Storage infrastructure at DESY

Belle II Germany Meeting

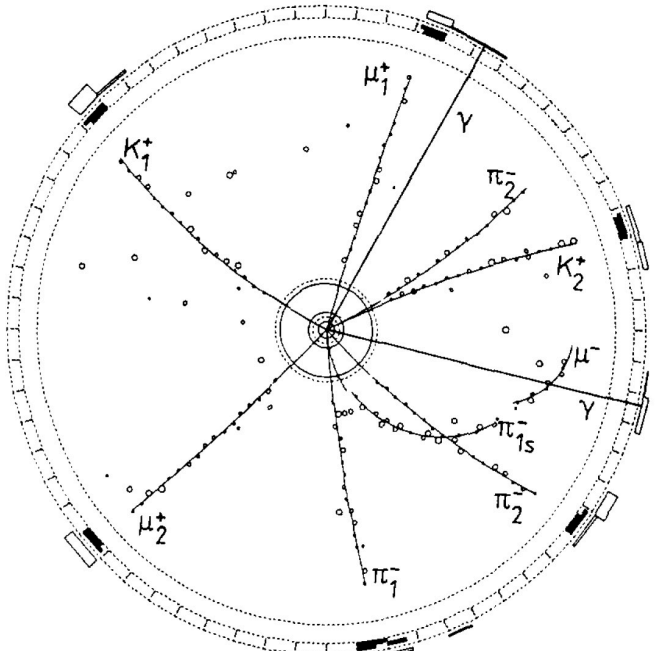
Christian Voß
Munich, 20th September 2022

HELMHOLTZ

DESY.

Overview of Scientific Research at DESY

Belle II in Context to other Communities



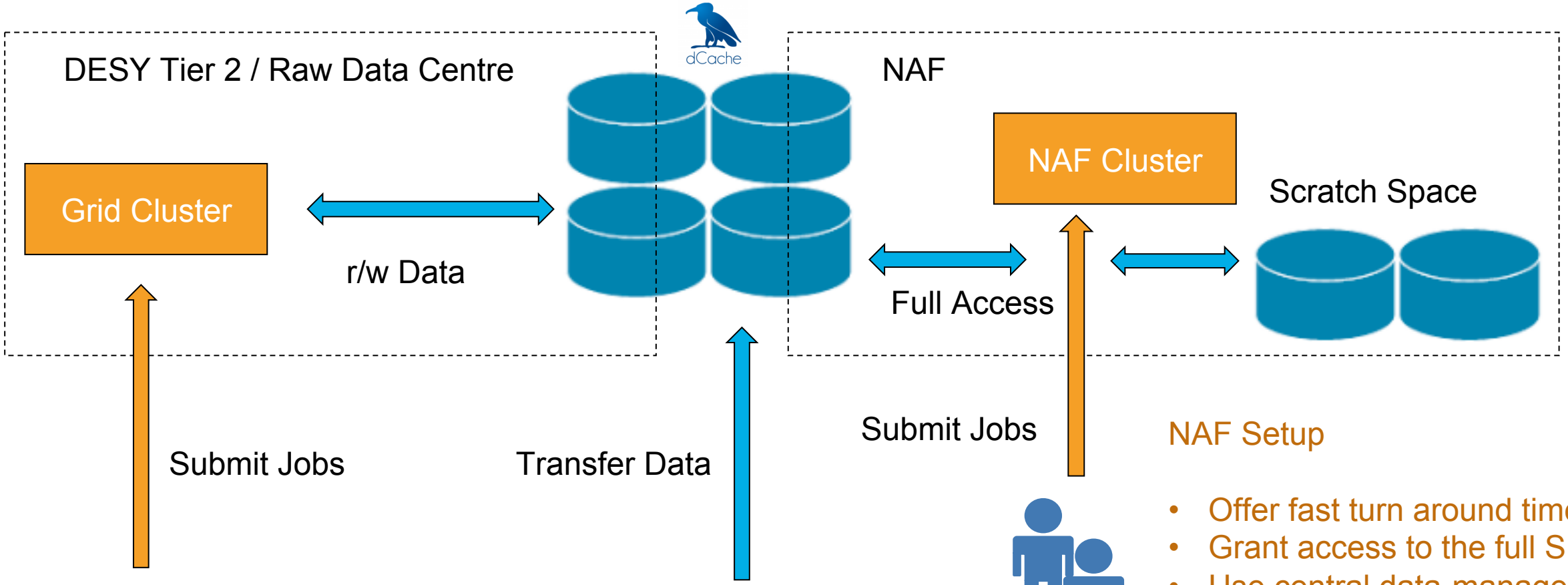
- DESY has a long tradition of on-site experiments
- Example: ARGUS experiments discovering many firsts in our field
  - B-Mixing
  - Semileptonic B-decay

- End of HERA in 2007 saw strategic shift focussing on research with photons present since the early 90s
- Today, the large on-site experiments support research with photons: PETRA III, FLASH, European XFEL
- Leads to direct competition for existing resources including compute resources
- Visible: re-calibration analysis pipeline using tape resources to greater degree

