INTERNATIONAL SUPERCOMPUTER CONFERENCE
JUNE 22-25, 2004 IN HEIDELBERG

Survey of *"Present and Future Supercomputer Architectures and their Interconnects"*

Jack Dongarra
University of Tennessee
and
Oak Ridge National Laboratory

Overview

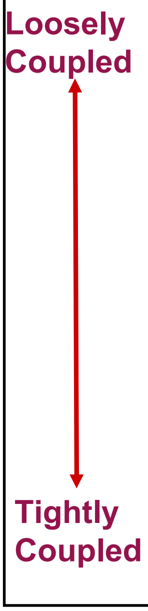
- ♦ **Processors**
- ♦ **Interconnects**
- ♦ **A few machines**
- ♦ **Examine the Top242**

Vibrant Field for High Performance Computers

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

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

Architecture/Systems Continuum



- Commodity processor with commodity interconnect
  - Clusters
    - Pentium, Itanium, Opteron, Alpha
    - 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
- Custom processor with custom interconnect
  - Cray X1
  - NEC SX-7
  - IBM Regatta
  - IBM Blue Gene/L

Commodity Processors

- Intel Pentium Xeon
  - 3.2 GHz, peak = 6.4 Gflop/s
  - Linpack 100 = 1.7 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

High Bandwidth vs Commodity Systems

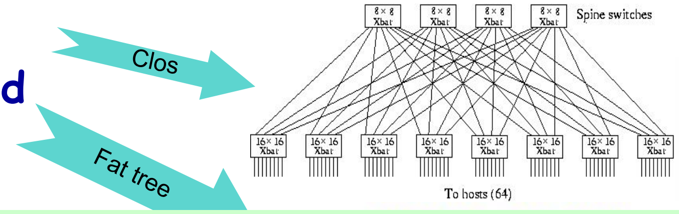
- High bandwidth systems have traditionally been vector computers
  - Designed for scientific problems
  - Capability computing
- Commodity processors are designed for web servers and the home PC market

  (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 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

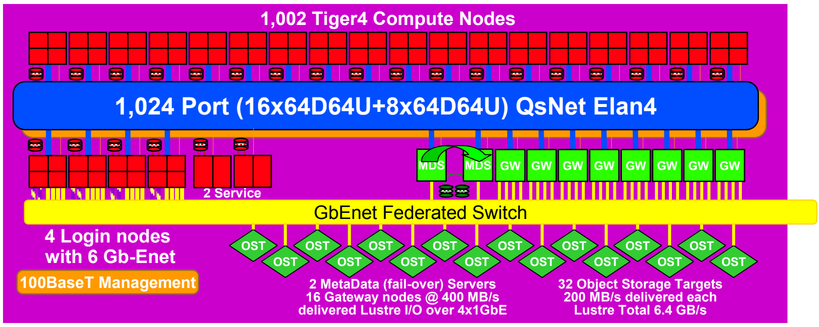
Commodity Interconnects

- ♦ Gig Ethernet
- ♦ Myrinet
- ♦ Infiniband
- ♦ QsNet
- ♦ SCI



					MPI Lat / 1-way / Bi-Dir
	Switch topology	$ NIC	$Sw/node	$ Node	(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

DOE - Lawrence Livermore National Lab's Itanium 2 Based Thunder System Architecture
1,024 nodes, 4096 processors, 23 TF/s peak


1,002 Tiger4 Compute Nodes
1,024 Port (16x64D64U+8x64D64U) QsNet Elan4
MDS
MDS
GW
GW
GW
GW
GW
GW
GW
GW
2 Service
GbEnet Federated Switch
4 Login nodes
with 6 Gb-Enet
100BaseT Management
OST
2 MetaData (fail-over) Servers
16 Gateway nodes @ 400 MB/s
delivered Lustre I/O over 4x1GbE
32 Object Storage Targets
200 MB/s delivered each
Lustre Total 6.4 GB/s

