GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Transformation of the university-based WLCG-Tier-2 structure in Germany: technical aspects and status of the implementation

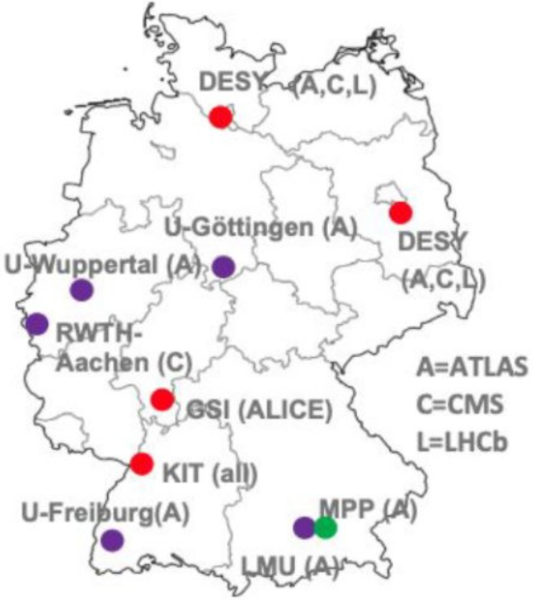
Sebastian Wozniewski
on behalf of the German ATLAS&CMS Joint Computing Project

WLCG/HSF Workshop, Hamburg – 13.05.24

ATLAS & CMS Computing in Germany

- Major accumulation of storage & compute resources at **Helmholtz Centres (WLCG Tier-1 GridKa at KIT and Tier-2s at DESY)**

- Additional contributions to ATLAS & CMS computing via **university-based Tier-2 centres**
    - run by local research institutes affiliated with the respective LHC collaboration and supported by the universities
    - operating cost: states; hardware cost + add. personell: project consortium of the federal government



Helmholtz Centres
Max-Planck-Institute
Universities

GridKa 25%

DESY / MPP: 32%

Universities: 43%

SPONSORED BY THE

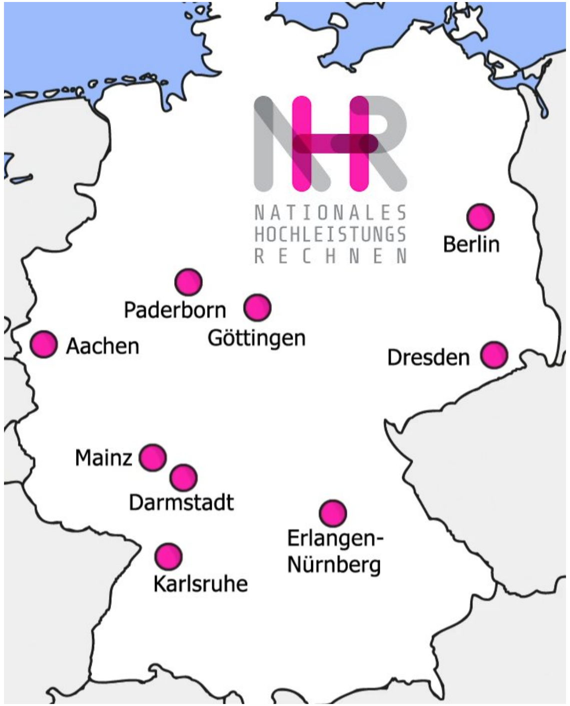
Federal Ministry of Education and Research

Funding recently approved for the coming three years

National High Performance Computing (NHR)

- association of large, university-based, **multi-disciplinary HPC centres** (independent from supranational Supercomputing Centres in Jürich, Stuttgart, Munich)
- founded in 2020
- funded by the federal government and most of the federal states (~60 Mio. Euro / year)
- **provides compute time to university research groups** passing the review of a scientific committee (applications every year)

→ Our perspective for a sufficient and sustainable provision of compute power towards HL-LHC instead of university-based Tier-2 centres (energy & resource usage efficiency, synergy)



