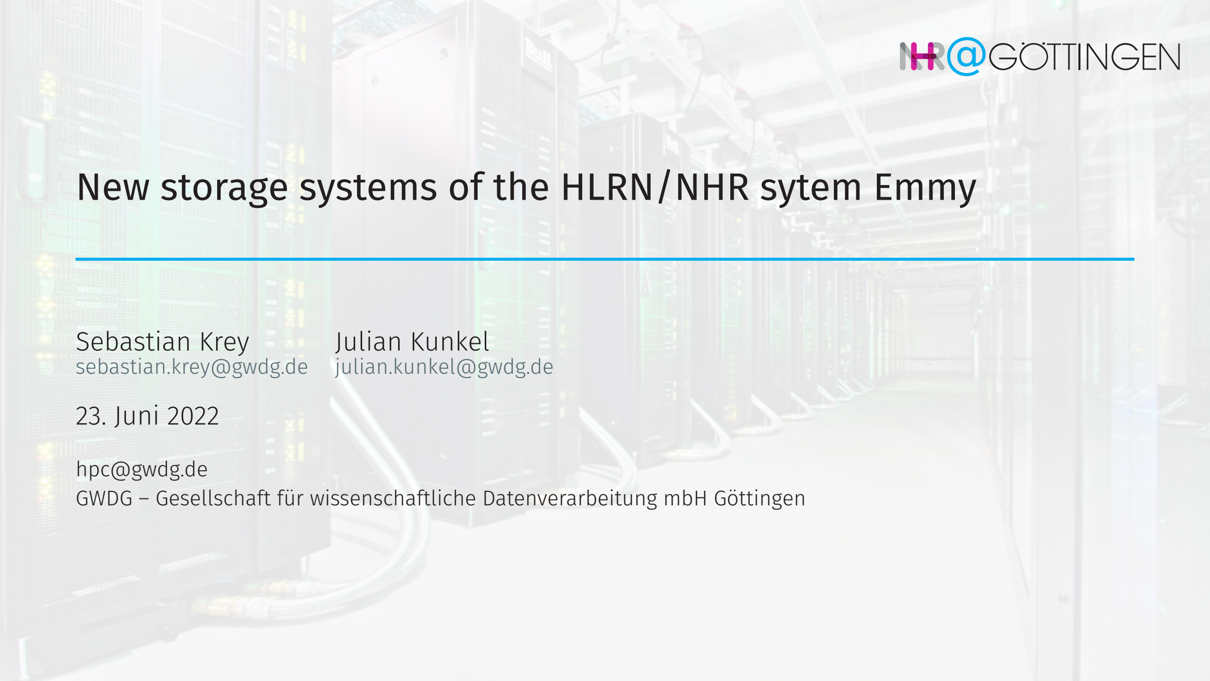


New storage systems of the HLRN/NHR sytem Emmy

Sebastian Krey
sebastian.krey@gwdg.de

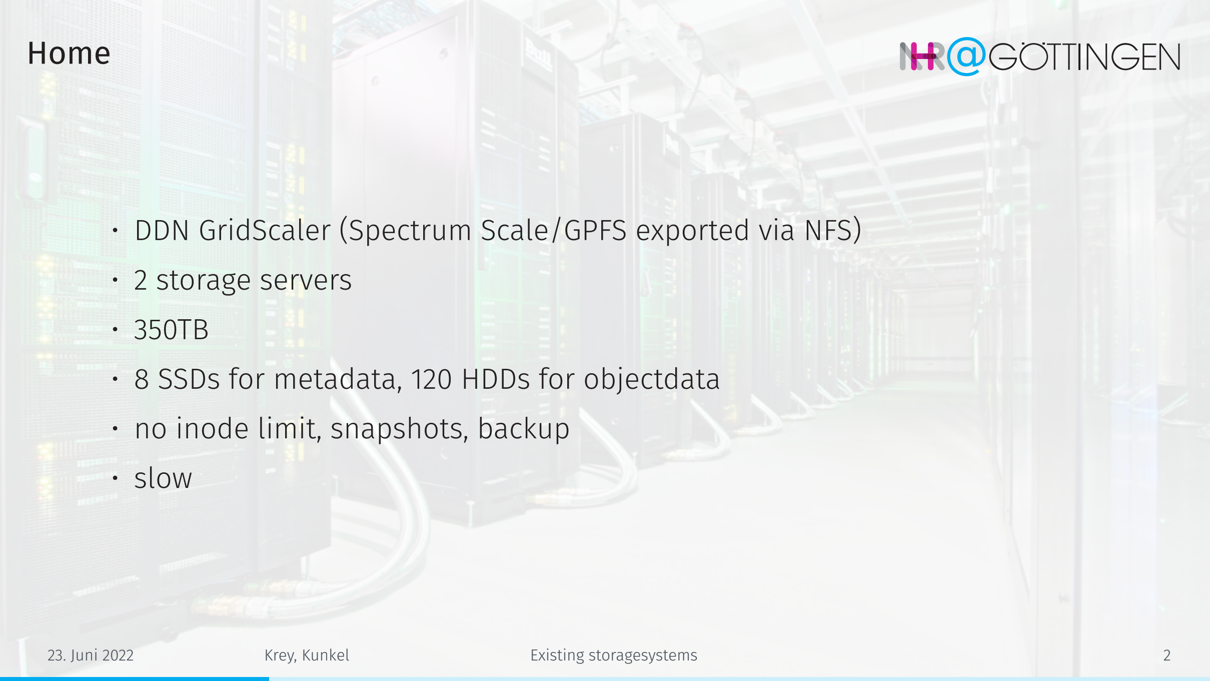
Julian Kunkel
julian.kunkel@gwdg.de

23. Juni 2022

hpc@gwdg.de
GWDG – Gesellschaft für wissenschaftliche Datenverarbeitung mbH Göttingen

Existing storagesystems

Home



- DDN GridScaler (Spectrum Scale/GPFS exported via NFS)
- 2 storage servers
- 350TB
