DC 27:
Estimates for Rate Requirements

Katy Ellis (RAL), Christoph Wissing (DESY)

DOMA General Meeting

Oct 2025

Introduction

- CMS is in the process of writing a Conceptual Design Report (CDR) for HL-LHC computing
  - Present planning numbers used as input also for DC27
- Focus is the modelling for Run-4, most relevant for DC27
  - Run-5 will be approximately double Run-4 rates
  - But we have a few more years to achieve those rates
- CMS is exposing preliminary network AND TAPE planning numbers
  - We would like to encourage other experiments to do similar exercises (and expose results)
  - A common set of planning numbers to be composed within DOMA
  - Arrive at a DC27 program and a planning path towards it (including mini-DCs etc)

How to estimate?

How much data do we have and how much do we want/need to move it around?

- Quantity of data depends on:
  - Number of events (real and MC)
  - Size of events (varying data formats)
  - Also the number of copies,
  - Data tiers kept/discarded, etc.
- The amount of network/tape activity also depends on:
  - How much data remains at CERN (or where it was generated)
  - How tight is the disk situation?
  - Other things?

CMS Scenarios & Parameters

Present CMS planning (corresponding CDR document not yet published)

- Recent numbers (as of October 2025):
  - **Run-4 (relevant for DC27)**
    - PU: 150
    - Trigger rates:
      Primary: 10kHz; HLT Scout: 150kHz;
    - Events sizes:
      RAW: 4.6MB; AOD(SIM): 1.5MB; MiniAOD(SIM): 190kB; NanoAOD(SIM): 4kB; HLT Scout: 18kB
    - L1Scouting at max 5GB/s
  - Run-5 (listed for completeness)
    - PU: 200
    - Trigger rates:
      Primary: 15kHz; HLT Scout: 225kHz;
    - Events sizes:
      RAW: 5.9MB; AOD(SIM): 2.0MB; MiniAOD(SIM): 250kB; NanoAOD(SIM): 4kB; HLT Scout: 32kB
    - L1Scouting at max 5GB/s

CMS Scenarios & Parameters

Data handling aspects

- RAW of primary datasets: Full tape copy at CERN; 50% with tape replica at T1s
- Scouting data: Single tape copy at CERN
- AOD(SIM):
  - Produced for all primary datasets in prompt and re-reco, archived (primarily) at T1
  - AODSIM is produced and archived for all MC
  - Usually not on disk, staging required for access, e.g. re-Mini campaigns
- MiniAOD(SIM):
  - Like for AOD, fully produced and archived at T1s
  - Recent version on disk
- Export for processing after staging
  - A significant fraction of staged data will be transferred for processing, e.g. to strong T2s - fraction is unclear
  - Assumption: Products of processing are archived, where input was originally staged

Exact flows are unknown as of today
Some reasonable assumptions are made

CMS Scenarios & Parameters

Present considerations about sites

T1 ingress	Fractions[%]
KIT	10
PIC	5
IN2P3	10
CNAF	10
RAL	7
Serbian T1	5
NCBJ	5
JINR	8
FNAL	40

- Tier-1s
  - Shares mainly derived from 2026 tape pledges
  - Well established: KIT, PIC, CNAF, IN2P3, RAL, FNAL, expected to stay
  - JINR: political status uncertain, considered in planning with 2026 pledge fraction
  - Serbia: No 2026 pledge yet
  - NCBJ (Poland): Slight ramp up expected compared to 2026 pledge
- Tier-2s
  - What might change by Run-4?
    - Concentration of storage at few locations?
    - Fewer end points in CMS DM system?
  - So far only a rough guess about the sum for all T2 - currently handle a lot of User activity
- Analysis Facilities
  - Might consume substantial bandwidth?
  - So far not included in any modeling

