LCSC
5th Annual Workshop on
Linux Clusters for Super Computing
October 18-21, 2004
Linköping University, Sweden

Cluster Computing: You've Come A Long Way In A Short Time

Jack Dongarra
University of Tennessee
and
Oak Ridge National Laboratory

National Supercomputer Centre in Linköping Sweden
NOTUR Norwegian High Performance Computing Consortium

Vibrant Field for High Performance Computers

- Cray X1
- SGI Altix
- IBM Regatta
- IBM Blue Gene/L
- IBM eServer
- Sun
- HP
- Bull NovaScale
- Fujitsu PrimePower
- Hitachi SR11000
- NEC SX-7
- Apple

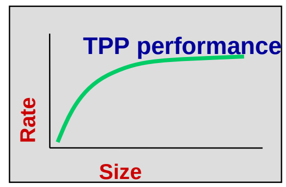
- Coming soon …
  - Cray RedStorm
  - Cray BlackWidow
  - NEC SX-8

2

TOP500 supercomputer

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

- Listing of the 500 most powerful Computers in the World
- Yardstick: Rmax from LINPACK MPP
  Ax=b, dense problem



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

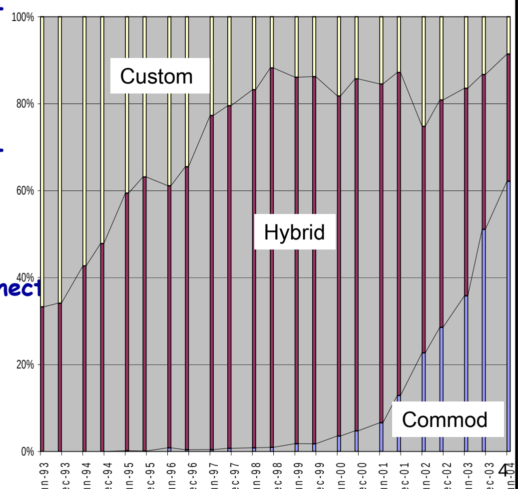
3

Architecture/Systems Continuum

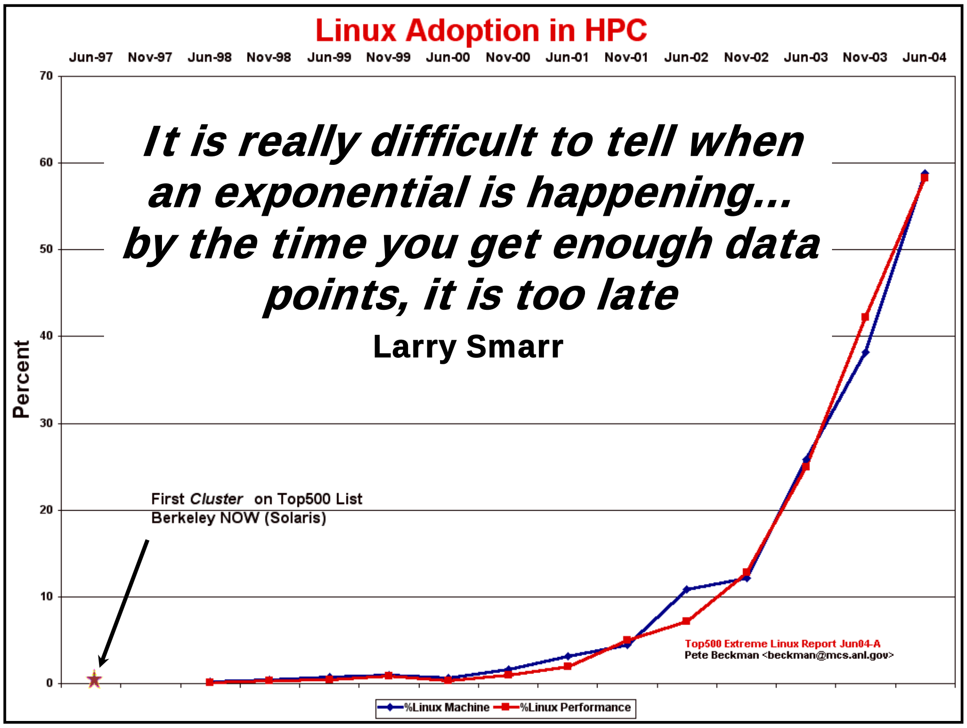
Tightly Coupled

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

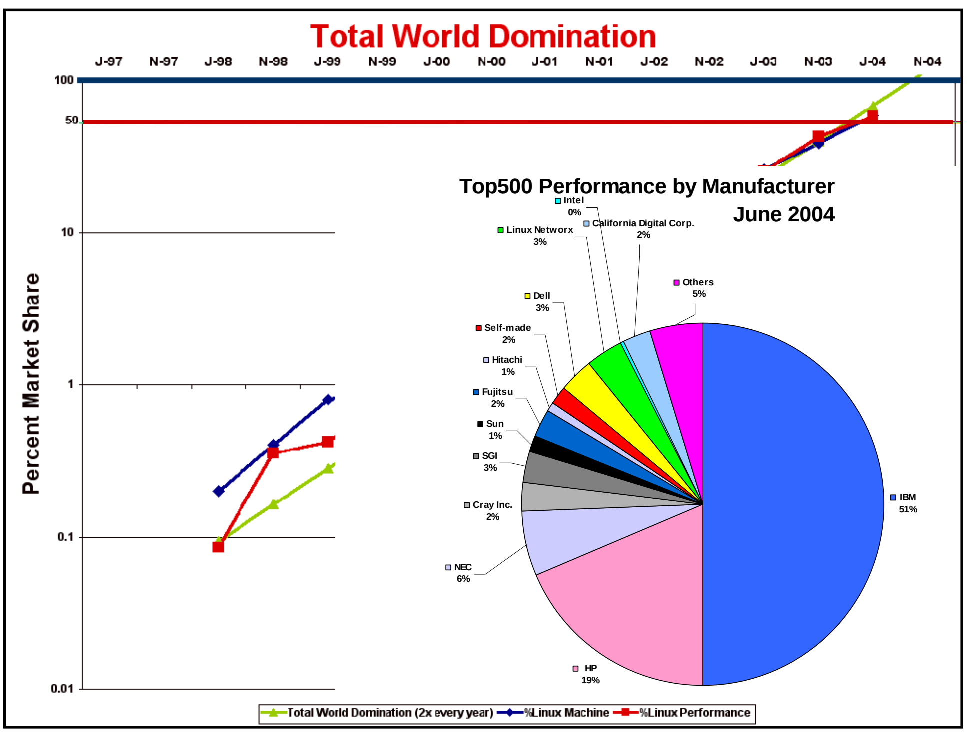
Loosely Coupled



4



Linux Adoption in HPC
Jun-97 Nov-97 Jun-98 Nov-98 Jun-99 Nov-99 Jun-00 Nov-00 Jun-01 Nov-01 Jun-02 Nov-02 Jun-03 Nov-03 Jun-04
Percent
70
60
50
40
30
20
10
0
It is really difficult to tell when an exponential is happening… by the time you get enough data points, it is too late
Larry Smarr
First Cluster on Top500 List
Berkeley NOW (Solaris)
Top500 Extreme Linux Report Jun04-A
Pete Beckman <beckman@mcs.anl.gov>
%Linux Machine
%Linux Performance



Total World Domination
J-97 N-97 J-98 N-98 J-99 N-99 J-00 N-00 J-01 N-01 J-02 N-02 J-03 N-03 J-04 N-04
