The Challenge of Designing and Using HPC Programs

Man vs Machine

Petascale Computing Insitute

Gordon Bell

The Argonne National Laboratory, the Blue Waters project at NCSA, the National Energy Research Scientific Computing Center, Oak Ridge Leadership Comp Facility, Pittsburgh Supercomputing Center, SciNet at the University of Toronto, and the Texas Advanced Computing Center.

You will be surprised…

- In hindsight, FORTRAN c1957 has hidden parallelism --until 90s
  - Last 25 years: High placed fruit--f (ladders, balloons, aircraft)
- Performance gain of 1000 per decade … 10K hour or 5 year mastery **change in speed, form of parallelism, algorithms**
- New computing paradigms after experimentation and theory
  - 3rd Paradigm: Ken Wilson, Simulation …mid 80's
  - Visualization rediscovered every decade to deal with data
  - 4th Paradigm: Data discovery …Gray 2008
  - AI, Machine learning, neural nets, to uncover or expose the phenomena …2018
- Enables new kinds of science and engineering

Topics

- Moore's Law vs Algorithms … What can go wrong?
- The Ideal supercomputer. Quick visit to the machines, parallelism, and performance….
- **Capability Computng**: Sunway and Global Simulation (Bell Prize, 2017)
- Juggling: Mastering the options for parallelism? **Capability? or Capacity?**
  Job Stream and Ensemble (embarasing parallelism), lots of jobs
  Shared Memory, **the first 34 years of my career until I gave up**
  MPI (Commpunicating Sequential Processes),
  SIMD
- Paradigms …new systems and careers
- The Petascale program
- Architecture will continue: Specialized chips versus programmable chips

Climate Change Doesn't Just Happen

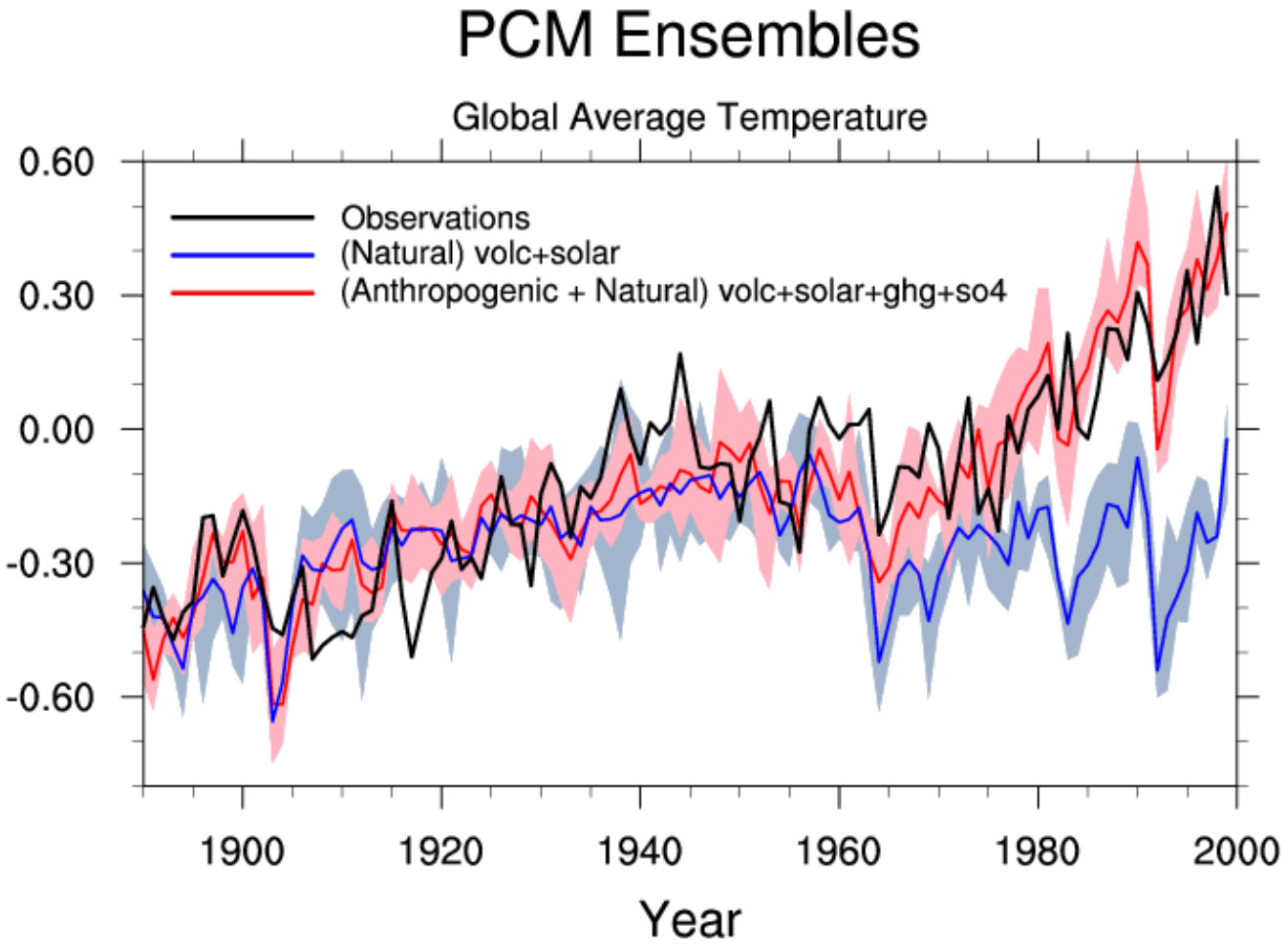
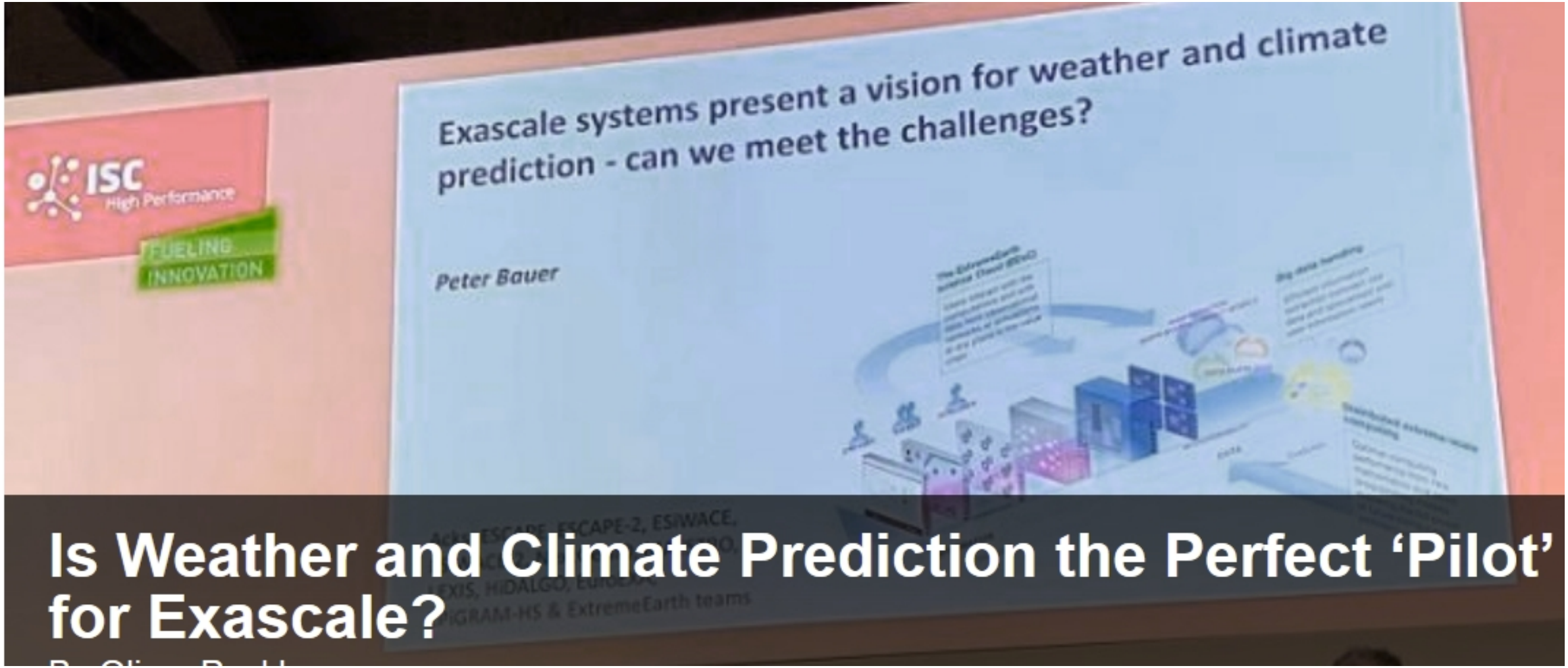
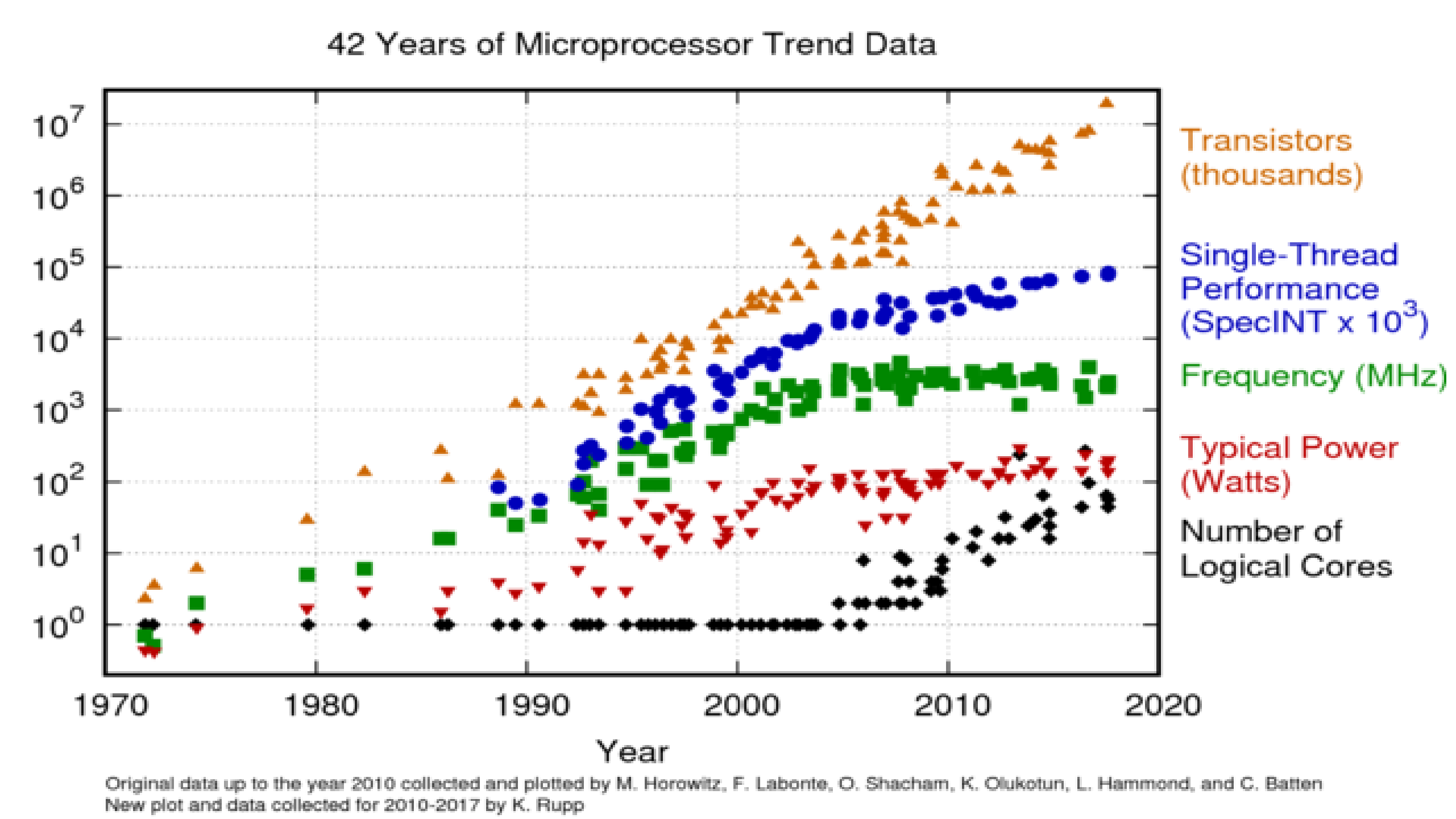


Fig. 5. An artist's impression of Richardson's Forecast Factory (© François Schuiten).

Blue Waters hurricane simulation & visualization

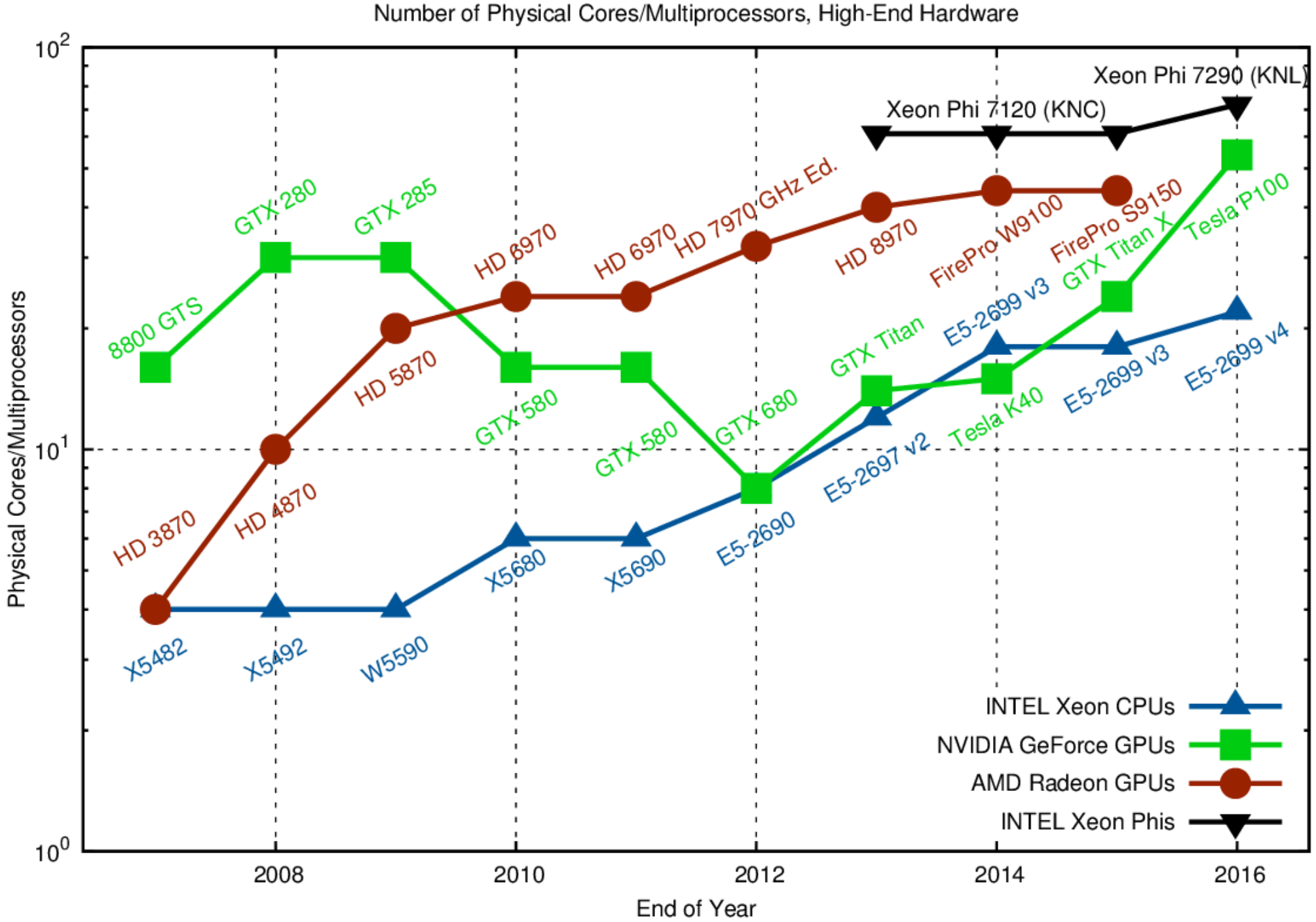
European Centre for Medium-Range Weather Forecasts (ECMWF) ISC19




42 Years of Microprocessor Trend Data
10^7
10^6
10^5
10^4
10^3
10^2
10^1
10^0
1970
1980
1990
2000
2010
2020
Year
Transistors (thousands)
Single-Thread Performance (SpecINT x 10^3)
Frequency (MHz)
Typical Power (Watts)
Number of Logical Cores
Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten
New plot and data collected for 2010-2017 by K. Rupp

