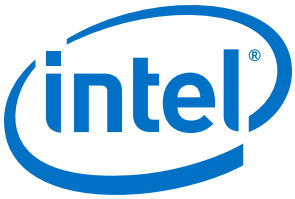

HPC Cluster Reference Design

Cluster installation based on Intel® HPC Orchestrator Advanced for systems with Intel® Xeon® Gold 6148 Processor and Intel® Omni-Path Fabric.

2017.08.03

Legal Notices

Contents

Chapter 1

This Reference Design

1.1 Summary

The Reference Design is a verified implementation example of a given Reference Architecture, complete with hardware and software Bill of Materials information and cluster configuration instructions. It can confidently be used “as is”, or be the foundation for enhancements and/or modifications.

1.2 Hardware Bill of Materials

Hardware Bill of Materials for the Head Node

Quantity	Item	Manufacturer	Model
1	Intel® Server Chassis	Intel	R2000WT
1	Intel® Server Board (w/10Gb Intel® Ethernet Controller)	Intel	S2600WFD
	(2x) Intel® Xeon® Gold Processor	Intel	6148
	(8x) 8GB ECC DDR4 2666MHz	Micron	MTA18ASF1G72PZ
	(8x) Intel® SSD 800GB, 2.5-inch SATA	Intel	S3700 Series
	(1x) Intel® Omni-Path Host Fabric Interface Adapter	Intel	100 Series 1 Port PCIe x16
1	10-Gigabit Smart Managed Ethernet Switch	Netgear	XS748T ProSAFE Switch
1	Intel® Omni-Path Edge Switch	Intel	100 Series

Hardware Bill of Materials for the Compute Nodes

Quantity	Item	Manufacturer	Model
32	Intel® Server Chassis	Intel	R2000WT
32	Intel® Server Board (w/10Gb Intel® Ethernet Controller)	Intel	S2600WFD
	(2x) Intel® Xeon® Gold Processor	Intel	6148
	(8x) 8GB ECC DDR4 2666MHz	Micron	MTA18ASF1G72PZ
	(1x) Intel® Omni-Path Host Fabric Interface Adapter	Intel	100 Series 1 Port PCIe x16

1.3 Software Bill of Materials

Software	Version
CentOS* Linux* Installation DVD	7.3.1611
Intel® HPC Orchestrator Advanced	17.01.006.ga
Intel® Parallel Studio XE Cluster Edition	2018 Beta Update 1
Intel® Omni-Path Fabric Software	10.3.1.0.22
(Compliance to) Intel® Scalable System Framework	2016.0

1.4 Conventions

Certain conventions used in this reference design are contingent upon the hardware and software listed in the bill of materials, the document date of release, and developer preference. These conventions are distinguished by red text throughout this document. In theory, other suitable values may be used in place of these conventions.

Cluster/Subcluster

- Internal Subnet (Ethernet): `192.168.1.0`
- Internal Netmask (Ethernet): `255.255.255.0` (/24)
- BMC Subnet: `192.168.100.0`
- BMC Netmask: `255.255.255.0` (/24)
- Internal Subnet (Intel® Omni-Path): `192.168.5.0`
- Internal Netmask (Intel® Omni-Path): `255.255.255.0` (/24)

Head Node

- Hostname: `frontend`
- Internal IP Address (Ethernet): `192.168.1.254`
- Internal Network Device (Ethernet): `enp61s0f0`
- External Network Device (Ethernet): `enp61s0f1`
- BMC IP Address: `192.168.100.254`
- Hostname (Intel® Omni-Path): `frontend-ib0`
- Internal IP Address (Intel® Omni-Path): `192.168.5.254`
- Internal Network Device (Intel® Omni-Path): `ib0`

Compute Node

The prescribed networks allow for a convenient pairing of node hostnames to IP addresses – e.g. `cX` to `192.168.1.X`. Hostname representation in this document varies to best fit the situation. When used, `N` and `X` are meant represent a number without zero-padding (e.g. `c1, c42, c222`). As a means of disambigutation, `N` is used to specify the last host, and `X` used to denote a single host.

- Hostnames: `cX`
- Internal IP Addresses (Ethernet): `192.168.1.X`
- Internal Network Device (Ethernet): `enp61s0f0`
- BMC IP Addresses: `192.168.100.X`
- Internal IP Addresses (Intel® Omni-Path): `192.168.5.X`
- Internal Network Device (Intel® Omni-Path): `ib0`

Chapter 2

How to Read This Document

2.1 Prefacing Comments About Interpreting This Document

Often times prose is added to provide an overview of the purpose of the section.

ADVISORY: These messages are meant to draw the attention of the reader.

