Evolution and Broadening of the National Analysis Facility at DESY

CHEP 2024

Christoph Beyer, Stefan Dietrich, Martin Flemming, Sandro Grizzo, Thomas Hartmann, Jürgen Hannappel, Yves Kemp, Joja Meyn, Johannes Reppin, Krunoslav Sever, Christian Sperl, Alexander Trautsch, Christian Voß

https://naf.desy.de https://docs.desy.de/naf

NAF Scope and History

>15 years of ongoing development

- Starting in 2007 as compute infrastructure for German HEP (ATLAS, CMS, LHCb, ILC) scientists
  - Complementary to Grid production
    - Direct user access
    - Fast job turn around
- Distributed between Hamburg and Zeuthen sites, AFS & LUSTRE storage backends, SGE cluster engine, PROOF I/O
  - Basic design aims reappearing today ~ remote/parallel/… fast data processing
- Lessons for the NAF
  - Avoiding being technology/implementation driven
  - Aiming for generic concept driven approach wrt. user needs

NAF Scope and History

>15 years of ongoing development

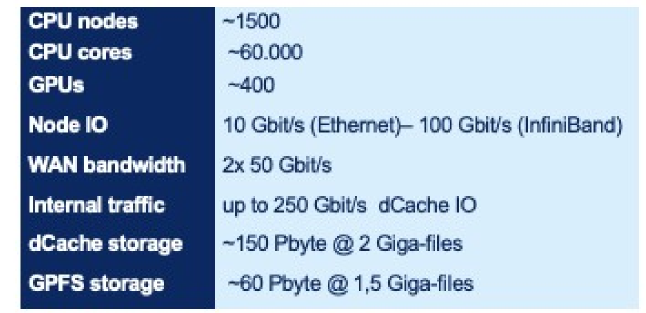
- NAF technology ongoing evolving from its initial technology implementation
  - Core Fundamentals
    - User focused
    - Data centric
    - Integrated Storage and Compute Infrastructure

$$i\hbar\frac{\partial}{\partial t}|\Psi(t)\rangle = \hat{H}|\Psi(t)\rangle$$

- NAF is a *whitelabel* Analysis Facility
  - Users today from a broad spectrum of communities end experiences
    - User support crucial
      - Any technology can only be auxiliary
    - Evolved beyond HEP
      - ***Interdisciplinarity*** central

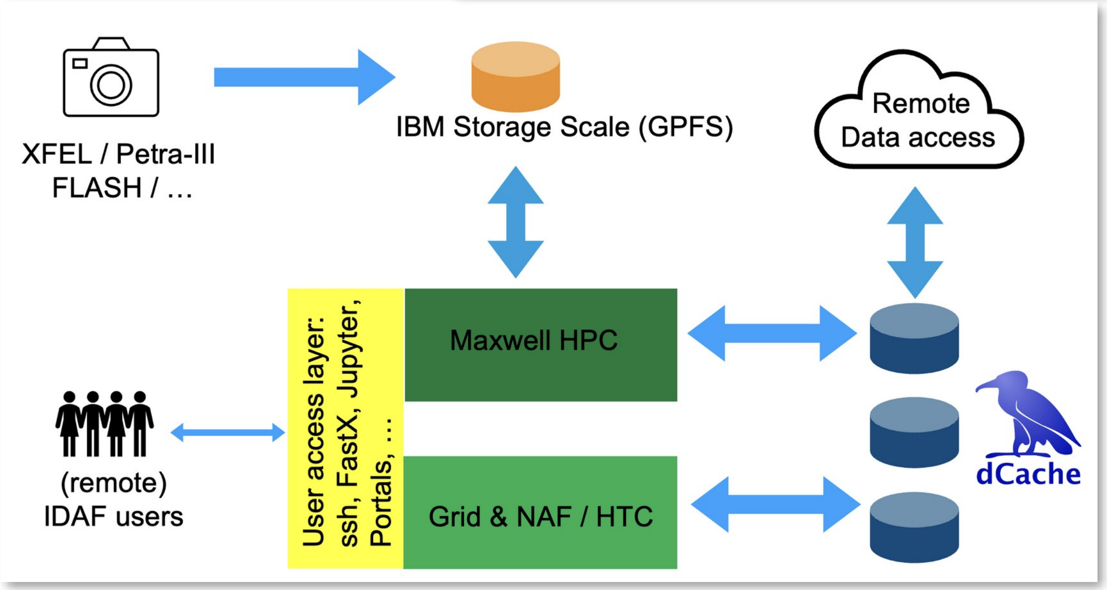
IDAF: Interdisciplinary Analysis Facility (IDAF)

Umbrella for the NAF (HTC) and Maxwell (HPC)



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

- NAF is part of the encompassing DESY "***Interdisciplinary* Data and Analysis Facility**" (IDAF)
  - NAF/HTC + Maxwell/HPC + Grid/HTC
- umbrella for analyses and production by DESY communities
  - **HEP**, **photon science**, **acclerator R&D**, **theory** & **operations**
- **Data centric**
  - Experiments data at the core
  - Local & global experiment data stored@DESY
  - Common namespaces
- Synergies between communities, operations and solutions whenever possible, special solutions when needed