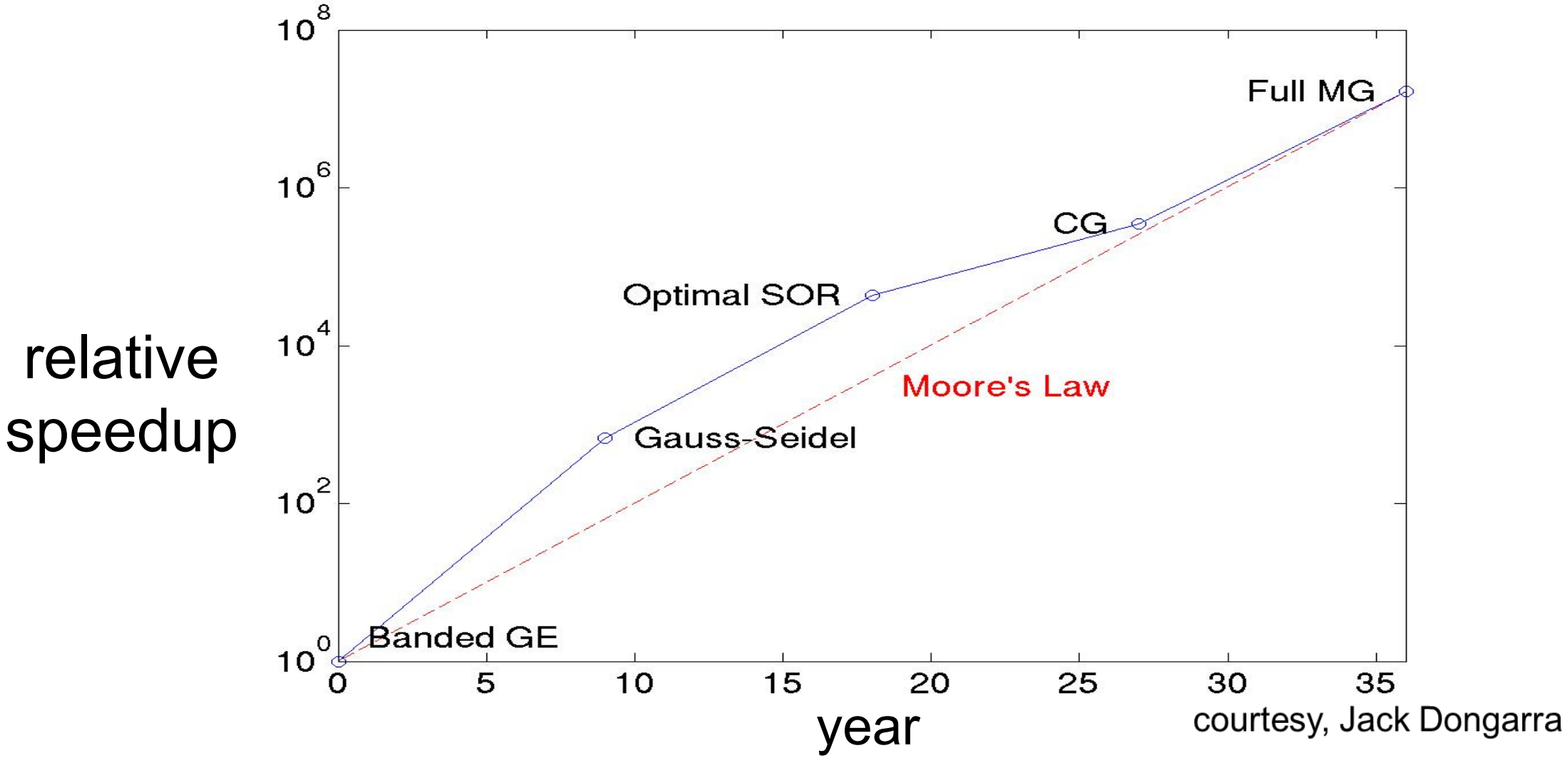
No magic bullets for exa-ops

- Op rate is constrained by a 2 -3 GHz clock
- Exa ops: 10^{18} / 2 x10^{9} or a half a billion-fold parallelism
- 1000s -10,000s Independent, interconnected computers
- Intra-computer parallelism
  - SIMD e.g. GPU
  - Specialized chip e.g. Google TPU
  - FPGA e.g. unrolling the loops
- Simpler ops versus double precision floating helps
  - 8 or 16 bit ops -> 16x


Number of Physical Cores/Multiprocessors, High-End Hardware
Physical Cores/Multiprocessors
End of Year
10^2
10^1
10^0
2008
2010
2012
2014
2016
Xeon Phi 7290 (KNL)
Xeon Phi 7120 (KNC)
8800 GTS
GTX 280
GTX 285
GTX 580
GTX 580
GTX 680
GTX Titan
GTX Titan X
Tesla P100
Tesla K40
HD 3870
HD 4870
HD 5870
HD 6970
HD 6970
HD 7970 GHz Ed.
HD 8970
FirePro W9100
FirePro S9150
X5482
X5492
W5590
X5680
X5690
E5-2690
E5-2697 v2
E5-2699 v3
E5-2699 v3
E5-2699 v4
INTEL Xeon CPUs
NVIDIA GeForce GPUs
AMD Radeon GPUs
INTEL Xeon Phis

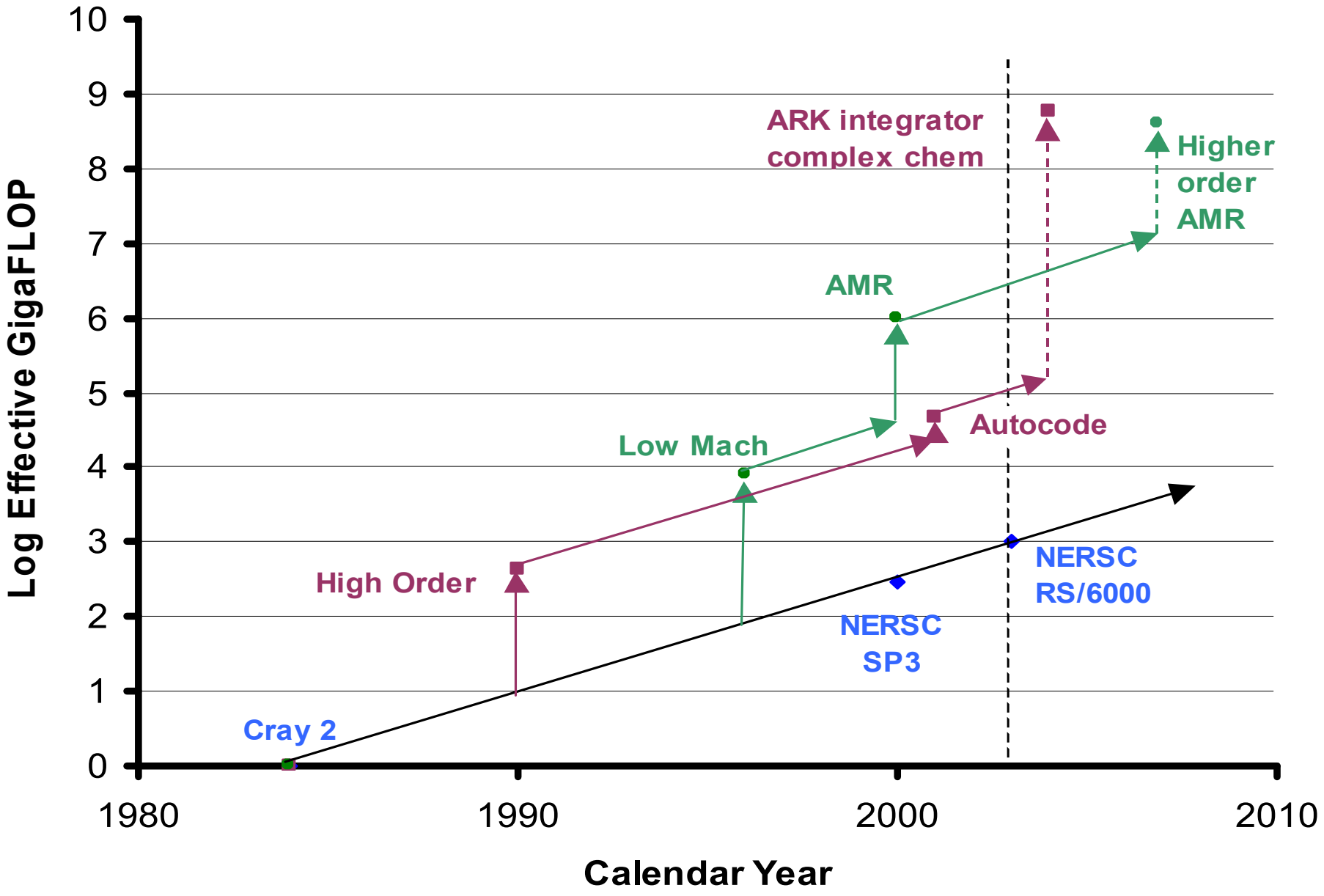
Algorithms and Moore's Law

- Advances over 36 years, or 24 doubling times for Moore's Law
- $2^{24} \approx 16$ million $\Rightarrow$ the same as the factor from algorithms alone!



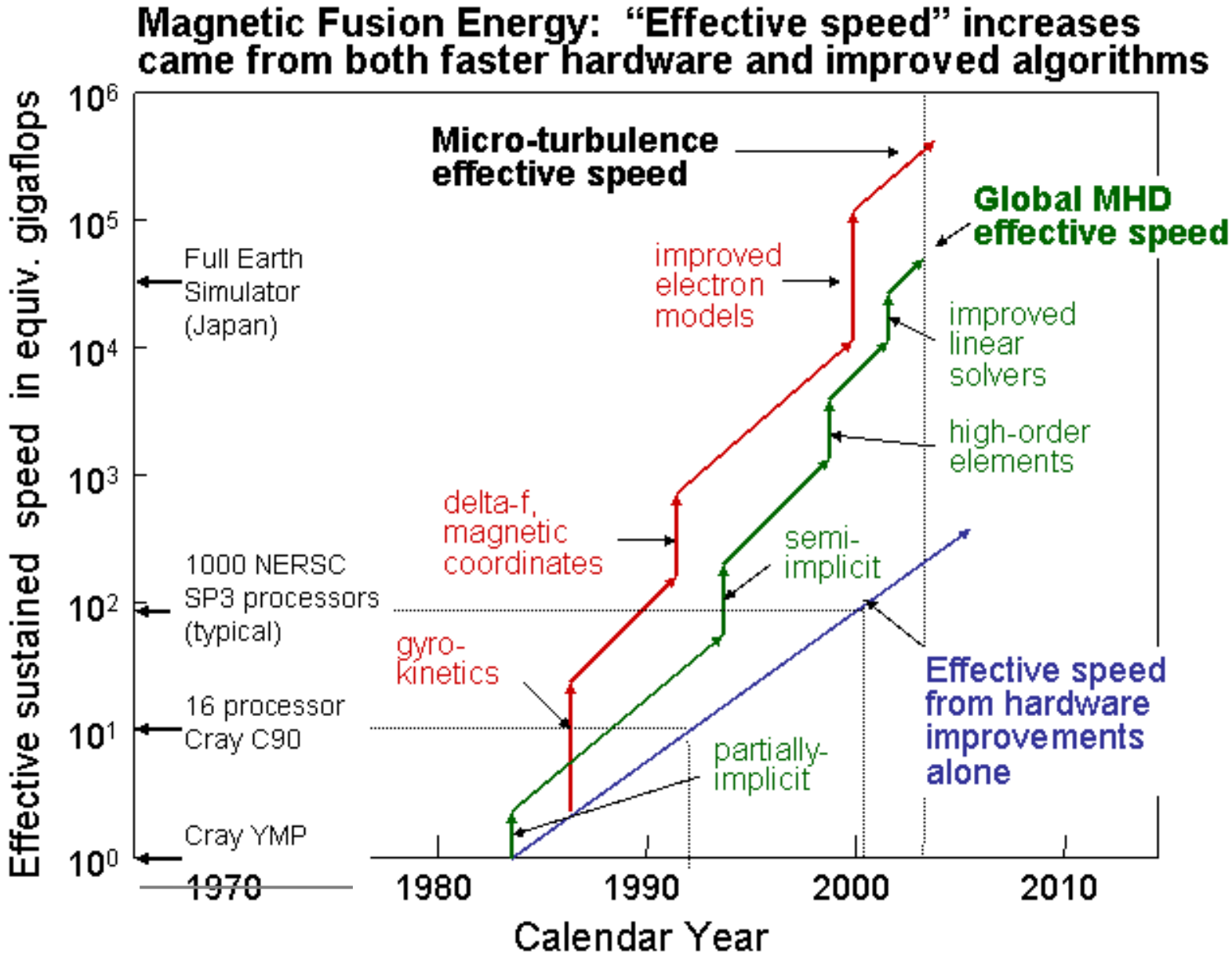
courtesy, Jack Dongarra

“Moore’s Law” for combustion simulations



ack Dongarra

“Moore’s Law” for MHD simulations



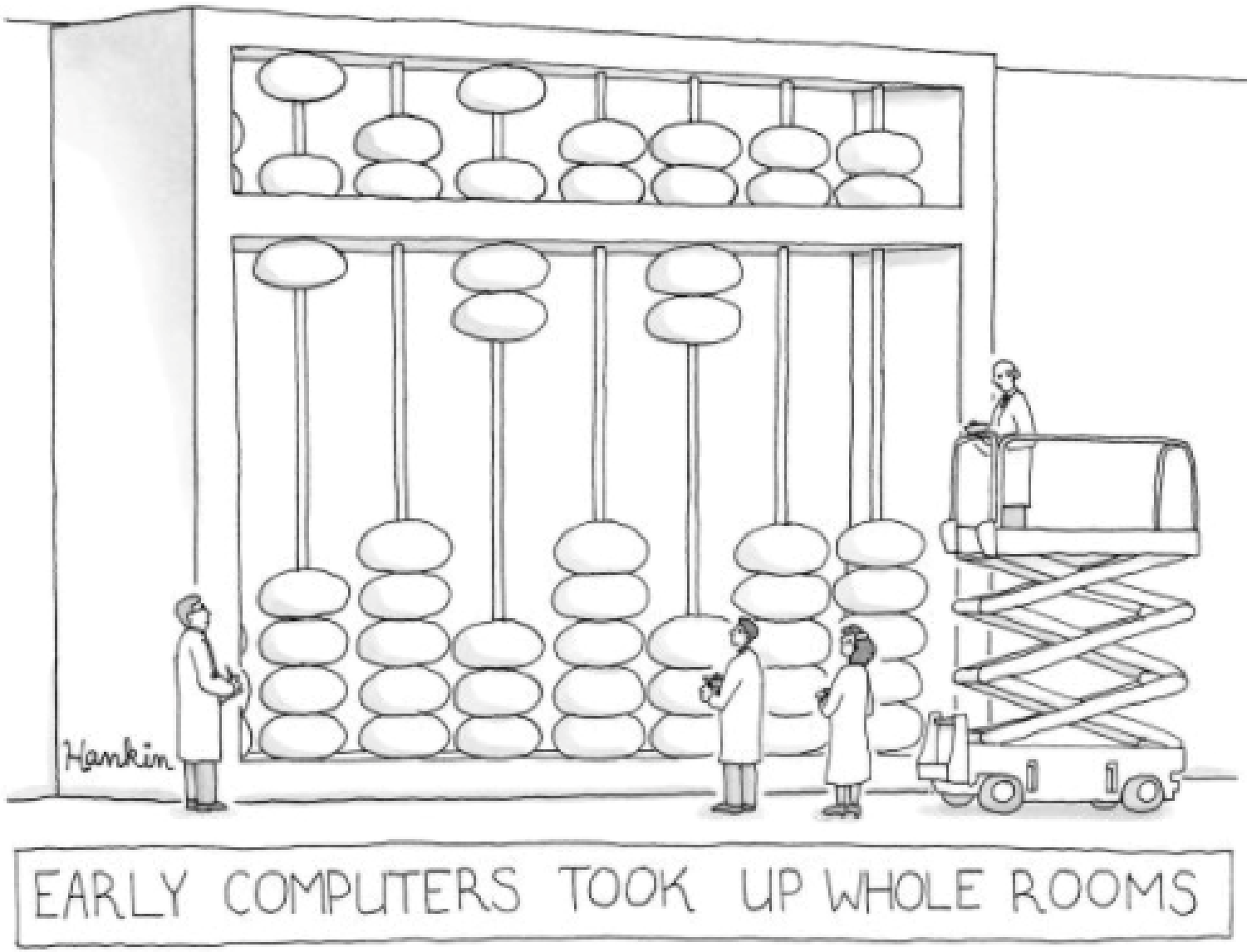
“Semi-implicit”:

All waves treated implicitly, but still stability-limited by transport

“Partially implicit”:

Fastest waves filtered, but still stability-limited by slower waves

courtesy, Jack Dongarra



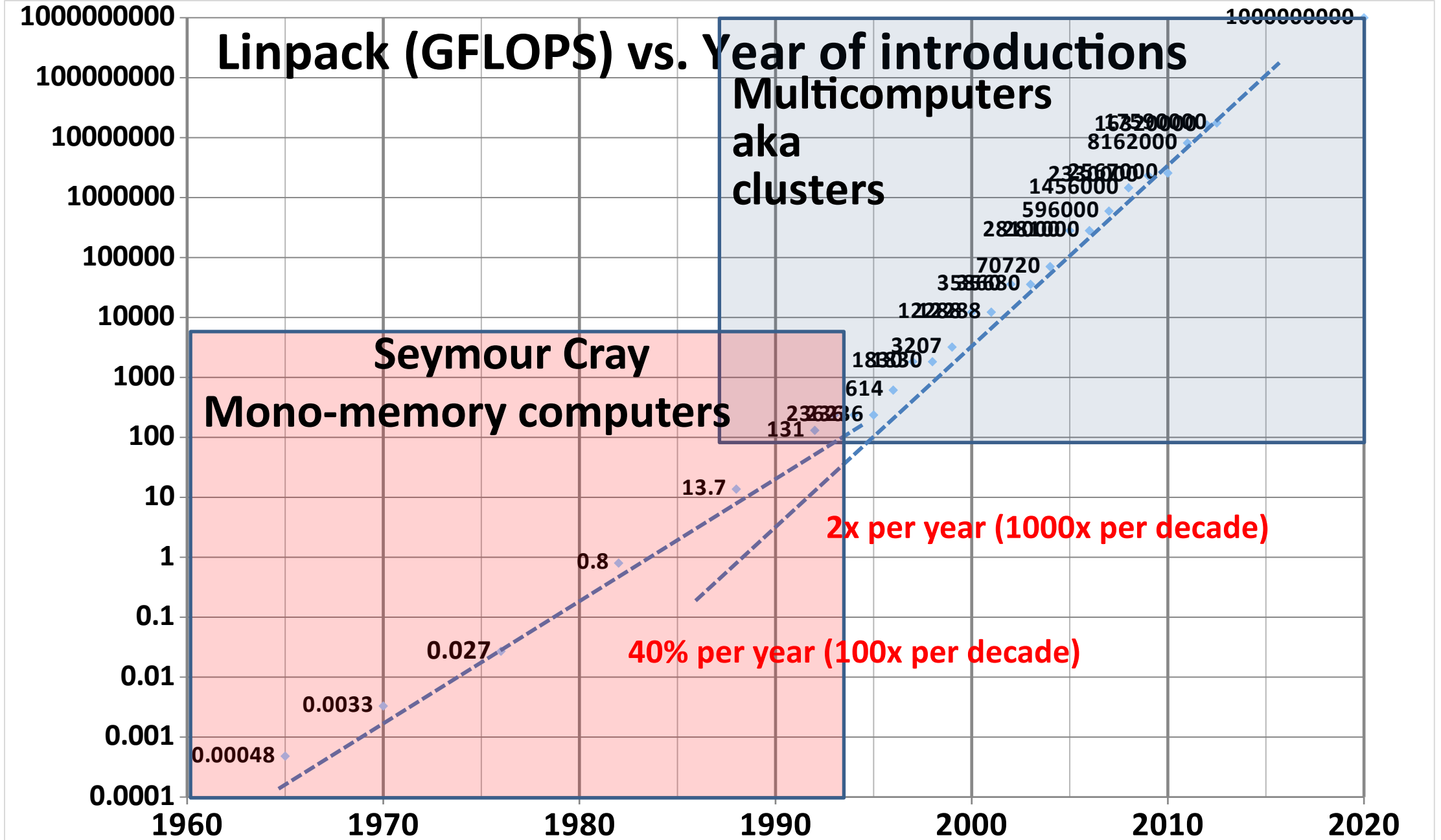
… and we started calling them supercomputers

The *ideal* Supercomputer:
Speed, memory, and parallelism (scaling)