Advisory statements are meant to highlight that further action may be required in certain circumstances or to explain how and why a step is performed.

1. Enumerated instructions indicate a required step.

When file content or terminal output is referenced, it appears in the following format:

```
#!/bin/sh

# This is an inline file.
echo "Hello world!"
```

Commands to execute are often grouped into logical execution blocks. From these blocks, it is simple to copy and paste several commands to the terminal at once. Blocks of commands appear in the following format:

```
cat >>./count.py <<_EndOfFile_
#!/usr/bin/python
print([i for i in range(10)])
_EndOfFile_
chmod 755 ./count.py
python count.py
```

Chapter 3

Preparation

3.1 Assembly

The system is a simple Beowulf style cluster, consisting of a single head node managing all cluster functions and one or more compute nodes for processing. On every compute node, the first Ethernet port (corresponding to network device name `enp61s0f0`) is used to connect to the cluster private network. Similarly on the head node, the first Ethernet port (corresponding to network device name `enp61s0f0`) is used to connect to the cluster private network. On the head node, the second Ethernet port (corresponding to network device name `enp61s0f1`) is used to connect to networks outside the cluster.

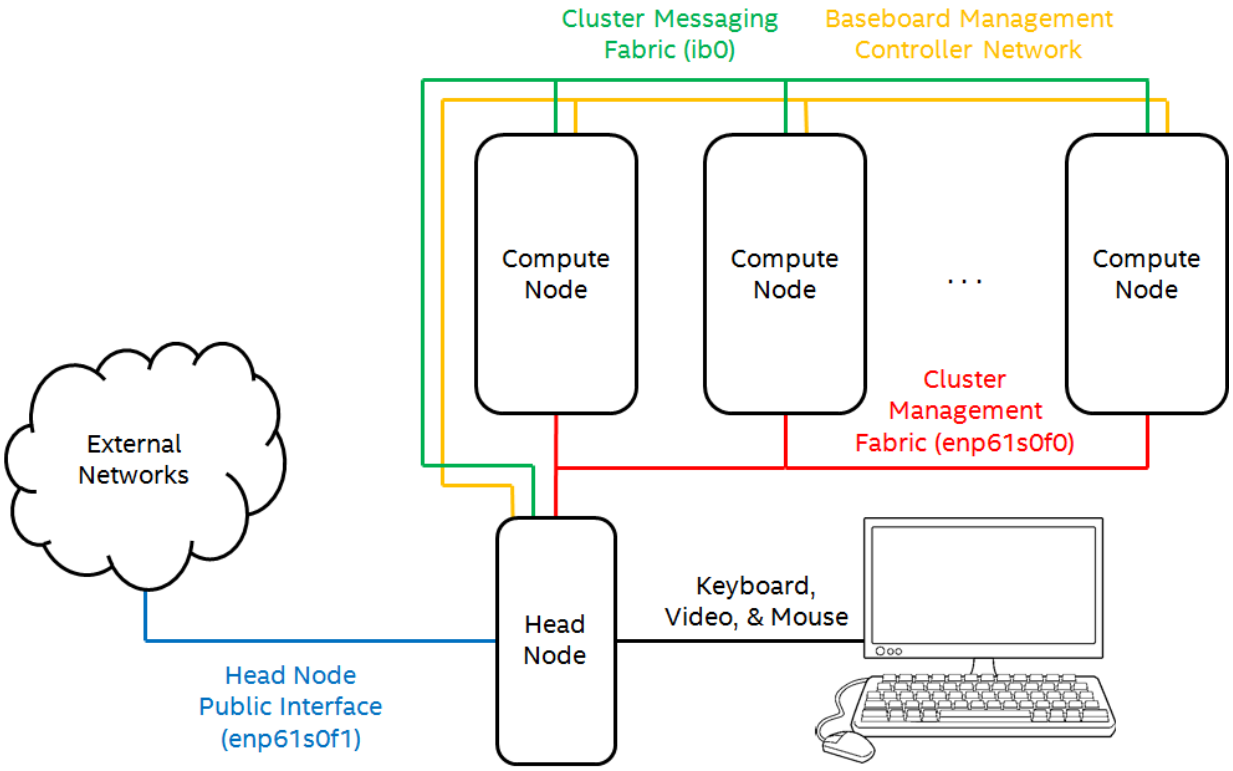

Figure 3.1: A visualization of system networks.

1. Collect MAC addresses.

Before beginning this reference design, record the MAC address for the first Ethernet port of each node. The MAC address may be found printed on a physical sticker on the rear of the server. This information may otherwise be acquired from the BIOS.

