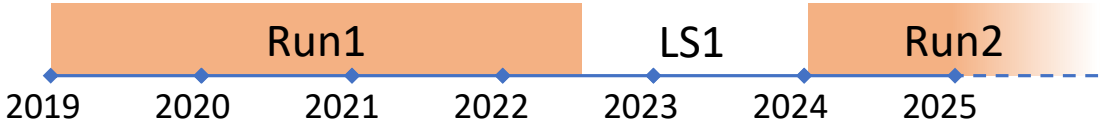
The Belle II Raw Data Transfer System

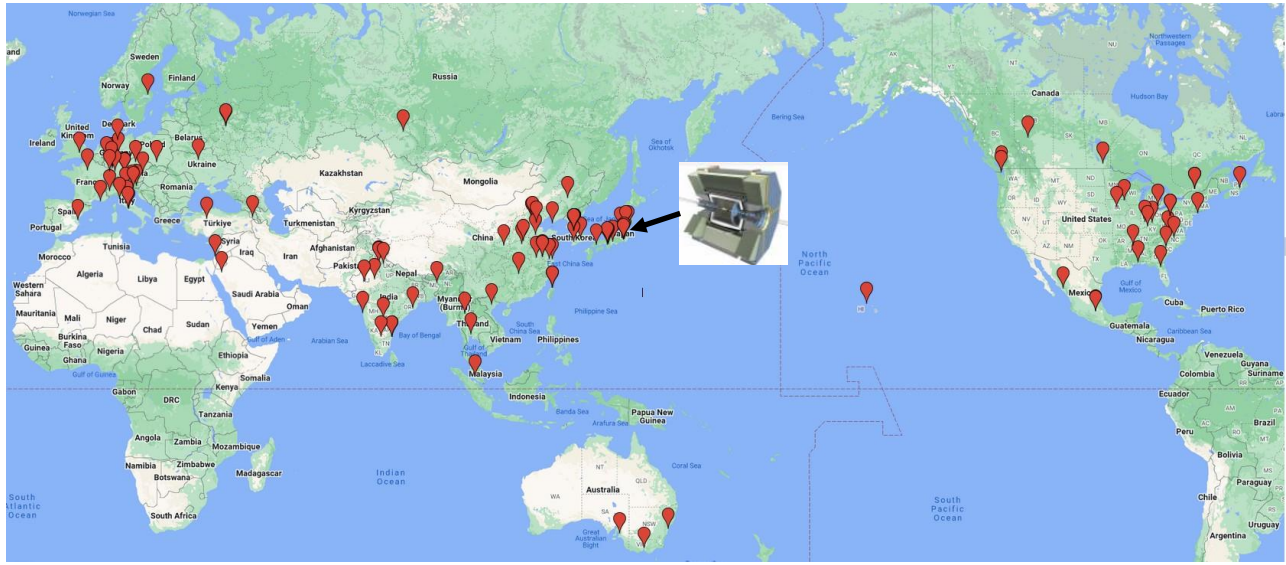
Matt Barrett[1], Tristan Bloomfield[1], Hara Takanori[1,2], Dhiraj Kalita[1], Cedric Serfon[3], Ueda Ikuo[1]

[1]High Energy Accelerator Research Organization (KEK)
[2]The Graduate University for Advanced Studies (SOKENDAI)
[3]Brookhaven National Laboratory (BNL)

The Belle II Experiment

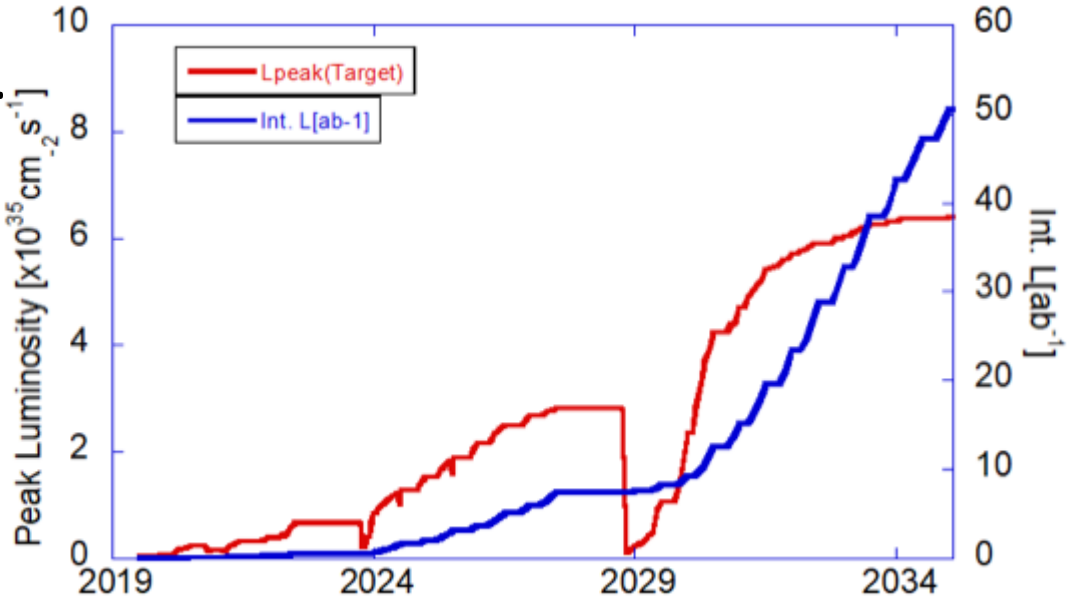
- Belle II is a particle physics experiment located at the KEK laboratory in Tsukuba, Japan.
- The Belle II collaboration is made up of:
  - >1000 members
  - from over 100 universities and research institutions
  - in 28 countries.
- Operation started 2019.
  - Recently ended 1.5 year Long Shutdown 1.



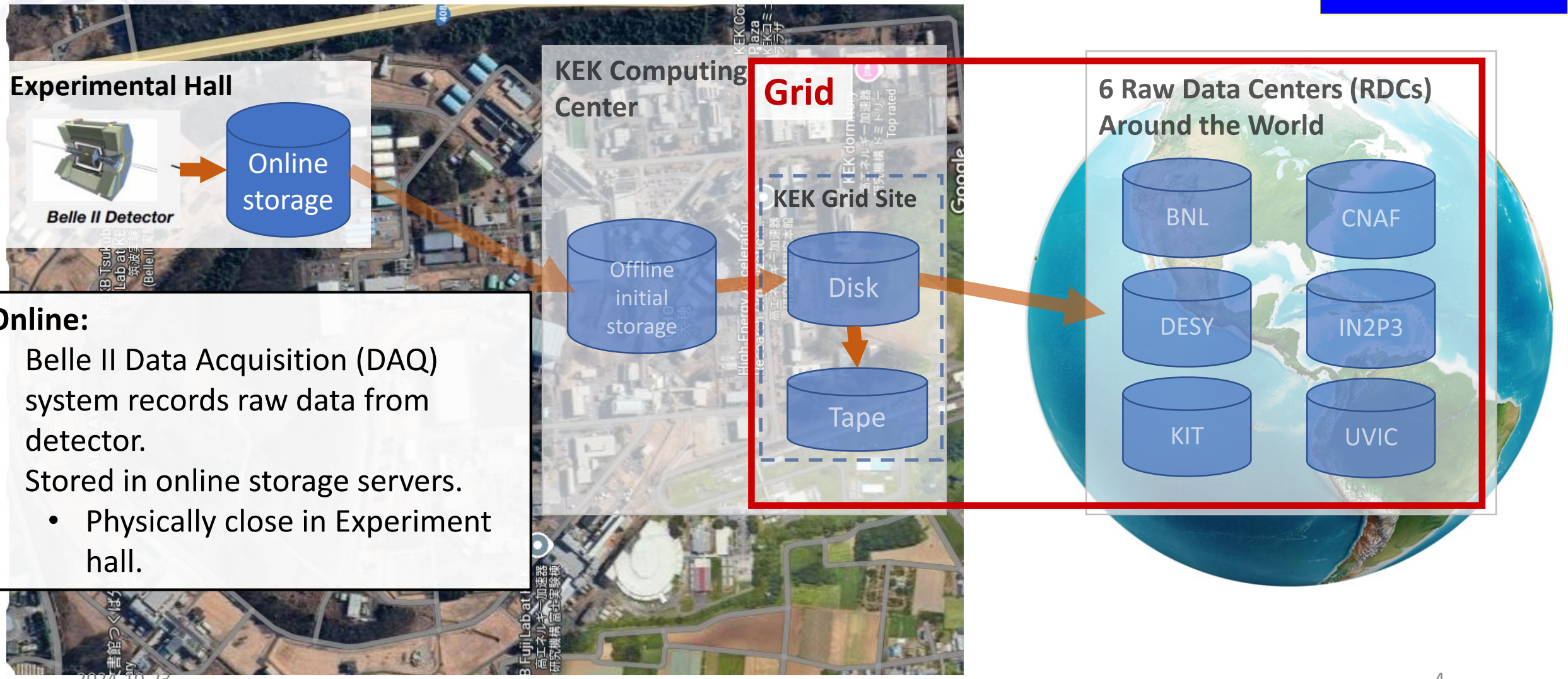


The Belle II Experiment

- Belle II aims to collect a dataset of 50 ab^{-1} (50x Belle)
  - Corresponds to approximately 60 PB.
  - Data rate (proportional to red line in plot) to increase significantly in later years.
- Due to data size and wide collaborator distribution, Belle II uses a grid computing model:
  - DIRAC main framework, Rucio data management.
  - 26 grid sites with storage.
- Distributed raw data storage.
  - 2 geographically separate copies for data redundancy.
  - Dedicated DIRAC extension for raw data, BelleRawDIRAC.



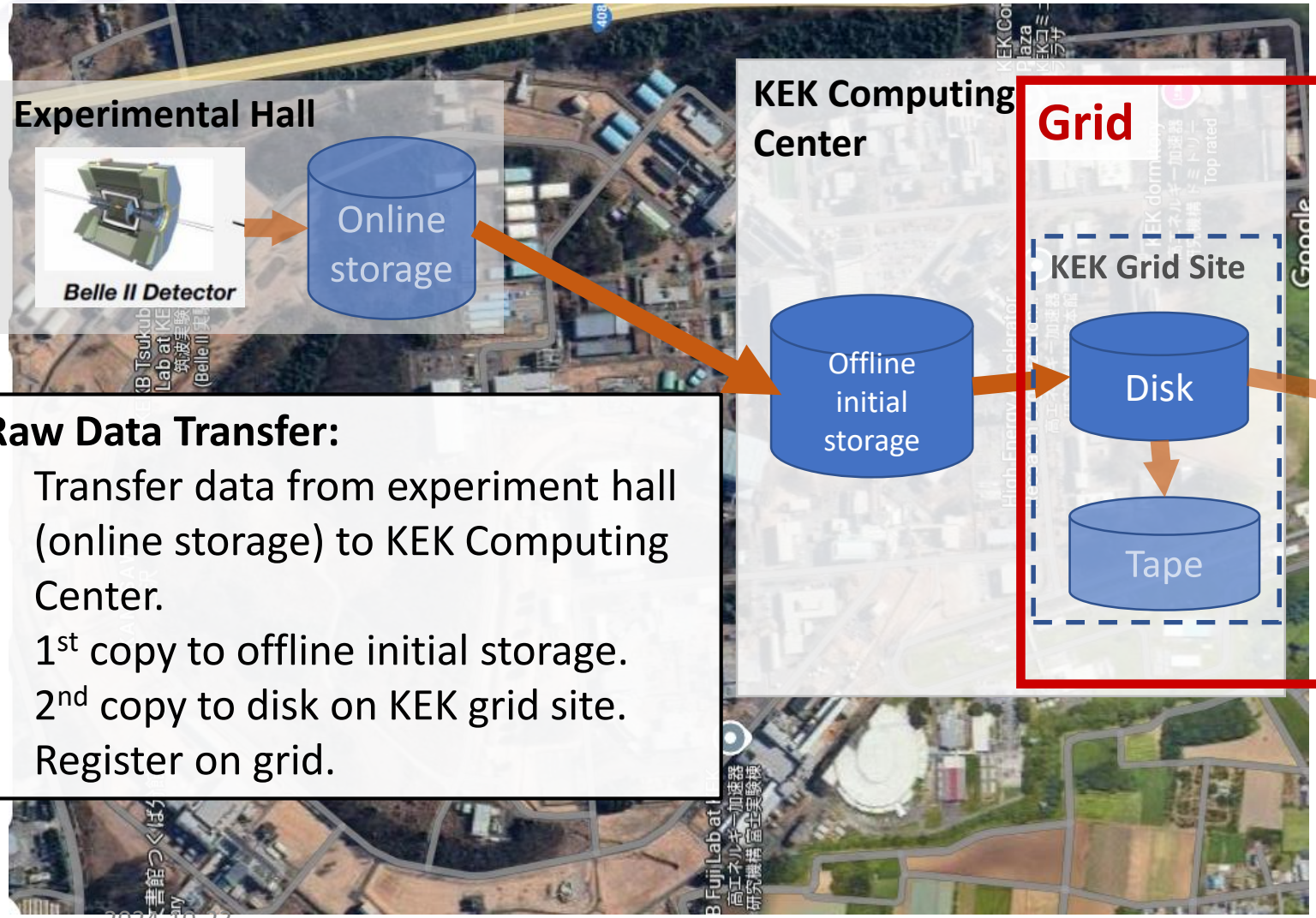
Belle II Raw Data Flow



Online:

- Belle II Data Acquisition (DAQ) system records raw data from detector.
- Stored in online storage servers.
  - Physically close in Experiment hall.

