~~NAF~~ IDAF at DESY

Interdisciplinary Data and Analysis Facility

Yves Kemp et al., DESY IT

Analysis Facilities Workshop
Munich 3.6.2024

HELMHOLTZ RESEARCH FOR GRAND CHALLENGES

DESY.

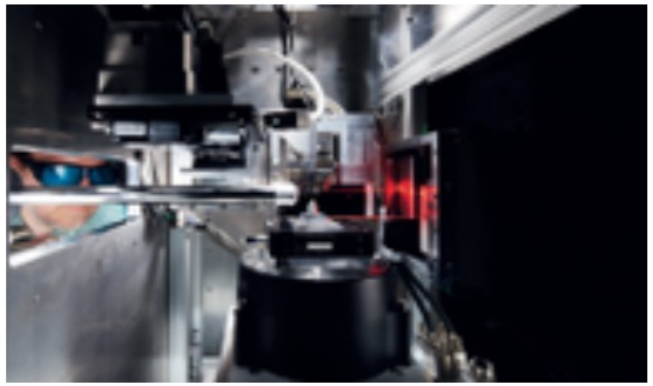
DESY research divisons ... In a nutshell (those in Hamburg)

Accelerators »

Running / Operating:
- Petra III, FLASH, XFEL, ...

Planning:
- Petra IV

General Accelerator R&D



Photon science »

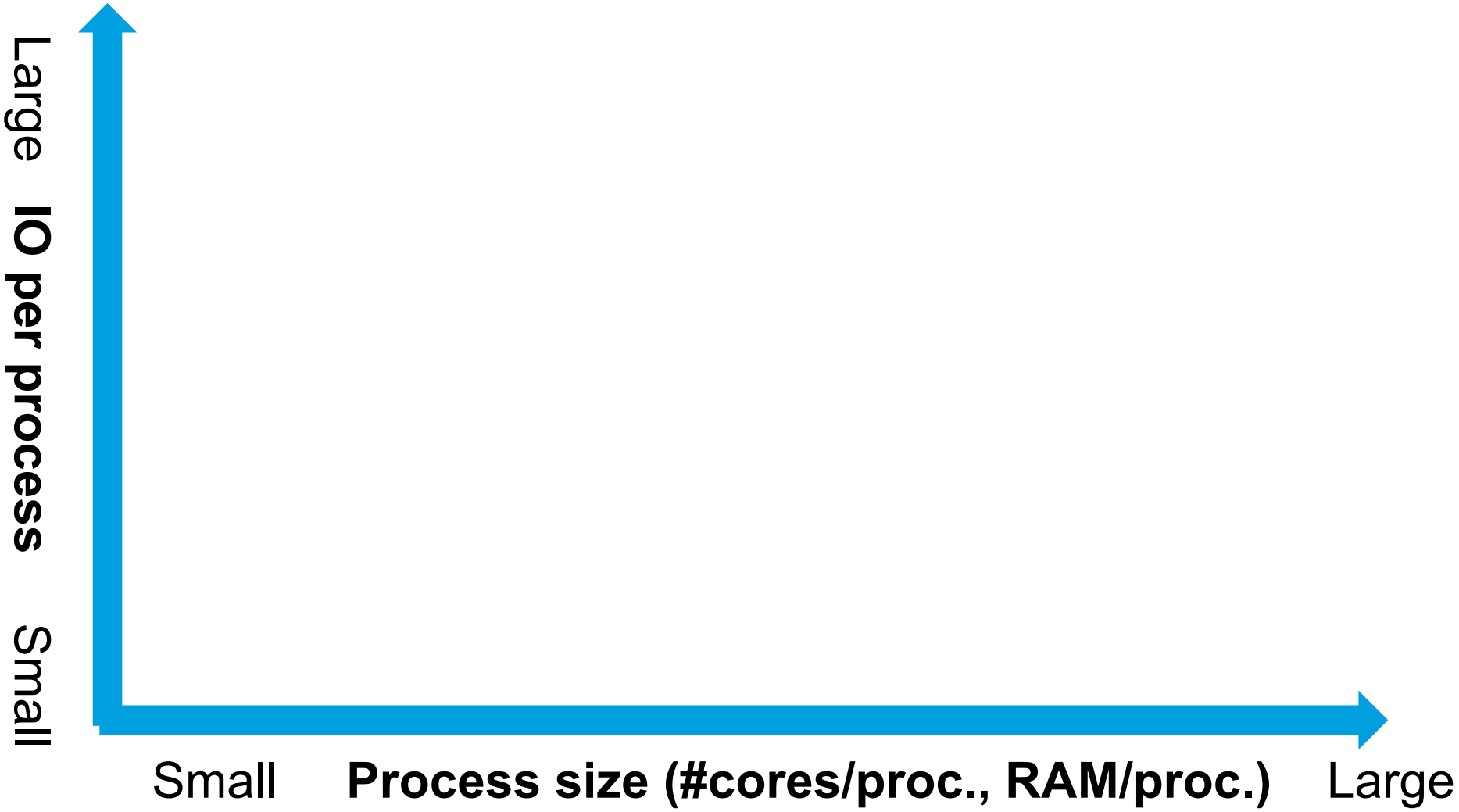
Petra III, FLASH, EXFEL, CFEL, CSSB, EMBL, HZG

Particle physics »

- LHC, HL-LHC
- Belle II
- ILC, ALPS, ...
- Theory division

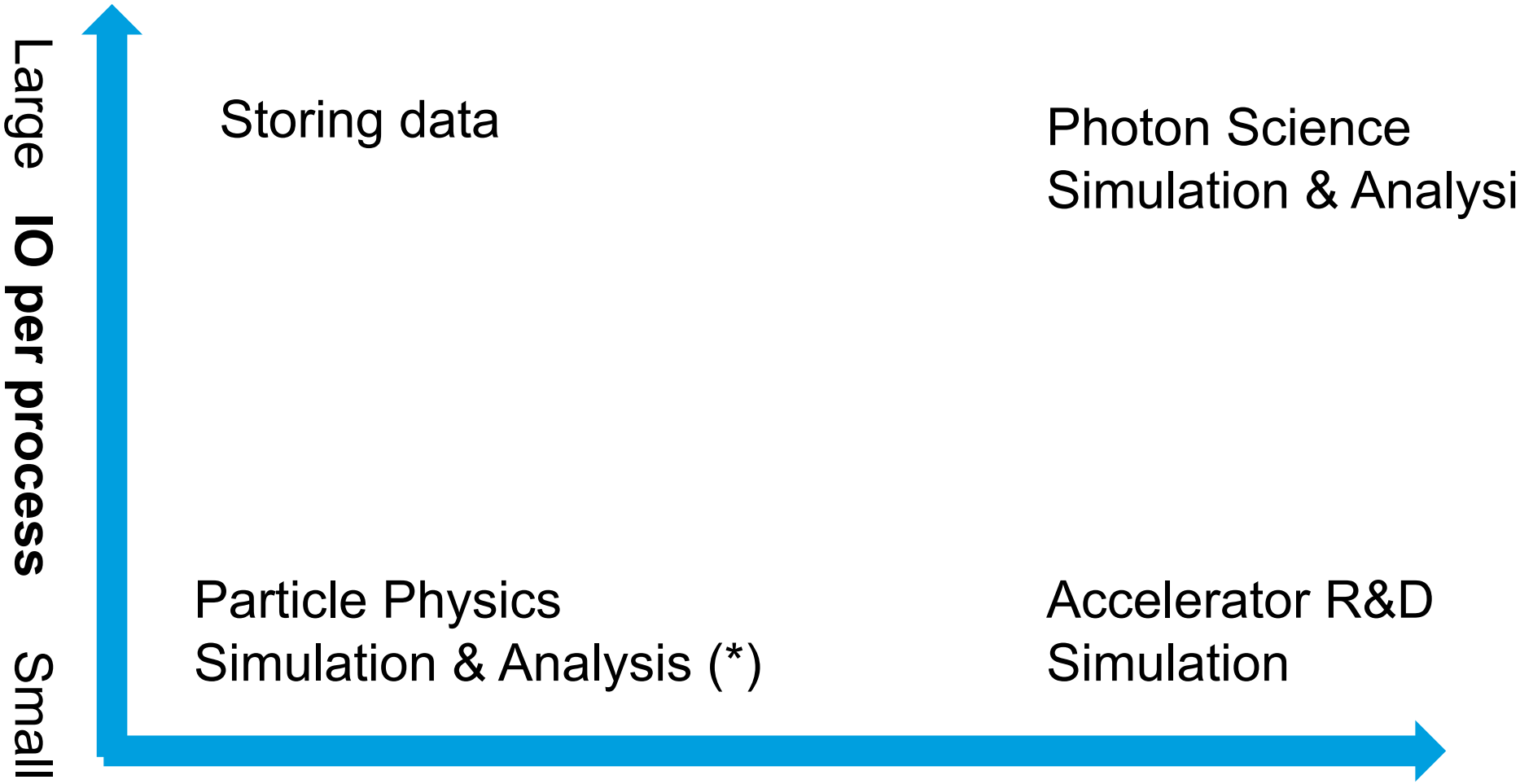
Computational requirements: Job size vs IO needs