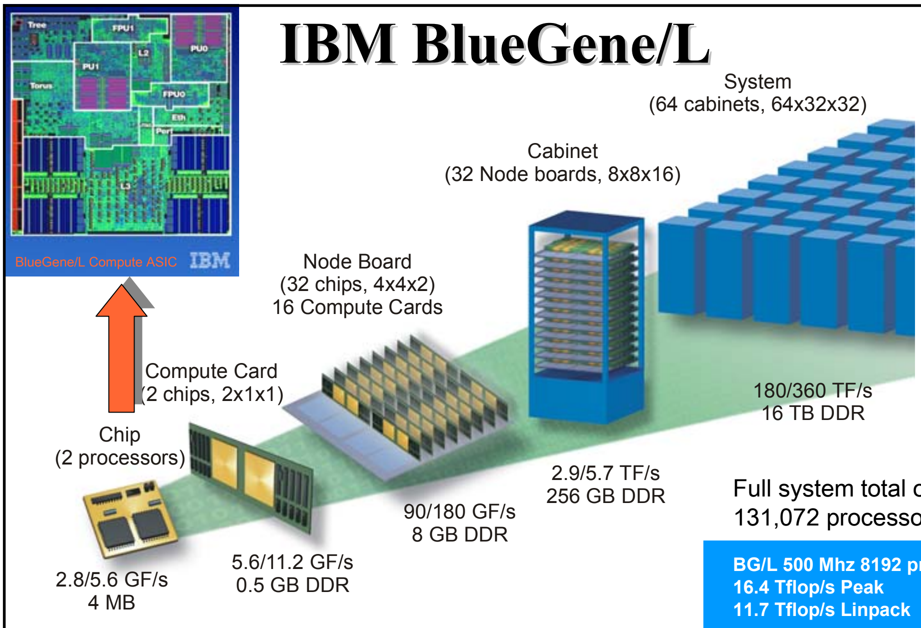
System Parameters

- Quad 1.4 GHz Itanium2 Madison Tiger4 nodes with 8.0 GB DDR266 SDRAM
- <3 µs, 900 MB/s MPI latency and Bandwidth over QsNet Elan4
- Support 400 MB/s transfers to Archive over quad Jumbo Frame Gb-Enet and QSW links from each Login node
- 75 TB in local disk in 73 GB/node UltraSCSI320 disk
- 50 MB/s POSIX serial I/O to any file system
- 8.7 B:F = 192 TB global parallel file system in multiple RAID5
- Lustre file system with 6.4 GB/s delivered parallel I/O performance
  - MPI I/O based performance with a large sweet spot
  - 32 < MPI tasks < 4,096
- Software RHEL 3.0, CHAOS, SLURM/DPCS, MPICH2, TotalView, Intel and GNU Fortran, C and C++ compilers

4096 processor
19.9 TFlop/s Linpack
87% peak

Contracts with
- **California Digital Corp for nodes and integration**
- **Quadrics for Elan4**
- **Data Direct Networks for global file system**
- **Cluster File System for Lustre support**

IBM BlueGene/L



System
(64 cabinets, 64x32x32)

Cabinet
(32 Node boards, 8x8x16)

Node Board
(32 chips, 4x4x2)
16 Compute Cards

Compute Card
(2 chips, 2x1x1)

Chip
(2 processors)

2.8/5.6 GF/s
4 MB

5.6/11.2 GF/s
0.5 GB DDR

90/180 GF/s
8 GB DDR

2.9/5.7 TF/s
256 GB DDR

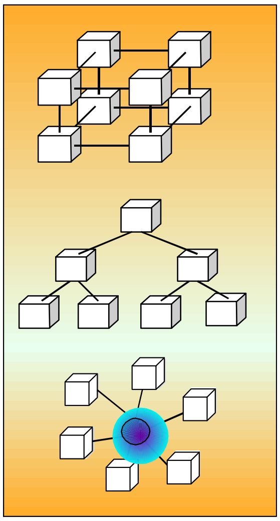
180/360 TF/s
16 TB DDR

Full system total of
131,072 processors

BG/L 500 Mhz 8192 proc
16.4 Tflop/s Peak
11.7 Tflop/s Linpack

BG/L 700 MHz 4096 proc
11.5 Tflop/s Peak
8.7 Tflop/s Linpack

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

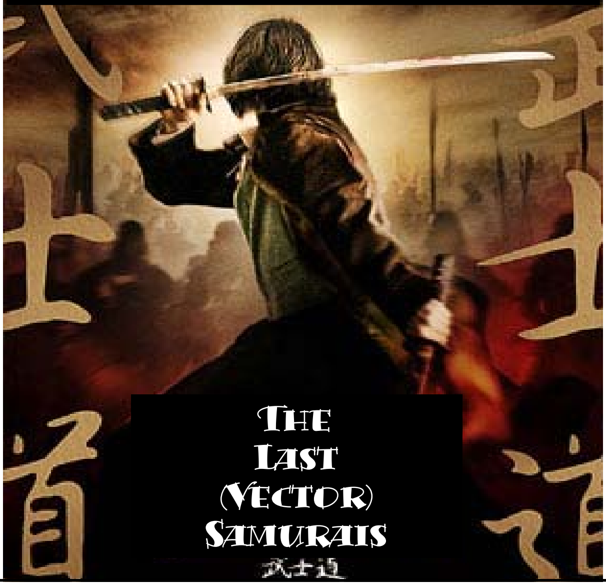
- **Incorporated into every node ASIC**
- **Active in the I/O nodes (1:64)**
- **All external comm. (file I/O, control, user interaction, etc.)**

Low Latency Global Barrier and Interrupt

- **Latency of round trip 1.3 µs**

Control Network

Cray and NEC



Cray X1 Vector Processor

- Cray X1 builds a victor processor called an MSP
  - 4 SSPs (each a 2-pipe vector processor) make up an MSP
  - Compiler will (try to) vectorize/parallelize across the MSP
  - Cache (unusual on earlier vector machines)

