安装了Intel Enterprise Edition for Lustre软件的戴尔HPC存储设备

戴尔技术白皮书

Quy Ta

Dell HPC工程部

2014年11月 | 版本1.0

目录

图

表

1. 简介

在高性能计算领域，高效地将数据传入和传出计算节点至关重要，这通常会涉及到一些复杂的因素。研究人员在HPC系统中能够以极高的速度产生和使用数据，不过与这种强大的计算能力相比，存储组件往往会成为整个HPC系统的瓶颈。此外，管理和监控复杂的存储系统也会增加存储管理员和研究人员的负担。存储的扩展也是个挑战，在HPC系统中，数据对性能和容量的需求会持续地快速增加，因此系统管理员需要进行精心的规划和充分的配置，才能不断提高存储的吞吐量和性能，从而为整个HPC系统提供有力的支持。

在本文档的其余部分，我们将装有Intel Enterprise Edition for Lustre软件的戴尔HPC存储设备称为安装了Intel EE for Lustre软件的戴尔HPC存储设备，此款存储设备是为那些需要部署完全受支持、易于使用、具有高吞吐量、能够横向扩展且经济实惠的并行文件系统存储解决方案的学术和行业用户而设计的。安装了Intel EE for Lustre软件的戴尔HPC存储设备是一款可横向扩展的存储解决方案设备，能够提供高性能和高可用的存储系统。

该解决方案利用智能、丰富且直观的管理界面（即Intel Manager for Lustre (IML)），大大简化了所有硬件和存储系统组件的管理和监控。它在容量或/和性能方面易于扩展，从而为未来增长提供了方便的途径。该存储解决方案使用Lustre®这一领先的HPC开源并行文件系统1。

该存储解决方案利用第十三代企业级Dell PowerEdge™服务器和最新一代的高密度PowerVault™存储产品。安装了Intel EE for Lustre软件的戴尔HPC存储设备解决方案由戴尔和英特尔提供全面的硬件和软件支持，是兼具高性能、高可靠性、高密度、高易用性和成本效益的出色产品。

本白皮书的后续部分将依次介绍Lustre文件系统、安装了Intel EE for Lustre软件的戴尔HPC存储设备解决方案、性能分析结果以及总结。附录A：基准命令参考

2. Lustre文件系统

Lustre是一种并行文件系统，它通过并行数据访问和分布式锁定功能来提供高性能。一个Lustre安装实例由以下三个关键要素组成：元数据子系统、对象存储子系统（数据）和用来访问和操作数据的计算客户端。

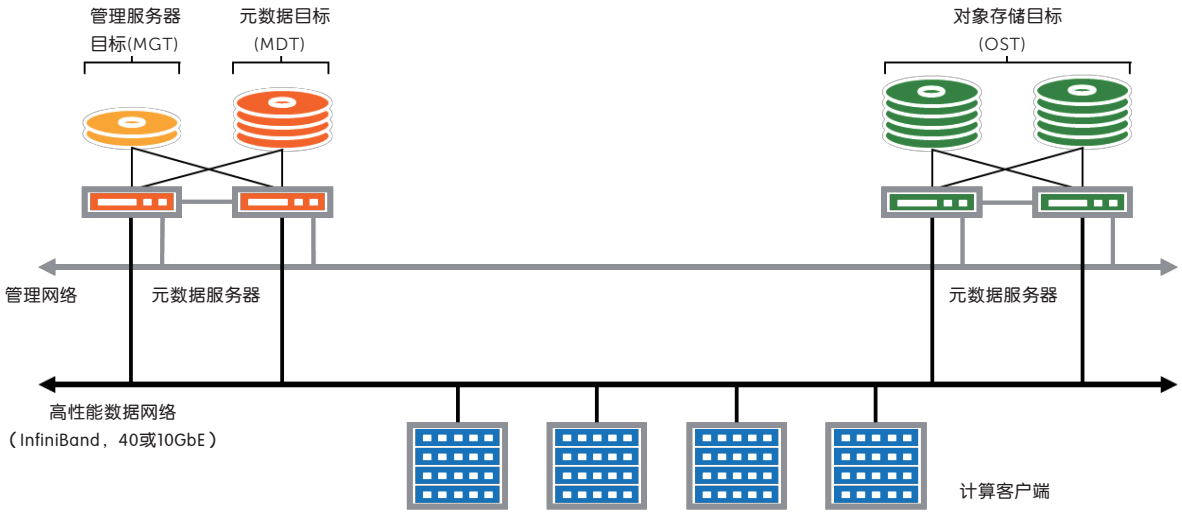
元数据子系统由元数据目标(MDT)、管理目标(MGT)和元数据服务器(MDS)组成。MDT存储文件系统的所有元数据，其中包括文件名、权限、时间戳和数据对象在对象存储系统中的位置。MGT存储管理数据（如配置信息和注册表）。MDS是用来管理MDT的专用服务器。

对象存储子系统由一个或多个对象存储目标(OST)和一个或多个对象存储服务器(OSS)组成。OST为文件

新一代的Lustre软件将HPC扩展到商业企业中

对象数据提供存储，而每个OSS则管理一个或多个OST。通常，在任何时候都有多个OSS处于活动状态。Lustre能够通过增加活动OSS（和相关OST）的数量来提高吞吐量。每增加一个OSS都会提高现有的网络吞吐量，而每增加一个OST都增加存储容量。图1显示典型的Lustre配置中MDS、MDT、MGS、OSS和OST组件之间的关系。该图中的客户端是HPC群集的计算节点。

图1：基于Lustre的存储解决方案的组件



并行文件系统（如Lustre）通过跨多个对象存储目标(OST)分布数据（"条带化"数据）来提供高性能和可扩展性，从而使多个计算节点能够同时以高效方式访问数据。在设计Lustre时，需要考虑的一个关键因素就是将元数据访问与IO数据访问分开以改善总体系统性能。

Lustre客户端软件安装在计算节点上以允许访问Lustre文件系统上存储的数据。对于客户端来说，文件系统显示为单个在装入后可进行访问的命名空间。由于只需要这一个装入点，因此为访问应用程序数据提供一个简单的起点，并且还可以通过本机客户端操作系统工具进行访问，从而更方便管理。

Lustre包括一个先进的增强型存储网络协议，该协议的名称是Lustre 网络（简称LNet）。LNet能够利用某些类型的网络功能。例如，当安装了Intel EE for Lustre软件的戴尔HPC存储设备使用InfiniBand作为网络来连接客户端、MDS和OSS时，LNet使Lustre得以利用InfiniBand结构的RDMA功能，提供比典型网络协议更快的I/O传输和更短的延迟时间。

总之，Lustre文件系统中包含以下要素：

- 元数据目标(MDT)——存储数据"条带"的位置、文件名、时间戳等。
- 管理目标(MGT)——存储管理数据（如配置和注册表）。
- 元数据存储服务器(MDS)——管理MDT并为Lustre客户端提供对文件的访问。
- 对象存储目标(OST)——存储文件系统上文件的数据条带或扩展区。
- 对象存储服务器(OSS)——管理OST并为Lustre客户端提供对数据的访问。

