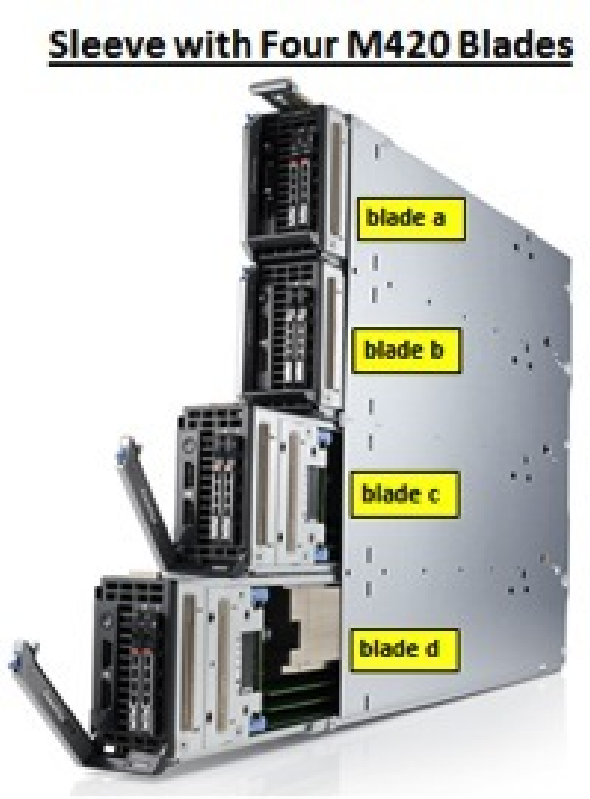
Deploy a PowerEdge M420 Cluster in a M1000e Chassis with a single Force10 MXL Switch

By Ishan Singh, Calvin Jacob

In an M1000e chassis populated with 32 M420 blade servers and a single Force10 10GbE MXL switch, 16 blades have connectivity to the MXL switch through NIC1 and the other 16 blades have connectivity through NIC2. For servers that are configured with the HPC SKU, the first device in boot sequence is set to NIC1 in PXE mode. By default, only 16 blades can PXE boot when using a single Force10 MXL Switch. This currently does not support the remaining 16 blades to be added to the cluster as they will have no connectivity on NIC1.

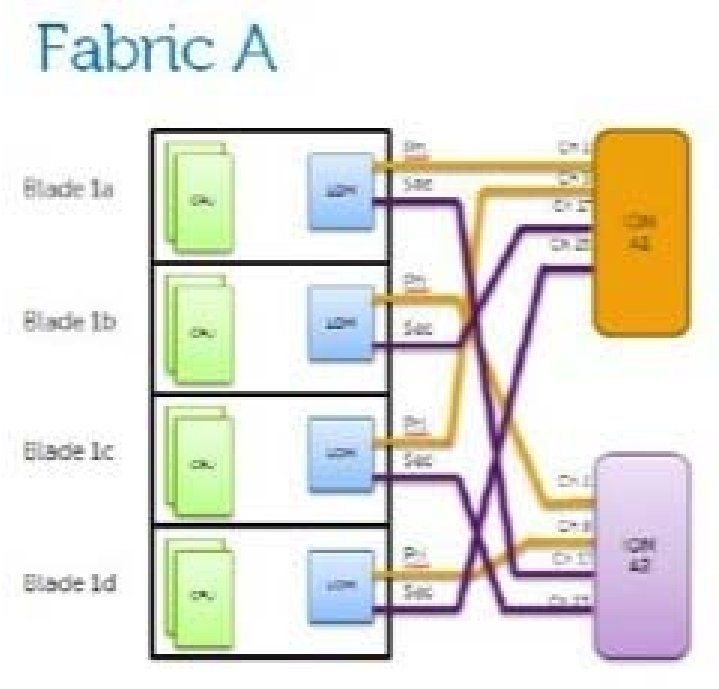
Some additional configuration is clearly needed and that's where this blog comes in!This blog describes the steps with the set of sample scripts that can be used to deploy a 32 node PowerEdge M420 cluster using a single Force10 MXL switch.

For solutions that use two MXL switches as IOM1 and IOM2 in A1 and A2 slots of the chassis, there is no additional configuration needed. All blades use NIC1 to PXE boot through the switches. Blades in slot "a" and "c" use the switch in Slot A1 and blades in slot "b" and "d" use switch in Slot A2 to PXE boot. In this configuration the servers have their NIC1 enabled with PXE, and use both the switches to run as a cluster.



The Force10 MXL switch has 32 internal ports which has a capacity to map to all 32 M420 blades within a chassis. However, as seen in the figure below, all blades in “a” and “c” slots have NIC1 mapped to the internal ports of IOM A1 and, all blades in “b” and “d” slots have NIC2 mapped to the internal ports of IOM A1.We want to provision blades in slots “b” and “d” via NIC2 and use only 1 MXL switch in Slot A1 to PXE boot all the servers, thus providing a cost efficient bundled solution.

