(PXE)Boot over IB

OFA user day workshop
Monterey, CA
2013-04-19

Florent.Parent@calculquebec.ca
Boot over IB

• Why?
  • Colosse: Sun 6048 cluster, 960 nodes (installed late 2009)
  • Node has no local hard drive
  • Node has Infiniband interface only (Ethernet for ILOM only)
  • Boot over Ethernet not even an option :)

Significant decrease on cabling infrastructure requirement

Cables per rack:
• 16 CXP 12x
• 6 Ethernet
IB network

- M9 648-ports core switch
- 24 QDR links
- 12 QDR links
- 40 shelves in 10 racks
- 960 2-sockets nodes
- 11 infrastructure nodes
- Lustre MDS
- 5 Lustre OSS HA pairs
- 7 Lustre OSS HA pairs
- 11 infrastructure nodes
Compute node

- **IB HW/FW**
  - ConnectX MT26428 (rev a0) (on board)
  - firmware: 2.7.8100
  - FlexBoot 3.0.000, gPXE 0.9.9+ (Sun/Oracle patch SW 2.7)

- **Software**
  - CentOS 5.8 (kernel 2.6.32)
  - OFED 1.4.2
  - OpenSM 3.2.6 (+ sun_patch_2.4)
Provisioning

• OneSIS and Cobbler

• One boot server per 192 nodes

• Boot server PXEboot from boot “master”

• Local disks on boot server NFS shared to compute nodes
Boot process

**IB link up**
Node boots

**DHCP discover (bcast)**

**DHCP offer (bcast)**

**request**

**ACK**

**TFTP**

**load "pxelinux.0"**

**exec "pxelinux.0"**

**load "pxelinux.cfg/config"**

**load kernel + initrd**

**exec kernel**

**IB link down**

**IB link up**

NOTE: iPXE supports http for file download
DHCP config

- DHCP entry for each node (Cobbler)
- Client identifier field holds IPoIB HW address
- DHCP patch required

```plaintext
host r102-n28 {
    option dhcp-client-identifier = ff:00:00:00:00:02:00:02:c9:00:50:80:02:00:00:8d:64:51;
    fixed-address 10.225.102.28;
    option subnet-mask 255.255.0.0;
    filename "pxelinux.0";
    next-server 10.225.101.100;
    option pxelinux.configfile "pxelinux.cfg/r102-n28";
}
```

RFC 4390: Dynamic Host Configuration Protocol (DHCP) over InfiniBand
* Mellanox FlexBoot User Manual*
Some config files

/etc/dhcpd.conf

host r102-n28 {
  option dhcp-client-identifier = ff:00:00:00:00:02:00:02:c9:00:50:80:02:00:00:8d:64:51;
  fixed-address 10.225.102.28;
  option subnet-mask 255.255.0.0;
  filename "/pxelinux.0";
  next-server 10.225.101.100;
  option pxelinux.configfile "pxelinux.cfg/r102-n28";
}

/tftpboot/pxelinux.cfg/r102-n28

prompt 0
timeout 1
default cn-20130412
label cn-20130412
  kernel /images/cn-20130412/vmlinuz-2.6.32.40-clumeq
  append initrd=/images/cn-20130412/initrd-2.6.32.40-clumeq.img root=10.225.101.100:/var/lib/oneSIS/image/cn-20130412
MLNX FlexBoot 3.0.000 (PCI 02:00.0) starting execution
MLNX FlexBoot 3.0.000 initialising devices...

Mellanox ConnectX FlexBoot v3.0.000
gPXE 0.9.9+ -- Open Source Boot Firmware -- http://etherboot.org

net0: 50:80:02:00:00:8d:64:51 on PCI02:00.0 (open)
  [Link:down, TX:0 TXE:0 RX:0 RXE:0]
  [Link status: Not connected (0x30006001)]
Waiting for link-up on net0... ok
DHCP (net0 50:80:02:00:00:8d:64:51).... ok
net0: 10.225.102.28/255.255.0.0 gw 10.225.3.14
Booting from filename "/pxelinux.0"
tftp://10.225.101.100//pxelinux.0..._
Observations

- Catch-all boot image: minimal kernel
  - Outputs IB interface information to syslog
  - Useful for new blade (replacement): Old PXE code. Boot timeouts. Need reflashing
- A lots of timeouts when booting too many nodes (e.g. after short power outage)
  - Timeouts on “Waiting for link-up on net0”. Often solved by kicking OpenSM. Too many “link state changes’?
In the todo stack...

- Test more recent PXE and IB FW code
- Upgrade to recent OFED release
  - Need to test SM (ibsim?)
Open discussion

- Boot firmware [http://ipxe.org/](http://ipxe.org/) (replaces gPXE)
- [http://www.mellanox.com/related-docs/prod_software/Linux_PXE_Installation_over_IPoIB_README.txt](http://www.mellanox.com/related-docs/prod_software/Linux_PXE_Installation_over_IPoIB_README.txt)