

CMS/FTS Community Talk

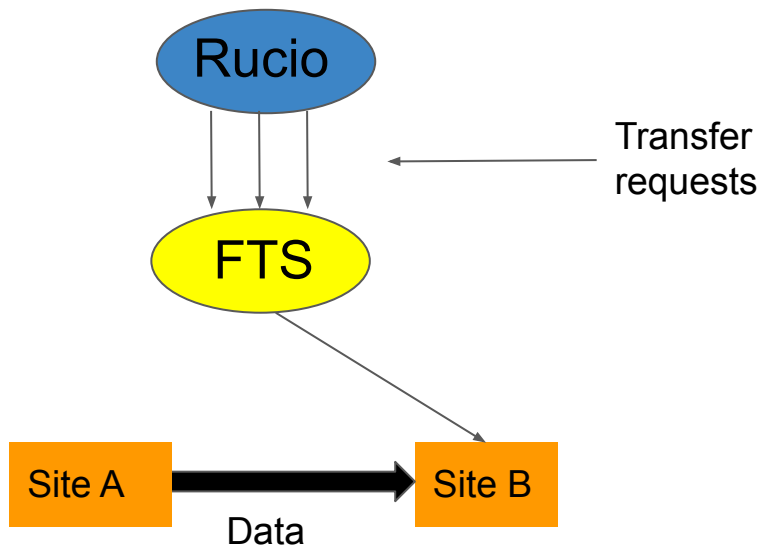
Katy Ellis, 28/03/23
XrootD/FTS workshop, Ljubljana

CMS Data Storage and Management

- The CMS experiment keeps ~200PB of data for short-term use on disk at around 60 sites, with an additional ~350PB of tape storage at 8 sites
- CMS switched in production to the Rucio data management software at the end of 2020
- Rucio manages data placement according to 'rules' and submits site-to-site transfer (copy) requests to FTS
- CMS also makes use of streamed data reads which do not use FTS

CMS data movement (1 / 2)

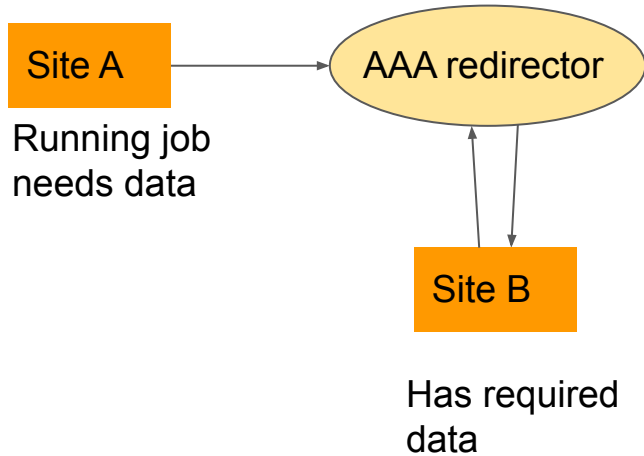
“Scheduled” transfers



- Rucio manages data placement according to ‘rules’, determines source and destination URLs based on site config and submits transfer requests to FTS
- FTS handles file transfers using...
- ...a range of protocols, such as davs/srm/xrootd/(gsiftp being phased out)

CMS data movement (2 / 2)

