Hewlett Packard Enterprise

HPE Scalable Storage with Intel Enterprise Edition for Lustre*

Featuring HPE Apollo 4520 System with EDR InfiniBand networking

Technical white paper

Contents

Executive summary

The HPE Scalable Storage Lustre* solution reference architecture is designed to remove the complexity of properly configuring an HPE Scalable Storage Lustre* solution using the HPE Apollo 4520 System and Intel® Enterprise Edition for Lustre* (Intel EE for Lustre*). This cluster configuration is designed especially for large-block sequential data sets at petabyte scale. It delivers maximum I/O bandwidth and is designed for high availability. The HPE Scalable Storage Lustre* solution delivers high-performance, purpose-built clustered storage that scales as your storage needs grow.

Target audience: It is for solution architects who are looking for information about HPE Lustre* solution on HPE Apollo 4520 Systems. This paper assumes a general knowledge of Lustre*, as well as typical HPC workloads requiring a high-performance parallel file system solution.

This white paper describes testing performed by Hewlett Packard Enterprise in June 2016.

Introduction

Reference architecture guidance

This paper describes a Lustre* solution assembled, tuned, and benchmarked by Hewlett Packard Enterprise. It describes component choices, cluster topology, features, and performance of this configuration. The configuration described in this document is intended to be viewed as a set of recommendations rather than the only supported configuration; however, the performance documented in this paper will not be guaranteed if a selected configuration deviates from these recommendations.

This paper will not attempt to describe in detail how to install and configure Intel EE for Lustre*. HPE installation support services will be provided to customers deploying production level HPE Scalable Storage Lustre* solutions.

Reference architecture structure

This document initially introduces the reader to the solution before progressing through a more in-depth description of the Intel EE for Lustre* software architecture and its features. Hardware components of the solution are described, focusing on the HPE Apollo 4520 System. Optional solution configurations are also described, with guidance regarding when you might want to select one of these options. Details of the specific reference architecture assembled for this paper include diagrams, a bill of materials for the components and required software licenses needed to replicate this solution are provided. Performance test results and information necessary to reproduce the tests will be provided. Appendices provide additional details about the solution.

Reference configuration summary

The HPE Scalable Storage Lustre* solution is a combination of HPE hardware components and Intel EE for Lustre* software. The core hardware component of the solution is the HPE Apollo 4520 System, which is purpose-built for software-defined, clustered storage.

This document describes a minimal Lustre* cluster and is composed of the following items. The solution scales by adding additional cluster nodes and storage.

- An HPE Apollo 4520 System containing two HPE ProLiant XL450 Gen9 Server nodes and 46X 8 TB 12 Gbps SAS drives
- Two HPE D6020 Disk Enclosures containing 70X 8 TB 12 Gbps SAS drives each
- Three HPE ProLiant DL360 Gen9 Servers
- An HPE MSA 2040 or an HPE D3700 Disk Enclosure (use of both are described) populated with 300 GB 12 Gbps SAS 15K rpm drives
- Intel EE for Lustre* version 3.0.0.0
- Red Hat® Enterprise Linux® 7.2 as the base OS installed on all of the cluster nodes
- 1GbE networking for HPE Integrated Lights-Out (iLO) and management traffic

In this solution, the HPE Apollo 4520 System serves as a redundant active/active pair of Lustre* object storage servers (OSSs). It hosts disk drives configured as Lustre* object storage targets (OSTs). The HPE D6020 Disk Enclosure hosts additional drives configured as OSTs. One of the HPE ProLiant DL360 Gen9 Servers is configured as an Intel Manager for Lustre* (IML) node, which is used to help manage and monitor the cluster. At this time, the IML supports only monitor mode for Lustre* configurations using ZFS as the underlying file system. The other two HPE ProLiant DL360 Gen9 Servers are configured as a redundant active/standby Lustre* management server (MGS) and as a Lustre* metadata server (MDS).

Regarding the high-performance Lustre* network (LNET), this document describes the use of EDR InfiniBand for the Lustre* OSSs which are configured to utilize the 100 Gbps bandwidth provided by EDR, and the use of 56 Gbps FDR InfiniBand for the Lustre* MGS and MDS servers which benefit more from low-latency IOPS than from additional bandwidth.

This document describes two options for the Lustre* metadata target (MDT). One utilizes the HPE MSA 2040. The other option uses an HPE D3700 Enclosure. The benefits and trade-offs of each option are described in this document. More details describing the roles of the various Lustre* servers and targets will be provided in the next section.

The software aspect of this reference architecture is Intel EE for Lustre* installed on Red Hat Enterprise Linux 7.2 operating system. EDR and FDR InfiniBand are used as the high-performance, low-latency cluster network for this paper. It is worth noting that the solution architecture does not exclude other high-speed network technologies for which Lustre* Network Drivers (LNDs) exist.

Overview

Solution architecture

Today's applications demand a storage solution that is capable of handling extremely large amounts of data and huge numbers of files shared concurrently across clustered servers while also providing a POSIX-compliant file system interface. The Lustre* file system is an ideally distributed, parallel file system for high-performance computing. With Intel EE for Lustre* software, Intel provides a commercial-grade version of Lustre* optimized to address the key storage and data throughput challenges of HPC.

