Interdisciplinary` Data and Analysis Facility

IDAF:

- what it is in a nutshell
- brainstorm: IDAF & ST1 federated infrastructures

Christian Voß & Yves Kemp

MT-DMA ST1 synergy workshop

DESY 10.11.2023

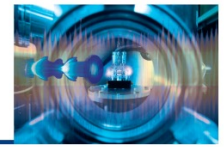
HELMHOLTZ

DESY.

IDAF in the PoF IV Proposal

Goal of the IDAF

The Evolution of the LK II Tier-2 Facility



- **Recommendation**: Tier-2 LK II Facility should support additional user communities
- Observation throughout all Programs in Matter
  - Growing data deluge ➡ Important to **access** and **analyse** large amounts of data

Necessity for a facility to store and analyse data with access for all scientists within Matter.

MT IDAF

↳ From LK II Tier-2 ➡ Interdisciplinary Data and Analysis Facility

- Association with MT
  - MT is interdisciplinary ➡ IDAF is moved from MU to MT
  - Current setup planned at DESY (very broad matter community, experience with Tier-2)

Interdisciplinary Data and Analysis Facility

Supported Communities

- **Accelerator Data**

FLASH. Free-Electron Laser FLASH

European XFEL

FF

- **Accelerator Development Data**

SINBAD

- **HPC simulations**
- **Test-beam data**

Detector and Accelerator R&D

- **Facility User Data**

European XFEL

PETRA III

FLASH. Free-Electron Laser FLASH

- **Data of external Partners**

CFEL SCIENCE

CSSB Centre for Structural Systems Biology

EMBL

Helmholtz-Zentrum hereon

Research with Photons

- **Particle Physics Data**

CMS

ATLAS EXPERIMENT

LHCb

Belle II

ALPS II

ilc international linear collider

- **Astro-Particle Data**

IceCube

H.E.S.S.

cta cherenkov telescope array

Astro- Particle Physics

Services in the IDAF

Small Overview over all Customers

For particle physics communities:

- WLCG-Tier2 & Belle II raw data center
- Complete data lifecycle for local experiments

For photon science communities:

- Direct connection & Tier-0 for large scale facilities at DESY: FLASH / Petra III / EuXFEL
- Complete data lifecycle for these facilities

For accelerator/detector communities:

- Offer storage resources to accelerator division for operating and simulation resources for R&D
- Support for DESY TEST BEAM.

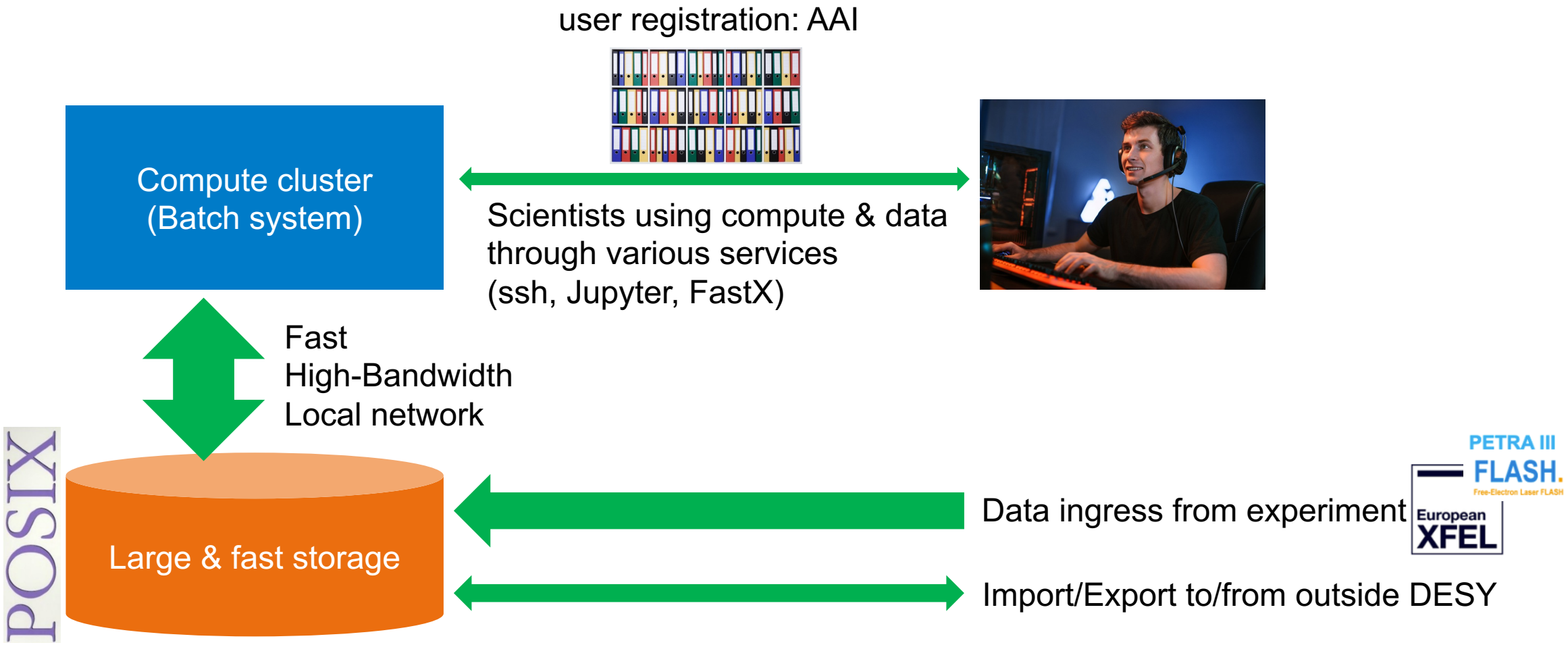
Services for all communities

- Interactivity & fast turn-around: Login-nodes, Jupyter, FastX remote desktop
- GPU resources
- Software installation & distribution, support
- Support of custom containers on clusters

Services on the roadmap

- Integrate data flow pipelines incl. data reduction
- Offer modern analysis tools(e.g. Dask/Spark)
- Integration of catalogues & portals
- Support for OpenData & FAIR

Concept of clusters for data driven science:

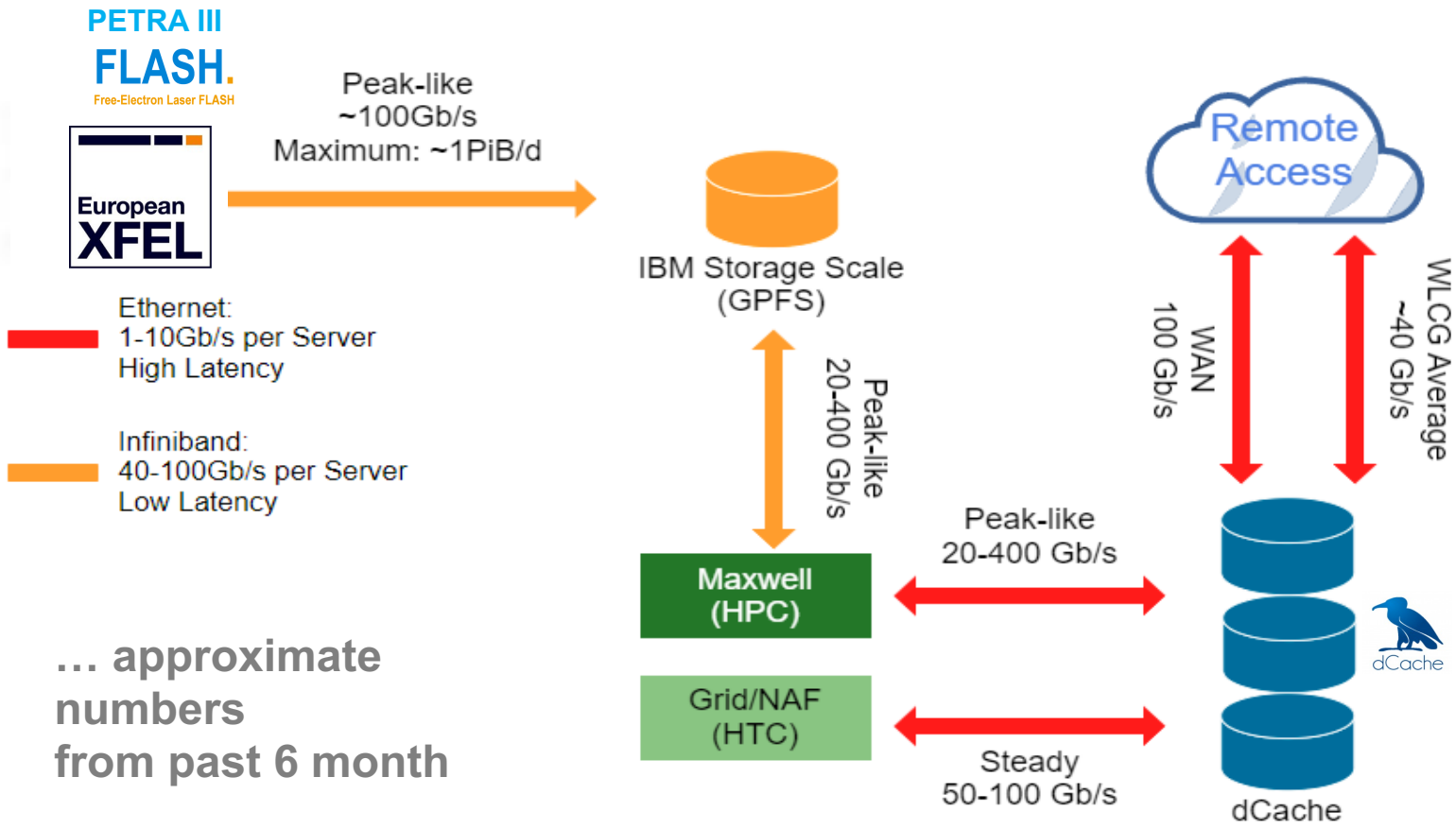


"A strongly interconnected Storage & Compute form the core of the infrastructure.
Users interact with Storage & Compute through well integrated services & portals."

