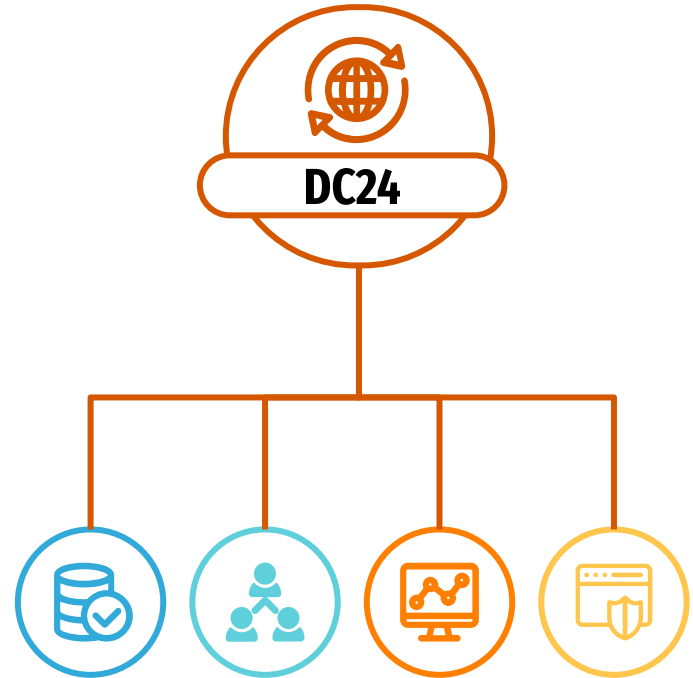


WLCG Data Challenge 24

Katy Ellis, on behalf of the DC
Community

CHEP 2024, 23/10/24



WLCG - Worldwide LHC Computing Grid



Global collaboration of
~170 computing centre in
40+ countries



Provide resources to store,
distribute and analyse
data



Raw data comes from
LHC experiments: ATLAS,
CMS, LHCb, ALICE



Manages grid-wide
operations and
deployments

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

WLCG Data Management

Copy raw data from CERN disk

Ensure multiple copies and free disk space at CERN



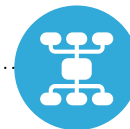
Job data placement

Ensure jobs have access to input data, copy output data and rebalancing