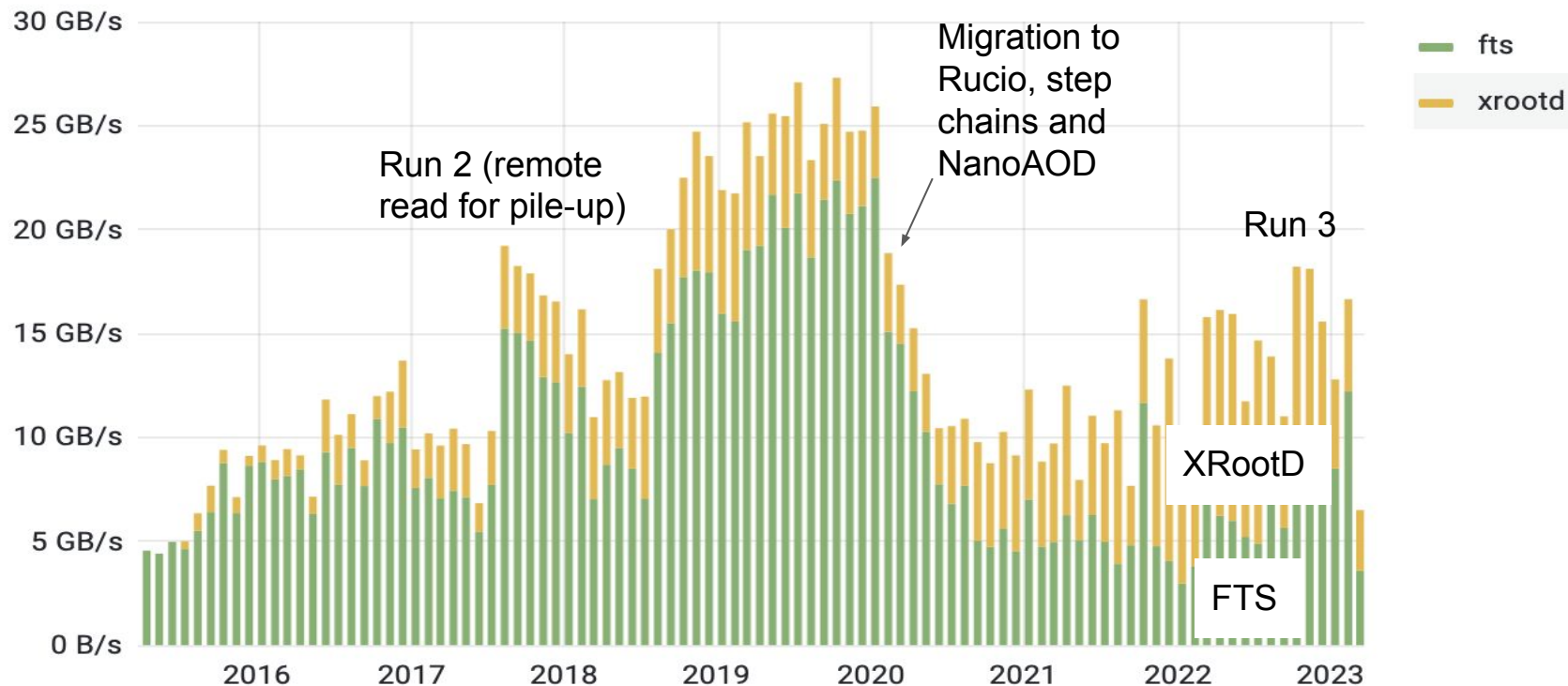
“Unscheduled” transfers



- CMS jobs stream their input data
 - Usually from the local storage system;
 - if that data is unavailable, then the job requests from AAA (“any data, anywhere, any time”)
 - Quickest site serves the data via XRootD
 - Since Run 2 used for significant secondary inputs (pile-up libraries)

Historical rates, 2015 - present

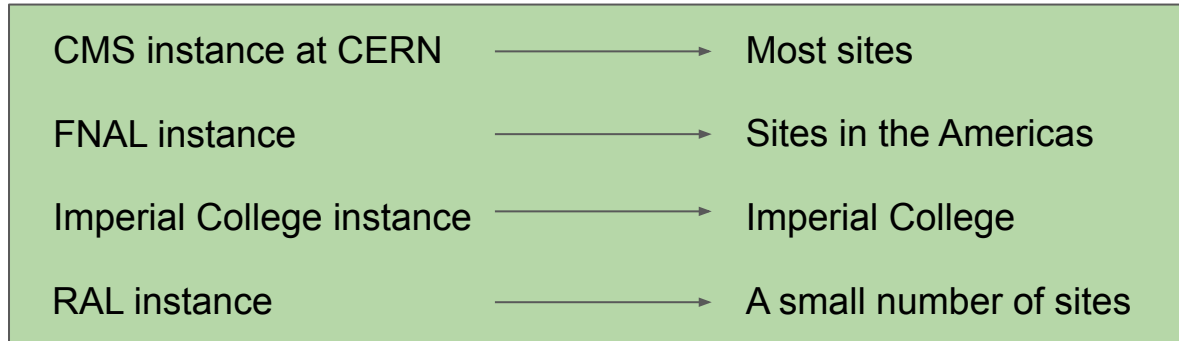
Transfer Throughput



CMS usage of FTS

CMS uses up to 4 instances of FTS

- CERN and Fermilab (doing most of the work) plus Imperial College and RAL
- Rucio selects which FTS to use based on the final destination of the data



- Multiple instances allow redundancy in the system

Recent FTS developments requested by CMS 1

Example 1 - file on tape:

- It was important for CMS to know that a file had been fully archived to physical tape, and not just copied to buffer
- FTS introduced an 'archiving' status
- A file must be archived according to FTS before it is 'OK' in Rucio
- This has been useful for quickly determining if tape systems are working

