WHITE PAPER

Implementing Storage in Intel® Omni-Path Architecture Fabrics

Rev 1

A rich ecosystem of storage solutions supports Intel Omni-Path Architecture-based systems.

Executive Overview

The Intel® Omni-Path Architecture (Intel® OPA) is the next-generation fabric architected to deliver the performance and scaling needed for tomorrow's high performance computing (HPC) workloads.

A rich ecosystem of storage offerings and solutions is key to enabling the building of high performance Intel OPA-based systems. This white paper describes Intel OPA storage solutions and discusses the considerations involved in selecting the best storage solution. It is aimed at the OEM solution architect and others interested in understanding storage options for Intel OPA-based systems.

Storage Overview

An Intel OPA-based system solution often requires connectivity to a parallel file system, enabling the Intel OPA-based compute nodes to access the file system storage. These storage solutions involve storage servers, routers, block storage devices, storage networks, and parallel file systems. This overview discusses the components that make up the storage solutions and describes some typical configurations.

Table of Contents

Figures

Tables

Storage Components

This section describes the terminology that will be used in this paper to discuss the components of a storage solution.

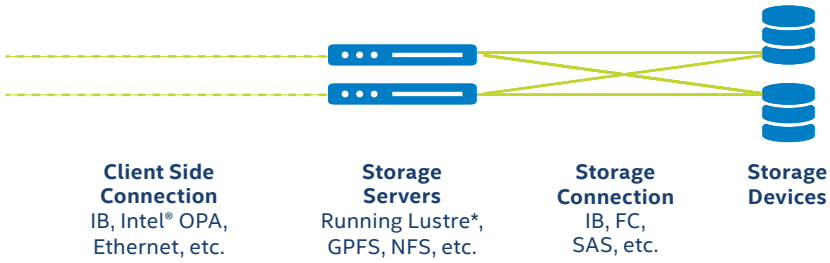

Figure 1: Storage Components

The client side connections are the connections to the compute cluster fabric.

The storage servers run the file system server software, such as Lustre* Object Storage Server (OSS) software, or General Parallel File System (GPFS) Network Shared Disk (NSD) software. Storage servers can take different forms. The storage servers are often implemented as standalone Linux* servers with adapter cards for connectivity to both the client side connections and the storage connections. These are sometimes productized by storage vendors, and in some cases the storage servers are integrated into an appliance offering.

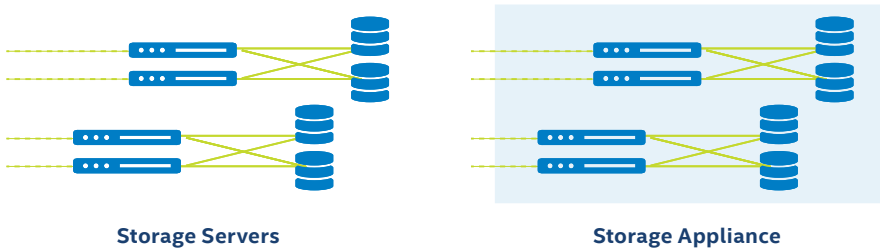

Figure 2: Storage Server Configurations

The storage connection and storage devices generally take the form of a block storage device offered by the storage vendors. It is expected that there will be block storage devices with Intel OPA storage connections; however, this is not critical to enabling Intel OPA storage solutions. The client side connection is the focus for Intel OPA enablement.

Storage Configurations

The connections from the storage servers to the Intel OPA fabric can take different forms, depending on the requirements of the system installation.

- **Direct attached** – the storage servers are directly attached to the Intel OPA fabric with Intel OPA adapter cards in the storage servers.
- **Dual-homed** – the storage servers are directly attached to the Intel OPA fabric and to another fabric, typically InfiniBand* (IB) or Ethernet. Adapter cards for both fabrics are installed in the storage servers.
- **Routed** – the storage servers are connected to the Intel OPA fabric through routers that carry traffic between the Intel OPA fabric to the client side connection of the storage servers, typically InfiniBand or Ethernet.

The dual-homed and routed configurations are typically used to provide connectivity to legacy storage or to share storage across multiple clusters with different fabrics.

