# IDAF — Strategy for PoF V

## Strategy Meeting MT for PoF V

Christian Voß (presented by Yves Kemp)
Kassel Wilhelmshöhe, 28th June 2023

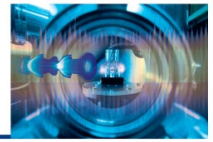
HELMHOLTZ

DESY.

# Start with Review of 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

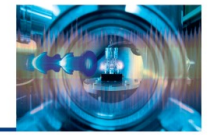
- Association with MT
  - MT is interdisciplinary ➡ IDAF is moved from MU to MT
  - Current setup planned at DESY (very broad matter community, experience with Tier-2)

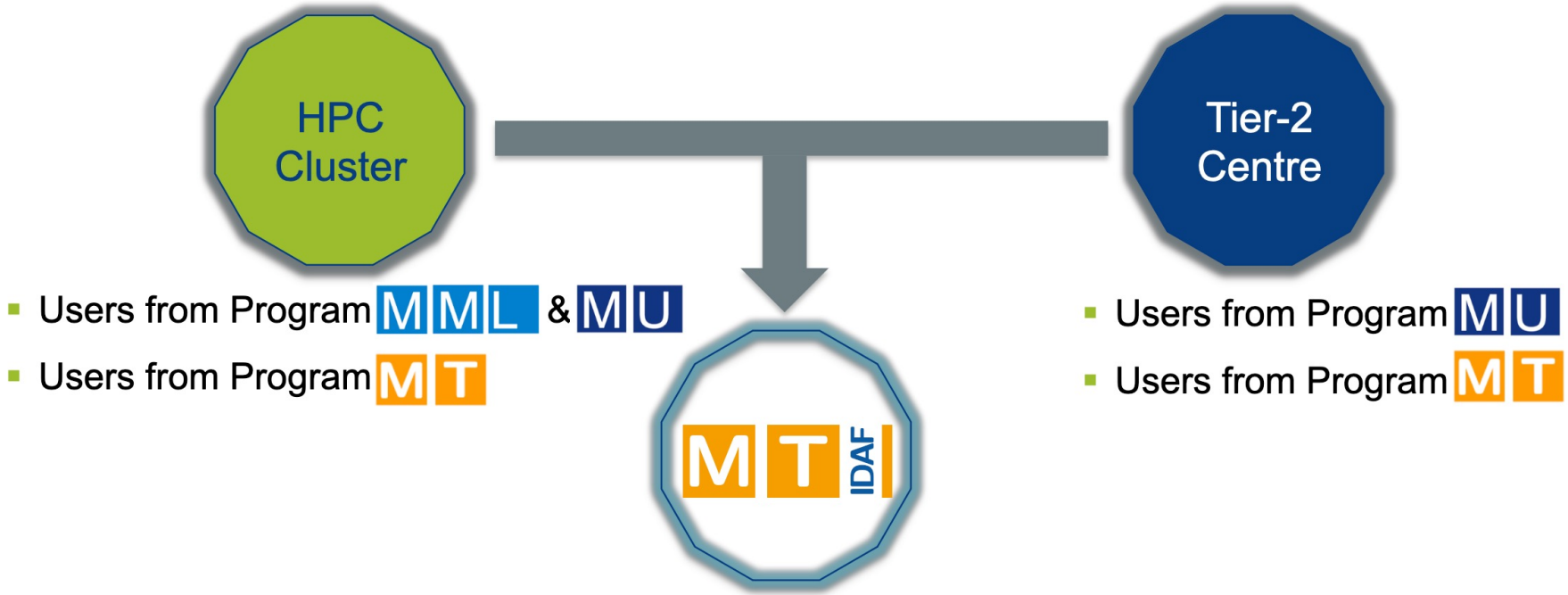
# Start with Review of PoF IV Proposal

## Plans for the IDAF

### Building the IDAF:

Merging High Performance Computing (HPC) and Tier-2 Clusters



