高能物理大规模数据管理

高能所计算中心 程耀东

chyd@ihep.ac.cn

2019-5-30

数据量快速增加

- BEPCII/BESIII：~1PB/年
- 中微子实验
  - DYB：数百TB/年
  - JUNO，2020年运行，2PB/年
- 宇宙线实验
  - LHAASO，2018年开始取数，目前3TB/day
  - 2021年开始：6PB/年
- 空间天文实验, 数百TB/年
- LHC：50PB/年，传输到高能所3-5PB/年
  - ATLAS, CMS, LHCb
- IHEPCC现有资源：~20000CPU cores, ~15PB disk storage, ~8PB tape
- 很快将达到百PB规模，如何高效管理科学数据

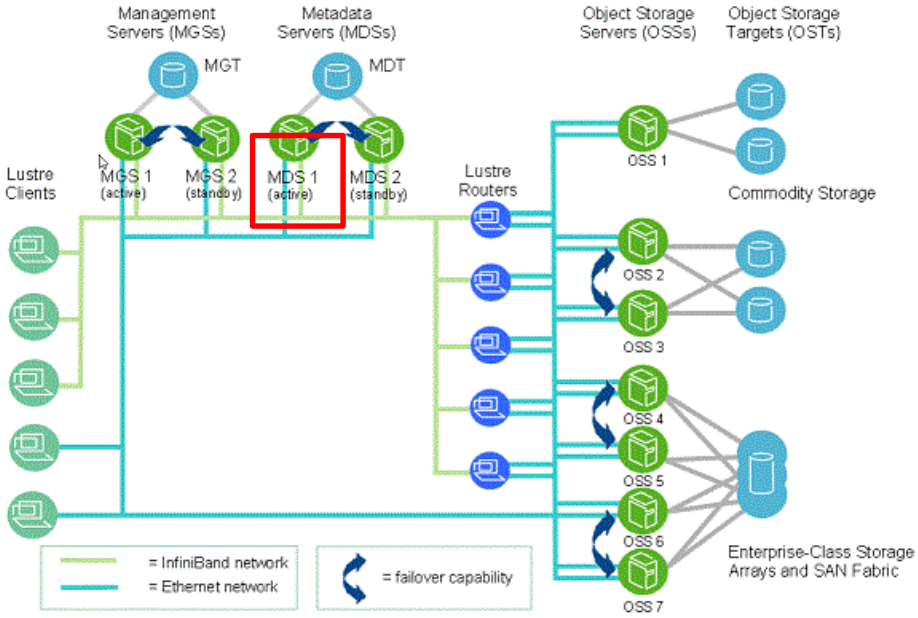
- BESIII, dayabay, HXMT, CSNS
- JUNO, LHAASO, AliCPT, GECAM, HEPS, …
- eXTP, HERD, CEPC, …

数据存储系统

- 开源并行文件系统
  - Lustre, Ceph, Hadoop, …
- 商业文件系统
  - GPFS, …
- 高能物理领域开源的系统
  - CERN EOS
  - dCache
- 面向对象的存储
  - S3, CEPH, Azure, aliyun, …
  - RUCIO

Lustre

- 高性能并行文件系统，内核级实现，提供完善的文件系统语义
- 在超级计算机领域得到非常广泛的应用
- 国际高能物理领域：GSI 14PB
- 高能所应用
  - BESIII, dayabay, juno, cepc
  - hpcfs, publicfs, sharefs, workfs, …
  - **>15PB**
- 面临的问题
  - 元数据服务器基于EXT2本地文件系统，文件数量及访问性能受到限制
  - 需要依赖特定的内核，硬件驱动兼容性等
  - 多个mount点，人工做均衡（比如：链接），管理复杂

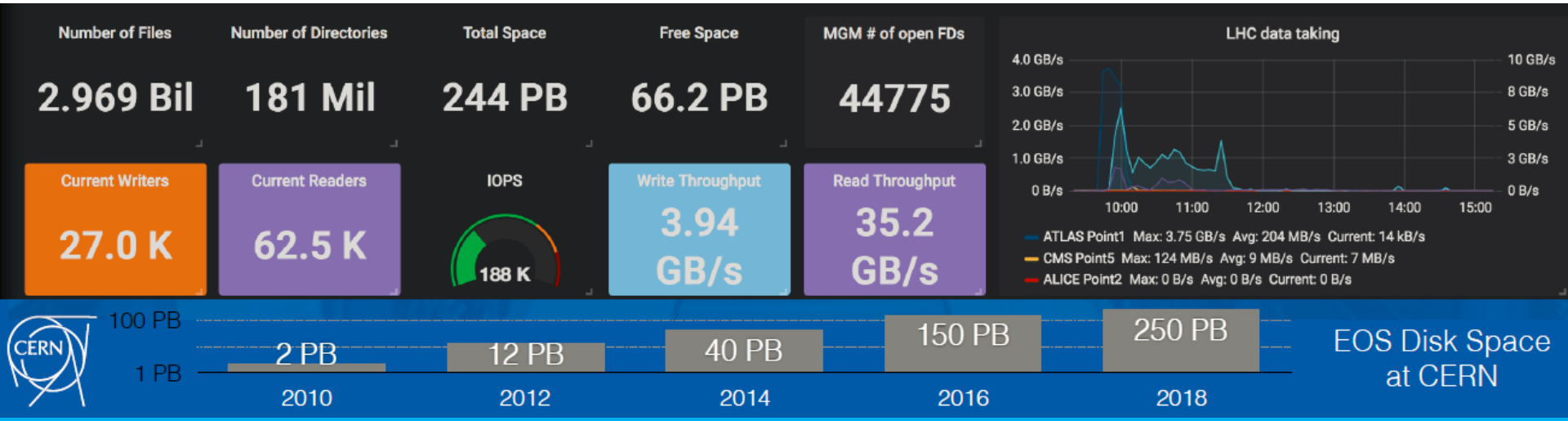


CERN EOS

- CERN开发分布式文件系统，已经管理超过200PB的磁盘
- 全部服务都基于Xrootd，架构简单
- 采用本地内存或者内存集群数据库存放元数据，响应快，可扩展性好