- Lustre客户端——访问MDS以确定文件所在的位置，然后访问OSS以读取和写入数据。

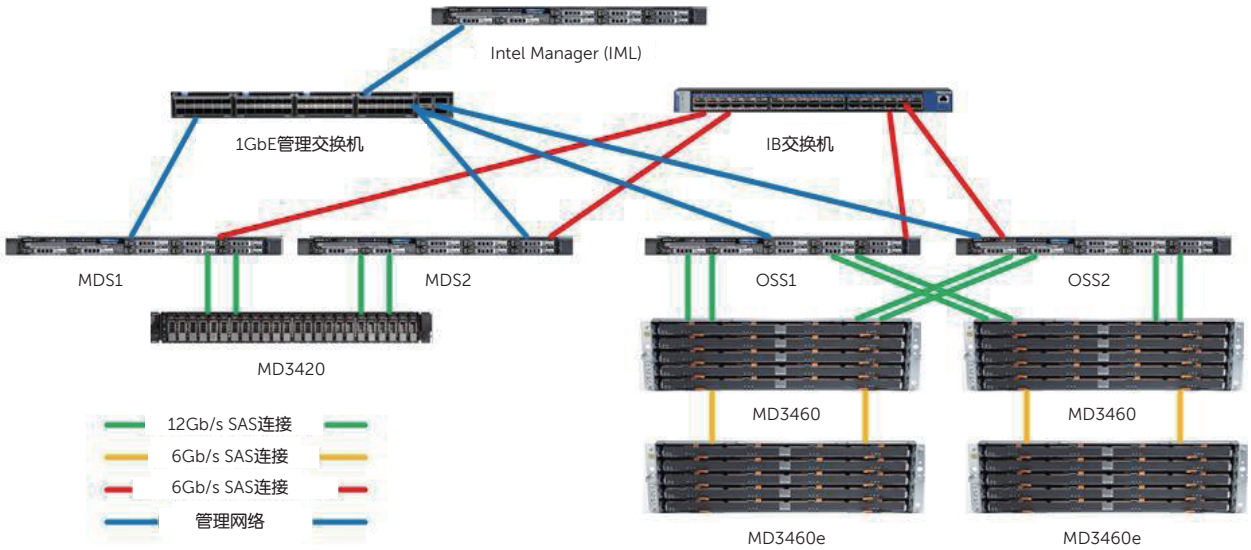
通常，Lustre部署和配置被视为非常复杂且相当耗时的任务。通常，Lustre安装和管理是通过命令行界面(CLI)完成的，这需要全面了解文件系统操作、辅助工具（如LNet）和锁定机制。另外，一旦Lustre存储系统就绪，保持系统和性能优化可能是一项艰巨的任务。这样的要求以及与它们相关联的陡峭学习曲线可能会阻止不熟悉Lustre的系统管理员执行安装，有可能会阻止他们所在的组织体验并行文件系统的好处。即便对于有经验的Lustre系统管理员来说，维护Lustre文件系统也可能会占据其很大一部分时间。

3. 安装了Intel EE for Lustre软件的戴尔HPC存储设备说明

通常，Lustre部署和配置被视为非常复杂且相当耗时的任务。通常，Lustre安装和管理是通过命令行界面(CLI)完成的，这需要全面了解文件系统操作、辅助工具（如LNet）和锁定机制。另外，一旦Lustre存储系统就绪，保持系统和性能优化可能是一项艰巨的任务。这样的要求以及与它们相关联的陡峭学习曲线可能会阻止不熟悉Lustre的系统管理员执行安装，有可能会阻止他们所在的组织体验并行文件系统的好处。即便对于有经验的Lustre系统管理员来说，维护Lustre文件系统也可能会占据其很大一部分时间。

- 管理服务器(IML)
- 元数据服务器对(MDS)
- 对象存储服务器对(OSS)

图2：安装了Intel EE for Lustre软件的戴尔HPC存储设备组件概述



安装了Intel Enterprise Edition for Lustre软件的戴尔HPC存储设备

在该配置中，安装了Intel EE for Lustre软件的戴尔HPC存储设备解决方案将Dell PowerEdge R630服务器平台用作管理服务器、对象存储服务器和元数据服务器。该解决方案支持Mellanox ConnectX-3 InfiniBand FDR (56 Gb/s)适配器，该适配器利用戴尔的第十三代服务器所支持的PCIe 3.0。或者，还支持使用10 Gb/s以太网连接到客户端。为了说明安装了Intel EE for Lustre软件的戴尔HPC存储设备解决方案的最大能力，本研究仅重点关注基于InfiniBand的配置在不同FDR速度下的性能。通过图3中显示的PowerEdge R630能够在1U外形规格中实现服务器密度、性能和可维护性。

图3：Dell PowerEdge R630

3.1 管理服务器

Intel Manager Server是通过内部1GbE网络与元数据服务器和对象存储服务器相连的单个服务器。

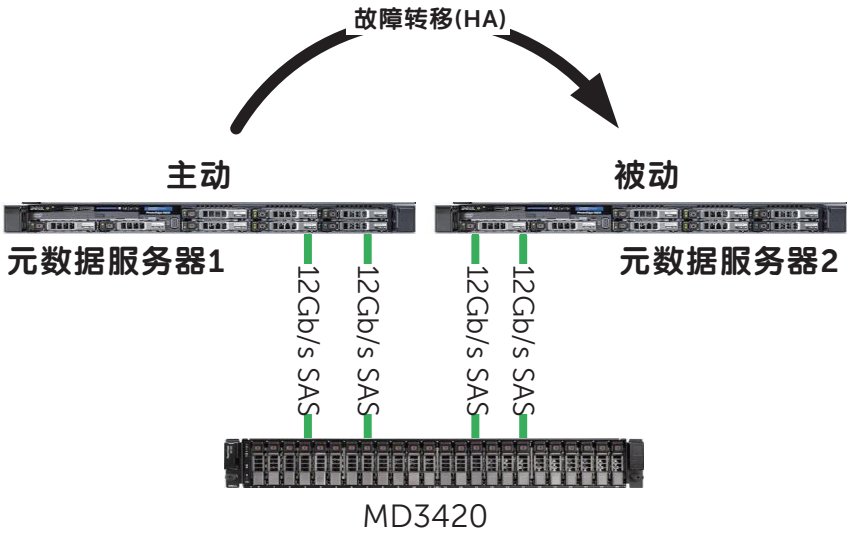
管理服务器负责用户交互、系统运行状况管理和基本的监控数据，这些数据是通过交互式Web GUI控制台Intel Manager for Lustre来收集和提供的。对于安装了Intel EE for Lustre软件的戴尔HPC存储设备的所有管理访问都将通过此服务器来执行。虽然管理服务器负责收集与Lustre文件系统相关的数据并为该解决方案提供管理，但它在Lustre文件系统或数据路径本身中不扮演主动操作角色。

Intel Manager for Lustre (IML) GUI可降低安装的复杂程度并尽可能缩短Lustre部署和配置时间。它还自动监控不同组件的运行状况和性能。缩减部署和配置工作所需的时间和工作量可加快为投入生产所做的一般准备。自动监控可为最终用户提供更好的服务，而不会增加系统管理员的负担。另外，使用该解决方案所提供的工具，可帮助解决与文件系统性能相关的问题。最后，监控工具能够保留历史信息，通过监控工具可以为扩展、维护和升级存储设备做更好的规划。

3.2 元数据服务器

元数据服务器对（如图4中所示）由两个Dell PowerEdge R630服务器组成，这两个服务器配置为主动/被动高可用性群集。每个服务器都直接连接到单个用来存放Lustre MDT和MGT的Dell PowerVault MD3420存储阵列。Dell PowerVault MD3420用24个300 GB、15K RPM、2.5英寸近线SAS驱动器完全填充，这些驱动器配置于带有2个热备盘的22磁盘RAID10中。在该元数据目标(MDT)中，该解决方案为文件系统元数据提供大约3TiB空间。MDS负责处理文件和目录请求并将任务路由到相应的对象存储目标上执行。使用单个具有此大小的MDT时，最多可以为16亿多文件提供服务。在该解决方案中，存储请求由单个56 Gb/s FDR InfiniBand连接通过LNet来处理。

