***Seven* Paths to High Performance** (x)

Ari Patrinos: “we don’t want to just beat the Japanese, we want to leapfrog”

DOE Science Computing Conference
Washington, DC
19 June 2003

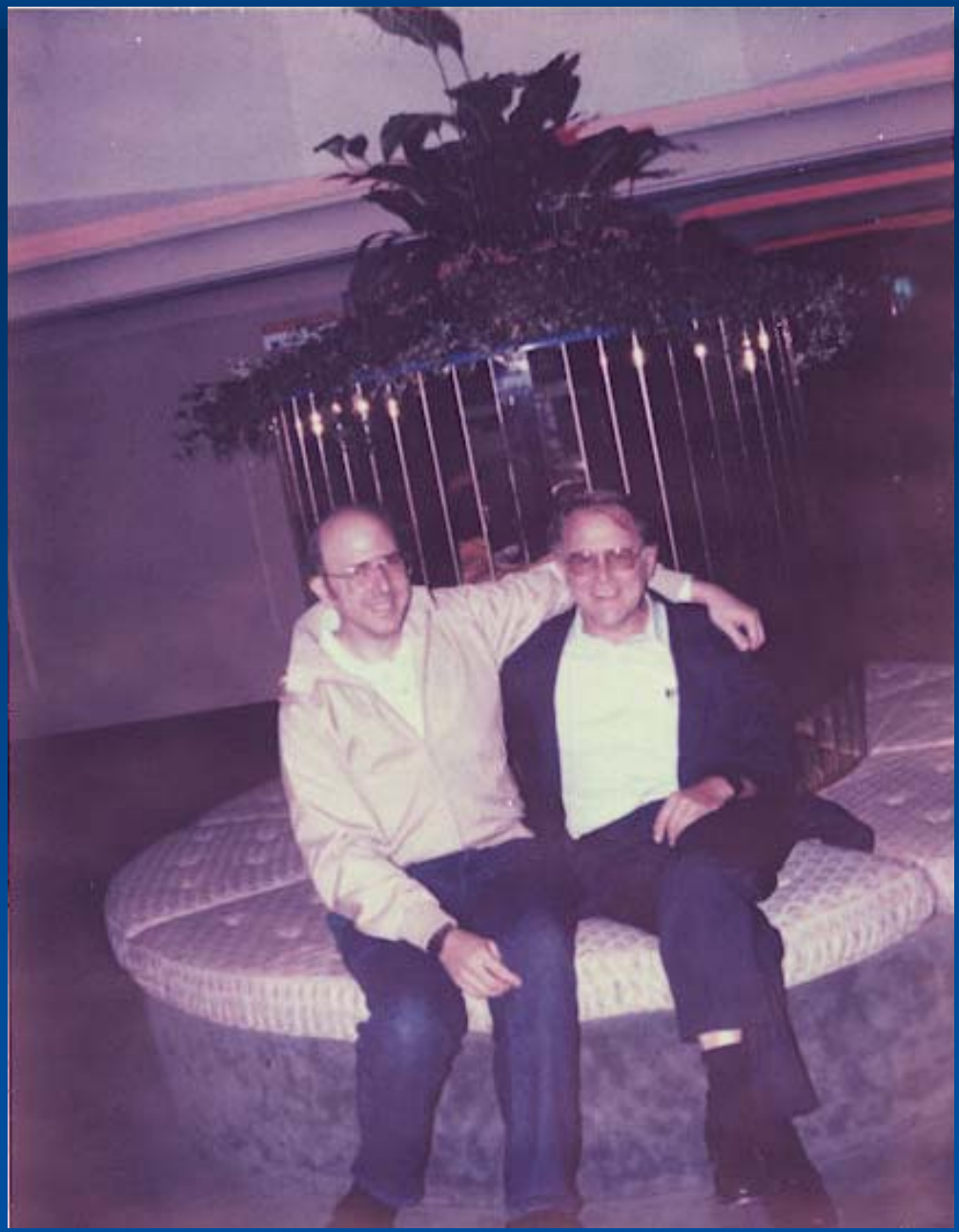
Gordon Bell
Microsoft Research

The Paths

- DARPA's HPC for National Security …5 >3 > 2
  - Standards paradox: the greater the architectural diversity, the less the learning and program market size
- COTS evolution…if only we could interconnect and cool them, so that we can try to program it
- Terror Grid– the same research community that promised a clusters programming environment
- Response to Japanese with another program
- …and then a miracle happens

A brief, simplified history of HPC

1. ***Cray* formula evolves *smPv* for *FORTRAN.* 60-02 (US:60-90)**
2. **1978: VAXen threaten computer centers…**
3. **1982 NSF response: Lax Report. Create 7-Cray centers**
4. **1982: The Japanese are coming with the 5th AI Generation**
5. **DARPA SCI response: search for parallelism w/scalables**
6. **Scalability is found: "bet the farm" on micros clusters**
   - **Beowulf standard forms. (In spite of funders.)>1995**
   - **"Do-it-yourself" Beowulfs negate computer centers since everything is a cluster enabling "do-it-yourself" centers! >2000.**
   - **Result >95 : EVERYONE needs to re-write codes!!**
7. **DOE's ASCI: petaflops clusters => "arms" race continues!**
8. ***2002: The Japanese came! Just like they said in 1997***
9. ***2002 HPC for National Security response: 5 bets &7 years***
10. ***Next Japanese effort? Evolve? (Especially software)***
    red herrings or hearings
11. **1997: High speed nets enable peer2peer & Grid or Teragrid**
12. ***2003 Atkins Report-- Spend $1.1B/year, form more and larger centers and connect them as a single center…***



Steve Squires & Gordon Bell at our "Cray" at the start of DARPA's SCI program c1984.

<u>20 years later: Clusters of Killer micros become the single standard</u>

“In Dec. 1995 computers with 1,000 processors will do most of the scientific processing.”

