HTC @ DESY

Introduction and overview

Beyer, Christoph
Madison, 12-07-2024

HELMHOLTZ

DESY.

DESYs research areas

The cornerstones of DESY's mission

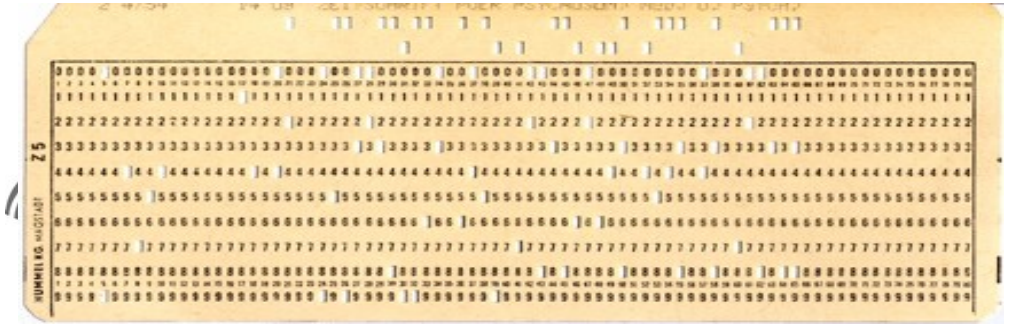
'Batch' ?

> 'Batch' = engl. fuer 'der Stapel'

> At ‘mainframe’ times meant more than one punch card in a row

> Charcteristics of a batchjob today:

- No inter action with running job
- Runtime typically between 1h and a week
- User/owner of the job has no access to the actual hardware the job is running on (ssh-to-job is possible at NAF)



What can be processed as a ‘batchjob’ (esp. as a HTC job) ?

> Anything that does not rely on interaction during runtime

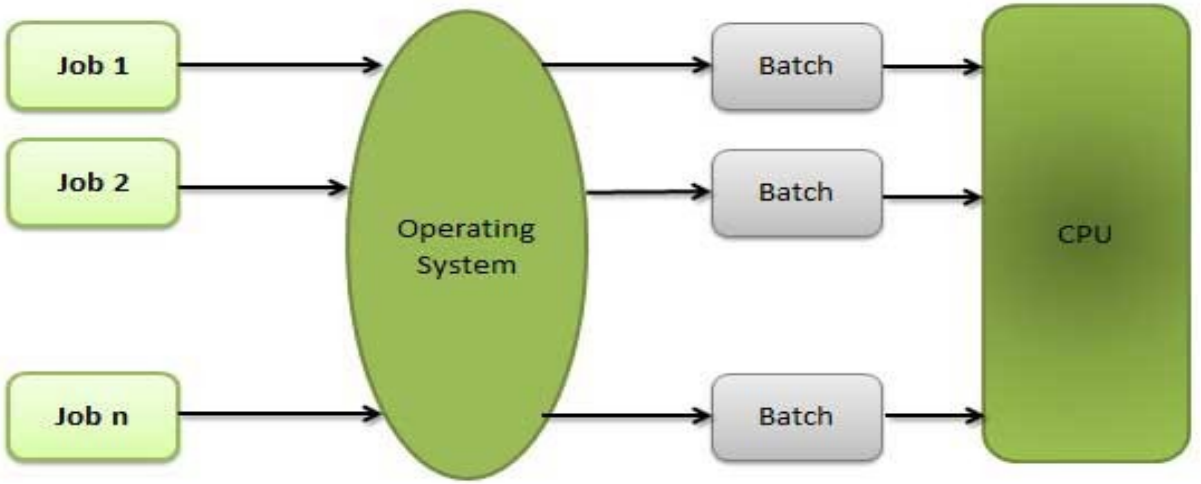
Typical batch-workflows:

> Annual accounts

> Formatconversions

> Data analysis

> Simulations



HPC vs HTC

HELMHOLTZ ASSOCIATION

DESY

HPC vs HTC

HEL
ASSOCIATION

DESY

What the batch system actual does – in case of HTCondor

- HTCondor manages Jobs
- HTCondor manages Workernodes
- HTCondor manages Datasources & Datatransfers
- HTCondor scales and runs encrypted and secure
- HTCondor simulates workflows
- HTCondor provides a monitoring interface
- HTCondor administers user access, quotas and priorities

HTCondor - Classadds


```
[
Type = “Apartment”;
SquareArea = 3500;
RentOffer = 1000;
HeatIncluded = False;
OnBusLine = True;
Requirements =
  TARGET.RentOffer >
  (MY.RentOffer – 150);
]
```

negotiator

```
[
Type = “Renter”;
RentOffer = 900;
Rank = 1/other.RentOffer;
Requirements =
  OnBusLine &&
  SquareArea > 1500;
]
```

HELMHOLTZ | ASSOCIATION

DESY

HTCondor – Classadds type ‘machine’ vs. ’jobs’

```
MyType = "Machine”
Name = "batch0930.desy.de"
START = TRUE
VirtualMemory = 342696
Disk = 28728536
Memory = 160
Cpus = 8
Arch = "INTEL"
OpSys = "RHEL_6“
```

```
MyType = "Job"
Owner = "chbeyer"
Cmd = "/usr/users/chbeyer/mem_stress.sh"
Output = “mem_stress.out”
Args = “100”
ImageSize = 2544
DiskUsage = 2544
Rank = Memory/Mips
Requirements = (Arch == "INTEL") && (OpSys == "RHEL_6")
                && (Disk >= DiskUsage) &&
                (VirtualMemory >= ImageSize)
```

