An Overview of High Performance Computing and Trends

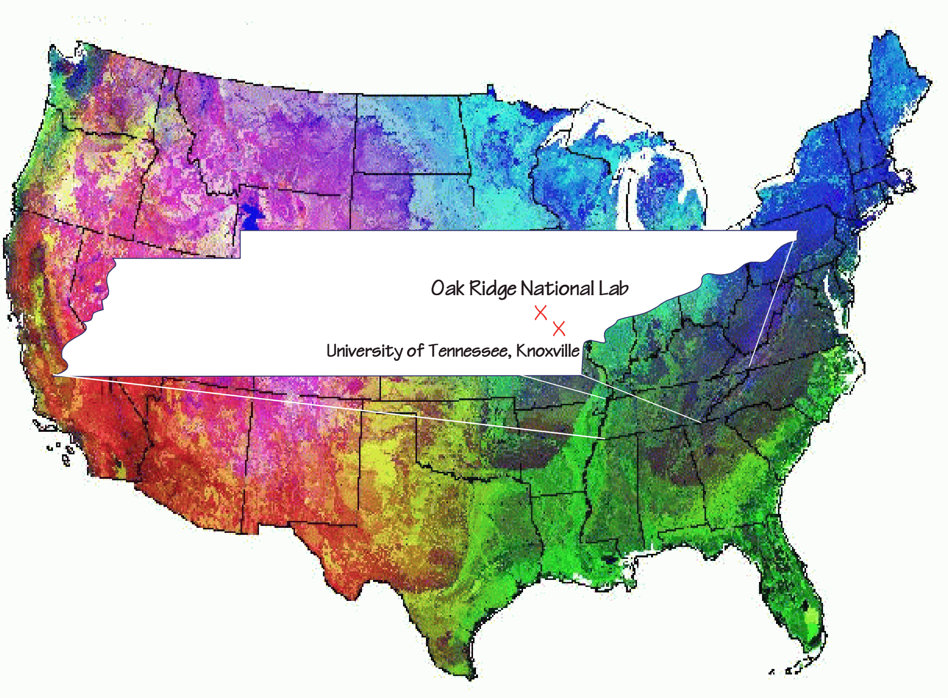
Jack Dongarra
Computer Science Department
University of Tennessee

1

Outline for the Next 3 Days

- Overview of High Performance Computing
- Impact of HPC on Linear Algebra Algorithms and Software
- Grid Computing

2



Innovative Computing Laboratory

- Numerical Linear Algebra
- Heterogeneous Distributed Computing
- Software Repositories
- Performance Evaluation

Software and ideas have found there way into many areas of Computational Science

Around 40 people: At the moment...

16 Researchers: Research Assoc/Post-Doc/Research Prof

15 Students: Graduate and Undergraduate

8 Support staff: Secretary, Systems, Artist

1 Long term visitors (Japan)

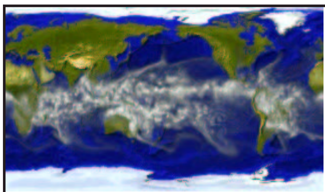
Responsible for about $4M/years in research funding from NSF, DOE, DOD, etc

4

Computational Science

- High Performance Computing offers a new way to do science:
  - Experiment - Theory - Computation
- Computation used to approximate physical systems - Advantages include:
  - Playing with simulation parameters to study emergent trends
  - Possible replay of a particular simulation event
  - Study systems where no exact theories exist

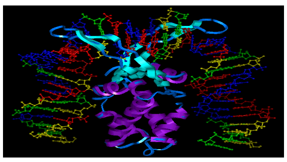
5

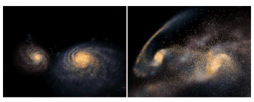


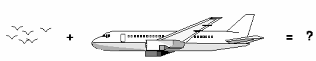
Why Turn to Simulation?

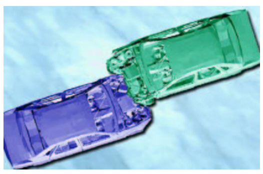
- When the problem is too . . .
  - Complex
  - Large / small
  - Expensive
  - Dangerous
- to do any other way.

- Climate / Weather Modeling
- Data intensive problems (data-mining, oil reservoir simulation)
- Problems with large length and time scales (cosmology)

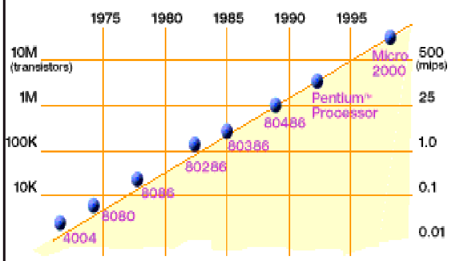








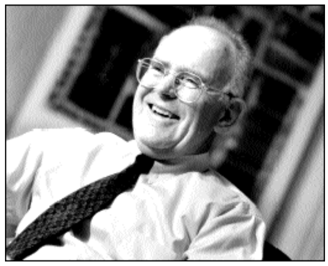
Technology Trends: Microprocessor Capacity



2X transistors/Chip Every 1.5 years

Called "Moore's Law"

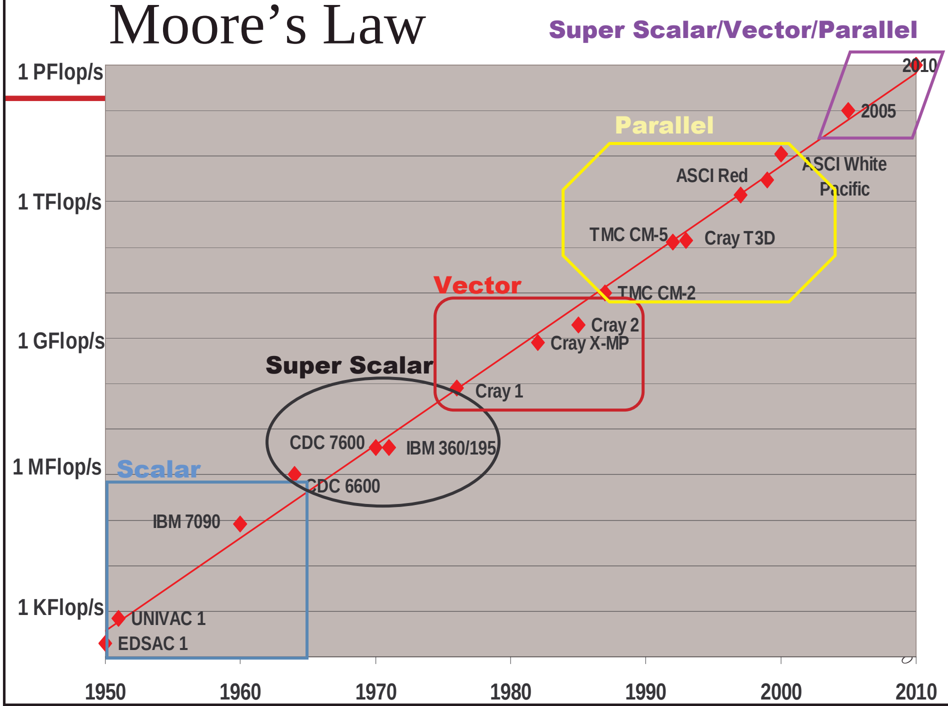
Microprocessors have become smaller, denser, and more powerful.
Not just processors, bandwidth, storage, etc



Gordon Moore (co-founder of Intel) predicted in 1965 that the transistor density of semiconductor chips would double roughly every 18 months.

7

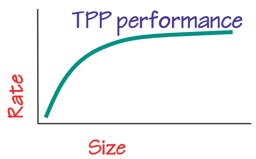
Moore's Law



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

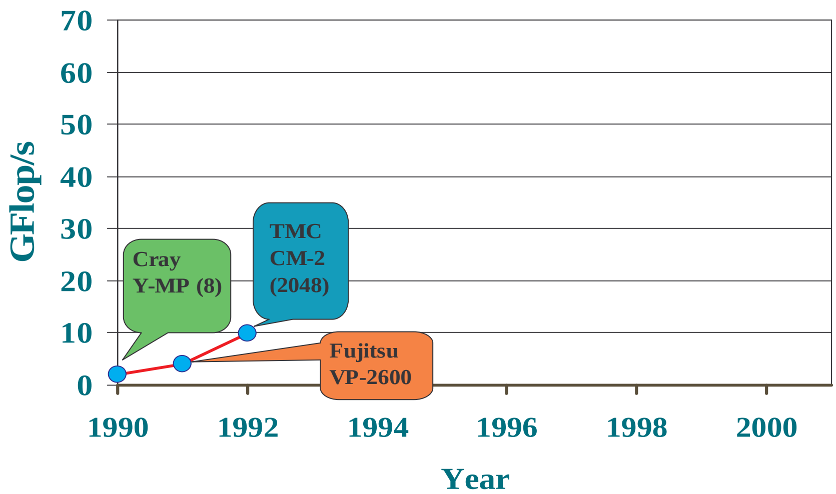
- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

Ax=b, *dense problem*



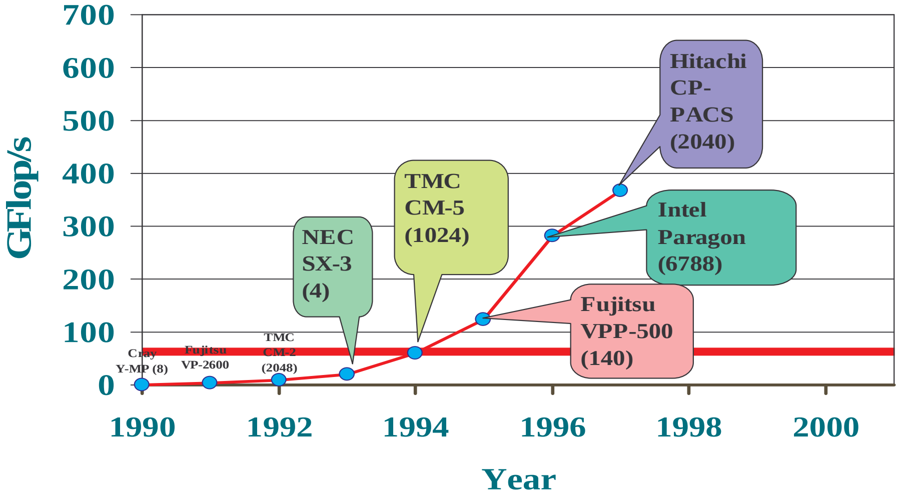
- Updated twice a year
  SC'xy in the States in November
  Meeting in Mannheim, Germany in June
- All data available from **www.top500.org**

Fastest Computer Over Time



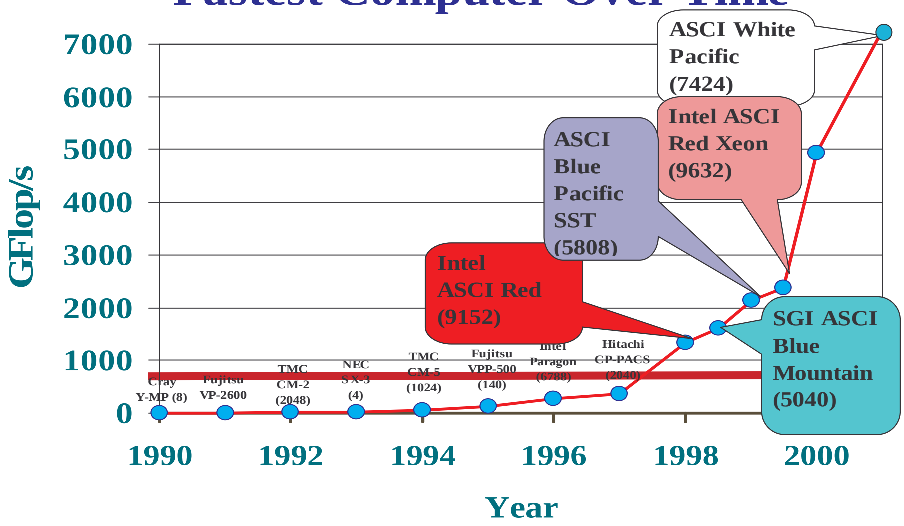
In 1980 a computation that took 1 full year to complete can now be done in ~ 10 hours!

Fastest Computer Over Time



In 1980 a computation that took 1 full year to complete
can now be done in ~ 16 minutes!

Fastest Computer Over Time

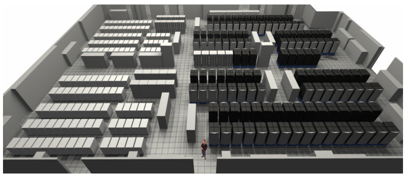


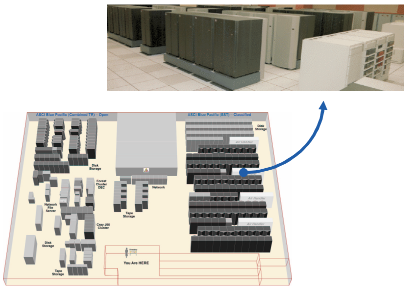
In 1980 a computation that took 1 full year to complete
can today be done in ~ 27 seconds!

Livermore National Laboratory –
IBM Blue Pacific and White SMP Superclusters

4TF Blue Pacific SST running

- 3 x 480 4-way SMP nodes
- 3.9 TF peak performance
- 2.6 TB memory
- 2.5 Tb/s bisectional bandwidth
- 62 TB disk
- 6.4 GB/s delivered I/O bandwidth



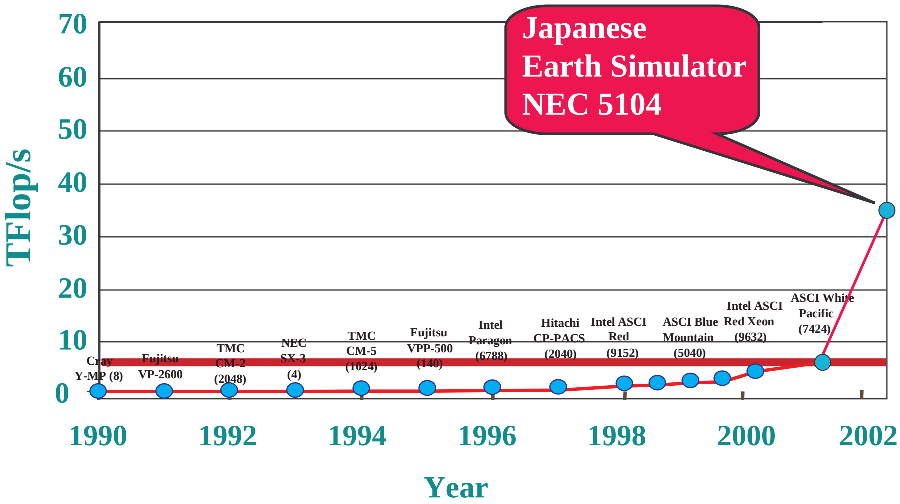


10TF ASCI White

- 512 Nighthawk 16-way SMP nodes
- 12. TF peak performance
- 4.0 TB memory
- 159 TB disk
- *2x I/O size and delivered bw over SST*
- *2.5x external network improvement*
- *Sufficient swap for GANG scheduling*

13

Fastest Computer Over Time



In 1980 a computation that took 1 full year to complete can today be done in ~ 5.4 seconds!

