

A Commodity Cluster for Lattice QCD Calculations at DESY

Andreas Gellrich*, Peter Wegner, Hartmut Wittig
DESY

CHEP03, 25 March 2003
Category 6: Lattice Gauge Computing

* e-mail: Andreas.Gellrich@desy.de

Initial Remark

This talk is being held in conjunction with two other talks in the same session:

- K. Jansen: *Lattice Gauge Theory and High Performance Computing: The LATFOR initiative in Germany*
(NIC/DESY)
- A. Gellrich: *A Commodity Cluster for Lattice QCD Calculations DESY*
(DESY)
- P. Wegner: *LQCD benchmarks on cluster architectures*
(DESY)

Andreas Gellrich, DESY

CHEP03, 25 March 2003

1

Contents

DESY Hamburg and DESY Zeuthen operate high-performance cluster for LQCD calculations, exploiting commodity hardware. This talk focuses on system aspects of the two installations.

- Introduction
- Cluster Concept
- Implementation
- Software
- Operational Experiences
- Conclusions

Andreas Gellrich, DESY

CHEP03, 25 March 2003

2

Cluster Concept

In clusters a set of main building blocks can be identified:

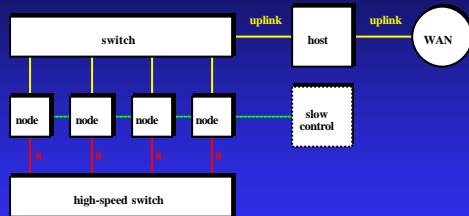
- computing nodes
- high-speed network
- administration network
- host system:
 - login, batch system
- slow network:
 - monitoring, alarming

Andreas Gellrich, DESY

CHEP03, 25 March 2003

3

Cluster Concept (cont'd)



Cluster Concept (cont'd)

Aspects for the cluster integration and software installation:

- Operating system (Linux)
- Security issues (login, open ports, private subnet)
- User administration
- (Automatic) software installation
- Backup and archiving
- Monitoring
- Alarming

Racks

Rack-mounted dual-CPU PCs.

Hamburg:

- 32 dual-CPU nodes
- 16 dual-Xeon P4 2.0 GHz
- 16 dual-Xeon P4 1.7 GHz

Zeuthen:

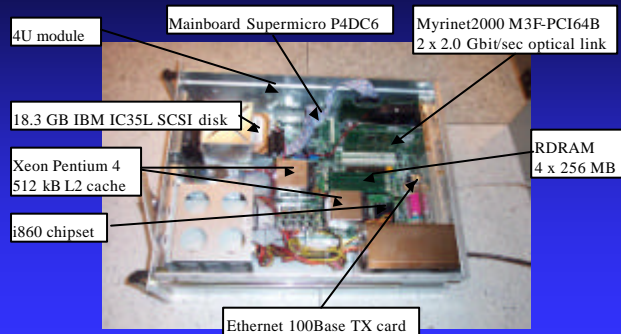
- 16 dual-CPU nodes
- 16 dual-Xeon P4 1.7 GHz



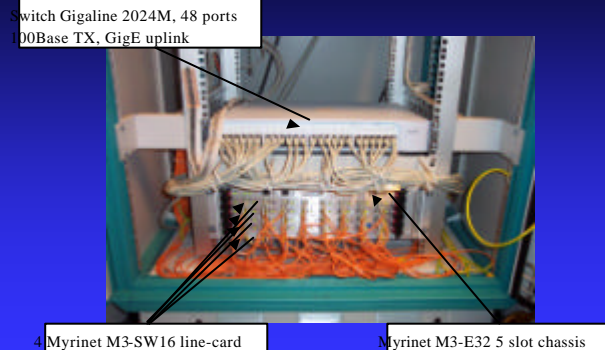
Boxes



Nodes



Switches



Software

- Network: nodes in private subnet (192.168.1.0)
- Operating system: Linux (S.u.S.E. 7.2); 2.4 SMP kernel
- Communication: MPI-based on GM (Myricom low level communication library)
- Compiler: Fortran 77/90 and C/C++ in use
GNU, Portland Group, KAI, Intel
- Batch system: PBS (OpenPBS)
- Cluster management: Clustware, (SCORE)
Time synchronization via XNTP