Danny Hillis
1990 (1 paper *or 1 company*)

PC+

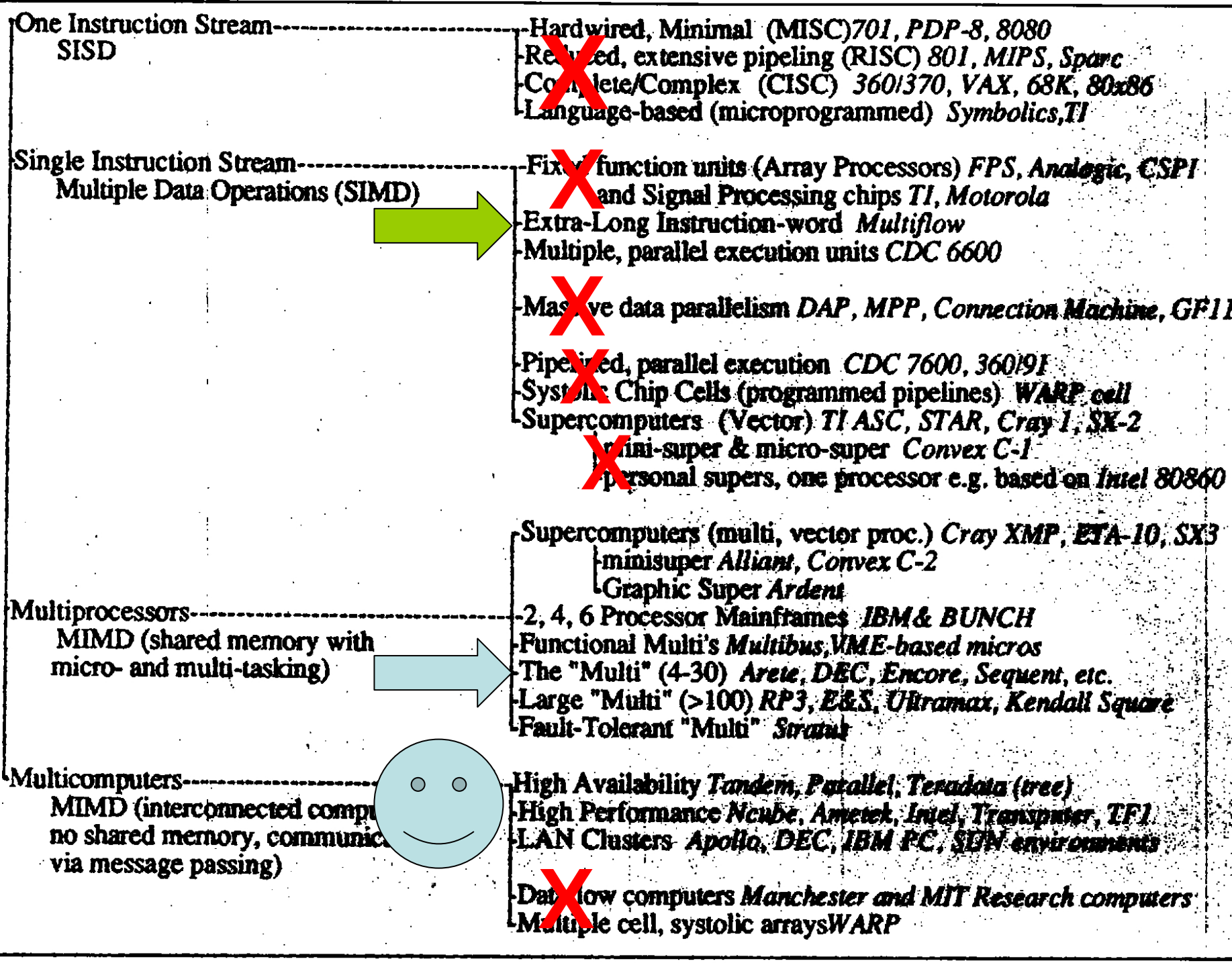
Lost in the search for parallelism

- ACRI
- Alliant
- American Supercomputer
- Ametek
- Applied Dynamics
- Astronautics
- BBN
- CDC
- Cogent
- Convex > HP
- Cray Computer
- Cray Research > SGI > Cray
- Culler-Harris
- Culler Scientific
- Cydrome
- Dana/Ardent/Stellar/Stardent
- Denelcor
- Encore
- Elexsi
- ETA Systems
- Evans and Sutherland Computer
- Exa
- Flexible
- Floating Point Systems
- Galaxy YH-1

- Goodyear Aerospace MPP
- Gould NPL
- Guiltech
- Intel Scientific Computers
- International Parallel Machines
- Kendall Square Research
- Key Computer Laboratories *searching again*
- MasPar
- Meiko
- Multiflow
- Myrias
- Numerix
- Pixar
- Parsytec
- nCube
- Prisma
- Pyramid
- Ridge
- Saxpy
- Scientific Computer Systems (SCS)
- Soviet Supercomputers
- Supertek
- Supercomputer Systems
- Suprenum
- Tera > Cray Company
- Thinking Machines
- Vitesse Electronics
- Wavetracer

1987 Interview July 1987 as first CISE AD

- Kicked off parallel processing initiative with 3 paths
  - Vector processing *was totally ignored*
  - Message passing multicomputers including distributed workstations and clusters
  - smPs (multis) -- main line for programmability
  - *SIMDs might be low-hanging fruit*
- Kicked off Gordon Bell Prize
- Goal: common applications parallelism
  - 10x by 1992; 100x by 1997


One Instruction Stream
SISD
Hardwired, Minimal (MISC)701, PDP-8, 8080
Reduced, extensive pipeling (RISC) 801, MIPS, Sparc
Complete/Complex (CISC) 360/370, VAX, 68K, 80x86
Language-based (microprogrammed) Symbolics,TI
Single Instruction Stream
Multiple Data Operations (SIMD)
Fixed function units (Array Processors) FPS, Analogic, CSPI
and Signal Processing chips TI, Motorola
Extra-Long Instruction-word Multiflow
Multiple, parallel execution units CDC 6600
Massive data parallelism DAP, MPP, Connection Machine, GF11
Pipelined, parallel execution CDC 7600, 360/91
Systolic Chip Cells (programmed pipelines) WARP cell
Supercomputers (Vector) TI ASC, STAR, Cray 1, SX-2
mini-super & micro-super Convex C-1
personal supers, one processor e.g. based on Intel 80860
Multiprocessors
MIMD (shared memory with
micro- and multi-tasking)
Supercomputers (multi, vector proc.) Cray XMP, ETA-10, SX3
minisuper Alliant, Convex C-2
Graphic Super Ardent
2, 4, 6 Processor Mainframes IBM& BUNCH
Functional Multi's Multibus,VME-based micros
The "Multi" (4-30) Arete, DEC, Encore, Sequent, etc.
Large "Multi" (>100) RP3, E&S, Ultramax, Kendall Square
Fault-Tolerant "Multi" Stratus
Multicomputers
MIMD (interconnected comput
no shared memory, communic
via message passing)
High Availability Tandem, Patallel, Teradata (tree)
High Performance Ncube, Ametek, Intel, Transputer, TF1
LAN Clusters Apollo, DEC, IBM PC, SUN environments
Data flow computers Manchester and MIT Research computers
Multiple cell, systolic arraysWARP

IEEE Software launches annual Gordon Bell Award

Editor-in-Chief Ted Lewis has announced the First Annual Gordon Bell Award for the most improved speedup for parallel-processing applications. The two $1000 awards will be presented to the person or team that demonstrates the greatest speedup on a multiple-instruction, multiple-data parallel processor.