Authentication

Protect data appropriately, giving access to authorised users



Network bandwidth

Sufficient for expected rates



Storage ingress/egress

Able to write and read data files at expected rates



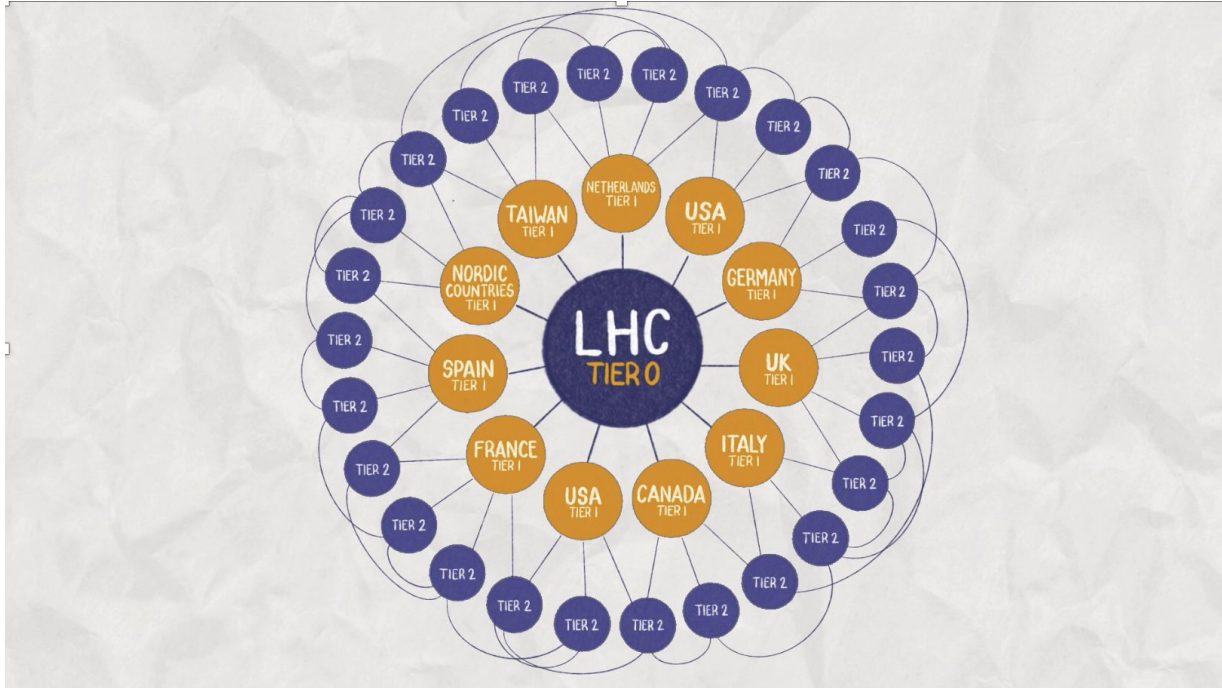
Transfer software

Able to feed the system at expected rates

Distributed computing

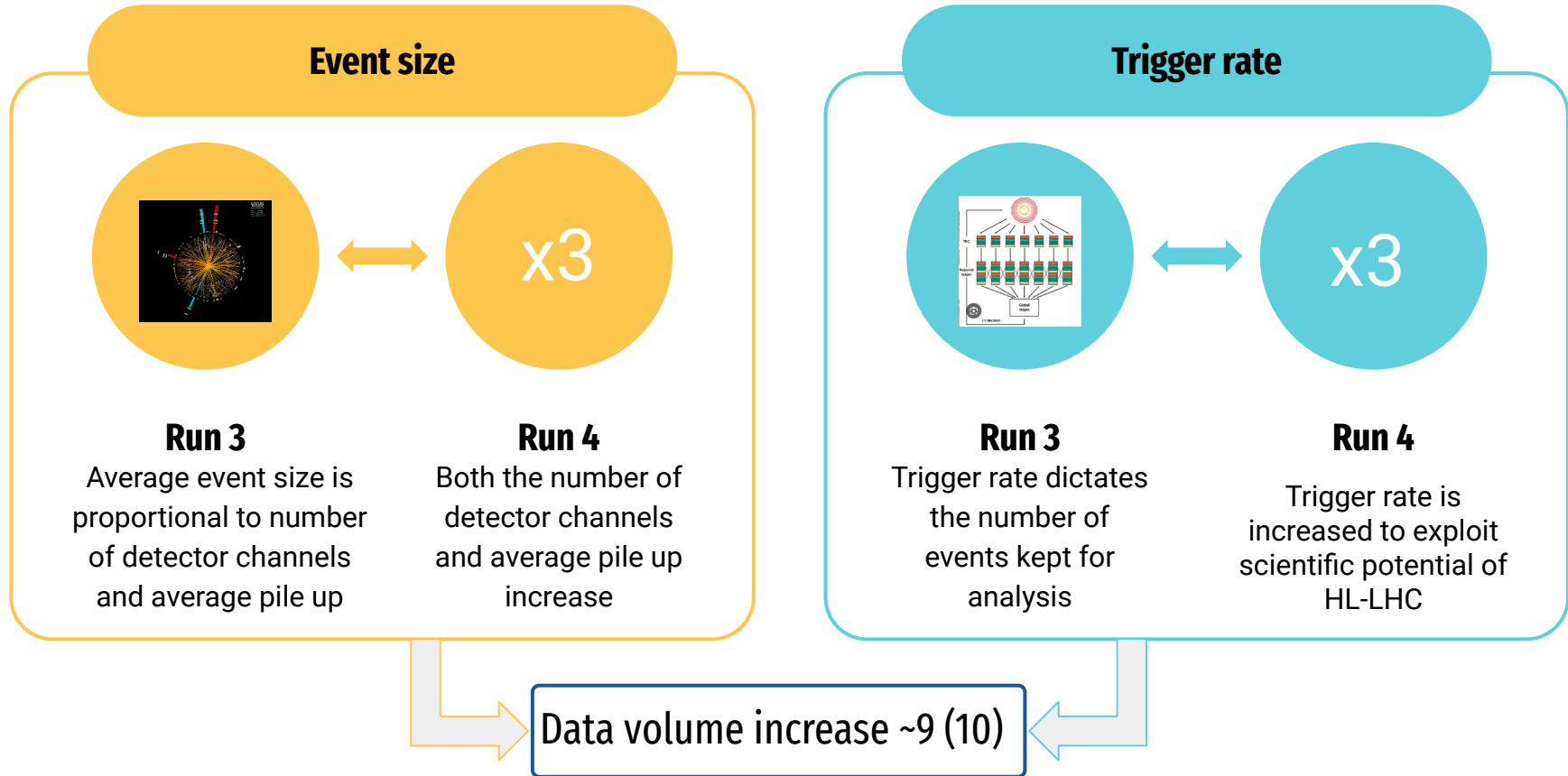


Distributed computing

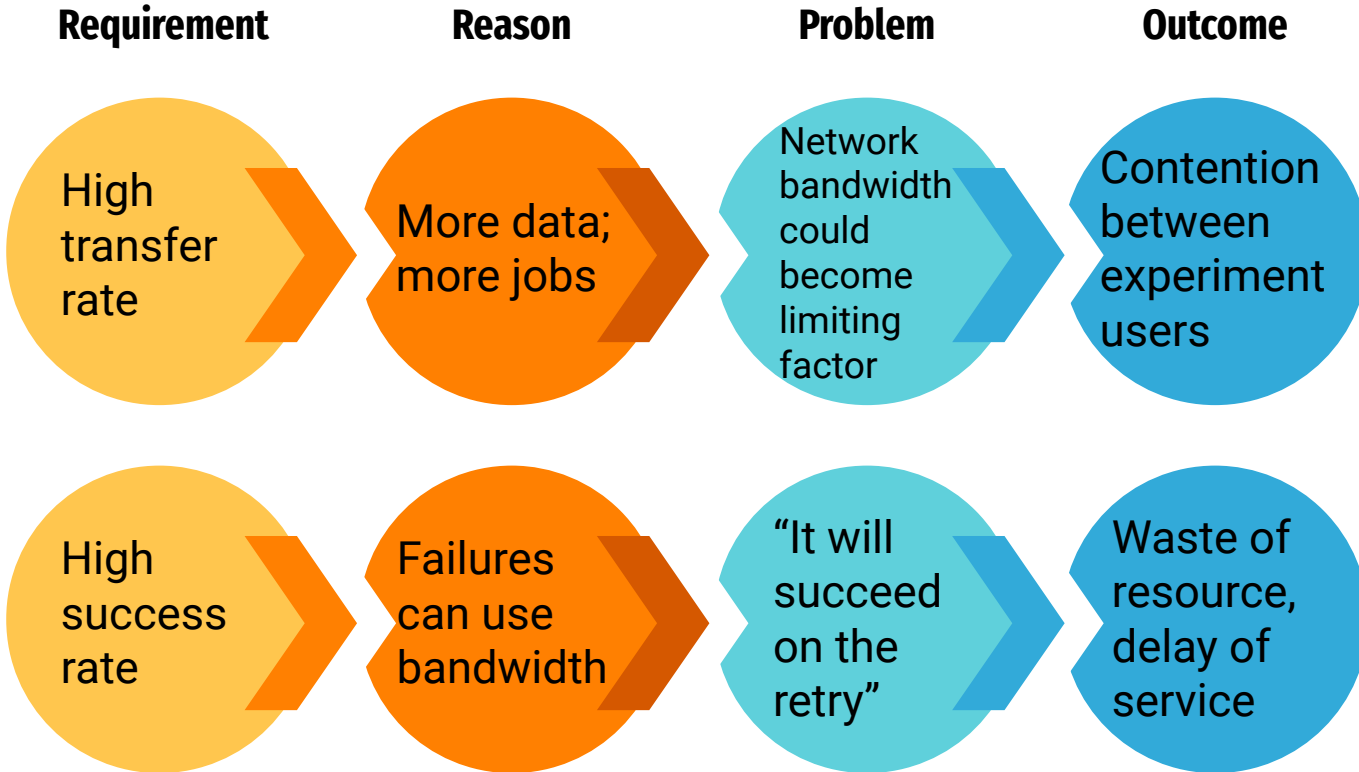


The networking is now a full mesh where all sites can talk to each other via a sophisticated network irrespective of tier or region.
Network provision tends to be ahead of the experiment requirements.

Expected increase in data volume (ATLAS and CMS)