Paradigm: HEP Analyses are Data Driven

As Underlying Principle of the NAF

- Almost all HEP data analyses require access to large amounts of data



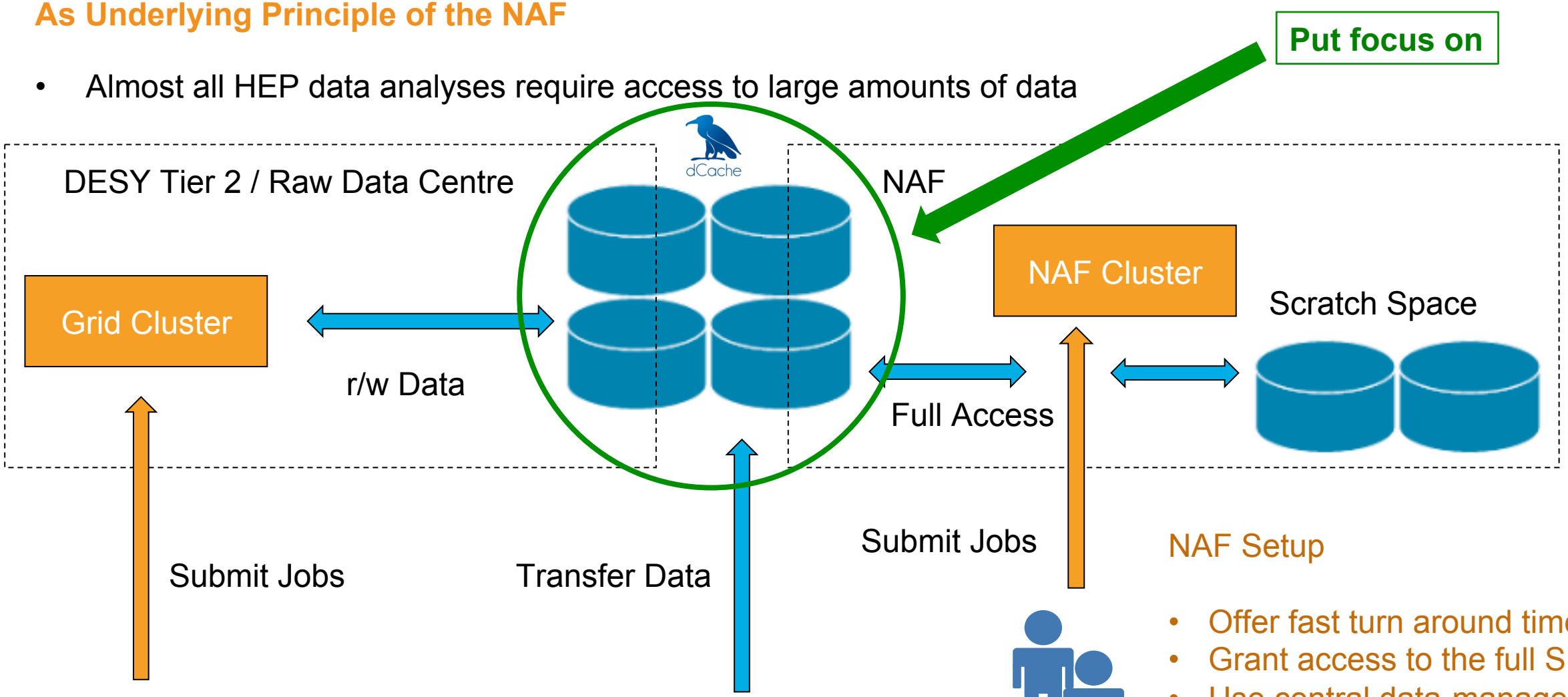
NAF Setup

- Offer fast turn around times
- Grant access to the full SEs
- Use central data-management of experiments to store data at DESY

Paradigm: HEP Analyses are Data Driven

As Underlying Principle of the NAF

- Almost all HEP data analyses require access to large amounts of data



NAF Setup

- Offer fast turn around times
- Grant access to the full SEs
- Use central data-management of experiments to store data at DESY

Mass-Storage for Belle II in Grid and NAF

dCache as Central Mass Storage for HEP communities

- Central element in overall storage strategy
- Collaborative development under open source licence by
  - DESY
  - Fermilab
  - Nordic E-Infrastructure Collaboration (inoffically NDGF)
- Particle Physics in general
  - In production at 9 of 13 WLCG Tier-1 centres
  - In use at over 60 Tier-2 sites world wide
  - 75% of all remote LHC data stored on dCache
  - In addition: Tevatron and HERA data
- Belle II among others
  - German Raw-Data-Centres DESY and KIT
  - Brookhaven National Laboratory
  - University of Victoria

