Around the sites

Torsten Harenberg - Bergische Universität Wuppertal
Jörg Marks - Universität Heidelberg

Categories of sites

University clusters

- size varies from small to larger than some Tier-2s
- funding usually by DFG and/or state ressources
- Usually no wLCG enabled disc space
- successful integration for example in Dortmund and Bonn

wLCG Tier-2 (+1)

- certain “minimal” size (typical ~2kSlots)
- pledged disc space with agreed protocols important for LHC experiments
- Funded by BMBF (Uni) and Helmholtz (DESY, GridKa)
- MoU

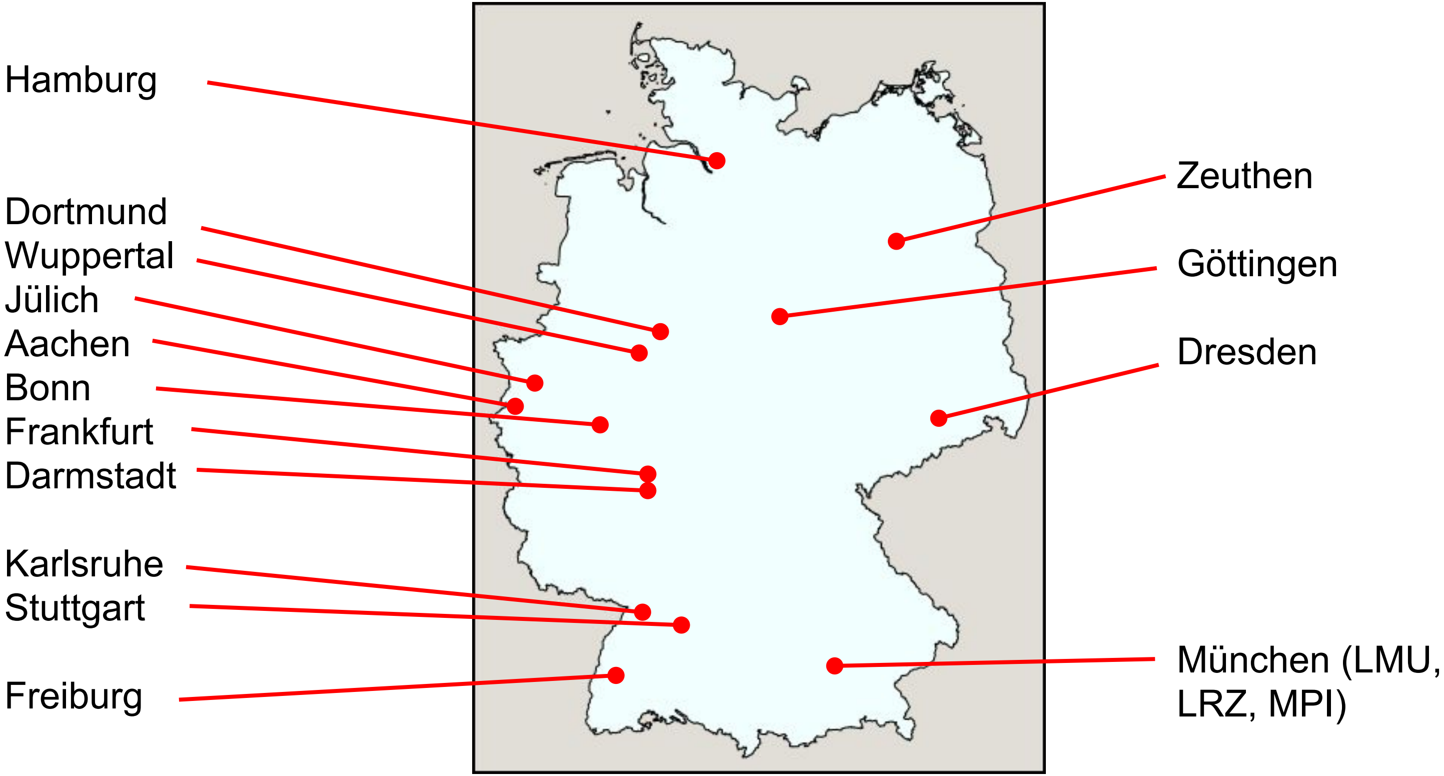
NAF

- DESY + GSI

HPC centers

- Examples: Jülich, Stuttgart
- massive resources, usually full-MPI, application needed
- sometimes non-standard x86 hardware
- Integration into Grid middleware successful @ LRZ Munich