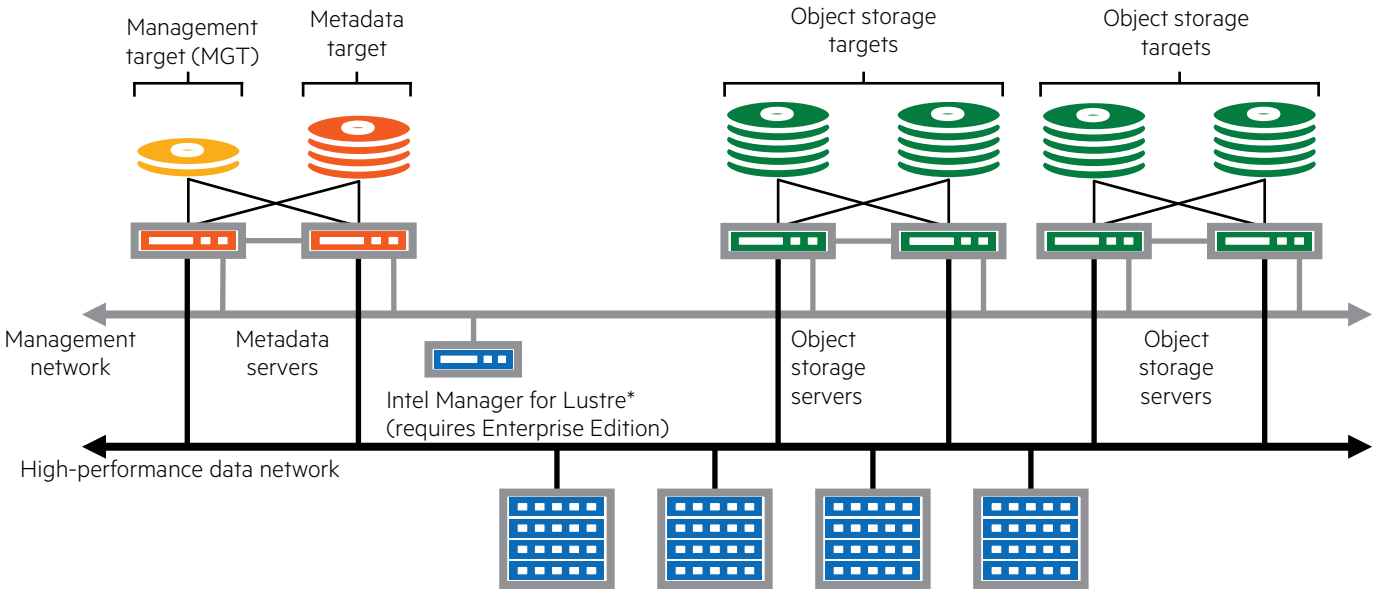

Figure 1. Generic Intel EE for Lustre* cluster diagram

The high-level configuration of an HA file system running Intel EE for Lustre* software consists of the following:

- IML: A dedicated manager server that hosts the IML software and dashboard.
- MGS: The MGS provides access to the Lustre* management target (MGT) storage. To support failover, the MGS is also configured as the backup for the MDS.
- MDS: The MDS provides access to the MDT storage. To support failover, the MDS is also configured as the backup management server.
- OSS: At least two servers provide access to the OSTs, which store the file system's data. OSSs are configured in failover pairs sharing access to the same data storage so that if an OSS fails, service is automatically failed over to its peer server.

- MGT: The MGT stores configuration information for all the Lustre* file systems in a cluster and provides this information to other Lustre* components. The MGT is accessed by the primary MGS and, if the MGS fails, by the MDS operating in failover mode. The MGT need not be larger than 10 GB in capacity.
- MDT: The MDT stores metadata (such as file names, directories, permissions, and file layout) for attached storage and makes them available to clients. The MDT is accessed by the primary MDS and, if the MDS fails, by the MGS operating in failover mode.
- OST: Client file system data is stored in one or more objects that are located on separate OSTs. The number of objects per file is configurable by the user and can be tuned to optimize performance for a given workload.
- Management network: The management network is 1GbE, connecting every server in the file system. This network is used with secure shell (SSH) to install and update IML software on each server. It is also used to manage the servers and make separate connections to the HPE iLO 4 port on each managed server.
- Lustre* network: The Lustre* network (LNET) provides high-speed file system access for each client. The file system size, the number of clients, and the average throughput requirements for each client drives the required data rate of this network.

Intel EE for Lustre* software unleashes the performance and scalability of the Lustre* parallel file system as an enterprise platform for organizations both large and small. It allows businesses that need large-scale, high-bandwidth storage to tap into the power and scalability of Lustre*—with additional features and capabilities including worldwide 24x7 technical support from the Lustre* experts at Intel.

Figure 2. IML performance monitoring

IML simplifies installation, configuration, and monitoring. This purpose-built management solution is a key component of Intel EE for Lustre* software. It reduces management complexity and costs, enabling storage administrators to exploit the performance and scalability of Lustre* storage. The administrator's dashboard speeds real-time monitoring including tracking usage, performance metrics, events, and errors at the Lustre* software layer.

The HPE Storage Plugin for Intel Manager for Lustre* provides additional insight into the status of underlying hardware and automatically triggers alerts when hard disks or controllers in the I/O path experience a failure. The plug-in provides all the information an administrator needs to manage their hardware both remotely and physically.

Why this solution technology?

Intel EE for Lustre* can run ZFS on Linux as the underlying back-end file system. ZFS is a robust, scalable file system with features not available in other file systems such as the patched version of ext4, commonly referred to as ldiskfs, which is the conventional choice of Lustre* back-end file system. ZFS enables a cheaper storage solution for Lustre* and increases the reliability of data for the next generation of fat HDDs.

- L2ARC—uses SSD devices to speed up random and small files read I/O.
- Copy on write (COW)—helps to ensure that write requests are aligned when they reach the storage hardware, improving write performance.
- Checksum on data block—in conjunction with the LNET, checksum provides end-to-end data protection for Lustre*.
- Always consistent on disk—waiting for a very long, offline file system consistency check is no longer needed.

- Resilvering—very fast rebuild times, which are based on the space used on the disk rather than the size of the disk.
- Scrubbing—automatic and online data correction.
- Manageability—simple to manage, simple to troubleshoot.
- Compression—compression can be enabled to maximize the ROI and potentially improve file system performance.
- Efficient snapshotting—enables Lustre* to create snapshots of the entire file system.
- RAID drive failure resiliency—enables efficient JBOD solutions.

Features of the HPE Apollo 4520 System

- **Simplified high availability**—two server nodes in the chassis make an ideal cluster in a box, allowing two Lustre* OSSs to be configured as a failover pair.
- **Dual domain SAS connectivity**—all disk drives internal to the HPE Apollo 4520 System are accessible by either server node, providing support for Lustre* OST failover in high-availability configurations.
- **Dense storage capacity**—over 368 TB raw capacity per 4U system using 8 TB disk drives, or up to 1488 TB per Lustre* OSS pair when two HPE D6020 Disk Enclosures are attached to the HPE Apollo 4520 System.
- **Flexible I/O configurations**—the HPE Apollo family of servers offers two I/O module options, each having two 1GbE ports. One option offers four low profile x8 PCIe Gen3 slots along with support for an additional FlexibleLOM. The other option offers three x8 PCIe Gen3 slots and a single x16 PCIe Gen3 slot in order to support high-bandwidth network options such as adapters for EDR InfiniBand and Intel OPA.
- **Power management**—the HPE Advanced Power Manager 2.0 provides dynamic power capping and asset management features that are standard across the HPE Apollo line. The converged HPE Apollo System chassis also yields power savings via shared cooling and power resources.
- **Solution integration and data center acceptance**—HPE hardware described in this white paper has been qualified together. This means no work building, maintaining, and qualifying white box architectures for the cluster. HPE hardware can be validated with confidence.
- **Enterprise support**—get dedicated solution and support resources from Hewlett Packard Enterprise, a trusted enterprise partner. At massive scale, system failures become part of the design even with the most reliable components. Therefore, it is critical to have good support infrastructure to keep system reliability and availability at acceptable levels.
- **Enterprise-class storage components**—HPE Smart HBAs provide a robust storage solution within the server. Hewlett Packard Enterprise also qualifies and supports hard drives to minimize failure impacts.
- **HPE iLO 4**—is an industry-leading embedded monitoring solution. Its agentless management, diagnostic tools, and remote support allow entire data centers to be managed with ease.
- **Enterprise-class management**—HPE Insight Cluster Management Utility is an efficient, customizable, and robust hyperscale-cluster lifecycle management framework and suite of tools for managing operations and performance of large clusters such as those found in HPC.

