Supercomputer Class
CMU-Qatar

Gordon Bell

Microsoft Research, Silicon Valley Laboratory

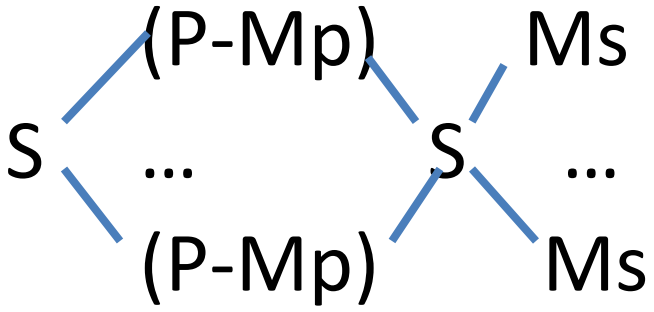
13 February

Topics

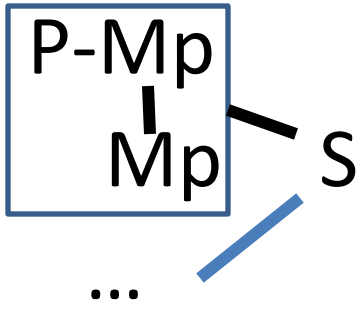
1. Evolution: B.C. to 2011
2. Parallelism, Scalability and Performance
3. Performance: Measurement, benchmarks, and kernels
4. Top500 (2011, 2012): Alternative approaches
5. HPC aka Supercomputers versus Cloud computing

HPC & Cloud: Twins, Separated at Birth (Computation versus Storage Centric)

HPC Separate Storage Area Network, two switches



Cloud Attached Storage, single switch



Supercomputer Evolution

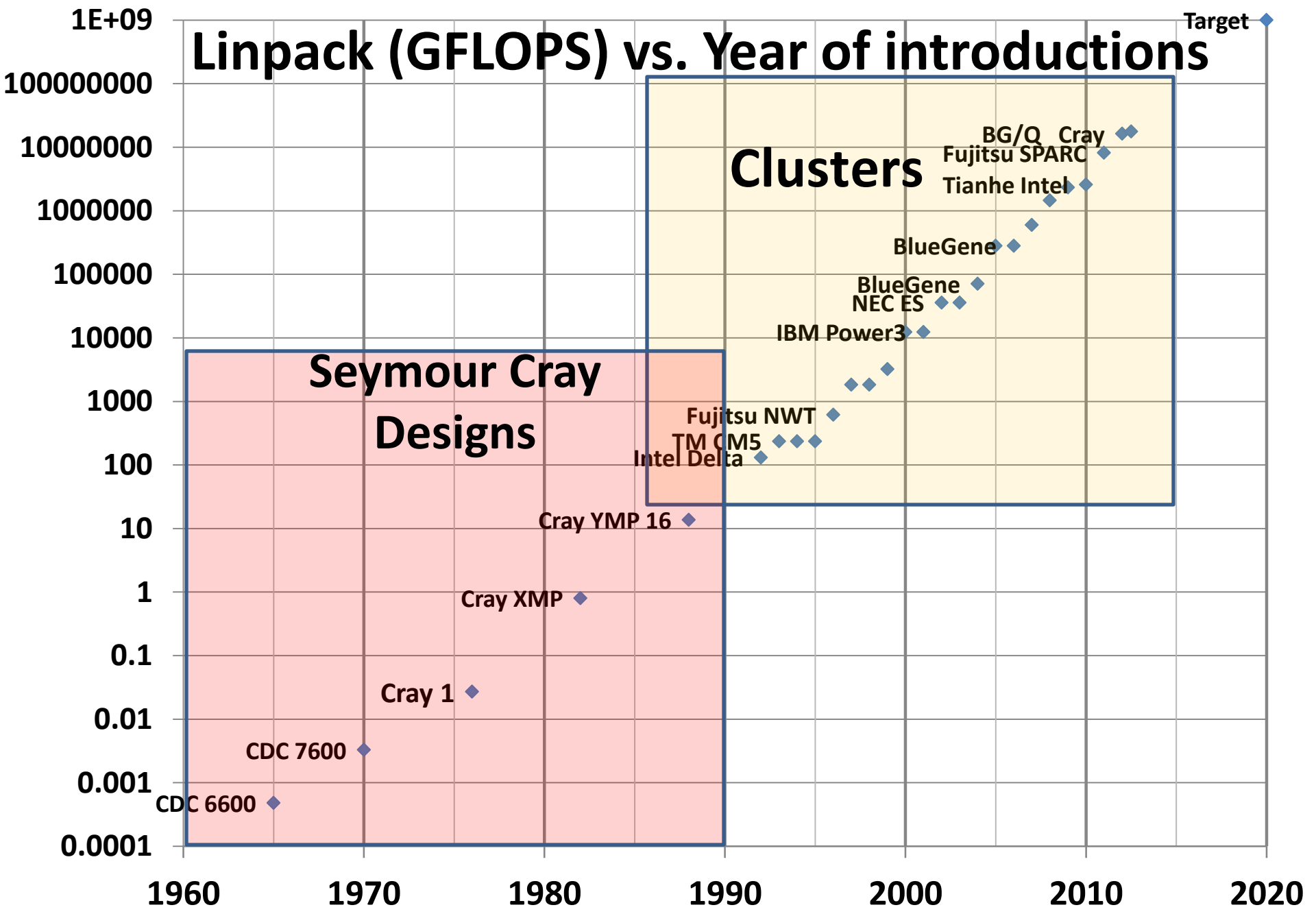
- What defines a supercomputer?
  - What is its function: calculate, run FORTRAN
  - Quest for performance: Who can build the fastest?
  - Price: How much do you have to spend?
    - To buy, to build the building, to power, to run
    - To program
  - Programming environment (standards): Beowulf
  - Users (market): climate models, science (simulate phenomena, engineering design
  - Applications: 3d time varying. Code breaking.
- Calculation versus record processing

Supercomputing: Speed & parallelism

- Clock speed
- Parallelism within a single instruction stream including wider instruction word
- Pipelining
- Vector processing
- Scalability across multiple streams & multi-threading
- Multiprocessors—Scale up
- Multiple computers—Scale out
- Stream processing using GPUs

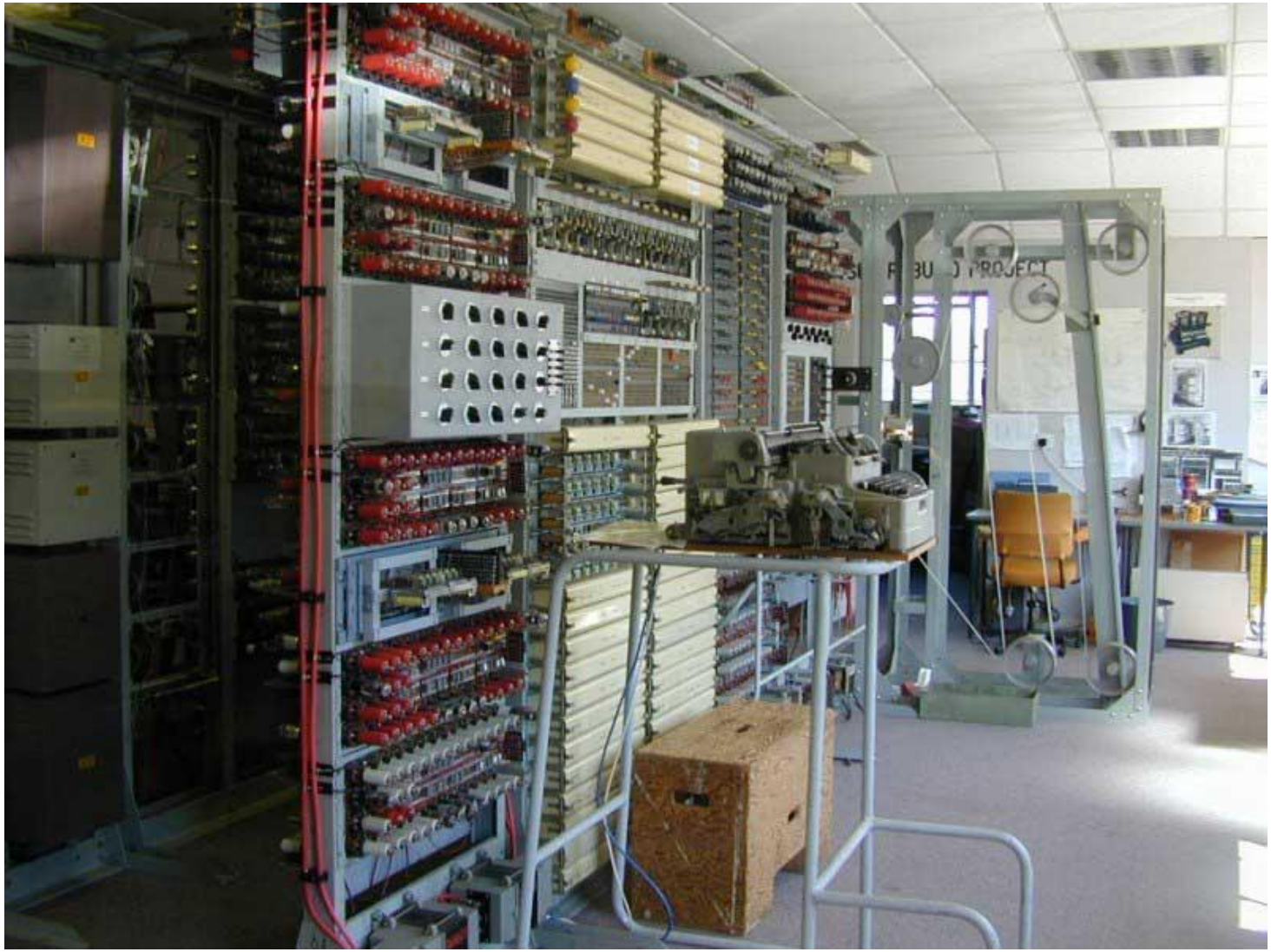
Five eras of Scientific Computing

Period	*Technology*	Machines (artifacts)
193x-1947	***Electromechanical-vacuum tubes; one-of machines Search for "the computer"***	**Computing with cards** at Los Alamos; IBM Multiplying calculator. Atanasoff, Harvard Marks, BTL, Zuse, culminating in ENIAC and **the EDVAC Report.**
1947-1950s	***Electronic Computing Era Vacuum Tube Scientific Calculators including von Neumann X-iacs***	**The Big Bang**. First stored program computers that just work (Univac, IBM 701 and ERA); Illiac, Maniac etc. Amdahl's WISC, First
1960s	***Discrete transistors. Supercomputer Class forms.*** Build fast single instruction stream processors; **FORTRAN** established.	**FORTRAN; LARC, STRETCH (61), plus 7090 and CDC 1604 workhorses Seymour Cray wins: CDC 6600 (64) & 7600 (71)**
Mid70s-mid 90s	***ICs (bipolar) …CMOS. Vector processor Era***	**Intro of Cray 1, vector processor 1975 and evolution takes over using multiple processors vector XMP, YMP, C-90 , T-90**
Mid 80s to the present	***Scalables era (commodity killer micros including "game" processors)***	**Scalable computers using micros**: **How much money? Seitz Cosmic Cube c1985, move to Intel and others. 45 companies casualties.**


Linpack (GFLOPS) vs. Year of introductions
1E+09
100000000
10000000
1000000
100000
10000
1000
100
10
1
0.1
0.01
0.001
0.0001
1960
1970
1980
1990
2000
2010
2020
Target
Clusters
BG/Q
Cray
Fujitsu SPARC
Tianhe Intel
BlueGene
BlueGene
NEC ES
IBM Power3
Fujitsu NWT
TM CM5
Intel Delta
Seymour Cray Designs
Cray YMP 16
Cray XMP
Cray 1
CDC 7600
CDC 6600

JUMP to Parallelism..

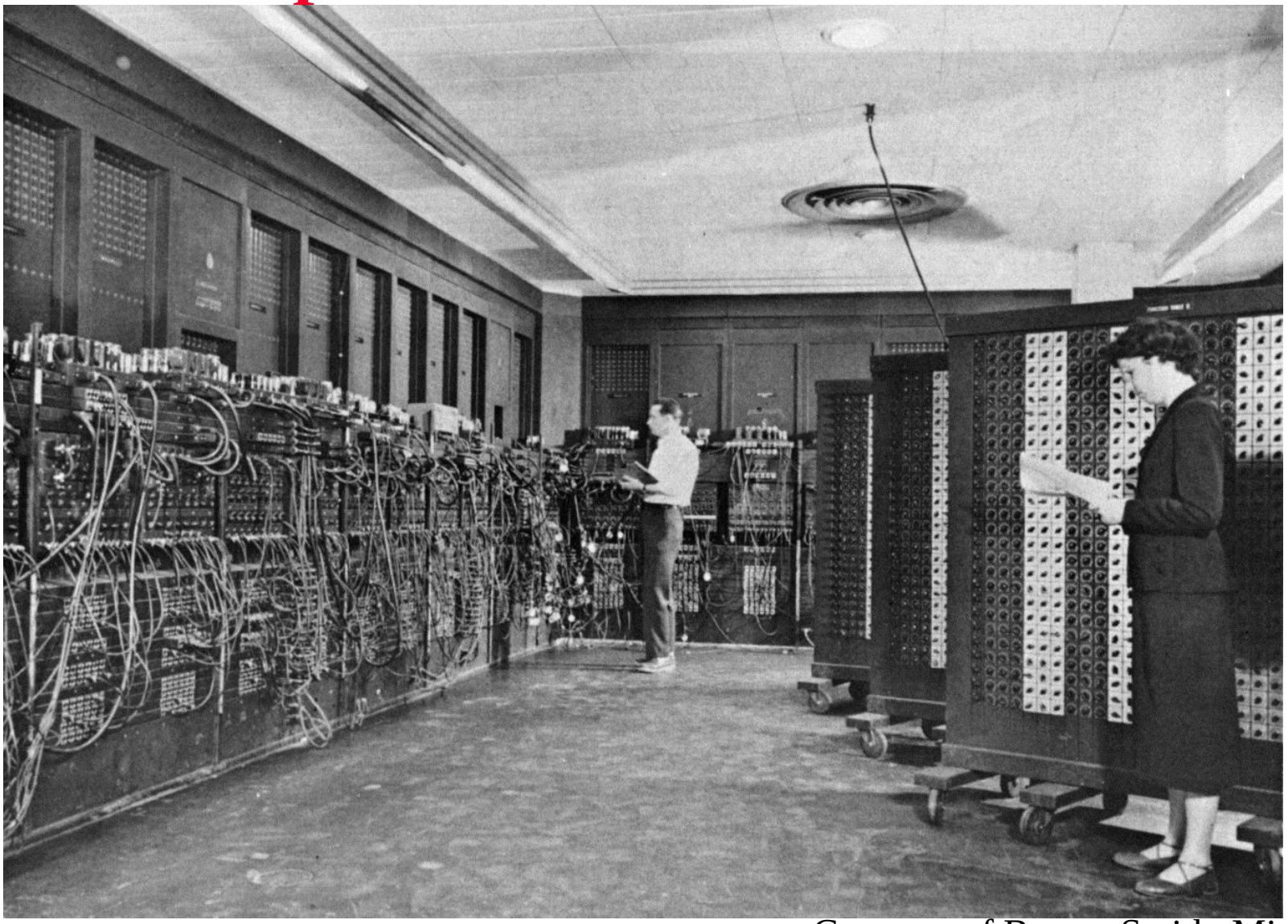
Colossus: 1943, 1944 10 produced



Courtesy of Burton Smith, Microsoft

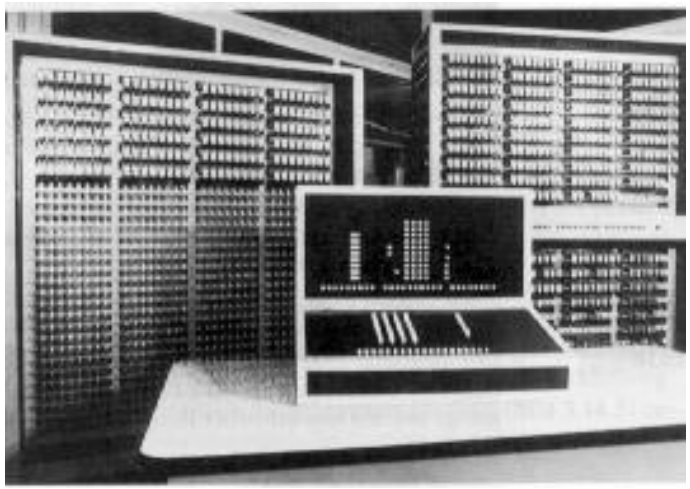
Bletchley Park “Bombe”

ENIAC: Electronic Numerical Integrator and Computer 1946-1955; Cost $500,000

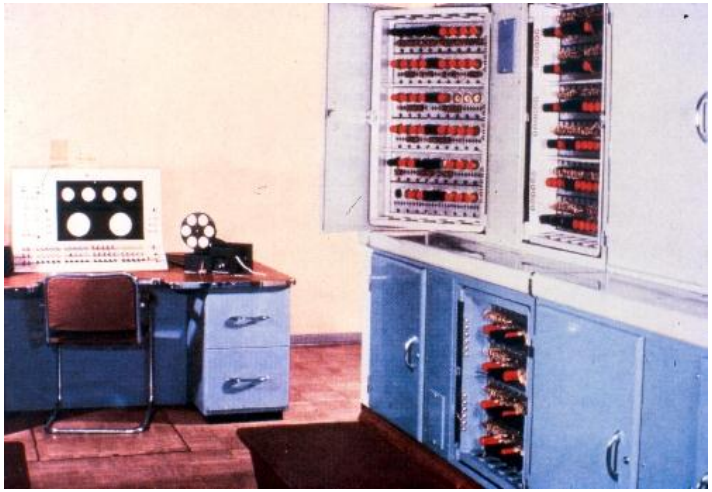


Courtesy of Burton Smith, Microsoft

Other early supercomputers

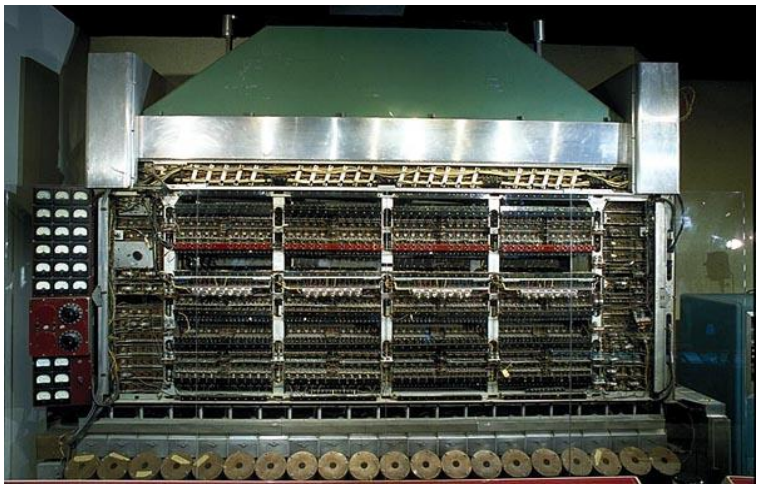


Zuse Z3 (1941)



Manchester/Ferranti Mark I (1951)

Univac 1 (1951)

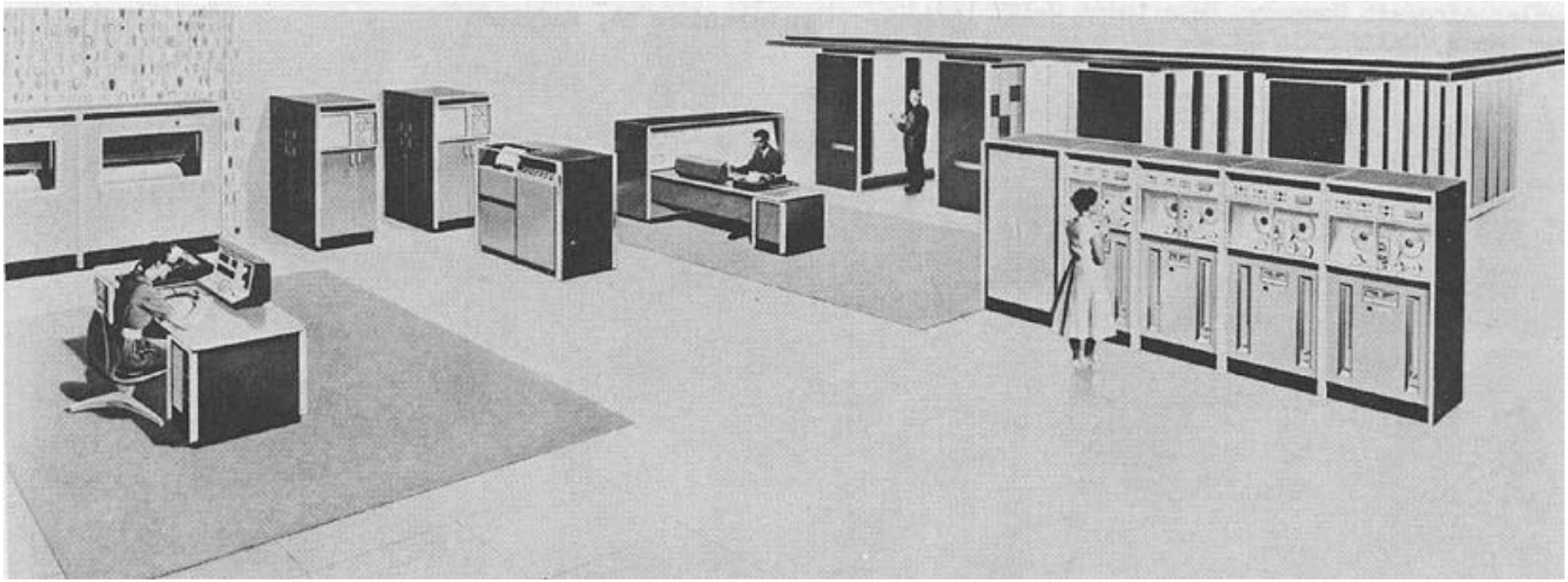


The IAS machines (1952)

Courtesy of Burton Smith, Microsoft

Mainframes: LARC

- **Begun in 1955 for Livermore and delivered in 1960**
- **Had dual processors and decimal arithmetic**
- **Employed surface-barrier transistors and core memory**



Courtesy of Burton Smith, Microsoft

Mainframes: Stretch and Harvest

- **IBM 7030 (STRETCH)**
- **Delivered to Los Alamos 4/61**
- **Pioneered in both architecture and implementation at IBM**

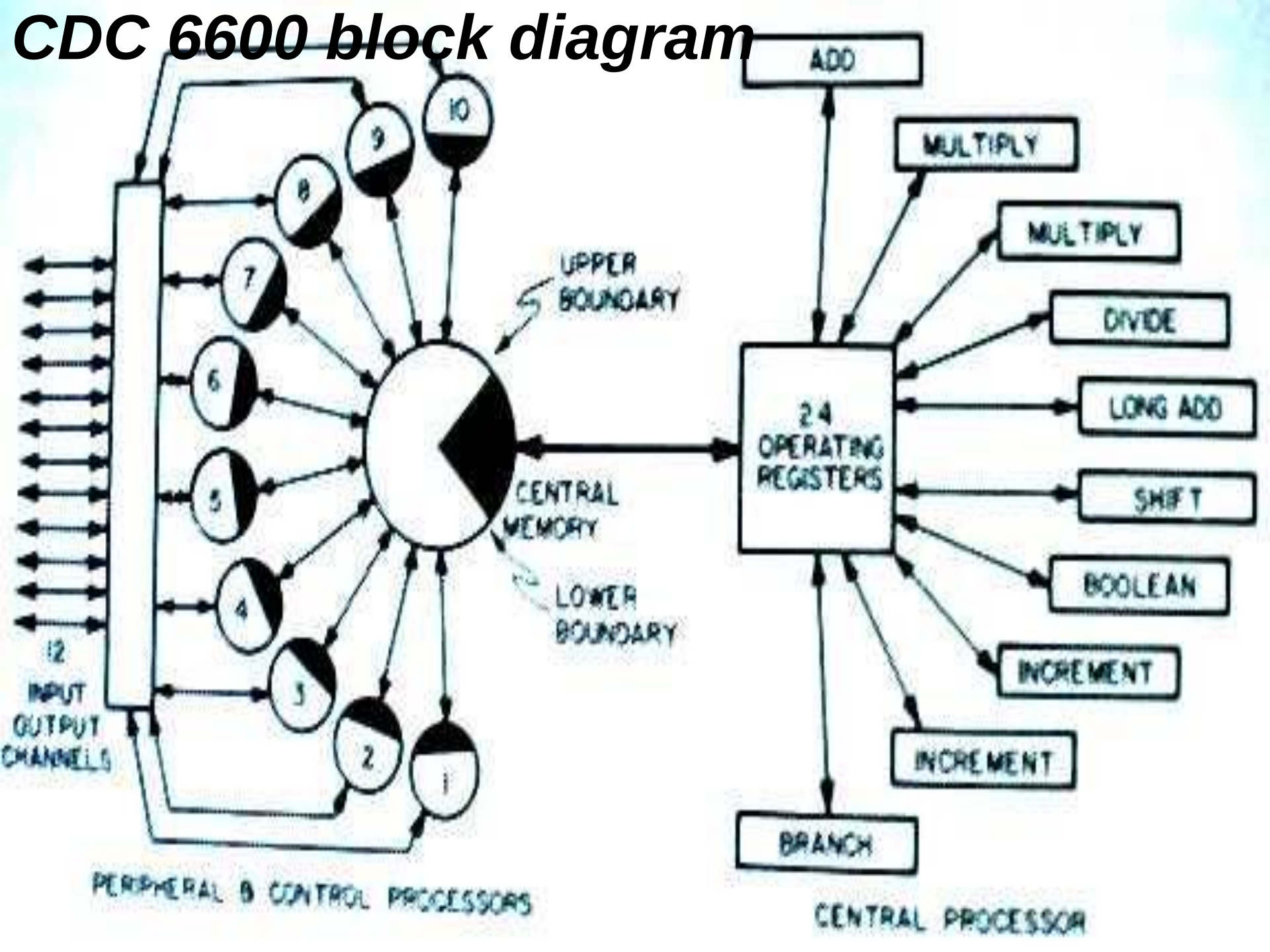
- **IBM 7950 (HARVEST)**
- **Delivered to NSA 2/62**
- **Was STRETCH + 4 boxes**
  - **IBM 7951 Stream unit**
  - **IBM 7952 Core storage**
  - **IBM 7955 Tape unit**
  - **IBM 7959 I/O Exchange**

Courtesy of Burton Smith, Microsoft

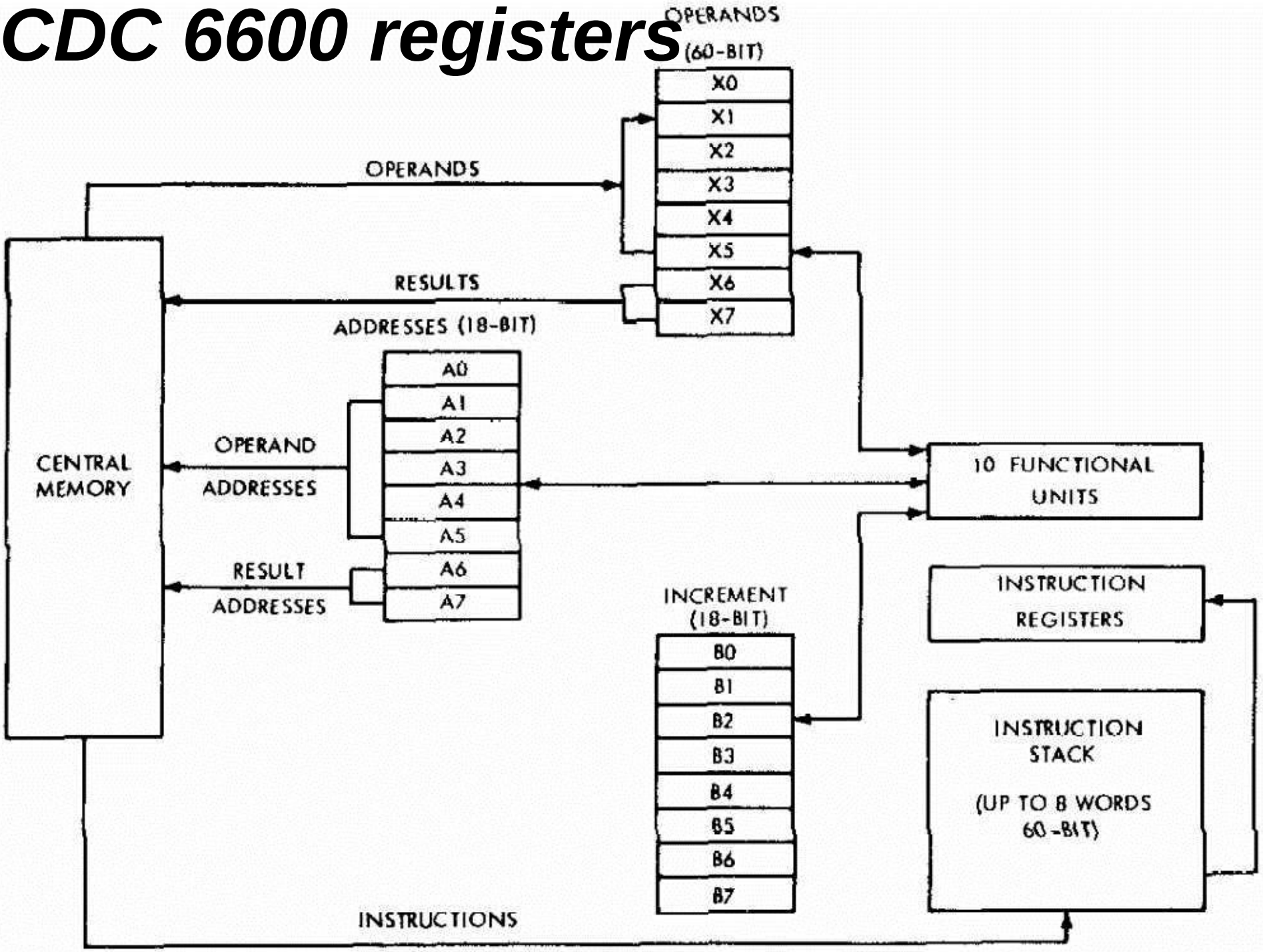
CDC 6600 Console c1964

Courtesy of Burton Smith, Microsoft

CDC 6600 block diagram



CDC 6600 registers



Two CDC 7600s and LLNL c1969

Courtesy of Burton Smith, Microsoft

CDC 7600 block diagram

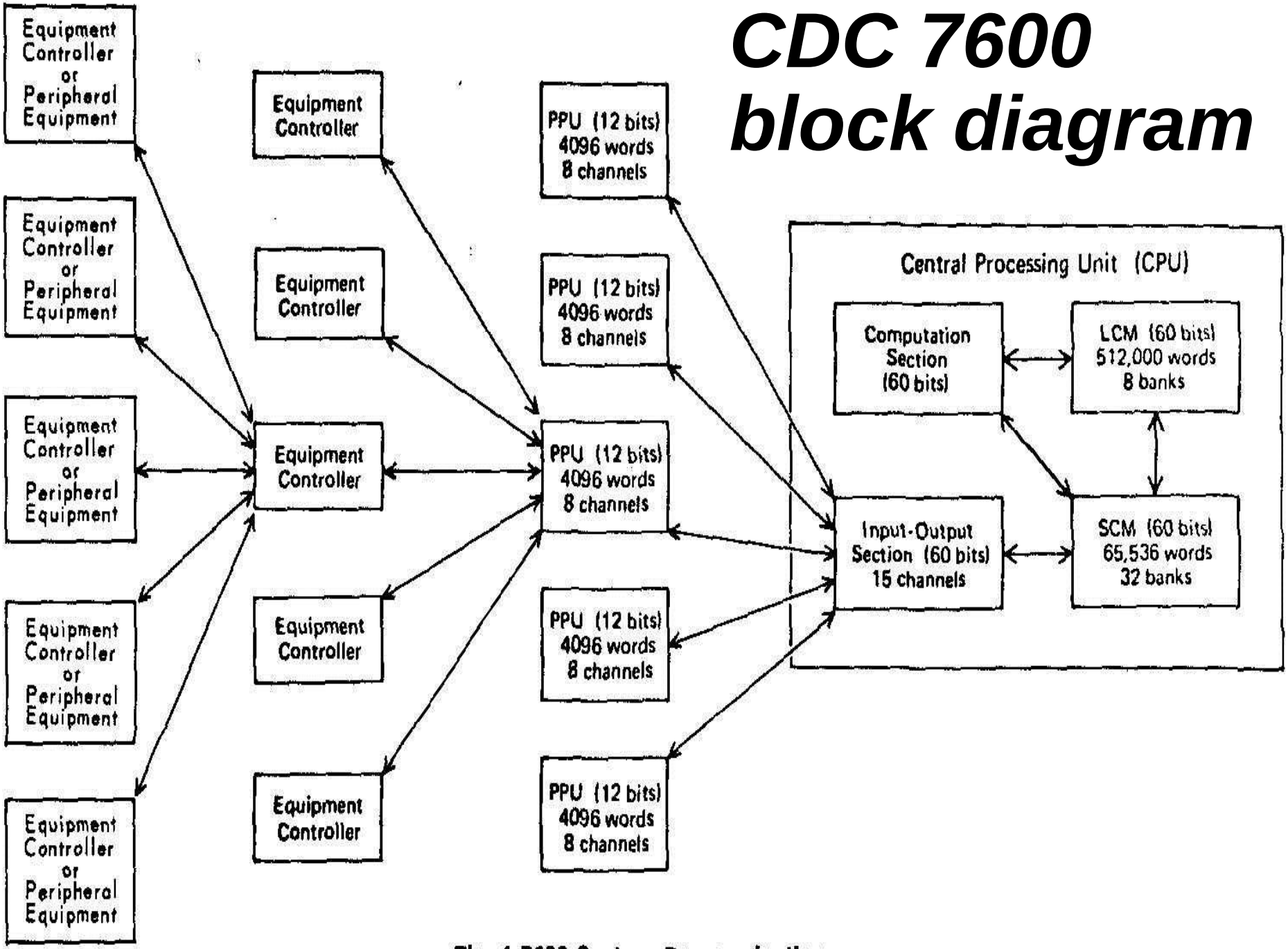


Fig. 1 7600 System Communication

Amdahl's law… the limit

- **If w_1 work is done at speed s_1 and w_2 at speed s_2, the average speed s is $(w_1+w_2)/(w_1/s_1 + w_2/s_2)$**
  - **This is just the total work divided by the total time**
- **For example, if $w_1 = 9$, $w_2 = 1$, $s_1 = 100$, and $s_2 = 1$ then $s = 10/1.09 \cong 9$ (speed)**
  - **This is obviously not the average of s_1 and s_2**

