CMS
Compact Muon Solenoid

CMS O&C
Offline Software and Computing

DC24 - CMS plan

Katy Ellis, Lisa Paspalaki, Christoph Wissing

DC24 workshop, 9 Nov 2023

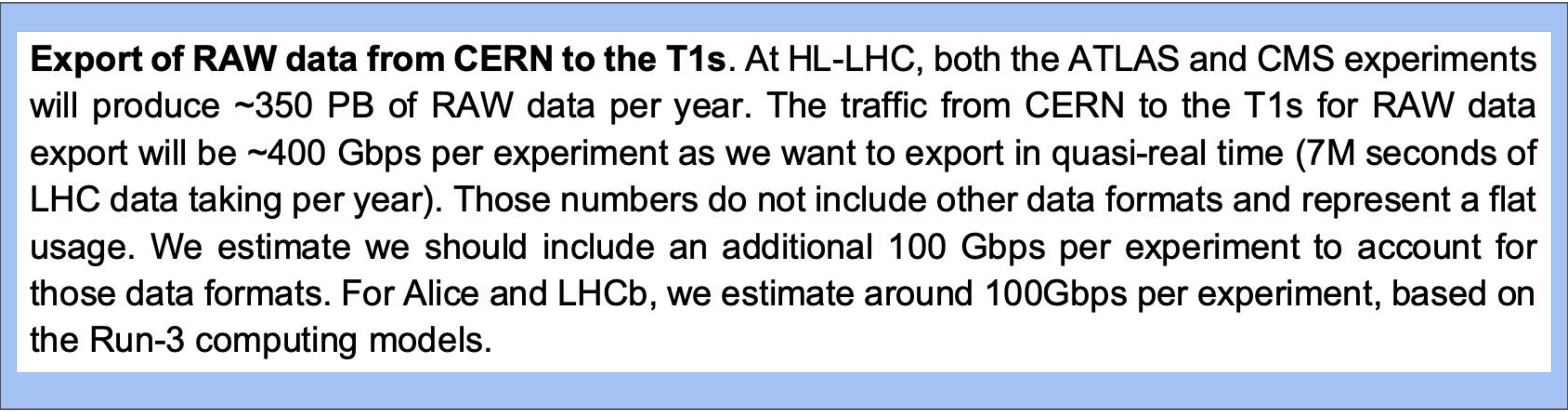
Outline

- CMS transfer scenarios - transfer rate goals
- High-level plan day-by-day
- Organisational details
- Rate requests by site
- Fitting in with ATLAS
- Site tests in progress
- Future tests

Transfer scenarios

What are the expected HL-LHC rates? 1. T0 export

WLCG gave some targets in the Data Challenge document:



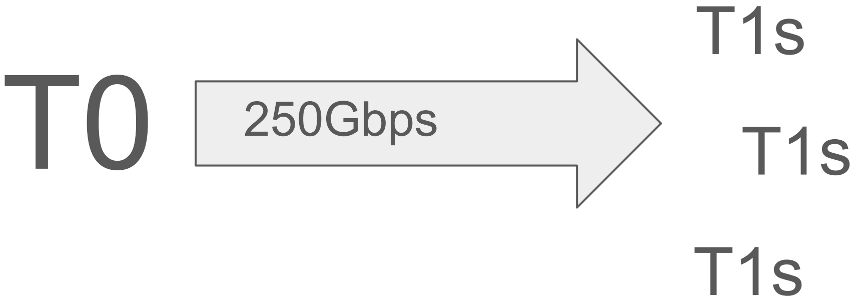
Export of RAW data from CERN to the T1s. At HL-LHC, both the ATLAS and CMS experiments will produce ~350 PB of RAW data per year. The traffic from CERN to the T1s for RAW data export will be ~400 Gbps per experiment as we want to export in quasi-real time (7M seconds of LHC data taking per year). Those numbers do not include other data formats and represent a flat usage. We estimate we should include an additional 100 Gbps per experiment to account for those data formats. For Alice and LHCb, we estimate around 100Gbps per experiment, based on the Run-3 computing models.

So the Tier 0 export rate (CERN to T1 tape and disk) estimate is :
500Gbps (ATLAS) + **500Gbps (CMS)** + 100Gbps (LHCb) + 100Gbps (ALICE) = 1200Gbps

Network usage and DC24

- Reminder: Network usage is ‘bursty’ so for the challenge we over-provision by a factor of 2
- Reminder: DC24 is a challenge at the 25% of HL-LHC level
- So the T0 export goal in DC24 for CMS is actually **250Gbps**

1. “T0 export”



T0

250Gbps

T1s

T1s

T1s

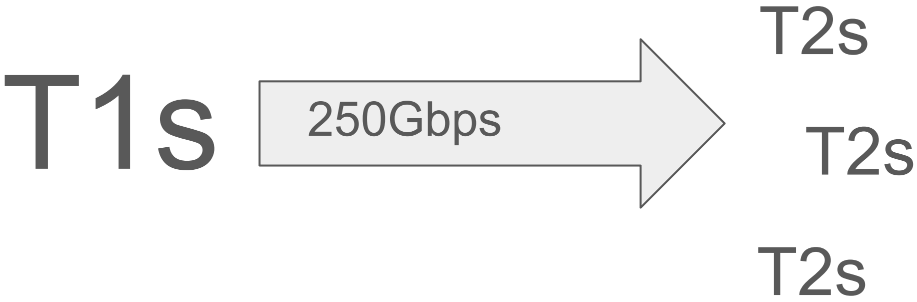
Is this a sufficient challenge?

