

# Large scale computing infrastructure at DESY Interdisciplinary Data Analysis Facility (IDAF)

Christian Voss, Yves Kemp, for DESY IT  
PunchLunch  
2025-01-15 DESY

With slide contributions from DESY and XFEL people

# DESY research divisions ... In a nutshell



## Accelerators »

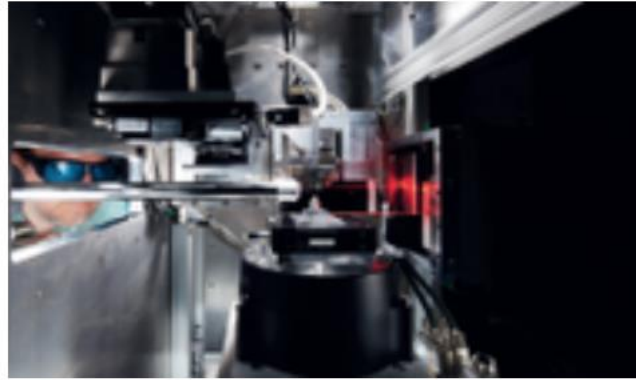
Running / Operating:

- 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

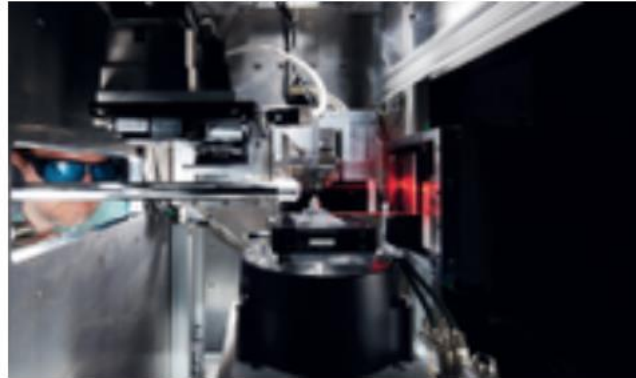
# DESY research divisions ... IT involvement in scientific computing

... An incomplete view



## Accelerators »

- Storage operational data
- Simulation & computational infrastructure for R&D
- Support



## Photon science »

- Online DAQ
- Offline storage & analysis infrastructure
- Simulation & computational infrastructure for R&D
- Support



## Particle physics »

- Global and national tasks within LHC and BELLE II
- Simulation & computational Infrastructure for ILC and detector R&D
- Support

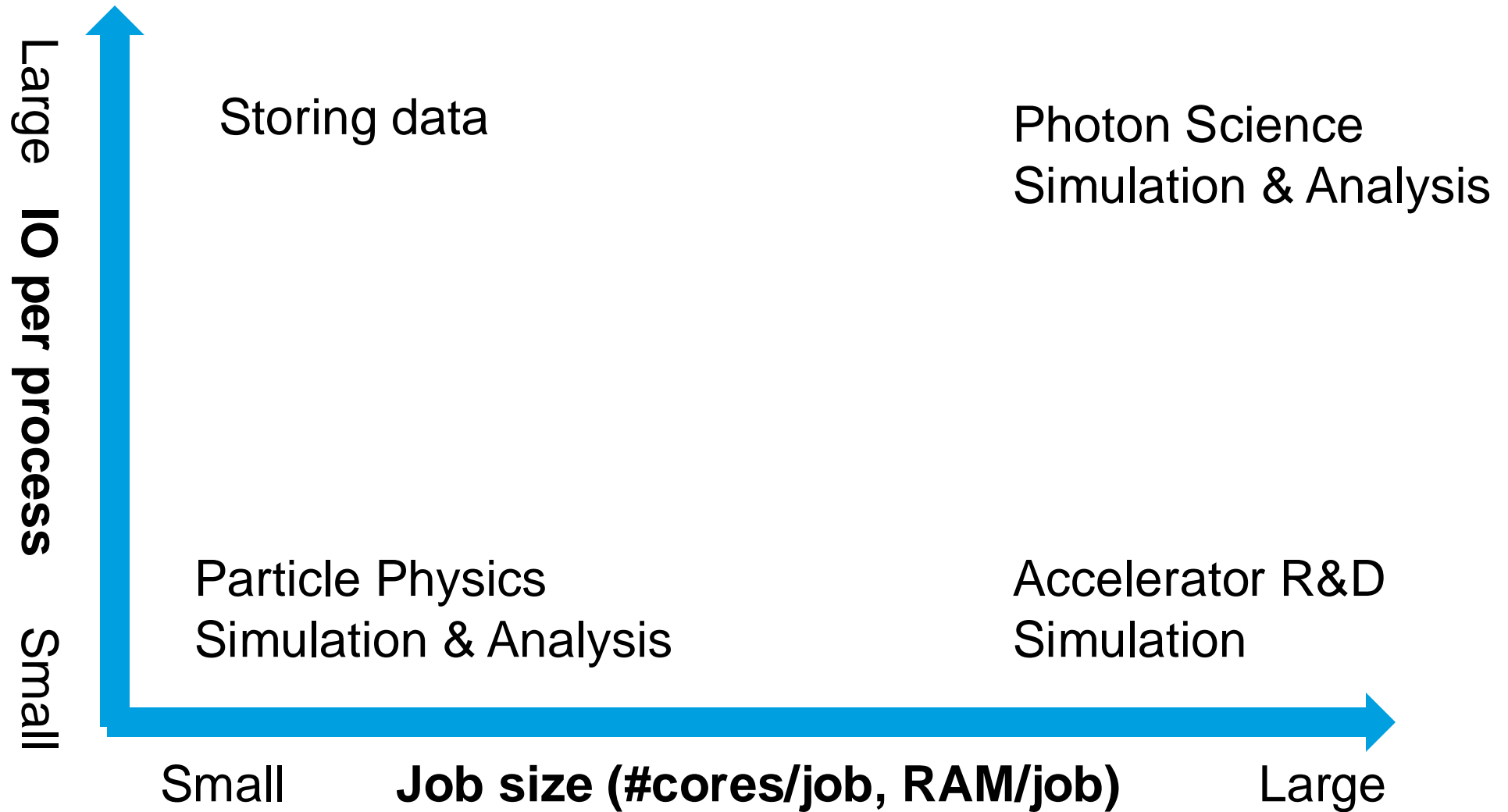


# Now, designing the compute and storage infrastructures



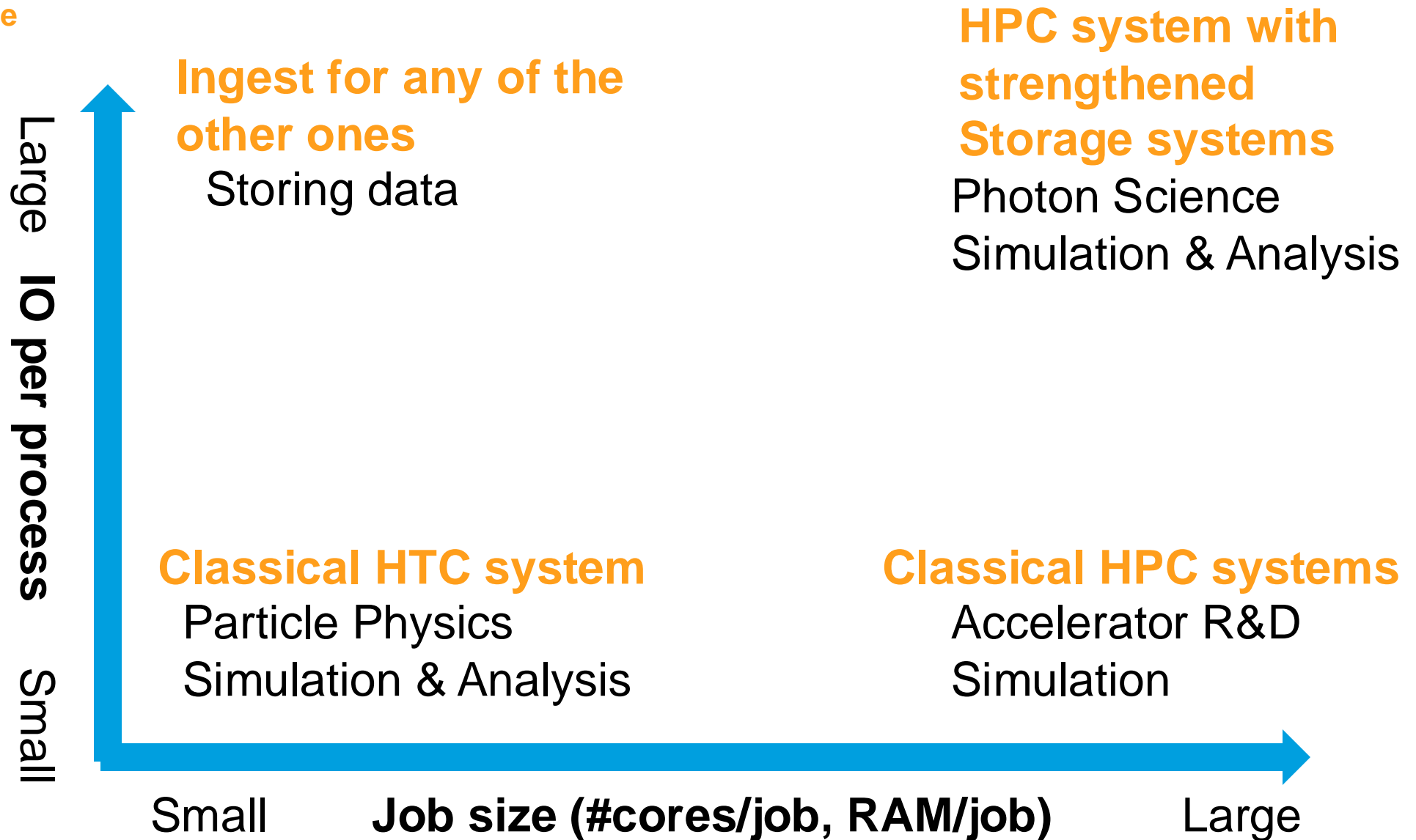
# Computational requirements: Job size vs IO needs

