Fabric Computing That Works™

Installation Guide for Intel® Omni-Path Fabric Suite - Debian Release

IntelOPA-IFS.DEBIAN9-x86_64.10.10.1.0.36

Prepared for Intel, Inc.
by System Fabric Works, Inc.

April 24, 2020

Contact Information:

http://www.systemfabricworks.com

email fabricsupport@intel.com

Table of Contents

1 Overview

This guide is a supplement to the standard documentation for Intel® Omni-Path Fabric Suite (IFS). These guides include the README for Intel® IFS 10.10.1.0.36, Intel® Omni-Path Architecture (OPA) Fabric Administrator's Guide and Intel® Omni-Path Fabric Suite FastFabric User Guide. The focus of this guide is to document special procedures and considerations for installation and configuration of the software on Debian Linux.

This guide is also relevant to the IntelOPA-BASIC distribution, which does not include opa-fm or opa-fastfabric packages.

2 Requirements

This software release is intended to be installed on a Debian 8.x or 9.x host. Third-party InfiniBand software such as OFED or Mellanox OFED should not be installed on the host.

IFS only supports the x86_64 architecture.

3 IFS Installation

Installation of IFS is performed via the INSTALL script provided within the software distribution. This script checks for installation dependencies, installs the IFS packages, creates configuration files for RDMA and OPA, rebuilds the initramfs and enables the opa service.

Note: Execute "INSTALL -h" first in order to see command-line options, such as "enable fabric manager".

Note: The activities of the installer are recorded in /var/log/opa.log.

1. Unpack the software distribution.
   `tar xf IntelOPA-IFS.DEBIAN9-x86_64.10.10.1.0.36.tgz`
2. Change your working directory to the distribution folder.
   `cd IntelOPA-IFS.DEBIAN9-x86_64.10.10.1.0.36`
3. Execute the installer as root
   `sudo ./INSTALL -a`
4. Reboot the host.
   `sudo reboot`

The installer may report an error after it installs the required packages from apt-get. If that occurs, execute INSTALL again and it should execute without errors. During installation, the Debian packaging rules will not overwrite existing configuration files from a previous install. The packager will output messages which indicate "Keeping old config file as default" and the packaged configuration file will be installed with a dpkg-dist suffix.

Once the installer completes, the administrator should reboot the host. After reboot, the hfi1 and ib modules should be loaded and the fabric may be tested according to procedures specified within the Administrator's Guide.

The .deb packages are all located under `packages` in the distribution folder. This folder also includes Debian build artifacts and source packages.

4 Kernel Support

kmod-ifs-kernel-updates is tied to a specific kernel revision, which is included in the package name (e.g. kmod-ifs-kernel-updates-4.9.0-4).

If IFS does not ship with a kmod-ifs-kernel-updates that supports the currently booted kernel version, INSTALL will report that the administrator must install a supported kernel or build the package for the booted kernel.

Invoke INSTALL -k to build kmod-ifs-kernel-updates for the booted kernel. If the build succeeeds, the binaries will be copied to the distro packages directory. INSTALL may then be invoked to install the distribution normally. The distribution directory may be copied to any host that needs the kernel support that this process created.

The currently supported kernels are of the 4.9.x variety for Debian 9 (stretch) and 3.16.x for Debian 8 (jessie).

Please contact fabricsupport@intel.com with any questions or issues that arise from this process.

5 IFS Uninstallation

IFS is uninstalled via the INSTALL script in the software distribution. This script removes the installed IFS packages. It does not remove the dependencies that were installed via apt-get.

1. Change your working directory to the distribution folder.
   `cd IntelOPA-IFS.DEBIAN9-x86_64.10.10.1.0.36`
   Execute the installer as root, with the -u argument
   `sudo ./INSTALL -u`

Uninstallation does not modify `/etc/security/limits.conf`. The administrator may remove the OPA entries from this file before rebooting the host.

If you wish to remove all packages and their configurations (i.e. "purge"), execute `./INSTALL -u -p`.

