GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

GoeGrid meets Emmy

Integration of HPC clusters into WLCG workflows

Sebastian Wozniewski, II. Physikalisches Institut, Georg-August-Universität Göttingen

GöHPC Coffee – 07.02.24

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

WLCG Tier-2 Site @Göttingen

NHR/HLRN HPC center @Göttingen

GoeGrid meets Emmy

Integration of HPC clusters into WLCG workflows

Sebastian Wozniewski, II. Physikalisches Institut, Georg-August-Universität Göttingen

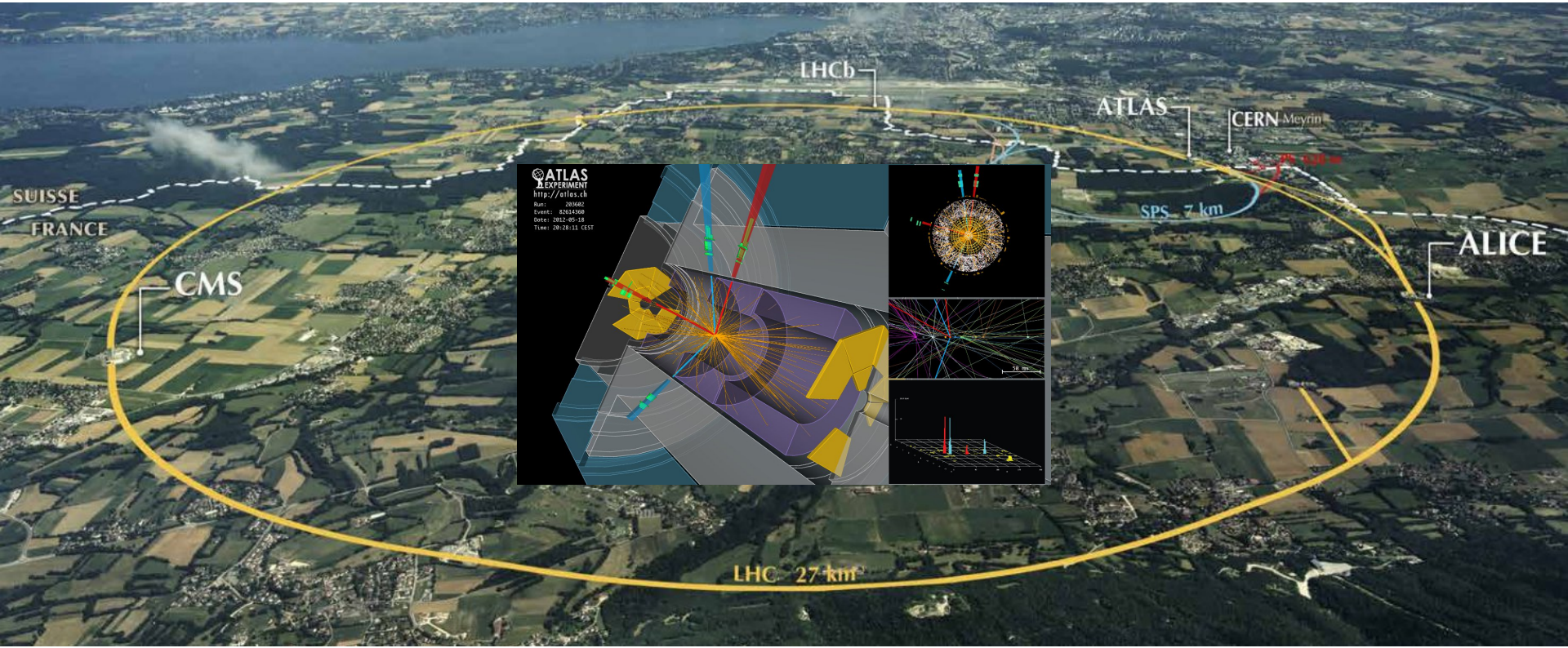
GöHPC Coffee – 07.02.24

Worldwide LHC Computing Grid (WLCG)