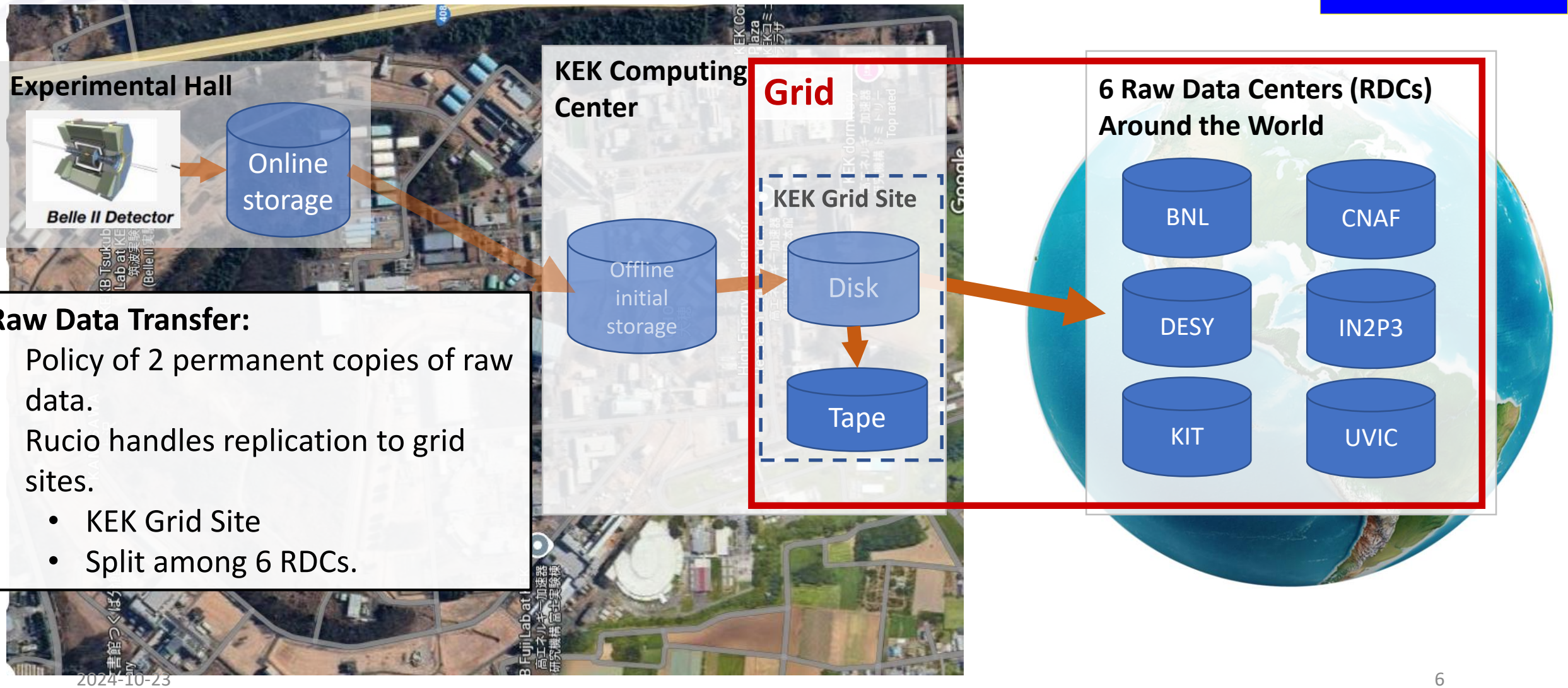
Belle II Raw Data Flow



Raw Data Transfer:

- Transfer data from experiment hall (online storage) to KEK Computing Center.
- 1st copy to offline initial storage.
- 2nd copy to disk on KEK grid site.
- Register on grid.

Belle II Raw Data Flow

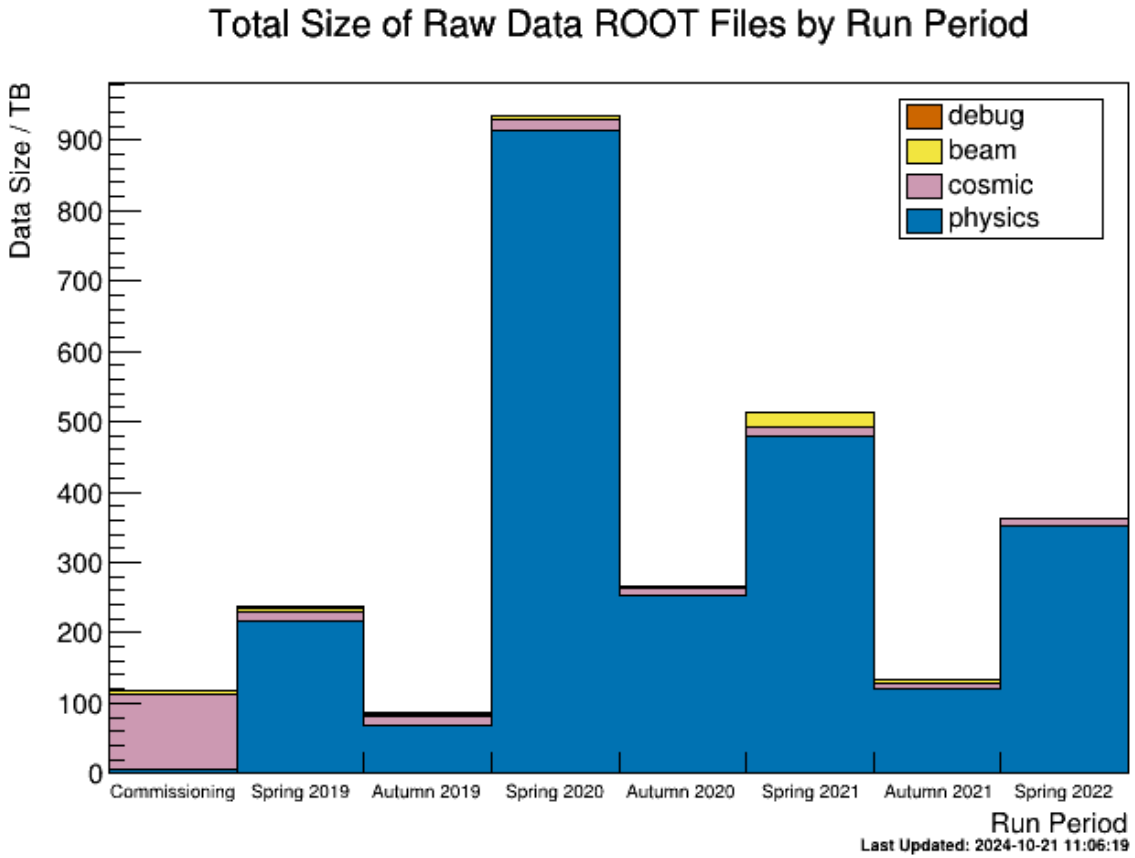


Raw Data Transfer:

- Policy of 2 permanent copies of raw data.
- Rucio handles replication to grid sites.
  - KEK Grid Site
  - Split among 6 RDCs.

Raw Data Transfer System – Run1

- >5PB of raw data transferred since 2019.
- DAQ recorded data in custom data format, converted to ROOT files offline.
  - 5PB -> 2PB after conversion.
- Generally smooth operation.
  - Details see: https://doi.org/10.1007/s41781-020-00045-9
- Operational experience identified areas for improvements over long shutdown 1.

