



Intel® Omni-Path Fabric

Setup Guide

Rev. 7.0

October 2017



You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest Intel product specifications and roadmaps.

The products described may contain design defects or errors known as errata which may cause the product to deviate from published specifications. Current characterized errata are available on request.

Copies of documents which have an order number and are referenced in this document may be obtained by calling 1-800-548-4725 or visit <http://www.intel.com/design/literature.htm>.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with your system manufacturer or retailer or learn more at intel.com.

No computer system can be absolutely secure.

Intel, the Intel logo, Intel Xeon Phi, and Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.

Copyright © 2016–2017, Intel Corporation. All rights reserved.



Revision History

For the latest documentation, go to: <http://www.intel.com/omnipath/FabricSoftwarePublications>.

Date	Revision	Description
October 2017	7.0	Updates to this document include: <ul style="list-style-type: none"> Updated Before You Begin, Installation Prerequisites, Install Intel® Omni-Path Software, and Link Debug CLI Commands. Changed bios_images filepath in Program and Verify Option ROM EEPROM Device to /usr/share/opa/bios_images. The <i>Intel® Omni-Path Fabric Suite FastFabric Command Line Interface Reference Guide</i> has been merged into the <i>Intel® Omni-Path Fabric Suite FastFabric User Guide</i>. In this document, all references have been updated appropriately. See the Intel® Omni-Path Documentation Library for details.
August 2017	6.0	Updates to this document include: <ul style="list-style-type: none"> Updated Configure BIOS Settings. Added NVIDIA* bullet in the Overview section of Install Intel® Omni-Path Software. Clerical changes.
April 2017	5.0	Updates to this document include: <ul style="list-style-type: none"> Changed the title of this document from <i>Intel® Omni-Path Fabric Staging Guide</i> to <i>Intel® Omni-Path Fabric Setup Guide</i>. Updated Decode the Physical Configuration of an HFI section with TMM details. Added Program and Verify Option ROM EEPROM Device section. Globally, updated the following filepaths: <ul style="list-style-type: none"> from /etc/sysconfig/opafm.xml to /etc/opa-fm/opafm.xml from /usr/lib/opa/src/ to /usr/src/opa/ from /etc/sysconfig/ to /etc/ Added Intel® Omni-Path Documentation Library to Preface.
December 2016	4.0	Updates to this document include: <ul style="list-style-type: none"> Updated Take State Dump of a Switch procedure. Globally, updated the following filepaths: <ul style="list-style-type: none"> from /opt/opa to /usr/lib/opa from /var/opt/opa to /var/usr/lib/opa from /opt/opafm to /usr/lib/opa-fm from /var/opt/opafm to /var/usr/lib/opa-fm Added Cluster Configurator for Intel® Omni-Path Fabric to Preface.
October 2016	3.0	Document has been updated to add the following sections: <ul style="list-style-type: none"> Perform Initial Fabric Verification. Edit Hosts and Allhosts Files. Defining Type in the Topology Spreadsheet. Configure Intel® Omni-Path Director Class Switch 100 Series. Configure Host Setup.
continued...		



Date	Revision	Description
		<ul style="list-style-type: none">Fabric Manager Routing Algorithm.
August 2016	2.0	Updates to this document include: <ul style="list-style-type: none">Updated Generate Cable Map Topology Files to include Intel® Omni-Path Director Class Switch 100 Series information.
May 2016	1.0	Initial release.

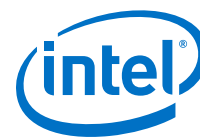


Contents

Revision History.....	3
Preface.....	8
Intended Audience.....	8
Intel® Omni-Path Documentation Library.....	8
Cluster Configurator for Intel® Omni-Path Fabric.....	10
Documentation Conventions.....	10
License Agreements.....	11
Technical Support.....	11
1.0 Introduction.....	12
2.0 Installation Prerequisites.....	14
2.1 Configure BIOS Settings.....	14
2.2 Configure OS Settings	14
3.0 TCP/IP Host Name Resolution.....	16
4.0 Install Intel® Omni-Path Software.....	17
4.1 Perform Initial Fabric Verification.....	18
4.2 Edit Hosts and Allhosts Files.....	18
5.0 Generate Cable Map Topology Files.....	19
5.1 Generate Cable Map Topology Files.....	19
6.0 Configure FastFabric.....	21
6.1 Format for IPoIB Host Names.....	21
6.2 Specify Test Areas for opaallanalysis.....	21
6.3 Location of mpi_apps Directory.....	21
7.0 Configure Managed Intel® Omni-Path Edge Switches.....	23
8.0 Configure Intel® Omni-Path Director Class Switch 100 Series.....	26
9.0 Configure Externally-Managed Intel® Omni-Path Edge Switches.....	27
10.0 Configure Host Setup.....	29
11.0 Verify Cable Map Topology.....	30
12.0 Verify Server and Fabric.....	31
13.0 Best Known Methods (BKMs) for Site Installation.....	32
13.1 Enable Intel® Omni-Path Fabric Manager GUI.....	32
13.2 Review Server and Fabric Verification Test Results.....	33
13.3 Debug Intel® Omni-Path Physical Link Issues.....	34
13.3.1 OPA Link Transition Flow.....	34
13.3.2 Verify the Fabric Manager is Running.....	35
13.3.3 Check the State of All Links in the System.....	35
13.3.4 Check the State of HFI Links from a Server.....	35
13.3.5 Link Width, Downgrades, and opafm.xml.....	36
13.3.6 How to Check Fabric Connectivity.....	36



13.3.7 Physical Links Stability Test using opacabletest.....	37
13.3.8 How to Debug and Fix Physical Link Issues.....	38
13.3.9 Link Debug CLI Commands.....	39
13.4 Use opatop for Bandwidth and Error Summary.....	40
13.5 Use the Beacon LED to Identify HFI and Switch Ports.....	41
13.6 Decode the Physical Configuration of an HFI.....	41
13.7 Program and Verify Option ROM EEPROM Device.....	42
13.8 Verify Fabric Manager Sweep.....	43
13.9 Verify PM Sweep Duration.....	44
13.10 Check Credit Loop Operation.....	44
13.11 Fabric Manager Routing Algorithm.....	45
14.0 Run Benchmark and Stress Tests.....	46
14.1 Run Bandwidth Test.....	46
14.2 Run Latency Test.....	46
14.3 Run MPI Deviation Test.....	46
14.4 Run run_mpi_stress.....	46
15.0 Take State Dump of a Switch.....	48
16.0 BKM for OPA Commands.....	49
16.1 Retrieve Host Fabric Interface (HFI) Temperature.....	49
16.2 Read Error Counters.....	49
16.3 Clear Error Counters.....	50
16.4 Load and Unload Intel® Omni-Path Host HFI Driver.....	50
16.5 Analyze Links.....	50
16.6 Trace Route between Two Nodes.....	51
16.7 Analyze All Fabric ISLs Routing Balance.....	51
16.8 Dump Switch ASIC Forwarding Tables.....	51
16.9 Configure Redundant Fabric Manager (FM) Priority.....	51
16.9.1 Configure FM Priority from a Local or Remote Terminal.....	52
16.9.2 Configure FM Elevated Priority.....	52
16.9.3 Configuration Consistency for Priority/Elevated Priority.....	52
16.9.4 Display FM states from the Management Node.....	52
17.0 Final Fabric Checks.....	53



Tables

1	HFI Temperature Output Definitions.....	49
2	Link Quality Values and Description.....	51



Preface

This manual is part of the documentation set for the Intel® Omni-Path Fabric (Intel® OP Fabric), which is an end-to-end solution consisting of Intel® Omni-Path Host Fabric Interfaces (HFIs), Intel® Omni-Path switches, and fabric management and development tools.

The Intel® OP Fabric delivers a platform for the next generation of High-Performance Computing (HPC) systems that is designed to cost-effectively meet the scale, density, and reliability requirements of large-scale HPC clusters.

Both the Intel® OP Fabric and standard InfiniBand* are able to send Internet Protocol (IP) traffic over the fabric, or *IPoFabric*. In this document, however, it is referred to as *IP over IB* or *IPoIB*. From a software point of view, IPoFabric and IPoIB behave the same way and, in fact, use the same `ib_ipoib` driver to send IP traffic over the `ib0` and/or `ib1` ports.

Intended Audience

The intended audience for the Intel® Omni-Path (Intel® OP) document set is network administrators and other qualified personnel.

Intel® Omni-Path Documentation Library

Intel® Omni-Path publications are available at the following URLs:

- Intel® Omni-Path Switches Installation, User, and Reference Guides
<http://www.intel.com/omnipath/SwitchPublications>
- Intel® Omni-Path Software Installation, User, and Reference Guides (includes HFI documents)
<http://www.intel.com/omnipath/FabricSoftwarePublications>
- Drivers and Software (including Release Notes)
<http://www.intel.com/omnipath/Downloads>

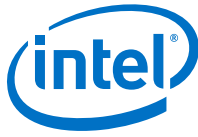
Use the tasks listed in this table to find the corresponding Intel® Omni-Path document.

Task	Document Title	Description
Key: Shading indicates the URL to use for accessing the particular document.		
• Intel® Omni-Path Switches Installation, User, and Reference Guides	http://www.intel.com/omnipath/SwitchPublications	
• Intel® Omni-Path Software Installation, User, and Reference Guides (includes HFI documents):	http://www.intel.com/omnipath/FabricSoftwarePublications (no shading)	
• Drivers and Software (including Release Notes):	http://www.intel.com/omnipath/Downloads	
continued...		