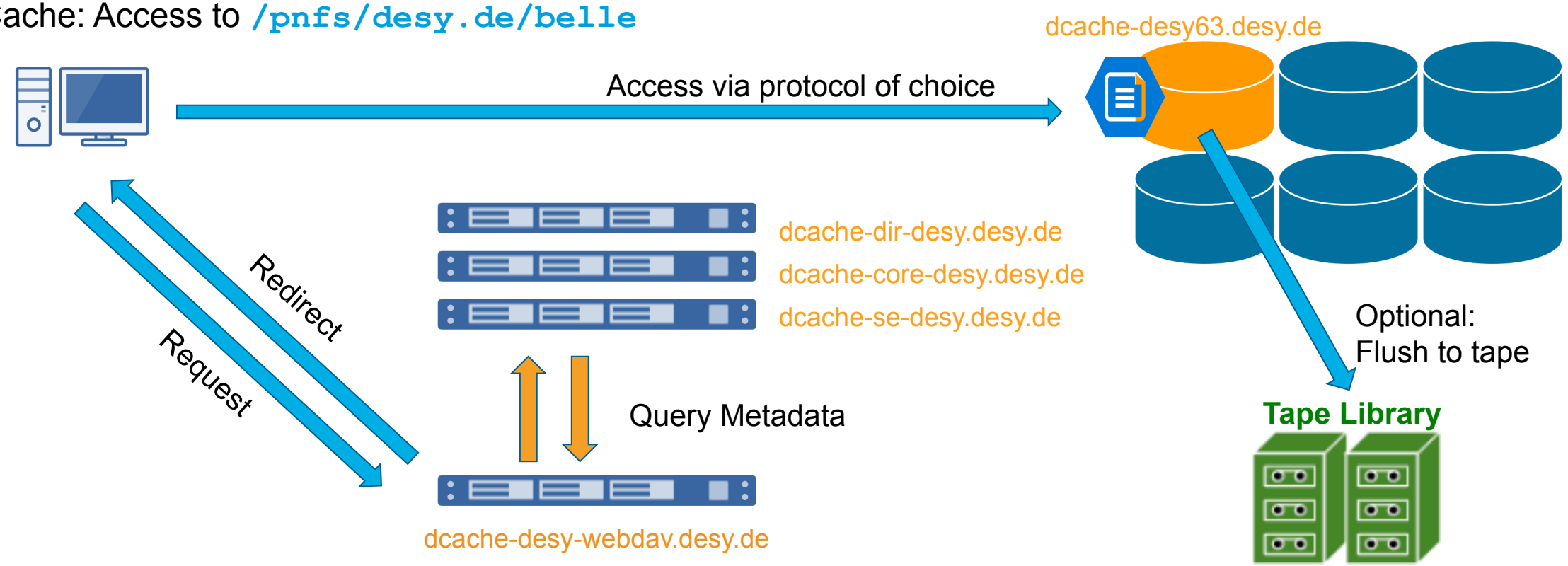
Features

- Highly horizontally scalable storage system
- Expose a single unified namespace
- Supports many protocols
- Supports many authorisation schemes
- Micro-service architecture

Basic Setup for dCache at DESY

How to Store and Access Data to dCache

- Use dCache: Access to `/pnfs/desy.de/belle`

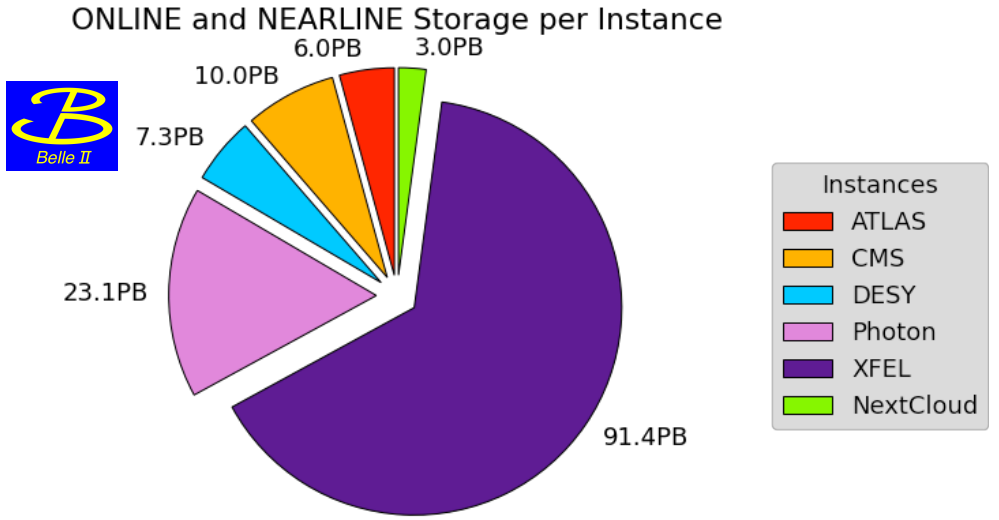


- Access done through doors: several load balanced door for each protocol to ensure availability
- Access controlled via Grid-certificate (to be replaced by tokens) and POSIX (NFS@NAF)
- Data streamed to/from pool, never through doors: allows horizontal scaling
- Namespace is uniform and independent of protocol

Supported Communities at DESY

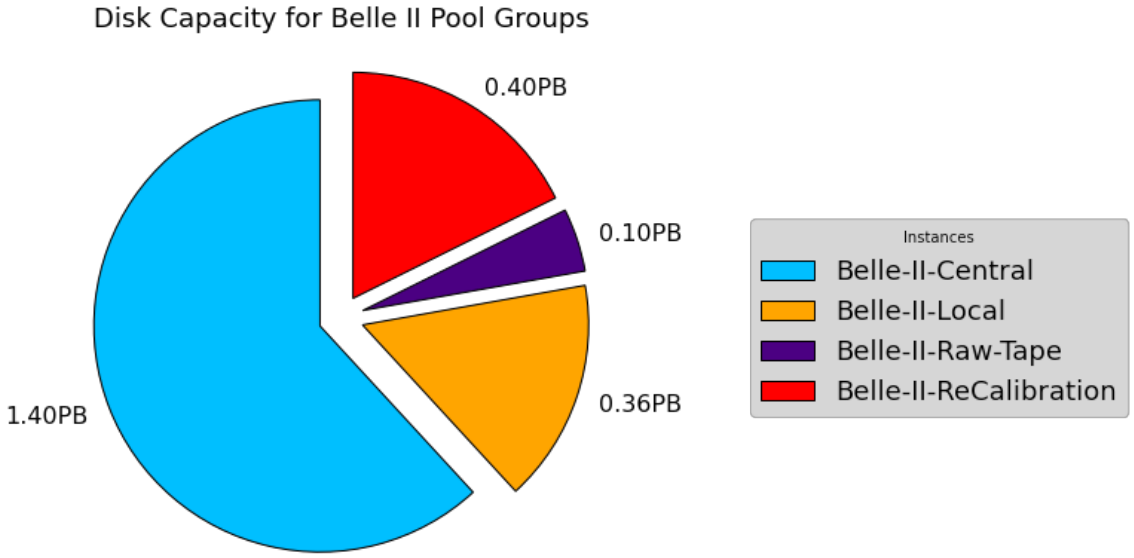
dCache Instances at DESY

- ATLAS, CMS, XFEL, Nextcloud
- Photon for PETRA III, FLASH et al., and Machine Group
- DESY for Belle II, ILC, LHCb, small on-site experiments, IT services



