

Status of Computing and Storage Activities at DESY

Annual Meeting of the BMBF-funded Research Compound 2024 March 26

Christoph Beyer, Andreas Gellrich, Yves Kemp, Andreas Haupt, Thomas Hartmann, Kai Leffhalm, Tigran Mrktchyan, Kilian Schwarz, Christian Voß

<https://dcache.org> <https://grid.desy.de> <https://naf.desy.de>



DESY HEP Computing Activities

Activities Fields

- dCache Storage Development
- Storage and Compute
 - ATLAS (HH + ZN)
 - CMS (HH)
 - Belle II Raw-Data Centre (HH)
 - Community Tools
 - Icecube, CTA (ZN)
- National Analysis Facility
 - DE HEP User Compute Cluster

dCache Development



DESY one of the main contributors to dCache storage system

- in house developers at DESY
- dCache instances serving wide range of use cases
 - 120PB instance with 20-40GB/s ingest (XFEL)
 - instance with 1.2e9 objects w. DB ~2.5TB (Photon)
 - file lifetimes <1s (DESY S&S)

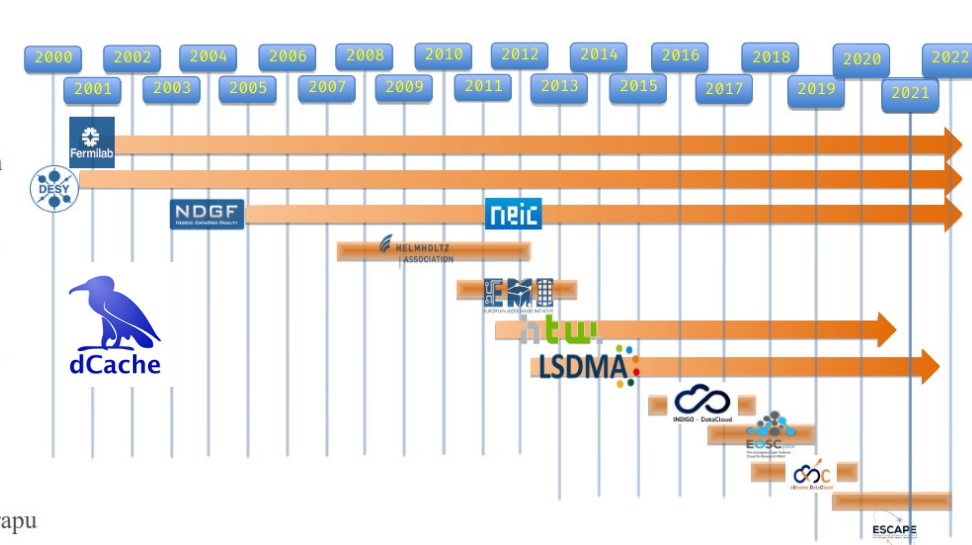
Recent Developments

- CTA tape backend integration
- Tape Rest API
- flexible metadata/namespace views
- storage event integration for alarming & self-healing load optimization
- Quotas, Tokens,...

- **DESY**
 - Svenja Meyer
 - *Paul Millar**
 - Tigran Mkrтчhyan
 - Lea Morschel
 - Marina Sahakyan

