The Future of Scientific Computing

Horst D. Simon

NERSC, Division Director

June 6, 2000

NERSC Overview

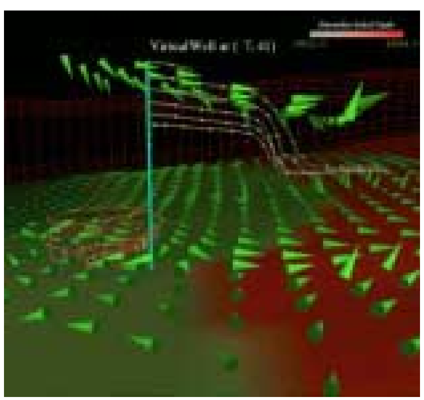
- Located in the hills next to University of California, Berkeley campus
- close collaborations between university and NERSC in computer science and computational science

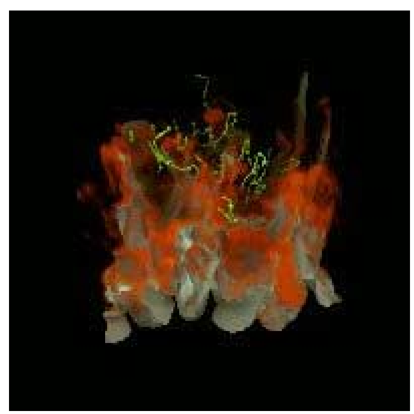
NERSC - Overview

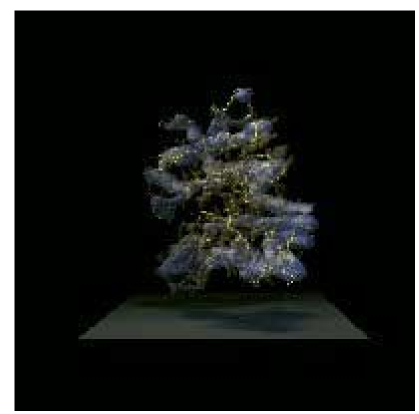
• the Department of Energy, Office of Science, supercomputer facility

• unclassified, open facility; serving >2000 users in all DOE mission relevant basic science disciplines

• 25th anniversary in 1999

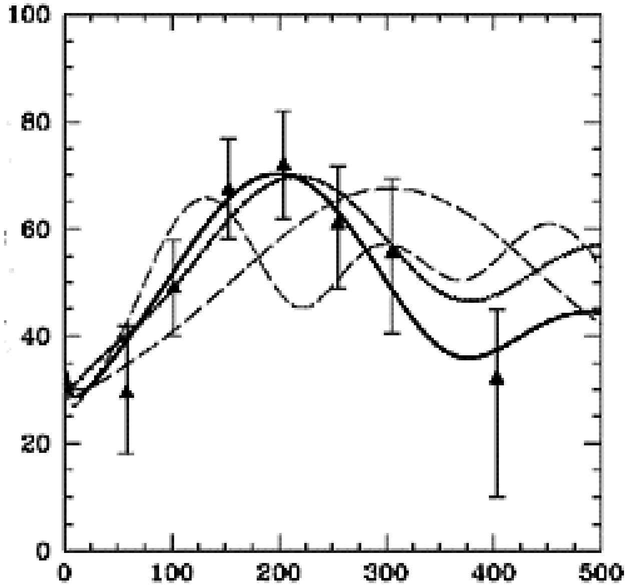






Computational Science at NERSC

- Analysis of the North American test flight data of BOOMERANG done at NERSC
- MADCAP analysis package developed by J. Borrill at NERSC
- First indication that universe is flat, more detailed analysis of Antarctic Flight data from 1998/99 to follow in 2000
- Analysis of future flight data (MAXIMA, PLANCK) are Petaflops problems and beyond
- added completely new set of users to NERSC community



“It’s hard to make predictions, especially about the future.”

Yogi Berra

"Technology does not drive change at all. Technology merely enables change. It's our collective cultural response to the options and opportunities presented by technology that drives change."

Paul Saffo
Institute for the Future
Menlo Park, California

Overview

- Retrospective: changes in the 1990s
- Extrapolation to the near future up to 2010
- The end of Moore’s Law in about 2020
- Beyond 2025

Things That Did Not Happen in the 1990s

1992 predictions (after Forest Baskett, SGI):

- TV and PC converge
- Interactive TV
- Video servers instead of video stores
- Apple/IBM/Motorola
- Intel makes a mistake
- MPPs go mainstream

1990s: Technology

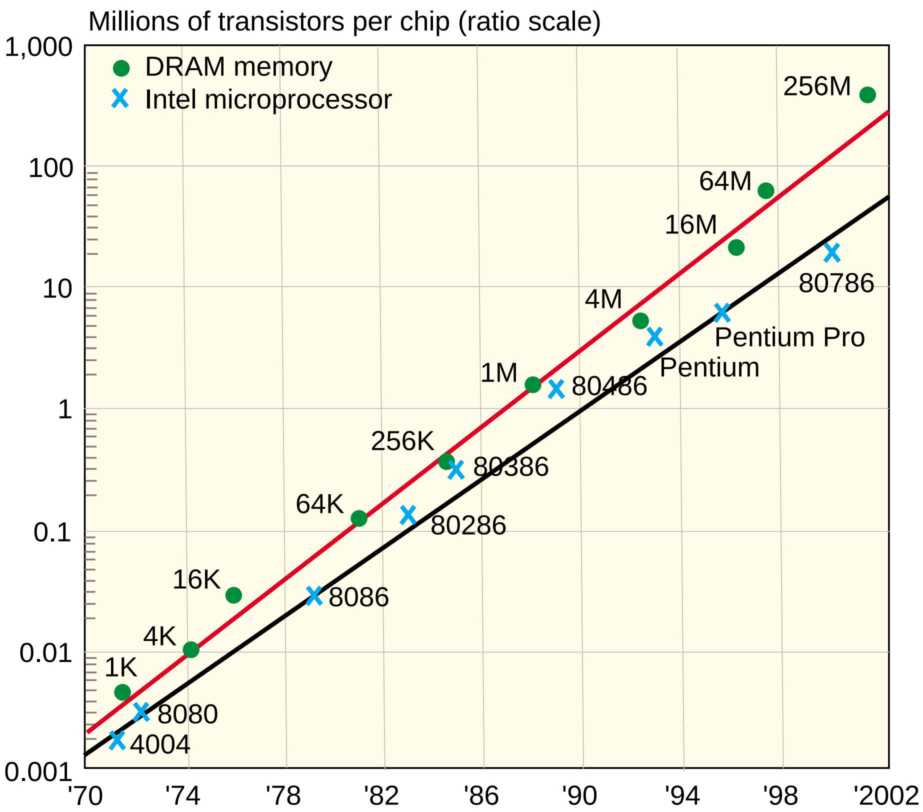
In the 1980's there were fundamental changes in microprocessor development ("killer micros")

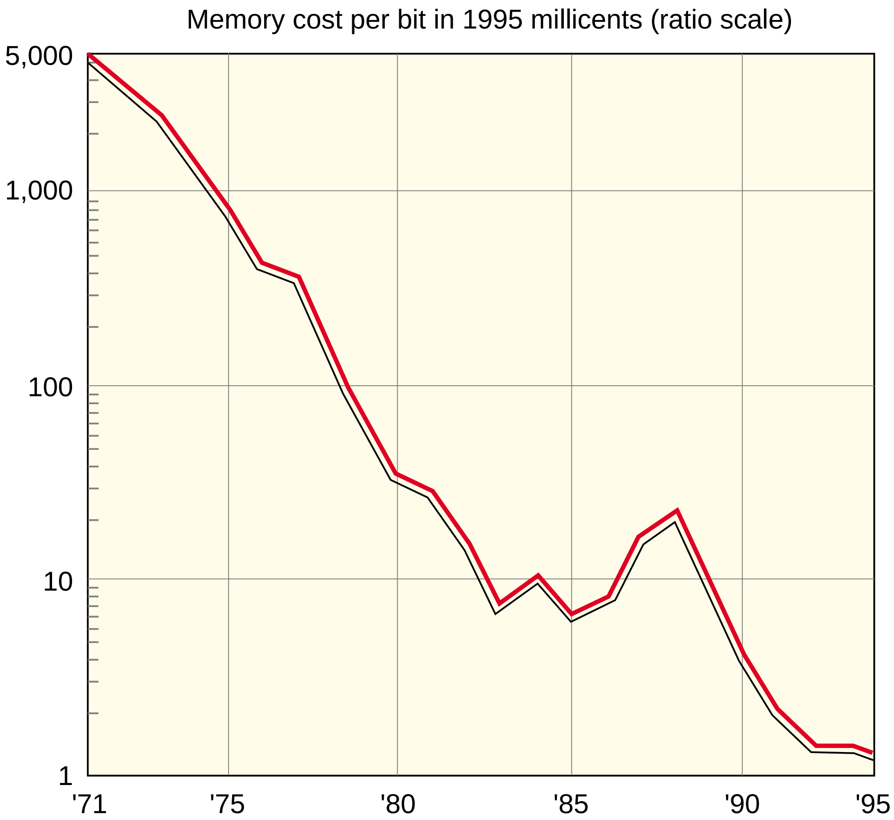
- — dramatic increase in number of transistors available per chip
- — architectural advances including the use of RISC ideas, pipelining and caches
- — as a result CPU performance has improved by a factor of 1.5 to 2.0 per year

Maturation in the late 80s

Full impact in the early 90s

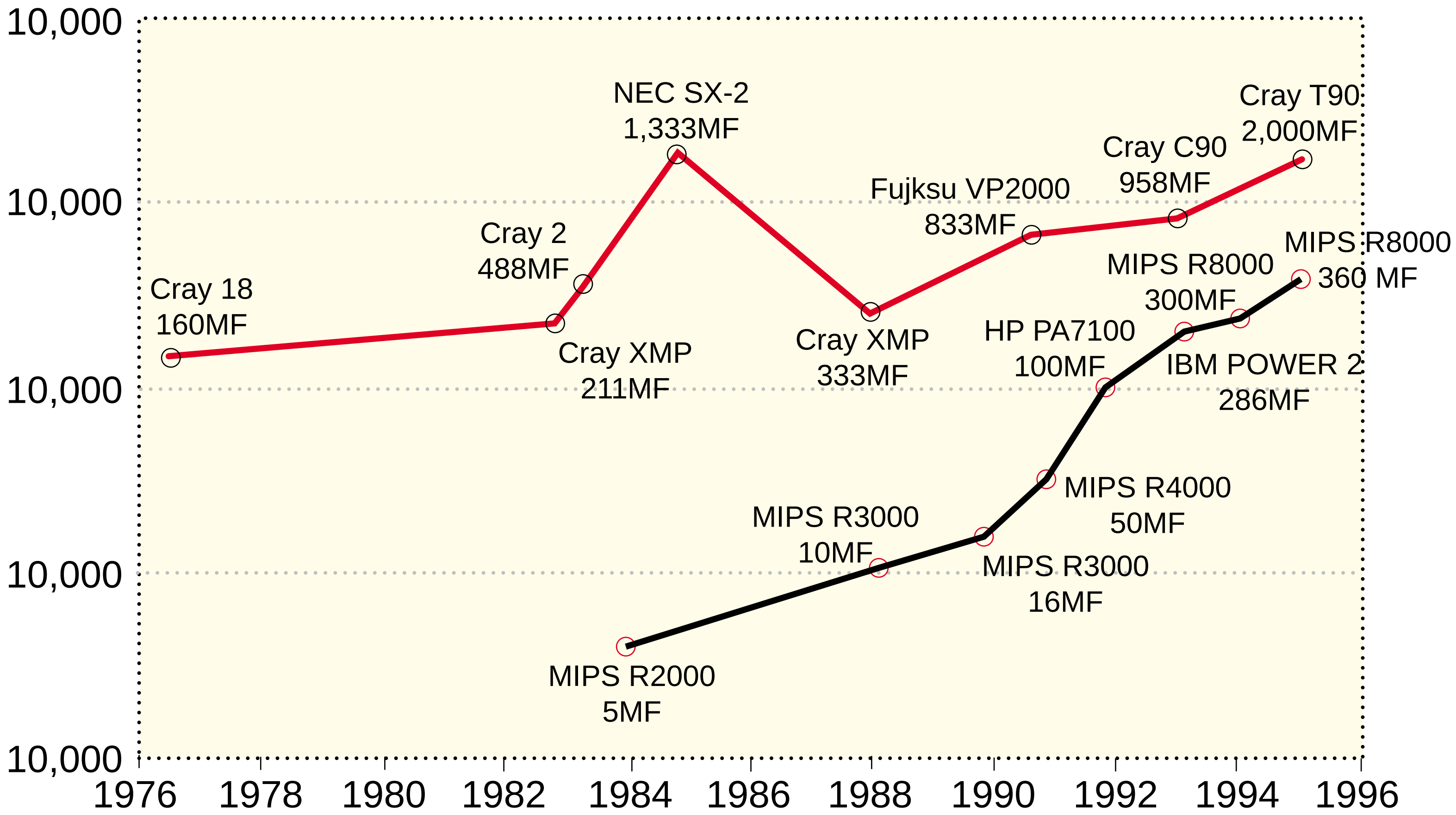
Moore’s First Law





Source: VLSI Research Inc.

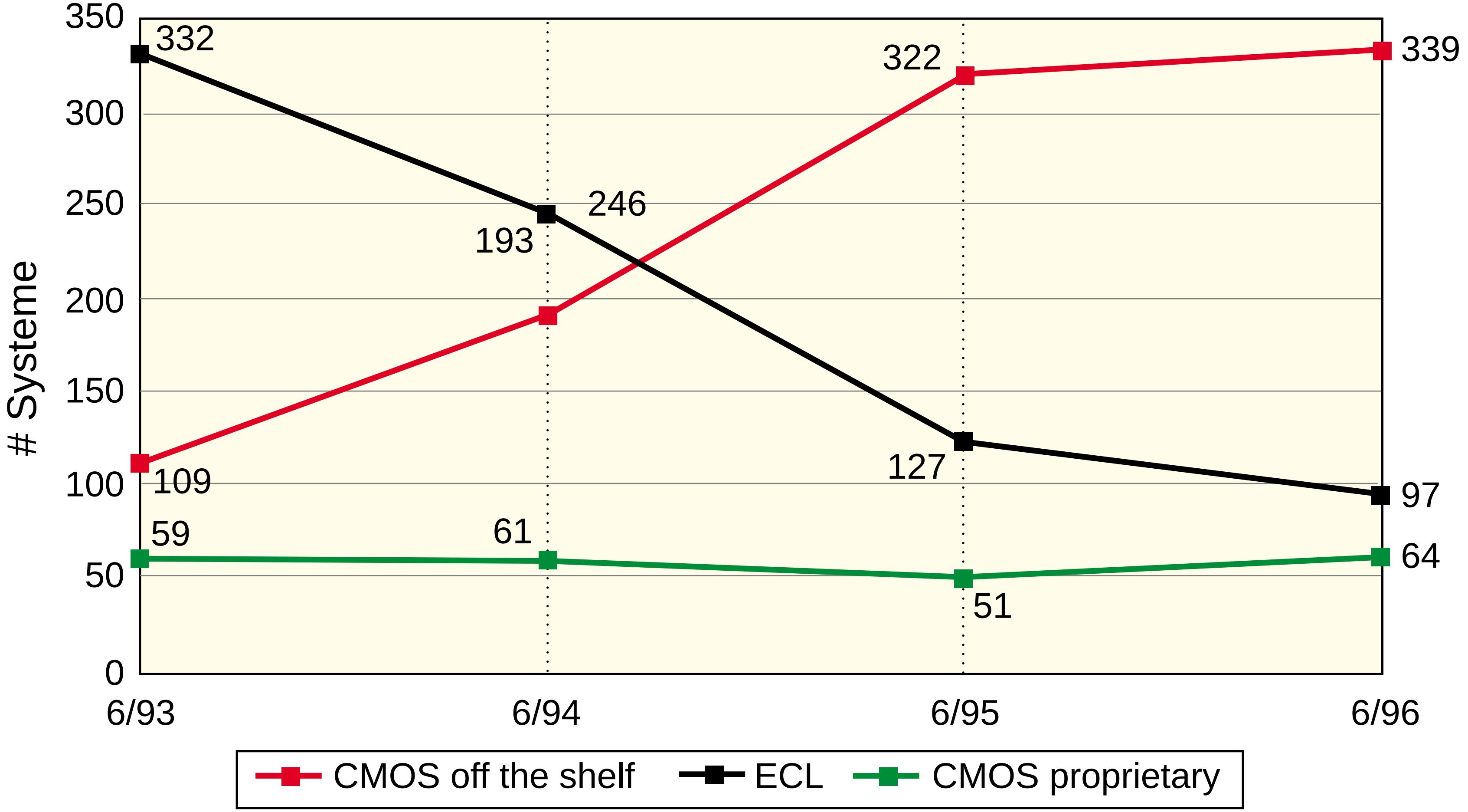
Microprocessors vs. Vector Supercomputers (ca. 1994)



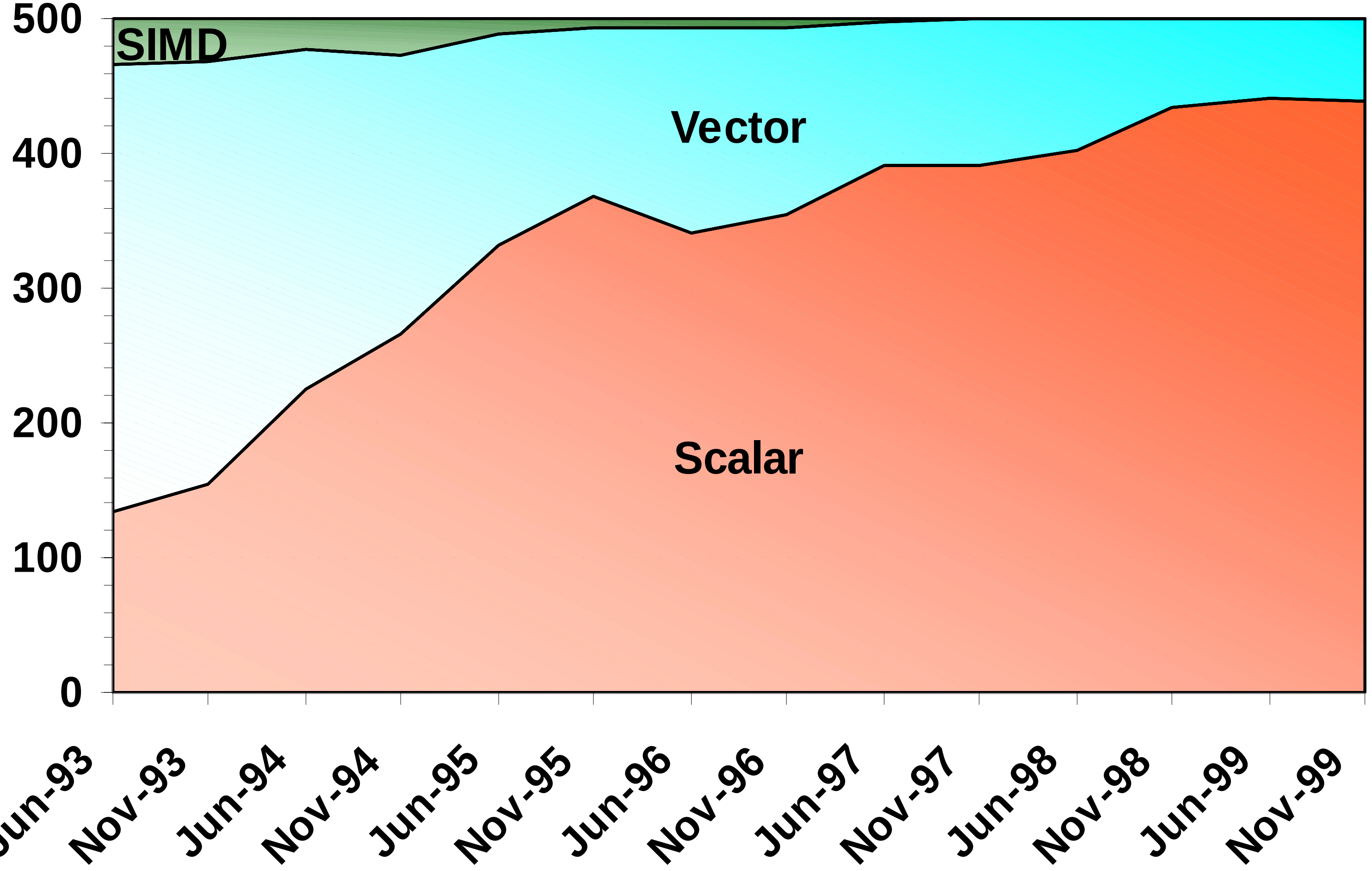
TOP500 List

- Published twice a year with the 500 most powerful supercomputers in actual use
- Ranked according to LINPACK R_max
- Powerful tool to evaluate trends in HPC
- Data available since 1993
- For details see http://www.top500.org/

Top 500—CPU Technology

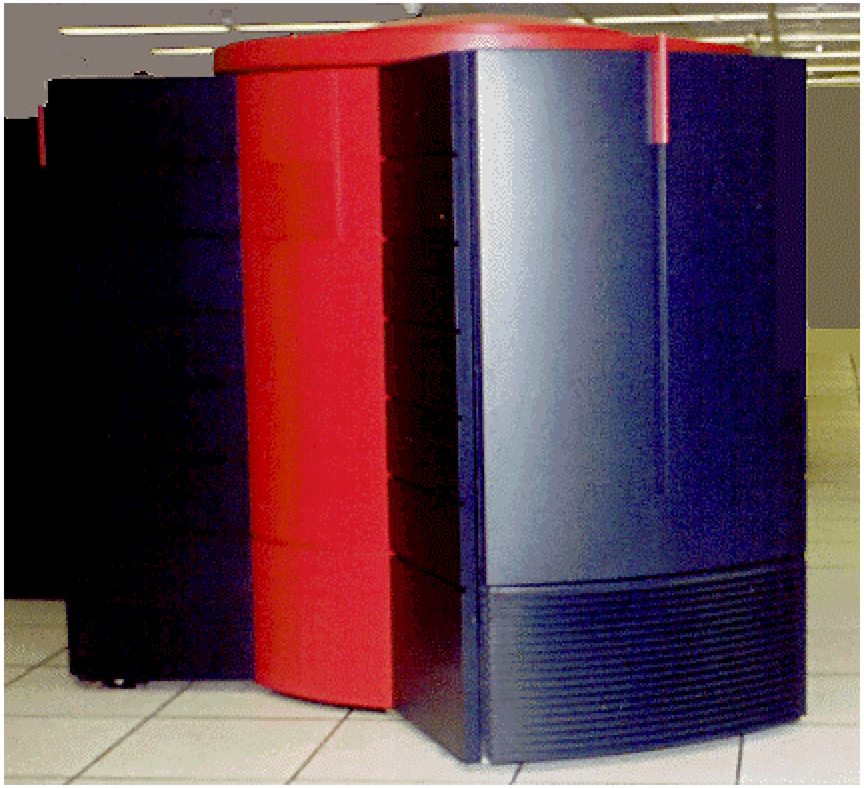


Processor Design as Seen in the TOP500



NERSC-1 Cray C90 Installed in Dec. 1991

- Cray C90 installed in December 1991
- Ended contract with CCC for a Cray-3
- Stable high-end production platform for seven years until 12/31/98



NERSC-2 Cray T3E-900 Installed in 1996

The 644 processor T3E-900 is one of the most powerful unclassified supercomputers in the U.S.

