

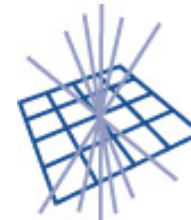
Taming the protocol zoo

Wahid Bhimji

GDB 14th January 2015



THE UNIVERSITY *of* EDINBURGH



GridPP
UK Computing for Particle Physics

Introduction

<https://indico.cern.ch/event/319817/>

- Very active participation – impossible to reliably summarise (and done quickly; and not as interesting) – so see original slides
- Michel's excellent notes –

<https://twiki.cern.ch/twiki/bin/view/LCG/GDBMeetingNotes20150113>

Central Topics:

- Whether we are on track to allow non-SRM disk-only sites within WLCG in Run 2
- Reducing / limiting the overall number of services needed by WLCG storage (so-called 'protocol zoo')

History – non-SRM sites

- Recommendation from Storage/Data ‘TEGs’ (see backup)
 - And followed up in ‘Storage interfaces (SI)’ working group
- Annecy pre-GDB (Oct 2012!) summarized quite well:
 - <https://indico.cern.ch/event/208241/>
 - Not designing an replacement to SRM but bring together people and activities to communicate and guide
 - Needed activities were defined (see backup).. SI group became just one part of more general ‘data’ activity. Some short updates since e.g. [May 2014 ‘data access’ GDB](#)
 - Things have progressed in that time ...
 - FTS3, Federations, gfal2, Rucio all in production; and Davix available
 - Much development for all activities in table done ..
 - BUT largely **not pushed into production use**




















Protocol Zoo – History and problem

- Linked somewhat with SRM item - sites currently still asked for SRM as well as new protocols.
- E.g DPM/ATLAS site currently needs SRM, gridftp, xrootd, rfio, WebDav..
- Also increase interest in (for example) Ceph– and don't want to develop (too many) plugins for different wlcg specific protocols.
- That's a site perspective but also load on storage system providers and locks us in to storage systems with wlcg-specific effort.
- Also Experiments – e.g. ATLAS has expressed (in Dec Jamboree – see slides later) that zoo of local access protocols is something they could do without
 - This is a zoo across sites rather than within so it's slightly different concern.

The meeting

- Experiments – progress towards non-SRM (if applicable) and protocol requirements.
- Sites – protocol opinions – RAL as example but also other comments.
- Storage providers – current protocol development plus roadmap and opinions.
- Discussion -> Action Plan -> WLCG Ops

13:30 - 17:40

13:30	Taming the protocol zoo	15'	▼
	Speaker: Wahid Bhimji (University of Edinburgh (GB))		
	Material: Slides  		
13:45	CMS - protocol use, needs and plans	15'	▼
	Speaker: Dr. Tony Wildish (Princeton University (US))		
	Material: Slides  		
14:05	LHCb - protocol use, needs and plans	15'	▼
	Speaker: Philippe Charpentier (CERN)		
	Material: Slides 		
14:25	ATLAS protocol use, needs and plans	15'	▼
	Speakers: Alessandro Di Girolamo (CERN), Wahid Bhimji (University of Edinburgh (GB))		
	Material: DraftTwikiWithActions  Slides 		
	SlidesFromAtlasDecJamborree 		
14:45	Site perspectives: including plans for proposed or current non-SRM sites: RAL; Cern..	15'	▼
	Speaker: Shaun De Witt		
	Material: Slides 		
15:05	StoRM - protocol developments and plans	10'	▼
	Speakers: Andrea Ceccanti, Andrea Ceccanti (CERN)		
	Material: Slides 		
15:20	Coffee / Tea	20'	
15:40	Ceph - protocols	10'	▼
	Speaker: Dan van der Ster (CERN)		
	Material: Slides 		
15:55	Xrootd - protocol developments and plans	10'	▼
	Speaker: Lukasz Janyst (CERN)		
	Material: Slides 		
16:10	DPM - protocol developments and plans	10'	▼
	Speaker: Oliver Keeble (CERN)		
	Material: Slides  		
16:25	dCache - protocol developments and plans	10'	▼
	Speaker: Paul Millar (Deutsches Elektronen-Synchrotron (DE))		
	Material: Slides  		
16:40	EOS and Castor protocol developments and plans	10'	▼
	Speaker: Mr. Andreas Joachim Peters (CERN)		
	Material: Slides 		
16:55	FTS / GFAL2 / Davix	10'	▼
	Speaker: Alejandro Alvarez Ayllon (CERN)		
	Material: Slides  		
17:10	Discussion, Solution and Action Plan	20'	▼

