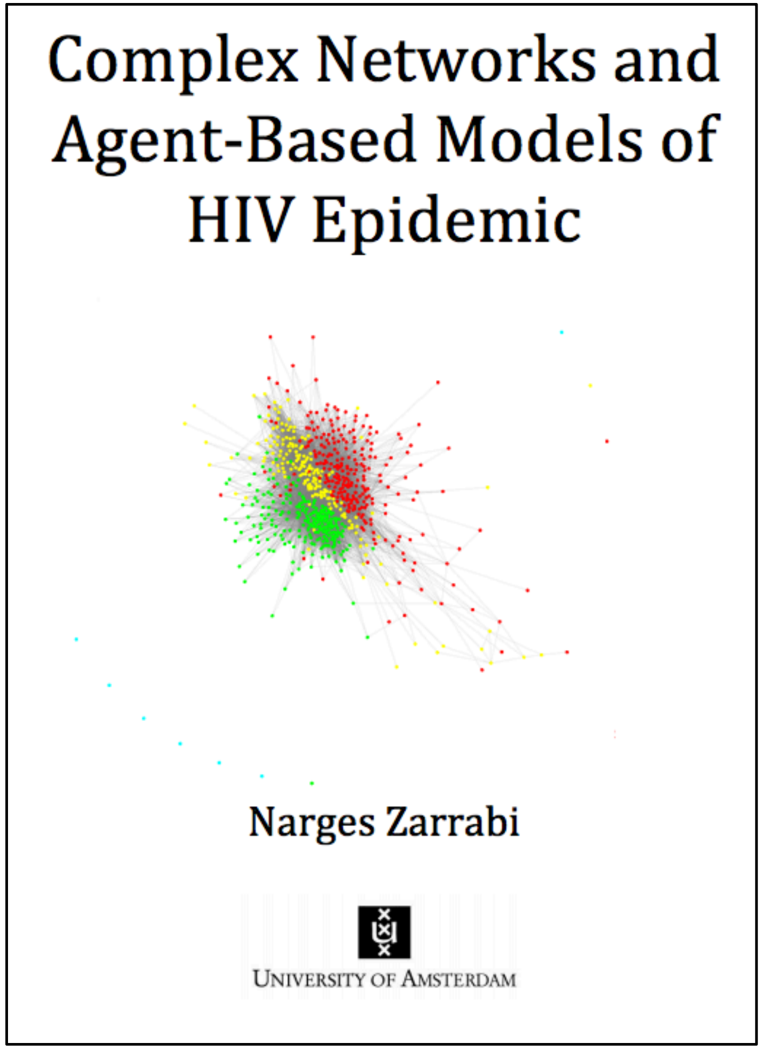
Data Archive
Infrastructure and GUI Access

Narges Zarrabi, SURFsara

UvA HPC Course 2017: Data Management

My Background

- Masters in Computational Science at UvA (2008)
- PhD in Computational Science at UvA (2010-2013)