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

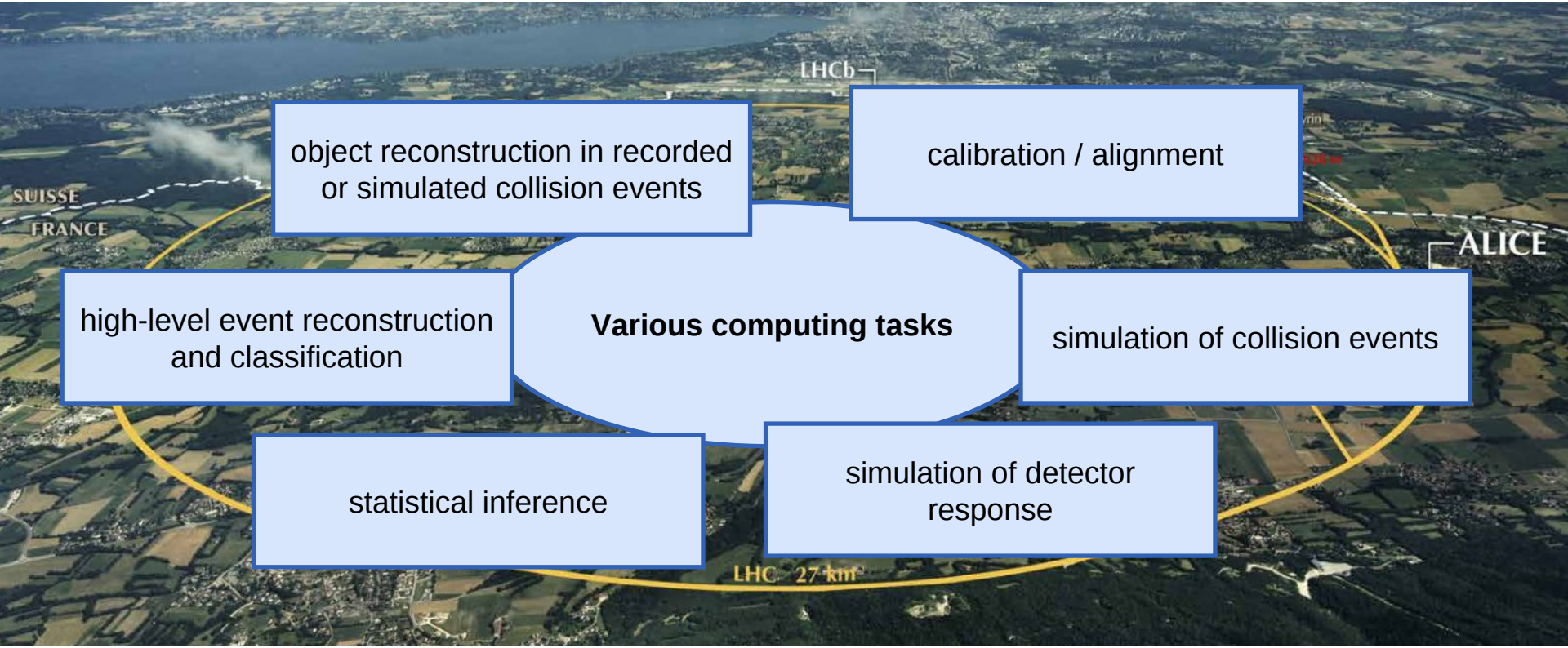
LHC Experiments and data processing

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737



LHC Experiments and data processing

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737



object reconstruction in recorded or simulated collision events

calibration / alignment

high-level event reconstruction and classification

Various computing tasks

simulation of collision events

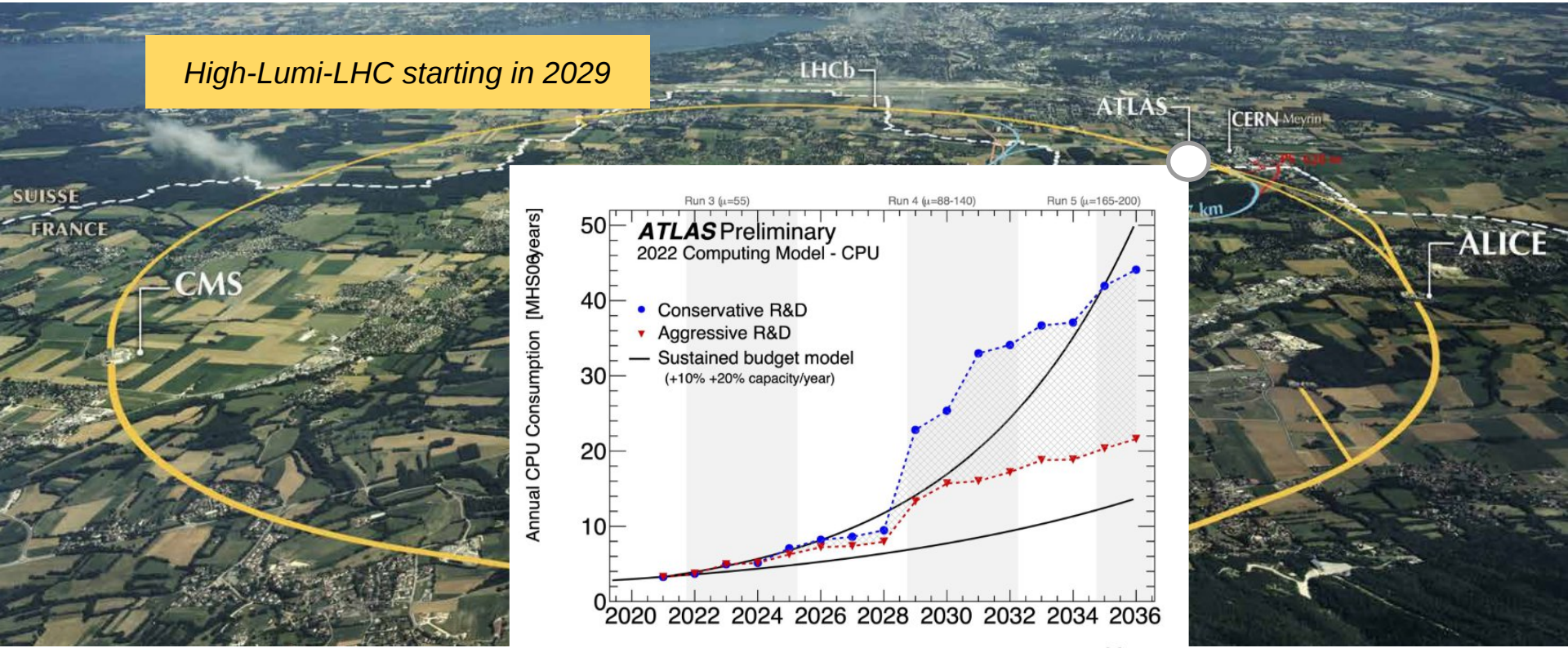
statistical inference

simulation of detector response

LHC Experiments and data processing

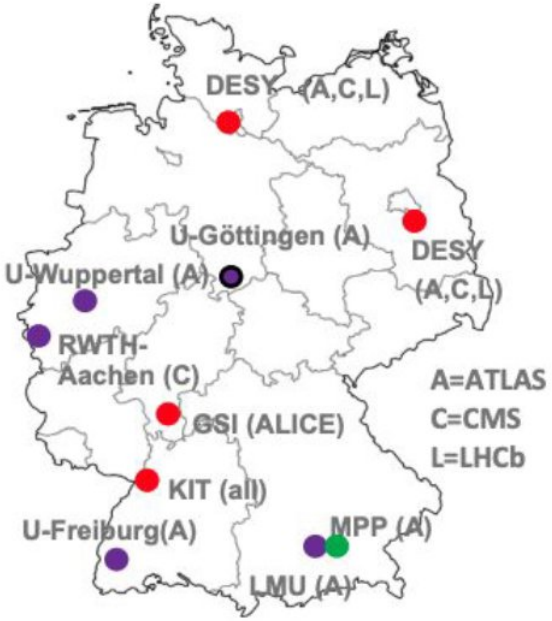
GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

High-Lumi-LHC starting in 2029



[ATLAS Software and Computing HL-LHC Roadmap]

