COFS
China Open File System

开源Lustre文件系统技术路线及前景展望

李希

CHINA LUG
2021

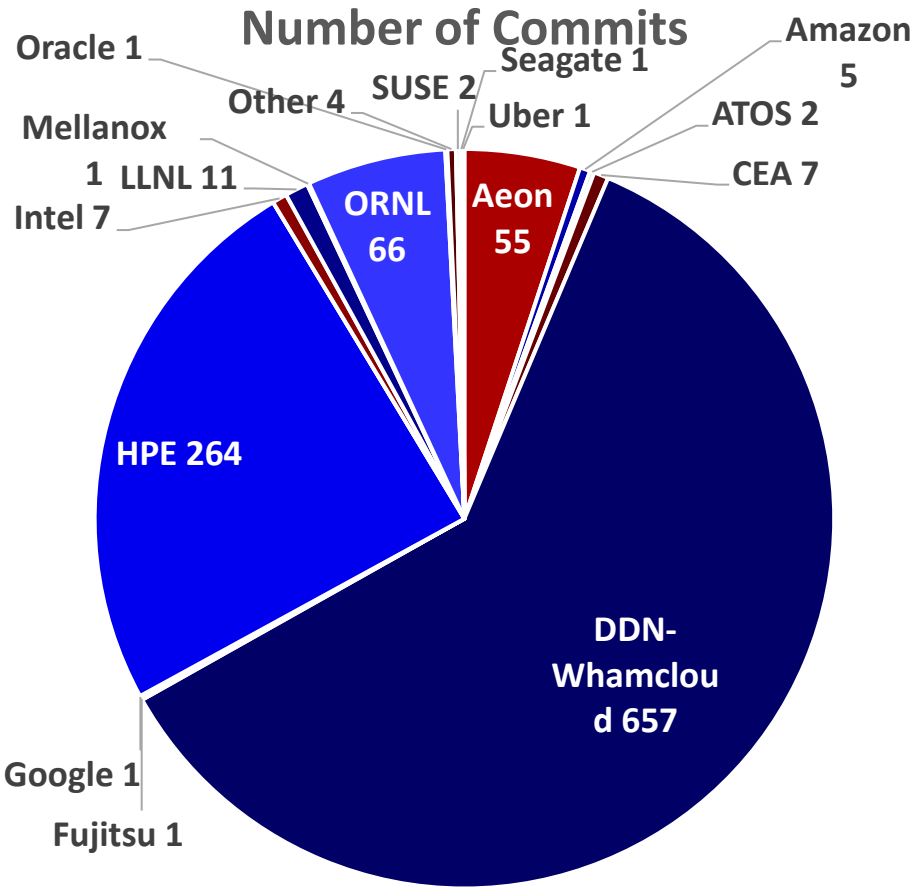
广泛采用的长线支持版：2.12.x

Whamcloud

- Lustre 2.12.7已发布
  - 服务器发行版支持： RHEL 7.9
  - 客户端发行版支持： RHEL 7.9、RHEL 8.4、SLES12 SP5、 Ubuntu 18.04
  - MOFED 4.9 （MOFED 5.x也可运行）
  - ZFS 0.7.13
  - E2fsprogs推荐版本： v1.46.2.wc3
  - 交互版本： 最新的2.10.X和2.11.X版本
- Lustre 2.12.8将在近期发布
  - 增加对RHEL 8.5客户端的支持

新的长线支持版：2.15

- Lustre 2.15将成为长线支持版
- 发布时间：2021年年底
- 服务器发行版支持：RHEL 8.5
- 交互版本：最新的2.12.X和Latest 2.14.X版本
- 客户端发行版支持：
  - RHEL 8.5
  - SLES15 SP5
  - Ubuntu 20.04



