The 2021 WLCG Data Challenges

Riccardo DI MARIA

on behalf of the WLCG DOMA Data Challenges Working Group

March 24th, 2022 - Data Management & Big Data, Session I, ISGC 2022

2021 WLCG Data Challenges

- First of a series of Data Challenges (DCs)
  - assess readiness of infrastructure for HL-LHC
  - pivotal for Run3 preparation

- Characterised by 2 parts
  - Network Data Challenge
  - Tape Challenge

Network and Tape DCs Common Monitoring

ESCAPE/WLCG existing expertise/resources leveraged and improved

- FTS-based data source (ElasticSearch) is mostly used for the **common** dashboard
  - currently including traffic from **ATLAS**, **CMS** & **LHCb**
  - special **activity** label for Data Challenge to filter traffic
  - matrices & plots in place for e.g. files/transfers and throughput; data volume; transfer error and failures (live links to latest FTS logs)
- WLCG data source including ALICE and CMS XRootD traffic also used
  - WLCG Global Throughput
- Tape-related metrics integrated and filtered accordingly, e.g. read traffic

Documentation

Network Challenge

Objectives

- Goal to demonstrate HL-LHC scale: 10% bandwidth, increasing in subsequent years
  - commission **HTTP Third Party Copy**
- Network target metrics (e.g per site and experiment) available at HL-LHC document

Centralised Submission Infrastructure

- Coordination across experiments: ATLAS and CMS
  - standardised procedures and injection → **flexibility** to modify strategy on-the-fly
  - experiment-agnostic framework for future challenges
- ATLAS and CMS injected data in parallel to production activities
- Alice and LHCb filled the network with normal production activities

Network Challenge Targets

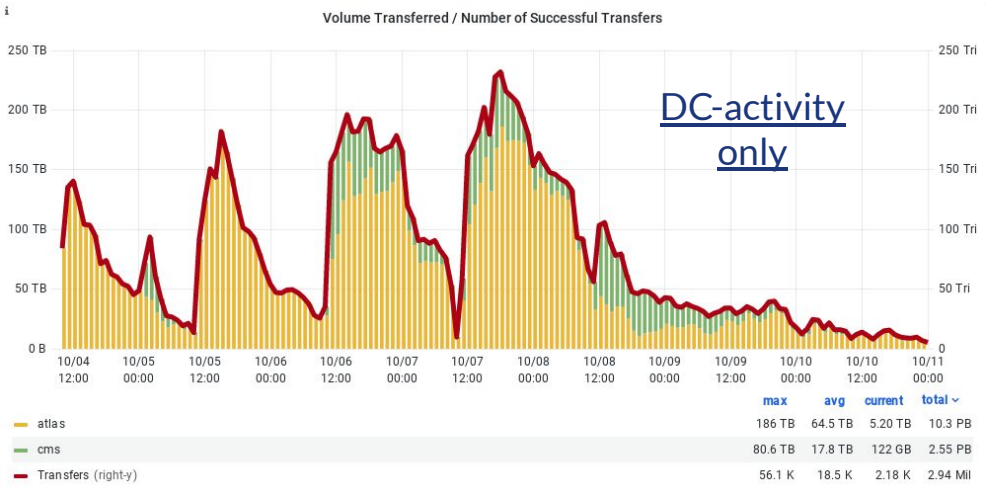
[HL-LHC document]

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-T1-BNL	450	900	230	140	70	20
US-FNAL-CMS	800	1600	400	240	120	40
(atlantic link)	1250	2500	630	380	190	60
Sum	4810	9620	2430	1450	730	240

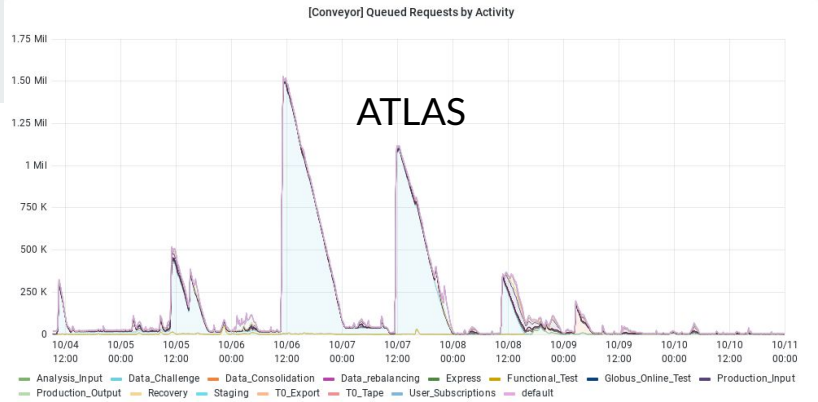
T1	%ATLAS	%CMS	% Alice	% LHCb	ATLAS+CMS Network Needs (Gbps) Minimal Scenario in 2027	Alice Network Needs (Gbps) Minimal Scenario in 2027	LHCb Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Minimal Scenario in 2027	LHC Network Needs (Gbps) Flexible Scenario in 2027
CA-TRIUMF	10	0	0	0	200	0	0	200	400
DE-KIT	12	10	21	17	450	80	70	600	1200
ES-PIC	4	5	0	4	180	0	20	200	400
FR-CCIN2P3	13	10	14	15	450	60	60	570	1140
IT-INFN-CNAF	9	15	26	24	480	110	100	690	1380
KR-KISTI-GSDC	0	0	12	0	0	50	0	50	100
NDGF	6	0	8	0	110	30	0	140	280
NL-T1	7	0	3	8	140	10	30	180	360
NRC-KI-T1	3	0	13	5	50	50	20	120	240
UK-T1-RAL	15	10	3	27	490	10	110	610	1220
RU-JINR-T1	0	10	0	0	200	0	0	200	400
US-T1-BNL	23	0	0	0	450	0	0	450	900
US-FNAL-CMS	0	40	0	0	800	0	0	800	1600
(atlantic link)					1250	0	0	1250	2500
Sum	100	100	100	100	4000	400	410	4810	9620

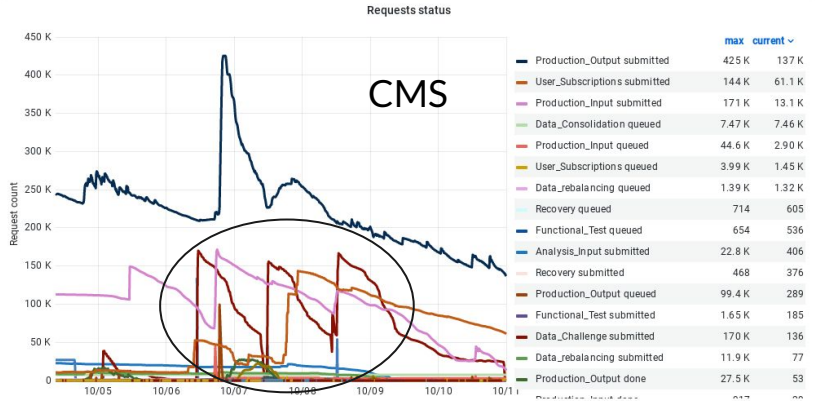
- Numbers referring to ingress and egress (disjoint)
- Total ingress+egress
- Minimal: hierarchical model T0-T1-T2 traffic
- Flexible: chaotic model currently more realistic