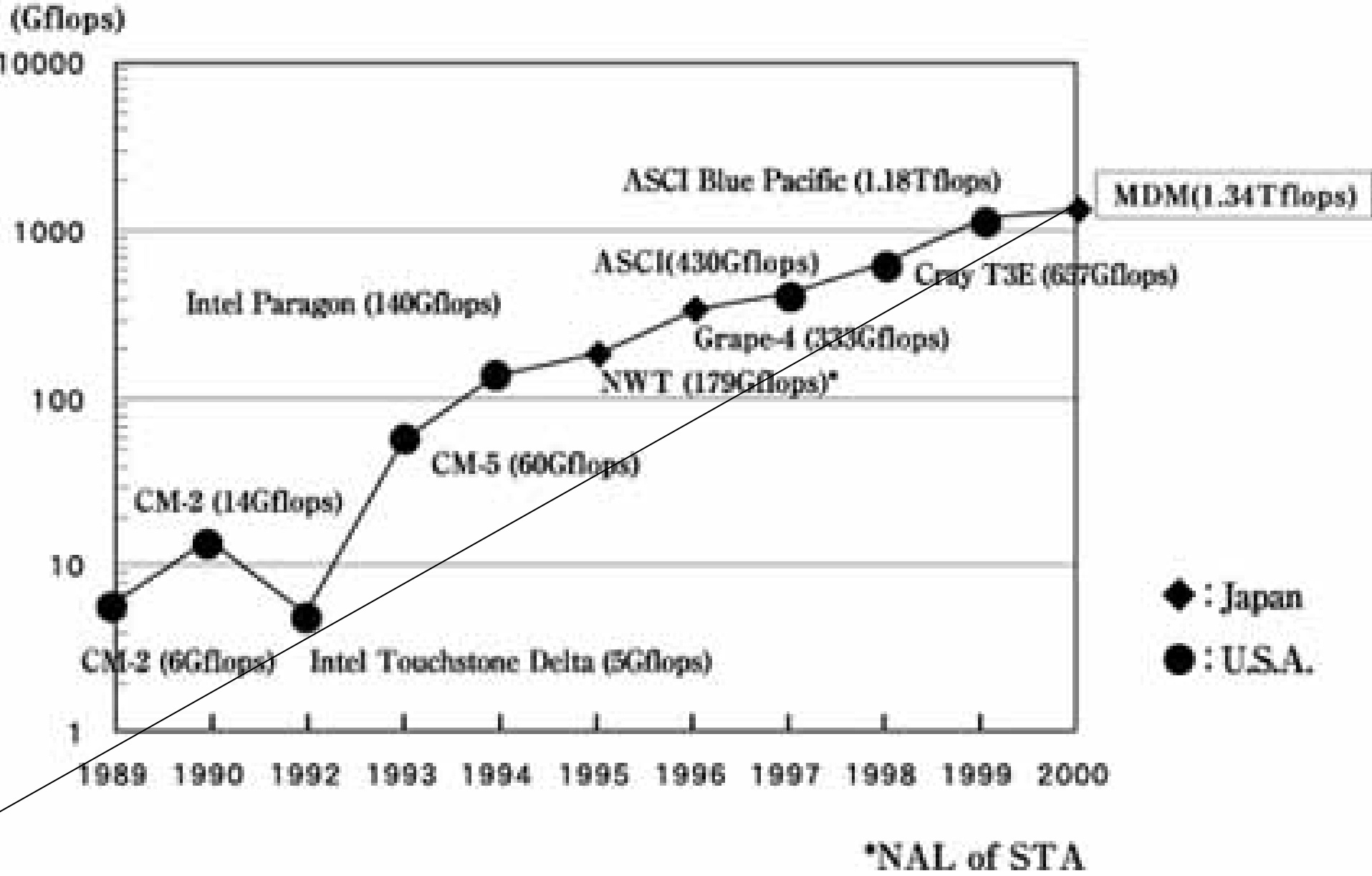
One award will be for most speedup on a general-purpose (multiapplication) MIMD processor, the other for most speedup on a special-purpose MIMD processor. Speedup can be accomplished by hardware or software improvements, or by a combination of the two.

To qualify for the 1987 awards, candidates must submit documentation of their results by Dec. 1. The winners will be announced in the March 1988 issue. This year's judges are Alan Karp of IBM's Palo Alto Scientific Center, Jack Dongarra of Argonne National Laboratory, and Ken Kennedy of Rice University.

For a complete set of rules, definitions, and submission guidelines, write to the Gordon Bell Award, *IEEE Software*, 10662 Los Vaqueros Cir., Los Alamitos, CA 90720.

