UK mini DC: T1

A. Forti, K. Ellis, A. Rogovskiy

GridPP53

11 April 2025

UK miniDC Goals

- **T1 tests:**
  - Read/Writes to Echo (disk)
    - Capacity tests
      - DC24 T0➔RAL
      - Fill 200 Gb/s T1⟷RAL
    - Cutting LHCOPN to/from RAL
      - Test failover to LHCONE
    - Test T1 JANET link to UK T2s
    - Stress test new checksum code
      - xrootd gateways
  - Read/Write to Antares (tape)
    - Newly attached LHCOPN link
    - New EOS buffer
  - Stress test 100Gb gateways at RAL
    - One writing to Ceph
    - One writing to CephFS
  - .....

- **T2 tests:**
  - Capacity tests T2s
    - Replicate DC24 rates using only UK sites
      - Sustained rates for at least 6h
    - Increase rates to max bandwidth at 100Gb/s sites
      - **CMS results**
      - **ATLAS results**
  - Capability testing under discussion
    - Jumbo frames
    - Token AAI
  - ....

– done
– postponed
– to be done

Schedule and high-level results

Day	Plan	Executed
Monday	Half-fill the OPN: writes to Echo from CERN	More than half-filled
Tuesday	Simulate ‘cutting’ of OPN; fallback to LHCONE	Mixed results
Wednesday	Revert to OPN and saturate; data from CERN/T1s	Some disturbance to gws
Thursday	Reads from Echo; data to CERN/T1s	Easily achieved
Friday	Tests with UK Tier 2 sites; postponed	Tests with chksum-on-the-fly

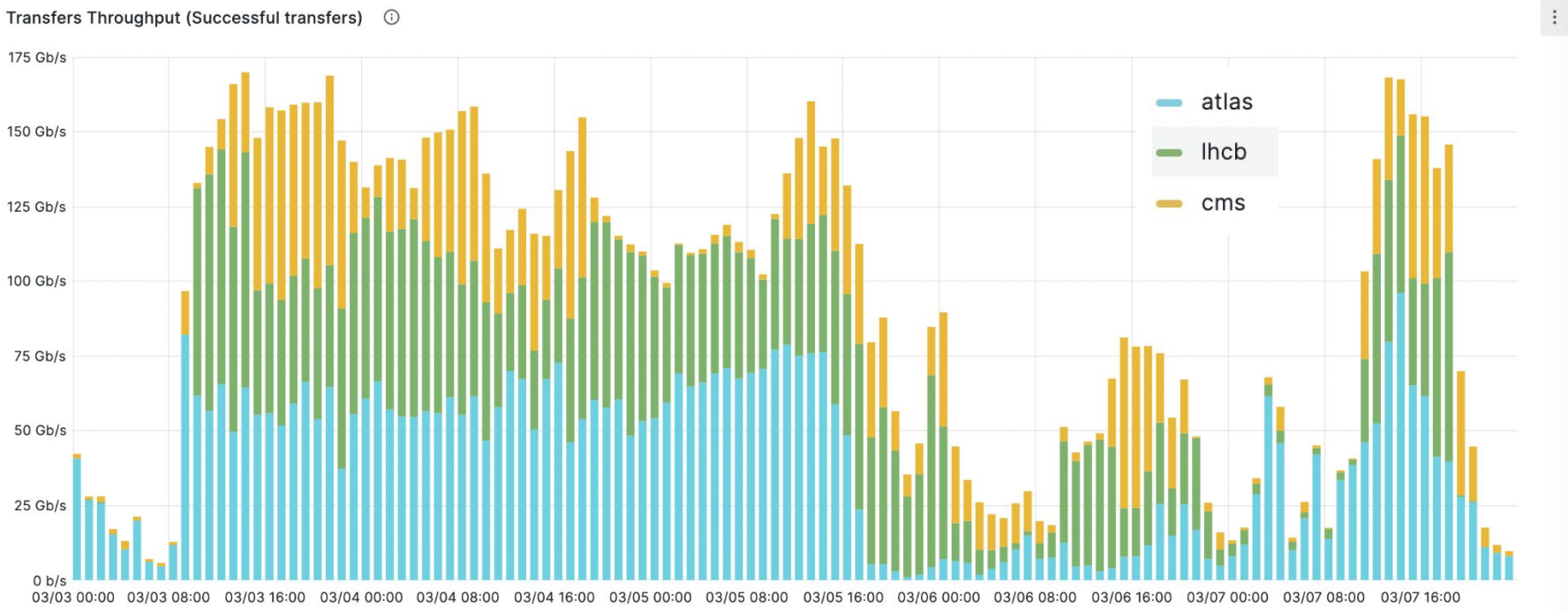
Setup&Rates

	CERN to RAL	CERN to RAL	CERN/T1 to RAL	RAL to CERN/T1	CERN/T1 to RAL
	MON	TUES	WED	THU	FRI
ATLAS	44000	44000	66667	36000	66667
LHCb	31680	31680	66667	21120	66667
CMS	33208	33208	66667	28408	66667
Total	108888	108888	200000	85528	200000

- Setup:
  - Each VO used similar file size ~5GB
  - ATLAS and CMS ran the dc_inject tool
    - creates and deletes rules in Rucio periodically, and continuously until stopped
  - LHCb created a new tool to inject data continuously (unlike previous methods)
    - Could be used in subsequent DCs
- Replicate the DC24 rates
  - Fill bandwidth where much larger than DC24 rates
    - LHCOPN 200 Gb/s

