Why Use Lustre

Why Use Lustre™

Executive Summary

High Performance Computing (HPC) clusters are created to provide extreme computational power to large scale applications. This computational power often results in the creation of very large amounts of data as well as very large individual files. For quite some time, the speed of processors and memory have risen sharply, but the performance of I/O systems has lagged behind.

While processors and memory have improved in cost/performance exponentially over the last 20 years, disk drives still essentially spin at the same speeds, and drive access times are still measured in numbers of milliseconds. As such, poor I/O performance can severely degrade the overall performance of even the fastest clusters. This is especially true of today's multi-petabyte clusters.

The Lustre file system is a parallel file system used in a wide range of HPC environments, small to large, such as AI/ML, oil and gas, seismic processing, the movie industry, and scientific research to address a common problem they all have and that is the ever increasing large amounts of data being created and needing to be processed in a timely manner. In fact it is the most widely used file system by the world's Top 500 HPC sites.

With Lustre in use it's common to see end-to-end data throughput over 100GigE networks in excess of 10 GB/sec and InfiniBand EDR links reach bandwidths up to 10 GB/sec. Lustre can scale to tens of thousands of clients. At Oak Ridge National Laboratory their production file system, Spider, runs Lustre with over 25,000 clients, over 20PB of storage, and achieves a peak aggregate IO throughput of 2TB/sec.

This page addresses the many advantages that the Lustre file system offers to High Performance Computing clusters and how Lustre, a parallel file system, improves the overall scalability and performance of HPC clusters.

HPC Building blocks

When designing a High Performance Computing (HPC) cluster, the HPC architect has three common file system options for providing access to the storage hardware. Perhaps the most common is the Network File System (NFS). NFS is a standard component in cloud computing environments. NFS is commonly included in what is known as a NAS, or Network Attached Storage architectures. A second option available to choose is SAN file systems, or Storage Area Network file systems. And last, but not least, are parallel file systems. Lustre is the most widely used file system on the top 500 fastest computers in the world. Lustre is the file system of choice on 7 out of the top 10 fastest computers in the world today, over 70% of the top 100, and also for over 60% of the top 500. This paper describes why Lustre dominates the top 500 (www.top500.org) and why you would use it for your high performance IO requirements.

NFS (Network File System)

NFS has been around for over 20 years, is very stable, easy to use and most systems administrators, as well as users, are generally familiar with its strengths and weaknesses. In low end HPC storage environments, NFS can still be a very effective medium for distributing data, where low end HPC storage systems are defined as capacity under 100TB and high end generally above 1PB. However, for high end HPC clusters, the NFS server can quickly become a major bottleneck as it does not scale well when used in large cluster environments. The NFS server also becomes a single point of failure, for which the consequences of it crashing can become severe.

SAN (Storage Area Networks)

SAN file systems are capable of very high performance, but are extremely expensive to scale up since they are implemented using Fibre Channel and therefore, each node that connects to the SAN must have a Fibre Channel card to connect to the Fibre channel switch.

Lustre (a Distributed Parallel File System)

The main advantage of Lustre, a global parallel file system, over NFS and SAN file systems is that it provides; wide scalability, both in performance and storage capacity; a global name space, and the ability to distribute very large files across many nodes. Because large files are shared across many nodes in the typical cluster environment, a parallel file system, such as Lustre, is ideal for high end HPC cluster I/O systems. This is why it is in such widespread use today, and why at Whamcloud we have an organization of Lustre engineers dedicated to its support, service, and continued feature enhancements.

Lustre is implemented using only a handful of nodes connected to the actual storage hardware. These are known as Lustre server nodes, or sometimes as Lustre I/O nodes, because they serve up all the data to the rest of the cluster, which are typically known as compute nodes and often referred to as Lustre clients.
Lustre is used primarily for Linux based HPC clusters. Lustre is an open source file system and is licensed under the GPLv2. There are two main Lustre server components of a Lustre file system; Object Storage Servers (OSS) nodes and Meta Data Servers (MDS) nodes. File system Meta data is stored on the Lustre MDS nodes and file data is stored on the Object Storage Servers. The data for the MDS server is stored on a Meta Data Target (MDT), which essentially corresponds to any LUN being used to store the actual Meta data. The data for the OSS servers are stored on hardware LUNs called Object Storage Targets (OSTs). OST ldiskfs targets can currently be a maximum size of 1PB. Since we usually configure OSS/OST data LUNs in an 8+2 RAID-6 configuration, a common LUN configuration is ten 16 TB SATA drives. These are the most common configurations, but some sites do use SAS or NVMe drives for their OSTs.

For the most part, most all Meta data information and transactions are maintained by and on the Meta Data Servers. So, all common file system name space changes and operations are handled by the MDS node. This would include common things such as looking up a file, creating a file, most file attribute lookups and changes. When a file is opened, the client contacts the MDS server, which in turn looks up the file and returns to the client the location of that file on any and all OSS servers and their corresponding OSTs. Once the client has the location of the file data the client can do any and all I /O directly between the client and any OSS nodes without having to interact with the MDS node again. 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 simply the addition of more hardware. For example, Lustre throughput exceeds 2 TB/sec set at Oak Ridge National Labs (ORNL). This performance number is essentially limited only by the amount and characteristics of the storage hardware available. In this case, the file system size was 10.7 Petabytes.

