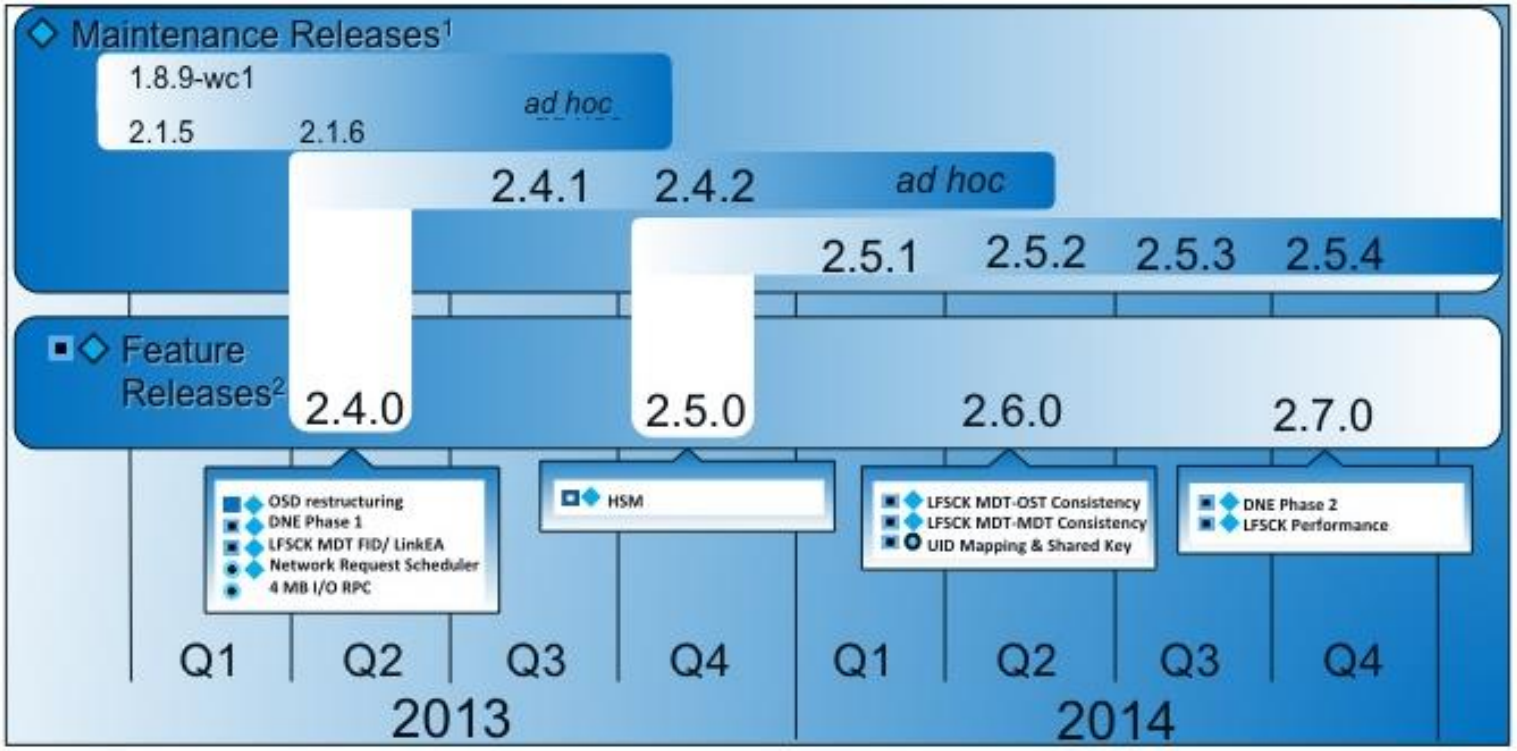
Lustre*发行路线图及部分在开发项目简介

范勇(fan.yong@intel.com)
英特尔高性能数据事业部

- 概览
  - Lustre*发行路线图
  - 集群式元数据服务器(DNE)
  - 文件系统在线校验(LFSCK)
  - 文件多副本(Replication)
  - 层次化存储管理(HSM)
  - 小文件存储/访问优化
  - 英特尔企业版Lustre*软件(IEEL)