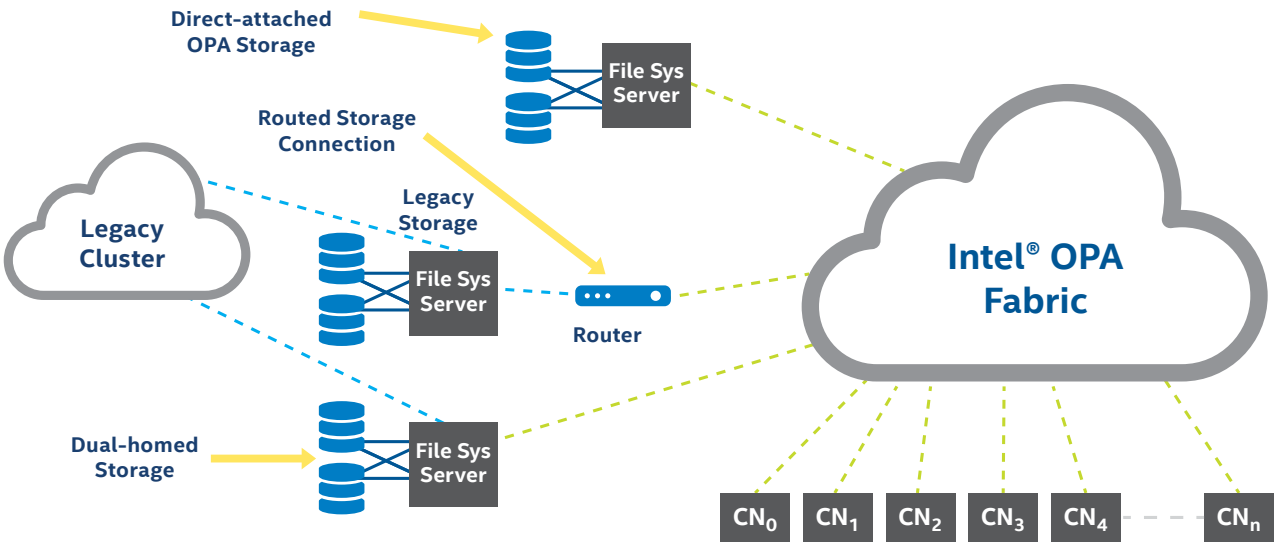

Figure 3: Storage Configurations Overview

Intel OPA Software and Storage Considerations

Intel OPA Host Software

Intel's host software strategy is to utilize the existing OpenFabrics Alliance interfaces, thus ensuring that today's application software written to those interfaces runs with Intel OPA with no code changes required. This immediately enables an ecosystem of applications to "just work." ISVs may over time choose to implement changes to take advantages of the unique capabilities present in Intel OPA to further optimize their offerings.

All of the Intel Omni-Path host software is open source. Intel is working with major operating system vendors to incorporate Intel OPA support into future releases. Prior to being in-box with these distributions, Intel will release a delta package to support Intel OPA. The Intel software will be available on the Intel® Download Center: https://downloadcenter.intel.com.

Table 1: Intel® OPA Linux* Support

LINUX* DISTRIBUTION	VERSIONS SUPPORTED
RedHat	RHEL 7.1 or newer
SuSE	SLES 12 SP0 or newer
CentOS	TBD version in 2016
Scientific Linux	TBD version in 2016

Table 2: Intel® OPA Lustre* Support

LUSTRE* DISTRIBUTION	VERSIONS SUPPORTING INTEL OPA
Community	2.8 or newer
Intel Foundation Edition	2.7.1 or newer
Intel Enterprise Edition	• 2.4 (client support only) • 3.0 or newer (client and server)

Table 3: Intel® OPA GPFS Support

GPFS SOFTWARE	VERSION SUPPORTING INTEL OPA
GPFS over IP	Supported with current versions of GPFS
GPFS over RDMA with Intel OPA	TBD, targeting 1H16

Intel OPA HFI coexistence with Mellanox* InfiniBand HCA

In a multi-homed file system server, or in a Lustre Networking (LNet) or IP router, a single OpenFabrics Alliance (OFA) software environment supporting both an Intel OPA HFI and a Mellanox* InfiniBand HCA is required. The OFA software stack is architected to support multiple targeted network types. Currently, the OFA stack simultaneously supports iWARP for Ethernet, RDMA over Converged Ethernet (RoCE), and InfiniBand networks, and the Intel OPA network has been added to that list. As the OS distributions implement their OFA stacks, it will be validated to simultaneously support both Intel OPA Host Fabric Adapters and Mellanox Host Channel Adapters.