TOP500 list - Data shown

Manufacturer	Manufacturer or vendor
Computer Type	indicated by manufacturer or vendor
Installation Site	Customer
Location	Location and country
Year	Year of installation/last major update
Customer Segment	Academic,Research,Industry,Vendor,Class.
# Processors	Number of processors
R_{max}	Maxmimal LINPACK performance achieved
R_{peak}	Theoretical peak performance
N_{max}	Problemsize for achieving R_{max}
$N_{1/2}$	Problemsize for achieving half of R_{max}
N_{world}	Position within the TOP500 ranking

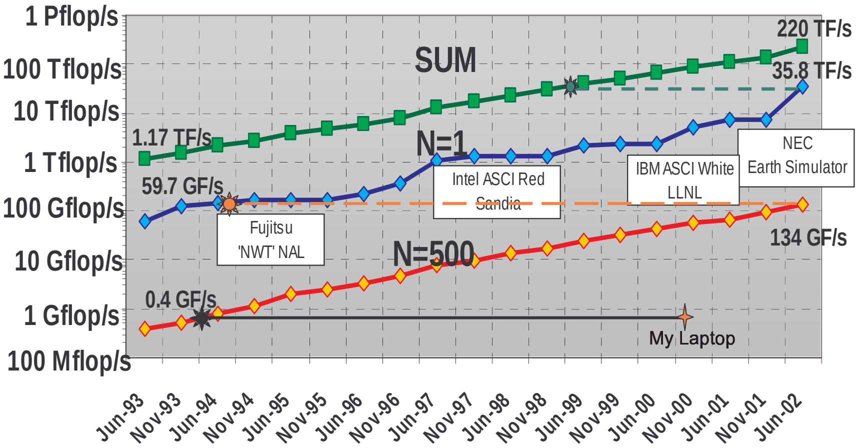
16

TOP10

Rank	Manufacturer	Computer	R_{max} [TF/s]	Installation Site	Country	Year	Area of Installation	# Proc
1	NEC	Earth-Simulator	35.86	Earth Simulator Center	Japan	2002	Research	5120
2	IBM	ASCI White SP Power3	7.23	Lawrence Livermore National Laboratory	USA	2000	Research	8192
3	HP	AlphaServer SC ES45 1 GHz	4.46	Pittsburgh Supercomputing Center	USA	2001	Academic	3016
4	HP	AlphaServer SC ES45 1 GHz	3.98	Commissariat a l'Energie Atomique (CEA)	France	2001	Research	2560
5	IBM	SP Power3 375 MHz	3.05	NERSC/LBNL	USA	2001	Research	3328
6	HP	AlphaServer SC ES45 1 GHz	2.92	Los Alamos National Laboratory	USA	2002	Research	2048
7	Intel	ASCI Red	2.38	Sandia National Laboratory	USA	1999	Research	9632
8	IBM	pSeries 690 1.3 GHz	2.31	Oak Ridge National Laboratory	USA	2002	Research	864
9	IBM	ASCI Blue Pacific SST, IBM SP 604e	2.14	Lawrence Livermore National Laboratory	USA	1999	Research	5808
10	IBM	pSeries 690 1.3 Ghz	2.00	IBM/US Army Reseach Lab (ARL)	USA	2002	Vendor	768

17

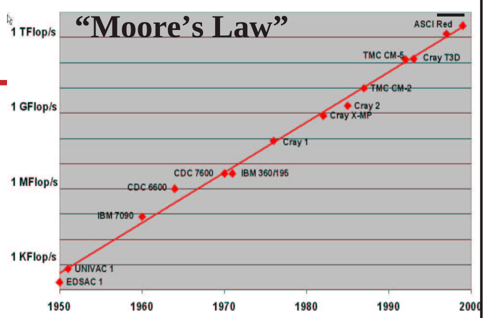
TOP500 - Performance



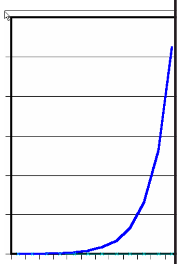
18

"Moore's Wall"

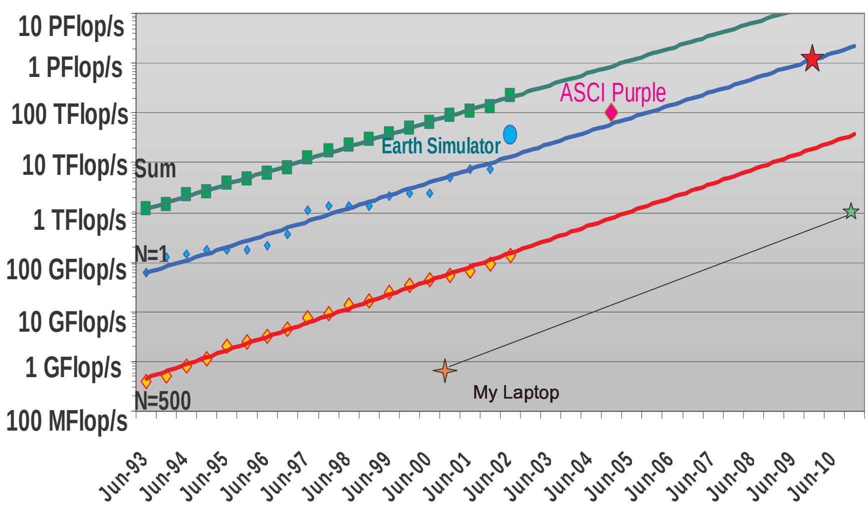
Horst Simon, NERSC



- Moore's Law predicts exponential growth
  - Performance doubling every 18 months
  - Usually plotted on semi-log scale,appears as straight line
- Human experience has a hard time deal with log scale
  - In 1980 a computation that took 1 full year to complete can now be done in minutes!
  - We are sitting at the bend of an exponential curve
- From our perspective Moore's Law appears as a "wall"
  - In a few years technology will again be completely different
  - Hard to predict what the future will be.

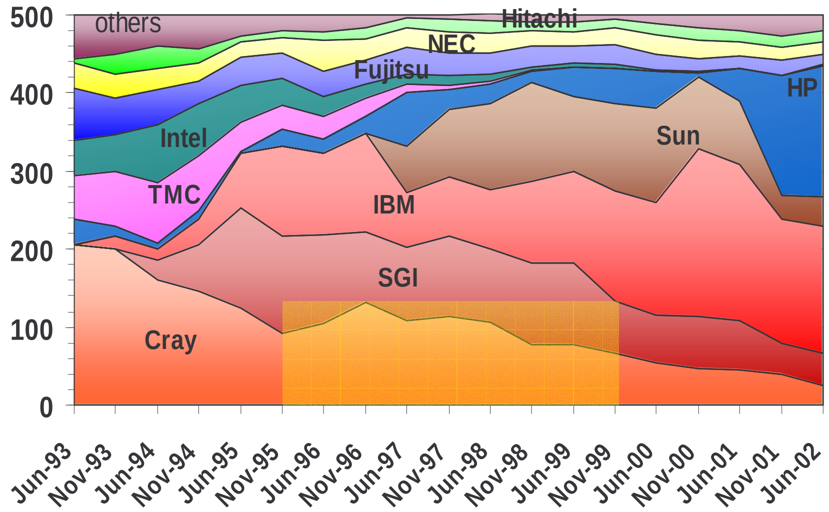


Performance Extrapolation



20

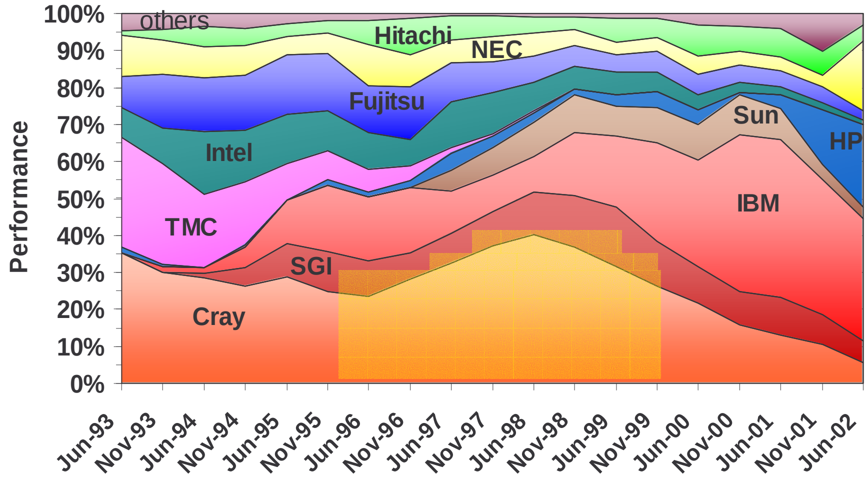
Manufacturers



21

HP 168, IBM 164

Manufacturers



22

IBM 33%, HP 22%, NEC 19%

Sun Systems on the Top500

List	Rank	Manufacturer	Computer	R_{max}(GFlops)	Installation Site	Country	Year	Installation Type	Processors
June 2002	113	Sun	HPC 4500 400 MHz Cluster	420.44	Service Provider	USA	2000	Industry	896
June 2002	115	Sun	HPC 4500 400 MHz Cluster	420.44	Sun	USA	2000	Vendor	896
June 2002	114	Sun	HPC 4500 400 MHz Cluster	420.44	Service Provider	USA	2000	Industry	896
June 2002	112	Sun	HPC 4500 400 MHz Cluster	420.44	Defense	Sweden	1999	Classified	896
June 2002	130	Sun	Fire 15K	357.10	Universitaet Aachen/RWTH	Germany	2002	Academic	288
June 2002	278	Sun	Fire 15K	197.30	Kyoto University	Japan	2002	Academic	144
June 2002	277	Sun	Fire 15K	197.30	Government	USA	2002	Classified	144
June 2002	271	Sun	Fire 15K	197.30	Automotive Manufacturer	France	2002	Industry	144
June 2002	276	Sun	Fire 15K	197.30	DaimlerChrysler	Germany	2002	Industry	144
June 2002	275	Sun	Fire 15K	197.30	DaimlerChrysler	Germany	2002	Industry	144
June 2002	274	Sun	Fire 15K	197.30	DaimlerChrysler	Germany	2002	Industry	144
June 2002	273	Sun	Fire 15K	197.30	BMW AG	Germany	2002	Industry	144
June 2002	272	Sun	Fire 15K	197.30	Automotive Manufacturer	France	2002	Industry	144
June 2002	309	Sun	Fire 6800/Sun Fire Link	195.80	High Performance Computing Virtual Laboratory	Canada	2002	Research	192
June 2002	358	Sun	Fire 6800	186.50	Universitaet Aachen/RWTH	Germany	2002	Academic	192
June 2002	359	Sun	Fire 6800	186.50	Universitaet Aachen/RWTH	Germany	2002	Academic	192
June 2002	496	Sun	HPC 10000 400 MHz Cluster	137.10	Telecommunication Company	South Africa	2001	Industry	256
June 2002	495	Sun	HPC 10000 400 MHz Cluster	137.10	Telecommunication Company	South Africa	2001	Industry	256
June 2002	482	Sun	HPC 10000 400 MHz Cluster	137.10	US Army Research Laboratory (ARL)	USA	1999	Research	256
June 2002	491	Sun	HPC 10000 400 MHz Cluster	137.10	Gateway	USA	2000	Industry	256
List	Rank	Manufacturer	Computer	R_{max}(GFlops)	Installation Site	Country	Year	Installation Type	Processors
June 2002	494	Sun	HPC 10000 400 MHz Cluster	137.10	MobilCom	Germany	2001	Industry	256
June 2002	492	Sun	HPC 10000 400 MHz Cluster	137.10	Information Technology Company	Germany	2001	Industry	256
June 2002	488	Sun	HPC 10000 400 MHz Cluster	137.10	Ford Motor Company	USA	2000	Industry	256
June 2002	486	Sun	HPC 10000 400 MHz Cluster	137.10	E-commerce	USA	2000	Industry	256
June 2002	484	Sun	HPC 10000 400 MHz Cluster	137.10	Clearstream Services	Luxembourg	2002	Industry	256
June 2002	479	Sun	HPC 10000 400 MHz Cluster	137.10	Motorola	USA	2000	Industry	256
June 2002	483	Sun	HPC 10000 400 MHz Cluster	137.10	Bank	USA	2000	Industry	256
June 2002	490	Sun	HPC 10000 400 MHz Cluster	137.10	GTE Communications	USA	2000	Industry	256
June 2002	489	Sun	HPC 10000 400 MHz Cluster	137.10	GTE Communications	USA	2000	Industry	256
June 2002	481	Sun	HPC 10000 400 MHz Cluster	137.10	Sun	USA	2000	Industry	256
June 2002	480	Sun	HPC 10000 400 MHz Cluster	137.10	New York City - Human Resources	USA	1999	Government	256
June 2002	498	Sun	HPC 10000 400 MHz Cluster	137.10	Telecommunication Company	USA	2001	Industry	256
June 2002	493	Sun	HPC 10000 400 MHz Cluster	137.10	MobilCom	Germany	2001	Industry	256
June 2002	497	Sun	HPC 10000 400 MHz Cluster	137.10	Telecommunication Company	South Africa	2001	Industry	256
June 2002	487	Sun	HPC 10000 400 MHz Cluster	137.10	EDS	Canada	2002	Industry	256
June 2002	485	Sun	HPC 10000 400 MHz Cluster	137.10	Clearstream Services	Luxembourg	2002	Industry	256
June 2002	499	Sun	HPC 10000 400 MHz Cluster	137.10	Telecommunication Company	USA	2001	Industry	256

French Top500 Computers

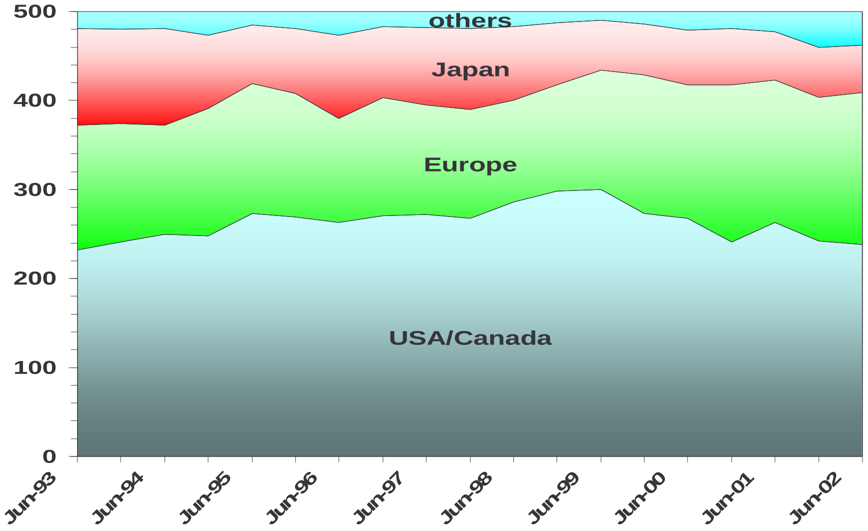
Country: France

ck to form

ere to show the explanation of the fields

List	Rank	Manufacturer	Computer	R_{max}(GFlops)	Installation Site	Country	Year	Installation Type	Processors
June 2002	4	Hewlett-Packard	AlphaServer SC ES45/1 GHz	3980.00	Commissariat a l'Energie Atomique (CEA)	France	2001	Research	2560
June 2002	59	IBM	pSeries 690 Turbo 1.3GHz GigEth	590.20	CNRS/IDRIS	France	2002	Academic	256
June 2002	90	IBM	SP Power3 375 MHz	494.00	Centre Informatique National (CINES)	France	2001	Academic	472
June 2002	140	Hewlett-Packard	AlphaServer SC ES40/833 MHz	326.40	Commissariat a l'Energie Atomique (CEA)	France	2000	Research	300
June 2002	162	NEC	SX-5/40M3	303.00	CNRS/IDRIS	France	2000	Academic	40
June 2002	174	Fujitsu	VPP5000/31	286.00	Meteo-France	France	1999	Research	31
June 2002	185	SGI	ORIGIN 3000 500 MHz	259.00	Centre Informatique National (CINES)	France	2001	Academic	320
June 2002	194	Hewlett-Packard	SuperDome 750 MHz/HyperPlex	245.30	France Telecom	France	2001	Industry	128
June 2002	222	Hewlett-Packard	SuperDome 750 MHz/HyperPlex	243.80	CIE Gegetel SI	France	2002	Industry	128
June 2002	231	IBM	pSeries 690 Turbo 1.3GHz GigEth	234.00	BOUYGTEL	France	2002	Industry	96
June 2002	250	IBM	SP Power3 375 MHz	214.00	PSA Peugeot Citroen	France	2001	Industry	212
June 2002	254	Hewlett-Packard	AlphaServer SC ES40/EV67	211.00	Commissariat a l'Energie Atomique (CEA)	France	1999	Research	232
June 2002	271	Sun	Fire 15K	197.30	Automotive Manufacturer	France	2002	Industry	144
June 2002	272	Sun	Fire 15K	197.30	Automotive Manufacturer	France	2002	Industry	144
June 2002	313	Hewlett-Packard	SuperDome/HyperPlex	195.80	France Telecom	France	2001	Industry	128
June 2002	343	Hewlett-Packard	SuperDome 750 MHz/HyperPlex	191.70	France Telecom	France	2001	Industry	96
June 2002	342	Hewlett-Packard	SuperDome 750 MHz/HyperPlex	191.70	France Telecom	France	2001	Industry	96
June 2002	373	IBM	SP Power3 375 MHz	179.00	CNRS/IDRIS	France	2001	Academic	176
June 2002	374	Hewlett-Packard	AlphaServer SC ES40/833 MHz	178.00	Commissariat a l'Energie Atomique (CEA)	France	2000	Research	160
June 2002	420	IBM	SP Power3 375 MHz	156.00	Dassault Aviation	France	2001	Industry	152
List	Rank	Manufacturer	Computer	R_{max}(GFlops)	Installation Site	Country	Year	Installation Type	Processors
June 2002	430	Fujitsu	VPP5000/16	149.00	Commissariat a l'Energie Atomique (CEA)	France	1999	Research	16
June 2002	454	Hewlett-Packard	SuperDome/HyperPlex	147.10	Hutchison Telecom	France	2001	Industry	96
June 2002	469	IBM	Netfinity Cluster PIII 1 GHz - Eth	138.00	Bank	France	2001	Industry	320