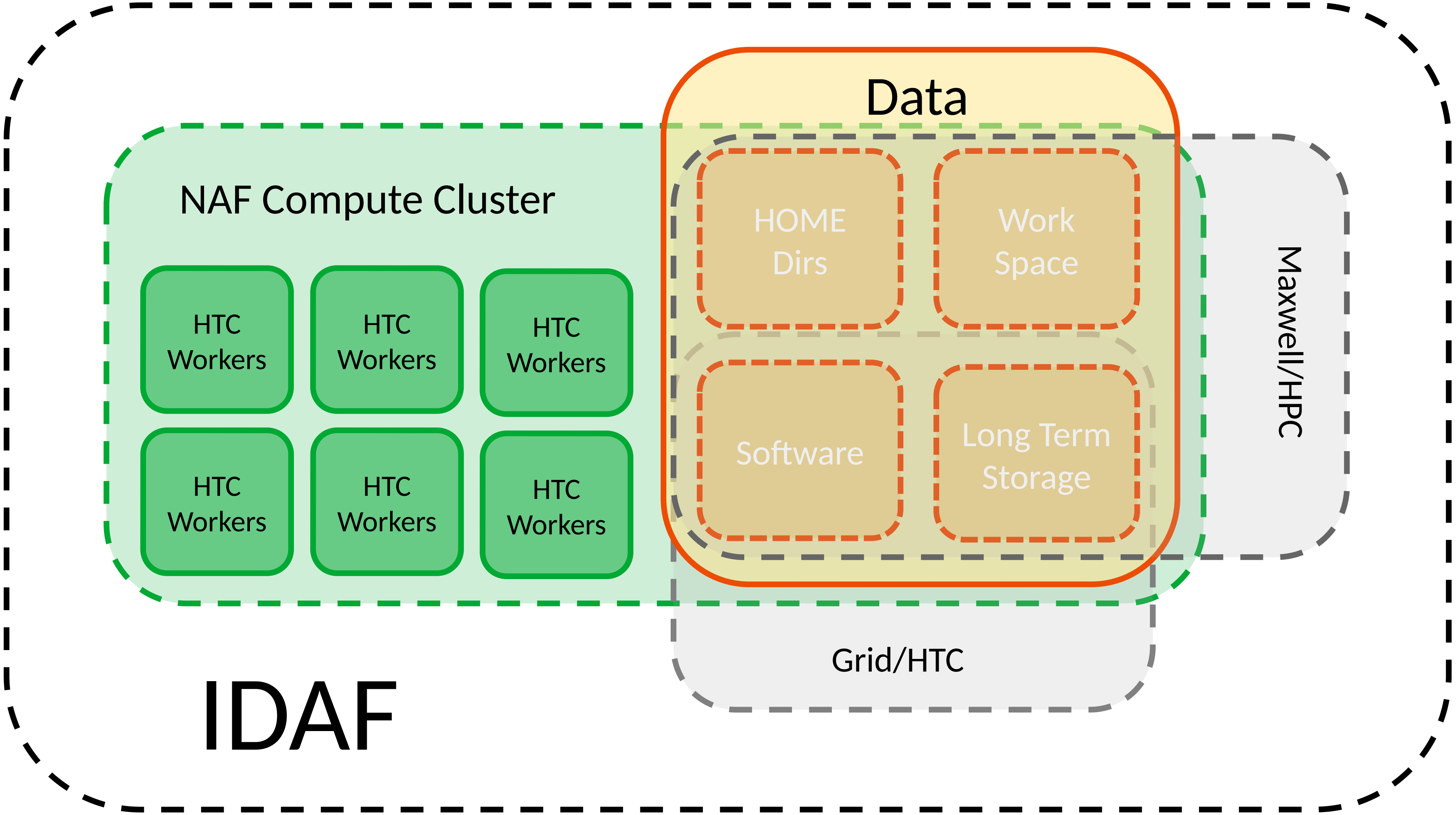
567. Serving Photon Science and HEP at the same facility