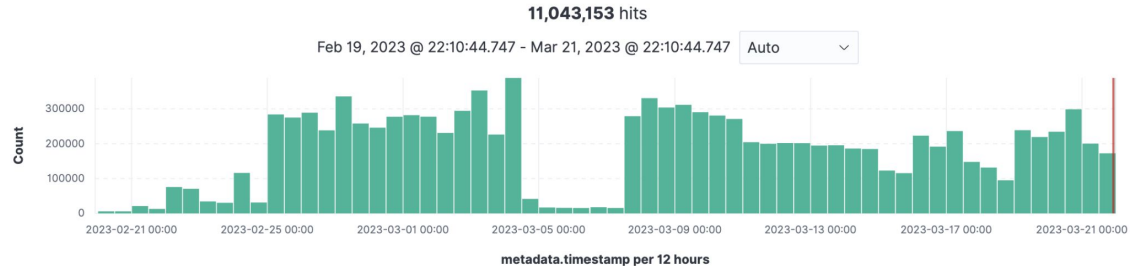
Source	Destination	VO	Submitted	Active	Staging	S.Active	Archiving	Finished	Failed	Cancel
+ davs://eoscms.cer	srm://cmssrm- kit.gridka.de	cms	2	-	-	-	4	331	-	-

Recent FTS developments requested by CMS 2

Example 2 - 'destination file exists':

- CMS files have the same logical file name wherever they are stored
- CMS do not allow automatic overwrites on tape
- Files are written to tape...but sometimes create error:
- `DESTINATION [17] Destination file exists and overwrite is not enabled`
- This can mean one of two things:
 - File is corrupt and needs re-writing
 - File is perfectly fine, but FTS did not receive confirmation of transfer

'Destination file exists'
errors in Rucio
monitoring last 30 days



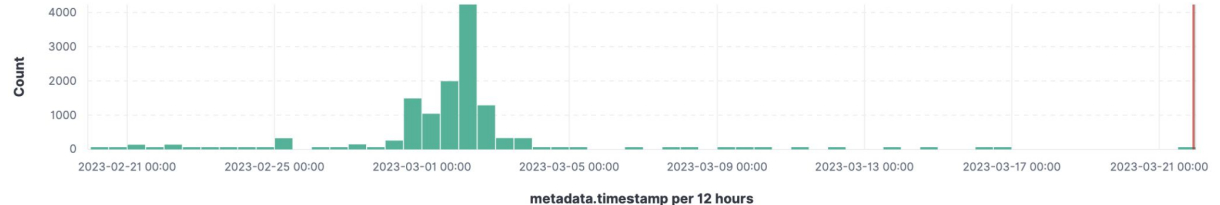
Recent FTS developments requested by CMS 2

Example 2 - 'destination file exists':

- FTS developed a feature to verify if the file was correct - here is an example of it working well:

Time ▾	data.event_type	data.purged_reason	data.dst_rse
> Mar 21, 2023 @ 22:02:17.017	transfer-done	DESTINATION [17] Destination file exists and overwrite is not enabled	T1_US_FNAL_Tape
> Mar 21, 2023 @ 15:44:20.834	transfer-failed	DESTINATION [70] srm-ifce err: Communication error on send, err: [SE][Ls][] IP HOST:PORT PATH HTTP Error	T1_US_FNAL_Tape
> Mar 21, 2023 @ 15:37:45.484	transfer-failed	DESTINATION [70] srm-ifce err: Communication error on send, err: [SE][Ls][] IP HOST:PORT PATH HTTP Error	T1_US_FNAL_Tape
> Mar 21, 2023 @ 15:35:05.256	transfer-failed	DESTINATION [70] srm-ifce err: Communication error on send, err: [SE][Ls][] IP HOST:PORT PATH HTTP Error	T1_US_FNAL_Tape
> Mar 21, 2023 @ 15:29:25.032	transfer-failed	ARCHIVING [70] srm-ifce err: Communication error on send, err: [SE][Ls][] IP HOST:PORT PATH HTTP Error	T1_US_FNAL_Tape

