新型分布式磁盘存储系统EOS在LHAASO实验上的应用

李海波，毕玉江，程耀东
中科院高能所计算中心
2019.7

内容

- EOS磁盘存储系统介绍
- LHAASO实验存储需求
- EOS在LHAASO实验上的建设效果
- EOS面临的挑战与发展
- 小结

EOS是什么？

- EOS： Exabyte scale storage
- 基于XRootD框架实现的磁盘文件存储系统
- 项目始于 2010年5月
  - 时间尚短（AFS 32年， CASTOR 16年， Lustre 16年）
- 主要用于高能物理实验数据存储和分析

http://eos-docs.web.cern.ch

EOS系统核心特性

- 元数据存储于内存设备，访问延迟低
- 提供类POSIX文件访问
  - Xroot，gridFTP，FUSE
- 强认证模式
  - strong (Kerberos5, X509) external clients
  - shared secret (sss) internal clients
- Quota管理
  - 支持用户/组配额
- 回收站机制
- 支持存储节点内文件系统间负载均衡及调度组内节点间负载均衡，提高系统吞吐率及文件访问效率
- GEO基于位置的数据调度

```
[root@eos01 ~]# eos ns
# ------------------------------------------------------------------------------------
# Namespace Statistics
# ------------------------------------------------------------------------------------
ALL      Files                            67575344 [booted] (155s)
ALL      Directories                      1170613
ALL      Total boot time                  161 s
# ------------------------------------------------------------------------------------
ALL      Compactification                 status=off waitstart=0 interval=0 ratio-file=1.3:1 ratio-di
r=18.2:1
# ------------------------------------------------------------------------------------
ALL      Replication                      mode=master-rw state=master-rw master=eos01.ihep.ac.cn conf
igdir=/var/eos/config/eos01.ihep.ac.cn/ config=default mgm:eos02.ihep.ac.cn=down mq:eos02.ihep.ac.cn:
1097=down
# ------------------------------------------------------------------------------------
ALL      File Changelog Size              11.29 GB
ALL      Dir  Changelog Size              880.89 MB
# ------------------------------------------------------------------------------------
ALL      avg. File Entry Size             167 B
ALL      avg. Dir  Entry Size             752 B
# ------------------------------------------------------------------------------------
ALL      files created since boot         29372577
ALL      container created since boot     270561
# ------------------------------------------------------------------------------------
ALL      current file id                  195632337
ALL      current container id             2529925
# ------------------------------------------------------------------------------------
ALL      eosxd caps                       0
ALL      eosxd clients                    1
# ------------------------------------------------------------------------------------
ALL      memory virtual                   71.12 GB
ALL      memory resident                  57.19 GB
ALL      memory share                     11.51 MB
ALL      memory growths                   28.17 GB
ALL      threads                          500
ALL      fds                              2348
ALL      uptime                           5597933
# ------------------------------------------------------------------------------------
ALL      drain info                       id=default, thread_pool_min=40, thread_pool_max=400, thread
_pool_size=40, queue_size=0
# ------------------------------------------------------------------------------------
```

