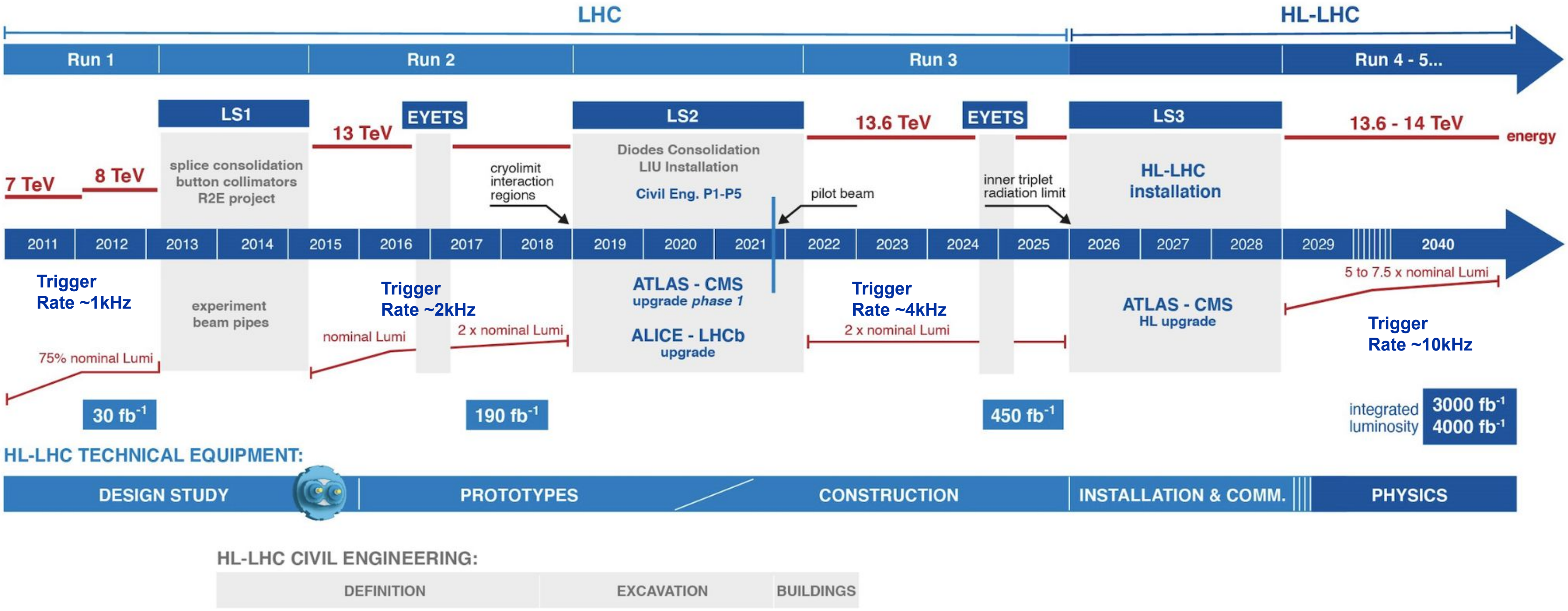
WLCG Data Challenge 24

GDB

2024-03-27

Christoph Wissing (DESY), Mario Lassnig (CERN)
On behalf of the DC24 community

Introduction: LHC and its High Luminosity upgrade



LHC

HL-LHC

Run 1

Run 2

Run 3

Run 4 - 5...

LS1

EYETS

LS2

EYETS

LS3

7 TeV

8 TeV

13 TeV

13.6 TeV

13.6 - 14 TeV

energy

splice consolidation
button collimators
R2E project

cryolimit
interaction
regions

Diodes Consolidation
LIU Installation

Civil Eng. P1-P5

pilot beam

inner triplet
radiation limit

HL-LHC
installation

2011 2012 2013 2014 2015 2016 2017 2018 2019 2020 2021 2022 2023 2024 2025 2026 2027 2028 2029 2040

Trigger
Rate ~1kHz

experiment
beam pipes

Trigger
Rate ~2kHz

ATLAS - CMS
upgrade *phase 1*

ALICE - LHCb
upgrade

Trigger
Rate ~4kHz

ATLAS - CMS
HL upgrade

5 to 7.5 x nominal Lumi

Trigger
Rate ~10kHz

75% nominal Lumi

nominal Lumi

2 x nominal Lumi

2 x nominal Lumi

30 fb^{-1}

190 fb^{-1}

450 fb^{-1}

integrated
luminosity

3000 fb^{-1}
4000 fb^{-1}

HL-LHC TECHNICAL EQUIPMENT:

DESIGN STUDY | PROTOTYPES | CONSTRUCTION | INSTALLATION & COMM. | PHYSICS

HL-LHC CIVIL ENGINEERING:

DEFINITION | EXCAVATION | BUILDINGS

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

Data Challenges for HL-LHC

- WLCG has been 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
  - Increase 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** needs and challenges
  - For the DCs find agreements across the LHC experiments and beyond
    - Suited dates
    - Reasonable targets
    - Functionalities
  - Help in orchestration