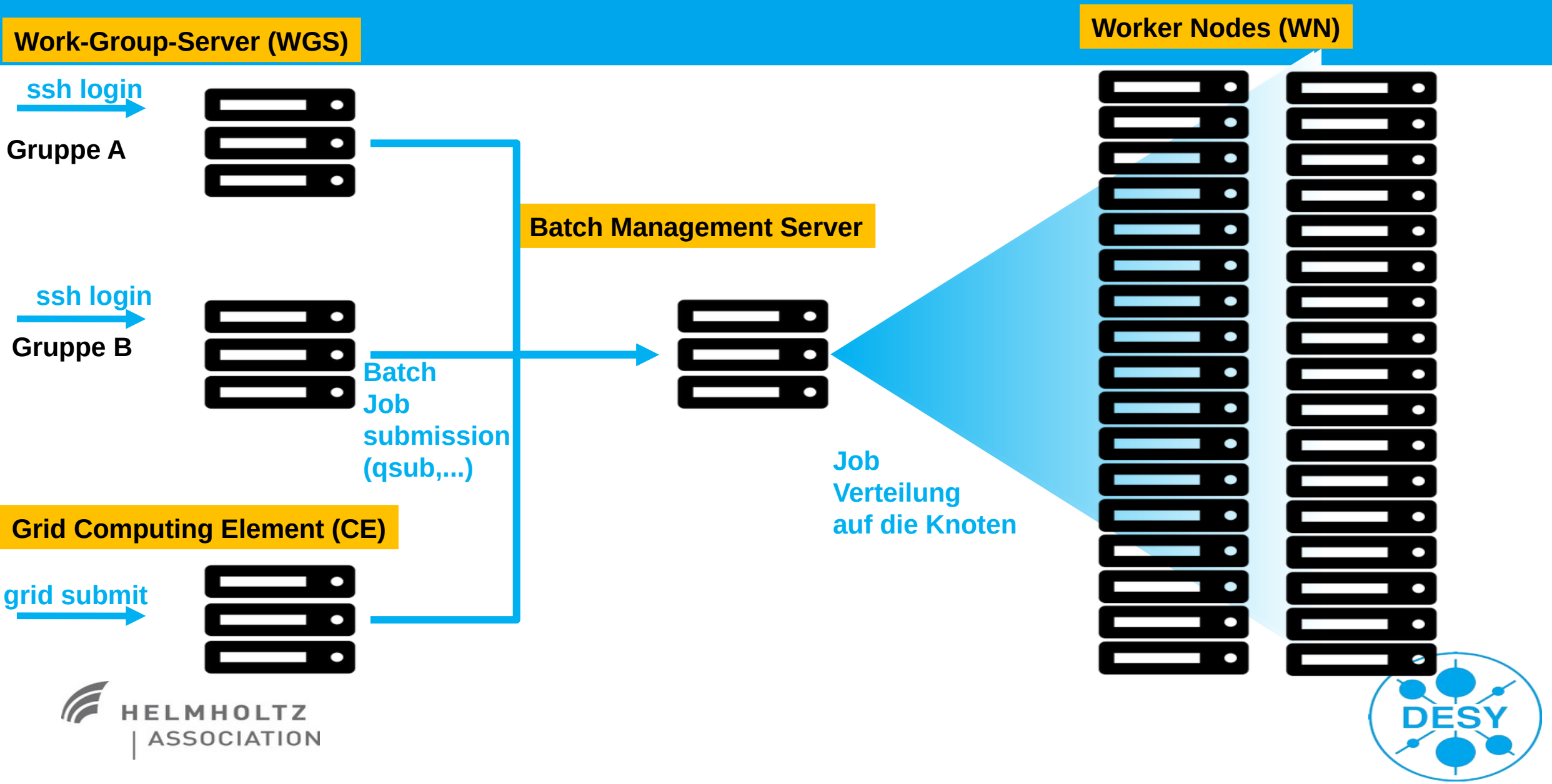
HELMHOLTZ | ASSOCIATION

DESY

On the Workernode

- Partitionable, dynamic Slots
- Meaning: The workernode provides one partitionable slot with all available ressources
- After the 'negotiation cycle' a 'Child slot' with the needed ressources gets carved off the parent slot
- The remaining parent slot will be negotiated again in the next cycle
- The usual mix of multi- and singlecore jobs is always demanding and some more flexible concepts are in place to cope with that

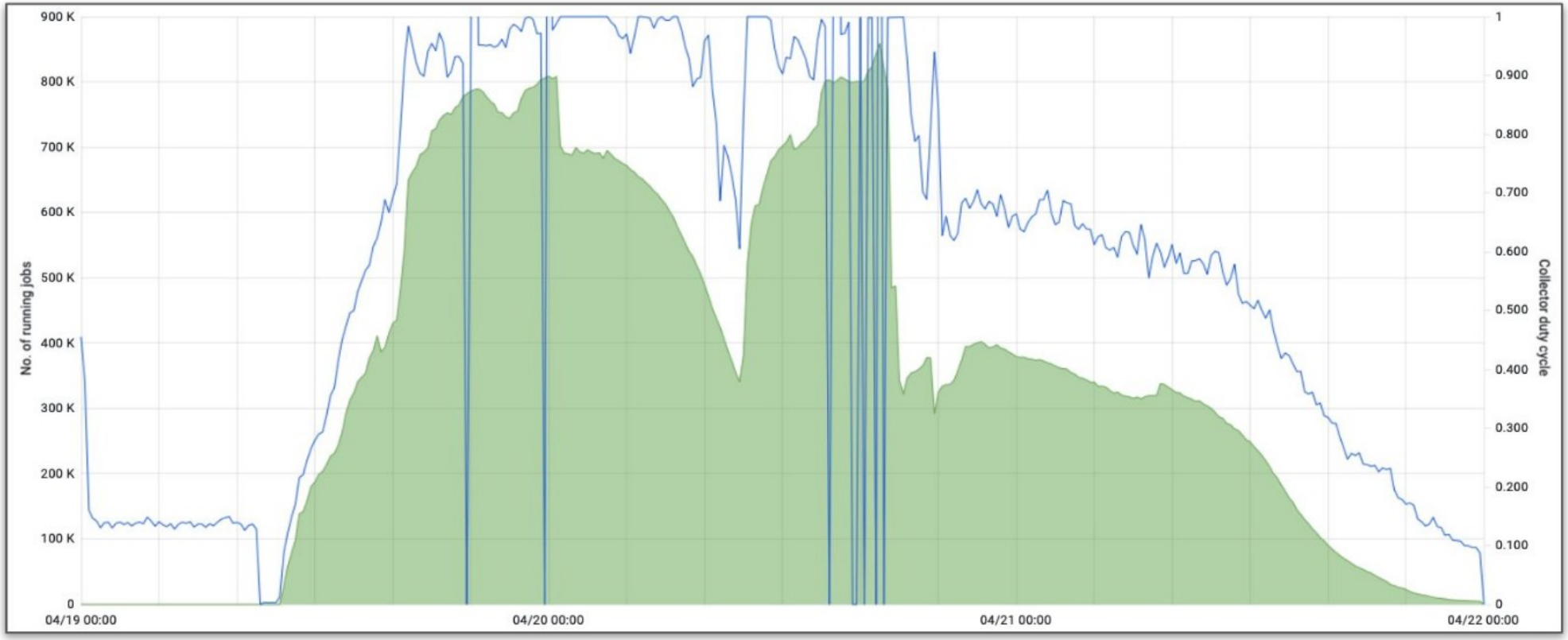
Principal setupof a Batch-Farm



The global CMS pool – HTCondor in large scale

1.000.000 cores 800.000 running jobs

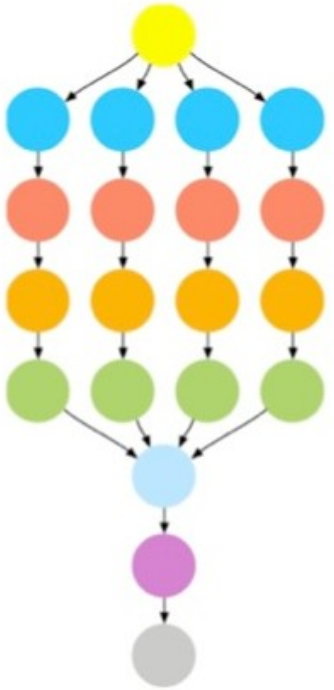
- Pushed the scalability of our Global Pool to about **800k simultaneously running jobs**
  - Factor of 8 away of current pool size considering 4 core jobs



condor_dagman

directed acyclic graph - a cool feature

- DAGMan (Directed Acyclic Graph Manager) is a workflow management system based on graphs
- Complex dependencies of tasks can be realized in DAGs
- Define dependencies and let DAGMan manage them for you
- Linear and parallel computation models
- Resubmitting of incomplete jobs with one command