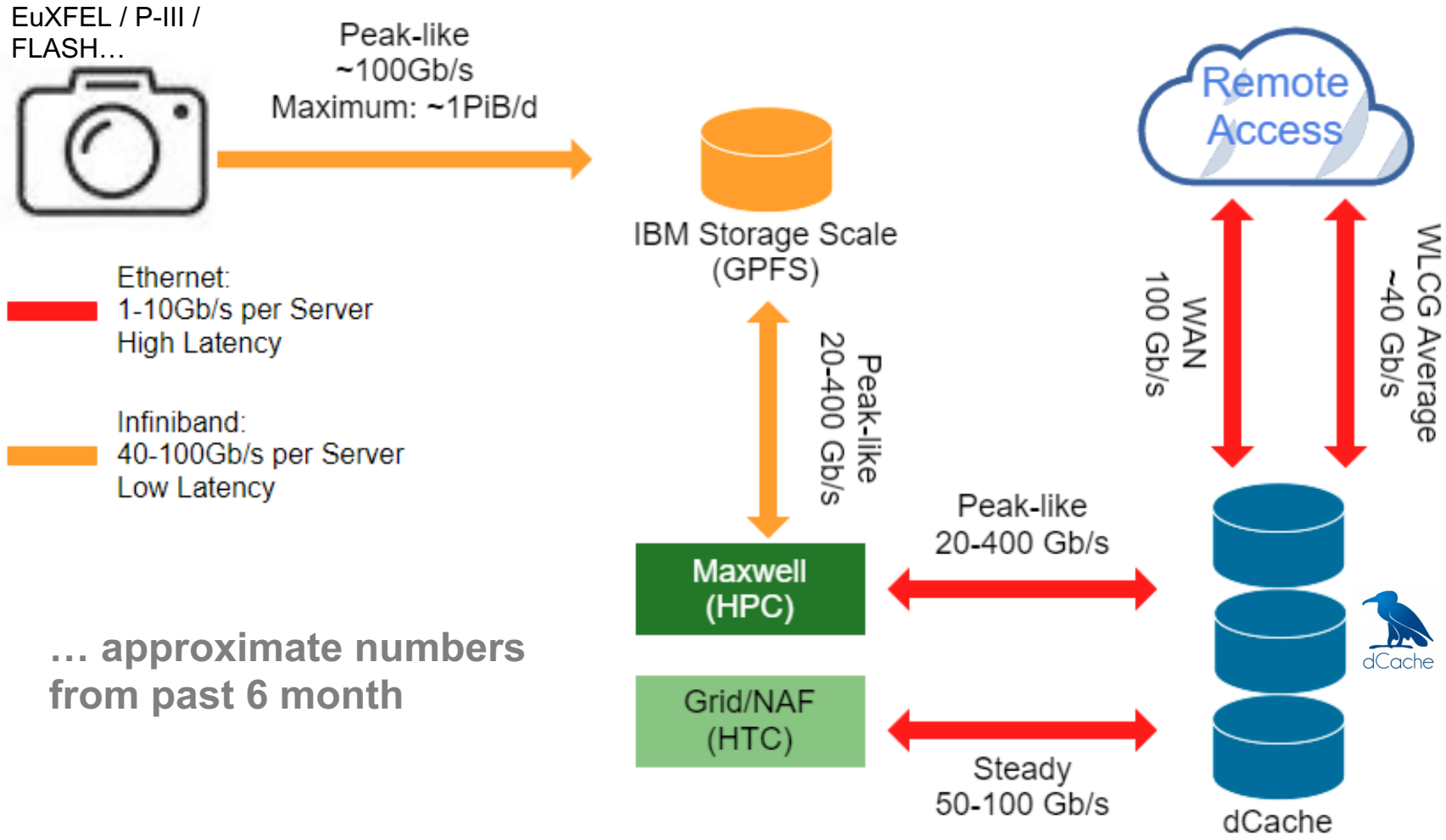



Single infrastructure open for all scientists in Matter

# Paradigm: Scientific Analyses are Data Driven

## Strategy: Keep the Paradigm that Made the Tier-2 Successful

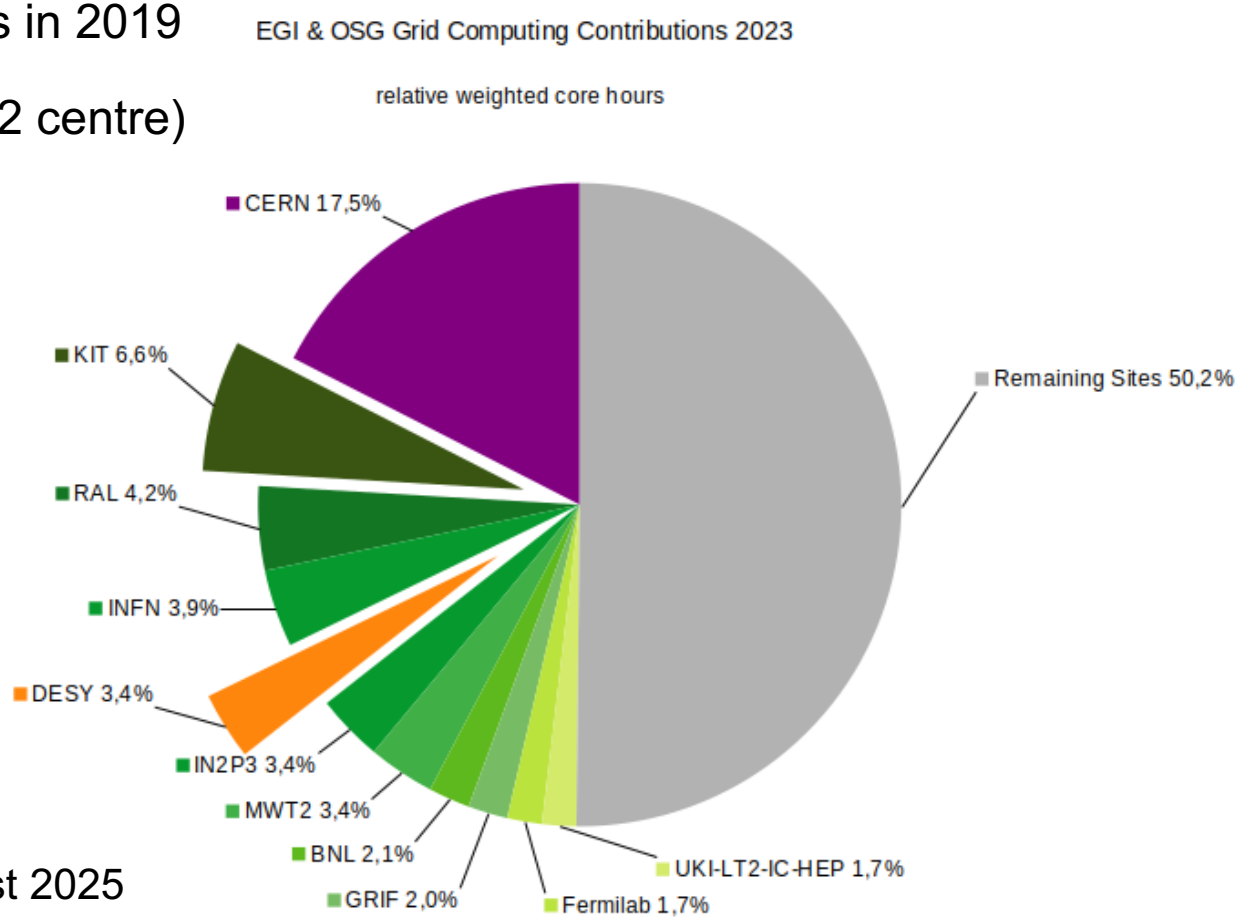
- Example: Traffic pattern in IDAF, approximate numbers from 2023H1



# Continued International & new National Commitments

## IDAF Inherited Previous MU Commitments for the (Astro-)Particle Physics Experiments from Tier-2

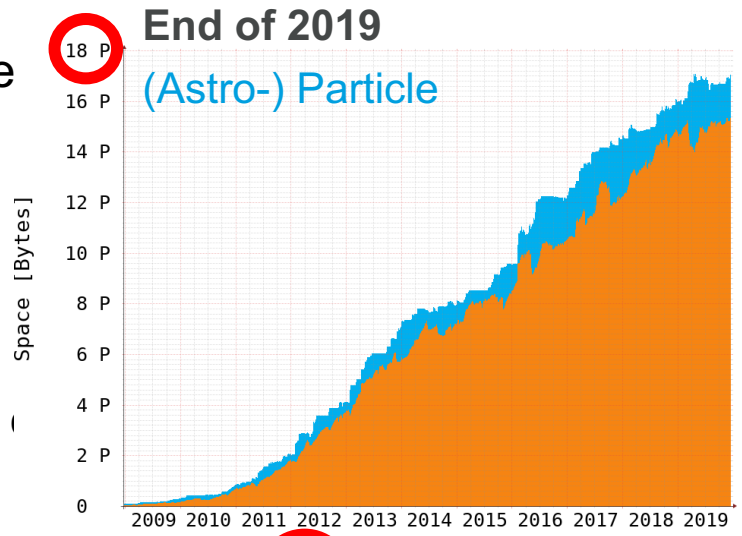
- IDAF contributed around 4% to (Astro-)Particle Physics in 2019
- 2022: share ~3.4% (IDAF still largest contributing Tier-2 centre)
- Expanded responsibilities
  - Raw Data Centre (Tier-1 equivalent) for Belle II
  - **Offer tape storage to LHC experiments (compensate affected Russian Tier-1 sites)**
- **Take over storage share from German universities**
  - KET: University Tier-2 centres to be discontinued
  - CPU shares to be taken over by some NHR sites
  - Storage to be split among Helmholtz Sites (KIT/DESY)
  - Investment in part covered by the BMBF (Verbundantrag)
  - Some additional investment expected in kind by DESY past 2025
  - New workflows expected. Will need research, and support. Close eye on network, might need expansion

