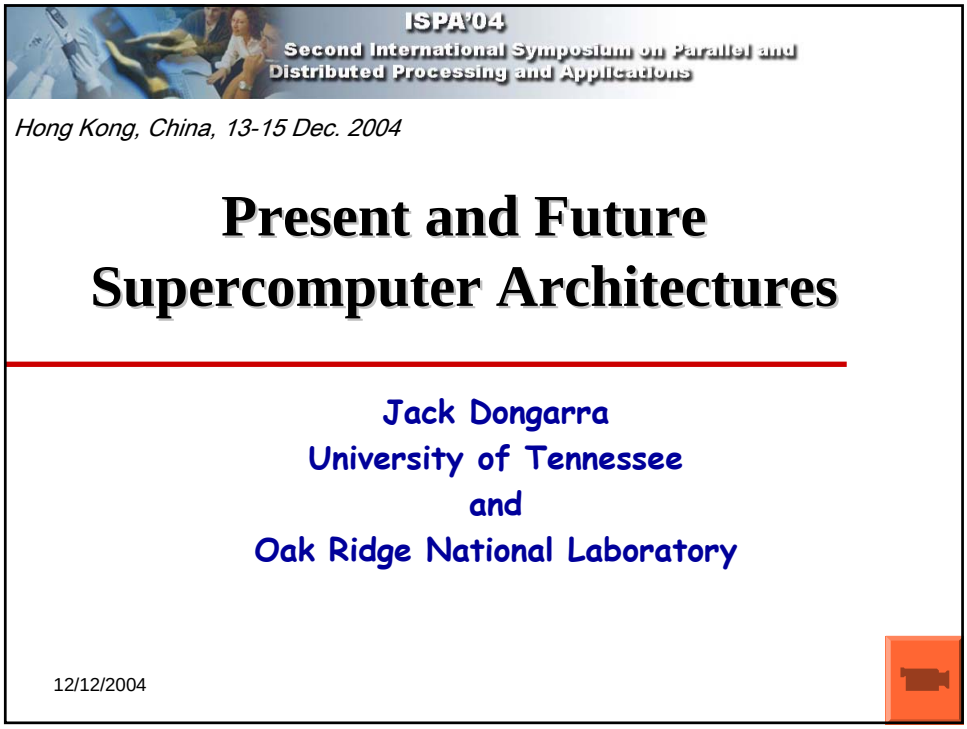
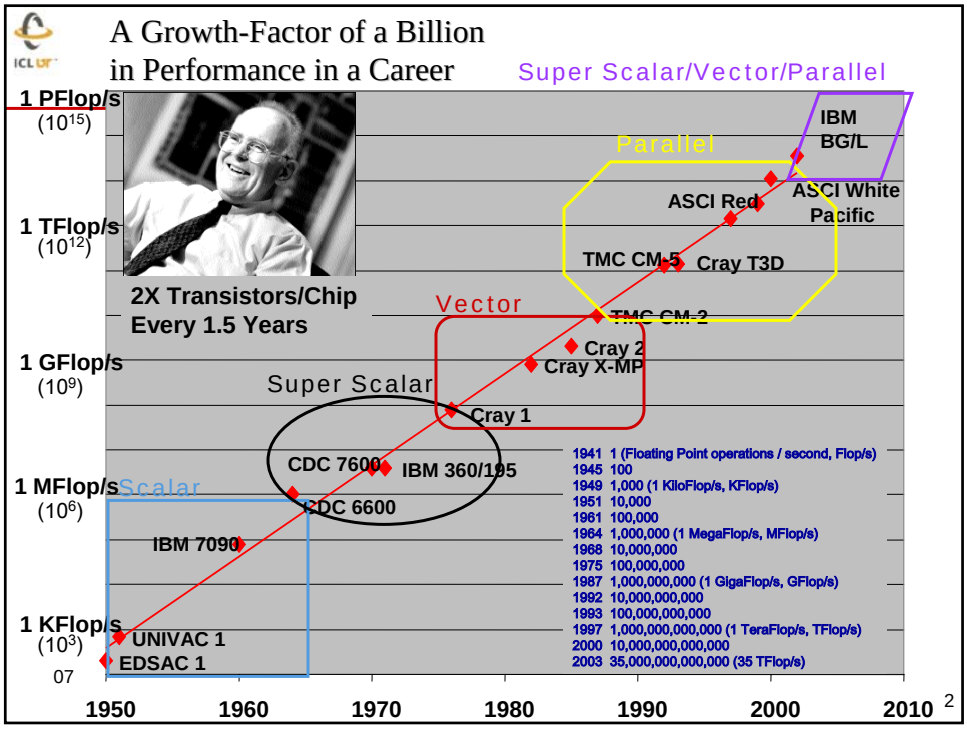

ISPA'04
Second International Symposium on Parallel and Distributed Processing and Applications
Hong Kong, China, 13-15 Dec. 2004
Present and Future Supercomputer Architectures
Jack Dongarra
University of Tennessee
and
Oak Ridge National Laboratory
12/12/2004

