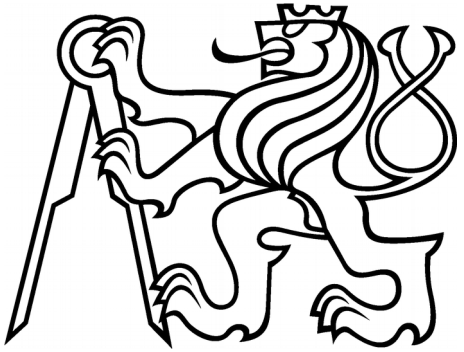


Third-party-copy alternatives to GridFTP

Petr Vokáč

NEC2019

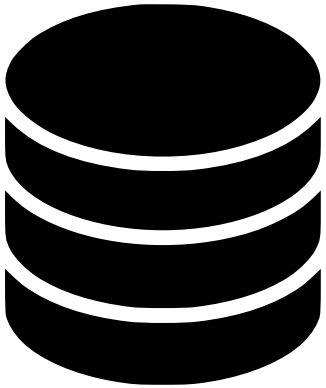
2nd October 2019

EUROPEAN UNION
European Structural and Investment Funds
Operational Programme Research,
Development and Education

MINISTRY OF EDUCATION, YOUTH AND SPORTS

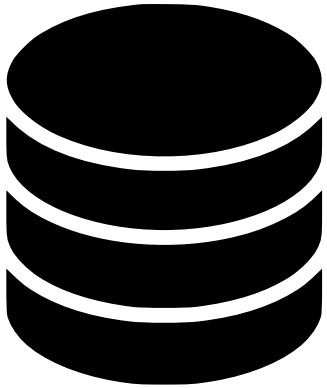
Third-party-copy

Source



party 1

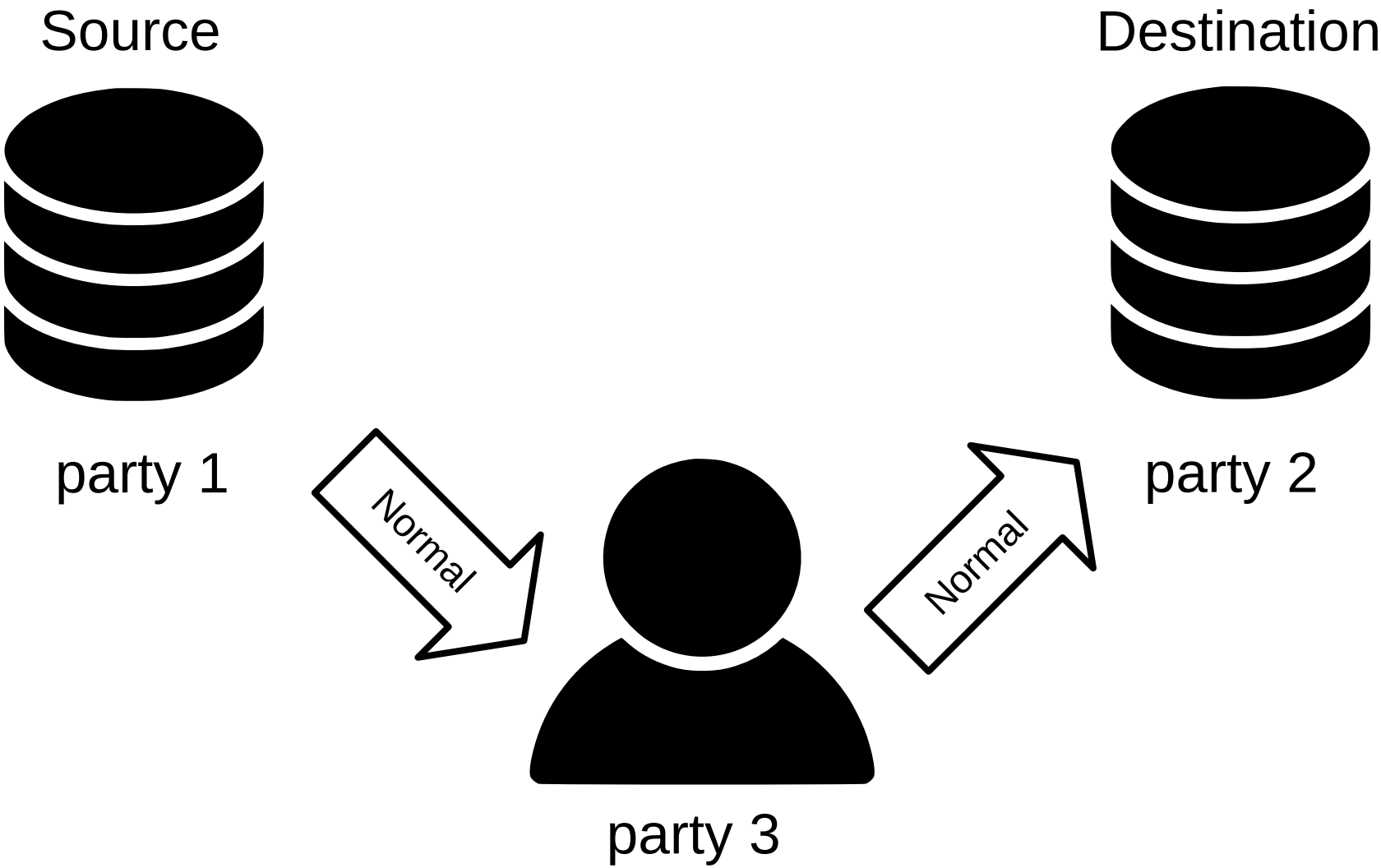
Destination



party 2

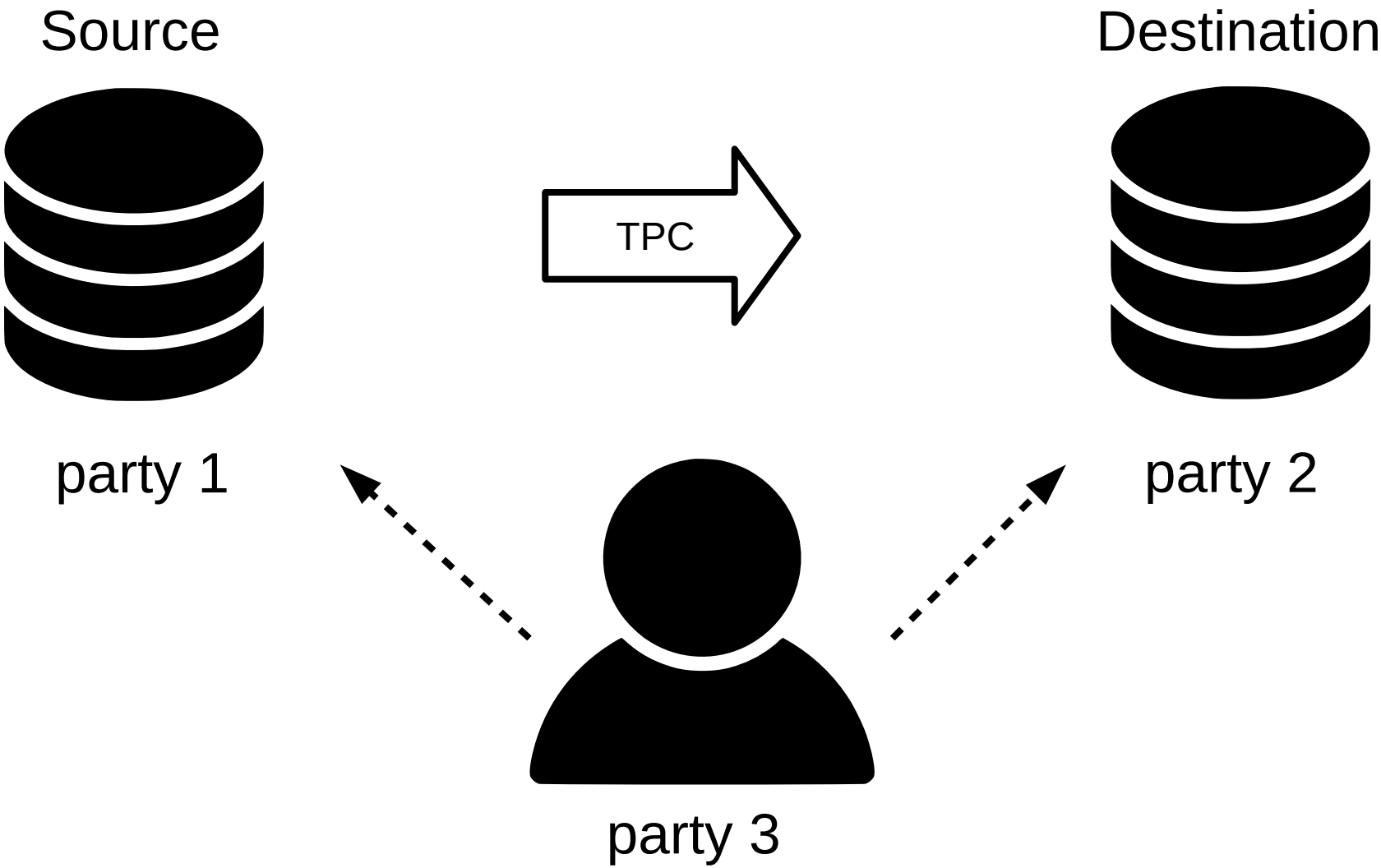
party 3

Third-party-copy



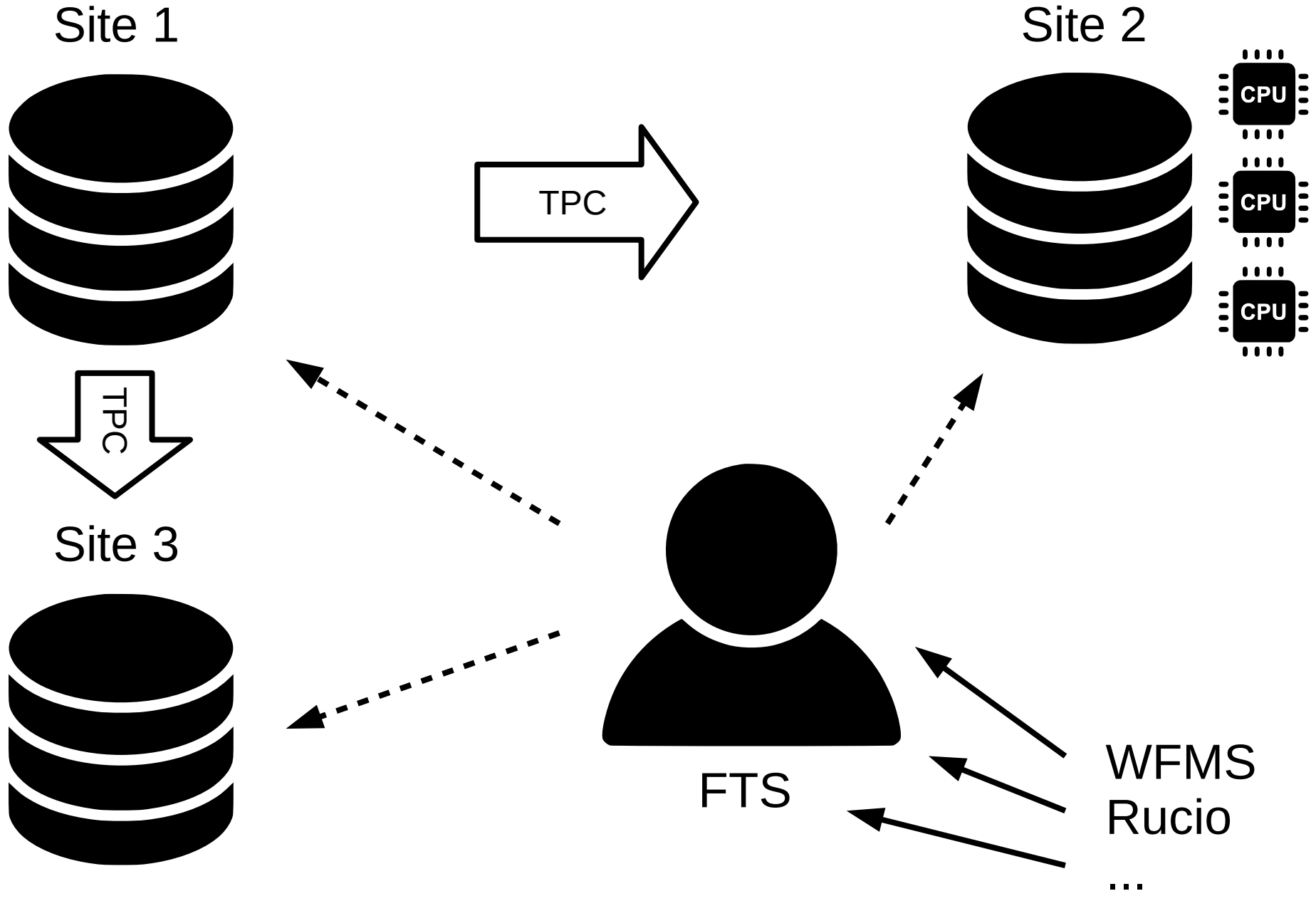

```
xrdcp root://party1/file root:/party2/file
gfal-copy -D "HTTP PLUGIN:DEFAULT_COPY_MODE=streamed" \
    davs://party1/file davs://party2/file
```

Third-party-copy



xrdcp -tpc only root://party1/file root:/party2/file
gfal-copy root://party1/file root://party2/file
gfal-copy davs://party1/file davs://party2/file

Third-party-copy



GridFTP status

- Supports TPC since ~ 2005
- Used almost exclusively for production TPC transfers
- General support for Globus Toolkit ended in 2017
  - Commercial Globus Connect
  - Grid Community Toolkit
    - fork of original open source Globus Toolkit
    - maintain existing tools including GSI and GridFTP
      - support at least till 2021
      - significant effort necessary with each OpenSSL ABI changes
  - catalyst to think about modernizing whole storage infrastructure
- WLCG DOMA working group
  - Access, content delivery and caching
  - QoS
  - TPC

WLCG DOMA – TPC

- TPC subgroup – find alternative protocol(s) for GridFTP
  - ***Phase 1*** (end of 2018) – survey replacement protocols available in common storage implementation, prototype / implement support TPC
  - ***Phase 2*** (mid 2019) – early deployment phase to ensure alternative protocol at all WLCG sites with > 3PB storage
  - ***Phase 3*** (end 2019) – widespread deployment when all WLCG storages must support non-GridFTP protocol
- GridFTP still considered for transfers between sites without matching alternative protocol
- Participants (developers, testers, site / storage admins)
  - XRootD, dCache, DPM, EOS, StoRM, Echo, also Rucio, FTS, gfal
- Related WLCG task forces
  - DPM Upgrade task force
    - GGUS ticket with request to upgrade submitted recently
  - dCache upgrade task force