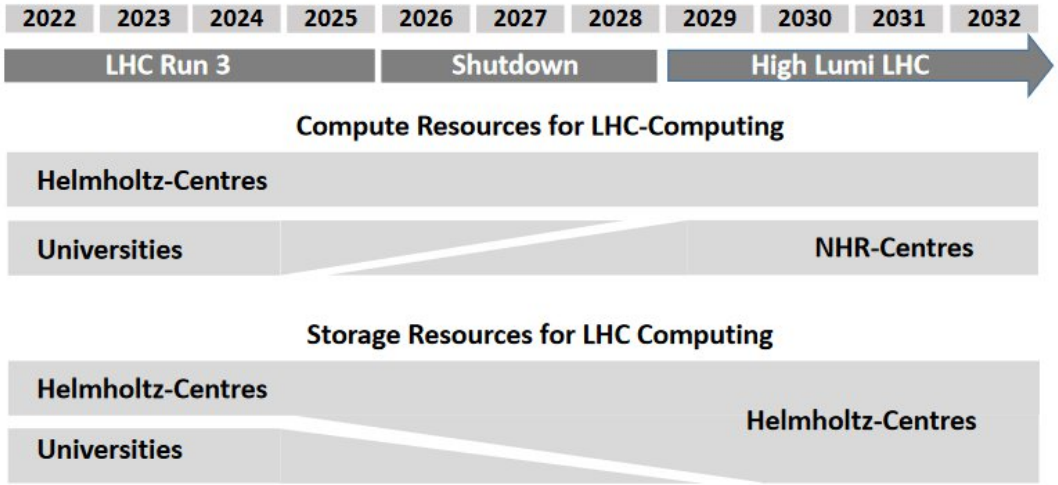
LHC Computing in Germany



- Mostly ATLAS and CMS in a research compound funded by BMBF “Föderiertes Computing für die ATLAS- und CMS-Experimente am Large Hadron Collider in Run 3”
- Tier 1 centre at KIT
- Various Tier 2 centers at Helmholtz Centres and Universities

Transformed Model for WLCG Resources in Germany

Markus Schuhmacher @ NHR-Symposium 2022



Moving Tier-2 resources from universities to NHR-Centers (for computation) and Helmholtz-Centers (for storage)

- Better cost and energy efficiency at fewer large sites
- Foster synergies with other science fields

HPC clusters in the WLCG

- Various cases of HPC usage over the past years, e.g. SuperMUC (Garching), CSCS (Lugano), HoreKa (Karlsruhe)...