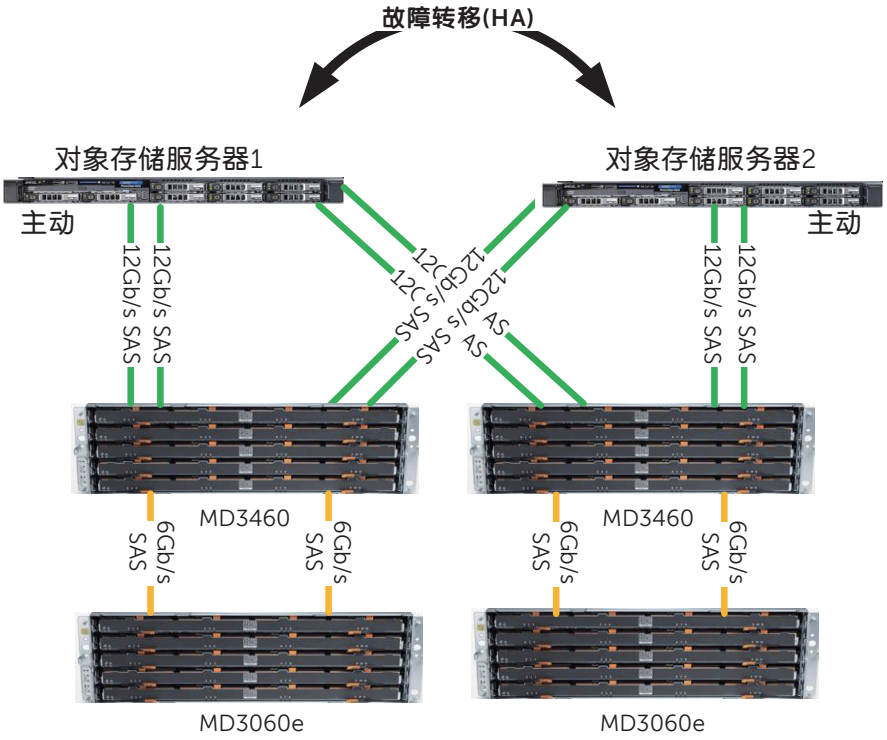
图4：元数据服务器对



3.3 对象存储服务器

对象存储服务器（如图5中所示）排列在双节点高可用性(HA)群集中，它们提供对两个Dell PowerVault MD3460高密度存储阵列（每个阵列都有MD3060e扩展盘柜）的主动/主动访问。每个PowerVault MD3460阵列都用60个4TB的3.5英寸NL SAS驱动器完全填充。每个PowerVault MD3460阵列的容量都用一个额外的PowerVault MD3060e高密度扩展阵列进行扩展。此配置为每个OSS对提供960TB的原始存储容量。

图5：对象存储服务器对

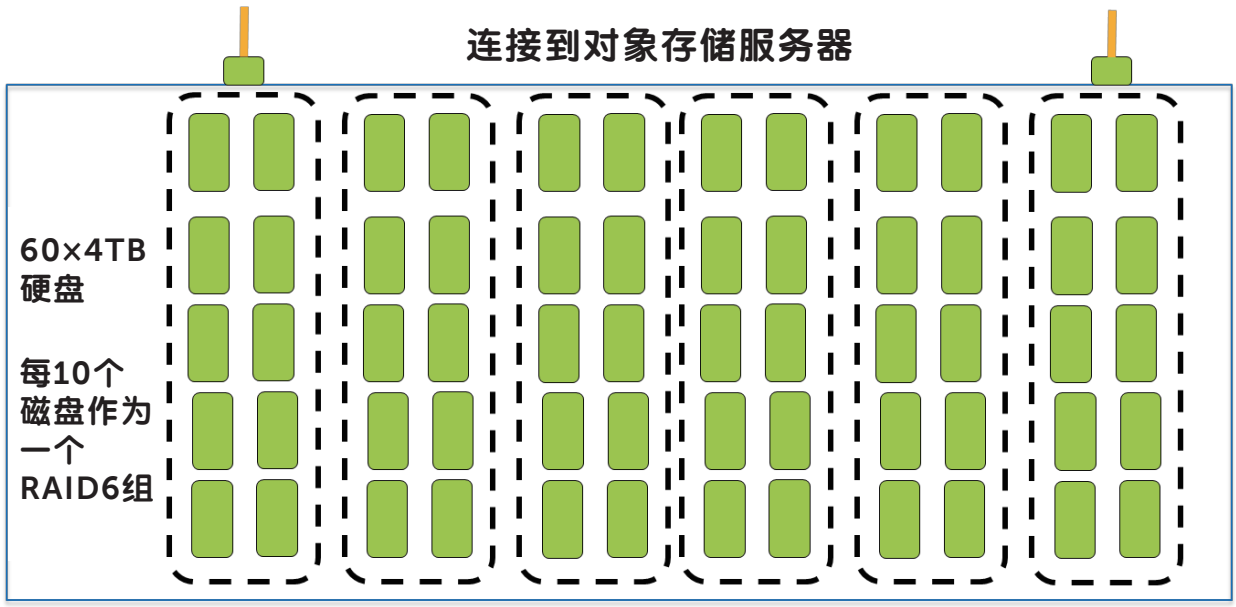


对象存储服务是该解决方案的构建块。利用每个PowerEdge R630中的两个双端口12Gb/s SAS控制器，两个服务器均以冗余方式连接到两个PowerVault MD3460高密度存储阵列。

图6说明了如何将每个存储阵列分成六个RAID 6虚拟磁盘（每个虚拟磁盘中包含八个数据磁盘和两个奇偶校验磁盘），并在每个阵列托架中使用两个磁盘。这会在每个盘柜中生成六个对象存储目标。通过使用RAID 6，该解决方案以边际成本针对写入性能提供更高的可靠性（由于每个RAID 6需要一组额外的奇偶校验数据）。每个OST提供大约29TiB的格式化对象存储空间。使用安装了Intel EE for Lustre软件的戴尔HPC存储设备解决方案，单个OSS对通过向MD3460阵列中添加PowerVault MD3060e扩展阵列拥有24个OST。OST通过56 Gb/s Infiniband FDR或10Gb/s以太网连接，使用LNet提供给客户端。