-
- For simple example See: bird.desy.de &

  https://swc-osg-workshop.github.io/OSG-UserTraining-AHM18/novice/DHTC/04-dagn

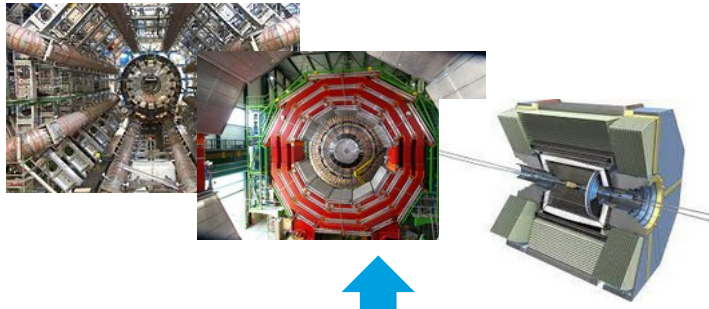
GRID & NAF

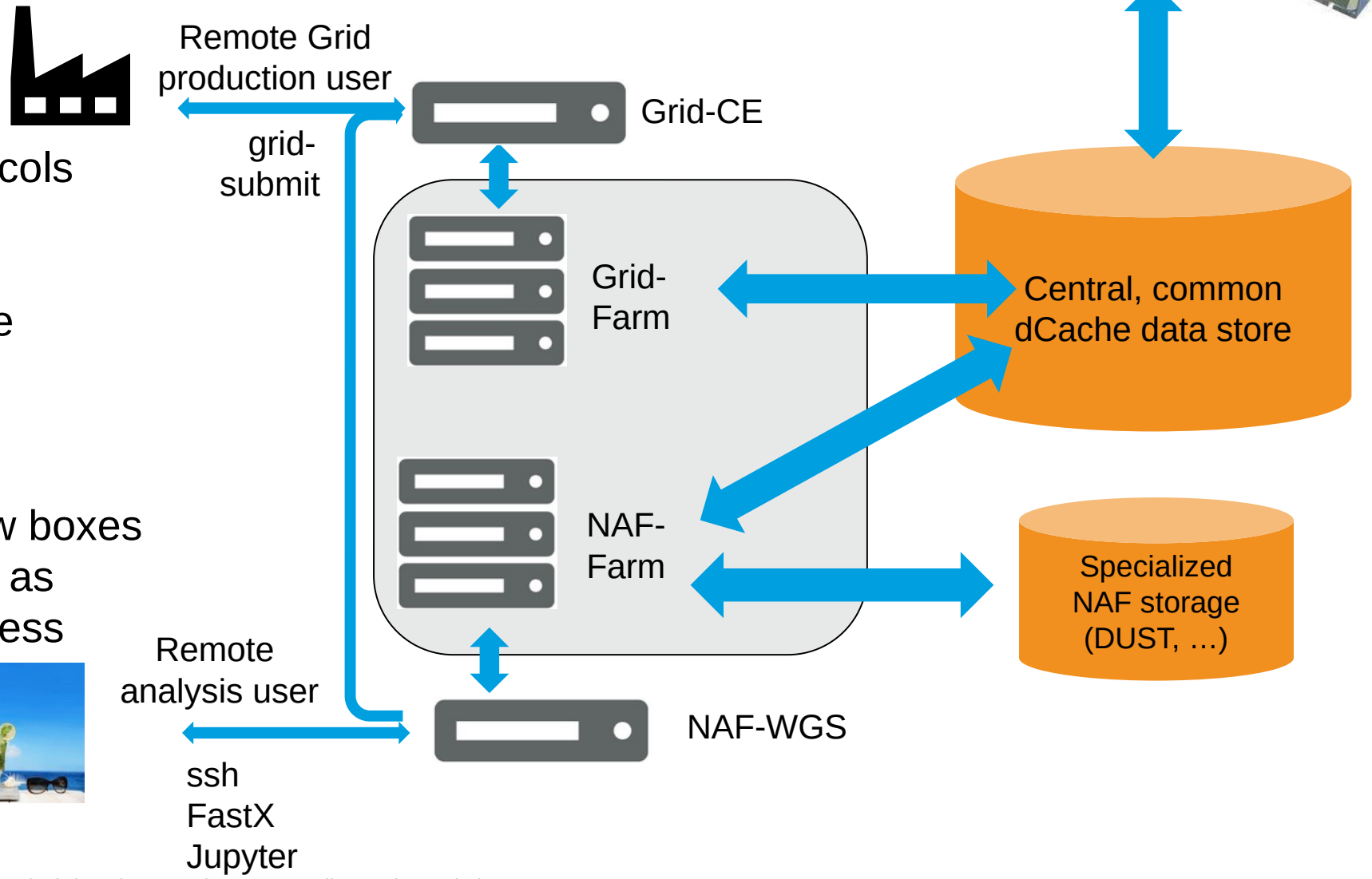
Two HTC AFs

Grid: Serves worldwide HEP community through Grid protocols

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

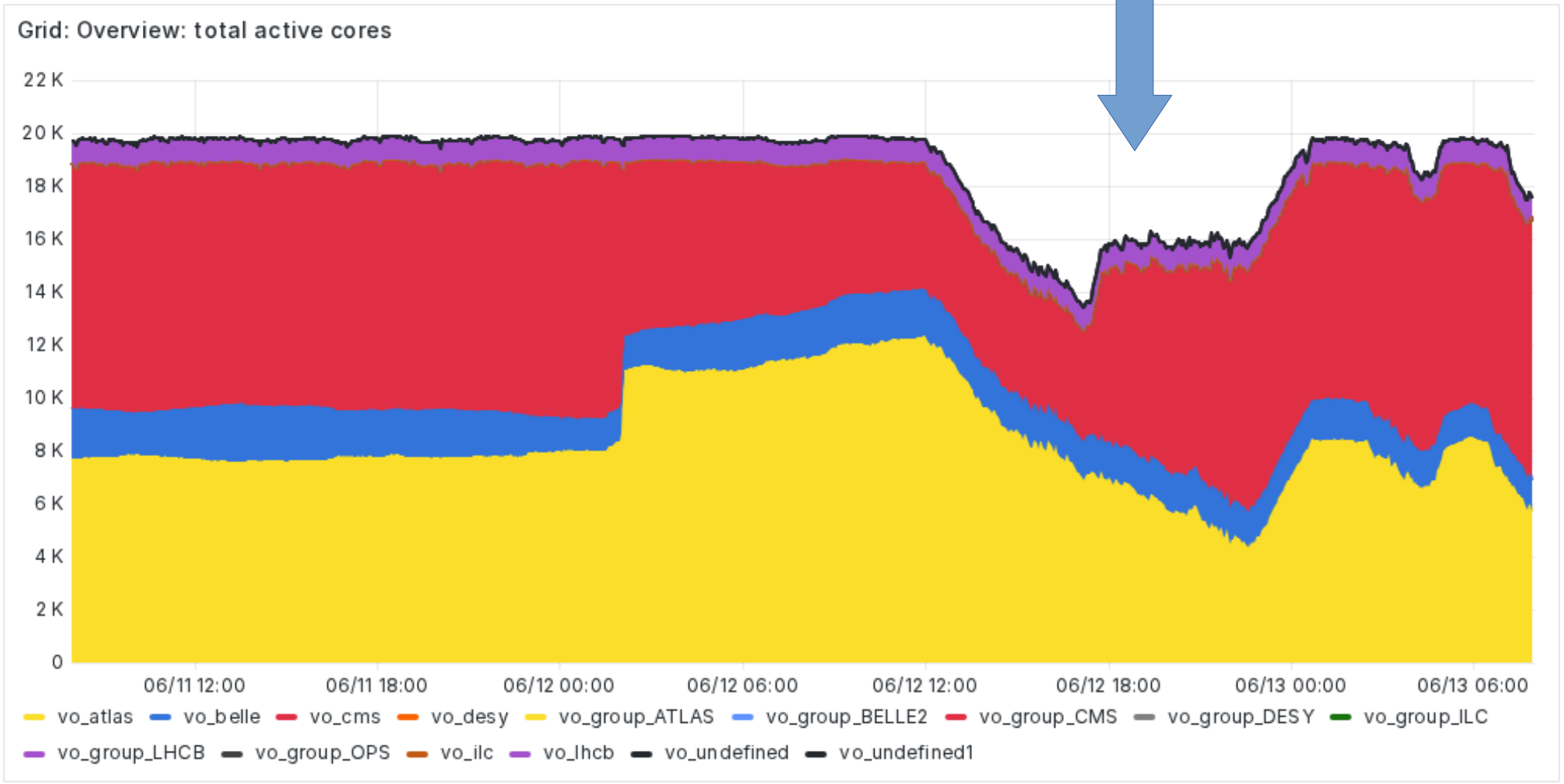
Access protocol is just one/few boxes large compute behind, as well as storage infrastructure and access is (mostly) identical