- 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!**

GoeGrid meets Emmy

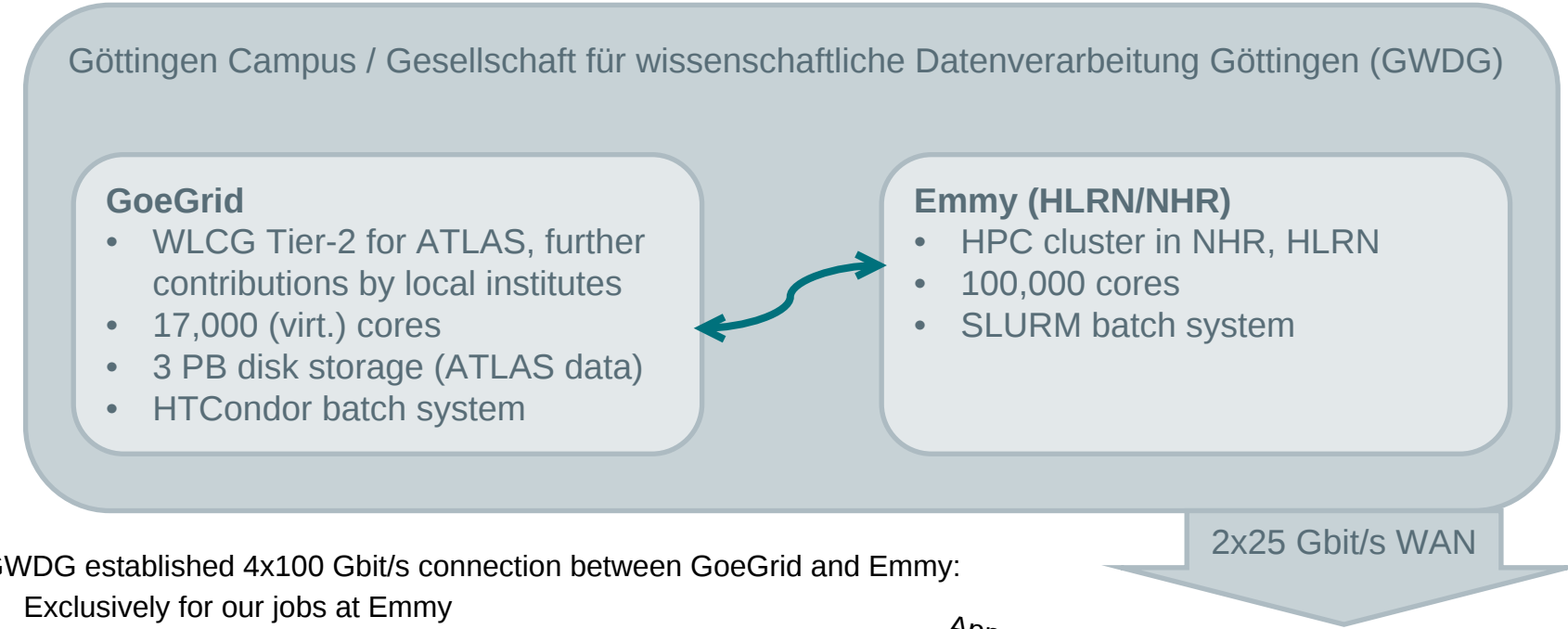
Göttingen Campus / Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG)

GoeGrid
- WLCG Tier-2 for ATLAS, further contributions by local institutes
- 17,000 (virt.) cores
- 3 PB disk storage (ATLAS data)
- HTCondor batch system

Emmy (HLRN/NHR)
- HPC cluster in NHR, HLRN
- 100,000 cores
- SLURM batch system

2x25 Gbit/s WAN

GoeGrid meets Emmy



GWDG established 4x100 Gbit/s connection between GoeGrid and Emmy:

- Exclusively for our jobs at Emmy
- Good access to existing grid storage
- Synergetic use of existing grid and cluster services

Long term vision: high-bandwidth WAN access with remote data lake

Approved NHR application for this R&D

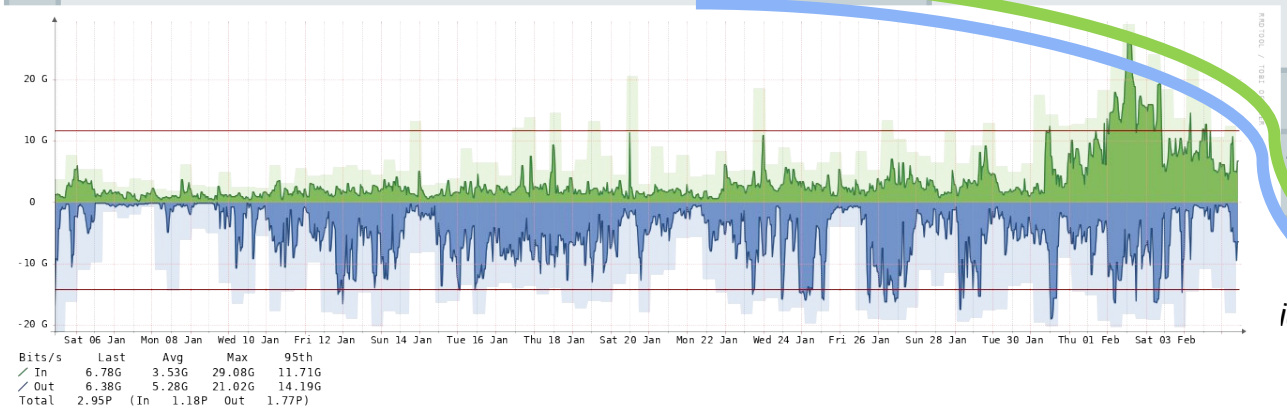
GoeGrid meets Emmy

Göttingen Campus / Gesellschaft für wissenschaftliche Datenverarbeitung Göttingen (GWDG)

GoeGrid
- WLCG Tier-2 for ATLAS, further contributions by local institutes
- 17,000 (virt.) cores
- 3 PB disk storage (ATLAS data)

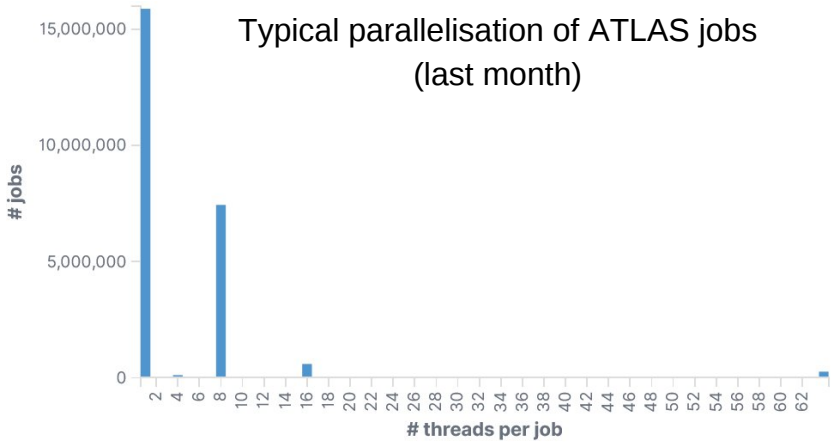
Emmy (HLRN/NHR)
- HPC cluster in NHR, HLRN
- 100,000 cores
- SLURM batch system

