Hewlett Packard Enterprise

HPE Cray EX Series System Administration with HPE Performance Cluster Manager

Lab guide add nodes

Confidential - For Training Purposes Only

Contents

Add nodes to cluster

Computer models handle UEFI, BIOS, and UEFI combined with Legacy BIOS differently due to product design goals and product and technology history.

In this lab exercise, you will discover nodes, and in this environment the HPE Proliant DL models sometimes require a PXE boot instruction and reboot after the discover command configures the iLO device in the cluster database.

Add the nodes

After you configure cluster attributes with the configure-cluster and cadmin commands and update the cluster with the latest patches, the next step is to discover the nodes that will participate in the cluster--identify and provision cluster nodes.

For this exercise, work with your LabGroup.

1. Log in to the admin node.
2. Show the available images:

   cm image show

3. Set the image for the node:

   cm node set -i sles15sp3 -n <node>

4. Make a directory for your work (replace <my-code> with your initials or a code that uniquely identifies you);

   mkdir /class/<my-code>

5. Collect discover show output (the command wraps to a second line; do not type the \ character; replace **<date>** with the date):

```
discover --show-configfile --images --kernel --bmc-info \
--kernel-parameters --ips > /class/<my-code>/discover-show-<date>.txt
```

> NOTICE: On a production cluster, periodically copy this output to a non-cluster node.

The following form of the command uses the date command to embed the date in the file name.

```
discover --show-configfile --images --kernel --bmc-info \
--kernel-parameters --ips > /class/<my-code>/discover-show-$(date +%F%H%M).txt
```

6. Review the archive that you created (replace **<date>** with the date that you specified in the ls discover --show-configfile command above).

   less /class/<my-code>/discover-show-<date>.txt

7. Quit the less command.

   q

8. Create a configuration file for your labgroup node output (the command wraps to a second line; do not type the \ character; replace <date> with the date; replace <node> with the name of your LabGroup node):

   ```
   discover --show-configfile --images --kernel --bmc-info \
   --kernel-parameters --ips | grep <node> > /class/<my-code>/cfg-<node>
   ```

9. Insert a new first line [discover] to the file /class/<my-code>/cfg-<node>.

   The file will contain 2 lines similar to:

   [discover]

   hostname1=x3019c0s31b0n0, internal_name=service31, mgmt_bmc_net_name=head-bmc, mgmt_bmc_net_macs="b4:7a:f1:48:88:24", mgmt_bmc_net_ip=172.24.1.31, mgmt_net_name=head, mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=active-backup, mgmt_net_macs="d4:f5:ef:3a:89:6c,d4:f5:ef:3a:89:6d", mgmt_net_interfaces="eno5,eno6", mgmt_net_ip=172.23.1.31, data1_net_name=hsn0, data1_net_interfaces="ens1f0", data1_net_ip=10.150.0.12, ib_0_ip=10.148.0.31, ib_1_ip=10.149.0.31, rootfs=disk, transport=udpcast, conserver_logging=yes, conserver_ondemand=no, dhcp_bootfile=ipxe-direct, disk_bootloader=no, predictable_net_names=yes, redundant_mgmt_network=yes, switch_mgmt_network=yes, tpm_boot=no, console_device=ttyS0, architecture=x86_64, card_type="iLO", baud_rate=115200, bmc_username=xx, bmc_password=xxx

10. Inspect cfg-node and ensure that the BMC/ILO credentials and other elements are correct.
11. Use cadmin to show and set cluster configuration values to omit the switch configuration steps during future discover commands—this would only be run once on the cluster—if the value is no, set it to yes:

    ```
    cadmin --show-discover-skip-switchconfig

    cadmin --enable-discover-skip-switchconfig

    cadmin --show-discover-skip-switchconfig
    ```

    By default, the discover command performs top-level switch configuration operations each time it runs. The cadmin --enable-discover-skip-switchconfig command directs the discover command to omit the switch configuration. When you add nodes on a system that is configured, skipping the switch configuration process saves time.

12. Work with your LabGroup for the remainder of this exercise:
13. Power off your LabGroup node (replace <node> with the name of your LabGroup node).