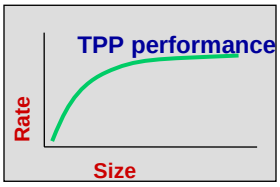

A Growth-Factor of a Billion in Performance in a Career
Super Scalar/Vector/Parallel
1 PFlop/s (10^15)
1 TFlop/s (10^12)
1 GFlop/s (10^9)
1 MFlop/s (10^6)
1 KFlop/s (10^3)
2X Transistors/Chip Every 1.5 Years
Parallel
Vector
Super Scalar
Scalar
IBM BG/L
ASCI Red
ASCI White Pacific
TMC CM-5
Cray T3D
TMC CM-2
Cray 2
Cray X-MP
Cray 1
CDC 7600
IBM 360/195
CDC 6600
IBM 7090
UNIVAC 1
EDSAC 1
1941 1 (Floating Point operations / second, Flop/s)
1945 100
1949 1,000 (1 KiloFlop/s, KFlop/s)
1951 10,000
1961 100,000
1964 1,000,000 (1 MegaFlop/s, MFlop/s)
1968 10,000,000
1975 100,000,000
1987 1,000,000,000 (1 GigaFlop/s, GFlop/s)
1992 10,000,000,000
1993 100,000,000,000
1997 1,000,000,000,000 (1 TeraFlop/s, TFlop/s)
2000 10,000,000,000,000
2003 35,000,000,000,000 (35 TFlop/s)
1950
1960
1970
1980
1990
2000
2010
07
2

TOP500 supercomputer

H. Meuer, H. Simon, E. Strohmaier, & JD

- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

  $Ax=b$, *dense problem*



- Updated twice a year
  SC'xy in the States in November
  Meeting in Mannheim, Germany in June
- All data available from **www.top500.org**

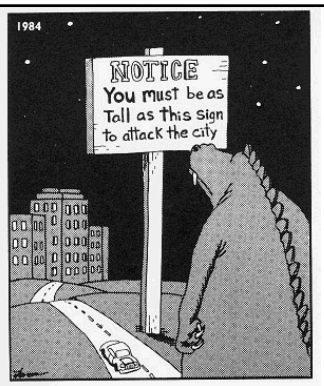
07

3

What is a Supercomputer?

- A supercomputer is a hardware and software system that provides close to the maximum performance that can currently be achieved.
- Over the last 10 years the range for the Top500 has increased greater than Moore's Law
- 1993:
  - #1 = 59.7 GFlop/s
  - #500 = 422 MFlop/s
- 2004:
  - #1 = 70 TFlop/s
  - #500 = 850 GFlop/s

07

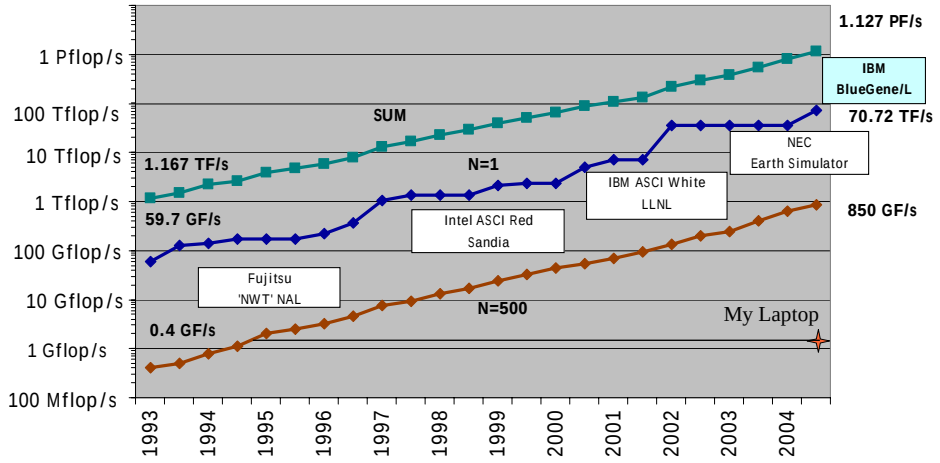


Why do we need them?
Almost all of the technical areas that are important to the well-being of humanity use supercomputing in fundamental and essential ways.

Computational fluid dynamics, protein folding, climate modeling, national security, in particular for cryptanalysis and for simulating nuclear weapons to name a few.

4

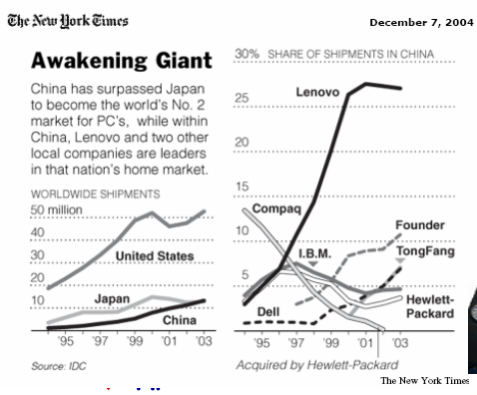
TOP500 Performance – November 2004



07

5

erformance



- Coming soon …
  - Cray RedStorm
  - Cray BlackWidow

- Dawning
- Bull NovaScale
- Lanovo
- Fujitsu PrimePower
- Hitachi SR11000
- NEC SX-7
- Apple

Steve Chen, Founder and CEO of Galactic Computing Ltd.

