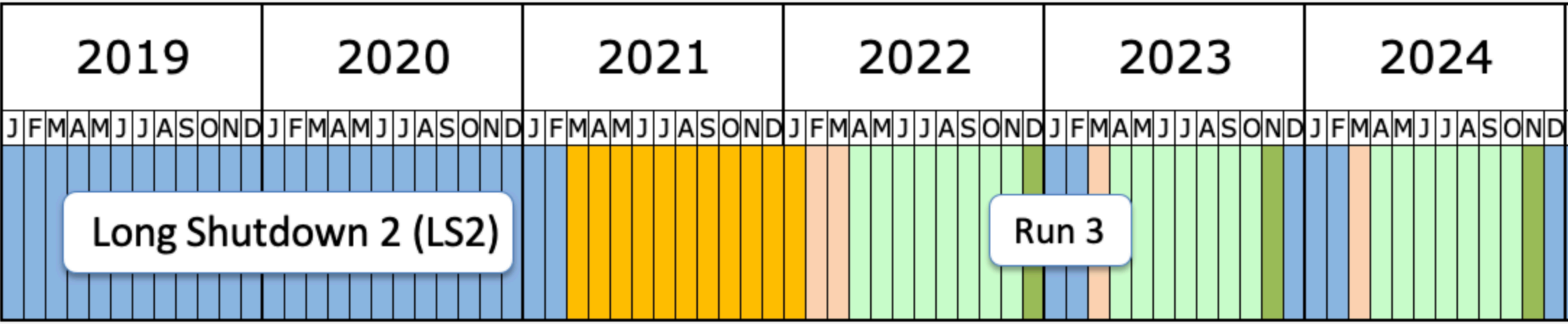


LHC Data Storage: RUN 3 Data Taking Commissioning

Maria Arsuaga-Rios EOS Workshop 07/03/22

CERN
Tape Archive

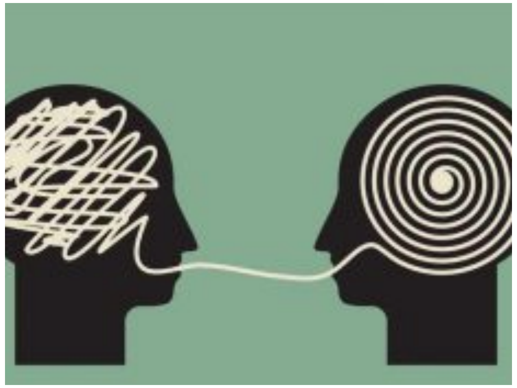
EOS

FTS
File Transfer Service

The CERN IT Storage Group ensures the symbiotic development and operations of storage and data transfer services for all CERN physics data, in particular the data generated by the four LHC experiments (ALICE, ATLAS, CMS and LHCb).

ALICE ATLAS EXPERIMENT CMS LHCb

RUN 3 Data Taking Commissioning: Main goals



Full understanding

- Experiment workflows
- RUN 3 objectives per experiment and component

Excellence and Innovation

- Test the storage solutions
- Identify all possible issues before RUN3

Trust

- Involve the experiment teams on participating, defining and validating the tests

RUN 3 Data Taking Commissioning

Planning and Communication

LHC workflows for RUN 3

Testing individual components

Individual data challenge

Combined data challenge

Follow up tests

Planning and Communication: Covid-19 Impact

- Multiple delays in hardware procurement
- Only virtual communications

Standard framework definition

Approve and validate the goals, dates, and blocking factors for individual tests, per component and per experiment

Individual tests per component and experiment

Data challenge tests with more than 3 experiments together

Dedicated storage meetings

Planning data challenges for RUN3

ATLAS

- Data challenges plan in ATLAS
- ATLAS Software week presentation
- New ATLAS network dashboard
- ATLAS Software week presentation

CMS

- Data challenges plan in CMS
- New CMS network dashboard
- CMS short outcome presentation after 1st combined Data Challenge

LHCB

- Data challenges plan in LHCB
- LHCB network dashboard

ALICE

- Data challenges plan in ALICE
- ALICE network dashboard

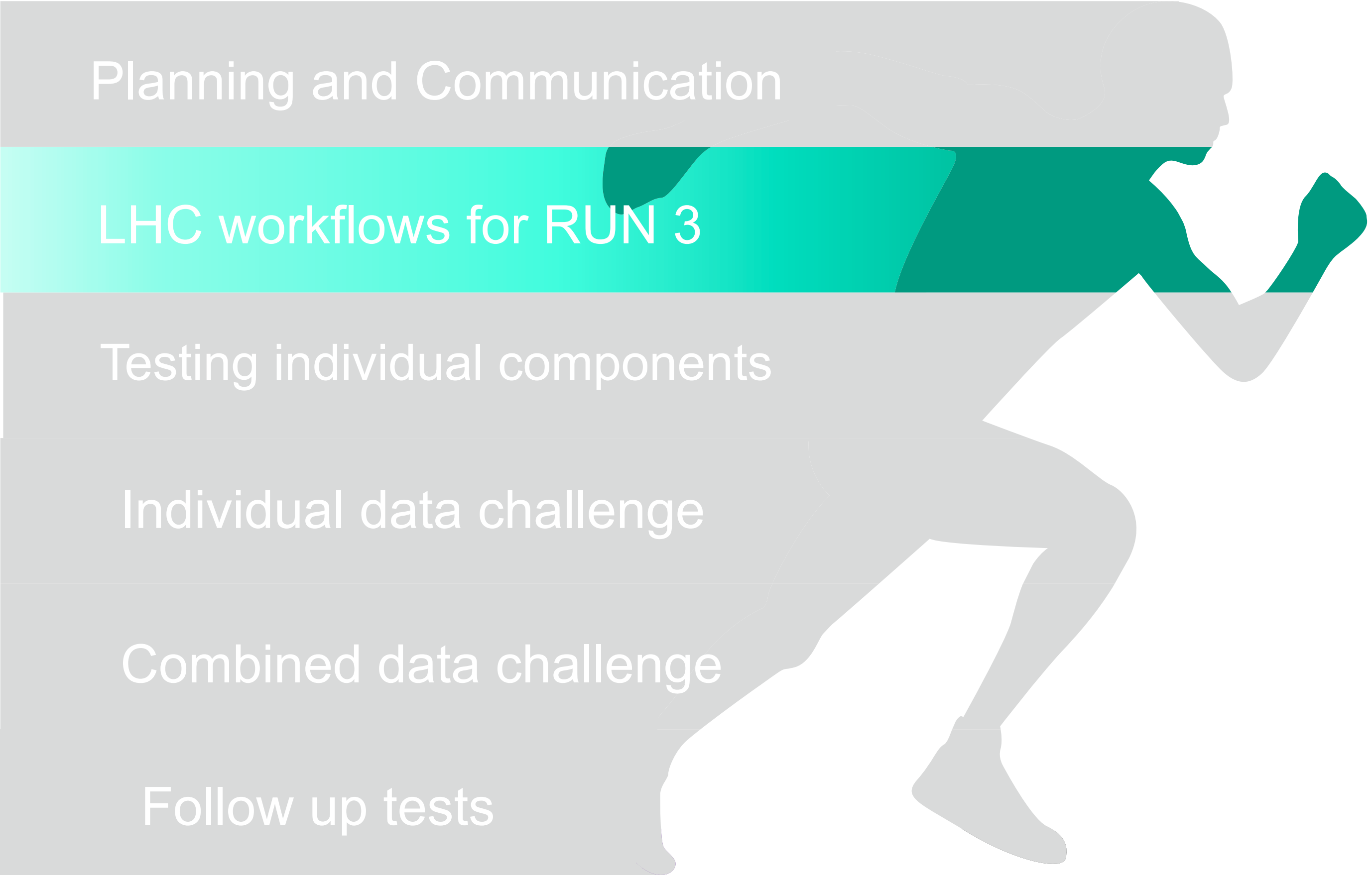
Combined tape data challenge including Tier1s

- Data challenges for tapes from ALL experiments

Internal ST tests

- Internal summary for data challenges
- Internal ST tests

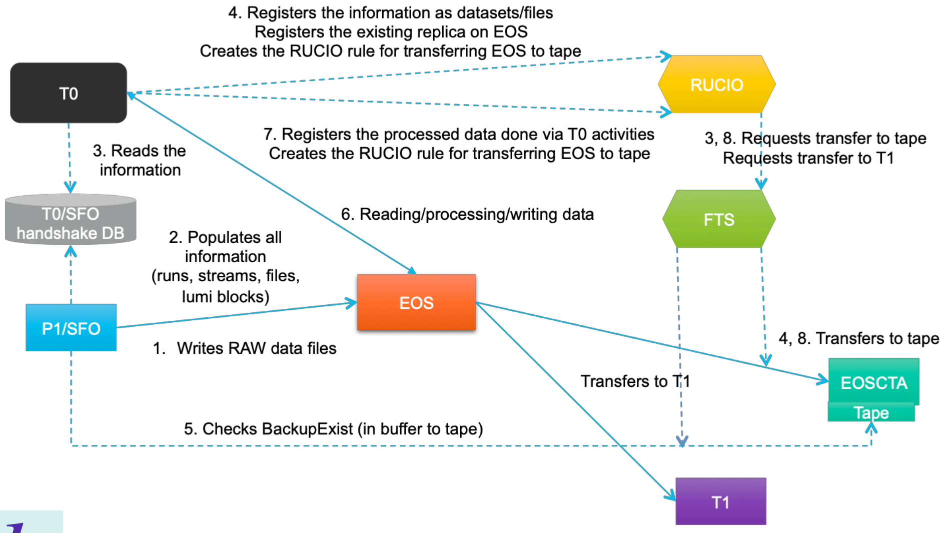
RUN 3 Data Taking Commissioning

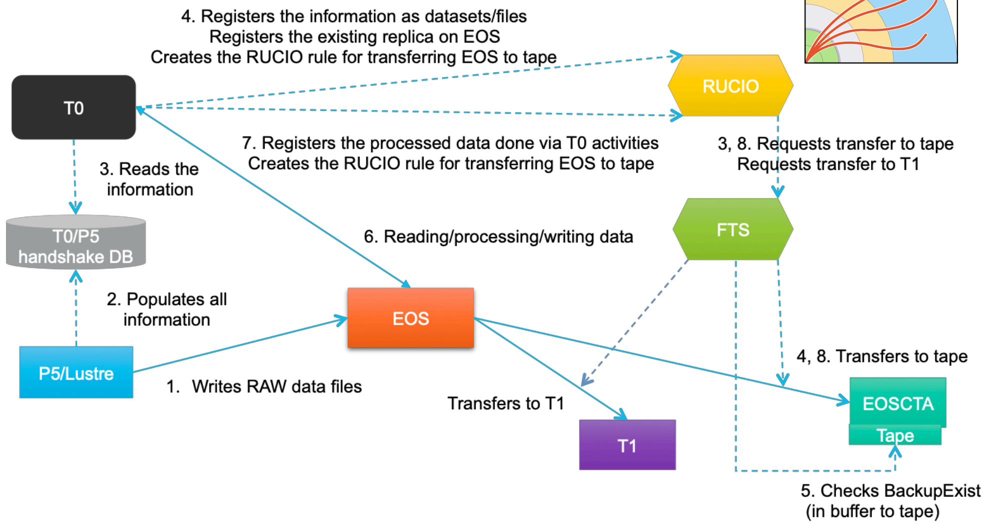

Planning and Communication
LHC workflows for RUN 3
Testing individual components
Individual data challenge
Combined data challenge
Follow up tests

