White Paper

Architecting Storage for Semiconductor Design: Manufacturing Preparation

March 2012 | WP-7157

EXECUTIVE SUMMARY

The manufacturing preparation phase of semiconductor design—especially mask data preparation—is extremely I/O intensive. With file sizes measured in terabytes and bandwidth requirements approaching or exceeding 40GB/sec for large, complex designs, architecting a storage system capable of addressing today's requirements and scaling to address future needs has become a challenge. The NetApp® High-Performance Computing Solution for Lustre is the answer. This solution uses modular, flexible NetApp E-Series storage systems as the building block for Lustre deployments that provide the performance, scalability, and reliability to meet your current and future bandwidth and capacity needs while keeping overall storage costs to a minimum.

TABLE OF CONTENTS

1 INTRODUCTION 3

MEETING MANUFACTURING PREPARATION NEEDS WITH LUSTRE 3

2 UNDERSTANDING THE LUSTRE FILE SYSTEM 4

THE LUSTRE ARCHITECTURE 4

BANDWIDTH AND CAPACITY REQUIREMENTS FOR MANUFACTURING PREPARATION 6

LUSTRE STORAGE CHECKLIST 6

3 CREATING A LUSTRE SOLUTION FOR MANUFACTURING PREPARATION 6

THE NETAPP E-SERIES 7

HIGH-PERFORMANCE COMPUTING SOLUTION FOR LUSTRE ARCHITECTURE 8

4 MEASURED PERFORMANCE 9

5 APPLICATION SOFTWARE 10

6 ADVANCED SERVICE AND SUPPORT FOR LUSTRE 10

7 CONCLUSION 10

LIST OF FIGURES

Figure 1) Semiconductor design workflow 3

Figure 2) Lustre File System components 5

Figure 3) I/O patterns to MDS and OSS are different 5

Figure 4) E-Series 60 drive disk shelf 7

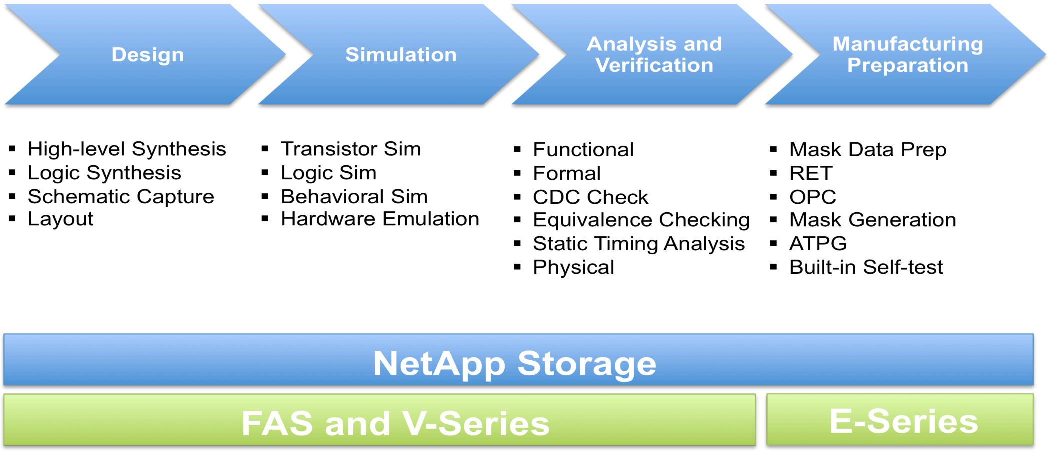
Figure 5) Components of the NetApp High Performance Computing Solution for Lustre 8

Figure 6) Demonstrated scalability of the High Performance Computing Solution for Lustre. 9

1 INTRODUCTION

The complexity and density of advanced semiconductor designs continue to increase with no end in sight. Although all phases of the design process are affected by tremendous data growth, the manufacturing preparation phase has been particularly affected. You need an effective strategy to provide the necessary bandwidth and capacity if you're going to continue to drive innovation, reduce cycle time, and remain competitive.

