WLCG Data Challenge 2024
Focus su Belle II

Dr. Silvio Pardi

Workshop sul Calcolo nell'I.N.F.N.

20 Maggio 2024

Worldwide LHC Computing Grid (WLCG)

The Worldwide LHC Computing Grid (**WLCG**) is a global collaboration of around 170 computing centres in more than 40 countries, linking up national and international grid infrastructures.

The mission of the WLCG project is to provide global computing resources to store, distribute and analyse the data expected every year of operations from the Large Hadron Collider (LHC) at CERN on the Franco-Swiss border.

Other experiments: Belle II, DUNE

The Worldwide LHC Computing Grid is partnered with EGI Fundating, OSG (Open Science Grid), and NeIC (Nordic e-Infrastructure Collaboration).

https://wlcg.web.cern.ch/

Introduction: LHC and its High Luminosity upgrade

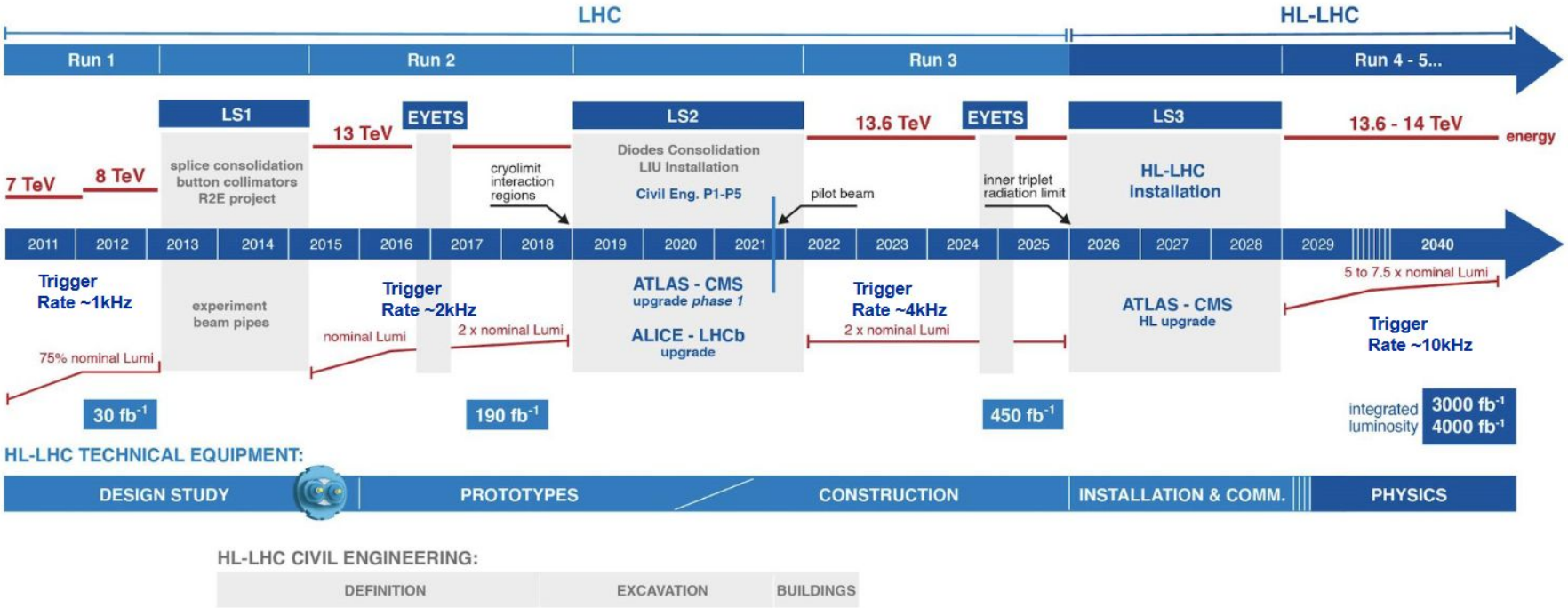


Figure adapted from:
Zerlauth, Markus & Bruning, Oliver. (2024). Status and prospects of the HL-LHC project. DOI; 615. 10.22323/1.449.0615.

WLCG Data Challenge 2024

- WLCG has mandated to execute data challenges (DC) for HL-LHC
  - Demonstrate readiness for expected HL-LHC data rates by a series of challenges
  - Increasing volume/rates
  - Increasing complexity (e.g. additional technology)
  - A data challenge roughly every two years
- DOMA is the coordination and execution platform
  - Data Organization Management & Access
    - Forum across all LHC experiments to address **technical** challenges
  - DC coordination across the LHC experiments and beyond
    - Suited dates
    - Reasonable targets
    - Functionalities
  - Help in orchestration
- No pressure on sites to increase their capacity
  - But can we improve the existing infrastructure?

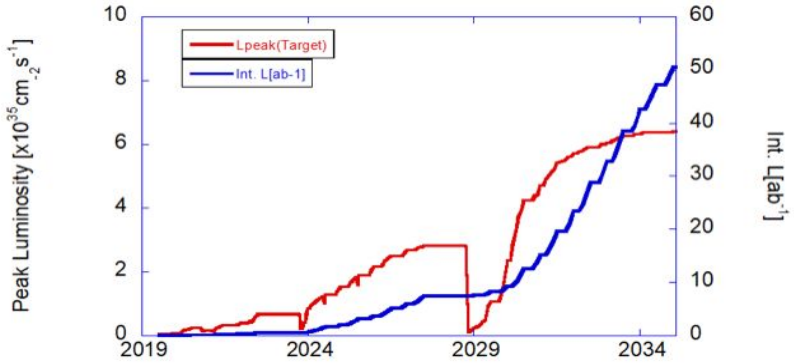
Year	% of HL-LHC
2021	10
2024	25
2026?	50
2028?	100

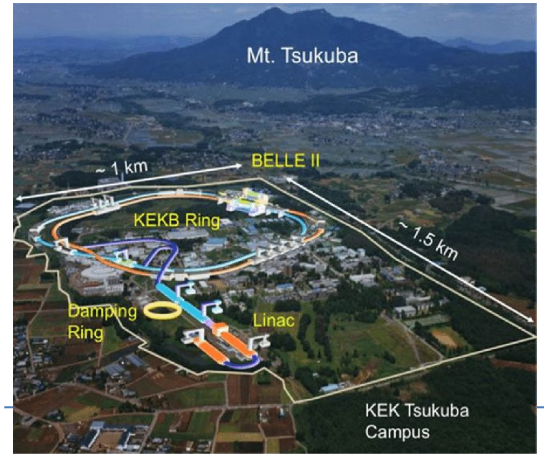
Modelling the rates for HL-LHC (T0 export)

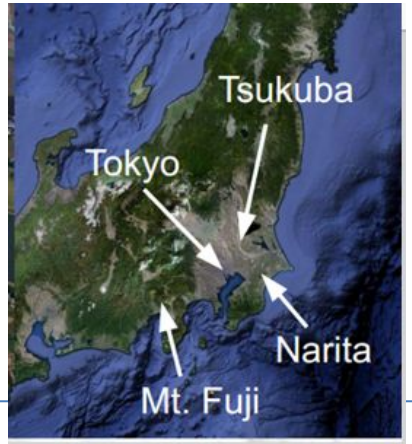
- **ATLAS & CMS T0 export (T0 to T1s)**
  - 350PB RAW per experiment, per year, taken and distributed during typical LHC uptime of 7M seconds
  - => 50GB/s or 400Gbps
  - Plus 100Gbps estimated for prompt, derived data
  - 1Tbps for CMS and ATLAS combined
- **ALICE & LHCb T0 Export**
  - 100 Gbps per experiment estimated from Run-3 rates
- **Network needs to be bigger than the average, estimated rates:**
  - Factor of 2 for bursts
  - Another factor of 2 for overprovisioning
- **Minimal Model:** Sum (ATLAS,ALICE,CMS,LHCb)*2(for bursts)*2(overprovisioning) = **4.8Tbps** per HL-LH
- **Flexible Model:** Duplication the Minimal Model up to **9.6Tbps** adding the other traffics:: T1s to T2s and among T1s

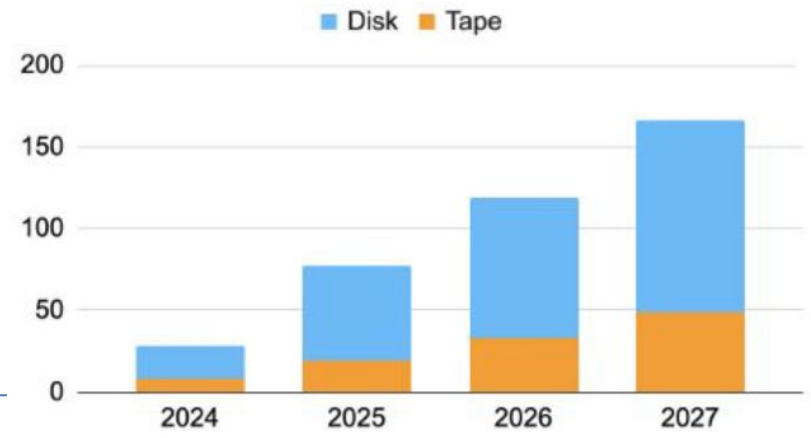
GOAL of WCLG DC24 is to reach the 25% of the traffic

Belle II Experiment









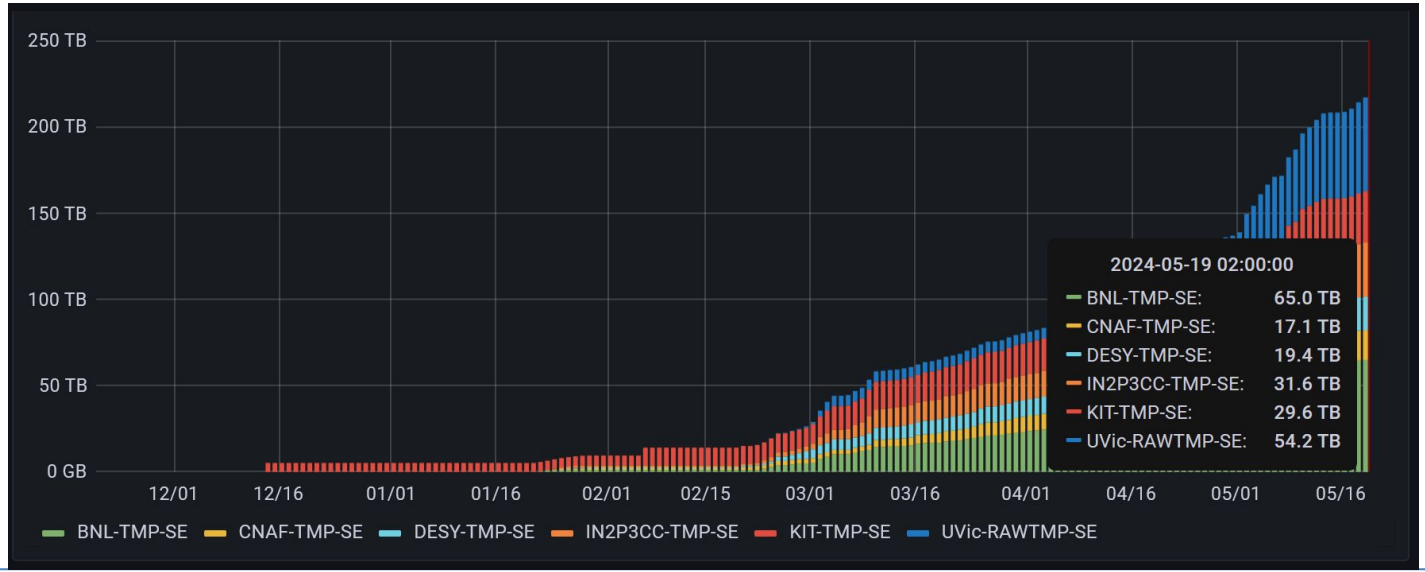
Belle II Status

Data taking started in 2019.

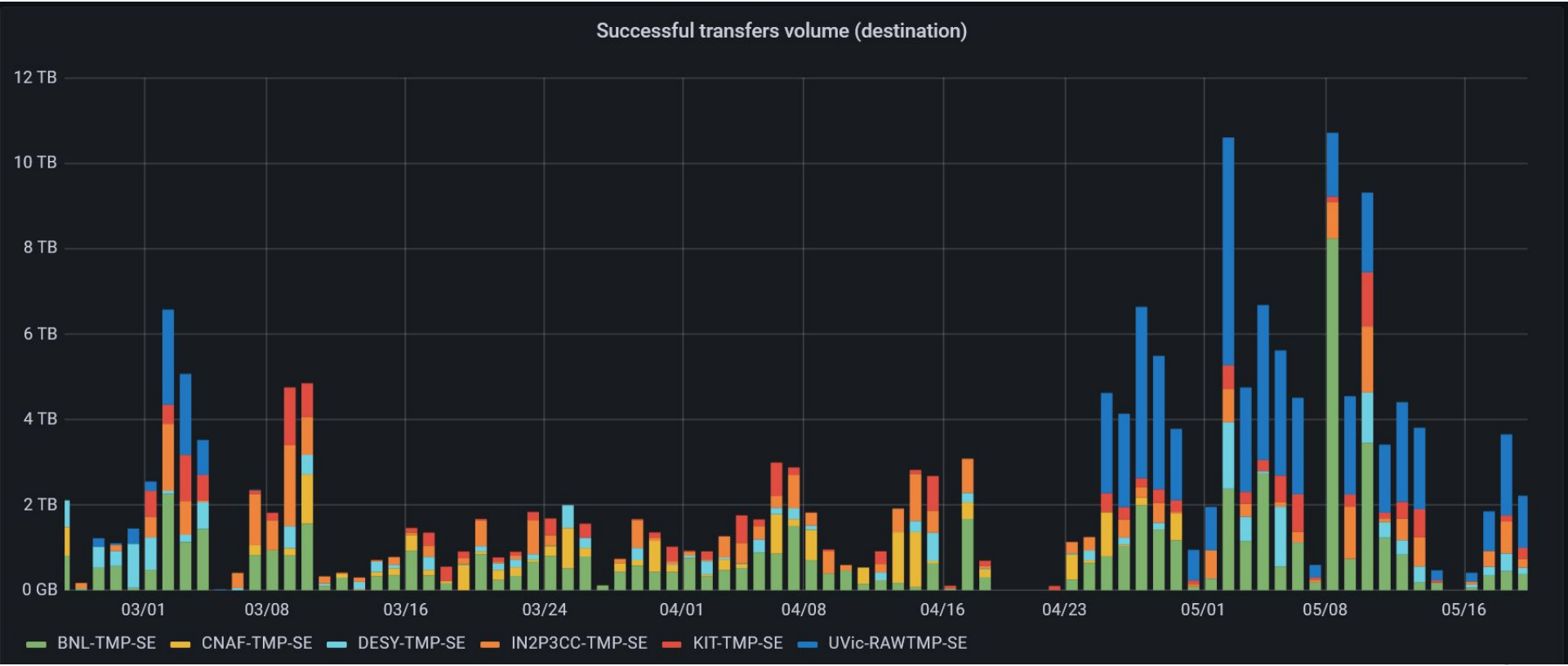
In July 2022 we started the Long Shutdown 1

Data taking restarted early 2024, first collision 20 February 2024.