Backup and Archiving

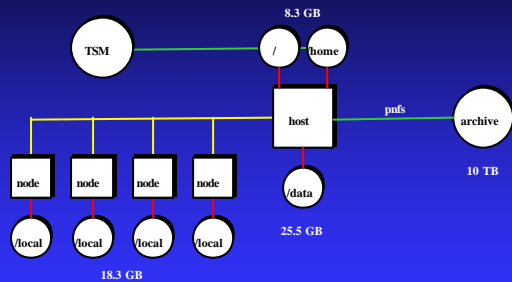
Backup:

- system data
- user home directories
- (hopefully) never accessed again ...
- incremental backup via DESY's TSM environment (Hamburg)

Archive:

- individual storage of large amounts of data O(1 TB/y)
- DESY product dCache
- pseudo-NFS directory on host system
- special *dccp* (dCache-copy) command

Backup and Archiving (cont'd)



Monitoring

Requirements:

- web-based
- history
- alarming (e-mail, sms)

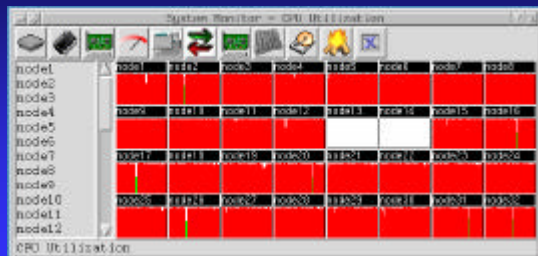
clustware: (by MEGWARE)

- no history kept
- exploits udp

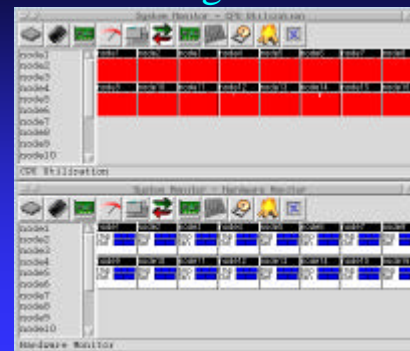
clumon: (simple; home-made; Perl-based; exploits NFS)

- deploys MRTG for history
- includes alarming via e-mail and/or sms

Monitoring: *clustware*



Monitoring: *clustware*



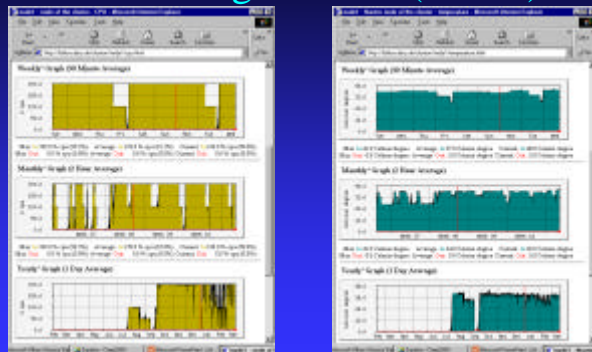
Monitoring: *clumon*



Monitoring: *clumon*



Monitoring: *clumon* (cont'd)



Operational Experiences

Zeuthen:

- running since December 2001
- 18 user accounts; 6 power users
- batch system is used to submit jobs
- hardware failures of single nodes (half of the disks replaced)
- Myrinet failures (1 switch line-cards; 2 interface-cards)
- uptime: server 362 days, nodes 172 days

Operational Experiences (cont'd)

Hamburg:

- running since January 2002
- 18 user accounts; 4 power users

- batch system is NOT used; manual scheduling

- hardware failures of single nodes (all local disks were replaced)
- Myrinet failures (3 switch line-cards; 2 interface-cards)
- 32 CPUs upgraded, incl. BIOS
- some kernel hang-ups (eth0 driver got lost, ...)

- uptime: server 109 days, 63 days (since switch repair)

Operational Experiences (cont'd)

In summary:

- several broken disks
- 1 broken motherboard
- 4 of 6 broken Myrinet switch line-cards
- 4 of 48 broken Myrinet interface-cards
- some kernel hang-ups

Possible Improvements:

- server is single point of server
- Linux installation procedure
- exploit serial console for administration

Conclusions

- Commodity hardware based PC clusters for LQCD
- Linux as operating system
- MPI-based parallelism
- Batch system optional

- Clusters in Hamburg and Zeuthen in operation for > 1 year
- Hardware problem occur, but repairs are easy

Successful model for LQCD calculations!