Experiences setting up a Rocks based Linux Cluster

Tomas Lindén
Helsinki Institute of Physics
CMS Programme

23.10.2003

4th Annual Workshop on Linux Clusters for Super Computing, October 22-24, 2003
National Supercomputer Centre (NSC), Linköping University, Linköping, Sweden

Contents

1. System requirements

- Hardware optimized for **Monte Carlo simulations** in Computational Material Science (MD-simulations) and High Energy Physics (CMS Monte Carlo production runs for CMS Data Challenges).
- The applications are mostly of the **embarassing parallell** type.
- **Network bandwidth** between the nodes and the **latency** for message passing is therefore not an issue, so only a 100 Mb/s network is required. A Gb/s network would be nice in the long run, however.
- One phase of the simulations is I/O bound and rather tough to meet because the data needs to be **read randomly**.

The **budget** for the system was $\approx$ 50 keuros (without VAT).

Main design goals

- Maximize CPU power
- Maximize random disk read capacity
- Minimize costs

The ideal computer

- CPU
- memory
- I/O-interfaces

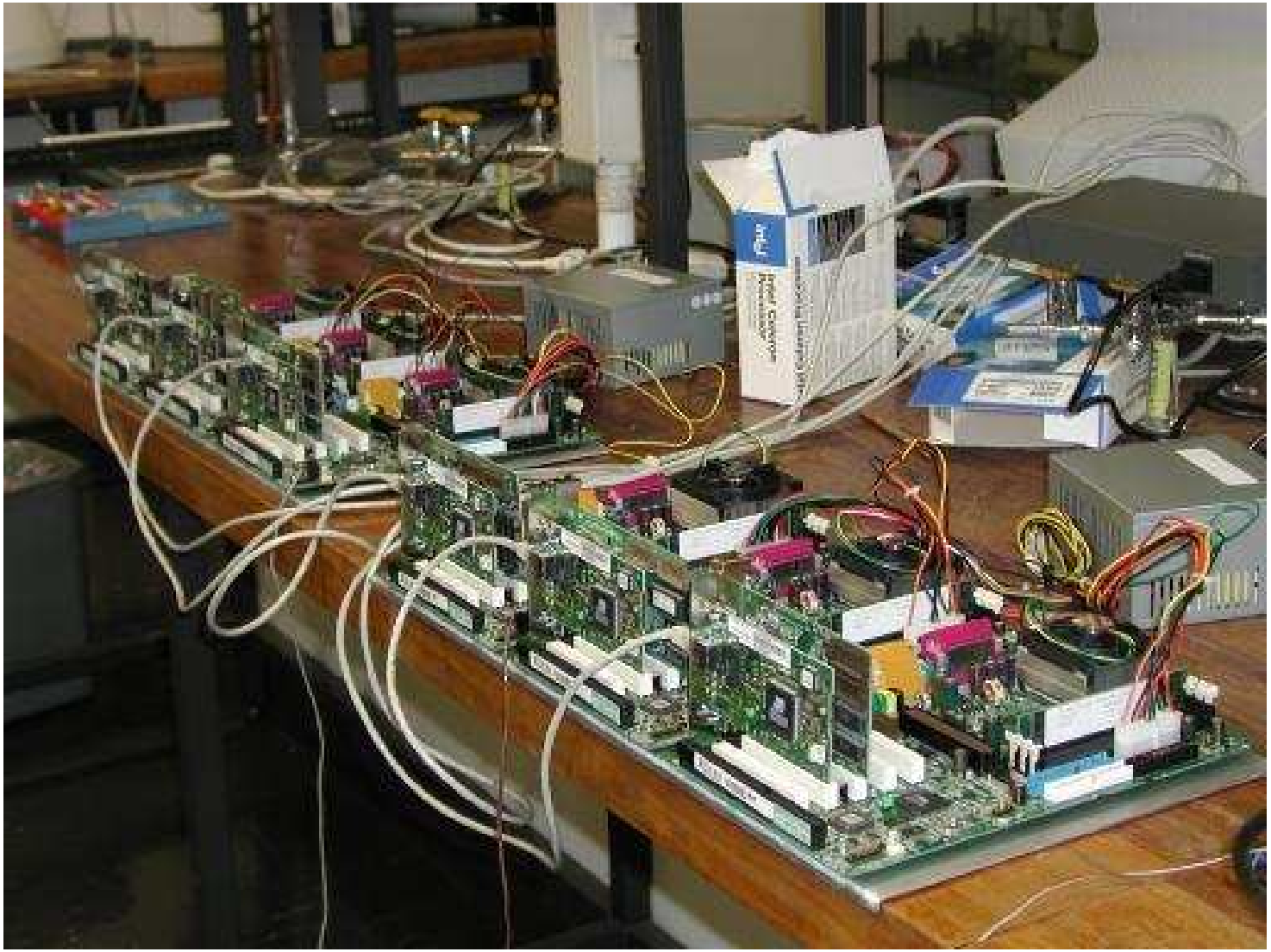
A real computer needs additional components like

- casing or mechanical support
- motherboard
- power supply
- bootable mass storage devices
- network interface card(s)
- graphics adapter
- monitor
- keyboard
- expansion slots
- I/O ports (USB, FireWire, parallell, serial, ...)