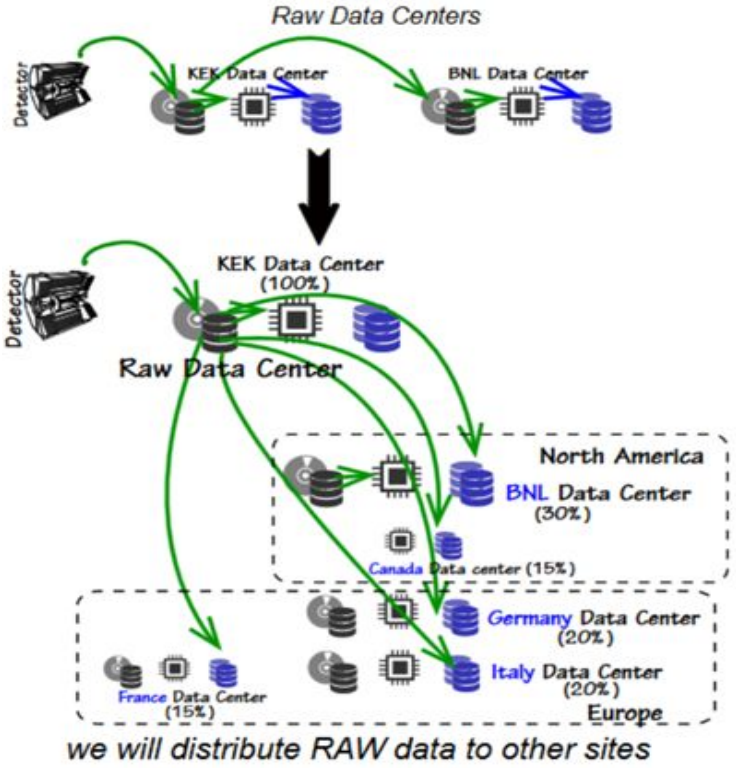
Restarted Copy of RAW Data from KEK to RAW Data Centers



RAW Data Export from KEK vs RAW DC



Belle II RAW Data distribution

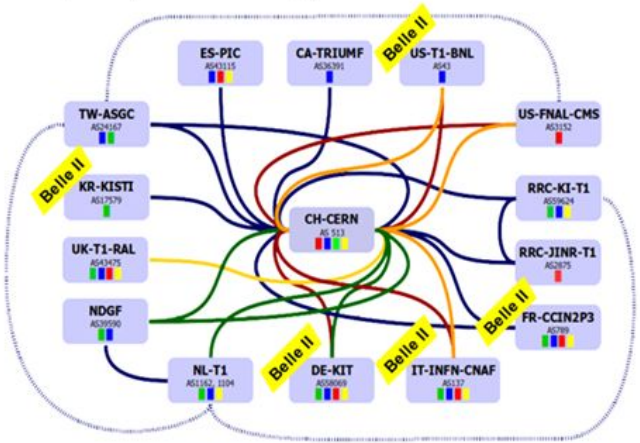


SITE	2019-2020	2021-2024
BNL - USA	100%	30%
CNAF - Italy	0%	20%
DESY - Germany	0%	10%
KIT - Germany	0%	10%
IN2P3CC - France	0%	15%
UVIC - Canada	0%	15%

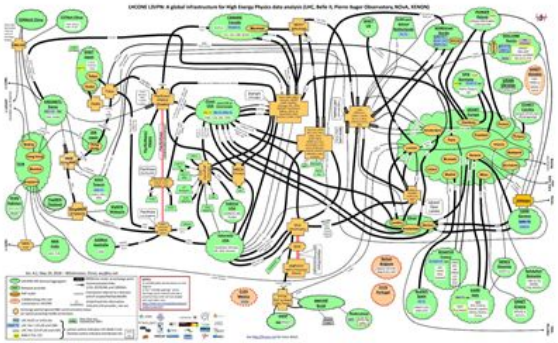
Belle II Network

100G Global Ring via SINET

LHCOPN Optical infrastructure that can be used without jeopardizing resources



LHCONE L3 VPN Connecting all the major Data Centres



Belle II in WLCG Data Challenge 2024

What should be exercised during DC24:

Technology that can be stressed: Network, DDM, FTS, Storages, Monitoring System, Protocols.

Main goal: Emulate data transfer conditions in a Belle II future scenario

Our current estimation we should produce 40 TB per day.

Transfers from KEK to RAW Data Centers according to our distribution schema (30%BNL, 20%CNAF, 15% IN2P3CC, 15%UVic, 10%DESY, 10%KIT)

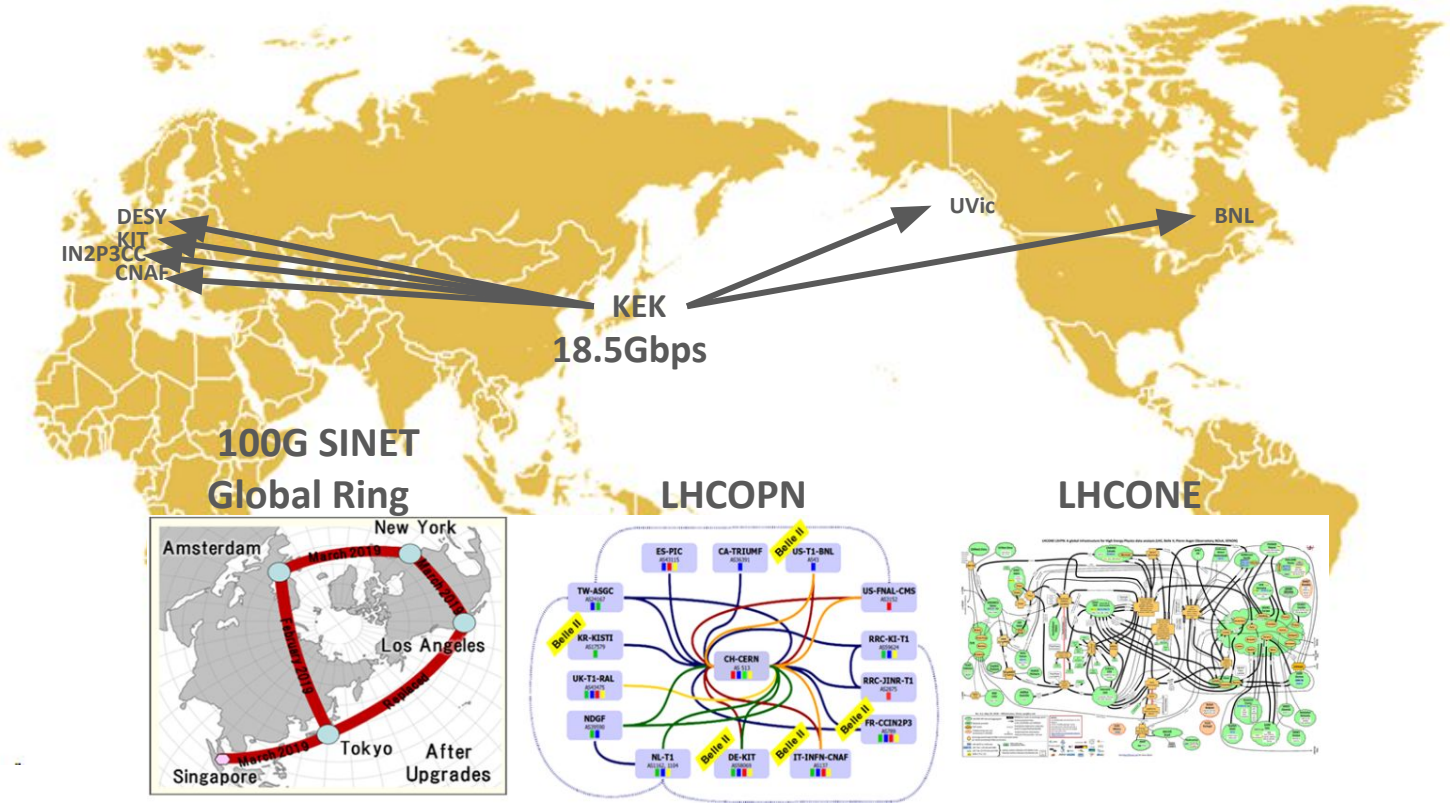
Considering that the average speed needed to transfer 40TB/day is 3.7Gbit/s in outbound at KEK vs all the Raw Data Centers.

- Min - The target speed to achieve is 3x3.7Gbit/s = **11.1 Gbit/s**
- Max - The target speed to achieve is 5x3.7Gbit/s = **18.5 Gbit/s**

Belle II Data Challenge 2024

				Minimal x3		Maximal x5	
Storage Name	Site	Country	#5G Files	Ingress (Gbps)	Egress (Gbps)	Ingress (Gbps)	Egress (Gbps)
KEK-TMP-SE	KEK	JP	8000	0,0	11,1	0,0	18,5
BNL-TMP-SE	BNL	US	2400	3,3	0	5,6	0
CNAF-TMP-SE	CNAF	IT	1600	2,2	0	3,7	0
DESY-TMP-SE	DESY	DE	800	1,1	0	1,9	0
KIT-TMP-SE	KIT	DE	800	1,1	0	1,9	0
IN2P3CC-TMP-SE	IN2P3CC	FR	1200	1,7	0	2,8	0
UIVc-RAWTMP-SE	UIVc	CA	1200	1,7	0	2,8	0
Napoli-TMP-SE	Napoli	IT	TBD	TBD	TBD	TBD	TBD
SIGNET-TMP-SE	SIGNET	SL	TBD	TBD	TBD	TBD	TBD

Belle II DC Test



Testing script

We started by a predefined dataset stored at KEK and reused multiple times for transfers.

All transfers have been done using DAVS protocol and the RUCIO+FTS production infrastructure.

Test automation done via a Python script that it operates on a cyclical base as follow:

At each cycle, the script checks for existing replication rules associated with specific datasets.

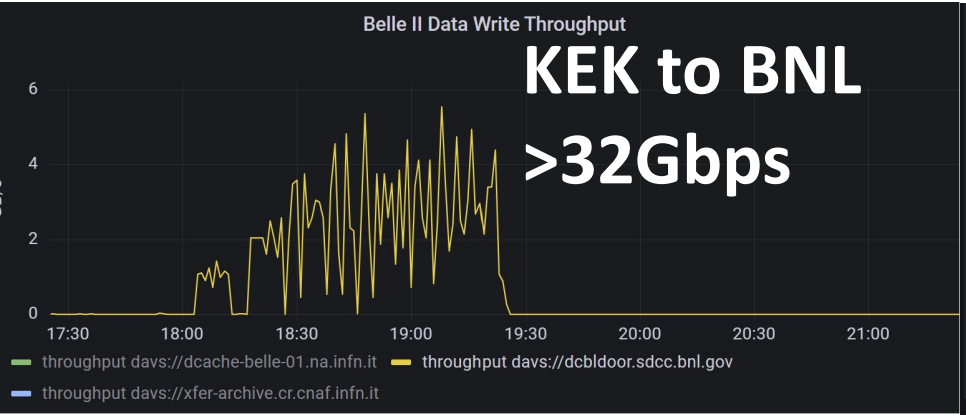
If no rule is found for a particular site, it verifies the presence of data replicas at that site.

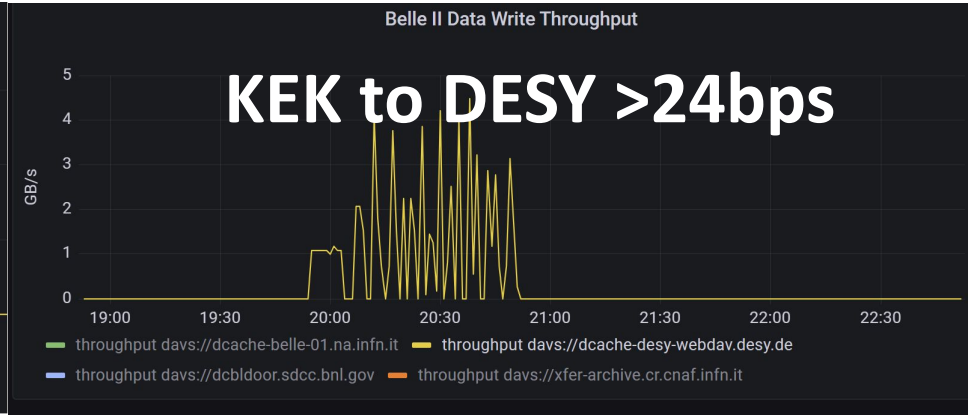
If replicas are absent, a new replication rule is created.

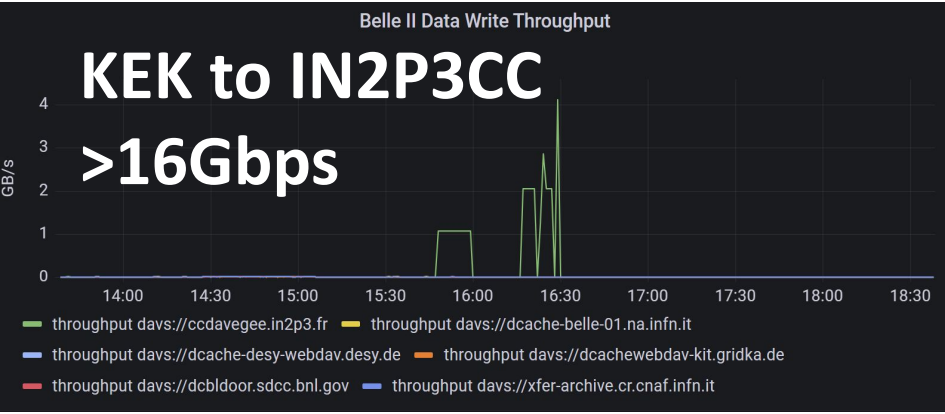
When a replication rule exists but the replication is completed, the script triggers a deletion instruction.