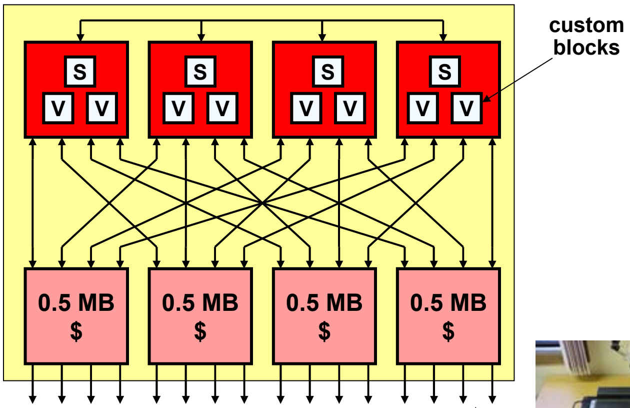
12.8 Gflops (64 bit)

25.6 Gflops (32 bit)

51 GB/s ↑

25-41 GB/s ↓

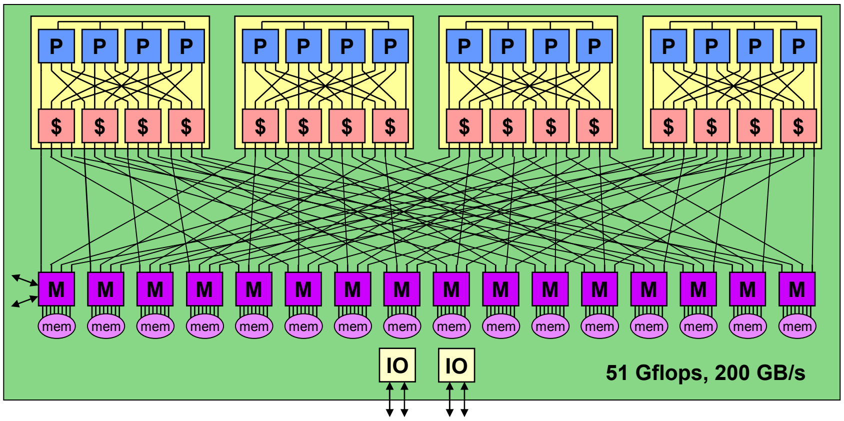
2 MB Ecache



At frequency of 400/800 MHz

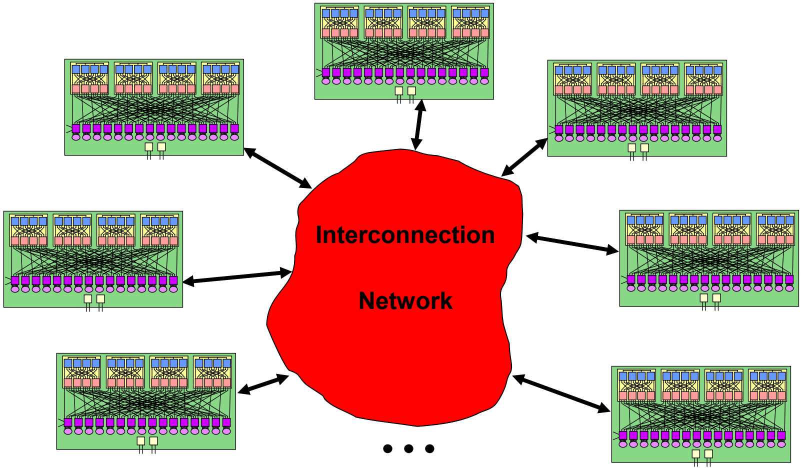
To local memory and network: 25.6 GB/s ↑
12.8 - 20.5 GB/s

Cray X1 Node



- Four multistream processors (MSPs), each 12.8 Gflops
- High bandwidth local shared memory (128 Direct Rambus channels)
- 32 network links and four I/O links per node

NUMA Scalable up to 1024 Nodes

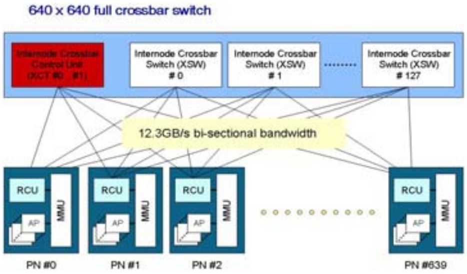


- 16 parallel networks for bandwidth

At Oak Ridge National Lab 128 nodes,
504 processor machine, 5.9 Tflop/s for Linpack
(out of 6.4 Tflop/s peak, 91%)

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)
- A record 5 times #1 on Top500
- H. Miyoshi; architect
  - NAL, RIST, ES
  - Fujitsu AP, VP400, NWT, ES


640 x 640 full crossbar switch
Internode Crossbar Control Unit (XCT #0 #1)
Internode Crossbar Switch (XSW) # 0
Internode Crossbar Switch (XSW) # 1
Internode Crossbar Switch (XSW) # 127
12.3GB/s bi-sectional bandwidth
RCU
AP
MMU
PN #0
PN #1
PN #2
PN #639

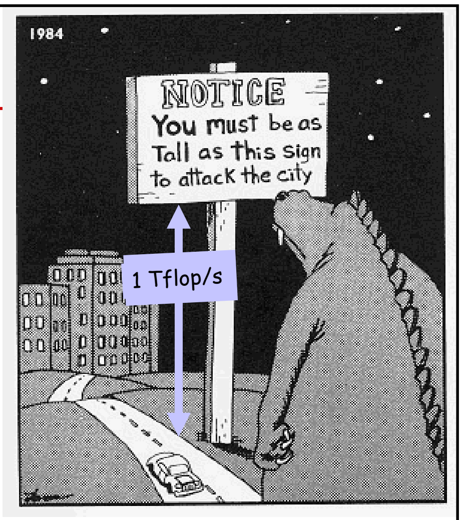
- Footprint of 4 tennis courts
- Expect to be on top of Top500 for another 6 months to a year.

