

Photon Science and HEP at DESY

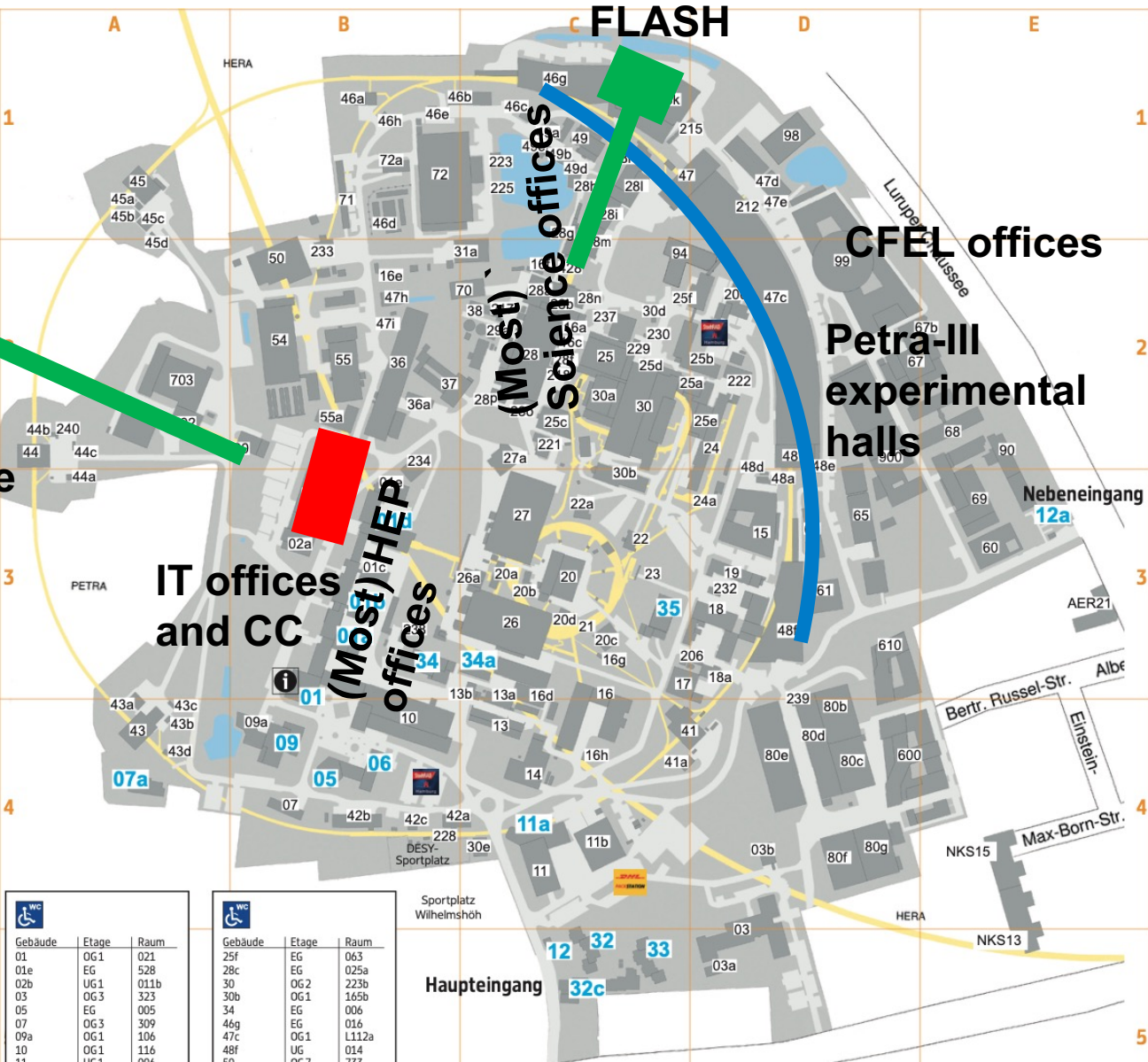
Christian Voß & Yves Kemp ... with lots of slides from other people

14.12.2023

Where is What?

European XFEL
3 km ... not to scale

... indicative only



Gebäude	Etage	Raum
01	OG1	021
01e	EG	528
02b	UG1	011b
03	OG3	323
05	EG	005
07	OG3	309
09a	OG1	106
10	OG1	116

Gebäude	Etage	Raum
25f	EG	063
28c	EG	025a
30	OG2	223b
30b	OG1	165b
34	EG	006
46g	EG	016
47c	OG1	L112a
48f	UG	014

Who is Who? – people involved in operations

On the experiment side

- Petra-III & FS division:
 - Beamline scientists
 - FS-EC (Experimental control)
 - FS-SC (Scientific computing)
- XFEL
 - IT&Data Mgmt group
 - “Data science teams”

On the IT side

- ASAP3 (Petra-III data taking) & XFEL GPFS team
- dCache operations team
- Maxwell HPC team
- + general IT services

(~Weekly) operations meetings between

- IT \leftrightarrow FS-EC
- IT \leftrightarrow XFEL-“DAQ ops team:”

and other topical meetings

Interdisciplinary Data and Analysis Facility

Supported Communities

- Accelerator Data

FLASH.

Free-Electron Laser FLASH



FF ▶▶

- Accelerator Development Data



- HPC simulations

- Test-beam data

Detector and
Accelerator R&D

- Facility User Data



PETRA III
FLASH.

Free-Electron Laser FLASH

- Data of external Partners



CSSB
Centre for Structural
Systems Biology

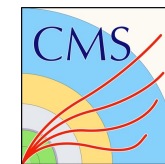
EMBL



Helmholtz-Zentrum
hereon

Research with
Photons

- Particle Physics Data



ALPS II



- Astro-Particle Data



Astro- Particle Physics

Interdisciplinary Data and Analysis Facility (IDAF)

