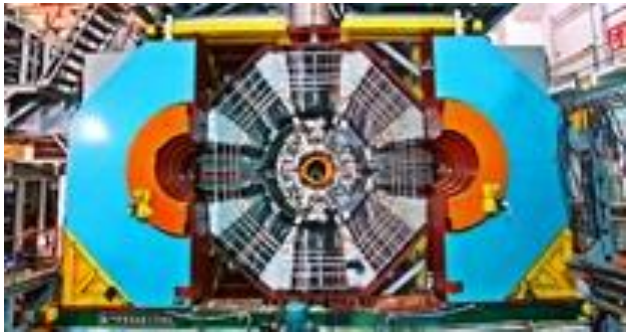
Cross-domain Data Access System for Distributed Sites in HEP

Qi XU (IHEP-CC, Chinese Academy of Sciences)

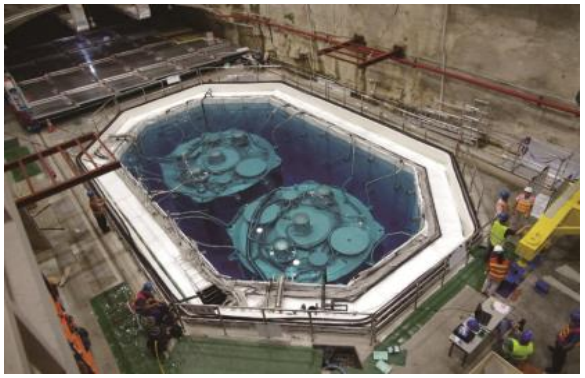
Outlook

- Data Storage in HEP
- Motivation
- System Architecture
- Performance Test
- Conclusion

Data Storage in HEP

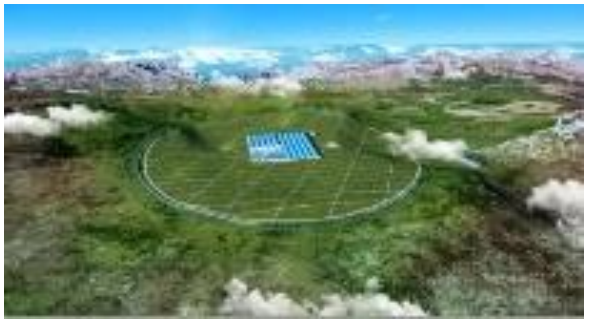


BESIII Experiment Data > 100TB/year



Daya Bay Experiment Data > 400TB in Total

Yangbajing Experiment Data > 200TB/year



LHAASO Experiment Data > 2PB/year

Data Storage in HEP

- Distributed Sites in HEP.
- Large-scale Data Sharing.
- Batch Mode with High Latency Response.
- Computing and Scheduling Based on Files.

Motivation

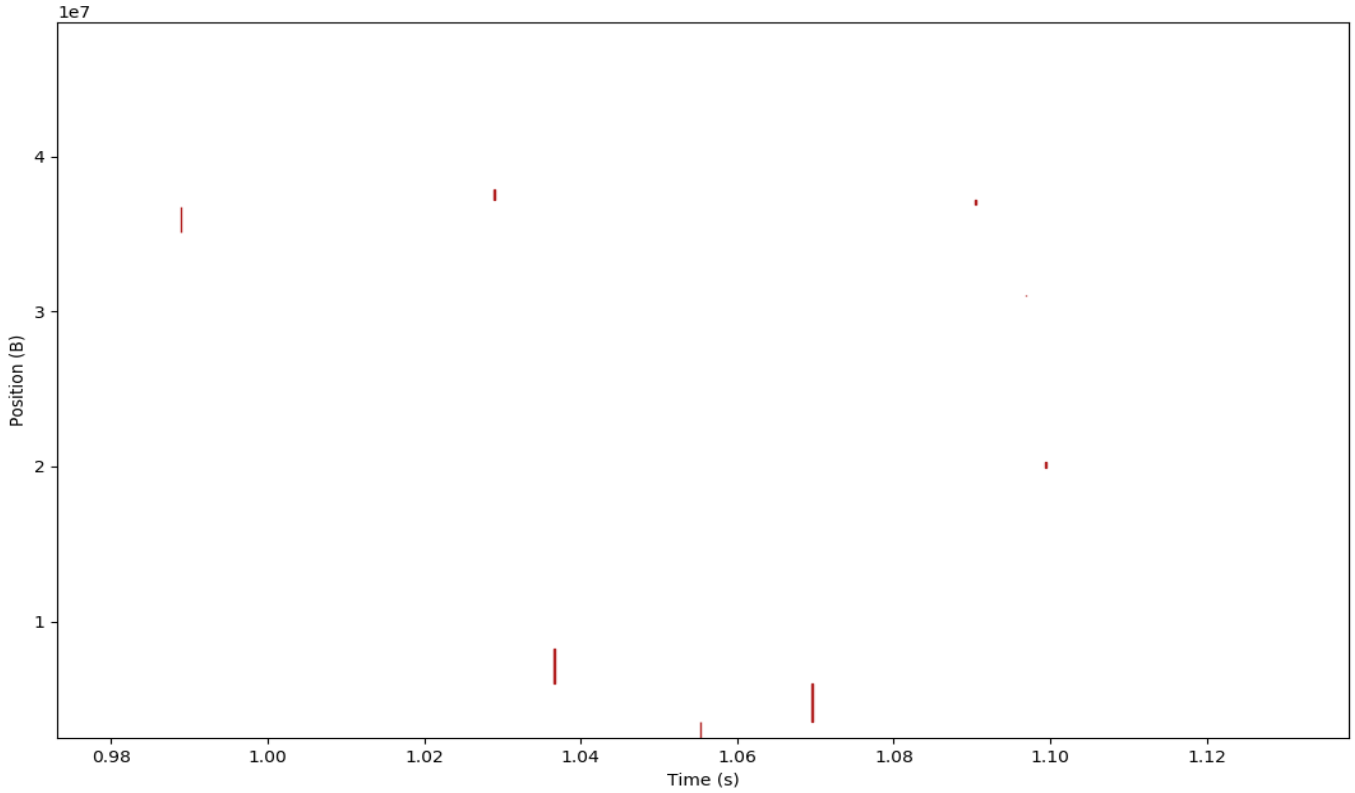
- Current data sharing mode is hard for unified management.
- A huge consumption of resources: network, storage and CPU.
- Hard to seek and read events based on files, scheduling is not flexible.
- Long time to get response in batch mode, inefficiently.