Christian Voss

24/10/2024, 11:00

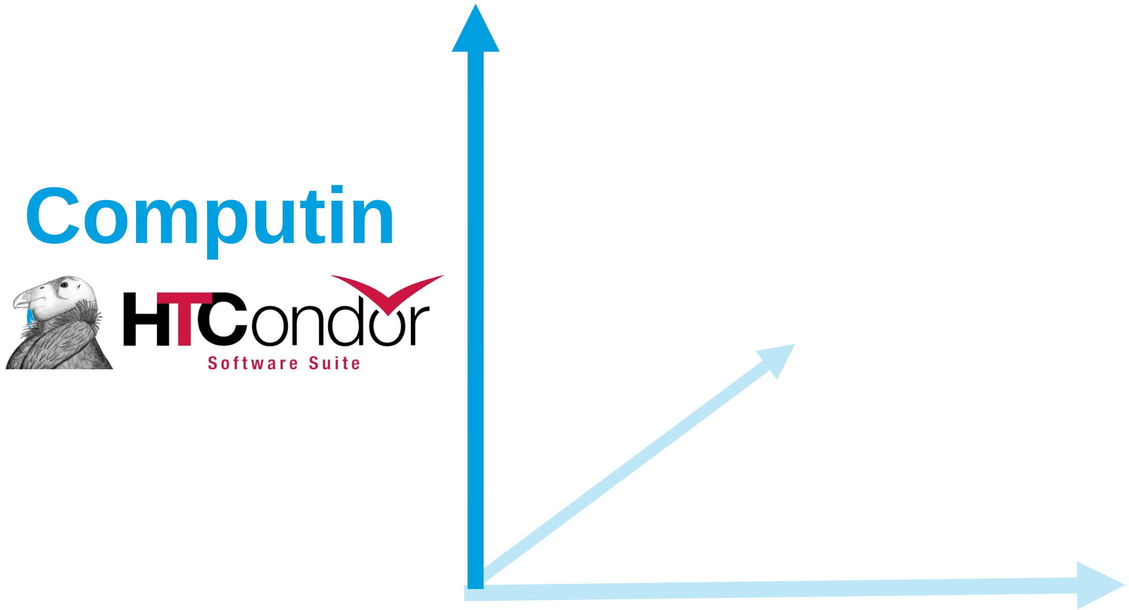
NAF in the IDAF

IDAF is Storage Centric



NAF (& IDAF) Dimension: Computing

number crunching scale out as HTC



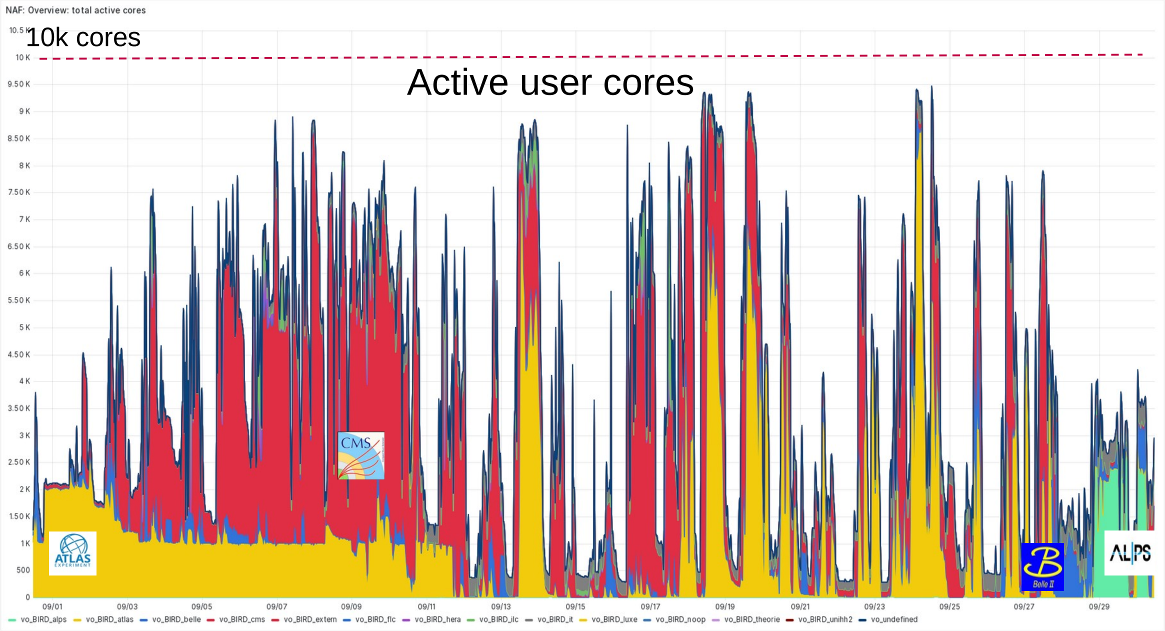
Computin

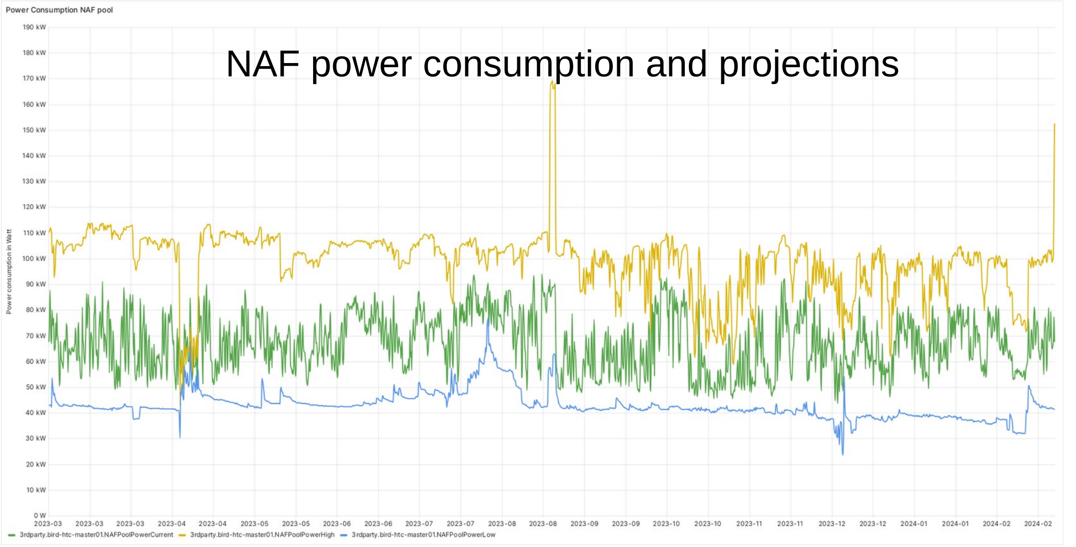
HTCondor
Software Suite

NAF Compute

Compute Cluster based on HTCondor

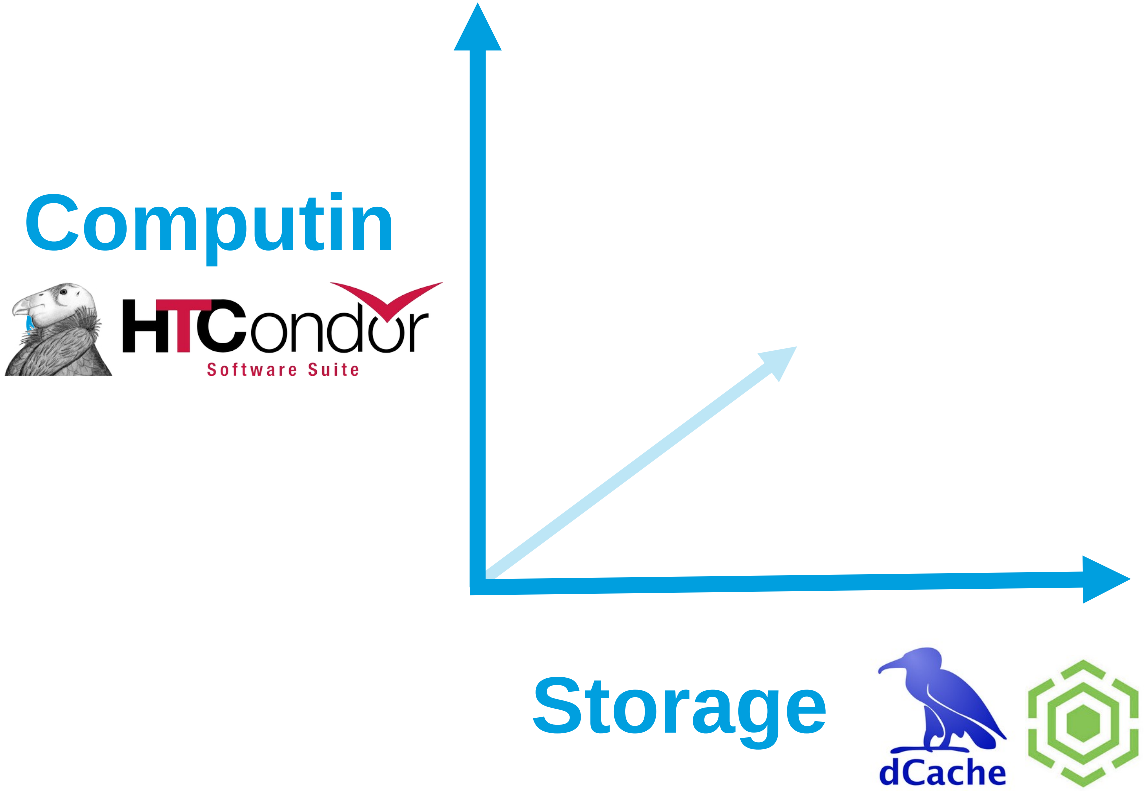
- HTC cluster on RHEL9 with HTCondor 23 with ~290 kHS23
  - Same puppet base as Grid HTC cluster
- Dynamic cluster utilization
  - Work day/week, conferences,…
  - Headroom wrt cluster utilization necessary to keep job start latencies low for interactivity
- Active power management
  - Horizontal job allocation for node load shedding