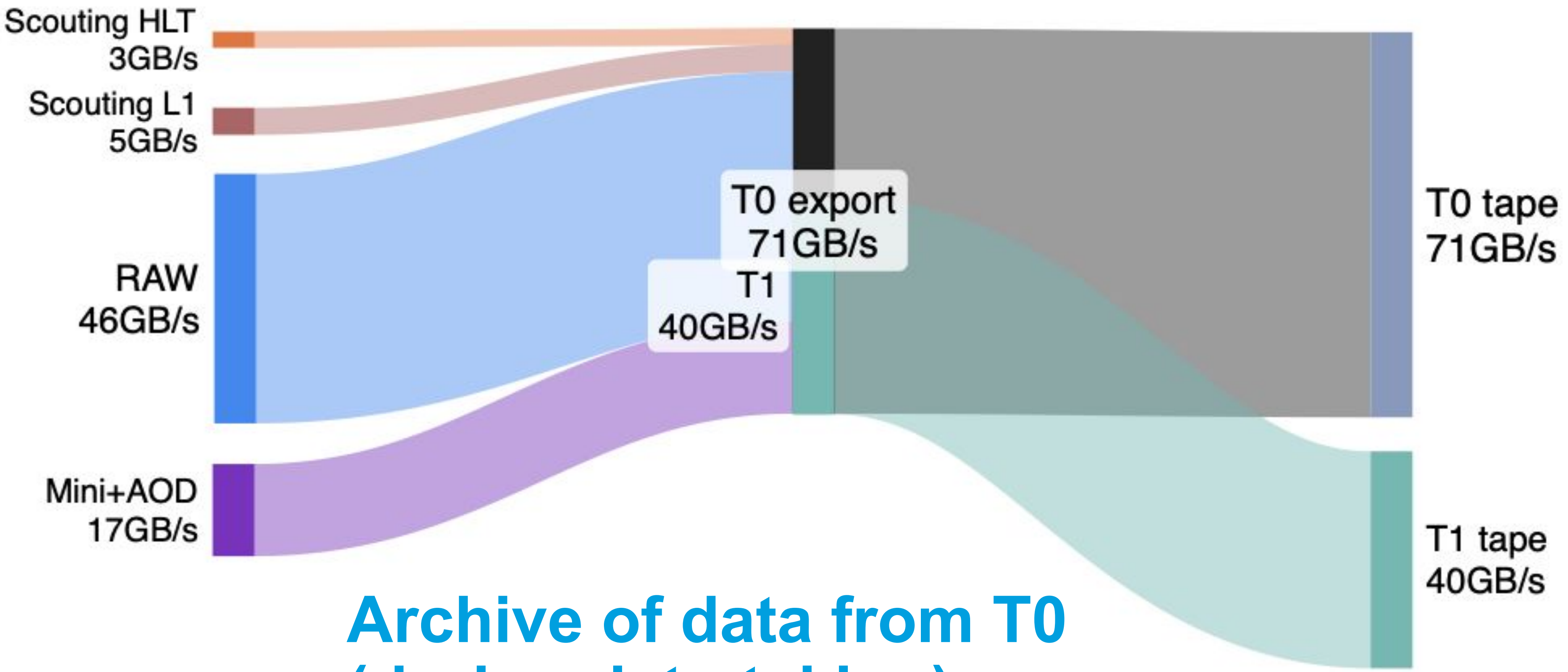
Tier-0 Archiving & Data Export

	Run-4 PU150 @10kHz				Run-5 PU200 @15kHz		
	Size[MB]	Exp. Logging[	Rate[GB/s]		Size[MB]	Logging[kHz]	Rate[GB/s]
RAW	4.6	10	46		5.9	15	88.5
(Mini+)AOD	1.69	10	16.9		2.25	15	33.75
HLTScouting	0.018	150	2.7		0.022	225	4.95
L1 Scouting	0.45	-	5		0.45	-	5
			70.6				132.2

	Run-4 PU150 @10kHz				Run-5 PU180 @15kHz	
T1 ingress	Fractions[%]	Rate [GB/s]	Connectivity [Gbs]		Rate [GB/s]	Connectivity [Gbs]
KIT	10	4.0	130		7.1	230
PIC	5	2.0	60		3.5	110
IN2P3	10	4.0	130		7.1	230
CNAF	10	4.0	130		7.1	230
RAL	7	2.8	90		5.0	160
Serbian T1	5	2.0	60		3.5	110
NCBJ	5	2.0	60		3.5	110
JINR	8	3.2	100			
FNAL	40	16.0	510		28.4	910
Summed T1s	100	39.9	1270.0		19.7	570
Summed T2s		8.5	270		15.2	490
CERN egress	100	48.4	1550		100.1	3200

- Connectivity assumes a factor 4 overprovisioning, consistent with WLCG planning
- 50% of RAW exported from CERN
- Overall connectivity still consistent with WLCG planning 2020: 2Tbs
- L1 Scouting set to 5GB/s
- No LHC duty cycle assumed. Full rate is exported ("upper limit")
- AOD and MiniAOD exported to T1s

New diagrams!