Motivation

Test 1:

Analysis Based on File System Log

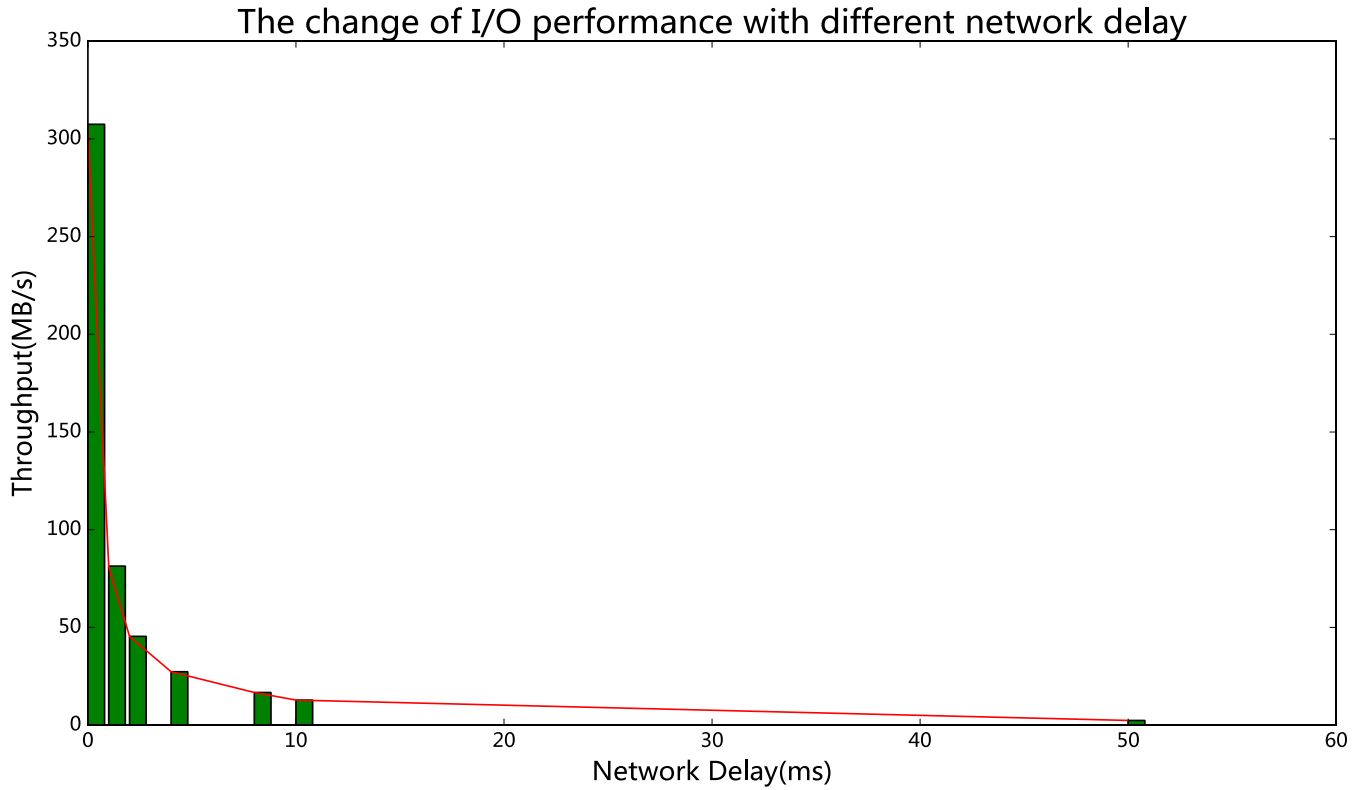
File Size is *478.75MB*, Read *22MB*, Read Ratio *4.6%*