- 8 SSDs for metadata, 120 HDDs for objectdata
- no inode limit, snapshots, backup
- slow

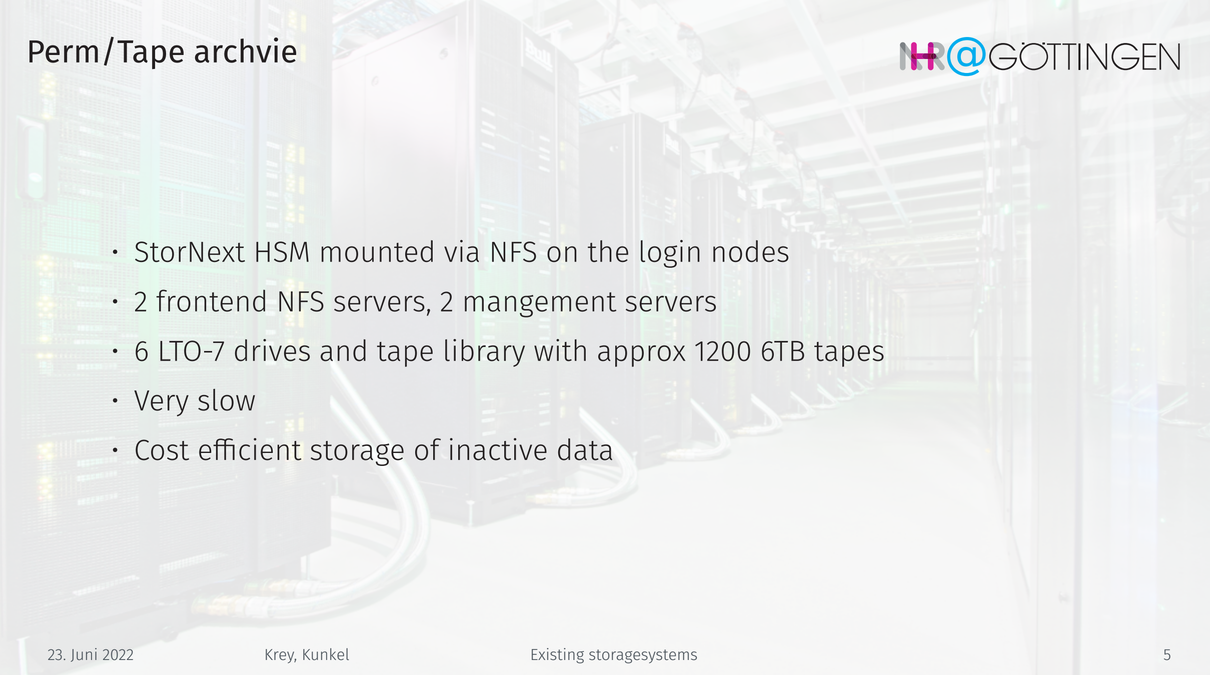
Work/Lustre HDD

- DDN ExaScaler (Lustre)
- 4 metadata servers with a DDN SFA7700X
- 8 object storage servers running on two DDN ES14KX
- 8.5PB
- 16 SSDs for metadata, 1000 HDDs for objectdata
- High performance for streaming IO (50-60GB/s)
- Full POSIX compliance
- Limited metadata performance