Task	Document Title	Description
Using the Intel® OPA documentation set	<i>Intel® Omni-Path Fabric Quick Start Guide</i>	A roadmap to Intel's comprehensive library of publications describing all aspects of the product family. It outlines the most basic steps for getting your Intel® Omni-Path Architecture (Intel® OPA) cluster installed and operational.
Setting up an Intel® OPA cluster	<i>Intel® Omni-Path Fabric Setup Guide</i> (Old title: <i>Intel® Omni-Path Fabric Staging Guide</i>)	Provides a high level overview of the steps required to stage a customer-based installation of the Intel® Omni-Path Fabric. Procedures and key reference documents, such as Intel® Omni-Path user guides and installation guides are provided to clarify the process. Additional commands and BKMs are defined to facilitate the installation process and troubleshooting.
Installing hardware	<i>Intel® Omni-Path Fabric Switches Hardware Installation Guide</i>	Describes the hardware installation and initial configuration tasks for the Intel® Omni-Path Switches 100 Series. This includes: Intel® Omni-Path Edge Switches 100 Series, 24 and 48-port configurable Edge switches, and Intel® Omni-Path Director Class Switches 100 Series.
	<i>Intel® Omni-Path Host Fabric Interface Installation Guide</i>	Contains instructions for installing the HFI in an Intel® OPA cluster. A cluster is defined as a collection of nodes, each attached to a fabric through the Intel interconnect. The Intel® HFI utilizes Intel® Omni-Path switches and cabling.
Installing host software Installing HFI firmware Installing switch firmware (externally-managed switches)	<i>Intel® Omni-Path Fabric Software Installation Guide</i>	Describes using a Text-based User Interface (TUI) to guide you through the installation process. You have the option of using command line interface (CLI) commands to perform the installation or install using the Linux* distribution software.
Managing a switch using Chassis Viewer GUI Installing switch firmware (managed switches)	<i>Intel® Omni-Path Fabric Switches GUI User Guide</i>	Describes the Intel® Omni-Path Fabric Chassis Viewer graphical user interface (GUI). It provides task-oriented procedures for configuring and managing the Intel® Omni-Path Switch family. Help: GUI online help.
Managing a switch using the CLI Installing switch firmware (managed switches)	<i>Intel® Omni-Path Fabric Switches Command Line Interface Reference Guide</i>	Describes the command line interface (CLI) task information for the Intel® Omni-Path Switch family. Help: -help for each CLI.
Managing a fabric using FastFabric	<i>Intel® Omni-Path Fabric Suite FastFabric User Guide</i> (Merged with: <i>Intel® Omni-Path Fabric Suite FastFabric Command Line Interface Reference Guide</i>)	Provides instructions for using the set of fabric management tools designed to simplify and optimize common fabric management tasks. The management tools consist of TUI menus and command line interface (CLI) commands. Help: -help and man pages for each CLI. Also, all host CLI commands can be accessed as console help in the Fabric Manager GUI.
Managing a fabric using Fabric Manager	<i>Intel® Omni-Path Fabric Suite Fabric Manager User Guide</i>	The Fabric Manager uses a well defined management protocol to communicate with management agents in every Intel® Omni-Path Host Fabric Interface (HFI) and switch. Through these interfaces the Fabric Manager is able to discover, configure, and monitor the fabric.
	<i>Intel® Omni-Path Fabric Suite Fabric Manager GUI User Guide</i>	Provides an intuitive, scalable dashboard and set of analysis tools for graphically monitoring fabric status and configuration. It is a user-friendly alternative to traditional command-line tools for day-to-day monitoring of fabric health. Help: Fabric Manager GUI Online Help.

continued...



Task	Document Title	Description
Configuring and administering Intel® HFI and IPoIB driver Running MPI applications on Intel® OPA	<i>Intel® Omni-Path Fabric Host Software User Guide</i>	Describes how to set up and administer the Host Fabric Interface (HFI) after the software has been installed. The audience for this document includes both cluster administrators and Message-Passing Interface (MPI) application programmers, who have different but overlapping interests in the details of the technology.
Writing and running middleware that uses Intel® OPA	<i>Intel® Performance Scaled Messaging 2 (PSM2) Programmer's Guide</i>	Provides a reference for programmers working with the Intel® PSM2 Application Programming Interface (API). The Performance Scaled Messaging 2 API (PSM2 API) is a low-level user-level communications interface.
Optimizing system performance	<i>Intel® Omni-Path Fabric Performance Tuning User Guide</i>	Describes BIOS settings and parameters that have been shown to ensure best performance, or make performance more consistent, on Intel® Omni-Path Architecture. If you are interested in benchmarking the performance of your system, these tips may help you obtain better performance.
Designing an IP or storage router on Intel® OPA	<i>Intel® Omni-Path IP and Storage Router Design Guide</i>	Describes how to install, configure, and administer an IPoIB router solution (Linux* IP or LNet) for inter-operating between Intel® Omni-Path and a legacy InfiniBand* fabric.
Building a Lustre* Server using Intel® OPA	<i>Building Lustre* Servers with Intel® Omni-Path Architecture Application Note</i>	Describes the steps to build and test a Lustre* system (MGS, MDT, MDS, OSS, OST, client) from the HPDD master branch on a x86_64, RHEL*/CentOS* 7.1 machine.
Building Containers for Intel® OPA fabrics	<i>Building Containers for Intel® Omni-Path Fabrics using Docker* and Singularity* Application Note</i>	Provides basic information for building and running Docker* and Singularity* containers on Linux*-based computer platforms that incorporate Intel® Omni-Path networking technology.
Writing management applications that interface with Intel® OPA	<i>Intel® Omni-Path Management API Programmer's Guide</i>	Contains a reference for programmers working with the Intel® Omni-Path Architecture Management (Intel OPAMGT) Application Programming Interface (API). The Intel OPAMGT API is a C-API permitting in-band and out-of-band queries of the FM's Subnet Administrator and Performance Administrator.
Learning about new release features, open issues, and resolved issues for a particular release	<i>Intel® Omni-Path Fabric Software Release Notes</i>	
	<i>Intel® Omni-Path Fabric Manager GUI Release Notes</i>	
	<i>Intel® Omni-Path Fabric Switches Release Notes</i> (includes managed and externally-managed switches)	

Cluster Configurator for Intel® Omni-Path Fabric

The Cluster Configurator for Intel® Omni-Path Fabric is available at: <http://www.intel.com/content/www/us/en/high-performance-computing-fabrics/omni-path-configurator.html>.

This tool generates sample cluster configurations based on key cluster attributes, including a side-by-side comparison of up to four cluster configurations. The tool also generates parts lists and cluster diagrams.

Documentation Conventions

The following conventions are standard for Intel® Omni-Path documentation:

- **Note:** provides additional information.
- **Caution:** indicates the presence of a hazard that has the potential of causing damage to data or equipment.



- **Warning:** indicates the presence of a hazard that has the potential of causing personal injury.
- Text in **blue** font indicates a hyperlink (jump) to a figure, table, or section in this guide. Links to websites are also shown in blue. For example:
See [License Agreements](#) on page 11 for more information.
For more information, visit www.intel.com.
- Text in **bold** font indicates user interface elements such as menu items, buttons, check boxes, key names, key strokes, or column headings. For example:
Click the **Start** button, point to **Programs**, point to **Accessories**, and then click **Command Prompt**.
Press **CTRL+P** and then press the **UP ARROW** key.
- Text in **Courier** font indicates a file name, directory path, or command line text. For example:
Enter the following command: `sh ./install.bin`
- Text in *italics* indicates terms, emphasis, variables, or document titles. For example:
Refer to *Intel® Omni-Path Fabric Software Installation Guide* for details.
In this document, the term *chassis* refers to a managed switch.

Procedures and information may be marked with one of the following qualifications:

- **(Linux)** – Tasks are only applicable when Linux* is being used.
- **(Host)** – Tasks are only applicable when Intel® Omni-Path Fabric Host Software or Intel® Omni-Path Fabric Suite is being used on the hosts.
- **(Switch)** – Tasks are applicable only when Intel® Omni-Path Switches or Chassis are being used.
- Tasks that are generally applicable to all environments are not marked.

License Agreements

This software is provided under one or more license agreements. Please refer to the license agreement(s) provided with the software for specific detail. Do not install or use the software until you have carefully read and agree to the terms and conditions of the license agreement(s). By loading or using the software, you agree to the terms of the license agreement(s). If you do not wish to so agree, do not install or use the software.

Technical Support

Technical support for Intel® Omni-Path products is available 24 hours a day, 365 days a year. Please contact Intel Customer Support or visit <http://www.intel.com/omnipath/support> for additional detail.



1.0 Introduction

This document provides a high-level overview of the steps required to set up an Intel® Omni-Path Fabric. Procedures and key reference documents, such as Intel® Omni-Path user guides and installation guides are provided to clarify the process. Additional commands and BKM s are defined to facilitate the installation process and troubleshooting.

Intel recommends that you use the Intel® Omni-Path FastFabric (FF) Text-based User Interface (TUI) as the initial tool suite for installation, configuration, and validation of the fabric. This tool includes a set of automated features that are specifically used for standalone host, Ethernet*, and Intel® Omni-Path Fabric connectivity validation.

This document includes recommendations for processes and procedures that complement the FF tools to reduce the time required to install and configure the fabric.

You should check applicable release notes and technical advisories for key information that could influence installation steps outlined in this document.

Before You Begin

- Familiarize yourself with the available documentation by referring to [Intel® Omni-Path Documentation Library](#) on page 8 in this document.
- Checklists that contain an overview of the required steps are located in the *Intel® Omni-Path Fabric Software Installation Guide, Checklists* section.
- Download software from: <http://www.intel.com/omnipath/Downloads>
 - Fabric Host Software: IFS package for management nodes.
 - Fabric Host Software: Basic package for compute nodes.
 - Switch Firmware *.emfw: for externally-managed switches.
 - Switch Firmware *.spkg: for managed switches.
 - Intel® Omni-Path Fabric Suite Fabric Manager GUI software.
 - Open Fabrics Interfaces (OFI) Software for Intel® Omni-Path.
- Make sure you have access to OS packages and some extended packages that are prerequisites for installing the Intel Fabric Suite software, as described in the *Intel® Omni-Path Fabric Software Installation Guide*.
- Intel recommends that you develop a cable map topology file to validate the physical installation of the cables and insure no accidental hot-spots are created through incorrect cabling. It is also a beneficial extension to any datacenter and rack layout diagrams developed for an installation. Refer to the instructions described in [Generate Cable Map Topology Files](#) on page 19 in this document.

This document assumes the following:

- Operating System (OS) Software: RHEL* 7.2 or later. See the *Intel® Omni-Path Fabric Software Release Notes* for the complete list of supported OSes.



- Single Management Node (with Fabric Manager running) configured with the Intel® Omni-Path Fabric Suite Software, also known as IntelOPA-IFS.
- Intel® Omni-Path Fabric Manager enabled on management nodes.
- Compute Nodes configured with the Intel® Omni-Path Fabric Host Software, also known as IntelOPA-Basic.
- Password-less access enabled for all hosts and switches.

Note: Before you run top500 HPL (High Performance Linpack) runs or customer acceptance tests, Intel recommends that you follow all steps outlined in this guide.



2.0 Installation Prerequisites

The recommended fabric installation prerequisites are defined in the *Intel® Omni-Path Fabric Software Installation Guide*, Installation Prerequisites section.

The RPMs required for the operating system you are using are defined in the *Intel® Omni-Path Fabric Software Installation Guide*, OS RPMs Installation Prerequisites section.

Complete the following steps before starting software installation:

1. Install HFIs in servers, install externally-managed and managed switches in racks, and connect cabling.

Details: *Intel® Omni-Path Fabric Switches Hardware Installation Guide* and *Intel® Omni-Path Host Fabric Interface Installation Guide*.