Data movement efficiency



WLCG Data Challenges

WLCG spoke! And declared that a series of data challenges should be run...

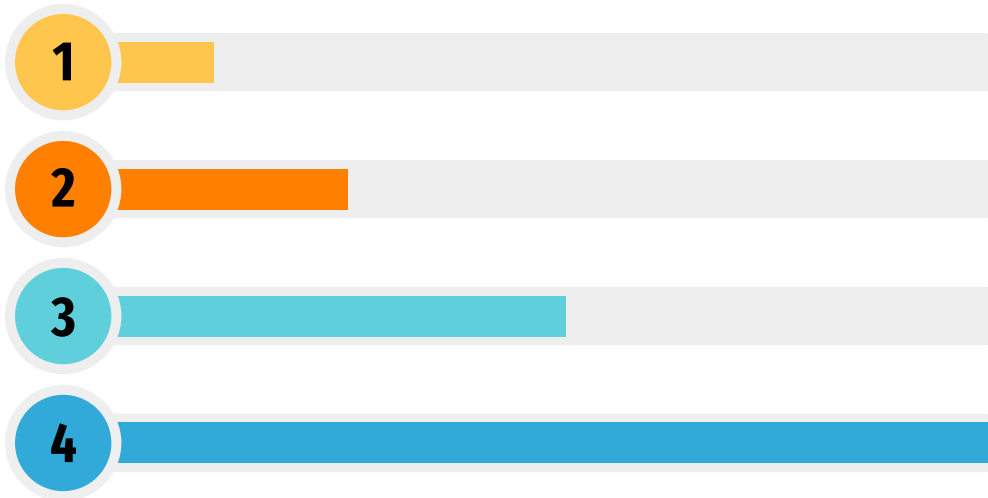


DCs - why do we need them?

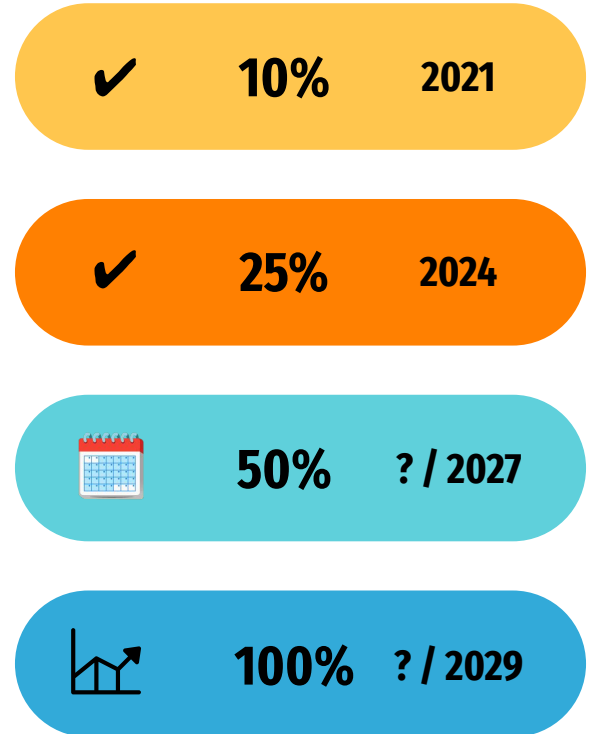


Data challenge series

Proportion of estimated Run 4 rates



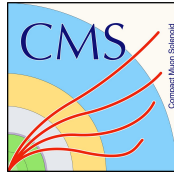
Run 4 is scheduled to start in 2030



DC 24 participants



Typically generates most data and highest current network user



Large-scale network user



Smaller-scale user; Run 4 usage similar to Run 3



ALICE

Smaller-scale user; Run 4 usage similar to Run 3



Small-scale users; using many of the same storage sites as LHC experiments but raw data source is not CERN