NAF (& IDAF) Dimension: Storage, Namespaces & I/O

Common namespace, scratch & LTS backends



Computin

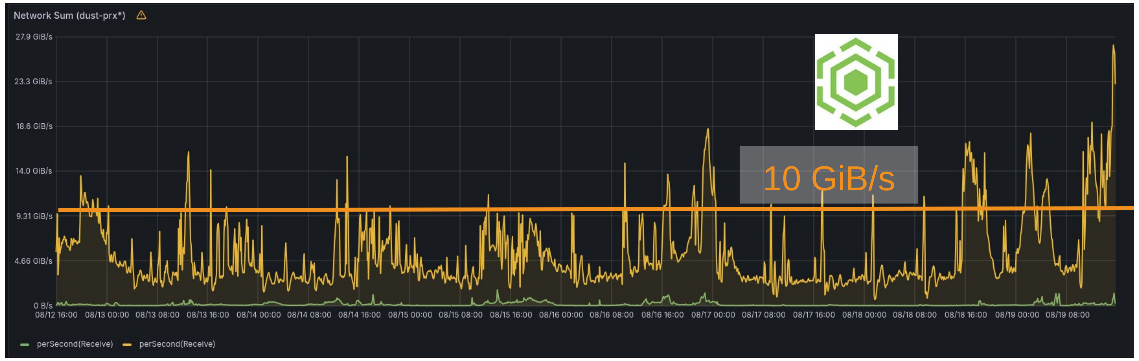
Storage

NAF Storage

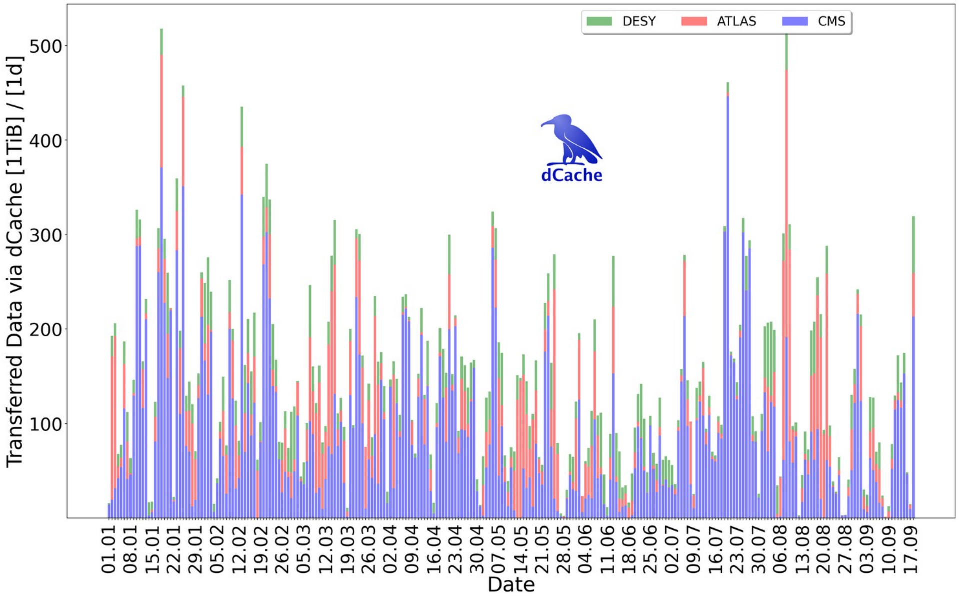
Tiered Storage Systems

- Local namespaces:
  - worker node local disks (rarely used by users)

- Cluster wide common namespace
  - HOMEs: AFS
    - low iops, remotely accessible
  - Software/Containers: CVMFS
    - scalable, globally available
  - Work Space: Spectrum Scale
    - fast I/O
  - Bulk Data: dCache
    - Long term storage, WAN~Grid I/O, Tape



throughput DUST scratch space



LTS data I/O from/to NAF

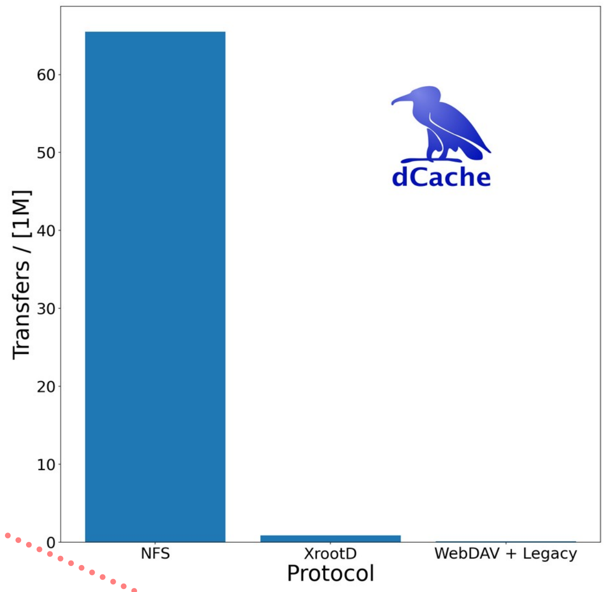
Application file interface: paths

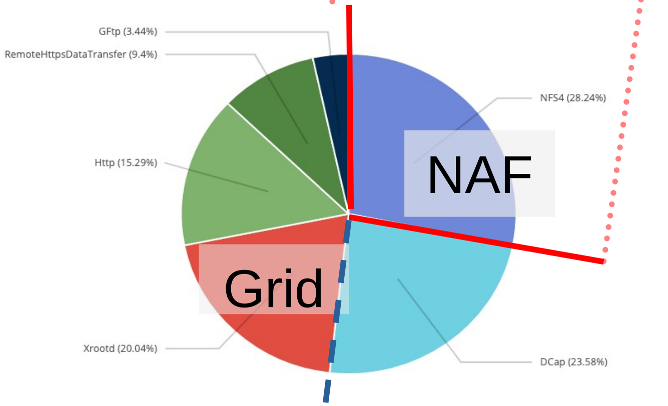
NFS as protocol of choice for file I/O

- User applications rely on paths for file addressing
  - Plethora of communities with common interface: POSIX (paths, ownership, calls…)
- NFS 4 protocol of choice
  - compatible with all NAF communities and storage backends (excl. AFS)

- Local access control on user/group level
  - Cluster-wide namespace via network file systems
  - Work In progress: Consolidation with Maxwell/HPC cluster towards a *common IDAF namespace*