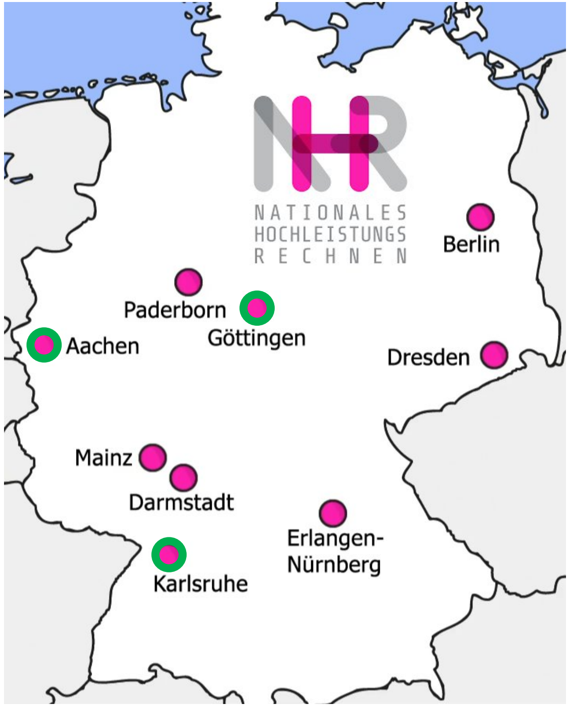
National High Performance Computing (NHR)

- association of large, university-based, **multi-disciplinary HPC centres** (independent from supranational Supercomputing Centres in Jürich, Stuttgart, Munich)
- founded in 2020
- funded by the federal government and most of the federal states (~60 Mio. Euro / year)
- **provides compute time to university research groups** passing the review of a scientific committee (applications every year)

→ Our perspective for a sufficient and sustainable provision of compute power towards HL-LHC instead of university-based Tier-2 centres (energy & resource usage efficiency, synergy)

3 NHR centres on a campus with WLCG Tier-1 or Tier-2 centre

→ local expertise + simplifies transition

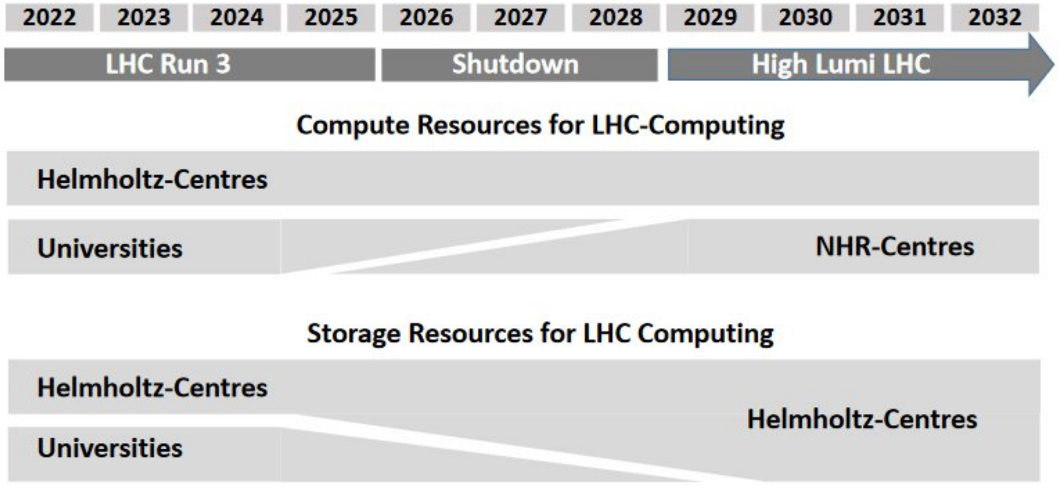


Transformation towards HL-LHC

Gradual transition from university-based Tier-2 centres **to NHR (CPU) and Helmholtz-Centres (mass storage)** towards beginning of HL-LHC, i.e. 20% per year.