Feb 19, 2023 @ 22:22:17.208 - Mar 21, 2023 @ 22:22:17.208 Auto ▾



'Destination file exists'
files 'fixed' in last 30
days

Recent FTS developments requested by CMS 2

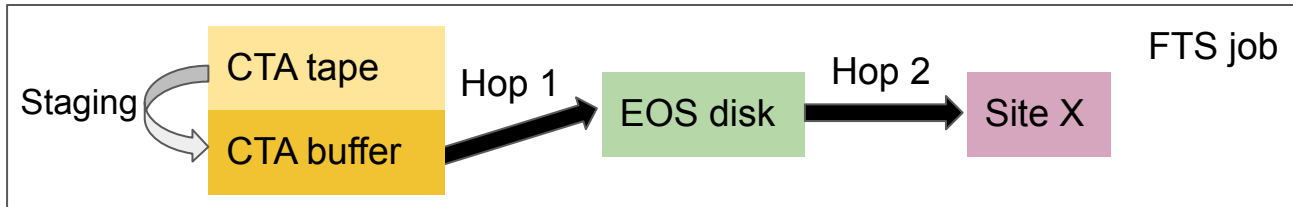
Example 2 - 'destination file exists':

- However, sometimes it is observed that the 'destination file exists' error repeats over and over before the check is performed and 'transfer-done'
 - Further work planned by FTS and Rucio to improve
- More recently files have been observed which **are** corrupted - CMS needs to decide how to deal with these effectively and also investigate the cause of file corruption

Recent FTS developments requested by CMS 3

Example 3 - multihop:

- ‘Multihop’ transfers were needed in FTS and Rucio because of the small buffer on the new CERN Tape Archive (CTA).
- Rucio works out the ‘route’ and communicates to FTS the ‘hops’.
- An FTS job contains all the hops for a particular file transfer from src->dest



Files archiving to tape take the same path in reverse

- Various improvements made in the last years, on FTS, CTA and Rucio (inc. CMS config)
 - Optimising timeouts, FTS notifying CTA of expired jobs, re-submitted jobs appearing as two FTS jobs, holding space on EOS, etc..

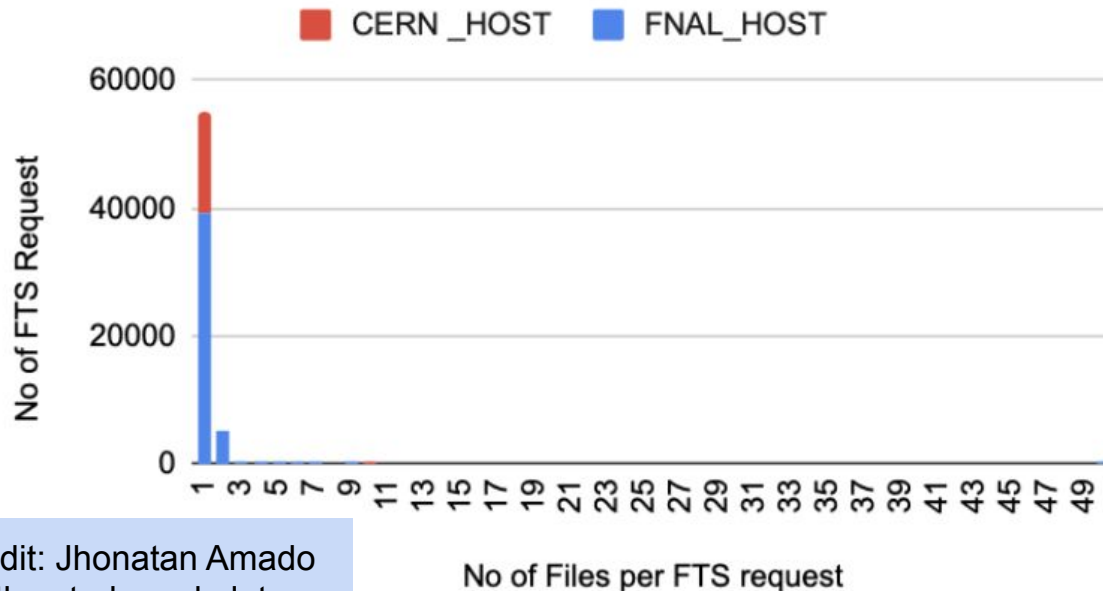
Recent CMS 'data storm'

- By the New Year 2023, CMS data transfers were struggling with a number of difficulties, leading to a huge number of requests in the system. E.g.
- Transfers to/from JINR Tape system
 - Recalling unique data to be replicated elsewhere
 - Attempting to write 12PB
- Consistency checking bug manifesting at Florida
 - Consistency checking found zero files present, and tried to re-copy the whole site
- Many sites needing attention after christmas break
 - Lots of failed transfers

