

Omni-Path Cluster Configurator

User Guide

October 2016



Legal Disclaimer

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 by visiting: <http://www.intel.com/design/literature.htm>

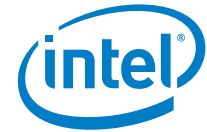
Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at <http://www.intel.com/> or from the OEM or retailer.

No computer system can be absolutely secure.

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

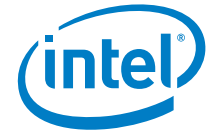
*Other names and brands may be claimed as the property of others.

Copyright © 2016, Intel Corporation. All rights reserved.



Contents

1.	Introduction	5
1.1	Intended Audience	5
2.	Designing Intel® OPA Fabric Cluster	6
2.1	Single 1U switch	6
2.2	Single Director	6
2.3	2-Tier	7
2.4	3-Tier fabric tree	7
3.	The User Interface	8
3.1	Quick Config	8
3.2	Detailed 'Config' with parameters	9
3.2.1	Ports:	9
3.2.2	OverSub:	9
3.2.3	Rails:	10
3.2.4	Switch Size:	10
3.2.5	EdgesPerTrunk:	10
3.2.6	HFI Type:	11
3.2.7	Ports per Cab:	11
3.2.8	Support	11
3.2.9	Years:	11
4.	Fabric Solution.....	12
4.1	Show Design	12
4.2	Show BOM	12
4.3	Configuration ID	12
4.4	Download Diagram	13
4.5	Cable length Estimator	13
4.6	Floor Plan Parameters	14



Revision History

Date	Revision	Description
October 2016	2.0	Updated for Configurator release 3.0.3
July 2016	1.0	Initial release.

§



1. Introduction

Intel Omni Path Cluster Configurator is a tool used by Fabric Marketing teams and OEM's pre-sales to design fabric solutions [topologies] based on requirements. This tool also provides accurate orderable solutions [Bill of Materials BOM] along with pricing calculations. This tool provides the best solution on how the cluster should be build based on customer requirements.

1.1 Intended Audience

This manual guides to use the configurator tool and also understanding solutions along with few detailed calculation examples and diagrams. Also provides explanation of rules/algorithms used in this configurator logic to aim authority on best solutions.

2. Designing Intel® OPA Fabric Cluster

There are several common topologies for Omni-path Fabric in different tiers and using different switches [directors and Edges] based on user requirement. Following is the list of some topologies:

- Single 1U switch
- 2-tier
- Single Director
- 3-Tier

2.1 Single 1U switch

This type of fabric solution has only 1 layer of switch which is 1U i.e., Edge (24p or 48p). This topology is selected only when number of ports required are less than 48.

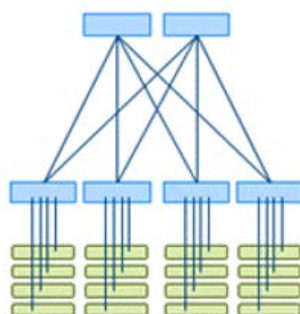
Single 1U switch



2.2 Single Director

This type of fabric uses one director chassis which has 768 ports. So this can be used when requested ports are up to 768 and switch size = AUTO [refer below]

2-tier

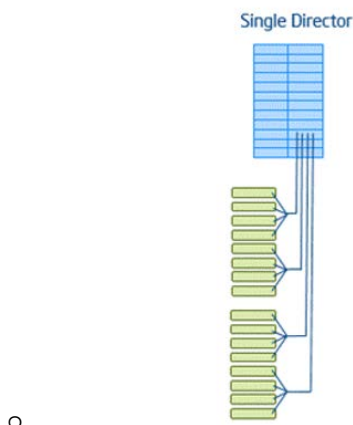


2.3 2-Tier

This fabric solution has two physical layers and one logical layer. This topology only uses Edge switch [24p or 48p], when switch size is restricted to 24 or 48 ports, and number of ports is more than 24 or 48 respectively, then 2-tier fabric solutions are provided.