Archive of data from T0 (during data-taking)

Re-Reco Scenario in Run-4 for one year

Scenario: CMS wants to run re-reco campaign during YETS, staging the data in **90 days**

Assumptions:
- LHC uptime 7e6 sec (note CDR: pp 6e6 sec + 1.2e6 sec HI)
- Half of RAW staged from T1s, other half from T0
- CERN processes 70% locally of what was staged there (lots of local CPU, e.g. HLT)
- Exported data processed at (strong) T2s
- Re-Reco products (40% of RAW) get archived where corresponding RAW was staged
- Only T1s absorb reconstructed (mini)AOD from T2s and CERN
- Archiving overlaps staging

	Staging Vol PB	Staging Rate [GB/s]	Export frac	egress Rate [GB/s]	Archiving Rate [GB/s]	ingress Rate [GB/s]	egress Connectivity [Gbs]	ingress Connectivity [Gbs]
CERN	161	20.70	0.3	12.01	0.00	0.00	380	0
KIT	16	2.10	0.5	1.05	1.65	1.31	30	40
PIC	8	1.00	0.5	0.50	0.83	0.65	20	20
IN2P3	16	2.10	0.5	1.05	1.65	1.31	30	40
CNAF	16	2.10	0.5	1.05	1.65	1.31	30	40
RAL	11	1.40	0.5	0.70	1.16	0.91	20	30
Serbian T1	8	1.00	0.5	0.50	0.83	0.65	20	20
NCBJ	8	1.00	0.5	0.50	0.83	0.65	20	20
JINR	13	1.70	0.5	0.85	1.32	1.04	30	30
FNAL	64	8.20	0.5	4.10	6.61	5.22	130	170
Summed T2s				8.26		20.65	260	660

Comment on tape

- Tapes throughput is limited by the number of tape drives and the drive speed
  - Typical drive speed at time of presentation: 400MB/s

- But the *effective* rate is always (a bit) lower due to e.g. having to change tapes, transfer failures, DM system not keeping the pressure at an optimum rate…

- For READS the effective rate can be *much* lower depending on
  - How well the data was ordered when written
  - Type of recall pattern, etc.

Re-MiniAOD Scenario in Run-4

Scenario: CMS wants to run re-mini campaign during YETS, staging the data in **60 days**

Assumptions:

- Campaign for one year of data and one year MC
- MC events per year 95B (likely a bit low)
- All AOD(SIM) staged at T1s, AODSIM perhaps not produced by default.
- Half of processing at (strong) T2s
- miniAOD (+potential nanoAOD) volume is ~20% of AOD volume, archived at T1s