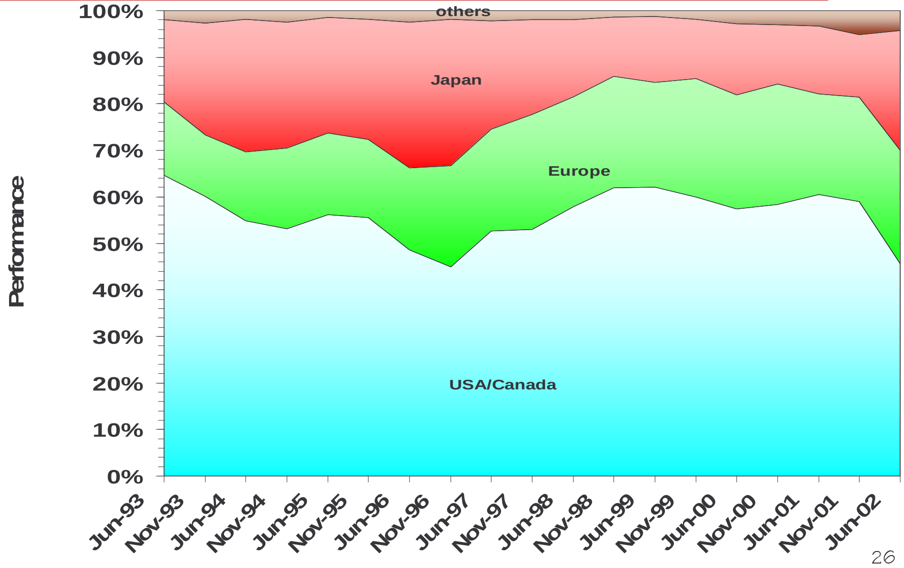
Continents



25

US 238 (242), Europe 171 (162), Japan 53 (56)

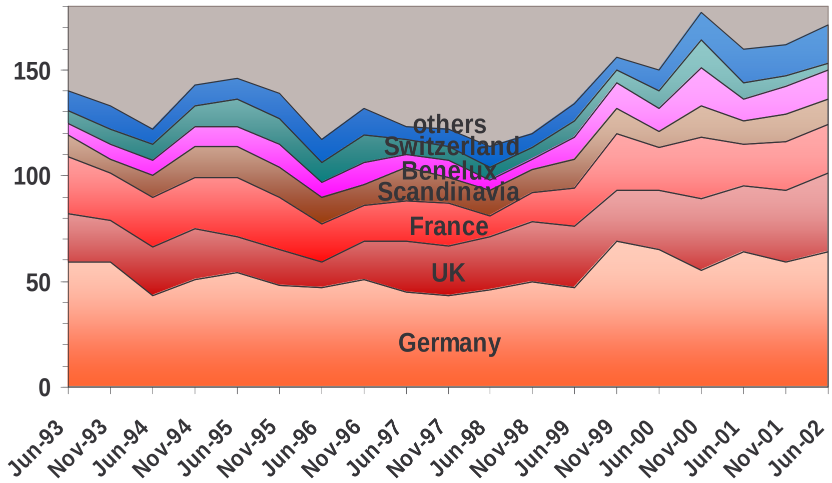
Continents - Performance



26

US 45% (59) Europe 24% (22) Japan 25% (13)

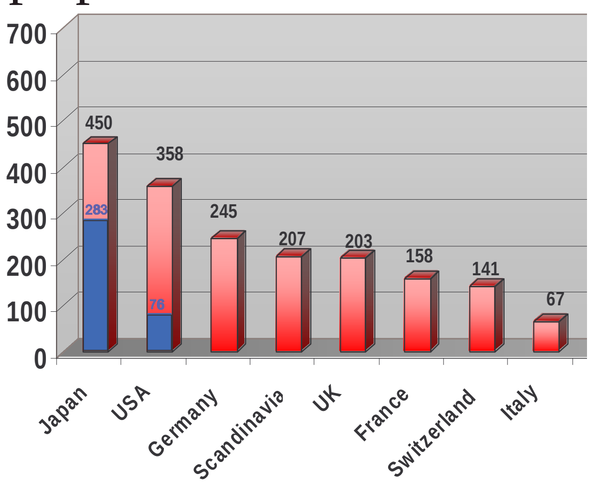
Europe - Countries



27

G 64, UK 37, F 23, SK 12, BEL 14, CH 3

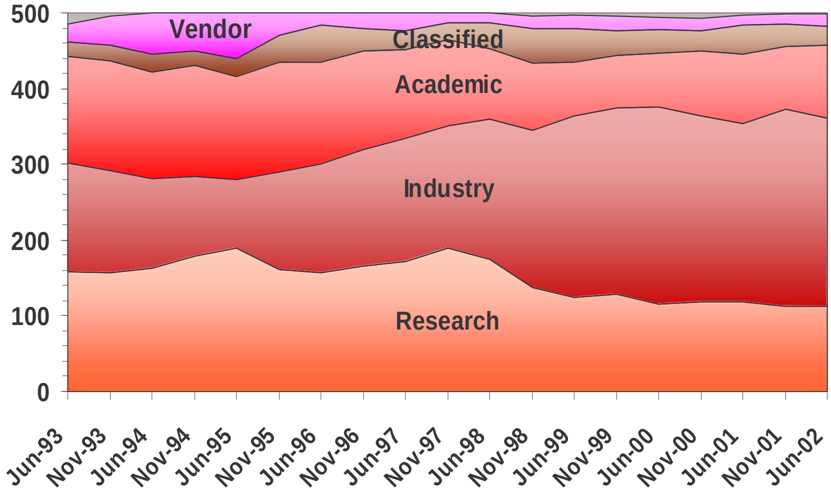
Kflops per Inhabitant



28

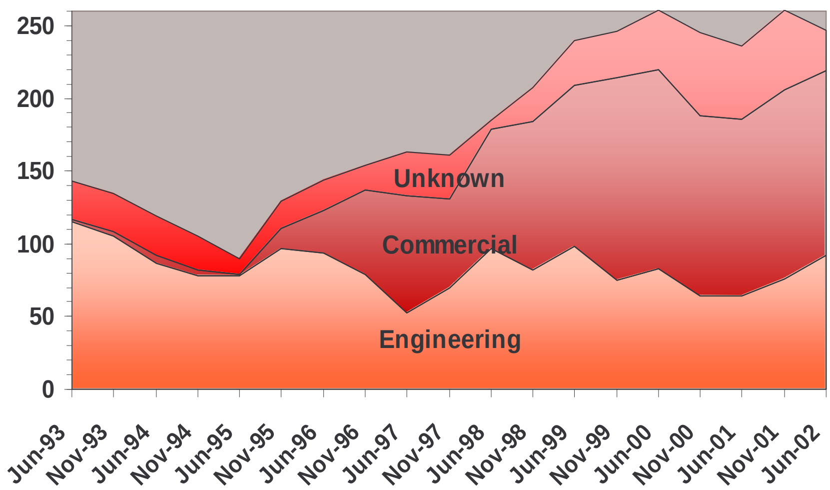
Japan 57 Tf/s US 99 Tf/s

Customer Type



29

Industrial Customer Segments



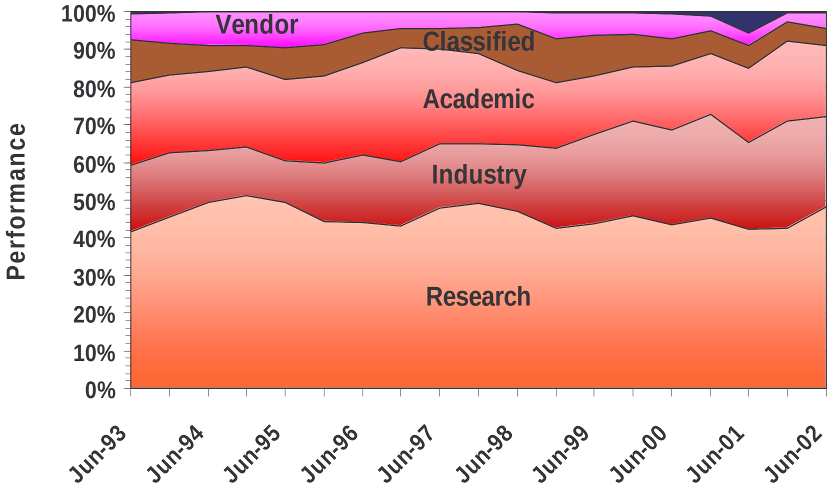
30

Excerpt from TOP500

Rank	Manufacturer	Computer	Rmax [GF/s]	Installation Site	Country	Area	# Proc
	…	…	…	…	…		…
40	IBM	SP Power3	795	Charles Schwab	USA	Finance	768
66	IBM	SP Power3	594	Sprint PCS	USA	Telecom	320
67	IBM	SP Power4	555	EDS General Motors	USA	Automotive	224
73	IBM	SP Power3	546	State Farm	USA	Database	520
125	IBM	Netfinity P3 Ethernet Cluster	366	WesternGeco	UK	Geophysics	1280
127	Hewlett-Packard	SuperDome HyperPlex	361	Centrica Plc	UK	Energy	196
…	…	…	…	…	…		…