- **FermiLab**
 - Dmitry Litvintsev
 - Albert Rossi

- **NeIC**
 - Krishnaveni Chitrapu

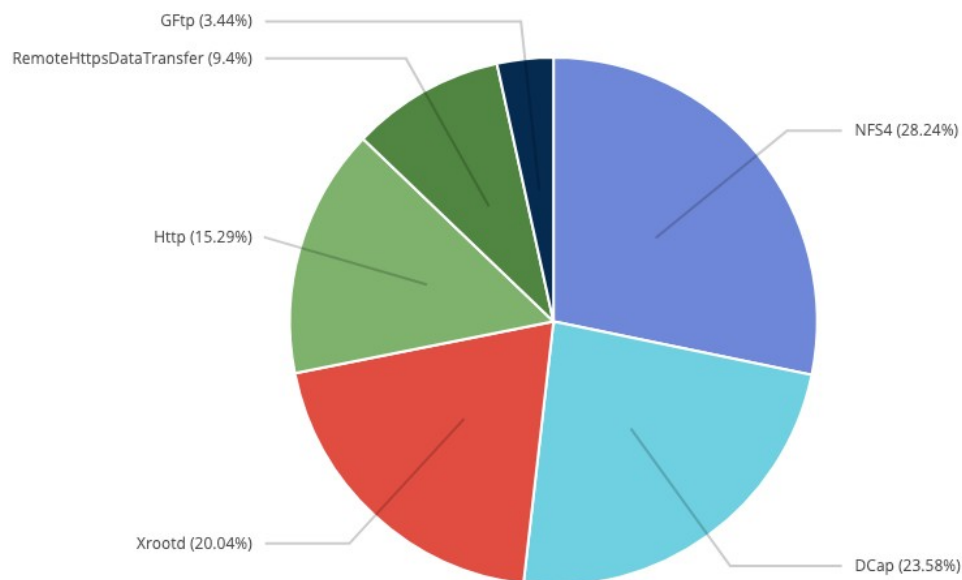


RSVP
dCache Workshop May.31-
June.1

Storage Operations

DOT Team HH (C. Voss), Team ZN (A. Haupt)

- about 29 PB HEP+Astro raw storage under management in HH+ZN
- dCache storage deployments at v9.2
- instance monitoring based on storage events
- unique challenges with Grid production and NAF user activities in parallel



VO	Raw Size
ATLAS DATADISK HH	2.6 PB
ATLAS SCRATCHDISK HH	0.2 PB
ATLAS LOCALGROUPDISK HH	3.4 PB
ATLAS DATADISK ZN	3.3 PB
ATLAS SCRATCHDISK ZN	0.05 PB
ATLAS LCOALGROUPDISK ZN	0.5 PB
Belle II prod+cal HH	2 PB
Belle local HH	0.4 PB
CMS Unmerged HH	0.2 PB
CMS Store HH	6.1 PB
CMS User HH	5.6 PB
ICECUBE ZN	2 PB
CTA ZN	2.2 PB
ILC HH	0.57 PB

Storage Operations

NAF Tape Support for User Data

Tape Support for ATLAS+CMS in Deployment at HH

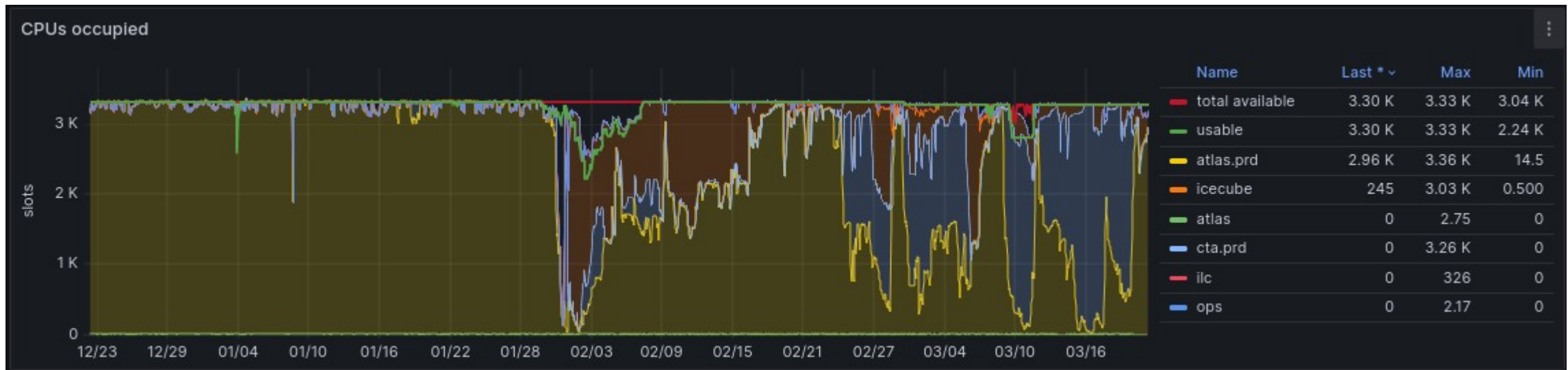
- (unique) user data had been single copies only
 - risk for data losses w/o Grid replicas
- tape support for ATLAS & CMS in preparation
- close collaboration with dCache developers (CTA – CERN Tape Archive)

