



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



Tutorials

Technical Papers

Planeries

Panels

Posters

Masterworks

Awards

BOFs

(Grid, Top500...)

Exhibits
Industry
Research

Education

SCNet

...
NIC Jülich/DESY
FNAL/SLAC
...

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.

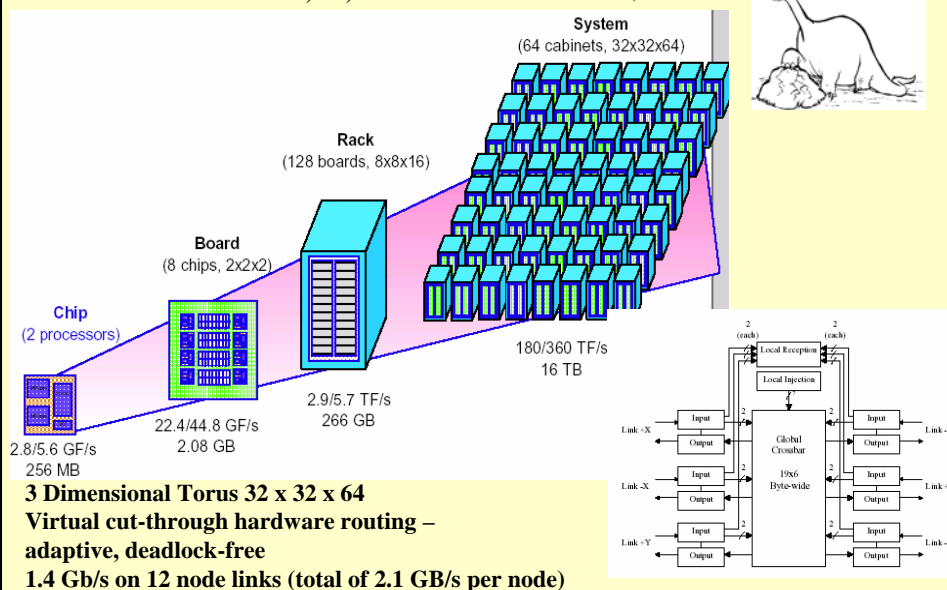


SC2002 Dinosaurs – BlueGene/L



IBM and Lawrence Livermore National Laboratory

...A Gara, ..., BD Steinmacher-Burow, ...





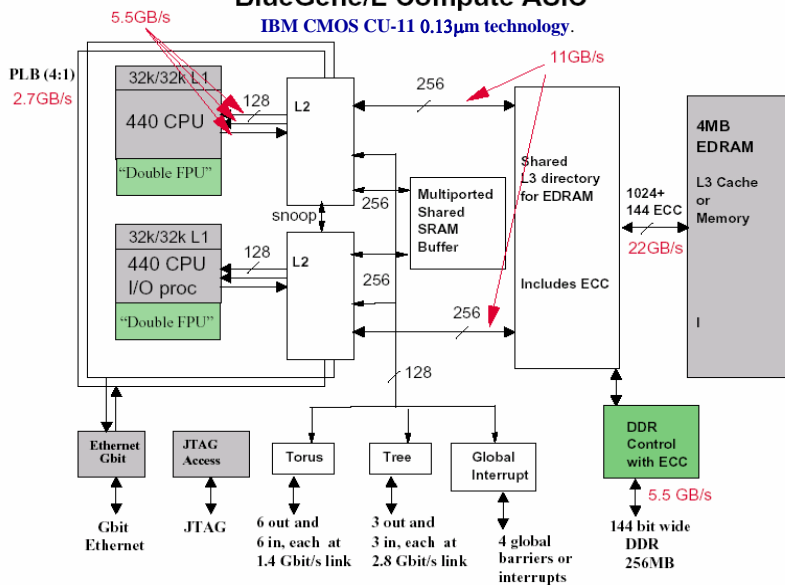
SC2002 Dinosaurs – BlueGene/L



“Big Brother of QDOC”

BlueGene/L Compute ASIC

IBM CMOS CU-11 0.13µm technology.