Amdahl, Gene M, "Validity of the single processor approach to achieving large scale computing capabilities", Proc. SJCC, AFIPS Press, 1967

Courtesy of Burton Smith, Microsoft

SIMD arrays: Illiac IV

- By the late 60's, it was clear mainframes weren't enough
- To improve performance, SIMD array machines were built or proposed with many arithmetic processing units
  - Solomon was an early Westinghouse SIMD array prototype
- The Illiac IV was a U. of Illinois/Burroughs project
  - Funded by DARPA from 1964 onward, usable in 1975
  - The chief architect, Dan Slotnick, from Westinghouse
- It was to have 256 arithmetic units, cut back to 64
- The thin-film memory system was a major headache
- After student demonstrations at Illinois in May 1970, the project was moved to NASA-Ames
- Languages, especially FORTRAN, aimed to use parallel loops to express parallelism

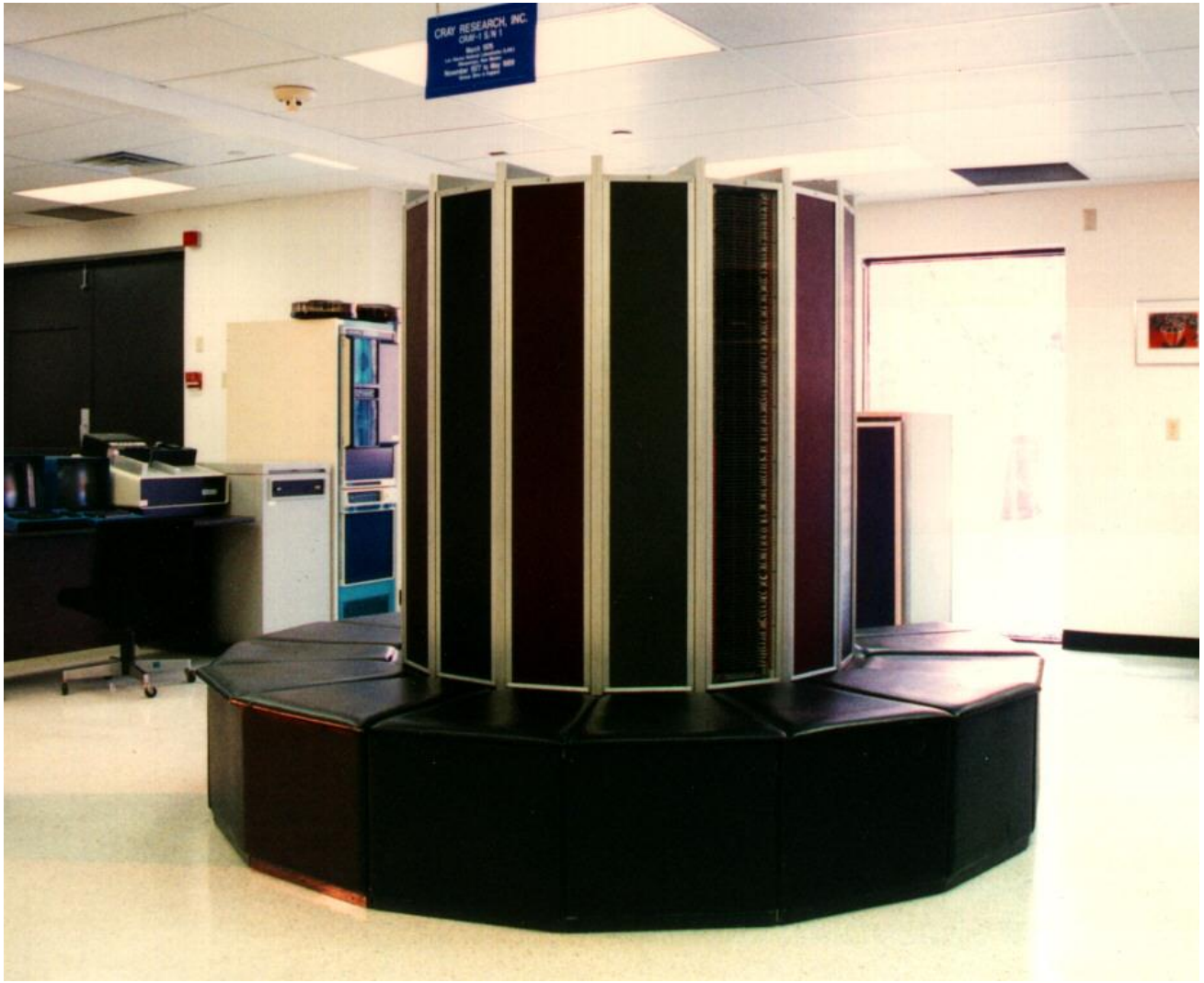
Courtesy of Burton Smith, Microsoft

ILLIAC IV: Uof IL at NASA in1971

Courtesy of Burton Smith, Microsoft

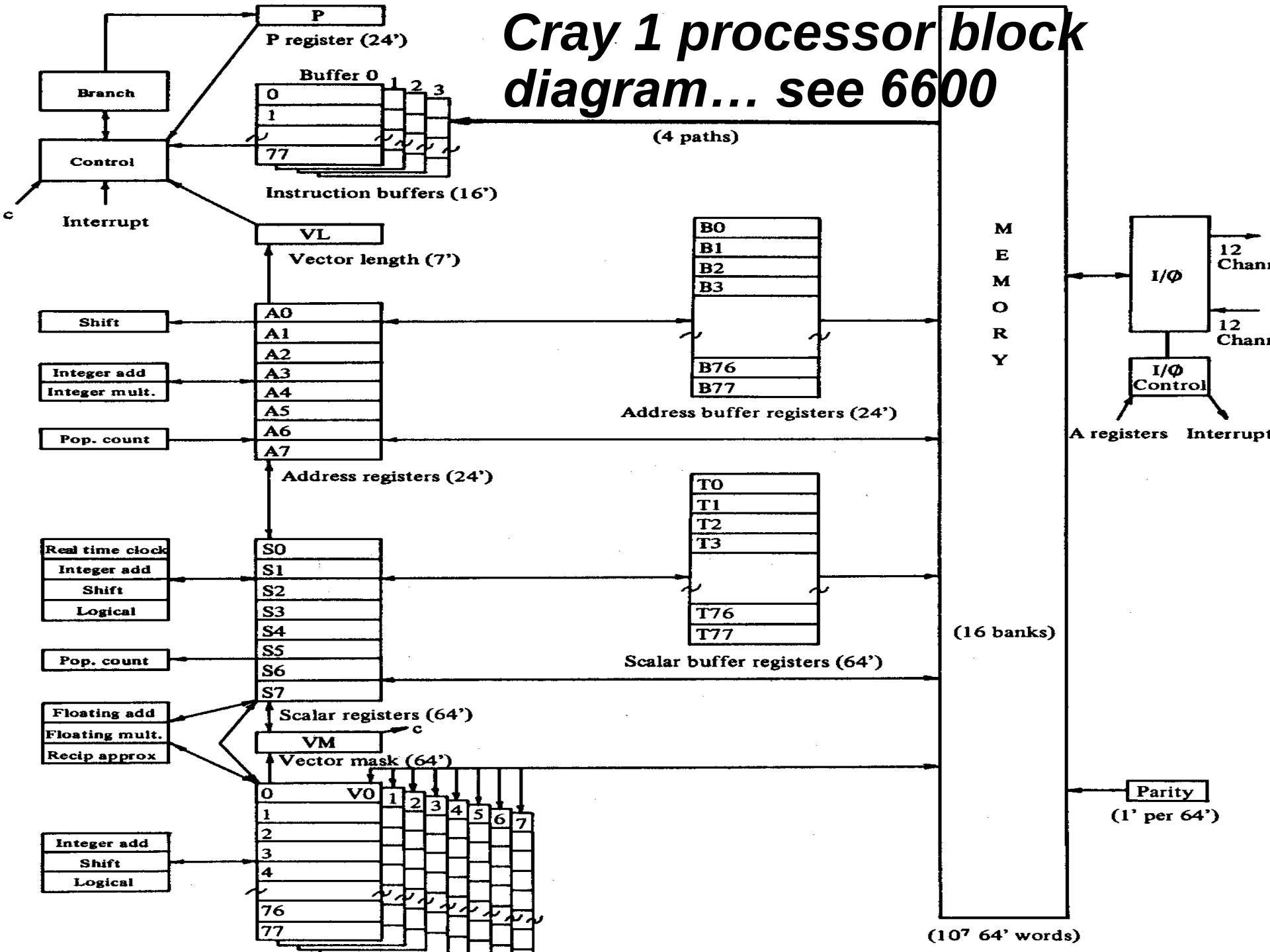
CMU C.mmp c1974: 16 processor, shared memory

Cray-1 c1976



Courtesy of Burton Smith, Microsoft

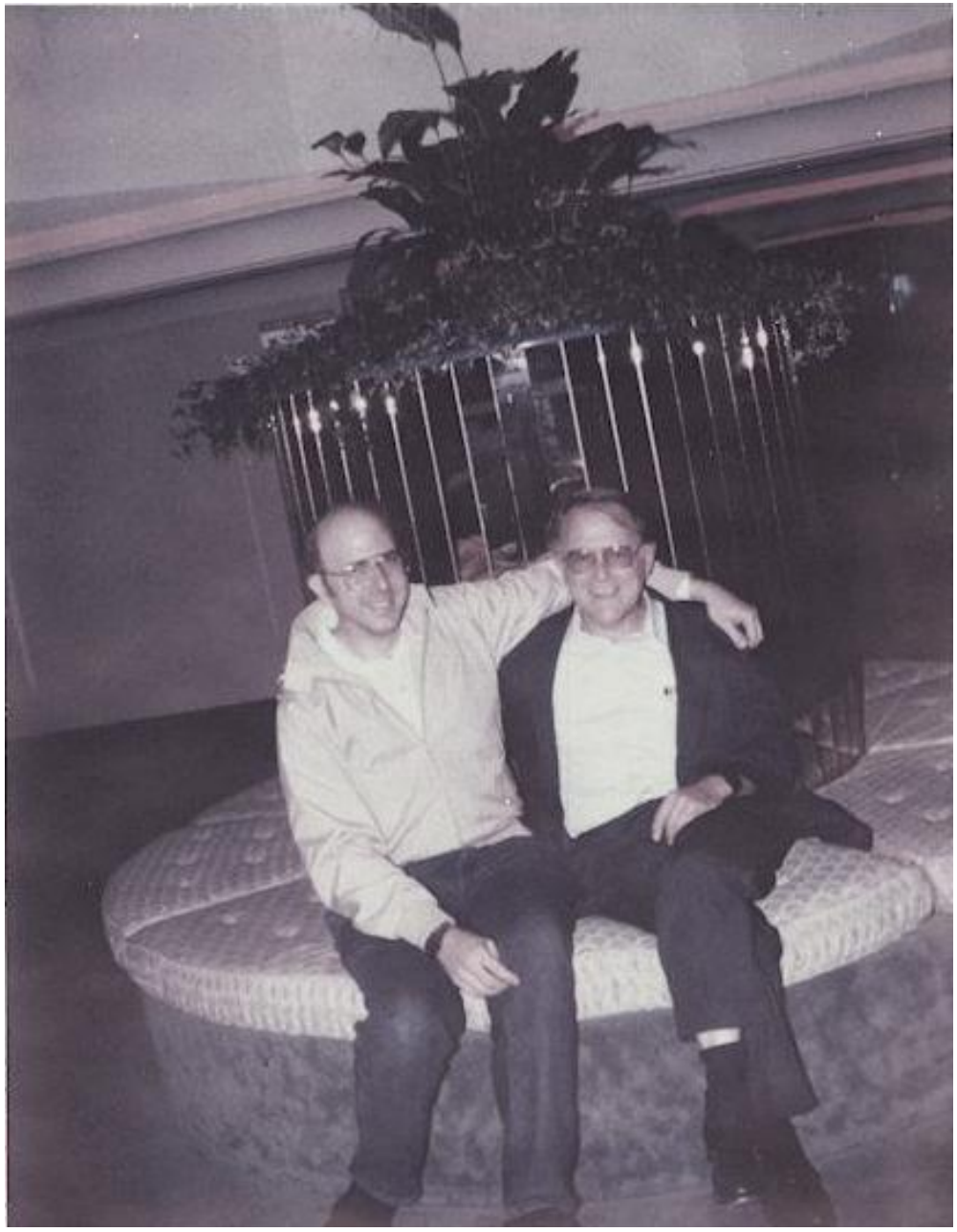
Cray 1 processor block diagram… see 6600



Vector Pipelining: Cray-1

- **Unlike the CDC Star-100, there was no development contract for the Cray-1**
  - **Cray disliked government's looking over his shoulder**
- **Instead, Cray gave Los Alamos a one-year free trial**
- **Almost no software was provided by Cray Research**
  - **Los Alamos developed or adapted existing software**
- **After the year was up, Los Alamos leased the system**
  - **The lease financed by a New Mexico petroleum person**
- **The Cray-1 did not suffer from Amdahl's law**
  - **Its scalar performance was twice that of the 7600**
  - **Once vector software matured, 2x became 8x or more**
- **The word "supercomputer" has connoted a Cray-1**

Courtesy of Burton Smith, Microsoft

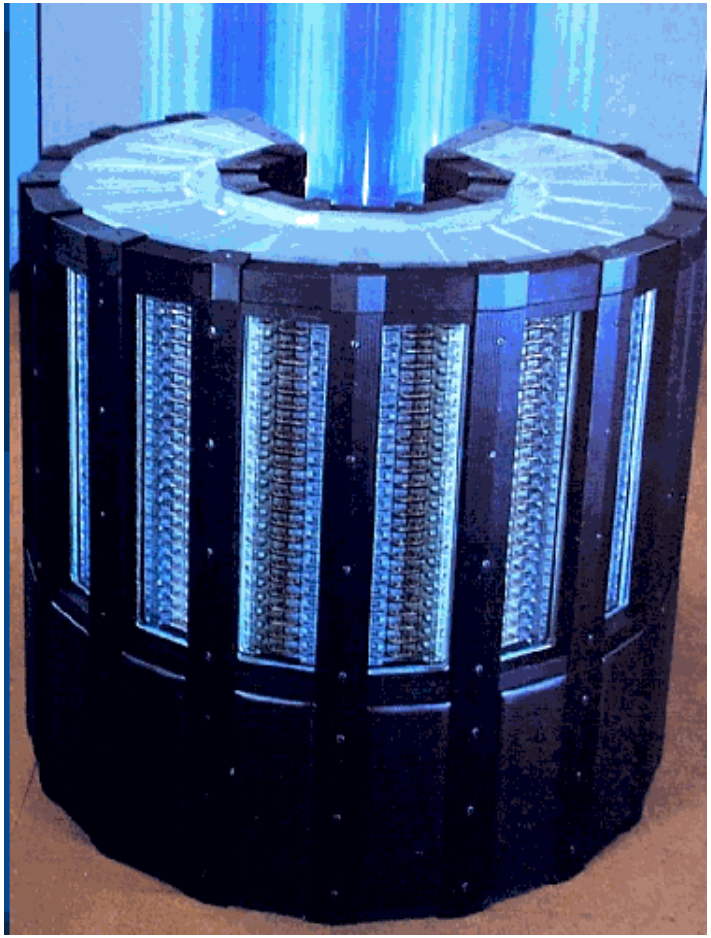


Steve Squires, DARPA & Gordon Bell, Encore seated at a“Cray”. Kickoff of DARPA’s SCI program c1984

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

Shared Memory: Cray Vector Systems

- **Cray Research, by Seymour Cray**
  - **Cray-1 (1976): 1 processor**
  - **Cray-2 (1985): up to 4 processors***
- **Cray Research, not by Seymour Cray**
  - **Cray X-MP (1982): up to 4 procs**
  - **Cray Y-MP (1988): up to 8 procs**
  - **Cray C90: (1991?): up to 16 procs**
  - **Cray T90: (1994): up to 32 procs**
  - **Cray X1: (2003): up to 8192 procs**
- **Cray Computer, by Seymour Cray**
  - **Cray-3 (1993): up to 16 procs**
  - **Cray-4 (unfinished): up to 64 procs**
- **All are UMA systems except the X1, which is NUMA**

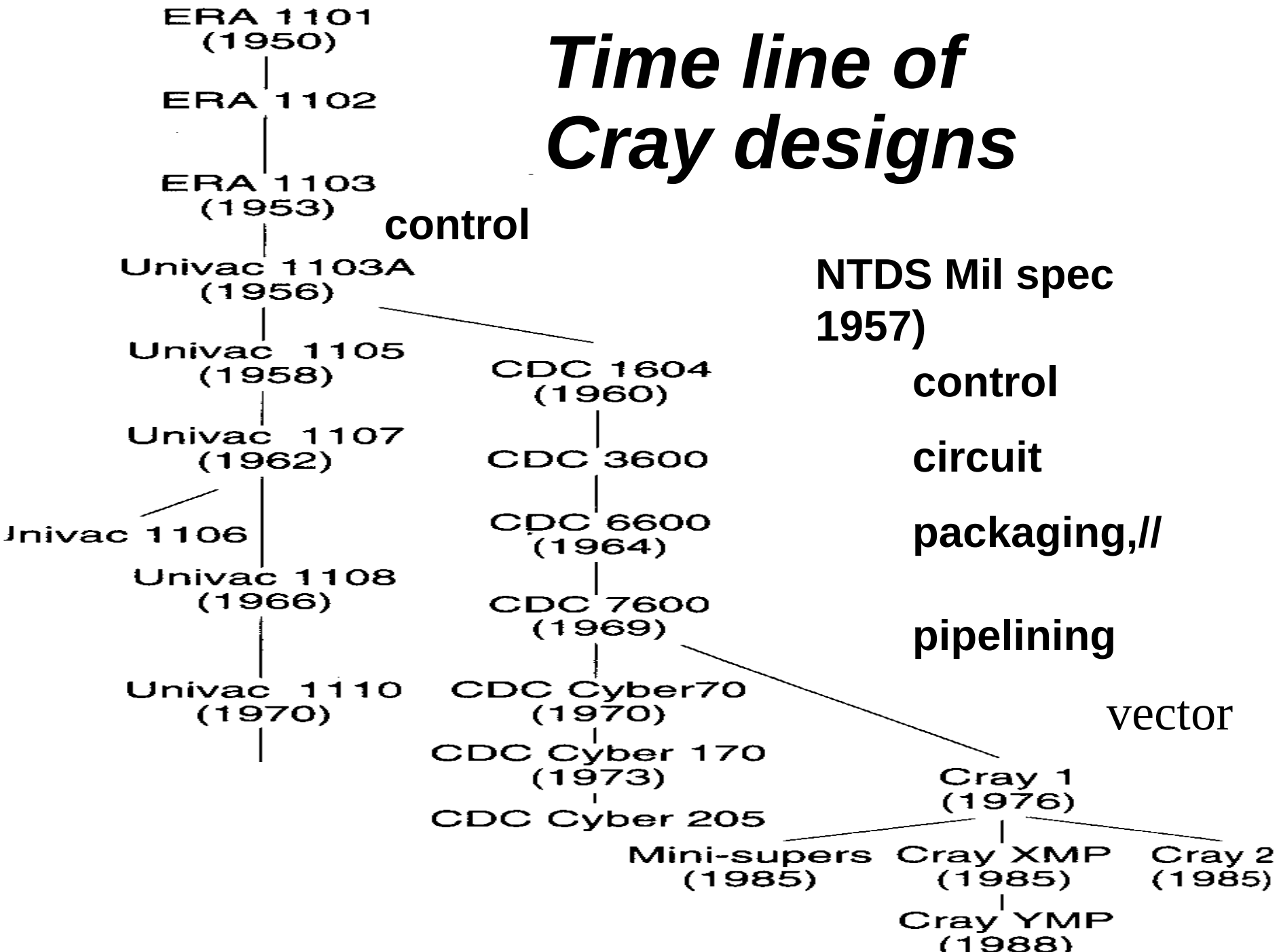


Cray-2

***One 8-processor Cray-2 was built**

Courtesy of Burton Smith, Microsoft

Time line of Cray designs



control

NTDS Mil spec 1957)

control

circuit

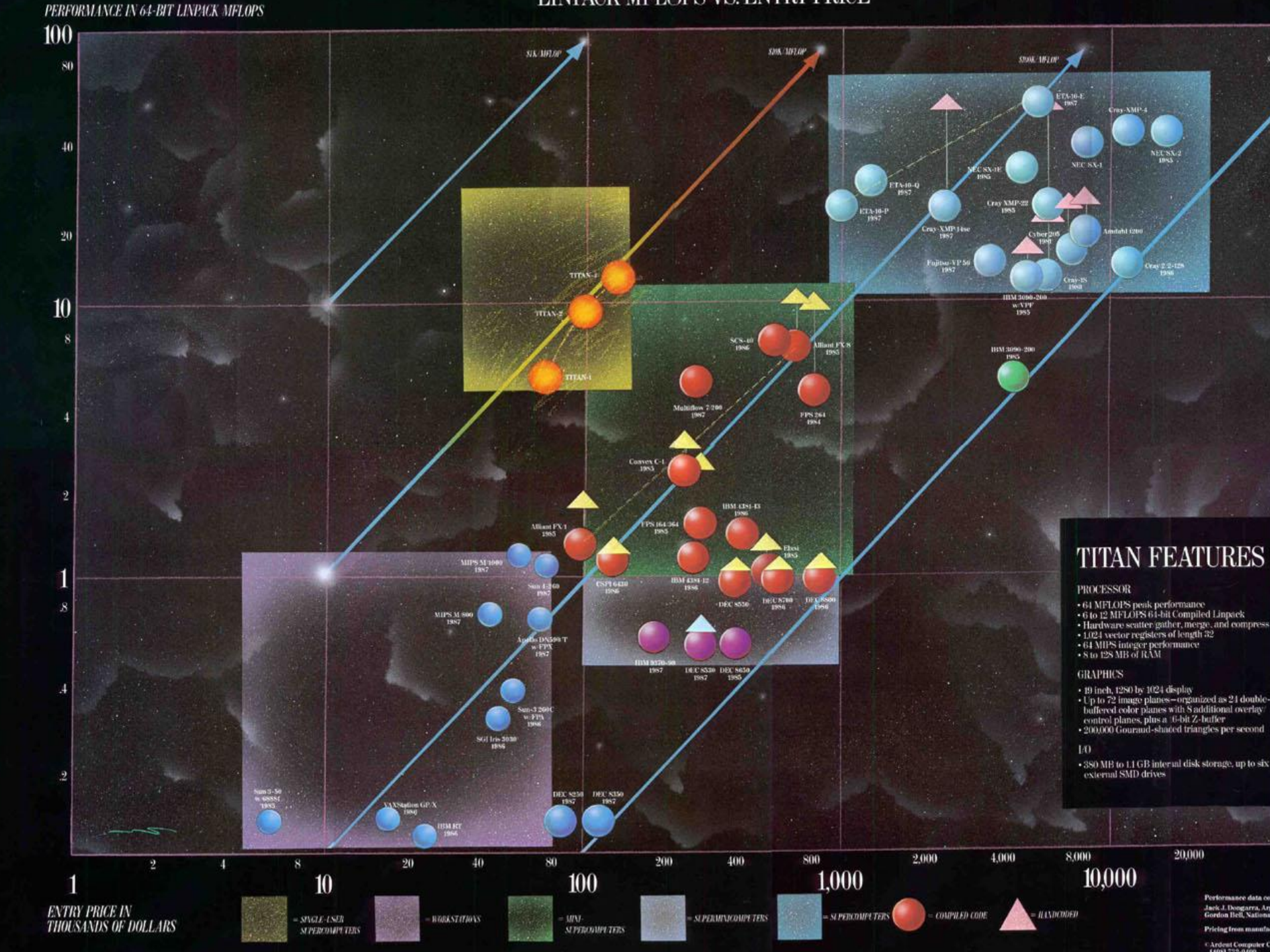
packaging,//

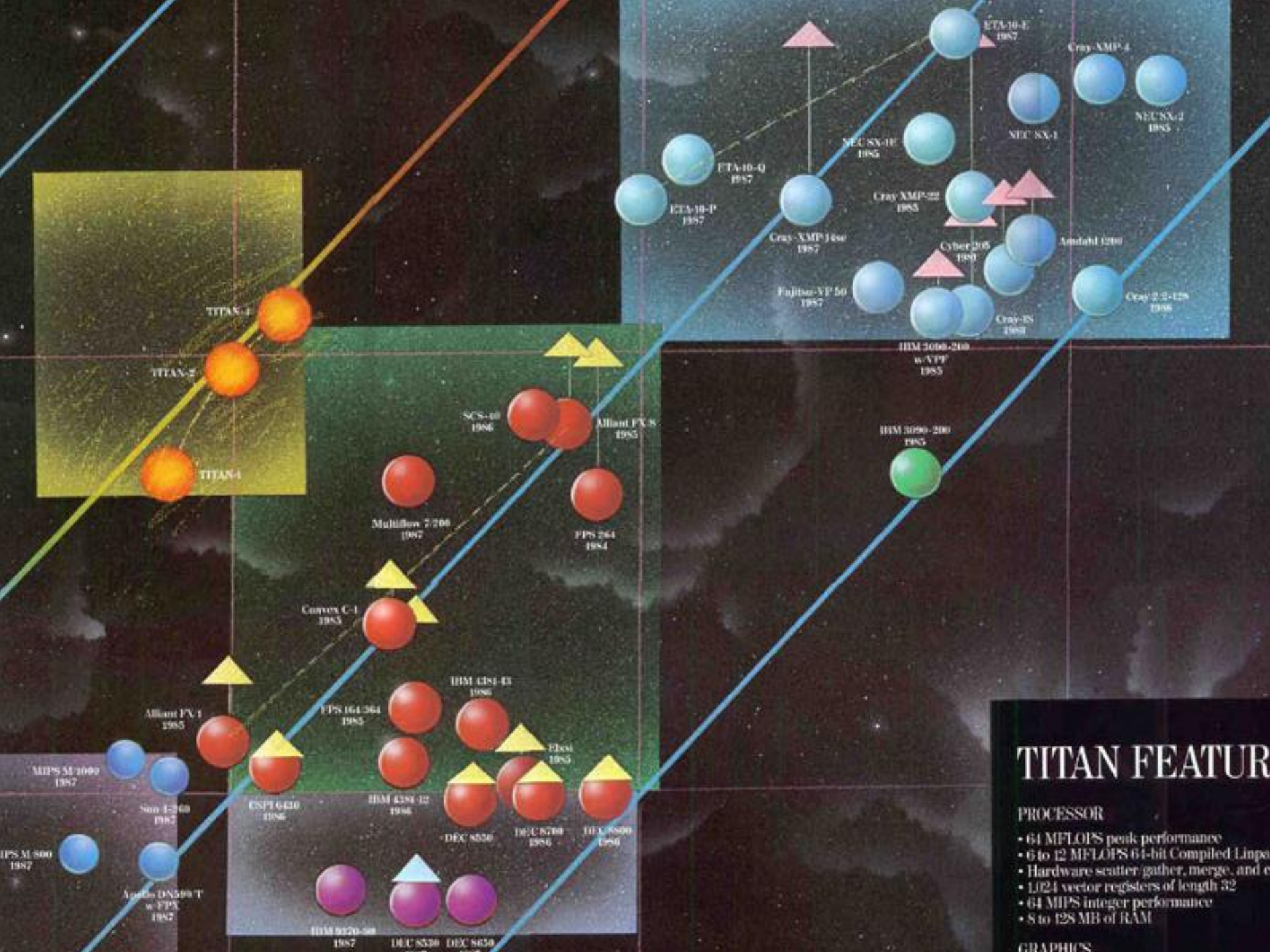
pipelining

vector

Alternative scale computers

- **Mini-supercomputers**
- **Personal supercomputers**


PERFORMANCE IN 64-BIT LINPACK MFLOPS
LINPACK MFLOPS VS. ENTRY PRICE
100
80
40
20
10
8
4
2
1
$1K/MFLOP
$20K/MFLOP
$500K/MFLOP
ETA-10-E 1987
Cray-XMP-4
NEC SX-2 1985
NEC SX-1
NEC SX-1E 1986
ETA-10-Q 1987
ETA-10-P 1987
Cray XMP-22 1985
Cray-XMP/14se 1987
Cyber 205 1985
Amdahl 6280
Fujitsu VP-50 1987
Cray-2S 1985
Cray 2-2-128 1986
IBM 3090-200 w/VF 1985
TITAN-1
TITAN-2
TITAN-4
SCS-40 1986
Alliant FX/8 1985
IBM 3090-200 1985
Multiflow 7/200 1987
FPS 264 1984
Convex C-1 1985
Alliant FX/1 1985
IBM 4381-13 1986
FPS 164/364 1985
Elxsi 1985
MIPS M/1000 1987
Sun 4/260 1987
CSPI 6420 1986
IBM 4381-12 1986
DEC 8550
DEC 8700 1986
DEC 8800 1986
MIPS M/800 1987
Apollo DN590/T w/FPX 1987
IBM 9370-90 1987
DEC 8530 1987
DEC 8650 1985
Sun-3 260C w/FPA 1986
SGI Iris 3030 1986
Sun-3 50 w/68881 1985
VAXStation GP/X 1986
IBM RT 1986
DEC 8250 1987
DEC 8350 1987
1
2
4
8
10
20
40
80
100
200
400
800
1,000
2,000
4,000
8,000
10,000
20,000
ENTRY PRICE IN THOUSANDS OF DOLLARS
= SINGLE-USER SUPERCOMPUTERS
= WORKSTATIONS
= MINI-SUPERCOMPUTERS
= SUPERMINICOMPUTERS
= SUPERCOMPUTERS
= COMPILED CODE
= HANDCODED
TITAN FEATURES
PROCESSOR
• 64 MFLOPS peak performance
• 6 to 12 MFLOPS 64-bit Compiled Linpack
• Hardware scatter/gather, merge, and compress
• 1,024 vector registers of length 32
• 64 MIPS integer performance
• 8 to 128 MB of RAM
GRAPHICS
• 19 inch, 1280 by 1024 display
• Up to 72 image planes—organized as 24 double-buffered color planes with 8 additional overlay/control planes, plus a 16-bit Z-buffer
• 200,000 Gouraud-shaded triangles per second
I/O
• 380 MB to 1.1 GB internal disk storage, up to six external SMD drives
Performance data co
Jack J. Dongarra, Ar
Gordon Bell, Nationa
Pricing from manufa
© Ardent Computer C