Experiments

- [CMS](#): Baseline gridFTP(FTS) and xrootd
 - Do now need deletion (by user) (and download).
 - Could use gfal2 with non-SRM if shown to work..
 - Can use non-SRM but site must discuss proposal and not many / none currently in production
- [LHCb](#): Currently use SRM but can use xrootd (or http) for 'url creation' and 'replication'
 - REQUIREMENT for stable single local redirector
 - Expect performance/ efficiency (e.g. balanced gridftp)
 - Need a way to select service class on T1s.. Could use namespace (as others do) but not practical / possible for existing disk/tape shared services (eg. RAL).

Experiments/ Sites

- ATLAS

- See advantage in rationalisation and presented a possible short/medium/long action plan. No hard deadlines – needs site/storage help
- Metadata ops (use SRM) : exploring WebDav for deletion; can use adhoc tool for query (e.g. json file method).
- Once this works and if gridFTP-only 3rd party and xrootd/dav download works then could allow (and encourage) non-SRM (non-tape) sites
- Zoo for local access. Would like this to move to xrootd / file in ‘short’ term

- Sites

- Sites increasingly need to serve other communities (gridFTP is widely used..)
- RAL “looking to deploy a more modern disk-only storage system” – based on object storage
- Want file put through protocol X –directly accessible by another.

Storage systems

- StoRM:
 - Improvements in http/Dav implementation – in testing at beta sites now
 - Future focus in Dav and non-GPFS quota support and space report via Dav. Interest in non-SRM (http/Rest) BringOnline
- Ceph
 - Scalable / useful object storage – offers various access protocols: RADOS, RBD, S3 and SWIFT, Posix CephFS – but *non-overlapping*:
 - E.g. Files written via S3 gateway cannot be directly accessed with RADOS
 - For WLCG data use needs another layer
 - CephFS could change / facilitate this but non-yet prod. ready.

Storage Systems

- DPM:

- 'Proactively support' HTTP/WebDAV, Xrootd, GridFTP
- Legacy inc SRM; rfiio – keep as is but like to deprecate
- Todo: rollout gridFTP redir; Implement Dav quota

- dCache:

- SRM investment; unique features; battle tested
- 3rd party copy have http support; no plan for xrootd
- For non-unique *propose* Accounting: Dav; WAN transfer: Dav; Local access: NFS; 3rd party:GridFTP; Namespace operations: SRM (bulk) or Dav (small)

Storage systems

- Eos

- Primary: Xrootd, Secondary: gridFTP/SRM; FUSE; http/https/WebDav/S3/OwnCloud http.
- Love to get rid of SRM; Performance benefit in xrootd (on EOS - from implementation). Deficits of HTTP reflected by standard extensions (none of these do provide all XRootD gridFTP semantics)
- no new protocols foreseen

- Castor

- Xrootd/gridFTP – Main protocols
- Rfio deprecated 2015, dropped 2016 (still ~15% use)
- SRM decommissioning in CASTOR would need planning and guideline how to provide space accounting

Storage systems

- Xrootd:

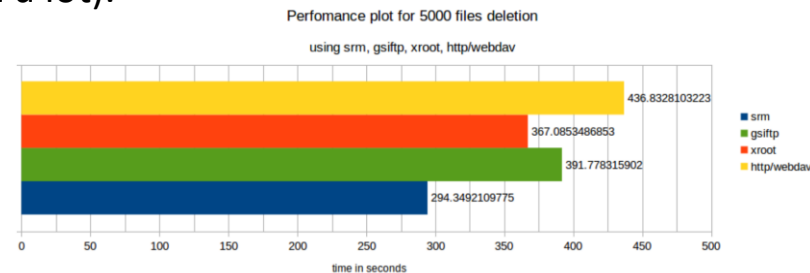
- 4.1 bring cross-protocol requests (start with root:// end up with file:// or http://) - works with ROOT 5.34/25
- 4.2 will bring CEPH/Rados plugin and Throttling

- FTS / GFAL2 / Davix

- Davix getting mature. Dav / Xrootd similar performance
- SRM deletion *is* quicker than anything else for bulk (> O(100)).
 - Need pipeline in http (even on client may win a lot).
- FTS gridFTP bulk transfers coming.

- Gfal2

- Lcg-util *fully deprecated*
- Supports srm, xrootd, gsiftp, http/dav, s3, rfio, dcap, file, lfc
- Third party copy : xrootd, gsiftp, http/dav (DPM and dCache partially)
- Checksum native on gridFTP, HTTP , xrootd (not on all storages)



Summary/Issues/Discussion/Actions

● Transfer

- 3rd party transfer:
 - GridFTP: still only feasible option for the short-term.
 - Xrootd or http: mix of support on storage. Should study / measure performance of these.
- Downloads SRM to Xrootd or http OK – but requirement on single/stable local redirector and c.f. other LHCb concerns

● Metadata

- Query used capacity OK: ‘adhoc’ solutions And DAV RFC 4331
 - Though in some cases storage development needed for namespace calculations (e.g. DPM/ non-GPFS StoRM) - in progress
- Deletion – need further study for SRM v Dav at bulk. For CMS use case, user deletion, either/any should be OK.

Summary/Issues/Discussion/Actions

- Local access
 - Transitions from rfio and dcap (to xrootd/nfs/http)(no brainer?)
 - BUT [email from Guenter Duckeck] Dcap and ATLAS sites – some issues when xrootd tried; nfs / davix still need operational proof.
 - Clarification from Patrick :

“although dcache.org would like to decommission dcap, certain preconditions must be met before that happens. One of which is that dCache can provide an alternative access protocol, sufficiently stable and performant. [...] As there is no pressure right now, the decision when to decommission dcap will be made by dcache.org in close collaboration with our customers”
 - Rfio – could decommission (as was already said in a previous GDB) but used internally in DPM anyway (work ongoing to remove).
 - Likely to be Xrootd and file in medium term.
- I will make a (new) table with storage system / protocol support for each required feature as well as what each experiment uses for each of these.

Discussion - miscellany

- Number of services + storages/sites where Dav suggested replacement
 - Therefore need “rollout of (existing) Http SAM test within WLCG” [Oliver] for those cases
- There was also mention of alternative for SRMBringOnline – interest in possibility expressed from CERN and StoRM.
 - Shouldn't tie with or confuse the disk-only issues here but an interesting point.

[My] Conclusion

- ‘No crisis’ [Markus] but do need to push for rationalisation [Me/Michel/Atlas[Ale]]
- I think continuing with current approach of (pre-)GDB discussion / reporting is OK
 - Could be more frequent tracking if need be.
 - Doesn't have to be me doing (or organising) it
 - Certain items (rfio – now ; dcap (medium-term), Dav space reporting (mid-term)) could move to WLCG ops
 - Other issues like local redirector could be raised there to
 - Longer term SRM (or gridFTP->Dav) transition could also be pushed there, but tests on e.g. dav deletion needed