SC2002 Dinosaurs – BlueGene/L



Machine Peak Speed (Tflop/s)	180 / 360*
Total Memory (Tbytes)	16–32
Footprint (m²)	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)

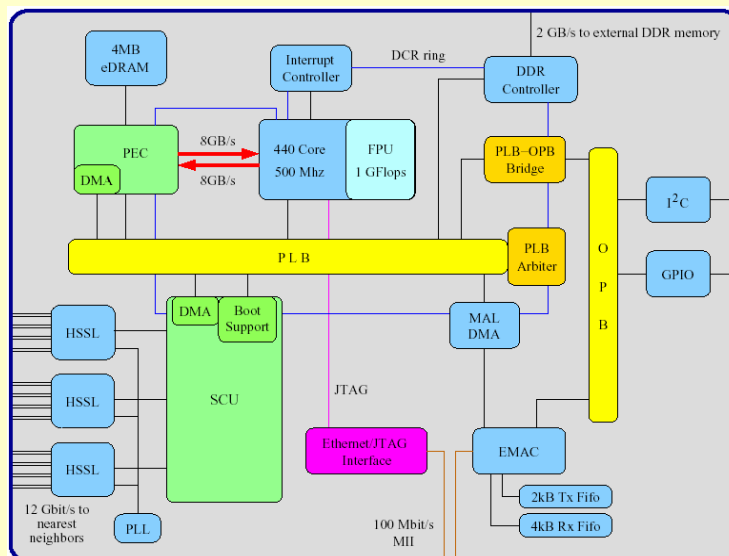
Silicon chip of about 11 mm square, consuming 2 W



Lattice QCD - QCDOC ASIC



Processor cores + IBM CoreConnect Architecture





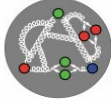
Lattice QCD, NIC/DESY Zeuthen poster : APEmille



APE TeraFlop Computers for Simulations of Elementary Particle Physics Applications running on APE Supercomputers

QCD

Quarks are the building blocks of particles like protons and neutrons. Quarks have never been observed as isolated particles. This is a key feature of the strong interactions among quarks. Quantum Chromodynamics (QCD) is the theory of these strong interactions.



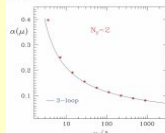
Schematic view of a proton with three valence quarks and two pairs of (virtual) sea quarks.

Algorithms

- Numerical problems and typical algorithms running on APE:
- Monte Carlo update algorithms: Metropolis, heatbath, Hybrid Monte Carlo
 - Solving large but sparse linear equations:
 - matrix inversion algorithms: conjugate gradient, BiCGstab, minimal residual
 - matrix preconditioning: LU preconditioning, SSOR
 - Calculating eigenvalues of large sparse matrices: Ritz functional

Example: α_s

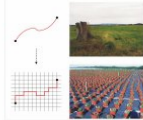
The strength of the strong interaction is not constant but depends on the distance (or on the equivalent energy). It becomes small at small (large) distances (energies) and strong at large (small) distances (energies).



Coupling constant of the strong interactions as a function of the energy in the Schrödinger functional scheme (M. Della Morte et al. (ALICE-Collaboration), 2002)

LATTICE

In the lattice formulation of QCD our usual four dimensional space-time is made discrete.



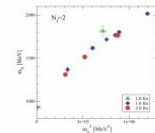
The lattice formulation allows to do numerical simulations.

Challenges

- Challenges for lattice QCD:
- Hadron spectroscopy:
 - identify resonances
 - Resonance singlet mesons
 - glueballs
 - Study of the running coupling
 - Determination of quark masses
 - Calculation of hadronic matrix elements
 - QCD thermodynamics:
 - equation of state
 - determination of the critical temperature

Example: nucleon mass

In simulations of QCD on the lattice quark masses are free parameters.

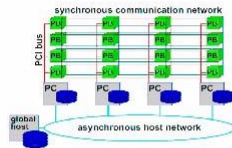


The nucleon mass as a function of the squared pseudo-scalar mass. The latter is proportional to the quark mass. This plot shows simulation results from various box sizes after correcting finite size effects.