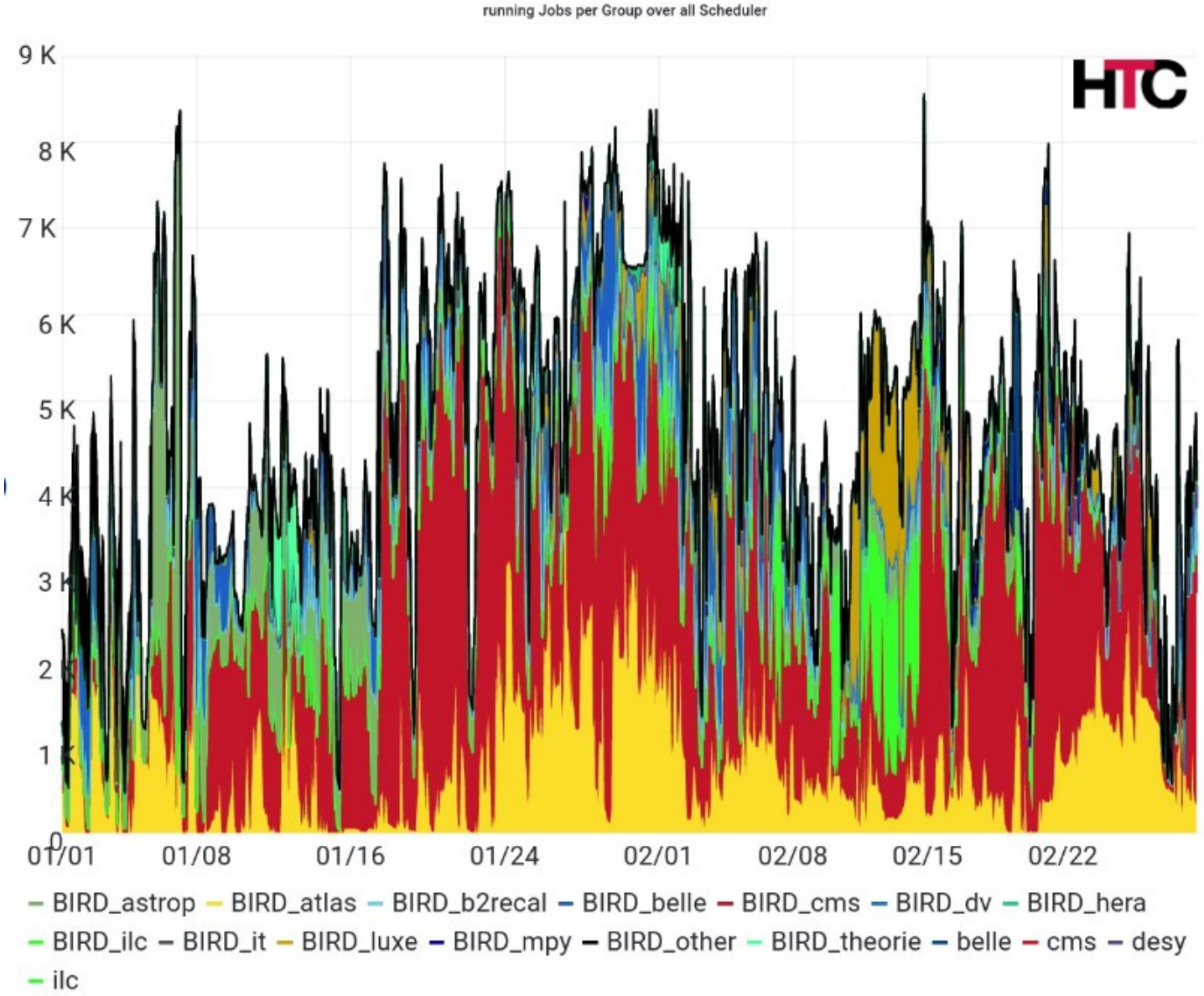


Usage in comparison – GRID

Something scary happened

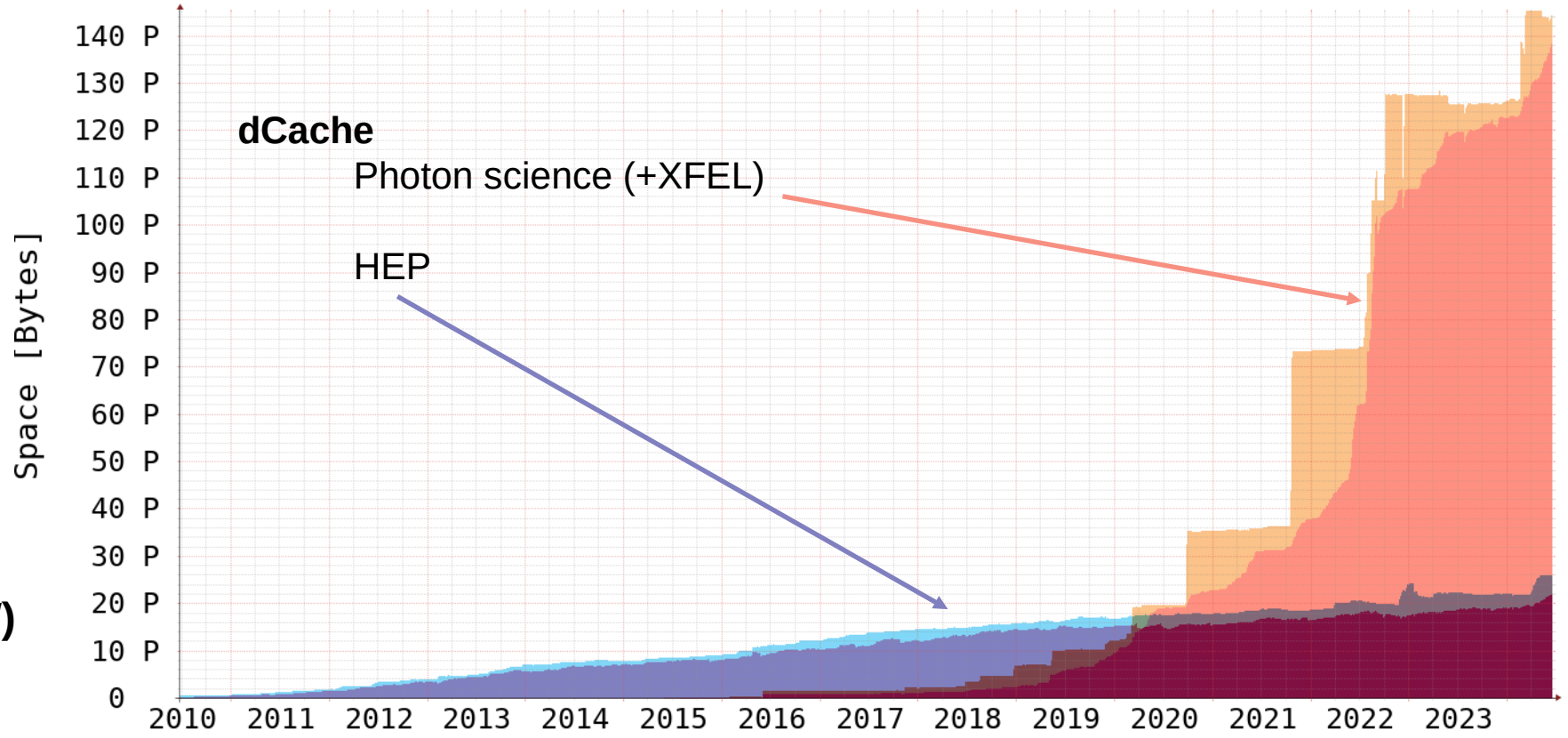


Usage in comparison – NAF



Everything is pretty much data driven

- **HPC + Grid + NAF**
- **~180 PB data on disk**
- **dCache + GPFS + BeeGFS**
- **~60.000 CPU cores, ~380 GPUs**
- **HTCondor, SLURM**
- **~2.700 server (compute, storage, management)**
- **~ >0.8 Megawatt (out of 1.3 MW)**