Data challenge (DC) specification

[Technical note from WLCG \(2021\)](#)

- For ATLAS and CMS in Run 4:
 - 350PB of raw data exported from CERN to Tier 1s in quasi-real time of 7 million seconds
 - This is 50GB/s == 400Gb/s
 - Plus additional data formats 100Gb/s => 500Gb/s for each ATLAS and CMS
- For LHCb and ALICE in Run 4:
 - 100Gb/s each estimated raw data exported from CERN

Data challenge (DC) specification

[Technical note from WLCG \(2021\)](#)

- For ATLAS and CMS in Run 4:
 - 350PB of raw data exported from CERN to Tier 1s in quasi-real time of 7 million seconds
 - This is 50GB/s == 400Gb/s
 - Plus additional data formats 100Gb/s => 500Gb/s for each ATLAS and CMS
- For LHCb and ALICE in Run 4:
 - 100Gb/s each estimated raw data exported from CERN
- Plus the same rate again for export of same data from Tier 1s to Tier 2s
- DC values should be doubled to allow for 'bursty' nature
- Networks should be provisioned at double the bursty rate...but data challenges are not required to fill them
- The document notes likely uncertainties in the numbers

Data challenge (DC) specification

[Technical note from WLCG \(2021\)](#)

- For ATLAS and CMS in Run 4:
 - 350PB of raw data exported from CERN to Tier 1s in quasi-real time of 7 million seconds
 - This is 50GB/s == 400Gb/s
 - Plus additional data formats 100Gb/s => 500Gb/s for each ATLAS and CMS
- For LHCb and ALICE in Run 4:
 - 100Gb/s each estimated raw data exported from CERN
- Plus the same rate again for export of same data from Tier 1s to Tier 2s
- DC values should be doubled to allow for 'bursty' nature
- Networks should be provisioned at double the bursty rate...but data challenges are not required to fill them
- The document notes likely uncertainties in the numbers

From experiment point of view

- What about other significant data flows, e.g. simulated (MC) data?

Data challenge (DC) specification

Technical note from WLCG (2021)

- For ATLAS and CMS in Run 4:
 - 350PB of raw data exported from CERN to Tier 1s in quasi-real time of 7 million seconds
 - This is 50GB/s == 400Gb/s
 - Plus additional data formats: 100Gb/s == 500Gb/s for each ATLAS and CMS
- For LHCb a
 - 100Gb/s
- Plus the same rate again for export of same data from Tier 1s to Tier 2s
- DC values should be doubled to allow for 'bursty' nature
- Networks should be provisioned at double the bursty rate...but data challenges are not required to fill them
- The document notes likely uncertainties in the numbers

“Minimal model”

From experimenter

- What about

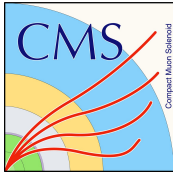
“Flexible model”

ed (MC) data?

(Additional) rate estimates for Run 4



Assumes Run 3 network usage represents 10% of Run 4 usage;
Detailed modelling on per-link basis. Flexible model target = 1.25Tb/s



Estimates that simulated data movement at least as big as raw data;
plus same again for AAA remote reads. Flexible model target = 1.0Tb/s



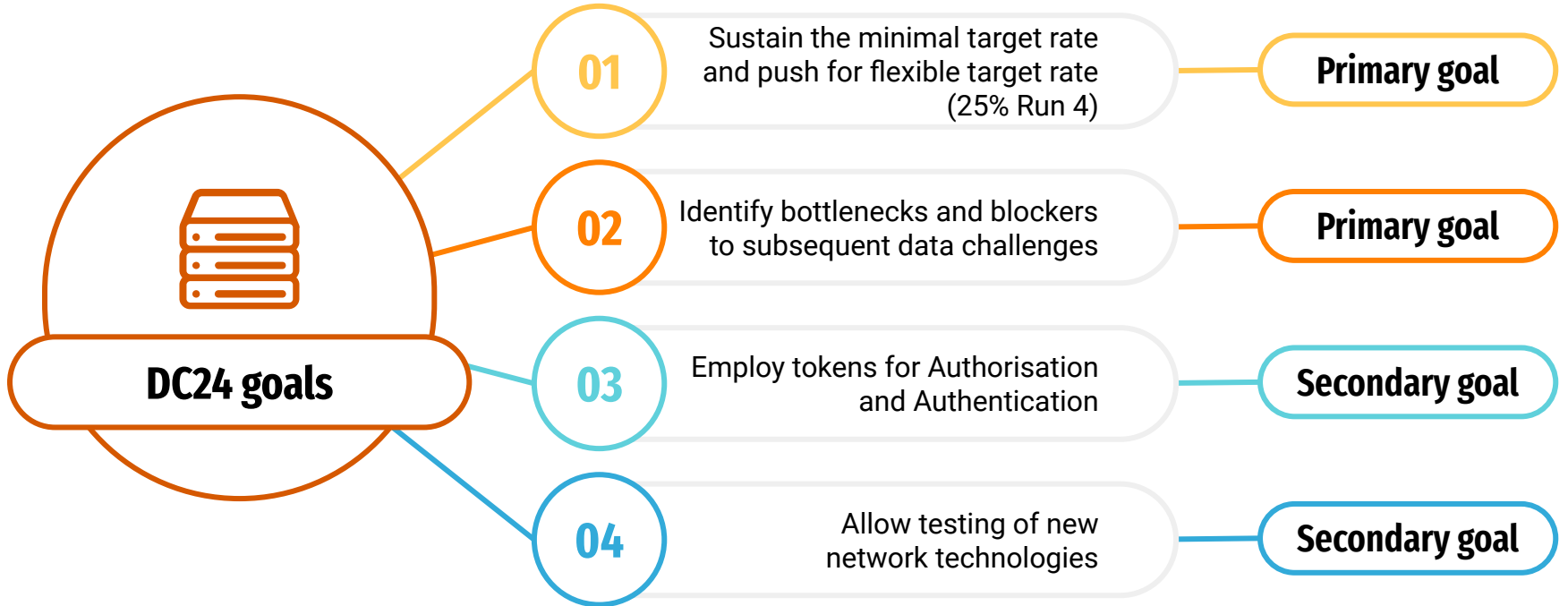
Simulate the high-lumi scenario (40TB/day) moving data from KEK to
data centres in Europe and North America. Target = 18Gb/s



