IDAF — Current Status

Interdisciplinary` Data and Analysis Facility

Christian Voß & Yves Kemp
HEPiX Autumn 2023
Victoria, BC 20.10.2023

HELMHOLTZ

DESY.

Interdisciplenary Data and Analysis Facility (IDAF)

Origins and Overview

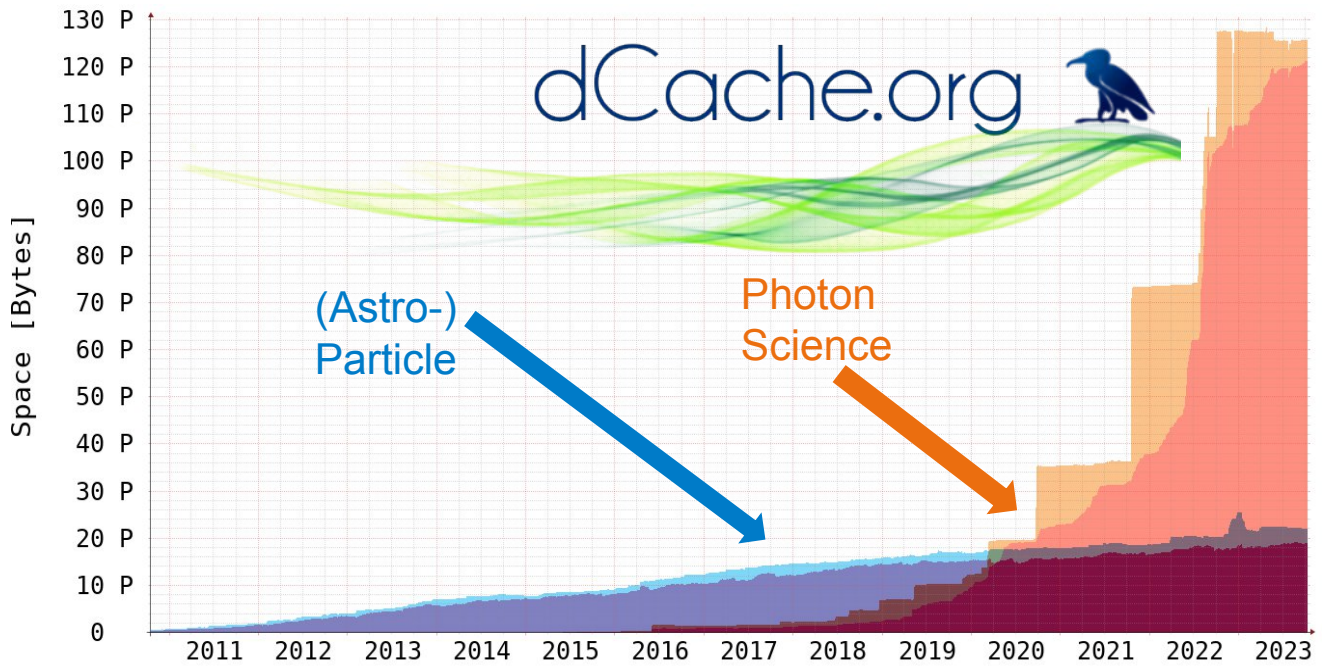
DESY historically centred on Particle Physics together with strong accelerator division:

- 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



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

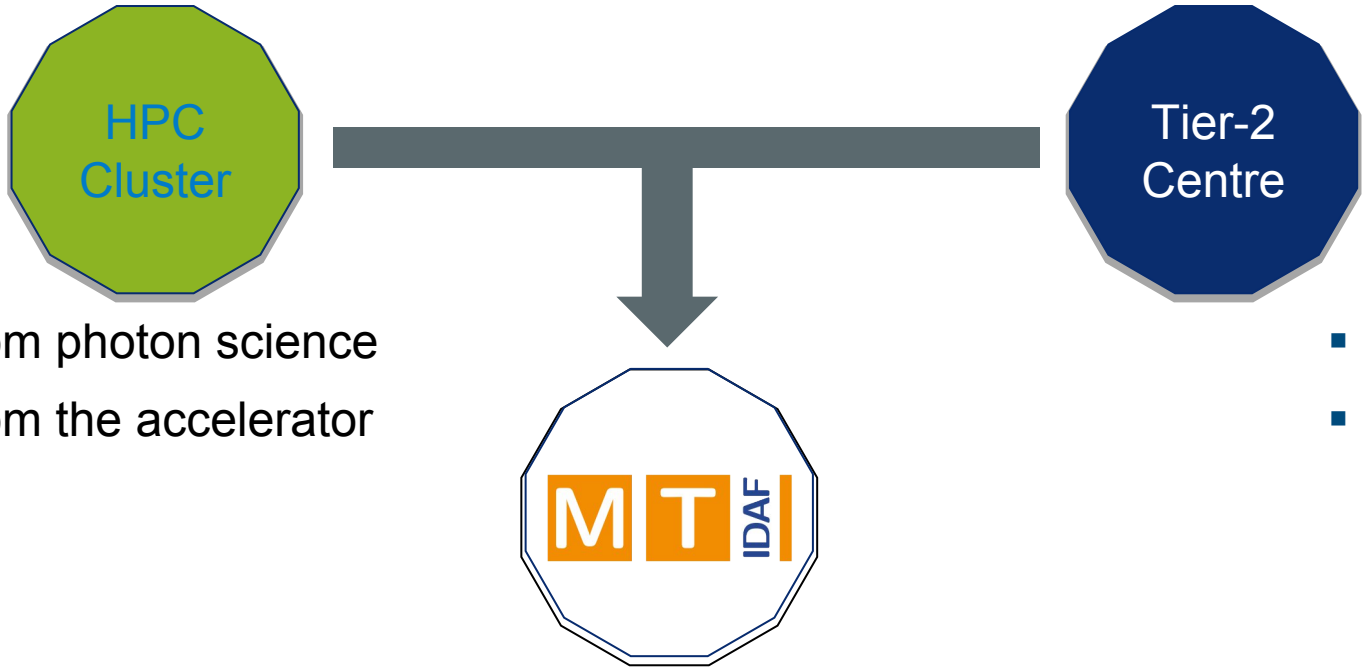
H.E.S.S.