Lustre* Software Roadmap



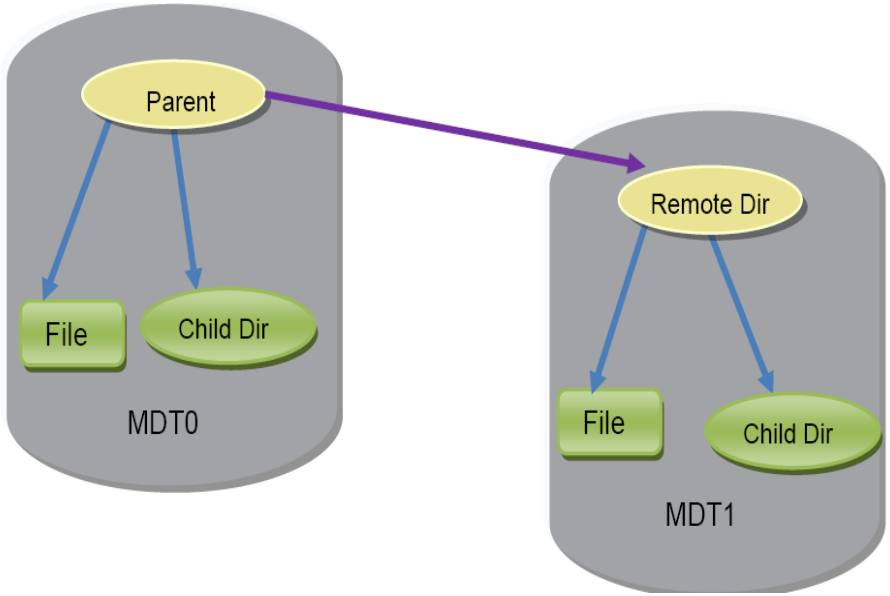
Sponsors for Development and Releases: ORNL, OpenSFS, LLNL, Intel, CEA, Xyratex, Indiana University

[1] 维护发行版主要为解决软件缺陷及稳定性问题，当前版本的更新周期大致为3个月，旧版本的更新视情况而定。

[2] 功能发行版主要为支持新的功能，其发行周期大致为6个月。维护发行版从其功能发行起周期性发行约18个月。

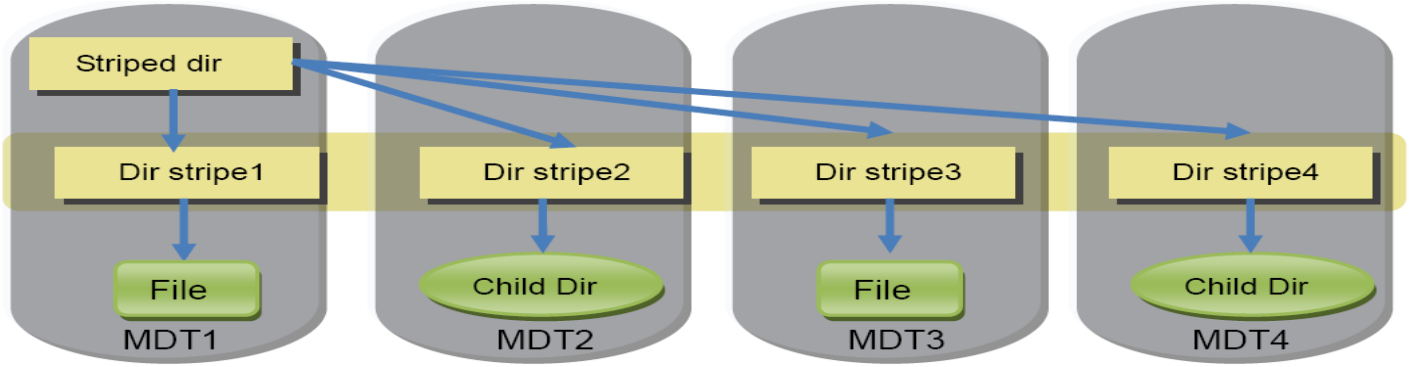
集群式元数据服务器

- OpenSFS基金支持，英特尔研发
  - DNE - Distributed NamespacE，不同于原有的CMD，DNE采用全新的架构设计与实现。
- DNE 1：远程目录
  - 创建时静态指定远程目录位置：lfs mkdir -i <MDT_index>
  - 普通文件创建在与父目录相同的元数据服务器上
  - 跨多个元数据服务器的元数据修改操作采用同步事务机制
  - 暂不支持跨多个元数据服务器的rename/link操作



集群式元数据服务器(cont’d)