Percent Market Share
100
50
10
1
0.1
0.01
Top500 Performance by Manufacturer
June 2004
Intel
0%
California Digital Corp.
2%
Linux Networx
3%
Others
5%
Dell
3%
Self-made
2%
Hitachi
1%
Fujitsu
2%
Sun
1%
SGI
3%
Cray Inc.
2%
NEC
6%
HP
19%
IBM
51%
Total World Domination (2x every year)
%Linux Machine
%Linux Performance

The Golden Age of HPC Linux

- **The adoption rate of Linux HPC is phenomenal!**
  - **Linux in the Top500 is (was) doubling every 12 months**
  - **Linux adoption is not driven by bottom feeders**
    - *Adoption is actually faster at the ultra-scale!*
- Most supercomputers run Linux
- **Adoption rate driven by several factors:**
  - **Linux is stable: Often the default platform for CS research**
  - **Essentially no barrier to entry**
  - Effort to learn programming paradigm, libs, devl env., and tools preserved across many orders of magnitude
  - **Stable, complete, portable, middleware software stacks:**
    - **MPICH, MPI-IO, PVFS, PBS, math libraries, etc**

7

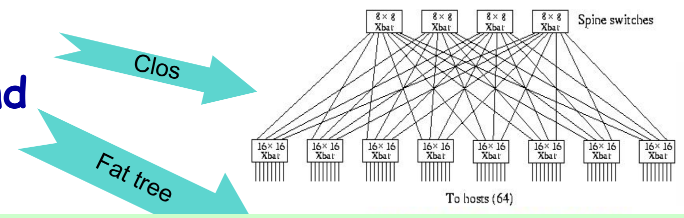
Commodity Processors

- **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

8

Commodity Interconnects

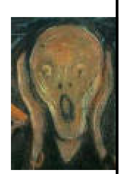
- Gig Ethernet
- Myrinet
- Infiniband
- QsNet
- SCI



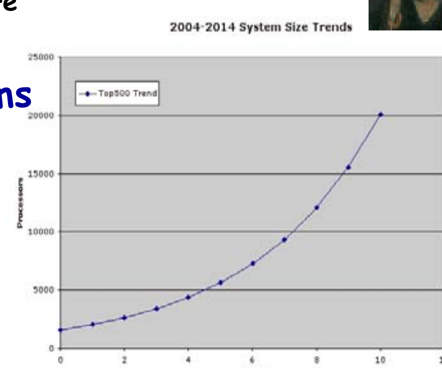
		Cost	Cost	Cost	MPI Lat / 1-way / Bi-Dir
	Switch topology	NIC	Sw/node	Node	(us) / MB/s / MB/s
Gigabit Ethernet	Bus	$ 50	$ 50	$ 100	30 / 100 / 150
SCI	Torus	$1,600	$ 0	$1,600	5 / 300 / 400
QsNetII (R)	Fat Tree	$1,200	$1,700	$2,900	3 / 880 / 900
QsNetII (E)	Fat Tree	$1,000	$ 700	$1,700	3 / 880 / 900
Myrinet (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

