

CS 594 Spring 2003

Lecture 2:

Top500

Jack Dongarra
Computer Science Department
University of Tennessee

High Performance Computers

- ◆ From the beginning of the digital age, supercomputers have been time machines that let researchers peer into the future, both intellectually and temporally.
 - Intellectually bring to life models of complex phenomena when economics and other constraints preclude experimentation.
 - Temporally, they reduce the time to solution by enabling us to evaluate larger and more complex models than would be possible on more conventional systems.

High Performance Computers

- ◆ ~ 25 years ago
 - 1×10^6 Floating Point Ops/sec (Mflop/s)
 - » Scalar based
- ◆ ~ 10 years ago
 - 1×10^9 Floating Point Ops/sec (Gflop/s)
 - » Vector & Shared memory computing, bandwidth aware
 - » Block partitioned, latency tolerant
- ◆ ~ Today
 - 1×10^{12} Floating Point Ops/sec (Tflop/s)
 - » Highly parallel, distributed processing, message passing, network based
 - » data decomposition, communication/computation
- ◆ ~ 5 years away
 - 1×10^{15} Floating Point Ops/sec (Pflop/s)
 - » Many more levels MH, combination/grids&HPC
 - » More adaptive, LT and bandwidth aware, fault tolerant, extended precision, attention to SMP nodes

Top 500 Computers

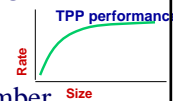
- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP

$$Ax=b, \text{ dense problem}$$

Updated twice a year

SC'xy in the States in November

Meeting in Germany in June

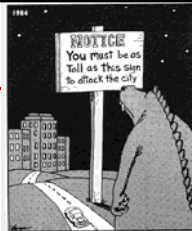


What is a Supercomputer?

- ◆ 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
- ◆ 2005:
 - #1 = 70 TFlop/s
 - #500 = 850 GFlop/s



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.