Map of sites



Hamburg

Dortmund
Wuppertal
Jülich
Aachen
Bonn
Frankfurt
Darmstadt

Karlsruhe
Stuttgart

Freiburg

Zeuthen

Göttingen

Dresden

München (LMU, LRZ, MPI)

Bergische Universität Wuppertal

#cores (x86 only) w/funding	1500 (BMBF) + 1024 (DFG)
SRM enabled storage (TB)	2 PB (dCache)
local storage (high performance)	120 TB Lustre
#FTEs (acad./tech) for maintenance	1 (2 people)/1
Network	2 GBit LHCone (contract cancelled end 2018 due to funding problems), 2x3 Git DFN
wLCG affiliation	Tier-2 (ATLAS)
Experiments served	ATLAS, Auger, IceCube
Remarks	DFG part is used by nearly all groups in university with computing needs

LMU Munich

#cores (x86 only) w/funding	1500 (BMBF)
SRM enabled storage (TB)	1500 TB ATLAS + 400 TB local (StorRM)
local storage (high performance)	(on opportunistically used HPC clusters)
#FTEs (acad./tech) for maintenance	approx 2 (4 people)
Network	2x130 GBit/s
wLCG affiliation	Tier-2
Experiments served	ATLAS
Remarks	Oppotunistic resources: backfill on SuperMUC (300kCores machine) + C2PAP (2kCores). Both machines will be decommissioned 2019. Group was very active in developing hacks to get these resources (completely different environment) -> MUC pledge fulfillment usually ~300%.

MPI Munich

#cores (x86 only) w/funding	
SRM enabled storage (TB)	
local storage (high performance)	
#FTEs (acad./tech) for maintenance	
Network	
wLCG affiliation	Tier-2
Experiments served	ATLAS
Remarks	

RWTH Aachen

#cores (x86 only) w/funding	2450 (HGF, BMBF) + 2970 (local)
SRM enabled storage (TB)	3.6 PB (50% official CMS, 50% german CMS groups) (dCache)
local storage (high performance)	
#FTEs (acad./tech) for maintenance	> 2.5
Network	80 GBit (LAN) -> >100 Gbit (WAN)
wLCG affiliation	Tier-2
Experiments served	CMS + (IceCube, Auger <5%)
Remarks	opportunistic usage of 275 Desktop PCs

DESY Zeuthen

#cores (x86 only) w/funding	~ 9100 together (4224 Grid,3000 local farm, 1920 HPC)
SRM enabled storage (TB)	~ 7.2 PB (dCache)
local storage (high performance)	~ 2.5 PB Lustre
#FTEs (acad./tech) for maintenance	~ 4-5
Network	2x10 GBit
wLCG affiliation	Tier-2
Experiments served	Atlas, CTA, HESS, Icecube, ZTF, Theoretical Astrophysics, (I)LDG, Theoretical Physics, PITZ
Remarks	

DESY Hamburg

#cores (x86 only) w/funding	~ 18000
SRM enabled storage (TB)	~ 16 PB
local storage (high performance)	
#FTEs (acad./tech) for maintenance	
Network	2x10 GBit LHCone + 2x15 GBit DFN
wLCG affiliation	Tier-2 (+ RAW data center for Belle 2)
Experiments served	Atlas, CMS, Belle 2, ILC, LHCb, CALICE
Remarks	

Uni Freiburg

#cores (x86 only) w/funding	1260 (BMBF) + 512 (DFG, phasing out this year) + 17.460 (NEMO Cluster, DFG, for whole BW - HEP and others)
SRM enabled storage (TB)	1.9 TB (dCache)
local storage (high performance)	500 TB BeeGFS + 30 TB (backuped)
#FTEs (acad./tech) for maintenance	1.5
Network	
wLCG affiliation	Tier-2 (+3)
Experiments served	Atlas,
Remarks	Support from university's compute center (providing virt. Machines, high performance storage and others). Using OpenStack. IPv6 is not supported by university.