- **Clock speed that increases with time**
- **One, very large and scalable memory for any number of processors**
  - **Overlap of memory access and instruction execution**
  - **Parallelism of a single instruction stream including look-ahead and execution**
  - **Pipelining**
  - **Vector processing**
  - **Multiprocessors—(Scale up)**
  - **Multiple streams & multi-threading scalability**
- **Multiple independent interconnected computers (Scale out) aka clusters aka multicomputers**
  - **Multiprocessor nodes aka constellations**
  - **Multi-threading and vector processing**
  - **Stream processing using GPUs**
  - **Direct compilation of algorithms into FPGAs and application specific chips**

HPC aka Supercomputing ideas & machines: It's about speed, parallelism, standards, & cost (COTS)

1. The Cray Era (1964-1993): Mono-memory computer
   - Increased clock: high Khz => Mhz => low Ghz (10,000x)
   - Intra-processor parallelism techniques (10x > 100x vector)
   - Shared memory, multi-processor parallelism (1>32)
2. *"Killer Micros" transition (1984-1993)*
   - *Searching for the way… scalability, CMOS micros, standards, failures*
   - *Similar to search for the first supers, and for xxMD way.*
   - **1987… Prize to acknowledge parallelism**
3. The Multicomputer aka Clusters era (1984-present)
   - Parallelism: <10 => 1000x => 10,000 => 100,000 => million… billion
   - Now it is up to programmers to exploit parallelism



Top 20 HPC seminal events: Machines, recipes, standards, plans, and prizes

- ***1945,6 EDVAC Recipe & IAS Architecture.***
  Over a dozen IACS were built e.g. Illiac, Johniac, Maniac,
- **1957 *FORTRAN*** establishes standard for scientific computing… **FORTRAN 2008**
- **1960 LARC** and **1961 Stretch**—response to customer demand; **1962 Atlas** commissioned
- **1964 CDC 6600** (.48 MF) "first super" parallel function units.
  S R Cray begins 30 year reign as *"the supercomputer designer"*
- **1965 Amdahl's Law:** Defines difficulty to speed up computiation with various forms of parallelism hardware
- **1976 Cray 1 (26 MF)** first, practical **vector processor architecture.**
- **1982 Cray XMP…C90** (.5-16 GF) **Intro of shared memory mPv.**

The beginning of the end of mono-memory computing

- ***1993 CM5, 1024 multicomputer beats mono memory Cray***

I never pass up an opportunity to visit a computing centet\

SDSC
flash Gordon

2 PFlops
47K Cores
247 Tbytes
64 Tbytes, flash

Bridget Gordon Bell, C. Gordon Bell, Robert Sinkovitz (SDSC)

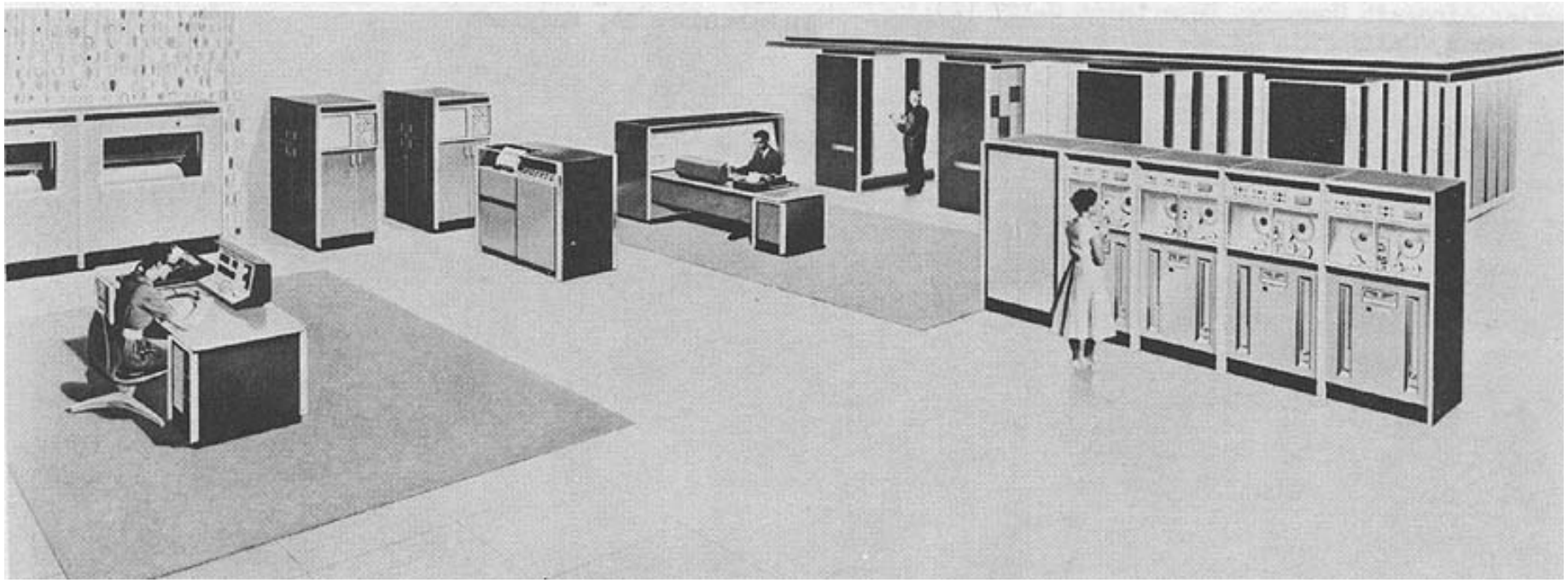
1960: “We need the largest computer you can build” UNIVAC, IBM, and Manchester U.

Three efforts to build the world’s largest computer establishes a unique computer class for scientific computing

- **UNIVAC LARC … Livermore /Univac spec’d, decimal (Univac/Eckert)**
- **IBM Stretch … lookahead, pipelining for LANL**
- **Manchester/Ferranti Atlas … paging and one-level store**

Mainframe? Supercomputer?: LARC

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



Courtesy of Burton Smith, Microsoft

LARC at LLNL c1960

Mainframe? Supercomputer?: Stretch, Harvest

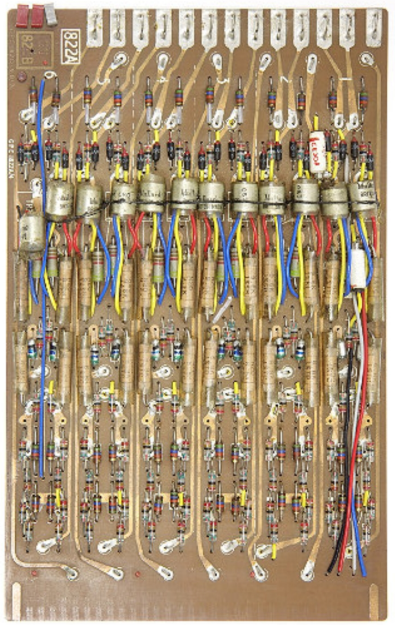
- **IBM 7030 (STRETCH)**
- **Delivered to Los Alamos 4/61**
- **Pioneered in both architecture and implementation at IBM**

- **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

IBM

Ferranti/Manchester Atlas c1961
(One million instructions per second)



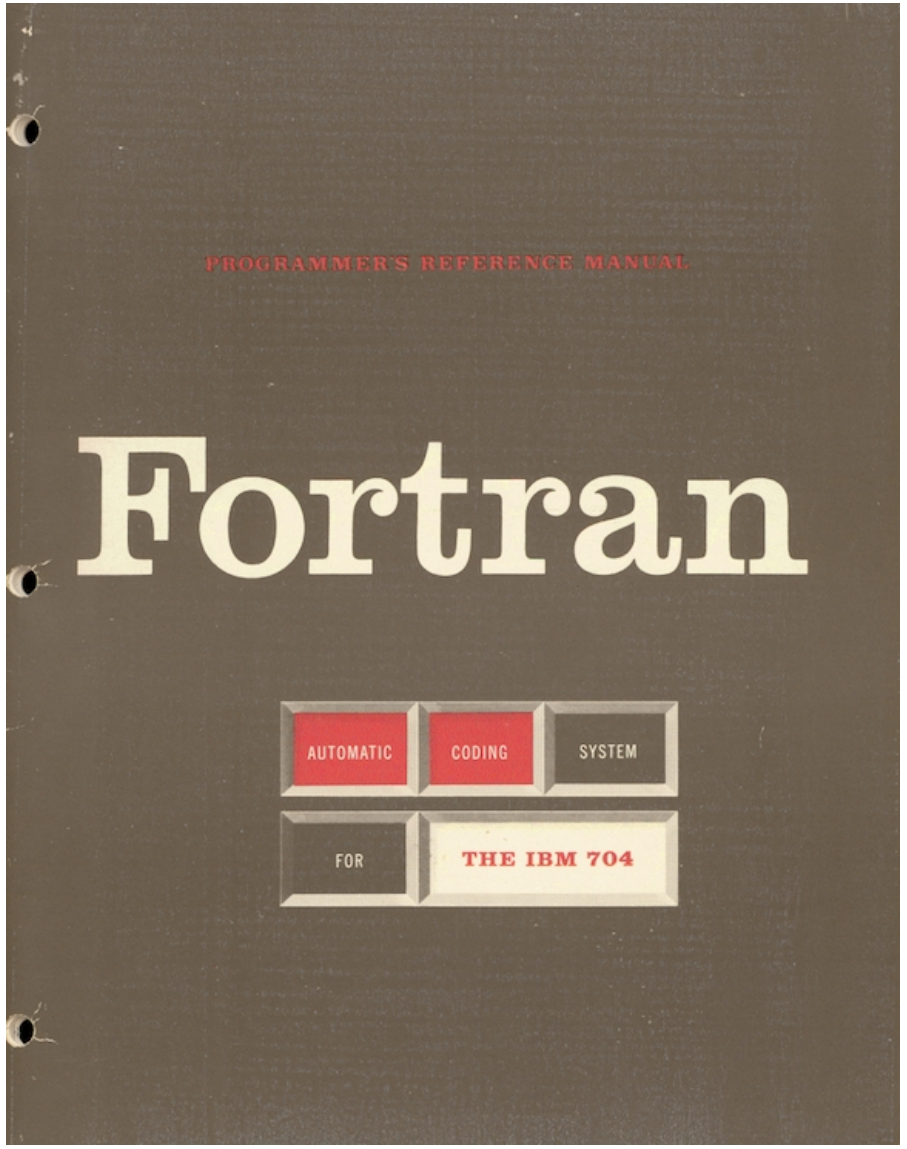
Mainframe? Supercomputer?

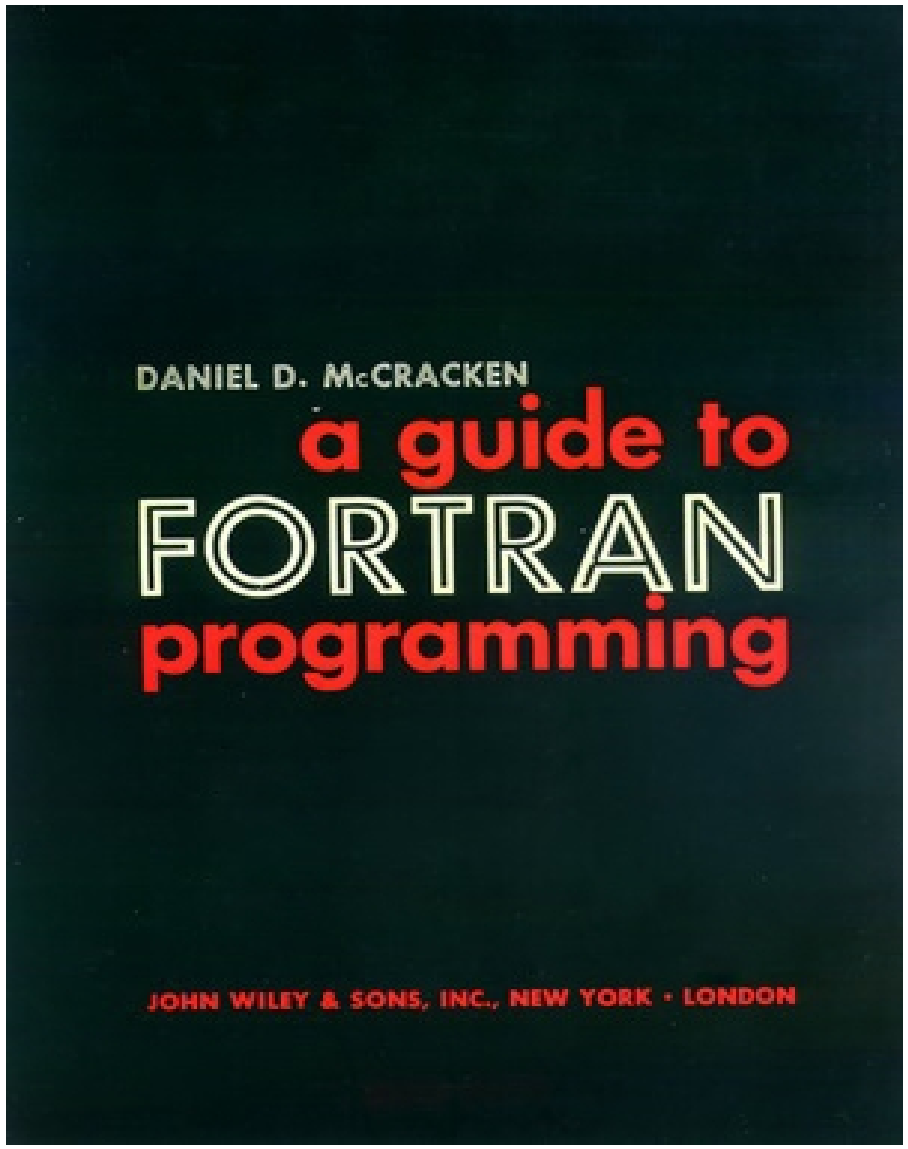
Mastering the skills of parallelistm

Pipelining

Look-ahead, …Multi-threading

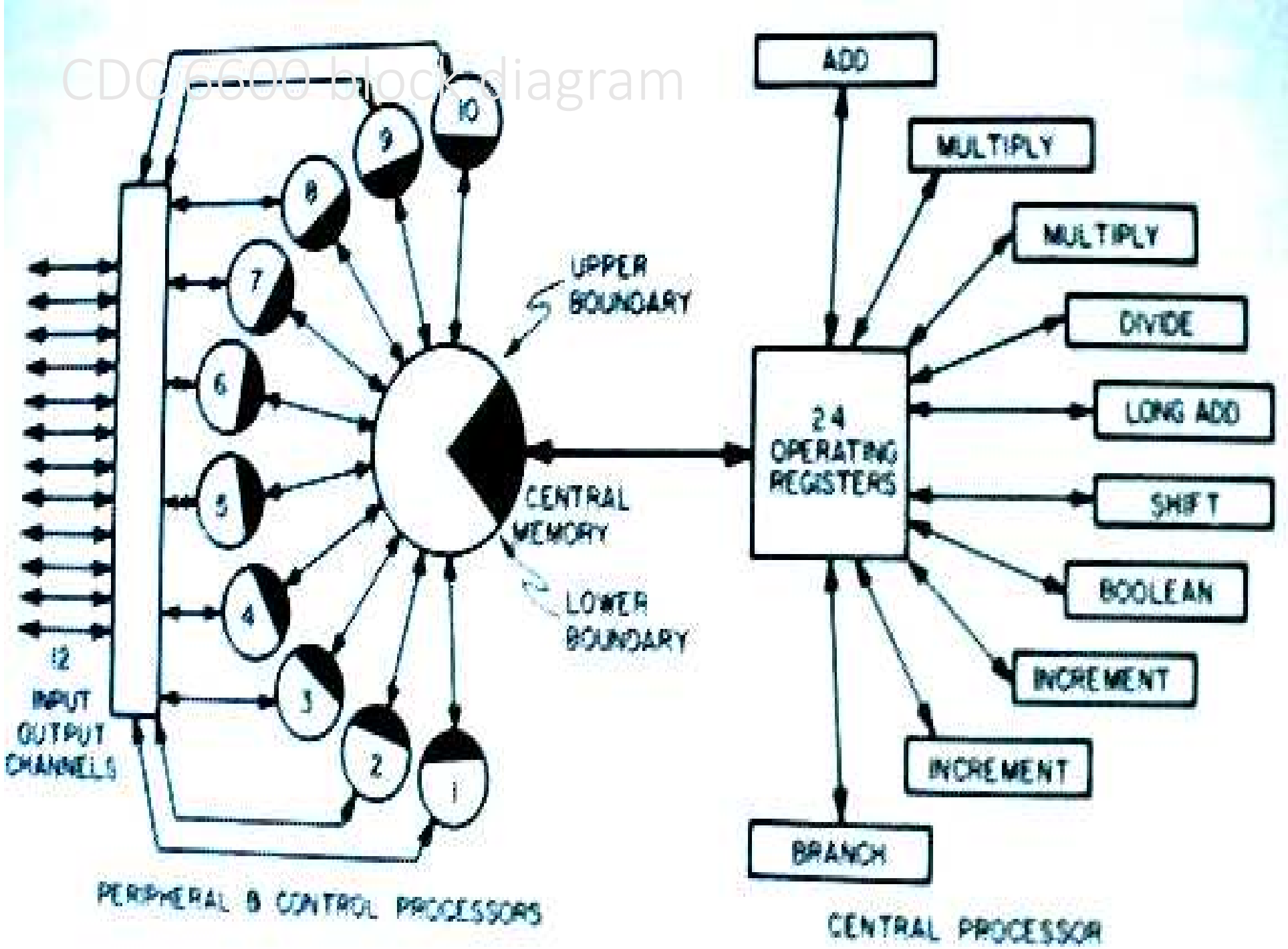
Fortran 1957, ’60, … ’08 Spec’ing a Supercomputer





CDC 6600 #1 Console & frame c1964 LLNL

First Supercomputer?