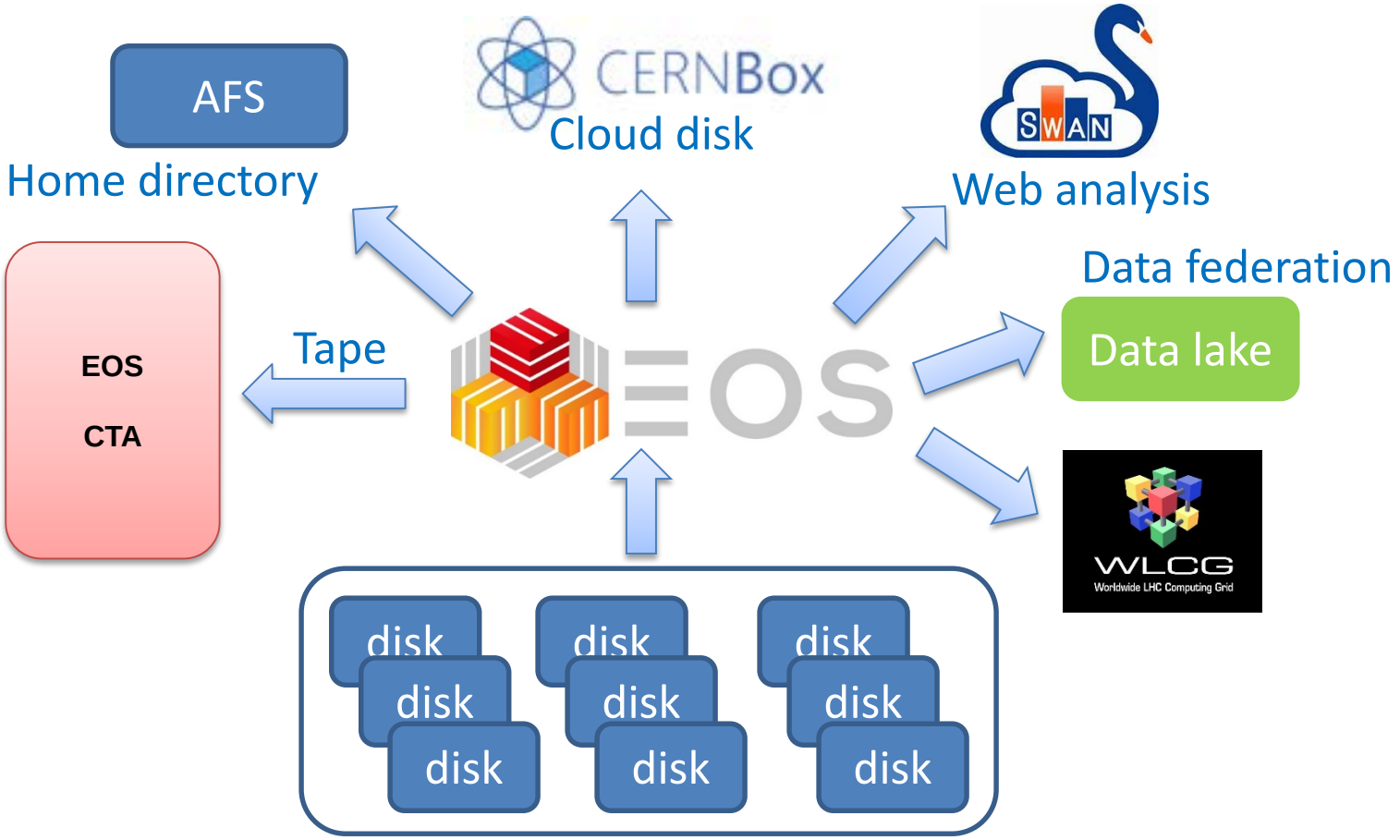
4.7P	1.8P	2.9P	38%	/eos/project
4.7P	1.8P	2.9P	38%	/eos/user
42P	32P	9.8P	77%	/eos/cms
39P	31P	7.5P	81%	/eos/experiment
39P	31P	7.5P	81%	/eos/ams
18P	14P	3.6P	80%	/eos/lhcb

Disk server	~1400
Hard Disks	~60K
Raw Capacity	~270PB
Number of files	~3.8Billion
Streams	~55K
Peak throughput	>100GB/s
File loss rate	~O(10^-6)