Intel is working closely with the major Linux distributors, including Red Hat* and SUSE*, to ensure that Intel OPA support is integrated into their OFA implementation. Once this is accomplished, then simultaneous Mellanox InfiniBand and Intel OPA support will be present in the standard Linux distributions.

Once support is present, it may still be necessary to update the OFA software to resolve critical issues. Linux distribution support is provided by the operating system vendor. Operating system vendors are expected to provide the updates necessary to address issues with the OFA stack that must be resolved prior to the next official Linux distribution release. This is the way that software drivers for other interconnects, such as Ethernet, work as well.

With the Lustre versions mentioned above, the software doesn't yet have the ability to manage more than one set of Lustre settings in a single node. There is a patch to address this capability that is being tracked by https://jira.hpdd.intel.com/browse/LU-7101. With QDR and FDR InfiniBand, there are settings that work well for both IB and Intel OPA. With the advent of Enhanced Data Rate (EDR) InfiniBand, there isn't a set of settings that work well for both the InfiniBand and the Intel OPA devices. Therefore, coexistence of Intel OPA and EDR InfiniBand isn't recommended until that Lustre patch is available. Once the patch is available, it will be available for Lustre source builds, and also included in all future Intel Lustre releases as well as community releases. It is expected that there will be an IEEL version in 1Q16 that includes this patch.

Intel OPA Storage Solutions

Intel OPA Direct Attached Storage

High performance file systems with connectivity to an Intel OPA compute fabric, including Lustre and GPFS, are a core part of end-to-end Intel OPA solutions. When an Intel OPA based system requires new storage, there are several options available.

- **OEM/Customer built** – OEMs or end customers put together the file system, procuring block storage from storage vendors, selecting an appropriate server, and obtaining the file system software either from the open source community or from vendors that offer supported versions. This option is straightforward with Intel OPA. See the Intel OPA Host Software section above for information about OS and file system software versions compatible with Intel OPA.
- **Storage vendor offering** – In some cases, the complete file system solution is provided by a storage vendor. These complete solutions can take the form of block storage with external servers running the file system software or fully integrated appliance-type solutions.

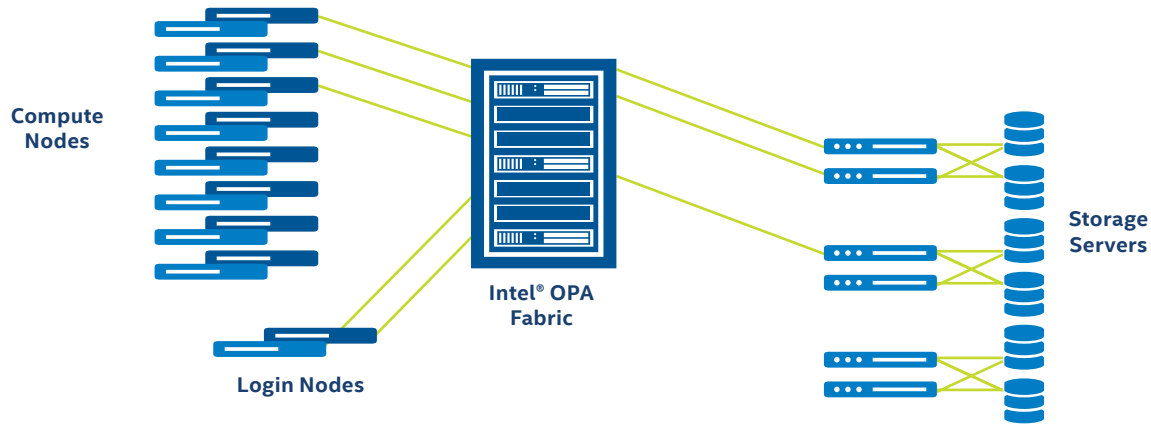

Figure 4: Direct Attached Storage

Interoperability with Existing Storage

In many cases, when a new Intel OPA-based system is deployed, there are requirements for access to existing storage. There are three main options to be considered:

