# File Systems

HLRN is operating at each site 3 central storage systems with their global file systems:

| File System | Capacity | Storage Technology and Function |
|---|---|---|
| HOME | 340 TiB | IBM Spectrum Scale file system, exported via NFS to compute and login nodes<br>• includes centrally managed software and the module system in `/sw` |
| WORK | 8 PiByte | DDN ExaScaler with Lustre parallel file system |
| PERM | | Tape archive with multiple petabyte capacity with additional harddisk caches |

The system Emmy has additional storage options for high IO demands:

- Phase 1 nodes (partitions `medium40` and `large40`): Local SSD for temporary data at `$LOCAL_TMPDIR` (400 GiB shared among all jobs running on the node). The environment variable `$LOCAL_TMPDIR` is available on all nodes, but on the phase 2 systems it points to a ramdisk.
- DDN IME based burst buffer with 48TiB NVMe storage (general availability together with the phase 2 nodes)

## HOME

Each user holds one HOMR directory on each compute site Emmy and Lise.

- directory `HOME=/home/${USER}`
- for a higher number of files
  - configuration files
  - source code and executables
- limited disk space
- backup is available

The home filesystem and `/sw` are mounted via NFS, so performance is medium. We take daily snapshots of the filesystem, which can be used to restore a former state of a file or directory. These snapshots can be accessed through the path `/home/.snapshots` or `/sw/.snapshots`. There are additional regular backups to restore the filesystem in case of a catastrophic failure.

## WORK

The Lustre based work filesystem `/scratch` is the main work filesystem for the HLRN clusters. Each user can distribute data to different directories.

- parallel input/output for production jobs
  - moderate number of files
  - transient nature of data
- no backup, no disaster recovery
- available directories
  - `WORK=/scratch/usr/${USER}`, for user data
  - project directory `/scratch/projects/<projectID>`, for project data
  - `TMPDIR=/scratch/tmp/${USER}`, applications and compilers store data temporarily

We provide no backup of this filesystem. The storage system of Emmy provides around 65GiB/s streaming bandwith and Lise around 85GiB/s during the acceptance test. With higher occupancy, the effective (write) streaming bandwidth is reduced.

The storage system is hard-disk based (with SSDs for metadata), so the best performance can be reached with sequential IO of large files that is aligned to the fullstripe size of the underlying RAID6 (Emmy 1MiB, Lise 16MiB).

If you are accessing a large file (1GiB+) from multiple nodes in parallel, please consider to activate striping of the file with the Lustre command `lfs setstripe` (specific to this file or for a whole directory, changes apply only for new files, so applying a new striping to an existing file requires a file copy) with a sensible `stripe_count` (recommendation: Emmy up to 32, Lise up to 8) and a `stripe_size`, which is a multiple of the RAID6 fullstripe size and matches the IO sizes of your job.

A general recommendation for network filesystems is to keep the number of metadata operations for open and closing files, as well as checks for file existence or changes as low as possible. These operations often become a bottleneck for the IO of your job and on large clusters, as the ones operated by HLRN, can easily overload the file servers.

## PERM, tape archive

The magnetic tape archive provides additional storage for inactive data to free up space on the WORK or HOME filesystem. It is directly accessible on the login nodes..

- directory `/perm/${USER}`
- secure file system location on magnetic tapes

- no solution for long-term data archiving
- no guarantee for 10 years according to rules for good scientific practice

Emmy provides the additional option to access the PERM archive via `ssh` to the archive nodes `gperm1` and `gperm2`, so you can use `rsync, scp, sftp` for file transfer.

For reasons of efficiency and performance, small files and/or complex directory structures should not be transferred to the archive directly. Please aggregate your data to compressed tarballs or other archive containers with a maximum size of 5,5TiB before copying your data to the archive.