

RAL's efforts towards an SE

Alastair Dewhurst, George Vasilakakos, Ian
Johnson, Bruno Canning, James Adams, Alison
Packer



Outline

- Overview
- Ceph backend status
 - Erasure Coding
- XrootD implementation
- GridFTP development
- S3 status
 - Dynafed
- Summary



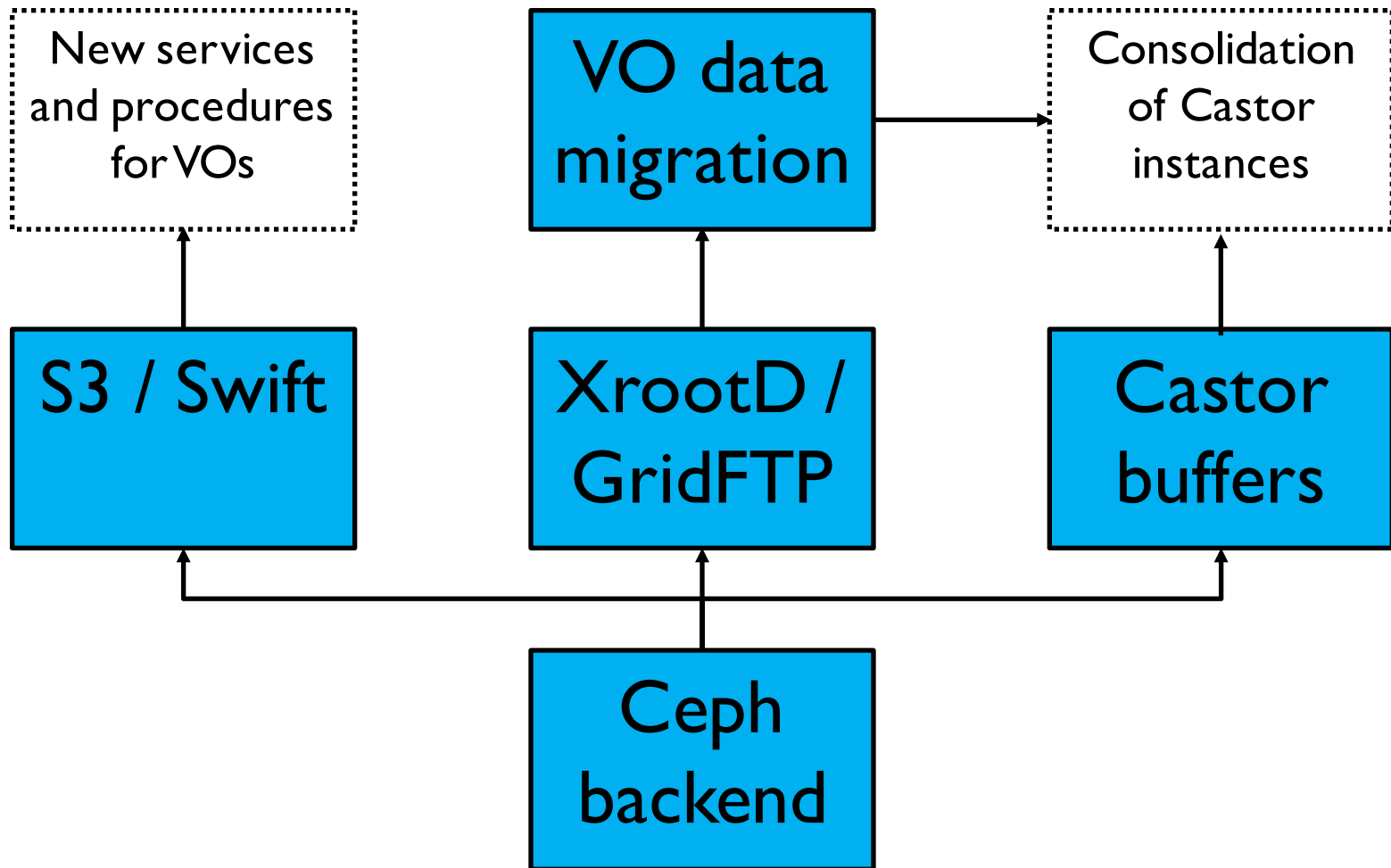
Motivation

- RAL are aiming to deploy a new disk only storage service to replace Castor for the LHC VOs.
 - It must be scalable to meet the data demands of the LHC to 2020 and beyond.
 - It must provide industry standard access protocols for use by future VOs.
 - Need to keep hardware costs in line with Castor – Erasure Coding required.
- The LHC VO are not currently able to run all their workflows through an SE providing only S3/Swift access.
 - Developing GridFTP + XrootD plugins
- Need to provide 5PB of usable storage to LHC VOs by April 2017.