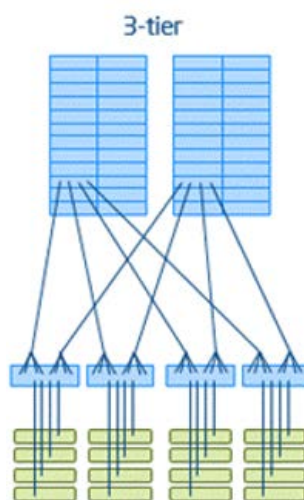
A director switch is simply a fat-tree-in-a-box, so the 2-tier and Single Director Fabric types are logically equivalent. A 2-tier may be a better solution if:

- The fabric size or the oversub ratio would require a significantly underpopulated director, and so a director would be relatively expensive.
- A blade system, where the first tier of the fabric is built into the blade chassis.



2.4 3-Tier fabric tree

This topology is similar to 2-tier topology with 2 physical layers [director and edge] and 1 logical layer [leaves and spines]. This topology is used only when requested ports are >768 and switch size = auto. Primary physical layer is Director Chassis and intermediate layer is by Edges.



3. The User Interface

Intel® OPA cluster configurator tool have various options on how user wants to configure the fabrics.

3.1 Quick Config

For a simple quick configure, enter the number of end ports required and click “Quick Config”. The Quick Config button preloads two or three of the Solution Columns with parameters likely to provide useful solutions for the number of ports you have specified.

Our configurator “quick Config” solution column rules are:

(Default of PCIE x16 HFI type, Single rail)

- For a fabric with 768 end-ports or fewer, it will provide
 1. a single director
 2. a 2-tier tree of 48p switches
 3. a 2-tier tree of 48p switches with 2:1 oversubscription.
- For a fabric with more than 768 end-ports, it will provide
 1. a 3-tier non-blocking solution
 2. a 3-tier solution with 2:1 oversubscription




3.2 Detailed 'Config' with parameters

Configurator provides more config parameters to build a fabric solution.

List of options user can input to build configurator:

Ports	<input type="text" value="2000"/>
Oversub	<input type="text" value="1:1"/>
Rail	<input type="text" value="Single-Rail"/>
Switch size	<input type="text" value="Auto"/>
Edges per Trunk	<input type="text" value="Min"/>
HFI Type	<input type="text" value="PCIe x16"/>
Ports per cab*	<input type="text" value="72"/>
Support	<input type="text" value="Basic"/>
Years	<input type="text" value="3"/>

Click this!  [Configure.1](#)

3.2.1 Ports:

The number of end-ports required for the fabric.

3.2.2 OverSub:

The level of oversubscription required for the fabric. Oversubscription can be expressed in any of these formats:

- Port Allocation. For 2-tier and 3-tier fabrics, how the ports on each edge switch are divided between EndPorts and ISLports.
Examples: 24:24, 32:16, 24:16.
- Ratio. Examples: 1:1, 2:1.
- Percentage. Examples: 100%, 50%.

Oversub will be interpreted slightly differently for different types of fabric.

- Single 1U Switch: oversub is ignored.
- Single Director Switch: The configurator will remove double spine modules to get near to, but not beyond, the requested oversub value. The Port Allocation format is not meaningful in this context, so it will be converted to a Ratio.
- 2-tier and 3-tier fabrics: If the Port Allocation format is used, the total number of ports (EndPorts+ISLports) must be less than or equal to the number of ports on an edge switch. If less than (example: 24:16), then some ports will be unused. The Ratio and Percentage formats will be converted to a Port Allocation value that uses all the ports on an edge switch.

3.2.3 Rails:

Dropdown has options: Single rail and Dual Rail. Default is single rail.