- DNE 2：目录条带化



  - 支持单个目录下的目录项（子目录及普通文件）跨越多个元数据服务器，用以提高大目录（共享）操作的性能。
  - 允许在多个元数据服务器之间迁移元数据对象（而无需数据复制），用以支持元数据服务器负载均衡。
  - 跨多个元数据服务器的元数据修改操作采用异步事务机制。
  - 支持跨多个元数据服务器的rename/link操作。

文件系统在线校验(1)

- OpenSFS基金支持，英特尔研发
  - 不同于已有的离线模式的Lustre*文件系统校验工具，新版LFSCK采用在线模式，在不停Lustre服务情况下扫描修复超大规模系统。
    - 以50K / 秒的速度扫描修复包含4G个文件的系统耗时约1天！
    - 文件数量越多，扫描频率越高，则相应的离线开销越大。
  - Lustre新旧版本数据兼容性问题

	Lustre-2.x	Lustre-1.8
全局文件标识	FID，通过索引文件映射到inode	本地文件系统inode索引号
LMA扩展属性	支持 (自身FID)	不支持
目录遍历加速信息	在目录项中存储FID	在目录项中存储inode索引号
linkEA扩展属性	支持 (文件名+父目录FID)	不支持

文件系统在线校验(2)

- LFSCK 1：Inode Iterator & OI Scrub
  - 实现基于服务器后端本地文件系统存储对象(inode)的线性迭代扫描工具，用于高效扫描该服务器节点上的全部存储对象，该迭代器被其他各LFSCK组件所共享。
  - 2.x版本使用本地索引文件(OI)来处理FID到后端本地文件系统inode的映射，该索引文件在经过文件系统操作(tar/untar)进行备份／还原处理后失效，需要借助OI Scrub达到还原后系统的可用性。
- LFSCK 1.5：FID-in-dirent & linkEA
  - 2.x版本所支持的FID-in-dirent信息无法通过文件系统操作进行备份，需借助LFSCK在还原后系统上重建该属性信息。
  - linkEA校验：文件系统目录项缺失；或目标对象linkEA缺失；或文件系统目录项与目标对象linkEA回指信息不匹配。

文件系统在线校验(3)

- LFSCK 2：文件条带信息一致性校验
  - 元数据对象存储的条带信息指向每一个数据对象，每一个数据对象储存回指对应元数据对象的信息，两者需一致：

悬空引用	元数据对象条带信息所指向的数据对象不存在。
多重引用	多个元数据对象条带信息指向同一个数据对象。
交错引用	元数据对象条带信息所指向的数据对象回指不存在的元数据对象。
孤儿引用	数据对象回指不存在的元数据对象，而且没有其他元数据对象条带信息指向该数据对象。

  - 元数据对象存储文件属主信息用于权限验证及quota，数据对象存储文件属主信息用于quota，两者需一致。

文件系统在线校验(4)

- LFSCK 3：多元数据服务器间一致性校验
  - 元数据对象的目录信息与该对象本身存储在不同的元数据服务器上，目录项信息应与元数据对象linkEA的回指信息相一致：

悬空引用	本地目录项所指向的远程元数据对象不存在。
孤儿引用	本地元数据对象linkEA回指不存在的远程目录项，而且没有其他目录项指向该元数据对象。
回指 / 引用缺失	本地目录项所指向的远程元数据对象不包含对应的linkEA回指项；或者反之。

  - 文件硬连接计数应与对应的目录项和linkEA回指项相一致。
- LFSCK 4：性能优化
  - 后台LFSCK速度可控，降低对前台正常服务的影响。

文件多副本

- OpenSFS基金支持，英特尔研发
  - Lustre*数据RAID0模式条带化隐含着硬件故障情况下的部分数据暂时 / 永久不可用的风险，该风险在廉价设备上更高。
  - 热点访问可能导致的负载失衡 / 性能抖动等问题。
  - 支持系统扩容，却不支持系统缩容或设备更换。
- Replication 1：只读模式副本
  - 通过用户层命令 / 工具拷贝目标文件，然后将多个副本的条带信息合并为一个对象属性，删除冗余元数据。
  - 每个副本的条带方式彼此独立，设备选择依策略而定，RAID0+1。
  - 支持副本的读操作 / 条带信息合并 / 故障处理，对写操作影响小。

