DATA CAROUSEL @CC-IN2P3

by **Aresh Vedaee** (12.12.19)

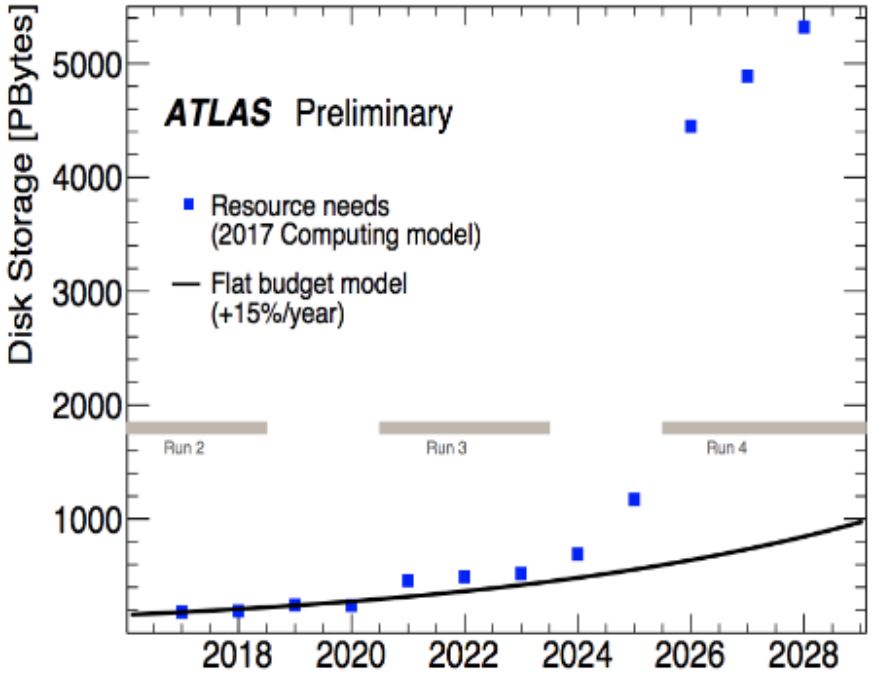
Outline

ATLAS Data Carousel R&D
Tests & results
Discussion points
Next steps

DATA CAROUSEL R&D : WHY ?

ATLAS perspective on the data storage challenge of HL-LHC:

- ‘Opportunistic storage’ basically doesn’t exist
- Format size reduction and data compression are both long-term goals, require significant efforts from the software and distributed computing teams
- Tape storage is 3~5 times cheaper than disk storage, increasing tape usage is a natural way to cut into the gap of storage shortage for HL-LHC



Exploit more tape usage

DATA CAROUSEL R&D : WHAT ?

- To study the feasibility to run various ATLAS workloads from tape
  - Facing the data storage challenge of HL-LHC, ATLAS started this R&D project this June 2018
- *By 'data carousel' we mean an orchestration between workflow management (WFMS), data management (DDM/Rucio) and tape services whereby a bulk production campaign with its inputs resident on tape, is executed by staging and promptly processing a sliding window of X% (5%?, 10%?) of inputs onto buffer disk, such that only ~ X% of inputs are pinned on disk at any one time.*

DATA CAROUSEL R&D : WHO?

- Rucio
  - Improve tape usage, e.g.bulk requests to tape, with size tailored to site parameters
- FTS
  - Optimize scheduling of transfers between tape and other storage endpoints, e.g. dedicated FTS instance for tape recall requests
- SE endpoints (dCache, StoRM, Castor, etc)
  - Any bottlenecks and possible improvements on interfacing with respective tape backend ?
- Evolving tape scheduler
  - Support high priority, low latency request ?
- PS2
  - Study and optimize prompt processing of data as it appears off of tape --- process immediately when X% of a dataset is staged ?

DATA CAROUSEL R&D : HOW ?

- First phase
  - Understand tape system performance at all T1 sites
  - Identify workloads (start with derivation), and evaluate performance based on current systems
    - Tape available at ~ 10 sites, while processing happens everywhere
    - Performance with tape vs disk
- Second phase ⇦
  - Address issues found in phase 1
  - Deeper integration between workload and data management systems (PanDA/PS2/Rucio)
- Third phase (for Run3)
  - Integrate with production system and run production, at scale, for selected workflows

DATA CAROUSEL TESTS - PHASE 1 : CONDITIONS

- Run the test:
  - Rucio → FTS → Site: staging files from tape to local disk (DATATAPE/MCTAPE to DATADISK)
  - Data sample
    - About 100TB~200TB AOD datasets, average file size 2~3GB
  - Bulk mode
    - Sites can request throttle on incoming staging requests (3 sites)
  - With concurrent activities (production tape writing/reading and other VOs)
- Status : all done at 10 T1s
  - BNL, FZK, PIC, INFN, TRIUMF, CCIN2P3, NL-T1 and RAL, NRC and NDGF

DATA CAROUSEL TESTS - PHASE 1 : RESULTS (ATLAS VIEW)