Motivation

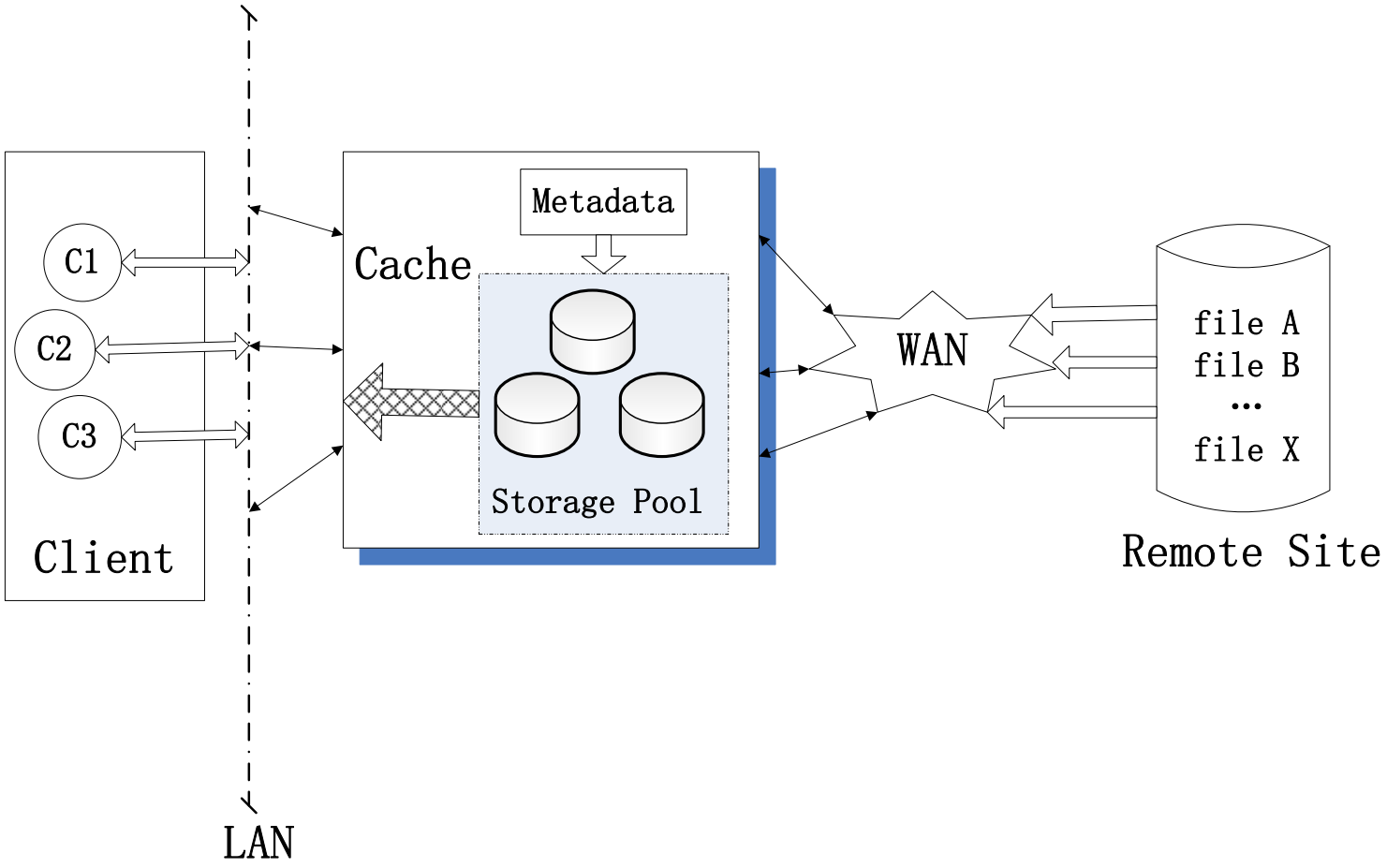
Test 2:

Throughput Test of Traditional Distributed File System with Different Latency in WAN