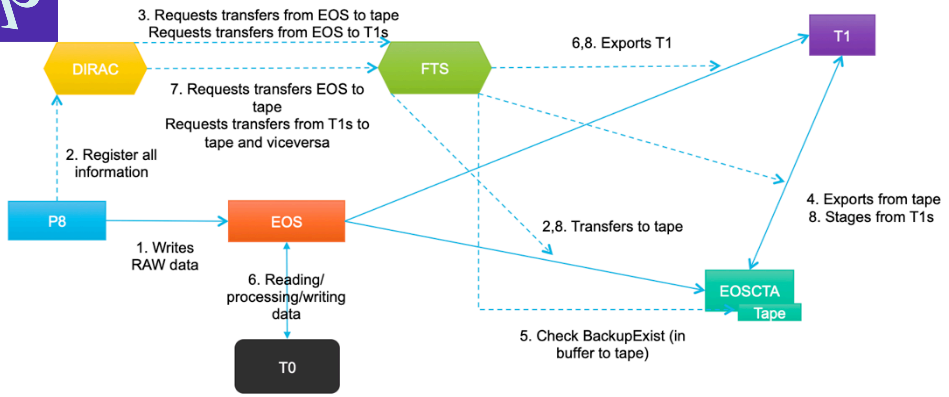
Goals

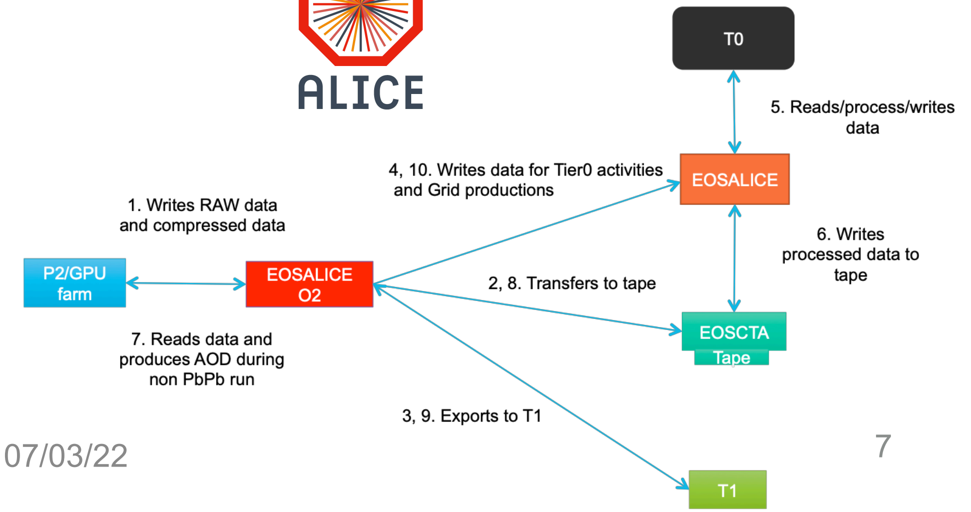
- Full understanding of all components activities and interactions
- Easier communication with stakeholders

LHC workflows

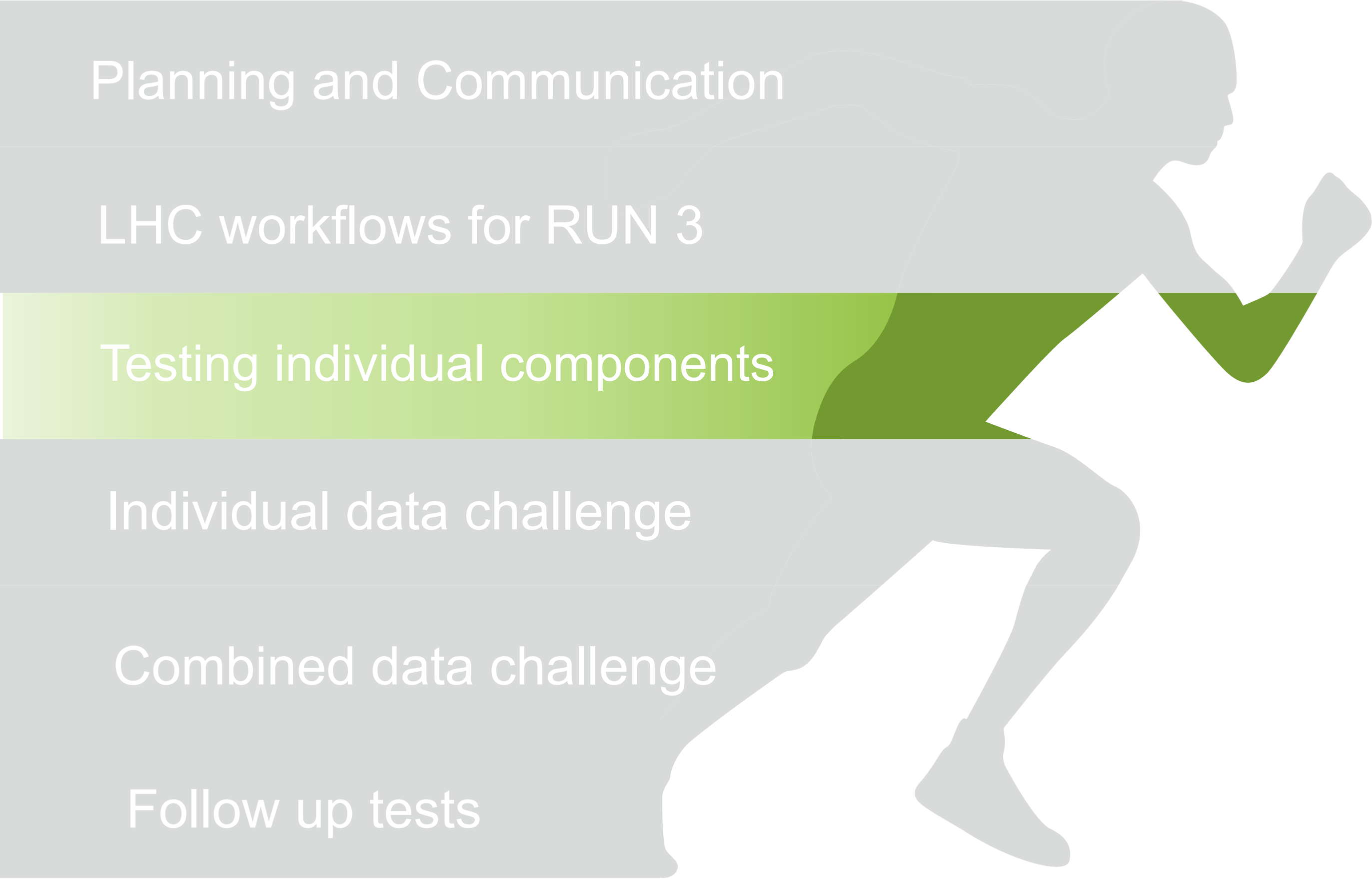

4. Registers the information as datasets/files
Registers the existing replica on EOS
Creates the RUCIO rule for transferring EOS to tape
T0
RUCIO
3. Reads the information
7. Registers the processed data done via T0 activities
Creates the RUCIO rule for transferring EOS to tape
3, 8. Requests transfer to tape
Requests transfer to T1
T0/SFO handshake DB
6. Reading/processing/writing data
FTS
2. Populates all information (runs, streams, files, lumi blocks)
EOS
P1/SFO
1. Writes RAW data files
4, 8. Transfers to tape
Transfers to T1
EOSCTA
Tape
5. Checks BackupExist (in buffer to tape)
T1


4. Registers the information as datasets/files
Registers the existing replica on EOS
Creates the RUCIO rule for transferring EOS to tape
RUCIO
T0
7. Registers the processed data done via T0 activities
Creates the RUCIO rule for transferring EOS to tape
3, 8. Requests transfer to tape
Requests transfer to T1
3. Reads the information
T0/P5 handshake DB
6. Reading/processing/writing data
FTS
2. Populates all information
EOS
P5/Lustre
1. Writes RAW data files
4, 8. Transfers to tape
Transfers to T1
EOSCTA
Tape
T1
5. Checks BackupExist (in buffer to tape)


3. Requests transfers from EOS to tape
Requests transfers from EOS to T1s
T1
DIRAC
FTS
6,8. Exports T1
7. Requests transfers EOS to tape
Requests transfers from T1s to tape and viceversa
2. Register all information
4. Exports from tape
8. Stages from T1s
P8
EOS
2,8. Transfers to tape
1. Writes RAW data
6. Reading/processing/writing data
EOSCTA
Tape
5. Check BackupExist (in buffer to tape)
T0


T0
5. Reads/process/writes data
4, 10. Writes data for Tier0 activities and Grid productions
EOSALICE
1. Writes RAW data and compressed data
6. Writes processed data to tape
P2/GPU farm
EOSALICE O2
2, 8. Transfers to tape
EOSCTA
Tape
7. Reads data and produces AOD during non PbPb run
3, 9. Exports to T1
T1

RUN 3 Data Taking Commissioning

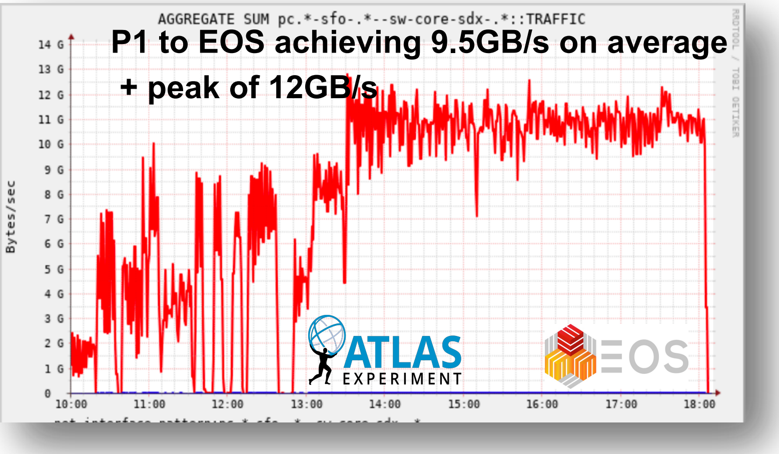


- Planning and Communication
- LHC workflows for RUN 3
- Testing individual components
- Individual data challenge
- Combined data challenge
- Follow up tests

Goals

- Obtain clear objectives per component
- Test our solutions with the new RUN 3 challenges per individual components

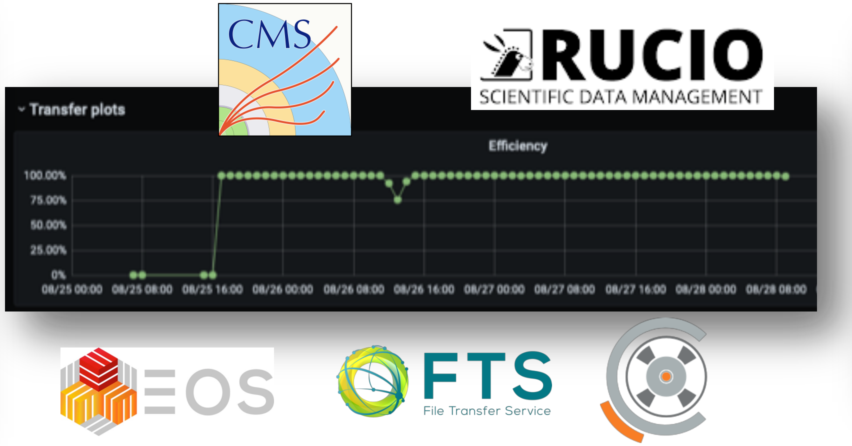
Testing Individual Components



P1 to EOS achieving 9.5GB/s on average + peak of 12GB/s

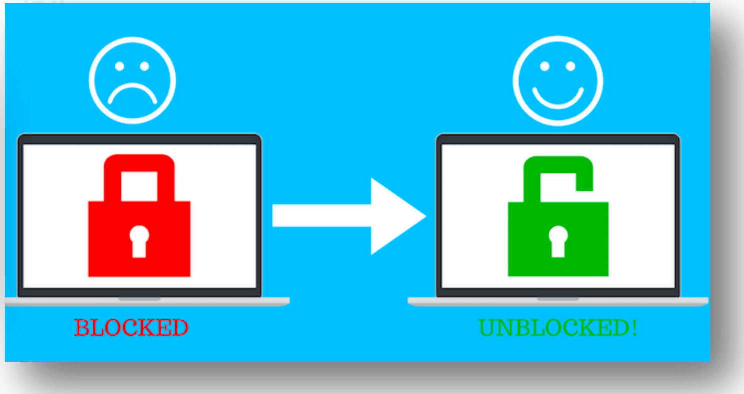
Stressing ATLAS Data Ingestion on EOS

- We achieved more than the tageted goal (7GB/s) with 9.5 GB/s average with peack of 12GB/s.
- The maximum peak found when testing and running RUN 2 was 7GB/s.
- **EOS handled this traffic without any problem**



