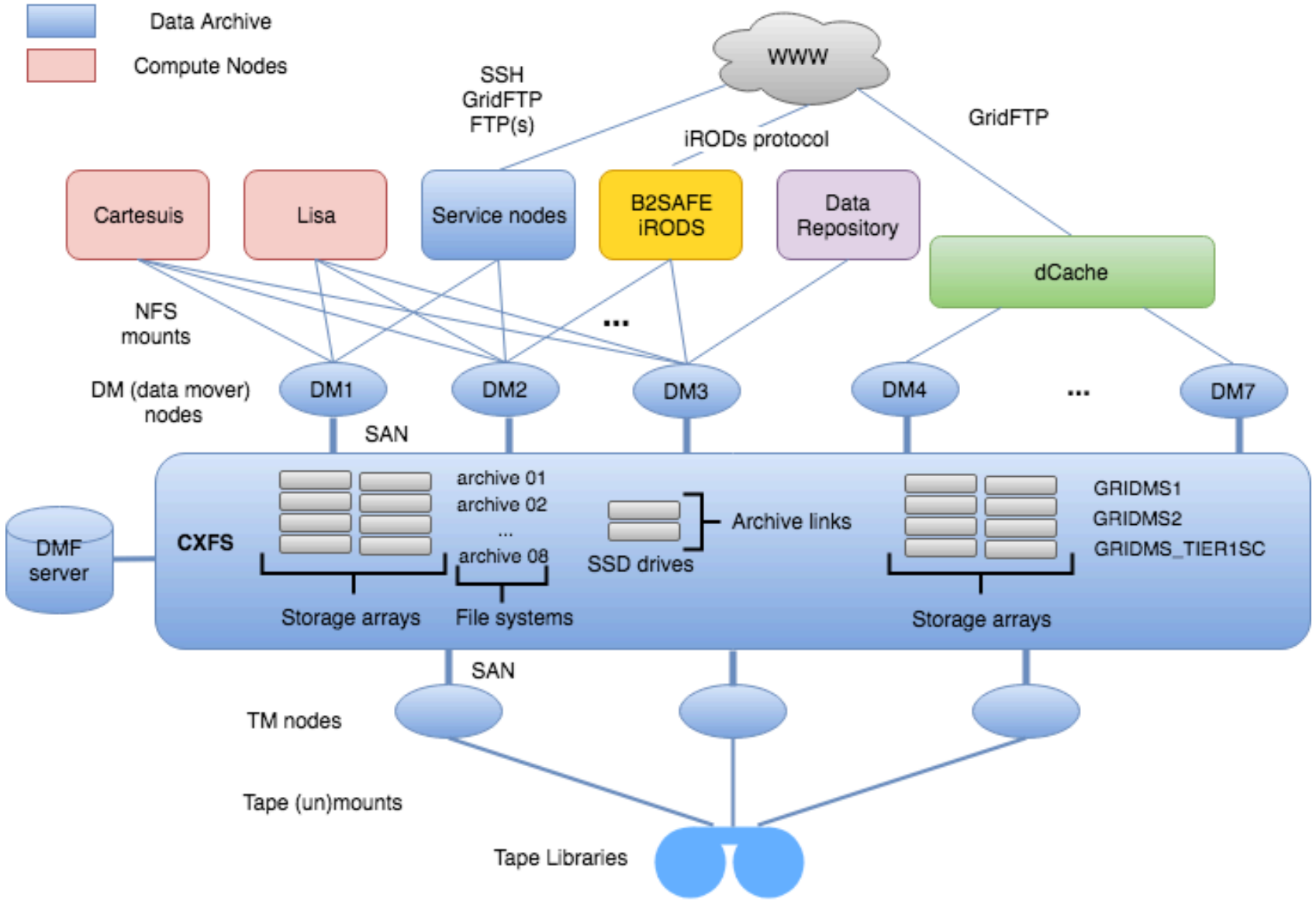
DATA ARCHIVE INFRASTRUCTURE

SURF SARA

Data Archive – Long-term storage

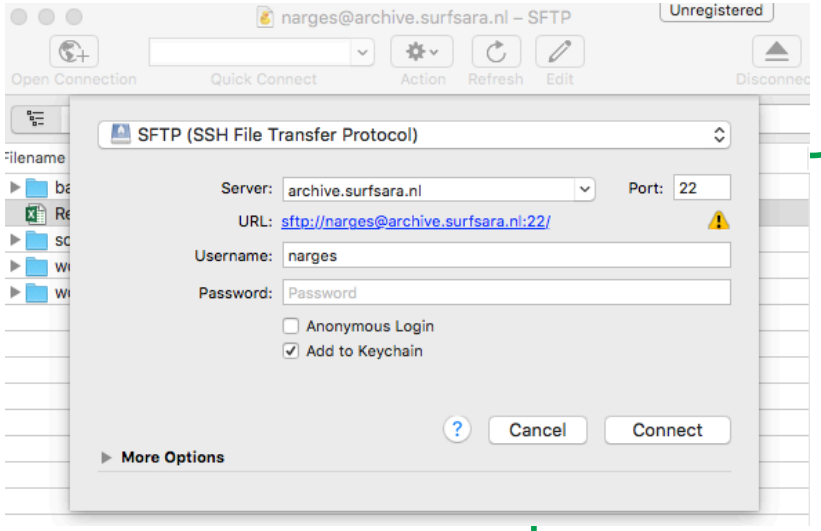
- Long-term storage of data
- Storage medium: Tape high latency
- Powerful transfer protocols (gridfTp, rsync, scp)
- Easy access from HPC services lisa and cartesius via NFS mounts use archive as yet another directory