Uni Göttingen

#cores (x86 only) w/funding	4428 (funding?)
SRM enabled storage (TB)	1.95 TB (dCache)
local storage (high performance)	
#FTEs (acad./tech) for maintenance	1
Network	10 GBit/s DFN (via GWDG)
wLCG affiliation	Tier-2 (+3)
Experiments served	Atlas,
Remarks	GWDG Cloud with OpenStack

Uni Mainz (mainzgrid+mainz)

#cores (x86 only) w/funding	Using share on University Cluster, no dedicated cluster
SRM enabled storage (TB)	2.7 PB (dCache)
local storage (high performance)	GPFS+Lustre
#FTEs (acad./tech) for maintenance	approx 0.5
Network	10 Gbit, non-DFN line to DE-CIX, several routing problems :-(
wLCG affiliation	Tier-3
Experiments served	ATLAS, NA62, Icecube
Remarks	

Uni Bonn (only 2017 resources)

#cores (x86 only) w/funding	1120 Cores
SRM enabled storage (TB)	220 TB (xrootd)
local storage (high performance)	CephFS
#FTEs (acad./tech) for maintenance	1
Network	10 GBit/s
wLCG affiliation	Tier-3, but no CPU resources offered to outside yet
Experiments served	ATLAS
Remarks	Singularity only jobs.

Observations (so far)

- Tier-2 sites run very well (Tier-1 still to come!)
  - esp. “other-than-CPU” services (Storage (dCache,..)) cannot be provided by opportunistic ressources.
- Most Tier-2 sites run with very little manpower (O(1 FTE))
- (At least in ATLAS):
  - most of the manpower in the “cloud squads” come from Tier-2s
  - very active mailing lists and expertise exchange
- We are already a well-established community
  - also between universities and centers

Experiments need these people

Issues raised

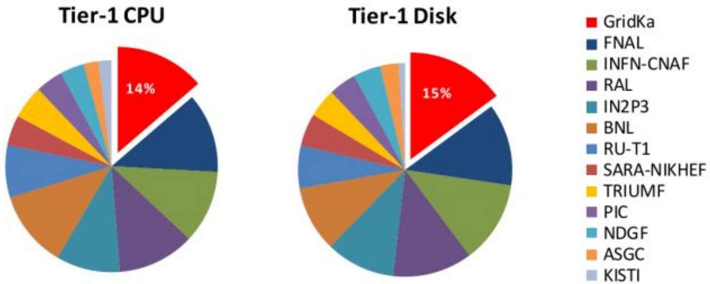
- Funding a “local mixture”, some site have trouble finding money for certain parts (for example LHCOne line, maintenance for Cisco,...)
- “Restmittel” is problematic with European public procurement laws (“Vergaberecht”)
    - Loss of DFG/HP master agreement made it more difficult
- non-DFN network lines are problematic (but probably cheaper)
- Grid middleware is mostly not documented, maintenance (updates) manpower intensive
    - “Grid site in a container”?
- RHEL 6 → 7 upgrade not done by all sites yet (experiments were ready very late)
- IPv6 not officially supported by every site

Advanced techniques

- Singularity seems to become an established container technology (CMS+ATLAS)
- More advanced container/virtualization R&D @ Karlsruhe, Freiburg, Aachen
    - Bonn (non-Grid jobs) already run “singularity only”
- opportunistic resources
    - successful integration of a large scale: from HPC centers (MUC) to desktop (AC, (BN))
    - HPC centers can be utilized by backfill jobs, esp. successful @ Munich HPC
    - no “one-fits-all” recipe available, depends on local circumstances
    - looks impossible without outbound IP from Worker Nodes (HPC)