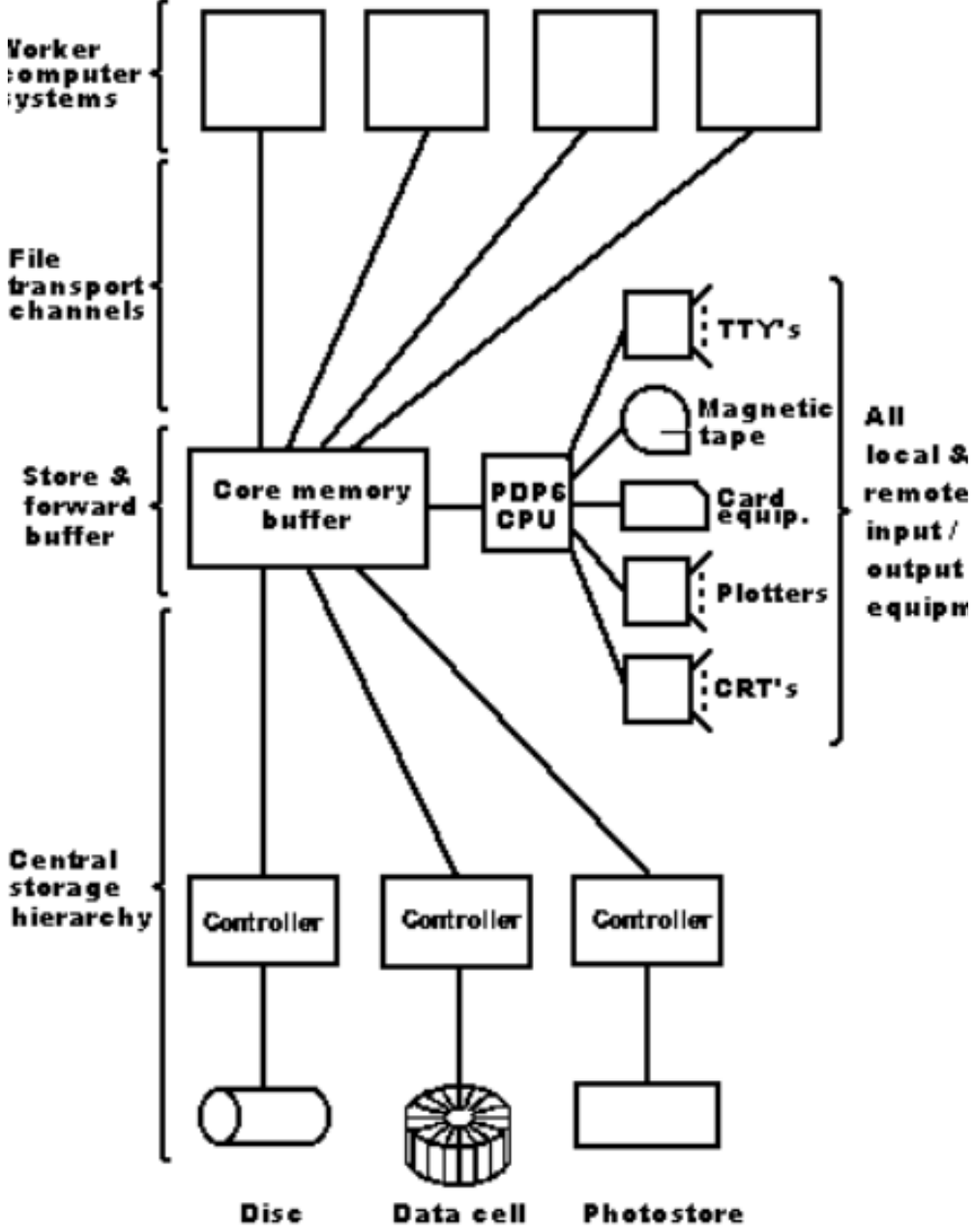
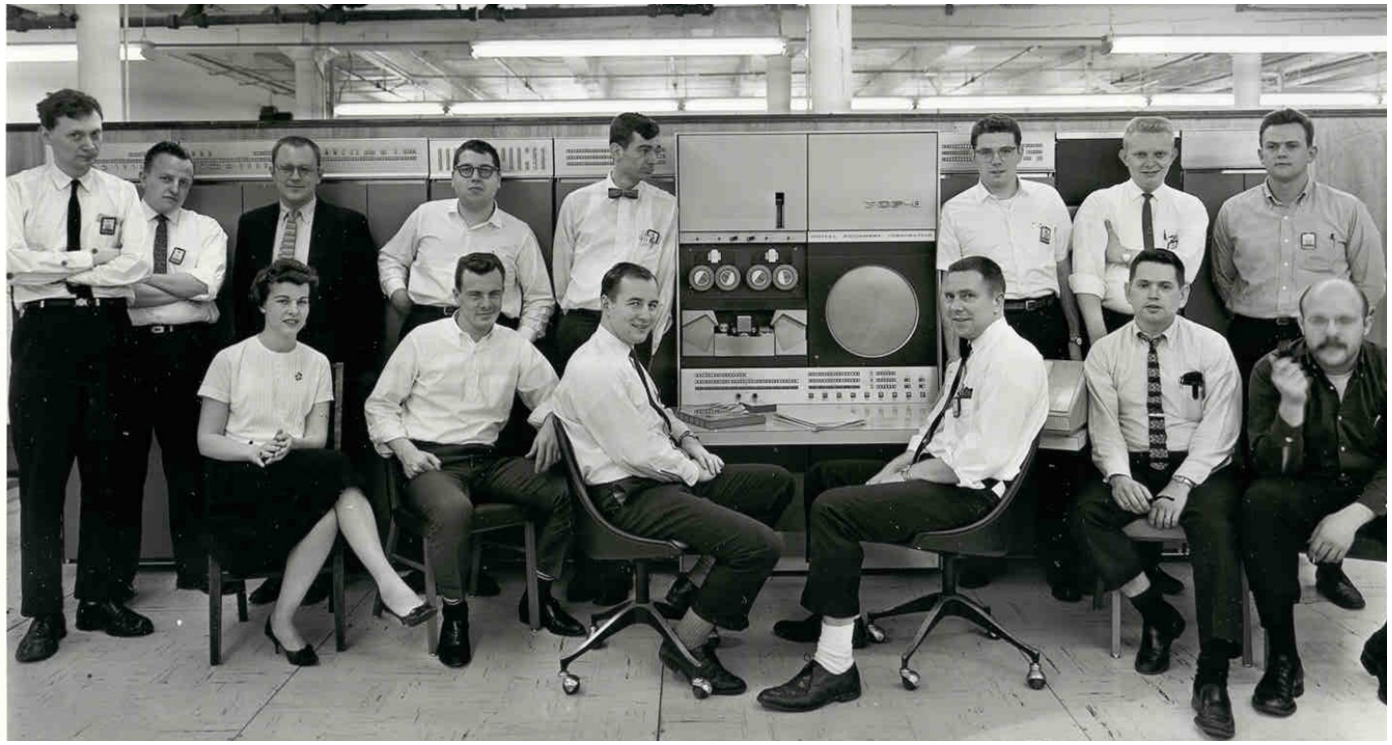


Figure 1

The Classical Octobus Concept

LLNL Octopus hub PDP-6 c1965
256K word, 36 bits/word



2 x 10 x 25 =500 modules; 5000 transistors?1900 watts

Thought in 1964
when I heard about the 6600:
"Holy s**t!" How did he do that?

- Digital's PDP- 6 was being built and introduced
- 10x less expensive ($300 K vs. $3 M)
- **6600 had 600K transistors**; 4 Phase, 10 Mhz clock
- **"6" had 5,000 transistors**, 2 bays x 10-5"crates x 25 = 500 modules, Clock ran asynchronously at 5 MHz.
  - Successor PDP-10 ran at 10 MHz

Two CDC 7600s and LLNL c1969

Courtesy of Burton Smith, Microsoft

Search for a Architecture parallel *constrained by Amdahl's Law*

Three approaches… that didn't work

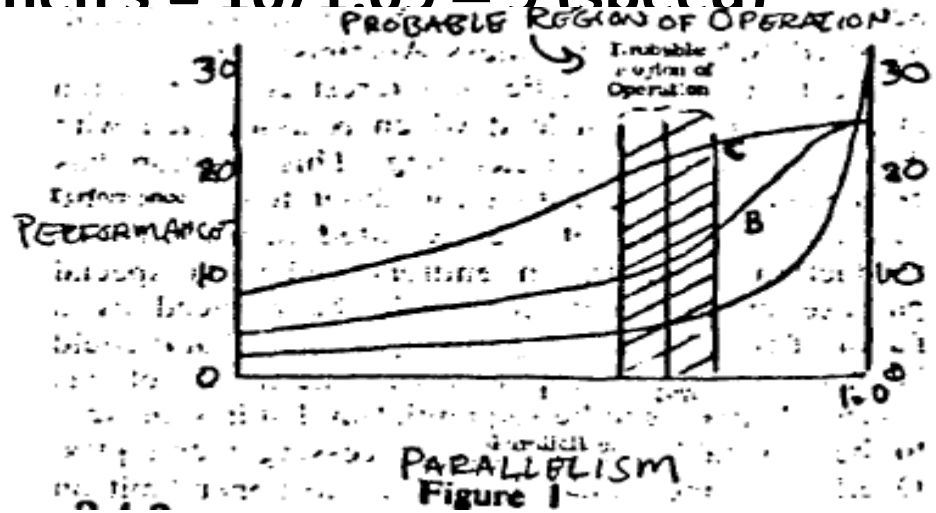
- **Illiac IV (SIMD) m=64**
- **CDC STAR and ETA10 vectors in memory**
- **TI ASC (Vector arch; Cray 1's 5x clock)**

Then success

- **Cray 1 vector architecture**

Amdahl's law c1967…the limit of parallelism

- **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 ≃ 9 (speed)**



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

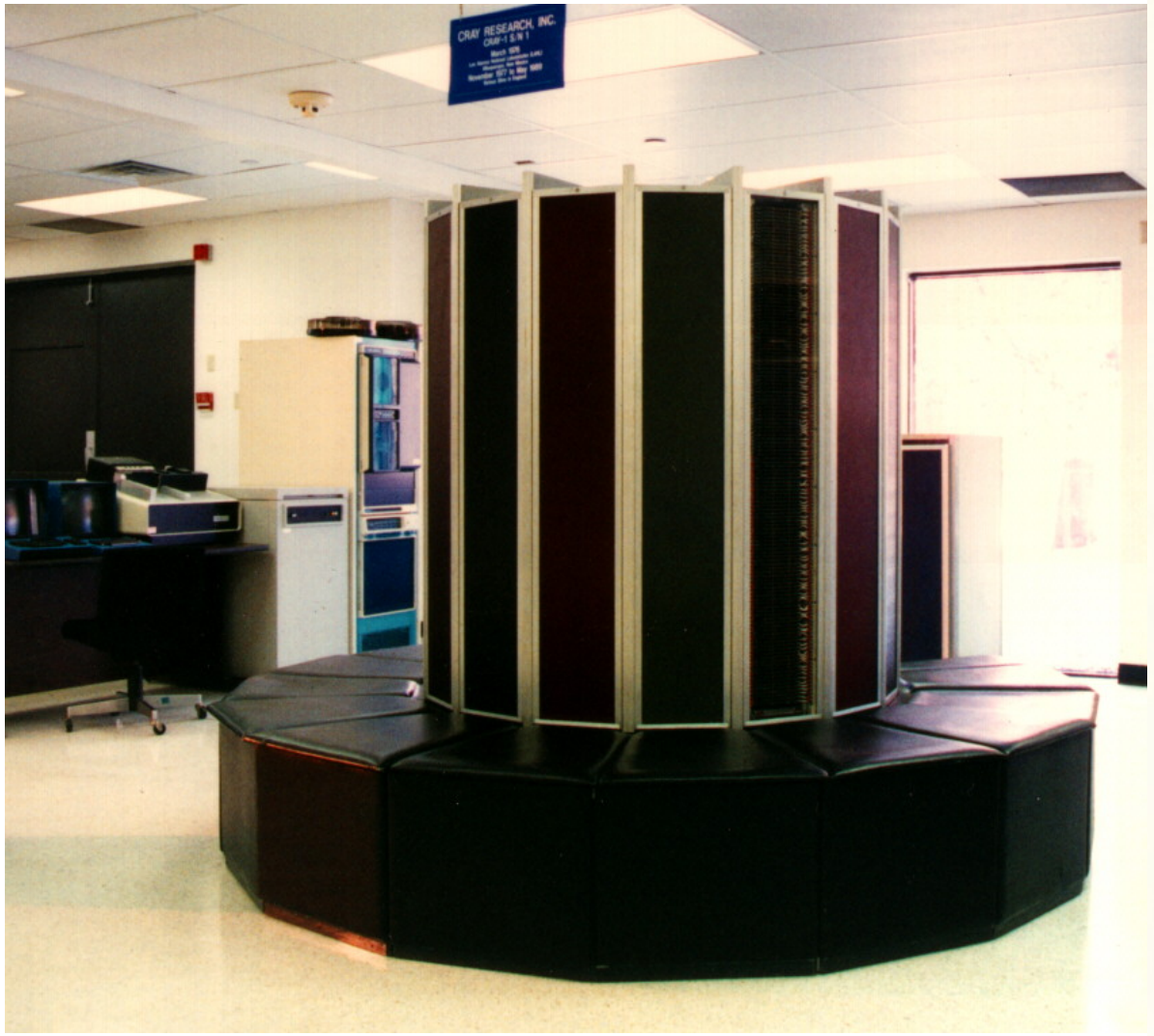
Courtesy of Burton Smith, Microsoft

ILLIAC IV: Uof IL at NASA in1971

- **1964 project (U. of IL)**
- **Burroughs contract to build**
- **SIMD 64 PEs**
- **10 MB disk/PE**
- **Moved to NASA**
- **1975 on ARPAnet**
- **Resides at the Computer History Museum**

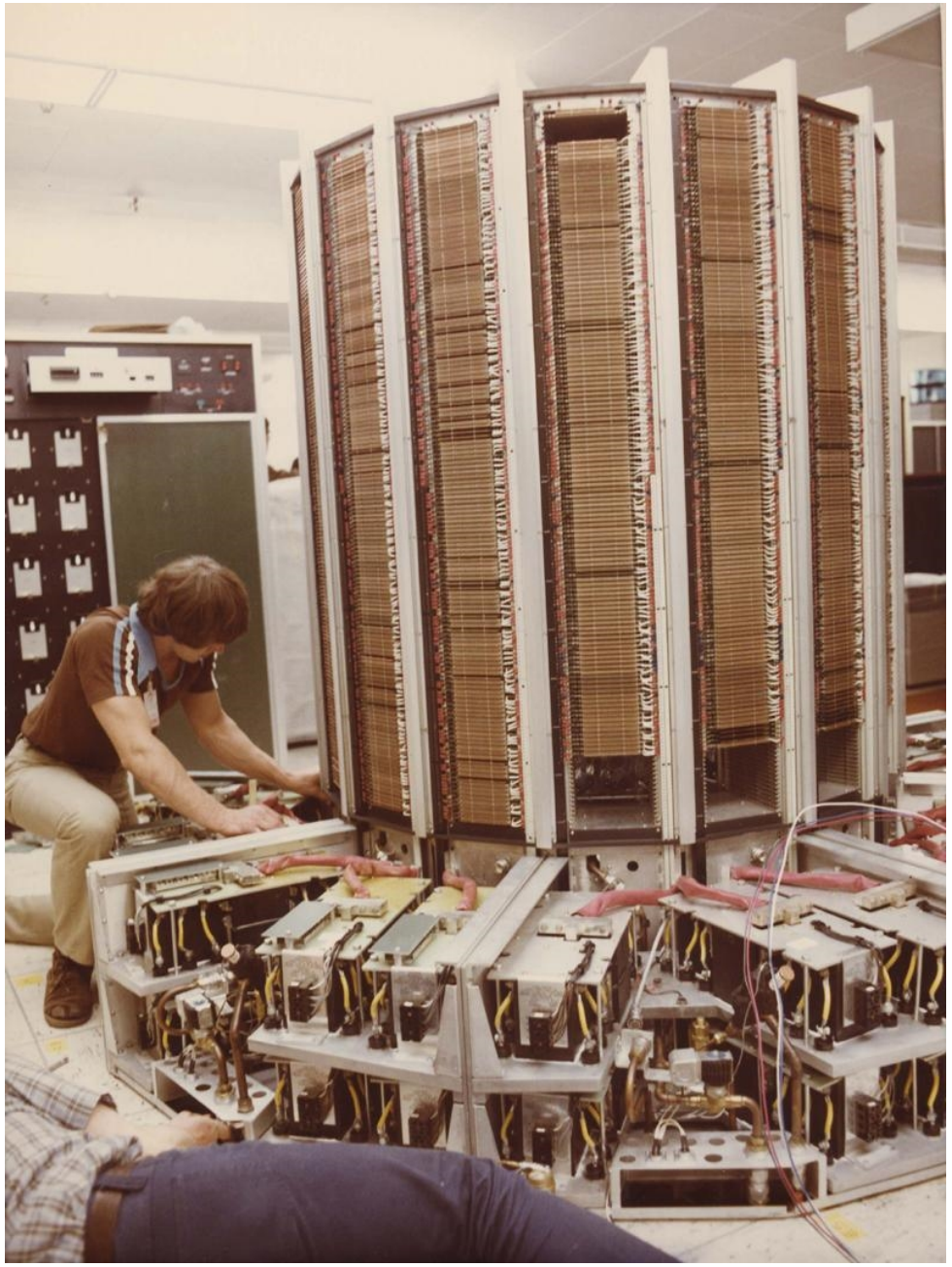
Courtesy of Burton Smith, Microsoft

Cray-1 c1976: Supercomputer vector era



Steve Squired, DARPA SCI 1985

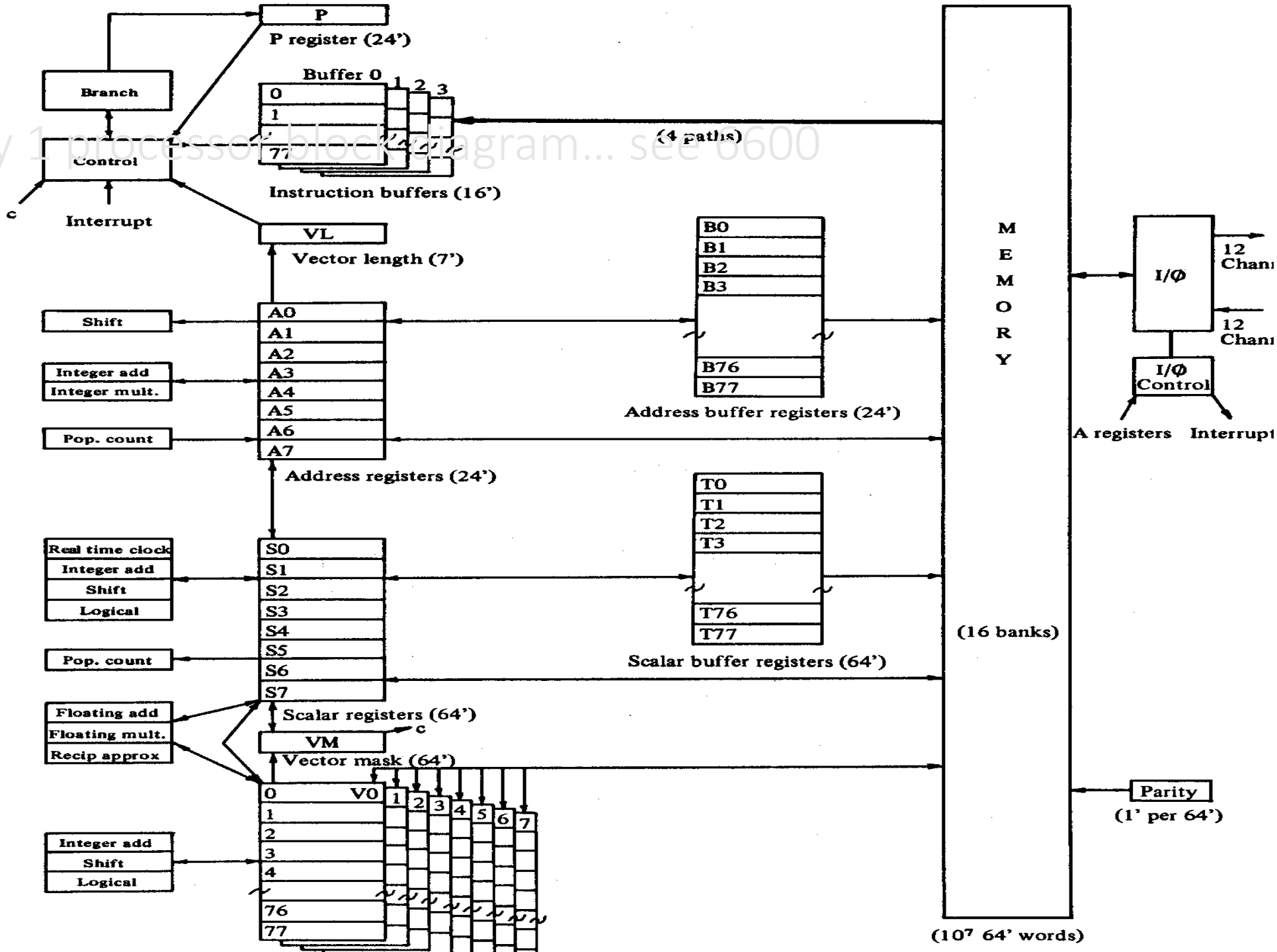
Courtesy of Burton Smith, Microsoft



Cray 1 sans covers
The Vector ISA

- Unlike the CDC Star-100, no development contract
- Los Alamos got a one-year free trial. Los Alamos leased the system.
- Los Alamos developed or adapted existing software
- Cray-1 and Amdahl's law
  - Scalar performance 2X the 7600
  - Vector 160 Mflops
  - 80 MHz clock
- Peak floating point ops vs. Instructions per second
- "Supercomputer" connotes a Cray-1