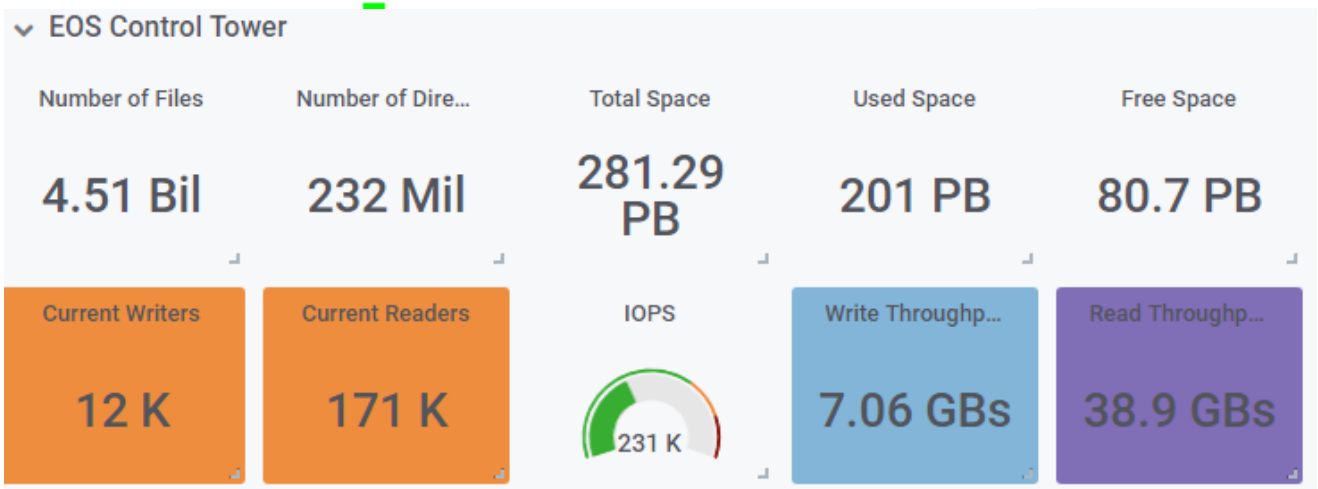
EOS @ CERN

	2017	2018
Nodes	~1200	~1400
Disks	~40000	~50000
Raw capacity	~150PB	~250PB

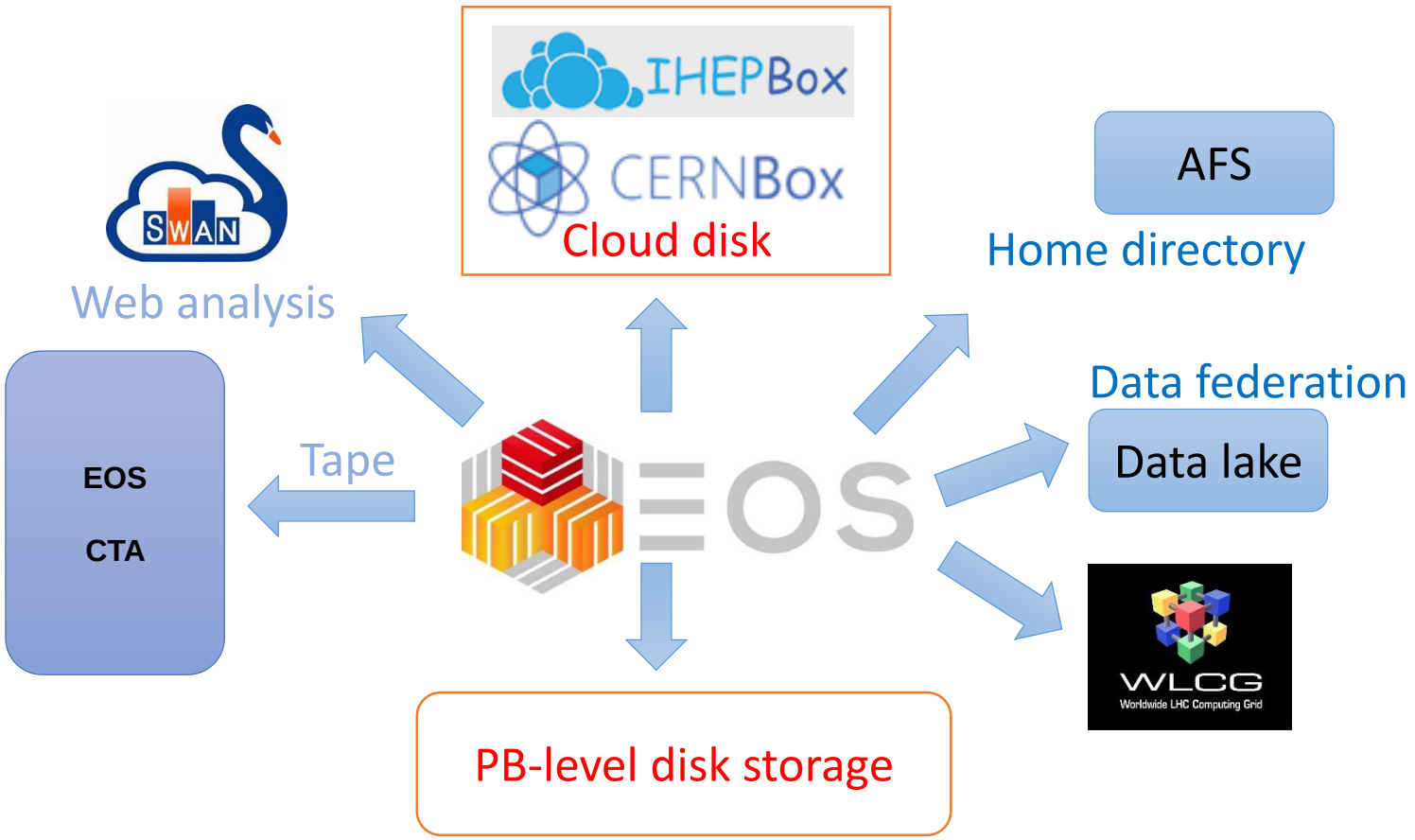
15 EOS instances

- 4 LHC
- 2 CERNBox (new home)
- EOSMEDIA (Foto, Video)
- EOSPUBLIC (non-LHC Experiments)
- EOSBACKUP (backup for CERNBox)
- 6 for various test infrastructures

```
eosams          40P    32P   7.4P   82% /eos/ams
eosexperiment   40P    32P   7.4P   82% /eos/experiment
eosproject     4.5P   1.5P   3.0P   33% /eos/project
eosuser        4.5P   1.5P   3.0P   33% /eos/user
```



EOS应用生态


IHEPBox
CERNBox
Cloud disk
SWAN
Web analysis
AFS
Home directory
Data federation
Data lake
EOS
CTA
Tape
EOS
WLCG
Worldwide LHC Computing Grid
PB-level disk storage

LHAASO实验数据处理需求

- 实验数据经过DAQ获取之后，进入离线计算平台
- 两个数据中心
  - 稻城在站数据中心
    - 提供在站快速重建
    - 计算集群：2500核
    - 磁盘存储容量：300TB
  - 北京高能所数据中心
    - 提供数据存储和处理
    - 计算集群：4500核
    - 磁盘存储容量：4PB
    - 磁带存储容量：20PB