- **Eight out of twelve DOE Grand Challenge Projects compute at NERSC**
- **50% of the resource dedicated to GC projects**
- **About 100 other projects allocated on the NERSC T3E-900**
- **1997 GAO report judged NERSC to have the best MPP utilization (75%) —1999 utilization >90%**

NERSC-3 IBM SP3 Installed in 6/99

- New contract with IBM announced in April 1999
- IBM was clearly the best value for the primary award
  - —provides the best absolute performance
  - —has lowest absolute cost
  - —provides the best price performance
  - —provides acceptable functionality
  - —guarantees performance - low risk

NERSC-3 IBM SP

- **IBM selected to provide NERSC-3 (IBM SP3/RS 6000)**
- **Phase I: June 1999 installation**
  - 608 processors
  - 410 gigaflop peak performance
  - Provides one teraflop NERSC capability

- **Phase II: December 2000 completion**
  - 2,432 processors
  - 3.2 teraflop peak performance
  - 4 teraflop total NERSC capability

HPC Systems at NERSC in the 90s

	NERSC-1 Cray C90	NERSC-2 Cray T3E	NERSC-3 IBM SP-3
Year of Installation	1991	1996	1999
Number of Processors	16	640	2048
Processor Technology	Custom ECL	Commodity CMOS	Commodity CMOS
Peak System Perform.	16 Gflop/s	580 Gflop/s	3000 Gflop/s
Architecture	Shared memory parallel vector	Distributed memory	128 nodes with 16 processor SMP
System	Fully integrated custom system	Fully integrated custom system with commodity CPU and memory	Loosely integrated system with commodity system components
System Software	Vendor supplied, ready on delivery	Vendor supplied, completed after nearly 3 years development	Vendor supplied, contractual complete in about 3 years
Floor space	588 ft	360 ft	4000 ft
Power consumption	500 kW	288 kW	1400 kW

Impact of Technology Transitions

	1994 – 1996 transition	1998 – 2000 transition
Economic Driver	Price performance of commodity processors and memory	16 – 64 CPU "sweet spot" for SMP technology in the commercial market place
Advantages of transition	Higher performance and better price performance	Higher performance
Challenges of transition	1) Applications transition to distributed memory, message passing model (MPI) 2) More complex system software (scheduling, checkpoint restarting)	1) Applications transition to hierarchical, distributed memory model (threads + MPI) 2) New development efforts for even more complex systems software 3) Increased cost of facilities

Table 2. Impact of the two technology transitions of the 1990s.

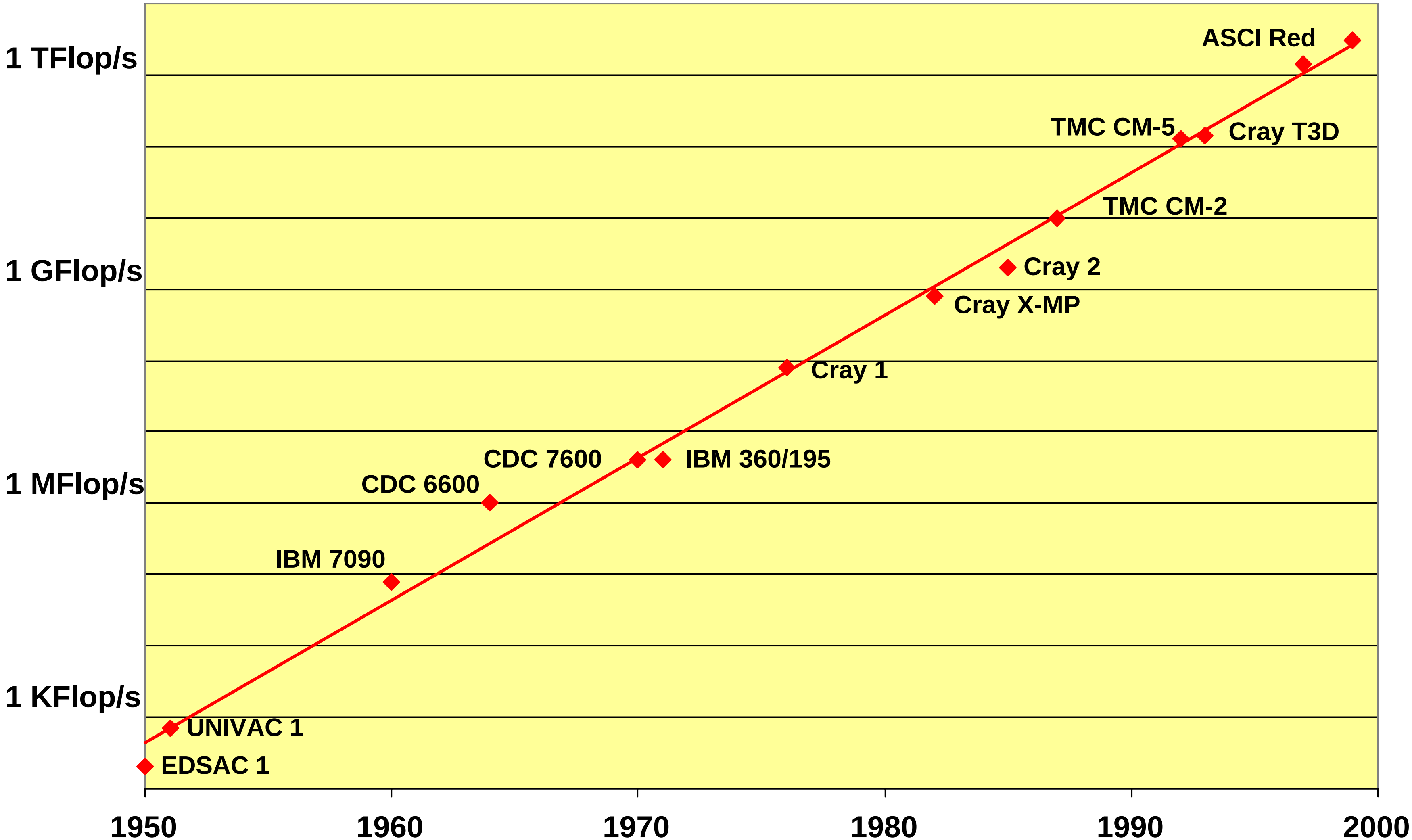
Three Challenges

- **Applications that can tolerate an increase in communication latency and parallelism as well as a distributed, hierarchical memory model need to be written**

- **System software for increasingly complex, more difficult to manage, one-of-a-kind systems will have to be developed anew**

- **Center management will be forced to take creative new approaches to solve the space and power requirements for the new systems**

Impact of Moore's Law on HPC



The Revolution of 1990s — Major HPC Market Realignment

1991 Newcomers with CMOS and MPP technology (Intel, TMC, KSR) gain mind share and market share

1993 Cray, IBM, Convex go CMOS (T3D, SP 1/2, SPP 1000)

1994 TMC, KSR go out of business; SGI's SMP success

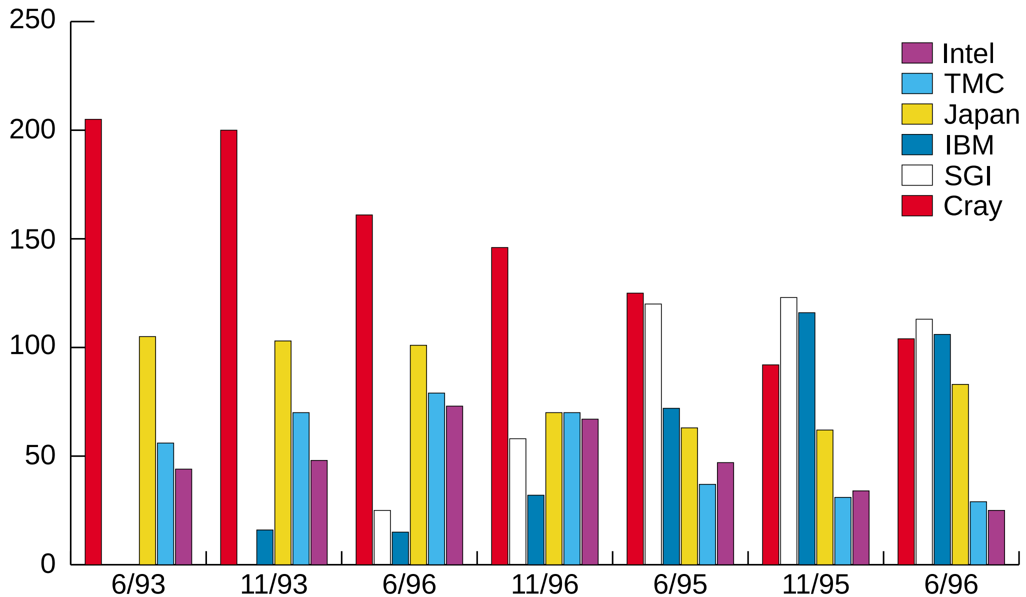
1995 HP buys Convex; Fujitsu, NEC introduce CMOS vector machines

1996 SGI buys Cray

1997 TOP500 takeover by CMOS complete

2000 Tera buys Cray Division from SGI and renames itself Cray Inc.

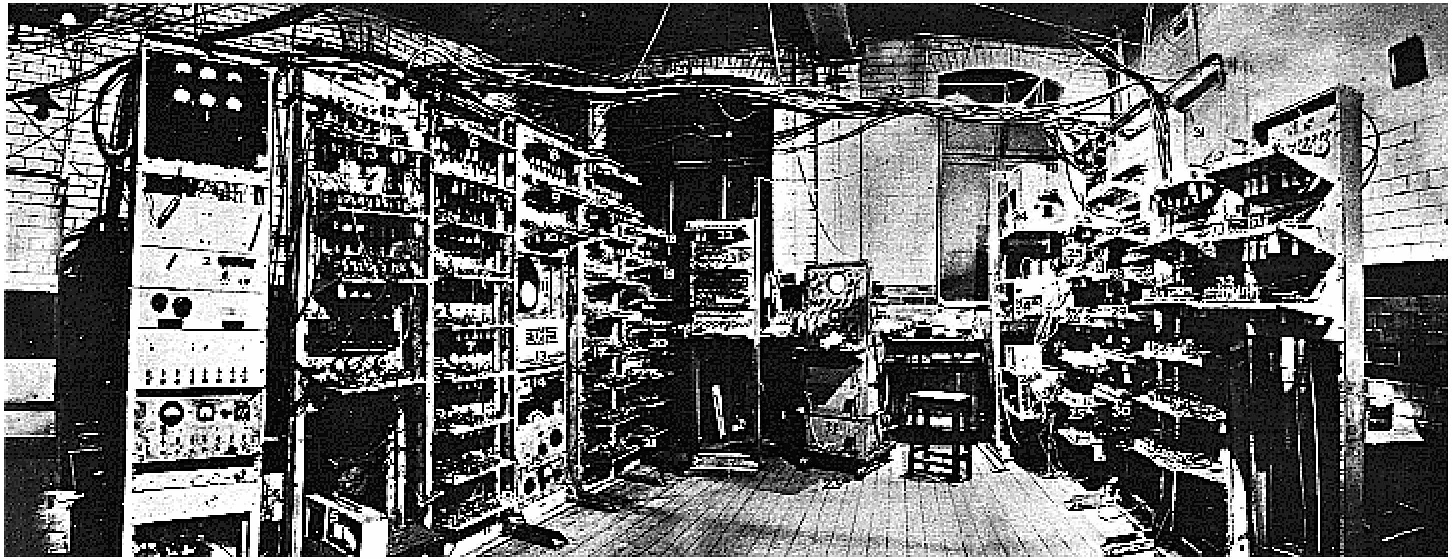
Top 500—Vendors



Data	Cray	SGI	IBM	Japan	TMC	Intel
6/93	205	0	0	105	56	44
11/93	200	0	16	103	70	48
6/96	161	25	15	101	79	73
11/96	146	58	32	70	70	67
6/95	125	120	72	63	37	47
11/95	92	123	116	62	31	34
6/96	104	113	106	83	29	25

The Dead Supercomputer Society

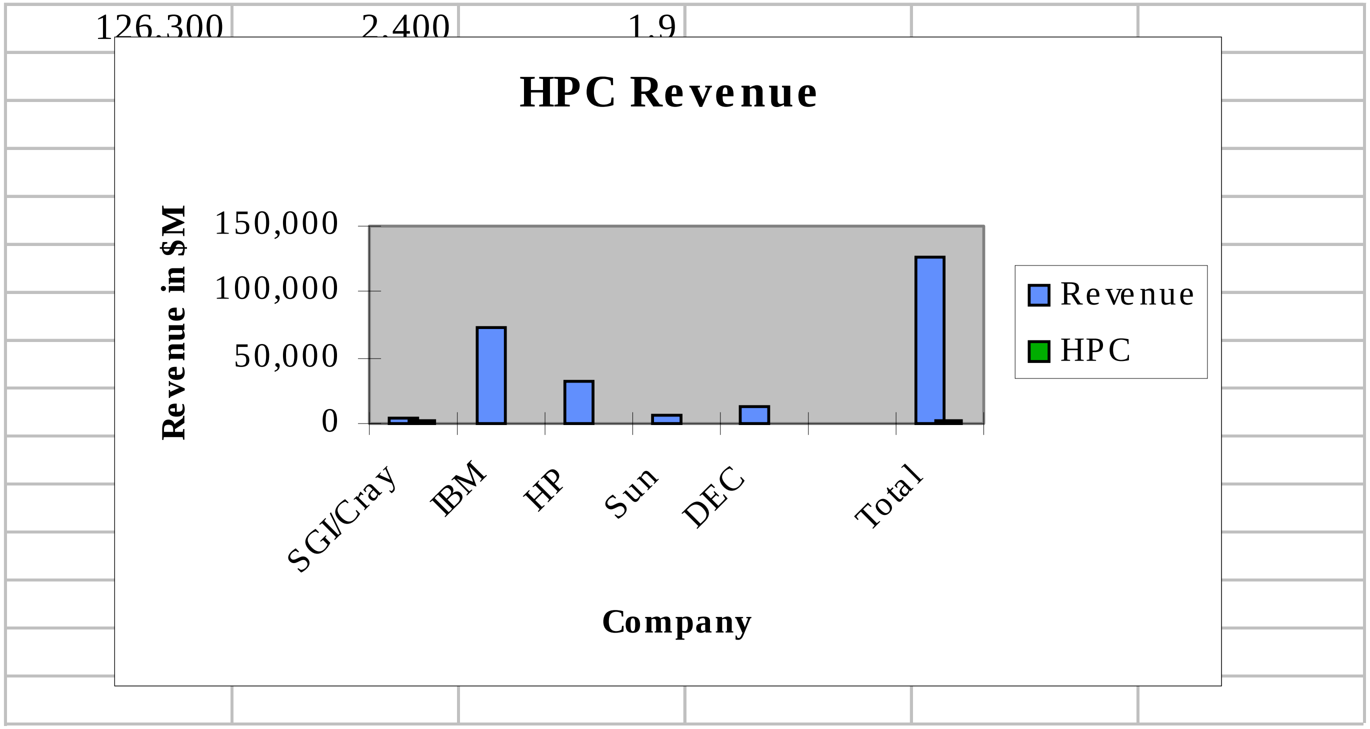
- See http://www.paralogos.com/DeadSuper/
- List of 42 dead companies or projects from 1975-today



Since 1997: The New HPC Marketplace

All major US HPC companies are now vertically integrated (SGI, IBM, HP, Sun, Compaq), with exception of Cray.

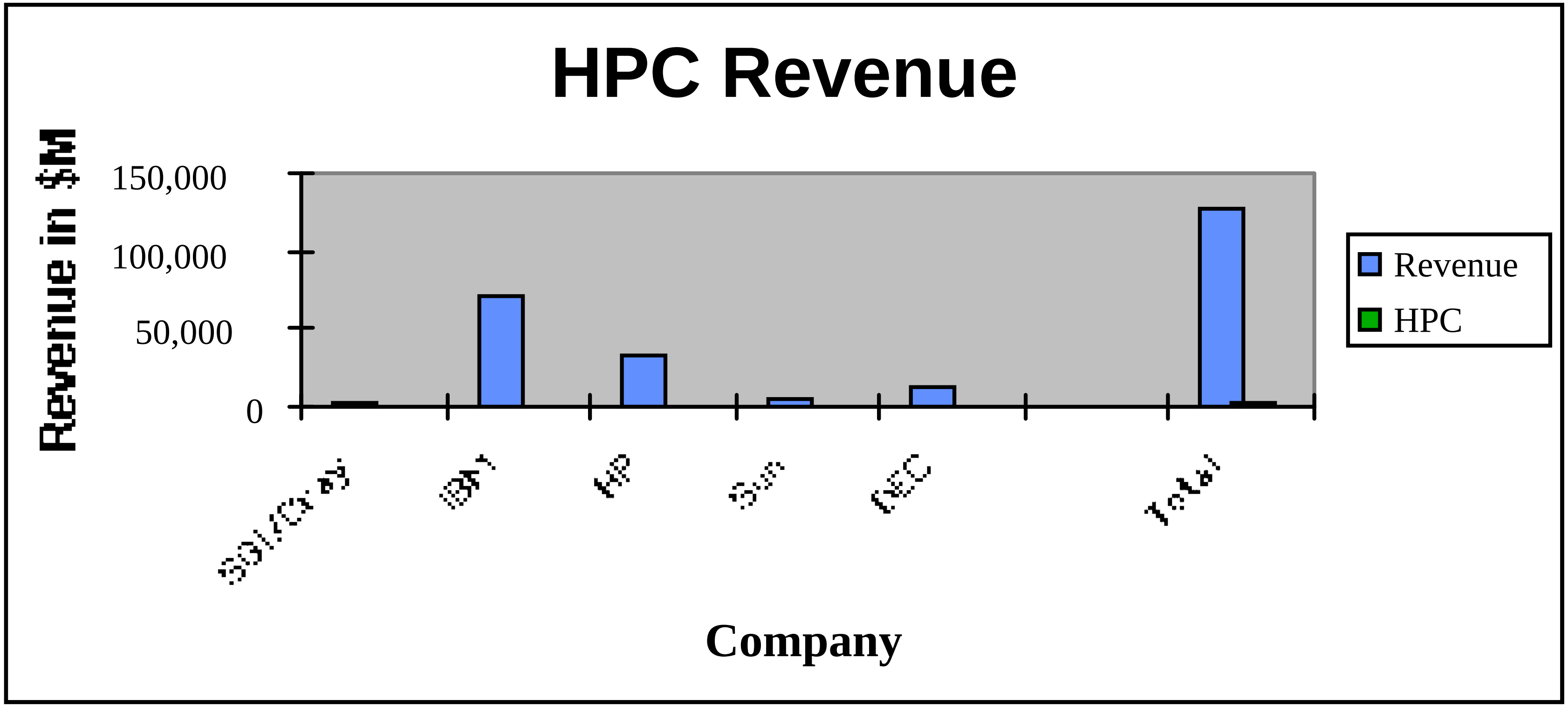
Almost all high-end procudcts are based onworkstation technology.



Since 1997:
The New HPC Marketplace

All major US HPC companies are now vertically integrated (SGI, IBM, HP, Sun, Compaq), with exception of Cray.

Almost all high-end products are based on workstation technology.