Very very coarse



Computational requirements: Job size vs IO needs

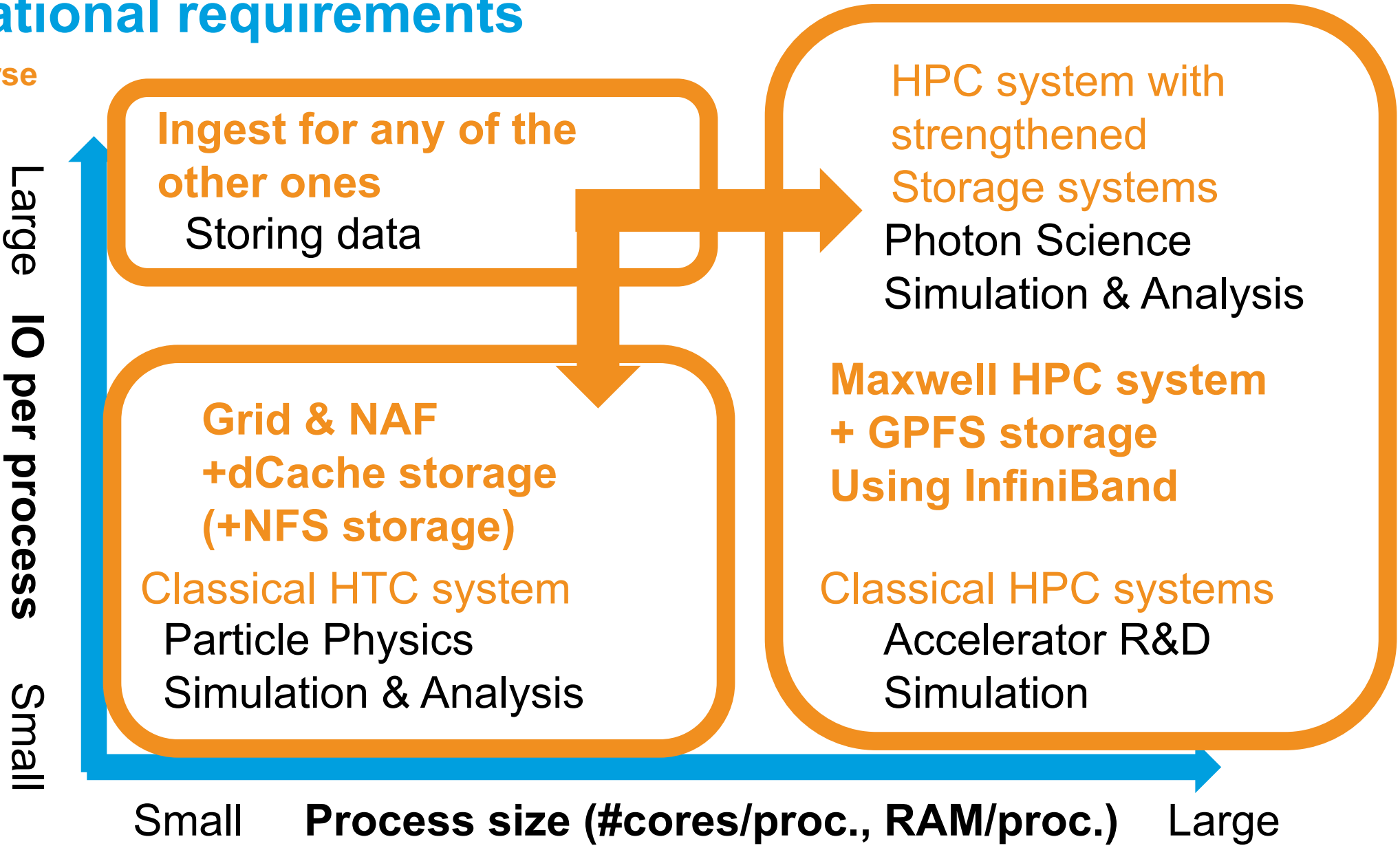
Very very coarse



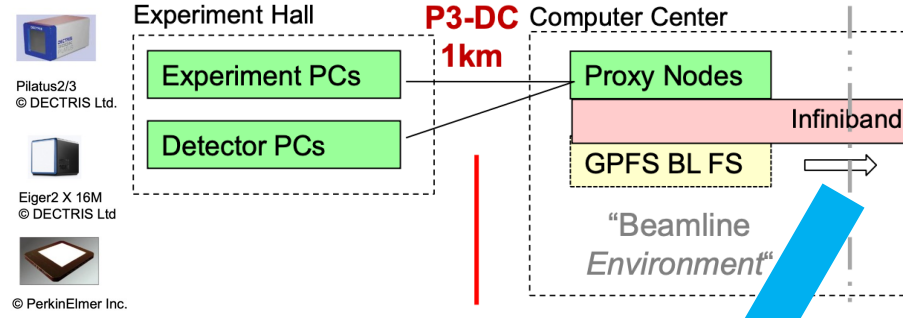
(*) small processes, but many of them

Computational requirements

Very very coarse



The Setup for Photon Science



ssh /
FastX /
Jupyter

Maxwell

Interactive

Job submit

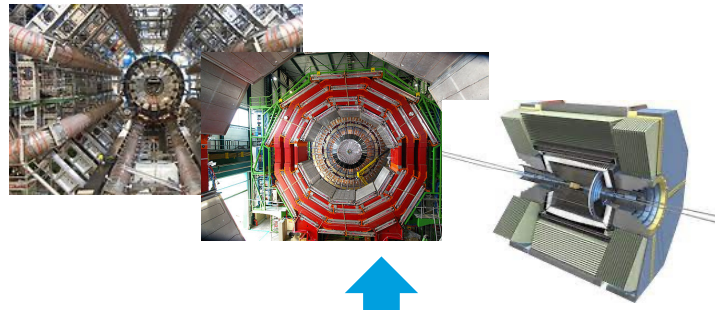
Batch

Storage

Extremly fast
Interconnect:
InfiniBand

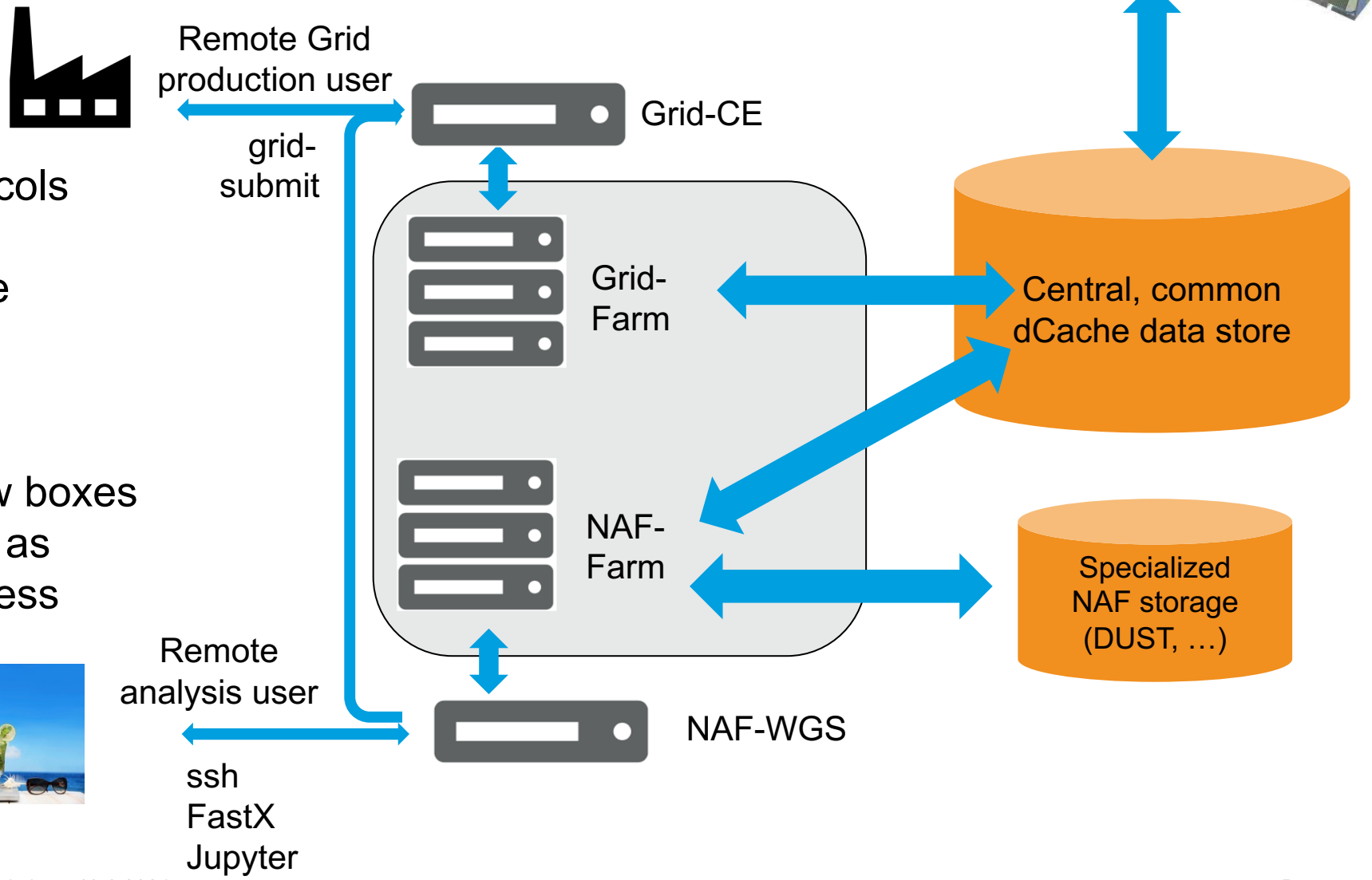
https://www.skyscanner.de/nachrichten/workations-reisetrend-2020