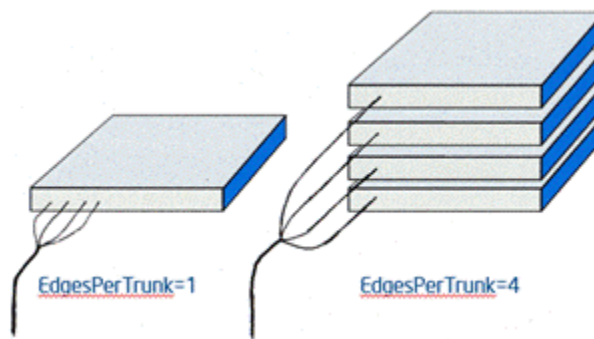
3.2.4 Switch Size:

The size of switch [number of ports] to use for the fabric. When set to Auto, configurator chooses the optimum size of switch or switches.

Dropdown menu has options of 24p, 48p, 192, 288, 768, 1152 port switch sizes.

3.2.5 EdgesPerTrunk:

This parameter is used in 3-tier fabrics. In a 3-tier fabric, edge switches are connected to one or more core director switches. The director switches use trunk ports which contain four individual links. To connect to an edge switch, each trunk cable has to split into four tail-cables. The EdgesPerTrunk parameter defines how many edge switches each trunk cable connects to, and can have values of 1, 2 or 4.



- EdgesPerTrunk=1 is very convenient for physical cabling because each trunk cable connects to a single edge switch, allowing for short tails and so reducing cable density by a factor of four.

Example: for fabrics with oversub=24:24 and EdgesPerTrunk=1, there would be six trunk cables connected to each edge switch.

However, it is not possible to build all fabrics, especially larger fabrics, using EdgesPerTrunk=1.

- EdgesPerTrunk=4 may be used for all fabrics. but it is less convenient for physical cabling because the tails of each trunk cable connect to a group of four separate edge switches, some of which may be in different cabinets.

Example: for fabrics with oversub=24:24 and EdgesPerTrunk=4, there would be twenty-four trunk cables connected to each group of four edge switches.

Also, EdgesPerTrunk=4 gives the fabric a slightly improved average latency, and may require slightly more parts.

With EdgesPerTrunk=Min, the configurator first determines the minimum number of core switches required for the fabric. Next, it chooses the minimum value of EdgesPerTrunk that can be used in that



topology. If you select EdgesPerTrunk=1 or 2, some larger fabrics cannot be built, and some fabrics may require more core switches than if using EdgesPerTrunk=Min.

*We recommend you always start with EdgesPerTrunk=Min. If desired, try different values of EdgesPerTrunk, and compare them with the initial solution given by EdgesPerTrunk=Min.

3.2.6 HFI Type:

The type of Host Fabric interface card (HFI) required. Drop down has options “PCIE x16”, “PCIE x8”, “KNL-F”.

KNL-F has 2 ports of integrated x16 HFI ports. So if Single rail is selected with KNL-F only 1 of 2 ports on KNL-F is used. If Dual rail is selected, both the ports on KNL-F is used.

3.2.7 Ports per Cab:

The number of fabric ports in each compute-node cabinet. Dual rail can fit in twice the number of ports per cabinet.

3.2.8 Support

The name of the service program [support level] required.

3.2.9 Years:

The number of years for which service coverage is required.

From all above described configuration params, you can set the ports and Oversub values and leave the other params at their default values.

4. Fabric Solution

Intel® Omni-Path Architecture
 Tags: Fabric Products Infraband High Performance Computing

INTEL® OMNI-PATH FABRIC CLUSTER CONFIGURATOR

Enables users to build and compare fabric configurations of various sizes and subscription rates

Resources

Intel is committed to making quality products and welcomes your comments and feedback. Please direct comments and feedback to fabricmarketing@intel.com.