Data Ingestion



DC-activity only

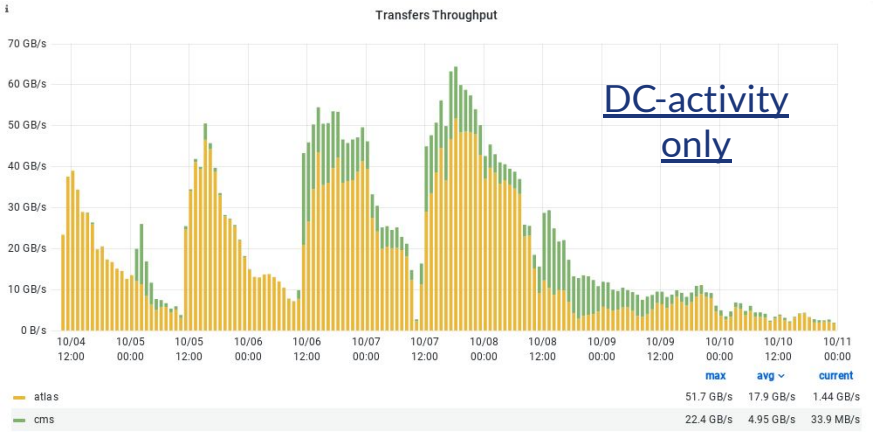




Usage of real data + end2end tests = unique data to select

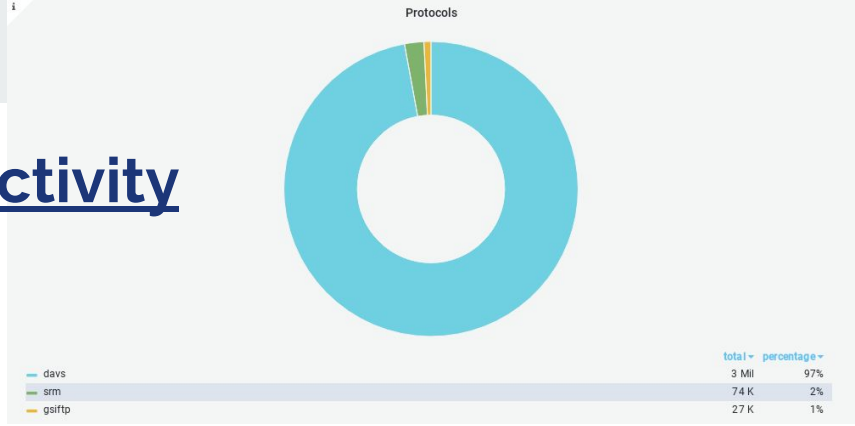
- Multiple Data-Bulk selections (not only a-priori) and no API used
- Spiky injections by design (vs. more sustained in Tape DC) and due to on-the-fly **flexibility**

DC Week - "Data Challenge" Activity



DC-activity only

- Transfers:
  - Attempted: 3.49 M - **Successful: 84.09%**
- Volume:
  - Attempted: 15.14 PB - **Successful: 12.81 PB**
- Average Throughput: 22.4 GB/s



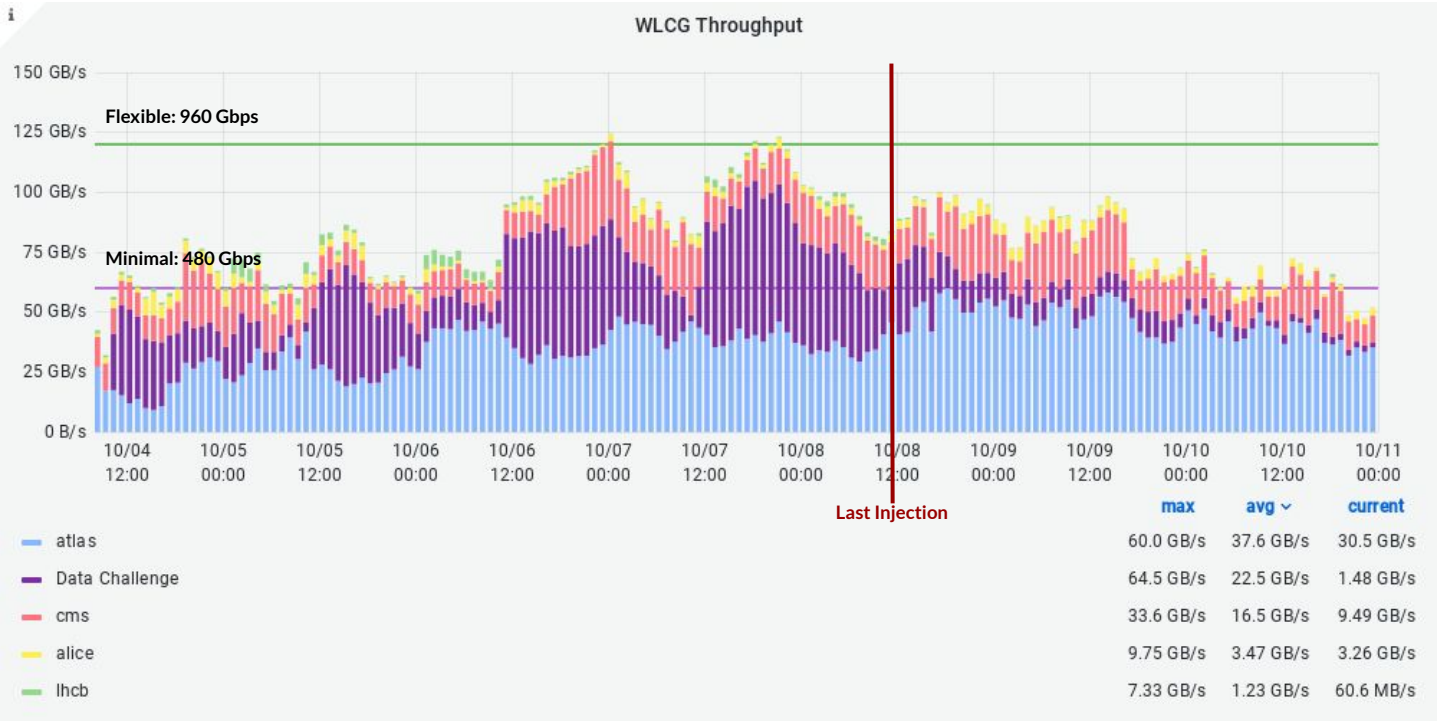
ATLAS

- Transfers:
  - Attempted: 2.88 M - **Successful: 86.88%**
- Volume:
  - Attempted: 11.54 PB - **Successful: 10.26 PB**
- Average Throughput: 17.9 GB/s

CMS

- Transfers:
  - Attempted: 616 k - **Successful: 71.06%**
- Volume:
  - Attempted: 3.60 PB - **Successful: 2.55 PB**
- Average Throughput: 4.95 GB/s

WLCG Global (FTS+XRootD) Throughput - DC Week Plot



October 4th 0900 - 11th 0000

- Reported values depending on the time-window
- Data injection methodology and tail taken into account when assessing sites
- T0-T1 FTS traffic vs. LHCOPN T1-T0-T1 to be considered during the assessment