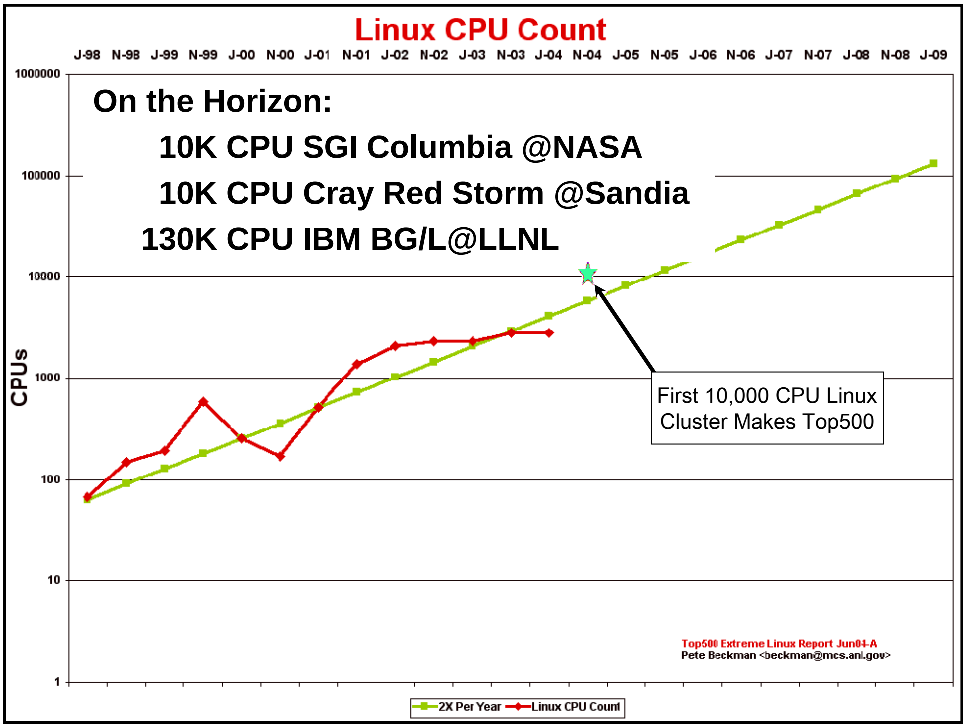
How Big Is Big?

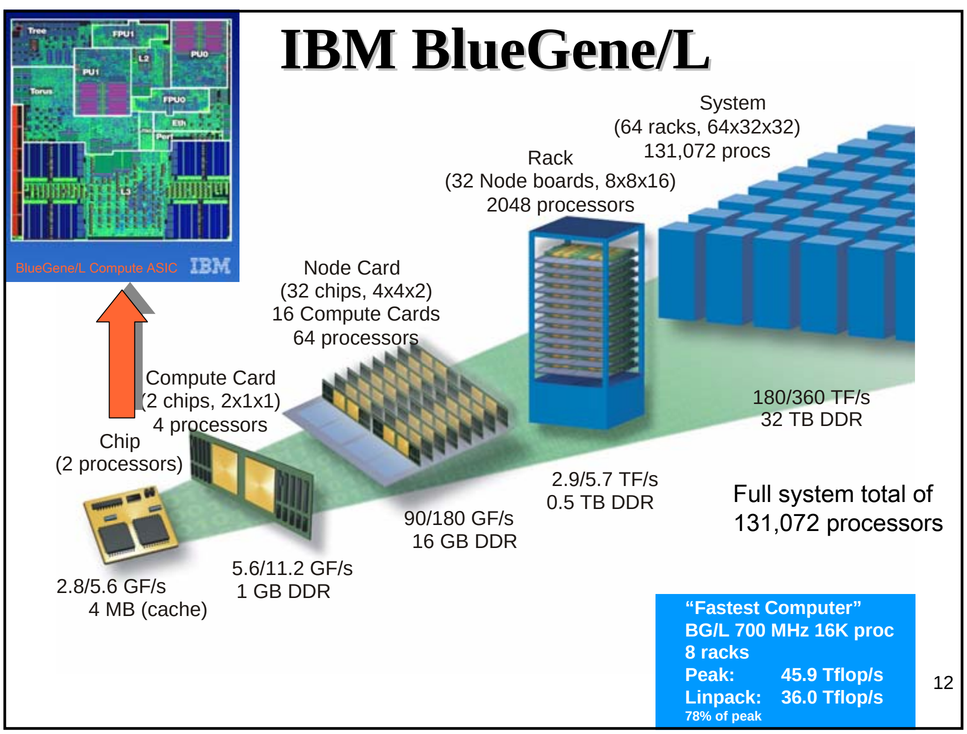
- Every 10X brings new challenges
  - 64 processors was once considered large
    - it hasn't been "large" for quite a while
  - 1024 processors is today's "medium" size
  - 2048-8096 processors is today's "large"
    - we're struggling even here



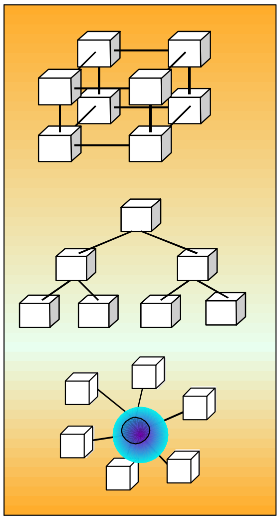
- 100K processor systems
  - are in construction
  - we have fundamental challenges …
  - … and no integrated research program




Linux CPU Count
J-98 N-98 J-99 N-99 J-00 N-00 J-01 N-01 J-02 N-02 J-03 N-03 J-04 N-04 J-05 N-05 J-06 N-06 J-07 N-07 J-08 N-08 J-09
1000000
100000
10000
1000
100
10
1
CPUs
On the Horizon:
10K CPU SGI Columbia @NASA
10K CPU Cray Red Storm @Sandia
130K CPU IBM BG/L@LLNL
First 10,000 CPU Linux Cluster Makes Top500
Top500 Extreme Linux Report Jun04-A
Pete Beckman <beckman@mcs.anl.gov>
2X Per Year
Linux CPU Count