0 [Quick Config](#) [Home](#) [Clear Config](#)

Ports: 2000
 Oversub: 1:1
 Rail: Single-Rail
 Switch size: Auto
 HFI Type: PCIe x16
 Ports per cab*: 72
 Support: Basic
 Years: 3

[Configure 1](#)

2016 ports
 3-tier non-blocking fat-tree
 84x48p + 3x768p switches

[Show Design](#) [Show BOM](#)

Ports: 0
 Oversub: 1:1
 Rail: Single-Rail
 Switch size: Auto
 HFI Type: PCIe x16
 Ports per cab*: 72
 Support: Basic
 Years: 3

[Configure 2](#)

Ports: 0
 Oversub: 1:1
 Rail: Single-Rail
 Switch size: Auto
 HFI Type: PCIe x16
 Ports per cab*: 72
 Support: Basic
 Years: 3

[Configure 3](#)

Ports: 0
 Oversub: 1:1
 Rail: Single-Rail
 Switch size: Auto
 HFI Type: PCIe x16
 Ports per cab*: 72
 Support: Basic
 Years: 3

[Configure 4](#)

[Download Diagram](#)
[Cable Length Estimator](#)

Configuration id: 334e fa90 **Version:** 3.0.3

Detailed Description

Overview
 This fabric is designed for 2000 server ports with Single-rail, and provides 2016. It is a 3-tier non-blocking fat-tree with 2 layers of physical switches using 84x 48p edge switches and 3x 768p core switches.

Edges Switches
 Each edge switch has 48 ports: 24 are end-ports and 24 are ISL ports.

Inter-switch cabling
 The 24 cables from each edge switch are split into 3 bundles of 8. Each of the 3 core switches takes 1 bundle of 8 cables from each of the 84 edge switches, using 672 of the 768 ports on each 768 port core switch.

*Dual rail can fill in twice 8ports per cabinet

Populating leaf slots
 Main cores: Qty 3, each with 42x 16-port leafs. LeafGroups are formed as follows: 5 groups, where each bundle of 8 cables connects to 8 leafs, consuming 80 bundles.
 1 group, where each bundle of 8 cables connects to 2 leafs, consuming 4 bundles.

Fabric hop-count
 Each server port has 23, 360 and 1632 other server ports within 1, 3 and 5 hops. True for all servers connected to edge switches that are connected to full leafGroups

Cable types and lengths
 There are 28 cabinets of servers, with 72 server ports per cabinet. Each server cabinet contains 3 edge switches.
 All cables are single-link QSFP-QSFP copper or active-optical.

Brief Description
 2016 ports at 24:24 (non-blocking), nCoresFilled/Factored/Packed=2.7/3/3, edgesPerTrunk=1
 84x24:24 + 3x672
 Main cores: Qty 3, each with 42x 16-port leafs. LeafGroups: Grouping: 5(8L)+8(8L, 1(2L)=48.

4.1 Show Design

This gives detailed description of building fabric solutions with requested number of fabric ports, including the inter-switch cabling for 2-tier and 3-tier fat trees.

4.2 Show BOM

BOM provides the line items to build a purchase order with latest MM's.

4.3 Configuration ID

[Show Design](#) [Show BOM](#)

Configuration id: XXXXXXXXXX Version: 3.0.3

Detailed Description

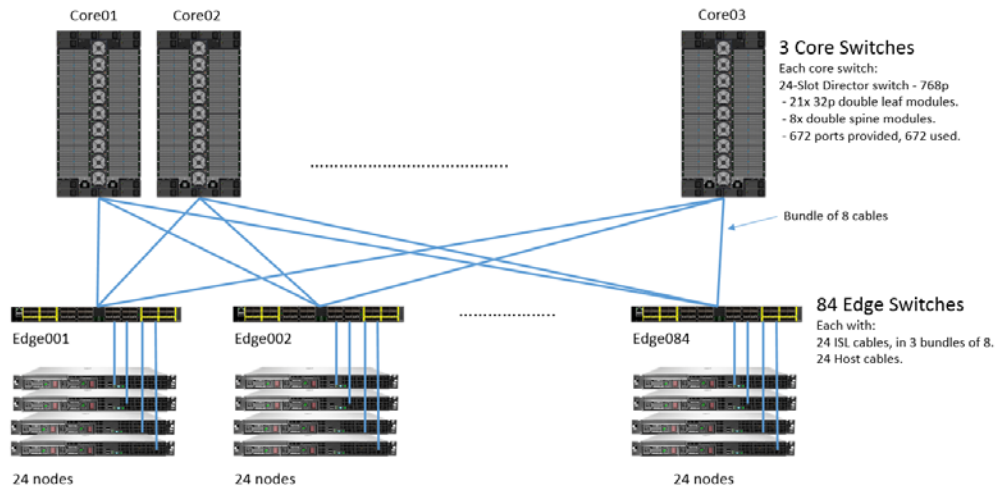
Unique ID is generated for each of the user configuration. This ID helps users to retrieve, share the configurations with others.

