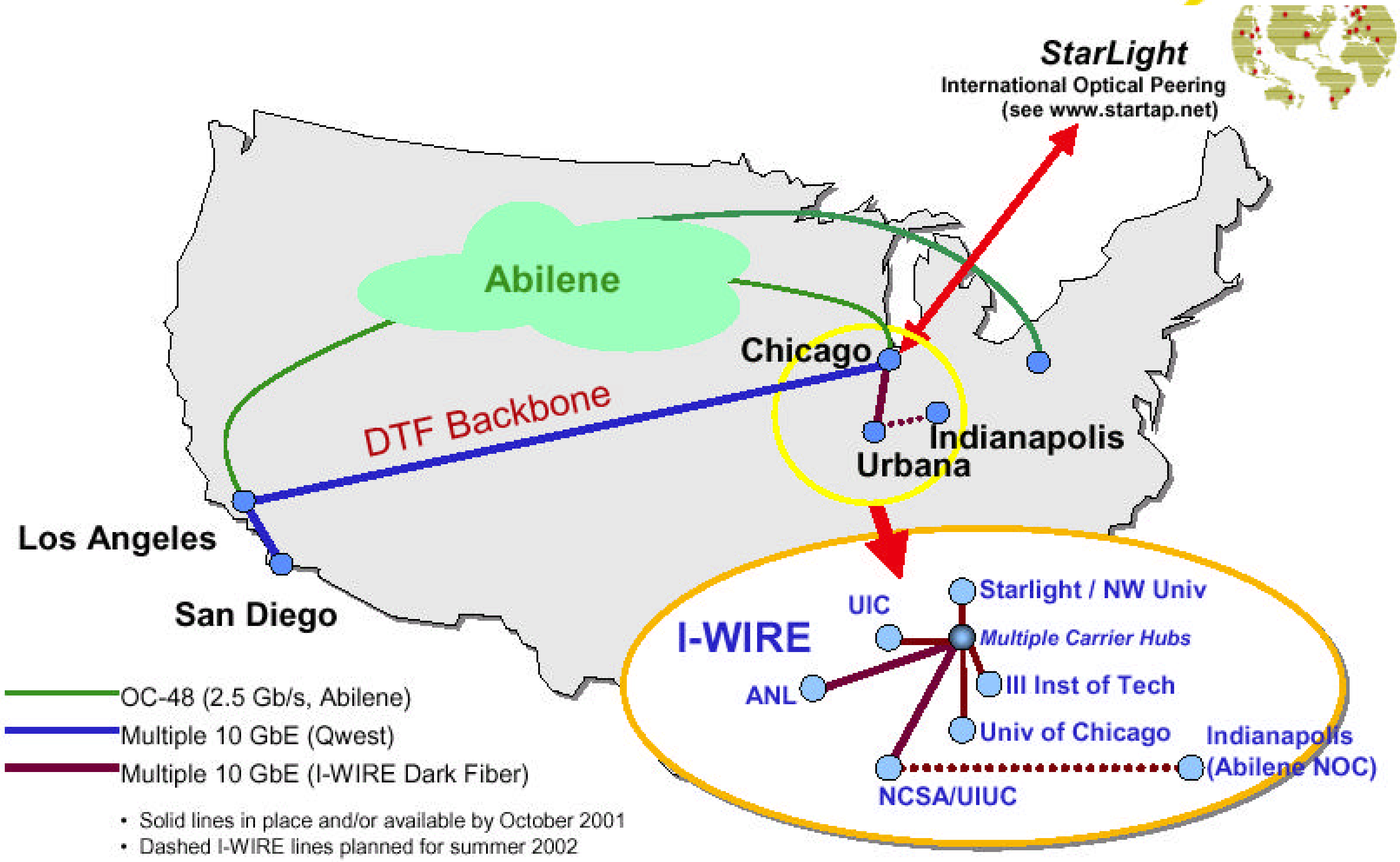


Two Sets of Goals

Our overall goal is to facilitate the establishment of a DOE Science Grid (“DSG”) that ultimately incorporates production resources and involves most, if not all, of the DOE Labs and their partners.

A “local” goal is to use the Grid framework to motivate the R&D agenda of the LBNL Computing Sciences, Distributed Systems Department (“DSD”).