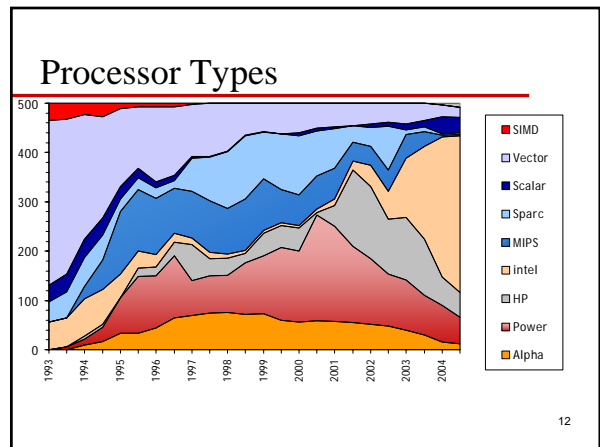
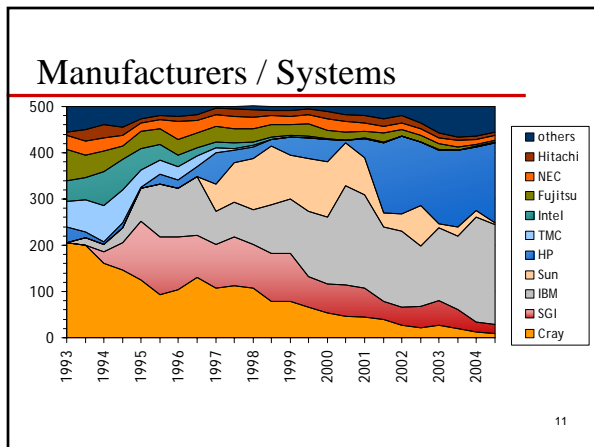
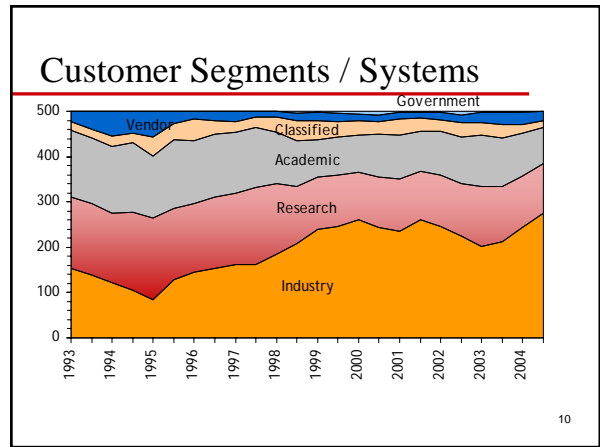
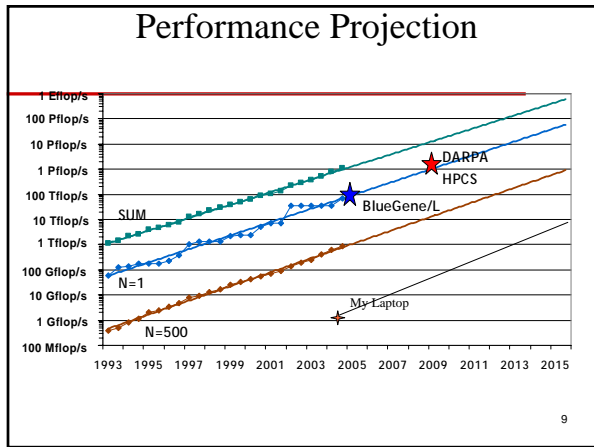
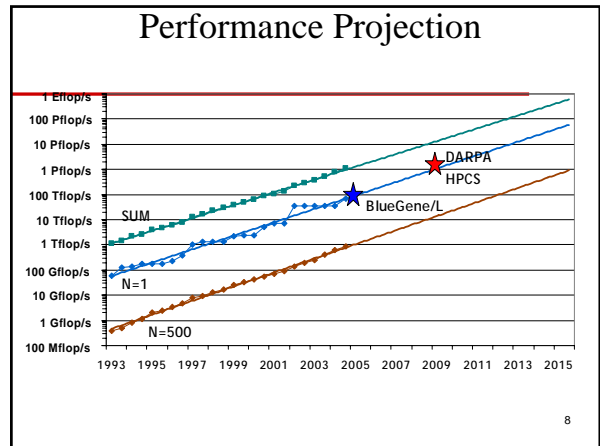
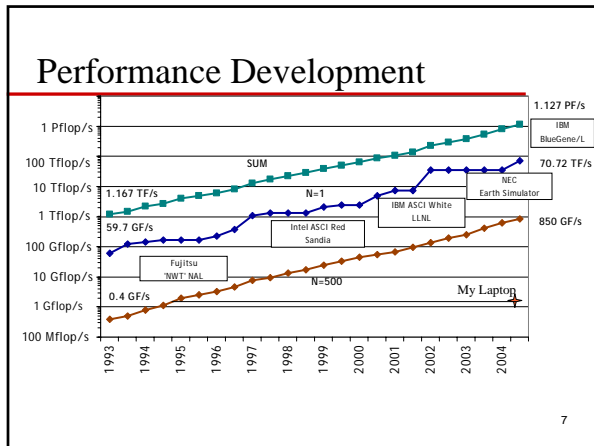
Computational fluid dynamics, protein folding, climate modeling, national security, in particular for cryptanalysis and for simulating nuclear weapons to name a few.