System Architecture

Streaming Transmission & Cache Service in Cross-domain Data Access

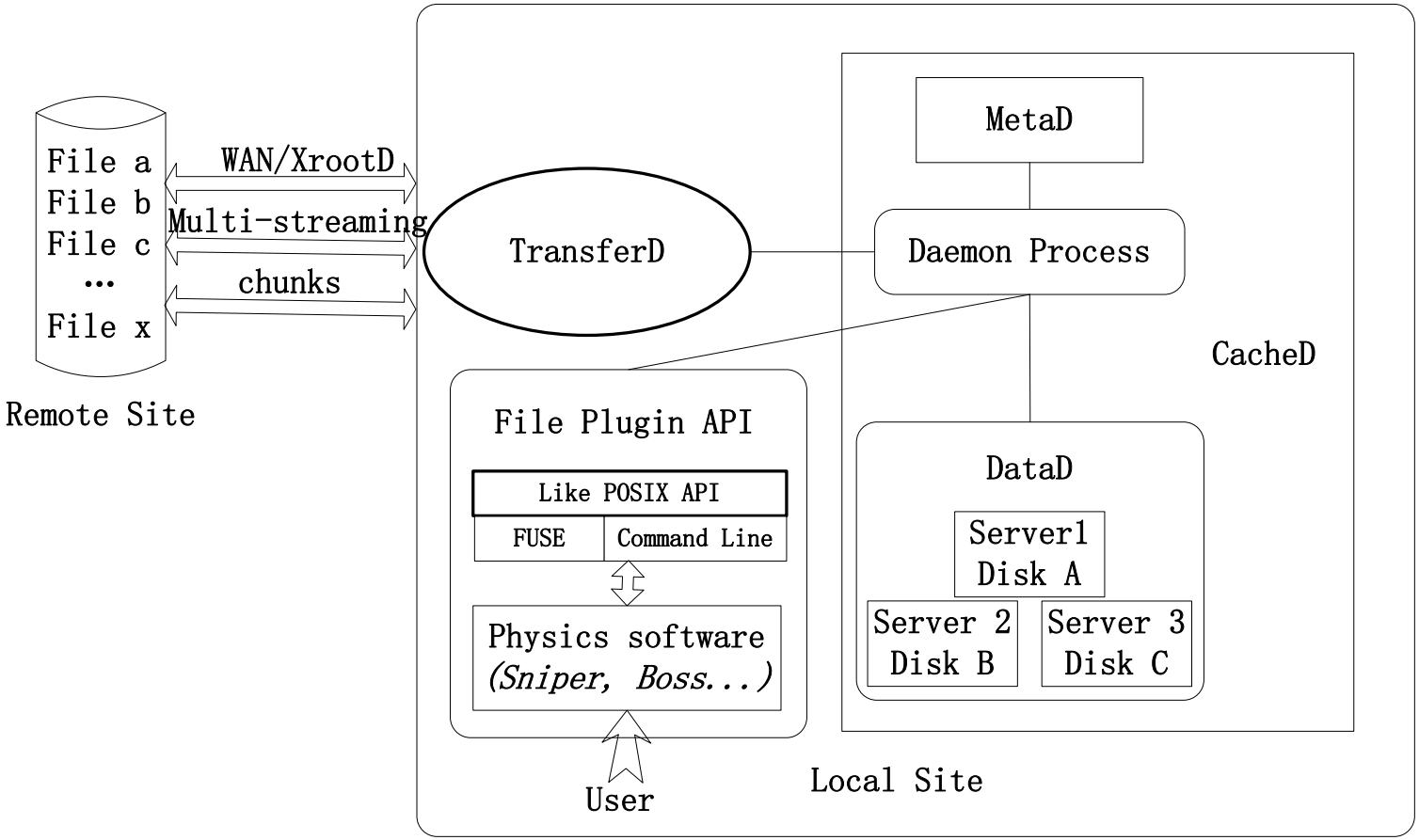


System Architecture

System is Consist of Three Loose Coupling Unix Services

- CacheD: Consists of three parts: MetaD, DataD and Daemon Process.
- TransferD: XrootdProtocol
- File Plugin: Similar to POSIX File System API