Beyond Teraflops: A Chinese Supercomputing Blade System to power China Information Services Grids and Accelerate China Information Services Economics

- Galactic Computing is now building its third generation Supercomputing Blade Systems in China to target at broad and deep, high-performance and high-productivity, scientific, engineering and commercial applications.
- 100 TFlop/s

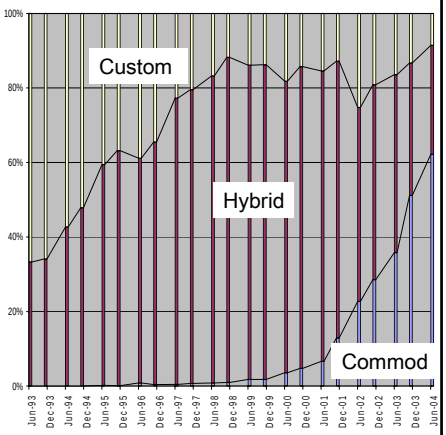
07

Architecture/Systems Continuum

Tightly Coupled

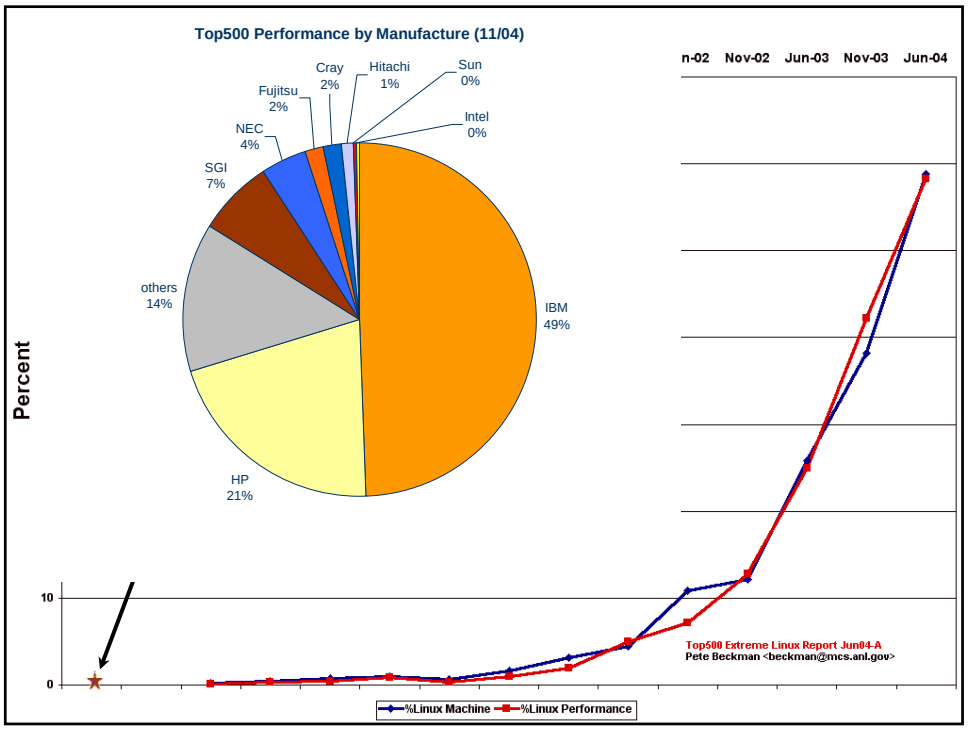
- Custom processor with custom interconnect
  - Cray X1
  - NEC SX-7
  - IBM Regatta
  - IBM Blue Gene/L
- Commodity processor with custom interconnect
  - SGI Altix
    - Intel Itanium 2
  - Cray Red Storm
    - AMD Opteron
- Commodity processor with commodity interconnect
  - Clusters
    - Pentium, Itanium, Opteron, Alpha
    - GigE, Infiniband, Myrinet, Quadrics
  - NEC TX7
  - IBM eServer
  - Dawning

Loosely Coupled



07

7

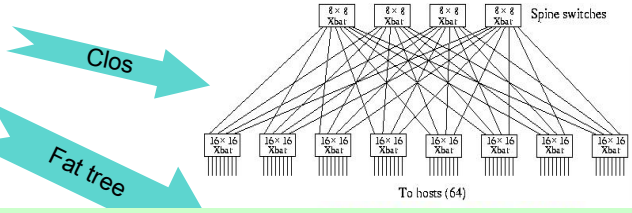


Commodity Processors

- Intel Pentium Nocona
  - 3.6 GHz, peak = 7.2 Gflop/s
  - Linpack 100 = 1.8 Gflop/s
  - Linpack 1000 = 3.1 Gflop/s
- AMD Opteron
  - 2.2 GHz, peak = 4.4 Gflop/s
  - Linpack 100 = 1.3 Gflop/s
  - Linpack 1000 = 3.1 Gflop/s
- Intel Itanium 2
  - 1.5 GHz, peak = 6 Gflop/s
  - Linpack 100 = 1.7 Gflop/s
  - Linpack 1000 = 5.4 Gflop/s
- HP PA RISC
- Sun UltraSPARC IV
- HP Alpha EV68
  - 1.25 GHz, 2.5 Gflop/s peak
- MIPS R16000

Commodity Interconnects

- Gig Ethernet
- Myrinet
- Infiniband
- QsNet
- SCI



	Switch topology	Cost NIC	Cost Sw/node	Cost Node	MPI Lat / 1-way / Bi-Dir (us) / MB/s / MB/s
Gigabit Ethernet	Bus	$ 50	$ 50	$ 100	30 / 100 / 150
SCI	Torus	$1,600	$ 0	$1,600	5 / 300 / 400
QsNetII (R)	Fat Tree	$1,200	$1,700	$2,900	3 / 880 / 900
QsNetII (E)	Fat Tree	$1,000	$ 700	$1,700	3 / 880 / 900
Myrinet (D card)	Clos	$ 595	$ 400	$ 995	6.5 / 240 / 480
Myrinet (E card)	Clos	$ 995	$ 400	$1,395	6 / 450 / 900
IB 4x	Fat Tree	$1,000	$ 400	$1,400	6 / 820 / 790