- Upgrade existing file system to Intel OPA
- Dual-home the existing file system to Intel OPA
- LNet and IP Router solutions

If the existing file system can be upgraded to support Intel OPA connections, this is often the best solution. There are multiple viable options.

For cases where the existing storage will only be accessed by the Intel OPA based system, this upgrade can take the form of replacing existing fabric adapter cards with Intel OPA adapter cards. Software upgrades may also be required. See the Intel OPA Host Software section above for information about OS and file system software versions compatible with Intel OPA.

In other cases, the existing file system will need to continue to be accessed from an existing (non-Intel OPA) cluster and also will require access from the new Intel OPA-based system. In these cases, the file system can be upgraded to be dual homed by adding Intel OPA host adapters to the file system servers.

In some cases, there are requirements for accessing existing storage from new Intel OPA-based systems and it is not possible to upgrade the existing file system. This can happen if the customer is unwilling to tolerate either the down time that would be required to upgrade the existing file system or the risk in doing so.

In these cases, a router-based solution can solve the interoperability challenge. For Lustre file systems, the LNet router component of Lustre can be used for this purpose. For other file systems, such as GPFS, the Linux IP router can be used.

Table 4: Storage Options Considerations

	DIRECT-ATTACHED/DUAL HOMED	ROUTED SOLUTION
Pros	• Excellent bandwidth • Predictable performance • No additional system complexity	• Easy to add to existing storage • Minimal downtime of existing storage
Cons	• Requires downtime of the existing file system to perform software and hardware upgrades • May be viewed as higher risk	• Bandwidth requirements may mean that multiple routers are required • Complexity to manage extra pieces in the system

Dual-homed Storage

In the dual-homing approach, an Intel OPA connection is provided directly from the file system server, providing the best possible bandwidth and latency solution. This is a good option when:

- The file system servers have a PCIe slot available to add the Intel OPA adapters
- The file system servers utilize OS and file system software versions compatible with Intel OPA, or can be upgraded to do so

For OEMs and customers who have built the file system themselves, this solution will be supported through the OS and file system software arrangements that are already in place. When the file system solution was provided by a storage vendor, that vendor can be engaged to perform and support the upgrade to dual-homed.