IDAF: Bandwidth for Flow of Data

Connecting Detectors, Storage and Compute

- Ingest rates up to several PiB/day
- Split between HPC and HTC both in compute and storage
- Photon science centred on HPC
- More steady analysis patterns of particle physics centred on HTC
- Overall about
  - 80k cores / 250GPUs
  - 200PiB GPFS/dCache storage
  - Recently extended tape system (stored currently ~150PiB)
  - 1.5k servers
- Very heterogeneous hardware



Glossary

dCache Storage

- Large data storage
- Interface to tape
- Access protocols for WAN & local

GPFS

- High performance data storage
- Interface to photon science DAQ
- POSIX access

Maxwell HPC

- High Performance Compute
- PhotonScience & Accelerator R&D+operations
- Access: interactive services (ssh, jhub, fastx)
- Offline analysis, simulation

DESY Grid Cluster

- High Throughput Compute
- WLCG Tier-2 and Belle 2 Raw Data center
- Access through Grid means
- Production type jobs

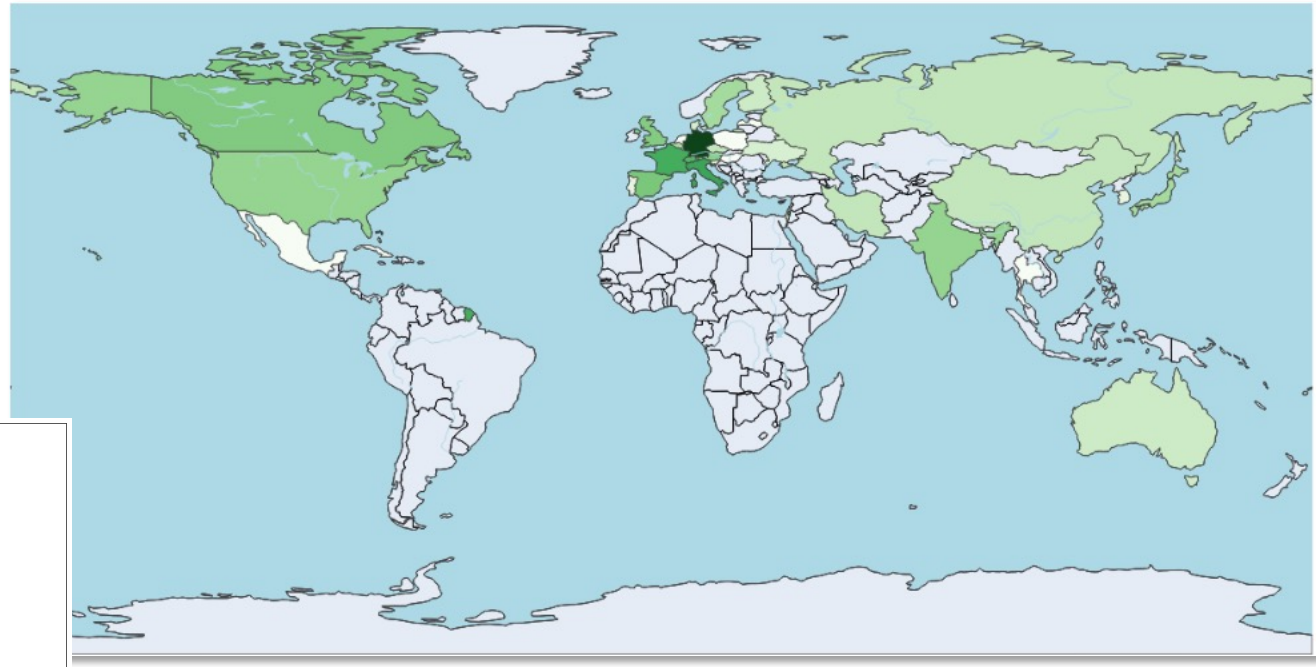
National Analysis Facility NAF

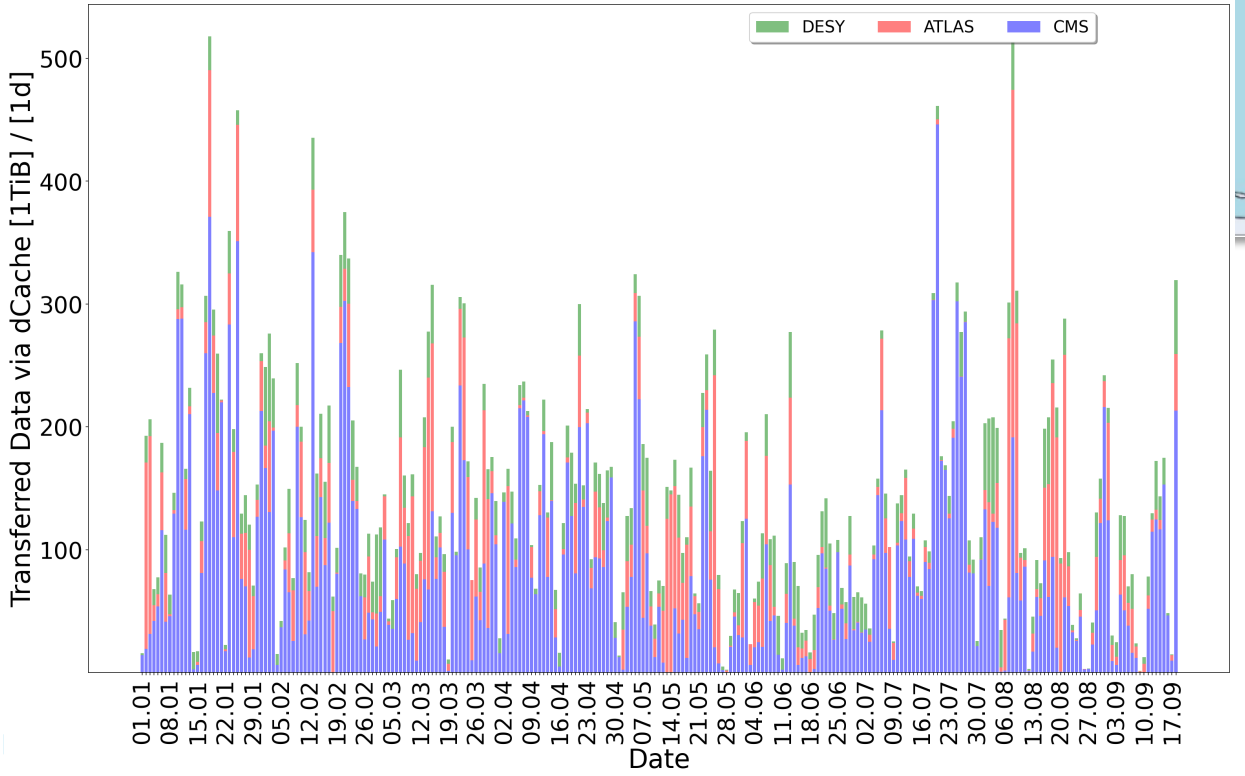
- High Throughput Compute
- German HEP users, Belle 2 users, DESY local
- Access: interactive services (ssh, jhub, fastx)
- Analysis type of jobs

Users of the NAF

Example for a Service with large Number of (inter-)national Users

- Interactive usage of the NAF
  - Most users from German universities
  - All Belle II scientists are potential NAF users
  - Large number of international users





- Data access inside NAF (only dCache shown)
- Additional storage space for NAF (linked to experiment frameworkd)
- CMS as largest contributor
- Jobs do almost exclusively POSIX

Ideas for IDAF ←→ MT-DMA ST1 interplay

Disclaimer:

- The following slides are meant for illustration and discussion purpose
- They do not represent current status of implementation and do not show an approved setup or work plan

Federating the users

Status now

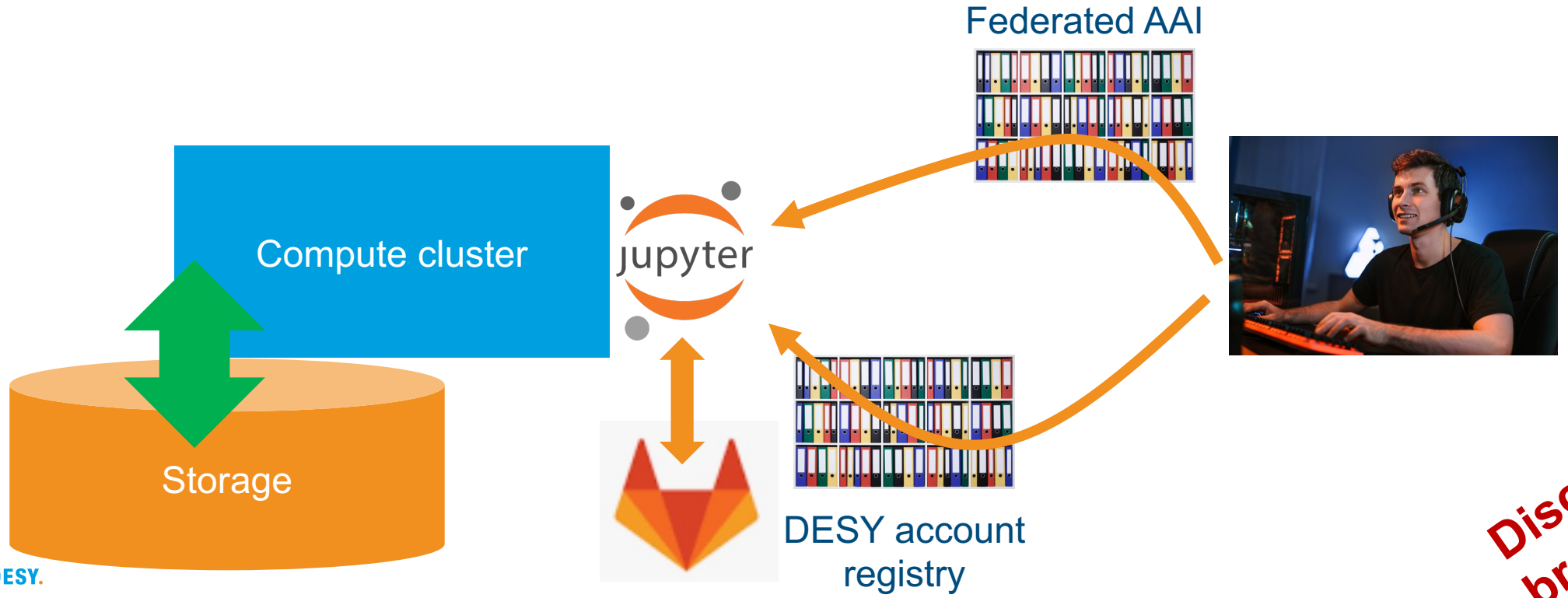
- Users of Maxwell & NAF need a DESY account
- Various options exists for external people
- Idea: Use federated (Helmholtz?) AAI for accessing services?