The Echo Project

4



Ceph backend




- Current Echo cluster is ~5PB RAW size made from last years procurement
 - Castor disk servers with extra memory and CPU.
- Performance is ok...
 - RAID cards limiting performance.
 - Believe current problems with thread count and load should be resolved in Jewel.
- **63 new disk servers available this week!**
 - New cluster (running Jewel) will be created ASAP.



Erasure Coding

- Typical Castor disk server with 36 drives, 30 are storing data.
 - 2 disk for OS, RAID60 (i.e. $2 \times 15 + 2$) = 83% of raw storage is usable.
- EC breaks data into 'k' chunks and creates 'm' additional parity chunks.
 - Can lose any 'm' chunks without losing data.
- For Ceph we want $m = 3$ (at least).
 - Allows us to take advantage of self healing (reducing effort required to maintain).
- To keep overhead down, need k to be as large as possible without affecting performance.

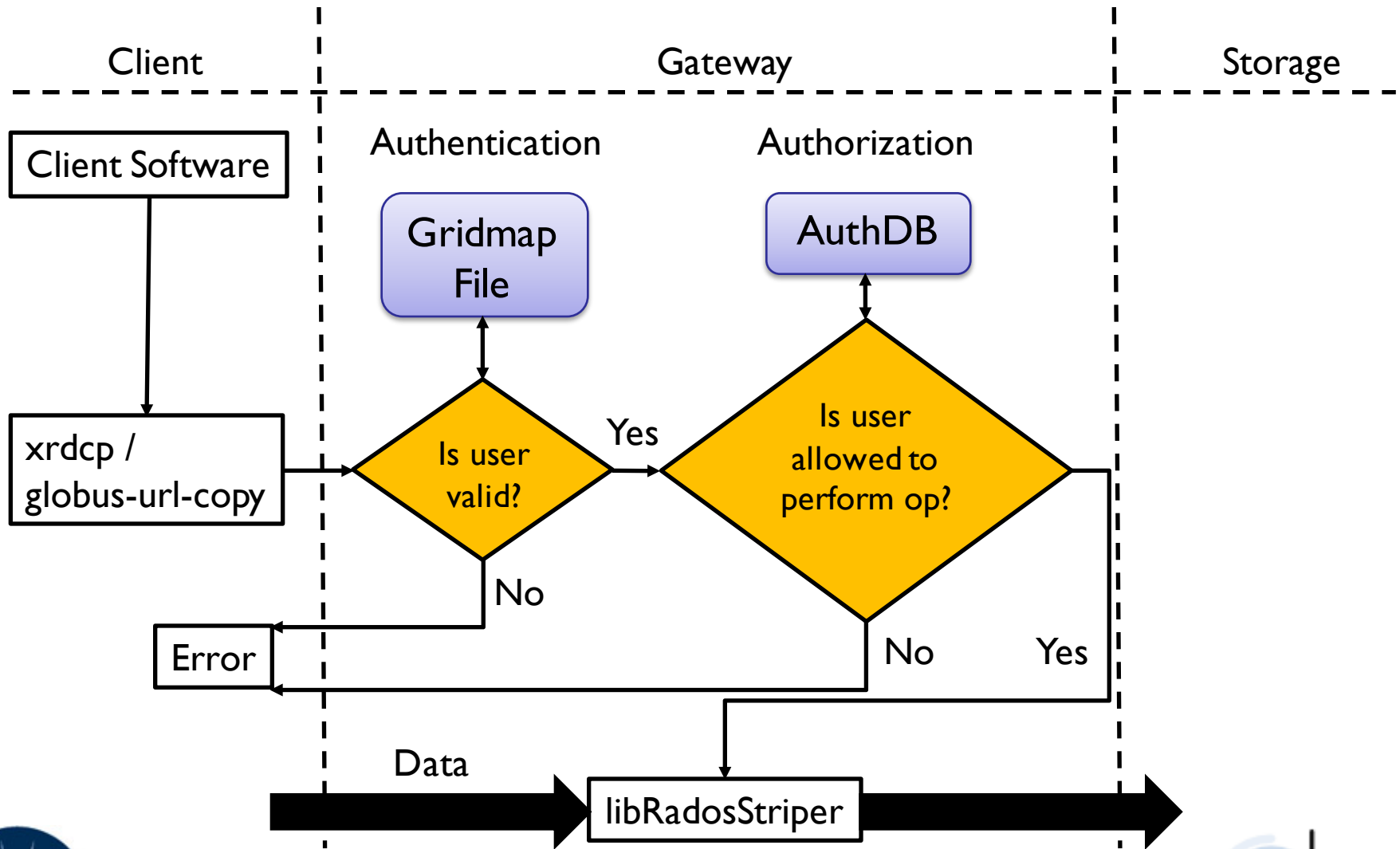
		k					
		8	10	12	14	16	
m	2	80	83	86	88	89	percentage of raw storage usable
	3	73	77	80	83	84	
	4	67	71	75	78	80	