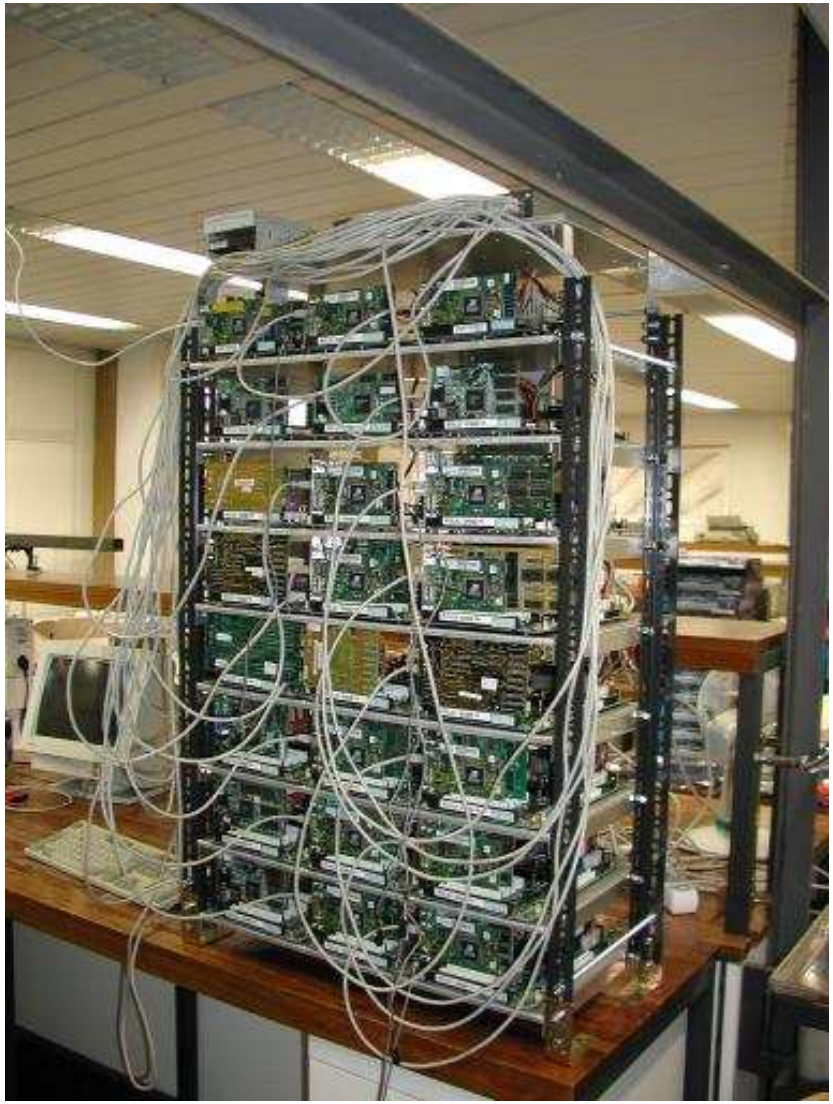
The ideal computer can be approximated by minimizing the number of components in the worker nodes, which will also minimize costs. Some money can be saved by recycling hardware for non critical items.

Casing comparison

- *minitower*
  - \+ standard commodity solution
  - \+ inexpensive
  - \- big space requirement
- *1U or 2U 19" rack case*
- \+ more compact than minitower
- \- more expensive than minitower
- home made *"ATX blade server"*
- \+ more compact than minitower
- \+ least expensive if assembly costs are neglected
- \- labour intensive
- \- cooling problems can be an issue
- *blade server*
- \+ most compact
- \- most expensive



3 node 1.2 GHz, 512 MB RAM Celeron Tualatin ATX blade built by A. Sandvik.



The 24 nodes of the Celeron ATX blade server [1].

Here the power supply tower can be seen.

AMD Athlon XP ATX blade at the Institute of Biotechnology at the University of Helsinki.

A 32 node AMD Athlon XP Mosix ATX blade server.

Heat dissipation is obviously a major problem with any sizeable cluster. Other factors affecting the choice of **casing** are space and cost issues.

The idea of building a *"ATX-blade server"* was very attractive to us in terms of needed space and costs, but we were somewhat discouraged by the heat problems with the previously shown cluster (the heat problem was subsequently solved with more effective CPU coolers).

It was also felt that one would have more problems with warranties with a completely home built system.

Also the mechanical design of a ATX blade server would take some extra effort compared to a more standard solution.

1U cases are very compact, but the only possibilty to add expansion cards is trough a PCI-riser card, which can be problematic. The cooling of a 1U case needs to be designed very carefully.

The advantage of a **2U** case is that one can use *half height* or *low-profile* PCI-cards without using any PCI-riser card (Intel Gb/s NICs are available in *half height PCI size*). Heat dissipation problems are probably less likely to occur in a 2U case because of the additional space available for airflow.

The advantage of **4U** cases is that standard PCI-expansion cards can be used.

Because of space limitations a *mini tower solution* was not possible so we chose a **rack based solution** with **2U** cases for the nodes and a **4U** case for the frontend for the Linux cluster *mill*.