Site	Tape Drives used	Average Tape (re)mounts	Average Tape throughput	Stable Rucio throughput	Test Average throughput
[1]BNL	31 LTO6/7 drives	2.6 times	1~2.5GB/s	866MB/s	545MB/s (47TB/day)
FZK	8 T10KC/D drives	>20 times	~400MB/s	300MB/s	286MB/s (25TB/day)
INFN	2 T10KD drives	Majority tapes mounted once	277MB/s	300MB/s	255MB/s (22TB/day)
PIC	5~6 T10KD drives	Some outliers (>40 times)	500MB/s	[2]380MB/s	400MB/s (35TB/day)
[1]TRIUMF	11 LTO7 drives	Very low (near 0) remounts	1.1GB/s	1GB/s	700MB/s (60TB/day)
CCIN2P3	[3]36 T10KD drives	~5.33 times	2.2GB/s	3GB/s	2.1GB/s (180TB/day)
SARA-NIKHEF	10 T10KD drives	2.6~4.8 times	500~700MB/s	640MB/s	630MB/s (54TB/day)
[4]RAL	10 T10KD drives	n/a	1.6GB/s	2GB/s	1.6GB/s (138TB/day)
[5]NDGF	10 IBM Jaguar/LTO-5/6 drives, from 4 sites	~3 times	200~800MB/s	500MB/s	300MB/s (26TB/day)

[1] dedicated to ATLAS
[2] with 5 drives, later increased to 6 drives
[3] 36 is the max number of drives, shared with other VOs who were not using them during the test
[4] 8 drives dedicated to this test. Will have 22 shared with other VOs in production.
[5] federated T1, 4 physical sites have tapes

@CC-IN2P3 :

Meilleurs résultats des T1 !

Mais :
- 36 drives utilisés
- Taux remontage : 5,33 x / bande

U.S. DEPARTMENT OF ENERGY

BROOKHAVEN NATIONAL LABORATORY

DATA CAROUSEL TESTS - PHASE 1 : COMMENTS (ATLAS VIEW)

- Results is better than expected
  - ~600TB/day total throughput from all T1s, under “as is” condition
  - Can we repeat it in real production environment ?

- ***Estimate*** on data volume required by current ATLAS derivation campaign
  - 260TB/day input AOD data, if run on 100k cores

DATA CAROUSEL TESTS - PHASE 2 : CONDITIONS

- The P2R2 test refers to the last 2018 RAW reprocessing campaign.
  - Timeline : started on the 8th August 2019
  - Data volume varies by site
  - No warning
  - Job released by ATLAS PS2 after 90% of staged input data
  - Monitoring tool
    - https://bigpanda.cern.ch/datacardash/

Source	Progress plot	DDM Dash	Datasets Active	Datasets Done (+ 90% readiness)	Files Remaining	Files Done
BNL-OSG2_DATATAPE		-->	3	0 (+0)	171	3
CERN-PROD_RAW		-->	3	0 (+3)	0	22886
FZK-LCG2_DATATAPE		-->	1	0 (+0)	1	0

Show 10 entries Search:

Campaign	Request ID	TaskID	Status	Total Files	Staged Files	Progress (%)	Source RSE	Time Elapsed	Started At	Rucio Rule
Archive	27866	19899431	staging	22299	22287	100		6 days, 18:11:08.719864	2019-11-28T13:56:03.295620	c5cf5e27841b4e3090bf42ff4e6a05eb
data15_13TeV	23576	18352838	done	58	1	2	BNL-OSG2_DATATAPE	1 day, 21:26:47.140179	2019-12-03T10:40:24.875153	19ff7762384443f081fa98ebb9a92ccc

DATA CAROUSEL TESTS - PHASE 2 : RESULTS (ATLAS VIEW)

sites	tape	SE	Tape drives (max. reading)	P1 avg throughput	P1 stable throughput	P2R2 avg throughput (100% staged)	P2R2 avg throughput (90% staged)	Capacity pledge (2019)
CERN	CTA	EOS	(expect 60GB/s for Run3)	2GB/s	2GB/s	763MB/s	1.2GB/s	100%
BNL	HPSS	dCache	22 LTO7 17 LTO6	545MB/s	866MB/s	900MB/s	1.4GB/s	23%
FZK	TSM → HPSS	dCache	30 T10KD	286MB/s	300MB/s	316MB/s	324MB/s	13%
RAL	CTA	Echo	21 T10KD	1.6GB/s	2GB/s	850MB/s	1.3GB/s	13%
CCIN2P3	HPSS	dCache	56 T10KD	2.1GB/s	3GB/s	401MB/s	524MB/s	12%
TRIUMF	Tapeguy	dCache	20 LTO8 12 LTO7	700MB/s	1GB/s	366MB/s	330MB/s	10%
INFN	TSM	StoRM	16 T10KD 19 TS1600	255MB/s	300MB/s	N/A	N/A	8%
NL-T1	DMF	dCache	8 T10KC, 2 T10KD, 10 LTO8	630MB/s	640MB/s	626MB/s	630MB/s	8%
NDGF	TSM	dCache	N/A	300MB/s	500MB/s	214MB/s	371MB/s	6%
PIC	Enstore	dCache	4~6 T10KD	400MB/s	380MB/s	179MB/s	170MB/s	4%

@CC-IN2P3 : 311K files, 0.6PB data in 21 days → AVG THROUGHPUT << 1GB/s << P1R2 (2.1GB/s)

DATA CAROUSEL TESTS - PHASE 2 : RESULTS (CC-IN2P3 VIEW)

