SC2002, Baltimore (http://www.sc-conference.org/sc2002)
From the Earth Simulator to PC Clusters

Structure of SC2002
- **Top500 List**

Dinosaurs Department
- **Earth simulator –**
- **US -answers**
- **(Cray SX1, ASCI purple), BlueGene/L**

QCD computing – QCDOC, apeNEXT poster

Cluster architectures
- **Low voltage clusters, NEXCOM, Transmeta-NASA**
- **Interconnects – Myrinet, QsNet, Infiniband**
- **Large installations –LLNL, Los Alamos**

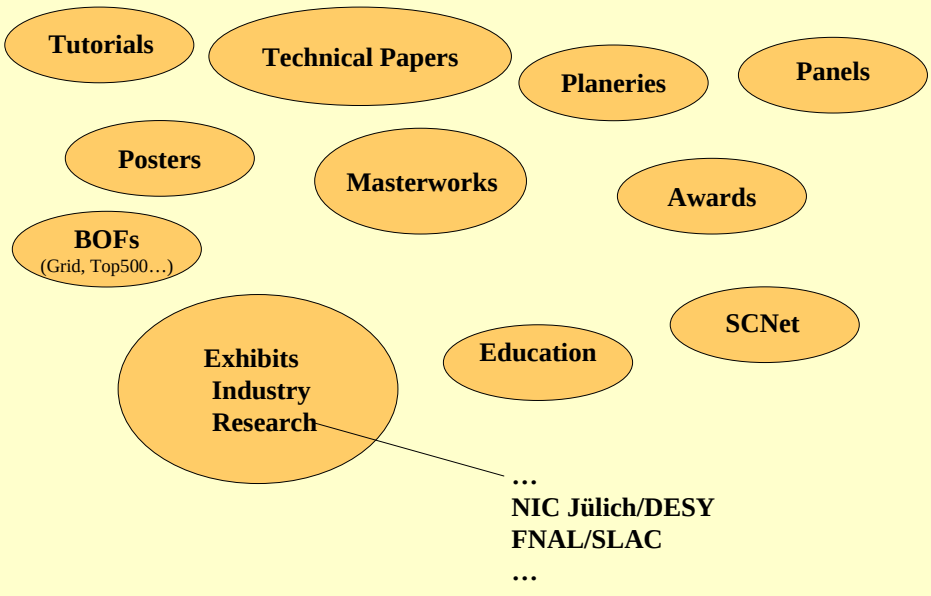
Cluster software
- **Rocks (NPACI), Linux BIOS (NCSA)**

Algorithms - David H. Bailey

Experiences - HPC in an oil company

Conclusions

SC2002
Structure

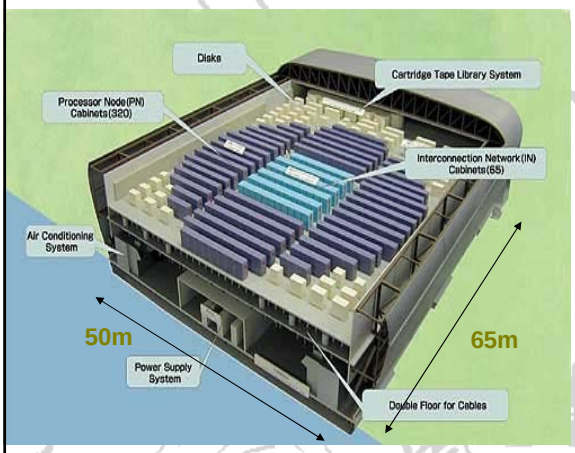


SC2002 Top500 list

Rank	Manufacturer Computer/Procs	R_{max} R_{peak}	Installation Site Country/Year	Inst. type Installation Area
1	**NEC,Vector** SX6 Earth-Simulator/ 5120	**35860.00** 40960.00	**Earth Simulator Center** Japan/2002	**Research**
2	**Hewlett-Packard** ASCI Q - AlphaServer SC ES45/1.25 GHz/ 4096	**7727.00** 10240.00	**Los Alamos National Laboratory** USA/2002	**Research**
3	**Hewlett-Packard** ASCI Q - AlphaServer SC ES45/1.25 GHz/ 4096	**7727.00** 10240.00	**Los Alamos National Laboratory** USA/2002	**Research**
4	**IBM** ASCI White, SP Power3 375 MHz/ 8192	**7226.00** 12288.00	**Lawrence Livermore National Laboratory** USA/2000	**Research** Energy
5	**Linux NetworX** MCR Linux Cluster Xeon 2.4 GHz - Quadrics/ 2304	**5694.00** 11060.00	**Lawrence Livermore National Laboratory** USA/2002	**Research**