- 存储需求：原始数据传输和存储数据量（压缩后）6 PB/year
- 目前国内产生数据量最大的科学实验装置之一

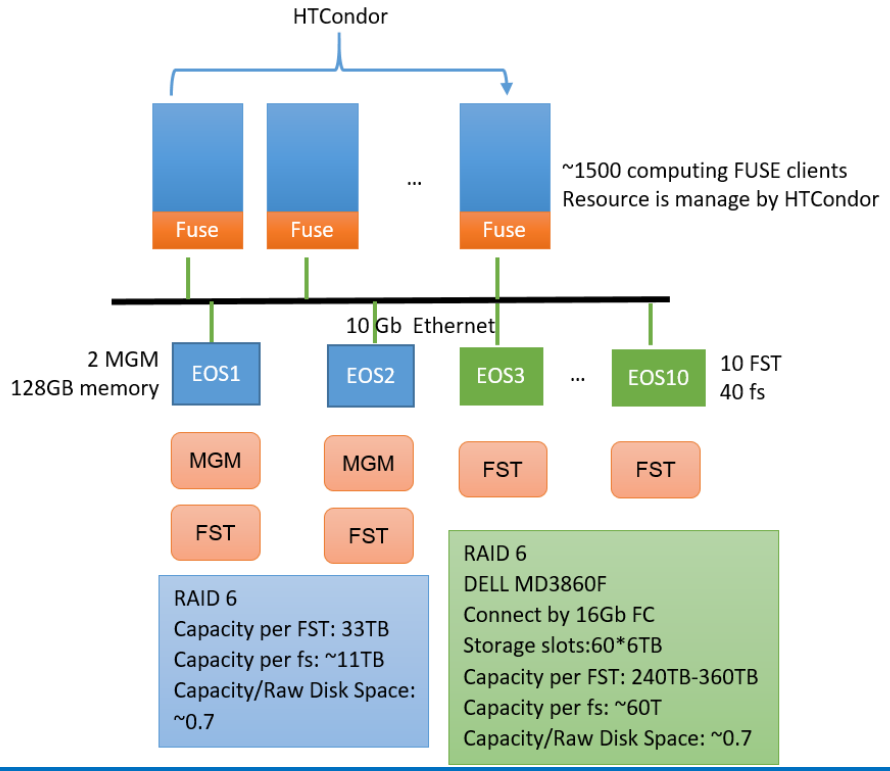
LHAASO实验存储系统

- 个人用户目录
  - /afs，默认500M空间，有备份
- 数据存储
  - /eos
    - 当前LHAASO实验主要数据存储盘，每人默认1TB空间，25万个文件数
  - /scratchfs
    - 临时文件存放，每人默认500GB空间，20万个文件数
  - /workfs
    - 可临时充当软件文件存储，每人默认5GB空间，50000个文件数，有备份
    - 不能作为作业提交目录
- 软件存储
  - /afs
  - /cvmfs

LHAASO实验EOS存储建设情况

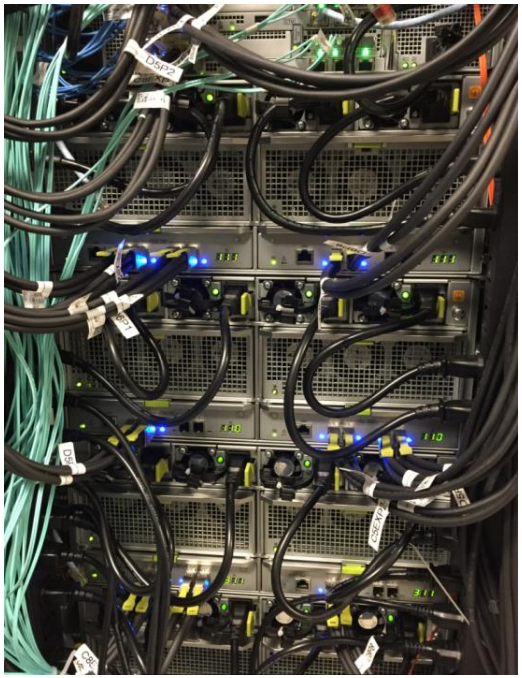
- 目前有两个实例
  - 稻城集群
    - 提供山上在线数据处理
    - 2台存储服务器
    - 共152.83TB空间
    - 即将扩容560TB空间
  - 本地集群
    - 提供本地离线数据处理
    - 10台存储服务器
    - 共2.39PB空间
    - 当前存储文件数：7053万
    - 当前存储目录数：120万
    - 聚合带宽达到8GB/s，跑满带宽

实例名	挂载点	总空间	使用场景	备注
Lhaaso本地集群	/eos	2.39PB	本地离线数据处理	本地登录节点可访问
Lhaaso稻城	/eos/daocheng	152.83 TB	山上在线数据处理	山上登录节点可访问



EOS存储硬件环境

- 存储服务器
  - 机架服务器
  - 10 Gbit/s网络接口
- 存储阵列
  - DELL ME3860F磁盘阵列
    - DDP动态磁盘池
    - 60*6TB 磁盘