WLCG DOMA – TPC

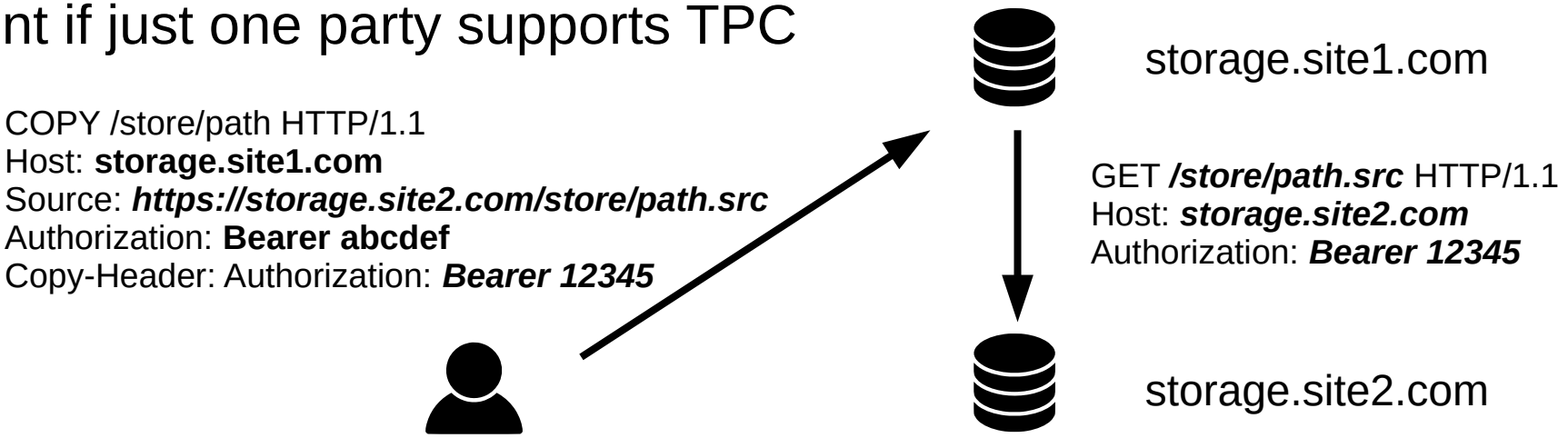
- Criteria for evaluating new protocols
  - Requirements
    - well documented (e.g. Open Grid Forum)
    - multiple implementation (necessary for standardization)
    - secure as GridFTP (as it is used by WLCG – no data encryption)
    - support multi-VO storage system
  - Desirable
    - improved security (stronger data integrity and privacy)
    - support universal endpoints (no VO specific gateway)
    - support for non-X.509 authentication (tokens)
    - support industry standard protocols (e.g. S3 via HTTP)
- Available alternative protocols already supported by storages
  - XRootD
  - WebDAV

XRootD TPC

- Basic support since ~ 2010
  - not sufficient for general use cases
  - missing support for credential delegation
    - local valid grid proxy necessary
    - not scalable and reliable enough
- XRootD 4.9.x with grid proxy delegation support
  - xrootd security protocol updated and documentation improved
  - implemented also by dCache 5.x (two implementation)
- TPC transfers
  - destination endpoint with delegated credentials pulls files from source
- Upcoming XRootD 5.x
  - encryption support
    - allows to specify which communication must be encrypted
    - including data transfer encryption

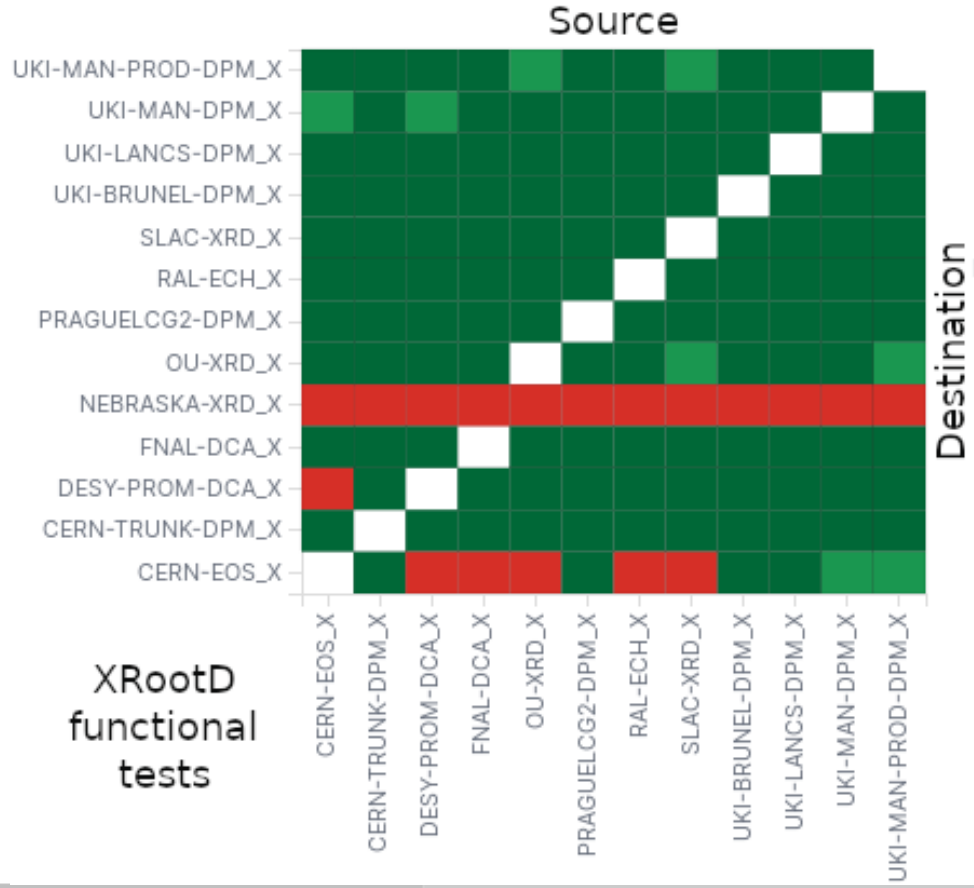
HTTP TPC

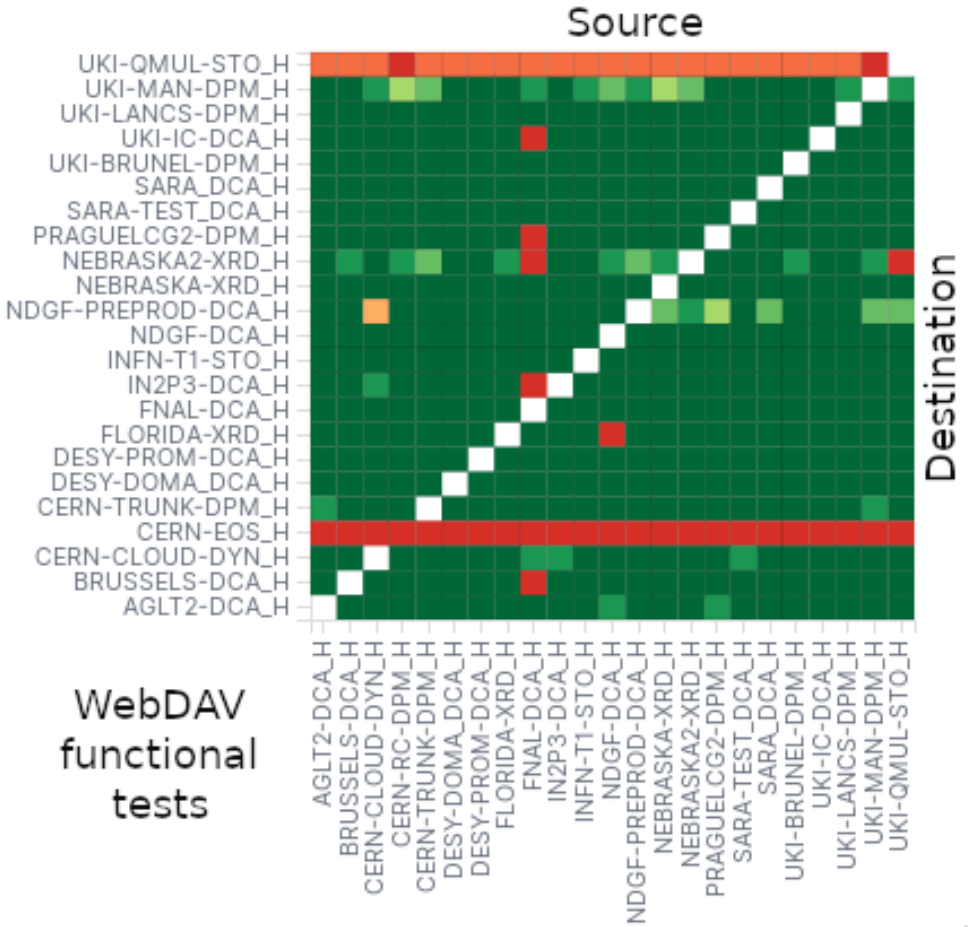
- Utilizing existing WebDAV “COPY” verb (RFC4918)
  - additional headers for AuthZ described in technical documentation
    - support for different AuthZ (gridsite proxy delegation, tokens)
    - performance markers for monitoring copy progress
    - communication finished with “success: Created” / “failure: msg“
  - implemented by DPM, dCache, StoRM, XRootD, Dynafed
- TPC transfers
  - pull mode – client ask destination to download data from source
  - push mode – client ask source to upload data to destination
  - sufficient if just one party supports TPC