http://www.top500.org/top5/2002/11/five/

SC2002 Dinosaurs – The Earth Simulator



640 nodes connected by a 640 x 640 single-stage crossbar switch

1 Node = 8 Arithm. Vector Processors + 16 GByte shared memory

Cost: > $ 350 Mio

SC2002 Dinosaurs – The Earth Simulator

In April 2002, the Earth Simulator became operational.
Peak performance of the Earth Simulator is 40 Teraflops (TF).

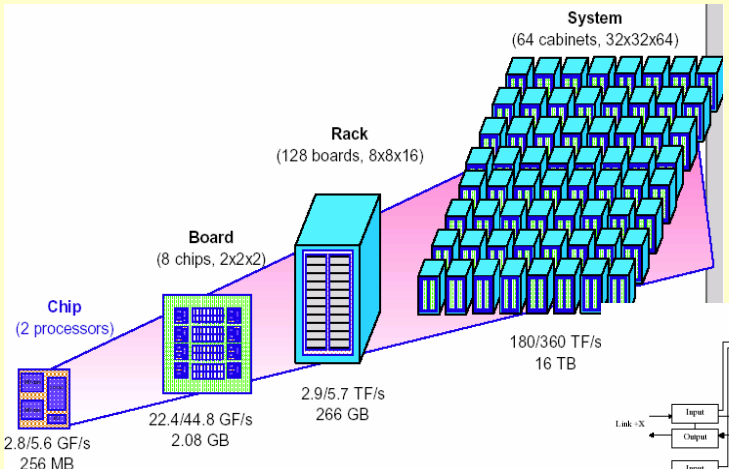
The Earth Simulator is the new No. 1 on the Top 500 list based on the LINPACK benchmark set (www.top500.org), it achieved a performance of 35.9 TF, or 90% of peak.

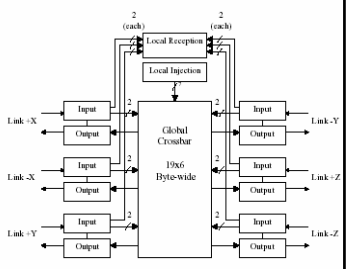
The Earth Simulator ran a benchmark global atmospheric simulation model at 13.4 TF on half of the machine, i.e. performed at over 60% of peak.

The total peak capability of all DOE (US Department of Energy) computers is 27.6 teraflops.
The Earth Simulator applies to a number of other disciplines such as fusion and geophysics as well.

SC2002 Dinosaurs – BlueGene/L

IBM and Lawrence Livermore National Laboratory
…A Gara, …, BD Steinmacher-Burow, …