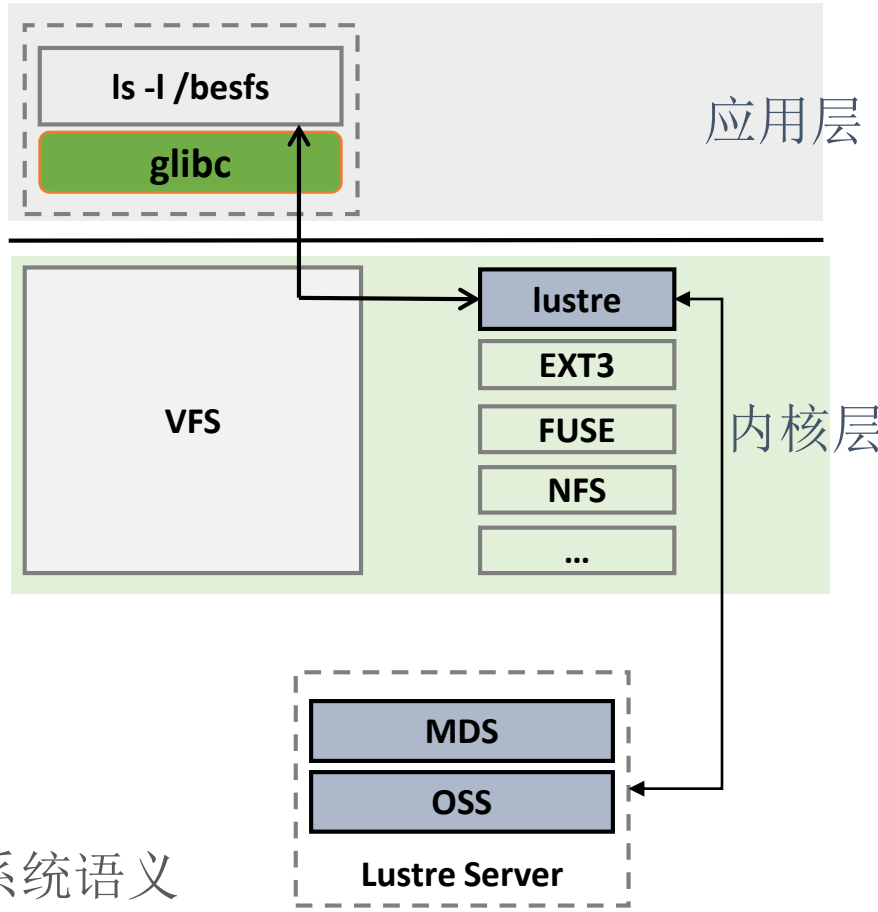


EOS使用方式

- FUSE方式（用户空间文件系统）
  - 本地挂载，比如LHAASO实验的挂载目录为/eos
  - 在登录节点和计算节点，可以通过访问/eos/访问数据
- Xrootd方式
  - 不需要本地挂载
  - 使用xrootd协议访问数据
  - Xrdfs root://eos01.ihep.ac.cn ls /eos/user/file.txt