- Dates and high level goals always approved by WLCG Management Board

Recap of (initial) modelling & resulting rates for HL-LHC

- ATLAS & CMS T0 to T1 per experiment
  - 350PB RAW per year, taken and distributed during typical LHC uptime of 7M seconds
  - 50GB/s or 400Gbps
  - Another 100Gbps estimated for prompt reconstruction data tiers (AOD, other derived output)
  - 1Tbps for CMS and ATLAS summed
    - ALICE & LHCb T0 Export
  - 100 Gbps per experiment estimated from Run-3 rates

WLCG data challenges for HL-LHC - 2021 planning
https://zenodo.org/records/5532452

- **Minimal Model**
  - Sum (ATLAS,ALICE,CMS,LHCb)*2(for bursts)*2(*overprovisioning*) = **4.8Tbps for the expected HL-LHC bandwidth needs**

- **Flexible Model**
  - Assumes reading of data from above for reprocessing/reconstruction in 3 months (about 7M seconds)

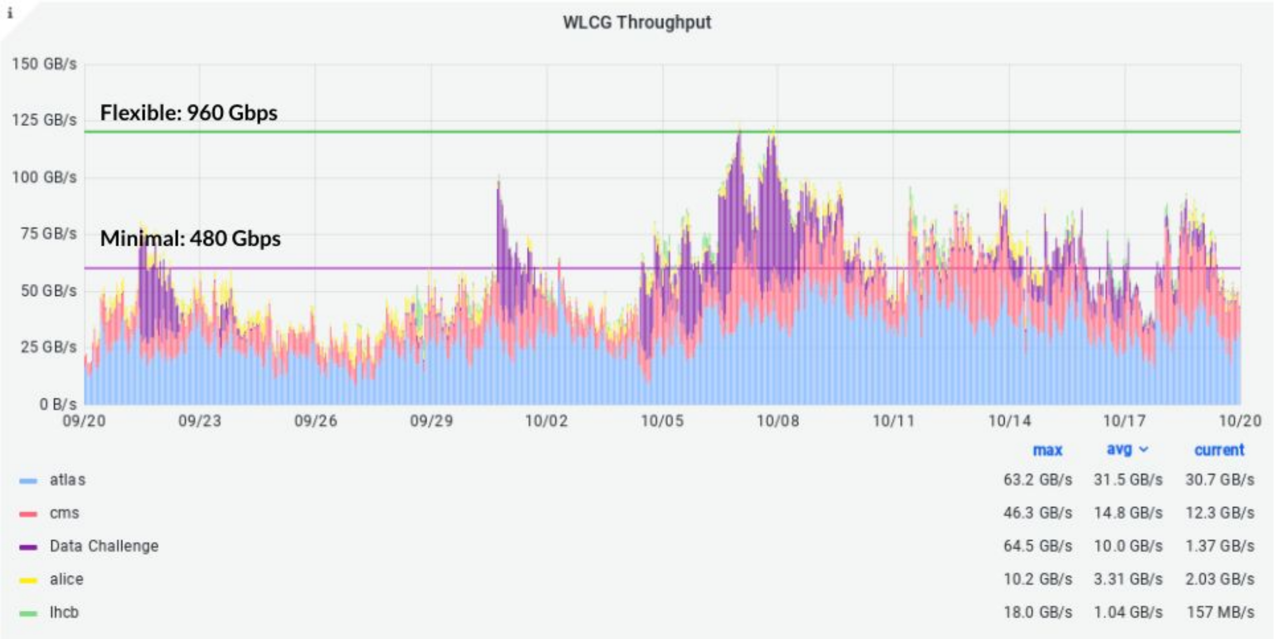
    Means doubling the Minimal Model: **9.6Tbps for the expected HL-LHC bandwidth needs**

    However data flows primarily from the T1s to T2s and T1s!

- Data Challenges target: **50% filling of expected** HL-LHC bandwidth needs

DC21 - 10% of HL-LHC Throughput

However, we managed to reach 100% of the (minimal) DC21 target!

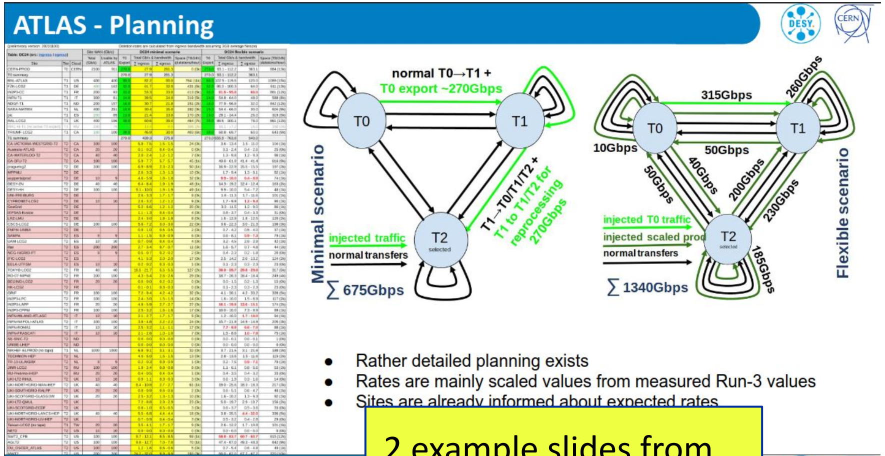


	max	avg	current
🟦 atlas	63.2 GB/s	31.5 GB/s	30.7 GB/s
🟥 cms	46.3 GB/s	14.8 GB/s	12.3 GB/s
🟪 Data Challenge	64.5 GB/s	10.0 GB/s	1.37 GB/s
🟨 alice	10.2 GB/s	3.31 GB/s	2.03 GB/s
🟩 lhcb	18.0 GB/s	1.04 GB/s	157 MB/s