In many cases a **motherboard** with as many integrated components as possible is desirable:

- Some motherboards come with integrated **graphics adapters**.
- No external graphics adapter is also needed if the BIOS supports a **remote serial console** [2].
- One (or two) integrated **network interface(s)** is also desirable.
- An integrated **SCSI- or ATA RAID-controller** might be useful.
- Each worker node does not need a **CDROM-drive** nor a **floppy disk drive** if the motherboard BIOS supports booting from a corresponding **USB device** and the network card and the cluster management software supports **PXE booting**. Worker nodes can then be booted for maintenance or BIOS upgrades from USB devices that are plugged into the node only when needed if the motherboard BIOS supports this.
- A **USB flash memory stick** might replace a floppy disk drive.

A cost effective solution to the random disk reading I/O problem is to equip each node with dual ATA disk drives in a software RAID 1 configuration. In this way we can get some 8 MB/s per node when reading *randomly* which is enough.

The motherboard requirements were:

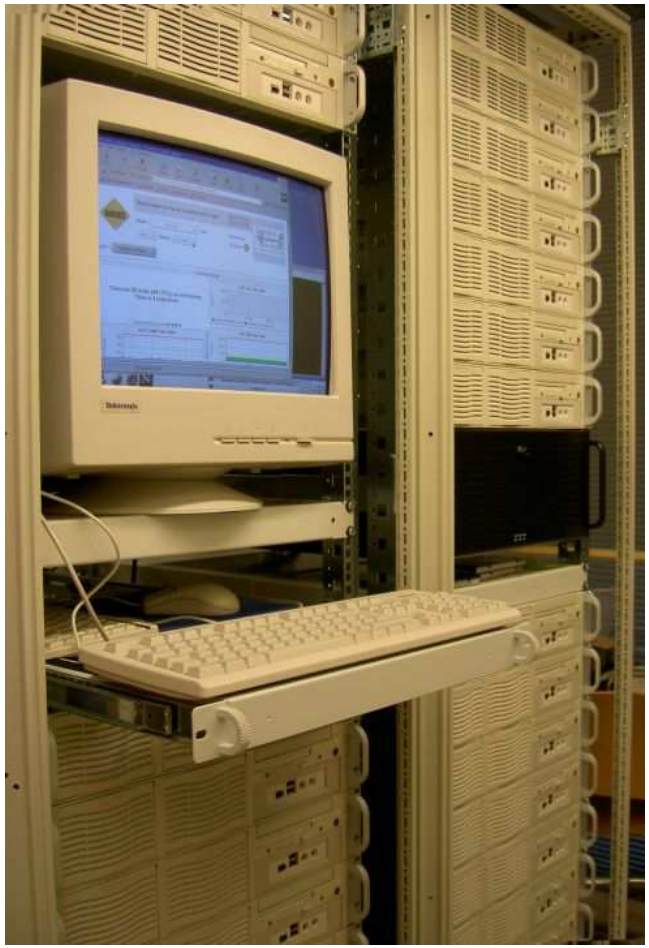
- Dual CPU support
  - minimizes the number of boxes to save space and maintenance effort
- At least dual ATA 100 controllers
- Integrated fast ethernet or Gb/s ethernet
- Serial console support or integrated display adapter
- Support for bootable USB-devices (CDROM, FDD)
- Support for PXE booting

2. Hardware setup

- CPU: 2 * AMD 2.133 GHz Athlon MP
- MB: Tyan Tiger MPX S2466N-4M
- Memory: 1 GB ECC registered DDR
- IDE disks: 2 * 80 GB 7200 rpm Hitachi 180 GXP DeskStar
- NIC: 3Com 3C920C 100 Mb/s
- Case: Supermicro SC822I-300LP 2U
- Power: Ablecom SP302-2C 300W
- FDD: integrated in the case
- Price: ≈1.4 kEuros / node with 0 % VAT

One USB-IDE case with a 52x IDE CDROM drive is common for all nodes.
The 100 Mb/s network switches existed already previously.

The 32+1 node dual AMD Athlon M-P 2U Rocks rack cluster *mill*.



Frontend hardware:

- *CPU: 2 * AMD 1.667 GHz Athlon MP*
- MB: Tyan Tiger MPX S2466N-4M
- Memory: 1 GB ECC registered DDR
- *IDE disks: 2 * 60 GB 7200 rpm IBM 180 GXP DeskStar*
- *NIC: 3Com 3C996-TX Gb/s*
- NIC: 3Com 3C920C 100 Mb/s
- *Graphics: ATI Rage Pro Turbo 8MB AGP*
- *Case: Compucase S466A 4U*
- *Power supply: HEC300LR-PT*
- *CDROM: LG 48X/16X/48X CD-RW*
- FDD: yes

Node cables:

- Power supply
- Ethernet
- Serial console CAT-5

Console monitoring:

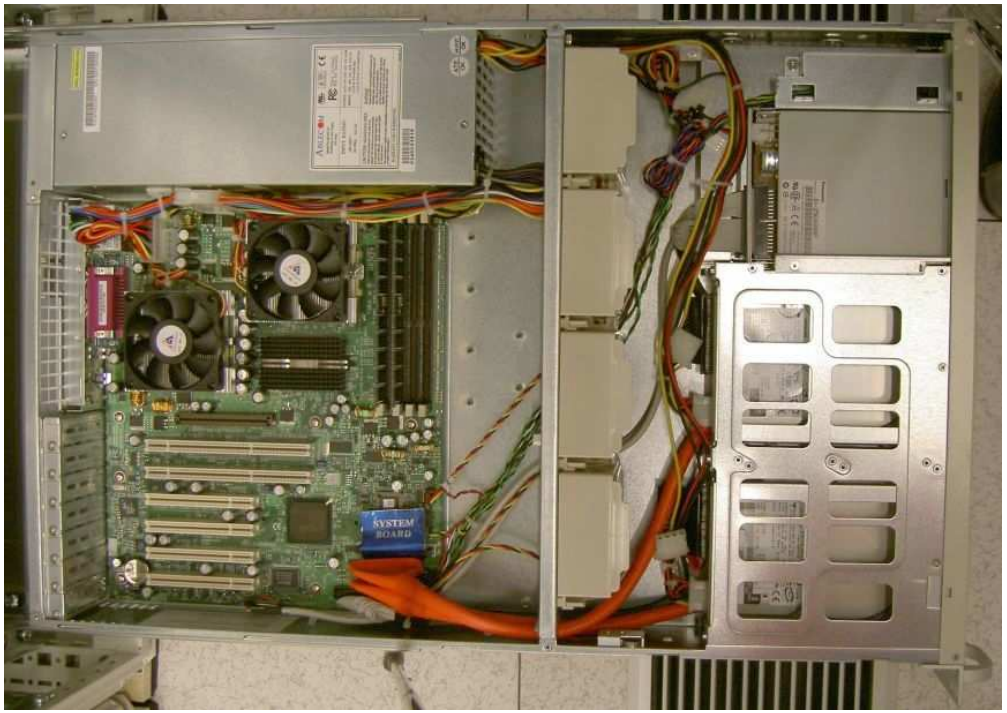
- Digi EtherLite 32 with 32 RS232 ports
- 2 port D-Link DKVM-2 switch

Recycled hardware:

- rack shelves
- serial console control computer
- 17” display
- keyboard
- mouse

Cooling

- The frontend draws 166 W (idle) — 200 W (full load).
- Rounded IDE cables on the nodes.
- The cases have large air exhaust holes near the CPUs.
- Node CPU temperatures are $\approx$ 40 °C, when idle.



The interior of one of the nodes.

About the chosen hardware

- **PXE booting** works just great.
- **USB CDROM** and **USB FD** booting works nicely with the latest BIOS.
- **USB 1.1** is a bit slow for large files.
- Booting from a **USB memory stick** is not so convenient, because the nodes sometimes hang when the flash memory is inserted. The BIOS boot order list has to be edited every time the memory stick is inserted, unlike the case for CDROM drives and FDDs. This also makes the flash memory unpractical for booting several nodes.
- The **remote serial** console is a bit tricky to setup, but it works. Unfortunately the BIOS does not allow interrupting the slow memory test from a serial console.
- The **Digi EtherLite 32** has worked nicely after the correct DB9-RJ-45 adapter wiring was used.

Node console

BIOS settings can be replicated with **/dev/nvram**, but to see BIOS POST messages or console error messages some kind of console is needed [2].

- keyboard and display is attached only when needed (headless)
- Keyboard Video Mouse (KVM) switch
- serial switch

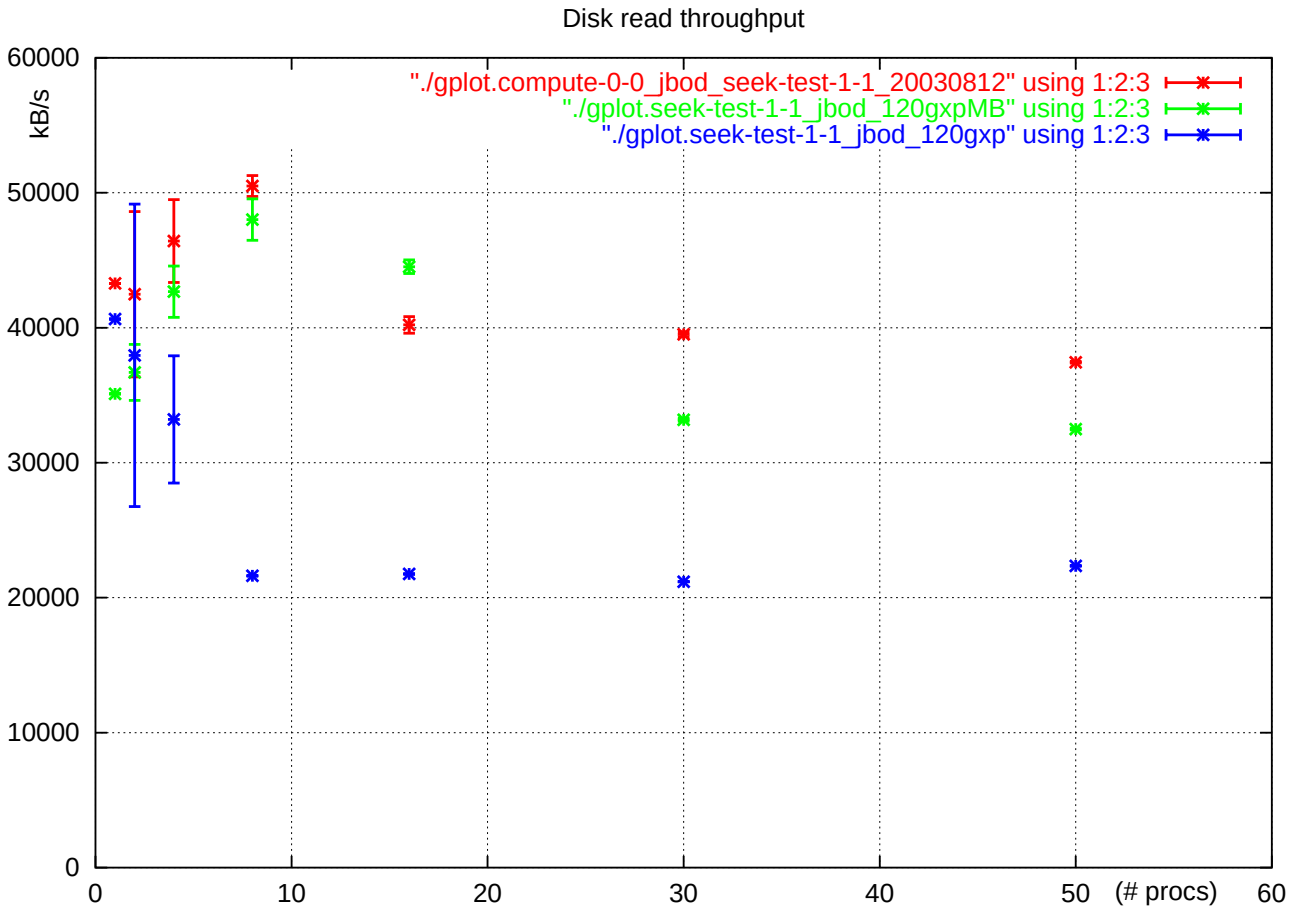
	KVM	**Serial**
cabling	-	+
LAN access	-	+
all consoles at once	-	+
graphics card	required	not required
speed	+	-
special keys	+	-
mouse	+	-