 **Yahoo**
 **Facebook**
 **Tier 1 goal**



Plugin architecture

7



For XrootD the Gateway could be the WN

Alastair Dewhurst, 13th June 2016



XrootD Implementation

grid-mapfile

"/C=UK/O=eScience/OU=CLRC/L=RAL/CN=alastair dewhurst" atlasprod

authDB

u atlasprod /test a

Pool needs to start with a / for AuthDB to accept it

Test#	Grid user mapping	RADOS user	Pool	Object	Result	Notes
	Command					
1	atlasprod	xrootd	/test	/128m5	SUCCESS	-
	xrdcp 128m root://ceph-gw2.gridpp.rl.ac.uk//test:/128m5					
2	atlasprod	xrootd	/test	/128m5	FAILED	Ran right after test #1. Error 3006; file exists
	xrdcp 128m root://ceph-gw2.gridpp.rl.ac.uk//test:/128m5					
3	atlasprod	xrootd	/test	128m5	SUCCESS	-
	xrdcp 128m root://ceph-gw2.gridpp.rl.ac.uk//test:128m5					
4	atlasprod	xrootd	rep2	128m56	FAILED	Error 3010; permission denied. Pool rep2 exists but cannot be placed in AuthDB because it doesn't start with a /
	xrdcp 128m root://ceph-gw2.gridpp.rl.ac.uk/rep2:128m56					
5	atlasprod	xrootd	rep2	128m57	FAILED	Error 3010; permission denied. Changed AuthDB: s/atlasprod/atlasanalysis.
	xrdcp 128m root://ceph-gw2.gridpp.rl.ac.uk//test:128m57					



Pool architecture

- The LHC VOs have a similar model for Grid storage:
 - One “space token” for central data – Users can read but not write
 - One “space token” for users.
- 2 possibilities:
 - Create pool for each “space token”.
 - Create pool for each VO and use AuthDB to determine “space token”
- Fewer pools are easier to manage.
 - Some basic accounting information is needed.



GridFTP plugin status

- Development started by Sebastien Ponce and Brian Bockelman.
- Development continued by Ian Johnson
 - <https://github.com/stfc/gridFTPCephPlugin>
- Authorization development to follow XrootD model.

Command	Reason
MKD – pretend to make a directory	FTS wants to create the parent directory (/directories) of a non-existent PUT target
STAT – return fake stat() information for '/'	FTS wants to find the status of root directory, '/'
DELE	Delete an object
CKSM – return the object attribute containing ADLER32 checksum	Support transfer commands which request the checksum stored for an object