Network Data Challenges 2021 wrap-up and recommendations
https://zenodo.org/records/5767913

Planning of DC24

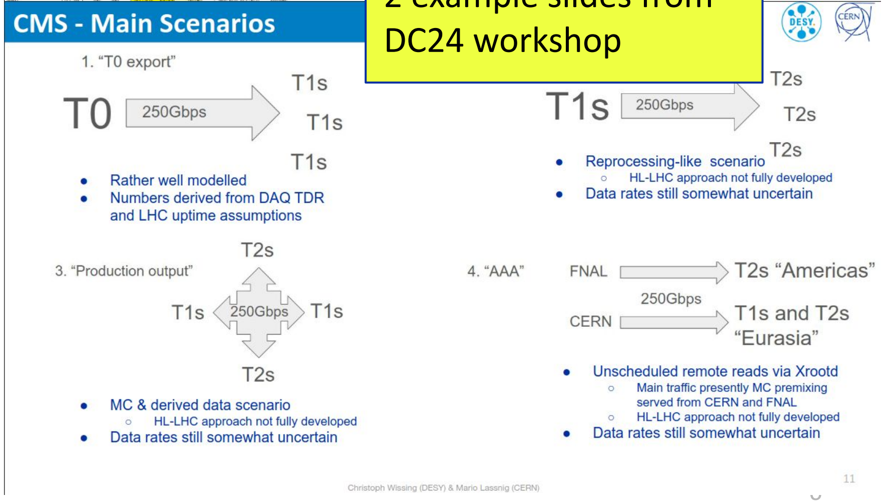
- Overall target: **25%** of HL-LHC throughput
  - Slightly lowered from originally 30% due to delayed start of HL-LHC
- Long way to towards the DC24 program
  - Agreement on dates
    - 2 weeks before beam operation in 2024
    - Full transfers from disk to disk
    - Not just network traffic
  - Experiments had room to define their goals
    - ALICE and LHCb involved tapes
    - ATLAS and CMS decided not to
  - Preparation of monitoring
  - Regular preparation started one year before
    - Monthly DOMA General checkpoints
    - Dedicated workshop in Nov 2023



ATLAS - Planning

- Rather detailed planning exists
- Rates are mainly scaled values from measured Run-3 values
- Sites are already informed about expected rates

2 example slides from DC24 workshop



CMS - Main Scenarios

1. "T0 export"

- Rather well modelled
- Numbers derived from DAQ TDR and LHC uptime assumptions

- Reprocessing-like scenario
  - HL-LHC approach not fully developed
- Data rates still somewhat uncertain

3. "Production output"

- MC & derived data scenario
  - HL-LHC approach not fully developed
- Data rates still somewhat uncertain

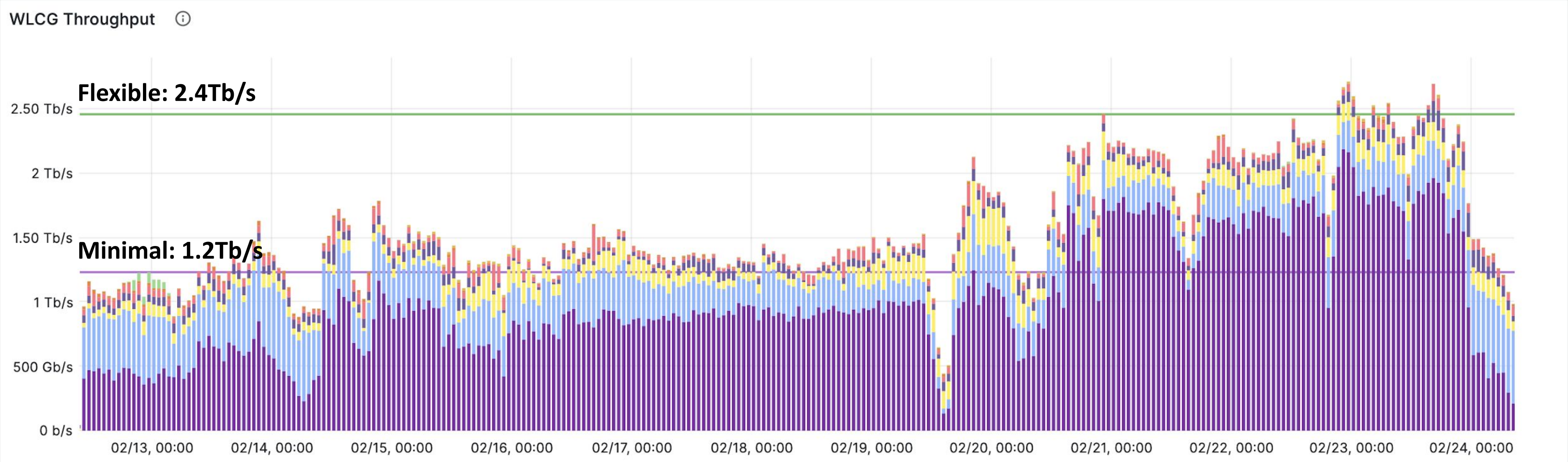
4. "AAA"