IBM BlueGene/L
BlueGene/L Compute ASIC
IBM
System
(64 racks, 64x32x32)
131,072 procs
Rack
(32 Node boards, 8x8x16)
2048 processors
Node Card
(32 chips, 4x4x2)
16 Compute Cards
64 processors
Compute Card
(2 chips, 2x1x1)
4 processors
Chip
(2 processors)
2.8/5.6 GF/s
4 MB (cache)
5.6/11.2 GF/s
1 GB DDR
90/180 GF/s
16 GB DDR
2.9/5.7 TF/s
0.5 TB DDR
180/360 TF/s
32 TB DDR
Full system total of
131,072 processors
"Fastest Computer"
BG/L 700 MHz 16K proc
8 racks
Peak: 45.9 Tflop/s
Linpack: 36.0 Tflop/s
78% of peak
12

BlueGene/L Interconnection Networks



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

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

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

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

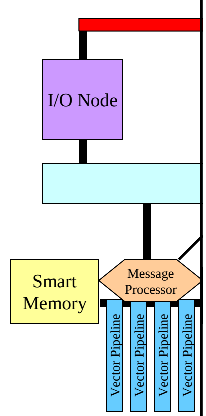
Control Network

13

OS for IBM's BG/L

- **Service Node:**
  - Linux SuSE SLES 8
- **Front End Nodes:**
  - Linux SuSE SLES 9
- **I/O Nodes:**
  - An embedded Linux
- **Compute Nodes:**
  - Home-brew OS

- **Trend:**
  - Extremely large systems run an "*OS Suite*"
  - Functional Decomposition trend lends itself toward a customized, optimized point-solution OS
  - Hierarchical Organization requires software to manage topology, call forwarding, and collective operations



14

Sandia National Lab's Red Storm

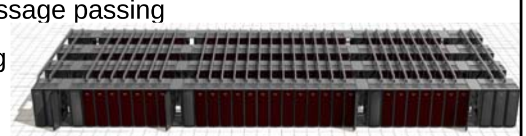
- Red Storm is a supercomputer system leveraging over 10,000 AMD Opteron™ processors connected by an innovative high speed, high bandwidth 3D mesh interconnect designed by Cray.
- Cray was awarded $93M to build the Red Storm system to support the Department of Energy's Nuclear stockpile stewardship program for advanced 3D modeling and simulation.
- Scientists at Sandia National Lab helped with the architectural design of the Red Storm supercomputer.

RED STORM

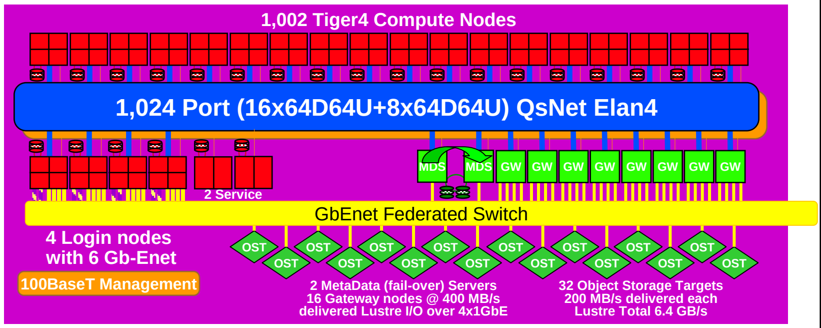
15

Red Storm System Overview

- 40TF peak performance
- 108 compute node cabinets, 16 service and I/O node cabinets, and 16 Red/Black switch cabinets
  – 10,368 compute processors - 2.0 GHz AMD Opteron™
  – 512 service and I/O processors (256P for red, 256P for black)
  – 10 TB DDR memory
- 240 TB of disk storage(120TB for red, 120TB for black)
- MPP System Software
  – Linux + lightweight compute node operating system
  – Managed and used as a single system
  – Easy to use programming environment
  – Common programming environment
  – High performance file system
  – Low overhead RAS and message passing
- Approximately 3,000 ft² including disk systems



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



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

17

High Bandwidth vs Commodity Systems

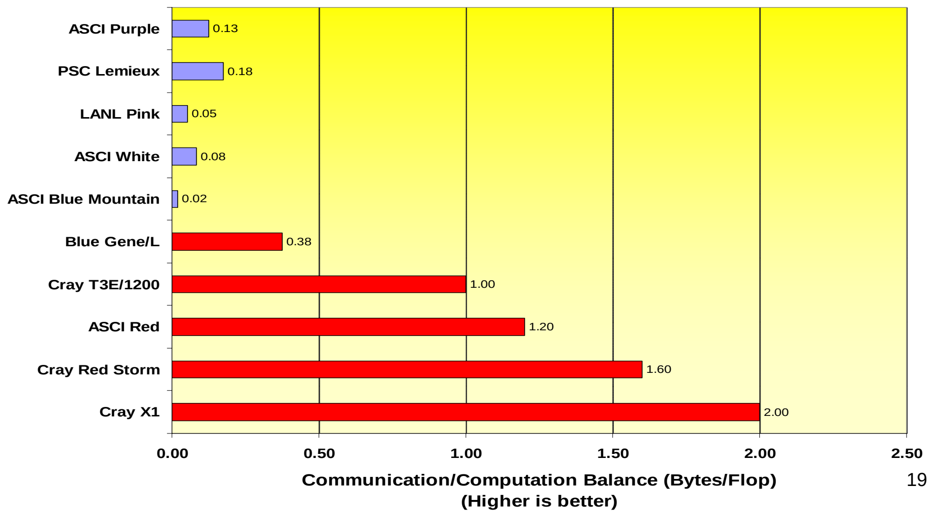
- High bandwidth systems have traditionally been vector computers
  - Designed for scientific problems
  - Capability computing
- Commodity processors are designed for web servers and the home PC market
  (should be thankful that the manufactures keep the 64 bit fl pt)
  - Used for cluster based computers leveraging price point
- Scientific computing needs are different
  - Require a better balance between data movement and floating point operations. Results in greater efficiency.