6 IFS Debug Packages

The INSTALL script does not install any debug packages. Debug packages exist under the packages directory and have -dbgsym in the package name.

```
ls packages/*-dbgsym_*.deb
```

The administrator may install debug packages using the `dpkg` utility. For example:

```
dpkg -i packages/libhfi1-dbgsym_0.5-27-2ifs+deb9_amd64.deb
```

7 IFS Source Code

Source code packages are included in the distribution tarball. The Debian packages and associated source code are under the `packages` directory.

To determine which source package is used to create a particular package, use the `dpkg --info` command and look for the Package attribute:

```
dpkg --info \
  hfi1-diagtools-sw_0.8-84-2ifs+deb9_amd64.deb
 Package: hfi1-diagtools-sw
```

In this example, hif1-diagtools-sw is the source package. If the source package name differs from the binary, there will be a Source attribute.

The related source code and build artifacts are:

- hfi1-diagtools-sw_0.8-84.orig.tar.gz - tarball of upstream code
- hfi1-diagtools-sw_0.8-84-2ifs+deb9_amd64.build - output of the build process that created the binary packages
- hfi1-diagtools-sw_0.8-84-2ifs+deb9_amd64.changes - summary of changes since the previous version of the package
- hfi1-diagtools-sw_0.8-84-2ifs+deb9.dsc - Debian source package description
- hfi1-diagtools-sw_0.8-84-2ifs+deb9.debian.tar.xz - tarball of debian source

To build a binary package, the `devscripts` package must be installed. Individual packages each have their own build requirements, which are specified in `debian/control` by the Build-Depends statement. Extract the debian source package and use `debuild` to create the binary:

```
# cd ~
# mkdir tmp
# cd IntelOPA-IFS.DEBIAN9-x86_64.10.10.1.0.36/packages
# dpkg-source -x \
   hfi1-diagtools-sw_0.8-84-2ifs+deb9.dsc \
   ~/tmp/hfi1-diagtools-sw
# cd ~/tmp/hfi1-diagtools-sw
# debuild -us -uc
```

After completion of the build, the generated package and artifacts will be present in the parent directory of the build process.

If the build process for a package differs this standard, it will be described in debian/README.debian.

8 Backports Repository

Installation of IFS may require components from the appropriate backports repository in the following scenarios:

- Jessie - jessie-backports
  - All installations - linux-cpupower
  - GPUDirect - nvidia-cuda-toolkit 7.5.18-4~bpo8+1
- Stretch - stretch-backports
  - GPUDirect - nvidia-cuda-toolkit 9.1.85-8~bpo9+1

Please refer to https://backports.debian.org/Instructions/

For Jessie, the Debian FTP site for jessie-backports has been archived. This article discusses the situation: https://www.lucas-nussbaum.net/blog/?p=947

Add the following to /etc/apt/sources.list:

```
deb http://archive.debian.org/debian/ jessie-backports main contrib non-free
deb-src http://archive.debian.org/debian/ jessie-backports main contrib non-free
```

Then, create /etc/apt/apt.conf.d/99no-check-valid-until, containing:

```
Acquire::Check-Valid-Until no;
```

apt should now be able to use jessie-backports.

9 GPUDirect

GPUDirect requires installation of NVidia CUDA libraries before executing INSTALL. It is recommended to install the correct version of nvidia-cuda-toolkit, as opposed to installing the individual packages.

Please note that the appropriate backports repository must be added to the host's apt configuration. The non-free component must also be enabled, at least for the appropriate backports repository indicated below. For more information apt configuration, refer to https://wiki.debian.org/SourcesList.

- Jessie - nvidia-cuda-toolkit 7.5.18-4~bpo8+1
  ```
  $ apt install -t jessie-backports nvidia-cuda-toolkit
  ```
- Stretch - nvidia-cuda-toolkit 9.1.85-8~bpo9+1
  ```
  $ apt install -t stretch-backports nvidia-cuda-toolkit
  ```

If desired to install individual packages, execute INSTALL -G -a and take note of the missing packages reported by the script.