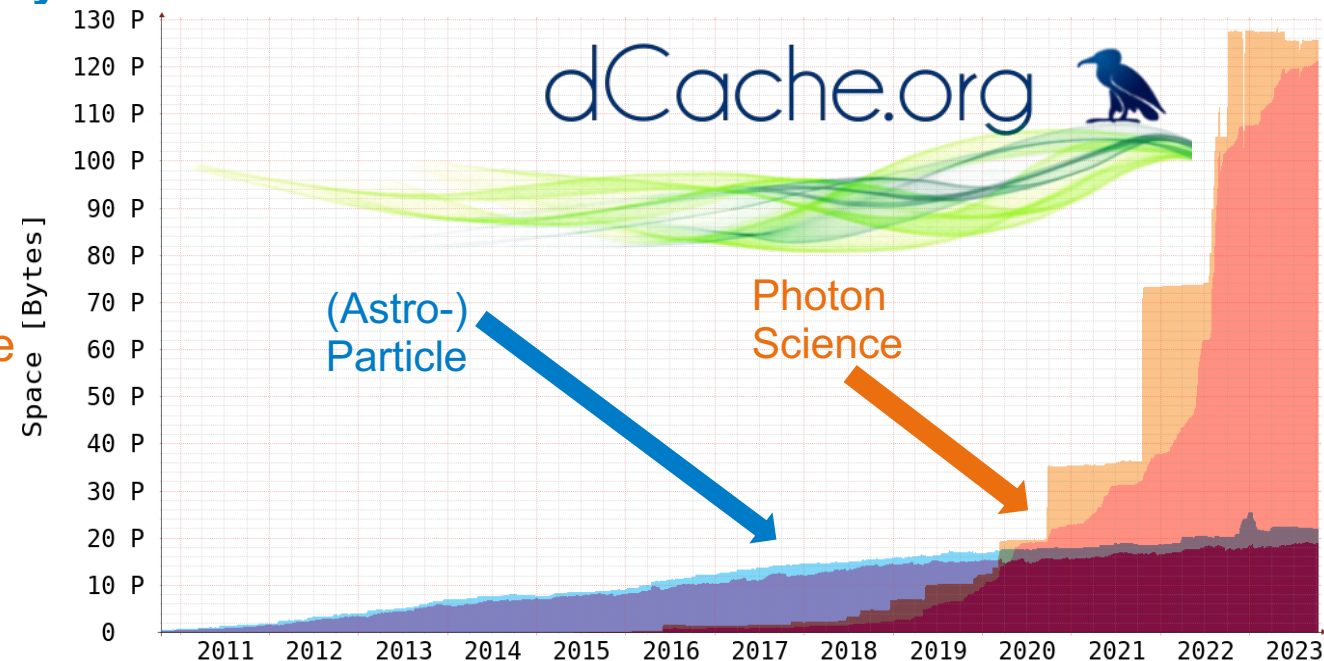
Origins and Overview

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

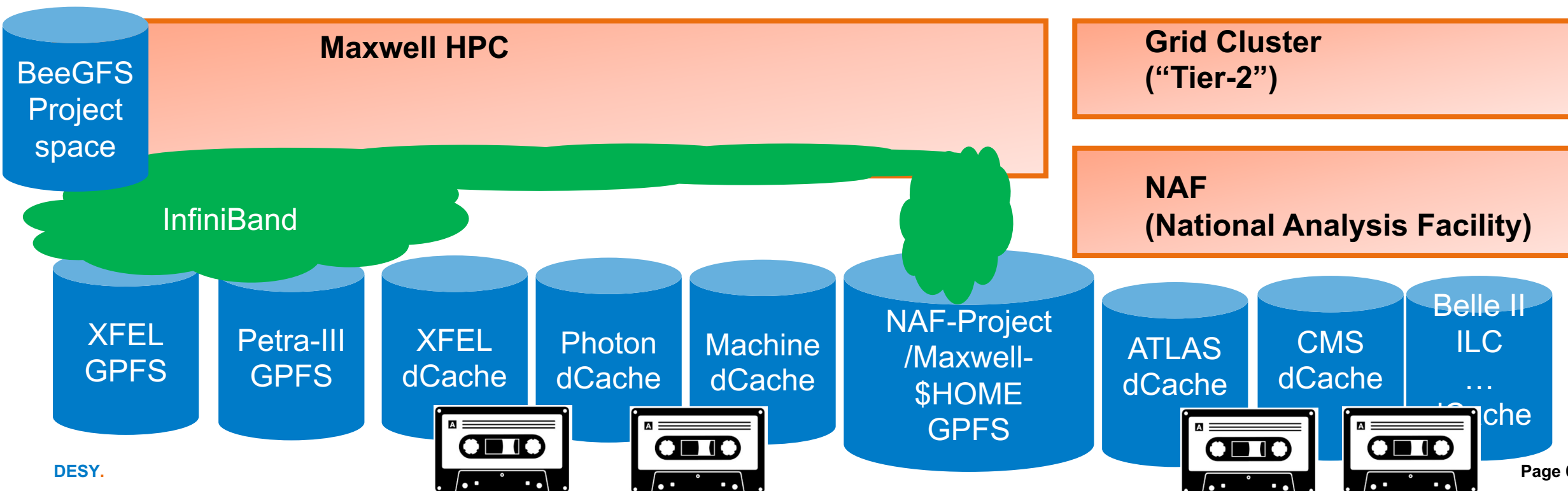
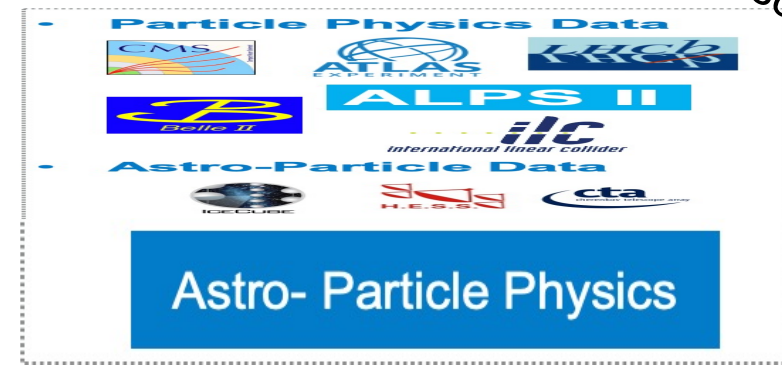
- Large photon science user facilities
- Large local particle physics groups
- Obvious when looking at provided and used storage



Interdisciplinary Data and Analysis Facility (IDAF)

... connecting the facilities to the communities

Blocks / Sizes not to scale!



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 



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

Maxwell HPC Cluster

What is the Maxwell cluster?



- **Lots of computers**
 - a variety of different models (typ. AMDs or Intel; different generations)
 - all equipped with 256GB up to 1.5TB+ of memory per node
 - all connected to Petra3 GPFS storage (and CFEL, EXFEL, CSSB storage)
 - quite a number of nodes with 1 to 4 GPUs (usually NVIDIA, different models like P100, V100, A100)
- **"Vast" amount of storage**
 - a variety of different options
- **Powerful network**
 - low latency, fast InfiniBand (IB)
 - good 10G ethernet
- **Lots of software**
 - never up-to-date ☹️
- **Quite a number of services**
 - not always what you'd wish for



What we got



What we sell

Maxwell HPC Cluster

What is the Maxwell cluster?



- **Main purpose**

- High Performance Computing
- Offline Data Analysis
- Simulations of all kind
- Remote Visualization
- Any application which can make use of the special features of Maxwell!

E.g. Ansys, Comsol, Fdmnes (MPI version), Matlab, OpenFOAM, Orca, Quantum espresso, Tensorflow, Xds, Xmimsim, XRT

- **Intention of the Maxwell cluster**

- Bring in group resources to become part of the cluster
 - Exclusive resources usage for jobs managed by SLURM
 - Efficient resource usage (batch queue, resource definitions, optimize costs etc.)

E.g. Conuss less well suited (single threaded/few mem.) and thus likely runs faster on office PCs given higher CPU core clock

- **All jobs are scheduled by the SLURM scheduler (via submission hosts)!**

- Usually jobs don't have to wait very long
- But it depends on the jobs requirements
- and there is no VIP fast lane ...

Maxwell HPC Cluster

in numbers

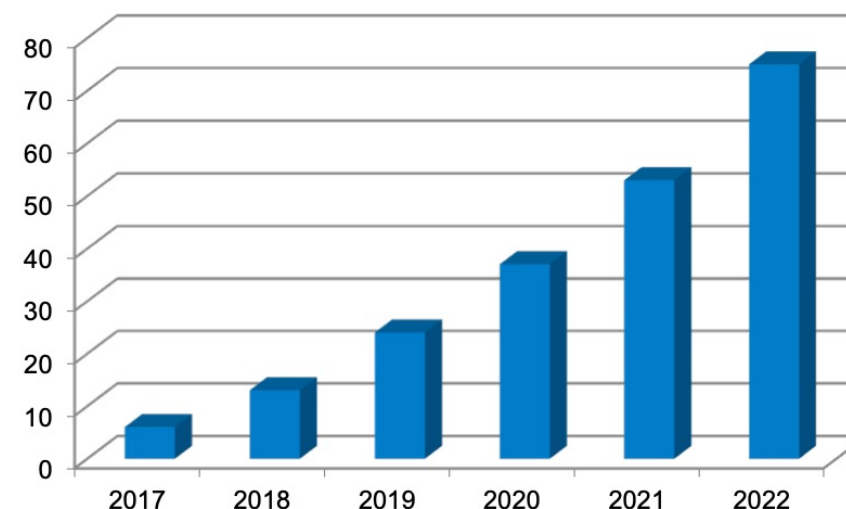


Nodes	800
Phys. CPU cores	32000
Memory	440 TB
CPU peak performance	1100 TFlops
GPU Nodes	180
GPUs	340
GPU peak performance	2250 Tflops
Total peak performance	3350 Tflops

Infiniband	
Infiniband switches	~60
Infiniband cables	~1600
Infiniband length	~8km
Storage	
IBM Spectrum Scale	4 instances with ~60PB
BeeGFS	1 instances with ~1.6PB
dCache	long term storage

Maxwell Users	~2500
Concurrent JupyterHub Users	up to 200
Concurrent interactive Users	up to 450
Jobs	
Number of batch jobs in 2022	3.913.869
Number of jupyter jobs 2022	38.949 (1%)

Citations/Acknowledgements per year



Maxwell HPC Cluster

Intention and organisation



- Bring in “group” / own resources to become part of the cluster
 - Exclusive resources usage for jobs managed by SLURM
 - Efficient resource usage (batch queue, resource definitions, optimize costs etc.)
 - access to / option to use resources brought/owned by others
- Homogeneous/common environment for ‘all groups’, e.g. rules, GPFS, software
- Organised in (slurm) partitions which can be/are tuned to meet the individual requirements
- A partition typ. ‘consists’ of
 - [a list] of compute nodes which can be used / are assigned for a compute job
 - A set features and boundaries/constraints
e.g. CPU Types, CPU Gen, amount of RAM, avail. GPU(s), num. of jobs, job duration
- Consequence is sophisticated setup
 - More than 40 partitions & more to come
 - Multiple overlapping partitions
 - Partitions with and without over-subscription; with and without pre-emption; with and without re-queuing; with all kind of different limits; ...
 - Heterogenic hardware
 - In general, **group partitions are configured to be available for all Maxwell users** (“-p all”), but **prioritising group members for their partitions, removing all "hostile" jobs** within an “adjustable period of time”
(*application/script has to catch first signal and then has Δt to end in defined manner*)

