

# CITEC Compute Cluster (C3)

Matthias Schöpfer

CITEC Central Labs Facilities, Bielefeld University, Germany

18.05.2010

# Outline

1. Cluster Components
2. Setup
3. Stats
4. Decisions

# Preposition

- ▶ Please interrupt anytime!!!
- ▶ Noob setup of HPC

## Err, Cluster Huh?!

- ▶ 16 Compute Nodes (Dell Blades)
  - ▶ 2 Intel Xeon E5420 @ 2.50GHz
  - ▶ 16 GB RAM
  - ▶ 2 x 250 GB Harddrive (10k?) (RAID 1 Config)
  - ▶ 2 x 1 GBit Ethernet
  - ▶ Dual-port DDR InfiniBand (10GBitE)
- ▶ Head Node (Dell 2950)
  - ▶ 2 Intel Xeon E5420 @ 2.50GHz
  - ▶ 16 GB RAM
  - ▶ 6 x 750 GB (10k) HDs (RAID 5 HotSwap)
  - ▶ 4 x 1 GBit Ethernet



### Sum

128 Cores - 256 GB Mem - 6 power supplies - 3 switches - 9 Fans

## Short History

- ▶ Ordered 2008
- ▶ Delivered Begin of 2009

## Short History

- ▶ Ordered 2008
- ▶ Delivered Begin of 2009
- ▶ Since Dec, 18th 2009  
Software Renovation

## Short History

- ▶ Ordered 2008
- ▶ Delivered Begin of 2009
- ▶ Since Dec, 18th 2009  
Software Renovation
- ▶ Since 15th Jan 2010  
Alpha-Testing

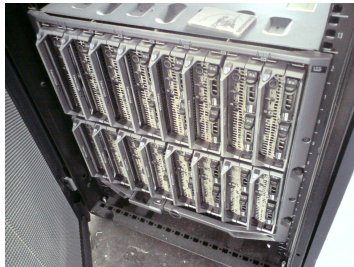
## Short History

- ▶ Ordered 2008
- ▶ Delivered Begin of 2009
- ▶ Since Dec, 18th 2009  
Software Renovation
- ▶ Since 15th Jan 2010  
Alpha-Testing
- ▶ Since 1st Feb 2010  
Beta-Testing



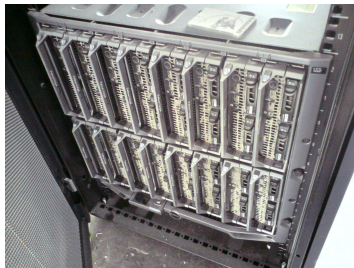
## Short History

- ▶ Ordered 2008
- ▶ Delivered Begin of 2009
- ▶ Since Dec, 18th 2009  
Software Renovation
- ▶ Since 15th Jan 2010  
Alpha-Testing
- ▶ Since 1st Feb 2010  
Beta-Testing
- ▶ Since 22nd Apr 2010 in  
TechFak Server Room



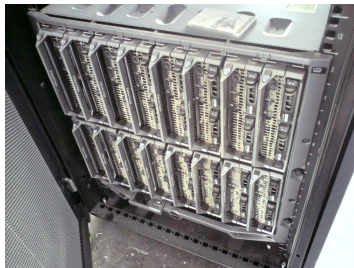
## Short History

- ▶ Ordered 2008
- ▶ Delivered Begin of 2009
- ▶ Since Dec, 18th 2009  
Software Renovation
- ▶ Since 15th Jan 2010  
Alpha-Testing
- ▶ Since 1st Feb 2010  
Beta-Testing
- ▶ Since 22nd Apr 2010 in  
TechFak Server Room
- ▶ Since 1st May 2010 Open to  
CITEC



## Short History

- ▶ Ordered 2008
- ▶ Delivered Begin of 2009
- ▶ Since Dec, 18th 2009  
Software Renovation
- ▶ Since 15th Jan 2010  
Alpha-Testing
- ▶ Since 1st Feb 2010  
Beta-Testing
- ▶ Since 22nd Apr 2010 in  
TechFak Server Room
- ▶ Since 1st May 2010 Open to  
CITEC



# Current Setup

- ▶ OS: Gentoo Linux (64 bit)
  - ▶ Extremely Scalable (USE and Compile-Flags)
  - ▶ Good, if not best performance
  - ▶ Science Overlay, “unstable” packages
  - ▶ I am at Home
- ▶ Downside
  - ▶ No “Version”
  - ▶ No binary software packages

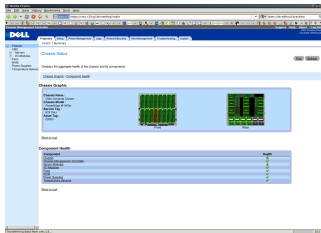
## Current Setup II

- ▶ Scheduler: Maui (Needed Bug Fix)
- ▶ Resource Manager TORQUE
  - ▶ Open Source Software
  - ▶ Sufficient for this “small” setup

What else ...