TeraGrid [40 Gbit/s] DWDM Wide Area Network



We Must Correct a Current Trend in Computer Science Research

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

Impact on HPC

- Internet Computing will stay on the fringe of HPC
  - — no viable model to make it commercially realizable

- Grid activities will provide an integration of data, computing, and experimental resources
  - — but not metacomputing

- More bandwidth will lead to aggregation of HPC resources, not to distribution

Five Computing Trends for the Next Five Years

- Continued rapid processor performance growth following Moore's law
- Open software model (Linux) will become standard
- Network bandwidth will grow at an even faster rate than Moore's Law
- Aggregation, centralization, co-location
- Commodity products everywhere

NERSC's Strategy Until 2010: Oakland Scientific Facility

New Machine Room — 20,000 ft^{2}, Option open to expand to 40,000 ft^{2}.

Includes ~50 offices and 6 megawatt electrical supply.

It's a deal: \$1.40/ft^{2} when Oakland rents are >\$2.50/ ft^{2} and rising!

The Oakland Facility Machine Room

Power and cooling are major costs of ownership of modern supercomputers

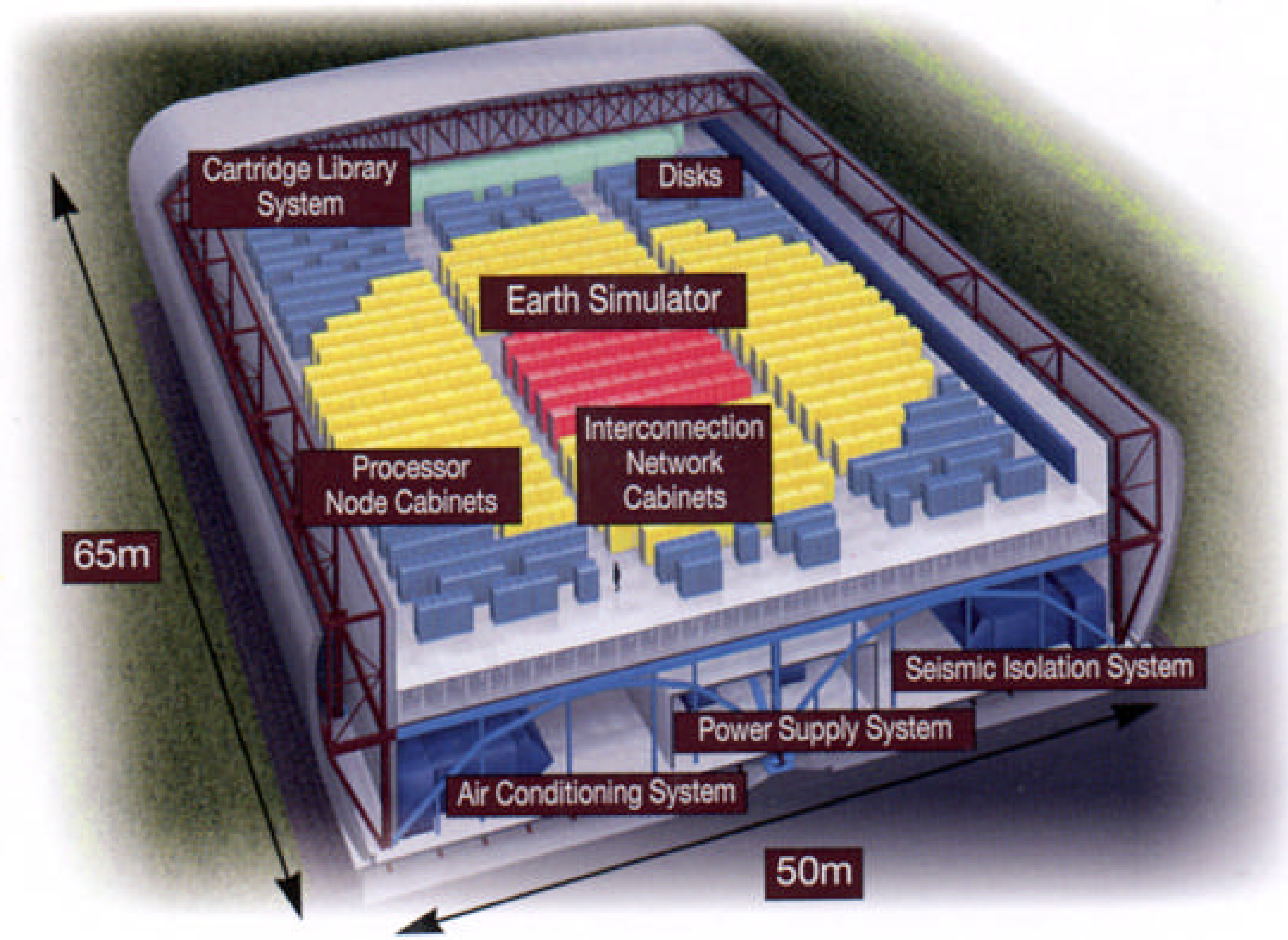
Expandable to 6 Megawatts

Metropolis Center at LANL – home of the 30 Tflop/s Q machine

Strategic Computing Complex at LANL

- 303,000 gross sq. ft.
- 43,500 sq. ft. unobstructed computer room
  - — Q consumes approximately half of this space
- 1 Powerwall Theater (6X4 stereo = 24 screens)
- 4 Collaboration rooms (3X2 stereo = 6 screens)
  - — 2 secure, 2 open (1 of each initially)
- 2 Immersive Rooms
- Design Simulation Laboratories (200 classified, 100 unclassified)
- 200 seat auditorium

Earth Simulator Building



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

This

Became

And will get bigger