Grid cluster and NAF cluster in a nutshell

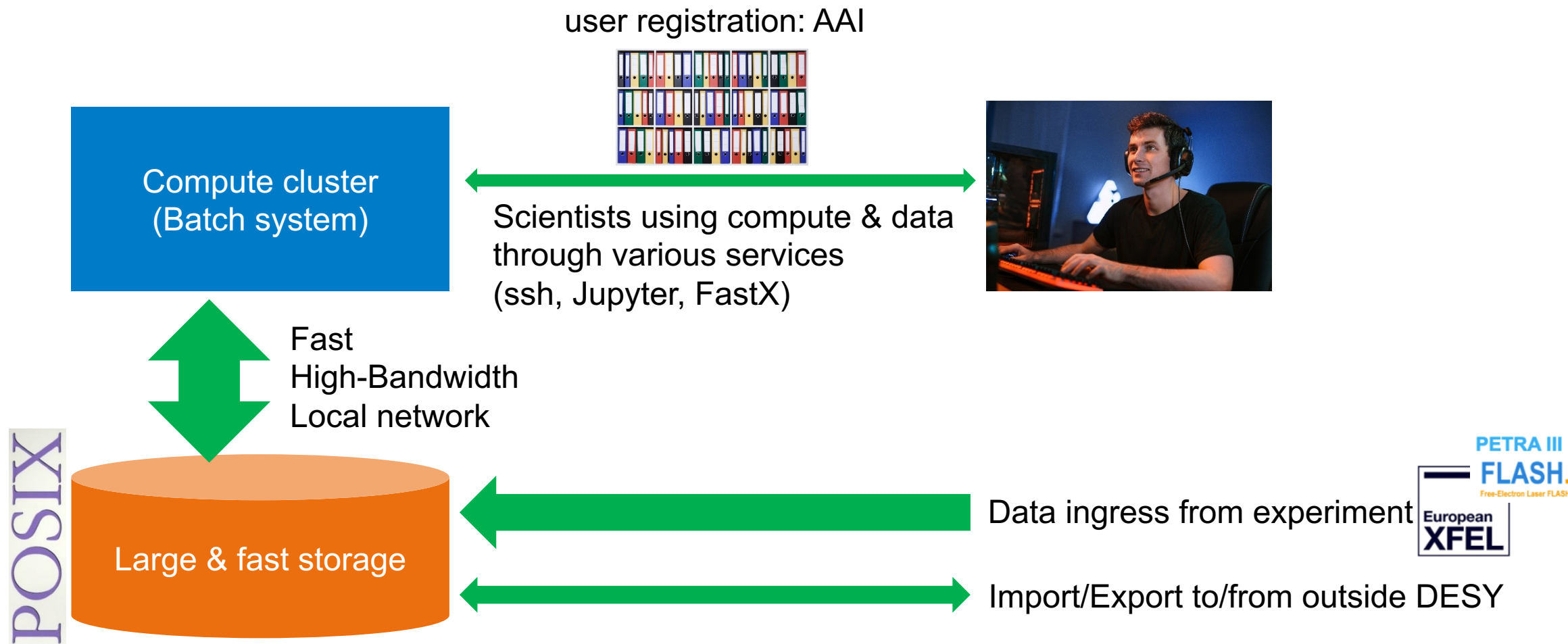
Grid cluster

- ~350 nodes, optimized for high throughput
- ~20k hyperthreaded cores, ~> 2GB RAM/core
- centrally purchased by IT, fairshare according to WLCG pledges (no partitions!)
- 10 GE (and some legacy 1 GE)
- Access to dCache, all protocols
- Behind CondorCE, governed by HTCondor
- Pool accounts (well, pilot jobs...)
 - Management by HTCondor group, 99% overlapping configuration
 - Users are mostly from HEP experiments
 - Per-core scheduling. Multicore possible. No multi-node scheduling!

NAF cluster

- ~300 nodes, O(10) GPU, optimized for fast turnaround
- ~9k physical cores, ~> 3-4 GB RAM/cores
- centrally purchased by IT, fairshare according to NAF experiments (no partitions!)
- 10 GE (and some legacy 1 GE)
- Access to dCache, all protocols
- Governed by HTCondor, access via SSH (WGS), Jupyter, FastX
- DESY accounts

Concept of clusters for data driven science:

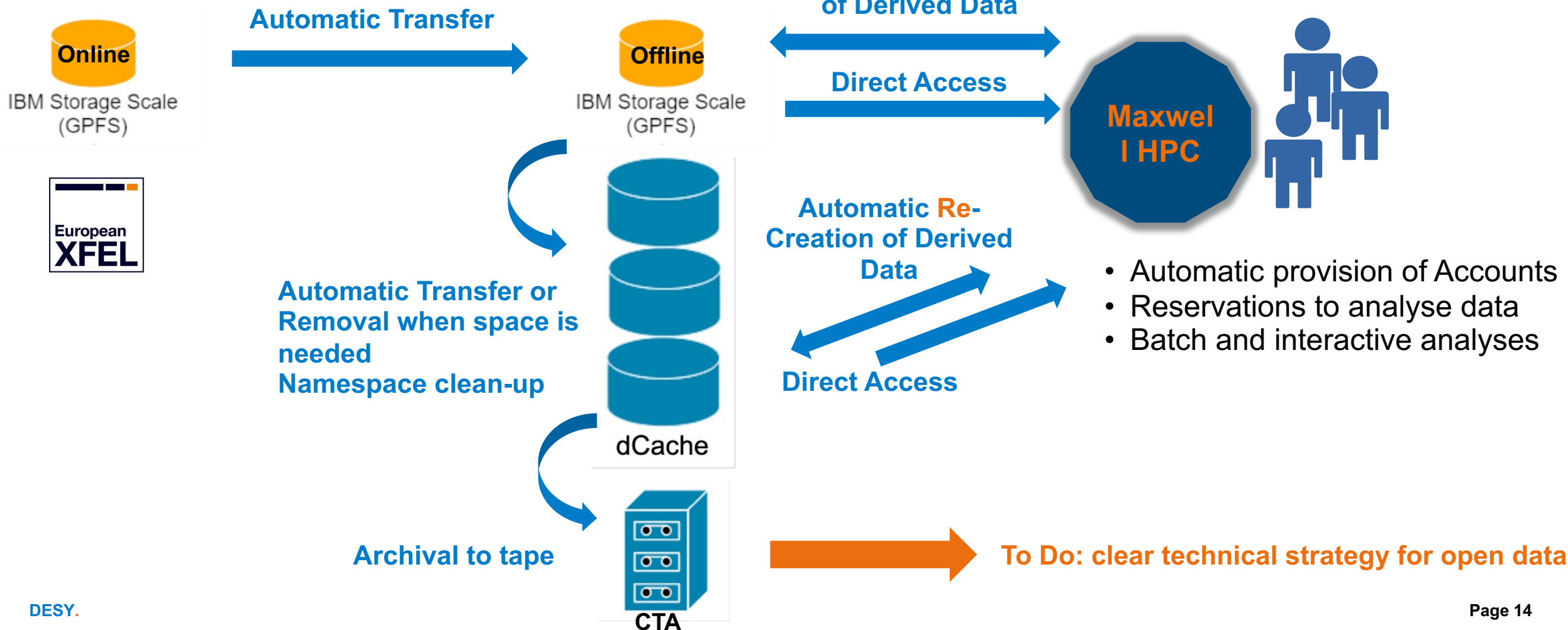


“A strongly interconnected Storage & Compute forms the core of the infrastructure. Users interact with Storage & Compute through well integrated services & portals.”