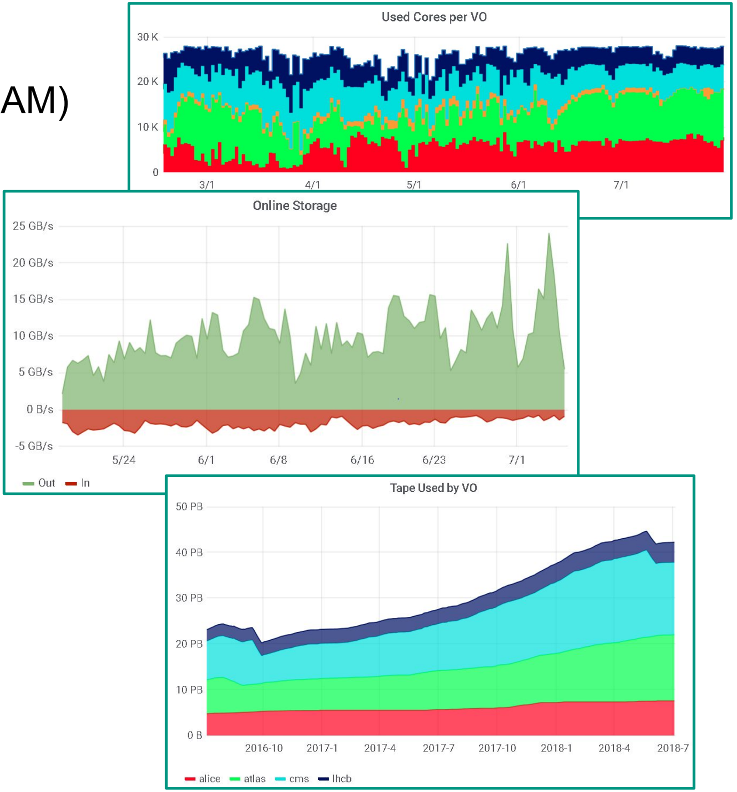
GridKa – A Cornerstone of WLCG

- Scientific data and computing center for HEP and astroparticle physics
  - WLCG Tier-1 center for 4 LHC experiments, 14% of Tier-1 resources
  - RAW data center for Belle II
- Resources
  - Computing: ~ 28,500 job slots
  - Disk: 27 PB (usable)
  - Tape: 45 PB (used)
  - 100 Gbit/s network connectivity to LHCOPN, LHCONE
- The largest and among best performing T1s
- LK-II in HGF



GridKa Current Resources 2018

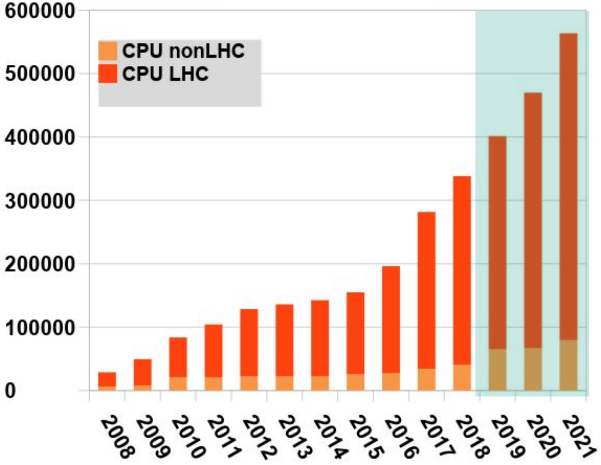
- Compute
  - 340kHS06 (950 WNs, 28500 usable cores, 80TB RAM)
- Online Storage
  - 27PB usable on GPFS (dCache + xrootd)
- Offline Storage
  - 45PB on tape, 1 tape library, 24+ drives
- Network
  - 20Gbit/s (soon 100+20) connection to CERN
  - 100Gbit/s connection to DFN (includes LHCONE)



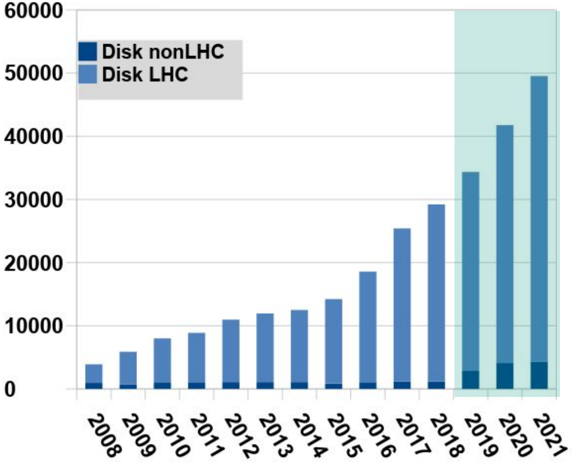
GridKa Resource Planning

- ~20% per year resource increase planned until 2021
  - CPU +220 kHS06, Disk +20 PB, Tape +40 PB
- Beyond 2021: new funding required

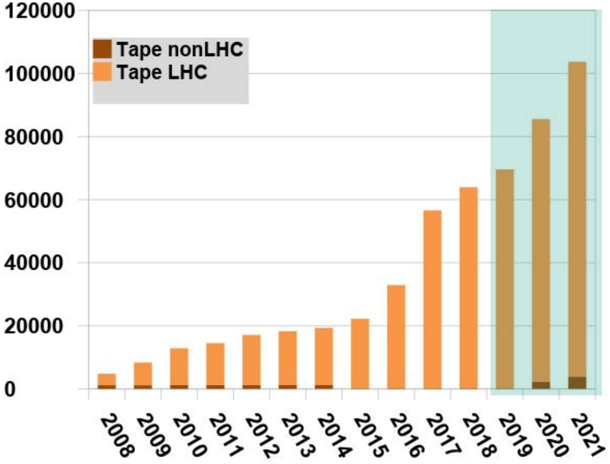
CPU [HS06]



