高海拔宇宙线观测站
Large High Altitude Air Shower Observatory

Institute of High Energy Physics
Chinese Academy of Sciences

NEC'2019

Distributed Data Management system for LHAASO

Haibo Li

On behalf of IHEP Computing Center, CAS

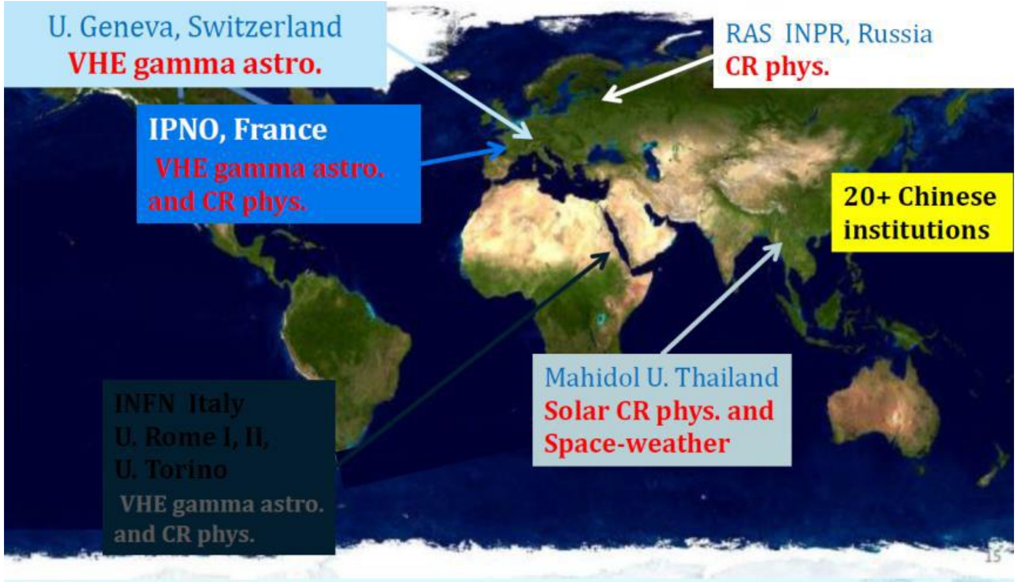
3 October, 2019

Outline

- Overview of the LHAASO project
- LHAASO offline data processing platform
- LEAF architecture and implementation
- Evaluation results
- Summary

The LHAASO Project

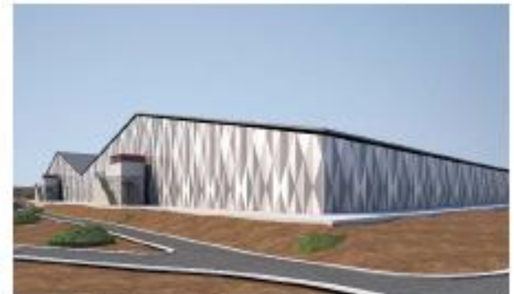
- Large High Altitude Air Shower Observatory (LHAASO)
- An major infrastructure project of 12th Five-Year Plan
- A new generation all-sky instrument to perform a combined study of cosmic rays and gamma-rays in the wide energy range 10 TeV -- 1 EeV
- Funded mainly by China, 20+ institutions joining the collaboration
- Investment of 1.2 billion RMB(174 million USD)



The LHAASO Project

Located in Daocheng, Sichuan, 4410m a.s.l

LHAASO



KM2A:
5195 EDs
1171 MDs

WCDA:
3120 cells
78,000 m^2

WFCTA:
18 telescopes
1024 pixels each

Future Enhancements: e.g., LHAASO-ENDA …

TBD …

Offline data processing workflow

- After the experimental data is acquired by DAQ, it enters the offline computing platform
- Provide support services for data storage, transmission, sharing, analysis and processing

Small on-site data center at Haizi Mountain observatory (4410m)