在从配有Lustre客户端的任何计算节点查看时，整个命名空间可以像任何其他文件系统那样查看和管理，但它具备Lustre管理增强功能。

图6：MD3460或MD3060e阵列上的RAID6布局



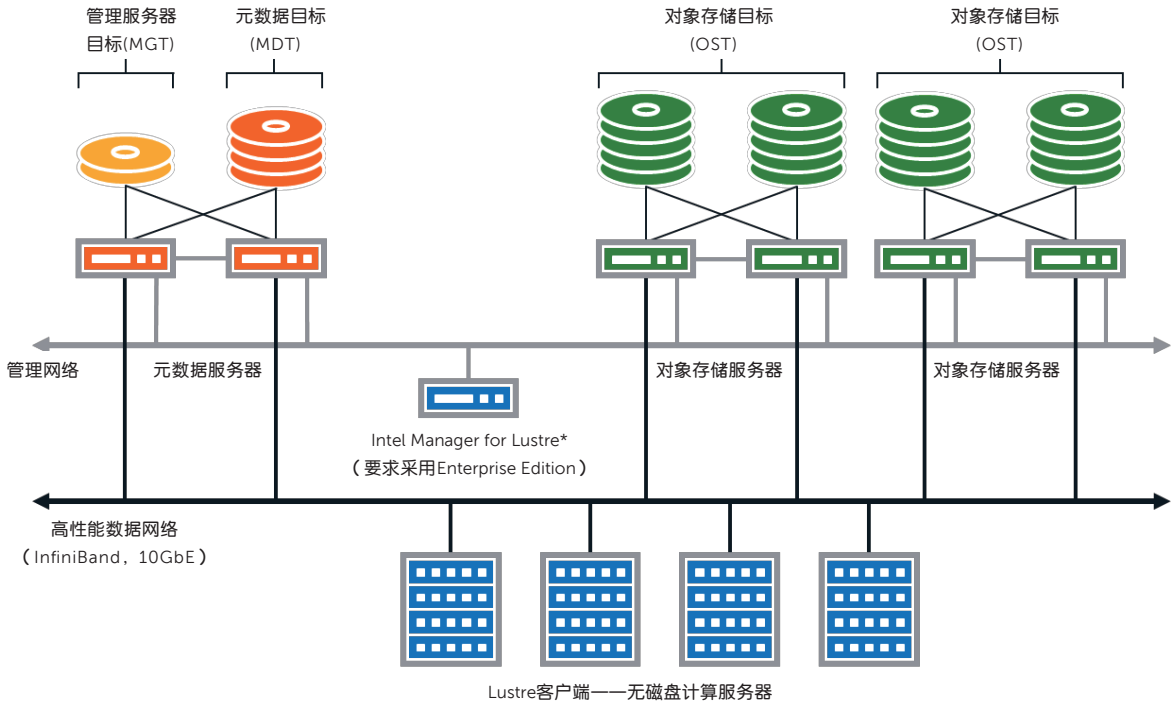
3.4 可扩展性

在主动/主动群集配置中提供对象存储服务器会生成更大的吞吐量和产品可靠性。此配置提供高可用性，它降低维护要求，并因此缩短潜在的停机时间。

PowerEdge R630服务器提供性能和密度。该解决方案为每个OSS对提供960TB的原始存储。该解决方案还利用FDR InfiniBand互连实现极高速度的低延迟存储事务，或者利用10Gb/s以太网实现高速、低成本并允许使用现有的10GbE基础架构。PowerEdge R630针对FDR InfiniBand使用PCIe Gen3接口，帮助实现更高的每OSS网络吞吐量。

对于具有Mellanox OFED版本2.2-1的RHEL6.5内核，基于RPM的Lustre 2.5.23版客户端可用于访问安装了Intel EE for Lustre软件的戴尔HPC存储设备（有关详细信息，请参见装有Intel EE Lustre的HPC存储设备配置指南）。

图7：OSS的可扩展性



可以通过添加具有存储后端的额外OSS对来扩展安装了Intel EE for Lustre软件的戴尔HPC存储设备，如**图7**中所示。因此，总网络吞吐量和存储容量将立即增加。这样，会在网络吞吐量保持最大的情况下增加可用存储量。

3.5 网络

3.5.1 管理网络

专用管理网络为Lustre和Lustre HA功能提供通信基础架构，还提供存储配置、监控和维护。此网络会创建为了便于执行日常操作和限制故障排除与维护范围而所需的分段。管理服务器使用此网络与不同的解决方案组件进行交互，以查询和收集系统运行状况信息，并执行由管理员启动的任何管理变更。OSS和MDS服务器与管理服务器进行交互，以便提供运行状况信息和性能数据并在执行管理操作期间进行交互。从带外（外部端口）访问PowerVault MD3420和MD3460控制器，以监控存储阵列的运行状况并针对存储后端执行管理操作。

经验不足的操作人员甚至可以使用这种级别的集成，毫不费力地高效监控和管理该解决方案。对所提供的信息进行了汇总以供用户快速查看，但是用户可以将服务器组件或存储组件的消息放大到所需的详细程度。

3.5.2 数据网络

Lustre文件系统通过InfiniBand FDR或/和10GbE上实施的Lustre网络(LNet)获得服务。客户端正是使用此网络来访问数据。Intel Manager for Lustre(IML) GUI界面提供了一个用来将MDS和OSS服务器

上的多个Lustre网络标识符(NID)服务器配置为参与Lustre网络的选项。例如，您应当在OSS服务器上将Infiniband接口（即ib0）和10GbE以太网接口（即eth0）配置为均参与Lustre网络。

在InfiniBand网络中，可以实现较快的传输速度和较短的延迟时间。LNet利用RDMA在MDT和OST与客户端之间快速传输数据和元数据。OSS和MDS服务器利用具有单端口Mellanox ConnectX-3 56 Gb适配器的InfiniBand结构。必要时，可以将FDR InfiniBand HBA集成到现有的QDR或DDR网络中。有了10GbE网络，Lustre仍可以在利用现有10GbE基础架构的情况下，从快速传输速度获益并利用以太网技术的较低成本和普遍性。

3.6 管理安装了Intel EE for Lustre软件的戴尔HPC存储设备

Intel Manager for Lustre (IML)通过提供一个用于进行管理的集中式Web GUI，使Lustre文件系统的管理不再复杂。例如，可以使用IML作为对以下操作进行标准化的工具：启动从一个节点到另一个节点（对于OSS或MDS）的文件系统故障转移、格式化文件系统、发出对目标的装入和卸载命令、监控Lustre文件系统的性能和其各个组件的状态。图8说明了几个IML监控界面。

IML是基于Web的管理控制台，可用来管理解决方案（假设满足了所有的安全要求）。它提供硬件、软件和文件系统组件的视图，而且还可用于进行监控和管理。

如果使用IML，那么，以前需要复杂CLI指令的许多任务现在只需单击几次鼠标即可轻松完成。IML可用于关闭文件系统、启动从一个MDS到另一个MDS的故障转移、进行监控等。

图8：Intel Manager for Lustre (IML)界面

4. 性能评估和分析

本白皮书中提供的性能研究概述了安装了Intel EE for Lustre软件的戴尔HPC存储设备的240驱动器配置的功能。该配置包含240个4TB的磁盘驱动器（960 TB原始空间）。其目标是量化该解决方案的功能、峰值性能点和最适合的扩展方法。用来提供I/O工作负载以测试该解决方案的客户端测试台是基于PowerEdge M610刀片服务器的Dell HPC计算群集，**表1**对该群集的配置进行了说明。

以具有不同类型工作负载的配置为重点进行了许多性能研究，以便确定性能极限并定义该性能的可持续性。之所以在这些研究中使用InfiniBand，是因为得益于它的高速和低延迟特点，可以从安装了Intel EE for Lustre软件的戴尔HPC存储设备获得最高性能，从而避免网络瓶颈。

表1：测试客户端群集的详细信息

组件	说明
计算节点：	Dell PowerEdge M610，64个节点
节点BIOS：	6.3.0
处理器：	两个英特尔至强™ X5650（频率为2.67GHz）六核处理器
内存：	24GiB DDR3 1333MHz
互连：	InfiniBand——Mellanox Technologies M3601Q (QDR)

Lustre：	Lustre 2.5.23——Mellanox OFED客户端
操作系统：	Red Hat Enterprise Linux 6.5 (2.6.32-431.el6.x86_64)
IB软件：	Mellanox OFED 2.2-1

性能分析侧重于三个主要的性能标记：

- 吞吐量，即按顺序传输的数据量(GB/s)。
- 每秒执行的I/O操作数(IOPS)。
- 每秒执行的元数据操作数(OP/s)。

我们的目标是广泛而准确地检查安装了Intel EE for Lustre软件的戴尔HPC存储设备的功能。我们选择以下三个基准来完成我们的目标：IOzone、IOR和MDtest。

可以将两种类型的文件访问方法与这些基准一起使用。第一种文件访问方法是多对多，在该方法中，基准的每个线程（多个客户端）都写入存储系统上一个不同的文件（多个文件）。IOzone和IOR都可以配置为使用多对多文件访问方法。在本研究中，我们针采用对多对多访问法的工作负载使用IOzone。第二种文件访问方法是多对一，在该方法中，所有的线程都写入同一个文件（多个客户端，一个文件）。在本研究中，我们针对采用多对一访问法的工作负载使用IOR。IOR可以使用MPI-IO、HDF5或POSIX运行多对一文件访问测试。为了进行分析，我们使用的是POSIX。多对一测试确定文件系统如何处理在多个线程写入或读取同一个文件时，由多个并发请求引入的开销。所遇到的开销来自那些处理Lustre文件锁定和序列化写入的线程。有关用来运行这些机制的命令的示例，请参见附录A。

会针对一定范围的客户端运行每组测试，以测试该解决方案的可扩展性。每个测试中涉及的并发物理客户端数从单个客户端变化到64个客户端。线程数量与物理服务器数量相对应，最多64个。高于64的线程总数是通过增加所有客户端中每个客户端的线程数来实现的。例如，对于128个线程，64个客户端中的每个客户端都运行2个线程。