The Force10 MXL IOM in Fabric A of the chassis would be mapped to the M420 servers as shown in the figure below:



For implementing the proposed solution, all the blades in slots “b” and “d” should have NIC2 enabled with PXE and have NIC2 first in the BIOS boot order. Blades in slots “a” and “c” need no configuration change and remain with NIC1 enabled with PXE and first in the BIOS boot sequence.

Remote system management commands are used to change the BIOS settings. These commands can be run from the cluster’s head node to make the above specified changes in the BIOS .

We use the `racadm` commands for these changes.`racadm` is a command line utility provided for **Dell OpenManage Server Administrator** .

`racadm` commands use the iDRAC IP of the blade servers to make the changes. So, in order to allocate the static iDRAC IP’s to all the servers and then make the desired BIOS changes, the following steps are needed. All the scripts referenced in the steps below are attached with this blog.

1. Install the head node. This is the master node of the cluster and will be used to deploy all the blade servers as compute nodes.

2. Create a multihomed interface eth0:0 with an IP 192.168.0.120/255.255.255.0 by executing the following script.

```
chmod +x network_config_mxl.sh
```

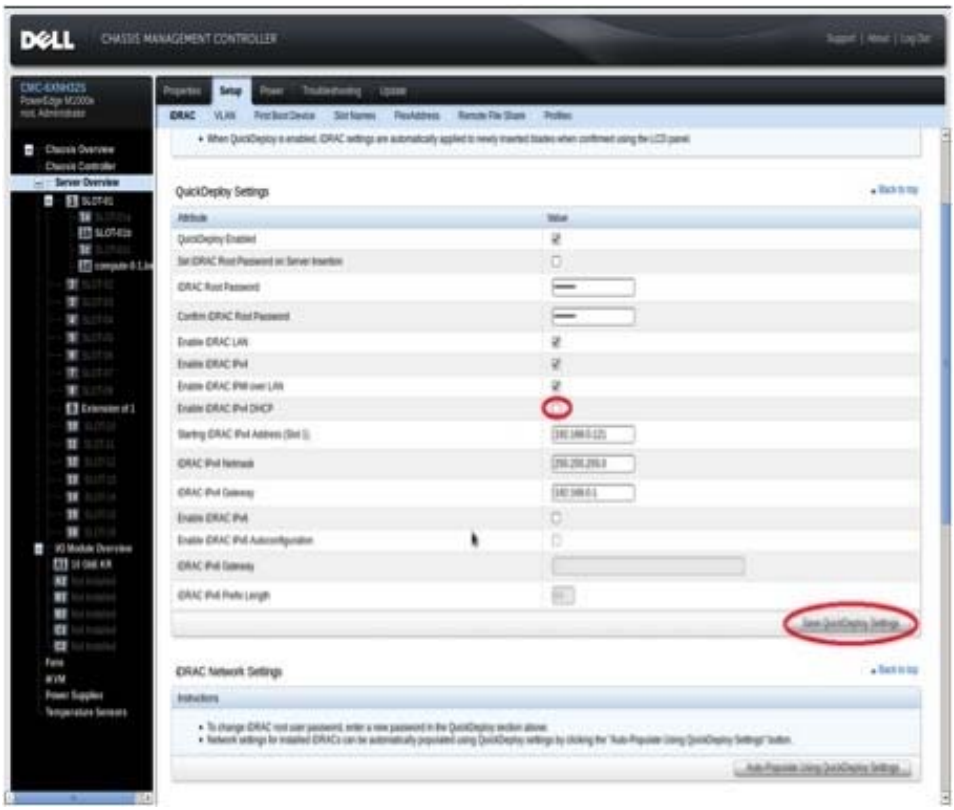
```
./network_config_mxl.sh
```

3. Power on the M1000e chassis and insert all the M420 blades in the chassis. Login to the CMC of the chassis and disable DHCP on the CMC by executing the following command on the CMC's CLI:

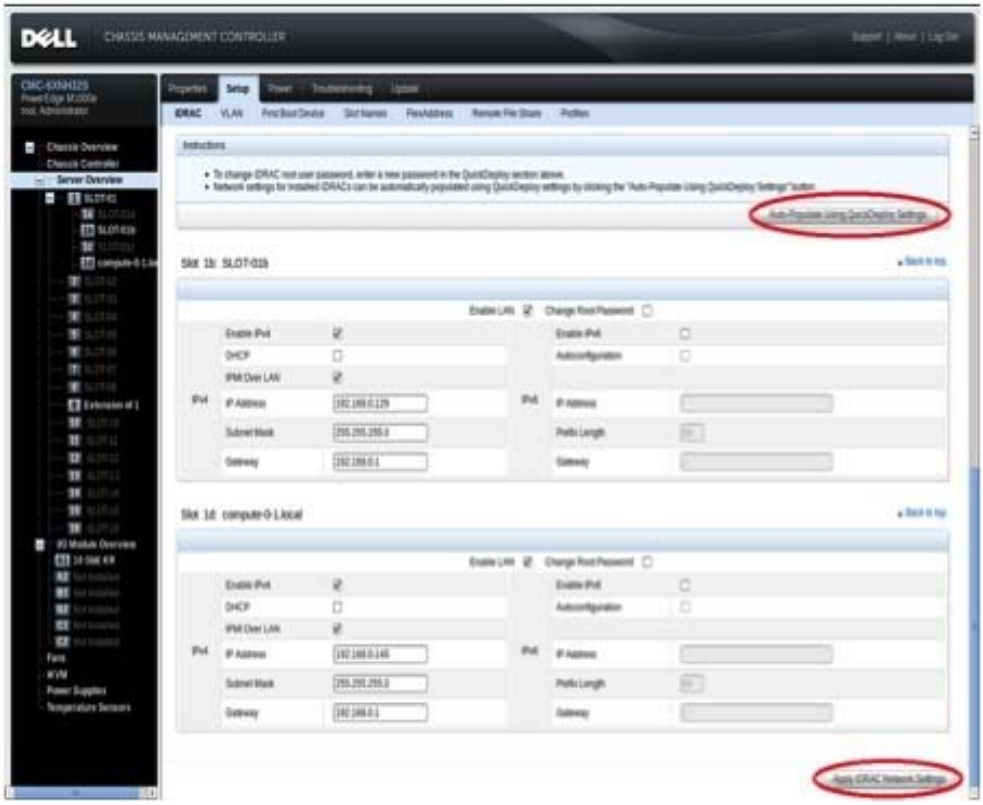
```
racadm config -g cfgLanNetworking -o cfgNicUseDHCP 0
```

4. Connect the MXL switch, the CMC of chassis and eth0 of the head node to the same network, i.e. the cluster's private network switch. From the head node, login to the CMC using a browser session. This can be done by using CMC's default Static IP of 192.168.0.120 as the URL.

In the CMC's GUI click on **Server Overview**, then click on the **Setup** tab and configure the QuickDeploy Settings by **disabling** the "**Enable iDRAC IPv4 DHCP**" option and then saving it by clicking at "**Save QuickDeploy Settings**".



5. Click on "**Auto-Populate Using QuickDeploy Settings**" and finally clicking on "**Apply iDRAC Network Settings**".



While applying the changes,make sure that all the servers in the chassis are listed in the CMC GUI.

Steps 4 and 5 will allocate Static iDRAC IP address to the blades starting from 192.168.0.121 for the blade in Slot 1a to 192.168.0.152 for blade in Slot 8d. That is a total of 32 blades.

6. After the IP addresses are allocated, install the remote `racadm` utility on the head node using the **Dell OpenManage Server Administrator** tar file or DVD.

To get the latest version of OMSA, please visit **Dell OpenManage Download Page** .

a) Mount the OMSA dvd to a directory(say to a directory /tmp/OM).

b) Go to `/tmp/OM/SYSMGMT/srvadmin/linux/supportscripts`

c) Execute the following script

```
./srvadmin-install.sh
```

d) Select Options 5 and 6 to install `racadm` on your system.

e) Press ‘y’ to start the Server Administrator services.

7. Now run **master_script_mxl.sh** to change the BIOS settings on the M420 servers in slot b and d and wait for the script to finish the execution.

```
chmod +x master_script_mxl.sh
```

```
./master_script_mxl.sh
```

This script will enable NIC2 with PXE, disable NIC1 with PXE and set NIC2 ahead of the hard drive in the server’s BIOS boot sequence.

At this point all changes needed to install the cluster with one MXL switch are complete!

8. Now start the dhcp server on your FE and execute the following script to add the M420 blades as compute nodes sequentially to the dhcp server running on your FE.

```
./mxl_powerup.sh
```

With this script,the M420 blades will automatically start powering on sequentially starting from the blades in slot a,then in slot b,slot c and finally in slot d, and will start getting listed as compute nodes.

Deploying larger than 32 node M420 cluster

- To apply this solution to a M420 cluster that is larger than 32 nodes using multiple M1000e chassis, the above mentioned solution can be used one chassis at a time.
- At a time only a single chassis CMC should be connected to the switch to which FE is connected. This is because all chassis have the same default CMC IP of 192.168.0.120
- As we are using the same static IP for every chassis, remove the cable connecting CMC of the first chassis to the private network switch to which head node is connected before connecting the CMC of the next chassis to that switch.
- Steps 1,2 and 6 are not required to be repeated on the other chassis after executing them on the first chassis**.**
- Let the servers in the first chassis get deployed as a cluster and turn off the DHCP server before moving on to the next chassis.

These scripts were tested in the lab on the following setup:

PowerEdge M420 iDRAC version: 1.30.30.

PowerEdge M420 BIOS version: 1.2

PowerEdge M1000e CMC firmware version: 4.30

Dell OpenManage Server Administrator version: 7.2

M1000e Chassis Midplane Version = 1.1

- All ports on the MXL switch were configured as “switchport” with default VLAN id 1.

Please refer to the attachments of this blog for scripts used in this solution.

References:

http://i.dell.com/sites/doccontent/shared-content/data-sheets/en/Documents/PowerEdge_M_Series_Blades_IO_Guide.pdf