NAF # Mega Transfers
in August-September





Ratio of protocol uses on
HEP dCache instances

Bridging the Monitoring Gap

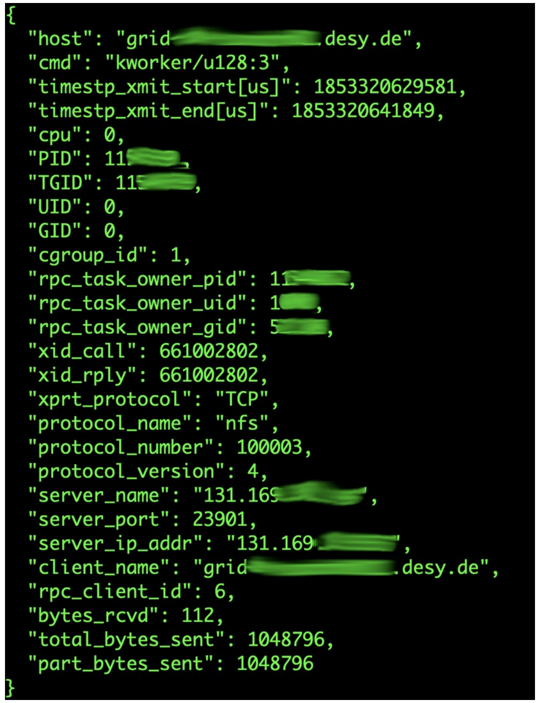
Storage and Compute Clusters as two sides of the same medal

- Before: monitoring separately for storage and compute clusters
- Not ideal for debugging
  - Compute Monitoring ↔ Storage Monitoring “*air gapped*”
  - Storage Side:
    - Worker as clients, User attribution only implicitly
  - Compute node side
    - Kernel NFS client in root NS &
    - http/xrootd application layer
- Consolidating monitoring into common view wrt. file requests
  - “*NAF Debug Mode*”

eBPF file request monitoring

Collecting kernel file info

- Ongoing project
  - Needed information on the compute workers:
    - On shared network fs: **what** file paths are opened by **whom** in **which** jobs
  - Should be agnostic wrt storage backend or LRMS (HTCondor, SLURM, K8s…
  - Querying the kernel via eBPF & reconstructing the paths
    - eBPF program extended to network/sockets,...



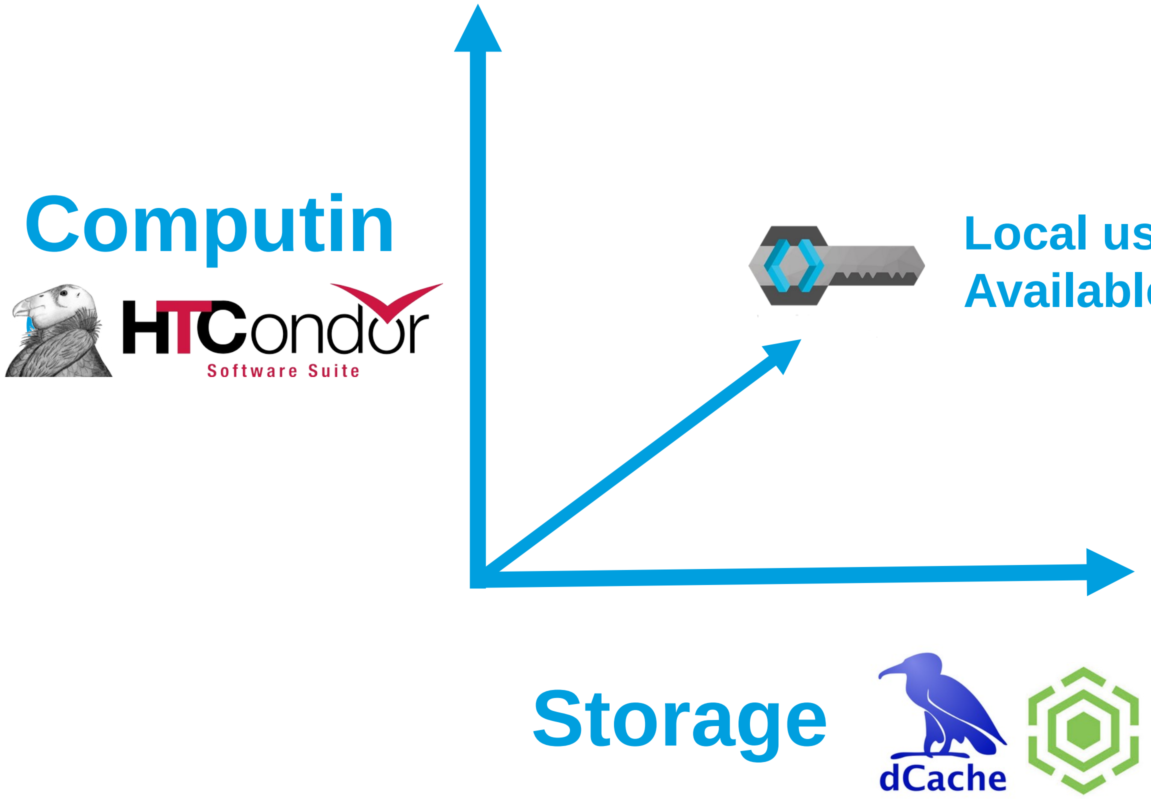
```
{
  "host": "grid[redacted].desy.de",
  "cmd": "kworker/u128:3",
  "timestp_xmit_start[us]": 1853320629581,
  "timestp_xmit_end[us]": 1853320641849,
  "cpu": 0,
  "PID": 11[redacted],
  "TGID": 11[redacted],
  "UID": 0,
  "GID": 0,
  "cgroup_id": 1,
  "rpc_task_owner_pid": 11[redacted],
  "rpc_task_owner_uid": 1[redacted],
  "rpc_task_owner_gid": 5[redacted],
  "xid_call": 661002802,
  "xid_rply": 661002802,
  "xprt_protocol": "TCP",
  "protocol_name": "nfs",
  "protocol_number": 100003,
  "protocol_version": 4,
  "server_name": "131.169[redacted]",
  "server_port": 23901,
  "server_ip_addr": "131.169[redacted]",
  "client_name": "grid[redacted].desy.de",
  "rpc_client_id": 6,
  "bytes_rcvd": 112,
  "total_bytes_sent": 1048796,
  "part_bytes_sent": 1048796
}
```

- Merge dCache storage events and NAF worker monitoring into common view
  - e.g., pseudo-query like ~

    “list the *workers where user Foo has file handles open with paths under /pnfs/desy.de/baz/… which are served by dCache pool dcache-baz314.desy.de*”