TPC Functional Tests

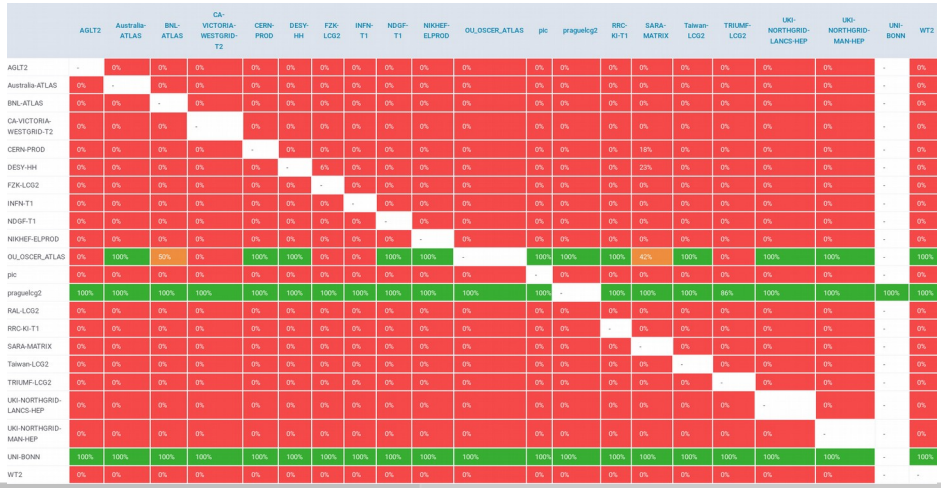
- Rucio dteam VO testbed for TPC transfers
  - all participating sites tested every hour with each other
  - Rucio transfer traces collected by MWT2 elasticsearch / kibana
    - filter failed transfers → provides link to FTS details
    - FTS with debug loglevel

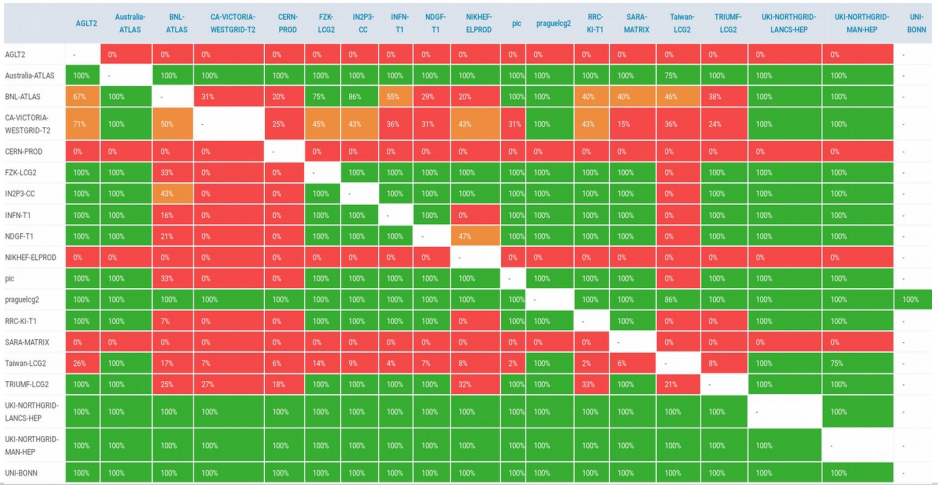




TPC Functional Tests

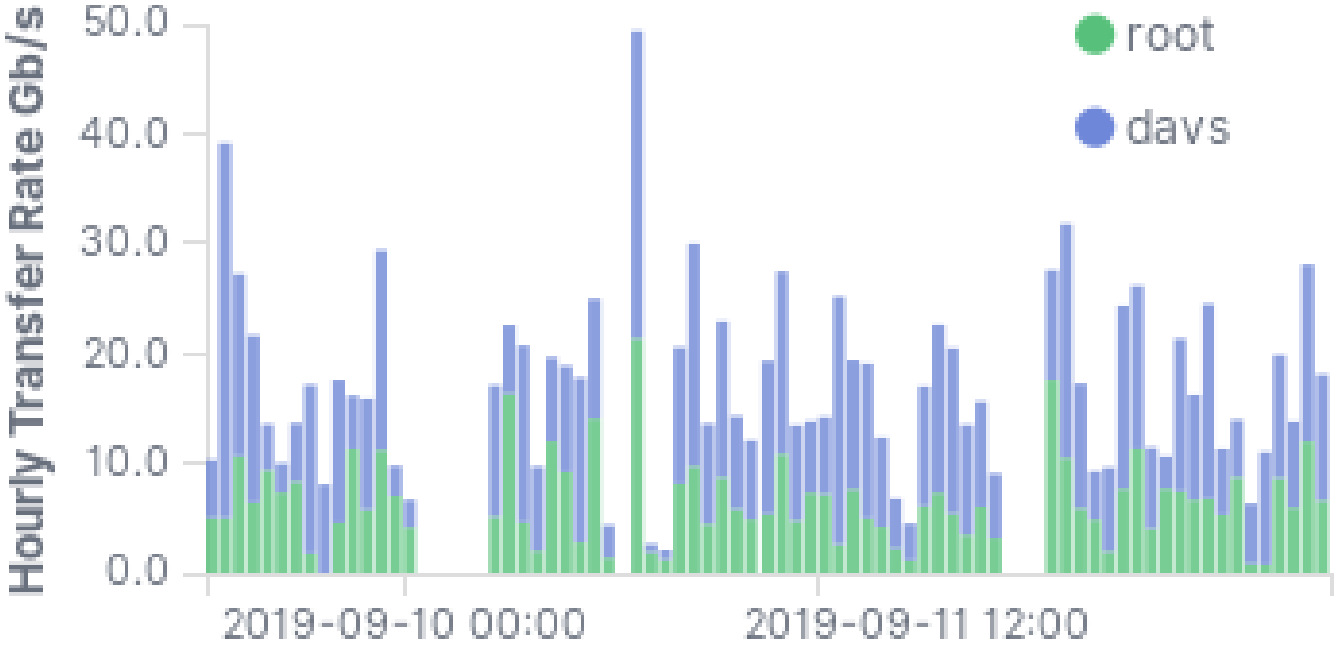
- ATLAS has own production transfer monitoring (NEC2019)
  - recently new functional tests for individual protocols
  - selected production endpoint probed with XRootD and WebDAV TPC
- Quite different picture from testbed
  - not all production FTS servers updated to version supporting XRootD
  - some FTS servers configured witch HTTP streaming
  - storage software with proper TPC support released recently (months)
  - number of storage endpoints without XRootD checksum support
    - checksum validation was disabled for DOMA TPC testbed