    ```
    cm power off -t node <node>
    ```

14. Wait a minute for the power off to take effect.
15. Confirm that the node is powered off.

    ```
    cm power status -t system
    ```

16. Delete your labgroup node from the cluster (replace <node> with the name of your LabGroup node):

    ```
    cm node delete -n <node>
    ```

17. Confirm that your labgroup node is no longer available (replace <node> with the name of your LabGroup node):

    ```
    cm node show -n <node>

    discover --show-configfile | grep <node>
    ```

18. Add your node to the cluster (use one of your LabGroup cfg-files, replace <node> with the name of your LabGroup node):

    cm node add -c /class/<my-code>/cfg-<node>

19. Provision the node:

    cm node provision -i sles15sp3 -n <node>

20. Monitoring node provisioning.

21. Confirm that your node is operational.

Troubleshooting tips

- Check that the node PXE booted.
- Is the node under the management of an SU leader?
- Change node to rootfs tmpfs; does it boot?
- Handle disk label problems using directions in `/tmp/si.log` or `si_monitor.log`.
- To install on a drive, overwrite the partition table with sgdisk --zap-all /dev/sd*X*
- Ping the node BMC.
- The imaging script is running from a memory resident mini-root that is `chroot'ed` boot order.
- If you get a `no sgi filesystem labels found` message, the installer could not determine which device is safe to install. From an install prompt, determine which drive is safe.
- You can use the --force-disk DEV option of cm node provision to specify the correct drive for installation.

Add Node as Hardware Type Other

Two hardware type options support nodes in the cluster network that are not managed by HPE Performance Cluster Manager administration tools. The "other" type option reserves cluster IP addresses that you can manually configure for an unmanaged computer node. The "generic" type reserves a single, cluster DHCP provided IP address, for a node that broadcasts for an IP address on boot but does not require any other cluster administration operations.

1. On the admin node, change to your working directory.

   cd /class/<my-code>

2. In your working directory, create the cfg-other file with contents below command—change ## to a number between 50-99 (2 places; the second wraps to a second line—ensure that all the attributes are on one line and do not include the \ character):

   ```
   [discover]
   internal_name=service##, mgmt_net_name=head, hostname1=othernode##, \
   discover_skip_switchconfig=yes, other
   ```

3. Confirm the contents of the file:

   cat /class/<my-code>/cfg-other

4. Add the hardware type other node to the cluster:

   cm node add -c cfg-other

Example output:

```
[root@admin1 ~]# cm node add -c cfg-other
Config file: cfg-other
Add - All nodes in the cfg-other will be added to the database.
admin1: fastdiscover: Config file parse step: , 0.08s
admin1: fastdiscover: Node othernode1 Management Network defined but
missing management network interface. Defaulting to eth0,eth1.
admin1: fastdiscover: new nodes step: , 0.15s
admin1: fastdiscover: Script time: , 0.24s

Refreshing the netboot environment for nodes in the config file...

Updating admin node configs...
Configuration manager initiating node configuration.
1 of 1 nodes completed in 2.0 seconds, averaging 1.1s per node
Node configuration complete.

Performing switch configuration...

Please view '/var/log/switchconfig.log' to verify no switch configuration
error occurred during this process.
```

5. Confirm that the new "other" type node does not appear in cm node show output.

   cm node show

   Example output:

```
[root@admin1 ~]# cm node show
n12
n13
n14
n15
n16
n17
n18
```

6. Print out the database entry for the "other" type node—substitute the unique number you used earlier for the ## characters:

   discover --show-configfile | grep service##

   Example output:

```
[root@admin1 ~]# discover --show-configfile | grep service53
hostname1=othernode1, internal_name=service##, mgmt_net_name=head,
mgmt_net_bonding_master=bond0, mgmt_net_bonding_mode=active-backup,
mgmt_net_interfaces="eth0,eth1", mgmt_net_interface_name="othernode1",
rootfs=disk, transport=bt, conserver_logging=yes, conserver_ondemand=no,
dhcp_bootfile=ipxe-direct, disk_bootloader=no, predictable_net_names=yes,
redundant_mgmt_network=yes, switch_mgmt_network=yes, tpm_boot=no,
console_device=ttyS1, architecture=x86_64, card_type="IPMI", other
```

7. Delete the "other" type node::

   cm node delete -c /class/<my-code>/cfg-other

   Example output:

   ```
   [root@admin1 ~]# cm node delete -c /class/<my-code>/cfg-other

   Checking power status of the nodes...Config file: /class/<my-code>/cfg-
   other
   Delete - Nodes in /class/<my-code>/cfg-other that exist will be deleted
   instead of added.
   admin1: fastdiscover: Config file parse step: , 0.08s
   After skipping nodes that don't exist, actually removing: 1 nodes.
   admin1: fastdiscover: new nodes step: , 0.04s
   admin1: fastdiscover: Script time: , 0.12s
   Configuration manager initiating node configuration.
   1 of 8 nodes completed in 7.0 seconds, averaging 1.1s per node
   8 of 8 nodes completed in 7.4 seconds, averaging 5.9s per node
   Node configuration complete.
   ```