GRID & NAF: The big detailed picture



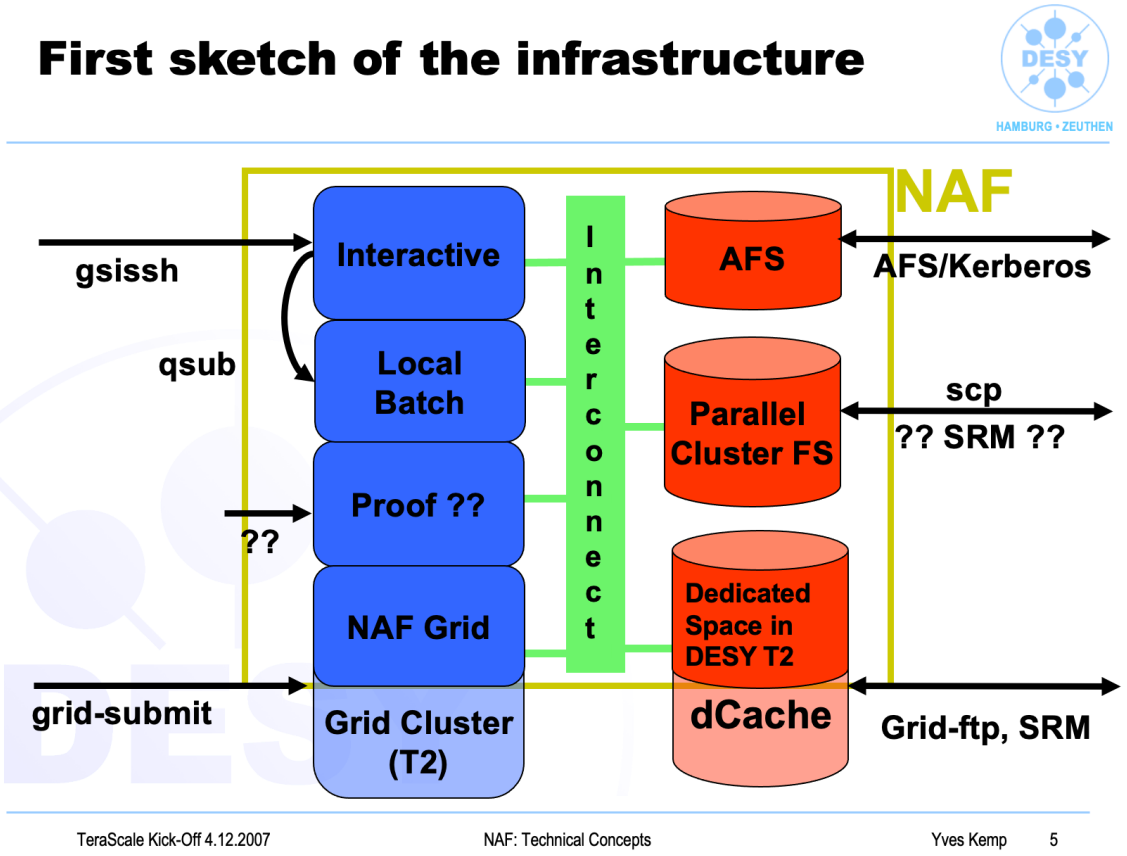
Grid: Serves worldwide HEP community through Grid protocols
NAF: Serves national HEP community through interactive protocols: Fast turn-around

Access protocol is just one/few boxes large compute behind, as well as storage infrastructure and access is (mostly) identical



… base concept in place and operated since 2007

PHYSICS AT THE TERA SCALE
Helmholtz Alliance

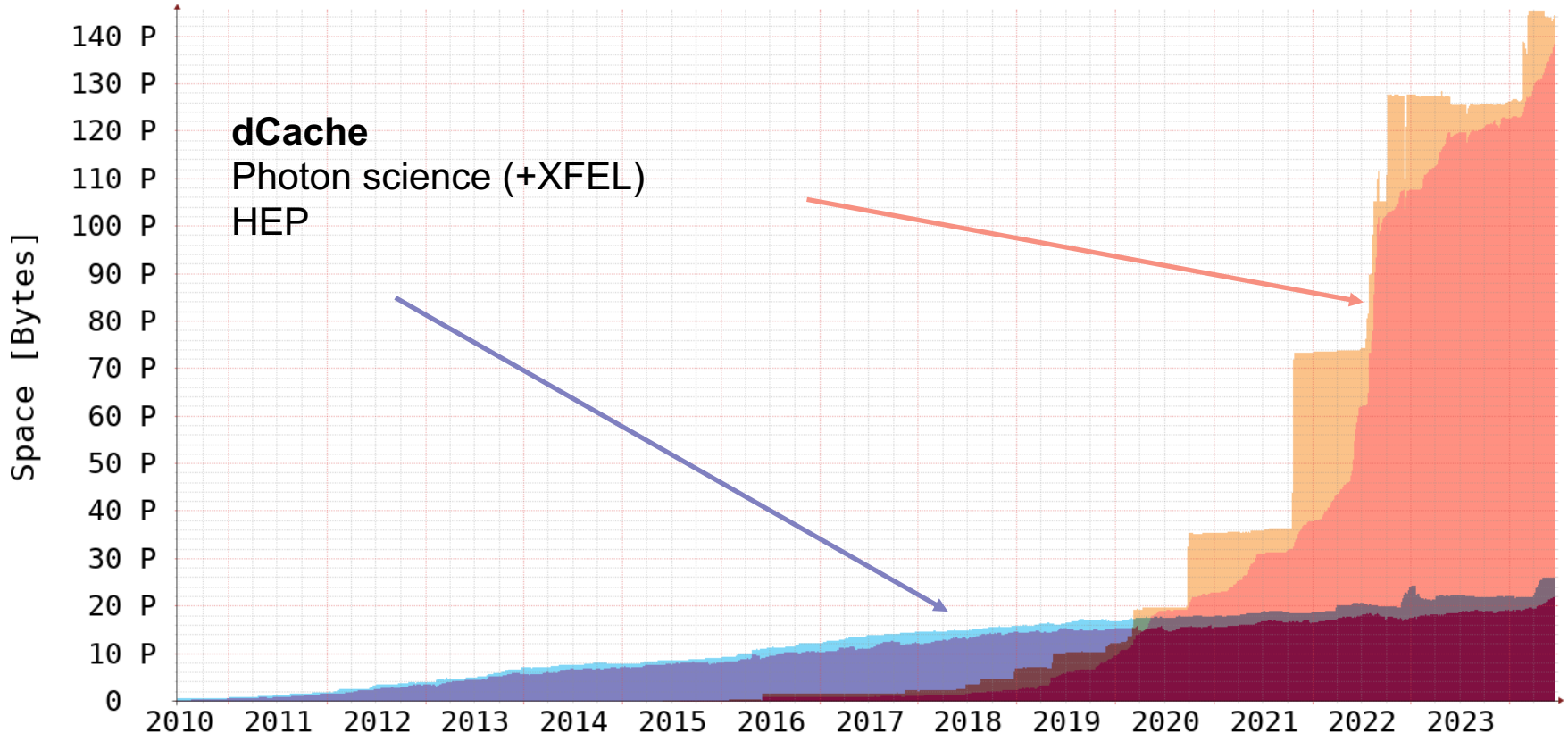


(some) design criteria:

- DESY science is data centric
- Generic setup for *all* participants
- Integration into experiments workflows

Facts and figures June 2024: IDAF

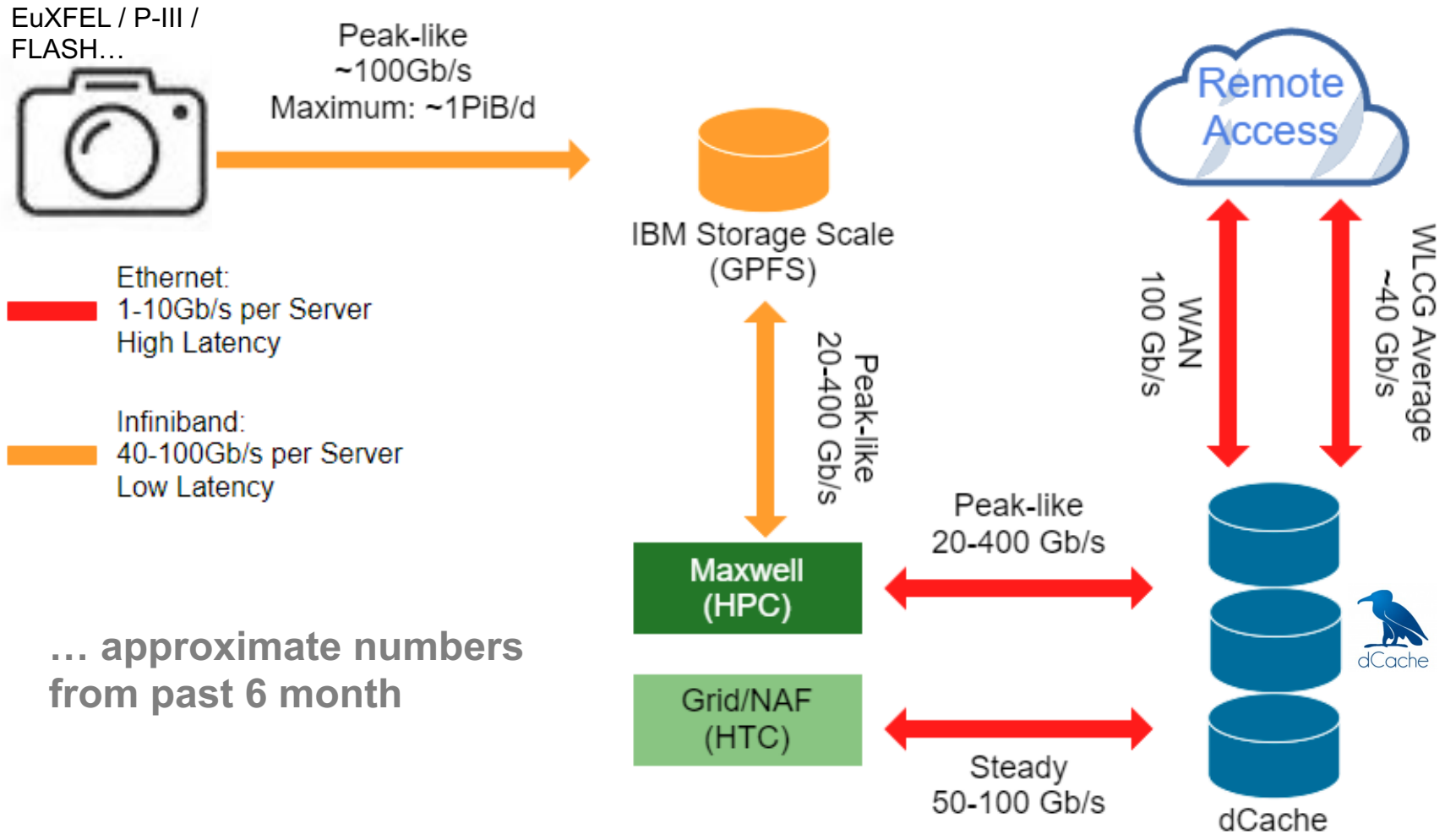
- **Maxwell + Grid + NAF**
- **~180 PB data on disk**
- **dCache + GPFS + BeeGFS**
- **~60.000 CPU cores, ~380 GPUs**
- **HTCondor, SLURM**
- **~2.700 server (compute, storage, management)**
- **~ >0.5 Megawatt**