Complex Networks and Agent-Based Models of HIV Epidemic

Narges Zarrabi

UNIVERSITY OF AMSTERDAM

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

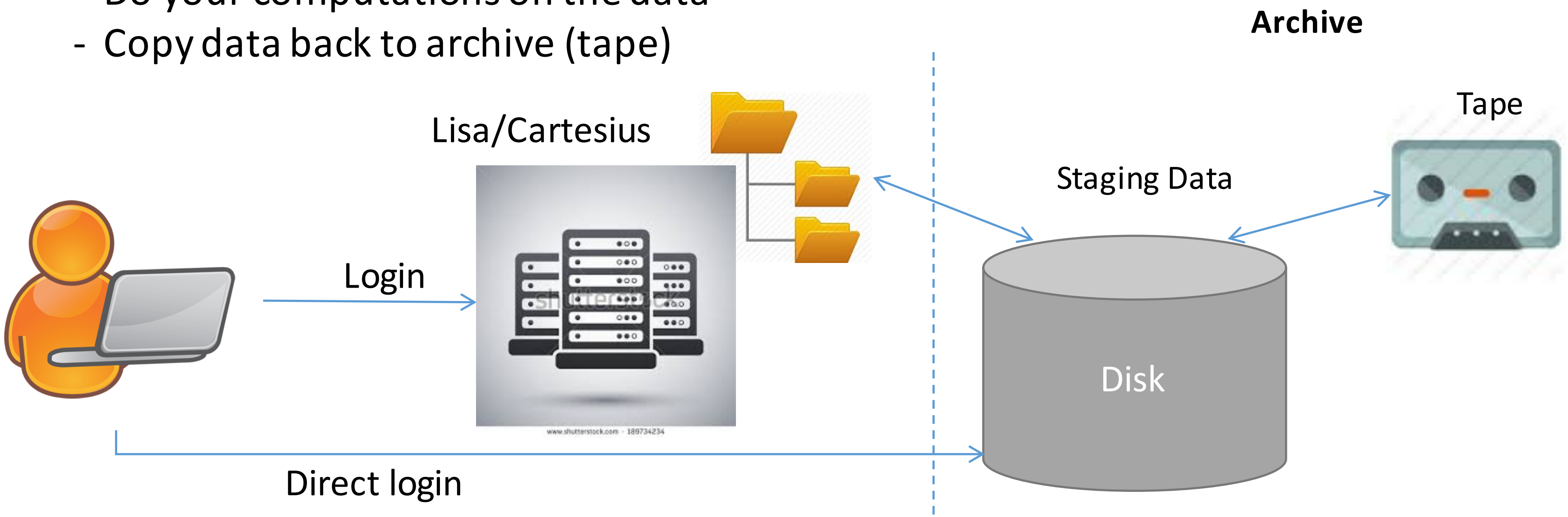
Data Archive Infrastructure

Data Archive infrastructure

- Direct access to Archive
- Access via HPC (NFS mounts, User sees the archive as another folder)

Workflow employing Archive from compute clusters at SURFsara:

- User logs in to Lisa/Cartesius
- Do your computations on the data
- Copy data back to archive (tape)



Archive Usage – Best practices

- 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.

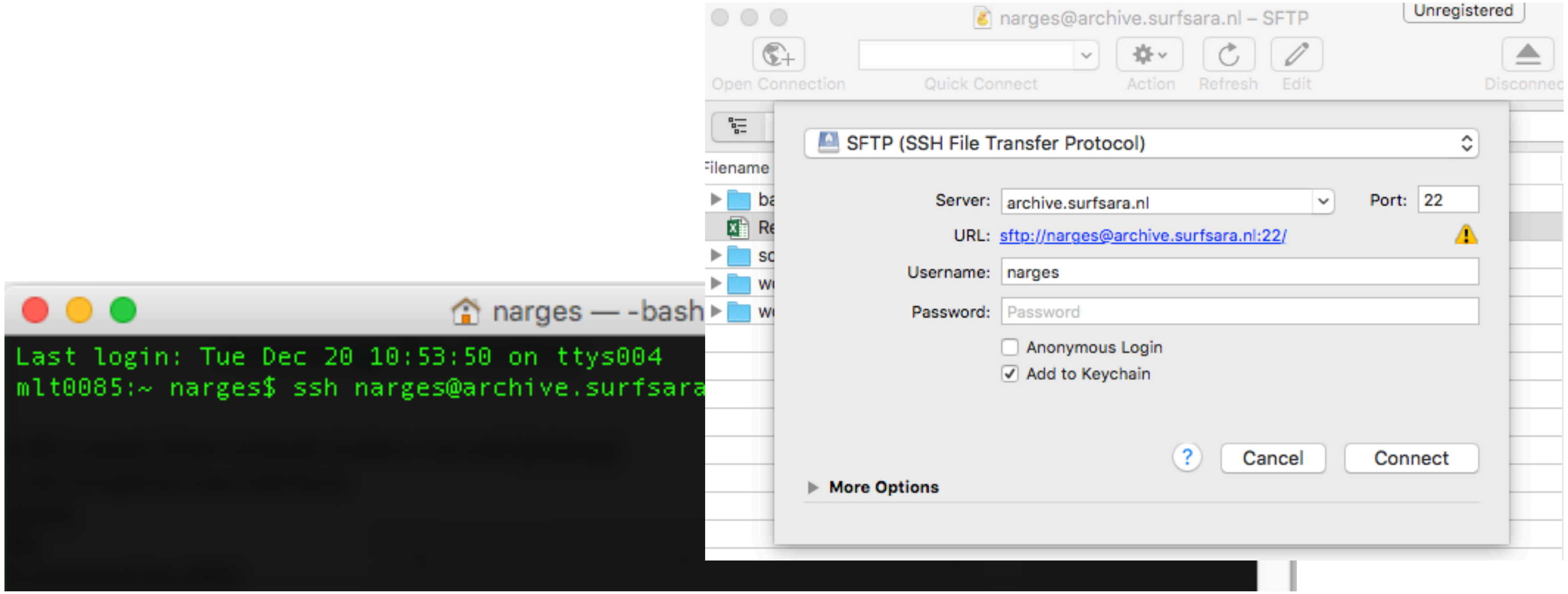
- Try to pack your files before uploading them to the archive.

- 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.