2. Collect BMC LAN information.

This reference design assumes that BMC LAN is configured for every compute node in the cluster. It may be necessary to cable these interfaces and configure the ports in each system BIOS. Before beginning this reference design, record the BMC LAN information for each node. If BMC LAN is already configured, this information may be acquired from the BIOS.

Chapter 4

Cluster Setup

4.1 Install the Linux* Operating System from Disc

ADVISORY: This section requires the following software component:

- `CentOS-7-x86_64-DVD-1611.iso`

1. Insert the CentOS* 7.3 install disc. Boot from disc and select "Install CentOS 7".

2. Select "English" as the language and click "Continue".

3. Select "DATE & TIME".

a. Select the timezone using the geographical location dropdown menus.

b. Click "Done" to return home.

4. Select "INSTALLATION DESTINATION".

a. Click all disks from "Local Standard Disks" to select them for installation. A selected disk will have a check mark over its icon. It may be necessary to scroll right to view all disks.

b. Under "Other Storage Options", select "I will configure partitioning.".

c. Click "Done" to go to the partitioning wizard.

d. In the left panel, expand any collapsed partition lists and delete any existing partitions by selecting each one and clicking "-" in turn. This will cause all local data to be deleted during installation.

e. Create a new partition by clicking "+".

(1) Select "/boot" from the mount point dropdown.

(2) Enter "1024 MiB" for desired capacity. Be sure to leave a space between the number and unit.

(3) Click "Add mount point".

f. Create a new partition by clicking "+".

(1) Select "/" from the mount point dropdown.

(2) Enter "200 GiB" for desired capacity. Be sure to leave a space between the number and unit.

(3) Click "Add mount point".

g. Create a new partition by clicking "+".

(1) Select "swap" from the mount point dropdown.

(2) Enter “32 GiB” for desired capacity. Be sure to leave a space between the number and unit.

(3) Click “Add mount point”.

h. Create a new partition by clicking “+”.

(1) Select “/home” from the mount point dropdown.

(2) Leave the desired capacity field empty. This will cause the partition to use any remaining space.

(3) Click “Add mount point”.

i. Click “Done” to return home. Accept all changes if prompted.

5. Select “NETWORK & HOST NAME”.

a. Enter “`frontend`” as the hostname.

(1) Click the “Apply” button for the hostname to take effect.

b. Select “Ethernet `enp61s0f0`” and click “Configure” to setup the internal cluster interface.

(1) From the “General” section, check “Automatically connect to this network when it is available”.

(2) From the “IPv4 Settings” section, select the “Manual” method and add the address `192.168.1.254` with netmask `24`. Save and exit.

c. Select “Ethernet `enp61s0f1`” and click “Configure” to setup the external cluster interface.

(1) From the “General” section , check “Automatically connect to this network when it is available”.

(2) Configure the external interface as necessary. Save and exit.

d. Set the toggle to “ON” for both interfaces.

e. Click “Done” to return home.

6. Select “SOFTWARE SELECTION”.

a. In the box labeled “Base Environment” on the left side, select “Infrastructure Server”.

b. Click “Done” to return home.

7. Wait until the “Begin Installation” button is not greyed out, which may take several minutes. Then click it to continue.

8. While waiting for the installation to finish, set the root password.

9. Click “Reboot” when the installation is complete.

10. Boot from the primary drive.

11. Login as root.

ADVISORY: The remaining steps in this document are all designed to be completed from a command line interface unless otherwise stated.

The command line interface is the default local login method since no GUI is provided in the prescribed operating system installation. Remote SSH login is also available by default. It is suggested to use the login method that makes it easiest to copy and paste commands from this document.

ADVISORY: This document includes the following best practices for package installation.

For package installations, it is preferable to use the YUM command line tool as opposed to RPM because YUM will autoresolve package dependencies.

The environment variable $CHROOT defines the path to the filesystem used to build the compute node image for provisioning. When installing packages to $CHROOT, it is preferable to utilize YUM repositories from the $CHROOT as opposed to YUM repositories from the head node.

4.2 Place Required Software Components and File on System

Place the following required software into `/root` for later installation.

- `Intel_HPC_Orchestrator-rhel7.2-17.01.006.ga.iso`
- `parallel_studio_xe_2018_beta_update1_cluster_edition.tgz`
- `psxe2017.lic` (license filenames will vary)
- `psxe2018.lic` (license filenames will vary)
- `IntelOPA-IFS.RHEL73-x86_64.10.3.1.0.22.tgz`

4.3 Configure YUM

ADVISORY: If necessary, configure an Internet proxy.