Pool Groups for Belle II

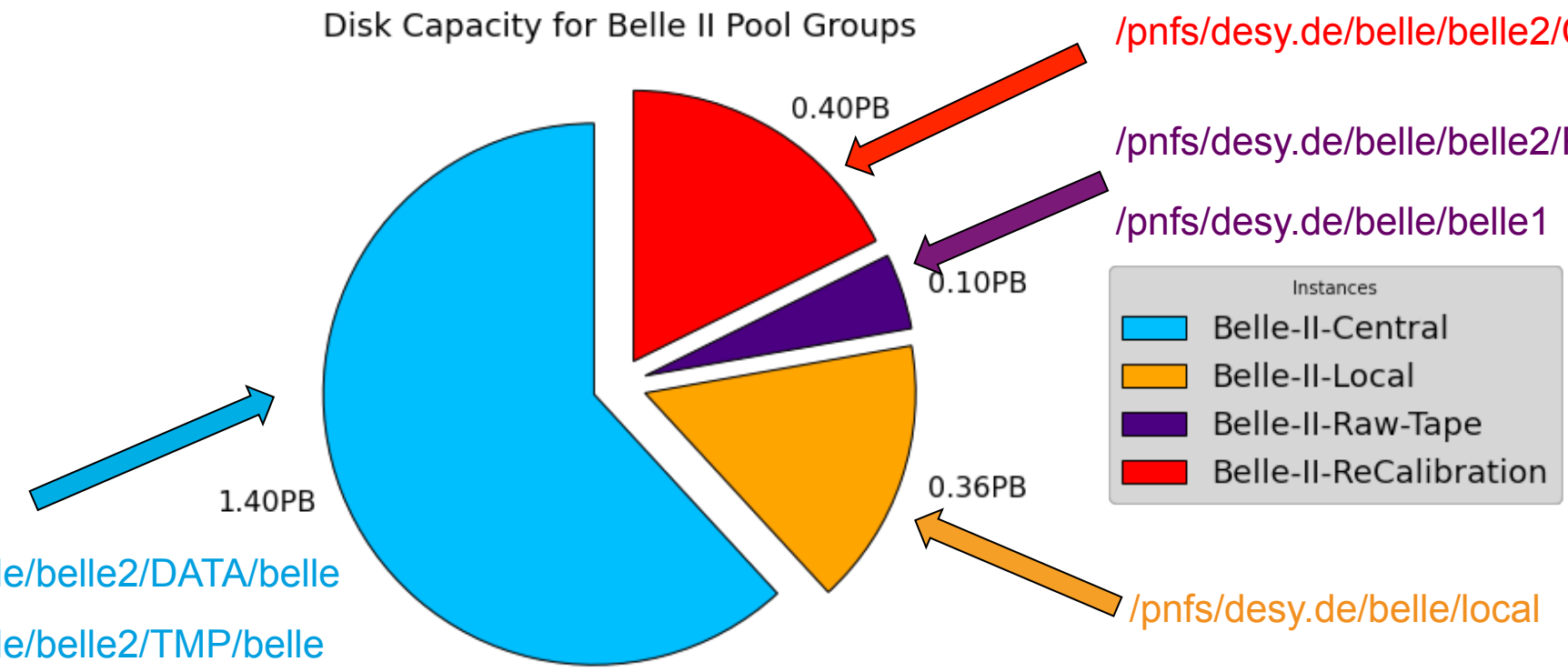
- Belle II takes a significant share of DESY dCache instance
- Pools mentioned before are organised in pool groups
- Pool configuration typically unified across pool group (e.g. enable (re-)storing to/from tape)
- Size of pool group limits the data that can be stored
- Hard limits and are not connected (i.e. free space in Belle-II-Central can not mitigate if Belle-II-Local is full)



Pool Groups from User Point of View

What e.g. Belle-II-Local Means for You

- Pool groups are connected to certain paths in the name space
- In theory: each directory incl. subdirectories can point to a different pool group