Figure 1) Semiconductor design workflow.



The mask data preparation applications used during the manufacturing preparation phase frequently run on high-performance computing (HPC) clusters and are extremely compute and I/O intensive. This creates significant IT challenges:

- **Huge data and bandwidth requirements.** The size of the individual data files used in mask data preparation is often measured in terabytes and will continue to increase as feature size shrinks to 22nm and beyond. With continued increases in storage capacity, storing files of this size might not seem like a problem, but making sure that you can deliver the necessary bandwidth to a compute cluster with thousands of cores likely will be. Bandwidth requirements today range from 1GB/sec to 40GB/sec depending on the size and complexity of the project. Although a single storage system might be able to meet your bandwidth and capacity needs today, there's no guarantee that it will continue to scale, and cost becomes a factor.
- **Shared access.** All nodes in a cluster need shared access to the same file systems and files.
- **Simultaneous reads.** All nodes in the compute cluster read simultaneously from the same input file during mask data preparation, resulting in a requirement for huge bandwidth to this single file.
- **Interconnect performance.** The interconnect fabric that connects compute nodes and storage must be able to satisfy these bandwidth requirements without bottlenecks or unacceptable latency.
- **Availability.** Given the time and resources that go into manufacturing preparation, data availability is critical. Any disruptions can affect schedules and affect time to market.

MEETING MANUFACTURING PREPARATION NEEDS WITH LUSTRE

Traditional, shared network file systems such as NFS are often unable to meet today's requirements. As a result, many in the EDA/semiconductor industry rely on parallel file systems, such as open source

Lustre or Quantum StorNext to address bandwidth and shared access needs. This paper addresses the use of Lustre for semiconductor manufacturing preparation.

Lustre is widely used among the world's top 500 supercomputing sites[1]. It is popular for its ability to support a wide range of clients, storage capacity, and bandwidth requirements. Lustre enables I/O performance and scaling beyond the limits of traditional storage and is becoming mainstream in storage environments that require very high bandwidth.

Parallel file systems such as Lustre provide shared access for multiple clients, but they allow a single file system to spread files across many servers and underlying storage systems such that aggregate data throughput for both reads and writes is significantly higher. Because Lustre is open source, it can be modified or extended to address site-specific needs.

Lustre puts unique requirements on underlying storage systems in terms of both bandwidth and capacity. These requirements are not always easily addressed by available storage solutions, and failure to achieve proper storage balance can result in a file system that chronically underperforms expectations.

Storage for Lustre must be tailored to address your requirements while at the same time delivering a high level of reliability and minimizing capital and operating costs. Optimizing Lustre to deliver the performance and availability needed for manufacturing preparation requires a storage partner that fully understands your environment and the unique requirements of mask data preparation and related applications.

NetApp offers a range of solutions capable of meeting your storage needs throughout the semiconductor development lifecycle (Figure 1). With the NetApp High-Performance Computing Solution for Lustre based on NetApp E-Series storage, NetApp provides an innovative storage approach that meets the unique needs of manufacturing preparation and that can be tailored to address your bandwidth requirements today and in the future while providing the balance necessary for efficient and cost-effective operation.

2 UNDERSTANDING THE LUSTRE FILE SYSTEM

Lustre enables I/O performance and scaling well beyond the limits of traditional storage technology and is becoming mainstream in storage environments that require very high bandwidth. Lustre provides a single file system namespace that is scalable to extremely high capacities and performance. Production file systems have scaled to 10PB with bandwidth of 300GB/sec. At least one system being actively deployed will scale to 55PB and up to 1,000GB/sec (1TB/sec). Lustre also supports large numbers of clients (tens of thousands) with concurrent read/write access. A distributed lock manager provides file coherency, and Lustre storage can be expanded on the fly.

