

How to configure multipath for high availability and performance on Debian and CentOS for storage at IBM DS8300 SAN

Author: André Felipe Machado<andremachado@techforce.com.br>

This detailed how to guides to achieve high availability and performance on Debian and CentOS for accessing storage space at IBM DS8300 Data Storage Systems.

Tested on Debian GNU/Linux 5.x Lenny 64 bits and CentOS 5.3 64 bits running on 8 cores blades, with Host Bus Adapters Qlogic and Emulex Light Pulse Fiber Channel in [deployed systems at SERPRO](#)

Observations showed that Debian Lenny has the best performance, for our app load profile and hardware.

Also, there are listed a number of previously not clearly documented critical pitfalls to avoid. STUDY whole articles, hints, implications, and cited resources before planning your deployment.

Every detail matters.

Before start, you **must** have LUNs at IBM DS8300 storage configured for high availability and performance as explained at the article [How to configure maximum performance storage space for Debian GNU/Linux on IBM DS.8300 Data Storage Systems](#)

Multipath and storage basic concepts

In order to achieve high availability and high performance connections with the IBM DS8300 storage, with load balancing and fail over, you must to configure multipath carefully and keep some concepts in mind.

- The storage presents LUN volumes to the host **as SCSI devices** through the fiber channel HBA interfaces.
- Each LUN has its own "universally" **unique wwid**.
- **Both** HBA are connected to the **same wwid LUN**.
- **Each** HBA presents the **same wwid LUN as different SCSI drives** to the host.
- **Each** HBA has **its own wwpn** towards the storage (as it sees them).
- Each storage connection also has its **own host port name** for each HBA.
- After being correctly configured, **each** HBA **"see" all other LUNs** connected through **all other** HBA at the **same host**. But **only enable** them, keeping them **inactive** unless needed.
 - Therefore, **a given LUN** is presented by **2 HBAs** as **4 SCSI drives** to the host.
- The unfriendly names `/dev/dm-XY` are **symbolic links** to the **real mappings** at `/dev/mapper/*`.
- **Do not mount the SCSI drives directly** or you will not leverage the high availability and performance of multipath.
 - Worse, if you mount same wwid, but different host SCSI drives, at different mount points and options, you will likely mess the underlying file system.

What is the Host Bus Adapter model?

At SERPRO, there are deployed HBA models that work well with Emulex or QLogic drivers.

You must be **absolutely sure** about which ones are installed in the host.

You

must not mix different boards.

Failure to identify correct model and suitable driver
WILL cause kernel lock up during next boot.

Configure APT to include network and QLogic HBA drivers

You must follow the configuration steps at [How to configure APT for multiple repositories and sections](#) AND only install the driver if they are the actually installed HBAs.

*Failure to identify correct model and suitable driver
WILL
cause kernel lock up during next boot.*

```
# apt-get install multipath-tools-boot multipath-tools firmware-qlogic  
firmware-bnx2 sysfsutils  
# reboot
```

The packages `firmware-qlogic` and `firmware-bnx2` (a network driver, if your host use this chipset) are in the non-free repository section, disabled by default Debian installation.

Configure APT to include network and Emulex Light Pulse Fiber Channel HBA drivers

The Emulex Light Pulse Fiver Channel (`lpfc`) is already a FOSS driver accepted into kernel source tree, and is available at Debian default installation.

You must follow the configuration steps at [How to configure APT for multiple repositories and sections](#) AND only install the driver if they are the actually installed HBAs.

Failure to identify correct model and suitable driver

WILL

cause kernel lock up during next boot.

```
# apt-get install multipath-tools-boot multipath-tools firmware-bnx2
sysfsutils
# reboot
```

The package firmware-bnx2 (a network driver, if your host use this chipset) is in the non-free repository section, disabled by default Debian installation.

Pitfalls

- *Failure to identify correct model and suitable driver*

WILL

cause kernel lock up during next boot.