- CMS agreed with ATLAS to extend the challenge from just the experiment data export from T0 to T1s…
- This significantly increases the requested data rates but is perhaps a better reflection of reality
- Brings in Tier 2 sites, reads from Tier 1s, etc.
- Accounts for other transfer workflows

CMS transfer scenario 2: “Reprocessing”

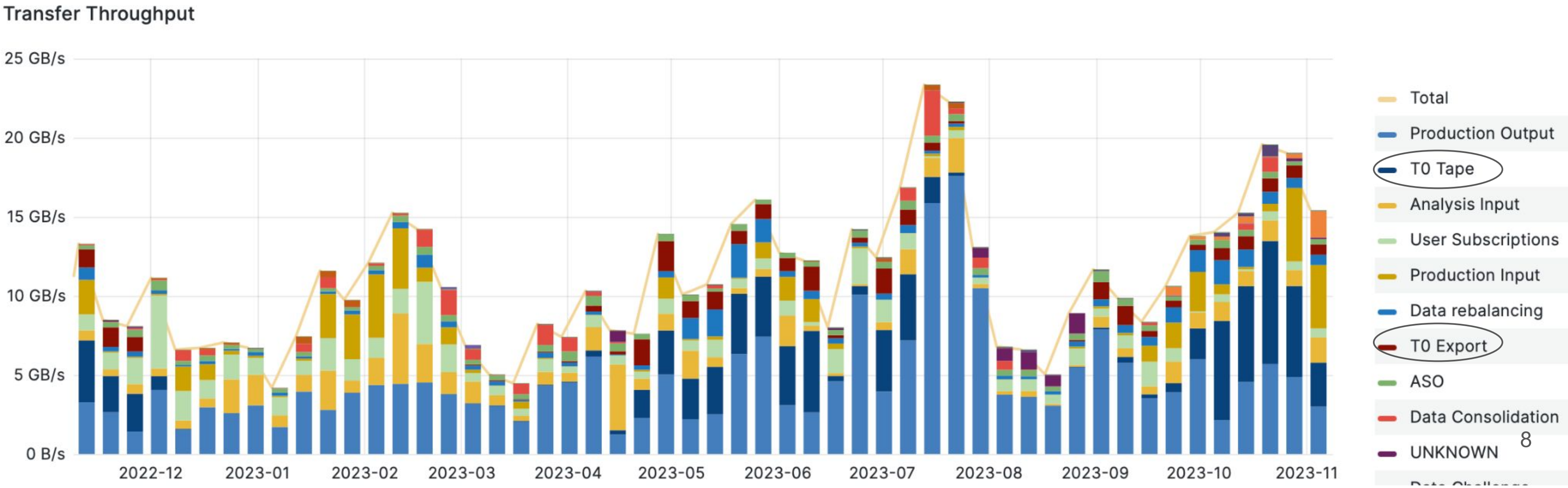
- After data-taking it is typical to reprocess data at the Tier 2s
- Much of the data stored at T1s is then moved to T2s
- Same data as T0 export, and transfer should be done over similar period
  - So the rate goal should be the same as T0 export

2. “T1 export”



What about Monte Carlo?

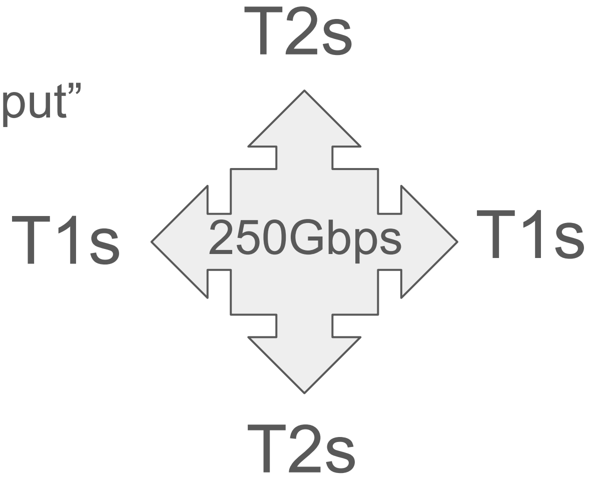
- Not mentioned in the DOMA document
- Decided by CMS and ATLAS to include estimated rates in the challenge anyway, to make it more realistic…



CMS transfer scenario 3: “Production output”