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



Users of the IDAF

- **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

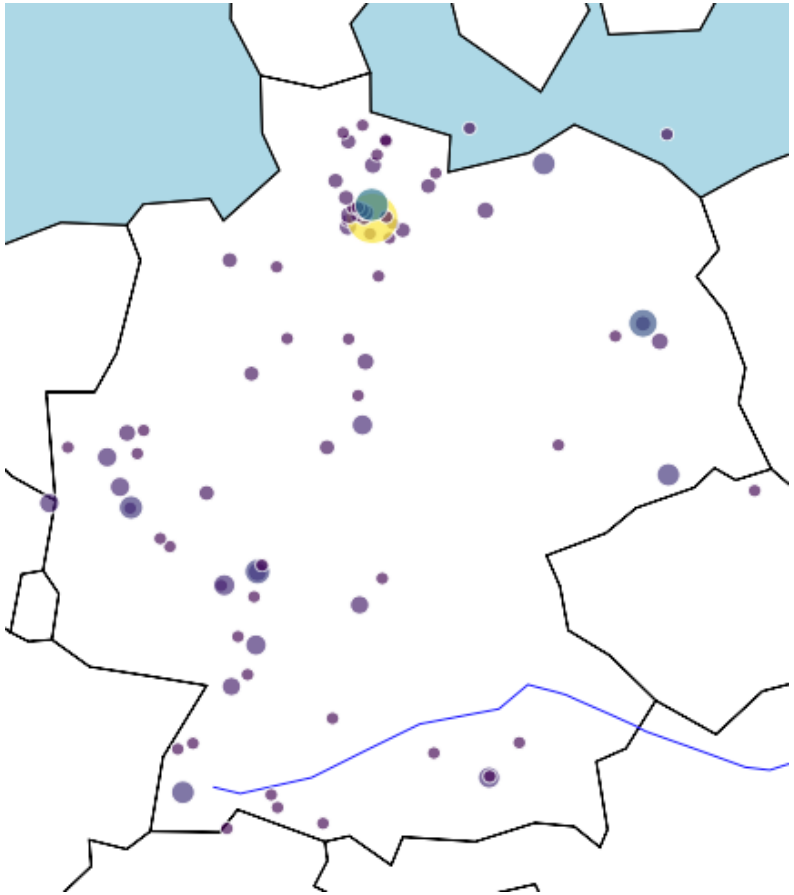
- **Astro-Particle Data**

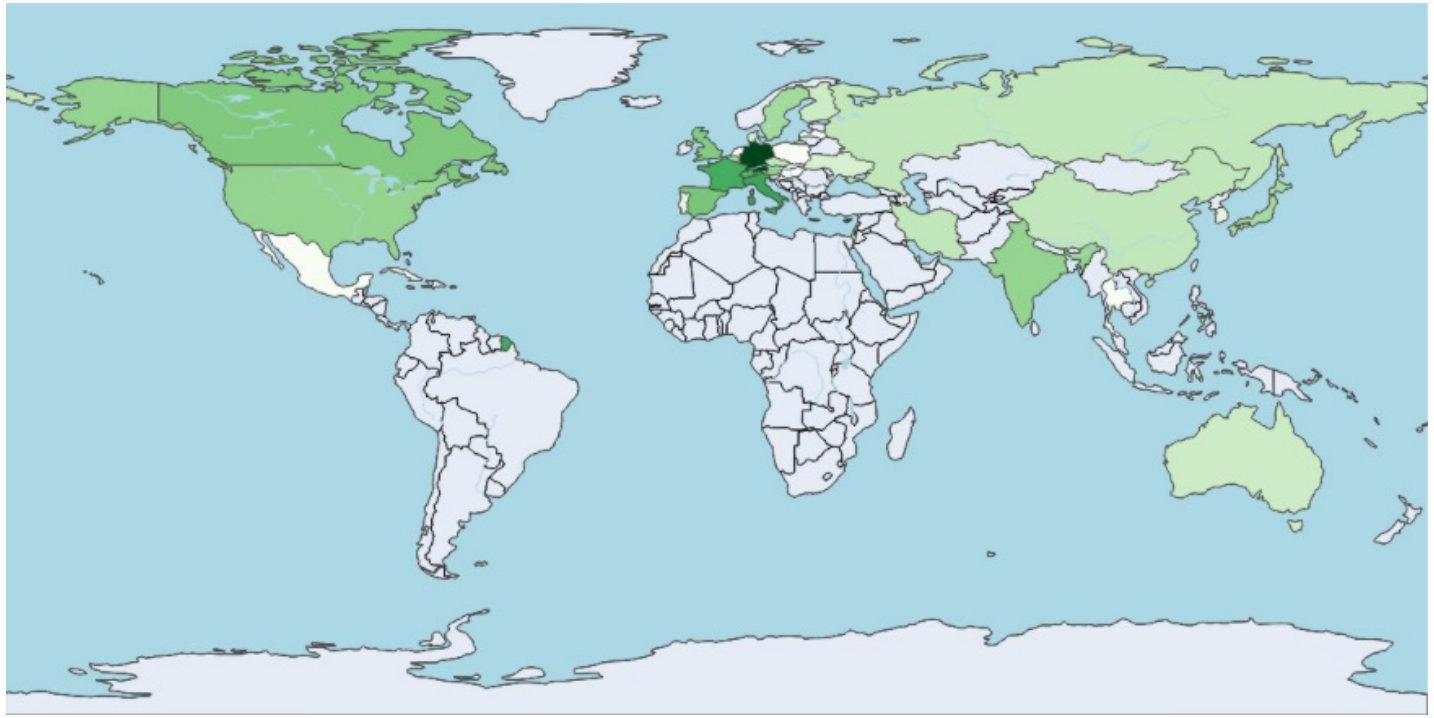
IceCube — H.E.S.S. — cta cherenkov telescope array

Astro- Particle Physics

… and where they come from

logins during two weeks in October 2023





Only NAF & Maxwell logins are accounted for (no Grid submission)

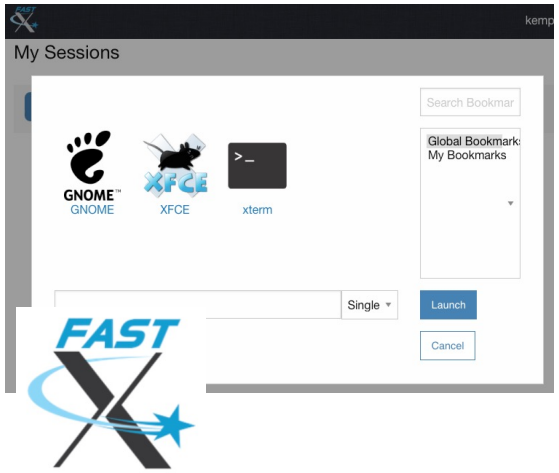
… mostly from academia (universities and institues)