~2000CPU cores and 700TB disk storage for calibration and rapid reconstruction

~300Mbps

Operation center at Daocheng city

~1Gbps

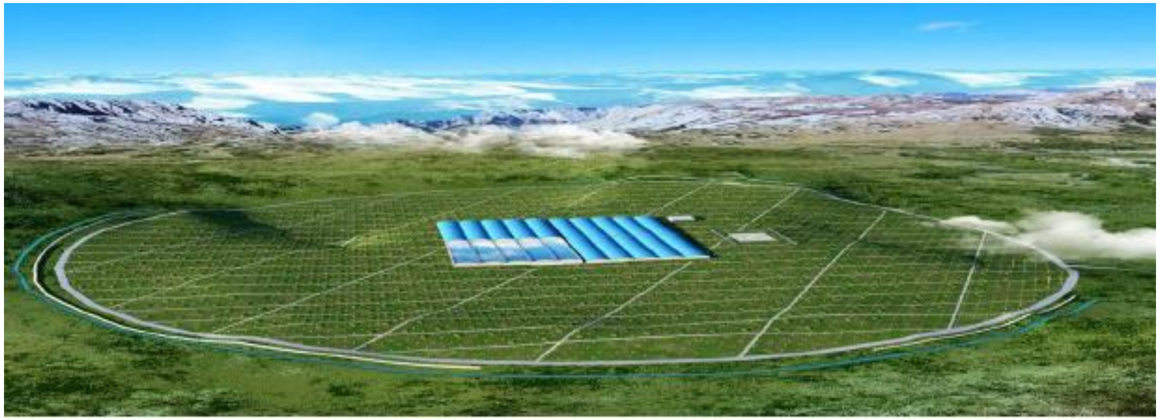
Distributed Computing sites (grid/cloud)

Large Offline data center at CC-IHEP

~4000 CPU cores and 4PB disk storage, 20PB tape storage for simulation, reconstruction, analysis, data storage and archive

LHAASO Computing requirements

- ~6 Petabytes of data annually generated by the LHAASO detectors
  - 6 PB of raw data, and >200TB of reconstruction data
  - Totally >60PB for ten years
- >2 Petabytes of data generated by MC simulation
- To build one distributed computing system containing about 6000 CPU cores to process the data
  - ~ 4500 CPU cores for reconstruction, analysis, …
  - ~ 1500 cores for production



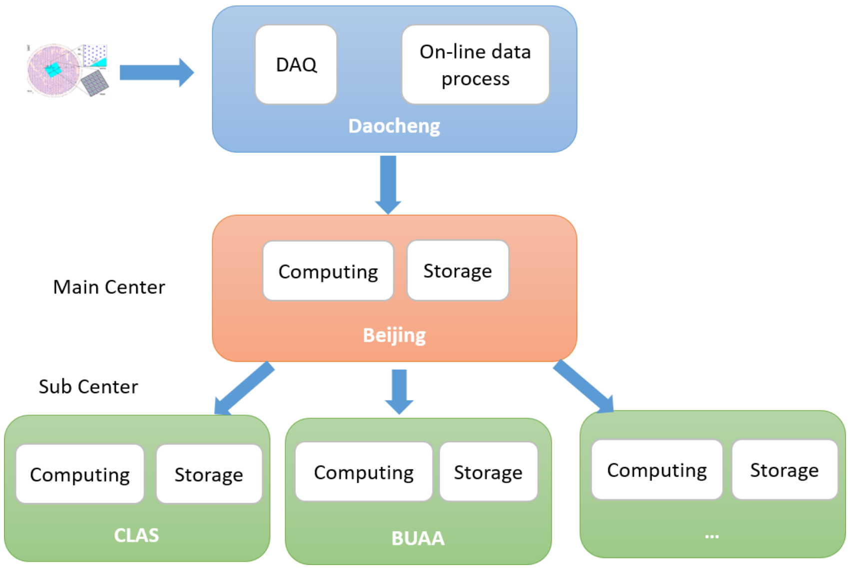
Current LHAASO computing environment

- Adopt the mode of "taking data while building", currently 1/2 stage
- Daocheng Observation Base
  - DAQ, data filtering, fast reconstruction, compression, etc.
  - Transfer raw data and fast reconstructed data to main center
- Beijing local cluster
  - Storage of all data (raw, reconstructed, simulated, analyzed, etc.)
  - All data reconstruction computation
  - Distribution of reconstructed data to sub-centers
  - Receiving simulation and analysis data from the sub-center
- Chengdu cluster
  - Simulation and analysis