该解决方案的测试环境具有单个MDS和单个OSS对，总原始磁盘空间为960TB。该OSS对包含两个PowerEdge R630（每个R630具有256GB内存）、两个12Gbps SAS控制器和单个Mellanox ConnectX-3 FDR HCA。有关布线和扩展卡位置的详细信息，请查阅《Dell Storage for HPC with Intel EE for Lustre Configuration Guide》（安装了Intel EE for Lustre软件的戴尔HPC存储设备配置指南）。MDS具有相同的配置，即具有256GB内存、单个Mellanox ConnectX-3 FDR HCA和两个12Gbps SAS控制器。

InfiniBand结构由一个32端口的Mellanox M3601Q QDR InfiniBand交换机（用于客户端群集）和一个36端口的Mellanox SX6025 FDR InfiniBand交换机（用于安装了Intel EE for Lustre软件的戴尔HPC存储设备服务器）组成。M3601Q交换机上的三个端口也连接到SX6025交换机。

表2显示了有关不同软件和硬件组件的特征的详细信息。

图8：Intel Manager for Lustre (IML)界面

配置大小	960TB原始空间
Lustre服务器版本	2.5.23
Intel EE for Lustre版本	v2.1
OSS节点	2个PowerEdge R630服务器
OSS内存	256GiB DDR4 2133MT/s
OSS处理器	2个英特尔至强™ E5-2660V3（频率为2.60GHz）十核处理
OSS服务器BIOS	0.3.28
OSS存储阵列	2个PowerVault MD3460，2个PowerVault MD3060e
OSS存储阵列中的驱动器	240个3.5英寸4 TB 7.2K RPM NL SAS
OSS SAS控制器	2个SAS 12Gbps HBA LSI 9300-8e
MDS节点	2个PowerEdge R630服务器
MDS内存	256GiB DDR4 2133MT/s
MDS处理器	2个英特尔至强™ E5-2660V3（频率为2.60GHz）十核处理
MDS服务器BIOS	0.3.28
MDS存储阵列	1个PowerVault MD3420
MDS存储阵列中的驱动器	24个2.5英寸300GB NL SAS
MDS SAS控制器	1个SAS 12Gbps HBA LSI 9300-8e
数据网络——InfiniBand	
OSS，MDS服务器	Mellanox FDR HCA MT27500
计算节点	Mellanox QDR HCA MT26428
客户端QDR IB交换机	Mellanox M3601Q
HSS5.5 FDR IB交换机	Mellanox 36端口SX6036
IB交换机连接	客户端：QDR电缆；服务器：FDR电缆 3个从QDR交换机到FDR交换机的上行链路

为了防止由于高速缓存效应而导致的膨胀结果，使用借助于以下技术建立的冷高速缓存执行测试。在执行每个测试之前，会重新装入待测试的Lustre文件系统。会执行同步，系统会使用以下命令指示内核丢弃所有客户端上的高速缓存：

- sync
- echo 3 > /proc/sys/vm/drop_caches

另外，为了模拟服务器上的冷高速缓存，在启动每个测试之前，在所有的主动服务器（OSS和MDS）上，执行"**同步**"，并使用与客户端上相同的命令指示内核丢弃高速缓存。

在衡量安装了Intel EE for Lustre软件的戴尔HPC存储设备解决方案的性能时，所有测试是使用相似的初始条件执行的。文件系统配置为完全正常，在执行每个测试之前会倒空所测试目标的文件和目录。

4.1 多对多顺序读取/写入

使用3.429版的IOzone测试工具执行顺序测试。在图9中显示的吞吐量结果已转换为MB/s。为该测试选择的文件大小可保证所有线程的聚合样本大小保持2TB不变。也就是说，顺序读写会在该测试中的所有线程之间平均划分聚合样本大小(2TB)。IOzone的数据块大小设置为1 MiB，以便与1 MiB Lustre请求大小相匹配。

所写入的每个文件都足够大，能够容纳OSS和客户端的高速缓存效应。另外，其他防止出现高速缓存效应的技术也帮助避免出现高速缓存效应。所写入的文件会在OST之间平均分布（采用循环法）。这是为了防止I/O负载在任一SAS连接或OST上不均匀，用户希望按照同样的方法平衡工作负载。

图9：按顺序读取/写入安装了Intel EE for Lustre软件的戴尔HPC存储设备

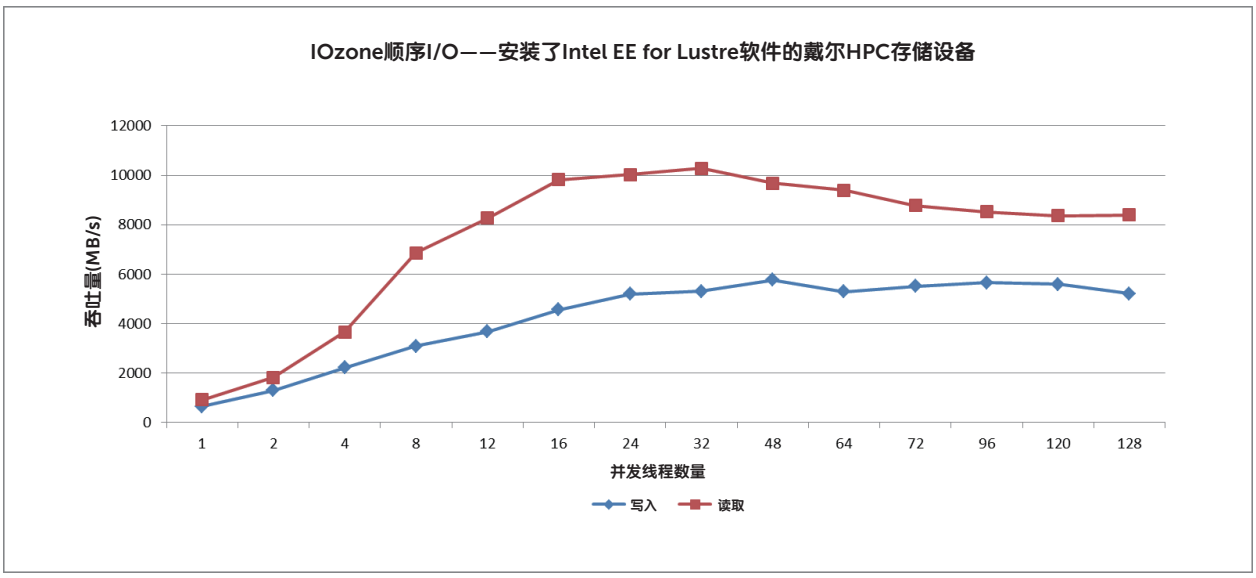


图9显示960TB测试配置的顺序读取/写入性能。在测试系统被使用时，写入性能的峰值接近6GB/秒，而读取性能的峰值接近10GB/秒。单个客户端的读取性能为926MB/秒，写入性能为645MB/秒。当我们将进程的线程数量持续增加到32（对于读取）和48（对于读取）时，读取和写入性能一直增加。其部分原因是，当线程数量增加时（在我们的系统中最多增加到24个OST），所利用的OST的数量也会增加。

为了针对更多的文件也保持较高的吞吐量，增加OST的数量可能会有所帮助。使用Dell PowerVault Modular Disk Storage Manager提供的工具检查存储阵列性能时，执行性能监视器以单独确认由基准工具生成的吞吐量值。