/pnfs/desy.de/belle/belle2/CAL/tape

/pnfs/desy.de/belle/belle2/RAW/tape

/pnfs/desy.de/belle/belle1

/pnfs/desy.de/belle/local

/pnfs/desy.de/belle/belle2/DATA/belle

/pnfs/desy.de/belle/belle2/TMP/belle

/pnfs/desy.de/belle/belle2/CAL/belle

Access these directories with all supported protocols in Grad and NAF

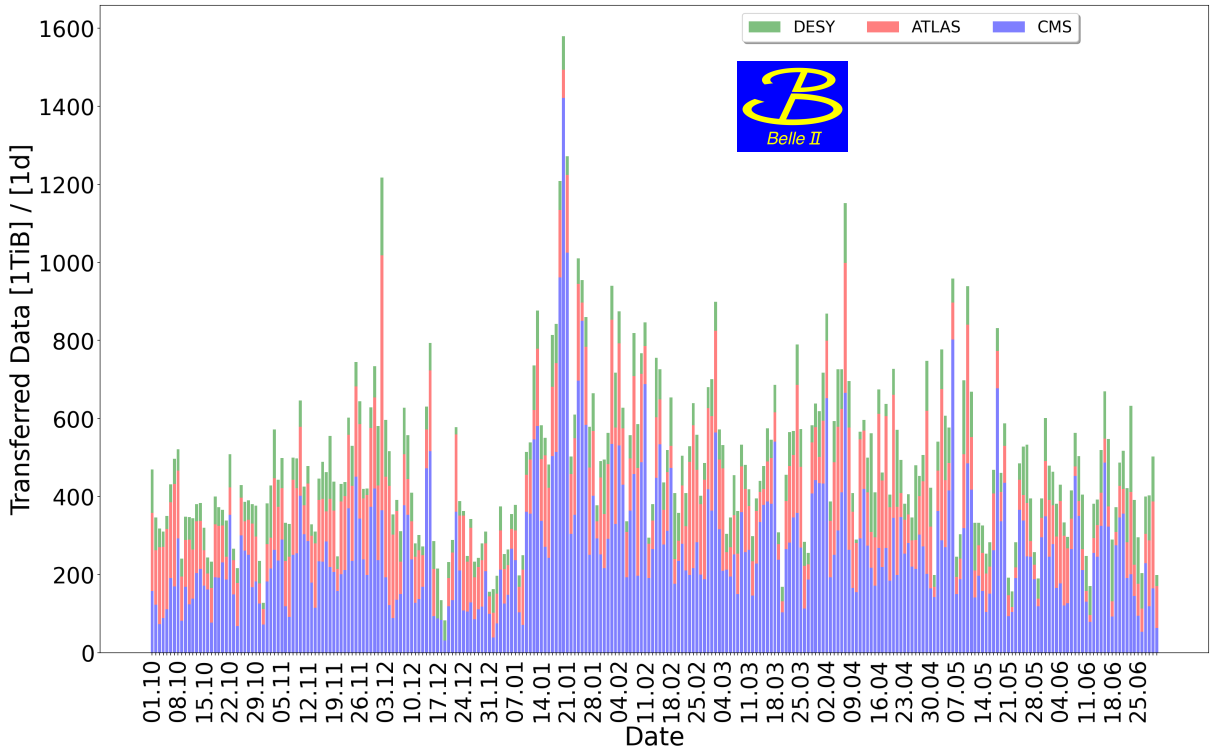
How to Access Data

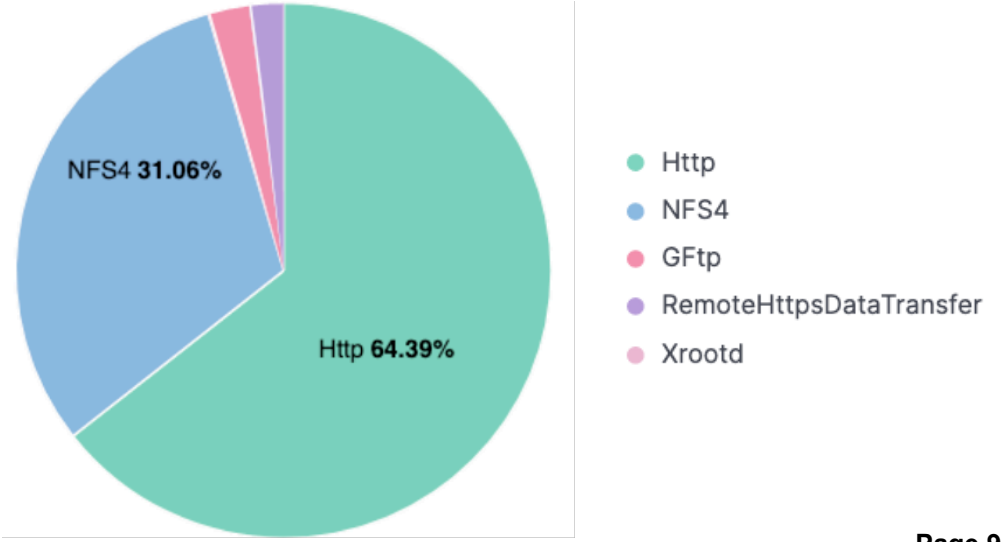
Making Use of Possible Protocols

- Need to know the directory a file is located in
  - The catalogue
  - Path otherwise known
- Access the unified namespace through any of the published endpoints

```
srm://dcache-se-desy
davs://dcache-desy-webdav.desy.de:2880
root://dcache-desy-xrootd.desy.de:1094
```

- Grid-workloads read/write with WebDAV
- Access on NAF dominantly NFS
- FTP/SRM → tape interactions





Additional Complications with Data on Tape

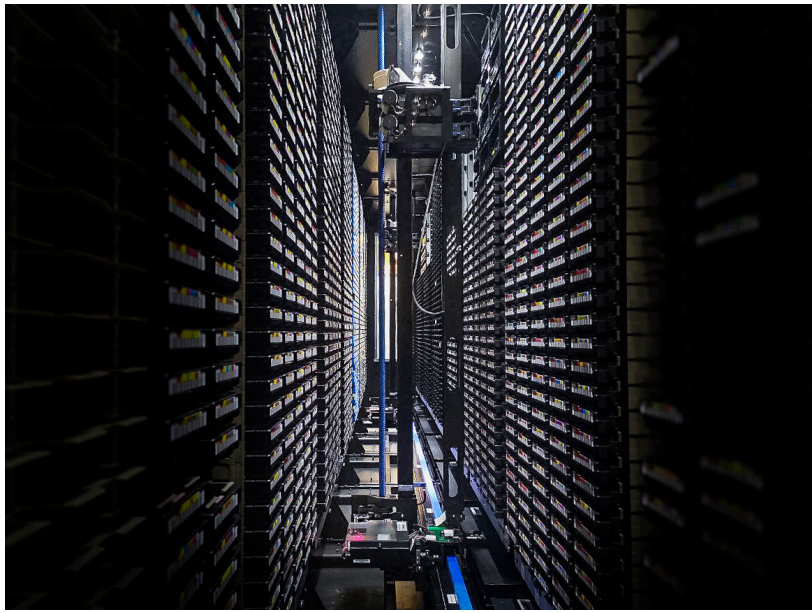
How to Access Data located on Tape

- DESY is a Raw Data Centre: provide tape storage for Belle II
  - Second tape copy for RAW data taken at KEK
  - Copy of the Belle I dataset
  - Re-calibration jobs stage data similar to ATLAS tape carousel
- Adds another layer of complexity for users: file locality:
  - ONLINE (copy on disk)
  - NEARLINE (copy on tape)
  - ONLINE_AND_NEARLINE (copy on disk and tape)
- Accessing file with a disk copy is safe; accessing a file only on tape can cause problems not just for you
  - NEARLINE files cause the NFS client on NAF to wait, blocking the whole node!
  - NEARLINE files cause grid-jobs to be idle wasting slots and CPUhs
- Check the locality before submitting your jobs: https://confluence.desy.de/display/BI/Germany+DESY
- Ensure the disk location is valid for the run-time of your jobs: ‘pinning’ the files

