中移苏研存储产品化之路

中国移动苏州研发中心

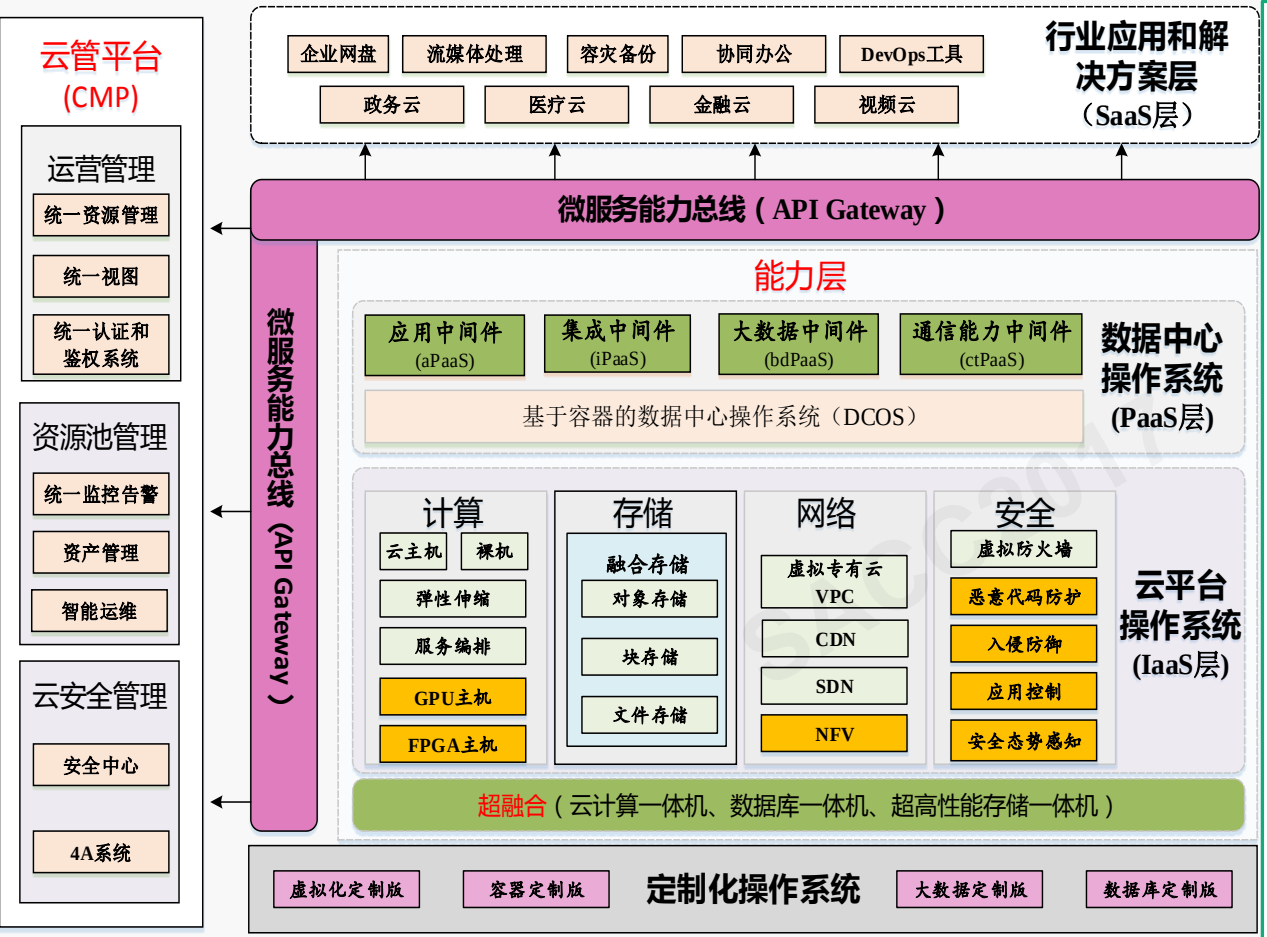
云计算产品部 刘军卫

中移（苏州）软件技术有限公司

苏州研发中心（对内），占地**480**亩，总建筑面积**36万平**，目前在职人数**850**人，远期规划**4500**人，聚焦**云计算**、**大数据**、**IT支撑系统**三大领域，中国移动IT能力内化和业务创新发展的中坚力量。

