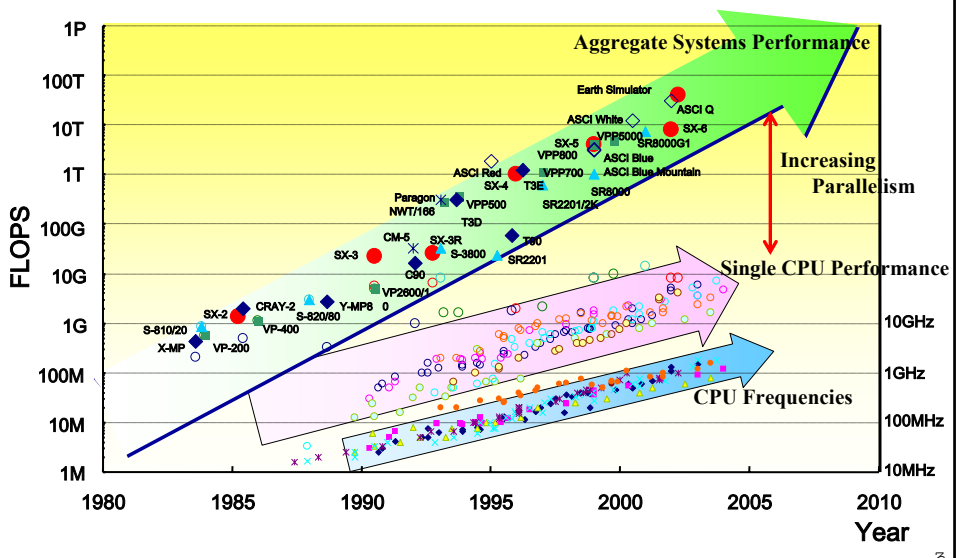
Survey of
"High Performance Machines"

Jack Dongarra
University of Tennessee
and
Oak Ridge National Laboratory

Overview

- **Processors**
- **Interconnect**
- **Look at the 3 Japanese HPCs**
- **Examine the Top131**

History of High Performance Computers



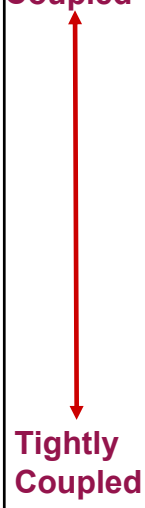
Vibrant Field for High Performance Computers

- Cray X1
- SGI Altix
- IBM Regatta
- Sun
- HP
- Bull
- Fujitsu PowerPower
- Hitachi SR11000
- NEC SX-7
- Apple

- Coming soon …
  - Cray RedStorm
  - Cray BlackWidow
  - NEC SX-8
  - IBM Blue Gene/L

JD1

Architecture/Systems Continuum



- Commodity processor with commodity interconnect
  - Clusters
    - Pentium, Itanium, Opteron, Alpha, PowerPC
    - GigE, Infiniband, Myrinet, Quadrics, SCI
  - NEC TX7
  - HP Alpha
  - Bull NovaScale 5160
- Commodity processor with custom interconnect
  - SGI Altix
    - Intel Itanium 2
  - Cray Red Storm
    - AMD Opteron
  - IBM Blue Gene/L (?)
    - IBM Power PC
- Custom processor with custom interconnect
  - Cray X1
  - NEC SX-7
  - IBM Regatta

Commodity Processors

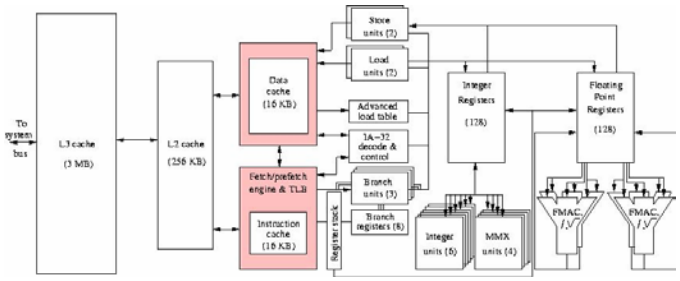
- AMD Opteron
  - 2 GHz, 4 Gflop/s peak
- HP Alpha EV68
  - 1.25 GHz, 2.5 Gflop/s peak
- HP PA RISC
- IBM PowerPC
  - 2 GHz, 8 Gflop/s peak
- Intel Itanium 2
  - 1.5 GHz, 6 Gflop/s peak
- Intel Pentium Xeon, Pentium EM64T
  - 3.2 GHz, 6.4 Gflop/s peak
- MIPS R16000
- Sun UltraSPARC IV

