Lustre

---------------------------------------

Lustre是一个大规模的、安全可靠的，具备高可用性的集群文件系统，它是由SUN公司开发和维护的。

该项目主要的目的就是开发下一代的集群文件系统，可以支持超过10000个节点，数以PB的数据量存储系统。

目前Lustre已经运用在一些领域，例如HP SFS产品等。

Luster file system. Sun whitepaper.

Understanding Lustre Filesystem Internals （Luster）

http://wiki.lustre.org/images/d/da/Understanding\_Lustre\_Filesystem\_Internals.pdf

第一篇文献介绍了luster的架构和特性。第二篇文献从设计角度详细描述了Luster的内部结构。

Luster是目前在全球top100高性能计算机中部署最多的高性能并行文件系统。支持一个MDS，存储节点可扩展到上千台。

Luster为开源软件，目前主要由Sun公司维护开发。Luster的存储节点的文件系统由ext3修改而成，这部分成果成为后来ext4的基础。修改内容包括支持extent的存储空间管理和文件inode管理、支持大目录等。

Luster系统中client端和server采用Portal RPC协议，其主要特性类似于RDMA，以零拷贝方式传递数据，具有极高的性能。

在实现方式上，Luster和PVFS/GoogleFS不同，client端和server都在OS内核实现。

Luster支持对 MDS节点的Active/Standby方式的容错，但缺乏数据冗余保护。

1. 什么是元数据

任何文件系统中的数据分为数据和元数据。数据是指普通文件中的实际数据，而元

数据指用来描述一个文件的特征的系统数据，诸如访问权限、文件拥有者以及文件数据

块的分布信息(inode...)等等。在集群文件系统中，分布信息包括文件在磁盘上的位置以及磁盘在集群中的位置。用户需要操作一个文件必须首先得到它的元数据，才能定位到文件的位置并且得到文件的内容或相关属性。

2． 元数据管理方式

元数据管理有两种方式。集中式管理和分布式管理。集中式管理是指在系统中有一个节点专门司职元数据管理，所有元数据都存储在该节点的存储设备上。所有客户端对文件的请求前，都要先对该元数据管理器请求元数据。分布式管理是指将元数据存放在系统的任意节点并且能动态的迁移。对元数据管理的职责也分布到各个不同的节点上。大多数集群文件系统都采用集中式的元数据管理。因为集中式管理实现简单，一致性维护容易，在一定的操作频繁度内可以提供较满意的性能。缺点是单一失效点问题，若该服务器失效，整个系统将无法正常工作。而且，当对元数据的操作过于频繁时，集中的元数据管理成为整个系统的性能瓶颈。

分布式元数据管理的好处是解决了集中式管理的单一失效点问题， 而且性能不会随着操作频繁而出现瓶颈。其缺点是，实现复杂，一致性维护复杂，对性能有一定影响。