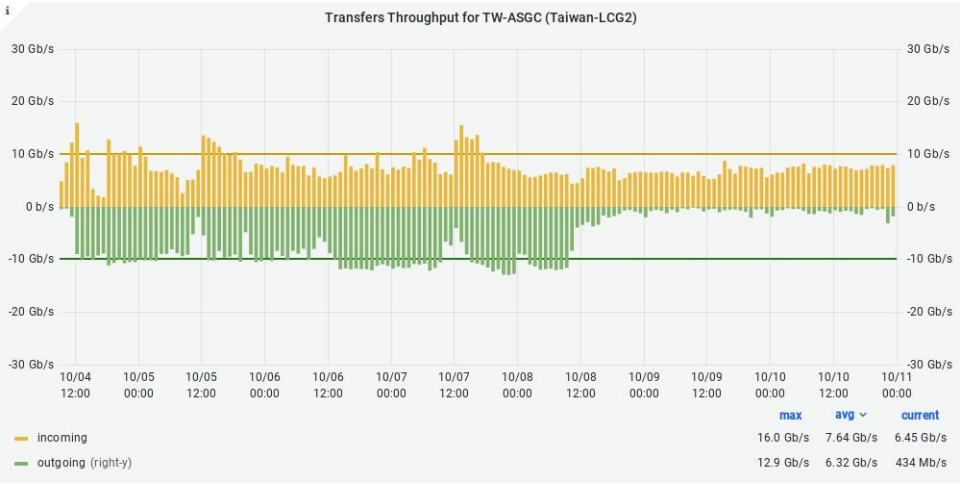
T1s Ingress/Egress

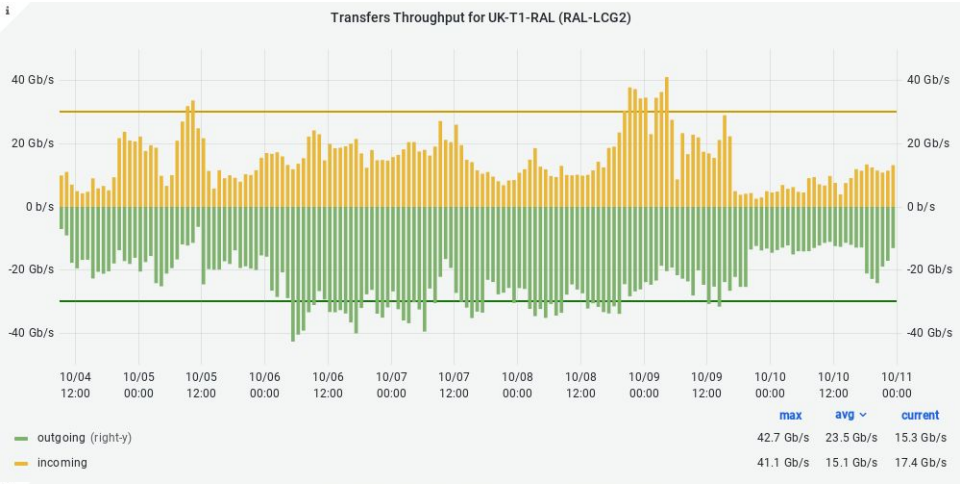
T1	Flexible Scenario 2027	Minimal Scenario 2027	10% Minimal Scenario (2021 Targets) ingress/egress	Average ingress/egress (hourly)	Maximum ingress/egress (hourly)
CA-TRIUMF	400	200	10/10	17/26	49/71
DE-KIT	1200	600	30/30	26/42	77/143
ES-PIC	400	200	10/10	9/11	18/17
FR-CCIN2P3	1140	570	30/30	34/41	70/80
IT-INFN-CNAF	1380	690	30/30	23/40	50/87
KR-KISTI-GSDC	100	50	-	-	-
NDGF	280	140	10/10	26/26	49/81
NL-T1 (NIKHEF)	-	-	10/10	10/12	38/53
NL-T1 (SARA)	360	180	10/10	12/16	51/79
RU-JINR-T1	400	200	10/10	10/12	26/31
RU-NRC-KI-T1	240	120	10/10	9/12	18/34
TW-ASGC	-	-	10/10	8/6	16/13
UK-T1-RAL	1220	610	30/30	15/24	41/43
US-FNAL-CMS	1600	800	40/40	19/16	49/64
US-T1-BNL	900	450	20/20	29/38	75/117

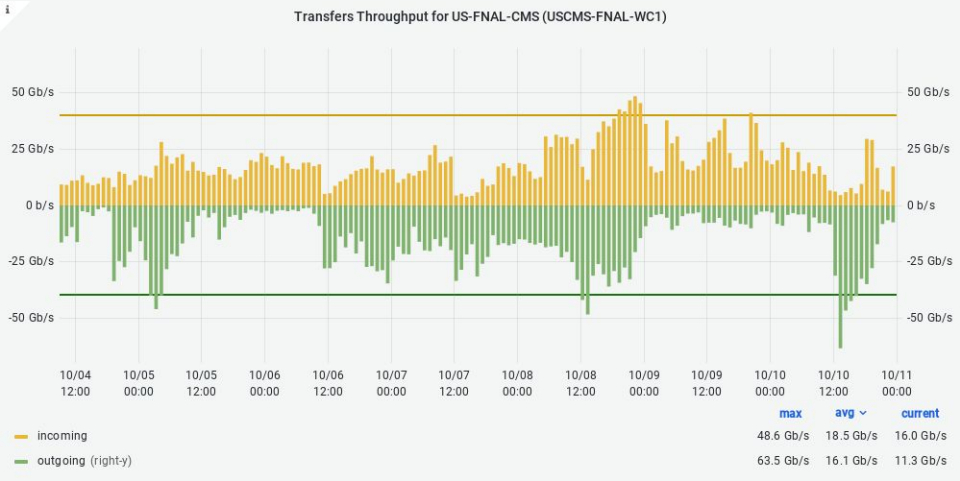
Global sum of average ingress/egress across T1s exceeds the targets (240/240): **242/309**

- **black** sites are OK
- **green** sites are OK after ingress+egress and/or dedicated time window (4d)
- **red** sites necessitate a closer inspection
  - TW-ASGC: small, no tape storage, farthest → considered OK
  - UK-T1-RAL: unique storage system (ECHO) reviewed for different-size buffer layers and network upgrade
  - US-FNAL-CMS: not enough production traffic to fill the bandwidth, lack of disk space and difficulty to free it
- → further tests recommended

FTS Global Traffic during DC Week







Highlighted Event: **NOTED SDN Algorithm**

- Enabled on PIC-CERN
- LHCOPN saturated → LHCONE bandwidth added
  - from 6 Gbps (LHCOPN monitoring) to 10 Gbps (FTS monitoring)
- Including site network monitoring data recommended

Network Challenge Lessons Learned

- Network not the bottleneck
  - additional tests recommended for few sites
  - mock challenges should play a leading role
  - test infrastructures should be avoided
- Central submission infrastructure could reduce need for experiment manpower
  - necessity for a change of culture for central coordination of multi-experiment challenges
- Common monitoring/DC dashboards useful for the wealth of information
  - priority on FTS and xrootd data structures and collections for traffic harmonisation
  - monitoring = N+1 dashboards and data sources → proven common solution feasibility
- Explore different scenarios towards HL-LHC target wrt usage of real data for DCs
  - prioritise T0 export rather than T2 import while planning future exercises
    → effect on global volume transferred

Tape Challenge

Objectives

- Goal to validate the maximum tape bandwidth needed for reads and writes for Run3
- October 11th-15th, all LHC experiments and Tier1 tape-sites participated → first of its kind
- Experiments expectations on T1 tape throughputs for Run3
  - values refer to delivered throughputs, not nominal tape bandwidth on the floor
- "Shared Clock" among experiments
  - Data-Taking (DT) mode for 2 days and After-Data-Taking (A-DT) mode for 3 days

Overall T1s objectives for RUN3:

Indicate in this table, the bandwidths required for all T1s for reads and writes during data taking (DT) and right after data taking(A-DT).