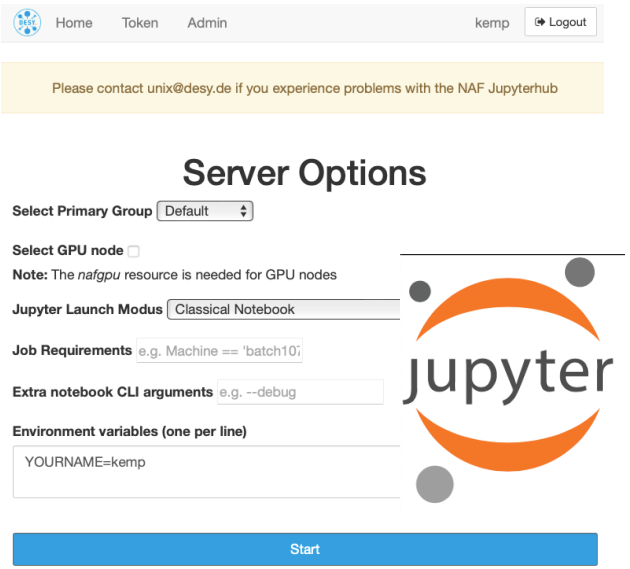
… some commercial users

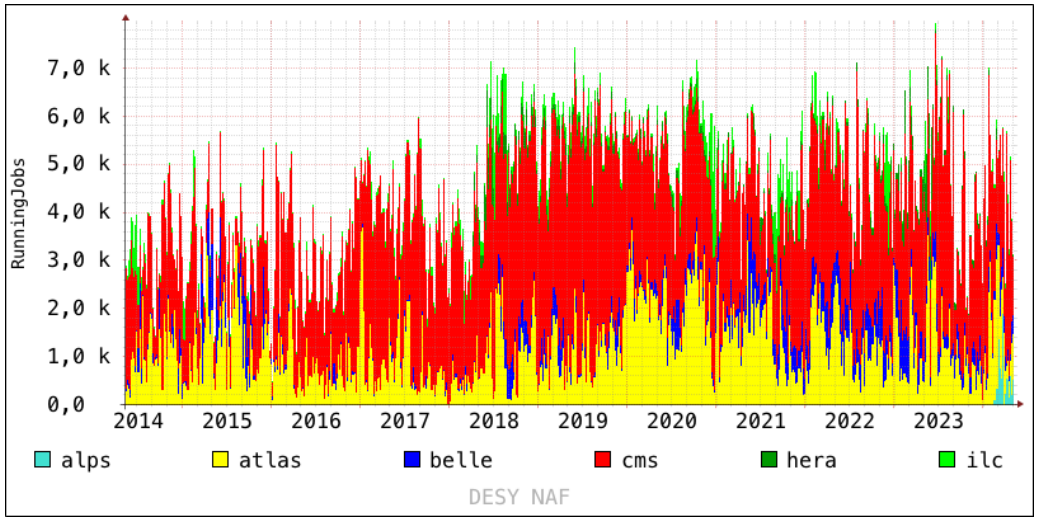
Some highlights of the current NAF (and Maxwell) setup

Hardware setup: compute

- Login nodes: NAF:
  - min. 2 nodes / VO, larger ones ~10 nodes
  - virtualized, around 8 cores / VM to spread load
  - ssh login
- FastX nodes: NAF: dedicated VM offer graphical login
- Maxwell: O(10) Nodes with GPU for ssh + FastX
- JupyterHub
  - a small VM handels external queries, and forwards to the batch system
- Batch:
  - NAF: ~300 nodes, ~10.000 physical cores, 14 GPUs, HTCondor
  - Maxwell: ~900 nodes, ~30.000 phys.cores, ~370 GPUs, SLURM



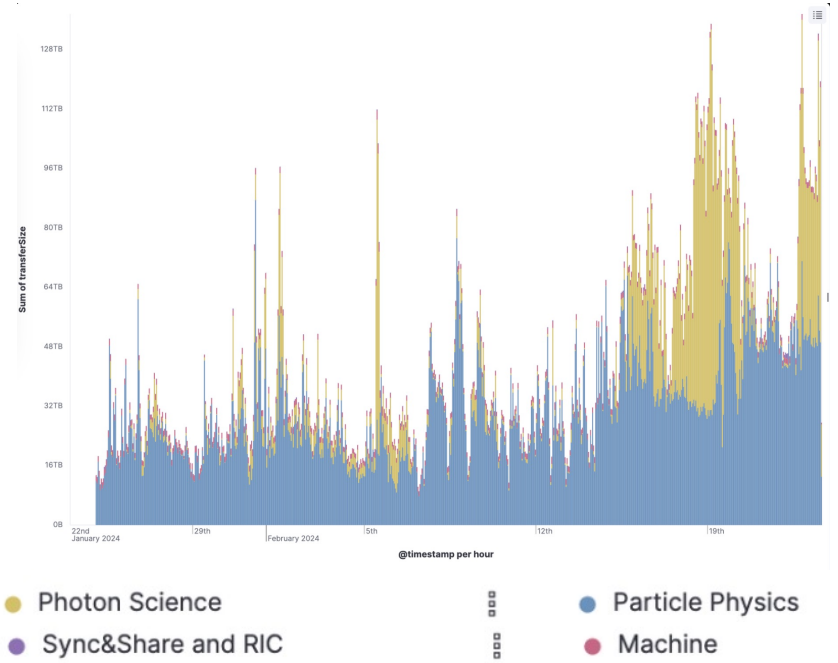




NAF storage:

- dCache:
  - Shard access from Grid & NAF to experiments' dCache
  - Dedicated space for non-pledged usage
  - Different protocols possible, NFS mount stands out
- Fast project space:
  - "DUST" (GPFS system, ~2,6 Pbyte) for users and groups
  - Typically 1+ TB quota per user
- DESY AFS cell for $HOME
- Different CVMFS repos

- Observed bandwidth to NAF dCache(s): Up to 250 Gbit/s
- Access governed by POSIX ACLs, and based on UID/GUID(s)