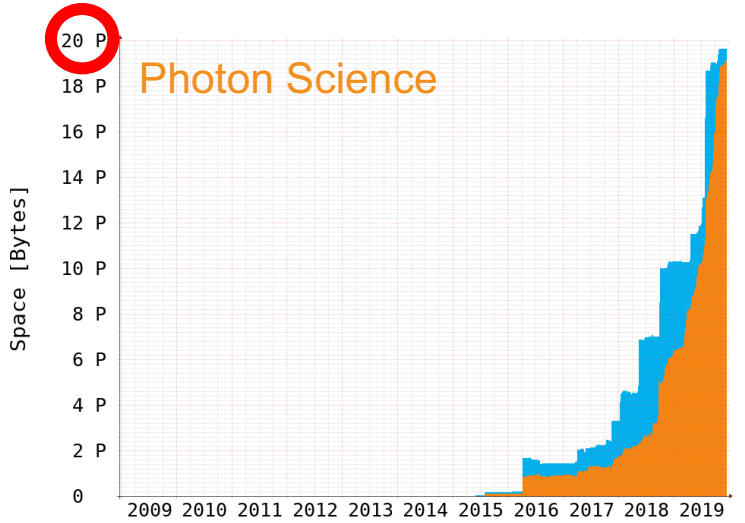


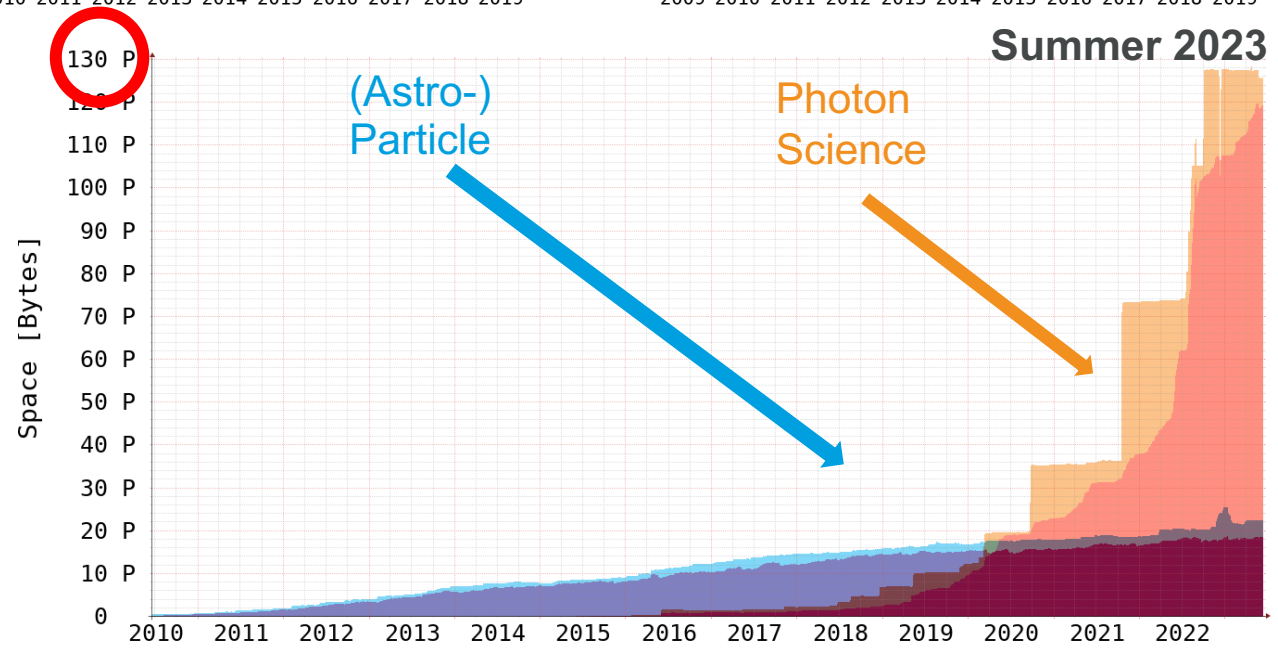
# Challenges: Data Deluge in Photon Science