APE TeraFlop Computers for Simulations of Elementary Particle Physics APEmille : Today's QCD Engine



Numerical simulations are an important tool for understanding the theory of strong interactions, called quantum chromodynamics (QCD), which remains one of the biggest challenges of modern physics. For this purpose the theory has to be discretized on a space-time lattice.



The APEmille computer racks at DESY Zeuthen. Each rack contains 100 nodes.

- 3-D communication topology
- SIMD (Single Instruction Multiple Data)
- Instruction scheduling by software
- communication by direct remote access to distributed data memory
- custom developed processors with:
 - optimized floating point units
 - complex normal operations : a * b + c
 - large register file instead of data cache
 - simple parallel programming model



Current APEmille installations:
Zeuthen (Germany): 850 Groups
Europe: ~2 Tflops total at 10 sites



Lattice QCD, NIC/DESY Zeuthen poster : apeNEXT



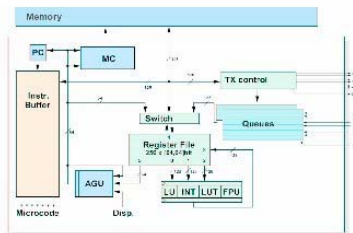
APE TeraFlop Computers for Simulations of Elementary Particle Physics

apeNEXT : Development for the Future



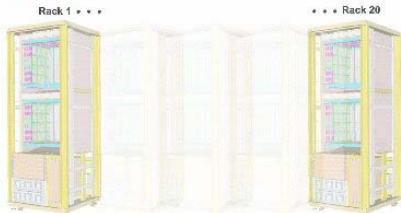
Aim : 0(3) Tflops/system peak in 2003

- APPE architecture:
 - silicon based and architectural O(3) processors
 - novice technology (upgrate)
- performance goals for local and remote data
- 8Tfl. still in use



by APC Collaboration
APE, DESY, NIC, UNIL, UNIV. PARIS SUD IX, Zeuthen

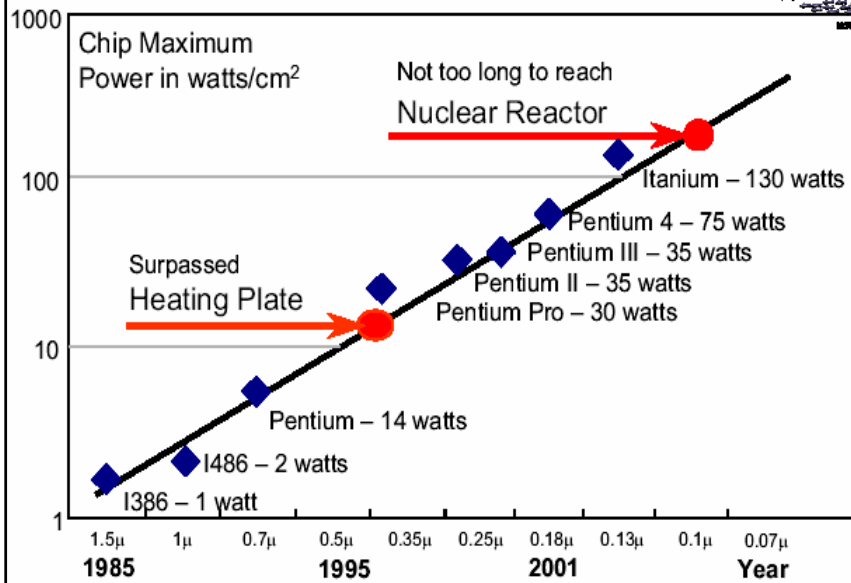
possible apeNEXT Installation



	apeNEXT (in development)
Peak perf./ rack	0.6 Tflops
Architecture	SIMD / SPMD
Component	novel neighbour
Bandwidth / connection	ca. 200 MB/s
processor	1.1 Gflops/peak
Architecture	32 bit / 4 word pipeline
Clock	200 MHz
Technology	1.5 micron chip, C 15 µ
Memory	256 - 1024 MByte / node
Power consumption	4-6 W / group
Density	~400 (chip/cm²)
Price	0.6 Euro / Mflops (peak)



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.