Full Validation of CMS DDM Infrastructure

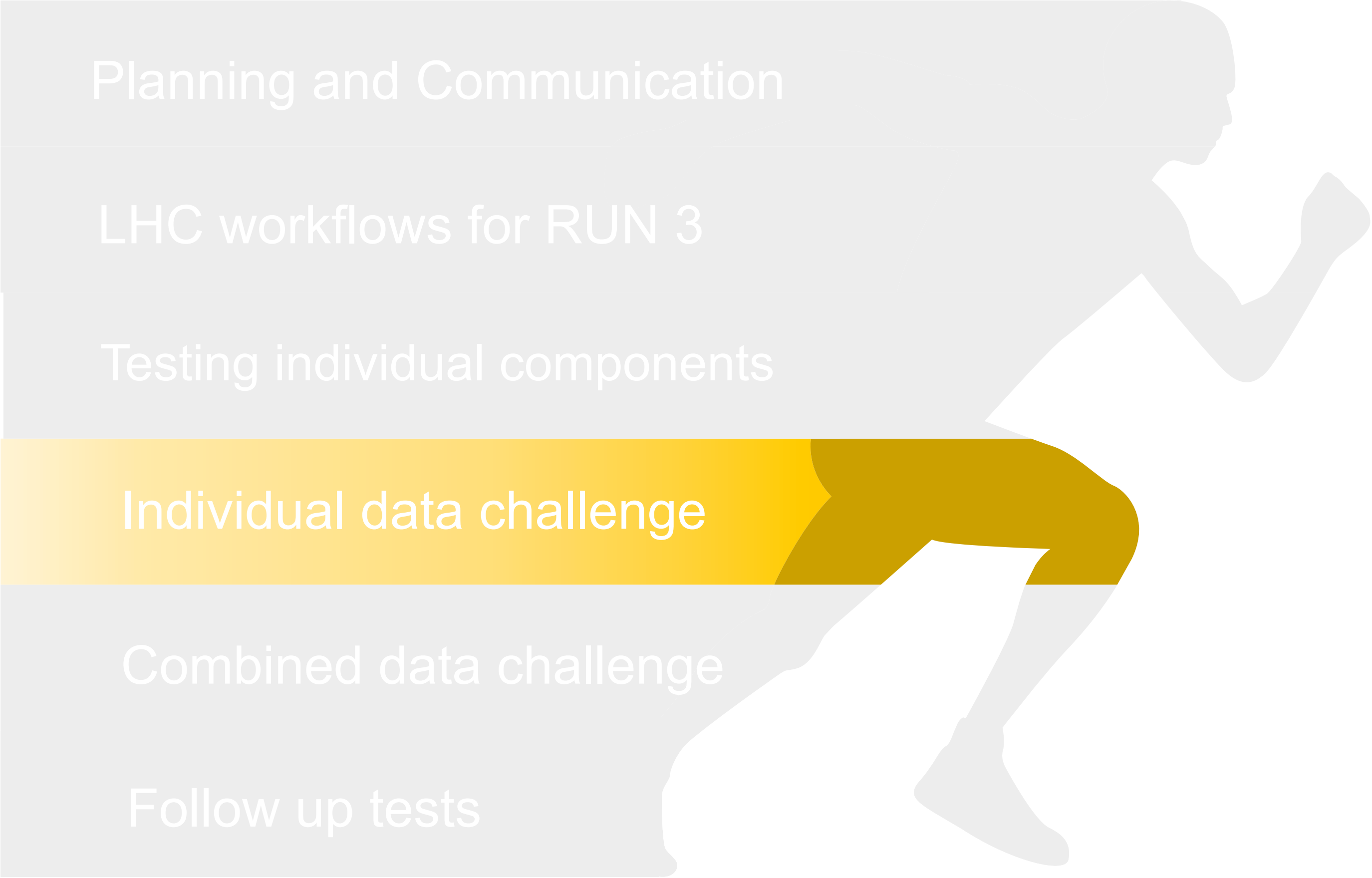
- Crucial follow up from ST during CMS migration from PhEDEx to Rucio
- Large test from EOS to CTA (675 TB) with the Rucio production instance



LHCb Validation of Data Export via HTTP TPC

- Follow up with Tier1s the deprecation of SRM-gridftp
- Deploy and configure XRootD or HTTP Third Party Copy (TPC)
- Validation test of 200 TB from EOSLHCb to CTA
- Multiprotocol submission model: XRootD stage-only and HTTP-TPC transfers from T0 to T1s

RUN 3 Data Taking Commissioning



Planning and Communication

LHC workflows for RUN 3

Testing individual components

Individual data challenge

Combined data challenge

Follow up tests

Goals

- Test the most complete workflow with realistic RUN 3 activity
- Incentive for other experiments to participate in the combined data challenge

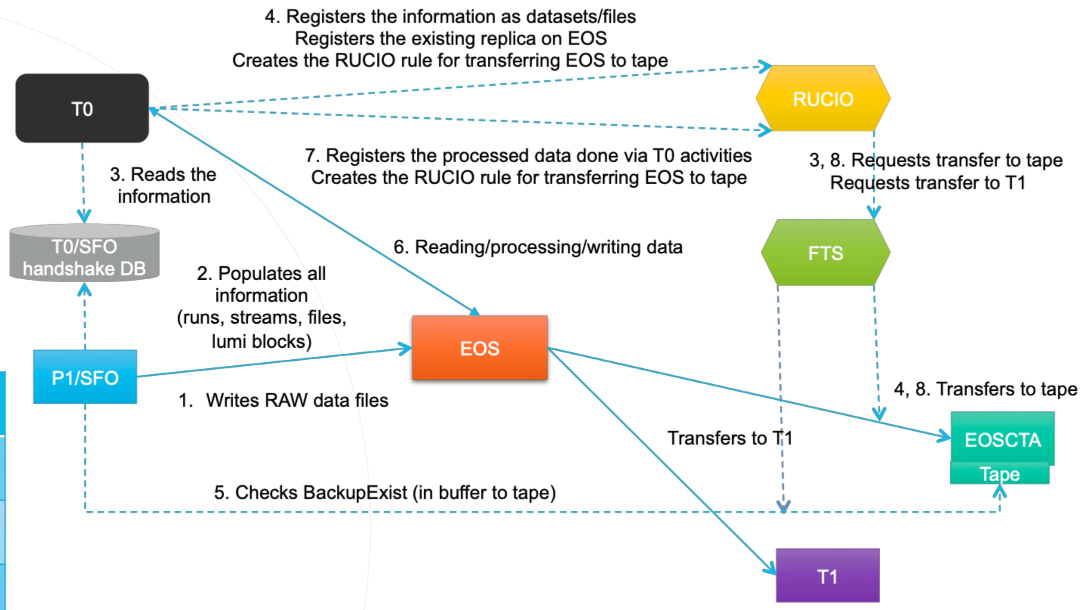
ATLAS data challenge with realistic RUN 3 activity

Purpose

Evaluate the whole processing chain, from SFO to EOS, repacking and storing into CTA and export to Tier 1 .

Throughput goals (see table)

Source	Destination	Operation	Throughput	Data type
SFO	EOS	write	8GB/s	RAW
EOS	batch	read	3GB/s	RAW
batch	EOS	write	2GB/s	AOD*
EOS	CTA	write	8GB/s + 2GB/s	RAW + AOD*
EOS	Tier1 Disk	export	8GB/s	RAW
EOS	Tier1 Disk	export	2GB/s	AOD*
EOS	Tier2 Disk	export	2GB/s	AOD*


4. Registers the information as datasets/files
Registers the existing replica on EOS
Creates the RUCIO rule for transferring EOS to tape
T0
RUCIO
3. Reads the information
7. Registers the processed data done via T0 activities
Creates the RUCIO rule for transferring EOS to tape
3, 8. Requests transfer to tape
Requests transfer to T1
T0/SFO handshake DB
6. Reading/processing/writing data
FTS
2. Populates all information (runs, streams, files, lumi blocks)
EOS
P1/SFO
1. Writes RAW data files
4, 8. Transfers to tape
Transfers to T1
EOSCTA
Tape
5. Checks BackupExist (in buffer to tape)
T1

ATLAS Tier0 Scenario:

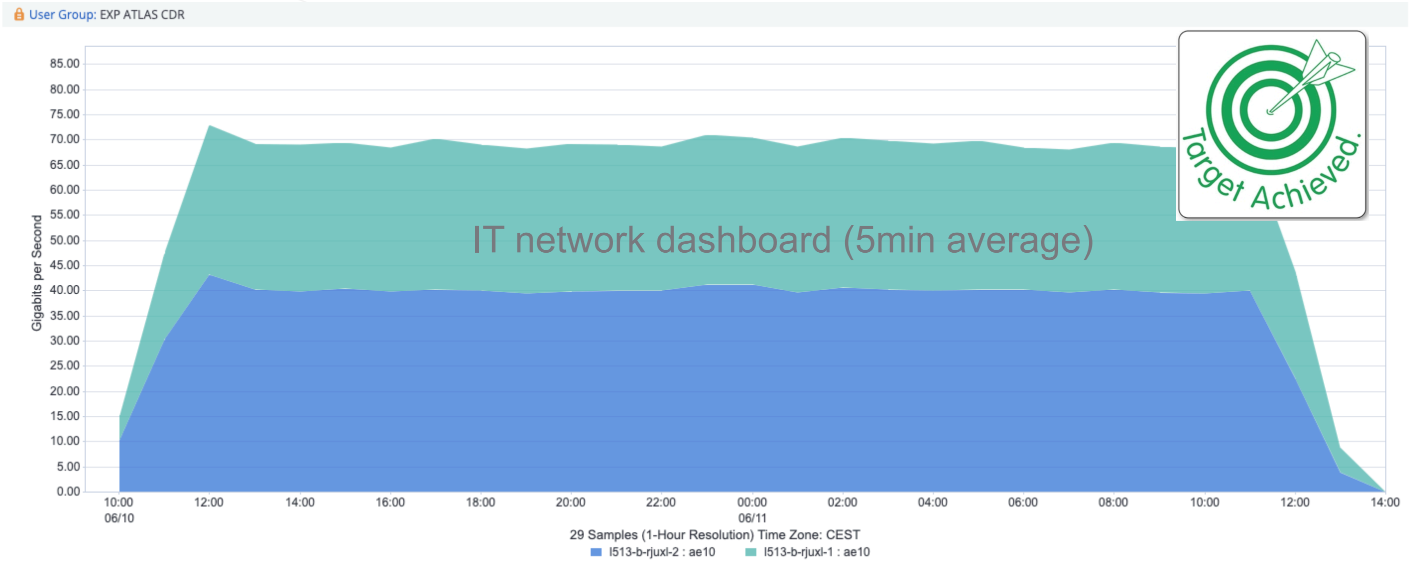
1. "Prompt reconstruction":
   - reading 3 GB/s RAW data EOS -> batch
   - writing 1 GB/s AOD batch -> EOS
2. "Merging":
   - reading 1 GB/s AOD EOS -> batch
   - writing 1 GB/s batch -> EOS
3. Registration of "merged derived products" in Rucio for CTA backup and export

