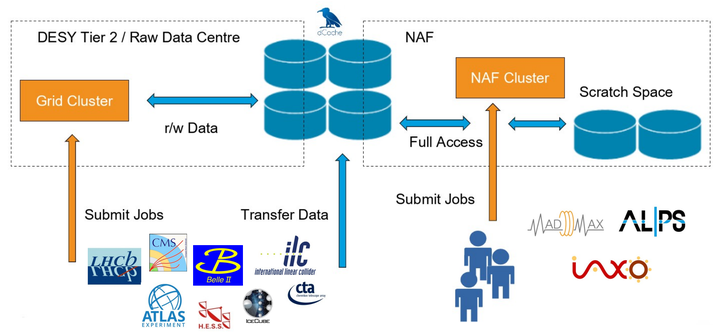
Operations Developments at the National Analysis Facility at DESY



The NAF as intergated Compute and Storage platform for diverse User Communities

Christoph Beyer, Stefan Dietrich, Martin Flemming, Elena Gapon, Sandro Grizzo, Thomas Hartmann, Joja Meyn, Yves Kemp, Johannes Reppin, Krunoslav Sever, Christian Voß

NAF User Communities and Requirements

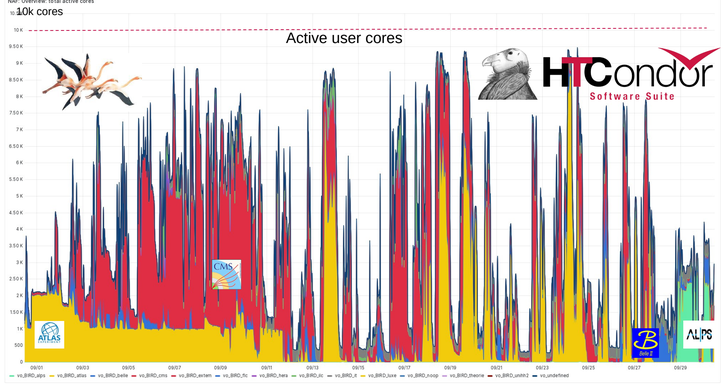
The NAF as *white label* Analysis Facility

- Broad range of scientific communities fromthe German particle physics groups
  - Ranging from large HEP to smaller astro-physics, weak interaction experiments, theory,... communities
  - Wide sets of requirements, tools, and experiences
  - User driven varying load patterns
- NAF has to be accessible to all its groups & users
  - Avoid focus on a single, experiment specific analysis-framework
  - User tools using paths for file access independent of protocol
    - NFS protocol of choice
      - single request protocols like http inefficient for various user I/O patterns
      - xrootd no established industry standard

User centric approach

- User support fundamental requriemend
  - User & group onboarding is manpower intensive
    - Accessibility hard requirement for AF
  - Regular user support sessions

Compute Core Utilization



Integrated Storage & Storage

- User workloads are data centric
  - NAF has to be data centric & storage first
  - Operated as integrated storage and compute facility
  - User & group authz and identities within the local ID namespace
    - Avoiding unresolved problems with respect to local vs. federated ID namespaces

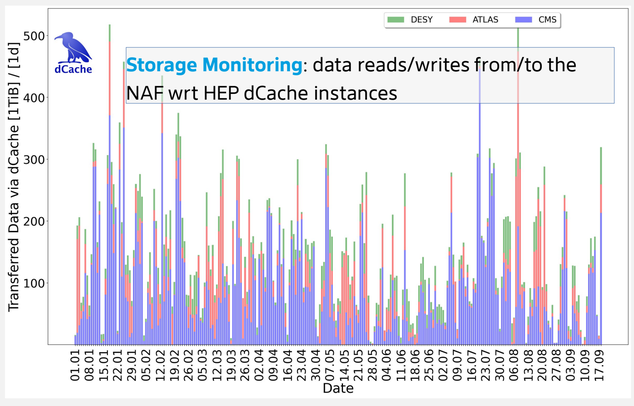
NAF Components

- compute cluster
  - ~300 kHS23 in compute resources managed by HTCondor
- Storage clusters
  - Common shared mount/path namespace
  - Multiple dCache instances for long-term storage with ~20PB in total
  - Shared GPFS filesystem as scratch space
  - AFS for home directories
  - CVMFS for group specific software distribution