Slide 5

JD1 check bgl status

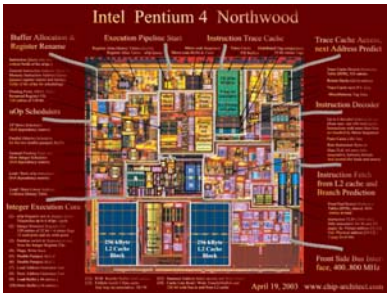
Jack Dongarra, 4/15/2004

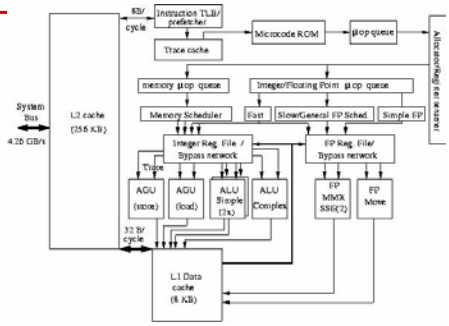
Itanium 2



- Floating point bypass for level 1 cache
- Bus is 128 bits wide and operates at 400 MHz, for 6.4 GB/s
- 4 flops/cycle
- 1.5 GHz Itanium 2
  - Linpack Numbers: (theoretical peak 6 Gflop/s)
    - 100: 1.7 Gflop/s
    - 1000: 5.4 Gflop/s

Pentium 4 IA32





- Processor of choice for clusters
- 1 flop/cycle
- Streaming SIMD Extensions 2 (SSE2): 2 Flops/cycle
- Intel Xeon 3.2 GHz 400/533 MHz bus, 64 bit wide(3.2/4.2 GB/s)
  - Linpack Numbers: (theorical peak 6.4 Gflop/s)
    - 100: 1.7 Gflop/s
    - 1000: 3.1 Gflop/s
- Coming Soon: "Pentium 4 EM64T"
  - 800 MHz bus 64 bit wide
  - 3.6 GHz, 2MB L2 Cache
    - Peak 7.2 Gflop/s using SSE2

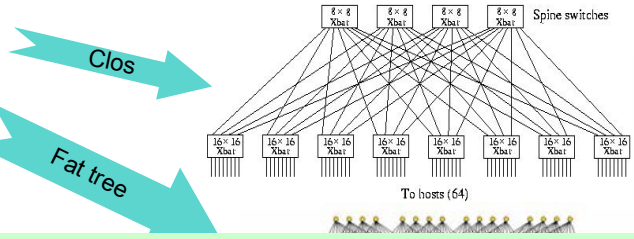
High Bandwidth vs Commodity Systems

- 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.

	Earth Simulator (NEC)	Cray X1 (Cray)	ASCI Q (HP EV68)	MCR (Dual Xeon)	VT Big Mac (Dual IBM PPC)
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
Bytes/flop (main memory)	4	2.6	0.8	0.44	0.5

Commodity Interconnects

- Gig Ethernet
- Myrinet
- Infiniband
- QsNet
- SCI

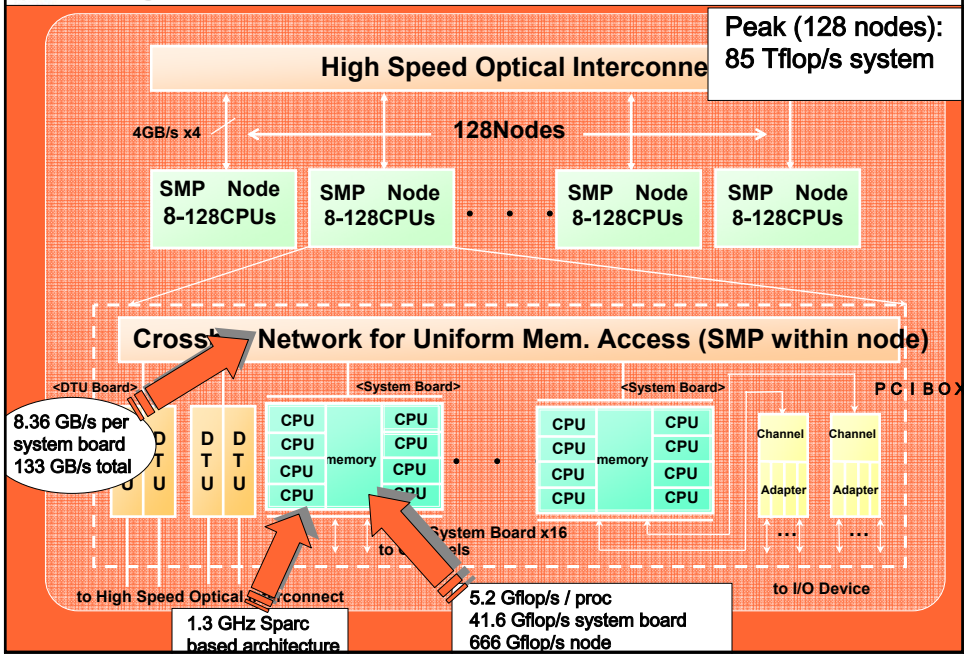


	Switch topology	$ NIC	$Sw/node	$ Node	Lt(us)/BW (MB/s) (MPI)
Gigabit Ethernet	Bus	$ 50	$ 50	$ 100	30 / 100
SCI	Torus	$1,600	$ 0	$1,600	5 / 300
QsNetII	Fat Tree	$1,200	$1,700	$2,900	3 / 880
Myrinet (D card)	Clos	$ 700	$ 400	$1,100	6.5/ 240
IB 4x	Fat Tree	$1,000	$ 400	$1,400	6 / 820