* AOD = transient derived products

ATLAS data challenge with realistic RUN 3 activity

Results

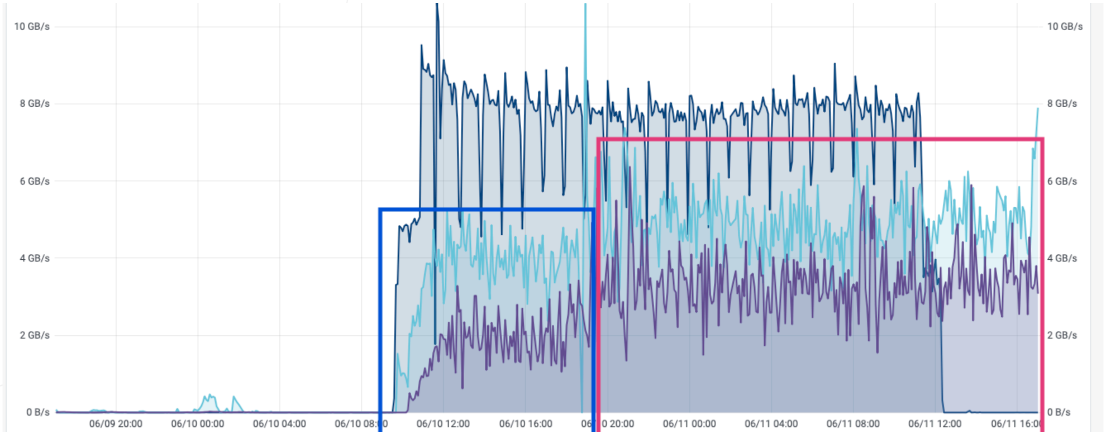
- **ATLAS Point 1 activity to IT EOS**:
  writes ~8GB/s average
  Expected throughput achieved
- **ATLAS Tier 0 activity to IT EOS:**
  writes 2GB/s average and reads 3GB/s
  Expected throughput was achieved but.. after the first 8 hours..



IT network dashboard (5min average)

How was it handled?

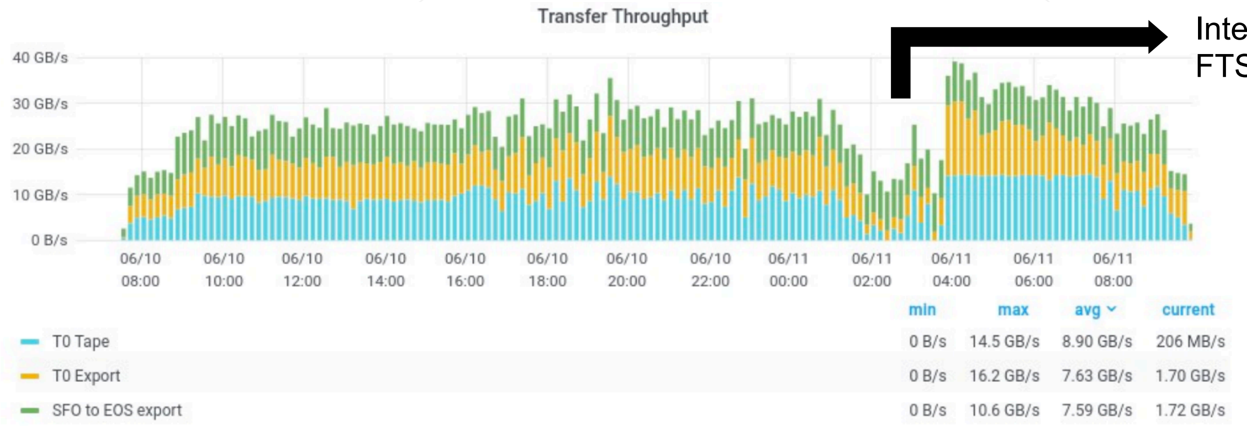
- **from ATLAS Tier0 side**: by changing the job configuration to produce correspondingly bigger AOD. output files (60% of the RAW input size)
- **from IT EOS side**: by adding automatic removal of overloaded storage nodes.



ATLAS data challenge with realistic RUN 3 activity

Results

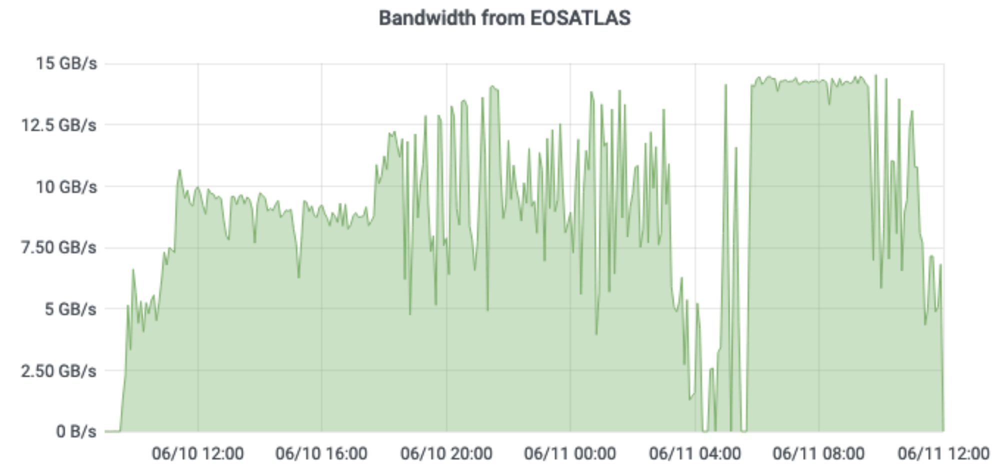
ATLAS Rucio exports activity: Expected exporting throughputs achieved

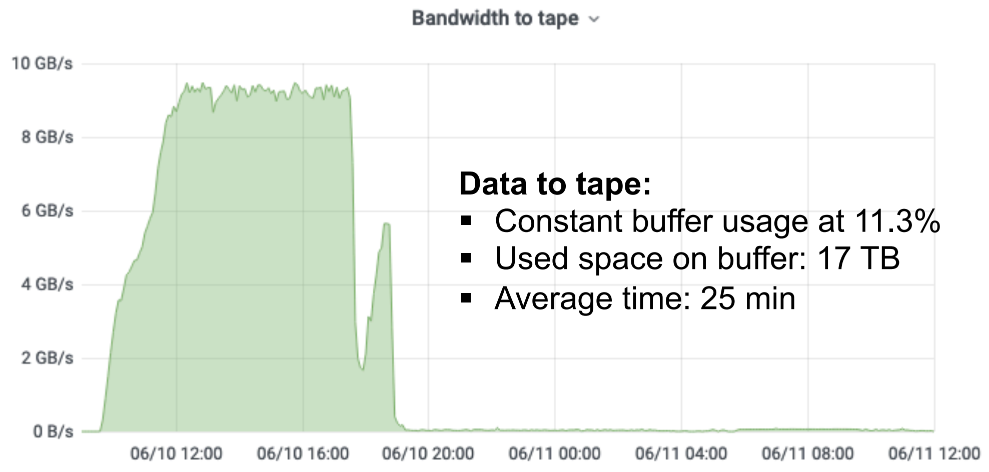


	min	max	avg	current
T0 Tape	0 B/s	14.5 GB/s	8.90 GB/s	206 MB/s
T0 Export	0 B/s	16.2 GB/s	7.63 GB/s	1.70 GB/s
SFO to EOS export	0 B/s	10.6 GB/s	7.59 GB/s	1.72 GB/s

Interference between DBoD backup and FTS monitoring under investigation

IT CTA activity: Expected throughputs achieved ~10GB/s





Data to tape:

- Constant buffer usage at 11.3%
- Used space on buffer: 17 TB
- Average time: 25 min

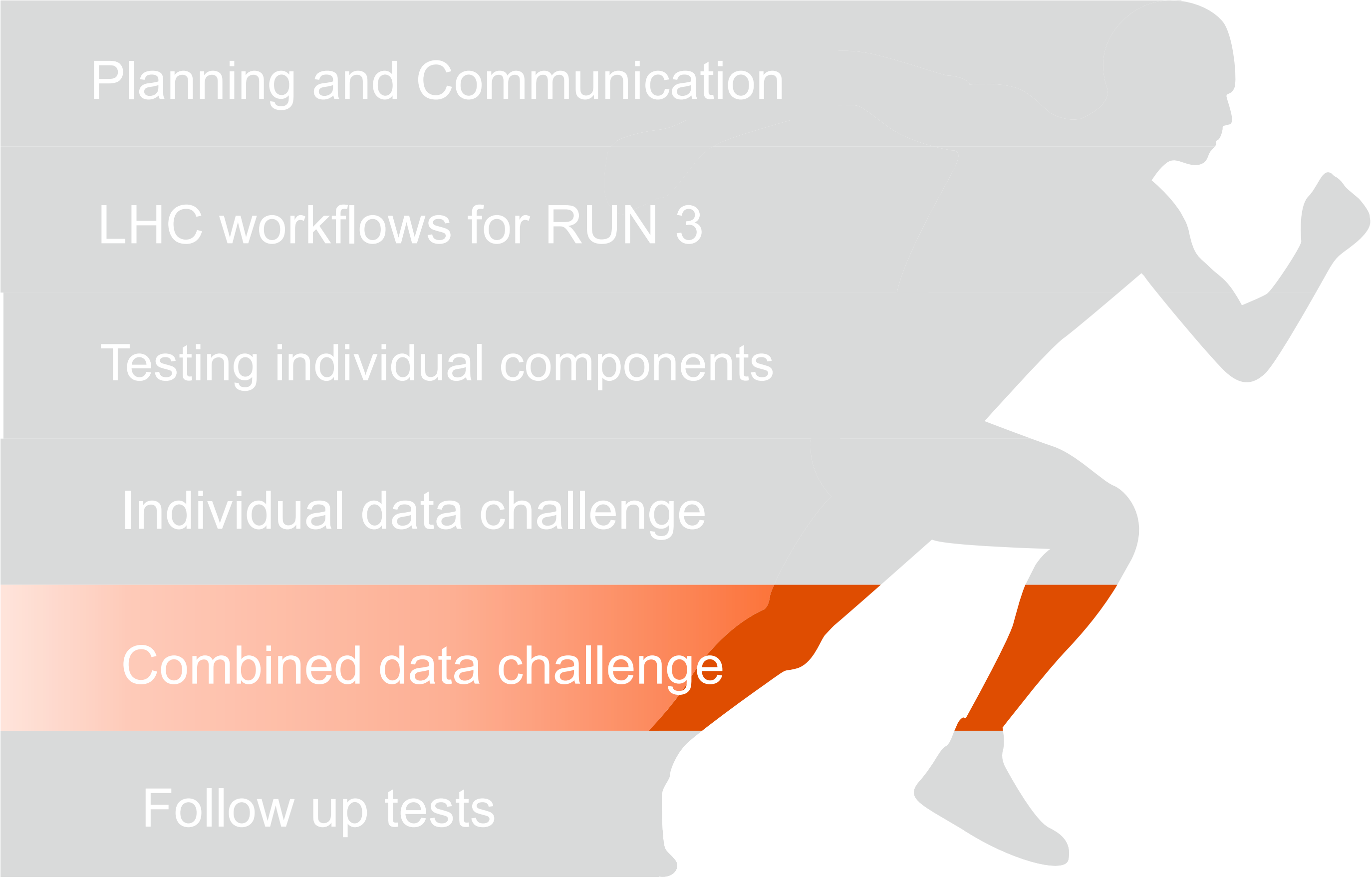
ATLAS data challenge with realistic RUN 3 activity

Conclusions: **Expected throughputs were finally achieved from all parties**

Improvements for next tests:

- **ATLAS SFO**: Correct configuration for the CTA directory.
- **ATLAS Tier0**: Use bigger files from SFO (tentative size 5GB) and include more files per job.
- **IT EOS**: Protecting measure to avoid slow storage nodes online.
- **IT FTS**: Workaround for force-kill hanging queries from the Web monitoring.