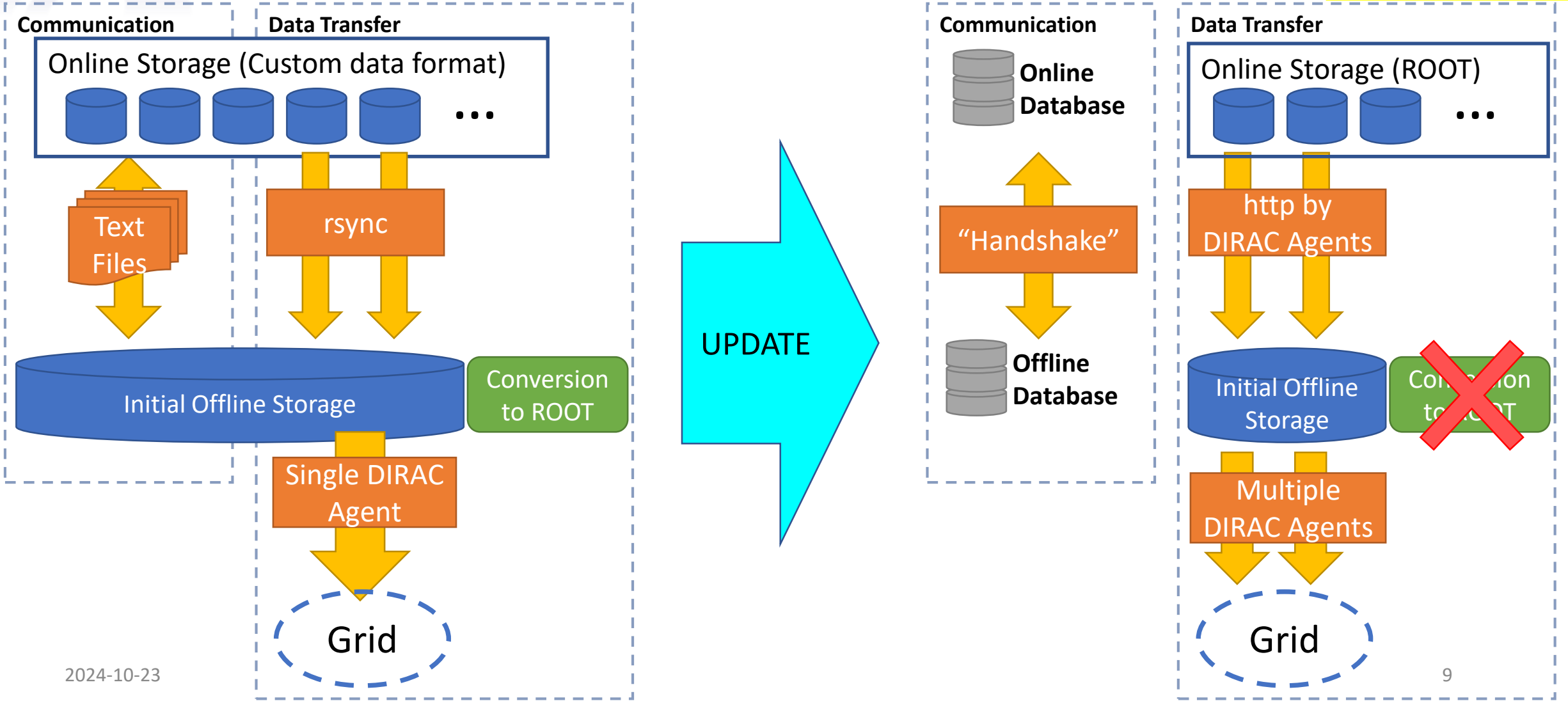


Goals of Update

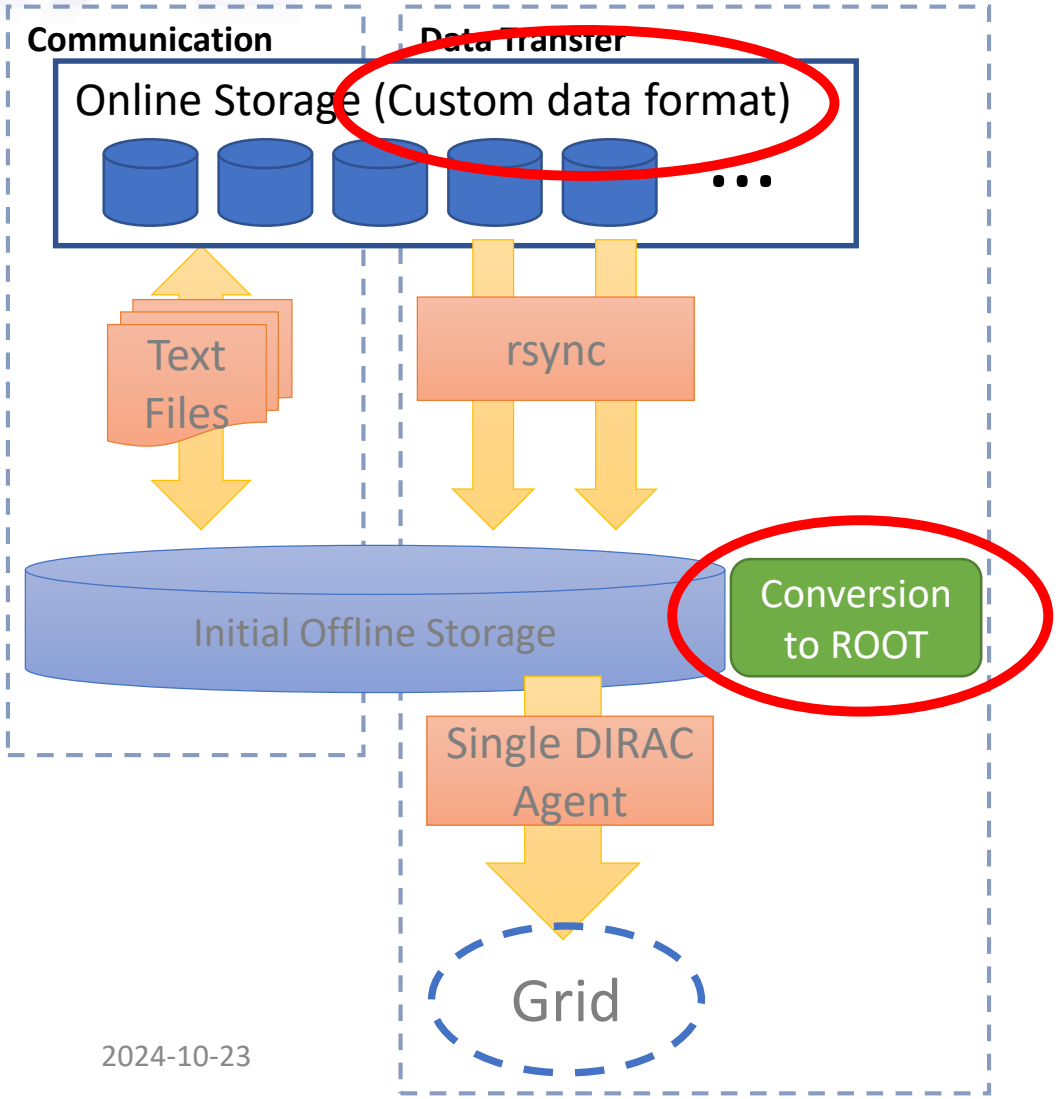
Preparing for future high data rates.

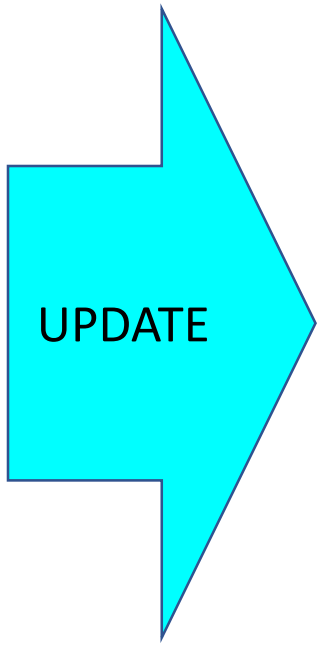
- Increase transfer efficiency.
  - Copying large custom format and converting to smaller ROOT was inefficient.
- Improve robustness and scalability.
  - Communication between online and offline system was infrequent.
    - Delays in starting transfers.
  - Upload to grid performed by single process.
- Reduce load on online storage.
  - rsync for transferring files places load on source server.