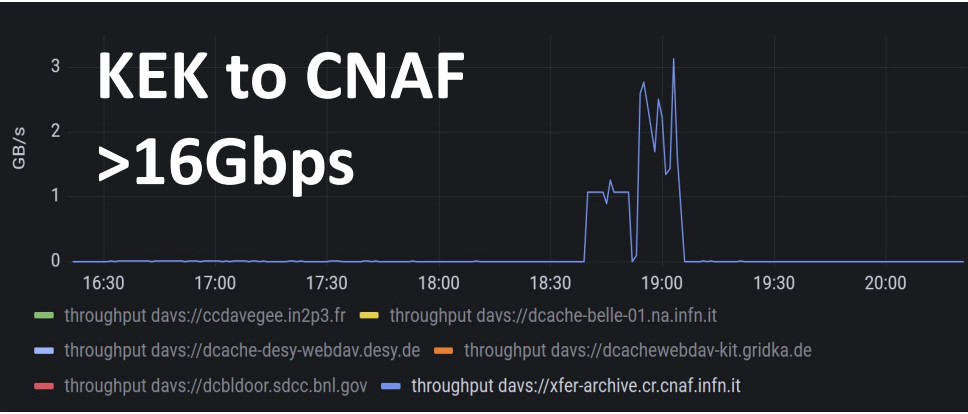
N.B. Only GSI authentication has been used

FTS pre-test in January









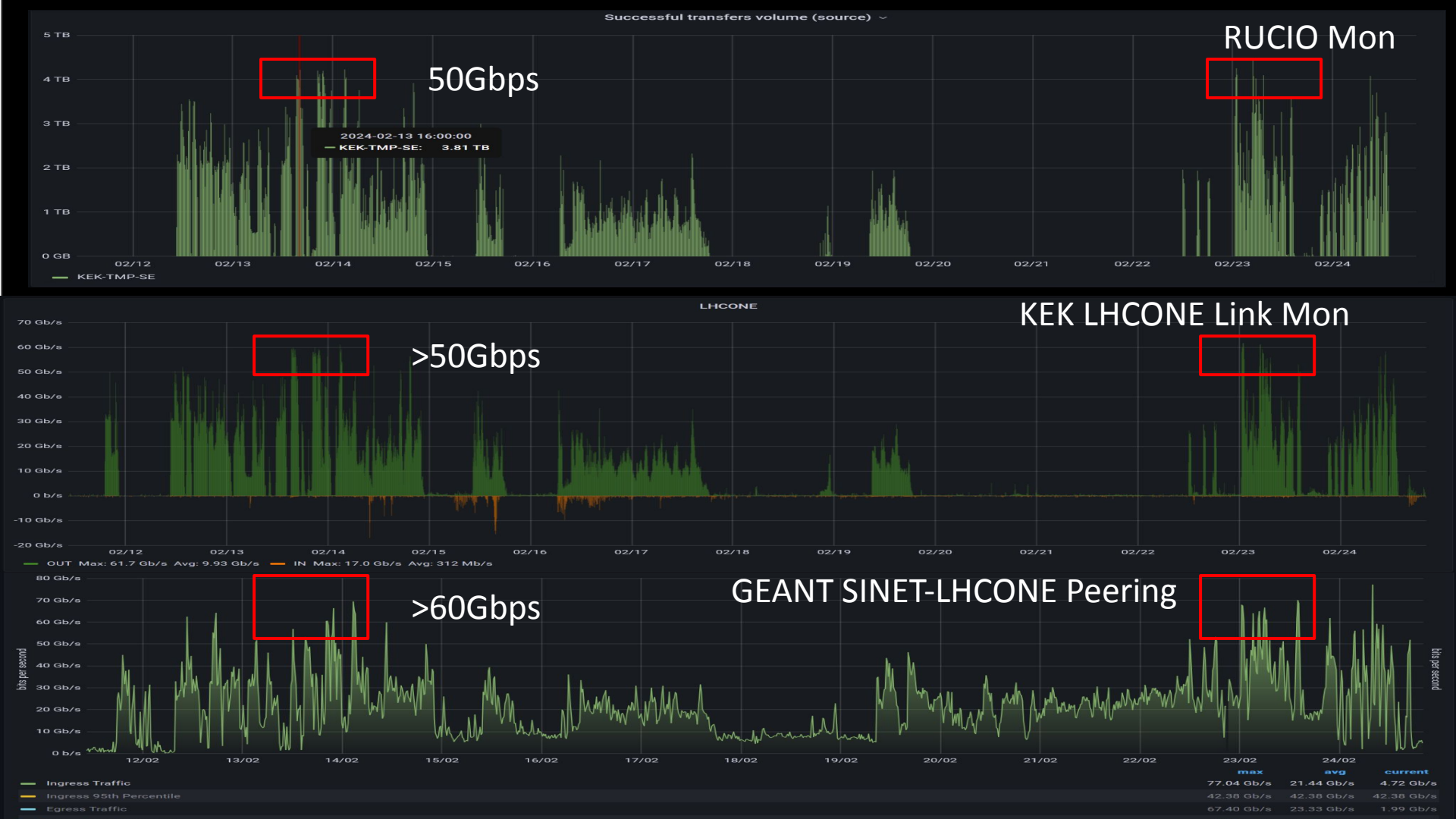
WLCG Data Challenge Program

	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
	12/02/2024	13/02/2024	14/02/2024	15/02/2024	16/02/2024	17/02/2024	18/02/2024
ALICE	T0 → T1	T0 → T1	T0 → T1	T0 → T1	T0 → T1	T0 → T1	T0 → T1
ATLAS	T0 → T1	T0 → T1	T0 → T1 → T2	T0 → T1 → T2	T0 → T1 → T2	T0 → T1 → T2	T0 → T1 → T2
CMS	T0 → T1	T0 → T1	T0 → T1 → T2	T1 → T2	T1 ↔ T2	T1 ↔ T2	T1 ↔ T2
LHCb		T0 → T1	T0 → T1	T0 → T1	T0 → T1	T0 → T1	T0 → T1
DUNE	T0 → T1 → T2	T0 → T1 → T2	T0 → T1 → T2	T0 → T1 → T2	T0 → T1 → T2	T0 → T1 → T2	T0 → T1 → T2
Belle II	T0 → T1	T0 → T1	T0 → T1	T0 → T1	T0 → T1	T0 → T1	T0 → T1
SUMMARY							
T0 exports minimal rates (ALICE+ATLAS+LHCB+CMS)	529.7 Gbps	650.3 Gbps	650.3 Gbps	650.3 Gbps	650.3 Gbps	650.3 Gbps	650.3 Gbps
T0 exports (DUNE + Belle II)	18.5 Gbps (belleII)	18.5 Gbps (belleII)	18.5 Gbps (belleII)	18.5 Gbps (belleII)	18.5 Gbps (belleII)	18.5 Gbps (belleII)	18.5 Gbps (belleII)

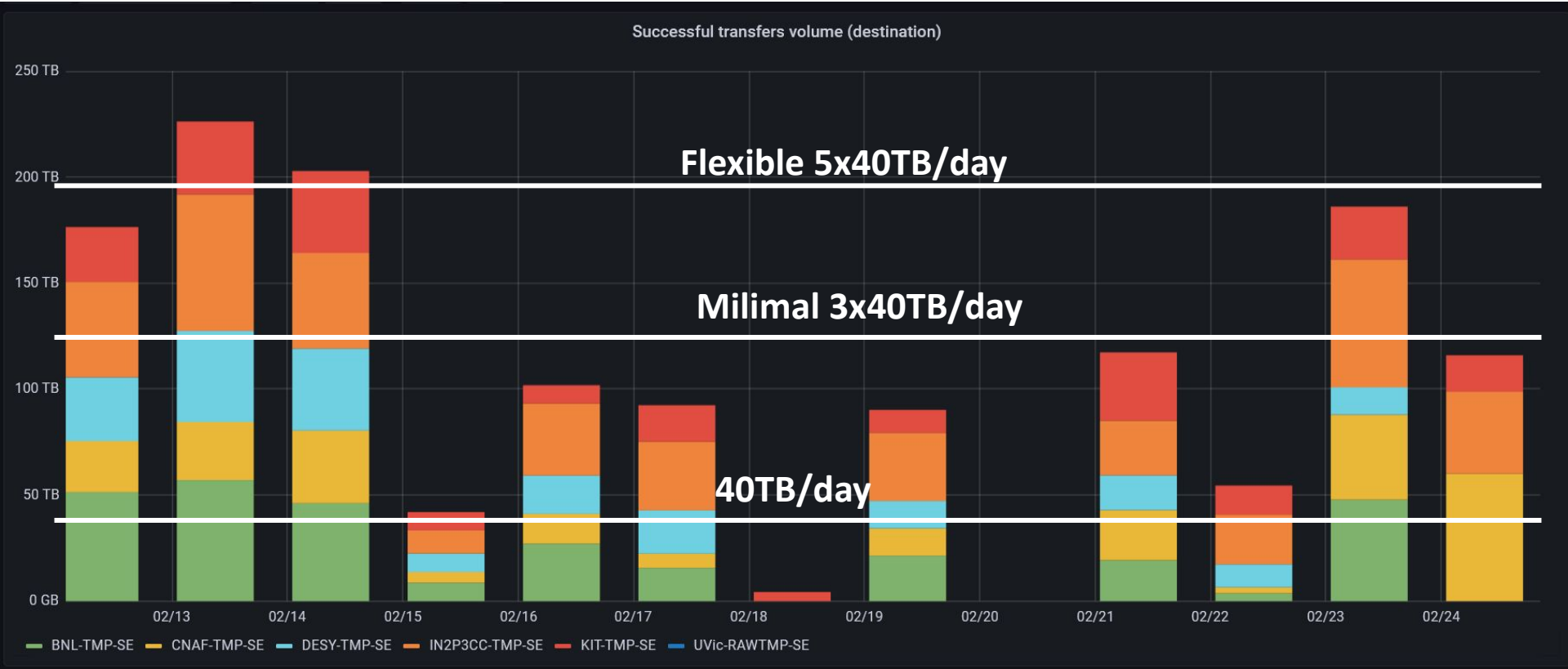
	Monday	Tuesday	Wednesday	Thursday	Friday	
	19/02/2024	20/02/2024	21/02/2024	22/02/2024	23/02/2024	yellow: "reduced minimal" (only T0 export)
ALICE	T0 → T1	T0 → T1	T0 → T1	T0 → T1	T0 → T1	blue: minimal scenario
ATLAS	T0 ↔ T1 ↔ T2	T0 ↔ T1 ↔ T2	T0 ↔ T1 ↔ T2	T0 ↔ T1 ↔ T2	T0 ↔ T1 ↔ T2	red: flexible scenario
CMS	AAA T1 → T2	T0 → T1 ↔ T2	T0 → T1 ↔ T2	T0 → T1 ↔ T2	T0 → T1 ↔ T2	
LHCb	T0 → T1	T1 Tape Recall	T1 Tape Recall	T1 Tape Recall	T1 Tape Recall	
DUNE	T0 → T1 → T2	T0 → T1 → T2	T0 → T1 → T2	T0 → T1 → T2	T0 → T1 → T2	
Belle II	T0 → T1	T0 → T1	T0 → T1	T0 → T1	T0 → T1	T0 == SURF , T1 == FNAL, T2 == Storage sites
SUMMARY						
T0 exports high rates (ALICE+ATLAS+LHCB+CMS)	449.56 Gbps	895.56 Gbps	895.56 Gbps	895.56 Gbps	895.56 Gbps	

Belle II DC24 Activities

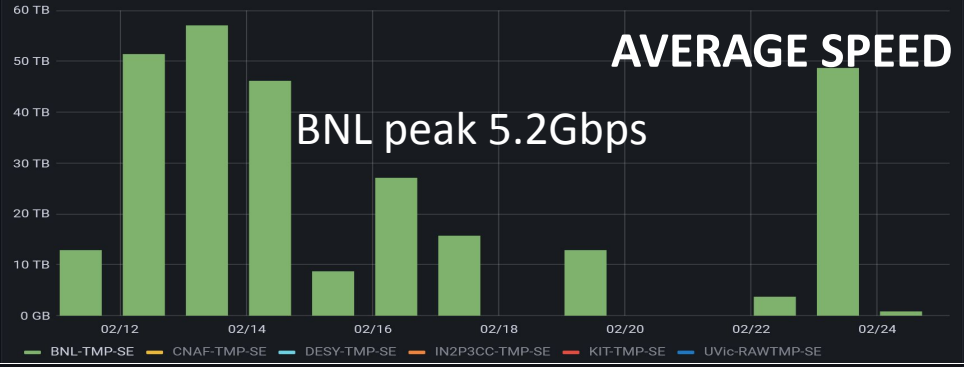
	DATE	Test	TOT	Peak (1h)	Average
1	12/02/2024 9:00 to 14/02/2004 23:00	KEK vs RAW DC (kek2-fts03 - v3.12.1)	606 TB/61h	50 Gbps	22,0 Gbps - Reached Max goal
2	15/02/2024 9:00 to 15/02/2024 16:00	KEK vs RAW DC (kek2-fts01 older)	39,9 TB/7h	25 Gbps	12,6 Gbps - Reached Min goal
3	16/02/2024 6:00 to 17/02/2024 19:00	KEK vs RAW DC (kek2-fts01)	194 TB/38h	24 Gbps	11,3 Gbps - Reached Min goal
4	19/02/2024 8:30 to 19/02/2024 21:30	KEK vs RAW DC + RAW DCs vs RAW DCs	80 TB/13h	27 Gbps	13,7 Gbps - Mixed traffic
5	21/02/2024 10:00 to 22/02/2024 9:00	RAW DCs vs RAW DCs (kek2-fts03)	141 TB/23h	46 Gbps	13,6 Gbps - Mixed traffic
6	23/02/2024 0:00 to 23/02/2024 14:00	KEK vs RAW DCs (kek2-fts03)	178 TB/15h	46 Gbps	26 Gbps - Reached Max goal

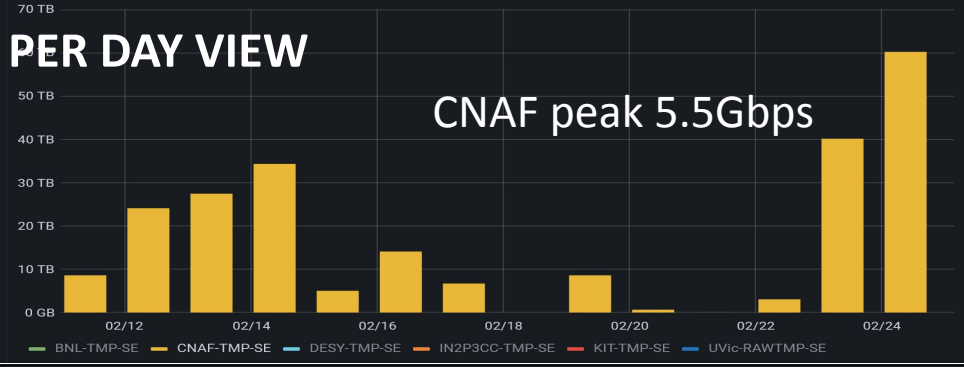

Successful transfers volume (source)
RUCIO Mon
50Gbps
2024-02-13 16:00:00
KEK-TMP-SE: 3.81 TB
5 TB
4 TB
3 TB
2 TB
1 TB
0 GB
02/12
02/13
02/14
02/15
02/16
02/17
02/18
02/19
02/20
02/21
02/22
02/23
02/24
KEK-TMP-SE
LHCONE
KEK LHCONE Link Mon
>50Gbps
70 Gb/s
60 Gb/s
50 Gb/s
40 Gb/s
30 Gb/s
20 Gb/s
10 Gb/s
0 b/s
-10 Gb/s
-20 Gb/s
OUT Max: 61.7 Gb/s Avg: 9.93 Gb/s
IN Max: 17.0 Gb/s Avg: 312 Mb/s
GEANT SINET-LHCONE Peering
>60Gbps
80 Gb/s
70 Gb/s
60 Gb/s
50 Gb/s
40 Gb/s
30 Gb/s
20 Gb/s
10 Gb/s
0 b/s
bits per second
12/02
13/02
14/02
15/02
16/02
17/02
18/02
19/02
20/02
21/02
22/02
23/02
24/02
max
avg
current
Ingress Traffic
77.04 Gb/s
21.44 Gb/s
4.72 Gb/s
Ingress 95th Percentile
42.38 Gb/s
42.38 Gb/s
42.38 Gb/s
Egress Traffic
67.40 Gb/s
23.33 Gb/s
1.99 Gb/s