GridFTP Performance

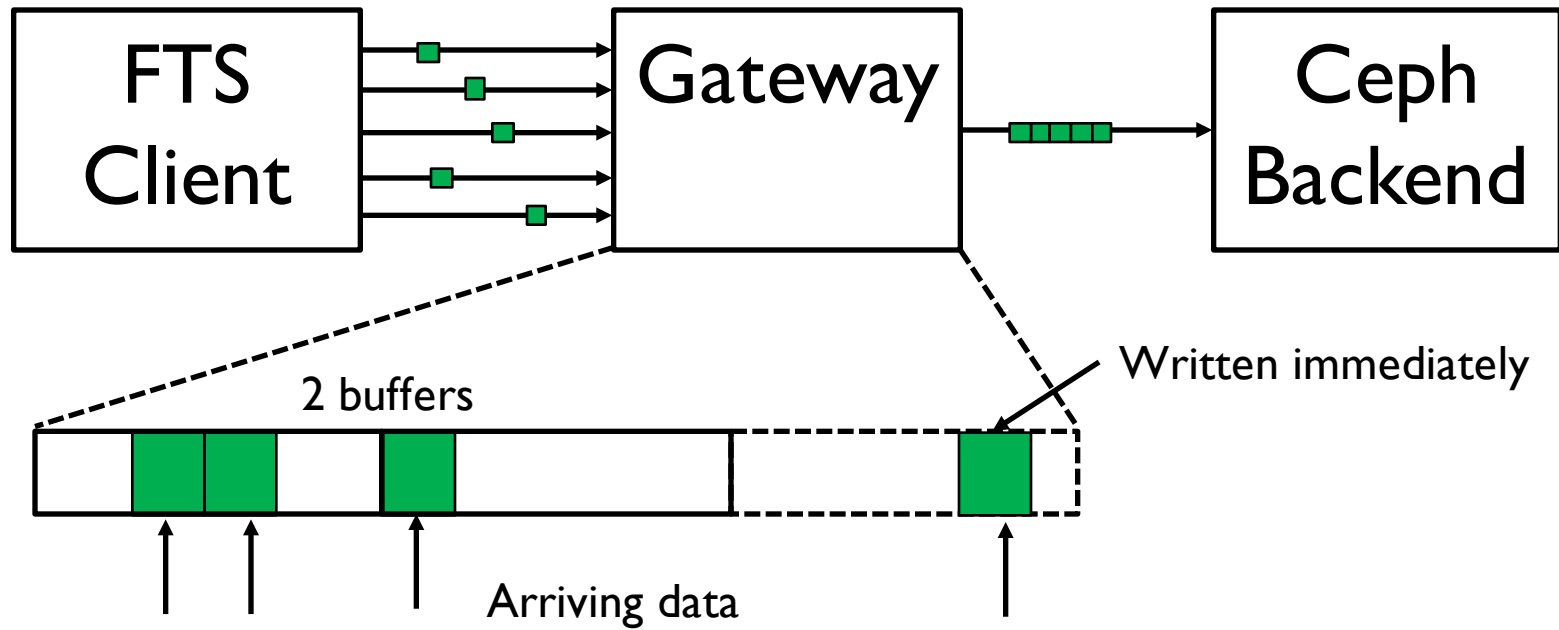
	Rados get/put	Globus-url-copy	FTS transfers
Reads	~80MB/s	~80MB/s	6-7MB/s
Writes	~80MB/s	~80MB/s	6-7MB/s

- Acceptable performance if using Globus-url-copy directly (Streaming).
- Poor performance using FTS (block mode).
 - Tried to get larger chunks to be sent but this is not viable over WAN.
- Brian Bockelman visited RAL last week.
 - Significant progress in design work



GridFTP design

12



- Patch from Brian Bockelman's to ensure arriving data chunks are not too far apart.



GridFTP alternative plan!

- On the plane home Brian wrote some different code!
- https://github.com/bbockelm/gridFTPCephPlugin/tree/async_writes
- Idea is to do all writes asynchronously and in parallel.
 - Only at close() time do you actually synchronize and wait for all outstanding writes to either finish or fail.
 - By issuing an asynchronous write per block received from GridFTP, you may be able to provide libradosstriper with sufficient parallelism to take full advantage of the underlying hardware.
- Cleaner approach if it works
 - Better to let someone else manage tricky buffer code.



Leading /

- There are several different issues with leading /.

1. Unable to delete paths (with XrootD) not starting with a /

- <https://github.com/xrootd/xrootd/issues/314>

2. XrootD authDB requires / else path is considered a template.

- Work around means creating a pool starting with a /.

3. GridFTP plugin adds a leading /.

- Ian Johnson has removed this from plugin.