All standard EMI protection should be adhered to whenever working with computer equipment. The same is true for switch and HFI installation.

HFIs should be installed in PCIe* x16 slots of the server. The same PCIe* slot should be used on all similar servers if possible to insure homogeneity in the cluster.

Switches should be installed based on switch guidelines and pre-designed rack placement.

Care must be taken when connecting the cables, please adhere to bend radius requirements and use cable trays for management. Never use zip-ties to cinch the cables together, use soft fabric Velcro* ties instead.

2. Power up all fabric hardware and verify LEDs operate as expected.

Details: *Intel® Omni-Path Fabric Switches Hardware Installation Guide* and *Intel® Omni-Path Host Fabric Interface Installation Guide*.

2.1 Configure BIOS Settings

Intel recommends that you use UEFI BIOS. For optimal performance, refer to a recommended BIOS configuration in the *Intel® Omni-Path Fabric Performance Tuning User Guide*, BIOS Settings sections for the following processors:

- Intel® Xeon® Processor E5 v3 Family and Intel® Xeon® Processor E5 v4 Family
- Intel® Xeon® Scalable Processors
- Intel® Xeon Phi™ x200 Product Family (codenamed Knights Landing)

Note: For Intel® Xeon Phi™ x200 Product Family, set the Snoop Holdoff Count to 9 as recommended in the *Intel® Omni-Path Fabric Performance Tuning User Guide*.

2.2 Configure OS Settings

Before you install Intel® Omni-Path software, perform the following tasks:



- Confirm Operating System (OS) versions match the versions listed in the *Intel® Omni-Path Fabric Software Release Notes*.
- Install OS RPM prerequisites listed in the *Intel® Omni-Path Fabric Software Installation Guide*, OS RPMs Installation Prerequisites section.
- Configure OS settings for optimal performance as described in the *Intel® Omni-Path Fabric Performance Tuning User Guide*, Linux* Settings section.



3.0 TCP/IP Host Name Resolution

For details on resolving TCP/IP Host Names, see the *Intel® Omni-Path Fabric Software Installation Guide*, Installation Prerequisites section. The following notes provide an example of the contents of the `/etc/hosts` file.

Create a `/etc/hosts` file before starting Intel® Omni-Path software installation to simplify the process. In a typical installation, the server and switch names follow a local convention to indicate physical location or purpose of the node.

- If using `/etc/hosts`, update the `/etc/hosts` file on the management node (the node with IFS installed) and copy to all hosts.
- If using DNS, all Management Network and IPoIB hostnames must be added to DNS `/etc/resolv.conf` and configured on the management node.
- The `/etc/hosts` file should contain:
 - Local host, required for subsequent single host verification using FastFabric TUI
 - Ethernet and IPoIB addresses and names for all hosts
 - Ethernet addresses and names of switches
 - Ethernet addresses of IPMI or remote management modules
 - Ethernet addresses of power domain

An example of these recommendations follows:

```
# /etc/hosts example
# localhost (required)
127.0.0.1 localhost localhost.localdomain localhost4 localhost4.localdomain
# Ethernet Addresses of hosts
10.128.196.14    node1
10.128.196.15    node2
10.128.196.16    node3
#IPoIB Address of hosts should be outside Ethernet network
10.128.200.14    node1-opa
10.128.200.15    node2-opa
10.128.200.16    node3-opa
#RMM IP Addresses
10.127.240.121   node1-rmm
10.127.240.122   node2-rmm
# Chassis IP Address
10.128.198.250   opaedge1
10.128.198.249   opaedge2
# OPA director switch IP Address
10.128.198.251   opadirector1
10.128.198.252   opadirector2
```

Other files that may need adjustment according to specific site requirements include:

`/etc/hostname`, `/etc/resolv.conf`, `/etc/network`, and

`/etc/network-scripts/ifcfg-enp5s0f0`



4.0 Install Intel® Omni-Path Software

You should configure at least one node to run the Intel® Omni-Path Management Software including Fabric Manager (FM). This node is used to configure and validate all of the other hosts, switches, and chassis fabric devices. You must install the Intel® Omni-Path Fabric Suite software on this node.

Overview

Driver installation is a two-step process. First, install the IFS package on your management servers, then use FastFabric to install the drivers on the rest of the systems in the cluster in parallel.

Note: If you use an existing cluster manager, installing the drivers to the image being installed on the compute nodes instead of using FastFabric is another option for quick driver installation.

- Install IntelOPA-IFS on management node(s) usually designated to run Subnet Manager (SM) and FastFabric Tools (including MPI applications) by changing the directory to `/IntelOPA-IFS.DISTRO.VERSION` and using the `./INSTALL` command.
 - Intel recommends that you use the Intel® Omni-Path FastFabric (FF) Text-based User Interface (TUI) as the tool suite for installation, configuration, and validation of the fabric. See *Install Using the TUI Menus* in the *Intel® Omni-Path Fabric Software Installation Guide*.
 - Alternatively, CLI commands can be used. See *Install Using CLI Commands* in the *Intel® Omni-Path Fabric Software Installation Guide*.

Examples include: `./INSTALL -n` or `./INSTALL -a -G`

- Intel recommends that you enable servers with IPMI interfaces to support ACPI or equivalent remote power management and reset control via an Ethernet network.
- Apply Technical Advisories as needed.
- Install NVIDIA* software (optional)

GPUDirect* RDMA is an NVIDIA* technology that allows third party network adapters to directly read and write to CUDA* host and device memory to enhance performance for latency and bandwidth. Intel provides CUDA* and GPUDirect* RDMA enhanced versions in both the Omni-Path HFI driver and PSM2 protocol libraries. Intel also provides a custom CUDA* enabled version of OpenMPI for PSM2.

For installation details: see the *Intel® Omni-Path Fabric Software Installation Guide*.

For usage details: see the *Intel® Omni-Path Fabric Host Software User Guide* and the *Intel® Performance Scaled Messaging 2 (PSM2) Programmer's Guide*.

References

The following document and sections describe the install procedures:



- *Intel® Omni-Path Fabric Software Installation Guide*, Download and Extract Installation Packages section
- *Intel® Omni-Path Fabric Software Installation Guide*, Install the Intel® Omni-Path Fabric Software section

Verify HFI speed and bus width using lspci

After the IFS installation, verify the Intel® OP HFI card is configured and visible to the host OS as Gen3 x16 slot speed (values are in **bold** text):

```
lspci -d 8086:24f0 -vv |grep Width
LnkCap: Port #0, Speed 8GT/s, Width x16, ASPM L1, Exit Latency L0s
<4us, L1 <64us
LnkSta: Speed 8GT/s, Width x16, TrErr- Train- SlotClk+ DLActive- BWMgmt- ABWMgmt-
```

4.1 Perform Initial Fabric Verification

Perform the following steps:

- Verify the port state of host is Active by running `opainfo`. If the command fails or returns other port state than Active, verify that the SM is running using `systemctl status opafm`.
- Use `opacmdall` or `pdsh` to run `opainfo` on all nodes in the fabric.
- Verify the OPA software version on all nodes using `opaconfig -V`. All nodes should be running the same version.
- Verify all nodes, switches, SM, and ISLs are up using `opafabricinfo` as shown in the following example.

```
opafabricinfo
Fabric 0:0 Information:
SM: node1 hfil_0 Guid: 0x001175010165b116 State: Master
Number of HFIs: 126
Number of Switches: 9
Number of Links: 252
Number of HFI Links: 126          (Internal: 0   External: 126)
Number of ISLs: 126              (Internal: 0   External: 126)
Number of Degraded Links: 0      (HFI Links: 0   ISLs: 0)
Number of Omitted Links: 0      (HFI Links: 0   ISLs: 0)
```

- Review the number of HFIs, number of switches, and external ISLs and confirm that they match the fabric design. The number of HFIs and external ISLs provide a fabric-blocking factor. If there are any degraded links, further troubleshooting is required.

4.2 Edit Hosts and Allhosts Files

Edit the following files, which are used by the `opafastfabric.conf` file.

- Edit `/etc/opa/hosts`
This file contains all hosts except the management node running IFS.
- Edit `/etc/opa/allhosts`
This file contains the statement `include /etc/opa/hosts`. Edit the file to add the node(s) running IFS.



5.0 Generate Cable Map Topology Files

For complete details on `topology.xlsx`, see the *Intel® Omni-Path Fabric Suite FastFabric User Guide*, `topology.xlsx` Overview section.

- `topology.xlsx` is a spreadsheet with 3 tabs:
Note: You should **not** modify tab 2 and 3.
 - Tab 1 Fabric is for the end user to define EXTERNAL links.
 - Tab 2 `swd06` contains the internal links for an Intel® OP Edge Switch 100 Series.
 - Tab 3 `swd24` contains the internal links for an Intel® OP Director Class Switch 100 Series.
- `README.topology` and `README.xlat_topology` describe best practices for editing the `topology.xlsx` file.

For descriptions of other sample files provided in the package, see the *Intel® Omni-Path Fabric Suite FastFabric User Guide*.

5.1 Generate Cable Map Topology Files

Defining Type in the Topology Spreadsheet

All host nodes should be defined Type = FI in column F of the spreadsheet. All Edge switches should be defined as Type = SW in column L (destination from host to Edge) and column F (source for Edge to core that is also Edge switch). The following example shows links between host and Edge switch.

```
R19 opahost1 1 FI R19 opaedge1 13 SW opahost1_opaelp13 1m Cable CU
```

All links between Edge switch to core that is also an Edge switch should be defined Type = SW as shown in the following example:

```
row1 rack01 opaedge1 1 SW row1 rack04 opaedgecore1 2 SW opaelp1_opac1p2 5M Cable Fiber
```

All Director switches should be defined as Type = CL in column L (destination from Edge switch to Director switch). Column J (Name-2) should have the destination leaf and column K should have the port number on that leaf. The following example shows a link between an Edge switch to core that is a Director switch.

```
R19 opaedge1 5 SW R72 opadirector1 01 L105B 11 CL opaelp5opad1L105Bp11 30m Fiber
```



All 24-leaf chassis Director switches should be defined as shown in the following example:

```
Core Name:opadirector1 Core Group:row1 Core Rack:rack72 Core Size:1152 Core Full:0
```

Set Core Full to 0 if the Director switch is not fully populated with all the leafs and spines. If it is fully populated, set Core Full to 1.

For complete details on `topology.xlsx`, see the *Intel® Omni-Path Fabric Suite FastFabric User Guide*.

Creating the Topology File

1. Copy and save the `/usr/share/opa/samples/topology.xlsx` file from the Fabric Manager node to your local PC for editing in Microsoft® Excel.
2. Edit tab 1 in the spreadsheet to reflect your specific installation details as described previously. Save tab 1 as `<topologyfile>.csv` and copy this `.csv` file back to the Fabric Manager node.