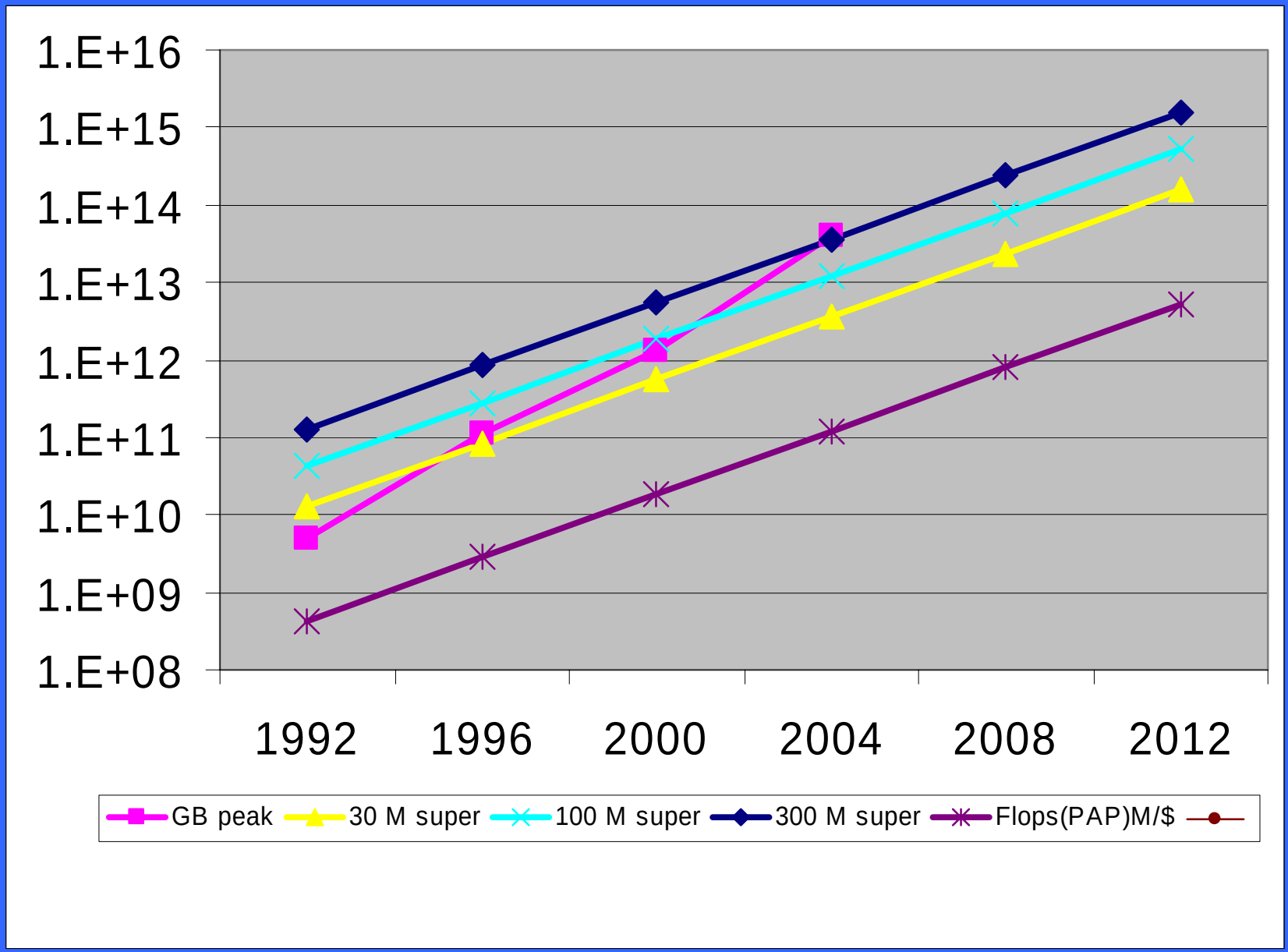
Gordon Bell Prize announced Computer July 1987

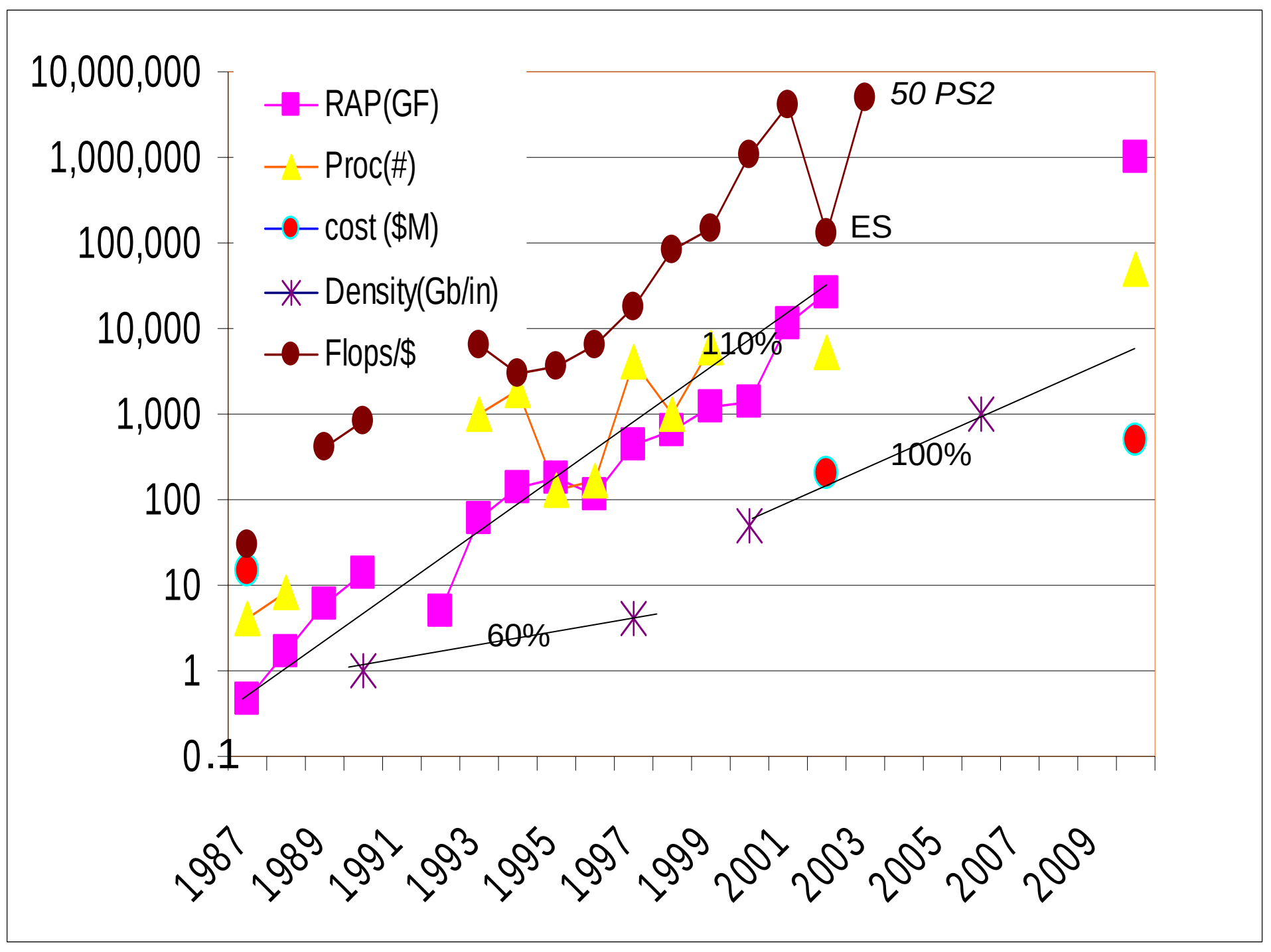
Trend of computing speed at Gordon Bell Prizes