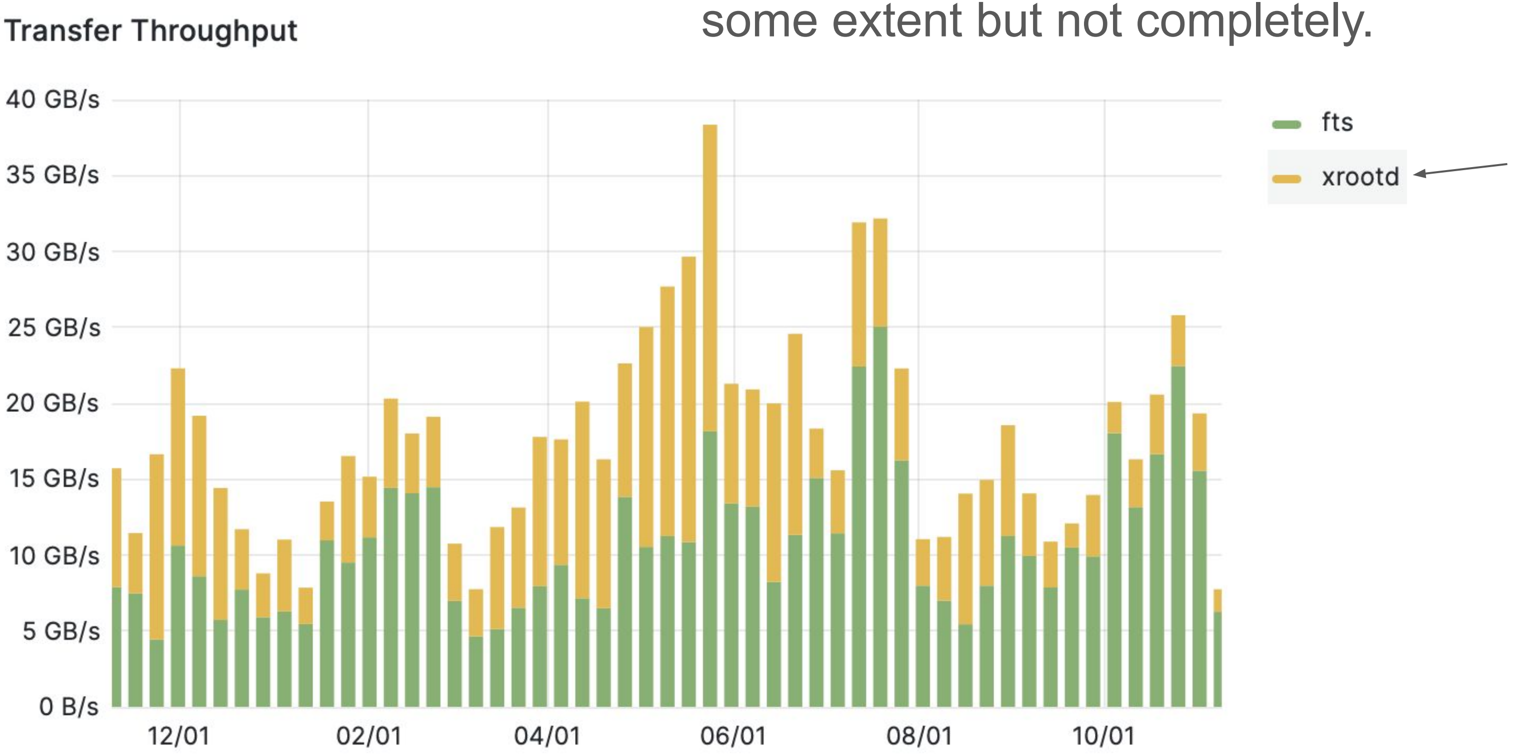
- Throughout the year we produce MC at T1s and T2s
- Data is moved between T1s and T2s
- As an estimate, I add an additional 250Gbps to the overall traffic to test links between T1s, between T2s, and from T2s and their local T1

3. “Production output”



What about AAA?

CMS jobs can access data remotely using the 'AAA' system. Reads are streamed through XRootD. This is monitored to some extent but not completely.

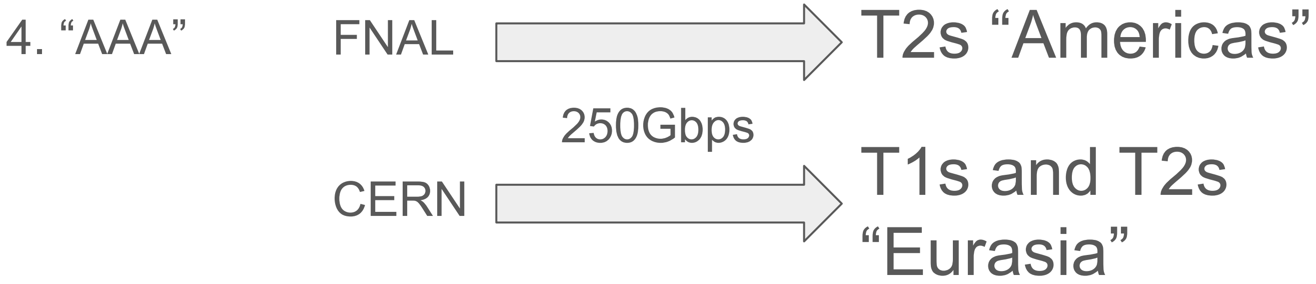


AAA

CMS transfers according to WLCG monit during the last year

CMS transfer scenario 4: “AAA”

- A major, planned user of AAA is MC jobs overlaying pile-up from ‘premix libraries’ generated separately from other jobs
- Premix is stored at CERN and FNAL and read by jobs at T1s and T2s
- As an estimate, I add an additional 250Gbps to the overall traffic



4. “AAA”

FNAL → T2s “Americas”

250Gbps

CERN → T1s and T2s “Eurasia”

So, what is the total required rate for CMS in DC24?