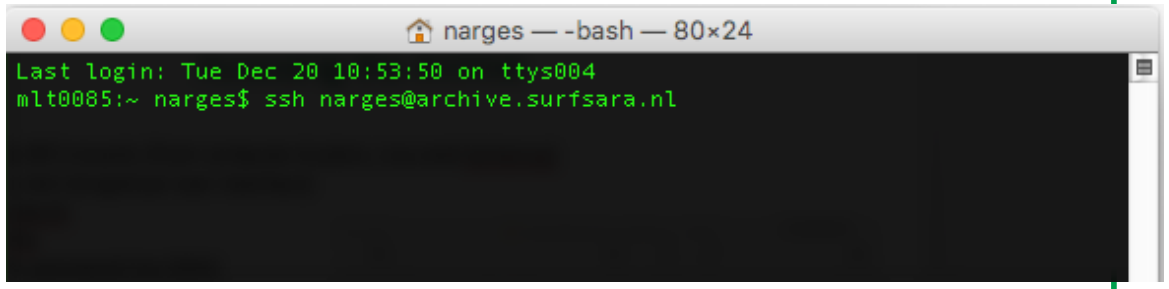
Data Archive Infrastructure



Data Archive Access

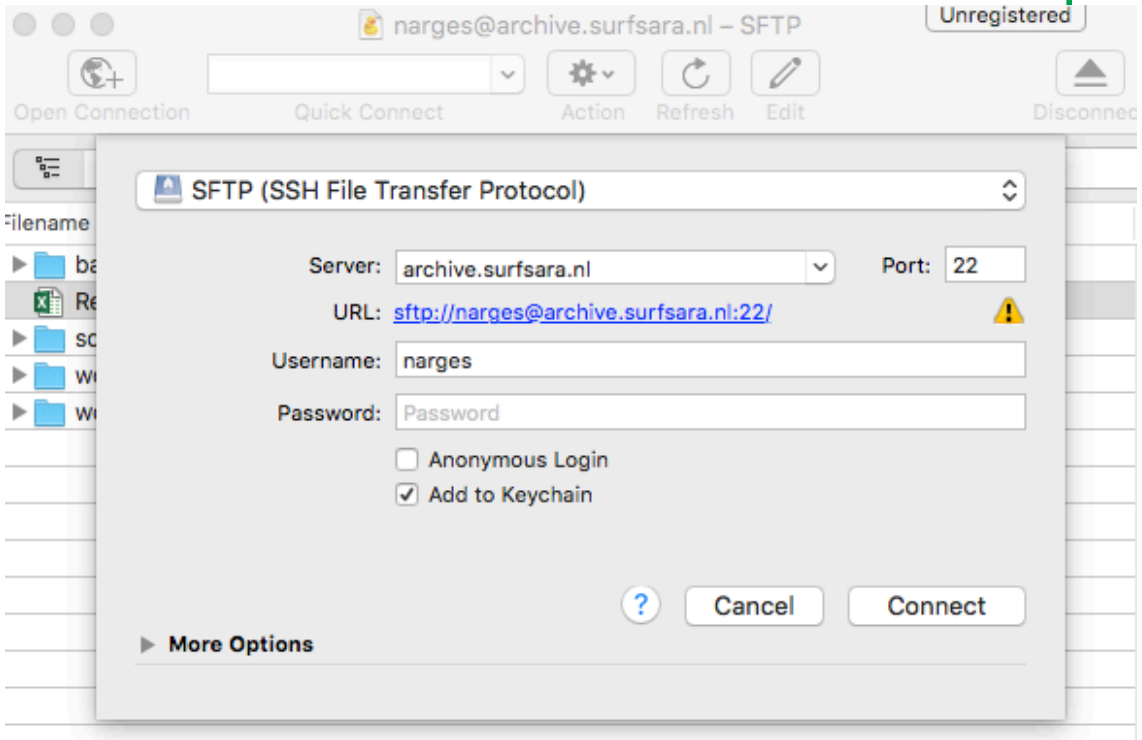
- Access via graphical user interface (GUI): A transfer client that support SSH File Transfer Protocol (SFTP)
  - Cyberduck (Mac and Windows)
  - Filezilla (Linux)
  - MobaXterm (Windows)
- Access via command line interface (CLI)
  - Terminal (preinstalled on Mac and Linux)
  - MobaXterm (Windows)
- Access via NFS mounts (Also via command line, only possible from compute clusters, Lisa and Cartesius)





