Lustre文件系统曲折而坚韧的发展史

2018-01-05 米搓 Lustre文件系统与DDN

Lustre是一种并行分布式文件系统，一般用于大规模集群。它的名字结合了Linux和Cluster。Lustre文件系统软件许可证为GNU General Public License（仅第2版）。Lustre可为从小型工作组群集到大型多站点群集提供高性能的文件系统。

因其具有高性能和开放的许可证，Lustre文件系统常被用于超级计算机 。 自2005年6月以来，它持续在超级计算机世界排名前十中占有至少一半的席位，前排名前100的超级计算机中，至少60个超级计算机使用Lustre，其中包括在2014年 TOP 500 超级计算机排名中的第二名和第三名，Titan和Sequoia。

Lustre文件系统具备可扩展性，可支持数万个客户端节点，数百个服务器，几十PB容量，以及每秒TB级的总I/O吞吐率。这使得Lustre文件系统成为大型数据中心企业（包括气象、仿真、石油和天然气，生命科学、富媒体和金融等行业）的热门选择。

历史演变

对Lustre文件系统体系结构的研究开始于当时在卡内基梅隆大学（CMU）工作的Peter J. Braam于1999年进行的一个研究项目。基于CMU的Coda项目中InterMezzo文件系统中的工作，Braam于2001年创建了自己的公司Cluster File Systems。当时，美国能源部联合Hewlett-Packard和Intel资助了 Accelerated Strategic Computing Initiative Path Forward项目，Lustre就是在这个项目下开发的。2007年9月，Sun Microsystems收购了Cluster File Systems Inc所有资产，包括知识产权。 Sun将Lustre和高性能计算硬件产品结合，目的是将Lustre技术引入到Sun的ZFS 文件系统和Solaris操作系统中 。 2008年11月，Braam离开了Sun Microsystems，Eric Barton和Andreas Dilger接管了这个项目。 2010年， 甲骨文公司收购了Sun，开始管理和发行Lustre。

2010年12月，甲骨文宣布将停止Lustre 2.x的开发，而针对Lustre 1.8则仅提供维护支持。这使得Lustre文件系统的未来发展充满了不确定性。 然而在这个公告发布之后，一些新机构开始如雨后春笋般涌现，它们以开放的社区发展模式提供支持和开发，如Whamcloud、Open Scalable File Systems Inc.（OpenSFS）、EUROPEAN Open File Systems （EOFS）等等。到2010年底，大部分Lustre开发者都离开了Oracle。在硬件导向的Xyratex公司收购ClusterStor之际，Braam和其他几个同事加入了Xyratex公司。而Barton、Dilger等人则成立了新兴软件公司Whamcloud，在那里他们继续着Lustre的研究开发。

2011年8月， OpenSFS把Lustre功能开发合同给了Whamcloud。 该合同包括：改进单个服务器元数据性能扩展从而使得Lustre能够更好地利用多核元数据服务器；在线Lustre分布式文件系统检查（LFSCK）以实现在文件系统挂载及运行时数据和元数据服务器之间的分布式文件

系统状态验证；分布式命名空间（DNE）及集群元数据（CMD），使Lustre元数据分布到多个服务器上。 基于ZFS的后端对象存储的开发同时也继续在Lawrence Livermore国家实验室进行着。 这些功能可在Lustre 2.2到2.4社区发布路线图中找到。 2011年11月，Whamcloud针对维护Lustre 2.x源代码签署了一份另外的单独合同，以确保在开发新功能的同时，Lustre代码将获得足够的测试和bug修复。

2012年7月，Whamcloud在赢得了FastForward DOE关于在2018年前将Luster扩展到E级计算系统的合同后，被英特尔收购。 随后， OpenSFS也就将Lustre开发合同转交给英特尔。

2013年2月，Xyratex Ltd.宣布从Oracle获得Lustre原始商标、logo、网站和相关知识产权。 2013年6月，英特尔开始将Lustre的使用扩展到传统的HPC之外，例如Hadoop 。 在2013年这一整年中， OpenSFS宣布了征求建议书（RFP），范围涵盖了Lustre功能开发，并行文件系统工具，Lustre技术债务处理，并行文件系统孵化器。 OpenSFS还建立了Lustre社区门户网站，这是一个集中提供信息和文档集，以供整个Lustre开源社区参考和指导的技术网站。 2014年4月8日，Ken Claffey宣布Xyratex/Seagate将lustre.org域名回赠给用户社区, 并于2015年3月完成。

发布历史

2003年3月，Lustre文件系统首次在Lawrence Livermore国家实验室的MCR Linux集群 （当时最大的超级计算机之一）上被安装用于生产。

2003年12月，Lustre 1.0.0发布，它提供了基本的Lustre文件系统功能，包括服务器故障切换和恢复。

2004年3月，Lustre 1.2.0发布，支持Linux内核 2.6，“size glimpse”功能避免了撤销正在进行写入文件的锁，实现了客户端的数据回写缓存的计量（grant）。

2004年11月，Lustre 1.4.0发布，提供了版本之间的协议兼容性，可使用InfiniBand网络，可利用ldiskfs磁盘文件系统中的extents/mballoc。