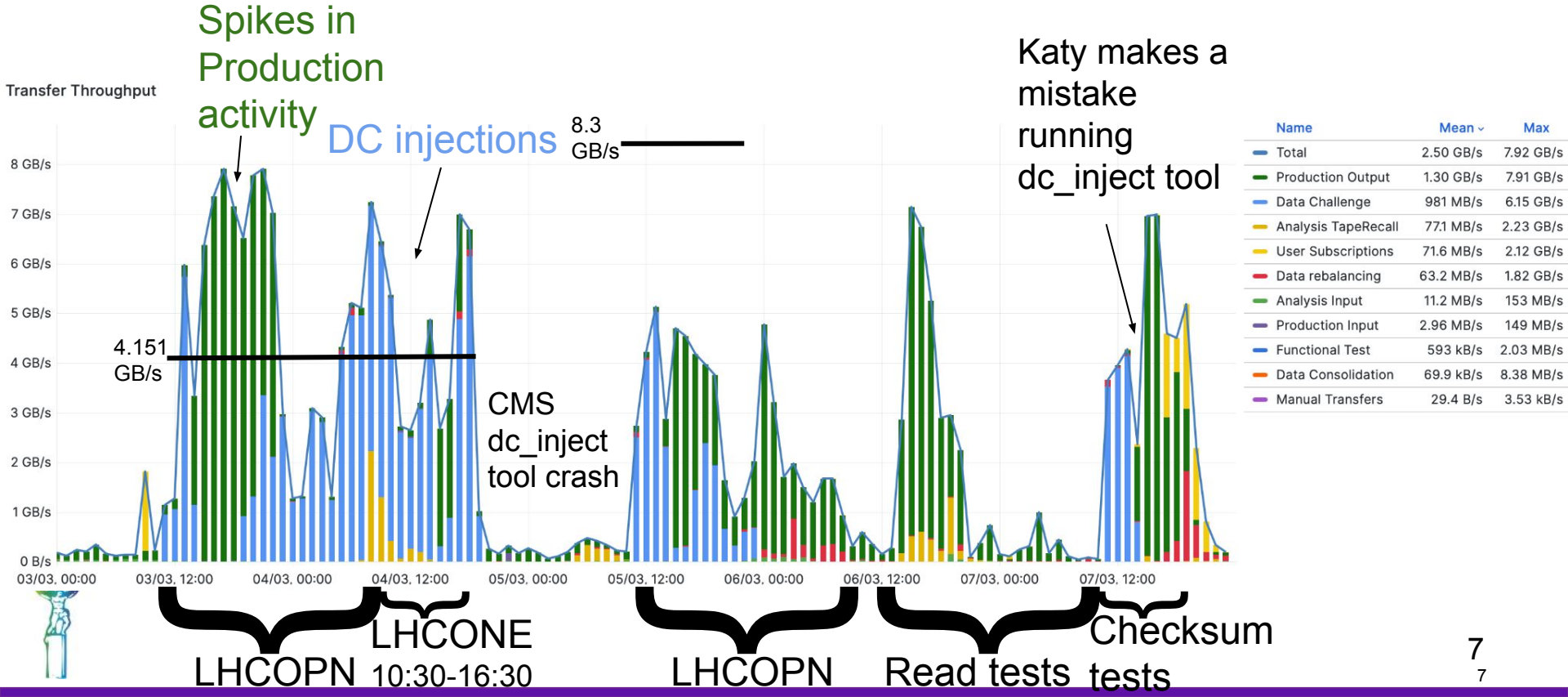
Tier1 results

Ingress



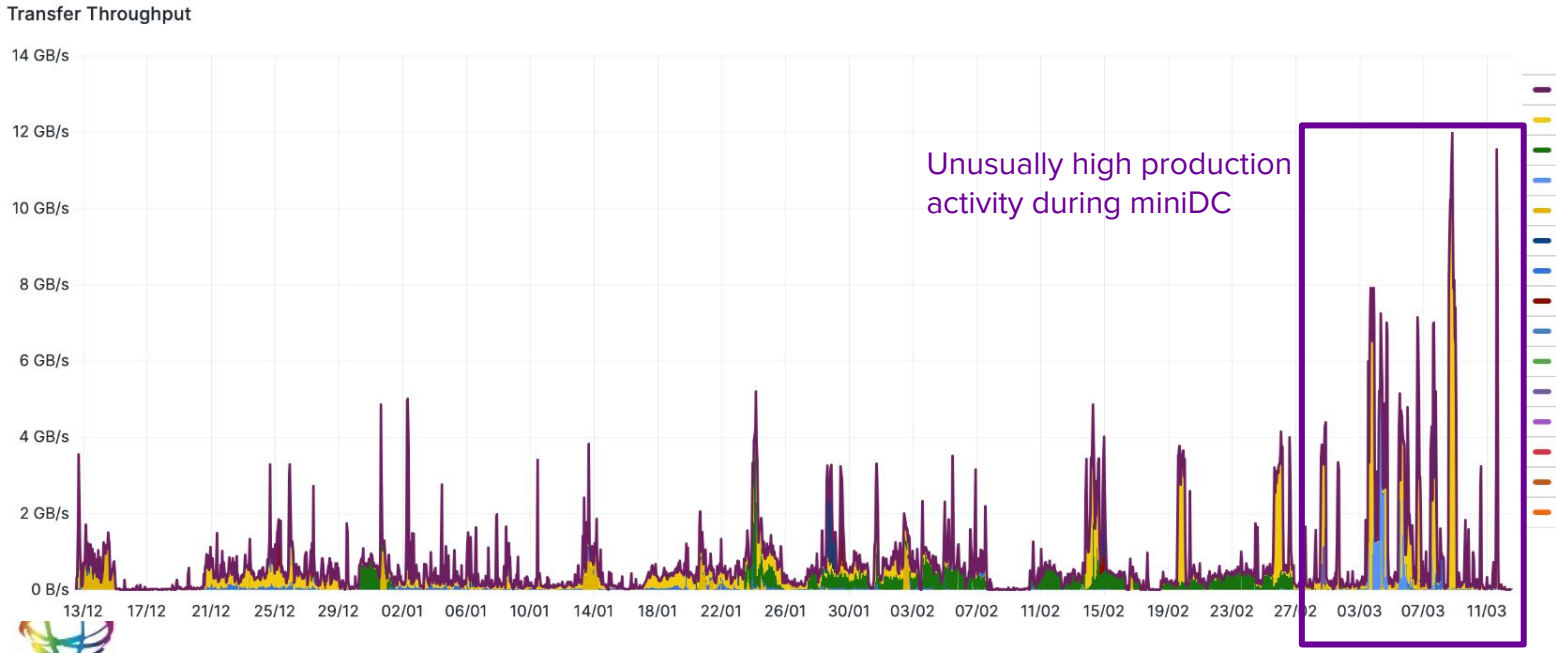
- Talk mostly about ingress

FTS Transfers to Echo: CMS



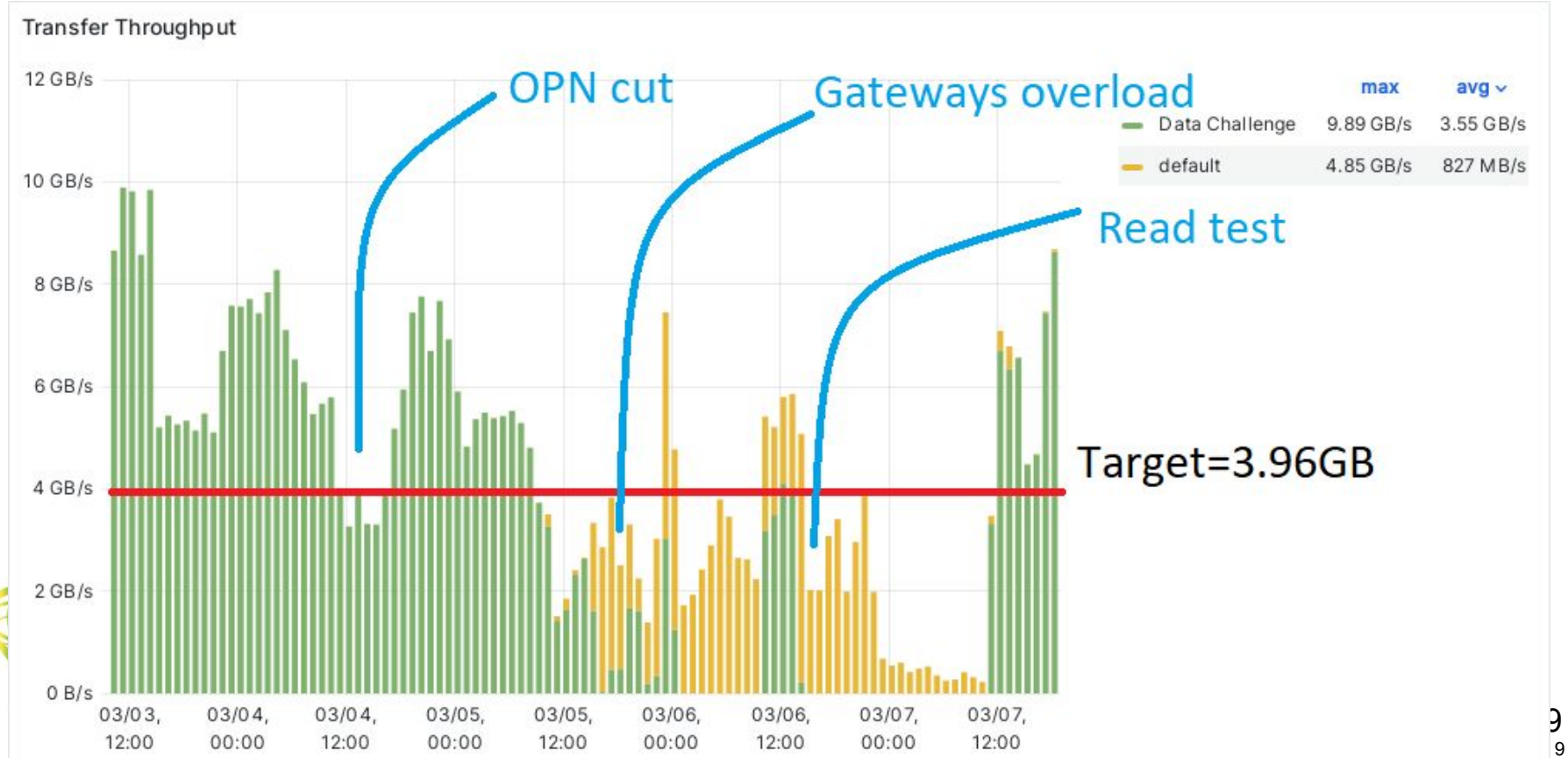
CMS FTS transfers to Echo

(last 3 months)



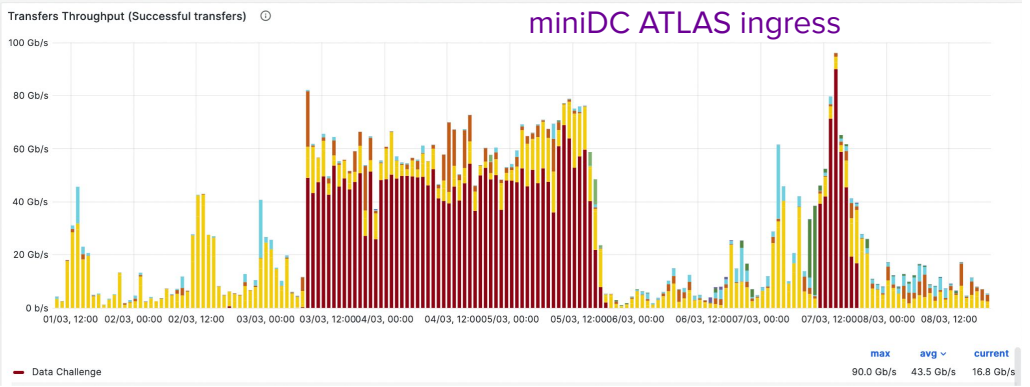
Name	Mean	Max
Total	725 MB/s	12.0 GB/s
Production Output	371 MB/s	11.9 GB/s
Production Input	136 MB/s	4.75 GB/s
Data Challenge	54.5 MB/s	6.15 GB/s
Analysis TapeRecall	51.1 MB/s	4.58 GB/s
User Subscriptions	38.2 MB/s	4.45 GB/s
Data rebalancing	36.5 MB/s	1.82 GB/s
User AutoApprove	14.7 MB/s	2.58 GB/s
Recovery	9.70 MB/s	2.76 GB/s
Analysis Input	6.44 MB/s	586 MB/s
Manual Transfers	5.82 MB/s	2.24 GB/s
Functional Test	518 kB/s	4.20 MB/s
Data Rebalancing	108 kB/s	196 MB/s
T0 Export	92.5 kB/s	27.4 MB/s
Data Consolidation	3.88 kB/s	8.38 MB/s