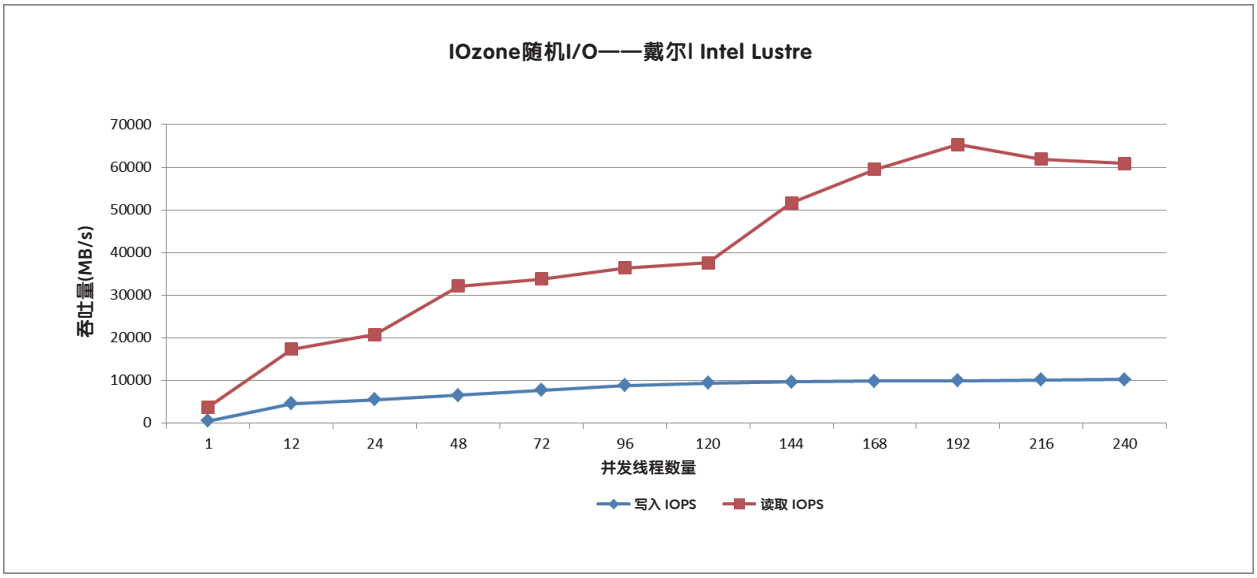
4.2 随机读取和写入

IOzone基准用于收集随机读取和写入指标。为该测试选择的文件大小可保证所有线程的聚合大小保持1TB不变。也就是说，随机读取和写入会在该测试中的所有线程之间平均划分聚合大小(1TB)。IOzone主机文件按照一定的方式排列，以便将工作负载平均分布到多个计算节点之间。存储作为单个卷进行

寻址，该卷的条带数为1，条带大小为4MB。之所以使用4KB请求大小，是因为它与Lustre的4KB文件系统数据块大小一致，而且代表随机工作负载的小数据块访问。性能用每秒I/O操作数(IOPS)度量。

图10显示当线程数量为240时，随机写入达到峰值（稍高于10K IOPS）；当线程数为192时，随机读取达到峰值(65K IOPS)。当线程数从120增加到192时，随机读取的IOPS迅速增加，之后，IOPS在稍微下降后趋于稳定。当写入操作要求所访问的每个OST都有文件锁定时，达到饱和并不意外。读取操作利用Lustre的如下功能：为部分或全部文件授予重叠的读取扩展区锁定。

图10：多对多随机读取和写入



4.3 IOR多对一读取和写入

安装了Intel EE for Lustre软件的戴尔HPC存储设备解决方案是使用IOR基准工具通过读取和写入单个文件进行性能分析的。IOR能够处理MPI通信以执行并行操作，而且支持操纵Lustre条带。IOR允许不同的IO接口处理文件。为了进行测试，我们使用POSIX接口排除其他可用IO接口的高级功能和相关开销。这使我们有机会独立于这些其他增强功能来检查文件系统和硬件性能。

本研究中使用的是IOR基准版本3.0.1。所使用的MPI堆栈是Intel MPI版本5.0 Update 1。

写入测试的配置中包括一个目录集，该集合具有旨在跨24个OST进行条带化（条带大小为4MB）的条带化特征。因此，所有的线程都写入一个跨全部24个OST条带化的文件。在本测试中，Lustre的请求大小设置为1MB，但是，使用4MB的传输大小来与目标文件上使用的特定大小相匹配。

为了降低服务器和客户端内存的高速缓存效应，我们决定使用的文件大小为OSS和客户端内存之和的两倍，其计算公式如下所示，必要时会四舍五入到整数值：

文件大小 = 2 * (2 OSSs*每个OSS 256GiB内存 + 物理客户端数量 * 每个客户端24GiB内存)。

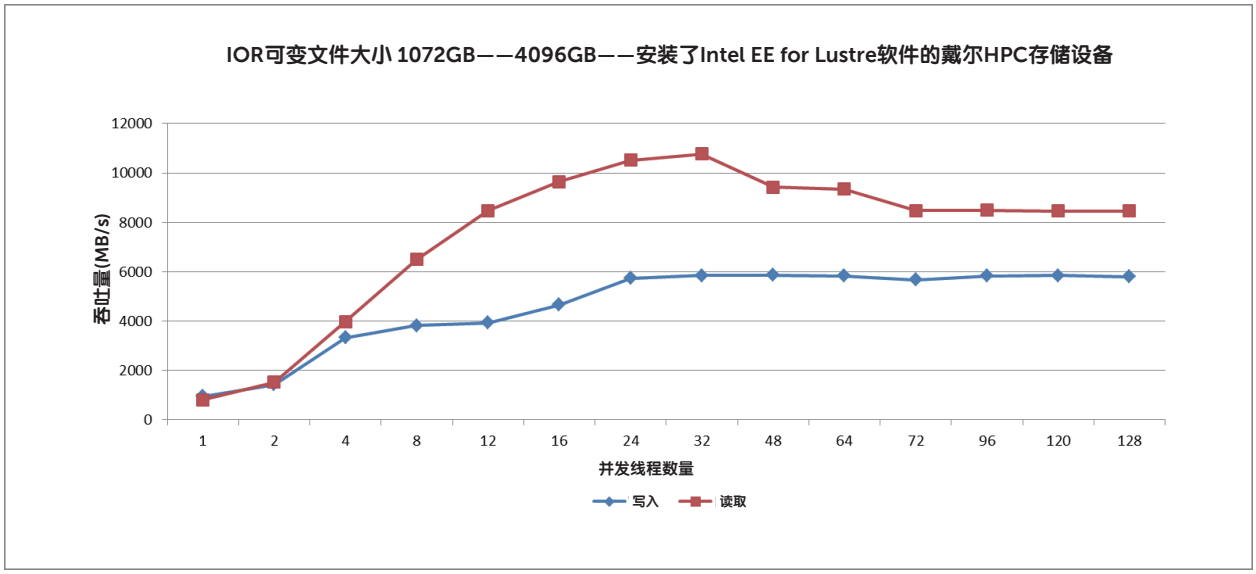
表3显示由每组客户端操纵的数据大小、线程数量和共享文件的总大小。

表3：IOR共享文件大小

线程数量	物理客户端数量	每个线程写入的数据量(GB)	共享文件大小(GB)
1	1	1072	1072
2	2	560	1120
4	4	304	1216
8	8	176	1408
12	12	133	1600
16	16	112	1792
24	24	91	2176
32	32	80	2560
48	48	69	3328
64	64	64	4096
72	64	57	4096
96	64	43	4096
120	64	34	4096
128	64	32	4096
144	64	28	4096
168	64	24	4096
192	64	21	4096
216	64	19	4096
240	64	17	4096
256	64	16	4096

图11显示IOR结果，其中读取操作占优势，当线程数量为32时，读取性能达到峰值(10.7GB/s)，其趋势与顺序多对多性能非常相似。当线程数与OST的比率持续增加时，性能一直增加。当线程数量为24（与测试系统中可用OST的数量相等）时，性能达到平稳。写入操作的峰值为6.2GB/s。单个客户端IOR的读取性能为806MB/秒，写入性能为948MB/秒。