System Balance - MEMORY BANDWIDTH

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

System Balance (Network)

Network Speed (MB/s) vs Node speed (flop/s)



19

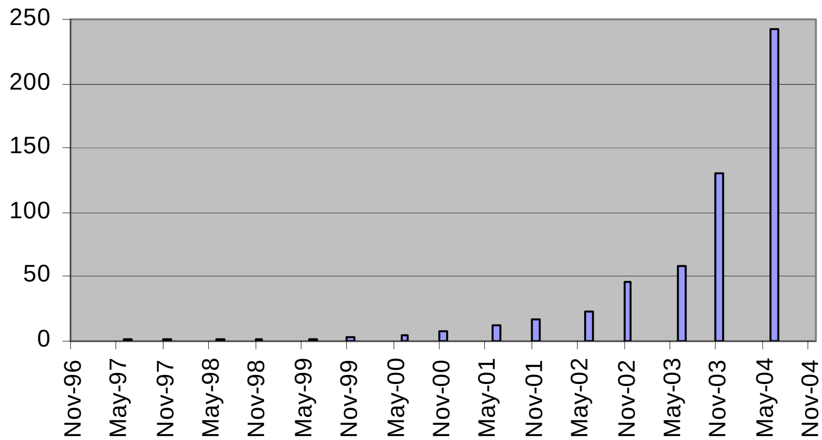
The Top242

- Focus on machines that are > 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

20

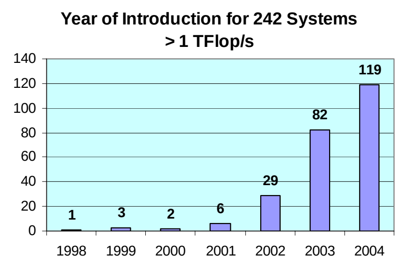
Number of Systems on Top500 > 1 Tflop/s Over Time

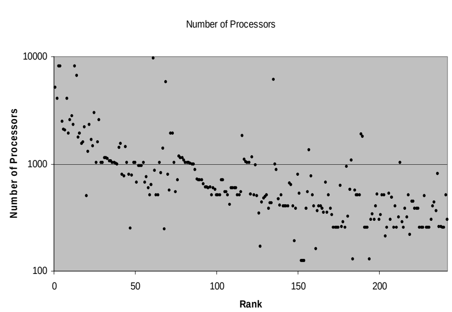


21

Factoids on Machines > 1 TFlop/s

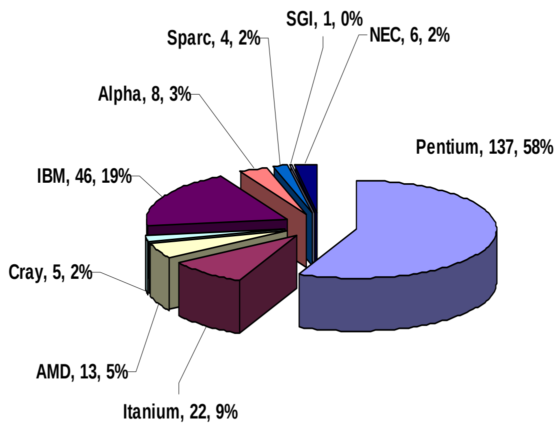
- 242 Systems
- 171 Clusters (71%)
- Average rate: 2.54 Tflop/s
- Median rate: 1.72 Tflop/s
- Sum of processors in Top242: 238,449
  - Sum for Top500: 318,846
- Average processor count: 985
- Median processor count: 565
- Numbers of processors
  - Most number of processors: 9632_{61}
    - ASCI Red
  - Fewest number of processors: 124_{152}
    - Cray X1

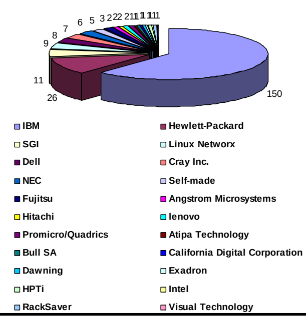




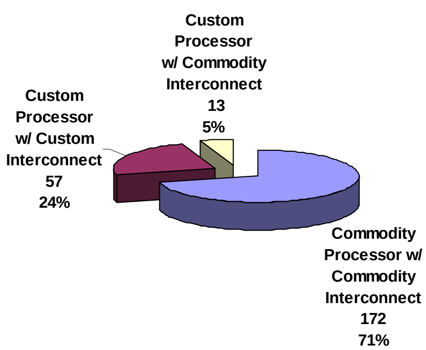
Percent Of 242 Systems Which Use The Following Processors > 1 TFlop/s

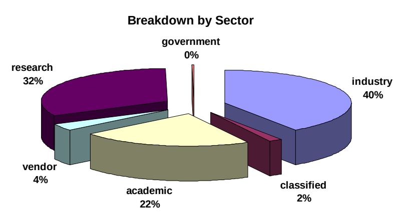
More than half are based on 32 bit architecture
11 Machines have a Vector instruction Sets