Contribution of Lustre 2.14

Lustre技术路线图

2.14 (已发布)
- **客户端数据加密**
- **OST存储池的限额**
- **DNE自动重分布**

2.15 (正在推进)
- **客户端目录加密**
- **GPU Direct支持**
- **网络选择策略**

2.16 (计划中)
- **FLR纠删码**
- **FLR同步镜像**
- **客户端元数据回写缓存**

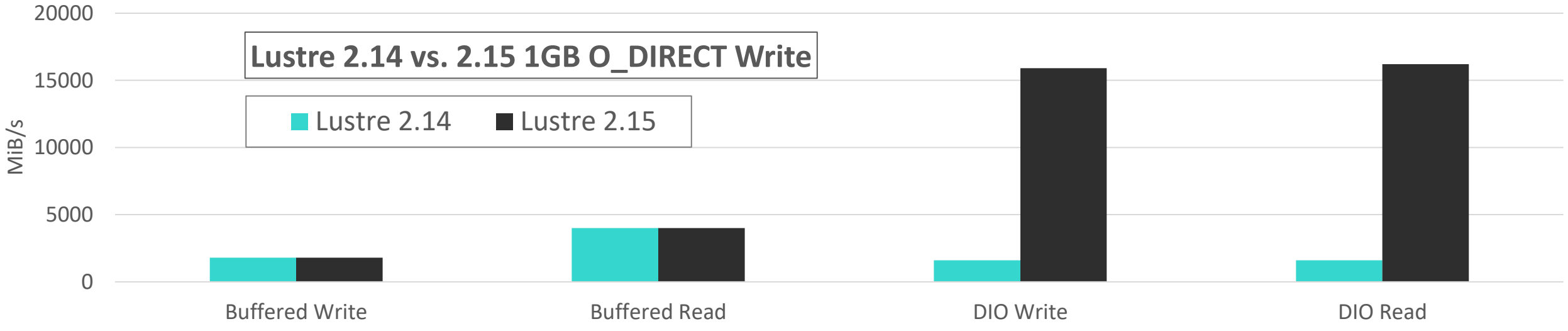
新版本特性：更强的单客户端/单线程性能 (2.15+)

Whamcloud

- **GPU Direct RDMA** - directly into GPU, bypass CPU (LU-14798, WC, NVIDIA)
  - Significant speedup for IO (A100 2x200Gb IB 25GB/s->**36GB/s** write, 23GB/s->**39GB/s** read @ 1MB)
  - Improve 110GB/s->**174GB/s** with 8x200Gb IB storage links (non-standard A100 config)
- **Parallel large DIO** optimization (LU-13798, LU-13799, HPE, WC)

2.15
  - Improve single-thread `read()`/`write()` (1.5GB/s->**15.8GB/s!**)

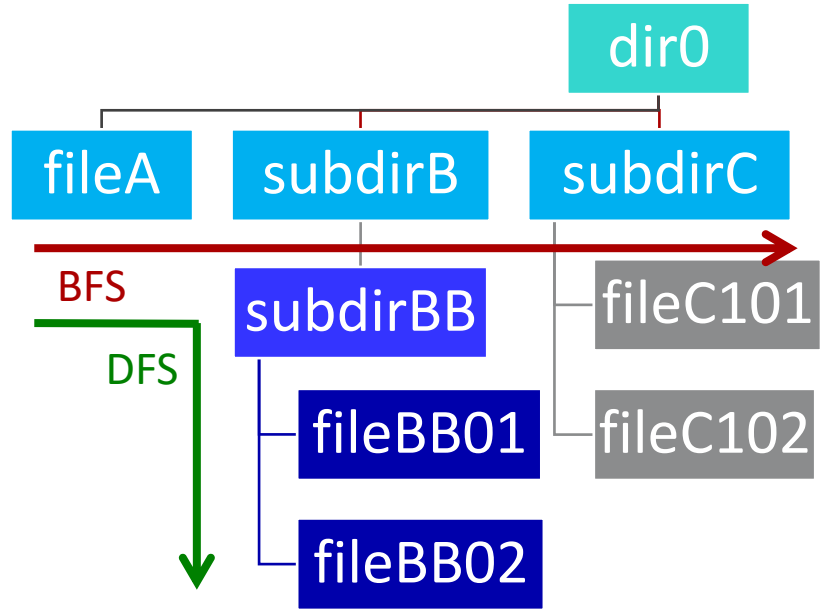
2.16
- Continued client performance optimizations



新版本特性：聚合的元数据操作RPC (2.16)

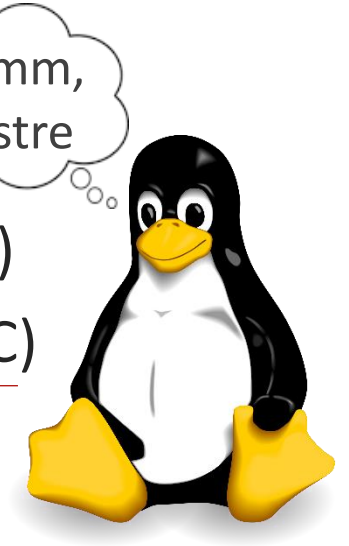
Whamcloud

- **Batched RPCs** for multi-update operations (LU-13045)
  - Allow multiple getattrs/updates packed into a single MDS RPC
  - More efficient network and server-side request handling
- **Batched statahead** for `ls -l`, `find`, etc. (LU-14139)
  - Aggregate getattr RPCs for existing statahead mechanism
- **Cross-Directory statahead** pattern matching (LU-14380)
  - Existing statahead only detects `readdir()`-ordered `stat()`
  - Detect pattern for alphanumeric ordered traversal + `stat()`
  - Detect breadth-first (BFS) depth-first (DFS) directory traversal
  - Direct statahead to next file/subdirectory based on pattern



