ALiCE – a Alpha Workstation Cluster for Lattice–QCD

Norbert Eicker

Department of Physics

University of Wuppertal

DESY Zeuthen: Technical Seminar

November 2nd 1999

http://www.theorie.physik.uni-wuppertal.de/∼eicker/zeuthen.ps

ALiCE

The Wuppertal Alpha-Linux-Cluster-Engine

Nodes:	128 Compaq DS10 Alpha workstations
Processors:	64×466 MHz EV6 / 64×667 MHz EV67
Cache:	2 MByte / 4 MByte
Memory:	128 MByte/node
Discs:	10 GByte/node
Connectivity:	64bit Myrinet with multistage crossbar

ALiCE Group

Wuppertal

A. Frommer	Numerical analyst, head IAI
K. Schilling	Group leader Computational Particle Physics
Th. Lippert	ALiCE concept
P. Ueberholz	System Administrator
N. Eicker	Linux, benchmarks

NIC

Ch. Best	NICse-TS & NICse

Collaboration with MIT and Jefferson Labs

J. Negele	Director Institute for Nuclear Physics MIT
C. Watson	Director CC Jefferson National Laboratories

Outline

I. Wuppertal Parallel Computing History

II. Benchmarks

III. Hardware

- Nodes
- Networks

IV. Software

- Compiler
- Networking

V. Summary

I. Wuppertal Parallel Computing History

- 1990 — 8k node Connection Machine CM2, operated by IAI
- 1993 — 32 node Connection Machine CM5 (5.6 Gflops peak)
- 1995 — 32 node Quadrics Q4, operated by Lattice QCD group (1.6 Gflops peak)
- 1997 — 64 proc QCDSP, operated by Lattice QCD group (3.2 Gflops peak)
- Fields — Computational Field Theory, Linear Algebra, Robotic Sim., Chemistry, Medical Imaging, Teaching HPC

- Mid '97 HBFG proposal: follow up system for CM5
- Mid–End '98 Benchmarks on commercial platforms
- Aug. '98 MIT/Wuppertal Project: Investigation of Clusters
- Sep. '98 First 4-node cluster at MIT
- Nov. '98 Cluster as follow up system for CM5 ?
- Jan. '99 Second 4-node cluster at NIC-Jülich (NICse-TS)
- Jan.–May '99 ALiCE Proposal / Benchmarks
- June '99 DFG recommends Alpha-Linux-Cluster

NICse-TS

- DEC Alpha 21164 EV56 533MHz.
- 2 MByte 3rd level cache.
- 256 MByte ECC SDRAM.
- Samsungs 164UX motherboards.
- 10 GB EIDE-harddisk.
- 4 D-Link DFE 500 TX fast Ethernet cards (DEC tulip 21140 based).
- Interconnected in torus topology.

System inspired by Avalon.

- Aug. '99 — Final bidding:
  - Compaq (128 DS10)
  - QuantX/API (48 UP2000)
  - IBM (SP-2 with Power3)
  - SUN (EP6000 / R24)
- We are heading for Compaq!
- Nov. '99 — First half of ALiCE (EV6 @ 466 MHz / 2MB cache)
- Apr. '00 — Second half of ALiCE (EV67 @ >667 MHz / 4MB)

II. Benchmarks

- **BAND** — Reduction of band matrix to bi-diagonal matrix (Scientific Computing): Efficiency of BLAS-2, scaling with fixed granularity, small messages
- **MULT** — Distributed (systolic) matrix multiplication (Parallel Algorithms): BLAS-3, max. global and local performance, dense matrix problems, scaling with fixed total size
- **SOLV** — Krylov solver with SSOR preconditioning (Numerical Field Theory): Memory-to-cache problems, short range interaction problems

SOLVE

Reference Application: SSOR-BICGStab solver for Lattice QCD

Wilson-Dirac linear system: > 90 % of qcd compute effort

$$MX = \phi$$

$$\begin{aligned} M &= \mathbf{I} - \kappa D, \qquad M \in \mathbb{C}^{12V \times 12V}, \\ D_{\mathsf{x},\mathsf{y}} &= \sum_{\mu=1}^{4} (\mathbf{I} - \gamma_\mu) \otimes U_\mu(\mathsf{x}) \otimes \delta_{\mathsf{x},\mathsf{y}-\mu} \\ &\qquad + (\mathbf{I} + \gamma_\mu) \otimes U_\mu^\dagger(\mathsf{x} - \mu) \otimes \delta_{\mathsf{x},\mathsf{y}+\mu}. \end{aligned}$$

- Two implementations: C and F77 both using MPI.
- About the same performance on T3E.
- `linsolv_BiCGStab`
  - no preconditioning.
  - communication not hidden behind computation.
- `linsolv_SSOR`
  - also BiCGStab, but with SSOR preconditioning.
  - slightly more FP operations and memory accesses.
  - communication hidden behind computation
- There are also versions without MPI.