*NAL of STA

Perf (PAP) = c x $s x 1.6**(t-1992); *c = 128 GF/$300M*
'94 prediction: c = 128 GF/$30M




10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
0.1
RAP(GF)
Proc(#)
cost ($M)
Density(Gb/in)
Flops/$
50 PS2
ES
110%
100%
60%
1987
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009

1987-2002 Bell Prize Performance Gain

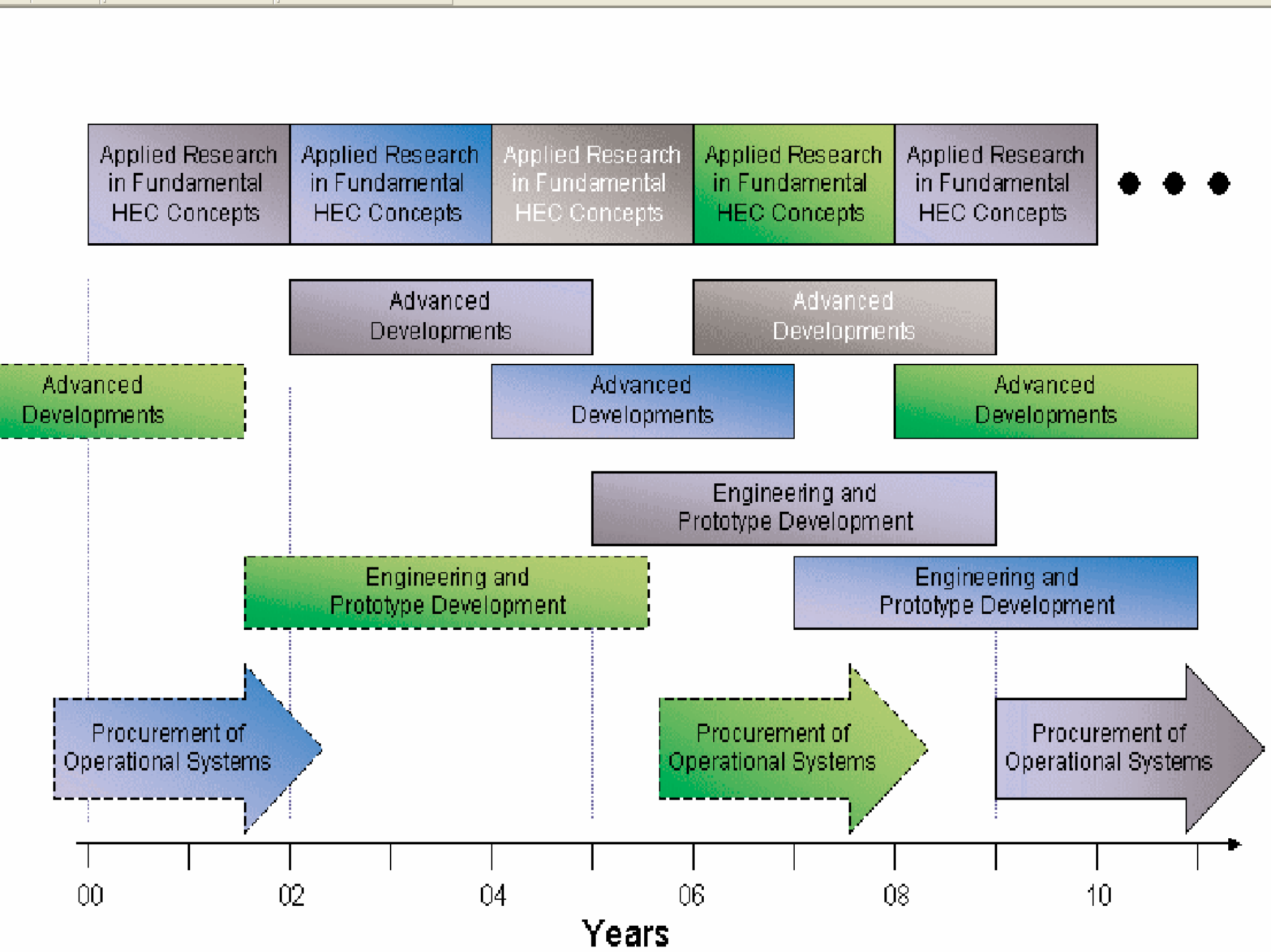
- **26.58TF/0.000450TF = 59,000 in 15 years**
  = 2.08^{15}
- **Cost increase $15 M >> $300 M? *say 20x***
- **Inflation was 1.57 X, so**
  effective spending increase 20/1.57 =12.73
- **59,000/12.73 = 4639 X**
  = 1.76^{15}
- **Price-performance 89-2002:**
  $2500/MFlops > $0.25/MFlops = 10^4
  = 2.04^{13} ***$1K/4GFlops PC = $0.25/MFlops***

1987-2002 Bell Prize Performance Winners

- **Vector: Cray-XMP, -YMP, CM2* (2), Clustered: CM5, Intel 860 (2), Fujitsu (2), NEC (1) = 10**
- **Cluster of SMP (Constellation): IBM**
- **Cluster, single address, very fast net: Cray T3E**
- **Numa: SGI… good idea, but not universal**
- **Special purpose (2)**
- **No winner: 91**
- **By 1994, all were scalable (Not: Cray-x,y,CM2)**
- **No x86 winners!**