2x25 Gbit/s WAN



Further challenges

- Data access done ✔ but that's not all!
- How efficiently **schedule variety of jobs** on HPC nodes? (whole-node scheduling policy)
- How provide **cvmfs access**? cvmfs:
  - used in WLCG for **distribution of software and configuration data**
  - implemented as POSIX read-only file system mounted in /cvmfs with files being hosted on remote webservers and local cache-instances



Extending the known concept of pilots

Pilot jobs widely used by LHC experiments:

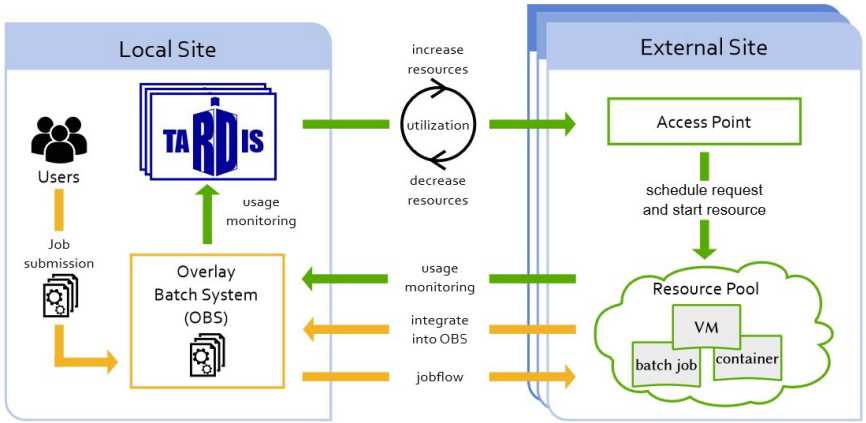
Ensures proper environment on worker node before pulling actual jobs and running them (**in a container if needed**).

Extending the known concept of pilots

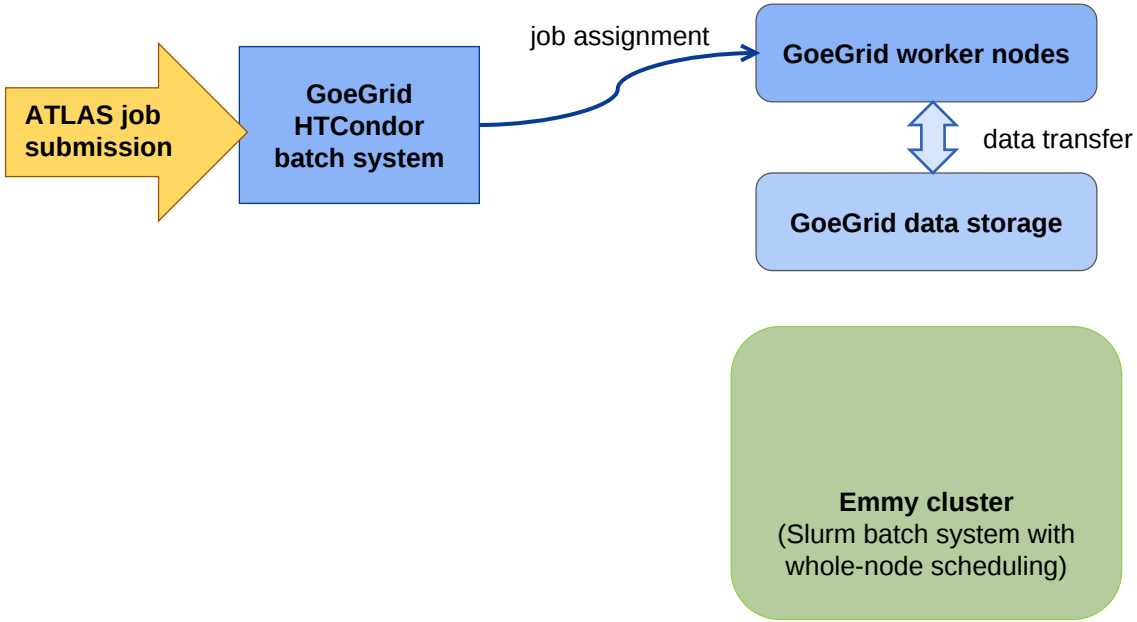
Pilot jobs widely used by LHC experiments:

Ensures proper environment on worker node before pulling actual jobs and running them (**in a container if needed**).