VO	Reads (DT) GB/s	Writes (DT) GB/s	Reads (A-DT) GB/s	Writes (A-DT) GB/s
ALICE	0	2.8	1.1	2.8
ATLAS	2.5	9.6	8.4	5.1
CMS	0.8	7.6	12.3	1.1
LHCb		11	3.38	
Total	2.5	24.78	25.18	8.3

Tape Challenge Driven by Experiments WFs

- Experiment tape-tests driven by production workflow and dataflow management systems
  - ATLAS and CMS write and read tape-tests in both DT and A-DT modes
  - ALICE and LHCb test only T0 export to T1s, i.e. only write tape-tests
- Some common data management infrastructure among experiments
  - Rucio for ATLAS and CMS
  - FTS for ATLAS, CMS, and LHCb
  - Several T1s support multiple experiments (multi-VO sites)
- Common monitoring dashboard
  - ALICE has its own MonALISA monitoring (different data source than FTS)

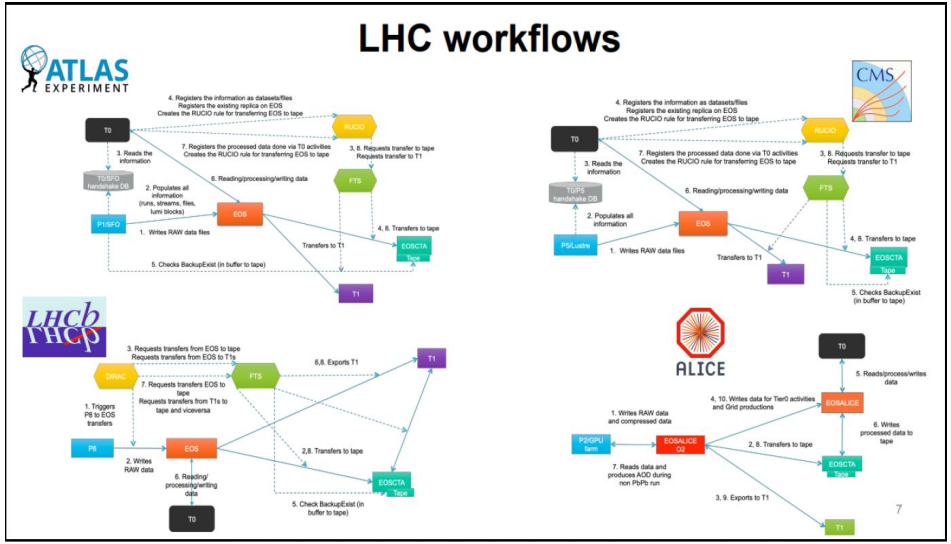
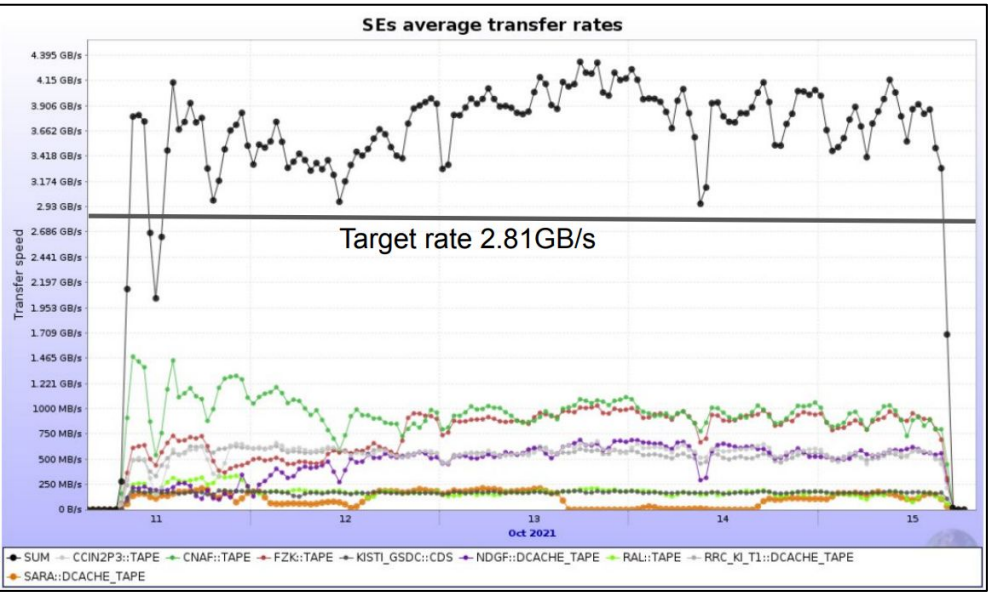


Diagram of the four LHC experiment workflows in Tape Challenge

ALICE Test Results

- Tested T0 export to T1 tape, DT and A-DT mode write test
- Smooth run, target rates achieved (exceeded), no particular issues found

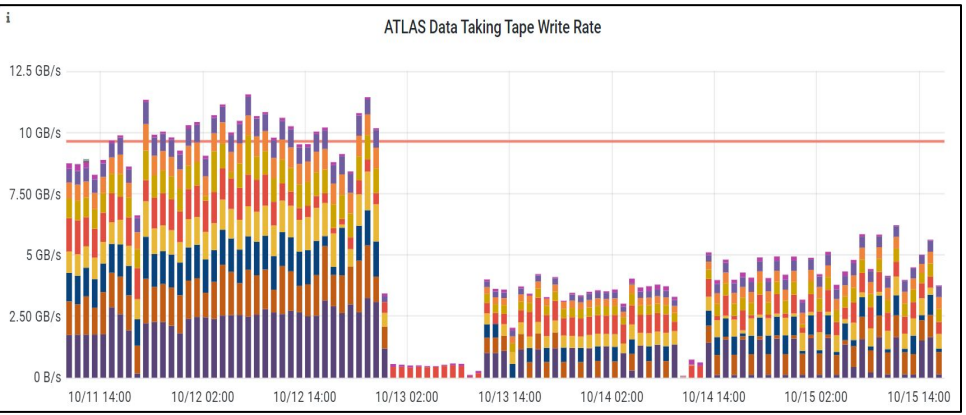


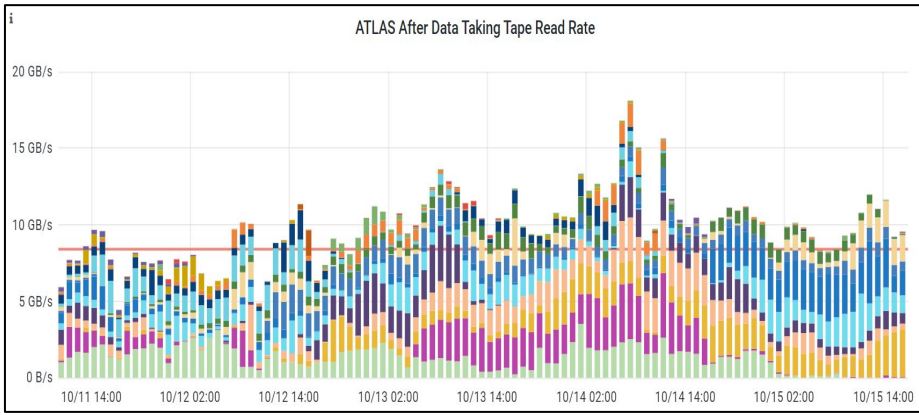
T1 Centre	Target rate GB/s	Achieved rate GB/s
CNAF	0.8	0.94 (116%)
IN2P3	0.4	0.54 (130%)
KISTI	0.15	0.16 (106%)
GridKA	0.6	0.76 (123%)
NDGF	0.3	0.47 (144%)
NL-T1	0.08	0.1 (122%)
RRC-KI	0.4	0.53 (128%)
RAL	0.08	0.17 (172%)