Heuristics

- **Use dense matrices, or almost embarrassingly // apps**
- **Memory BW… you get what you pay for (4-8 Bytes/Flop)**
- **RAP/$ is constant. Cost of memory bandwidth is constant.**
- **Vectors will continue to be an essential ingredient; the low overhead formula to exploit the bandwidth, stupid**
- **Bad ideas: SIMD; Multi-threading tbd**
- **Fast networks or larger memories decrease inefficiency**
- **Specialization really, really pays in performance/price!**
- **2003: 50 Sony workstations @6.5gflops for 50K is good.**
- **COTS aka x86 for Performance/Price BUT <u>not Perf</u>.**
- **Bottom Line: Memory BW, Interconnect BW <>Memory Size, FLOPs,**


Applied Research in Fundamental HEC Concepts
Applied Research in Fundamental HEC Concepts
Applied Research in Fundamental HEC Concepts
Applied Research in Fundamental HEC Concepts
Applied Research in Fundamental HEC Concepts
Advanced Developments
Advanced Developments
Advanced Developments
Advanced Developments
Advanced Developments
Engineering and Prototype Development
Engineering and Prototype Development
Engineering and Prototype Development
Procurement of Operational Systems
Procurement of Operational Systems
Procurement of Operational Systems
00
02
04
06
08
10
Years

Does the schedule make sense?

- Early 90s-97 4 yr. firm proposal
- 1997-2000 3 yr. for SDV/compiler
- 2000-2003 3+ yr. useful system

System Components and Technologies

Logic	Memory	Inter-connects On–Chip	Inter-connects Chip–Chip	Inter-connects Component - Component	Switches	Storage	Packaging	Thermal Manage-ment	Design and Test
Si-CMOS	Si-CMOS	Optical - Free Space, Guided	Optical - Free Space, Guided	Optical– Free Space, Guided, MEMs	Optical	Magnetic	MCM	Liquid-Single Phase, Multi Phase	Physical Design Tools
Si/Ge-HBT	Ferroelectrics	Cu-Low k	Low k	Stacked	S.C.	Optical	Stacking	Air	Layout
SC-RSFQ	Magnetic	RF	RF	RF	Semiconductor	Optical Tape	Wafer Scale	Cryogenic	Design Test
Optical	Optical	Heterogeneous Interconnects	Stacked	Electrical Coax Flat Ribbon	RF	Spintronics	HIST		Components Test
III-V-HBT	III-V	Carbon Nanotubes	Wafer Scale	Pliable		Scanning Probe	Boards & Modules		
Spintronics	Spintronics	Nano-electronics					Assembly		
Nano-electronics	Nano-electronics						Integrated Interconnect Technology		
S.C./Si-CMOS	S.C./Si-CMOS						Pliable Interconnects		

Potential <10 years (white)	*Potential in 10+ years (blue)*	*Not Suitable for HPC (red)*	*Mainstream development is adequate (green)*

S.C	superconductor	MCM	Multi Chip Module
S.C.–RSFQ	superconducting rapid single flux quanta	HIST	Heterogeneous Integrated Semiconductor Technology
III-V	three-five compound semiconductor	Cu	copper
HBT	heterojunction bipolar transistor	Low k	low dielectric constant insulation

What about software? Will we ever learn?

The working group did not establish a roadmap for software technologies. One reason for this is that progress on software technologies for HEC are less likely to result from focused efforts on specific point technologies, and more likely to emerge from large integrative projects and test beds: one cannot develop, in a meaningful way, software for high performance computing in absence of high performance computing platforms.

The State of HPC Software and It's Future

John M. Levesque
Senior Technologist
Cray Inc.

Courtesy of John Levesque, Cray ---- NRC/CTSB Future of Supers

Bottom Line

- Attack of the Killer Micros significantly hurt application development and overall performance has declined over their 10 year reign.
  - As Peak Performance has increased, the sustained percentage of peak performance has declined
- Attack of the "Free Software" is finishing off any hopes of recovery. Companies cannot build a business case to supplied needed software
- Free Software does not result in productive software.
  - While there are numerous "free software" databases, Oracle and other proprietary databases are preferred due to

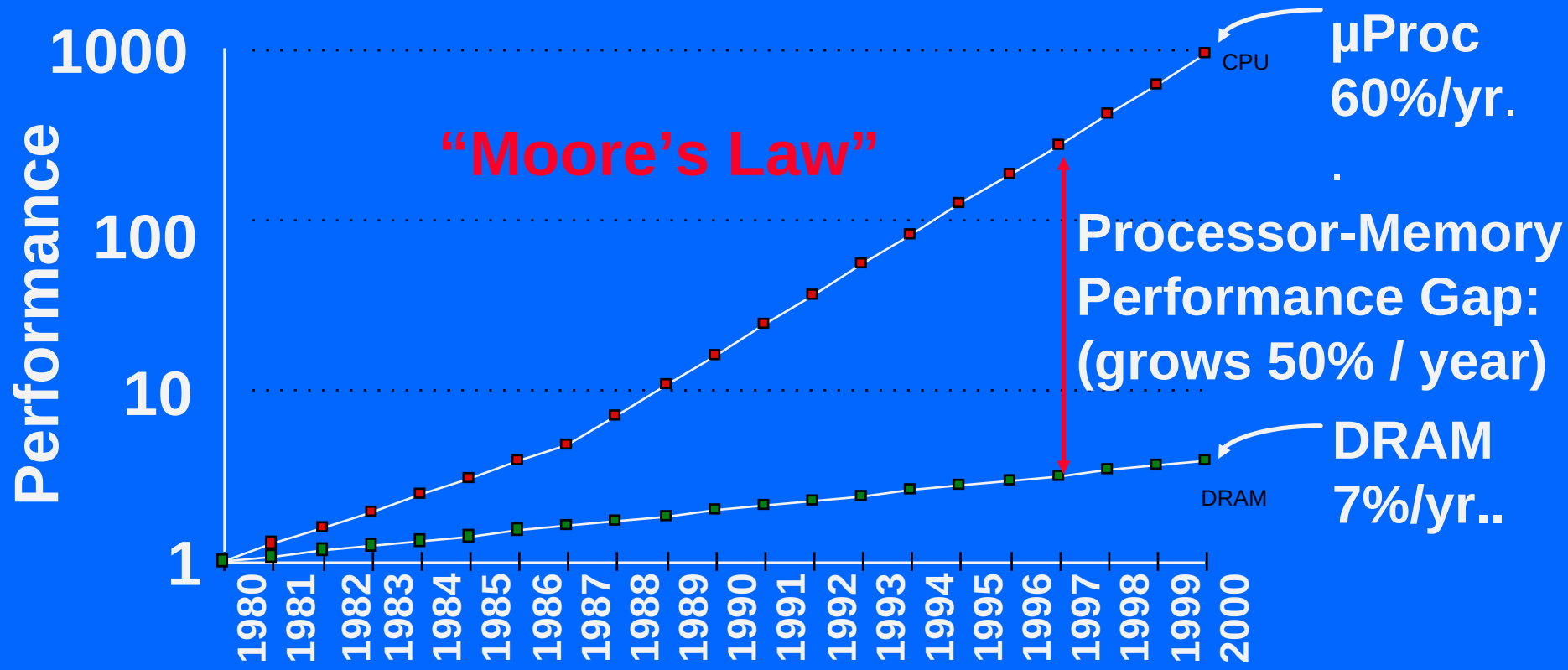
A Massive Public Works Program … but will it produce a HPC for NS?