Note: In release 10.3 and later, the `topology.csv` cable label field can be up to 57 characters.

3. Generate the topology file in `.xml` format using the following command and the `topology.csv` file as the source:

```
# opaxlattopology <topologyfile>.csv <topologyfile>.xml
```

If there are Director switches defined in the `.csv` file, then `opaxlattopology` includes all the ISL (internal chassis links between leafs and spines) in the `.xml` file.



6.0 Configure FastFabric

The list of configuration files that are used by FastFabric are contained in the *Intel® Omni-Path Fabric Suite FastFabric User Guide*, Configuration Files for FastFabric section.

The `opafastfabric.conf` file provides default settings for most of the FastFabric command line options.

6.1 Format for IPoIB Host Names

By default, FastFabric uses the suffix OPA for the IPoIB host name. You can change this to a prefix and you can also change from opa to another convention such as ib, as the customer requires in `/etc/opa/opafastfabric.conf`.

The following examples show how to change opa to ib as a prefix or suffix.

For suffix:

```
export FF_IPoIB_SUFFIX=${FF_IPoIB_SUFFIX:--opa} to export FF_IPoIB_SUFFIX=${FF_IPoIB_SUFFIX:--ib}
```

For prefix:

```
export FF_IPoIB_PREFIX=${FF_IPoIB_PREFIX:-opa-} to export FF_IPoIB_PREFIX=${FF_IPoIB_PREFIX:-ib-}
```

6.2 Specify Test Areas for opaallanalysis

By default, `opaallanalysis` includes the fabric and chassis. These can be modified to include host SM, embedded SM, and externally-managed switches in `/etc/opa/opafastfabric.conf` as follows:

```
# pick appropriate type of SM to analyze
#export FF_ALL_ANALYSIS=${FF_ALL_ANALYSIS:-fabric chassis hostsm esm}
export FF_ALL_ANALYSIS=${FF_ALL_ANALYSIS:-fabric chassis hostsm}
```

6.3 Location of mpi_apps Directory

By default, `opafastfabric` uses `mpi_apps` located in `/usr/src/opa/mpi_apps`. If a different path is set up for `mpi_apps`, then modify the following in `/etc/opa/opafastfabric.conf`:

```
export FF_MPI_APPS_DIR=${FF_MPI_APPS_DIR:-/usr/src/opa/mpi_apps}
```



Note: The default source location for `mpi_apps` is `/usr/src/opa/mpi_apps`, but the default compilation directory for `mpi_apps` is `$HOME/mpi_apps`.



7.0 Configure Managed Intel® Omni-Path Edge Switches

For a complete description of the configuration process, refer to the *Intel® Omni-Path Fabric Software Installation Guide*, Configure the Chassis section.

The following steps provide a summary:

1. Download and install the driver file CDM v2.12.00 WHQL Certified.exe from:
<http://www.ftdichip.com/Drivers/VCP.htm>
2. Set up USB serial port terminal emulator using the following serial options:
 - Speed: 115200
 - Data Bits: 8
 - Stop Bits: 1
 - Parity: None
 - Flow Control: None
3. Set up the switch TCP/IP address, gateway, netmask, and other options using a terminal emulator.
 - a. Set the chassis IP address:

```
setChassisIpAddr -h ipaddress -m netMask
```

where *ipaddress* is the new IP address in dotted decimal format (xxx.xxx.xxx.xxx), and *netMask* is the new subnet mask in dotted decimal format.

- b. Change the chassis default gateway IP address:

```
setDefaultRoute -h ipaddress
```

where *ipaddress* is the new default gateway IP address in dotted decimal format.

The changes are effective immediately.

For details, refer to the *Intel® Omni-Path Fabric Switches Hardware Installation Guide*.

4. Edit the chassis file using the command:

```
opagenchassis >> /etc/opa/chassis
```

The chassis file contains the node name of managed switches corresponding to TCP/IP addresses as defined in the `/etc/hosts` file.

5. Run the `opafastfabric` TUI (Text-based User Interface).
6. Select 1) Chassis Setup/Admin.



7. Select items 0–6 and press **P** to **Perform**.
 - a. Item 0: Edit Config and Select/Edit Chassis File.
 - i. Skip `opafastfabric.conf`, no changes needed.
 - ii. Skip `ports`, no changes needed.
 - iii. For `chassis` file, in the editor, review the list of chassis selected. The setup of this file should have occurred above when setting up the Management Node by editing `/etc/opa/chassis` with the name corresponding to the Ethernet IP address of the chassis.
 - b. Item 1: Verify Chassis via Ethernet Ping, should pass without error.
 - c. Item 2: Update Chassis Firmware.
Specify the location for the firmware file to use.
 - d. Item 3: Set Up Chassis Basic Configuration.
Provide answers as follows:
 - i. Password: - Press **Enter** (no password).
 - ii. Syslog (*y*)
 1. Syslog server (*n*)
 2. TCP/UDP port number (*n*) - Use default.
 3. Syslog facility (*n*) - Use default.
 - iii. NTP (*n*) - Customer to assign
 - iv. Timezone and DST (*y*)
Use local timezone of server (*y*).
 - v. Do you wish to configure OPA Node Desc to match Ethernet chassis name? (*y*) - Enter *y*.
 - vi. Do you wish to configure the Link CRC Mode? (*n*)
 - e. Item 4: Set Up Password-Less SSH/SCP.
 - f. Item 5: Reboot Chassis should pass without error.
 - g. Item 6: Get Basic Chassis Configuration.
Expected Summary output at end is shown below. Note that count should match the number of Edge switches.

```
Edgeswitch1:
Firmware Active       : 10.x.x.x.x
Firmware Primary      : 10.x.x.x.x
Syslog Configuration  : Syslog host set to: 0.0.0.0 port 514 facility
22
NTP                   : Configured to use the local clock
Time Zone             : Current time zone offset is: -5
LinkWidth Support     : 4X
Node Description      : switch1
Link CRC Mode         : 48b_or_14b_or_16b
```

To review the results, use an editor to view the files:

`/root/test.res` and `/root/test.log`



For more information, refer to the *Intel® Omni-Path Fabric Switches Hardware Installation Guide* and *Intel® Omni-Path Fabric Switches Release Notes*.



8.0 Configure Intel® Omni-Path Director Class Switch 100 Series

Most Intel® OP Director switches are supplied with two Management Modules (MMs) for redundancy. In addition, Intel® OP Director switches have the following additional features:

- The switch has two Ethernet ports (one for each MM) and requires two Ethernet cables.
- The switch requires three IP addresses: one for each MM and one for the chassis, which is bound to the MM that is currently Master.
- It is useful to understand all reboot modes: `reboot all|-s|-m [slot #]` and how that causes failover.
- Default IP addresses of the Management Modules are:
Chassis IP address: 192.168.100.9
Management Module M201: 192.168.100.10
Management Module M202: 192.168.100.11

The chassis file, located in `/etc/opa/chassis`, contains the node name of Intel® OP Director switches corresponding to TCP/IP addresses as defined in the `/etc/hosts` file. The chassis IP address is configured using the procedure for configuring managed switches, as described in [Configure Managed Intel® Omni-Path Edge Switches](#).

The MM IP addresses must be configured using a serial connection as described in the following procedure:

1. Ensure that the module is connected to a COM port on a serial terminal device through the USB port.
2. Get to a `[boot]` prompt by following either step a or b:
 - a. If the management module is running and displays `->` prompt, type the following command at the console: `reboot now` and press **ENTER**.
 - b. If the management module is not running, power on the switch.
3. When the system displays `image1`, press the spacebar to interrupt the autoloading sequence before the counter expires (within 5 seconds).
4. At the prompt, enter the command: `moduleip <ip_address>`
The module reboots itself within 5 seconds and comes back with the new IP assigned to it. This module becomes the slave and the other MM becomes the master.

Repeat these steps for the second management module.

For more information, refer to the *Intel® Omni-Path Fabric Switches Hardware Installation Guide* and *Intel® Omni-Path Fabric Switches Release Notes*.



9.0 Configure Externally-Managed Intel® Omni-Path Edge Switches

For a complete description of the install process, refer to the *Intel® Omni-Path Fabric Software Installation Guide*, Configure the Externally-Managed Switches section.

The 100SWE48QF Edge switches do not have an Ethernet* interface. Setup of these switches is performed using FastFabric via in-band commands.

Preferred approach:

1. Edit the switches file for externally-managed switches using the command:
`opagenswitches >> /etc/opa/switches`

The switches file contains a list of all the externally-managed switches in the fabric.

Edit the switches file to replace the default switch name with the actual name that corresponds to the GUID for each switch. For example:

Default: `0x00117501026a5683:0:0,OmniPth00117501ff6a5602,2`

Edited: `0x00117501026a5683:0:0,opaextmanagededge1,2`

2. Run `opafastfabric`.
3. Select 2) Externally Managed Switch Setup/Admin.
4. Select items 0–9 and press P to **Perform**.
 - a. Item 0: Edit Config and Select/Edit Switch File
 - i. Skip `opafastfabric.conf`. No changes needed.
 - ii. Skip `ports`. No changes needed.
 - iii. Edit the file `/etc/opa/switches` and review the list of chassis selected. The switches file specifies:
 - switches by node GUID
 - (optional) hfi:port
 - (optional) Node Description (nodename) to be assigned to the switch
 - (optional) distance value indicating the relative distance from the FastFabric node for each switch

The following snippet shows the switches file format and an example:

```
nodeguid:hfi:port,nodename,distance
0x00117501026a5683:0:0,opaextmanagededge1,2
```

- b. Item 1: Generate or Update Switch File.
 - i. Regenerate - Answer `n` if it was generated in step 1. Answer `y` if this is the first time or additional externally-managed switches have been added or replaced.