National Analysis Facility at DESY

Result of Continued Deployment and Development

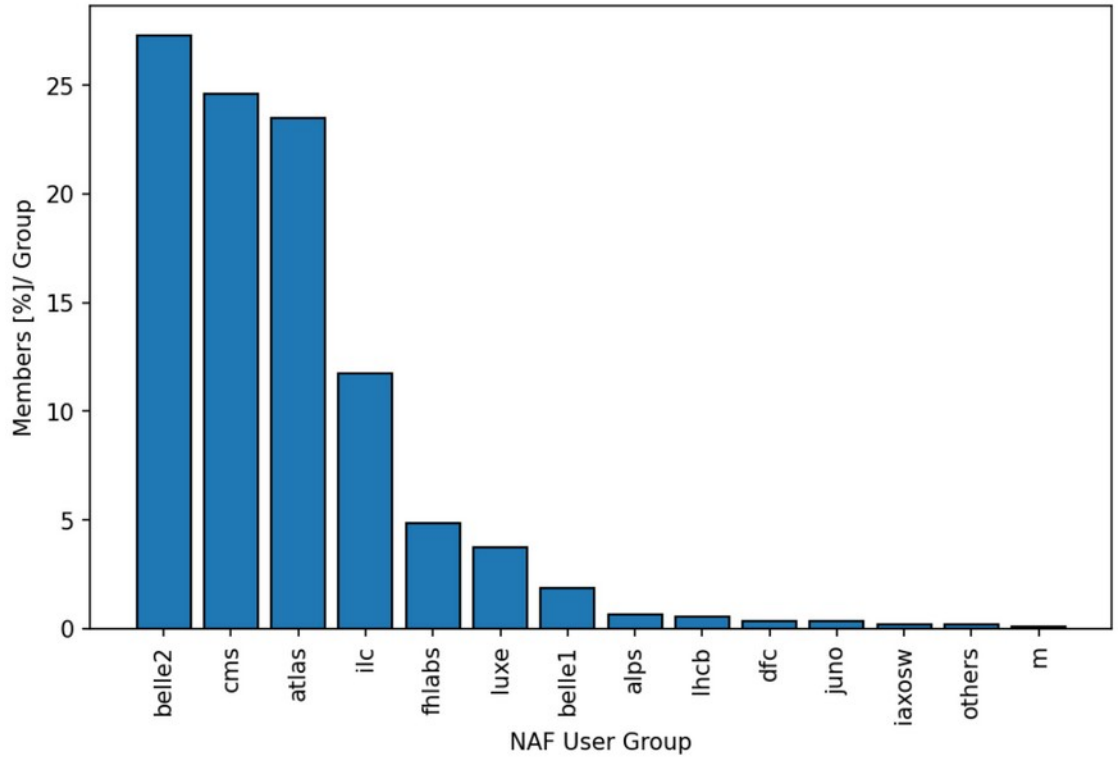
- 15 years user operations
  - NAF v1 started in 2007 in our two lab sites: Hamburg & Zeuthen (Berlin)
  - Reimplemented in 2013 as NAF 2.0 in Hamburg because
    - NAF v1 not integrated into central DESY IT
    - Data analysis over WAN was problematic
- Available to a broad user scope
  - Initial idea: for individual HEP users complementary to Grid workflows
    - faster job turn-around vs. high-latency grid jobs
  - Nowadays
    - Grid *primarily* production
    - NAF resources becoming even more relevant to individual users
    - Users beyond classic HEP experiments → Detector R&D, small scale Intensity Frontier Experiments

User Communities

Who is Working on the NAF

- DESY in general is a major site for the large HEP experiments
  - ATLAS, CMS, LHCb as Tier 2
  - Belle II (Raw Data Centre)
- Host large local groups for ATLAS, CMS, Belle 2
- Host ILC as Tier 0

- DESY hosts many smaller (on-site) communities
  - Detector/Accelerator R&D
  - On-site experiments: ALPS II, IAXO, MADMAX, LUXE ...
  - Theory groups, e.g. ILDG for Lattice QCD community
- DESY is a large Photon Science Laboratory
  - PS relies heavy on POSIX for data access
  - PS also data driven but using different tools
  - Look for synergies between the communities

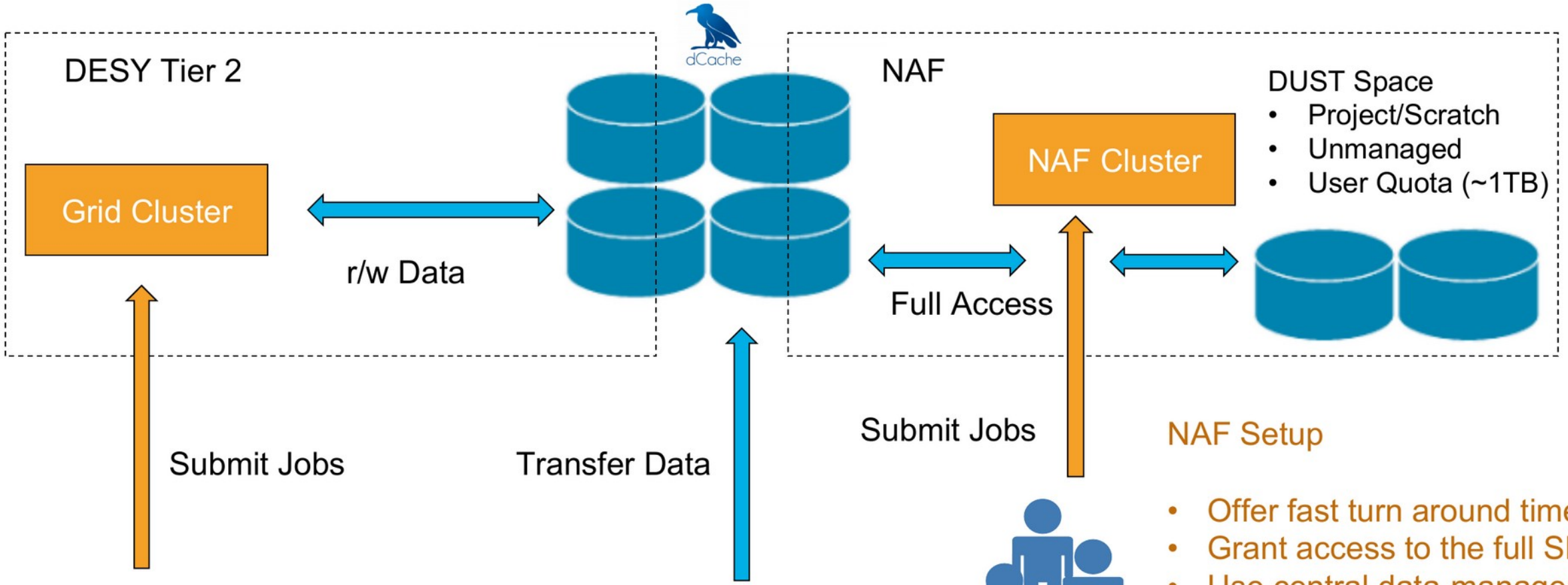


- Only the big four use full Grid workflows
- WLCG/HEP-only solution do not fit us
- E.g. many tools do not support XrootD