Grid Computing and Middleware @ ZN

Grid Production HTC Cluster

Grid HTC Cluster serving ATLAS, CTA, ICECUBE

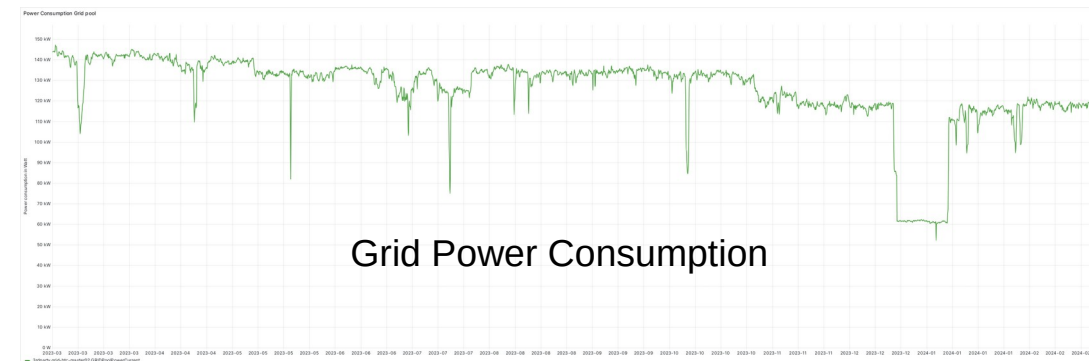
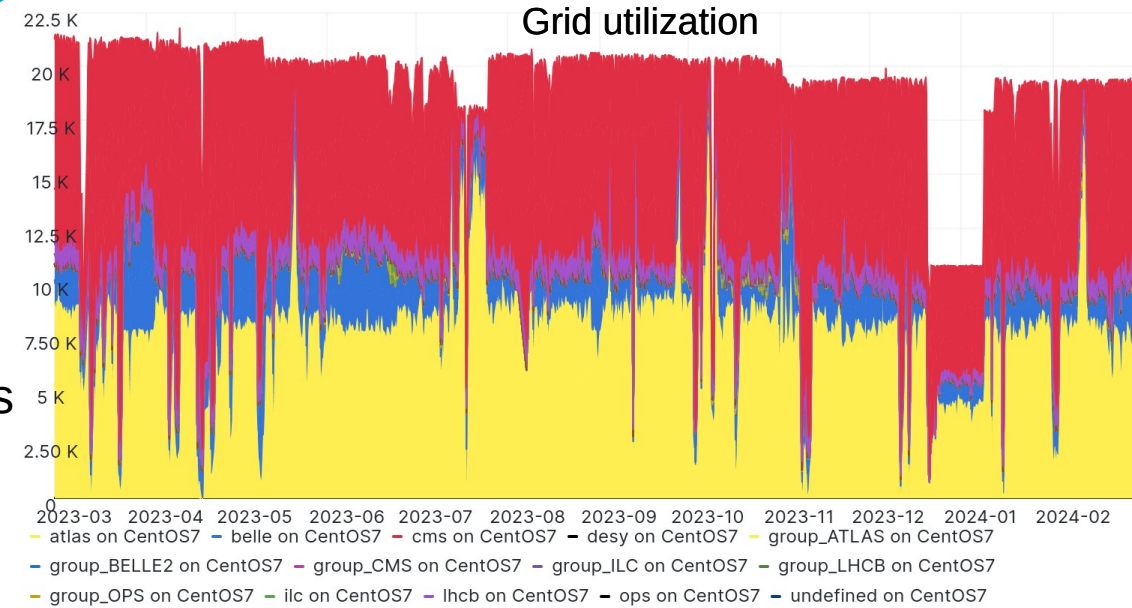
- ~68 kHS06 on 3300 cores
- Lost ~40% of resources to energy-saving switch-offs January 2023
- ICECUBE/CQTA GPU cluster with ~120 GPU cards