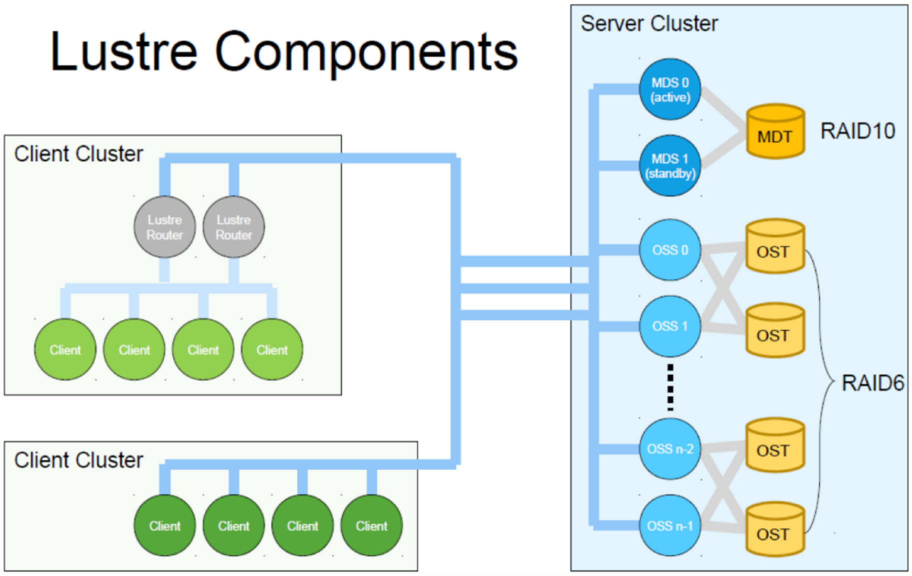
THE LUSTRE ARCHITECTURE

A key design feature of the Lustre file system is that it separates file system metadata—the bookkeeping information needed to keep track of files and directories—from data. (See Figure 2.) Metadata is managed by one or more separate metadata servers (MDSs) and underlying metadata targets (MDTs, one per file system). A standby MDS protects availability.

Actual file data is spread across multiple object storage servers (OSSs) and underlying object storage targets (OSTs). Up to 450 object storage servers and 10,000 object storage targets have been deployed in production. Clients run Lustre client software and mount Lustre file systems for read/write access.

[1] http://bigdata-io.org/lustre-used-in-top-70-big-computer-systems.

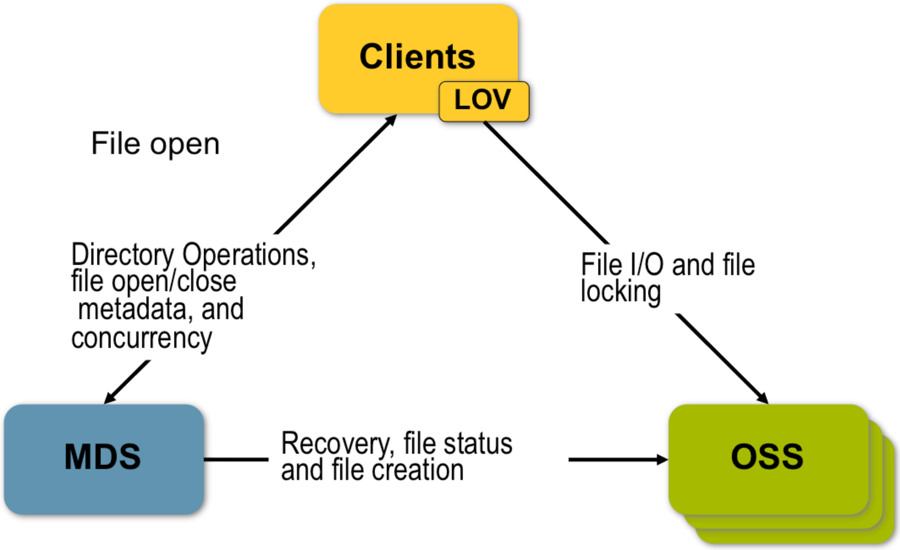
Figure 2) Lustre file system components.



