



How to build and run a 128 Processors Cluster

Stefano Martinelli
(System Management Group)
CINECA

January 31 -
February 15 2002



ICTP - School in HPC on Linux
Clusters

1



Overview

- Hardware components
- Software components
- Applications
- Problems
- Future plans

January 31 -
February 15 2002



ICTP - School in HPC on Linux
Clusters

2



IBM Linux Cluster (Beowulf)

Features

- Model bunch of x330
- Architecture SMP Cluster
- CPU Pentium III 1.13GHz
- Number of PEs 128
- Peak Perf. 289 Gflops
- RAM 64 GB
- OS Linux RedHat 7.1 based



January 31 -
February 15 2002



ICTP - School in HPC on Linux
Clusters

3



IBM Beowulf proposal

The Beowulf model, architecture and installation procedures followed what was illustrated in a very interesting IBM RedBook:

“Linux HPC Cluster Installation”

<http://www.redbooks.ibm.com/redbooks/SG246041.html>

That machine had no “product name”
⇒ Was considered just a bunch of nodes
(now it has a name: “1300 Linux Cluster”)

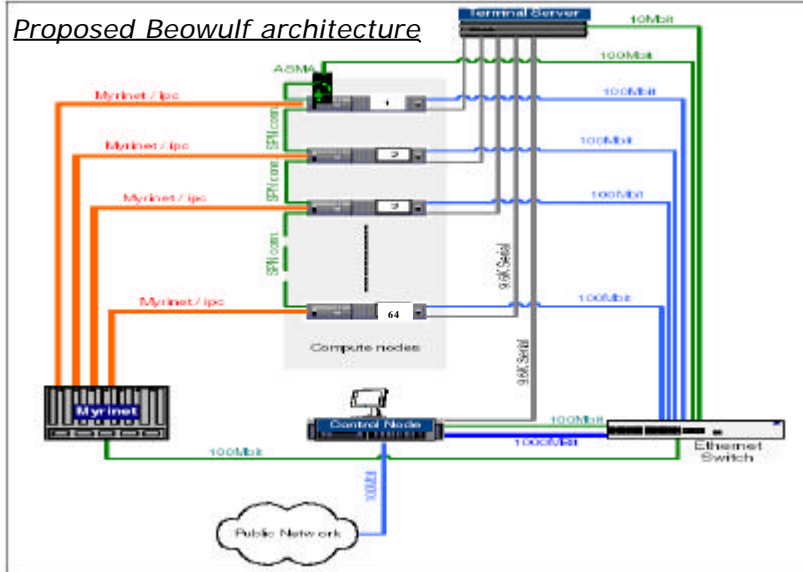
January 31 -
February 15 2002



ICTP - School in HPC on Linux
Clusters

4

Proposed Beowulf architecture



January 31 -
February 15 2002



ICTP - School in HPC on Linux
Clusters

5

Hardware Beowulf components

Quantity	Component
1	x340: 2-ways Pentium III 1 GHz, 2 GB RAM RAID 5 SCSI array
64	x330: 2-ways Pentium III 1.13 GHz ,1 GB RAM one 40 GB – 10000 rpm EIDE disk
1	64-port Myrinet switch with copper connections
64	Myrinet2000 LAN cards "C" version
2	48-port Cisco switches (46 fast + 2 Gbit)
4	16-port Equinox terminal servers
8	ASMA adapters
1	Gb ethernet NIC

January 31 -
February 15 2002

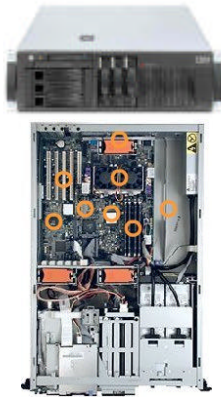


ICTP - School in HPC on Linux
Clusters

6

Nodes

Master node:
x340



January 31 -
February 15 2002



ICTP - School in HPC on Linux
Clusters

Compute nodes:
x330



7

Nodes common features

- Intel based motherboards
- 2 integrated Fast ethernet interfaces
- onboard Service Processors, also called "Advanced System Management" (ASM) processors, for:
 - ✓ Power management
 - ✓ Environmental monitoring
 - ✓ Alerts