## Photon Science and Especially European XFEL Continued to Grow Exponentially

- Date stored since beginning of PoF IV more than doubled
- **Accelerator division starts to contribute**
- HPC cluster storage similarly increased
- Capacity growth slow down/halt during end 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** on the horizon?
- **Observe scaling issues for the IDAF**



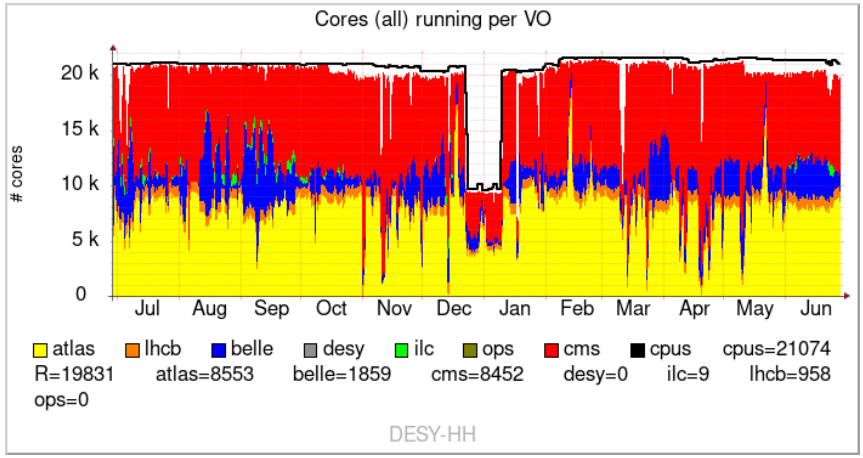



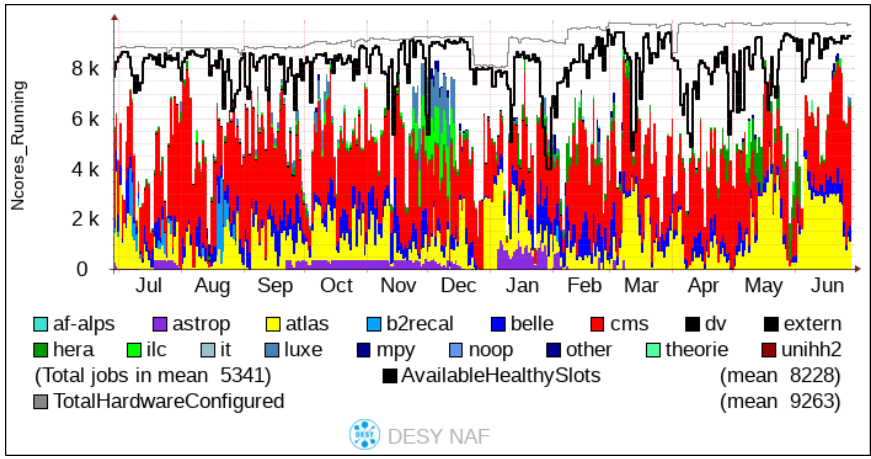



# Ressource and usage status IDAF

- **High Performance Cluster: Maxwell**
  - ~900 nodes (inkl. ~250 GPU), ~50k Cores. 2700 users (~1000 active in past 3 month)
  - Storage: GPFS, dCache, (BeeGFS). InfiniBand, SLURM scheduler
- **High Throughput, Production: Grid**
  - 400 nodes, 20.000 cores
  - Storage: dCache, CVMFS. Ethernet, HTCondor Scheduler – Integration in WLCG/Experiment frameworks.
- **High Throughput, Interactive: NAF**
  - 350 nodes, 8.000 cores.
  - Storage: dCache, DUST (GPFS/NFS), CVMFS, AFS. Ethernet, HTCondor Scheduler.