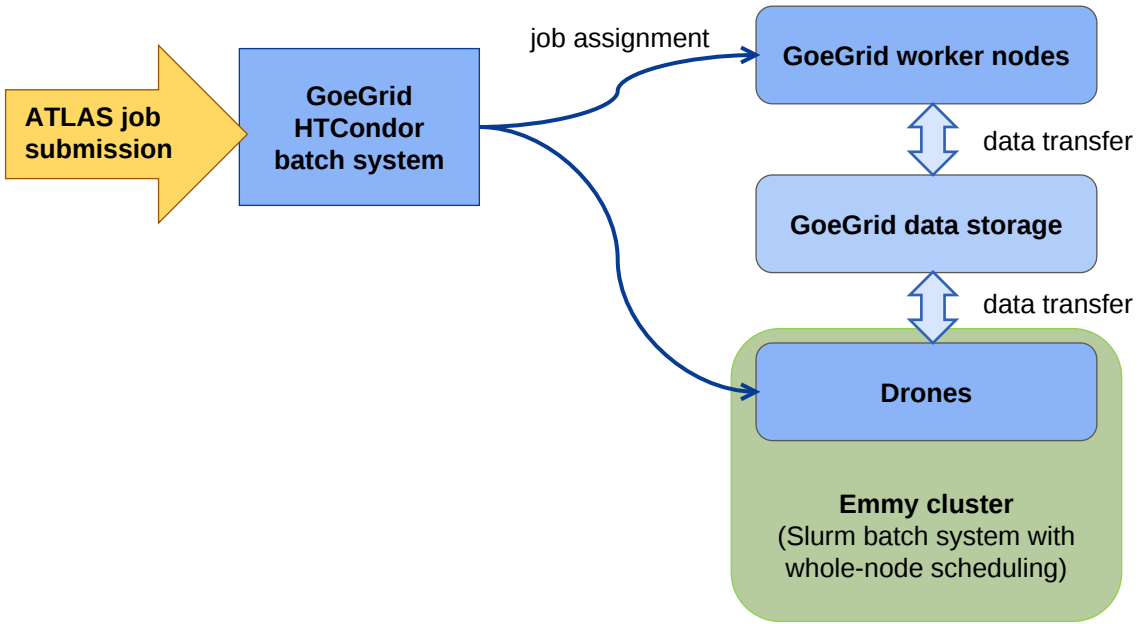
Adding **another layer of containers**, worker nodes can be turned into virtual worker nodes of an overlay batch system, called “**drones**”, which satisfy our needs.
The COBalD/TARDIS resource manager has been developed at KIT for automatically managing such drones ([presented at NHR-Symposium 2022 by Manuel Giffels])