Quick Look at …

- Fujitsu PrimePower2500
- Hitachi SR11000
- NEC SX-7

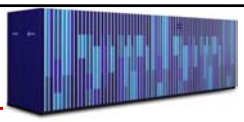
Fujitsu PRIMEPOWER HPC2500



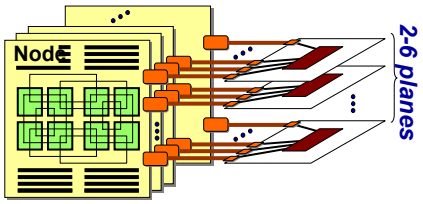
Latest Installation of FUJITSU HPC Systems

User Name	Configuration
Japan Aerospace Exploration Agency (JAXA)	PRIMEPOWER 128CPU x 14(Cabinets) (9.3 Tflop/s)
Japan Atomic Energy Research Institute (ITBL Computer System)	PRIMEPOWER 128CPU x 4 + 64CPU (3 Tflop/s)
Kyoto University	PRIMEPOWER 128CPU(1.5 GHz) x 11 + 64CPU (8.8 Tflop/s)
Kyoto University (Radio Science Center for Space and Atmosphere)	PRIMEPOWER 128CPU + 32CPU
Kyoto University (Grid System)	PRIMEPOWER 96CPU
Nagoya University (Grid System)	PRIMEPOWER 32CPU x 2
National Astronomical Observatory of Japan (SUBARU Telescope System)	PRIMEPOWER 128CPU x 2
Japan Nuclear Cycle Development Institute	PRIMEPOWER 128CPU x 3
Institute of Physical and Chemical Research (RIKEN)	IA-Cluster (Xeon 2048CPU) with InfiniBand & Myrinet
National Institute of Informatics (NAREGI System)	IA-Cluster (Xeon 256CPU) with InfiniBand PRIMEPOWER 64CPU
Tokyo University (The Institute of Medical Science)	IA-Cluster (Xeon 64CPU) with Myrinet PRIMEPOWER 26CPU x 2
Osaka University (Institute of Protein Research)	IA-Cluster (Xeon 160CPU) with InfiniBand

Hitachi SR11000



- Based on IBM Power 4+
- SMP with 16 processors/node
  - 109 Gflop/s / node(6.8 Gflop/s / p)
  - IBM uses 32 in their machine
- IBM Federation switch
  - Hitachi: 6 planes for 16 proc/node
  - IBM uses 8 planes for 32 proc/node
- Pseudo vector processing features
  - No hardware enhancements
    - Unlike the SR8000



- Hitachi's Compiler effort is separate from IBM
  - No plans for HPF
- 3 customers for the SR 11000,
  - 7 Tflop/s largest system 64 nodes
- National Institute for Material Science Tsukuba - 64 nodes (7 Tflop/s)
- Okasaki Institute for Molecular Science - 50 nodes (5.5 Tflops)
- Institute for Statistic Math Institute - 4 nodes

NEC SX-7/160M5

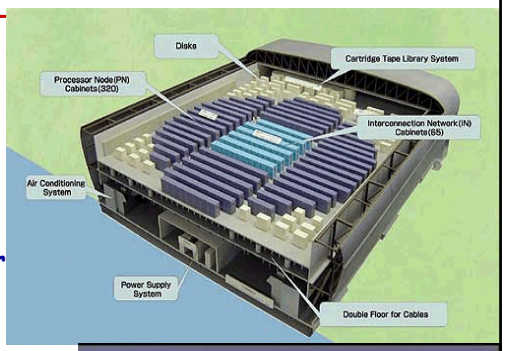
Total Memory	1280 GB
Peak performance	1412 Gflop/s
# nodes	5
# PE per 1 node	32
Memory per 1 node	256 GB
Peak performance per PE	8.83 Gflop/s
# vector pipe per 1PE	4
Data transport rate between nodes	8 GB/s

Rumors of SX-8
8 CPU/node
26 Gflop/s / proc

- SX-6: 8 proc/node
  - 8 GFlop/s, 16 GB
  - processor to memory
- SX-7: 32 proc/node
  - 8.825 GFlop/s, 256 GB,
  - processor to memory