Since 1997: The New HPC Marketplace

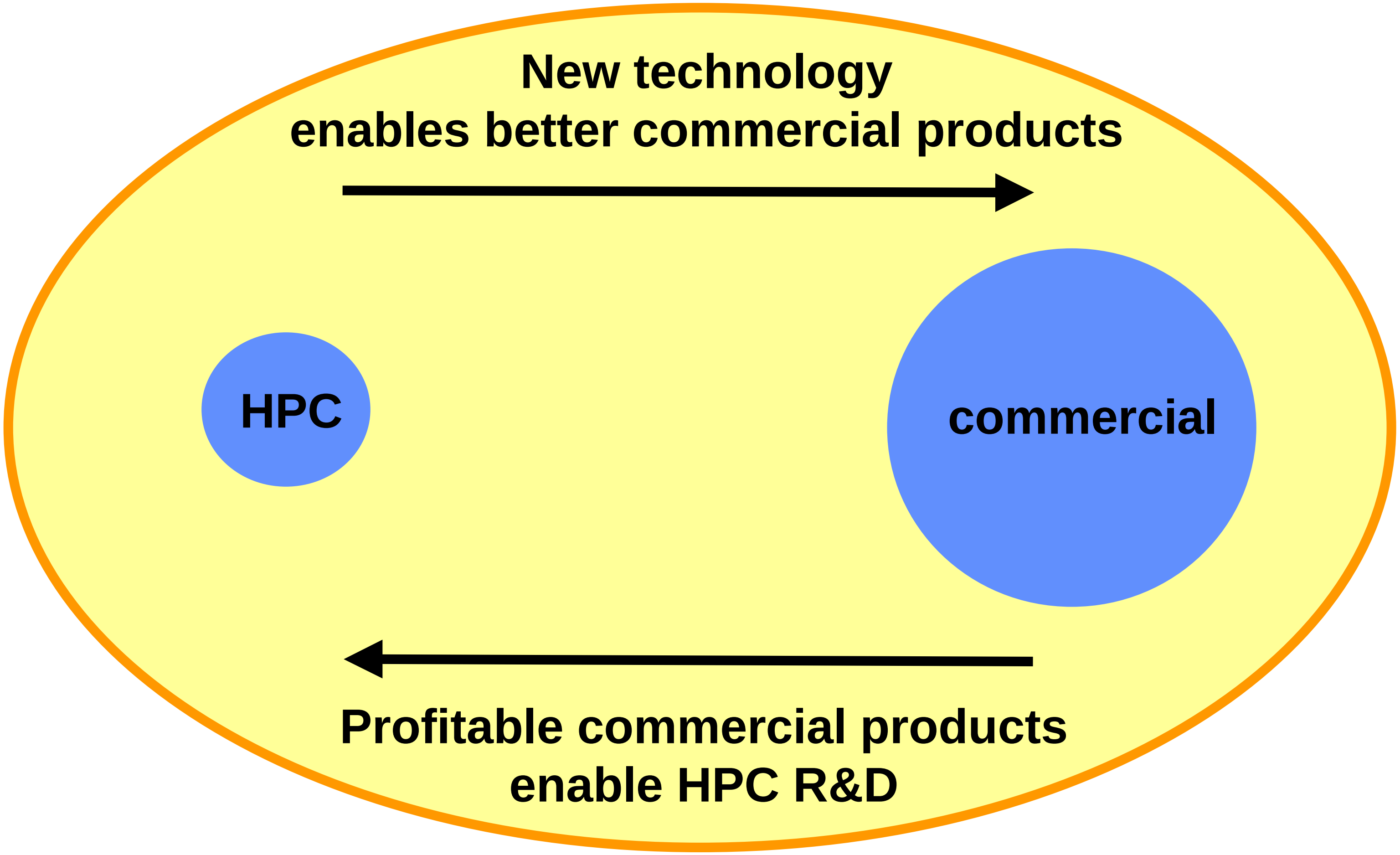
All these companies are in the computer business.

HPC customers must get used to a new role: they are no longer the center of attention.

Companies must have commitment to technology, and understand the potential of technology leverage from the high end, in order to remain in the HPC business.

Fortunately for us, the HPC users, they all do understand that (for now).

Since 1997:
The HPC Business Model



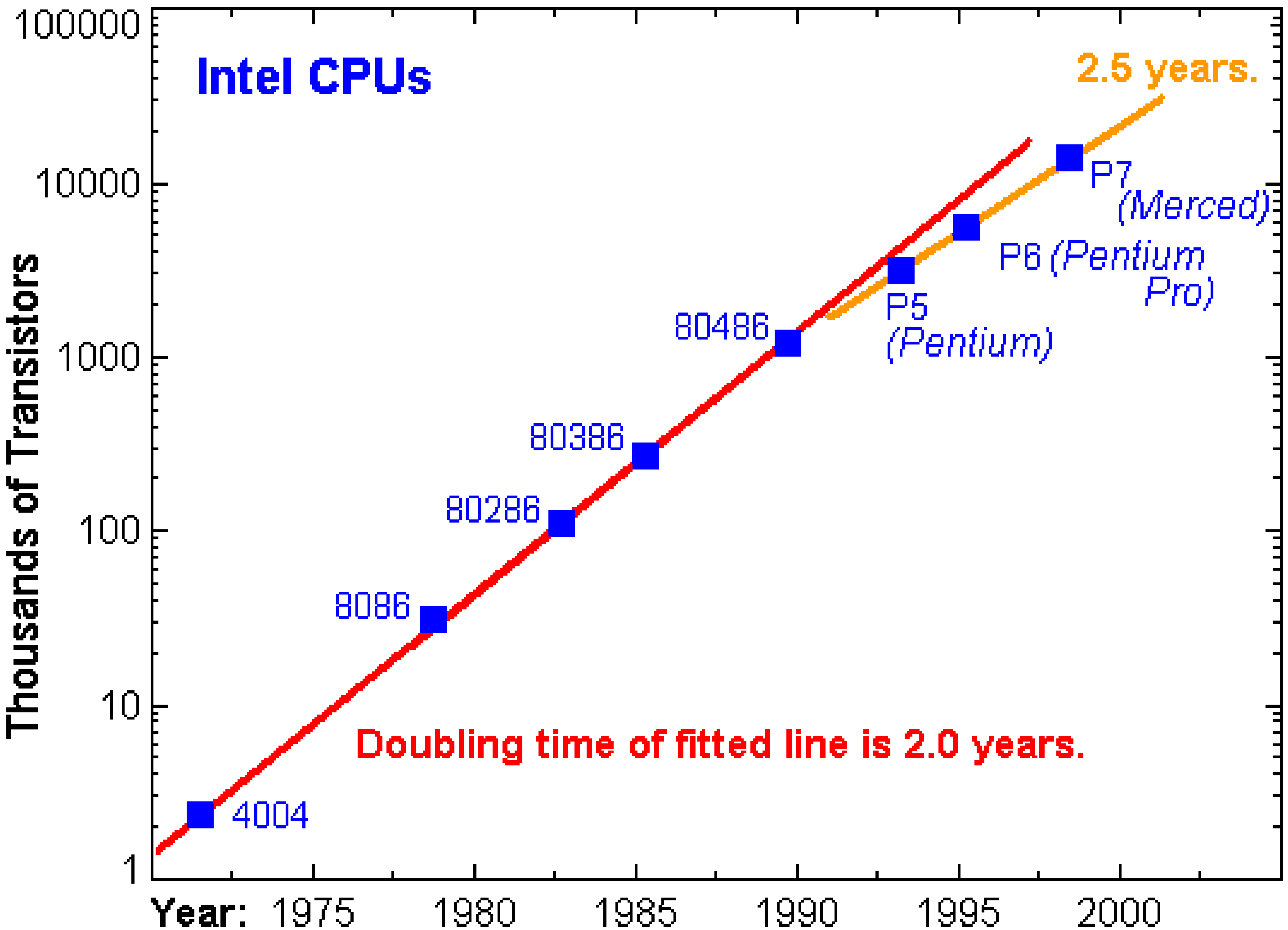
Summary on 1990s

- Moore's Law predicted technology change for HPC
- Major HPC trend of the 1990s was totally predictable
- Technology change radically changed business environment for HPC
- What was not predictable:
  - — rise of Internet, WWW
  - — impact of end of Cold War on HPC

Overview

- Retrospective: changes in the 1990s
- Extrapolation to the near future up to 2010
- The end of Moore's Law in about 2020
- Beyond 2025

Moore's Law —
The Traditional (Linear) View

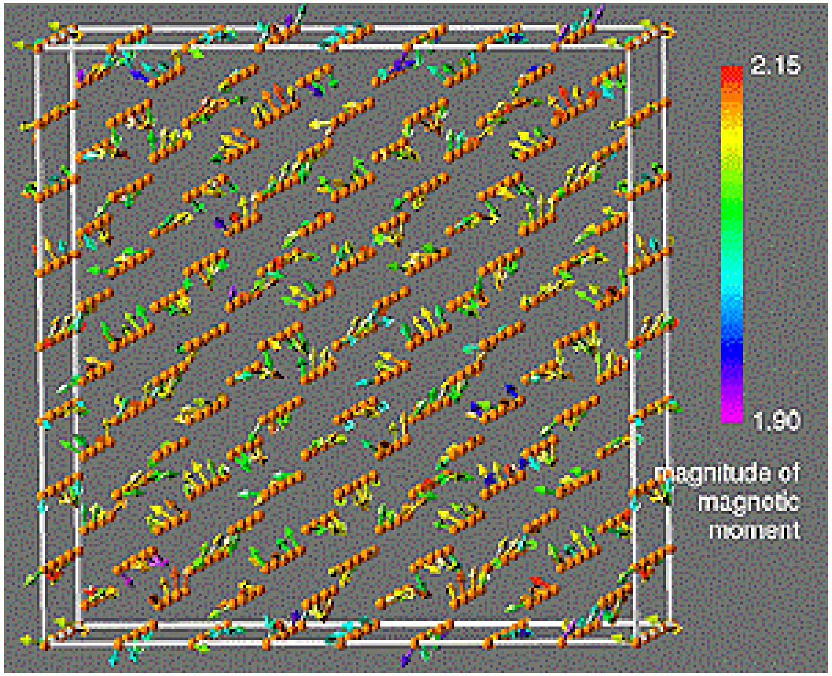


Reality Check on Real Applications

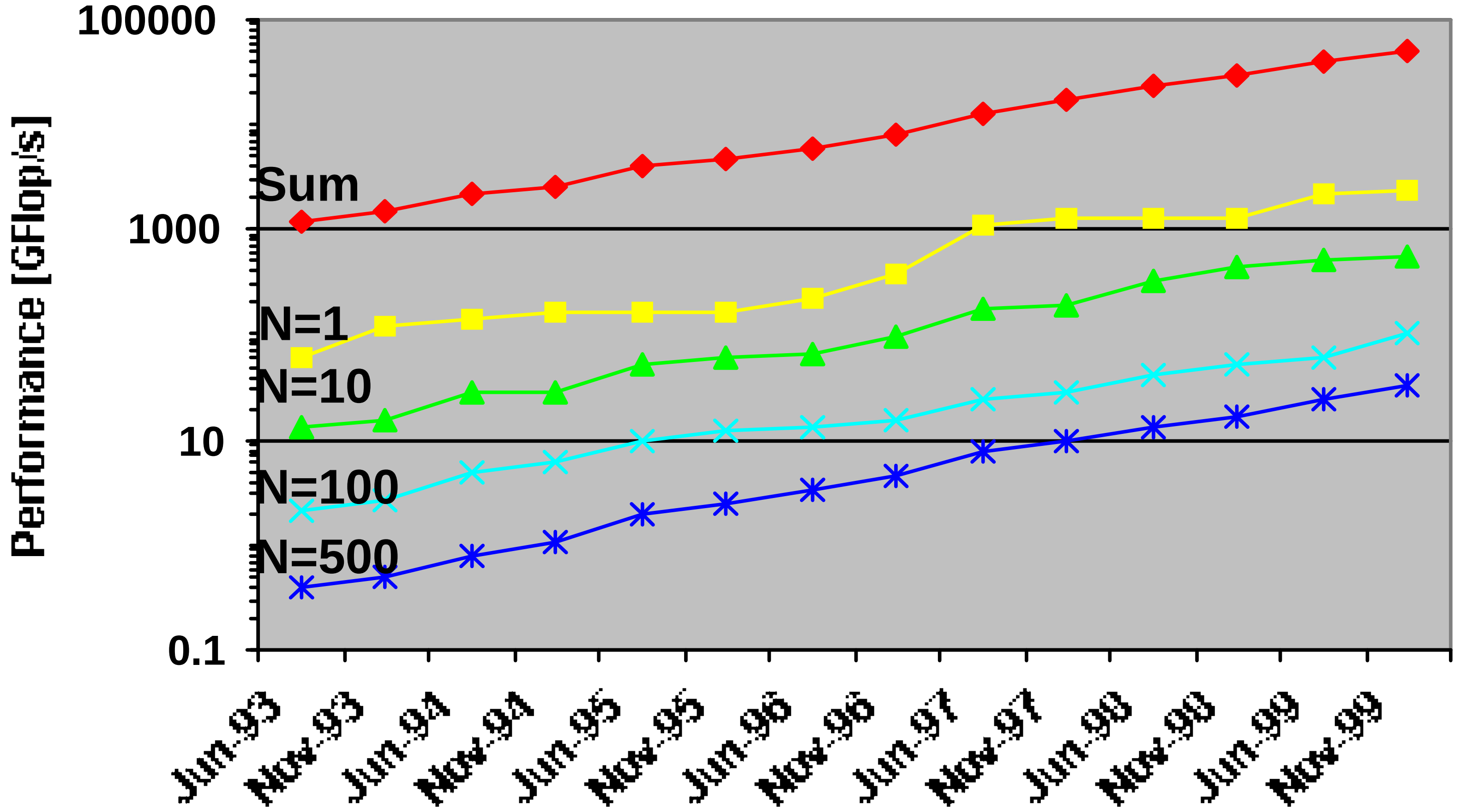
First complete application to break the 1Tflop/s sustained barrier in 1998.

Collaborators from DOE's Grand Challenge on Materials, Methods, Microstructure, and Magnetism.

1024-atom first-principles simulation of metallic magnetism in iron



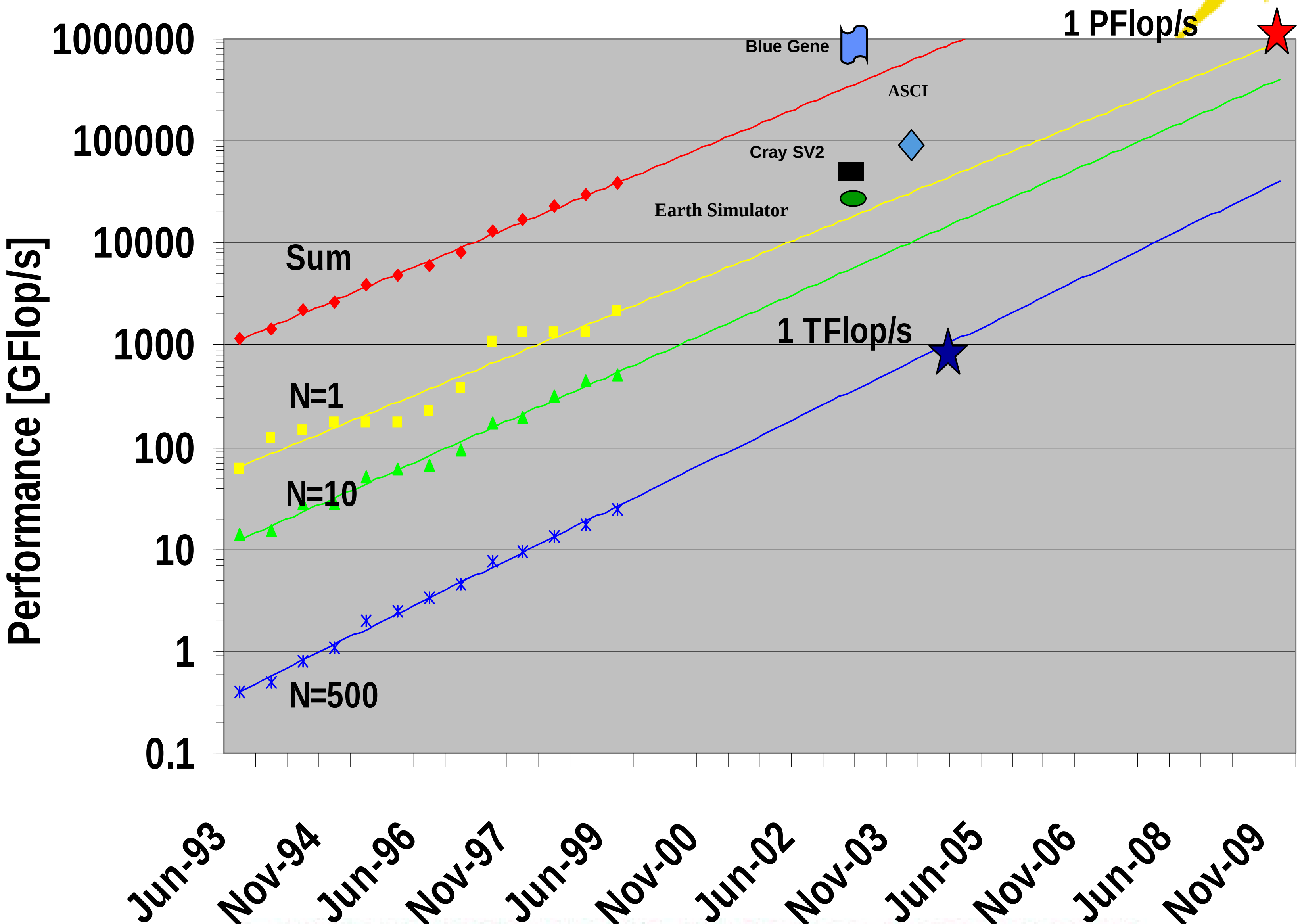
Performance Increases in the TOP500



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**
- **For more details see paper by Dongarra, Meuer, Simon, and Strohmaier in Parallel Computing (to appear)**

Extrapolation to the Next Decade



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 then 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**

2000-2005: Technology Options

- Clusters
  - — SMP nodes, with custom interconnect
  - — PCs, with commodity interconnect
  - — vector nodes (in Japan)
- Custom built supercomputers
  - — Cray SV-2
  - — IBM Blue Gene
  - — HTMT
- Other technology to influence HPC
  - — IRAM/PIM
  - — Computational and Data Grids

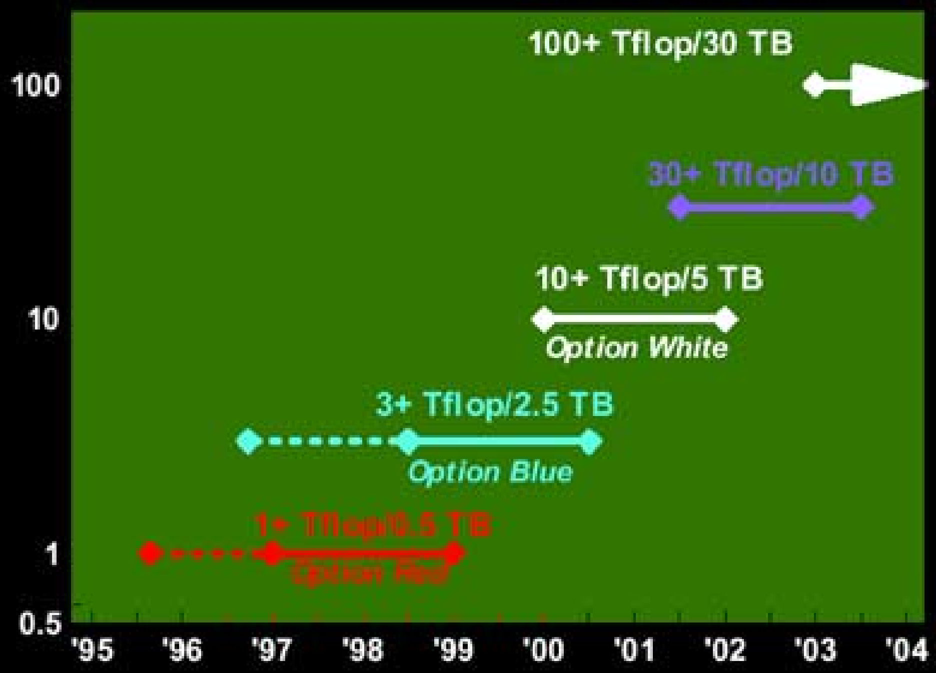
10 - 100 Tflop/s Cluster of SMPs

- The first ones are already on order
  - — LLNL has 10 Tflop/s on order for 2000
  - — NERSC will have a 3 Tflop/s system in 2000
- Systems are large clusters
  - — SMP nodes in US
  - — Vector nodes in Japan

- Programming model:
  - — OpenMP and/or vectors to maximize node speed
  - — MPI for global communication

Cluster of SMP Approach

- A Supercomputer is a "stretched" high-end server
  - parallel system, built by assembling nodes that are conventional, modest size, shared memory multiprocessor
  - just put more of them together