ASM processors can be controlled using an "Advanced System Management Adapter" (ASMA) card

January 31 -
February 15 2002

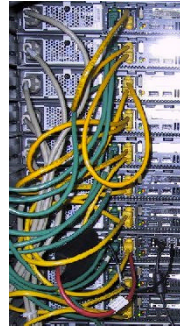


ICTP - School in HPC on Linux
Clusters

8

ASMA card

- Used for controlling ASM processors.
- Up to 11 ASM procs can be daisy chained and connected to an ASMA card
- An ASMA card can be physically inserted in a PCI slot, but can be (and usually is) independently powered
- An ASMA card has an ethernet connection from which can be reached



January 31 -
February 15 2002



ICTP - School in HPC on Linux
Clusters

9

Graphical Consoles

A KVM switch is supplied



x340 is directly plugged into
KVM switch

x330 has special interface for
KVM switch

- You daisy chain the x330 KVM ports
- Connect the chain to a central KVM switch
- Push on button on the node front to select the node



January 31 -
February 15 2002

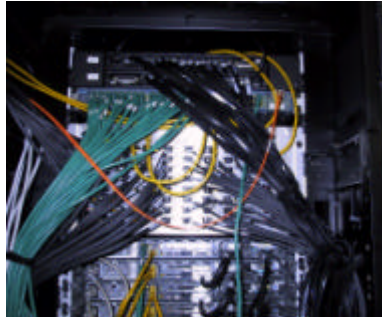


ICTP - School in HPC on Linux
Clusters

10

Myrinet IPC network

PCI slots on x330 are 64 bits 33 MHz
⇒ Maximum nominal Bandwidth is 264 MB/s
Measured MPI Bandwidth is ~150 MB/s
Measured MPI latency is ~40 μ s



January 31 -
February 15 2002



ICTP - School in HPC on Linux
Clusters

11

Equinox ELS Terminal Server/1

Goal: have high level of remote control on compute nodes

Terminal Server connects many terminals using a serial line to a LAN through a network connection

Benefits:

- We can access node console remotely even if TCP/IP is down or in single user mode
- MAC address collection during installation can be easily performed

⇒ Not vital, but useful



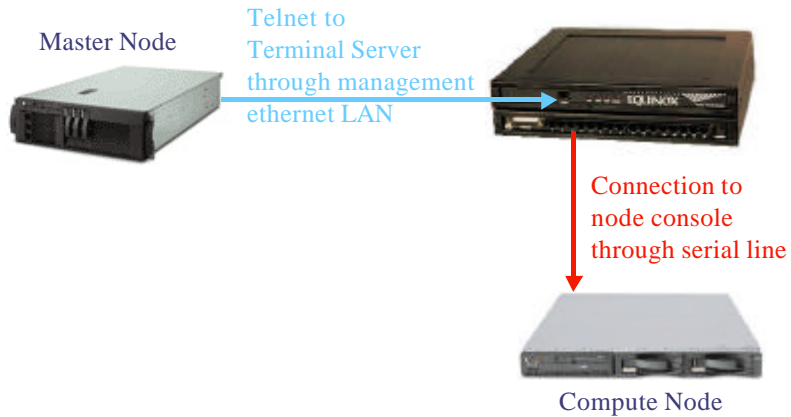
January 31 -
February 15 2002



ICTP - School in HPC on Linux
Clusters

12

Equinox ELS Terminal Server /2



Front and rear view





Software Beowulf architecture

- xCAT for installation and management
- Mpich_gm for parallel environment
- OpenPBS for batch system
- MAUI as scheduler
- PGI compilers
- NFS for HOMEs and local scratches



