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

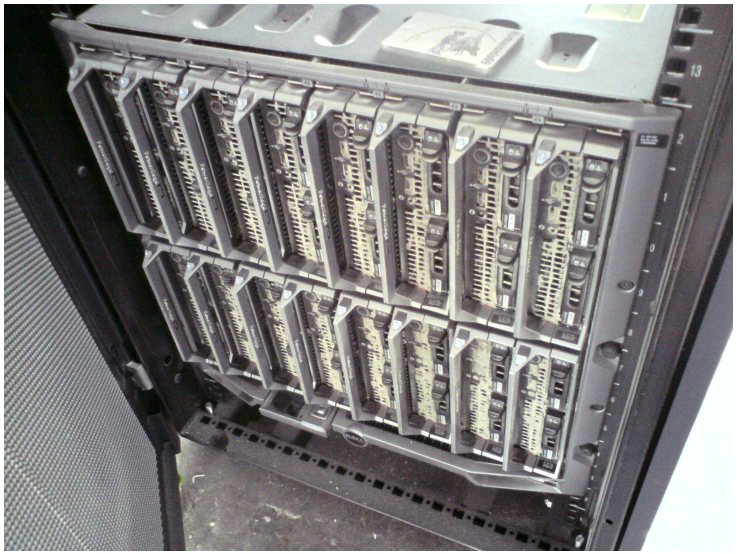
- Ordered 2008
- Delivered Begin of 2009
- Since Dec, 18th 2009
  Software Renovation
- Since 15th Jan 2010
  Alpha-Testing