Solution diagram

The reference architecture described in this document is depicted in figure 3. The HPE Apollo 4520 System provides a pair of Lustre* OSSs and OSTs. HPE D6020 Disk Enclosure provides expansion for additional OSTs. A pair of HPE ProLiant DL360 Gen9 Servers is used for the Lustre* MGS and for the Lustre* MDS, each acting as a standby failover server for the other.

This architecture describes two options for Lustre* MDTs, one using the HPE MSA 2040 Storage array and the other using HPE D3700 Disk Enclosure. Another HPE ProLiant DL360 Gen9 Server is used for the IML. Finally, the HPE Apollo 6000 System with HPE ProLiant XL230a Gen9 Servers were used as workload clients for benchmark testing.

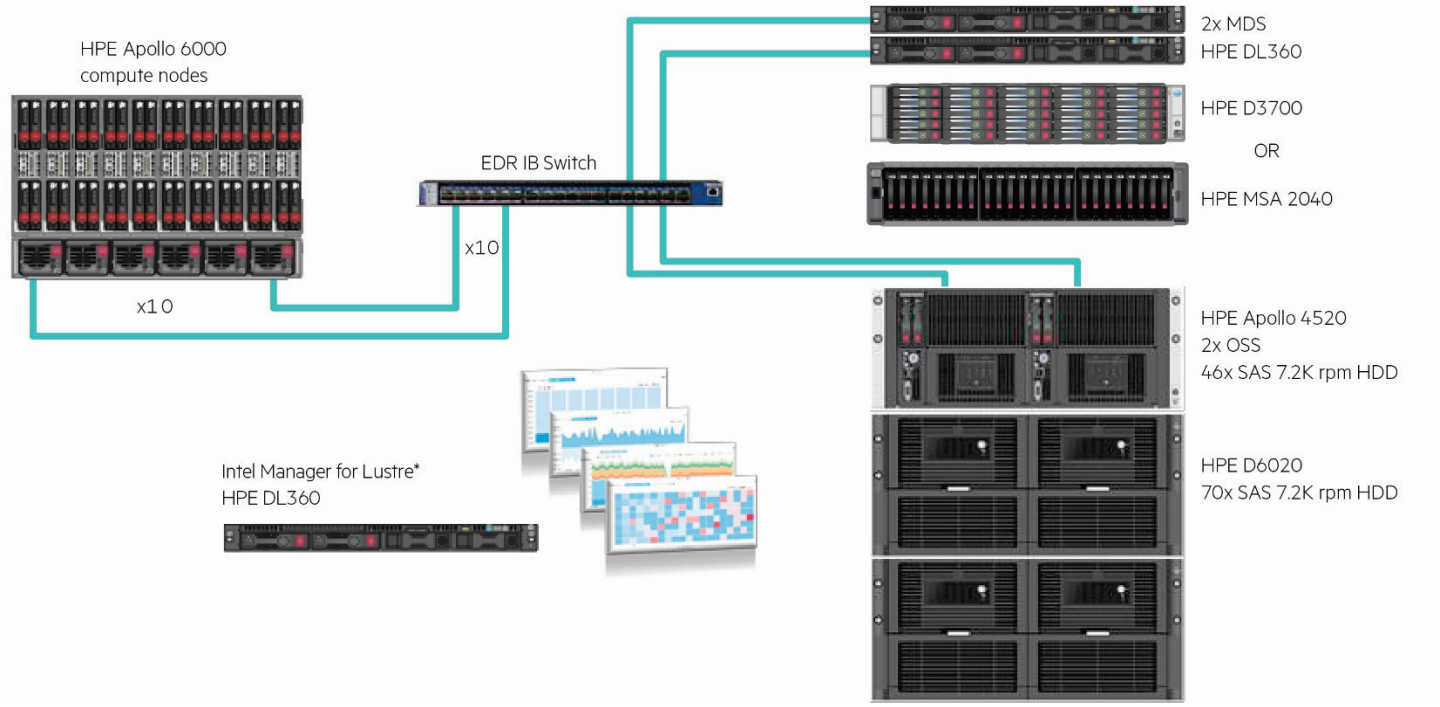

Figure 3. Hardware components in the reference architecture

Solution components

Component choices

Server selection

HPE Apollo 4520 System

HPE Apollo 4520 System is used for the Lustre* OSSs in this reference architecture. With its two HPE ProLiant XL450 Gen9 Servers and dual-domain capability, the HPE Apollo 4520 System provides both servers access to all 46 of the internal SAS data drives. This makes it a perfect platform for a highly available pair of clustered Lustre* OSSs. Three x8 and one x16 PCIe Gen3 slots provide a variety of configuration options for high-performance networking and external storage expansion.

Figure 4. HPE Apollo 4520 System

Each server supports dual Intel Broadwell-EP Xeon E5-2600 v4 processors and up to 16 memory slots (eight per processor), supporting registered DDR4-2400 DIMM with ECC.

HPE DL360 Gen9 Server

The 1U HPE ProLiant DL360 Gen9 Server is a dual socket server, with a choice of Intel® Xeon® E5-2600 v4 processors, up to 1.5 TB of memory in 24 DIMM slots and three expansion slots when populated with two processors. Many options are available for network connectivity using either FlexibleLOM or PCIe network adapters.

Figure 5. HPE ProLiant DL360 Gen9 Server

The HPE ProLiant DL360 Gen9 Server is used in this reference architecture to host the Lustre* metadata and management servers, as well as the Intel Manager for Lustre*. It provides the flexibility to configure a low-cost, lightly provisioned server for the IML. Alternatively, it can also be more generously provisioned with processing power and memory to meet the requirements of a high-performance metadata server. It provides sufficient slots and networking options in one rack unit to support high-performance network connections such as InfiniBand, while also providing slots for an HBA to connect to an external shared metadata target.

Compute

Both server nodes of the HPE Apollo 4520 System have dual Intel Xeon E5-2630L v4 CPUs installed. At 1.80 GHz and 10 cores, the E5-2630L v4 has sufficient processing capability to keep up with the demands of RAID 6 ZFS computations. Both processor sockets are populated to enable the use of all the available PCIe slots.

HPE ProLiant DL360 Gen9 Servers used as Lustre* management and metadata servers are each provisioned with dual Intel Xeon E5-2643 v4 CPUs, which at 3.40 GHz and 6 cores, enable these servers to keep up with the demands of compute clients making file metadata requests. The Lustre* MGS is typically configured to be the high-availability standby for the metadata server, so it should be provisioned to be just as capable as the primary metadata server.

The requirements of the Intel Manager for Lustre* are pretty lightweight for processing, so Intel Xeon E5-2603 v4 with 6 cores at 1.6 GHz are selected for the HPE ProLiant DL360 Gen9 Server hosting this service. Other choices for the IML processor will not impact the cluster's performance.