cta cherenkov telescope array

Astro- Particle Physics

IDAF in a Nutshell

Merging Existing Infrastructures



- Users from photon science
- Users from the accelerator division

- Users from particle physics
- Grid and National Analysis Facility clusters

Single infrastructure open for all scientists in Matter

- Currently mostly administrative and logical merger
- Iron out ideas for a full on merger (several pit falls: Namespace for data access)

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

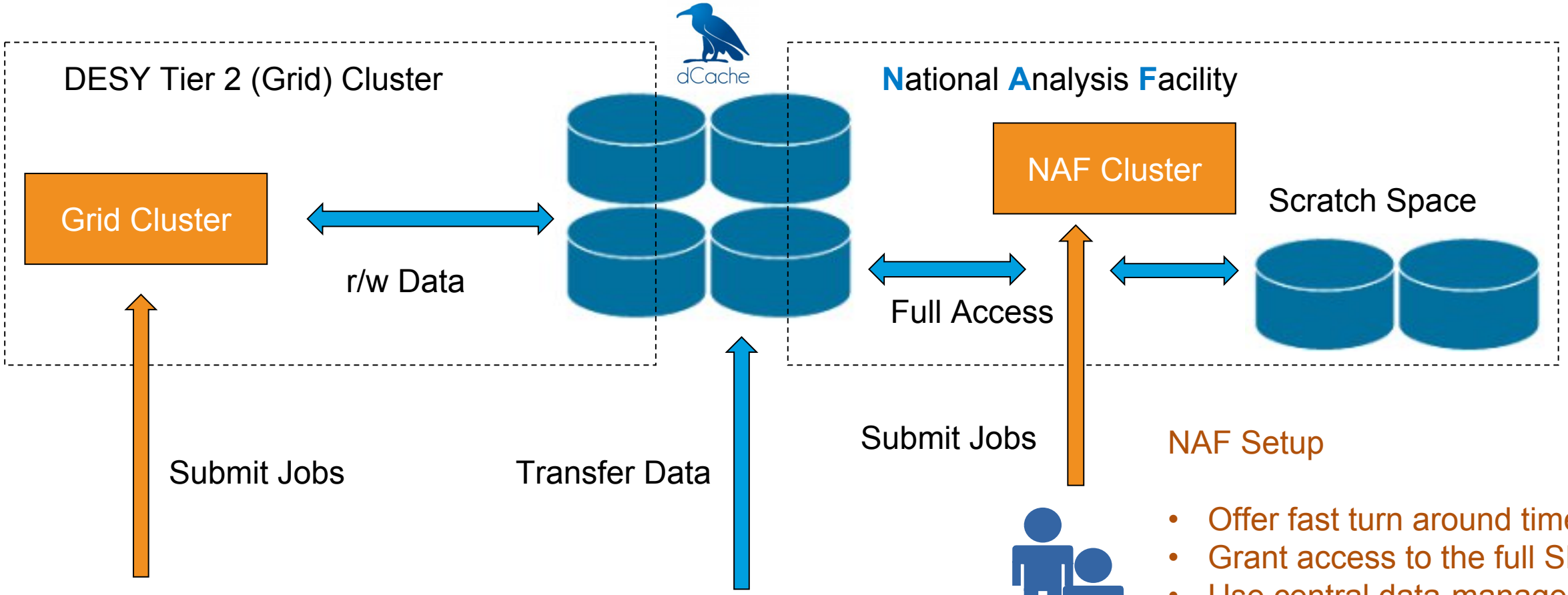
Services on the roadmap

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

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



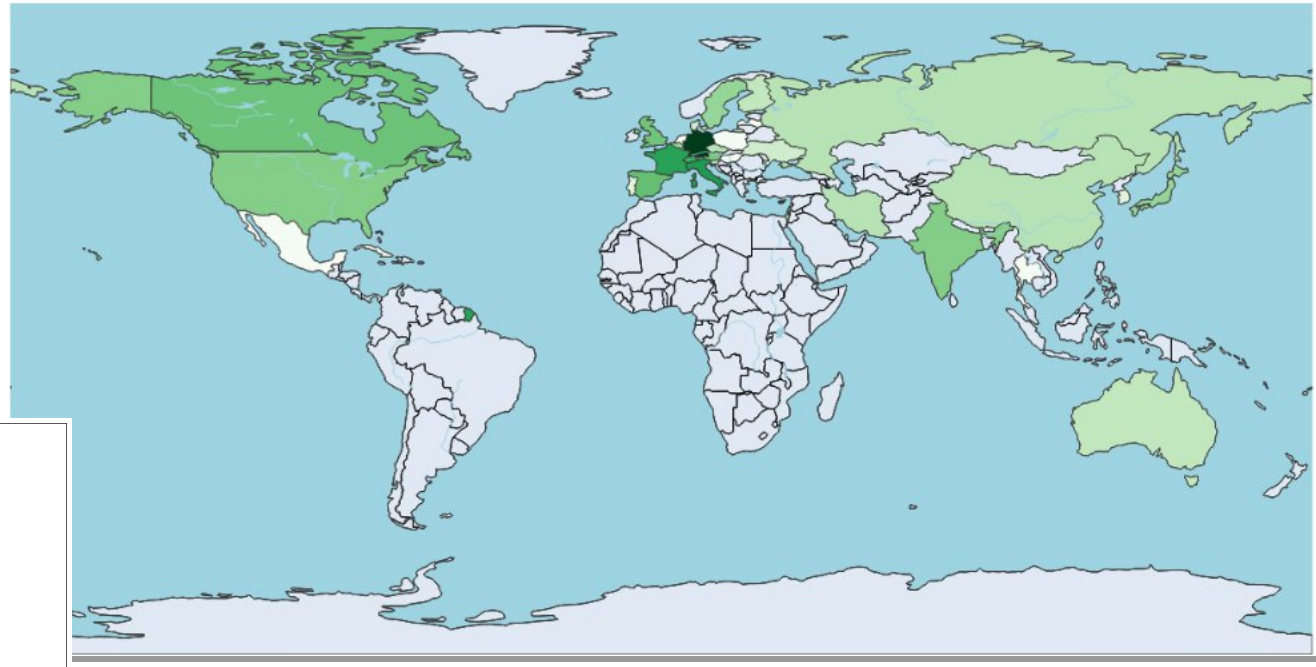
NAF Setup

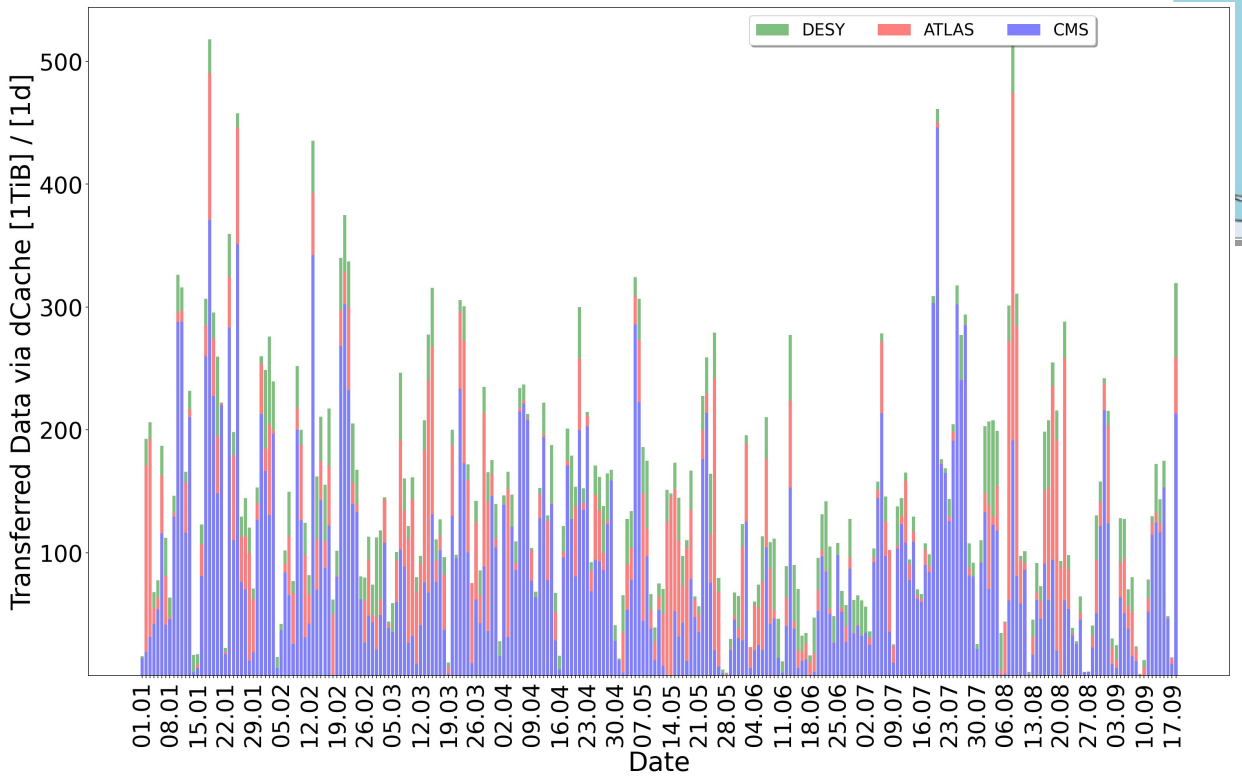
- Offer fast turn around times
- Grant access to the full SEs
- Use central data-management of experiments to store data at DESY