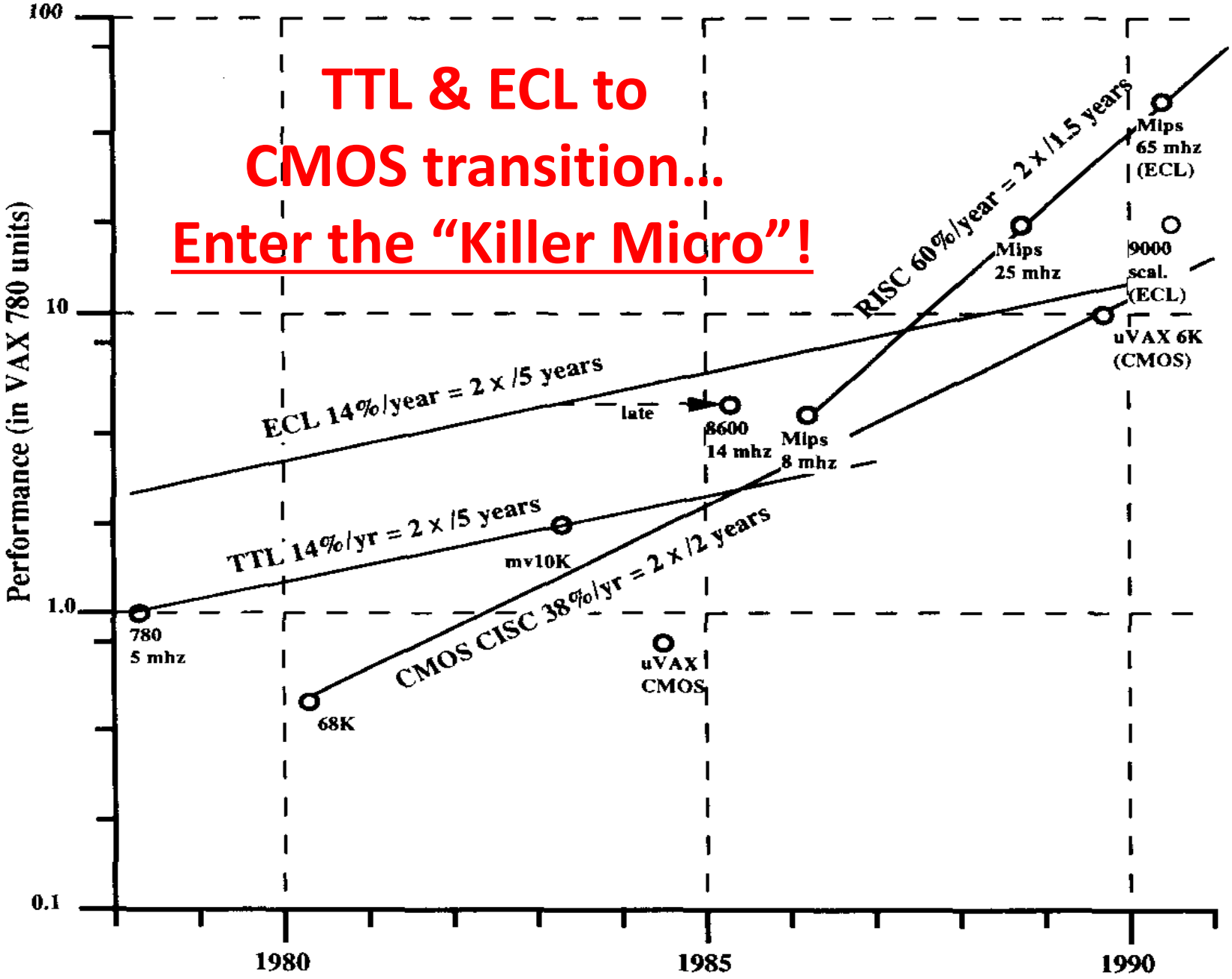


TITAN FEATUR

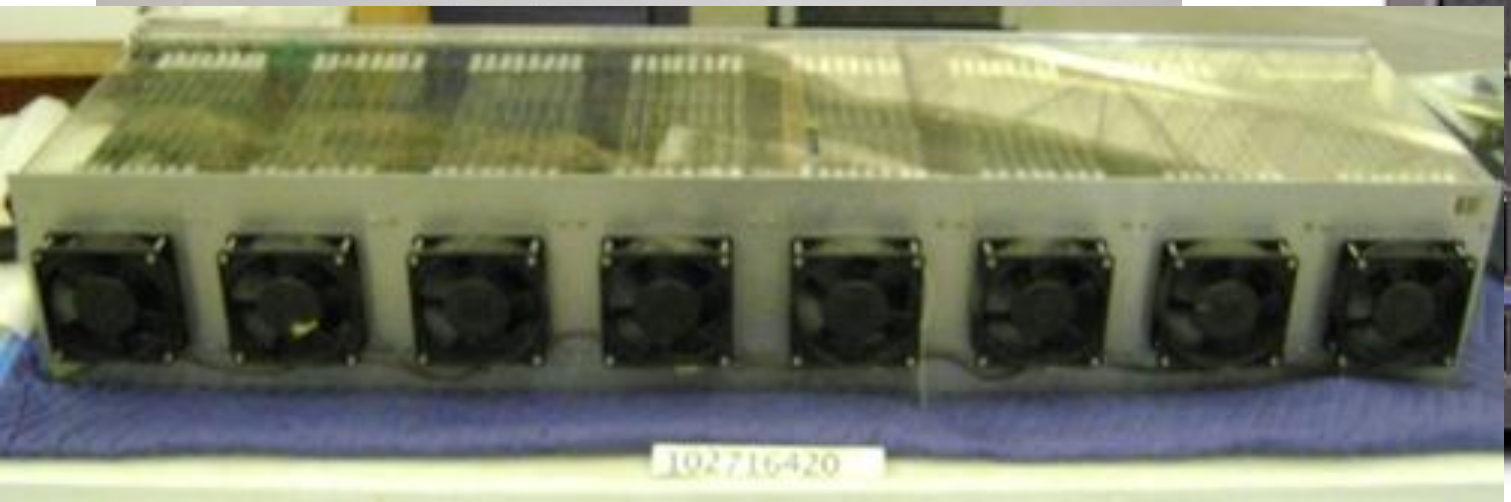
PROCESSOR

- 64 MFLOPS peak performance
- 6 to 12 MFLOPS 64-bit Compiled Linpa
- Hardware scatter/gather, merge, and c
- 1,024 vector registers of length 32
- 64 MIPS integer performance
- 8 to 128 MB of RAM

GRAPHICS


TTL & ECL to
CMOS transition...
Enter the "Killer Micro"!
Performance (in VAX 780 units)
100
10
1.0
0.1
1980
1985
1990
RISC 60%/year = 2 x /1.5 years
ECL 14%/year = 2 x /5 years
TTL 14%/yr = 2 x /5 years
CMOS CISC 38%/yr = 2 x /2 years
780
5 mhz
mv10K
68K
late
8600
14 mhz
uVAX
CMOS
Mips
8 mhz
Mips
25 mhz
Mips
65 mhz
(ECL)
9000
scal.
(ECL)
uVAX 6K
(CMOS)

Caltech Cosmic Cube
8 node prototype (‘82) & 64 node ‘83
Intel iPSC 64 Personal Supercomputer ‘85



Bell Prize for Parallelism, July 1987

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.

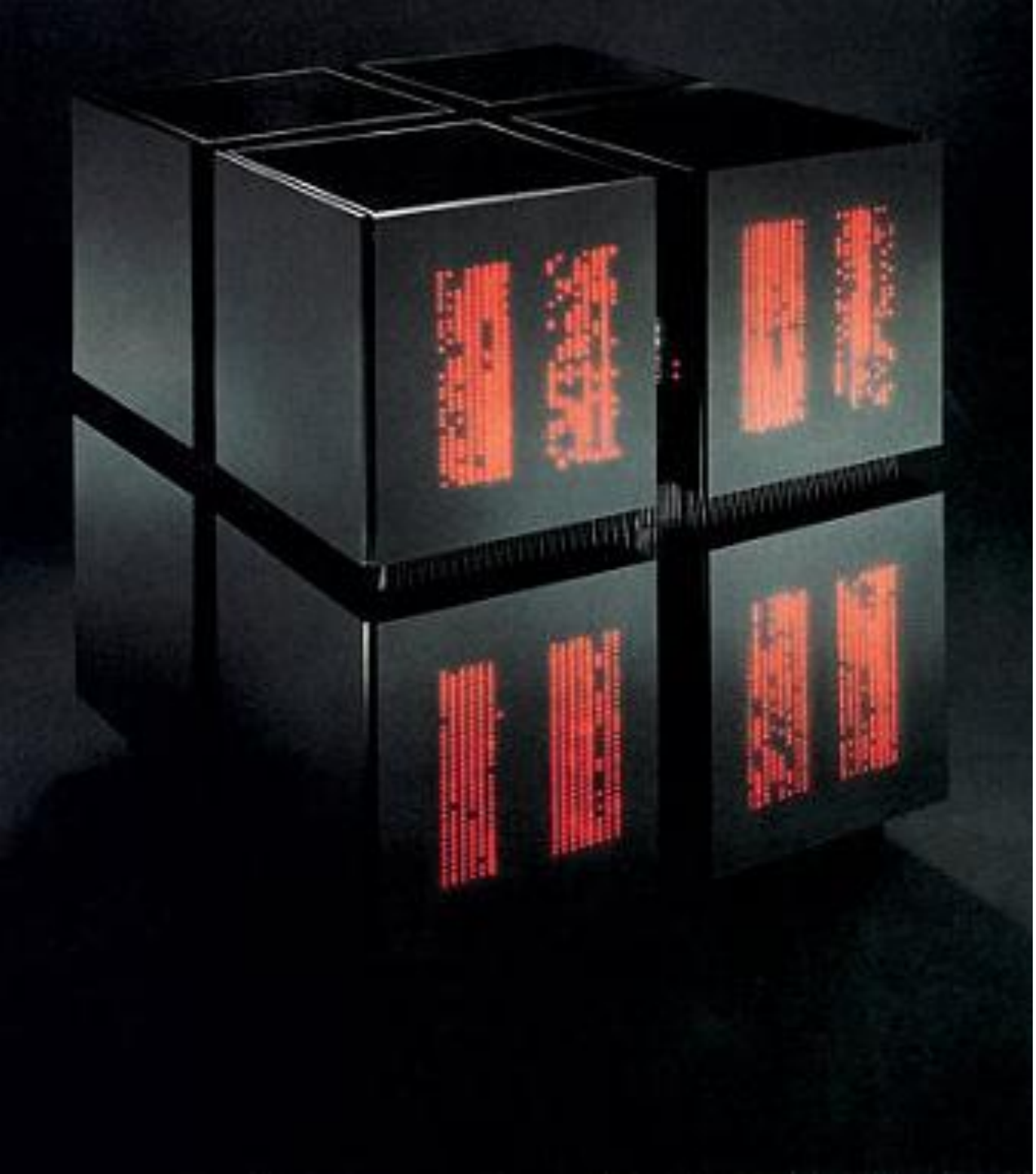
To qualify for the 1987 awards, candidates must submit documentation of their results by Dec. 1. The winners will be announced in the March 1988 issue. This year's judges are Alan Karp of IBM's Palo Alto Scientific Center, Jack Dongarra of Argonne National Laboratory, and Ken Kennedy of Rice University.

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.

Alan Karp:
Offers $100 for a program with 200 X parallelism by 1995.

Bell, 1987 goals:
10 X by 1992
100 X by 1997

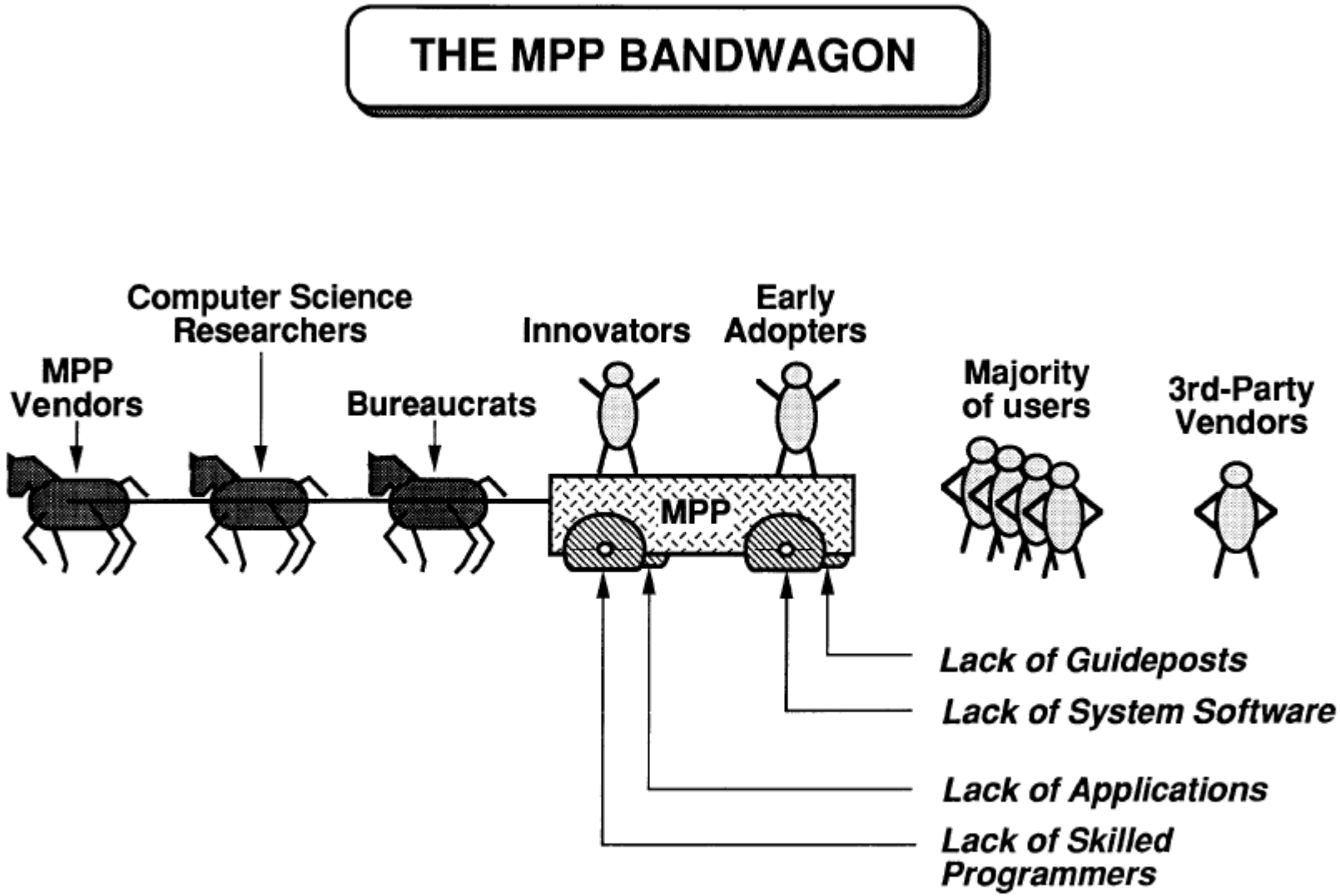
Researcher claims:
1 million X by 2002



Thinking Machines CM-2 1990

64K PE
SIMD
14 Gflops

Worlton view c 1991



Sandia Touchstone Delta c1992

Intel Touchstone
Delta 1992
30-120 Gflops
8.2- Gbytes
512-2048
computers

10.8 Million

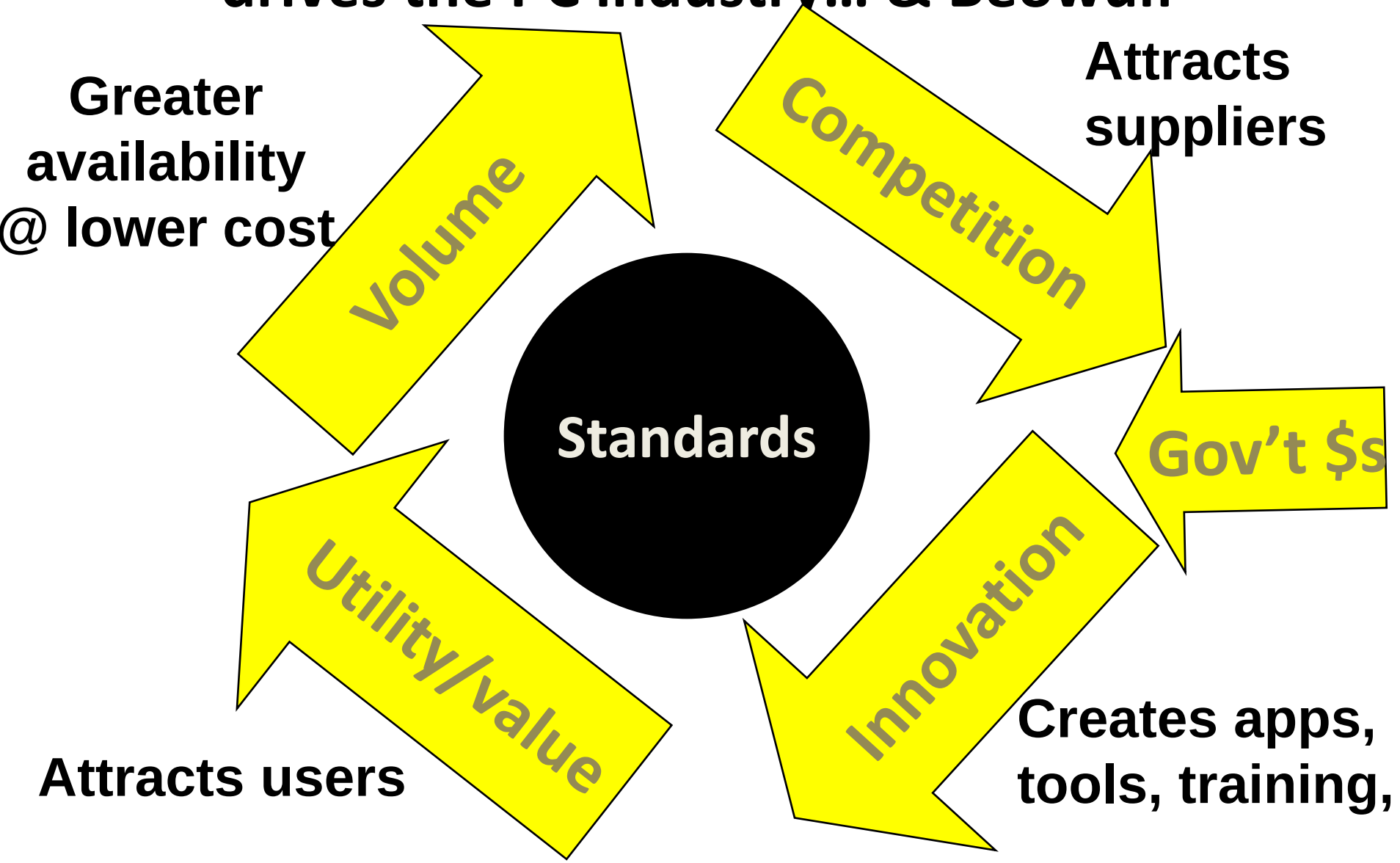
Beowulf: Computer Cluster by Don Becker & Tom Sterling, NASA 1994

BSD, LINUX, Solaris, and Windows Support for MPI and PVM

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

The Virtuous Economic Cycle drives the PC industry… & Beowulf

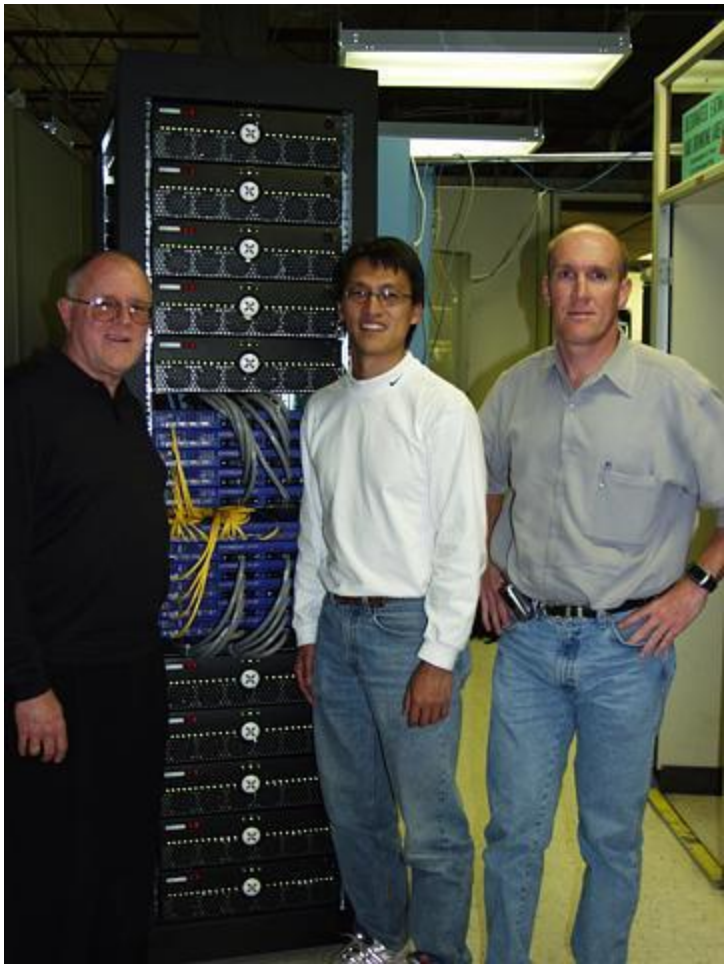


Lost: The search for parallelism c1983-1997

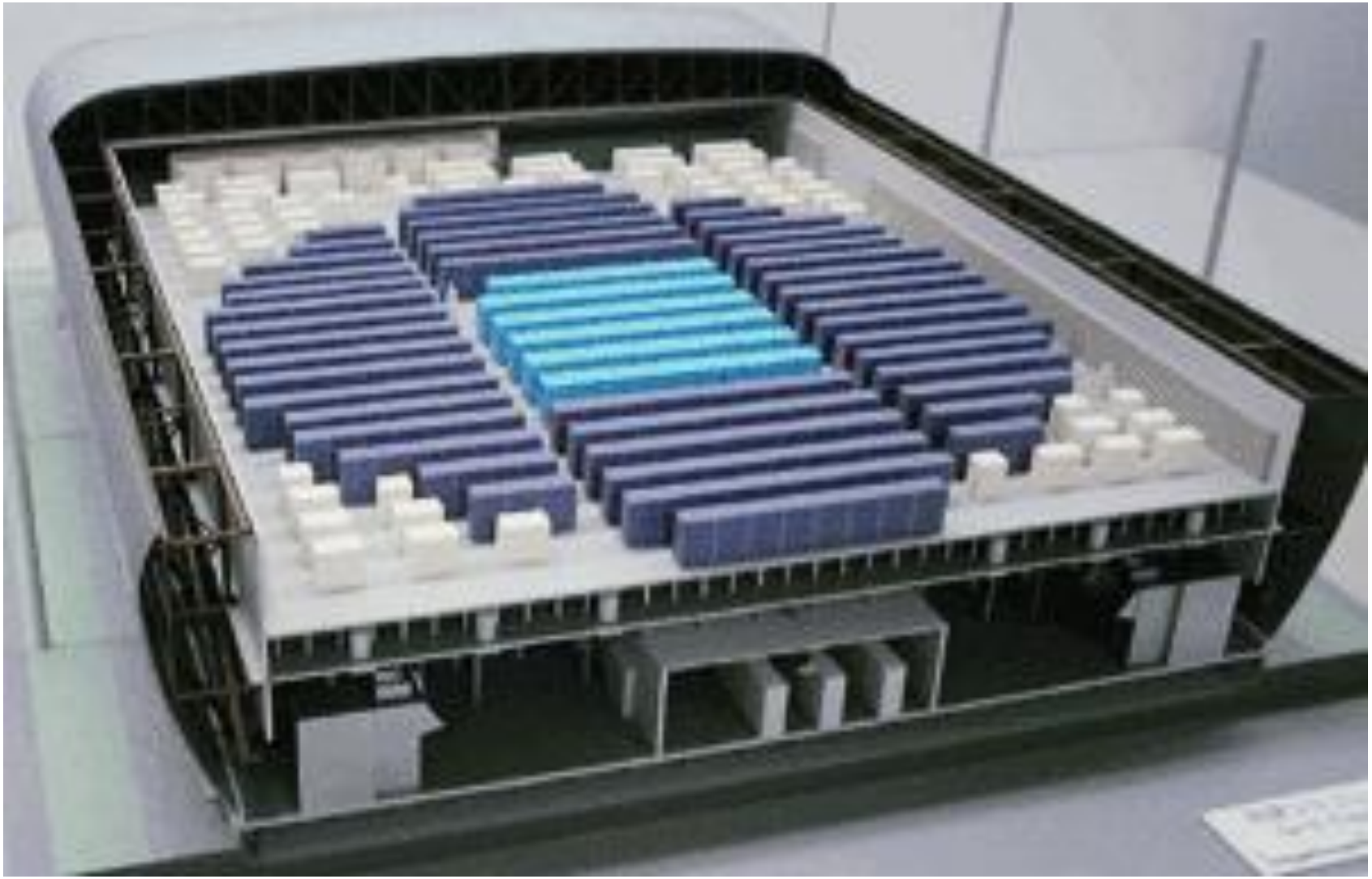
DOE and DARPA Adv. Sci Comp. Initiative

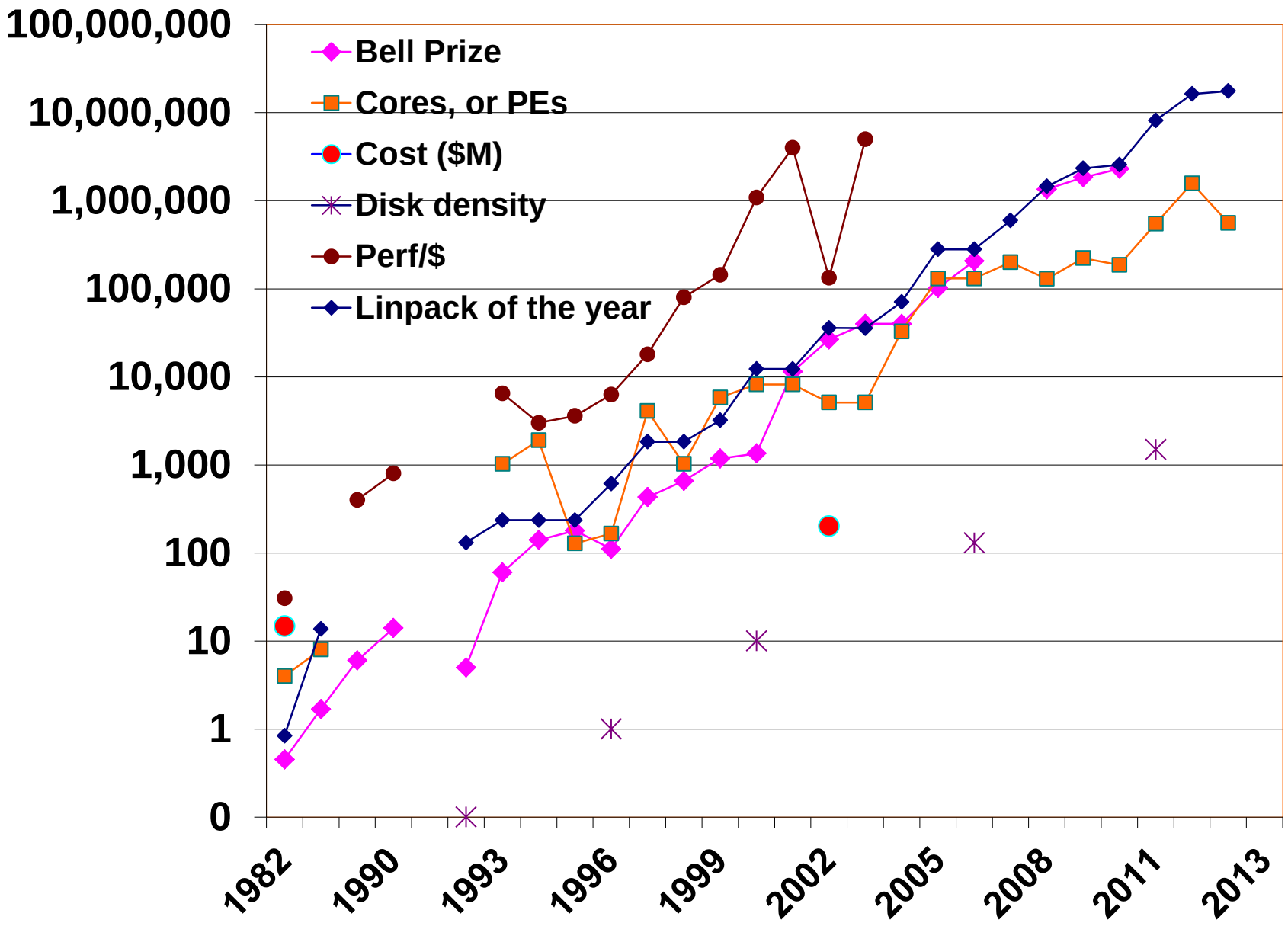
- ACRI *French-Italian program*
- Alliant *Proprietary Crayette*
- American Supercomputer
- Ametek
- Applied Dynamics
- Astronautics
- BBN
- CDC >ETA *ECL transition*
- Cogent
- Convex > HP
- Cray Computer > SRC *GaAs flaw*
- Cray Research > SGI > Cray *Manage*
- Culler-Harris
- Culler Scientific *Vapor…*
- Cydrome *VLIW*
- Dana/Ardent/Stellar/Stardent
- Denelcor
- Encore
- Elexsi
- ETA Systems *aka CDC;Amdahl flaw*
- Evans and Sutherland Computer
- Exa
- Flexible
- Floating Point Systems *SUN savior*
- Galaxy YH-1
- Goodyear Aerospace MPP *SIMD*
- 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 *Early RISC*
- Ridge
- Saxpy
- Scientific Computer Systems (SCS)
- Soviet Supercomputers
- Supertek
- Supercomputer Systems
- Suprenum
- Tera > Cray Company
- Thinking Machines
- Vitesse Electronics
- Wavetracer *SIMD*

First Clusters RLX Startup c2002 Defines blade…



Japanese Earth Simulator (NEC)
2002 35 Teraflops 5,000 vector processor




100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
0
Bell Prize
Cores, or PEs
Cost ($M)
Disk density
Perf/$
Linpack of the year
1982
1990
1993
1996
1999
2002
2005
2008
2011
2013

30+ year history

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 with Earth Simulator! 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…*
13. *DARPA HP 2010 project 5 >3 (Cray, IBM, SUN) > 1 winner*

Supercomputer Evolution

End

Performance

Performance is all about Parallelism!

Parallelism is all about scalability!

Three Scalabilities

Size scalable **computers are designed from a few components, with no bottleneck component.**

Generation scalable **computers can be implemented with the next generation technology with No rewrite/recompile**

Problem x machine scalability **- ability of a problem, algorithm, or program to exist at a range of sizes so that it can be efficiently or effectively used on a *given*, scalable computer. Run at affordable size, not largest size.**

Problem x machine space => run time: problem scale, machine scale (#p), run time, implies speedup and efficiency,

Application Taxonomy

If real rich
then SMP clusters
else PC Clusters

Scientific

General purpose, non-parallelizable codes (PCs have it!)

Vectorizable & //able (Supers & all SMPs)

Hand tuned, one-of MPP course grain.