24th List: The TOP10

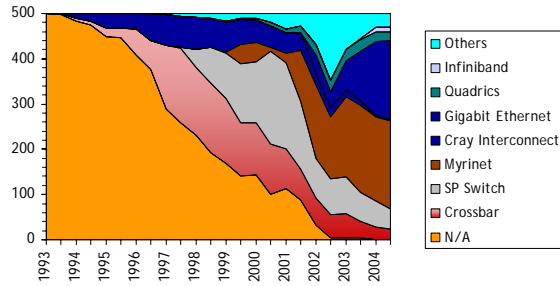


	Manufacturer	Computer	Rmax (TF/s)	Installation Site	Country	Year	#Proc
1	IBM	BlueGene/L β-System	70.72	DOE/IBM	USA	2004	32768
2	SGI	Columbia Athix, Infiniband	51.87	NASA Ames	USA	2004	10160
3	NEC	Earth-Simulator	35.86	Earth Simulator Center	Japan	2002	5120
4	IBM	MareNostrum BladeCenter JS20, Myrinet	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	UI C/U, NCSA	USA	2003	2500

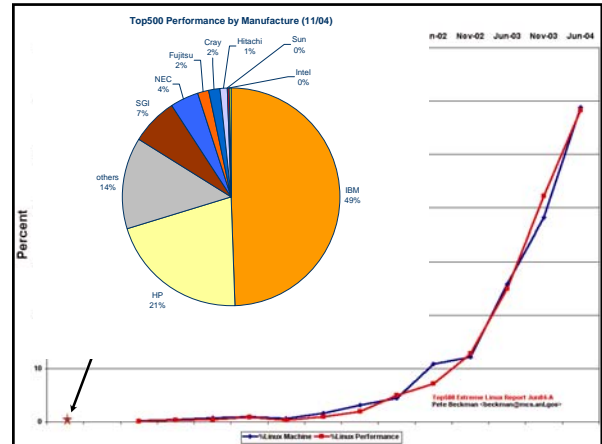
ORNL #29 Cray X1, 5.8 Tflop/s, 504 procs
399 system > 1 Tflop/s; 294 machines are clusters, top10 average 8K proc



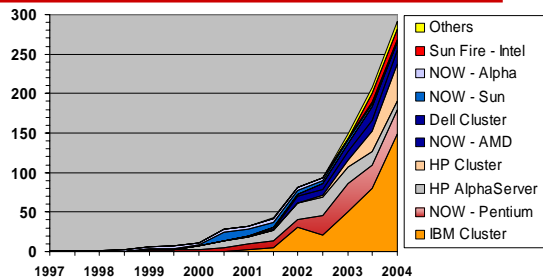
Interconnects / Systems



13



Clusters (NOW) / Systems



15

Top500 Conclusions

- ◆ Microprocessor based supercomputers have brought a major change in accessibility and affordability.
- ◆ MPPs continue to account of more than half of all installed high-performance computers worldwide.

16

Performance Numbers on RISC Processors

Processor	Cycle Time	Linpack n=100	Linpack n=1000	Peak
Intel P4	3600	1821 (25%)	4220 (59%)	7200
Intel/HP Itanium 2	1600	1765 (28%)	5953 (93%)	6400
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
NEC SX-8	2000	2177 (14%)	14960 (93%)	16000
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 J	80	27 (17%)	110 (69%)	160

High-Performance Computing Directions: Beowulf-class PC Clusters

Definition:

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

Advantages:

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

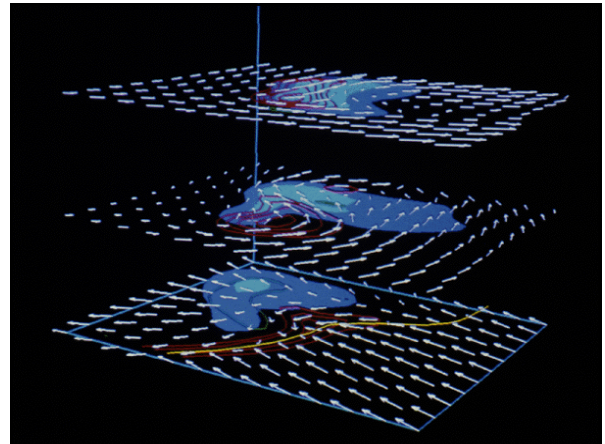
Enabled by PC hardware, networks and operating system achieving capabilities of scientific workstations at a fraction of the cost and availability of industry standard message passing libraries. However, much more of a contact sport.¹⁸