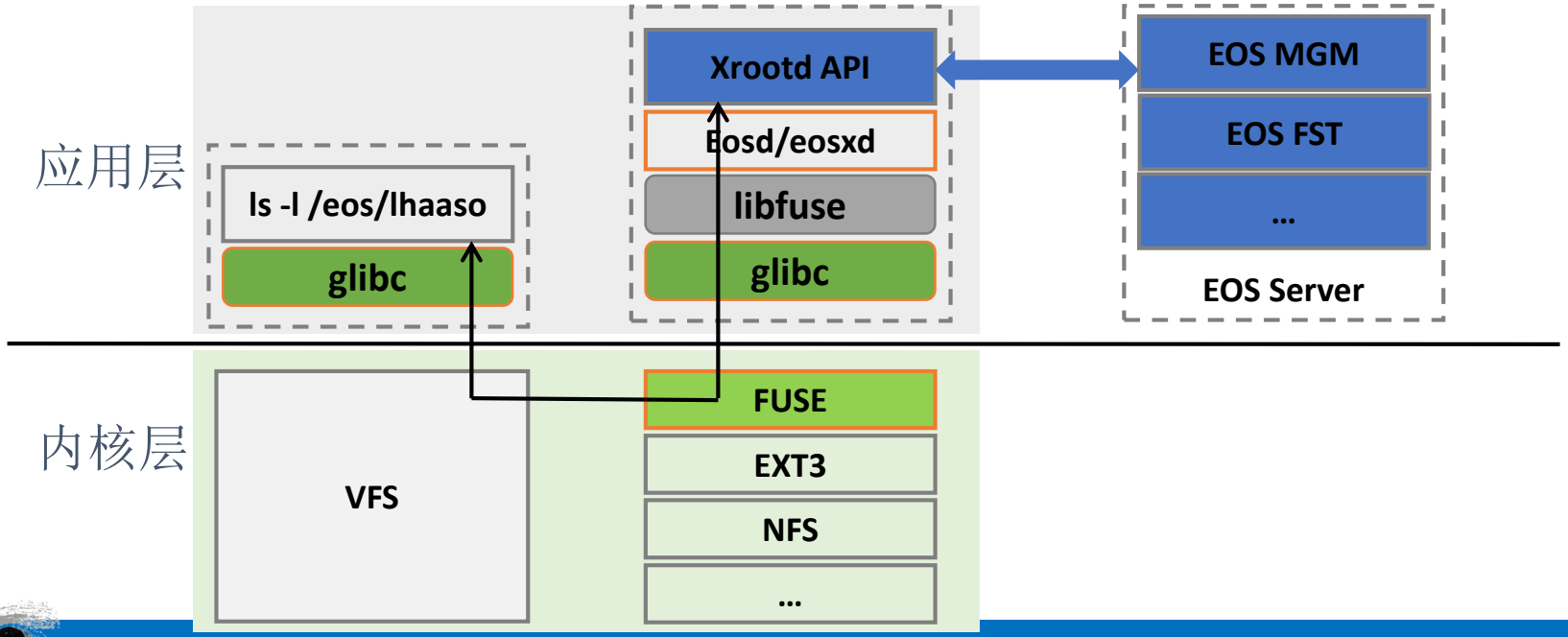
数据访问协议

- 内核级文件系统：Lustre
  - 并行性好，文件系统语义支持好
  - 内核依赖，管理复杂
- 应用级文件系统：EOS Fuse
  - 提供文件系统语义
  - 并行性支持差
  - 目前稳定性差
- 应用级数据访问：Xrootd
  - 基于文件访问API，不提供文件系统语义
  - 稳定性好，不受文件系统限制



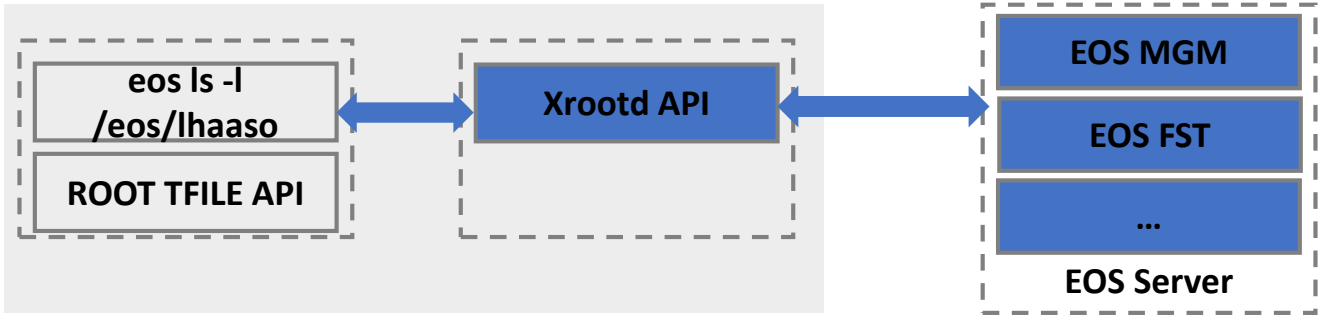
EOS文件系统

- 在XROOTD基础上开发文件系统接口，类似于本地文件系统
- 对于物理分析来说并不高效，但是比较灵活
- 基于FUSE（ Filesystem in Userspace ）实现，FUSE是Linux内核标准模块
  - 比内核级文件系统（eg. Lustre）实现简单
  - 但是并不是最高效的
  - 任何FUSE模块或者eosd的失败都会导致作业的失败



应用层访问接口

- 命令行方式，比如eos ls，直接调用xrootd API来访问EOS服务器，绕过任何内核模块
- 调用ROOT TFILE类的应用软件，也可以直接调用xrootd API
- 这种方式完全工作在应用层，不受文件系统及内核的影响，稳定好
- 用户使用不太灵活，没有本地文件系统的接口，cat等命令无法工作




```
[lihaibo@lxslc601 ~]$ eos ls -lh /eos/lhaaso
drwxr-xr-+   1 lhaasore lhaaso       122.68 M Mar  4 15:10 cal
drwxr-xr-+   1 lhaasore lhaaso       111.05 T Jul  2 09:16 decode
drwxr-xr-+   1 root     root         162.33 T Jun 20 14:11 experiment
drwxr-xr-+   1 lhaasore lhaaso        58.47 G Mar  4 15:11 monitor
drwxr-xr-+   1 root     root         213.66 T Jun  4 11:17 raw
drwxr-xr-+   1 lhaasore lhaaso        14.26 T May  5 08:47 rec
drwxr-xr-+   1 root     root         70.35 T Aug 29  2017 simulation
```

Xrootd使用

- 首先，物理软件（比如BOSS或者SNiPER ）调用ROOT库 File:: Open，比如：

  TFile* inputFiles[m_fileNum] = TFile::Open(m_fileNames[m_fileNum].c_str(),"READ");

  - 注意：以下两种调用方式不支持

  （1）简单声明：TFile file(fn.c_str());

  （2）New方法： TFile* inputFile = new TFile(m.c_str(),"READ");