Site	Function	Computing	Storage
Haizi Mountain observatory	fast reconstruction	468 Cores	700 TB
Beijing Local Cluster	Data reconstruction and analysis	15000 Cores	2.4 PB
Chengdu Cluster	Simulation and analysis	200 Cores	145 TB

LHAASO data processing platform architecture

- Adopt "main center + sub-center" distributed computing solution
- The sub-center uses cloud computing technology, using openstack and singularity technology to manage
- Manage job scheduling with HTCondor
- Local storage managed by EOS
- Software storage uses CVMFS
- Using LEAF to provide cross-domain access

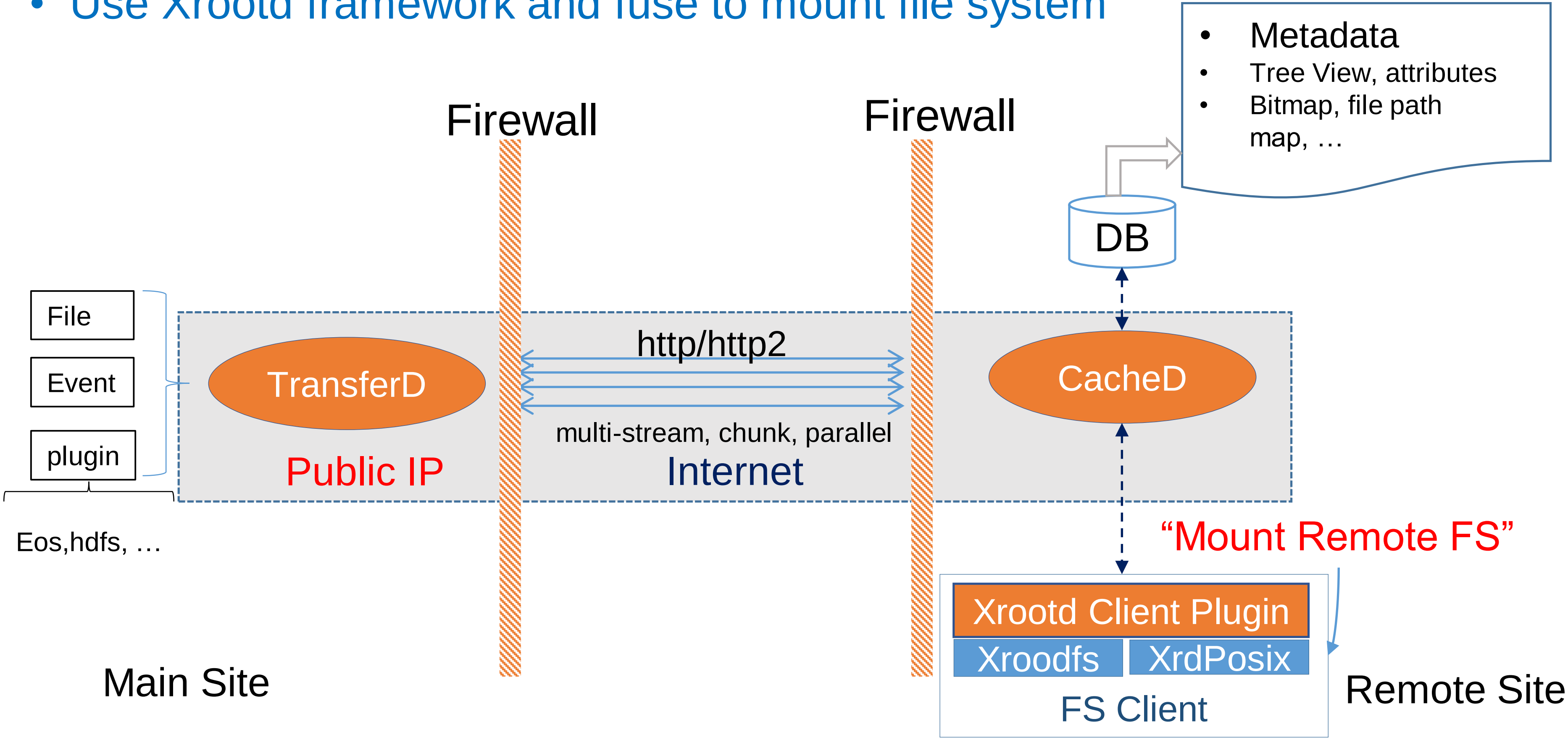


What's LEAF?

- LEAF is a data cache and access system across remote sites
  - Same file system view at local and the remote sites
  - Good access speed over WAN
  - Client requests are served as soon as one small fraction of file is available before one whole file is fully downloaded
  - Portable, compatible and scalable
  - Secure and reliable