In terms of the impact on storage, these components generate two significantly different patterns of I/O:

- All clients access file metadata (name, size, access times, data locations, and so on) stored by the MDS. Metadata transactions are small, highly random transactions, and since each Lustre file system typically relies on a primary MDS, response time for these I/O transactions is critical to overall Lustre performance.
- File data is striped across one or more object storage servers and underlying OSTs. Once a client gets the information it needs from the MDS, it negotiates locks with each object storage server it accesses, and then reads and writes data directly to each server in parallel. As a result, the I/O pattern to each OSS typically consists primarily of large, sequential reads and writes.

Figure 3) I/O patterns to MDS and OSS are different.


Clients
LOV
File open
Directory Operations, file open/close metadata, and concurrency
File I/O and file locking
MDS
Recovery, file status and file creation
OSS

These two distinct I/O patterns result in very different impacts on underlying storage. Traditional storage configurations might be good at one or the other, but probably won't provide great performance for both. Many Lustre installations run into problems when they try to use the same storage configuration to satisfy both requirements.

BANDWIDTH AND CAPACITY REQUIREMENTS FOR MANUFACTURING PREPARATION

Typical manufacturing preparation operations have bandwidth requirements that range from 1GB/sec to 40GB/sec and capacity requirements that range from 100TB to 5PB. Corner cases that require high bandwidth but relatively low capacity inevitably arise. These can be very difficult for underlying storage to address efficiently. For example, addressing a capacity requirement of a few petabytes using 3TB drives might simply not result in enough disk drive spindles to meet the bandwidth requirement.

LUSTRE STORAGE CHECKLIST

Storage for use with Lustre deployed in manufacturing preparation environments faces a unique set of requirements:

- Large, sequential reads and writes serviced by OSS and OST
- Small, random I/Os with high transaction rates serviced by MDS and MDT

In addition, there are a few additional considerations for Lustre storage in manufacturing preparation:

- **Reliability and availability.** The Lustre file system and the underlying storage must be as reliable as possible so that important jobs don't have to be restarted.
- **Cost.** The cost of storage is an important issue for most HPC environments, which would prefer to reserve budget dollars for computing resources. Lustre helps keep costs under control because it has no license fees, because it lets you to use any block storage device, thus eliminating vendor lock-in, and because Lustre's ease of scalability (performance and capacity) makes deploying a parallel file system more cost effective than scaling NAS implementations. Nevertheless, storage price per performance and price per capacity remain important considerations for a Lustre deployment.
- **Modularity.** Modular storage can make it much easier to achieve your Lustre performance and cost goals. Growing your storage environment (and it will grow) in small increments is much more cost effective than periodically adding large, monolithic storage systems. This modular approach to storage is very similar to the modular approach to growing HPC clusters and offers similar benefits in terms of achieving the balance necessary for the best storage performance.

3 CREATING A LUSTRE SOLUTION FOR MANUFACTURING PREPARATION

The NetApp High-Performance Computing Solution for Lustre uses NetApp E-Series platforms to create an optimized Lustre deployment for manufacturing preparation. This design uniquely satisfies the high transaction rates required by Lustre metadata servers and the sequential I/O needs of Lustre object storage servers.

The High-Performance Computing Solution for Lustre helps you:

- Achieve results faster by improving computational efficiency
- Start small and grow incrementally to better manage escalating capacity requirements
- Maximize productivity with a design for serviceability
- Decrease deployment and support costs with preconfigured and pretested solutions

THE NETAPP E-SERIES

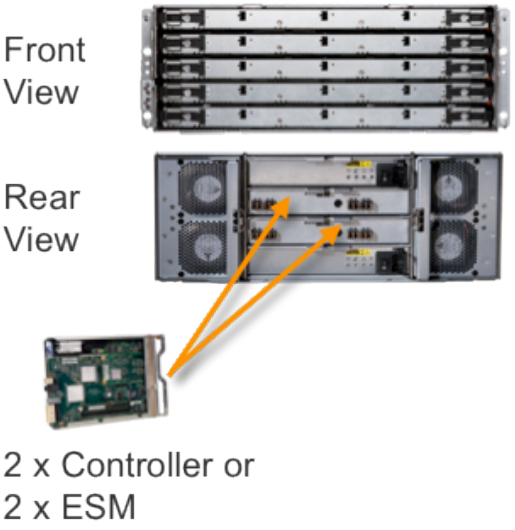
The modular, highly scalable NetApp E-Series provides the bandwidth, storage density, and reliability necessary for critical manufacturing preparation operations while minimizing both capital and operating costs. Advanced features reduce energy consumption and make your operations more efficient. The modular design of the E-Series simplifies scaling and increases flexibility. You simply buy the number of modules with the number and types of disks you need. Various types and numbers of host interconnects (SAS, FC, InfiniBand) give you the flexibility to build your Lustre storage environment in the way that makes the most sense for you.

