WHITE PAPER
Intel® Enterprise Edition for Lustre* Software
High Performance Data Division

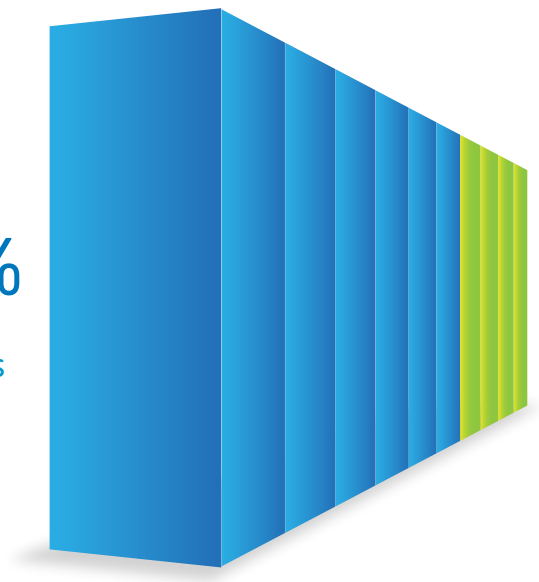
A New Generation of Lustre* Software Expands HPC Into the Commercial Enterprise

Not so long ago, storage for high performance computing (HPC) meant complexity and massive data sets, and was the concern of only a small group of computer users. Super-scale computing was the province of government-sponsored research in national labs, data-intensive simulations for weather forecasting and climate modeling, or certain information-intense industries, such as defense, aeronautics, and oil and gas.

Today, HPC is undergoing democratization: Extracting knowledge and information from ever-expanding flows of data is now seen to be a key source of competitive advantage for modern businesses of any size.

Enterprises of all kinds now generate huge volumes of data. They rely on high-performance data processing applications to analyze and derive value from their data flows. They require a storage infrastructure that can scale endlessly and deliver large volume I/O for high-throughput data processing.

Lustre powers over **60%** of the top 100 supercomputers worldwide[1]



However, roughly half of enterprise storage systems today are based on the Network File System (NFS), a type of distributed file system that consolidates data resources onto centralized networked servers, and allows remote clients to mount data files over the network and interact with these files as though they are available locally. While effective in smaller storage environments, NFS can quickly become a major bottleneck in high volume systems because it does not scale well, and requires increasingly costly management overhead, even as its performance diminishes.

Designed specifically for high performance computing, the open source Lustre* parallel file system is one of the most popular, powerful and scalable data storage system currently available, and is in widespread use today in super-computing scenarios where high performance and enormous storage capacity is required. However, from its early days of development in research labs and academia, Lustre has also

INTEL & OPEN SOURCE PARALLEL STORAGE

With Intel® EE for Lustre* software, Intel builds on its leadership in high performance server platforms, software tools for creating parallel applications, and datacenter optimization to extend its innovation to parallel storage. Intel, with the backing and support of OpenSFS and EOFS, leads the open, collaborative development of the Lustre file system; in this role Intel manages the source code repositories on behalf of the community, ensuring coordinated feature development, rigorous testing and predictable releases. Intel EE for Lustre software plays an essential role in the growing, Intel-supported open-source ecosystem that presents a multi-vendor approach to such leading-edge technical challenges as HPC, big data, and the cloud.

gained the reputation of being difficult to manage. Though high performance, scalable storage could be useful for many business applications, these early versions of Lustre were considered too difficult to manage for enterprise use.

With the release of **Intel® Enterprise Edition for Lustre* software (Intel® EE for Lustre* software)**, Lustre has moved beyond the lab and into the enterprise. Intel provides business customers with a commercial-grade version of Lustre optimized to address key storage and data throughput challenges of HPC-class computing in business. Intel EE for Lustre contains an open source distribution of Lustre, tested and validated by Intel, with the latest features and hardened for deployment in production. Intel® Manager for Lustre* software, a set of integrated management tools, simplifies installation, deployment, monitoring and management of Lustre. Intel EE for Lustre software delivers the Lustre parallel storage file system, backed with the full support and expertise of Intel, with the stability, efficiency, and reliability required by today's enterprises.

The limitations of NFS-based storage and I/O

HPC clusters are designed to provide extreme computational power to large-scale applications. This computational power often results in the creation of huge numbers of files and/or extremely large individual files that stress the capabilities of conventional file systems beyond their limits.

But while the speed of processors and memory has risen sharply in recent years, the performance of I/O systems has lagged behind. Even HPC infrastructures can operate only as fast as its slowest component, usually its storage file system. Disk drives still essentially spin at the same speeds as 20 years ago, and poor I/O performance can severely degrade the overall throughput of even the fastest clusters.