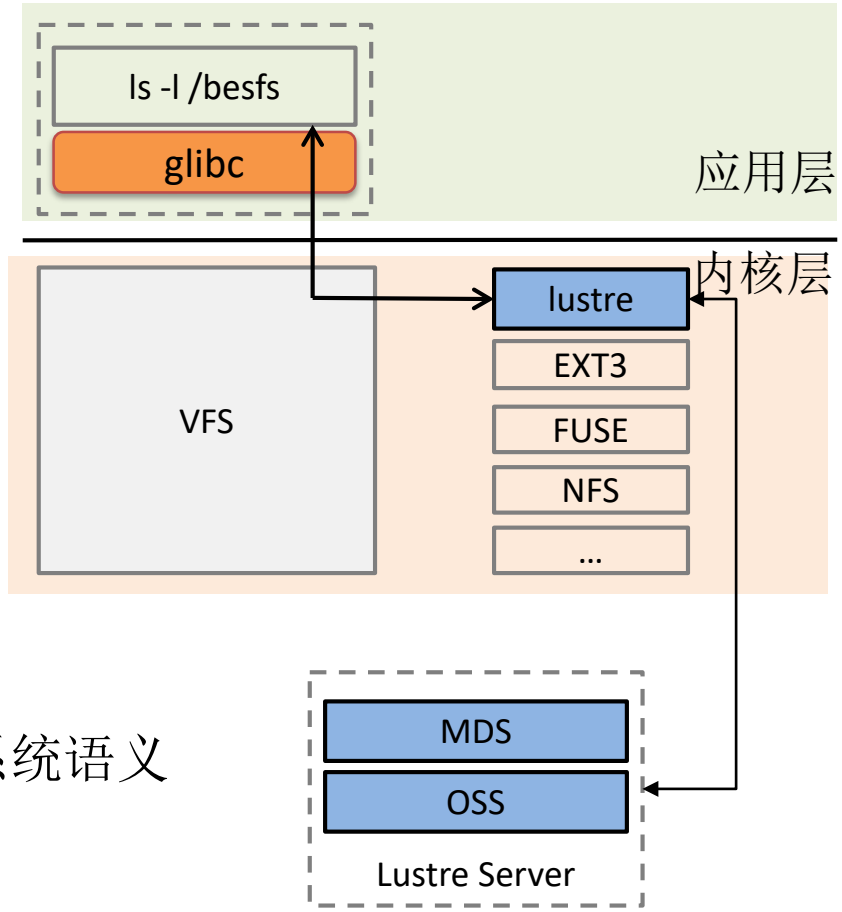
EOS Ecosystem

EOS has become an ecosystem in high energy physics



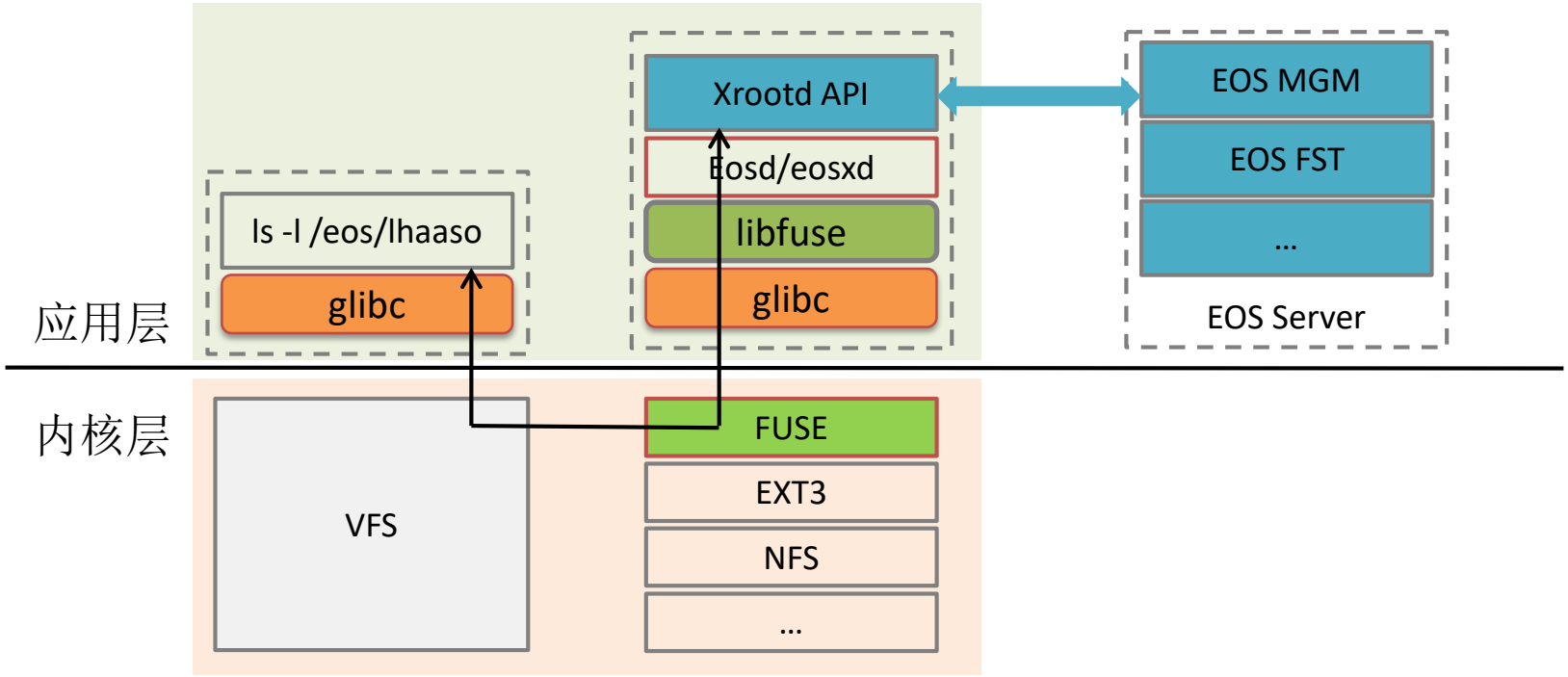
数据访问协议

- 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



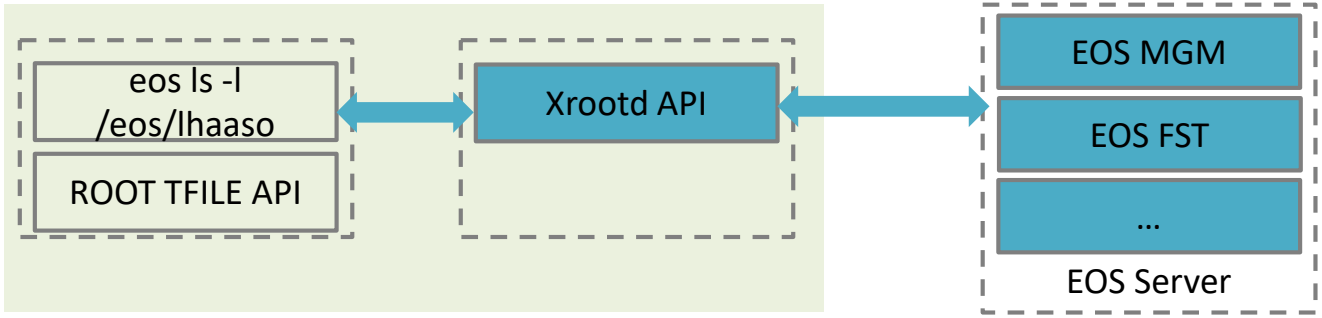
EOS文件系统

- 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




```
[chyd@lxslc602 ~]$ eos ls -l /eos/lhaaso
drwxr-xr-+   1 lhaasore lhaaso       85464334 Mar  4 15:10 cal
drwxr-xr-+   1 lhaasore lhaaso   27513723502514 Mar  4 15:07 decode
drwxr-xr-+   1 root     root     1625551928996 82 Feb 22 16:24 experiment
drwxr-xr-+   1 lhaasore lhaaso    27540788823 Mar  4 15:11 monitor
drwxr-xr-+   1 root     root     88469667087277 May 15 09:07 raw
drwxr-xr-+   1 lhaasore lhaaso   3269125778620 May  5 08:47 rec
drwxr-xr-+   1 root     root     70345464229928 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/c/chyd/703/mc/KKpi/phsp2/KKpi_phsp_0001_boss703.rtraw

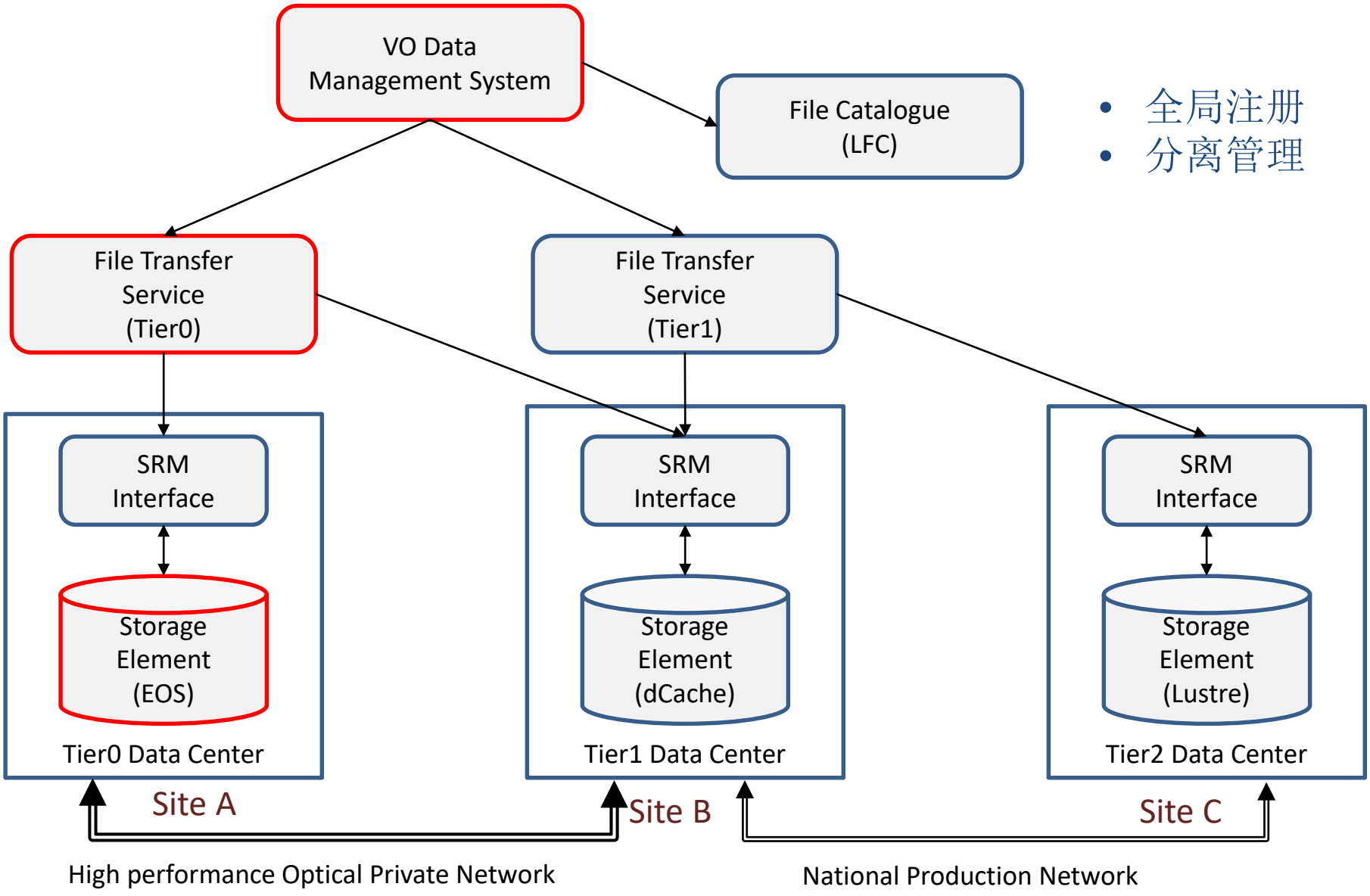
协议	服务器名	绝对路径