LEAF Architecture

- Full Metadata synchronization from main site periodically
- Data transfer technologies: multi-stream, chunk, non-block, etc
- Use HTTP protocol to go through firewall
- Use Xrootd framework and fuse to mount file system



File Transfer Service

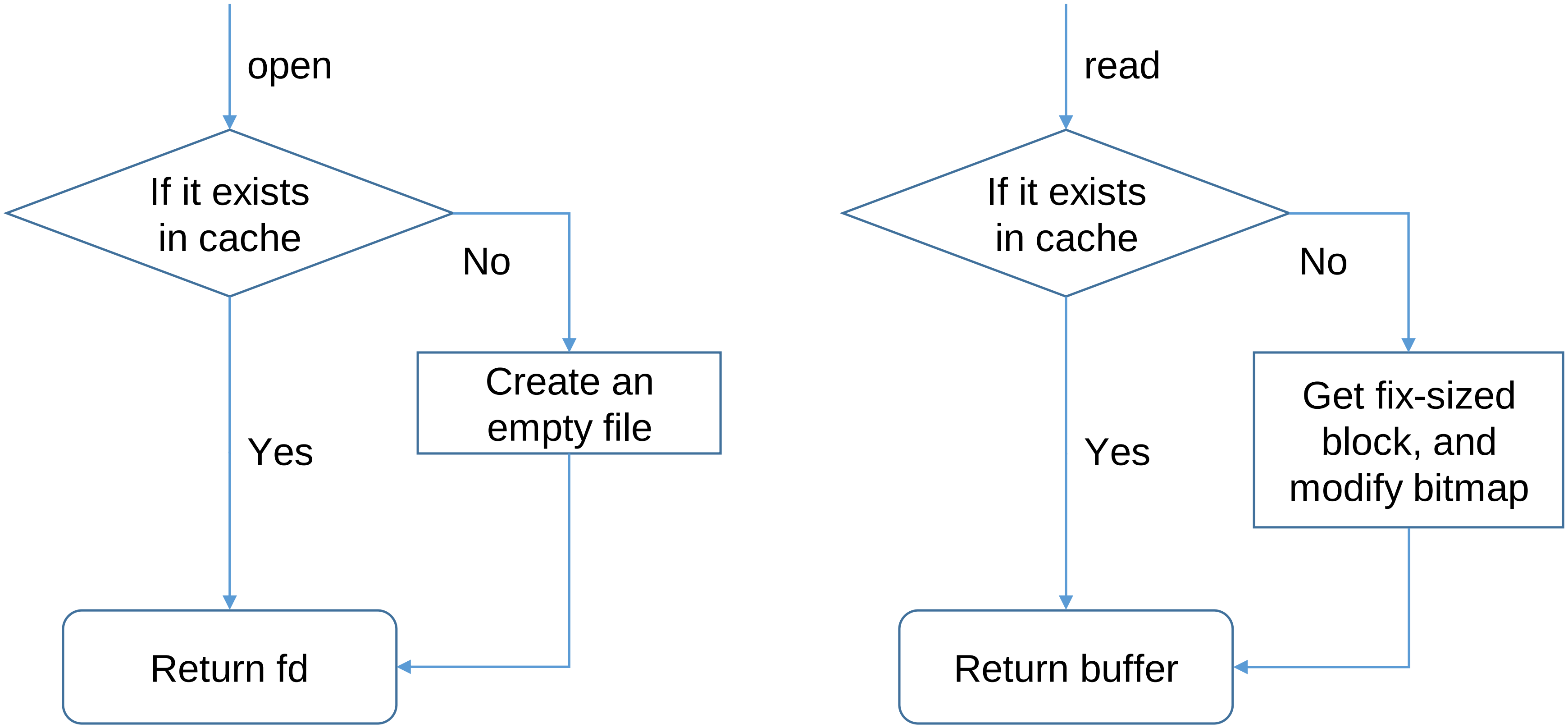
- Two components
  - TransferD: daemon running at Main site
  - Client library: deployed at remote site, called by CacheD
- Based on Tornado web framework
  - a python web framework and asynchronous networking library
  - support non-blocking network I/O, suitable for long polling, WebSockets, long-lived connection
- If file transfer service receives a request, it will download or upload data using multi-streams in parallel
- Client routines have theses parameters: file path, file operation (stat, getdir, read, write, …), mode, offset, …
- Easy to go through firewall using HTTP protocol
  - Usually client doesn't have public IP behind the firewall