8. Confirm that the node has been deleted—substitute the unique number you used earlier for the ## characters:

   discover --show-configfile | grep service##

Check management switch configuration

1. Review the switchconfig man page:

   man switchconfig

2. On the admin node, display all management switch IP routing tables.

   switchconfig info -s all -r

3. Display L2 FDB (mac-address-table) table information:

   switchconfig info -s mgmtsw0 -f

4. Display expected firmware release, bonding, bonding VLAN, and bonding database sanity check information on all management switches:

   switchconfig sanity_check -s all

5. Review output from a different switch model (a few highlighted sections indicate an issue and the procedure to fix):

```
# switchconfig sanity_check -s all
=== Beginning Sanity Check on mgmtsw0 ===

checking switch firmware on mgmtsw0 ...

        mgmtsw0 slot 1 (5510 24G 4SFP+ HI 1-slot Switch) has firmware '7.1.070 Release
3506P02-US' installed (recommended: 7.1.070, Release 3506P08 or 7.1.070 Release 3506P08-
US)
        mgmtsw0 slot 2 (5510 24G 4SFP+ HI 1-slot Switch) has firmware '7.1.070 Release
3506P02-US' installed (recommended: 7.1.070, Release 3506P08 or 7.1.070 Release 3506P08-
US)
```

```
checking Bridge-Aggregation (bonding) configuration on mgmtsw0...

        interface Bridge-Aggregation 15 contains ports (1/0/15, 2/0/15) in bonding mode =
manual
        interface Bridge-Aggregation 24 contains ports (1/0/24, 2/0/24) in bonding mode =
802.3ad
        interface Bridge-Aggregation 122 contains ports (1/0/22) in bonding mode =
802.3ad

checking Bridge-Aggregation VLAN configuration on mgmtsw0...

        === General Information about VLAN Configurations ===

        Note: this output will only show VLAN subscription for Bridge-Aggregation
interfaces
        Default VLAN (native) means untagged packets are put into the VLAN
        Tagged VLAN means the port allows 802.1Q tagged packets to pass on the port
        Admin Node - Untagged VLANs: 1(default vlan), Tagged VLANs: 3
        Service Node - Untagged VLANs: 1(default vlan)
        Rack Leader Controller - Untagged VLANs: 1(default vlan), Tagged VLANs: <rack #
VLAN> (101, 103, etc.)
        HPE/SGI ICE 8600 CMCs - Untagged VLANs: <[rack # + 100] VLAN> (101, 102, etc.),
Tagged VLANs: 3
        HPE Apollo 9K CMCs - Untagged VLANs: (2001, 2002, etc.)
        Interswitch Links - Untagged VLANs: 1(default vlan), Tagged VLANs: 3, 1998(RIP),
1999(OSPF)

        interface Bridge-Aggregation 15 has the following VLAN subscriptions:

                Tagged VLANs:   3
                Untagged VLANs: 101

        interface Bridge-Aggregation 24 has the following VLAN subscriptions:

                Tagged VLANs:   101
                Untagged VLANs: 1(default vlan)

        interface Bridge-Aggregation 122 has the following VLAN subscriptions:

                Tagged VLANs:   101
                Untagged VLANs: 1(default vlan)


checking configured VLAN settings on mgmtsw0...

                 VLAN ID: 1
                 VLAN type: Static
                 Route interface: Configured
                 IPv4 address: 172.23.255.254
                 IPv4 subnet mask: 255.255.0.0
                 IPv6 global unicast addresses:
                    FD3E:58FB:6B27:1::AC17:FFFE, subnet is FD3E:58FB:6B27:1::/64
                 Description: configured_by_switchconfig_v1.5.0
                 Name: VLAN 0001
```

```
Tagged ports:   None
Untagged ports:
   Bridge-Aggregation24          Bridge-Aggregation122
   GigabitEthernet1/0/1          GigabitEthernet1/0/2
   GigabitEthernet1/0/3          GigabitEthernet1/0/4
   GigabitEthernet1/0/5          GigabitEthernet1/0/6
   GigabitEthernet1/0/7          GigabitEthernet1/0/8
   GigabitEthernet1/0/9          GigabitEthernet1/0/10
   GigabitEthernet1/0/11         GigabitEthernet1/0/12
   GigabitEthernet1/0/13         GigabitEthernet1/0/14
   GigabitEthernet1/0/16         GigabitEthernet1/0/17
   GigabitEthernet1/0/18         GigabitEthernet1/0/19
   GigabitEthernet1/0/20         GigabitEthernet1/0/21
   GigabitEthernet1/0/22         GigabitEthernet1/0/23
   GigabitEthernet1/0/24         GigabitEthernet2/0/1
   GigabitEthernet2/0/2          GigabitEthernet2/0/3
   GigabitEthernet2/0/4          GigabitEthernet2/0/5
   GigabitEthernet2/0/6          GigabitEthernet2/0/7
   GigabitEthernet2/0/8          GigabitEthernet2/0/9