- ii. Update switch names - Answer `y`. Note that this step may take a few minutes.
- c. Item 2, 3:
Should pass without error.
- d. Item 4: Specify the location for the FW file (`.emfw`) to use.
- e. Item 5: Set up switch basic configuration and set the node description.

```
Performing Switch Admin: Setup Switch basic configuration
Executing: /usr/sbin/opa-switch-admin -L
/etc/opa/switches configure
Do you wish to configure the switch Link Width Options? [n]:
Do you wish to configure the switch Node Description as it is set in the
switches file? [n]: y
Do you wish to configure the switch FM Enabled option? [n]: Do you wish
to configure the switch Link CRC Mode? [n]: Executing configure Test
Suite (configure) Fri Jan 15 11:11:12 EST 2016 ...
Executing TEST SUITE configure CASE (configure.
0x00117501026a5683:0:0,OmniPth00117501ff6a5602.i2c
.extmgd.switchconfigure) configure switch
0x00117501026a5683:0:0,OmniPth00117501ff6a5602 ...
TEST SUITE configure CASE (configure.
0x00117501026a5683:0:0,OmniPth00117501ff6a5602.i2c
.extmgd.switchconfigure) configure switch
0x00117501026a5683:0:0,OmniPth00117501ff6a5602 PASSED
TEST SUITE configure: 1 Cases; 1 PASSED
```

- f. Item 6: Reboot.
- g. Item 7: Review results for redundant power and FAN status.
Expected summary output at end should be similar to the following (count should match number of externally-managed Edge switches):

```
0x00117501026a5683:0:0,opaextmanagededge1:
F/W ver:10.x.x.x.x H/W ver:003-01 H/W pt num:H89344-003-
01 Fan status:Normal/Normal/Normal/Normal/Normal PS1
Status:ONLINE PS2 Status: ONLINE Temperature
status:LTC2974:33C/MAX_QSFP:40C/PRR_ASIC:40C
```

Any non-redundant or failed fans or power supplies found during this step are also reported in `/root/punchlist.csv`.

- h. Item 8: Get Basic Switch Configuration.
Expected summary output at end should be similar to the following (count should match number of externally-managed Edge switches):

```
Link Width           : 1,2,3,4
Link Speed           : 25Gb
FM Enabled           : No
Link CRC Mode        : None
vCU                  : 0
External Loopback Allowed : Yes
Node Description      : Edgeswitch1
```

- i. Item 9: Save the `test.res` output for future reference.

To review results, view the `/root/test.res` and `/root/test.log` files.

For more information, refer to the *Intel® Omni-Path Fabric Switches Hardware Installation Guide* and *Intel® Omni-Path Fabric Switches Release Notes*.



10.0 Configure Host Setup

Perform the following steps:

1. Make sure all hosts are booted; this is required to identify switch names. If hosts are not available, you can perform all configuration steps except setting the switch names.
2. Run `opafastfabric`.
3. Select 3) Host Setup.
4. Select items 0–4 and press **P** to **Perform**.
5. Select item 5 and press **P** to **Perform**.

This installs IntelOPA-Basic on all compute nodes defined in `/etc/opa/hosts`. Be sure to exclude management node(s) with IFS installed and the node where you are running `opafastfabric`.

- a. Provide the path to `IntelOPA-Basic.DISTRO.VERSION.tgz` when prompted.
 - b. Enter directory to get `IntelOPA-Basic.DISTRO.VERSION.tgz` from (or none) : `/root`.
6. Select item 6 and press **P** to **Perform**. This performs the IPoIB ping test.
 7. Select item 7 and press **P** to **Perform**. This Builds Test Apps and Copy to Hosts.
 - a. Choose an MPI when prompted: Please Select MPI Directory
 - b. Select an MPI with `-hfi` extension, so it will build with PSM2. For example: `/usr/mpi/gcc/openmpi-x.x.x-hfi`.
 - c. When prompted to build base sample applications, select `yes`.

For more information, refer to the *Intel® Omni-Path Fabric Software Installation Guide* and *Intel® Omni-Path Fabric Software Release Notes*.

11.0 Verify Cable Map Topology

This section describes how to use the fabric `topology.xml` file created in [Generate Cable Map Topology Files](#) to verify that fabric topology (cabling) is consistent with the cable map.

The command `opareport -o verify* -T <topologyfilename>.xml` compares the live fabric interconnect against the topology file created based on the cable map. These commands test links, switches, and SM topology. If successful, the output reports a total of 0 Incorrect Links found, 0 Missing, 0 Unexpected, 0 Misconnected, 0 Duplicate, and 0 Different.

```
# opareport -o verifyfis -T <topologyfilename>.xml
# opareport -o verifyextlinks -T <topologyfilename>.xml
# opareport -o verifyall -T <topologyfilename>.xml
```

In most cases, links reported with errors are either due to incorrect cabling to the wrong port or the `topology.csv` file has incorrect source and port destinations. Verify the physical interconnect against the cable map using `opaextractsellinks` as in the following examples:

- List all the links in the fabric: `opaextractsellinks`
- List all the links to a switch named `OmniPth00117501ffffffff`:
`opaextractsellinks -F "node:OmniPth00117501ffffffff"`
- List all the connections to end-nodes: `opaextractsellinks -F "nodetype:FI"`
- List all the links on the second HFI's fabric of a multi-plane fabric:
`opaextractsellinks -h 2`

After all topology issues have been resolved, copy the `topologyfile.xml` from the local working directory to `cat /etc/opa/topology.0\0.xml`.

Refer to the *Intel® Omni-Path Fabric Suite FastFabric User Guide* for more information about using `opareport` in general, and using `opareport` for Advanced Topology Verification.



12.0 Verify Server and Fabric

Validation of servers and the fabric is initiated from the Management Node using the FastFabric TUI using `opafastfabric` on hosts defined in `/etc/opa/allhosts`.

Perform the following steps:

1. Choose item 4) Host Verification/Admin and run through all steps.
2. Perform 3) Perform Single Host Verification.
When prompted "Would you like to specify tests to run? [n]:" enter `y` for HPL test.
When prompted "View Load on hosts prior to verification? [y]:" enter `y`. This option checks CPU load by running `/usr/sbin/opacheckload -f /etc/opa/allhosts`.
Edit `hostverify.res` for results.
3. Perform 4) Verify OPA Fabric Status and Topology. This option goes through a fabric error and topology verification.
Choose the default for all prompts.
Edit `/root/linkanalysis.res` to view results.
4. Perform 6) Verify Hosts Ping via IPoIB. This option pings all IPoIB interfaces.
5. Perform 8) Check MPI Performance. This option tests Latency and Bandwidth deviation between all hosts.
Choose defaults for all prompts.
Edit `/root/test.log` for results.

For more information, refer to the *Intel® Omni-Path Fabric Software Installation Guide*.

Note: A punchlist file is generated during execution of the FastFabric TUI and CLI commands, which can be used to track issues identified by the Intel® OPA tools. The punchlist file is located in `$FF_RESULT_DIR /punchlist.csv`, typically `/root/punchlist.csv`.

Two additional files, `/root/test.res` and `/root/test.log`, are created during OPA test commands and are useful for tracking test failures and issues.



13.0 Best Known Methods (BKMs) for Site Installation

This section contains commands useful for configuring and debugging issues during fabric installation.

13.1 Enable Intel® Omni-Path Fabric Manager GUI

By default, the Intel® Omni-Path Fabric Suite Fabric Manager GUI is disabled after installation of the IFS software. To quickly enable for early debug, use the following steps. For complete details, refer to the *Intel® Omni-Path Fabric Suite Fabric Manager GUI User Guide*.

Note: This method bypasses the SSH key authorization and is not intended for end customer installs.

1. Edit `/etc/opa-fm/opafm.xml` file on the Management Node. Make the two changes shown in **bold** for `SslSecurityEnabled` and default FE startup:

```
<SslSecurityEnabled>0</SslSecurityEnabled>

<!-- Common FE (Fabric Executive) attributes -->
<Fe>
<!-- The FE is required by the Intel Omni-Path FM GUI. -->
<!-- To enable the FE, configure the SslSecurity parameters in this file -->
<!-- as desired. -->
<!-- For Host FM then set Start to 1. -->
<!-- For Embedded FM the Start parameter in this file is not used; -->
<!-- enable the FE via the smConfig and smPmStart chassis CLI commands. -->
<Start>1</Start> <!-- default FE startup for all instances -->
<!-- Overrides of the Common.Shared parameters if desired -->
<!-- <SyslogFacility>Local6</SyslogFacility> -->
```

2. Restart the Fabric Manager to enable the changes and start the FE process required by the Fabric Manager GUI.

```
# systemctl restart opafm
```

3. Download and install the Fabric Manager GUI application to a Windows* PC or Linux* system.
4. Start the Fabric Manager GUI application.
5. Open the **Configuration** tab and enter the hostname or IP address of the Management Node running the Fabric Manager in your system into the FE Connection.
6. Uncheck the Secure tab.
7. Select **Apply** to run the connection test and then **Run** to start the Fabric Manager GUI application.



Note: The Fabric Manager GUI does not operate through network proxies. Network firewall access may also need to be disabled. For a quick go/no-go verification, complete the connection test in the configuration tab as previously described.

13.2 Review Server and Fabric Verification Test Results

During fabric validation, unexpected loads on Host CPUs may result in inconsistent performance results. As a debug step, isolate the issue using the following:

Use the OPA tool to verify CPU host load. By default, it captures the top ten most heavily loaded hosts.

```
# /usr/sbin/opacheckload -f /etc/opa/allhosts
```

After the high load hosts have been identified, the next step is to root cause the issues.

Perform the following steps:

1. Check for HFI PCIe width or speed issues.

Are HFI cards operating in a degraded mode, narrow width, or less than PCIe Gen3? Use `lspci` or `opahfirev` to verify the PCIe operating speed and bus width:

- `lspci`: [Verify HFI speed and bus width using lspci](#)
- `opahfirev`: [Decode the Physical Configuration of an HFI](#)

Possible sources for narrow PCIe width:

- a. Be aware that OPA does support different width PCIe cards, including dual HFI cards using two x8 slices of a x16 physical connector. `opahfirev` is very useful for detecting this configuration.
 - b. HFI Card partial insertion into x16 slots. Initially this appears to be a narrow width issue but re-inserting the card often resolves the issue. This may occur after a server is shipped. This step has resolved most width issues.
 - c. Server physical configuration: Many servers support different PCIe logical widths based on riser card configuration. The slot may be physically x16 but internally limited to x8. Check other servers of the same configuration in the fabric. Check the server configuration. *This is also a common issue.*
 - d. Swap the HFI to another server to determine if the problem follows the card or the server.
2. Use the Linux* `top` command to identify the key CPU load processes:

```
# top
```

`opatop` may be useful for checking for loads that vary over time. Use the `r` (rev), `f` (forward), and `L` (live) options to look through PM snapshots of system activity. This is also helpful for monitoring application startup versus run time loads. The



PM captures high resolution statistics, with very low system overhead, over periods up to two days. The tools that harvest the PM stats are `opatop` and the FM GUI.

```
# opatop
```

3. Check for high CPU percent processes.

Examples of some issues:

ksoftirqd process - known issue in RHEL* 7.1, the workaround is to reboot the individual server. The fix is to update to a newer release.

Screen savers - when a Linux* GUI is enabled on hosts, the screen that runs when the user interface is idle may have a high CPU load.

Test applications - look for MPI jobs or similar applications running in the background. This is a common issue particularly in a shared fabric bring-up environment. Use `kill -p process` to stop orphan applications or reboot the server to debug the issue.