Very very coarse



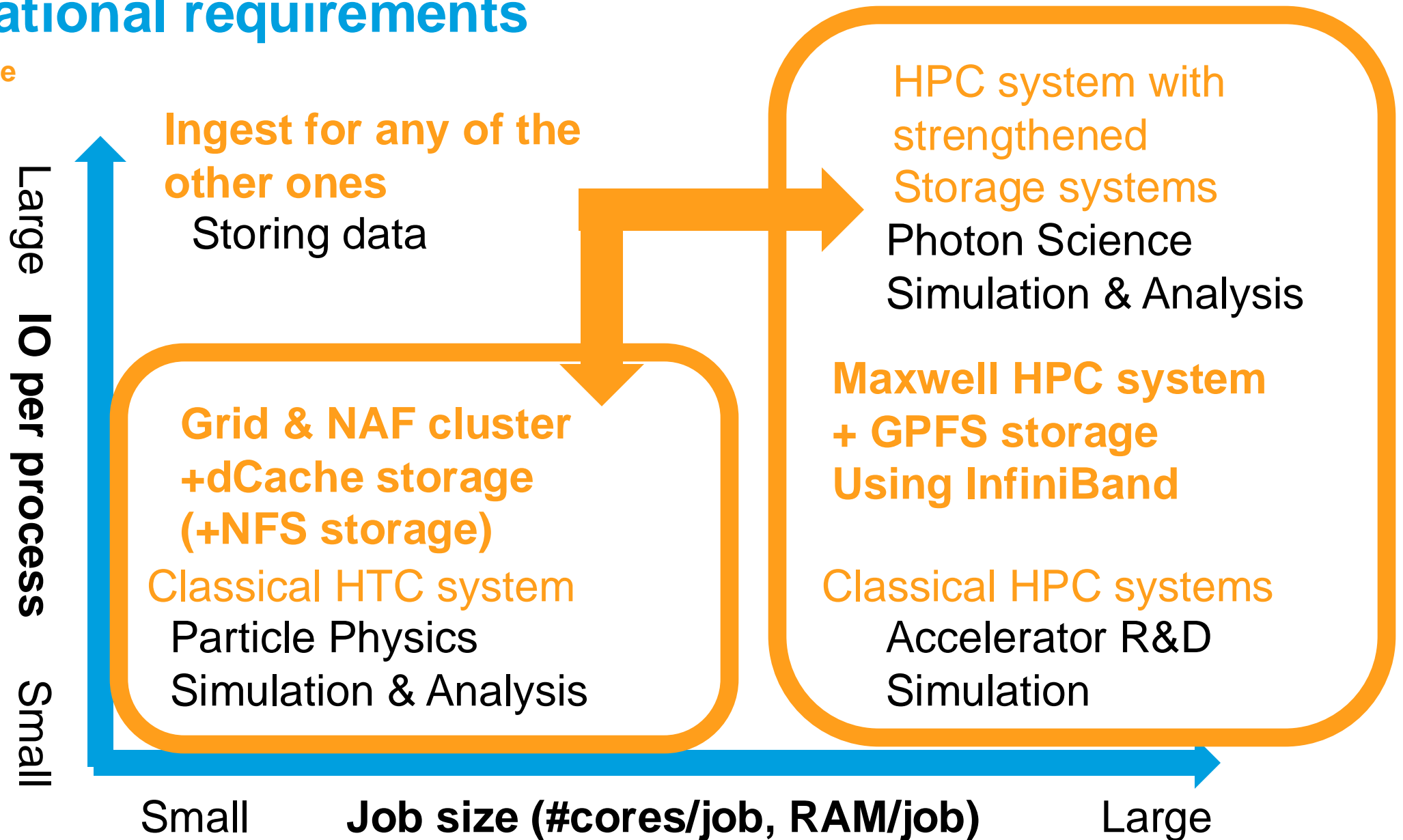
# Computational requirements

Very very coarse



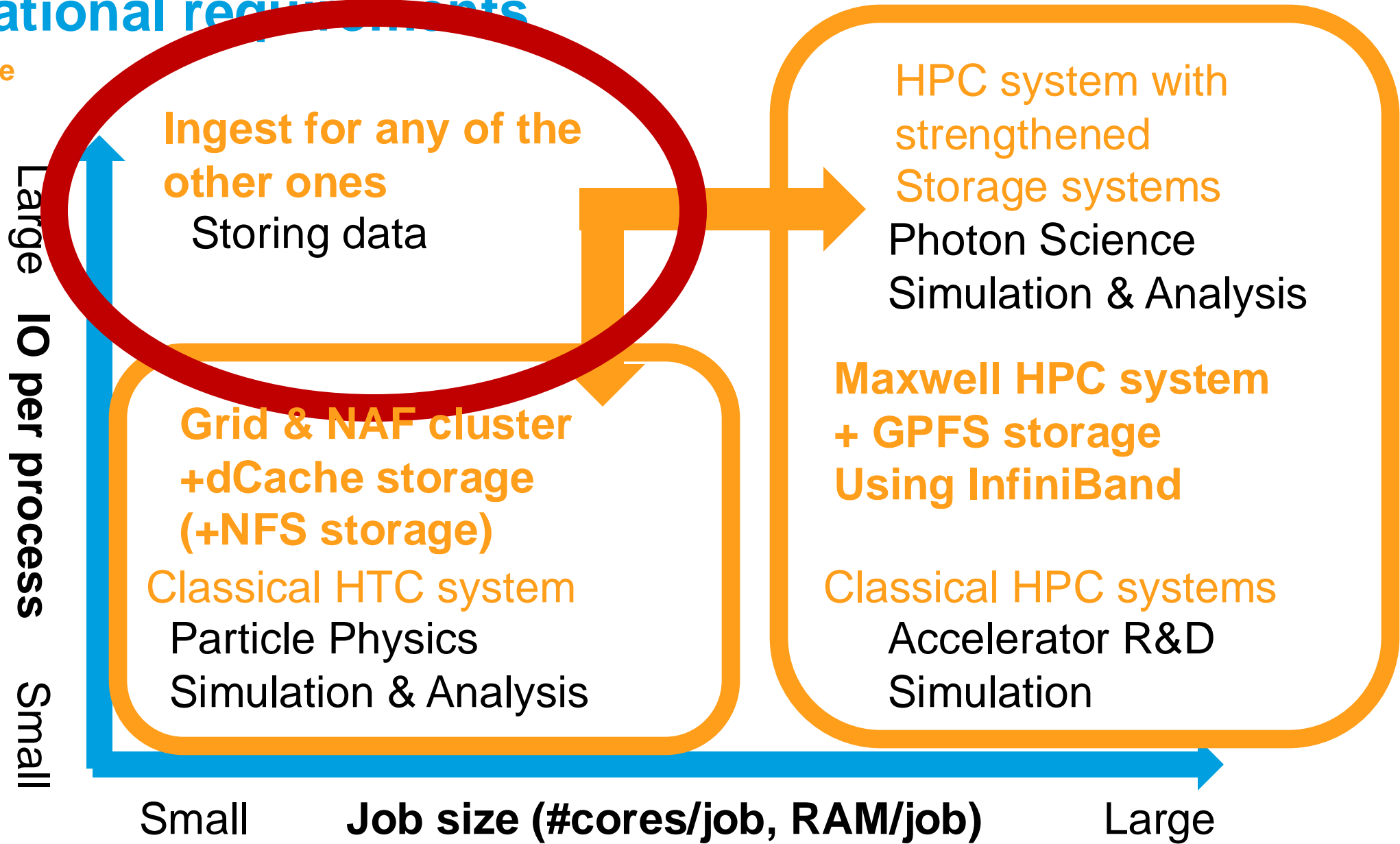
# Computational requirements

Very very coarse



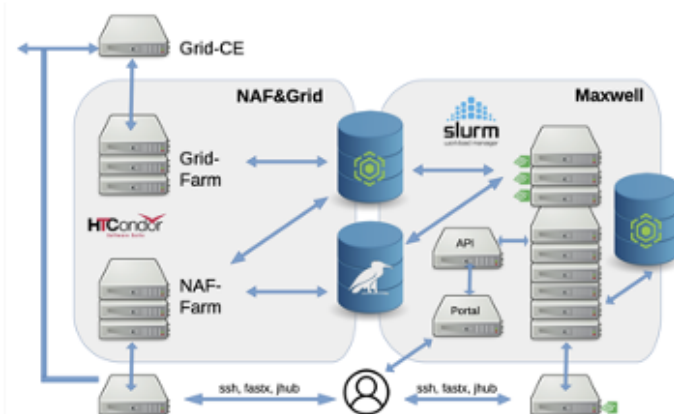
# Computational requirements

Very very coarse





# The IDAF in a nutshell



- Accelerator Data
  - Accelerator Development Data
  - HPC simulations
  - Test-beam data
- Detector and Accelerator R&D

- Facility User Data
  - Data of external Partners
- Research with Photons

- Particle Physics Data
  - Astro-Particle Data
- Astro- Particle Physics

<b>CPU nodes</b>	~1500
<b>CPU cores</b>	~60.000
<b>GPUs</b>	~400
<b>Node IO</b>	10 Gbit/s (Ethernet)– 100 Gbit/s (InfiniBand)
<b>WAN bandwidth</b>	2x 50 Gbit/s
<b>Internal traffic</b>	up to 250 Gbit/s dCache IO
<b>dCache storage</b>	~150 Pbyte @ 2 Giga-files
<b>GPFS storage</b>	~60 Pbyte @ 1,5 Giga-files