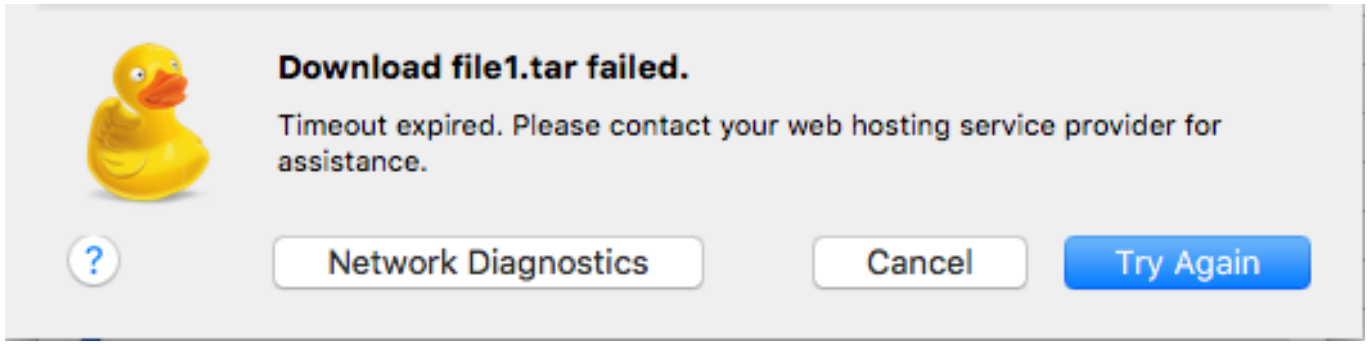
Access Data Archive via GUI

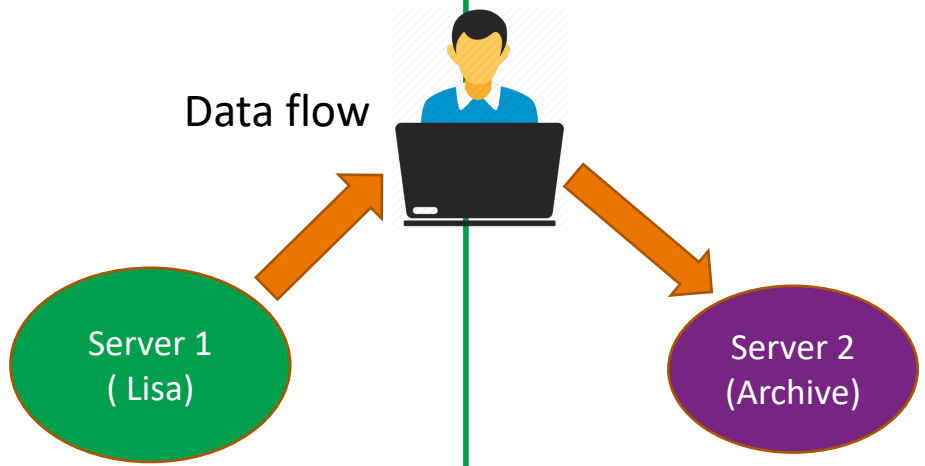
- Cyberduck is a standalone client that runs on Windows and Mac OSX
- Download and install: http://cyberduck.ch/
- To start an Archive session with Cyberduck:
  - Start Cyberduck
  - Click on 'Open connection': You now see this screen
  - Choose the following options:
    - Connection type: SFTP (SSh File Transfer Protocol)
    - Server: archive.surfsara.nl
    - port: 22