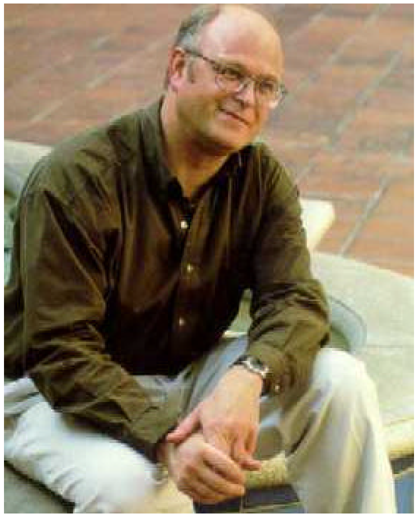
Power and cooling are also increasingly problematic, but there are limiting forces in those areas.

— 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.

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

Thomas Sterling, Salishan, 2001



Five Computing Trends for the Next Five Years

- Continued rapid processor performance growth following Moore's law
- Open software model (Linux) will become standard
- Network bandwidth will grow at an even faster rate than Moore's Law
- Aggregation, centralization, co-location
- Commodity products everywhere

…. the first ever coffee machine to send e-mails

"Lavazza and eDevice present the first ever coffee machine to send e-mails

On-board Internet connectivity leaves the laboratories

eDevice, a Franco-American start-up that specializes in the development of on-board Internet technology, presents a world premiere: e-espressopoint, the first coffee machine connected directly to the Internet. The project is the result of close collaboration with Lavazza, a world leader in the espresso market with over 40 million cups drunk each day.

Lavazza's e-espressopoint is a coffee machine capable of sending e-mails in order, for example, to trigger maintenance checks or restocking visits. It can also receive e-mails from any PC in the given service.

A partnership bringing together new technologies and a traditional profession …"

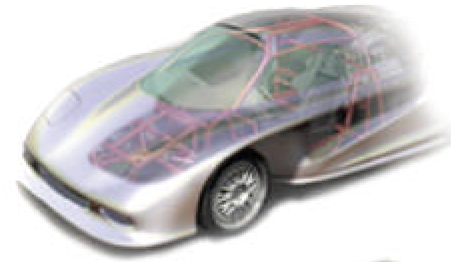
See http://www.cyperus.fr/2000/11/edevice/cpuk.htm

New Economic Driver: IP on Everything

Source: Gordon Bell, Microsoft, Lecture at Salishan Conf.

Information Appliances

- Are characterized by what they do
- Hide their own complexity
- Conform to a mental model of usage
- Are consistent and predictable
- Can be tailored
- Need not be portable