Lustre in the Marketplace

The Lustre architecture is used for many different kinds of clusters, but it is best known for powering seven of the ten largest high-performance computing (HPC) clusters in the world, with some systems supporting over ten thousands clients, tens or hundreds of petabytes (PB) of storage and many of these systems with hundreds of gigabytes per second (GB/sec) or even terabytes per second (TB/sec) of I/O throughput.

A great deal of HPC sites use Lustre as a site-wide global file system, servicing dozens of clusters on an unprecedented scale. IDC shows Lustre as being the file system with the largest market share in HPC.

Scalability

The scalability offered by the Lustre file system has made it a popular choice in the oil and gas, manufacturing, and finance sectors.

Lustre's scalability reduces and often eliminates the common requirement of needing to create many separate file systems, such as one for each cluster or, perhaps worse yet, one for each NFS file server head. This feature of Lustre leads to very significant storage management advantages. For example, Lustre sites are able to avoid the maintenance of multiple copies of data staged on multiple file systems. Large HPC datacenters indicate that for this reason alone they require much less aggregate storage space with a Lustre file system than with other solutions.

Along with aggregating file system capacity across many servers, I/O bandwidth is also aggregated and also scales with the addition of more OSS servers. In addition to that, I/O throughput and/or storage capacity can be easily adjusted after the cluster is initially installed by just adding more OSS servers dynamically. This ability to simply add more OSS servers and storage allows the system to grow over time as the demands for more storage space and more bandwidth capabilities increase.

The larger Lustre file systems in use have tens of thousands of clients.

Open Source

The Lustre file system software is open source software, and because of this it keeps the pricing competitive because no one has a monopoly and its continued existence is not dependent upon the economic success of any single company. Since the file system itself is free and one pays just for support, this keeps the pricing competitive as you would pay only for the support of your Lustre file system.

Lustre Networking (LNet)

In a cluster using a Lustre file system, the Lustre network is the network connecting the OSS and MDS servers and the clients. The disk storage backing the MDS and OSS server nodes in a Lustre file system is connected to these Lustre I/O server nodes usually using traditional SAN technologies, however the breakthrough architecture with Lustre is that it does not need a SAN to extend or connect to Lustre clients. LNet is only used over the Lustre system network, wherein it provides the entire communication infrastructure needed by the Lustre file system.

Though TCP/IP over Ethernet is certainly supported, one of the key features of the LNet driver is its support of RDMA when such capability is supported by the underlying networks such as InfiniBand, RDMA over Converged Ethernet (RoCE), and OmniPath Architecture (OPA). This provides near full bandwidth utilization of the underlying link versus other architectures. LNet also supports many common network types such as IP and OFED along with simultaneous support of multiple network types connected together for which LNet will route packets between the different types of networks. LNet also provides Multi-Rail for performance and high availability and recovery that enable transparent recovery when used with Lustre failover servers.

Lustre Networking (LNet) performance is extremely high. As pointed out earlier, it is very common to see end-to-end throughput over 100 GigE networks in excess of 10 GB/sec, InfiniBand EDR links reach bandwidths over 10 GB/sec. These are data rates of the bandwidth delivered to clients from just one interface on a Lustre server, higher performance is available with multiple interfaces!

High Availability

Servers in a Lustre cluster are often equipped with a very large number of storage devices which Lustre can serve and handle up to tens of thousands of clients. In fact, over 29,000 clients have been handled in the largest production Lustre system so far. Any cluster file system should be able to handle server reboots or failures transparently through a high-availability mechanism such as failover to remove or reduce any single points of failure in the system. When a server fails for example, applications should merely see a delay in the response to their system calls executed in accessing the file system.

The lack of a reliable failover mechanism can lead to hanging or failed jobs, requiring applications to be restarted and sometimes requiring cluster reboots to clean up the mess created, all of which are very undesirable.

The Lustre failover mechanism delivers call completion that is completely application transparent. The Lustre failover implementation is robust and works in conjunction with software that offers interoperability between versions, which is needed to support rolling upgrades of file system software on active clusters.

The Lustre recovery feature, in conjunction with its failover capabilities, allows servers to be upgraded without the need to take the system down. Any one OSS or MDS server is simply taken offline, upgraded, and restarted (or failed over to a secondary/backup OSS/MDS server prepared with the new software). All active jobs continue to run without failures. The active jobs merely experience a delay while one of the OSS/MDS servers is being upgraded.

Lustre MDS servers are configured as an active/passive pair, while OSS servers are usually configured in an active/active configuration that provides redundancy without extra overhead. In many instances, the standby MDS is the active MDS for a second Lustre file system. This allows both MDS servers to be actively working at the same time, such that there are then no idle nodes in the cluster. In the case where you only have one big Lustre file system, then the secondary or backup MDS node would remain idle until such time as the primary MDS server encountered a problem that triggered a failover event.

Summary

The Lustre parallel file system is well suited for large HPC cluster environments and has capabilities that fulfill important I/O subsystem requirements. The Lustre file system is designed to provide cluster client nodes with shared access to file system data in parallel. Lustre enables high performance by allowing system architects to use any common storage technologies along with high-speed interconnects. Lustre file systems also can scale well as an organization's storage needs grow. And by providing multiple paths to the physical storage, the Lustre file system can provide high availability for HPC clusters. And maybe best of all, Lustre is now supported by Whamcloud and its many partners.