Opportunities

- Access to compute service *integrated* with storage
- Putting away burden of DESY account creation & care

Challenges

- Getting authorization for external people right
- Mapping different identities of same user
- Resource planning and sharing with other users

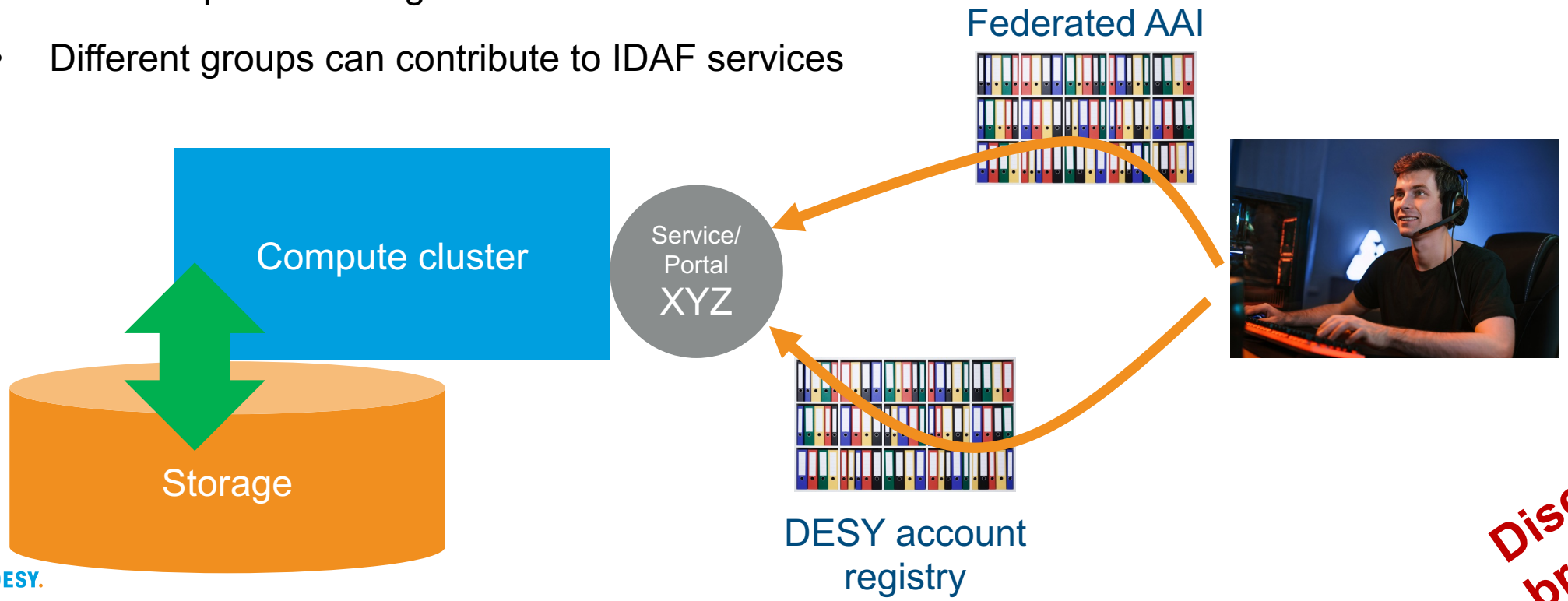


Disclaimer: Idea for brainstormingonly

Offering new services

Outlook

- Once users are federated
- Easy to offer new *integrated* services:
  - AAI established
  - Compute & storage access stable
- Different groups can contribute to IDAF services



Disclaimer: Idea for brainstormingonly

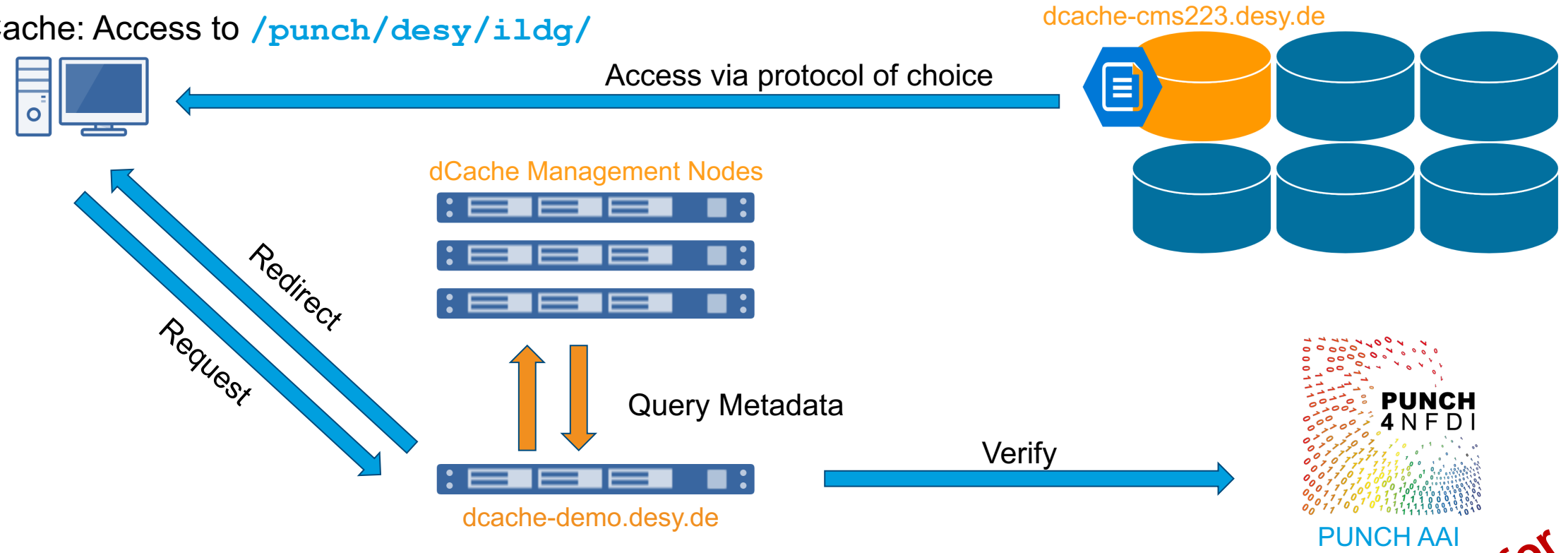
Federating the storage

- Remote integration into PUNCH storage instance at DESY
- Pools at different sites, single federated instance
- Concepts and following slides by Christian Voss

Disclaimer: Idea for brainstormingonly

Basic setup: Standard singe site setup

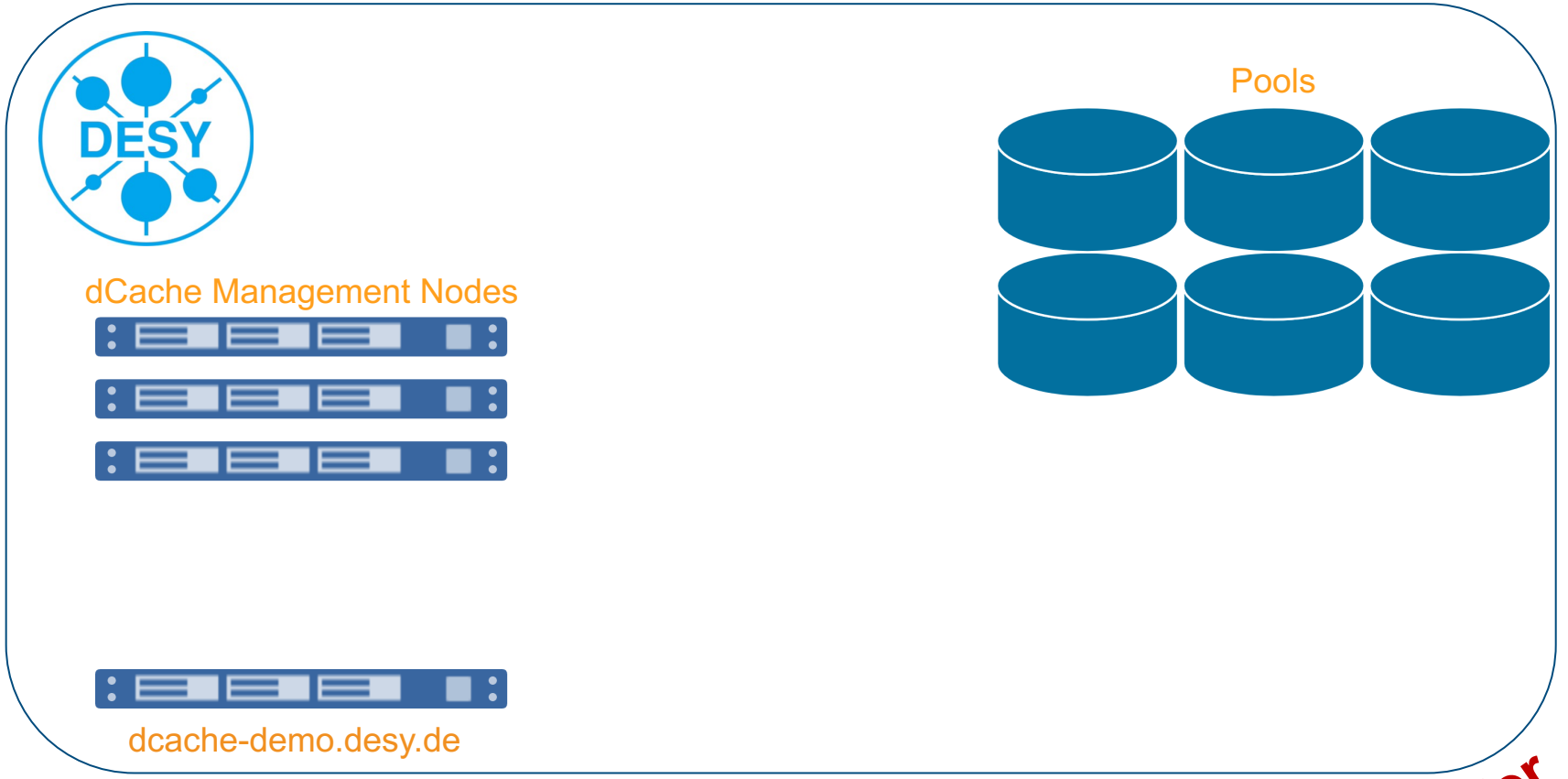
- Use dCache: Access to `/punch/desy/ildg/`