Memory

Our HPE ProLiant XL450 Gen9 Server nodes are each populated with 256 GB of DDR4-2400 Registered HPE SmartMemory. Memory on the Lustre* OSSs can be used for write buffering to coalesce sequential requests to the disk drives. Use of 2-stage commits by Lustre* for writes make this a safe operation. Memory is also used as an adaptive read cache and for read-ahead caching. Increasing the memory on these nodes improves performance for most application workloads.

The Intel Manager for Lustre* has very minimal requirements for memory, so the HPE ProLiant DL360 Gen9 Server we use for this purpose is populated with only 16 GB of DDR4-2400 memory.

Object storage

Storage is one of the key components of this solution. Disk drives used in configuring Lustre* OSTs are found within the HPE Apollo 4520 System and optionally in one or more HPE D6020 Disk Enclosures. The HPE Apollo 4520 System has 46 large form factor (LFF) drive bays. For this reference architecture, we have populated the bays with HPE 8TB, 12G SAS, 7200 rpm LFF 512e midline (MDL) hard drives. These drives provide excellent performance and capacity, along with support for dual domain, which is required to configure a highly available cluster.

Figure 6. HPE Apollo 4520 System—top view

The 12 Gbps SAS HPE D6020 Disk Enclosure is designed with two pull-out drawers to support 70 hot plug LFF SAS or SAS MDL drives in just 5U of rack space.

This reference architecture utilizes two HPE D6020 Disk Enclosures, populated with the same drive model described earlier for the HPE Apollo 4520 System. Use of two HPE D6020 Disk Enclosures, along with HPE Apollo 4520 System will fully utilize the bandwidth of EDR InfiniBand Lustre* network connections in the HPE ProLiant XL450 Gen9 Server nodes.

Figure 7. HPE D6020 Disk Enclosure

Hewlett Packard Enterprise currently offers 12 Gbps SAS, 7200 rpm MDL drives in 2, 4, 6, and 8 TB capacities. When using an HPE Apollo 4520 System with one or two attached HPE D6020 Disk Enclosures as your Lustre* object storage building block, these options result in the following raw and usable capacities. The difference between the two is capacity utilized for parity data and for spare drives.

Table 1. Available storage capacities

	Apollo 4520		Apollo 4520 + D6020		Apollo 4520 + 2x D6020	
	Raw capacity	**Usable capacity**	**Raw capacity**	**Usable capacity**	**Raw capacity**	**Usable capacity**
2 TB HDDs	92 TB	72 TB	232 TB	180 TB	372 TB	288 TB
4 TB HDDs	184 TB	144 TB	464 TB	360 TB	744 TB	576 TB
6 TB HDDs	276 TB	216 TB	696 TB	540 TB	1116 TB	864 TB
8 TB HDDs	368 TB	288 TB	928 TB	720 TB	1488 TB	1152 TB

The densest capacity configuration per rack is three blocks of the HPE Apollo 4520 System with two HPE D6020 Disk Enclosures each. This configuration will support raw capacity of 4464 TB and file system usable capacity of 3058 TiB in a 42U rack.

Note: Disk manufacturers calculate drive capacity using multiples of 1,000-byte kilobytes, while the file system reports capacity in 1,024-byte kibibytes. The calculation for usable capacity takes into account overheads, including data parity, file system reserve space, and the difference between TB vs. TiB.

Metadata storage

This white paper describes two options that are available for MDS in this reference architecture. First, the HPE MSA 2040 Storage array with dual 12 Gbps SAS controllers provides excellent performance and data protection. It is ideal for environments where the workload will involve large compute clusters doing simultaneous directory and file operations, typically on large numbers of relatively small files. The implication is that this environment will place high demands on the Lustre* metadata server and target while creating, opening, and closing directories and files.

Figure 8. HPE MSA 2040 Storage array

For this reference architecture, the HPE MSA 2040 Storage array is populated with 24X 300GB 12G SAS 15K rpm hard drives. These drives are configured as a RAID 10 array, providing fault tolerance against drive failure and maximizing performance. This configuration supports over 900 million files in the Lustre* file system. If support for greater numbers of files is required, consider using larger capacity drives, or using multiple Lustre* metadata targets.

Optionally, this reference architecture describes a solution using the HPE D3700 Disk Enclosure populated with 25X 300GB 12G SAS 15K rpm hard drives. This option is required to take advantage of the new snapshot feature of Intel EE for Lustre*, as well as for other benefits that come with the use of ZFS as the underlying file system for the MDT such as bit rot detection and compression. For this option, we used ZFS volume management to configure a RAID 10 MDT and a RAID 1 MGT.

Figure 9. HPE D3700 Disk Enclosure

The option of using the HPE D3700 does not perform as well as the option using the HPE MSA 2040 for the application workload described at the beginning of this section. It is, however, less expensive and will provide adequate performance for application workloads that require fewer simultaneous file metadata operations. These applications typically stream large amounts of write or read data into relatively few files at a time.

Performances of both options are provided in the Workload testing section of this document.

EDR InfiniBand Network option

The HPE Apollo 4520 OSS nodes in the reference architecture cluster are equipped with HPE InfiniBand EDR/Ethernet 100Gb 1P 840QSFP28 Adapters. Configured for InfiniBand connections, each port delivers up to 100 Gbps EDR for performance-driven server and storage clustering applications.

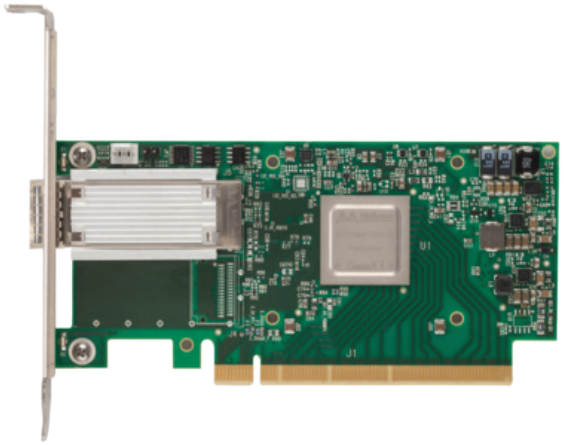

Figure 10. HPE InfiniBand EDR/Ethernet 100Gb 1P 840QSFP28 Adapter

EDR Switch option

For InfiniBand connectivity, a 648-Port Mellanox CS7500 EDR switch was used for testing this reference architecture. In the test configuration, both OSS servers (HPE Apollo 4520 System) were configured with EDR InfiniBand adapters, with the other cluster nodes configured with FDR InfiniBand adapters.

Image provided courtesy of Mellanox Technologies

Figure 11. Mellanox CS7500 648-Port EDR 100Gb/s InfiniBand Director Switch

Ethernet Switch

For connections to the "management" network and for iLO, an HPE 3500-48G-PoE+ yl Switch was used for testing. No traffic related to the benchmark results traverses this switch, so other 1GbE switch models can be substituted.