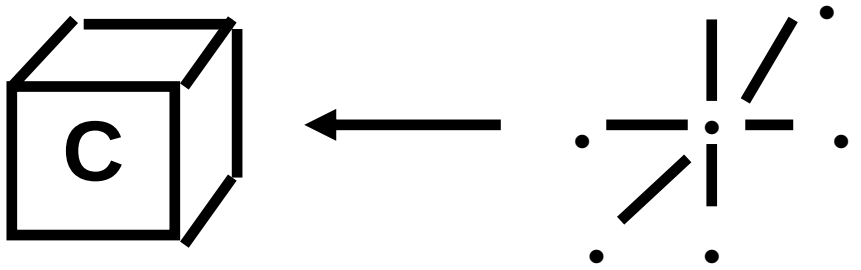
Ensembles & parameter sweeps → MPP embarrassingly // (Clusters of anythings)

Commercial Cloud

Database
Database/TP
Web Host
Stream Audio/Video

If real rich
then IBM Mainframes or large SMPs
else PC Clusters

Scalable Problems



Distributed computing node

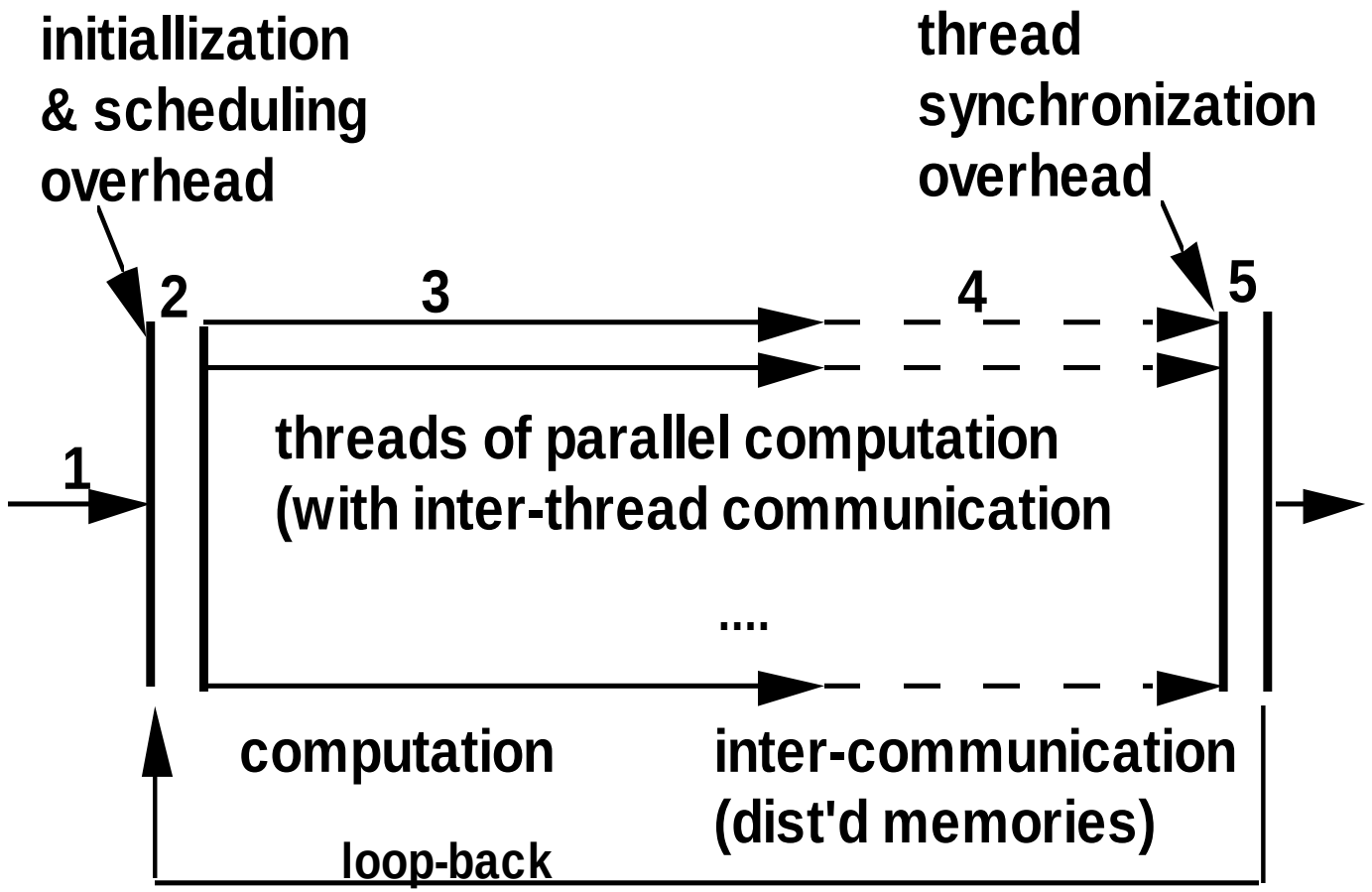
Computational grid point array

7 ops for average

6 values to communicate per time step

Is speed limited by: memory size, processing speed, interconnect bandwidth, message passing overhead time, or synchronization time

Parallel Computation: Granularity



Make long grains: unrolling, virtual processors, inf. //,

Amdahl's law… the limit

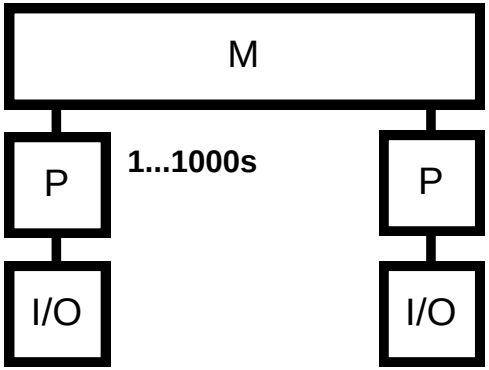
- **If w_1 work is done at speed s_1 and w_2 at speed s_2, the average speed s is $(w_1+w_2)/(w_1/s_1 + w_2/s_2)$**
  - **This is just the total work divided by the total time**
- **For example, if $w_1 = 9$, $w_2 = 1$, $s_1 = 100$, and $s_2 = 1$ then $s = 10/1.09 \cong 9$ (speed)**
  - **This is obviously not the average of s_1 and s_2**

Amdahl, Gene M, "Validity of the single processor approach to achieving large scale computing capabilities", Proc. SJCC, AFIPS Press, 1967

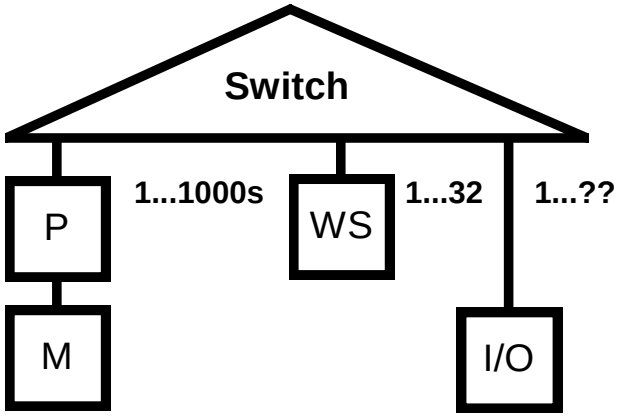
Courtesy of Burton Smith, Microsoft

Multiprocessors ° Multicomputers

Is it general, i.e., will it process an arbitrary workload?

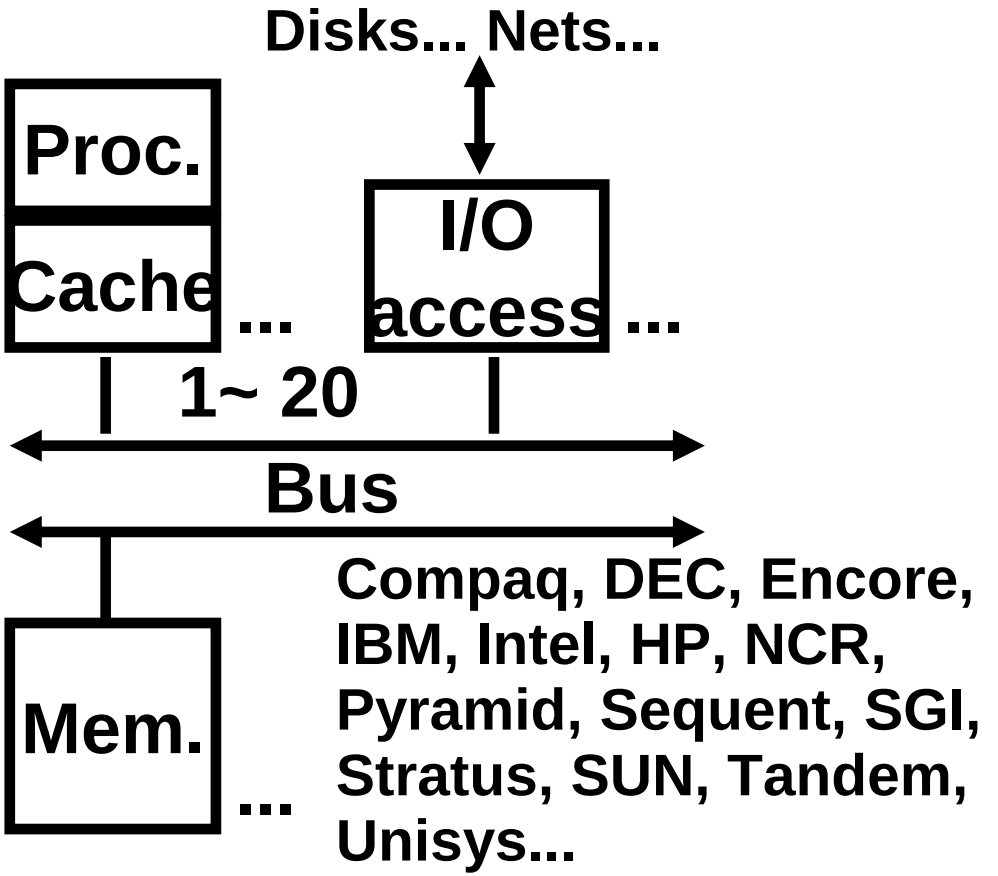


- 1 *large* address space
- direct data access
- = shared memory mP
- coherent memory
- "1" copy of OS kernel
- 1 work queue; 1 set of fungible resources
- automatic migration of data to processor
- non-trivial, related to mP
- port & tune for //

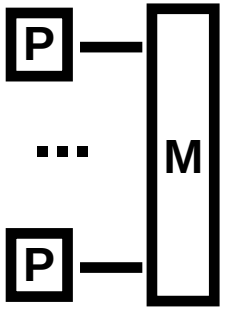


- n *small* address spaces
- message passing
- *simulates an mP*
- non-coherent memory
- n copies of dist'd OS
- work is bound to a WS; idle resources may remain
- software moves data around
- non-trivial, related to WS
- rethink, rewrite, & tune for //

The "Multi" (limited scalable mP) "mainline": PC, WS, & Servers

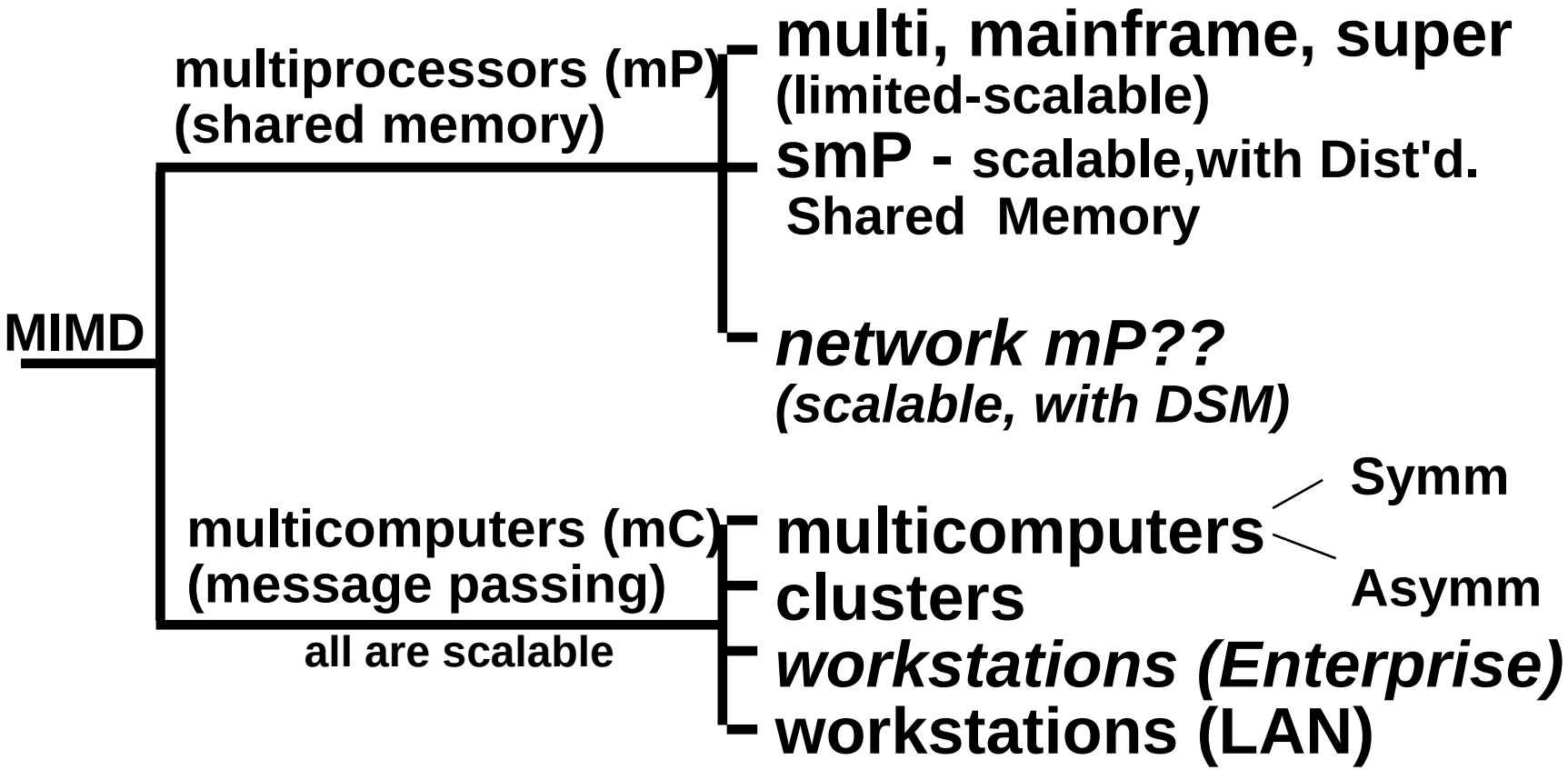


Physical structure



Program(mer) view

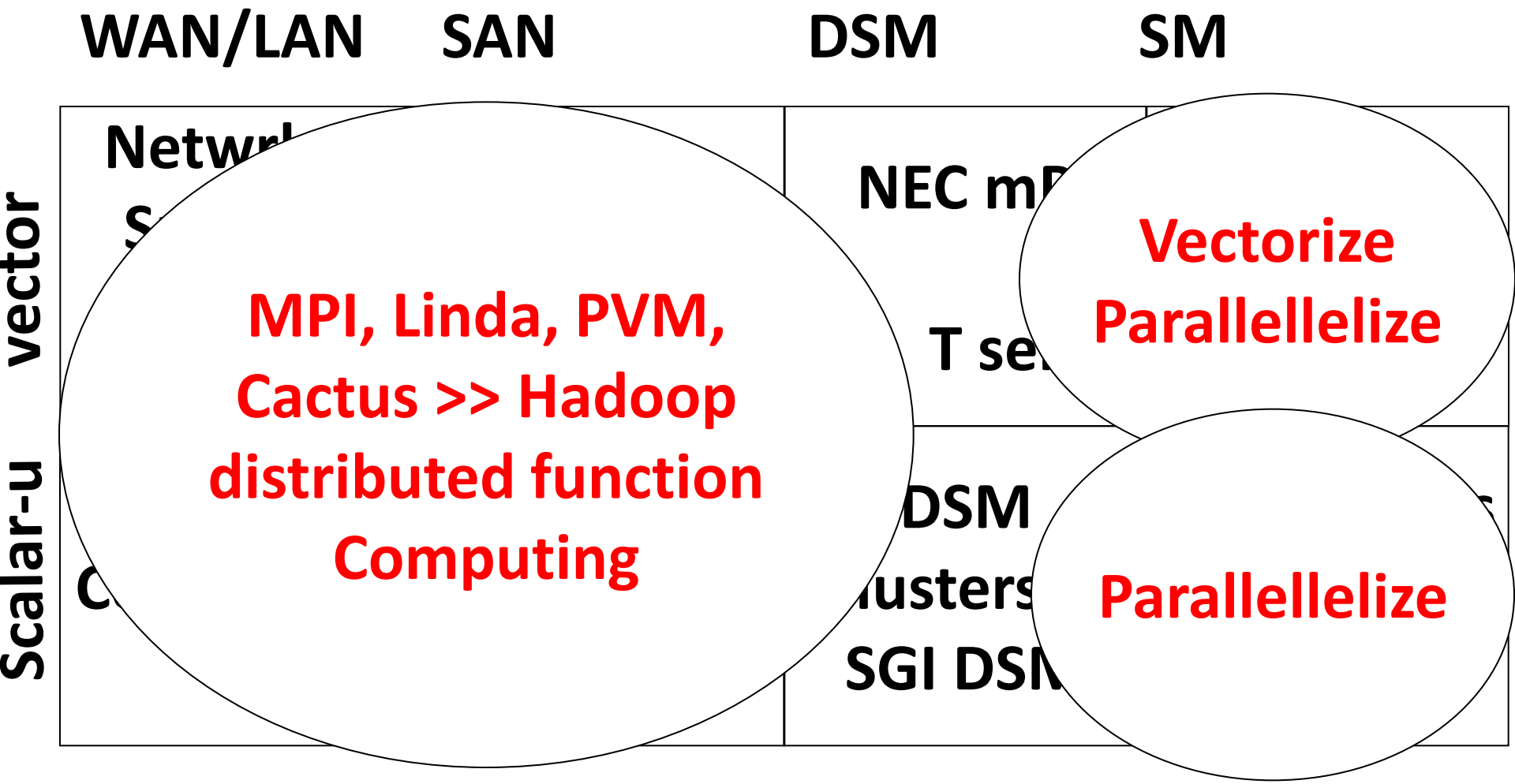
The Architectural Alternatives for scalablity & high performance



Technical computer types: Pick of: 4 nodes, 2-3 interconnects

	SAN	DSM	SMP
vector	Fujitsu Hitachi	NEC	NEC super Cray ???
Scalar-u	IBM ?PC? SGI cluster Beowulf	SGI DSM T3 HP?	HP IBM Intel SUN plain old PCs

Technical computer types



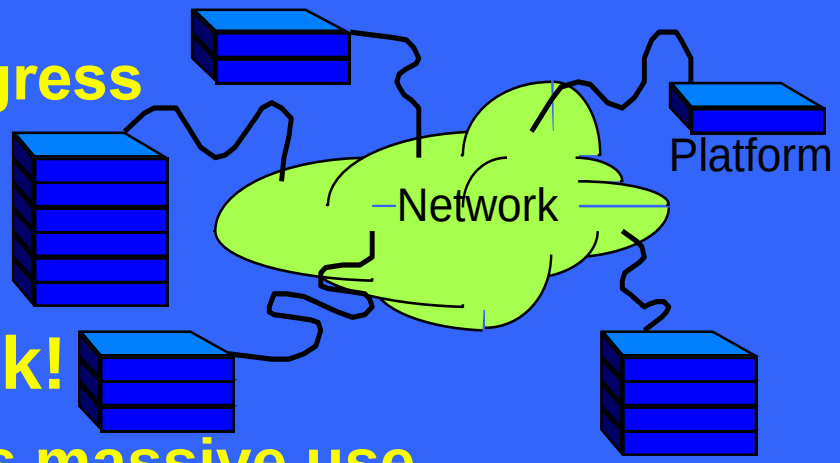
1994: Computers will All be Scalables

Thesis: SNAP: Scalable Networks as Platforms

- upsize from desktop to world-scale computer
- based on a few standard components

Because:

- Moore's law: exponential progress
- standards & commodities
- stratification and competition



When: Sooner than you think!

- massive standardization gives massive use
- economic forces are enormous

End Performance, parallelism, and scalability

- **End**

Performance– measuring

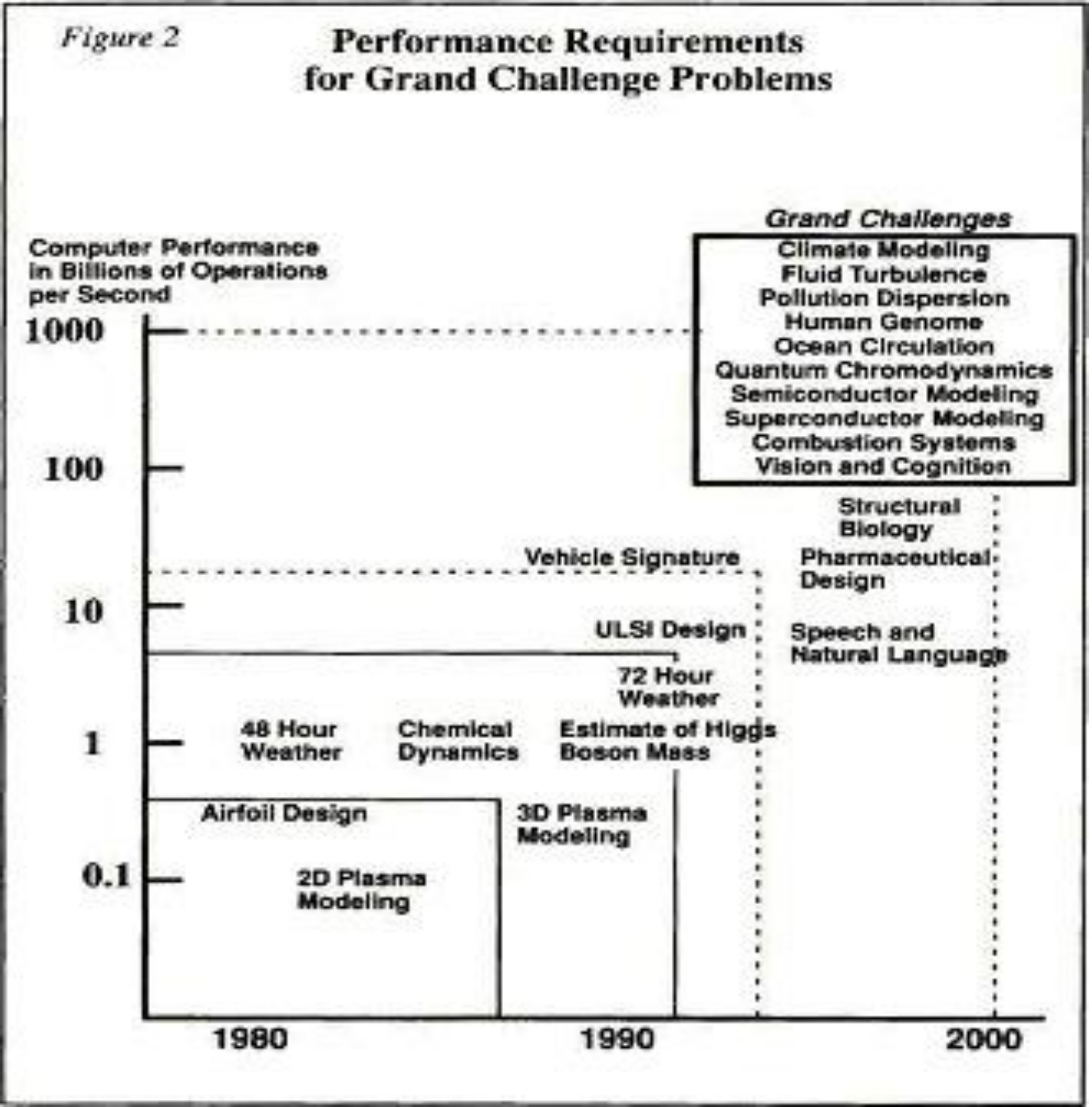
Grand Challenges

Benchmarks

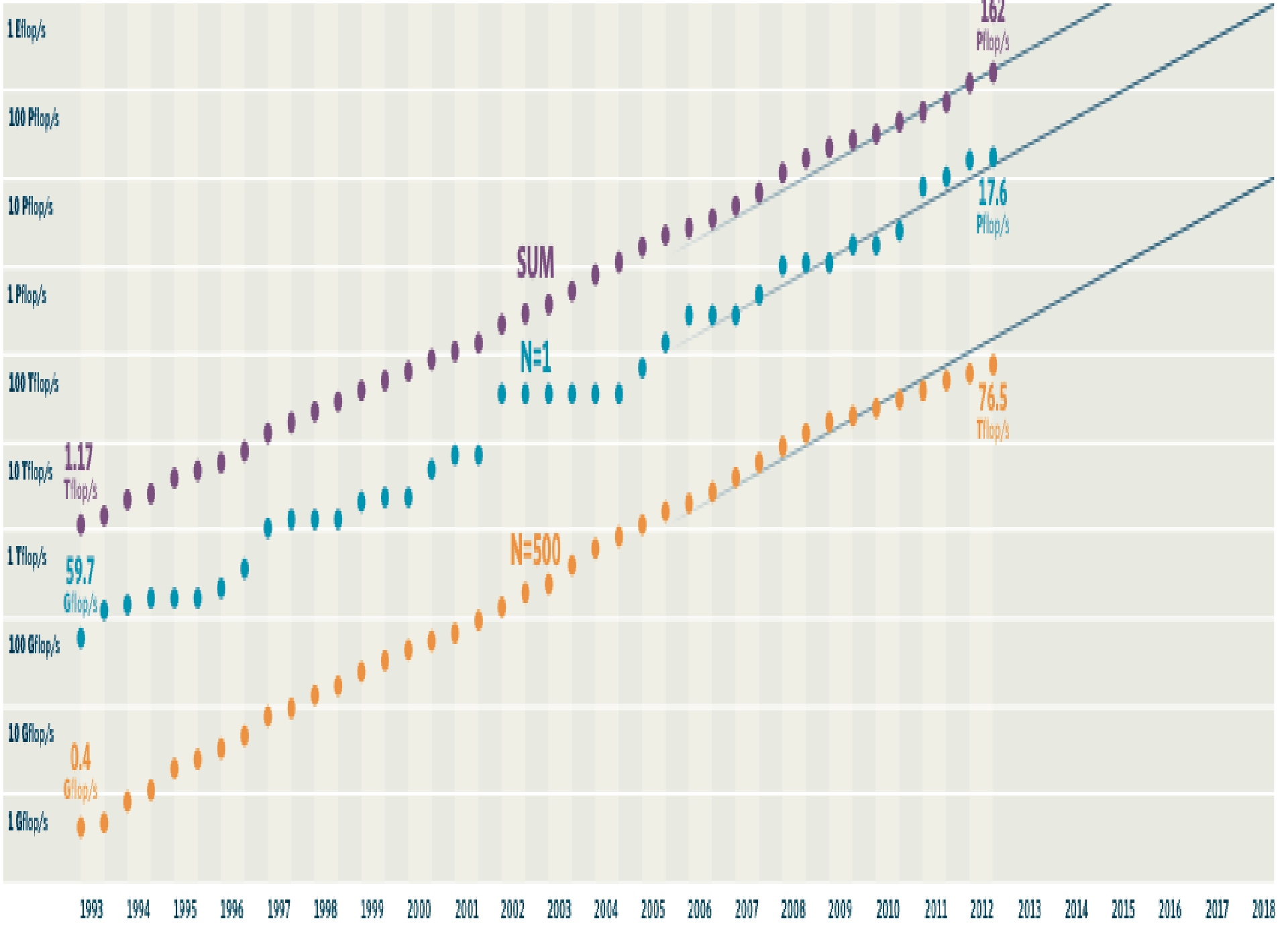
- LINPACK
- Graph500
- Green500

Bell Prize rewarding parallelism

UC/Berkeley Kernel methodology for estimating application performance



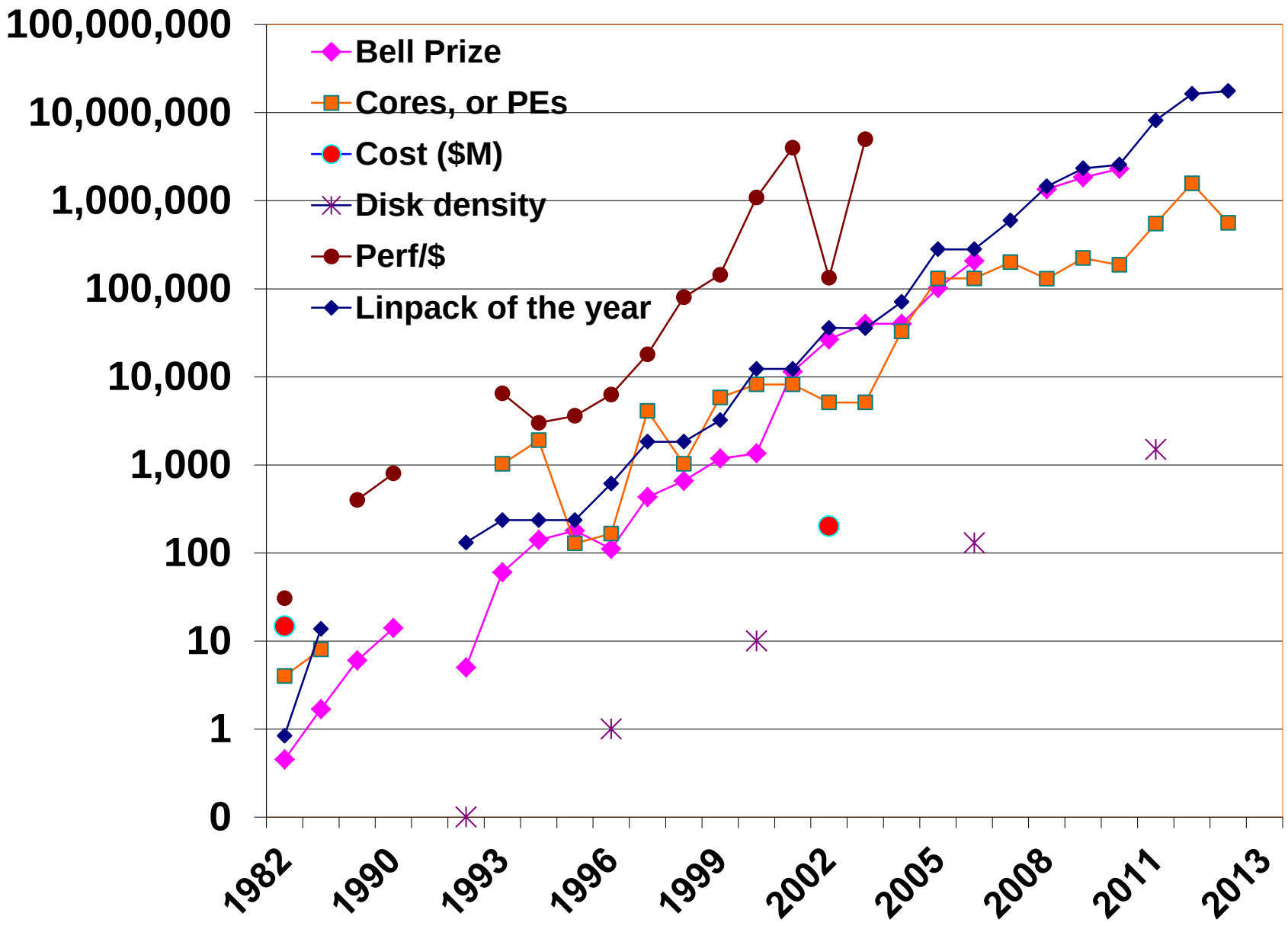
Grand Challenge problems c1992


1 Eflop/s
100 Pflop/s
10 Pflop/s
1 Pflop/s
100 Tflop/s
10 Tflop/s
1 Tflop/s
100 Gflop/s
10 Gflop/s
1 Gflop/s
SUM
N=1
N=500
1.17
Tflop/s
59.7
Gflop/s
0.4
Gflop/s
162
Pflop/s
17.6
Pflop/s
76.5
Tflop/s
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018

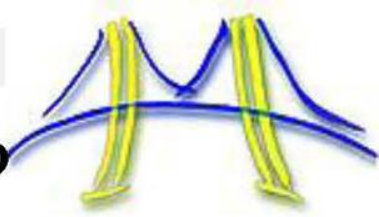
Graph500 Benchmark

Problem class	Scale	Edge factor	Approx. storage size in TB
Toy (level 10)	26	16	0.0172
Mini (level 11)	29	16	0.1374
Small (level 12)	32	16	1.0995
Medium (level 13)	36	16	17.5922
Large (level 14)	39	16	140.7375
Huge (level 15)	42	16	1125.8999

Data set sizes ("Scale") of the Graph 500 benchmark.


100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1
0
Bell Prize
Cores, or PEs
Cost ($M)
Disk density
Perf/$
Linpack of the year
1982
1990
1993
1996
1999
2002
2005
2008
2011
2013

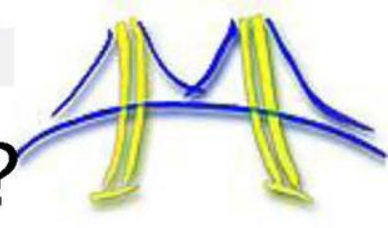
How do we describe apps and kernels?