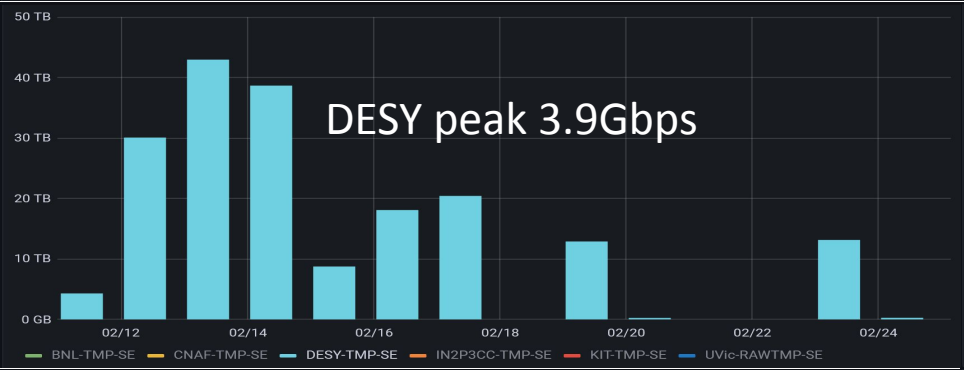
Traffic per Day View vs Goals

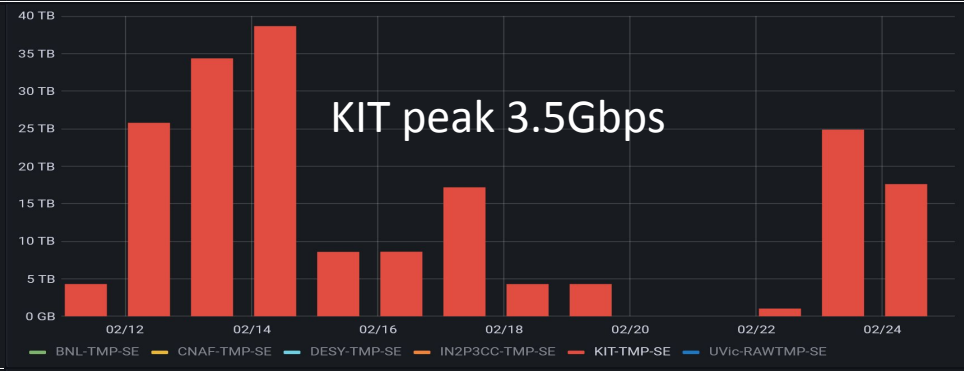


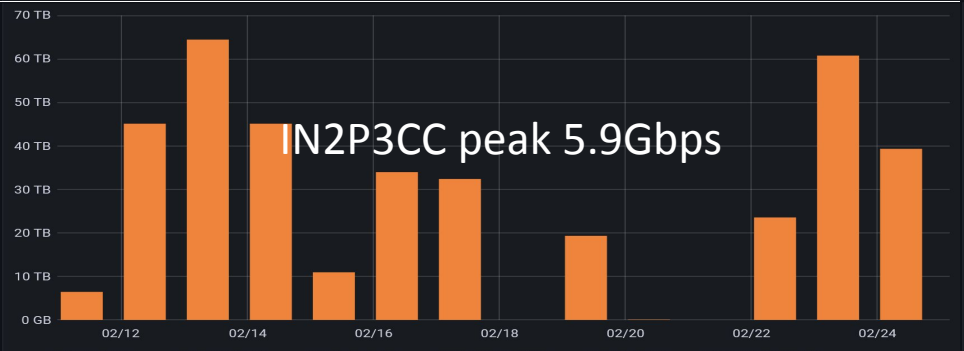
AVERAGE SPEED PER DAY VIEW

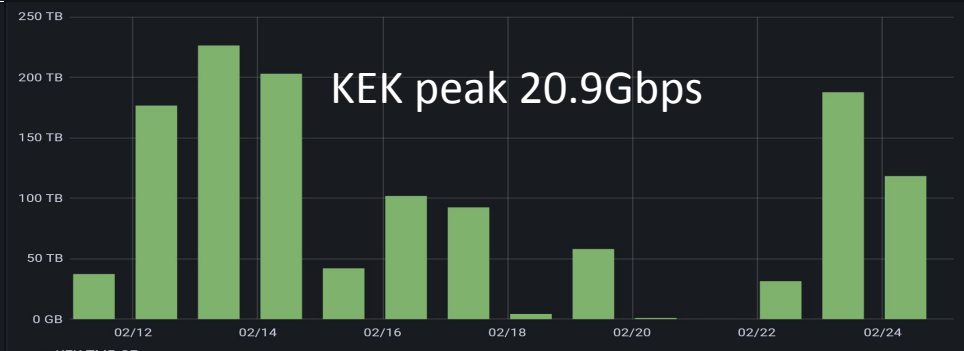




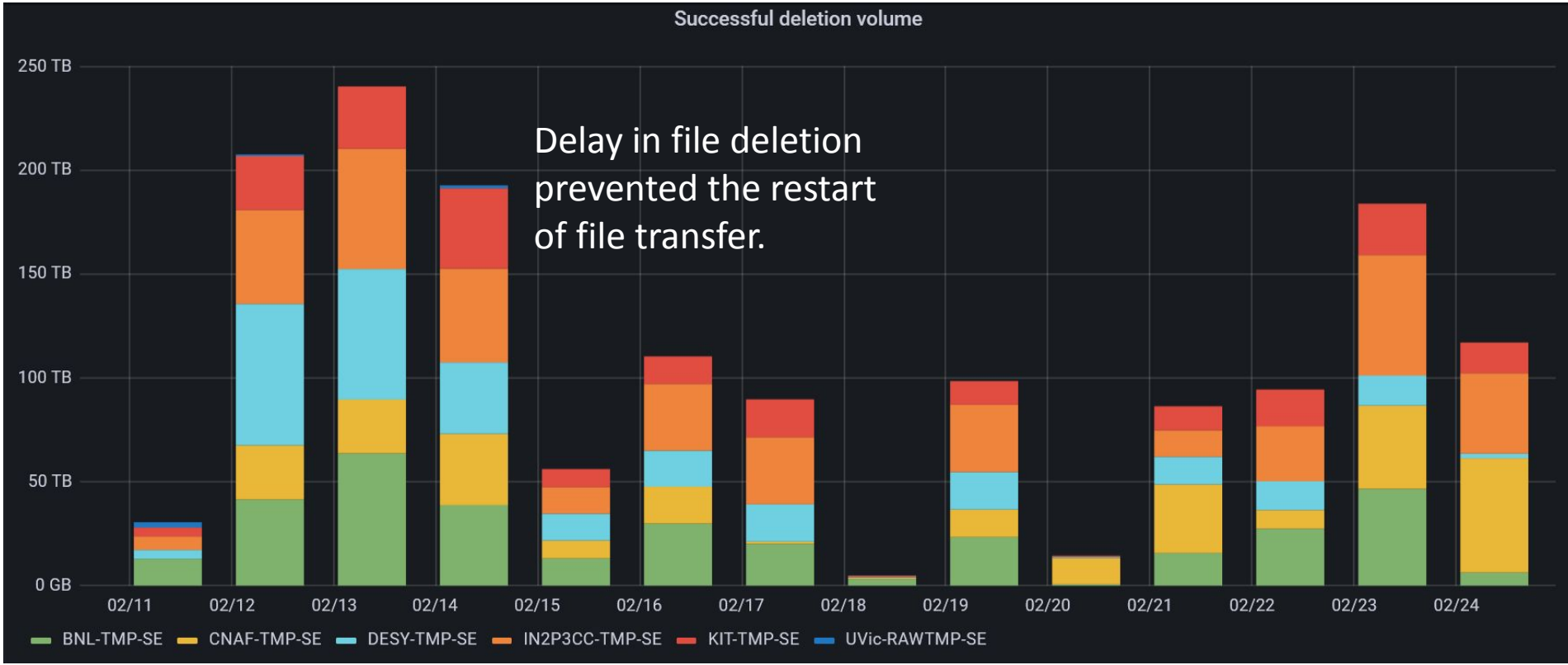








Deleted Volume

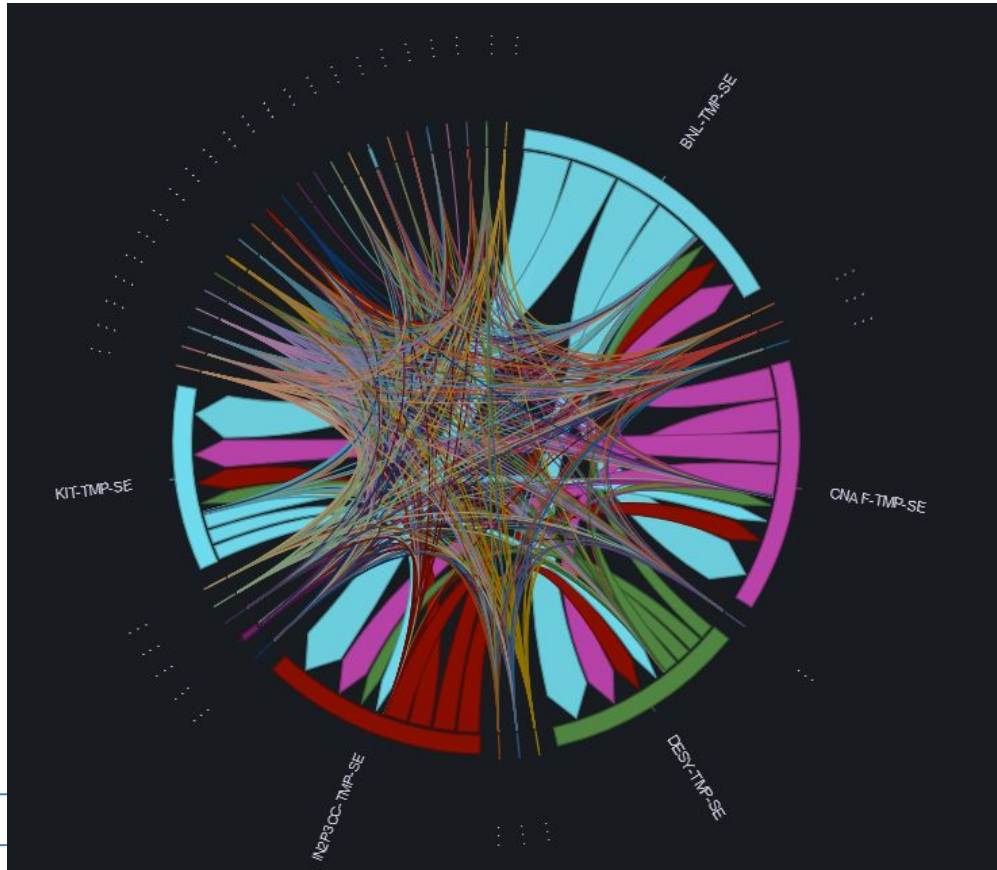


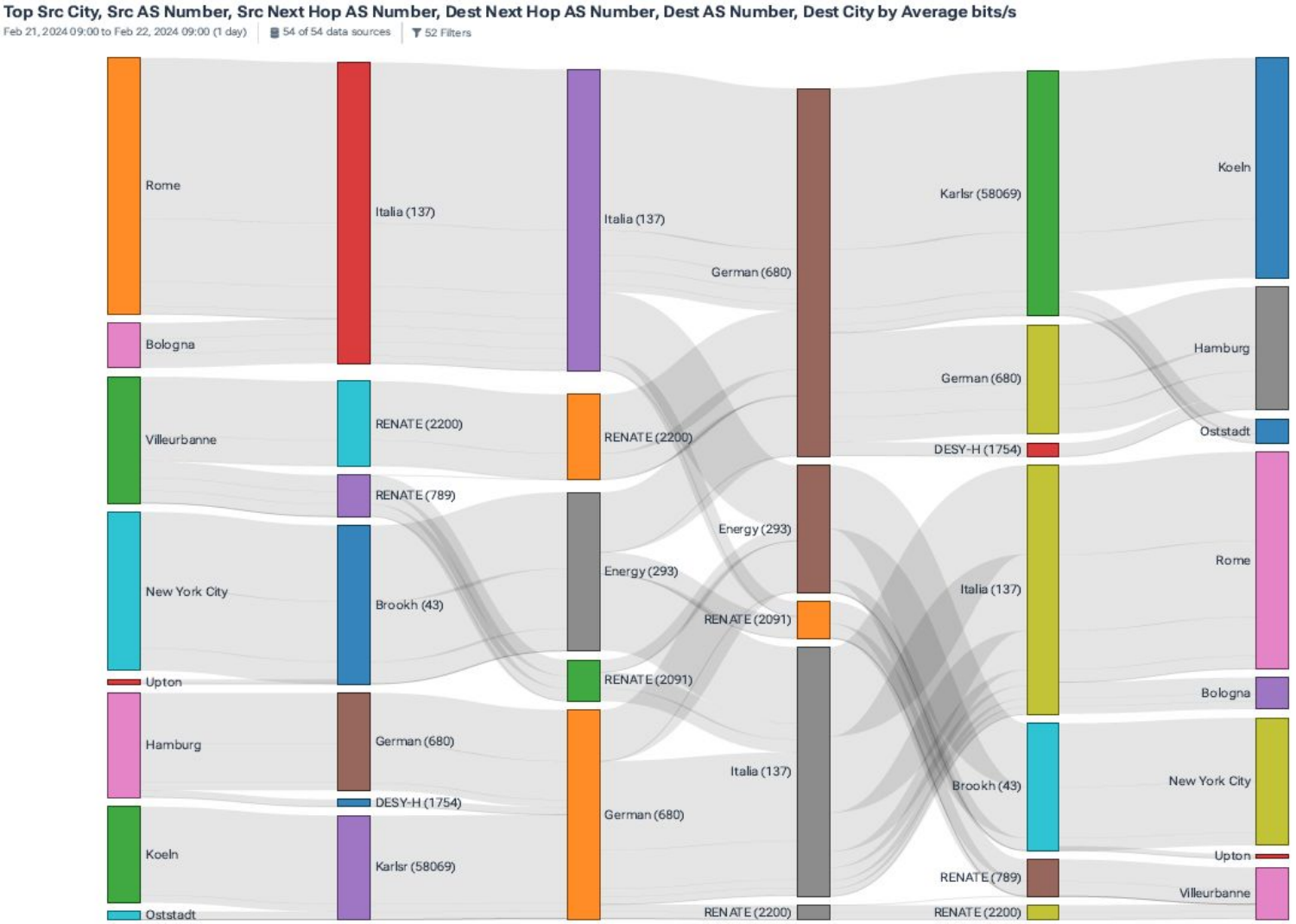
Delay in file deletion prevented the restart of file transfer.