Accessing the Archive

- Access via graphical user interface (GUI)
- Access via command line interface (CLI)
- Access via NFS mounts (only possible from compute clusters, Lisa and Cartesius)



```
Last login: Tue Dec 20 10:53:50 on ttys004
mlt0085:~ narges$ ssh narges@archive.surfsara
```

narges@archive.surfsara.nl – SFTP

SFTP (SSH File Transfer Protocol)

Server: archive.surfsara.nl

Port: 22

URL: sftp://narges@archive.surfsara.nl:22/

Username: narges

Password: Password

□ Anonymous Login

☑ Add to Keychain

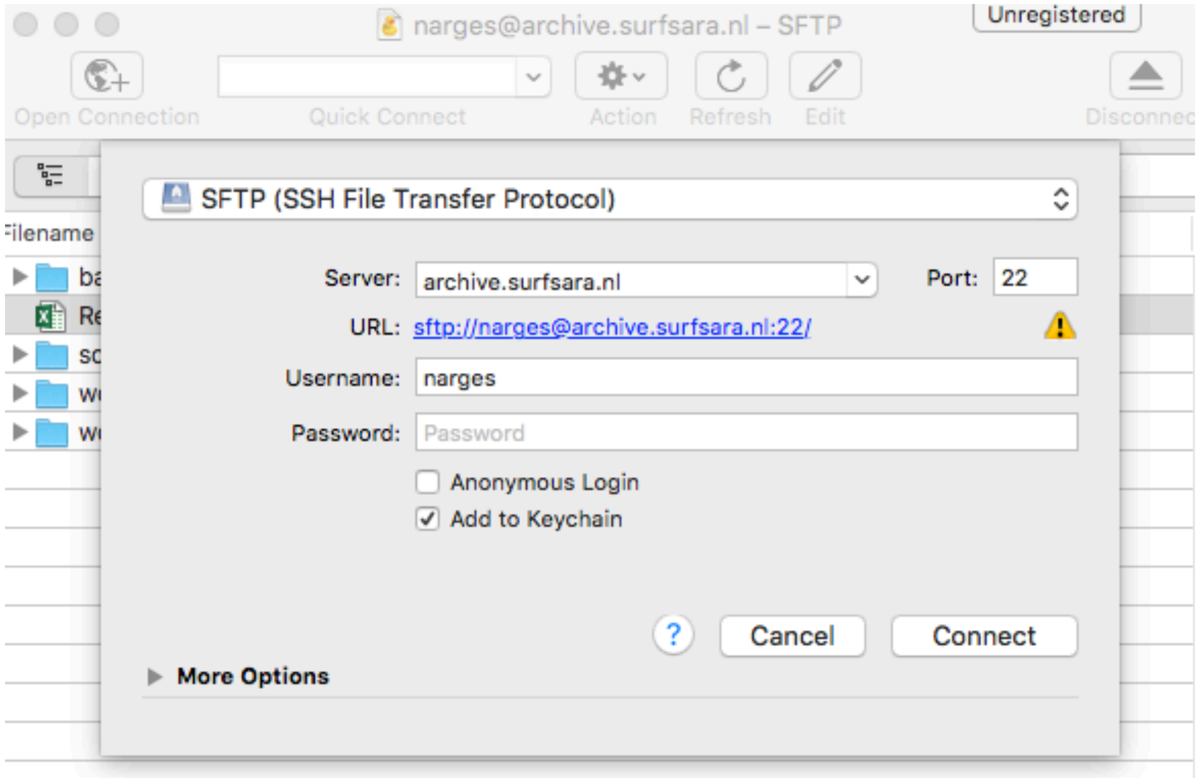
Access Archive via GUI

- Tools to access the Archive via GUI:
  - **Cyberduck** (Mac and Windows) →http://cyberduck.io/
  - **Filezilla** (Linux) → https://filezilla-project.org/
  - **MobaXterm** (Windows) → http://mobaxterm.mobatek.net/

Link to the hands on material:
https://github.com/sara-nl/dsdocs/blob/master/HPCcourse-UvA-20160131/Archive_GUI/achive-gui-handson.md

Access Archive via Cyberduck

- 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>)



narges@archive.surfsara.nl – SFTP

SFTP (SSH File Transfer Protocol)

