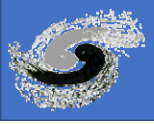
高能物理数据的存储和管理

汪璐

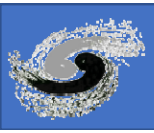
高能所计算中心

2021-8-17



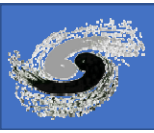
大纲

- 需求和挑战
- 高能所计算中心的海量存储系统
  - 分布式文件系统
  - 磁带管理系统
  - 访问接口和使用建议
- 分布式环境下的数据管理技术
- 更详细的课程及课件
- 问题和反馈



高能物理计算是数据密集型计算

- 对海量的实验和模拟数据进行重建处理、统计分析是验证理论模型和发现新物理的主要途径
- 数据相关的IT技术是高能物理计算绕不开的重要组成部分
  - 数据发现
  - 海量存储
  - I/O性能优化
  - 数据共享
  - 数据长期保存
- 快速增长的数据量和分布式的计算环境给数据的存储和管理提出了新的挑战



数据量的快速增长

- BESIII/BEPCII
  - ~1 PB/年
- LHC实验
  - 50 PB每年，传到高能所3-5PB/年
- 中微子实验
  - 大亚湾：数百TB/年
  - JUNO：2022年运行，预计3PB/年
- 宇宙线实验
  - LHAASO，目前3 TB/天， 2021年起，预计6PB/年
- 空间天文实验
  - HXMT，AliCPT， GECAM
  - HERD，eXTP (规划立项中)
  - 数百TB/年
- 光源实验
  - HEPS, 500TB/天
  - 数据保留半年，总量~100PB

对存储系统的需求

- 百PB存储容量, 百GB/s 聚合数据读写带宽
- 横向扩展的I/O性能
- 支持7x24小时读写访问的高可用性
- 高可靠性
- 性价比

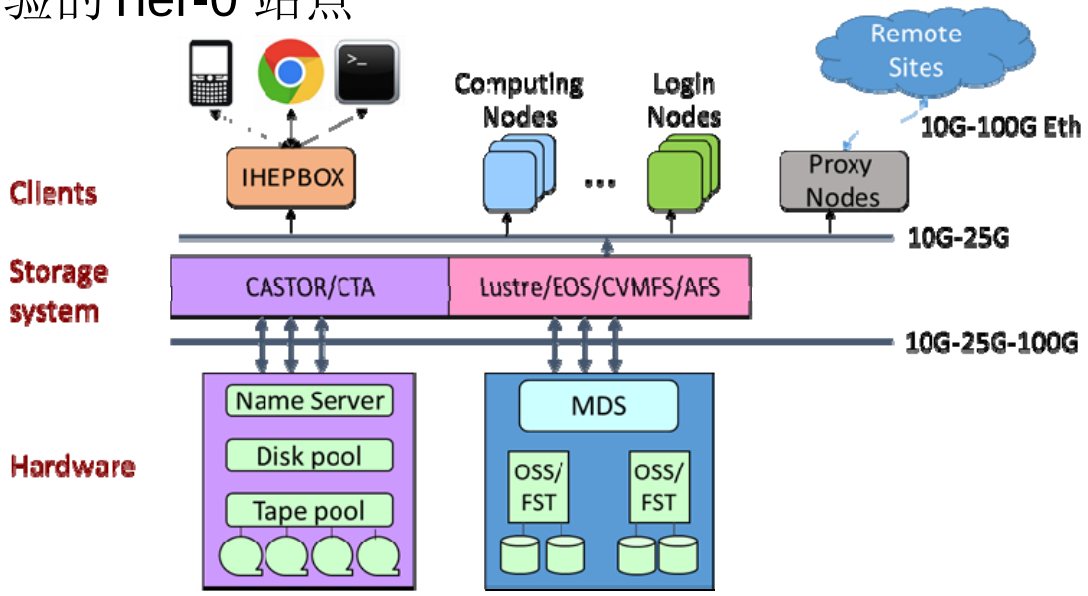
高能物理特色需求

- 混合多样的应用类型和数据访问模式
  - 模拟、刻度、重建、分析、机器学习...
  - 以后台作业大块读、一次写多次读为主，混合前台交互型小文件读写、后台随机读写
- 支持多种数据访问入口，统一视图
  - 用户数据
- 跨域数据统一视图，透明访问
  - 针对实验数据和软件存储
  - 有限的广域网带宽和延时条件下，保证二级远程站点对数据的高效访问
- 数据长期保存
  - 实验数据在磁盘上存放的时间长达数年
  - 数据在数十年的实验周期内，可读、可分析

高能所的海量存储系统

高能所的海量存储系统

- 高能所计算中心是中国高能物理数据处理中心
  - BESIII，JUNO，LHAASO大型实验的Tier-0 站点
  - LHC实验的Tier-2 站点
- 分布式文件系统 （磁盘）
  - 带宽、随机读写、集群共享
  - 热数据
- 磁带管理系统
  - 绿色节能、高性价比
  - 冷数据



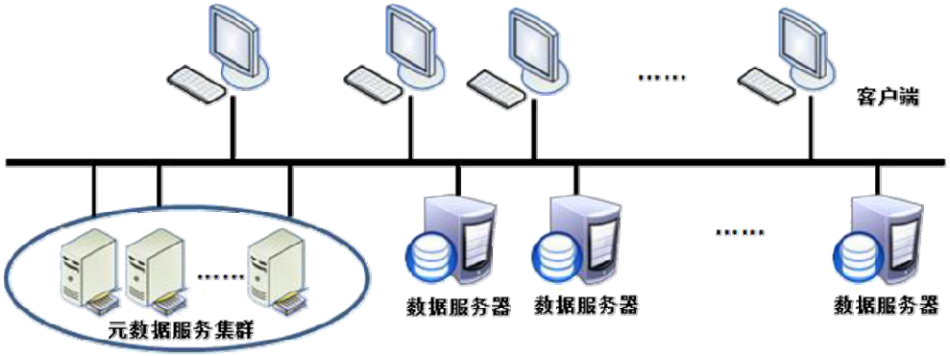
- 管理个人数据的云存储系统 IHEPBOX，备份系统AMANDA，软件和镜像管理系统CVMFS等

分布式文件系统

- 分布式文件系统将数千盘磁盘，数百台服务器组成单一系统镜像
  - 集群上所有计算节点和登录节点看到的是同一份视图，可以像访问单机文件系统一样访问海量的存储资源
  - 解决用户资源分配、访问控制、数据可靠性、服务高可用、分级存储等问题
- 高能所主要的分布式文件系统：
  - Lustre ：实验数据和用户数据，20 PB
  - EOS：实验数据，10 PB
  - AFS：用户认证，home目录，百TB