Traffic among RAW Data Centres 21/02/2024 9:00 to 22/02/2024 9:00

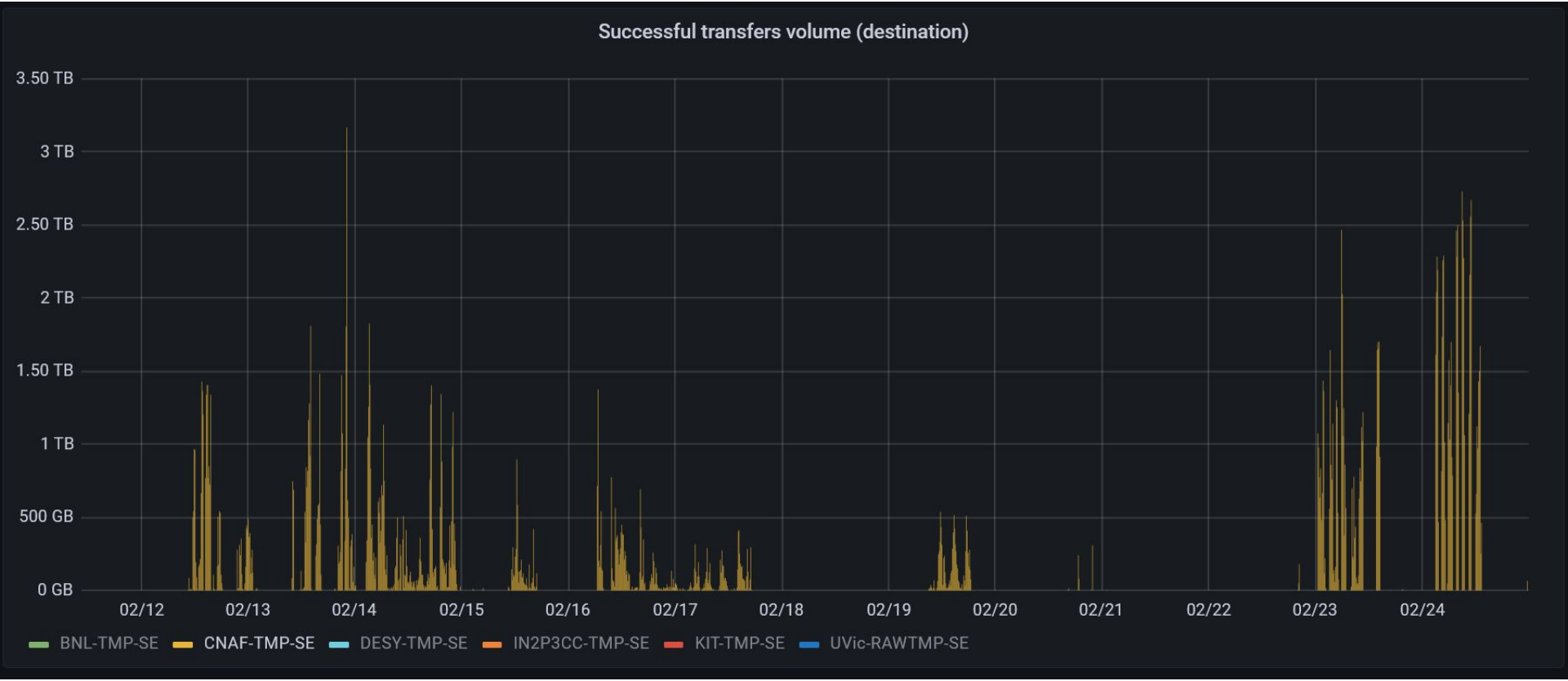
Expected routing table. Tentative of flow analysis ongoing.

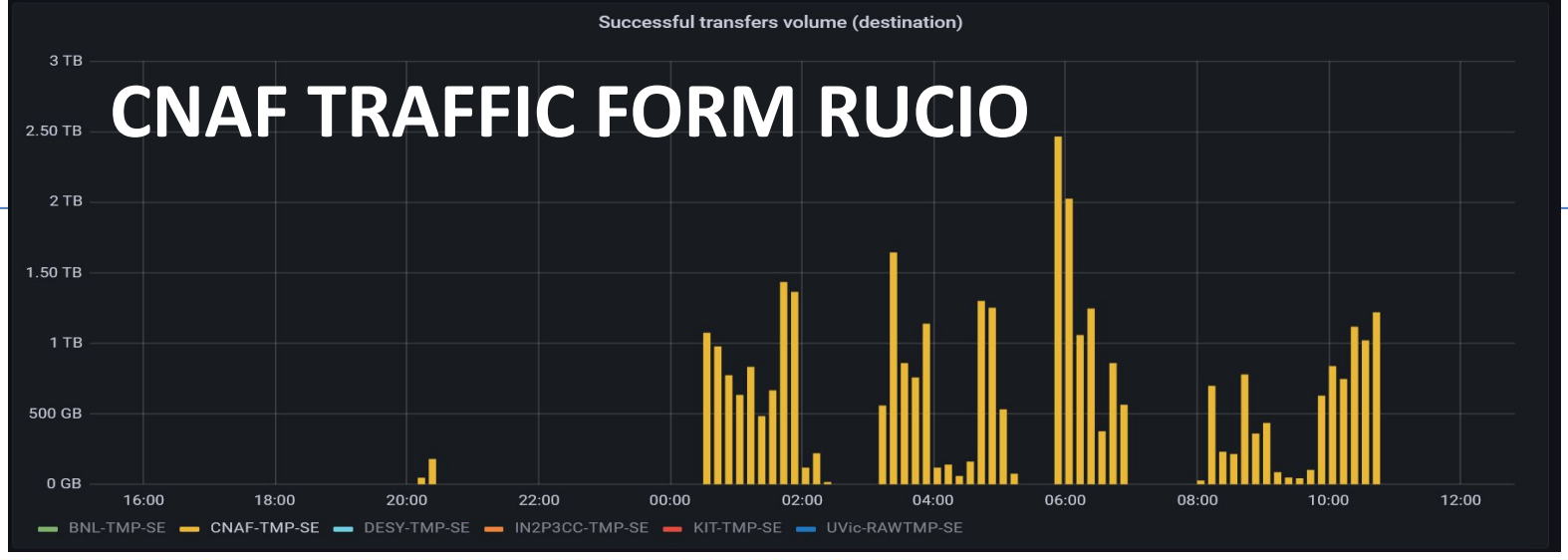
	BNL	KIT	CNAF	DESY	IN2P3CC	Uvic
BNL		LHCOPN	LHCOPN	LHCONE	LHCONE	GeneralIP
KIT	LHCOPN		LHCONE	LHCONE	LHCONE	GeneralIP
CNAF	LHCOPN	LHCONE		LHCONE	LHCONE	GeneralIP
DESY	LHCONE	LHCONE	LHCONE		LHCONE	GeneralIP
IN2P3CC	LHCONE	LHCONE	LHCONE	LHCONE		GeneralIP
Uvic	GeneralIP	GeneralIP	GeneralIP	GeneralIP	GeneralIP	




Top Src City, Src AS Number, Src Next Hop AS Number, Dest Next Hop AS Number, Dest AS Number, Dest City by Average bits/s
Feb 21, 2024 09:00 to Feb 22, 2024 09:00 (1 day)
54 of 54 data sources
52 Filters
Rome
Bologna
Villeurbanne
New York City
Upton
Hamburg
Koeln
Oststadt
Italia (137)
RENATE (2200)
RENATE (789)
Brookh (43)
German (680)
DESY-H (1754)
Karlsr (58069)
Italia (137)
German (680)
RENATE (2200)
Energy (293)
RENATE (2091)
Italia (137)
German (680)
RENATE (2200)
Karlsr (58069)
German (680)
DESY-H (1754)
Italia (137)
Brookh (43)
RENATE (789)
RENATE (2200)
Koeln
Hamburg
Oststadt
Rome
Bologna
New York City
Upton
Villeurbanne

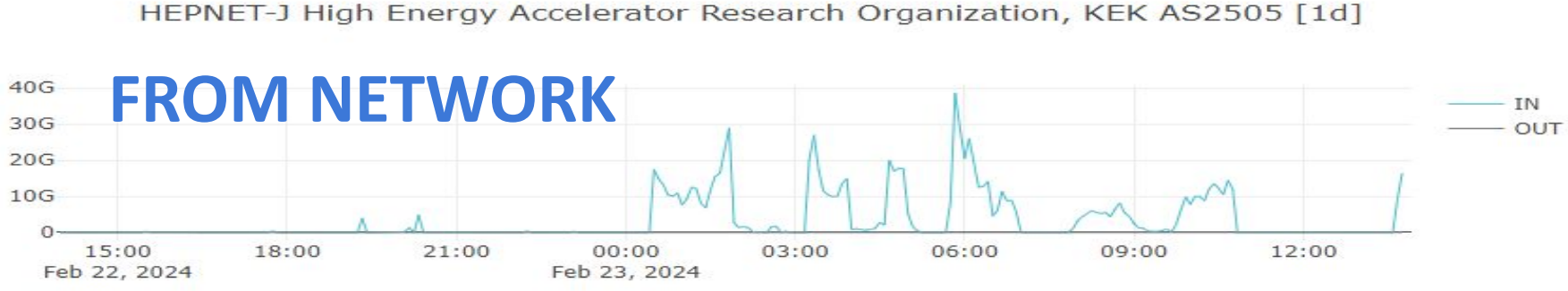
CNAF Traffic of Belle II from RUCIO (10min bin)





Belle II

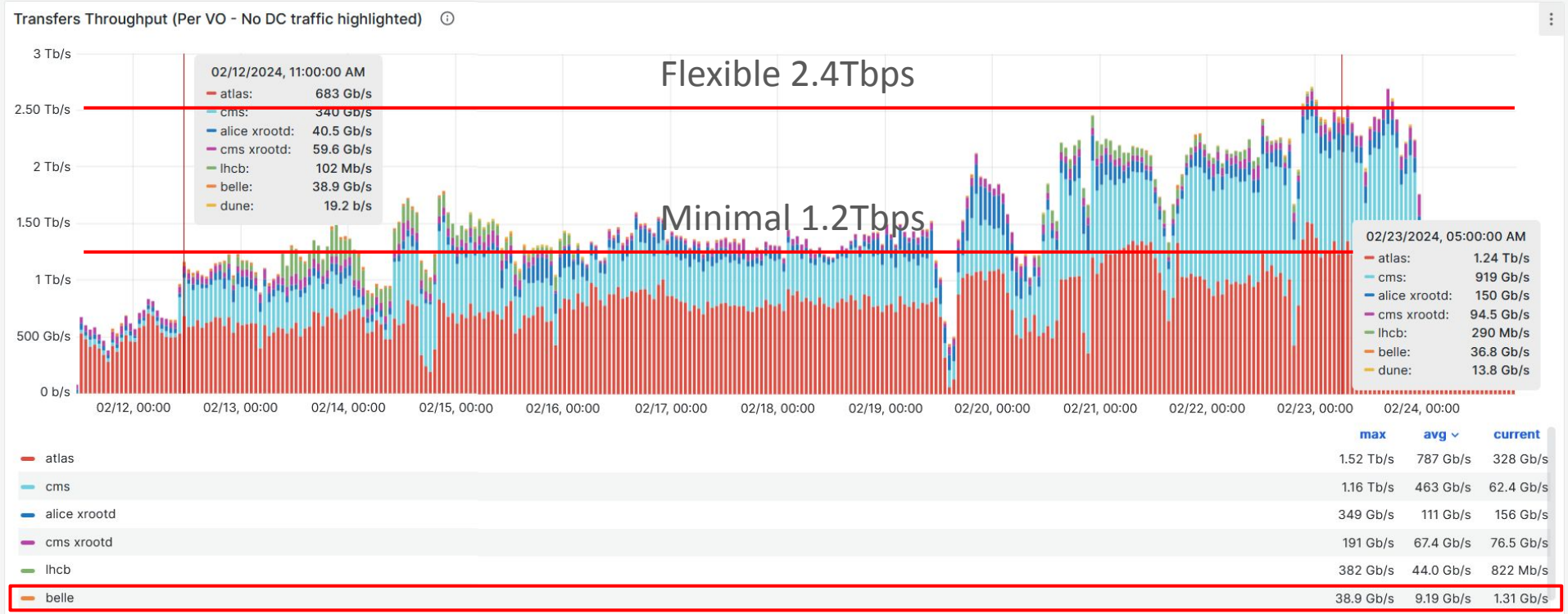
Peering internazionali ricerca: GEANT L3VPN LHCONE



thanks to Marco Marletta

Change time scale: *day - week - month - year*

WLCG Data Challenge 2024



	max	avg	current
atlas	1.52 Tb/s	787 Gb/s	328 Gb/s
cms	1.16 Tb/s	463 Gb/s	62.4 Gb/s
alice xrootd	349 Gb/s	111 Gb/s	156 Gb/s
cms xrootd	191 Gb/s	67.4 Gb/s	76.5 Gb/s
lhcb	382 Gb/s	44.0 Gb/s	822 Mb/s
belle	38.9 Gb/s	9.19 Gb/s	1.31 Gb/s

Not just bandwidth

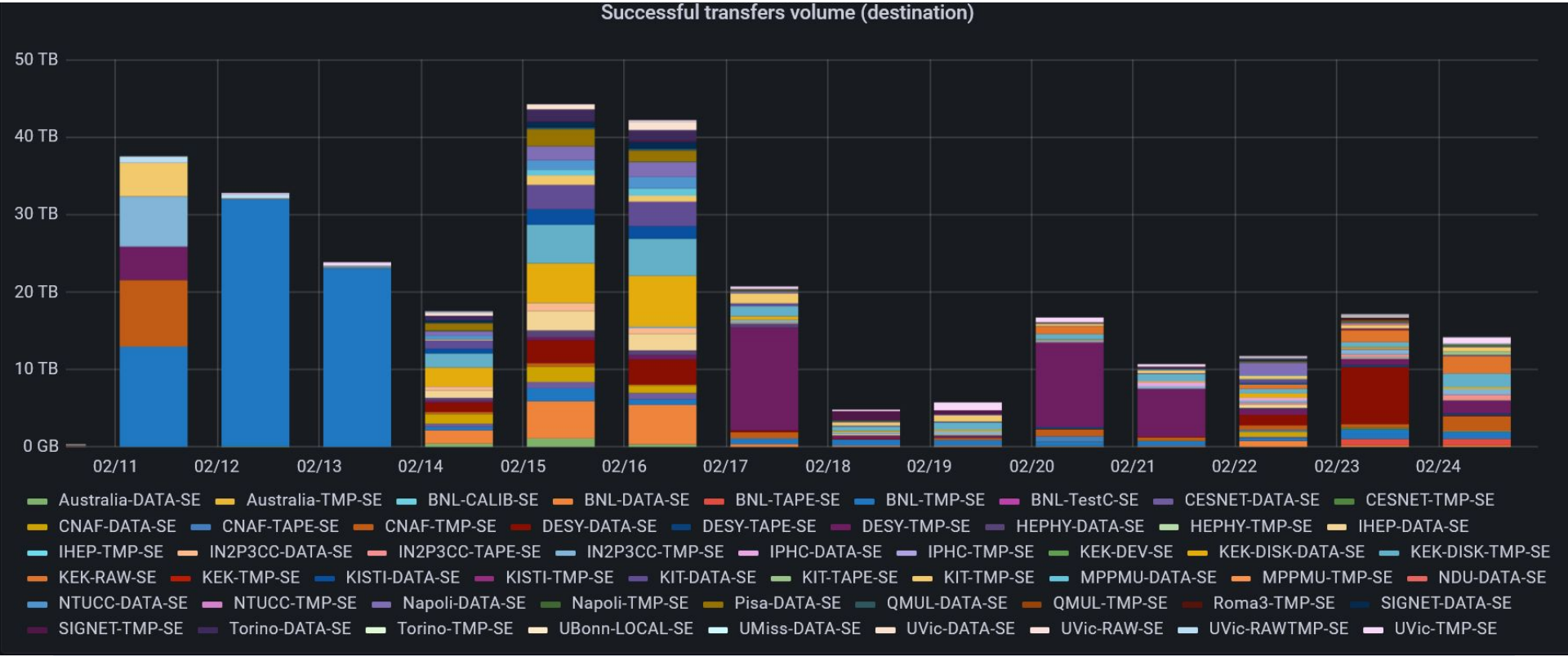
- Data Management System stress test RUCIO, FTS
- Token Access to Storage
- Jumbo frame
- Network CNAF-CERN DCI
- DC24 SENSE/Rucio (Network Orchestration)
- Monitoring systems

Conclusion

- WLCG Data Challenge 2024 has been a successful test for all experiments.
- Belle II has reached the Maximum Target (>18.6Gbps in outbound of KEK), during the test of Flexible model by LHC experiments.
- Lessons learned
  - What went well, where were bottlenecks, organizational improvements
  - Set priorities for ongoing developments
- Planning for DC 2026
  - So far nothing is set except the global target of 50% of expected HL-LHC throughput
  - Belle II will include other traffic than RAW DC export
- Key role of NRENs. I've had high support form GARR. Additionally, other communities have expressed interest in participating, with the Data Challenge being incorporated into the proposal of the European project JENNIFER3, in collaboration with T2K and HyperK.