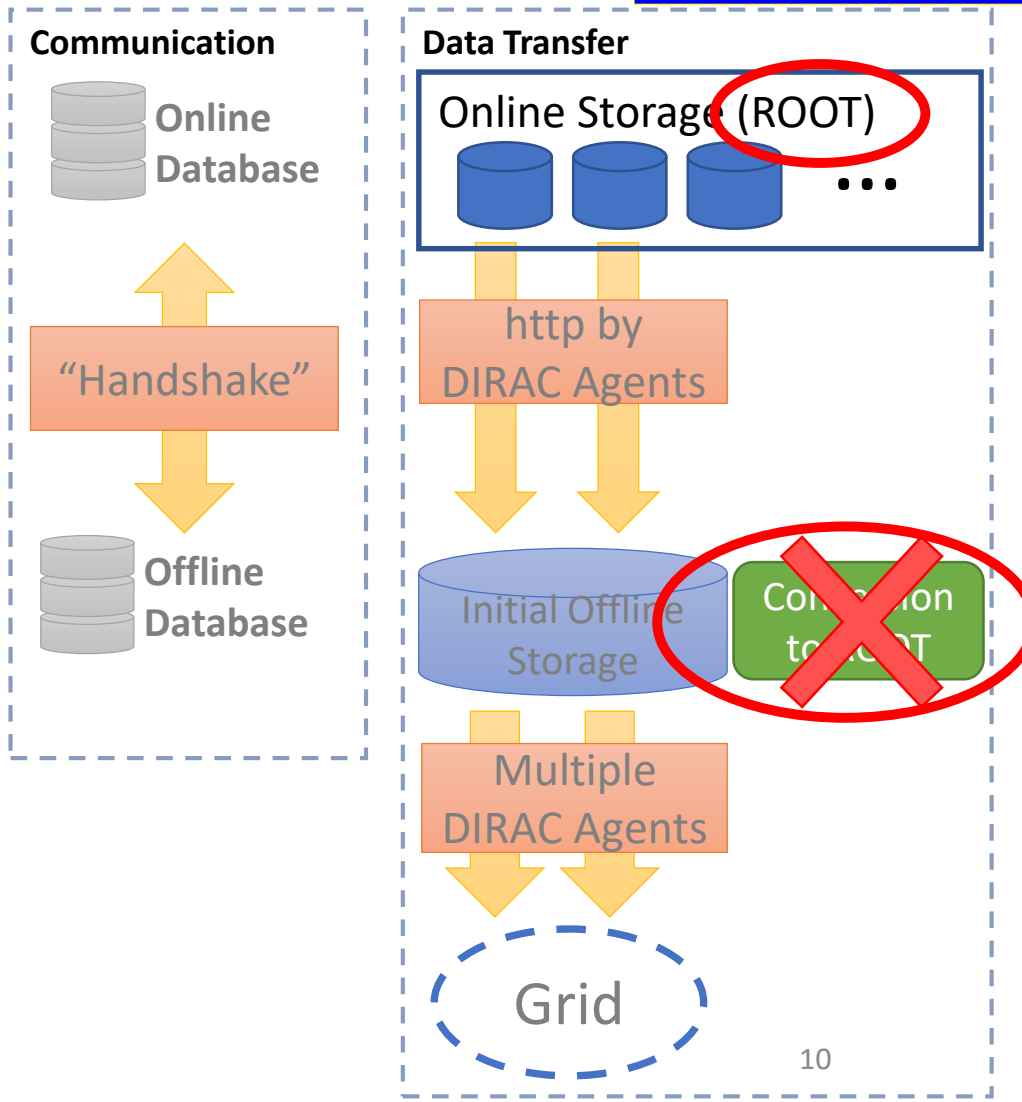
Design of Update



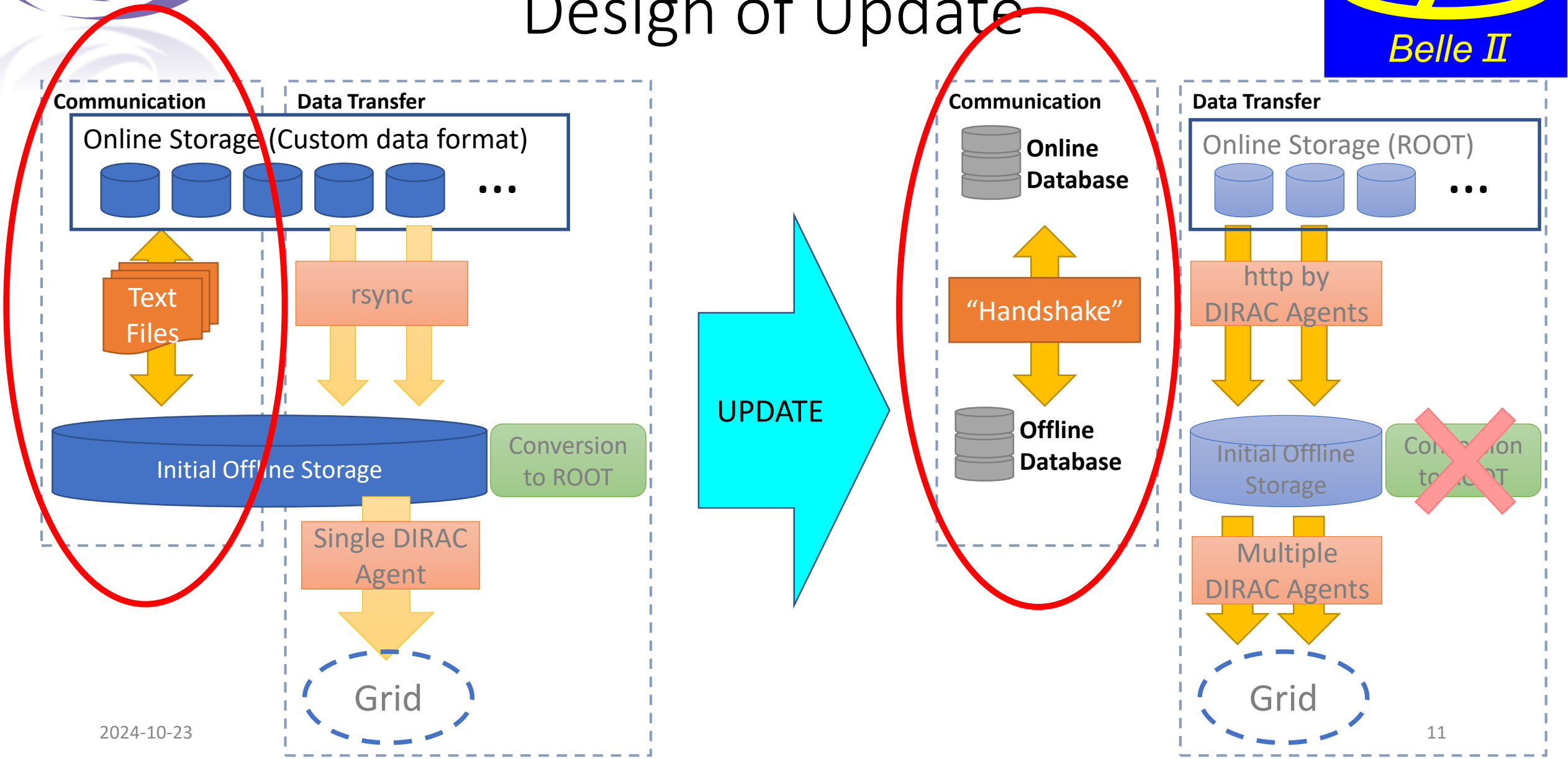
Design of Update



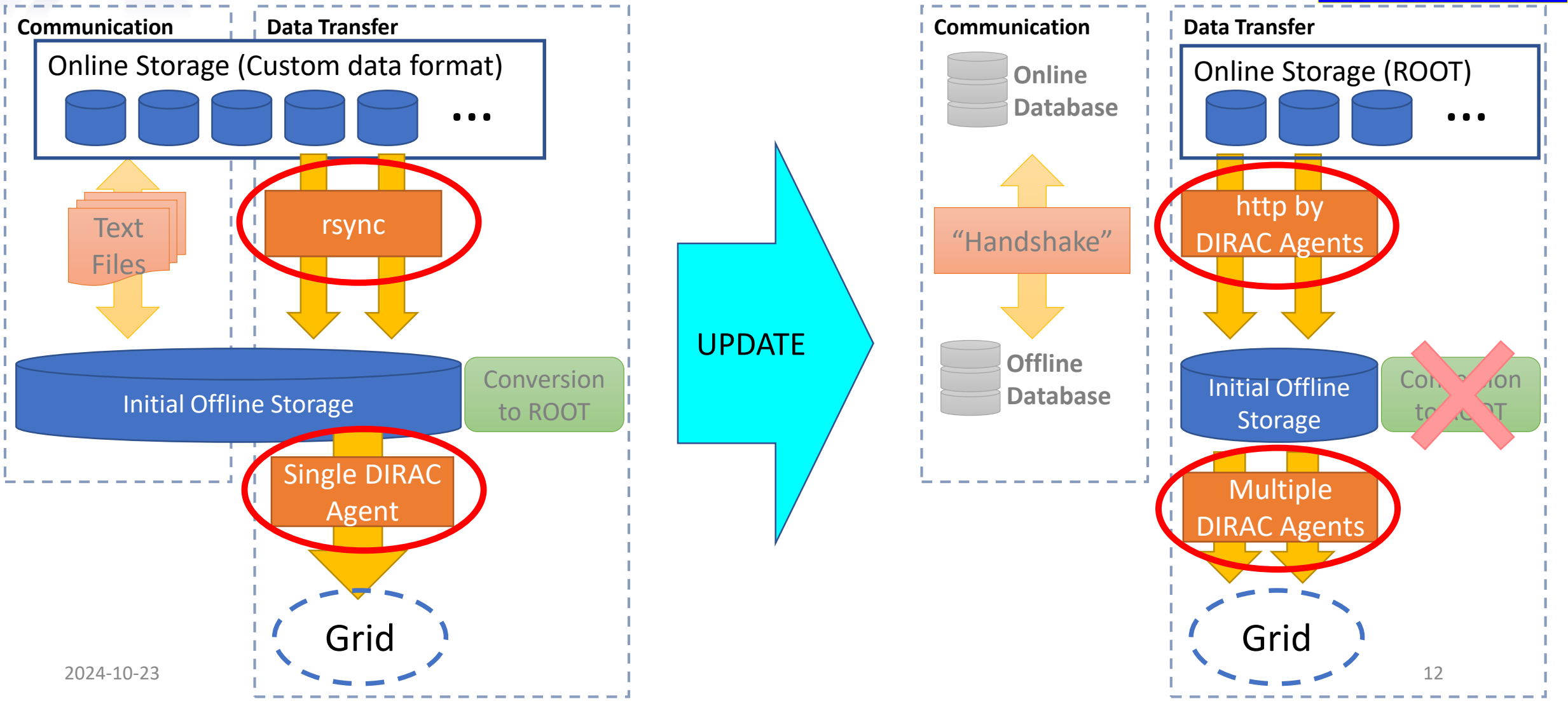




Design of Update



Design of Update



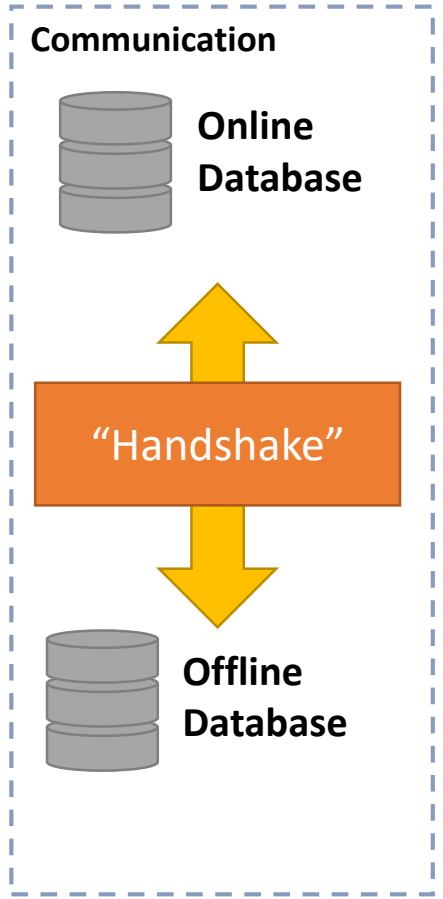
Online-Offline “Handshake”

Handshake initialisation.

- Find new runs to transfer in online database.
- Pre-transfer checks:
  - Detect issues with data before transfer starts.
  - Quickly inform experts to investigate issues before they propagate.
  - Issues raise alert, but transfer of good data continues.
- Insert new runs into offline database.

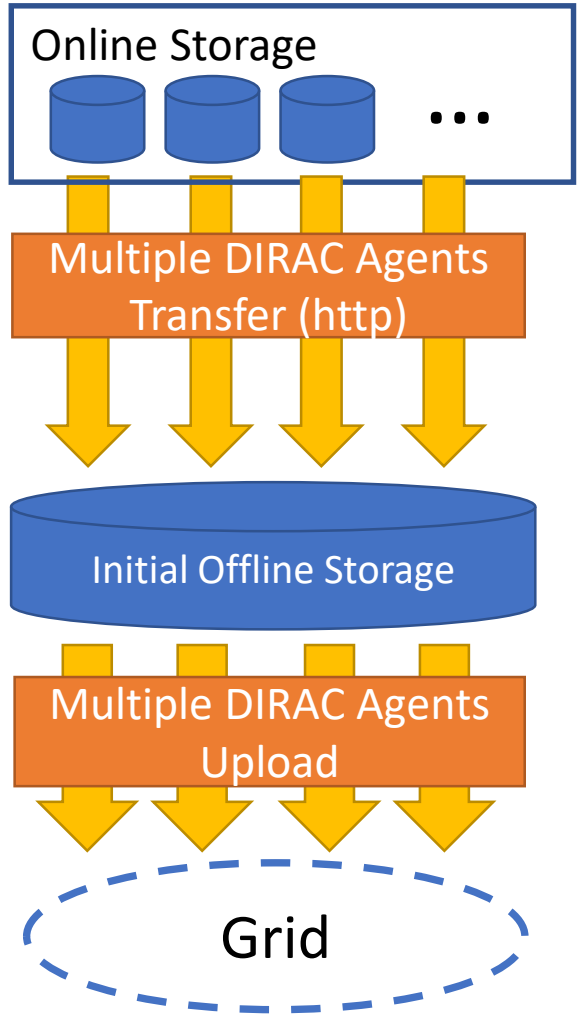
Handshake Close

- Check integrity of transferred data using checksum.
- Set flag in offline database to indicate 2 offline copies.
  - Checked by the online system to allow cleanup of online storage.



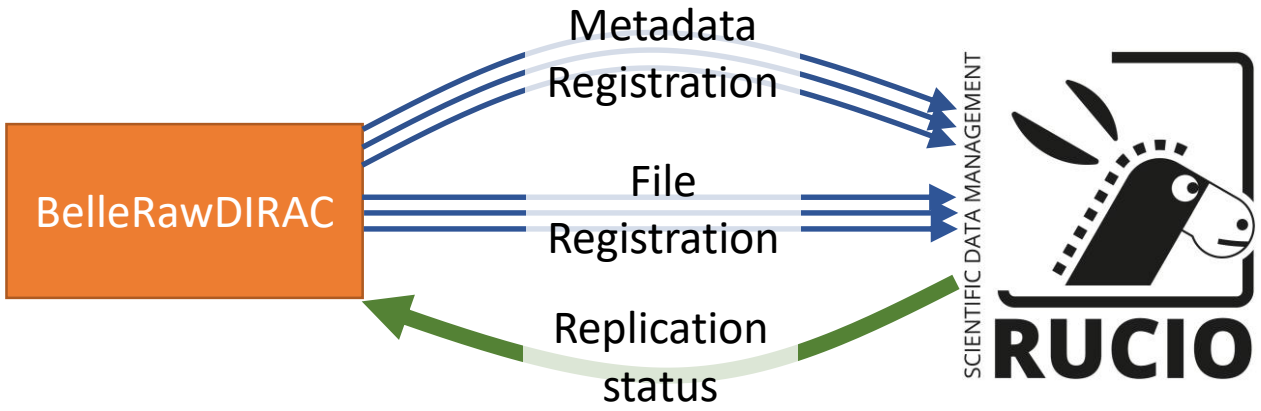
BelleRawDIRAC Upgrade - Parallelisation

- BelleRawDIRAC performs transfer from online storage to offline initial storage now.
- DIRAC agents that perform upload now parallelised.
  - Can scale with number of storage servers.



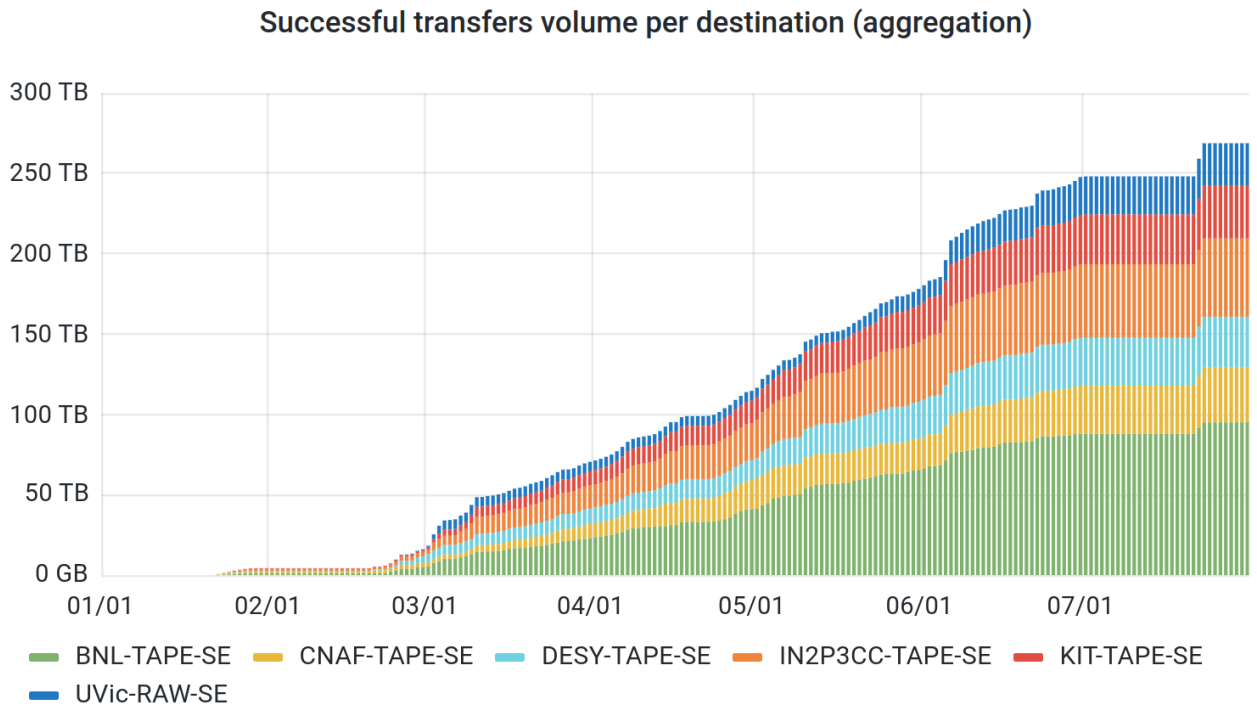
BelleRawDIRAC Upgrade – Rucio Integration

- Belle II grid moving to Rucio to store metadata.
  - Talk at CHEP2022: https://doi.org/10.1051/epjconf/202429501025
- All raw data metadata now sent to Rucio.
  - Scalable with usage of bulk registration API.
- Improved tracking of replication status thanks to better Rucio integration.