新版本特性：各种易用性优化 (ORNL, SuSE, WC)

Whamcloud



- `fallocate()` for regular files on ldiskfs (LU-3606, AEON, WC)
- `SEEK_HOLE/SEEK_DATA` to efficiently handle sparse files (LU-10810, WC)
- `statfs()` on directory with projid returns project quota limits (LU-9555, WC)

2.14
- `statx()` allows fetching specific inode attributes, lazy file size (LU-10934, WC)

2.15
- Automatic open lock caching on client (LU-10948, WC, ORNL)
- Handle large ACLs up to 8k entries (LU-14430, WC)
- `fallocate(FALLOCATE_FL_PUNCH_HOLE)` to free space (LU-14160, AEON)
- `fallocate()` for DoM files (LU-14382, WC)
- `llstat/llobdstat` usability improvements (LU-13705, WC)

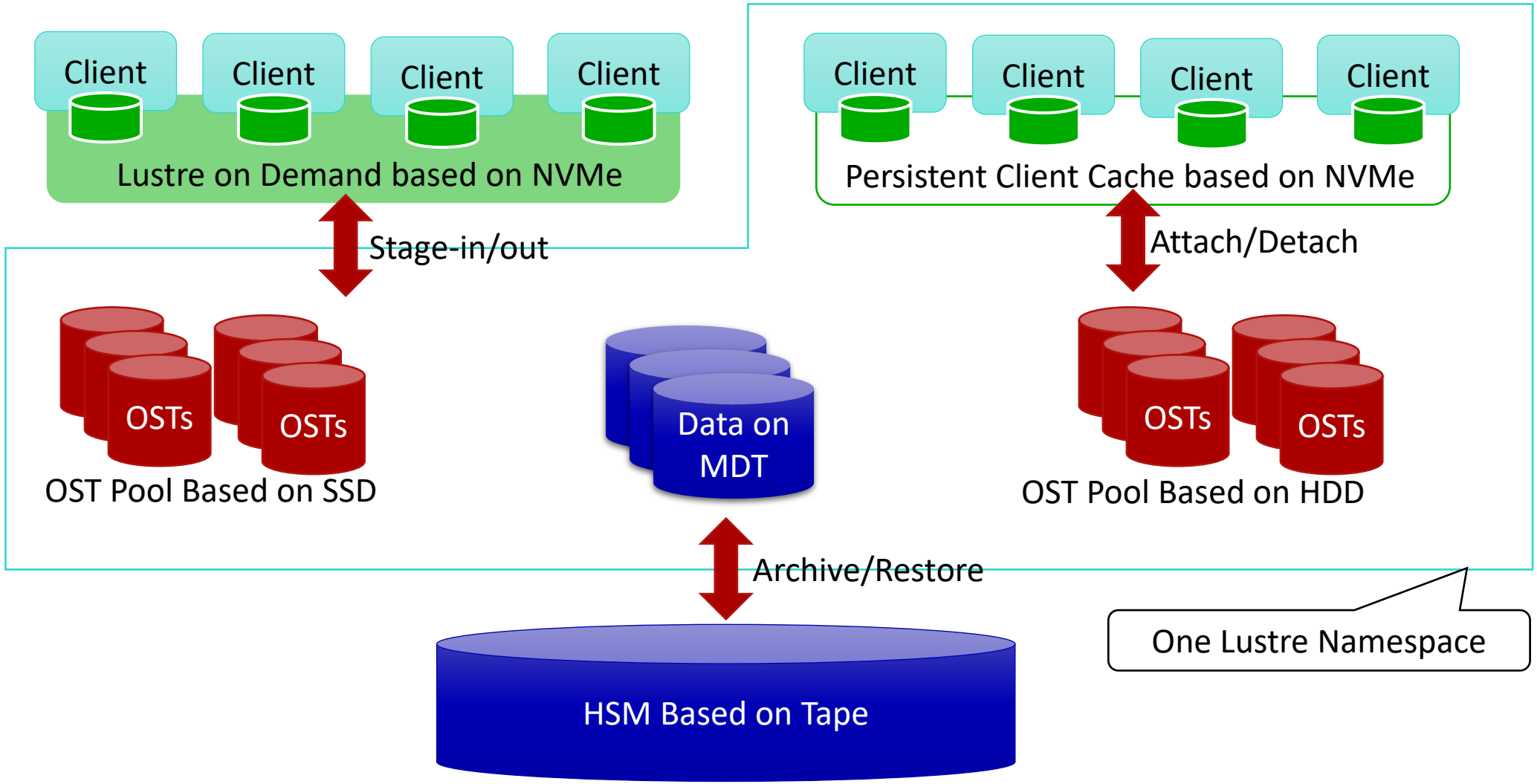
2.16
- Ongoing upstream kernel cleanups (ORNL, SUSE)
- `o2iblnd` cleanups for in-kernel OFED (LU-8874)