- 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

广域网数据管理

- 网格数据管理
- 数据联盟
- 数据湖

网格数据管理



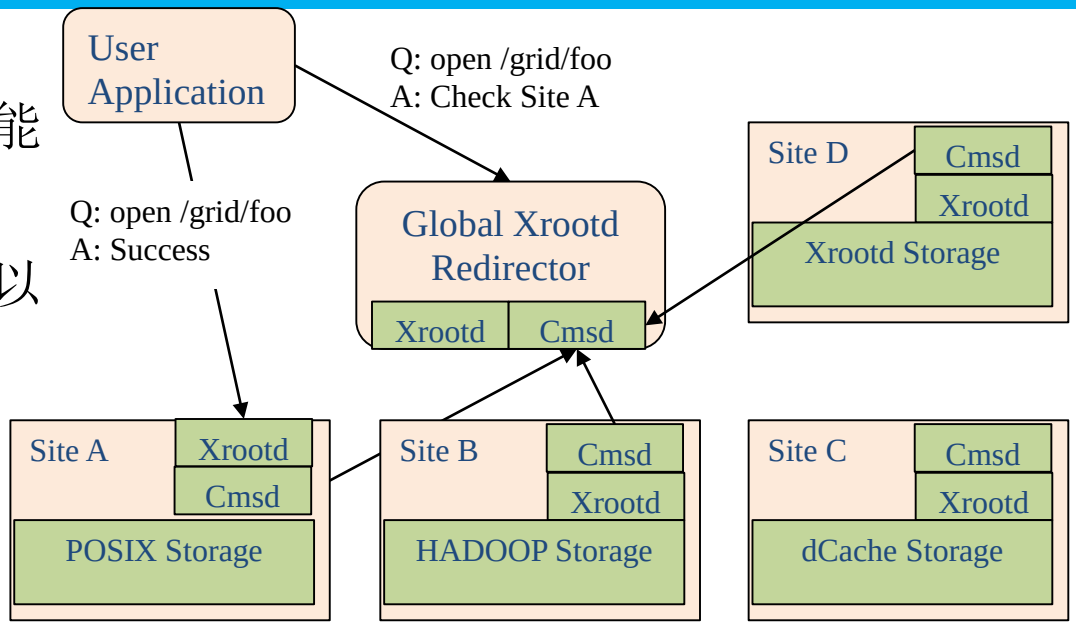
- 全局注册
- 分离管理

新的需求和模式

- 高能物理数据处理的基本单元是“事例（Event）”
- 多个事例能够“独立”处理，具有“天然并行性”
  - 比如：需要在10个站点模拟10亿个事例，那么每个站点模拟1亿
- 但是，今天数据处理系统的基本单位是“作业”和“文件”
  - 难以管理和平衡全局系统资源（网络、存储、CPU、IO等）
  - 那么未来的模式：面向事例的处理和面向对象的存储吗？
- 今天基本的工具：批处理和存储单位（SE）
  - 高成本
  - 未来：数据联盟或者数据湖吗？

数据联盟

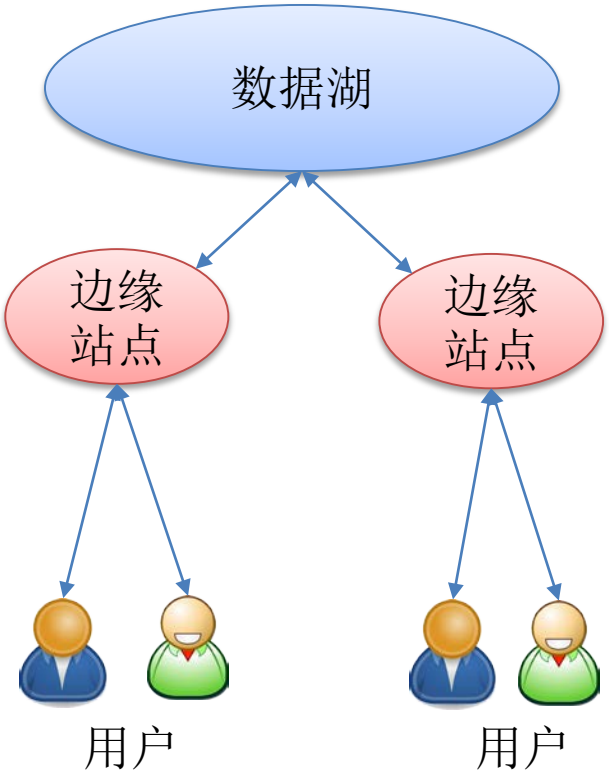
- 基于Xrootd系统及Redirector功能
- 数据访问模式的一种替代方案
- 提供透明的统一命名空间，可以访问多个、独立的存储系统
- 不需要复杂的FTS和SE



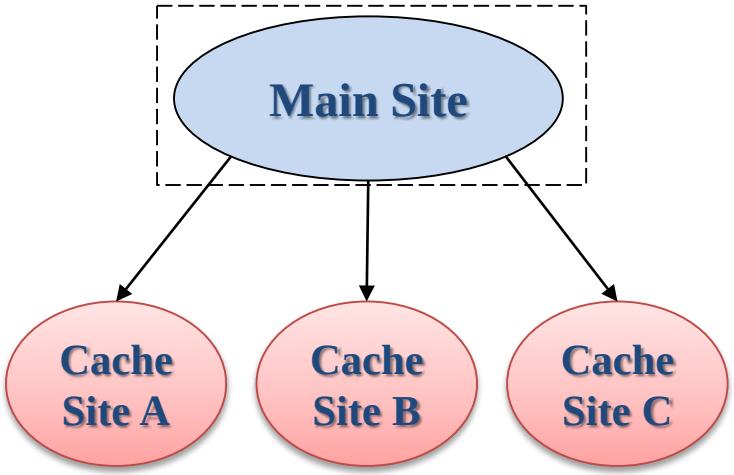
- 但是，这种方法缺乏一个权威的机构来决定谁可以加入联盟以及加入的内容
- 缺乏全局的系统，来管理跨站点的数据移动和复制
- 更适合具有密切关系的“私有”联盟

数据湖

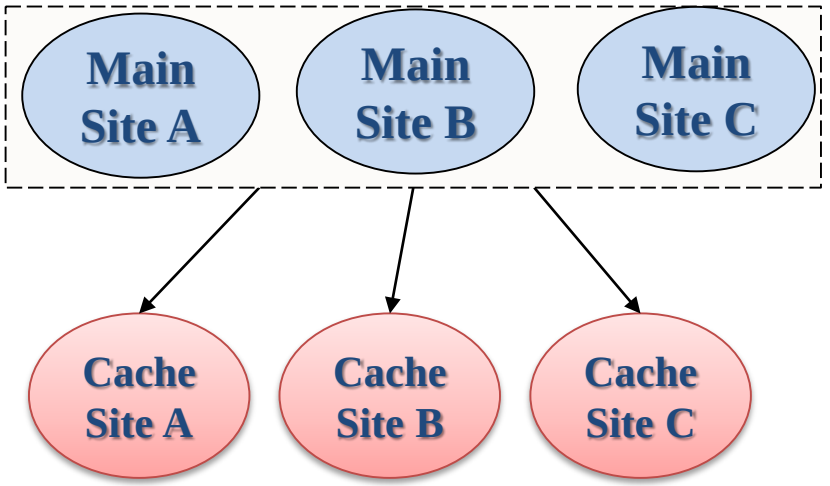
- 不同于计算机领域内的“数据湖”（多源异构数据的统一存储）
- 不同于传统的网格存储（一个站点配置一个或多个SE），数据湖有一个单一的逻辑SE，具有足够大的存储容量和访问性能
- 在数据湖之外的站点没有持久的存储
  - 例如：cached or streamed.
- 非实验的私有数据直接到用户所在的边缘站点
  - 不管这个站点在数据湖之内或者之外



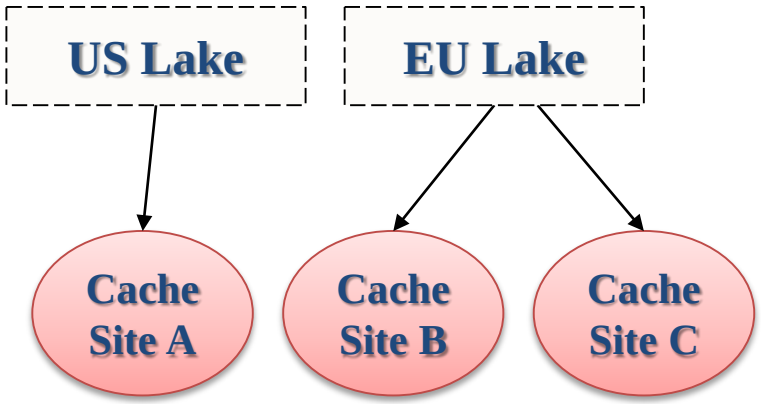
数据湖的一些实现模型



1. 所有数据都存储在单一的大型站点

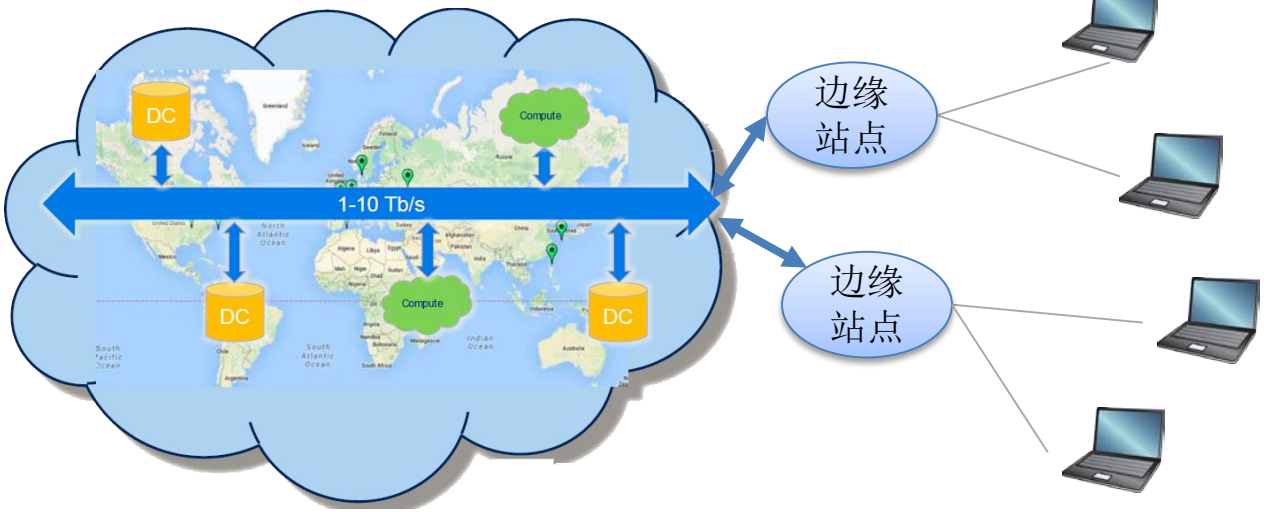