ATLAS ACTIVITY ON HPSS

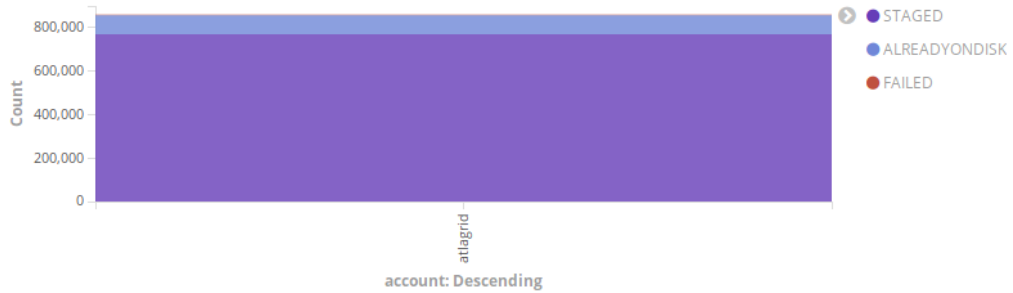
860K files, 1.7PB data in 21 days (~1GB/s) : quite different stats wtr ATLAS report !

860,523 Requests **1.724PB** Total Size

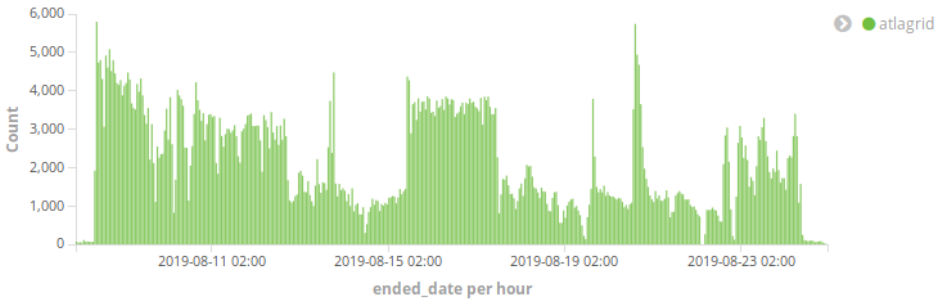
Status	Count	File size	Avg File size
STAGED	769,821	1.642PB	2.237GB
ALREADYONDISK	87,764	79.816TB	953.618MB
FAILED	2,938	4.027TB	1.403GB
	860,523	**1.724PB**	**4.572GB**

Export: Raw Formatted

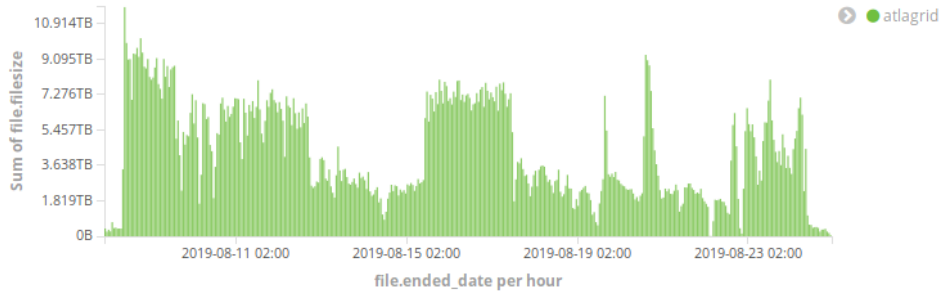
TREQS2 : Requetes par utilisateurs



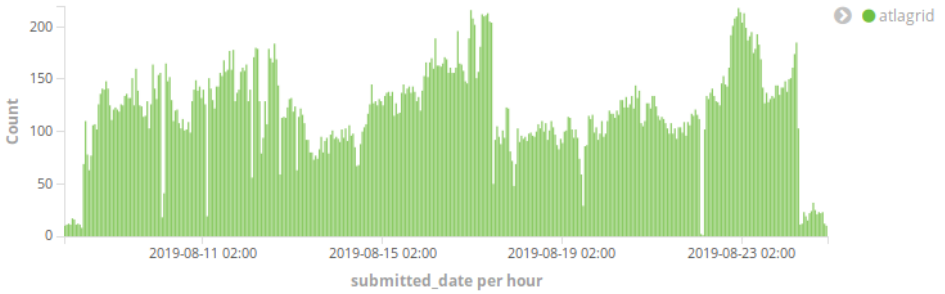
TREQS2: File requests by hour



TREQS2: Stage rate by users



TREQS2: Tape count by users



DATA CAROUSEL TESTS - PHASE 2 : COMMENTS (1/2)

- FTS issues @CERN:
  - FTS Scheduler degraded
    - FTS was not able to schedule transfers between the tape buffer to the disk on time → files got garbage collected → transfer requests failed → FTS optimizer throttled to the minimum parallel transfer requests
  - Fail nearline
    - Redundant transfer requests' failures → FTS optimizer throttling
  - FTS Daemons' crashes
    - Failing to recover "already started" requests → untracked staged files

The main consequence of this problem was the CC-IN2P3 poor staging stats !