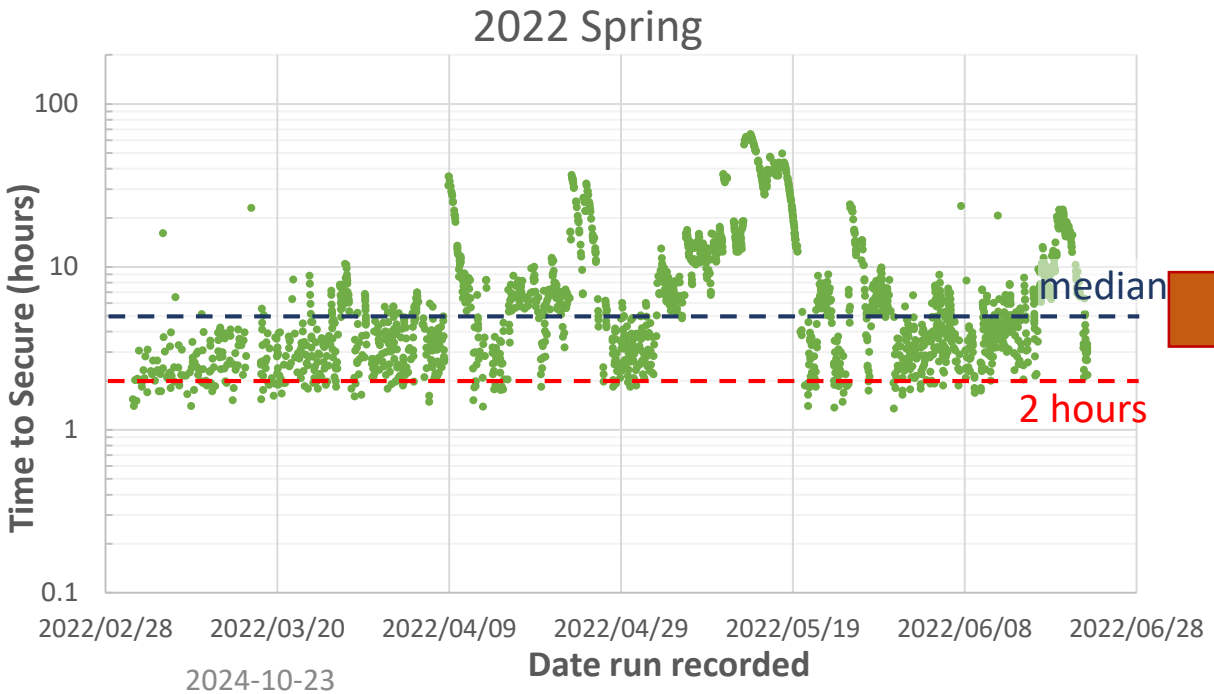
Operation in 2024

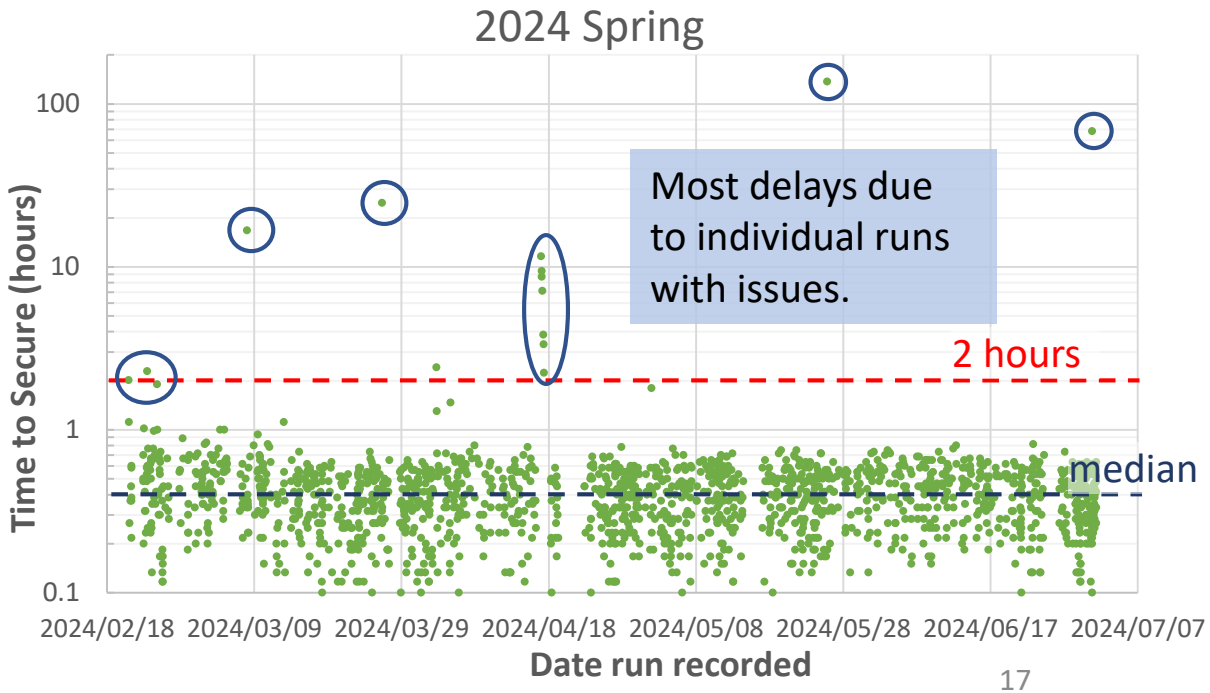
- Updated system was put into production from Feb. 2024 for the start of Run 2.
- >250TB of raw data transferred.
- Very few issues with system.
  - Successful in detecting most problems with data before transfer.
- Significant decrease in delays.



Time to Transfer Offline

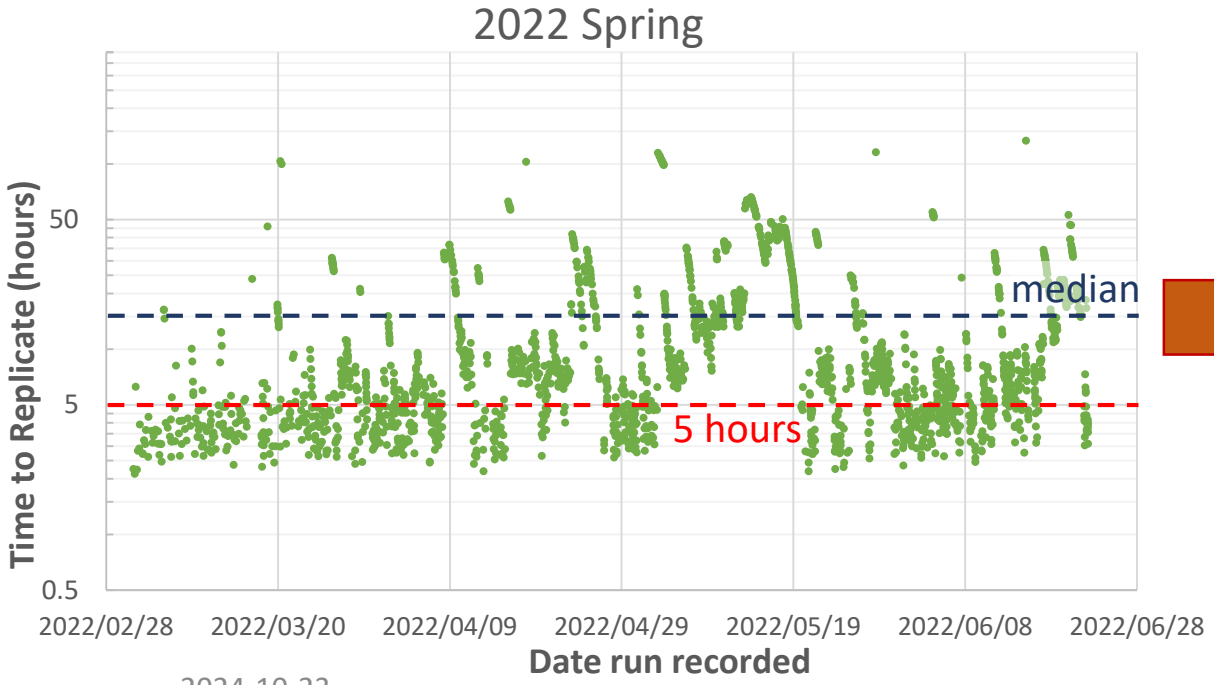
- Time between run end and having 2 offline copies.
- Median: 310 mins -> 26 mins.
- Almost all runs stored offline within 2 hours now.

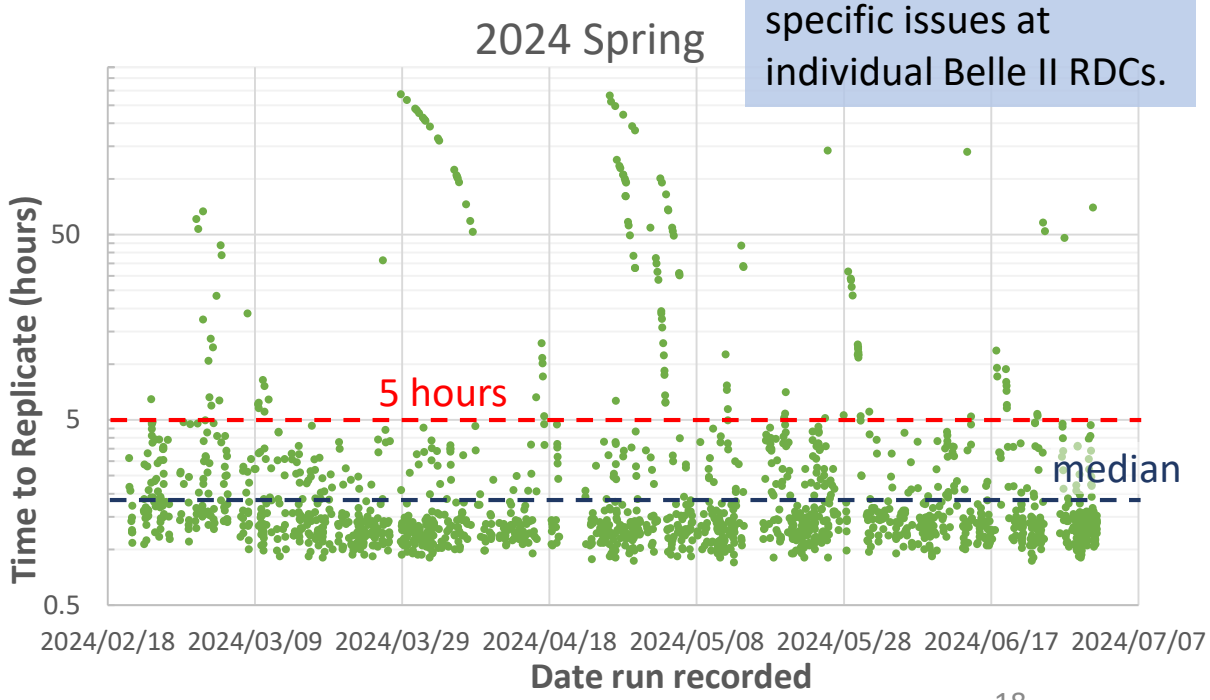




Time to Replicate to Grid

- Time taken to complete replication to final storage on grid.
- Median: 7 hours-> 1.5 hours.
- Most runs replicated within 5 hours.