3. Disk performance

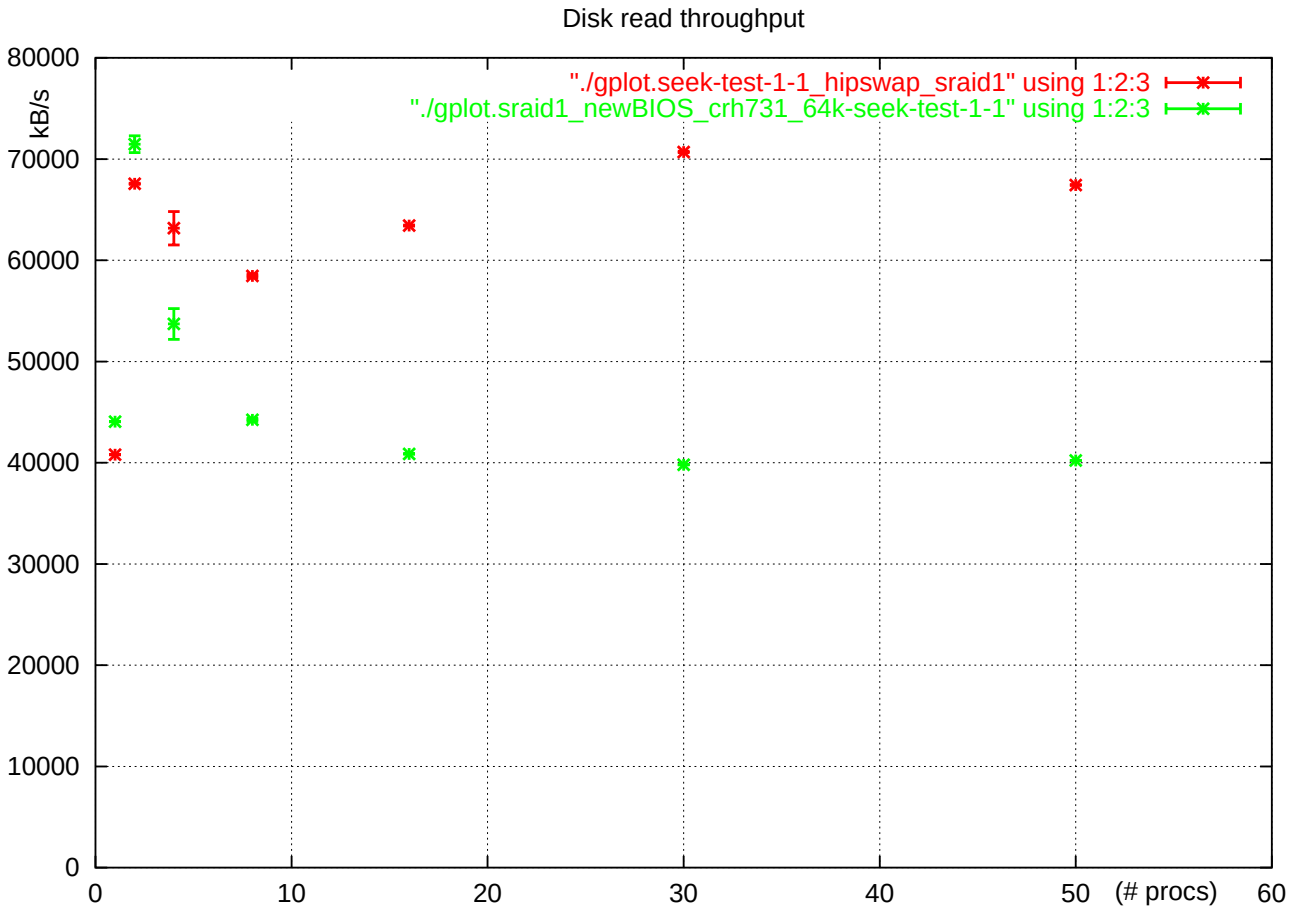
The **two disks** on each node can be used in a software RAID 1 configuration or a in a software RAID 0 configuration or as individual disks (with or without LVM). In the following performance for single disks and software RAID 1 are presented.