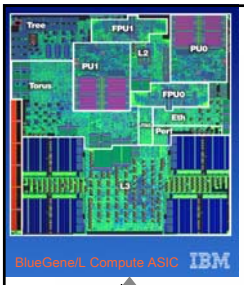
24th List: The TOP10

	Manufacturer	Computer	Rmax [TF/s]	Installation Site	Country	Year	#Proc
1	IBM	BlueGene/L β-System	70.72	DOE/IBM	USA	2004	32768
2	SGI	Columbia Altix, Infiniband	51.87	NASA Ames	USA	2004	10160
3	NEC	Earth-Simulator	35.86	Earth Simulator Center	Japan	2002	5120
4	IBM	MareNostrum BladeCenter JS20, Myrinet	20.53	Barcelona Supercomputer Center	Spain	2004	3564
5	CCD	Thunder Itanium2, Quadrics	19.94	Lawrence Livermore National Laboratory	USA	2004	4096
6	HP	ASCI Q AlphaServer SC, Quadrics	13.88	Los Alamos National Laboratory	USA	2002	8192
7	Self Made	X Apple XServe, Infiniband	12.25	Virginia Tech	USA	2004	2200
8	IBM/LLNL	BlueGene/L DD1 500 MHz	11.68	Lawrence Livermore National Laboratory	USA	2004	8192
9	IBM	pSeries 655	10.31	Naval Oceanographic Office	USA	2004	2944
10	Dell	Tungsten PowerEdge, Myrinet	9.82	NCSA	USA	2003	2500

07

399 system > 1 TFlop/s; 294 machines are clusters, top10 average 8K proc

11



IBM BlueGene/L

131,072 Processors

System
(64 racks, 64x32x32)
131,072 procs

Rack
(32 Node boards, 8x8x16)
2048 processors

Node Card
(32 chips, 4x4x2)
16 Compute Cards
64 processors

Compute Card
(2 chips, 2x1x1)
4 processors

Chip
(2 processors)

180/360 TF/s
32 TB DDR

2.9/5.7 TF/s
0.5 TB DDR

90/180 GF/s
16 GB DDR

5.6/11.2 GF/s
1 GB DDR

2.8/5.6 GF/s
4 MB (cache)

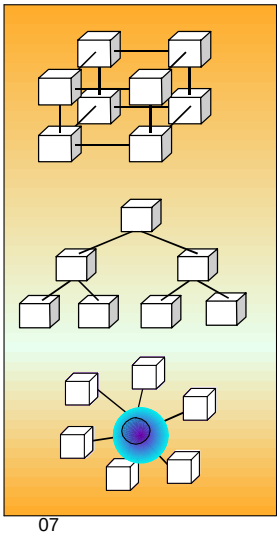
Full system total of
131,072 processors

"Fastest Computer"
BG/L 700 MHz 32K proc
16 racks
Peak: 91.7 Tflop/s
Linpack: 70.7 Tflop/s
77% of peak

07

12

BlueGene/L Interconnection Networks



3 Dimensional Torus
- Interconnects all compute nodes (65,536)
- Virtual cut-through hardware routing
- 1.4Gb/s on all 12 node links (2.1 GB/s per node)
- 1 µs latency between nearest neighbors, 5 µs to the farthest
- 4 µs latency for one hop with MPI, 10 µs to the farthest
- Communications backbone for computations
- 0.7/1.4 TB/s bisection bandwidth, 68TB/s total bandwidth

Global Tree
- Interconnects all compute and I/O nodes (1024)
- One-to-all broadcast functionality
- Reduction operations functionality
- 2.8 Gb/s of bandwidth per link
- Latency of one way tree traversal 2.5 µs
- ~23TB/s total binary tree bandwidth (64k machine)

Ethernet
- Incorporated into every node ASIC
- Active in the I/O nodes (1:64)
- All external comm. (file I/O, control, user interaction, etc.)

Low Latency Global Barrier and Interrupt
- Latency of round trip 1.3 µs

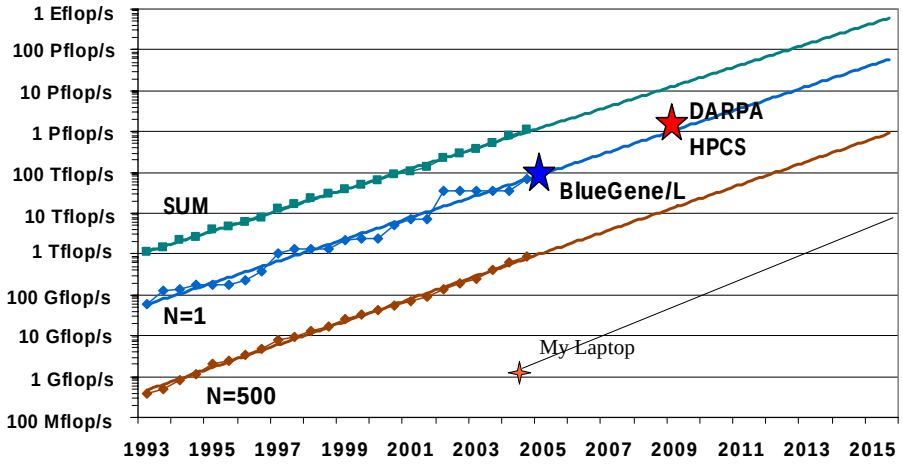
Control Network

NASA Ames: SGI Altix Columbia 10,240 Processor System

- Architecture: Hybrid Technical Server Cluster
- Vendor: SGI based on Altix systems
- Deployment: Today
- Node:
  - 1.5 GHz Itanium-2 Processor
  - 512 procs/node (20 cabinets)
  - Dual FPU's / processor