- Without reboot, it will not correctly identify the boards nor map them correctly at device-mapper. They will **SEEM** to work and will **change mappings** at next boot, confusing your configurations.
- Without the correct packages installed and suitable repository sections enabled, initrd system mappings will NOT be configured correctly during next kernel upgrade.
- During initial system installation, the boards with non-free drivers were detected, you may had installed the BLOB non-free drivers with a pen-drive or other secondary media/method, but the /etc/apt/sources.list may had not be [configured to enable non-free sections](#) **must** verify this.
- Maybe, a kernel module unload/load cycle could identify the boards, but this is a valuable opportunity to check your configuration files for an unattended reboot event.

Verifying that the correct Linux kernel module was loaded

Emulex driver

Below there is an example of what you should expect to find at your host:

```
debian:~# cat /var/log/dmesg | grep Emulex
[ 19.505676] Emulex LightPulse Fibre Channel SCSI driver 8.2.6
[ 19.505676] Copyright(c) 2004-2008 Emulex. All rights reserved.
debian:~#
debian:~# cat /var/log/dmesg | grep lpfc
[ 32.076790] lpfc 0000:10:00.0: 0:1303 Link Up Event x1 received
Data: x1 xf7 x8 x0
```

```

[ 54.856504] lpfc 0000:07:00.1: 2:1303 Link Up Event x1 received
Data: x1 xf7 x10 x0
debian:~#
debian:~# lsmod |grep lpfc
lpfc                243060  8
scsi_transport_fc   49668  1 lpfc
scsi_mod            160760  9
sr_mod,ses_sd_mod,sg,libata,lpfc,scsi_transport_fc,aacraid,scsi_tgt
debian:~#

```

QLogic driver

The qla2xxx driver is at the non-free repository section at Debian.

```

debian:~# cat /var/log/dmesg |grep qla
[ 3.201082] qla2xxx 0000:10:03.0: Found an ISP2312, irq 160, iobase
0xfffffc20000061a000
[ 3.201082] qla2xxx 0000:10:03.0: Configuring PCI space...
[ 3.201082] qla2xxx 0000:10:03.0: Configure NVRAM parameters...
[ 3.294125] qla2xxx 0000:10:03.0: Verifying loaded RISC code...
[ 3.438017] qla2xxx 0000:10:03.0: Allocated (412 KB) for firmware
dump...
[ 3.489479] scsi1 : qla2xxx
[ 3.492172] qla2xxx 0000:10:03.0:
[ 3.492172] qla2xxx 0000:11:04.0: Found an ISP2312, irq 224, iobase
0xfffffc20000061c000
[ 3.492172] qla2xxx 0000:11:04.0: Configuring PCI space...
[ 3.492172] qla2xxx 0000:11:04.0: Configure NVRAM parameters...
[ 3.585353] qla2xxx 0000:11:04.0: Verifying loaded RISC code...
[ 3.697266] qla2xxx 0000:11:04.0: Allocated (412 KB) for firmware
dump...
[ 3.841405] scsi2 : qla2xxx
[ 4.018499] qla2xxx 0000:11:04.0:
[ 4.019634] qla2xxx 0000:11:04.1: Found an ISP2312, irq 225, iobase
0xfffffc20000061e000
[ 4.019736] qla2xxx 0000:11:04.1: Configuring PCI space...
[ 4.019930] qla2xxx 0000:11:04.1: Configure NVRAM parameters...
[ 4.112595] qla2xxx 0000:11:04.1: Verifying loaded RISC code...
[ 4.271948] qla2xxx 0000:11:04.1: Allocated (412 KB) for firmware
dump...
[ 4.405866] qla2xxx 0000:10:03.0: LOOP UP detected (2 Gbps).
[ 4.453005] scsi3 : qla2xxx
[ 4.453005] qla2xxx 0000:11:04.1:
[ 5.368773] qla2xxx 0000:11:04.1: LIP reset occurred (f7f7).
[ 5.478246] qla2xxx 0000:11:04.1: LOOP UP detected (2 Gbps).
debian:~#