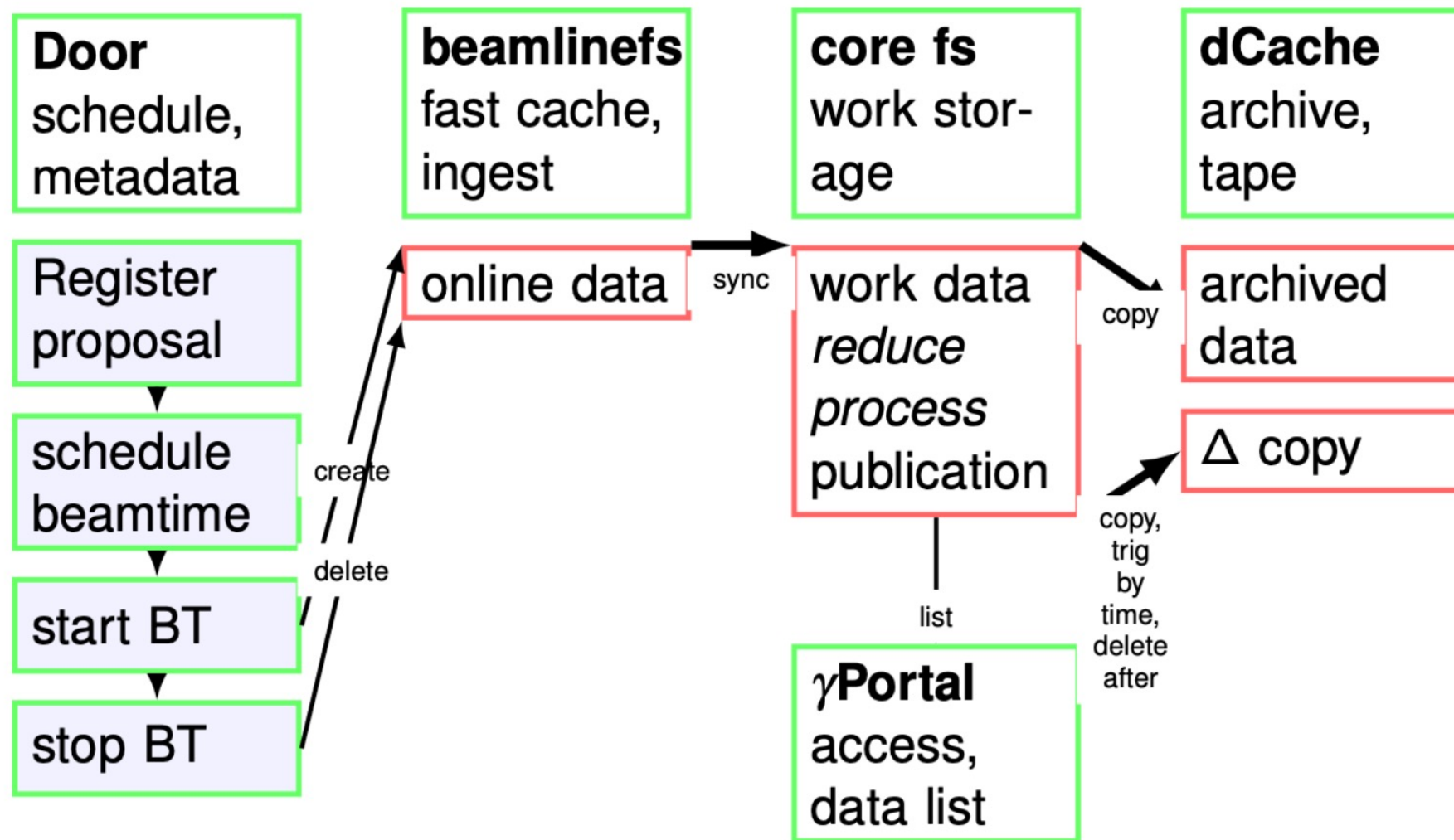
On-Site Example

User Proposals for European XFEL

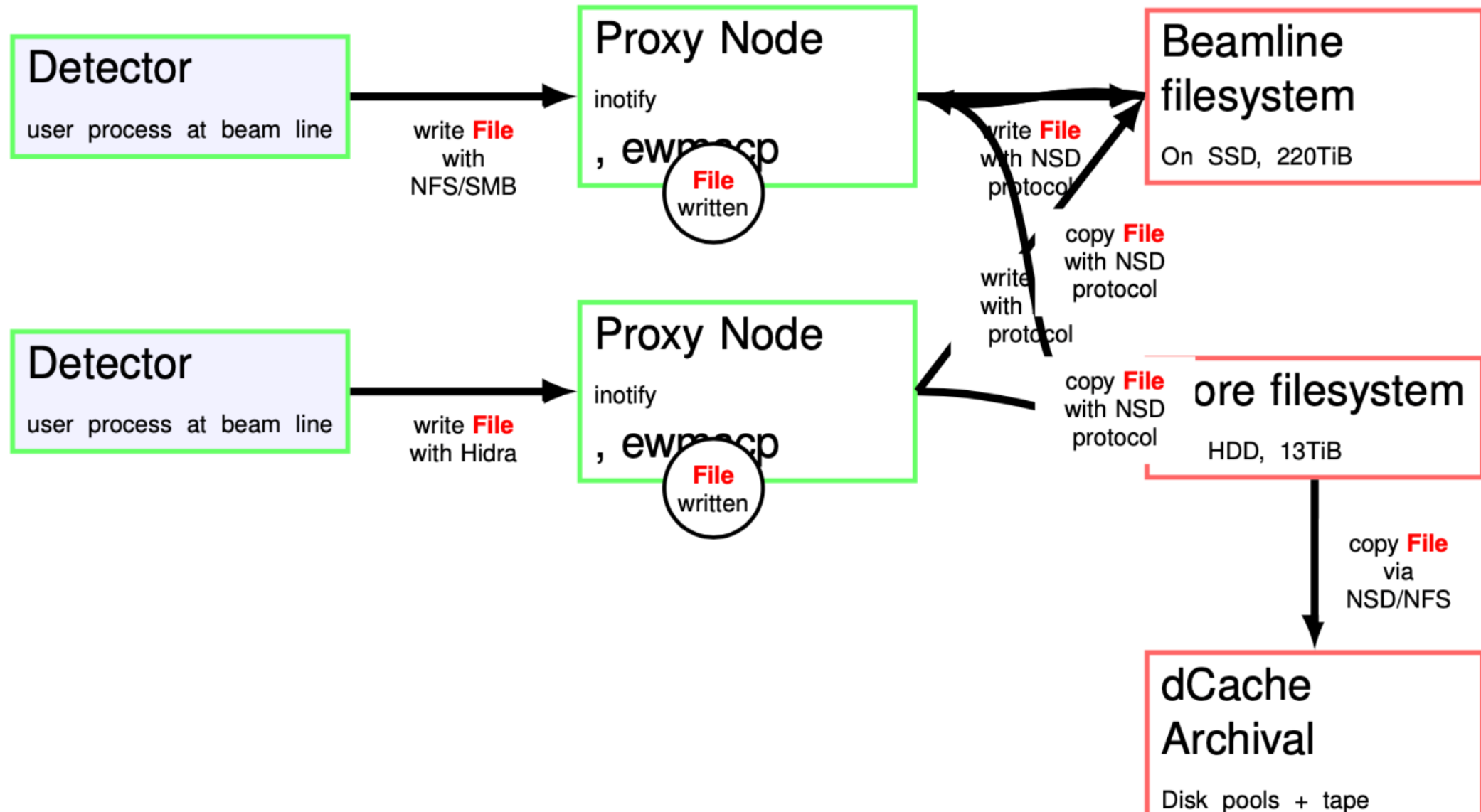
- Developed largely by our colleagues at the European XFEL → Analysis is centered on Maxwell HPC
- DESY Largely involved in data transfer and archival



On-Site Example: Workflows for Petra-III



Data taking architecture for Petra-III



User accounts in a nutshell

- **Grid:** Since pilot jobs: One VO → Mapped to O(10) DESY accounts
- **NAF:**
 - External user register using an X509 based form
 - Get normal DESY accounts
 - Normal Login, normal UID/GID, ...
- **Maxwell:**
 - Internal users: Have DESY account, get adequate permissions
 - External users: Petra-III DOOR account NOT sufficient
 - If needed, get a PSX / UPEX account ... which is basically a normal, stripped down version of a DESY account, attached to a beamtime/proposal

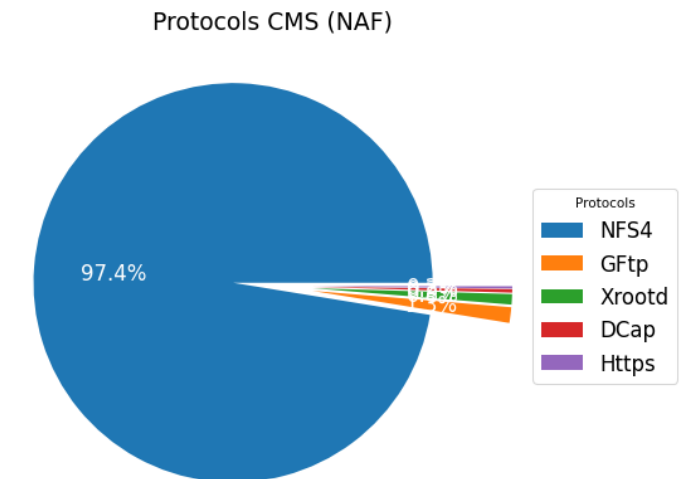
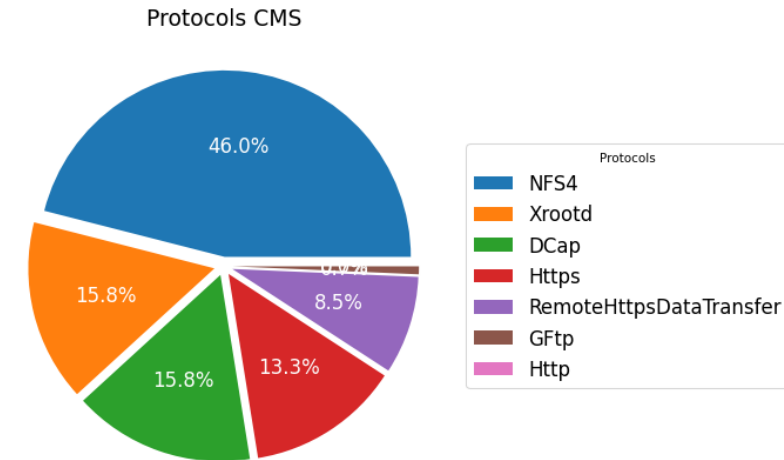


governed by DESY Registry

Challenges: The Return of POSIX

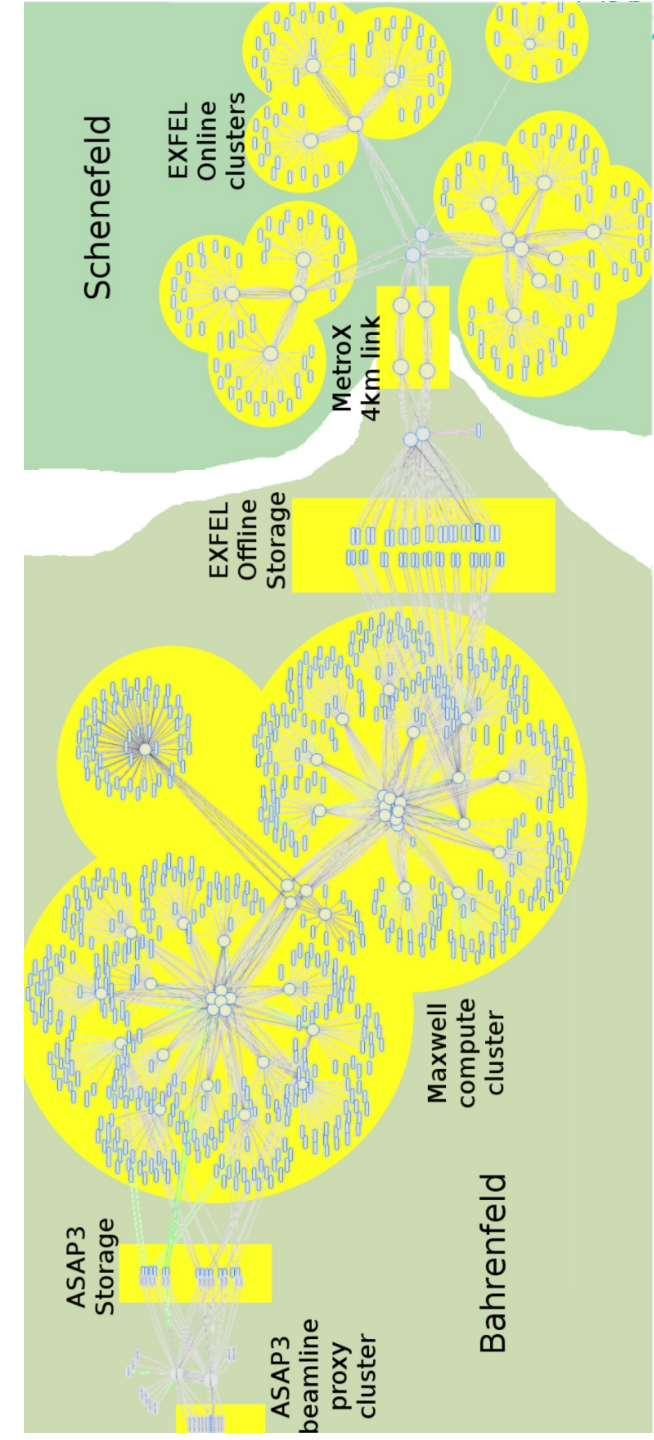
POSIX Reliance on Data Access

- We see ever increasing POSIX access pattern
 - Photon science software often can only read 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**