The *seek-test* by *Tony Wildish* is a benchmark tool to study sequential and random disk reading speed as a function of the number of processes [3].

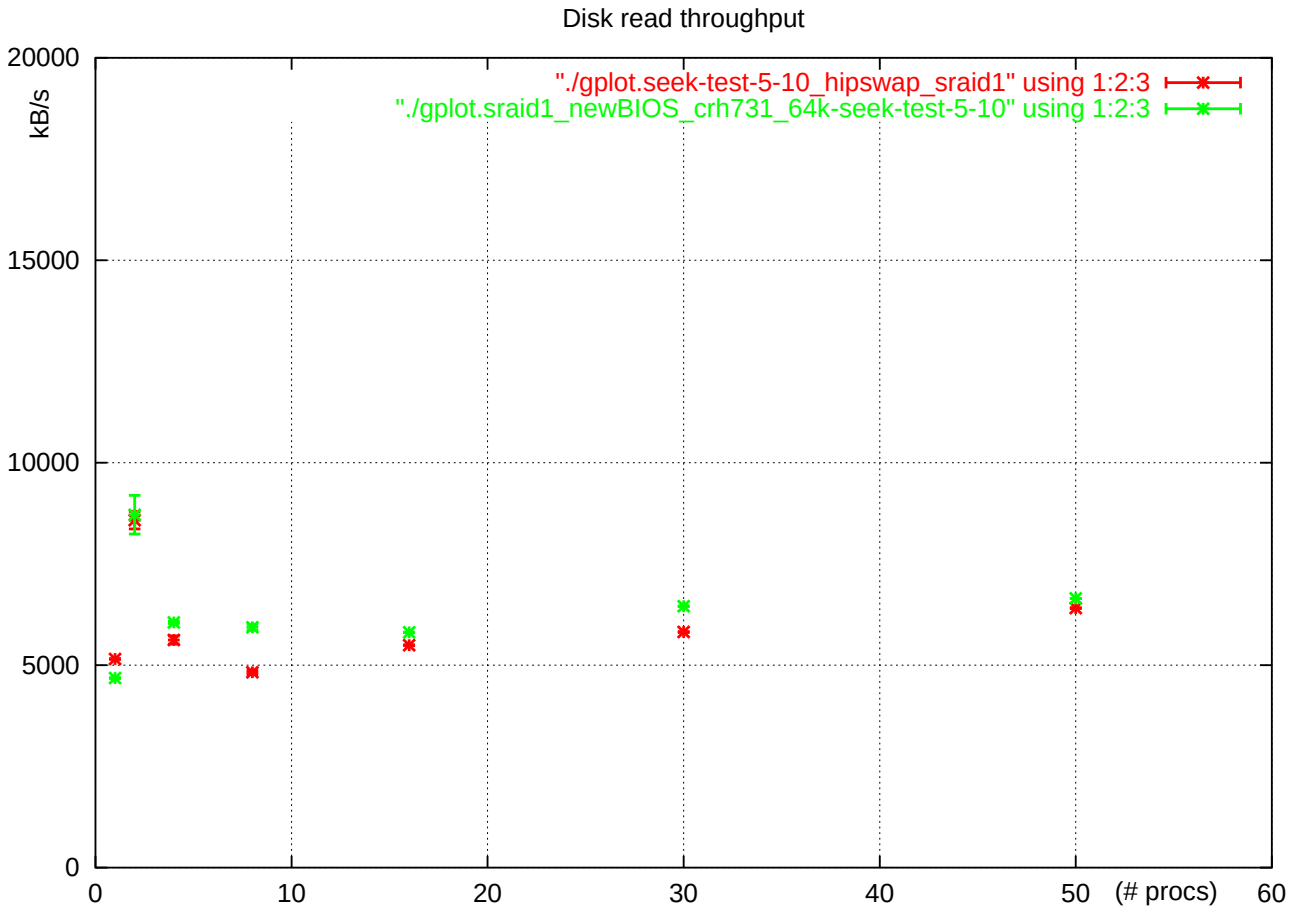
- Each filesystem was filled so the disk speed variation of different disk area regions were averaged over.
- The files were read sequentially (1-1) and randomly (5-10) (the test can randomly skip a random number of blocks within a interval of minimum and maximum number of blocks).
- All of the RAM available was used in these tests and each point reads data for 600 s.



Single disk sequential read performance Tiger MPX MB IDE-controller 180 GXP disk, Tyan MP MB IDE-controller 120 GXP disk and 3ware 7850 IDE-controller 120 GXP disk on a Tiger MP.



Software RAID 1 sequential read performance Tiger MPX MB IDE-controller and 3ware 7850 IDE-controller on a Tiger MP.



Software RAID 1 random read performance Tiger MPX MB IDE-controller and 3ware 7850 IDE-controller on a Tiger MP.