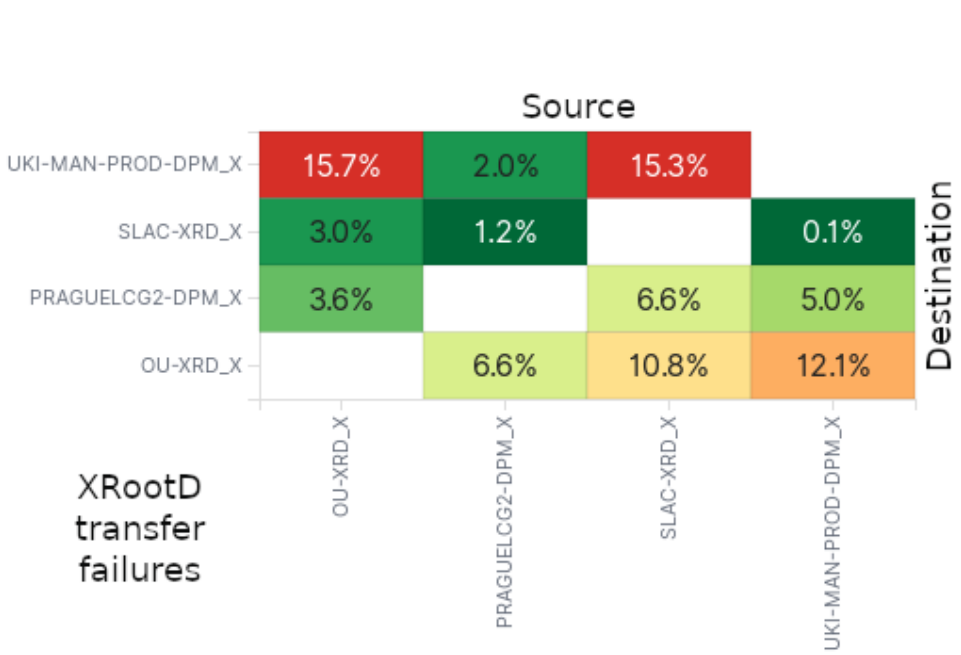
TPC Stress Tests

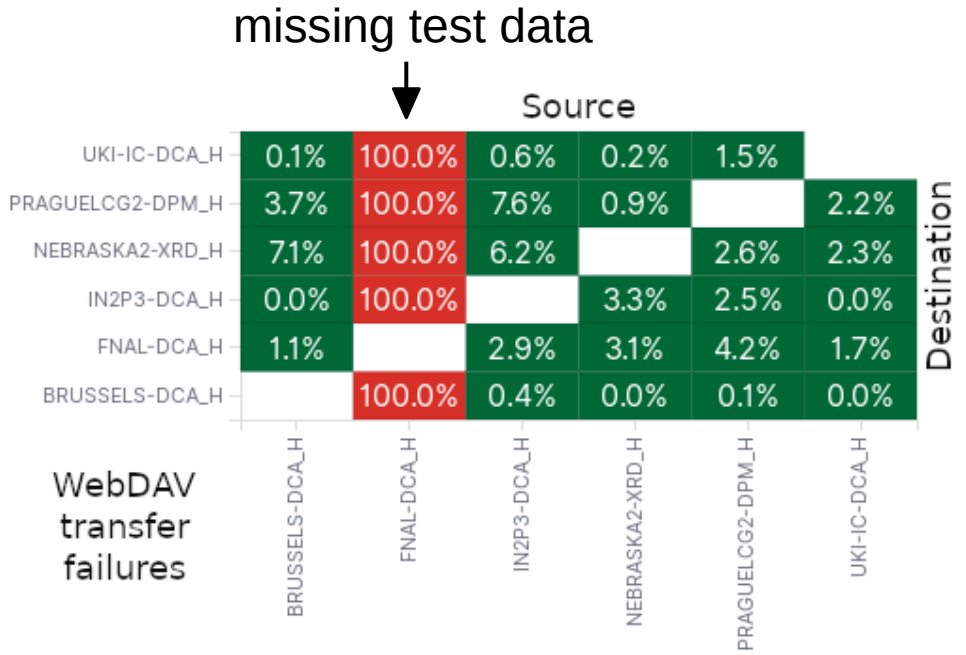
- Production endpoint stress tests with XRootD and WebDAV TPC
  - 250 transfers with 4GB files scheduled every hour between each site
  - 1.25PB transferred every week (more than 300k transfers)
    - still just ~ 5% of average transfer volume within single LHC experiment
    - reaching up to 50Gb/s hourly transfer rate
  - failure rate still needs to be better understood
  - throughput comparison GridFTP vs. XRootD vs. WebDAV not yet done



TPC Stress Tests

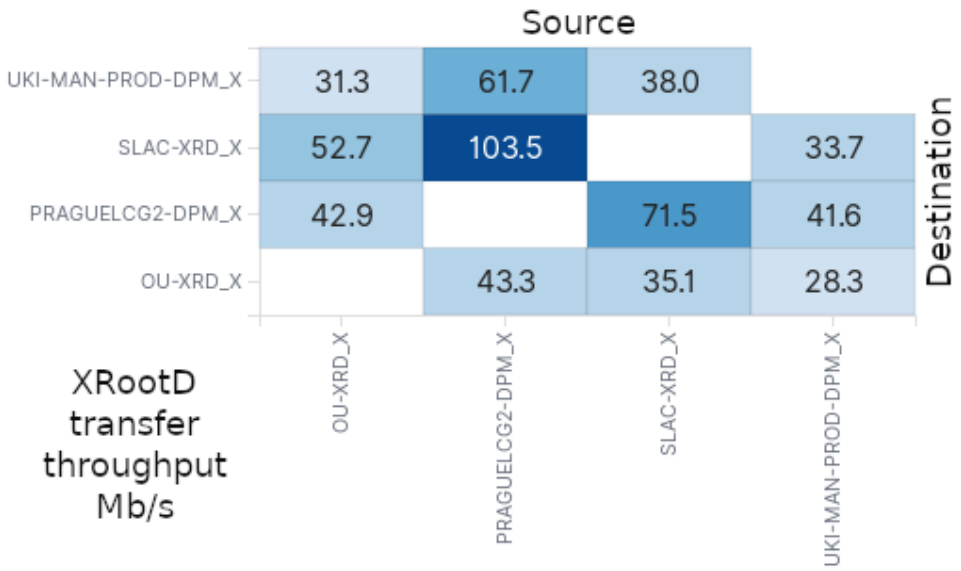
- Production endpoint stress tests with XRootD and WebDAV TPC
  - 250 transfers with 4GB files scheduled every hour between each site
  - 1.25PB transferred every week (more than 300k transfers)
    - still just ~ 5% of average transfer volume within single LHC experiment
    - reaching up to 50Gb/s hourly transfer rate
  - failure rate still needs to be better understood
  - throughput comparison GridFTP vs. XRootD vs. WebDAV not yet done