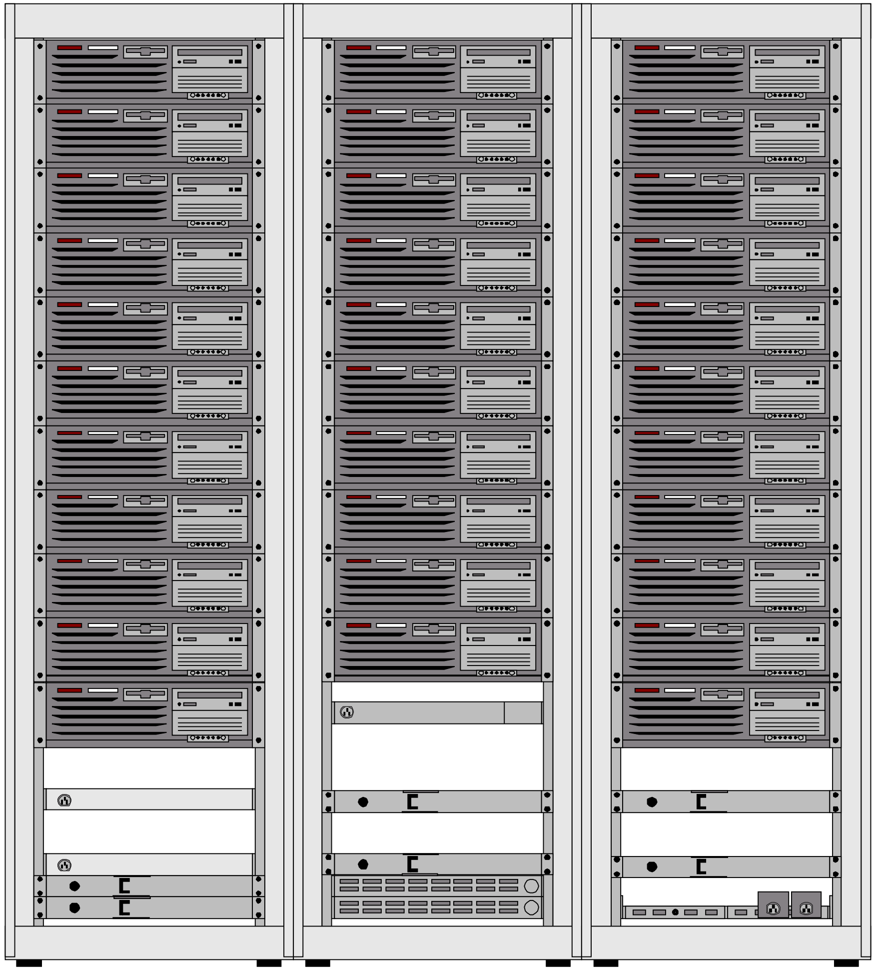
Results

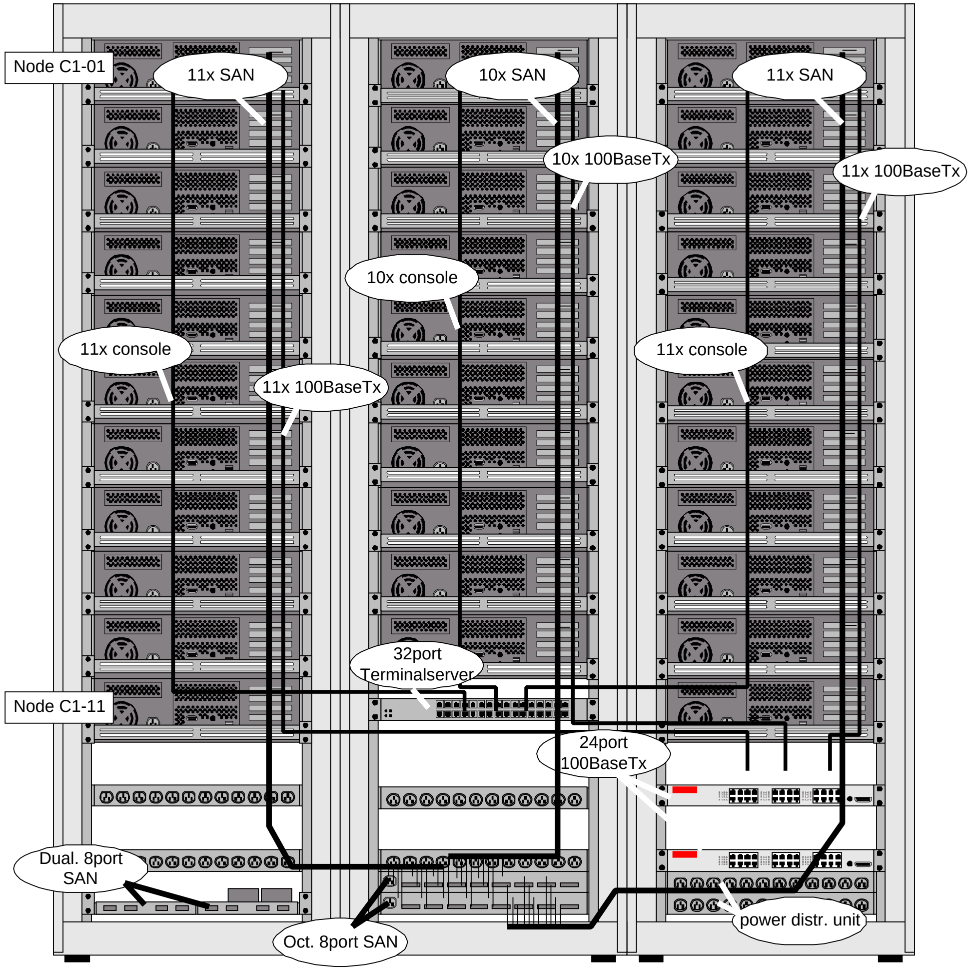
- Proprietary Systems Performance of SOLV not satisfactory !
- PC clusters (1998) Beowulf and Avalon: fast ethernet connectivity
