Node Provisioning: Timeouts during loading of InfiniBand drivers make my filesystem clients fail. Is there a workaround?

Yes.

Driver modules load in a hierarchy (a dependency chain), and delays due to non-standard use can cause timeouts as one particular driver module may not be able to find the one that it depends upon. This can prevent nodes or applications that rely on InfiniBand drivers from running.

As the administrator manual explains in the chapter on administration, in the section on "InfiniBand Provisioning", a list of modules that InfiniBand (IB) loads up with on a fully booted system can be seen with:

```
modlist(){ cut -f1 -d" " /proc/modules; }
IB=/etc/init.d/rdma
diff <($IB stop; modlist) <($IB start; modlist)
```

rdma in the above is replaced by openibd when using SLES or distributions based on versions of Red Hat prior to version 6.

For SLES the output may display something like:

```
1c1,25
< Unloading HCA driver:                                 [  OK  ]
---
> Loading HCA driver and Access Layer:                  [  OK  ]
> Setting up InfiniBand network interfaces:
> Setting up service network . . .                      [  done  ]
> rdma_ucm
> ib_srp
> scsi_transport_srp
> scsi_tgt
> ib_sdp
> rdma_cm
> iw_cm
> ib_addr
> ib_ipoib
> ib_cm
> ib_sa
> ib_uverbs
> ib_umad
> iw_cxgb3
> cxgb3
> mdio
> mlx4_en
> mlx4_ib
```

> mlx4_core
> ib_mthca
> ib_mad
> ib_core

The exact modules used depend on the hardware., and in some cluster configurations, the default order in which these modules are loaded may cause timeout issues for clients requiring services based on these modules. Juggling the allowed possibilities in module load order can solve the issue.

For example:

When using ipoib, the ib0 interface comes up very late with a default-image order for SLES 11. If using NFS or FhGFS client (both of these are filesystems), these take a long time to mount during boot. For FhGFS, /var/log/messages shows something like:

[67.357825] microcode: CPU11 sig=0x206c2, pf=0x1, revision=0x13
[67.361040] Microcode Update Driver: v2.00 <tigran@aivazian.fsnet.co.uk>, Peter Oruba
[73.343947] mlx4_ib: Mellanox ConnectX InfiniBand driver v1.0 (April 4, 2008)
[73.505545] NET: Registered protocol family 27
[73.976198] igb: eth0 NIC Link is Up 1000 Mbps Full Duplex, Flow Control: RX/TX
[73.977491] ADDRCONF(NETDEV_UP): eth0: link is not ready
[73.978611] ADDRCONF(NETDEV_CHANGE): eth0: link becomes ready
[74.315349] ib0: enabling connected mode will cause multicast packet drops
[74.317539] ib0: mtu > 2044 will cause multicast packet drops.
[74.320750] ADDRCONF(NETDEV_UP): ib0: link is not ready
[74.768008] fhgfs_opentk: modprobe: FhGFS OpenToolkit loaded. (ibverbs enabled)
[74.770530] fhgfs: module license 'Proprietary' taints kernel.
[74.770531] Disabling lock debugging due to kernel taint
[74.773613] fhgfs: modprobe(5401): File system registered. Type: fhgfs. Version: 2011.04-r12
[84.939662] eth0: no IPv6 routers present
[105.801741] fhgfs: mount(5415): Mount sanity check failed. Canceling mount. (Log file may provide additional information. Check can be disabled with sysMountSanityCheckMS=0 in the config file.)
[107.738263] ip_tables: (C) 2000-2006 Netfilter Core Team
[111.281702] BIOS EDD facility v0.16 2004-Jun-25, 1 devices found
[114.563957] ADDRCONF(NETDEV_CHANGE): ib0: link becomes ready
[125.356702] ib0: no IPv6 routers present
[126.628771] fhgfs: mount(6105): Mount sanity check failed. Canceling mount. (Log file may provide additional information. Check can be disabled with sysMountSanityCheckMS=0 in the config file.)
[148.973014] fhgfs: rmmod(6245): fhgfs_client unloaded.

[148.974373] fhgfs_opentk: rmmod: FhGFS OpenToolkit unloaded.
[148.987500] fhgfs_opentk: modprobe: FhGFS OpenToolkit loaded. (ibverbs enabled)
[148.991039] fhgfs: modprobe(6278): File system registered. Type: fhgfs. Version: 2011.04-r12
[149.034458] fhgfs: mount(6292): FhGFS mount ready.

The timestamps indicate the FhGFS client must be started later on, since the IB drivers take a long time to come up. Or equivalently, the IB drivers can be loaded up earlier.

If ib_ipoib is loaded in the initrd, it will shorten the time a bit, but ib0 will still come up too late.

However, affter adding all the ib_ipoib andmlx4_ib dependencies as well (and moving them up a bit in the order and moving ib_ipoib down), ib0 comes up immediately and there are no failures on boot for FhGFS-client and for NFS over IB mounts. The cmsh commands for this are something like this, and in this order:

[slcmg1->softwareimage[vnc-image]->kernelmodules]% add ib_core
[slcmg1->softwareimage*[vnc-image*]->kernelmodules*[ib_core*]]% add ib_cm
[slcmg1->softwareimage*[vnc-image*]->kernelmodules*[ib_cm*]]% add ib_sa
[slcmg1->softwareimage*[vnc-image*]->kernelmodules*[ib_sa*]]% add ib_mad
[slcmg1->softwareimage*[vnc-image*]->kernelmodules*[ib_mad*]]% add mlx4_core
[slcmg1->softwareimage*[vnc-image*]->kernelmodules*[mlx4_core*]]% add mlx4_ib
[slcmg1->softwareimage*[vnc-image*]->kernelmodules*[mlx4_ib*]]%

So the module list order is:
mlx4_core
mlx4_en
mlx4_ib
ib_core
ib_mad
ib_sa
ib_cm

and at the bottom:
ib_ipoib

/var/log/messages then shows

...

Node Provisioning: Timeouts during loading of InfiniBand drivers make my filesystem clients fail. Is there a workaround?

[78.709193] ib0: enabling connected mode will cause multicast packet drops
[78.711345] ib0: mtu > 2044 will cause multicast packet drops.
[78.714563] ADDRCONF(NETDEV_UP): ib0: link is not ready
[78.715726] ADDRCONF(NETDEV_CHANGE): ib0: link becomes ready
[79.152507] fhgfs_opentk: modprobe: FhGFS OpenToolkit loaded. (ibverbs enabled)
[79.164389] fhgfs: module license 'Proprietary' taints kernel.
[79.164390] Disabling lock debugging due to kernel taint
[79.167557] fhgfs: modprobe(5522): File system registered. Type: fhgfs. Version: 2011.04-r12
...
[94.080841] fhgfs: mount(5536): FhGFS mount ready.

Unique solution ID: #1022
Author: Frank Furter
Last update: 2012-05-22 15:22