Paradigm: HEP Analyses are Data Driven

As Underlying Principle of the NAF

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



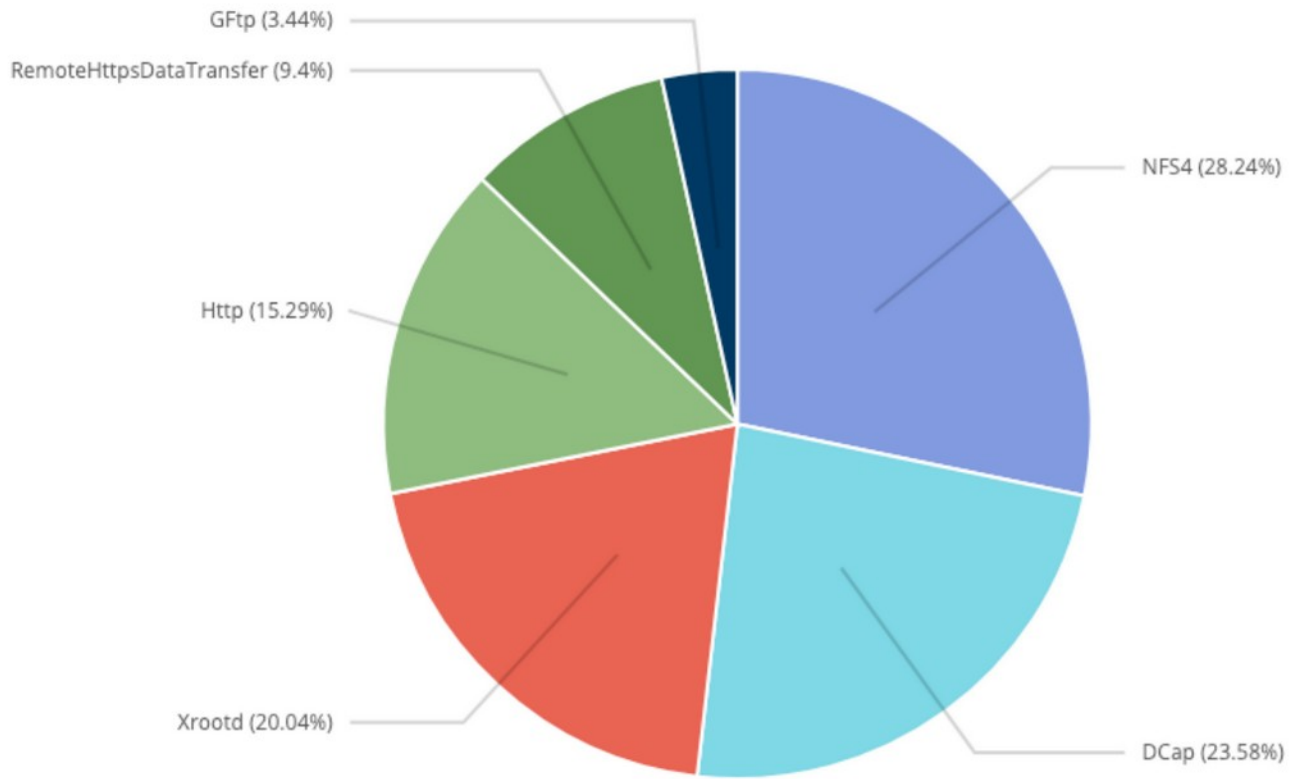
NAF Setup

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

Data Access

Protocol and Patterns

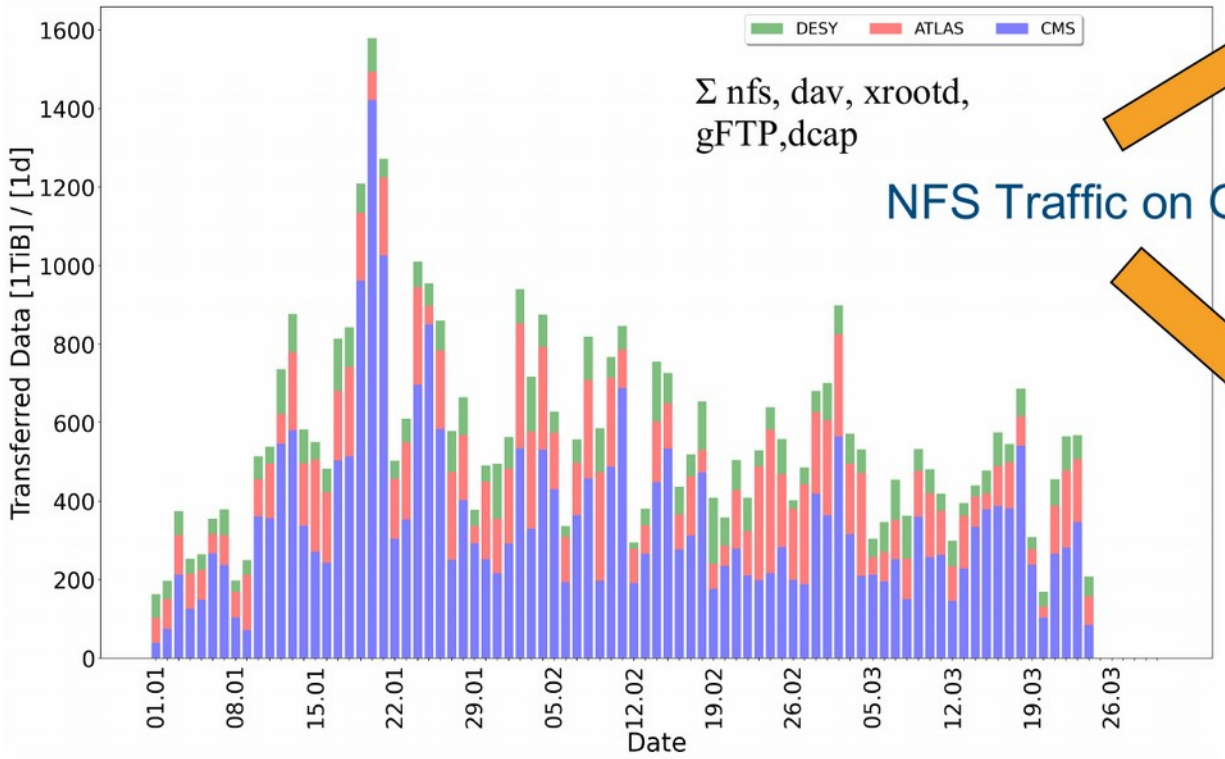
- Support all commonly available protocols for HEP
  - WebDav, XrootD & gsi-FTP, dCap as legacy
- Rely heavily on POSIX-like access
  - Storage infrastructure based on dCache and GPFS
  - Scalable NFS architecture across the NAF
  - Belle II group uses NFS almost to 100%
  - ATLAS users read (Rucio managed) LocalGroupDisk via NFS
  - Smaller groups use tools supporting only local files or industry formats like S3 or Hadoop
- Observed issues with NFS in the past, i.e. hanging clients (esp. SL6) → fixed with help of developers
- NFS makes testing of new tools very easy