- Unscheduled remote reads via Xrootd
  - Main traffic presently MC premixing served from CERN and FNAL
  - HL-LHC approach not fully developed
- Data rates still somewhat uncertain

Christoph Wissing (DESY) & Mario Lassnig (CERN) 11

DC24 in one plot

WLCG Throughput



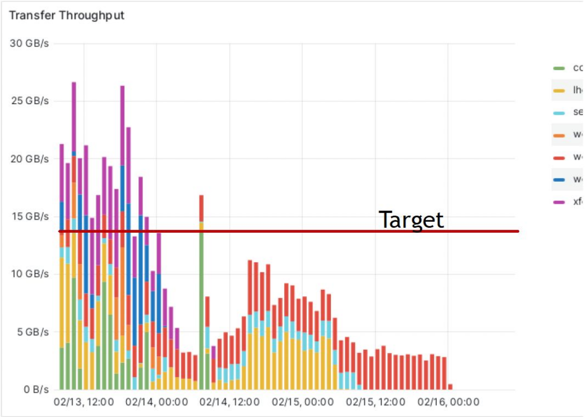
	max	avg	current
Data Challenge	2.19 Tb/s	1.02 Tb/s	211 Gb/s
atlas	625 Gb/s	304 Gb/s	567 Gb/s
alice xrootd	349 Gb/s	115 Gb/s	71.4 Gb/s
cms xrootd	191 Gb/s	67.4 Gb/s	42.7 Gb/s
cms	271 Gb/s	57.2 Gb/s	75.0 Gb/s
belle	38.9 Gb/s	9.45 Gb/s	17.1 Gb/s

DC24 met the (main) goals:

- **Achieved full throughput of minimal model (1st week)**
- **Push for flexible target (2nd week)**

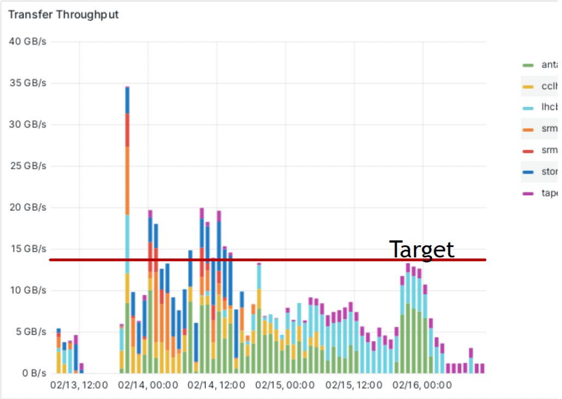
LHCb: Export from CERN to Tier-1s

EOS -> Disk link



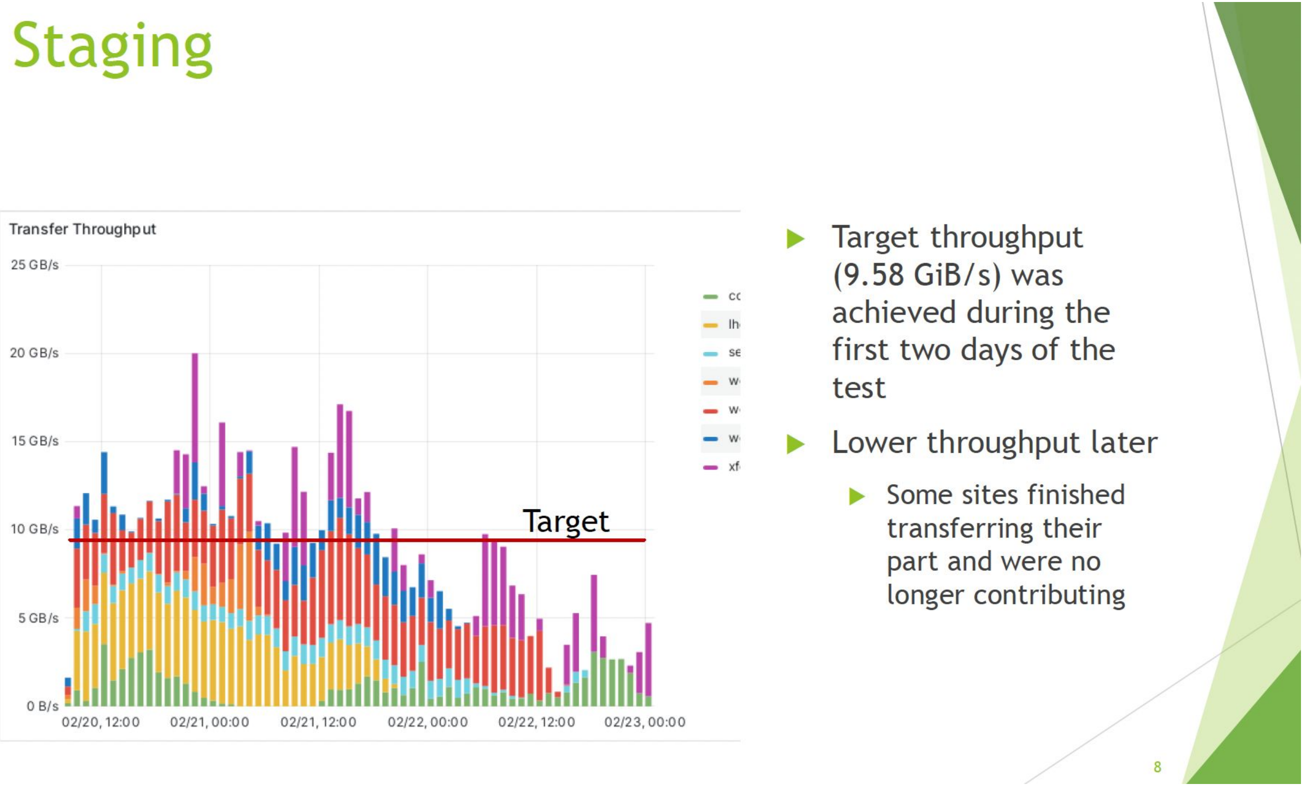
- Target throughput (14GiB/s) was achieved during the first day
- Lower throughput later
  - Some sites finished transferring their part during the first day so were no longer contributing to overall throughput
  - Submissions were slow and not optimal
  - Submission agent got stuck a few times, that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