Peculiarities of Data on Tape

Tape Behaviour in Comparison to Disks

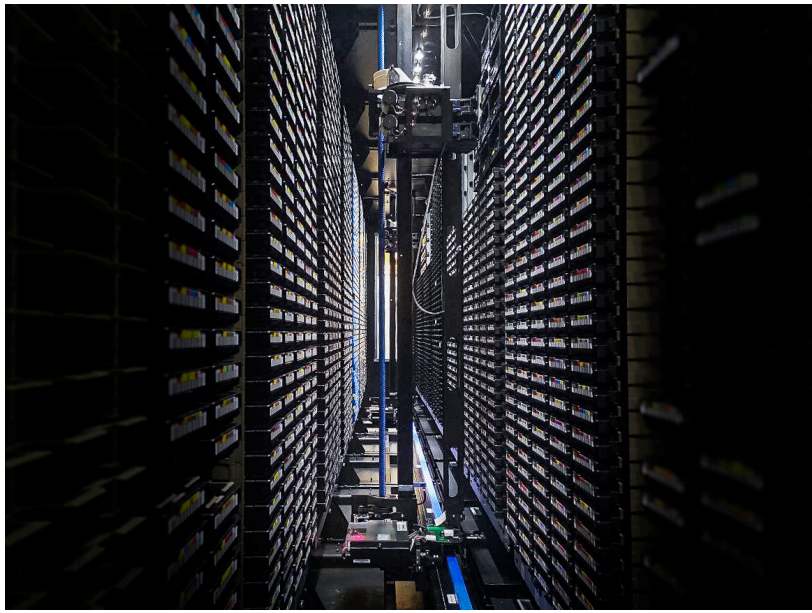
- Remember the days of old → tape operates in much the same way
- Tapes are a streaming media:
  - Streaming r/w performance of modern drives: up to 400MB/s
  - Ideal: write or read one tape in one single operation (listen to an album in one go)
  - Random I/O terminates performance → restores especially costly (remember looking for that one song)
  - Small files lead to massive performance issue in writing and reading
- Tape still in heavy use:
  - In 2022 we've stored ~50PiB on tape mostly for European
  - Overall: about 120 PiB on tape (~1PiB Belle I/II)



Peculiarities of Data on Tape

Tape Behaviour in Comparison to Disks

- Remember the days of old → tape operates in much the same way
- Tapes are a streaming media:
  - Streaming r/w performance of modern drives: up to 400MB/s
  - Ideal: write or read one tape in one single operation (listen to an album in one go)
  - Random I/O terminates performance → restores especially costly (remember looking for that one song)
  - Small files lead to massive performance issue in writing and reading
- Tape still in heavy use:
  - In 2022 we've stored ~50PiB on tape mostly for European
  - Overall: about 120 PiB on tape (~1PiB Belle I/II)



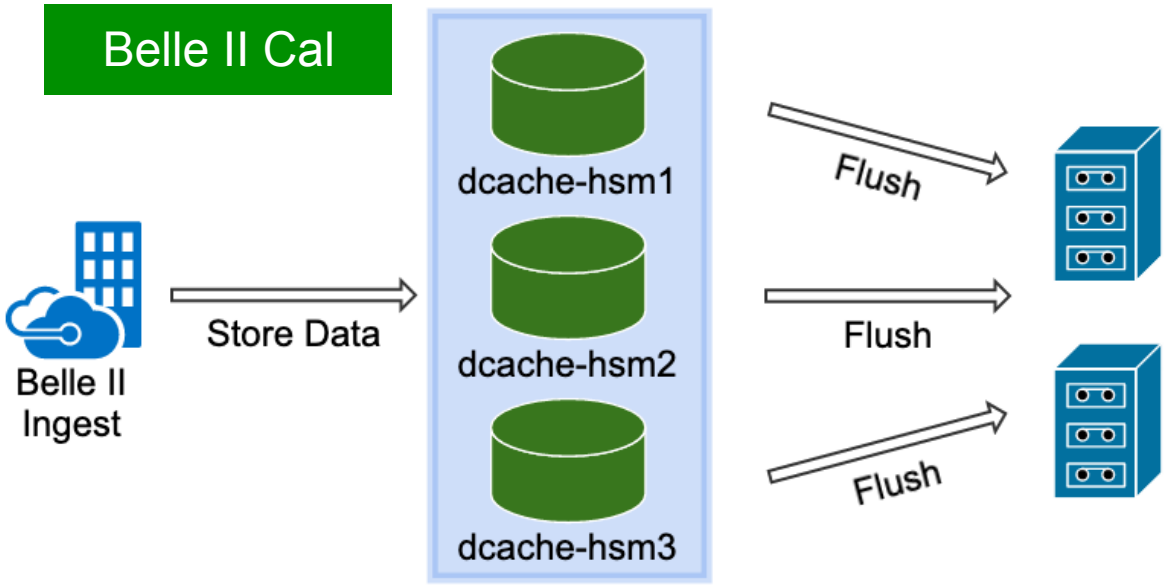
Why are tapes still so popular

- Other magnetic media replaced long ago
- Economics: high investment but very limited running costs compared to disks
- Per TB cost: 18TB of compressed data for 140Euros, disks: 50Euro per TB
- If used correctly: good streaming performance

DESY: Tape Setup for Belle II - Calibration

Optimising Contradicting Requirements/Workflows at DESY vs. Belle II

- Maximize throughput → have a large number of pools that can flush
- Utilize space → avoid empty storage areas due to data policies