More on POSIX / mounted files system access

Users prefer to use mounted netFS with (some) POSIX semantics

- 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
```

- Usually only option for applications from photon science and acceletor R&D
- Trend includes HEP despite remote read capabilities
- **Poses the challenge of having uniform name-space across the IDAF**

HPC

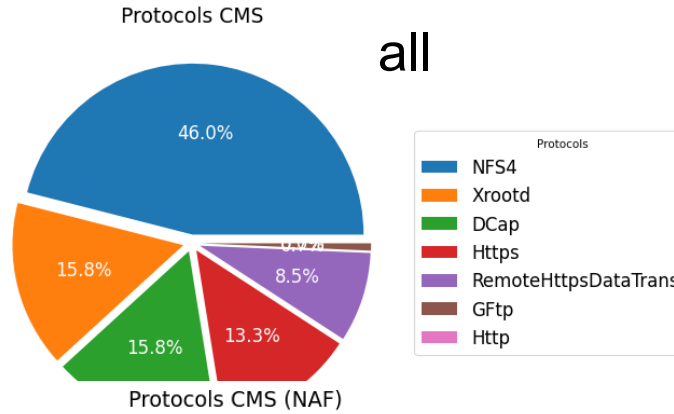
[vossc@max-display008] ~ $ md5sum /gpfs/dust/belle2/user/vossc/stage-rest-api.out
0108f37dbbb38103bba6d836f356d7b7 /gpfs/dust/belle2/user/vossc/stage-rest-api.out

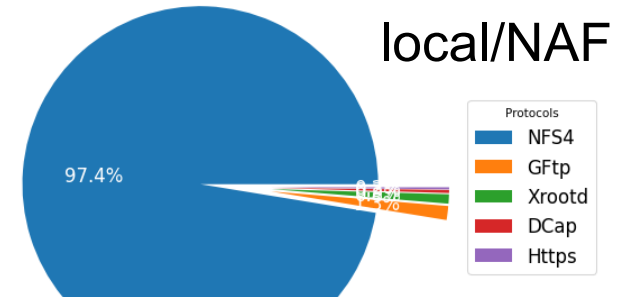
HTC

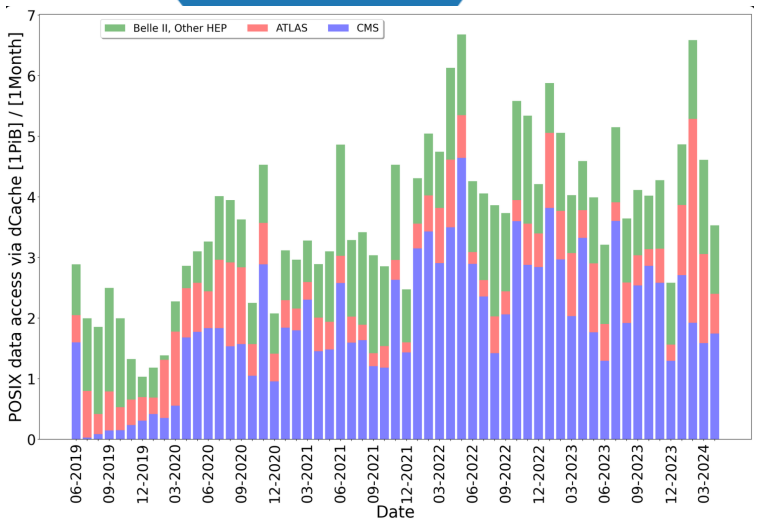
[vossc@naf-belle12] ~ $ md5sum /nfs/dust/belle2/user/vossc/stage-rest-api.out
0108f37dbbb38103bba6d836f356d7b7 /nfs/dust/belle2/user/vossc/stage-rest-api.out

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

Data Access CMS May 2023







Recent portal developments

Providing applications for Maxwell users

General purpose

Julia apps

Generic utilities

NoVNC EL9

Virtual Desktops

Debian + CrystFEL

mlExchange

Scientific Applications

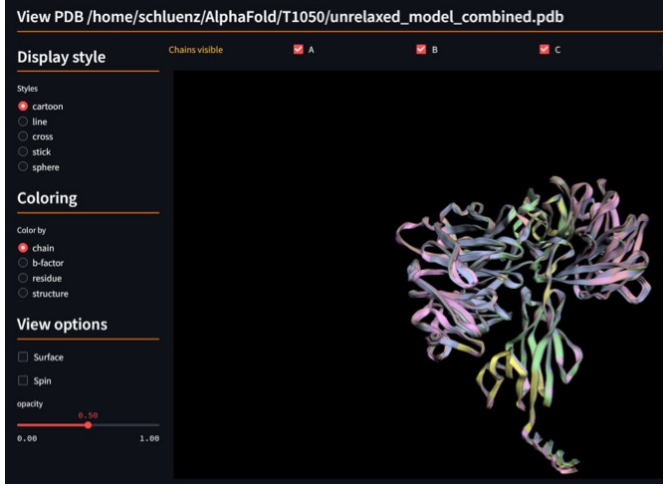
HDF5 Viewer

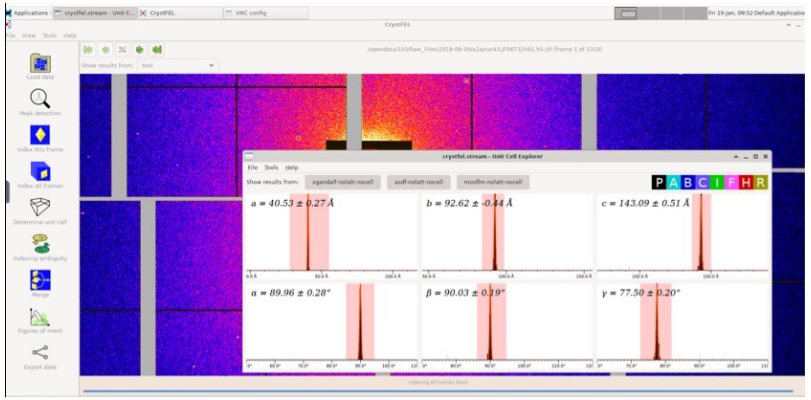
SciCat Frontend

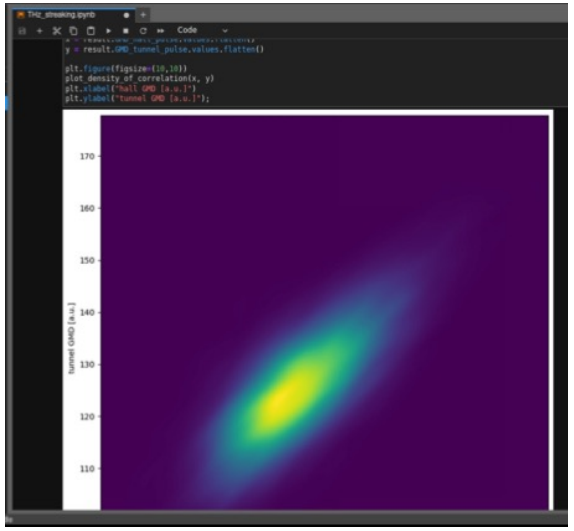
Named Servers

Note: this is a test setup. It won't work most of the time! Start a named server, or use one of the pre-configured applications.

Server name	URL	Last activity	Actions
Name your server Add New Server			
hdf5		a month ago	start delete
scicat		a month ago	start delete



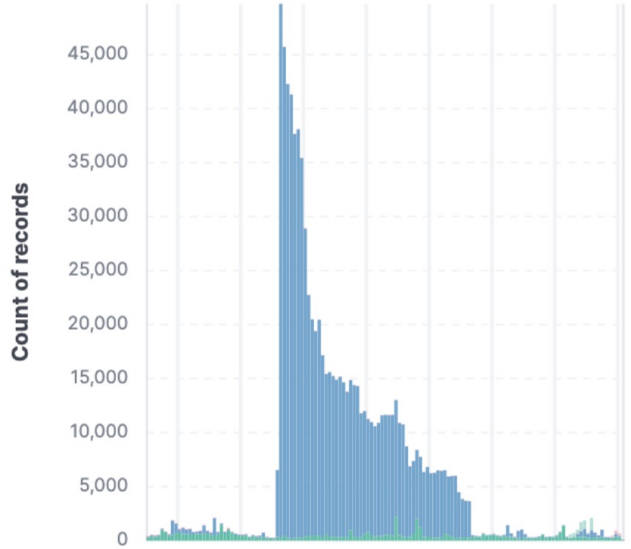


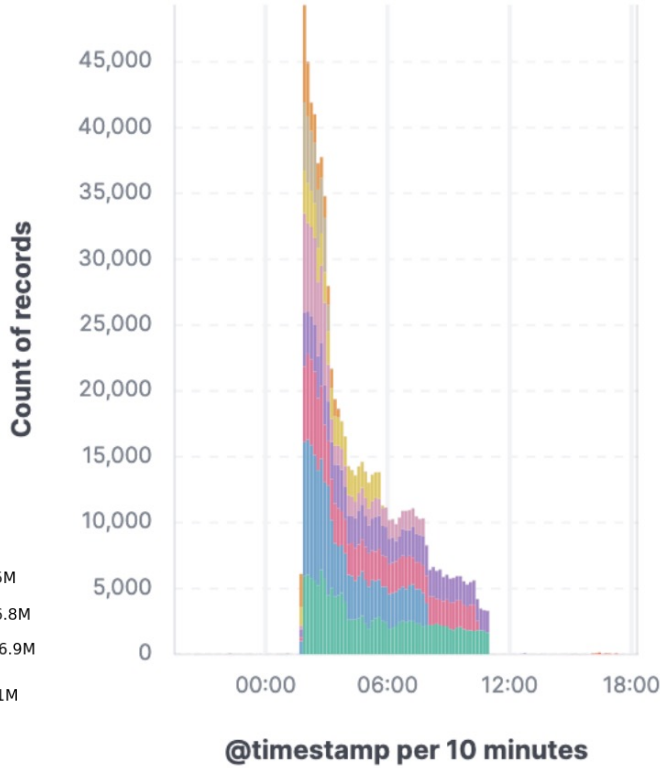


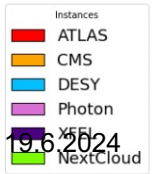
Improving Monitoring and Analytics

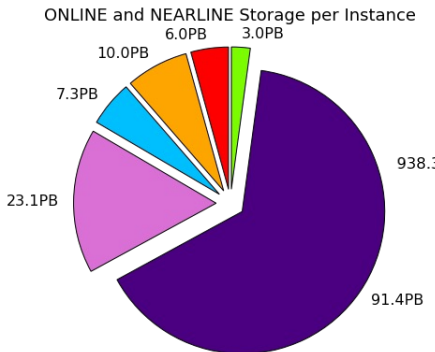
Managing and Understanding the Change User Access Patterns

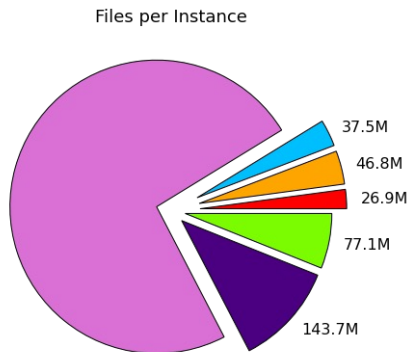
- Increasing capacity found to manageable
  → read/write patterns found to be more challenging
- Departure from classic C/C++ or FORTRAN driven batch analysis
- Ease-of-Use of Python leads to higher memory footprint and excessive, repetitive data access (open files to read <1MiB)
- Increased WAN/Tape 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











GPUs in NAF and Maxwell

- GPUs used for computational purposes on Maxwell started with Kepler generation (around 2014/2015)
- GPUs on NAF introduced in 2018 … only a small number, little usage
  - Some HEP users invested into GPU machines → Put into Maxwell
- GPUs on Maxwell: Taken up speed … currently around ~380 GPUs of different generations
  - plus installation and support of GPU / ML / AI related software
- Future of GPUs:
  - Get out of NVIDIA vendor lock-in → Generic tools beyond CUDA
  - Trend to larger and larger systems (NVIDIA DGX pods, …)
  - Unify NAF and Maxwell to make efficient use of GPUS