- More details :
  - https://indico.cern.ch/event/843988/contributions/3543611/attachments/1904532/3146356/FTS_Data_Carousel_PostMortem.pdf

DATA CAROUSEL TESTS - PHASE 2 : COMMENTS (2/2)

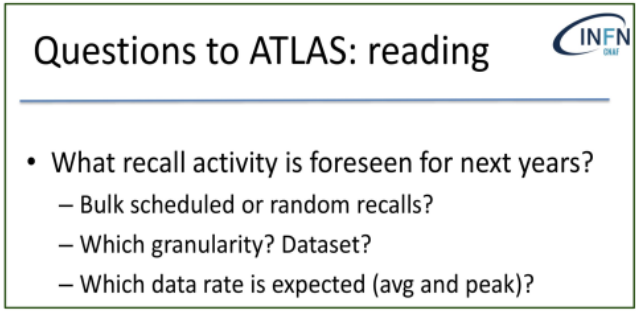
- During P2R2 the performance of CC-IN2P3 storage system was lower than P1R2.
- **But it wasn't a site issue nor a performance issue :**
  - As later understood, it was mainly due to several FTS issues @CERN
  - Besides, the number of requests per recall pool during P2R2 @CC-IN2P3 was even increased wrt P1R2
  - And site performance varies wrt the whole of concurrent activities within and outside a given VO, sharing same storage resources → **how reppresentative can be this and past tests for storage systems' performance assessments ?**
- Future improvements :
  - Revise dCache config for ATLAS : redundant intermediate pool-to-pool copies (but that depends on how ATLAS' actual tape usage) ? Increase # max requests per recall pools ?
  - Move to IBM Entreprise class tape cartridges&drives (Jaguar)

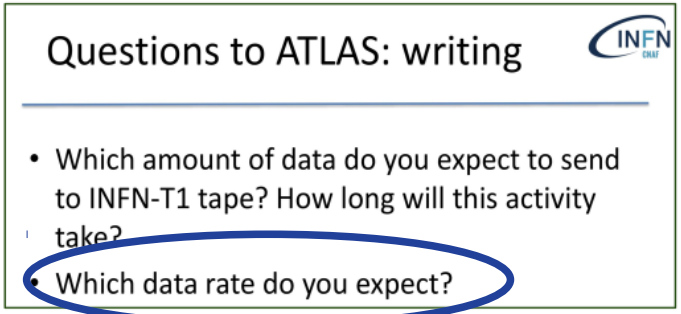
DISCUSSION POINTS

- ATLAS SW&Computing Week
  - 4 sessions dedicated to Data Carousel: Sites, FTS&Rucio, Storage (dCache, CTA), Discussions
  - Addressing two main gaps :
    - Gap 1 : between the throughput out of the tape system itself and the throughput delivered to users (rucio in this case)
      - This is to tackle the issues along the staging chain, dCache, FTS, Rucio, PS2, etc, to minimize performance penalties to the original tape throughput
    - Gap 2 : between the nominal tape throughput and the current throughput out of tape
      - This is about “smart writing”, by bigger files and/or better organizing files among tapes, to reach higher tape reading efficiency

DISCUSSION POINTS

- What are the expected tape throughput (both write and read) for my site ?





- Another aspect of expectation is the switch to " data carousel" model from a traditional archival model, which isn't what some sites had in mind when they initially designed their tape infrastructure

  - We're all different
    - Different technologies
    - Different hardware
    - Different software
    - Different setups
  - So we're on our own basically wrt tuning

DISCUSSION POINTS

- Improvements on hardware
  - Bigger disk buffer on the frontend
  - More disk pool servers
- Improvements on software
  - dCache request grouping mechanisms ?
  - Other HSM interface: ENDIT (dCache plugin to solve scalability issue of default dCache HSM interface)
- Write in the way you want to read later
  - File family is good feature provided by tape system
  - There are more...group by datasets ? If grouping by dataset is too small is container a solution?
  - Discussion between dCache/Rucio: Rucio provide dataset info in the transfer request ?
- File size
  - ADC working on increasing file size for tape writing (target at 10GB)

- Bulk request limit
  - Need knob to control bulk request limit : 3 sites requested a cap on the incoming staging requests from upstream (Rucio/FTS)
  - Consideration factors ---limit from tape system itself, size of disk buffer, load the SRM/pool servers can handle, etc
  - Three places to control the limit
    - Rucio can set limit per (activity&destination endpoint) pair
      - Adding another knob on limiting the total staging requests, from all activities
    - FTS can set limit on max requests
      - Each instance sets its own limit, need to orchestrate multiple instances
    - dCache sites can control incoming requests by setting limits on:
      - Total staging requests, in progress requests and default staging lifetime

NEXT STEPS

- New test with reprocessing campaign on the 6th January 2020 (after FTS upgrade ?)

- Accounting @CC-IN2P3 ?
  - Mounts per tape
  - Number of drives used per VO and R/W activity
    - Concurrent activities within and outiside a VO
  - Stats by dataset (dataset distribution on tape) and file request (# of times a given file was staged)
    - For recall pool configuration and buffer dimensioning