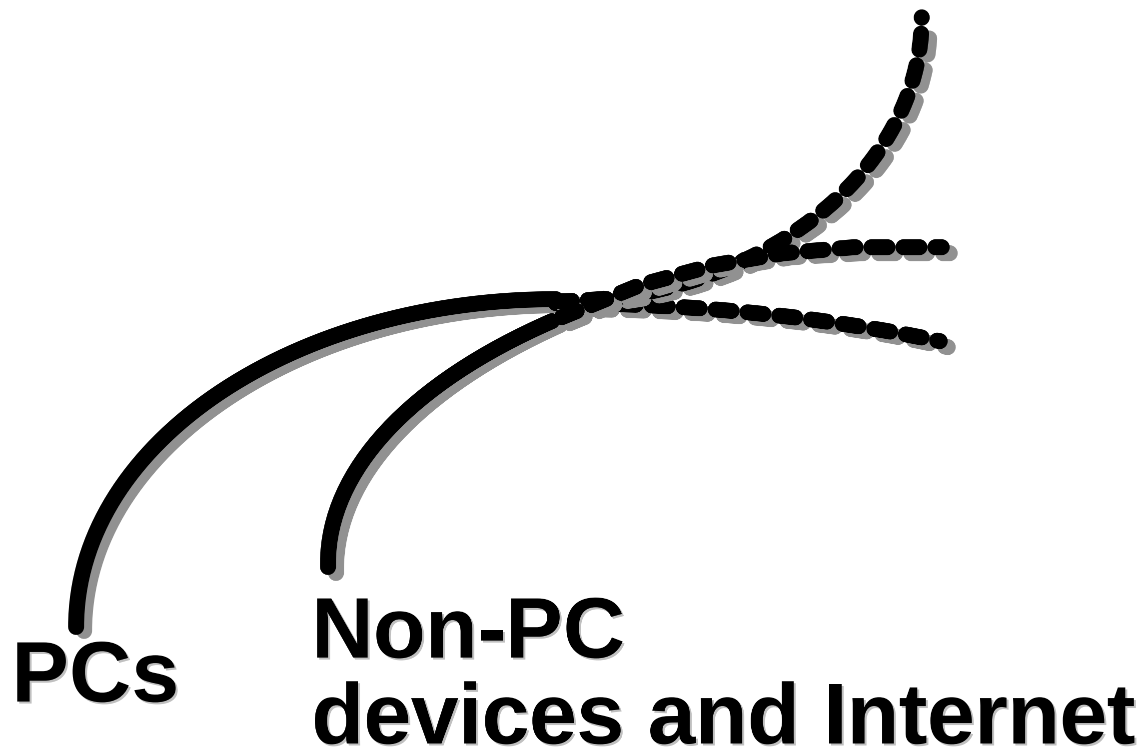



Source: Joel Birnbaum, HP, Lecture at APS Centennial, Atlanta, 1999

… but what does that have to do with supercomputing?

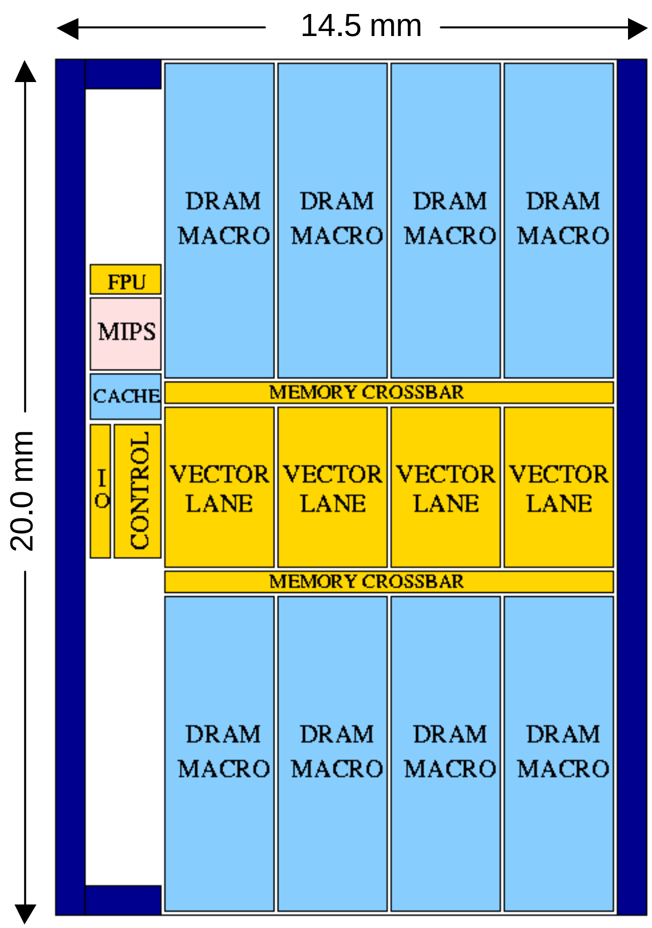
HPC depends on the economic driver from below:

- **Mass produced cheap processors will bring microprocessor companies increased revenue**
- **system on a chip will happen soon**