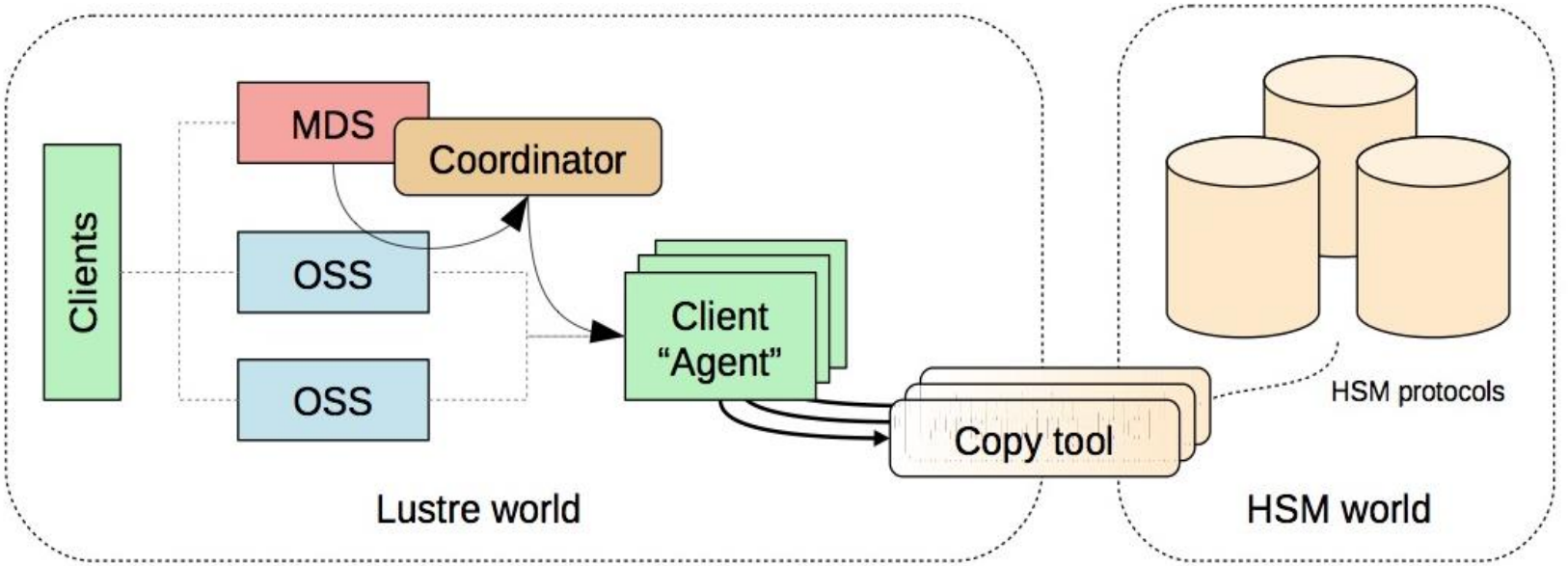
文件多副本(cont'd)

- Replication 2：同步生成文件副本
  - 文件多个副本的条带信息在文件创建时设定。
  - 客户端将每个写操作同步到各副本相应条带所对应的数据服务器，保证每次写操作返回后各副本之间的一致性。
  - 支持副本的修改操作，同步冗余模式，写开销较大。
- Replication 3：异步生成文件副本（可选）
  - 文件多个副本的条带信息在文件创建时设定或从其父目录继承。
  - 客户端将数据写到各副本相应条带所对应的本地缓存，再通过异步机制将缓存数据写回各自相应的数据服务器。
  - 写性能较前者为好，但副本间的一致性及故障恢复机制较为复杂。