```

```

debian:~# cat /var/log/dmesg | grep QLogic
[ 2.180303] QLogic Fibre Channel HBA Driver: 8.02.01-k4
[ 3.492172] QLogic Fibre Channel HBA Driver: 8.02.01-k4
[ 3.492172] QLogic QLA2340 - 133MHz PCI-X to 2Gb FC, Single Channel
[ 4.018501] QLogic Fibre Channel HBA Driver: 8.02.01-k4
[ 4.018503] QLogic QLA2342 - 133MHz PCI-X to 2Gb FC, Dual Channel
[ 4.453005] QLogic Fibre Channel HBA Driver: 8.02.01-k4
[ 4.457805] QLogic QLA2342 - 133MHz PCI-X to 2Gb FC, Dual Channel
debian:~#

```

Create the /etc/multipath.conf for the IBM

DS8300 storage

Starting from the [IBM DS8300 Reference](#) will modify it for our purposes.

Discover the wwid for the /etc/multipath.conf

As multipath is not yet correctly configured, the command below will return "undef" for some paths, as the example below. What we need now is to identify the wwid between parenthesis.

```
debian:~# multipath -l
mpath0 (36005076308ffc36c000000000000115f) dm-1 IBM ,2107900
[size=50G][features=1 queue_if_no_path][hwhandler=0]
\_ round-robin 0 [prio=0][active]
  \_ 3:0:0:0 sdb 8:16 [active][undef]
  \_ 3:0:1:0 sdc 8:32 [active][undef]
  \_ 5:0:0:0 sdd 8:48 [active][undef]
  \_ 5:0:1:0 sde 8:64 [active][undef]
```

Your /etc/multipath.conf could become similar to the example below for the IBM DS8300 storage system.

```
#####
#####
# Multipath.conf file for IBM Storage
#
# Version 3.01
# AFM 09dez2009 for DS8300 1.2.3.4
# SERPRO > SUPSI > SIETE > SIECE
#
# IMPORTANT: If you change multipath.conf settings after the DM MPIO
devices
# have already been configured, be sure to rerun "multipath".
#####
#####
#
#
# defaults:
#
#   polling_interval   : The interval between two path checks in
seconds.
#
#   failback           : The failback policy should be set to
"immediate"
#
#                       to have automatic failback, i.e. if a higher
#                       priority path that previously failed is
restored,
#
#                       I/O automatically and immediately fails back
to
#
#                       the preferred path.
#
#   no_path_retry      : Use this setting in order to deal with
transient
```

```

#           total path failure scenarios. Indicates that
the if
#           all paths are failed for 10 checks
# (iterations of
#           the checkerloop) then will set the device to
#           .fail_if_no_path. so that I/O will not stay
queued
#           forever and I/O errors are returned back to
the
#           application. This value should be adjusted
# based on
#           the value of the polling_interval.
Basically, with a
#           larger polling_interval, this means that the
# amount
#           of time of allowed total path failure will be
#           longer, since the tolerance time is
#           (no_path_retry * polling_interval) seconds.
#           SHOULD NOT BE USED WITH .features..
#
# rr_min_io           : The number of IOs to route to a path before
switching
#           to the next path in the same path group
#
# path_checker        : The default 'readsector0' path checker uses
SCSI
#           READ (opcode 0x28) which doesn't work in
clustered
#           environments. TUR (Test Unit Ready) does
work in
#           clustered environments with storage that
subscribes
#           to the SCSI-3 spec.
#
# user_friendly_names : With this value set to .yes., DM MPIO
devices will
#           be named as .mpath0., .mpath1., .mpath2.,
etc. ...
#           The /var/lib/multipath/bindings file is
#           automatically generated, mapping the
.mpathX. name
#           to the wwid of the LUN. If set to "no", use
the
#           WWID as the alias. In either case this be
will be
#           overridden by any specific aliases in this
# file.
#
#
defaults {
    polling_interval    30
    failback            immediate
    no_path_retry       5
    rr_min_io           100
    path_checker         tur
    user_friendly_names yes
}

#
# An example of blacklisting a local SCSI disk.
# Here a local disk with wwid SIBM-ESXSMAN3184MC_FUFR9P29044K2 is
# blacklisted and will not appear when "multipath -l(1)" is invoked.