前景： Lustre基于各类存储介质的多层级高速缓存

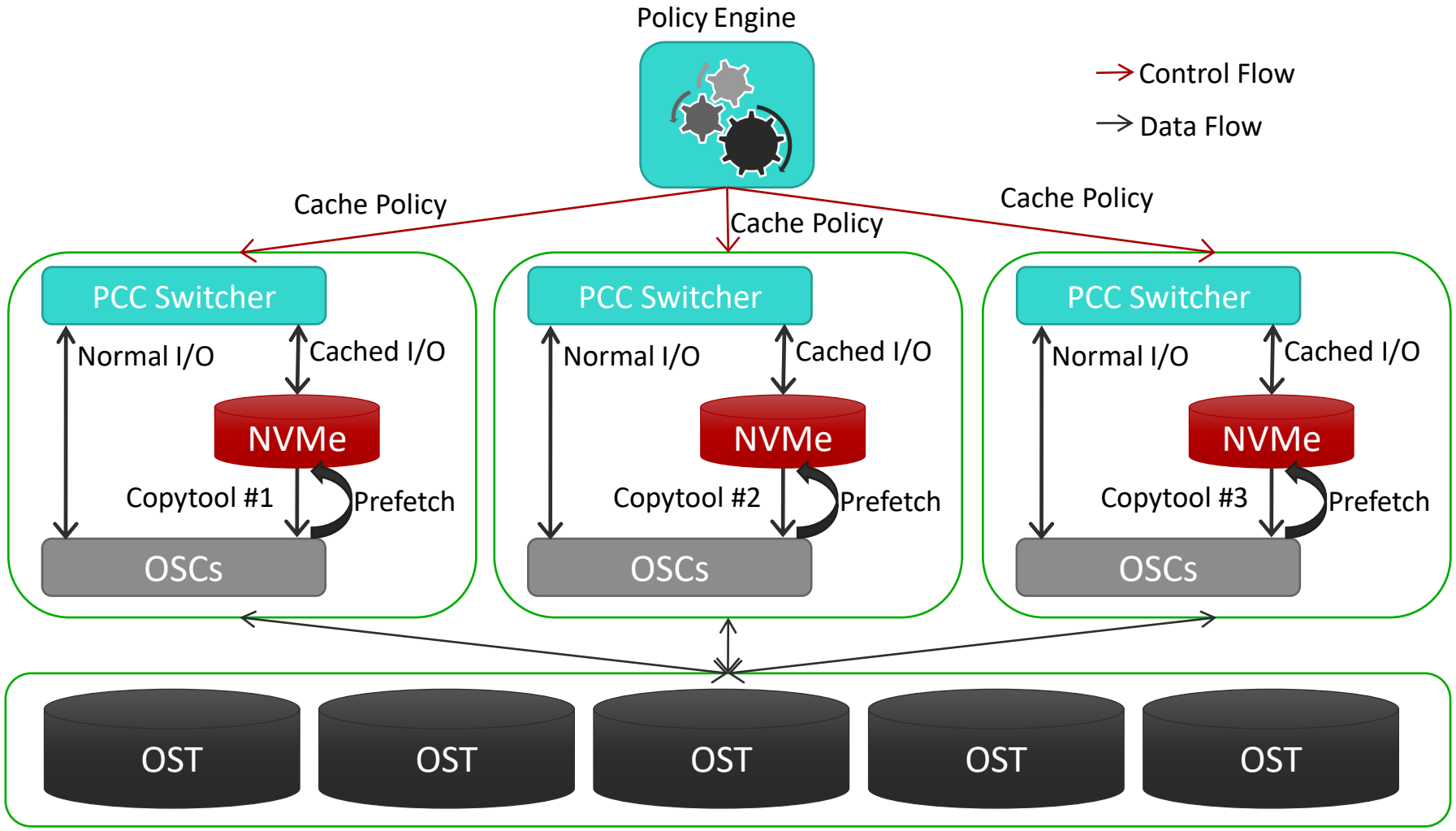
Whamcloud

- Lustre客户端持久缓存（Persistent Client Cache，PCC）
  - 客户端本地存储用作只读高速缓存或独占文件的高速缓存
- Lustre on Demand（LoD）作为活跃作业的数据集缓存
  - 活跃作业可利用更快的计算节点网络和存储
- Data on MDT（DoM）用以提速数据访问
  - MDT存储介质通常为SSD/NVMe，可减少RPC，提高性能
- OST存储池作为高速缓存池
  - 热数据存储在存储介质更快的OST池中
- 客户端元数据回写缓存（Write Back Cache，WBC）
  - 元数据所有修改都可缓存在客户端，可利用非易失内存