BACKUP

Belle II Traffic not DC24 related



Personal considerations

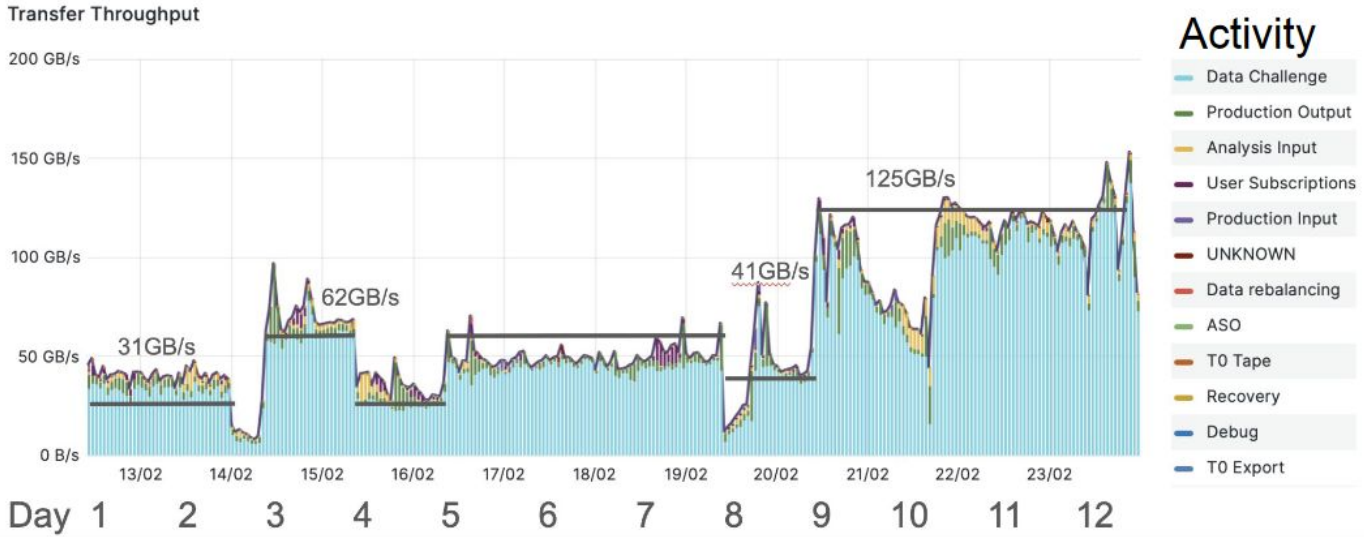
The Data Challenge has proven to be a potent tool for gaining a comprehensive understanding of network usage and to be a powerful technology deployment accelerator. This initiative has also fostered stronger collaboration among various experiment groups and teams.

Key role of NRENs. I've had high support form GARR. Additionally, other communities have expressed interest in participating, with the Data Challenge being incorporated into the proposal of the European project JENNIFER3, in collaboration with T2K and HyperK.

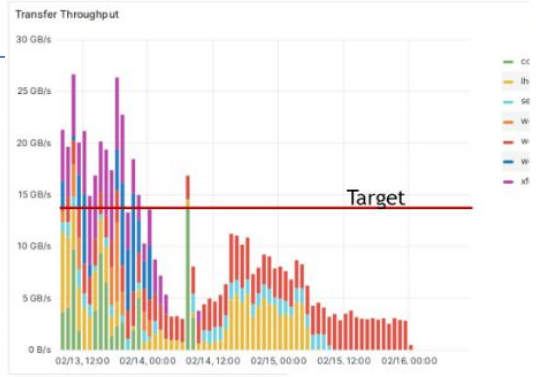
- Daily exercise menu with increasing complexity
- T0 export, T1s to T1s and T1s to T2s, AAA

Date	12 Feb	13 Feb	14 Feb	15 Feb	16 Feb	17 Feb	18 Feb	19 Feb	20 Feb	21 Feb	22Feb	23 Feb
	T0 export	T0 export	T0 export	T1 export	T1 export	T1 export	T1 export	AAA	T0 export	T0 export	T0 export	T0 export
			T1 export		Prod. output	Prod. output	Prod. output		T1 export	T1 export	T1 export	T1 export
									Prod. output	Prod. output	Prod. output	Prod. output
									AAA	AAA	AAA	AAA
Scenario(s)	1	1	1,2	2	2,3	2,3	2,3	4	1,2,3,4	1,2,3,4	1,2,3,4	1,2,3,4
Rate (GB/s)	**31**	**31**	**62**	**31**	**62**	**62**	**62**	**31**	**125**	**125**	**125**	**125**
Rate (Gb/s)	250	250	500	250	500	500	500	250	1000	1000	1000	1000

- First week targets were mostly met easily
- Overall target of ~125GB/s was reached with significant effort
  - A few hundred links maximum (Prod + DC)
  - More data injected than the target required

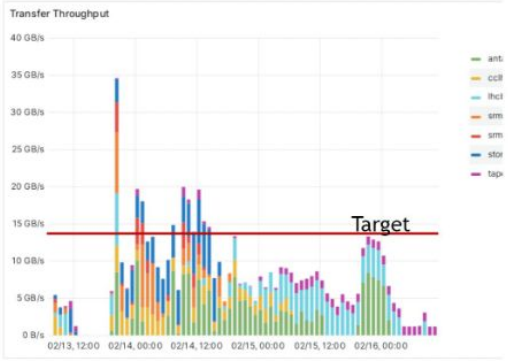


EOS -> Disk link



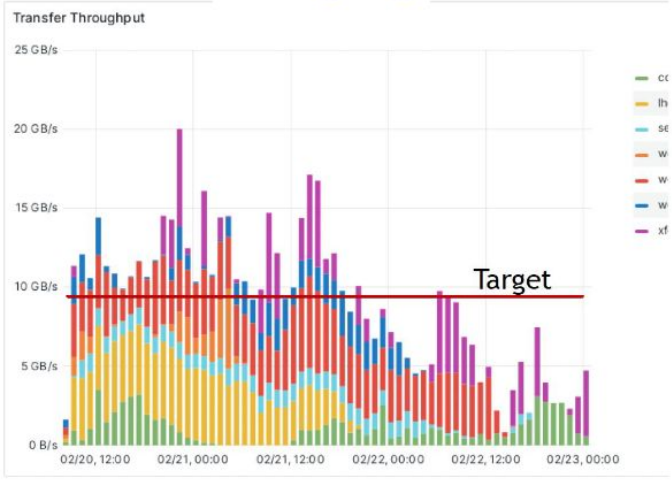
- Target throughput (14GiB/s) was achieved during the first day
- Lower throughput later
  - Some sites finished transferring their part during the first day so were no longer contributing to overall throughput
  - Submissions were slow and not optimal
  - Submission agent got stuck a few times, that was also a contributing factor

Disk -> Tape link



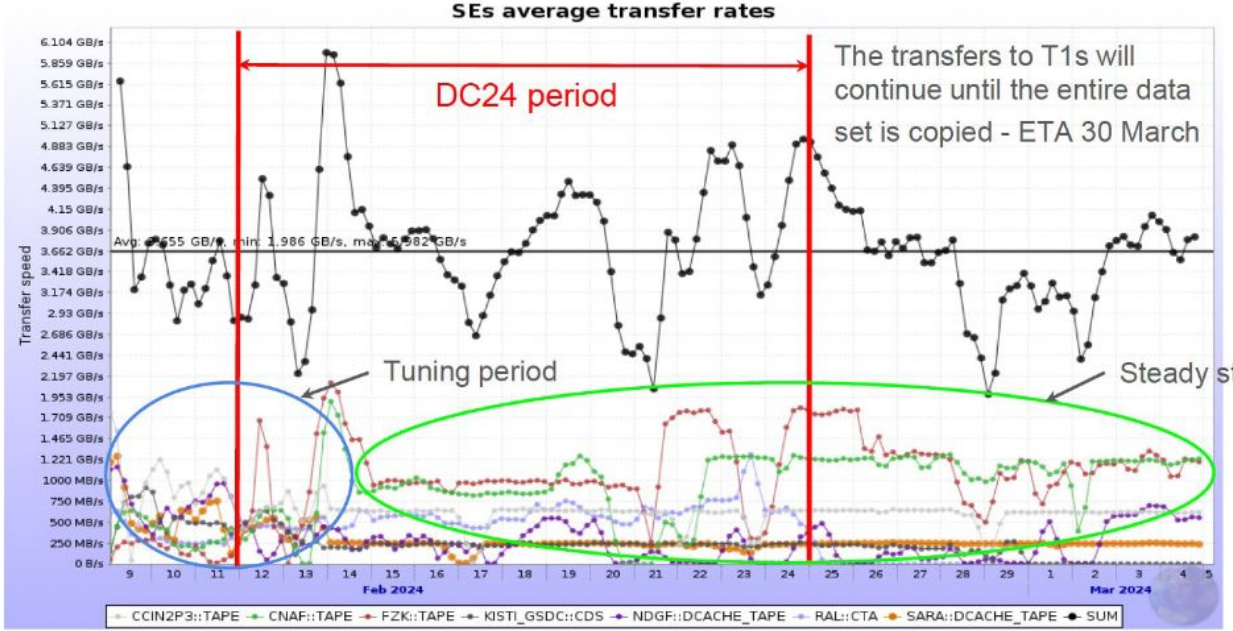
- Target threshold (14GiB/s) crossed several times
  - Max around 35GiB/s
  - Spikier throughput because of the nature of the link and submission agent problems

Staging



- Target throughput (9.58 GiB/s) was achieved during the first two days of the test
- Lower throughput later
  - Some sites finished transferring their part and were no longer contributing

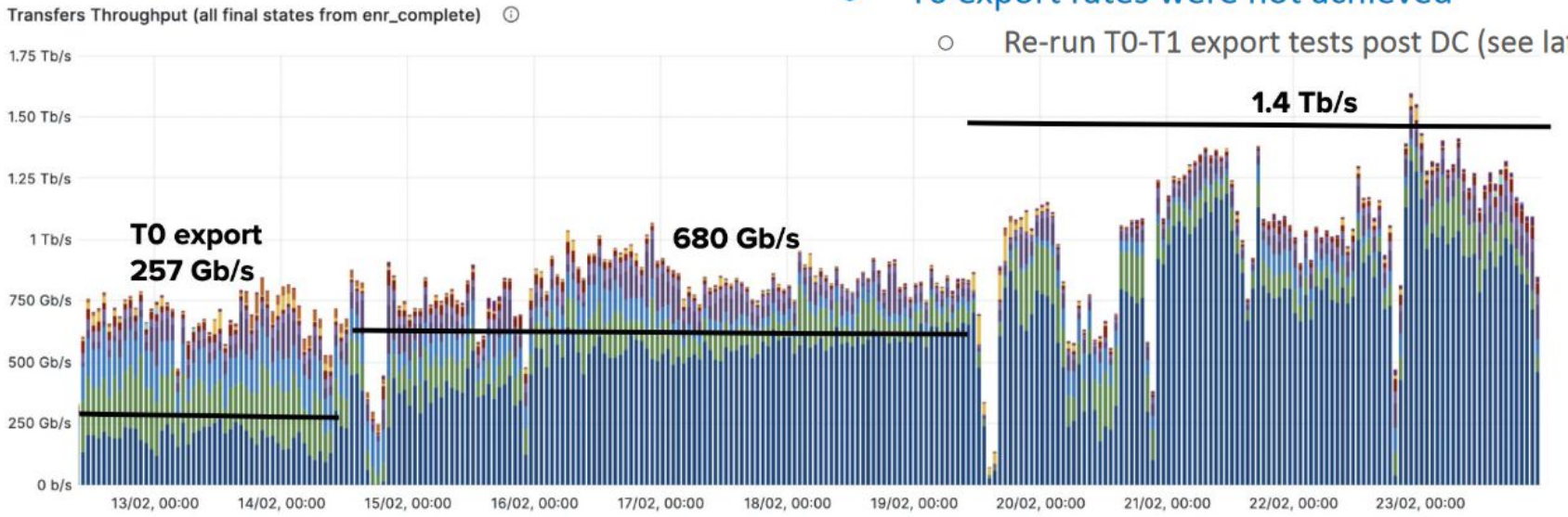
Time evolution T1s



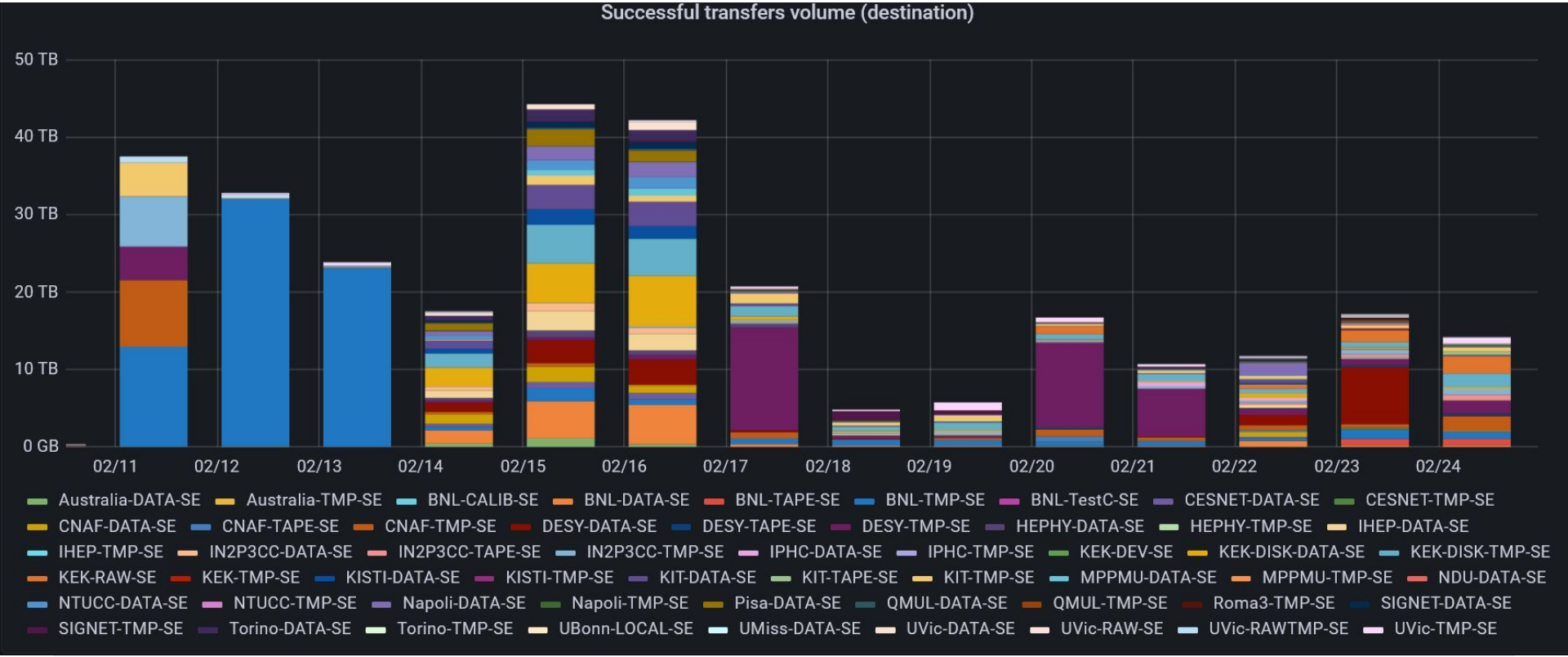
Centre	Target rate GB/s	Average achieved GB/s
CNAF	0.8	0.98 (+20%)
IN2P3	0.4	0.6 (+40%)
KISTI	0.2	0.25 (+22%)
GridKA	0.6	1.12 (+90%)
NDGF	0.3	0.35 (+15%)
NL-T1	0.1	0.25 (+150%)
RAL	0.1	0.58 (+500%)
CERN	*10*	*14.2 (+40%)*