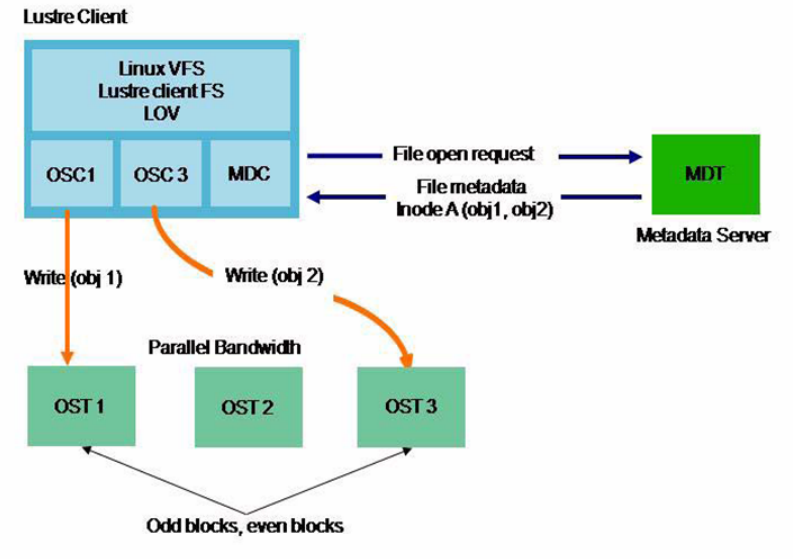
分布式文件系统的组成

- 元数据和数据
  - 元数据指文件的属主、访问权限和时间、扩展属性等
  - 数据是文件的内容
  - 元数据和数据分离的设计能够更好的保证数据的一致性和性能的可扩展性
- 元数据服务器（集群）
  - 解决文件的逻辑文件名和物理存放位置的映射关系
    - /home/myfile-> (obj1 on dataserver1, obj2 on dataserevr2 …)
  - 控制文件访问权限
    - (set/get attr, set/get acl, chmod …)
  - 管理名字空间
    - Lookup ,create, delete
- 数据服务器（集群）
  - Read, write, seek, create, delete

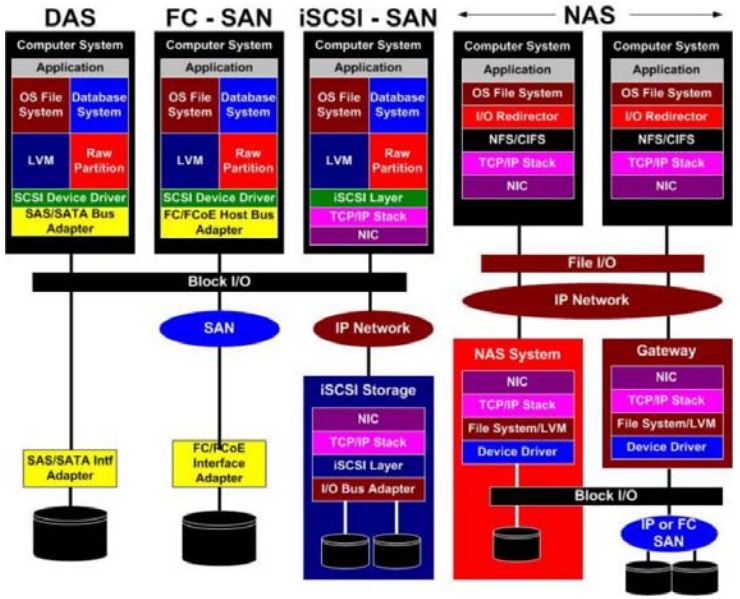


分布式文件系统的I/O路径

- 客户端
  - 实现POSIX或者XROOTD数据访问协议，与服务器集群通信，完成上层应用的I/O 请求调用



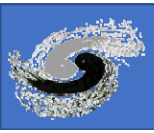
客户端的I/O路径



服务器的I/O路径

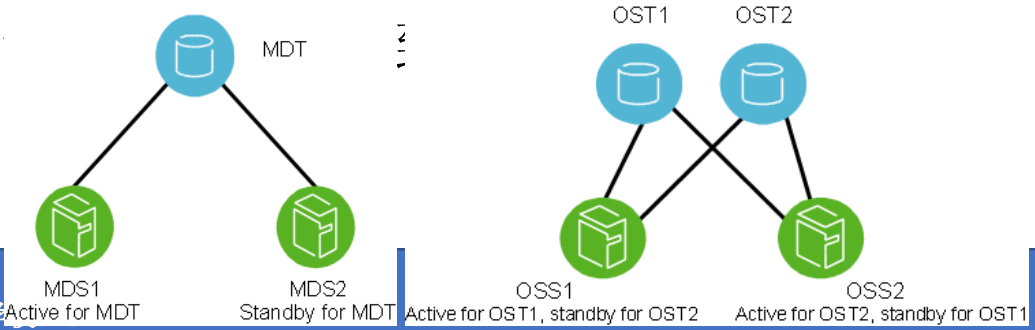
数据分布和均衡

- 文件创建时，元数据服务器会根据文件的分条规则在1个或多个存储设备上为文件创建1个或者多个对象
  - 当存储设备水位差不多时，采用循环放置法
  - 当设备水位相差很大时，采用加权循环放置法
- 新加入了存储设备，如何保证水位均衡和负载均衡？
  - 根据数据放置算法，持续的数据删除和写入可以保证存储水位最终均衡
  - 这种自然选择的方法会导致新数据集中分布在新设备，老数据集中分布在老设备上，一般管理员会通过后台数据迁移，手动均衡



服务高可用

- 数百台服务器组成的存储集群中，任何一台服务器故障，会导致全部或者部分数据无法访问，影响集群计算效率
- 高可用设计可以缩短服务器故障的影响时间
  - 在管理服务器中为每个存储设备注册多个可以访问到它的IP地址
  - 备机发现主机down掉的时候，mount 相关存储设备，接管存储服务
    - HA 软件，自动mount；人工检查，手动mount
  - 客户端发现主机IP不通后，会根据注册备机地址，联系备机，恢复通信
  - 存储系统通过transaction 保证故障
- 服务器之间可以互为备机

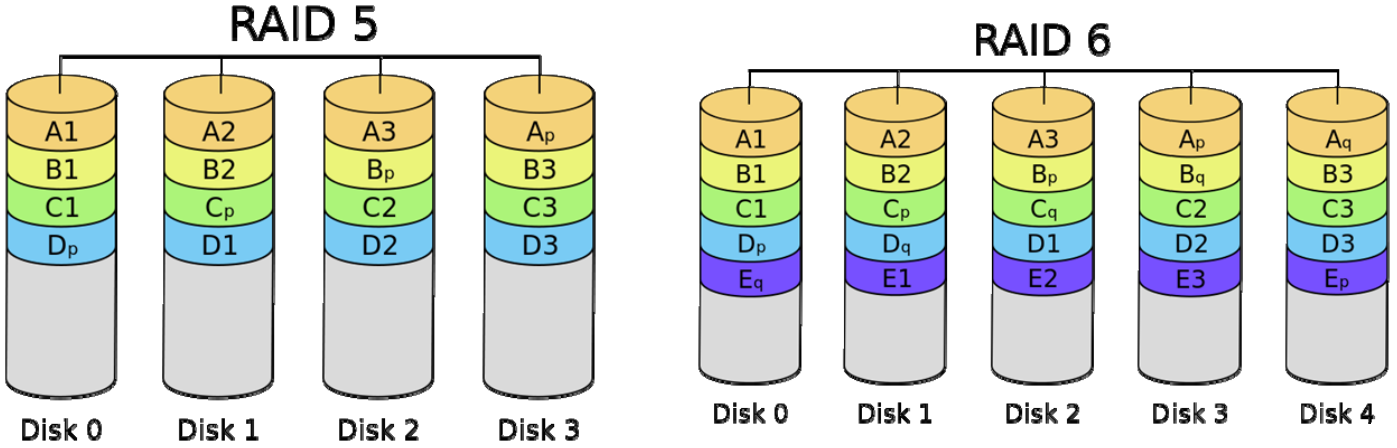


数据冗余

- 单个磁盘是很难保证数据可靠性的
  - 根据CERN-IT的统计，6万块磁盘的系统中，每天的坏盘数大约为10块，每年系统中有1%-10%的磁盘会出现损坏
- 常用的数据冗余
  - 数据副本：相同的数据存放多份
    - 块设备层 or 文件系统层，软件or 硬件
  - RAID：Redundant Array of Inexpensive Disks
    - 块设备层，软件or 硬件