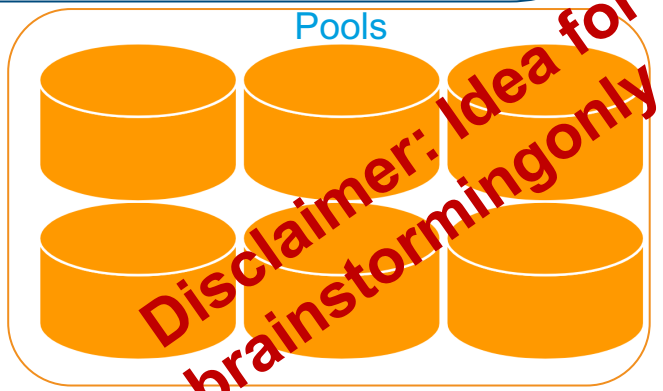
- Scaling horizontally through addition of more storage nodes
- Management of complete namespace
- Verification of login credentials

Disclaimer: Idea for brainstormingonly

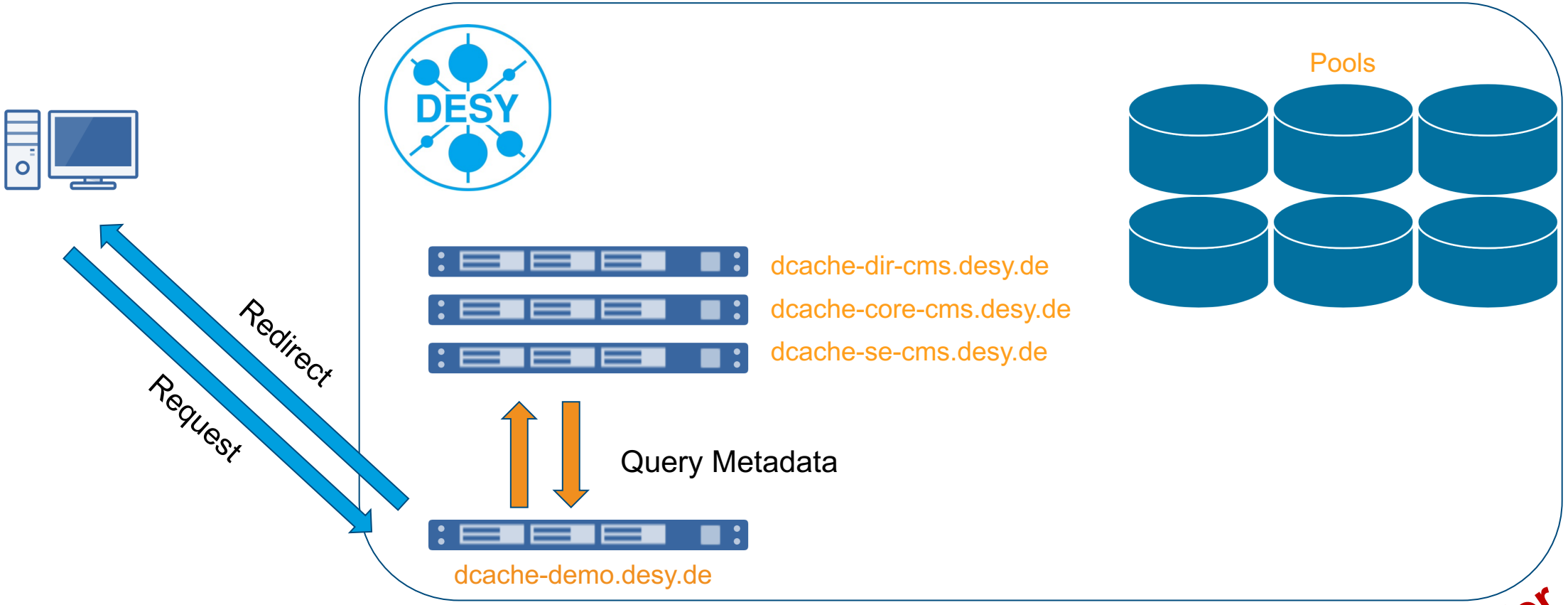
Layout of Federated dCache: Simplest, most centralized Layout



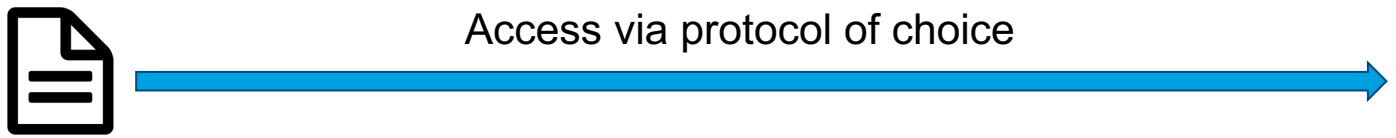
- At remote site only pools are deployed
- All management services located at central site
- Central access-point
- Centralised AAI-interface and namespace

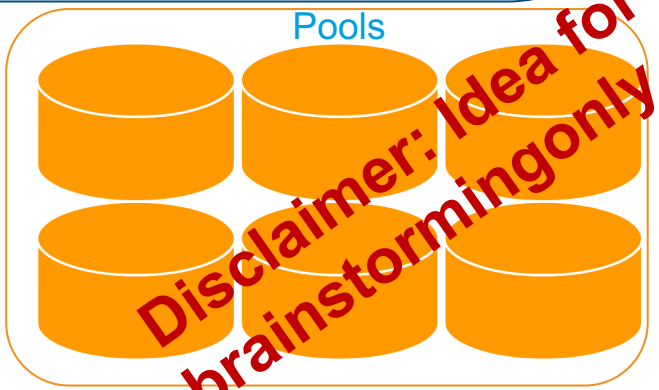


Layout of Federated dCache: Simplest, most centralized Layout



- Use dCache: Access to `/punch/<remote-site>/ildg`





General Features

Advantages and Disadvantages

Requirements

- Depending failure tolerance (networking issue at central site, maintenance at central site)
- Ports from DESY to remote sites (still in discussion)
- Basic: only pools on remote site
  - Ports to DESY: dCache communications (11111), Zookeeper (2181), Kafka (9092)
  - Optional local: data access to pools to/from other remote sites

Deployment

- Container based setup: provide a container for sites to deploy (WIP)
- Stick with established Singularity workflows using CVMFS for distribution
- Local admin offers data directory and ports

- DESY admins: map namespace entry-point to remote storage area

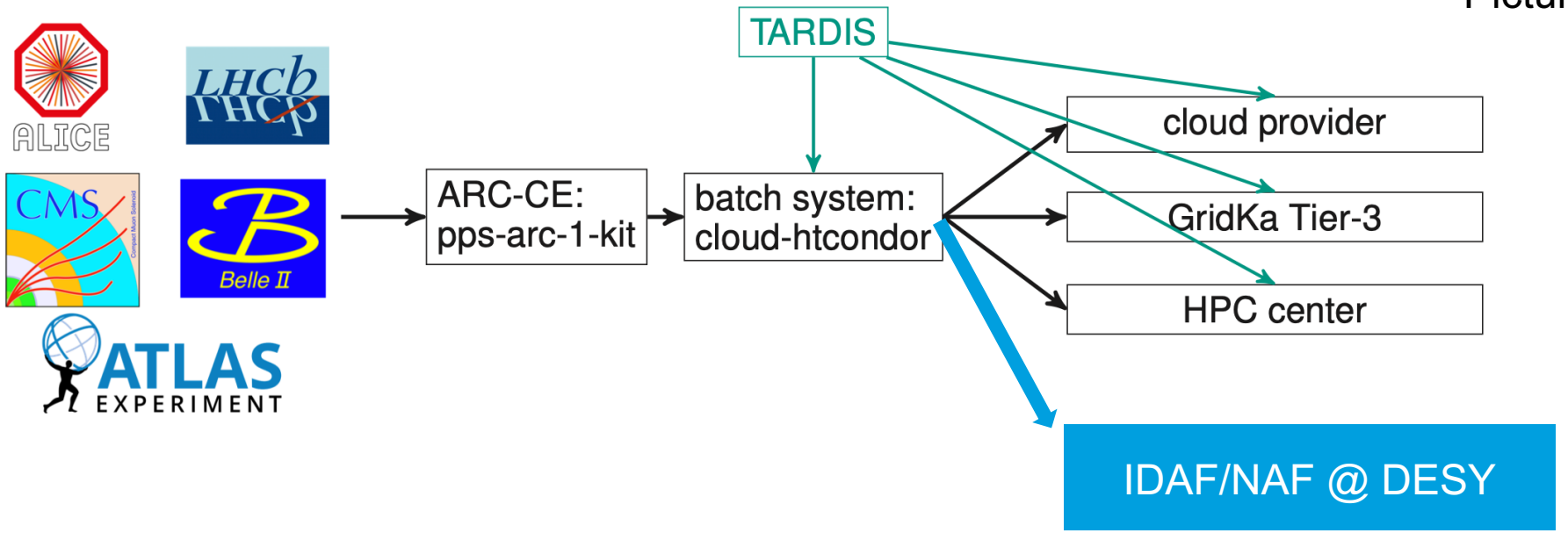
Policy Questions

- Security questions: security of Zookeeper/Kafka
- Opening ports to specific remote sites easier
- Generic openings more difficult/impossible