- System:
  - 20 Altix NUMA systems @ 512 procs/node = 10240 procs
  - 320 cabinets (estimate 16 per node)
  - Peak: 61.4 Tflop/s ; LINPACK: 52 Tflop/s
- Interconnect:
  - FastNumaFlex (custom hypercube) within node
  - Infiniband between nodes
- Pluses:
  - Large and powerful DSM nodes
- Potential problems (Gotchas):
  - Power consumption - 100 kw per node (2 Mw total)

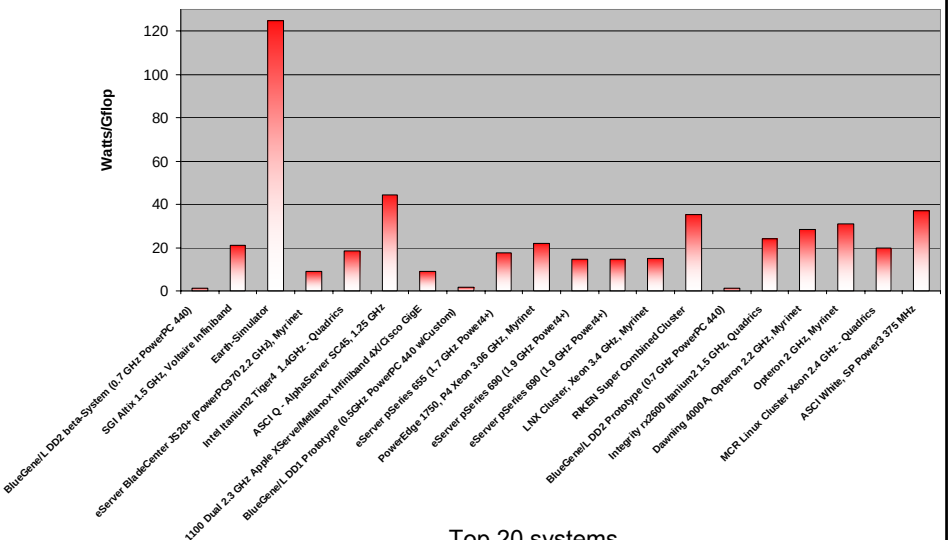
Performance Projection



07

15

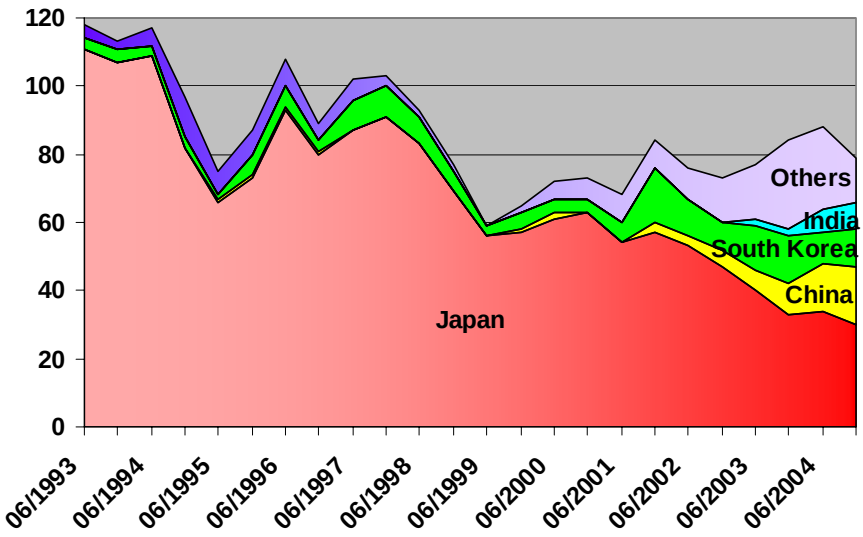
Power: Watts/Gflop (smaller is better)



Top 20 systems
Based on processor power rating only

16

Top500 in Asia (Numbers of Machines)



07

17

17 Chinese Sites on the Top500

Rank	Installation-site-name	Manufacturer	Computer	Area	Year	R max	Procs
17	Shanghai Supercomputer Center	Dawning	Dawning 4000A, Opteron 2.2 GHz, Myrinet	Research	2004	8061	2560
38	Chinese Academy of Science	lenovo	DeepComp 6800, Itanium2 1.3 GHz, QsNet	Academic	2003	4193	1024
61	Institute of Scientific Computing/Nankai University	IBM	xSeries Xeon 3.06 GHz, Myrinet	Academic	2004	3231	768
132	Petroleum Company (D)	IBM	BladeCenter Xeon 3.06 GHz, Gig-Ethernet	Industry	2004	1923	512
184	Geoscience (A)	IBM	BladeCenter Xeon 3.06 GHz, Gig-Ethernet	Industry	2004	1547	412
209	University of Shanghai	HP	DL360G3 Xeon 3.06 GHz, Infiniband	Academic	2004	1401	348
225	Academy of Mathematics and System Science	lenovo	DeepComp 1800 - P4 Xeon 2 GHz - Myrinet	Academic	2002	1297	512
229	Digital China Ltd.	HP	SuperDome 1 GHz/HPlex	Industry	2004	1281	560
247	Public Sector	IBM	xSeries Cluster Xeon 2.4 GHz - Gig-E	Governmn't	2003	1256	622
324	China Meteorological Administration	IBM	eServer pSeries 655 (1.7 GHz Power4+)	Research	2004	1107	1008
355	XinJiang Oil	IBM	BladeCenter Cluster Xeon 2.4 GHz, Gig-Ethernet	Industry	2003	1040	448
372	Fudan University	HP	DL360G3, Pentium4 Xeon 3.2 GHz, Myrinet	Academic	2004	1016	256
384	Huapu Information Technology	HP	SuperDome 875 MHz/HyperPlex	Industry	2004	1013	512
419	Saxony Developments Ltd	HP	Integrity Superdome, 1.5 GHz, HPlex	Industry	2004	971	192
481	Shenzhen University	Tsinghua U	DeepSuper-21C, P4 Xeon 3.06/2.8 GHz, Myrinet	Academic	2003	877	256
482	China Petroleum	HP	HP BL-20P, Pentium4 Xeon 3.06 GHz	Industry	2004	873	238
498	Digital China Ltd.	HP	SuperDome 875 MHz/HyperPlex	Industry	2004	851	416