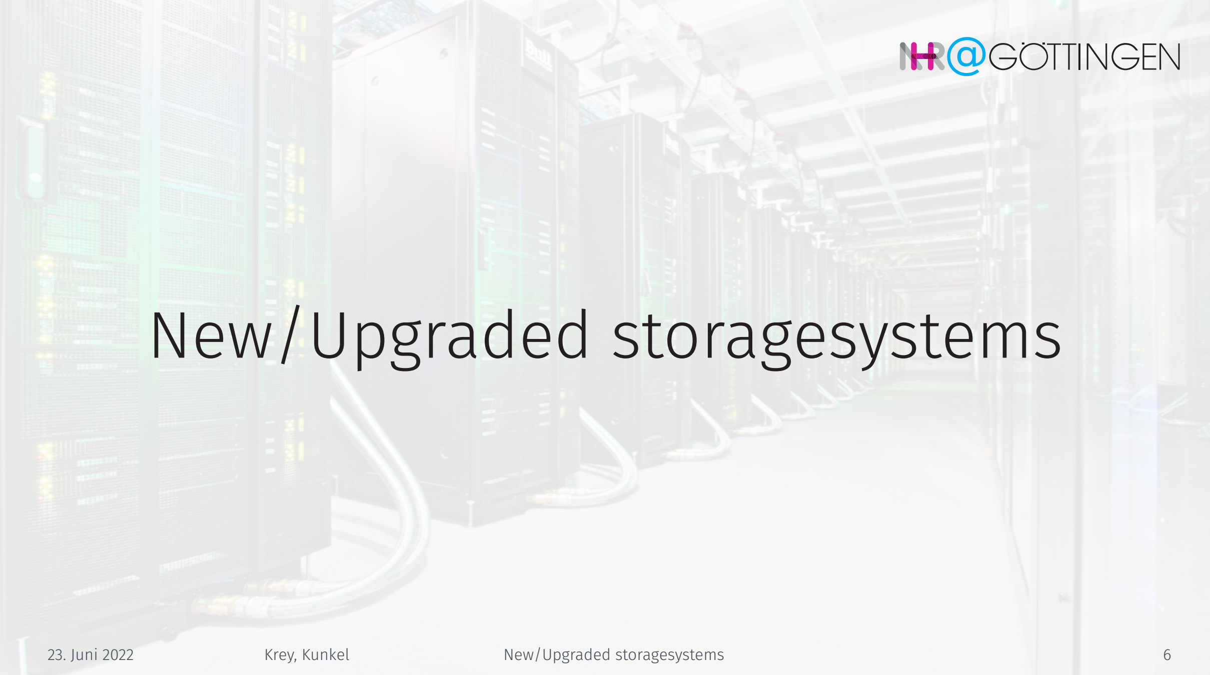
Node local SSDs

- Skylake nodes (medium40, large40): single 480GB Enterprise S-ATA SSD
- GPU nodes: two 2TB Enterprise S-ATA SSDs
- 150 Cascadelake nodes (standard96:ssd, large96, huge96): single 1TB Enterprise NVME SSD
- Very high metadata performance
- Dedicated performance
- Data not shared
- Automatic cleanup at the end of each job
- Usage with environment variable `$LOCAL_TMPDIR`

Perm/Tape archvie



- StorNext HSM mounted via NFS on the login nodes
- 2 frontend NFS servers, 2 mangement servers
- 6 LTO-7 drives and tape library with approx 1200 6TB tapes
- Very slow
- Cost efficient storage of inactive data



New/Upgraded storagesystems

Lustre SSD Pool

- Extension of Work Storage
- 4 additional object storage servers with two DDN SFA200NVX
- 120TB
- 46 NVME SSDs
- High random IO performance (especially reading)
- Higher sequential IO performance than HDD pool with low process counts
- Usage with environment variable `$SHARED_SSD_TMPDIR` (automatic cleanup at the end of each job)
- Long term usage via striping setting (see HLRN documentation) or as additional folder below `/scratch/fast/usr` (support ticket)

IME

- Additional burst buffer storage system (very fast storage)
- 10 DDN IME 140 servers 90 NVME SSDs
- 300TB
- Extreme high performance for sequential and random I/O (up to 190GB/s)
- Relaxed POSIX semantics, so please use MPIIO or dedicated libraries (e.g. HDF5, NetCDF) for shared file write IO to ensure consistency
- Metadata performance low as IME is an additional layer on top of the limited Lustre metadata performance
- Only accessible from compute nodes
- Access via environment variable `$IME_TMPDIR` (automatic cleanup at the end of each job) or `$IME` for longer usage with multiple jobs after adding the option `--constraint=ime` to your job.

Documentation and Performance comparison

- Details about the different storage systems and performance comparison
  `https://www.hlrn.de/doc/display/PUB/Special+Filesystems`
- IME usage information
  `https://www.hlrn.de/doc/display/PUB/IME+Burst+Buffer%2C+File+System+Cache`
- Storage overview
  `https://www.hlrn.de/doc/display/PUB/File+Systems`
- Example for metadata profiling (open calls) with `strace`:
  `https://www.hlrn.de/doc/display/PUB/Metadata+Usage+on+WORK`
- IO profiling tool for MPI jobs:
  `https://www.mcs.anl.gov/research/projects/darshan/`