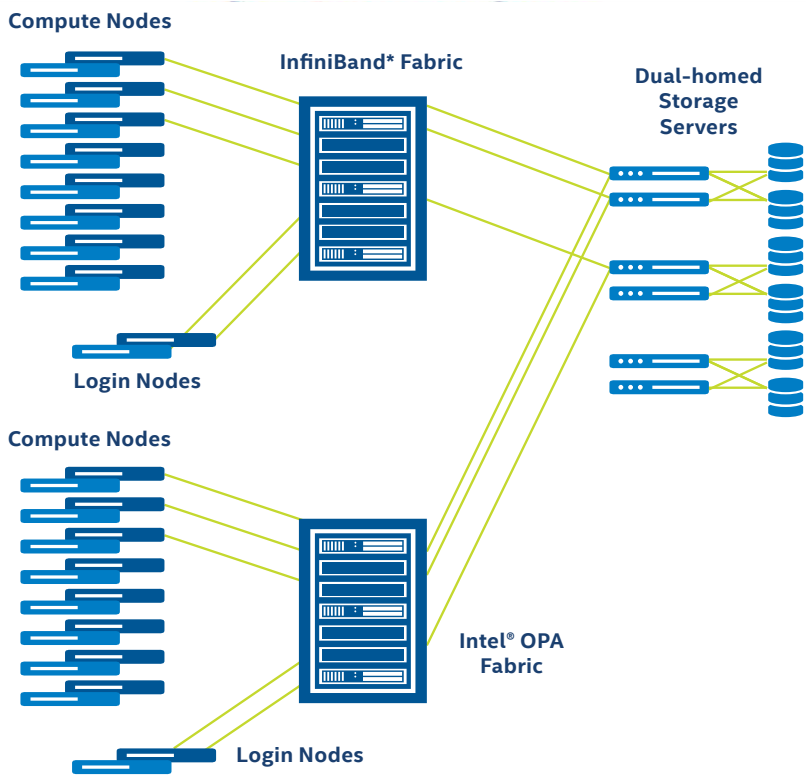

Figure 5: Dual-homed Configuration

LNet Router

The LNet router is a standard component of the Lustre stack. When configured to support routing between an Intel OPA fabric and a legacy fabric, it can provide routing of LNet storage traffic.

To facilitate the implementation of Lustre routers in Intel OPA deployments, a validated reference design recipe is provided. This recipe provides instructions on how to implement and configure LNet routers to connect Intel OPA and InfiniBand fabrics.

The Lustre software supports dynamic load sharing between multiple targets. This is handled in the client, which has a router table and does periodic pings to all of its end points to check status. This capability is leveraged to provide load balancing across multiple LNet routers. Round-robin load sharing is performed transparently. This capability also provides for failover because in the event of an LNet router failure, the load is automatically redistributed to other available routers.

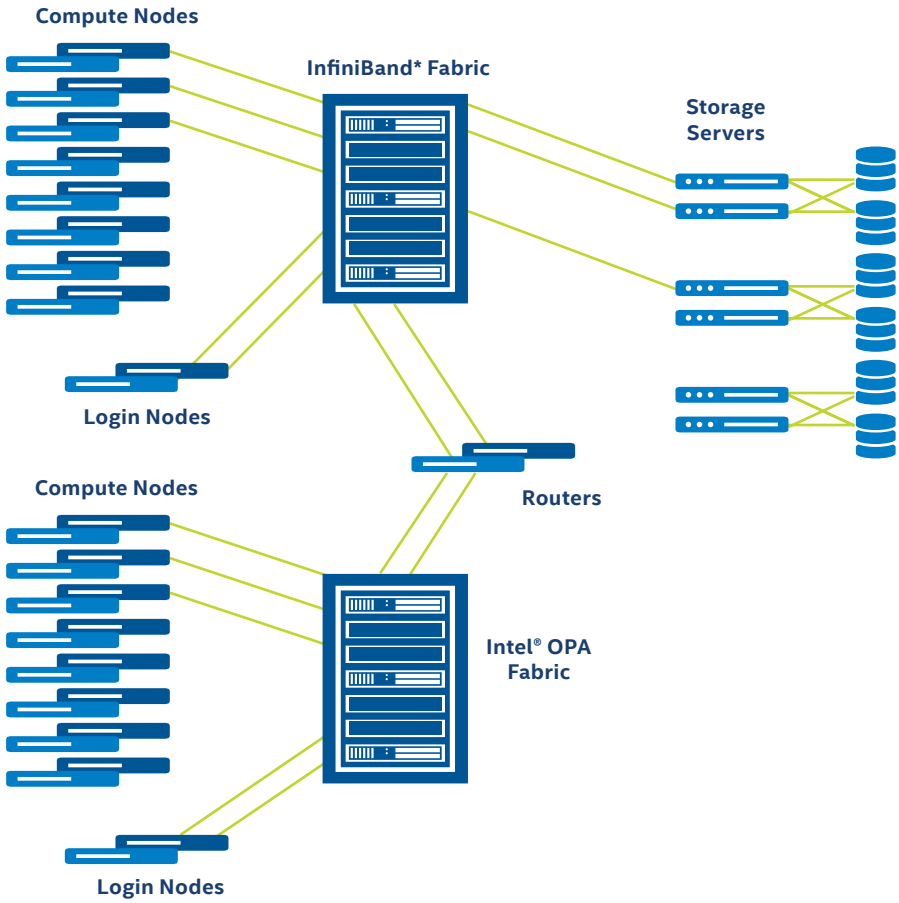

Figure 6: Router Configuration

Table 5: LNet Router Hardware Recipe

HARDWARE	RECOMMENDATION
CPU	Intel® Xeon® E5-2697 v3 (2.60 GHz, 1-2 core Turbo 3.6 GHz), 14 core
Memory	64 GByte RAM per node
Server Platform	1U rack server or equivalent form factor with two x16 PCIe* slots
Intel® OPA connection	One x16 Chippewa Forest HCA
IB/Ethernet connection	• Mellanox FDR • Mellanox EDR will be supported in the future • Ethernet – TBD