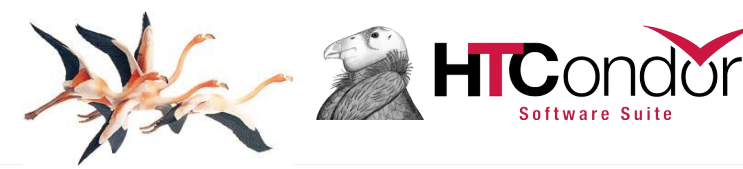
Grid Computing and Middleware @ HH

Grid Production HTC Cluster


- various CVMFS stratum-0 repositories
- EL7 cluster @ ~316 kHS06/23 on ~19.5k HT cores
 - EL9 pre-prod cluster @ ~100 kHS23 on ~4.5k HT cores
 - EGI middleware readiness questionable
 - Aiming for AUDITOR wrt. EL9 accounting
 - Accounting in the cpu:mem plane desirable(?)
- Energy shaping static (Christmas Load Shedding)
 - Planing wrt to “price coridor optimization” (~mystic HH site power pricing particularities)

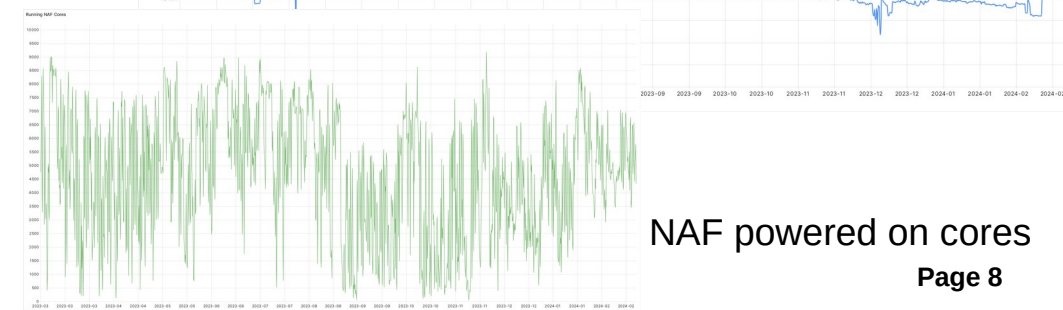
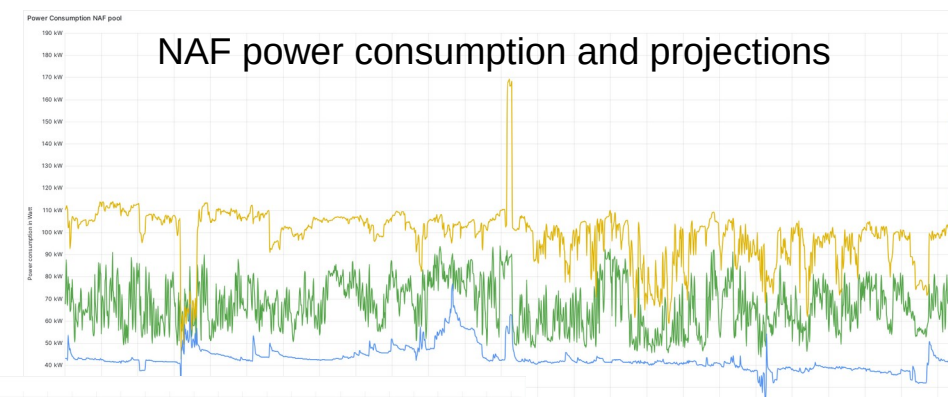
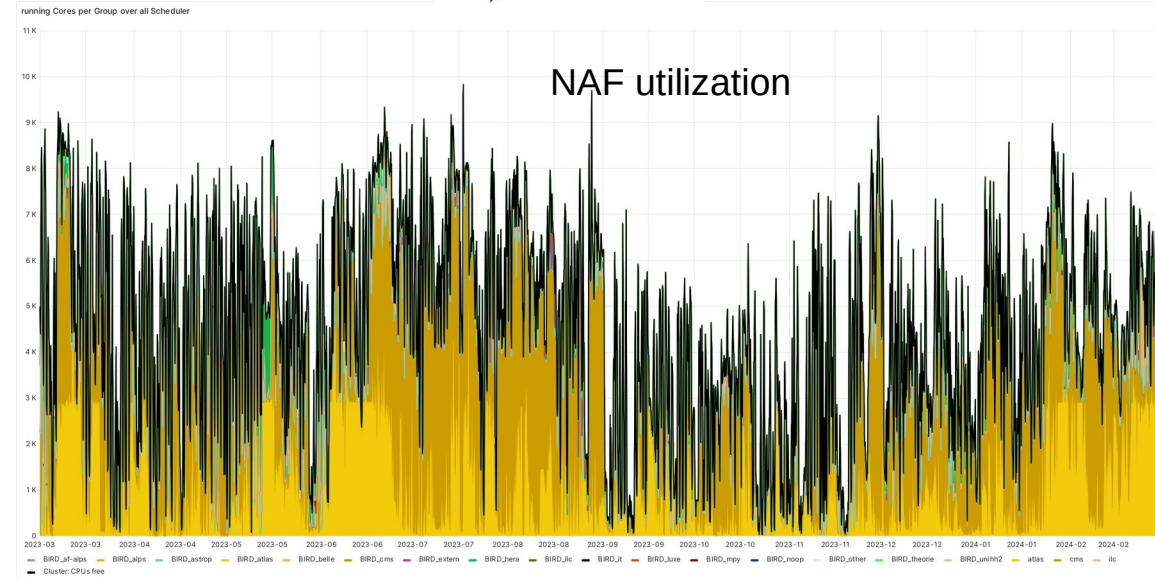