- Furthermore, *high-end computing laboratories are needed.*
- These laboratories will fill a critical capability gap … to
  - test system software on dedicated large-scale platforms,
  - support the development of software tools and algorithms,
  - develop and advance benchmarking and modeling, and
  - simulations for system architectures, and
  - conduct detailed technical requirements analysis.
- these functions would be executed by existing

What I worry about our direction

- Overall: tiny market and need.
  - Standards paradox: the more unique and/or greater diversity of the architectures, the less the learning and market for sofware.
- Resources, management, and engineering:
  - Schedule for big ideas isn't in the ballpark e.g. Intel & SCI (c1984)
  - Are the B & B working on the problem? Or marginal people T & M?
  - Proven good designers and engineers vs. proven mediocre|unproven!
  - Creating a "government programming co" versus an industry
- Architecture
  - Un-tried or proven poor ideas? Architecture moves slowly!
  - Evolution versus more radical, but completely untested ideas
  - CMOS designs(ers): poor performance… from micros to switches
  - Memory and disk access growing at 10% versus 40-60%
- Software
  - No effort that is of the scale of the hardware
  - Computer science versus working software
  - Evolve and fix versus start-over effort with new & inexperienced

Processor Limit: DRAM Gap

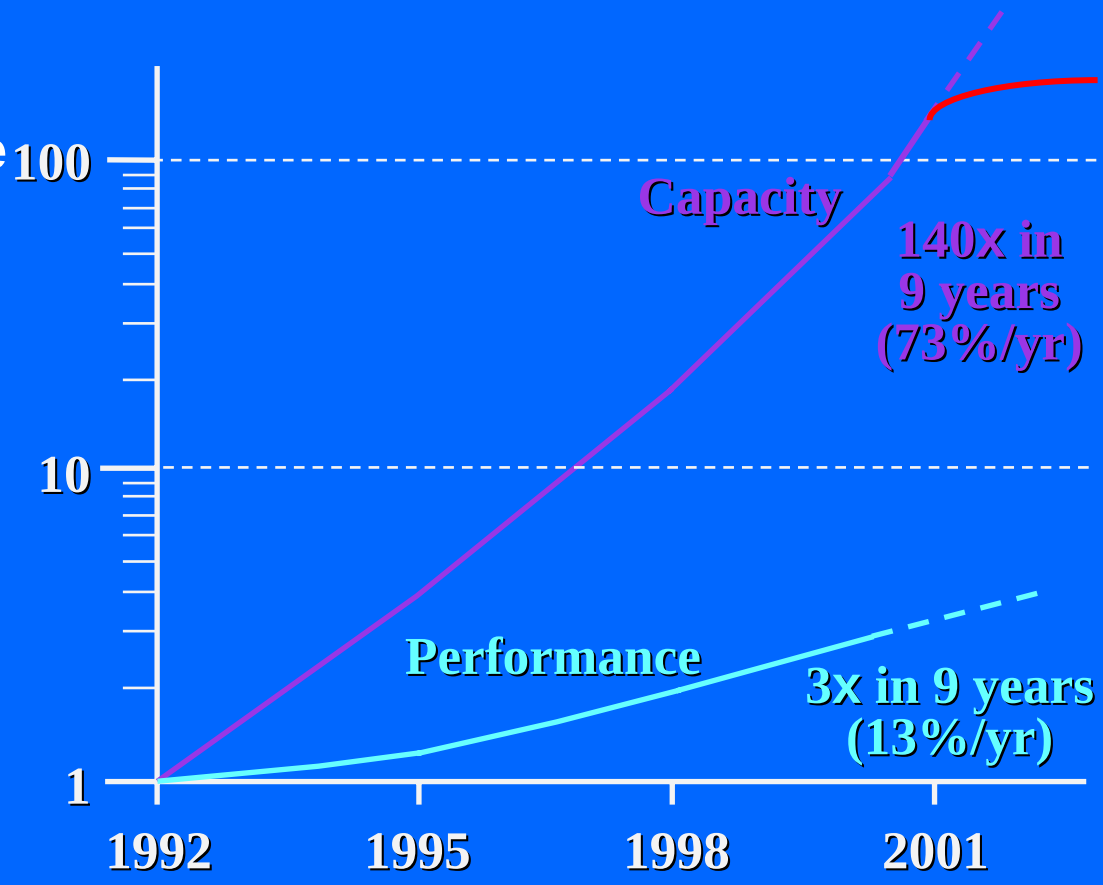


- Alpha 21264 full cache miss / instructions executed
  180 ns/1.7 ns =108 clks x 4 or 432 instructions
- Caches in Pentium Pro: 64% area, 88% transistors

*Taken from Patterson-Keeton Talk to SigMod

Disk Capacity / Performance Imbalance

- Capacity growth outpacing performance growth
- Difference must be made up by better caching and load balancing
- Actual disk capacity may be capped by market (red line); shift to smaller disks (already happening for high speed disks)



Courtesy of Richard Lary

An interesting design…

- A scalable(25:1000 nodes), low power (5w/Gflops*), high performance system
  - 1000 nodes: ~2000 Gflops for ~$1.1M
  - 100 nodes: ~200 Gflops for ~$126K
  - 25 nodes: ~50 Gflops for ~$31K
- Standard software & applications Beowulf environment
- Very good switch!
- Proven silicon engineers versus proven non-engineers