层次化存储管理

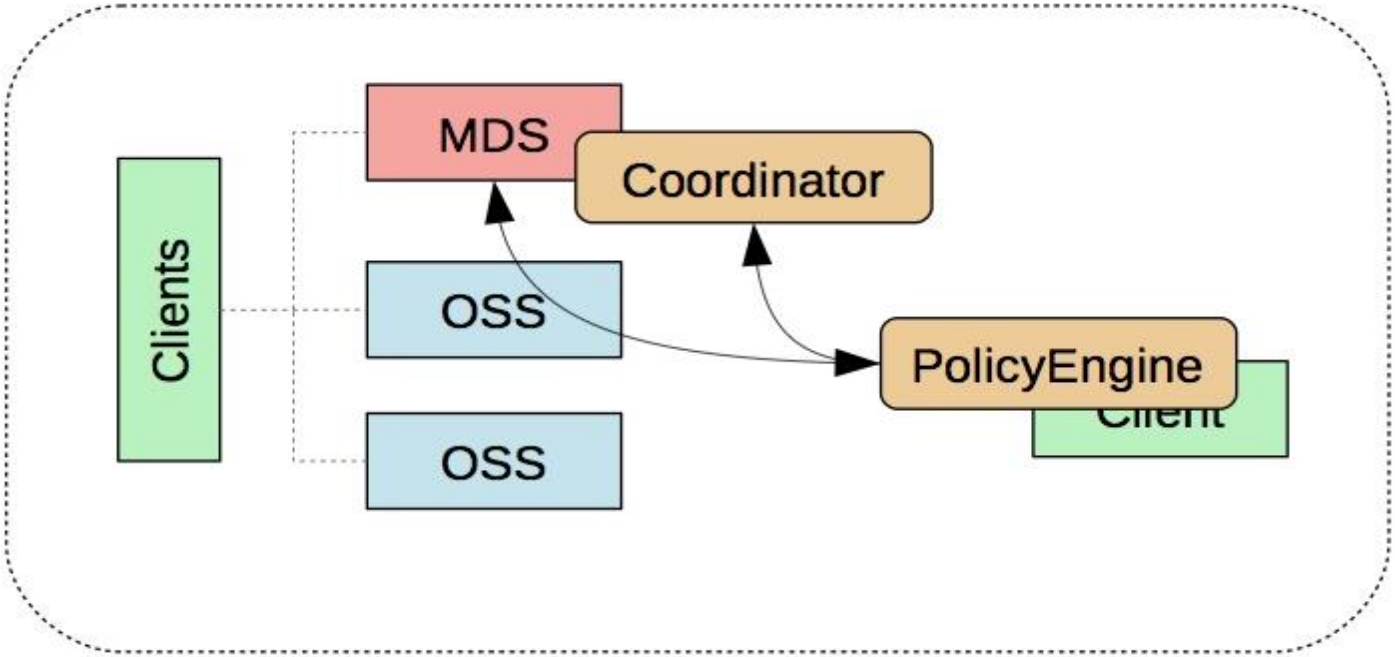
- 法国原子能委员会(CEA)主导，英特尔协助
  - http://www.eofs.eu/fileadmin/lad2013/slides/10_Aurelien_Degremont_lustre_hsm_lad13.pdf
- 功能
  - 归档：将数据从Lustre*系统迁移到HSM系统。
  - 释放：如果需要，释放Lustre系统的存储空间。
  - 还原：当Lustre不命中时，将数据从HSM系统迁移回Lustre系统。
  - 策略管理：数据迁移，数据清除等。
  - 灾难恢复：利用历史归档还原系统。