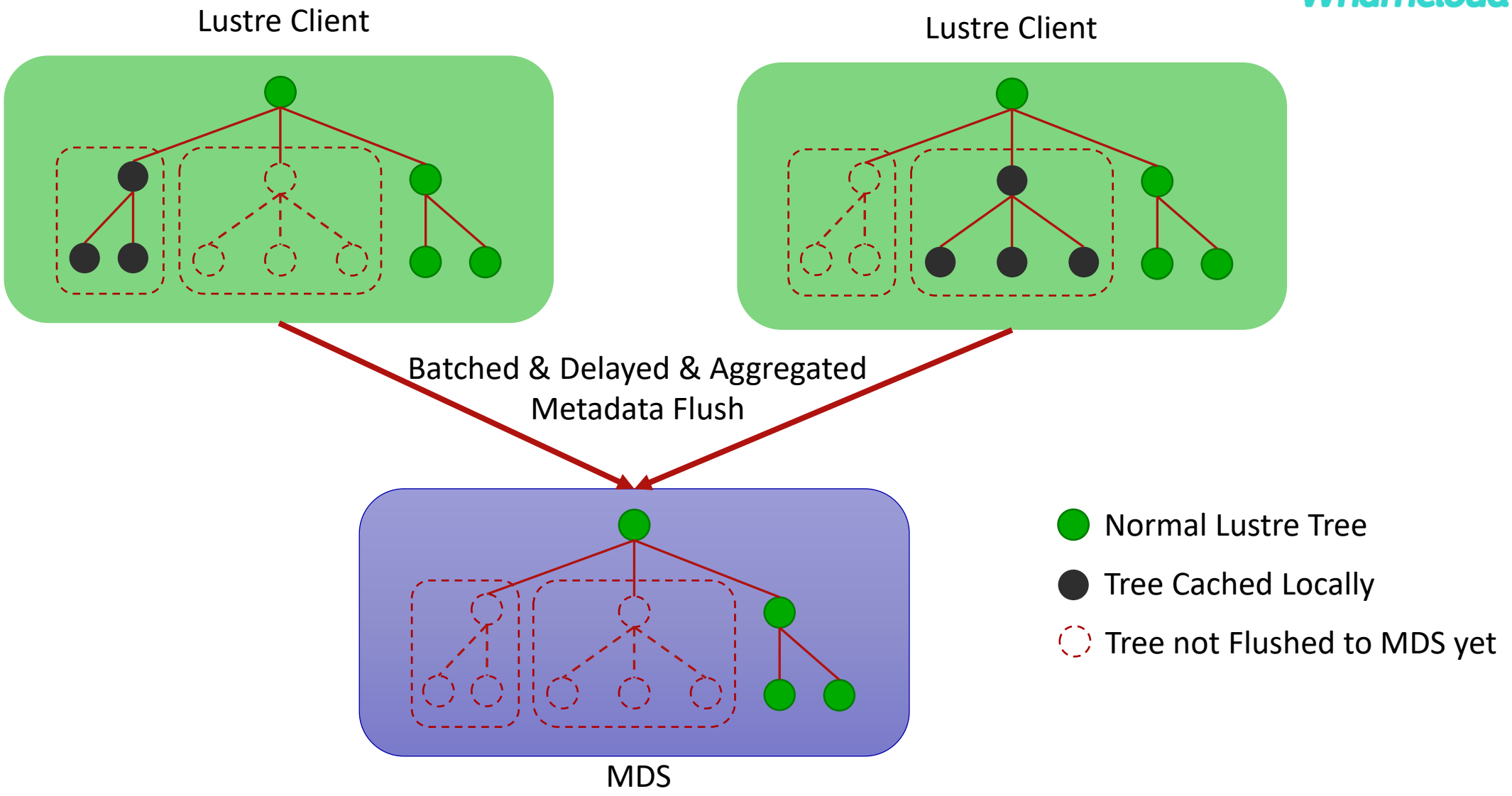
Lustre综合利用各种存储介质



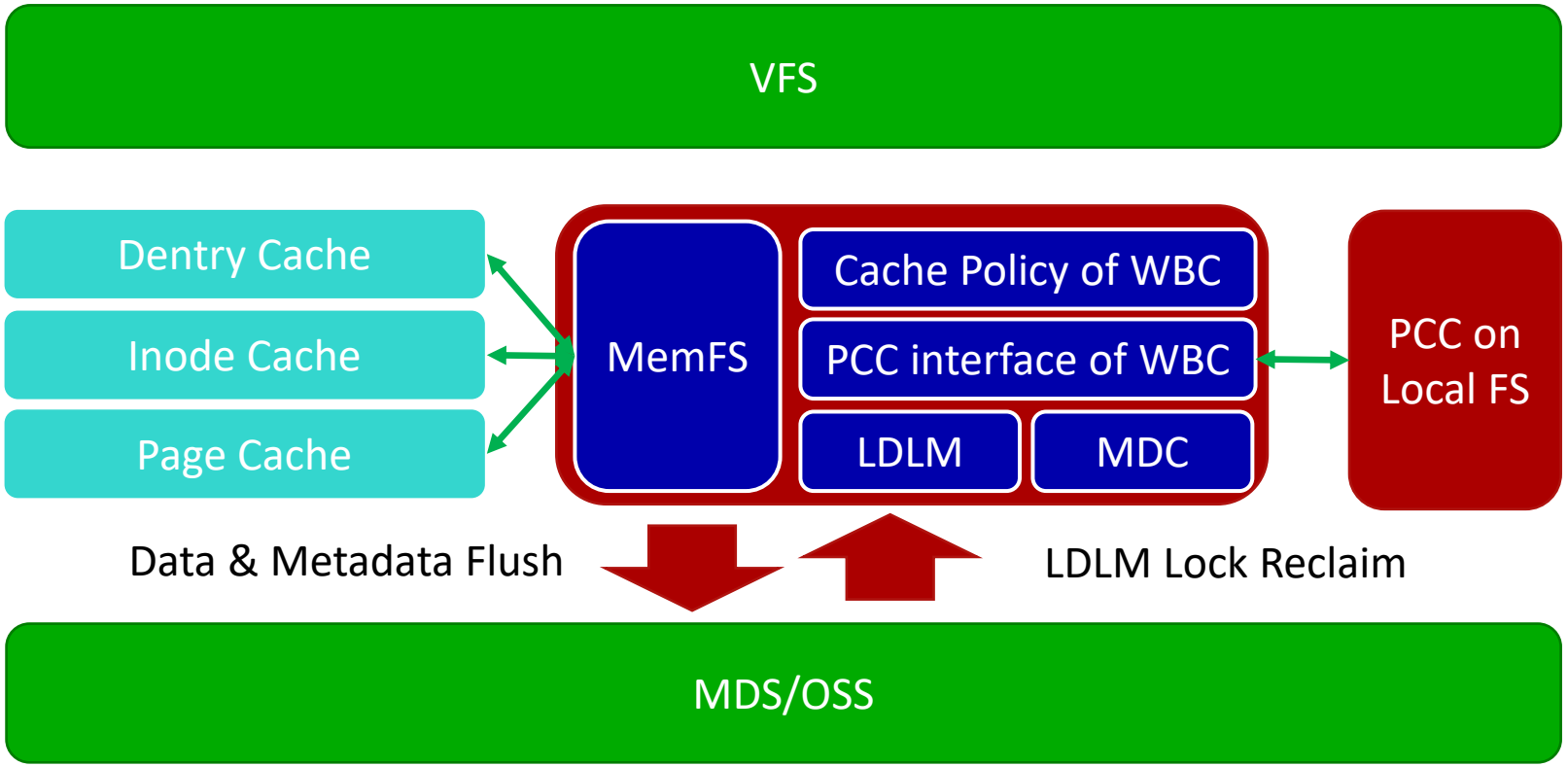
Lustre客户端持久缓存（PCC）的架构



Lustre客户端元数据回写缓存的基本思路