- From the Top500 (June 2004)
  - Performance of ESC > Σ Next Top 2 Computers

The Top242

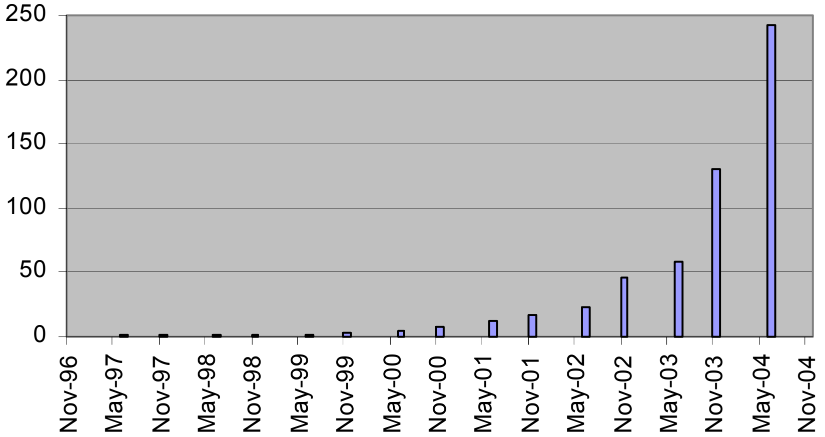
- Focus on machines that are at least 1 TFlop/s on the Linpack benchmark

- Linpack Based
  - 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
- 2004:
  - #1 = 35.8 TFlop/s
  - #500 = 813 GFlop/s

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



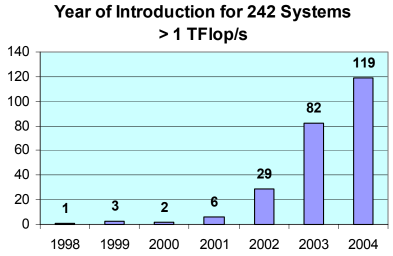
Factoids on Machines > 1 TFlop/s

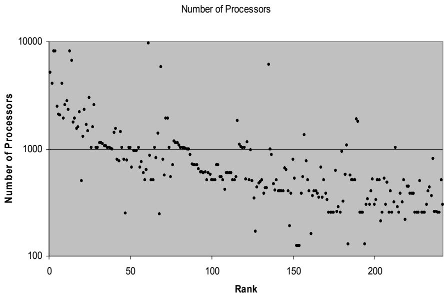
- 242 Systems
- 171 Clusters (71%)

- Average rate: 2.54 Tflop/s
- Median rate: 1.72 Tflop/s

- Sum of processors in Top242: 238,449
  - Sum for Top500: 318,846
- Average processor count: 985
- Median processor count: 565

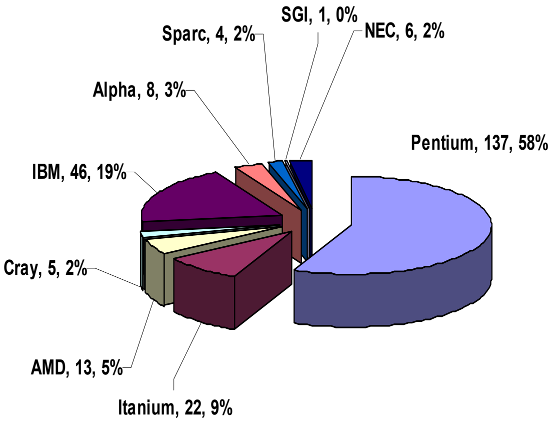
- Numbers of processors
  - Most number of processors: 9632_{61}
    - ASCI Red
  - Fewest number of processors: 124_{152}
    - Cray X1

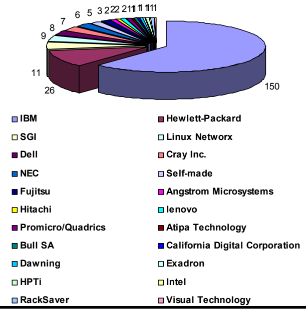




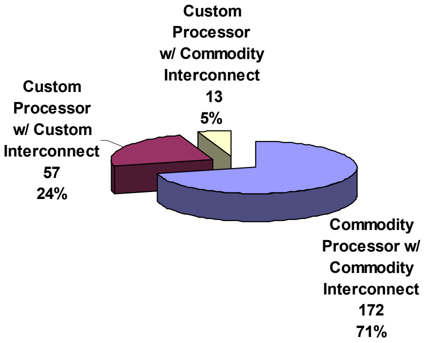
Percent Of 242 Systems Which Use The Following Processors > 1 TFlop/s