BACKUP SLIDES - HISTORY

Annecy GDB –development areas at the time Updates in Red

<https://indico.cern.ch/event/155073/other-view?view=standard>

Needed by?	Issue	Solution
ATLAS/ LHCb	Reporting of space used in space tokens.	JSON publishing currently used in some places on ATLAS – probably temporary. WebDav quotas? Use RFC 4331 (reporting). Calculating per directory also needs action.
ATLAS/ LHCb	Targeting upload to space token.	Could just use namespace but certain SEs would need to change the way they report space to reflect. (Or use e.g. http) SEs are enabling – but ongoing as above.
ATLAS/LHCb	Deletion	gFal2 will help. gFal2 in production. ATLAS deletion via dav in Rucio (not tested)
LHCb (ATLAS)	Surl->Turl	Require a redirecting protocol and SURL = Turl for sites that want no SRM. See LHCb update
Any?	Checksum check – confirm not needed?	Some service query is needed by ATLAS – as is some “srm-ls”. gFal2 will help (OK for gridftp)
All?	pure gridFTP on different storage types	DPM at least willing to look at this. Enabled for DPM (not used in prod.). dCache had it already

TEG recommendation reminder

- ☐ Maintain SRM at archive sites
- ☐ Experiments, middleware experts and sites should agree on alternatives to be considered for testing and deployment, targeting not the full SRM functionality but the subset detailed here, determined by its actual usage.
 - ... recommend a small working group be formed, reporting to the GDB, to evaluate alternatives as they emerge
- ☐ Develop future interfaces:
 - ... different approaches to integrate cloud-based storage resources need to be investigated ...

Table of used functions from TEG

	Is this feature used by ...				Tier	SRM function ²
	Atlas	CMS	LHCb	FTS only		
Transfer Management						
Upload / download a complete file	Yes	Yes	Yes	No	All	srmPrepareToPut/Get//Put/GetDone
Manage transfers.	Yes	Yes	Yes	Yes	T1/2	srmAbort/Suspend/ResumeRequest
Balance over multiple transfer servers.	Yes	Yes	Yes	Yes	T1/2	srmPrepareToGet ³
Manage third-party copy	Yes	Yes	Yes	Yes ⁵	T1/2	
Negotiating a transport protocol	No	No	No			srmGetTransferProtocols
Namespace Interaction						
Querying information about a file (stat)	No	No	Yes ¹	Yes ⁶	T1/2	srmLs
Upload data integrity information (chksums)	No	No	No	No	T1/2	
Check integrity information	Yes	Yes	Yes	Yes		srmLs
Creating/Deleting data and directories	Yes	Yes	Yes ¹	Yes ⁷	All	srmMkdir srmRmdir srmRm srmMv
Changing ownership, perms and ACLs	No	No	No	No	-	srmSet/Check/GetPermission
Storage Capacity Management						
Query used capacity (like df)	Yes	No	Yes	No	T1/2	srmGetSpaceMetaData/Tokens
Create/remove reservations; assign characteristics	No	No	No	No	-	srmReserve/Update/ReleaseSpace
Targeting uploads to specific reservation	Yes	Yes	Yes	No	T1/2	srmPrepareToPut
Moving files between reservations	No	No	Yes	No	T1/2	srmChangeSpaceForFiles
Server Identification						
Test service availability and information	Yes	Yes	No	No		srmPing

- Somewhat simplified and removed those only relevant for Archive/T1
- A couple of observations:
 - Not that much is needed – e.g. space management is only querying and not even that for CMS

Brief functionality table from Annecy pre-GDB: (see also LHCb talks)

Function	Used by ATLAS	CMS	LHCb	Is there an existing Alternative or Issue (to SRM)
Transfer: 3 rd Party (FTS)	YES	YES	YES	Using just gridFTP in EOS (ATLAS) and Nebraska (CMS) What about on other SEs?
Transfer: Job in/out (LAN)	YES	YES	YES	ATLAS and CMS using LAN protocols directly
Negotiate a transport protocol	NO	NO	YES	LHCb use lcg-geturls;
Transfer: Direct Download	YES	NO	NO	ATLAS use SRM via lcg-cp, Alternative plugins in rucio
Namespace: Manipulation / Deletion	YES	YES	YES	ATLAS: Deletion would need plugin for an alternative
Space Query	YES	NO	YES?	Development Required
Space Upload	YES	NO	YES?	Minor Development Required