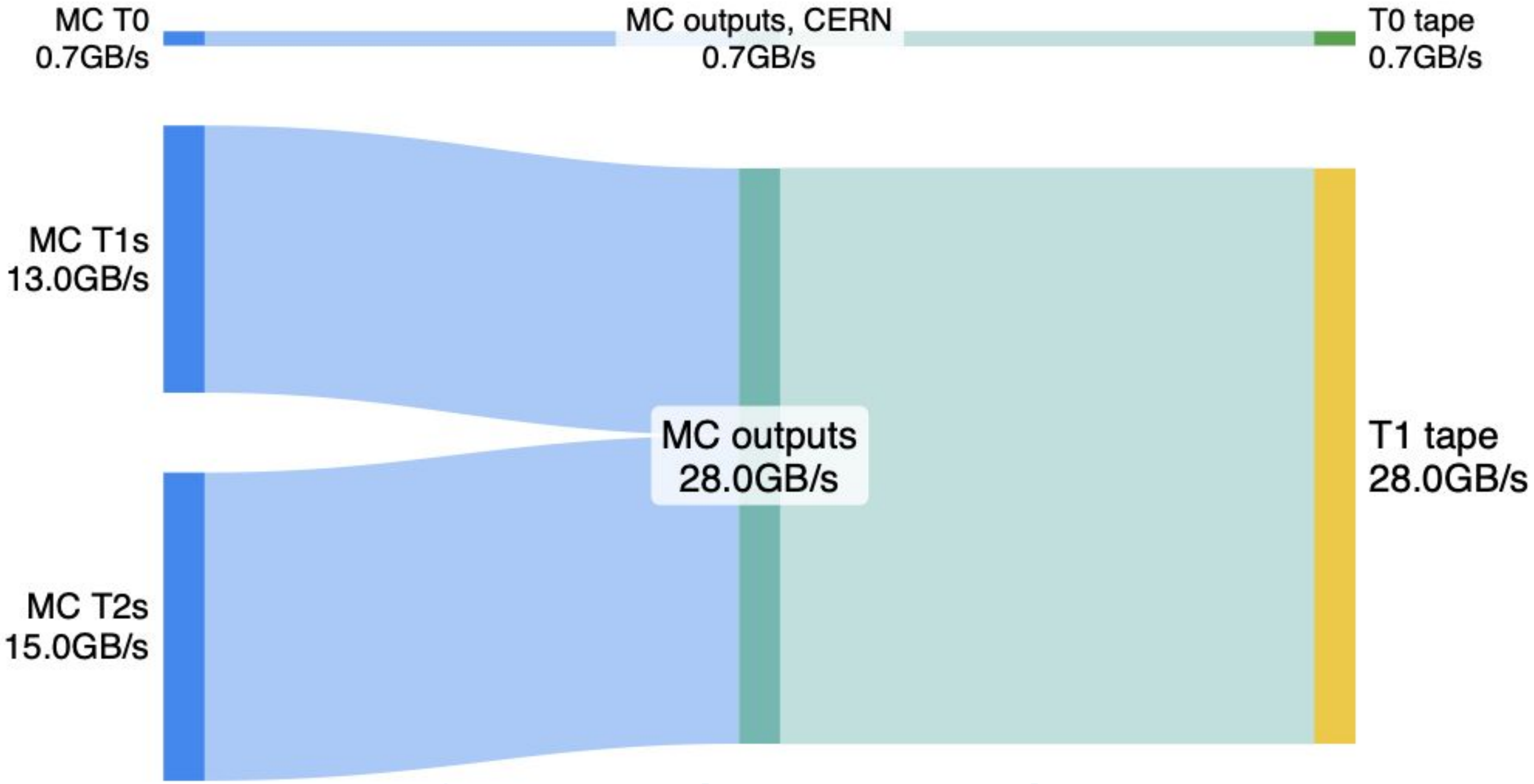
	Staging Vol PB	Staging Rate [GB/s]	Export frac	egress Rate [GB/s]	ingress Rate [GB/s]	egress Connectivity [Gbs]	ingress Connectivity [Gbs]
CERN (no AOD)	0	0.00	0.5	0.00	0.00	0	0
KIT	23	4.40	0.5	2.20	0.44	10	0
PIC	12	2.30	0.5	1.15	0.23	0	0
IN2P3	23	4.40	0.5	2.20	0.44	10	0
CNAF	23	4.40	0.5	2.20	0.44	10	0
RAL	16	3.10	0.5	1.55	0.31	10	0
Serbian T1	12	2.30	0.5	1.15	0.23	0	0
NCBJ	12	2.30	0.5	1.15	0.23	0	0
JINR	19	3.70	0.5	1.85	0.37	10	0
FNAL	93	17.90	0.5	8.95	1.79	40	5
Summed T1s	233	44.8		22.4	4.48	90	5
Summed T2s				4.48	22.4	20	90

MC Production

Comments:

- Model is super simple: flat production rate over the year
- Network and tape rates not very demanding (in step-chain like approach)
- Premix delivery is substantial
    - PileUp is served from FNAL and CERN
    - Likely underestimated in the following table
    - However there are efforts to reduce bandwidth at slightly increased CPU