Once the required NVidia packages have been installed, execute INSTALL -G -a to install IFS with CUDA support.

If the ifs-kernel-updates package needs to be rebuilt for the host's kernel, a few extra streps must be taken. The NVidia driver needs to be built in order to generate its Module.symvers and the build's location must be specified to INSTALL when it builds the ifs-kernel-updates package.

```
$ cd $HOME
$ cp -r /usr/src/nvidia-current-384.130 .
$ cd nvidia-current-384.130
$ make
$ cd path/to/IFS/distro
$ sudo NVIDIA_GPU_DIRECT=$HOME/nvidia-current-384.130 \
    ./INSTALL -G -k -a
$ sudo ./INSTALL -G -a
```

10 HFI ROM Image Location

In order to comply with Debian Filesystem Hierarchy Standard, HFI ROM images are not located in `/opt/opa/bios_images`. The images are installed under `/usr/share/hfi1-uefi/bios_images`.

ROM images are provided by the UEFI distribution, available at https://support.systemfabricworks.com/redmine/projects/intel-omni-path/files .

11 RDMA Module Loading

On the RedHat platform, RDMA kernel modules are usually loaded via `/usr/libexec/rdma-init-kernel. rdma-init-kernel` is a RedHat package, and the prescribed method for loading these modules on Debian is via a configuration file in /etc/modules-load.d. The `opa-scripts` package creates `/etc/modules-load.d/rdma.conf` to load the RDMA modules at boot time.

A consequence of this technique is that the settings in `/etc/rdma/rdma.conf` are not used for the RDMA modules (e.g. "IPOIB_LOAD=no does not prevent ib_ipoib from loading). However, the opa service does use this file for its configuration.

12 IPoIB Interface Configuration

An IPoIB interface is configured using the standard Debian network interface configuration techniques. An overview is provided at https://pkg-ofed.alioth.debian.org/howto/infiniband-howto-5.html

To temporarily bring up the ib0 interface, execute the following as root, using the appropriate CIDR:

```
ifconfig ib0 10.20.30.2/24
```

The ib0 device should then be configured with an IP address.

```
ifconfig ib0
ping 10.20.30.2
```

The opaconfig script may be used by the administrator to configure IPoIB interfaces. Use menu option "2) Reconfigure OFA IP over IB".

To manually configure the ib0 interface and bring it up at boot time, create the following file /etc/network/interfaces.d/ifcfg-ib0:

```
auto ib0
iface ib0 inet static
  address 10.20.30.2
  netmask 255.255.255.0
  broadcast 10.20.30.255
```

If connected mode is desired, and/or an MTU should be set, use a configuration that resembles the following:

```
auto ib0
iface ib0 inet static
  address 10.20.30.2
  netmask 255.255.255.0
  broadcast 10.20.30.255
  post-up echo connected > /sys/class/net/ib0/mode && ifconfig ib0 mtu 65520
```

The post-up directive in that configuration sets ib0's IPoIB mode to "connected" and sets the MTU to 65520. The directive may be modified to meet the system requirements.

Test the configuration:

```
$ ifup ib0
$ ifconfig ib0 #view and verify IP address
$ ping 10.20.30.2
```

The ib0 interface should come up automatically and be configured with the IP address when the host is rebooted.

13 TID-RDMA

TID-RDMA enables hardware offload for RDMA processing of messages sized 2MB or greater. It is not enabled, by default.

To enable TID-RDMA:

1. If this is not a virtual machine, add intel_iommu=off to the GRUB_CMDLINE_LINUX_DEFAULT setting in `/etc/default/grub`
2. Execute update-grub to update settings in `/boot/grub/grub.cfg`. WARNING: this will overwrite any customizations previously made

to /boot/grub/grub.cfg. If grub.cfg has been manually edited, then intel_iommu=off needs to be added to the linux kernel statements in grub.cfg.
3. In /etc/modprobe.d/hfi1.conf, add cap_mask=0x4c09a01cbba to the options for hfi1.
4. Reboot.