RAID算法

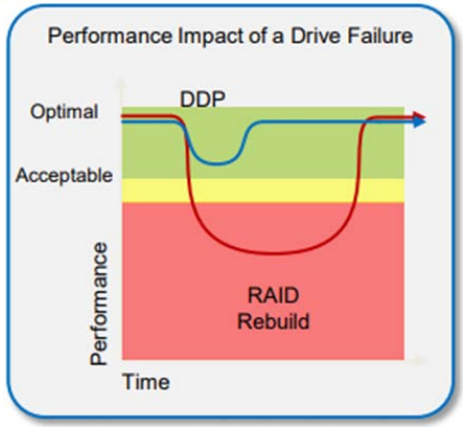
- RAID 5 ：将N磁盘分成长为N的条带，条带中的N-1个数据块存放数据，1个数据块存放数据块的XOR校验
  - 阵列中的一个磁盘损坏，数据可以通过其他磁盘上的结果重建

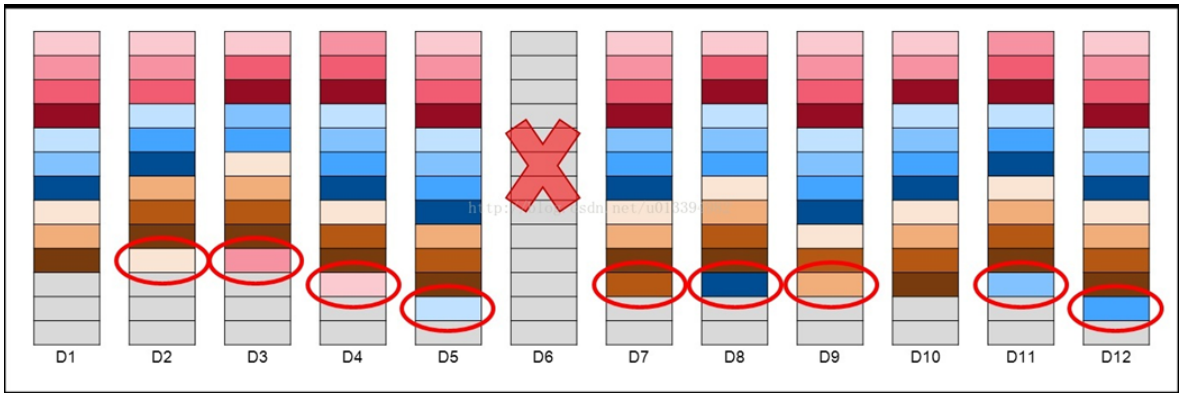


- RAID 6 :将N磁盘分成长为N的条带，条带中的N-2个数据块存放数据，2个数据块存放数据块的校验值，允许两块磁盘损坏

RAID改进

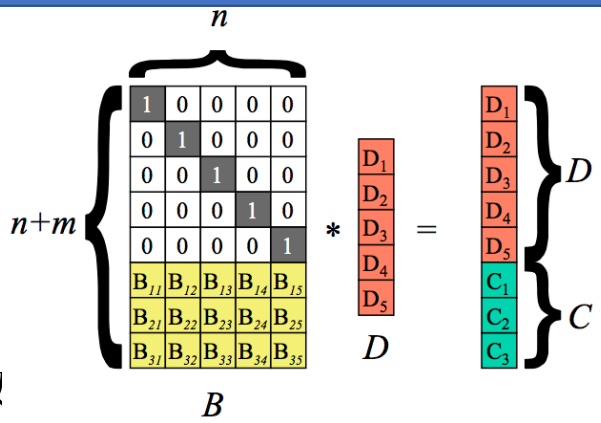
- 随着磁盘容量的增大，传统RAID中重建一块物理磁盘的时间越来越长
  - 按照200MB/s的写入速度，重建一块12TB的硬盘需要多长时间？
- 分布式RAID/RAID 2.0/DDP
  - 在一个更大的磁盘池M中划分长度为N的虚拟化条带，数据块和校验块被分散到更多的磁盘中，重建效率更高，丢失数据的风险

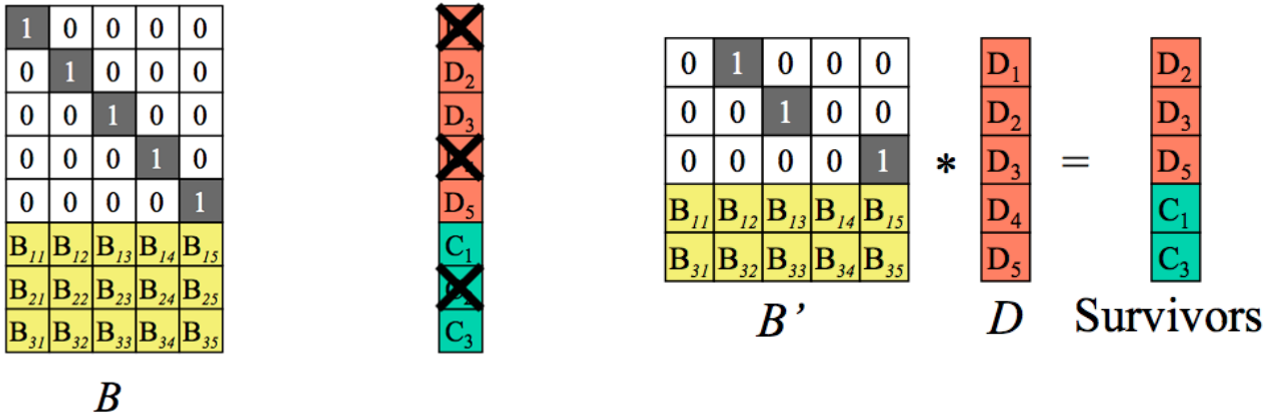


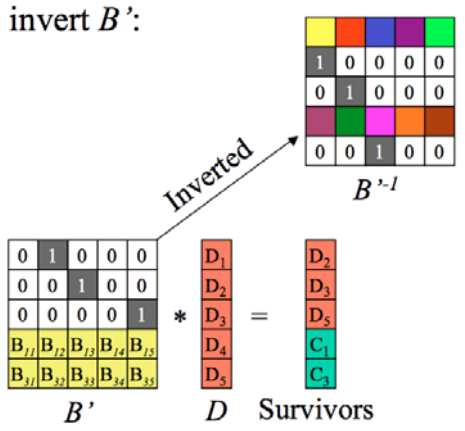


RAID改进

- RAIN: Redundant Array of Inexpensive Nodes
  - 分布式文件系统层实现的RAID
  - 兼顾的数据的可靠性和服务的高可用性
- 纠删码算法（以Reed-Solomon算法为例）
  - RS(n,m)：n代表原始数据块个数，m代表校验块个数