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

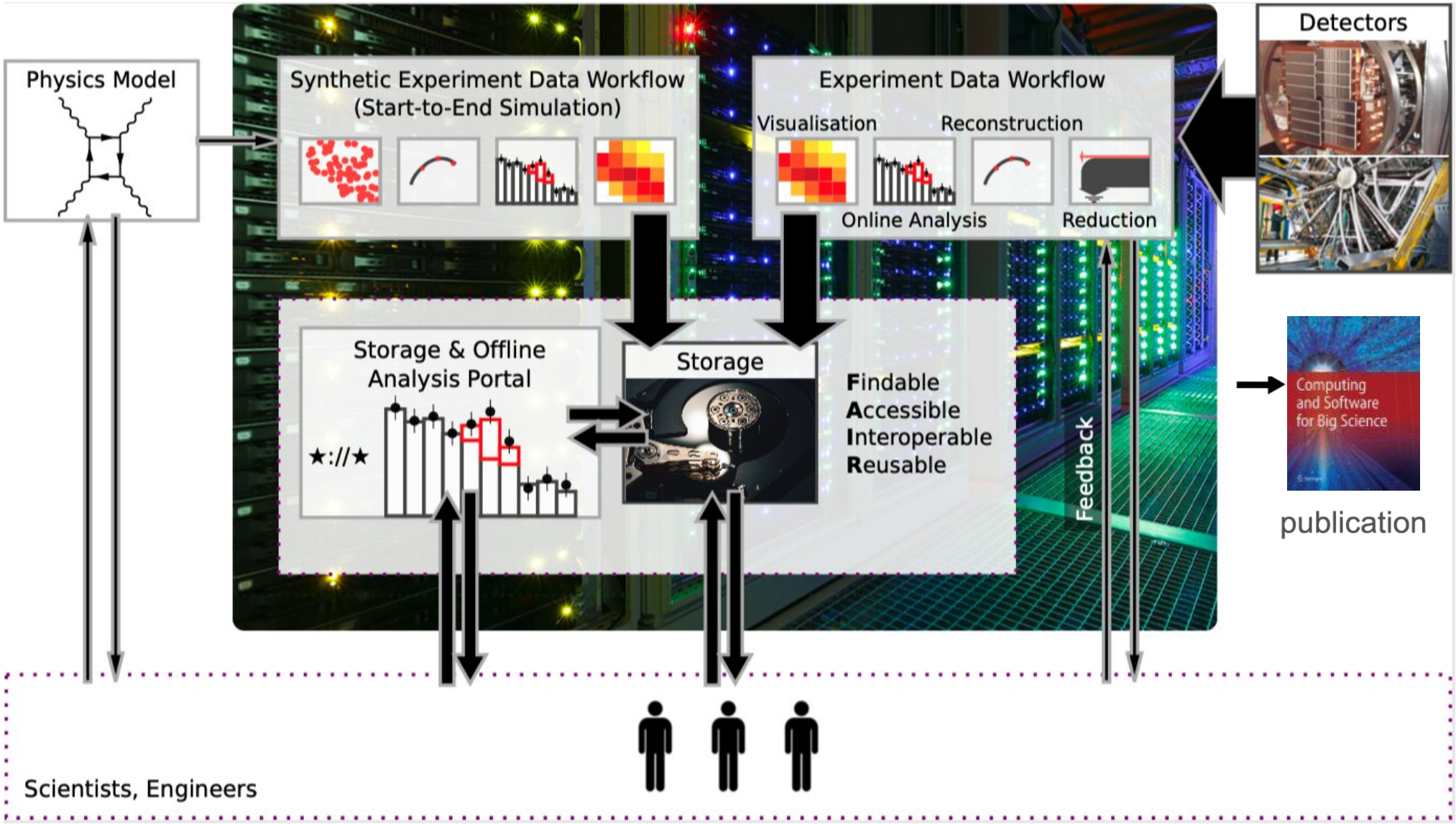




- Data access inside NAF (only dCache shown)
- Additional storage space for NAF (linked to experiment frameworkd)
- CMS as largest contributor
- Jobs do almost exclusively POSIX

On-Site: Particle Physics, Accelerators, Photon Science

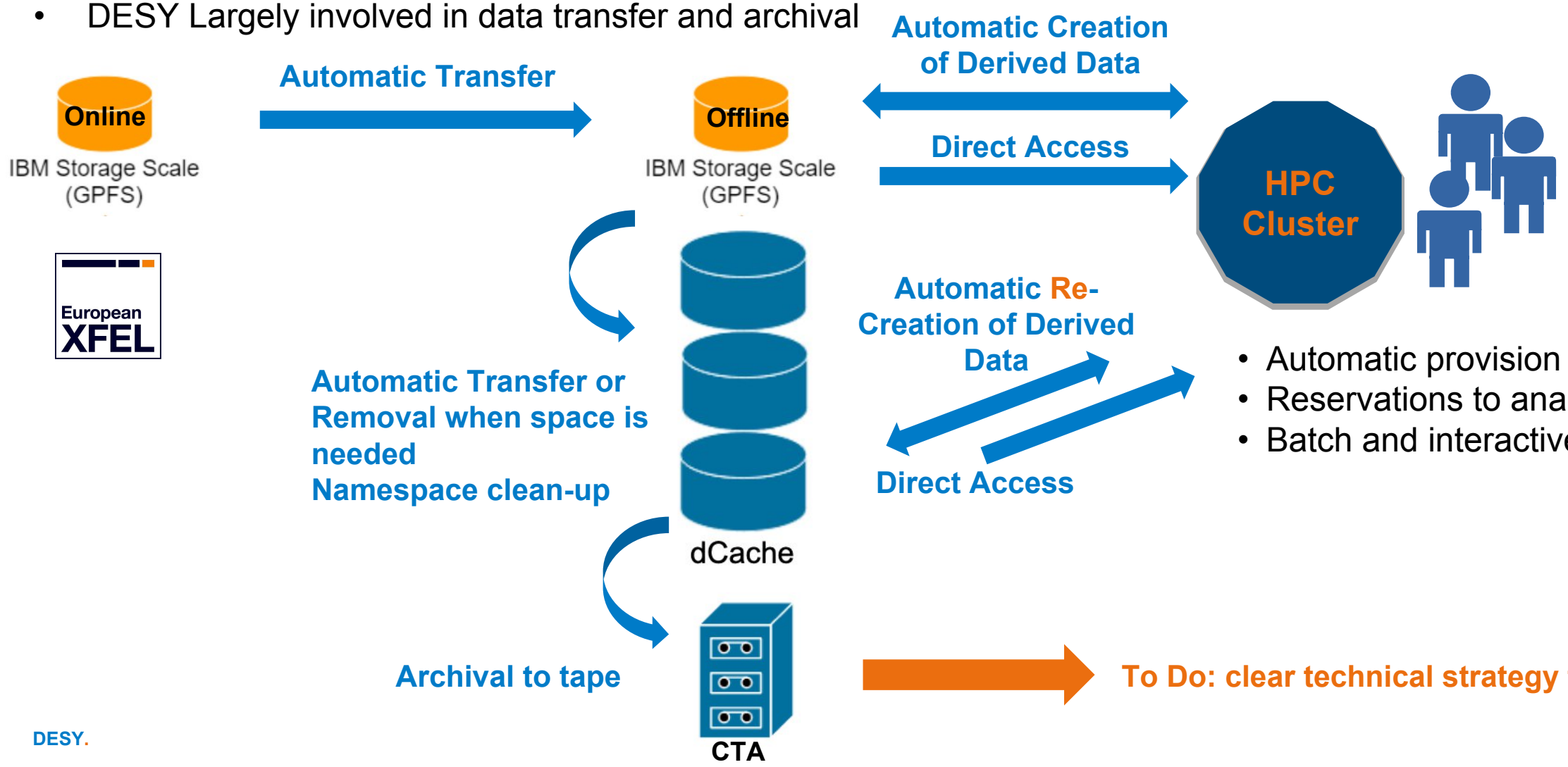
Enable the Full Analysis/Data Lifecycle: From Simulation to Publication and Archival