The Network File System, developed by Sun Microsystems in the 1980s, is the de facto distributed file system for Linux* based computing, and Linux is the predominant operating system used in HPC and large enterprise computing environments.

NFS is very stable, and for general-purpose enterprise and business computing, it is adequate for a wide variety of tasks. However, as commercial technical computing workloads scale upward into HPC realm and storage system capacity tops 100 terabytes, many users discover that traditional data management solutions based on distributed file systems like NFS come up short—due to their inherent scalability constraints.

NFS and other distributed file systems work by designating a single node to function as the I/O server for the storage cluster. All I/O—reading data and writing data—go through that single node. While this system is relatively simple to manage in a single cluster deployment, pushing all of an enterprise's I/O through one server

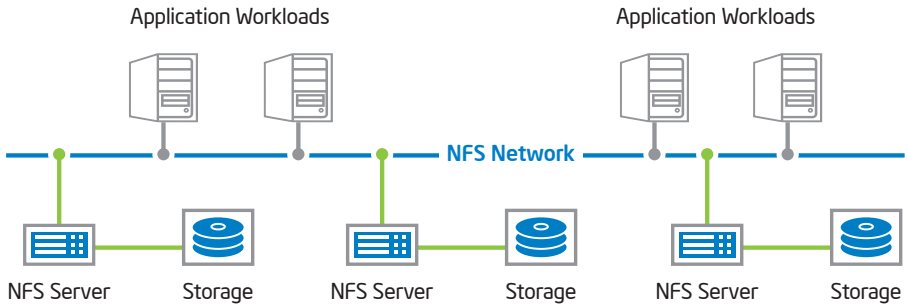


Figure 1. When scaling up an NFS-based environment, each of the disparate NFS server clusters must be managed individually, adding to data bottlenecks as well as management overhead and costs.

1,000,
000,
000,
000 Bytes per second

Leading-edge Lustre configurations can deliver data throughput in excess of 1 terabyte per second

node quickly presents a bottleneck for data-intensive workloads. With a single gateway between an application and its target storage system, storage performance remains restricted by the processing capacity of the node.

This single server approach can also mean a single point of failure, as a failed I/O results in data that's no longer available for applications.

Scaling up performance and capacity in NFS environments begins as an additive process—just add more server clusters to the network, each with its own I/O server node. This linear expansion of clusters can work well enough until management overhead and costs become onerous (each of the disparate NFS server clusters must be managed individually) and data bottlenecking compounds across the network.

Introduction to Lustre

The Lustre global parallel file system was first conceived in 1999 at Carnegie Melon University in response to growing awareness of the limits of NFS, and has become the predominant file system in HPC environments, or any supercomputing environment where high I/O bandwidth and scale are required. The Lustre file system is unmatched in speed, scalability, and availability and can support tens of thousands of client systems, tens of petabytes of storage, and more than a terabyte per second of aggregate I/O throughput. Lustre sotware powers over 60 percent of the top 100 supercomputers and is the most widely used file system for TOP500 supercomputing.

Lustre is an object-based file system that splits file metadata (such as the file system namespace, file ownership and access permission) from the actual file data and stores them on different servers. File metadata is stored on a Metadata Server (MDS), and the file data is split into multiple objects and stored on Object Storage Targets (OST).

When a client opens a file, Lustre contacts the MDS to look up the file in the filesystem namespace, verify access permissions and return the file layout. This layout tells the client how the file's data is distributed over OSTs. From then on, all I/O is transacted directly between the client and the OSTs without having to interact with the MDS again. Because metadata and object data are stored on separate servers, each system can be optimized for the different workloads they present.

This is the primary advantage that Lustre has over a file system such as NFS, where all I/O has to go through a single NFS server or head node. Lustre allows multiple clients to access multiple OSS nodes at the same time independent of one another, thereby allowing the aggregate throughput of the file system to scale with the simple addition of more hardware. Performance is essentially limited only by the amount and characteristics of the storage hardware available. Lustre does not slow down despite data growth and grows in storage capacity as computing needs grow. Redundant servers and failover-enabled storage results in a file system that is highly available, with no single point of failure.

The Lustre Network (LNet) is a set of protocols and APIs that provide a network for connecting clustered servers and clients to Lustre file systems. The LNet supplies pluggable drivers to support ultra low latency network infrastructures such as Infiniband (verbs) or Ethernet. The drivers are loaded into the driver stack for each network type that is in use. Key features of LNet include support for Remote Direct Memory Access (RDMA) to provide fast memory-to-memory interconnects over InfiniBand or Ethernet networks for maximum performance and reduced CPU overhead. LNet also delivers high availability and transparent recovery in coordination with failover storage servers, along with the ability to route across multiple network types.