DC24 was a great success for ALICE, achieving above target rates at every site, with minimal interference, and no effect on other activities

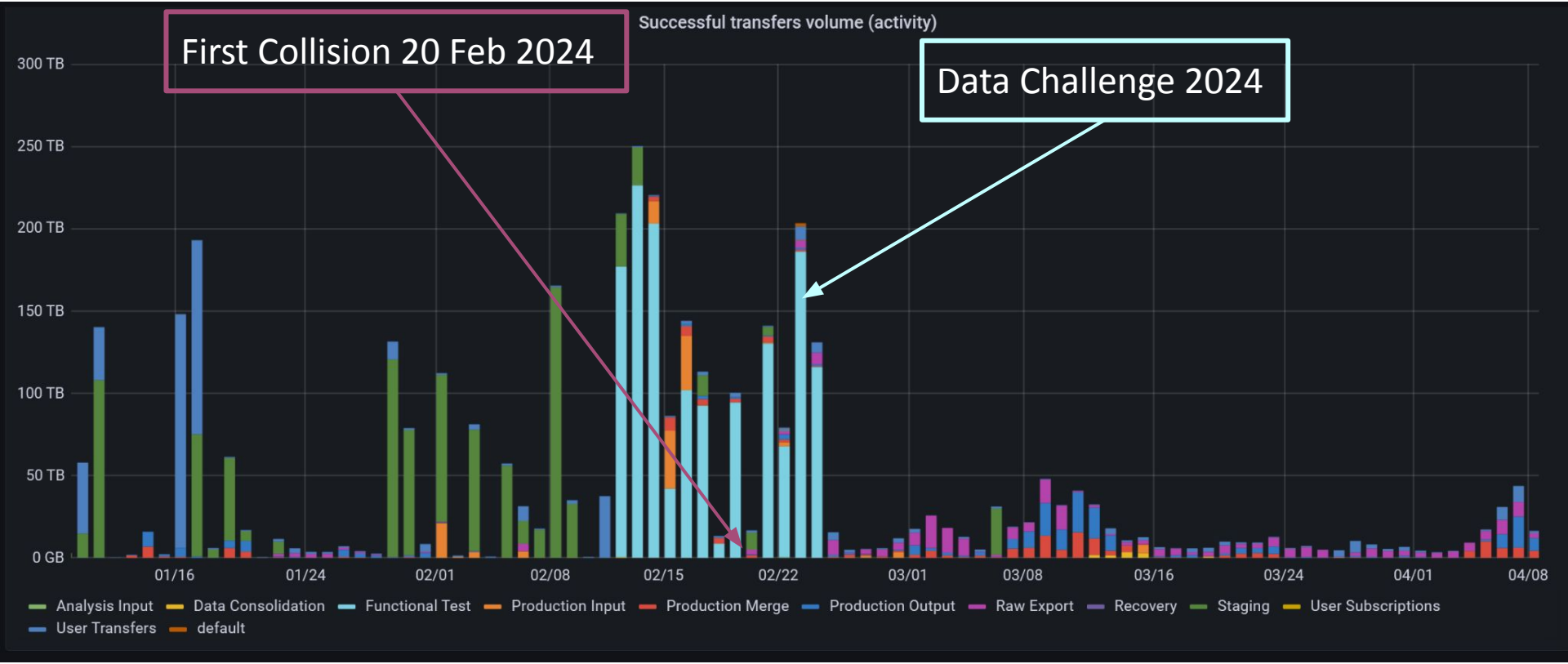
- Generally considered success for highlighting bottlenecks, though rates hampered by the really large number of links
  - Injections on >1200 links every 15m
    - ~2000 links with production
  - Short data sets lifetime 1h -> 2h -> 3h
  - Helped highlighting problems that wouldn't have been seen otherwise
- None of the bottlenecks were due to the network specifically
  - Some sites had the LHCOPN link down but had alternative paths
- Some sites struggled mostly due to storage limitations
  - 17 problems were reported on GGUS
- T0 export rates were not achieved
  - Re-run T0-T1 export tests post DC (see later)



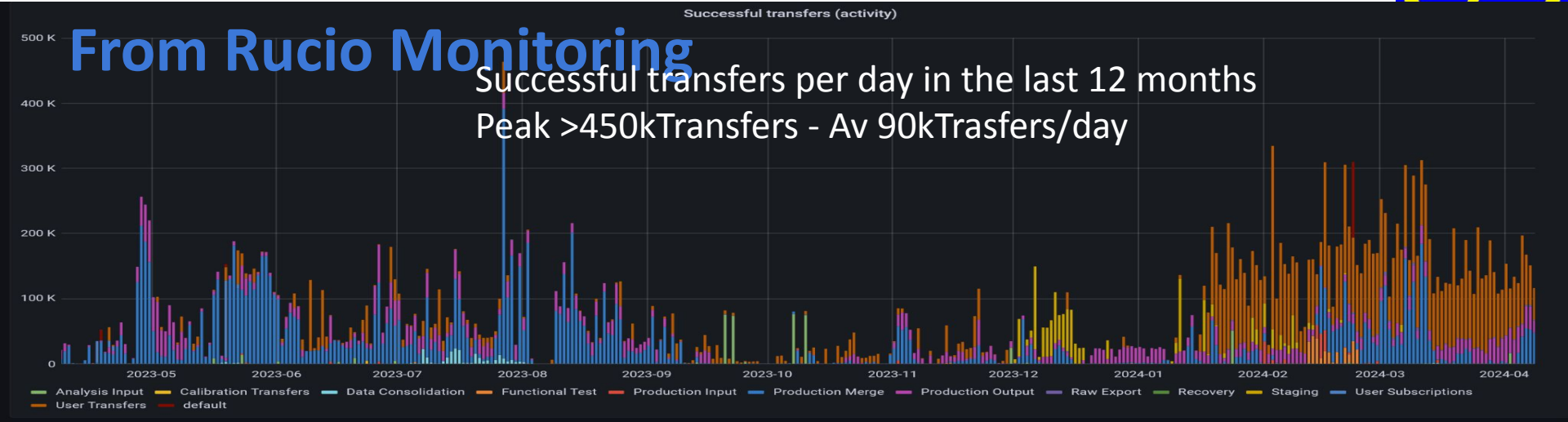
Traffic during DC24 excluding Functional test

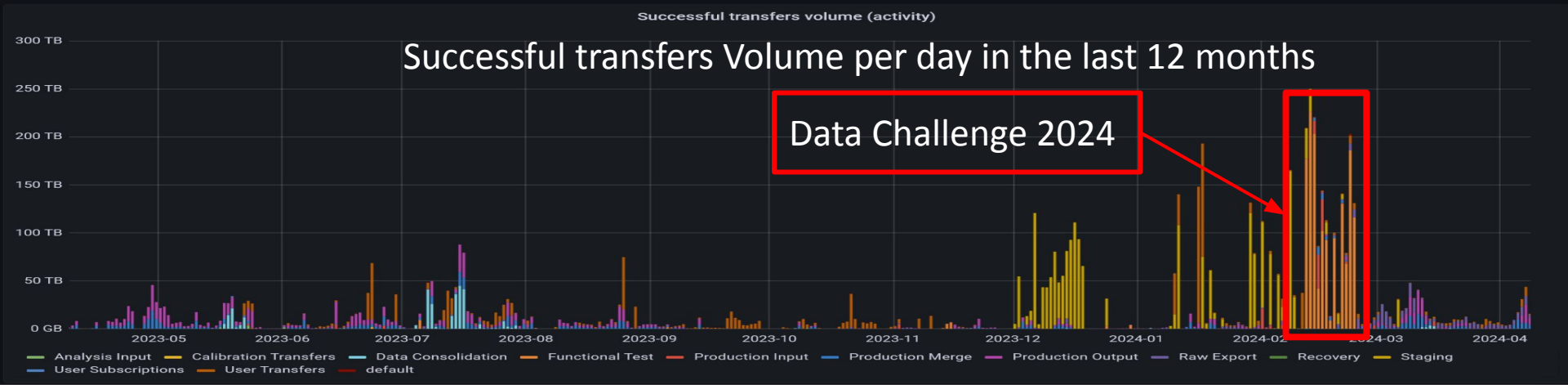


Successful Transfer Volume since January 2024

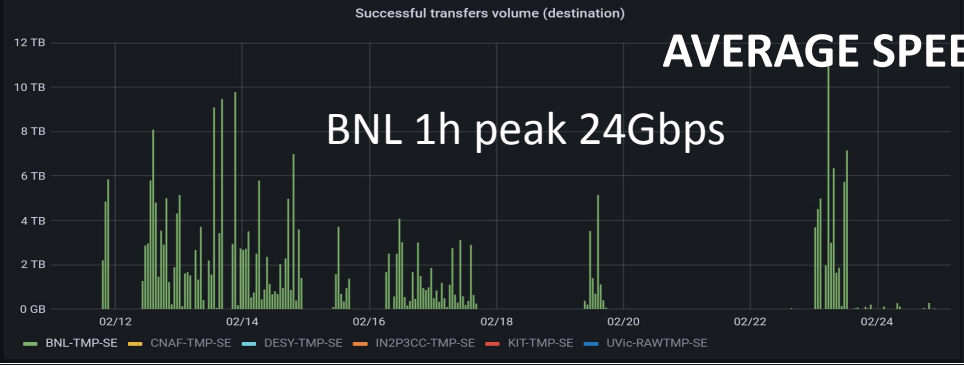


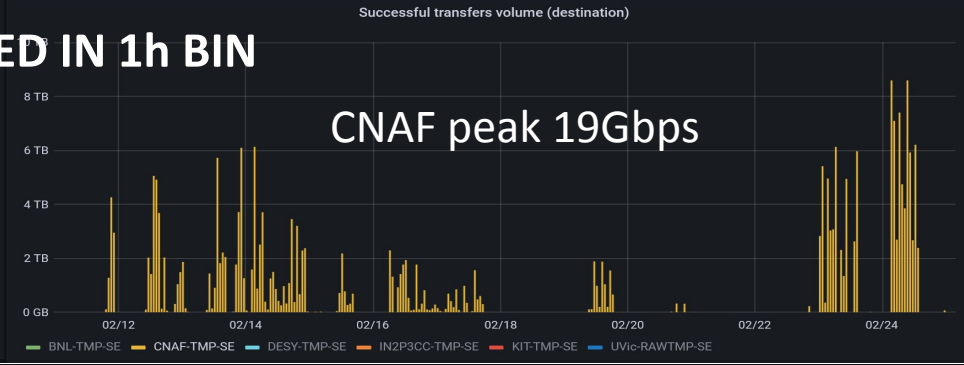
From Rucio Monitoring

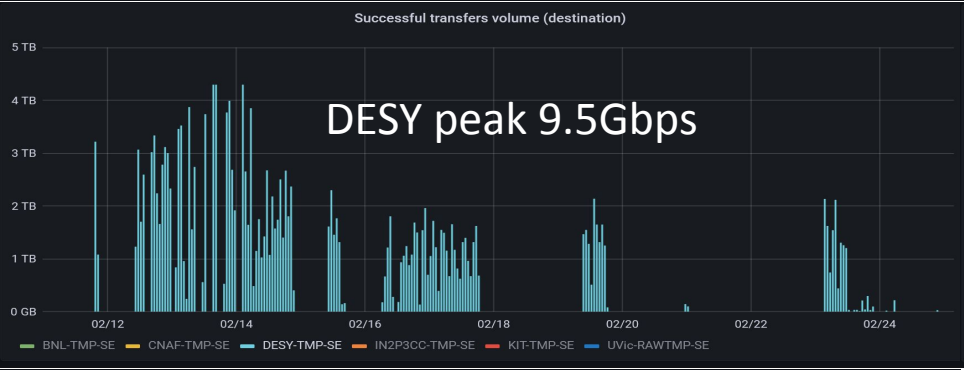


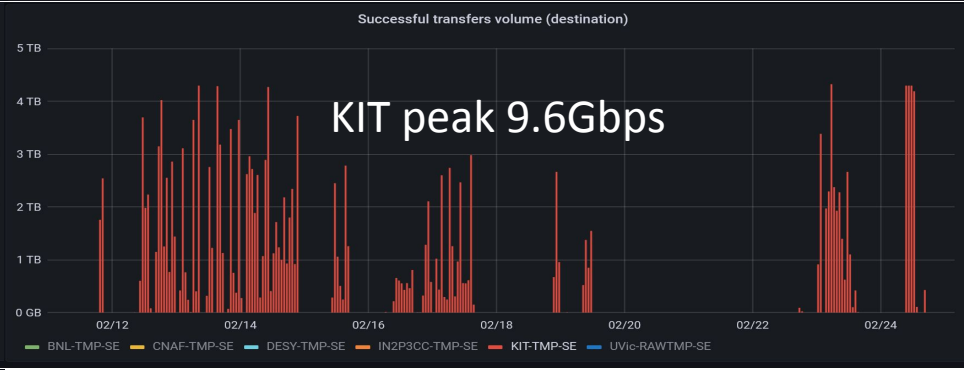


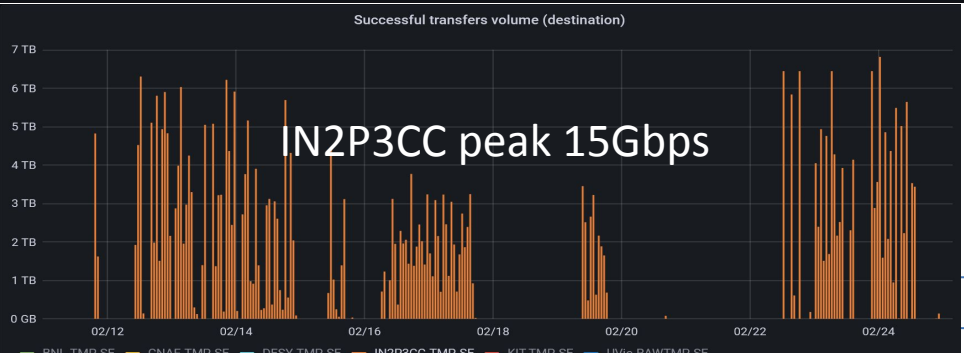
AVERAGE SPEED IN 1h BIN

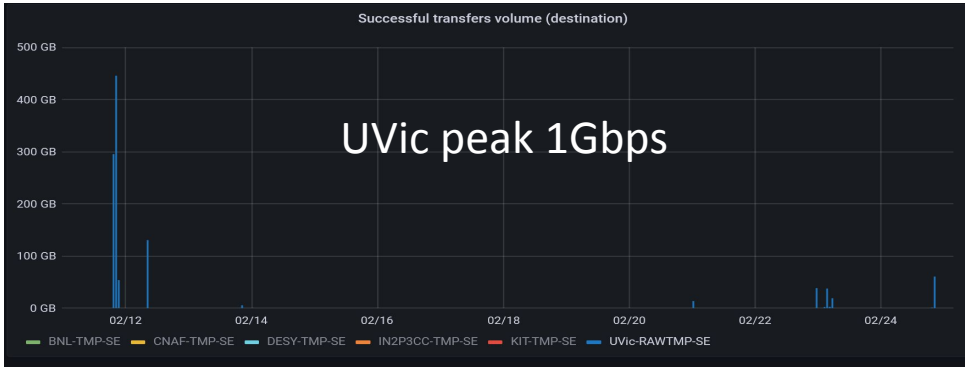




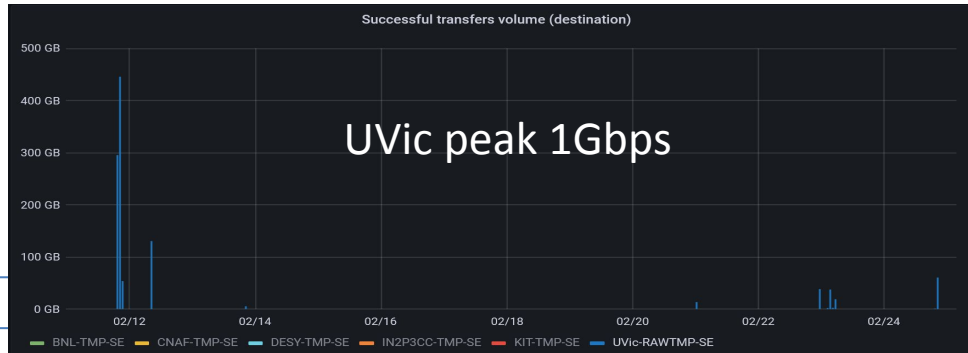








UVic peak 1Gbps



UVic peak 1Gbps

Zoom on 12/02/2024 9:00 to 14/02/2004 23:00

Successful transfers volume (destination)