LHCb: Staging exercise

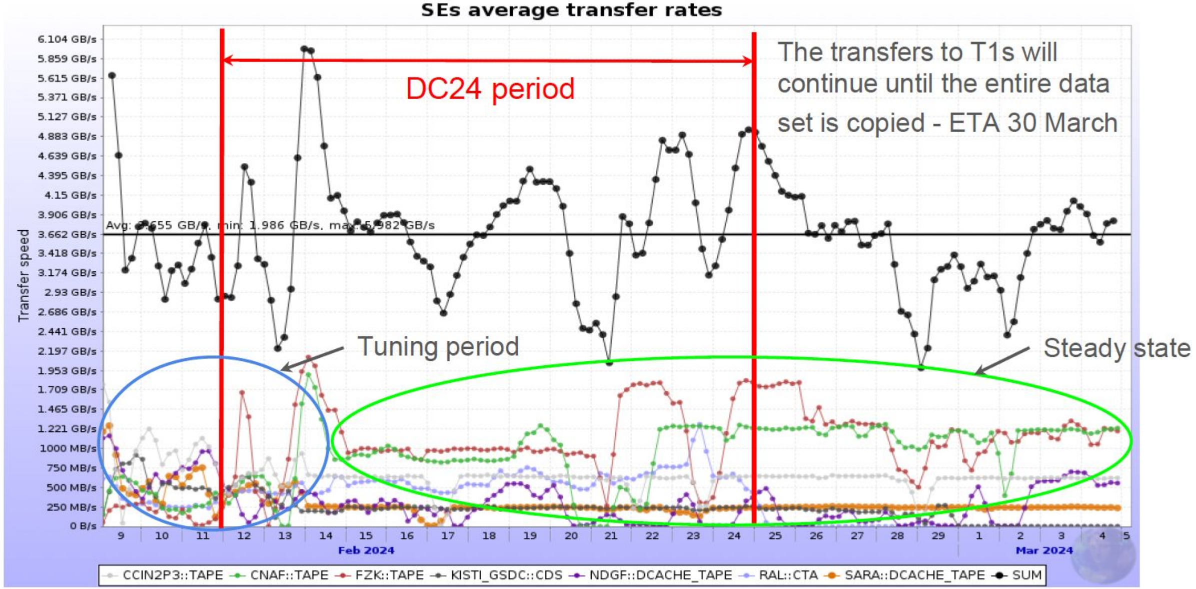


Staging

- 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

ALICE: Ongoing data export from 2023 HI run

Time evolution T1s



DC24 period

The transfers to T1s will continue until the entire data set is copied - ETA 30 March

Tuning period

Steady state

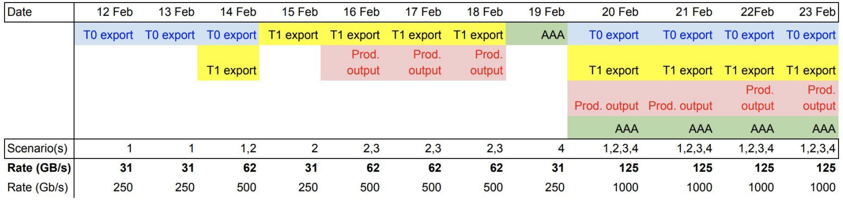
Centre	Target rate GB/s	Average achieved GB/s
CNAF	0.8	0.98 (+20%)
IN2P3	0.4	0.6 (+40%)
KISTI	0.2	0.25 (+22%)
GridKA	0.6	1.12 (+90%)
NDGF	0.3	0.35 (+15%)
NL-T1	0.1	0.25 (+150%)
RAL	0.1	0.58 (+500%)
CERN	*10*	*14.2 (+40%)*