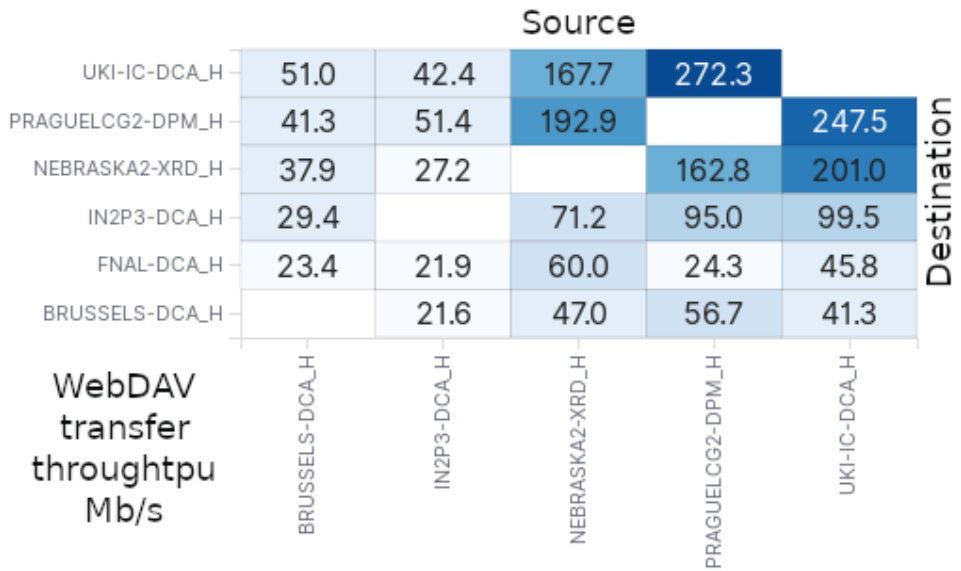




TPC Stress Tests

- Production endpoint stress tests with XRootD and WebDAV TPC
  - 250 transfers with 4GB files scheduled every hour between each site
  - 1.25PB transferred every week (more than 300k transfers)
    - still just ~ 5% of average transfer volume within single LHC experiment
    - reaching up to 50Gb/s hourly transfer rate
  - failure rate still needs to be better understood
  - throughput comparison GridFTP vs. XRootD vs. WebDAV not yet done





TPC Smoke Tests

- Available for both protocols – WebDAV and XRootD
- Provides much more diagnostic details about TPC storage support
  - test both pull and push mode for HTTP
  - testing compliance with TPC standard
  - different credential delegations
- Executed automatically every day
  - each storage participating in TPC testbed
  - all results sent by email to storage administrators
  - simple statistic collected including historical data
- Can be executed by site/storage admins
  - dteam VO X.509 proxy necessary to run smoke test
  - otherwise daily reports provides same info

Software with TPC support

- WLCG software baseline updated
  - minimal storage version with TPC support
- Storage baseline for TPC
  - XRootD: 4.10.0 (July 2019)
    - 4.11 brings fixes necessary for Echo (soon)
  - DPM: 1.13.2 (October 2019)
  - dCache: 5.2 (July 2019)
    - WebDAV TPC functional since 3.2
    - dCache < 4.2 already EOL
  - EOS: 4.5.6 (August 2019)
- Other software
  - gfal2: 2.16.3, gfal-utils: 1.5.3
  - davix: 0.7.2
  - FTS: 3.8.3

FTS

- TPC transfers (XRootD) requires at least 3.8.3
  - Current status
    - FTS prod @CERN → xrootd ok, http ok - pull & push
    - FTS devel @CERN → xrootd ok, http ok - pull & push
    - FTS pilot @CERN → xrootd ok, http ok - pull & push & streaming
    - FTS @BNL → xrootd ok, http ok - pull & push
    - FTS @RAL → xrootd plugin not installed, http ok - pull & push & streaming
    - FTS @FNAL → old version not supporting xrootd delegation, old gfal version, http ok - pull & push & streaming
- http streaming
  - fallback from TPC to normal copy
  - data transferred through FTS
- http pull & push – unexpected transfer “retry”
- http TPC was causing excessive logging with old dCache versions

Rucio

- Only third_party_copy activity in Rucio configuration
  - unable to distinguish configuration for active / passive party in TPC
  - non-GridFTP sites can’t work properly
    - workaround with distance config necessary to protect other transfers
  - Needs development and database schema update
    - postponed for next major release
    - Rucio coding camp
    - available in October
- Multi-hop support
  - necessary for CTA
  - two sites TPC protocols intersection empty → muti-hop
  - utilize existing FTS multi-hop functionality
    - temporary copy registered in Rucio (secondaries)
  - a lot of corner cases makes implementation non-trivial
  - ATLAS wants both protocols deployed to prevent multi-hop transfers