4. Software setup

BIOS settings

All nodes have serial console and PXE booting enabled. To turn on the system in a controlled way after a power loss the nodes have the BIOS setting Stay off enabled.

Boot discs

For rescue booting and testing the headless nodes a set of boot discs supporting a serial console were created:

- Memtest 3.0 (compiled with serial support enabled)
- Rocks 2.3.2 bootnet.img (syslinux.cfg edited)
- RH9 CD 1 (syslinux.cfg edited and CDROM burned)

Useful kernel parameters:

- RedHat kernels accept the option *apm=power_off*, which is useful on SMP-machines.

Cluster software requirements

- Cluster software should work with CERN RedHat Linux
- No cluster software with a modified kernel (OpenMosix, Scyld)

Because of a favorable review and positive experiences within CMS NPACI Rocks was chosen as the cluster software instead of OSCAR [4].

Compilers to install:

- C/C++ and F77/F90 compilers by Intel and Portland Group

Grid tools to install:

- NorduGrid software
- Large Hadron Collider Grid software

Application software to install:

- Compact Muon Solenoid software
- Compact Muon Solenoid production software,
- Molecular dynamics simulation package PARCAS

5. Experiences using NPACI Rocks

NPACI Rocks Cluster Distribution is a RPM based cluster management software for scientific computation based on RedHat Linux [5]. The most recent version is 3.0.0 and the previous one is 2.3.2. Both are based on RedHat 7.3.

Features of Rocks

- Supported architectures IA32, IA64, Opteron coming, (no Alpha, SPARC)
- Network: Ethernet, Myrinet
- Relies heavily on Kickstart and Anaconda (works only with RedHat)
- XML configuration scripts
- RPM software is supported, other packages can be handled with XML postconfiguration scripts
- Can handle a cluster with heterogenous hardware

- eKV a telnet based tool for remote installation monitoring
- Supports headless nodes
- Supports PXE booting and installation
- Installation is very easy on well behaved hardware
- Services and libraries out of the box
  - Ganglia (nice graphical monitoring)
  - PBS (batch queue system)
  - Maui (scheduler)
  - MPICH (parallell libraries)
  - DHCP (node ip-addresses)
  - NIS (user management) 411 SIS is beta in v. 3.0.0
  - NFS (global disk space)
  - MySQL (cluster internal configuration bookkeeping)
  - HTTP (cluster installation)

The Rocks manual covers the minimum to get one started. Rocks has a very active mailing list with a web archive for users.

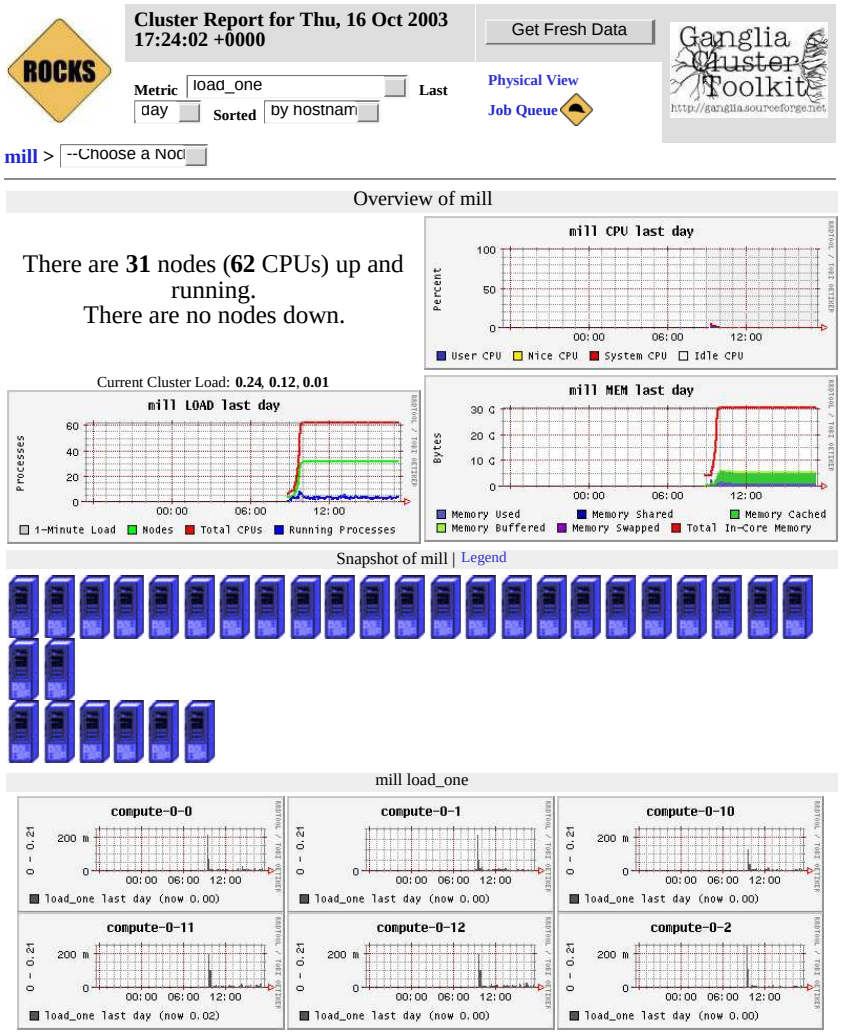
All nodes are considered to have *soft state* and any upgrade, installation or configuration change is done by node reinstallation, which takes about 10 min for a node.