图11：多对一IOR读取/写入



4.4 元数据测试

对于某些将返回属性的文件操作或目录操作，可以使用元数据测试来度量这些操作的完成时间。MDtest是由MPI协调的基准，用来针对文件或目录执行"创建"、"统计"和"删除"操作。本研究使用的是MDtest版本1.9.3。本研究使用的MPI堆栈是Intel MPI版本5.0 Update 1。由MDtest报告的指标是用每秒操作数（OP/秒）表示的完成速率。MDtest可配置为对目录和文件的元数据性能进行比较。在本测试中，我们针对文件操作执行一遍测试，然后针对目录操作执行另一遍操作。

在Lustre文件系统中，在OST中查询对象标识符以分配或查找与元数据操作相关联的扩展区。在大多数元数据操作中，此交互要求间接涉及OST。在使用测试台的实验室试验中，我们发现，对于大多数元数据操作，当OST=1时效率更高，当线程较多时尤其如此。这些试验涉及到对于下列Lustre拓扑，针对多达64个客户端执行测试（此处未显示结果），以便为针对该版本的解决方案执行元数据性能测试选择最高效的配置：

- 1个OST，1MB条带
- 1个OST，4MB条带
- 24个OST，1MB条带
- 24个OST，4MB条带

我们发现，"1个OST，1MB条带"配置最高效，因此，本部分中显示的结果针对的是这样的Lustre拓扑。

而且，在初步元数据测试期间，我们发现每个目录的文件数量对结果造成的影响非常明显，当创建的文件总数保持恒定时也是如此。例如，当使用64个线程进行测试时，每个线程在5个目录中的每个目录中创建3125个文件，或者每个线程在25个目录中的每个目录中创建625个文件，都会导致创建

100万个文件，但是用IOPS度量的性能不同。这是由于在更改目录时，针对OST执行寻道所产生开销不同。为了显示一致的结果，我们将每个目录的文件数固定在3125，并改变每个线程的目录数，使生成的文件总数不小于100万并随着线程数增加而一直增加。表4表示每个测试中使用的值。对于文件操作测试和目录操作测试，我们都使用八次迭代执行测试，每次迭代都对所记录的结果取平均值。

表4：IOR共享文件大小

线程数量（多个）	每个目录的文件数量	每个线程的目录数量	文件总数
1	3125	320	1000000
2	3125	160	1000000
4	3125	80	1000000
8	3125	40	1000000
12	3125	27	1012500
16	3125	20	1000000
24	3125	14	1050000
32	3125	10	1000000
48	3125	7	1050000
64	3125	5	1000000
72	3125	5	1125000
96	3125	4	1200000
120	3125	3	1125000
128	3125	3	1200000
144	3125	3	1350000
168	3125	2	1050000
192	3125	2	1200000
216	3125	2	1350000
240	3125	2	1500000

图12：文件元数据操作

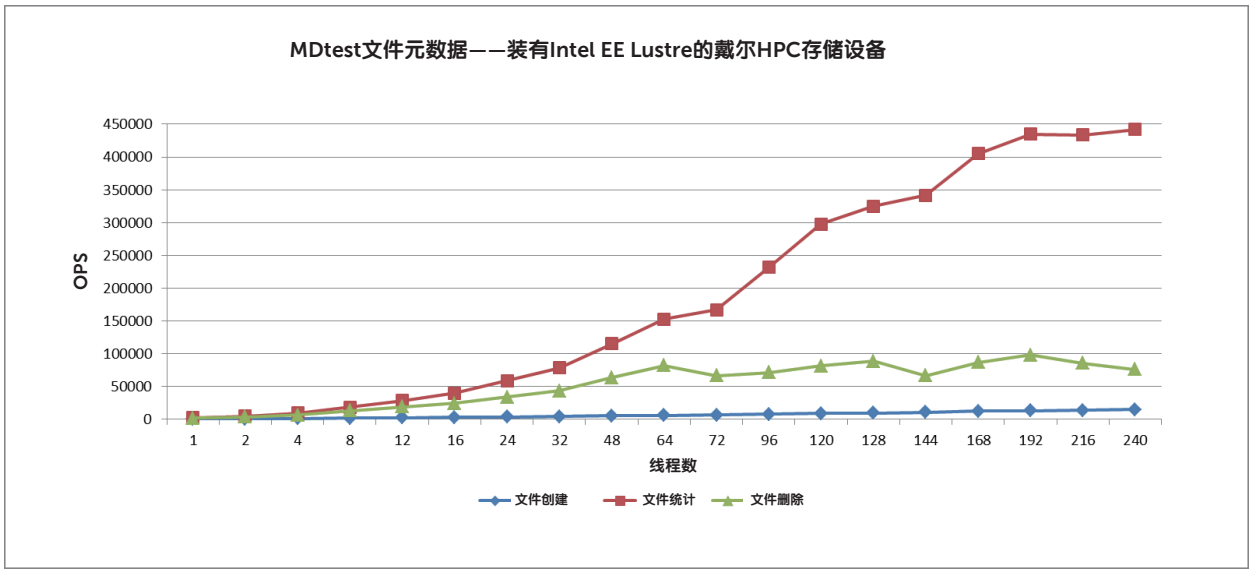


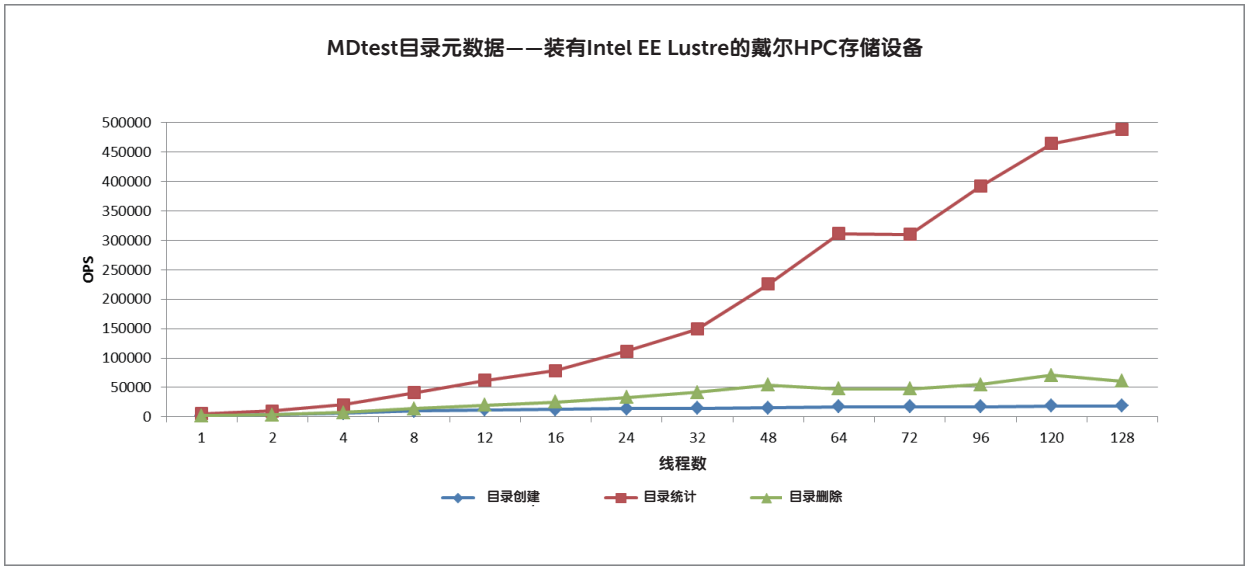
图12说明使用MDtest的文件元数据结果。从该图可以看出，文件创建元数据操作的性能最初为504 OPS（此时线程数为1），之后，当并发线程数为240时，性能增加到大约15K OPS。其原因可能是：条带数为24时性能会显著下降，这会导致MDT和OST均需要进行Lustre锁定。当线程数为240时，我们有2个目录(-b 2)，共创建了220万个文件。