To determine if TID-RDMA is enabled, examine the cap_mask module parameter. If TID-RDMA is enabled, it should be 0x4c09a01cbba.

```
$ cat /sys/module/hfi1/parameters/cap_mask
0x4c09a01cbba
```

14 Accelerated IPoIB (AIP)

IFS for Debian 9 includes Accelerated IPoIB (AIP) in the installed ib_ipoib module. Please see the Intel Readme for IFS for more information about AIP features.

AIP is not offered for Debian 8.

15 Use of PSM and /dev/ipath, MPI

If an MPI implementation has not been compiled with PSM2 support, and PSM is desired instead of ibverbs, please review section 4 of:

http://www.intel.com/content/dam/support/us/en/documents/network-and-i-o/fabric-products/Intel_OP_Fabric_Host_Software_UG_H76470_v4_0.pdf

/dev/ipath is created by udev according to /lib/udev/rules/40-psm.rules when the hfi1 module is loaded by the operating system.

psm2-compat is offered as a Debian alternative. To select psm2-compat's libpsm_infinipath, execute:

```
sudo update-alternatives --set libpsm_infinipath.so.1 \
 /usr/lib/libpsm2-2/libpsm_infinipath.so.1
```

16 openmpi and Debian

There are two issues when running openmpi on Debian:
1. openmpi does not specify absolute paths for certain utilities that it executes via ssh (e.g. orted).
2. Debian's .bashrc terminates early when executed in a non-interactive shell (e.g. a remote ssh command).

The consequence of this is that mpirun will report an error "bash: orted: command not found". In order to correct this:

1. Use mpi-selector-menu to select openmpi on each node.
2. Add "`. /etc/profile.d/mpi-selector.sh`" to the beginning of `$HOME/.bashrc` on all nodes.

17 Fabric Manager Service

The Fabric Manager installed by opa-fm is not enabled by default. This service is managed via the `systemctl` utility.

Note: `INSTALL` will enable opafm if the -f flag is specified.

To start the Fabric Manager:

```
systemctl start opafm
```

To stop the Fabric Manager:

```
systemctl stop opafm
```

To enable the Fabric Manager on boot:

```
systemctl enable opafm
```

18 Build mpi_apps

In order to build mpi_apps, which are required for certain OPA utilities (e.g. `opacabletest`), the administrator should install the necessary development tools and execute the make process.

```
$ sudo bash
$ apt-get install build-essential gfortran
$ cp -a /usr/src/opa/mpi_apps $HOME/mpi_apps
$ cd $HOME/mpi_apps
$ make
```

After the build completes, these final messages should appear on the console:

```
Built subset of sample applications
Built sample applications
```

19 Known Issues Specific to Debian IFS

Issue	Description	Notes
345	hfi1-firmware conflicts with firmware-misc-nonfree	hfi1-firmware installs /lib/firmware/hfi1_dc8051.fw, which conflicts with firmware-misc-nonfree 20161130-3. The workaround is to

		remove firmware-misc-nonfree.
624	cannot mmap PCI device when driver is loaded	On kernel 4.9, hfi1_eprom and hfidiags report Unable to mmap, Invalid argument. This is due to https://lkml.org/lkml/2015/11/25/691 . Either unload the hfi1 module or add iomem=relaxed to the kernel command line.
693	IPoIB allows illegal MTU for non-AIP datagram	When AIP is disabled and the IPoIB mode is datagram, the maximum allowed MTU should be 4092. If set to greater than 4092, communication failures will be present on the IPoIB interface. Solutions are to set MTU <= 4092, use connected mode or enable AIP.

20 References

- Intel® Omni-Path Architecture (OPA) Fabric Administrator's Guide
- Intel® Omni-Path Fabric SuiteFastFabric User Guide
- Infiniband HOWTO: Setting up a basic infiniband network
- Infiniband HOWTO: IP over Infiniband (IPoIB)
- Intel® Omni-Path Fabric Suite FabricManager User Guide