NAF (& IDAF) Dimension: Identities

Local identities with federated user access



Computin

Local user NS
Available to federated users

Storage

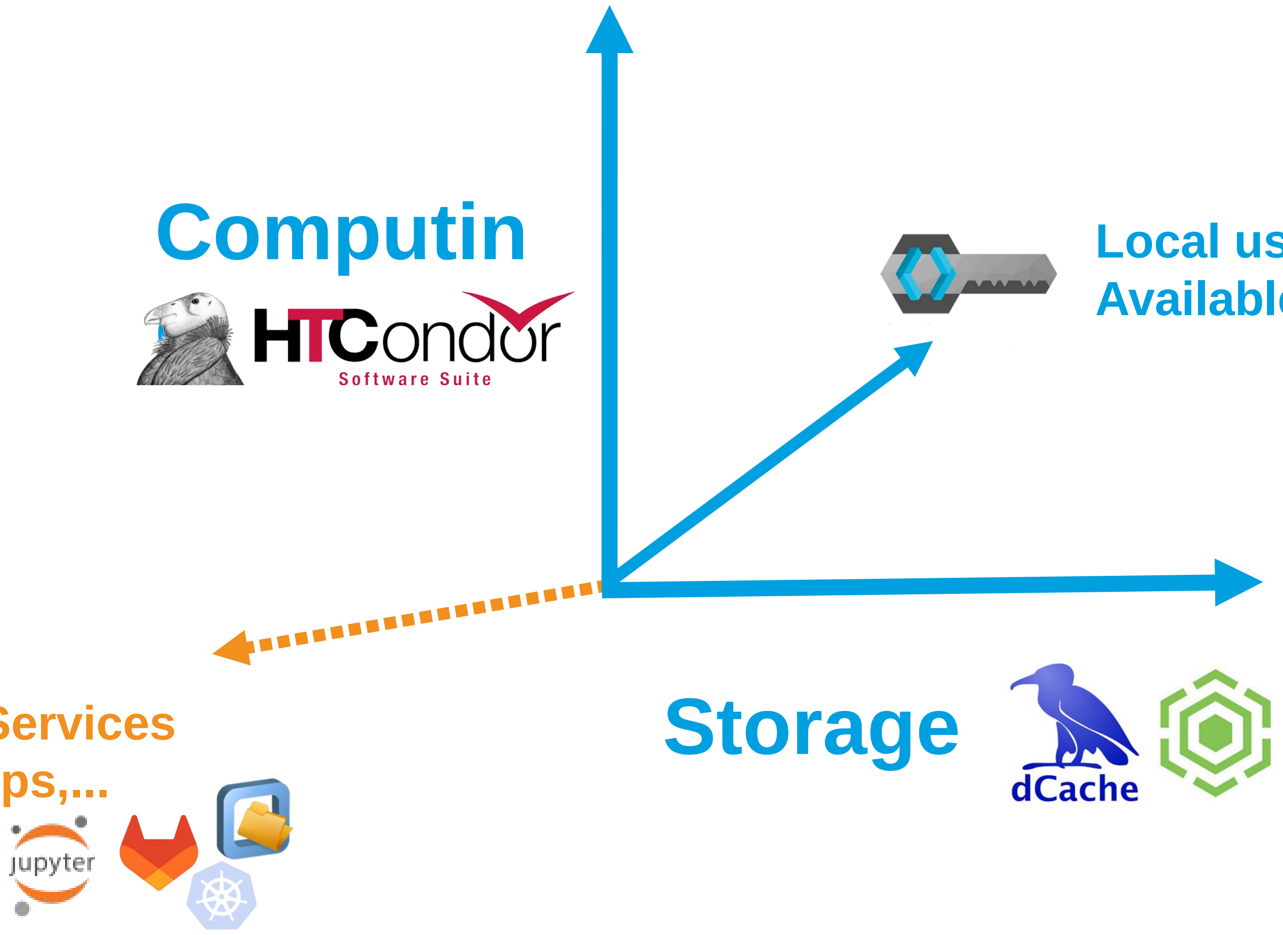
Federated Users – Local Identities

Keeping ID namespace intrinsic local

- In NAF 1.0 negative expriences with fully federated IDs/compute/storage
- ownership and capabilities wrt files is local
- Ongoing project:
  - enable remote users to board the NAF via federated AAIs
  - Open questions at fed level: IDs persistent wrt. data/files → clean up?
- WAN data ingress/egress via dCache instances
  - Grid workflows
  - non-HEP solutions/protocols

NAF (& IDAF) Dimension: Auxilliary Services

Local identities with federated user access



Computin

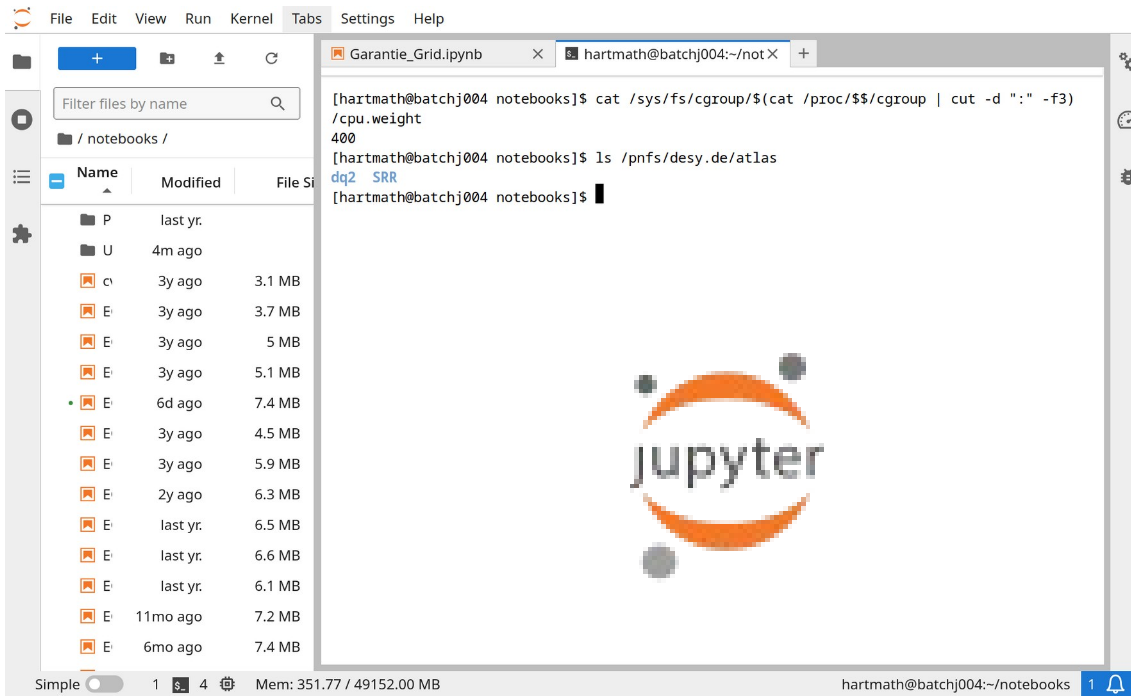
Local user NS
Available to federated users

Aux Services
DevOps,...