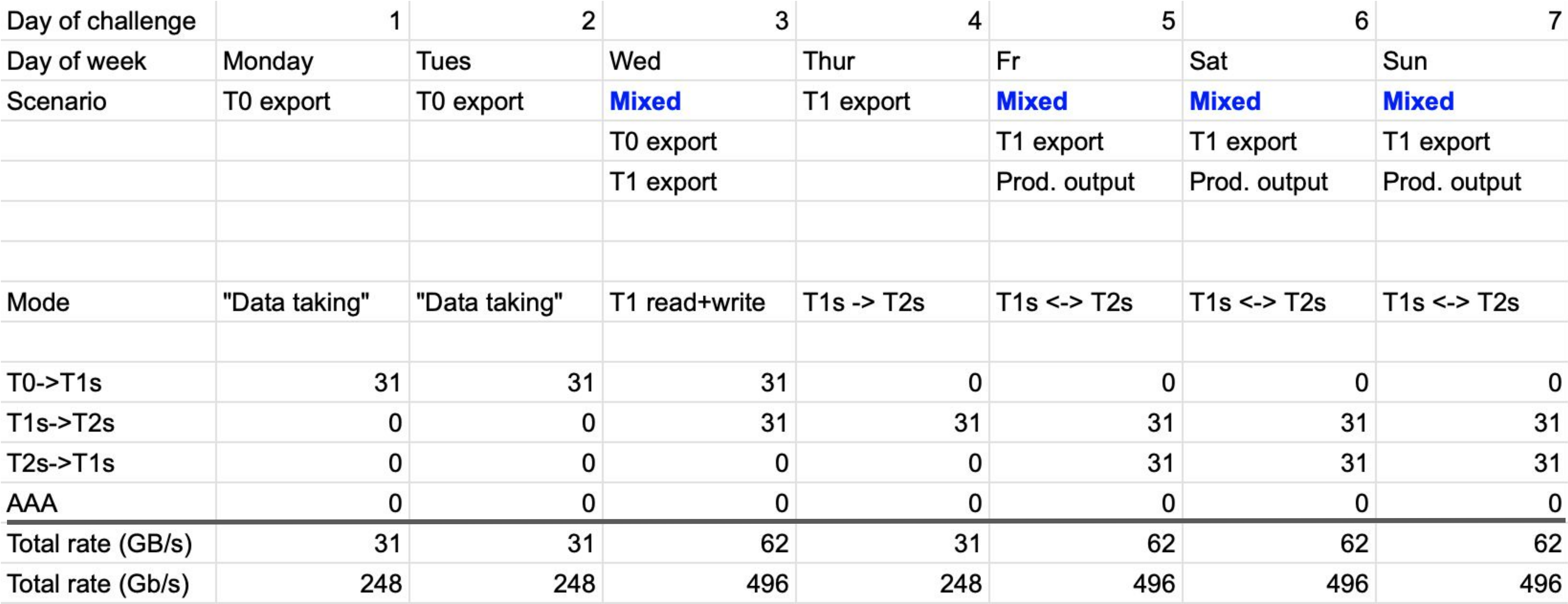
- Does it make sense to sum the rates and try to achieve total CMS traffic of 750-1000Gbps?
  - Maybe not…the rates we have been estimating apply to different infrastructure and different time periods.
  - However, this is a challenge! So we aim high.
- In summary, we do both
  - Start with individual tests, then work up to ‘everything, all at once’