MC tape archiving



Occurs (averaged) all year round

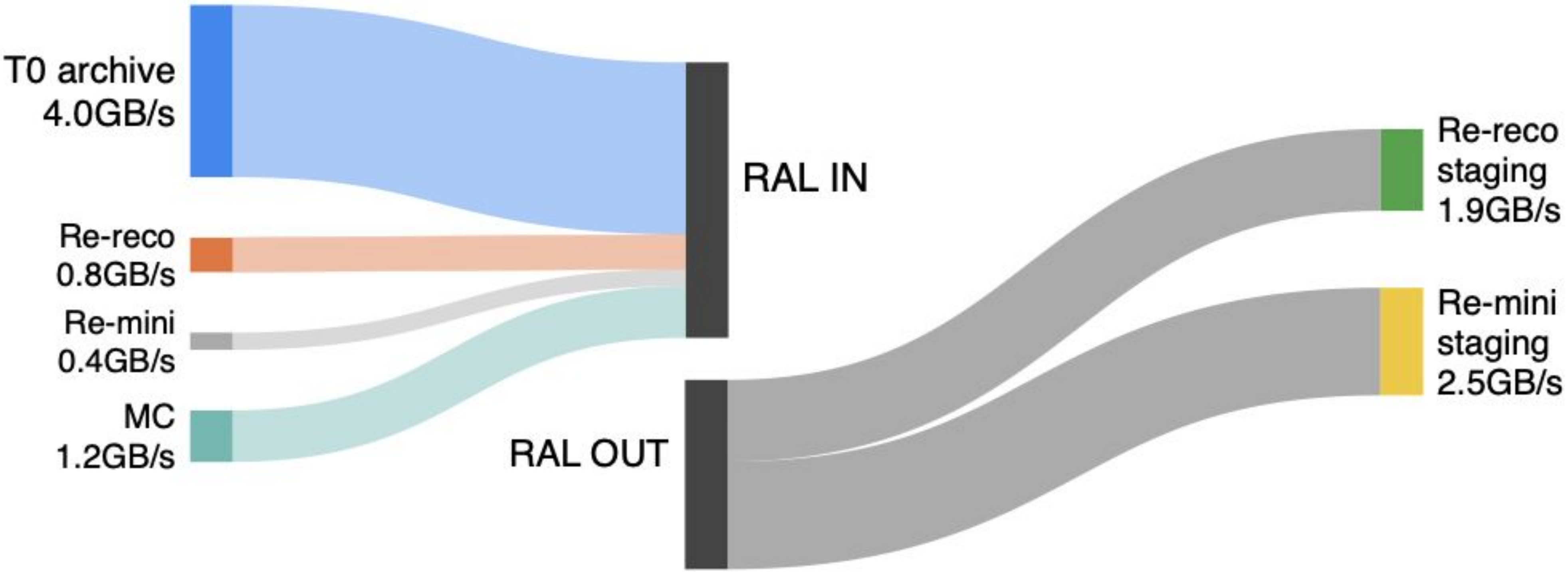
Tape Staging/Archiving Overview

Run-4								
	T0		Re-Reco		Re-Mini		MC (incl. PU)	
Data Flow	archiving	staging	archiving	staging	archiving	staging	archiving	staging
CERN	70.6	--	8.3	20.7	0.0	0.0	0.6	--
KIT	4.0	--	0.8	2.1	0.7	4.4	1.3	--
PIC	2.0	--	0.4	1.0	0.3	2.3	0.8	--
IN2P3	4.0	--	0.8	2.1	0.7	4.4	1.3	--
CNAF	4.0	--	0.8	2.1	0.7	4.4	1.3	--
RAL	2.8	--	0.6	1.4	0.5	3.1	1.3	--
Serbia	2.0	--	0.4	1.0	0.3	2.3	0.8	--
NCBJ	2.0	--	0.4	1.0	0.3	2.3	0.8	--
JINR	3.2	--	0.7	1.7	0.6	3.7	1.3	--
FNAL	16.0	--	3.3	8.2	2.7	17.9	1.4	--
Summed T1s	39.9	--	8.24	20.6	6.7	44.8	10.3	--

Throughput in units of GB/s

Certainly some flows are not included, e.g. staging for user analysis jobs

Site example for key tape scenarios



Network Connectivity Overview

Run-4								
All in Gbs	T0 Export		Re-Reco		Re-Mini		MC (incl. PU)	
Data Flow	egress	ingress	egress	ingress	egress	ingress	egress	ingress
CERN	1550		380	0	0	0	380	20
KIT		130	30	40	10	0	5	40
PIC		60	20	20	5	0	5	25
IN2P3		130	30	40	10	0	5	40
CNAF		130	30	40	10	0	5	40
RAL		90	20	30	5	0	5	40
Serbia		60	20	20	5	0	5	25
NCBJ		60	20	20	5	0	5	25
JINR		100	30	30	5	0	5	40
FNAL		510	130	170	35	5	380	45
Summed T1s		1270	330	410	90	5	420	320
Summed T2s		270	260	660	20	90	130	485

Connectivity is given in Gbs including a factor 4 for over provisioning
- Divide by 32 to get throughput in GB/s (no overprovisioning)