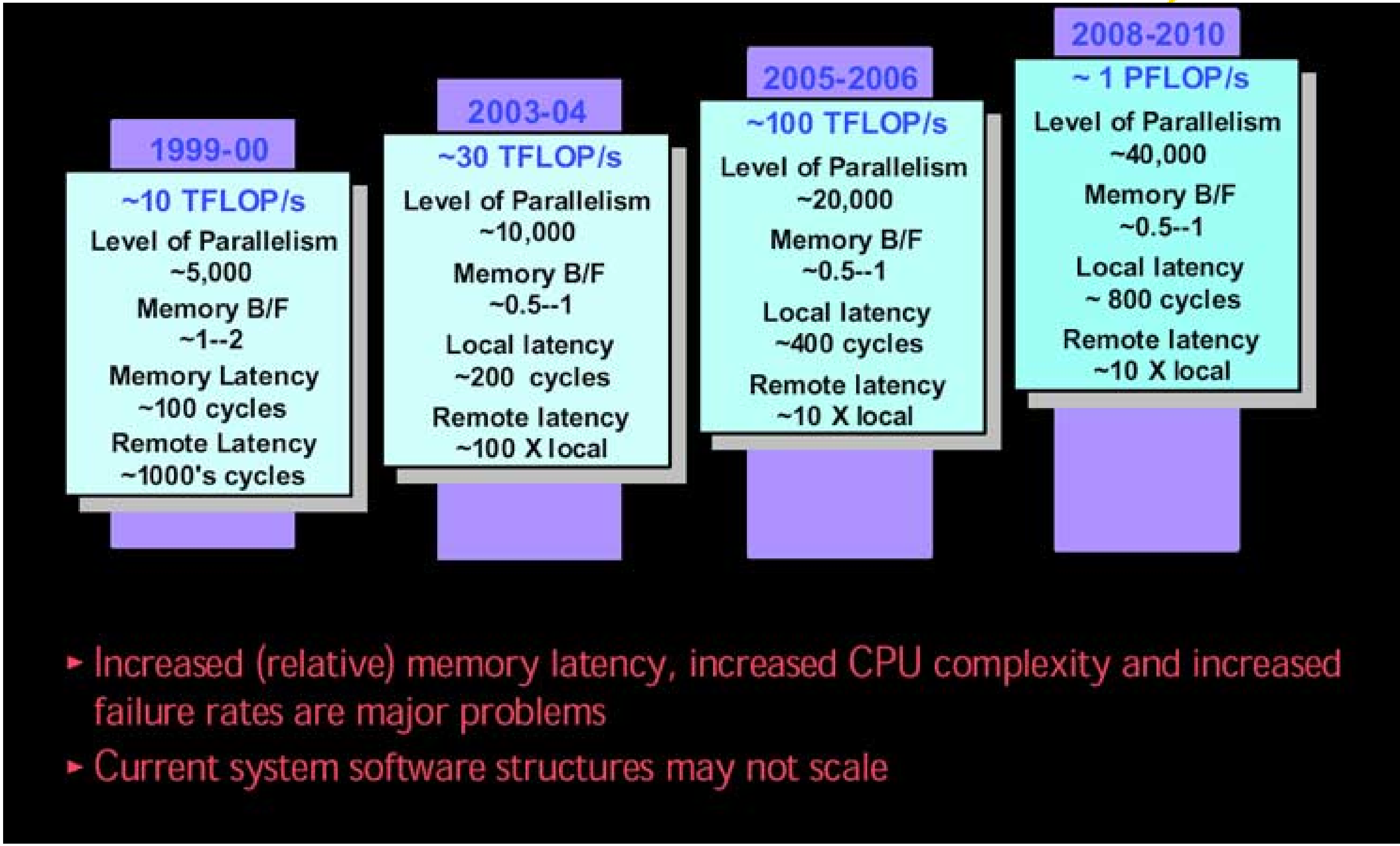


Accelerated Strategic Computing Initiative

ASCI Blue Pacific -- LLNL

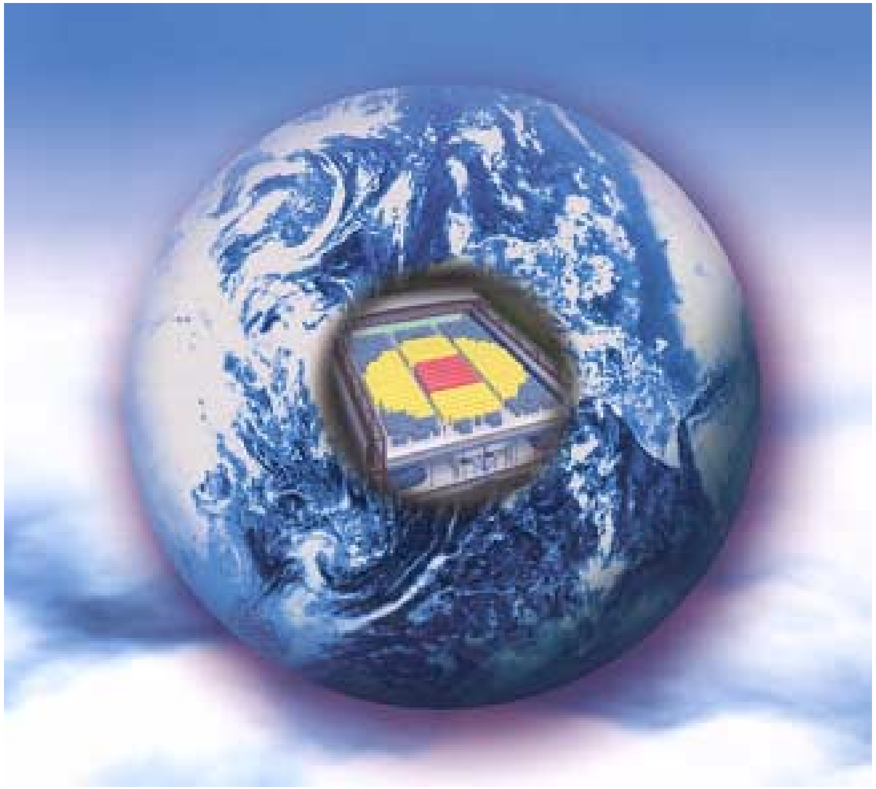
1,464 nodes; 5,856 CPUs
2.6 TB memory
80 TB disk
3.3 TFlop/s demonstrated

100 - 1000 Tflop/s Cluster of SMPs (IBM Roadmap)

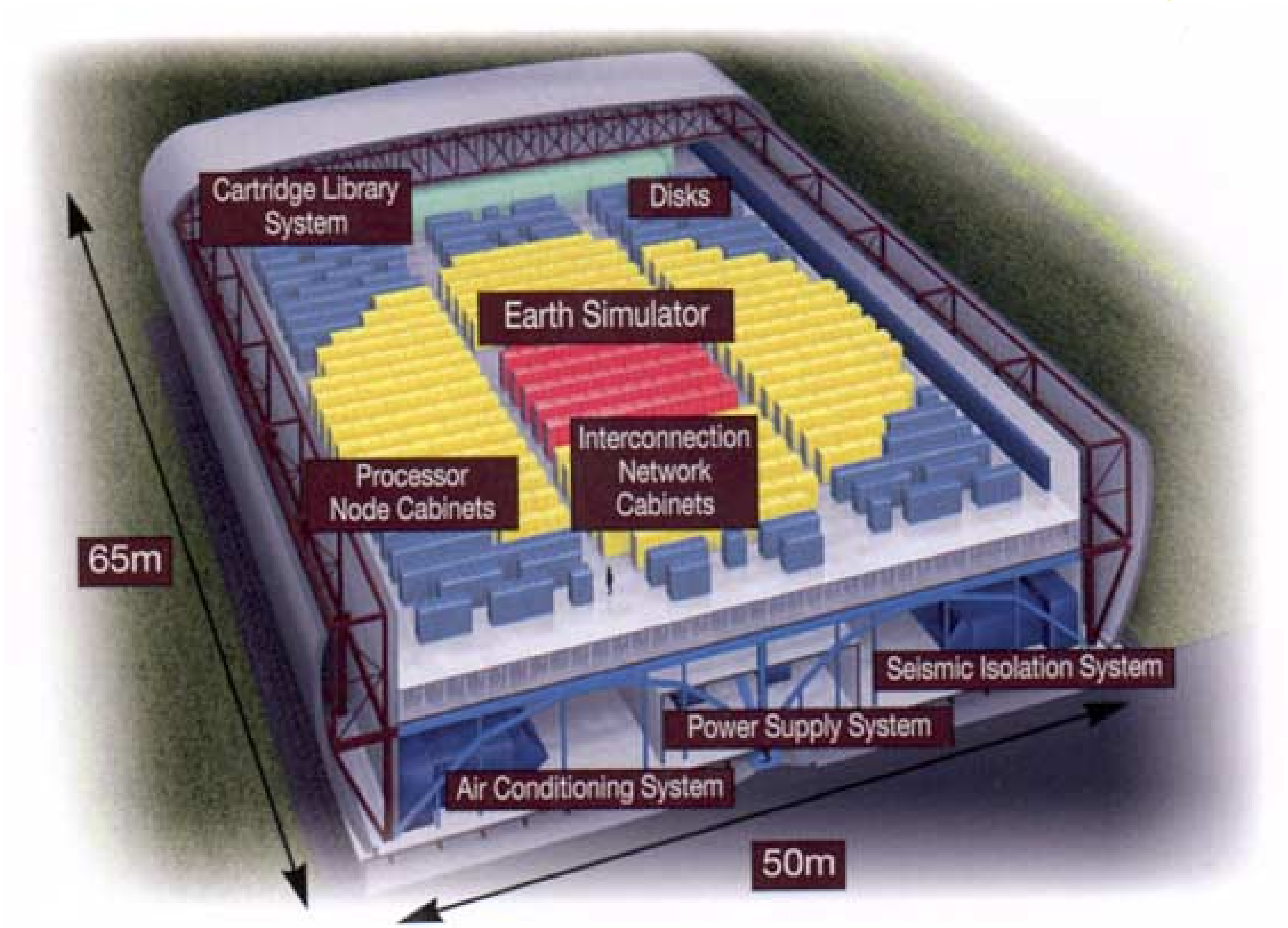


Global Earth Simulator

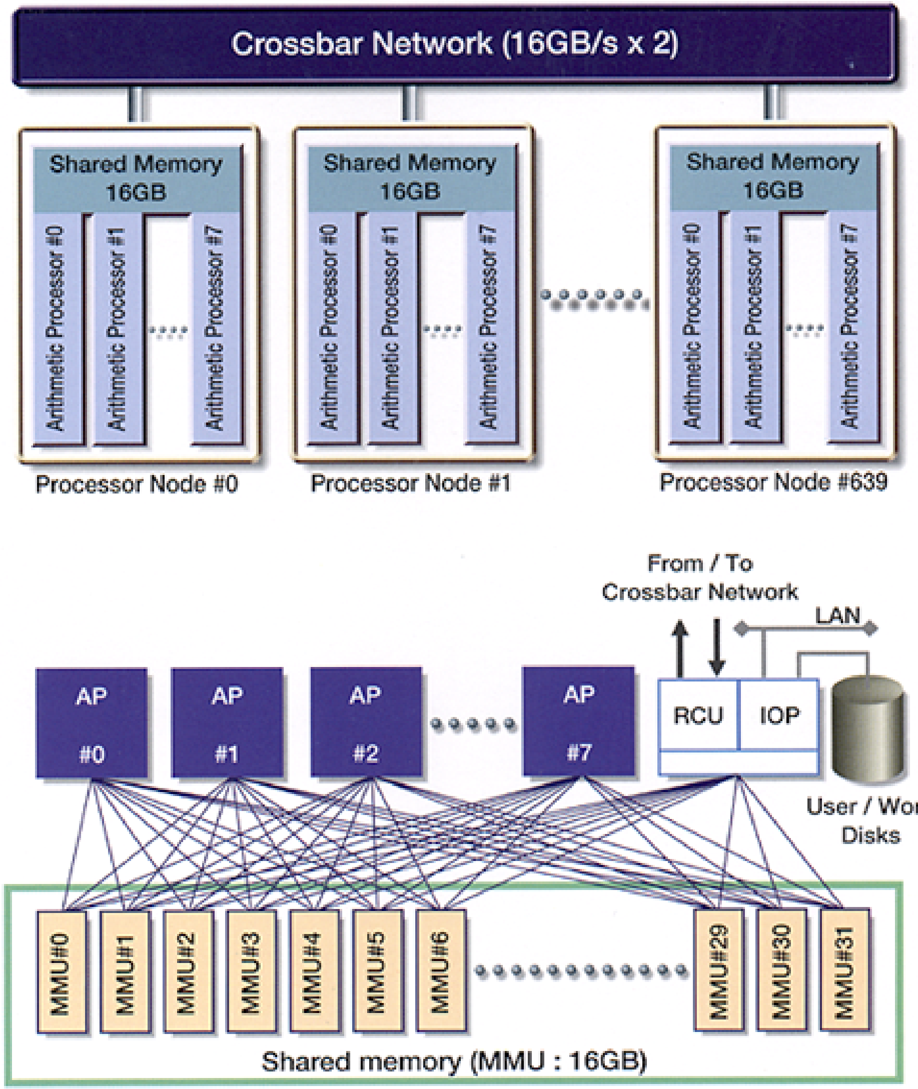
- 30 Tflop/s system in Japan
- completion 2002
- driven by climate and earthquake simulation requirements
- built by NEC
- CMOS vector nodes



Global Earth Simulator Building



Earth Simulator



Cray SV2 Overview:

— Basic building block is a 50/100 GFLOPs node:

— 4 x CPUs per node. IEEE. Design goal is 12.8 GFLOPs per CPU.

— 8, 16 or 32 GB of coherent flat shared memory per CPU

— SSI to 1024 nodes: 50/100 TFLOPs, 32TB:

— 100 GB/sec interconnect capacity to/from each node

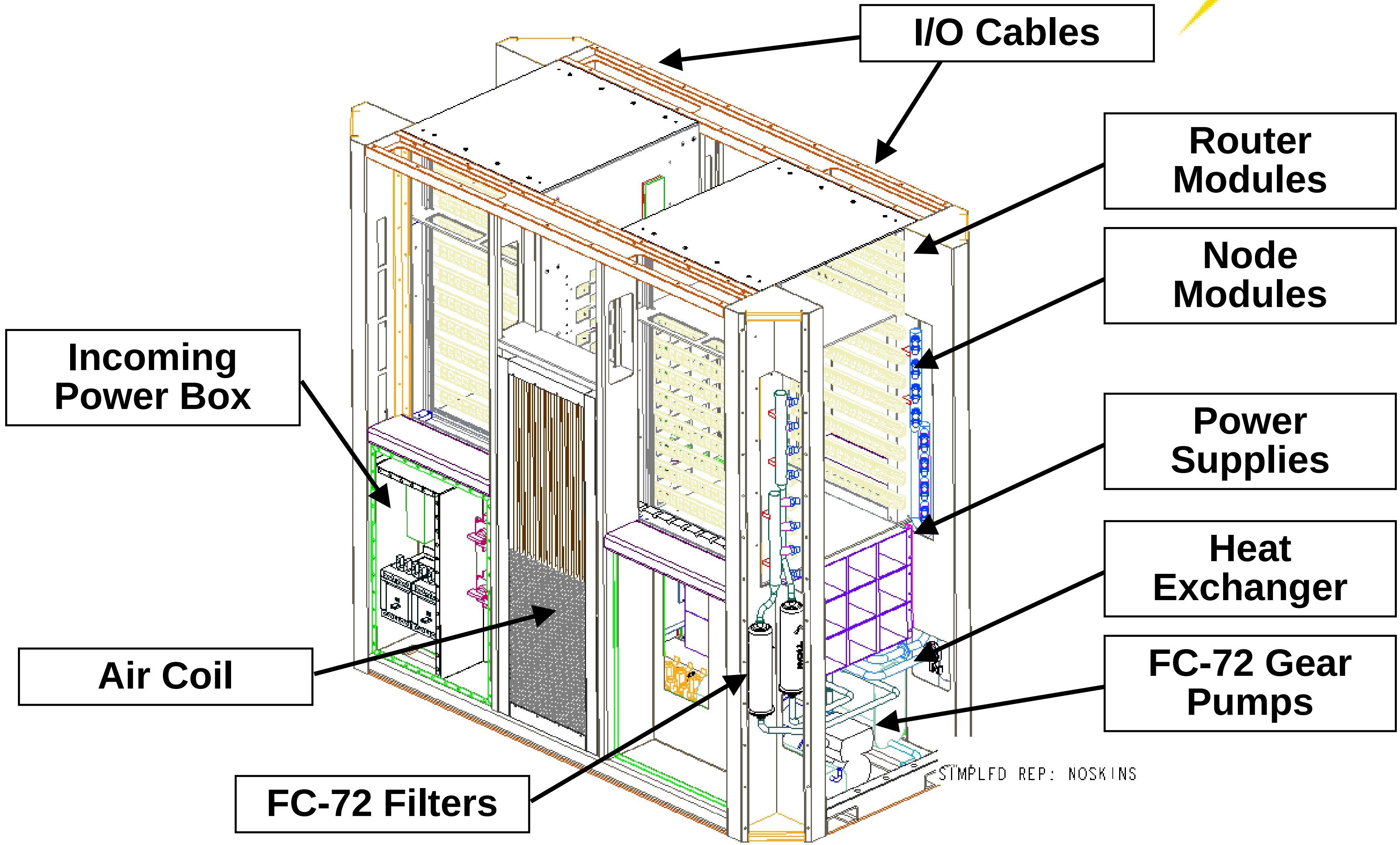
— ~1 microsecond latency anywhere in hypercube topology

— Targeted date of introduction, mid-2002.

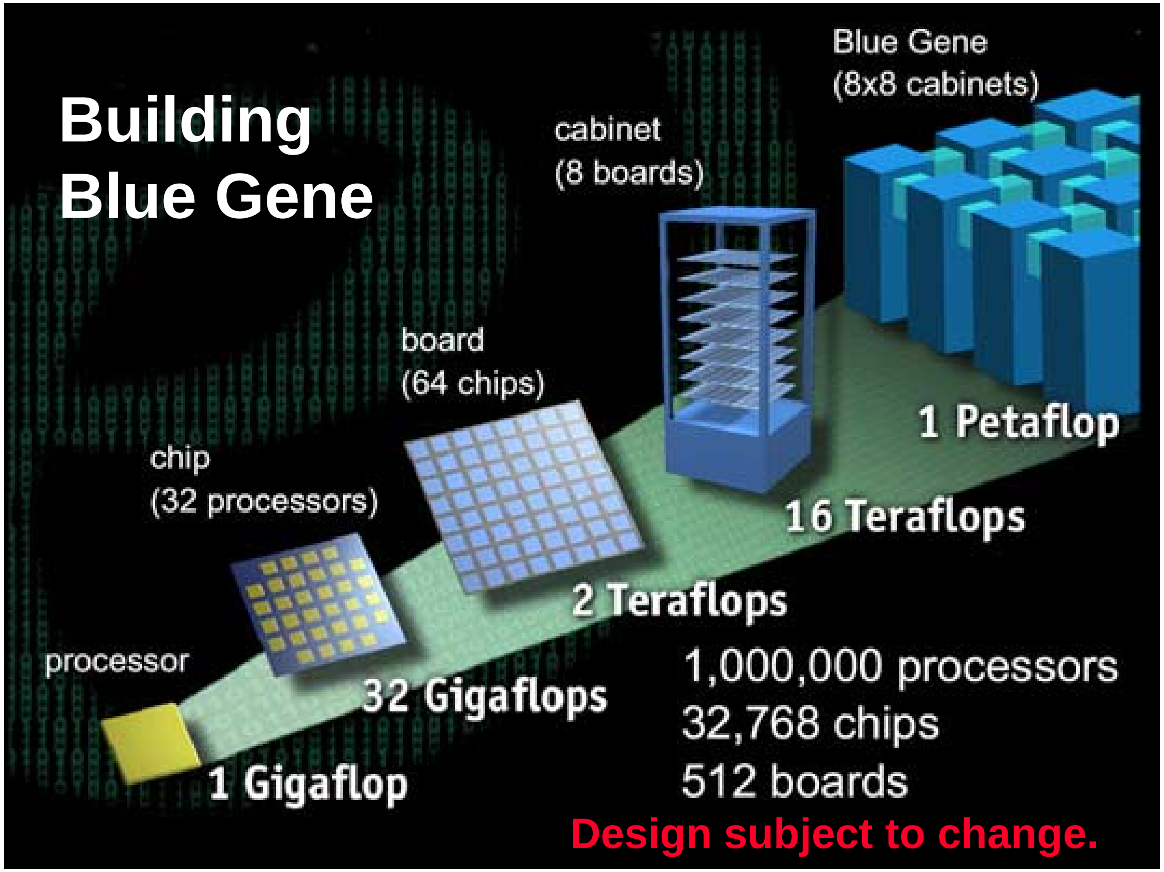
— LC cabinets; Integral HEU (heat exchange unit)

— Up to 64 cabinets (4096 CPUs/50 TFLOPS) mesh topology

Liquid-Cooled Cabinet — 64 CPUs



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**

Deep Blue Architecture

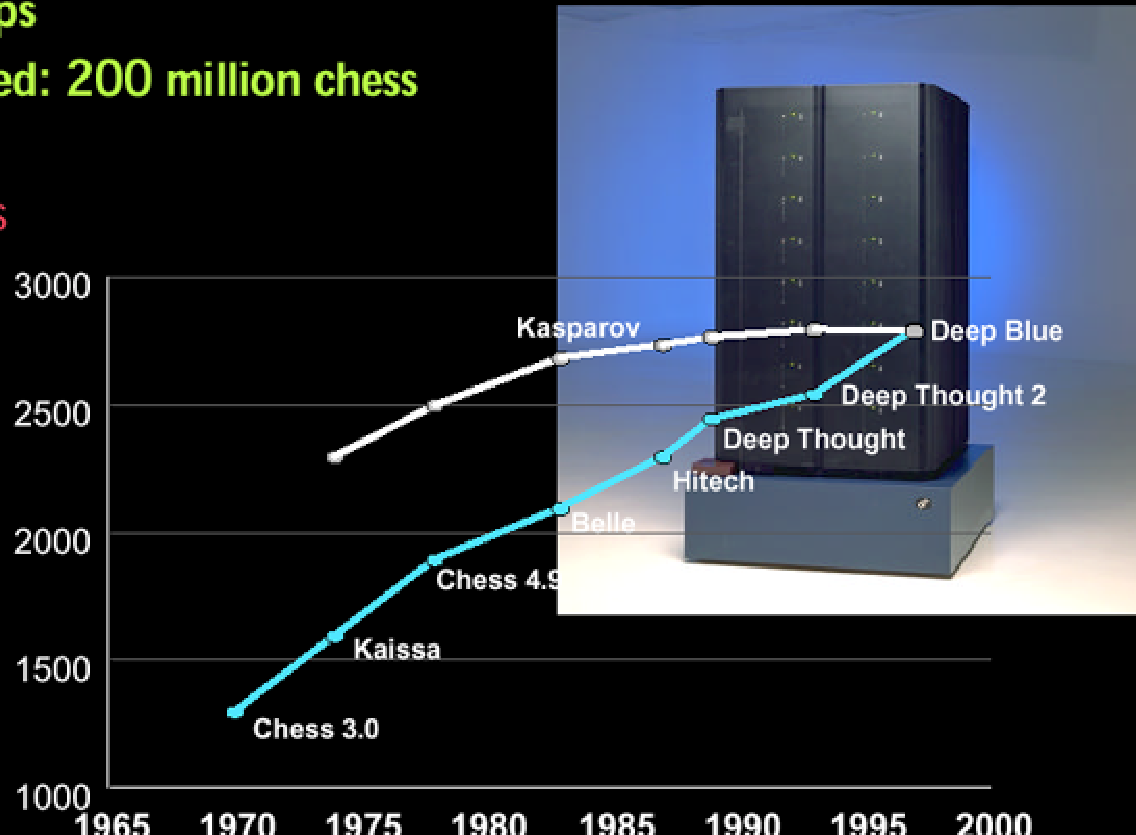
- A general purpose parallel computer + custom accelerators for specific algorithm
  - 30 node IBM SP system
  - 480 accelerator chips
  - Average system speed: 200 million chess positions per second

very good at playing chess (the best!)