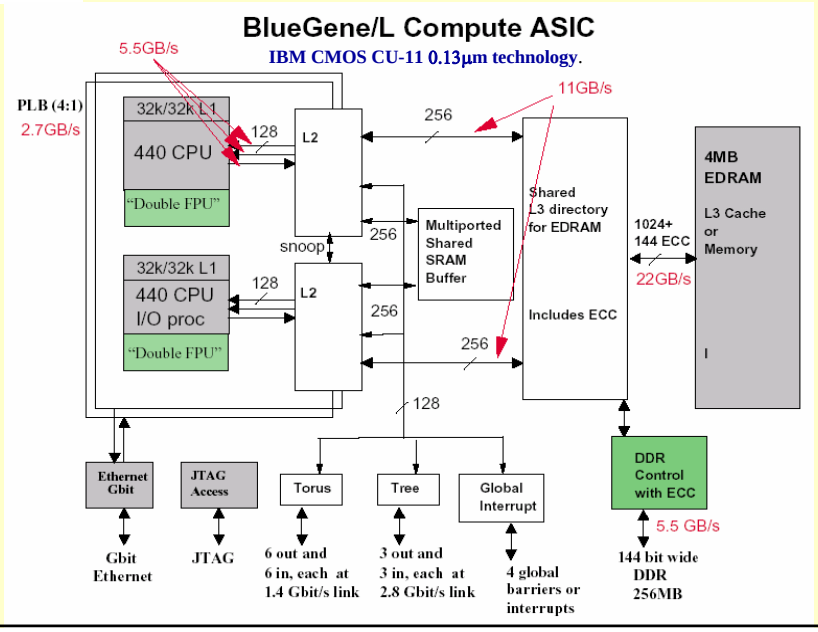




3 Dimensional Torus 32 x 32 x 64
Virtual cut-through hardware routing – adaptive, deadlock-free
1.4 Gb/s on 12 node links (total of 2.1 GB/s per node)

SC2002 Dinosaurs – BlueGene/L

"Big Brother of QCDOC"



SC2002 Dinosaurs – BlueGene/L

Machine Peak Speed (Tflop/s)	**180 / 360***
Total Memory (Tbytes)	**16–32**
Footprint (m^2)	**230**
Total Power (MW)	**1.2**
Cost (M$)	**<< $100 Mio**
Installation Date	**~12/2004**
No. of Nodes	**65,536**
CPUs per Node	**2**
Clock Frequency (MHz)	**700**
Power Dissipation/Node (W)	**15**
Peak Speed/Node (Gflop/s)	**2.8**
Memory/Node (GB)	**0.25–0.5**
MPI Latency (µs)	**7**

Lattice QCD - QCDOC

Columbia University
RIKEN/Brookhaven National Laboratory (BNL)
UKQCD
IBM

MIMD machine with distributed memory system-on-a-chip design

(QCDOC = QCD on a chip)Technology:

IBM SA27E = CMOS 7SF = 0.18 μm lithography process

ASIC combines existing IBM components and QCD-specific, custom-designed logic:

500 MHz PowerPC 440 processor core with 64-bit, 1 GFlops FPU

4 MB on-chip memory (embedded DRAM)

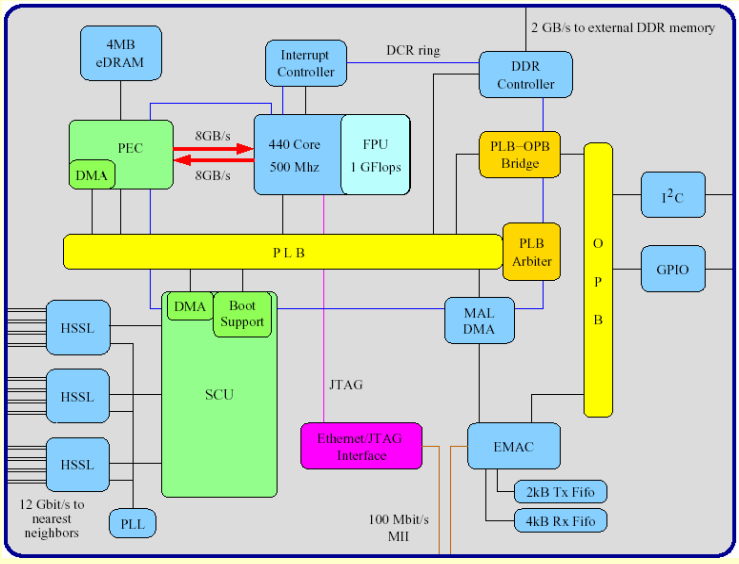
Nearest-neighbor serial communications unit (SCU)

6-dimensional communication network (4-D Physics, 2-D partitioning)

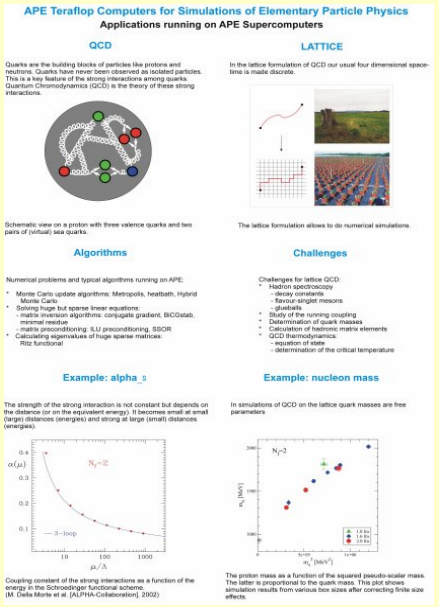
Silicon chip of about 11 mm square, consuming 2 W