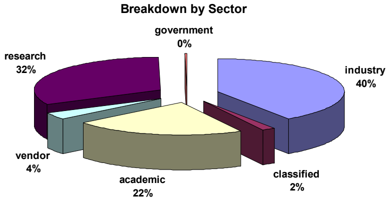
More than half are based on 32 bit architecture
11 Machines have a Vector instruction Sets





Percent Breakdown by Classes

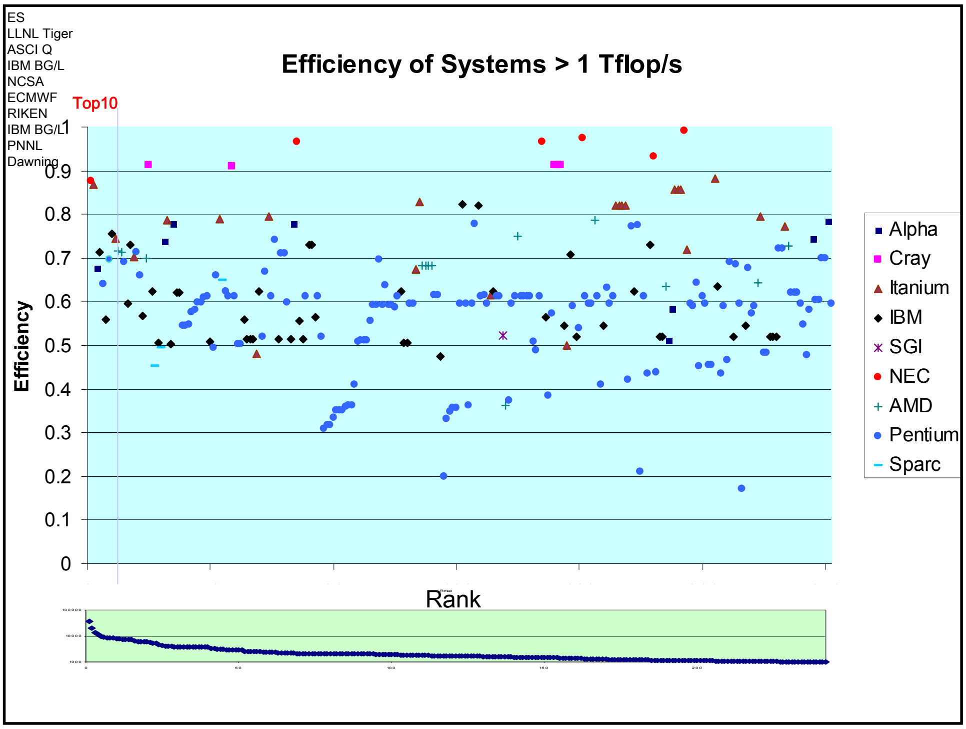


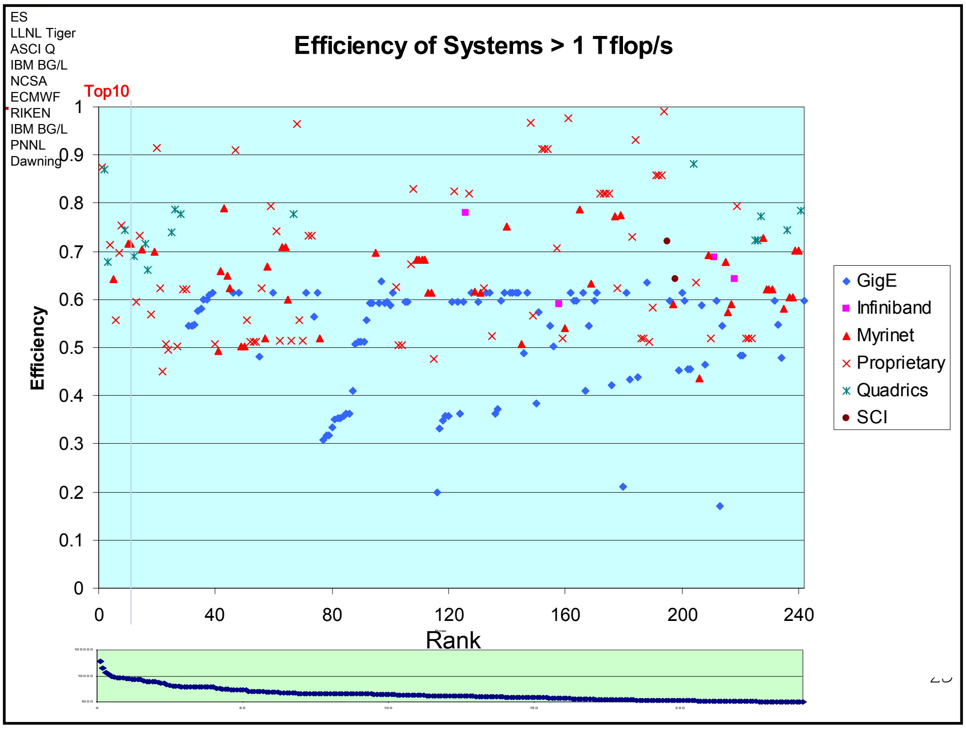


What About Efficiency?