Percent Breakdown by Classes



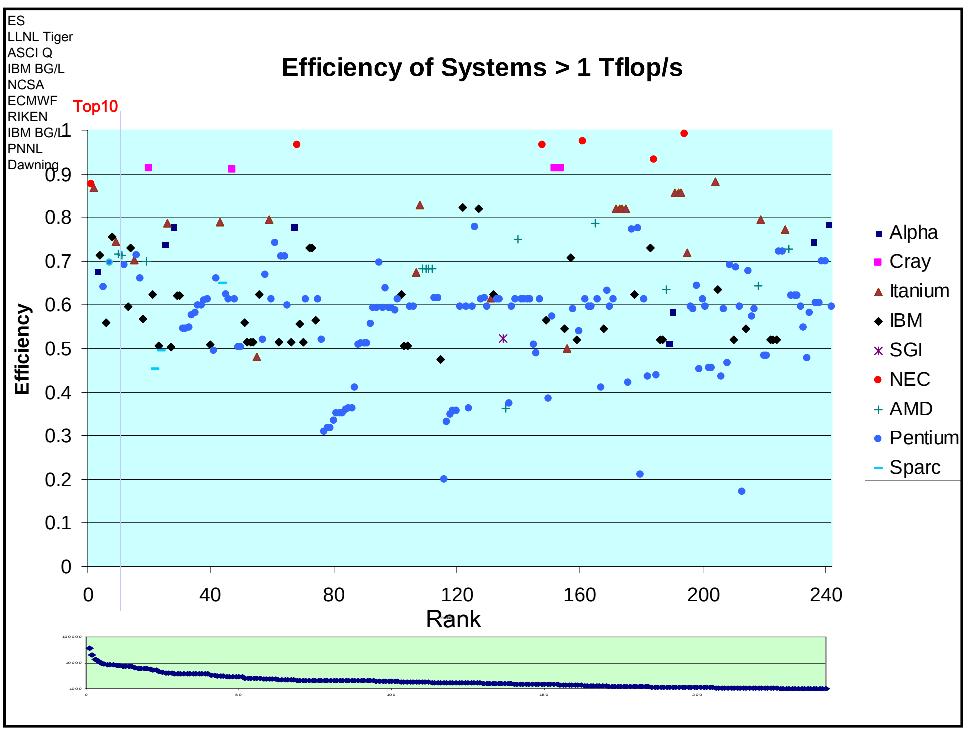


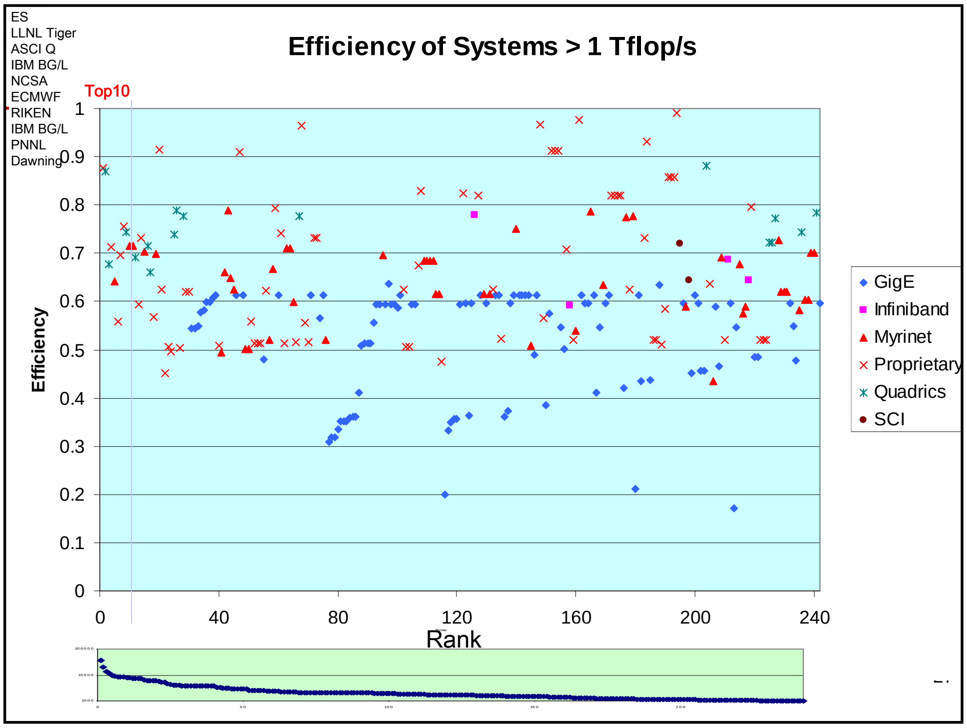
What About Efficiency?