Disclaimer: Idea for brainstormingonly

Federating the compute: e.g. COBaID/TARDIS

Picture from Matthias Schnepf



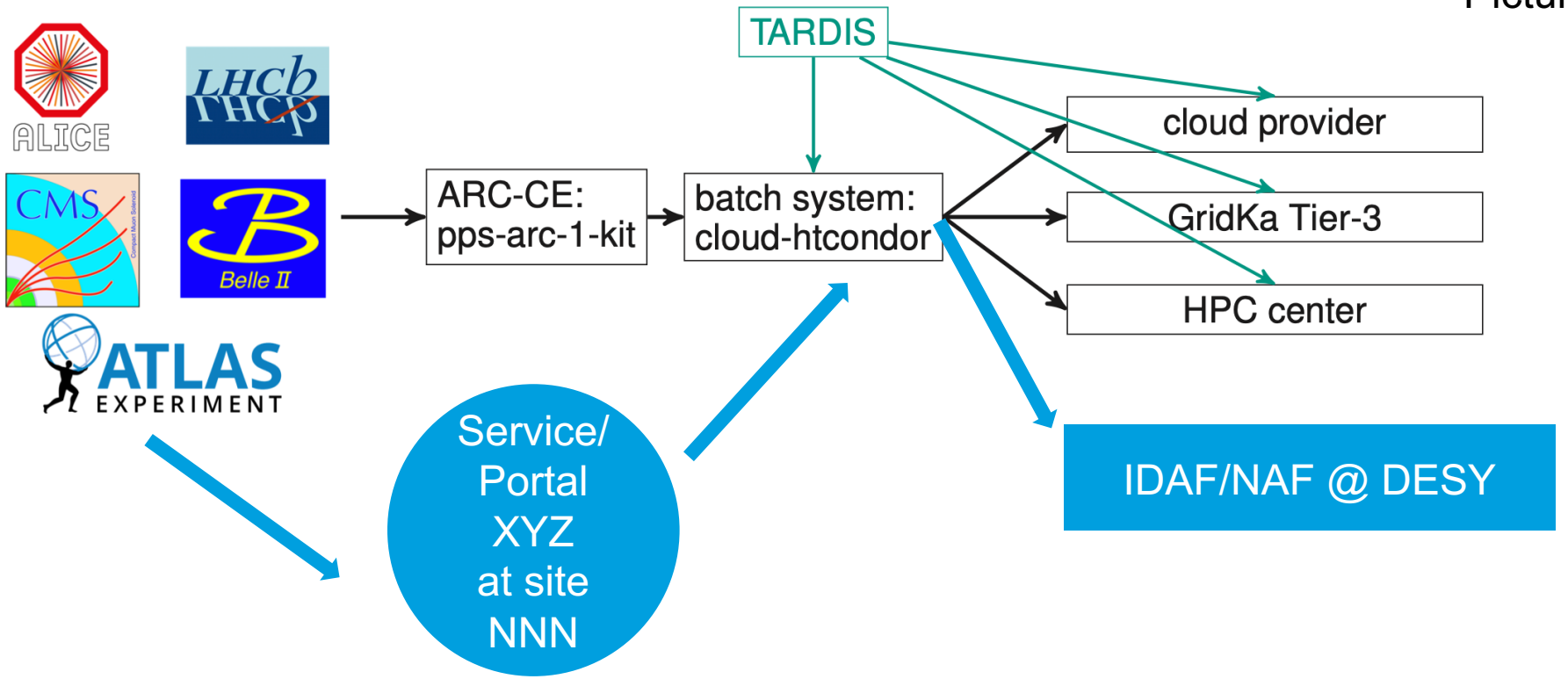
Status & Outlook

- Currently being setup up
- Pilot integration into production NAF setup planned

Disclaimer: Idea for brainstormingonly

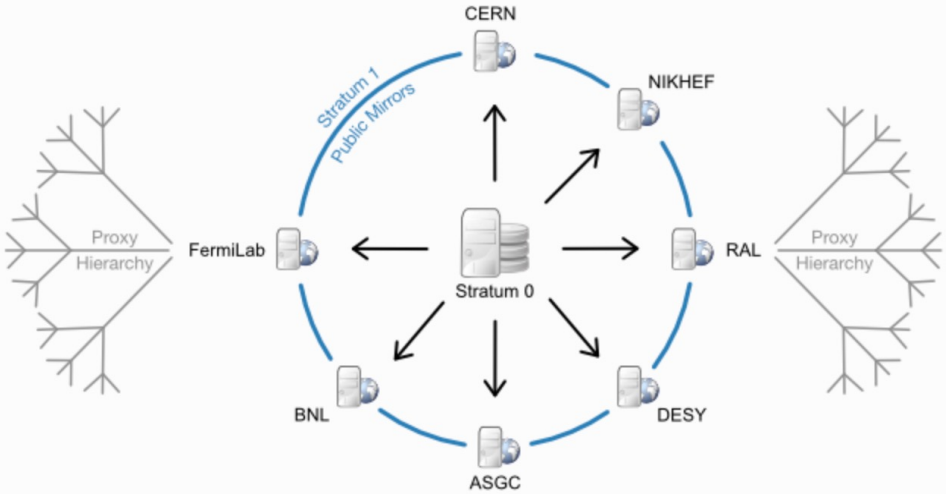
Enabling remote compute services: An IDEA

Picture from Matthias Schnepf



Disclaimer: Idea for brainstormingonly

The Software



Containers & registry

- Running containers must be supported
- Integrated in IDAF
- CI/CD workflows to be established in a federated way
- Federated Gitlab

CVMFS

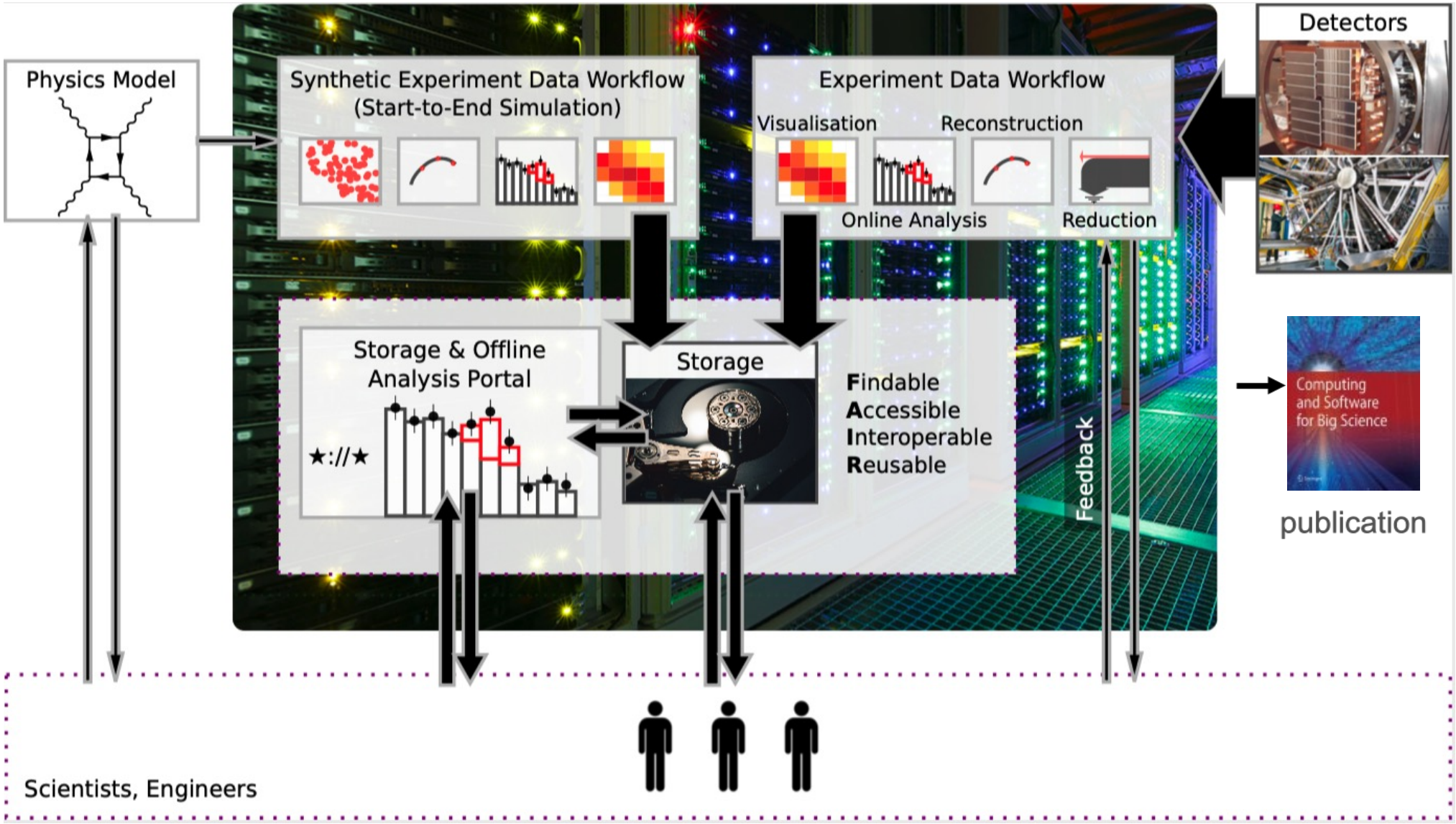
- e.g. central Stratum 0
- Software available at all connecting sites
- Analyze: non-public software