x100 -- x1000 a general purpose system of similar complexity

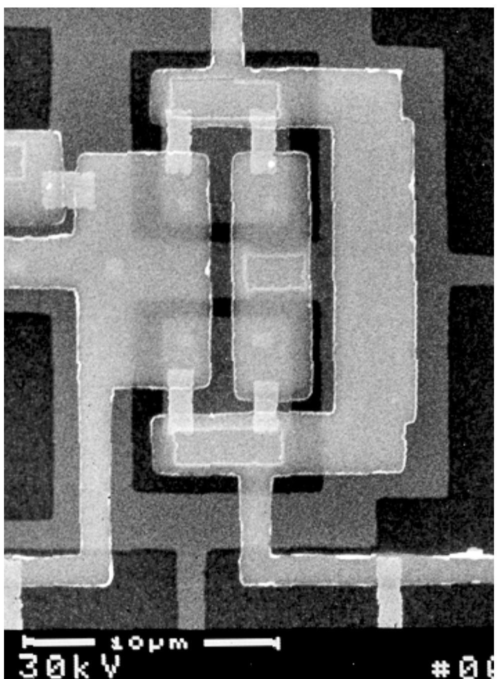
- not good at anything else

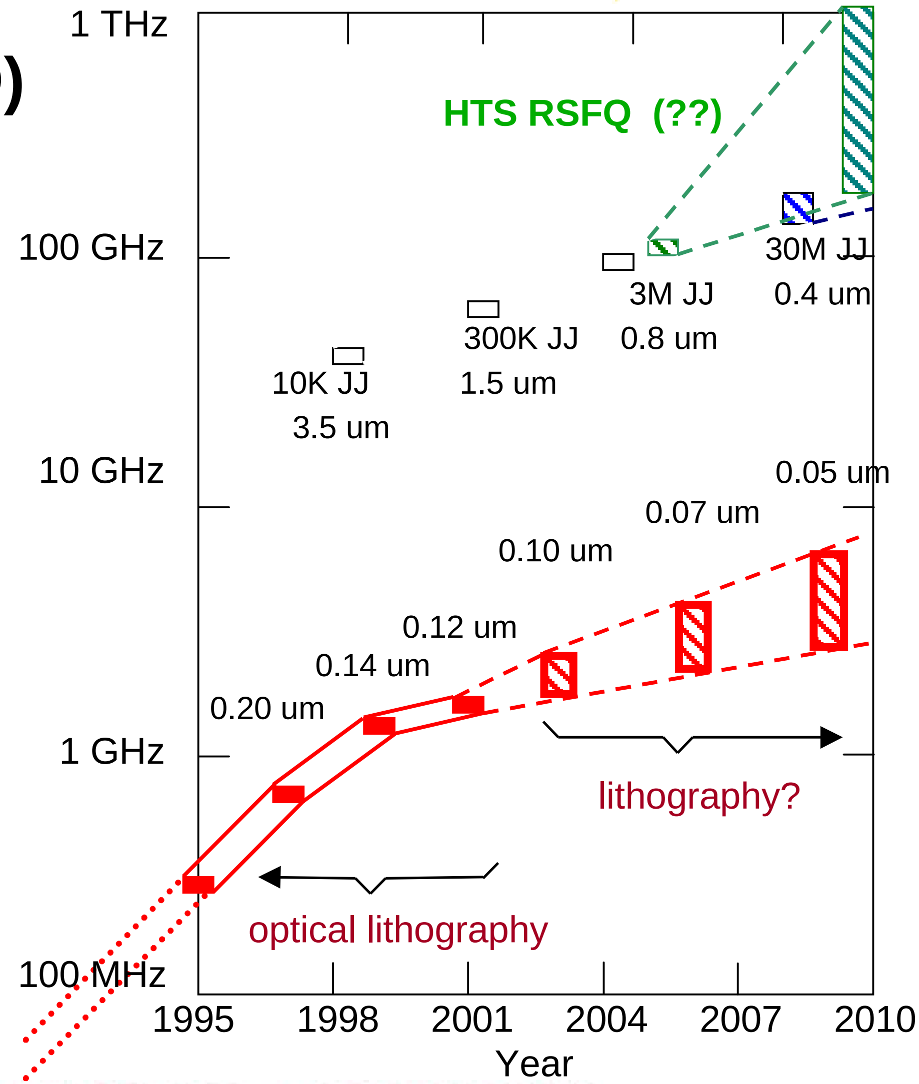
chess rules do not change, but simulation rules change!



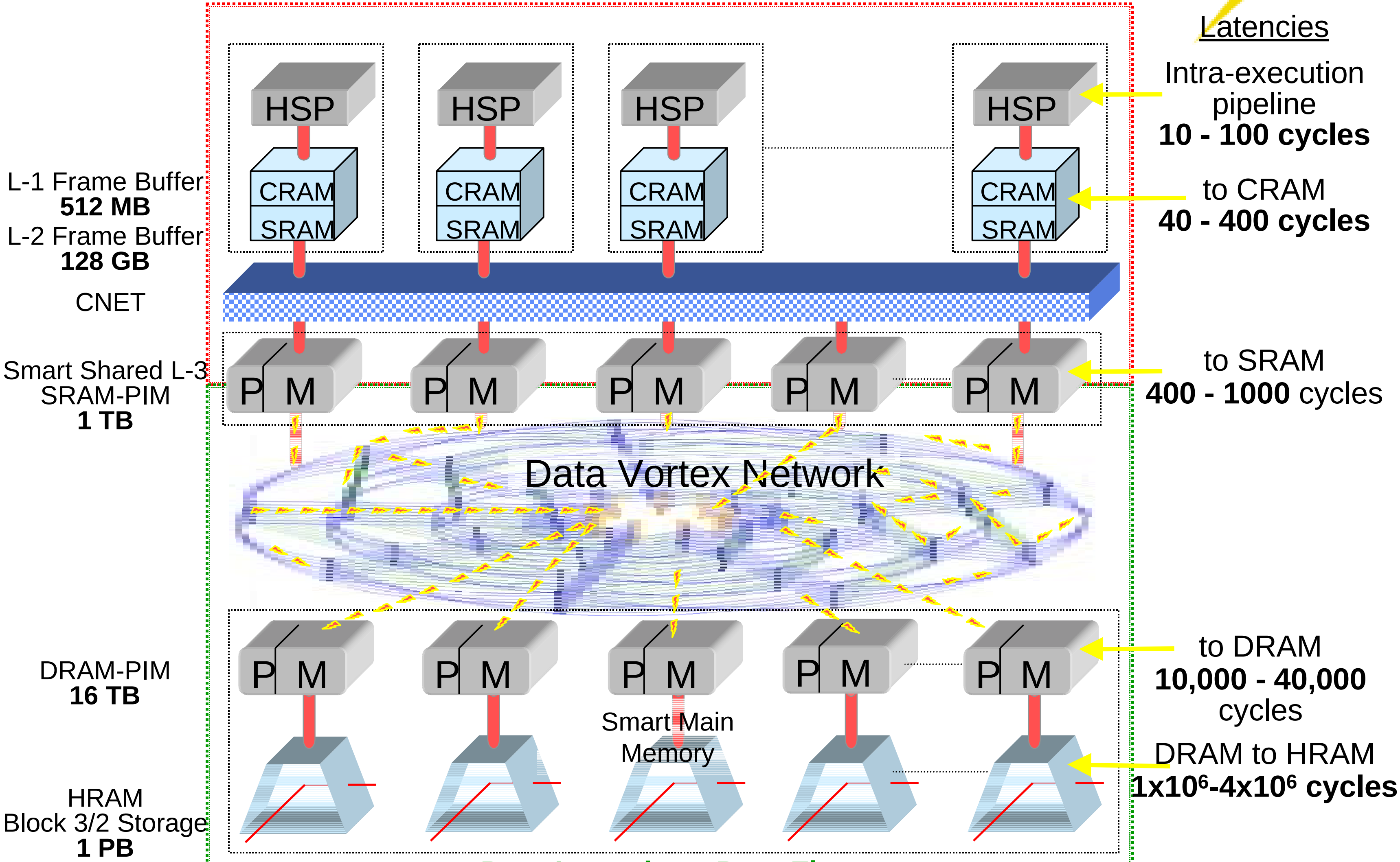
An Alternate Technology?

- **Single Flux Quantum (SFQ)**
- **Operates at 4 Kelvin**





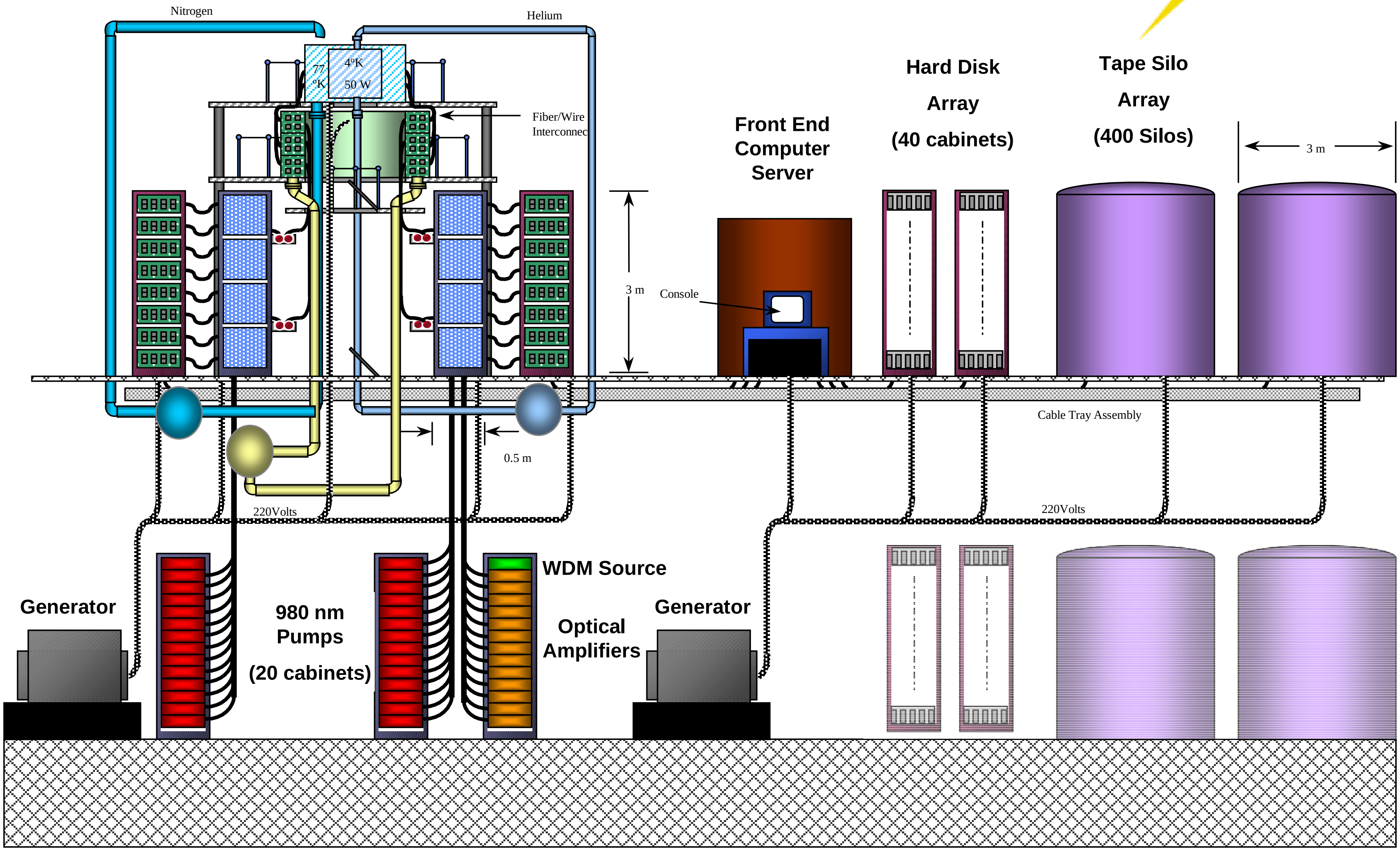
Hybrid Technology, Multithreaded Architecture



Data Intensive—Data Flow Driven

HTMT Machine Room

HTMT Cross-Section



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)

Open Source Will Change the Rules!

- Stage 1: (40s and 50s): every computer different, every program unique
- Stage 2: (60s and 70s): software is unbundled from hardware, commercial software companies arise
- Stage 3: (80s and 90s): mass market computers and mass market software, the notions of software copyright and privacy are born
- Stage 4: (2000 and beyond): software migrates to the WWW, OSS communities provide high quality software

Commercially Integrated Clusters Are Already Happening

- **Forecast Systems Lab procurement (Prime contractor is High Performance Technologies Inc., subcontractor is Compaq)**

- **Los Lobos Cluster (IBM with University of New Mexico)**

Linux Supercomputer Howls

by John Gartner and Mi

3:00 a.m. Mar. 22, 2000 PST

The University of New Mexico and IBM are teaming up to build the world's fastest Linux-based supercomputer.

Named "LosLobos", the new supercomputer is scheduled to be fully operational by the summer.

CHNOLOGY

Also:
SGI Sells a Piece of Its Heart
Caldera Set to Ride Linux Wave
News from the Linux front
Read more Technology news

ay's Headlines
) p.m. Mar. 22,
) PST

ux
ercomputer
vls

bit, Don't Run

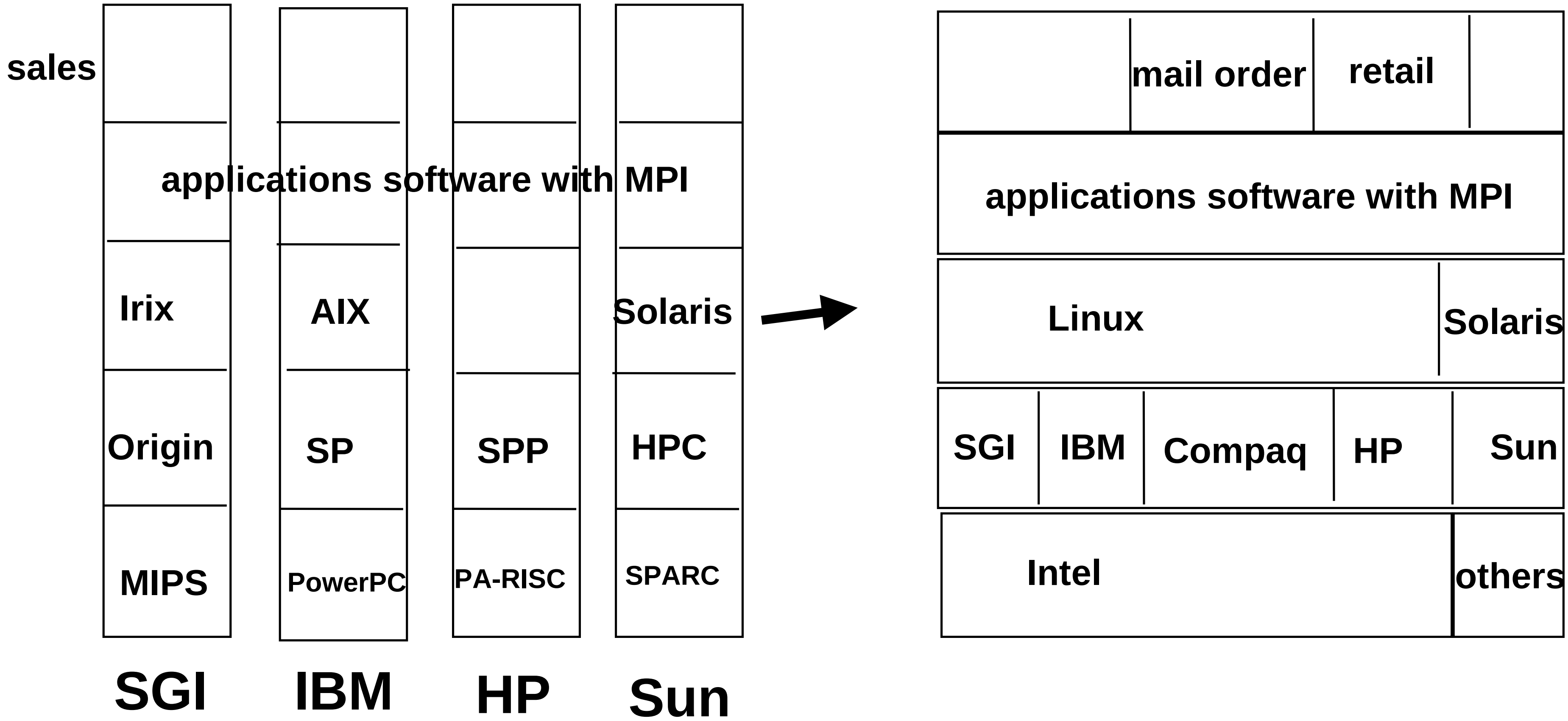
LosLobos is a departure from the traditional supercomputer set-up. It's built from 256 IBM Netfinity servers.

The Netfinity servers are linked together using special clustering software and high-speed networking hardware, which causes the separate units to act as one computer, delivering a processing speed of 375 gigaflops, or 375 billion operations per second.

2000-2005: Market Issues

From vertical to horizontal companies—

the Compaq model of High Performance Computing



Until 2010: Market Issues

Business transition will be more fundamental than previous technology transition.

Tremendous impact on HPC community — no more business as usual (e.g., how do we procure machines)

Extremely difficult to pick winner

Tumultuous transition may make it difficult for boutique companies such as Cray, Inc. to survive

Open source is not just a PC cluster issue, it will be industry wide

Until 2010: A New Parallel Programming Methodology? - NOT

The software challenge: overcoming the MPI barrier

- **MPI created finally a standard for applications development in the HPC community**
- **Standards are always a barrier to further development**
- **The MPI standard is a least common denominator building on mid-80s technology**

Programming Model reflects hardware!

"I am not sure how I will program a Petaflops machine, but I am sure that I will need MPI somewhere"

New Economic Driver: IP on Everything

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

Enablers of Pervasive Technologies

- General accessibility through intuitive interfaces
- A supporting infrastructure, perceived valuable, based on enduring standards
- MOSAIC browser and World Wide Web are enablers of global information infrastructure

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

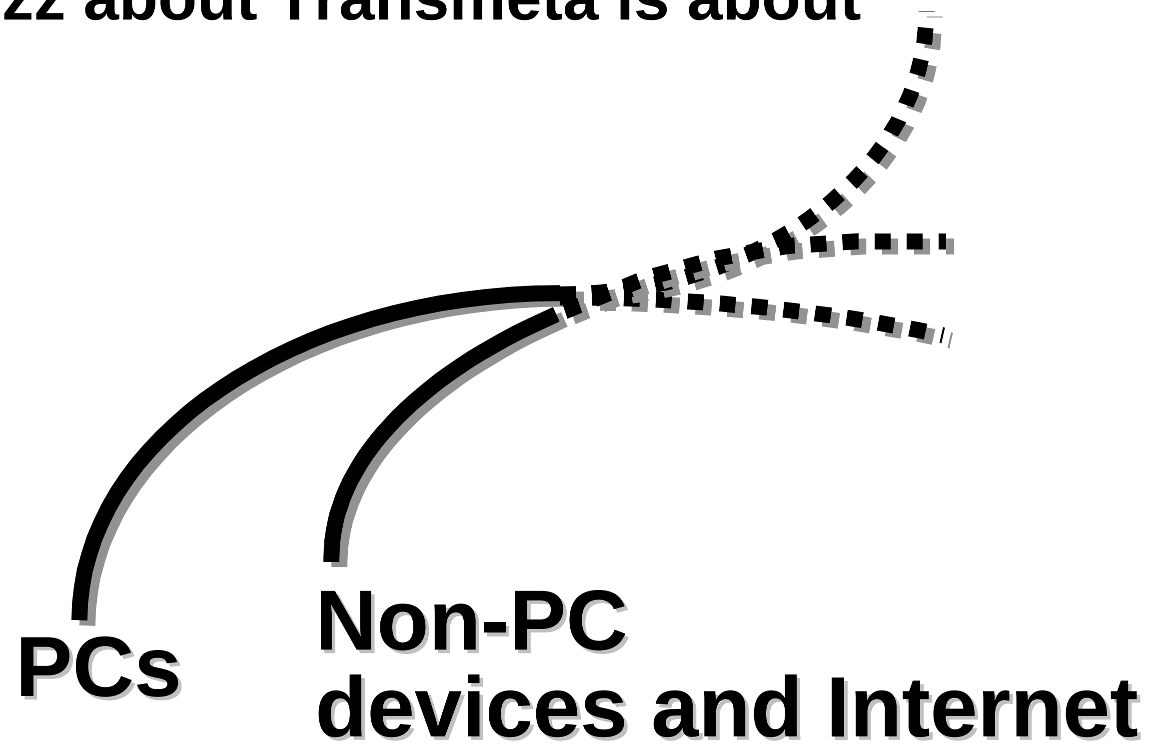
Information Appliances

- 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

Economic driver from below

- **Mass produced cheap processors will bring microprocessor companies increased revenue**
- **system on a chip will happen soon**
- **that is what the buzz about Transmeta is about**



“PCs at Inflection Point”,
Gordon Bell, 2000

ISTORE Hardware Vision

- System-on-a-chip enables computer, memory, without significantly increasing size of disk
- 5-7 year target:

- MicroDrive:1.7” x 1.4” x 0.2”
  - 2006: ?
  - 1999: 340 MB, 5400 RPM,
    - 5 MB/s, 15 ms seek
  - 2006: 9 GB, 50 MB/s ? (1.6X/yr capacity, 1.4X/yr BW)
- Integrated IRAM processor
  - 2x height
- Connected via crossbar switch
  - growing like Moore’s law
- 16 Mbytes; ; 1.6 Gflops; 6.4 Gops
- 10,000+ nodes in one rack! 100/board = 1 TB; 0.16 Tf

Questions for 2010?