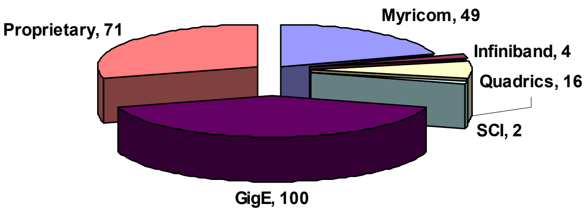
- Talking about Linpack
- What should be the efficiency of a machine on the Top242 be?
  - Percent of peak for Linpack
    - 90% ?
    - 80% ?
    - 70% ?
    - 60% ?

    …
- Remember this is O(n³) ops and O(n²) data
  - Mostly matrix multiply



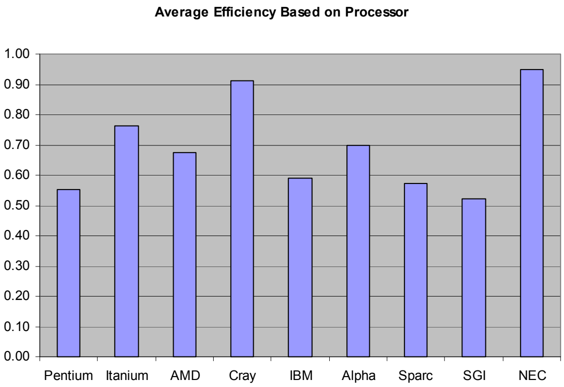


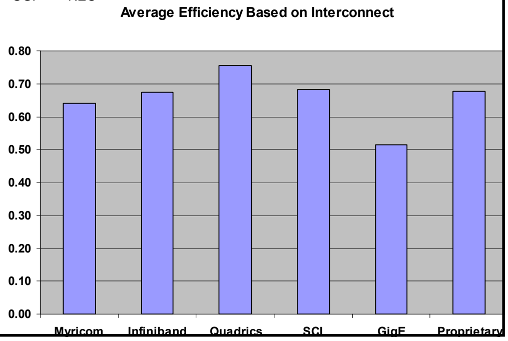
Interconnects Used in the Top242



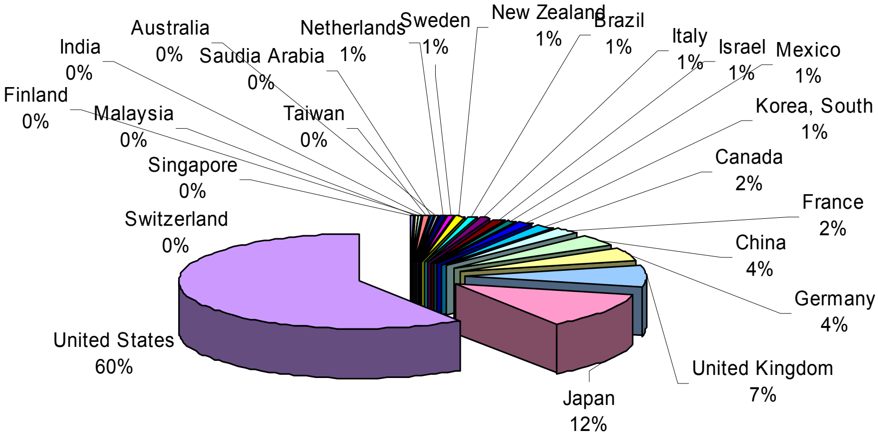
Efficiency for Linpack

	Largest node count	min	max	average
GigE	1128	17%	64%	51%
SCI	400	64%	68%	72%
QsNetII	4096	66%	88%	75%
Myrinet	1408	44%	79%	64%
Infiniband	768	59%	78%	75%
Proprietary	9632	45%	99%	68%