If the public network implements a proxy server for Internet access, YUM must be configured to use it.

a. Open the `/etc/yum.conf` file for editing.

b. Under the `[main]` section, append the following line, replacing `<address>` and `<port>`.

```
proxy=http://<address>:<port>
```

c. Save the file and exit.

12. Update the system to latest package versions.

```
yum -y update
```

ADVISORY: Take caution with future OS updates.

Certain procedures in this reference design require packages to be built against the kernel. A future kernel update is likely to break the compatibility of these built packages with the new kernel, so take heed when updating packages following the deployment of this reference design. To provide longevity to this reference design while simultaneously incoporating security updates, it may be convenient to perform only security-critical updates as follows.

```
yum -y --security update
```

Even so, it may be necessary to occasionaly rebuild kernel-dependent packages.

13. Reboot the server.

Reboot the server to run the updated kernel for the remainder of the installation. Boot to the primary disk.

```
init 6
```

4.4 Enable EPEL* and Intel® HPC Orchestrator Advanced Repositories

ADVISORY: This section requires the following software components:

- `Intel_HPC_Orchestrator-rhel7.2-17.01.006.ga.iso`
- `psxe2017.lic`

14. Enable the EPEL* repository.

The Intel® HPC Orchestrator Advanced release package provides GPG keys for package signing as well as YUM repository configuration files. The release package may be installed from the mounted ISO. Intel® HPC Orchestrator Advanced builds on the base of CentOS* and also requires some packages from EPEL*.

```
yum -y install epel-release
```

15. Enable the Intel® HPC Orchestrator Advanced repository.

```
if [ ! -d "/mnt/hpc_orch_iso" ]; then mkdir -p /mnt/hpc_orch_iso; fi
echo -n "/root/Intel_HPC_Orchestrator-rhel7.2-17.01.006.ga.iso " >>/etc/fstab
echo "/mnt/hpc_orch_iso auto loop 0 0" >>/etc/fstab
mount /mnt/hpc_orch_iso
rpm -Uvh /mnt/hpc_orch_iso/x86_64/Intel_HPC_Orchestrator_release-*.x86_64.rpm
rpm --import /etc/pki/pgp/HPC-Orchestrator*.asc
rpm --import /etc/pki/pgp/PSXE-keyfile.asc
```

16. Install the Intel® Parallel Studio XE 2017 license.

```
mkdir -p /opt/intel/licenses
mv /root/psxe2017.lic /opt/intel/licenses
```

4.5 Setup Intel® HPC Orchestrator Advanced Automation Scripts

17. Make a copy of the scripts to edit and use.

Intel® HPC Orchestrator Advanced also provides automation scripts to simplify building a cluster. These scripts are copied to an alternate directory to be later edited and used in this reference design.

```
yum -y install docs-orch
mkdir -p /root/hpc-orch-install
cp -pr /opt/intel/hpc-orchestrator/pub/doc/recipes/ /root/hpc-orch-install
chmod 700 /root/hpc-orch-install
```