- What will the system-on- a-chip and IP-on-everything-trend do to HPC?
- Imagine:
  - engineering/scientific workstations deliver 100 Gflop/s, but …
  - the engineering/scientific market is marginal
  -

What am I willing to predict?

2010:

- 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 to "SMP servers" are possible)
- The "grid" will have happened, because a killer app made it commercially viable
- An incredible tale like:
  - Microsoft will be split into three companies; in 2004 the Microsoft applications company buys the near bankrupt Cray Inc.; $$ are spent in revamping the Tera MTA; the company loses focus on its key applications; word processing, spreadsheets etc. are provided by open source competitors ...

Overview

- Retrospective: changes in the 1990s
- Extrapolation to the near future up to 2010
- The end of Moore's Law in about 2020
- Beyond 2025

In the 2010s: Pervasive Computational Modeling

Commodity consumer products

Example:

MOTOROLA, Pager Division, Boynton Beach, Florida

Applications: Radioss/Parallel Solids
ABAQUS Standard/Explicit
Alias - Render Industrial Designs
EFMASS, MDS, from H.P., MCSPICE

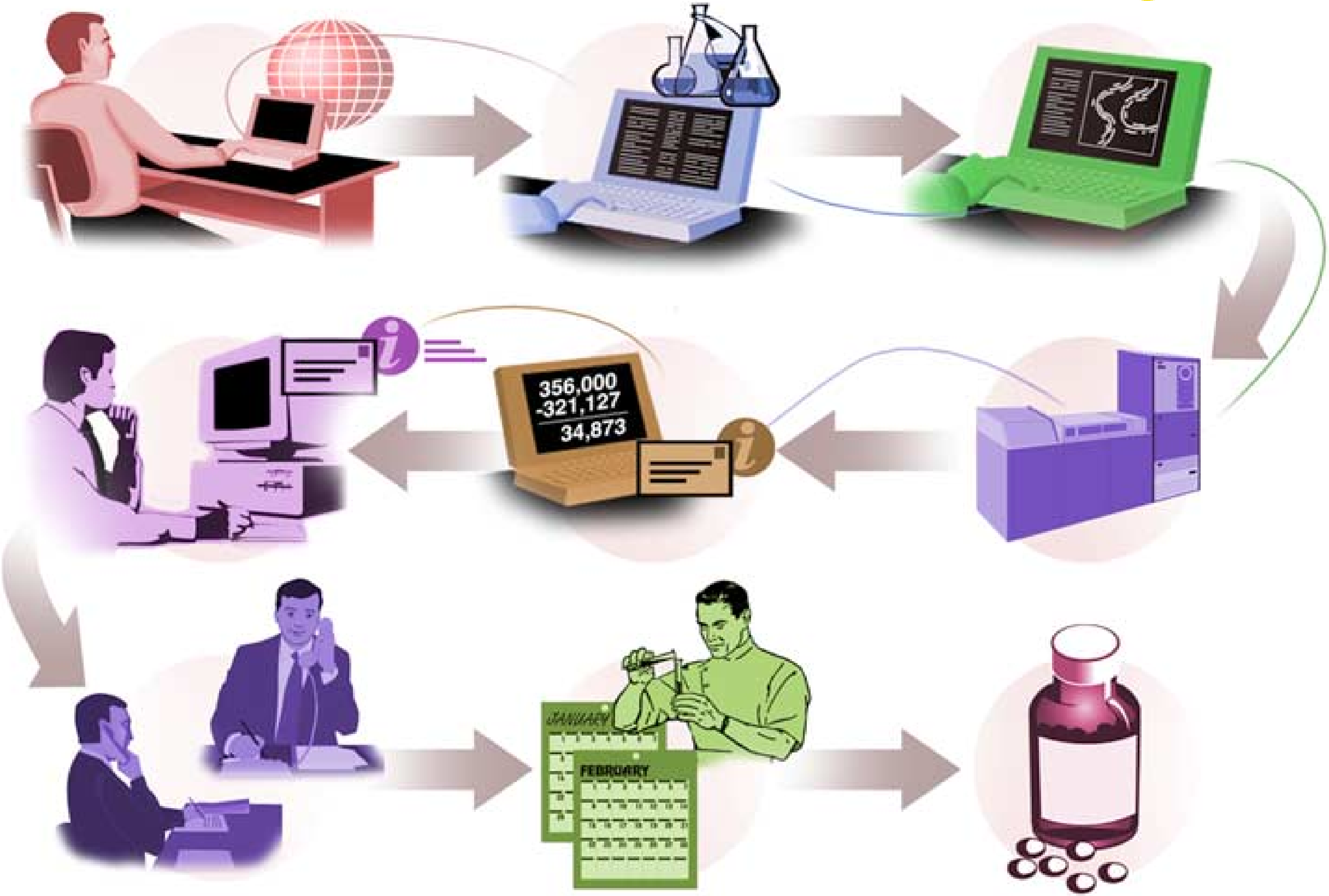
System: 8 CPU POWER CHALLENGE
2 GB Memory, 40GB Disk

Problem: Pager Case
- Battery Containment
- Electronics Integrity
- Display Life

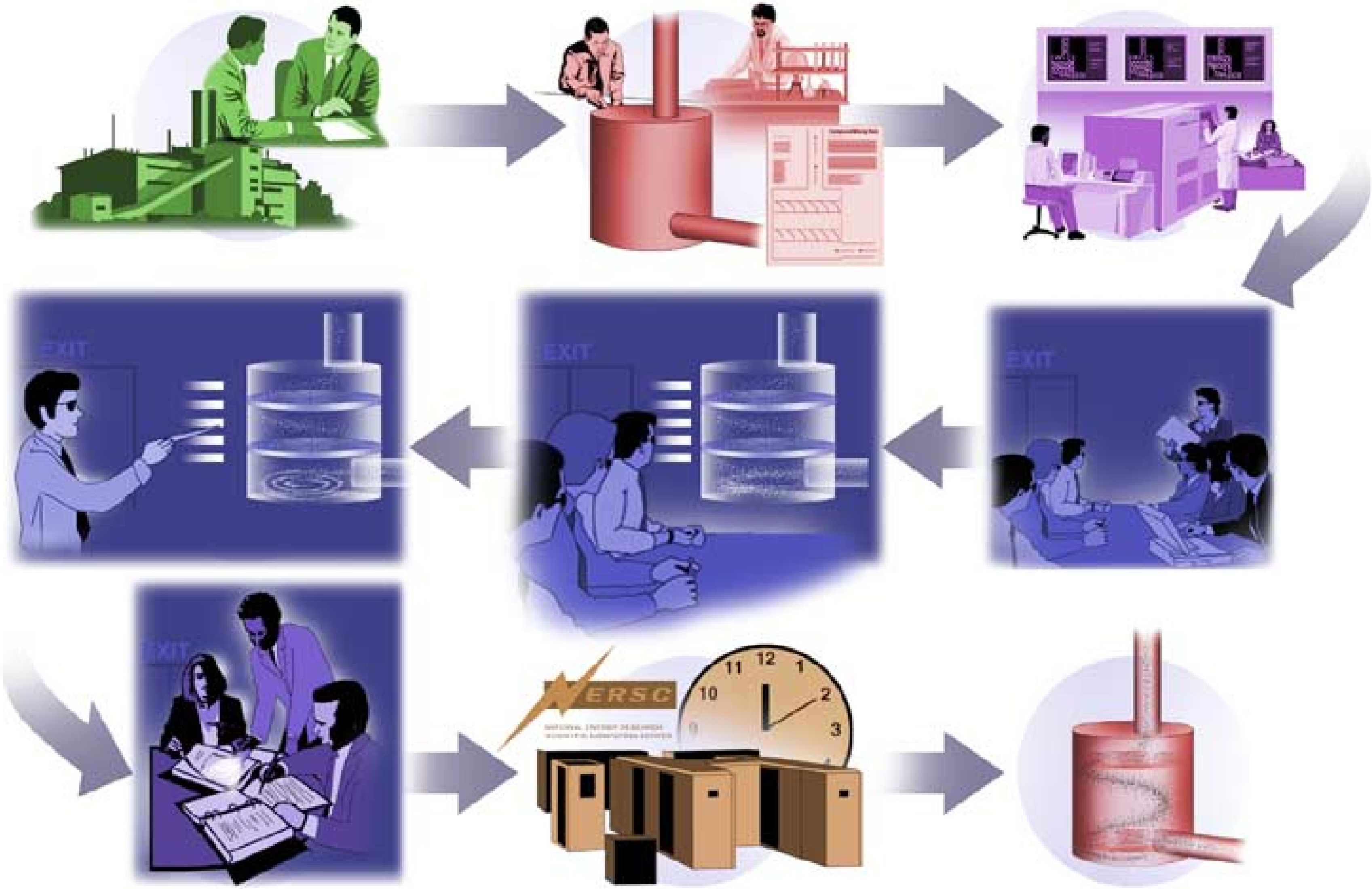
Towards Ubiquitous Computational Modeling

1985	1990	1995
specialized hardware	specialized hardware	commodity hardware
Cray X-MP	Cray Y-MP	POWER CHALLENGE XL
nuclear weapons lab.	industrial company unique control resource	industrial company decentralized divisional resource
unique multimillion $ product (weapons impact)	expensive consumer product $10K (car crash)	mass consumer product $1.99 (pager/cellular phone)