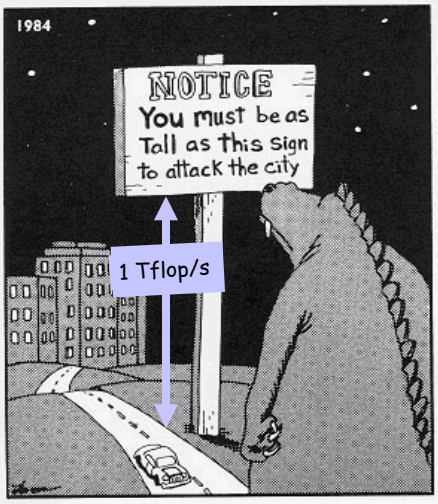
After 2 years, Still A Tour de Force in Engineering

- Homogeneous, Centralized, Proprietary, Expensive!
- Target Application: CFD-Weather, Climate, Earthquakes
- 640 NEC SX/6 Nodes (mod)
  - 5120 CPUs which have vector ops
  - Each CPU 8 Gflop/s Peak
- 40 TFlop/s (peak)
- H. Miyoshi; master mind & director
  - NAL, RIST, ES
  - Fujitsu AP, VP400, NWT, ES
- ~ 1/2 Billion $ for machine, software, & building
- Footprint of 4 tennis courts
- Expect to be on top of Top500 for at least another year!
- From the Top500 (November 2003)
  - Performance of ESC > Σ Next Top 3 Computers



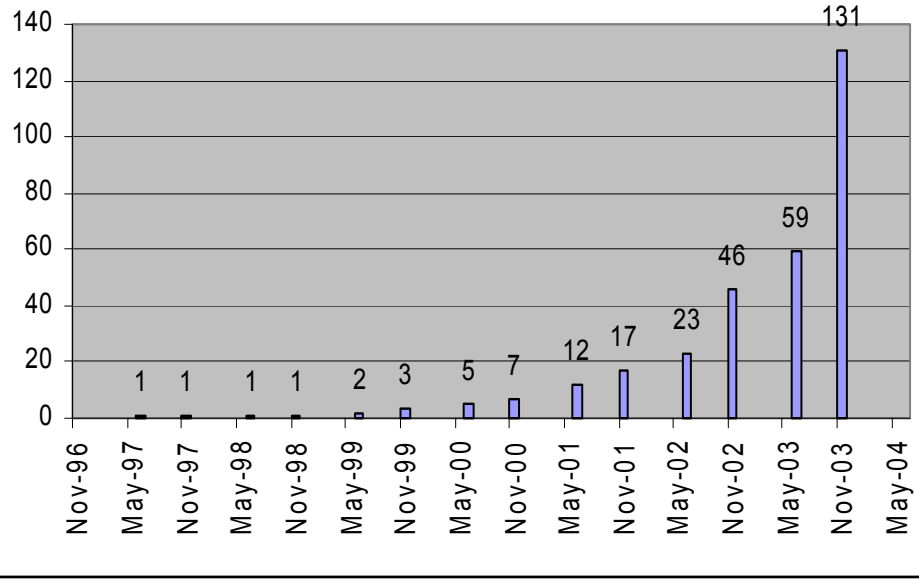
The Top131

- Focus on machines that are at least 1 TFlop/s on the Linpack benchmark
- Pros
  - One number
  - Simple to define and rank
  - Allows problem size to change with machine and over time
- Cons
  - Emphasizes only "peak" CPU speed and number of CPUs
  - Does not stress local bandwidth
  - Does not stress the network
  - Does not test gather/scatter
  - Ignores Amdahl's Law (Only does weak scaling)
  - …



- 1993:
  - #1 = 59.7 GFlop/s
  - #500 = 422 MFlop/s
- 2003:
  - #1 = 35.8 TFlop/s
  - #500 = 403 GFlop/s

Number of Systems on Top500 > 1 Tflop/s Over Time



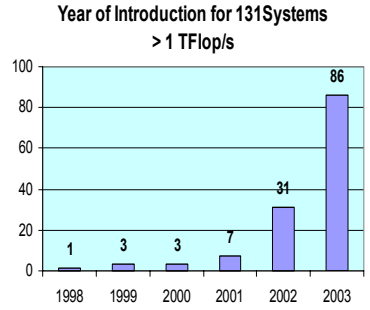
Factoids on Machines > 1 TFlop/s

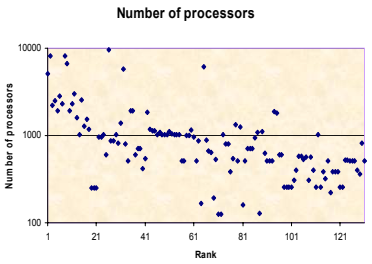
- 131 Systems
- 80 Clusters (61%)