Day-by-day plan

Converting to GB/s

250Gbps ~= 31GB/s

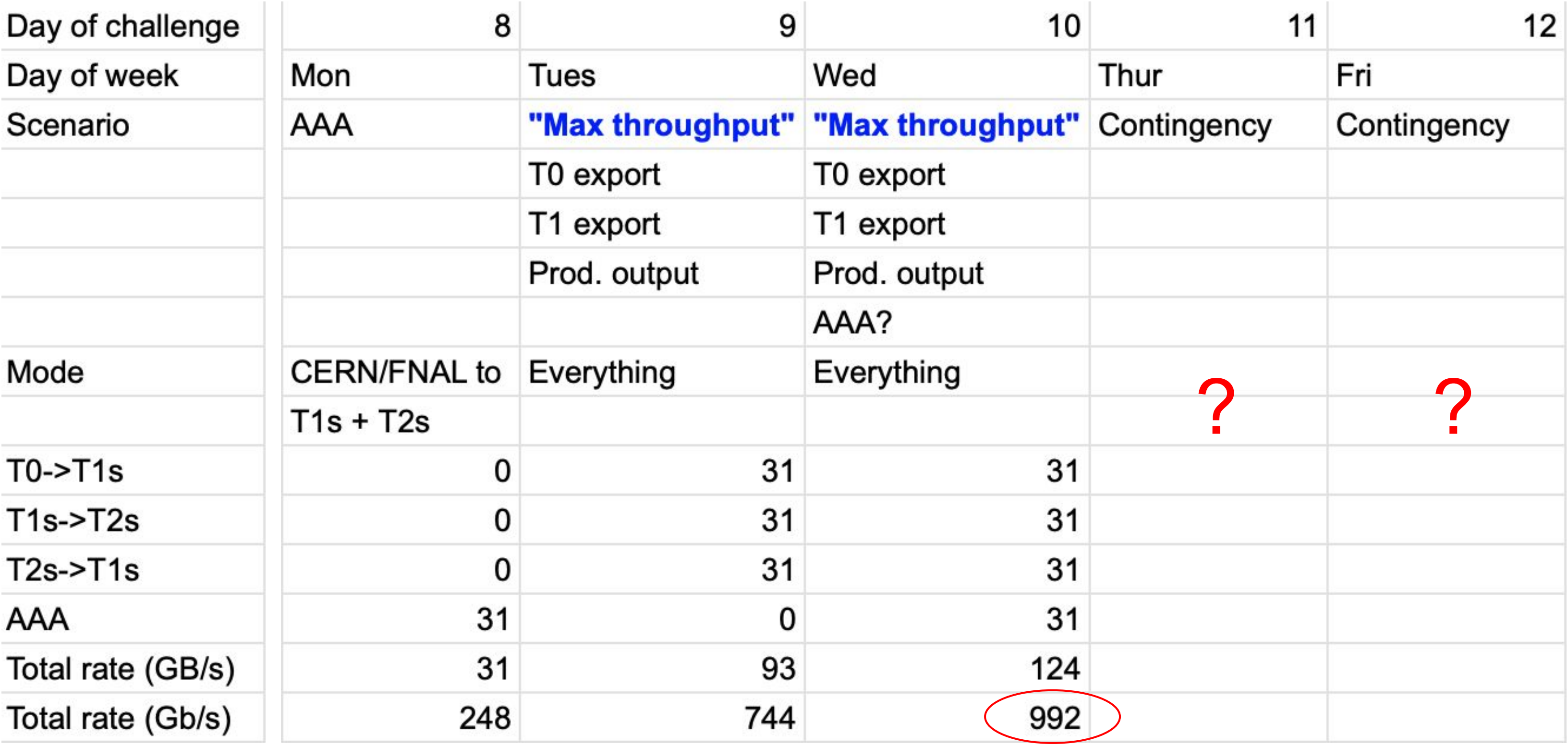
Running the challenge week 1 proposal



Day of challenge	1	2	3	4	5	6	7
Day of week	Monday	Tues	Wed	Thur	Fr	Sat	Sun
Scenario	T0 export	T0 export	**Mixed**	T1 export	**Mixed**	**Mixed**	**Mixed**
			T0 export		T1 export	T1 export	T1 export
			T1 export		Prod. output	Prod. output	Prod. output
Mode	"Data taking"	"Data taking"	T1 read+write	T1s -> T2s	T1s <-> T2s	T1s <-> T2s	T1s <-> T2s
T0->T1s	31	31	31	0	0	0	0
T1s->T2s	0	0	31	31	31	31	31
T2s->T1s	0	0	0	0	31	31	31
AAA	0	0	0	0	0	0	0
Total rate (GB/s)	31	31	62	31	62	62	62
Total rate (Gb/s)	248	248	496	248	496	496	496

N.B. zero rates imply zero additional traffic injected; production traffic continues

Running the challenge week 2 proposal



Day of challenge	8	9	10	11	12
Day of week	Mon	Tues	Wed	Thur	Fri
Scenario	AAA	**"Max throughput"**	**"Max throughput"**	Contingency	Contingency
		T0 export	T0 export		
		T1 export	T1 export		
		Prod. output	Prod. output		
			AAA?		
Mode	CERN/FNAL to	Everything	Everything	?	?
	T1s + T2s				
T0->T1s	0	31	31		
T1s->T2s	0	31	31		
T2s->T1s	0	31	31		
AAA	31	0	31		
Total rate (GB/s)	31	93	124		
Total rate (Gb/s)	248	744	992		

Organisational details

Running the challenge

- Propose with ATLAS a daily ‘run meeting’ where we confirm the plan for the day, communicate a summary to sites
- Remote personnel may choose to attend CERN in person during the second week of the challenge
- CMS will use the ‘injector’ tool to generate DC load with Rucio
    - Mario will describe in more detail tomorrow
    - Supply the tool with info about datasets and locations
    - Create a list of links to be tested, and rate required
    - CMS will come up with a ‘daily menu’ for each day, in advance of the challenge, of links and goal rates for that day, which should be consistent with our overall plans

Pre-challenge

- Share with (CMS) sites the target rates and ask if any sites wish to opt-out
- Become familiar with the injector tool and DC bespoke monitoring
    - Panos and Hasan from CMS DM, plus Diego from USCMS - all planning improvements
    - Will we be able to determine from monitoring which sites are using tokens?
    - Synchronise with ATLAS tests on common sites, including CERN
- Pre-challenge tests

Number of links

- Should we test every link between T0, T1s and T2s?
  - **No!** There are 55 sites, so 1540 unique links
  - We don’t have effort to analyse transfers on every link
- Focus attention on a small number of links per site

Rate requests by site

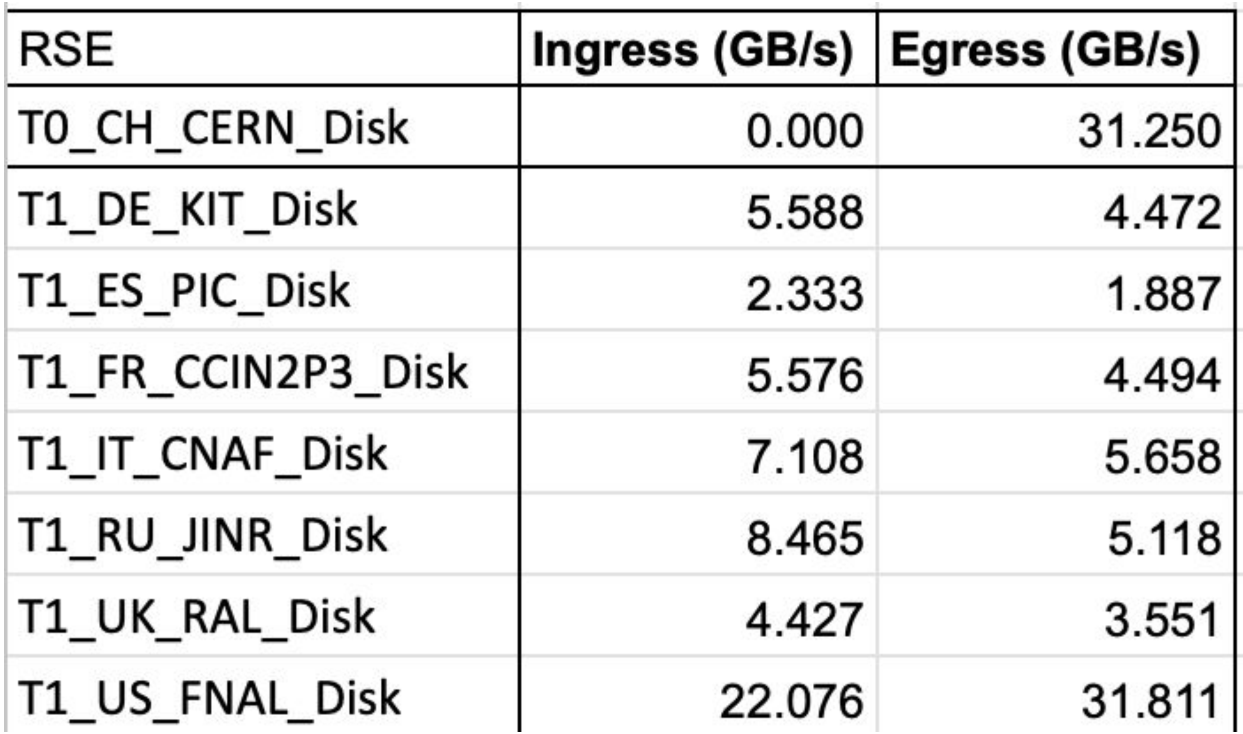
Splitting the data rates by site

1. T0 export: Split by tape pledge
2. T1 export: Split by tape pledge at T1s; split by disk pledge at T2s
3. Production output: Split by disk pledge at all sites
4. AAA: Split by ‘useable cores’ pledged at all sites