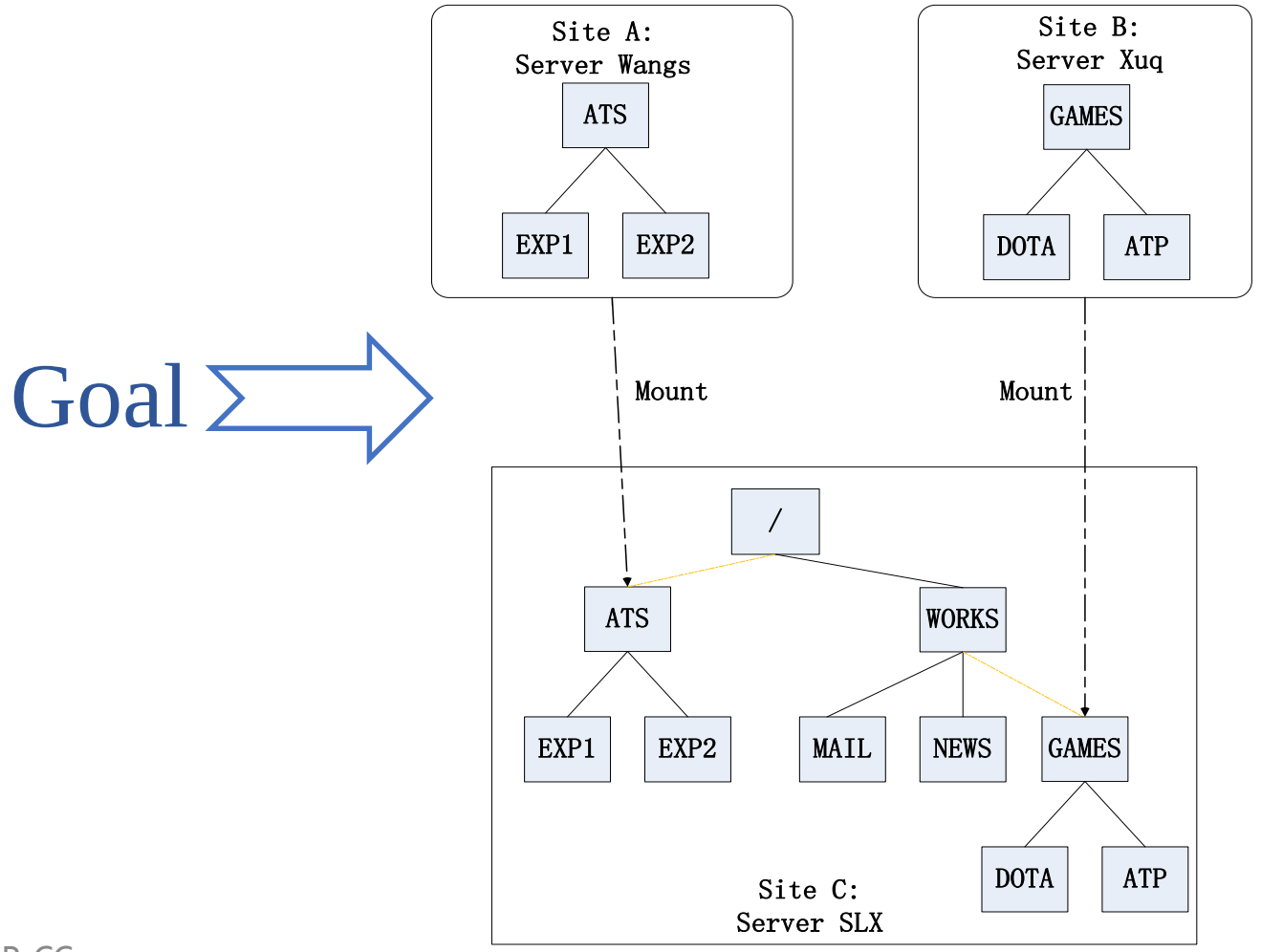
System Architecture



The structure of the cross-domain data access system

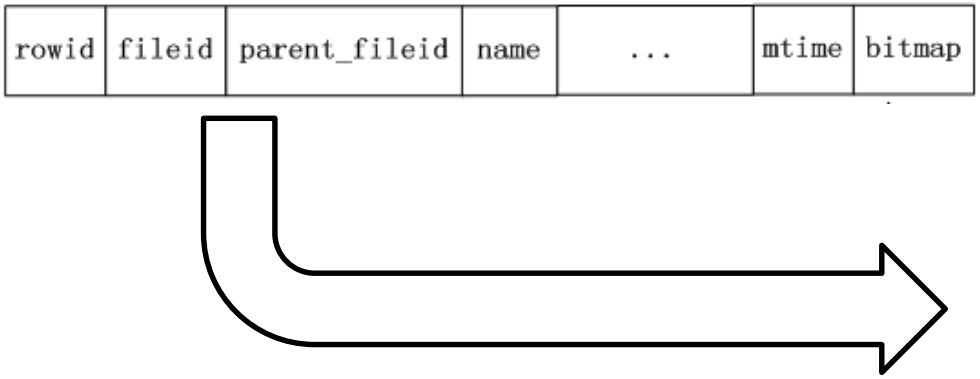
System Architecture

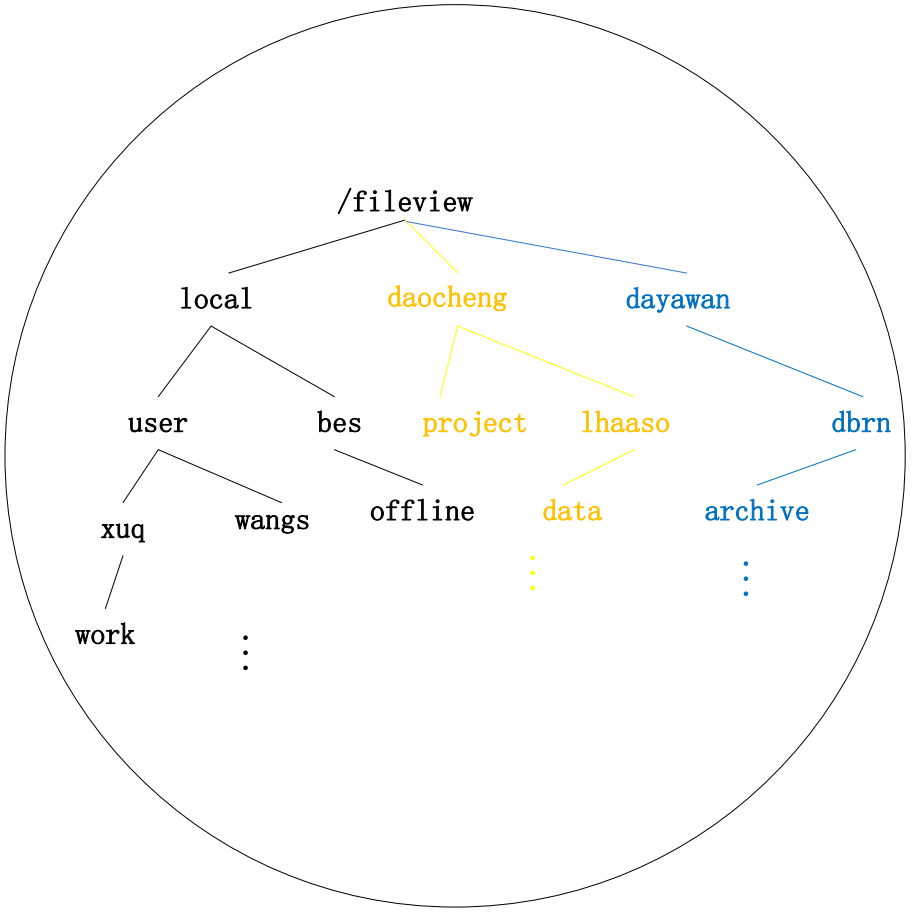
MetaD: Storage and Management for metadata