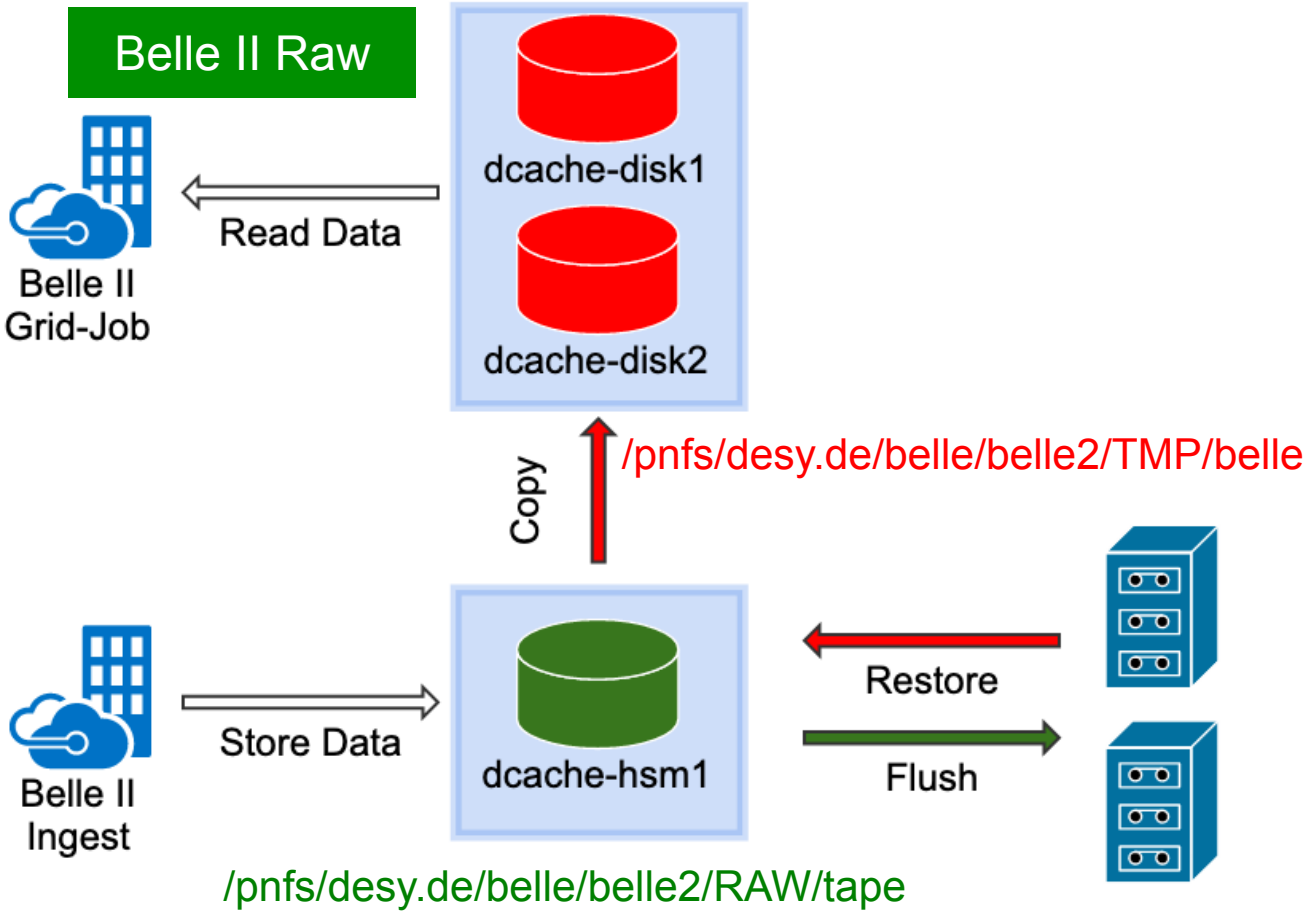
/pnfs/desy.de/belle/belle2/CAL/tape

- Data written into single directory tree
- All files scheduled to tape
- Awareness of file location:
  - After store, files become cached: ready to be removed
  - Files that should remain in disk: pinning
- Planned and coordinated which cycle needs to be on disk during analysis
- Post campaign: unpin the files
- Overall: good fit between throughput and space usage

DESY: Tape Setup for Belle II - Raw

Optimising Contradicting Requirements/Workflows at DESY vs. Belle II

- Maximize throughput → have a large number of pools that can flush
- Utilize space → avoid empty storage areas due to data policies



- Data written into single directory tree
- All files scheduled to tape
- No awareness of file location:
  - After store, files become cached: get removed

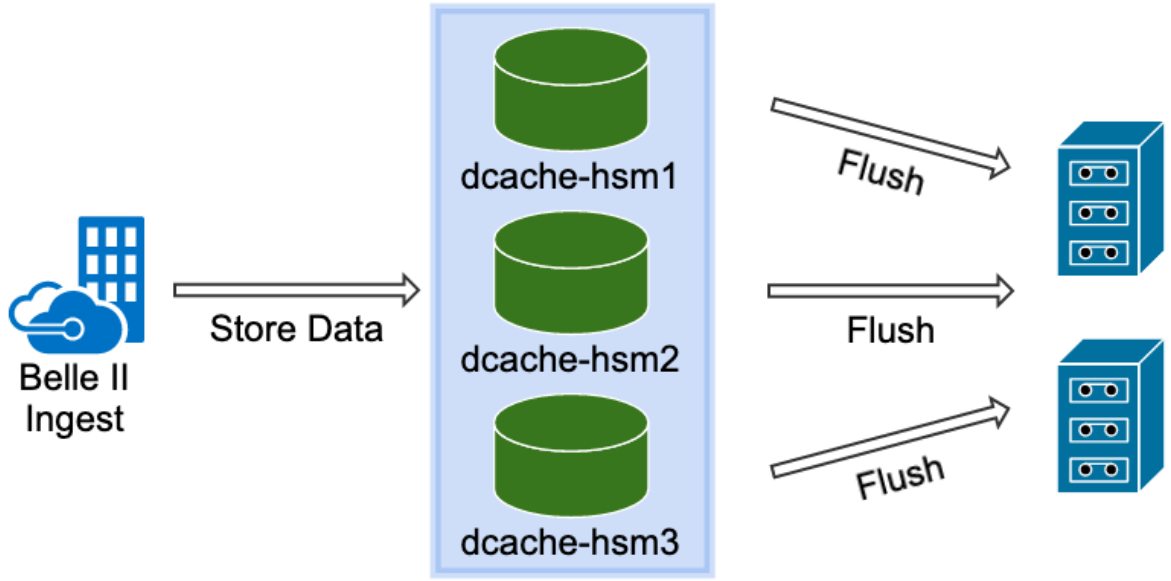
- Centrally managed restore campaigns
- Limited pinning
- Copy from tape area to regular disk area