Other rates and questions

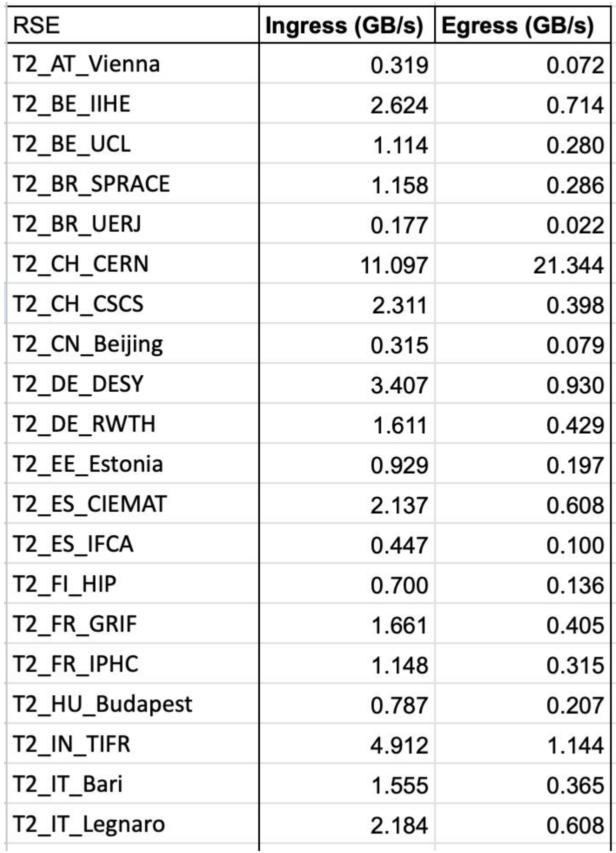
- Test all T1 and T2 sites? (my preference, but allow opt-out)
- Do not test T3s
- Is it fair to split all network rates according to the size of the storage?
- What about distant sites?

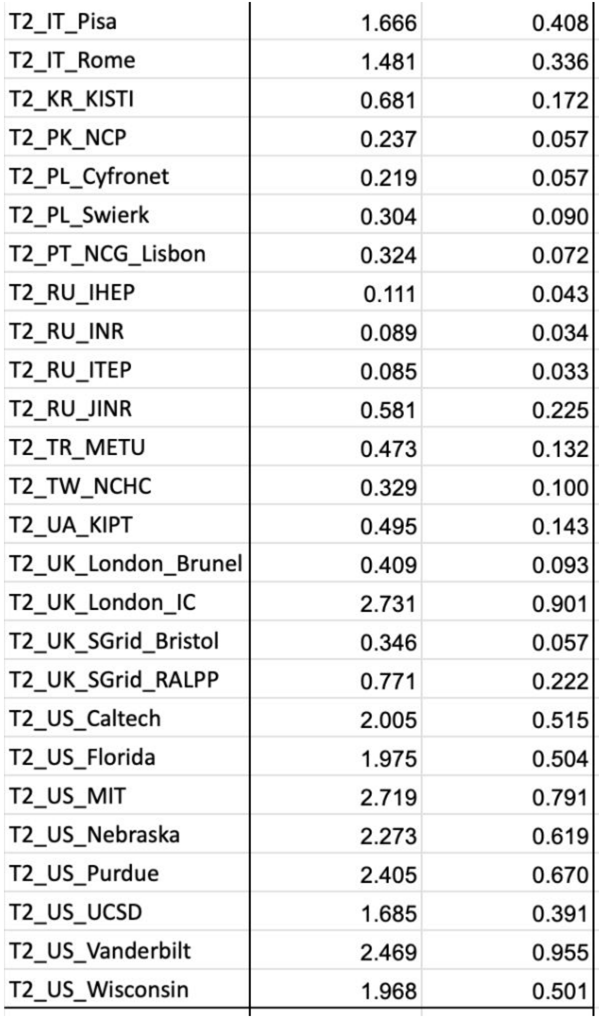
Tier 1 total rate goals (all 4 scenarios summed)



RSE	Ingress (GB/s)	Egress (GB/s)
T0_CH_CERN_Disk	0.000	31.250
T1_DE_KIT_Disk	5.588	4.472
T1_ES_PIC_Disk	2.333	1.887
T1_FR_CCIN2P3_Disk	5.576	4.494
T1_IT_CNAF_Disk	7.108	5.658
T1_RU_JINR_Disk	8.465	5.118
T1_UK_RAL_Disk	4.427	3.551
T1_US_FNAL_Disk	22.076	31.811