National Analysis Facility



User Analysis HTC Cluster (C. Beyer)

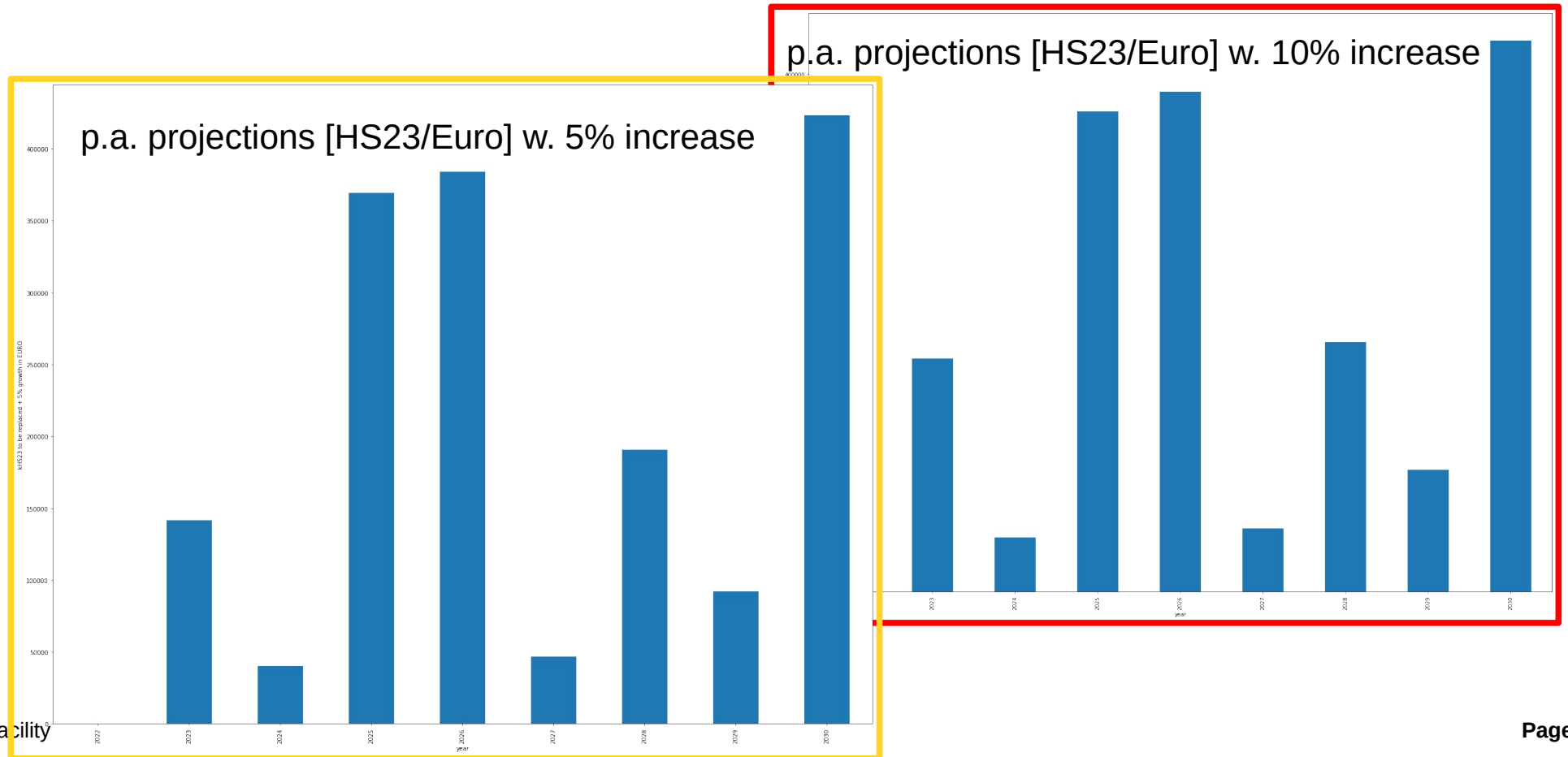
- HTC cluster for DE HEP communities with ~290 kHS23
 - EL9 pre-prod like Grid cluster
- shared scratch & grid storages available
- dynamic Jupyter notebooks 
 - User memory easily footprints >> 2GB
 - preparing highmem notebook options
- utilization more dynamic than Grid HTC cluster
 - updated scheduling approach for more aggressive node shedding
- rolled out MFA