Table 6: LNet Router Software Recipe

SOFTWARE	RECOMMENDATION
Base OS	• RHEL 7.1 + Intel® OPA delta distribution OR • SLES 12 SP0 + Intel OPA delta distribution
Lustre*	• Community version 2.8 OR • IEEL 2.4 or newer OR • FE 2.7.1 or newer

Note: *The Net Router Hardware Recipe (Table 5) and the Net Router Software Recipe (Table 6) information is preliminary and based upon configurations that have been tested by Intel to date. Further optimization of CPU and memory requirements is planned.*

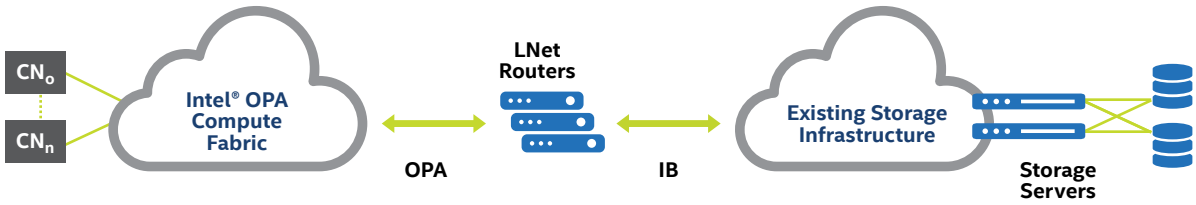

Figure 7: LNet Router

The target peak aggregate forwarding rate is 85 percent of the InfiniBand line rate or 47 Gbps per LNet router server with FDR InfiniBand when following the above recipe. Other generations of InfiniBand are expected to work with this recipe; however, performance will vary. Ethernet connectivity to storage is also supported by the recipe.

Additional bandwidth can be achieved by instantiating multiple LNet routers.

The target message rate is at least 5 Million messages per second.

The LNet router solution package will include:

- Hardware and software recipes for router solution (as shown here)
- Hardware requirement specifications
- Performance expectations
- Documentation/User Guides

Support for the LNet router solution is provided through the customer's Lustre support path, as it is a standard part of the Lustre software stack.

IP Router

The IP router is a standard component in Linux. When configured to support routing between an Intel OPA fabric and a legacy fabric, it provides IP based routing that can be used for GPFS, NFS, and other file systems that use IP based traffic.

To facilitate the implementation of IP routers in Intel OPA deployments, a validated reference design recipe is provided. This recipe provides instructions on how to implement and configure IP routers to connect Intel OPA and InfiniBand or Ethernet fabrics.

The Virtual Router Redundancy Protocol (VRRP) v3 software in Linux is used to enable failover and load balancing with the IP routers. The VRRP is a computer networking protocol that provides for automatic assignment of available Internet Protocol (IP) routers to participating hosts. This increases the availability and reliability of routing paths via automatic default gateway selections on an IP subnetwork.

IP routers can be configured for high availability using VRRP. This can be done with an active and a passive server. In a system configured with multiple routers, routers can be configured to be master on some subnets and slaves on the others, thus allowing the routers to be more fully utilized while still providing resiliency.

The load-balancing capability is provided by VRRP using IP Virtual Server (IPVS). IPVS implements transport-layer load balancing, usually called Layer 4 LAN switching, as part of the Linux kernel. IPVS is incorporated into the Linux Virtual Server (LVS), where it runs on a host and acts as a load balancer in front of a cluster of real servers. IPVS can direct requests for TCP- and UDP-based services to the real servers, and make services of the real servers appear as virtual services on a single IP address.