System Architecture

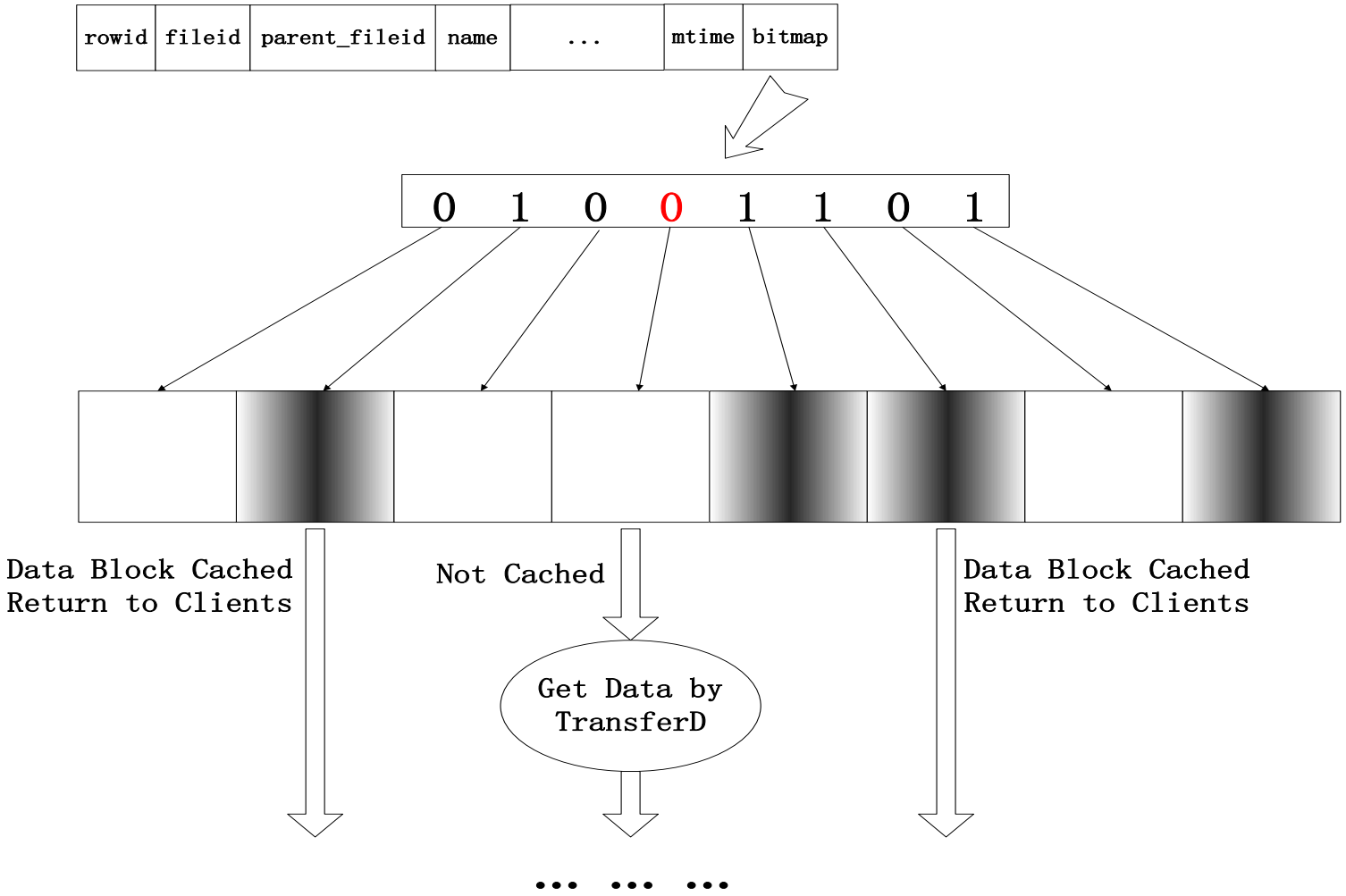
Fileid Parent_fileid: Uniform File View





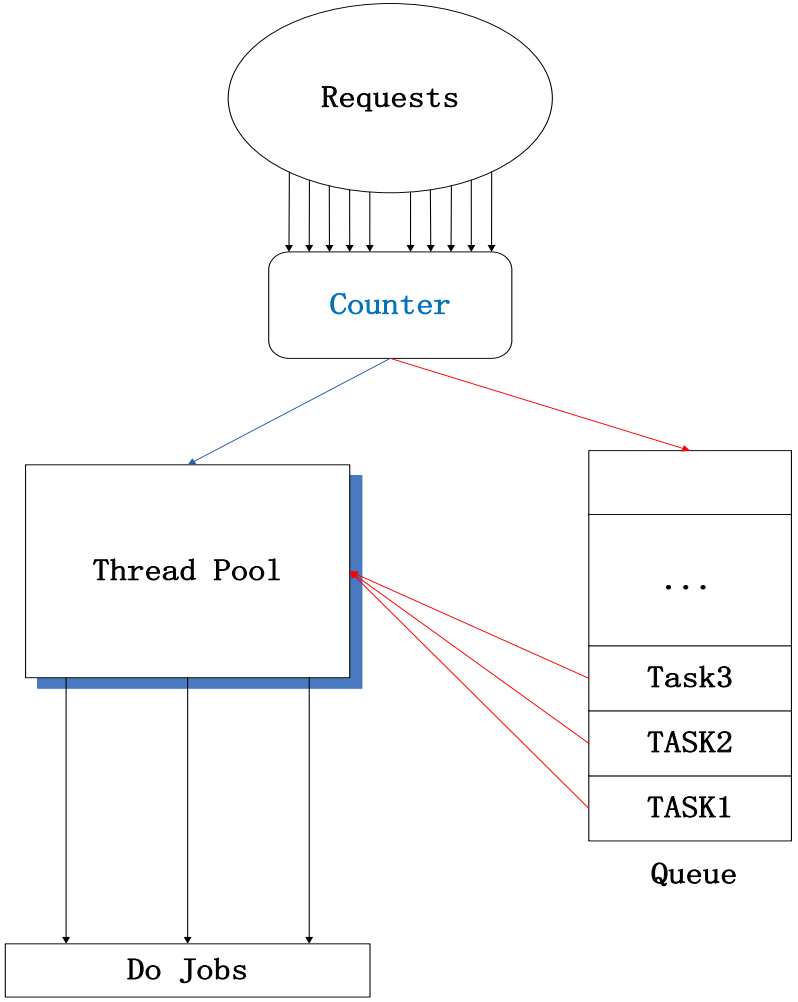
System Architecture

Bitmap: On-deman Data Access



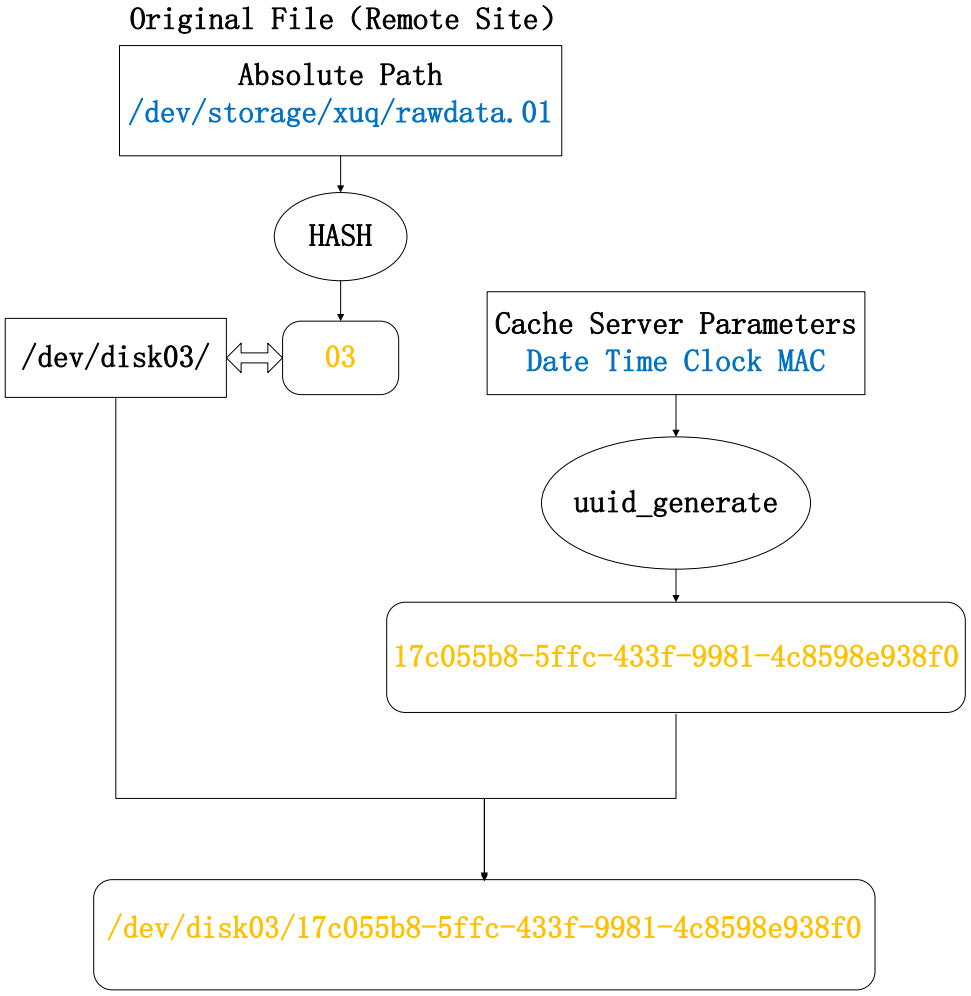
System Architecture

Daemon Process: Concurrent Message Process



System Architecture

DataD: Storage for Cache Files



System Architecture

TransferD: Xrootd Protocol

Optimization

- Multiple streams are supported on a single socket.
- Clients can be redirected to another server at any time.
- Clients may be asked to delay server contact.
- Clients may piggy-back read-ahead lists with any read request.
- Servers may ask clients to perform certain actions at any time.

System Architecture

File Plugin: Similar POSIX File System API

int cdas_open()	Open file (OW/OR)
int cdas_close()	Close file (Auto)
int cdas_getattr()	Get metadata from remote site
int cdas_read()	Read file (Transfer data block, if not cached)
int cdas_access()	Whether file is accessible
int cdas_opendir()	Open directory, get DIR_ID
int cdas_readdir()	Read directory (Metadata of files in it)
int cdas_rfsync()	Sync files to remote site
int cdas_refresh()	Update cache files
int cdas_unlink()	Delete cache files