RUN 3 Data Taking Commissioning



Planning and Communication

LHC workflows for RUN 3

Testing individual components

Individual data challenge

Combined data challenge

Follow up tests

Goals

- Evaluate the IT infrastructure: Network, EOS, CTA and FTS.
- Detect possible interactions between experiments.
- Trust on the readiness of the storage team

Combined Data Challenge (ALICE, ATLAS,CMS, LHCb)

Experiment	DAQ System Throughput	Experiment Tier0 Activity	EOS to CTA (no tape writes)	Tier1s Export
ALICE	25 GB/s (1/4 of the capacity)	Standard data analysis workflow: Expected load during the first year of RUN3	10 GB/s + current load for reprocessing/ staging	None
ATLAS	8 GB/s	Read -> 3GB/s Write -> 2GB/s	10 GB/s	Export to Tier1 DISK : 10 GB/s Export to Tier2 DISK: 2 GB/s
CMS	20 GB/s *	Stressing EOS with repacking jobs	10 GB/s *	None
LHCb	10 GB/s *	None	10 GB/s*	None

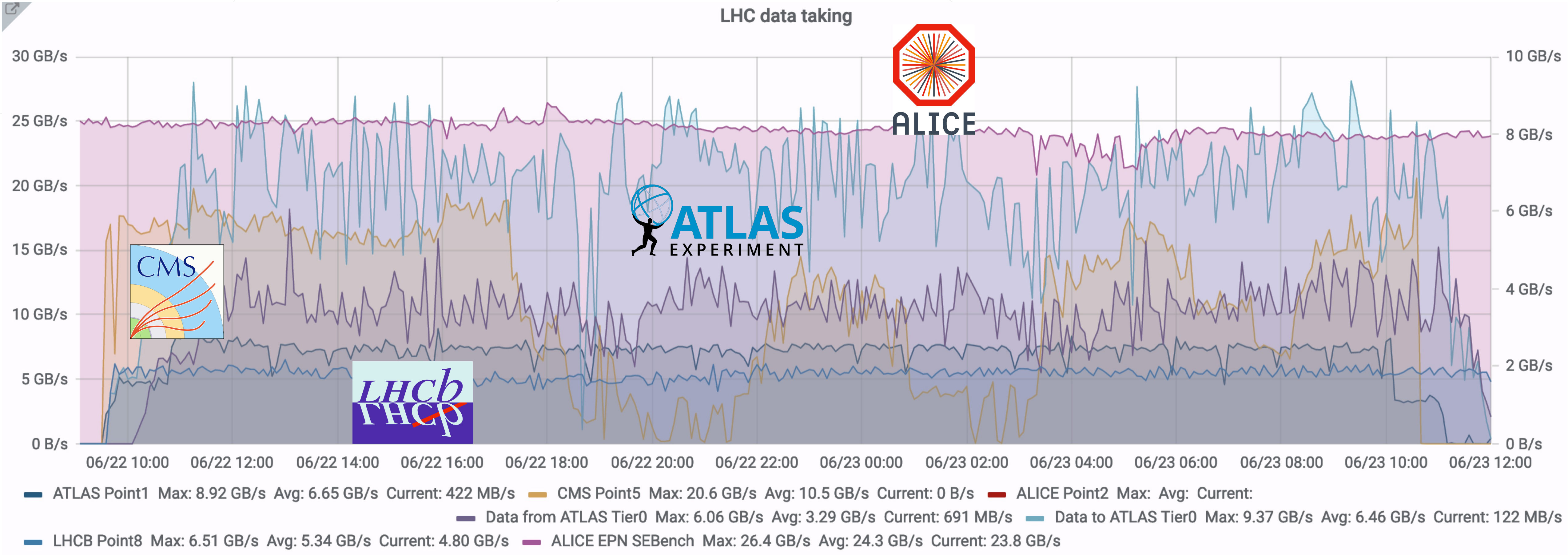
*Synthetic load

Combined Data Challenge: DAQ Systems

Target Achieved.

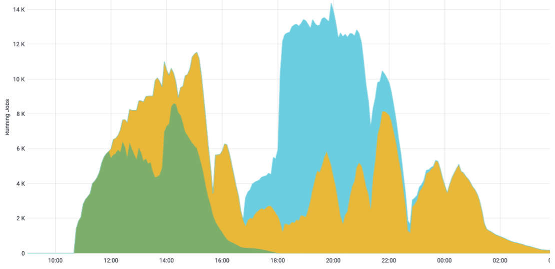
The four LHC experiments achieved their objectives

* only CMS had some disruptions



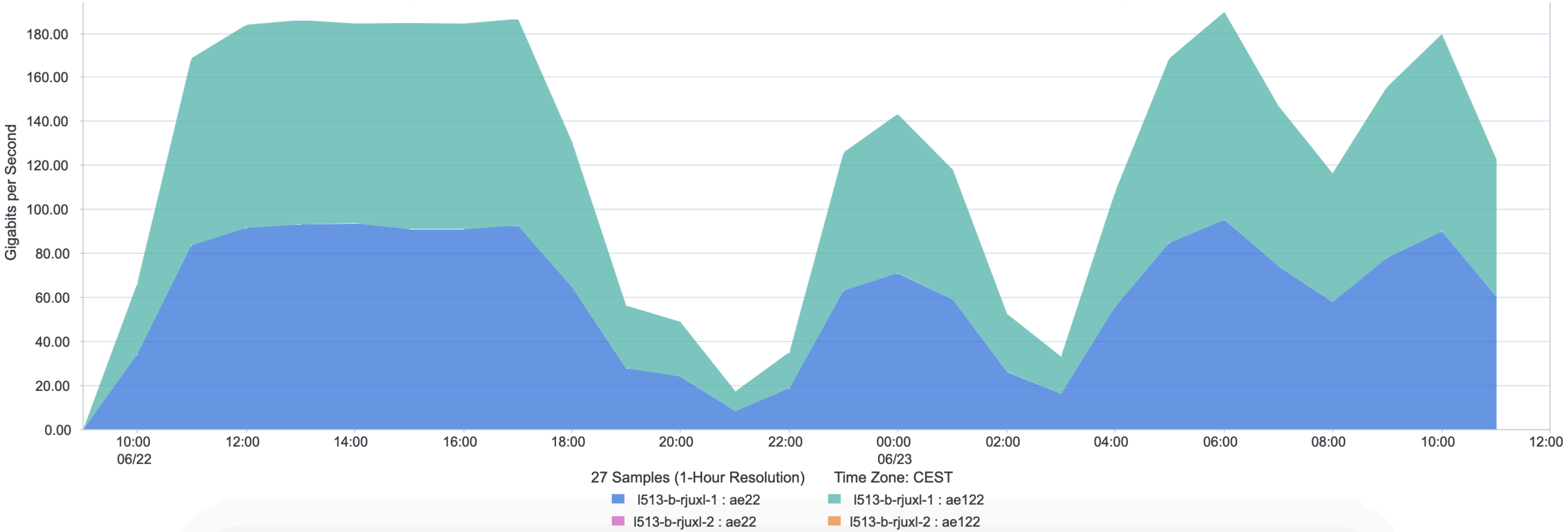
Combined Data Challenge: CMS disruptions

CMS Tier0 activity:
Simulate worst scenario
with repacking jobs



Disruptions factors:

1. CMS Point 5 directory configuration changed replication factor (1->2 replica)
2. CMS Tier0 repacking jobs impact: 2.6x actual production
3. Unbalanced scheduling on storage nodes



Combined Data Challenge: CMS disruptions

Solution:

- **Improvement from EOS side:**

  Scheduling improvements to avoid starvation:

  1. Flat out scheduling – independent from physical location (since Wigner Computer Center no longer exists)
  2. Introduce scheduling threshold values for reads and writes to avoid stream aggregations in the storage nodes (EOS-4762)

  Ensure high priority for CMS Point 5 writes:

  1. Allow to tag IO priority on streams: writes over reads (EOS-4759)
  2. Might configure a dedicated pool of disk for CMS Point 5

- **Improvement from CMS Point 5 side:**

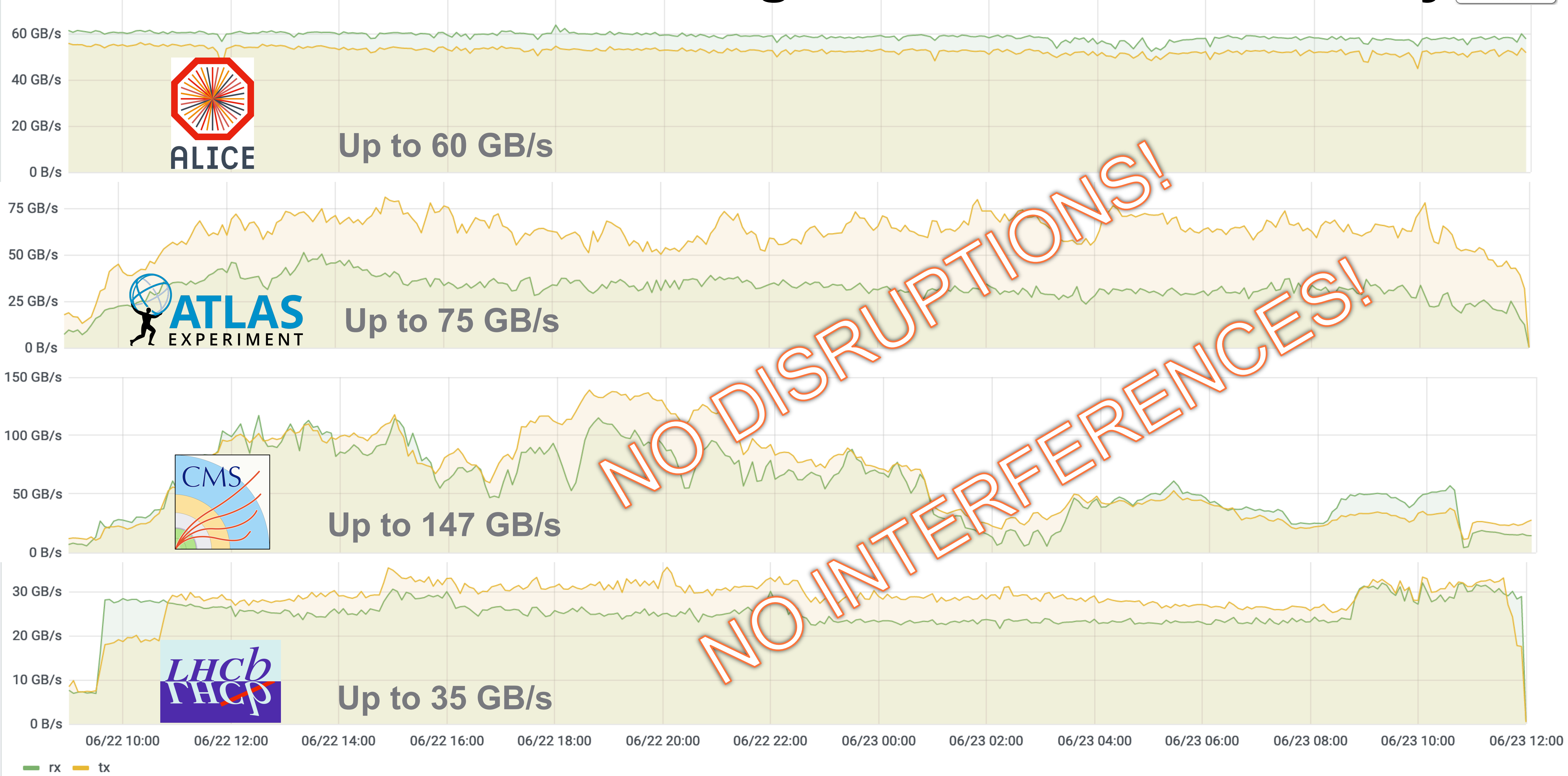
  Introduce retries and adaptive timeouts. This is already done by the other experiments.