```

```

#
#
#blacklist {
#    wwid SIBM-ESXSMAN3184MC_FUFR9P29044K2
#}

#
# An example of using an alias.
# NOTE: this will override the "user_friendly_name" for this LUN.
#
# Here a LUN from IBM storage with wwid
3600507630efffe32000000000000120a
# is given an alias of "IBM-1750" and will appear as "IBM-1750
#(3600507630efffe32000000000000120a)", when "multipath -l(l)" is
invoked.
#
#
#multipaths {
#    multipath {
#        wwid 3600507630efffe32000000000000120a
#        alias IBM-1750
#    }
#}

#
# devices      : List of per storage controller settings, overrides
default
#               settings (device_maps block), overridden by per multipath
#               settings (multipaths block)
#
# vendor       : Vendor Name
#
# product      : Product Name
#
# path_grouping_policy : Path grouping policy to apply to multipath
hosted
#               by this storage controller
#
# prio_callout : The program and args to callout to obtain a path
#               weight. Weights are summed for each path group to
#               determine the next PG to use case of failure.
#               NOTE: If no callout then all paths have equals weight.
#
#
devices {
# These are the default settings for 2145 (IBM SAN Volume Controller)
# Starting with RHEL5, multipath includes these settings be default
    device {
        vendor           "IBM"
        product          "2145"
        path_grouping_policy group_by_prio
        prio_callout     "/sbin/mpath_prio_alua /dev/%n"
    }

# These are the default settings for 1750 (IBM DS6000)
# Starting with RHEL5, multipath includes these settings be default
    device {
        vendor           "IBM"
        product          "1750500"
        path_grouping_policy group_by_prio
        prio_callout     "/sbin/mpath_prio_alua /dev/%n"
    }
}

```

```

# These are the default settings for 2107 (IBM DS8000)
# Uncomment them if needed on this system
    device {
        vendor                "IBM"
        product                "2107900"
        path_grouping_policy   group_by_serial
    }

# These are the default settings for 2105 (IBM ESS Model 800)
# Starting with RHEL5, multipath includes these settings be default
    device {
        vendor                "IBM"
        product                "2105800"
        path_grouping_policy   group_by_serial
    }
}

multipaths {
    multipath {
        wwid 36005076308ffc36c000000000000115f
        # alias disk1_50Gb
        path_grouping_policy failover
        path_selector "round-robin 0"
    }
}

devnode_blacklist {
    devnode "*"
}

blacklist {
    devnode "sda"
}

```

Please, notice that we are NOT using the friendly name defined at the alias parameter, because a personal configuration management preference. But you could define a friendly name given that is universally unique, *at least unique at your corporation.*

Also, notice the "blacklist" and "path_checker" parameters. They are suitable for this hardware.

Remove invalid LVM volumes and mappings

If you do not remove invalid LVM volumes, the "multipath -F" command will not release the invalid /dev/mapper/mpath* ones.

Before, you must move the /var/lib/multipath/bindings to another place. The file is useful for cluster setups, mostly, when all hosts must have same bindings to avoid confusion.

If you do not move it, old parameters will be read again,

```
# mv /var/lib/multipath/bindings /var/lib/multipath/bindings.bak
```

The following command, will remove the LVM mappings that do not exist anymore, and are inactive, allowing "multipath -F" to release them and allowing re-mappings.

