作者	时间	QQ技术交流群
perrynzhou@gmail.com	2022/09/18	672152841

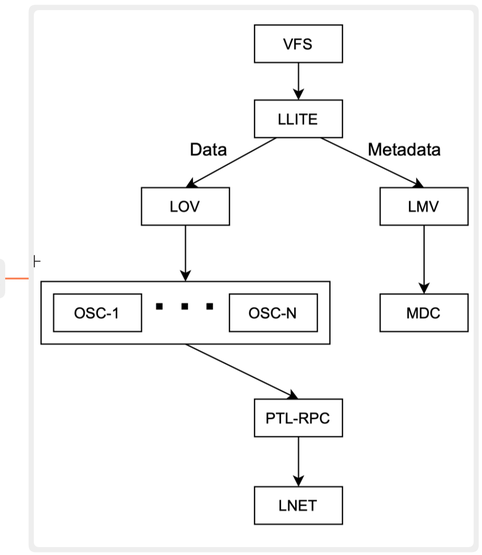
存储内核技术交流

微信扫描二维码，关注我的公众号

Lustre整体实现概览

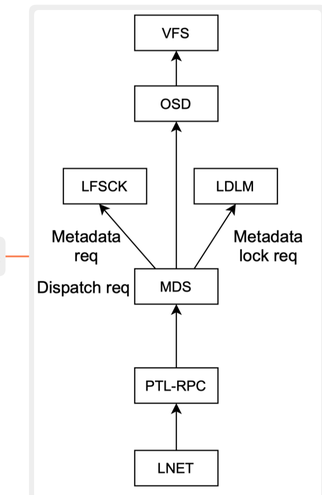
核心架构

Lustre Client

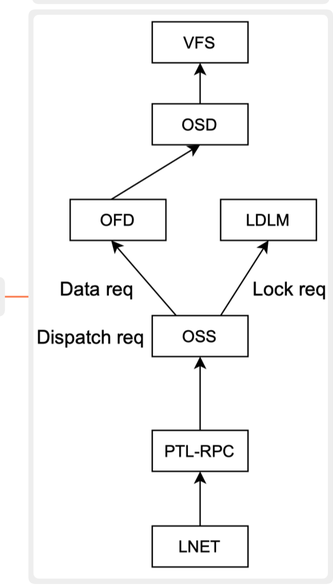


Lustre MGS

Lustre MDS



Lustre OSS



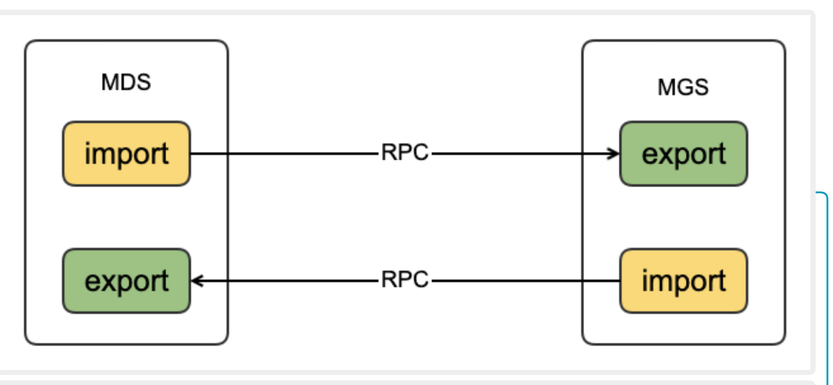
工具链

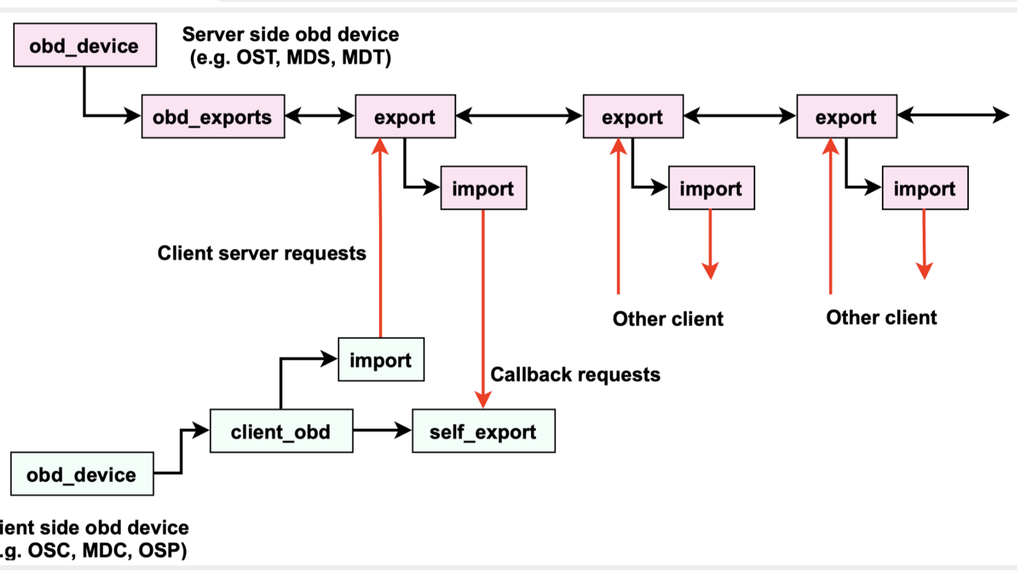
- mkfs.lustre — 格式化lustre后端文件系统
- tunefs.lustre — 后端文件系统配置信息的调整
- lctl — lustre服务的控制工具
- mount.lustre — 挂载(启动)lustre客户端和后端文件系统
- lfs — 配置和查询文件/目录/空间等工具
- lfs_migrate — 在不同OST之间迁移数据，用来均衡不同OST之间数据
- llog_reader — 解析lustre配置日志