Computing pledges

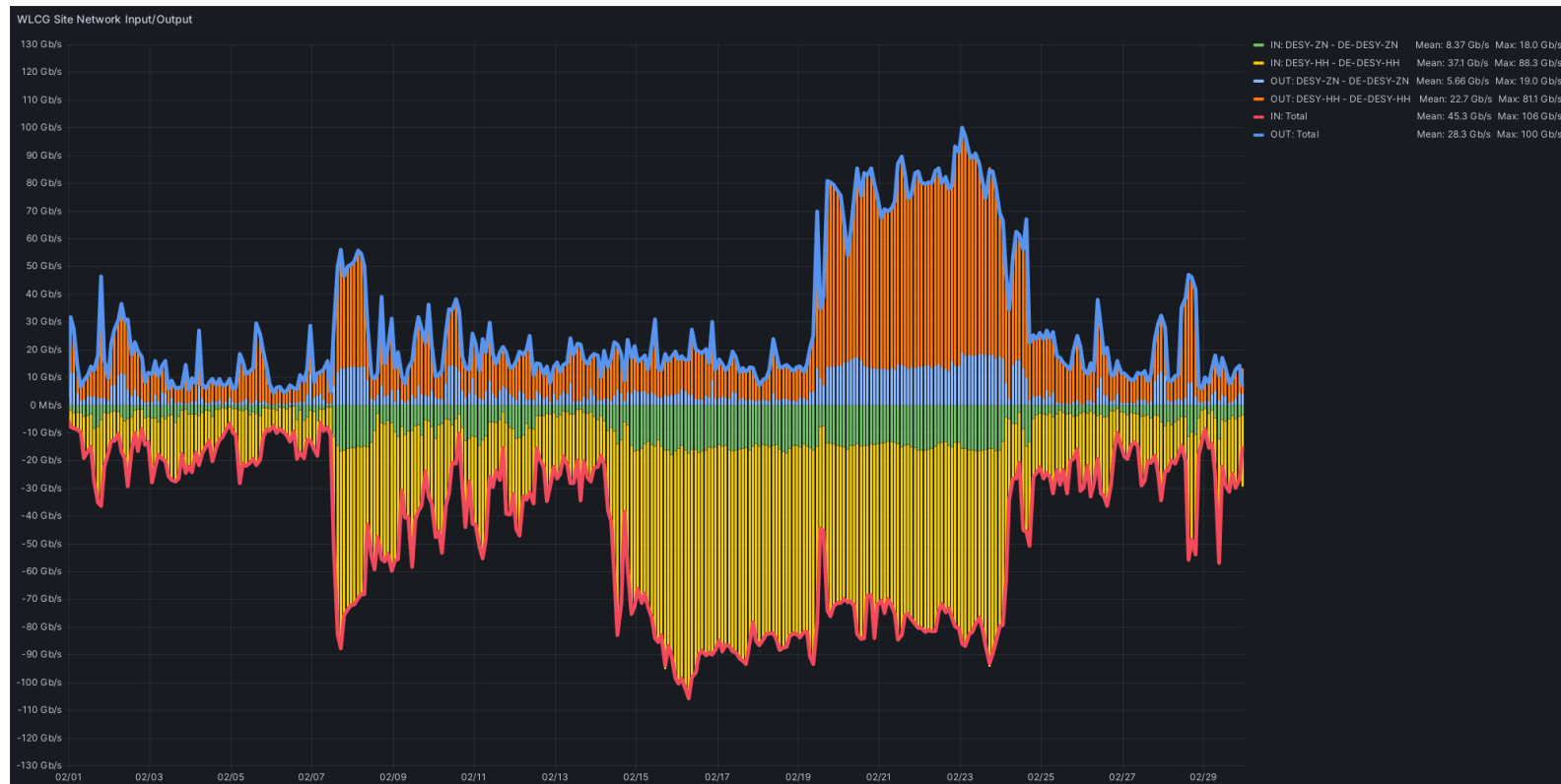
Investments / replacements over the next years