Virtual Environments

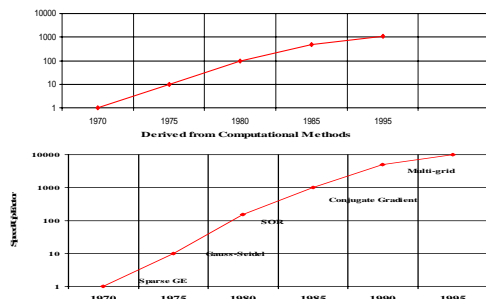
0.32E-08 0.00E+00 0.00E+00 0.00E+00 0.38E-04 0.13E-05 0.22E-05 0.33E-05 0.59E-05 0.11E-04
 0.18E-04 0.23E-04 0.23E-04 0.21E-04 0.67E-04 0.38E-03 0.90E-03 0.18E-02 0.30E-02 0.43E-02
 0.50E-02 0.51E-02 0.49E-02 0.44E-02 0.39E-02 0.35E-02 0.31E-02 0.28E-02 0.27E-02 0.26E-02
 0.24E-02 0.27E-02 0.28E-02 0.30E-02 0.33E-02 0.34E-02 0.38E-02 0.39E-02 0.39E-02 0.38E-02
 0.34E-02 0.30E-02 0.27E-02 0.24E-02 0.21E-02 0.18E-02 0.16E-02 0.14E-02 0.11E-02 0.96E-03
 0.79E-03 0.63E-03 0.48E-03 0.35E-03 0.24E-03 0.15E-03 0.80E-04 0.34E-04 0.89E-05 0.16E-05
 0.18E-04 0.34E-04 0.00E+00 0.00E+00 0.00E+00 0.00E+00 0.00E+00 0.00E+00 0.00E+00 0.00E+00
 0.00E+00 0.00E+00 0.00E+00 0.00E+00 0.24E-08 0.00E+00 0.00E+00 0.00E+00 0.39E-06 0.11E-05
 0.19E-05 0.30E-05 0.53E-05 0.96E-05 0.15E-04 0.20E-04 0.20E-04 0.18E-04 0.27E-04 0.23E-03
 0.65E-03 0.14E-02 0.27E-02 0.40E-02 0.49E-02 0.51E-02 0.49E-02 0.45E-02 0.40E-02 0.35E-02
 0.31E-02 0.28E-02 0.27E-02 0.26E-02 0.26E-02 0.27E-02 0.28E-02 0.30E-02 0.33E-02 0.36E-02
 0.38E-02 0.39E-02 0.39E-02 0.37E-02 0.34E-02 0.30E-02 0.27E-02 0.24E-02 0.21E-02 0.18E-02
 0.16E-02 0.14E-02 0.12E-02 0.98E-03 0.81E-03 0.65E-03 0.51E-03 0.38E-03 0.27E-03 0.17E-03
 0.99E-04 0.47E-04 0.16E-04 0.36E-05 0.62E-06 0.41E-07 0.75E-10 0.00E+00 0.00E+00 0.00E+00
 0.00E+00 0.00E+00 0.00E+00 0.00E+00 0.00E+00 0.00E+00 0.00E+00 0.00E+00 0.15E-08 0.00E+00
 0.00E+00 0.00E+00 0.19E-06 0.84E-04 0.16E-05 0.27E-05 0.47E-05 0.82E-05 0.13E-04 0.17E-04
 0.17E-04 0.15E-04 0.16E-04 0.10E-03 0.41E-03 0.11E-02 0.23E-02 0.37E-02 0.48E-02 0.51E-02
 0.49E-02 0.45E-02 0.40E-02 0.35E-02 0.31E-02 0.28E-02 0.27E-02 0.26E-02 0.26E-02 0.27E-02
 0.28E-02 0.31E-02 0.33E-02 0.36E-02 0.38E-02 0.39E-02 0.38E-02 0.36E-02 0.33E-02 0.29E-02

Do they make any sense?