Installation adventure /1

- 1/10/2001: Components arrived (to be assembled) but the Equinox serial cables (which are not standard serial cables) were wrong
- 15/10/2001: nodes were patched and myrinet card inserted
- 29/10/2001 : beginning of assembling, cabling, and software installation. Software installation had problems for:
 - ✓ Inappropriate official RedHat 7.1 CD images (not correctly read by CD drive => solved using a laptop and subsequent download)
 - ✓ Problems with OpenSSH on which xCAT is based for cluster wide commands
- 15/11/2001 : We (CINECA) solved SSH configuration problems: use of ssh protocol 2 was not foreseen!!
- 21/11/2001: We (CINECA) solved substituted ASMA configuration problem: blanks were appended to node names!!
- 28/11/2001: We (CINECA) proposed access problem resolution (mpiexec)





Installation adventure /2

- 18/12/2001: Equinox cables arrived
- 19/12/2001: After some tests, we realized that there was a problem with mpich_gm. Asking to myrinet support people, we realized that xCAT scripts for mpich_gm compilation did not contained a '-noromio' flag, which have well known bad consequences
- 20/12/2001: Preparing for GPFS installation
We (CINECA) discovered that DMA was not active on compute nodes: I/O performance was ~3 MB/s on local scratch FS!!! The kernel provided by IBM people (2.4.5-hpc2) was compiled basically for SCSI, with no EIDE DMA support!!
We (CINECA) discovered that the IP myrinet address and interface names were wrongly set: the xCAT script could not configure the IP interface for Myrinet cards!!



Installation adventure /3

- 21/12/2001: Equinox configuration "apparently" completed. The day after, we (CINECA) discovered that the same console applications (very bad in quality) worked even if Equinox cables were unplugged !!! (we discovered ASMA can provide some console functionality)
- 22/12/2001: GPFS installation
During GPFS rpm installation, we realized that on some nodes the 'rpm -qa' command exited with a very bad "segmentation fault: core dump"
We decided to use a 2.4.2 base RedHat kernel for the compute nodes!!!!





Installation summary

14 February 2002:

- installation is not complete yet!!!
- The system is not in a complete production regime due to instability problems

Broken parts found within the first month (child mortality):

- 1 disk
- 1 ASMA adapter
- 2 motherboards
- 1 myrinet card
- 1 myrinet switch port

14 February 2002: NO more broken parts found since then.



Assistance and support

Hardware:

- Standard IBM for IBM parts
- CISCO for switches
- Myricom for Myrinet components
- Equinox resellers for Equinox Terminal servers

Software:

- PGI assistance for compilers
- IBM for... (not clear yet)





Our contribution

Hardware:

- **Added FC board for connection between x340 and IBM ESS (Shark) for storage**

Software

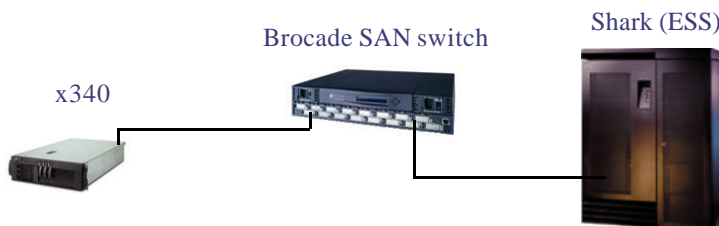
- OpenPBS and MAUI configuration
- **xCAT customization**
- LVM on master node
- Rdist for account files propagation
- Mpiexec as PBS parallel job launcher
- DMA (Direct Memory Access) activation
- **"Unix stile" daily Accounting for individual nodes and for the entire cluster (written by our own)**
- **IBM GPFS configuration for global scratch**



FC connection to storage

- A Qlogic 2200 FC adapter has been inserted in x340 Master node, and connected to our SAN through a Brocade switch.
- Some disk space has been reserved from our ESS
- The driver has been downloaded (<http://www.qlogic.com/support/>) and installed successfully

NO problems have been found: Linux sees the LUN as another SCSI device





xCAT customization

We customized:

- Kickstart installation file in different parts
- Ssh host key collection
- GM routes file distribution
- Installation “firstboot” scripts have been introduced for post install activities:
 - ✓ After installation a “firstboot” script is written in /etc of client node
 - ✓ During the first boot after installation this script is executed



“Unix style” daily accounting

What was present:

- Every node has ‘pacct’ fileset installed.
- Accounting activation records in a file (usually /var/log/pacct) all data relative to ALL completed processes
- The ‘sa’ command can be used to read and manipulate the accounting file

What had to be done :

- A ‘runacct’ script to be run daily by each compute node:
 - ✓ Summarize pacct data for that day on a user basis
- A ‘crunacct’ script to be run daily by the master node
 - ✓ Collect pacct data from all compute nodes, and produce a single accounting file for each day





GPFS

Goal:

- Have a scalable global scratch FS
- Consolidate the local disk space (~40 GB x 64 !!)

Features:

- Release 1.1.1 has been tried
- RedHat 7.1 kernels are required
- Limit of 32 nodes was bad for full use (we could not get the 1.2 beta from the labs, where limit is 128 nodes!)

Activity:

- We just played with it, and tested it
- GPFS has been installed on half of the whole machine (32 nodes)
- We've taken a 20 GB disk partition from each node
- We configured GPFS for replicating twice data and metadata for reliability => if one node is down, the FS should be UP!

Benchmarks were stopped because of our problems on local disks



FS configuration on compute nodes

- HOME directories are NFS mounted
- Local scratches have 2 path names:
 - /scratch_local
 - /global/nodeXX/scratch_local -> /scratch_local
- For first 32 nodes, a GPFS FS is available:
/scratch_p1



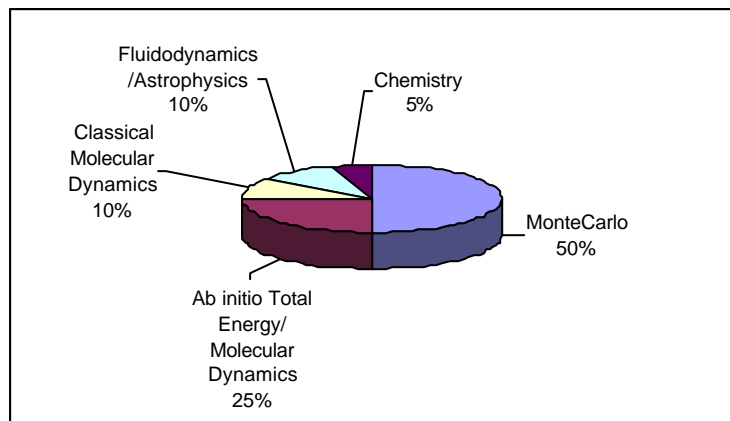


FS configuration on master node

- HOME directories are local
- Local scratch available through NFS using Autofs with path name:
/global/nodeXX/scratch_local
- GPFS /scratch_p1 is NFS mounted



Typical Applications





Actual Problems

- Checksum error and DMA
 - On some nodes, when I/O activity on local disks is intense:
 - a Checksum Error appears
 - DMA is consequently disabled
 - Hardware or software problem??
- Nodes hang
 - Sometimes some compute nodes “hang” in strange way:
 - You can login, but...
 - ‘ps’ command never finishes
 - ⇒kernel problem on process table
 - ⇒A power cycle is needed
- Client NFS activity of master node
 - If connection with servers stop (eg. for reboot):
 - Sometimes NFS client hangs
 - ⇒no more mounts are possible
 - ⇒Reboot is the only solution



Future plans and activities

- Solve actual problems first!!
- Move login node on a x330 node, away from the x340 management node
- Install proper monitoring tool (netsaint)
- Setup a reliable recovery procedure
- Install GPFS 1.2 on all 64 nodes
- Take GPFS disk from external storage using Fibre Channel connectivity



Information about the system

- News and configurations:
 - ✓ <http://www.cineca.it/sistemi>
- Configurations, HOWTOs and manuals:
 - ✓ <http://www.cineca.it/HPSystems/Resources/>



Load information