CONFIGURATION OPTIONS

The NetApp High-Performance Computing Solution for Lustre includes two E-Series controller options. The NetApp E2600 storage controller is a good choice to meet the transactional performance needs of Lustre MDTs. It can also be deployed in the OST role in Lustre installations with more modest bandwidth requirements. The E5400 storage controller supports up to 360 disks per controller pair (using external disk shelves) and is designed to serve as the ideal OST in most Lustre deployments with optimal bandwidth for modular, scalable Lustre deployments and can also serve as an MST in very large or high-demand environments. Performance of the E5400 in the NetApp High-Performance Computing Solution is described in section 4.

E-Series controllers (in redundant pairs for reliability) combine with several disk shelf options. For maximum storage density, an innovative 4U disk shelf holds up to 60 disk drives in five drawers, providing up to 180TB per enclosure or 1.8PB per rack. Controlled drawer movement allows each 12-drive drawer to be extended while its drives remain active, allowing for individual drives to be replaced without affecting the operation of other drives (in either the drawer or the enclosure) for superior availability and serviceability. A 2U disk shelf supports 24 2.5” SAS drives. Configuring 20 of these shelves in a 40U rack provides write performance of more than 30GB/sec. (See section 4 for more information on performance.)

Figure 4) E-Series 60-drive disk shelf. The chassis can be configured with redundant 5400 controllers or redundant shelf electronics (ESM), allowing it to serve as either a standalone storage system or an expansion shelf.



SAS-ENABLED

The E-Series uses the latest 6Gb/sec SAS technology. To connect to disk, SAS uses a 4-lane, wide port delivering up to 24Gb/sec full duplex. A point-to-point topology offers better isolation for each device, and improvements in redundant drive paths further protect against failures that could take data offline.

ALWAYS ON

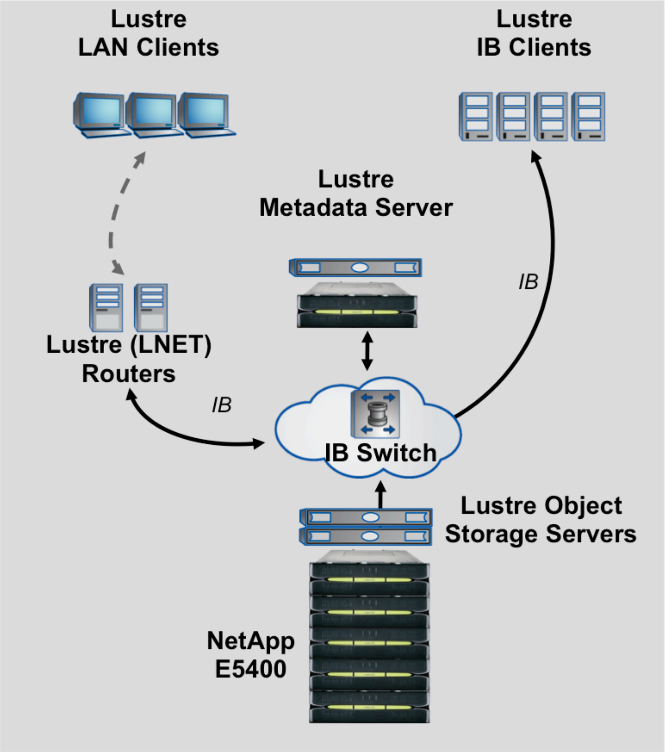
Fully redundant components and I/O path failover allow the E-Series to deliver “always-on” availability. Additional reliability and availability features include:

- Media scan with automatic parity check and optional correction
- Extensive diagnostic data capture and statistics collection
- Proactive drive and I/O monitoring and automated repair
- Optional RAID parity verification
- Embedded system health check
- Enhanced drive recovery
- Redundancy checks during RAID group configuration

HIGH-PERFORMANCE COMPUTING SOLUTION FOR LUSTRE ARCHITECTURE

Typical components of the High-Performance Computing Solution for Lustre are illustrated in Figure 5. The diagram shows InfiniBand as the client interconnect, but 10 Gigabit Ethernet can also be used. The storage fabric can be InfiniBand, Fibre Channel, or SAS.

Figure 5) Components of the NetApp High-Performance Computing Solution for Lustre.



Because of the modularity and flexibility of the E-Series, it serves the needs of both MDT and OST components. To meet MDT storage requirements, a single E-Series (often an E2624) is configured with either SAS or solid-state drives (SSDs) in a RAID 10 configuration, while for OST needs, multiple E5460 systems are typically configured with 3TB or 2TB SAS disks. A key element of the design is the ability to scale capacity and performance incrementally for either metadata or data by adding additional E-Series systems or additional drive enclosures.

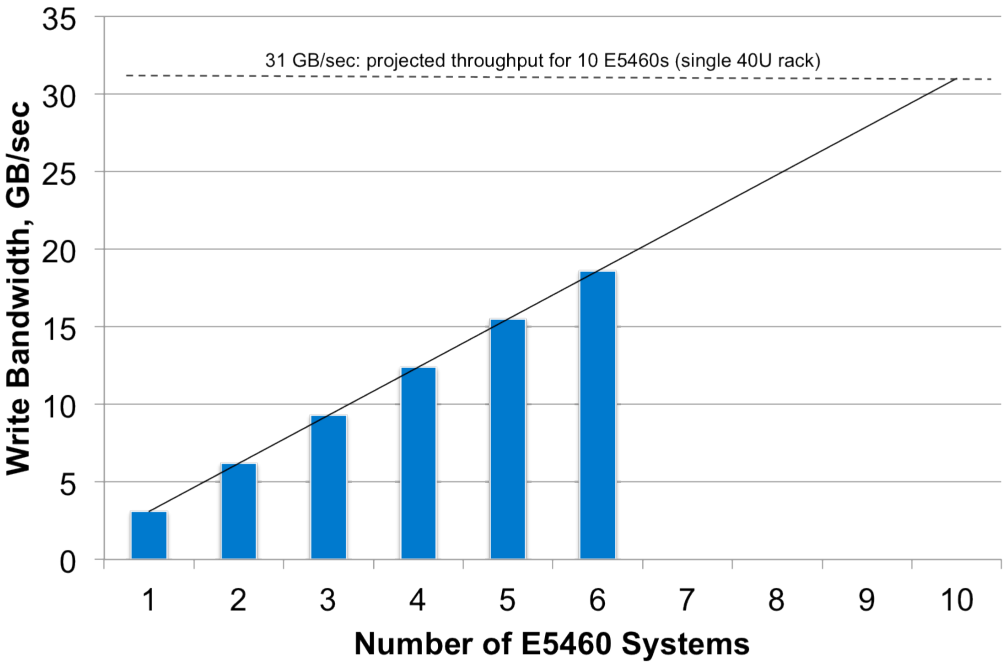
As illustrated, LAN clients can be supported using one or more Lustre distributed LAN servers, which act as a gateway to allow clients that aren't directly connected to InfiniBand to read and write data stored in a Lustre file system.

