Davide Tacchella
CSCS

CUG – Lugano 2006

Extending the connectivity of XT3 system nodes

1 XT3 nodes

On XT3 system there are 2 kinds of blades, **Compute blades** and **Service blades**.

Service blade nodes can be specialized for different purposes, main destinations are:

- **Boot node** (1)

  This node is the first node that is booted on the system, all Linux nodes mounts then the shared root file system from this host
- **SDB node** (1)
  This node is the second node booted on the system, it holds the databased with all informations about compute node status and part of service node software configuration.
- Luster OSS/MDS
- Login node
- PBS MOM node

This paper is focusing mainly on **Boot node** and **SDB node** and shows advantages that CSCS had by extending network connectivity on those nodes.

2 Boot node

2.1 Original configuration

Hardware, PCI- X adapters:

- 1 dual port FC adapter
  XT3 service nodes do not have internal disk, this adapter allows the boot node to have access to disks; these disks holds system OS and shared root file system needed for all others Linux nodes on the system.
- 1 single port Gigabit Ethernet
  This adapter provides connectivity to boot node from SMW, management software requires this link in order to obtain and update node status from/to system database.

Software:

- NTP
  NTP is required since XT3 hardware do not have any persistent memory for maintaining time informations across reboots.
- PGI license manager
  PGI is the default compiler for XT3 system

Default network connectivity:

- connection to SMW
- connection to XT3 High Performance Network

Connectivity limitations:

- NTP server used by boot node is SMW.
- Backup of local partitions is possible only when machine is down and file systems are mounted on SMW.
  All system disks can be mounted from **SMW, boot node, SDB node**, since used file system (ext3) is not a shared file system, in order to back it up the machine must be shut-down.
- Access to boot node is possible from SMW or service nodes with external connectivity only.

2.2 Possible solutions

- Use SMW or an XT3 service node as router
  Using SMW as router could be feasible, but since we have 2 XT3 systems and all systems use the same private subnets, this is not a possible solution; if the kernel provided with SMW would have NAT support this could have been a possible solution.
- Configure NAT on SMW or on an XT3 service node
  As previously said, the SMW default kernel do not have NAT support, while using other XT3 service nodes as NAT host could have been done, we have decided to try first the dual port Ethernet card solution.
- **Replace Ethernet card with a dual port**
  Boot node is delivered with Intel MT Server Gigabit card, Intel builds with the same chipset other two models of this card, a Dual port card and a Quad port card.
  At the moment of the decision we first choose the 2 port card, since we didn't know what to do with additional 2 ports. We are pretty confident that the quad port card will work as well, but we didn't test it.

2.3 Adopted solution

Replace single port Gigabit Ethernet card with dual port card; installed card is “Intel PRO/1000 MT dual port Server Adapter”.

First port is connected to SMW

Second port is connected to site network; IP address assigned to this interface is from CSCS public range.

2.4 Gains from extending connectivity

- Backup of local partitions while the machine is up and running
- Direct connection to site NTP servers
- Remote boot node administration
- Email functionality

3 SDB node

3.1 Original configuration

Hardware, PCI- X adapters:

- 1 dual port FC adapter
  This adapter provides connectivity to disks used on SDB node for storing MySQL database files and syslog informations from all Linux nodes

Software:

- NTP
- MySQL
- centralized syslog server
- PBS pro server
- PBS pro scheduler

Default network connectivity:

- connection to XT3 High Performance Network

Connectivity limitations:

- NTP server is Boot node
  This is not a real limitation, but in order to use boot node as NTP server the boot process has to wait until boot node is able to become a server suitable for synchronization, and this process takes time (if **iburst** parameter is set it takes approximately 1 minute, if **iburst** is not set it takes longer)
- Backup of local partitions is possible only when machine is down and

filesystems are mounted on SMW

- PBS can't send email notification
  One functionality that users appreciate is email notification feature, if SDB node do not have external connectivity, this feature cannot be used.
- Job statistics are held on MySQL database; this database is accessible from XT3 nodes only
  XT3 system maintains job statistics on SDB, we need these information to update our accounting DB.
- PBS jobs priority can't be set using CSCS accounting DB
  CSCS has an accounting DB that holds information about allocated and used CPU time; information from this DB are used by the modified PBS scheduler to define job priority.

3.2 Possible solutions

- Configure NAT on an XT3 service node
  SDB node has connection to Seastar network only, configuring NAT on a service node with external connectivity could be feasible, but since this node still has a free PCI-X slot we decided first to add a Gigabit card.
- **Add Ethernet card**

3.3 Adopted solution

Insert into PCI-X free slot Gigabit Ethernet card removed from Boot node, IP address assigned to this interface is from CSCS public range.

3.4 Gains from extending connectivity

- Access from external accounting system
- Access to external accounting system
- Direct connection to site NTP servers
  This way SDB node has connection to site NTP servers and the boot process do not require to wait for boot node to be suitable for time synchronization.
- Increased SMW functionality
  - XT3 CPU and job status from SMW (xtprocadmin, xtshowcabs)
    with default configuration, xtprocadmin and xtshowcabs do not work from SMW because they do query the DB on SDB node, with the installed Ethernet card on SDB node and by modifying the file **/etc/xt.conf** on SMW these two commands can be run on SMW too.
- PBS email functionality
  PBS is able to send email notification as jobs starts/ends, this feature is appreciated by many users, and with the additional node connectivity this feature is functional.

4 How CSCS operates today

At this moment we can do 90% of management work from SMW, only administration of shared root and OS upgrade still needs to be done from boot node.

Users home directories are hosted by an external server and are shared among our 2 XT3 systems. (single cabinet system, 12 cabinet system)

5 Acronyms used

SMW	: System Management Workstation
SDB	: System DataBase
NTP	: Network Time Protocol
NAT	: Network Address Translation
PBS	: Portable Batch System
MOM	: Mom is a PBS service that starts and controls jobs
OSS	: Lustre Object Storage Server
MDS	: Lustre Meta-Data Server