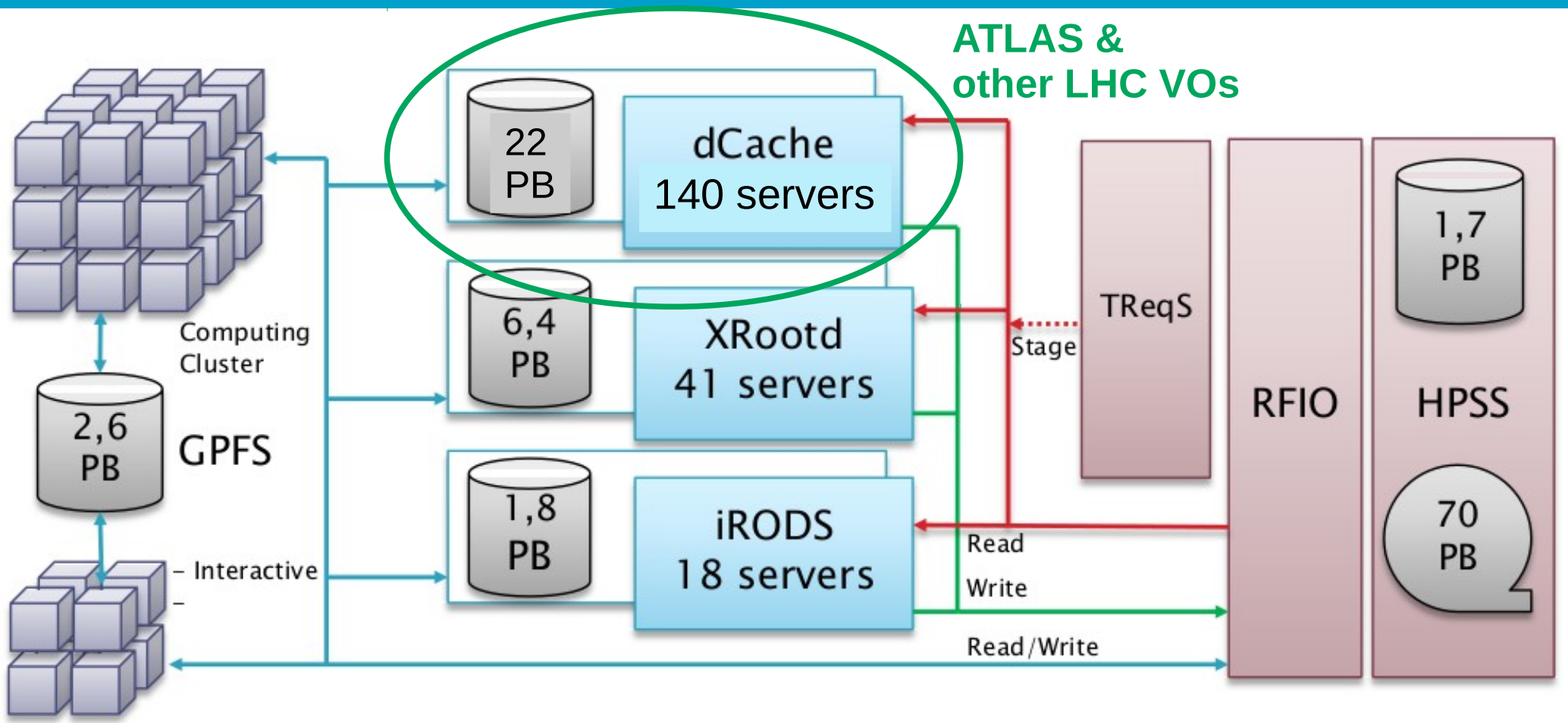
USEFUL LINKS

- https://indico.cern.ch/event/756338/attachments/1723845/2784624/update-atlas-data-carousel-wlcg-wg.pdf
- https://indico.cern.ch/event/651359/contributions/3208536/attachments/1752789/2840658/atlas-data-carousel-GDB-nov2018.pdf
- https://indico.cern.ch/event/865577/contributions/3646827/attachments/1951569/3240027/CC_IN2P3_DATA_CAROUSEL_2019.pdf
- https://indico.cern.ch/event/823341/


Centre de Calcul de l'Institut National de Physique Nucléaire et de Physique des Particules

BACKUP

STORAGE SYSTEM : TAPE (2/3)



- HPSS hpss-7.5.1.2-20190116.u9
- 85 % of HPSS access are performed through storage middleware
  - **dCache** (LCG/egee),
  - **Xrootd** and **iRods**
- Still some direct access to HPSS but decreasing

- HPSS Interface :
  - RFIO with HPSS extensions
  - Read operations from storage middleware are handled by TREQS 2

STORAGE SYSTEM : TAPE (3/3)