在所观察的这三个元数据操作中，文件统计元数据操作的开销最低。当线程数为1时，此测试的性能高于2K OPS；当线程数为240时，性能会扩展到超过400K OPS。
性能增加可能是由于Lustre版本2.5对元数据操作进行了改进。另请注意的是，在MDT卷中使用的是15K RPM驱动器。

文件删除还受到对OST访问的限制，这与创建操作相似。但是，当总线程数增加时，删除操作优于创建操作：它最初的性能高于1.8K OPS（此时线程数为1），之后，当并发线程数为192时，性能增加到大约100K OPS。

图13说明使用MDtest的目录元数据结果。从该图可以看出，在大多数情况下，目录创建元数据操作的开销也是最高，它最初的性能为1.8K OPS（此时线程数为1），当线程数为128时，性能达到最大值（大约19K）。目录操作也受所使用的顶层目录数(-b)的影响，但比文件操作受影响的程度低。删除操作的开销与目录创建操作的开销几乎一样高，删除操作最初的性能为大约1.9K OPS（此时线程数为1），当线程数为120时，性能达到最大值（超过70K OPS）。在所观察的这三个操作中，目录统计操作的开销也是最低。当线程数为1时，此测试的性能高于5K OPS；当线程数为128时，性能会增加到超过450K。

图13：目录元数据操作



5. 总结

对于可扩展的高性能群集文件系统解决方案有一个众所周知的要求。安装了Intel EE for Lustre软件的戴尔HPC存储设备使用易于管理、完全受支持且设计完善的解决方案满足此需求。此解决方案包括Dell PowerEdge™第十三代服务器平台、PowerVault™存储产品和Lustre®技术所带来的附加好处，对于并行文件系统来说，它是领先的开源解决方案。Intel Manager for Lustre (IML)将对于Lustre文件系统和解决方案组件的管理统一到单个控制和监控面板中，以便于使用。

打包组件能够满足高性能计算环境的需求，它对原始存储（每个对象存储服务器对960TB）进行扩展，具有高达10GB/s的读取吞吐量和高达6GB/s的写入吞吐量。安装了Intel EE for Lustre软件的戴尔HPC存储设备还能够像扩展容量那样方便地扩展吞吐量。

性能研究发现，对于多对多和多对一文件类型访问来说，读取和写入吞吐量都很高。MDtest结果表明，元数据文件操作的容量有所提升。使用PCI-e 3.0界面，IB FDR HCA在高带宽应用程序中表现出众。

通过持续使用广泛可用且符合行业标准的基准工具（如IOzone、IOR和MDtest），将使当前性能和预期性能与所概述的性能轻松匹配。其中每个工具所报告的概况都提供足够的信息，通过这些信息可以使安装了Intel EE for Lustre软件的戴尔HPC存储设备的配置符合许多应用程序或应用程序组的要求。

安装了Intel EE for Lustre软件的戴尔HPC存储设备解决方案可提供基于并行文件系统的横向扩展存储的所有好处，从而满足您的高性能计算需求。

附录A：基准命令参考

本部分描述用来对安装了Intel EE for Lustre软件的戴尔HPC存储设备解决方案进行基准测试的命令。

IOzone

IOzone顺序写入——

```
iozone -i 0 -c -e -w -r 1024K -I -s $Size -t $Thread -+n -+m /root/list.$Thread
```

IOzone顺序读取——

```
iozone -i 1 -c -e -w -r 1024K -I -s $Size -t $Thread -+n -+m /root/list.$Thread
```

IOzone IOPS随机读取/写入——

```
iozone -i 2 -w -c -O -I -r 4K -s $Size -t $Thread -+n -+m /root/list.$Thread
```

IOzone命令行	说明
–i 0	写入测试
–i 1	读取测试
–i 2	随机IOPS测试
–+n	不重新测试
–c	在计时计算中包括关闭
–e	在计时计算中包括刷新
–r	记录大小
–s	文件大小
–+m	当处于群集模式时，要在其上运行IOzone的客户端的位置
–I	使用O_Direct
–w	不取消链接（删除）临时文件
–+n	不选择重新测试
–O	在OPS中返回结果

O_Direct命令行参数（"-I"）允许我们在运行IOzone线程的计算节点上绕过高速缓存。

IOR

IOR写入——

```
mpirun -np $Threads -rr --machinefile /root/list.$Threads /cm/share/IOR -a POSIX -v -i $rep -d 3
-e -k -o /mnt/boulder/perf/mytestfile -w -s 1 -t 4m -b $SizePerThread
```

IOR读取——

```
mpirun -np $Threads -rr --machinefile /root/list.$Threads /cm/share/IOR -a POSIX -v -i $rep -d 3
-e -k -o /mnt/boulder/perf/mytestfile -r -s 1 -t 4m -b $SizePerThread
```

IOR 命令行参数	说明
-a S	api——用于I/O的API [POSIX|MPIIO|HDF5|NCMPI]
-v	verbose——输出信息（重复标志增加级别）
-I N	repetitions——测试的重复次数
-d N	interTestDelay——相邻重复测试之间的延迟时间（秒）
-e	fsync——针对POSIX写入关闭执行fsync
-k	keepFile——在程序退出时不删除测试文件
-o S	testFile——测试的完整名称
-w	writeFile——写入文件
-r	readFile——读取现有文件
-s N	segmentCount——段数
-t N	transferSize——用字节数表示的传输大小（例如：8、4k、2m、1g）
-b N	blockSize——每个任务写入的连续字节数（例如：8、4k、2m、1g）

MDtest——元数据

文件操作——

```
mpirun -np $Threads -rr --hostfile /share/mdt_clients/mdtlist.$Threads /share/mdtest/mdtest.intel
-v -d /mnt/lustre/perf_test24-1M -i $Reps -b $Dirs -z 1 -L -I $Files -y -u -t -F
```

目录操作——

```
mpirun -np $Threads -rr --hostfile /share/mdt_clients/mdtlist.$Threads /share/mdtest/mdtest.intel
-v -d /mnt/lustre/perf_test24-1M -i $Reps -b $Dirs -z 1 -L -I $Files -y -u -t -D
```

IOR 命令行参数	IOR 命令行参数
-d	将在其中运行测试的目录
-v	详细程度（选项的每个实例按一递增）
-i	测试将运行的迭代次数
-b	分层目录结构的分支因素
-z	分层目录结构的深度
-L	仅限位于树叶层的文件
-I	树中每个目录中的项数
-y	写入之后同步文件
-u	每个任务的独特工作目录
-t	与时间有关的工作目录开销
-F	仅针对文件执行测试（不针对目录测试）
-D	仅针对目录执行测试（不针对文件测试）

参考资料

安装了Intel EE for Lustre软件的戴尔HPC存储设备解决方案简介
http://salesedge.dell.com/doc?id=0901bc82808e334f&ll=d&pm=160376162

安装了Intel EE for Lustre软件的戴尔HPC存储设备配置指南
* 如需本文档，请与您的戴尔销售代表联系

Dell PowerVault MD3420
http://www.dell.com/support/home/us/en/04/product-support/product/powervault-md3420/research

Dell PowerVault MD3460和MD3060e
http://www.dell.com/support/home/us/en/04/product-support/product/powervault-md3460/research

Lustre主页
http://www.whamcloud.com/lustre/
http://wiki.lustre.org/index.php/Main_Page

Dell HPC解决方案主页
http://www.dell.com/hpc

Dell HPC Wiki
http://www.HPCatDell.com

英特尔主页
http://www.intel.com

安装了Intel Enterprise Edition for Lustre软件的戴尔HPC存储设备

Intel HPDD Wiki
https://wiki.hpdd.intel.com/display/PUB/HPDD+Wiki+Front+Page

Mellanox Technologies主页
http://www.mellanox.com

LSI 12Gb/s SAS HBA
http://www.lsi.com/downloads/Public/Host%20Bus%20Adapters/LSI_PB_SAS9300_HBA_Family.pdf