University of California, Berkeley Applications can be composed from a set of standard kernels.

Courtesy of David Patterson, et al

How do we describe apps and kernels?



- **Observation: Use Dwarfs. Dwarfs are of 2 types**

Algorithms in the dwarfs can either be implemented as:

- Compact parallel computations within a traditional ***library***
- Compute/communicate pattern implemented as ***framework***

Libraries

- Dense matrices
- Sparse matrices
- Spectral
- Combinational
- Finite state machines

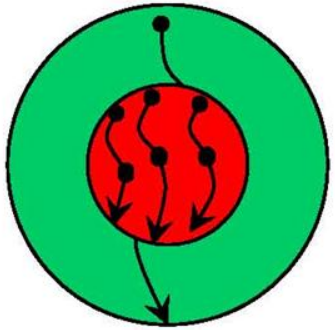
Patterns/Frameworks

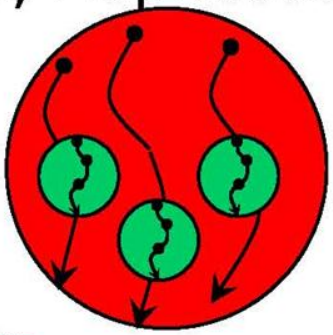
- MapReduce
- Graph traversal, graphical models
- Dynamic programming
- Backtracking/B&B
- N-Body
- (Un) Structured Grid

- Computations may be viewed a multiple levels: e.g., an FFT library may be built by instantiating a Map-Reduce framework, mapping 1D FFTs and then transposing (generalize reduce)

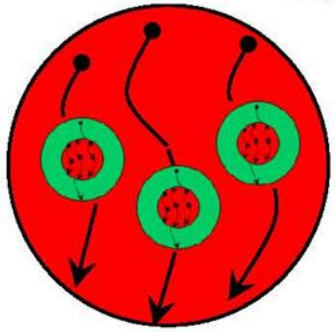
Composing dwarfs to build apps

- **Any** parallel application of **arbitrary complexity** may be built by composing parallel  and serial  components
  - Serial code invoking parallel libraries, e.g., FFT, matrix ops.,…
  - Parallel patterns with serial plug-ins e.g., MapReduce





- Composition is hierarchical

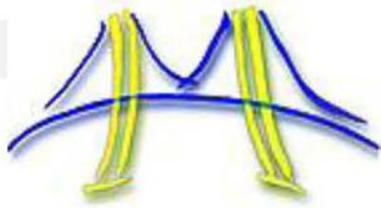


Dwarf Popularity (Red Hot → Blue Cool)

Embed SPEC DB Games ML HPC

1 Finite State Mach.
2 Combinational
3 Graph Traversal
4 Structured Grid
5 Dense Matrix
6 Sparse Matrix
7 Spectral (FFT)
8 Dynamic Prog
9 N-Body
10 MapReduce
11 Backtrack/ B&B
12 Graphical Models
13 Unstructured Grid

13 Dwarfs (so far)

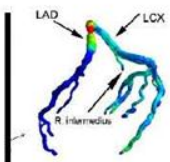


1. Finite State Machine
2. Combinational Logic
3. Graph Traversal
4. Structured Grids
5. Dense Linear Algebra
6. Sparse Linear Algebra
7. Spectral Methods (FFT)
8. Dynamic Programming
9. N-Body Methods
10. MapReduce
11. Back-track/ Branch & Bound
12. Graphical Model Inference
13. Unstructured Grids

- **Claim: parallel arch., lang., compiler … must do at least these well to do future parallel apps well**
- Note: MapReduce is embarrassingly parallel; perhaps FSM is embarrassingly sequential?

Dwarf Popularity (Red Hot → Blue Cool)

- How do compelling apps relate to 13 dwarfs?

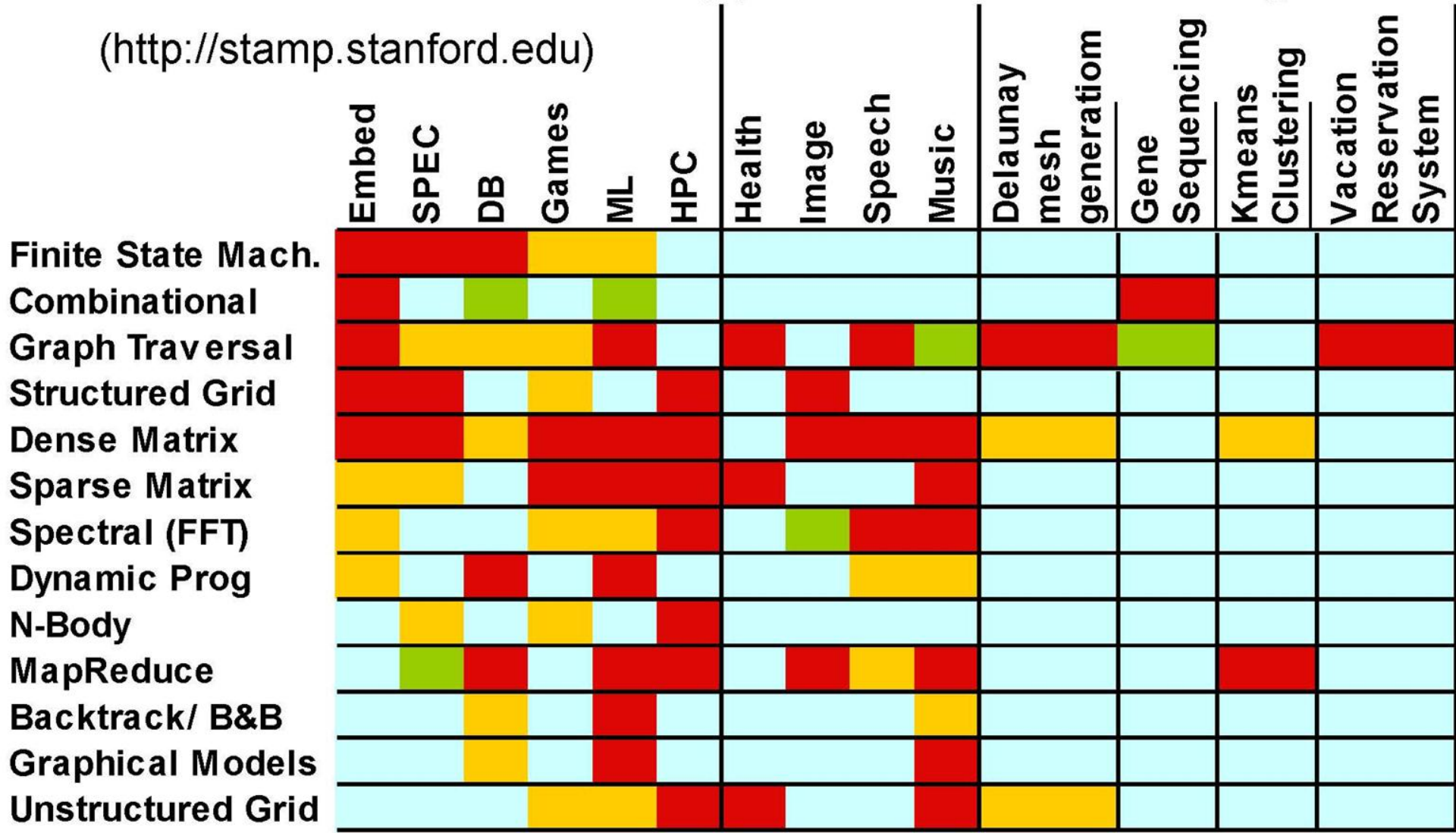


	Embed	SPEC	DB	Games	ML	HPC	Health	Image	Speech	Music
Finite State Mach.										
Combinational										
Graph Traversal										
Structured Grid										
Dense Matrix										
Sparse Matrix										
Spectral (FFT)										
Dynamic Prog										
N-Body										
MapReduce										
Backtrack/ B&B										
Graphical Models										
Unstructured Grid										

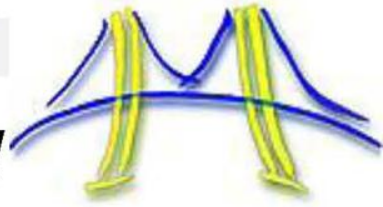
Dwarf Popularity (Red Hot → Blue Cool)

- Stanford Transactional Apps for Multi Processing?

(http://stamp.stanford.edu)



HW features supporting Parallel SW



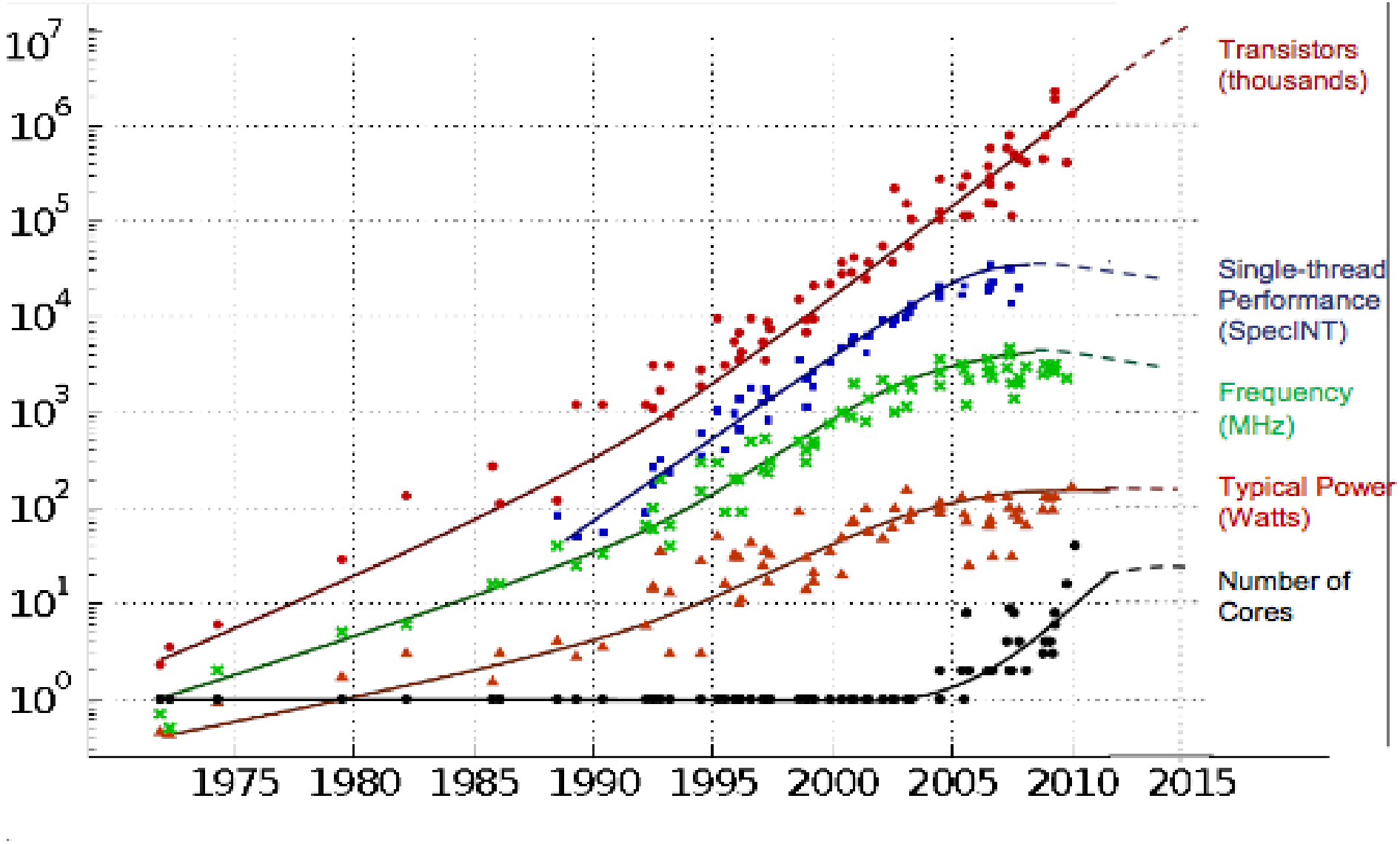
- Want Composable Primitives, Not Packaged Solutions
  - Transactional Memory is usually a Packaged Solution
- Partitions
- Fast Barrier Synchronization & Atomic Fetch-and-Op
- Active messages plus user-level event handling
  - Used by parallel language runtimes to provide fast communication, synchronization, thread scheduling
- Configurable Memory Hierarchy (Cell v. Clovertown)
  - Can configure on-chip memory as cache or local store
  - Programmable DMA to move data without occupying CPU
  - Cache coherence: Mostly HW but SW handlers for complex cases
  - Hardware logging of memory writes to allow rollback

End Benchmarks

Top500 Computers

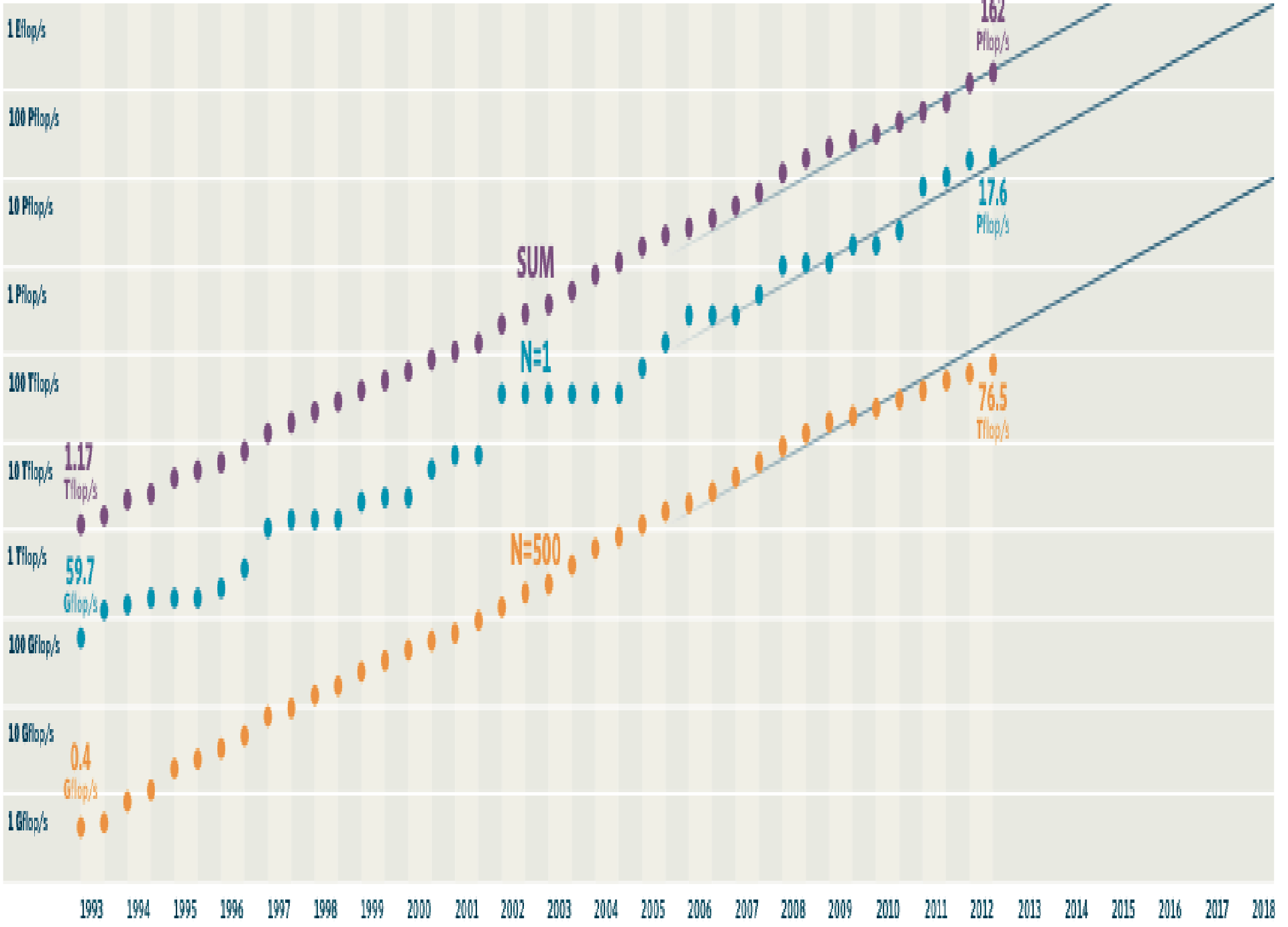
- Architectures
  - Fujitsu 2011 Fall
  - IBM 2012 Spring
  - Cray 2012 Fall– the emergence of NVIDIA
- Parallela
- Convey

Result: The End of Historic Scaling

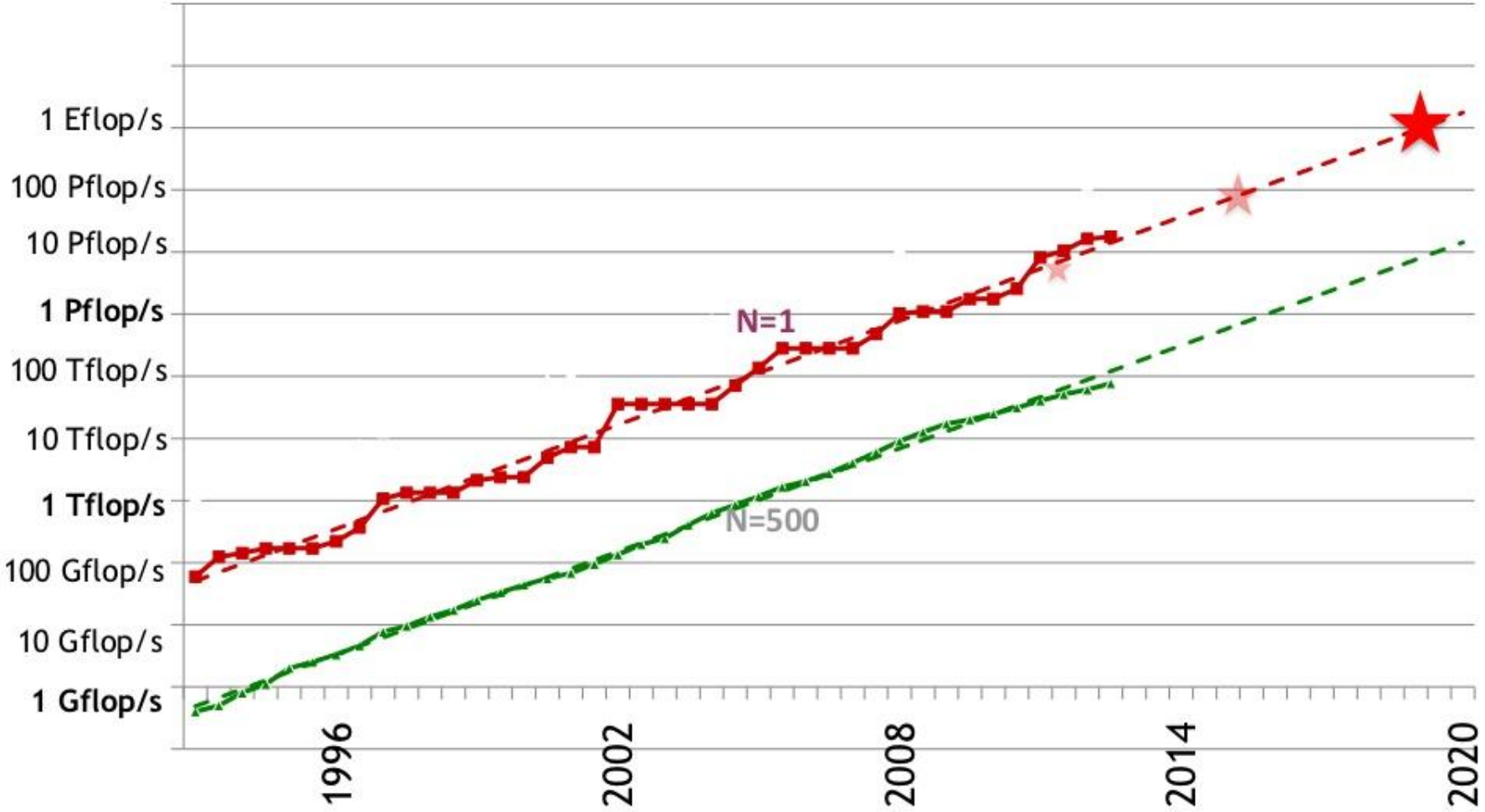


Original data collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond and C. Batten
Dotted line extrapolations by C. Moore

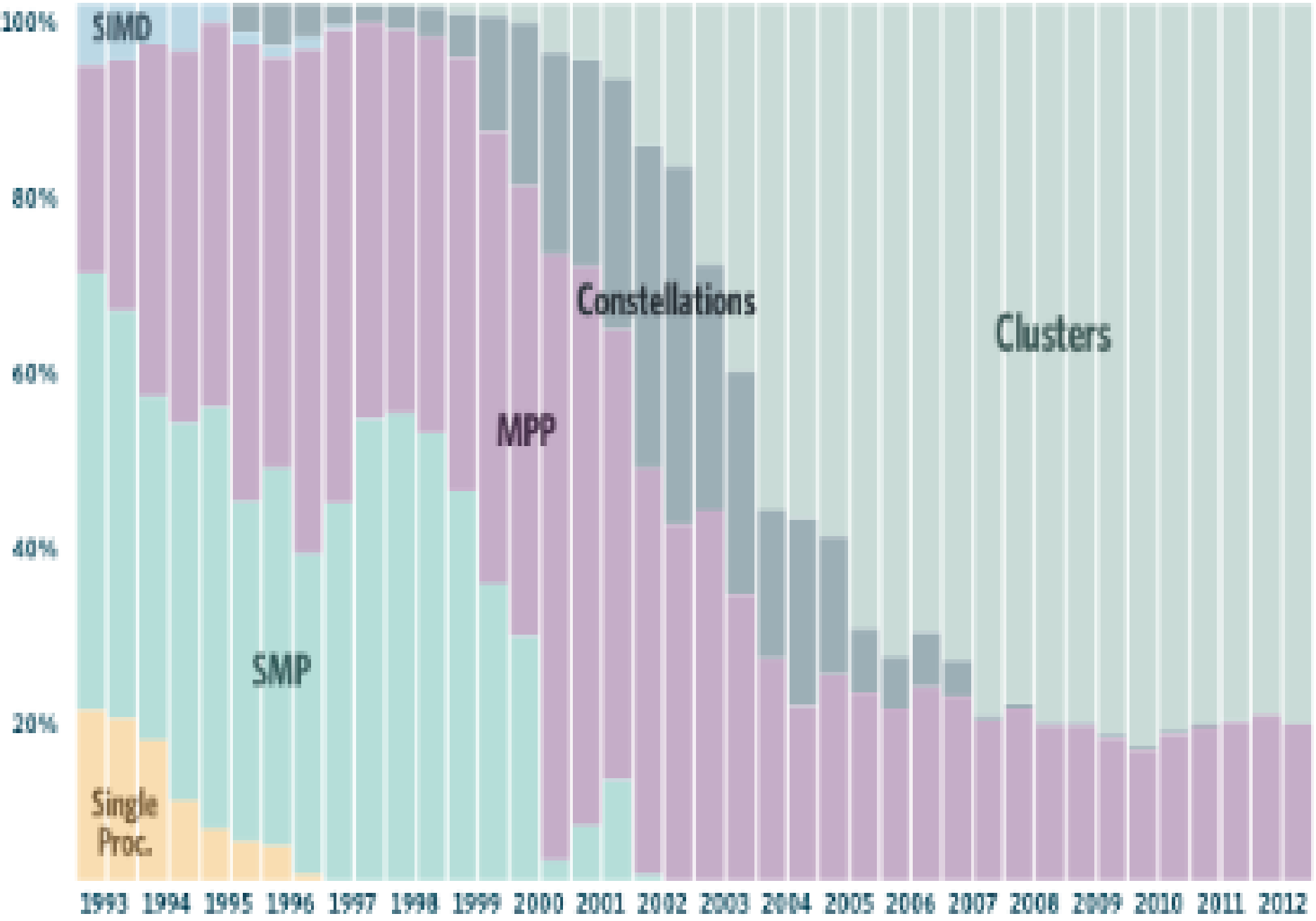
C Moore, *Data Processing in ExaScale-ClassComputer Systems,* Salishan, April 2011

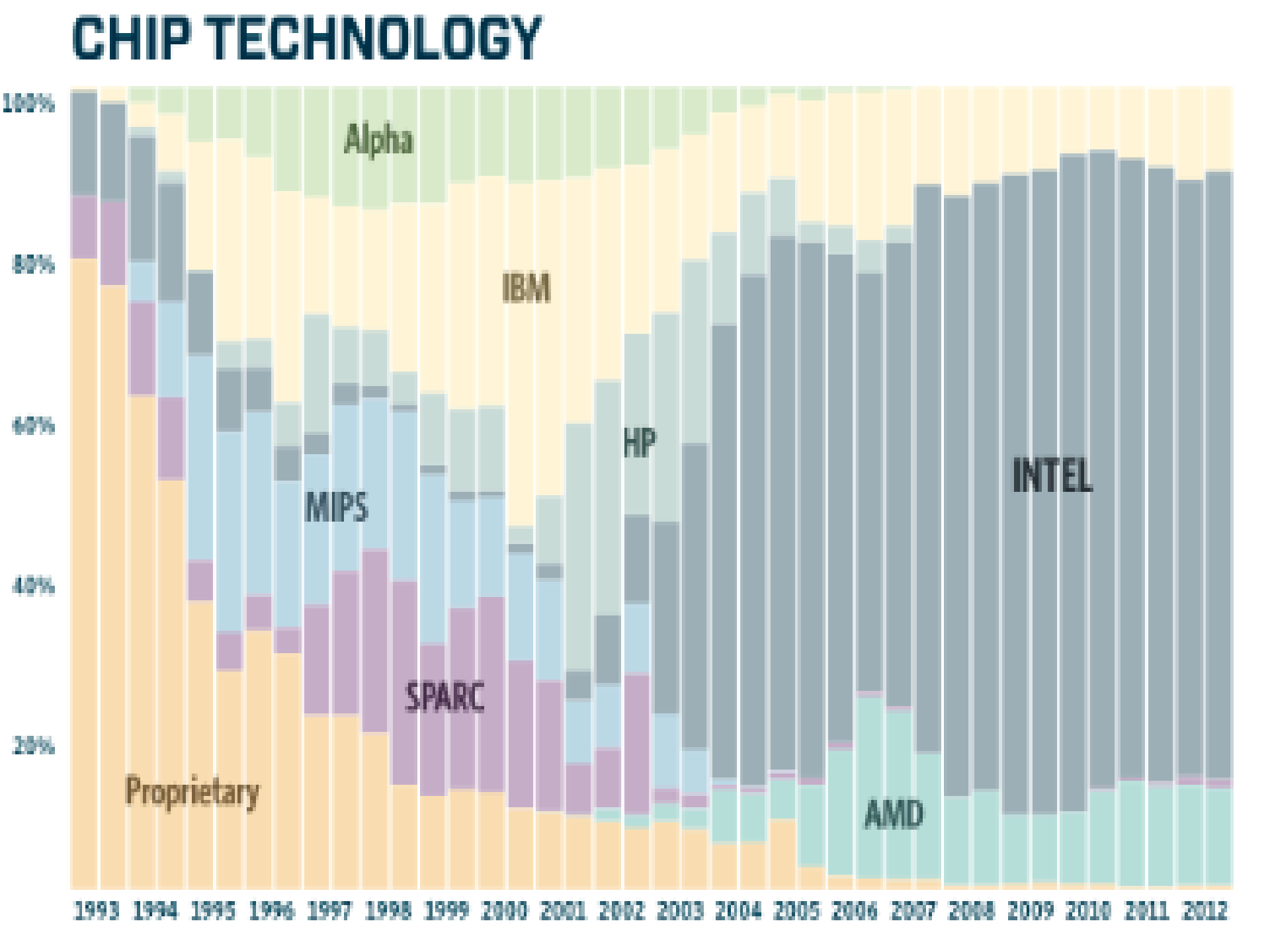

1 Eflop/s
100 Pflop/s
10 Pflop/s
1 Pflop/s
100 Tflop/s
10 Tflop/s
1 Tflop/s
100 Gflop/s
10 Gflop/s
1 Gflop/s
SUM
N=1
N=500
1.17
Tflop/s
59.7
Gflop/s
0.4
Gflop/s
162
Pflop/s
17.6
Pflop/s
76.5
Tflop/s
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018

Performance Development in Top500

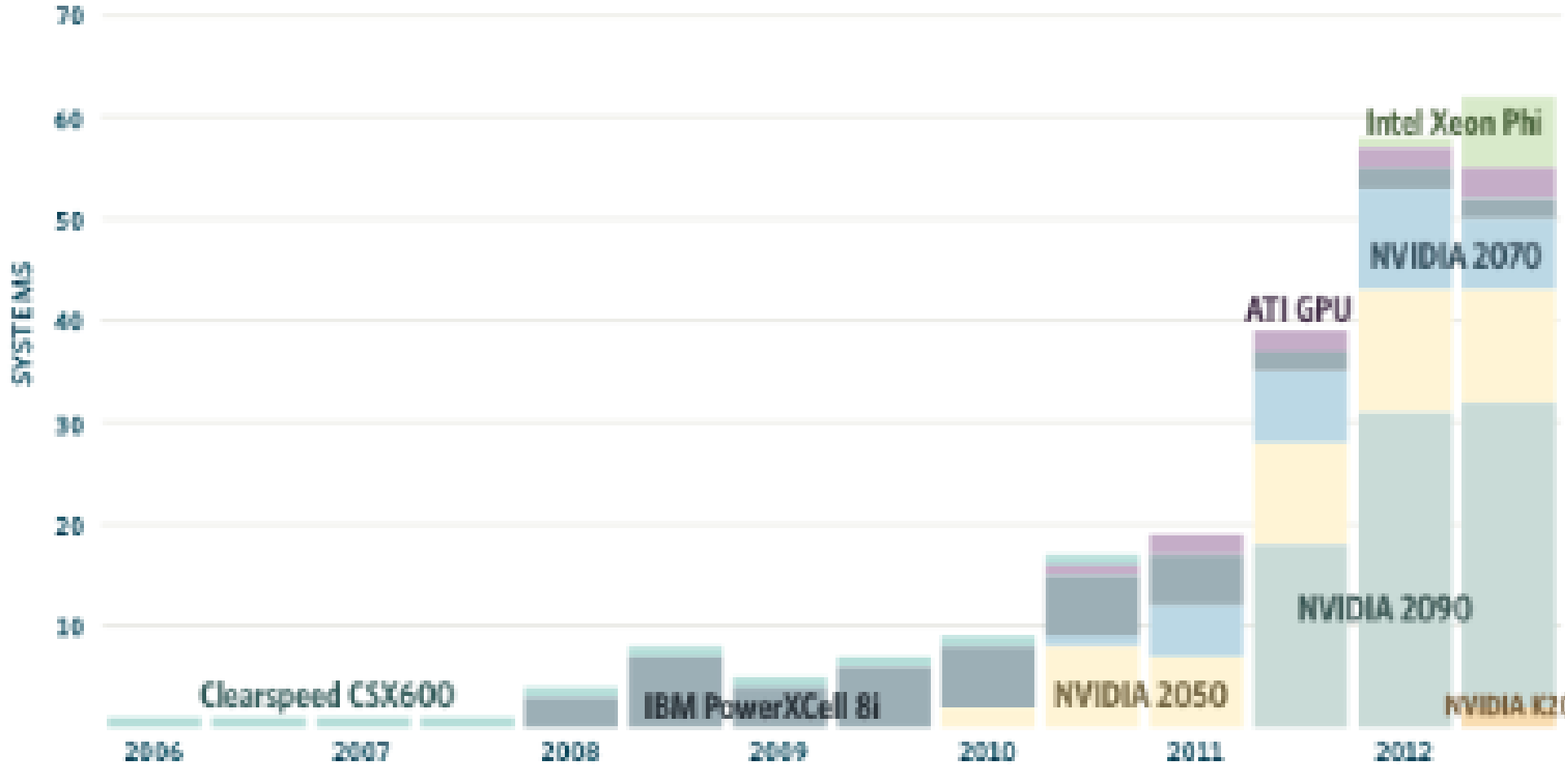


ARCHITECTURES




CHIP TECHNOLOGY
100%
80%
60%
40%
20%
Alpha
IBM
HP
INTEL
MIPS
SPARC
Proprietary
AMD
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012

ACCELERATORS/CO-PROCESSORS



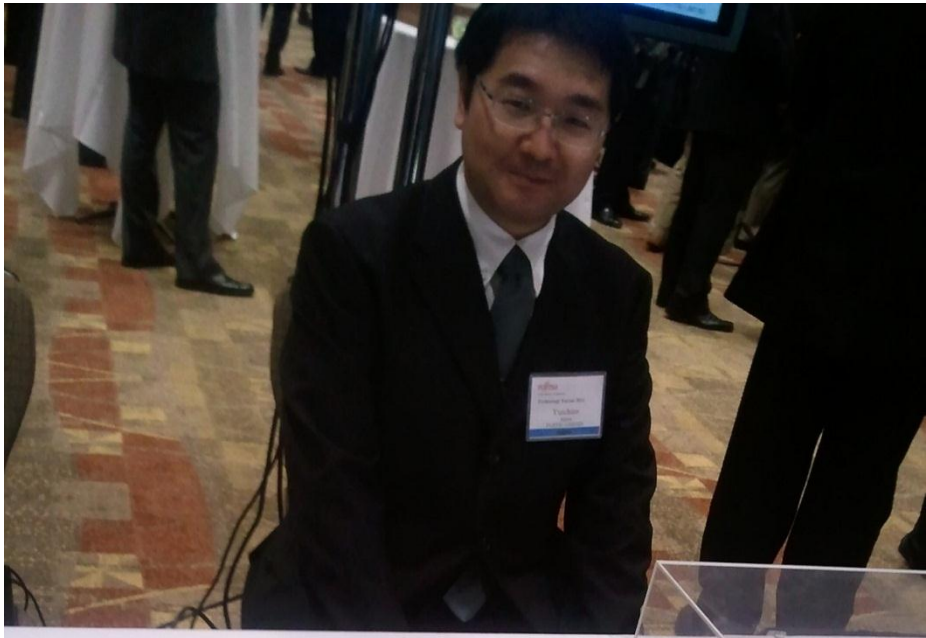
	Cray	IBM	Fujitsu
Highest in	2012 Nov	2012 June	2011 June
Power (Mwatts)	8.2	7.9	12.6
Space Sq. meters	404	280	
Memory (Pbytes)	0.6+0.1	1.6 PB	
Storage	10 PB, 240 GB/s IO[2]		
Speed (Pflops)	17.59	16.32	10.51
Cost	$97 M		
Cabinets|Racks	200	96	864
Blades/cab|Cnode(aka chip)/rack	24	1024	102
Nodes/blade|Core/Cnode	4	16	8
Cores/Node(aka chip)|Cnode/chip	16	1	-
Mp/node|Mp/Cnode	32	16	16
TF/cabinet|TF/rack	100	209	
KW/rack	41	80	
Processor Total	307,200	1,572,864	705,024
Mp (TB)	614.4	1,573	1,410
Nvidis PE total	51,609,600 Mp 100 TB		

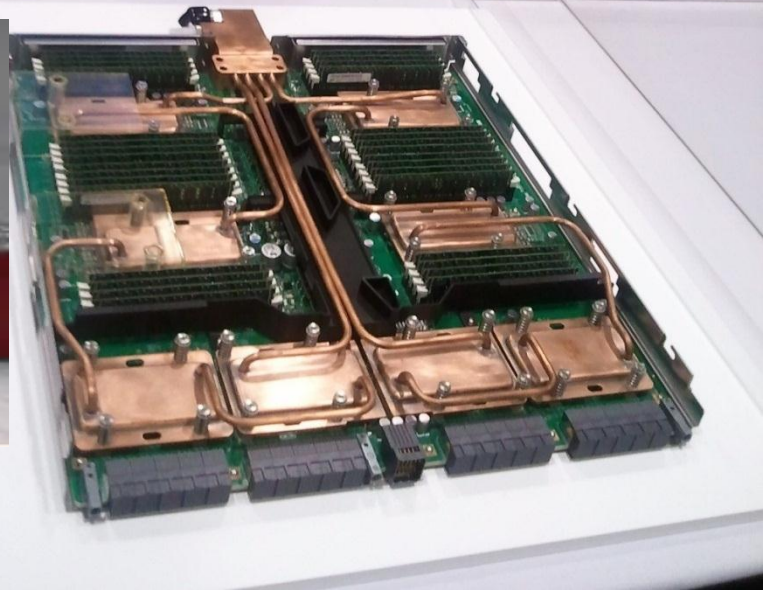
Fujitsu K Riken

Fujitsu K Computer c1/2012


Each rack houses 24 system boards, 12 at the t
and weighs approximately 1000 kg.