Courtesy of Burton Smith, Microsoft

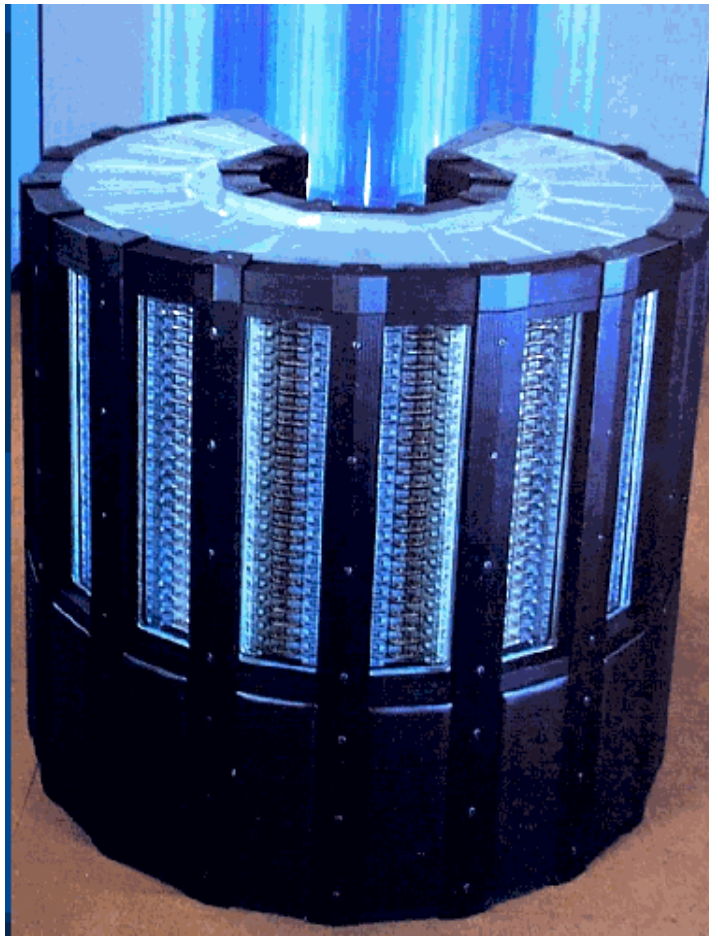


Cray 1 processor block diagram... see 6600

Shared Memory: Cray Vector Systems

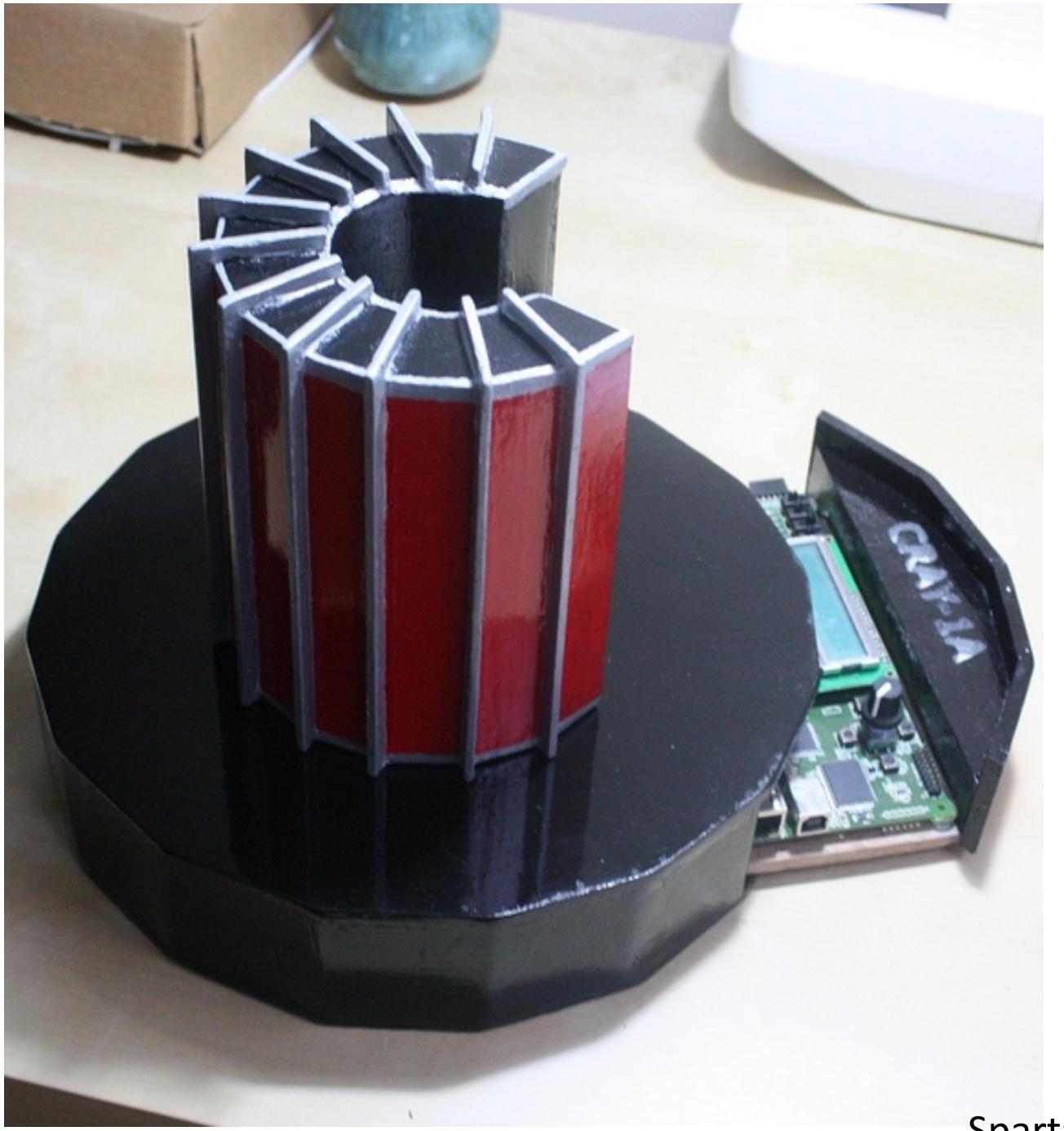
- **Cray Research, by Seymour Cray**
  - **Cray-1 (1976): 1 processor**
  - **Cray-2 (1985): up to 4 processors***
- **Cray Research, not by Seymour Cray**
  - **Cray X-MP (1982): up to 4 procs**
  - **Cray Y-MP (1988): up to 8 procs**
  - **Cray C90: (1991?): up to 16 procs**
  - **Cray T90: (1994): 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**

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

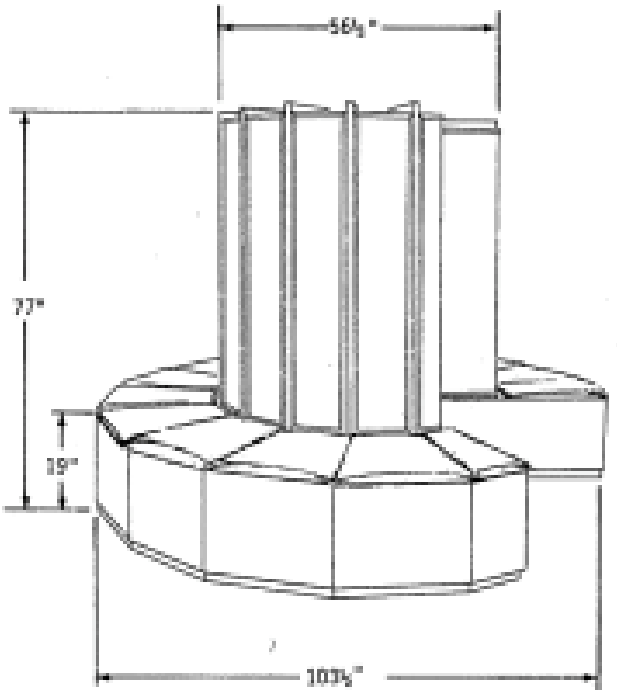


Cray-2

Courtesy of Burton Smith, Microsoft



Four Decades 1976…2016
a million, million
Fenton-Tantos
Desktop Cray 1 or XMP
at 0.1 size.



Spartan-3E-1600-Development-Board

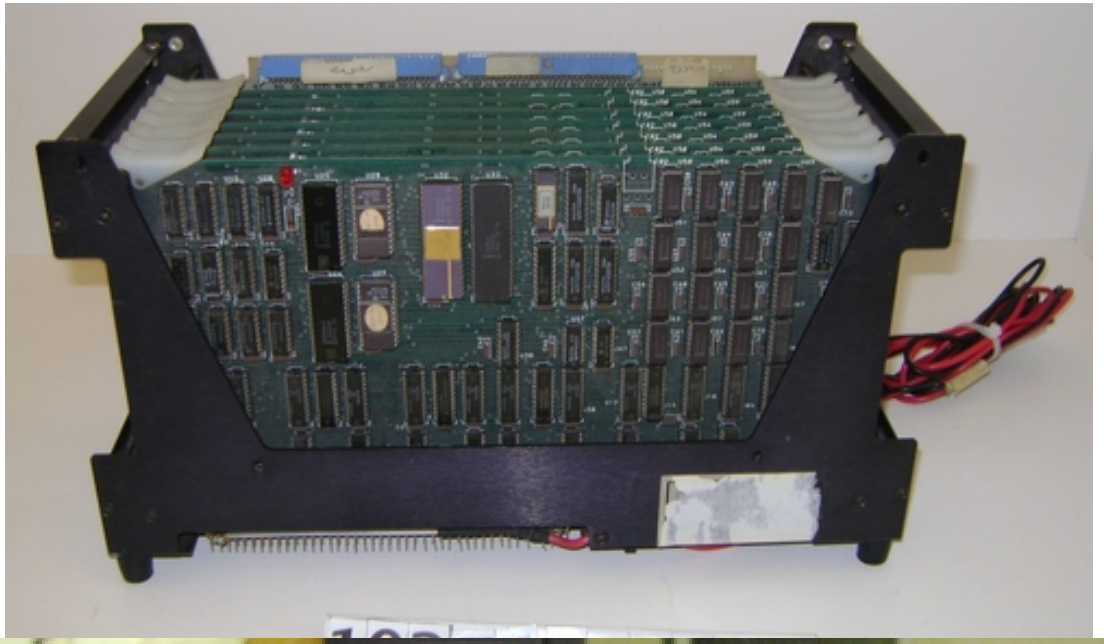
A new beginning: "the killer micros:"
A transition decade

- **1982, 83 Caltech Cosmic Cube** multicomputer with 8, 64 computers.
  DARPA Strategic Computing Initiative and Japanese Fifth Gen.
- **1984 NSF Establishes Supercomputer Centers at Illinois & San Diego**
- **1987** First **Bell Prize** for parallel programing awarded to Sandia **nCUBE** (1K computers) 400-600 speedup,
- **1988 Gustafson's Law** as Amdahl's Law Corollary (Simon #1)
- 1989 Parallel Virtual Machine *distributed memoryprogramming*
- **1992 Intel Touchstone Delta** at Sandia Reaches 100 GF
- **1993 CM5** (60 GF Bell Prize) 1024 Sparc computers.
  Cray C90 was 16! No way/plan to compete with ECL, shared memory
- **1993 Top500** established , using LINPACK Benchmark. (Simon #10)
  Begin multicomputer era
- **1994 Beowulf kit** and recipe for multicomputers and
  MPI-1 Standard released

Capacity Computing … independent parallelism

Capability … parallelism for a single job

1983 Caltech Cosmic Cube
8 node prototype (‘82) & 64 node ‘83
Intel iPSC 64 Personal Supercomputer ‘85



1982-1984 : The Lax Report to NSF's NSB for NSF Advanced Scientific Computing

- **Gresham's Law: VAXen are KILLING supercomputers**
- **NSF needs to fund supercomputer access and centers**
- **1984: NSF Establishes Office of Scientific Computing**
  - **NCSA at U IL (1985)**
  - **SDSC at UCSF**
  - **Cornell National Supercomputer Facilility**
  - **Pittsburgh Supercomputer Center**
  - **John Von Neumann Center at Princeton etc.**
- **1986 CISE (Computer and Information Scinece and Engineering) Directorate**
- **Research focus on parallelism!**

  Knuth , Thompson, Karp disagree

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

Development of Parallel Methods For a 1024-Processor Hypercube

John L. GUSTAFSON, Gary R. MONTRY, and Robert E. BENNER
Sandia National Laboratories, Albuquerque, New Mexico

March 1988

As printed in SIAM Journal on Scientific and Statistical Computing
Vol. 9, No. 4, July 1988, pp. 609–638.
(Minor revisions have been made for the Web page presentation of this paper. JLG 1995)

EDITOR'S NOTE

[This paper] reports on the research that was recognized by two awards, the Gordon Bell Award and the Karp Prize, at IEEE's COMPCON 1988 meeting in San Francisco on March 2.

The Gordon Bell Award recognizes the best contributions to parallel processing, either speedup or throughput, for practical, full-scale problems. Two awards were proposed by Dr. Bell: one for the best speedup on a general-purpose computer and a second for the best speedup on a special-purpose architecture. This year the two awards were restructured into first through fourth place awards because of the nature of the eleven December 1987 submissions. Bell presented the first place award of $1,000 to the authors of [this paper].

Ed Barsis, Director of Computer Sciences and Mathematics at Sandia in 1988 kneels in front of the nCUBE/10.

Gustafson's Law

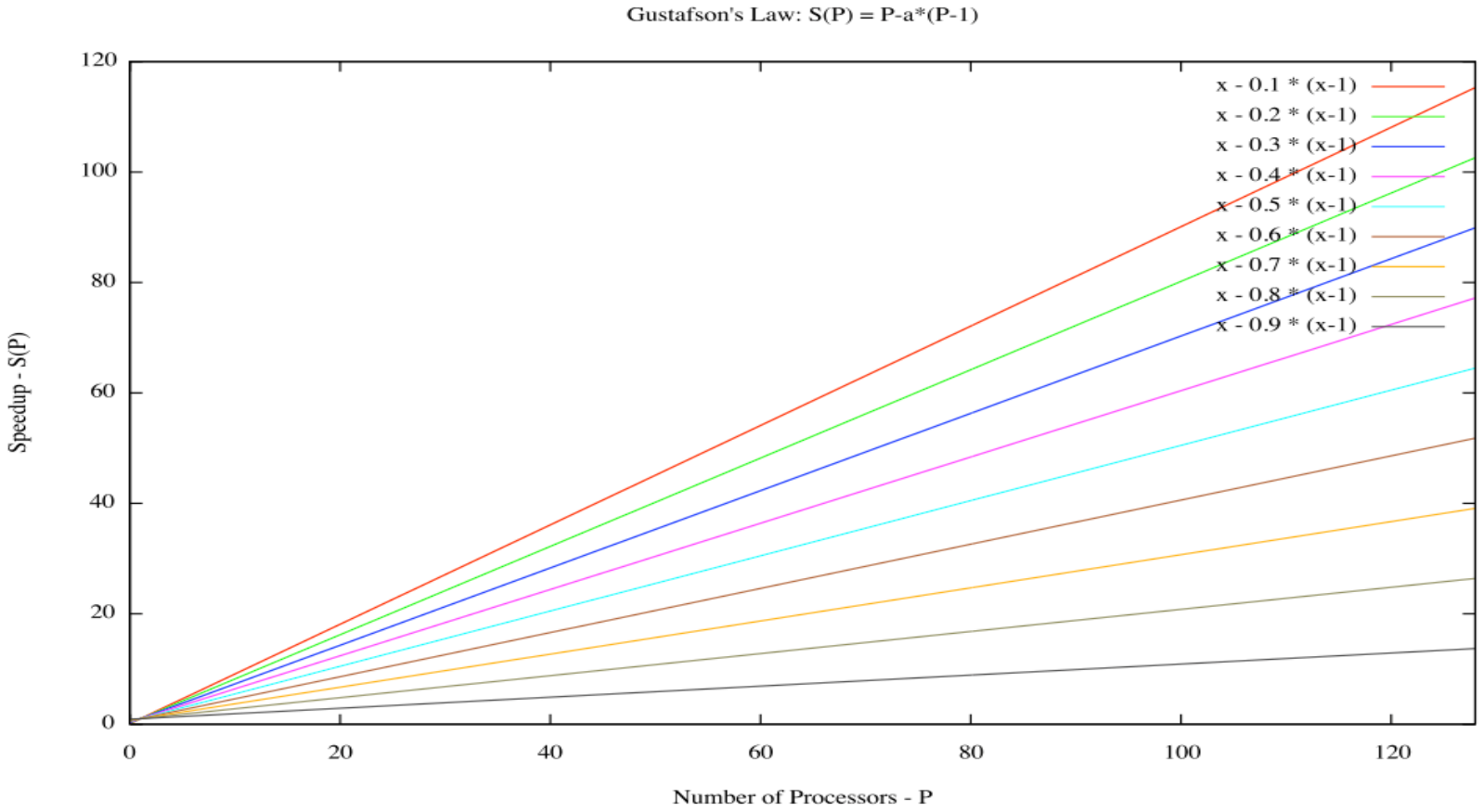
Benner, Gustafson, Montry winners of first Gordon Bell Prize

$S(P) = P - \alpha \times (P-1)$

***P* is the number of processors,**

***S* is the SPEEDUP, and**

α, the non-parallelizable fraction of any parallel process



Top500 #1 @ 60 Gflops: June 1993

Home | TOP500 Supercomputer Sites - Google Chrome

top500.org

Apps | Web Slice Gallery | Windows Live Hotm... | GB stuff | Imported From IE | Google | Home - Windows Live | Suggested Sites | bing | Google (2)

Home | Project | Features | Lists | Statistics | Contact | Search | Go

Powered by Intel® Xeon® processors

CRAY

FUJITSU

Like 3,614 people like this.