Challenge: Infiniband

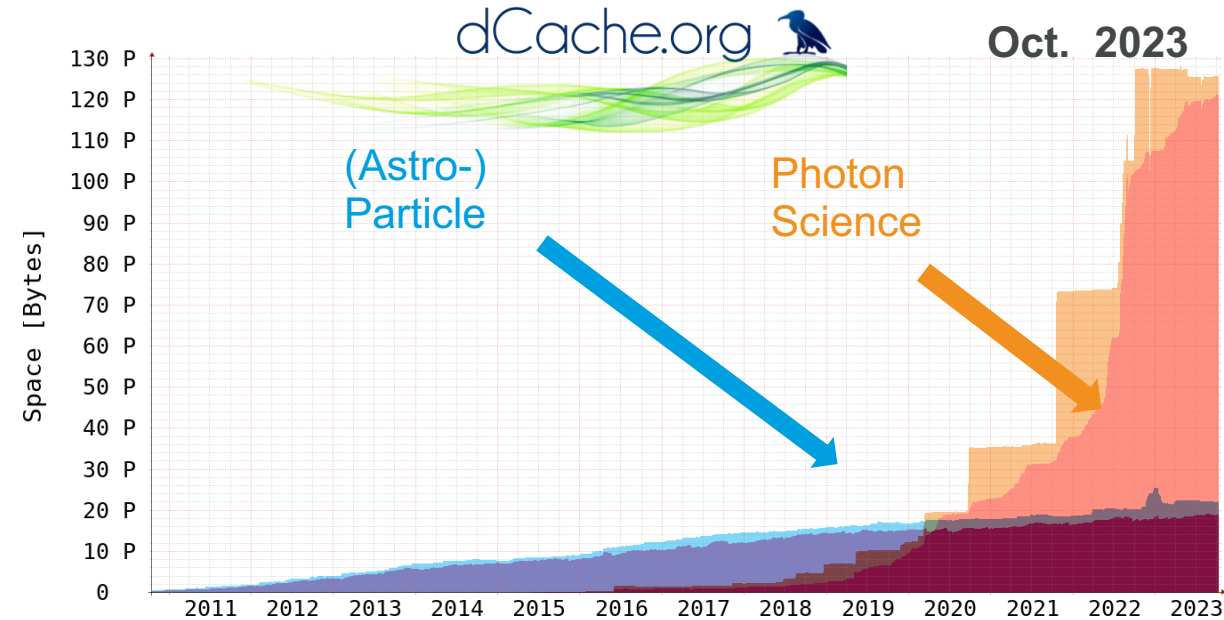
- Would like to get rid of it ... (vendor lock-in, upgrades, ..., mgmt. by Maxwell team) ... but:
- Our storage experts say, GPFS (+NativeRaid) is currently still the storage system that is
 - best performing
 - most secure
 - reasonably expensive
- Our GPFS experts say: GPFS works best using InfiniBand
 - RoCE is still living on the edge
 - When Ethernet: Use Cluster Export Services + NFS (Ganesha) (e.g. in the NAF)
- At the moment, we do not see an alternative to InfiniBand when fast POSIX cluster storage is needed.



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**

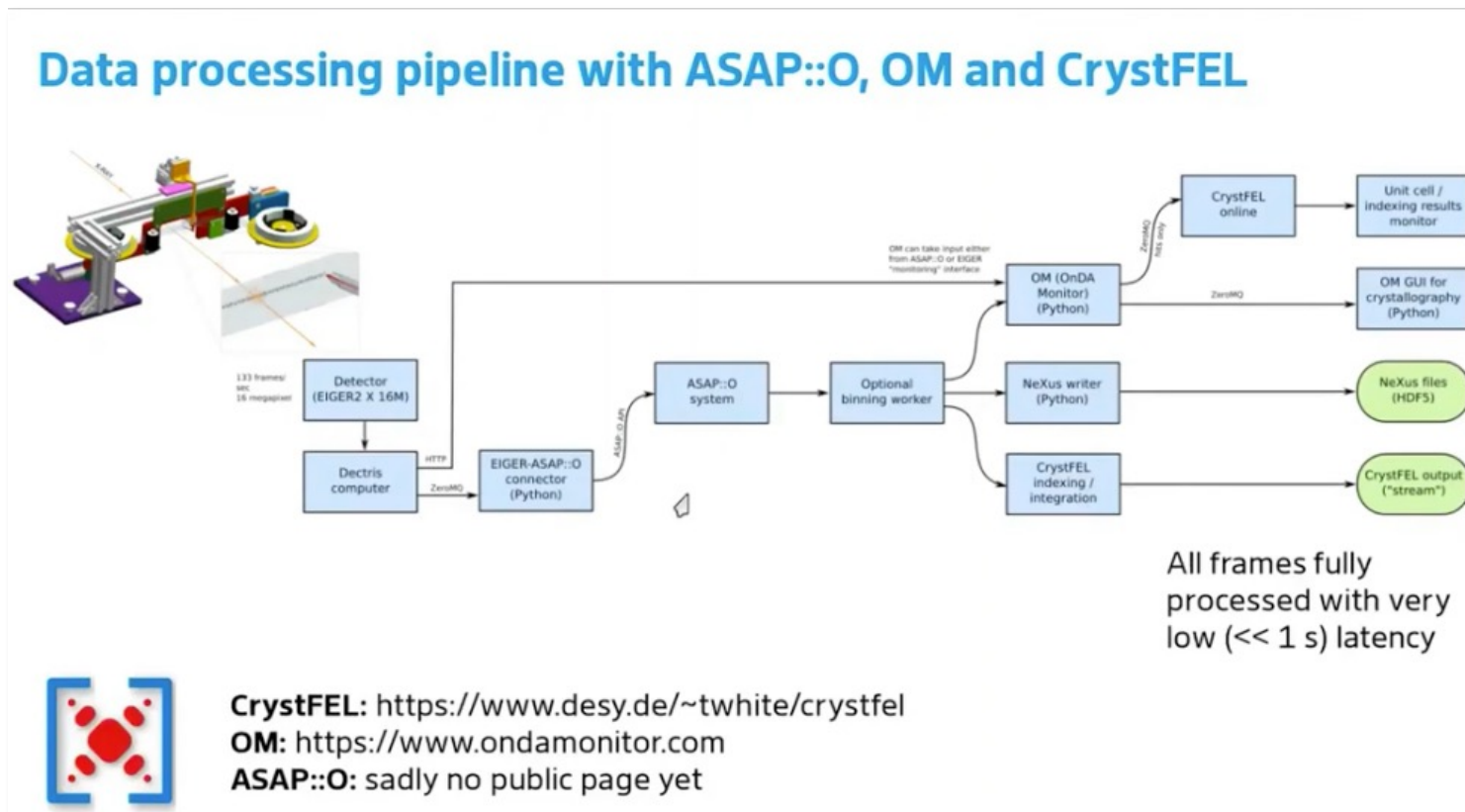
Challenges: Data reduction ... and getting rid of (some) storage during DAQ?

- Data reduction: Work in progress: Combined effort: Scientists, beamline scientists, photon science community (experts), Facility
- **Data reduction** essential! Integrate data reduction workflows

One example:

CrystFEL

Thomas White



Designed to be Fast



Focus on What Matters



Easy to Use

Challenges: Our Maxwell HPC partition setup

- Grown historically: Group buy-in
 - Groups have their own partitions
 - XFEL basically owns $\sim\frac{1}{2}$ of the cluster
- If you decide for a HPC setup, plan for centrally purchased systems, resource sharing
- Maybe one of the largest obstacle for technically integrating Grid + NAF + Maxwell

Challenge: Namespaces

- /pnfs/desy.de ... is dCache obviously
- /nfs/dust ... is an NFS mount (of a GPFS system)
- /gpfs/... is a GPFS mount, via NSD
- Cross mounting e.g. /gpfs via NFS would either lead to:
 - inconsisting naming scheme (/gpfs being both NSD and NFS)
 - inconsiting path (/gpfs/some/data == /nfs/gpfs/some/data depending on system)
- Luckily, we got rid of “/data” or “/experiment” long time ago for central systems

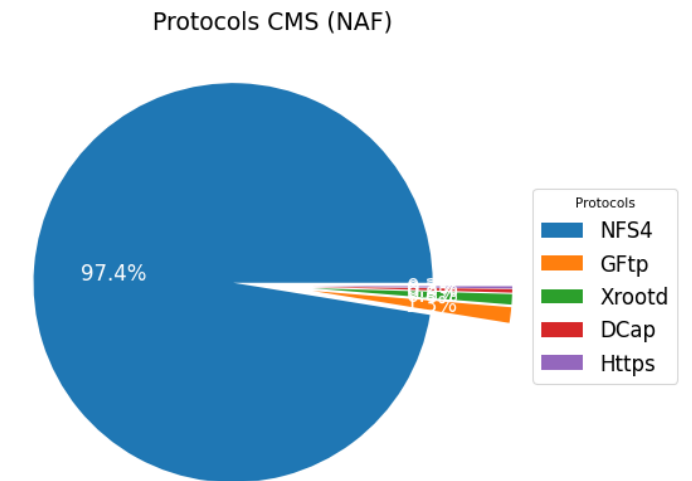
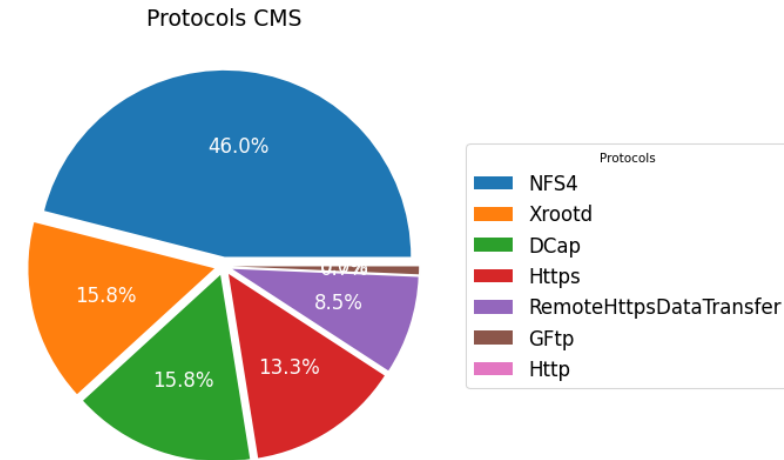
Challenges: Commercial users (Photon science only!)

