

Lustre™ File System

High-performance storage architecture
and scalable cluster file system



lustre

Highlights

- **Unparalleled scalability:** Employs an object-based storage architecture that scales to tens of thousands of compute nodes and petabytes of data storage — a file system virtually without limits
- **Reliable:** Deployed in production on many large and small clusters, meeting the uptime requirements of business and national security applications
- **Proven performance:** Delivers dramatic increases in throughput and I/O by intelligent serialization and separation of metadata operations from data manipulation
- **Cost effective:** Significantly reduces deployment and support costs through support for industry-standard platforms and heterogeneous networking environments
- **Open source, open standards:** Developed and maintained as open source software with an open networking protocol and POSIX file system semantics, ensuring broad support for industry-standard platforms and heterogeneous networking environments



The Lustre™ file system redefines I/O performance and scalability standards for the world's largest and most complex computing environments. Ideally suited for data-intensive applications that require the highest possible I/O performance, Lustre is an object-based cluster file system that scales to tens of thousands of nodes and petabytes of storage with groundbreaking I/O and metadata throughput.

Unique storage architecture

The highly scalable and distributed Lustre file system combines open standards, the Linux operating system, an open networking API, and innovative protocols. Together, these elements create the world's largest “network-neutral” data storage and retrieval system.

Applying intelligence throughout its architecture, the Lustre file system turns commodity hardware into smart storage devices that manage data objects. The objects are dynamically distributed horizontally across the servers, shattering the performance limitations of traditional storage systems.

Building a Lustre cluster requires a Lustre MetaData Server and Lustre Object Storage Servers, each with disk storage. A pool of client systems can access these servers through one of many supported networks. Lustre I/O bypasses the MetaData Server completely and fully utilizes the parallel data paths to all Object Storage Servers in the cluster. This unique approach of separating metadata operations from data operations results in significantly enhanced performance.

Open source

Lustre technology has been developed and maintained as open source software under the GNU General Public License (GPL), enabling broad support for industry-standard platforms.

Heterogeneous networking

The Lustre network architecture provides flexible support for a wide variety of networks and high-performance features. The Lustre file system, based on open standards interoperates with networking stacks as well as vendor-supplied libraries through Lustre Network Drivers, utilizing advanced features such as RDMA and OS-bypass to maximize throughput. Lustre Network Drivers exist for many networks, including TCP/IP, many flavors of InfiniBand OFED, and Myrinet MX — each with performance exceptionally close to the raw device throughput.

Rugged high availability

The Lustre file system organizes all servers in active-active failover pairs. Together with protocol interoperability between versions, live cluster upgrades are now routine.

POSIX compliance

The Lustre file system provides a tested and fully compliant file system interface in accordance with the POSIX standard.

Simple configuration

With version 1.8, the Lustre file system revolutionizes configuration simplicity. Routine formatting and mounting of server devices aggregates them into a global high-availability cluster file system.

Extreme parallel computing

The Lustre lock manager automatically adapts its policies to minimize overhead for the current application. Files being used by a single node are covered by a single lock, eliminating additional lock overhead. Nodes sharing files get the largest possible locks, which still allows all nodes to write at full speed.

To dramatically reduce bottlenecks and to increase overall data throughput, the Lustre file system uses an intent-based locking mechanism, where file and directory lock requests also provide information about the reason the lock is being requested. For example, if a directory lock is being requested to create a new unique file, the Lustre file system handles this as a single request. In other file systems, this action requires multiple network requests for lookup, creation, opening, and locking. This ensures file system coherency without sacrificing overall performance.

Lustre File System Specifications

Operating systems

- Red Hat Enterprise Linux 5
 - SuSE Linux Enterprise Server 11, i686 and x86_64 only (Lustre 1.8.1)
 - Linux kernel 2.6.16 or greater
- NOTE: Lustre does not support security-enhanced (SE) Linux (including clients and servers).

Hardware platforms

- Intel Architecture 32-bit/64-bit
- x86, IA-64, x86-64 (EM64 and AMD64)
- PowerPC architectures (for clients only) and mixed-endian clusters

Networking

- TCP/IP
- InfiniBand OFED
- Myrinet MX
- Cray XT3/4/5

Proven performance

Lustre is best known for powering the largest HPC clusters in the world, with tens of thousands of clients per system, petabytes (PB) of storage, and hundreds of gigabytes per second (GB/sec) of I/O throughput. Many HPC sites use Lustre as a site-wide global file system, serving dozens of clusters on an unprecedented scale.

Recent results

- File I/O % of raw bandwidth: >90%
- Achieved single OSS I/O: >2.5 GB/s
- Achieved single client I/O: >2.0 GB/s
- Achieved aggregate IO = 240 GB/s
- Metadata transaction rate: 15,000 ops/s
- Maximum clients supported: 100,000
- Maximum file/file system size: 320 TB/>32 PB

Sun Support and Services

Sun Spectrum Support:

Sun provides comprehensive software support for your Lustre application with the SunSpectrum Software Service Plan at the Standard level. Having access to Sun's expert technicians, in-depth knowledge base, software updates and releases helps ensure the success of your deployment and the availability of your solution.

Services:

- Sun Lustre Implementation Services provide a combination of pre-defined Lustre configurations and a choice of three preselected service offerings.
- Sun HPC Services provides help to architect, implement, and manage your HPC infrastructure quickly and easily, with reduced cost and risk.
- Sun Datacenter Express Services for HPC gets your HPC solution up and running smoothly while reducing operational cost and complexity and improving service levels.

For more information:

sun.com/software/products/lustre/support.xml

Learn More

For more information on the Lustre File System, visit sun.com/lustre or the Lustre Community: www.lustre.org

And, to further ensure the success of your project, take advantage of Lustre administration training. In a packed 3-day course, Sun Learning Services covers best practices in deployment and administration as well as configurations from very simple to very complex.

Leading performance and scalability

Using the Lustre file system and Sun Open Storage products, the Sun Lustre Storage System eliminates I/O bottlenecks and provides very high rates of I/O bandwidth, scaling, and capacity. Through this simple to deploy, integrated system, the Sun Lustre Storage System, enables you to benefit from faster time to results, higher cluster utilization, completion of more jobs per a given period while delivering results to users more cost effectively.

For more information:

sun.com/servers/hpc/storagecluster/index.jsp