Setting up iSCSI Multipath in Ubuntu Server 12.04

A Dell Technical White Paper

Jose De la Rosa

Linux Engineering

August 2013

Contents

1. Introduction

1.1 Purpose of this document

This whitepaper describes how to setup and configure iSCSI Multipathing in Ubuntu Server 12.04 using a Dell PowerEdge server and Dell EqualLogic storage. I don't make recommendations or evaluate strategies for deploying a highly-available environment, but simply describe a step-by-step set of instructions to give you a working deployment. I leave all final tweaks up to you.

1.2 Assumptions & Disclaimers

It is assumed that the reader is familiar with iSCSI and multipathing concepts, as I do not discuss any theory or concepts here. Expertise with iSCSI is not required to successfully follow these instructions; however some practical experience will make it easier understanding all steps.

It is assumed that the reader is familiar with Ubuntu Server and with the Linux operating system in general. You don't have to be an expert, but some past background will be useful. I don't cover advanced topics such as booting from a multipath device or configuring a preseed file for automated Ubuntu installations onto a multipath device.

It is assumed that you are familiar with Dell EqualLogic arrays and have experience managing them. I do not cover the steps needed to create and configure storage LUNs and I do not cover adding your array to pools and groups. All of this information can be found in the official documentation at http://www.dell.com/storage.

Because I did not use the EqualLogic Host Integration Tools, you could potentially apply these instructions if you are instead using Dell PowerVault iSCSI arrays or software iSCSI targets hosted on a Dell PowerEdge server. However, I only used Dell EqualLogic storage.

1.3 Terms & Conventions

- LUN: iSCSI storage target on the storage array.
- Initiator: iSCSI client connecting to the iSCSI storage target (LUN).
- All commands are run as root. If you use a non-root account, prepend 'sudo' to each command.

2. Requirements

2.1 Hardware

- A Dell PowerEdge server with at least two network ports. For this whitepaper, I used different 11G and 12G PowerEdge servers,
- One LUN on a Dell EqualLogic array (I used a PS4100E array). LUN capacity does not matter but try with at least several hundred MBs. For simplicity purposes, I only used one array, which was a member of the default pool.

2.2 Network Topology

The ideal network configuration in a multipath environment is to connect each network port on your server to a different subnet. That way, you have additional resilience in case one of your subnets goes down (i.e. bad switch or router). So ideally, you would have something like Figure 1:

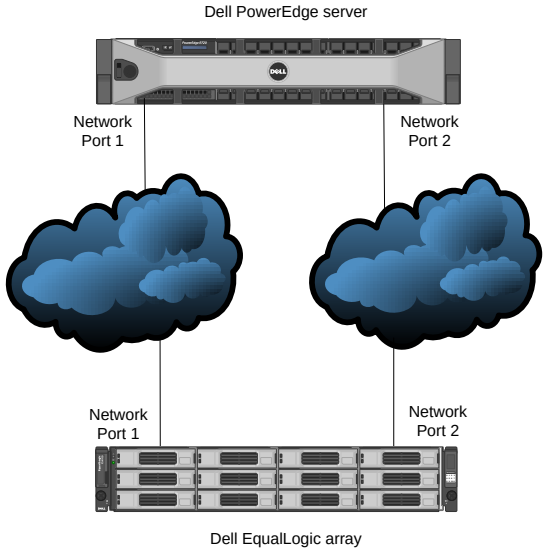

Figure 1: Multipath environment using one subnet per network port

However, you can also connect both of your network ports to the same subnet if that is all you have, as depicted in Figure 2. In this case, your network subnet becomes a single point of failure, but you still have high-availability capabilities in case one of your network ports fails. To increase resiliency in this scenario, connect each network port to a different switch in your subnet.

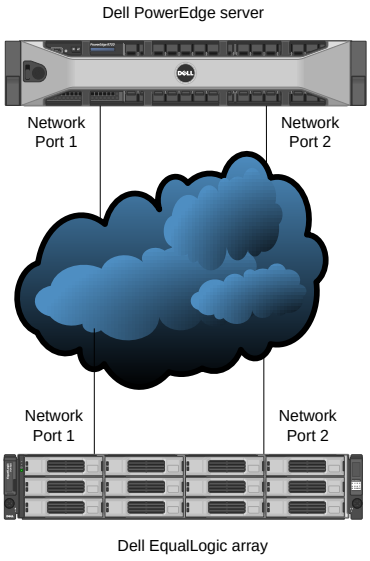

Figure 2: Multipath environment using one subnet for both network ports

2.3 Network Configuration

For simplicity purposes, I used the network topology shown in Figure 2 with only one subnet. I have a Class C network (192.168.1.0/24) and I used the following IP addresses:

Interface	Interface name	IP address
Server network port 1	eth0	192.168.1.2
Server network port 2	eth1	192.168.1.3
EqualLogic Group port		192.168.1.100
EqualLogic data port 1	eth0	192.168.1.101
EqualLogic data port 2	eth1	192.168.1.102

It's up to you whether you use static IP addresses vs. DHCP. In this example I used static addresses for demonstration purposes. Best practice documents don't appear to prefer one method over the other, but my recommendation is to use static IP addresses since that's what I used. If you use DHCP, be sure you set the LUN access permissions accordingly. So for example if you are restricting access by IP address, restrict to "192.168.1.*".

The Group IP address is used for administrative and host access to your LUNs. It is the IP address that we will use in this whitepaper to communicate with the storage array. For details on how to configure your array network ports, please refer to the official Dell EqualLogic documentation.

2.3.1 Reverse Path Filtering

You might have to edit the reverse-path filter settings if you have problems pinging the Group IP through both of your network interfaces:

```
# ping -I eth0 192.168.1.100
# ping -I eth1 192.168.1.100
```

If you get a response from both interfaces, then move on. Otherwise, you will have to change this setting in /etc/sysctl.conf:

```
net.ipv4.conf.eth0.rp_filter=2
net.ipv4.conf.eth1.rp_filter=2
```

If these entries are commented out, uncomment them and set to "2" as shown above (the default value is 1). Here's a short snippet on values for this parameter:

0: No source validation.
1: Strict mode as defined in RFC3704 Strict Reverse Path. Each incoming packet is tested against the forwarding table and if the interface is not the best reverse path the packet check will fail. By default failed packets are discarded.
2: Loose mode as defined in RFC3704 Loose Reverse Path. Each incoming packet's source address is also tested against the forwarding table and if the source address is not reachable via any interface the packet check will fail.

After making the change and saving the file, load the new settings:

```
# sysctl -p
```

3. Setting up iSCSI connections

Before we setup multipathing, we must first establish the iSCSI connection to the LUN. In order to walk you through the iSCSI configuration process, we will illustrate with an example using the same steps that I used in my lab.

1. Install required packages:

   ```
   # apt-get install open-iscsi
   ```

2. Assign a name to the initiator in */etc/iscsi/initiatorname.iscsi*. The 'iscsi-iname' tool can be used to generate a random initiator name that you can later add to the file, but the name already in the configuration file can also be used since it will already be unique. The file content will look something like this:

   ```
   InitiatorName=iqn.1993-08.org.debian:01:ecff9348ab3f
   ```

3. Edit parameters in */etc/iscsi/iscsid.conf*:

   Edit node startup from 'manual' to 'automatic' so that logins to the iSCSI LUNs are automatic after a system reboot:

   ```
   node.startup = automatic
   ```

   If you configured CHAP authentication in your Dell EqualLogic array, uncomment and edit these parameters:

   ```
   node.session.auth.authmethod = CHAP
   node.session.auth.username = <chap-user>
   node.session.auth.password = <chap-password>
   ```

   Restart open-iscsi service so that new values take effect:

   ```
   # service open-iscsi restart
   ```

4. Create iSCSI interfaces. To create the multiple logins for multipathing to work, we need to create an interface file for each network interface you wish to use to connect to the array.

   ```
   # iscsiadm -m iface -I eth0 -o new
   # iscsiadm -m iface -I eth1 -o new
   ```

   Add interface name to each network port:

   ```
   # iscsiadm -m iface -I eth0 --op=update -n iface.net_ifacename -v eth0
   # iscsiadm -m iface -I eth1 --op=update -n iface.net_ifacename -v eth1
   ```

   Verify settings:

```
# iscsiadm -m iface -I eth0

# BEGIN RECORD 2.0-871
iface.iscsi_ifacename = eth0
iface.net_ifacename = eth0
iface.ipaddress = <empty>
iface.hwaddress = <empty>
iface.transport_name = tcp
iface.initiatorname = <empty>
# END RECORD

# iscsiadm -m iface -I eth1

# BEGIN RECORD 2.0-871
iface.iscsi_ifacename = eth1
iface.net_ifacename = eth1
iface.ipaddress = <empty>
iface.hwaddress = <empty>
iface.transport_name = tcp
iface.initiatorname = <empty>
# END RECORD.
```

You can leave the initiator, IP address and HW address fields empty.