2. 几个大型站点联合组成一个数据湖



3. 有些大型实验，比如HL-LHC具有少量的数据湖

未来的高能物理数据处理平台

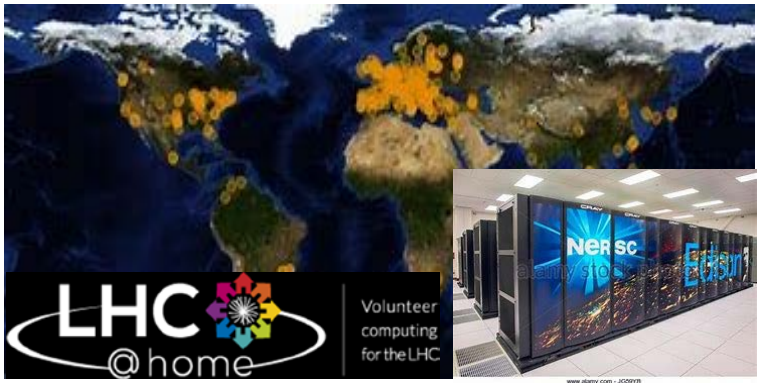


基于WEB/Cloud
的数据分析

SWAN

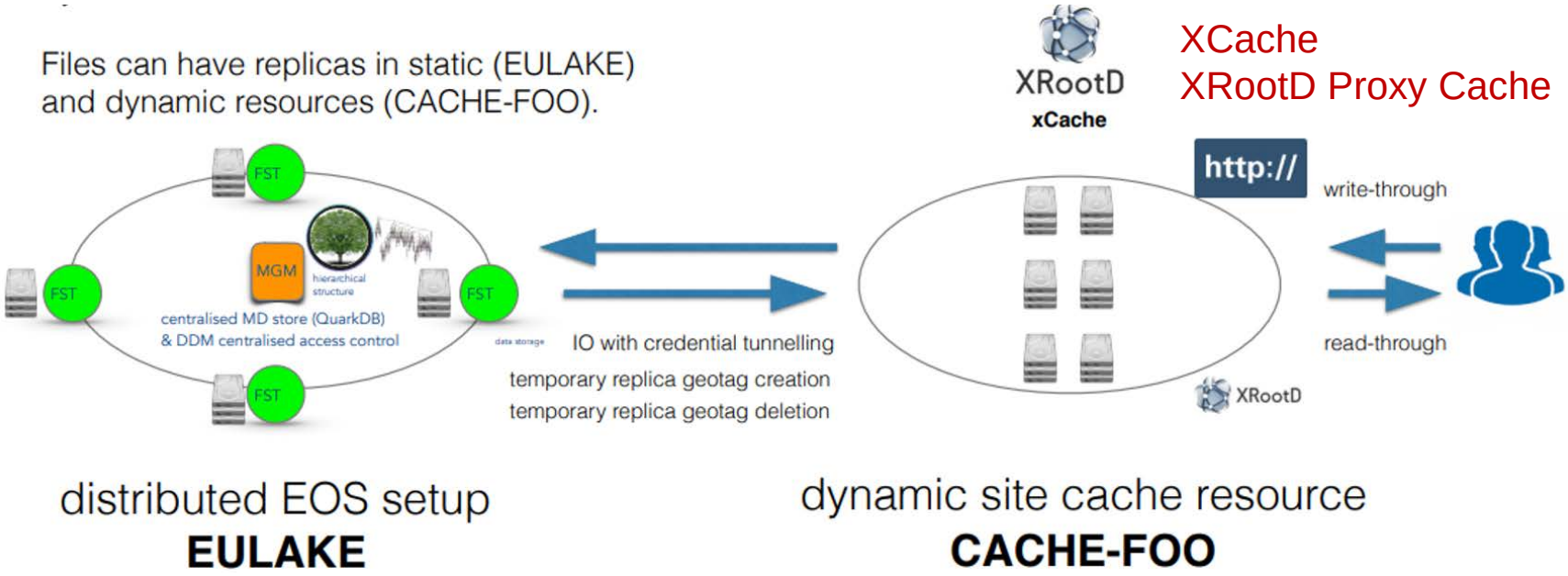
HEP Data lake

Storage and compute



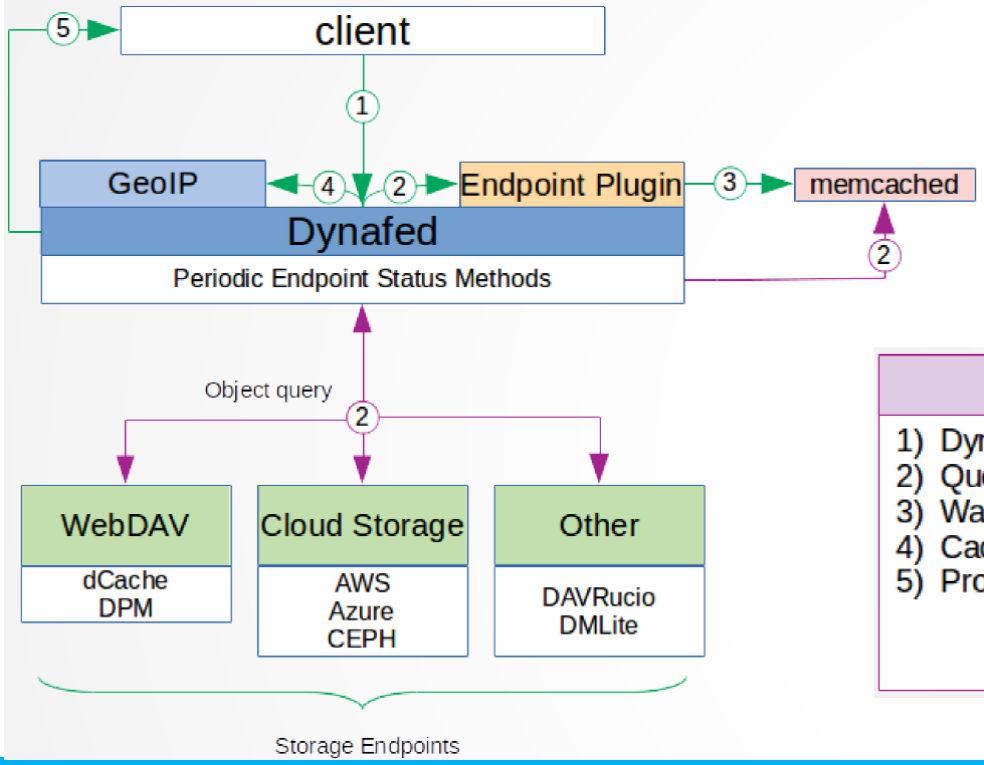
EULake

- CERN提出的数据湖实现，基于EOS存储系统
  - EOS是CERN开发的新型EB级分布式文件系统，用于管理超过200PB的LHC实验数据
- 针对分站点资源的动态存储缓存技术
- 远程站点保存主站点数据的副本（静态/动态）



dynafed

- 将远程的多个数据服务端“endpoint”（比如S3, WebDAV, Rucio等）动态构建成统一视图
- 将数据重定向到“最近的”存储，加速访问
- Belle-II使用RO（只读）模式（生产系统）
- ATLAS使用RW（读写）模式（测试阶段）



Client Request Processing

1) GET request is received
2) Poll connected SEs
3) Pass positive results to memcached
4) Choose SE according to GeoIP
5) Provide redirect link to client

Object Query

1) Dynafed checks cache for existing entry
2) Query all connected endpoints
3) Wait for returns until timeout
4) Cache all positive results
5) Provide list of file replicas

事例数据库

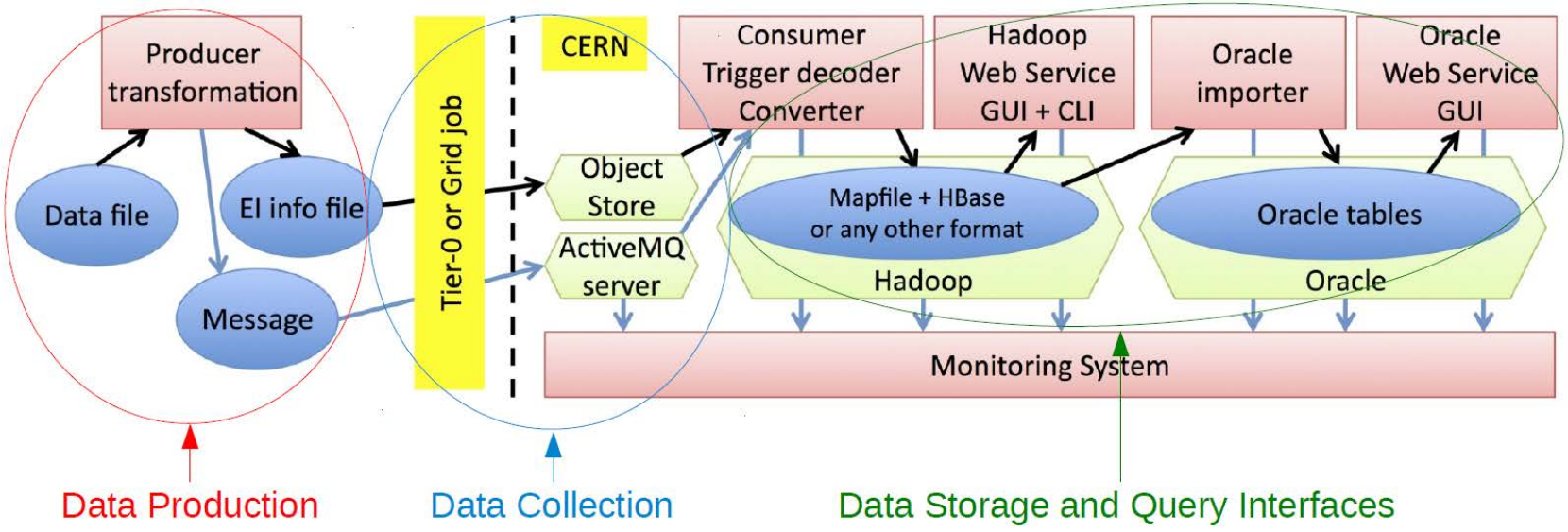
- 使用场景
  - 事例筛选：海量事例数据中快速筛选和发现极少量特殊事例
  - 加速分析：基于事例的并行化操作