→ **Planning for consolidation, unification**



# Challenges: Accessing Data

## Users Prefer to Use POSIX — IDAF Needs to Adapt to that Fact

- Continued trend to access data ‘directly’

```
def read_frame_from_file(frame_id: int, data_file: str):
    start_time = time.time()
    with h5py.File(data_file, 'r') as h5in:
        tmp_arr = h5in['/PATH:xtdf/image/data'][frame_id]
    read_time = time.time() - start_time
    return read_time
```

European XFEL

- Usually only option for MML and MT
- Trend includes MU despite remote read capabilities
- **Poses the challenge of having uniform name-space across the IDAF**

HPC

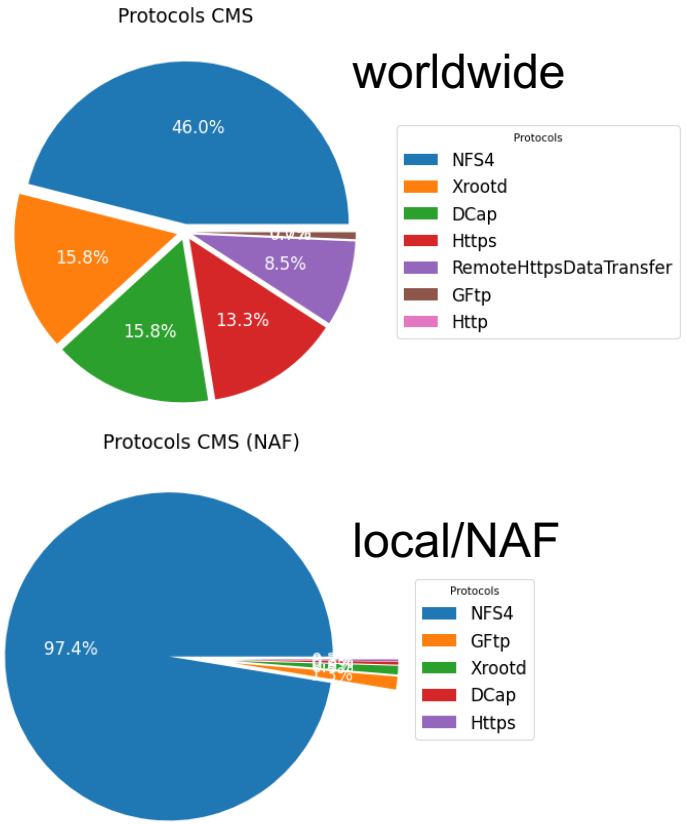
**[vossc@max-display008] ~ $ md5sum /gpfs/dust/belle2/user/vossc/stage-rest-api.out**
**0108f37dbbb38103bba6d836f356d7b7 /gpfs/dust/belle2/user/vossc/stage-rest-api.out**

HTC

**[vossc@naf-belle12] ~ $ md5sum /nfs/dust/belle2/user/vossc/stage-rest-api.out**
**0108f37dbbb38103bba6d836f356d7b7 /nfs/dust/belle2/user/vossc/stage-rest-api.out**