- 其次，将输入输出文件采用ROOT的命令方式，比如：

root://eos01.ihep.ac.cn///eos/user/l/lihaibo/lhaasotest/DAT000011.root

协议	服务器名	绝对路径

- 由于没有本地文件系统接口，脚本中不能出现通配符，不能采用相对路径，不能使用操作系统命令来遍历目录，比如for f in `ls /eos/lhaaso/raw/wcda` 之类的语句

使用ROOT访问eos文件

- ROOT中操作过程
  - TFile打开文件
    - $root [0] TFile *file0 =

      TFile::Open("root://eos01.ihep.ac.cn//eos/user/l/lihaibo/xrootd_test/DAT.root")
    - 或 $root -l root://eos01.ihep.ac.cn//eos/user/l/lihaibo/xrootd_test/DAT.root
  - $root[1].ls 查看文件信息
  - $root[2] t_runh->Show(0) 显示第0个event的信息
  - $root[3]t_runh->GetEntries() 总事例数
  - $root[4]t_runh->Scan()
  - $root[5]t_runh->Print()
- 可以在lxslc登录节点使用xrootd访问稻城的数据

```
[lihaibo@lxslc601 ~]$  root -l root://lhmteos01.lhaaso.ihep.ac.cn//eos/daocheng/user/l/lihaibo/tV6_nfit100-200.root
root [0]
Attaching file root://lhmteos01.lhaaso.ihep.ac.cn//eos/daocheng/user/l/lihaibo/tV6_nfit100-200.root as _file0...
(TFile *) 0x2674360
root [1]
```

EOS 异地副本

- LHAASO实验数据处理是多中心架构，对数据异地访问有很强的需求
- EOS异地副本策略，可提供基于GEO位置标签的访问策略
- EOS异地副本测试效果
  - 测试方法：在北京、稻城各搭建1台EOS服务器形成集群，设置不同GEO Tag，设置同一目录的双副本，分别进行10次文件写入和读出测试
  - 结论：基于 GeoTag 的调度暂时不会完全匹配相同的 tag，即使设定了强制匹配规则，仍需优化

location	Replica	Max	Min	Mean
Beijing	beijing+daocheng	8.0	6.8	7.5
Daocheng	beijing+daocheng	8.6	7.5	8.0

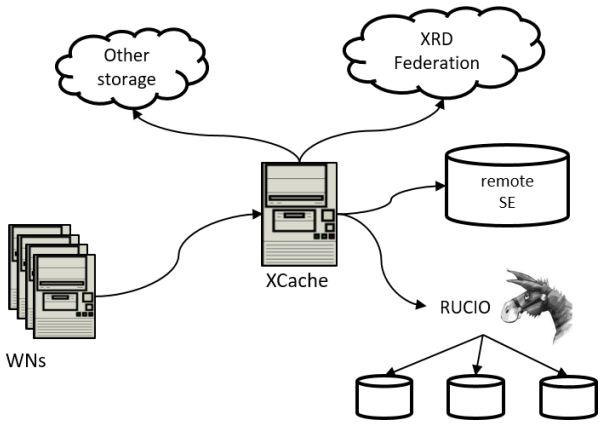
文件写入（MB/s）

location	Replica	target	Max	Min	Mean
Beijing	beijing+daocheng	beijing	113.8	111.3	112.55
Daocheng	beijing+daocheng	daocheng	232.7	146.3	166.11

文件读出（MB/s）

EOS+Xcache

- 异地数据访问时，对于对于在本地服务器上没有的数据，如何提高访问速度？
- Xcache技术
  - 一种优化数据异地访问技术
  - 使用Squid缓存，基于Xrootd协议
- 验证xcache方案
  - 测试方案：在北京、稻城两地搭建EOS集群，在稻城配置xcache proxy，对不同大小的文件进行10次读测试
  - 结论：大文件会缓存到硬盘上，小文件同时会缓存到内存中