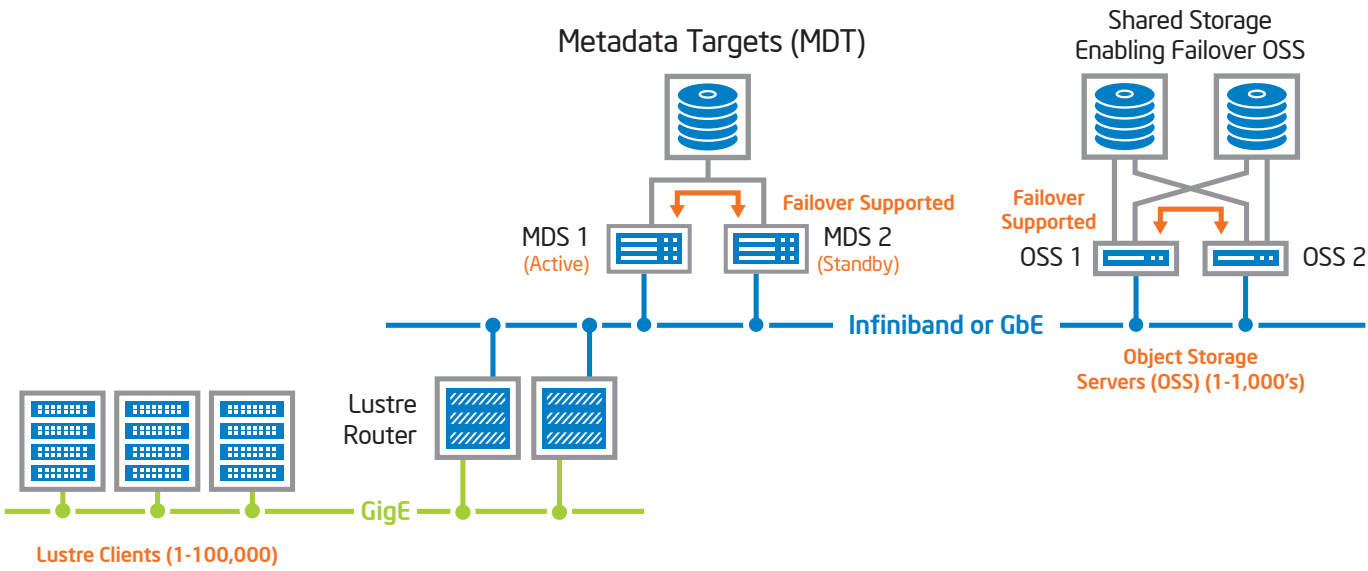


Figure 2. Typical Lustre configuration

HPC AND THE MISSING MIDDLE

In 2008, the Council on Competitiveness released a series of studies on U.S. manufacturing that found the vast majority of small to medium companies in the U.S. were missing out on high performance computing's potential to drive innovation and make American companies more competitive.[2] Of some 300,000 manufacturing companies in the nation, five percent are global industrial giants that have been using HPC for years for advanced modeling, simulation and analysis. The study found, however, that most small and medium companies in the U.S. not only lack sophisticated HPC capabilities, but some 65 percent didn't even use rudimentary desktop-based modeling in their manufacturing research and development, instead relying on physical prototyping.

The Council on Competitiveness coined the term "the missing middle" to identify this large group of manufacturers who lack HPC capabilities and, increasingly, global competitiveness. According to the council, "HPC represents a crucial edge that can build and sustain competitive advantage through innovative product design, production techniques, cost savings, improved time-to-market cycles, and overall quality."[3] Making HPC R&D more widely available to the missing middle has the potential to transform U.S. manufacturing while creating jobs and allowing a wider range of companies to compete globally.

Intel Enterprise Edition for Lustre Software

Intel EE for Lustre software optimizes the Lustre parallel file system as an enterprise platform for a broad spectrum of commercial organizations. It allows enterprises with large scale, high-bandwidth storage to tap into the power and scalability of Lustre, but with the simplified installation, configuration and monitoring features of Intel Manager for Lustre, a management solution purpose-built for the Lustre file system. Intel Manager for Lustre helps bring the performance benefits of Lustre to data-intensive businesses and organizations without the need for highly specialized technical administrators.