  up to 256 nodes

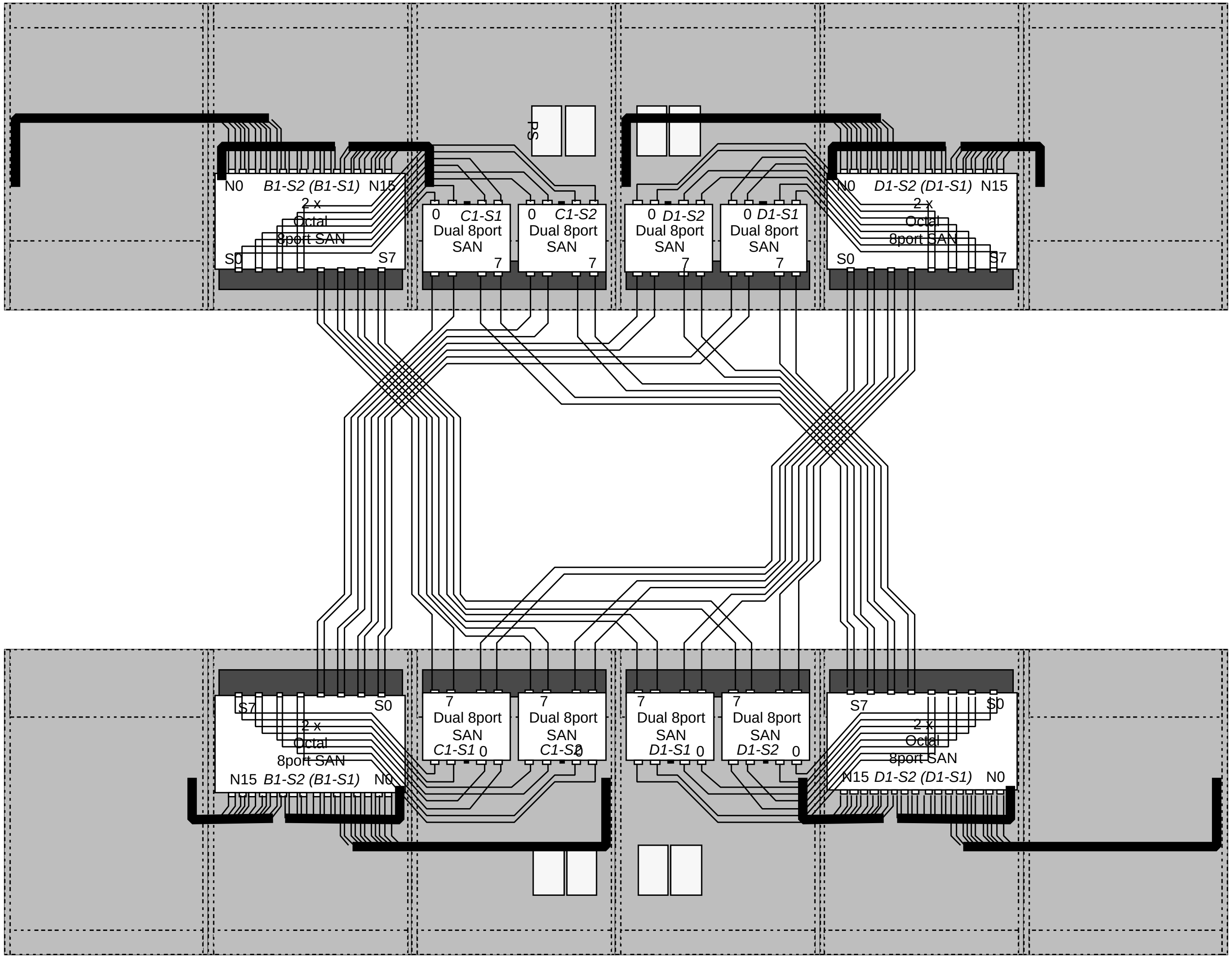
  $\rightarrow$ \$ 10 - \$ 20 / Mflop for Lattice QCD ?
- NICse-TS 6 % efficiency for QCD reference benchmark