Sustainability: Green-IT

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?

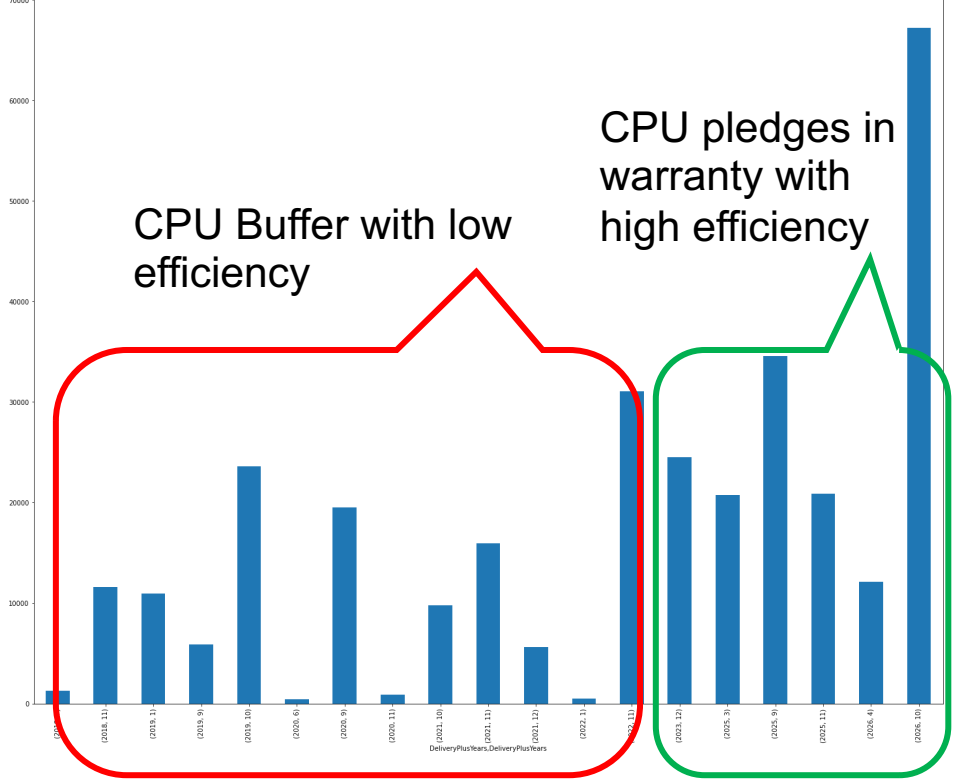
Raising **awareness** of users

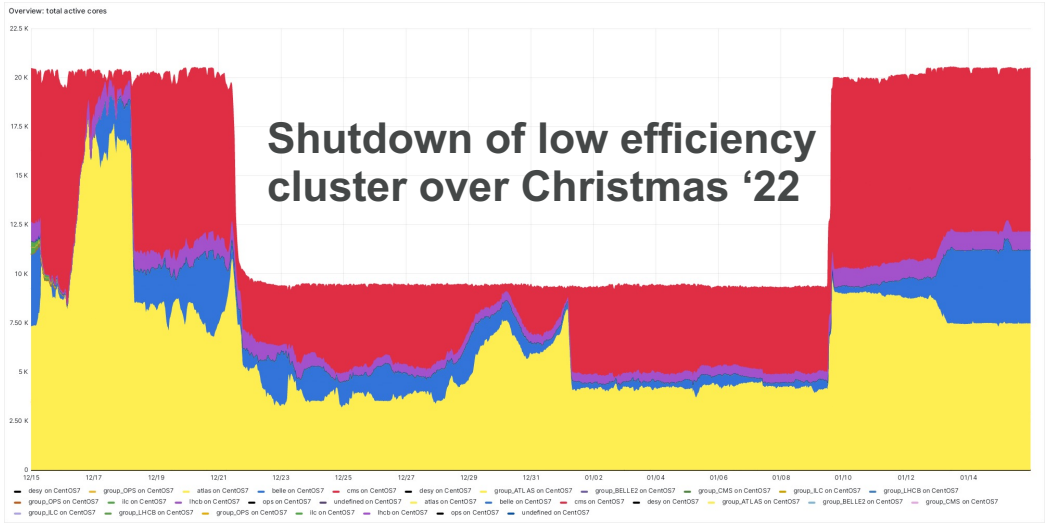
Train users on most efficient use of IDAF

Train users on tooling and optimal algorithms

Interactivity and fast reaction come with inefficiencies:

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





Provided by T. Hartmann

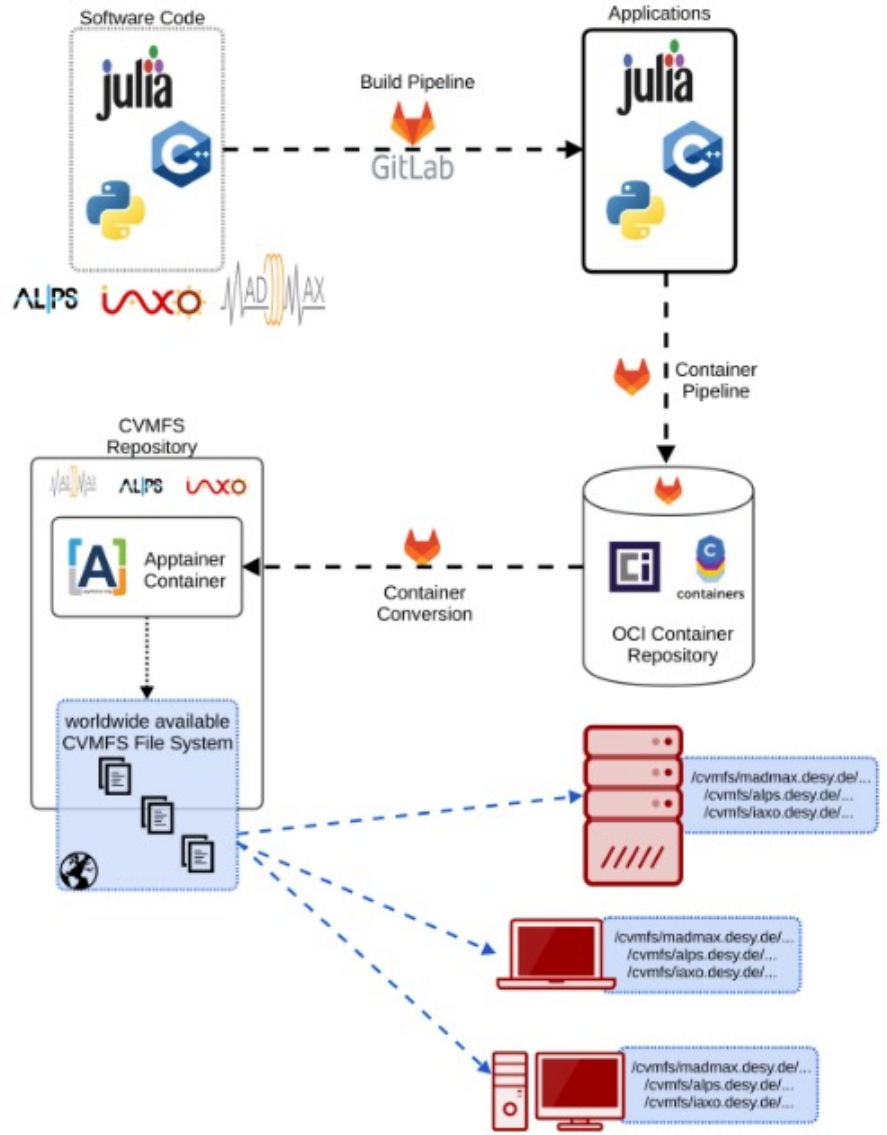
Software and containers

- Software provisioning for HEP users a topic since the beginning of the **NAF**
  - Mostly solved using CVMFS for the large VOs. DESY provides CVMFS for small VOs.
  - If needed, also installation on shared filesystem is possible.

  → Experiments provide their software ; DESY-IT provides standard software
- Software provisioning for photon science users still a topic with **Maxwell**
  - No CVMFS to draw from, some non-free software → minor role of CVMFS

  → DESY-IT provides abundant list of photon science software, mostly via shared filesystem … and some application support
- Users can run containers on the batch system since several years
  - integrated both in NAF (HTCondor) and Maxwell (SLURM)
  - preferably Apptainer or SingularityCE
  - Build pipelines for Containers incl. CVMFS are possible

  → Distribute those artefacts wider than just the DESY NAF



Support, documentation, training, consulting, governance

- User support one of the most crucial pillars of success for an Analysis Facility
  - Split support model: DESY-IT facility questions, experiment expert for their topics
  - Kind of works … but is manpower intensive … can make this setup more efficient?
- Documentation: Tedious work, but important.
  - AI & LLM can offer a new level of interaction with documentation and eventually support
- Training: IDAF experts involved in training for newcomers as well as experienced users
- Consulting: IDAF experts consult new groups, or discuss new requirements and workflows
- Governance: NAF: regular NAF Users Committee, yearly NAF Users Meeting, review by DESY Physics Research Committee

Ideas for the Future

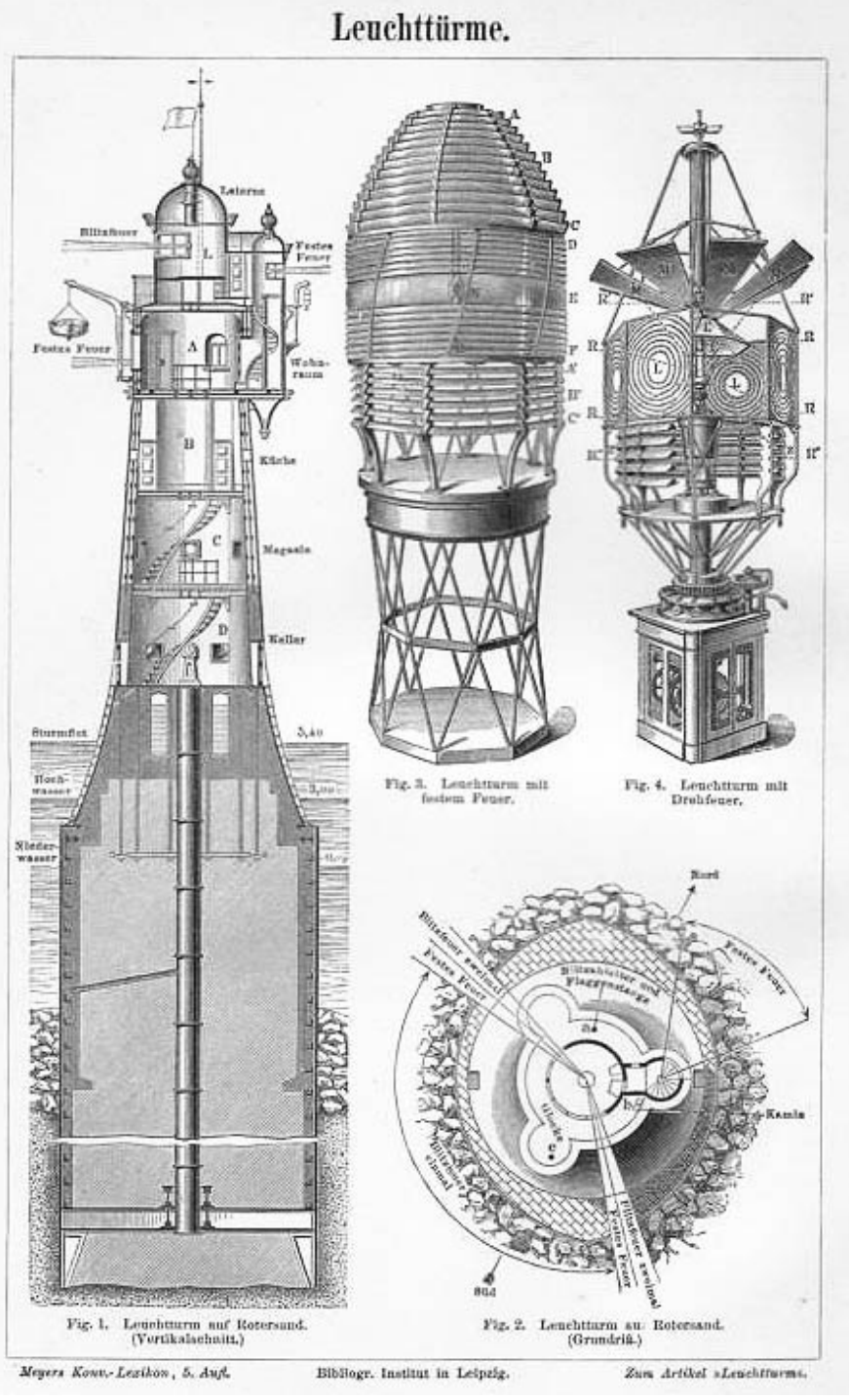
Plans for evolution … users view

- Seamless integration of application portals
- Seamless federated user access
- Seamless online & interactive resource and data access
- Seamless integration with FAIR and open data repositories

Wikimedia
Leuchtturm in Westerheversand.jpg
Marco Leiter

Plans for evolution … under the hood

- Balance user friendliness and accessibility with security, scalabililty and sustainability
- Towards a more homogeneous IDAF concept to support heterogeneous user communities and heterogeneous compute hardware
- Disruptive compute hardware evolution
  - GPU becoming more and more ”mainframe”
  - (maybe quantum on the far horizon?)