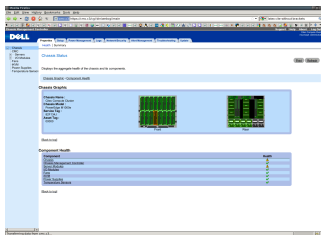
- ▶ NFS
- ▶ DNS/DHCP (dnsmasq)
- ▶ Routing (iptables)
- ▶ Monitoring (nagios/lighttpd)
- ▶ LDAP (using TechFak LDAP)
- ▶ rsync (for portage)

# Dell Management Tools



- ▶ Chassis Management Controller
  - ▶ SSH
  - ▶ Serial console
  - ▶ Hardware sensors
  - ▶ Web interface
  - ▶ Support for boot image & ssh-key upload
  - ▶ From linux sometimes cumbersome
  - ▶ Debugging cumbersome (email alert) (Error No. xxx)

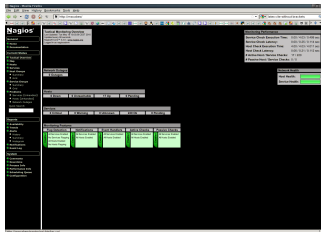
# Dell Management Tools



- ▶ Chassis Management Controller
  - ▶ SSH
  - ▶ Serial console
  - ▶ Hardware sensors
  - ▶ Web interface
  - ▶ Support for boot image & ssh-key upload
  - ▶ From linux sometimes cumbersome
  - ▶ Debugging cumbersome (email alert) (Error No. xxx)

# Nagios

- ▶ Nagios
  - ▶ Monitors Services & Hosts
  - ▶ Substitute for CMC





# How to get my software running

- ▶ Log in (Host: macabeo, PW is same as Techfak (LDAP is same))
- ▶ You will get a new `$HOME`
- ▶ `scp` your software in your `$HOME`
- ▶ RECOMPILE!
- ▶ Software packages missing? Please contact me ([c3-admins@lists.cit-ec.uni-bielefeld.de](mailto:c3-admins@lists.cit-ec.uni-bielefeld.de))

# Submitting a job

- ▶ Please, do not put macabeo under (heavy) load! This is NOT the compute cluster!
- ▶ Use `qsub` to submit a job
- ▶ You need to tell `qsub` your needs (nodes, time, etc.)
- ▶ Use `qstat` or `showq` to check status of your job(s)
- ▶ Use `qdel` to delete a job
- ▶ Use `showbf` of what is available rightaway

```
bin Edit View Terminal Help
ACTIVE JOBS-----
JOBNAME      USORNAME      STATE      PROC      REMAINING      STARTTIME
7298          qgrssok        Running    1      00:57:24      Mon May 17 02:05:06
7299          qgrssok        Running    4      4:38:25      Sat May 15 17:44:07
7221          qgrssok        Running    4      12:28:22      Sun May 09:09:09-09
7232          qgrssok        Running    4      26:43:43      Sun May 09:00:25:25
7233          qgrssok        Running    4      1:04:43:20      Sun May 09:17:49:03
7234          qgrssok        Running    4      1:04:41:31      Sun May 09:17:49:13
7226          qgrssok        Running    4      1:04:41:09      Sun May 09:17:49:18
7237          qgrssok        Running    4      1:04:41:35      Sun May 09:17:49:18
7236          qgrssok        Running    4      1:04:41:35      Sun May 09:17:49:18
7229          qgrssok        Running    4      1:04:41:46      Sun May 09:17:49:28
7248          qgrssok        Running    4      1:04:42:07      Sun May 09:17:49:28
7241          qgrssok        Running    4      1:04:42:47      Sun May 09:17:49:29
7242          qgrssok        Running    4      1:04:42:47      Sun May 09:17:49:29
7243          qgrssok        Running    4      1:04:49:49      Sun May 09:17:55:31
7240          qgrssok        Running    4      1:05:04:23      Sun May 09:18:12:41
7246          qgrssok        Running    4      1:05:06:40      Sun May 09:18:14:27
7247          qgrssok        Running    4      1:05:06:45      Sun May 09:18:14:27
7249          qgrssok        Running    4      1:05:07:38      Sun May 09:18:15:12
7250          qgrssok        Running    4      1:05:07:38      Sun May 09:18:15:12
7251          qgrssok        Running    4      1:05:07:47      Sun May 09:18:15:29
7252          qgrssok        Running    4      1:05:07:47      Sun May 09:18:15:29
7253          qgrssok        Running    4      1:05:07:49      Sun May 09:18:15:31
7254          qgrssok        Running    4      1:05:07:53      Sun May 09:18:15:35
7255          qgrssok        Running    4      1:05:07:53      Sun May 09:18:15:35
7256          qgrssok        Running    4      1:05:08:30      Sun May 09:18:18:12
7257          qgrssok        Running    4      1:05:08:30      Sun May 09:18:18:12

20 Active Jobs      199 of 128 Processors Active (85.16%)
10 of 30 Nodes Active (100.00%)

IDLE JOBS-----
JOBNAME      USORNAME      STATE      PROC      HELDPT      QHOLDPT
7300          walter        Idle       5      00:28:00      Mon May 17 14:27:44
7312          walter        Idle       34      00:28:00      Mon May 17 20:08:04
7313          walter        Idle       20      00:28:00      Mon May 17 20:08:09
7314          walter        Idle       24      00:28:00      Mon May 17 20:08:31
7315          walter        Idle       28      00:28:00      Mon May 17 20:08:52
7316          walter        Idle       26      00:28:00      Mon May 17 20:08:58
7317          walter        Idle       40      00:28:00      Mon May 17 20:09:41
7318          walter        Idle       40      00:28:00      Mon May 17 20:09:59
7319          walter        Idle       38      00:28:00      Mon May 17 20:10:59
7320          walter        Idle       6      00:28:00      Mon May 17 20:11:03
7322          walter        Idle       9      00:28:00      Mon May 17 20:12:02
7323          walter        Idle       31      00:28:00      Mon May 17 20:12:21

12 Idle Jobs

BLOCKED JOBS-----
JOBNAME      USORNAME      STATE      PROC      HELDPT      QHOLDPT
Total Jobs: 40      Active Jobs: 20      Idle Jobs: 12      Blocked Jobs: 0
```