# The IDAF is about *people and experiments*

- Accelerator Data



- Accelerator Development Data



- HPC simulations
- Test-beam data

Detector and  
Accelerator R&D

- Facility User Data



- Data of external Partners



Research with  
Photons

- Particle Physics Data



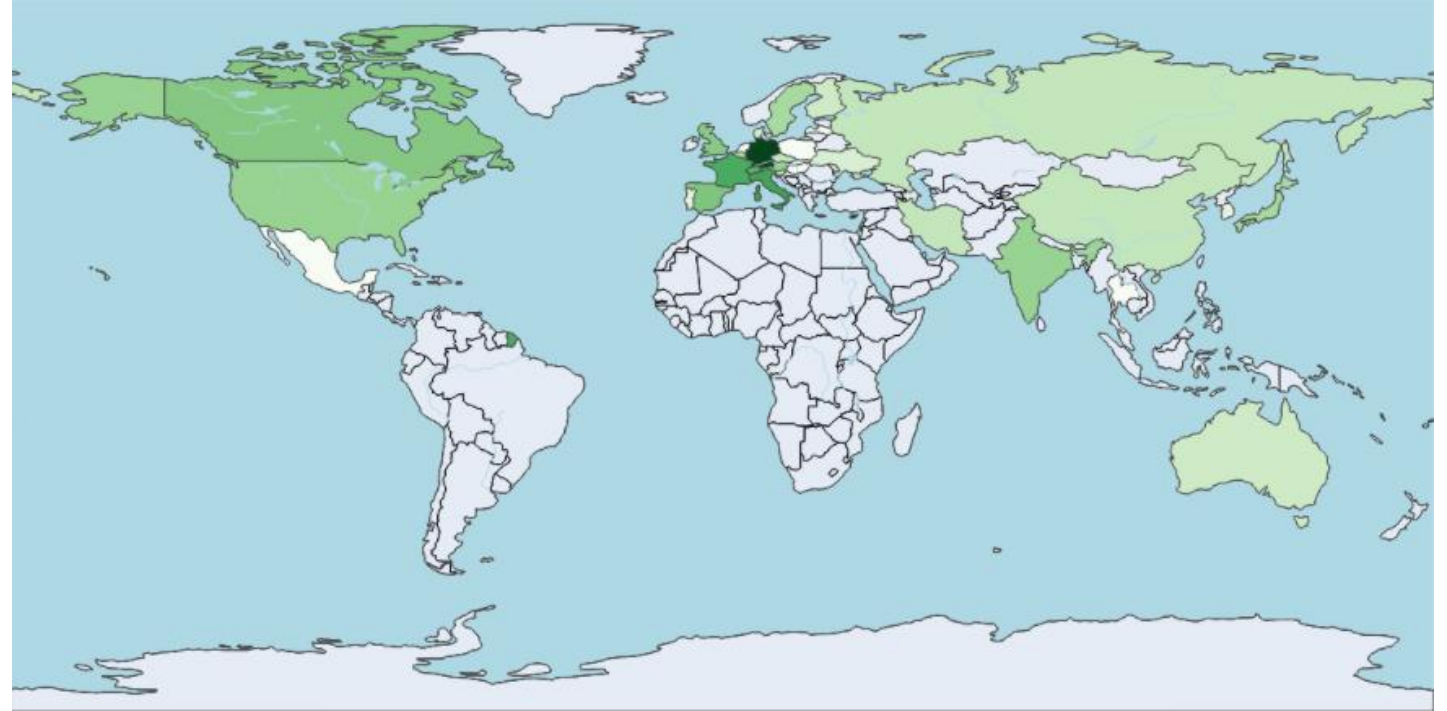
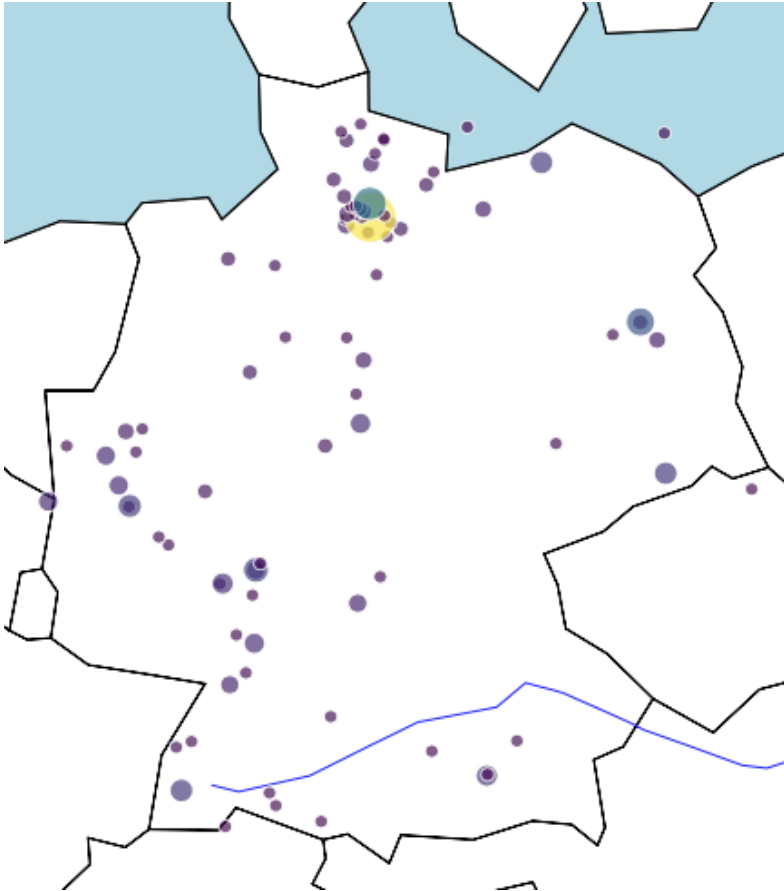
- Astro-Particle Data



Astro- Particle Physics

# ... and where they come from

logins during two weeks in October 2023

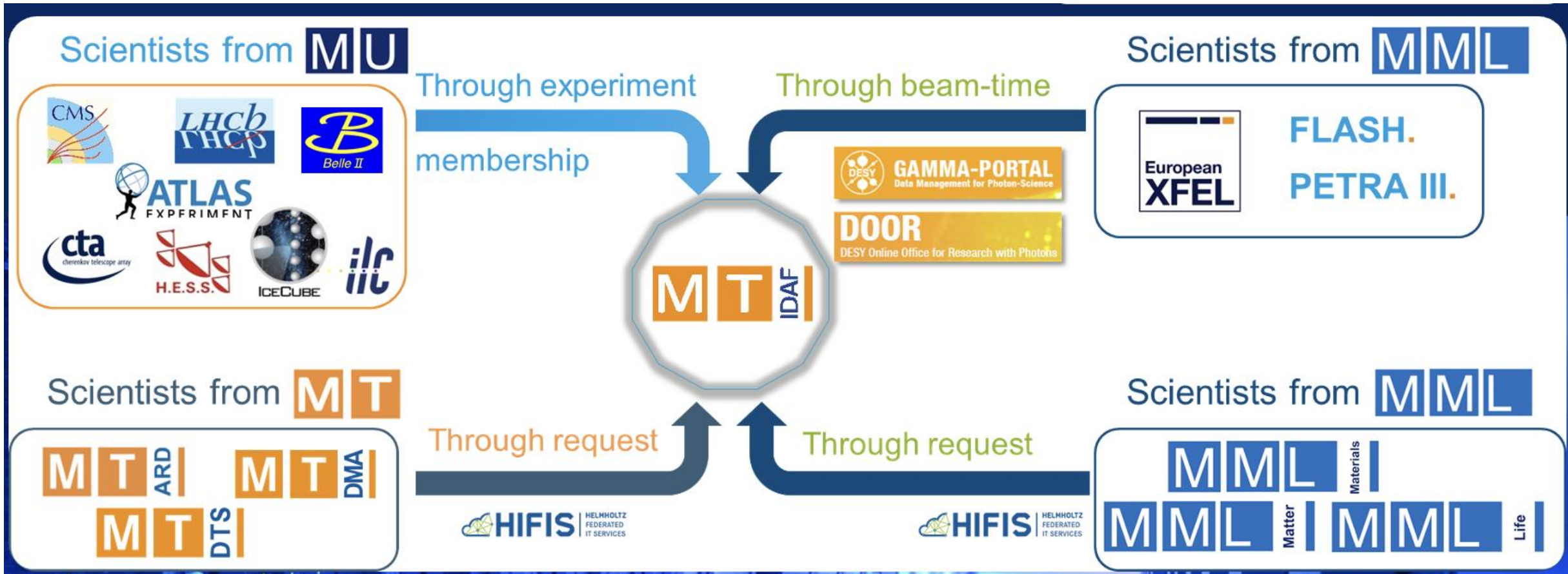


Only NAF & Maxwell logins are accounted for (no Grid submission)  
... mostly from academia (universities and institutes)  
... some commercial users

# ... who can access it and why?

## User access:

- Grid communities
- scientists from PETRA III, FLASH, EuXFEL
- Helmholtz Matter users



# The IDAF is about *services*

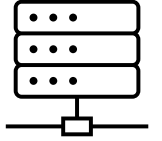
Interactive  
access and  
compute



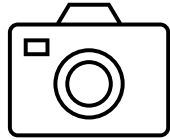
Federated  
access to  
compute &  
data



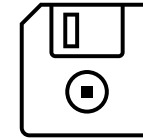
Large-scale  
compute and  
workflows



Integration  
with DAQ



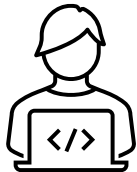
**Data  
storage &  
management**



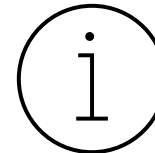
Metadata  
management



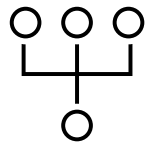
Software  
and  
containers



Documentation,  
support,  
training

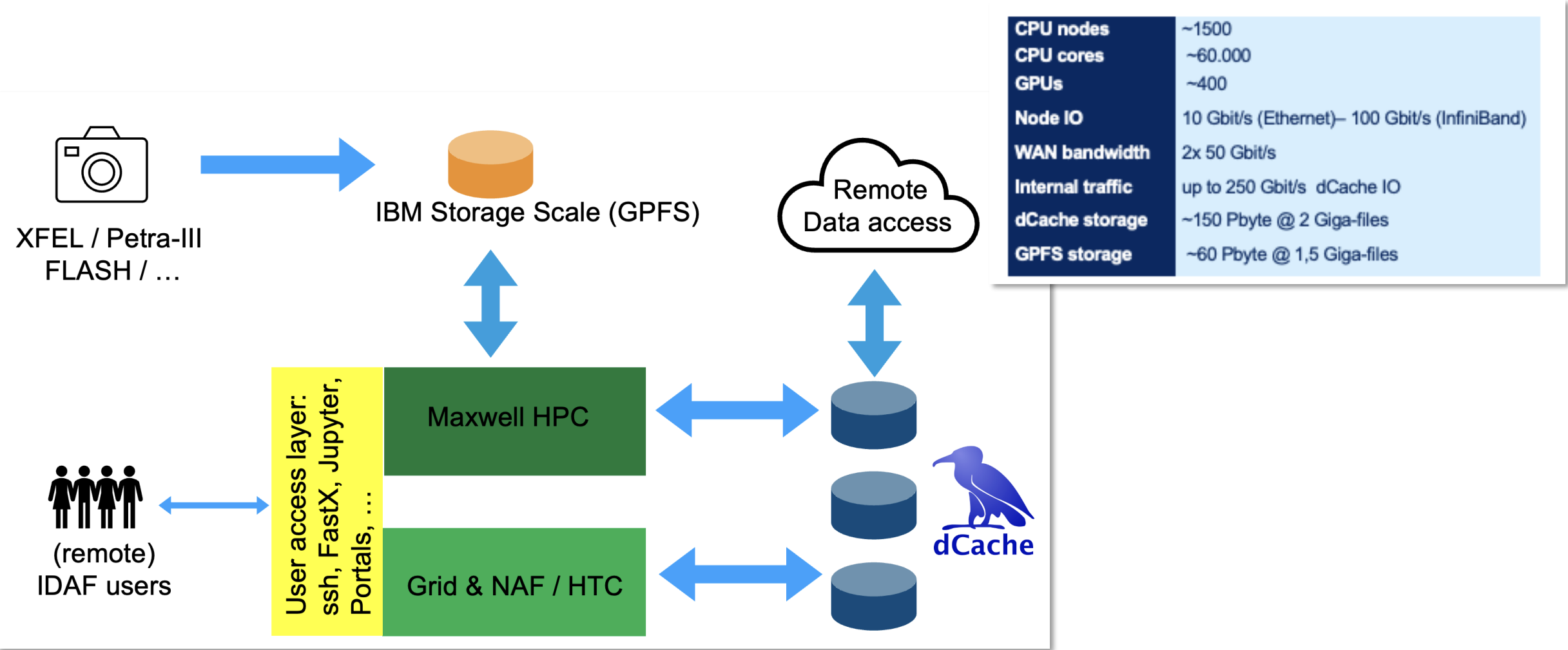


Code  
management  
and CI/CD





# ... and the IDAF is also about infrastructure



# Data Storage: Essential for Science@DESY

# Data Management

Today: Most Scientific Endeavours Produce Large Amounts of Data

- **Computing@DESY**: Storage of data for all departments and communities

## (Astro-)Particle Physics

- Store and archive raw data
- Store and archive simulated data
- Store pre-processed data for
  - Experiment specific workflows
  - Dedicated user analyses