访问接口

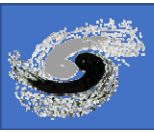
- 大部分用户盘和数据盘可以通过POSIX接口访问
  - POSIX：Portable Operating System Interface
- 数据访问接口与单机文件系统没有区别
  1. cat, vim, cp, mv, ln, chmod, chown,get/set acl, get/set fattr …
  2. 编程接口

```
#include <fcntl.h>
#include <stdio.h>
#include <unistd.h>
```

  3. 从ROOT 里面访问文件

```
TFile file(fn.c_str());
TFile* inputFile = new TFile(m.c_str(),"READ");
TFile*  inputFiles[m_fileNum] = TFile::Open(m_fileNames[m_fileNum].c_str(),"READ");
```

- /EOS/xxx的存储系统需要 EOS 命令和XROOTD 接口访问

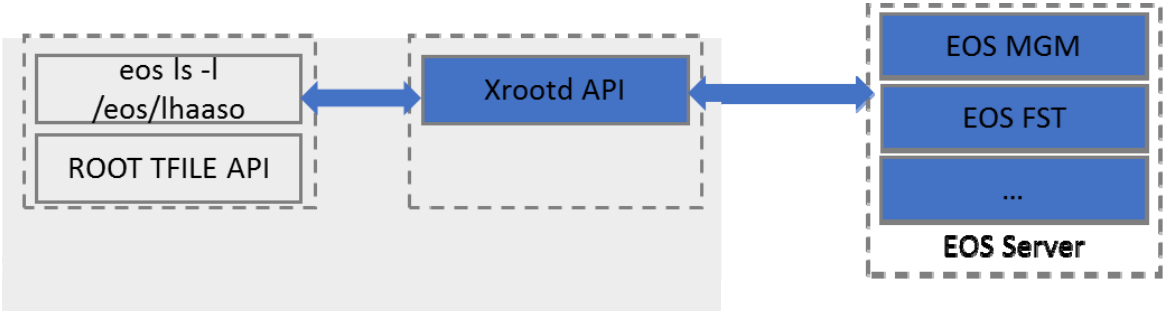


EOS 命令

- 在标准的文件系统命令前增加一个eos 前缀

```
-bash-4.2$ eos ls -lh /eos/lhaaso
drwxr-sr-+   1 lhaasore lhaaso       419.26 G Dec 15  2019 cal
drwxr-sr-+   1 lhaasore lhaaso       722.73 T Dec 15  2019 decode
drwxr-sr-+   1 root     root          51.41 T Dec 15  2019 experiment
drwxr-sr-+   1 lhaasore lhaaso       529.19 G Dec 15  2019 monitor
drwxr-sr-+   1 root     root         786.12 T Dec 15  2019 raw
drwxr-sr-+   1 lhaasore lhaaso       219.16 T Dec 15  2019 rec
drwxr-sr-+   1 root     root         915.70 T Dec 15  2019 simulation
```

```
$ eos cp /eos/lhaaso/cal/km2a/ADCcal/ReadDrawADC.C /tmp
[eoscp] ReadDrawADC.C            Total 0.00 MB   |==================| 100.00 % [0.0 MB/s]
```



命令格式：

```
$ eos recycle help
可支持ls（列表），purge（清空），restore（恢复）三种操作。
```

查看用户个人回收站中的文件

```
$ eos recycle ls
# Deletion Time          UID      GID      SIZE        TYPE          RESTORE-KEY      RESTORE-PATH
# ------------------------------------------------------------------------------------------------------------------------
Sat Dec 29 16:32:03 2018   zhangsan   u07      13          file          000000008b0f7bf /eos/user/z/zhangsan/test.txt
```

清空用户个人回收站中的文件

```
$ eos recycle purge
```

恢复用户个人回收站中的某个文件

```
$ eos recycle restore [--force-original-name|-f] [--restore-versions|-r] <recycle-key>
```

示例：

```
$ eos recycle restore 000000008b0f7bf
```

注意：目前/eos回收站中的文件只保留3天时间。

- 直接调用XROOTD API来访问EOS服务器，绕过任何内核模块
  - 参数是全路径名，没有$PWD, 不能使用通配符
  - 回收站，ls -lh直接显示目录子树的容量（POSIX系统需要du -hs）

XROOTD接口

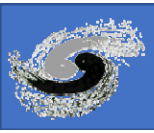
- 对于ROOT 格式的文件

```
TFile *filein = TFile::Open(root://eos01.ihep.ac.cn//eos_absolute_path_filein_name.root)
```

- 另外两种打开方式不支持
- 非ROOT格式的文件

```
TFile *rf = TFile::Open("root://eos01.ihep.ac.cn//eos/user/c/chyd/set.log?filetype=raw");
size = rf->GetSize();
printf("size is %d\n", size);
memset(buf, 0, 1024);
rf->ReadBuffer(buf, 1024);
printf("%s\n", buf);
rf->Close();
```

- 协议名称
- 管理服务器地址，可以设置成环境变量，在这里省略
- 绝对路径名
- 非ROOT格式文件的标识



分布式文件系统使用建议

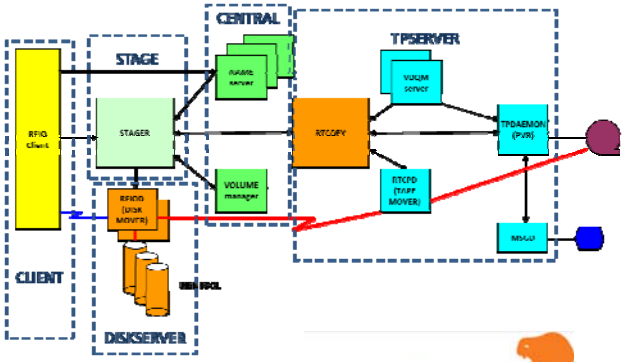
- 尽量使用软件框架访问数据
- 尽量避免在单个目录下存放过多的数据文件
  - 如果避免不了，可以生成一个文件列表，之后的数据处理直接访问该列表，不要使用ls *, rm * hadd *.root等命令
- 专盘专用，不要在用户盘做数据分析，生成大量的.root结果文件
  - 数据盘的物理资源比较少，数据分析产生的大量负载会严重拖慢其它用户的vi等交互式操作响应
- 不要把文件系统当成消息通信管道
  - 会给Lustre带来额外的负载EOS系统的一致性达不到该场景的要求
  - 考虑MPI等数据通信和同步协议
- 使用XROOTD协议访问EOS
- 程序中打开了文件一定要关闭