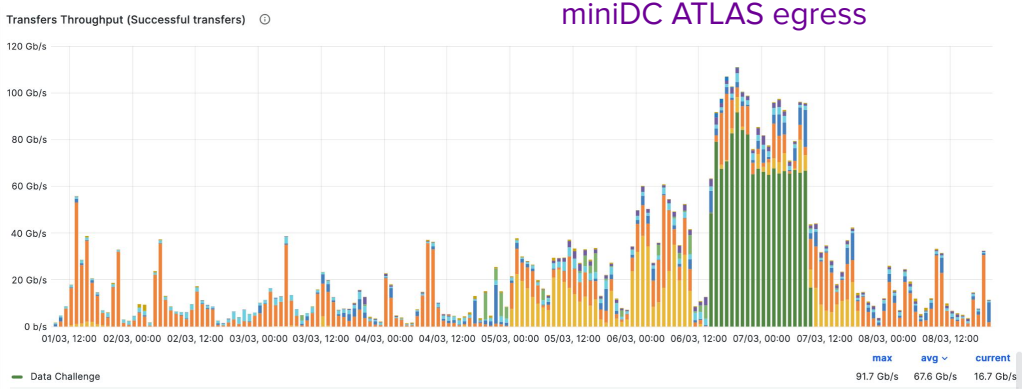
FTS Transfers to Echo: LHCb



ATLAS general results

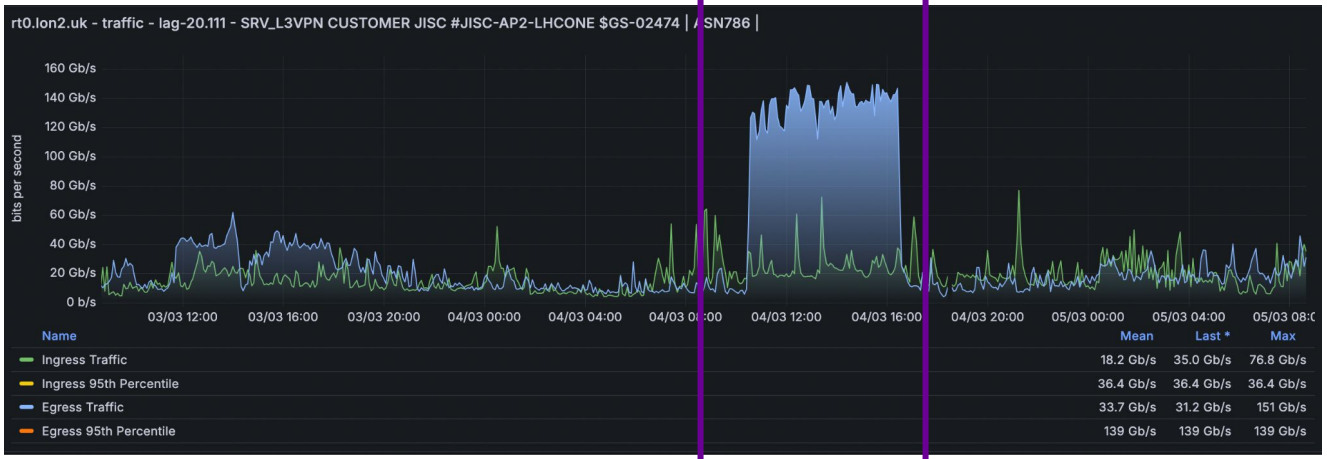
- UK T1 disk miniDC was a success
  - Most of the problems ATLAS had during DC24 have disappeared
    - T0 export rates
    - General rates
    - Tokens
  - LHCOPN cut didn't affect rates
  - Checksum new code doesn't have obvious side effects



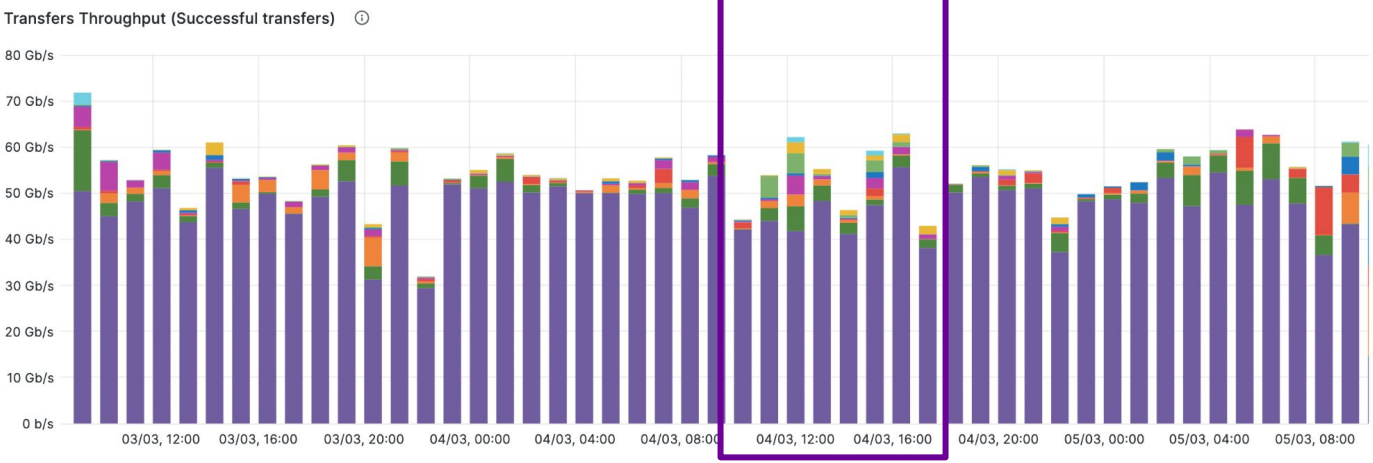


LHCOPN cut

LHCOPN UK-RAL Total - 200Gbps
100 Gb
50 Gb
0 b
-50 Gb
-100 Gb
-150 Gb
-200 Gb
-250 Gb
03/03, 12:00
03/03, 16:00
03/03, 20:00
04/03, 00:00
04/03, 04:00
04/03, 08:00
04/03, 12:00
04/03, 16:00
04/03, 20:00
05/03, 00:00
05/03, 04:00
05/03, 08:00
min max avg current total
in total - RAL to CERN 0 b 75.5 Gb 11.1 Gb 14.1 Gb 6.34 Tb
out total - CERN to RAL 0 b 230 Gb 132 Gb 72.1 Gb 75.0 Tb