  - 一致性检查：不同的文件或者数据集中出现重复的事例ID
- ATLAS EventIndex
  - Atlas事例索引数据库，目前已经索引万亿级事例，存储在CERN及相关网格站点上的文件中，总数据量超过100PB
  - 平均每天索引35亿个事例，峰值可达1000亿
- EventDB
  - 国家重点研发计划支持的万亿级半结构化大数据管理系统
  - 针对高能物理的需求，已经支持BOSS软件
  - 已经索引BESIII、HXMT等实验的1万亿个事例

EventIndex

Atlas事例索引数据库，基于Hadoop实现

1. 事例存储在文件中，通过GUID来标识
2. 文件聚合成数据集（dataset）
3. 事例索引大小从300到1KB不等

- Event identifiers (run / event numbers, trigger stream, luminosity block)
- Online trigger pattern (l1, l2, ef)
- References (pointers) to the events at each processing step
- (RAW, ESD, AOD, DAOD) in all permanent files on storage



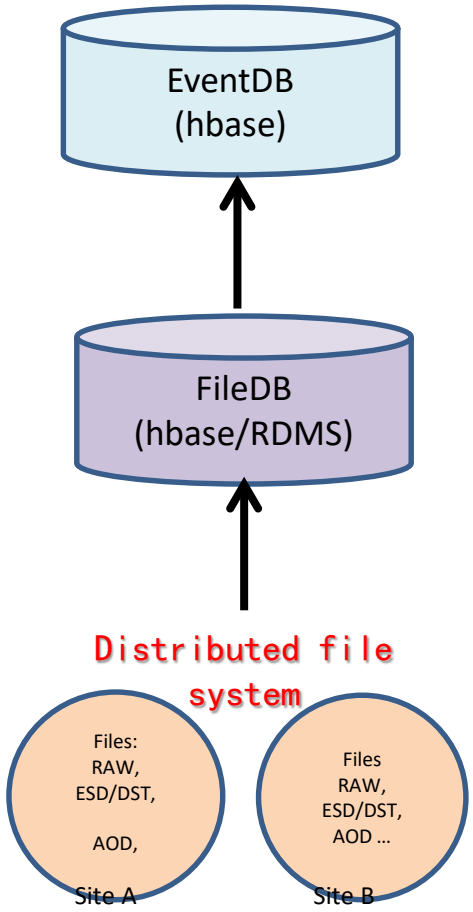
EventDB

- Event-level metadata system intended to discover and select events of interest to an analysis
- Store event TAGs and its location in files
- Export index file after selection

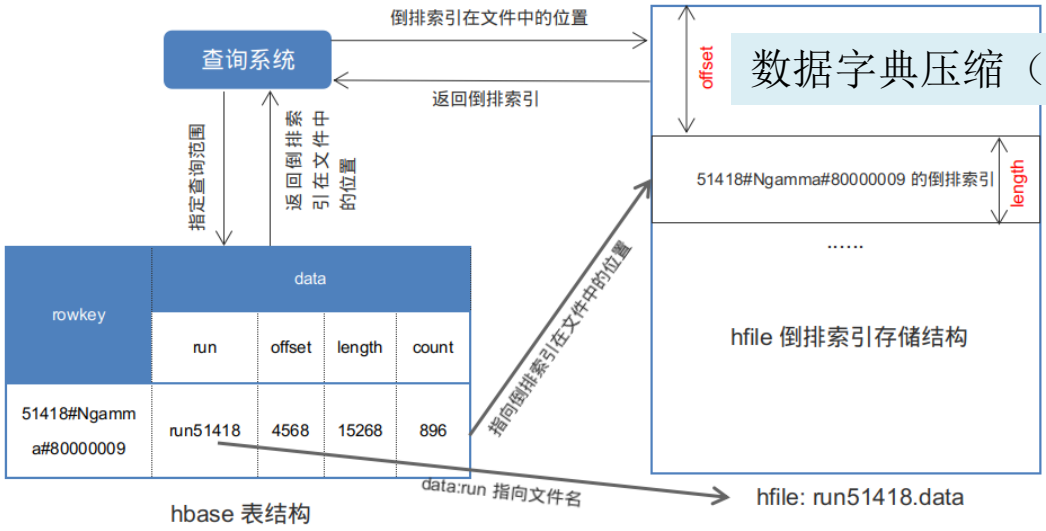
Rowkey	EventIndex
-8026#Neutral#003	pips1.dst-0
-8026#Neutral#004	pips1.dst-0, pips1.dst-5, pips2.dst-8
-8026#Neutral#005	pips1.dst-0, pips1.dst-5, pips2.dst-9