5. Discover LUN using the Group IP address:

```
# iscsiadm -m discovery -t st -p 192.168.1.100

192.168.1.100:3260,1 iqn.2001-05.com.equallogic:8-cb2b76-05afa6b6d-
7f2f3e9dc2d52096-dm-ubuntu
192.168.1.100:3260,1 iqn.2001-05.com.equallogic:8-cb2b76-05afa6b6d-
7f2f3e9dc2d52096-dm-ubuntu
```

This will give you the target ID for your LUN (will be different for you). Since you have two connections to the iSCSI array, you should see the LUN listed twice.

6. Login to LUN (replace target ID below with your own!):

```
# iscsiadm -m node -T iqn.2001-05.com.equallogic:8-cb2b76-05afa6b6d-
7f2f3e9dc2d52096-dm-ubuntu --login

Logging in to [iface: eth0, target: iqn.2001-05.com.equallogic:8-cb2b76-05afa6b6d-
7f2f3e9dc2d52096-dm-ubuntu, portal: 192.168.1.100,3260]
Logging in to [iface: eth1, target: iqn.2001-05.com.equallogic:8-cb2b76-05afa6b6d-
7f2f3e9dc2d52096-dm-ubuntu, portal: 192.168.1.100,3260]
Login to [iface: eth0, target: iqn.2001-05.com.equallogic:8-cb2b76-05afa6b6d-
7f2f3e9dc2d52096-dm-ubuntu, portal: 192.168.1.100,3260]: successful
Login to [iface: eth1, target: iqn.2001-05.com.equallogic:8-cb2b76-05afa6b6d-
7f2f3e9dc2d52096-dm-ubuntu, portal: 192.168.1.100,3260]: successful
```

If you list your storage devices with 'fdisk –l", you will see two additional storage devices. It's the same device, but it's listed twice (one for each network port).

7. Verify the iSCSI connections:

```
# iscsiadm -m session -P 1
```

```
Target: iqn.2001-05.com.equallogic:8-cb2b76-05afa6b6d-7f2f3e9dc2d52096-dm-ubuntu
     Current Portal: 192.168.1.101:3260,1
     Persistent Portal: 192.168.1.100:3260,1
             **********
             Interface:
             **********
             Iface Name: eth0
             Iface Transport: tcp
             Iface Initiatorname: iqn.1993-08.org.debian:01:fce6d6dfed6e
             Iface IPaddress: 192.168.1.2
             Iface HWaddress: (null)
             Iface Netdev: eth0
             SID: 1
             iSCSI Connection State: LOGGED IN
             iSCSI Session State: LOGGED_IN
             Internal iscsid Session State: NO CHANGE
     Current Portal: 192.168.1.102:3260,1
     Persistent Portal: 192.168.1.100:3260,1
             **********
             Interface:
             **********
             Iface Name: eth1
             Iface Transport: tcp
             Iface Initiatorname: iqn.1993-08.org.debian:01:fce6d6dfed6e
             Iface IPaddress: 192.168.1.3
             Iface HWaddress: (null)
             Iface Netdev: eth1
             SID: 2
             iSCSI Connection State: LOGGED IN
             iSCSI Session State: LOGGED_IN
             Internal iscsid Session State: NO CHANGE
```

Notice the two IP addresses in **blue** above. These are the IP addresses for the two network data interfaces in my Dell EqualLogic array.

8. Issue with making LUN login persistent

   I noticed that the initiator wasn't logging in to the LUN automatically after restarting the open-iscsi service or rebooting the server, despite having changed the 'node.startup' parameter in */etc/iscsi/iscsid.conf* to 'automatic'. I looked around and came across this bug:

   https://bugs.launchpad.net/ubuntu/+source/open-iscsi/+bug/1001535

   After looking at the startup script */etc/init.d/open-iscsi*, I applied the patch mentioned there and that fixed it (you can download the patch I used here). When I look in */etc/iscsi/nodes/*/*/*, I see the two iSCSI interfaces I created in step 4, but I don't see a 'default' interface:

```
# ls -l /etc/iscsi/nodes/*/*
total 8
-rw------- 1 root root 1671 Aug 16 10:16 eth0
-rw------- 1 root root 1671 Aug 16 10:16 eth1
```

   Verify auto login:

```
# service open-iscsi restart
* Disconnecting iSCSI targets                                          [ OK ]
```

```
* Stopping iSCSI initiator service                                    [ OK ]
* Starting iSCSI initiator service iscsid                             [ OK ]
* Setting up iSCSI targets                                            [ OK ]

# iscsiadm –m session
tcp: [1] 192.168.1.100:3260,1 iqn.2001-05.com.equallogic:8-cb2b76-05afa6b6d-
7f2f3e9dc2d52096-dm-ubuntu
tcp: [2] 192.168.1.100:3260,1 iqn.2001-05.com.equallogic:8-cb2b76-05afa6b6d-
7f2f3e9dc2d52096-dm-ubuntu
```

We are currently working with Canonical to get clarification on this issue and what the proper resolution should be, but the applied patch appears to work well with Ubuntu Server 12.04.