Disk Cache Service

- Three components
  - CacheD: daemon running at remote site
  - DB: store file metadata and bitmap
  - Client tool and library: called by xrootd client plugin
- CacheD will get all entries periodically from main site once the “exported” file system is defined
- DB supports Mysql and Ramcloud currently

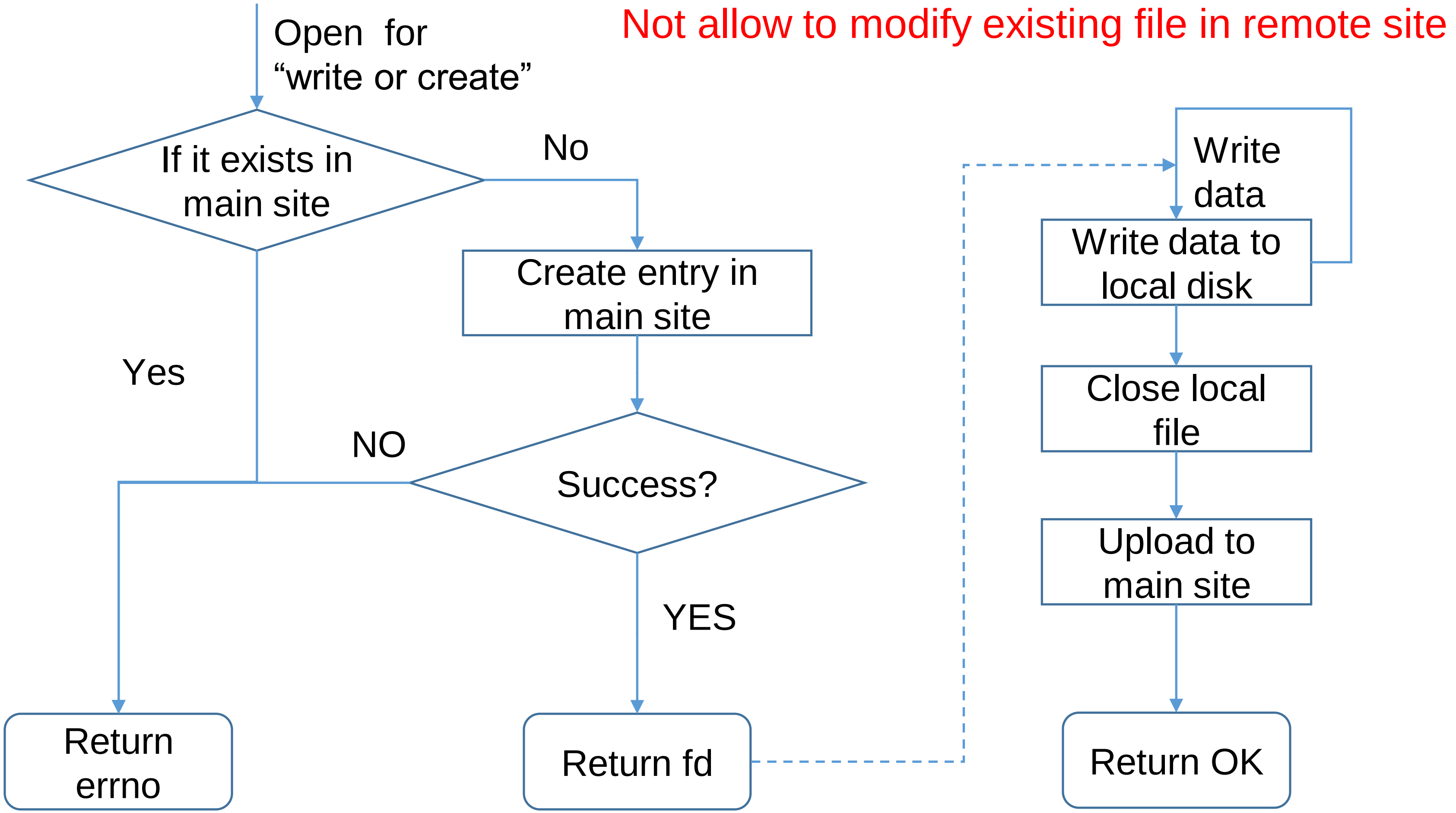
Open/read workflow

- CacheD creates an empty file on local disk once it receives 'open' request from client
- CacheD gets fixed-size block (1MB) from offset specified by 'read' operation