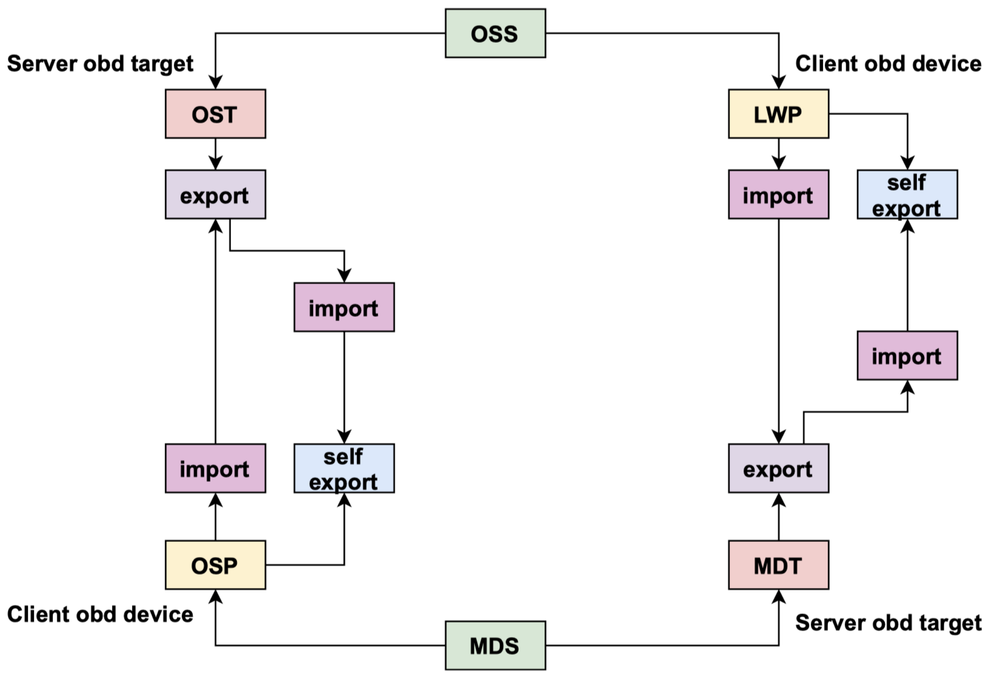
Client Kernel Module

- MGC
  - 介绍
    - 一个MGS对应一个MGC
    - lustre虚拟文件系统层和MGS之间的通信接口
  - 功能
    - 处理lustre log
    - 分布式锁管理
    - 文件系统设置
- MDC — 每一个MDT对应一个MDC
- OSC — 每一个OST对应一个OSC

服务通信模型

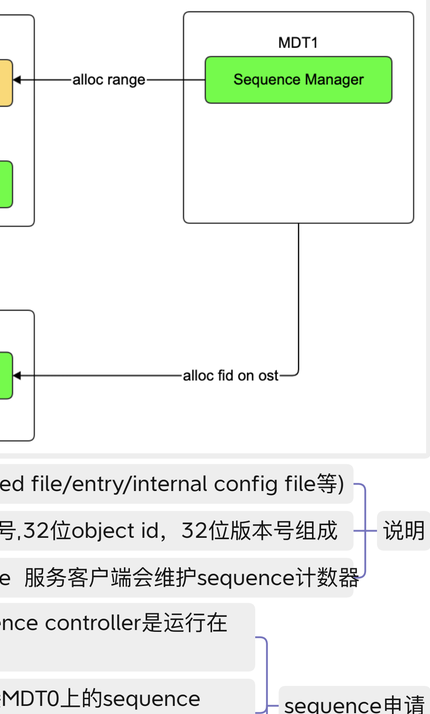




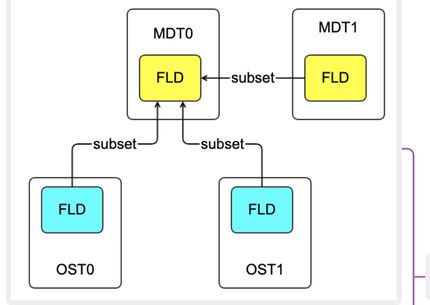


FID

- Lustre中所有数据的唯一标识(striped file/entry/internal config file等)
- 说明
  - FID长度是128位的，是由64位sequence序号,32位object id，32位版本号组成
  - lustre中的fid不会复用，fid被申请的时候，sequence 服务客户端会维护sequence计数器
- sequence申请
  - sequence服务分为sequence controller和sequence manager.sequence controller是运行在mdt0上，MDT和OST运行sequence manager.
  - 启动时候MDT和OST中运行的sequence manager会连接MDT0上的sequence controller申请一个唯一的sequnece 的范围
  - lustre客户端在首次数据写入都会像运行seqneunce 的MDT和OST申请sequence序号
- 接口
  - OSP子系统申请新的sequence number(mdt之间和ost之间) — seq_client_get_seq
  - lustre客户端申请新的sequence number — seq_client_alloc_fid



FLD



- fid用来表示对象的唯一性，但是fid在后端的osd中定位的对象的位置映射关系靠的是object index
- object index (oi)
- 说明
  - fid的存储数据库,mdt0存储了全量的fid,其他的lustre server存储了FLD的数据子集
  - 客户端请求的fid都会在server的FLD进行lookup