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

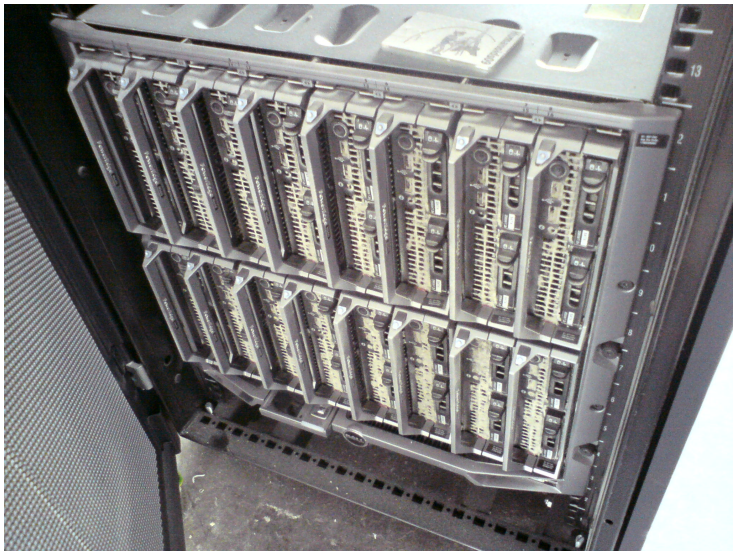
Short History

- 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



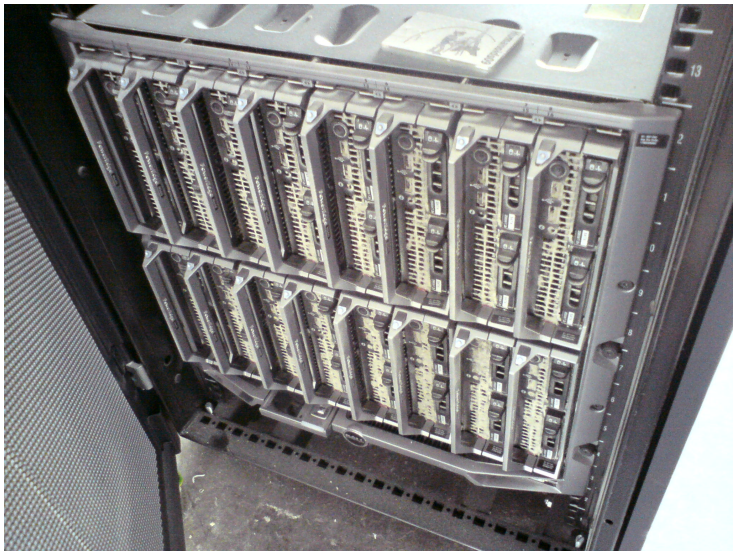
Short History

- 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

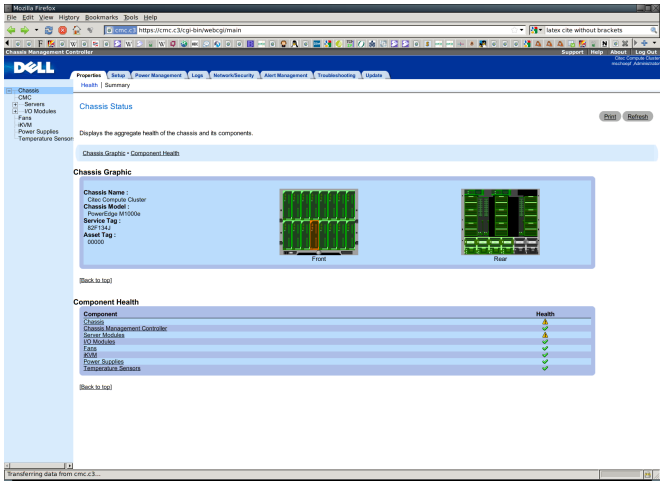
- OS: Gentoo Linux (64 bit)
  - Extremly Scalable (USE and Compile-Flags)
  - Good, if not best performance
  - Science Overlay, “unstable” packages
  - I am at Home
- Downside
  - No “Version”
  - No binary software packages

Current Setup II

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

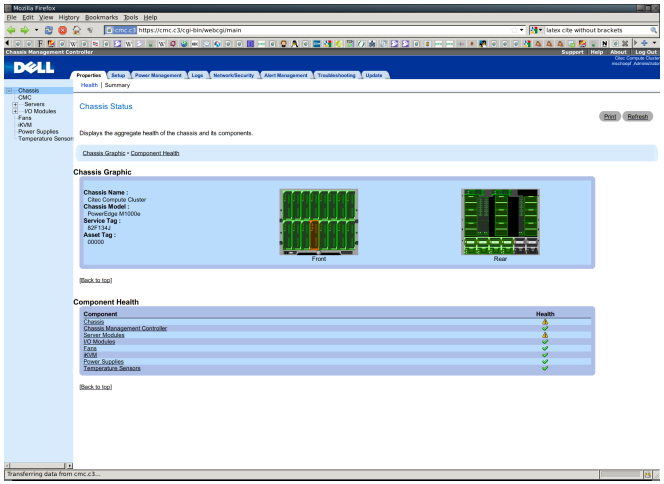
  What else ...
  - NFS
  - DNS/DHCP (dnsmasq)
  - Routing (iptables)
  - Monitoring (nagios/lighttpd)
  - LDAP (using TechFak LDAP)
  - rsync (for portage)

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)

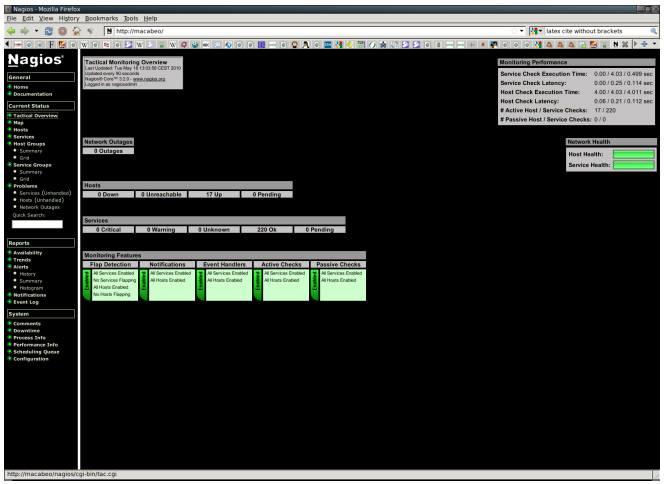
Dell Management Tools



- 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

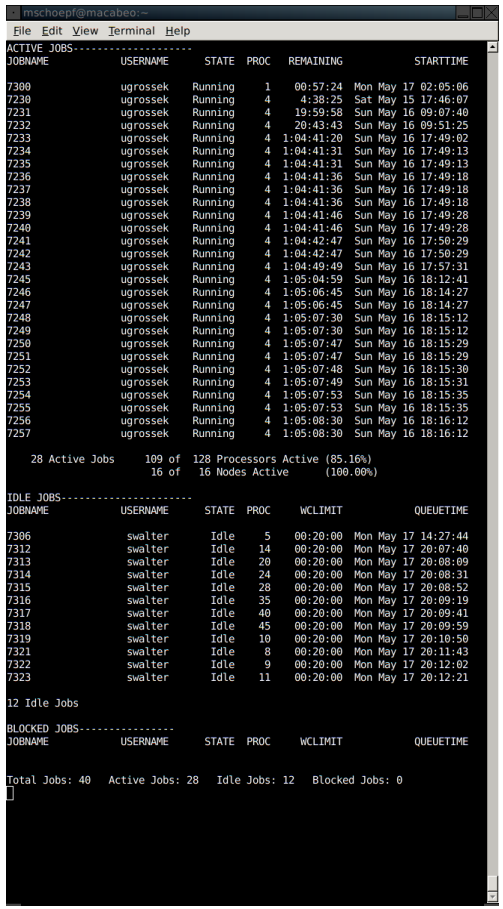


How to get my software running

- 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)