Storage

NAF User Interfaces

Support and Auxilliary Infrastructure for smaller Groups



Name	Modified	File Si
P	last yr.	
U	4m ago	
c\	3y ago	3.1 MB
E	3y ago	3.7 MB
E	3y ago	5 MB
E	3y ago	5.1 MB
E	6d ago	7.4 MB
E	3y ago	4.5 MB
E	3y ago	5.9 MB
E	2y ago	6.3 MB
E	last yr.	6.5 MB
E	last yr.	6.6 MB
E	last yr.	6.1 MB
E	11mo ago	7.2 MB
E	6mo ago	7.4 MB

```
[hartmath@batchj004 notebooks]$ cat /sys/fs/cgroup/$(cat /proc/$$/cgroup | cut -d ":" -f3)
/cpu.weight
400
[hartmath@batchj004 notebooks]$ ls /pnfs/desy.de/atlas
dq2  SRR
[hartmath@batchj004 notebooks]$ 
```

- Classic ssh
  - Workgroup server pooled behind remote submit nodes
- Jupyter Notebooks
  - Notebooks scale out as batch jobs
  - Memory heavy notebooks
    - Highmem slots
    - Spark/Dask overlay cluster
- FastX browser X UI
- Seeing more & more VSCode
  - Neat solution from Uni Bonn [https://indico.cern.ch/event/1386170/contributions/6118491/]

    investigating how we can adapt it to the NAF

Experiment Onboarding

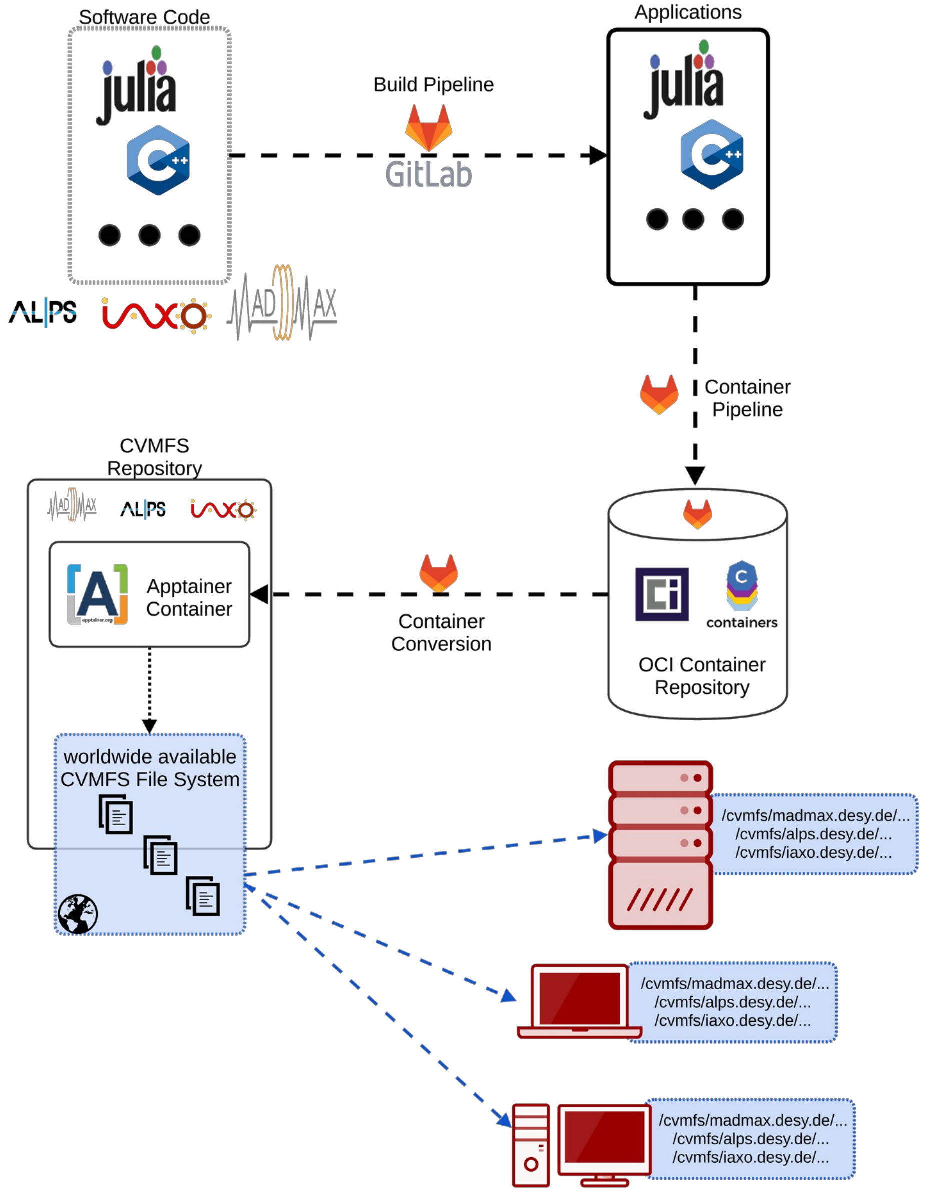
Support and Auxilliary Infrastructure for smaller Groups

- User support crucial
  - on/off-site experiments with limited manpower
  - Utilizing shared infrastructure and experiences for serving computing & storage needs
- NAF auxilliary services becoming more prominent
  - DevOps, sw distribution, collaborator authz,…

ALPS, MADMAX & IAXO

NAF as platform for Experiments

- ALPS choose NAF as platform for compute & storage
  - Collaboration with DESY IT to integrate axilliary services (many thanks @ **Rachel Wolf**!)
  - Sofware build and deployment as Gitlab pipeline to CVMFS
- Cooperating with MADMAX & IAXO to use & share platform, experiences,…
- User support & interaction
  - Significant gain from feedback
    - MADMAX as test user on EL9 preprod cluster discovered critical issue & commited fix (many thanks @ **David Leppla-Weber**!)