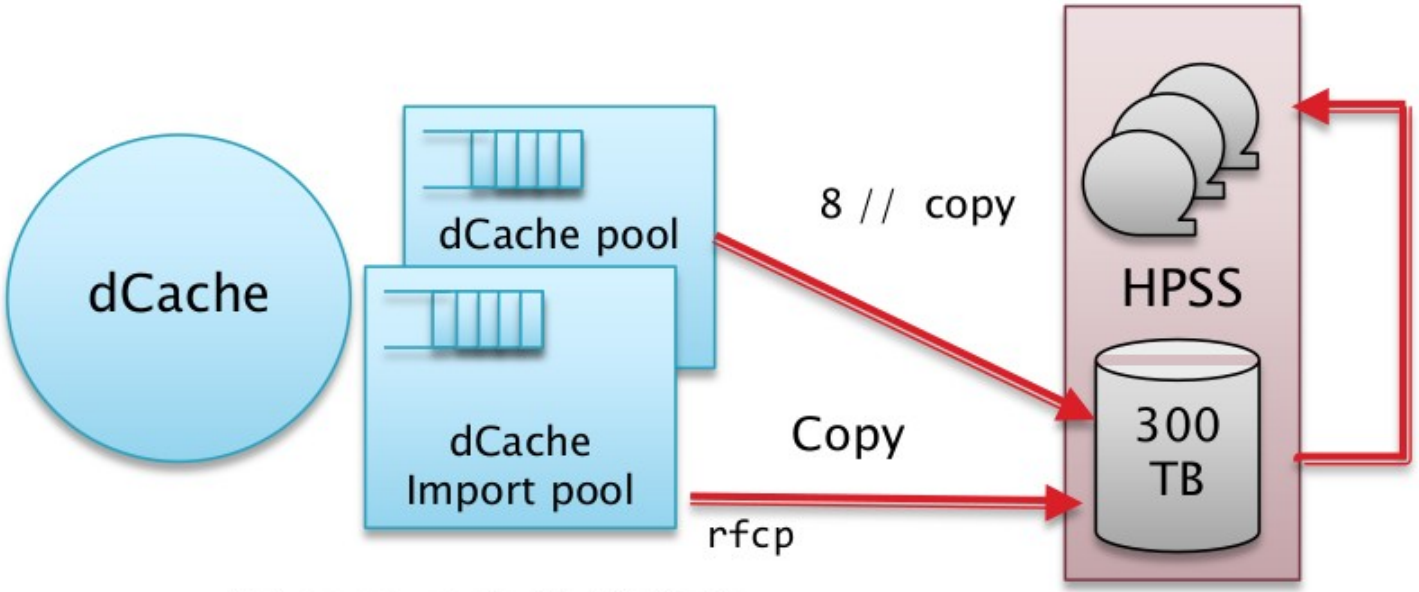
- Tape Libraries
  - 4 Oracle SL8500 Libraries
  - Interconnected (with PTP)
  - Collocated with TSM (backup)
- 116 Tapes drives
  - 22 T10K-C (EOL)
  - 20 LTO 4 (TSM)
  - 17 LTO 7 (TSM)
  - 56 T10K-D (HPSS)
  - 1 LTO8 (test)

ATLAS & other LHC VOs

- ACSLS 8.5 on RHEL 7
- 20 000 Tapes
  - 13000 T10000T2 (8,5 TB)
  - 5 000 LTO 4
  - 2 000 LTO 6
- Daily tape mounts:
  - 2 000 average
  - > 6 000 peak (300/h)

- Store operation
  - 8 import pool , 1 connection , ~8TB per pool
  - 3 space tokens (mctape, datatape, archive)
  - 4 storage class (based on file size)
- Files written in the same time are spread over multiple tapes
  - IE : 6 tapes for XL files.
- Filename based on pnfsid:

/hpss5/dcache/atlas/**datatape**/2018/10/**000055C8CF15B9764B858C974D31928071D3**

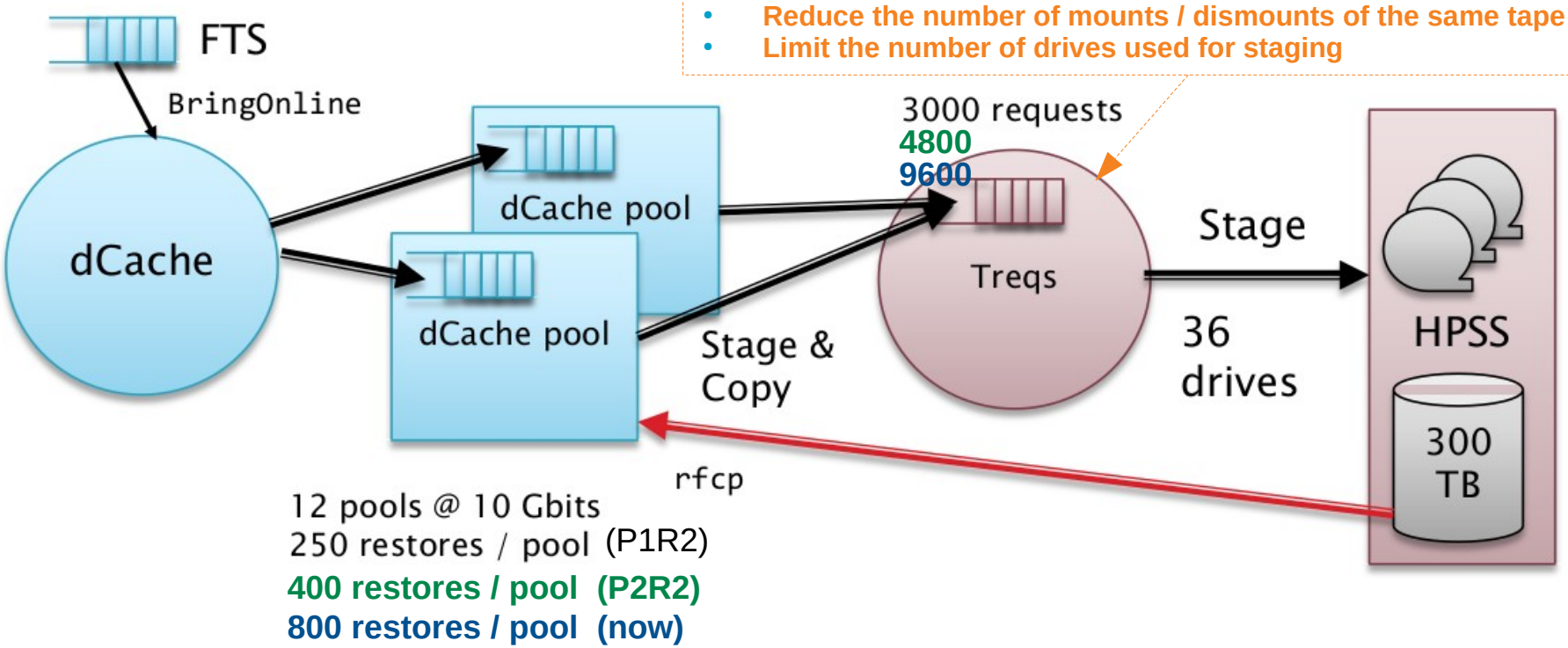


Migration :
Run every 4h to 24h (depends of file size)