On-Site Example

User Proposals for European XFEL

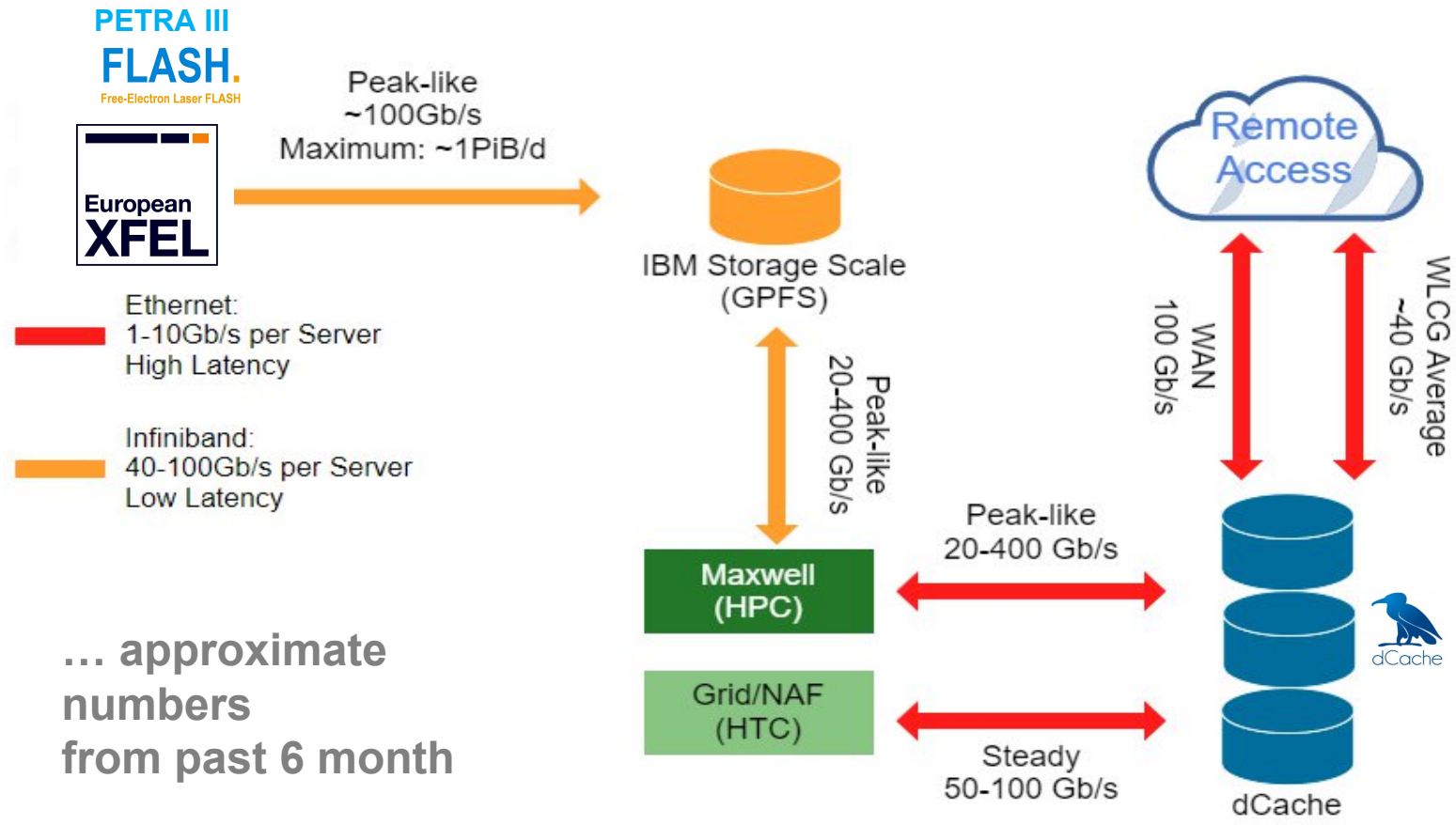
- Developed largely by our colleagues at the European XFEL ➜ Analysis is centered on IDAF
- DESY Largely involved in data transfer and archival