云计算产品部

目前在职**240**人，**开源**与**自主研发**相结合，打造产品化的**计算**、**存储**、**网络**、**安全**、**云管平台**等IaaS、PaaS全线云计算产品，产品部署规模超过**20000**台服务器，研发和工程实力业内领先。



核心设计理念：一级平台，两级管理

统一化（运营与运维统一）

- ✓ 统一用户管理与认证鉴权
- ✓ 统一资源管理和视图
- ✓ 统一监控、告警

分层解耦、微服务化

- ✓ 微服务总线，实现业务与能力前后端分离，实现业务的标准化接入
- ✓ 分层解耦，IaaS/PaaS/SaaS分层搭积木式累加设计，实现资源动态联动
- ✓ 功能组件和业务逻辑模块化、服务化，实现以应用为中心的能力化封装

智能化、自动化

- ✓ 智能化业务部署与运维
- ✓ 智能化资源分配和调度
- ✓ 智能化的服务发现和治理

控制平面容器化（CCP）

- ✓ Containerization Control Plane
- ✓ 基于Kubernetes的微服务化控制平面

- **开源SDN方案**
  - OpenDaylight (2013,java)
  - ONOS (2014,java)
  - RYU (2012,python)
- **商业SDN方案**
  - 阿朗，华为，华三，中兴
  - 思科，Juniper，NSX
- **存在问题**
  - underlay与overlay统一管理
  - 物理机/虚拟机/容器统一管理
  - 与Neutron对接问题
  - -设备兼容性适配

从中国移动看存储需求

- **共享硬盘**（块存储，替代SAN设备或者专用存储设备）
  - 数据库，和目视频，电信网性能、告警数据，归档，容灾等，量最大，百PB以上级别，绝大部分要求ISCSI支持
- **云存储**（对象存储）
  - 企业网盘，无纸化办公，CDN，归档，容灾等，预计18年需求超过100PB
- **NAS存储设备**（文件存储）
  - 139邮箱，和目视频，人工智能，大数据框架等，预计18年需求在200PB左右

中移苏研存储产品线

统一存储管理平台

- 自动化部署
- 用户管理
- 监控告警
- 资源统计
- 性能分析
- 存储池管理
- 块存储管理
- 对象存储管理
- 文件存储管理

网盘 | **存储网关** | **和目视频** | **无纸化办公** | **归档备份**

云存储 | **云硬盘** | **云化NAS** | **云化CDN**

S3 | Swift | Qemu | iSCSI | NFS | CIFS

对象存储（BC-oNest）

块存储（BC-EBS）

文件存储（BC-EFS）

超融合存储（BC-Cube）

一体机、定制化服务器

ceph


接口层
QEMU
ISCSI
NBD
服务层
快照
克隆
增量备份
热迁移
VAAI
实时镜像
跨存储迁移
QoS
引擎层
分布式哈希
强一致副本
全冗余设计
并行重建
自动均衡
智能修复
瘦存储
线性扩展
硬件感知
故障域隔离
存储池隔离
热点缓存
硬件层
X86服务器
SAS/SATA/SSD
磁盘控制器管理
1GE/10GE
磁盘错误/慢盘检测
SSD寿命监控
存储管理
资源监控
性能监控
滚动升级
在线扩容
告警管理
日志管理
磁盘定位
数据盘漫游
部件更换
可视化