- **Improvement from CMS Tier0 side:**

  Tag the traffic for easy impact detection. This is already done by ATLAS.

Combined Data Challenge: EOS network activity

Target Achieved.



Up to 60 GB/s

Up to 75 GB/s

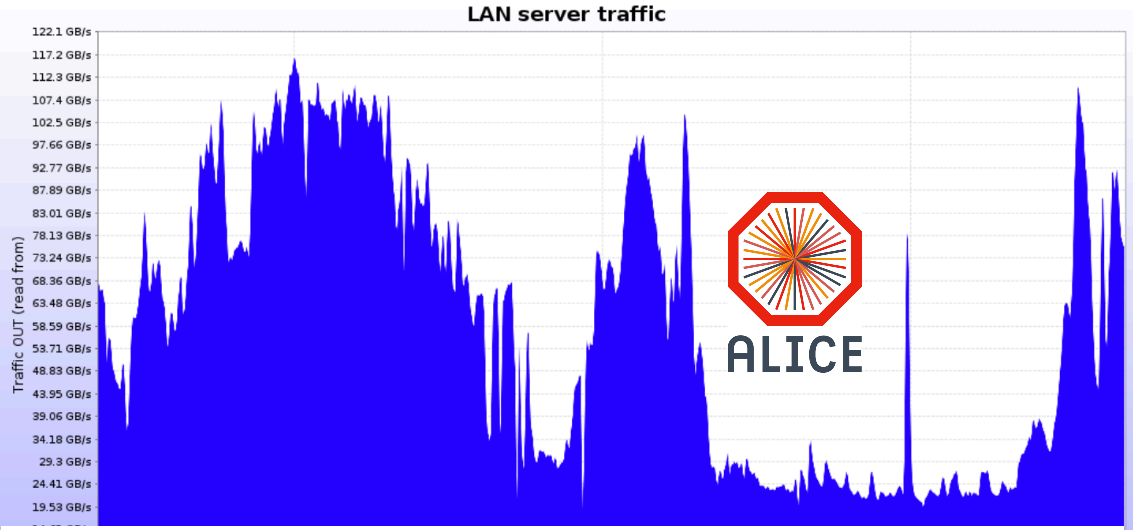
Up to 147 GB/s

Up to 35 GB/s

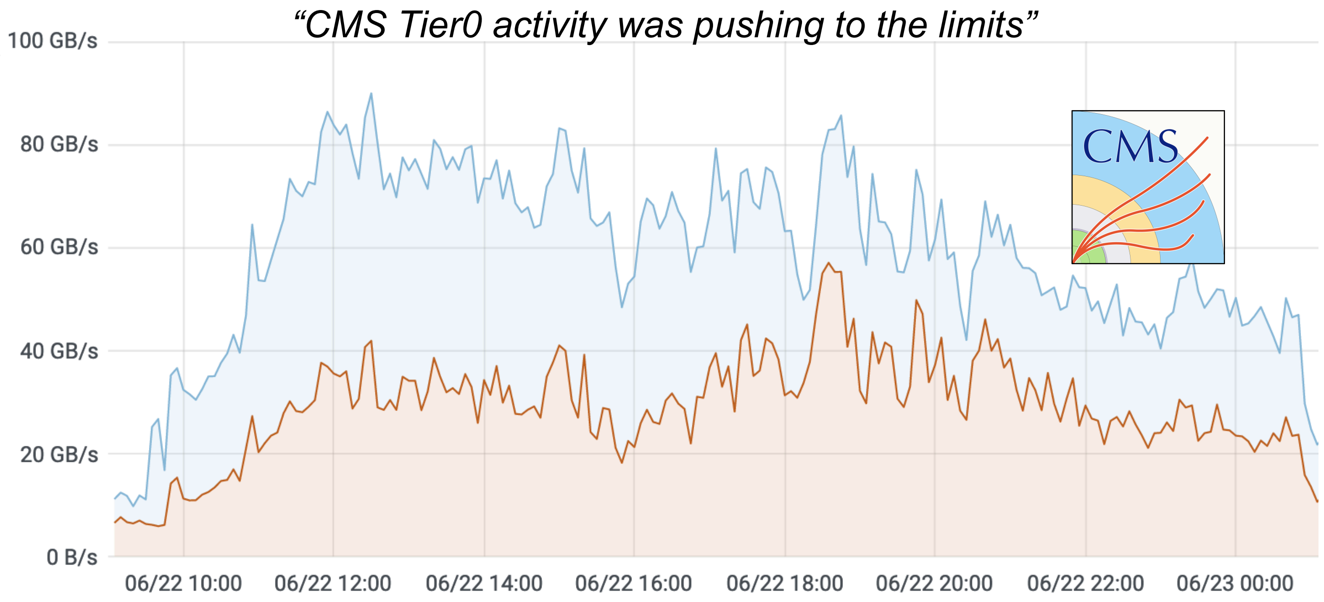
NO DISRUPTIONS!

NO INTERFERENCES!

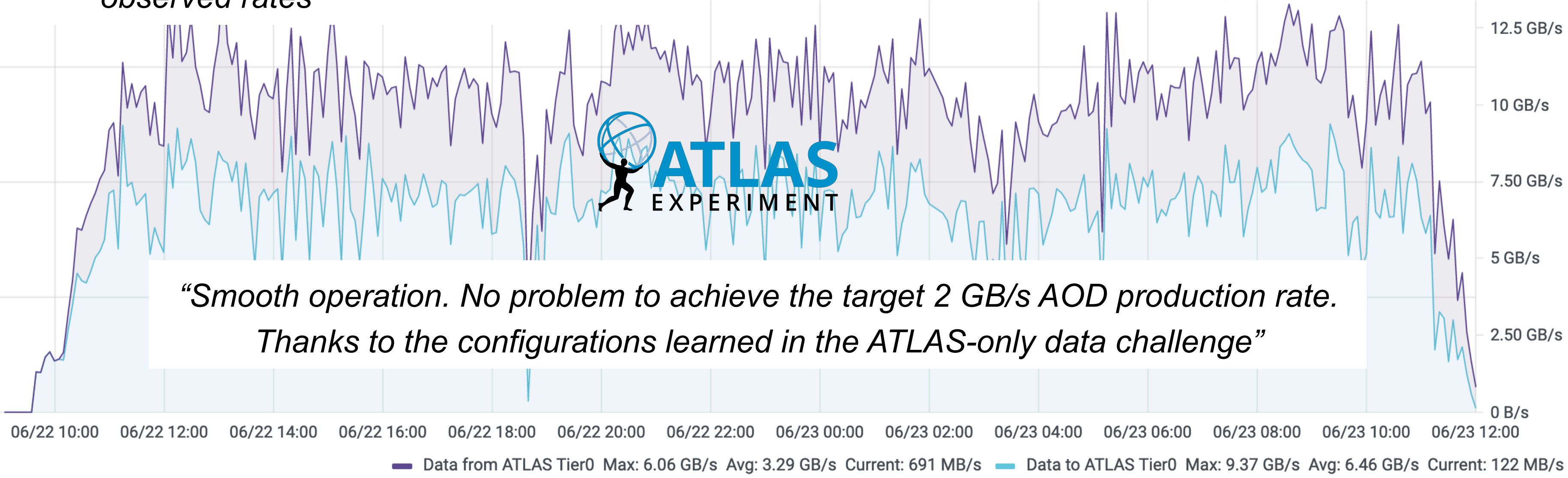
Combined Data Challenge: Experiment Tier0 activity



“There were no effects observed on the analysis workflow and no effect of the EOSALICE read/write rates on the EOSALICEO2 observed rates”



“CMS Tier0 activity was pushing to the limits”

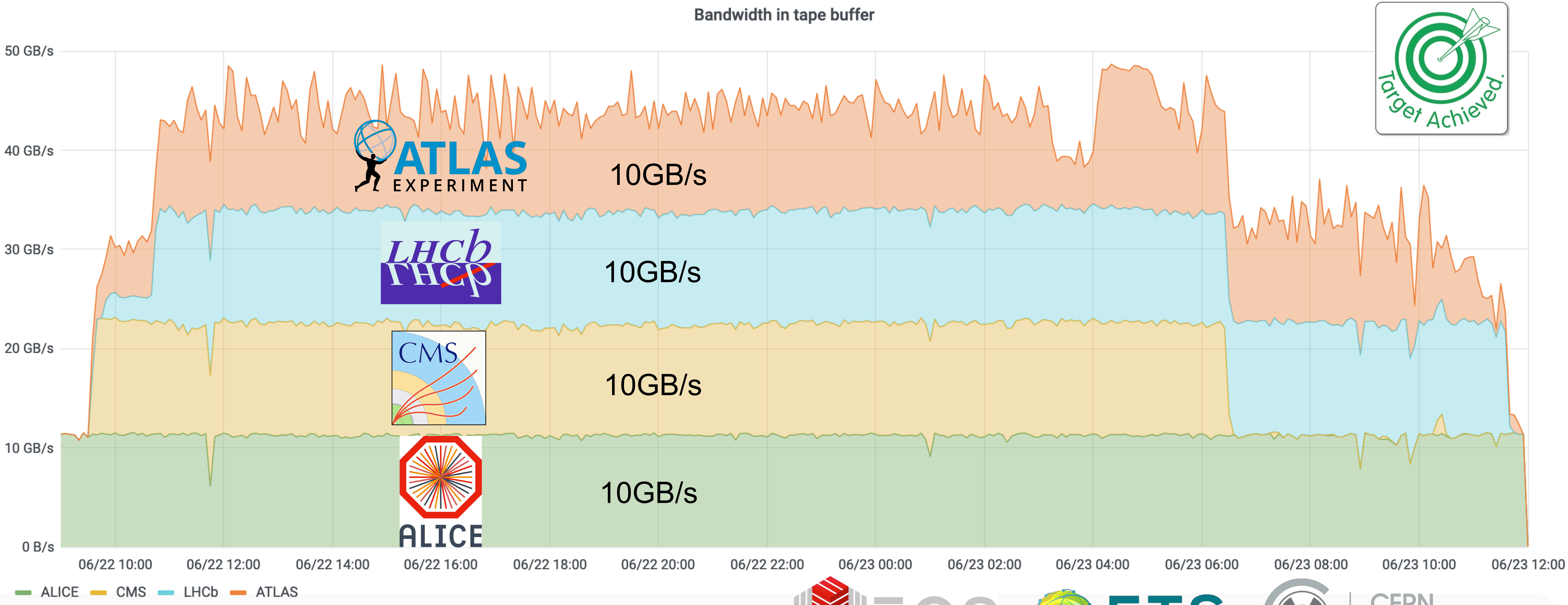


“Smooth operation. No problem to achieve the target 2 GB/s AOD production rate. Thanks to the configurations learned in the ATLAS-only data challenge”

Target Achieved.