- Cost predictions for the next years
 - Projections: 5%, 10% yearly pleade [HS23/Euro] increase + replacement



Networking / DC24

- Successfully finished DC24, however bottlenecks (bandwidth saturation) observed
- HH & ZN nominally separate sites in WLCG & wrt layer-3
 - one entity wrt. layer-2 (hoping for firefly probe extension)



Networking

Status and ideas for future DFN connectivity

- DESY-HH connected via 2x 50GE atm.
 - Throttled via traffic shaping on DFN side
 - Upgrading to 2x 100GE just a question of an additional yearly fee
 - 400GE most probably desirable in the long run
 - DC24 led to (temporary) saturation
- DESY-ZN connected indirectly via DESY-HH
 - 4x 10GE (2x 20GE via port channels)
 - Defacto only 20Gbit/s reliably available due to technical limitations
 - Upgrade to at least 2x 50GE desirable (upgrade ready within the fiscal year)
 - CTA & Icecube desire / will profit from upgrade, too
 - DFN wants to get rid of 10GE links – yearly fees would even decrease

Future networking improvements

Further ideas for future DFN connectivity

- Going to e.g. 400GE will lead to many follow-up costs
 - New WAN routers, line cards, firewall, SFPs, ...
- Really invest in ridiculously expensive firewalls?
 - rethink the idea of „science dmz“, i.e. let LHCone traffic bypass firewall?
 - only realizable within LHCONE sites
 - peering with non-LHCONE/NREN sites not applicable

EGI concerns

What does the G in EGI stand for?

- 3 months before final EL7 retirement
 - ... there is no middleware available for EL9
 - ... there is no Python3-compatible (APEL) accounting client available
- Status of APEL unclear
 - Transition to token-only submission not implemented
 - Does already affect EGI VOs who transitioned away from x509 like Icecube
- Status of VOMS until Token-transition period has finished?
 - limited EL9 support, only
 - IAM can act as VOMS server - general replacement?
- EGI as middleware & service provider to be questioned generally!

General concerns

Unclear future usage pattern

- Future job distribution mix will likely look different
 - less on-site simulation, more analysis
- (dCache) storage scaling with increased on-site analysis job mix an issue?
 - Will future data access from NHR centres scale, as well?
 - ... in terms of latency, bandwidth, iops, etc.?
- No full redundancy of Grid/storage operations personal atm.

Summary

DESY HEP Computing Activities

- active contributor to dCache development
- operational experiences for a wide range of storage use cases
- DESY one of the biggest WLCG Tier2 centres worldwide
 - co-hosts non-WLCG storage and compute resources for experiments like XFEL, BELLE-II, CTA, Icecube ...
- DESY has provided highly available and reliable services in the last years
 - We do not plan to change this in future ;-)

Contact

DESY. Deutsches
Elektronen-Synchrotron

www.desy.de