Figure 12. HPE 3500-48G-PoE+ yl Switch

The HPE 3500 yl Switch Series provides performance and ease of use for enterprise-edge and branch office deployments that require PoE+. Based on an HPE Networking ASIC, the 3500 yl Switch Series has low latency with increased packet buffering and an optional module for 10GbE ports.

It has a robust Layer-3 feature set with IPv4 BGP, policy-based routing, and IPv4/IPv6 OSPF, which can be seamlessly managed with the optional HPE Intelligent Management Center (IMC) software platform. Software-defined networking (SDN) is ready with support for OpenFlow.

Licensing and support

Each OSS and each active MDS in the Lustre* cluster will require an Intel EE for Lustre* software support license. Using this reference architecture as an example, three licenses would be required. Two for the HPE Apollo 4520 System, and one for the HPE ProLiant DL360 Gen9 Server used as the active MDS. HPE Installation Services will also be required for this solution.

Table 2. Intel EE Lustre* software support licenses

Part number	Description
P9L65AAE	Intel EE Lustre* SW L1–3 1yr E-LTU (worldwide)
P9L67AAE	Intel EE Lustre* SW L1–3 3yr E-LTU (worldwide)
P9L68AAE	Intel EE Lustre* SW L1–3 5yr E-LTU (worldwide)
P9L69AAE	Intel EE Lustre* SW EM L1–3 1yr E-LTU China and India only
P9L70AAE	Intel EE Lustre* SW EM L1–3 3yr E-LTU China and India only

If used, licenses for Red Hat Enterprise Linux will also be required on all nodes. In addition, a license for Red Hat High Availability add-on is required for all but Intel Manager for Lustre* node.

Workload testing

A principal guideline for this paper is to create a performance result that is simple to replicate. Readily available open source benchmark tools that are commonly used for Lustre* benchmarking have been employed to produce the results described in this paper.

Workload description

Tests have been performed on this reference architecture to determine the I/O bandwidth capabilities of the HPE Scalable Storage Lustre* solution, as well as to determine the number of directory and file operations the solution is capable of sustaining per second. As described previously in this document, the HPE Scalable Storage Lustre* solution offers two options for the Lustre* MDT. One is focused on supporting the maximum number of directory and file operations per second and utilizes the HPE MSA 2040 Storage array for the MDT. The other option uses the HPE D3700 Disk Enclosure that supports the full set of features offered by the solution, such as system snapshots.

Performance results for both options are offered in this paper.

All test results shown in this paper were achieved running the benchmark test against an empty file system. Results may vary if these tests are run against a file system that already contains data.

Efforts have been made to minimize the effects of caching by the Lustre* file system. For IOR bandwidth testing, the aggregate size of the test data sets has been calculated to be two to four times the size of memory available to the object servers. Cache on the client side has been explicitly dropped between successive tests.

No tuning of Lustre* client settings was performed in order to improve the results obtained by this testing.

Client configuration

All workload clients used for this testing were HPE ProLiant XL230a Gen9 server nodes in an HPE Apollo 6000 System. Each was equipped with dual Intel Xeon E5-2698 v3 16-core processors and 128 GB (8 x 16 GB) of dual-rank DDR4-2133 RDIMM. The boot device was a pair of mirrored 300 GB 15K rpm SAS drives.

InfiniBand FDR InfiniBand Connect-IB adapters were used for the LNET connections for the EDR workload test.

These nodes were installed with Red Hat ES 6.6, MLNX_OFED_LINUX-3.1-1.0.3, and IEEL client version 3.0.

Workload generator tools

IOR was used to perform client-based bandwidth tests against the HPE Scalable Storage Lustre* solution. IOR utilizes message-passing interface (MPI) to facilitate parallel operation among multiple workload clients. For this testing, we used IOR version 2.10.3 and Intel MPI Library for Linux, version 5.1.2 Build 10151015. IOR generates sequential write and read I/O from the workload clients, but with multiple clients executing multiple processes each, I/O will not typically be entirely sequential at the Lustre* OST or disk-drive levels.

Client-based metadata operation benchmark tests were performed using **mdtest**. It also utilizes MPI to facilitate parallel operations from multiple workload clients. **mdtest** has options to test directory creation, stat, and removal, as well as file operations such as creation, stat, read, and removal. Results of this test reflect the file system's ability to handle large numbers of application processes doing many concurrent directory and file operations. The results presented in this paper were obtained using mdtest version 1.9.3.

Both IOR and mdtest support a wide variety of options to customize the test to be executed. Options that were used for our benchmarking will be provided along with the results.

IOR may be obtained from sourceforge.net/projects/ior-sio/ and **mdtest** from sourceforge.net/projects/mdtest/.

Workload results and analysis

Bandwidth tests

This first set of charts display results of IOR testing against a Lustre* storage building block configured for both maximum capacity and bandwidth—a pair of object storage servers in an HPE Apollo 4520 System with two attached HPE D6020 Disk Enclosures. Storage for these OSSs is configured into 16 OSTs. Performance and capacity scales when additional blocks of these devices are added to the file system.

Tests were executed from one to 32 HPE ProLiant XL230a Gen9 Server nodes. The number of process threads per node was set to 16 in order to represent a reasonable number of cores per node in a compute cluster.

Additional XL230a nodes with 16 client processes were added to the test, scaling the total client processes for each test point. The IOR test was set for three iterations at each test point. The graph in figure 13 shows both the minimum and maximum values achieved in the three iterations for each data point.

The command line used with IOR uses these test parameters. The block size is adjusted for each test, which results in an aggregate data set size of 1,024 GiB. These tests were all executed using the file-per-process option.

```
IOR -posix -vv -i 3 -w -r -k -F -g -C -e -b <blksize> -t 1m -o /testfs01/ior-test/tmpfile
```

Note that the "-e" option is used on each write test to enable use of fsync, and the aggregate data set is sized to be twice the physical memory of the combined Lustre* OSSs to minimize or remove the advantage of server-side cache. In addition, there are three iterations for each IOR run and each run does a repeating **write, read, write** pattern so that test always reads new data.

Bandwidth values are expressed in MiB/s, which is 1,024x1,024.