Disk capacity [TB]



Pledged tape capacity [TB]



National Analysis Facilities (I)

Two analysis facilities (NAF) available at DESY and GSI for analysis computing of the german institutes

NAF DESY: infos Yves Kemp

➢ ~40 dedicated work group servers for login, interactive processing,

➢ testing and development

➢ - use FastX as remote desktop (via browser or client, desktop sharing)

➢ - jupyterhub with 2 hardware backends

➢ - container environment under development (needs AFS and Kerberos

➢ integration)

➢ Large batch farm with ~9000 cores (~130 kHS06)

➢ - recently fully migrated to HTCondor

➢ dCache Grid storage (ATLAS ~5 PB, CMS ~8 PB, Belle II ~0.5 PB)

➢ (includes pledged and non-pledged resources)

National Analysis Facilities (II)

NAF DESY: continued

- DUST - dedicated fast file space for scratch purpose
- (total 2.6 PB, 15 GB/s sustained output rate)

- AFS to be replaced in medium term future

- Integrate access to GPU computing

NAF GSI: analysis computing infrastructure for ALICE germany, which is strongly interleaved with ALICE T2 (infos Kilian Schwarz)

- NAF with CPU of 18 kHS06 and 1.7 PB disk space

- Scientific Linux environment provided with Singularity
- containers on Debian-based HPC cluster

- Data processing mainly via analysis trains optimizing throughput

National Analysis Facilities (III)

GSI NAF → Analysis Facility (AF) for run 3

In Run 3 ALICE processes analysis jobs as trains in dedicated analysis facilities (AF) using AODs. Currently setup a prototyp at GSI

- AF hardware requirement
- - process 5 PB in 0.5 days
- - minimize data transfer and optimize processing efficiency
- - only AODs on the AF storage element

- XRootD redirect plugin
- - open file directly from Lustre filesystem

- Prototype test of the AF setup performed
- - 600 TB of data
- - 1000 job slots
- - full AOD set of 2015 Pb-Pb data processing mainly via analysis trains

Backup Slides 1

remaining slides GridKa

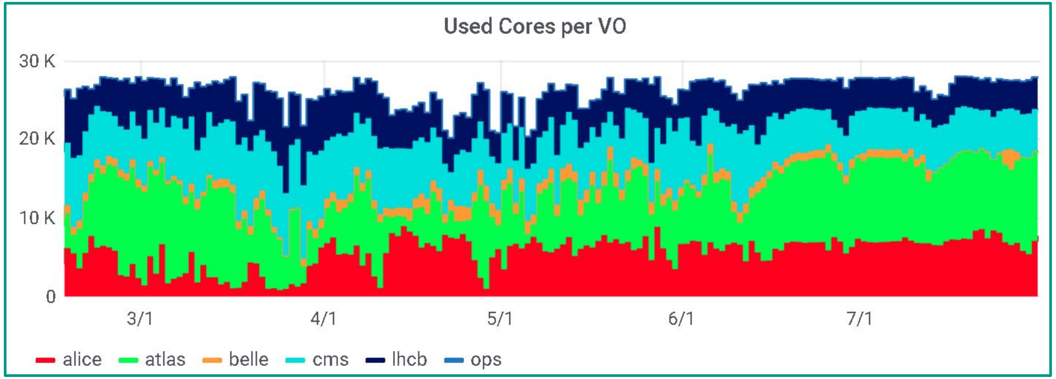
All GridKa slides provided by Andreas Petzold (thanks!).