Tier 2 total rate goals





RSE	Ingress (GB/s)	Egress (GB/s)
T2_AT_Vienna	0.319	0.072
T2_BE_IIHE	2.624	0.714
T2_BE_UCL	1.114	0.280
T2_BR_SPRACE	1.158	0.286
T2_BR_UERJ	0.177	0.022
T2_CH_CERN	11.097	21.344
T2_CH_CSCS	2.311	0.398
T2_CN_Beijing	0.315	0.079
T2_DE_DESY	3.407	0.930
T2_DE_RWTH	1.611	0.429
T2_EE_Estonia	0.929	0.197
T2_ES_CIEMAT	2.137	0.608
T2_ES_IFCA	0.447	0.100
T2_FI_HIP	0.700	0.136
T2_FR_GRIF	1.661	0.405
T2_FR_IPHC	1.148	0.315
T2_HU_Budapest	0.787	0.207
T2_IN_TIFR	4.912	1.144
T2_IT_Bari	1.555	0.365
T2_IT_Legnaro	2.184	0.608
T2_IT_Pisa	1.666	0.408
T2_IT_Rome	1.481	0.336
T2_KR_KISTI	0.681	0.172
T2_PK_NCP	0.237	0.057
T2_PL_Cyfronet	0.219	0.057
T2_PL_Swierk	0.304	0.090
T2_PT_NCG_Lisbon	0.324	0.072
T2_RU_IHEP	0.111	0.043
T2_RU_INR	0.089	0.034
T2_RU_ITEP	0.085	0.033
T2_RU_JINR	0.581	0.225
T2_TR_METU	0.473	0.132
T2_TW_NCHC	0.329	0.100
T2_UA_KIPT	0.495	0.143
T2_UK_London_Brunel	0.409	0.093
T2_UK_London_IC	2.731	0.901
T2_UK_SGrid_Bristol	0.346	0.057
T2_UK_SGrid_RALPP	0.771	0.222
T2_US_Caltech	2.005	0.515
T2_US_Florida	1.975	0.504
T2_US_MIT	2.719	0.791
T2_US_Nebraska	2.273	0.619
T2_US_Purdue	2.405	0.670
T2_US_UCSD	1.685	0.391
T2_US_Vanderbilt	2.469	0.955
T2_US_Wisconsin	1.968	0.501

Comparing with ATLAS

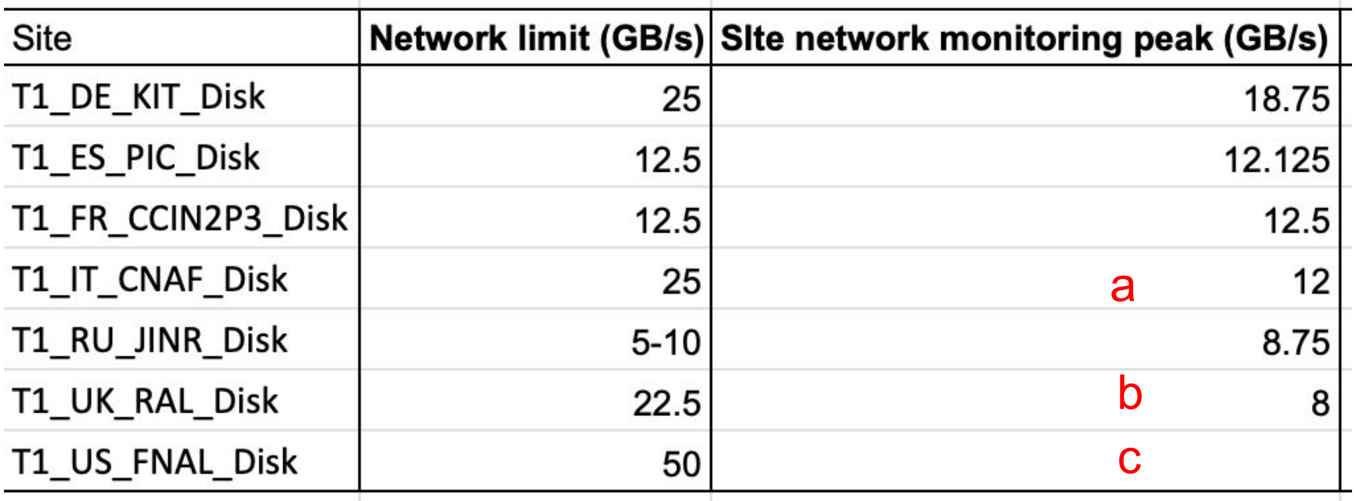
A few observations

1. According to the challenge, CMS need to move the same amount of data from T0 to T1s as ATLAS…but we have a smaller number of T1s. A bigger challenge for CMS in some ways.
2. ATLAS' current production traffic is much more significant than that of CMS, doubly so if you only count FTS transfers (not AAA).
3. CMS shares several Tier 1s with ATLAS, LHCb and others. We need to sum our combined DC24 goals and check if they are within the physical limitations of sites' network bandwidth and storage write/read rates.

Site tests in progress

Tier 1 tests

- Performed a series of tests to each Tier 1, with dataset sourced at CERN
- Asked for info from T1 sites, worked with several sites to make improvements
- Compared transfer rate with expected rate on LHCOPN



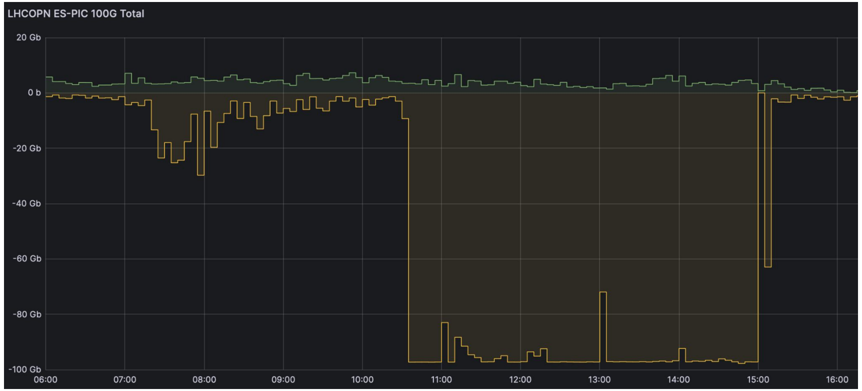
Site	Network limit (GB/s)	SIte network monitoring peak (GB/s)
T1_DE_KIT_Disk	25	18.75
T1_ES_PIC_Disk	12.5	12.125
T1_FR_CCIN2P3_Disk	12.5	12.5
T1_IT_CNAF_Disk	25	a 12
T1_RU_JINR_Disk	5-10	8.75
T1_UK_RAL_Disk	22.5	b 8
T1_US_FNAL_Disk	50	c