A stable, commercial-ready version of Lustre, bundled with smart management tools, helps bring the benefits of high performance computing to a broad range of businesses, including smaller and mid-sized companies with more limited technology resources. HPC capacity can deliver a higher level of computing power and throughput, and make available information and insights derived from big data and compute-intensive applications—such as advanced modeling, simulation, and data analysis—to a new tier of enterprise users. Intel EE for Lustre software, in combination with other super-computing technologies, offers the potential to drive innovation, deliver higher quality products and designs, and sustain competitive advantage for a wide array of businesses.

PUSHING STORAGE BOUNDARIES: THE FASTFORWARD PROGRAM

Administered by the United States Department of Energy, the FastForward program is designed to accelerate the research and development of critical technologies to enable extreme scale, or exascale, computing. The program seeks to address the nation's most pressing scientific challenges by advancing simulation-based scientific discovery made possible by exascale supercomputers.

FastForward is contracted through a consortium of seven national laboratories, with subcontracts awarded to a network of private companies with expertise in high performance computing (HPC). In 2012, Whamcloud was awarded the Storage and I/O Research & Development subcontract for the FastForward program. Whamcloud, a global leader in parallel storage for HPC and Lustre research and support, was shortly afterward acquired by Intel. The Intel® High Performance Data Division (Intel® HPDD), the new home of Whamcloud at Intel, has continued the storage and I/O research for the FastForward subcontract. The two-year project includes key R&D necessary for a new object storage paradigm for HPC exascale computing; the new storage technology will also address next-generation storage mechanisms required by the Big Data market. All components developed in the project will be open sourced and benefit the entire Lustre community.

Benefits of Intel Enterprise Edition for Lustre Software for Commercial Organizations

Performance

Intel EE for Lustre software is designed to enable fully parallel I/O throughput across many clients, servers and storage devices. Many Lustre configurations are running in production at 500 to 750 gigabytes per second, with leading edge installations achieving throughput in excess of 1 terabyte per second.

This means extremely high volumes of data can be delivered to critical high performance applications, leading to improved decision-making based on near real-time analysis. High performance data flows allow an enterprise to run larger and more complex applications faster and easier, providing an innovative edge for businesses. Intel EE for Lustre can also scale down efficiently to provide fast parallel storage for smaller organizations.

High throughput data flows can help contribute to a higher return on investment (ROI) for HPC infrastructure. Massive data flows can utilize a high percentage of underlying storage and networking hardware performance, for low performance overhead.

Capacity

Lustre was developed to handle the demands of scientific data, and has been tested and trusted at extreme levels of throughput. The object-based storage architecture of Intel EE for Lustre software can scale to tens of thousands of clients; and at 512 petabytes of storage for the current version of Lustre, storage capacity is essentially unlimited.

Affordability

The Lustre distribution in Intel EE for Lustre software is open source, hardware neutral, and supports storage and servers from multiple vendors. There is no vendor lock-in, so administrators can customize the storage infrastructure to accommodate individual budgets.

Intel Manager for Lustre software provides browser-based tools for fast, efficient management, greatly reducing the once-imposing level of care and nurturing that Lustre deployments formerly demanded. Intel EE for Lustre software does not require specialized training or expertise to operate, and businesses can scale their Lustre storage deployments horizontally, yet continue to manage them with the same efficiency and precision as before.

Maturity

Intel EE for Lustre software delivers the most advanced Lustre features, rigorously tested and proven for diverse markets.

Intel EE for Lustre software brings together the best and brightest of Lustre expertise coupled with the resources, support and credibility of Intel. Intel EE for Lustre software has a clear product roadmap, with predictable releases. It includes best-in-class support from the Lustre experts at Intel, including worldwide 24X7 technical support services.

Intel Manager for Lustre Software

Intel Manager for Lustre software provides a unified, consistent view of Lustre storage systems and simplifies the installation, configuration, monitoring and overall management of Lustre. The manager consolidates all Lustre information in a central, browser-accessible location for ease of management and reduced com-

plexity. It helps lower the management costs of Lustre while accelerating the benefits of parallel storage software.