大云1.0发布

- 基于IPSAN的块存储
- 自研对象存储

1

大云3.0发布

- 基于Sheepdog，容量盘
- 基于IPSAN，性能盘
- Cinder统一管理性能盘、容量盘

5

- 全面基于Ceph提供块、对象存储
- 40PB的对象存储集群
- 双集群20PB块存储集群

10

大云4.0发布

- 优化Ceph性能，SSD性能盘
- 支持ISCSI，开始试点
- 支持物理机挂载
- 存储一体机，支持高性能场景

21

2010年 **2015年** **2016年** **2017年8月**

- 2016年5月对象存储从oNest转向Ceph RGW
- 2016年10月块存储从Sheepdog转向Ceph RBD
- 生产环境块存储**400+**节点，**15PB+**容量，对象存储**600+**节点，**30PB+**
- **首个**对象存储多数据中心生产环境案列

- 16年11月发送第1个Ceph补丁
- 累计至今**13**人共计被接受**150**个补丁，提交**6**个特性，修复**50**多个Bug
- Ceph社区排名**5**位，国内第**2**位
- Ceph RBD iSCSI项目(TCMU)**最大**贡献者之一

产品化特性（1）– ISCSI

- **LIO：**LinuxIO（LIO）是 Linux 里面一个标准、开源的 SCSI Target 子系统。LIO 是下一代基于软件实现的各种 SCSI Target 主流解决方案，其支持 的SAN 技术中所有流行的存储协议。
- **TCMU：**通过 UIO（用户态驱动实现技术）把 SCSI 命令从 LIO Core 透传到用户空间，使得可以在用户空间实现各种 Target 驱动。
- **TCMU-Runner：**tcmu-runner 是 TCMU 在用户态下的驱动部分，也是 TCMU 模块的主要处理逻辑单元。其主要工作是从 TCMU 内核模块映射 ring buffer 到用户空间，然后读取、处理、并更新各个 SCSI 命令。