CMS

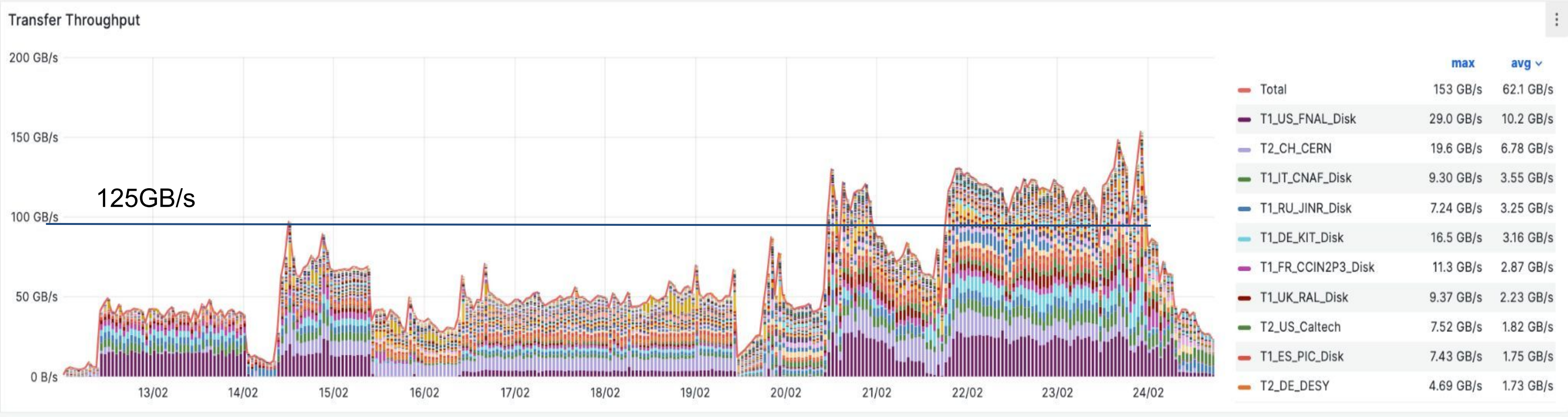
- Daily exercise menu with increasing complexity
- T0 export, T1s to T1s and T1s to T2s, AAA

- Overall target of ~125GB/s could be met
  - A few hundred links in total (Prod + DC)
  - Performance of individual links still under analysis

- Some limitation in 'deletion performance'
  - Tuning of Rucio deletion pods



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

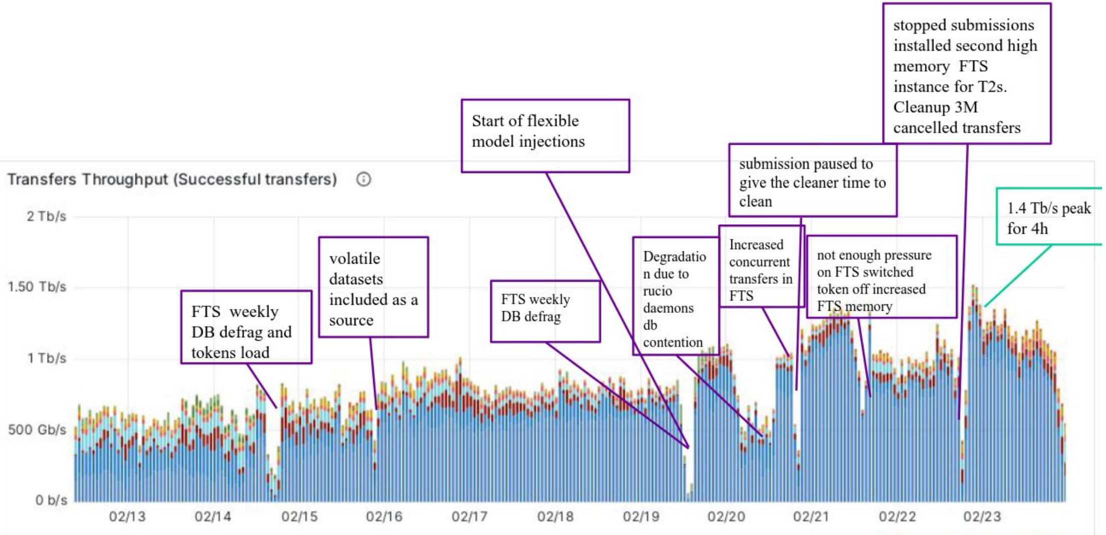


125GB/s

	max	avg
Total	153 GB/s	62.1 GB/s
T1_US_FNAL_Disk	29.0 GB/s	10.2 GB/s
T2_CH_CERN	19.6 GB/s	6.78 GB/s
T1_IT_CNAF_Disk	9.30 GB/s	3.55 GB/s
T1_RU_JINR_Disk	7.24 GB/s	3.25 GB/s
T1_DE_KIT_Disk	16.5 GB/s	3.16 GB/s
T1_FR_CCIN2P3_Disk	11.3 GB/s	2.87 GB/s
T1_UK_RAL_Disk	9.37 GB/s	2.23 GB/s
T2_US_Caltech	7.52 GB/s	1.82 GB/s
T1_ES_PIC_Disk	7.43 GB/s	1.75 GB/s
T2_DE_DESY	4.69 GB/s	1.73 GB/s

ATLAS

- Same "run scheme" as CMS
- Generally considered success, though with some homemade issues
    - 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
    - In the meantime, we have successfully re-run T0-T1 export tests