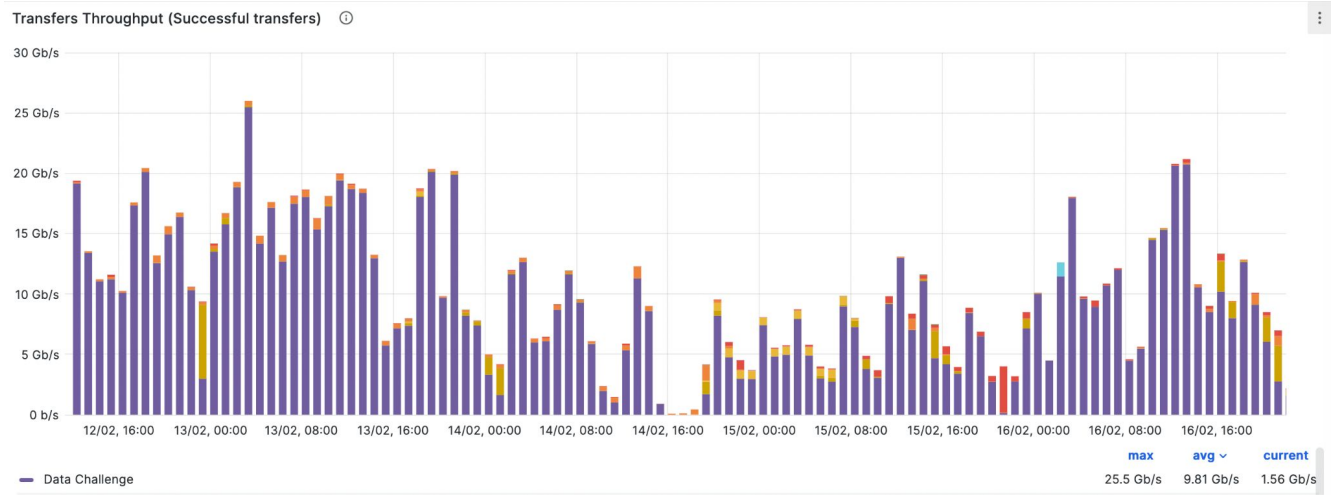

rt0.lon2.uk - traffic - lag-20.111 - SRV_L3VPN CUSTOMER JISC #JISC-AP2-LHCONE $GS-02474 | / SN786 |
bits per second
160 Gb/s
140 Gb/s
120 Gb/s
100 Gb/s
80 Gb/s
60 Gb/s
40 Gb/s
20 Gb/s
0 b/s
03/03 12:00
03/03 16:00
03/03 20:00
04/03 00:00
04/03 04:00
04/03 12:00
04/03 16:00
04/03 20:00
05/03 00:00
05/03 04:00
Name Mean Last * Max
Ingress Traffic 18.2 Gb/s 35.0 Gb/s 76.8 Gb/s
Ingress 95th Percentile 36.4 Gb/s 36.4 Gb/s 36.4 Gb/s
Egress Traffic 33.7 Gb/s 31.2 Gb/s 151 Gb/s
Egress 95th Percentile 139 Gb/s 139 Gb/s 139 Gb/s

LHCOPN cut

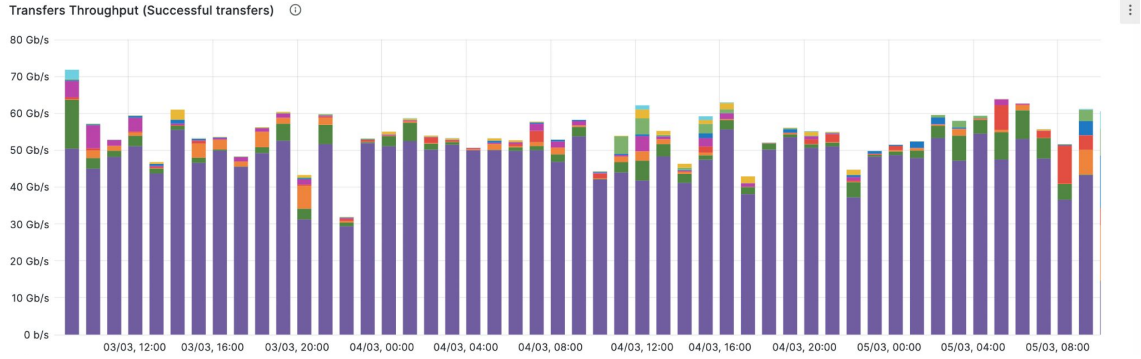


ATLAS LHCOPN (T0+T1) rates remained in average

ATLAS T0 export rates

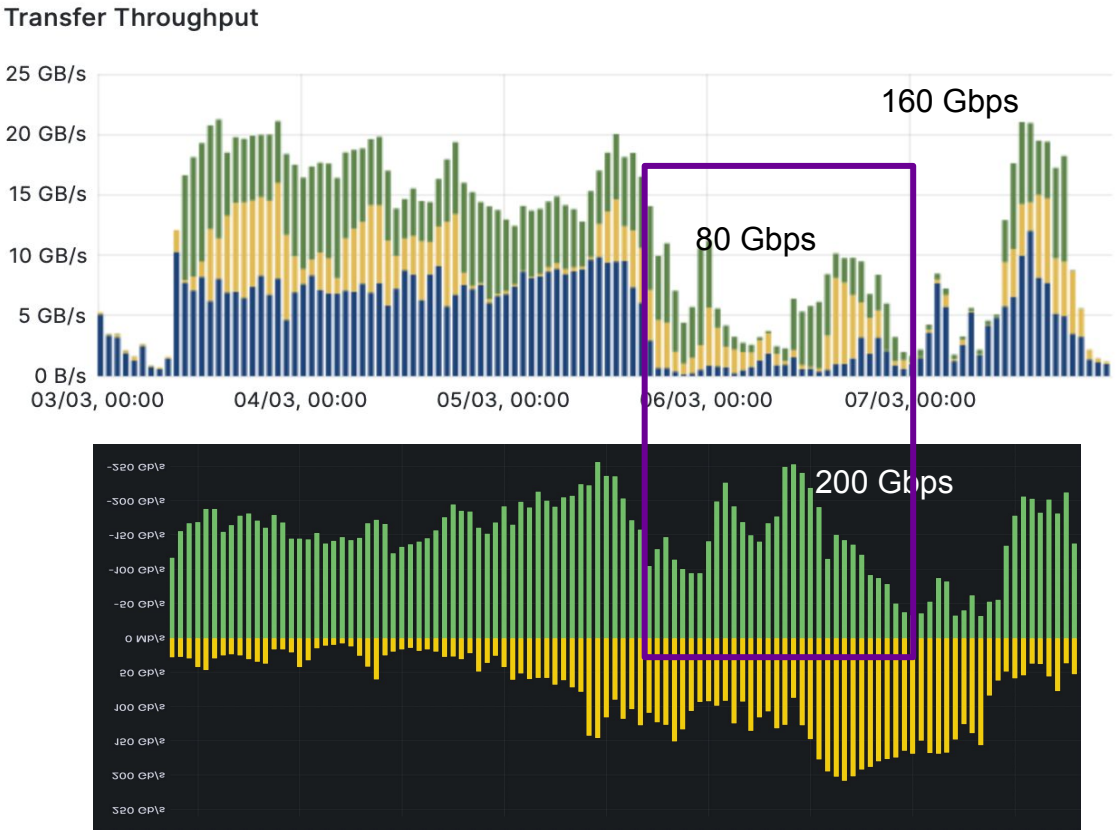


DC24 T0 export expected rates 44 Gb/s achieved ~ 10 Gb/s wobbling due to several concurrent problems



miniDC T0 export achieved ~51 Gb/s sustained. Expected still 44 Gb/s

FTS vs total (IN to RAL)