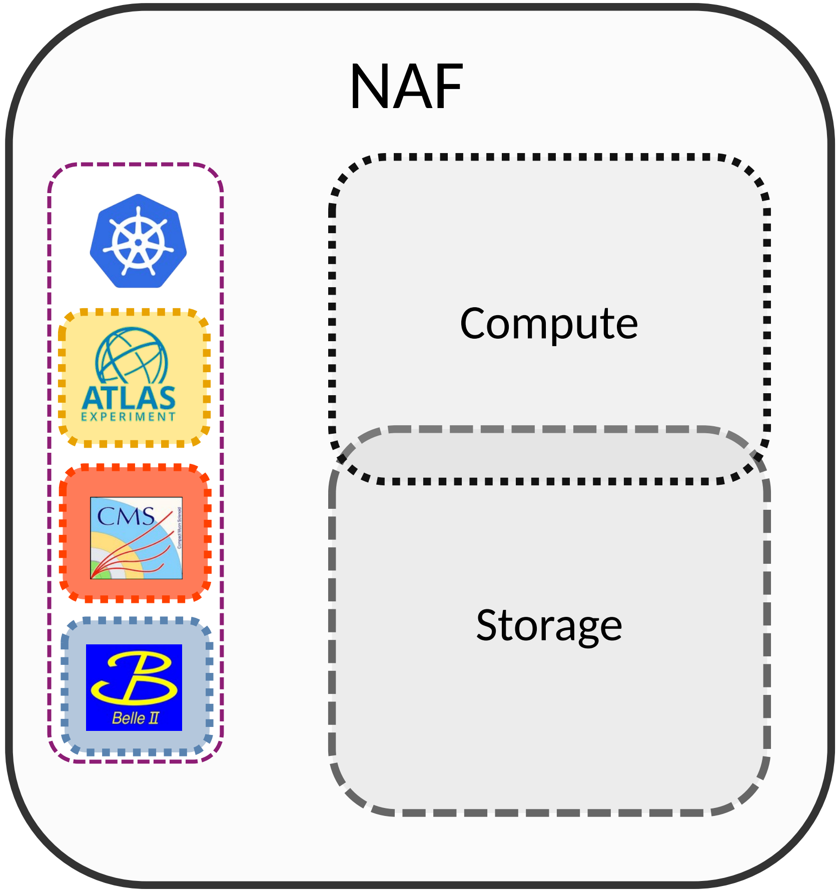
ALPS SW build & deployment workflow

K8s Group Addon Service

Ongoing work

- NAF *white label* ~ not specific to single group use case
    - NAF **interdisciplinary** → **I**_DAF_
  - Groups looking for deployment of own services,
    i.e., longer running, persistent applications

- Auxilliary Kubernetes cluster *under construction*
  - end point for *friendly groups* to deploy their flavoured services
  - **No heavy lifting**
    - CPU heavy workloads→ **scale out via batch LRMS**
  - **Strong isolation**, cap constraints
    - no native mounts → heavy data I/O? → **scale out via batch LRMS**
  - Commitment for dedicated service admin required
      - Subject to good security practices etc.



Czy masz pytania?

Operations Developments at the National Analysis Facility at DESY
The NAF as intergated Compute and Storage platform for diverse User Communities

485. Operations Developments at the National Analysis Facility at DESY (THU 25)

Christoph Beyer, Thomas Hartmann (Deutsches Elektronen-Synchrotron (DE)), Yves Kemp

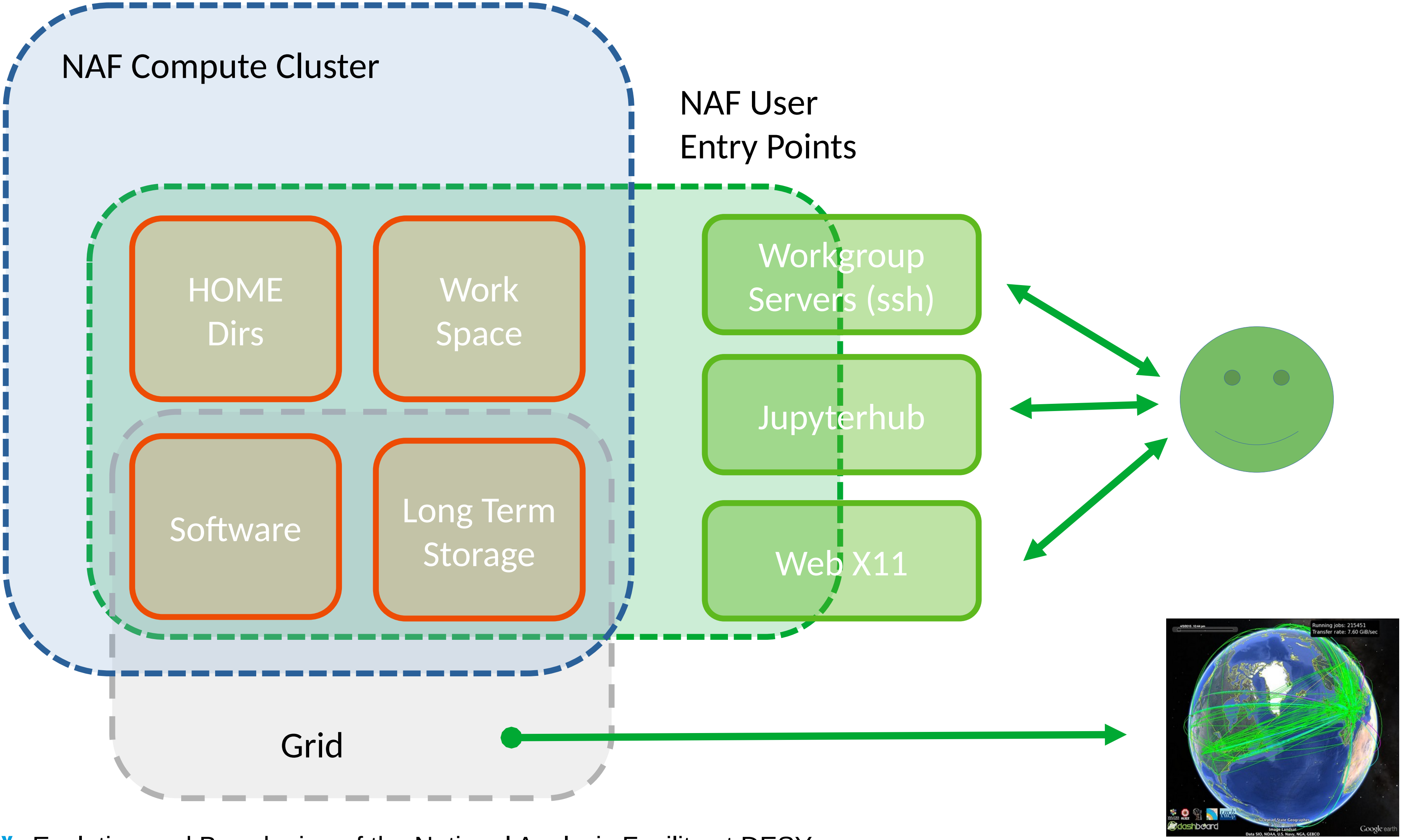
24/10/2024, 15:18

Track 9 - Analysis faciliti... Poster Poster session

Appendix

Integrated Storage and Compute

NAF is storage centric



Contact

DESY. Deutsches
Elektronen-Synchrotron

www.desy.de