## Accelerators and Detectors

- Store and archive simulated data
- Store and archive test-beam data
- Store and archive telemetry data

## Photon Science

- Store simulated data
- Store and archive raw data
- Store pre-processed data for analyses

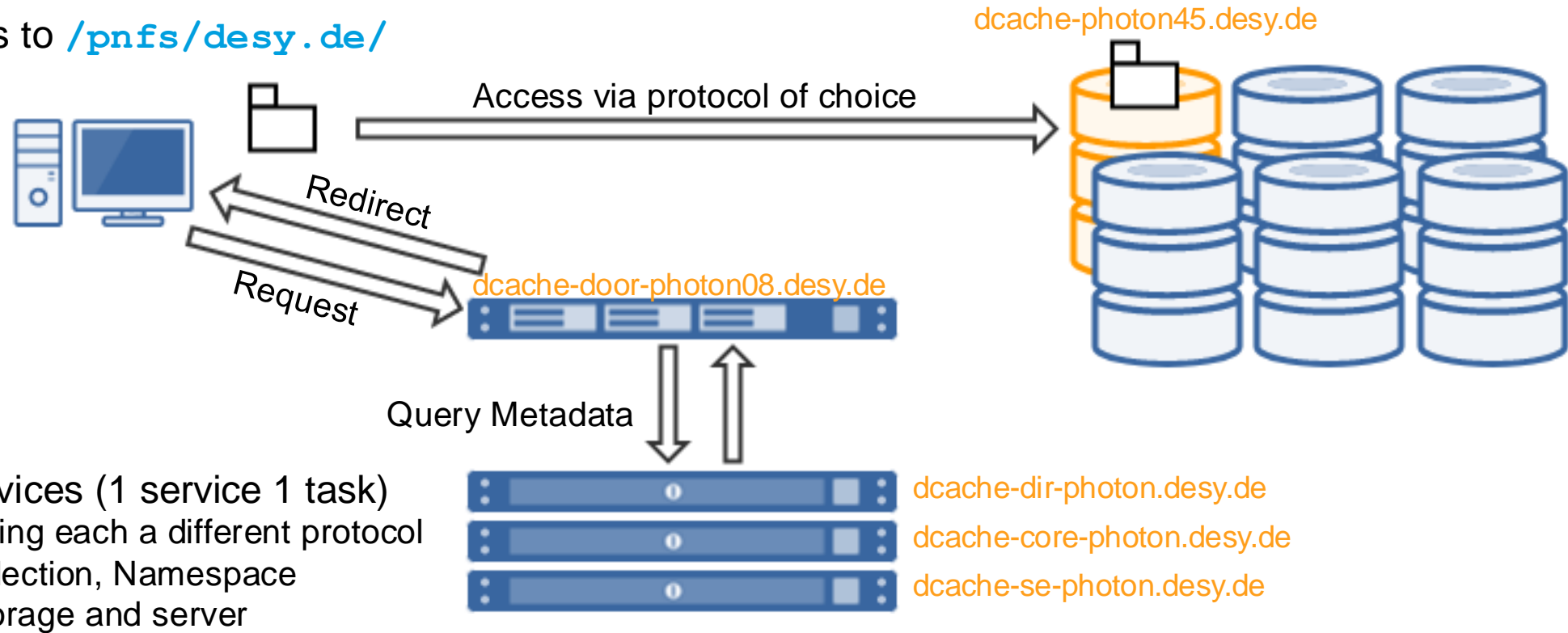
- Data as central element for most research
- Make data the central hub and **trigger** for scientific workflows



# dCache: Architecture

## User Access to dCache Responsible to Store Machine Data

- Use dCache: Access to [pnfs/desy.de/](https://pnfs.desy.de/)



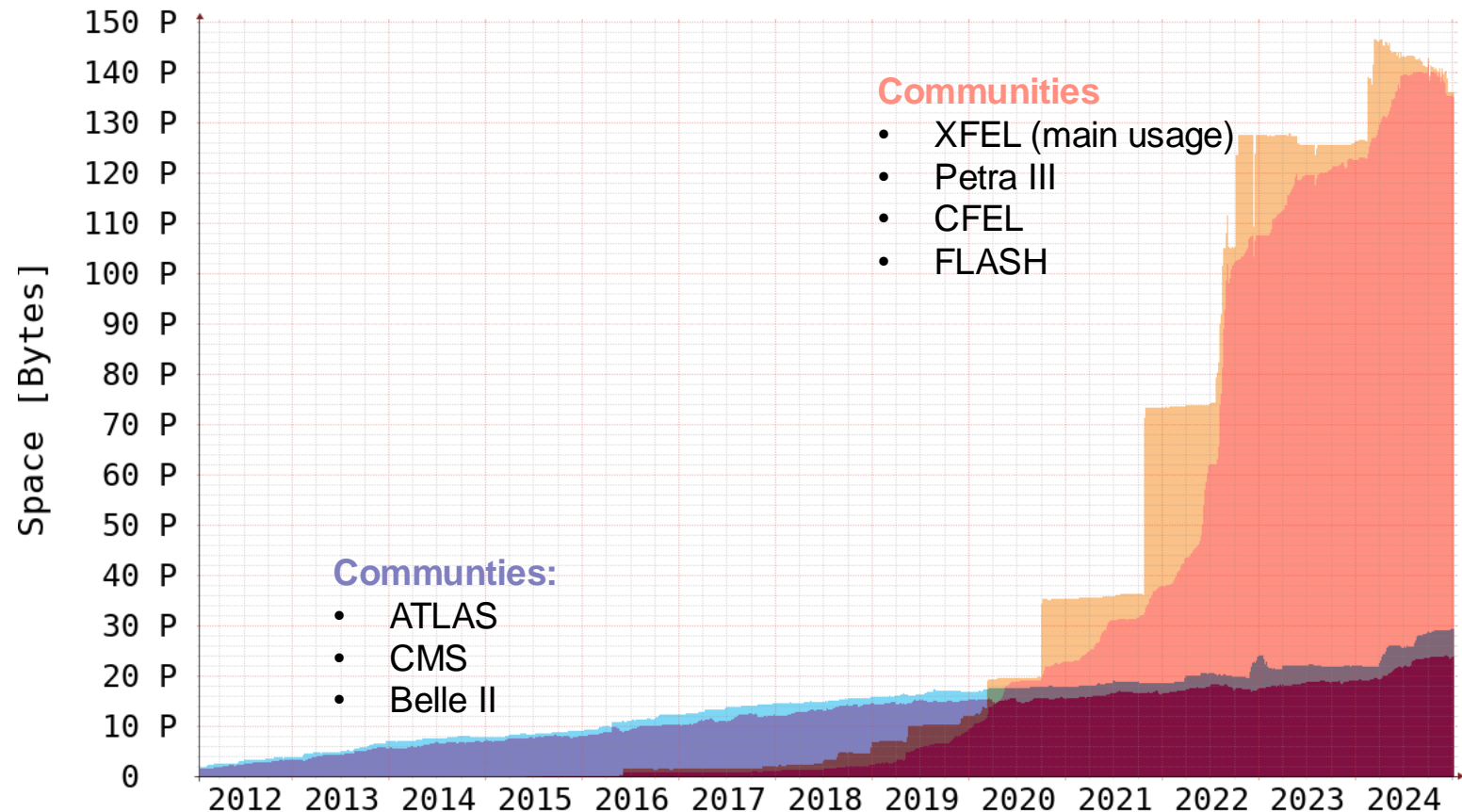
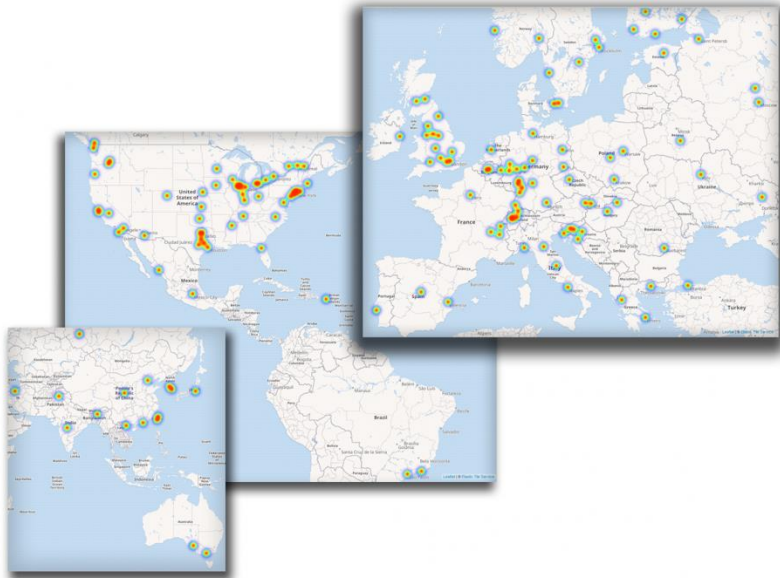
- Based on Micro-Services (1 service 1 task)
  - **Doors** – supporting each a different protocol
  - **Heads** – pool selection, Namespace
  - **Pools** – data storage and server
- dCache instances for Photon Science/Machine, European XFEL, ATLAS, CMS, Belle/ILC/DPHEP, Sync&Share
- Similar layout: three head-nodes, doors for requested protocols and pools nodes
- Scale-out horizontally: 10 pool nodes for Sync&share and 200 for European XFEL with 100 more ordered
- Scale-out horizontally: client always to connect to pools for transfer, no data access through doors



# dCache: Capacity of Local DESY Instances

## Available and Used dCache Storage

- Steady increase for HEP since inception of dCache
- Exponential increase for Photon science since start of European XFEL
- HEP dCache is connected to the WLCG
- Transfers all over the world
- Read access to HEP dCache (excluding DESY-HH):