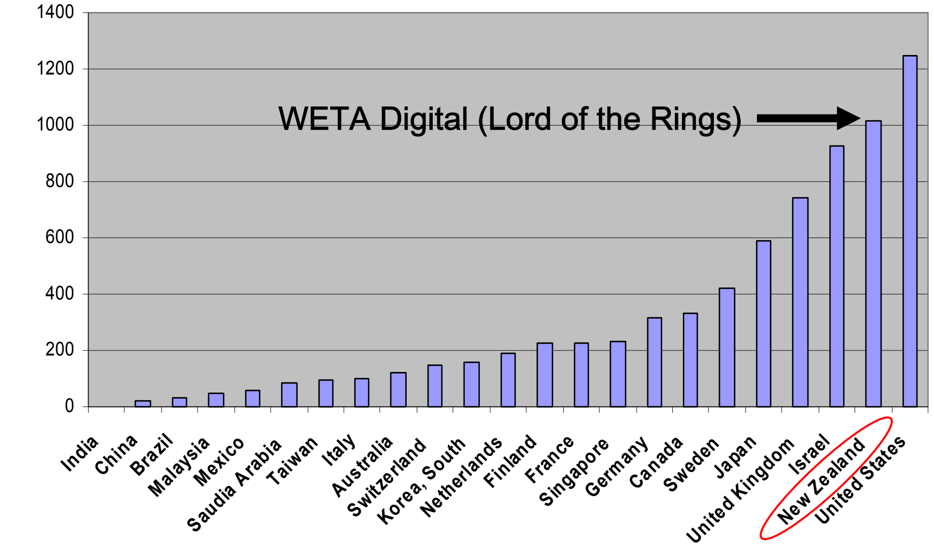




Country Percent by Total Performance

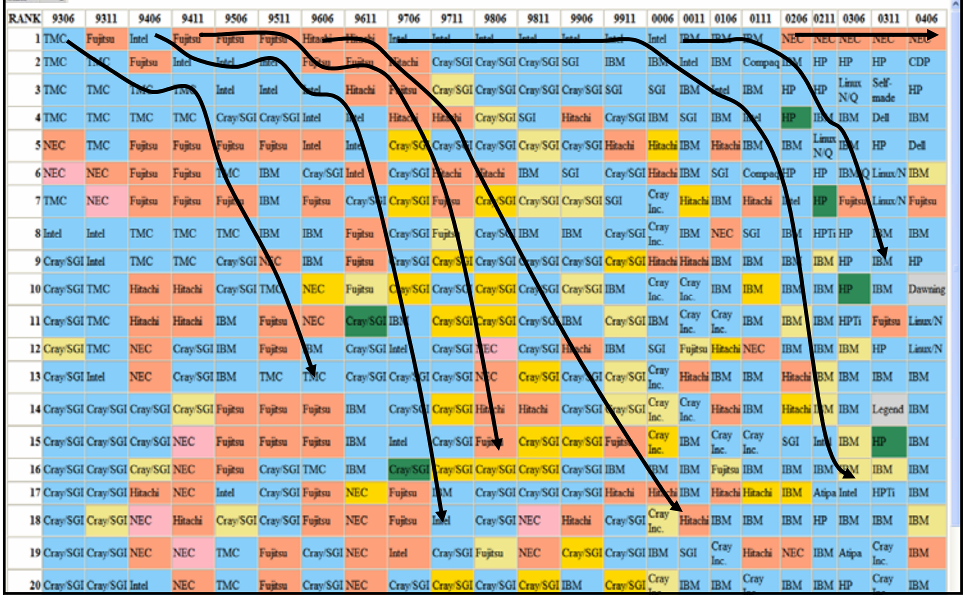


KFlop/s per Capita (Flops/Pop)



Color Codes	
Country	Count
USA	282
Japan	101
UK	32
Germany	30
Canada	7
France	6
China	2