Local ATLAS/CMS groups keep supervising the NHR resources and apply for funding from federal government for **dedicated personnel**:

- ATLAS:
  - NHR@KIT - Freiburg group
  - NHR@Göttingen - Göttingen group
- CMS:
  - NHR@KIT - Karlsruhe group
  - NHR@Aachen - Aachen group



Strategy paper by KET from 2022:
https://www.ketweb.de/sites/site_ketweb/content/e199639/e312771/KET-Computing-Strategie-HL-LHC-final.pdf

HPC clusters in the WLCG

- Various cases of HPC usage over the past years, e.g. Perlmutter (Berkeley), SuperMUC (Garching), FORHLR2 (Karlsruhe), Piz Daint/Alps (Lugano), Vega (Maribor), Karolina (Ostrava)...

- Often restricted to certain workflows / job types due to boundary conditions not meeting all WLCG needs, but still valuable contributions of compute power,

  e.g. highly-parallelisable simulation jobs can be used to fill an entire node if required (whole-node scheduling) and are less I/O-intense requiring no high-bandwidth data storage access.

- **For a regular usage of NHR resources we need to avoid such restrictions. All job types should run efficiently!**

R&D Project IDIUM

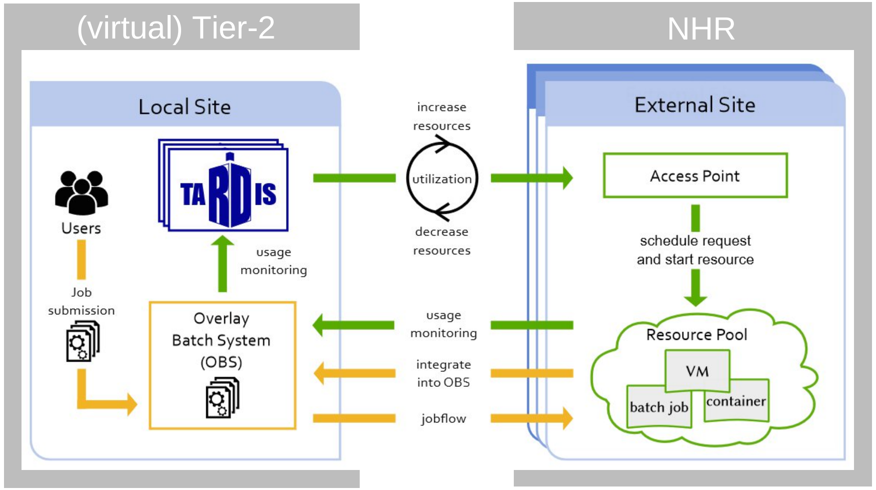
SPONSORED BY THE

Federal Ministry
of Education
and Research

Dedicated project funded by the federal government for development and testing of technologies for a federated computing infrastructure (including the transformation, but not exclusively).

- Tools for the **integration of heterogenous resources**, e.g. HPC centres:
    - Resource management: COBalD/TARDIS
    - Accounting: AUDITOR *(talk by Michael Böhler tomorrow WLCG-Session 17:00)*
    - ...
- Tools for **distributed data storage**:
    - Caching
    - Monitoring
    - Improved authorization mechanisms
    - ...
- **Testing and optimization** under realistic conditions

Virtual Worker Nodes (Drones)



Basic concept being implemented for all three NHR centres (twice for NHR@KIT)

Provides necessary flexibility on HPC nodes for:

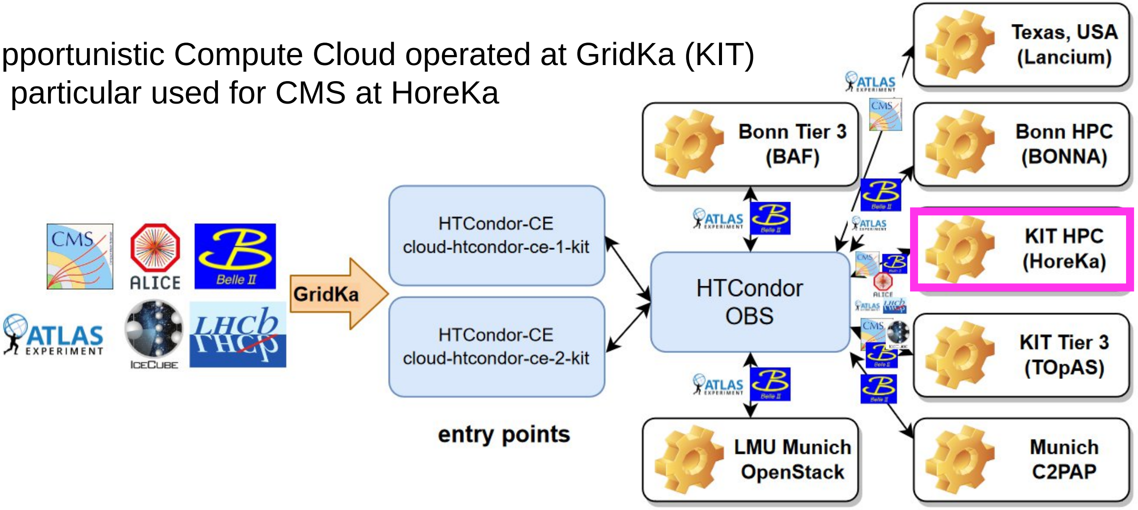
- Job slot design: Overlay batch system takes care of partitioning host nodes (dealing with whole-node scheduling)
- cvmfs: no cvmfs-installation on host required with cvmfs-exec + less dependence on HPC site admins
- Software
- Network configuration

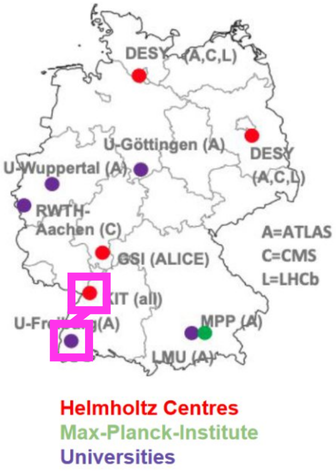
Minimal dependence on HPC administration. However, some dependencies remain, e.g. enabled user namespaces, network, downtimes - therefore two (or more in the future) NHR centres going to serve ATLAS / CMS respectively.

Integration at the WLCG/NHR-Sites

NHR@KIT “HoreKa”

Opportunistic Compute Cloud operated at GridKa (KIT)
in particular used for CMS at HoreKa

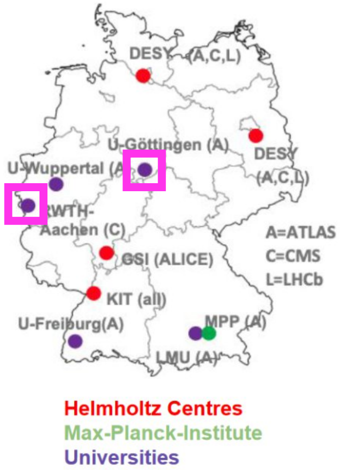




- KIT CMS group developed COBalD/TARDIS and has included various heterogenous resources in an Opportunistic Compute Cloud for many years, including the NHR@KIT HPC cluster HoreKa. (Note: Future usage of HoreKa not (just) opportunistic)
- Freiburg ATLAS group recently prepared an independent drone-based setup to integrate HoreKa resources transparantly into the Freiburg Tier-2 batch system for ATLAS workflows (i.e. job submission via Freiburg ARC-CEs and batch system).
- Bandwidth for data access minor problem due to Tier-1 and future additional storage in same building.