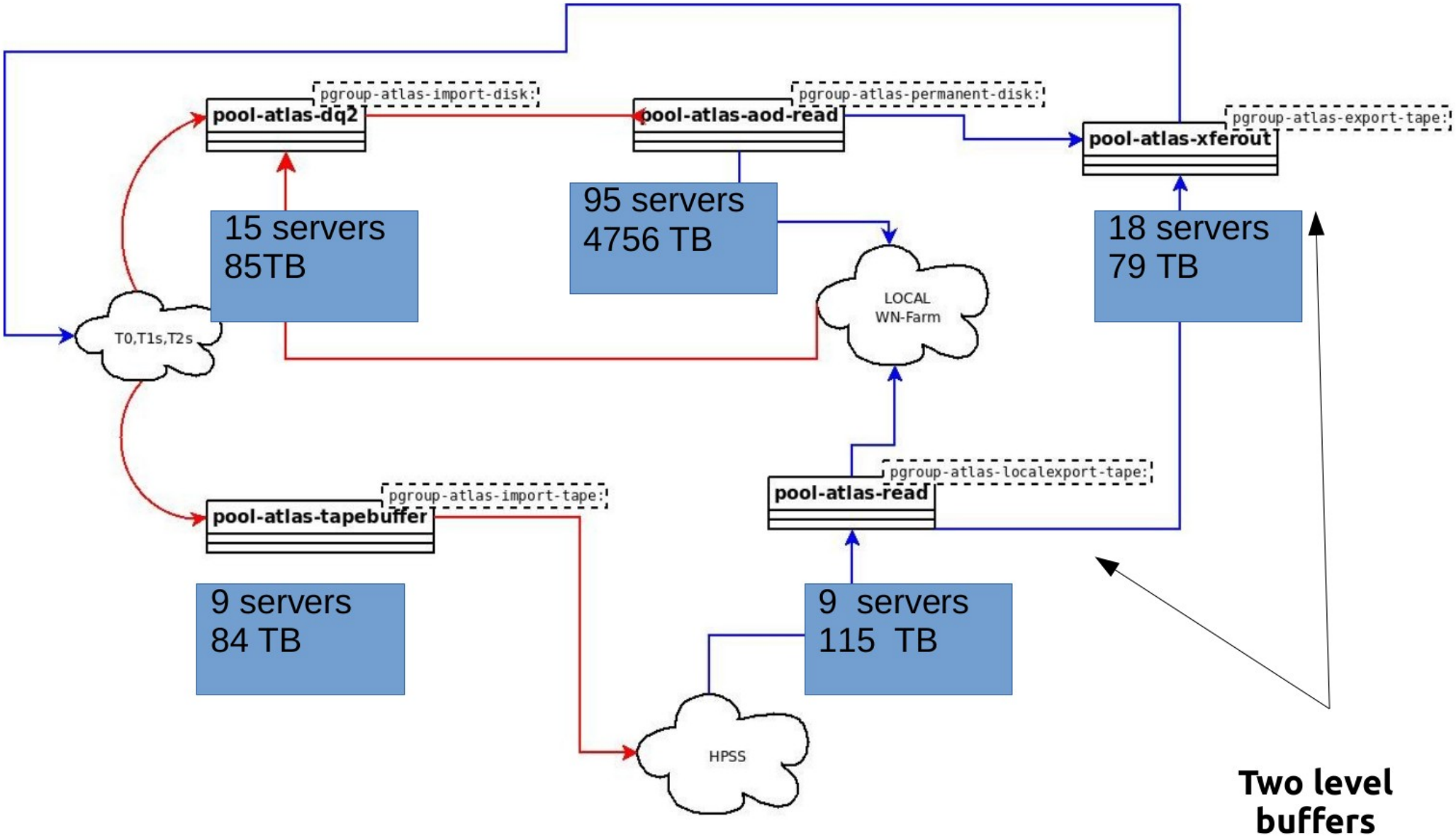
Drive used
S Files : 1 drive
M Files : 2 drives
L Files : 5 drives
XL Files : 6 drives (≥2GB)

- Restore operation
  - dCache submit requests over 12 pools (~10TB per pool)
  - Each pool can handle **800 requests max (250 for P1R2, 400 for P2R2**)
  - Treqs schedule requests and stage files
    - 36 drives configured (**shared by all experiments and r/w activities !**)
  - HPSS handle only 1 stage requests a time per drive

- **Optimize read operations by sorting files by tapes and positions**
- **Reduce the number of mounts / dismounts of the same tape**
- **Limit the number of drives used for staging**



dCache Pool Diagram



DATA CAROUSEL TESTS - PHASE 2 : RESULTS (ATLAS VIEW)