  128 node cluster: \$ 100 / Mflop
- ALiCE \$ 50 / Mflop
- T3E, IBM, SUN etc. roughly \$ 150 –\$300 / Mflop
  (IBM's final bid: \$ 85 / MFlop)

III. Hardware






PS
N0
B1-S2 (B1-S1)
N15
2 x
Octal
8port SAN
S0
S7
0
C1-S1
Dual 8port
SAN
7
0
C1-S2
Dual 8port
SAN
7
0
D1-S2
Dual 8port
SAN
7
0
D1-S1
Dual 8port
SAN
7
N0
D1-S2 (D1-S1)
N15
2 x
Octal
8port SAN
S0
S7
S7
S0
2 x
Octal
8port SAN
N15
B1-S2 (B1-S1)
N0
7
Dual 8port
SAN
C1-S1
0
7
Dual 8port
SAN
C1-S2
0
7
Dual 8port
SAN
D1-S1
0
7
Dual 8port
SAN
D1-S2
0
S7
S0
2 x
Octal
8port SAN
N15
D1-S2 (D1-S1)
N0

More facts

- Peak performance: 150 GFlops.
- Total memory: 16 GB.
- Storage: 1.2 TB.
- Three networks:
  - Serial network
  - Fast Ethernet
  - MyriNet