# Complementing storage for DAQ and project use cases

## Cooperation with IBM on GPFS / Spectrum Scale

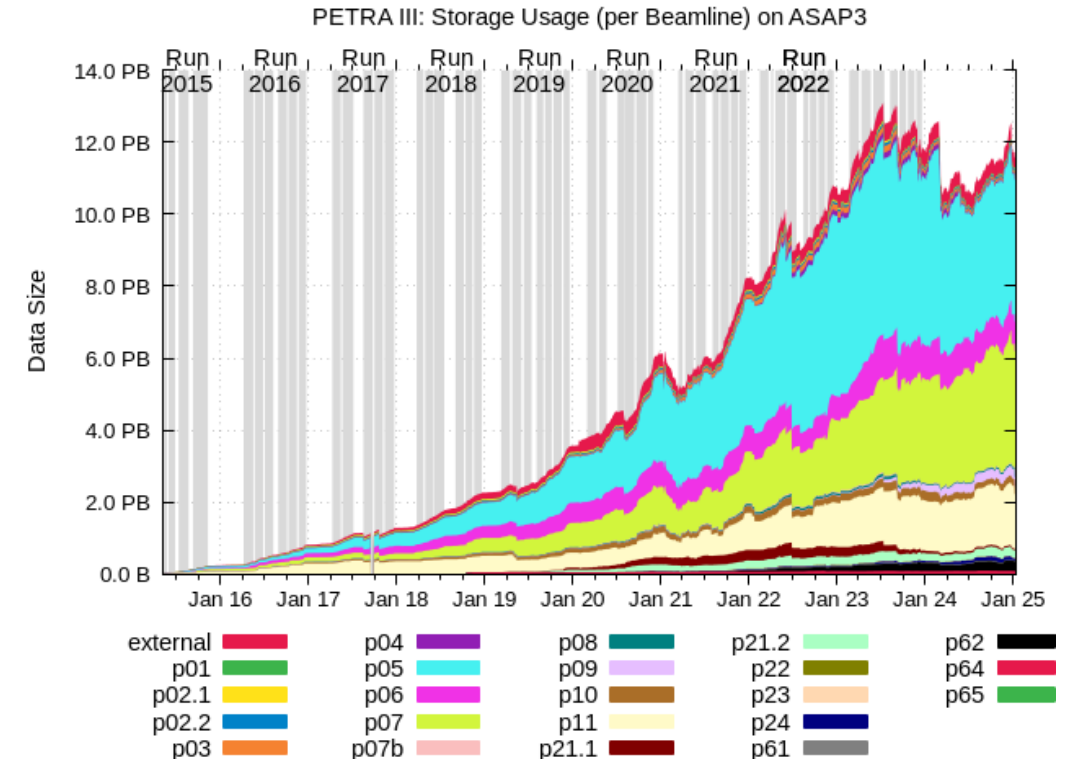
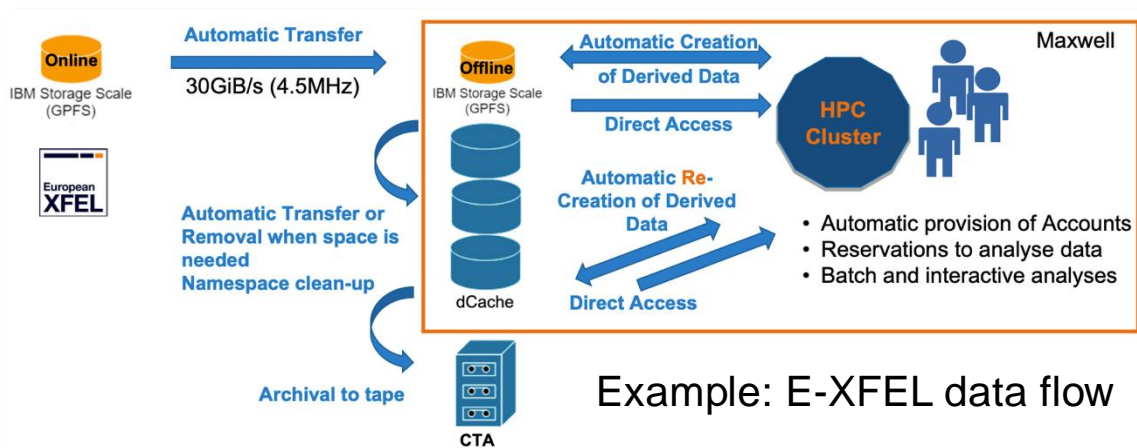
- dCache storage is optimized for throughput
- Several applications need different profiles:
- Photon science data taking:
  - Lots of small files, high bandwidth
  - Integrated (e.g.) into ASAP::O
- End user analysis
  - Single file performance, many small files, meta data heavy

→ Complementing using  
(previously: GPFS)



IBM **Spectrum Scale**

- ~17 PB (Petra-III), ~64 PB (E-XFEL), ~5 PB (project space)
- Collaboration with IBM