Used the period to test entire 'keep up processing' workflow including data
movement; 25% of raw data rate from Far Detector (8Gb/s)

Learn more about the [ATLAS](#), [CMS](#) and [Belle II](#) rate estimates and other
experiment-specific details in other contributions this week

DC24 goals



DC24 timeline



Systems used during DC24

- Experiments should use their standard production system tools if possible

Experiment	Direct transfer	FTS?	Rucio?	Auth	Tape transfer?
ALICE	XRootD			ALICE token	✓
LHCb		✓		WLCG token	✓
ATLAS		✓	✓	Token & cert	
CMS		✓	✓	Token & cert	
DUNE		✓	✓	X509 cert	
Belle II		✓	✓	X509 cert	

Large-scale data movement using Rucio

dc_inject tool

A bespoke script to continuously inject rules into Rucio based on individual links and requested rates

01



02

Rucio

Scalable data management software submits transfers to FTS and handles deletions



03

FTS

File Transfer Service issues commands to move data between sites

04



Monitor

Human operators monitor the rate of transfers and alter the input to the dc_inject tool

Major improvements in monitoring since DC21



Improved joint monitoring dashboard, incorporating input from all DC24 experiments, including those not using FTS
Able to distinguish DC traffic from production for larger experiments



New network monitoring developed and deployed at many sites
Gives the experiment and network teams eyes on what is happening at the sites, where there may be multiple activities ongoing



SciTags enabled for some sites
Will allow future network monitoring to split the traffic by VO and activity



New XRootD throughput monitoring 'Shoveler'
Improves on the previous XRootD monitoring; however this was not validated in time for DC24

Pre-DC24 tests



Test ingress at individual Tier 1s

Set expectations and give high-pressure sites the chance to make improvements



Test egress from CERN

Gain confidence that CERN can serve data at the required rate



Rehearse tools and the team

Allow teams to use the dc_inject tool, putting pressure on Rucio and FTS

Basic token testing

Tokens not used in production before DC24!



**Build relationships
with stakeholders**

DC24 schedule

	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		

Running the challenge

ATLAS and CMS:

- Week 1 - not too difficult; initial problem with FTS/token interaction
- Week 2 - push for flexible rate required constant baby-sitting

LHCb:

- Struggled to keep up continuous rate during the first week

ALICE:

- After tuning period all went smoothly

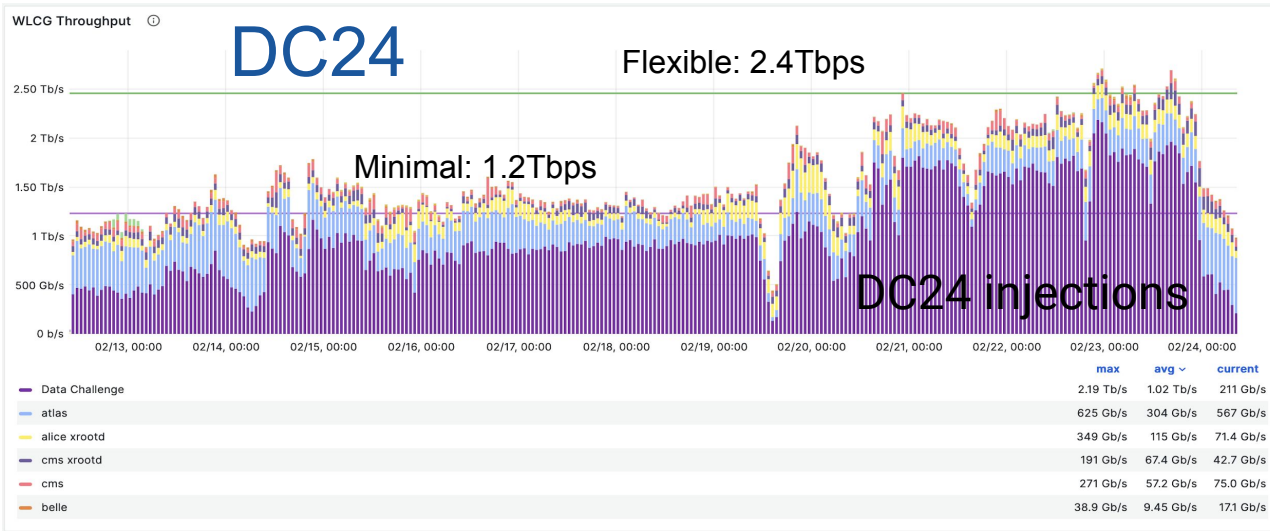
DUNE:

- Significant progress on Rucio setup and operation

Belle II:

- After solving issues with deletions and tuning the FTS parameters, everything went smoothly

Main Result



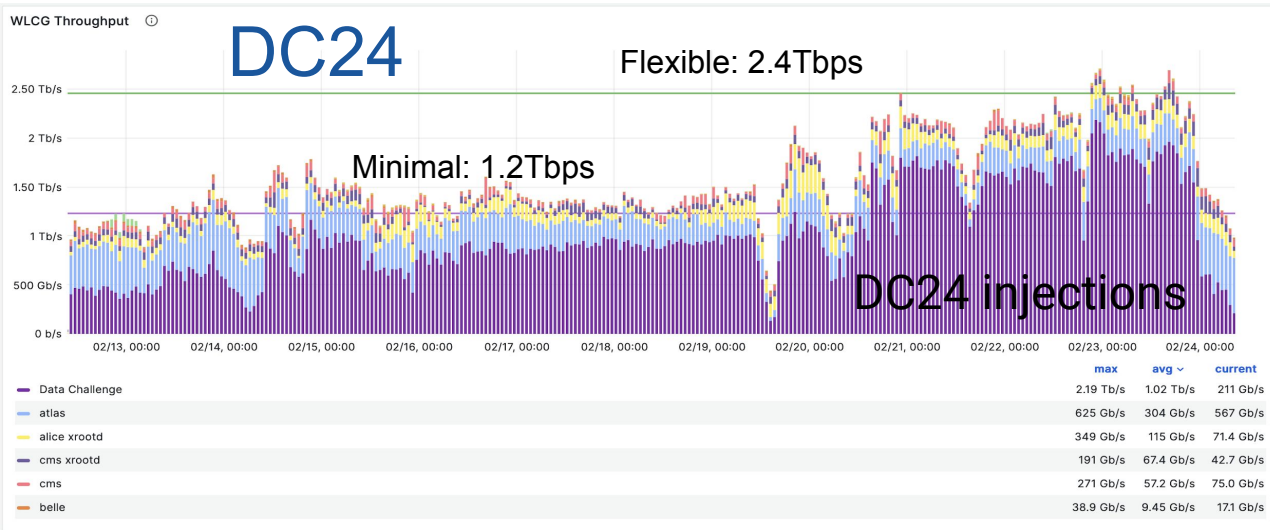
DC24 achieved the main goal:

- Full throughput of minimal model (week 1)
- Push for flexible target (week 2)

Data Challenges 2024 report:

<https://zenodo.org/records/11444180>

DC24 vs DC21



DC24 achieved the main goal:

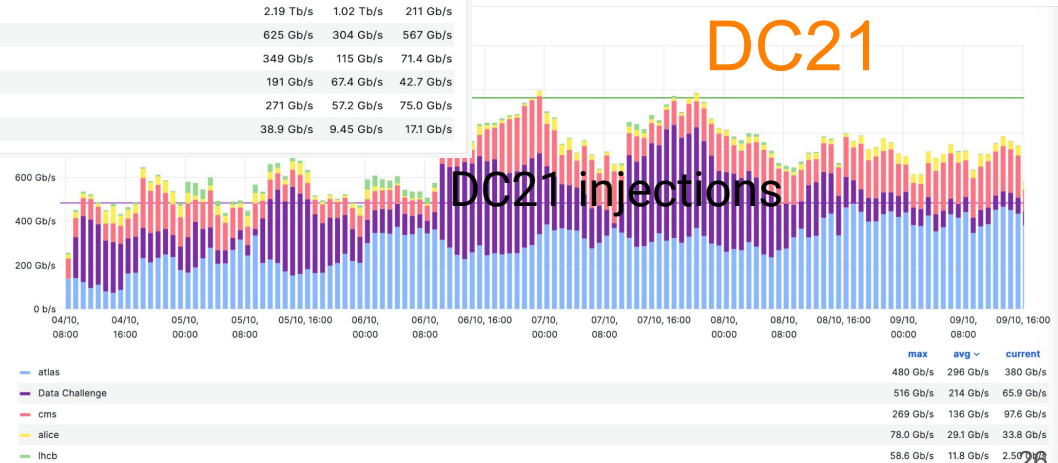
- Full throughput of minimal model (week 1)
- Push for flexible target (week 2)

Data Challenges 2024 report:

<https://zenodo.org/records/11444180>

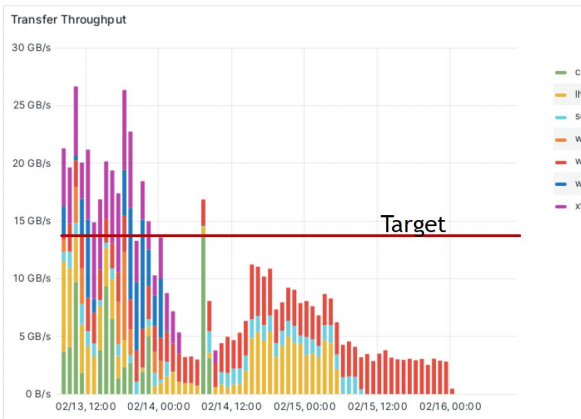
Data Challenges 2021 report:

<https://zenodo.org/records/5767913>



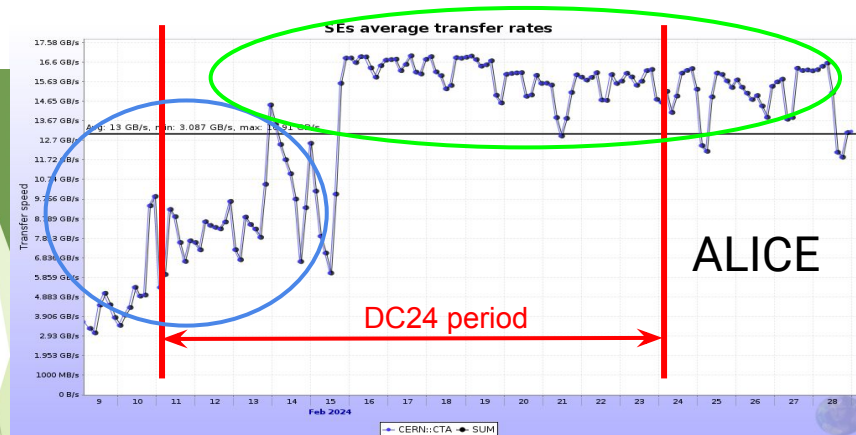
Summary results from LHCb and ALICE

EOS -> Disk link



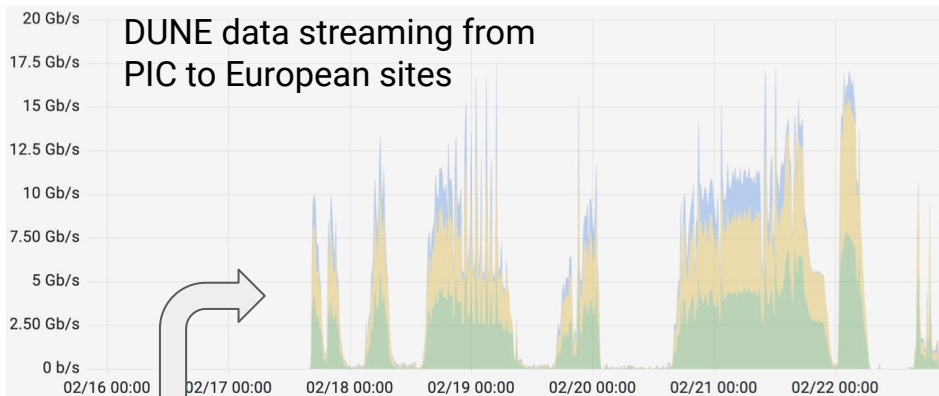
- ▶ Target throughput (14GiB/s) was achieved during the first day
- ▶ Lower throughput later
 - ▶ Some sites finished transferring their part during the first day so were no longer contributing to overall throughput
 - ▶ Submissions were slow and not optimal
 - ▶ Submission agent got stuck a few times,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

3



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%)
<i>CERN</i>	<i>10</i>	<i>14.2 (+40%)</i>

Summary results from DUNE and Belle II



DUNE used the DC24 period to exercise complete workflows - moving data AND running jobs (“keep-up processing”) A lot of progress made on the production system as a whole, including major steps forward in the setup of Rucio

Belle II simulated a high-lumi scenario 40TB per day
Transfers from KEK to RAW Data Centers according to the distribution schema (30%BNL, 20%CNAF, 15%IN2P3CC, 15% UVic, 10%DESY, 10% KIT)



Technical challenges and bottlenecks identified

(ATLAS, CMS, LHCb)



Submissions in FTS

Sustaining the rate with manageable entries in FTS

FTS tuning and optimiser

What is optimum number of parallel transfers on each link?

FTS prioritisation

Better system-wide throughput if FTS processed 'fast' links first

File Deletions

Deletions must keep up with transfers; can sites and Rucio keep up?

Token refresh rate

Can the token provider issue sufficient tokens for experiment requests?

[FTS tokens talk](#)

Focus on Tier 1s

- DC24 had a focus on **Tier 1** disk endpoints
 - This was revealing - some sites worked well as a source but not as a destination, or vice versa.
 - Even if they did well in the pre-testing
 - LHCb also tested Tier 1 tape endpoints, using local disk as a via point
- Some Tier 2 sites were disappointed with their observed rates
 - ATLAS and CMS would have liked to have pushed more on Tier 2s, however:
 - They were protecting their FTS instances
 - Focus was on keeping up the rates at the Tier 1s and there was only so much time in the day

Networks and new technologies

Thanks to significant preparation by the experts, the network was not a bottleneck during DC24!