Some values are rounded down to 0

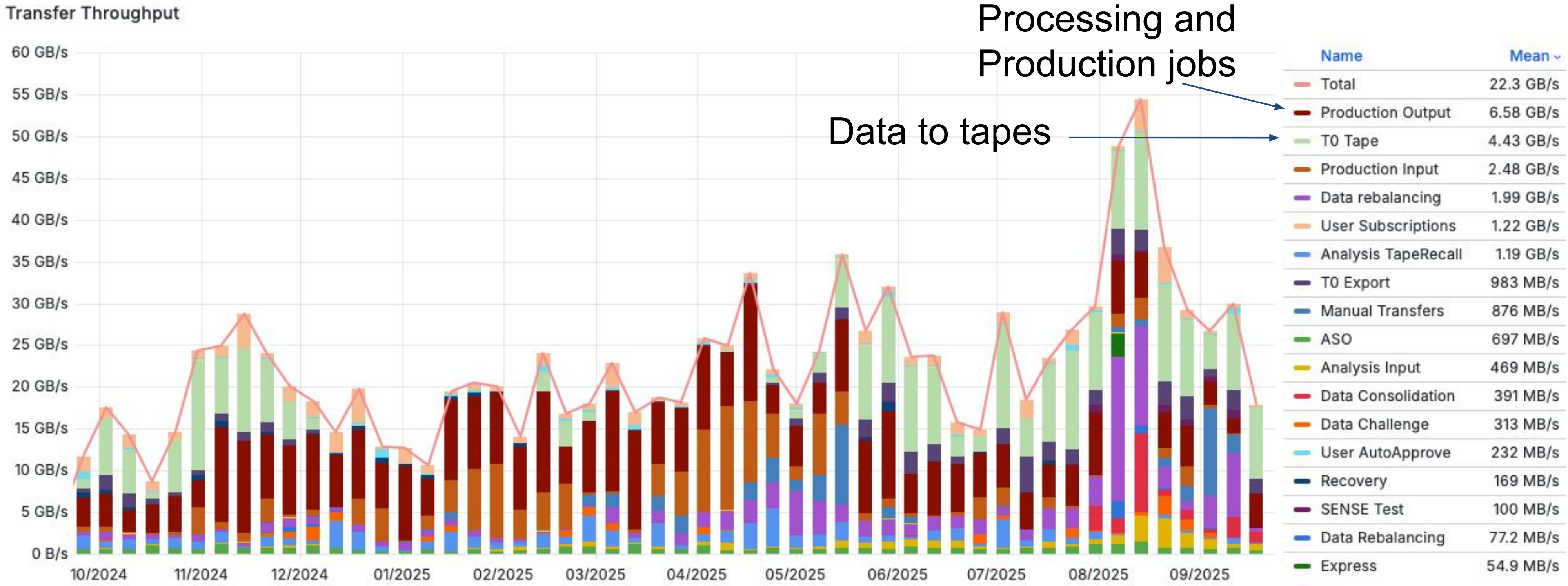
Certain flows are missing
- Can be guessed by extrapolating from Run-3?

Network Throughput in 2024/2025 (Run 3)