The default configuration is to reinstall a node also after each power down. Settings like this can be changed according to taste.

Rocks makes it possible for nonexperts to setup a Linux cluster for scientific computation in a short amount of time.

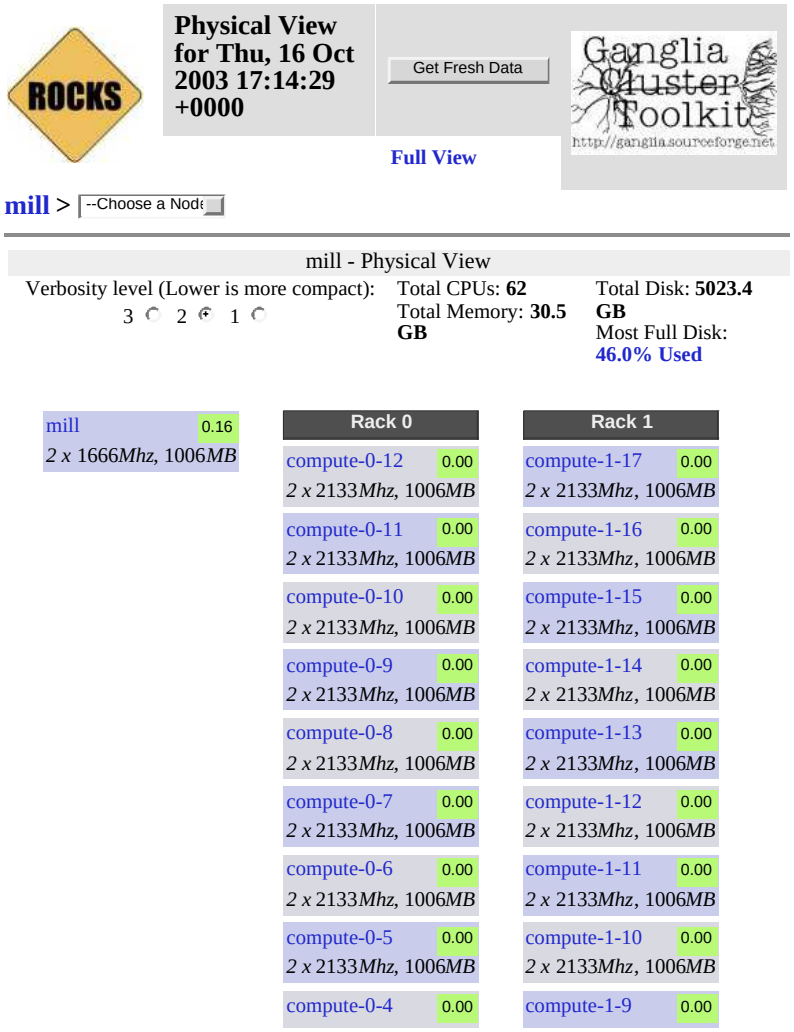
Recent RedHat Linux licensing changes will affect both NPACI Rocks and CERN RedHat Linux.

Ganglia Cluster Toolkit:: Cluster Report



Overview produced with Ganglia of the Rocks cluster *mill*.

Ganglia Cluster Toolkit:: Physical View



Physical view over the Rocks cluster *mill*.

6. Conclusions

- The hardware chosen for the cluster works as expected.
- The BIOS has still room for improvements.
- BIOS writers should add remote console capabilities to something faster than the serial line (LAN or USB). This would be simpler than using a boot card.
- Software RAID 1 gives a good sequential and random (2 processes) disk reading performance.
- The 3ware IDE-controller driver or the Linux SCSI driver has some room for improvement compared to the IDE-driver.
- The NPACI Rocks cluster distribution has shown to be a powerful tool enabling nonexperts to set up a Linux cluster.
- The Rocks documentation level is not quite up to the software quality. This is mostly compensated by the active Rocks user mailing list.

7. Acknowledgements

The Institute of Physical Sciences has financed the cluster nodes.

The **Kumpula Campus Computational Unit** hosts the *mill* cluster in their machine room.

N. Jiganova has helped with the software and hardware of the cluster.

P. Lähteenmäki has been very helpful in clarifying network issues and setting up the network for the cluster.

Damicon Kraa the vendor of the nodes has given very good service.

References

[1] *ATX blade server cluster built by A. Sandvik at Åbo Akademi* http://www.abo.fi/~physcomp/cluster/cluster.html

[2] *Remote Serial Console HOWTO*, http://www.dc.turkuamk.fi/LDP/HOWTO/Remote-Serial-Console-HOWTO/index.html.

[3] *Seek-test*, by T. Wildish http://wildish.home.cern.ch/wildish/Benchmark-results/Performance.html.

[4] *Analysis and Evaluation of Open Source Solutions for the Installation and Management of Clusters of PCs under Linux*, R. Leiva http://heppc11.ft.uam.es/galera/doc/ATL-SOFT-2003-001.pdf.

[5] *Rocks homepage*, http://rocks.sdsc.edu/Rock.

[6] *NPACI All Hands Meeting, Rocks v2.3.2 Tutorial Session*, March 2003, http://rocks.sdsc.edu/rocks-documentation/3.0.0/talks/npaci-ahm-2003.pdf