- Average rate: 2.44 Tflop/s
- Median rate: 1.55 Tflop/s

- Sum of processors in Top131: 155,161
  - Sum for Top500: 267,789
- Average processor count: 1184
- Median processor count: 706

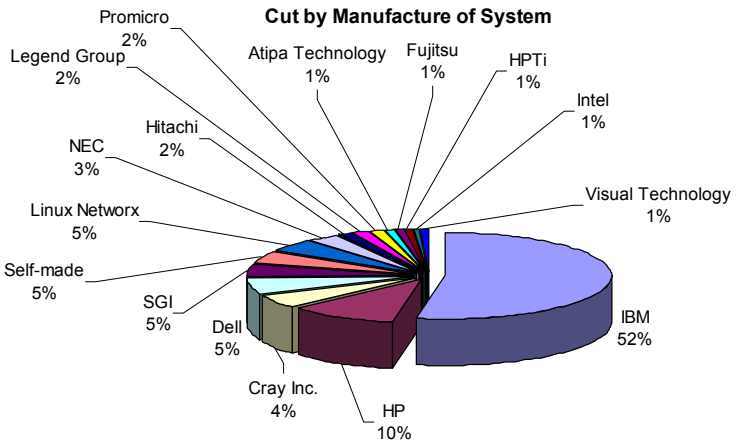
- Numbers of processors
  - Most number of processors: 9632_{26}
    - ASCI Red
  - Fewest number of processors: 124_{71}
    - Cray X1



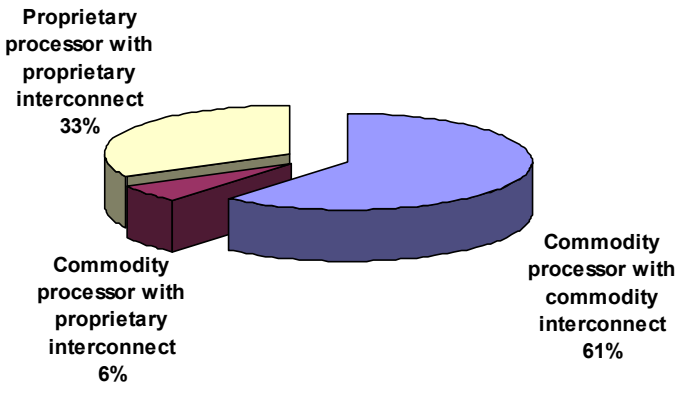


Percent Of 131 Systems Which Use The Following Processors > 1 TFlop/s

About a half are based on 32 bit architecture
9 (11) Machines have a Vector instruction Sets