Disclaimer: Idea for brainstormingonly

Backup slides

On-Site: Particle Physics, Accelerators, Photon Science

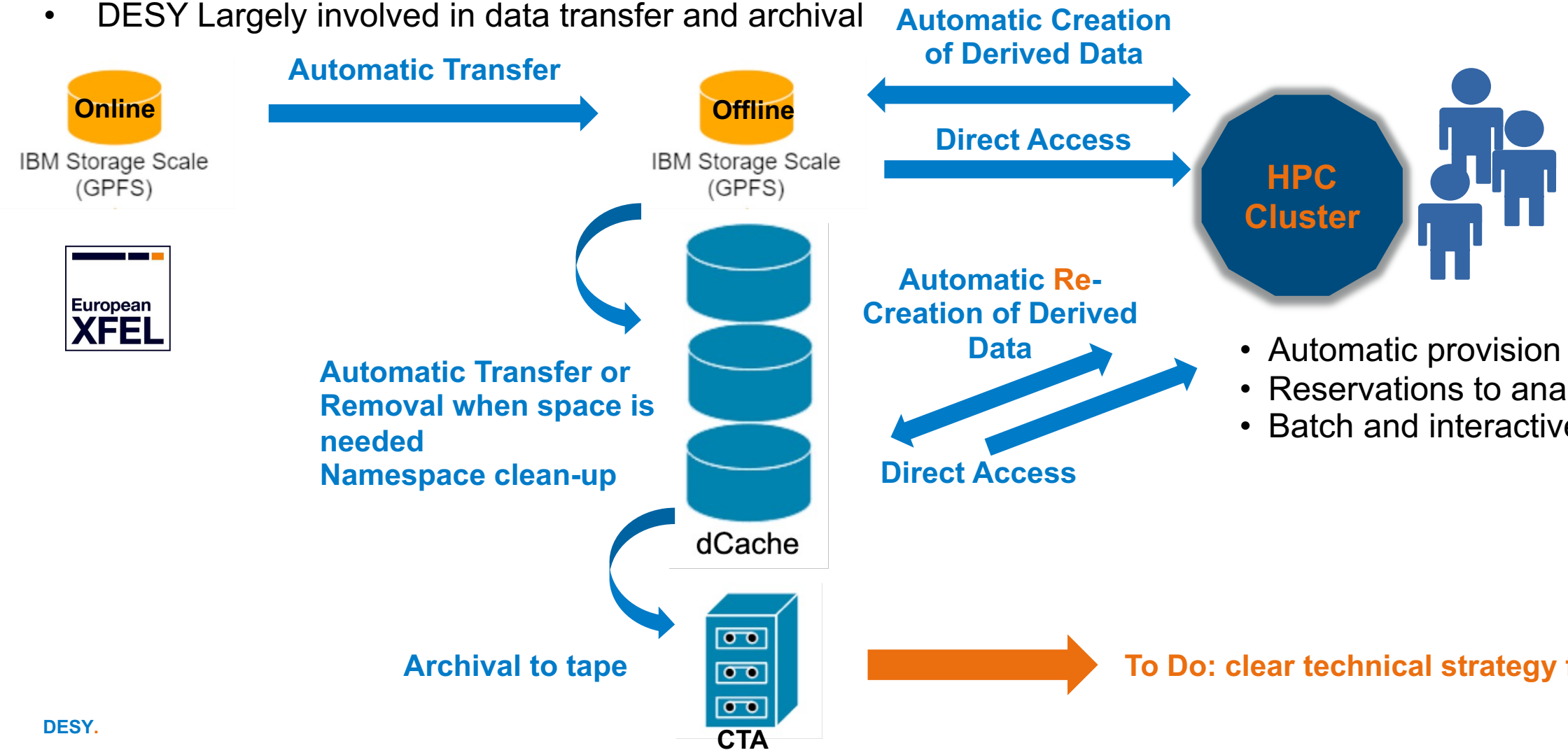
Enable the Full Analysis/Data Lifecycle: From Simulation to Publication and Archival



On-Site Example

User Proposals for European XFEL

- Developed largely by our colleagues at the European XFEL → Analysis is centered on IDAF
- DESY Largely involved in data transfer and archival

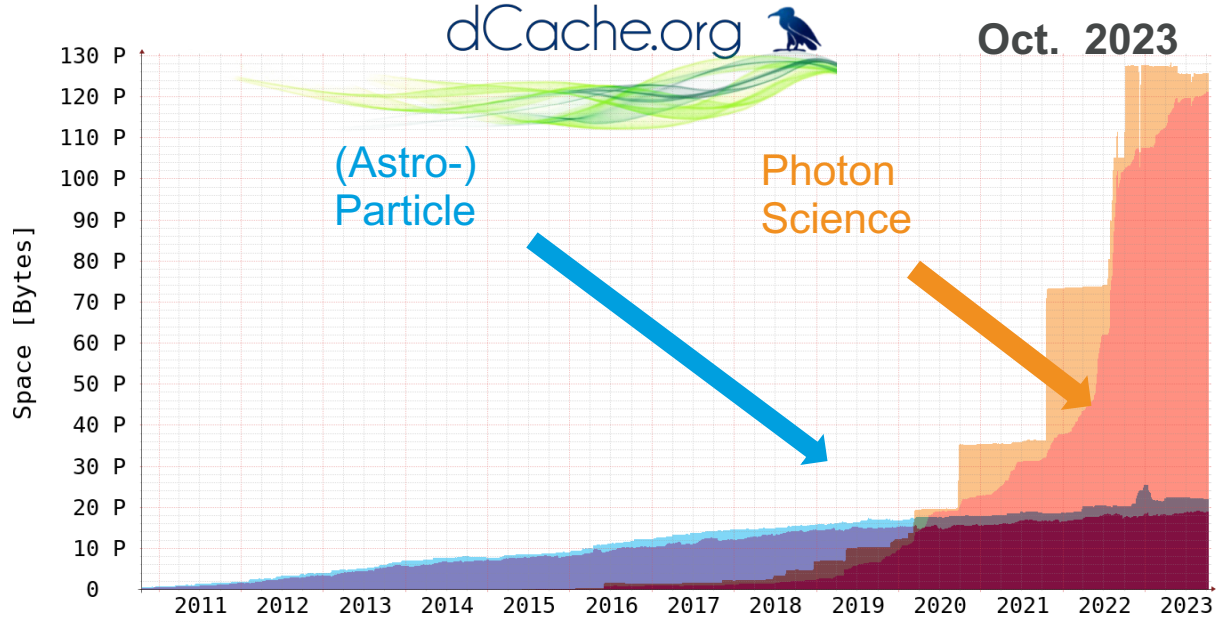


- Automatic provision of Accounts
- Reservations to analyse data
- Batch and interactive analyses

Challenges: Data Deluge in Photon Science

Photon Science and Especially European XFEL Continued to Grow Exponentially

- **Exponential growth for photon science!**
- **Accelerator division starts to contribute (2 weeks of XFEL Linac operation: ~1PiB)**
- HPC cluster storage similarly increased
- Capacity growth slow down/halt during end of 2022 due to funding situation
- Alternative usage of existing capacity
- **More heavy involvement of tape storage** (as done by ATLAS in the WLCG)
- European XFEL still expects to collect 50PiB in 2024



- **Observe scaling issues for the IDAF**
- Number of dCache pools causes issues when rebalancing after introducing new pools
- Pool nodes start pile up in the computing centre: **start experience limits to rack space**

Interdisciplenary Data and Analysis Facility (IDAF)

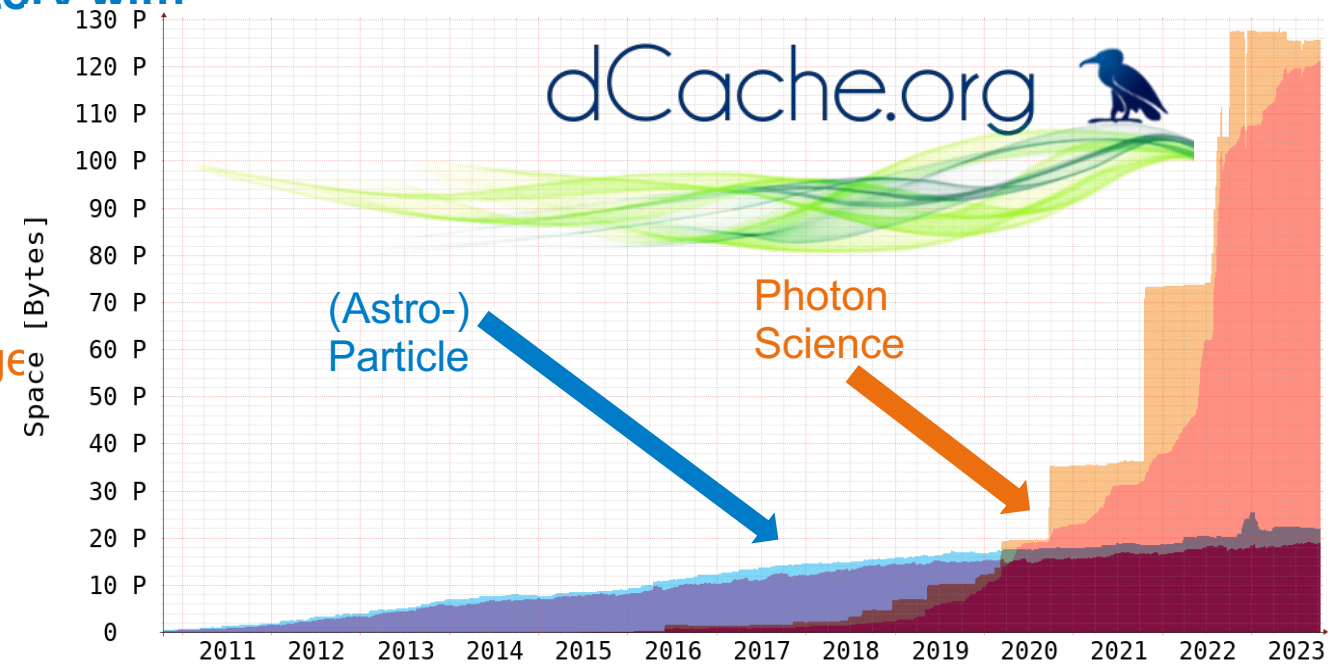
Origins and Overview

DESY historically centred on Particle Physics together with strong accelerator division:

- HERA and original PETRA accelerators
- Discoveries: Gluon and B-mixing

Accelerated transition to an accelerator laboratory with

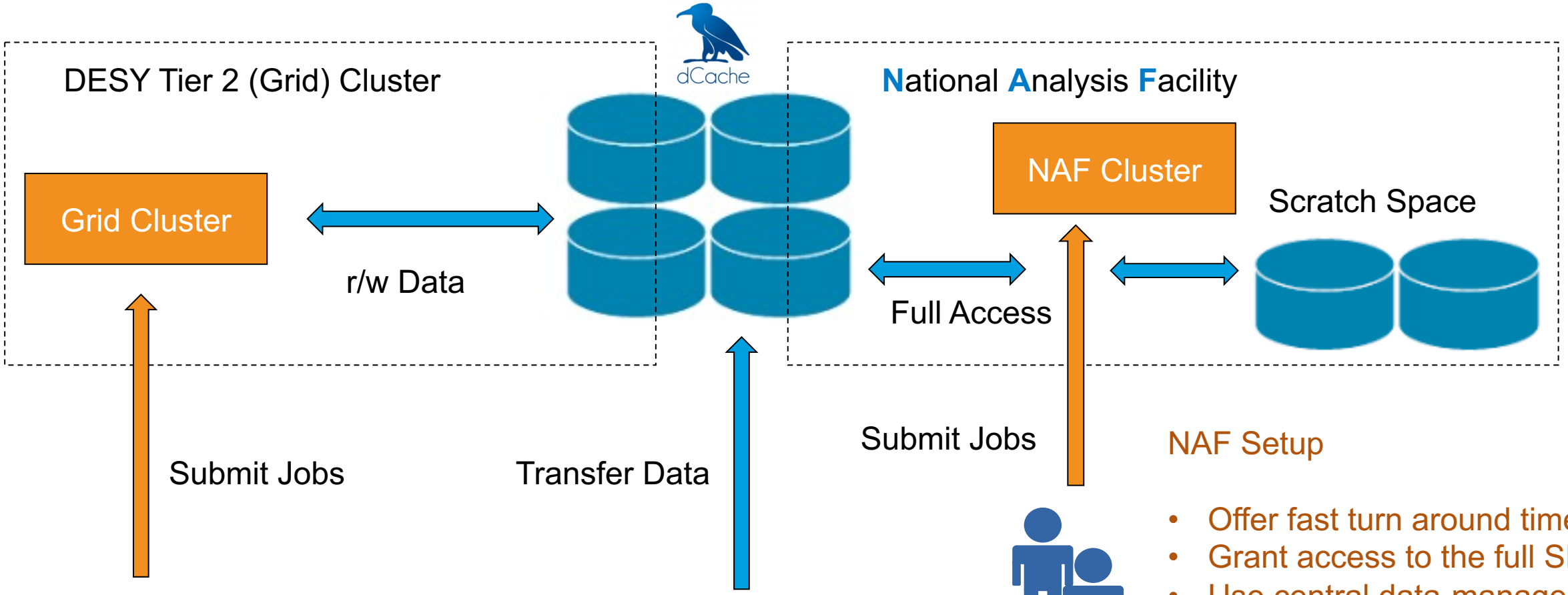
- Large photon science user facilities
- Large local particle physics groups
- Obvious when looking at provided and used storage



Paradigm: Data Analyses are Data Driven

As Underlying Principle of the Particle Physics Infrastructure

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



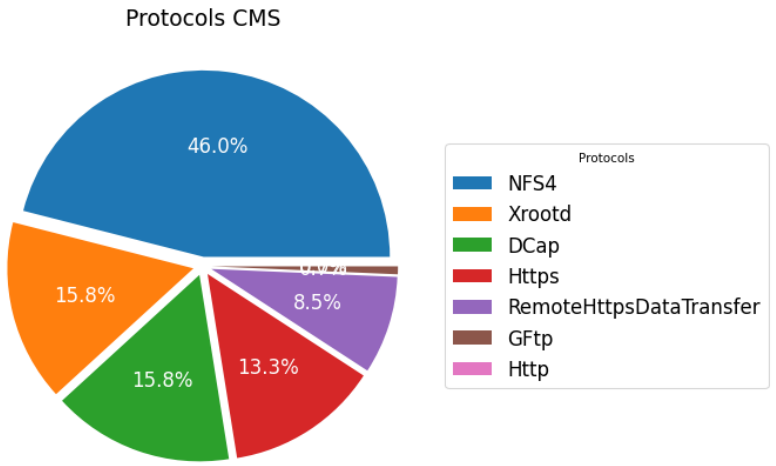
NAF Setup

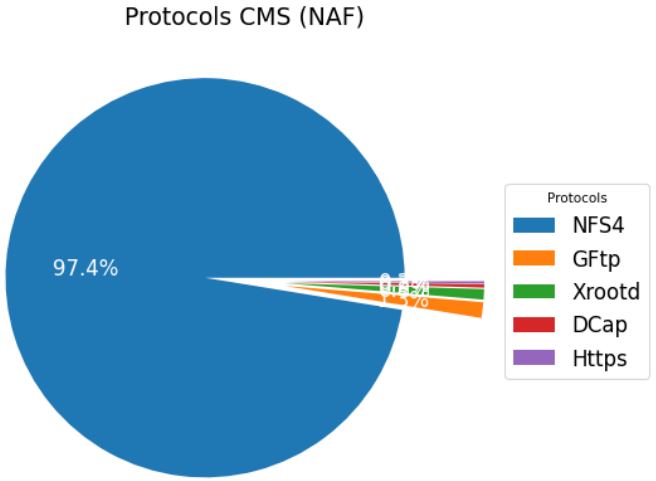
- Offer fast turn around times
- Grant access to the full SEs
- Use central data-management of experiments to store data at DESY

Challenges: The Return of POSIX

POSIX Reliance on Data Access

- We see ever increasing POSIX access pattern
  - Photon science software often can only ready via POSIX (native GPFS mount or through dCache-NFS-mounts)
  - Becomes more and more true for particle physics as well (despite XrootD): On Grid we see XrootD/WebDAV, but on NAF we see >90% NFS (dCache and GPFS)
  - ATLAS less prone, CMS and Belle II use POSIX almost exclusively
  - Depend a lot on the NFS client: Linux discussion from yesterday
  - Strange interaction e.g. with ATLAS Rucio namespace
  - Complicates merging of HPC and HTC part ➜ make sure both share the same namespace
  - How to treat native GPFS on HPC on HTC (again NFS?)
- **Not sure how well the upcoming Analysis Facilities deal with it**