*10Kw/cabinet is a limiter for all systems

Comparing Approaches

Cost	$375,000 100 2P	$377,000 324 2P	comparable
Software	Linux, Beowulf, MPI	Linux, Beowulf, MPI	comparable
Power	25 kilowatts	3 kilowatts	**8x**
Memory bandwidth	320 GB/sec	1000 GB/sec	**3x**
Inter-P Network bi-BW	12 GB/sec	324 GB/sec	**27x**
Inter-P Network latency	120 microsecond	0.5 microsecond	**240x**
Floor space	24 sq ft	4 sq ft	**6x**
Weight	5000 lbs	500 lbs	**10x**
Delivered Performance*	910 Megaflops;	49,600 Megaflops	**54x**

*Based on NAS benchmarks. Assumes: cache miss/15 flops, 1000 small msgs/mflops
Approximately 600 GB primary memory. Node cost 1200 + 100 for Ethernet. If Myrinet, divide by 2. (1500/node)
Linpack 8 Gflops/node; 2 Gflops/node; Or 800 for Penita vs 650 for other.

PC Nodes Don't Make Good Large-Scale Technical Clusters

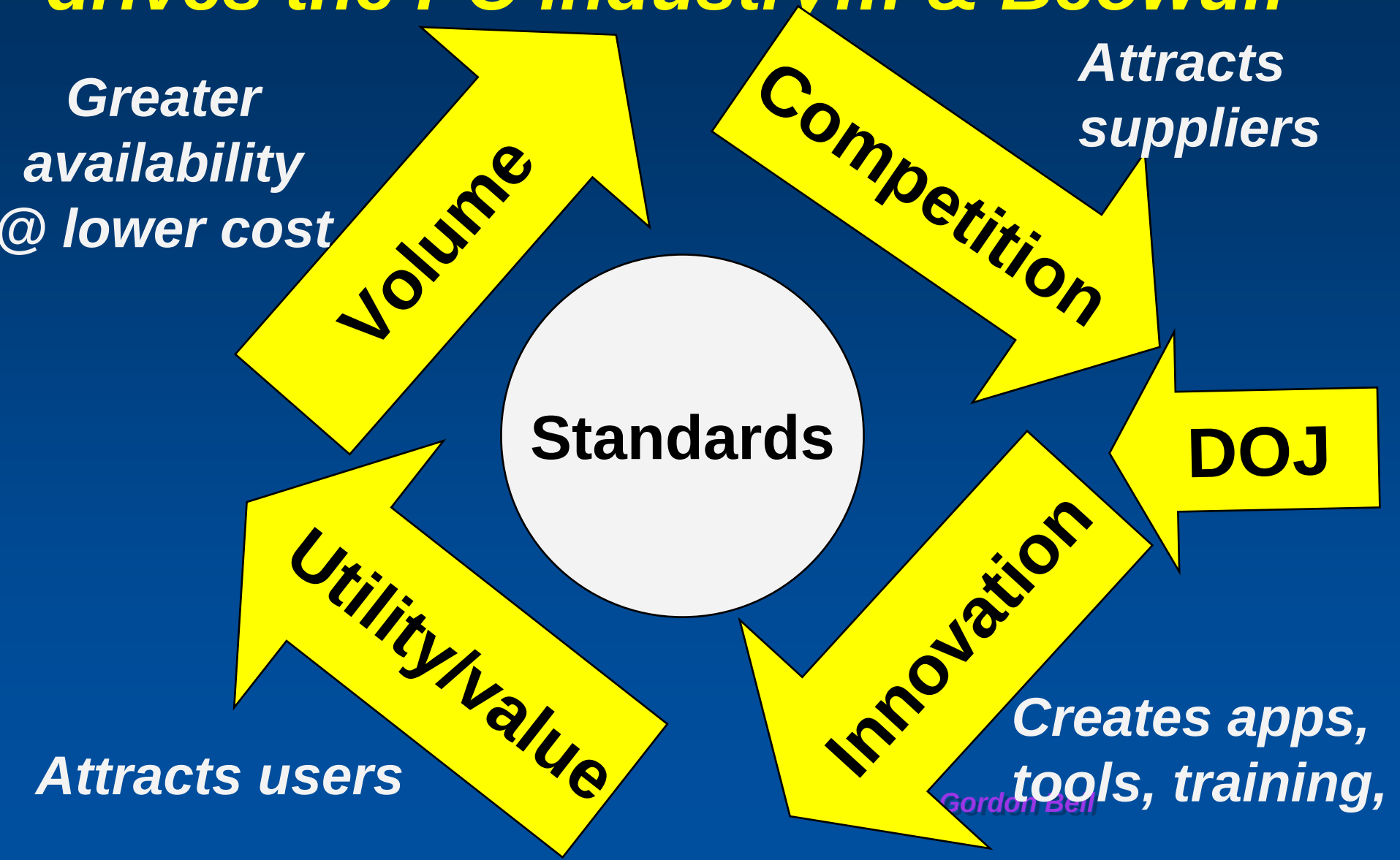
- PC microprocessors are optimized for desktop market (the highest volume, most competitive segment)
  - Have very high clock rates to win desktop benchmarks
  - Have very high power consumption and (except in small, zero cache miss rate applications) are quite mismatched to memory
- PC nodes are physically large
  - To provide power and cooling – Papadopoulous "computers… suffer from excessive cost, complexity, and power consumption"
  - To support other components
- High node-count clusters must be spread over many cabinets with cable-based multi-cabinet networks
  - Standard (Ethernet) and faster, expensive networks have quite inadequate performance for technical applications …Switching cost equals node cost
  - Overall size, cabling, and power consumption reduce reliability …Infrastructure cost equals node cost

Lessons from Beowulf

- An experiment in parallel computing systems ‘92
- Established vision- low cost high end computing
- Demonstrated effectiveness of PC clusters for some (not all) classes of applications
- Provided networking software
- Provided cluster management tools
- Conveyed findings to broad community
- Tutorials and the book
- Provided design standard to rally community!
- Standards beget: books, trained people, software … virtuous cycle that allowed apps to form
- Industry began to form beyond a research project

Courtesy, Thomas Sterling, Caltech.

The Virtuous Economic Cycle drives the PC industry… & Beowulf



The End