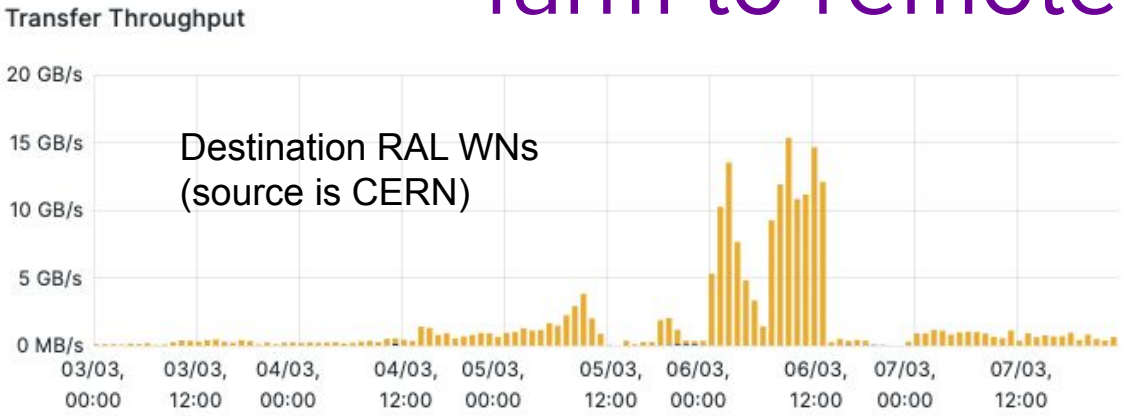


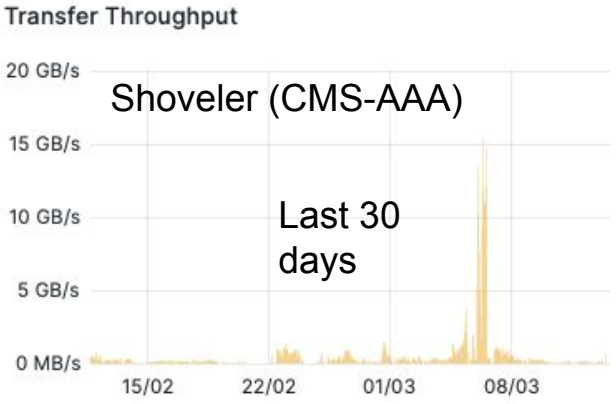
Non FTS traffic to RAL.

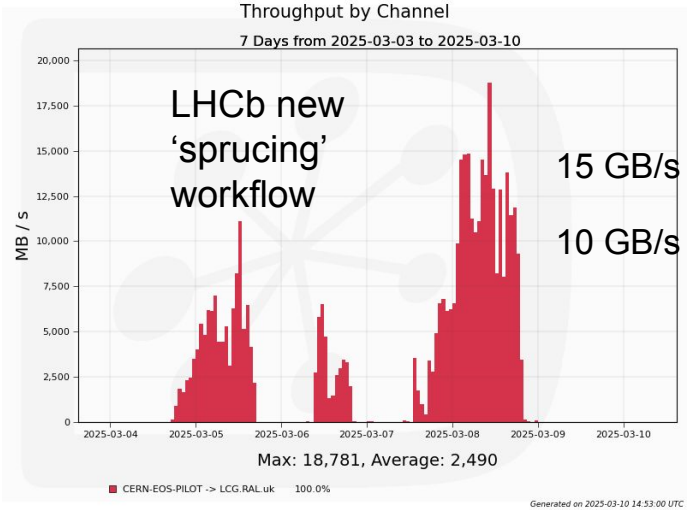
CMS AAA jobs pulling data over root

New LHCb workflow pulling data over https

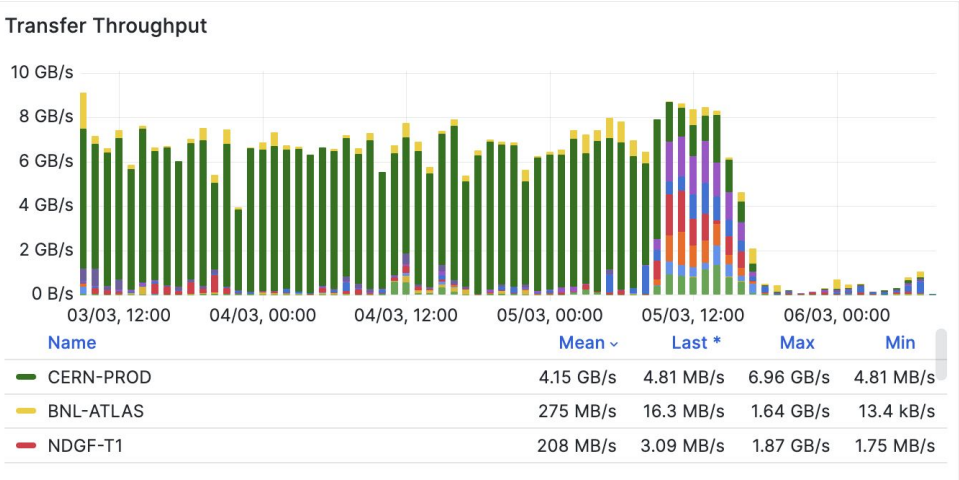
Direct reads from batch farm to remote site