- Difficult
- Licensing
- Usage of tools
- Usage of infrastructure!
- Currently: Dedicated nodes (<10)
- Very similar setup to Maxwell (minus some parts)
- Petra-IV calls for more commercial participation
- Probably will enlarge Maxwell copy for that purpose

Challenges: The Return of POSIX

POSIX Reliance on Data Access

- We see ever increasing POSIX access pattern
 - Photon science software often can only read 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: Software and support

HEP:

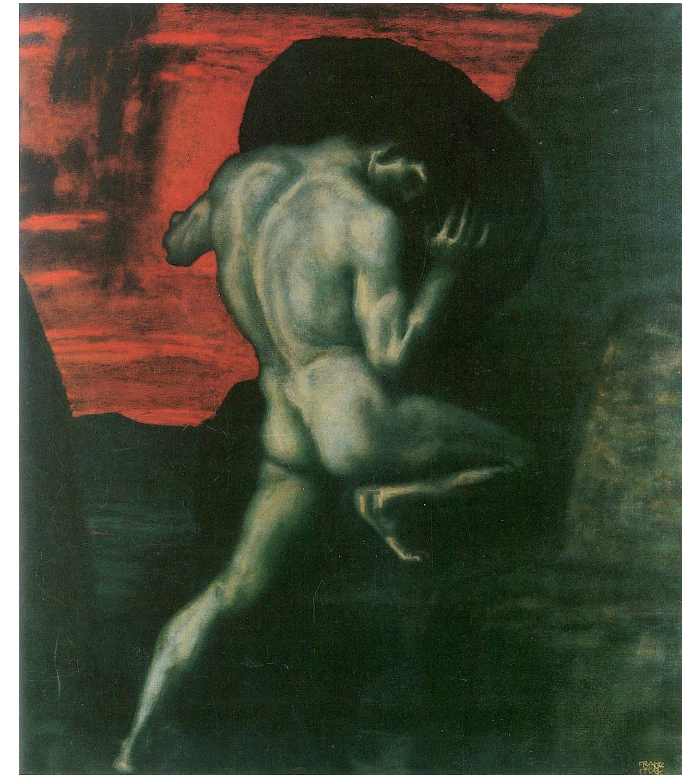
- Software: Mostly CVMFS. Most of it from CERN, some DESY
- DESY CVMFS: Infrastructure managed by IT, content managed by experiment experts (ILC)

Photon Science

- Lot of SW managed by Maxwell team (well, one person out of it) → central GPFS installation
- Some group software → group GPFS space

User support

- End-users tend to be support intensive. End-user support is critical to facility success!
- often little know-how level with users
- Try to hide some complexity → e.g. specialized application portals through Jupyter
- ... but in the end: Lot of work





$$\mathbf{B} = 0$$

$$\nabla \cdot \mathbf{D} = 0$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \partial \mathbf{D} / \partial t$$

Pages / Maxwell Cluster / Software

Applications

Brief alphabetical list of documented applications

A-B	C	D-F	G-J	K-N	O-P	Q-U	V-Z
abinit	ccp4	dawn	gamess	Comsol Multiphysics	O	qchem	vmd
adxv	chimera	DeepQMC	gaussian	keras	oasys	QuantumEspresso	XChemExplorer
albula	CMIstark	demon2k	gbench	lammps	Octave	quanti	xcrysden
alphafold	cns	dials	genesis	Lumerical	olex2	R	xds
alphapulldown	Comsol Multiphysics	dioplas	geopixe	Mathematica	opal	rapids	xdsapp
Amira-Avizo	comsyl	DirAx	gromacs	MATLAB	openbabel	RF2NA	xdsgui
Ansys	condor	dpdak	hdf5	Maud	openfoam	rosetta	xia2
arp_warp	conuss	elegant	hdfview	mosfilm	openmolcas	rosettafold	xmimsim
atompaw	cp2k	elk	hexrd	namd	openstructure	sasfit	xop
atsas	CrystFEL	fasta	IceNine	nwchem	Oracle	scikit-image	xrt
autodock		FastX2	IDL		orca	scikit-learn	
balbes		FastX3	ilastic		Origin	Scilab	
blast		fdmnes	ImageD11		PanDDA	shelx	
blender		ffmpeg	imod		phenix	spark	
BornAgain		fiji	Impact-Z		platon	srw	
		firefly			puffin	TensorFlow	
		FLUKA			pyfabio	TeX	
					pyFAI	tomopy	
					pyMca		
					pynx		
					pytorch		

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 \leftrightarrow internal network
- At the entrance: Introduction of MFA planned for end of 2023 for all interactive logins to IDAF

Summary

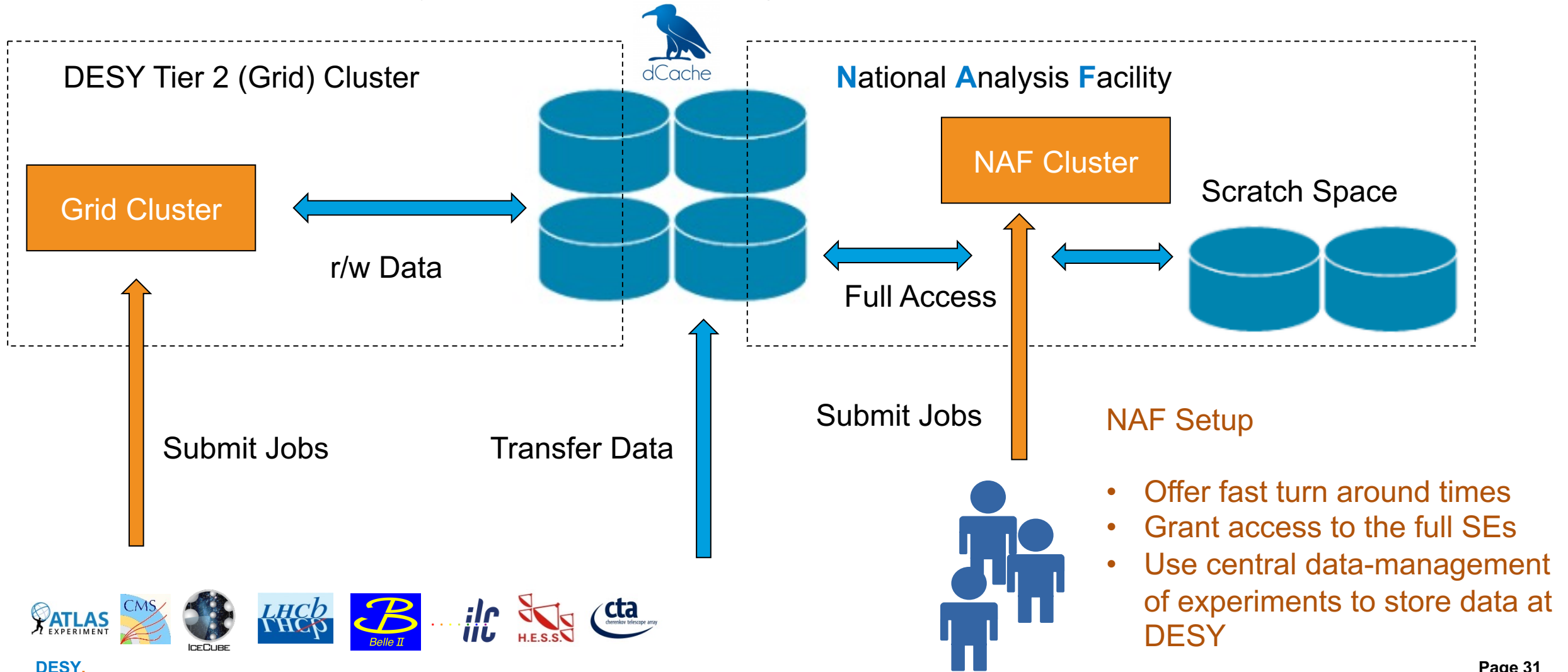
- HEP and Photon Science are different
- So are the solutions we implement
- Some commonalities
- Some differences are inherent, some because of past choices, DESY reasons ...
- **Get back to us for more details!**
- **Visit PhotonScience User Days End of January 2024 at DESY**