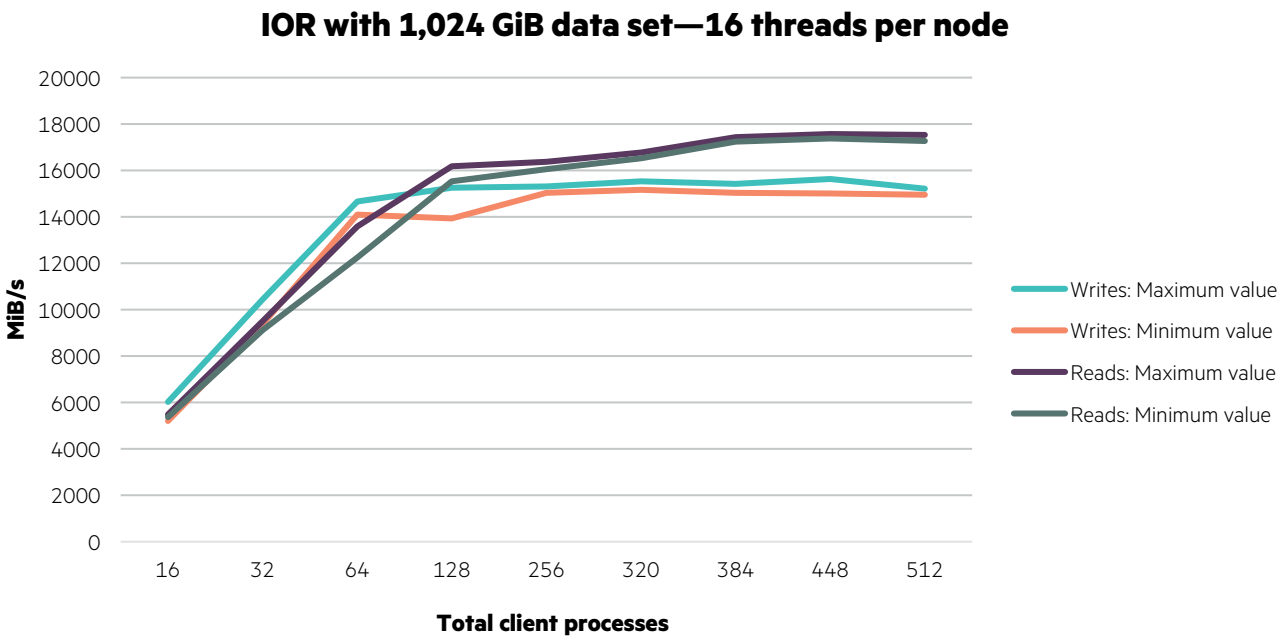

Figure 13. EDR option: IOR performance, 16 threads per node

In this test, writes peaked at 15,630 MiB/s and reads peaked at 17, 568 MiB/s.

The following chart shows performance for a single ProLiant XL230a Gen9 Server node. The number of processes was scaled from one to 32. Again, three iterations of each test were performed with the block size calculated to result in an aggregate data set size of 1,024 GiB or greater. For a single node configured with FDR InfiniBand, writes peaked at 6,049 MiB/s and reads peaked at 5, 797 MiB/s.

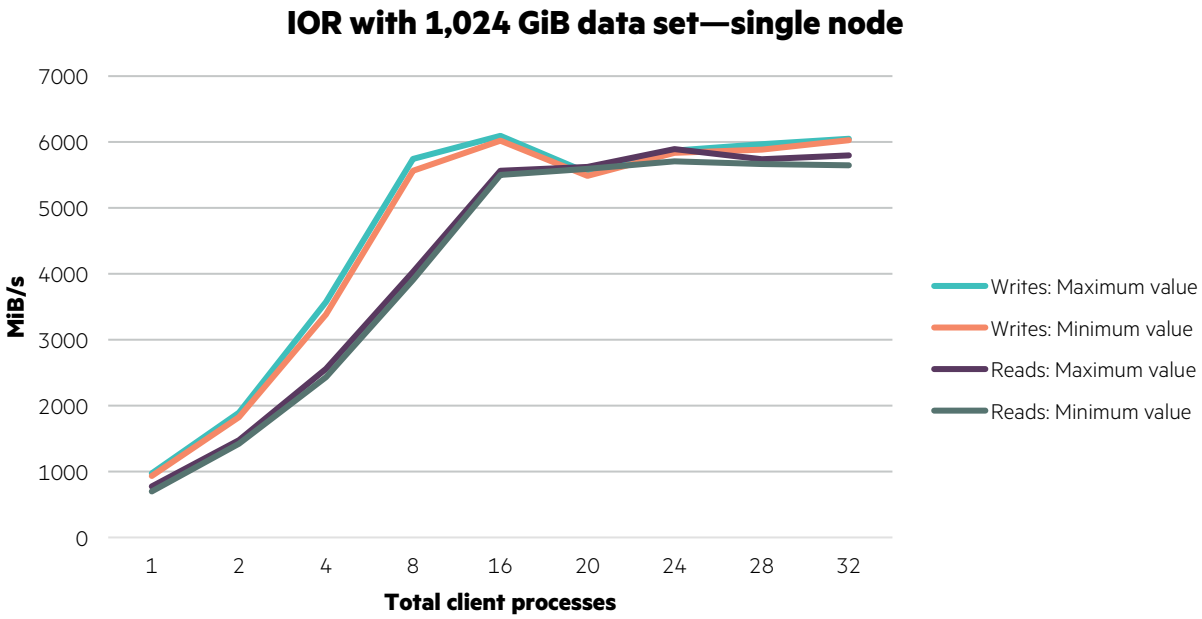

Figure 14. IOR performance, single node

Similar testing has been performed where the tests were limited to only those OSTs configured from drives within the HPE Apollo 4520 System, as well as tests limited to those OSTs configured within the HPE Apollo 4520 System and a single attached HPE D6020 Disk Enclosure.

Tests against the four OSTs within the HPE Apollo 4520 System resulted in peak writes of 6737 MiB/s (minimum writes of 6609 MiB/s) and peak reads of 6891 MiB/s (minimum reads of 6857 MiB/s) with an IOR test of three iterations.

Tests against the 10 OSTs within the HPE Apollo 4520 System and a single HPE D6020 Disk Enclosure resulted in peak writes of 12,332 MiB/s (minimum writes of 12,316 MiB/s) and peak reads of 15,049 MiB/s (minimum reads of 14,778 MiB/s) with an IOR test of three iterations.

Metadata performance tests

This document has described two different options for the Lustre* metadata target (MDT). One utilizes the HPE MSA 2040 Storage array and the other an HPE D3700 Disk Enclosure. The features of each have already been described in this document. mdtest, using the same test parameters, has been executed against the same test cluster, with the exception that the two MDT options have been tested in the cluster independently.