LIO + TCMU + LIBRBD

- 目前社区主流
- Redhat，Suse，IBM大力推进
- 代码易于维护

LIO + KRBD

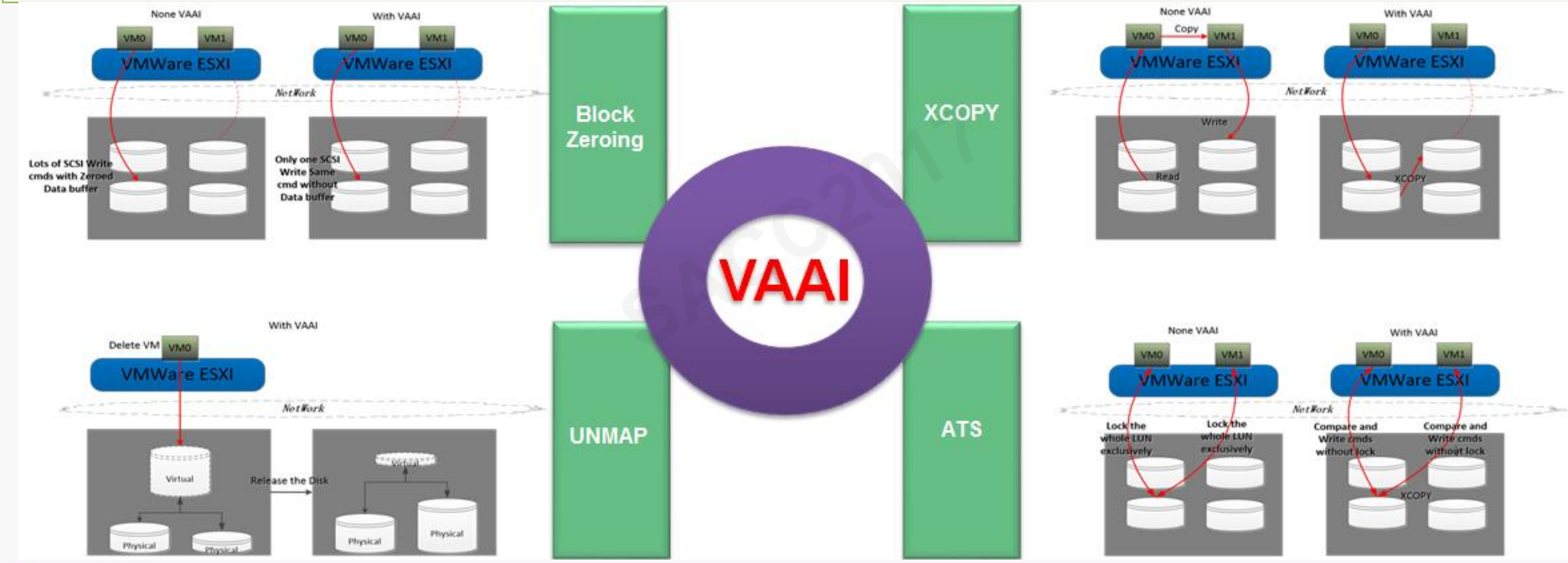
- stgt无社区
- 国内厂商采用毕源定制版本

STGT + LIBRBD

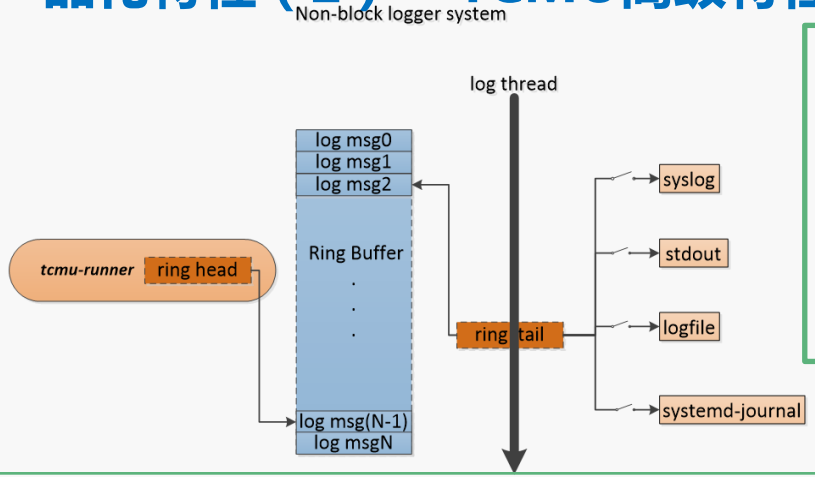
- krbd功能、性能全方面落后librbd
- 需要高版本内核

产品化特性（1）– ISCSI：VAAI高级特性支持

苏研主导了TCMU+LIO对VAAI特性支持的开发，在TCMU社区合并了**50+Commit**，在Ceph的Librbd端合并了**Writesame**与**CompareAndWrite**两大特性，**XCOPY**与**UNMAP**优化。



产品化特性（2）– TCMU高级特性开发

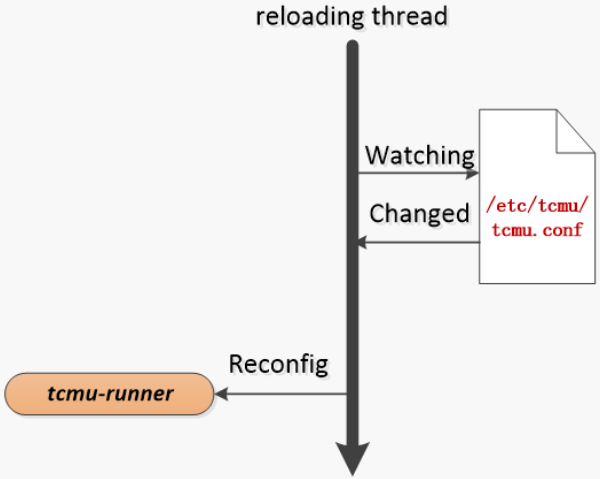


TCMU-runner Logger

- 原来的实现强依赖于syslog接口，难以维护，且存在出错时阻塞工作线程的问题
- 独立实现Non-block Logger子系统，引入自己的**ring-buffer，跟syslog实现解耦**从而避免了阻塞问题，并支持**多种方式**的日志输出（syslog、stdout、logfile）
- 代码少于1000行，易于维护