Outliers arise from specific issues at individual Belle II RDCs.

Summary

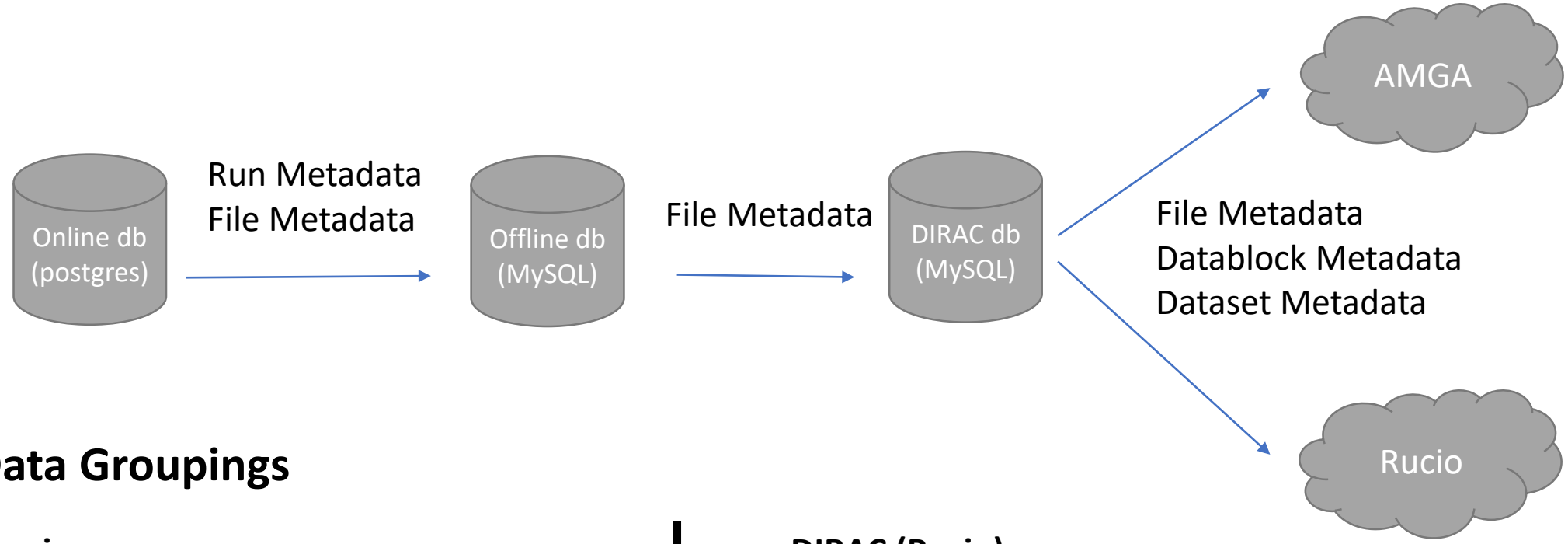
- The Belle II Raw Data Transfer System moves raw data from near the detector to final grid storage.
- During the long shutdown period significant improvements were made to the system.
  - More size efficient file format.
  - New, more scalable and timely online-offline communications system.
  - Improved scalability of our DIRAC extension.
  - Better integration with Rucio.
- In production since Feb. 2024.
  - Improvements in both transfer times and reliability.

Backup

Raw Data Transfer System – Pre LS1

- Data transfer initiated by ‘list-send’ system.
  - Each online storage server periodically sends text file listing files to transfer.
- Data from online to offline initial storage via rsync.
- Files are converted from SROOT to ROOT format.
- Files registered in BelleRawDIRAC (DIRAC extension).
  - Uploaded to KEK permanent storage.
  - Replicated to RDCs.
- Send text file back to storage server, informing data copied.
- Once online storage server confirms copied, start cycle again.

Raw Data Metadata



Data Groupings

Physics:

Run: Data gather period (up to 8 hours)

Files

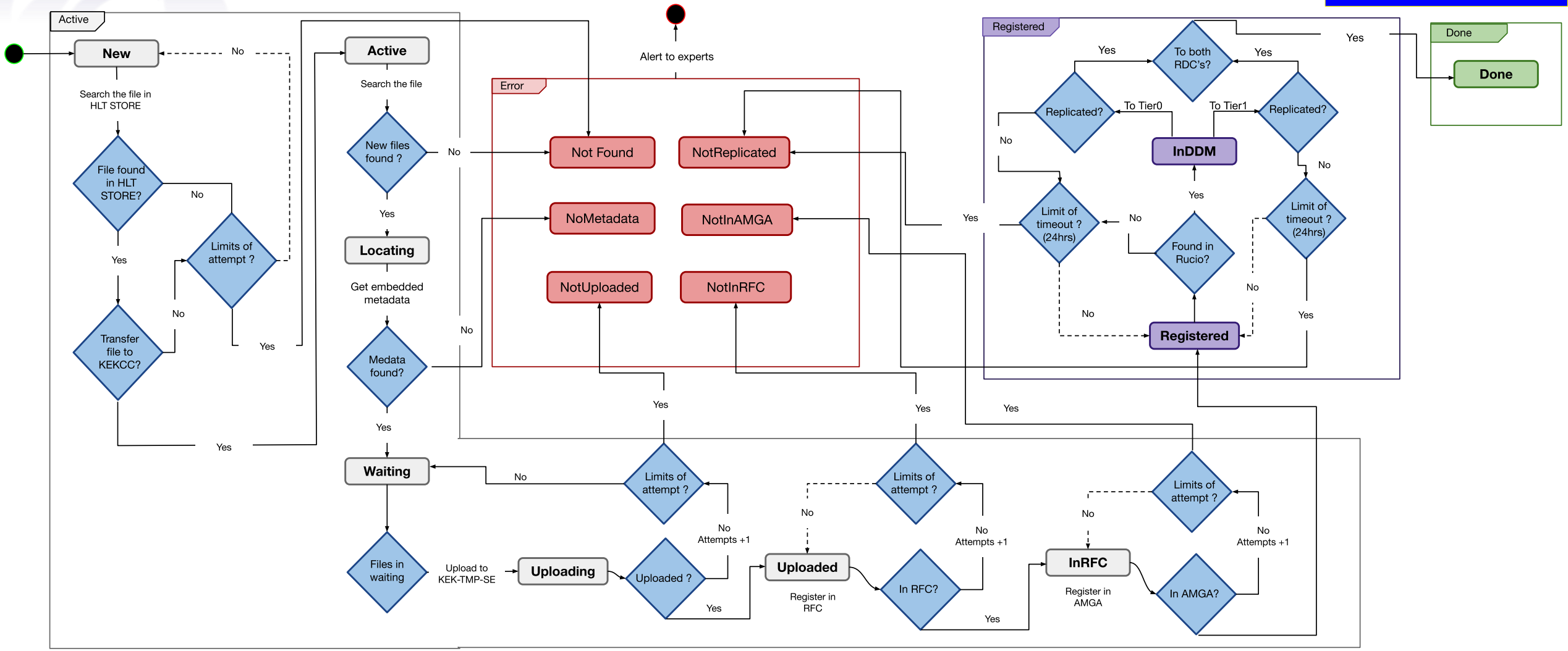
DIRAC (Rucio):

Dataset (container): All data in a run.

Datablock (dataset): 1TB grouping.

Files (file)

BelleRawDIRAC - Detailed

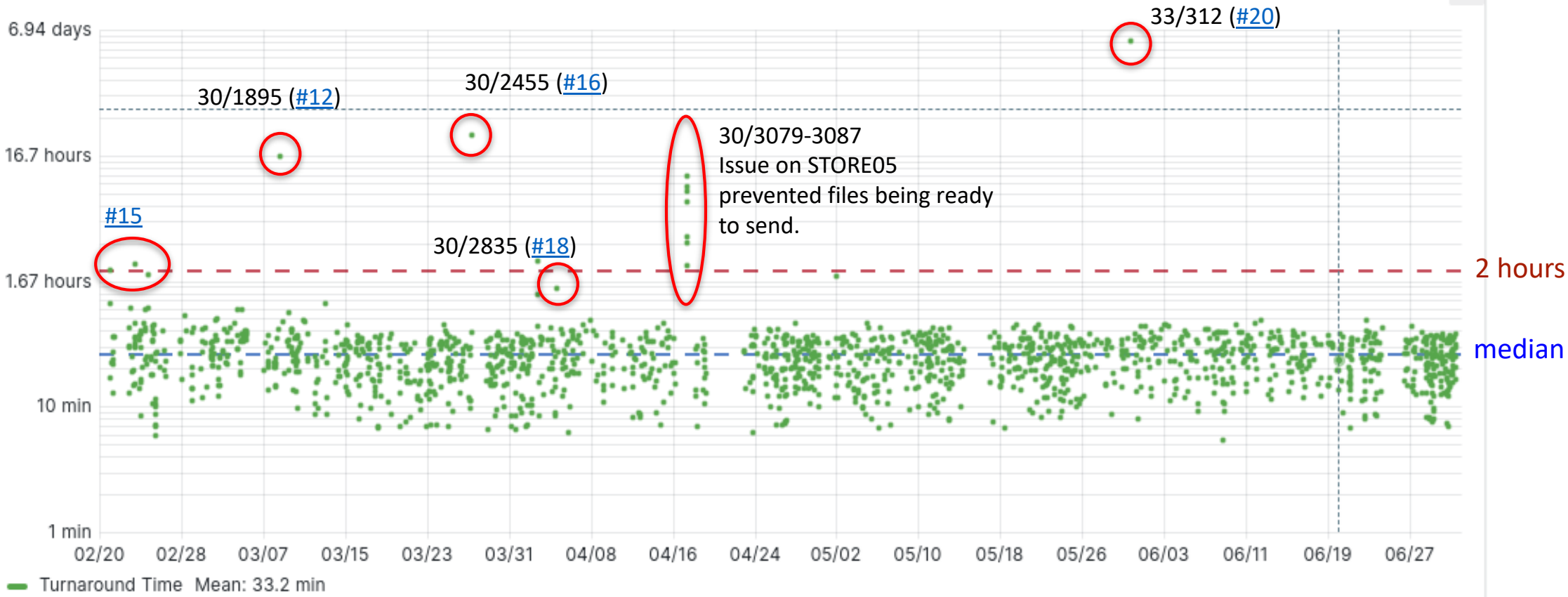


Operational issues - selected specific issues

- #6 , #21: A mix of global and local run files (e.g. *debug* and *trg*) produced for a single run.
  - Manual cleanup of the extra files (by DAQ expert) and database(s) was required.
- #12, #20, #22: runs were not initially transferred, as they were not marked as ready to be transferred.
  - Notified DAQ expert, and fix by them allowed for the transfer of files.
- #13 and #16: wrong run type.
  - #13: (30/2068, 30/2069) set as *cosmic*, intended to be *debug*.
    - Runs were not transferred due to another issue - allowed the run type to be fixed.
  - #16: (30/2455) set as *debug*, intended to be *physics.* All other setting were correct for physics.
  - It was possible to change the run type, and transfer/process as a physics run.
  - **Reminder**: Once a run has been registered on the grid it is (almost) impossible to change the run type.