磁带管理系统CASTOR

- 磁带是非常适合存放冷数据副本的存储介质，各实验都有基于磁带的数据长期保存系统和策略
- 高能所的CASTOR系统已经运行了15+年
  - 3个主柜，共15个磁带柜，26 个LTO6/7 磁带驱动器
  - 聚合访问带宽 2.3 GB/s（可扩容）

	Volume (TB)	No. of Files(Million)
BESIII online*	1000	0.9
BESIII	1700	1.4
DYB *	1200	2.3
YBJ	535	0.6
Total	3618	0.7

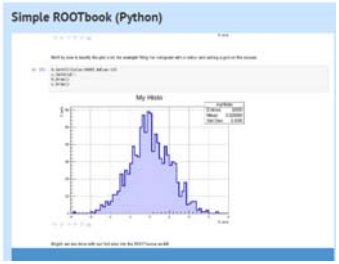
- 使用模式
  - 原始数据先写入CASTOR保存副本，再拷贝到磁盘做重建
  - 重建后数据、用户数据、备份数据定期写入CASTOR
  - 多试验共享，个人用户不可直接访问
  - 顺序读写，ALDER32 校验，远程离线副本
  - 以欧洲核子中心的磁带管理流程为主要参考
- 正在进行
  - 从CASTOR->EOSCTA的升级

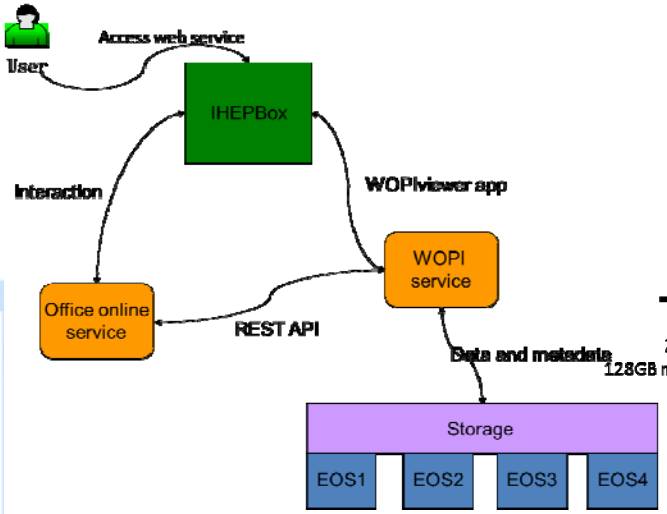


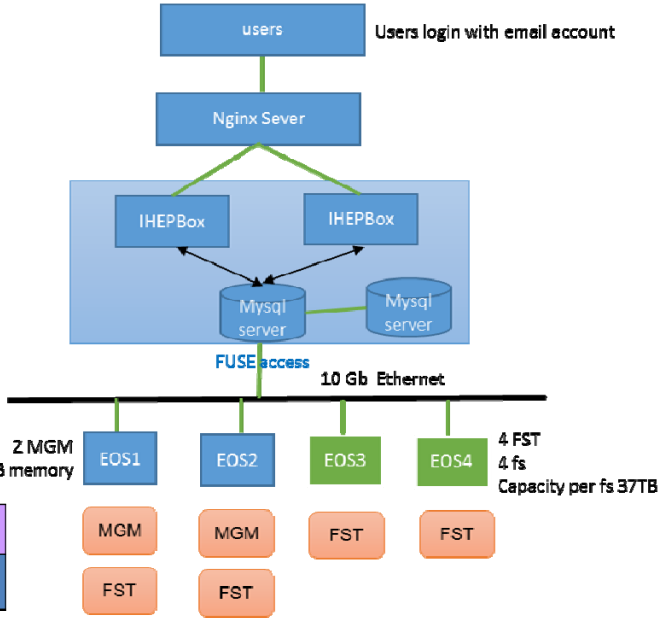
CASTOR
CERN Advanced STORage manager

IHEPBOX

- IHEPBOX是一个基于私有云技术Owncloud的可扩展的个人云存解决方案
- 用户的在多个设备中的数据可以融合为统一视图
- 使用相同的接口协议访问（HTTP）并进行数据同步
- 授权用户可以进行协同文档编辑
- 为更多的组织内APP应用提供共享数据空间
  - 交互分析、画图、视频等







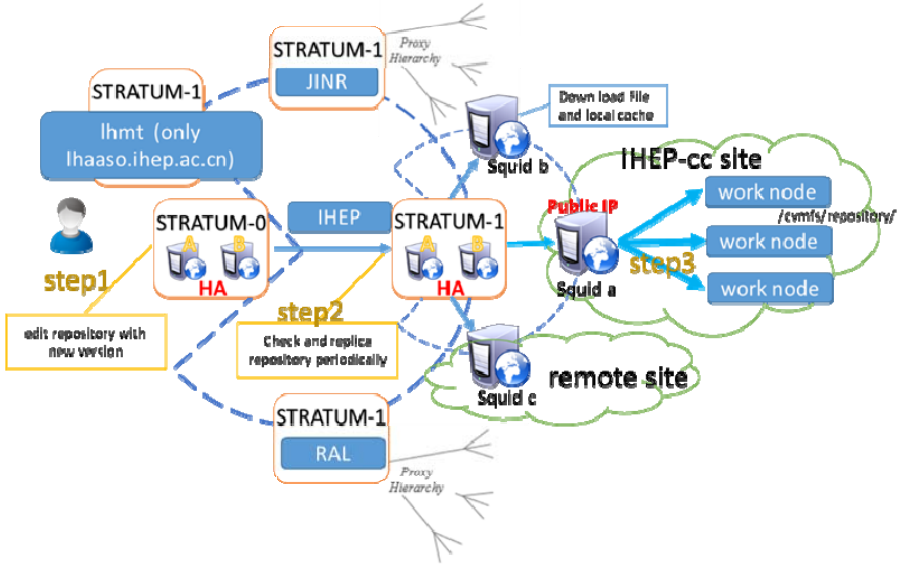
备份系统

- 备份系统对关键的用户目录进行全备份/增量备份
  - 解决用户误操作造成的文件版本丢失问题
  - rm –rf * ☹
  - 注意： /scratchfs 没有备份

应用	目录	备份策略
BES	/ihepbatch/bes	每周一次全备份，每天一次增量备份
/afs/ihep.ac.cn/bes3	每周一次全备份	
YBJ	/ihepbatch/home-ybj	每周一次全备份，每天一次增量备份
/ihepbatch/work	每月一次全备份，每周一次增量备份	
/afs/ihep.ac.cn/soft/YBJ	每月一次全备份，每周一次增量备份	
BSRF	/home/bsrf	每周一次全备份，每天一次增量备份
CMS	/afs/ihep.ac.cn/soft/CMS	每月一次全备份，每周一次增量备份
CC	/home/cc	每周一次全备份，每天一次增量备份
/afs/ihep.ac.cn/soft/common	每月一次全备份，每周一次增量备份	
ATLAS	/afs/ihep.ac.cn/soft/atlas	每周一次全备份，每天一次增量备份
公共目录	/afs/ihep.ac.cn/users	每周一次全备份，每天一次增量备份
/workfs	每周一次全备份，每天一次增量备份	