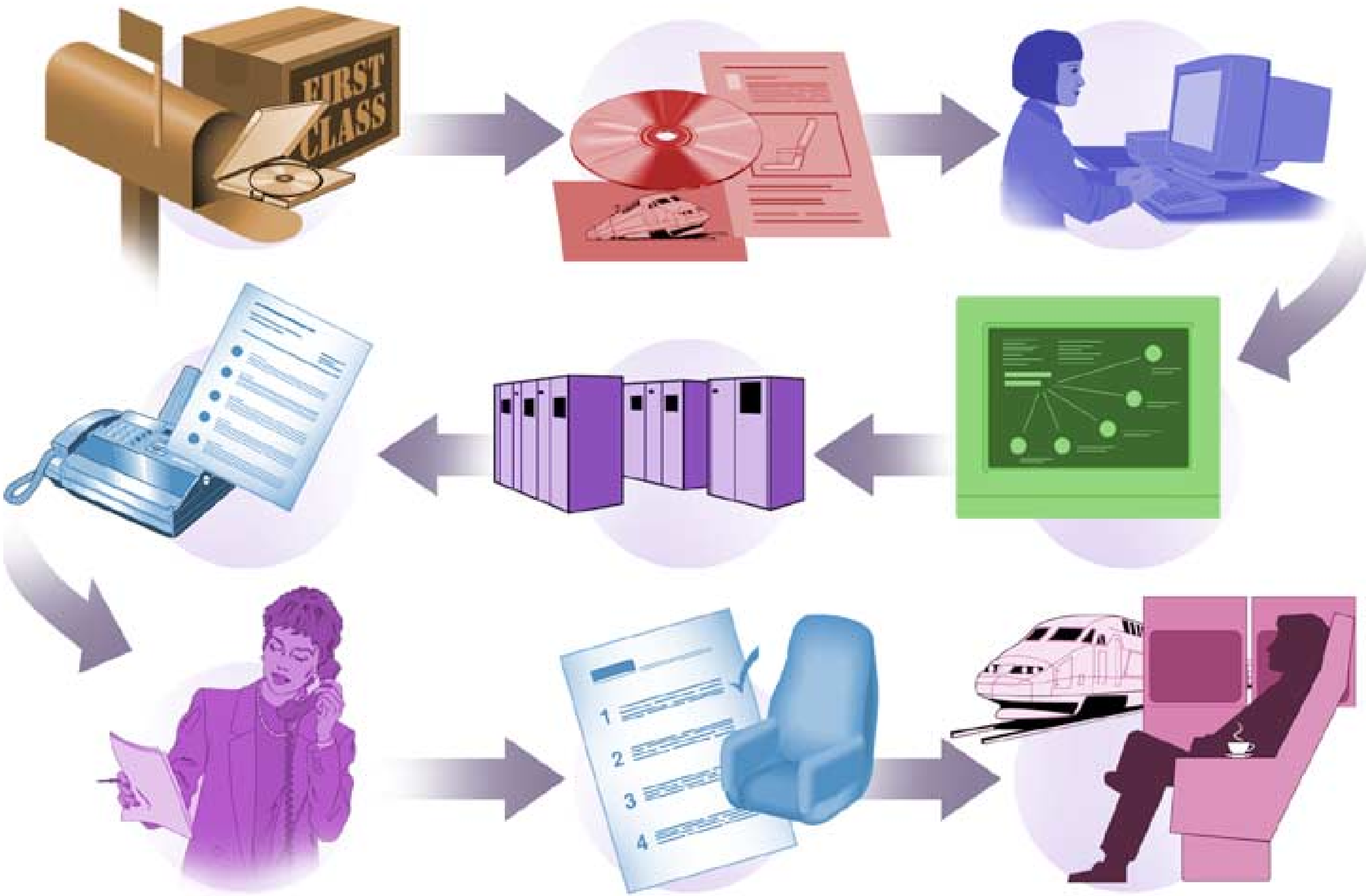
Invention:
Design a New Drug



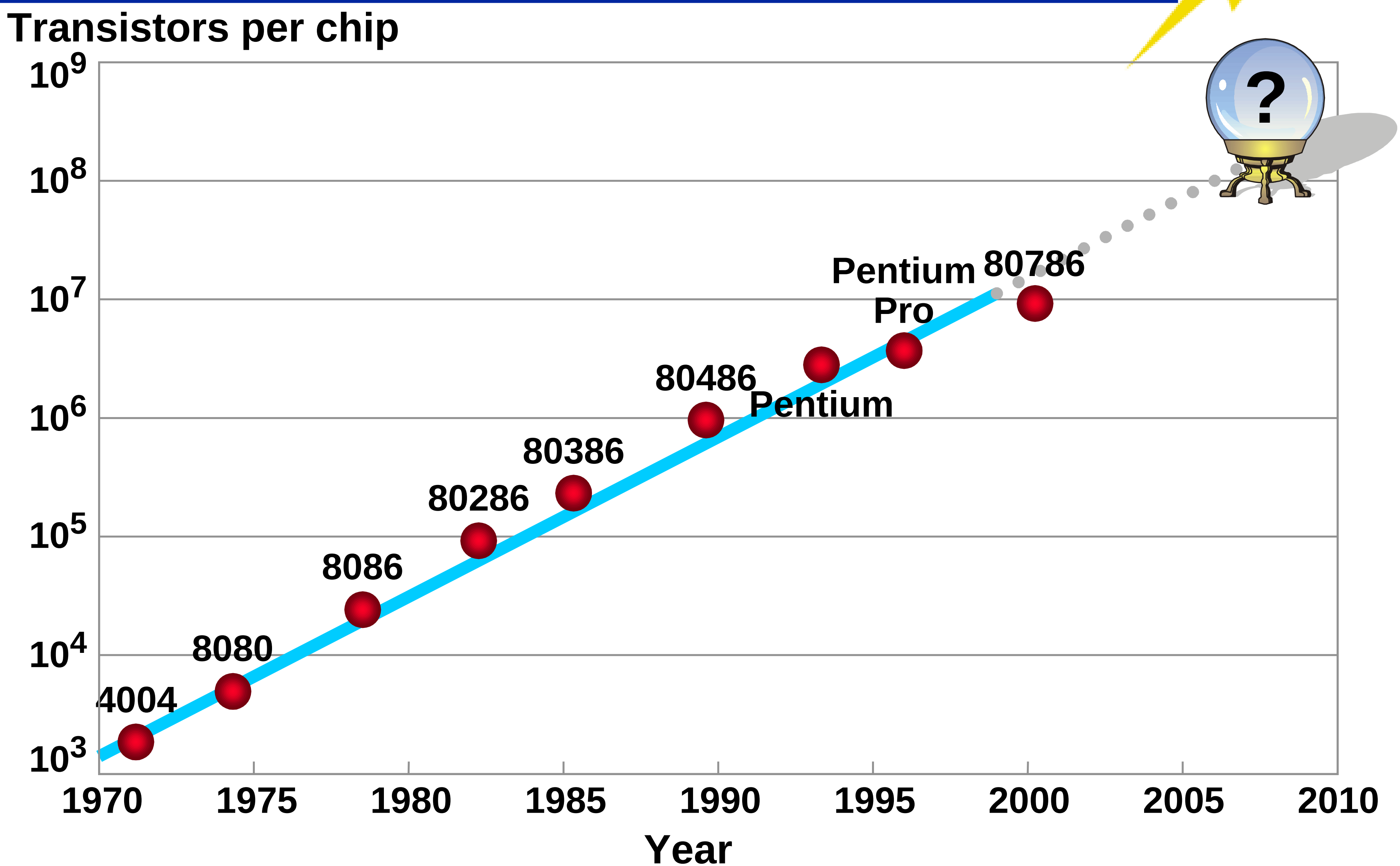
Discovery:
Efficient Compound Mixing



Design:
Seat Frame Innovation

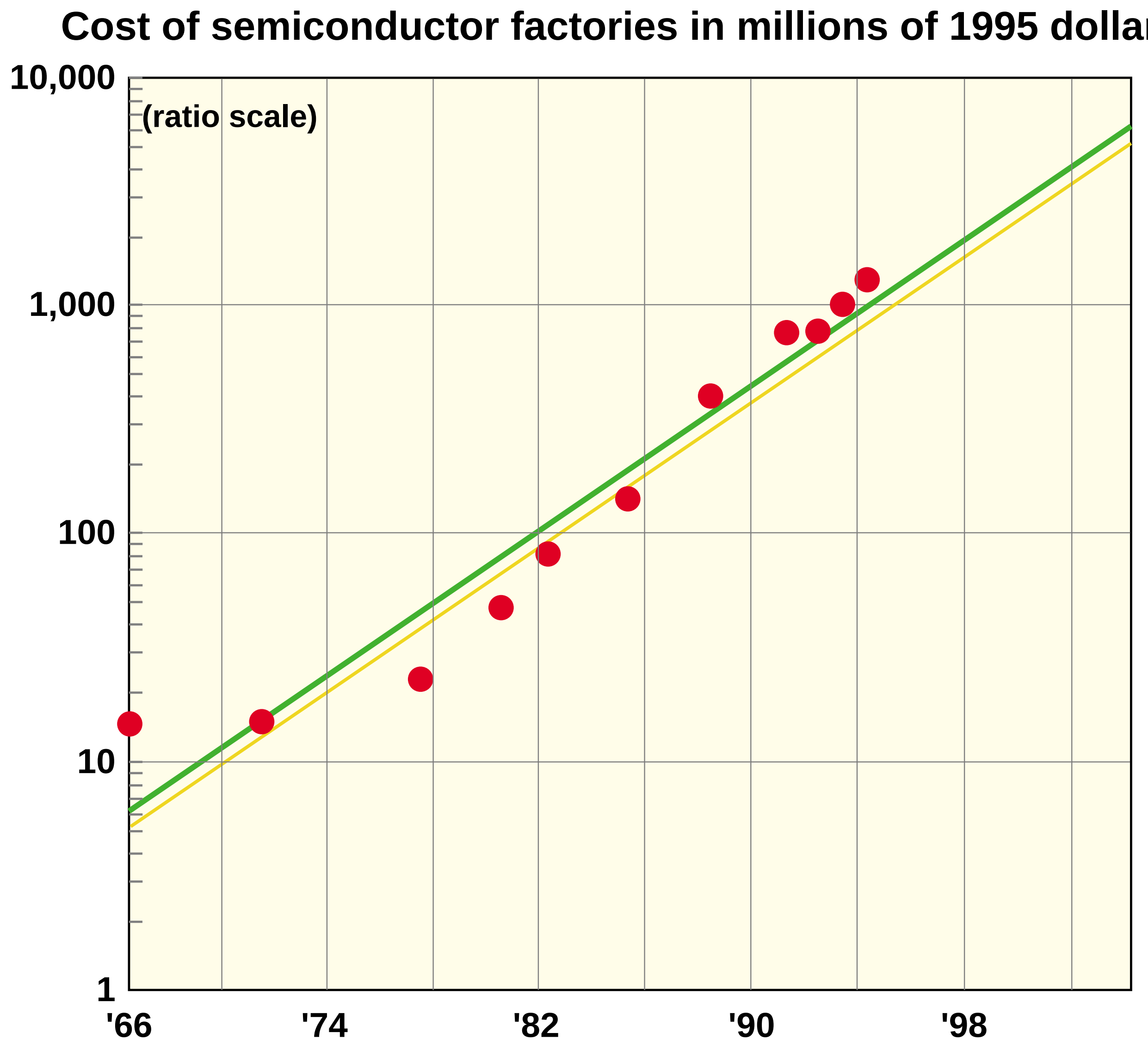


Moore's Law



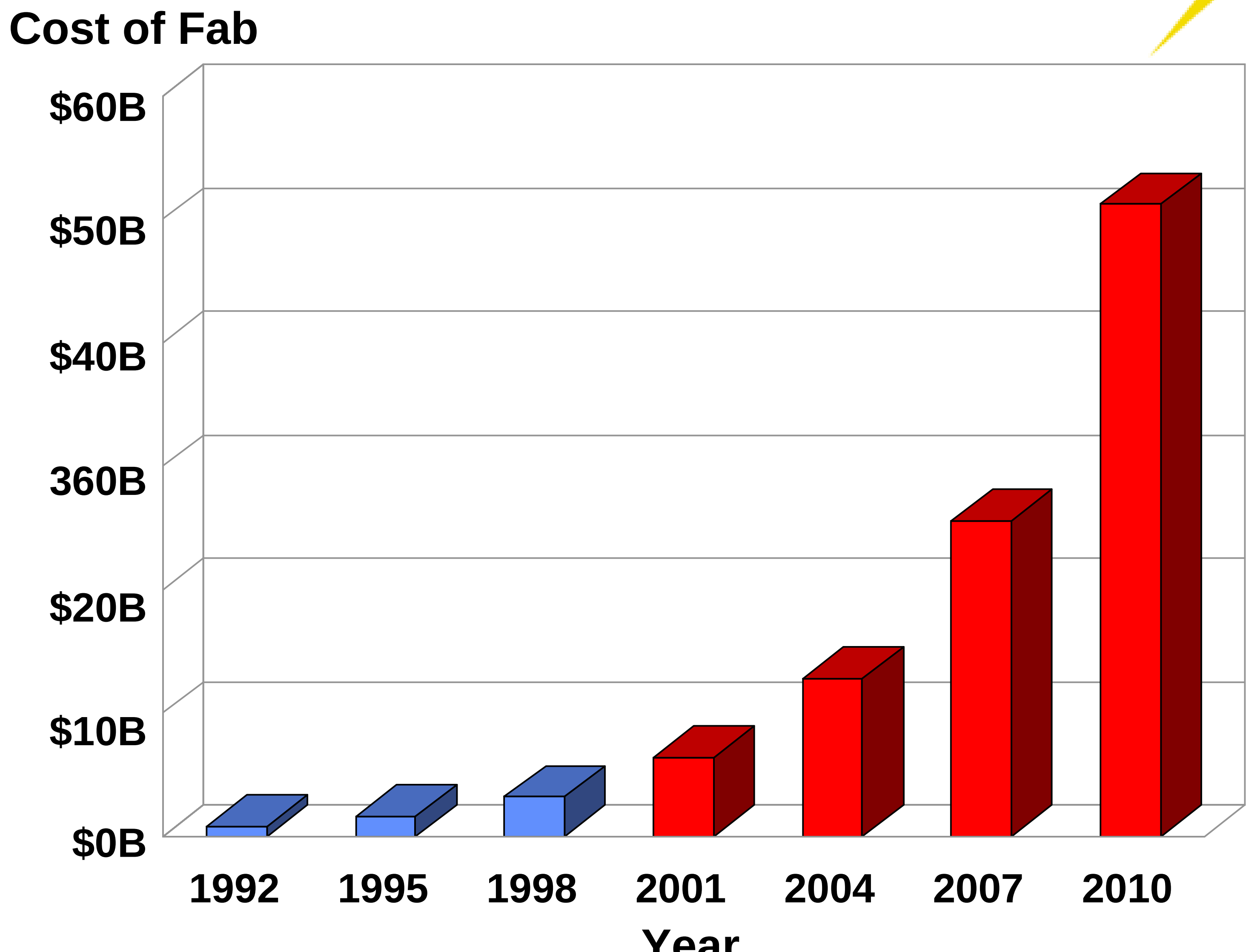
Source: Joel Birnbaum, HP, Lecture at APS Centennial, Atlanta, 1999

Moore’s Second Law (Rock’s Law)



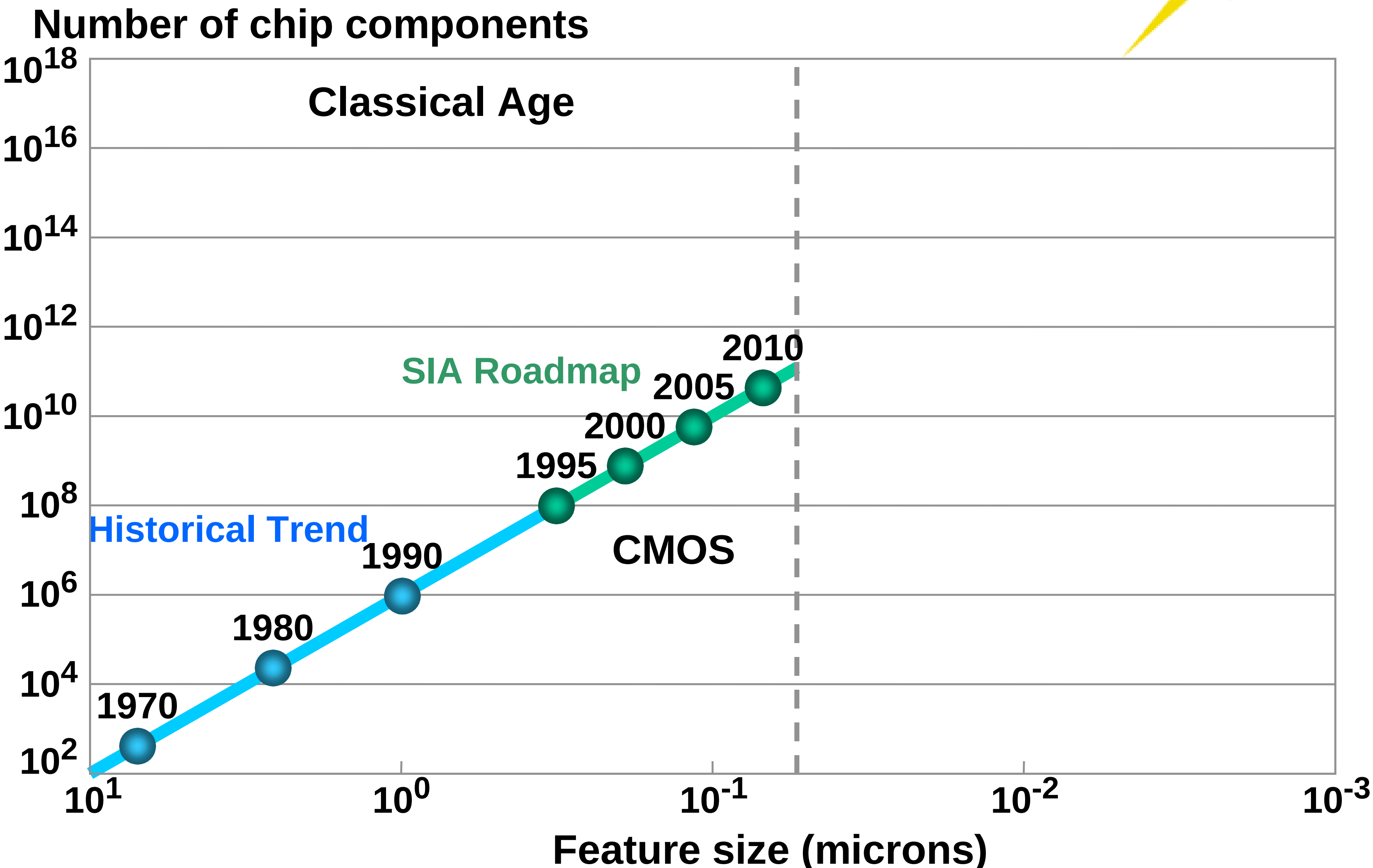
Source: Forbes Magazine

Moore’s Second Law



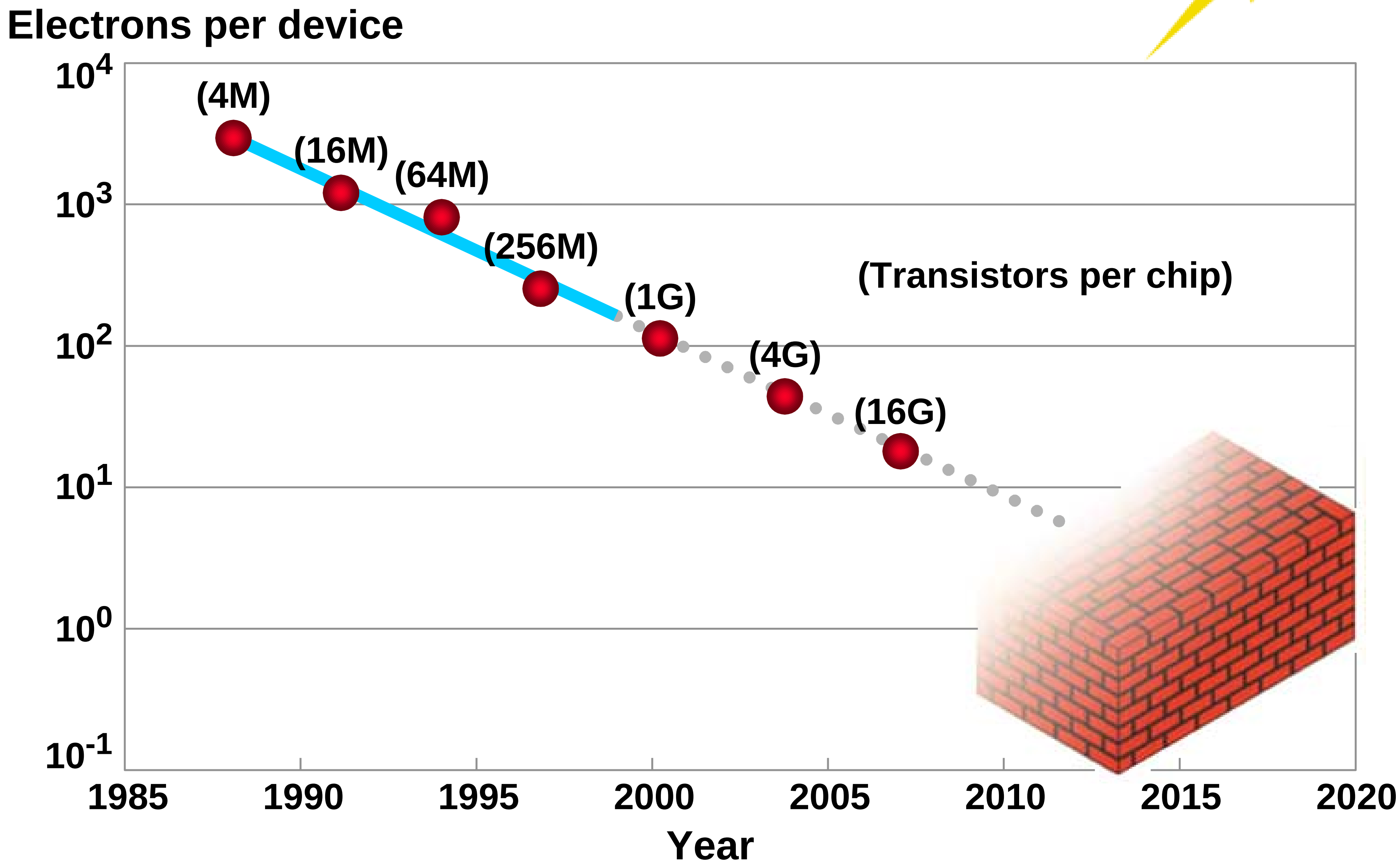
Source: Joel Birnbaum, HP, Lecture at APS Centennial, Atlanta, 1999

Scaling of Electronic Devices



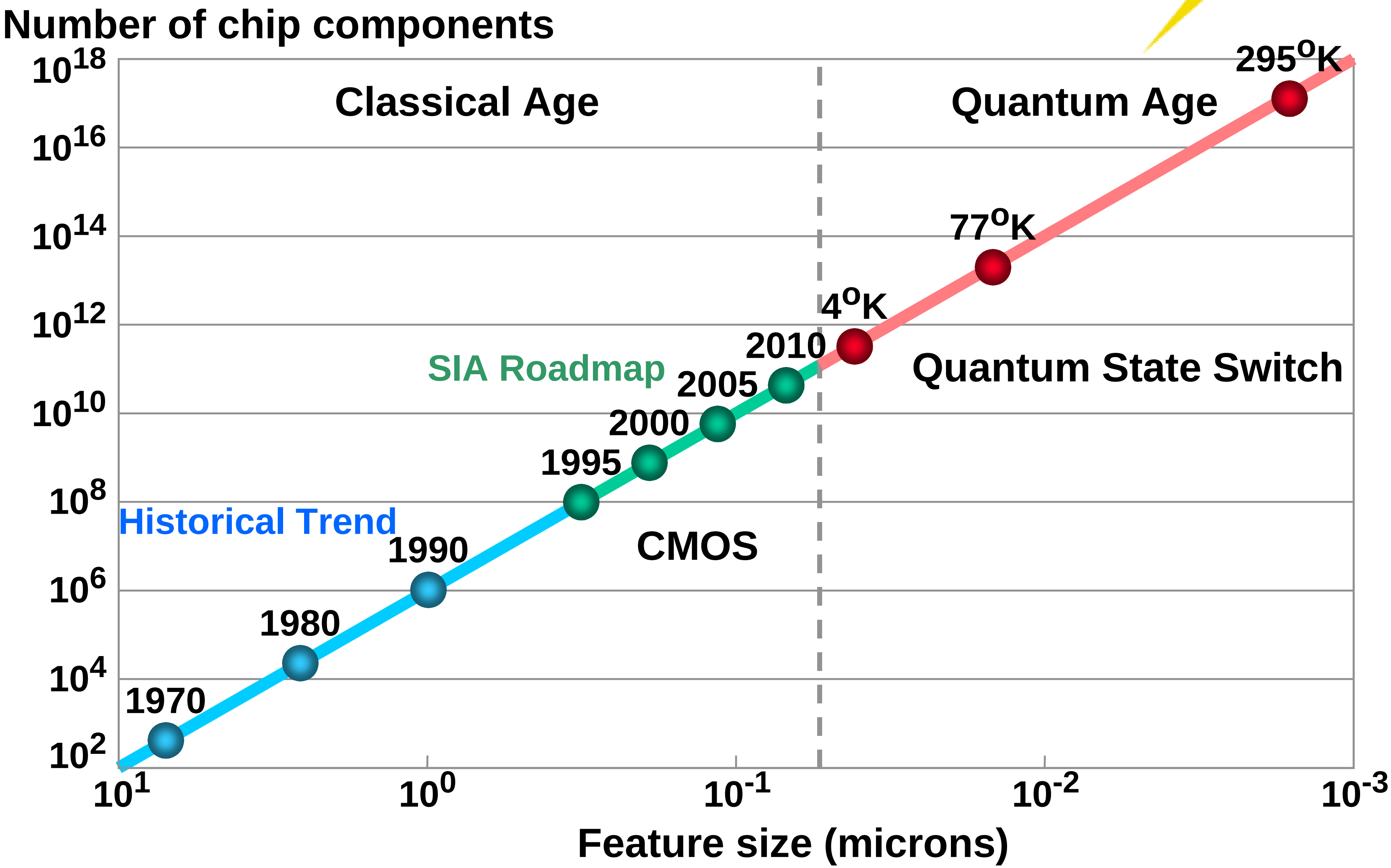
Source: Joel Birnbaum, HP, Lecture at APS Centennial, Atlanta, 1999

Vanishing Electrons



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

Scaling of Electronic Devices



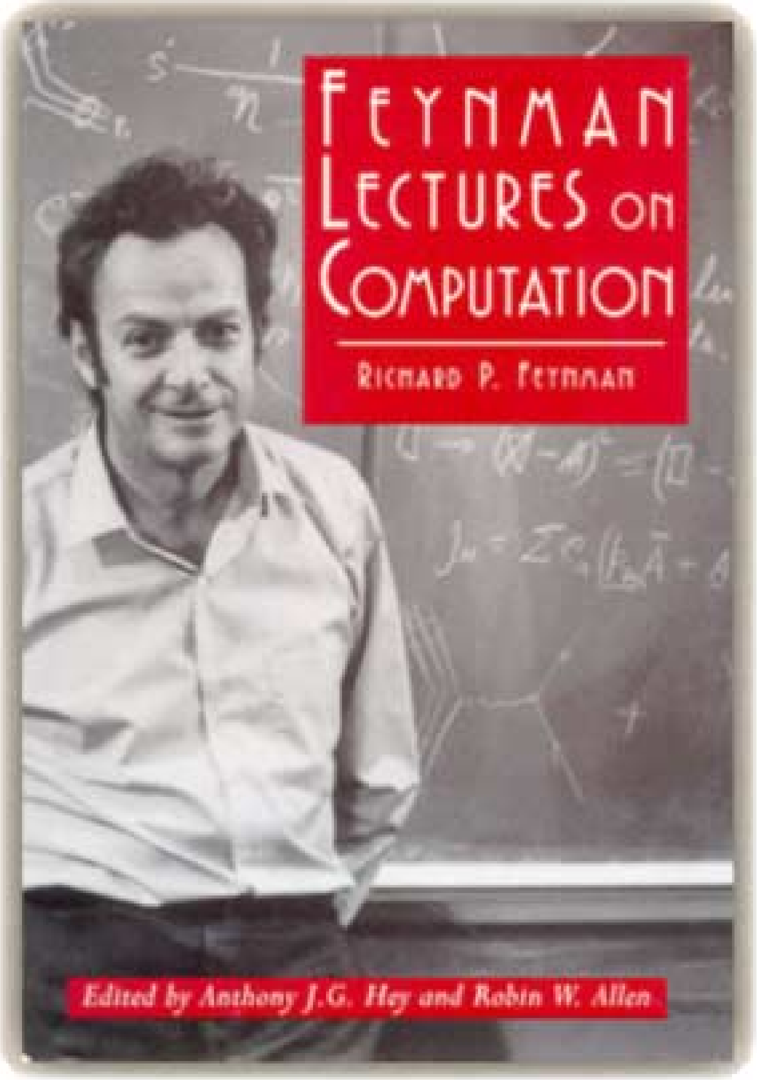
Computation Limit for Nonreversible Logic

- **Assume a power dissipation of 1W at room temperature**

How many bit operations/second can be performed by a nonreversible computer executing Boolean logic?

$$\nu = P/kT\, \mathit{ln}(2) = 3.5 \times 10^{20} \text{ bit ops/sec}$$

Power Cost of Information Transfer?



$$P = nk_B T \frac{d}{c} \nu^2$$

P = power

k_B = Boltzman constant

T = temperature

d = transmission distance

c = speed of light

ν = operating frequency

n = number of parallel operations

Rate of Nonreversible Information Transfer

- **Assume a power dissipation of 1W and a volume of 1cm^3**

How many bits/second can be transferred?

$$\nu = \sqrt{\frac{cP}{k_B T d}} = 10^{18} \text{ ops/sec}$$

This is roughly the equivalent of 10^9 Pentiums!

Other Possibilities?

Molecular nanomechanics:

- **DNA, mechanical, chemical, biological**

Quantum cellular automata:

- **Arrays of quantum dots**

Molecular nanoelectronics:

- **Chemically-synthesized circuits**

Will History Repeat Itself?

	1939	1999
Technology engine	Vacuum tube	CMOS FET
Disruptive technology	Solid state switch	Quantum state switch?
Fundamental research	Purity of materials	Size & shape of materials
Impact	Demise of vacuum tubes	Demise of semiconductors

Thinking About 2025

- Extrapolation
- “Reading the Clearing” (Denning)
- Scenario Planning
- Science Fiction and Wishful Thinking

Extrapolation: The Long Boom

- **Peter Schwartz and Peter Leyden, Wired, July 1997**
- **Global economic boom of unprecedented scale**
- **Continued sustained economic growth**
- **Managing ecological problems**
- **Globalization and openness**
- **Five waves of technology (computers, telecommunication, biotech, nanotechnology, alternative energy)**

Reading the Clearing: J. Coates, The Highly Probable Future circa 2025

- 8.4 B, English speaking, personally tagged & identified, prosthetic assisted and/or mutant, tense people who have access & control of their medical records
- Everything will be smart, responsive to environment.
  - Sensing of everything… challenge for science & engineering!
  - Fast broadband network
  - Smart appliances & AI
  - Tele-all: shop, vote, meet, work, etc.
  - Robots do everything, *but there may be conflict with labor…*
- A "managed", physical and man-made world
  - Reliable weather reports
  - *"Many natural disasters e.g. floods, earthquakes, will be mitigated, controlled or prevented"*
- No surprises. We can see 10 years, but not 20!