Integration at the WLCG/NHR-Sites

NHR@Aachen “CLAIX”, NHR@Göttingen “Emmy”



Similar drone-based approaches implemented in Aachen CMS and Göttingen ATLAS:

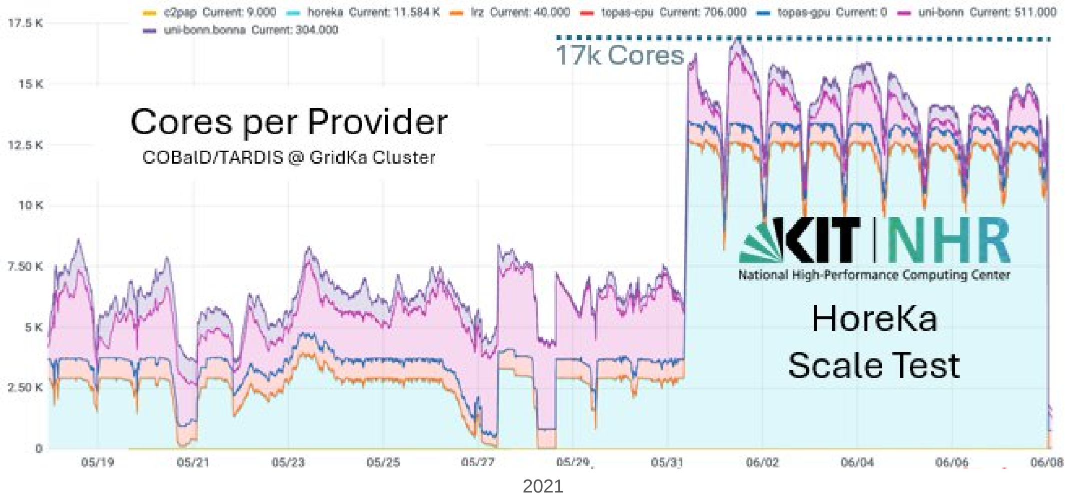
- in Göttingen with cvmfs-exec, in Aachen cvmfs on bare metal nodes
- transparently expanding the existing Tier-2 batch systems (as OBS), reusing CEs and squid proxies
- Aachen already equipped with 100Gbit/s shared WAN access and direct outbound connections allowed; in Göttingen, fast and direct access to local Tier-2 mass storage has been established, remaining outbound traffic via proxy servers

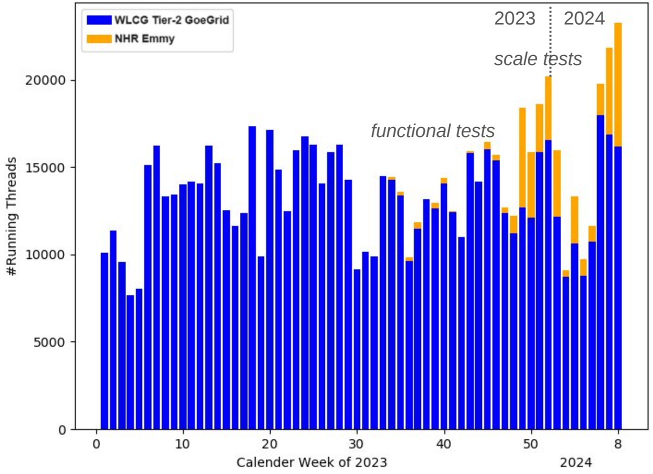
Future data access to KIT and DESY envisaged:

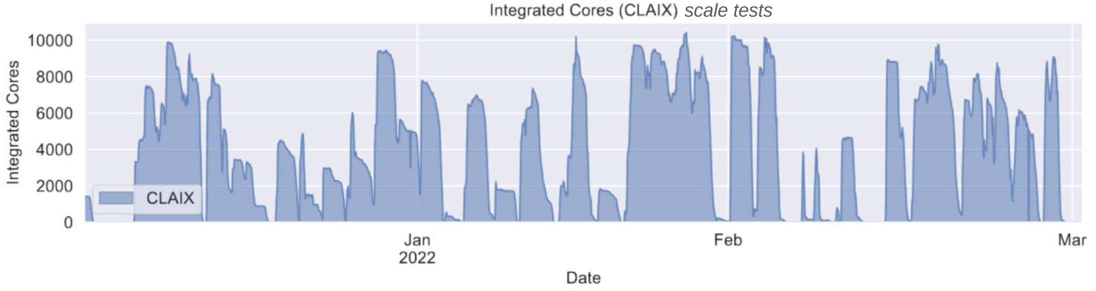
- pushing for high-bandwidth WAN
- caching mechansims to be implemented as much as needed