Clean the LVM ghost mappings similarly to the example:

```
# dmsetup remove grupol-labpostfix--exp--02
```

Configure /etc/lvm/lvm.conf

Search for a similar snippet as the example below. But adapt it for your host, as it may have another local volumes outside the SAN.

```
# This is an example configuration file for the LVM2 system.
# By default we accept every block device:
# AFM 16oct2009 trying to filter duplicate volumes allowing only dm
#filter = [ "a./.*/" ]
# filter = [ "a|/dev/dm(\-[0-9])|", "r/sd.*/" , "r|.*|" ]
# AFM 22oct2009 use correct multipath devices instead of dm-X
filter = [ "a|^/dev/mapper/mpath.*|" , "r|.*|" ]
```

There are some useful commands to identify and remove invalid LVM volumes.
STUDY the man pages. You must use them ONLY if there are invalid mappings, and being extremely careful with such powerfull commands.

```
# vgreduce --test --removemissing vg01
# dmsetup ls --tree
# dmsetup info grupol-lvstripe14
# dmsetup info mpath0
# dmsetup targets
# dmsetup table --target multipath mpath0
# dmsetup table --target striped grupol-lvstripe14
```

Immediately after, you must generate a new initrd boot image.

```
# update-initramfs -u
```

******YOU HAVE TO REBOOT HOST****.**

The multipath tools daemon command line is too slow for correct mappings of all devices at the device-mapper.

Therefore, is not enough to execute:

```
debian:~:# invoke-rc.d multipath-tools restart
```

or

```
[centos]:# service multipathd restart
```

You MUST execute the following sequence:

```
debian:~# invoke-rc.d multipath-tools stop
debian:~# multipath -F
debian:~# reboot
```

or

```
# service multipathd stop
Stopping multipathd daemon: [ OK ]
# multipath -F
# ls /dev/mapper/
control
# reboot
```

Verify the mappings

You have to execute the following command and compare output with previous results.

This time, priorities should be assigned, devices activated, alternate paths enabled and ready.

```
debian:~# multipath -v2 -ll
mpath0 (36005076308ffc36c000000000000115f) dm-1 IBM ,2107900
[size=50G][features=1 queue_if_no_path][hwhandler=0]
\_ round-robin 0 [prio=1][active]
  \_ 1:0:0:0 sdb 8:16 [active][ready]
\_ round-robin 0 [prio=1][enabled]
  \_ 1:0:1:0 sdc 8:32 [active][ready]
\_ round-robin 0 [prio=1][enabled]
  \_ 3:0:1:0 sde 8:64 [active][ready]
\_ round-robin 0 [prio=1][enabled]
  \_ 3:0:0:0 sdd 8:48 [active][ready]
```

How discover the wwpn that SAN receive from your host

In order to configure the SAN, you must know the identifier that HBA presents to outside.

```
debian:~# cat /sys/class/fc_host/host1/port_name
0x210000e08b8e3020
debian:~# cat /sys/class/fc_host/host2/port_name
0x210000e08b18f6bb
```

How discover the interface id of the SAN

```
debian:~# cat /sys/class/fc_remote_ports/rport-1\:0-0/port_name
0x500507630808c36c
debian:~# cat /sys/class/fc_remote_ports/rport-1\:0-1/port_name
0x500507630838c36c
debian:~# cat /sys/class/fc_remote_ports/rport-3\:0-0/port_name
0x500507630803c36c
debian:~# cat /sys/class/fc_remote_ports/rport-3\:0-1/port_name
0x500507630833c36c
```

Create high performance LVM, RAID and file systems at the storage space

You have to create LVM, RAID and file systems for high performance over the multipathed devices mapped at /dev/mapper/mpath* as explained at another article, Configuration and tuning of LVM, RAID and filesystems for high performance and availability on SAN (to be published).

Resources

[IBM DS8300 /etc/multipath.conf reference file for RHEL 5.](#)

[Emulex IBM branded HBA drivers.](#)

[How to get wwpn from Emulex HBA on RHEL 5.3.](#)

[Calivia - Linux multipath IO.](#)

[Oracle VM Server Configuration- multipathed SAN storage.](#)

[FC SAN and lpfc driver.](#)