"PCs at Inflection Point",
Gordon Bell, 2000

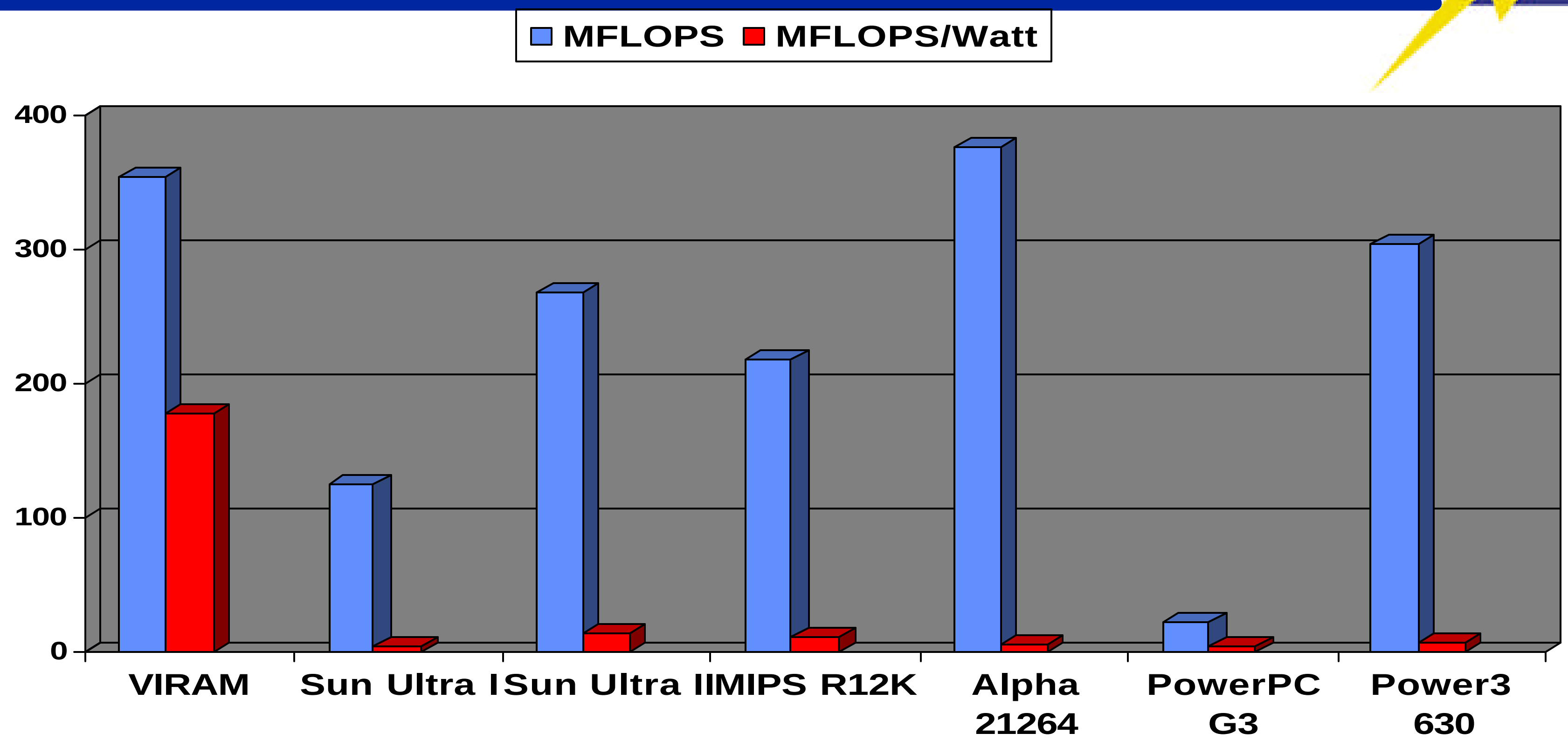
VIRAM Overview (UCB)