18. Specify the installation parameters.

Intel® HPC Orchestrator Advanced provides a template `input.local` file to enter and store the installation parameters. This file is copied to `/root/hpc-orch-install/recipes/input.local`. Edit this file to reflect the conventions and purposes of this reference design. The contents of this file used in testing are listed below (with sensitive information redacted). Note that many of these parameters - NTP server, BMC username and password, MAC addresses, BMC (IPMI) IP addresses - must be set to fit the environment in which the cluster is setup. Where `# ...` appears, the variable array may be extended to accomodate additional compute nodes.

```
# -*-sh-*-
# ------------------------------------------------------------------------------
# ------------------------------------------------------------------------------
# Template input file to define local variable settings for use with
# an installation recipe.
# ------------------------------------------------------------------------------

# ---------------------------
# cluster fabric technology
# ---------------------------
```

```
# set to 1 for OPA, otherwise 0 for IB
opa_fabric=1

# ---------------------------
# SMS (master) node settings
# ---------------------------

# Flag to recreate ssh keys on new install
recreate_keys=0

# Hostname for master server (SMS)
sms_name=frontend

# Local (internal) IP address on SMS
sms_ip=192.168.1.254

# Internal ethernet interface on SMS
sms_eth_internal=enp61s0f0

# Subnet netmask for internal cluster network
internal_netmask=255.255.255.0

# Provisioning interface used by compute hosts
eth_provision=enp61s0f0

# Local ntp server for time synchronization
ntp_server=<ntp_server>

# BMC user credentials for use by IPMI
ipmi_username=<bmc_username>
IPMI_PASSWORD=<bmc_password>

# Additional time to wait for compute nodes to provision (seconds)
provision_wait=300

# Optional Stateful install device
stateful_dev="${stateful_dev:-sda}"

# Flags for optional installation/configuration

enable_clustershell=0
enable_ipmisol=1
enable_ipoib=1
enable_ganglia=0
enable_kargs=0
enable_lustre_client=0
enable_mrsh=0
enable_nagios=0
enable_powerman=1
enable_stateful=0
enable_ssf=1


# -------------------------
```

```
# compute node settings
# ------------------------

# Set location of local BOS mirror
BOS_MIRROR=http://mirror.centos.org/centos/7.3.1611/os/x86_64

# Prefix for compute node hostnames
nodename_prefix=c

# compute node IP addresses
c_ip[0]=192.168.1.1
c_ip[1]=192.168.1.2
c_ip[2]=192.168.1.3
c_ip[3]=192.168.1.4
# ...

# compute node MAC addreses for provisioning interface
c_mac[0]=00:1a:2b:3c:4f:56
c_mac[1]=00:1a:2b:3c:4f:57
c_mac[2]=00:1a:2b:3c:4f:58
c_mac[3]=00:1a:2b:3c:4f:59
# ...

# compute node BMC addresses
c_bmc[0]=192.168.100.1
c_bmc[1]=192.168.100.2
c_bmc[2]=192.168.100.3
c_bmc[3]=192.168.100.4
# ...

#-------------------
# Optional settings
#-------------------

# additional arguments to enable optional arguments for bootstrap kernel
kargs="${kargs:-acpi_pad.disable=1}"

# Lustre MGS mount name
mgs_fs_name="${mgs_fs_name:-192.168.100.254@o2ib:/lustre1}"

# Subnet netmask for IPoIB network
ipoib_netmask=255.255.255.0

# IPoIB address for SMS server
sms_ipoib=192.168.5.254

# IPoIB addresses for computes
c_ipoib[0]=192.168.5.1
c_ipoib[1]=192.168.5.2
c_ipoib[2]=192.168.5.3
c_ipoib[3]=192.168.5.4
# ...
```

4.6 Run Intel® HPC Orchestrator Advanced Automation Scripts

19. Preparation

Intel® HPC Orchestrator Advanced installs and configures NTP as the system time daemon. Before configuring the system, remove the default chrony time daemon from the head node.

```
yum -y remove chrony
```

20. Run the Intel® HPC Orchestrator Advanced recipe script.

The `recipe.sh` script will serially invoke subscripts, each requiring the user to select between continuing and aborting.

After executing the steps in the following commands, pay close attention to the prompts. Press "y" followed by "Enter" to proceed through (Section 3.1)-(Section 3.2), (Section 3.4)-(Section 3.7), and (Section 3.8)-(Section 3.10.2). When prompted to take action for (Section 3.11)-(Section 3.11.9), do not enter input. Instead, allow the session to remain idle at this prompt and proceed to the next section of this reference design.

```
cat >>/root/.bashrc <<_EndOfContent_
export INPUT_LOCAL=/root/hpc-orch-install/recipes/input.local
export CHROOT=/opt/intel/hpc-orchestrator/admin/images/rhel7.2
_EndOfContent_
source /root/.bashrc
cd /root/hpc-orch-install/recipes/warewulf/slurm
./recipe.sh
```

4.7 Switch to a New Session

With the session running `recipe.sh` left running idle, launch a new separate session for additional system configuration. This is done to circumvent having to abort and later relaunch `recipe.sh`.

4.8 Install Intel® Omni-Path Software on Head Node

ADVISORY: This section requires the following software component:

- `IntelOPA-IFS.RHEL73-x86_64.10.3.1.0.22.tgz`

This package installs the Intel® Omni-Path Software components needed to set up compute, I/O, and service nodes with drivers, stacks and basic tools for local configuration and monitoring. It also installs the Intel® Omni-Path Fabric Suite FastFabric tools and the Intel® Omni-Path Fabric Suite Fabric Manager.

21. Extract the installer.

```
tar -zxvf IntelOPA-IFS.RHEL73-x86_64.10.3.1.0.22.tgz -C /usr/src
```

22. Install prerequisite packages.

```
yum -y install \
opa-support-orch opa-support-devel-orch \
libhfi1 libuuid-devel papi elfutils-libelf-devel libpfm
```

23. Install the Intel® Omni-Path Software.

The `infiniband-diags` package from the Intel® Omni-Path Fabric Suite supplies the RDMA node description daemon (`rdma-ndd`) that generates meaningful node descriptions used by fabric tools.

```
cd /usr/src/IntelOPA-IFS.RHEL73-x86_64.10.3.1.0.22
yum -y install IntelOPA-OFED_DELTA.RHEL73-x86_64.10.3.1.0.20/RPMS/redhat-ES73/\
infiniband-diags-1.6.7-2.el7.x86_64.rpm
./INSTALL -i opa -i ipoib -i psm_mpi -i pgas -i fastfabric -i opafm
./INSTALL -s
./INSTALL -E opafm
systemctl enable rdma-ndd.service
systemctl start rdma-ndd
cd ~
```

In the final lines of the installer output, a warning is printed:

```
A System Reboot is recommended to activate the software changes
```

Despite this warning, it is **NOT** necessary to reboot the system at this time. The system is rebooted in a later step.