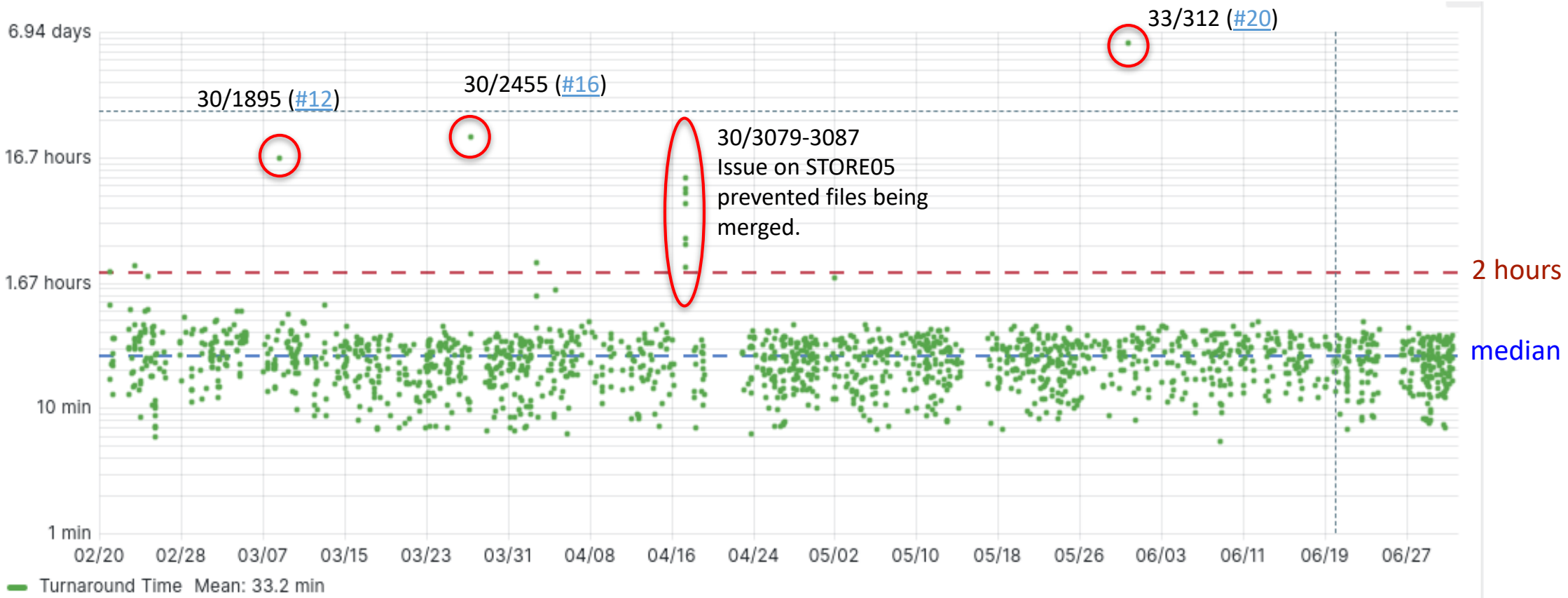
Operational Issues - Turnaround time

- Time taken to secure run (2 offline copies, DAQ can clean HLT STORE) from run stop.
  - physics runs, exp 30-33: mean: 33 min, median: 26 min.



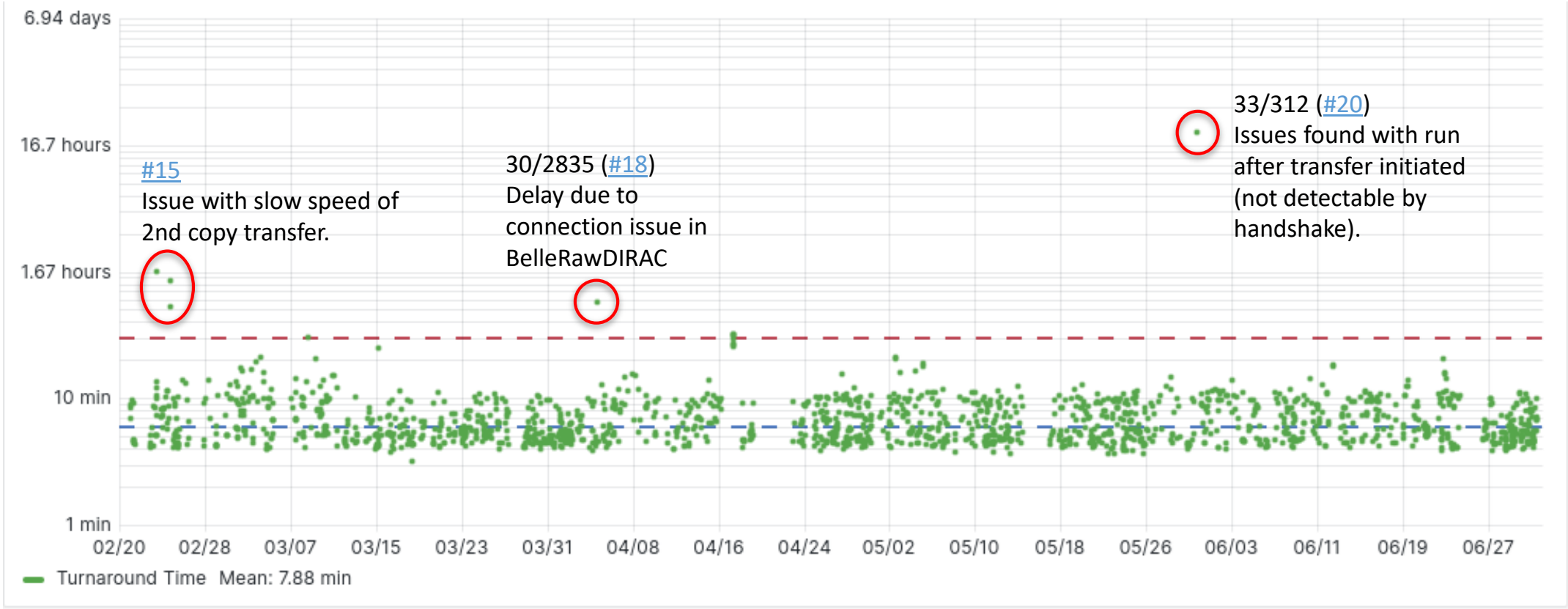
Operational Issues - Turnaround time

- Delays were caused by issues with a run being detected by the handshake protocol, or a run not marked as ready for transfer.
  - Later runs were not delayed.

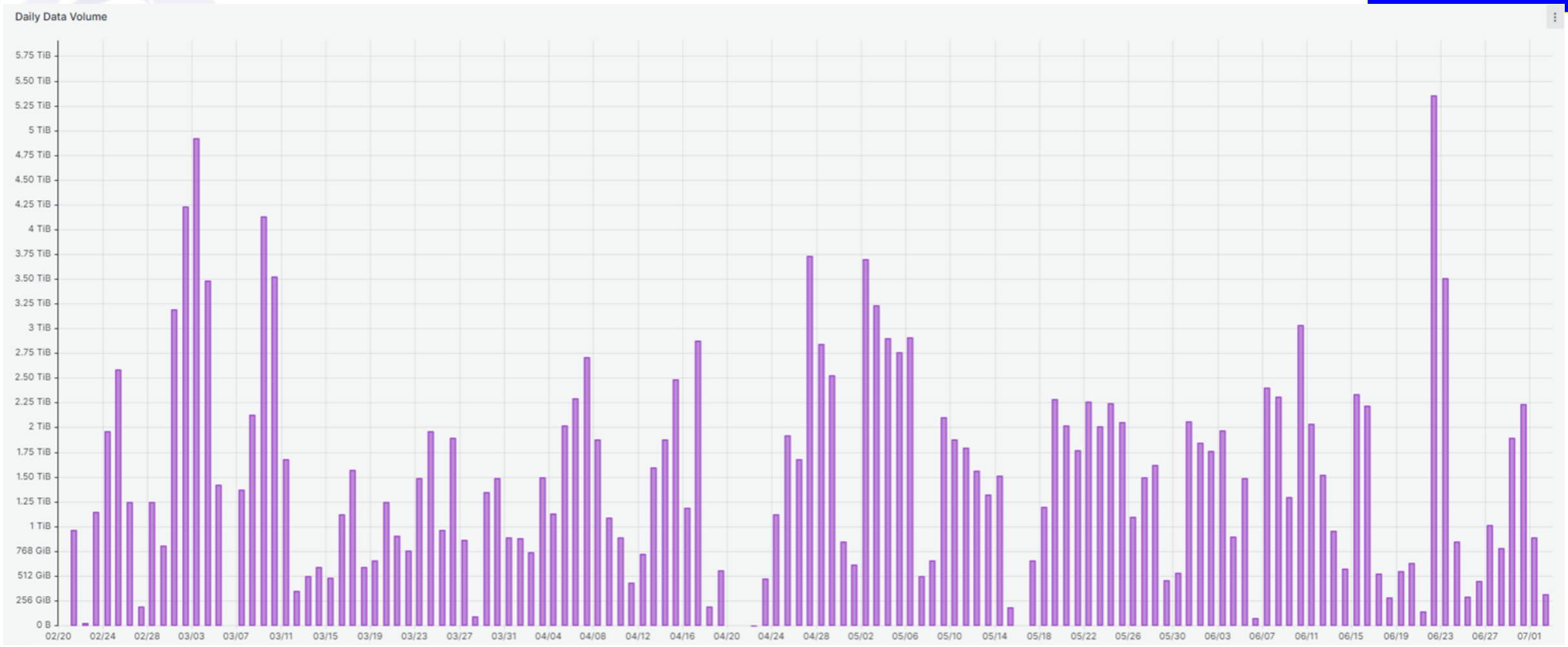


Operational Issues - Transfer time

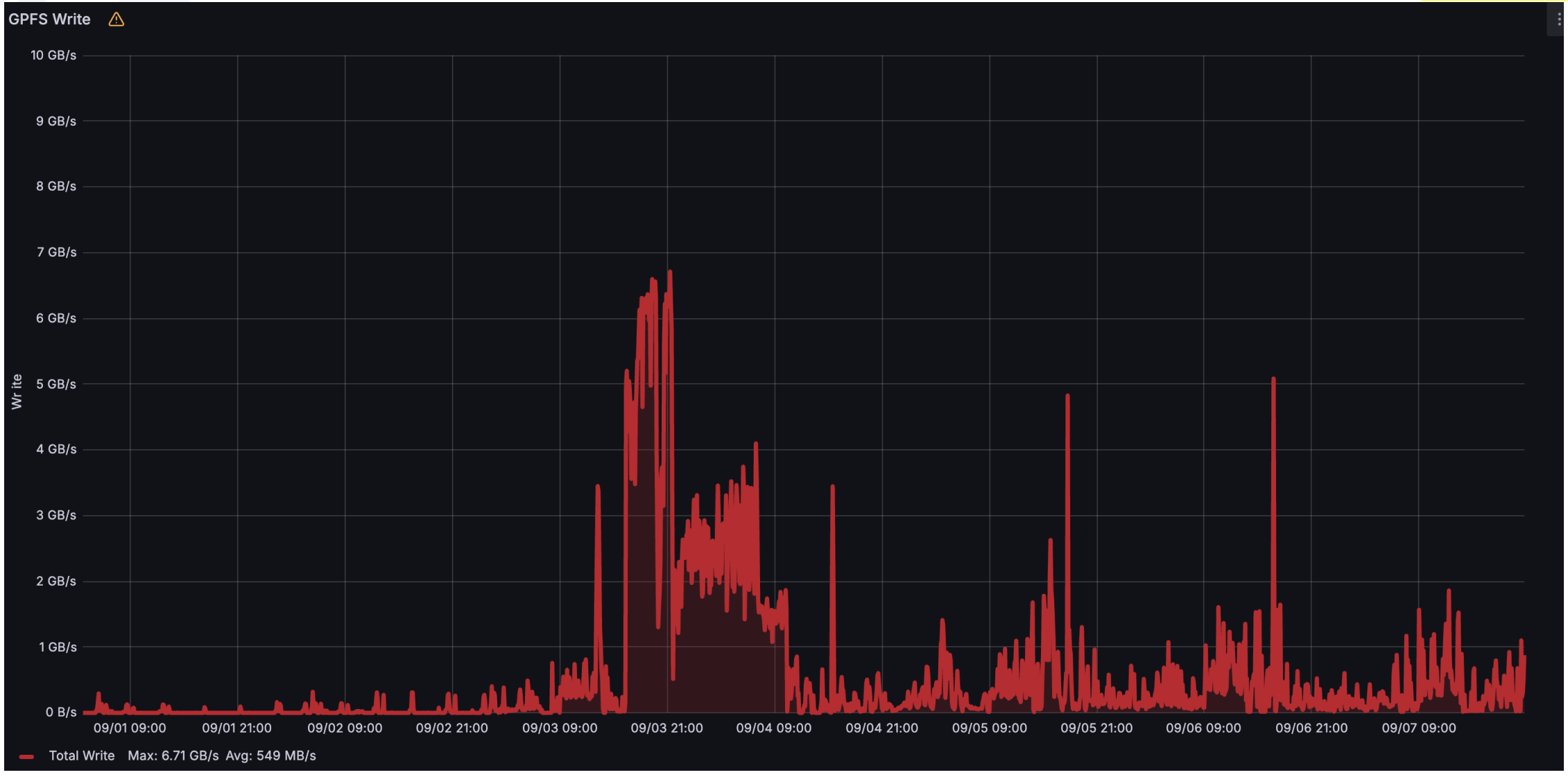
- Time taken to transfer 2 copies of each run offline. physics runs, exp 30-33.
  - Start of transfer until 2 copies secured.
  - Mean: 8min, median: 6min
- Only a few minor delays due to issues with the transfer system