4. Setting up Multipath

Now that the iSCSI connections are established through both network ports, we are ready to proceed with setting up multipath.

1. Install required package

```
# apt-get install multipath-tools
```

2. Set up configuration file */etc/multipath.conf*. There are a myriad of configuration options. The Ubuntu documentation indicates you can leave it empty and use the defaults. Below are the options that worked for me, along with a brief comment for each one, Most default options will probably be ok, but it is highly recommend you read the documentation (man page for multipath.conf) to decide which options are right for you:

```
defaults {
       user_friendly_names     yes
       # Use 'mpathn' names for multipath devices
       path_grouping_policy    multibus
       # Place all paths in one priority group
       path_checker            readsector0
       # Method to determine the state of a path
       polling_interval        3
       # How often (in seconds) to poll state of paths
       path_selector           "round-robin 0"
       # Algorithm to determine what path to use for next I/O operation
       failback                immediate
       # Failback to highest priority path group with active paths
       features               "0"
       no_path_retry           1
       # These two options go hand-in-hand. The documentation states that
       # the only value available for 'feature' is '1 queue_if_no_path'
       # which is the same as setting no_path_retry to 'queue'. However
       # after trying different values for both, this is what worked for me.
       # Refer to the multipath.conf man page for details.
       }

blacklist {
     # devnode "^sd[a]$"
     # I highly recommend you blacklist by wwid instead of device name
```

```
      wwid                360024e80551ed500160e317e08963b8b
}

multipaths {
      multipath {
              wwid              368b7b2dcb6a6af059620d5c29d3e2f7f
              # alias here can be anything descriptive for your LUN
              alias             mylun
      }
}
```

To get the WWID of a storage device to either blacklist (i.e. local drive) or to specify an alias for an iSCSI LUN, use the following:

```
# /lib/udev/scsi_id --whitelisted --device=/dev/sdX
```

Where X is your storage device name; for example sda is usually a local drive, whereas the iSCSI LUN might be sdb and sdc.

3. Detect the paths to your iSCSI LUN. If the following command doesn't print anything after running, verify your iSCSI connections:

```
# multipath -v2
mylun (368b7b2dcb6a6af059620d5c29d3e2f7f) dm-2 EQLOGIC,100E-00
size=60G features='0' hwhandler='0' wp=undef
`-+- policy='round-robin 0' prio=1 status=undef
  |- 3:0:0:0 sdb 8:16 undef ready running
  `- 4:0:0:0 sdc 8:32 undef ready running
```

Display multipath topology:

```
# multipath -ll
mylun (368b7b2dcb6a6af059620d5c29d3e2f7f) dm-2 EQLOGIC,100E-00
size=60G features='1 queue_if_no_path' hwhandler='0' wp=rw
`-+- policy='round-robin 0' prio=1 status=active
  |- 7:0:0:0 sdb 8:16 active ready running
  `- 8:0:0:0 sdc 8:32 active ready running
```

This command will print the name of your multipath device(s), something like mpath0 or in this case, the alias 'mylun' defined in */etc/multipath.conf*. As you can see, there are 2 paths going to the iSCSI LUN, both of which are active and running.

4. Tweak iSCSI failover properties to your liking. The only value I changed (by choice) was the timeout value for when the operating system will fail over to the other path after one of the two paths fails. The default value is 120 seconds, which I changed to 10 seconds. I changed it using the iscsiadm command, as setting it in */etc/iscsi/iscsid.conf* somehow did not take effect.