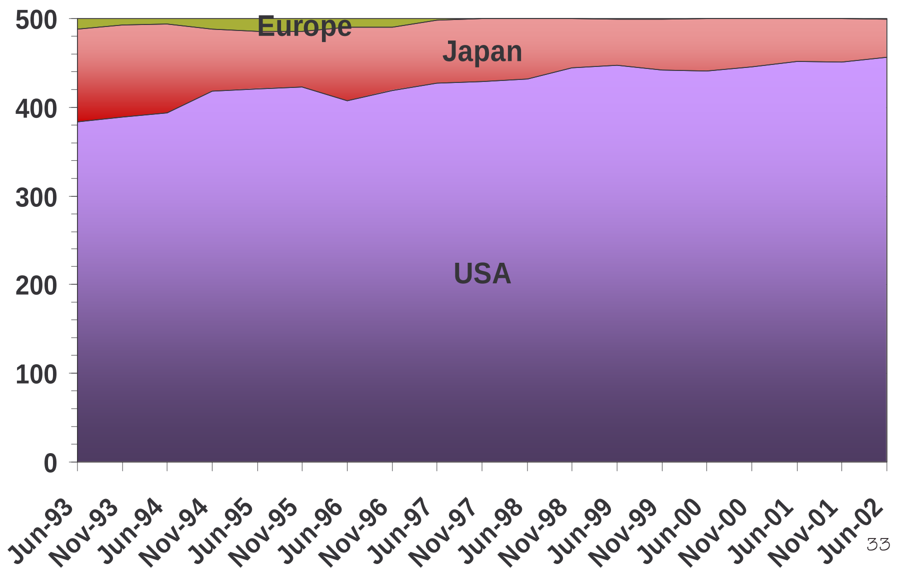
31

Customer Types - Performance



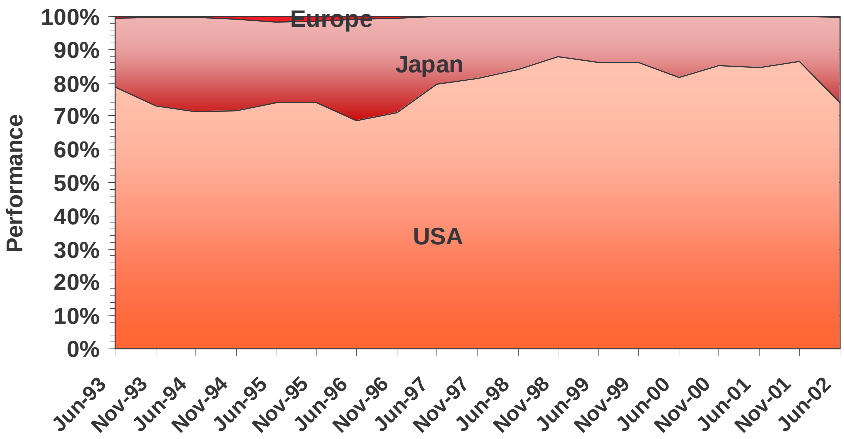
32

Producers



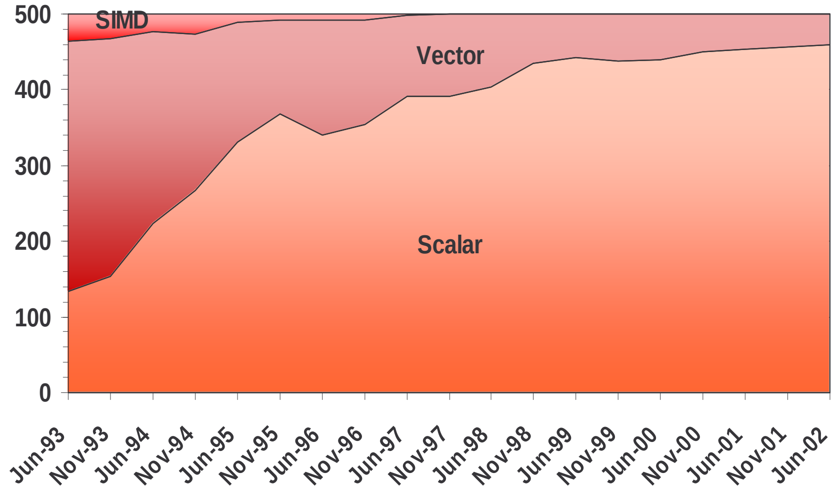
33

Producers - Performance



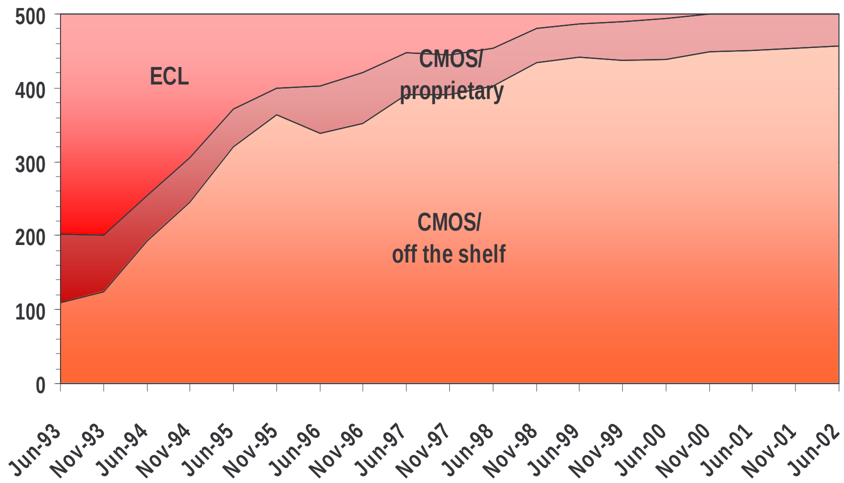
34

Processor Type



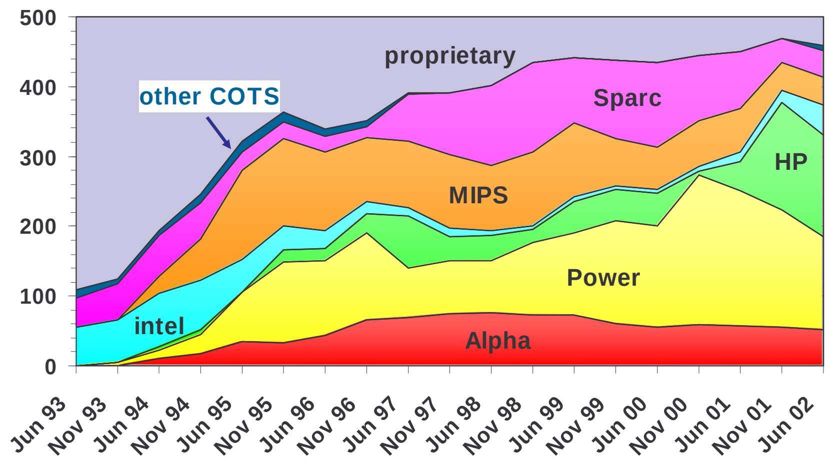
35

Chip Technology



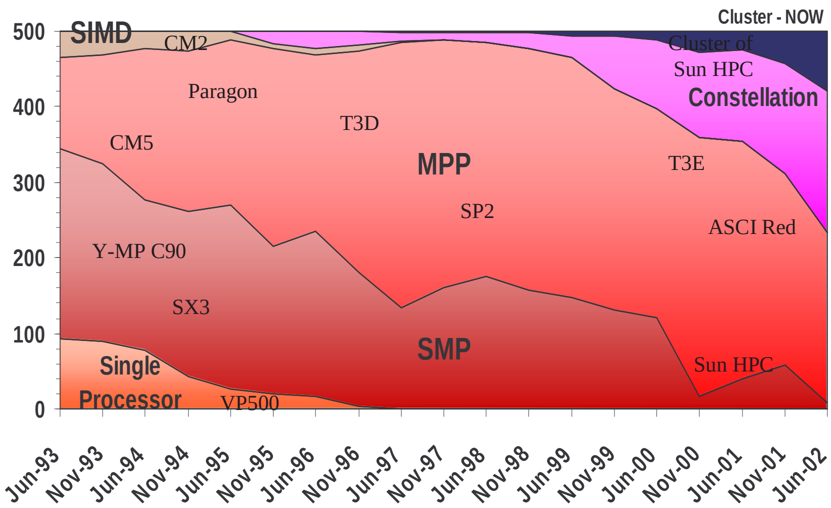
36

Chip Technology



37

Architectures

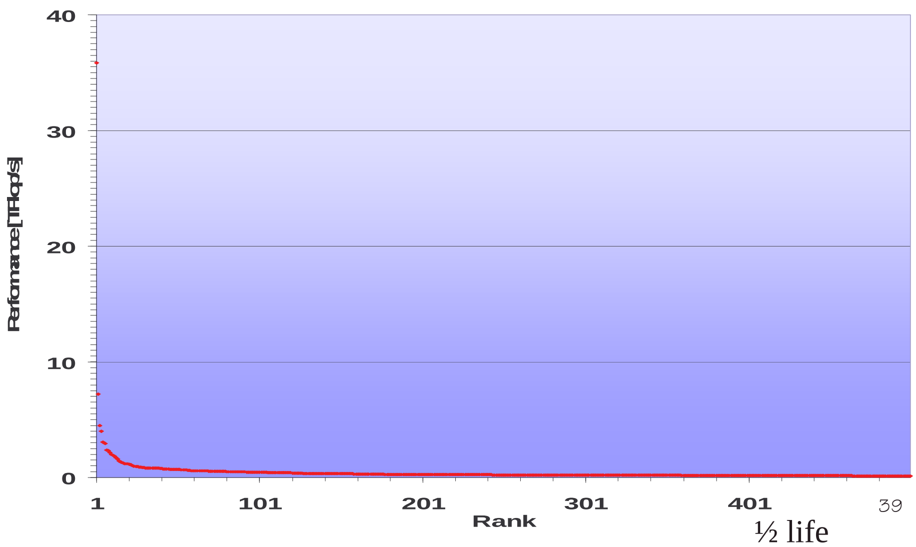


38

Constellation: # of p/n ≥ n

Performance Distribution

June 2002

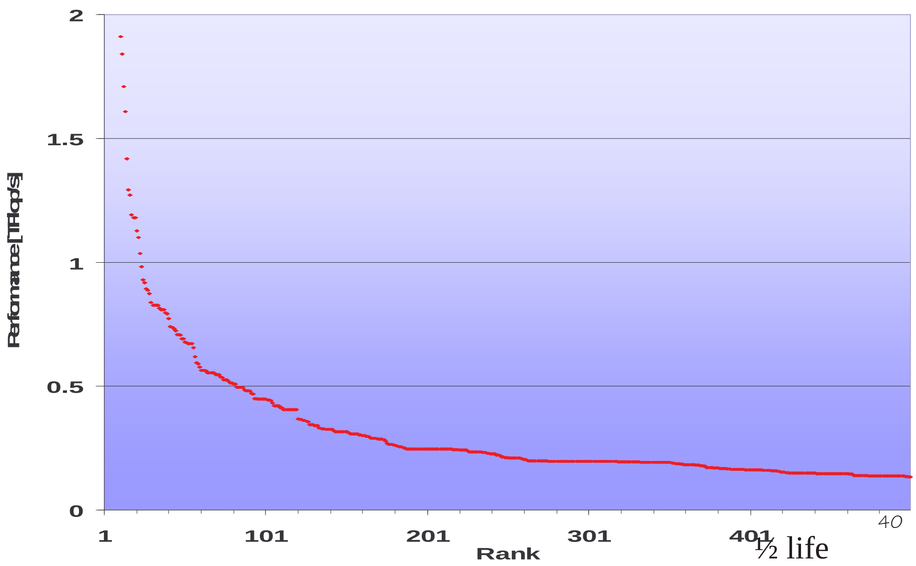


½ life

39

Performance Distribution

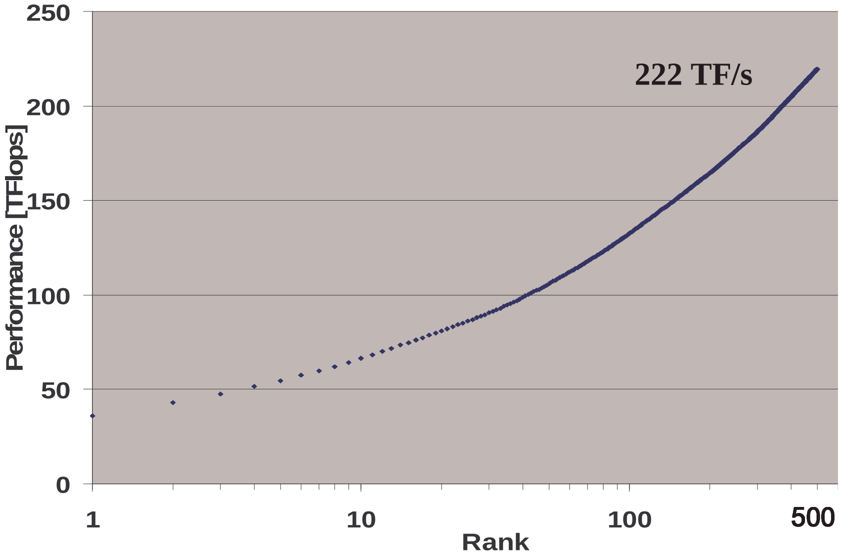
June 2002



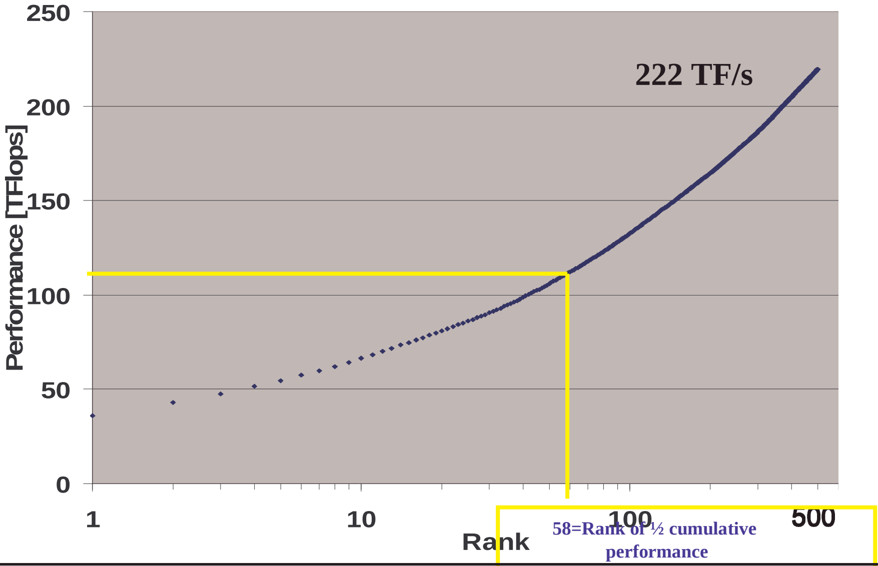
½ life

40

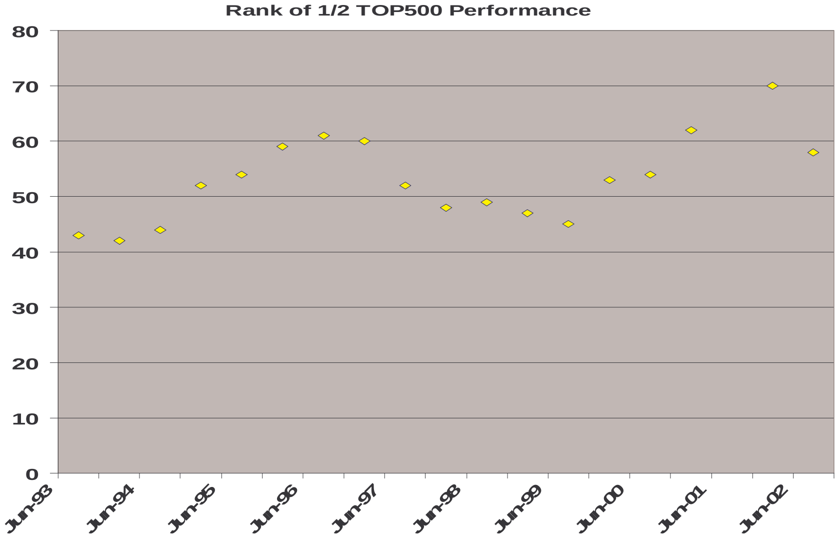
Cumulative Performance
June 2002



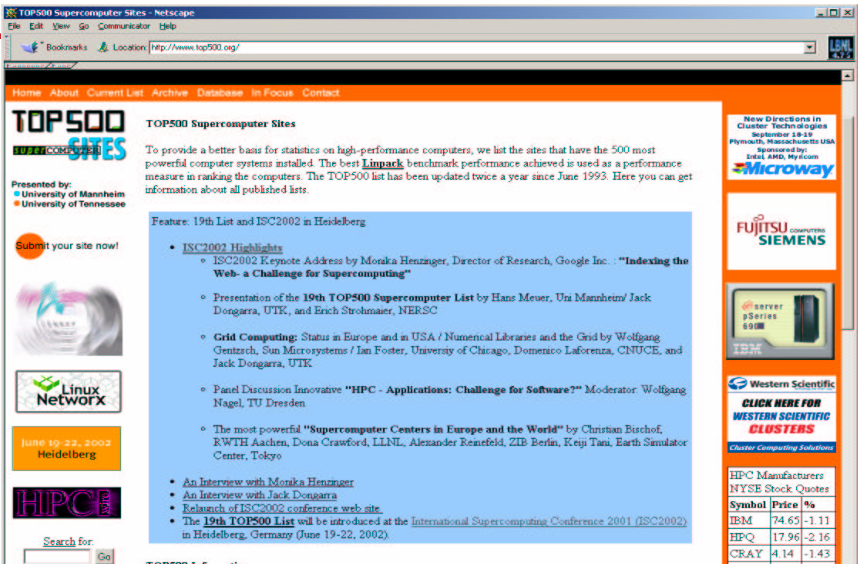
Cumulative Performance
June 2002



Performance Distribution



43



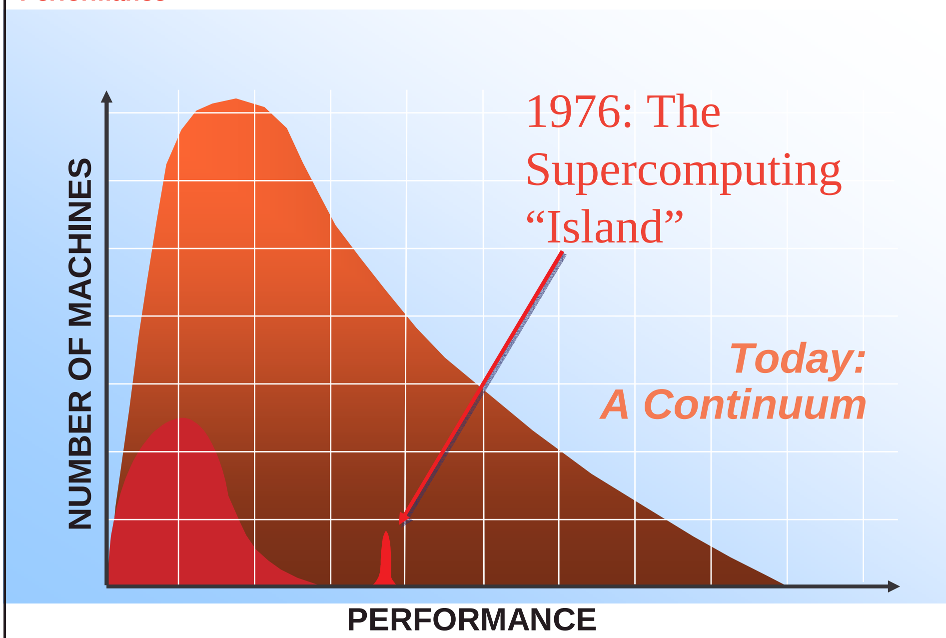
44

To Run Benchmark for TOP500

- HPL: High Performance Linpack
  Antoine Petitet and Clint Whaley, ICL, UTK
  - icl.cs.utk.edu/hpl
  - Needs only
    - MPI
    - BLAS or VSIPL
  - Highly scalable and efficient for the whole range of system sizes we see

45

Performance



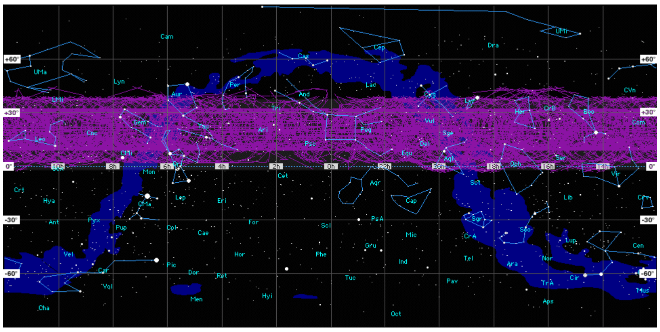
Petaflop Computers Within the Next Decade

- **Five basis design points:**
  - Conventional technologies
    - » 4.8 GHz processor, 8000 nodes, each w/16 processors
  - Processing-in-memory (PIM) designs
    - » Reduce memory access bottleneck
  - Superconducting processor technologies
    - » Digital superconductor technology, Rapid Single-Flux-Quantum (RSFQ) logic & hybrid technology multi-threaded (HTMT)
  - Special-purpose hardware designs
    - » Specific applications e.g. GRAPE Project in Japan for gravitational force computations
  - Schemes utilizing the aggregate computing power of processors distributed on the web
    - » SETI@home ~26 Tflop/s

47

SETI@home: Global Distributed Computing

- Running on 500,000 PCs, ~1000 CPU Years per Day
  - 485,821 CPU Years so far
- Sophisticated Data & Signal Processing Analysis
- Distributes Datasets from Arecibo Radio Telescope



48

SETI@home

- Use thousands of Internet-connected PCs to help in the search for extraterrestrial intelligence.
- Uses data collected with the Arecibo Radio Telescope, in Puerto Rico
- When their computer is idle or being wasted this software will download a 300 kilobyte chunk of data for analysis.
- The results of this analysis are sent back to the SETI team, combined with thousands of other participants.