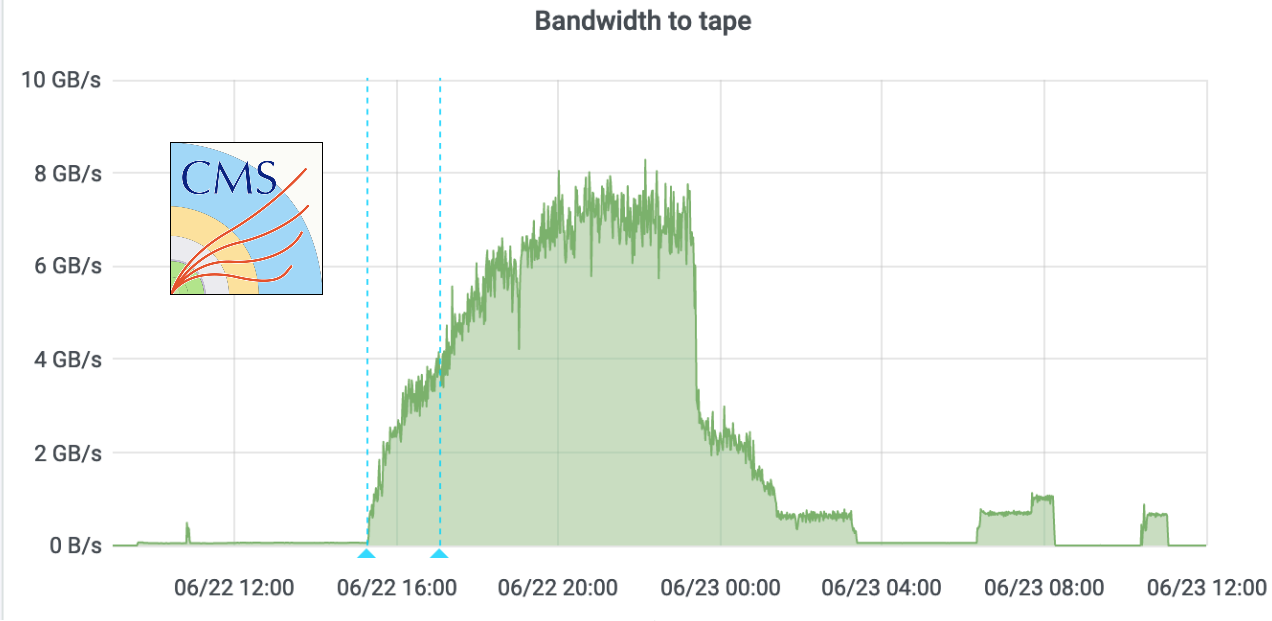
CERN

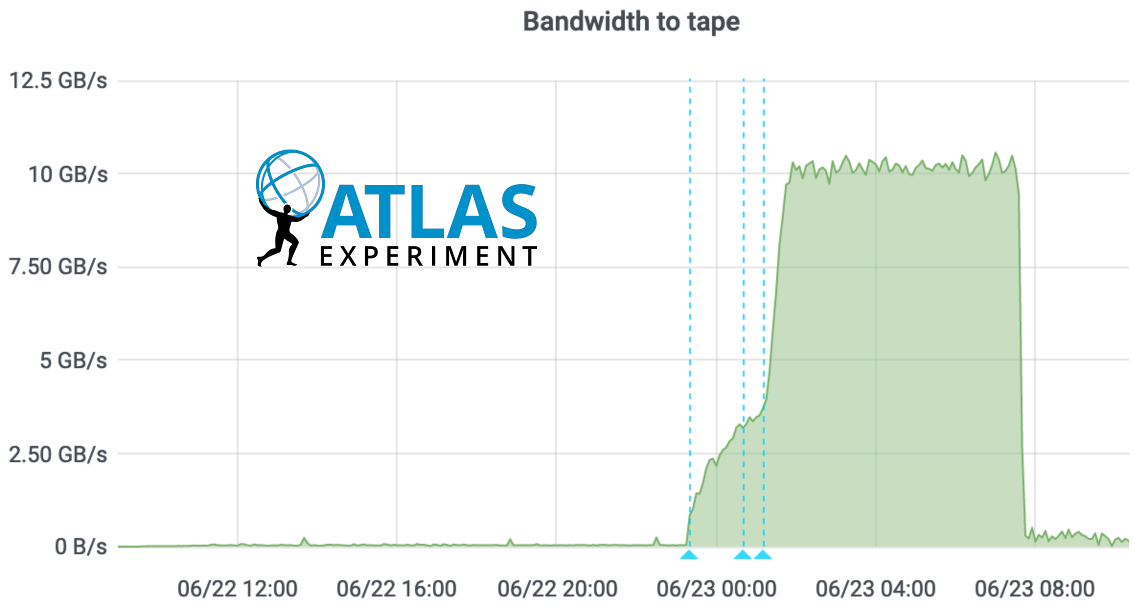
Combined Data Challenge: Data export to EOSCTA



Combined Data Challenge: Data export to CTA

Target Achieved.





Tape momentum at 16:00 getting up to 8GB/s according to the available tape hardware

Tape momentum at 00:00 was achieving the target rate 10GB/s

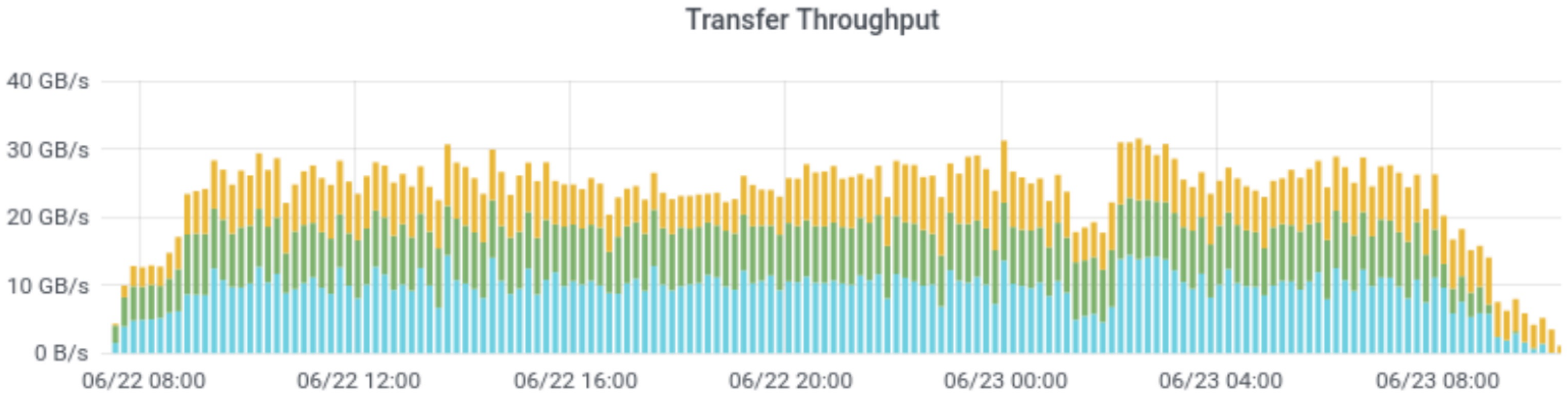
* Simultaneous tape data challenge still to be done

EOS FTS File Transfer Service CERN Tape Archive

Combined Data Challenge: Data Export to Tier1s

Successful export rates

* As during the AtlasDataChallenge test, the DBoD backup at 3-4AM intervened with the FTS service. However, due to the mitigation measures, this time the impact was greatly reduced.



Target Achieved.

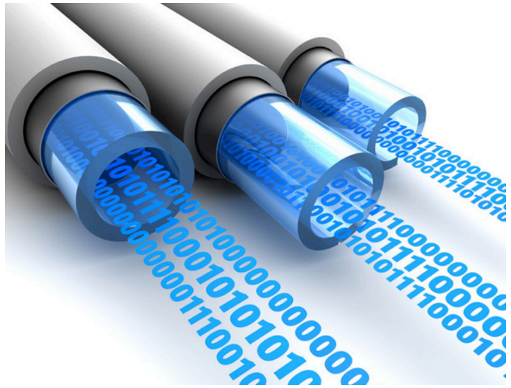
	min	max	avg	current
🟦 T0 Tape	0 B/s	14.4 GB/s	9.29 GB/s	9.53 MB/s
🟩 SFO to EOS export	0 B/s	9.03 GB/s	7.37 GB/s	0 B/s
🟨 T0 Export	0 B/s	9.84 GB/s	6.65 GB/s	1.07 GB/s

Conclusions

Great experiment participation and collaboration with all parties

Participation of the 4 LHC experiments:

All responsible for the different workflow components were having a direct collaboration with the storage team