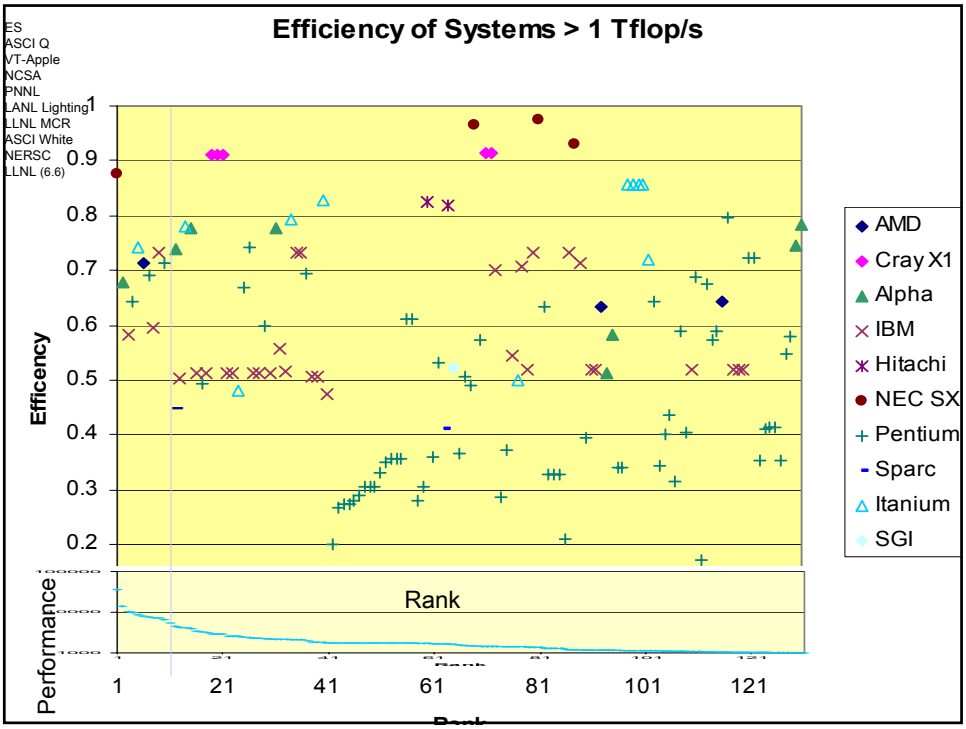


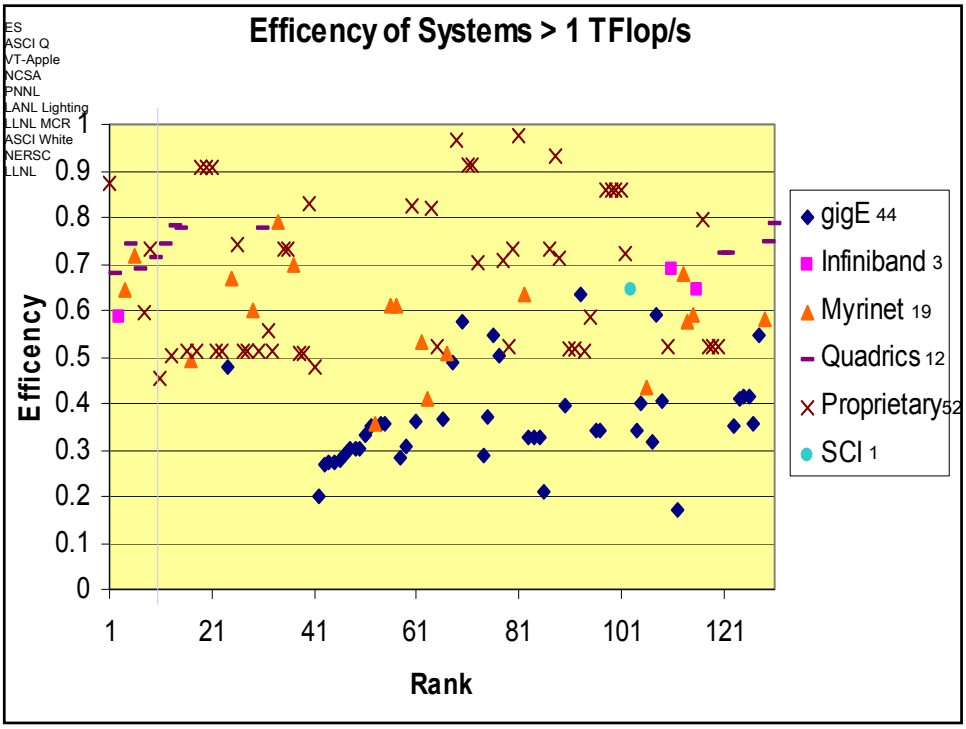
Percent Breakdown by Classes



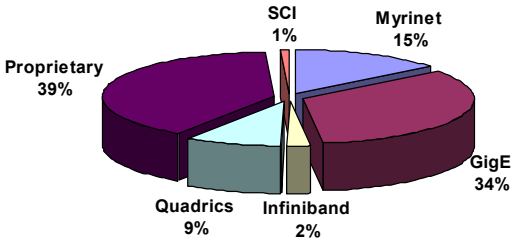
What About Efficiency?

- Talking about Linpack
- What should be the efficiency of a machine on the Top131 be?
  - Percent of peak for Linpack
  - > 90% ?
  - > 80% ?
  - > 70% ?
  - > 60% ?
  - ...
- Remember this is $O(n^3)$ ops and $O(n^2)$ data
  - Mostly matrix multiply


Efficiency of Systems > 1 Tflop/s
ES
ASCI Q
VT-Apple
NCSA
PNNL
LANL Lighting
LLNL MCR
ASCI White
NERSC
LLNL (6.6)
Efficency
Performance
Rank
AMD
Cray X1
Alpha
IBM
Hitachi
NEC SX
Pentium
Sparc
Itanium
SGI


Efficency of Systems > 1 TFlop/s
ES
ASCI Q
VT-Apple
NCSA
PNNL
LANL Lighting
LLNL MCR
ASCI White
NERSC
LLNL
Efficency
Rank
gigE 44
Infiniband 3
Myrinet 19
Quadrics 12
Proprietary 52
SCI 1