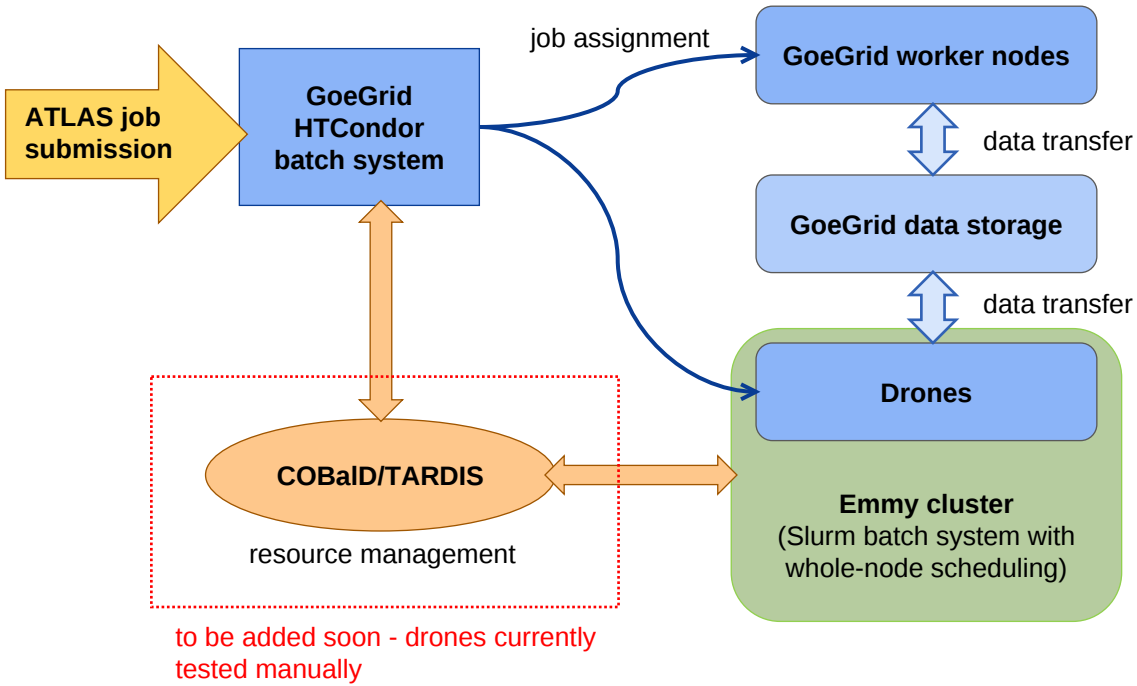
GoeGrid drones on Emmy



GoeGrid drones on Emmy

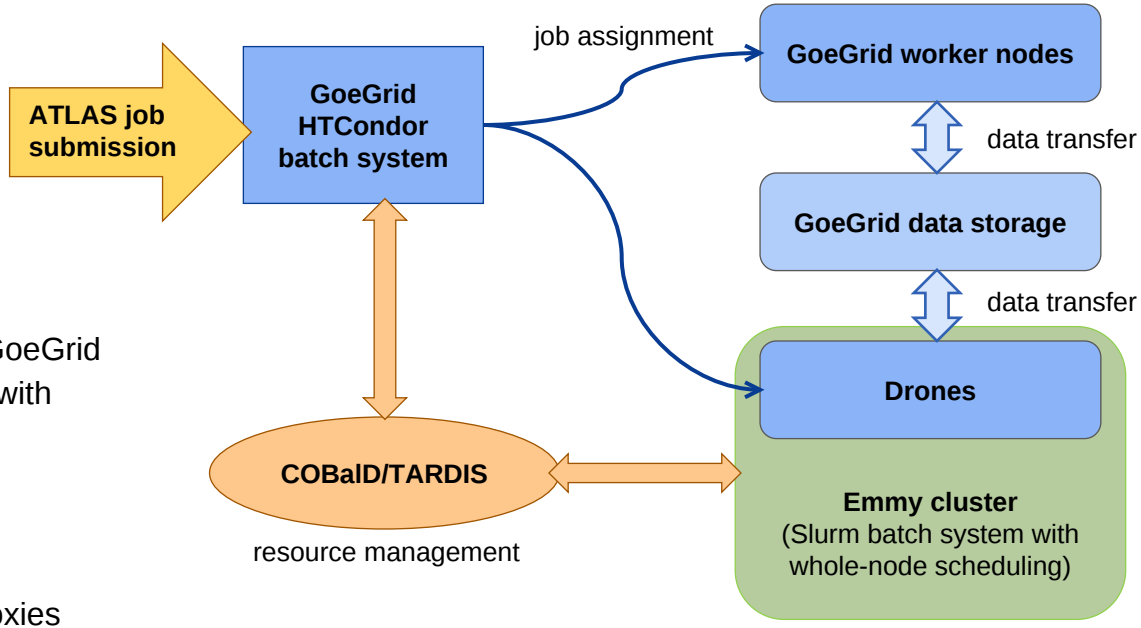


GoeGrid drones on Emmy



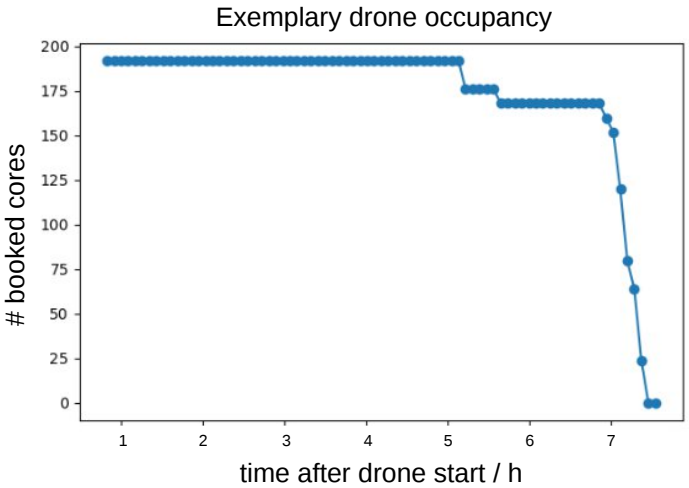
GoeGrid drones on Emmy

- connect to HTCondor batch system of GoeGrid
  - **dynamically partitionable slots** with continuous job execution
  - additional flexibility
- cvmfs made available in the container (cvmfs-exec)
  - reuses existing GoeGrid squid-proxies
- scratch space assigned on shared SSDs / HDDs depending on availability