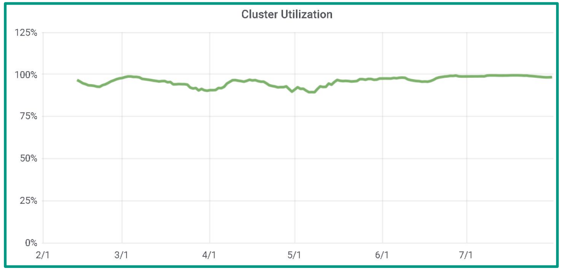
GridKa Components

- Compute Farm
- Online Storage
- Offline Storage
- Network
- GridKa Team

GridKa Batch Farm

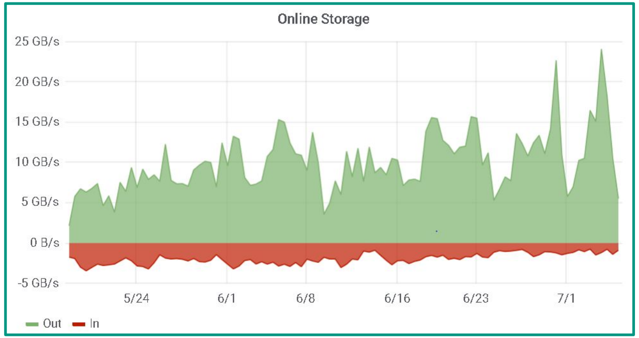
- 950 Worker Nodes
- no low-latency interconnect (inexpensive)
- 28500 usable cores, 80TB RAM
- 93-98% average utilization
- R&D to integrate dynamic external WNs





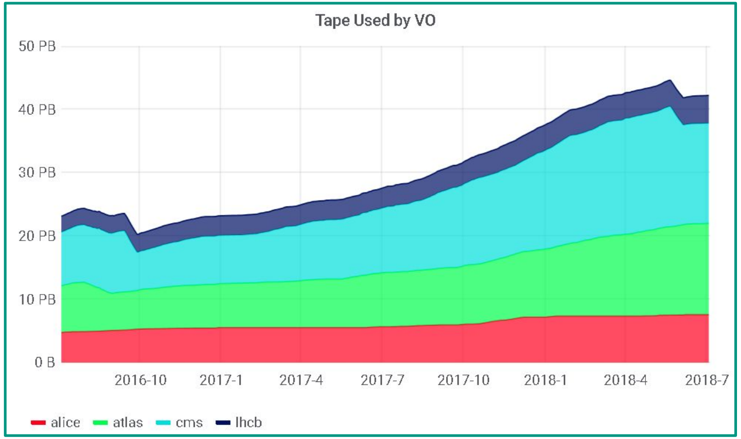
GridKa Online Storage

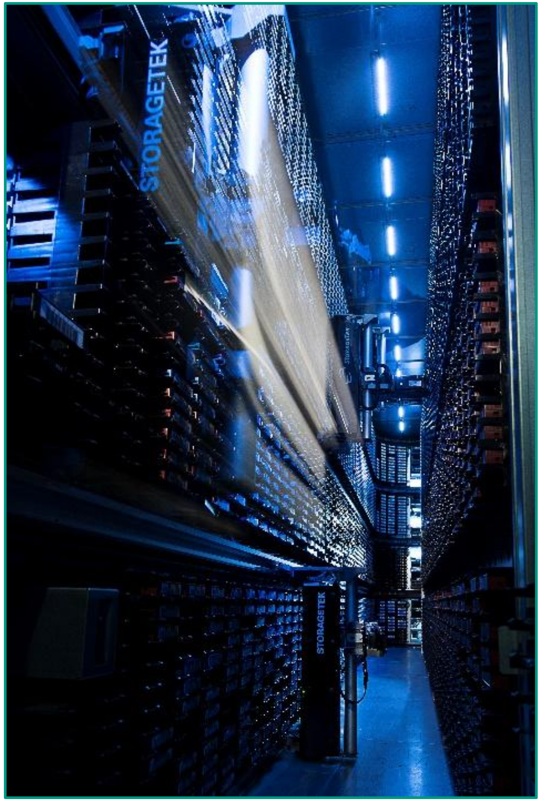
- 27PB available storage
  on hard drives
- >4000 HDDS (8 /10TB)
- 28TB on SSDs
- Up to 100GB/s combined
  read + write performance



GridKa Offline Storage

- Offline storage on magnetic tapes
- 45PB currently stored
- 1 tape library with 10000 slots capacity 85TB today
- 24+ tape drives





GridKa Network

- 20Gbit/s dedicated link to CERN (100+20Gbit/s end of 2018)
- 100Gbit/s to German Research Network
- 200Gbit/s internal backbone
- 40Gbit/s connections to storage

Backup Slide 2

Data Volume Run 3

Output Data Volume Run 3

Data output bandwidth of the experiments for run 3 are currently under discussion and vary depending on concepts (not all numbers are up to date).

- CMS: 2 GB/s + 2 GB/s for parking stream
- ATLAS: 2 GB/s
- ALICE: 100 GB/s
- LHCb: 5 - 10 GB/s depending on the fraction of data in reduced
- format (TURBO)