Total performance growing by a factor of 3 every 6 months for the past 24 months

18

Important Metrics: Sustained Performance and Cost

- **Commodity processors**
  - Optimized for commercial applications.
  - Meet the needs of most of the scientific computing market.
  - Provide the shortest time-to-solution and the highest sustained performance per unit cost for a broad range of applications that have significant spatial and temporal locality (good caches use).

- **Custom processors**
  - For bandwidth-intensive applications that do not cache well, custom processors are more cost effective
  - Hence offering better capacity on just those applications.

07

High Bandwidth vs Commodity Systems

- High bandwidth systems have traditionally been vector computers
  - Designed for scientific problems
  - Capability computing
- Commodity processors are designed for web servers and the home PC market

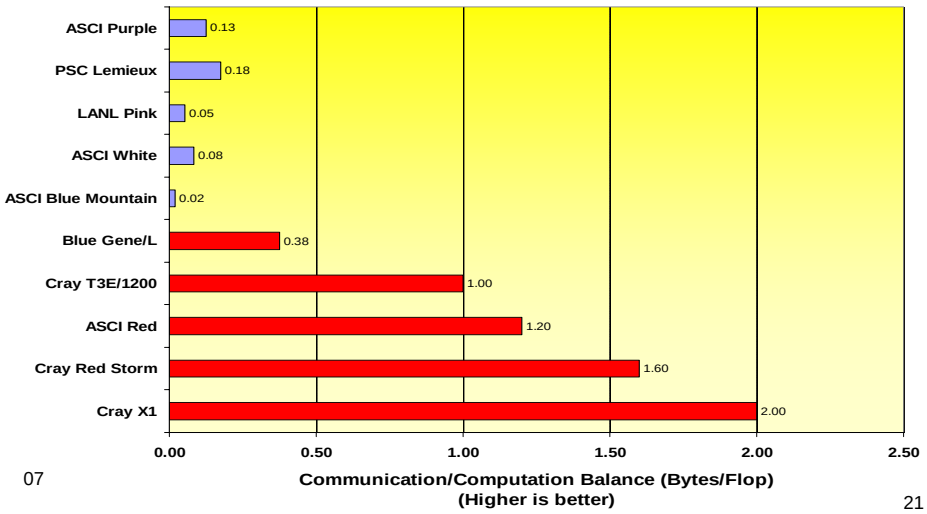
  (should be thankful that the manufactures keep the 64 bit fl pt)
  - Used for cluster based computers leveraging price point
- Scientific computing needs are different
  - Require a better balance between data movement and floating point operations. Results in greater efficiency.

System Balance - MEMORY BANDWIDTH

	Earth Simulator (NEC)	Cray X1 (Cray)	ASCI Q (HP EV68)	MCR Xeon	Apple Xserve IBM PowerPC
Year of Introduction	2002	2003	2002	2002	2003
Node Architecture	Vector	Vector	Alpha	Pentium	Power PC
Processor Cycle Time	500 MHz	800 MHz	1.25 GHz	2.4 GHz	2 GHz
Peak Speed per Processor	8 Gflop/s	12.8 Gflop/s	2.5 Gflop/s	4.8 Gflop/s	8 Gflop/s
Operands/Flop(main memory)	0.5	0.33	0.1	0.055	0.063

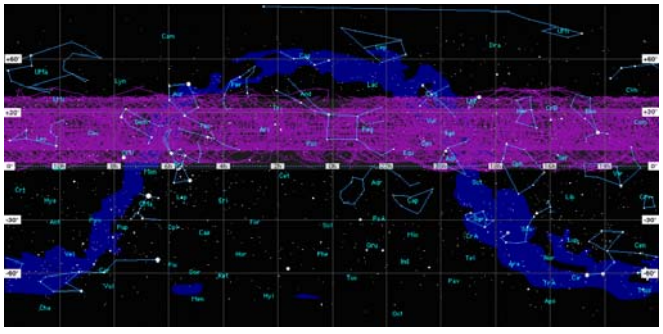
07

System Balance (Network)



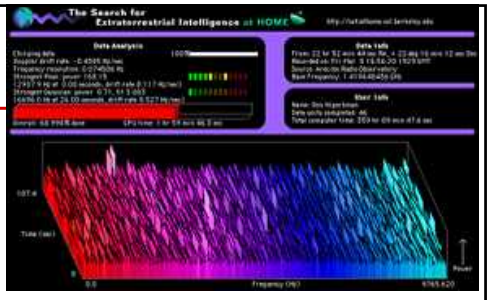
SETI@home: Global Distributed Computing

- Running on 500,000 PCs, ~1300 CPU Years per Day
  - 1.3M CPU Years so far
- Sophisticated Data & Signal Processing Analysis
- Distributes Datasets from Arecibo Radio Telescope