FROM DATA TO KNOWLEDGE

Aug. 14, 2014, 5:30 a.m.

Last year we organized the first ISC Big Data Conference and the acceptance of the conference across the high performance and enterprise computing communities has encouraged us to bring you this event once more. This year's theme, From Data to Knowledge, reflects the fundamental value of all big data applications, encapsulating the challenges as well as the solutions.

CLOUD MEETS HPC

Aug. 14, 2014, 5:16 a.m.

We launched the ISC Cloud Conference series five years ago when cloud computing was still an unfamiliar concept to a large portion of high performance computing (HPC) practitioners. As many as 250 members of the HPC community attended our first cloud meeting in 2010. These early adopters continue to support the ISC Cloud Conference with their attendance as well as sponsorships.

TOP500 List Presented at ISC'14

July 1, 2014, 6:01 a.m.

Jack Dongarra, Erich Strohmaier and Michael Resch discussing the current TOP500 list at ISC'14.

Slides for the 43rd TOP500 list now available.

June 24, 2014, 7:54 a.m.

1. **Tianhe-2 (MilkyWay-2)** - TH-IVB-FEP Cluster, Intel Xeon E5-2692 12C 2.200GHz, TH Express-2, Intel Xeon Phi 31S1P
   NUDT
2. **Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x
   Cray Inc.
3. **Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom
   IBM
4. K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect
   Fujitsu
5. **Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom
   IBM
6. **Piz Daint** - Cray XC30, Xeon E5-2670 8C 2.600GHz, Aries interconnect , NVIDIA K20x
   Cray Inc.
7. **Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi SE10P
   Dell

1994: MPI 1.0 Massage Passing Interface

MPI: A Message-Passing Interface Standard
Version 3.0

Message Passing Interface Forum

September 21, 2012

Version 3.0: September 21, 2012. Coincident with the development of MPI-2.2, the MPI Forum began discussions of a major extension to MPI. This document contains the MPI-3 Standard. This draft version of the MPI-3 standard contains significant extensions to MPI functionality, including nonblocking collectives, new one-sided communication operations, and Fortran 2008 bindings. Unlike MPI-2.2, this standard is considered a major update to the MPI standard. As with previous versions, new features have been adopted only when there were compelling needs for the users. Some features, however, may have more than a minor impact on existing MPI implementations.

Version 2.2: September 4, 2009. This document contains mostly corrections and clarifications to the MPI-2.1 document. A few extensions have been added; however all correct MPI-2.1 programs are correct MPI-2.2 programs. New features were adopted only when there were compelling needs for users, open source implementations, and minor impact on existing MPI implementations.

Version 2.1: June 23, 2008. This document combines the previous documents MPI-1.3 (May 30, 2008) and MPI-2.0 (July 18, 1997). Certain parts of MPI-2.0, such as some sections of Chapter 4, Miscellany, and Chapter 7, Extended Collective Operations, have been merged into the Chapters of MPI-1.3. Additional errata and clarifications collected by the MPI Forum are also included in this document.

Version 1.3: May 30, 2008. This document combines the previous documents MPI-1.1 (June 12, 1995) and the MPI-1.2 Chapter in MPI-2 (July 18, 1997). Additional errata collected by the MPI Forum referring to MPI-1.1 and MPI-1.2 are also included in this document.

Version 2.0: July 18, 1997. Beginning after the release of MPI-1.1, the MPI Forum began meeting to consider corrections and extensions. MPI-2 has been focused on process creation and management, one-sided communications, extended collective communications, external interfaces and parallel I/O. A miscellany chapter discusses items that do not fit elsewhere, in particular language interoperability.

Version 1.2: July 18, 1997. The MPI-2 Forum introduced MPI-1.2 as Chapter 3 in the standard "MPI-2: Extensions to the Message-Passing Interface", July 18, 1997. This section contains clarifications and minor corrections to Version 1.1 of the MPI Standard. The only new function in MPI-1.2 is one for identifying to which version of the MPI Standard the implementation conforms. There are small differences between MPI-1 and MPI-1.1. There are very few differences between MPI-1.1 and MPI-1.2, but large differences between MPI-1.2 and MPI-2.

Version 1.1: June, 1995. Beginning in March, 1995, the Message-Passing Interface Forum reconvened to correct errors and make clarifications in the MPI document of May 5, 1994, referred to below as Version 1.0. These discussions resulted in Version 1.1. The changes from Version 1.0 are minor. A version of this document with all changes marked is available.

Version 1.0: May, 1994. The Message-Passing Interface Forum (MPIF), with participation from over 40 organizations, has been meeting since January 1993 to discuss and define a set

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

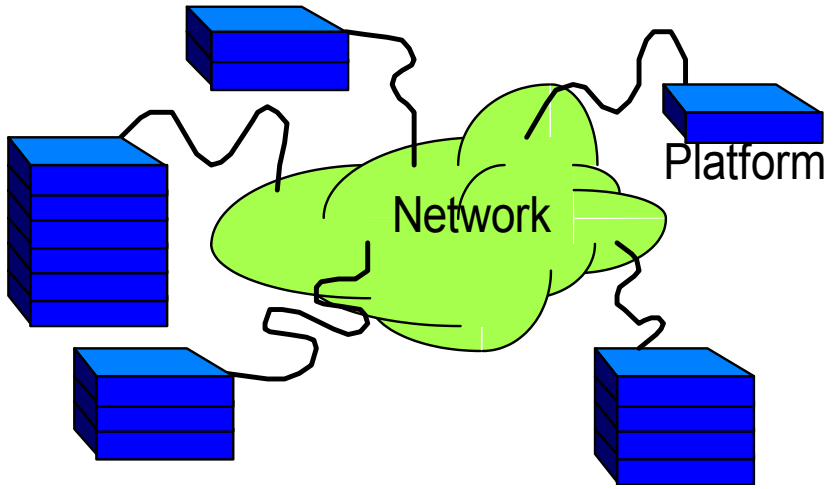
BSD, LINUX, Solaris, and Windows Support for MPI and PVM

Lives Lost: The search for parallelism c1983-1997 DOE and DARPA Strategic Computing 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*

1994: Computers will All be Scalable

- 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



1994 Meeting with Jim Gray

"the day I gave up on shared memory computers"

The Multicomputer aka Clusters Era

- **1993 CM5 , 1024 computer cluster . Top500**
- **1995 ASCI >** Advanced Simulation and Computing (ASC) Program
- **1996 Seymour R. Cray** is killed in a car accident…. Was uilding a shared memory computer using itanium
- **1997 ASCI Red (1 TF) at Sandia, 9K computers**
- **2008 IBM BlueGene (1.5 PF)**
- **2012 Cray Titan** (17.6 PF) **GPU and CUDA**
- **Tiahne-2 at NUDT, 2016 Sunlight achieves 93 PF with > 10M core**
- **ORNL Summit Top500 2018;**
  148.5 PF, 2.4 Mcores, 10 Mwatts, 3 Ghz

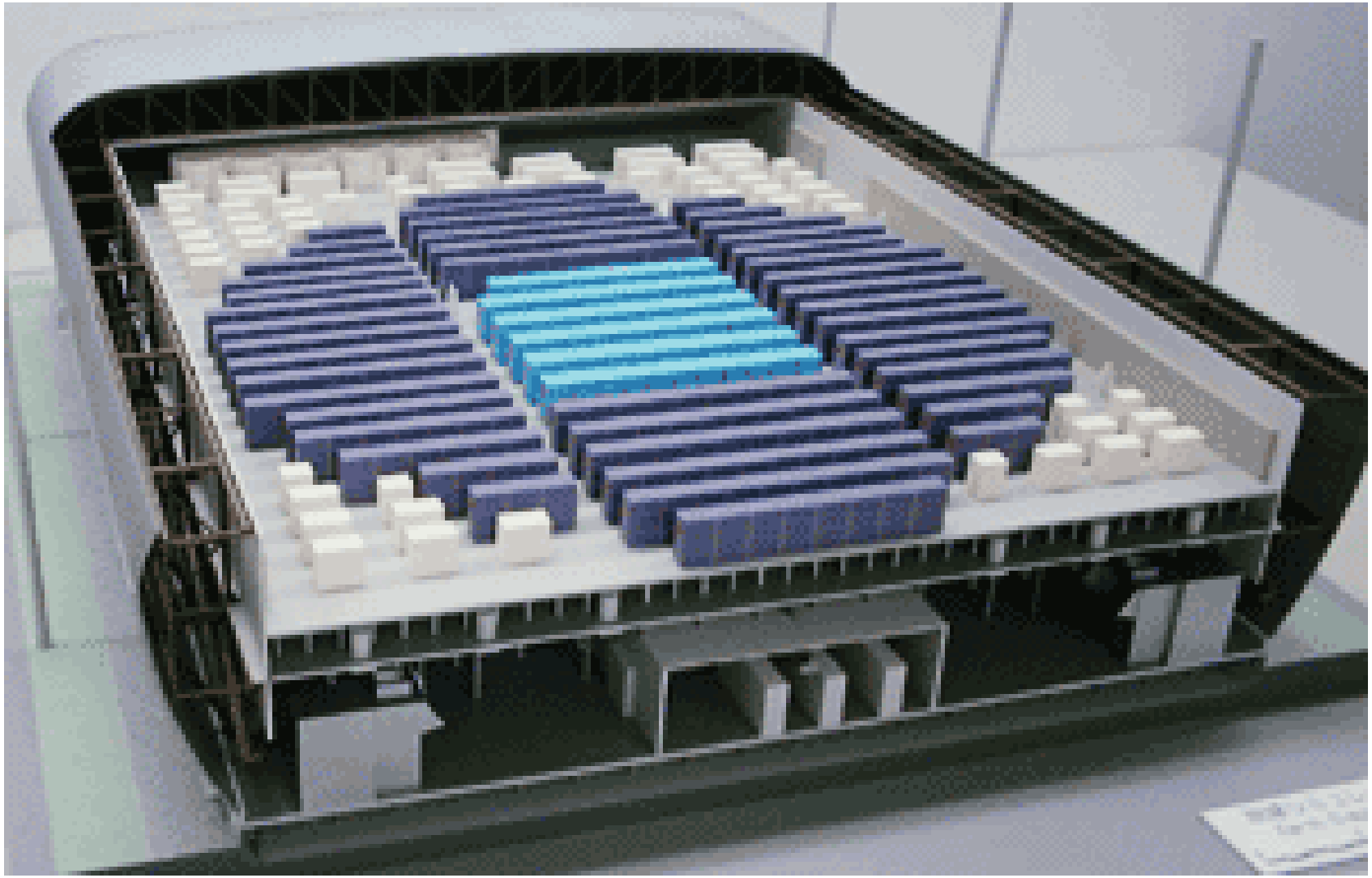
- **2018 $500 M commitment to deliver one+ exaflops** to ANL

ASCI Red 1997-2005 at Sandia National Lab

Figure 4-1. The ASCI Red system at Sandia.

- June 1997-2000
- 1.3-2.5 Tflops
- 9,216-9632 proc
  640 disks,
  1,540 PS,
  616 interconnect

Japanese Earth Simulator (NEC)
2002-2004 35 Teraflops 5,000 vector processor computers
…Stimulant for ASCI



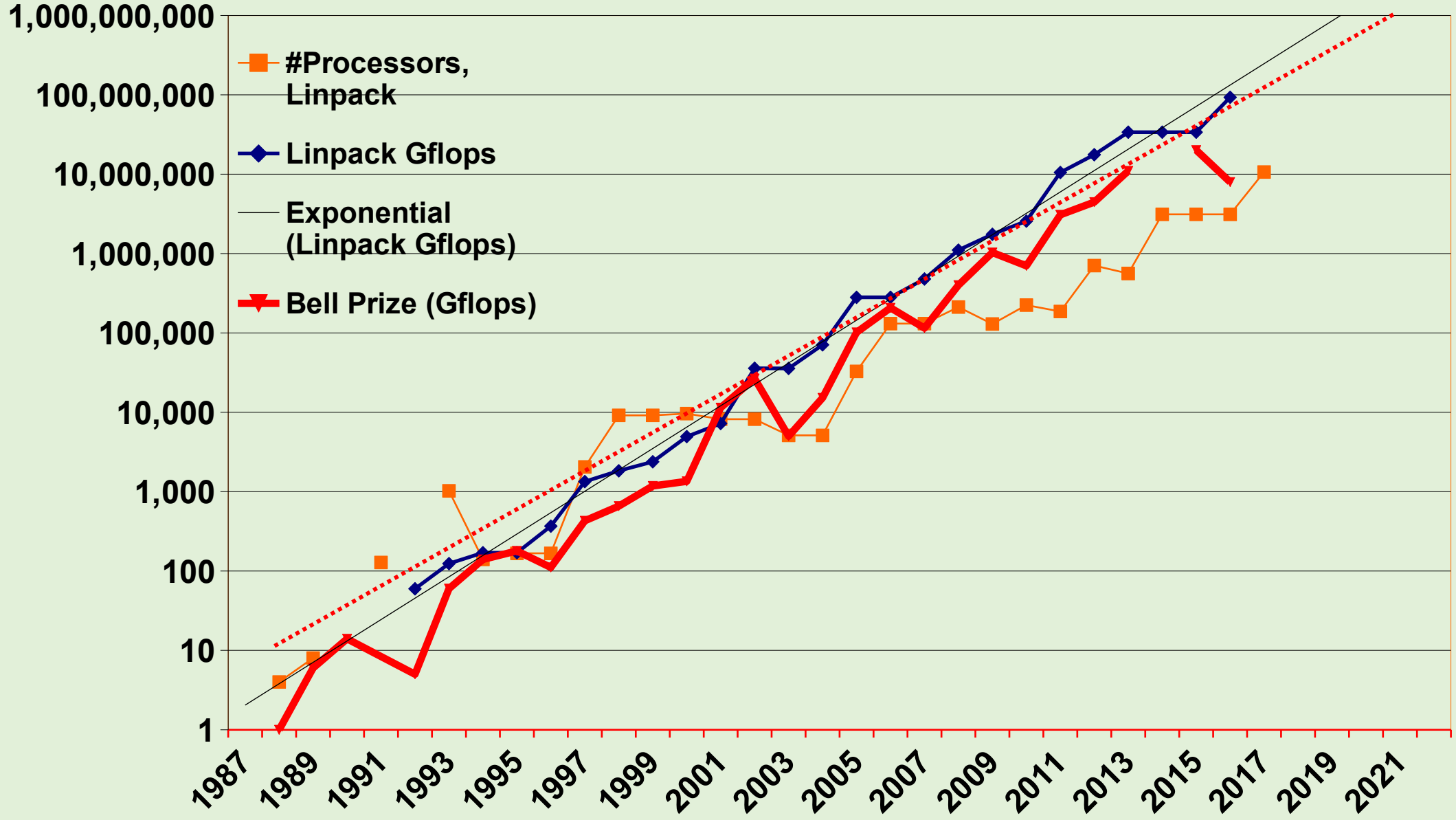
LLNL Sequoia 17 Petaflops…
Tops500 #1 June 2012