Source site	NB files (Share)	Nb days for 90% staging	NB days for 100% staging
BNL-OSG2_DATATAPE	606333 (19.58%)	10.10	18.29
CERN-PROD_RAW	325140 (10.50%)	6.17	11.93
FZK-LCG2_DATATAPE	345232 (11.15%)	>28	>33
IN2P3-CC_DATATAPE	311457 (10.06%)	16.36	21.02
NDGF-T1_DATATAPE	212794 (6.97%)	13.25	26.47
PIC_DATATAPE	176536 (5.70%)	21.85	25.12
RAL-LCG2_DATATAPE	455127 (14.70%)	8.44	14.27
SARA-MATRIX_DATATAPE	345130 (11.15%)	12.54	14.50
TRIUMF-LCG2_DATATAPE	260537 (8.42%)	16.33	18.27

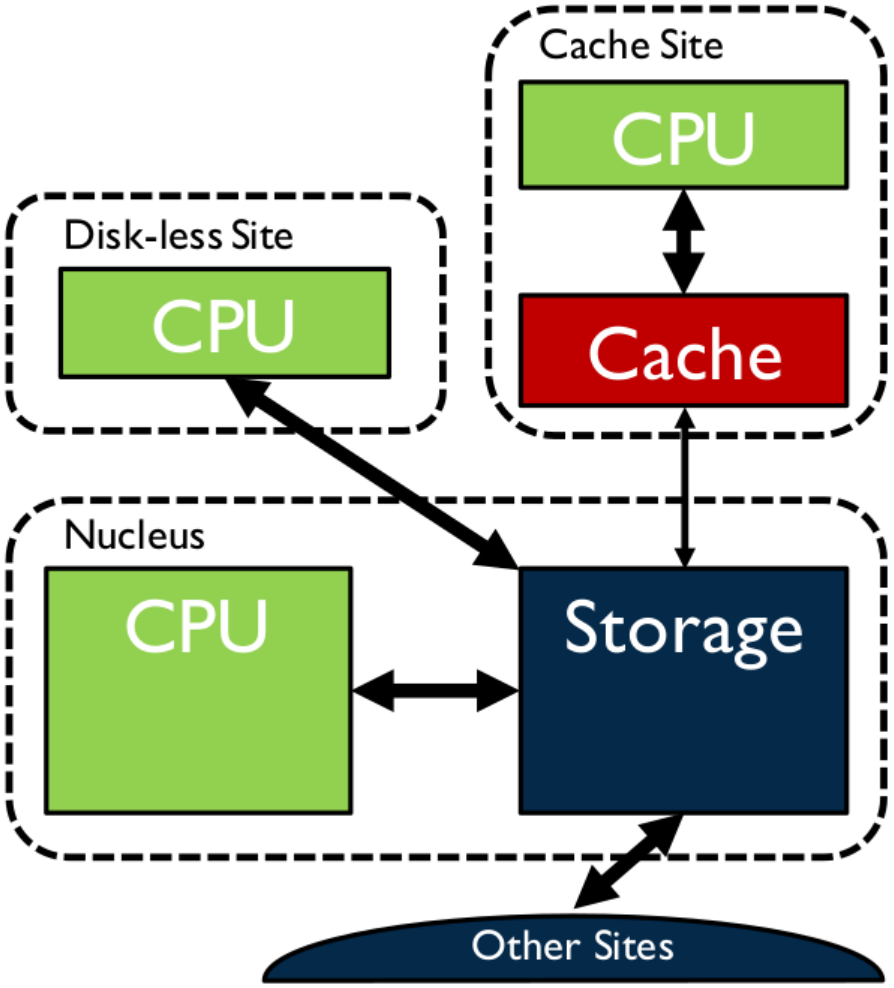
@CC-IN2P3 :
311K files, 0.6PB data in 21 days → AVG THROUGHPUT << 1GB/s << P1R2 (2.2GB/s)

From Cedric

From RAL

Tape	Total size	Avg. size per dataset	Avg. files per dataset	Avg. dataset staging time (days)	Avg. task executing time(days)
BNL-OSG2_DATATAPE	1.2 PB	26 TB	12000	15	22
FZK-LCG2_DATATAPE	0.8 PB	31 TB	14000	33	34
IN2P3-CC_DATATAPE	0.6 PB	29 TB	13000	15	27
NDGF-T1_DATATAPE	0.5 PB	30 TB	13000	21	24
PIC_DATATAPE	0.4 PB	34 TB	15000	22	27
RAL-LCG2_DATATAPE	1 PB	32 TB	14000	11	16
SARA-MATRIX_DATATAPE	0.7 PB	31 TB	14000	13	30(?)
TRIUMF-LCG2_DATATAPE	0.6 PB	32 TB	15000	18	25

Network transfers



Network Requirements

Nucleus	Now	5 year (2022)	10 year (2027)
Storage Capacity (PB)	2	5	12.5
Total CPU (kHS06)	40	100	250
LAN (Gb/s)	40	200	1000
WAN (Gb/s)	20	60	200

Disk-less	Now	5 year (2022)	10 year (2027)
Total CPU (kHS06)	20	50	125
WAN (Gb/s)	4	20	100

- Storage and CPU numbers are example values, scale appropriately for different sites.
- Storage and CPU increasing by 20% year (flat cash assumption).
  - We expect network usage to grow slightly faster.
- We expect Nucleus storage to be dual stack by 2019 (End of Run 2).
- In 5 years we recommend 100Gb/s WAN connectivity for Tier 2s.