FTS weekly DB defrag and tokens load

volatile datasets included as a source

Start of flexible model injections

FTS weekly DB defrag

Degradation due to rucio daemons db contention

Increased concurrent transfers in FTS

submission paused to give the cleaner time to clean

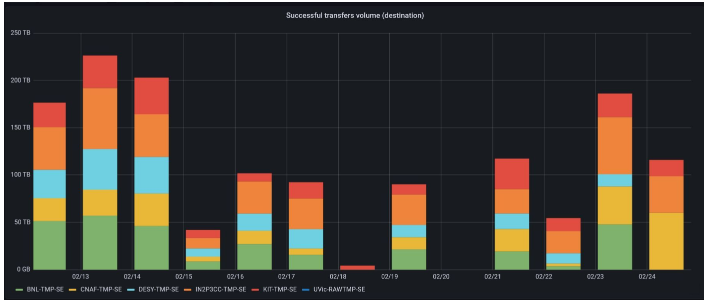
not enough pressure on FTS switched token off increased FTS memory

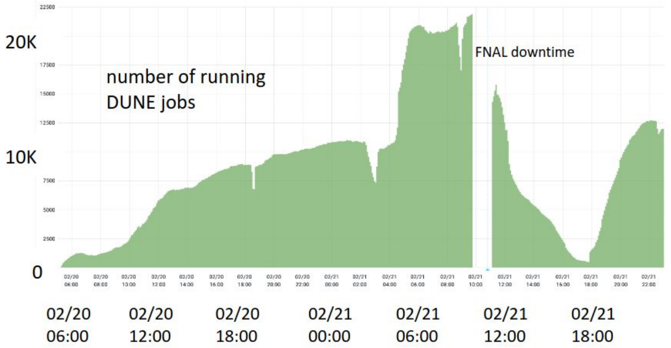
stopped submissions installed second high memory FTS instance for T2s. Cleanup 3M cancelled transfers

1.4 Tb/s peak for 4h

Belle II & DUNE

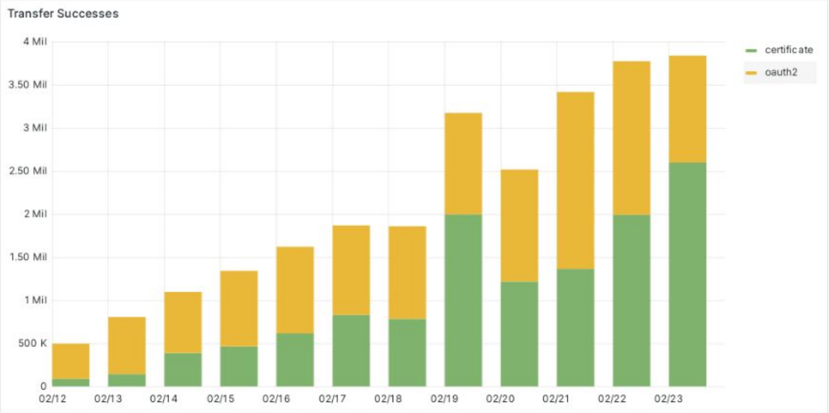
- Participation of non-LHC experiments in WLCG challenge for the first time
  - Belle II and DUNE fully included in planning process
  - Rates order of magnitude slower compared to LHC, flows often in opposite direction
- Belle II
  - Focus on traffic from KEK to RAW data centers and between RAW data centers
  - Targets were met
  - No obvious interference with LHC experiments
- DUNE
  - Focus on RAW data archiving & processing
  - Identified and improved some bottlenecks
  - Participation considered extremely useful





Token based Authentication

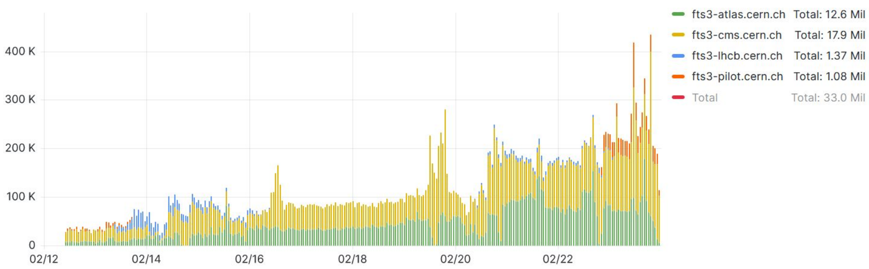
- Distributed infrastructure just became ready for DC24
  - FTS pre-release with token support
  - Rucio with base set of features for ATLAS and CMS
  - Deployment campaign to prepare storage elements
- About half of the transferred DC injected traffic via token authentication
  - Very high load on IAM for LHCb
    - Used 1 token per transfer
  - ATLAS switched tokens off at the end of 2nd week
    - Refresh very expensive for FTS
  - Valuable experiences gained with token usage at production scale
- Follow-up discussions in relevant forums to come
- **Dedicated talk on tokens by Maarten in this session!**



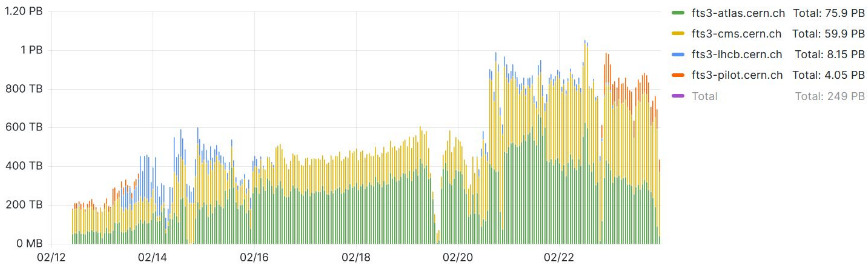
FTS operating at unprecedented scales

- Particularly FTS ATLAS instance survived thanks to permanent baby sitting by FTS team
    - ○ Database surgery in production
    - ○ Increase of hardware resources
- Improved understanding of current FTS scaling
    - ○ Optimizer cycle needed several hours
    - ○ FTS has no concept of storage back pressure
    - ○ FTS treats all links with the same activity with equal priority
- FTS team started to iterate developments items and related priorities with stakeholders of the community
- First official FTS release with token support this spring