Difficult to optimize flush throughput vs. underutilised disk space

Changing Configuration to Match Belle-Workflows better

From Two Pool Groups to One using Features in dCache

- Unify the centrally managed pool group with HSM pool group
- Make use of directory configuration to split tape and disk files
  → improve bandwidth and utilize complete disk space
- Pinning procedure found to be error prone (lifetime of pin too short)
- Propose to a more flexible solution
  → Quality of Service to set/update state when needed
- Offer the HSM resources to different Belle users



A Final Slide on Tape: Efficient Restores

Optimise Tape Families

- Reminder: tape is most efficient when used in streaming → connected files on the same tape
- E.g. spent long discussions with local experiments on most efficient tape patterns
- Almost no influence on organization in our Belle namespace
- Users have no way of knowing/checking this
- Need tools to optimize restores, optimize the few free moments during other stores

Tape Recall Scheduler in dCache

- Initially developed for KIT for the ATLAS tape carousel
- DESY helped in testing using the re-calibration data → experience massive blocks due to inefficient recalls
- Uses the information of which file is on which tape
- Schedules only, when a certain percentage of files are requested (or timeout is met)
- If your recalls take long to finish this might be the reason

Summary

dCache Storage for Belle II at DESY

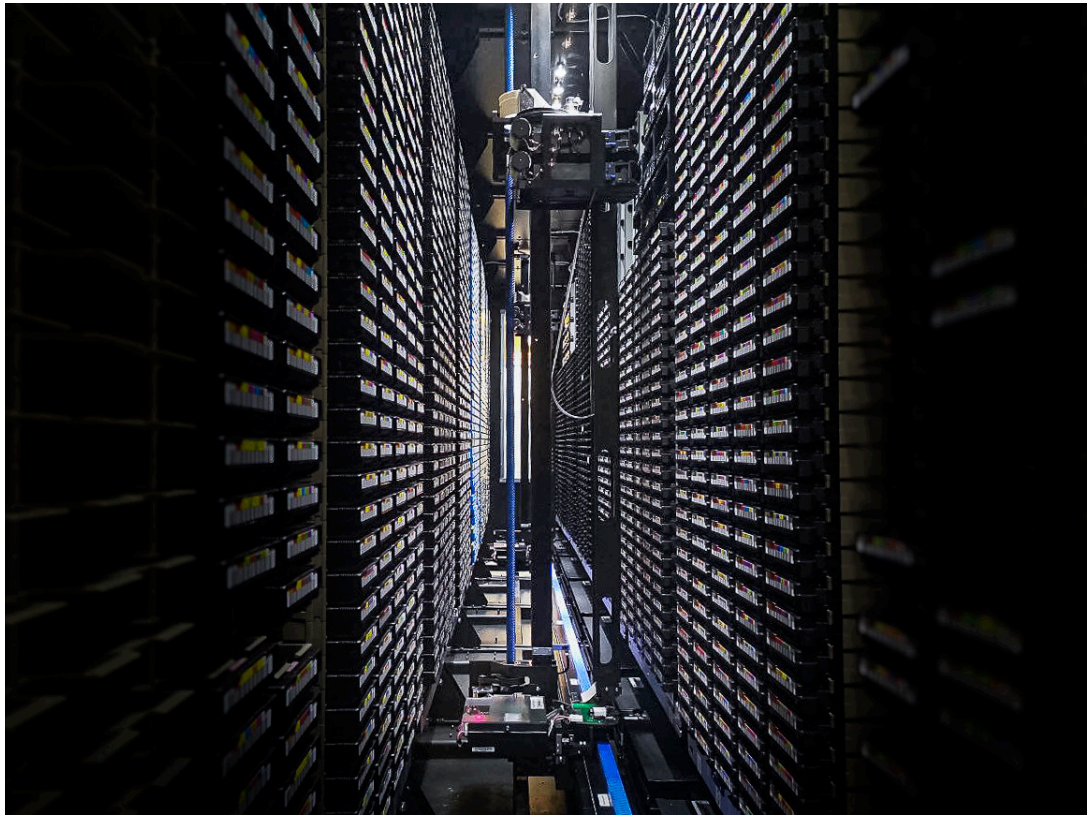
- DESY hosts together with KIT a large share of the German storage pledges for Belle II
- dCache storage offers a uniform namespace available via NAF and Grid (DESY and remote)
- Support all protocols requested by Belle II
- DESY offers local storage for user data
- DESY offers storage for recalibration workflow run on NAF
- DESY offers tape storage for raw and calibration data

Thank you

Tape Resources at DESY

Library and Drives

- Three tape libraries, one to be decommissioned
- Two libraries in production since 2020, IBM TS4500
- Support open standard: LTO in generation 8 (11TiB) and 9 (18TiB) → in use for Belle II
- Support enterprise standard: IBM Jaguar (20TiB) → in use for Photon Science
- Drives:
  - 20 Jaguar drives
  - 16 LTO 9 drives
- Peak Performance: ~4GB/s



Computational Requirements of Different Communities

High Throughput Computing vs. High Performance Computing