Sum 2.81GB/s

ATLAS Test Results

- Tests according to plan
  - Staging test piggybacked on the Data Carousel traffic (real production requests from the ongoing reprocessing and derivation campaigns)
- Overall target rates reached in both DT and A-DT modes, and staging rate exceeded target
  - Pressure on particular sites varied depending on the real production needs at the time

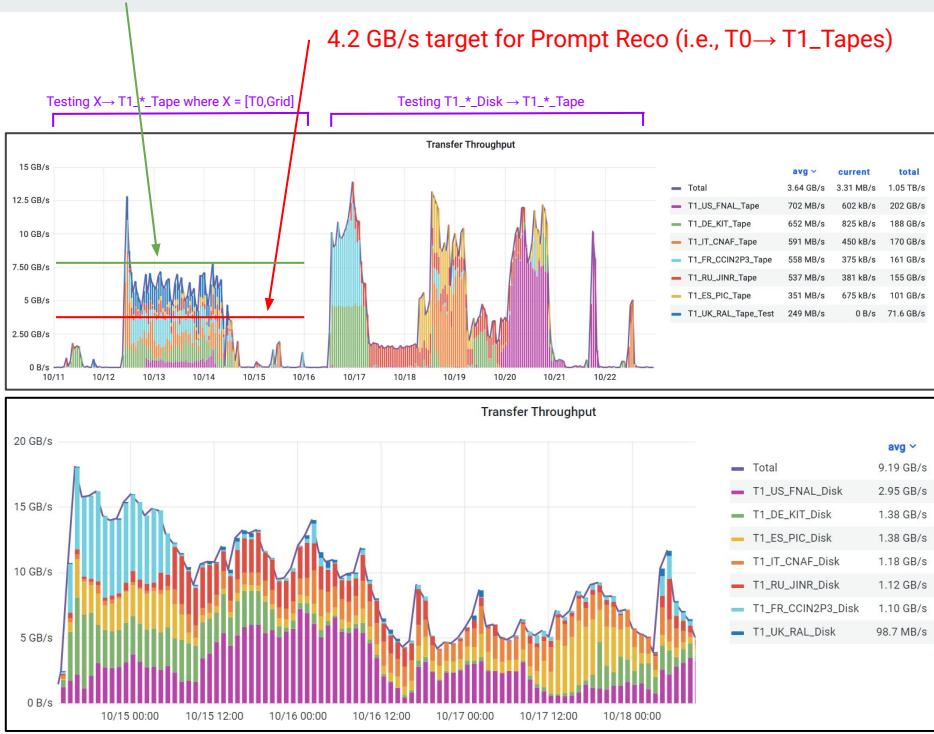




7.6 GB/s target is sum of estimated Prompt Reco (4.2 GB/s) and from MC production (3.4 GB/s) → almost no production traffic

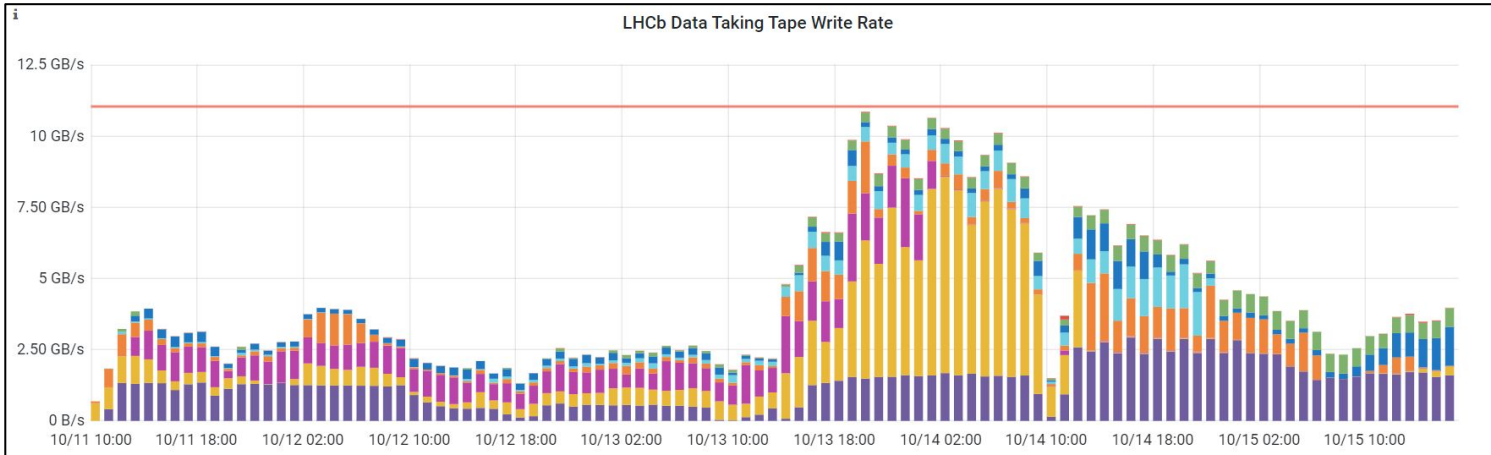
CMS Test Results

- Write test
  - 7.6 GB/s target not achieved
    - 4.2 GB/s target achieved (T0 export) despite large queue at FNAL
  - T1_Disk → T1_Tape
    - useful to measure max. tape write capacity
    - sites tested at different time (time zone difference vs. approval model)
- Read test → no specific target (stress system)
  - Staging with 300TB data bulks→ KIT, PIC, IN2P3 performed better wrt Spring Tape tests



LHCb Test Results

- Tested T0 export to T1 tape (DT mode write test - 11 GB/s target)
- Slow start
  - default FTS configuration settings, e.g. transfer limits per link and storage
  - not enough gridftp gateways at CERN EOS
  - misunderstanding of expected tape write rate by sites → not enough tape drives allocated
- On-the-fly fixing → close target



Tape Challenge Lessons Learned

- Common monitoring crucial → providing additional information to individual exp. dashboards
  - necessity of consistent representation of exp. data and inclusion of xrootd traffic
- Site tape activities information to be integrated into central monitoring (ongoing in WLCG)
- Future challenges to use srm+https protocol (gridftp deprecated)
- Site staging profile (also in ATLAS Data Carousel) useful for sites to customize incoming staging requests to fit site resource allocation
- Site feedback on optimal use of tape resources
  - write data to tape the way it will be read back and vice versa
  - file size affecting tape bandwidth utilisation in both migration and staging
  - match I/O of disk buffer and network with tape drive I/O

- **Close collaboration between experiments and sites** to face the storage challenge of the HL-LHC

Conclusions

- 2021 Data Challenges an overall success
  - very fruitful exercise, first of its kind
  - time and effort from several actors - e.g. central operations team, experiments, sites - pivotal for this success
- WLCG, experiments, and sites now with a clearer idea on what to expect during Run3
  - 2021 DCs essential step in the learning process towards future challenges and HL-LHC
- Further studies and dedicated tests as follow-up
  - 2022 Tape Challenge ongoing