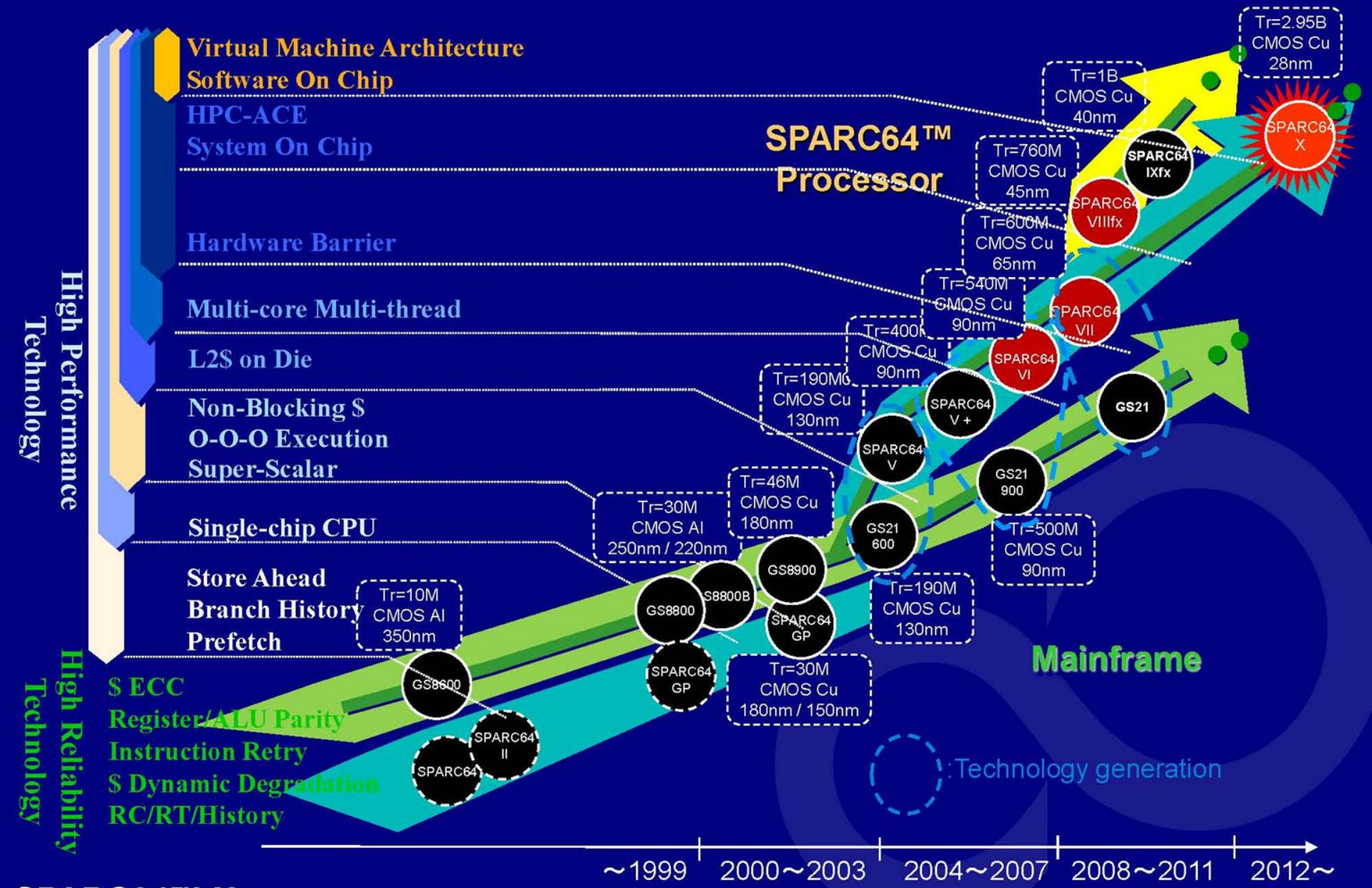
((4 processors x 24 boards)+6i/o x 864 cabs x (88,128 @2GHz) 8 cores per chip = 705,024 proc. Mp(1.4 PB; 2 GB/core)





Fujitsu Supercomputer Team 2012

Fujitsu Processor Development



SPARC64™ X Design Concept

- Combine UNIX and HPC FJ processor features to realize an extremely high throughput UNIX processor.
  - SPARC64 VII/VII+ (UNIX processor) feature
    - High CPU frequency (up-to 3GHz)
    - Multicore/Multithread
    - Scalability : up-to 64sockets
  - SPARC64 VIIIfx (HPC processor) feature
    - HPC-ACE: Innovative ISA extensions to SPARC-V9
    - High Memory B/W: peak 64GB/s, Embedded Memory Controller
- Add new features vital to current and future UNIX servers
  - Virtual Machine Architecture
  - Software On Chip
  - Embedded IOC (PCI-GEN3 controller)
  - Direct CPU-CPU interconnect

Software on Chip 1/2

- HW for SW
  Accelerates specific software function with HW

- The targets
  - Decimal operation (IEEE754 decimal and NUMBER)
  - Cypher operation (AES/DES)
  - Database acceleration

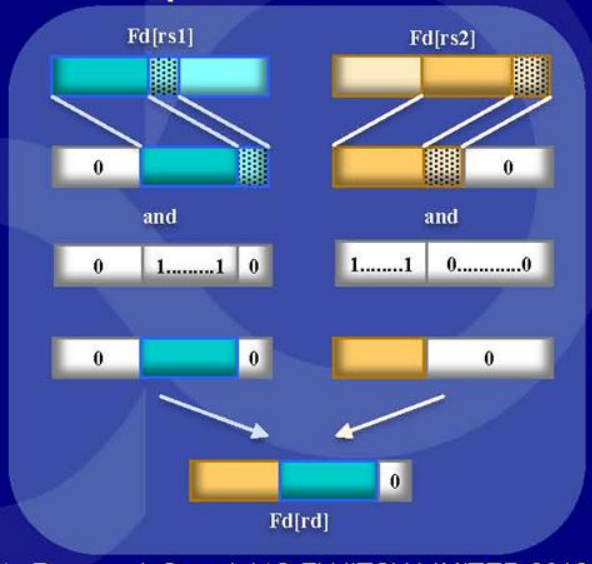
- HW implementation
  - The HW engines for SWoC are implemented in FPU
    - To fully utilize 128 FP registers & software pipelining
  - Implemented as instructions rather than dedicated co-processor to maximize flexibility of SW.
  - Avoid complication due to “CISC” type instructions
    - Various “RISC” type instructions are newly defined, instead.
    - 18 insts. for Decimal, and 10 insts. for Cypher operation

Software on Chip 2/2

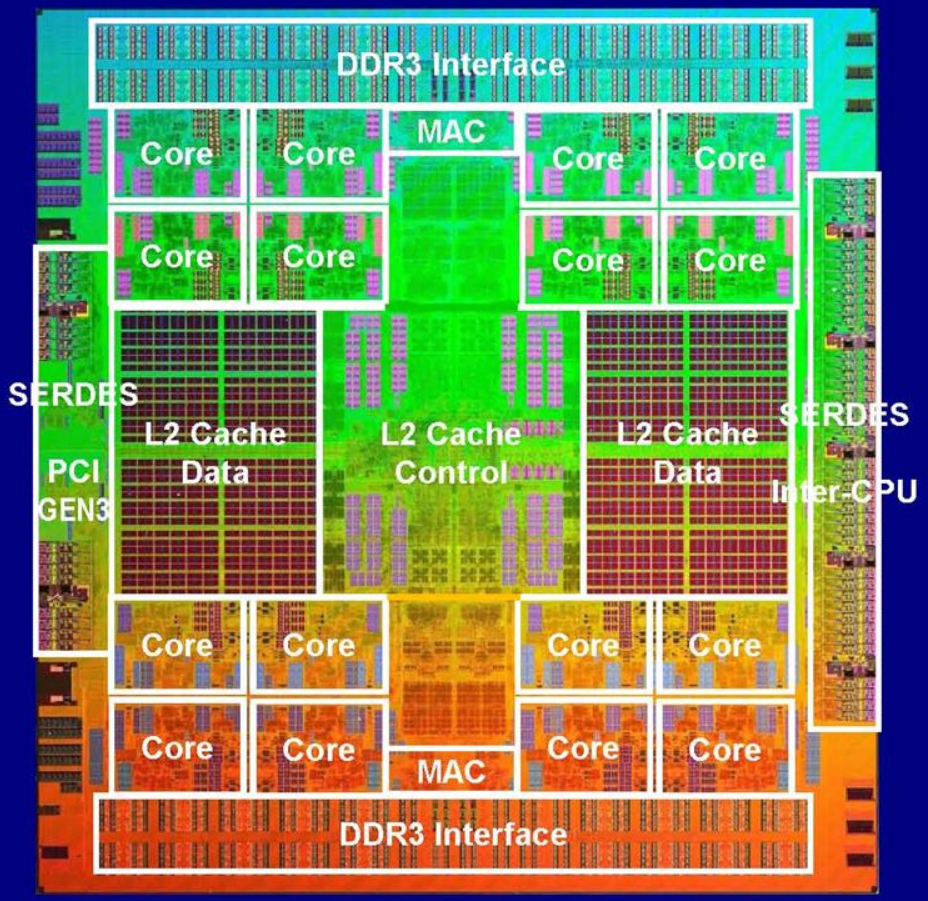
Decimal Instructions

- Supported data type
  - IEEE754 DPD(Densely Packed Decimal)
    8B fixed length
  - NUMBER
    Variable length (max 21Byte)

- Instructions
  - Both DPD/NUMBER instructions are defined as 8B operation (add/sub/mul/div/cmp) on FP registers
    - To maximize performance with reasonable HW cost
    - When the data length is > 8byte, multiple such instructions will be used.
  - An instruction for special byte-shift on FP registers is newly added to support unaligned NUMBER

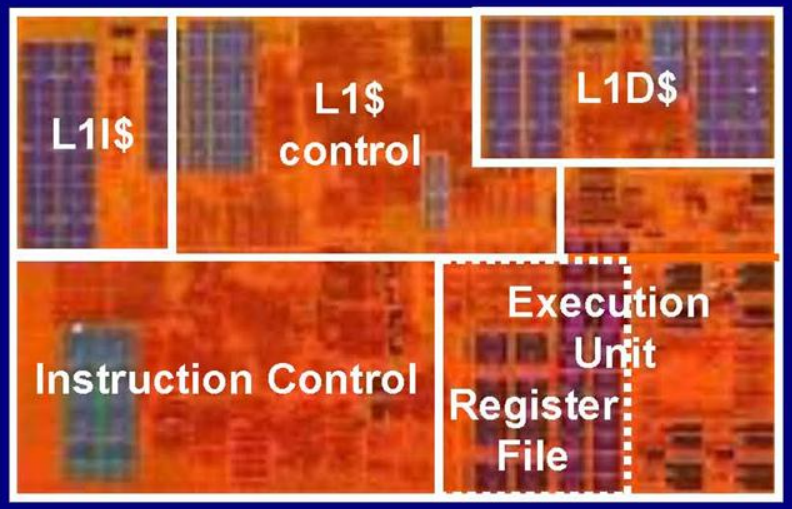


SPARC64™ X Chip Overview



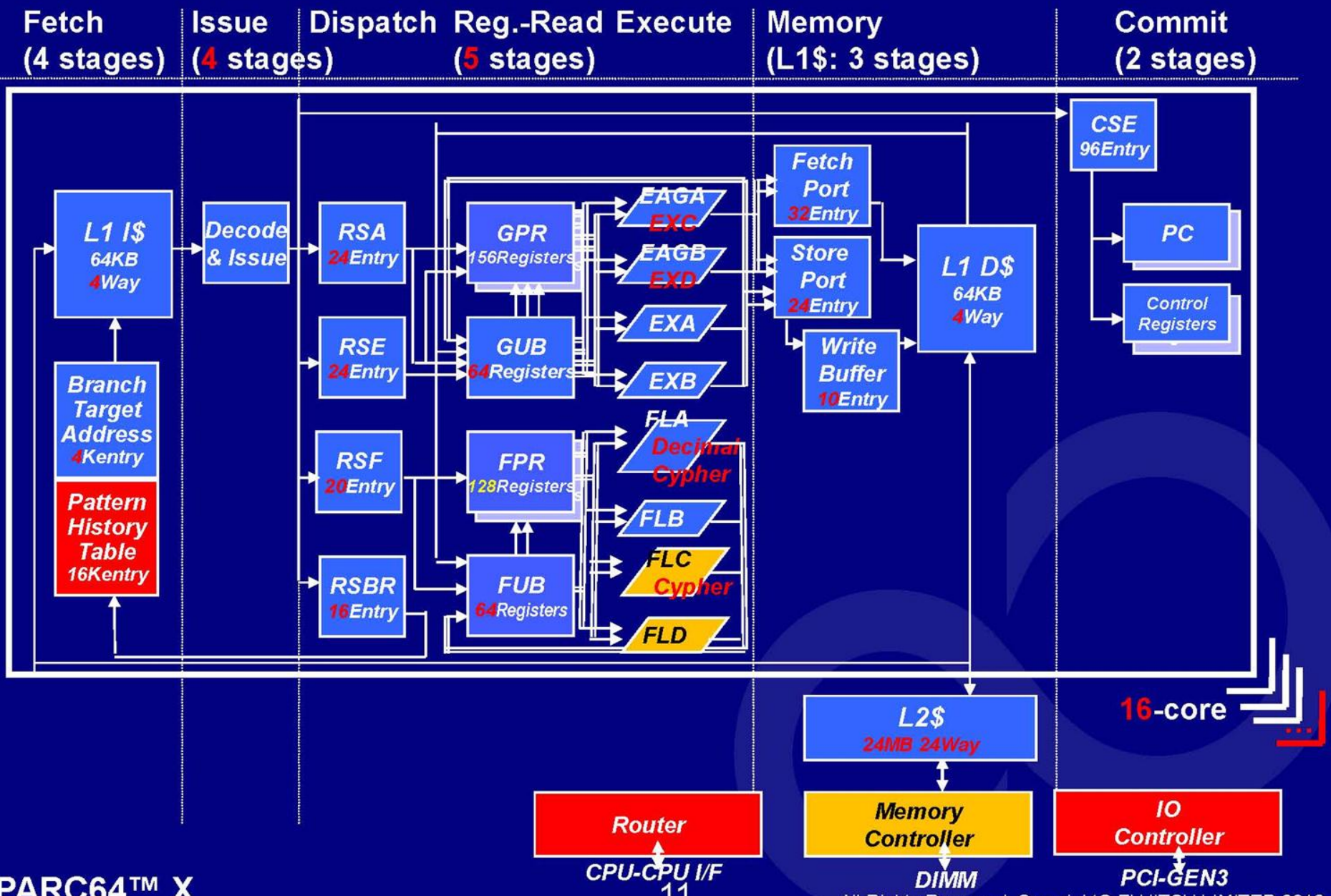
- **Architecture Features**
  - 16 cores x 2 threads
  - SWoC (Software on Chip)
  - Shared 24 MB L2$
  - Embedded Memory and IO Controller
- **28nm CMOS**
  - 23.5mm x 25.0mm
  - 2,950M transistors
  - 1,500 signal pins
  - 3GHz
- **Performance (peak)**
  - 288GIPS/382GFlops
  - 102GB/s memory throughput

SPARC64™ X Core spec



Instruction Set Architecture	SPARC-V9/JPS HPC-ACE VM SWoC
Branch Prediction	4K BRHIS 16K PHT
Integer Execution Units	156 GPR x 2 + 64 GUB
	ALU/SHIFT x2 ALU/AGEN x2 MULT/DIVIDE x1
FP Execution Units	128 FPR x 2 + 64 FUB
	FMA x4, FDIV x2 IMA/Logic x4 Decimal x1 / Cypher x2
L1$	L1I$ 64KB/4way L1D$ 64KB/4way

SPARC64™ X Pipeline



IBM Blue Gene/Q Summary

Processor	IBM PowerPC® A2 1.6 GHz, 16 cores per node
Memory	16 GB SDRAM-DDR3 per node (1333 MTps)
Networks	5D Torus—40 GBps; 2.5 µsec latency Collective network—part of the 5D Torus; collective logic operations supported Global Barrier/Interrupt—part of 5D Torus PCIe x8 Gen2 based I/O 1 GB Control Network— System Boot, Debug, Monitoring
I/O Nodes (10 GbE or InfiniBand)	16-way SMP processor; configurable in 8,16 or 32 I/O nodes per rack
Operating systems	Compute nodes—lightweight proprietary kernel
Performance	Peak performance per rack—209.7 TFlops
Power	Typical 80 kW per rack (estimated) 380-415, 480 VAC 3-phase; maximum 100 kW per rack; 4x60 amp service per rack
Cooling	90 percent water cooling (18°C - 25°C, maximum 30 GPM); 10 percent air cooling
Acoustics	7.9 bels
Dimensions	Height: 2095 mm Width: 1219 mm Depth: 1321 mm Weight: 4500 lbs with coolant (LLNL 1 IO drawer configuration)

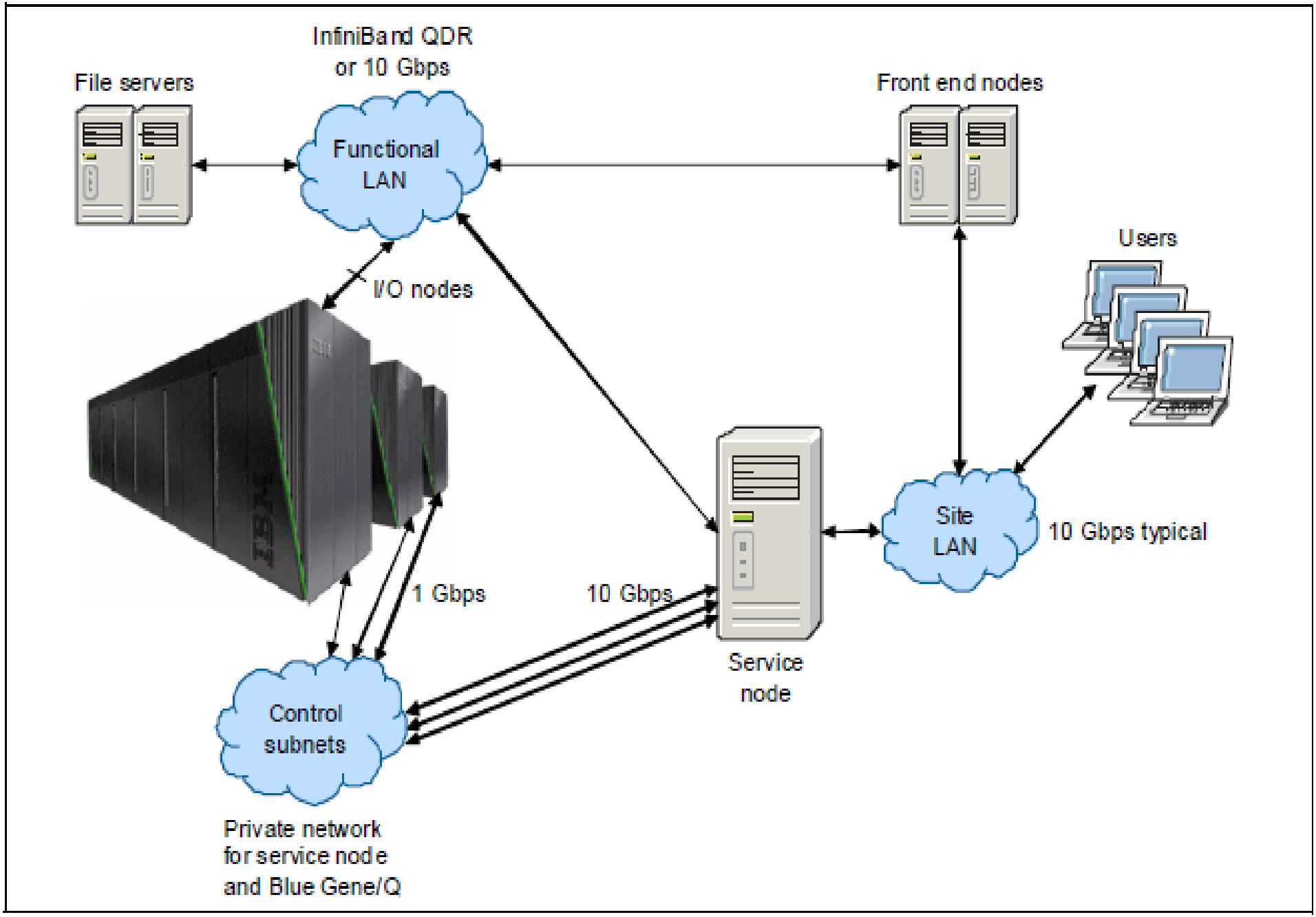


Figure 1-1 IBM Blue Gene/Q system architecture

IBM Blue Gene/Q

Scales to 512 racks or 100 Pflops

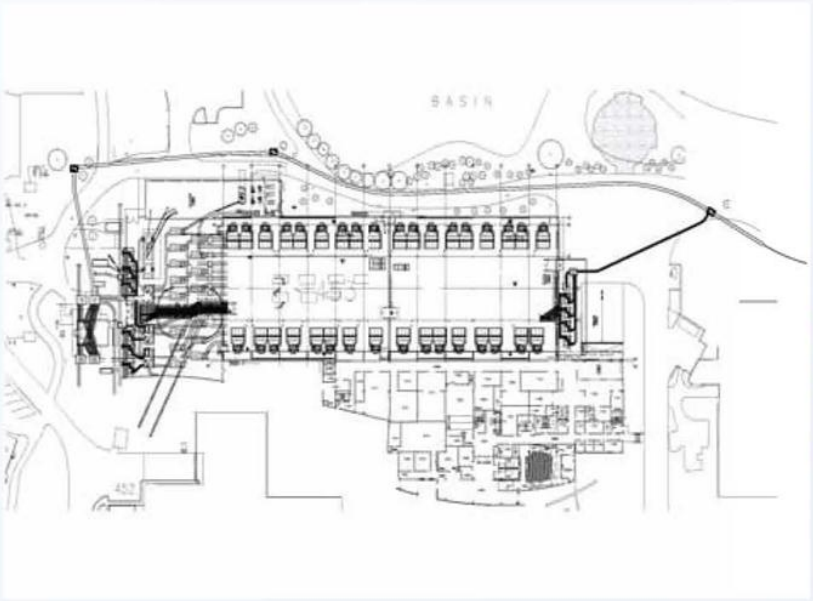
C.Node
16 cores, 16 GB,
10 Gflops/core
1.6 TF/node

1024 C.nodes/rack
209 TF/rack
16 TB/rack?
80 KW/rack

96 rack system
98,304 C.nodes or
1.57 M proc. cores
16 PF. 1.6 PB
7.9 Mwatts
280 m^2

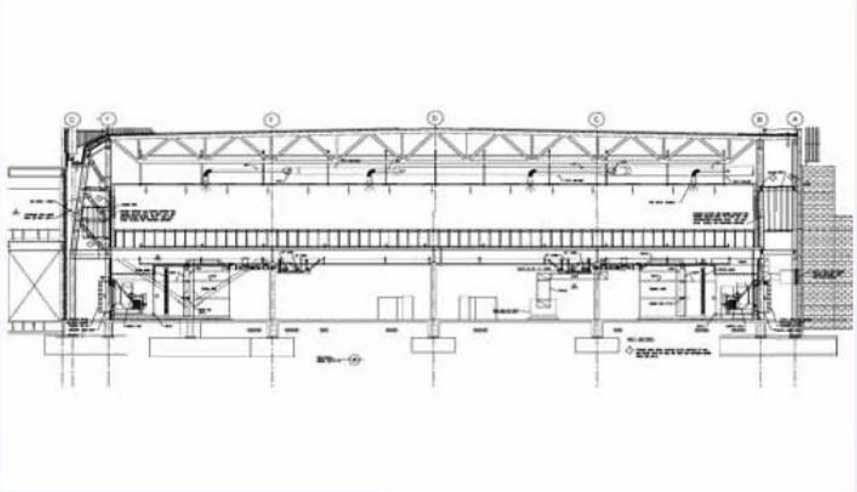
TSF computer room power is being scaled from 15MW to 30MW

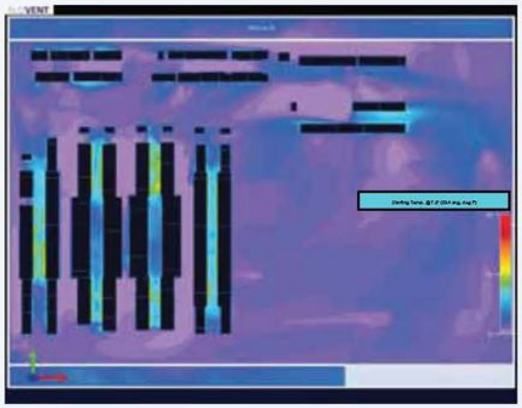
- Capitalize on the computational efficiencies (TF/MW and SF/TF)
- Capitalize on the electrical/mechanical system efficiencies
- Adding an additional 15MW into the TSF

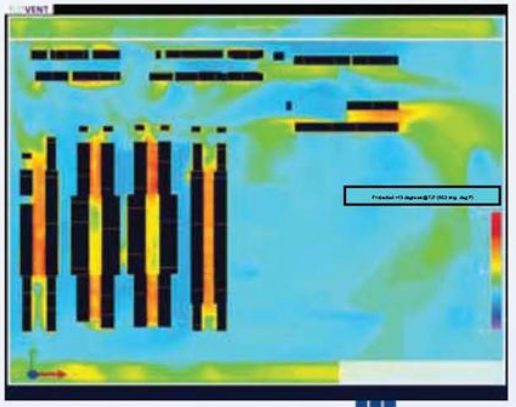


A comprehensive computational fluid dynamic (CFD) model was performed to analyze airflow patterns in the TSF

- Physical layouts imported
- Baseline CFD
  - Starting temperature 53.4°F
  - Study temperature 66.3°F
- Modeled airflow
  - 2" above finished floor (AFF) – inlet of racks
  - 7.5' AFF – above racks
  - 10.5' AFF – ceiling