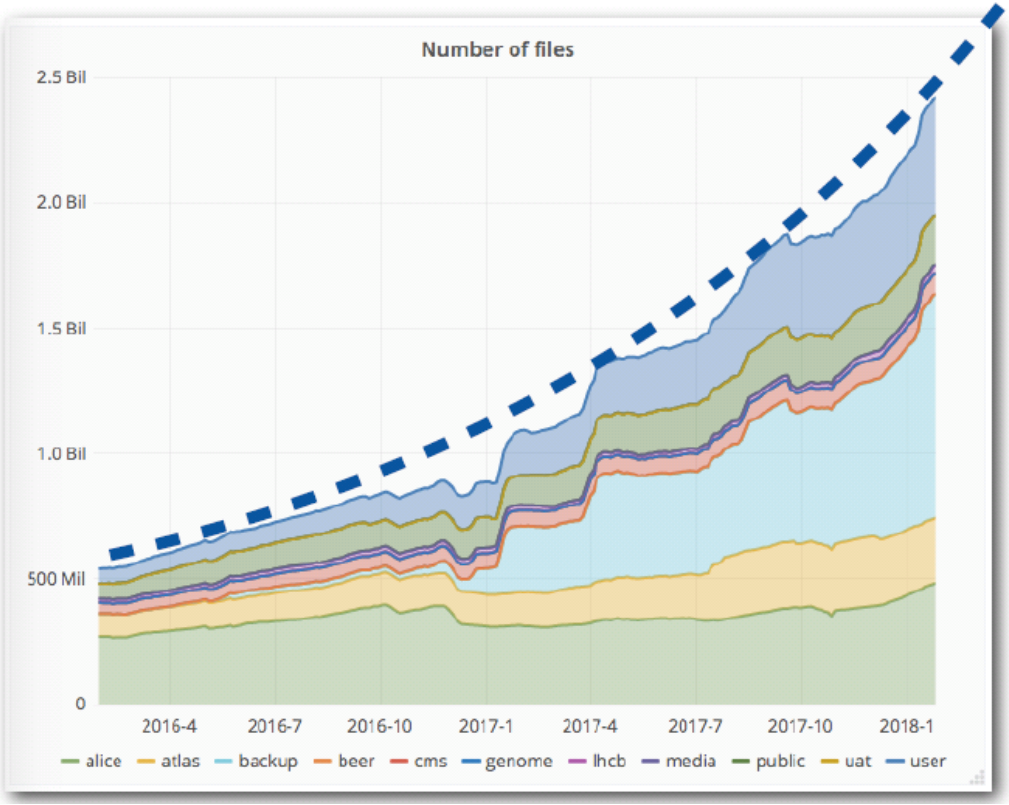


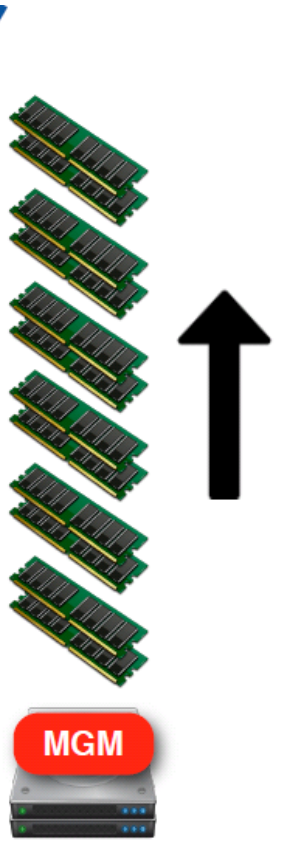
size	mode	1st Max	1st Min	1st Mean	2nd Max	2nd Min	2nd Mean
500M	Disk Cache	11.38	10.24	10.95	512	512	512
5G	Disk Cache	5.095	5.095	5.095	341.3	301.2	320