HSM协调器及代理



- 与HSM系统之间转移数据。
- HSM协调器负责收集归档请求，并将其分发至相应的HSM代理。

HSM策略引擎



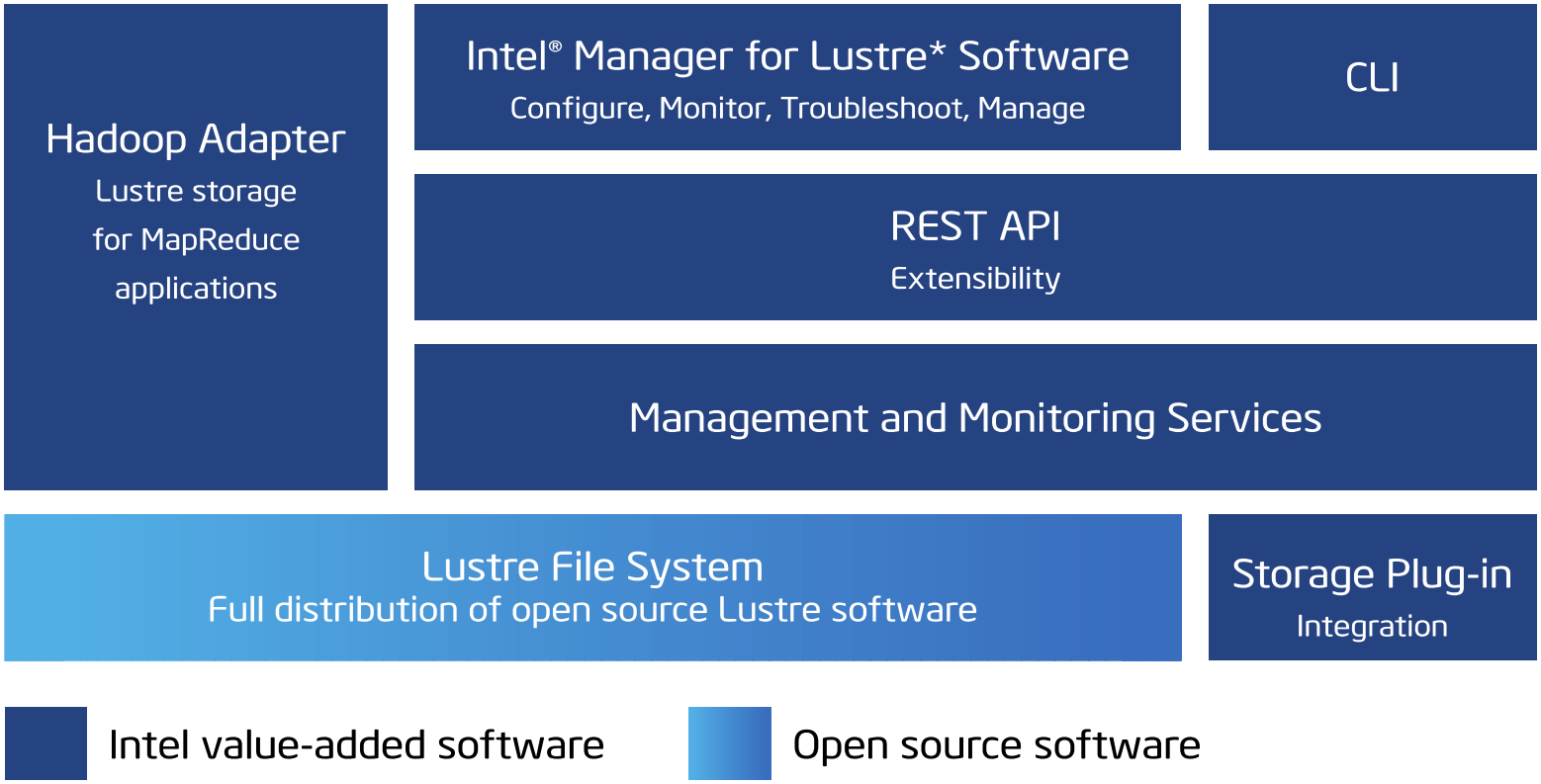
– 策略引擎是运行于用户空间的HSM工具，它通过与元数据服务器通讯来监视文件系统的状态，并依策略向HSM协调器发起相应的归档 / 释放 / 销毁等请求。

小文件存储/访问优化

- OpenSFS基金支持，英特尔研发
  - Lustre*现有的数据与元数据分离模式更适合大文件读写，而非小文件，后者瓶颈主要在于RPC自身开销。
  - 单纯提高元数据服务器的可扩展性与单个元数据服务器的处理效率都难以有效改善小文件的读写性能。
- 构建统一的数据与元数据对象
  - 一致的接口，结构更简洁，代码更干净。
  - 允许客户端直接从 / 向元数据服务器读 / 写数据。
- 将小文件数据存储在高速元数据服务器上
  - 减少小文件操作的RPC（锁，文件尺寸，RAID5/6读-修改-写）。
  - 随着文件变大，自动将数据从元数据服务器迁移到数据服务器。

英特尔企业版Lustre*软件

- IEEL - Intel Enterprise Edition for Lustre* software



IEEL与传统Lustre*方案对比

Product	Open source	Intel® Enterprise Edition for Lustre* software
Open source base	Yes	Yes (2.3 servers, 2.4 clients)
Supported operating systems	Storage servers: Red Hat and CentOS Clients: Red Hat, CentOS, SUSE	
Processor support	Intel® Xeon®, x86 and IBM PowerPC	
Additional validation testing		✔
Intel® Xeon Phi support		Future
Selected fixes		✔
Selected features		✔
Includes Intel® Manager for Lustre		✔
Storage plug-in for array integration		✔
REST API		✔
Advanced Monitoring		✔
Simplified Management		✔
Level 3 Technical Support	Direct to end users	Via OEM/SI/resellers
Integration with Intel Distribution for Apache Hadoop		✔
Enterprise Service Level Agreement support	Optional	Standard
Intel training courses	✔	✔
Intel Professional Services	✔	✔

* Some names and brands may be claimed as the property of others.

Q&A

谢谢!

范勇(fan.yong@intel.com)