Source: Gordon Bell and J. Coates, Futurist, Vol. 84, 1994

Scenario Planning: Air Force 2025

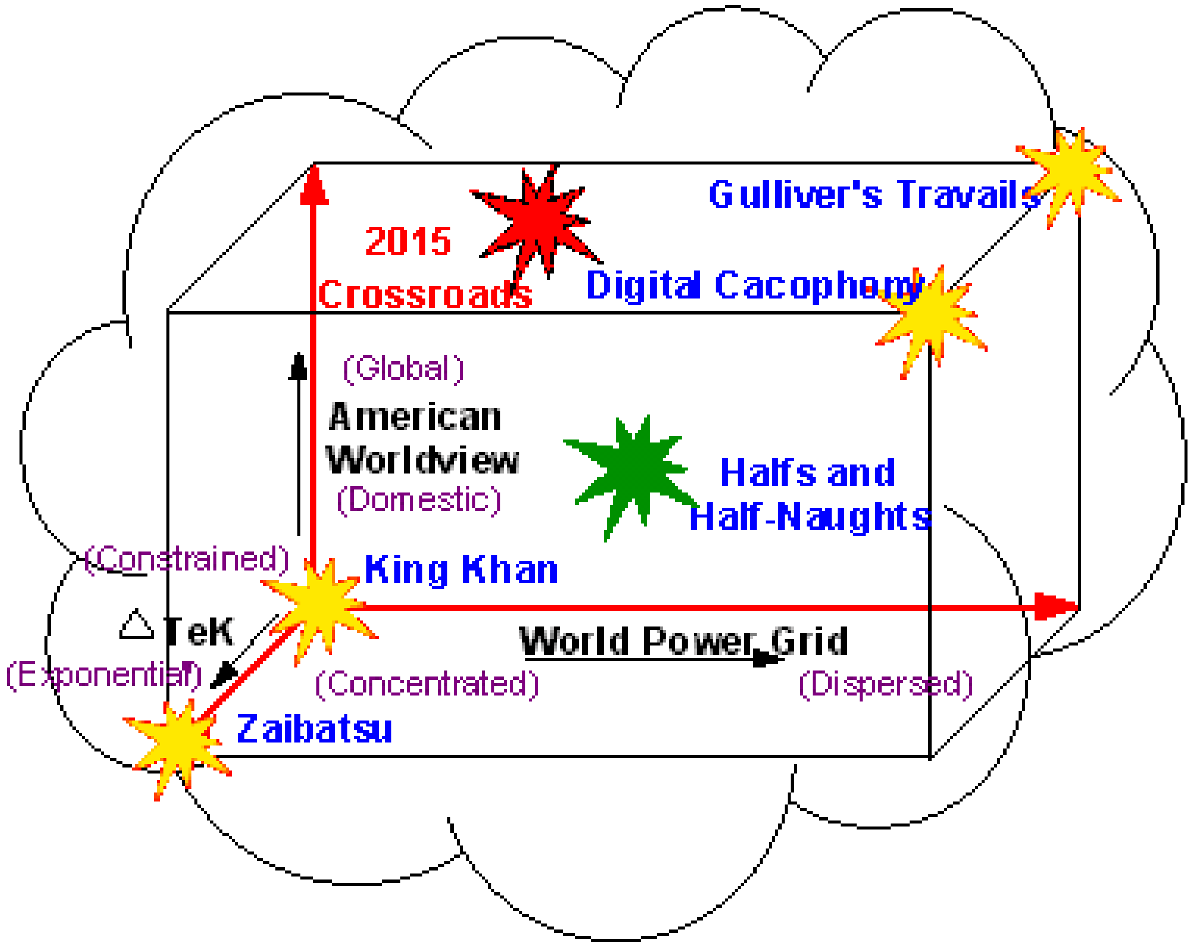
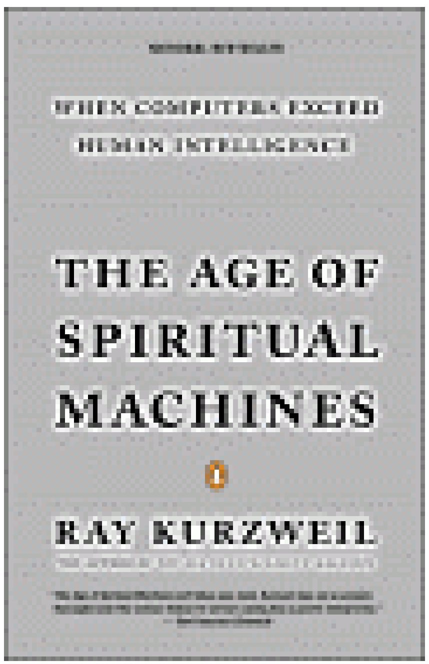


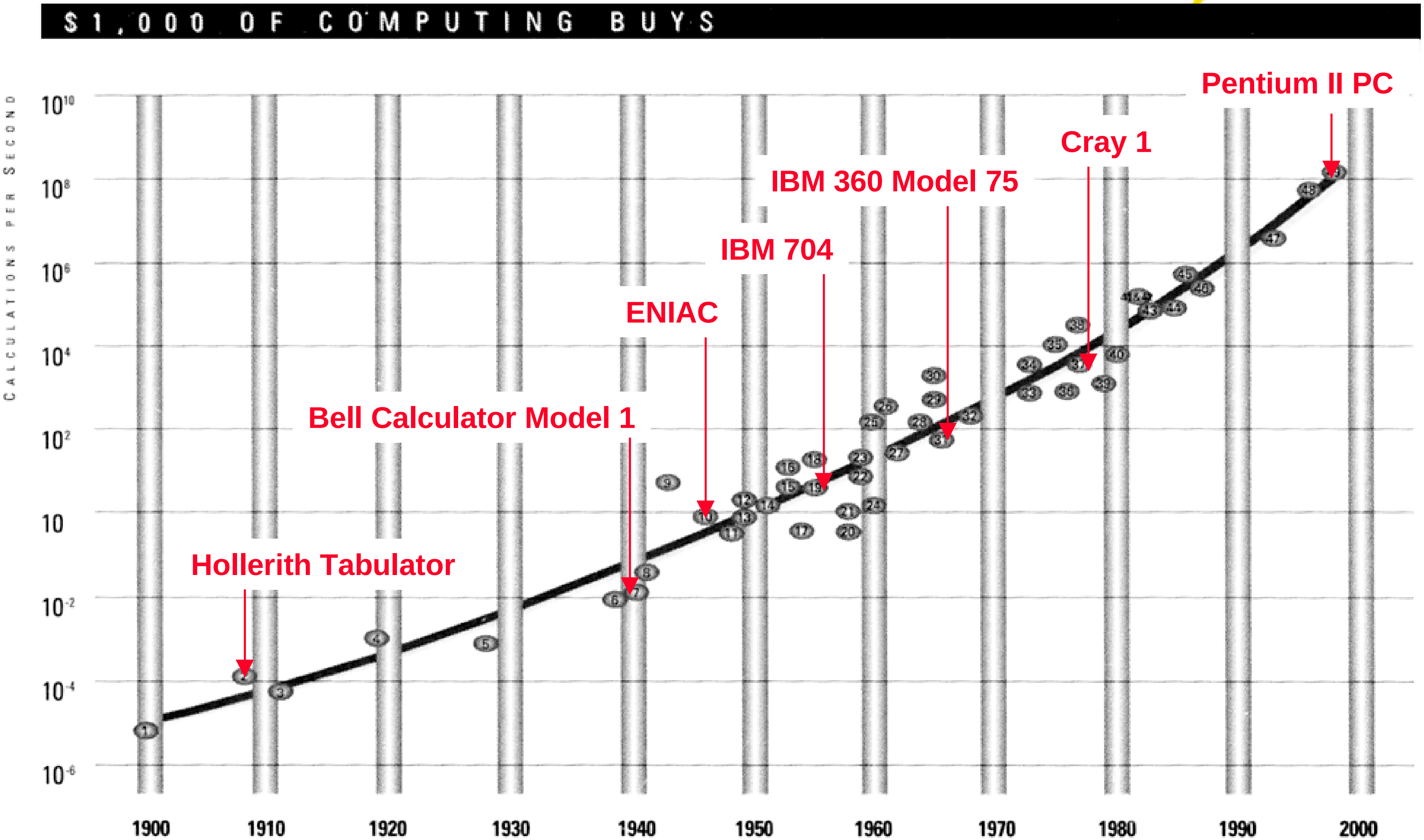
Figure ES-1. Strategic Planning Space

Science Fiction and Wishful Thinking

- **R. Kurzweil, *The Art of Spiritual Machines***
- **Bill Joy, *Why the Future Does Not Need Us*, Wired March 2000**

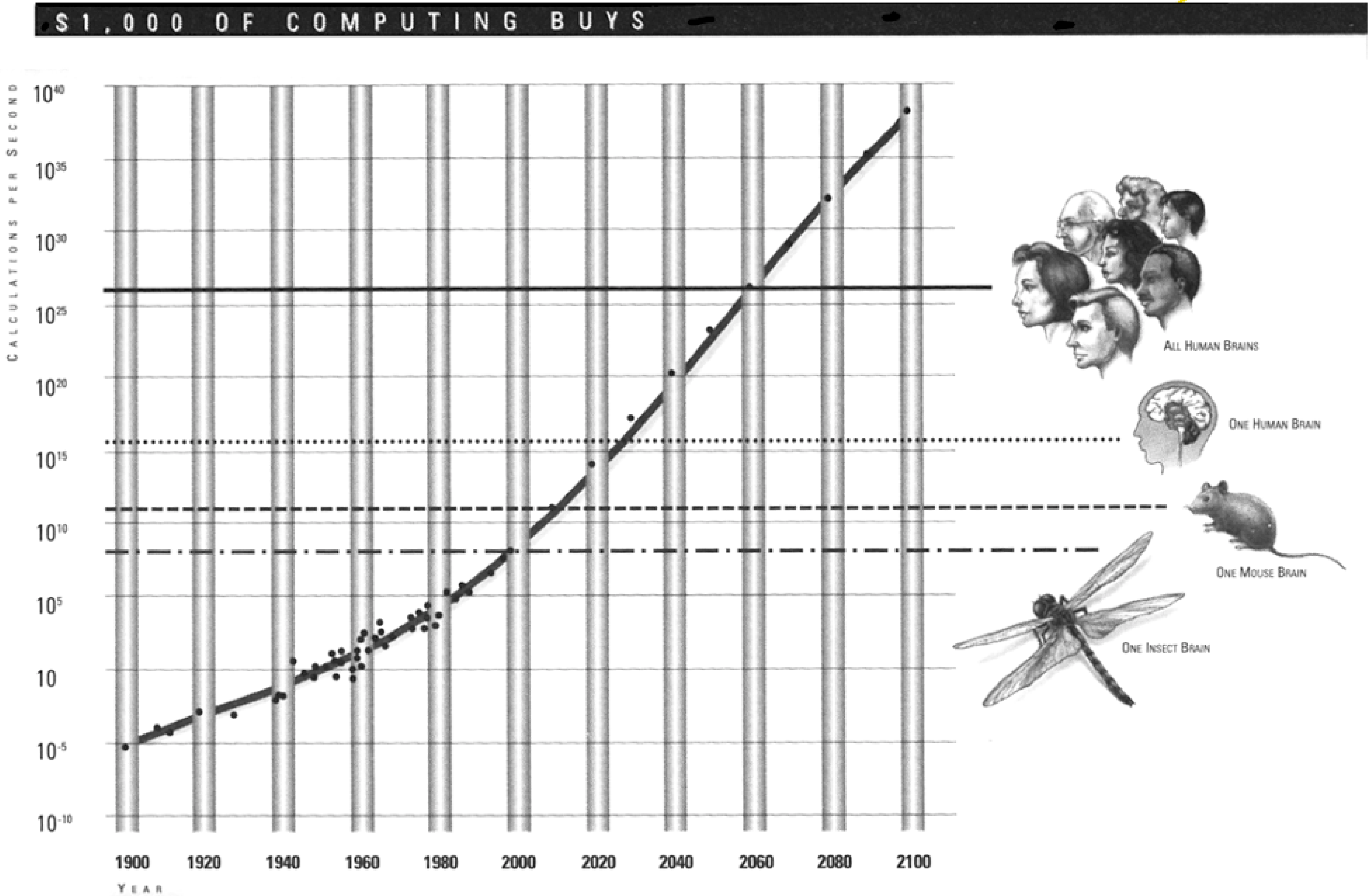


The Exponential Growth of Computing, 1900-1998



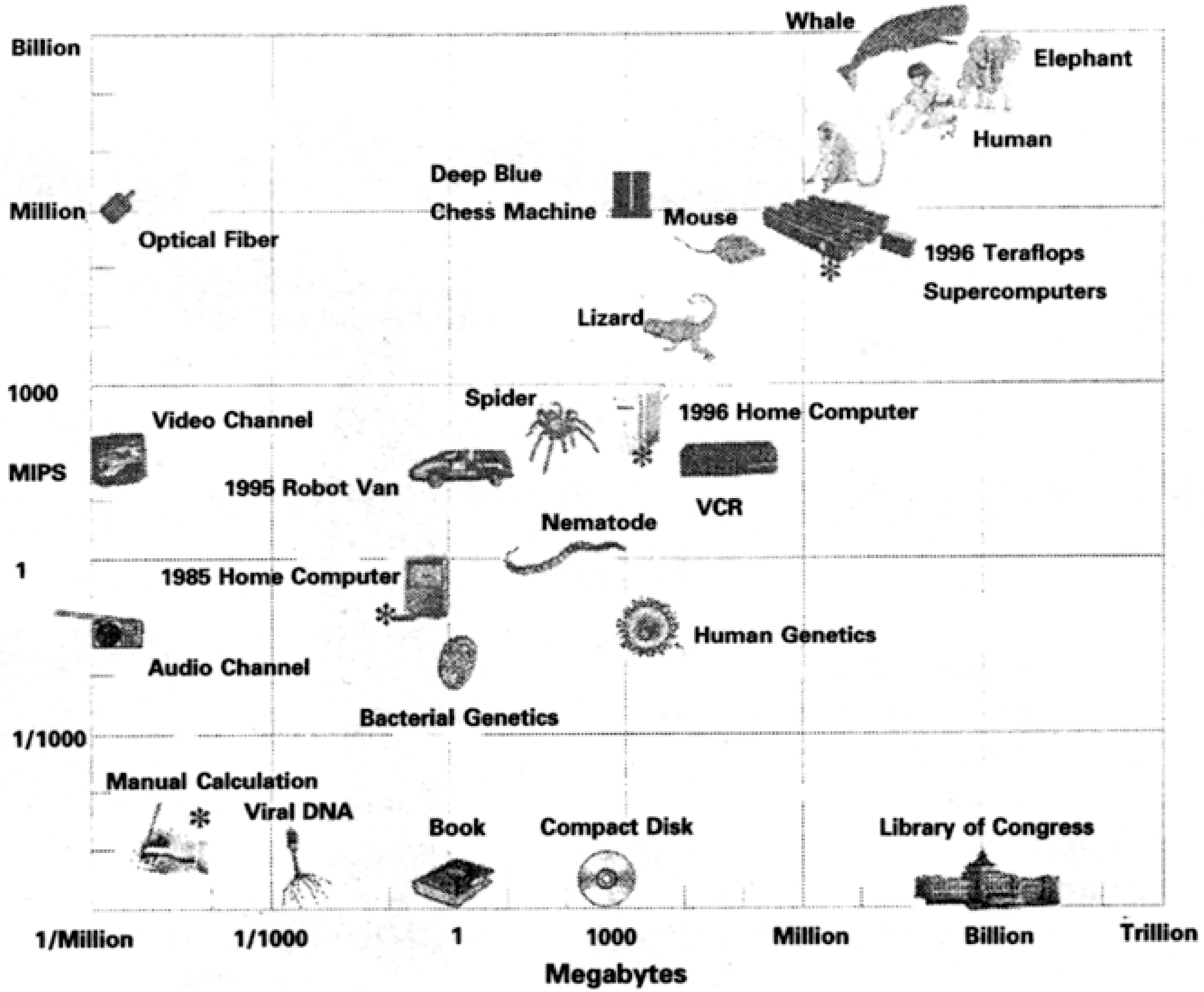
Adapted from Kurzweil, *The Age of Spiritual Machines*

The Exponential Growth of Computing, 1900-2100



Adapted from Kurzweil, *The Age of Spiritual Machines*

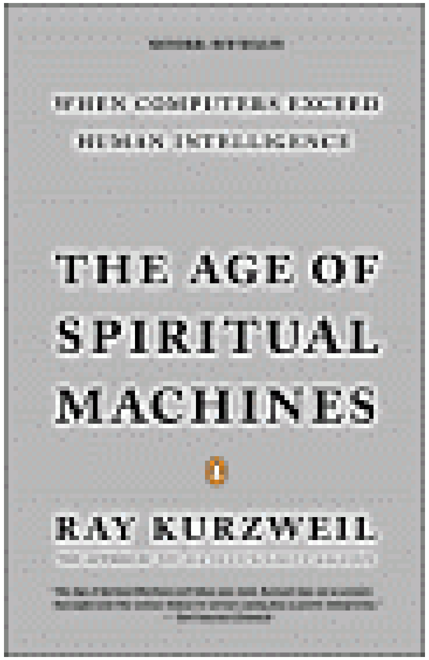
Megabytes-per-MIPS

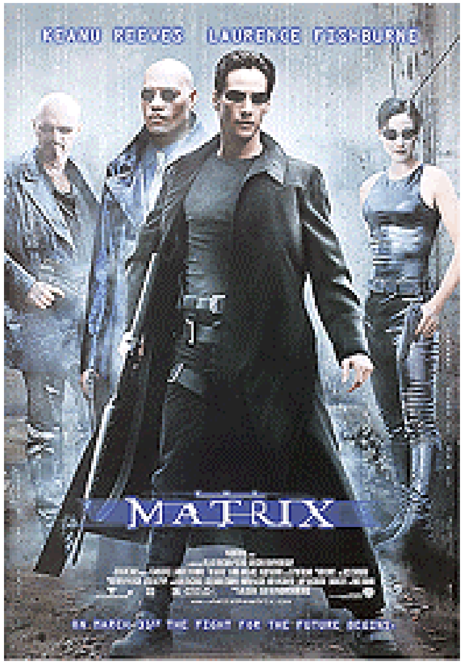


Adapted from Moravec, *Progress in Robotics: An Acclerated Rerun Of Natural History?*

Science Fiction and Wishful Thinking

- **The meme of "powerful computers controlling the future of mankind" has been with us for about a decade**

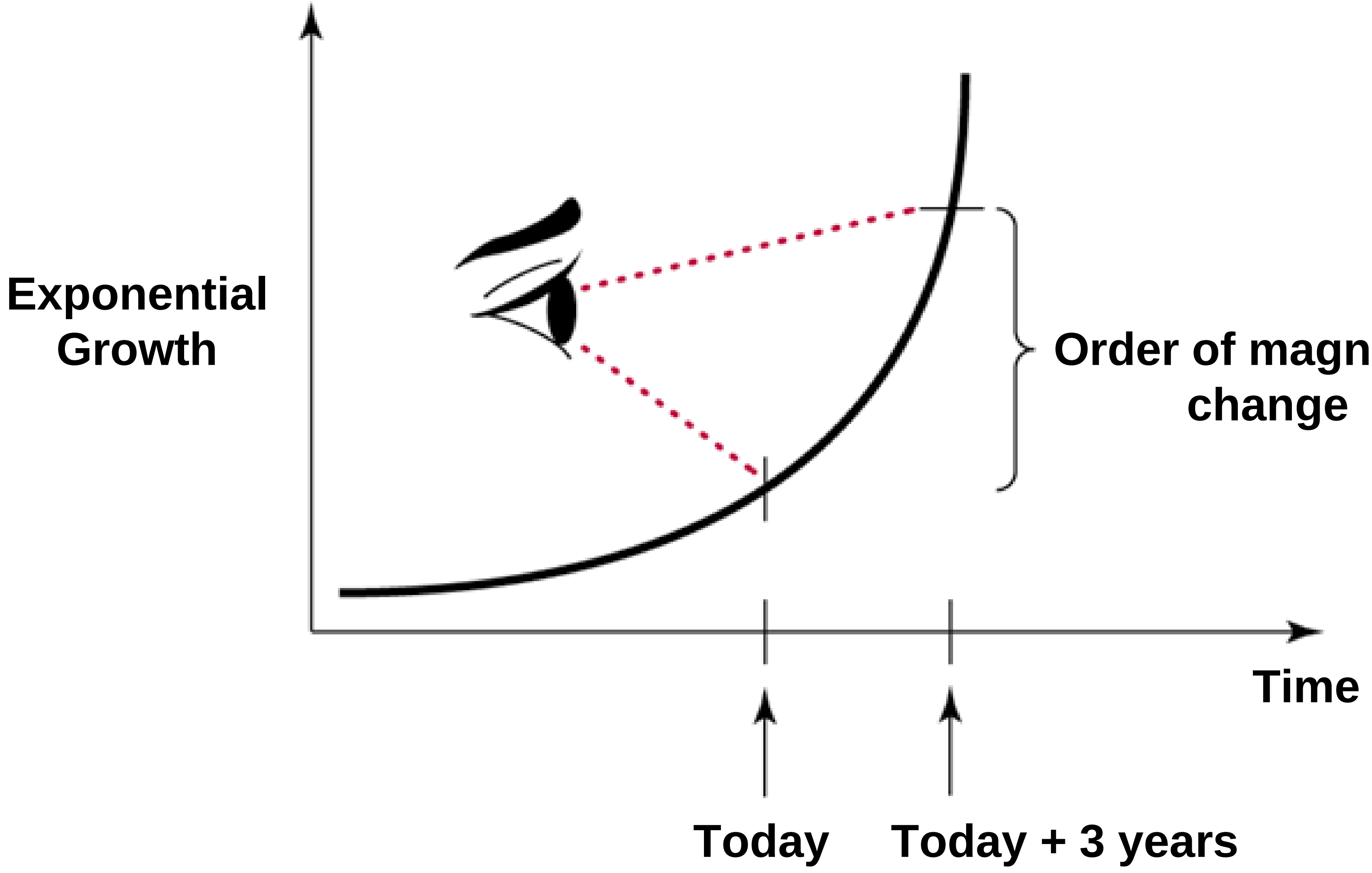




Why I don't share this vision

- 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
- past experience: people bring technology into their lives, not vice versa
  — the slow rate of change of human behavior and processes will work against this change

Moore's Law — The Exponential View



Moore's Wall — The Real (Exponential) View

What am I willing to predict?

2025:

- Moore’s Law for semiconductors will have ended
- There is room for a replacement technology
- Robots and intelligent machines won’t eat our lunch