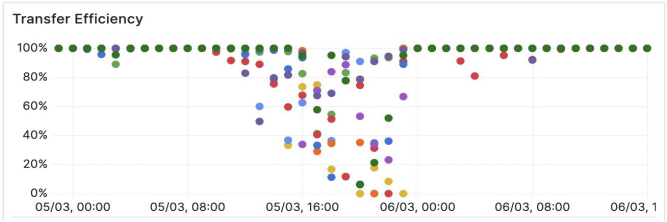




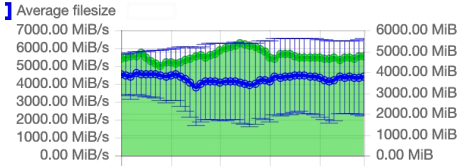
ATLAS Traffic from/to T1

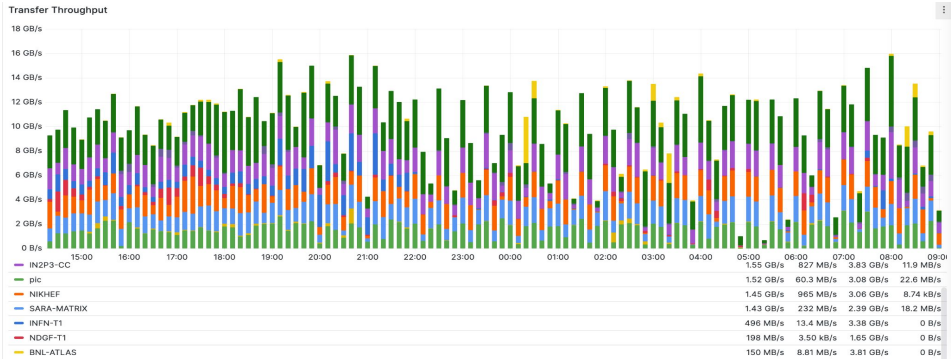


Ingress smooth switch from CERN only miniDC traffic to mixture of T1s but added load on gateways added a tail of time outs. 100 Gb/s nics should solve this

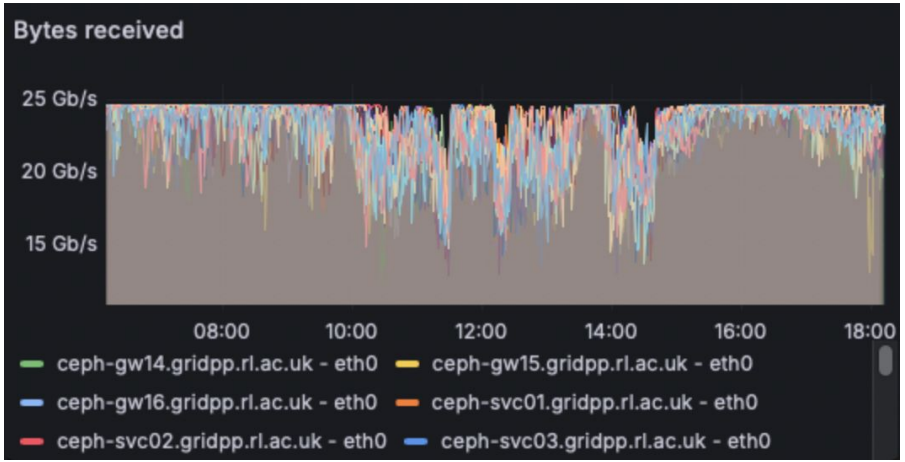


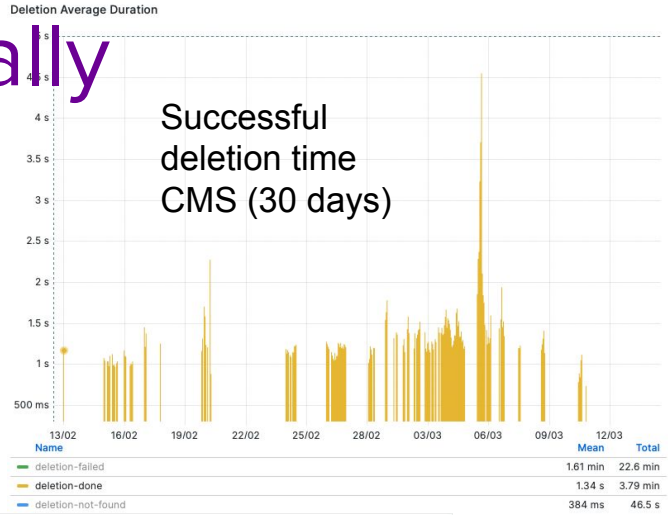
egress tooth comb pattern due to T1s absorbing files too fast. Positive but might be also due to the system injecting smaller files.





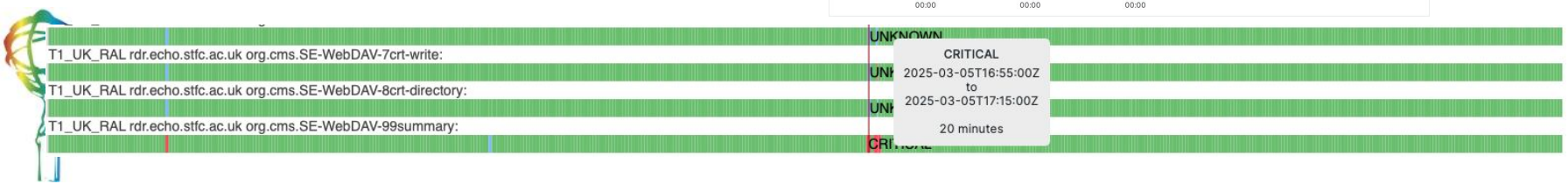
Gateways maximally stressed (Wed)





Successful deletion time CMS (30 days)