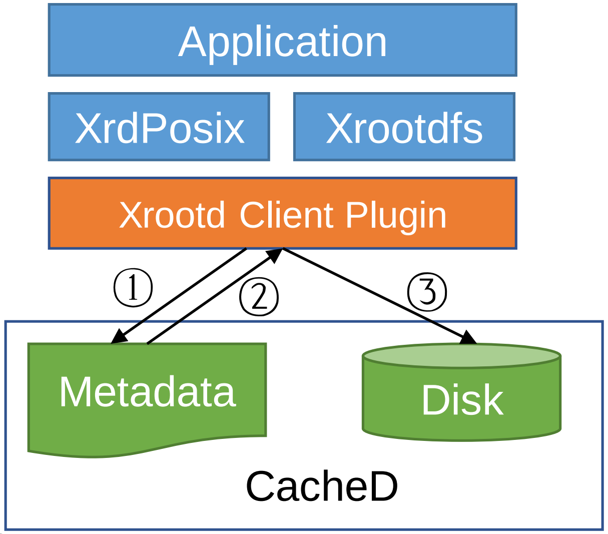
Write workflow

- CacheD puts the whole file in local disk, then upload it to the main site later in case of 'write' operation



Xrootd client plugin

- Application access data using xrdposix API or xrootdfs
- Implement a xrootd client plugin
  - 1) check if the block is in cache. If not, it calls cached to get the block from main site
  - 2) return physical path of the file
  - 3) get real data from disk using xrootd



- Xrootd client plugin manager
  - /etc/xrootd/client.plugins.d
  - Manage a map between URLs and plug-in factories