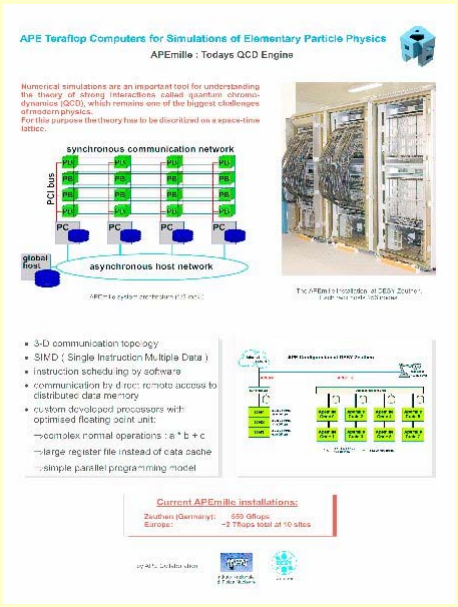
Lattice QCD – QCDOC ASIC

Processor cores + IBM CoreConnect Architecture

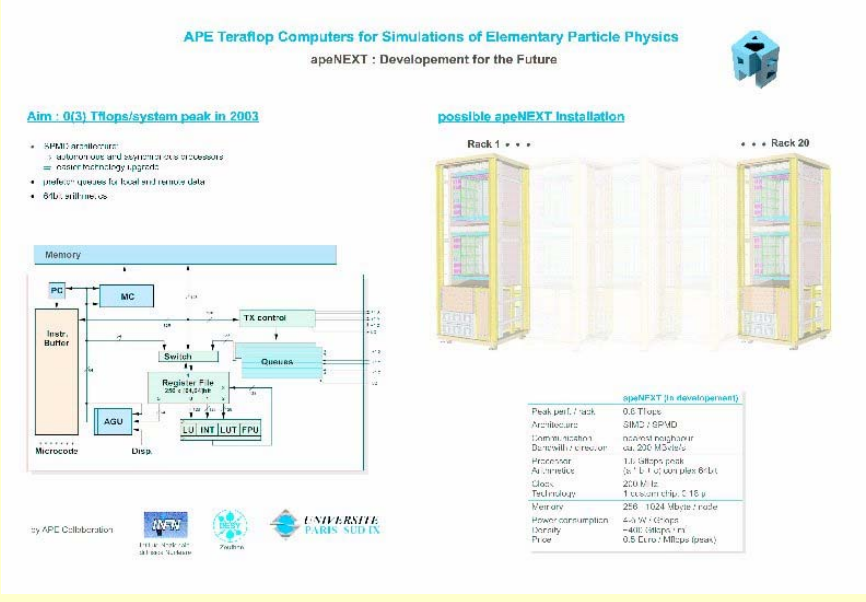


Lattice QCD, NIC/DESY Zeuthen poster : APEmille

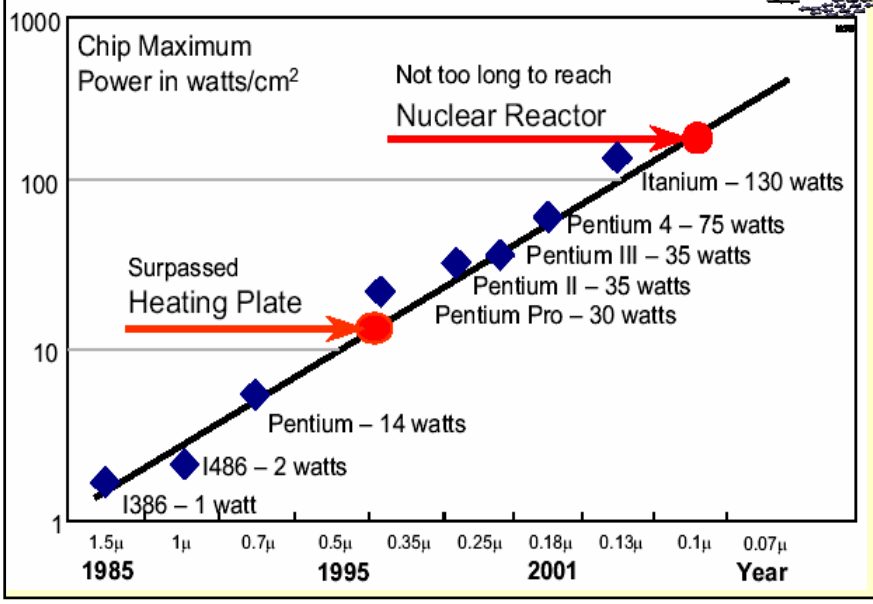




Lattice QCD, NIC/DESY Zeuthen poster : apeNEXT



Low Power Cluster Architectures
sensitivity to power consumption



Low Power Cluster Architectures
good old Beowulf approach - LANL

'Green Destiny' could be a model for future high-performance computing ?!

- **240 Transmeta TM5600 CPUs (667 MHz) mounted onto blade servers.**
- **24 blades then mount into a RLX Technologies System 324 chassis**
- **10 chassis, with network switches, are mounted in a standard computer rack.**

Peak rate of 160 Gflops, Sustained - average of 38.9 Gflops on a 212 nodes system in a gravitational treecode N-body simulation of galaxy formation using 200 million particles.

Power dissipation

1 Blade:	**22W**
240 node Cluster:	**5.2 kW**

Cluster Architectures