Increasing number of active file from 200 to 500

206TB/23h - 19Gbps
kek2-fts03 max active 200/200

203TB/22h - 20Gbps
kek2-fts03 max active 500/500

4.50 TB
4 TB
3.50 TB
3 TB
2.50 TB
2 TB
1.50 TB
1 TB
500 GB
0 GB

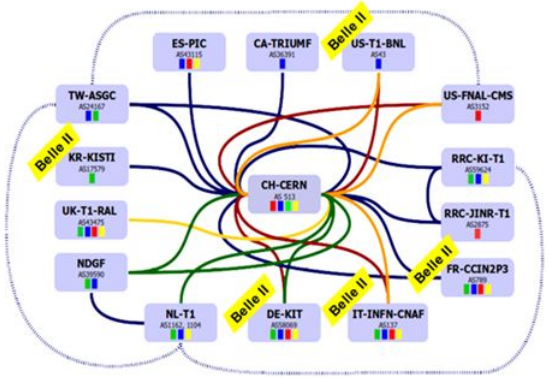
02/12 08:00 02/12 12:00 02/12 16:00 02/12 20:00 02/13 00:00 02/13 04:00 02/13 08:00 02/13 12:00 02/13 16:00 02/13 20:00 02/14 00:00 02/14 04:00 02/14 08:00 02/14 12:00 02/14 16:00 02/14 20:00 02/15 00:00

BNL-TMP-SE CNAF-TMP-SE DESY-TMP-SE IN2P3CC-TMP-SE KIT-TMP-SE UVic-RAWTMP-SE

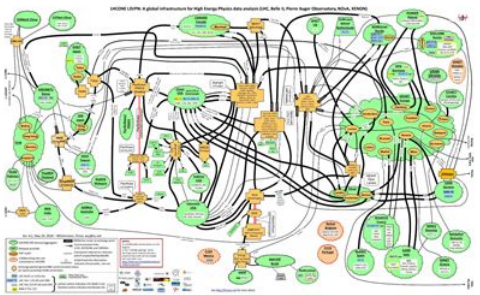
Belle II Network

100G Global Ring via SINET

LHCOPN Optical infrastructure that can be used without jeopardizing resources

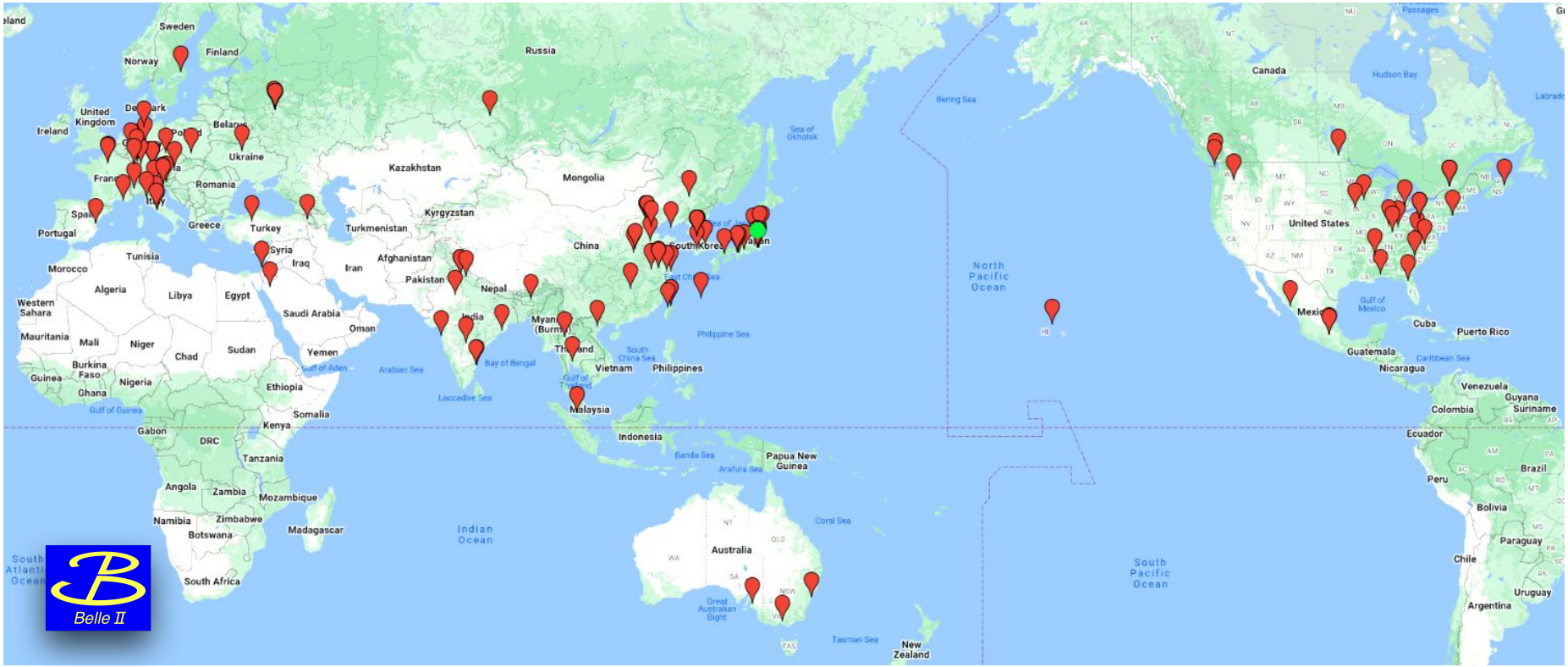


LHCONE L3 VPN Connecting all the major Data Centres



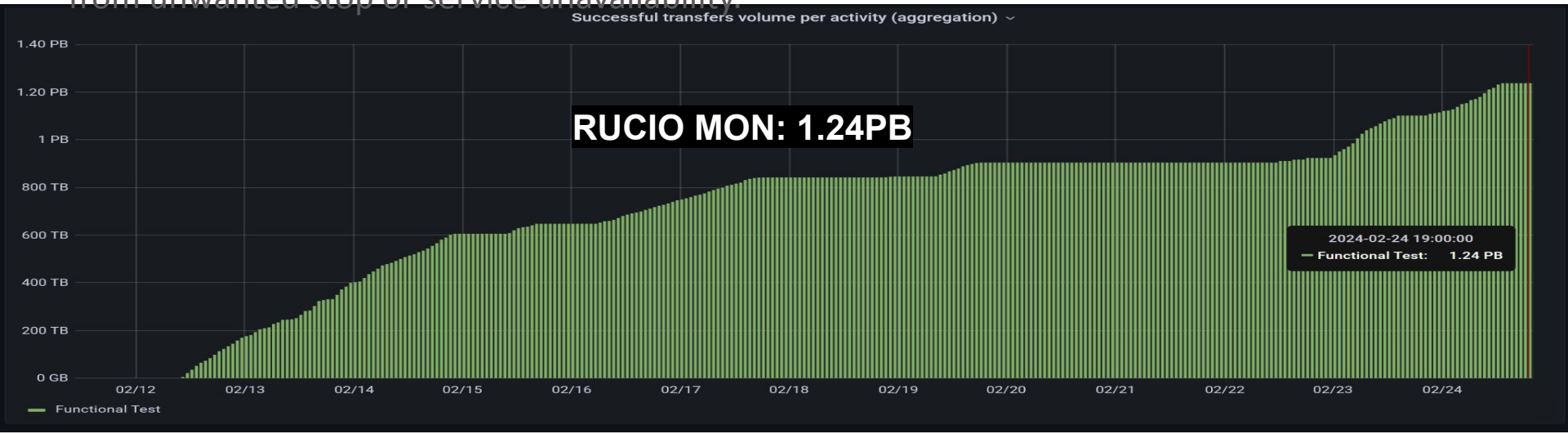
The Belle II Experiment

Around 1200 members, 131 institutions, 28 countries



Total Data transfered.

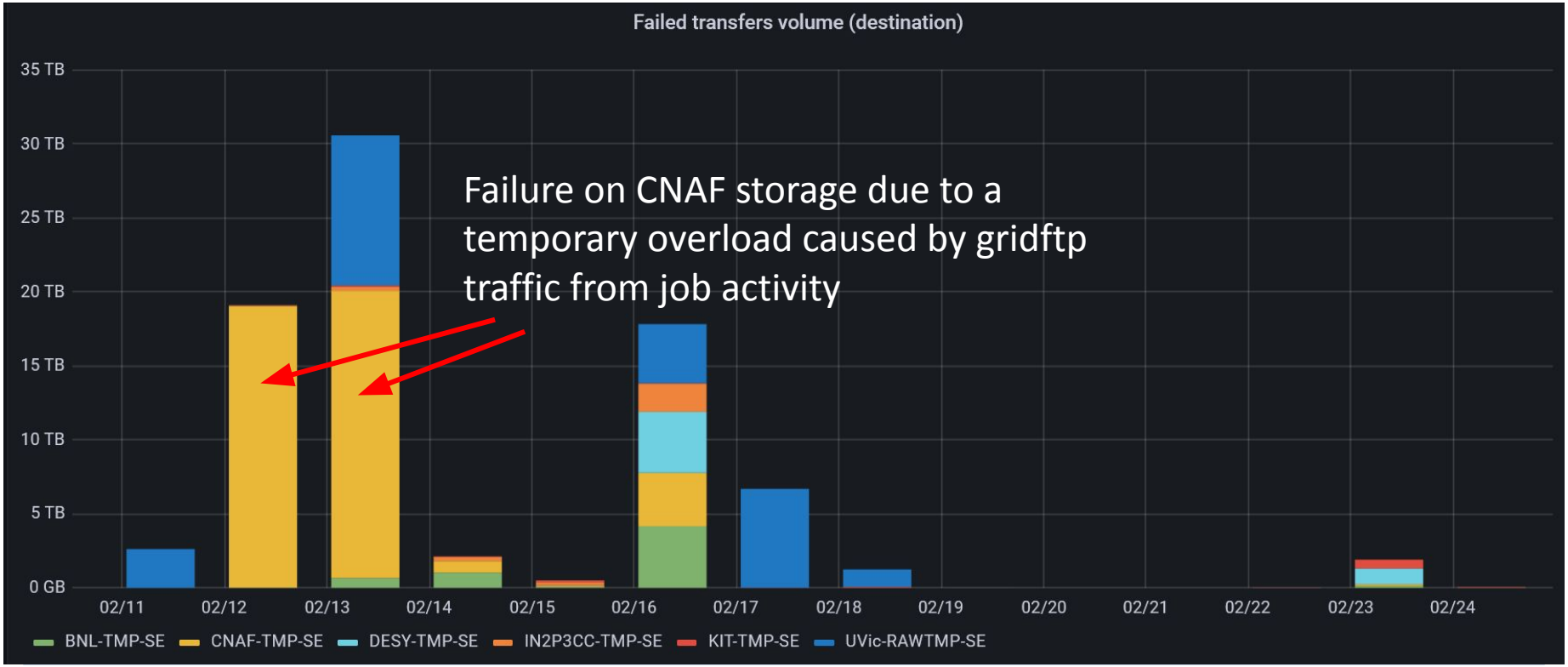
1.24 PB of synthetic data copied from KEK to RAW DC in 12 days of tests performed in burst. Average of 103TB per day, more than 2 times the needed throughput from KEK to RAW DC at maximum luminosity. This demonstrate how the capability to reach 5x40TB/day may protect from unwanted stop or service unavailability.



Zoom on 23/02/2024 0:00 to 23/02/2024 14:00

178TB/15h - 26Gbps
kek2-fts03 max active src 500 and max 200 per dest

Failed Transfers



IAM Performance

Experiment	Issued tokens	Max. number of tokens in DB	Peak token request rate	Typical token request rates
ATLAS	2.6 M	1.03 M	5 Hz	3 Hz (12 days)
CMS	2.7 M	0.97 M	200 Hz	60 Hz (6 hours), 20 Hz (10 hours), 1 Hz (11 days)
LHCb	3.4 M	1.65 M	120 Hz	25 Hz (2 days), 1 Hz (10 days)

- LHCb tested the “1 token per file transfer” configuration for 2 days which increased their token request rate.
- CMS had high token request rates for ~16h
- During these peak token requests rates on CMS and LHCb, IAM slowed down on issuing tokens
- ALICE instance isn’t included in the summaries, as it was not used for any data management operations.

IAM Performance - Challenges Faced

- **Database Overload:**
  - Increased token request rates led to database overload, impacting response times.
  - Token lifetimes of up to 30 days delayed cleanup processes during DC24 which led to the database being filled with tokens.
  - The database cleanup algorithm was running slowly and filled up the database connection pool.
- **Token Management:**
  - Suboptimal token usage patterns, especially concerning refresh tokens.

Lessons Learned:

- **Token Lifecycle Management:**
  - Implement shorter token lifetimes to facilitate quicker cleanup processes during peak usage periods.
- **Token Management Enhancements:**
  - Stop storing access tokens in the DB to improve the performance. This needs a modification of token management engine (MitreID). IAM developers are working on this.
- **Collaborative Discussions:**
  - Foreseen discussions between Rucio, DIRAC, FTS, and IAM experts to explore more efficient token orchestration methods for large-scale data transfers.
- **Performance Testing:**
  - Enhance IAM performance tests to make them closer to the real use-cases and include closer examination of latency issues.