CVMFS

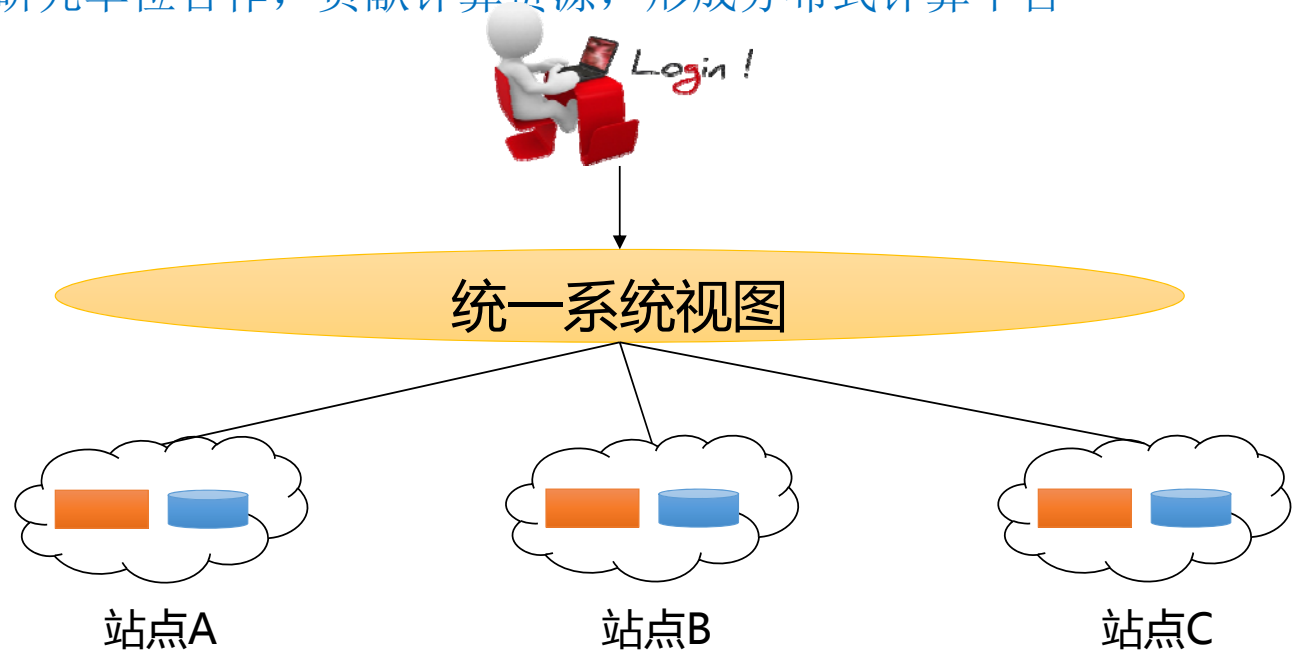
- CernVM-File System简称CVMFS, 提供了一种可扩展的，可靠的，低维护成本的软件分发服务
  - 大部分实验的公共软件库
  - 虚拟机操作系统镜像
  - 老的OS版本
- 使用POSIX接口只读访问
- 底层是HTTP协议
- 两级服务器
  - Stratum0: 中心服务，软件管理员发布软件，设置权限等
  - Stratum1: 广域网上分布的副本服务，可以有任意多个
- 客户端节点上还有本地cache



分布式数据管理技术

分布式数据管理的需求

- 多种计算模式，分布在多个数据中心进行
- 多种探测器及模拟程序，数据在多个地方产生
- 单个实验，全球的高能物理学家共同分享数据，数据分布到全球
- 全球高能物理研究单位合作，贡献计算资源，形成分布式计算平台



分布式数据管理的目标

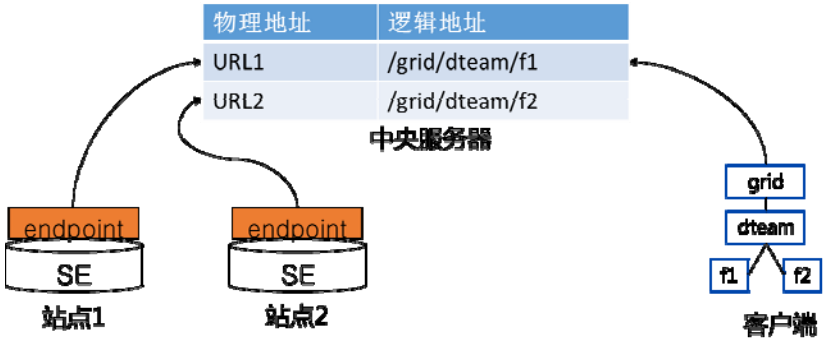
- 将动态变化、异构的、全球分布的存储资源虚拟成一个稳定的、单一的存储系统视图
- 透明性
  - 位置：客户端不需要知道文件存放在哪个站点哪台服务器
  - 迁移：文件可透明在不同的站点之间移动
  - 副本：一个文件可存在多个副本
- 全局命名
  - 唯一标识符
  - 树状的文件名字空间
- 可访问性
  - 通用协议标准，比如Xrootd, Http、WebDAV、S3等
  - POSIX访问接口或者文件系统接口
- 访问性能
  - 延迟管理、性能优化、缓存、安全保证等
- 统一认证和权限控制

- 组件众多
  - 安全，存储，广域网数据传输等
- 每个实验有不同的解决方案
  - Rucio: ATLAS&CMS
  - DMS（Dirac Data management）
    - LHCb, BELLLE II, JUNO, BES III
  - 自研：LHAASO …

统一名字空间

●集中注册

- 类似于文件系统的元数据服务器
- 将站点中的文件注册到中央数据服务器中，生成唯一的标识符
- 分布在全球的站点都可以看到统一的命名空间
- 特点：模型简单，易于实现
- 不足：性能瓶颈；文件不注册，在网格上就看不到



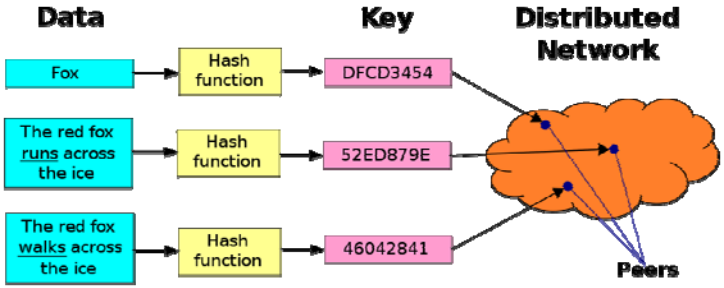
●DHT（ Distributed Hash Table ）内容寻址

- 在不需要服务器的情况下每个客户端负责一个小范围的路由，并负责存储一小部分数据，从而实现整个DHT网络的寻址和存储
- 例如， hash(A) = f(A) % N, 把A这个值映射到了[0~N-1]的范围之中