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

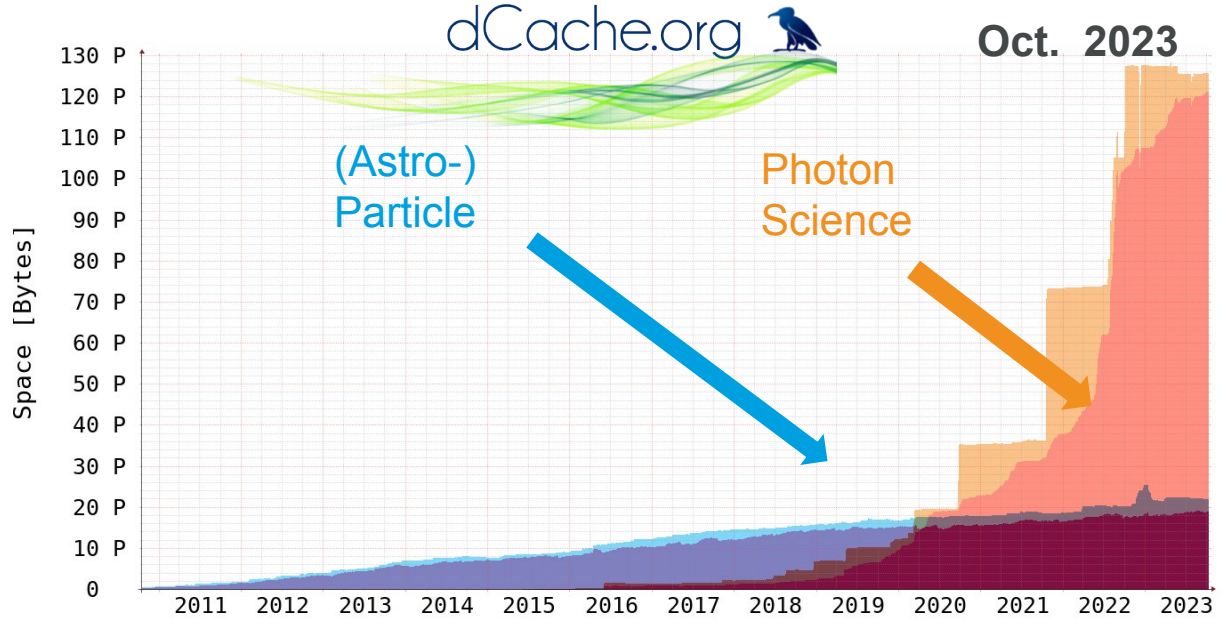


… approximate numbers from past 6 month

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
- **Data reduction** essential! Integrate data reduction workflows

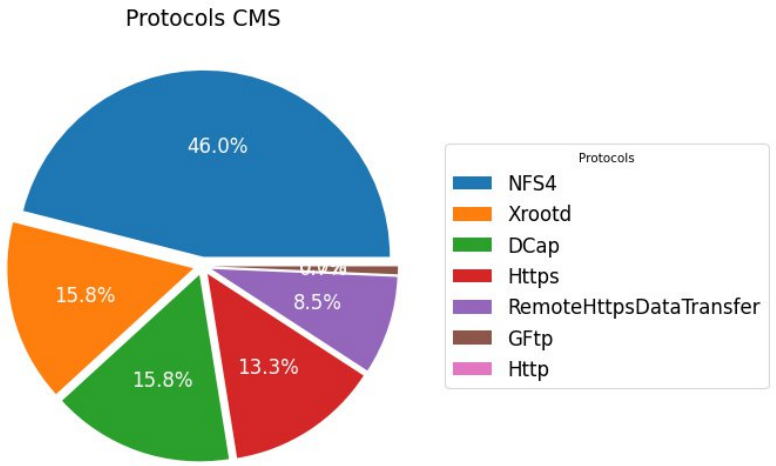


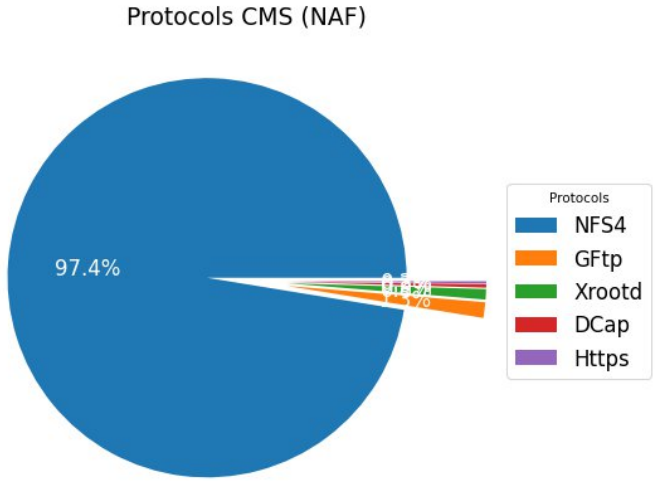
- **Observe scaling issues for the IDAF**
- Number of dCache pools causes issues when rebalancing after introducing new pools
- Pool nodes start pile up in the computing centre: **start experience limits to rack space**