4 MEASURED PERFORMANCE

To assess the performance of the High-Performance Computing Solution for Lustre, we utilized the Lustre I/O kit, a collection of benchmark tools for assessing the performance of a Lustre deployment. We wanted to demonstrate not only the performance of a single E5460 storage system deployed as an OST, but also the scalability as additional E5460 systems are added. For all tests, two object storage servers were connected to one or more E5460 storage systems. Each E5460 was configured with redundant 5400 controllers in a 4U, 60-disk-drive chassis in a dual-parity RAID 6 configuration. Cache mirroring was enabled between controllers, and 7200 RPM, high-capacity near-line SAS drives were used in each storage system. Up to 10 E5460 storage systems fit in a single 40U rack.

The obdfilter_survey script, provided as part of the Lustre I/O kit, is designed to assess and size OST throughput performance over the network. It was employed to characterize the performance of the solution starting with a single OST and then incrementing up to a total of 6 OSTs. Results are shown in Figure 6.

Figure 6) Demonstrated scalability of the High-Performance Computing Solution for Lustre.



Each E5460 delivers 3.1GB/sec write bandwidth, and the scaling is extremely linear as you add E5460s. What this means in practice is that you can use the E5460 to create a Lustre configuration that closely matches your peak bandwidth requirements and scale performance in increments of 3.1GB/sec as your requirements grow. If you need more capacity than bandwidth, you can add capacity to each E-Series system by attaching additional disk shelves.

5 APPLICATION SOFTWARE

The Lustre file system and the NetApp High-Performance Computing Solution for Lustre are independent of the application software that utilizes them. Popular manufacturing preparation applications including:

- Mentor Graphics Calibre product line
- Synopsys Proteus and CATS
- Cadence Mask Compose Reticle and Wafer Synthesis Suite

and other applications that run on Linux™ should all work without changes. Simply installing Lustre client software on underlying compute servers will provide applications with transparent high-speed, block-level access to Lustre file systems.

6 ADVANCED SERVICE AND SUPPORT FOR LUSTRE

The right technology is important, but having the right partners to stand behind that technology can also be critical to success. NetApp Global Services provides 24x7x365 support and parts dispatch. To enhance your success with NetApp storage and the Lustre file system, NetApp has partnered with Whamcloud, Inc. to provide Lustre support services for NetApp storage customers. Together, the companies provide Level 1–3 Lustre support from a single source. Support services will be available 24x7x365 and are designed to enable you to get your Lustre installations back to full operation more quickly than with other alternatives.

The collaboration between NetApp and Whamcloud brings together a storage industry leader in NetApp, with hundreds of petabytes of storage deployed in HPC environments, with Whamcloud's HPC and Lustre industry veterans, who are focused on evolving Lustre and advancing HPC storage. If you have deployed or are considering deploying Lustre, you can now get a combination of product and support services from a single source that fully understands the interaction between storage and Lustre.

Additionally, NetApp Professional Services can help you with storage solution design, performance optimization, data migration, and strategic consulting.

7 CONCLUSION

The NetApp High-Performance Computing Solution for Lustre has been designed to deliver the I/O performance required for manufacturing preparation using a modular, building block approach that simplifies Lustre deployment while delivering the reliability and scalability you need at an extremely competitive price. Based on NetApp E-Series storage, the High-Performance Computing Solution for Lustre offers great performance density and maximum storage density—up to 1.8PB in a single 40U rack—giving you the flexibility to create a Lustre deployment tailored to your needs.

It's clear that the I/O and capacity requirements for manufacturing preparation will continue to increase as feature size shrinks to 20nm and beyond. The NetApp High-Performance Computing Solution for Lustre is designed to provide exceptional investment protection and grow incrementally with your needs without breaking your budget. For more information on the NetApp High-Performance Computing Solution for Lustre, visit [www.netapp.com/lustre.](http://www.netapp.com/lustre)

NetApp provides no representations or warranties regarding the accuracy, reliability, or serviceability of any information or recommendations provided in this publication, or with respect to any results that may be obtained by the use of the information or observance of any recommendations provided herein. The information in this document is distributed AS IS, and the use of this information or the implementation of any recommendations or techniques herein is a customer's responsibility and depends on the customer's ability to evaluate and integrate them into the customer's operational environment. This document and the information contained herein may be used solely in connection with the NetApp products discussed in this document.

Go further, faster®

www.netapp.com