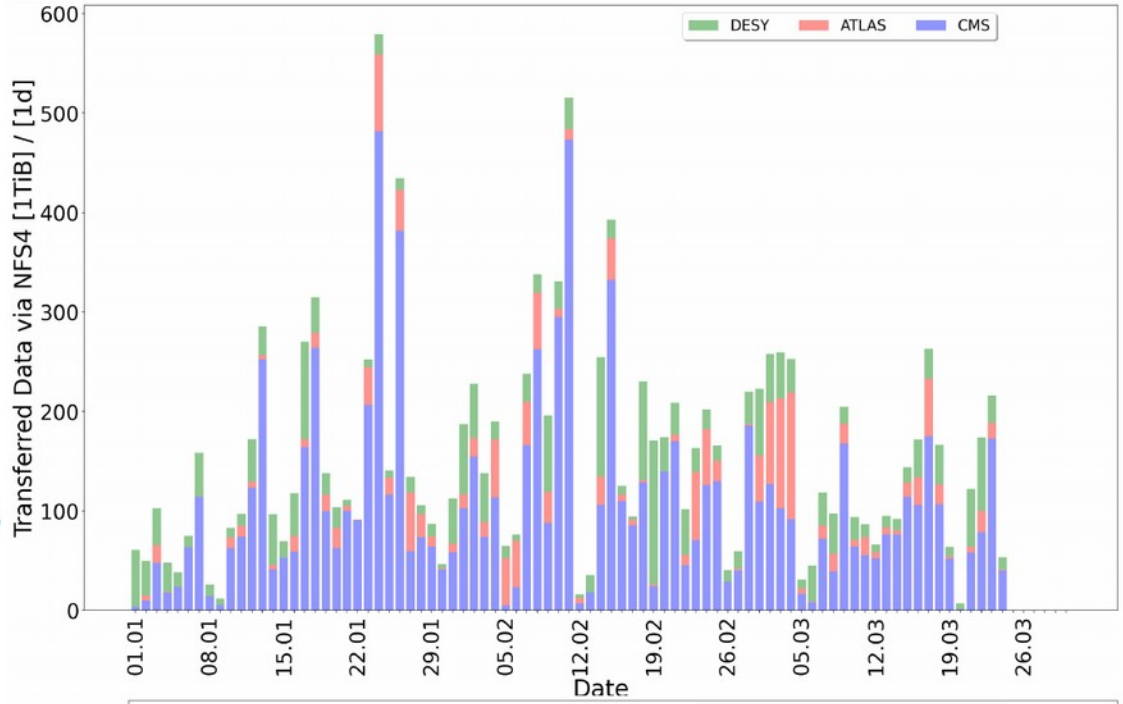
Data Usage on NAF

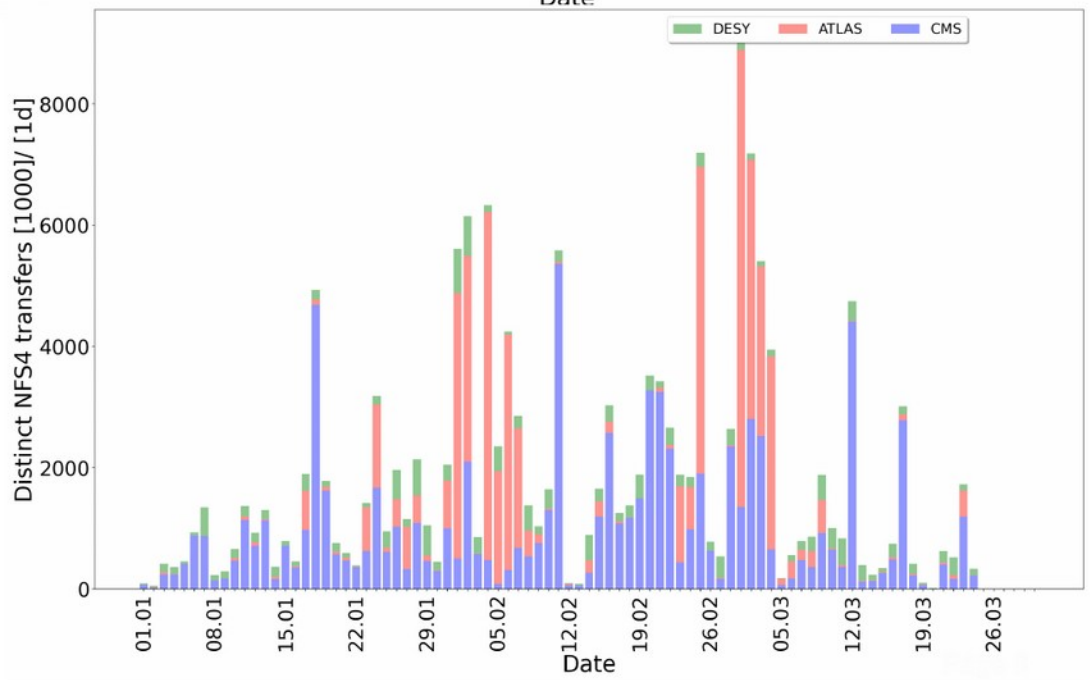
Focus on NFS Access

Whole Traffic on Grid SEs



NFS Traffic on Grid SEs





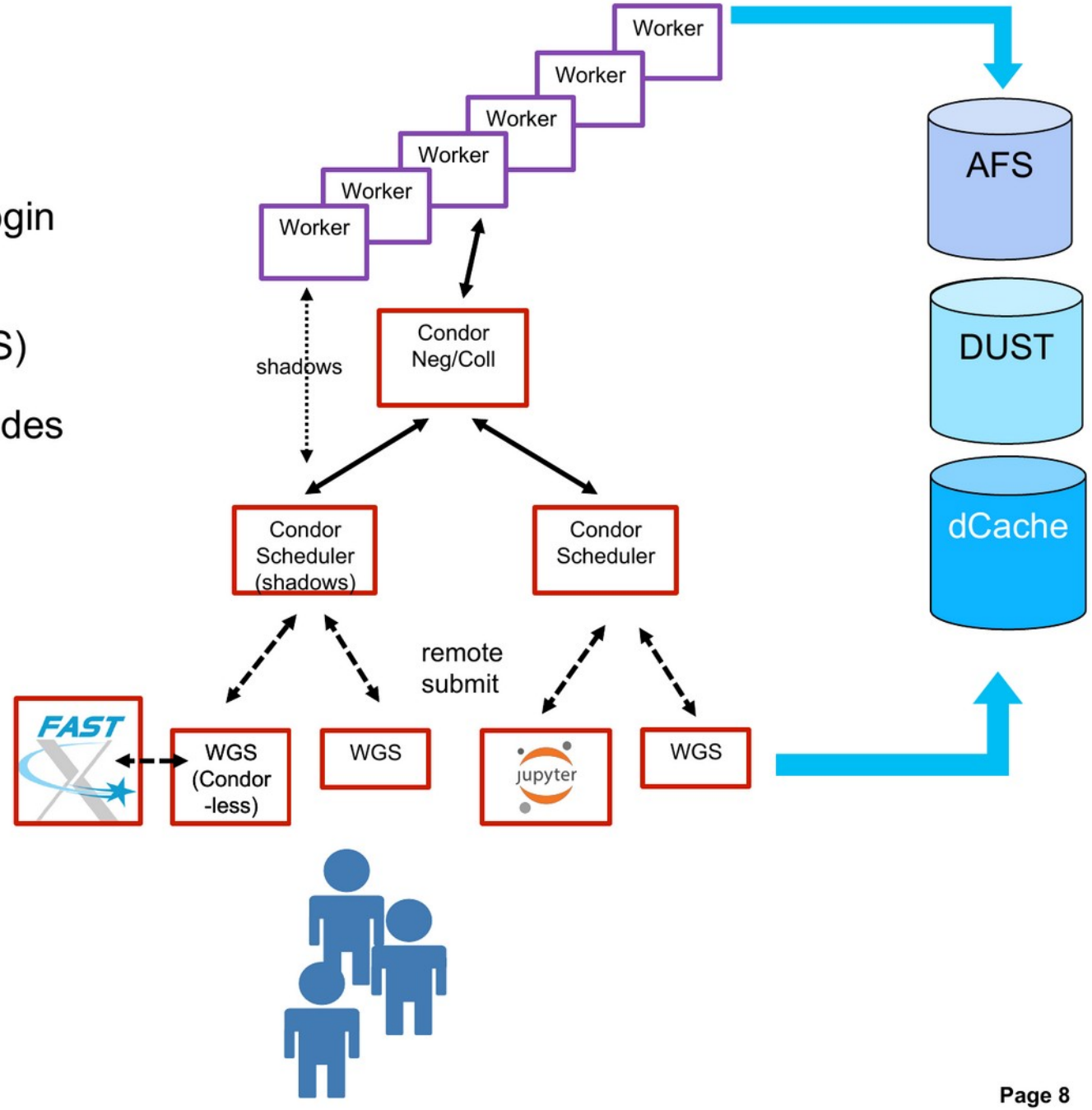
Whole Data transfers for 2022 till yesterday