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**

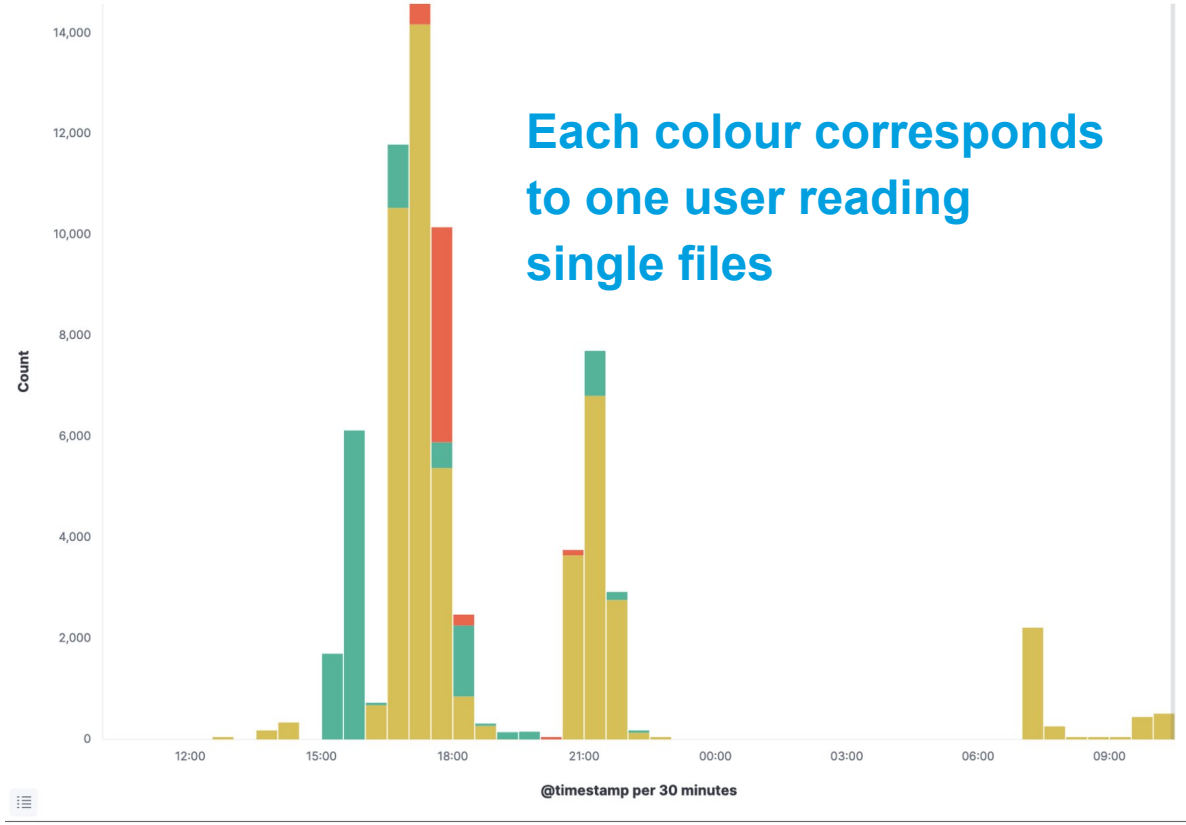


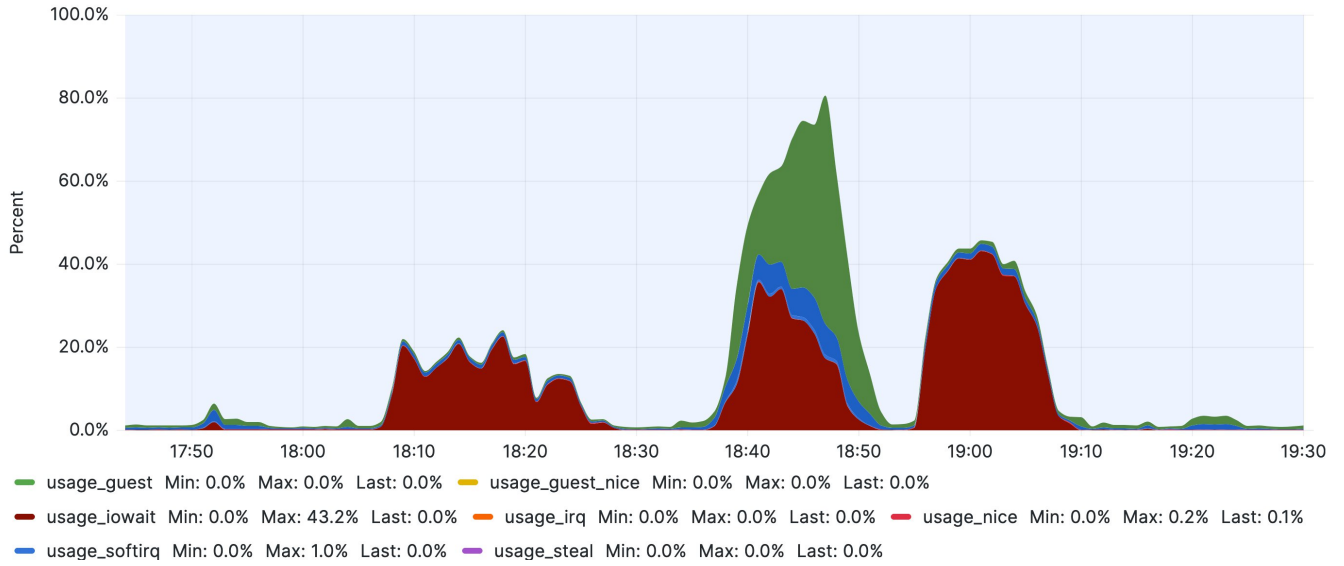


Challenges: Using HTC as HPC

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?