RadosGW

15

- RAL provide RadosGW.
 - Open for developers although still partially behind a firewall.
 - ATLAS Event Service jobs use it.
- Want to encourage people to use S3 / Swift as it is much easier for site to support.
 - Does anyone develop for Swift?
- Two problems:
 - Integrate with Grid security (x509).
 - Integrate with FTS to allow transfers to other site SE.



Dynafed

- Umbrella name for multiple services/products:
 - http://svnweb.cern.ch/world/wsvn/lcgdm/ugr/trunk/doc/whitepaper/Doc_DynaFeds.pdf
- Allows FTS transfers to S3 endpoints:
 - Third party should work for Webdav enabled DPM and dCache sites.
 - Other SE use FTS service as proxy (RAL is running separate test FTS instance for this).
- Dynafed instance can act as an authorization layer above S3.
 - Grid proxy → pre-signed URL.



Summary

17

- Hardware available for new cluster.
- XrootD works.
- Active GridFTP development
- Exploring possibilities with VOs for S3/Swift.



Backup

18



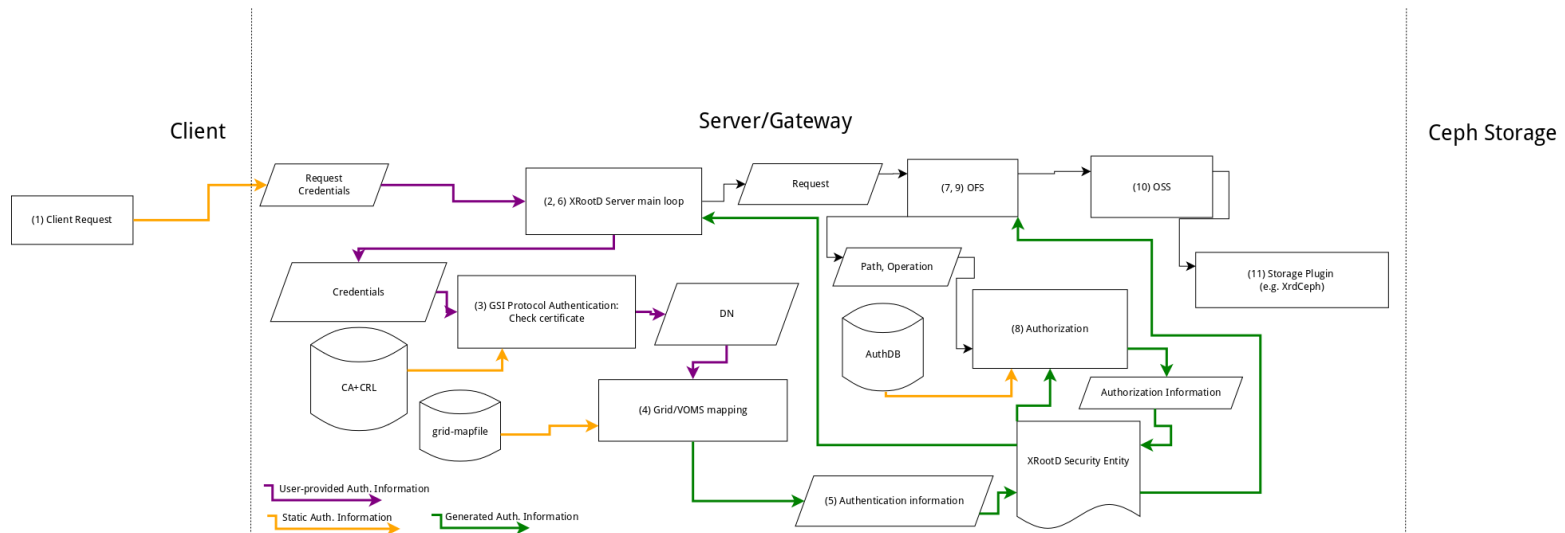
Alastair Dewhurst, 13th June 2016



Hardware

- 3 × monitor nodes: Dell R420, RAM: 64GiB, CPU: 2 × Intel Xeon E5-2430v2, 6 core, 2.50GHz.
- 3 × gateway nodes: Dell R430, RAM: 128GiB, CPU: 2 × Intel Xeon E5-2650v3, 10 core, 2.30GHz.
- 63 × storage nodes: XMA (Supermicro X10DRi), RAM: 128GiB, CPU: as gateways, OS Disk: 1 × 233GiB SSD, Data Disks: 36 × 5.46TiB HDD (WD6001F9YZ) via a SAS HBA.
- Total Raw Storage = 12.1PiB, 13.6PB.