Top20 Over the Past 11 Years



RANK	9306	9311	9406	9411	9506	9511	9606	9611	9706	9711	9806	9811	9906	9911	0006	0011	0106	0111	0206	0211	0306	0311	0406
1	TMC	Fujitsu	Intel	Fujitsu	Fujitsu	Fujitsu	Hitachi	Hitachi	Intel	Intel	Intel	Intel	Intel	Intel	Intel	IBM	IBM	IBM	NEC	NEC	NEC	NEC	NEC
2	TMC	TMC	Fujitsu	Intel	Intel	Intel	Fujitsu	Fujitsu	Hitachi	Cray/SGI	Cray/SGI	Cray/SGI	SGI	IBM	IBM	Intel	IBM	Compaq	IBM	HP	HP	HP	CDP
3	TMC	TMC	TMC	TMC	Intel	Intel	Intel	Hitachi	Fujitsu	Cray/SGI	Cray/SGI	Cray/SGI	Cray/SGI	SGI	SGI	IBM	Intel	IBM	HP	HP	Linux N/Q	Self-made	HP
4	TMC	TMC	TMC	TMC	Cray/SGI	Cray/SGI	Intel	Intel	Hitachi	Hitachi	Cray/SGI	SGI	Hitachi	Cray/SGI	IBM	SGI	IBM	Intel	HP	IBM	IBM	Dell	IBM
5	NEC	TMC	Fujitsu	Fujitsu	Fujitsu	Fujitsu	Intel	Intel	Cray/SGI	Cray/SGI	Cray/SGI	Cray/SGI	Cray/SGI	Hitachi	Hitachi	IBM	Hitachi	IBM	IBM	Linux N/Q	IBM	HP	Dell
6	NEC	NEC	Fujitsu	Fujitsu	TMC	IBM	Cray/SGI	Intel	Cray/SGI	Hitachi	Hitachi	IBM	SGI	Cray/SGI	Hitachi	IBM	SGI	Compaq	HP	HP	IBM Q	Linux/N	IBM
7	TMC	NEC	Fujitsu	Fujitsu	Fujitsu	IBM	Fujitsu	Cray/SGI	Cray/SGI	Fujitsu	Cray/SGI	Cray/SGI	Cray/SGI	SGI	Cray Inc.	Hitachi	IBM	Hitachi	Intel	HP	Fujitsu	Linux/N	Fujitsu
8	Intel	Intel	TMC	TMC	TMC	IBM	IBM	Fujitsu	Cray/SGI	Fujitsu	Cray/SGI	IBM	IBM	Cray/SGI	Cray Inc.	IBM	NEC	SGI	IBM	HPTi	HP	IBM	IBM
9	Cray/SGI	Intel	TMC	TMC	Cray/SGI	NEC	IBM	Fujitsu	Cray/SGI	Cray/SGI	Cray/SGI	Cray/SGI	Cray/SGI	Cray/SGI	Hitachi	Hitachi	IBM	IBM	IBM	IBM	HP	IBM	HP
10	Cray/SGI	TMC	Hitachi	Hitachi	Cray/SGI	TMC	NEC	Fujitsu	Cray/SGI	Cray/SGI	Cray/SGI	Cray/SGI	Cray/SGI	IBM	Cray Inc.	Cray Inc.	IBM	IBM	IBM	IBM	HP	IBM	Dawning
11	Cray/SGI	TMC	Hitachi	Hitachi	IBM	Fujitsu	NEC	Cray/SGI	IBM	Cray/SGI	Cray/SGI	Cray/SGI	IBM	Cray/SGI	IBM	Cray Inc.	Cray Inc.	IBM	IBM	IBM	HPTi	Fujitsu	Linux/N
12	Cray/SGI	TMC	NEC	Cray/SGI	IBM	Fujitsu	IBM	Cray/SGI	Intel	Cray/SGI	NEC	Cray/SGI	Hitachi	IBM	SGI	Fujitsu	Hitachi	NEC	IBM	IBM	IBM	HP	Linux/N
13	Cray/SGI	Intel	NEC	Cray/SGI	IBM	TMC	TMC	Cray/SGI	Cray/SGI	Cray/SGI	NEC	Cray/SGI	Cray/SGI	Cray/SGI	Cray Inc.	Hitachi	IBM	IBM	Hitachi	IBM	IBM	IBM	IBM
14	Cray/SGI	Cray/SGI	Cray/SGI	Cray/SGI	Fujitsu	Fujitsu	Fujitsu	IBM	Cray/SGI	Cray/SGI	Hitachi	Hitachi	Cray/SGI	Cray/SGI	Cray Inc.	Cray Inc.	Hitachi	IBM	Hitachi	IBM	IBM	Legend	IBM
15	Cray/SGI	Cray/SGI	Cray/SGI	NEC	Fujitsu	Fujitsu	Fujitsu	IBM	Intel	Cray/SGI	Fujitsu	Cray/SGI	Cray/SGI	Fujitsu	Cray Inc.	IBM	Cray Inc.	Cray Inc.	SGI	Intel	IBM	HP	IBM
16	Cray/SGI	Cray/SGI	Cray/SGI	NEC	Fujitsu	Cray/SGI	TMC	IBM	Cray/SGI	Cray/SGI	Cray/SGI	Cray/SGI	Cray/SGI	IBM	IBM	IBM	Fujitsu	IBM	IBM	IBM	IBM	IBM	IBM
17	Cray/SGI	Cray/SGI	Hitachi	NEC	Intel	Cray/SGI	Fujitsu	NEC	Fujitsu	IBM	Cray/SGI	Cray/SGI	Cray/SGI	Hitachi	Hitachi	IBM	Hitachi	Hitachi	IBM	Atipa	Intel	HPTi	IBM
18	Cray/SGI	Cray/SGI	NEC	Hitachi	Cray/SGI	Cray/SGI	Fujitsu	NEC	Fujitsu	Intel	Cray/SGI	NEC	Hitachi	Cray/SGI	Cray Inc.	Hitachi	IBM	IBM	IBM	HP	IBM	IBM	IBM
19	Cray/SGI	Cray/SGI	NEC	NEC	TMC	Fujitsu	Cray/SGI	NEC	Intel	Cray/SGI	Fujitsu	NEC	Cray/SGI	Cray/SGI	IBM	SGI	Cray Inc.	Hitachi	NEC	IBM	Atipa	Cray Inc.	IBM
20	Cray/SGI	Cray/SGI	Intel	NEC	TMC	Fujitsu	Cray/SGI	NEC	Cray/SGI	Cray/SGI	Cray/SGI	Cray/SGI	IBM	Cray/SGI	Cray Inc.	IBM	IBM	Cray Inc.	IBM	IBM	HP	Cray Inc.	IBM

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.

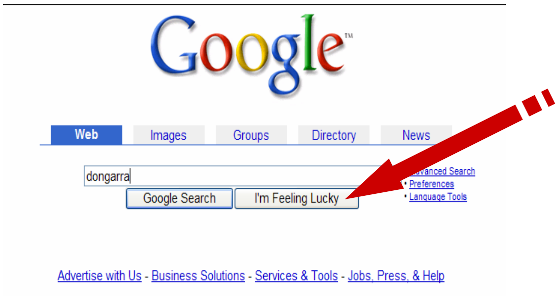
Some Current Unmet Needs

- 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.

Collaborators / Support

- Top500 Team
  - Erich Strohmaier, NERSC
  - Hans Meuer, Mannheim
  - Horst Simon, NERSC

- For more information:
  - Google “dongarra”
  - Click on “talks”



dongarra