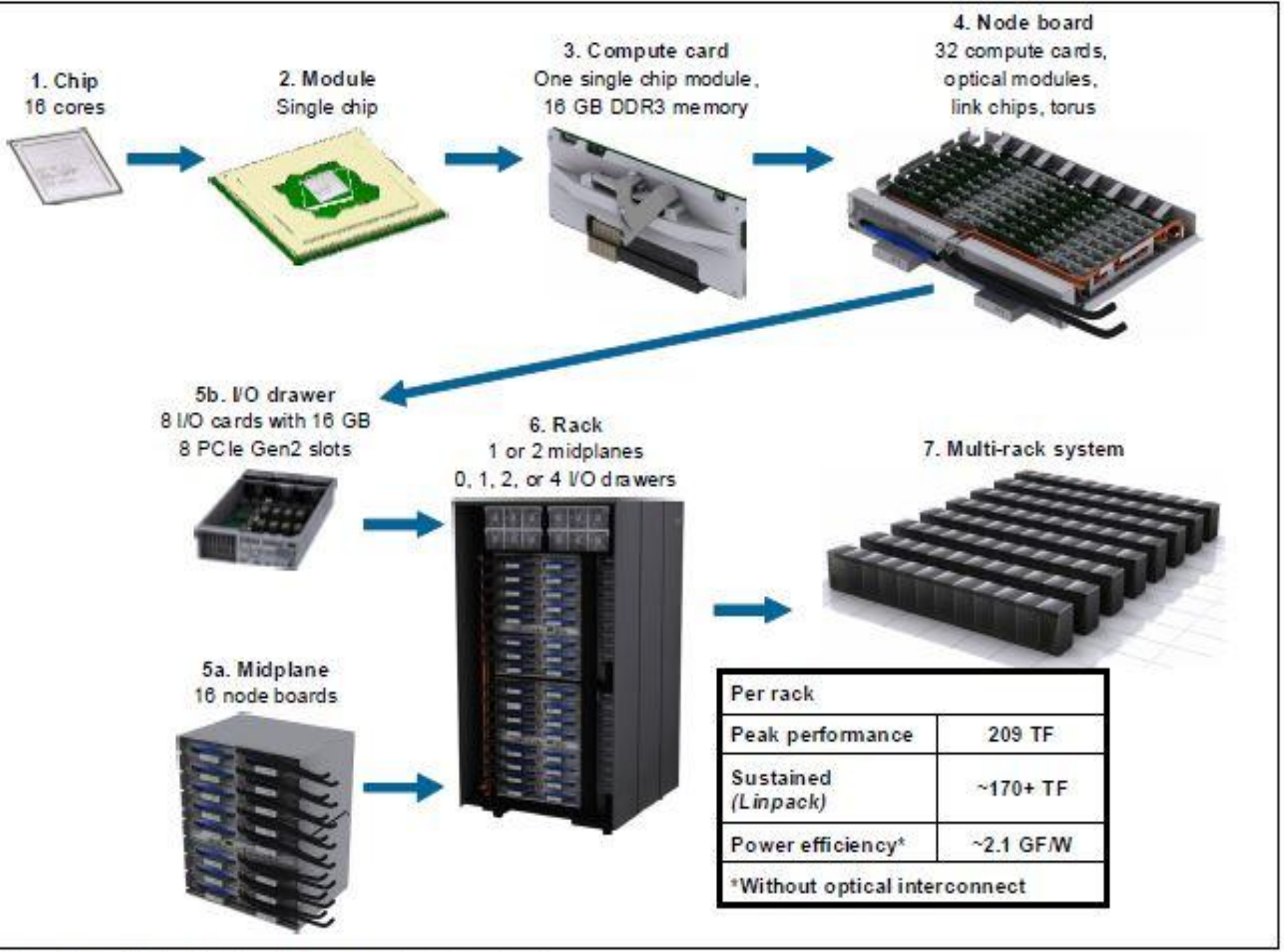
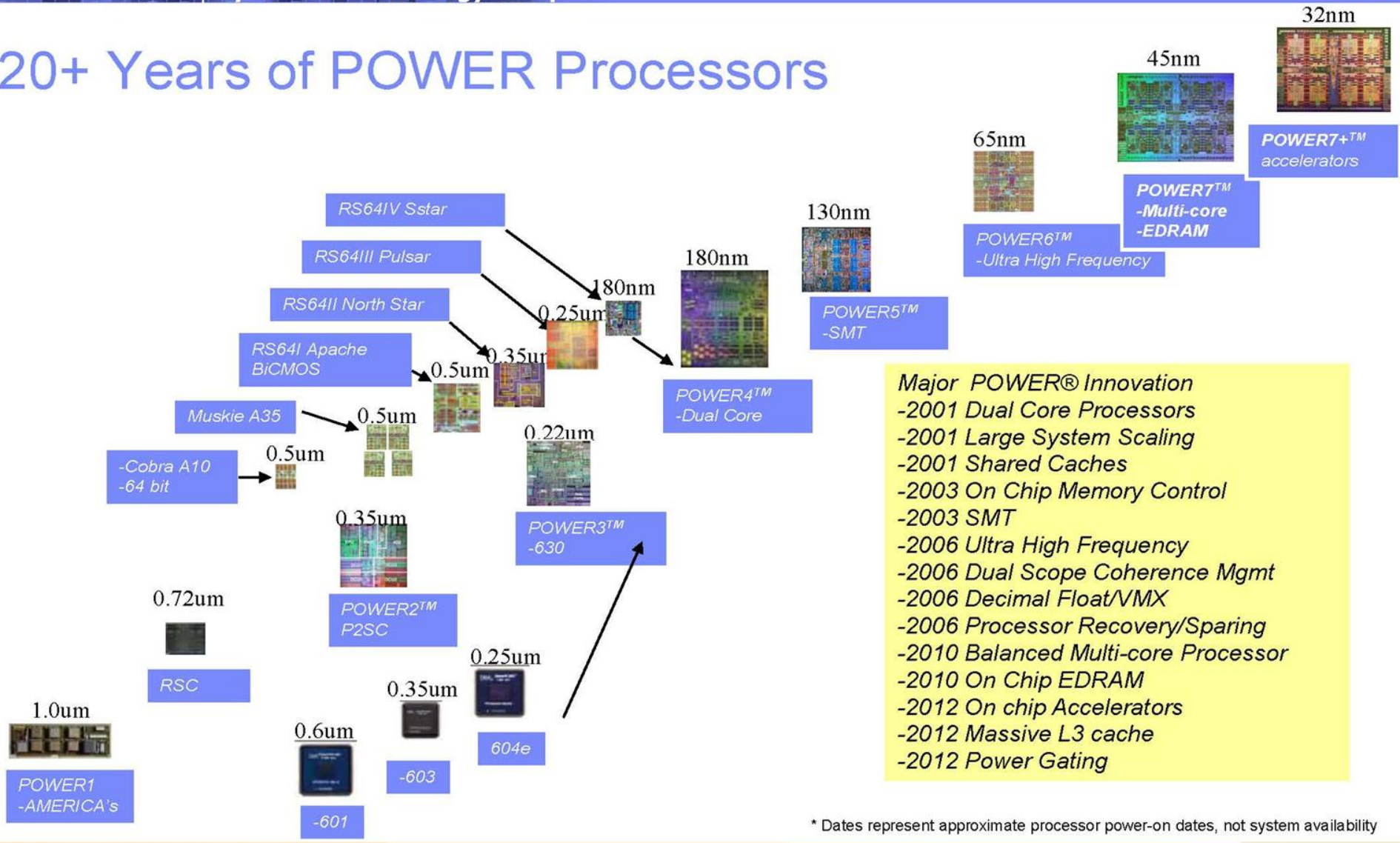


Figure 1-2 Blue Gene/Q hardware overview

20+ Years of POWER Processors

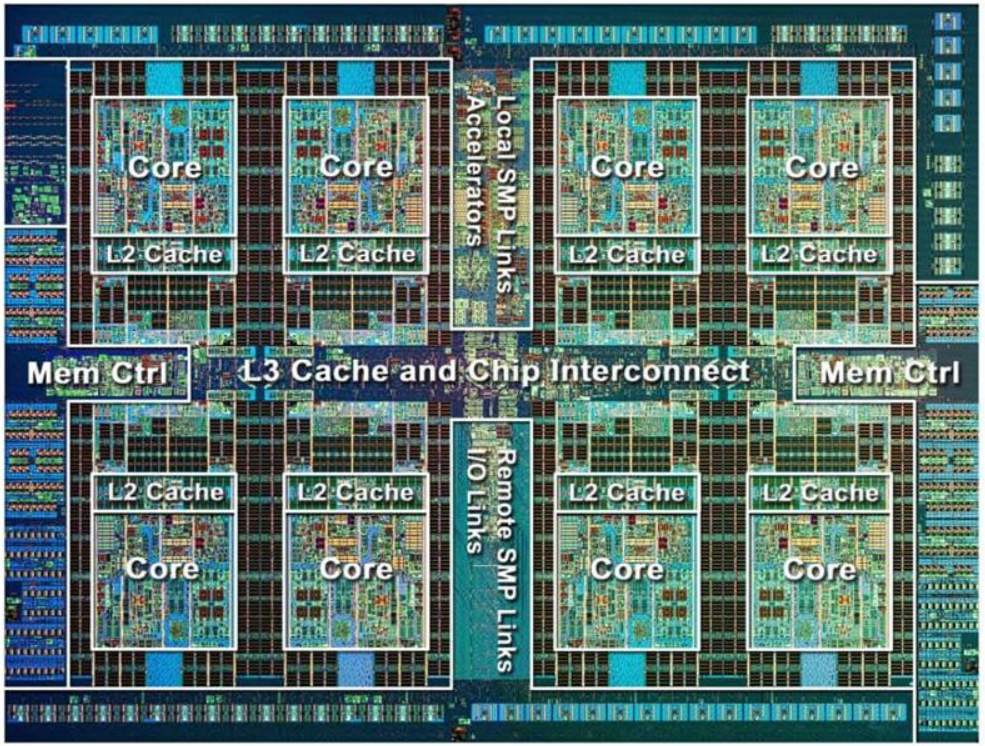


Major POWER® Innovation
-2001 Dual Core Processors
-2001 Large System Scaling
-2001 Shared Caches
-2003 On Chip Memory Control
-2003 SMT
-2006 Ultra High Frequency
-2006 Dual Scope Coherence Mgmt
-2006 Decimal Float/VMX
-2006 Processor Recovery/Sparing
-2010 Balanced Multi-core Processor
-2010 On Chip EDRAM
-2012 On chip Accelerators
-2012 Massive L3 cache
-2012 Power Gating

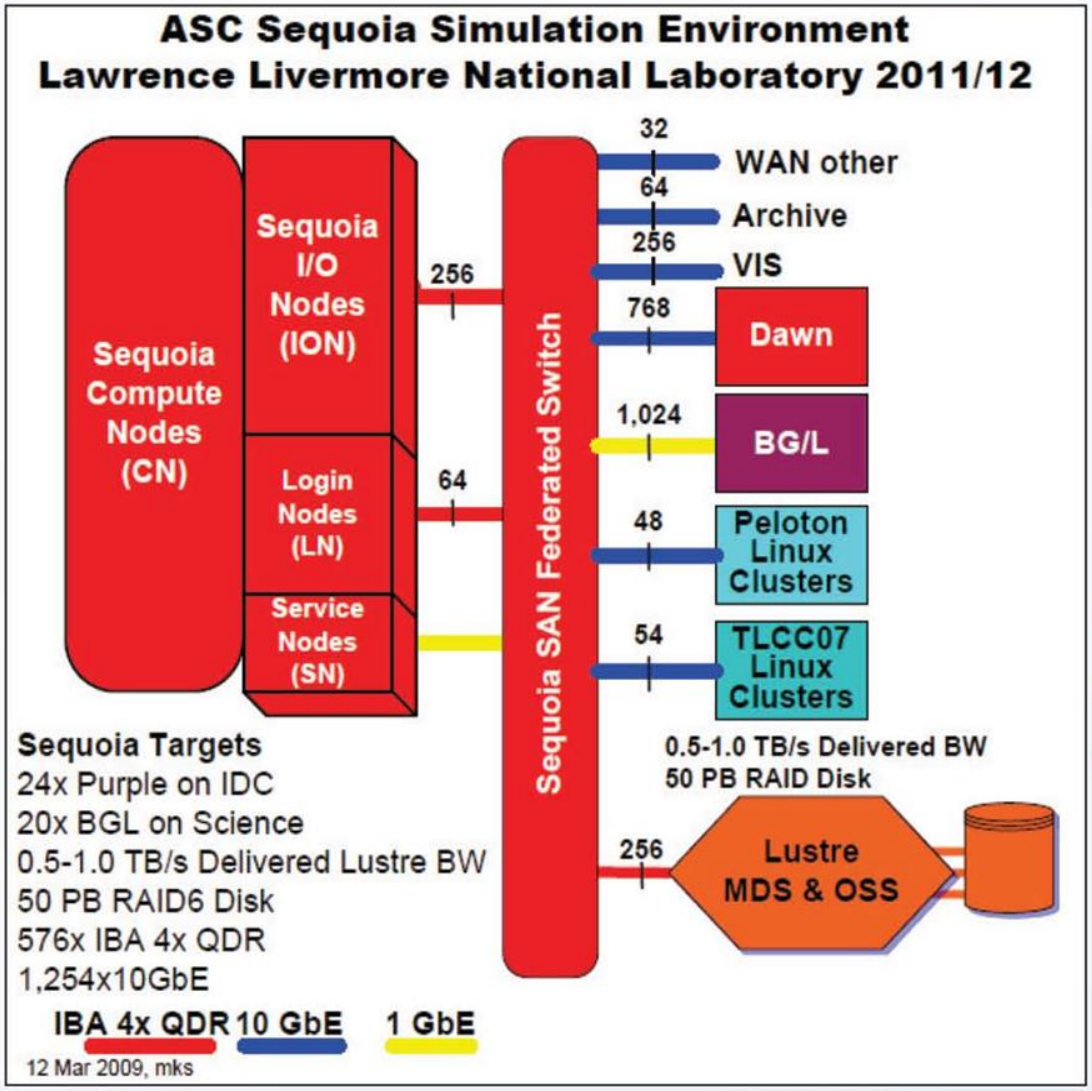
* Dates represent approximate processor power-on dates, not system availability

POWER7+ Processor Chip

- Area: 567mm2
- Eight processor cores
  - 12 execution units per core
  - 4 Way SMT per core
  - 32 Threads per chip
  - 256KB L2 per core
- Scalability up to 32 Sockets
  - 360GB/s SMP bandwidth/chip
  - 20,000 coherent operations in flight
- Technology: 32nm lithography, Cu, SOI, eDRAM, 13 metal levels
- 2.1B transistors
  - Equivalent function of 5.4B
- 80MB on chip eDRAM shared L3
- Accelerators
- Enhanced Power management
- Binary Compatibility with POWER6/7



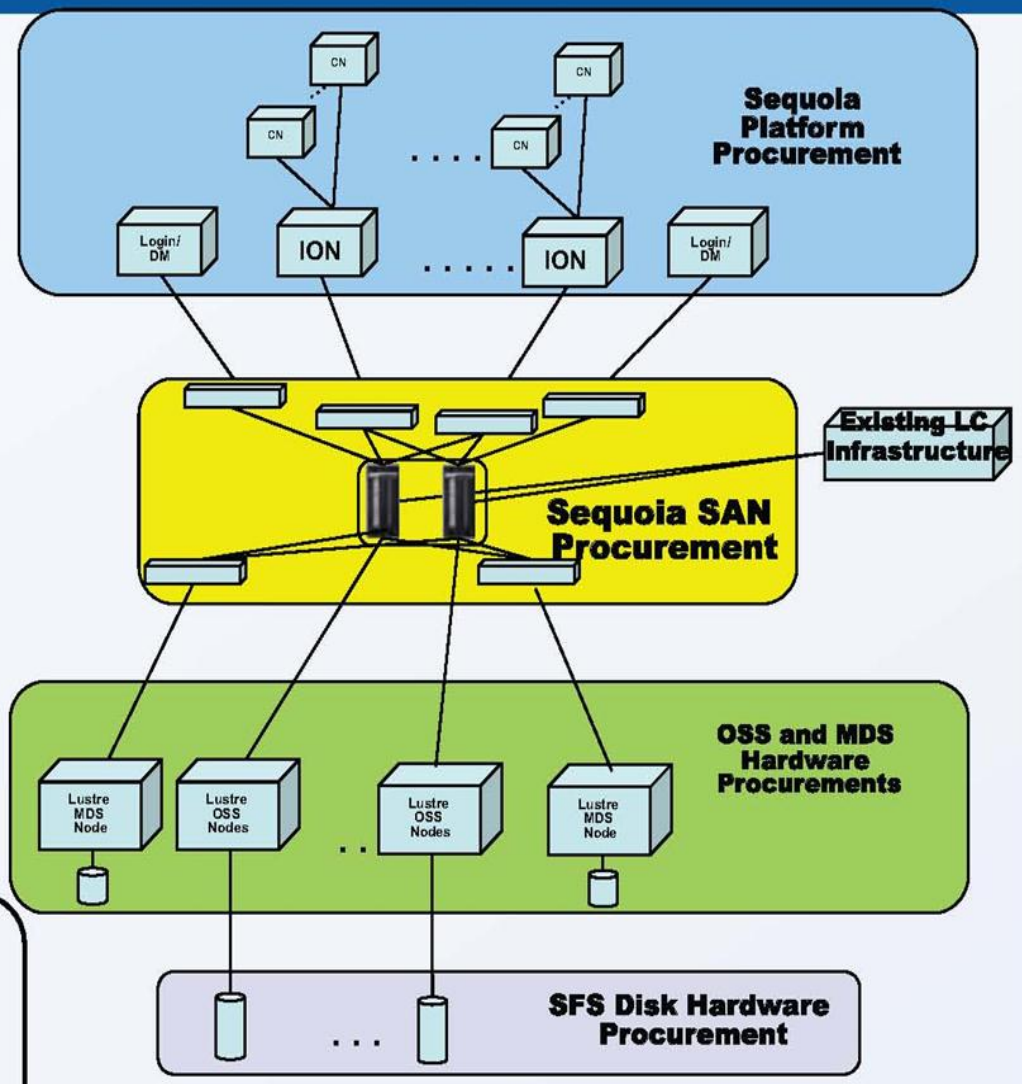
Sequoia Hierarchal Hardware Architecture in Integrated Simulation Environment



- Sequoia Statistics
  - 20 PF/s target
  - Memory 1.6 PB, 4 PB/s BW
  - 1.5M Cores
  - 3 PB/s Link BW
  - 60 TB/s bi-section BW
  - 0.5-1.0 TB/s Lustre BW
  - 50 PB Disk
- ~8.0MW Power, 3,500 ft^2
- Third generation IBM BlueGene
- Challenges
  - Hardware Scalability
  - Software Scalability
  - Applications Scalability

Sequoia I/O Infrastructure

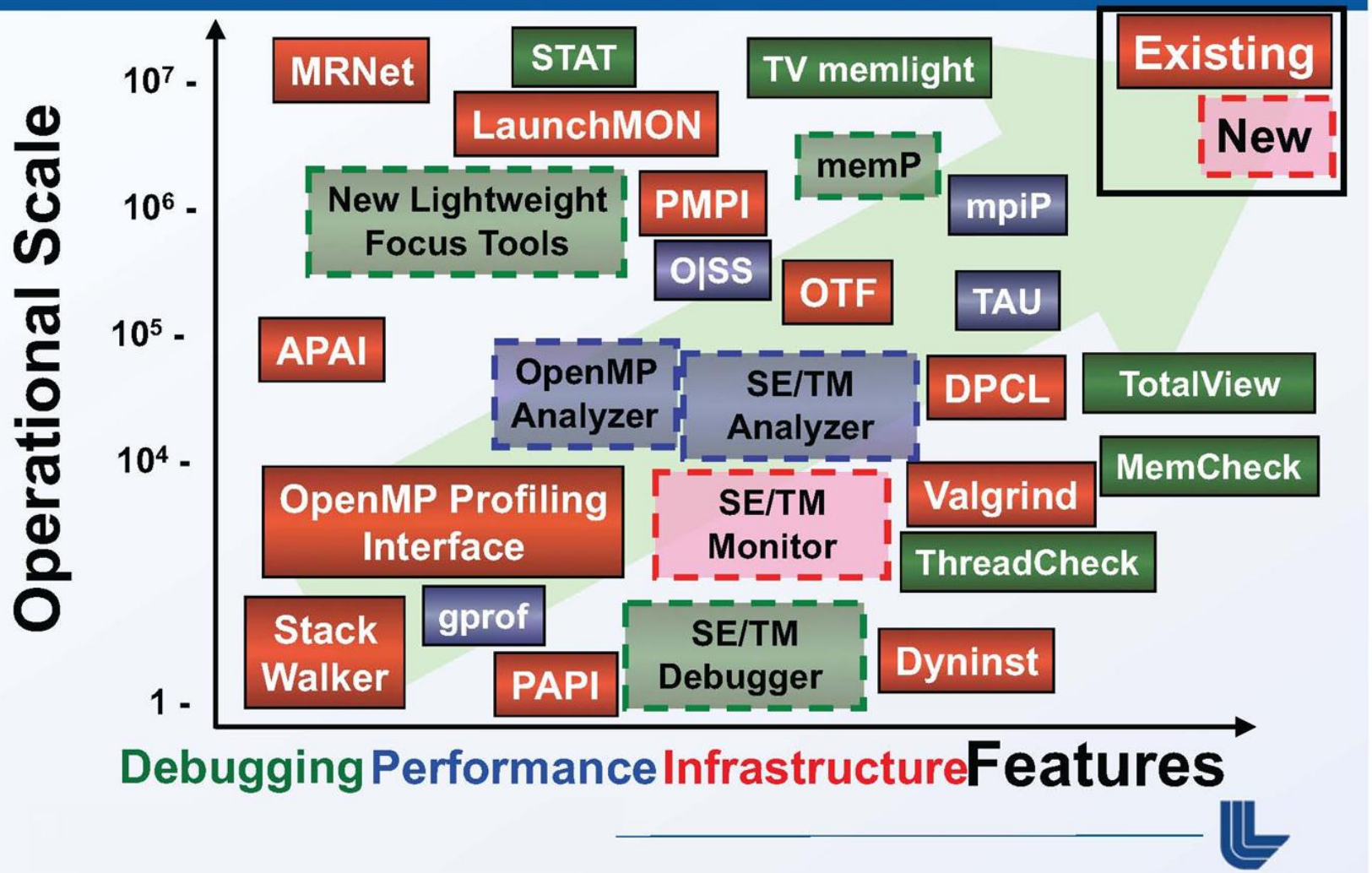
- **Requirements**
  - 50PB file system
  - 500GB/s minimum, 1TB/s stretch goal
  - QDR InfiniBand SAN connection to Sequoia
  - Must integrate with existing Ethernet infrastructure



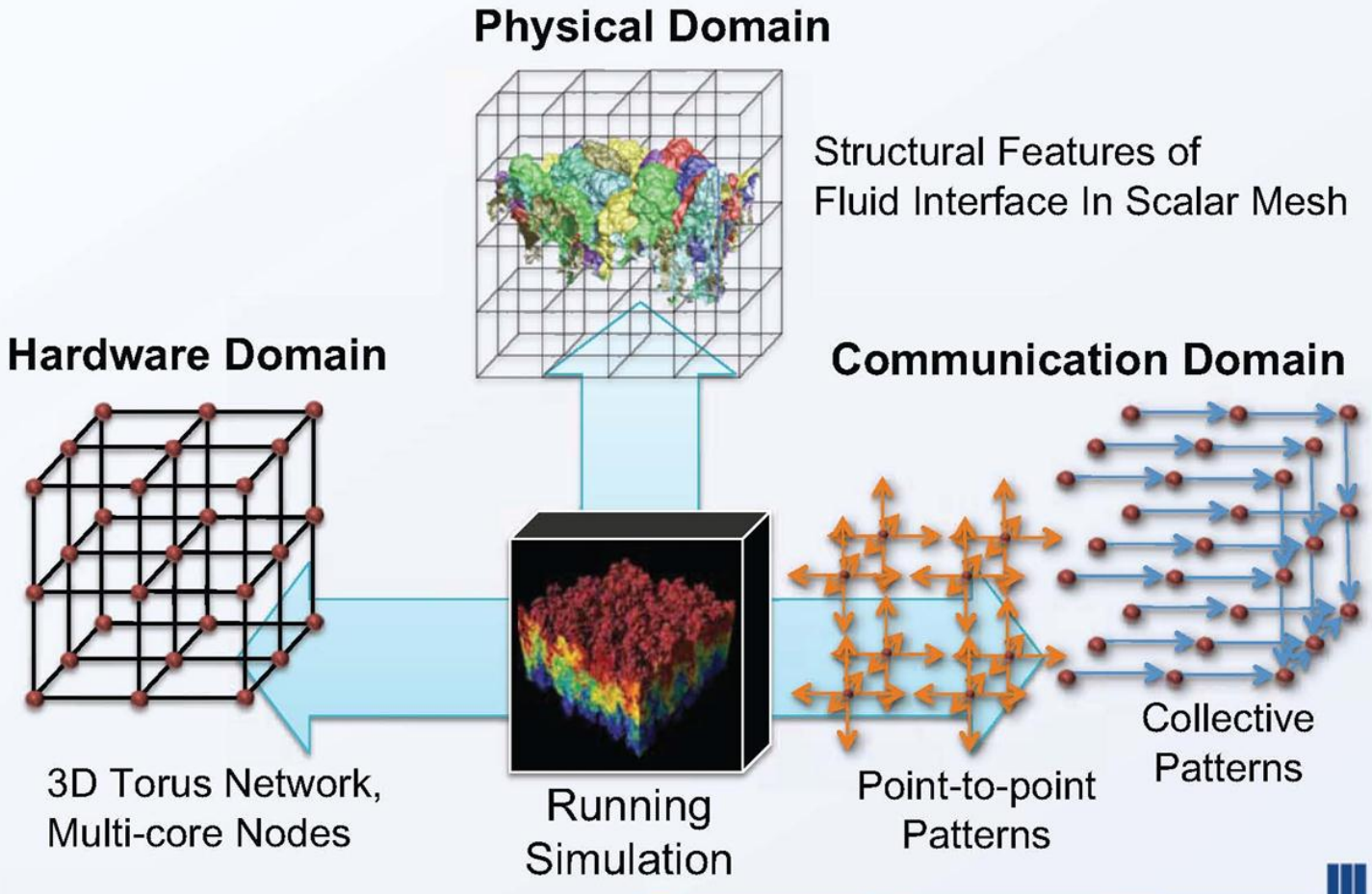
Phased Bandwidth Delivery

- **Phase 1: 10% Sep 2011**
- **Phase 2: 50% April 2012**
- **Phase 3: 100% July 2012**

The tools that users know and expect will be available on Sequoia with improvements and additions as needed



An application's performance should be understood in light of its three interdependent domains



Cray Titan at ORNL

Active	**Became operational October 29, 2012**
Sponsors	**US DOE and NOAA (<10%)**
Operators	**Cray Inc.**
Location	**Oak Ridge National Laboratory**
Architecture	**18,688 AMD Opteron 6274 16-core CPUs** **18,688 Nvidia Tesla K20 GPUs** **Cray Linux Environment**
Power	**8.2 MW**
Space	**404 sqm (4352 sq ft)**
Memory	**710 TB (598 TB CPU and 112 TB GPU)[1]**
Storage	**10 PB, 240 GB/s IO[2]**
Speed	**17.59 petaFLOPS[3]**
Cost	**US$97 million**
Ranking	**TOP500: 1, November 12, 2012[3]**
Purpose	**Scientific research**
Legacy	**First GPU based supercomputer to perform over 10 petaFLOPS**
Web site	**http://www.olcf.ornl.gov/titan/**

ORNL Titan Cray Supercomputer

- 18,688 AMD 16-core Opteron 6274 CPUs = 299K nodes@2.2 GHz
  - 200 cabinets x 4 nodes/blade x 24 blades/cabinet x 16 cores/node = 299K core
  - Pc: 18.7K proc x 16 core/proc. x 2.2 GHz. = 658 Tticks; ?? Flops/tick
  - Mp: 200 x 96 nodes/cab x 32 GB/node = 600 TB
- 18,688 Nvidia 2.5K-core K20 GPUs. 732 MHz = 46.5 M cores
  - 1.3 TFlops per chip??
  - Mp: 112 TB; 6 GB/proc? …1/40 of a byte per FLOP on GPU
- Ms: 13.6 PB driven by 140-Dell servers
- 9 Mwatts; PUE=?? 404 m^2
- 1-5 weather years per day of simulation

Processor	**16-core 64-bit AMD Opteron 6200 Series processors, up to 96/cab; NVIDIA® Tesla® K20 GPU Accelerators, up to 96/cab**
Memory	**16 GB or 32 GB registered ECC DDR3 SDRAM and 6 GB GDDR5 per compute node** **Memory bandwidth: 4 channels of DDR3 memory per compute node**
Compute Cabinet	**AMD processing cores: 1,536 processor cores per system cabinet** **Peak performance: 100+ Tflops per system cabinet**
Interconnect	**1 Gemini routing and communications ASIC per two compute nodes**
	48 switch ports per Gemini chip (160 GB/s internal switching capacity per chip) 3D torus INterconnect
System Administration	**Cray System Management workstation** **Graphical and command line system administration** **Single-system view for system administration**
Reliability Features (Hardware)	**Cray Hardware Supervisory System (HSS) with independent 100 Mb/s management fabric between all system blades and cabinet-level controllers** **Full ECC protection of all packet traffic in the Gemini network**
	Redundant power supplies; redundant voltage regulator modules; Redundant paths to all system RAID
Reliability Features (Software)	**HSS system monitors operation of all operating system kernels** **Lustre file system object storage target failover; Lustre metadata server failover** **Software failover for critical system services including system database, system logger, and batch subsystems** **NodeKARE (Node Knowledge and Reconfiguration)**
Operating System	**Cray Linux Environment (components include SUSE Linux SLES11, HSS and SMW software)** **Extreme Scalability Mode (ESM) and Cluster Compatibility Mode (CCM)**
Compilers, Libraries & Tools	**PGI compilers, Cray Compiler Environment, PathScale, CUDA, CAPS, support for Fortran 77, 90, 95; C/C++, UPC, Co-Array Fortran, MPI 2.0, Cray SHMEM, OpenACC directives-based programming, other standard MPI libraries using CCM**
Job Management	**PBS Professional, Moab Adaptive Computing Suite, Platform LSF**
External I/O Intface	**InfiniBand, 10 Gigabit Ethernet, Fibre Channel (FC) and Ethernet**
Disk Storage	**Full line of FC-attached disk arrays with support for FC and SATA disk drives**
Parallel File System	**Lustre, Data Virtualization Service allows support for NFS, external Lustre and other file systems**
Power	**45-54.1 kW (45.9 - 55.2 kVA) per cabinet, depending on configuration** **Circuit requirements: three-phase wye, 100 AMP at 480/277 and 125 AMP at 400/230 (three-phase, neutral and ground)**
Cooling	**Air-cooled, air flow: 3,000 cfm (1.41 m3/s); intake: bottom; exhaust: top** **Optional ECOphlex liquid cooling**
Dimensions (Cab)	**H 93 in. (2,362 mm) x W 22.50 in. (572 mm) x D 56.75 in. (1,441 mm)**
Weight (Maximum)	1,600 lbs. per cabinet (725 kg) air cooled; 2,000 lbs. per cabinet (907 kg) liquid cooled

Titan has 200 cabinets, 18,688 nodes (4 nodes per blade, 24 blades per cabinet=96 nodes/cab),[24] each node containing a 16-core AMD Opteron 6274 CPU with 32 GB of DDR3 ECC memory and

an Nvidia Tesla K20X GPU with 6 GB GDDR5 ECC memory.[25] The total number of processor cores is 299, 008 and the total amount of RAM is over 710 TB.[21]

10 PB of storage (made up of 13, 400 7200 rpm 1 TB hard drives)[26] is available with a transfer speed of 240 GB/s.[21][18] The next storage upgrade is due in 2013, it will up the total storage to between 20 and 30 PB with a transfer speed of approximately 1 TB/s.[21][27]

Titan runs the Cray Linux Environment, a full version of Linux on the login nodes but a scaled down, more efficient version on the compute nodes.[28] GPUs were selected for their vastly higher parallel processing efficiency over CPUs.[25] Although the GPUs have a slower clock speed than the CPUs, each GPU contains 2, 688 CUDA cores at 732 MHz,[29] resulting in a faster overall system.[30][18] Consequently, the CPUs cores are used to allocate tasks to the GPUs rather than for directly processing the data as in previous supercomputers for well optimized codes

- Titan has 200 cabinets, 18,688 nodes (4 nodes per blade, 24 blades per cabinet=96 nodes/cab), and AMD 16-core Opteron 6274 CPU with 32 GB of DDR3 ECC memory for 299K processors and
- An Nvidia Tesla K20X GPU with 6 GB GDDR5 ECC memory and 2, 688 CUDA cores at 732 MHz.[25]
- The total amount of RAM is over 710 TB.[21]
- CPUs cores are used to allocate tasks to the GPUs rather than for any processing
- 10 PB of storage (made up of 13,400 7200 rpm 1 TB hard drives)[26] is with a transfer speed of 240 GB/s.[21][18]
- The next storage upgrade provides20 and 30 PB at1 TB/s
- Titan runs the Cray Linux Environment,

NVIDIA Kepler 1 TF DPFP

	KEPLER GK110
Compute Capability	3.5
Threads / Warp	32
Max Warps / Multiprocessor	64
Max Threads / Multiprocessor	2048
Max Thread Blocks / Multiprocessor	16
32-bit Registers / Multiprocessor	65536
Max Registers / Thread	255
Max Threads / Thread Block	1024
Shared Memory Size Configurations (bytes)	16K 32K 48K
Max X Grid Dimension	2^32-1
Hyper-Q	Yes
Dynamic Parallelism	Yes