4. Review the following section of this document to isolate nodes with different or incorrect settings.

- [Configure BIOS Settings](#) on page 14

13.3 Debug Intel® Omni-Path Physical Link Issues

After you have run the FastFabric tool suite and identified issues with links, then it is useful to start root-causing the issues. This section focuses on Intel® Omni-Path Fabric physical links and not PCIe bus link issues.

OPA reporting tools are robust, but it can be confusing for new users to understand the difference between error counters and actual failures.

From an installation perspective, it is important to watch for physical issues with cabling, both copper and optical. In general, bend radius, cable insertion issues, and physical compression or damage to cables can result in transmission issues. OPA recovers from many issues transparently. This section helps root-cause solid failures as well as marginal links. Most often the issue is resolved simply by re-installing a cable and verifying that it clicks into the connector socket on the HFI or switch.

View the QSFP/cable details of a specific switch port using the command:

```
opasmaquery -o cableinfo -d 10 -l <lid> -m <switch portnumber>
```

To debug a particular switch, a useful technique is to get a snapshot of it, using the command:

```
opareport -o snapshot -F portguid:0x001175010265bb1d
```

13.3.1 OPA Link Transition Flow

To debug link issues, it is helpful to understand the four key link states, starting from Offline and running properly in the final Active state.



Note: The Fabric Manager, `opafm`, must be running to transition physical links from the Init state to the Active state. If you subsequently stop the Fabric Manager when a link is in the Active state, the link remains active. You can safely make changes to the `opafm.xml` file for the Fabric Manager and restart the service without dropping active links. As of the 10.0.0.696 software release, by default, the `opafm` service is not configured for autostart after IFS FULL installation.

PortState:

- Offline: link down. QSFP not present or not visible to the HFI driver.
- Polling: physical link training in progress. At this point you do not know whether the other end of the QSFP is connected to a working OPA device.
- Init: Link training has completed, both sides are present. Typically waiting for the Fabric Manager to enable the link.
- Active: Normal operating state of a fully functional link.

13.3.2 Verify the Fabric Manager is Running

From the Management Node, run the following command to report all HFIs and Switches.

```
# opafabricinfo
```

If it fails, try the following steps:

- Check status of the Fabric Manager process using the command:

```
# systemctl status opafm
```

- Restart the Fabric Manager using the command:

```
# systemctl start opafm
```

13.3.3 Check the State of All Links in the System

The `opaextractsellinks` command generates a CSV output representing the entire link state of the fabric.

```
# opaextractsellinks > link_status.csv
```

For links with errors, run the `opaextracterror` command.

```
# opaextracterror > link_status.csv
```

13.3.4 Check the State of HFI Links from a Server

If you are debugging server link issues, the `opainfo` command may be useful for a single server view.



opainfo captures a variety of data useful for debugging server related link issues. Multiple OPA commands can be used to extract individual data elements, however, this command is unique in the combination of data it provides.

- PortState: see [OPA Link Transition Flow](#) on page 34.
- LinkWidth: a fully functional link should indicate Act:4 and En:4.
- QSFP: Physical cable information for the QSFP, in this case a 5M Optical (AOC) Finisar cable.
- Link Quality: Range = 0 - 5 where 5 is Excellent.

```
# opainfo
hfil_0:1          PortGID:0xfe80000000000000:001175010165b19c
  PortState:      Active
  LinkSpeed       Act: 25Gb          En: 25Gb
  LinkWidth       Act: 4             En: 4
  LinkWidthDnGrd ActTx: 4 Rx: 4      En: 3,4
  LCRC            Act: 14-bit        En: 14-bit,16-bit,48-bit Mgmt: True
  LID: 0x00000001-0x00000001      SM LID: 0x00000002 SL: 0
  QSFP: PassiveCu, 1m FCI Electronics P/N 10131941-2010LF Rev 5
  Xmit Data:      22581581 MB Pkts: 5100825193
  Recv Data:      18725619 MB Pkts: 4024569756
  Link Quality: 5 (Excellent)
```

13.3.5 Link Width, Downgrades, and opafm.xml

By default, OPA links run in x4 link width mode. OPA has a highly robust link mechanism, as compared to InfiniBand*, and it allows links to run in reduced widths with no data loss.

Three things to know:

1. By default, the opafm.xml configuration file requires links to start up in x4 link width mode. This is configurable separately for HFI and ISL links using the **WidthPolicy** parameter.
2. Link downgrade ranges are also configurable in the opafm.xml file, using the **MaxDroppedLanes** parameter.
3. Default configuration example - A link that successfully starts up in x4 width and subsequently downgrades to x3 width continues to operate. If the link is restarted, by a server reboot, for example, and attempts to run by less than x4 width, then the link is disabled by the Fabric Manager and does not enter the Active state.

The opainfo command for HFIs is useful for checking the link width and link downgrade configuration on servers.

For a system view of all links that are running in less than x4 width mode, use the command:

```
# opareport -o errors -o slowlinks
```

13.3.6 How to Check Fabric Connectivity

For large fabrics, follow the flow described in [Generate Cable Map Topology Files](#) on page 19.



13.3.7 Physical Links Stability Test using opacabletest

Intel® Omni-Path Architecture uses a quality metric for reporting status (opainfo). The quality metric ranges from 5 (excellent) to 1 (poor). For a more quantitative metric, use `cabletest` to generate traffic from on the HFI and ISL links, and `opaextractperf` and `opaextracterrors` to harvest the data.

Before you begin:

- Clear error counters prior to test using `opareport -o none -clearall` and check the error counters after the test.
- Check to make sure there are no errors in fabric using: `opareport -o errors`
- Use `opatop` to monitor fabric utilization.

Detailed procedure:

1. Start and stop cable test on the management node either from the `opafastfabric` TUI or using CLI commands:

- a. `# opafastfabric`
- b. 4) Host Verification/Admin
- c. a) Start or Stop Bit Error Rate cable Test

Or to run manually, use the following tests for hosts, then ISLs. Test each one for a reasonable time, typically 5 - 15 minutes.

```
# /usr/sbin/opacabletest -A -n 3 -f '/etc/opa/allhosts' stop-fi stop-isl
# opareport -o none -clearall
# /usr/sbin/opacabletest -A -n 3 -f '/etc/opa/allhosts' start-fi
```

Run the previous command for 5 - 15 minutes for the hosts.

```
# /usr/sbin/opacabletest -A -n 3 -f '/etc/opa/allhosts' stop-fi start-isl
```

Run the previous command for 5 - 15 minutes for the ISLs.

```
# /usr/sbin/opacabletest -A -n 3 -f '/etc/opa/allhosts' stop-isl stop-fi
# opaextractperf > link_stability_perf.csv
# opaextracterrors > link_stability_counters.csv
```

Use `opatop` to view link utilization.

2. For large fabrics, check stability using a long run of `opacabletest` (typically 4-8 hours). Short runs of 10-15 minutes are fine for initial validation.

How to interpret the results:



The `opaextracterrors` command is a misnomer, it captures interesting statistics for evaluating links, but most of the content is not indicative of failures. The OPA fabric has robust end-to-end recovery mechanisms that handle issues.

Suggest looking specifically at the following columns:

- `LinkWidthDnGradeTxActive` - expect to see x4 Width
- `LinkWidthDnGradeRxActive` - expect to see x4 Width
- `LinkQualityIndicator` - 5 is excellent, 4 is acceptable, 3 is marginal and clearly an issue.
- `LinkDowned` - when an HFI is reset, the link down count increases, so rebooting a server results in small increments. If you see a link with significantly higher counts than its reboot expectations, then take a look at the server `/var/log/messages` file to determine whether the server is rebooting or the link is re-initializing.

For the other error counters, run a column sort and look for high error counts (greater than 100x) versus other links and take a look at the link types. Optical links have higher retry rates. This is not typically an issue unless they far exceed their peers.

The output is useful for verifying that every link is being tested. Unusual fabric `opaextractperf` topologies may result in non-optimum cabletest results. One workaround is to separately run `isl` and `fi` (HFI) link tests, then look at the total error results.

13.3.8 How to Debug and Fix Physical Link Issues

Check the topology before and after each of the debug steps using:

```
# opareport -o verifyall -T test_topology.xml
```

If the original issue was marginal operation rather than a hard failure, then re-run `cabletest` and analyze the `opaextracterrors` results to verify whether the issues were resolved.

At this point, you have a list of links with issues. Intel recommends the following approach for physical link resolution:

1. Unplug and re-insert each end of a physical cable. Check that the cable actually clicks into place. It may be useful to do this step separately for each end of the cable. Re-run `opacabletest` and verify whether the issue has been resolved or not.
Note: This step has resolved more link issues in fabric installs than all others.
2. Swap the questionable cable with a known good cable to isolate whether it is an HFI/Switch issue or cable issue.
3. If step 2 worked, then install the questionable cable into another location and verify whether it works.
4. If the issue is corrected, then the issue may be a mechanical latching issue on the HFI/Switch connector.
5. If the original issue was marginal operation rather than a hard failure, then re-run `opacabletest` and analyze the `opaextracterrors` results to verify whether the issues were resolved.
6. Re-run the physical links stability test using `opacabletest`.



13.3.9 Link Debug CLI Commands

- Identify fabric errors:

```
# opareport -o errors
```

- Identify slow links (< x4 width):

```
# opareport -o slowlinks
```

- To obtain lid of the switch, use `opaextractlids` for the links.
To obtain the port and links, use `opaextractsellinks` for the links.
- If a link is not coming up as Active, first bounce the link and check the link state:

```
# opaportconfig -l <lid> -m <port> bounce
```

- Get detailed link info for all nodes connected to an Edge switch or leaf and their neighbor:

```
# opareport -M -m -A -s -o comps -d 20 -F lid:<lid of edge switch>:node
```

- Find links that are not plugged in or not seen by the interface. Find all links stuck in the Offline state:

```
# opareport -A -m -F portphysstate:offline -o comps -d 5
```

- A link stuck in Polling may indicate that the other end of the cable is not inserted correctly. In this case, typically, one end is Polling and the other end is Offline.
- Find all links stuck in the Polling state:

```
# opareport -A -m -F portphysstate:polling -o comps -d 5
```

- Identify bad links:

```
# opaextractbadlinks
```

- As a debug step, temporarily disable all bad links and append `/etc/opa/disabled.0:0.csv` with a list of all bad links disabled.

```
# opaextractbadlinks | opadisableports
```

- To enable links previously disabled:

```
# cat /etc/opa/disabled.0:0.csv | opaenableports
```

- To bounce a link, simulating a cable pull and re-insert on a server. It may take up to 60 seconds for the port to re-enter the active state.

```
# opaportconfig bounce
```



- Check status using:

```
# opainfo
```

- `opaportconfig` and `opaportinfo` are key commands for port debugging. Run the commands with the `-help` option to see available parameters.
- To disable a set of links, extract them to a `csv` file using `opaextractsellinks`. In the following example, links are extracted to `linkstodisable.csv`. To disable a set of links, run:

```
opadisableports < linkstodisable.csv
```

By default, all disabled links are appended to the file `/etc/opa/disabled\:1\:1.csv`.