文件读出（MB/s）

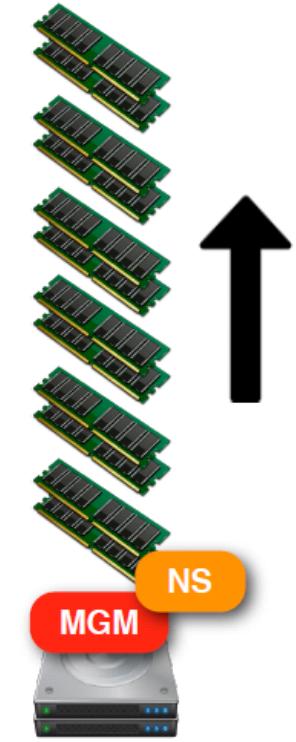
EOS元数据面临的挑战

- EOS 元数据面临的挑战
  - 从scale-up 到scale-out

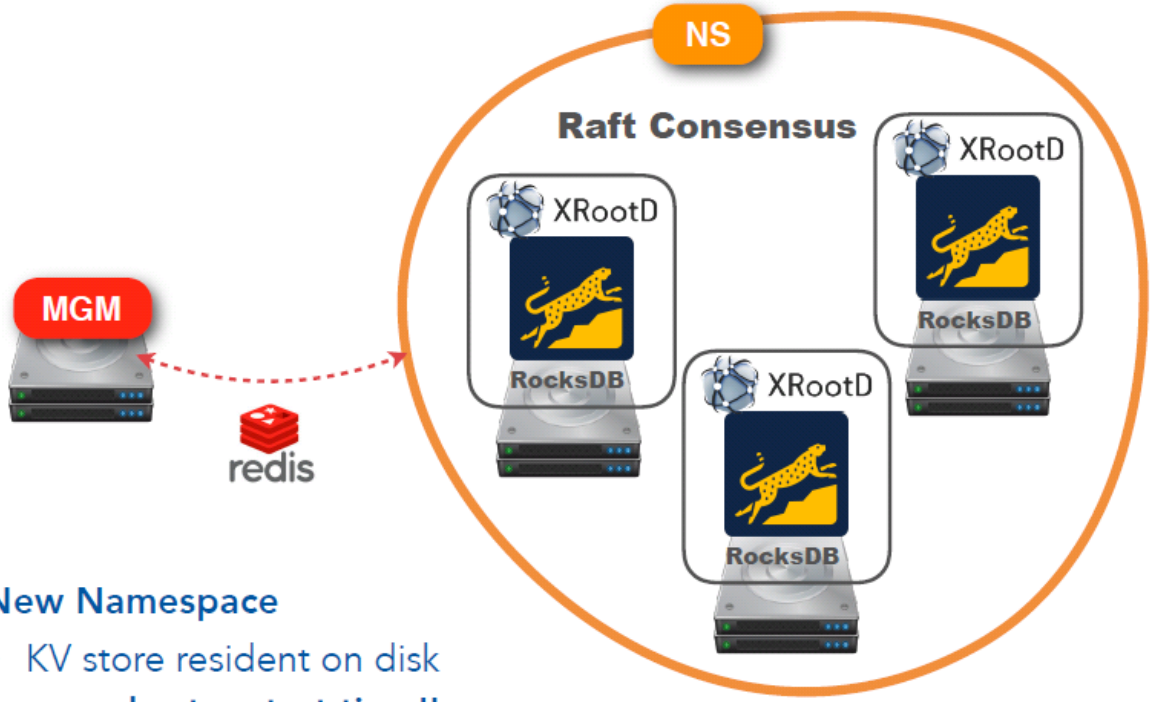




EOS元数据面临的挑战



namespace



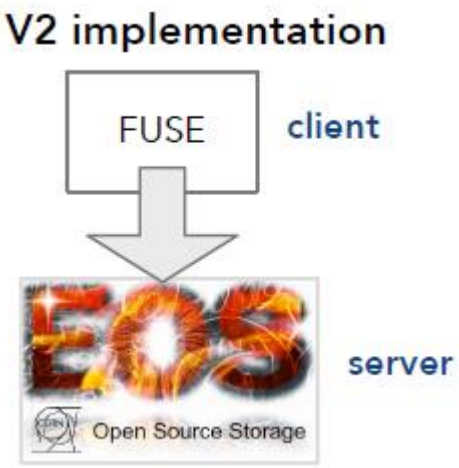
namespace

New Namespace

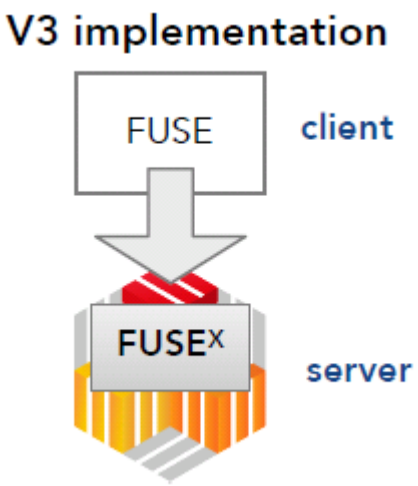
- KV store resident on disk
- **very short restart time!!**
  not based on #files and #dirs
- namespace caching in MGM memory

EOS fuse面临的挑战

- FUSE方式中，任何FUSE模块或eosd的失败都会导致作业的失败



Fuse文件系统实现方式为纯客户端应用程序

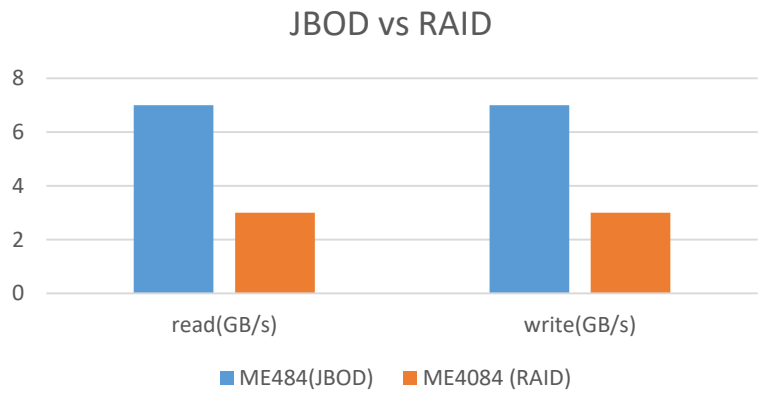


专用的服务器端支持提供完全异步的服务器->客户机通信、锁、文件内联、本地元数据和数据缓存

下一步计划：EOS+JBOD

- 需求
  - 当前存储使用RAID阵列，无法充分利用阵列的性能，将来可能成为瓶颈
  - 使用JBOD代替RAID阵列可以提供更好的性能
- RAID vs JBOD
  - 同等可用存储容量下，费用相当
  - JBOD的IO路径比RAID短，理论上会有提升
- 设备选型
  - DELL ME484 JBOD
- ME484测试
  - 70块盘下，FIO测试读写性能最高达到7GB/s
  - 初步测试显示速度提升2倍左右
  - 需进一步与应用结合测试

	RAID	JBOD
Raid controller	yes	no
Configuration	RAID5 or RAID6	Replica
Disk usage	(67%) RAID6	Replica (50%)
Price	Expensive	Cheap



小结

- EOS是一个开源的分布式文件系统，可提供几十PB的单实例存储能力，能够满足LHAASO实验存储需求
- 加大推广使用xrootd方式提交作业
- EOS可扩展性好，可使用低成本的JBOD磁盘作为存储介质
- EOS在高能物理领域正在形成完整的生态
- EOS还在不断演进发展

谢谢！