The chart in figure 15 shows results when tests were executed against the HPE MSA 2040 Storage array MDT. The test parameters used for mdtest were as follows. Workload clients were added in multiples of five for each test, with each executing 16 client processes. Values shown in the chart are the lowest of five iterations for each test point.

```
mdtest -d /testfs01/mdtest -i 5 -b 1 -z 1 -L -I 1000  -u -t
```

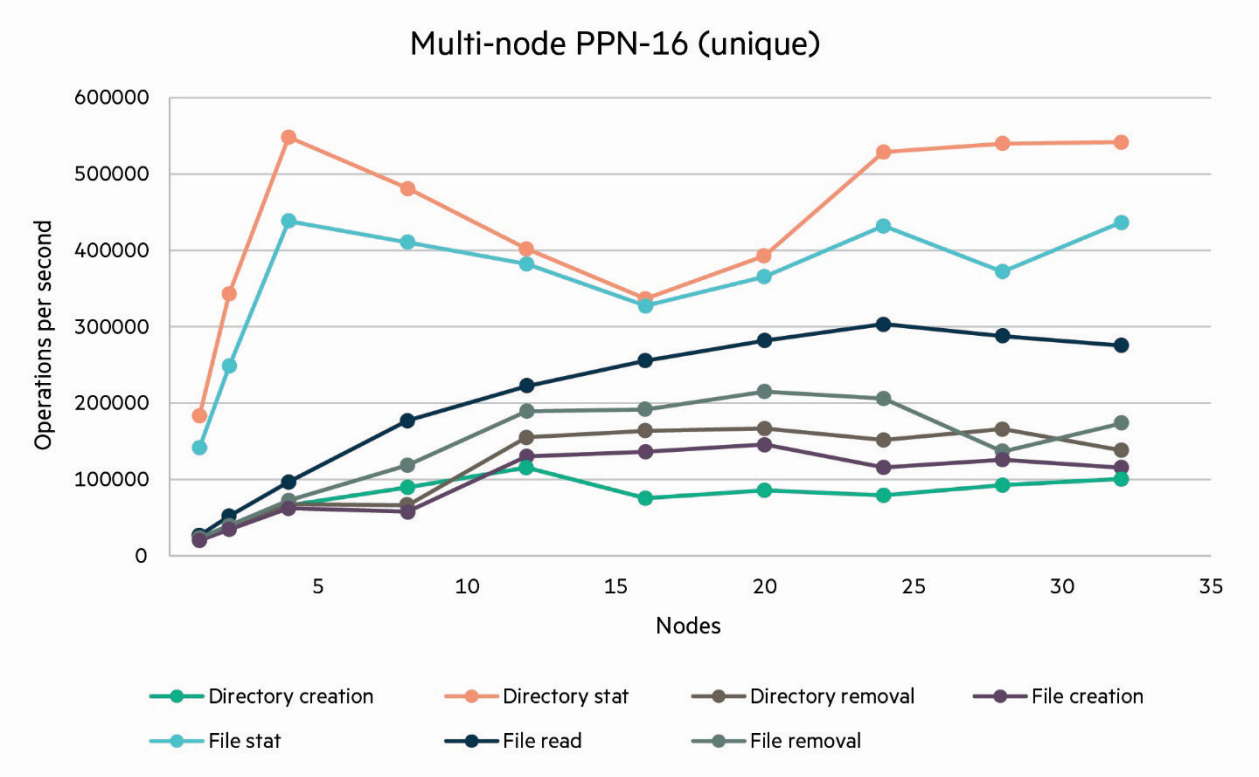

Figure 15. Metadata performance using HPE MSA 2040 Storage array for MDT

We then replaced the HPE MSA 2040 Storage array with the HPE D3700 Disk Enclosure and configured the equivalent of a RAID 10 MDT. The JBOD contained 25X 300 GB 12G SAS 15K rpm hard drives. The same test parameters have been used as with the other MDT.

```
mdtest -d /testfs01/mdtest -i 5 -b 1 -z 1 -L -I 1000  -u -t
```

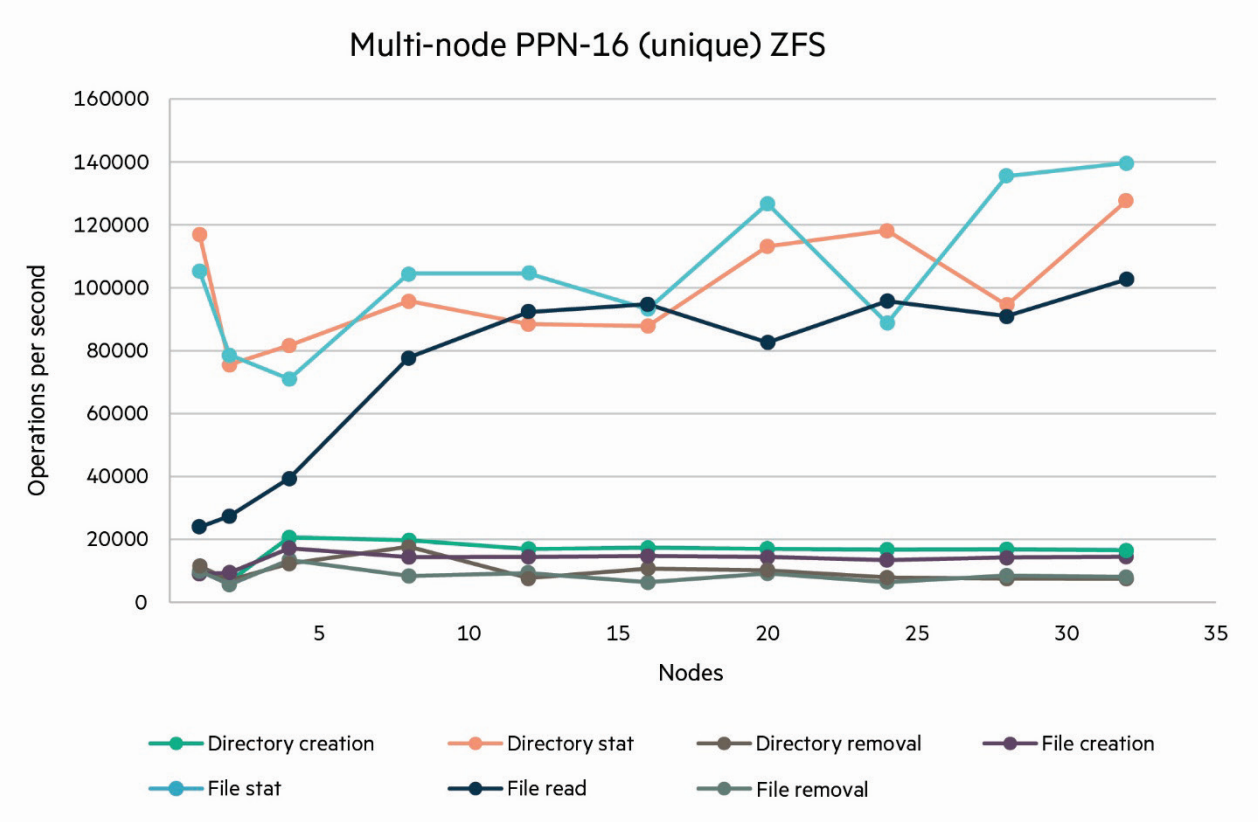

Figure 16. Metadata performance using HPE D3700 Disk Enclosure and ZFS MDT

Your application requirements will determine which metadata target option is right for you.

Bill of materials

This BOM represents the hardware components of the reference architecture described by this document. See the "Licensing and support" section of this document for information regarding software and support licenses required to deploy this solution.