Key features of Intel Manager for Lustre software include:

- **Intuitive, GUI-based administration.** The easily navigated user interface centralizes definitions and management of common administrative tasks and provides insights across the entire storage system using simple but powerful graphical views and scriptable command line interfaces. Cluster configuration, provisioning, and management are performed with point-and-click simplicity.
- **Real-time system monitoring.** Keep abreast of end-to-end storage system health and access key performance indicators (KPIs) in real time. The manager interface allows insights into high-level system performance or in-depth focus into individual components. Robust reporting tools make it easy to generate historical and real-time charts and reports.
- **Advanced troubleshooting tools.** Proactively identify and correct faults before they become larger issues. Deal quickly with events and solve problems through a consolidated view of cluster-wide storage log files, with intelligent log-scanning capabilities for efficient problem isolation and analysis. Use repeatable self-test metrics to monitor incremental changes within the system. Configurable event notifications make it easy to track and schedule automated patches and security fixes.
- **Open, documented APIs.** Deploy Intel EE for Lustre software in existing networks using open and scriptable REST-compliant APIs for simplified integration with other storage systems and software management tools. Storage plug-in architecture provides easy extensibility.

USING LUSTRE IN CONJUNCTION WITH APACHE HADOOP*

From Wall Street to the Great Wall, enterprises and institutions of all sizes are faced with the benefits – and challenges – promised by 'Big Data'. But before users can take advantage of the near limitless potential locked within their data, they must have affordable, scalable and powerful software tools to manage the data.

High performance infrastructure workloads have expanded and are now key technologies used by today's forward-looking commercial computer users. Parallel storage solutions powered by Lustre storage software have found a new home in these data-intensive business operations, and Apache Hadoop* has become the framework of choice for big data analytics. Hadoop transforms enormous amounts of data into manageable distributed datasets that applications can more easily analyze.

When organizations operate both Lustre and Hadoop within a shared infrastructure, there is a strong case for using Lustre as the file system for Hadoop analytics as well as HPC storage. Hadoop users can access any Lustre files directly from Hadoop, without the need to copy them over to the Hadoop environment. Using Lustre in combination with Hadoop also makes storage management simpler—since the platform will be running a single Lustre file system instance rather than Hadoop instances for each cluster—and makes more productive use of storage assets.

Moreover, Hadoop's own file system, referred to as HDFS, is inconsistent with the HPC paradigm of decoupling computation from storage, as HDFS expects storage disks to be locally attached to individual compute nodes.

In addition, since HDFS is not POSIX-compliant—meaning that it does not conform to standards that maintain compatibility between operating systems—it entails the performance overhead of moving extremely large datasets in and out of Lustre for the purposes of staging I/O throughput. Fortunately, Hadoop uses a storage abstraction layer for accessing persistent data, thus allowing the potential for plugging in different types of file systems. Lustre can be made to comply with Hadoop's storage requirements by implementing its Java* file system API. Since Lustre is POSIX-compliant and can be mounted like an NFS, it is able to exploit Java's inherent support for native file systems.

The only additional step for mounting Lustre as the file system for Hadoop analytics is to convey to the Hadoop task scheduler that Lustre is indeed a distributed file system and the input data are accessible uniformly from all the compute nodes. This allows tasks to be scheduled on any node independent of data locality, so all Hadoop "compute" nodes can access any data, eliminating the need to move the data itself between nodes. Additional optimization is possible by allowing reducers to read intermediate map outputs directly from the shared file system and eliminating the overhead of streaming large files over HTTP.

Intel EE for Lustre software includes an adapter for Apache Hadoop* which allows users to run Map/Reduce* applications directly on Lustre. This optimizes the performance of Map/Reduce operations while delivering faster, more scalable and easier to manage storage.

Lustre is Open for Business

With the release of Intel Enterprise Edition for Lustre software, Lustre is in a stronger position than it's ever been, as high performance storage is set to become a vital competitive differentiator for a broad range of businesses and commercial organizations. With its stable performance, simplified management, and support from Intel, Intel EE for Lustre software is poised to help move HPC into new markets, helping companies to improve data analysis; speed design, production and decision-making; and spur competitive advantage. Intel EE for Lustre software also stands ready to extend high-volume data processing into technology areas such as big data, business intelligence, and private cloud computing, where high data throughput is required.

Intel EE for Lustre software is built to solve today's most demanding storage challenges and accelerate performance of critical applications and workflows to bring the benefits of HPC to a broader community of businesses. With veteran Lustre engineers and developers working at Intel and contributing to the Lustre code, Lustre will continue its growth in both HPC and commercial environments. It remains the best breakthrough technology for addressing the exascale and emerging big data challenges of tomorrow.

For more information on Intel® Enterprise Edition for Lustre software, please contact us at **hpdd-sales@intel.com**

[1] www.top500.org

[2] Council on Competitiveness and USC-ISI In-Depth Study of Technical Computing End Users and HPC

[3] High Performance Computing to Enable Next-Generation Manufacturing

Printed in USA 0613/CM/MB/PDF Please Recycle 329079-001US