- **What transfers did CMS do in the last 12 months? (FTS only)**

Includes Processing and Production jobs

Data to tapes

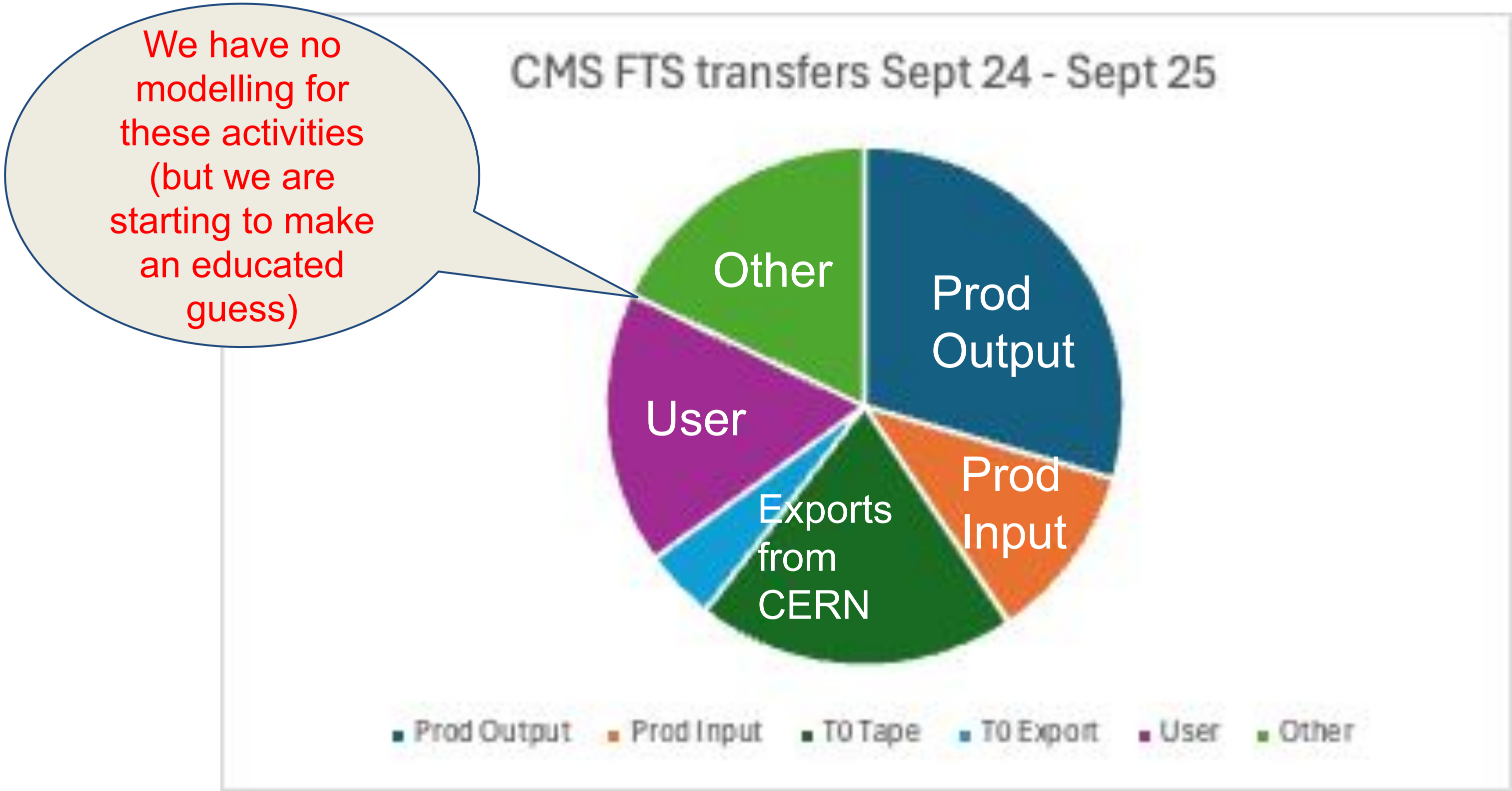


Tight storage situation -> increased transfers

- Recently there was a big ‘rebalancing’ campaign for the MINIAOD
  - During a 90 day period ~15% of FTS transfer bandwidth was this activity
- “We’ll pull that off tape”
- As with our usage of AAA, the more we struggle with storage capacity, the more we need to use network

Network Activities in 24/25 by proportion

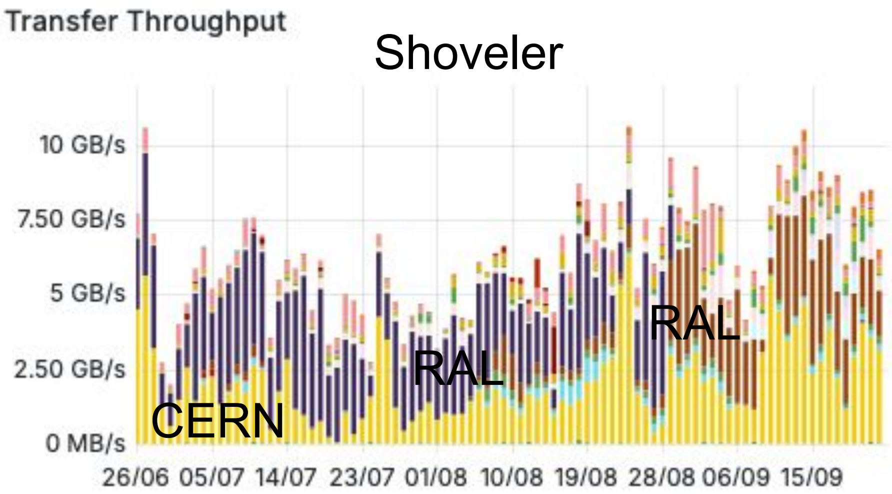
- **What FTS transfers did CMS do in the last year? (By activity)**