-8026#Neutral#006	pips1.dst-0, pips1.dst-5, pips2.dst-10, …
-8026#Neutral#007	pips1.dst-0, pips1.dst-5
-8026#Neutral#008	pips1.dst-0, pips1.dst-5, pips2.dst-12
-8026#Neutral#009	pips1.dst-0, pips1.dst-5, pips2.dst-13, …

Part of the clustered data

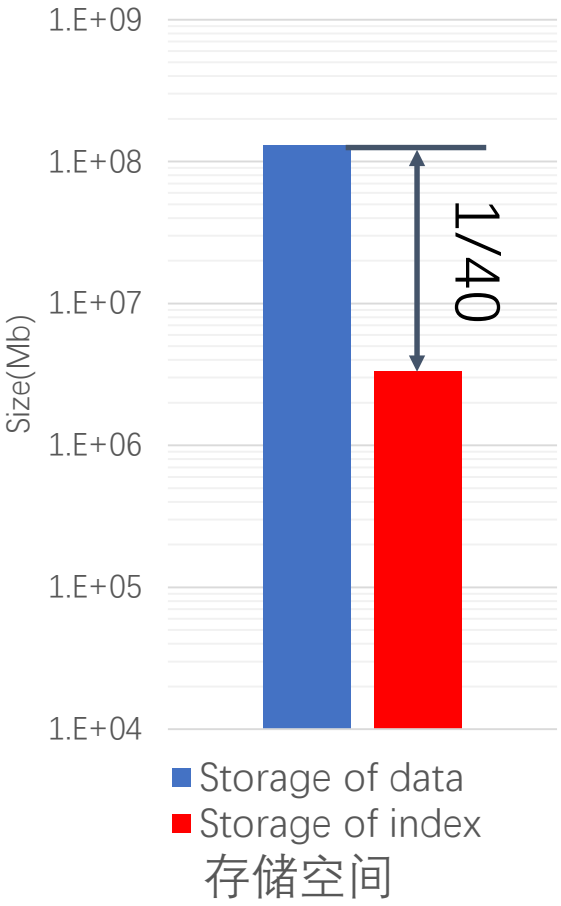
NoSQL database： HBase
Rowkey: runNo#PropertyName#Value
Value: Filename-Eventoffset,……

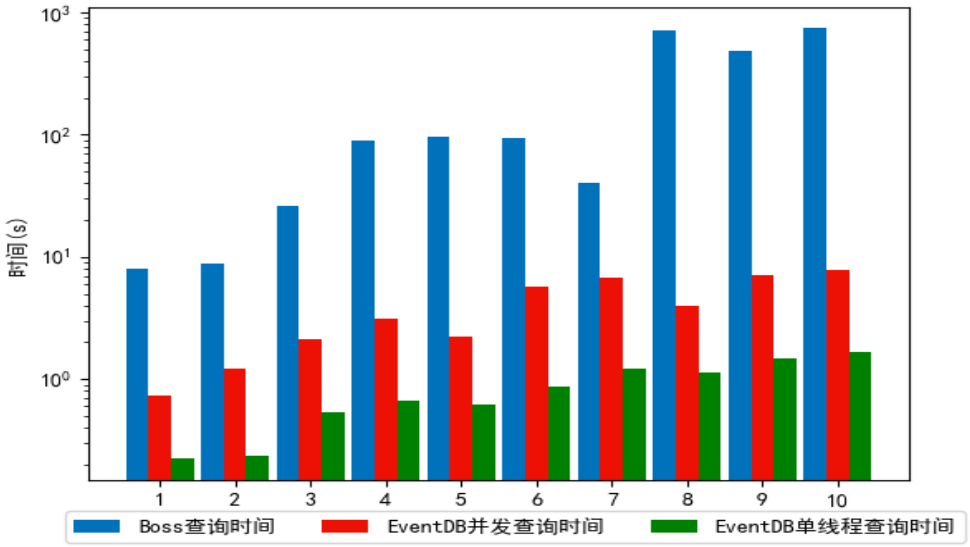


初步测试



数据字典压缩（压缩率1:8）





How to use it：

1、Use your afs account to log in to lxslc6.ihep.ac.cn;

2、Copy the environment variable script file to the home directory;

cp /afs/ihep.ac.cn/soft/common/EventDB/command_tools/CLIToolSet/initEventDB.sh ~/

3、Use the script file;

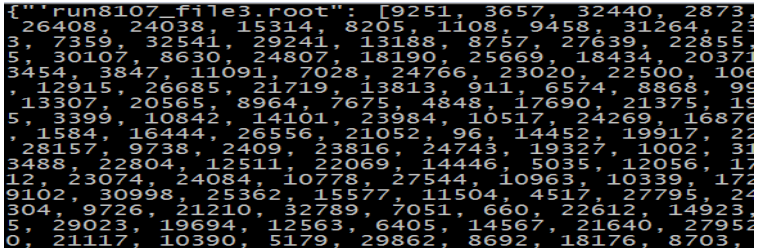
source initEventDB.sh

4、Use the event extractor to do the extraction;

EvtQuery -v 700 -r -8107 -q "range(NTracks,4,5)" -f 700.json

5、Use the json index file as input of joboption to do the analysis.

(Boss version 702p02)



Json file