书名	人名
荒原	小明
最后一天	小红
变形记	小红

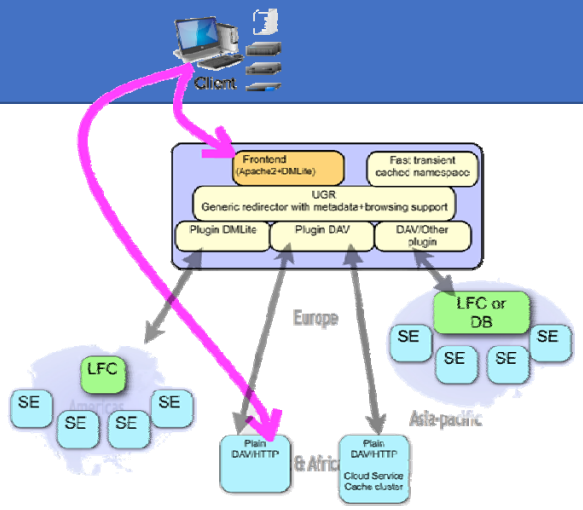
hash(荒原) = 2
hash(最后一天) = 0
hash(变形记) = 1

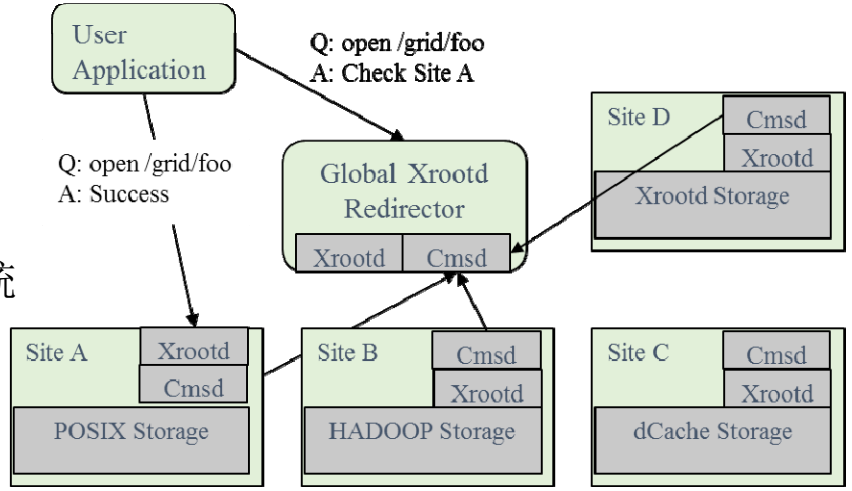
0	小红
1	小红
2	小明



统一名字空间

- 动态聚合（dynafed）
  - 每个站点独立，没有中央服务器
  - 文件不需要注册，即可访问
  - 客户端启动动态聚合服务，指定聚合哪些站点
  - 不完全的全局视图
  - 需要全局协调和发布站点信息
  - 基于GeoIP计算距离
  - 支持WebDav/HTTP/S3等新型存储接口
- 数据联盟
  - 基于Xrootd系统及Redirector功能
  - 数据访问模式的一种替代方案
  - 提供透明的统一命名空间，可以访问多个独立的存储系统
  - 结合XROOTD Cache，还可以做成Cache联盟



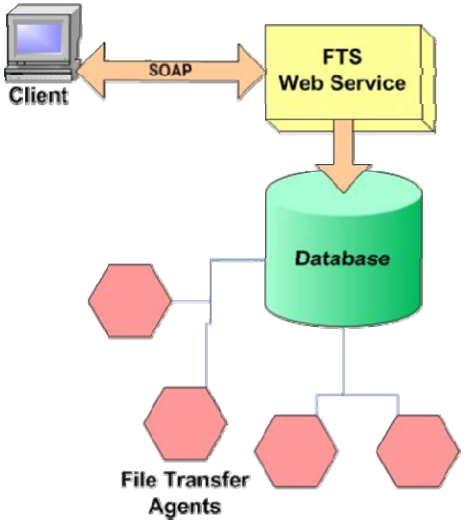


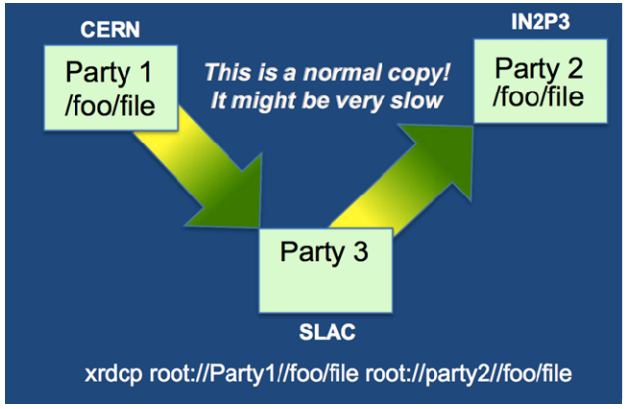
统一的存储访问API

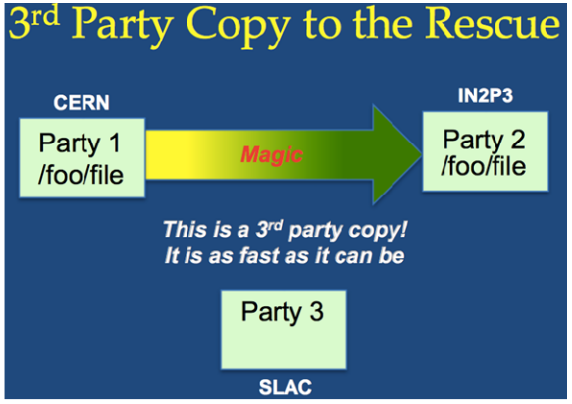
- 封装各站点存储技术的异构性，向用户提供统一透明的接口
  - Lustre, GPFS, EOS, HDFS，CASTOR, HPSS, UNICORE， …
- 将网格数据操作转化成本地存储上的操作
  - 管理本地磁盘，并支持常见的海量存储系统MSS
  - 支持基本的文件传输协议
    - GridFTP，HTTP，FTP，HTTPS，XROOTD等
  - 支持本地IO和远程文件访问协议
    - POSIX及远程文件访问库GFAL（Grid File Access Library）
- 加上了统一访问API代理服务存储系统，在分布式环境中叫做SE
  - Storage Element

数据传输系统

- 代表系统：FTS，SPADE 等
- 两种形式：
  - 通过传输Agent 进行的数据传输
  - TPC（Third Party Copy）
- 高级的大规模数据传输服务
  - 批量传输、消息传递、多源排序、重传…

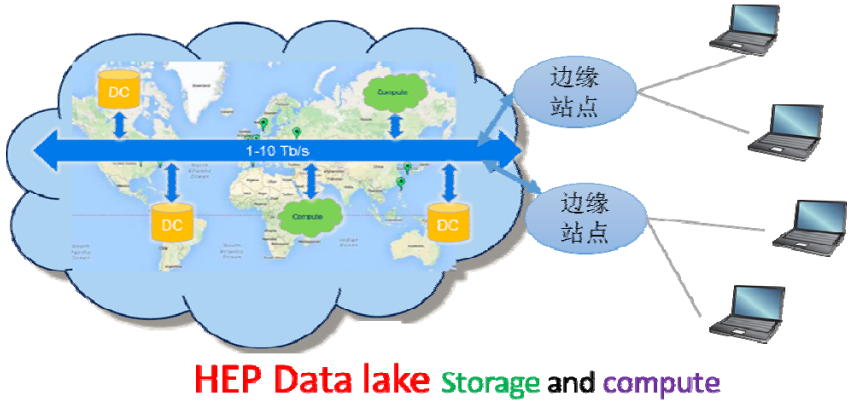






一个中心多个站点

- 大数据中心+边缘计算中心是未来高能物理实验数据处理的趋势
  - 传统的网格层级计算模型，二级站点在维护人力和数据冗余上负担过重
  - 新的模型性价比更高，二级站点只负责计算、只Cache不存储数据
  - 灵活接入超算、云计算等应对突发计算需求的机会计算资源