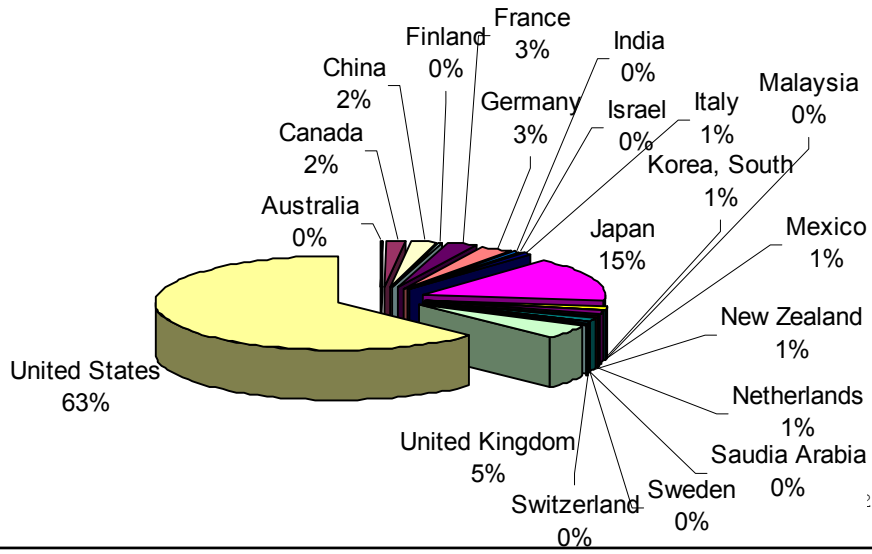
Interconnects Used



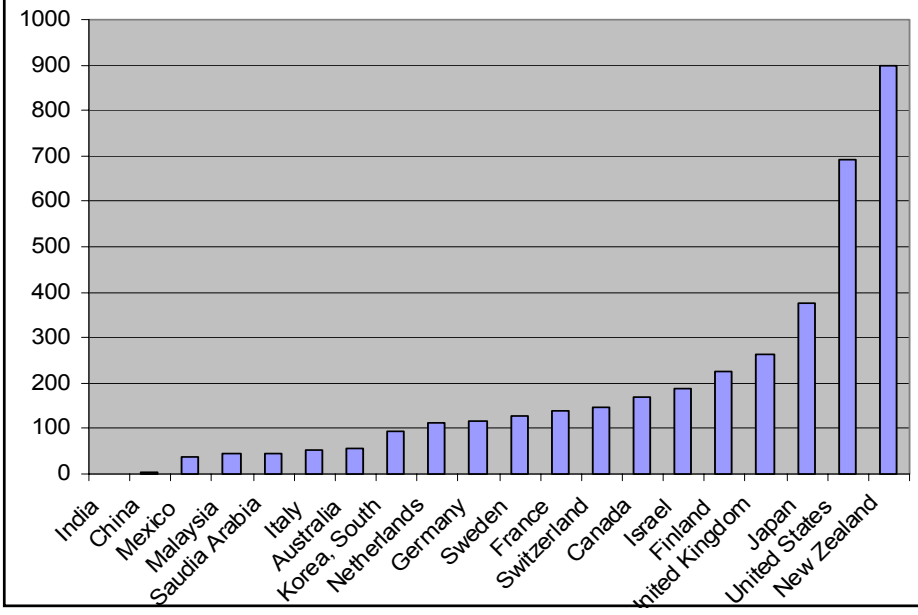
Efficiency for Linpack

	Largest node count	min	max	average
GigE	1024	17%	63%	37%
SCI	120	64%	64%	64%
QsNetII	2000	68%	78%	74%
Myrinet	1250	36%	79%	59%
Infiniband 4x	1100	58%	69%	64%
Proprietary	9632	45%	98%	68%

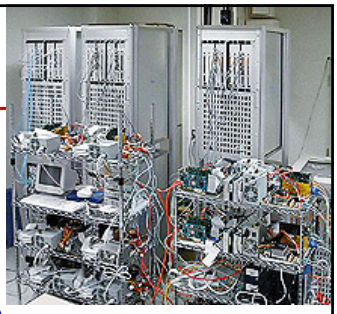
Country Percent by Total Performance



KFlop/s per Capita (Flops/Pop)



Special Purpose: GRAPE-6



- The 6th generation of GRAPE (Gravity Pipe) Project
- Gravity (N-Body) calculation for many particles with 31 Gflops/chip
- 32 chips / board - 0.99 Tflops/board
- 64 boards of full system is installed in University of Tokyo - 63 Tflops
- On each board, all particles data are set onto SRAM memory, and each target particle data is injected into the pipeline, then acceleration data is calculated
  - No software!
- Gordon Bell Prize at SC for a number of years (Prof. Makino, U. Tokyo)

Sony PlayStation2

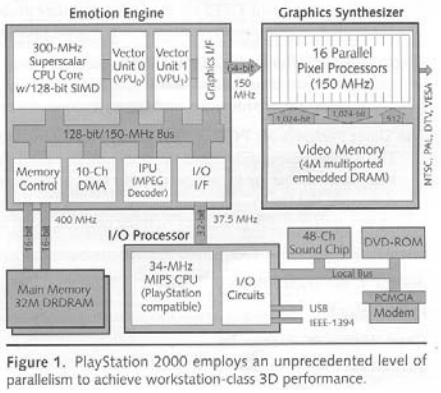


Figure 1. PlayStation 2000 employs an unprecedented level of parallelism to achieve workstation-class 3D performance.