10

Performance Test

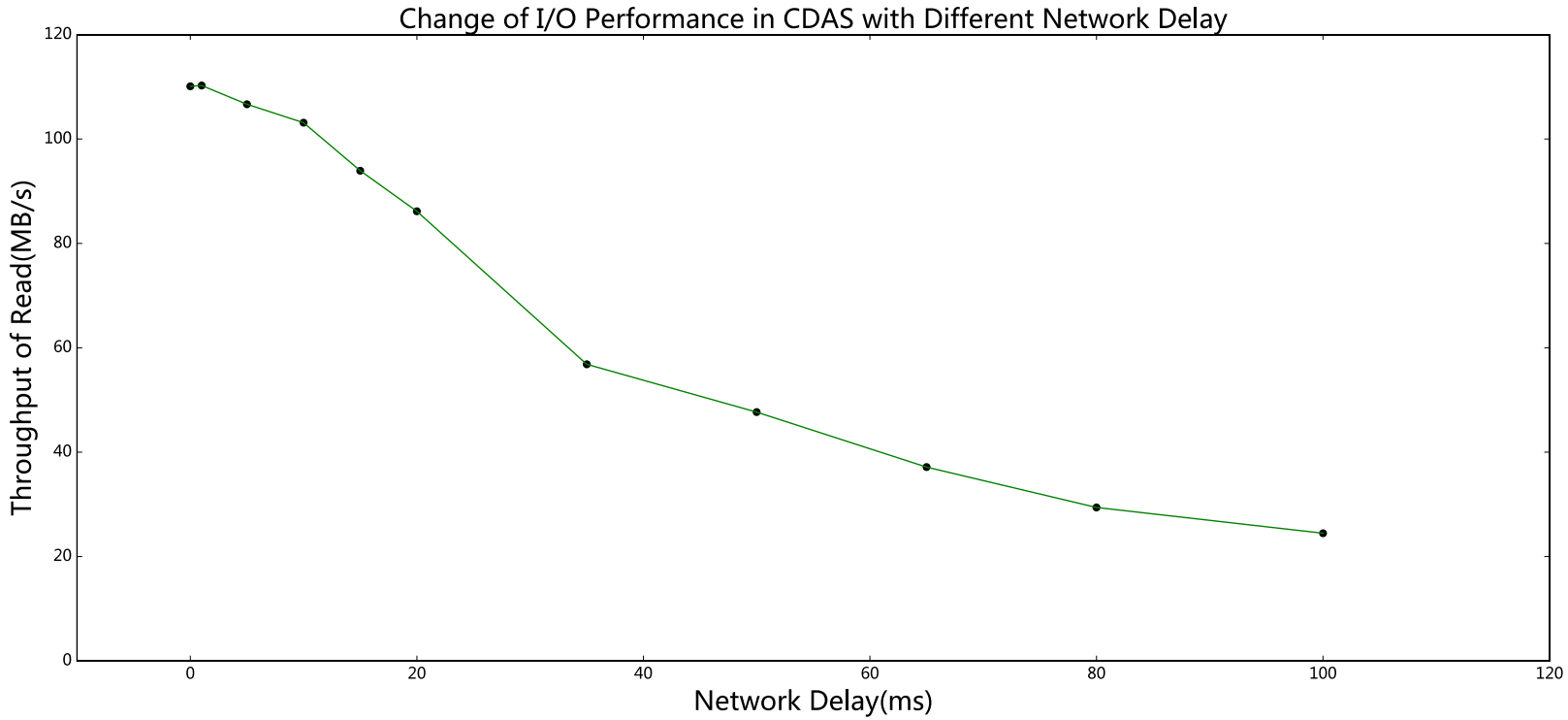
I/O Performance of Lustre EOS and the System are Tested in Bandwidth 1000Mbps

Read performance (Throughput) of different file system (MB/s)

Network Delay (ms)	Lustre	EOS	CDAS (file not cached)
0	79.85	30.12	110.10
1	45.93	7.26	110.25
10	13.87	1.03	103.13
100	1.49	0.12	24.46

Performance Test

I/O Performance with Network Delay Jitter is tested in Bandwidth 1000Mbps



The change of I/O performance with different network delay

Performance Test

Test of Command Line Interface

- ilist: show metadata of files in the remote site.
- idelete: delete metadata or data blocks of files in cache.
- iget: get data blocks of files from the remote site.
- iput: put new files to the remote site.

Performance Test

```
[root@bigdata07 client]# ./ilist  -l /xrootdD
-rw-------   1 root     root              104857600 Nov 22 11:21 100ce_ul
-rw-------   1 root     root              104857600 Nov 22 11:34 100cs_ul
-r--------   1 root     root              104857600 Jun 07 20:50 100m
-rw-------   1 root     root              104857600 Nov 22 20:48 100mb_ul
-r--------   1 root     root              104857600 Nov 18 17:17 100mm
-rw-------   1 root     root              104857600 Sep 23 21:28 100upup_ul
-rw-------   1 root     root              104857600 Sep 23 21:43 100upupup_ul
-rw-------   1 root     root              104857600 Nov 22 13:06 101_ul
-rw-------   1 root     root                      0 Nov 22 16:54 110_ul
-r--------   1 root     root                      5 Sep 24 17:28 1728
-rw-------   1 root     root              104857600 Sep 24 17:54 1up_ul
-r--------   1 root     root             2147483648 Nov 18 15:41 2000m
-r--------   1 root     root              209715200 Nov 18 15:43 200m
-rw-------   1 root     root                      0 Nov 22 20:49 200mb_ul
-rw-------   1 root     root                      0 Nov 22 13:21 202_ul
-r--------   1 root     root                     28 Sep 24 17:27 ceshi
-r--------   1 root     root              104857600 Sep 14 13:55 ceshi100
drwx------   1 root     root                     64 Nov 06 15:49 dir
-rw-------   1 root     root                     27 Sep 23 21:23 upup_ul
```

```
[root@bigdata07 client]# ./iget -fvm /xrootdD/100m /tmp
DEBUG: Cache file: /cdfs_data/0/429546d7-a073-4a9b-a3fa-3b03db0f401b filesize 104857600
 10485760 bytes     67.61 MB/sec avg     67.61 MB/sec instDEBUG: read 10485760 bytes of size 10485760
DEBUG: read 10485760 bytes of size 10485760
DEBUG: read 10485760 bytes of size 10485760
DEBUG: read 10485760 bytes of size 10485760
DEBUG: read 10485760 bytes of size 10485760
DEBUG: read 10485760 bytes of size 10485760
DEBUG: read 10485760 bytes of size 10485760
DEBUG: read 10485760 bytes of size 10485760
DEBUG: read 10485760 bytes of size 10485760
DEBUG: read 10485760 bytes of size 10485760

104857600 bytes transferred in 0.97 seconds (103.45 MB/s)
```

Conclusion

- Streaming Transmission and Cache Service are adopted to the system.
- The system is consist of TransferD, CacheD and File Plugin.
- Cross-domain data access is localized and transparent for users in the system.
- Data access is on demon and effective in the system.
- Excellent I/O performance of the system with high network latency in WAN.
- Stable I/O performance of the system with Network Delay Jitter.
- Make the most of resources in distributed sites.
- The system is suitable for cross-domain data access between distributed sites.

Q & A

Thank you for your attention