Display current value (you see the value listed twice since you are connected to the LUN from two paths, but remember it's the same LUN):

```
# iscsiadm -m node -T iqn.2001-05.com.equallogic:8-cb2b76-05afa6b6d-
7f2f3e9dc2d52096-dm-ubuntu | grep node.session.timeo.replacement_timeout
node.session.timeo.replacement_timeout = 120
```

```
node.session.timeo.replacement_timeout = 120
```

Update value:

```
# iscsiadm -m node -T iqn.2001-05.com.equallogic:8-cb2b76-05afa6b6d-
7f2f3e9dc2d52096-dm-ubuntu -o update -n
node.session.timeo.replacement_timeout -v 10
```

Display new value:

```
# iscsiadm -m node -T iqn.2001-05.com.equallogic:8-cb2b76-05afa6b6d-
7f2f3e9dc2d52096-dm-ubuntu | grep node.session.timeo.replacement_timeout
node.session.timeo.replacement_timeout = 10
node.session.timeo.replacement_timeout = 10
```

5. The iSCSI multipath setup is complete. If you mount a file system on the iSCSI LUN, **don't forget to pass the _netdev option when you mount it**. This parameter tells the operating system that this is a network file system, so that on a reboot it unmounts it before the network services are stopped. Otherwise, your system will hang on a reboot.

   To add an entry in /etc/fstab, you would use something like:

   ```
   /dev/mapper/mylun   /share     ext4   _netdev   0 0
   ```

5. Testing path failover

Before you are ready to deploy your new iSCSI multipath environment, it's a good idea to do some sanity checking by simulating a network failure and verifying that the path to the iSCSI LUN is not interrupted.

1. We create an ext4 file system on the iSCSI LUN, mount it on directory /share and then do some disk I/O on it by using a script to continually create dummy files in that directory:

   ```
   # mkfs.ext4 /dev/mapper/mylun
   # mkdir /share
   # mount –o _netdev /dev/mapper/mylun /share
   ```

   Here's the script I used:

   ```
   #!/bin/bash
   interval=1
   while true; do
      ts=`date "+%Y.%m.%d-%H:%M:%S"`
      echo $ts > /share/file-${ts}
      echo "/share/file-${ts}...waiting $interval second(s)"
      sleep $interval
   done
   ```

   The script creates a dummy file every second and echoes a message to the console. The file name and the echoed message include a time stamp down to the second, so you can easily verify if there was an interruption writing to the LUN.

2. Run the script and then simulate a network port failure by unplugging the cable from either one of the network ports. After 5 seconds, you will see an error message on the screen (and in */var/log/syslog*) from the iSCSI subsystem that one of the iSCSI connections has failed. This is normal and it makes sense since indeed one of the connections has been severed. However, the multipath connection should remain active and the script should continue writing to the iSCSI LUN. Verify multipath status:

```
# multipath -ll
mylun (368b7b2dcb6a6af059620d5c29d3e2f7f) dm-2 EQLOGIC,100E-00
size=60G features='1 queue_if_no_path' hwhandler='0' wp=rw
`-+- policy='round-robin 0' prio=1 status=active
  |- 4:0:0:0 sdc 8:32 active ready  running
  `- 3:0:0:0 sdb 8:16 failed faulty running
```

   Is the script still running? Are files still being written to /share? The answer should be 'Yes' for both.

3. Plug the network cable back in, wait a few seconds and then unplug the other network cable. Keep looking at the console for messages. You will again see the message from the iSCSI subsystem about a failed connection, but again, this is normal. Check the multipath status:

```
# multipath -ll
mylun (368b7b2dcb6a6af059620d5c29d3e2f7f) dm-2 EQLOGIC,100E-00
size=60G features='1 queue_if_no_path' hwhandler='0' wp=rw
`-+- policy='round-robin 0' prio=1 status=active
  |- 4:0:0:0 sdc 8:32 failed faulty running
  `- 3:0:0:0 sdb 8:16 active ready  running
```

   Again, one of the two paths has failed (although now is the other one). However, the path to the LUN remains active.

4. Connect the network cable back in and verify the multipath device. Everything should be back to normal.

```
# multipath -ll
mylun (368b7b2dcb6a6af059620d5c29d3e2f7f) dm-2 EQLOGIC,100E-00
size=60G features='1 queue_if_no_path' hwhandler='0' wp=rw
`-+- policy='round-robin 0' prio=1 status=active
  |- 4:0:0:0 sdc 8:32 active ready running
  `- 3:0:0:0 sdb 8:16 active ready running
```

6. Conclusion

Setting up iSCSI Multipath with Dell EqualLogic storage is not drastically different than other enterprise Linux operating systems. There were a couple of workarounds, but as you can see most implementation steps are the same.

Please note that the multipath configuration settings in */etc/multipath.conf* and the ISCSI timeout values I used here were sufficient for me and will most likely be sufficient for you too, but be sure you fully understand all options available before you deploy your environment.