# Qsub Script

Example command: `qsub -l walltime=00:20:00 script.sh`

## Script:

```
#PBS -l ncpus=4
#PBS -N mytestprogramm
#PBS -o testprog.out
#PBS -e testprog.err
#PBS -M mschoepf@techfak.uni-bielefeld.de
#PBS -m be
echo $PBS_JOBID
echo "Start time :"
date
cd work/hpcc-1.3.1/
pwd
uname -a
sleep 30
echo "End Time :"
date
```

## What if I want ...

- ▶ Use `qsub -t` option to run a program multiple times (`$PBS_ARRAYID`)
- ▶ Usage of OpenMPI: `mpirun <binary> ergo NO HOSTFILE!!`

# A word about queues

Available queues (priorities in descending order):

- ▶ debug (max. 5 min)
- ▶ short (max. 2 hrs)
- ▶ medium (max. 3 days)
- ▶ long (max. 30 days)
- ▶ default (routing queue)

# Infiniband

## ► Updating Firmware and Software (Mellanox)

```
$ mpirun -mca btl self,tcp pingpong  
Max rate = 114.410304 MB/sec Min latency =  
51.975250 usec  
$ mpirun -mca btl self,openib pingpong  
Max rate = 763.152266 MB/sec Min latency =  
2.026558 usec
```

# Infiniband

- ▶ Updating Firmware and Software (Mellanox)

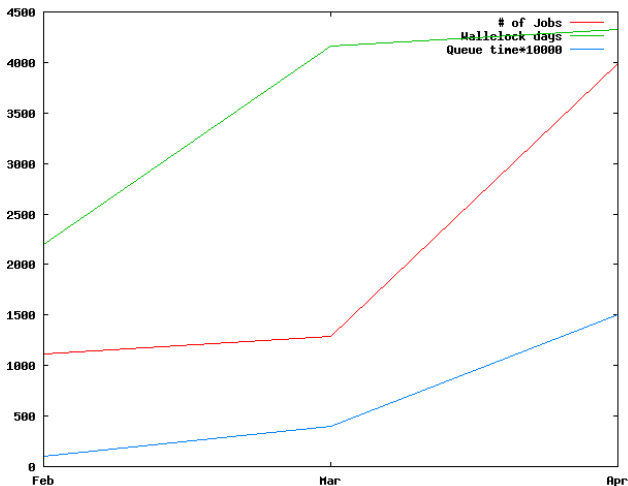
```
$ mpirun -mca btl self,tcp pingpong  
Max rate = 114.410304 MB/sec Min latency =  
51.975250 usec  
$ mpirun -mca btl self,openib pingpong  
Max rate = 763.152266 MB/sec Min latency =  
2.026558 usec
```

# To Dos

- ▶ Export /vol/compute
- ▶ Import /vol/matlab



# Statistics



# MPI Infiniband

Detailed benchmark results:

Ping Pong:

Latency min / avg / max: 0.000864 / 0.003192 /  
0.004053 msec

Bandwidth min / avg / max: 935.446 / 1070.490 /  
4019.458 MByte/s

Ring:

On naturally ordered ring: latency= 0.004292  
msec, bandwidth= 182.348160 MB/s

On randomly ordered ring: latency= 0.018119 msec,  
bandwidth= 116.242569 MB/s

# Decisions

- ▶ Gentoo vs. RHEL, Debian, Ubuntu Server...
- ▶ Local Install vs. Netboot
- ▶ Queue Management vs. direct access
- ▶ Maui/TORQUE vs. SGE
- ▶ NFS, User Management, ...

# Questions?!

**Finisch is not the End**

Thank you for your attention!