DC24 file transfers per FTS instance per hour



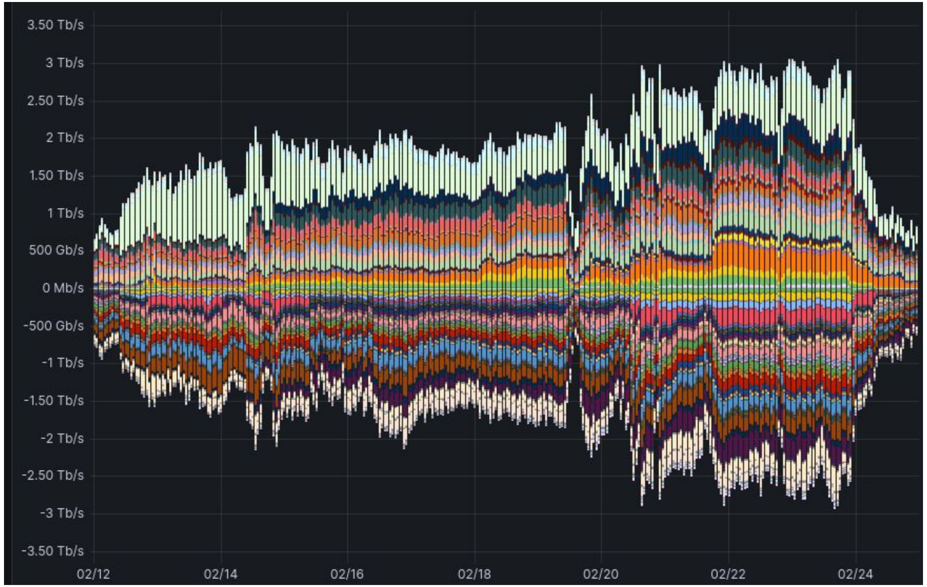
DC24 data volume transfered per hour



Plots show DC injected 'activity' only
Parallel ongoing production not included!

Beyond throughput

- WLCG DCs should also (scale) test new technologies
  - Deployment can vary depending on level of matureness
- Some technical topics addressed in the context of DC24
  - Measures to improve monitoring
    - Site based network monitoring
      - Captures all traffic
    - Network flow marking
      - with *SciTags* and UDP *Fireflies*
  - Software Defined Networking (SDN)
    - NOTED
    - SENSE/Rucio
  - Low level network stack
    - Jumbo frames
    - BBRv2, BBRv3 TCP stacks



After the Challenge is before the next Challenge

- Aftermath of DC24
  - Derive 'lessons learned'
    - What went well, where were bottlenecks, organizational improvements …
  - Set priorities of for ongoing developments
    - VO & community specific tools, e.g. Rucio, FTS,
    - Storage middleware
    - Network equipment
- Planning of next DC
  - So far nothing is set except the global target of **about 50%** of expected HL-LHC throughput
  - Dates
    - Likely in 2026 or even later
    - Almost for sure in LS3, which makes scheduling much easier for LHC experiments
  - Participating experiments
    - LHC experiments, hopefully again Belle-2 and DUNE
    - Interest (already expressed during DC24) by JUNO, SKA, Neutrino experiments in Japan
  - Experience shows that planning needs to start early (1 year before at least)

Some random preliminary observations & remarks

- There are other bottlenecks than network bandwidth
  - Maintenance of DC injections was challenging
    - FTS instances got pushed to their limits, particular the ATLAS one
    - Keeping up with deletions is not trivial, systems not designed for best scaling here
    - Already ideas how to integrate data injector natively into Rucio
  - It needs time before a complex system reacts to parameter changes
    - The parameter space is huge
    - Not many attempts to re-adjust (very few per day)
  - A number of CMS sites asked for more (than planned) traffic to exercise their WAN connectivity

Final report

- To be delivered in time for the DESY Workshop **(NO EXTENSION!)**
  - Pre-structured document is here
  - If you were involved in DC24, please fill the appropriate sections
  - We would like to close edits two weeks before the workshop
    - **End of April!**
  - CW & ML will then edit the final document

Joint WLCG/HSF workshop at DESY

- May 13-17th in Hamburg
- Topics include data challenges, analysis facilitates, software tools and training
- Lots of details and information is available now on the workshop indico
- Book your accommodation now, DESY Hostel rooms are still available
- Dinner at the Altes Mädchen Craft Beer Brewery and Restaurant
- Registration is now open: 250€ until April 8th, rising to 275€, registration closes on April 26th

What AI thinks we are doing …

Bing Image Creator: "Worldwide LHC Computing Grid, Data Challenge Workshop, Happy Mood"

Bing Image Creator: "Worldwide LHC Computing Grid, Data Challenge Workshop, Serious Mood"