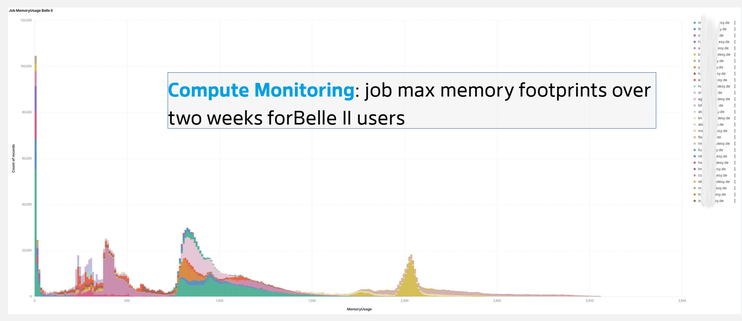
Consolidated Access and I/O Monitoring

Merging Storage and Compute Monitoring

- Treating compute & storage monitoring separately caused inefficiencies and friction
  - Compute cluster with N ~ 430 worker nodes
  - Storage cluster with M ~ 500 storage pools
  - N x M matrix of possible interactions (issues)...



Storage Monitoring: data reads/writes from/to the NAF wrt HEP dCache instances



Compute Monitoring: job max memory footprints over two weeks forBelle II users

eBPF based per User/Job file path I/O monitoring

- NFS client in root/kernel space
  - Job I/O appearing on storage pools only per compute worker
  - On storage pool user/job specific I/O details not easily available
- Need job/user request monitoring on path and socket level
- Zoo of storage systems specific solutions prohibitive

- Per worker node aggregating file request information under development (**Sandro Grizzo**)
- Project aims:
  - eBPF based: harnessing all necessary kernel information
  - By cgroup task list → LRMS agnostic: HTCondor, SLURM, K8s,...
  - Job requests as json for aggregation in Elastic Search
  - Intended to be extend to TCP/IP based I/O protocols

Compute Monitoring: job max memory footprints over two weeks forBelle II users

Auxiliary Services, Software Pipelines & Community onboarding

Group onboarding

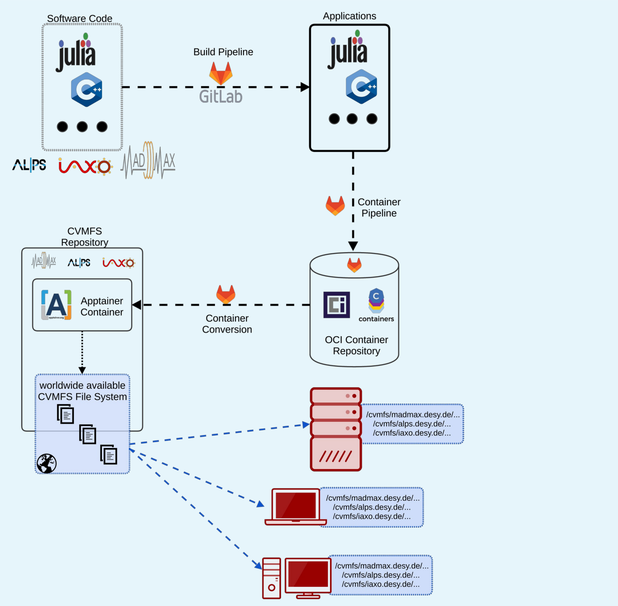
- Smaller experiments on/off-site physics focused
  - Limited personel for computing tasks
  - NAF offering common infrastructure for computing and storage needs including long term tape archival
  - NAF as end-to-end infrastructure for computing needs

Auxiliary Services

- Need for code development, management, deployment,...
- Interconnecting Jupyter, Gitlab and CVMFS within the NAF storage & compute
- Workload scale out onto NAF compute

Example: ALPS Software Deployment

- ALPS: on-site expriment searching for hidden sector photons
  - Experiment software in Julia developed primarily for Debian
  - Gitlab pipelines for building sw containerized as OCI compatible container images
- Scalable deployment via CVMFS
  - Dedicated ALPS CVMFS repo node
  - Gitlab Runner for deployment stage as Apptainer dir image



Experiments' SW Deployment Pipelines: job max memory footprints over two weeks forBelle II users

Future: Interdisciplinary Data Analysis Facility IDAF

Serving HEP and Photon Communities under one roof

- Currently Particle Physics and Research with Photons on independent compute clusters and storage backends
- Integrating all into the IDAF serving all communities at DESY

A common interface for HTC & HPC workloads

- NAF & Grid Hight Throuput Computing
- Photon science High Performance Computing
- IDAF as umbrella for HTC and HPC needs

Consolidating Namespaces

- Storage namespaces over NAF, HPC and Grid being consolidated

Deutsches Elektronen-Synchrotron DESY
Ein Forschungszentrum der Helmholtz-Gemeinschaft

https://naf.desy.de