TCMU-runner Dynamic Reloading

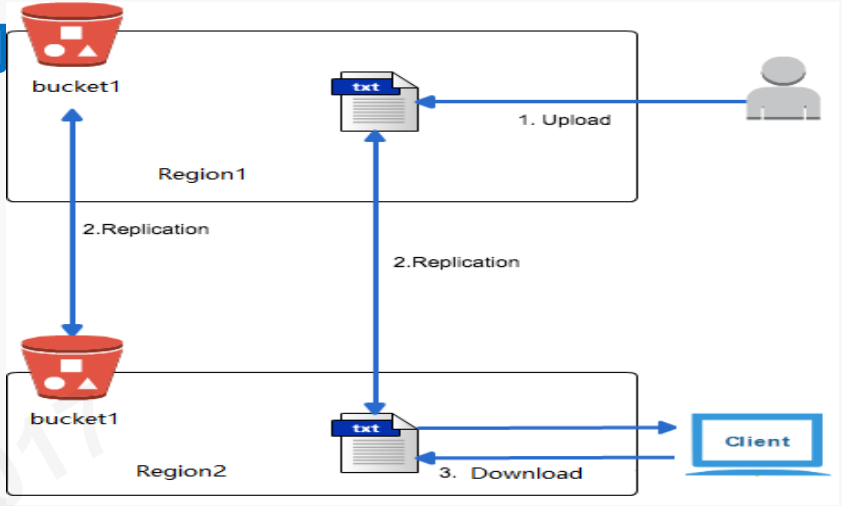
- 原来的tcmu-runner修改配置文件之后，需要重启服务使配置生效，修改配置会影响业务。
- 独立实现Dynamic Reloading技术，通过引入新的**独立线程**reloading thread**监听配置文件的修改**，从而支持TCMU配置的**动态修改**。

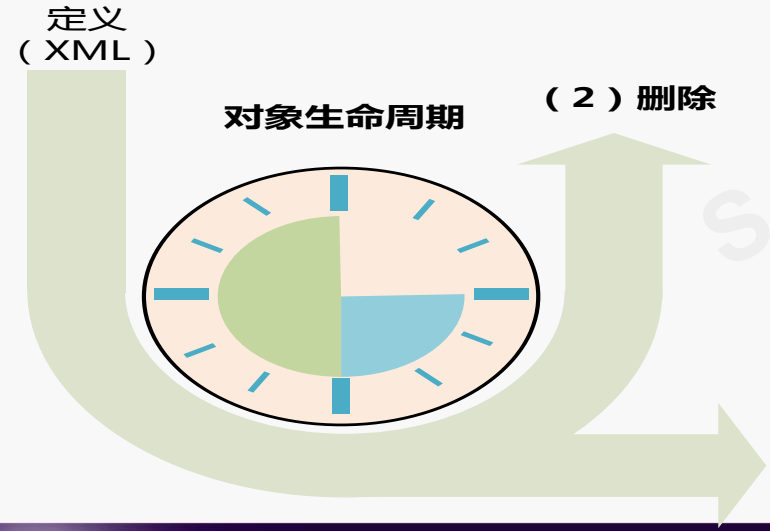


产品化特性（3）– 生命周期管理&桶级别同

S3对象生命周期

- 为RGW添加非当前版本对象生命周期管理机制
  https://github.com/ceph/ceph/pull/13385
- 为RGW添加冗余delete marker清除机制
  https://github.com/ceph/ceph/pull/14703





桶级别同步

- 原来RGW只支持**Zone**级别同步，S3也并不支持桶级别同步
- 苏研和社区协作开发了**桶级**别同步
  https://github.com/ceph/ceph/pull/15801