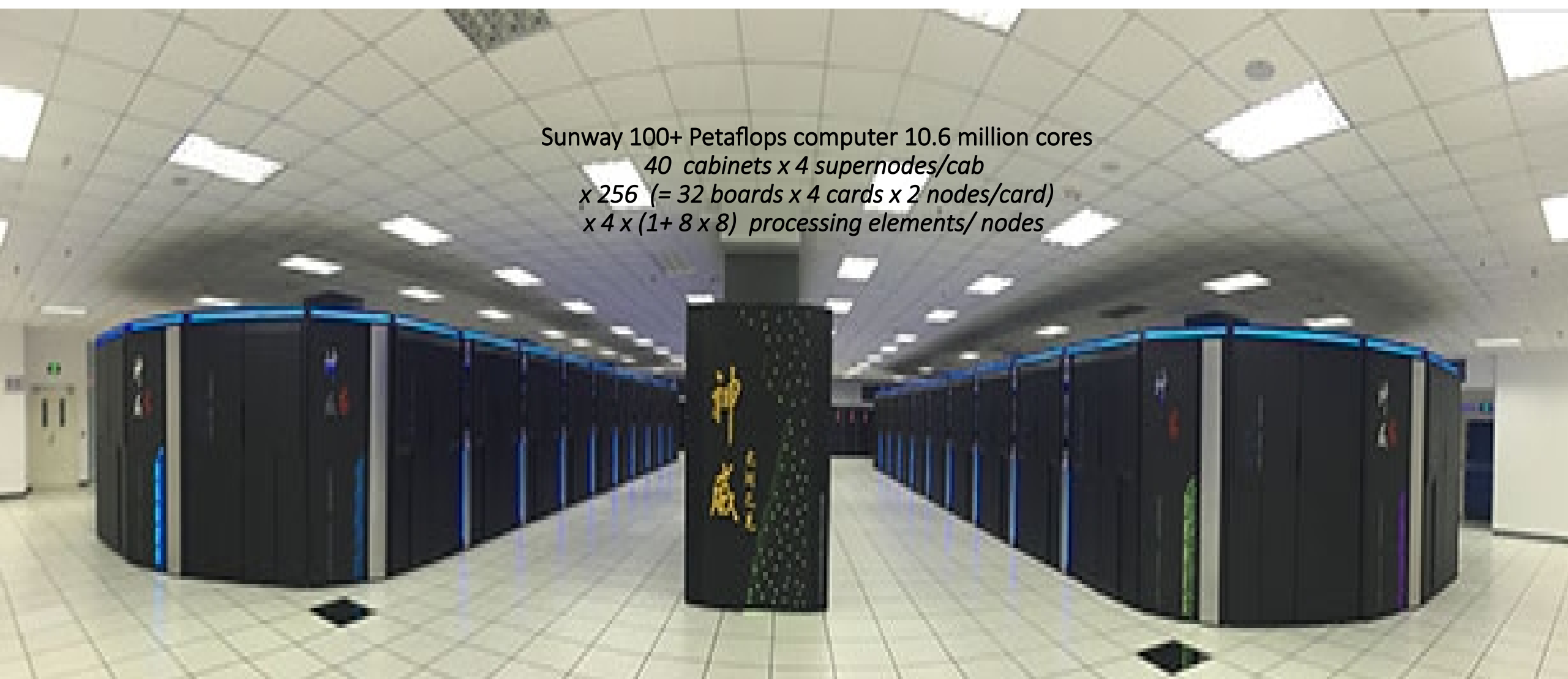
4 threads/core
16 cores/chip

1024 chips/rack
96 racks per system
1.57 million processors

1 GigaBytes/processor
1.6 PetaBytes primary memory

80 KWatt/rack
7.7 MWatts

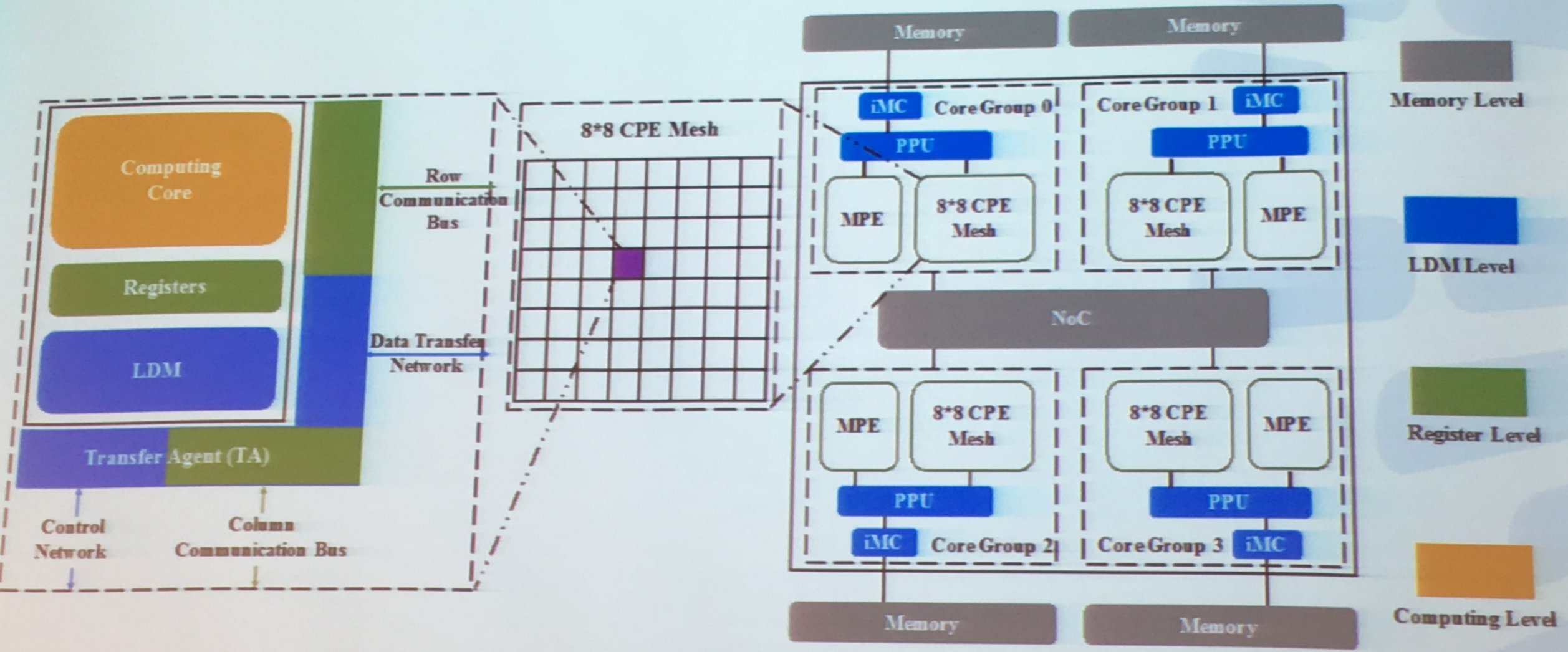




Sunway 100+ Petaflops computer 10.6 million cores
40 cabinets x 4 supernodes/cab
x 256 (= 32 boards x 4 cards x 2 nodes/card)
x 4 x (1+ 8 x 8) processing elements/ nodes

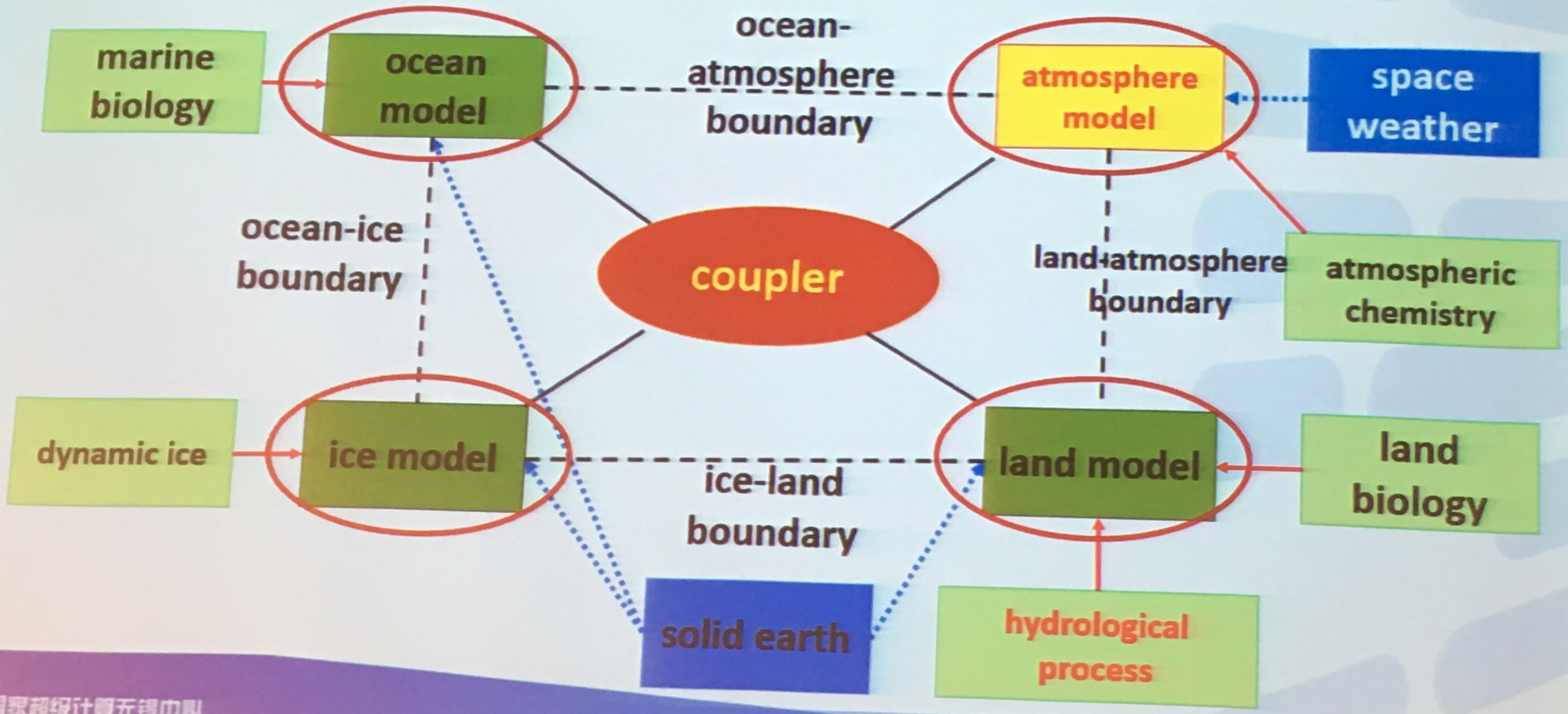
Node peak performance of (256 cores * 8 flops/cycle * 1.45 GHz) +
(4 core * 16 flops/cycle * 1.45 GHz) = 3.0624 Tflop/s per node
Node memory 4 x 8 Gbytes. 32 x 109 / 3 x 10**12 flops /sec**
= .01 Bytes/Operation

SW26010: Sunway 260-Core Processor



CPUs, GPUs, … all those registers

More and more component models



Highly-Scalable Atmospheric Simulation Framework

2016 Bell Prize: Climate Modeling

Yang, Chao
Institute of Software, CAS
computational mathematics

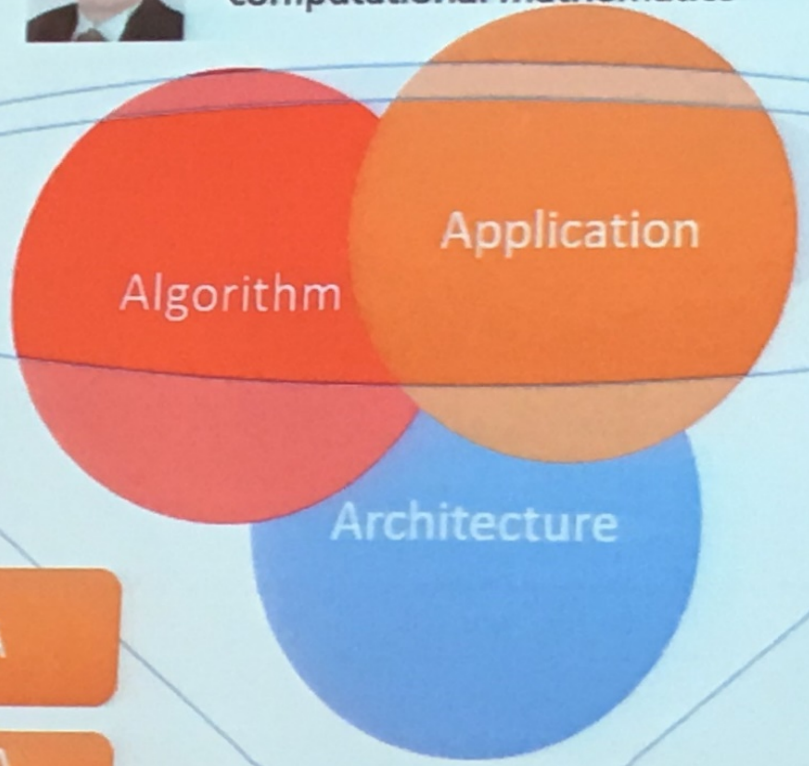
cube-sphere grid or other grid

explicit, implicit, or semi-implicit method

cloud resolving

Wang, Lanning
Beijing Normal University
climate modeling

Xue, Wei
Tsinghua University
computer science



Sunway, GPU, MIC, FPGA

C/C++, Fortran, MPI, CUDA, Java, ...

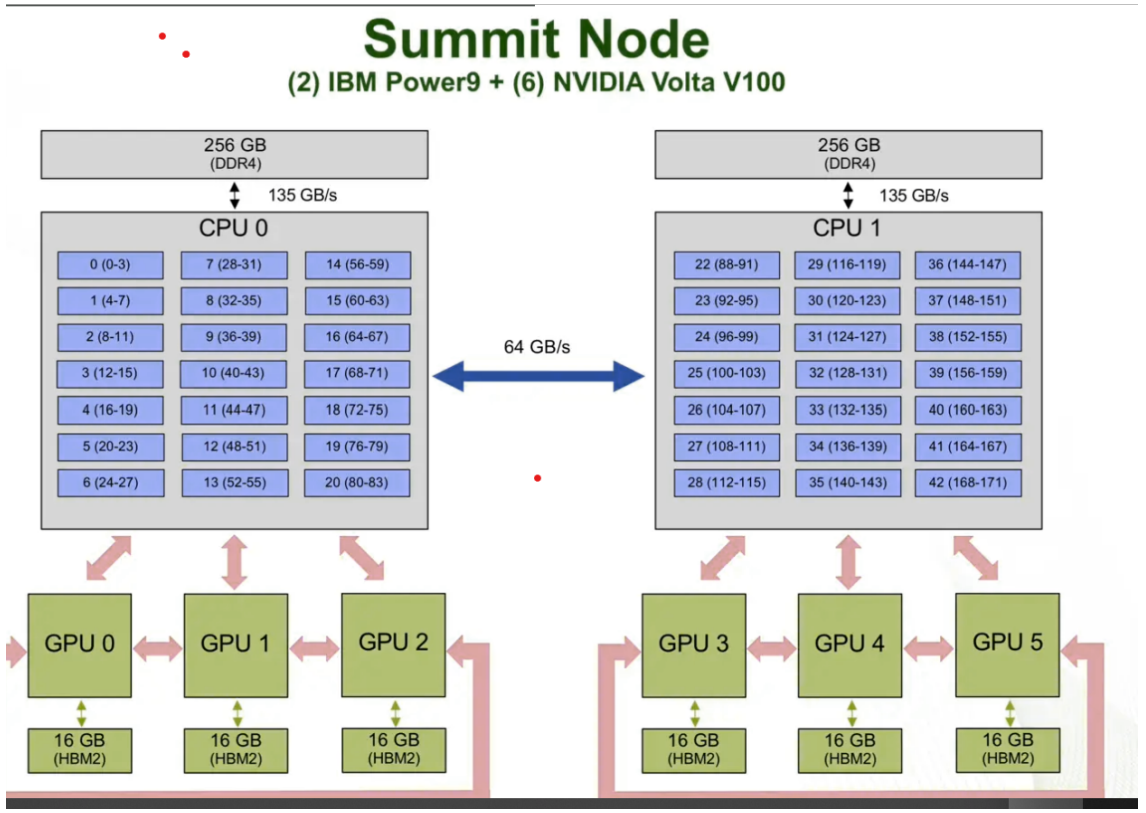
Fu, Haohuan
Tsinghua University
geo-computing

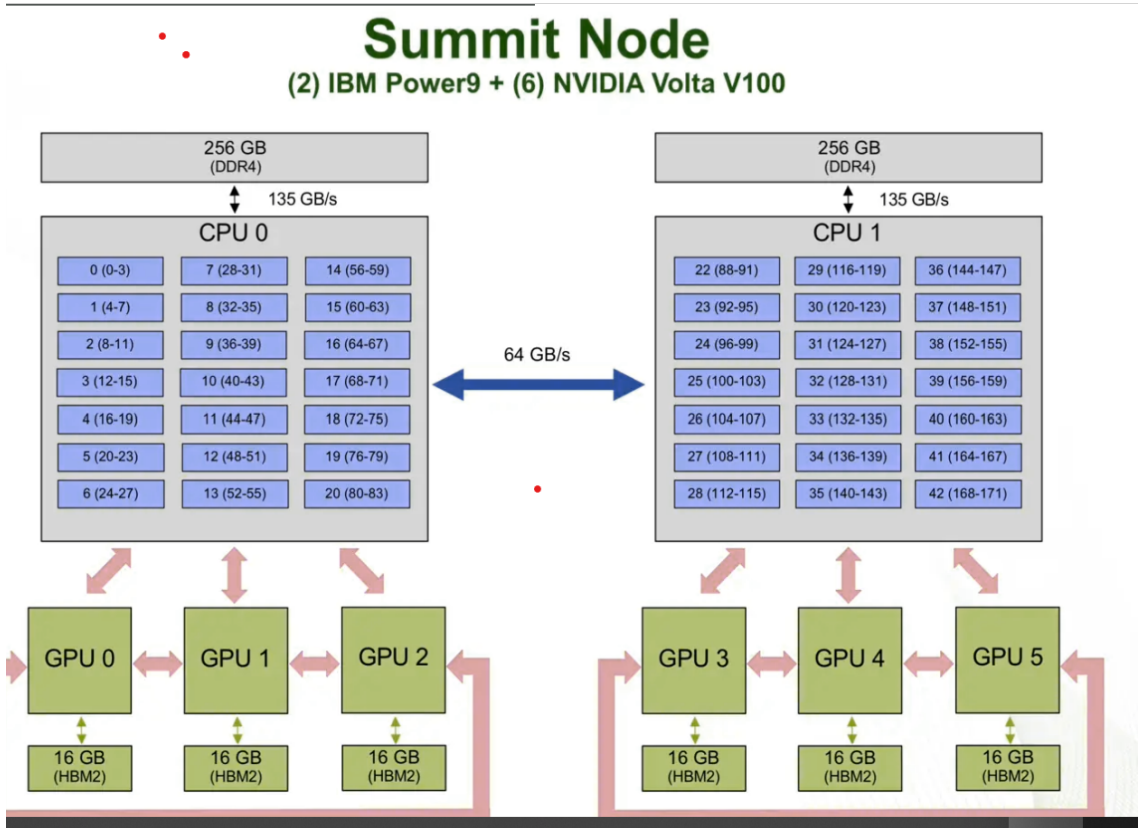
The "Best" Computational Solution

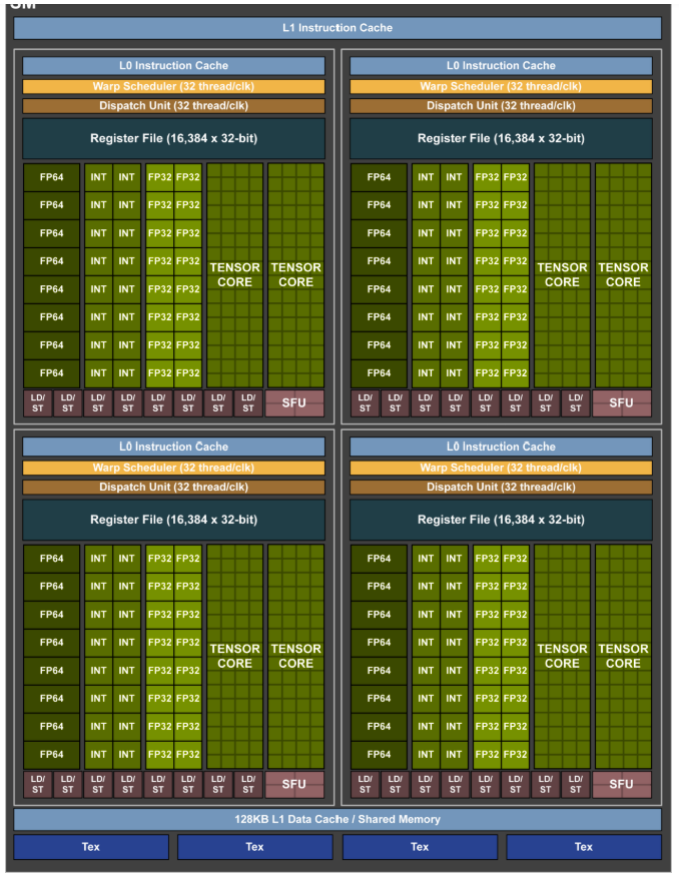
Yifeng Cui, SDSC
team member

ORNL Summit Top500 2018 #1
148.5 PF, 201 peak, 2.4 Mcores, 10 Mwatts, 3 Ghz

- **IBM Power System AC922 node**
- **4,600 compute nodes**
  - **22 GB/s non blocking links**
  - **two IBM POWER9 processors and**
    - **Four threads**
    - **2 SIMD Multi-Core (SMC)**
  - **512 GB of DDR4 memory**
  - **six NVIDIA Volta V100 accelerators**
    - **80 streaming multiprocessors (SMs)**
    - **32 FP64 (double-precision) cores, 64 FP32 (single-precision) cores, 64 INT32 cores, and 8 tensor cores.**
  - **96 GB for accelerators**
  - **1.6TB of non-volatile memory**