19



Performance Improvements for Scientific Computing Problems



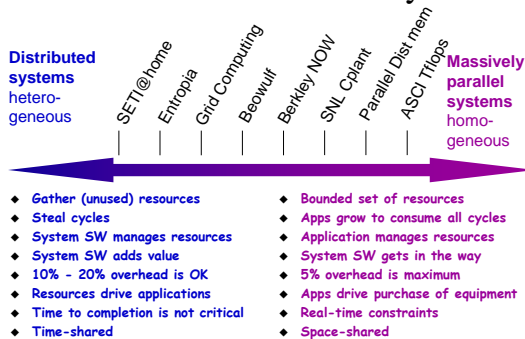
21

Different Architectures

- ◆ **Parallel computing:** single systems with many processors working on same problem
- ◆ **Distributed computing:** many systems loosely coupled by a scheduler to work on related problems
- ◆ **Grid Computing:** many systems tightly coupled by software, perhaps geographically distributed, to work together on single problems or on related problems

22

Distributed and Parallel Systems

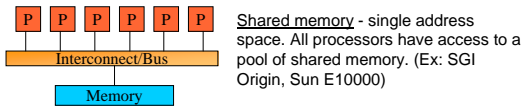


24

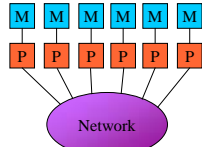
Types of Parallel Computers

- ◆ The simplest and most useful way to classify modern parallel computers is by their memory model:
 - shared memory
 - distributed memory

Shared vs. Distributed Memory

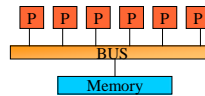


Distributed memory - each processor has its own local memory. Must do message passing to exchange data between processors. (Ex: CRAY T3E, IBM SP, clusters)



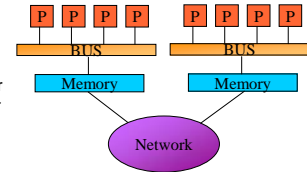
25

Shared Memory: UMA vs. NUMA



Uniform memory access (UMA): Each processor has uniform access to memory. Also known as **symmetric multiprocessors** (Sun E10000)

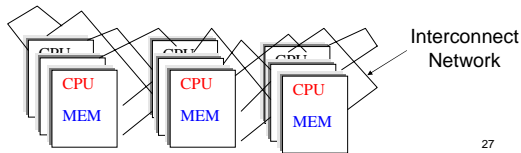
Non-uniform memory access (NUMA): Time for memory access depends on location of data. Local access is faster than non-local access. Easier to scale than SMPs (SGI Origin)



26

Distributed Memory: MPPs vs. Clusters

- ◆ **Processors-memory nodes are connected by some type of interconnect network**
 - **Massively Parallel Processor (MPP):** tightly integrated, single system image.
 - **Cluster:** individual computers connected by switch



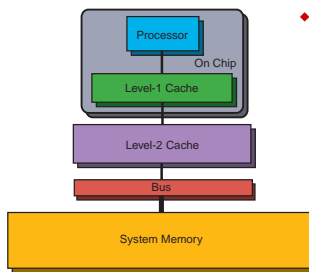
27

Processors, Memory, & Networks

- ◆ **Both shared and distributed memory systems have:**
 1. **processors:** now generally commodity RISC processors
 2. **memory:** now generally commodity DRAM
 3. **network/interconnect:** between the processors and memory (bus, crossbar, fat tree, torus, hypercube, etc.)