```
#include "$ROOTIOROOT/share/jobOptions_ReadRec.txt"
#include "$VERTEXFITROOT/share/jobOptions_VertexDbSvc.txt"
#include "$MAGNETICFIELDROOT/share/MagneticField.txt"
#include "$ABSCORROOT/share/jobOptions_AbsCor.txt"
#include "$RHOPIALGROOT/share/jobOptions_Rhopi.txt"

// Input REC or DST file name
//EventCnvSvc.digiRootInputFile = {"rhopi.dst"};
TagFilterSvc.tagFiles = {
"/home/cc/wangcong/boss702p02/workarea/Event/TagFilterSvc/TagFilterSvc-00-00-04/share/1.json",
"/home/cc/wangcong/boss702p02/workarea/Event/TagFilterSvc/TagFilterSvc-00-00-04/share/2.json",
"/home/cc/wangcong/boss702p02/workarea/Event/TagFilterSvc/TagFilterSvc-00-00-04/share/1.json",
"/home/cc/wangcong/boss702p02/workarea/Event/TagFilterSvc/TagFilterSvc-00-00-04/share/2.json"
};
EventCnvSvc.selectFromTag = 1;

// Set output level threshold (2=DEBUG, 3=INFO, 4=WARNING, 5=ERROR, 6=FATAL )
MessageSvc.OutputLevel = 5;

// Number of events to be processed (default is 10)
ApplicationMgr.EvtMax = -1;

ApplicationMgr.HistogramPersistency = "ROOT";
NTupleSvc.Output = { "FILE1 DATAFILE='rhopi_ana.root' OPT='NEW' TYP='ROOT'"};
```

总结

- 文件系统接口是最简单的使用方式，但是却不是最高效的
- 基于xrootd的应用层接口访问模式，稳定高效，同时支持远程直接打开、缓存等多种功能
- HL-LHC等新型实验的需求远远超过目前能够提供的资源，必须要采用更为精细的计算模式和简单的存储模型
- “数据湖”是其中一种比较可行的解决方案，类似于边缘云计算，能够有效降低分布式数据管理的成本
  - 统一视图以及高效的存储和传输是其中要解决的一些关键问题
- RUCIO、Dynafed、事例数据库等数据管理工具都需要物理软件系统的支持
- 以应用需求为牵引，实现数据管理平台和物理软件的深度融合，是解决未来大规模数据管理的一个重要方案

讨论

- 数据存储系统
- 数据访问协议
- 广域网数据管理
- Bookkeeping & 数据集管理
- 事例数据库
- 其它