# 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
```



- Usually only option for photon science and accelerator R&D (and commercial software)
- Trend includes HEP despite remote read capabilities

- Poses the challenge of having uniform name-space across the IDAF

HPC

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

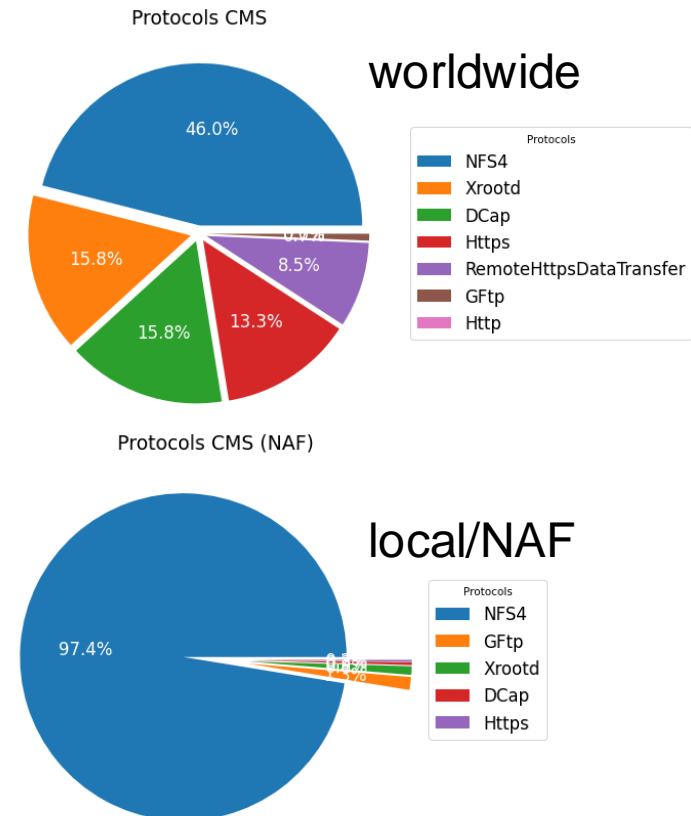
HTC

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

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

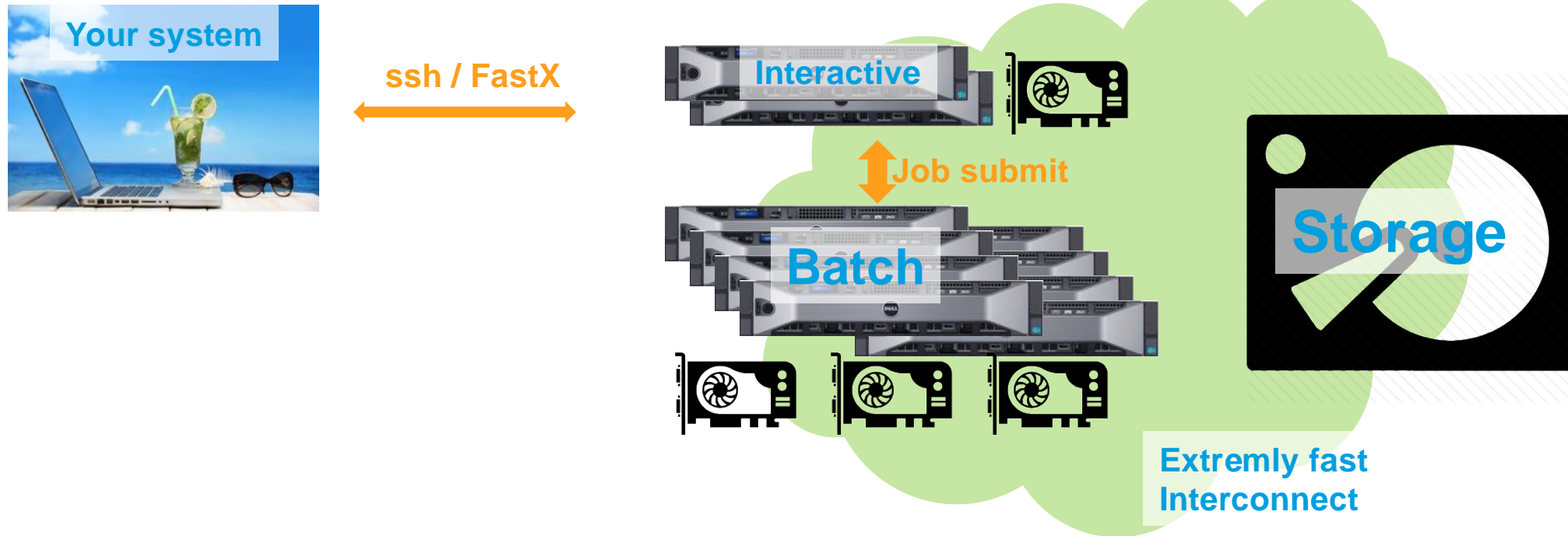
→ **SOLVED** now: in heroic act, we unified the user/project areas to one single namespace

Data Access CMS May 2023



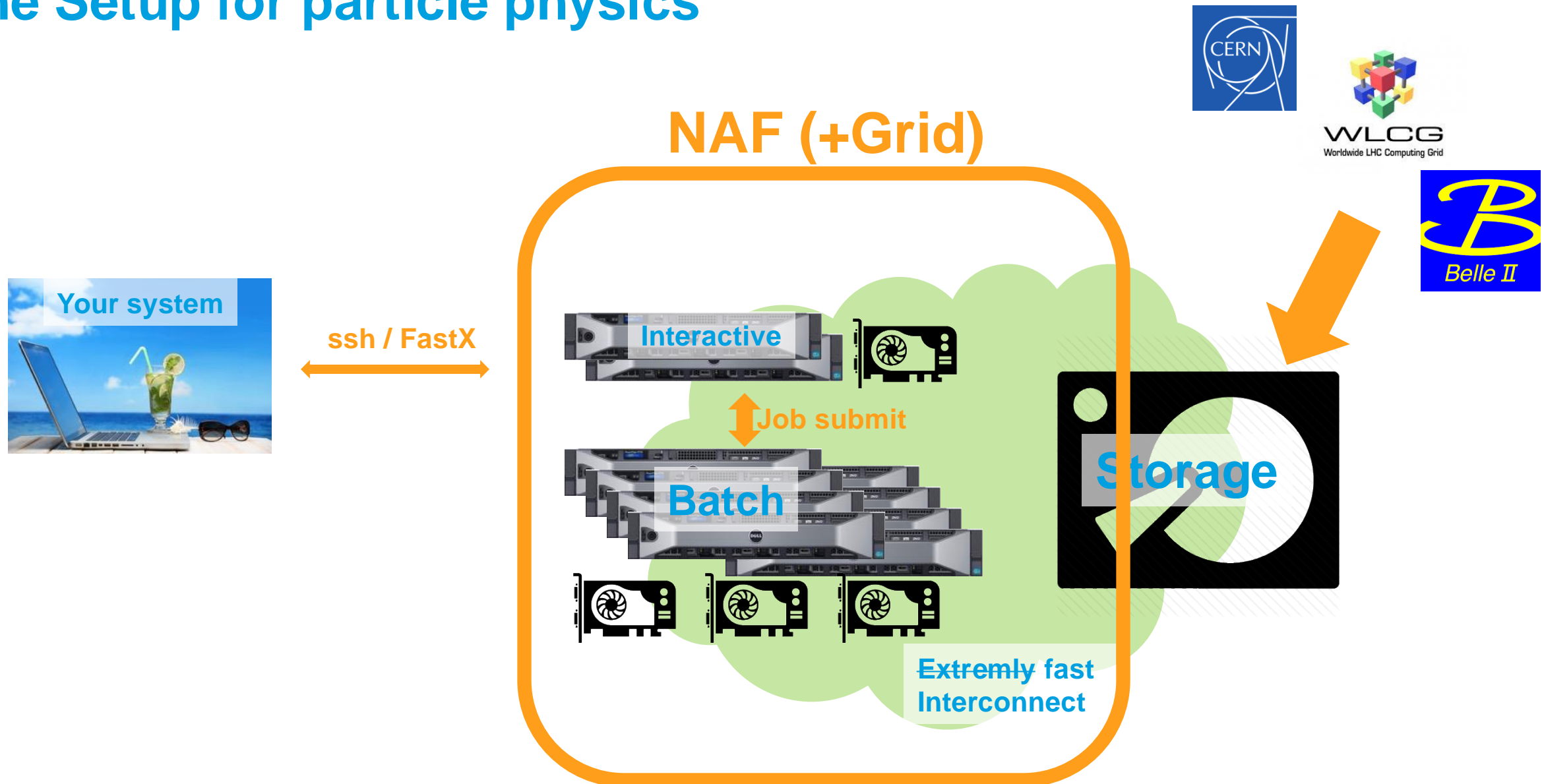
# Analysis & simulation infrastructure at DESY

# Basic setup at DESY

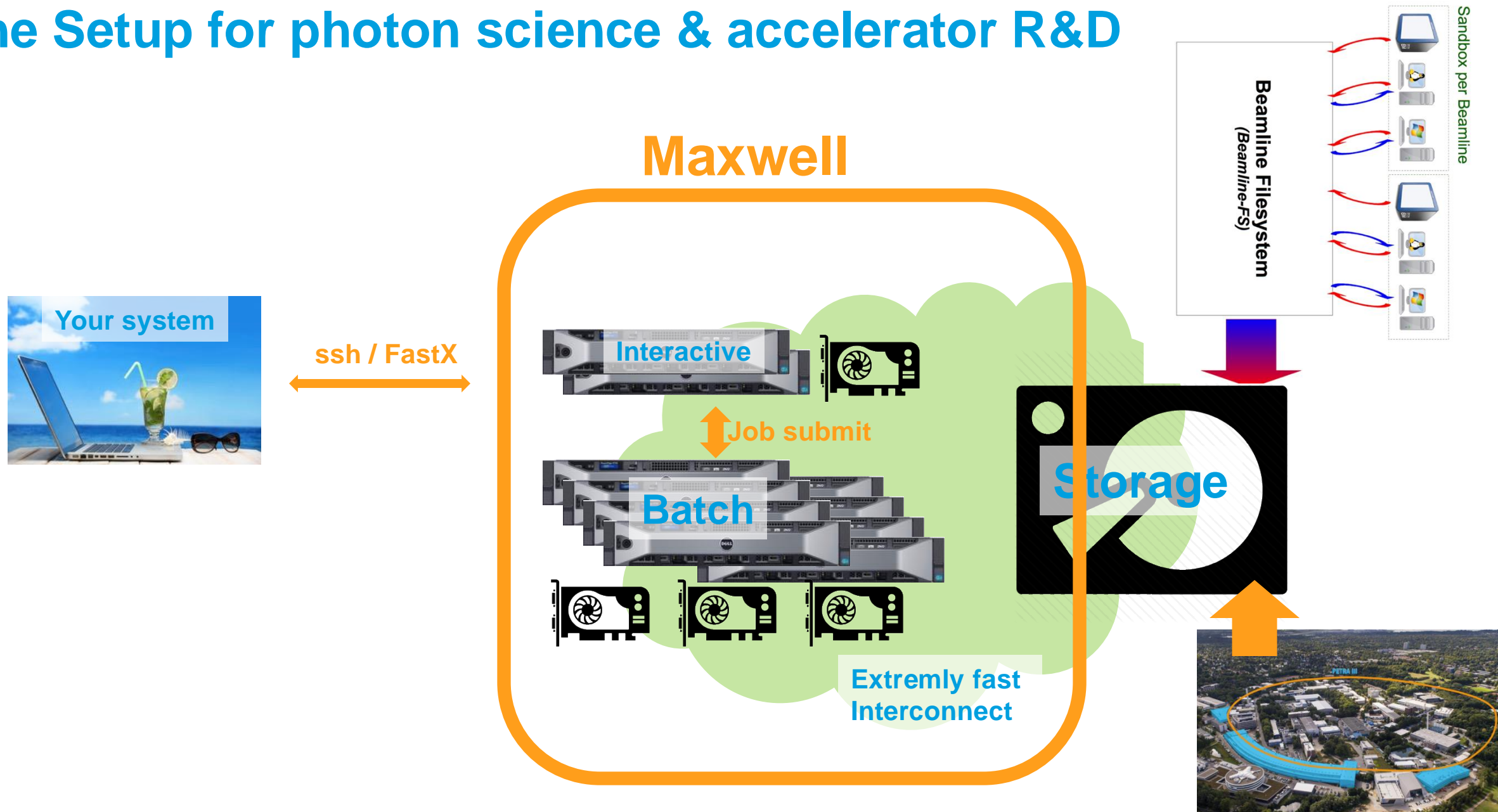




# The Setup for particle physics



# The Setup for photon science & accelerator R&D



# Architecture of a supercomputer

Number-crunching compute nodes + interconnect + file system

- **Compute node**

*Homogeneous within a partition of a supercomputer*

*Accelerated computing (Graphics Processing Unit)*

*More on that later*

**Maxwell:** - Homogeneous within partition: We try...  
- GPUs: Yes!

- **Interconnect**

*Invisible to the user (send a message)*

*No all-to-all connections*

*Multiple topologies (Fat Tree, Torus, Dragonfly)*

*InfiniBand is a widespread communication standard*

**Maxwell:** Using InfiniBand in a  
blocking fat tree topology

- **Parallel file system (I/O)**

*GPFS, Lustre*

**Maxwell: InfiniBand based storage:**  
- GPFS for \$HOME , P-3 and XFEL  
- BeeGFS as "project space"

- **Software**

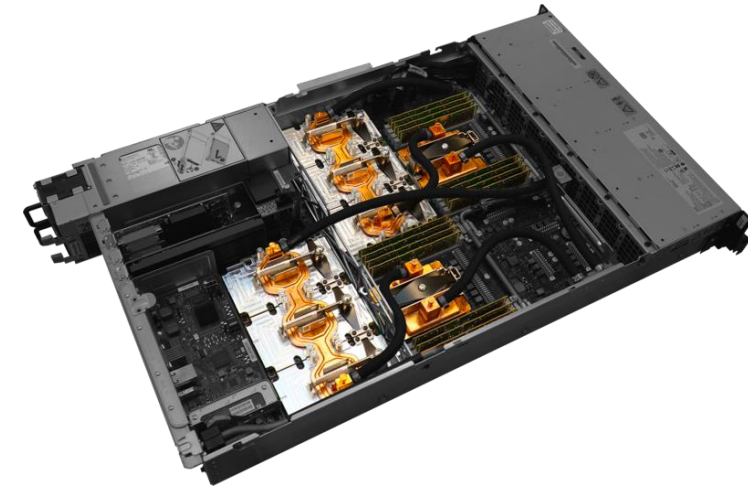
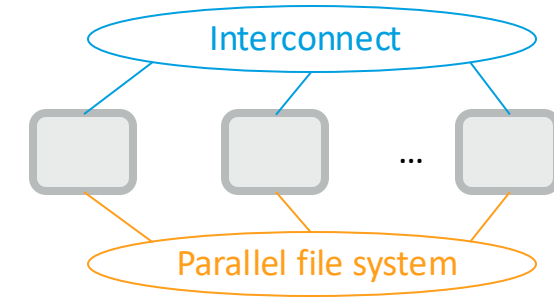
*Open-source, Linux-based*

*Job scheduler: Slurm, LSF*

*Launcher (resource allocation, placement): mpirun, srun*

**Maxwell: Slurm  
Supporting MPI**

**Many other applications available, incl.  
commercial ones**

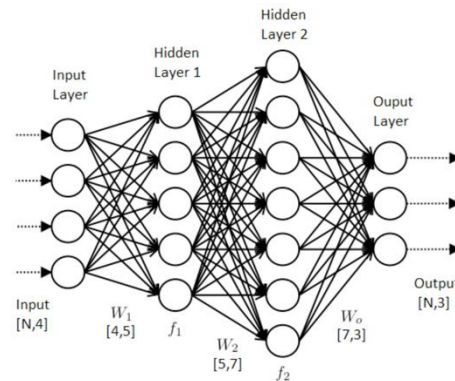
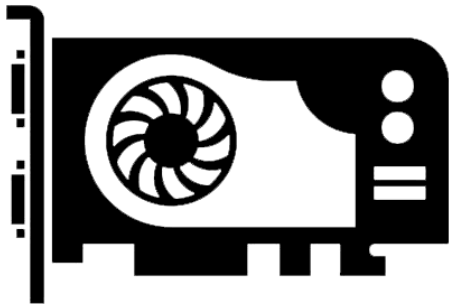


# Comparing Maxwell HPC & GRID/NAF HTC systems

Feature	Maxwell	GRID/NAF
Size	~950 nodes / 48k cores / 560 TB RAM ~200 nodes with GPU	~ 600 nodes / ~30k cores / ~100 TB RAM
Network	InfiniBand for fast data & IPC, 10 GE Ethernet	1 GE - 10 GE Ethernet
Storage	Access to GPFS data (IB), dCache (NFS, Ethernet).	dCache (NFS, Ethernet), GPFS (NFS, Ethernet)
Batch strategy	Whole/Multi-node-scheduling. Integration of private resources possible, with prioritized access.	Per-core-scheduling, no multi-node. Centrally procured resources. Fairshare on group basis.
Product	SLURM	HTCondor

# GPU computing & Machine Learning

- **General GPU computing** established in HPC systems
  - ... so in Maxwell: ~200 nodes equipped with GPUs
  - Different generations, different setups: From one GPU/Server to four GPU/Server
- Maxwell HPC cluster natural candidate for hosting GPU computing
  - Users have applications profiting from GPUs
  - GPUs benefit from “HPC-like” environment