Note that the experiments do not make requests for network capacity

More information on new technologies in the [DC24 Final Report](#)

Network routing

Flow labelling and packet tagging: Fireflies and SciTags

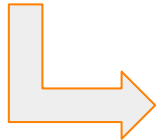
Load-balancing between networks: NOTED

Software Defined Networking in Rucio: [SENSE](#)

IPv6

TCP congestion protocols: BBRv1 vs CUBIC

Also see CHEP24
talks on [Network
Analytics with ML](#)
and [SciTags](#)



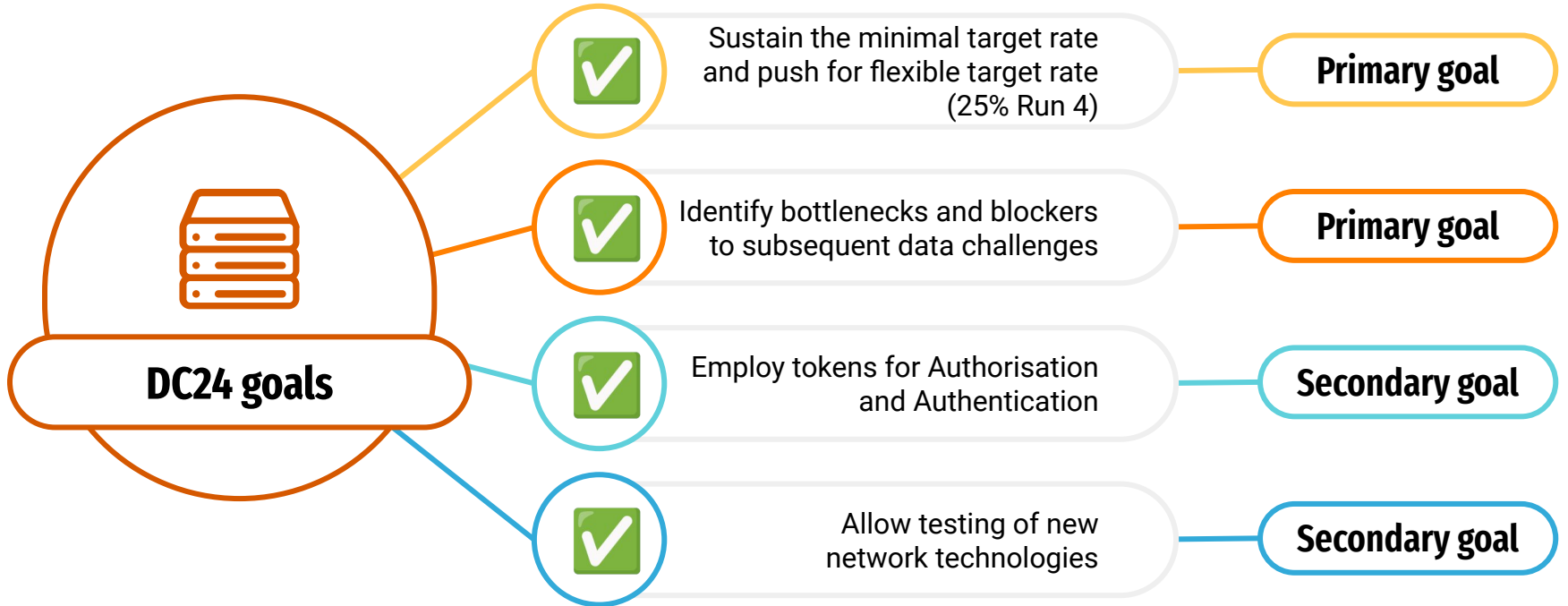
Significant research ongoing, but difficult to demonstrate effectiveness when **network was not congested!**

And tokens..?

Around 50% of DC24 transfers were completed using the new token auth



DC24 goals



Summary

 DC24 was a success!



What did we learn?

ATLAS and CMS could push systems up to the 25% of Run 4 estimates...but only just...ATLAS found the limits of their FTS
Other experiments achieved required performance even when the system was busy
Current network provision was sufficient
Rucio scaled well



What were the benefits?

Improvements to FTS are already being made, with others in the pipeline
Gave Tier 1 sites the chance to see their system under pressure and make changes
Brought the communities together - experiments, sites, storage and network experts



What plans do we have?

Already thinking about the third WLCG Data Challenge!
50% of the Run 4 will be the target
Even more scientific communities are interested in joining



What changes for next time?

Big jump in target rates
Include tape test for all experiments
More emphasis on Tier 2 sites
All transfers using tokens for auth

Instructions for use (free users)

In order to use this template, you must credit [Slidesgo](#) by keeping the Thanks slide.

You are allowed to:

- Modify this template.
- Use it for both personal and commercial purposes.

You are not allowed to:

- Sublicense, sell or rent any of Slidesgo Content (or a modified version of Slidesgo Content).
- Distribute this Slidesgo Template (or a modified version of this Slidesgo Template) or include it in a database or in any other product or service that offers downloadable images, icons or presentations that may be subject to distribution or resale.
- Use any of the elements that are part of this Slidesgo Template in an isolated and separated way from this Template.
- Delete the “Thanks” or “Credits” slide.
- Register any of the elements that are part of this template as a trademark or logo, or register it as a work in an intellectual property registry or similar.

For more information about editing slides, please read our FAQs or visit Slidesgo School:

<https://slidesgo.com/faqs> and <https://slidesgo.com/slidesgo-school>