- Largest distributed computation project in existence
  - ~ 400,000 machines
  - Averaging 27 Tflop/s
- Today many companies trying this for profit.

49

CNN.com / SCI-TECH

SEARCH GO

MAIN PAGE
WORLD
U.S.
WEATHER
BUSINESS
SPORTS
POLITICS
LAW
SCI-TECH
SPACE
HEALTH
ENTERTAINMENT
TRAVEL
EDUCATION
IN-DEPTH

CNN money XstreamQuotes POWERED BY STANDARD & POOR'S

QUICK NEWS
LOCAL
COMMUNITY
MULTIMEDIA
E-MAIL SERVICES
CNNtoGO
ABOUT US

CNN TV
what's on
show transcripts
CNN Headline News
CNN International
askCNN

EDITIONS
CNN.com Asia
CNN.com Europe
CNNenEspanol.com
set your edition

Languages

Time, Inc.

Try 4 issues of TIME

PCs tapped to help fight anthrax

January 22, 2002 Posted: 12:16 PM EST (1716 GMT)

SAVE THIS EMAIL THIS PRINT THIS MOST POPULAR

SAN JOSE, California (AP) -- A coalition of scientists and technology companies is asking people around the world to use their computers' extra processing power to help search for a cure for anthrax.

The project follows similar efforts to use "distributed computing" to hunt for extraterrestrial life and a cure for cancer. It is being launched Tuesday to help Oxford University researchers find ways to treat anthrax that can no longer be treated by antibiotics.

The project is based on the premise that the average personal computer uses between 13 percent and 18 percent of its processing power at any given time. It employs "peer-to-peer" technology, in which millions of computers can share files over the Internet.

Participants download a screen-saver that runs whenever their computers have resources to spare, and uses that power to perform computations for the project. When the user connects to the Internet, the computer sends data back to a central hub and gets another assignment.

The company that designed the program, United Devices Inc. of Austin, Texas, promises that no personal information on participants' PCs can be compromised while they take part.

If the project attracts more than 160,000 participants, it can give researchers more computational power than the world's 10 best supercomputers combined, said United Devices spokesman Andy Prince.

With enough participants, the project would provide researchers 10 times more power than the world's best supercomputer, said Graham Richards, the Oxford professor leading the study.

"The screen-saver doesn't cost you anything, and at least you're taking part in something, adding your bit," he said.

Intel, Microsoft involved

Scientists have discovered that the anthrax toxin is made up of three proteins that are not toxic on th eir own but become toxic after binding together

Grid Computing - from ET toAnthrax

UNITED DEVICES™

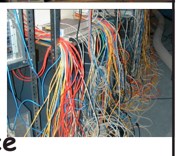
Petaflops (10^{15} flop/s) Computer Today?

2 GHz processor (O(10^9) ops/s)

- 1/2 Million PCs O(10^6)
- ~$2K each, O($10^3$) → $1B
- 100 Mwatts
- 5 acres
- 500,000 Windows licenses!!
- PC failure every second

51

High-Performance Computing Directions: Beowulf-class PC Clusters



Definition:

- COTS PC Nodes
  - Pentium, AMD, Alpha, PowerPC, SMP
- COTS LAN/SAN Interconnect
  - Ethernet, Myrinet, Giganet, ATM
- Open Source Unix
  - Linux, BSD
- Message Passing Computing
  - MPI, PVM

Advantages:

- Best price-performance
- Low entry-level cost
- Just-in-place configuration
- Vendor invulnerable
- Scalable
- Rapid technology tracking

Enabled by PC hardware, networks and operating system achieving capabilities of scientific workstations at a fraction of the cost and availability of industry standard message passing libraries. However, much more of a contact sport.

Excerpt from TOP500

Rank	Manufacturer	Computer	Rmax [GF/s]	Installation Site	Country	# Proc
	…	…	…	…	…	…
30	Self-made	Cplant/Ross	707	Sandia National Lab	USA	1369
34	IBM	Titan Cluster Itanium 800 MHz	594	NCSA	USA	320
39	NEC	Magi Cluster PIII 933 MHz	654	CBRC – Tsukuba Advanced Computing Center	Japan	1024
40	Self-made	SCoreIII PIII 933 MHz	618	Real World Computing, Tsukuba	Japan	1024
41	IBM	Netfinity Cluster PIII 1 GHz	594	NCSA	USA	1024
320	Dell	PowerEdge Cluster Windows2000	121	Cornell Theory Center	USA	252
…	…	…	…	…	…	…

53

Performance Numbers on RISC Processors

Processor	Cycle Time	Linpack n=100	Linpack n=1000	Peak
Intel P4	2540	1190 (23%)	2355 (46%)	5080
Intel/HP Itanium 2	1000	1102 (27%)	3534 (88%)	4000
Compaq Alpha	1000	824 (41%)	1542 (77%)	2000
AMD Athlon	1200	558 (23%)	998 (42%)	2400
HP PA	550	468 (21%)	1583 (71%)	2200
IBM Power 3	375	424 (28%)	1208 (80%)	1500
Intel P3	933	234 (25%)	514 (55%)	933
PowerPC G4	533	231 (22%)	478 (45%)	1066
SUN Ultra 80	450	208 (23%)	607 (67%)	900
SGI Origin 2K	300	173 (29%)	553 (92%)	600
Cray T90	454	705 (39%)	1603 (89%)	1800
Cray C90	238	387 (41%)	902 (95%)	952
Cray Y-MP	166	161 (48%)	324 (97%)	333
Cray X-MP	118	121 (51%)	218 (93%)	235
Cray J-90	100	106 (53%)	190 (95%)	200
Cray 1	80	27 (17%)	110 (69%)	160

54

Pentium 4 - SSE2
Today's "Sweet Spot" in Price/Performance

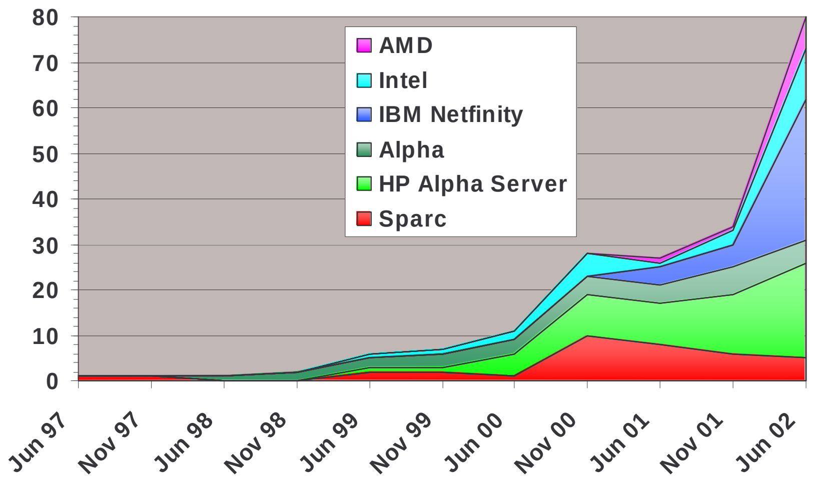
- 2.53 GHz, 400 MHz system bus, 16K L1 & 256K L2 Cache, theoretical peak of 2.53 Gflop/s, high power consumption
- Streaming SIMD Extensions 2 (SSE2)
  - which consists of 144 new instructions
  - includes SIMD IEEE double precision floating point
    - » Peak for 64 bit floating point 2X (5.06 Gflop/s)
    - » Peak for 32 bit floating point 4X (10.12 Gflop/s)
  - SIMD 128-bit integer
  - new cache and memory management instructions.
  - Intel's compiler supports these instructions today
  - ATLAS was trained to probe and detect SSE2

55

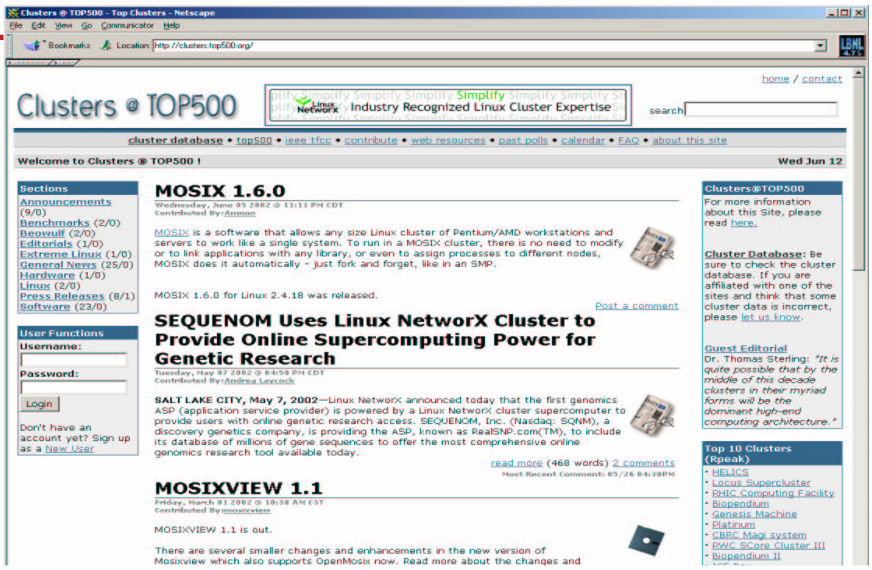
Table 1: **Performance in Solving a System of Linear Equations**

Computer	"LINPACK Benchmark" n = 100 OS/Compiler	Mflop/s	"TPP" Best Effort n=1000, Mflop/s	"Theoretical Peak" Mflop/s
Intel Pentium 4 (2.53 GHz)	ifc -O3 -xW -ipo -ip -align	1190	2355	5060
NEC SX-6/8 (8proc. 2.0 ns)			41520	64000
NEC SX-6/4 (4proc. 2.0 ns)			23680	32000
NEC SX-6/2 (2proc. 2.0 ns)			13350	16000
NEC SX-6/1 (1proc. 2.0 ns)	R12.1 -pi -Wf"-prob_use"	1161	7575	8000
Fujitsu VPP5000/1(1 proc.3.33ns)	frt -Wv,-r128 -Of -KA32	1156	8784	9600
Cray T932 (32 proc. 2.2 ns)			29360	57600
Cray T928 (28 proc. 2.2 ns)			28340	50400
Cray T924 (24 proc. 2.2 ns)			26170	43200
Cray T916 (16 proc. 2.2 ns)			19980	28800
Cray T916 (8 proc. 2.2 ns)			10880	14400
Cray T94 (4 proc. 2.2 ns)	f90 -O3,inline2	1129	5735	7200
HP RX5670 Itanium 2(4 proc 1GHz)			11430	16000
HP RX5670 Itanium 2(2 proc 1GHz)			6284	12000
HP RX5670 Itanium 2(1 proc 1GHz)	f90 +DSmckinley +O3 +Oinline_budget=100000 +Ono_ptrs_to_globals	1102	3534	4000
HP RX2600 Itanium 2(2 proc 1GHz)			6251	8000
HP RX2600 Itanium 2(1 proc 1GHz)	f90 +DSmckinley +O3 +Oinline_budget=100000 +Ono_ptrs_to_globals	1102	3528	4000

NOW - Cluster



57



58

Clusters @ TOP500

home / contact

search

cluster database • top500 • ieee tfcc • contribute • web resources • past polls • calendar • FAQ • about this site

Welcome to Clusters @ TOP500 ! Thu Nov 29

Sections
Announcements (8/0)
Benchmarks (2/0)
Beowulf (2/0)
Editorials (1/0)
Extreme Linux (1/0)
General News (23/0)
Hardware (1/0)
Linux (2/0)
Press Releases (7/0)
Software (20/0)

User Functions
Username:
Password:
Login
Don't have an account yet? Sign up as a New User

Cluster Sublist

This is **no official ranking**. Please read here to learn more about the results and the benchmarks.

Number of results: 171

Go back to form

#	Site	Country	System Name	Integrator	Node Number	Total Processors	Total Peak Performance	Interconnect
1	Locus Discovery	USA	Locus Supercluster	Self, Western Scientific, VA L	708	1416	1416.00	Fast Ethernet
2	Inpharmatica Ltd.	United Kingdom	Biopendium	In house	800	1220	1061.00	Fast Ethernet
3	Shell Technology Exploration and Production	Netherlands	Genesis Machine	IBM	1030	1038	1037.10	Gigabit Ethernet
4	NCSA	USA	Platinum	IBM	516	1032	1032.00	Myrinet 2000
5	Brookhaven National Laboratory	USA	RHIC Computing Facility	VA Linux and IBM	638	1276	990.80	Fast Ethernet
6	AIST - Computational Biology Research Center	Japan	CBRC Magi system	NEC	520	1040	967.20	Myrinet 2000
7	Real World Computing Partnership	Japan	RWC SCore Cluster III	Self-made	512	1024	955.40	Myrinet 2000
8	University of Utah, Center for High Performance Computing	USA	ICE Box	Self Made	303	388	814.66	Fast Ethernet
9	Incyte Genomics	USA	Incyte Genomics	In house	767	1511	754.00	Gigabit Ethernet
10	Sandia National Lab	USA	CPlant Siberia	Self-made	628	628	628.00	Myrinet

- Peak performance
- Interconnection
- http://clusters.top500.org
- Benchmark results to follow in the coming months