Lustre客户端元数据回写缓存的架构

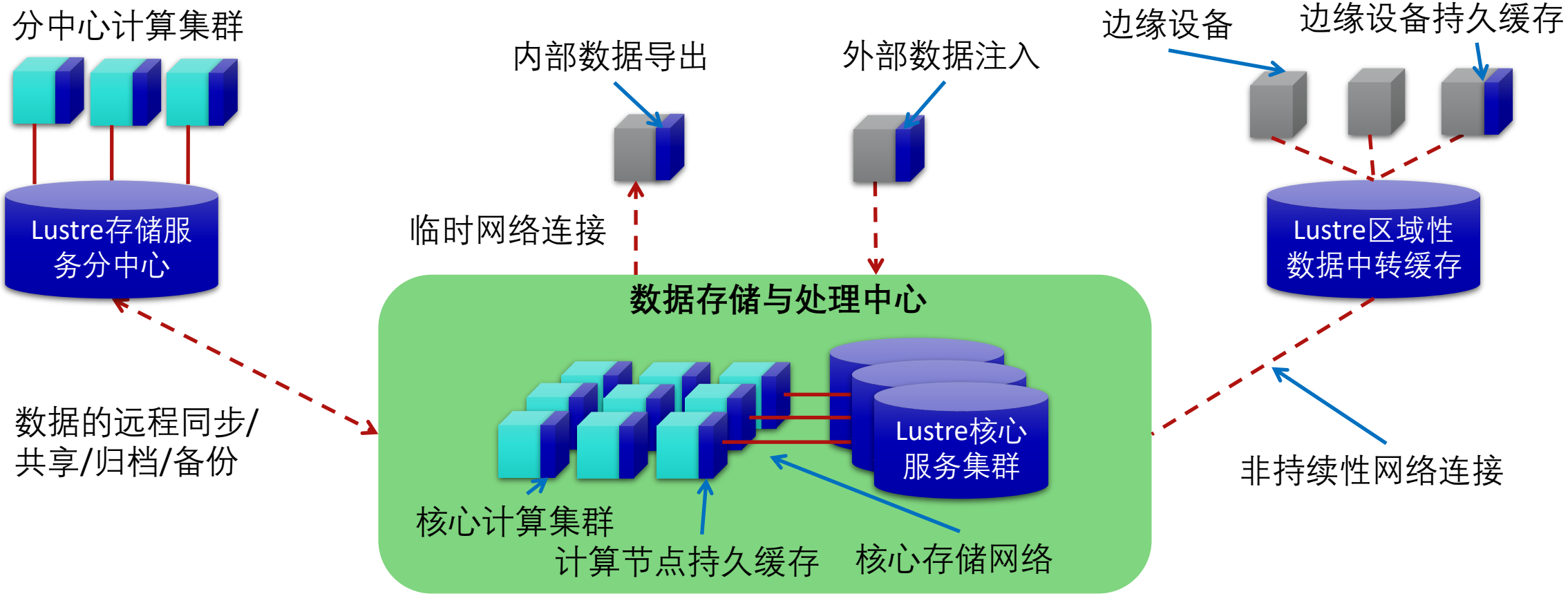


Lustre存储架构演进趋势

Whamcloud

- 新存储介质引导Lustre架构的持续演进
  - 更丰富的高速缓存层级
  - 更强大的客户端能力
- 未来趋势：客户端离线持久缓存
  - 客户端存储性能和容量持续增长
  - 数据访问的客户端局部性十分明显
  - 边缘设备日益增多
  - 网络的持续高速互通面临挑战
  - 跨地域数据同步需求强劲
  - 基于项目/文件集的数据管理机制日渐丰富