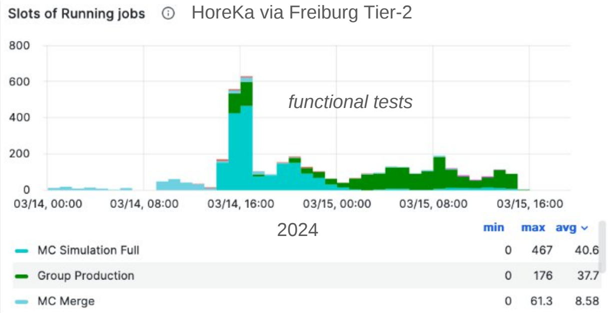
NATIONALES HOCHLEISTUNGS RECHNEN

Testing and Consolidation









	min	max	avg
MC Simulation Full	0	467	40.6
Group Production	0	176	37.7
MC Merge	0	61.3	8.58

Federated data access

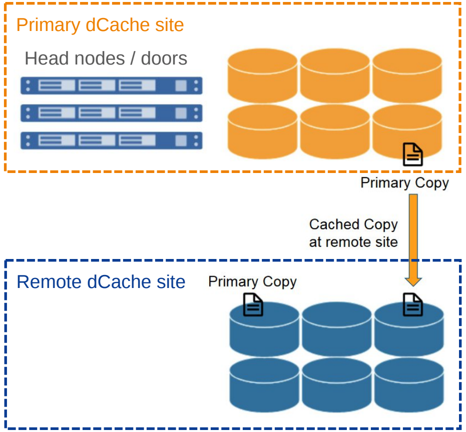
Tier-2 mass storage to be hosted at Helmholtz Centres KIT and DESY - no permanent storage at NHR, just smaller caches to be applied for if needed.
However, remaining storage at univ.-based Tier-2 centres facilitates transition phase until 2029!

Caching solutions being developed and tested, e.g.

- XRootD buffer prototype installed at HoreKa
- ATLAS Pre-Caching to be tested on smaller storage instance in Göttingen
- more activities in Freiburg, Munich, Wuppertal

Federated dCache solutions being developed at DESY

- Allows for decentralized data pools, i.e. primary dCache site with data pools (e.g. for caching) at other sites
- If using dCache for data caches, this would avoid running core dCache services at the NHR sites
- Various degrees of ‘independence’ of the dCache satellite site possible, e.g. shared/separate namespaces



Summary and Outlook

- Upcoming transition from German university-based Tier-2 centres to NHR centres (for CPU) and Helmhotz-Centres (for mass storage) in order to profit from their resource efficiency and synergies
- Basic setups for integration of NHR compute resources are in place at all three envisaged NHR sites
- Regular jobs from ATLAS/CMS can be run and **scaling tests have reached size of a Tier-2** respectively
- Optimization and long-term testing advanced but ongoing
- Data access currently supported by local WLCG centres

→ WAN bandwidth improvements and caching solutions to be pursued where necessary

- Current tests and job processing at NHR sites in scope of a one-year ‘pilot-phase’ (following previous R&D projects)
- Starting in 2025, extension and replacement of university-based Tier-2 CPU resources to be covered by applications for CPU time at NHR centres
- **Transition phase** foreseen for the **next 5 years** till beginning of HL-LHC
- NHR sites are hosting GPUs, which could be included in project applications if requested by ATLAS/CMS