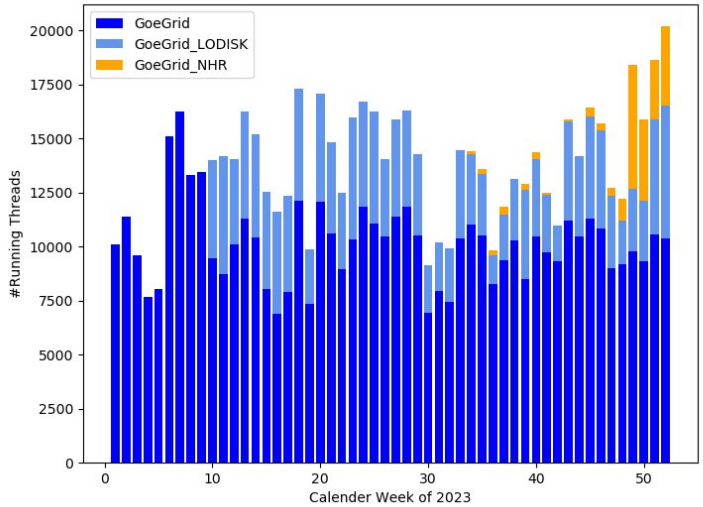


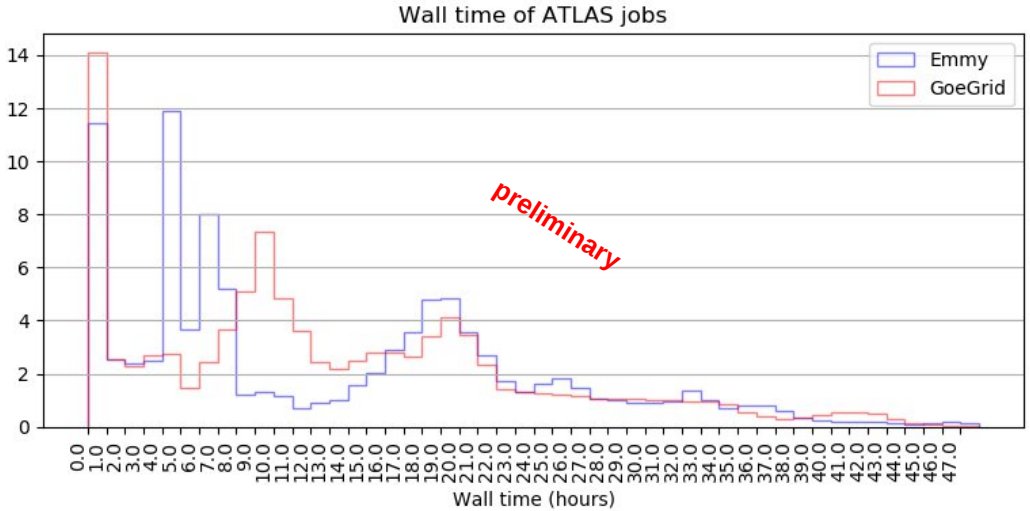
Drone usage efficiency

- Most relevant losses due to draining of drones
- Continuous job submission in a drone **clearly beneficial** compared to bundled job submission approaches
  if drone lifetime >> average job wall time
- Initial default of 12h now extended to 7 days.
- Continuous resource usage by ATLAS with multiple drones envisaged => even with 7d lifetime drones would finish at an hourly basis.
- Nevertheless, working on solutions to use remaining cores while draining.



Successfully running real ATLAS jobs





- Testing processing of regular ATLAS jobs since Aug., incl. larger permanent load during Dec.
- Performance comparable to GoeGrid
- Planning for a one-year test phase: Some issues only show up afte some time - special jobs / other useres / ...

Conclusion & Outlook

- Making NHR site Emmy ready for WLCG workloads
- Implemented access to GoeGrid data storage as well as containers with necessary environment and sub-WN-job-scheduling via GoeGrid batch system
- ATLAS jobs running successfully on Emmy nodes - detailed tests ongoing
- Soon ready for transition to regular job production - ATLAS Germany planning to hand in two NHR production applications by March 2024, one by Göttingen group for Emmy and one by Freiburg group for HoreKa

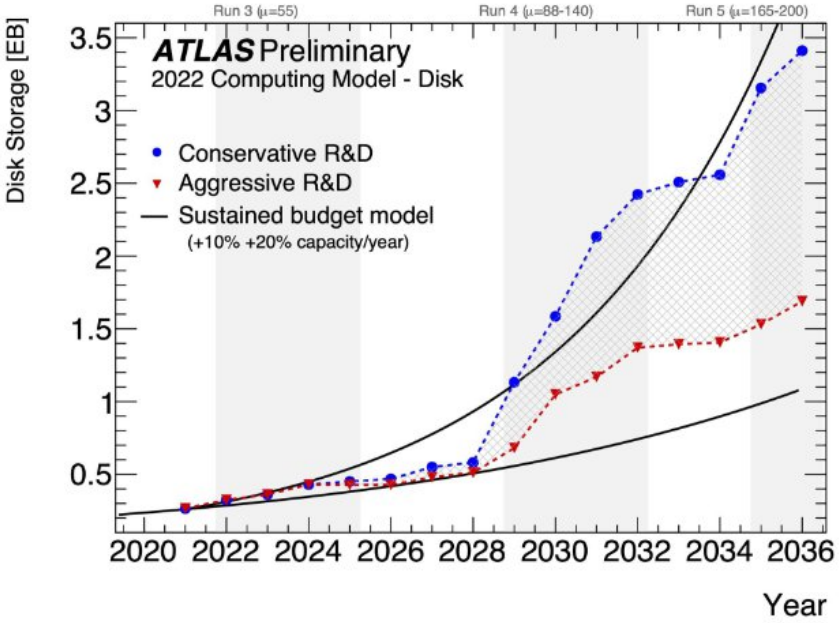
Acknowledgements

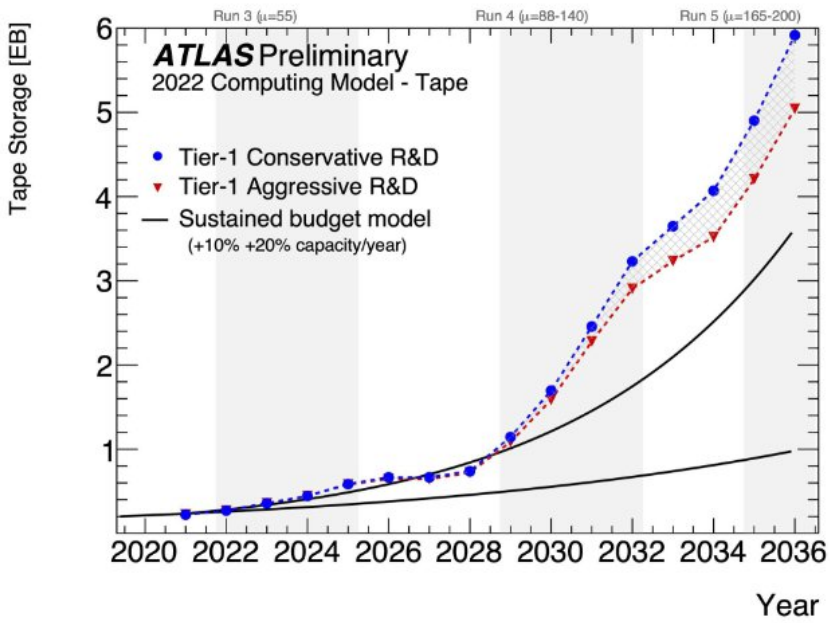
- Partners of the FIDIUM (federated infrastrucures) project and its funding agency BMBF
- Network and Emmy teams at GWDG
- NHR

GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

backup

ATLAS data storage needs





ATLAS Software and Computing HL-LHC Roadmap