Lustre客户端离线持久缓存的应用场景



Lustre客户端离线持久缓存的技术储备

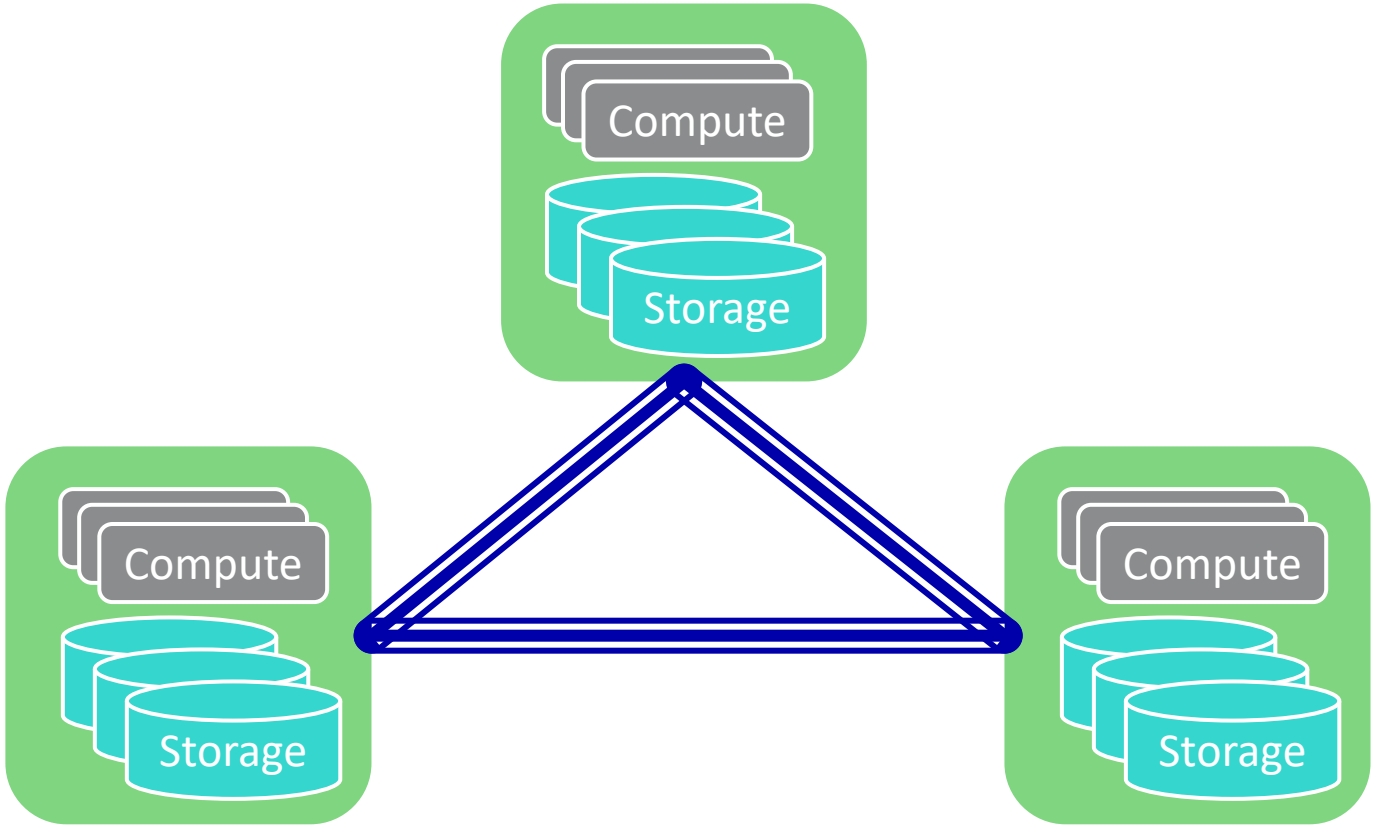
Whamcloud

- Lustre客户端离线持久缓存是PCC和WBC的自然延伸
- 基于Project Quota/ID的新扩展提供了数据管理工具
  - 文件按Project ID被分为不同的数据集
  - 灵活地将文件加入或移出数据集
  - 数据集内的所有目录/文件可在O(1)时间内被设置上一组可选标志
  - 只读标志：标识已归档数据集
  - 跟踪标志：标识需跟踪审计的数据集
  - 离线标志：标识离线客户端缓存中的数据集
  - 同步标志：标识需要增量同步到S3等存储的数据集
- 带有离线标志的数据集将不可访问，直至客户端重新连接

Lustre客户端离线持久缓存对多数据中心的支持

Whamcloud

- 高效的数据共享
  - 减少数据的无效传输
  - 共享数据的按需传输
  - 充分利用网络的并发度和带宽
- 清晰的数据管理
  - 避免数据一致性的紊乱
  - 避免脑裂问题
- 充分利用数据局部性
  - 允许各副本的本地修改
  - 各数据副本的并发访问



Lustre的技术演进目标

Whamcloud

- 分布式文件系统 + 并行文件系统
  - 并行文件系统的强一致性语义和并行访问能力
  - 分布式文件系统的高可靠性和横向扩展能力
- 极致性能 + 综合的数据处理和管理能力
  - 利用各类客户端高速缓存获得极致的性能
  - 利用服务端丰富的功能获得全面的数据处理和管理能力
- 强一致性语义 + 便捷的数据共享与同步
  - 应用免受弱一致性语义的困扰
  - 数据可在各数据中心间、数据中心和边缘设备间自由流转与共享

Whamcloud

Thank You!
Questions?