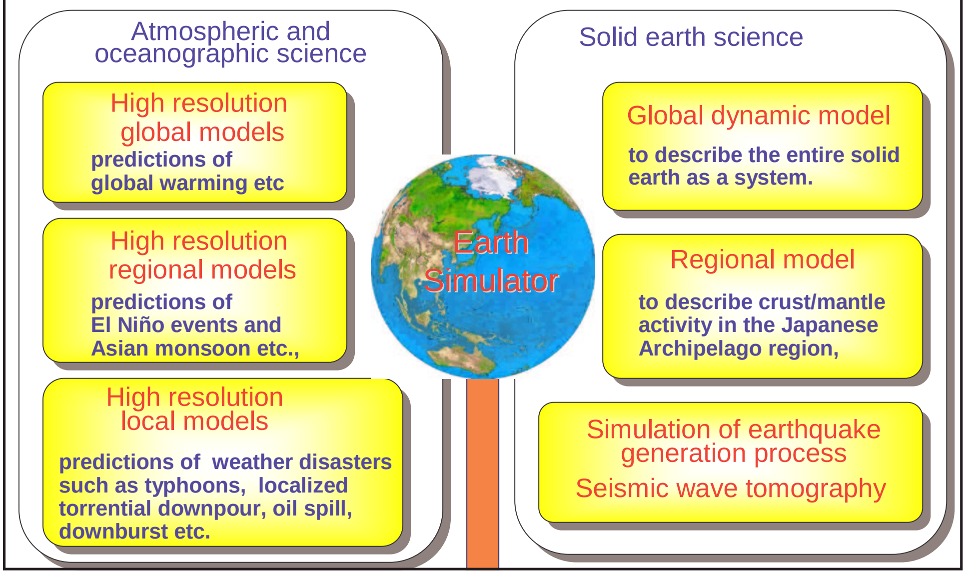
59

Notes on the Earth Simulator

Jack Dongarra
Computer Science Department
University of Tennessee

60

Development Center
Japan Atomic Energy Research Institute

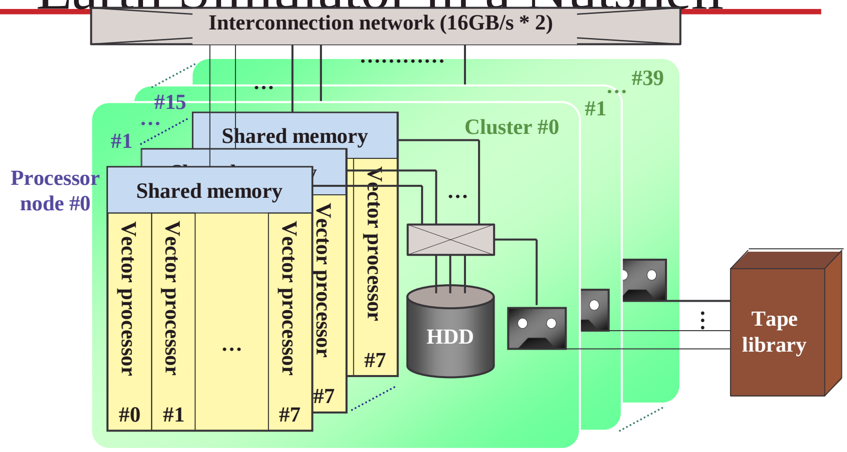


Earth Simulator

- Based on the NEC SX architecture, 640 nodes, each node with 8 vector processors (8 Gflop/s peak per processor), 2 ns cycle time, 16GB shared memory.
  - Total of 5104 total processors, 40 TFlop/s peak, and 10 TB memory.
- It has a single stage crossbar (1800 miles of cable) 83,000 copper cables, 16 GB/s cross section bandwidth.
- 700 TB disk space
- 1.6 PB mass store
- Area of computer = 4 tennis courts, 3 floors

62

Earth Simulator in a Nutshell

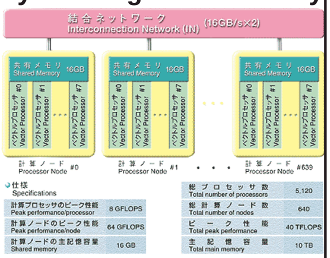


Specifications	
Peak performance / processor	8 Gflops
Peak performance / node	64 Gflops
Shared memory	16 GB

Total number of processors	5,120
Total number of nodes	640
Total peak performance	40 Tflops
Total main memory	10 TB

Outline of the Earth Simulator Computer

- Architecture : A MIMD-type, distributed memory, parallel system consisting of computing nodes in which vector-type multi-processors are tightly connected by sharing main memory
  - Total number of processor nodes: 640
  - Number of PE's for each node: 8
  - Total number of PE's: 5120
  - Peak performance of each PE: 8 GFLOPS
  - Peak performance of each node: 64 GFLOPS

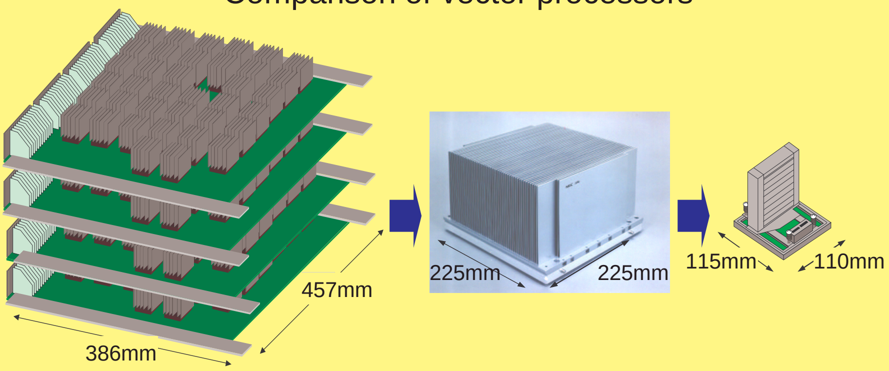


- Main memory : 10 TB (total).
  Shared memory / node : 16 GB
- Interconnection network: Single-Stage Crossbar Network
- Performance : Assuming the efficiency 12.5%, the peak performance 40 TFLOPS (the effective performance for an atmospheric circulation model is more than 5 TFLOPS).

Earth Simulator Research and Development Center

R&D results

Comparison of vector processors

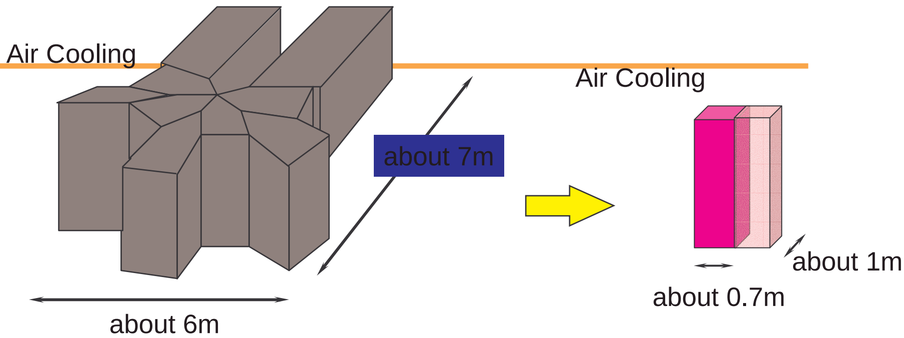


SX4	SX5	Earth Simulator
8 Gflops (2 Gflop/s x 4) Clock :125MHz LSI: 0.35μm CMOS 37x4=148 LSIs	8 Gflop/s Clock :250MHz LSI: 0.25μm CMOS 32 LSIs	8 Gflop/s Clock :500MHz/1GHz LSI: 0.15μm CMOS 1 chip processor

R&D results

Comparison of cabinets for 1 node

Present distributed-memory supercomputer (SX-4) 1 node	Earth Simulator 1 node
Peak Performance : 64 Gflops	Peak Performance : 64 Gflops
Main Memory : 16GB	Main Memory : 16GB
Electric Power : 90KVA	Electric Power : 8KVA



R&D results

R&D Issues on Hardware Technologies

(1) LSI Technology

- Enhancement of clock cycle 150MHz ⇒ 500MHz (partly 1GHz)
- Development of high density LSI
  - 0.15μm CMOS + Cu interconnection (8 layers)
  - 1.50-2.0 million transistors/cm^2 ⇒ 10 million transistors/cm^2
- Enlargement of chip size (about 2cm ×2cm)

High performance one-chip vector processor: OCVP-ES

(2) Packaging Technology

- Build-up PCB (110mm x 115mm)
  - Line width / Spacing : 25μm / 25μm
  - 6 core layers + 4 build-up layers on both surfaces
- number of pins/chip <1000 (present) ⇒ 4000 - 5000

(3) Cooling Technology

- Air cooling using heat pipe technology (Max. 170W per chip)

(4) Board to Board Interconnection Technology

- Interface connector 0.5mm pitch surface mount
- Interface cable 0.6mm diameter coaxial cable and 3.8ns/m delay time

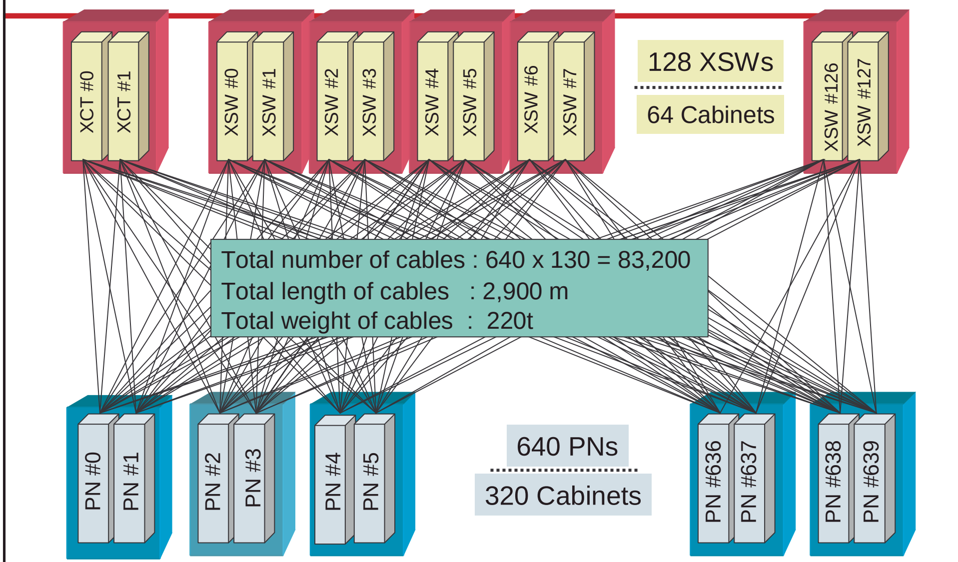
(5) PN-IN Interconnection Technology

- 40m transmission distance with fine tuned equalizer circuit

Earth Simulator Research and Development Center

R&D results

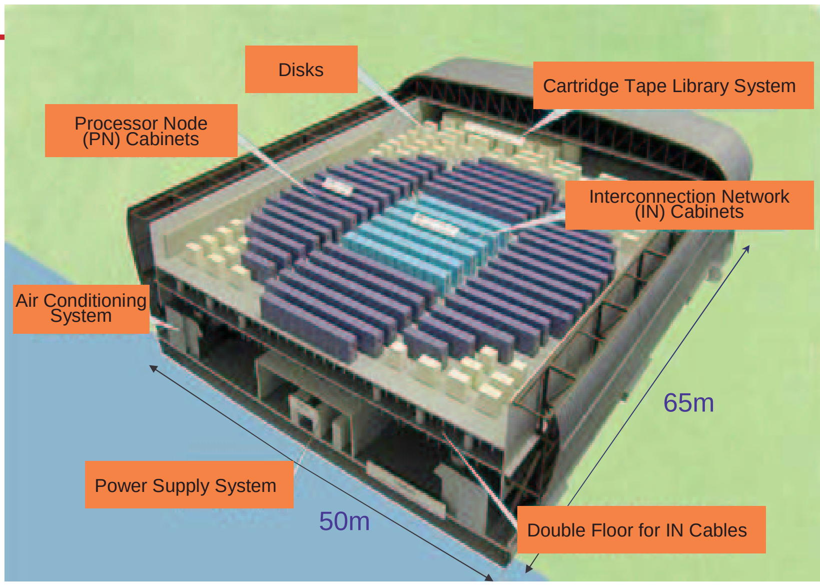
Connection between processor nodes (crossbar network)



68

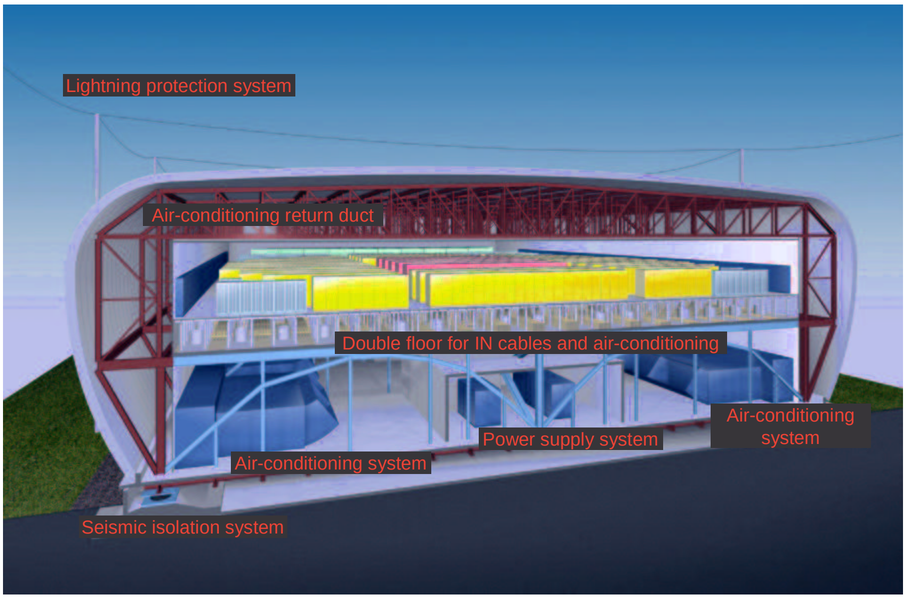
Earth Simulator Research and Development Center

Bird's-eye View of the Earth Simulator System

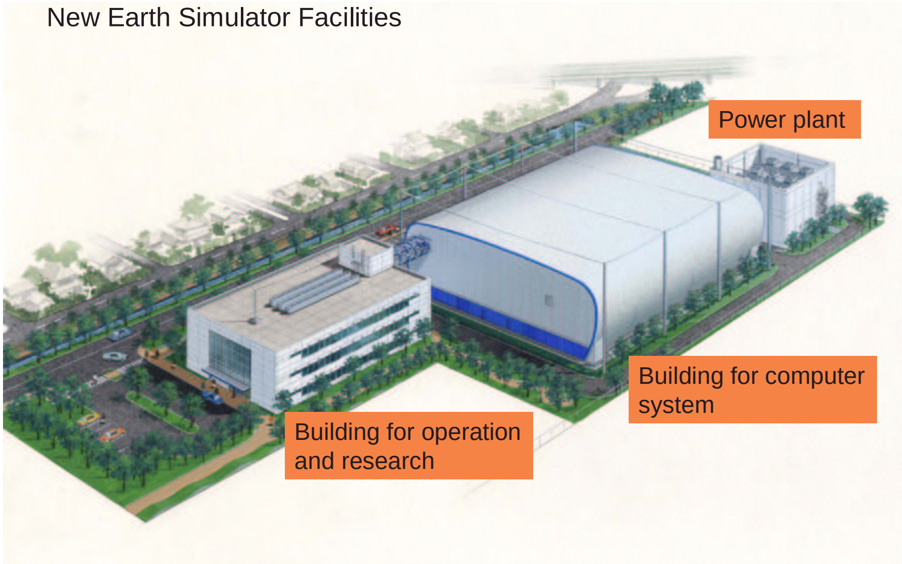


69

Cross-sectional View of the Earth Simulator Building



New Earth Simulator Facilities



Wiring of interconnection network cables

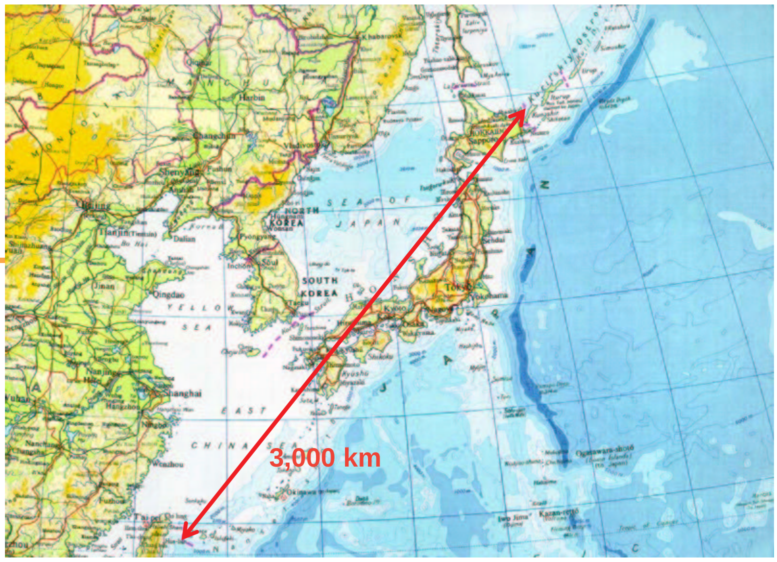
Earth Simulator Research and Development Center