To enable the disabled ports, run:

```
opaenableports < /etc/opa/disabled\:1\:1.csv
```

After enabling the ports, the file `/etc/opa/disabled\:1\:1.csv` purges the links that are enabled.

Note: For each listed link, the switch port closer to this node is disabled.

Run `opaportinfo -l <lid of switch> -m <port number>`. Check the port state by running:

```
opaportinfo -l 3 -m 0x10
```

Note: Be sure to exclude the SM node on the Edge switch you are on and run `disableports` from the `linkstodisable` file to prevent cutting off this node from the fabric.

13.4 Use opatop for Bandwidth and Error Summary

Use the `opatop` Textual User Interface (TUI) to look at bandwidth and error summary of HFIs and switches.

This section provides a high-level overview of `opatop`.

- 1) selects HFIs and 2) selects SW.
- Intel recommends selecting 2) SWs. In this display, HFIs show up as Send/Rcv and ISLs show up as Int.
- On the Group Information screen:
 - Select (W) for Bandwidth.
 - Select (E) for Error summary.
- Use `u` to move to an upper level.
- Use `2)` to view SWs Bandwidth and error summary.

For details, see the *Intel® Omni-Path Fabric Suite FastFabric User Guide*, `opatop` Fabric Performance Monitor section.



13.5 Use the Beacon LED to Identify HFI and Switch Ports

The LED beaconing flash pattern can be turned ON/OFF with the `opaportconfig` command. This can be used to identify the HFI and switches/ports installed in racks that need attention.

For HFI:

```
opaportconfig -l 0x001 ledoff
Disabling Led at LID 0x00000001 Port 0 via local port 1 (0x0011750101671ed9)

opaportconfig -l 0x001 ledon
Enabling LED at LID 0x00000001 Port 0 via local port 1 (0x0011750101671ed9)
```

For Switch port:

```
opaportconfig -l 0x002 -m 40 ledon (where -m 40 is port number)
Enabling LED at LID 0x00000002 Port 40 via local port 1 (0x0011750101671ed9)

opaportconfig -l 0x002 -m 40 ledoff
Disabling Led at LID 0x00000002 Port 40 via local port 1 (0x0011750101671ed9)
```

13.6 Decode the Physical Configuration of an HFI

The `opahfirev` command provides a quick snapshot of an Intel® Omni-Path Host Fabric Interface (HFI), providing both PCIe status and physical configuration state, complementary to the `opainfo` command.

```
# opahfirev
#####
phsmpri07.ph.intel.com - HFI 0000:81:00.0
HFI: hfi1_0
Board: ChipABI 3.0, ChipRev 7.17, SW Compat 3
SN: 0x0063be82
Location:Discrete Socket:1 PCISlot:00 NUMANode:1 HFI0
Bus: Speed 8GT/s, Width x16
GUID: 0011:7501:0163:be82
SiRev: B1 (11)
TMM: 10.4.0.0.146
#####
```

Note the field for Thermal Monitoring Module (TMM) firmware version, an optional micro-controller for thermal monitoring on vendor-specific HFI adapters using the SMBus. For more information on the `opatmmtool`, see the *Intel® Omni-Path Fabric Host Software User Guide*.

- Check the current TMM firmware version using: `opatmmtool -fwversion`.
- Check the TMM firmware version in the `hfi1_smbus.fw` file using:

```
opatmmtool -f /lib/firmware/updates/hfi1_smbus.fw fileversion
```

- If the `fwversion` is less than `fileversion`, then update the TMM firmware version using:

```
opatmmtool -f /lib/firmware/updates/hfi1_smbus.fw update
```



- After the TMM is updated, restart the TMM using:

```
opatmmtool reboot
```

Note: Data traffic is not interrupted during a reboot of the TMM.

13.7 Program and Verify Option ROM EEPROM Device

This section describes how to program and verify the Option ROM EEPROM device on an Intel® Omni-Path Host Fabric Interface (HFI).

Before you Begin

- If you have an Intel® Omni-Path Host Fabric Interface, continue with this process.
- If you have an HFI from another manufacturer, contact the manufacturer's support team. Your HFI may require different modifications to the `hfil_platform.dat` file.

Overview

There are three files available for Option ROM EPROM partitions. These default files are packaged with Intel Fabric Suite (IFS) and Basic releases. See the *Intel® Omni-Path Fabric Software Release Notes* for the version provided in the release. The files are:

- HFI1 UEFI Option ROM: `HfiPcieGen3_x.x.x.x.x.efi`
- UEFI UNDI Loader: `HfiPcieGen3Loader_x.x.x.x.x.rom`
- HFI1 platform file: `hfil_platform.dat`

Note: The `hfil_platform.dat` file is for Intel® Omni-Path Host Fabric Interface cards.

To find the file locations, enter the following Linux* commands:

```
find / -name hfil_platform.dat
find / -name <UEFI UNDI file name>.rom
find / -name <HFI1 UEFI file name>.efi
```

Single Rail (one HFI) Example

The following example shows how to program a single HFI with three partitions.

1. Enter the following commands, replacing `x.x.x.x.x` with the versions provided in the release:

```
# hfil_eprom -w -c /lib/firmware/updates/hfil_platform.dat
# hfil_eprom -w -o /usr/share/opa/bios_images/HfiPcieGen3Loader_x.x.x.x.x.rom
# hfil_eprom -w -b /usr/share/opa/bios_images/HfiPcieGen3_x.x.x.x.x.efi
```

Note: Ensure you select the correct partitions for your files.



2. Verify programmed versions using the following commands:

```
# hfi_eprom -e
# hfi1_eprom -V -c
# hfi1_eprom -V -o
# hfi1_eprom -V -b
```

3. Reboot the server for the firmware updates to take effect.

Dual Rail (two HFIs) Example

The following example shows how to program two HFIs with three partitions for each HFI.

1. Obtain the device assignment using the following Linux* command:

```
# lspci | grep HFI
05:00.0 Fabric controller: Intel Corporation Omni-Path HFI Silicon 100 Series
[discrete] (rev 11)
81:00.0 Fabric controller: Intel Corporation Omni-Path HFI Silicon 100 Series
[discrete] (rev 11)
```

2. Program the devices using the following commands, replacing x.x.x.x.x with the versions provided in the release:

Note: Ensure you use the correct device assignment.

```
# hfi1_eprom -d /sys/bus/pci/devices/0000:05:00.0/resource0 -w -c /lib/
firmware/updates/hfi1_platform.dat
# hfi1_eprom -d /sys/bus/pci/devices/0000:05:00.0/resource0 -w -o /usr/
share/opa/bios_images/HfiPcieGen3Loader_x.x.x.x.x.rom
# hfi1_eprom -d /sys/bus/pci/devices/0000:05:00.0/resource0 -w -b /usr/
share/opa/bios_images/HfiPcieGen3_x.x.x.x.x.efi

# hfi1_eprom -d /sys/bus/pci/devices/0000:81:00.0/resource0 -w -c /lib/
firmware/updates/hfi1_platform.dat
# hfi1_eprom -d /sys/bus/pci/devices/0000:81:00.0/resource0 -w -o /usr/
share/opa/bios_images/HfiPcieGen3Loader_x.x.x.x.x.rom
# hfi1_eprom -d /sys/bus/pci/devices/0000:81:00.0/resource0 -w -b /usr/
share/opa/bios_images/HfiPcieGen3_x.x.x.x.x.efi
```

3. Verify programmed versions using the following commands:

```
# hfi1_eprom -d /sys/bus/pci/devices/0000:05:00.0/resource0 -V -c
# hfi1_eprom -d /sys/bus/pci/devices/0000:05:00.0/resource0 -V -o
# hfi1_eprom -d /sys/bus/pci/devices/0000:05:00.0/resource0 -V -b

# hfi1_eprom -d /sys/bus/pci/devices/0000:81:00.0/resource0 -V -c
# hfi1_eprom -d /sys/bus/pci/devices/0000:81:00.0/resource0 -V -o
# hfi1_eprom -d /sys/bus/pci/devices/0000:81:00.0/resource0 -V -b
```

4. Reboot the server for the firmware updates to take effect.

13.8 Verify Fabric Manager Sweep

By default, Fabric Manager sweeps every five minutes as defined in the `/etc/opa-fm/opa_fm.xml` file. Sweeps are triggered sooner if there are fabric changes such as hosts, switches, or links going up or down. Edit `/var/log/messages` and search for `CYCLE START`. Each cycle start has a complementary cycle end. Any links with errors are noted during this sweep cycle.



An example of a clean SM sweep follows:

```
Feb 16 16:12:08 hds1fnb8261 fm0_sm[3946]: PROGR[topology]: SM: topology_main: TT:
DISCOVERY CYCLE START - REASON: Scheduled sweep interval
Feb 16 16:12:08 hds1fnb8261 fm0_sm[3946]: PROGR[topology]: SM: topology_main:
DISCOVERY CYCLE END. 9 SWs, 131 HFIs, 131 end ports, 523 total ports, 1 SM(s),
1902 packets, 0 retries, 0.350 sec sweep
```

Compare the sweep result with `opafabricinfo` and the fabric topology.

13.9 Verify PM Sweep Duration

To show the sweep duration, open `opatop` then select `i`.

```
opatop: Img:Tue Feb 16 01:54:43 2016, Hist Now:Tue Feb 16 09:53:26 2016
Image Info:
Sweep Start: Tue Feb 16 01:54:43 2016
Sweep Duration: 0.001 Seconds

Num SW-Ports:      3  HFI-Ports:      2
Num SWs:           1  Num Links:      2  Num SMs:           2

Num Fail Nodes:    0  Ports:          0  Unexpected Clear Ports: 0
Num Skip Nodes:    0  Ports:          0
```

Select `r` to traverse the previous sweep duration time from history files. By default, PM sweeps every ten seconds. The latest ten image files (100 sec) are stored in RAM and up to 24 hours of history is stored in `/var/usr/lib/opa-fm`.

13.10 Check Credit Loop Operation

For details on credit loops, see the *Intel® Omni-Path Fabric Suite Fabric Manager User Guide* QoS Operation section.