   GigabitEthernet2/0/10         GigabitEthernet2/0/11
   GigabitEthernet2/0/12         GigabitEthernet2/0/13
   GigabitEthernet2/0/14         GigabitEthernet2/0/16
   GigabitEthernet2/0/17         GigabitEthernet2/0/18
   GigabitEthernet2/0/19         GigabitEthernet2/0/20
   GigabitEthernet2/0/21         GigabitEthernet2/0/22
   GigabitEthernet2/0/23         GigabitEthernet2/0/24
   Ten-GigabitEthernet1/0/27
   Ten-GigabitEthernet1/0/28
   Ten-GigabitEthernet2/0/27
   Ten-GigabitEthernet2/0/28

VLAN ID: 3
VLAN type: Static
Route interface: Configured
Description: VLAN 0003
Name: vlan0003
Tagged ports:
   Bridge-Aggregation15
   GigabitEthernet1/0/10         GigabitEthernet1/0/15
   GigabitEthernet1/0/16         GigabitEthernet1/0/20
   GigabitEthernet2/0/10         GigabitEthernet2/0/15
   GigabitEthernet2/0/16         GigabitEthernet2/0/20
Untagged ports: None

VLAN ID: 101
VLAN type: Static
Route interface: Configured
IPv4 address: 10.159.3.254
IPv4 subnet mask: 255.255.252.0
Description: VLAN 0101
Name: vlan0101
Tagged ports:
   Bridge-Aggregation24          Bridge-Aggregation122
   GigabitEthernet1/0/22         GigabitEthernet1/0/24
   GigabitEthernet2/0/24
```

```
                 Untagged ports:
                    Bridge-Aggregation15
                    GigabitEthernet1/0/15         GigabitEthernet2/0/15

                 VLAN ID: 1998
                 VLAN type: Static
                 Route interface: Configured
                 IPv4 address: 1.2.255.254
                 IPv4 subnet mask: 255.255.0.0
                 Description: VLAN 1998
                 Tagged ports:   None
                 Untagged ports: None
=== Neighboring Switch Cabling Information on mgmtsw0 ===

checking configured IRF(Stacking) settings on mgmtsw0...

=== IRF Link Information on mgmtsw0 ===

Running command - `display irf link`...

                Member 1
                 IRF Port  Interface                            Status
                 1         Ten-GigabitEthernet1/0/25            UP
                 2         Ten-GigabitEthernet1/0/26            UP
                Member 2
                 IRF Port  Interface                            Status
                 1         Ten-GigabitEthernet2/0/25            UP
                 2         Ten-GigabitEthernet2/0/26            UP

=== IRF General Information on mgmtsw0 ===

Running command - `display irf`...

                MemberID    Role    Priority  CPU-Mac         Description
                 *+1        Master  32        00e0-fc0f-8c02  ---
                   2        Standby 16        00e0-fc0f-8c03  ---
                ------------------------------------------------
                 * indicates the device is the master.
                 + indicates the device through which the user logs in.

                 The bridge MAC of the IRF is: ec9b-8b82-56b4
                 Auto upgrade               : yes
                 Mac persistent             : always
                 Domain ID                  : 0

=== IRF Configuration Information on mgmtsw0 ===

Running command - `display irf configuration`...

                 MemberID NewID    IRF-Port1                     IRF-Port2
                 1        1        Ten-GigabitEthernet1/0/25     Ten-
GigabitEthernet1/0/26
                 2        2        Ten-GigabitEthernet2/0/25     Ten-
GigabitEthernet2/0/26
```

```
checking configured/preferred fan direction settings on mgmtsw0...

=== Fan Direction Status Information on mgmtsw0 ===

Running command - `display fan`...

                Slot 1:
                Fan 1:
                State    : Normal
                Airflow Direction: Port-to-power
                Prefer Airflow Direction: Port-to-power
                Fan 2:
                State    : Normal
                Airflow Direction: Port-to-power
                Prefer Airflow Direction: Port-to-power

                Slot 2:
                Fan 1:
                State    : Normal
                Airflow Direction: Port-to-power
                Prefer Airflow Direction: Port-to-power
                Fan 2:
                State    : Normal
                Airflow Direction: Port-to-power
                Prefer Airflow Direction: Port-to-power


=== General Rules and Tips ===

- CMCs wired to stacked/IRF-enabled switches must be connected to the same port on both
switches. IE: 1/0/20+2/0/20, NOT 1/0/20+1/0/21