24. Install OFI fabric on the head node.

```
yum -y install libfabric libfabric-devel
```

25. Configure the Intel® Omni-Path Software.

Add an IPoIB entry for the head node to `/etc/hosts`.

```
source $INPUT_LOCAL
systemctl enable opafm
echo "${sms_ipoib} ${sms_name}-ib0" >>/etc/hosts
```

Ensure supporting drivers are included in the compute node bootstrap image.

```
cat >>/etc/warewulf/bootstrap.conf <<_EndOfFile_
drivers += updates, extra/ifs-kernel-updates, hfi1, rdmavt
firmware += updates/hfi1*
_EndOfFile_
```

4.9 Install Intel® Omni-Path Software on Compute Node Image

26. Install prerequisite base packages.

Multiple base packages are required on the compute nodes for the Intel® Omni-Path Software.

```
chroot $CHROOT yum -y install \
opa-support-orch libhfi1 libuuid-devel elfutils-libelf-devel papi
```

27. Install the Intel® Omni-Path Software.

```
mkdir -p $CHROOT/mnt/opa/IntelOPA-IFS.RHEL73-x86_64.10.3.1.0.22
mount -o bind {/usr/src/,$CHROOT/mnt/opa}/IntelOPA-IFS.RHEL73-x86_64.10.3.1.0.22
chroot $CHROOT <<_EndOfCommands_
cd /mnt/opa/IntelOPA-IFS.RHEL73-x86_64.10.3.1.0.22
yum -y install IntelOPA-OFED_DELTA.RHEL73-x86_64.10.3.1.0.20/RPMS/redhat-ES73/\
infiniband-diags-1.6.7-2.el7.x86_64.rpm
./INSTALL -i opa -i ipoib -i psm_mpi -i pgas
./INSTALL -s
```

```
systemctl enable rdma-ndd.service
_EndOfCommands_
umount $CHROOT/mnt/opa/IntelOPA-IFS.RHEL73-x86_64.10.3.1.0.22
```

In the final lines of the installer output, a warning is printed:

```
A System Reboot is recommended to activate the software changes
```

Despite this warning, it is **NOT** necessary to reboot the system at this time. The system is rebooted in a later step.

28. Install OFI fabric on the compute node.

```
chroot $CHROOT yum -y install libfabric
```

4.10 Configure Hardware Details in Slurm*

```
sed -i \
-e "/^NodeName/ s/\(Sockets\)=[0-9]*/\1=2/" \
-e "/^NodeName/ s/\(CoresPerSocket\)=[0-9]*/\1=20/" \
-e "/^NodeName/ s/\(ThreadsPerCore\)=[0-9]*/\1=2/" \
/etc/slurm/slurm.conf
```

4.11 Enable Filesystem Hybridization in Warewulf*

```
sed -i \
-e 's/# \(hybridpath = \)/\1/' \
-e '/hybridpath/ s%/var/chroots%/opt/intel/hpc-orchestrator/admin/images%' \
/etc/warewulf/vnfs.conf
```

4.12 Return to the Idle Session

It is now safe to close the active session and return to the session left idle running `recipe.sh`.

4.13 Run Intel® HPC Orchestrator Advanced Automation Scripts (Continued)

29. Run the Intel® HPC Orchestrator Advanced recipe script.

Return to the session left idle running `recipe.sh`. Proceed through the rest of the sections by pressing “y” followed by “Enter” at each prompt.

Do not proceed beyond this point until `recipe.sh` has finished execution. Take caution not to prematurely reboot the system.

30. Reboot the server.

For some of the configuration settings to take effect, a reboot is necessary.

```
init 6
```

4.14 Run Intel® HPC Orchestrator Advanced Test Suite

Before performing additional customizations or updates, it is necessary to validate the installation using Intel® HPC Orchestrator Advanced’s test suite.

ADVISORY: Wait a few minutes to run these tests.