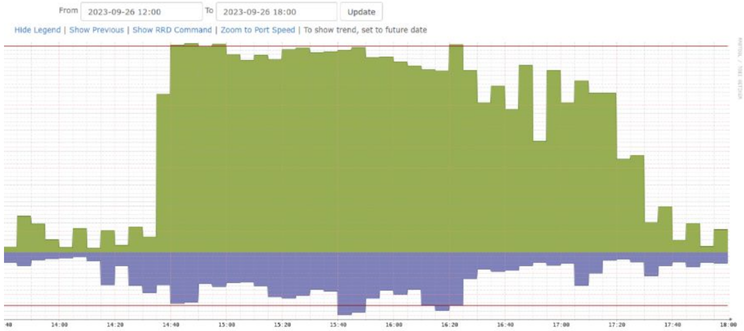
a. Storm devs looking at why load is not balanced

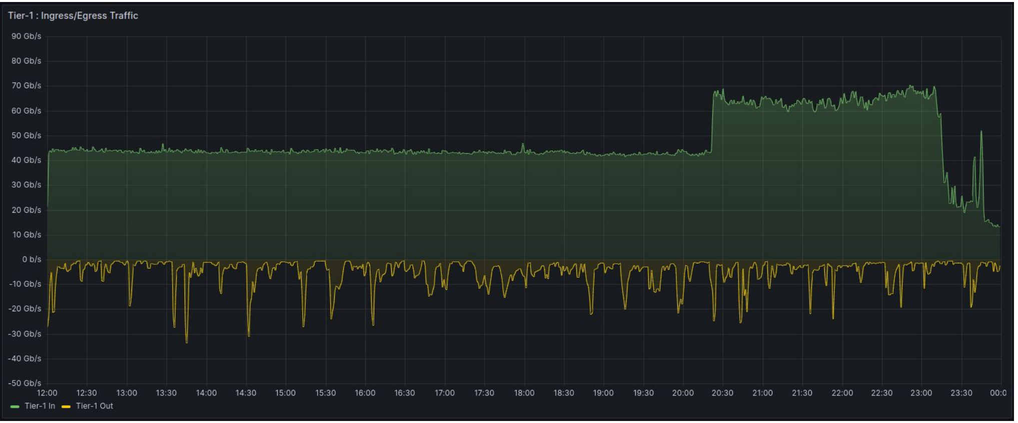
b. Investigating single file transfer speed

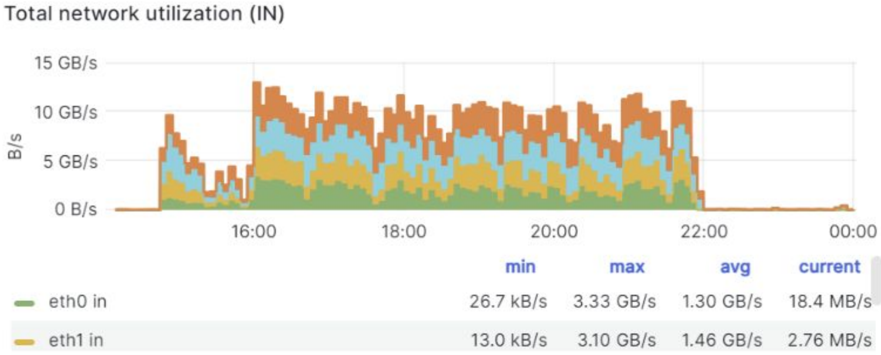
c. Ran out of time for repeated tests

Tier 1 test example network plots




From 2023-09-26 12:00 To 2023-09-26 18:00





	min	max	avg	current
🟩 eth0 in	26.7 kB/s	3.33 GB/s	1.30 GB/s	18.4 MB/s
🟨 eth1 in	13.0 kB/s	3.10 GB/s	1.46 GB/s	2.76 MB/s

FTS configuration for parallel transfers

- In FTS you can specify the maximum concurrent transfers into a site, out of a site, and on any link (site A -> site B)
- Some sites can write files fast
    - They can fill their pipes with ~200 concurrent transfers
- Other sites don’t write files so fast (or have really big pipes)
    - But some of them can still fill their pipes if you increase the number of concurrent transfers to ~500+

Future test ideas

- Egress from CERN (what FTS configuration is needed?)
- Read tests at T1s?
- T2 site tests?
- Joint tests with ATLAS?
- Collaborative tests with new network technology groups?
- Using the injector tool to do a CMS mini-challenge?
    - Is it easy to find sufficient datasets?
    - How much do we need to scale up the Rucio infrastructure
    - Can data be deleted and re-transferred at the required rate?
    - Do all our transfers appear in the monitoring?

Summary

- Data Challenge 2024 should try to represent 25% of HL-LHC network traffic
- Example rates per site have been calculated
  - To be confirmed
- CMS will inject additional transfers on top of usual production activities
- CMS will use the DC inject tool, Rucio and FTS, and monitor using monit
- Pre-tests are in progress and there is scope for more
- Benefits will come if the challenge encourages VOs, network experts and sites to investigate anomalies and plan for the future