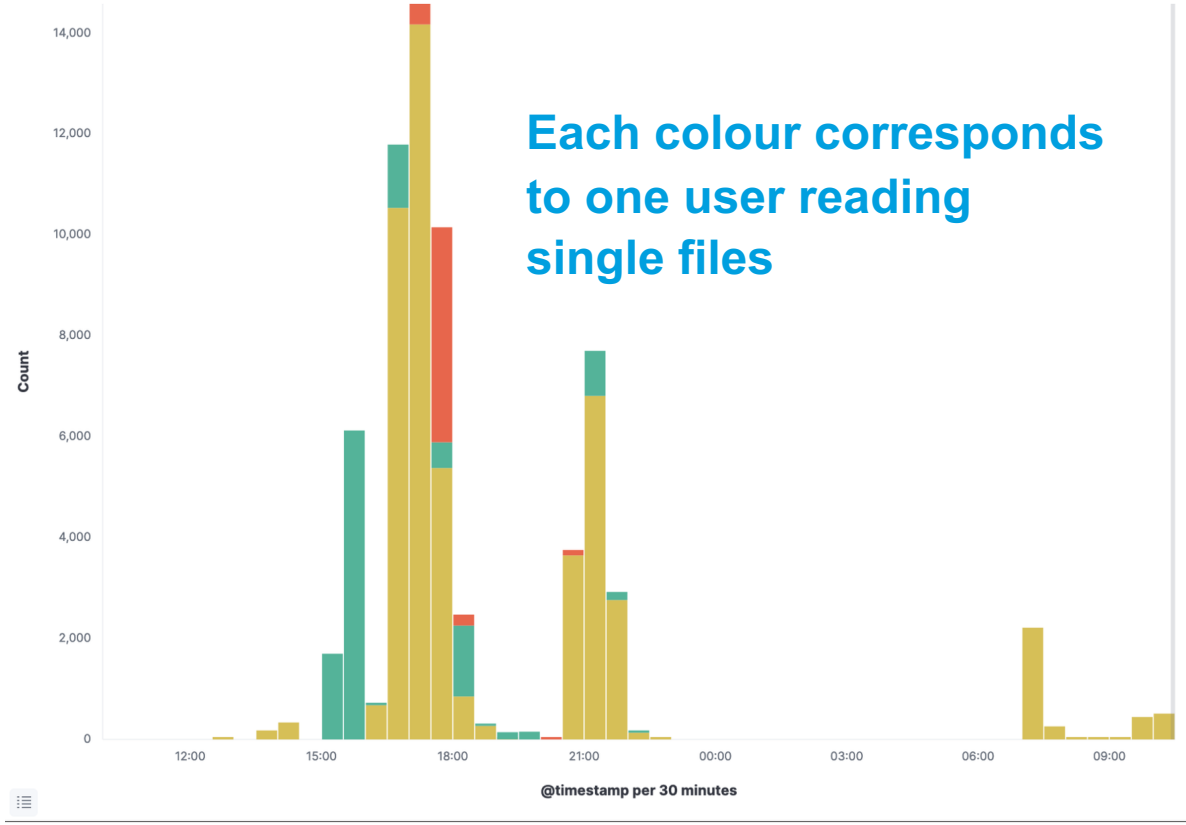


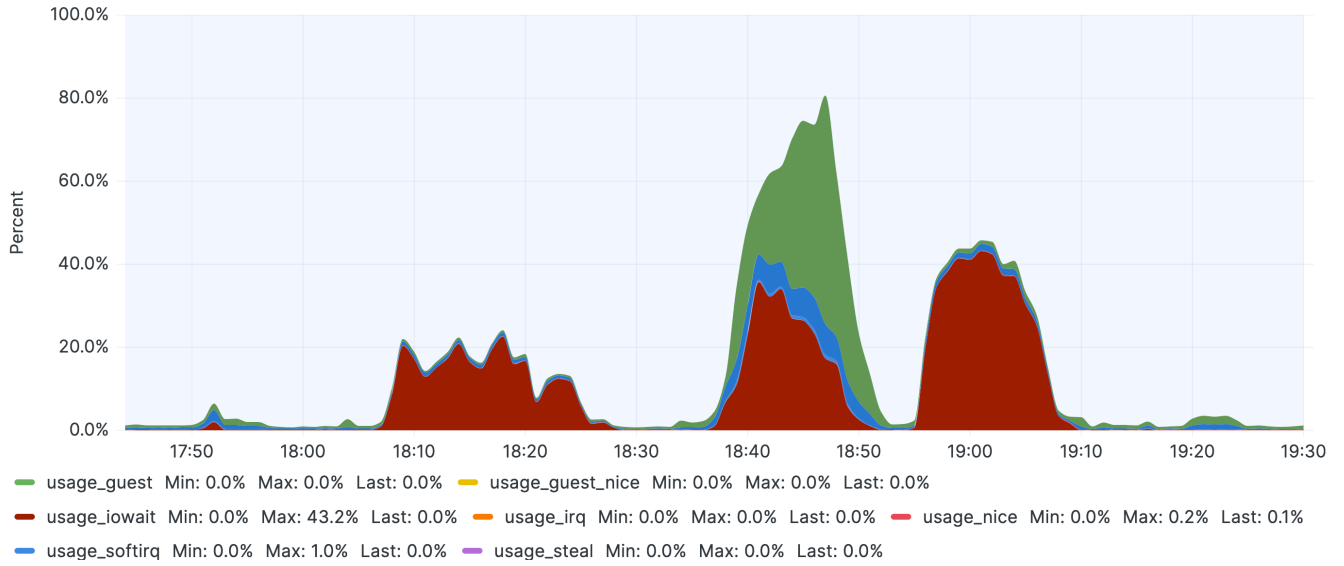


Challenges: Using HTC as HPC

Excessive Access Pattern from HEP Users on NAF

- Classically ideal read pattern: 1 job reads 1 file
- Experience quite aggressive job patterns on NAF
  - CMS users submitting 100k jobs at once
  - Job starting together leads to large number of reads
- Custom frameworks of local trigger many parallel reads
- Overloads dCache storage nodes, turning pools unresponsive
- Causes snowball effect on the worker nodes
- One user can cause the whole NAF to become unresponsive





Challenges: Security

Harden the IDAF against External Threats

- Several German universities and institutes have been hacked recently – also in Helmholtz Association:
  - E.g. Helmholtz Zentrum Berlin (also operates photon science user facilities with external users)
- In the era of federations, a hacked account at $REMOTE poses a danger also at $HOME
  - The communication channels in federations w.r.t. security are brought to life
  - Found some federations especially lacking in that regard (e.g. EGI-Checkin)
  - See how token transition from X.509 certificates changes this
- In case of a whole center being hacked, other players have other communication
  - Federal police communicates differently than befriended admins laboratory wide strategy on incidents
- Security effort increases:
  - On system level: Hardening of systems in the IDAF (`root` login only through intranet, MFA logins)
  - On network: Reduce connections IDAF ←→ internal network
  - At the entrance: Introduction of MFA planned for end of 2023 for all interactive logins to IDAF

Challenges: Sustainability

How to Make the Infrastructure more Sustainable

Constant improvement in DESY computing centre and infrastructure on DESY Campus w.r.t. energy efficiency

- Energy price becomes an additional incentive to be more efficient
- Hardware life cycle under close watch

Compute: Adapt hardware availability to power availability and/or user needs

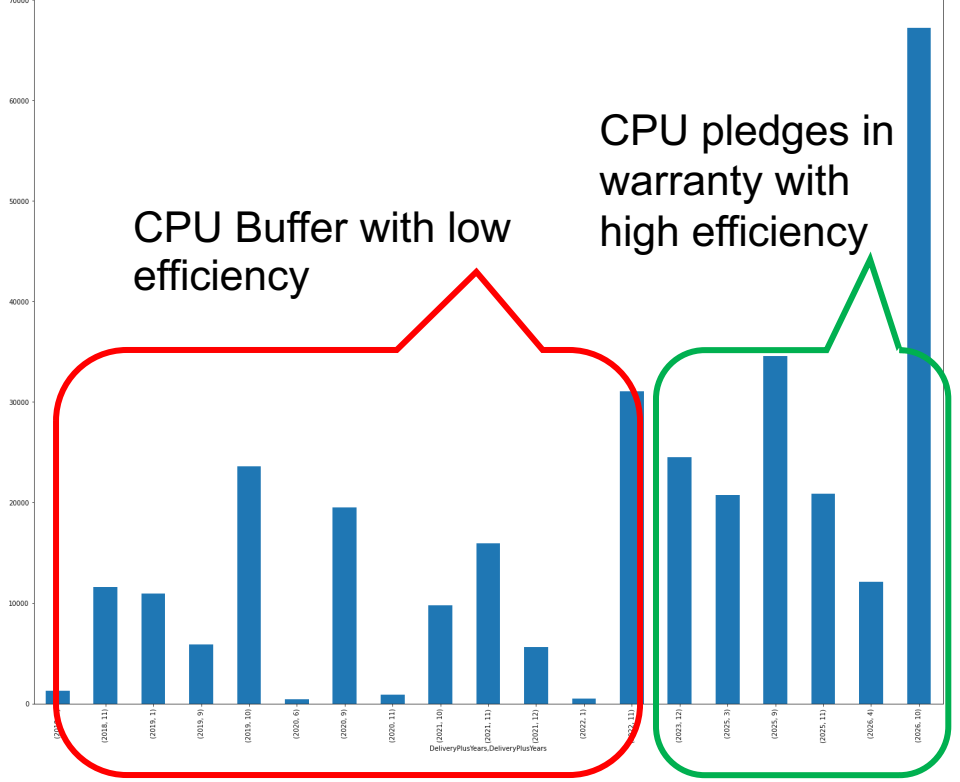
Storage: Unused data on tape → Tape?

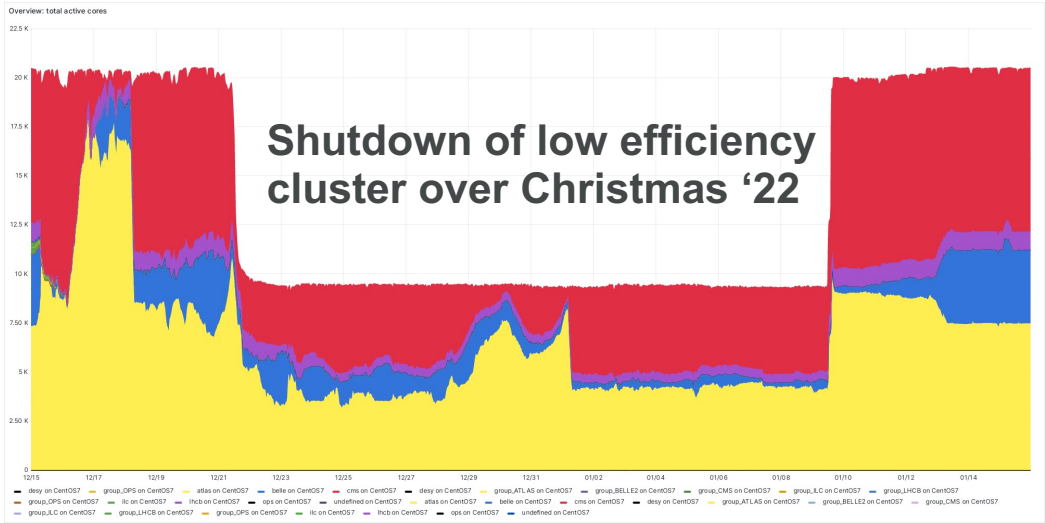
Raising **awareness** of users

Train users on most efficient use of IDAF

Train users on tooling and optimal algorithms

Interactivity and fast reaction come with inefficiencies:

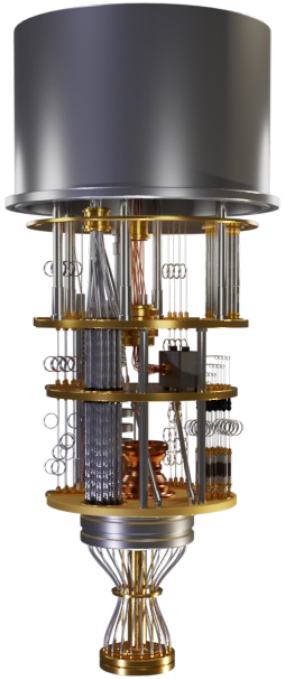




Provided by T. Hartmann

Challenges: Hardware evolution and Person Power

Difficulty Acquiring Hardware and Filling Open Positions



Hardware evolution

- Short-term: Supply chains have still not returned to full capacity after end of pandemic
- Short/mid-term: GPU: NVIDIA dominance is, scientific communities should be more open/flexible
  - Many interesting architectures / accelerator products out there vs. CUDA convenience
- Mid/long-term: Cloud providers driving technology
  - Started to offer tape for ‘ultra-cold storage’ ➔ profound effect on design of tape libraries not well suited to the IDAF
  - Some architectures already now only available in commercial clouds
- Mid/long-term: First quantum computer commercially available. Bring QC into the IDAF

Person Power

- More and more difficult to fill open positions and attract people for IDAF operation & development

Detailes Cobald/Tardis setup … e.g. NAF