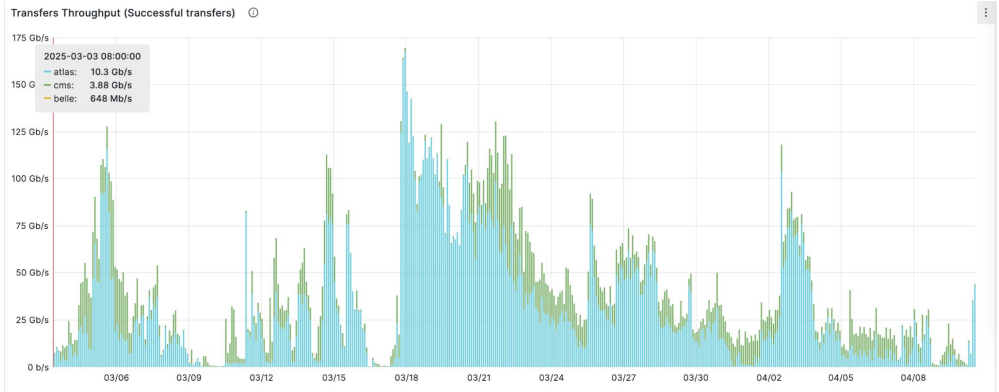
LHCb transfer efficiency



Next tests: Antares

Reasons to test tape

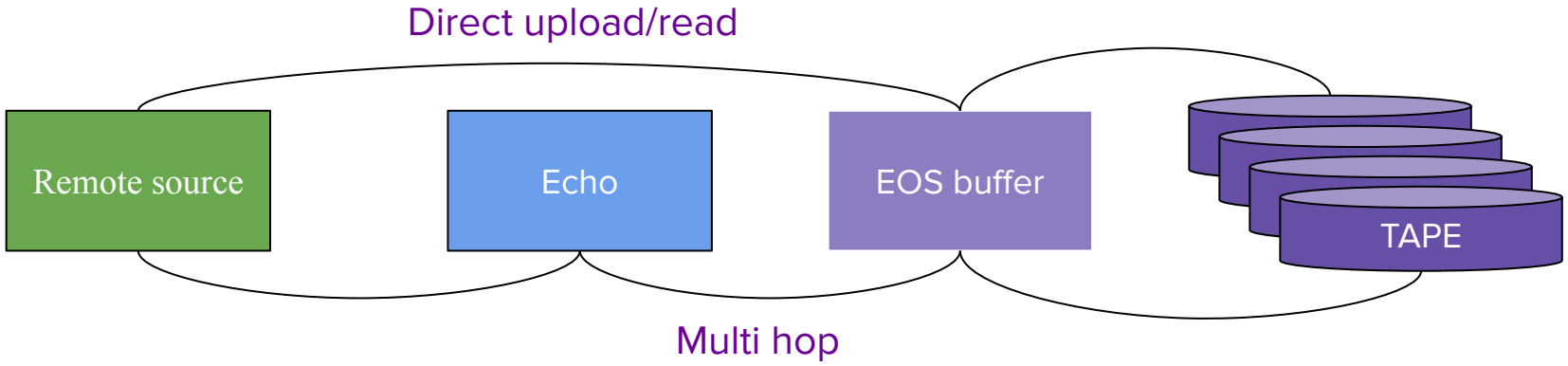
- Tape is a storage system for all the LHC experiments
  - We tested T0 export disk to disk but in production it is disk to tape
  - Archiving rates for HL-LHC might/will go dramatically up (CMS numbers later)
  - ATLAS also uses tape dynamically for prod jobs and now also for analysis (data carousel)
- Tape will likely be included in DC27
- Antares will soon have new EOS buffer (from Alastair's talk
  - 7 x new EOS nodes with NVMEs and 100 Gb/s
- New nodes will be accessible via LHCOPN and LHCONE.
- The rest API was deployed
  - ➜ http transfer protocol for reading



atlas dominant staging rates

Testing

- Tape testing more complicated
  - Write
    - To EOS buffer and then packing on tape
    - Sometimes multi-hop through echo
      - LHCb DC24 was done hopping
  - Read
    - Stage data to EOS buffer and then copy to destination
    - Sometimes multihop through echo



Questions to answer

- Tape cannot be easily deleted/rewritten
  - How much should we write Alastair in November didn't think we need huge rates, but…
    - we probably should try DC24 rates with tape or understand where we can get
  - Is we cannot delete do we have enough space?
- Read tests should be easier…recall efficiency is also important
- What tape families to write to?
  - Do we need to make it realistic or can just write to one 'DC' test family?
- A mixture of direct transfer from CERN to Antares and multihop through Echo?
- Can we use all the tape drives simultaneously for writing?
- What about concurrent reading/writing?
- What happens if/when we fill the buffers?
- Tune to FTS behaviour
- Is the monitoring correct to assess the rates?
  - Rucio monitoring might differ from FTS and Dirac…

Tape staging (early) modelling

- In the longer term experiments should look at the archiving/staging modelling as part of reviewing HL-LHC numbers
  - CMS already started to model tape staging numbers based on doing a re-reco and a re-mini in the 5 months between data-taking
  - Using current data model. Numbers are scary…
    - Don't even ask about Run-5!

Run-4

GB/s	T0		Re-Reco		Re-Mini		MC (incl. PU)	
Data Flow	archiving	staging	archiving	staging	archiving	staging	archiving	staging
CERN	61.6	--	7.8	19.4	2.4	16.2	0.5	--
KIT	2.6	--	0.8	1.9	0.4	2.5	1.2	--
PIC	1.3	--	0.4	1.0	0.2	1.4	0.7	--
IN2P3	2.6	--	0.8	1.9	0.4	2.5	1.2	--
CNAF	3.9	--	1.2	3.0	0.6	3.9	1.7	--
RAL	2.6	--	0.8	1.9	0.4	2.5	1.2	--
Serbia	2.6	--	0.8	1.9	0.4	2.5	1.2	--
FNAL	10.5	--	3.1	7.7	1.5	10.2	1.4	--

UNOFFICIAL!

Conclusions

- Echo tests were a success
  - Many of the problems we had in DC24 have been solved
    - Some have still to be looked at, particularly the GW overload
- Initial framing of tape tests
  - what we want to get out of it
  - how to make experiments results comparable
  - how to monitor correctly the rates
- Longer term staging modelling
  - As part of the validation of current HL-LHC general rates tape traffic big part.

Backup

LHCb Inject tool (backup)

- Similar concept to Atlas and CMS, but the implementation is a bit different
  - Test dataset is passed to the script (not selected by the script)
  - FTS jobs are then submitted for individual files
  - Once the file is copied, it is deleted and re-submitted again
  - So, “constant” load is created on the given channel
- Channels can be multi-hop
  - Managed by DIRAC though, not FTS
- No concept of the desired throughput
  - Can be controlled on the FTS side to some extent
- Can be useful e.g. for smaller sites in the future challenges, or when the free space is limited
  - Small sites used to finish their chunk of work before others during DC24

LHCb Mini-DC goals (backup)

- Prove that the DC24 target for RAL can be achieved
  - Had not happened during the actual DC24, for various reasons
- Try new injection tool
  - Concept similar to Atlas and CMS
- Test system's performance when the OPN is cut
  - Applies to all VOs, probably
- *(optional)* test RAL->CERN channel
  - Is not used in production very often