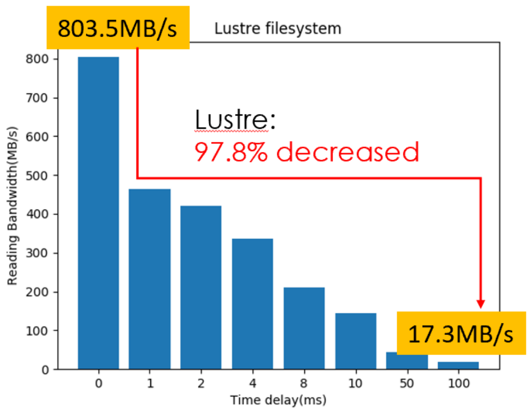
    url = root://cached.domain:1094

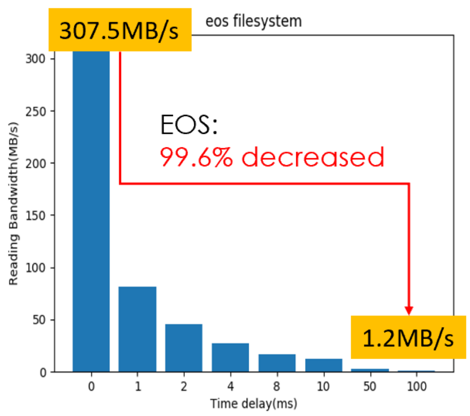
    lib = /usr/lib/libXrdLeafClient.so

    enable = true

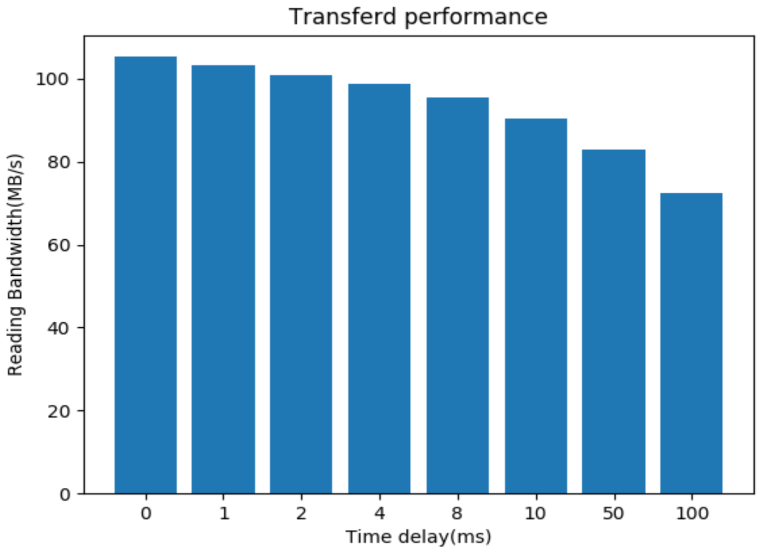
Performance evaluation

- Bandwidth: 1Gbps
- Latency: 1~100ms using tc simulation
- Transfer parameters: long-lived, 1M block, 10 streams