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



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

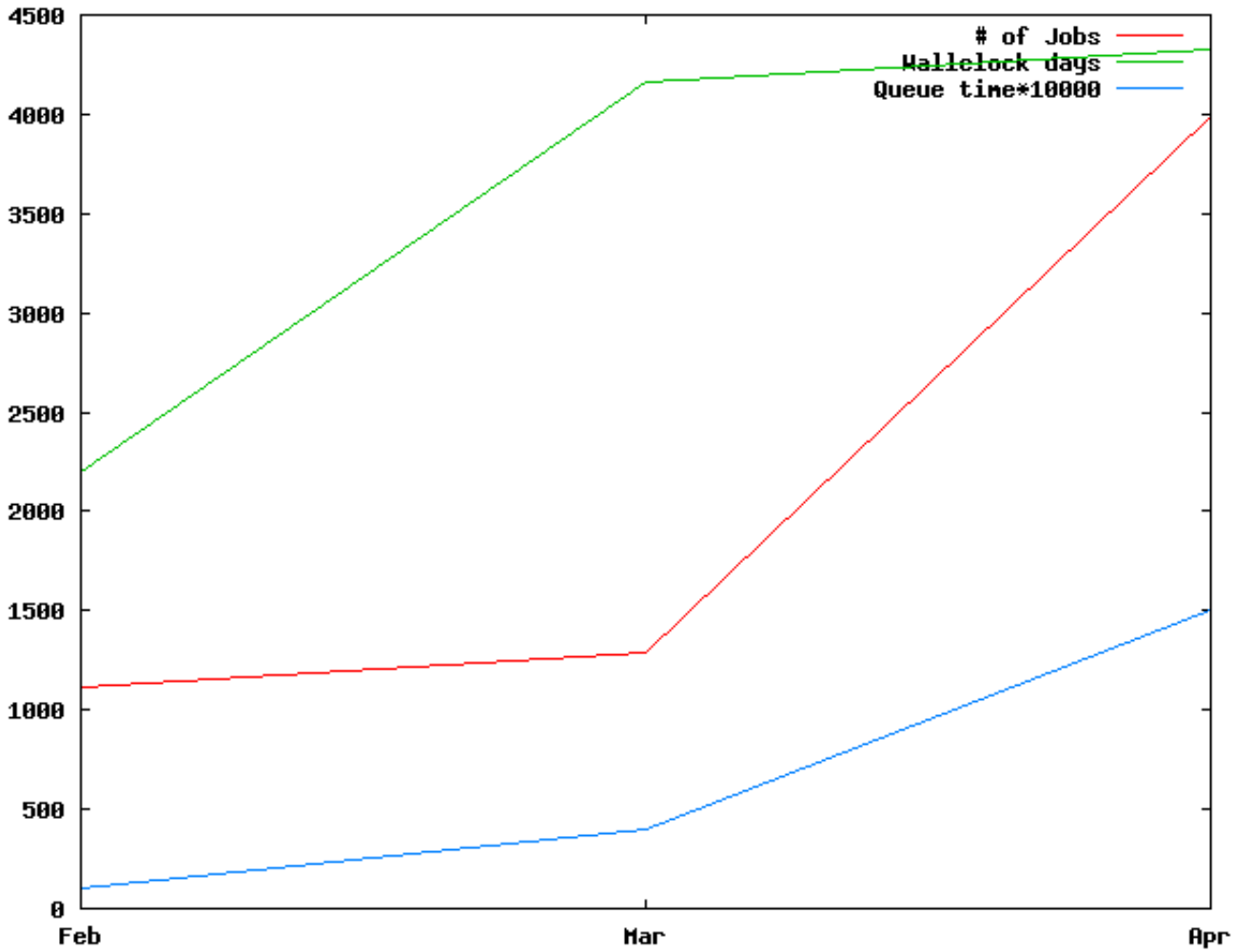
- Updating Firmware and Software (Mellanox)

```
$ 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


4500
4000
3500
3000
2500
2000
1500
1000
500
0
Feb
Mar
Apr
# of Jobs
Wallclock days
Queue time*10000

MPI Infiniband

```
Detailed benchmark results:
Ping Pong:
Latency min / avg / max:  0.000864 / 0.003192 /
0.004053 msecs
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!