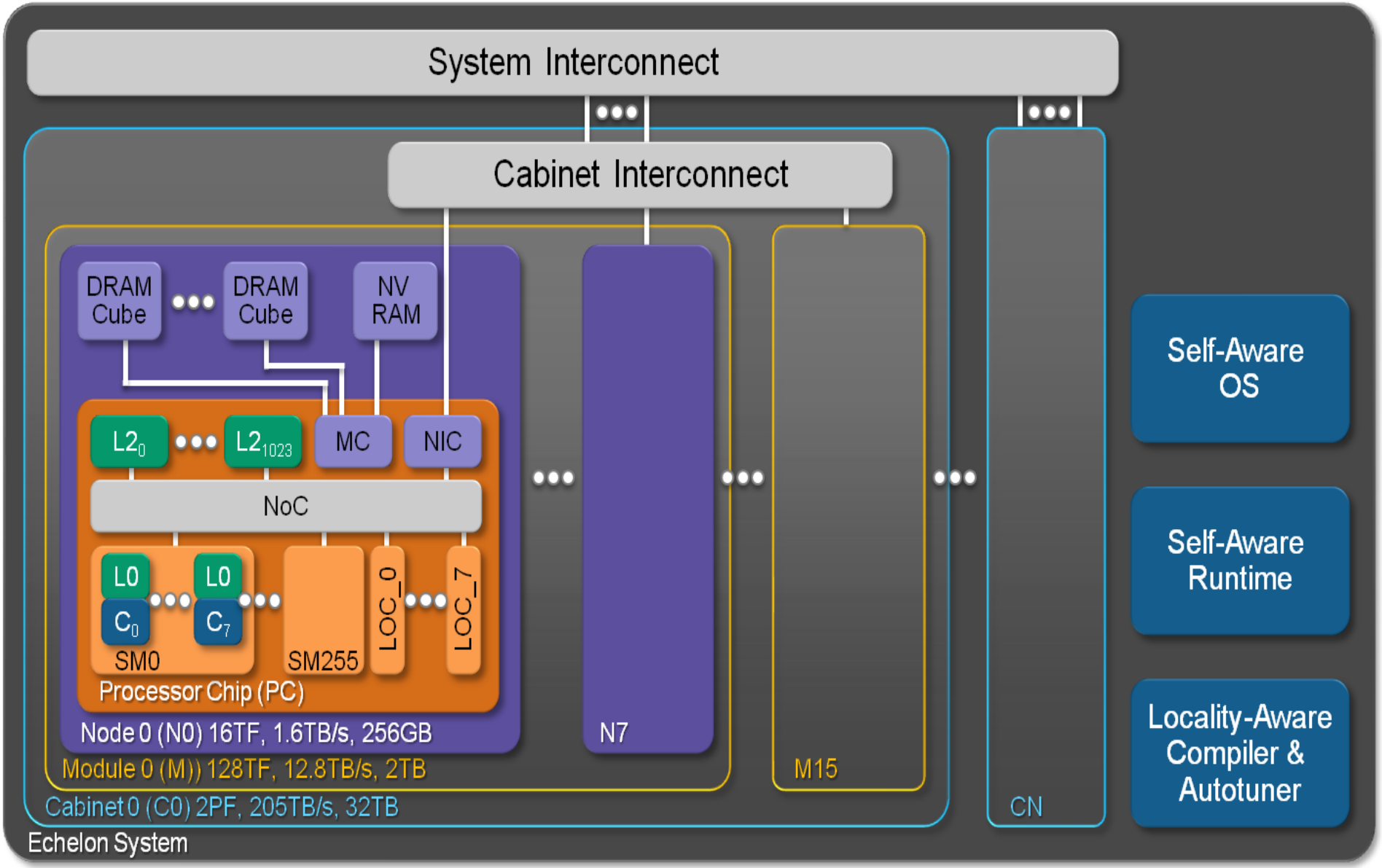
Compute Capability of Fermi and Kepler GPUs

Nvidia TESLA

TECHNICAL SPECIFICATIONS	TESLA K10[a]	TESLA K20	TESLA K20X
Peak double precision floating point performance (board)	0.19 teraflops	1.17 teraflops	1.31 teraflops
Peak single precision floating point performance (board)	4.58 teraflops	3.52 teraflops	3.95 teraflops
Number of GPUs	2 x GK104s	1 x GK110	
Number of CUDA cores	2 x 1536	2496	2688
Memory size per board (GDDR5)	8 GB	5 GB	6 GB
Memory bandwidth for board (ECC off)[b]	320 GBytes/sec	208 GBytes/sec	250 GBytes/sec
GPU computing applications	Seismic, image, signal processing, video analytics	CFD, CAE, financial computing, computational chemistry and physics, data analytics, satellite imaging, weather modeling	
Architecture features	SMX	SMX, Dynamic Parallelism, Hyper-Q	
System	Servers only	Servers and Workstations	Servers only

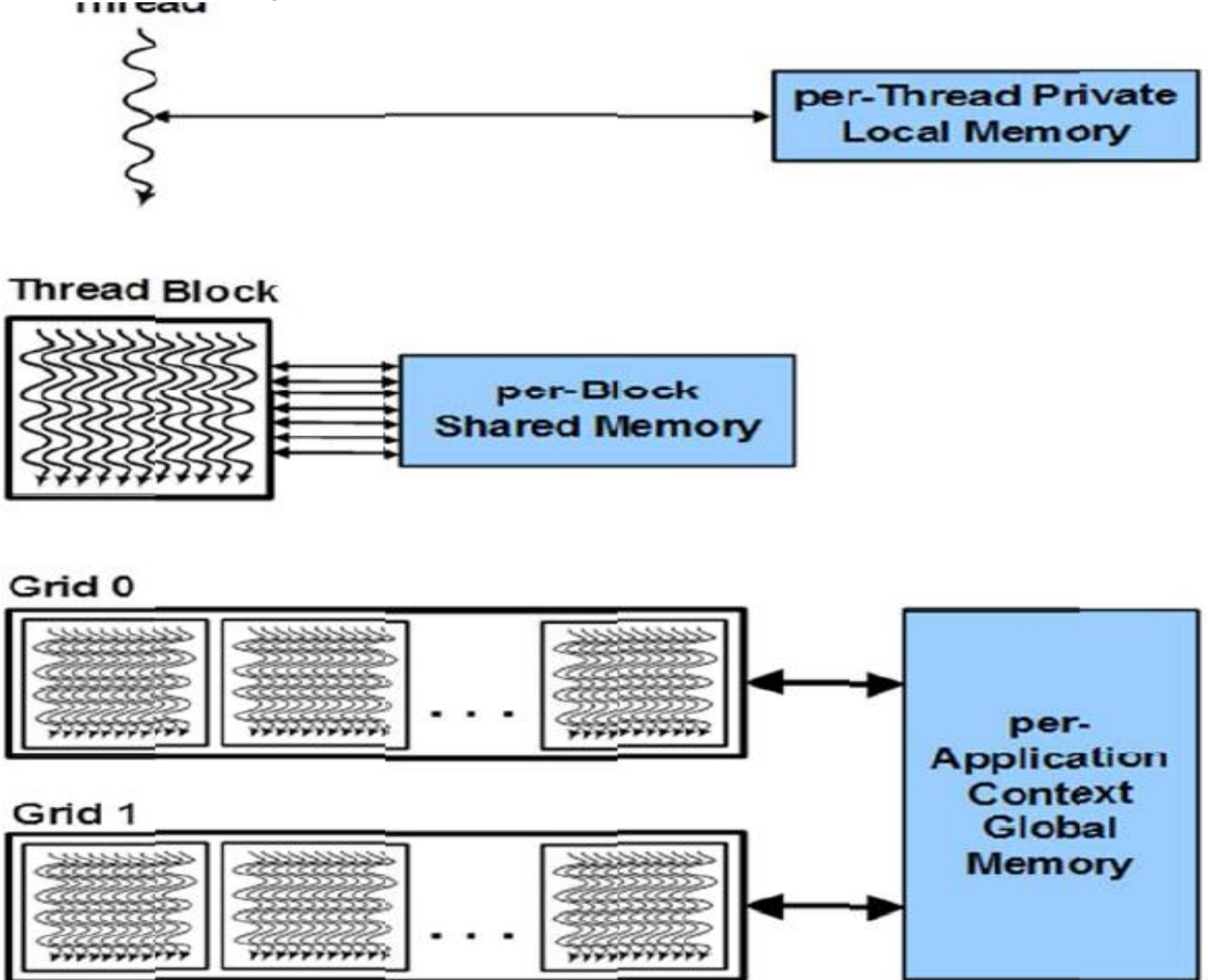
Tesla K10 specifications are shown as aggregate of two GPUs.
With ECC on, 12.5% of the GPU memory is used for ECC bits. So, for example, 6 GB total memory yields 5.25 GB of user available memory with ECC on.

System Sketch

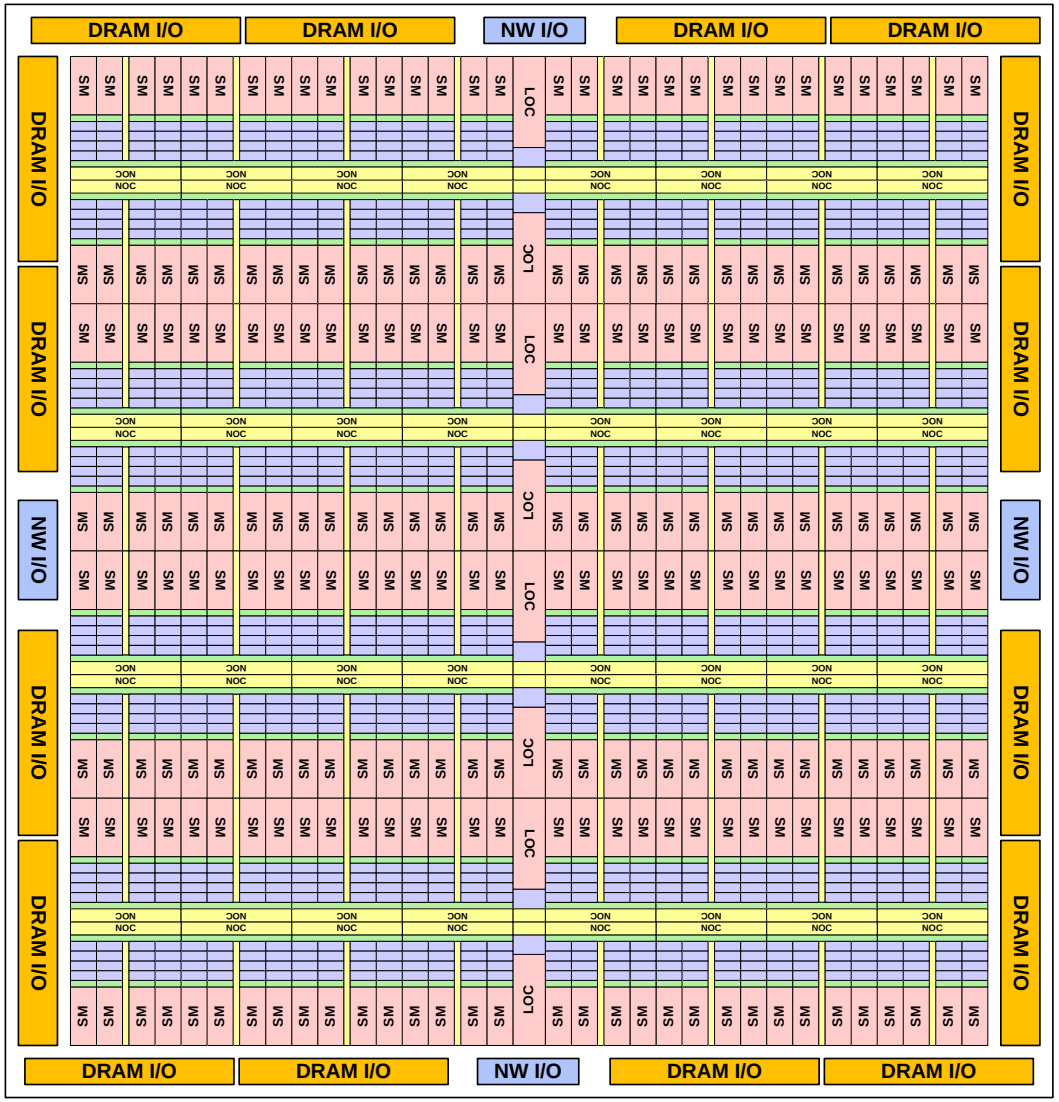


CUDA Programming model

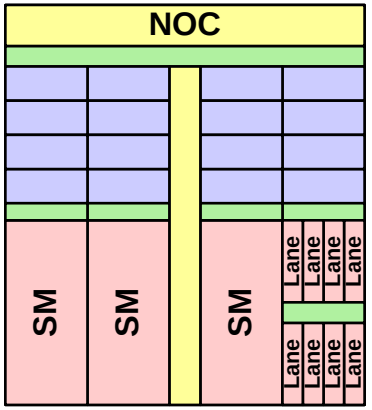
Compute Unified Device Architecture



Echelon Chip Floorplan



17mm



10nm process
290mm²


PCI Express 3.0 Host Interface
GigaThread Engine
Memory Controller
Memory Controller
Memory Controller
Memory Controller
Memory Controller
Memory Controller
SMX
SMX
SMX
SMX
SMX
SMX
SMX
SMX
L2 Cache
SMX
SMX
SMX
SMX
SMX
SMX
SMX

SMX: 192 single-precision CUDA cores, 64 double-precision units, 32 special function units (SFU), and 32 load/store units (LD/ST).

Share Mem

192 SP cores,

64 dp cores

32 ldst units

Interesting machines

- Convey: Alan Wallach, Convex founder
- Parallela Personal super: Kickstart project
- Blue Brain

Convey Architecture

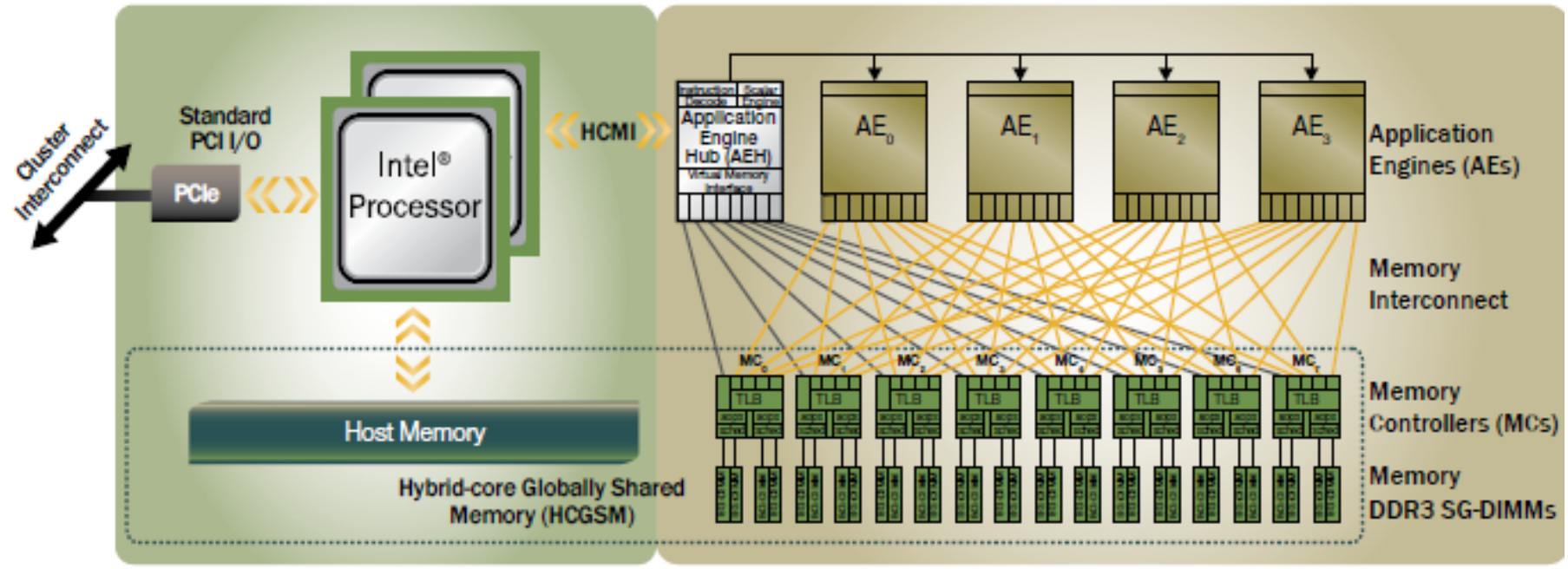


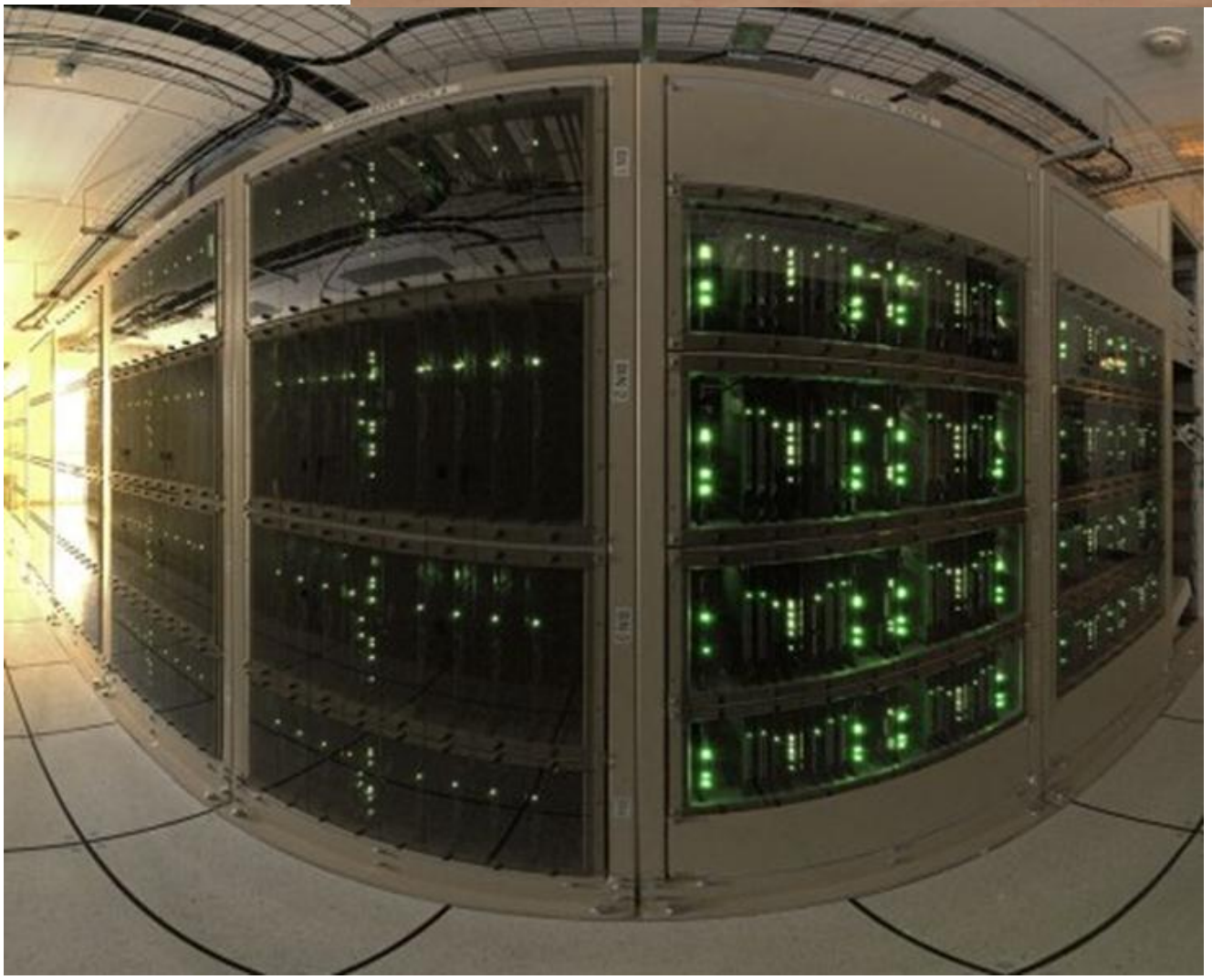
Figure 3. Overview of the Convey hybrid-core computing architecture.

64 …1,024 GB memory

	Rank	Machine	Installation	Nodes	Cores	Scale	GTEPS	GTEPS/kW
★	48	thunder4	Convey Computer Corporation	1	12	27	11.61	16.59
★	49	Celero	Argonne National Laboratory	1	8	27	11.45	16.35
★	50	Convey HC-2ex	UC Riverside	1	8	27	11.45	16.35
★	45	fox6	Convey Computer Corporation	1	16	29	14.56	13.87
★	62	Vortex	Convey Computer Corporation	1	4	27	6.64	9.48
	22	Vesta	DOE/SC/Argonne National Laboratory	1024	16384	34	382.00	6.10
	123	Scott Beamer's iPad	UC Berkeley	1	2	14	0.03	6.08
	2	DOE/SC/Argonne National Laboratory Mira	Argonne National Laboratory	32768	524288	39	10461.00	5.22
	30	BlueGene/Q	Brookhaven National Laboratory	1024	16384	34	294.29	4.70
	1	DOE/NNSA/Lawrence Livermore National I	Lawrence Livermore National Laborat	65536	1048576	40	15363.00	2.91
	77	Intel(R) Xeon(R) CPU E7-4870 @ 2.40GHz	Chuo University	1	40	30	2.16	2.21
	3	JUQUEEN	Forschungszentrum Juelich (FZJ)	16384	262144	38	5848.00	1.95
*	52	Grace	Mayo Clinic, Rochester	64	64	31	10.94	0.50
	21	TSUBAME 2.0	GSIC Center, Tokyo Institute of Techno	1,366	16,392	35	462.25	0.48
	111	ultraviolet	Sandia National Laboratories		32	29	0.42	0.42
	46	Altix ICE 8400EX	SGI	256	1024	31	13.96	0.36
	72	GPU-based cluster	Seoul National University, Korea	8	192	26	4.06	0.25
	93	PowerEdge R815 Opteron 6174	STE Lab, Nagoya University	4	192	22	1.16	0.17
*	101	XMT2	CSCS	64		31	0.86	0.04
**	56	Lonestar	TACC	1024	12288	35	9.23	0.03
	32	DOE/SC Hopper	NERSC/Lawrence Berkeley National La	4817	115600	35	254.07	0.03
	90	XMT	Sandia National Laboratories	128		29	1.26	0.02
*	91	cougarxmt	Pacific Northwest National Laboratory	128		29	1.18	0.02
*	94	graphstorm	Sandia National Laboratories	128		29	1.07	0.02
	108	Hyperion + FusionIO	Lawrence Livermore National Laborat	64	512	36	0.60	0.01
	124	Gordon	San Diego Supercomputing Center	7	84	29	0.02	0.00

Figure 4. Performance and power on the Graph 500 benchmark.

NRAO One-of: 17 peta-ops

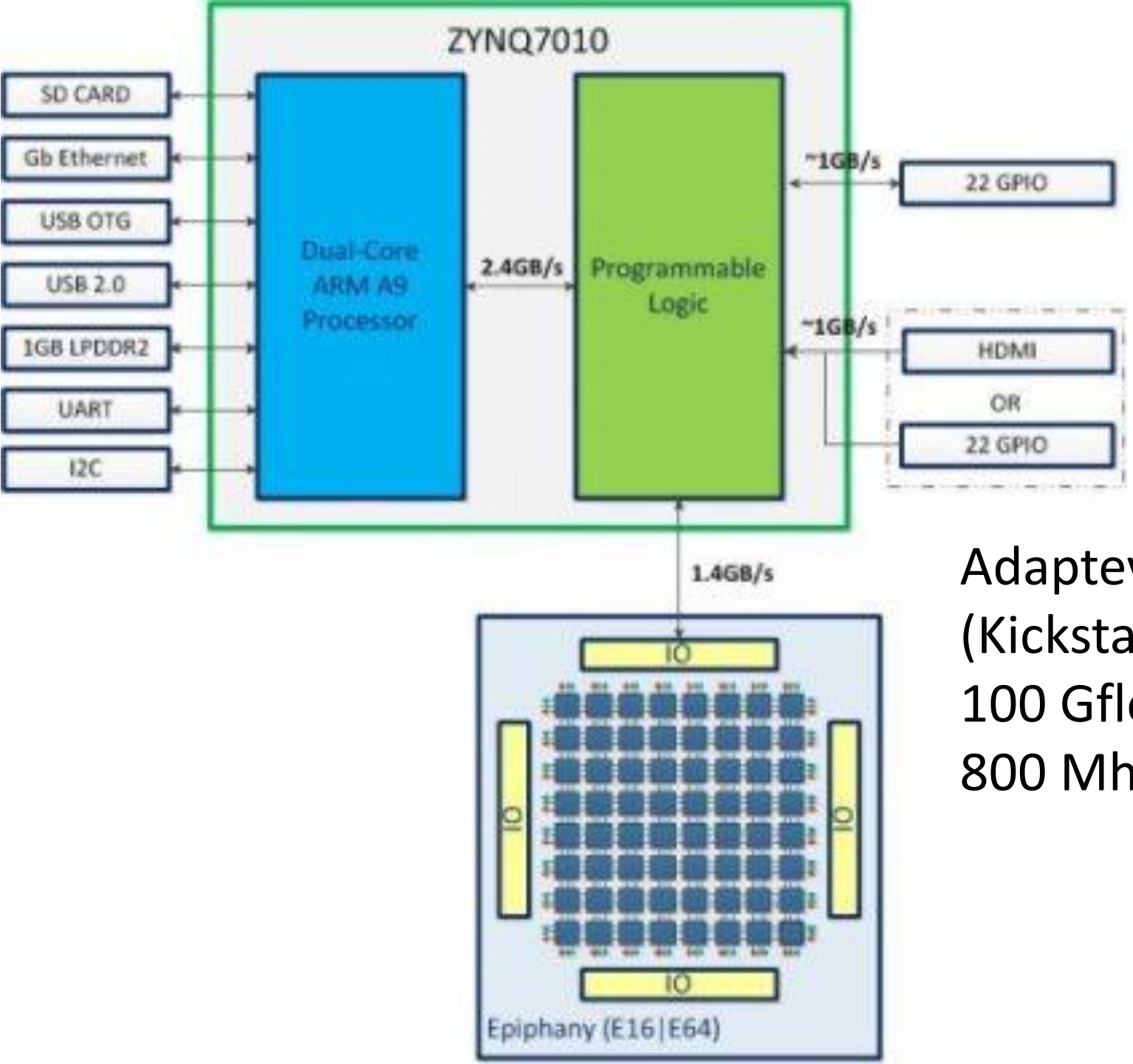


Cost: $10 Million
Power: 140 KW
32K custom
@125 MHz;
500 Gops, 1.8 w
64/board

Gains:
Power 100 X
Cost 10 X

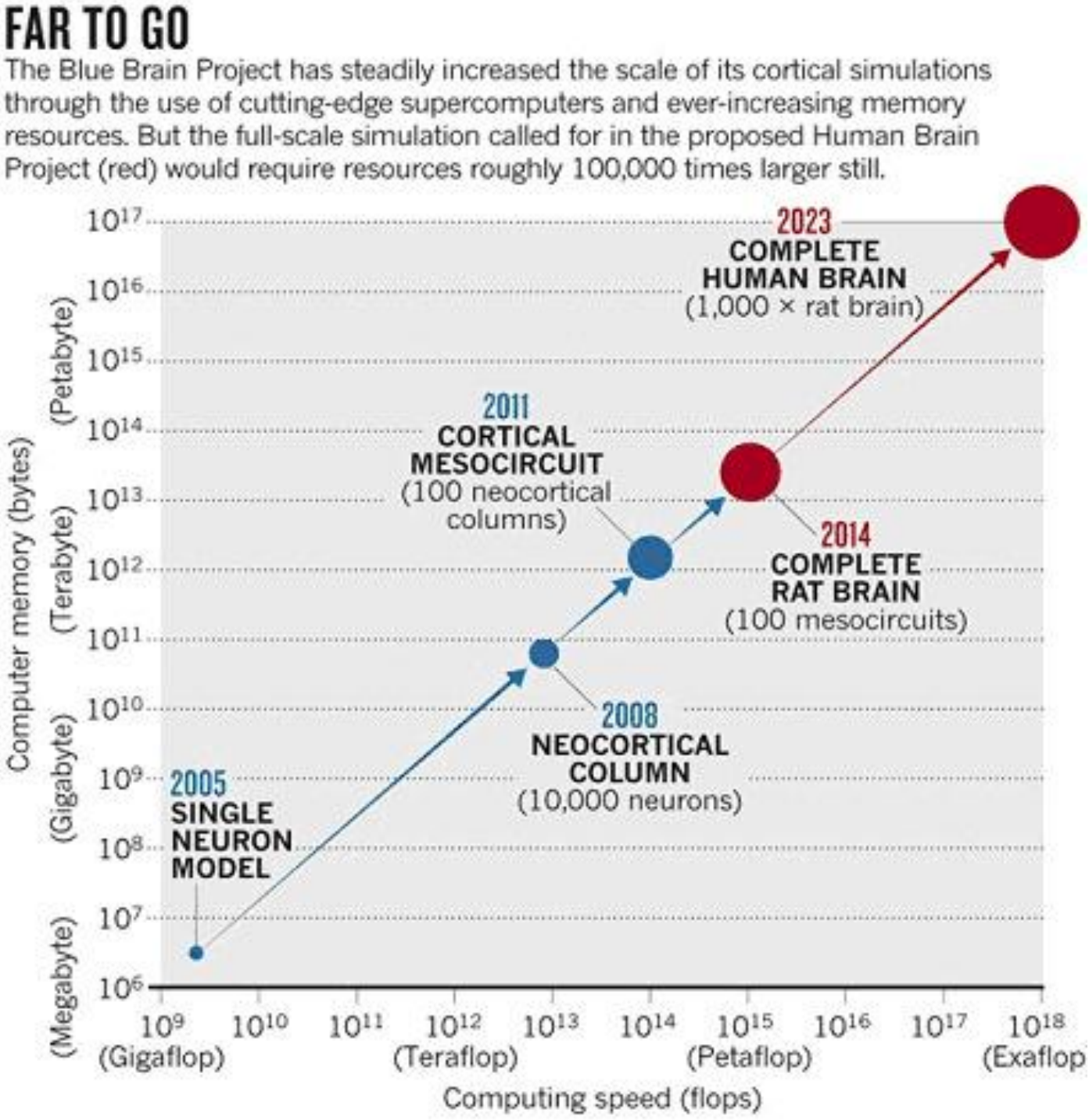
Parallela Computer

- Zynq-7010 Dual-core ARM A9 CPU
- Epiphany Multicore Accelerator (16 or 64 cores)
- 1GB RAM; MicroSD Card
- USB 2.0 (two) ; Ethernet 10/100/1000; HDMI connection
- Ubuntu OS and open source Epiphany development tools that include C compiler, multicore debugger, Eclipse IDE, OpenCL SDK/compiler, and run time libraries.
- Dimensions are 3.4'' x 2.1''
- A 64-core version of the Parallella computer delivers over 90 GFLOPS, comparable to a theoretical 45 GHz CPU [64 CPU cores * 700MHz] on a credit card size boardwhile consuming only 5 Watts.
- Epiphany-IV and Epiphany-III processors http://www.coremark.org and blog post here.
- Epiphany-IV processor was designed in a leading edge 28nm process and started sampling in July, demonstrating 50 GFLOPS/Watt. Epiphany energy efficiency specs are near 2018 DARPA Exascale goals



Adapteva: Parallela
(Kickstarter $s)
100 Gflops
800 Mhz; 2 Watts


XILINX



Blue Brain Project…

Build a complete human brain

End Top500 etc.

Potential System Architecture

Systems	2012 Titan Computer
System peak	27 Pflop/s
Power	8.3 MW (2 Gflops/W)
System memory	710 TB (38*18688)
Node performance	1,452 GF/s (1311+141)
Node memory BW	232 GB/s (52+180)
Node concurrency	16 cores CPU 2688 CUDA cores
Total Node Interconnect BW	8 GB/s
System size (nodes)	18,688
Total concurrency	50 M
MTTI	?? unknown

20xx

1 Exaflop/s
??

??

Fill in the blanks for each of the characteristics

Sketch the interconnect

To think about

Where has the performance come from in x?

How many operations can be in process for the various machines?

What's differentiates IBM, Fujitsu, Cray and Convey architectures? Start with some metrics…

Which one would you select assuming an operation i.e. flops/$ are the same? For what?

Metrics from a user pov? Time|Cost for x|Cost to program.

Going out x years, when is an exaflops computer?

What will an exaflops computer look like in ?

More to think about

- Given an environment e.g. body area, home, car, small business, an industrial structure, what is the structure and IT taxonomy of the network, computers, storage, etc. showing function
  - Now, in 5 years, in 10 years
- How will Moores's Law change computing
  - In 5 years, in 10 years
- What new computers could you envision that Bell's Law might enable
  - In 5 years, in 10 years

More to think about

- Name, classify, and construct a taxonomy of all the platforms i.e. dominant programming environments after the PC, WIMP*
  - What year
  - Programming environment
  - Key apps
- What will IoT add to the platform and classes?

*These could encompass Internet 1.0 and 2.0

The end