28

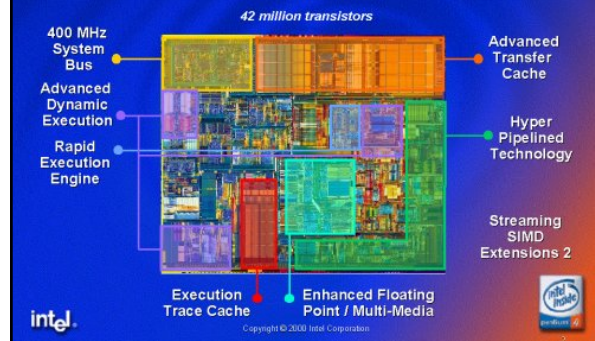
Standard Uniprocessor Memory Hierarchy



- ◆ **Intel Pentium 4 3.2 - 3.46 GHz processor**
 - 130 nm technology
 - 12 Kbytes of 4 way assoc. L1 instruction cache with 32 byte lines.
 - 8 Kbytes of 4 way assoc. L1 data cache with 32 byte lines.
 - 512 Kbytes of 8 way assoc. L2 cache 32 byte lines.
 - 2 MB L3 cache
 - 400 MB/s bus speed
 - 3.2 GB/s
 - SSE2 provide peak of 6.4-6.92 Gflop/s

29

The Intel® Pentium® 4 Processor Takes A Leap Forward, Delivering...



Processor-Related Terms

Clock period (cp): the minimum time interval between successive actions in the processor. Fixed, depends on design of processor. Measured in nanoseconds (~1-5 for fastest processors). Inverse of frequency (MHz)

Instruction: an action executed by a processor, such as a mathematical operation or a memory operation.

Register: a small, extremely fast location for storing data or instructions in the processor

31

Processor-Related Terms

Functional Unit: a hardware element that performs an operation on an operand or pair of operations. Common FUs are ADD, MULT, INV, SQRT, etc.

Pipeline : technique enabling multiple instructions to be overlapped in execution

Superscalar: multiple instructions are possible per clock period

Flops: floating point operations per second

32

Processor-Related Terms

Cache: fast memory (SRAM) near the processor. Helps keep instructions and data close to functional units so processor can execute more instructions more rapidly.

TLB: Translation-Lookaside Buffer keeps addresses of pages (block of memory) in main memory that have recently been accessed (a cache for memory addresses)

33

Memory-Related Terms

SRAM: Static Random Access Memory (RAM). Very fast (~10 nanoseconds), made using the same kind of circuitry as the processors, so speed is comparable.

DRAM: Dynamic RAM. Longer access times (~100 nanoseconds), but hold more bits and are much less expensive (10x cheaper).

Memory hierarchy: the hierarchy of memory in a parallel system, from registers to cache to local memory to remote memory. More later.

34

Interconnect-Related Terms

◆ **Latency:** How long does it take to start sending a "message"? Measured in microseconds.

(Also in processors: How long does it take to output results of some operations, such as floating point add, divide etc., which are pipelined?)

◆ **Bandwidth:** What data rate can be sustained once the message is started? Measured in Mbytes/sec.

35

Interconnect-Related Terms

Topology: the manner in which the nodes are connected.

➤ Best choice would be a fully connected network (every processor to every other). Unfeasible for cost and scaling reasons.

➤ Instead, processors are arranged in some variation of a grid, torus, or hypercube.



3-d hypercube



2-d mesh



2-d torus

36

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.

37

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.

38

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: (PVM, MPI & HPF), which ...
 - ... allows users & manufacturers to protect software investment
 - ... encourage growth of a "third party" parallel software industry & parallel versions of widely used codes

39

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**

40

The Future of HPC

- ◆ 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

41

Achieving TeraFlops

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

42

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

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

- ◆ A P flop for 1 second \approx 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

43

Petaflop (10^{15} flop/s) Computers Within the 5 Years

- ◆ **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

44

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

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

- 1/2 Million PCs
- \$1B (\$2K each)
- 100 Mwatts
- 2.5 acres
- 1/2 Million Windows licenses!!
- PC failure every second

45