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

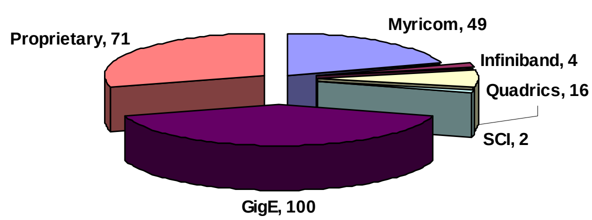
    …
- Remember this is $O(n^3)$ ops and $O(n^2)$ data
  - Mostly matrix multiply

25





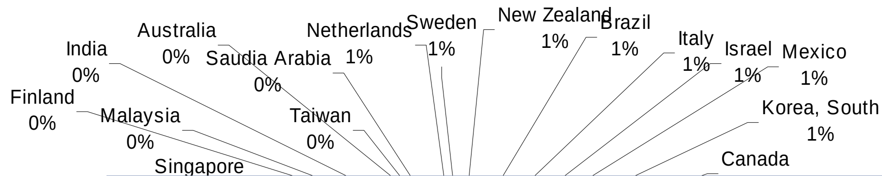
Interconnects Used in the Top242



Efficiency for Linpack

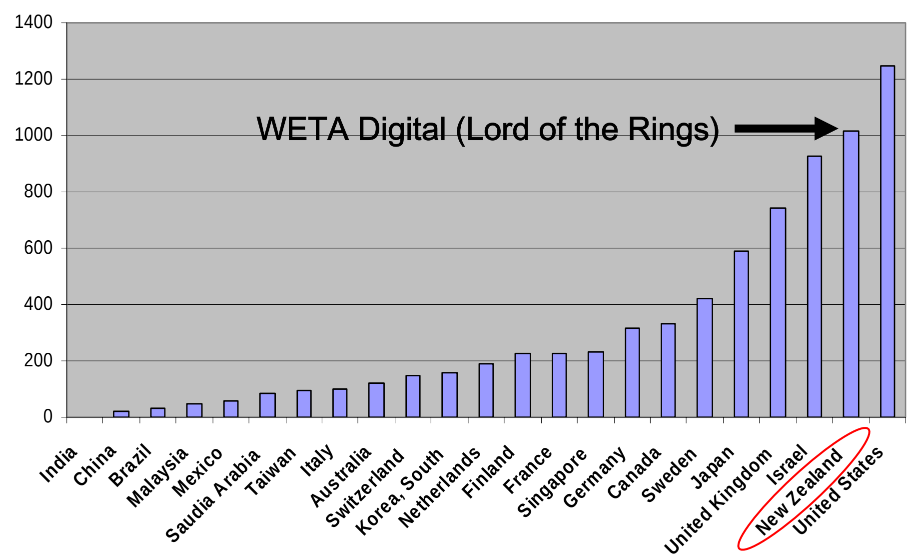
	Largest node count	min	max	average
GigE	1128	17%	64%	51%
SCI	400	64%	74%	68%
QsNetII	4096	66%	88%	75%
Myrinet	1408	44%	79%	64%
Infiniband	768	59%	78%	75%
Proprietary	9632	45%	99%	68%

Country Percent by Total Performance



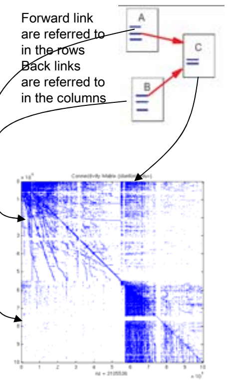
Rank	Site Country/Year	Computer / Processors Manufacturer	R_{max} R_{peak}
165	Umea University / HPC2N Sweden/2004	HP Opteron 2.2 GHz, Myrinet 384 HP	1329 1689.6
166	Evergrow Grid Sweden/2004	xSeries Xeon 3.06 GHz - Gig-E 352 IBM	1321.76 2154.24
198	National Supercomputer Centre (NSC) Sweden/2002	Pentium Xeon Cluster 2.2 GHz - SCI 3D 400 Self-made	1132 1760
263	Ericsson Sweden/2004	Integrity Superdome, 1.5 GHz, HPlex 192 HP	940.2 1152

KFlop/s per Capita (Flops/Pop)



Google

- Google query attributes
  - 150M queries/day (2000/second)
  - 100 countries
  - 4.2B documents in the index
- 60 Data centers
  - 100,000 Linux systems in data centers around the world
    - 15 TFlop/s and 1000 TB total capability
    - 40-80 1U/2U servers/cabinet
    - 100 MB Ethernet switches/cabinet with gigabit Ethernet uplink
  - growth from 4,000 systems (June 2000)
    - 18M queries then
- Performance and operation
  - simple reissue of failed commands to new servers
  - no performance debugging
    - problems are not reproducible



Eigenvalue problem; Ax = λx
$n=4.2\times10^9$
(see: MathWorks Cleve's Corner)

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

Source: Monika Henzinger, Google & Cleve Moler

Sony PlayStation2

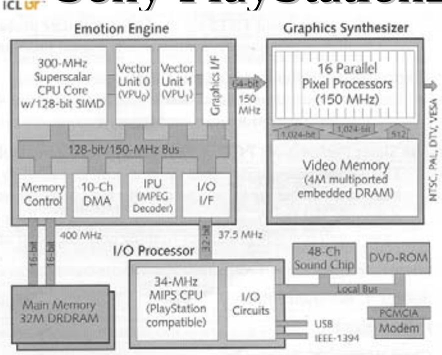
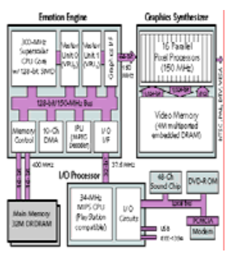


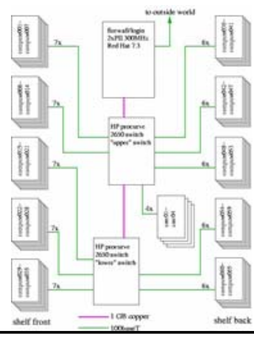
Figure 1. PlayStation 2000 employs an unprecedented level of parallelism to achieve workstation-class 3D performance.