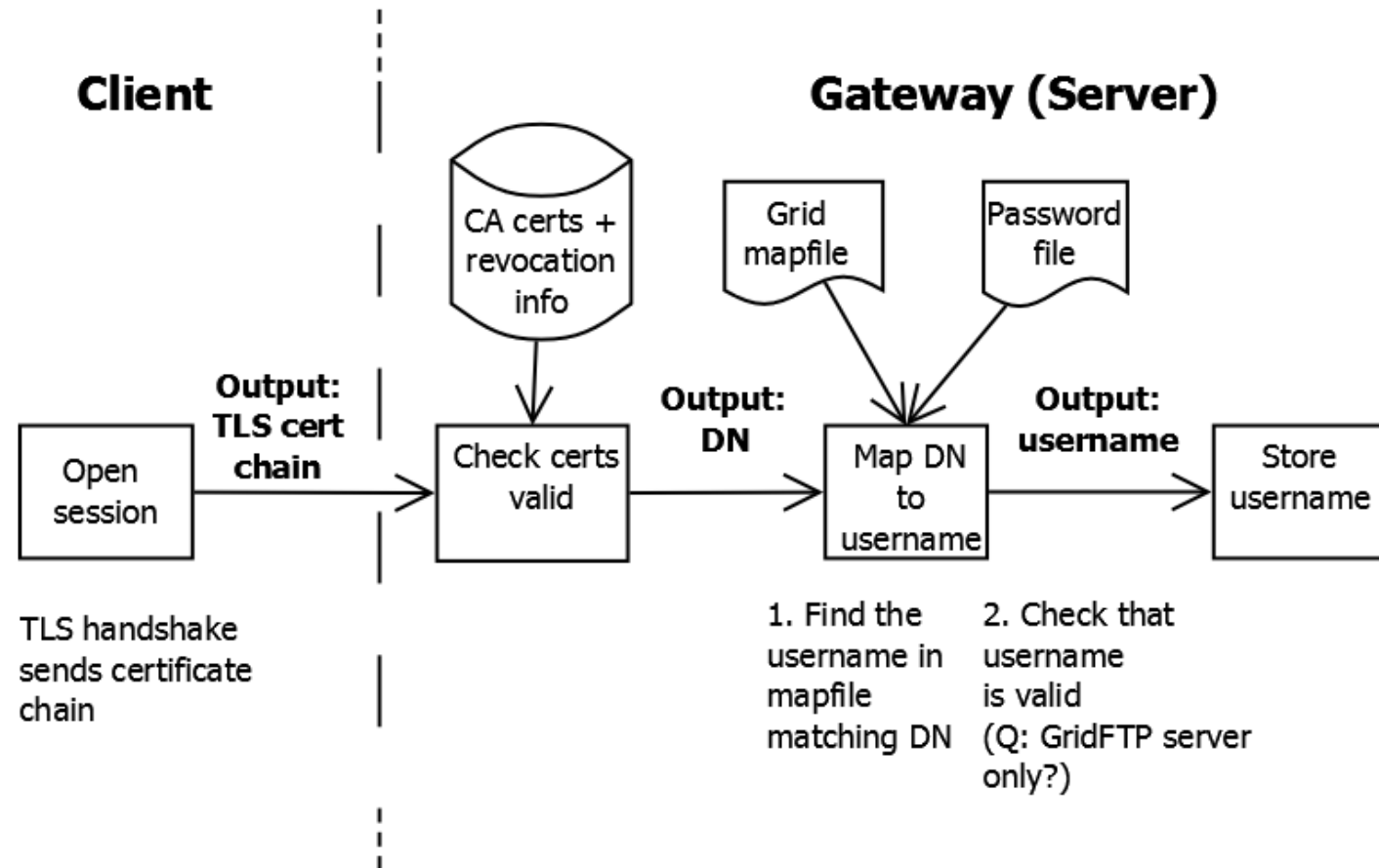




Backup

21

1. GSI Authentication - performed at session start-up



Backup

22

2a. Current GridFTP authorization - performed for every operation