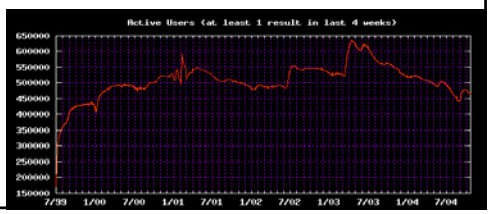


SETI@home

- Use thousands of Internet-connected PCs to help in the search for extraterrestrial intelligence.
- When their computer is idle or being wasted this software will download ~ half a MB chunk of data for analysis. Performs about 3 Tflops for each client in 15 hours.
- The results of this analysis are sent back to the SETI team, combined with thousands of other participants.
- About 5M users

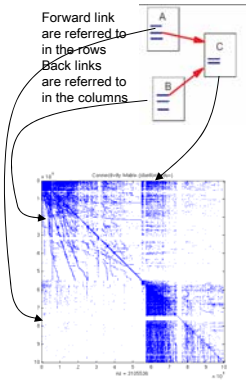


- **Largest distributed computation project in existence**
  - Averaging 72 Tflop/s



Google

- Google query attributes
  - 150M queries/day (2000/second)
  - 100 countries
  - 8.0B documents in the index
- Data centers
  - 100,000 Linux systems in data centers around the world
    - 15 TFlop/s and 1000 TB total capability
    - 40-80 1U/2U servers/cabinet
    - 100 MB Ethernet switches/cabinet with gigabit Ethernet uplink
  - growth from 4,000 systems (June 2000)
    - 18M queries then
- Performance and operation
  - simple reissue of failed commands to new servers
  - no performance debugging
    - problems are not reproducible



Eigenvalue problem; $Ax = \lambda x$
$n=8x10^9$
(see: MathWorks Cleve's Corner)

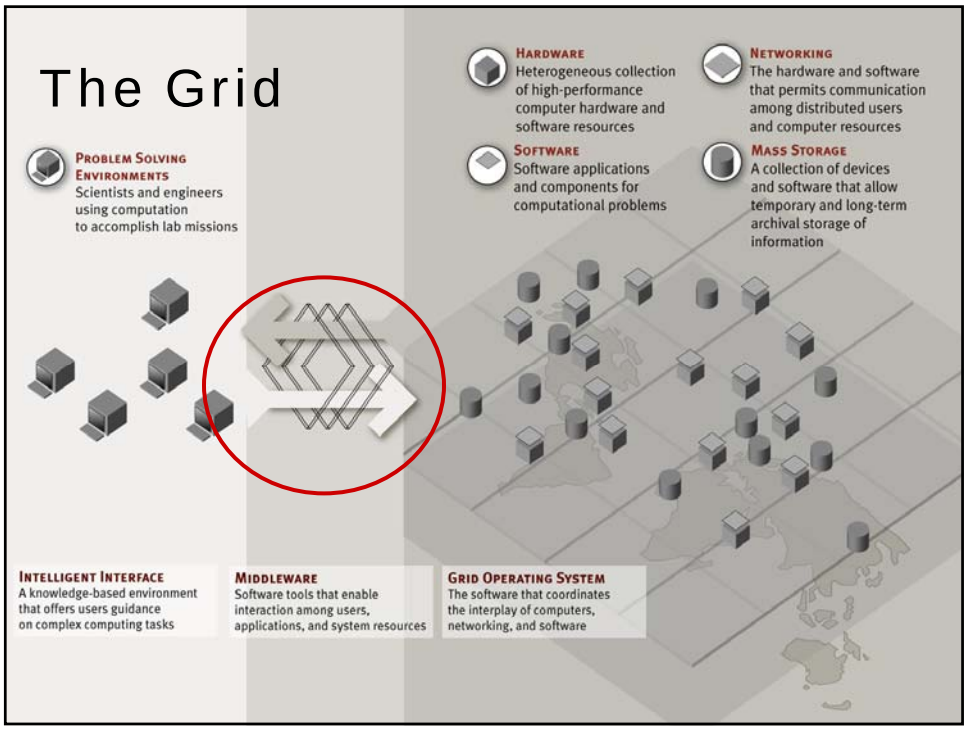
The matrix is the transition probability matrix of the Markov chain; $Ax = x$

Source: Monika Henzinger, Google & Cleve Moler

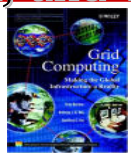
The Grid

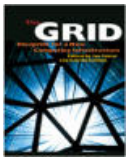
- The Grid is about gathering resources …
  - run programs, access data, provide services, collaborate
- …To enable and exploit large scale sharing of resources
- Virtual organization
  - Loosely coordinated groups
- Provides for remote access of resources
  - Scalable
  - Secure
  - Reliable mechanisms for discovery and access
- In some ideal setting:
  - User submits work, infrastructure finds an execution target
  - Ideally you don't care where.