    - Login with your credentials (sdemo<xxx>)



Limitations of GUI Access

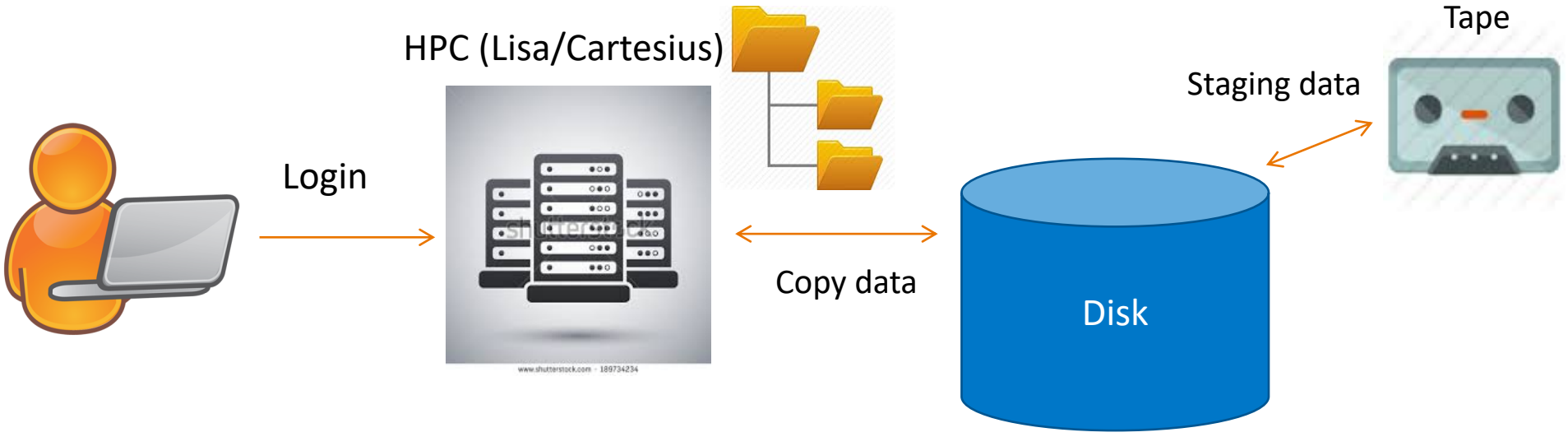
- The data flows via the user laptop. Therefore the transfer depends on your local storage and connectivity (If the connection is lost, the transfer is lost).
- Only for small data files
- Does not always work for fetching data (data needs to be staged first)
- You can't see the status of the data (i.e. weather the data is on disk or on tape).
- Unclear error messages