Parallelism

Compute o(1000s)

- Nodes (many jobs, many instances of a job (ensemble computing)
- Internode communication or communicating sequential processes
  - Multiple sockets (chips) per node 1…8 may share memory
    - Multiple cores 2-44
      - Multiple threads 1-2
        - » Instruction level parallelism pipeling into GPU ALUs o(50)

Parallel levels by grain size

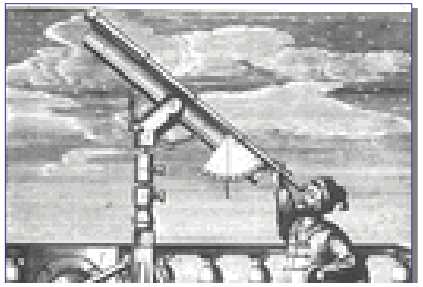
- Job stream parallelism aka ensembles, capacity computing
- Communicating, sequential processes with Multiple threads
- Shared Memory
- GPU, CUDA, Accelerated computing

Four Science Paradigms, Jim Gray, Jan. 2007

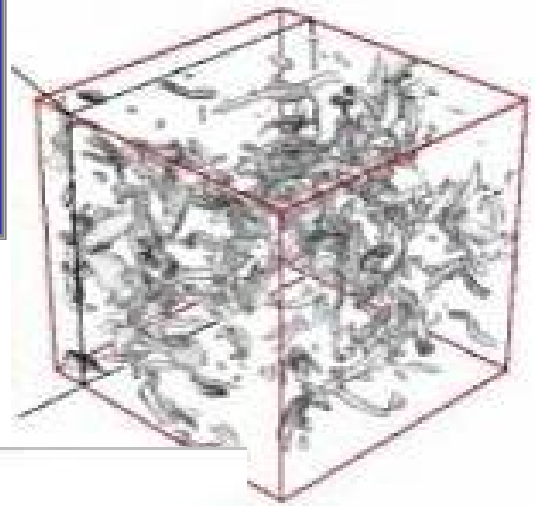
1. Thousand years ago:
   science was **empirical**
   describing natural phenomena
2. Last few hundred years:
   theoretical branch
   using models, generalizations
3. 1987, since ENIAC & FORTRAN:
   a **computational** branch
   simulating complex phenomena
4. 2010 **Data-intensive science** :
   data exploration (eScience)

unify theory, experiment, and simulation

- **Data captured by instruments**
  Or generated by simulation
- Processed by software
- Information/Knowledge stored in computer
- Scientist analyzes database / files
  using data management and statistics



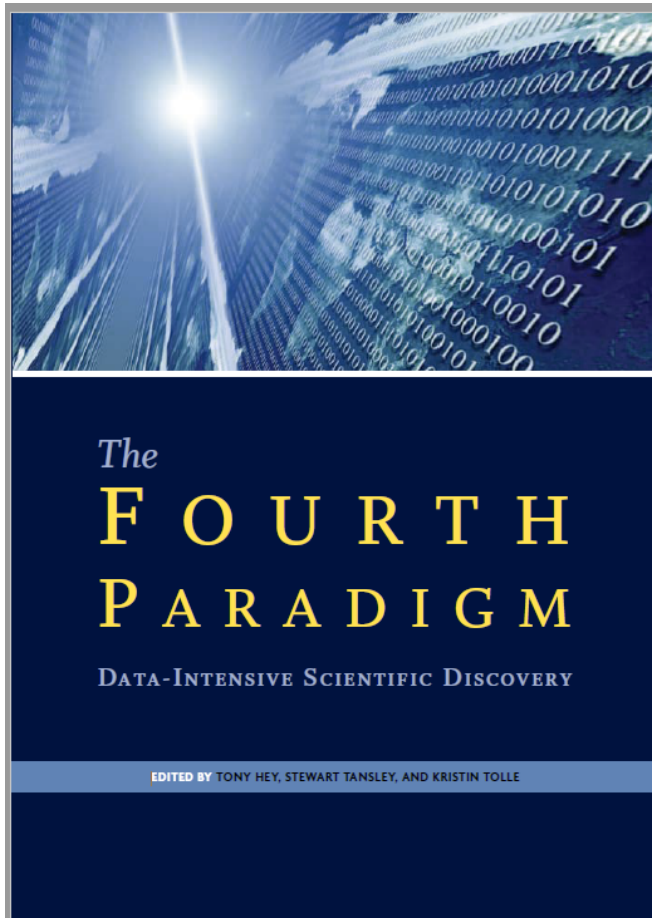
$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi G\rho}{3} - \mathrm{K}\frac{c^2}{a^2}$$



Third and Fourth Paradigms of science

- **1987 Ken Wilson, Nobel Prize Winner declares: "Compuation is 3rd Paradigm"**
- **2008 DOE Dir. Of Science discovers 3rd Paradigm**
- **Nov 2010 "The Big Idea: The Next Scientific Revolution - Harvard Business Review"**

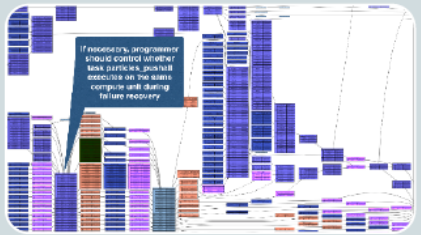
2007 Jim Gray @NRC CSTRB : Data Science is 4th Paradigm

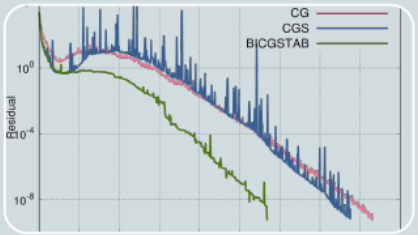


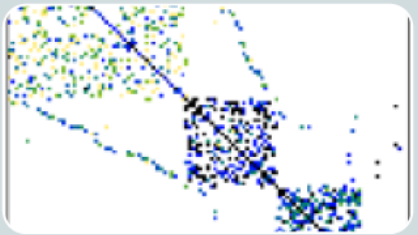
More paradigms? What's Next?

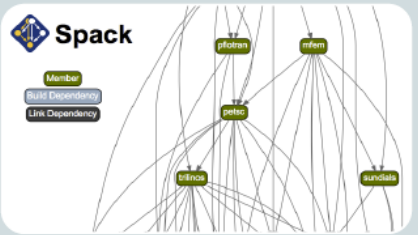
- Visualization described in 1987, Re-discovered 2000, and Re-re-discover (Every 20 years something is rediscovered)
- Very large, coupled, complete, & complex models e.g. climae simulation
- Data Science recognized in 2010 to manage data
- Data Scientist has become a profession
- AI for Machine Learning is next big thing for HPC

ECP investments in software technologies help ensure the exascale computers will be a success









Programming Models & Runtimes

- Enhance and get ready for exascale the widely used MPI and OpenMP programming models (hybrid programming models, deep memory copies)
- Development of performance portability tools (e.g. Kokkos and Raja)
- Support alternate models for potential benefits and risk mitigation: PGAS (UPC++/GASNet) ,task-based models (Legion, PaRSEC)
- Libraries for deep memory hierarchy and

Development Tools

- Continued, multifaceted capabilities in portable, open-source LLVM compiler ecosystem to support expected ECP architectures, including support for F18
- Performance analysis tools that accommodate new architectures, programming models, e.g., PAPI, Tau

Math Libraries

- Linear algebra, iterative linear solvers, direct linear solvers, integrators and nonlinear solvers, optimization, FFTs, etc
- Performance on new node architectures; extreme strong scalability
- Advanced algorithms for multi-physics, multiscale simulation and outer-loop analysis
- Increasing quality, interoperability, complementarity of math libraries

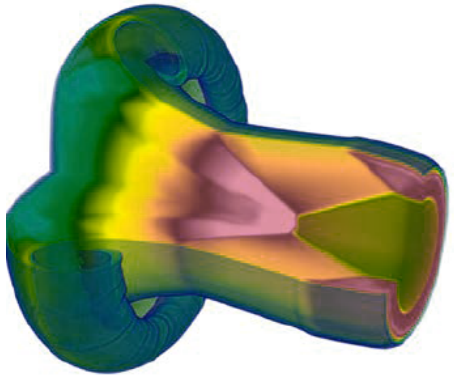
Data and Visualization

- I/O via the HDF5 API
- Insightful, memory-efficient in-situ visualization and analysis – Data reduction via scientific data compression
- Checkpoint restart

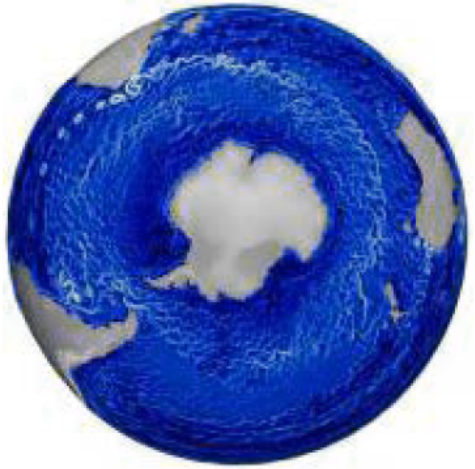
Software Ecosystem

- Develop features in Spack necessary to support all ST products in E4S, and the AD projects that adopt it
- Development of Spack stacks for reproducible turnkey deployment of large collections of software
- Optimization and interoperability of containers on HPC systems
- Regular E4S releases of the ST software stack and SDKs with regular integration of new ST products

Exascale Apps

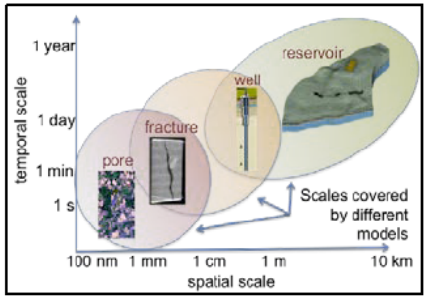


Compressible flow (MARBL)

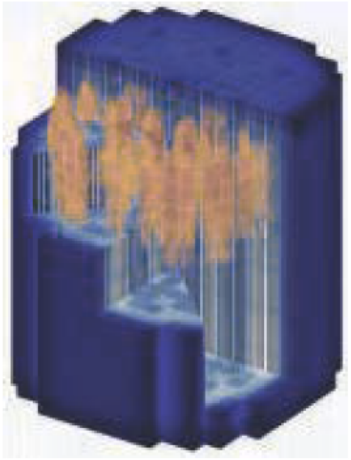


Climate (E3SM)

Urban systems (Urban)



Subsurface (GEOS)

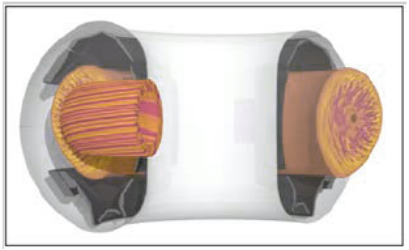


Modular Nuclear Reactors (ExaSMR)

Wind Energy (ExaWind)

Additive Manufacturing (ExaAM)

Combustion (Nek5000)



Magnetic Fusion (WDMApp)

Exascale program goals

The software must be:

- interoperable
- sustainable
- maintainable
- adaptable
- portable
- scalable
- deployed at DOE computing facilities

Work

- Easy o use
- Understanable
- Perform well
- Outperform anything out there
- Competititve

Architectures for Apps, “MSFT Consulting Eng”

- Ending of a 50 year exponential obviously has huge ramifications
- We will see a Cambrian explosion of new hardware in the cloud
- This heterogeneity will be disruptive in many respects
- We can back into it with massive amounts of programmable hardware
- We need ways of applying “architecture” that are not manual or ad hoc
- Innovation will continue but will be more surprising / less predictable
- We should celebrate this world-changing era but keep going!

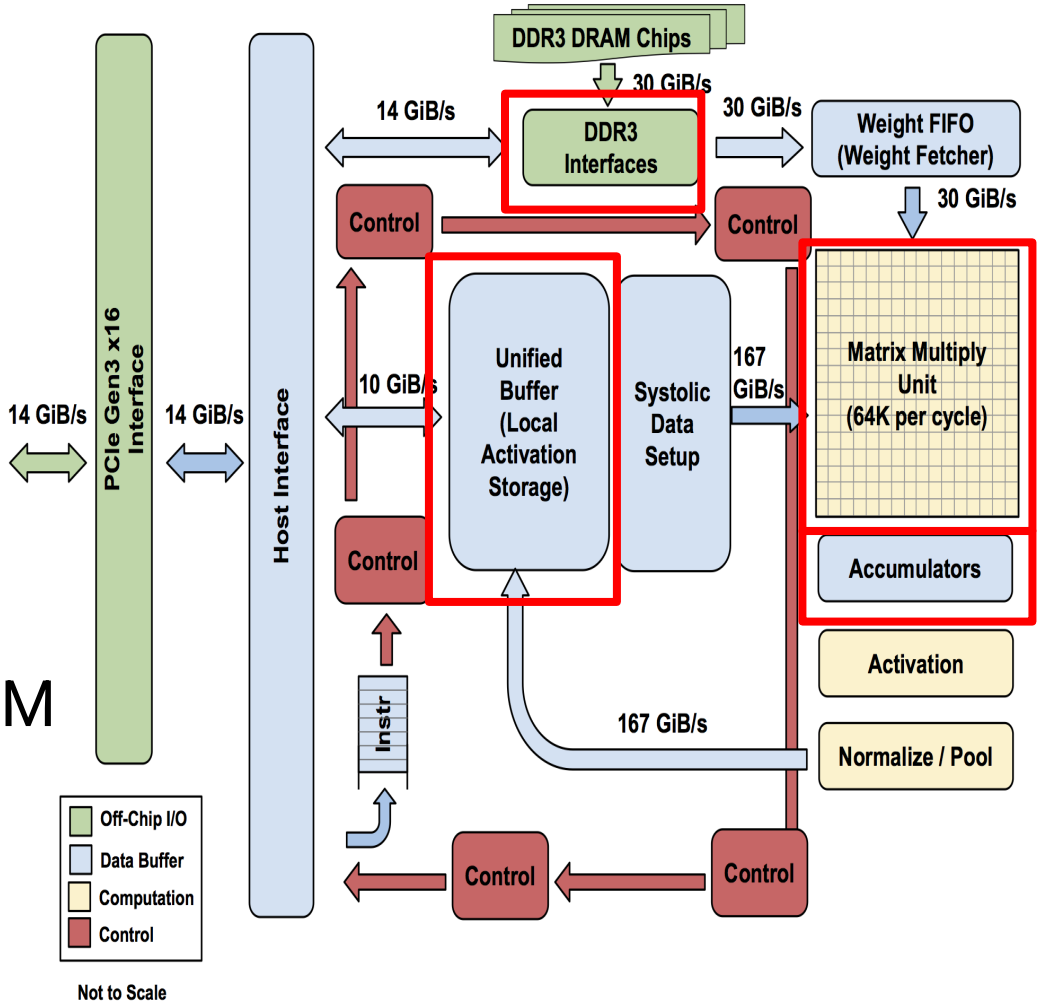
Doug Burger, Microsoft

TPU: High-level Chip Architecture

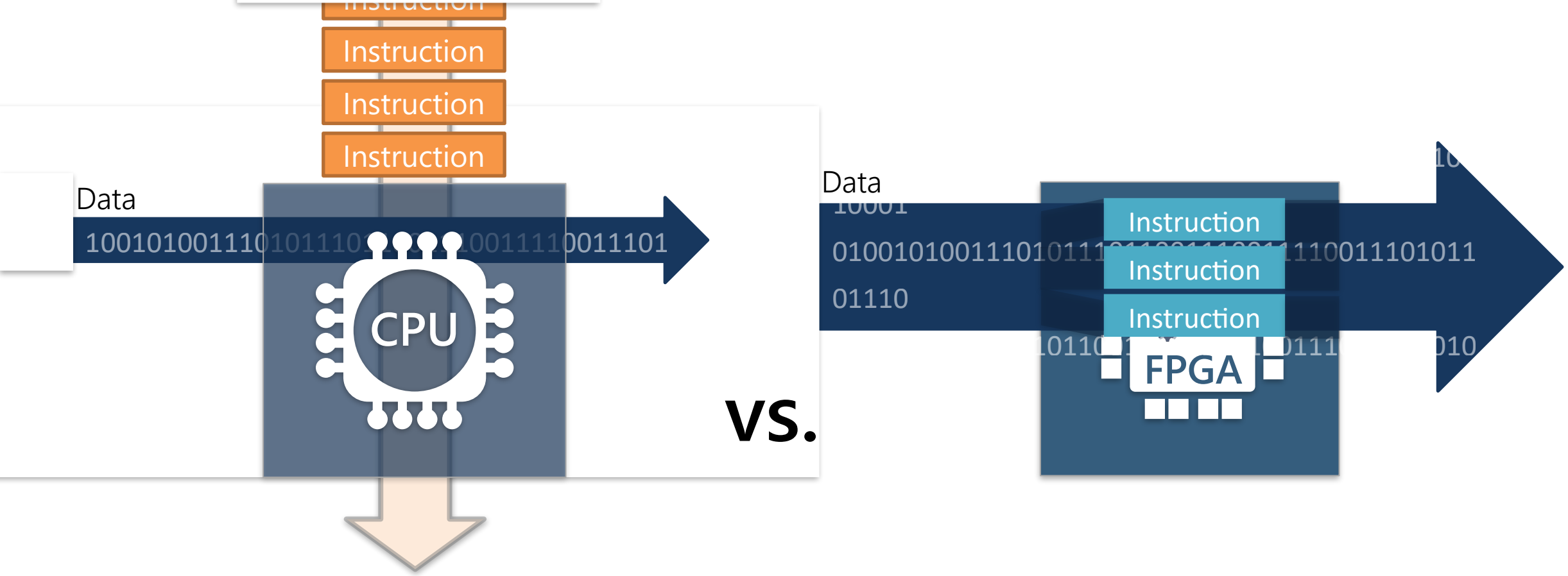
- The Matrix Unit: 65,536 (256x256) 8-bit multiply-accumulate units
- 700 MHz clock rate
- Peak: 92T operations/second
  - 65,536 * 2 * 700M
- >25X as many MACs vs GPU
- >100X as many MACs vs CPU

- 4 MiB of on-chip Accumulator memory
- 24 MiB of on-chip Unified Buffer (activation memory)
- 3.5X as much on-chip memory vs GPU

- Two 2133MHz DDR3 DRAM channels
- 8 GiB of off-chip weight DRAM memory



Temporal versus Spatial Computing



CPU: temporal compute

FPGA: spatial compute

Doug Burger, Microsoft

The end