- When a node is configured for 'active-backup' bonding, the management switch doesn't
require bonding configuration
- Any nodes bonding must match the switch configured bonding (LACP with LACP, Static with
Static)
- Switch-to-Node bonding matches are as follows:
        Active-backup bonding requires no switch bonding, so `switchconfig set <other
details ommitted> --bonding none`
        802.3ad bonding requires LACP switch bonding, so `switchconfig set <other details
ommitted> --bonding lacp`
        Static/Manual bonding requires manual bonding, so `switchconfig set <other
details ommitted> --bonding manual`
- ICE XA CDU/CRC Cooling Hardware needs to be in native VLAN 3 and manually assigned in
most cases
        Example: 1 CDU and 2 CRC's are plugged into mgmtsw5's ports 1/0/33,1/0/34,1/0/35
        Solution: run the following switchconfig command:
                switchconfig set --switches mgmtsw5 --ports 1/0/33,1/0/34,1/0/35 --
default-vlan 3 --bonding none --redundant no

=== Troubleshooting ===
A compute/leader node is not booting or bonding mode is mismatched
        Example: A compute/leader connections = eno1<->1/0/1 & eno2<->2/0/1 on mgmtsw1
and will not PXE boot/DHCP/etc.
```

```
        Reason: Some power actions may cause bonding configuration to remain active after
a node reboots, bonding + PXE causes failed booting
        Solution: Run the following switchconfig commands:
        1.) Reset ports 1/0/1 & 2/0/1 back to factory settings
                switchconfig unset --switches mgmtsw1 --ports 1/0/1 --redundant yes
        2.) Reboot the problem node and wait for it to boot all the way
                `cm power reset -t leader <rXlead/leaderX>` or `cm power reset -t node
<node>`
        3.) Once booted, configure the switch for the problem node
                switchconfig_configure_node --node <rXlead/leaderX/node>

=== Database / Node Sanity Check Below ===
        Host admin w/mac-address 20:67:7c:ef:9a:de found on mgmtsw0 port(s) 2/0/10 - DB
Bonding = active-backup , Current Switch Bonding = active-backup
        Host admin w/mac-address 20:67:7c:ef:9a:de found on mgmtsw0 port(s) 1/0/10 - DB
Bonding = active-backup , Current Switch Bonding = active-backup
        PASS - Host admin passes switch sanity_check, DB bonding + VLAN config looks
correct

        Host leader1 w/mac-address 48:df:37:c6:d3:5c found on mgmtsw0 port(s) 1/0/13 - DB
Bonding = active-backup , Current Switch Bonding = active-backup
        PASS - Host leader1 passes switch sanity_check, DB bonding + VLAN config looks
correct

        Host leader2 w/mac-address 48:df:37:c4:84:18 found on mgmtsw0 port(s) 1/0/14 - DB
Bonding = active-backup , Current Switch Bonding = active-backup
        PASS - Host leader2 passes switch sanity_check, DB bonding + VLAN config looks
correct

        Host leader3 w/mac-address 48:df:37:c4:0c:40 found on mgmtsw0 port(s) 1/0/15 - DB
Bonding = active-backup , Current Switch Bonding = manual
        FAIL - WARNING!! - Host leader3 DB Bonding or VLAN config is not correct!!

        INFO -  Run the command directly below to fix host leader3 switch configuration
(w/--dry-run for test)
                switchconfig_configure_node --node leader3 [--dry-run]

=== Sanity Check Summary ===
NOTE: See above sections for a complete review of different components of the switch
configuration

Component                                   Result
------------------------------------------------------------
Firmware on Slot 1                          WARN - firmware on mgmtsw0 slot 1 does not match
one of the recommended firmware versions for this release
Firmware on Slot 2                          WARN - firmware on mgmtsw0 slot 2 does not match
one of the recommended firmware versions for this release
Bridge-Aggregation Interfaces(LACP)         INFO - mgmtsw0 has 2 LACP Bridge-
Aggregation(Bonded) Interfaces
Bridge-Aggregation Interfaces(Static)       INFO - mgmtsw0 has 1 Static Bridge-
Aggregation(Bonded) Interfaces
IRF(stacking) Status                        INFO - mgmtsw0 is an IRF-enabled switch with 2
slots
Fan Direction Status                        PASS - mgmtsw0 has the correct fan-direction
status
```

```
Node-to-Switch Configuration              ERROR - mgmtsw0 has at least 1 mismatched node-
to-switch configuration, see above detailed output for more details
```

This completes lab exercise add nodes.