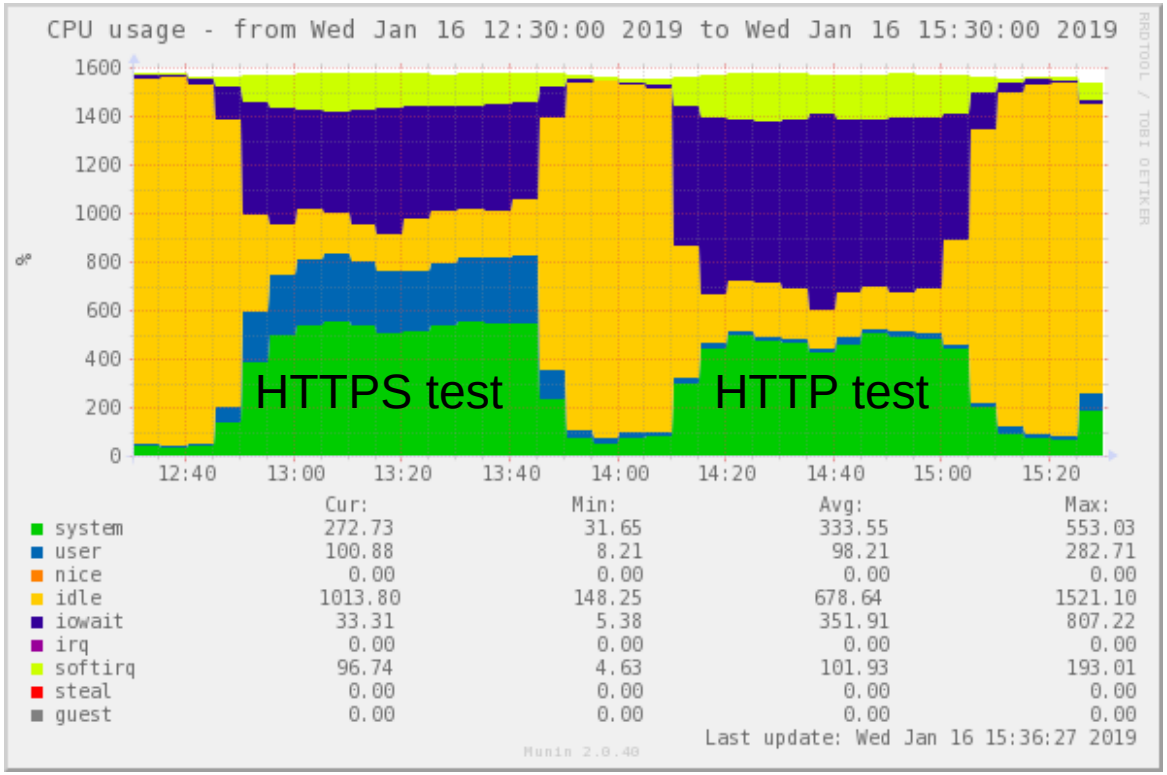
Data transfers encryption

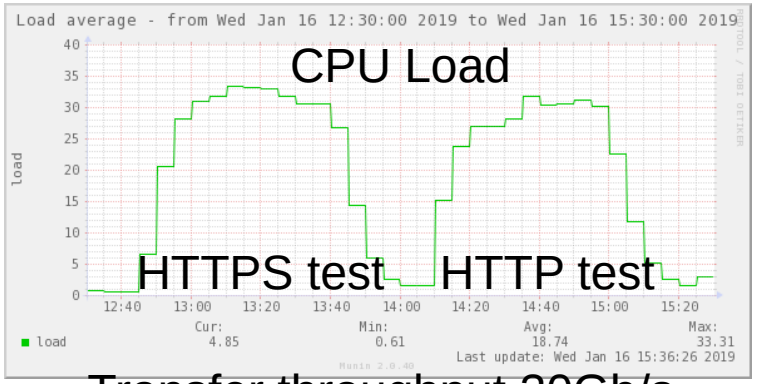
- In future all transfers will be probably encrypted
  - HTTPS is necessary for TPC
  - XRootD will come with data encryption soon
- Server CPU has build-in support for encryption – AES-NI
  - usually 1 encryption unit per physical core
  - 5Gb/s with single HTTPS connection on low-end modern CPU
    - 16 cores saturate easily 40Gb from mem
    - real file transfers limited by disks
  - 1Gb/s on our oldest storage servers
    - can become quite busy with 10Gb
- Less resources for BEER (NEC2017, CERN-IT-Note-2019-001)

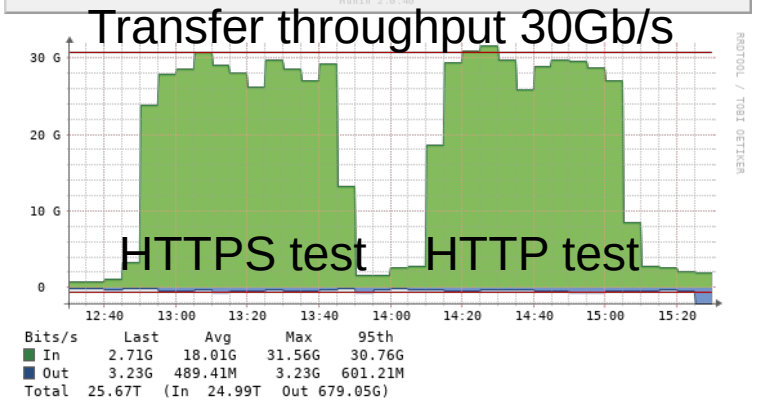
CPU	openssl	HTTPS one	HTTPS mem	HTTPS disk
2x8core Intel Silver 4108	279.8Gb	4.2Gb	40Gb on 40Gb	30.0Gb disk lim.
2x6core Intel E5-2620	77.7Gb	2.3Gb	10Gb on 10Gb	N/A
2x4core Intel E5620	8.6Gb	0.9Gb	N/A	N/A

AES-NI CPU utilization test – disk

- CPU utilization while reading 1GB files from disk and sending them using apache with average speed ~ 30Gb/s (limited by disk read throughput). HTTPS stream encrypted with TLSv1.2,ECDHE-RSA-AES256-GCM-SHA384,2048,256 vs. simple HTTP test
  - CPU load details from /proc/stat
  - one minute load average, network transfer throughput







Summary

- Third-party-copy now available for WebDAV and XRootD
- Implementation exists for all grid storage implementations
- Functional tests works between all implementation
- Stress tests in progress – already use production instances
- Only very recent storage releases provides sufficient TPC support
  - WLCG ask sites to upgrade (GGUS)
  - Provide at least one non-GridFTP protocol
- Most of FTS servers supports TPC with WebDAV and XRootD
- Rucio should be ready with next major version released in October
- Upgrading majority of storages → GridFTP could become optional
- WLCG DOMA activities not limited to TPC
  - caching, quality of services, authorization, storage organizations, …
  - more changes coming in near future, TPC TF continue with tokens

BACKUP

ATLAS Rucio TPC Configuration Test

- Rucio doesn’t currently properly support active / passive TPC party
  - ATLAS configuration often fails TPC with lower priority protocols
  - doesn’t affect production transfers that always use SRM/GridFTP
    - protocol with highest priority