Cables

73

R&D results

Total length of IN cables



74

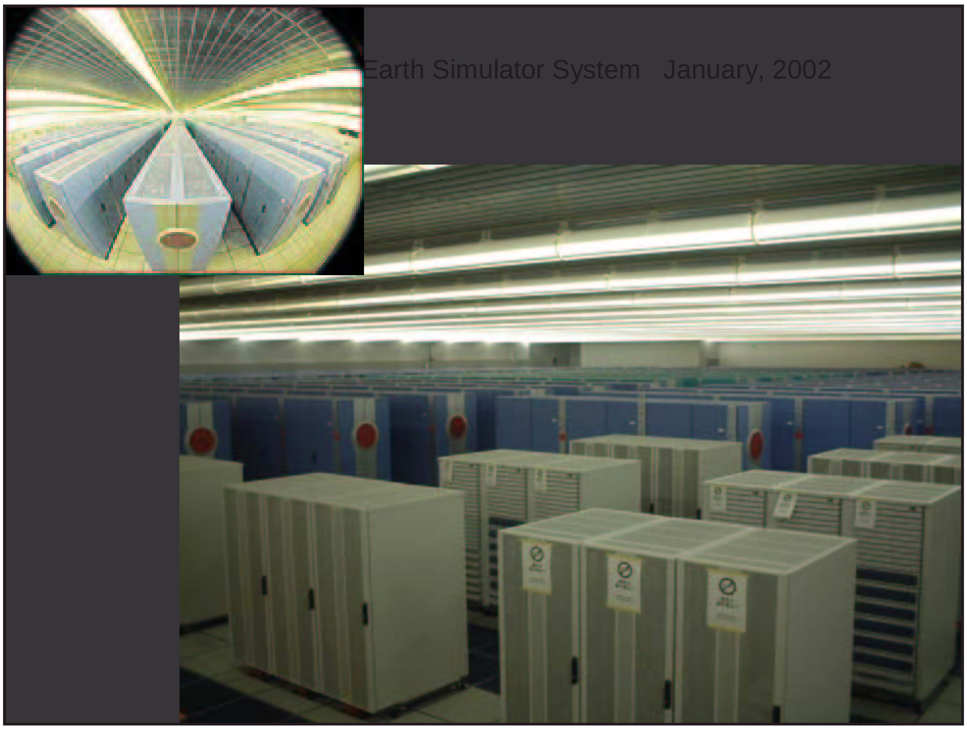
Wiring of interconnection network cables

Earth Simulator Research and Development Center

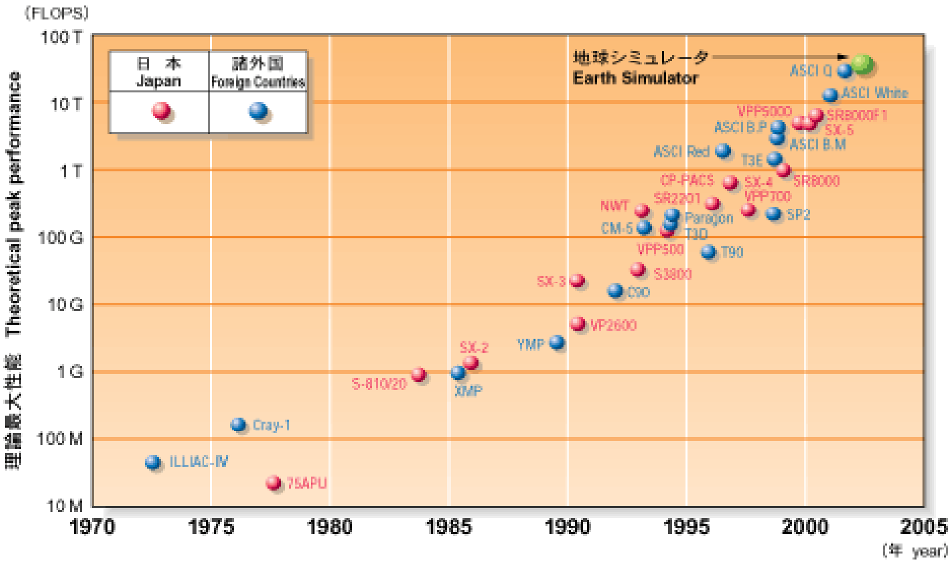
Processor Cabinets

76

Earth Simulator System January, 2002

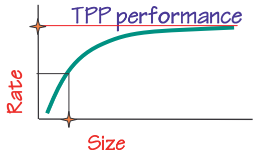


Peak Performance



Earth Simulator Computer (ESC)

- **Rmax from LINPACK MPP Benchmark** *Ax=b, dense problem*
  - Linpack Benchmark = 35.6 TFlop/s
  - Problem of size n = 1,041,216; (8.7 TB of memory)
  - Half of peak ($n_{\frac{1}{2}}$) achieved at $n_{\frac{1}{2}}$ = 265,408
  - Benchmark took 5.8 hours to run.
  - Algorithm: LU w/partial pivoting
  - Software: for the most part Fortran using MPI



- **For the Top500**
  - Σ of all the DOE computers = 27.5 TFlop/s
  - Performance of ESC ~ 1/6 Σ(Top 500 Computers)
  - Performance of ESC > Σ(Top 12 Computers)
  - Performance of ESC > Σ(Top 15 Computers in the US)
  - Performance of ESC > All the DOE and DOD machines (37.2 TFlop/s)
  - Performance of ESC >> the 3 NSF Center's computers (7.5 TFlop/s)

SETI@home ~ 27 TFlop/s

79

Year	Computer	Measured Gflop/s	Factor Δ from Pervious Year	Theoretical Peak Gflop/s	Factor Δ from Pervious Year	Number of Processors	Size of Problem
2002	Earth Simulator Computer, NEC	35610	4.9	40832	3.7	5104	1041216
2001	ASCI White-Pacific, IBM SP Power 3	7226	1.5	11136	1.0	7424	518096
2000	ASCI White-Pacific, IBM SP Power 3	4938	2.1	11136	3.5	7424	430000
1999	ASCI Red Intel Pentium II Xeon core	2379	1.1	3207	0.8	9632	362880
1998	ASCI Blue-Pacific SST, IBM SP 604E	2144	1.6	3868	2.1	5808	431344
1997	Intel ASCI Option Red (200 MHz Pentium Pro)	1338	3.6	1830	3.0	9152	235000
1996	Hitachi CP-PACS	368.2	1.3	614	1.8	2048	103680
1995	Intel Paragon XP/S MP	281.1	1	338	1.0	6768	128600
1994	Intel Paragon XP/S MP	281.1	2.3	338	1.4	6768	128600
1993	Fujitsu NWT	124.5		236		140	31920

80

LINPACK Benchmark List

Computer (Full Precision)		Number of Processors	R_{max} Gflop/s	N_{max} order	$N_{1/2}$ order	R_{peak} Gflop/s
☆Earth Simulator, NEC processors****	esc	5104	35610	1041216	265408	40832
ASCI White-Pacific, IBM SP Power 3(375 MHz)	llnl	8000	7226	518096	179000	12000
☆Compaq AlphaServer SC ES45/EV68 1GHz	psc	3016	4463	280000	85000	6032
Compaq AlphaServer SC ES45/EV68 1GHz	psc	3024	4059	525000	105000	6048
☆Compaq AlphaServer SC ES45/EV68 1GHz	cea	2560	3980	360000	85000	5120
IBM SP Power3 208 nodes 375 MHz	lbnl	3328	3052.	371712		4992
☆Compaq Alphaserver SC ES45/EV68 1GHz	lanl	2048	2916	272000		4096
☆IBM SP Power3 158 nodes 375 MHz	lbnl	2528	2526.	371712	102400	3792
ASCI Red Intel Pentium II Xeon core 333MHz	snl	9632	2379.6	362880	75400	3207
ASCI Blue-Pacific SST, IBM SP 604E(332 MHz)	llnl	5808	2144.	431344	432344	3868
ASCI Red Intel Pentium II Xeon core 333MHz	snl	9472	2121.3	251904	66000	3154
Compaq Alphaserver SC ES45/EV68 1GHz	lanl	1520	2096	390000	71000	3040
☆IBM SP 112 nodes (375 MHz POWER3 High)	ibm	1792	1791	275000	275000	2688
HITACHI SR8000/MPP/1152(450MHz)	u toyko	1152	1709.1	141000	16000	2074
☆HITACHI SR8000-F1/168(375MHz)	leibniz	168	1653.	160000	19560	2016
ASCI Red Intel Pentium II Xeon core 333Mhz	snl	6720	1633.3	306720	52500	2238
SGI ASCI Blue Mountain	lanl	5040	1608.	374400	138000	2520
IBM SP 328 nodes (375 MHz POWER3 Thin)	noo	1312	1417.	374000	374000	1968
Intel ASCI Option Red (200 MHz Pentium Pro)	snl	9152	1338.	235000	63000	1830
NEC SX-5/128M8(3.2ns)	osaka	128	1192.0	129536	10240	1280
CRAY T3E-1200 (600 MHz)	us government	1488	1127.	148800	28272	1786
HITACHI SR8000-F1/112(375MHz)	leibniz	112	1035.0	120000	15160	1344

Performance of AFES Climate Code

82

Physics Model of AFES

Cumulous convection	Condensation, precipitation, convection - Simplified Arakawa-Schubert (Arakawa and Schubert, 1974;Moorthi & Suarez, 1992) - Kuo scheme + shallow convection - Manabe's moist convection
Large-scale condensation	Other cloud processes and prediction of cloud water (Le Treut & Li, 1990)
Radiation	2-stream k-distribution scheme (Nakajima&Tanaka, 1986)
Vertical diffusion	Transport of heat, momentum, and moisture in PBL Level 2 turbulence scheme (Mellor & Yamada, 1974,1982)
Surface flux	Fluxes in surface boundary layer (Louis, 1979) (Mellor et al., 1992)
Ground process	Multi-layer heat conduction, Hydrology (Manabe, 1979) Ground moisture (Manabe et al., 1965) Frozen soil process (Clapp & Hornberger, 1978) Bucket model (Kondo, 1993)
Ocean mixing layer	Ocean temperature (Wilson et al, 1987) Sea ice
Gravity wave-induced drag	Orographic effect (McFarlane, 1987)
Others	Dry convection adjustment

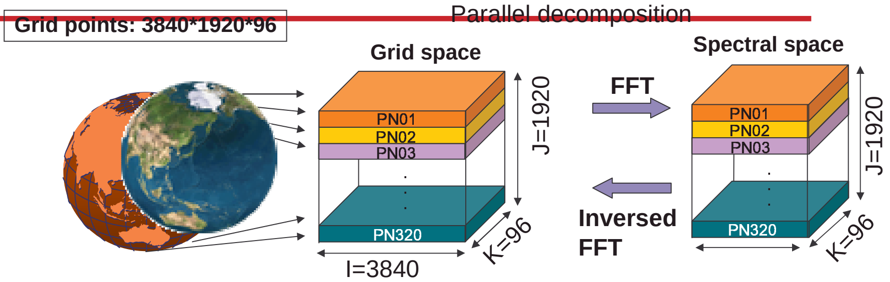
83

Parallelization of AFES

- MPI (Top-down approach) --> among processor nodes
  - Domain decomposition w.r.t. latitude in grid space (S.Pole to N.Pole)
  - Decomposition w.r.t. wave number of Fourier transform in wave domain
- Microtasking (Bottom-up approach) --> within node
  - Parallel decomposition of collapsed DO-loop to maximize the length of vector loop
  - Parallelism
    - Vertical direction for Legendre transform
    - Column-wise (2-dimensional) for physical process
- Vectorization (Bottom-up approach) --> with 1PE
  - Optimization of vector loop
  - Maximization of loop length with DO-loop collapse

84

Optimization Strategies for AFES Climate Model



- High resolution (10km) resulting in increased cost concentration on vector-tailored dynamics part (>75%)
- MPI among nodes / Microtasking within node
- Domain decomposition that fully exploits parallel nodes (>99% parallelization ratio) with less communication
- Reduced load imbalance due to improved algorithms (e.g., Use of increasingly popular Kuo cloud physics model)
- Improved vector performance with DO-loop optimization
- Combined use of assembler coding for part of matrix operations

85

Strategy for Performance Enhancement for the ES

- **Minimization of serial sections**
  - Most dominant factors affecting the total performance of applications
- **Pursuit of reduced communication overhead**
- **Increase of vector performance**
  - Effective combination of vector and parallel processing efficiency

86

Simulator
T1279L96 (3840x1920x96, 10.4km)

Total CPUs	Nodes	CPUs /Node	Elapse time (sec)	TFLOPS		
				Peak	Sustained	Ratio(%)
80	80	1	238.04	0.64	0.52	81.1
160	160	1	119.26	1.28	1.04	81.0
320	320	1	60.52	2.56	2.04	79.8
640	80	8	32.06	5.12	3.86	75.3
1280	160	8	16.24	10.24	7.61	74.3
2560	320	8	8.52	20.48	14.50	70.8

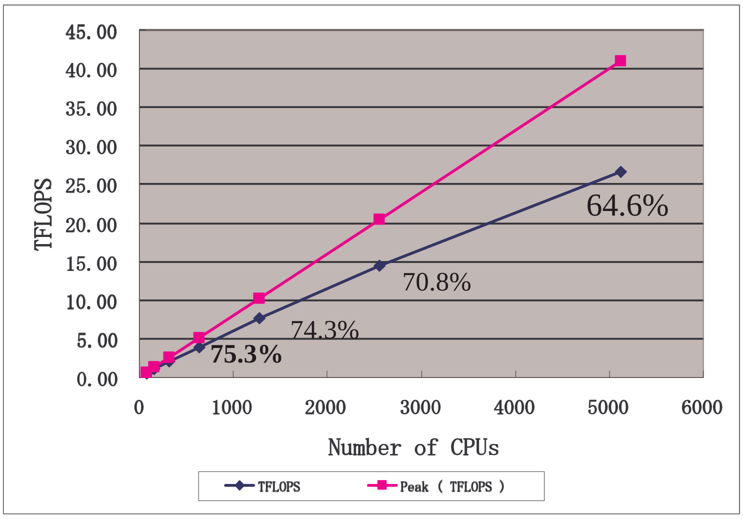
Measurement for 10 time integration steps

26.6 TFLOP/S sustained performance
with the 640 full nodes (5120 CPUs/ Peak 40 TFLOP/S)

87

Effective Performance of the AFES Climate Code on the ES with the Kuo's Cumulus Convection Scheme for a T1279L96 Resolution Model