After rebooting, it may take a number of minutes for the compute nodes' NTP services to trust the head node as a time server. The indicator for established trust is an asterisk next to the frontend hostname when running `ntpq -p` on a compute node.

```
pdsh -g compute ntpq -p
```

If trust is not yet established on one or more compute nodes, the "admin/run" root-level test from the Intel® HPC Orchestrator Advanced test suite will fail.

31. Make compute nodes available in Slurm*.

After rebooting, the Slurm* controller service will put nodes in a drain state. This may be observed via `sinfo`. Make all the nodes idle again.

```
scontrol update NodeName=$(nodeattr -c compute) State=idle
```

32. Prepare the environment.

```
export I_MPI_SCALABLE_OPTIMIZATION=off
export IPMI_PASSWORD=<bmc_password>
module load testsuite
```

33. Run the test suite.

There are 5 root-level tests. Verify they all pass. Refer to `/opt/intel/hpc-orchestrator/pub/tests/testsuite/test-suite.log` for results.

```
run-tests -r
```

There are 10 user-level tests. Verify they all pass. Refer to `/opt/intel/hpc-orchestrator/pub/tests/testsuite/test-suite.log` for results.

```
run-tests -u
```

4.15 Install Intel® Parallel Studio XE 2018 Beta Update 1 Cluster Edition

ADVISORY: This section requires the following software components:

- `parallel_studio_xe_2018_beta_update1_cluster_edition.tgz`
- `psxe2018.lic`

ADVISORY: Support for Intel® Parallel Studio XE 2018 Beta Update 1.

Intel® Parallel Studio XE 2018 Beta Update 1 is not a component of Intel® HPC Orchestrator Advanced 17.01.006. Please direct all support questions directly to Intel® Parallel Studio support.

Some applications included with Intel® HPC Orchestrator Advanced may not compile or run correctly with Intel® Parallel Studio XE 2018 Beta Update 1. These applications are using compiler or environment options that are no longer supported. Application source must be updated to replace discontinued or deprecated options. To find a list of these options, refer to Intel® Cluster Tools Depcrecation Information and Deprecated and Removed Compiler Options.

To load components from Intel® Parallel Studio XE 2017 Update 1, use the enrivonment modules from Intel® HPC Orchestrator Advanced (e.g. `module load clck`). To load components from Intel® Parallel Studio XE 2018 Beta Update 1, use the packaged environment variable scripts (e.g. `source /opt/intel/clck/2018.0.001/bin/clckvars.sh`).

34. Extract the installer.

```
tar -zxvf parallel_studio_xe_2018_beta_update1_cluster_edition.tgz -C /usr/src
```

35. Move the license file to expected location.

```
mv /root/psxe2018.lic /opt/intel/licenses
```

36. Configure the installation parameters.

Modify the configuration file for the silent install of Intel® Parallel Studio XE 2018 Beta Update 1 Cluster Edition. The changes to be made include: accept EULA, change the architecture to "INTEL64" and change the options related to MPSS to "no". Also, for this installation, it is expected to use the license file and that the license file be present in /opt/intel/licenses. However, there are other methods of providing the license as mentioned in the configuration file for silent install. Look into these configurations if the license file needs to be provided in a different way. To access the EULA, refer to /usr/src/parallel_studio_xe_2018_beta_update1_cluster_edition/license.txt

```
sed -i \
  -e "/^ACCEPT_EULA/ s/decline/accept/" \
  -e "/^ARCH_SELECTED/ s/ALL/INTEL64/" \
  -e "/^MPSS/ s/yes/no/" \
  -e "/^ACTIVATION_TYPE/ s/exist_lic/license_file/" \
  -e "s%#\(ACTIVATION_LICENSE_FILE=\)%\1/opt/intel/licenses/psxe2018.lic%" \
  /usr/src/parallel_studio_xe_2018_beta_update1_cluster_edition/silent.cfg
```

37. Start the installer.

The installation will take a few minutes.

```
cd /usr/src/parallel_studio_xe_2018_beta_update1_cluster_edition/
./install.sh -s silent.cfg
cd ~
```

Chapter 5

Validation

5.1 Configure Intel® Cluster Checker

1. Modify the Intel® Cluster Checker configuration file.

Open the `/opt/intel/clck/2018.0.001/etc/clck.xml` file for editing.

Modify the file so that it looks like this. Review the inline comments for justification:

```
<?xml version="1.0" encoding="UTF-8"?>
<configuration>
<analyzer>
  <config>
  </config>
  <suppressions>
    <suppress>
      <!-- compute nodes are diskless -->
      <id>iozone-ran-no-bandwidth</id>
    </suppress>
    <suppress>
      <!-- dapl not needed for Intel® Omni-Path -->
      <id>datconf-no-dapl-providers</id>
    </suppress>
    <suppress>
      <!-- Intel® Xeon® Processor Scalable Family detection feature is
           not available in this 2018 beta release -->
      <id>cpu-not-compliant</id>
    </suppress>
  </suppressions>
</analyzer>
<collector>
  <!-- specify the head node private network interface -->
  <network_interface>enp61s0f0</network_interface>
</collector>
<database>
</database>
</configuration>
```

Save the file and exit.