- Contact: doma-data-challenges-development@cern.ch

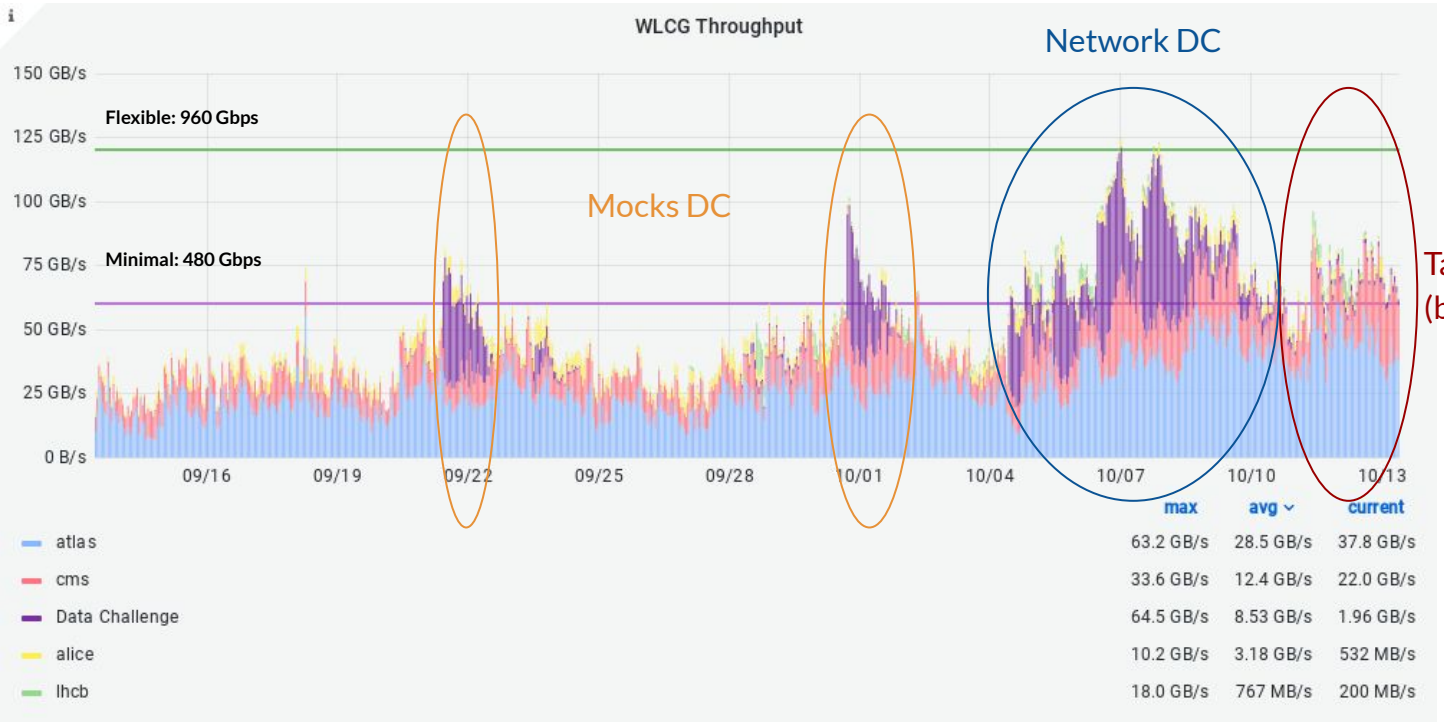
References and Authors

- WLCG Network Data Challenges 2021: wrap-up and recommendations
- WLCG Network Data Challenges 2021: DC week
- Tape Data Challenge 2021

ARSUAGA RIOS Maria
BETEV Latchezar
CAMPANA Simone
CHRISTIDIS Dimitrios
DI MARIA Riccardo
DONA Rizart
ELLIS Katy
FORTI Alessandra
GARRIDO BEAR Borja
GARZON Fernando
GOMEZ-CORTEZ Felipe
HAEN Christophe
LASSNIG Mario
MAIER Benedikt
MASCETTI Luca
PASPALAKI Garyfalia
PATRASCOIU Mihai
ZHAO Xin

Backup Slides

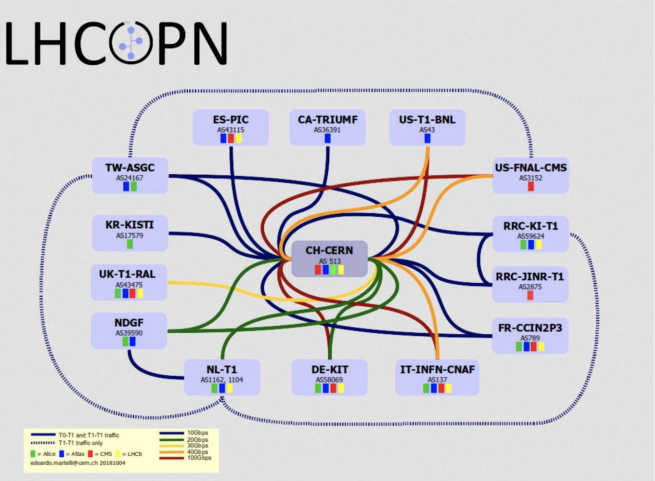
WLCG (FTS+XRootD) Throughput - 30-Day Plot

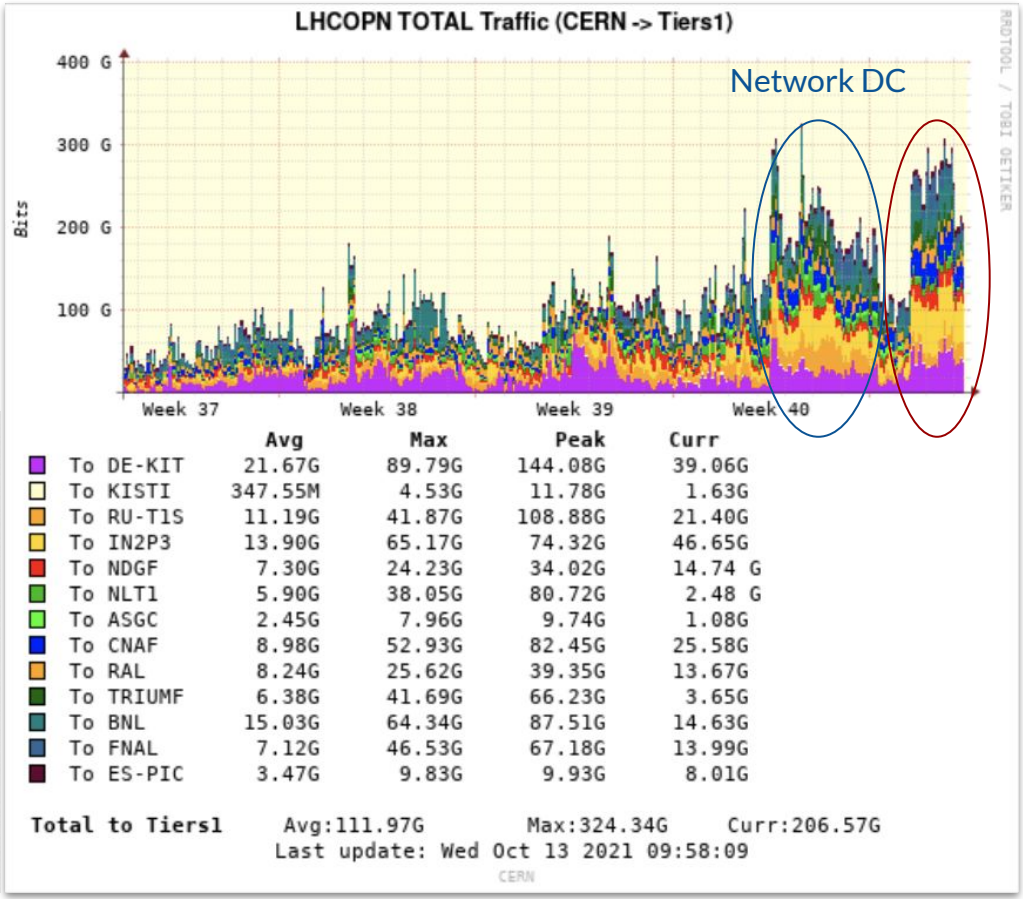


Tape Challenge (beginning of)