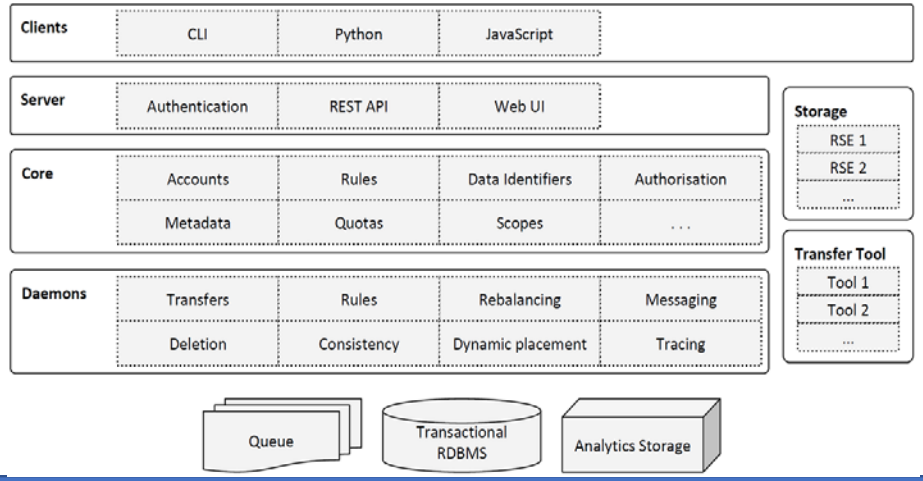


HEP Data lake Storage and compute

- 全局调度服务、数据管理和传输服务
  - 将分布式的计算资源和数据组织成松耦合的、逻辑统一的计算环境

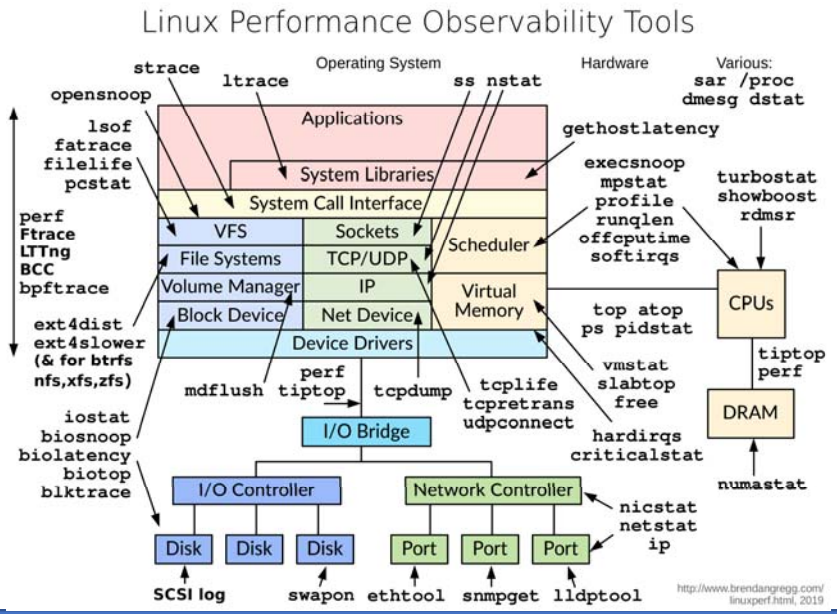
Rucio

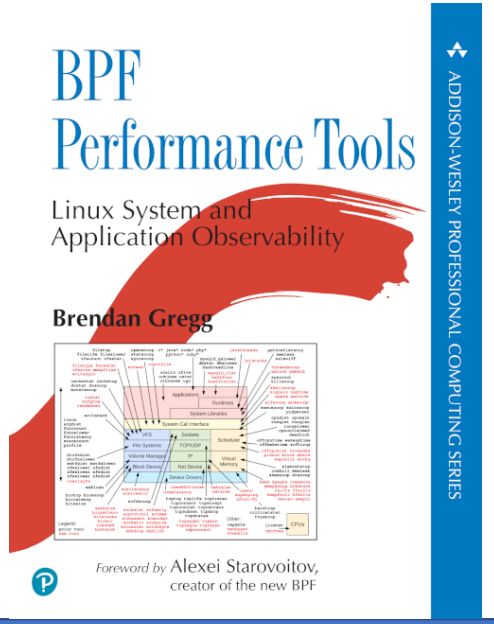
- RUCIO是一套开源的分布式科学数据管理系统，用于组织、管理和访问大规模的实验数据
- 封装了领域最先进的分布式数据管理的概念和方法
- 最初由LHC Altas实验组开发，用来满足ATLAS实验的需求，目前已经扩展到很多其它的实验，甚至天文、生物等领域
  - https://rucio.cern.ch/



更详细的课程及课件

- I/O 性能决定了大部分高能物理计算的整体性能
- I/O 性能的优化需要系统层和软件层的密切配合
  - I/O模式监控 ->报警->调参->测试





更详细的课程及课件

- CERN 暑期学校(CSC)数据技术课程：
  - https://indico.cern.ch/event/769356/contributions/3197035/attachments/1743595/3151809/2019-09-CSC-Alberto-Pace-Data-Technologies.v54_4_UP.pdf
- CERN暑期学校(CSC)数据技术上机：
  - https://apeters.web.cern.ch/apeters/csc2018/_downloads/acd7f44e664e4e45a7ce2d82679edd5f/CSC-DT-2018-Introduction.pdf
  - https://apeters.web.cern.ch/apeters/csc2018/Introduction.html
- CERN 专题计算学校(T-CSC)数据存储部分课程：
  - http://sponce.web.cern.ch/sponce/CSC/slides/
- http://www.brendangregg.com/bpf-performance-tools-book.html

问题反馈渠道

- 官方渠道
  - helpdesk.ihep.ac.cn(推荐）
- 民间渠道
  - QQ群
  - 微信群
  - 时效性比较高，但是问题处理过程没有追溯和记录
- 常见问题
  - 用户手册： http://afsapply.ihep.ac.cn/cchelp/zh/
  - 类似于今天的培训

小结

- 需求
  - 海量存储，单一名字空间
  - 横向性能和容量扩展
  - 数据分布和水位均衡
  - 服务高可用
  - 数据高可靠
  - 数据长期保存和绿色存储
  - 分布式数据管理
  - 跨域数据传输
  - 性能监控和调优

- 技术
  - 分布式文件系统
  - 服务器集群、数据元数据分离
  - 存储分配策略、数据均衡工具
  - HA软件和服务器互备
  - RAID、RAIN、checksum
  - 磁带管理系统
  - Rucio, XrootD 联盟,Xcache …
  - 更详细的课程和课件

Q+A