26.6 TFLOP/S sustained performance
with the 640 full nodes (5120 CPUs/ peak 40 TFLOP/S)

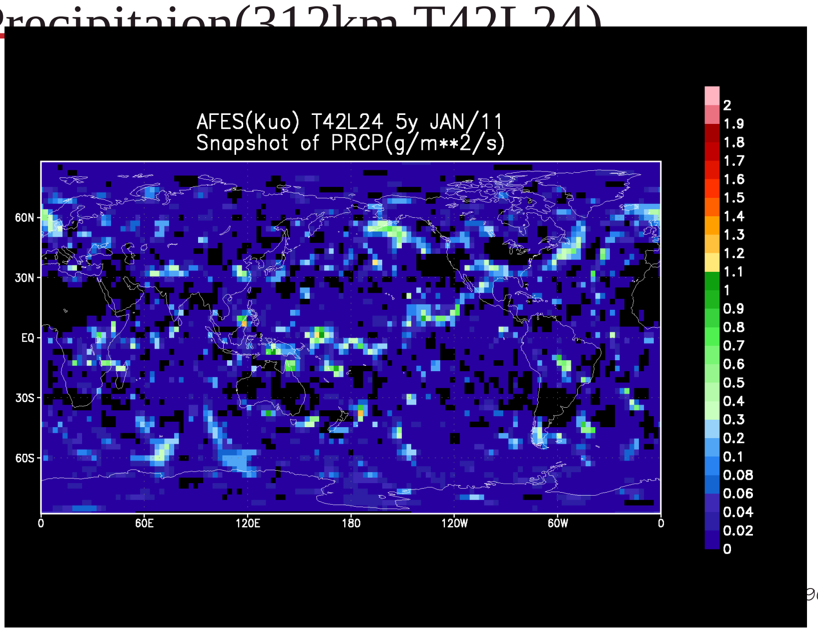


88

Results from AFES

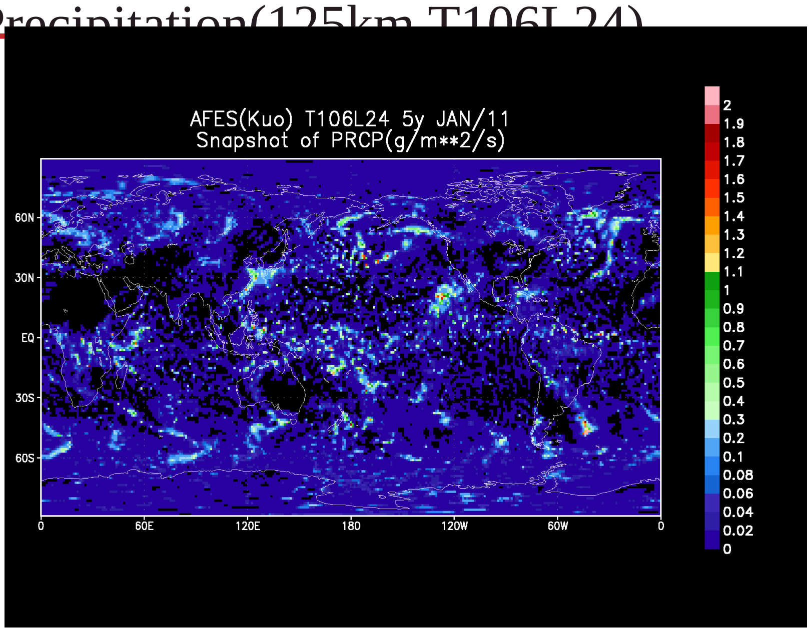
89

Precipitaion(312km T42L24)



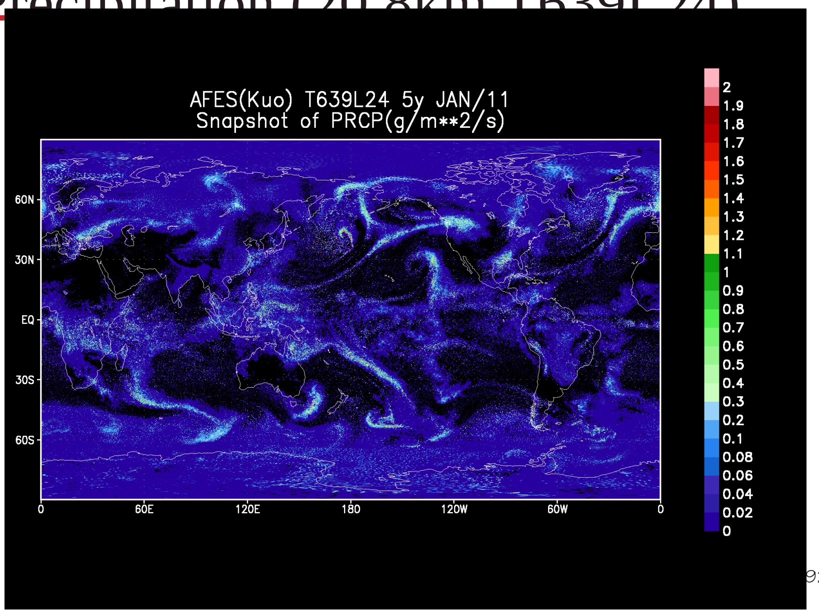
90

Precipitation(125km T106L24)



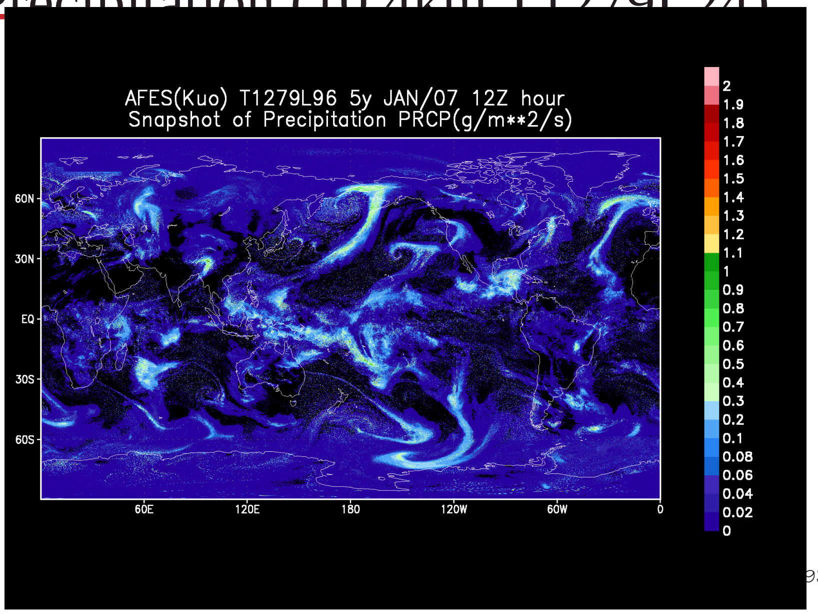
91

Precipitation (20.8km T639L24)



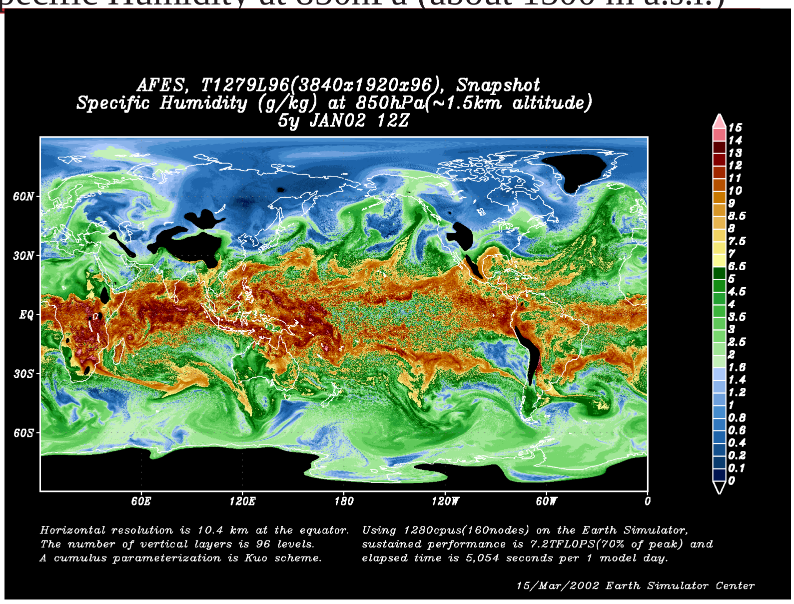
92

Precipitation (10.4km T1279L24)



93

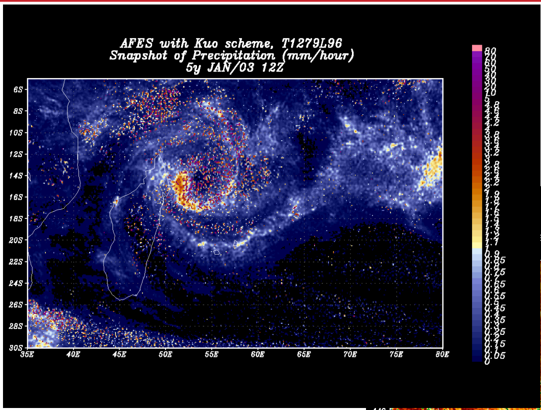
Specific Humidity at 850hPa (about 1500 m a.s.l.)

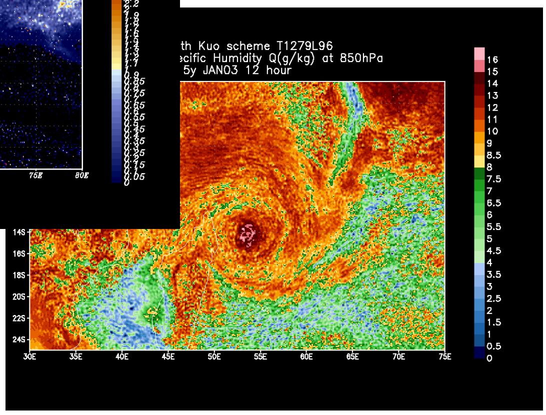


94

Cyclones around the Madagascar Islands

Specific humidity

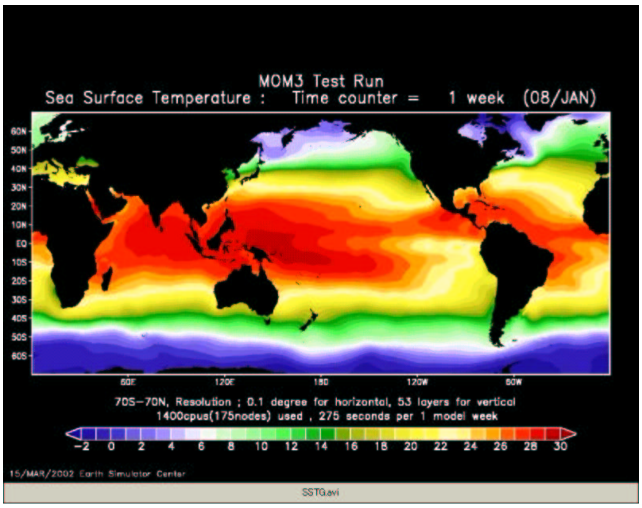




Precipitation

Seasonal Variation of Sea Surface Temperature

10 k m resolution for oceans (previously 100km)



MOM3 Oceanic Model of GFDL/Princeton Univ.

96

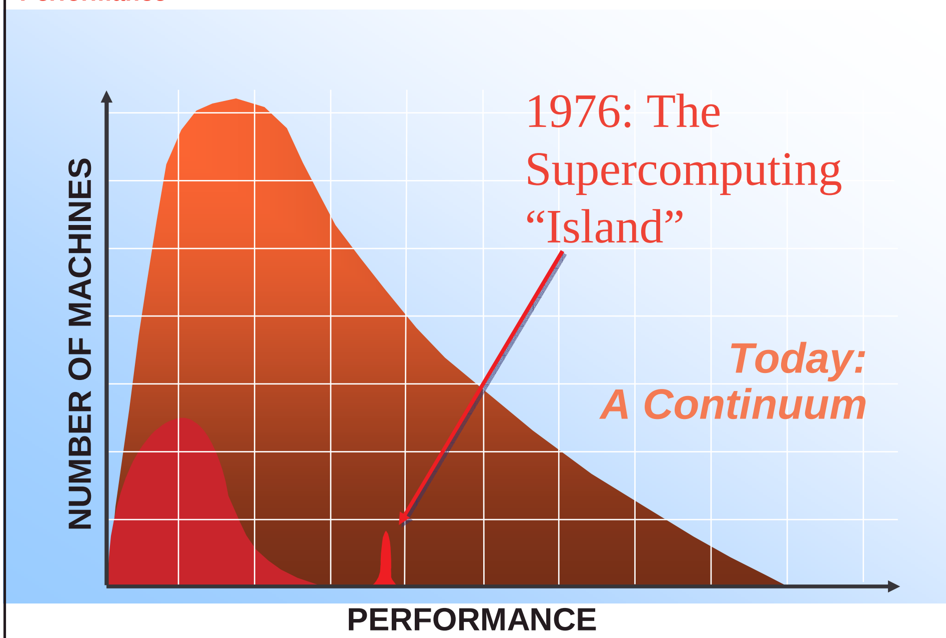
Distributed and Parallel Systems

- Gather (unused) resources
- Steal cycles
- System SW manages resources
- System SW adds value
- 10% - 20% overhead is OK
- Resources drive applications
- Time to completion is not critical
- Time-shared
- SETI@home
  - ~ 400,000 machines
  - Averaging 27 Tflop/s

- Bounded set of resources
- Apps grow to consume all cycles
- Application manages resources
- System SW gets in the way
- 5% overhead is maximum
- Apps drive purchase of equipment
- Real-time constraints
- Space-shared
- ASCI White LLNL
  - 8000 processors
  - Averaging 7.2 Tflop/s

97

Performance



The Future of HPC

- Great excitement in the area of High Performance Computing
- The expense of being different is being replaced by the economics of being the same
- HPC needs to lose its "special purpose" tag
- Still has to bring about the promise of scalable general purpose computing ...
- ... but it is dangerous to ignore this technology
- Final success when MPP technology is embedded in desktop computing
- Yesterday's HPC is today's mainframe is tomorrow's workstation

99

Highly Parallel Supercomputing: Where Are We?

- Performance:
  - Sustained performance has dramatically increased during the last year.
  - On most applications, sustained performance per dollar now exceeds that of conventional supercomputers. But...
  - Conventional systems are still faster on some applications.
- Languages and compilers:
  - Standardized, portable, high-level languages such as HPF, PVM and MPI are available. But ...
  - Initial HPF releases are not very efficient.
  - Message passing programming is tedious and hard to debug.
  - Programming difficulty remains a major obstacle to usage by mainstream scientist.

100

Highly Parallel Supercomputing: Where Are We?

- **Operating systems:**
  - Robustness and reliability are improving.
  - New system management tools improve system utilization. But...
  - Reliability still not as good as conventional systems.
- **I/O subsystems:**
  - New RAID disks, HiPPI interfaces, etc. provide substantially improved I/O performance. But...
  - I/O remains a bottleneck on some systems.

101

The Importance of Standards - Software

- Writing programs for MPP is hard ...
- But ... one-off efforts if written in a standard language
- Past lack of parallel programming standards ...
  - ... has restricted uptake of technology (to "enthusiasts")
  - ... reduced portability (over a range of current architectures and between future generations)
- Now standards exist: (MPI, OpenMP, PVM, & HPF), which ...
  - ... allows users & manufacturers to protect software investment
  - ... encourage growth of a "third party" parallel software industry & parallel versions of widely used codes

102

The Importance of Standards - Hardware

- Processors
  - commodity RISC processors
- Interconnects
  - high bandwidth, low latency communications protocol
  - no de-facto standard yet (ATM, Fibre Channel, HPPI, FDDI)
- Growing demand for total solution:
  - robust hardware + usable software
- HPC systems containing all the programming tools / environments / languages / libraries / applications packages found on desktops

103

Achieving TeraFlops

- In 1991 we had, 1 Gflop/s
- Today, 1000 fold increase
  - Architecture
    - exploiting parallelism
  - Processor, communication, memory
    - Moore's Law
  - Algorithm improvements
    - block-partitioned algorithms

104

Future: Petaflops (10^{15} fl pt ops/s)

Today $\approx \sqrt{10^{15}}$ flops for our workstations

- A Pflop for 1 second ≈ a typical workstation computing for 1 year.
- From an algorithmic standpoint
  - concurrency
  - data locality
  - latency & sync
  - floating point accuracy
  - dynamic redistribution of workload
  - new language and constructs
  - role of numerical libraries
  - algorithm adaptation to hardware failure

105

A Petaflops Computer System

- 1 Pflop/s sustained computing
- Between 10,000 and 1,000,000 processors
- Between 10 TB and 1PB main memory
- Commensurate I/O bandwidth, mass store, etc.
- If built today, cost $40 B and consume 1 TWatt.
- May be feasible and "affordable" by the year 2010

106