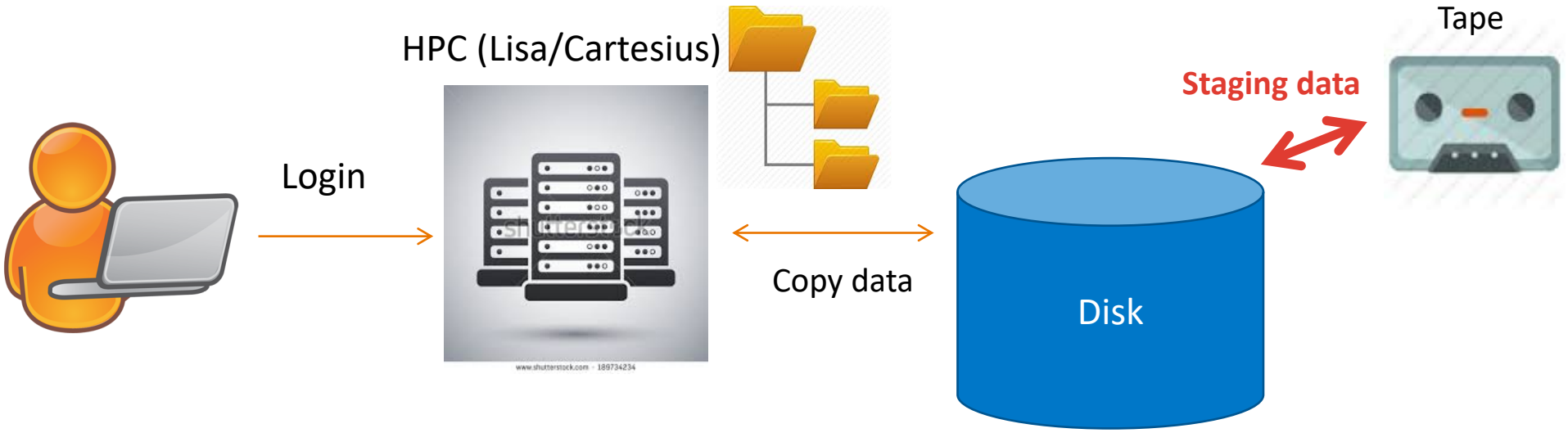
Archiving Workflows on HPC

- Login to the HPC system (Lisa / Cartesius)
  - User's archive home directory is mounted as folder /archive/<username>
- Some data to work with
  - Retrieve data from the Archive/or copy from any other source
- Process and Analyze your Data on HPC
- Archive the data
  - Using tar or "dmftar" tool



Staging Data

- Data on the Archive is stored on Tape
- Staging data: copy the data from tape to disk
- Always stage the data first from the archive, before you start to work with the data (read/write actions)
- Use the dm commands for staging data (“dmget” command)



Archive Usage – Best practicies

- Try to store files of significant size (> 1 GB) as much as possible. Smaller files will always be accepted, but will lower the performance of restoring your files from tape.
- If you have many small files, make sure to pack them using a file archiving tool like tar or dmftar.
- Try to pack your files before uploading them to the archive.
- Organize your files in such a way that in case the files are needed again only parts of the data set need to be restored from tape.
- Avoid storing unpacked software packages, these usually contain a lot of small files. Instead pack these as well, or refer to a specific software repository.

Optimal Archiving with dmftar

- Wrapper for GNU tar, developed in-house by SURFsara.
- Creates archive files of any size (default 10 GB).
- Can be used remotely to transfer data to and from the archive file system.
- Available on Data Archive, Lisa cluster or Cartesius supercomputer. Also made opensource.
- Contains the same information as tarballs, plus more:
  - Checksum of each tarball (default checksum algorithm is md5, but others are supported as well, i.e sha1, sha224,…)
  - File index: list of files and directory structure
  - Understands underlying storage infrastructure: ‘tape-aware’ and automatically stages your archived files

SURF SARA

Archiving tools: syntax

- Staging data from tape on the archive:

```
dmget –a [file]
```

- Pushing data to tape on the archive:

```
dmput [–r] [file]
```

- dmftar syntax:

```
dmftar [TASK] [OPTIONS] –f <dmftar–archive> <input–files>
```

(Note: always use the right extension (‘.dmftar’ for your archive files!)

Hands-on Archiving Data

- Archiving data using dmftar
  - Login to LISA (command: ssh sdemoXXX@lisa.surfsara.nl)
  - Explore the environment
    - Connection to archive
    - DMF commands
- Start an archiving workflow with dmftar