User Access Methods

SSH to Dedicated Login Nodes

- Worldwide access through SSH to dedicated Login Nodes (WGS)
- Per group dedicated Work-Group Servers (WGS)
- Used as interactive work and job submission nodes
- Full access to the storage of the group
- HW for larger groups, VMs for smaller

- Remote submission from WGS
  - Dedicated scheduler nodes
  - Easier draining & rotation

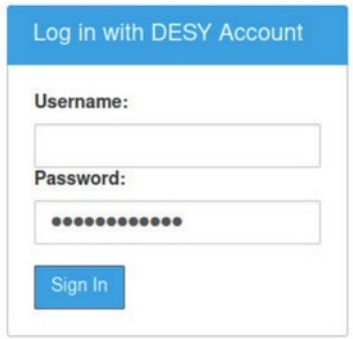


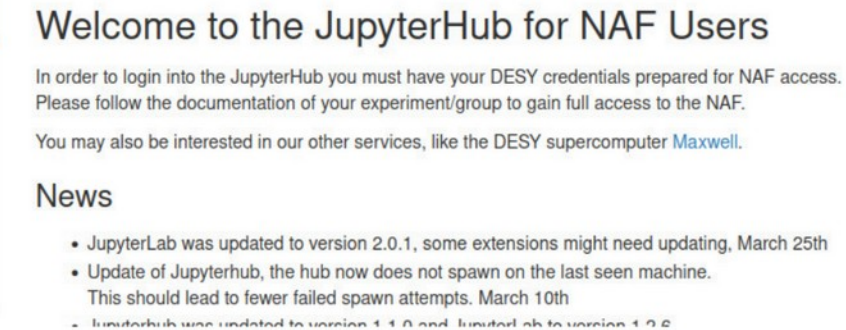
User Access Methods

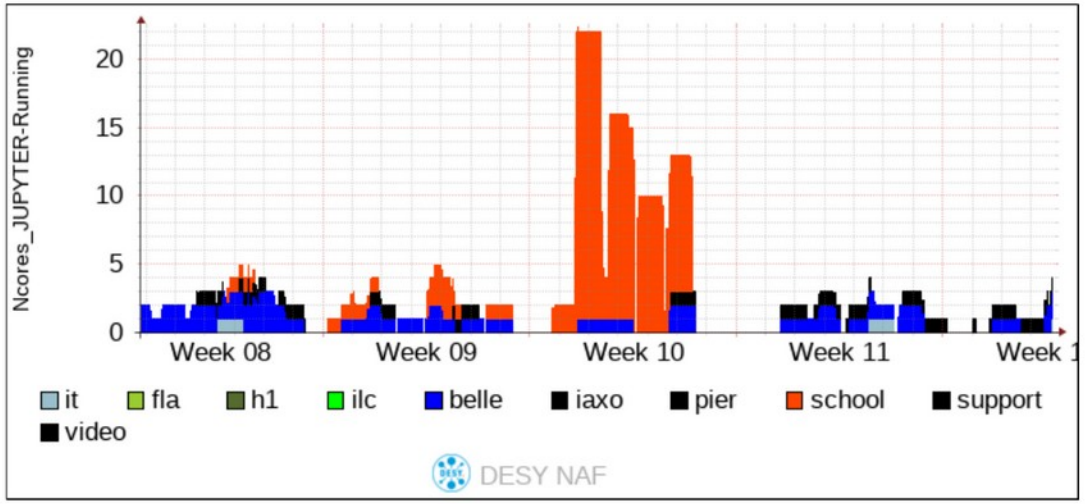
Jupyter Notebooks through Jupyter-Hub

- Jupyterhub integration in the NAF
  - Spawn notebooks as Condor jobs
    - Limit run time to 24h to avoid abandoned notebooks
    - Standard job limits (1core/2GB)
  - Full storage access
- Upstart responsiveness even more significant
    - Added *fast lane* for Condor notebook jobs
    - Reserve slots exclusively for notebooks (1/node)
- Route notebooks to nodes with smaller loads
- Notebooks as alternative to ssh sessions
  - World-wide access
- Limited adaption so far (more from Photon Science)

Deutsches Elektronen-Synchrotron DESY
A Research Centre of the Helmholtz Association







Authentication & Authorization

Login and Access to Compute and Storage

- Access to the NAF requires a DESY account
- Use account as login
- Access based on POSIX users:groups
  - Storage authorization via AFS token
  - Classic file ownership
- Jobs running under user IDs
- AFS access requires forwarding of Kerberos tokens
  - Implemented token renewal for long running jobs

- Grid part support classic VOMS access
  - Grid Condor Cluster
  - dCache based Grid SEs
- Watch closely the token development in WLCG and elsewhere
- Prototype infrastructure supporting tokens
- Federated identities requires support as well in the future (opportunity for easier user provisioning)
- *long* running tokenised user jobs realistic?

- See users preferring classic name-space based file access
- Migration to e.g. tokens: probably not easy to implement it due to users' learning curves

User Support and Governance

Interaction between Admins and Scientists

- NAF is not operated in a vacuum
- NAF Users Committee (NUC) to plan and decide on questions concerning NAF with members from
  - DESY IT
  - All supported experiments
- Discussions on
  - Size of additional procurements
  - Type of procurements
  - Shares between groups
  - Technology evolution
- Infrastructure support
  - Multi-level support within IT
  - Dedicated Queues in our Ticket system
- Experiment support
  - Organized by experiments
  - Mailing lists
- Good communication between experts from IT and experiments needed

User Support and Governance

Interaction between Admins and Scientists

- NAF is not operated in a vacuum
- NAF Users Committee (NUC) to plan and decide on questions concerning NAF with members from
  - DESY IT
  - All supported experiments
- Discussions on
  - Size of additional procurements
  - Type of procurements
  - Shares between groups
  - Technology evolution
- Infrastructure support
  - Multi-level support within IT
  - Dedicated Queues in our Ticket system
- Experiment support
  - Organized by experiments
  - Mailing lists
- Good communication between experts from IT and experiments needed