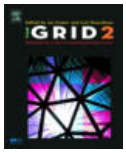
07



The Grid: The Good, The Bad, and The Ugly

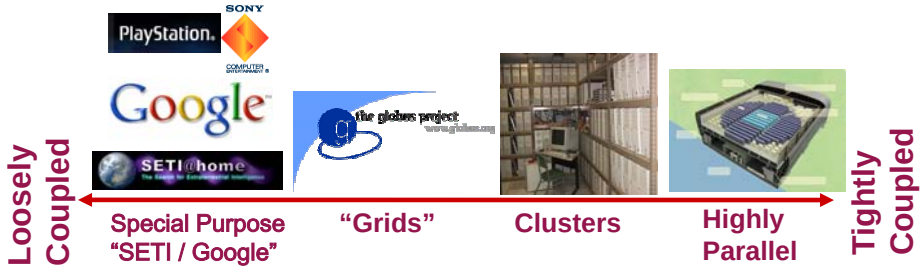






- **Good:**
  - Vision;
  - Community;
  - Developed functional software;
- **Bad:**
  - Oversold the grid concept;
  - Still too hard to use;
  - Solution in search of a problem;
  - Underestimated the technical difficulties;
  - Not enough of a scientific discipline;
- **Ugly:**
  - Authentication and security

The Computing Continuum

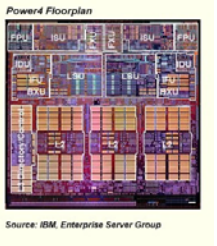


- **Each strikes a different balance**
  - computation/communication coupling
- **Implications for execution efficiency**
- ***Applications for diverse needs***
  - *computing is only one part of the story!*

Grids vs. Capability vs. Cluster Computing

- Not an "either/or" question
  - Each addresses different needs
  - Each are part of an integrated solution
- Grid strengths
  - Coupling necessarily distributed resources
    - instruments, software, hardware, archives, and people
  - Eliminating time and space barriers
    - remote resource access and capacity computing
  - Grids are not a cheap substitute for capability HPC
- Highest performance computing strengths
  - Supporting foundational computations
    - terascale and petascale "nation scale" problems
  - Engaging tightly coupled computations and teams
- Clusters
  - Low cost, group solution
  - Potential hidden costs
- Key is easy access to resources in a transparent way

Petascale Systems In 2008



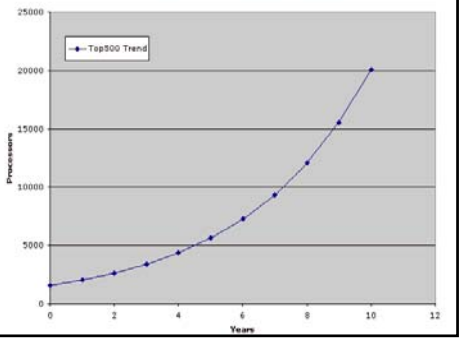
- Technology trends
  - multicore processors perhaps heterogeneous
    - IBM Power4 and SUN UltraSPARC IV
    - Itanium "Montecito" in 2005
    - quad-core and beyond are coming
  - reduced power consumption
    - laptop and mobile market drivers
  - increased I/O and memory interconnect integration
    - PCI Express, Infiniband, …
- Let's look forward a few years to 2008
  - 8-way or 16-way cores (8 or 16 processors/chip)
  - ~10 GFlop cores (processors) and 4-way nodes (4, 8-way cores/node)
  - 12x Infiniband-like interconnect, perhaps heterogeneous
- With 10 GFlop processors
  - 100K processors and 3100 nodes (4-way with 8 cores each)
  - *1-3 MW of power, at a minimum*
- To some extent, Petaflops systems will look like a *"Grid in a Box"*

How Big Is Big?

- Every 10X brings new challenges
  - 64 processors was once considered large
    - it hasn't been "large" for quite a while
  - 1024 processors is today's "medium" size
  - 8096 processors is today's "large"
    - we're struggling even here

2004-2014 System Siz

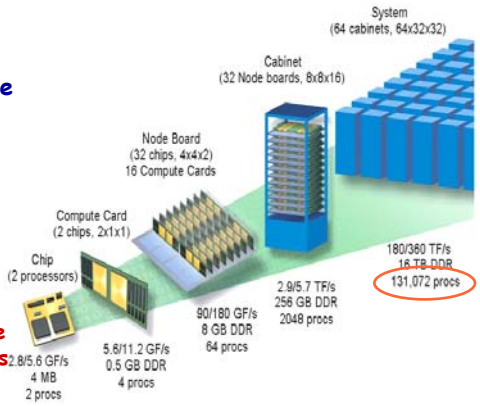
- 100K processor systems
  - are in construction
  - we have fundamental challenges in dealing with machines of this size
  - … and little in the way of programming support



07

Fault Tolerance in the Computation

- Some next generation systems are being designed with > 100K processors (IBM Blue Gene L).
- MTTF 10^5 - 10^6 hours for component.
  - sounds like a lot until you divide by 10^5!
  - Failures for such a system can be just a few hours, perhaps minutes away.



- Problem with the MPI standard, no recovery from faults.
- Application checkpoint / restart is today's typical fault tolerance method.
- Many cluster based on commodity parts don't have error correcting primary memory.

07

32

Real Crisis With HPC Is With The Software

- Programming is stuck
  - Arguably hasn't changed since the 60's
- It's time for a change
  - Complexity is rising dramatically
    - highly parallel and distributed systems
      - From 10 to 100 to 1000 to 10000 to 100000 of processors!!
    - multidisciplinary applications
- A supercomputer application and software are usually much more long-lived than a hardware
  - Hardware life typically five years at most.
  - Fortran and C are the main programming models
- Software is a major cost component of modern technologies.
  - The tradition in HPC system procurement is to assume that the software is free.
- We don't have many great ideas about how to solve this problem.

07

33

Collaborators / Support

- TOP500
  - H. Meuer, Mannheim U
  - H. Simon, NERSC
  - E. Strohmaier