HPE Apollo 4520 System

Quantity	Part number	Description
1	799581-B22	HPE Apollo 4520 Gen9 CTO Chassis
2	855128-B21	HPE Apollo 4500 x16 FIO I/O Module
2	786594-B22	HPE XL450 Gen9 2x Node Svr
2	842993-L21	HPE XL450 Gen9 Intel Xeon E5-2630Lv4 (1.8GHz/10-core/25 MB/55W) FIO Processor Kit
2	842993-B21	HPE XL450 Gen9 Intel Xeon E5-2630Lv4 (1.8GHz/10-core/25 MB/55W) Processor Kit
16	805351-B21	HPE 32GB 2Rx4 PC4-2400T-R Kit
2	761878-B21	HP H244br 12Gb 2-ports Int FIO Smart Host Bus Adapter
4	655710-B21	HP 1TB 6G SATA 7.2K rpm SFF (2.5-inch) SC Midline 1yr Warranty Hard Drive
2	726907-B21	HP H240 12G 2-ports Int Smart Host Bus Adapter
2	808963-B21	HPE Apollo 4510/4520 H240 Cable Kit
46	834031-B21	HPE 8 TB 12G SAS 7.2K LFF 512e LP MDL Hard Drive
2	825110-B21	HPE IB EDR/EN 100Gb 1P 840QSFP28 Adapter
2	845406-B21	HPE 100Gb QSFP28 to QSFP28 3m DAC Cable
2	720620-B21	HP 1400W Flex Slot Platinum Plus Hot Plug Power Supply Kit
1	681254-B21	HP 4.3U Server Rail Kit
4	726903-B21	HP Smart Array P841/4GB FBWC 12Gb 4-ports Ext SAS Controller
2	P9L6xAAE	Intel EE Lustre* SW License (see options in Licensing and support section of this document)

HPE D6020 Disk Enclosure and cabling (Note: the below numbers reflect quantities for one enclosure. Two enclosures are described in this reference architecture.)

Quantity	Part number	Description
1	K2Q28A	HPE D6020 Enclosure with Dual I/O Modules
1	K2Q23A	HPE D6020 Dual I/O Module Kit
70	858384-B21	HPE 8TB 12G SAS 7.2K rpm LFF 512e MDL Hard Drive
16	716197-B21	HP External 2.0m (6ft) Mini-SAS HD 4x to Mini-SAS HD 4x cable

HPE ProLiant DL360 Gen9 Server used as MDS and MGS (Two were used in this reference architecture)

Quantity	Part number	Description
1	755259-B21	HP ProLiant DL360 Gen9 4LFF CTO Server
1	818194-L21	HP DL360 Gen9 E5-2643v4 (3.4GHz/6-core/20MB/135W) FIO Kit
1	818194-B21	HP DL360 Gen9 E5-2643v4 (3.4GHz/6-core/20MB/135W) Processor Kit
8	805347-B21	HP 8GB (1x8GB) SR x8 DDR4-2400 CAS 17-17-17 Registered Memory Kit
1	764285-B21	HP IB FDR/EN 40Gb 2P 544+FLR-QSFP Adapter
1	670759-B25	HP 3M IB FDR QSFP DAC Cable
1	726911-B21	HP H241 12Gb 2-ports Ext Smart Host Bus Adapter

Bill of materials (continued)

1	749976-B21	HP H240ar 12Gb 2-ports Int FIO Smart Host Bus Adapter
1	766211-B21	HP DL360 Gen9 LFF Smart Array P440ar/H240ar SAS Cable
2	657750-B21	HP 1TB 6G SATA 7.2K rpm LFF (3.5 inch) SC Midline 1yr Hard Drive
2	720478-B21	HP 500W Flex Slot Platinum Hot Plug Power Supply Kit
1	789388-B21	HP 1U LFF Gen9 Easy Install Rail Kit
1	P9L6xAAE	Intel EE Lustre* SW License (see options in Licensing and Support section of this document)

HPE ProLiant DL360 Gen9 Server used as IML

Quantity	Part number	Description
1	755259-B21	HP ProLiant DL360 Gen9 4LFF CTO Server
1	818168-L21	HP DL360 Gen9 Intel Xeon E5-2603v4 FIO Kit
2	805347-B21	HP 8GB (1x8GB) Single Rank x8 DDR4-2400 CAS-17-17-17 Registered Memory Kit
1	749976-B21	HP H240ar 12Gb 2-ports Int FIO Smart Host Bus Adapter
1	766211-B21	HP DL360 Gen9 LFF Smart Array P440ar/H240ar SAS Cable
2	657750-B21	HP 1TB 6G SATA 7.2K rpm LFF (3.5 inch) SC Midline 1yr Hard Drive
2	720478-B21	HP 500W Flex Slot Platinum Hot Plug Power Supply Kit
1	789388-B21	HP 1U LFF Gen9 Easy Install Rail Kit

HPE MSA 2040 Storage and cabling

Quantity	Part number	Description
1	K2R84A	HP MSA 2040 Energy Star SAS Dual Controller SFF Storage
24	J9F40A	HP MSA 300GB 12G SAS 15K SFF (2.5-inch) Enterprise 3yr Warranty Hard Drive
4	716197-B21	HP External 2.0m (6ft) Mini-SAS HD 4x to Mini-SAS HD 4x cable

HPE D3700 Storage and Cabling

Quantity	Part number	Description
1	K2Q10A	HP D3700 w/25 300GB 12G SAS 15K SFF (2.5-inch) ENT SC HDD 7.5TB Bundle
4	716197-B21	HP External 2.0m (6ft) Mini SAS HD 4x to Mini SAS HD 4x cable

EDR InfiniBand TOR switch options

Quantity	Part number	Description
1	834976-B21	Mellanox IB EDR 36P unmanaged switch
1	834977-B21	Mellanox IB EDR 36P unmanaged RAF Switch
1	834978-B21	Mellanox IB EDR 36P Managed Switch
1	834979-B21	Mellanox IB EDR 36P RAF Managed Switch

Summary

The HPE Lustre* Scalable Storage solution featuring the HPE Apollo 4520 System offers excellent parallel file system performance at an attractive price. The configuration of the servers and storage in this solution are not fixed into an appliance model, providing you freedom to select processor, memory, and options to suit your application requirements. Installation Services from HPE ensures that your cluster will be configured and tuned for maximum availability and performance.

Appendix A: Software versions

The following section describes the firmware, driver, and software versions used for the reference platform described in this document.

- OS version: Red Hat Enterprise Linux 7.2
- H240/H241 firmware version: 3.56
- HPSA driver version: 3.4.10-0-RH1
- FDR IB adapter firmware: 10.12.0780
- EDR IB adapter firmware: 12.14.2036
- IB driver version: 3.1-1.0.3
- Mellanox OFED version: 3.1.1
- Intel EE for Lustre* 3.0.0.0

Resources

HPE Apollo Systems

HPE ProLiant DL360 Gen9 Server

Intel Enterprise Edition for Lustre*

HPE InfiniBand offerings

HPE MSA 2040 Storage

HPE Disk Enclosures

Learn more at

hpe.com/servers/proliant

4AA6-7430ENW, September 2016