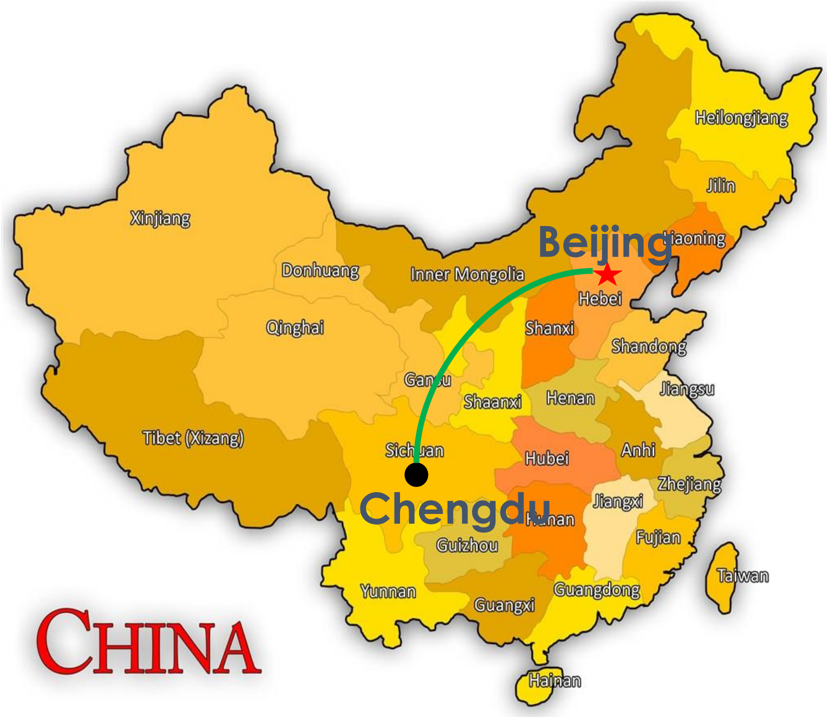
Round trip latency	Transfer performance (MB/sec)
0 ms	105.3
10 ms	90.7
50 ms	82.8
100 ms	72.5

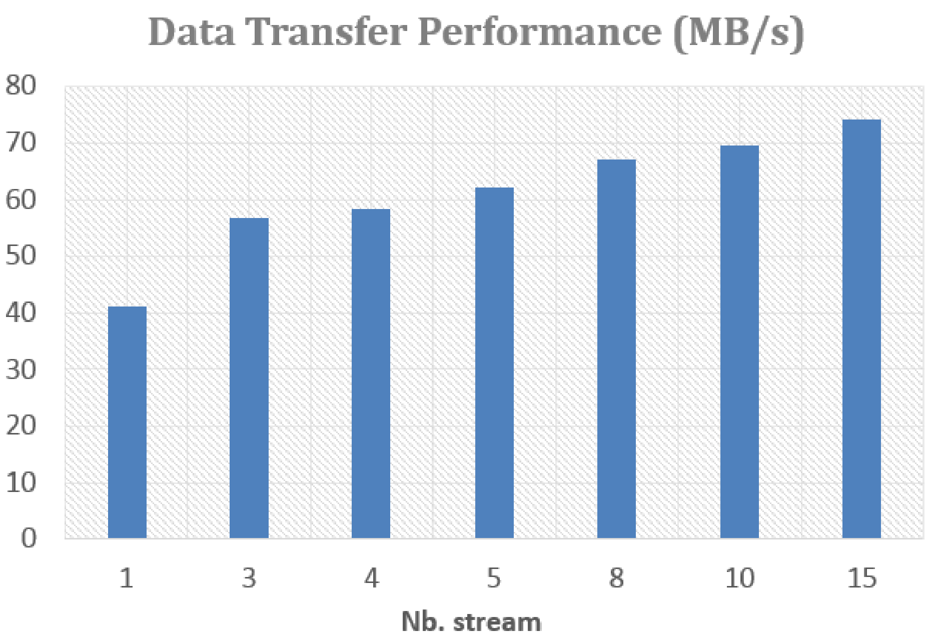


Results: decreased by 31% (105MB->72.5MB), better than EOS/Lustre

Testbed

- Two sites： IHEP (Beijing) <-> CLAS (Chengdu)
- Distance: ~2000KM， Latency: ~35ms
- Bandwidth: ~1Gbps, Iperf: ~80MB/s
- Performance is getting better with the increasing of stream number





Summary

- LHAASO distributed computing has a need for remote data transmission
- LEAF provides a data cache and access solution for accessing data directly from remote site
- Implemented as a xrootd plugin supporting most of HEP applications transparently
- Adding new functions, eg HTTP2 support, event-level transfer, etc

Thanks for your attention!