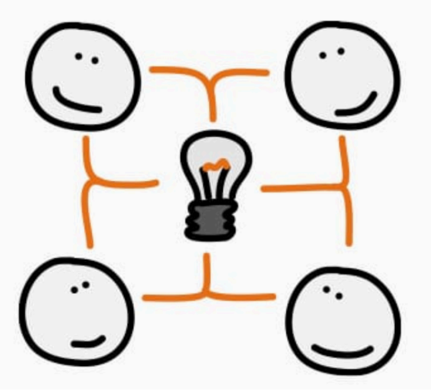


Non throughput disruptions between experiments

No network disruption with combined RUN

Well balanced with no overload

Thanks to the network team



Objectives achieved and some lessons learned

Scheduling improvements on EOS

Database optimization on FTS

Tagged traffic, retries and adaptive timeouts from CMS

RUN 3 Data Taking Commissioning

Planning and Communication

LHC workflows for RUN 3

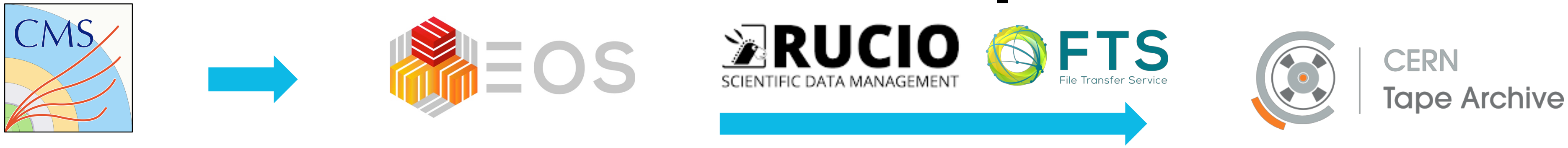
Testing individual components

Individual data challenge

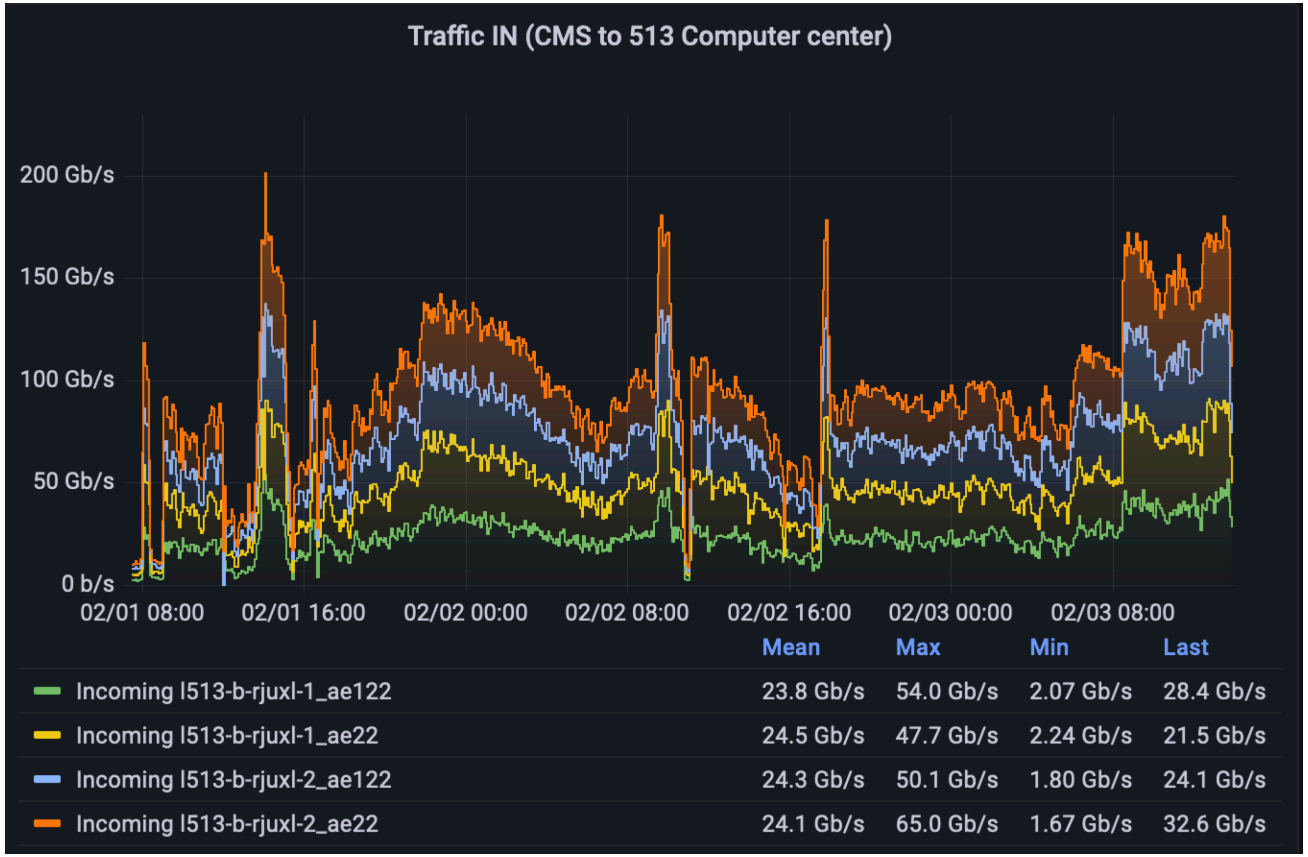
Combined data challenge

Follow up tests

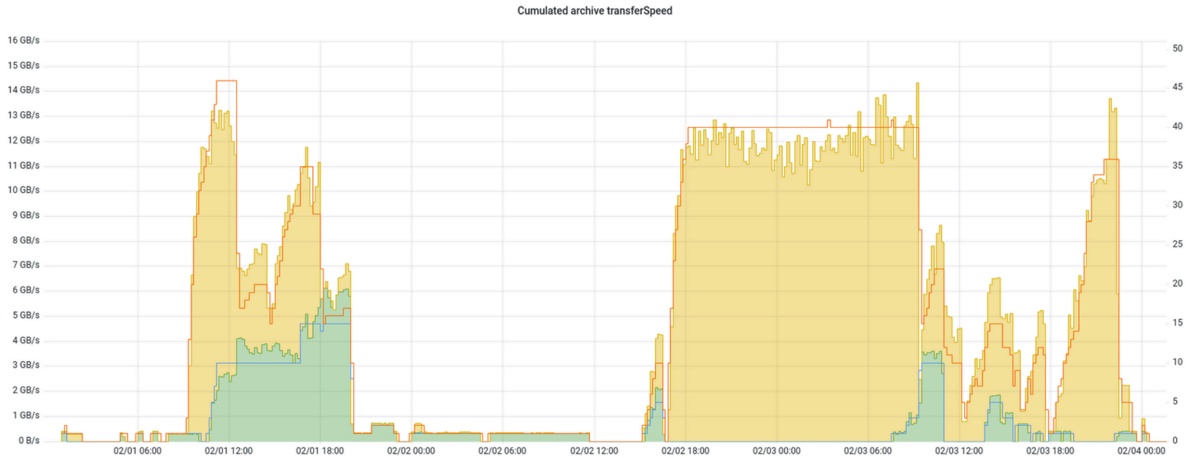
Follow up tests



Target Achieved.


Traffic IN (CMS to 513 Computer center)

	Mean	Max	Min	Last
🟩 Incoming l513-b-rjuxl-1_ae122	23.8 Gb/s	54.0 Gb/s	2.07 Gb/s	28.4 Gb/s
🟨 Incoming l513-b-rjuxl-1_ae22	24.5 Gb/s	47.7 Gb/s	2.24 Gb/s	21.5 Gb/s
🟦 Incoming l513-b-rjuxl-2_ae122	24.3 Gb/s	50.1 Gb/s	1.80 Gb/s	24.1 Gb/s
🟧 Incoming l513-b-rjuxl-2_ae22	24.1 Gb/s	65.0 Gb/s	1.67 Gb/s	32.6 Gb/s


Cumulated archive transferSpeed

Goal: Emulate at 2 rates:

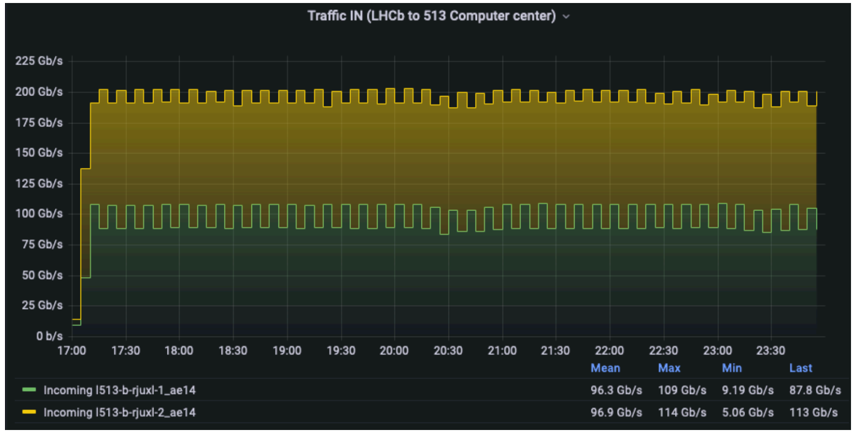
- 13GB/s (normal pp run) with ~23GB filesize
- 17GB/s (HI runs) with ~29GB filesize
- CTA nominal rate: 14GB/s

- Date: 1-3/02/2022

New EOS configurations tested:

- Flat scheduling

Follow up tests

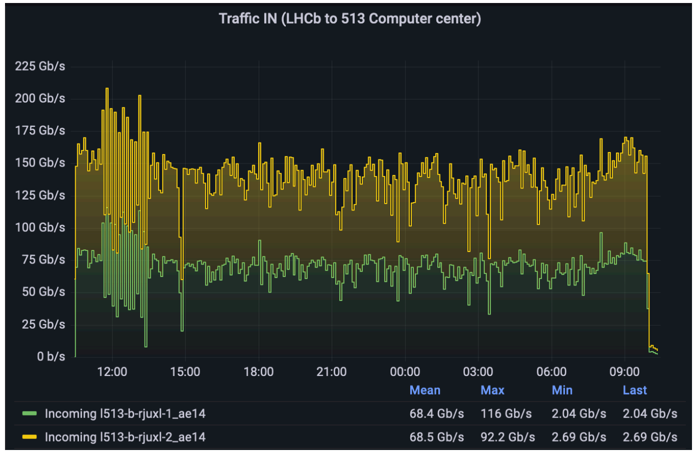


	Mean	Max	Min	Last
🟩 Incoming l513-b-rjuxl-1_ae14	96.3 Gb/s	109 Gb/s	9.19 Gb/s	87.8 Gb/s
🟨 Incoming l513-b-rjuxl-2_ae14	96.9 Gb/s	114 Gb/s	5.06 Gb/s	113 Gb/s

Goal: Pushing to the limits (max 20GB/s throughput)
Mean throughput obtained (P8 to EOS): 24GB/s
File size: 10GB
Date: 24/02/2022

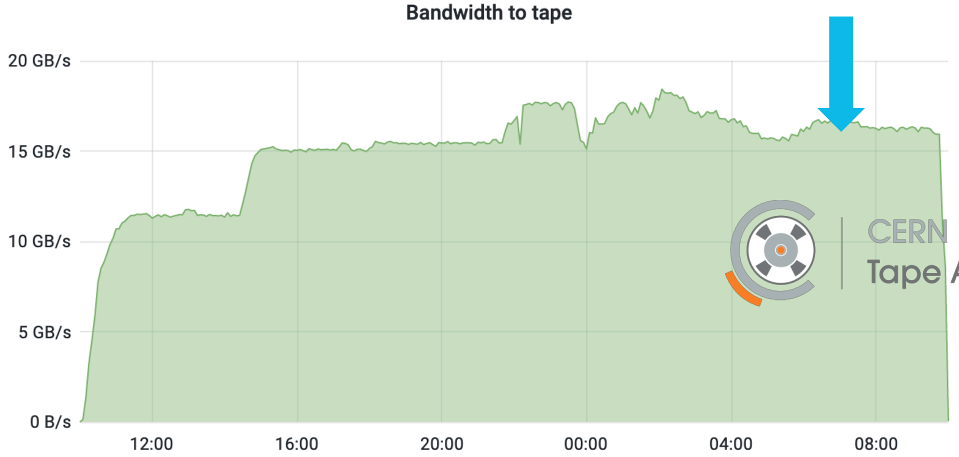
New EOS configurations tested:

- Async write replicas
- Flat scheduling



	Mean	Max	Min	Last
🟩 Incoming l513-b-rjuxl-1_ae14	68.4 Gb/s	116 Gb/s	2.04 Gb/s	2.04 Gb/s
🟨 Incoming l513-b-rjuxl-2_ae14	68.5 Gb/s	92.2 Gb/s	2.69 Gb/s	2.69 Gb/s

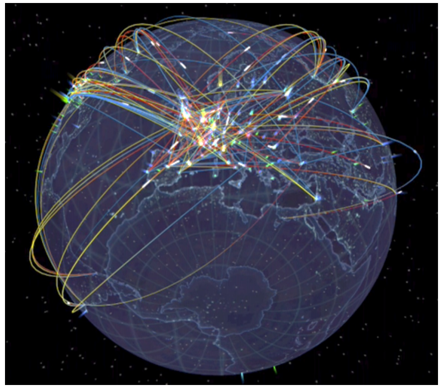
Goal: Sustained throughput
Mean throughput (P8 to EOS): 17GB/s
File size: 75% of 5GB files, 25% of 10GB files
Date: 1-2/03/2022

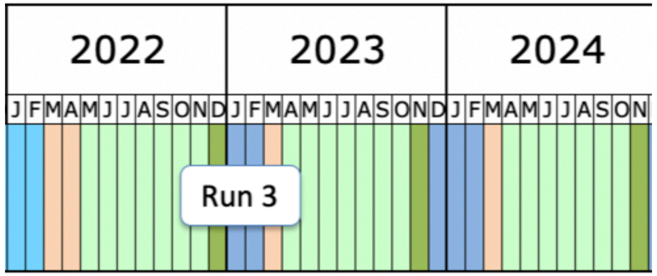


EOS → DIRAC FTS File Transfer Service

CERN Tape Archive

Follow up tests




2022
2023
2024
J F M A M J J A S O N D J F M A M J J A S O N D J F M A M J J A S O N
Run 3

Evaluate storage and experiments improvements (decoupled tests)

Scheduling improvements on EOS

Database optimization on FTS

Tagged traffic, retries and adaptive timeouts from CMS

New hardware procurement from IT and experiments

Combined tape data challenge

Evaluate the tape infrastructures for T0 and T1s according with experiments' expectations

Final Commissioning with RUN 3 Hardware

Evaluate real RUN 3 workflows with the final hardware from the experiments, T0 and T1s

Conclusion

Storage and transfer services successfully delivered the required performance to accommodate experiments' demand for RUN 3.

Identified areas to enhance the reliability of the storage and transfer systems.

We are confident about our readiness for RUN3.

“Thanks to all the people involved in these tests because without them wouldn’t have been possible”