Nodes

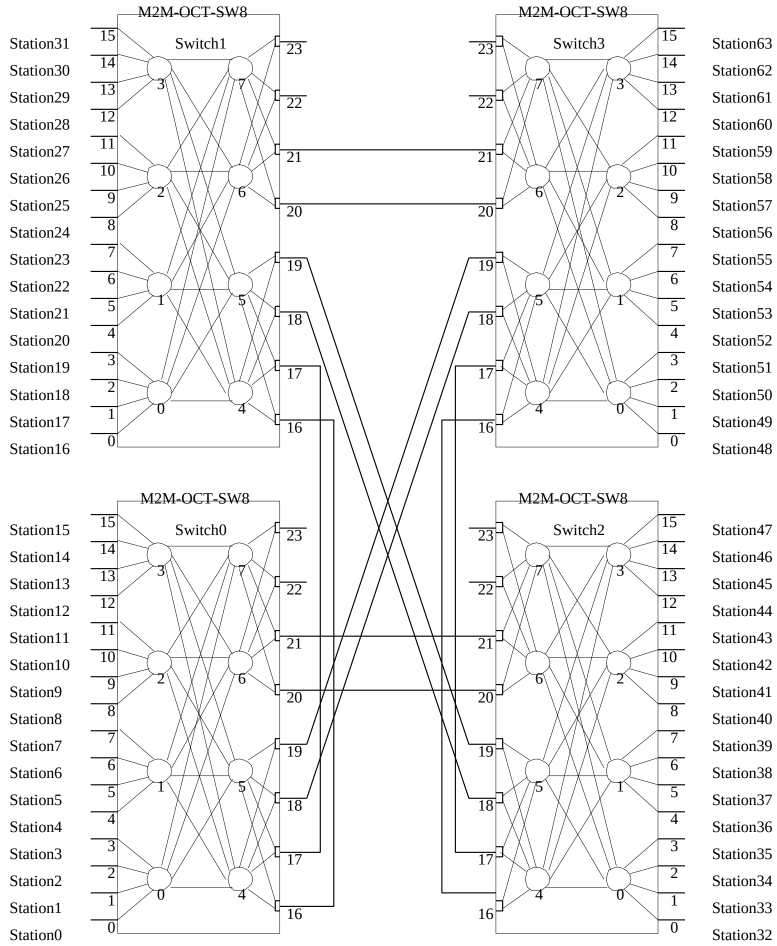
- Compaq DS10 (rackmounted).
- Alpha EV6 @466MHz (EV67 @667MHz).
- 2MB (4MB) level 2 cache.
- 128MB ECC RAM @ 1.3 GB/sec memory to CPU bandwith.
- 10 GB IDE disc.
- 3× 64bit, 1× 32bit PCI slots @33MHz.
- 64bit MyriNet SAN interface (1.28 Gbit/sec).

Networks

- Serial network:
  - 4×32 port DECserver 900TM terminal server.
  - All consoles available on one managment-station.
  - Used for booting, maintenance.
  - Backup network.
- Fast Ethernet:
  - 6×24 port 100BaseTX HUBS.
  - Used for login, NFS, normal operation.

MyriNet

- Full-duplex 1.28+1.28 Gbit/sec
- multistage crossbar switches
- Latencies:
  - Hardware: $\sim$ 100ns/stage
  - GM: $\sim 18\mu$s PM: $< 10\mu$s
- Any network topology allowed
- Open-Source Software
- cheap SAN cables up to 3m.



Alternative Networks

- QS-Net (Quadrics)
  - Very fast; used in Compaq's ASCI-Projekt
  - Very expensive $\rightarrow$ fewer nodes
  - No Linux support yet (but in 3-6 month)
- (Fast-/Gigabit-) Ethernet
  - Big protocol-overhead for TCP/IP.
  - No MPI support for user-level protocols (but GAMMA)
- SCI
  - Less flexible in choosing topology

IV. Software

Start with:

- SuSE 6.1 / AXP (glibc 2.1 based)
- Linux 2.2.10
- egcs 1.1.2 (C, C++, F77)
- MyriCom's gm-1.1.1 and MPIch-GM 1.1.2..8

Compilers

- gcc/egcs's optimization is know to be poor on Alpha, maybe even worse on 21264.
- Compaq ports its True64 compilers to Linux.
- Fortran (f77 & f95) and C compilers already available as beta.
- C++ is announced.
- $\sim$ 20% better performance.
- Compaq's `cxml`–library also available (BLAS etc.).

Networking

start with:

- MyriCom's gm.

Later:

- SCore
  - no error-correction in the protocol (needed?)
  - 64bit cards not yet supported
- ParaStation2
  - Very expensive but with error-correction.
  - Proposed DFG project with Prof. Tichy (University Karlsruhe).

Parallel Filesystem

- Need a parallel filesystem for efficient use of disk-space (1.2 TB):
  - easy to use parallel IO for user programs.
  - transparent access even between partitions.
  - transparent serial access of parallel stored data.
  - possibility of creating parallel data from serial files.
- Only very poor (free) implementations yet.
- Available implementations are based on TCP/IP (slow).
- Proposed DFG project with Karlsruhe to develop a parallel filesystem.
  - implementation should use the (fast) MyriNet.
  - portable code (usable with ParaStation, SCore and GM).

V. Summary

- Workstation clusters can represent cost efficient mid-range systems.
- Price/performance ratios of 40 – 50 k$/GFlop for Lattice QCD can be achieved.
- Clusters beat proprietary systems by a factor of 2.
- They can't compete with dedicated machines like APEmille yet.
- Linux is ready for clusters.
- Administration tools are available, parallel filesystems on the way.
- ALiCE will be a general purpose Supercomputer.

ALiCE
in Wuppertal