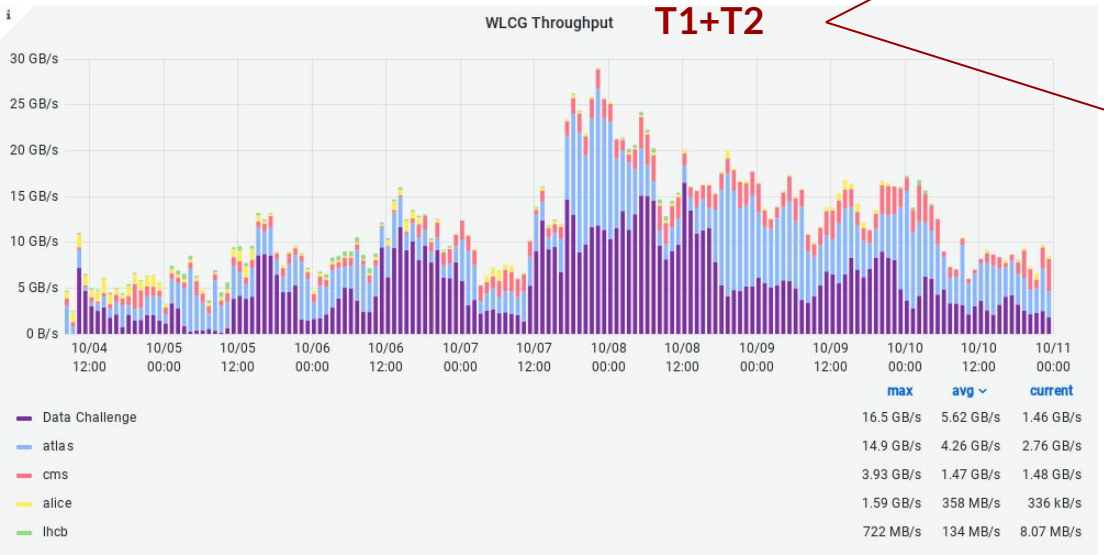
LHCOPN

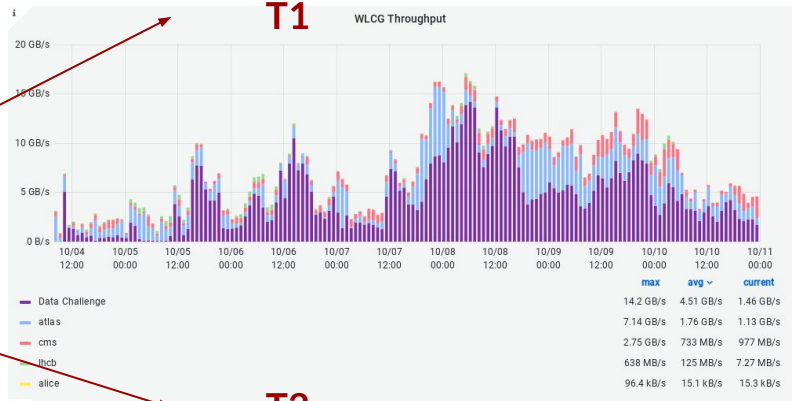
T0-T1 FTS traffic
vs
LHCOPN T1-T0-T1

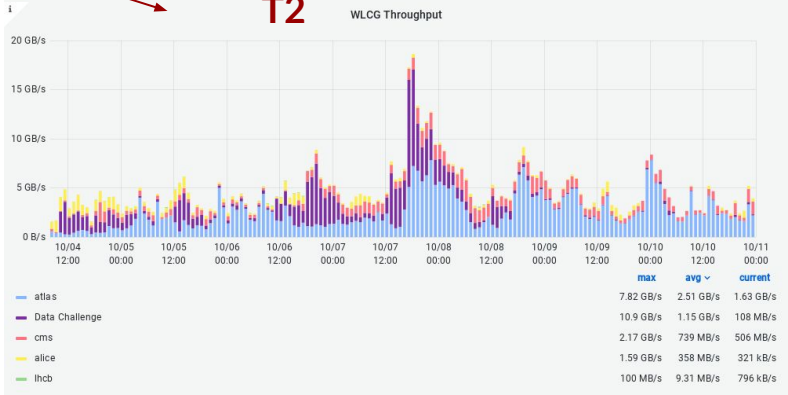




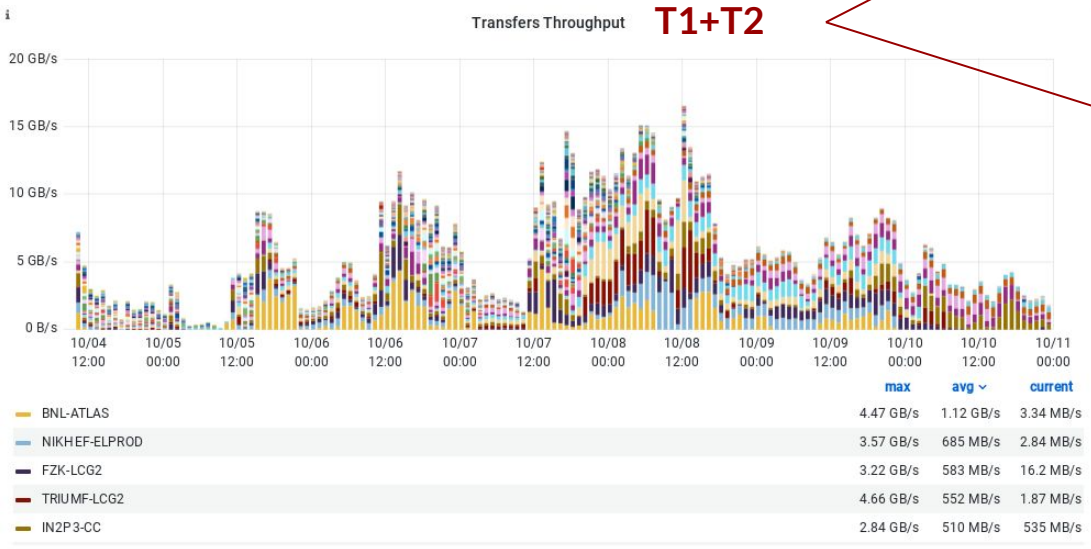
WLCG FTS+XRootD - T0->T1+T2

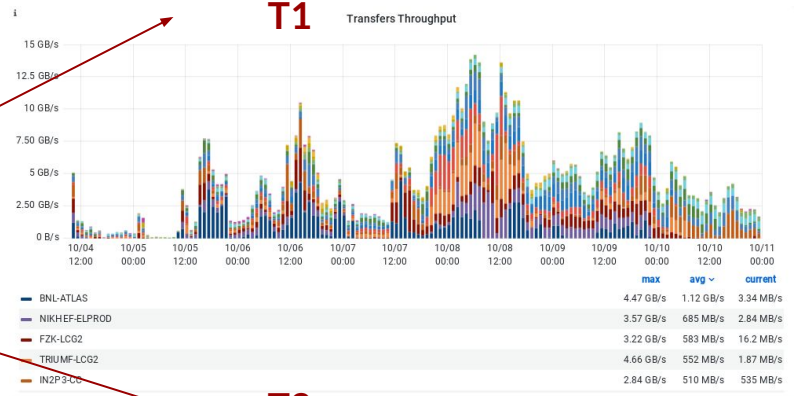


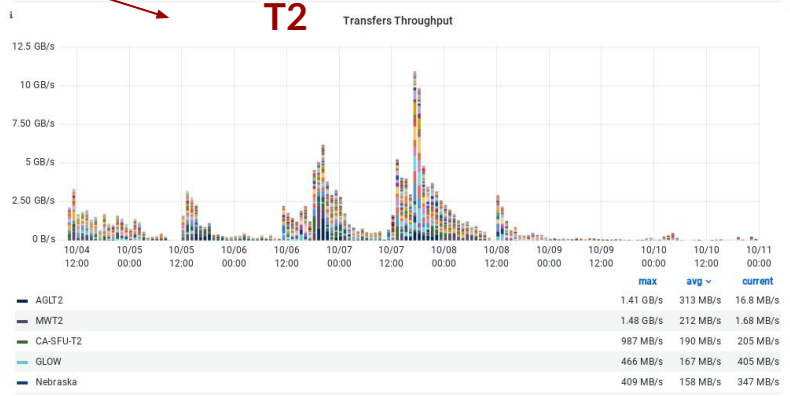




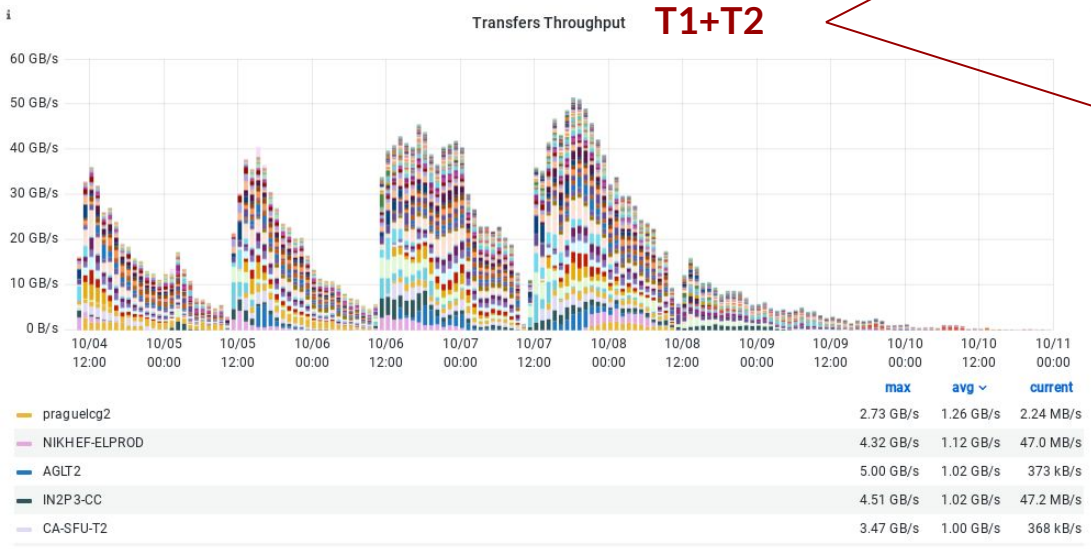
DC-Activity-only - T0->T1+T2



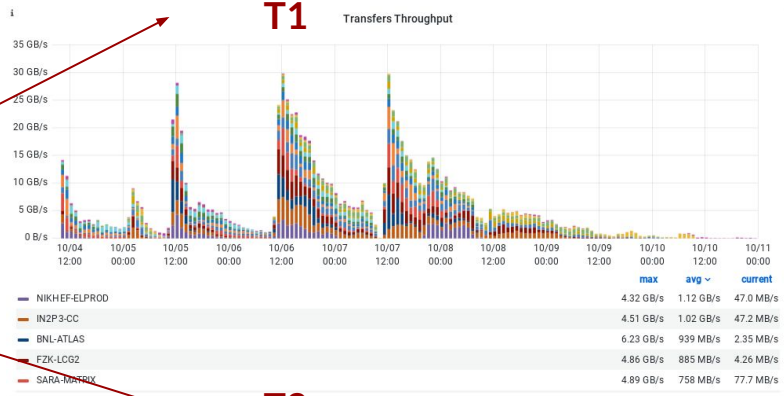




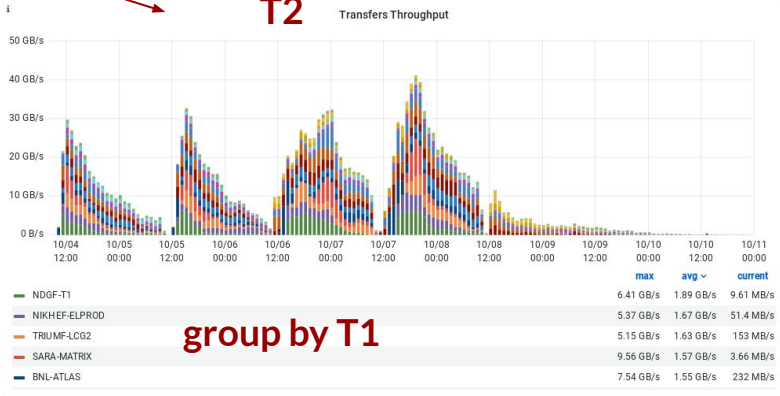
DC-Activity-only - T1->T1+T2



T1



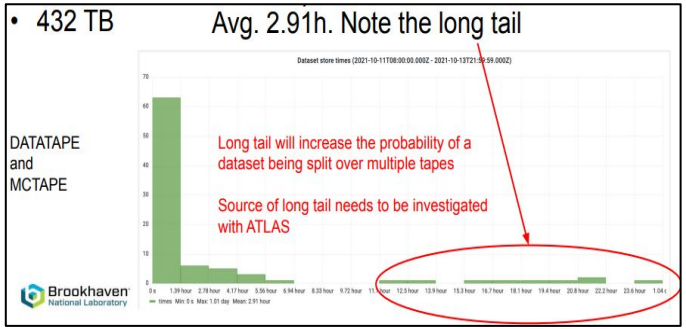
T2



group by T1

Tape Challenge Lessons Learned and Recommendations

- Site feedback on optimal use of tape resources
  - Write data to tape the way it will be read back; read data from tape the way it is written
    - Group files on tape during migration
    - Reduce time spent in mounting/dismounting/positioning during recall
  - Reduce the dataset write windows and the dataset bringonline windows, coming to sites.
  - Bigger files leads to higher tape bandwidth utilization in both migration and staging
  - File aggregation helps with migration but hurts staging performance (unless the whole aggregate is recalled)
  - Collaboration between sites and experiments is needed for more optimal tape resources utilization.



- CMS staging performance during A-DT better than ATLAS b/c
  - Bigger files (good for both migration and staging)
  - Less scattered across tapes
  - Less competing activities
  - Better data distribution across drive sets (more staging drives for CMS and underperforming drives for ATLAS)