- I would need to change my analysis depending on the cluster I’m on

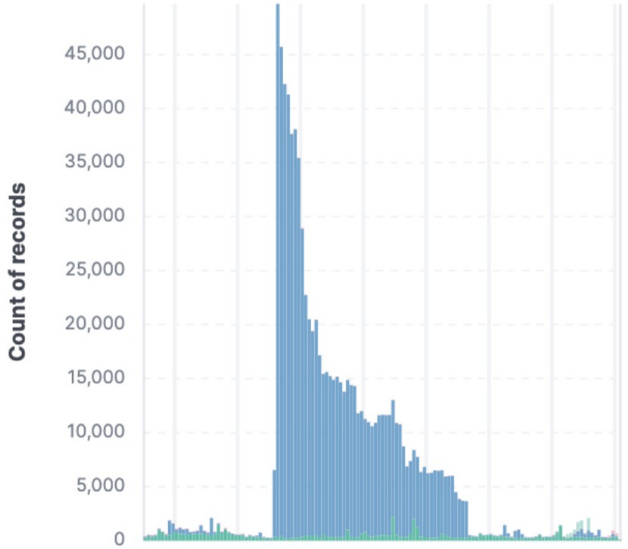
### Data Access CMS May 2023

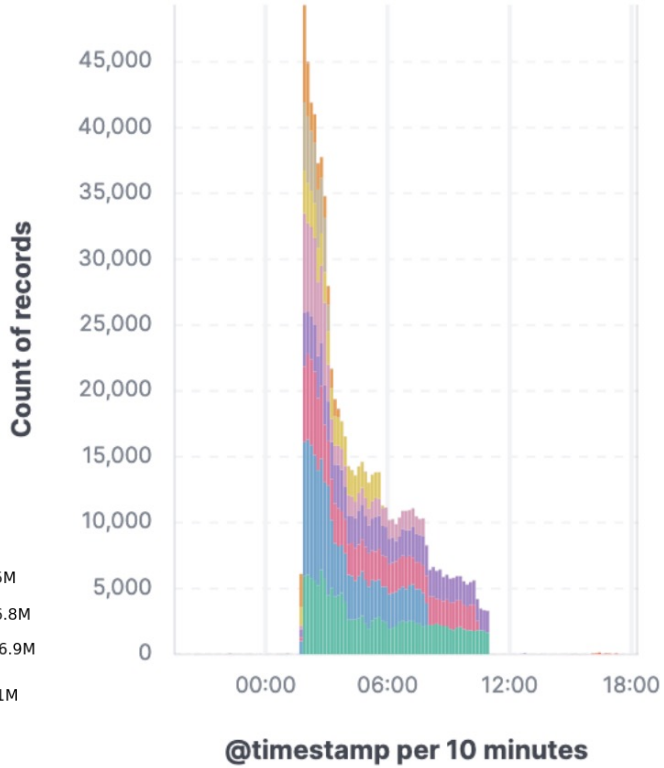


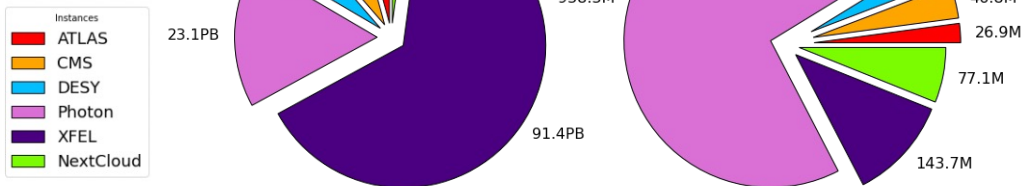
# Challenge: Improved Monitoring and Analytics

## Managing and Understanding the Change User Access Patterns

- Increasing capacity found to manageable
  → read/write patterns found to be more challenging
- Departure from classic C/C++ or FORTRAN driven batch analysis
- Ease-of-Use of Python leads to higher memory footprint and excessive, repetitive data access (open files to read <1MiB)
- Increased WAN/Tape access will escalate this further
- Profit from research in :
  - Self adapting systems (e.g. Smart file replication) MTDMA
  - Improved I/O pattern, e.g. through portals (Coffea-Casa)
- Profit from research in MTDTS / MTDMA
  - Reasonable file sizes/numbers
  - Streaming/Online Analysis