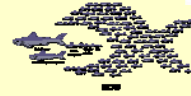
Power dissipation

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



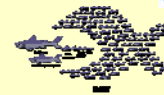
Mellanox – Infiniband blade server

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

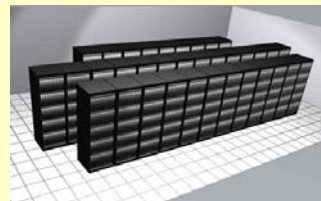
NCSA, Ohio State University



Top500 Cluster



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

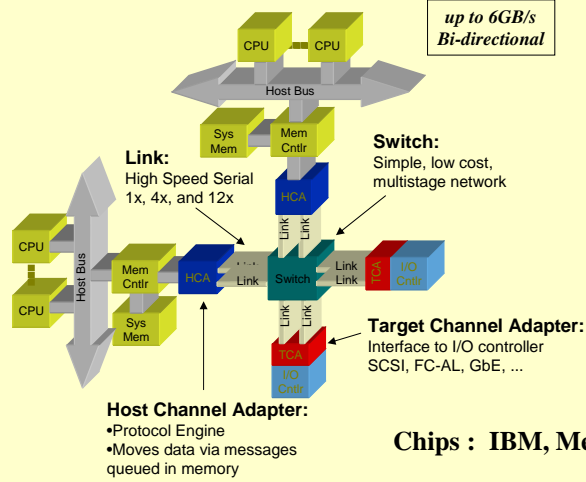


- 11.2 Tflops Linux cluster
- 4.6 TB of aggregate memory
- 138.2 TB of aggregate local disk space
- 1152 total nodes plus separate hot spare cluster and development cluster
- 2,304 Intel 2.4 GHz Xeon processors
- Cluster File Systems, Inc. supplied the Lustre Open Source cluster wide file system
- Cluster 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

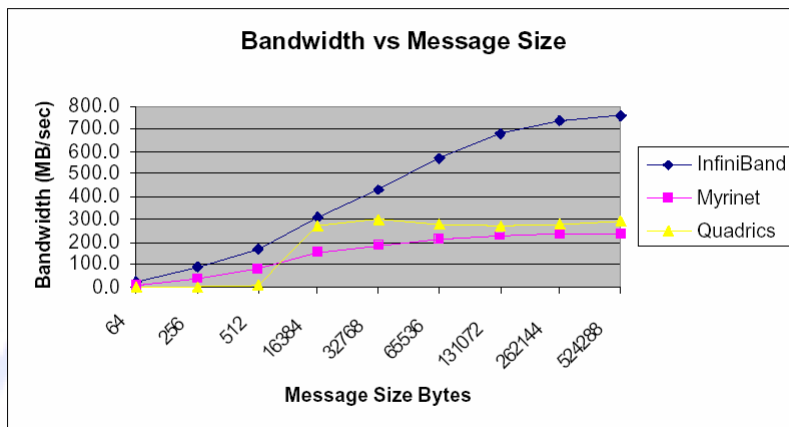
PCI-X cards: Fujitsu, Mellanox, JNL, IBM



Clusters Infiniband interconnect



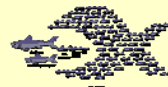
InfiniBand MPI Throughput Comparison



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



Cluster/Farm Software



NPACI Rocks

(National Partnership for Advanced Computational Infrastructure)

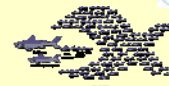
Open-source enhancements to Red Hat Linux (uses RedHat Kickstart).
100% automatic Installation – Zero hand configuration
One CD installs all servers and nodes in a cluster (PXE)
Entire Cluster-Aware Distribution
Full Red Hat release De-facto standard cluster packages (e.g., MPI, PBS)
NPACI Rocks packages
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,
Northwestern 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)