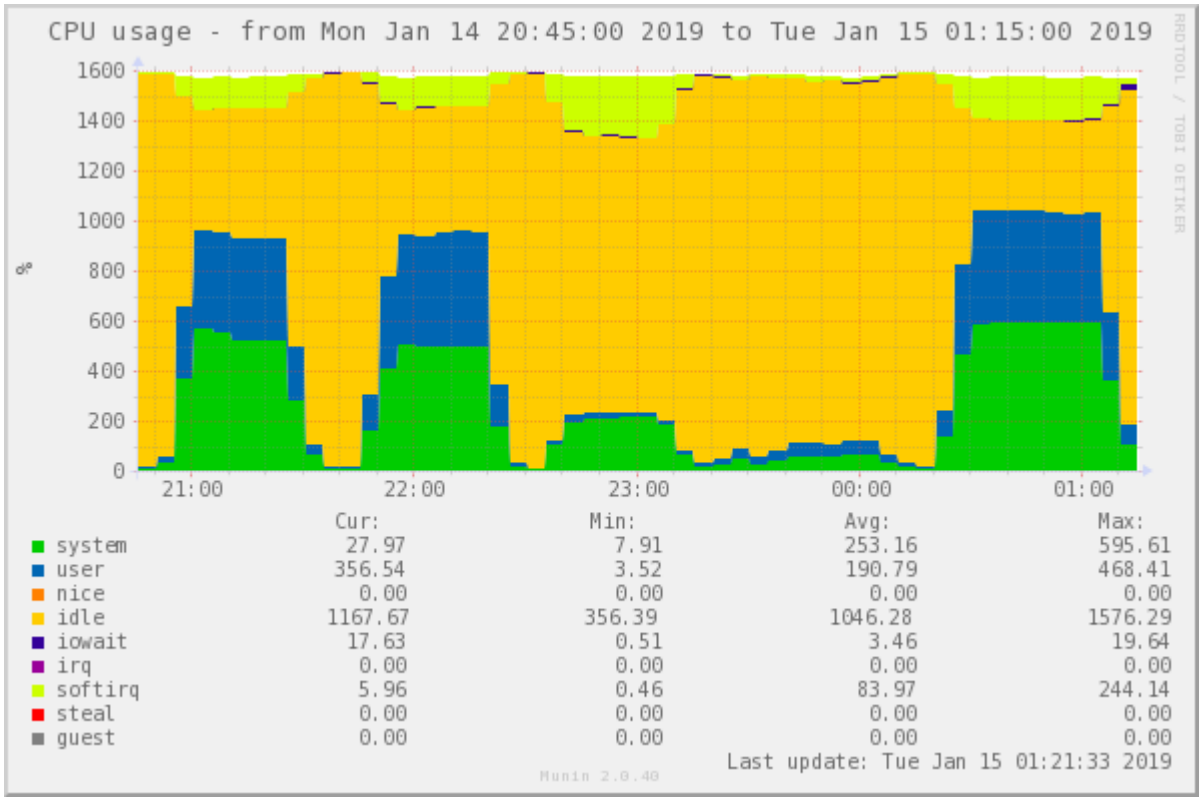
praguelcg2 INFN-NAPOLI-ATLAS [srm,root,davs]	2	1902	show	srm[DPM DOME/1.13.1]	**ok/1**	**ok/1**	ok/2	**ok/1**	**ok/1**	ok/2	**ok/1**	**ok/1**	**ok/1**
				gsiftp[gt6.2/13.11]	ok/4			ok/4					
				root[xrootd/0x40000]	ok/2	ok/3		ok/2	ok/3		ok/3	ok/3	ok/3
				davs[Apache/2.4.6]	ok/3	ok/2	**ok/1**	ok/3	ok/2	**ok/1**	ok/2 push: ok/2 pull: ok/2 streamed: ok/2	ok/2 push: ok/2 pull: ok/2 streamed: ok/2	ok/2 push: ok/2 pull: ok/2 streamed: ok/2
praguelcg2 INFN-ROMA1 [srm]	2	1264	show	srm[DPM/1.10.0-1]	**ok/1**	**ok/1**	ok/2	**ok/1**	**ok/1**	ok/2	**ok/1**	**ok/1**	**ok/1**
				root[xrootd/0x10030000]	ok/2			ok/2			ok/3	error/3	error/3
				davs[Apache/2.2.15]	ok/3	N/A/2	**ok/1**	ok/3	error/2	**ok/1**	ok/2 push: ok/2 pull: ok/2 streamed: ok/2	error/2 push: error/2 pull: error/2 streamed: error/2	error/2 push: error/2 pull: error/2 streamed: error/2
praguelcg2 INFN-ROMA2 [srm]	3	31	show	srm[StoRM/1.11.13]	**ok/1**	**ok/1**	**ok/1**	**ok/1**	**ok/1**	**ok/1**	**ok/1**	**ok/1**	**ok/1**
praguelcg2 INFN-ROMA3	3	51	show	srm[StoRM/ERROR]	test failed								
				davs[unknown/unknown]	test failed								
praguelcg2 INFN-T1 [srm]	1	8480	show	srm[StoRM/1.11.15]		**ok/1**		**ok/1**	**ok/1**		**ok/1**	**ok/1**	**ok/1**
				root[xrootd/0x40000]	ok/2			ok/2			ok/3	error/3	error/3
				davs[unknown/unknown]		ok/2	**ok/1**	ok/3	ok/2	**ok/1**	ok/2 push: ok/2 pull: ok/2 streamed: ok/2	ok/2 push: ok/2 pull: ok/2 streamed: ok/2	ok/2 push: error/2 pull: error/2 streamed: ok/2
praguelcg2 INFN-TRIESTE [srm]	3	26	show	srm[StoRM/1.11.11]	**ok/1**	**ok/1**	**ok/1**	**ok/1**	**ok/1**	**ok/1**	**ok/1**	**ok/1**	**ok/1**
praguelcg2 JINR-LCG2 [srm,davs]	2	1020	show	srm[dCache/5.2.4]	**ok/1**	**ok/1**		**ok/1**	**ok/1**		**ok/1**	**ok/1**	**ok/1**
				root[xrootd/0x89020000]	ok/2			ok/2			ok/3	error/3	error/3
				davs[dCache/5.2.5]	ok/3	ok/2	**ok/1**	ok/3	ok/2	**ok/1**	ok/2 push: ok/2 pull: ok/2 streamed: ok/2	ok/2 push: ok/2 pull: ok/2 streamed: ok/2	ok/2 push: ok/2 pull: ok/2 streamed: ok/2
praguelcg2 LRZ-LMU [srm]	2	2302	show	srm[dCache/4.2.39]		**ok/1**	ok/2	ok/2	**ok/1**	ok/2	**ok/1**	**ok/1**	**ok/1**
				root[xrootd/0x89020000]	**ok/1**			**ok/1**			ok/3	error/3	error/3
				davs[dCache/4.2.39]		N/A/2	**ok/1**	ok/3	error/2	**ok/1**	ok/2 push: ok/2 pull: ok/2 streamed: ok/2	error/2 push: error/2 pull: error/2 streamed: error/2	error/2 push: error/2 pull: error/2 streamed: error/2

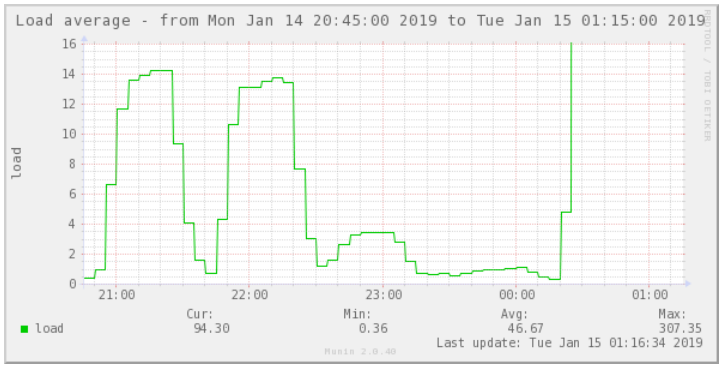
AES-NI & OpenSSL performance

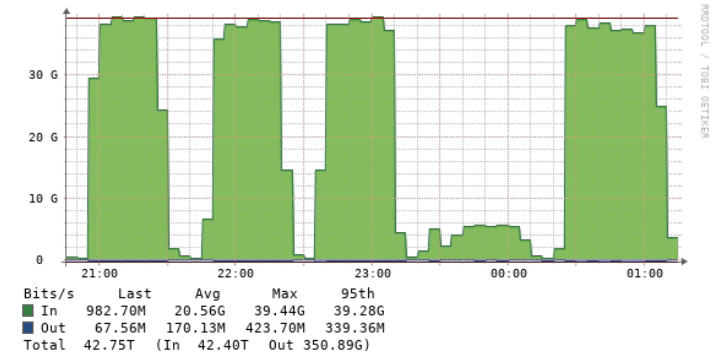
	2x Intel Xeon Silver 4108		2x Intel Xeon E5-2620 v2		2x Intel Xeon E5620	
cipher	#streams	speed Gb/s	#streams	speed Gb/s	#streams	speed Gb/s
aes-128-cbc	1	6.0	1	3.6	1	2.0
	16	95.1	12	43.0	8	15.5
			24	78.4		
aes-128-gcm	1	33.8	1	6.7	1	1.4
	16	385.8	12	70.9	8	10.3
			24	83.0		
aes-256-cbc	1	5.7	1	2.8	1	1.5
	16	68.9	12	31.0	8	12.3
			24	61.3		
aes-256-gcm	1	24.0	1	6.0	1	1.2
	16	279.8	12	61.3	8	8.6
			24	77.7		

AES-NI CPU utilization test - memory

- Machine utilization transferring data from memory with Apache
  - bumps in graph corresponds to this settings
    - https TLSv1.2,ECDHE-RSA-AES128-GCM-SHA256,2048,128, 32 connections, CPU load ~ 14
    - https TLSv1.2,ECDHE-RSA-AES256-GCM-SHA384,2048,256, 32 connections, CPU load ~ 14
    - http (no encryption), 32 connections, CPU load ~ 3.5
    - https TLSv1.2,ECDHE-RSA-AES256-GCM-SHA384,2048,256, 1 connection with throughput 4.2Gb/s
    - https TLSv1.2,ECDHE-RSA-AES256-GCM-SHA384,2048,256, 320 connections ~ load 310 but 40Gb still full







SRM-less operation

- Only LHCb still working on SRM-less
- ALICE use only XRootD
- ATLAS (CMS) needs storage space information formerly provided by SRM
- WLCG Storage Resource Reporting (SRR) format proposed
  - json format with basic data related to storage “spacetokens” (directories)
  - file provided by at least one supported protocol (GridFTP, XRootD, WebDAV)
- WLCG SRR implementation
  - DPM 1.10.3, since 1.13.2 available automatically via HTTP CGI at https://dpmheadnode.example.com/static/srr
  - dCache 4.2
  - StoRM 1.11.13