5.2 Run Intel® Cluster Checker

2. Switch to the test user.

To run Intel® Cluster Checker, switch to the `test` user.

```
su - test
```

3. Create the nodelist.

The application makes use of the shared home directory for storing temporary files. The application requires a nodelist – a file containing node and node group information. This nodelist is easily generated by reusing the genders node data.

```
echo "$(nodeattr -c sms) # role:head" > nodelist
echo "$(nodeattr -c compute | tr ',' '\n')" >> nodelist
```

4. Source environment scripts for Intel® Cluster Checker.

```
source /opt/intel/clck/2018.0.001/bin/clckvars.sh
```

5. Run the collector.

Intel® Cluster Checker separates the tasks of data collection and data analysis into two separate tools. The nodes may only be accessed through Slurm*, so it is necessary to secure an allocation of all nodes for which collection is to occur. Run these commands below to collect the data.

```
salloc --nodelist=$(nodeattr -c compute)
clck-collect -f nodelist -a
exit
```

The `clck-collect` command above instructs Intel® Cluster Checker to collect information for all checks for every node in the nodelist.

ADVISORY: Perform HPC cluster health checks via Intel® CLCK framework definitions.

Prior to performing validation, it is recommended to identify any general HPC cluster health issues. This may be done using framework definitions with the following command:

```
clck-analyze -f nodelist -F $CLCK_ROOT/etc/rulesets/health.xml \
-p diagnosed_signs
```

The analyzer returns the list of checks performed, the list of nodes checked, and also the results of the analysis. The results are essentially a synopsis of the cluster health. Fix issues as needed before performing validation.

6. Perform Intel® SSF validation via Intel® SSF framework definitions.

Run the analyzer to validate against the Intel® Scalable System Framework Reference Architecture Specification. This may be done using framework definitions with the following command:

```
clck-analyze -f nodelist -F $CLCK_ROOT/etc/rulesets/IntelSSF/compat-hpc.xml \
-p diagnosed_signs
```

The analyzer results highlight all instances of non-compliance against the Intel® Scalable System Framework Reference Architecture Specification. Analysis of this reference design results with a “PASS” output indicative of compliance.

Appendix A

Design Considerations

1. Intel® HPC Orchestrator Advanced Validation and Support

The version of Intel® HPC Orchestrator Advanced from this reference design is validated using RHEL* 7.2 as the operating system. However, both RHEL* 7.2 and CentOS* 7.X (which defaults to latest, presently 7.3) are supported. Instances of RHEL* 7.2 (e.g. the name of the compute node $CHROOT and VNFS) are indicative of the OS used for Intel® HPC Orchestrator Advanced validation, not the OS used in this reference design.

2. Intel® Omni-Path Fabric – Subnet Manager Version Compatibility

The Intel® Omni-Path Edge Switch used in this reference design does not run a subnet manager. For switches with that capability, ensure the subnet manager version matches the version from the Intel® Omni-Path Software distribution. The subnet manager belongs to the fabric manager subcomponent of the software distibution. In many cases, the version of the fabric manager differs from that of the overarching software distribution.

In this reference design, the overarching Intel® Omni-Path Software distribution version is **10.3.1.0.22** (e.g. IntelOPA-IFS.RHEL73-x86_64.**10.3.1.0.22**), but the fabric manager version is **10.3.1.0.5** (e.g. IntelOPA-FM.RHEL73-x86_64.**10.3.1.0.5**).

3. Drivers for Intel® Ethernet Controller X722

CentOS* 7.3 provides version 1.5.10-k of the i40e and i40evf network drivers to support Intel® Ethernet Controller X722. Older versions of this driver (e.g. 1.3.21-k from CentOS* 7.2) may not support this device.

To obtain the latest drivers for Intel® Ethernet Controller X722, please visit the Intel download center.

4. Memory Configuration

In development of this reference design, it was observed that systems utilizing all 6 memory channels per CPU socket (i.e. 48GB per CPU) had considerably better performance than those utilizing only 4 memory channels per CPU socket (i.e. 32GB per CPU) when running DGEMM with AVX-512 enabled.

To achieve best performance for a specific application with this server platform, please contact your Intel representative about obtaining document number 569458.