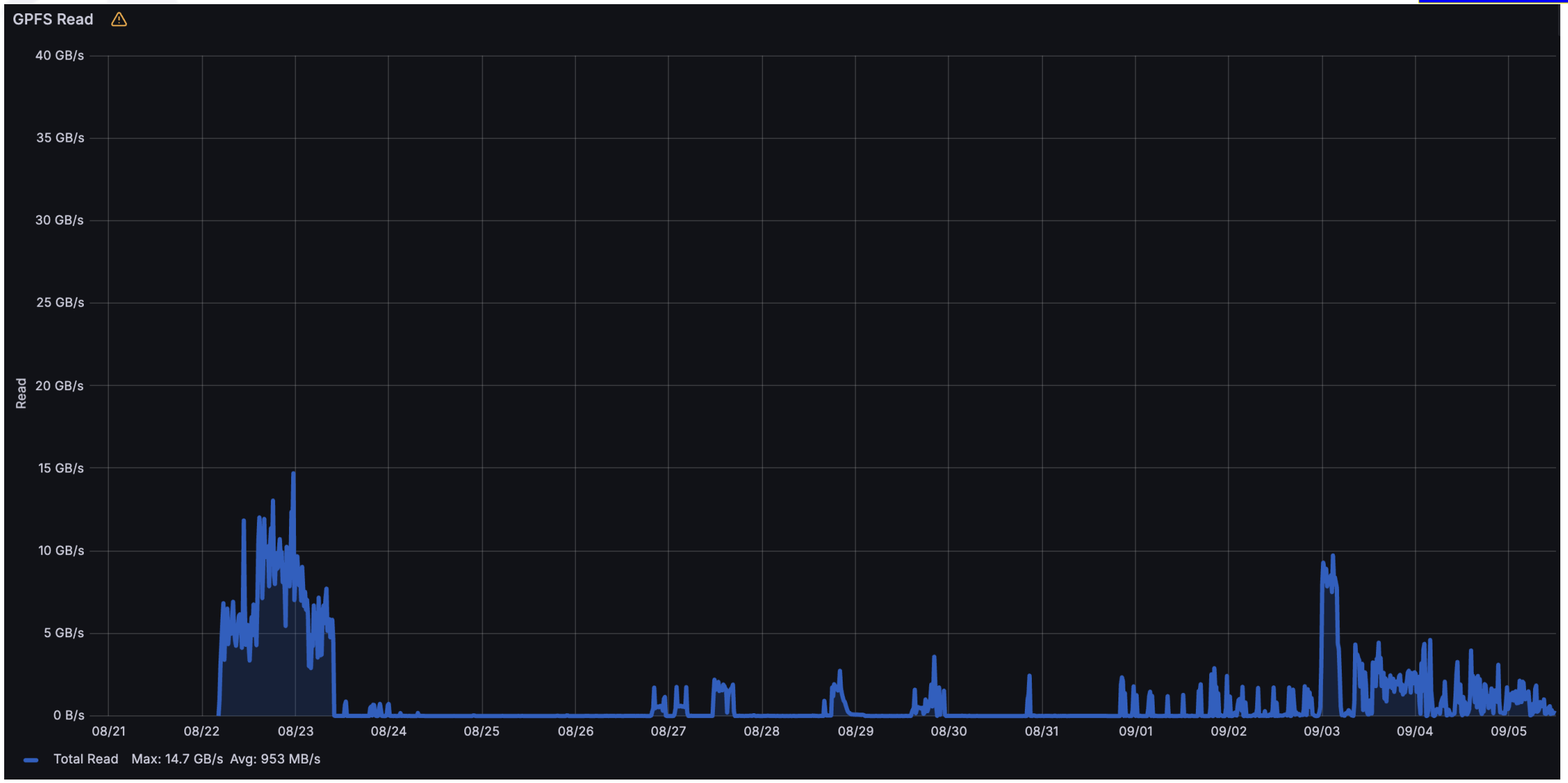
Daily Data Volume - Run 2024ab



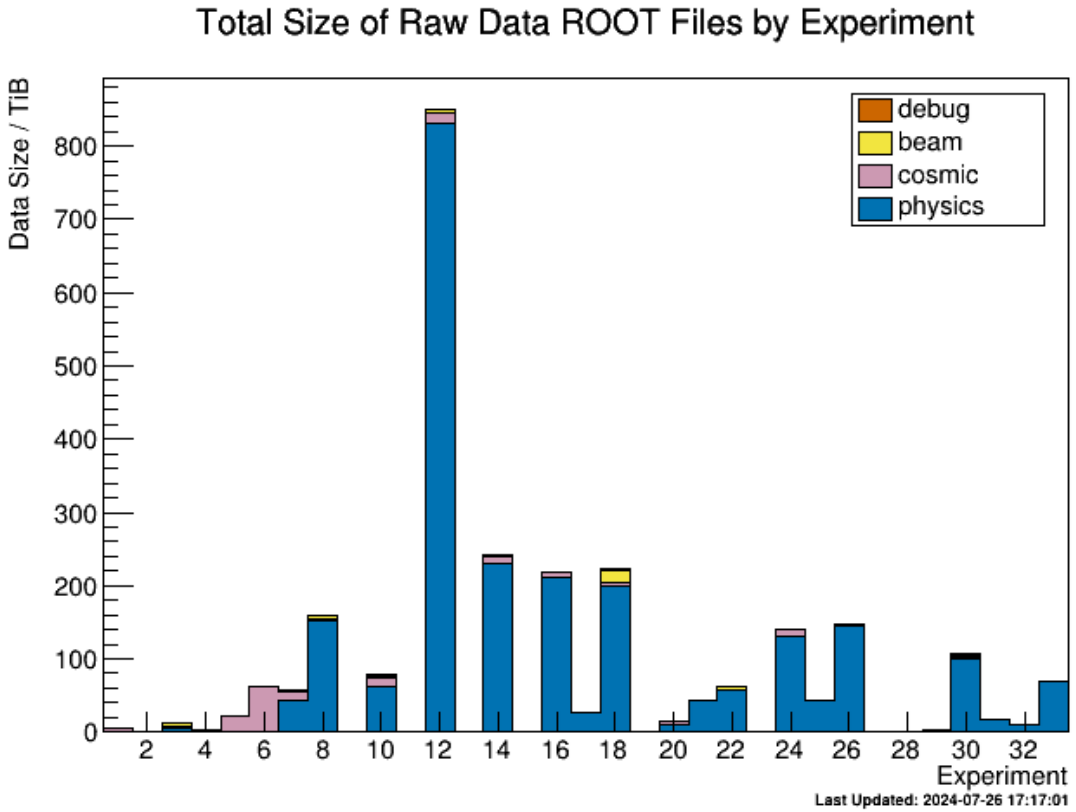
Transfer Bandwidth

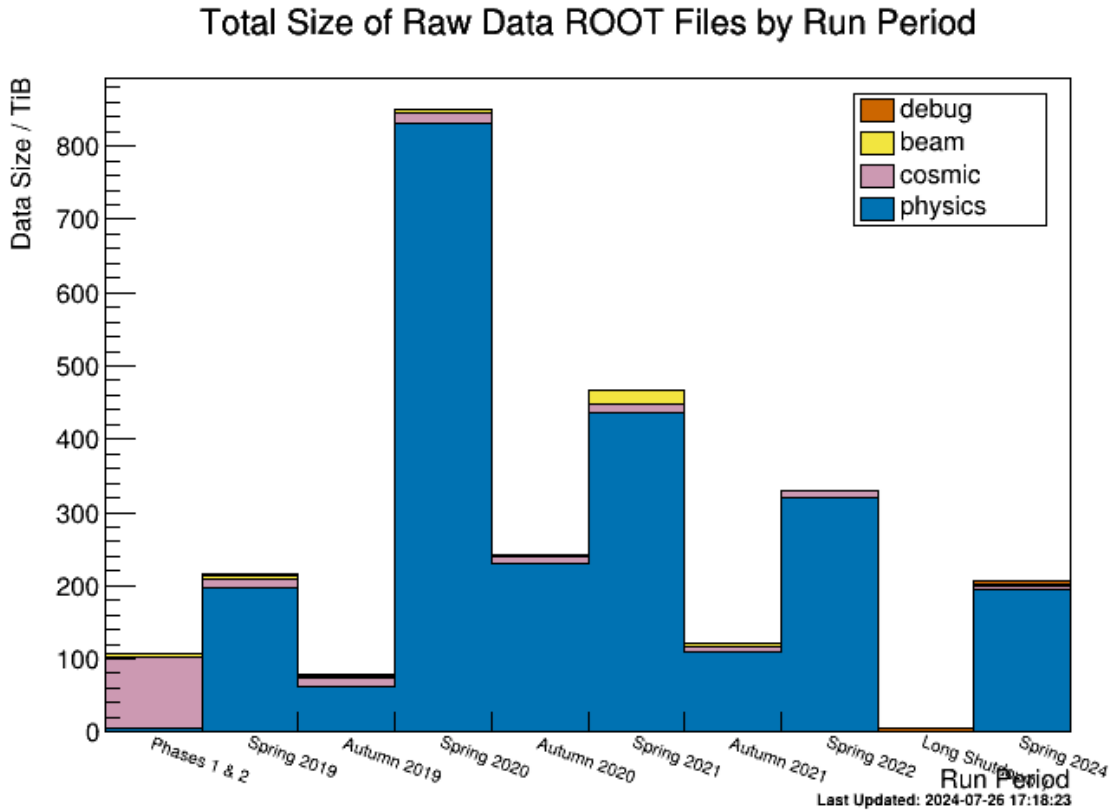


Upload Bandwidth



Raw data size by Experiment and Run Period





Control Room Monitoring

- Updated monitoring page available, and updated instructions.
- Check if connection is down between Tsukuba Hall and KEKCC.
- Check if free HLT STORE space is becoming dangerously low (not expected in Run 2).

The status and trend of the different stages of the data transfers.
- **Task 1**: If we are taking ***physics*** runs, the Data Transfer status is "ON", and these numbers show no change more than 4 hours after the end of a run, please report this.

Note: *null*, *hlttest*,and local run numbers will not appear here; *debug* runs are not registered on the grid, and thus will not update the "latest run on grid" value.

The connection status between the HLTs and KEKCC.
- **Task 2**: If the status for any HLT is DOWN and the HLT is included in data taking, please report this.

Note: HLT12 is masked here for Run2024ab.

The time taken to transfer a run to KEKCC.
- **Task 4**: If this is above 1 hour for any run please report this.

How long it would take to completely fill the HLT STORE with continuous data taking, assuming no transfers and thus no clean up take place.
This will be a large number for (early) Run 2, but will become increasingly important as Belle II reaches full luminosity.
- **Task 3**: Please report if this number is less than 1 week (this is not expected to occur during Run 2).

How to report: send an email to comp-b2cc-ops@belle2.org, and post a comment in #daqcore in rocketchat. **Do not make a phone call.**

https://confluence.desy.de/display/BI/Belle+II+Operation+Mar

Belle II Raw Data Flow