2007年4月，Lustre 1.6.0发布，允许使用“mkfs”和“mount”进行挂载配置（“mountconf”），允许动态添加对象存储目标（OST），实现了Lustre分布式锁管理器（LDLM ）在对称多处理 （SMP）服务器上的可扩展性，并为对象分配提供了空闲空间管理。

2009年5月，Lustre 1.8.0发布，提供了OSS读取缓存，提高了在多种故障下的恢复能力，通过OST池增加了基本的异构存储管理，自适应网络超时和基于版本的恢复。 这是一个过渡版本，可与Lustre 1.6和Lustre 2.0兼容。

2010年8月，Lustre 2.0发布，进行了重要的内部代码重构，为主要的架构改进做好了准备。

Lustre 2.x *客户端*无法与1.8或更早版本的服务器进行交互。 但是，Lustre 1.8.6及更高版本的客户端可以与Lustre 2.0及更高版本的服务器进行交互。 1.8的元数据目标（MDT）和OST磁盘格式可以升级到2.0及更高版本，而无需重新格式化文件系统。

2011年9月，Lustre 2.1发布，是整个社区对Oracle暂停Lustre 2.x开发的回应。 它增加了在Red Hat Linux 6上运行服务器的能力，并将基于ext4的OST的最大容量从24TB增加到了128TB，同时还提高了性能和稳定性。 Lustre 2.1服务器与1.8.6及更高版本的客户端保持兼容。

2012年3月，Lustre 2.2发布，其重点在于改进元数据性能和加入新功能。 它增加了并行目录操作，允许多个客户端同时遍历和修改单个大目录， 能更快地从服务器故障中恢复，单个文件的条带数增加（最多可达2000个OST），并且改进了单客户端目录遍历性能。

2012年10月，Lustre 2.3发布，继续改进了元数据服务器代码，以消除具有多个CPU核心（超过16个）的节点上的内部锁定瓶颈。 对象存储层初步具备了使用ZFS作为后端文件系统的能力。 Lustre文件系统检查（LFSCK）功能可以在文件级备份/恢复之后，或MDS损坏的情况下，当文件系统正在使用时，进行MDS对象索引（OI）的验证和修复。 服务器端IO统计信息功能得到了增强，允许与批处理作业调度程序（如SLURM）集成，以跟踪每个作业的统计信息。 客户端软件已经升级到可以使用3.0版本的Linux内核。

2013年5月，Lustre 2.4发布，增加了相当多的主要特性，许多项目直接通过OpenSFS获得资助。 通过允许单个名称空间的子目录树位于单独的MDT上，分布式名称空间（DNE）实现了元数据容量和性能的线性可扩展性。ZFS可用作MDT和OST存储的后端文件系统。 LFSCK功能添加了扫描和验证MDT FID和LinkEA属性的内部一致性的功能。 网络请求调度程序（NRS）添加策略来优化客户端请求处理，以进行磁盘排序或提高公平性。 客户端可以选择发送大小为4 MB的批量RPC。 客户端软件已经升级到可以使用3.6版本的Linux内核，并且仍然可以与1.8客户端兼容。

2013年10月，Lustre 2.5发布，增加了备受期待的功能——分级存储管理 （HSM）。 HSM是企业环境中的核心要求，它使得客户可以在其操作环境中轻松实施分层存储解决方案。 该版本是Lustre当前OpenSFS指定的维护版本分支。 最新的维护版本是2.5.3，于2014年9月发布。