Bulk transfer investigations


CMS don't appear to be using FTS bulk transfers effectively

wmcore_output last_7_days



- Plot shows the vast majority of wmcore_output account FTS requests contain only a single file
- The same applies to other accounts
- **We think this needs to be improved via better configuration in CMS-Rucio**

Timeout on tape staging extended

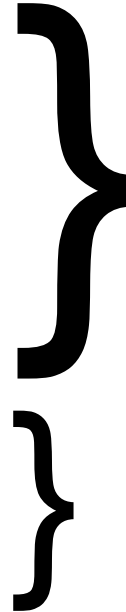
- Tape recalls have an FTS `bring-online` timeout which is configurable via Rucio
 - Previously set to 7 days...since CMS wanted to re-submit requests if they did not succeed in that time
 - However, it was not considered that this would cause massive failures when the staging 'queue' was much longer than 7 days!
-  Recently extended to 30 days, to cope with e.g. B-parking staging

Future improvements

- Long gap between FTS complete and Rucio OK
 - There has been a fix made for this - CMS Rucio should pick it up in next upgrade
- Using different Rucio 'activities' for tape recall priority
 - Could all FTS transfers be better prioritised?
- Better grouping of file transfers
 - Although things to try on the CMS-Rucio side
- Better handling of 'destination file exists'
- **Request:**
 - CMS would like to understand better large (>20GB) file transfer failures
 - Could 100GB file transfers be feasible by Run-4?
 - Would it be possible for FTS to resume a failed transfer rather than start from scratch?

Increase in transfer volume?

- Event size is driven by pile-up
 - Increased from 2022 to 2023
 - But no expected increase in 2024
 - Run 4 will see a significant increase
- Number of events is driven by luminosity
 - Increases from 90 fb^{-1} in 2023 to 110 fb^{-1} in 2024
 - Big increase expected in Run 4
- More jobs = increased need for input and output data transfers (FTS and AAA)
- More dynamic usage of disks



Increase in data size

Increase in the need to move data around

Assessing Future Needs (stolen from comp coord talk)

As a baseline, WLCG & experiments did back-of-the-envelope estimates of HL-LHC needs by extrapolating Run 2 network usage by the experiments to PU=200 scales. A lot has changed since then:

- Run 4 start has slipped from 2027 to 2029, with the first full production year 2030 with PU=140 instead of PU=200.
- PU=200 will be reached in Run 5, more than a decade from now.

Still, it's a very good starting point:

T1	LHC Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Flexible Scenario in 2027	Data Challenge target 2027 (Gbps)	Data Challenge target 2025 (Gbps)	Data Challenge target 2023 (Gbps)	Data Challenge target 2021 (Gbps)
CA-TRIUMF	200	400	100	60	30	10
DE-KIT	600	1200	300	180	90	30
ES-PIC	200	400	100	60	30	10
FR-CCIN2P3	570	1140	290	170	90	30
IT-INFN-CNAF	690	1380	350	210	100	30
KR-KISTI-GSDC	50	100	30	20	10	0
NDGF	140	280	70	40	20	10
NL-T1	180	360	90	50	30	10
NRC-KI-T1	120	240	60	40	20	10
UK-T1-RAL	610	1220	310	180	90	30
RU-JINR-T1	200	400	100	60	30	10
US-FNAL	450	900	230	140	70	20
US-FNAL-CMS (atlantic link)	800 1250	1600 2500	400 630	240 380	120 190	40 60
Sum	4810	9620	2430	1450	730	240

The CDR process over the next year should clarify the CMS needs more precisely.

Table 2: data challenge target rates.

Data challenges

- CMS participated in recent WLCG data challenges to test our data movement system as a whole.
- It was particularly important to check tape write speeds, with target rates set at 10% of those estimated during HL-LHC.
- **All sites exceeded or were close to the target rate**

CMS site	March 2022 rates (GB/s)	Target Rates (GB/s)
CTA	5.72	3.2 <i>(to TAPE)</i>
KIT	0.59	0.29
PIC	0.58	0.15
Fermilab	1.97	0.73
RAL	0.8	0.29
IN2P3	0.59	0.29
JINR	0.50	0.55
CNAF	0.32	0.37

Summary

- CMS would like to thank the FTS team for their continued collaboration and support
- Together we have identified a number of improvements
- CMS usage of FTS has not risen above 2019/2020 levels
 - Rates are not expected to rise significantly until Run-4
- CMS has participated in the various WLCG data challenges, and intends to take part in those over the next years as we prepare for HL-LHC