https://www.etsy.com/de/listing/255985292/antike-kontenplan-alte-leuchtturme

Plans for evolution … actual doing

DESY IDAF is a unique environment for research and innovation

- Large, diverse user community, requirements, workflows
- Large, scalable infrastructure: Storage, Compute, Network

DESY IDAF is open to integrate, leverage and scale novel concepts and developments

- Also integrate other partners, also with universities and other institutes
- Philipp Neumann is both new DESY-IT head and professor at Universität Hamburg on High Performance Computing and Data Science

https://www.greetsiel.de/sehenswuerdigkeiten/campener-leuchtturm

Summary and outlook

- DESY offers analysis facilities for several communities
- NAF in operation since 2007
- IDAF puts NAF, Grid and Maxwell-HPC under a common umbrella

Strong, data centric core enables

- stable, scaling, sustainable operation
- flexible developments and adaption

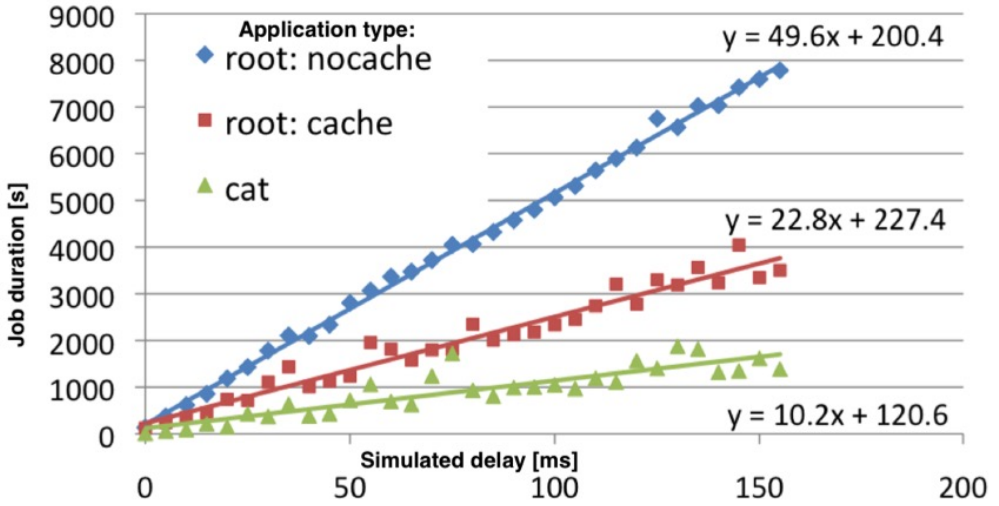
(we’re still laking a logo)

Backup slides

NAF, NAF 1 and NAF 2.0

- Original design and setup in 2007
- Rework in 2013

(Some) changes and their background:



Experience with HEP analysis on mounted filesystems, *J.Phys.Conf.Ser.* 396 (2012) 042020

NAF (1)

- Spread ”transparently” between Hamburg and Zeuthen
- Separate user registry, non-DESY accounts, X509 based logins
- Separate admin tools, separated from DESY network

NAF 2.0

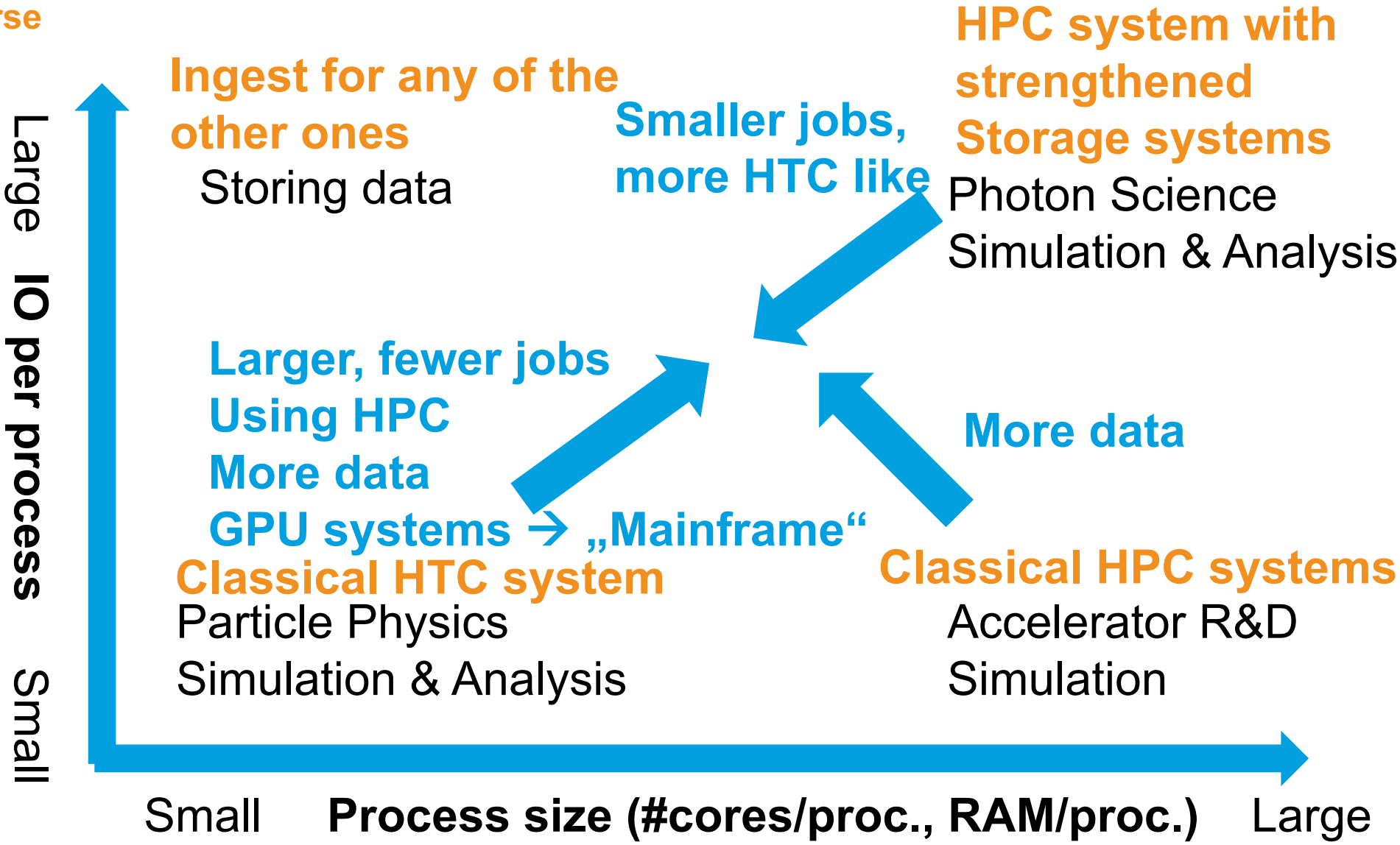
- Located only in Hamburg
- Normal DESY accounts, incl. passwords
- Hamburg admin tools, integrated into Hamburg network

Role Based Access for Photon Science data @ Maxwell

- Static ACL configuration for ASAP3 and XFEL
- Roles based on unix group membership
  - <beamtime id>-dmgt → Data Manager, allows read/write/delete in all folders
  - <beamtime id>-part → Participant, allows read/write/delete, except write/delete in raw folder
  - <beamtime id>-clbt → Collaborator, Read-only access
- Same scheme used for ASAP3 and XFEL
- Group memberships are managed via
  - Gamma Portal for ASAP3
  - Meta Data Catalog for XFEL
- Example
  - ASAP3 → 10000000-dmgt
  - XFEL → 60900009-dmgt

Computational requirements are changing

Very very coarse



Getting data from the experiment to Maxwell

ASAP::O Docs Changelog API 23.12.0

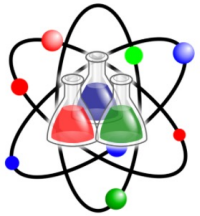
ASAP::O

High performance distributed streaming platform

Get Started

Designed to be Fast

ASAP::O was designed to be able to keep up with huge data volumes and frame rates of next generation high-speed detectors.



Focus on What Matters

ASAP::O lets you focus on science, while we'll take care of nasty details like storage and network and deliver your data right where you need it.

Easy to Use

ASAP::O API is avaiable in Python or C/C++ and is quite simple. Just couple lines of code and you can start using your data.