- 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

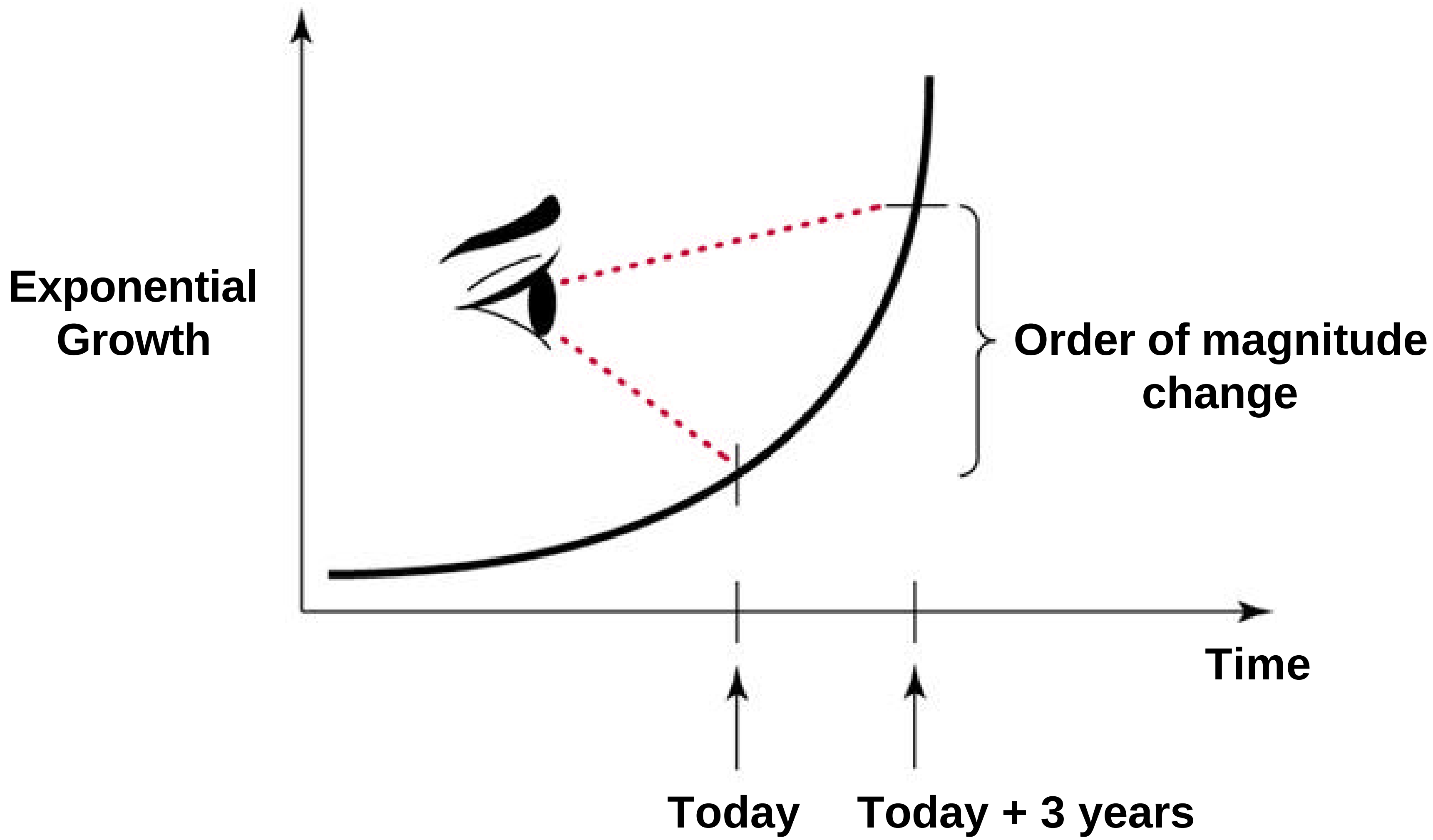
Power Advantage of PIM+Vectors



- 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

Moore's Law — The Exponential View



Moore's Wall —
The Real (Exponential) View

What am I willing to predict?

In 2007:

- **Clusters of SMPs will hit (physical) scalability issues**
- **PC clusters will not scale to the very high end, because**
  - **Immature systems software**
  - **Lack of communications performance**
- **We will need to look for a replacement technology**
  - **Blue Gene/L ; Red Storm, SV-2 …**

In 2010: **"Per Aspera Ad Astra"**

- **Petaflop (peak) supercomputer before 2010**
- **We will use MPI on it**
- **It will be built from commodity parts**
- **I can't make a prediction from which technology (systems on a chip is more likely than commodity PC cluster or clusters of SMPs)**
- **The "grid" will have happened, because a killer app made it commercially viable**

Disruptive Technology – non linear effects

- **In spite of talk about the "information superhighway" in 1992 it was impossible to predict the WWW**
- **Technologic and economic impact of disruptive technology not predictable**
- **Candidate technology:**

  robotics ?

Berkeley RAGE robot just won R&D 100 award