# Challenges: Sustainability

## How to Make the Infrastructure more Sustainable

**Constant improvement** on PUE in DESY CC and infrastructure on DESY Campus … ongoing since years

- Hardware life cycle under close watch

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

**Storage:** Unused data on tape → Tape?

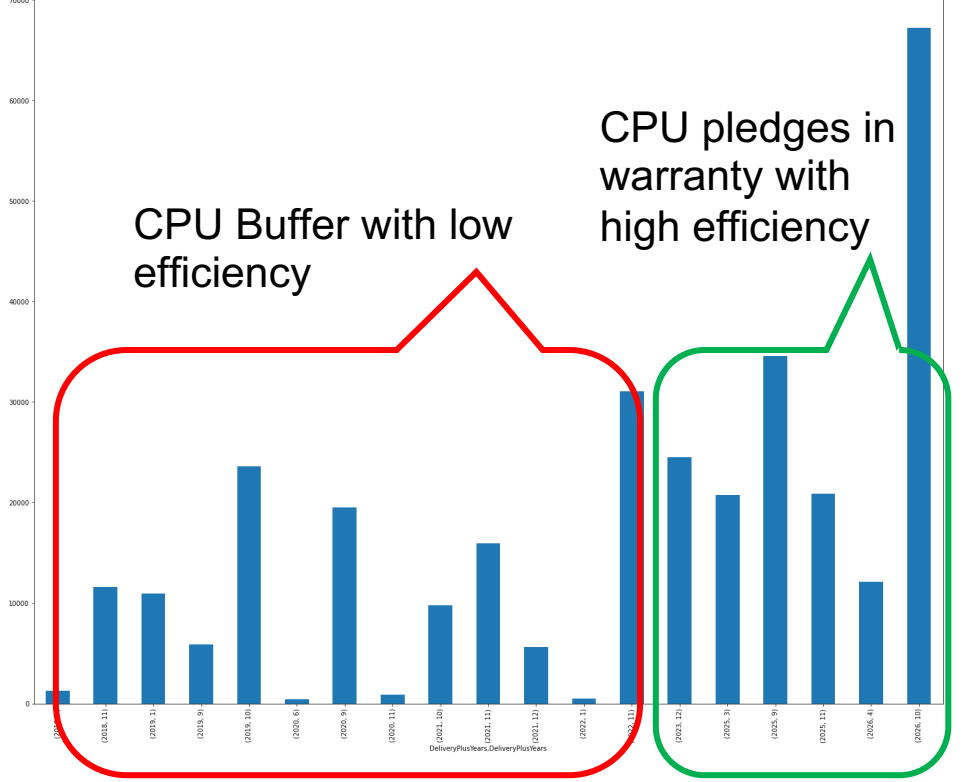
Raising **awareness** of users

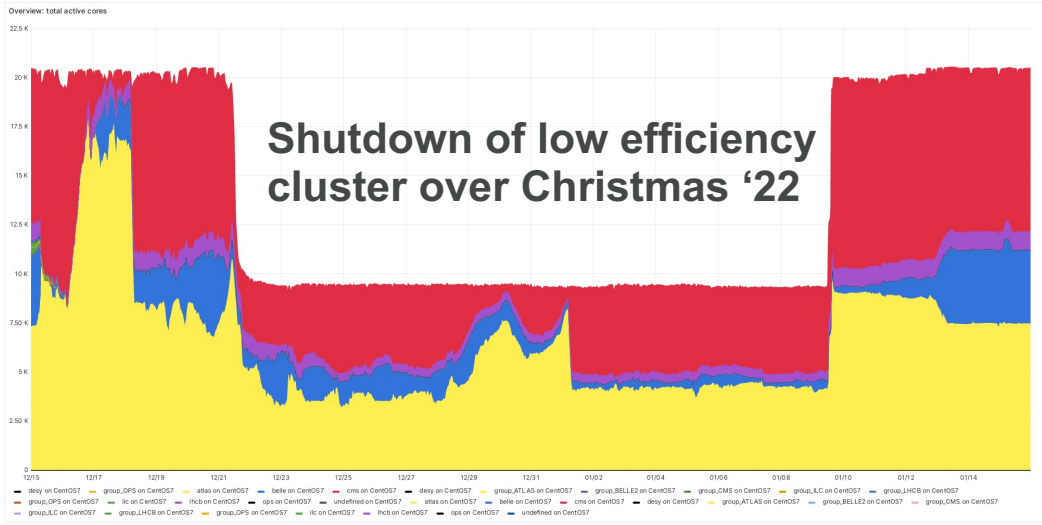
**Train users** on most efficient use of IDAF

**Train users** on tooling and optimal algorithms

**Interactivity and fast reaction** come with inefficiencies:

- Re-evalute how much is needed
- Eventually tax users
- Work on scheduling and availability

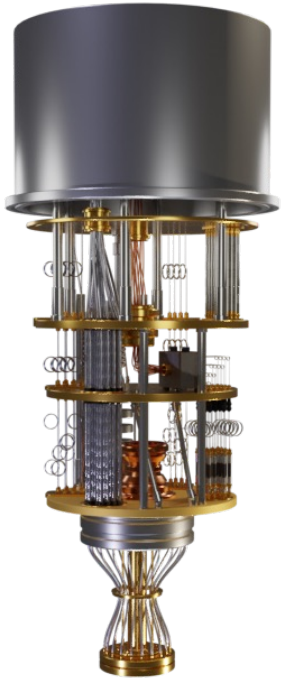



Provided by T. Hartmann

# 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 not healthy, need combined effort to overcome
  - many interesting architectures / accelerator products out there, we should be more open and flexible
- Mid/long-term: Cloud providers driving technology … and making it private
  - 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. QC will become an additional IDAF platform

### Person Power

- More and more difficult to fill open positions
- ML/AI can be filled eventually
- Regular IT positions often cannot be filled and get cancelled

# Thank you, any Questions