0 [Discount](#) 0 [Warranty Discount](#)

4.4 Download Diagram

This option gives graphical representation of the cluster taking user inputs from configurator.

2016 ports Single -rail for 2000 nodes/hosts, 3 tier, non-blocking fat-tree



4.5 Cable length Estimator

This feature is used to calculate the lengths of cables used to build the real solution. Also provide user option to enter the cable lengths to generate the quantities for the requested fabric cluster. Solution gives required cable quantities for Host and ISL cables.

Cable lengths

Length	Quantity	Cumulative Length (meters)		
		Supplied	Required	Excess
4 m	32	128	119	9
5 m	224	1120	995	125
6 m	576	3216	2950	266
7 m	480	3360	3175	185
8 m	360	2880	2673	207
9 m	328	2952	2752	200
10 m	56	560	529	31
Totals:	2016	14216	13190	1026

Available lengths: 1, 5, 10.5, 15

1 m	2080	Host cables: QSFP/QSFP, 1m - 2m		
		ISL cables: QSFP/QSFP ISL:		
5 m	256		1280	1113
10.5 m	1760		18480	12077
Totals:	2016		19760	13190

6570m (33%) of the 19760m of cable supplied is excess and will need to be coiled under the floor.
 Selecting lengths from [1, 5, 10.5, 15] provides 9544m more cable than selecting lengths in 1m increments.

[Note: Totals may not equal the sum of the displayed lengths due to rounding at different stages in the calculations.]

4.6 Floor Plan Parameters

Based on the few assumptions floor plan is designed for the user displaying hot and cold aisles with the cabinet arrangement. Click on each of the cabinet to get the detailed view of cable information required to connect racks and cabinets across the aisles.

Floor Plan

Click on a cabinet to view its cables.



User also have a table of floor plan parameters (rack units per end port, width of aisles etc.) which can edited to generate custom floor plan and length of cables needed to build the solution.

Parameters

Refresh		
Rack Units per end port	0.5	dRUpperNode
Max RUs to use in each cabinet	39	maxRU
Max end ports per cabinet	72	maxPortsPerCab
Width of cabinets (side to side)	0.7	dRackWidth
Depth of cabinets (front to back)	1.0	dRackDepth
Width of Hot Aisles	1.2	dHotAisle
Width of Cold Aisles	1.2	dColdAisle
From cable tray to first slot in compute cabinet	0.5	dGap1
From cable tray to first slot in switch cabinet	0.5	dGap2
Floor plan: number of rows	7	floorRows
Floor plan: number of cabinets in each row	17	floorCols
One rack unit	0.04445	dRU
Type of topology	2-tier	topoType
Number of end ports	2000	nPorts
Number of edge switches that each trunk cable connects to	1	edgesPerTrunk
Number of links in each [trunk] cable	1	trunkFactor
Number of edge switches	84	nEdges
Number of end ports on each edge switch	24	portsPerEdge
Size of core switch(es)	768	coreSize
Number of main core switches	3	n1
Number of cables in the bundle from each edge to the main core switches	8	b1
Number of remainder core switches	0	n2
Number of cables in the bundle from each edge to the remainder core switch	0	b2