- Emotion Engine:
- 6 Gflop/s peak
- Superscalar MIPS 300 MHz core + vector coprocessor + graphics/DRAM
  - About $200
  - 70M sold

- 8K D cache; 32 MB memory not expandable OS goes here as well
- 32 bit fl pt; not IEEE
- 2.4GB/s to memory (.38 B/Flop)
- Potential 20 fl pt ops/cycle
  - FPU w/FMAC+FDIV
  - VPU_1 w/4FMAC+FDIV
  - VPU_2 w/4FMAC+FDIV
  - EFU w/FMAC+FDIV

High-Performance Chips Embedded Applications

- The driving market is gaming (PC and game consoles)
  - which is the main motivation for almost all the technology developments.
- Demonstrate that arithmetic is quite cheap.
- Not clear that they do much for scientific computing.
- Today there are three big problems with these apparent non-standard "off-the-shelf" chips.
  - Most of these chips have very limited memory bandwidth and little if any support for inter-node communication.
    - Integer or only 32 bit fl.pt
  - No software support to map scientific applications to these processors.
  - Poor memory capacity for program storage
- Developing "custom" software is much more expensive than developing custom hardware.

Real Crisis With HPC Is With The Software

- Programming is stuck
  - Arguably hasn't changed since the 70'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.

Some Current Unmet Needs

- Performance / Portability
- Fault tolerance
- Better programming models
  - Global shared address space
  - Visible locality
- Maybe coming soon (since incremental, yet offering real benefits):
  - Global Address Space (GAS) languages: UPC, Co-Array Fortran, Titanium)
    - "Minor" extensions to existing languages
    - More convenient than MPI
    - Have performance transparency via explicit remote memory references
- The critical cycle of prototyping, assessment, and commercialization must be a long-term, sustaining investment, not a one time, crash program.

Thanks for the Memories and Cycles

ACRI
Alex AVX 2
Alliant
Alliant FX/2800
American Supercomputer
Ametek
Applied Dynamics
Astronautics
Avalon A12
BBN
BBN TC2000
Burroughs BSP
Cambridge Parallel Processing DAP Gamma
C-DAC PARAM 10000 Openframe
C-DAC PARAM 9000/SS
C-DAC PARAM Openframe
CDC
Convex
Convex SPP-1000/1200/1600
Cray Computer
Cray Computer Corp Cray-2
Cray Computer Corp Cray-3
Cray J90
Cray Research
Cray Research Cray Y-MP, Cray Y-MP M90
Cray Research Inc APP
Cray T3D
Cray T3E Classic
Cray T90
Cray Y-MP C90
Culler Scientific
Culler-Harris
Cydrome
Dana/Ardent/Stellar/Stardent
DEC AlphaServer 8200 & 8400
Denelcor HEP
Digital Equipment Corp Alpha farm
Elxsi
ETA Systems
Evans and Sutherland Computer Division
Floating Point Systems
Fujitsu AP1000
Fujitsu VP 100-200-400
Fujitsu VPP300/700
Fujitsu VPP500 series
Fujitsu VPP5000 series
Fujitsu VPX200 series
Galaxy YH-1
Goodyear Aerospace MPP
Gould NPL
Guiltech
Hitachi S-3600 series
Hitachi S-3800 series
Hitachi SR2001 series
Hitachi SR2201 series
HP/Convex C4600
IBM RP3
IBM GF11
IBM ES/9000 series
IBM SP1 series
ICL DAP
Intel Paragon XP
Intel Scientific Computers
International Parallel Machines
J Machine
Kendall Square Research
Kendall Square Research KSR2
Key Computer Laboratories
Kongsberg Informasjonskontroll SCALI
MasPar
MasPar MP-1, MP-2
Meiko
Matsushita ADENART
Meiko CS-1 series
Meiko CS-2 series
Multiflow
Myrias
nCUBE 2S
NEC Cenju-3
NEC Cenju-4
NEC SX-3R
NEC SX-4
NEC SX-5
Numerix
Parsys SN9000 series
Parsys TA9000 series
Parsytec CC series
Parsytec GC/Power Plus
Prisma
S-1
Saxpy
Scientific Computer Systems (SCS)
SGI Origin 2000
Siemens-Nixdorf VP2600 series
Silicon Graphics PowerChallenge
Stern Computing Systems SSP
SUN E10000 Starfire
Supercomputer Systems (SSI)
Supertek
Suprenum
The AxilSCC
The HP Exemplar V2600
Thinking Machines
TMC CM-2(00)
TMC CM-5
Vitesse Electronics