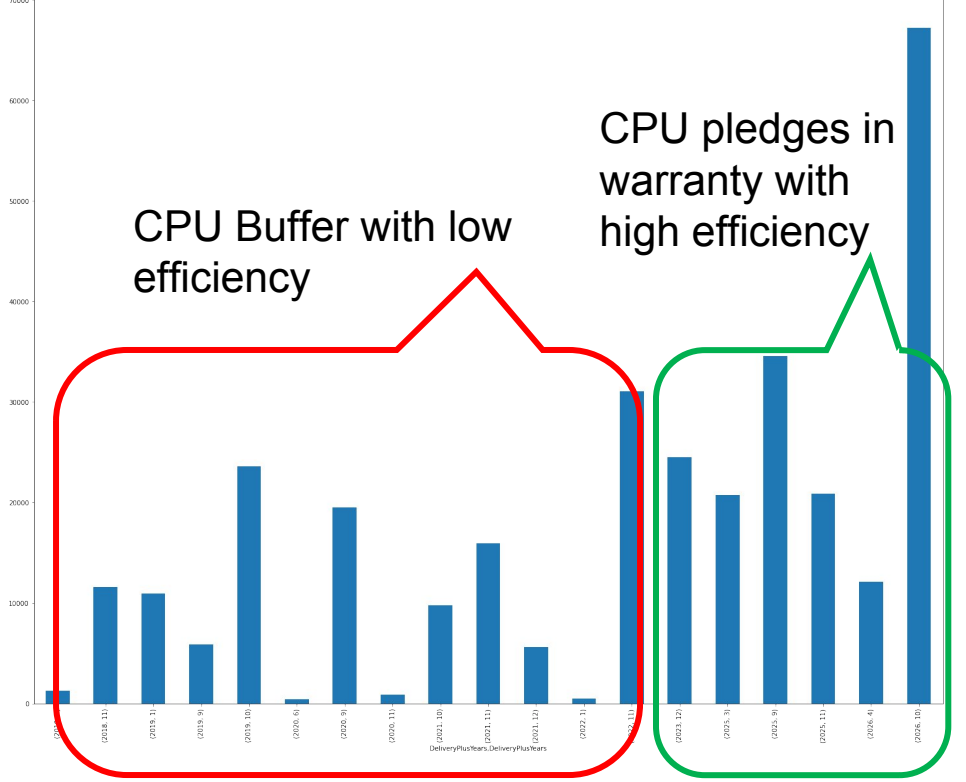
Raising **awareness** of users

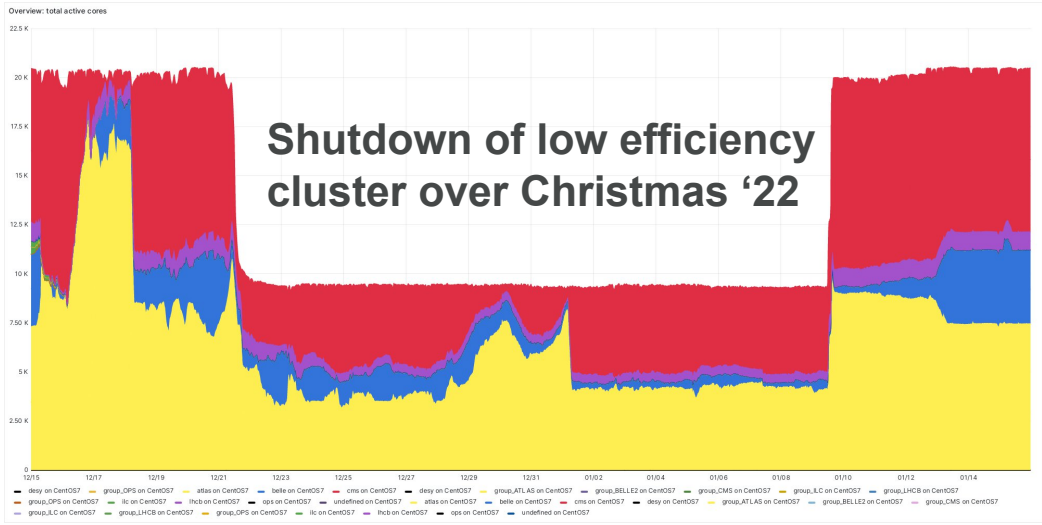
Train users on most efficient use of IDAF

Train users on tooling and optimal algorithms

Interactivity and fast reaction come with inefficiencies:

- Re-evaluate how much is needed
- Eventually tax users
- Optimize scheduling and availability

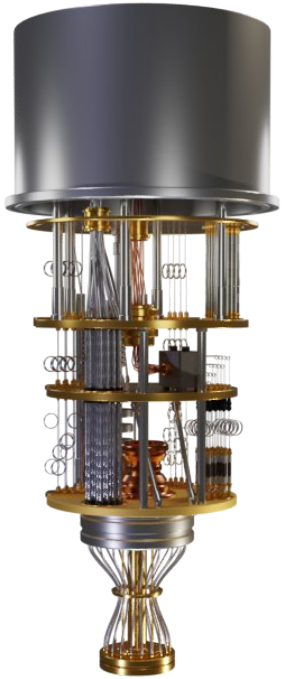




Provided by T. Hartmann

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
- Danger that certain key services lack fall-back admins in case of sickness/holidays

Summary

From a WLCG Tier-2 Centre to an Interdisciplinary Facility

- Currently in Progress of consolidating the compute infrastructure of our communities to a single facility
- We do this while data rates are ever increasing
- How do we deal with the reliance on POSIX
- POSIX probably complicates future development to more cloud-like workflows
- Sustainability and hardware selection/person power as new challenges
- Outstanding: true unification ➜ transparent access to data from HTC/HPC and resource assignment based on needs rather than community
- On the Horizon: **PETRA IV** with data rates similar or surpassing XFEL