- **Machine Learning**
  - Boosted by the usage of GPUs for training (and inference)
  - Benefits heavily from fast access to (large amounts of) data, and high-RAM machines
  - Maxwell is natural environment
- **Future** of GPU computing & Machine Learning
  - We see an increase in demand for “multi-GPU nodes” (~4 GPU/node)
    - Expensive, few nodes, challenging from scheduling point of view
  - Look for alternatives to NVIDIA. Have some examples in the lab. Dependency from CUDA challenging



# Making batch more user-friendly – and maybe overcome it?

- Select a good scheduler ... With active developer
- Containers healing the OS & software incompatibilities
  - Started on Maxwell in 2016, using Docker technology
- Interactivity & access: Jupyter
  - Integrate interactivity into batch
  - „Tragedy of the commons“
- git based workflows & CI/CD
- And ever and ever again, do training, taking by the hand, ...



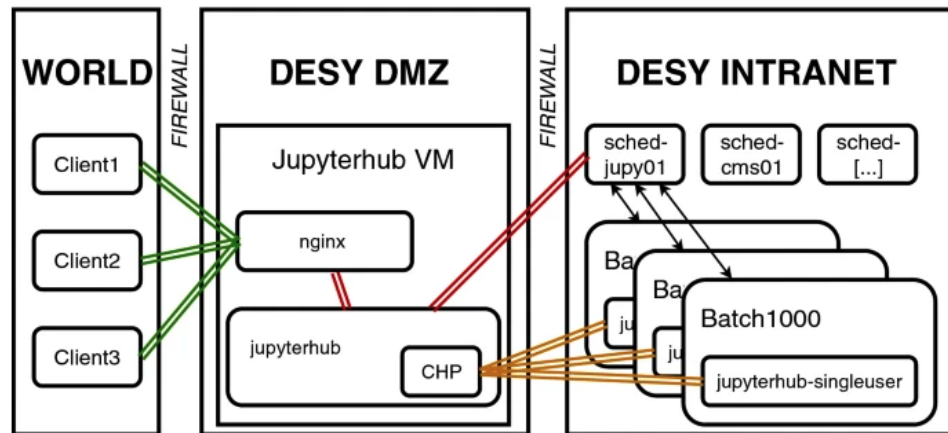
# Jupyter: Interactive & easy remote access

Jupyter notebooks and Maxwell and NAF



**What are Jupyter Notebooks? Data analysis and simulation in your browser**

- Python based interpreter for Python, Matlab, ...
- Access via web-browser through portal
- Computation itself happens on Maxwell or NAF: Integration with SLURM / HTCondor scheduler



## Maxwell Jupyter Job Options

Maxwell partitions ⓘ

Choice of GPU ⓘ

**Note:** For partitions without GPUs (or choice of GPUs) the GPU selection will be set to 'none'

Constraints ⓘ

**Note:** This will override GPU selections!

Number of Nodes ⓘ

**Note:** Number of nodes will be set to 1 on shared jhub partition!

Job duration ⓘ

**Note:** on the shared Jupyter partition (jhub) the time limit is always 7 days!

Launch modus ⓘ

Remote Notebook ⓘ

Upload New

Notebook:

Bash  
Matlab R2018b  
Python 3  
Python [conda env:Spyder]  
Python [conda env:Tensorflow-GPU]  
Python [conda env:Tensorflow2]  
Python [conda env:pyFAI]  
Python [conda env:pytorch]  
Python [conda env:tomopy]  
Pytorch  
Tensorflow-GPU

Other:


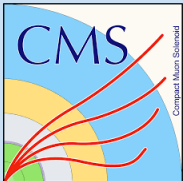






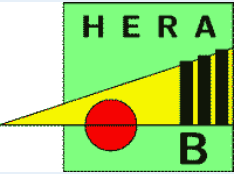

Text File  
Folder  
Terminal

Interactive analysis notebooks on DESY batch resources  
Johannes Reppin et al. (2021) , CSBS/Springer

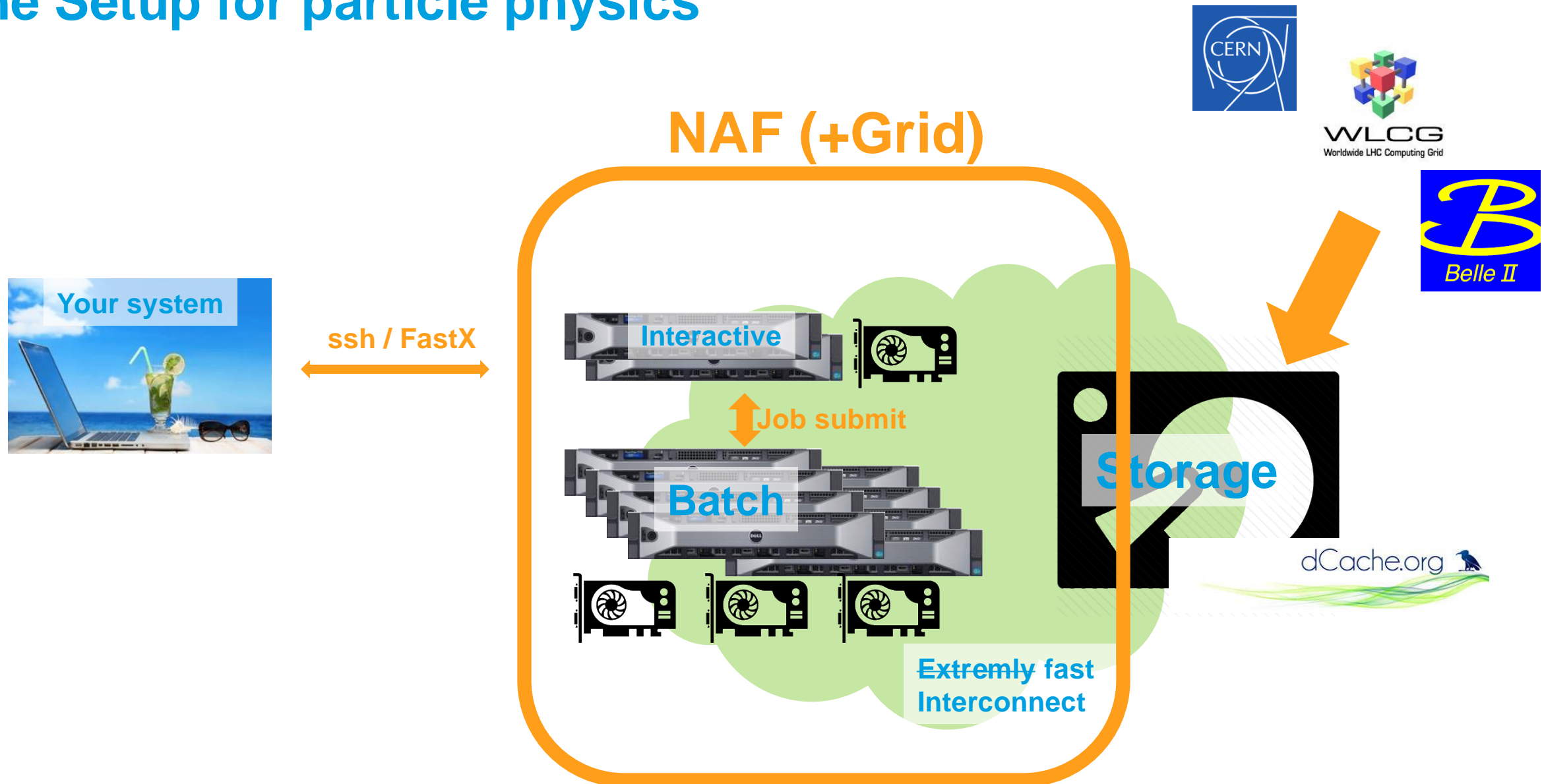
DESY. | IDAF @DESY | Christian Voss, Yves Kemp, PunchLunch 2025-01-16

# A view to particle physics analysis

# HEP communities at DESY

Community / Experiments	Compute activities
  	Grid Tier-2, German NAF users
	Compute & Storage, Management services, Collaborative tools, ...
 	Compute & Storage, Management services
   	Compute & Storage, Management services

# The Setup for particle physics



# Analysis Facility Evolution



IDAF integrated into global experiments and global workflows  
Yet, the final steps of the analysis happen locally, at one place: The Analysis Facility

How should an Analysis Facility look like in the future?

Discussions and proposals going on at different levels:  