To verify that a fabric does not have a credit loop issue, use:

```
# opareport -o validatecreditloops
```

The output should report similar to the following where no credit loops are detected:

```
Fabric summary: 135 devices, 126 HFIs, 9 switches,
504 connections, 16880 routing decisions,
15750 analyzed routes, 0 incomplete routes
Done Building Graphical Layout of All Routes
Routes are deadlock free (No credit loops detected)
```



13.11 Fabric Manager Routing Algorithm

If long Fabric Manager (FM) sweep times are observed or FM sweeps do not finish when a large number of nodes are bounced, consider changing the FM routing algorithm to `fattree` from the default `shortestpath`. You can do this by updating the `/etc/opa-fm/opafm.xml` file as shown in the following example:

```
<!-- ***** Fabric Routing ***** -->
<!-- The following Routing Algorithms are supported -->
<!-- shortestpath - pick shortest path and balance lids on ISLs -->
<!-- dgshortestpath - A variation of shortestpath that uses the -->
<!-- RoutingOrder parameter to control the order in which -->
<!-- switch egress ports are assigned to LIDs being routed -->
<!-- through the fabric. This can provide a better balance -->
<!-- of traffic through fabrics with multiple types of end -->
<!-- nodes. -->
<!-- See the <DGShortestPathTopology> section, below, for -->
<!-- more information. -->
<!-- fattree - A variation of shortestpath with better balancing -->
<!-- and improved SM performance on fat tree-like fabrics. -->
<RoutingAlgorithm>fattree</RoutingAlgorithm>
```



14.0 Run Benchmark and Stress Tests

For details on the tests provided with the Intel® Omni-Path Software, refer to the *Intel® Omni-Path Fabric Suite FastFabric User Guide*, MPI Sample Applications.

For optimal performance when running benchmark tests, configure systems according to the *Intel® Omni-Path Fabric Performance Tuning User Guide*.

14.1 Run Bandwidth Test

From `/usr/src/opa/mpi_apps` run:

```
# ./run_bw3
```

This test uses hosts defined in the `mpi_hosts` file, however, only the first two hosts in the file are used.

14.2 Run Latency Test

From `/usr/src/opa/mpi_apps` run:

```
# ./run_lat3
```

This test uses hosts defined in the `mpi_hosts` file, however, only the first two hosts in the file are used.

14.3 Run MPI Deviation Test

From `/usr/src/opa/mpi_apps` run:

```
# ./run_deviation 20 20 50
```

This test uses hosts defined in the `mpi_hosts` file.

14.4 Run run_mpi_stress

The default traffic pattern is "all-to-all" for this test.

Refer to *Intel® Omni-Path Fabric Suite FastFabric User Guide* for detailed information.

The test is located in `/usr/src/opa/mpi_apps`.

Note: These steps assume that you have defined hosts in the `/usr/src/opa/mpi_apps/mpi_hosts` file.



1. Clear error counters:

```
# opareport -o none --clearall
```

2. Confirm no errors exist:

```
# opareport -o error
```

3. Run `mpi_stress` test using a 60 minute duration:

```
# ./run_mpi_stress all -t 60
```

4. Run `opatop` to monitor the link utilization during the test.

5. Check error counts after the test:

```
# opareport -o errors
```

6. View the log file that is available for analysis in `/usr/src/opa/mpi_apps/logs`. The log filename format is `mpi_stress.date_time`.

7. Extract the log file in CSV format for errors and performance.

```
# opaextracterror
# opaextractperf
```

Example

The following example demonstrates this test being run on four nodes.

```
# ./run_mpi_stress all -t 60
Running MPI tests with 4 processes
logfile /usr/src/opa/mpi_apps/logs/mpi_stress.12Apr17150628
OpenMPI Detected, running with mpirun.
Running Mpi Stress ... Running Mpi Stress ...
Using hosts list: /usr/src/opa/mpi_apps/mpi_hosts
Hosts in run:
node1
node2
node3
node4

+ /usr/mpi/gcc/openmpi-1.10.4-hfi/bin/mpirun -np 4 -map-by node
--allow-run-as-root -machinefile /usr/src/opa/mpi_apps/mpi_hosts
-mca plm_rsh_no_tree_spawn 1
/usr/mpi/gcc/openmpi-1.10.4-hfi/tests/intel/mpi_stress -t 60
```

15.0 Take State Dump of a Switch

Note: Taking a state dump is a disruptive process and requires reboot of the switch after the state dump is taken. A state dump should only be taken if required to debug an issue.

You can take a state dump of any switch in the fabric, using its LID.

Prerequisites

- Find the LID of the switch whose state you want to dump by running the `opaextractlids|grep switch name` command.
- Contact Intel Customer Support to get the correct username and password for the `supportLogin` command.

Procedure

The following example describes how to take a state dump of a switch.

- Log in to a managed switch. The default username and password are `admin` and `adminpass`.
- Run the `supportLogin` command using the support username and password obtained in **Prerequisites**.

```
-> supportLogin
username: support
password:
```

- For a local managed switch, run the `ismTakeStateDump` command.
 - For a remote managed or externally-managed switch, run the command below, where `<lid>` identifies the desired switch.

```
-> ismTakeStateDump -lid <lid>
Dumping state of the switch at lid 4 to /firmware/prr-LID0004.gz
```

- From the Management Node, SFTP to the managed switch used for running the state dump command to retrieve the log:

```
sftp admin@<managed switch> with password adminpass.
admin@10.228.222.20's password:
Connected to 10.228.222.20.
sftp> dir
admin      operator      prr-LID0004.gz  prr-LID0005.gz  prr-LID0015.gz
get prr-LID0004.gz
```

- Reboot the switch on which the state dump was taken to clear the state dump. For externally-managed switches, use FastFabric to reboot the switch.



16.0 BKM for OPA Commands

Note: OPA commands should be issued from the Management Node where the IFS Full package was installed.

16.1 Retrieve Host Fabric Interface (HFI) Temperature

Use the command:

```
# cat /sys/class/infiniband/hfi1_X/tempsense
```

where *X* represents the device number.

When you send the command, the information is acquired at that specific time. Do not be concerned with the file's date/time.

An example of the output and the definition for each group of numbers follows:

```
# cat /sys/class/infiniband/hfi1_0/tempsense
68.50 0.00 105.00 105.00 0 0 0
```

Table 1. HFI Temperature Output Definitions

Example Value	Definition
68.50	Actual temperature Temperature steps are 0.25 °C increments.
0.00	Low limit
105.00	Upper limit
105.00	Critical limit
0	Low limit flag 1 = flag is set.
0	Upper limit flag 1 = flag is set.
0	Critical limit flag 1 = flag is set.

16.2 Read Error Counters

To use the default thresholds defined in the `/etc/opa/opamon.conf` file, use the command:

```
# opareport -o errors
```



To run against a different threshold file, for example `/etc/opa/opamon.si.conf`, use the command:

```
# opareport -o errors -c /etc/opa/filename.conf
```

16.3 Clear Error Counters

Use the command:

```
# opareport -o none --clearall
```

16.4 Load and Unload Intel® Omni-Path Host HFI Driver

If the configuration is changed, you may need to reload the HFI driver.

Unload the HFI driver using:

```
# modprobe -r hfil
```

Load the HFI driver using:

```
# modprobe hfil
```

Note:

The HFI driver should not be reloaded on SM nodes due to unloading other required dependencies and restarting them. On all other nodes, you may have to restart the IPoIB interface with `ifup` after the HFI driver is reloaded.

16.5 Analyze Links

To include the link quality of the local HFI port, use the command:

```
# opainfo
```

To include links with lower quality, use the command:

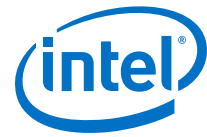
```
# opareport -o errors
```

To output the ports with a link quality less than or equal to *value*, use the command:

```
# opareport -o links -F linkqualLE:value
```

To output the ports with a link quality greater than or equal to *value*, use the command:

```
# opareport -o links -F linkqualGE:value
```



To output the ports with a link quality equal to *value*, use the command:

```
# opareport -o links -F linkqual:value
```

Table 2. Link Quality Values and Description

Link Quality Value	Description
5	Working at or above preferred link quality, no action needed.
3	Working on low end of acceptable link quality, recommended corrective action on next maintenance window.
2	Working below acceptable link quality, recommend timely corrective action.
1	Working far below acceptable link quality, recommend immediate corrective action.
0	Link down

16.6 Trace Route between Two Nodes

Use the command:

```
# opareport -o route -S nodepat:"hds1fnb6101 hfil_0" -D nodepat:"hds1fnb6103 hfil_0"
```

To trace using LID, use the command:

```
# opareport -o route -S lid:5 -D lid:8
```

16.7 Analyze All Fabric ISLs Routing Balance

Use the command:

```
# opareport -o treepathusage
```

16.8 Dump Switch ASIC Forwarding Tables

To display all switch unicast forwarding tables DLIDs and Egress ports, use the command:

```
# opareport -o linear
```

To display multicast groups and members, use the command:

```
# opareport -o mcast
```

16.9 Configure Redundant Fabric Manager (FM) Priority

This section describes several configuration methods.

16.9.1 Configure FM Priority from a Local or Remote Terminal

Perform the following steps:

1. Edit the `/etc/opa-fm/opafm.xml` file.
2. Select the `<Priority>0</Priority>` value and change 0 to the number you want (0 - 15).
3. Save the file.
4. Start or restart the Fabric Manager to load the new file, using the command:

```
# systemctl restart opafm
```

Note: If you set a Fabric Manager to a higher priority, it becomes the master Fabric Manager automatically. The sticky finger option is disabled by default.

16.9.2 Configure FM Elevated Priority

Perform the following steps:

1. Edit the `/etc/opa-fm/opafm.xml` file.
2. Select the `<ElevatedPriority>0</ElevatedPriority>` value and change 0 to the number you want (0 - 15).
3. Save the file.
4. Start or restart the Fabric Manager to load the new file, using the command:

```
# systemctl restart opafm
```

16.9.3 Configuration Consistency for Priority/Elevated Priority

Priority and Elevated Priority are not part of the `opafm.xml` configuration consistency checksum calculation. This makes standby Fabric Managers with mismatched configuration inactive because they are not valid to take over as Master in case of failover.

Having different values for Priority and Elevated Priority settings for SM instances is allowed and failover works as documented per Priority/ElevatedPriority settings. In normal failover without elevated priority, if the original Master Fabric Manager goes down, the Standby Fabric Manager becomes Master. When the original Master comes back up, it again takes over as Master.

Note: In sticky failover, Elevated Priority is used and with sticky failover enabled, when the original Master comes back up, it does NOT take over.

16.9.4 Display FM states from the Management Node

Run the `opafabricinfo` command to view the new active master SM.



17.0 Final Fabric Checks

After addressing all issues, perform final fabric checks as described in [Verify Server and Fabric](#) on page 31.