BACKUP Material

Paradigm: Data Analyses are Data Driven

As Underlying Principle of the Particle Physics Infrastructure

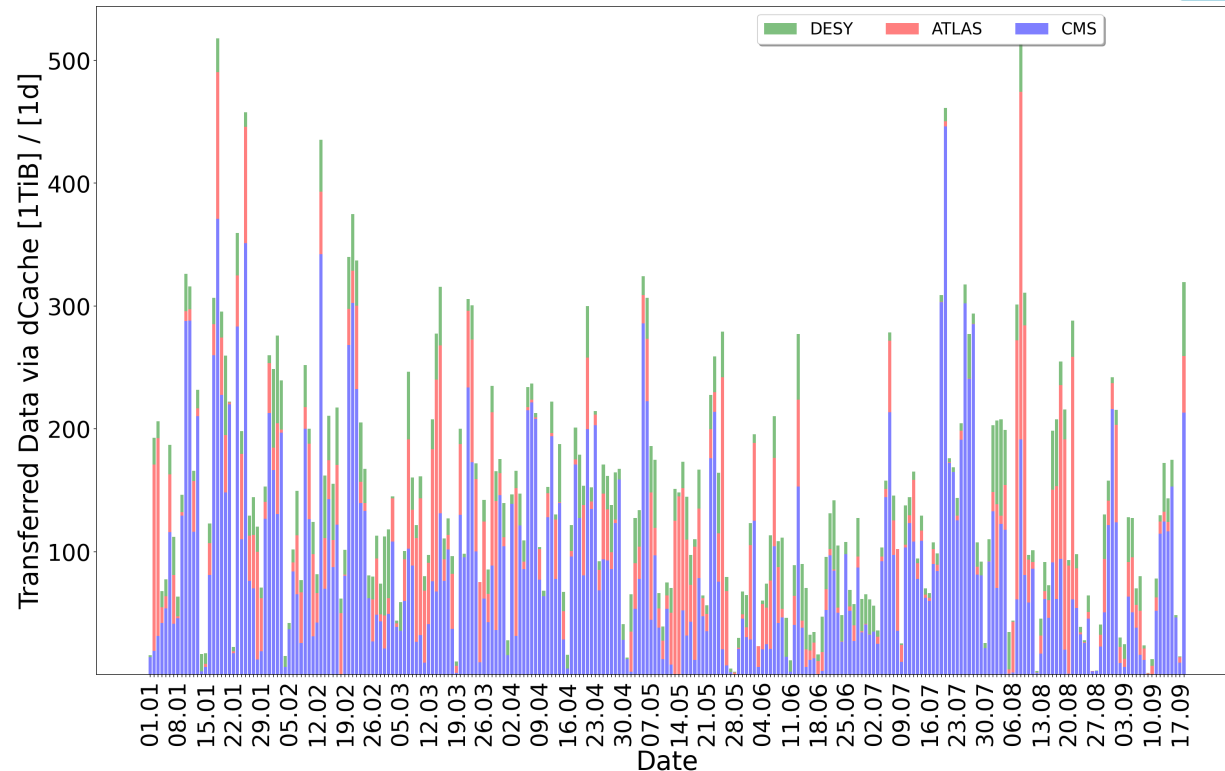
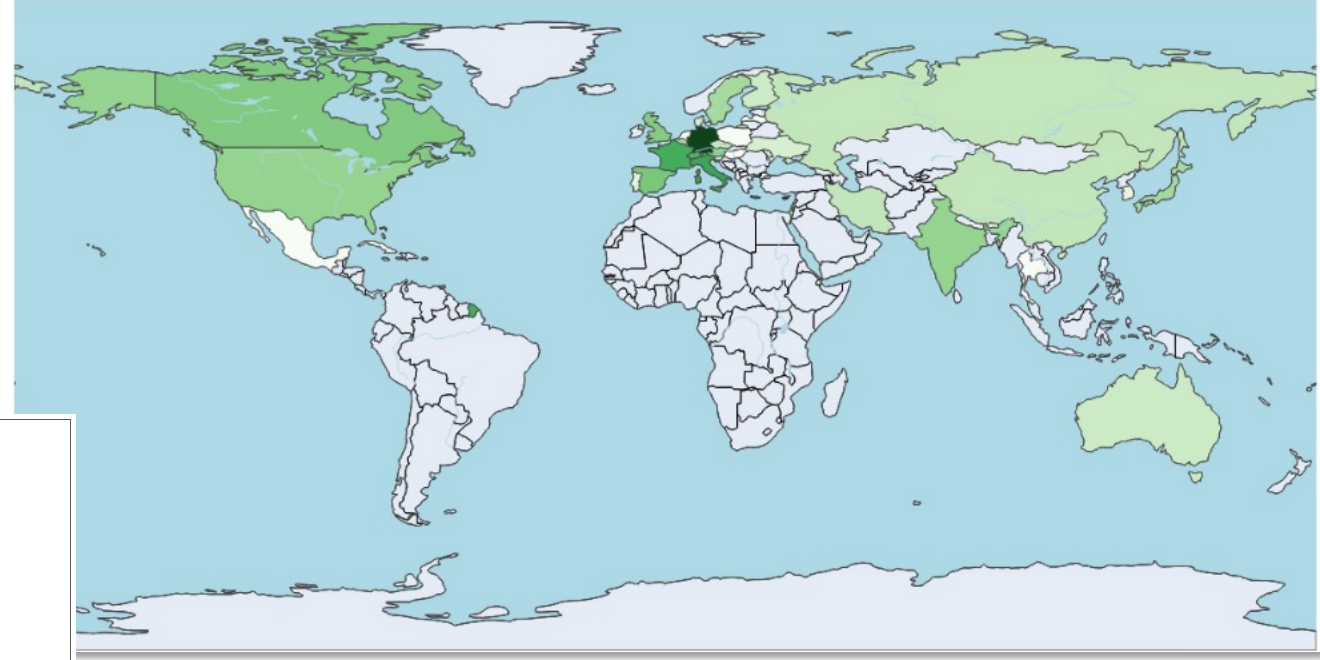
- Almost all HEP data analyses require access to large amounts of data



Users of the NAF

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

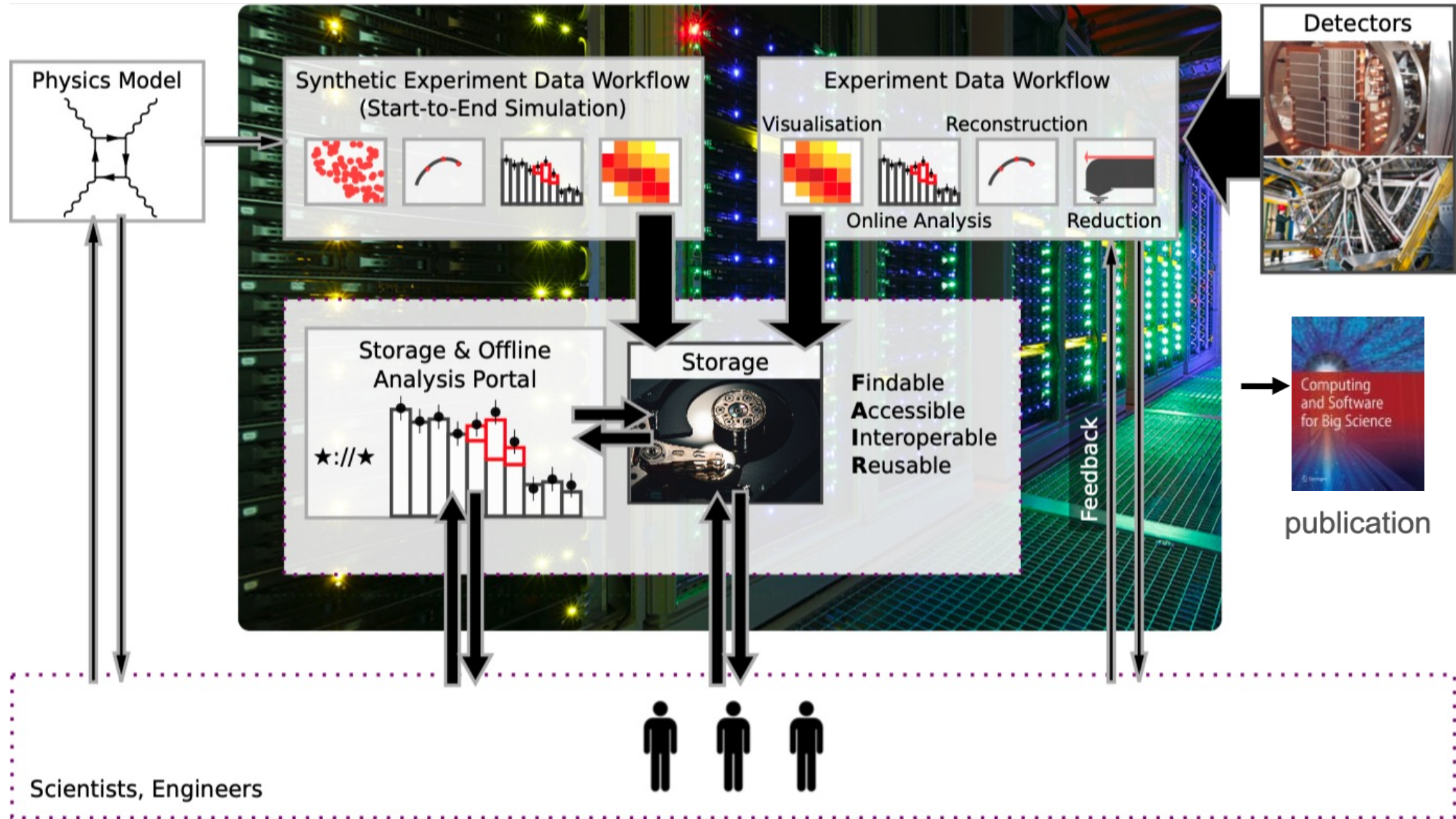
- Interactive usage of the NAF
 - Most users from German universities
 - All Belle II scientists are potential NAF users
 - Large number of international users