Blade Servers

NEXCOM – Low voltage blade server
200 low voltage Intel XEON CPUs (1.6 GHz – 30W) in a 42U Rack
Integrated Gbit Ethernet network

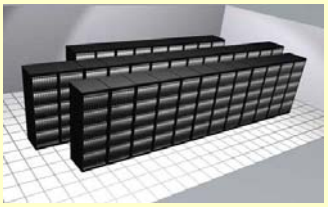
Mellanox – Infiniband blade server

Single XEON Blades connected via a 10 Gbit (4X) Infiniband network

NCSA, Ohio State University

Top500 Cluster

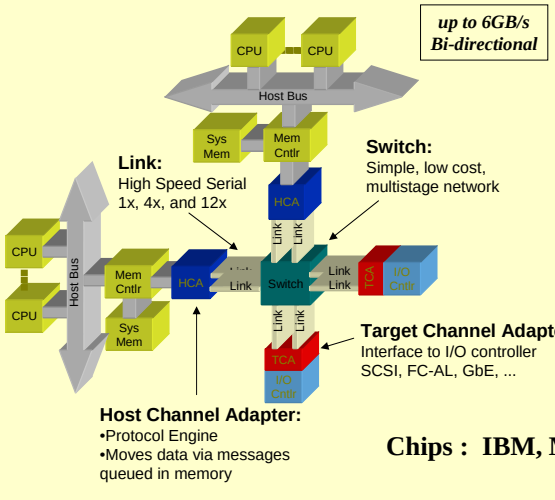
MCR LINUX CLUSTER
LLNL, LIVERMOORE
LINUX NETWORX/QUADRICS
Rmax: 5.69 TFlops



- **11.2 Tflops Linux cluster**
- **4.6 TB of aggregate memory**
- **138.2 TB of aggregate local disk space**
- **1152 total nodes plus separate hot spare cluster and development cluster**
- **2,304 Intel 2.4 GHz Xeon processors**
- **Cluster File Systems, Inc. supplied the Lustre Open Source cluster wide file system**
- **Cluster connect: QsNet ELAN3 by Quadrics,**
- **4 GB of DDR SDRAM memory per node, 120 GB Disk Space.**

A similar cluster with a Myrinet connection was announced for the Los Alamos National Laboratory, at Fermilab planned for 2006