Don’t forget AAA

Plots split by **source** site

- **What AAA transfers did CMS do recently?**

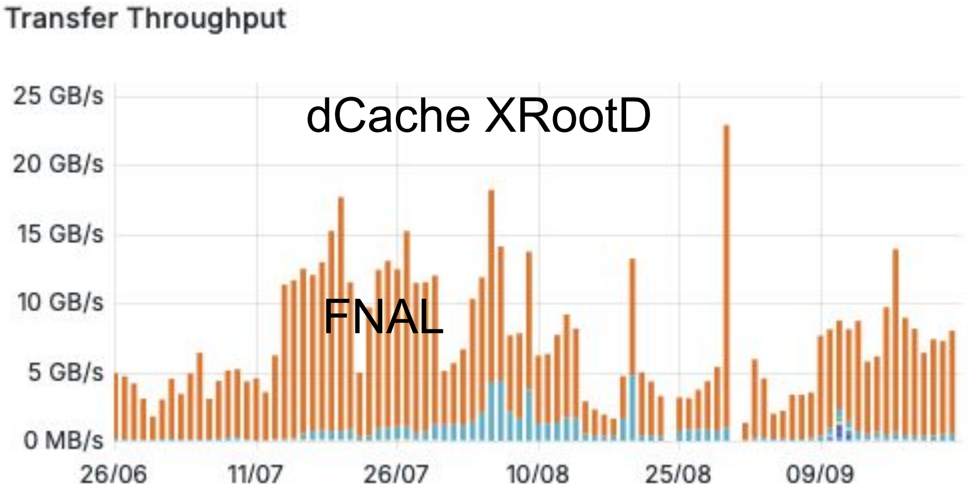


Caveats to this monitoring!

- Not fully verified
  - But seems better
- Not all sites have enabled
  - CERN and FNAL have

Average usage approx.:
6GB/s + 7GB/s = 13GB/s compared with ~22GB/s transferred via FTS

- Lower XRootD average than last presentation (only showing 3 months)
- Less MC prod while processing?



Data Challenge 2027

- The third, major DC, DC27 will take place in the first half of 2027 (less than 18 months from now?)
- Primary goal: Demonstrate **50%** of estimated Run-4 requirement for data transfer
  - Potentially difficult to achieve depending on how sites decide to provision in preparation for Run-4, still 3 years in the future
- Secondary goals:
  - Demonstrate all transfers using tokens, not certificates
  - Demonstrate tape performance (writes and reads)
  - Stress-test new FTS4
  - etc.
- Differing from DC24
  - Tests are more difficult for various reasons
    - Target rate much higher
    - Tape tests are more difficult
    - Big test for tokens
- Work can be done in advance to check tape performance

Setting DC targets

Approach from initial WLCG DC21 document, used again in DC24

- Model or estimate rates for Run-4
- Go big! SUM all rates together
  - This assumes everything happening all at once - not realistic (although was useful to do in DC24)
- Apply a <u>factor of two for spikes</u>

Modelling and estimation

- Some obvious network flows:
  - RAW data export from CERN
  - Staging of RAW/AOD for Re-Reco/Re-MiniAOD
- Many other flows that are harder to model
  - MC production
  - Distribution and archiving of re-processing
  - Anything about end-user driven analysis flows
  - Re-distribution of data by Data Management

Having the (approximate) numbers ready early

- Helps sites to know what we expect from them
  - Network, disk, and tape capabilities
- We, CMS, find the bottlenecks in our data model
- Encourage other VOs with shared infrastructure to also do this work early
  - Including VOs who are still ramping up (e.g. DUNE, SKA, Juno)
  - Sum all VO requirements for DC27 and see which sites will (in theory) be able to cope
  - Need input from sites on their estimated capabilities for DC27
- Decide what to do IF getting to approximately double DC24 seems impossible

Mini-DCs

- We only just, briefly, hit our ‘flexible’ targets in DC24 (25% Run-4)
- Mini-DCs are smaller in scope, intended to maximise performance of existing systems and maintain momentum of continuous improvement
  - MAY involve multiple VOs, but only as necessary
- Split into two categories:
  - Capacity (e.g. what rates can a particular storage endpoint achieve?)
  - Capability (e.g. what new technologies could help us? how will transfers change in future?)
- Several mini-DCs have been run; others are in planning stage
- Results reported at the DOMA-General meetings https://indico.cern.ch/category/10360/

Mini-DC proposal

- Katy would like to plan a tape staging mini-DC for CMS sites…
- Outside of data-taking - January 2026?
- What performance do we get in real life when staging?
  - Set a baseline for staging rate expectations
  - Help us to help sites that have slow performance
- Choose similar datasets that exist on all CMS tapes
  - E.g. same total size, same number of files, regular file size, similar age?
- 100 TB recall? (100TB recalled at 1GB/s would take 28 hours)
- Ask sites to dedicate a fixed number of drives to the test, e.g. 4
- How quickly can sites perform the recall to their Disk sites?
  - E.g. at RAL: T1_UK_RAL_Tape -> T1_UK_RAL_Disk
- Can be done site-by-site

Summary

- Basic modelling for major network flows for CMS is done/in progress
  - Should serve CDR and DC27 planning
  - Rates are derived from computing model parameters
  - …or alternatively extrapolated from Run-3 for less deterministic flows (e.g. user analysis)
- Sharing this with DOMA to encourage similar activities in other experiments
- Derive more detailed planning for DC27 with other VOs
  - What can be achieved?
- Preparing mini-challenges especially around tape