- 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

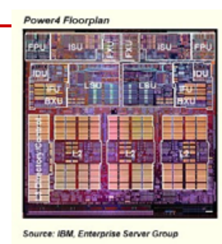
Computing On Toys

- **Sony PlayStation2**
  - 6.2 GF peak
  - 70M polygons/second
  - 10.5M transistors
  - superscalar RISC core
  - plus vector units, each:
    - 19 mul-adds & 1 divide
    - each 7 cycles
- ***$199 retail***
  - *loss leader for game sales*
- **100 unit cluster at U of I**
  - Linux software and vector unit use
    - over 0.5 TF peak
  - but hard to program & hard to extract performance …





Petascale Systems In 2008



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

34

Software Evolution and Faults

- Cost dynamics
  - people costs are rising
  - hardware costs are falling
- Two divergent software world views
  - parallel systems
    - life is good – *deus ex machina*
  - Internet
    - evil everywhere, trust no one, we'll all die horribly
- What does this mean for software?
  - abandon the pre-industrial "craftsman model"
  - adopt an "automated evolution" model

35

Fault Tolerance: Motivation

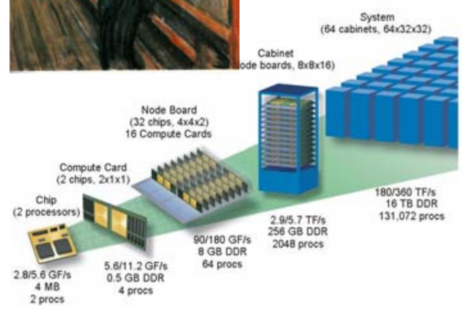
- Trends in HPC:
  - High end systems with thousand of processors
- Increased probability of a node failure
  - Most systems nowadays are robust
- MPI widely accepted in scientific computing
  - Process faults not tolerated in MPI model

Mismatch between hardware and (non fault-tolerant) programming paradigm of MPI.

36

Fault Tolerance in th ation

- Some next generation systems are being designed with 100K processors (IBM Blue Gene L)
- MTTF 10^5 - 10^6 hours for component
  - sounds like a lot until you divide by 10^5!
  - Failures for such a system is likely to be just a few hours perhaps minutes away.
- Application checkpoint /restart is today's typical fault tolerance method.
- Problem with MPI, no recovery from faults in the standard



- Many cluster based on commodity parts don't have error correcting primary memory
- Caches are not SECDED

37

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.

38

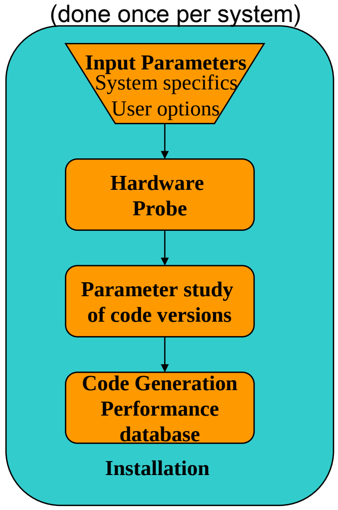
Motivation Self Adapting Numerical Software (SANS) Effort

- Optimizing software to exploit the features of a given system has historically been an exercise in hand customization.
  - Time consuming and tedious
  - Hard to predict performance from source code
  - Must be redone for every architecture and compiler
    - Software technology often lags architecture
    - Best algorithm may depend on input, so some tuning may be needed at run-time.
- There is a need for quick/dynamic deployment of optimized routines.

39

Performance Tuning Methodology

Software Installation
(done once per system)



Software Generation Strategy - ATLAS BLAS

- Parameter study of the hw
- Generate multiple versions of code, w/difference values of key performance parameters
- Run and measure the performance for various versions
- Pick best and generate library
- Optimize over 8 parameters
  - Cache blocking
  - Register blocking (2)
  - FP unit latency
  - Memory fetch
  - Interleaving loads & computation
  - Loop unrolling
  - Loop overhead minimization
- Similar to FFTW

40

Self Adapting Numerical Software - SANS Effort

- Provide software technology to aid in high performance on commodity processors, clusters, and grids.
- Pre-run time (library building stage) and run time optimization.
- Integrated performance modeling and analysis
- Automatic algorithm selection – polyalgorithmic functions
- Automated installation process
- Can be expanded to areas such as communication software and selection of numerical algorithms

Generic Code Optimization

- Follow on to ATLAS
  - Take generic code segments and perform optimizations via experiments
- Collaboration with ROSE project (source-to-source code transformation / optimization) at Lawrence Livermore National Laboratory
  - Daniel Quinlan and Qing Yi
  - LoopProcessor -bk3 4 -unroll 4 ./dgemv.c
  - We generate the test cases and also the timing driver.
- Also collaboration with Jim Demmel and Kathy Yelick at Berkeley under an NSF ITR effort.

Some Current Unmet Needs

- Performance / Portability
- Fault tolerance
- Better programming models
  - Global shared address space
  - Visible locality
- Maybe coming soon (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.

43

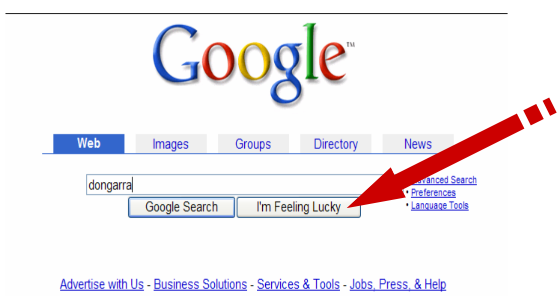
Collaborators / Support

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

LACSI

SciDAC Scientific Discovery through Advanced Computing

- Slides are online:
  - Google "dongarra"
  - Click on "talks"



44