e.g. FIDIUM, HSF, WLCG



Some ideas, e.g.

- Federate users, access and portals
- Make federated data available



New tools, e.g.

- Gitlab based pipelines
- Columnar analysis



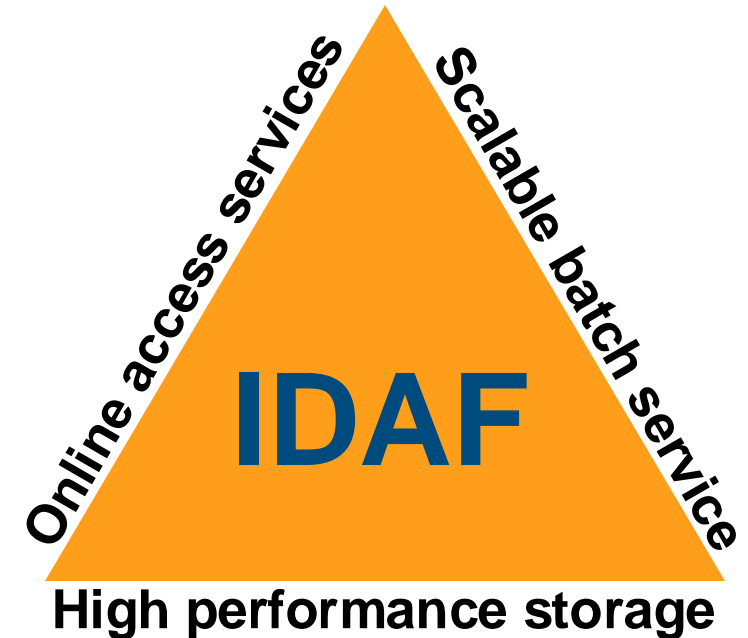
# HEP and Batch?

- Batch based computing Ansatz long established in HEP
- Nevertheless: Alternatives are being investigated:
- NAF: Augment Grid with interactive resources
- Jupyter as new access method is being rolled out successively
  - Investigation on Jupyter resource scaling
  - „Tragedy of the commons“
- Investigating DASK & Spark as non-batch compute organization
- gitlab / CI/CD workflows ... connection to batch?
- Batch ↔ Cloud integration ?



# What is an Analysis Facility?

- Basic concept of the IDAF are:
  - Data locality
  - Access services and compute integrated with storage
  - Present a holistic service to the analyst
- With the advent of machine learning and new compute technologies and storage concepts:
  - Does data locality still hold?
  - Is current storage integration still OK?
  - Do people need an integrated service, or rather flexible infrastructure?



# Discussions going on at different levels

- Sometimes user driven ... who drives?
  - 10% pioneer users?
  - 10% special requirements users?
  - 80% normal users?
- Research at facilities needed:
  - E.g. PhD students / young postdocs that spend some time doing their analysis in a novel way – in close interaction with IDAF experts
  - Participation in ERUM data call on

THE HEP SOFTWARE FOUNDATION (HSF)

HSF-TN-2024-01  
April 2024

## Analysis Facilities White Paper

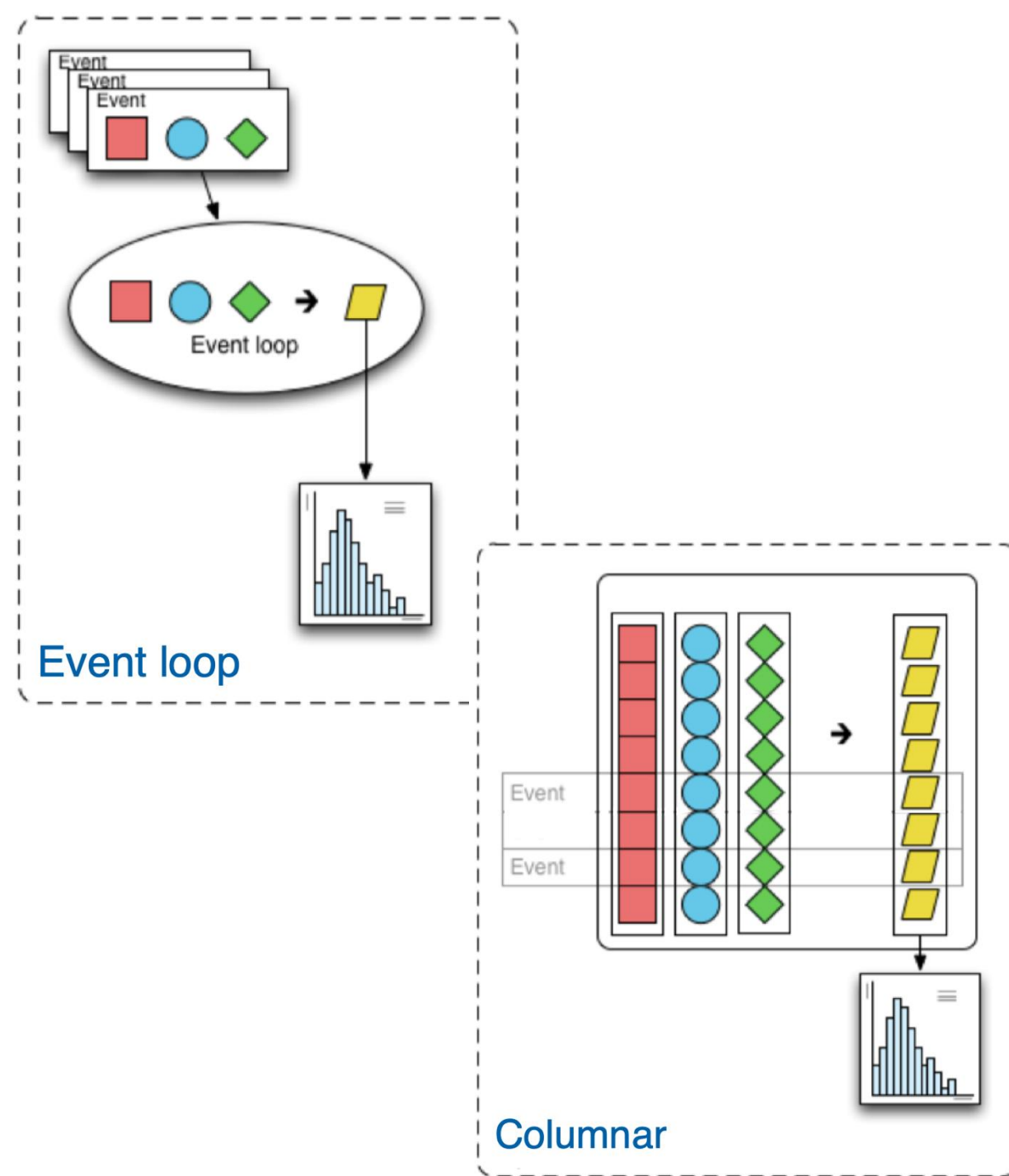
### Analysis Facilities Workshop

 Jun 18, 2024, 2:00 PM → Jun 20, 2024, 4:00 PM Europe/Berlin

 MIAPbP (Garching)

# Columnar Analysis

- Orthogonal ansatz to compute
- Optimal with data stored in-memory
- Poses radically new challenges
- IDAF has pioneered columnar analysis on small scale
- Currently seeing that users are taking up this techniques
  - > 50% in some groups
- **Need further R&D to offer CA reliably and performing, at-scale, experiment-generic, in a multi-user environment**

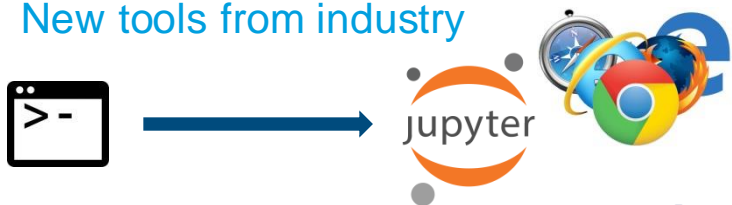


# Future: Following Changes in Communities

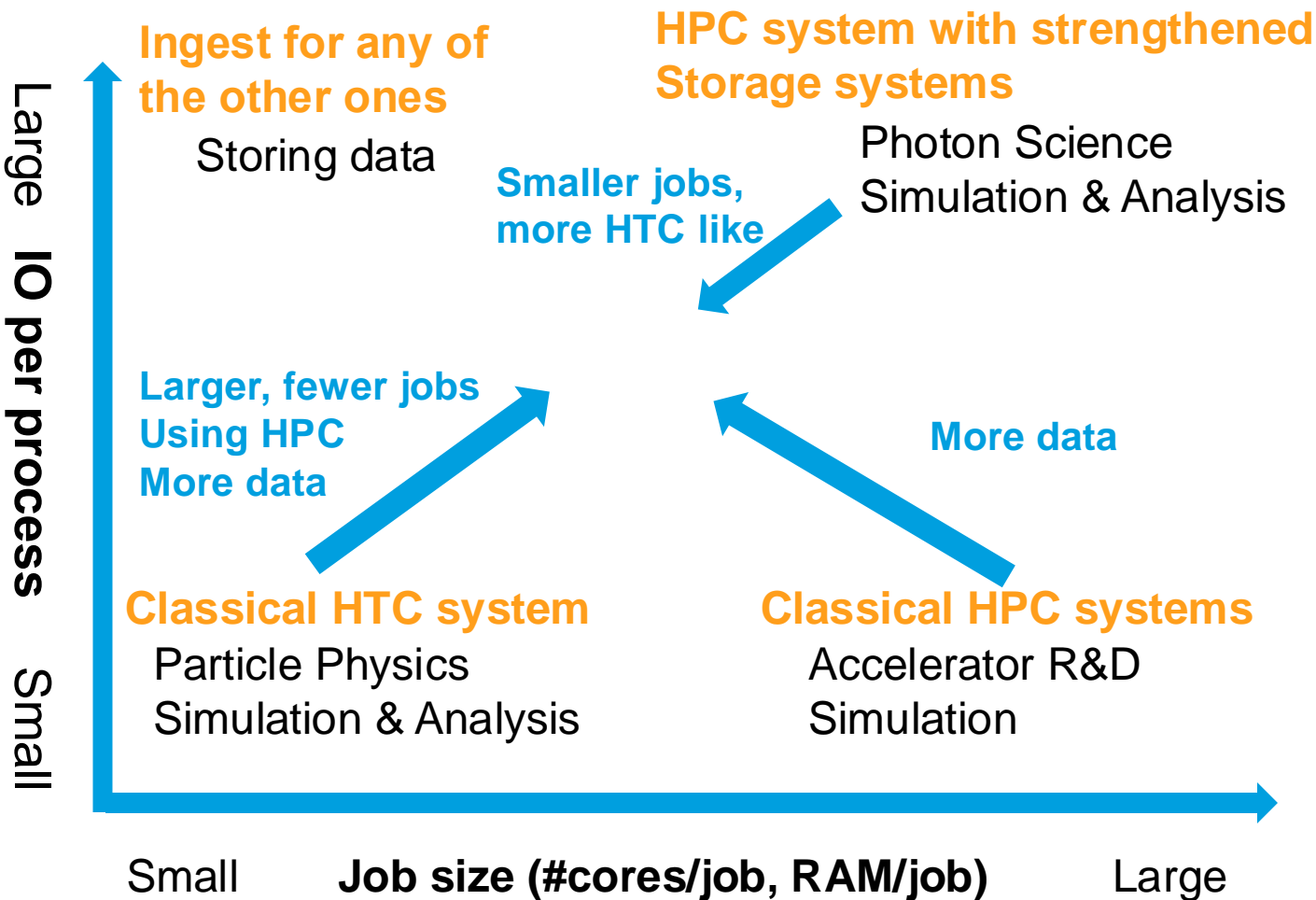
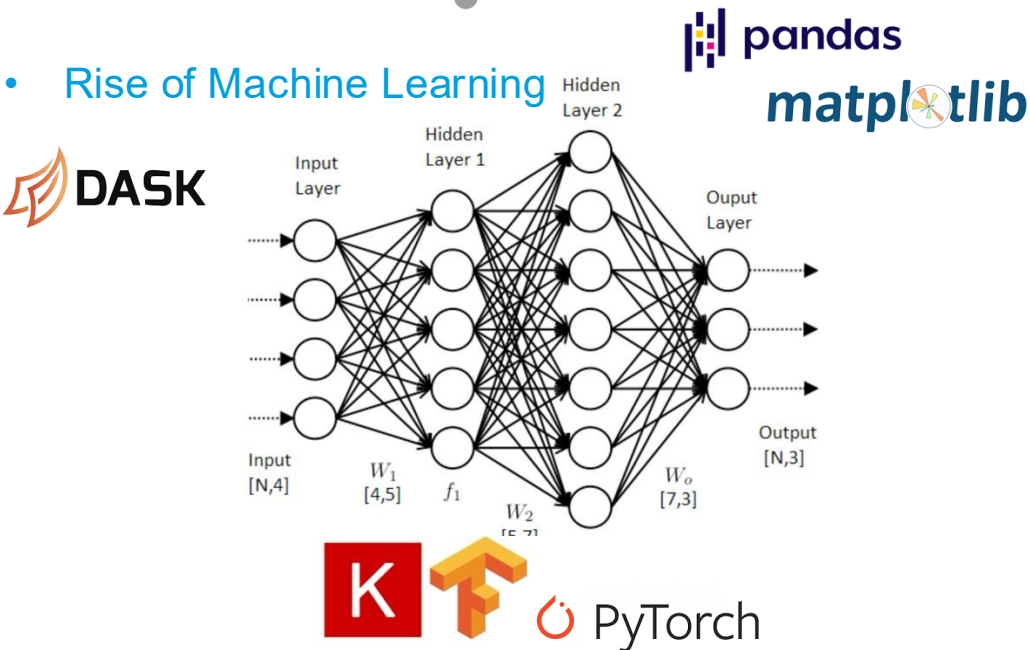
New Tools and Workflows Bring Communities Closer Together

Revisit diagram from beginning

- New tools from industry



- Rise of Machine Learning



# Summary & Outlook

- Interdisciplinary**    DESY IT already serves all branches of Science at DESY
- Infrastructures are there, and working well ...
- Data**    Science produces large amount of data
- Detectors, Accelerators and Simulation
- Analysis**    Main goal is to provide best possible analysis infrastructure for all our users.
- Large scale offline, and fast online ... overcome online/offline barrier for analysis
- Facility**    Not an institute cluster: *Facility* for internal *and* external users
- Full service for entire data lifecycle