- Data access inside NAF (only dCache shown)
- Additional storage space for NAF (linked to experiment framework)
- 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



Challenges: Hardware evolution and Person Power

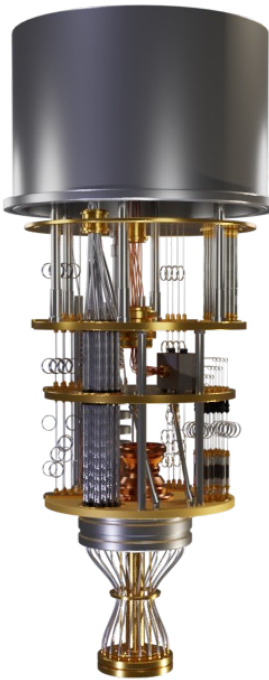
Difficulty Acquiring Hardware and Filling Open Positions

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

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



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

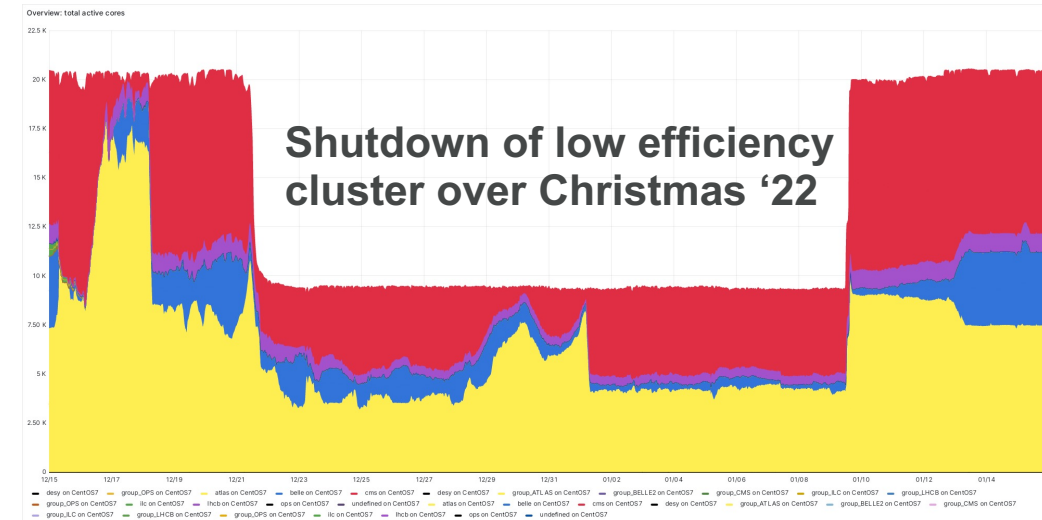
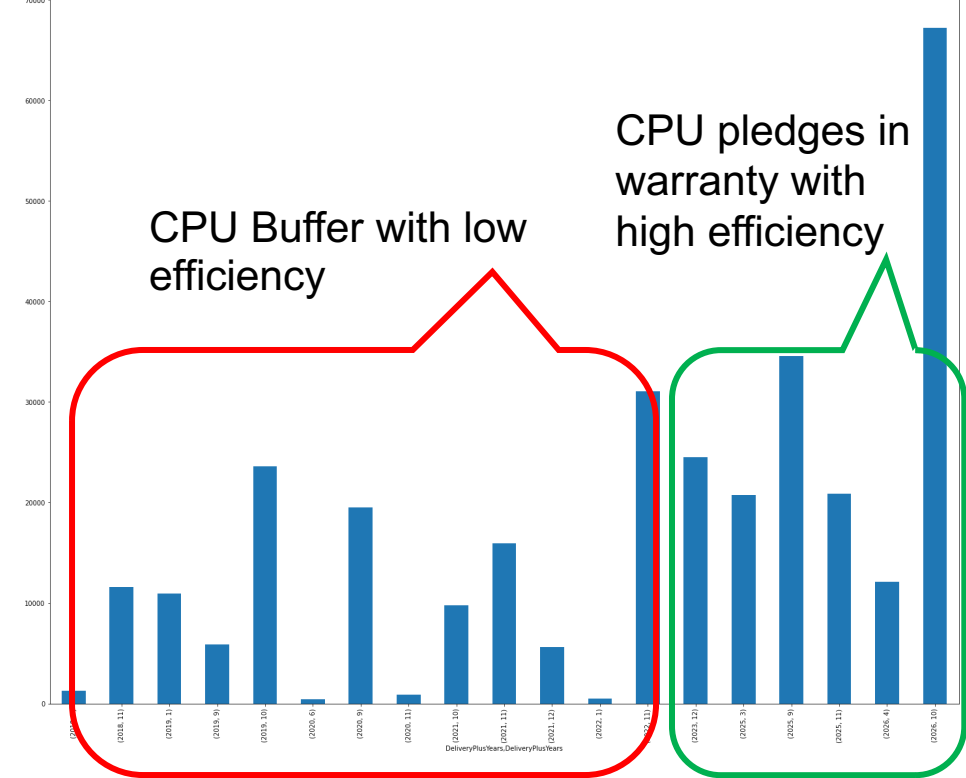
Storage: Unused data on tape → Tape?

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:

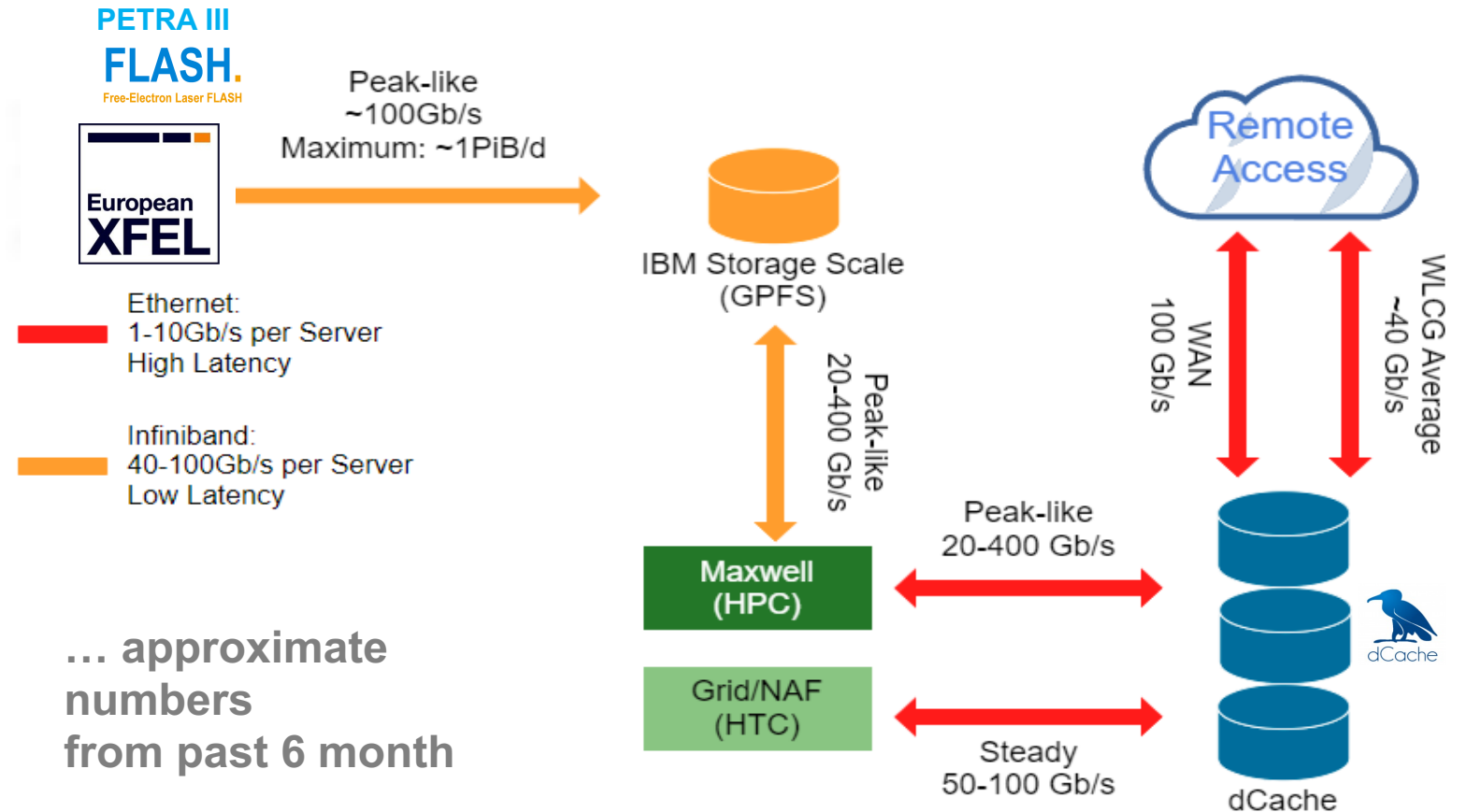


Provided by T. Hartmann

IDAF: Bandwidth for Flow of Data

Connecting Detectors, Storage and Compute

- Ingest rates up to several PiB/day
- Overall about
 - 80k cores / 250GPUs
 - 200PiB GPFS/dCache storage
 - Recently extended tape system (stored currently ~150PiB)
 - 1.5k servers
- Very heterogeneous hardware



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