产品化特性（4）– 流式存储

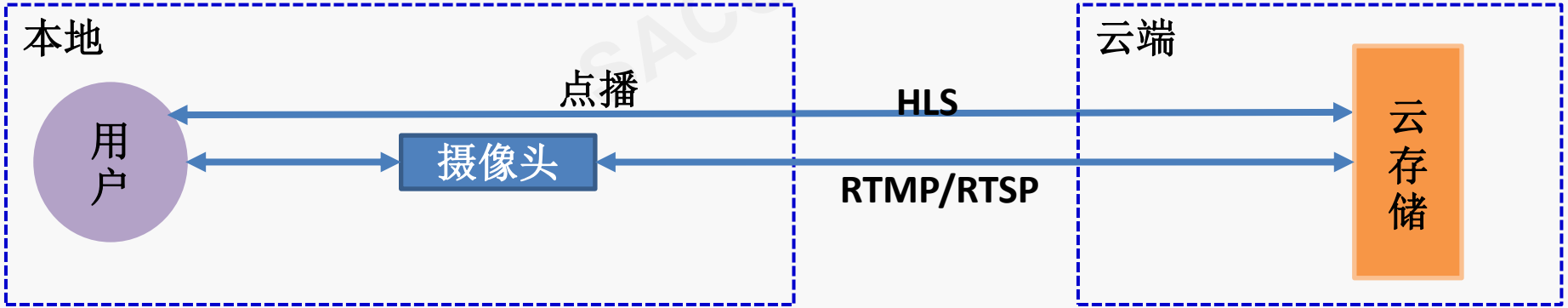
□ 支持多种协议
- 支持RTMP推流上传
- 支持RTSP推流上传
- 支持HLS观看视频