Clusters: Infiniband interconnect

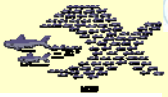


http://www.infinibandta.org

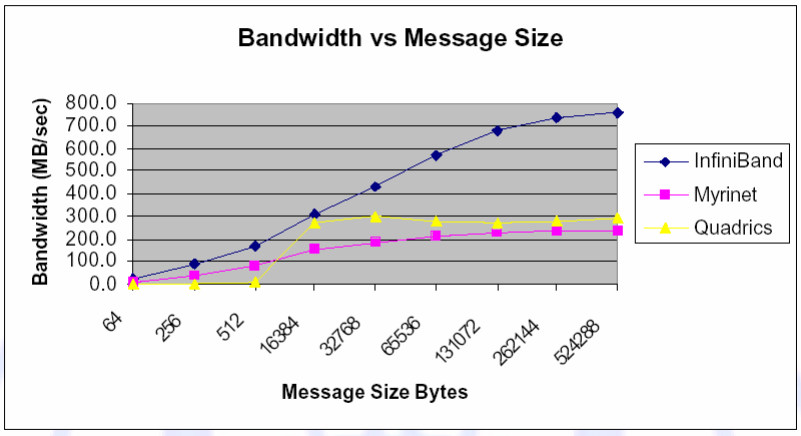
Chips : IBM, Mellanox

PCI-X cards: Fujitsu, Mellanox, JNI, IBM

Clusters Infiniband interconnect



InfiniBand MPI Throughput Comparison



Source: Ohio State University, Xeon 2.2 GHz up processor platform

Cluster/Farm Software

NPACI Rocks (National Partnership for Advanced Computational Infrastructure)

Open-source enhancements to Red Hat Linux (uses RedHat Kickstart).
100% automatic Installation – Zero hand configuration
One CD installs all servers and nodes in a cluster (PXE)
Entire Cluster-Aware Distribution
Full Red Hat release De-facto standard cluster packages (e.g., MPI, PBS)
NPACI Rocks packages
Initial configuration via simple web page, integrated monitoring (Ganglia)
Full re-installation instead of configuration management (cfengine)
Cluster configuration database (based on MySQL)

Sites using NPACI Rocks

Germany:	ZIB Berlin, FZK Karlsruhe
US:	SDSC, Pacific Northwest National Laboratory, Nothwestern University, University of Texas, Caltech, …

Cluster/Farm Software

LinuxBIOS

Replaces the normal BIOS found on Intel-based PCs, Alphas, and other machines with a Linux kernel that can boot Linux from a cold start.
Primarily Linux—about 10 lines of patches to the current Linux kernel.
Additionally, the startup code—about 500 lines of assembly and 1500 lines of C—executes 16 instructions to get into 32-bit mode and then performs RAM and other hardware initialization required before Linux can take over.
Provides much greater flexibility than using a simple netboot.

LinuxBIOS is currently working on several mainboards

http://www.acl.lanl.gov/linuxbios/

PC Farm real life example

High Performance Computing in a Major Oil Company (British Petroleum)
Keith Gray

Dramatically increased computing requirements in support of seismic imaging researchers

SGI-IRIX, SUN-Solaris, IBM-AIX, SUN-Solaris
→ PC-Farm-Linux

Hundreds of Farm PCs, Thousands of desktops
GRD (SGEEE) batch system,
cfengine - configuration update

Network Attached Storage

SAN – too expensive switches

Algorithms

High Performance Computing Meets Experimental Mathematics
David H. Bailey

The PSLQ Integer Relation Detection Algorithm :
Recognize a numeric constant in terms of the formula that it satisfies.
PSLQ is well-suited for parallel computation, and lead to several examples of new mathematical results, some of these computations performed on highly parallel computers, since they are not feasible on conventional systems (e.g. the identification of Euler-Zeta Sum Constants).

New software package for performing arbitrary precision arithmetic, which is required in this research:

http://www.nersc.gov/~dhbailey/mpdist/index.html

SC2002 Conclusions

•Big supercomputer architectures are back – Earth simulator, Cray SX1, BlueGene/L

•Special architectures for LQCD computing are continuing

•Clusters are coming, number of special cluster vendors providing hard and software is increasing – LinuxNetworx (former AltaSystems), RackServer, …

•Trend to blade servers – HP, Dell, IBM, NEXCOM, Mellanox …

•Intel Itanium Server processor widely accepted – surprising success (also for Intel) at SC2002

•Infiniband could be a promising alternative to the special High Performance Link products Myrinet and QsNet

•There may come standard tools for cluster and farm handling – NPACI Rocks, Ganglia, LinuxBios (for RedHat)

•Batch systems: increasing number of SGE(EE) users (Ratheon)

•Increasing number of GRID projects (no one mentioned here – room for another talk)