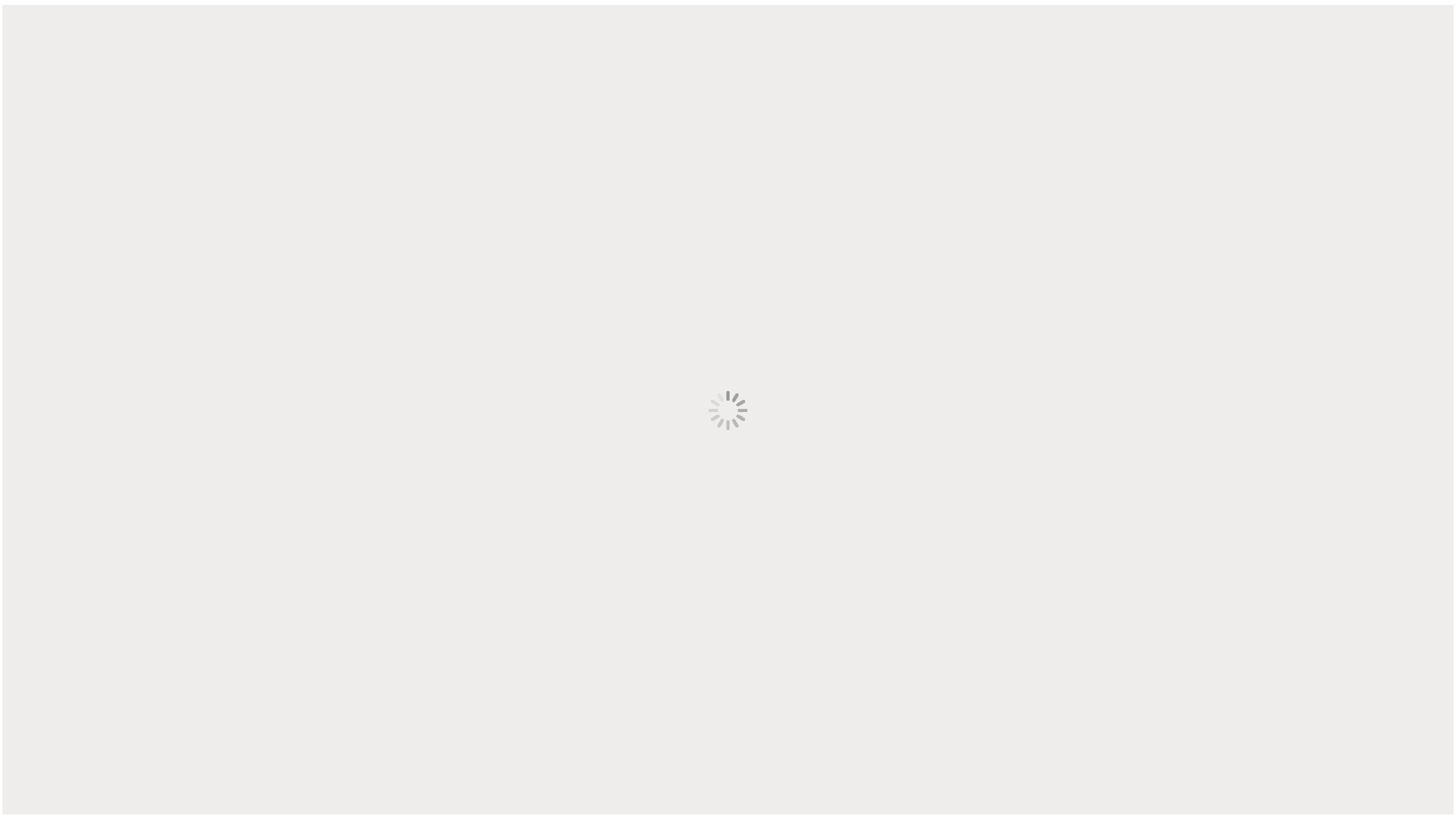
2014年7月，Lustre 2.6发布， Lustre 2.6版本功能上讲是一个朴素的版本，增加了LFSCK功能来对OST进行本地一致性检查以及MDT和OST对象之间的一致性检查。 与以前的版本相比，单客户端IO性能得到了提高。 此版本还添加了DNE分条目录的预览，使得将单个大目录可以存储在多个MDT上，以提高性能和可扩展性。

2015年3月，Lustre 2.7发布， 增加了LFSCK功能来验证多个MDT之间远程目录和条带目录的DNE一致性。 动态LNet配置添加了在运行时配置和修改LNet网络接口，路由和路由器的能力。 为具有不同管理域的客户端的UID / GID映射添加了新的评估功能，同时还改进了DNE条带化目录功能。

2016年3月，Lustre 2.8发布， 完成了DNE条带目录功能，包括支持在MDT之间迁移目录，以及跨MDT硬链接和重命名。 此外，还包括对客户端上安全增强型Linux身份验证和RPC加密以及LFSCK的性能改进。

2016年12月，Lustre 2.9发布，包含了许多与安全性和性能有关的功能。 共享密钥安全特性使用与Kerberos相同的GSSAPI机制来提供客户机和服务器节点身份验证以及RPC消息完整性和安全性（加密）。 Nodemap功能允许将客户端节点分组，然后映射这些客户端的UID/GID，使得远程管理的客户端可以透明地使用共享文件系统，而无需为所有客户端节点都配置一组UID/GID。 子目录挂载功能允许客户端从MDS挂载文件系统命名空间的一个子集。 此版本还增加了对最大16MiB RPC的支持，以便实现更高效的磁盘I/O，并添加了ladvise接口以允许客户端向服务器提供I/O提示，以便将文件数据预取到服务器缓存或从服务器刷新文件数据缓存。 改进了对指定文件系统范围的默认OST池的支持，并改进了OST池继承以及其他文件布局参数。

2017年7月，Lustre 2.10发布，有了许多重大改进。 LNet Multi-Rail功能允许在客户端和服务器上绑定多个网络接口（InfiniBand，OPA和/或以太网），以增加总的I/O带宽。 文件格式现在可以由多个组件构成，基于文件偏移量，允许根据文件大小确定不同的参数，例如条带数，OST池等。NRS令牌桶过滤器（TBF）服务器端调度器已经实现了新的规则类型，包括RPC类型的调度以及指定多个参数的能力，例如用于规则匹配的JobID和NID。 添加了用于管理Lustre文件系统的ZFS快照的工具，它作为单独的Lustre加载点简化了MDT和OST ZFS快照的创建，安装和管理。



长按下图，点击“识别图中二维码”，关注微信公众号“Lustre文件系统与DDN”，随时获取

Lustre文件系统的相关资讯。