□ 丰富的API
- 签名API
- 推流API
- 点播、直播API

□ 简化视频存储方案
- 视频采集客户端直接推流到对象存储
- 支持点播/直播

□ 支持第三方软件
- 支OBS/ffmpeg
- 其他第三方客户端

核心代码开发 | 推流/观流实现 | 推流/观流日志记录 | 自定义配置



产品化特性（5）– 桶级日志

记录内容丰富

- 请求类型、访问对象名称、请求时间、请求处理时间、客户端IP、请求URI、用户自定义参数都可记录

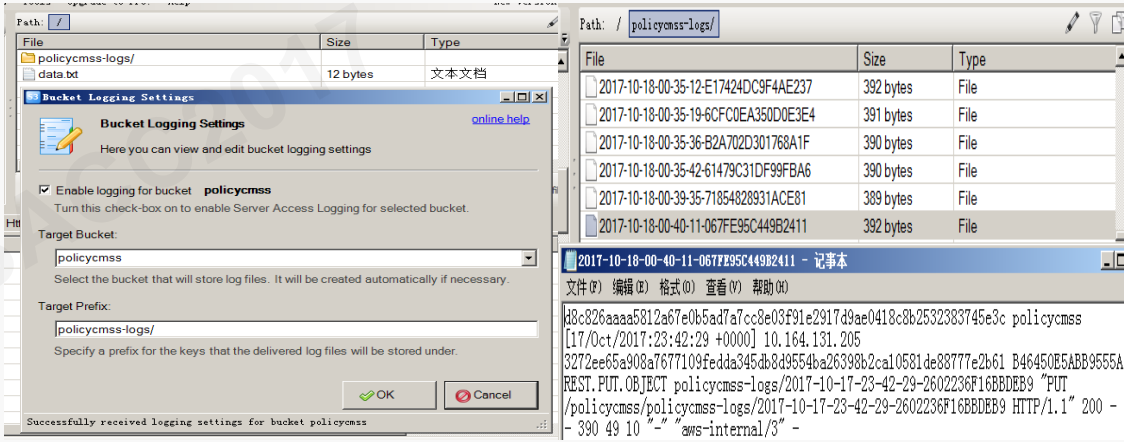
自定义日志存放位置

- 可指定日志存放在位置
- 存放在其他桶可选性能优先或者容量优先

CLI和REST接口

- CLI接口查看生成状态
- REST接口配日志置桶日志更方便

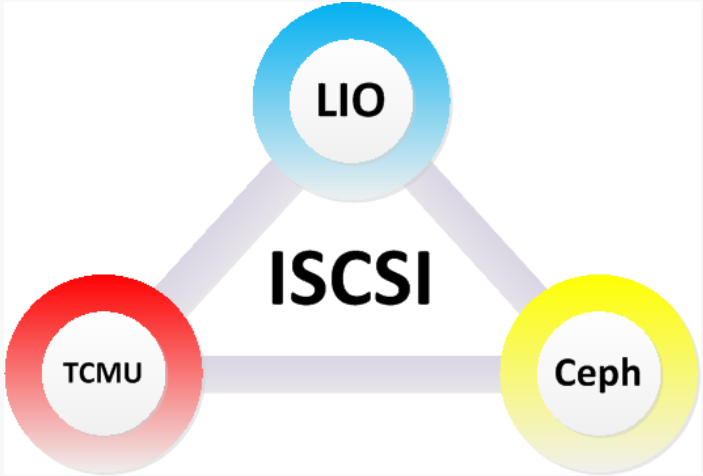
共享资源的统计分析，例如
- 下载次数最多的文件是哪个？
- 下载次数最多客户端IP是哪个？



后期规划（1）

基于TCMU+LIO的 下一代iSCSI解决方案

- 强化集成基于TCMU+LIO+Ceph的iSCSI解决方案，并大力推广，使之成为**业内首选**的**标准化解决方案**。
- TCMU **Ring Buffer CMD Area**的**Dynamic Grow/Shrink**开发，优化CMD处理效率和节省内存使用。
- **SCSI命令集**完整支持。

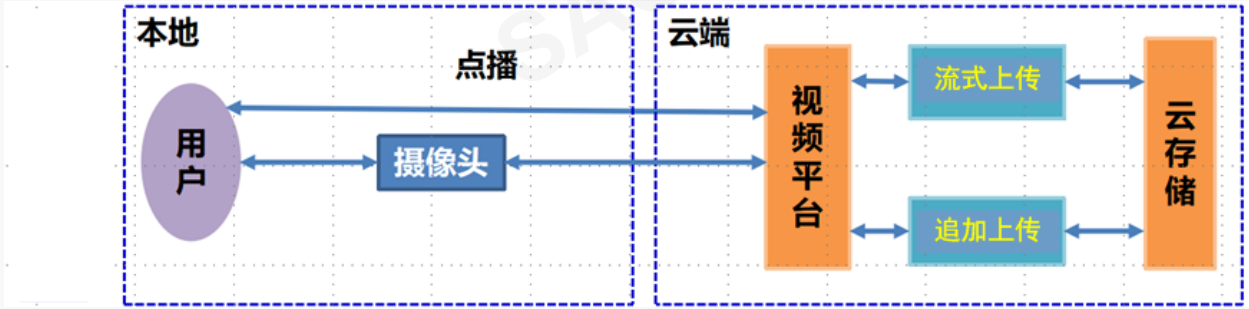


后期规划（2）

针对视频应用优化的对象存储方案

流式上传。采用**RTMP/RTSP 协议进行**推流上传，将视频数据直接存放到对象存储中，转储成HLS文件，可用于**视频的点播或直播。**

追加上传。提供对象的追加上传功能，可以在对象的**尾端追加数据**，提高传输效率，满足视频应用的需求。



后期规划（3）

基于Key/Value存储存储的小文件性能优化方案

- 开源的Glusterfs小文件性能提升是一大难题，苏研计划使用**Key/Value存储**加速元数据处理，提升**小文件性能**。


收获云计算最新鲜优质的资源分享！
抢先知晓中移苏研云计算最新动向！
苏研大云人


THANKS