A weighted round-robin algorithm is used and different weights can been added to distribute load across file system servers that have different performance capabilities.

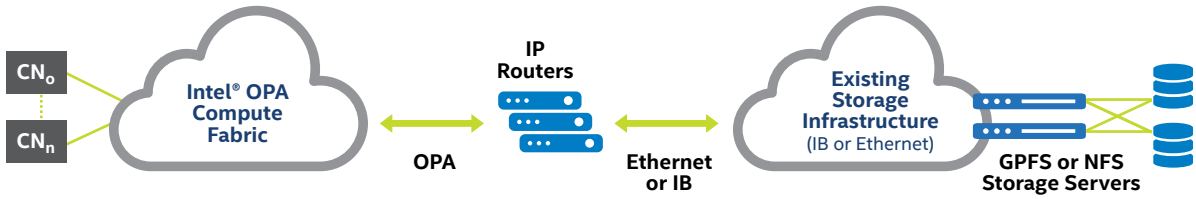

Figure 8: Router

Table 7: IP Router Hardware Recipe

HARDWARE	RECOMMENDATION
CPU	Dual-socket current or future generation Intel® Xeon® Processors which have their 1-2 core Turbo frequency above 3.3 GHz, and at least 10 cores per CPU **Example:** • Intel® Xeon® E5-2697 v3 (2.60 GHz, 1-2 core Turbo 3.6 GHz), 14 core
Memory	64 GByte RAM per node
Server Platform	1U rack server or equivalent form factor with two x16 PCIe slots
Intel® OPA connection	One x16 Chippewa Forest HCA
IB/Ethernet connection	• Mellanox FDR and EDR (Other generations are expected to work, performance will vary) • Ethernet - TBD

Table 8: IP Router Software Recipe

SOFTWARE	RECOMMENDATION
Base OS	• RHEL 7.1 + Intel® OPA delta distribution OR • SLES 12 SP0 + Intel OPA delta distribution

Note: The IP Router Hardware Recipe (Table 7) and the IP Router Software Recipe (Table 8) information is preliminary and based upon configurations that have been tested by Intel to date. Further optimization of CPU and memory is planned.

The target peak aggregate forwarding rate is 40 Gbps per IP router server with either EDR or FDR IB when following the above recipe. Other generations of InfiniBand are expected to work with this recipe, however performance will vary. Ethernet connectivity to storage is also supported by the recipe.

Additional bandwidth can be achieved by instantiating multiple IP routers.

The target message rate is at least 5 Million messages per second. The IP router solution package will include:

- Hardware and software recipes for router solution (as shown here)
- Hardware requirement specifications
- Performance expectations
- Documentation/User Guides

Support for the IP router solution is provided through the customer's Linux support path, as it is a standard part of the Linux software stack.

For more information about Intel Omni-Path Architecture and next-generation fabric technology, visit:
www.intel.com/hpcfabrics
www.intel.com/omnipath

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.

THE INFORMATION PROVIDED IN THIS PAPER IS INTENDED TO BE GENERAL IN NATURE AND IS NOT SPECIFIC GUIDANCE. RECOMMENDATIONS (INCLUDING POTENTIAL COST SAVINGS) ARE BASED UPON INTEL'S EXPERIENCE AND ARE ESTIMATES ONLY. INTEL DOES NOT GUARANTEE OR WARRANT OTHERS WILL OBTAIN SIMILAR RESULTS.

INFORMATION IN THIS DOCUMENT IS PROVIDED IN CONNECTION WITH INTEL PRODUCTS AND SERVICES. NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. EXCEPT AS PROVIDED IN INTEL'S TERMS AND CONDITIONS OF SALE FOR SUCH PRODUCTS AND SERVICES, INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO SALE AND/OR USE OF INTEL PRODUCTS AND SERVICES INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

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.

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

Copyright © 2015 Intel Corporation. All rights reserved. Intel, Intel Xeon, and the Intel logo are trademarks of Intel Corporation in the U.S. and other countries.

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

Printed in USA 1115/EON/HBD/PDF ♻ Please Recycle 333509-001US