Server: archive.surfsara.nl

Port: 22

URL: sftp://narges@archive.surfsara.nl:22/

Username: narges

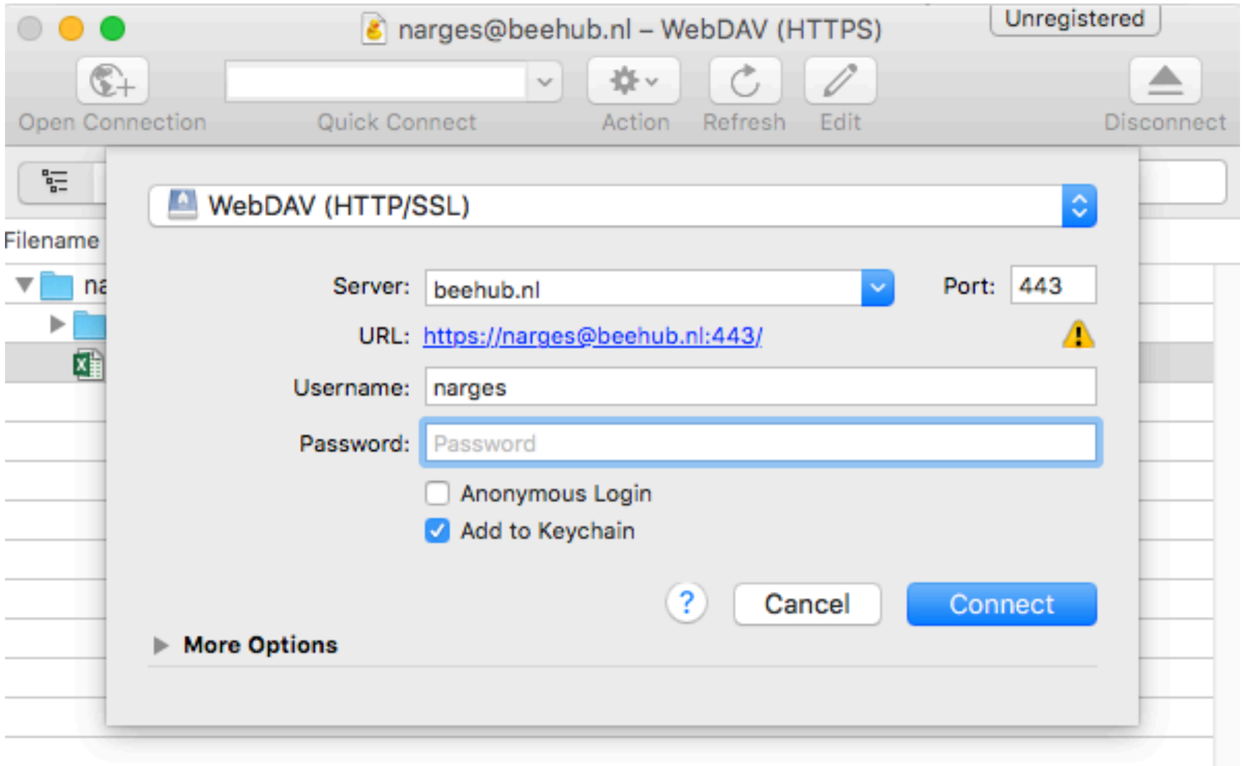
Password: Password

☐ Anonymous Login

☑ Add to Keychain

Access BeeHub via Cyberduck

- To start a BeeHub session with Cyberduck:
  - Start Cyberduck
  - Click on 'Open connection'
  - You now see this screen
  - Choose the following options:
    - Connection type: WebDAV (HTTP/SSL)
    - Server: beehub.nl
    - port: 443
    - Enter your BeeHub username and password as you use them on the website (not your sdemo credentials!)



narges@beehub.nl – WebDAV (HTTPS)

WebDAV (HTTP/SSL)

Server: beehub.nl

Port: 443

URL: https://narges@beehub.nl:443/

Username: narges

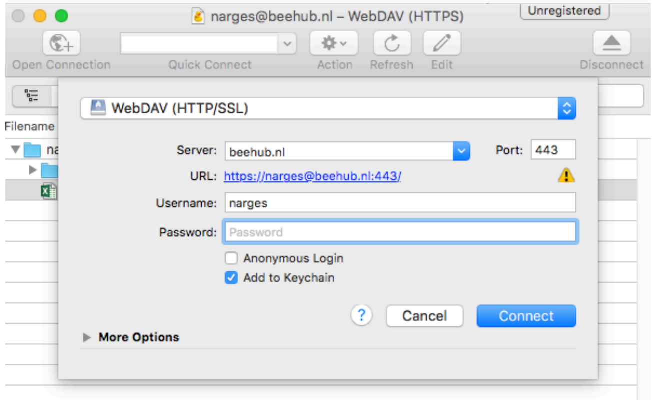
Password: Password

☐ Anonymous Login

☑ Add to Keychain

Transfer Data using Cyberduck

- To transfer data between services using Cyberduck:
  - Start Cyberduck
  - Establish a connection to the Archive
  - Establish another connection to BeeHub
  - Simply drag and drop files to transfer data



WebDAV (HTTP/SSL)

Server: beehub.nl Port: 443

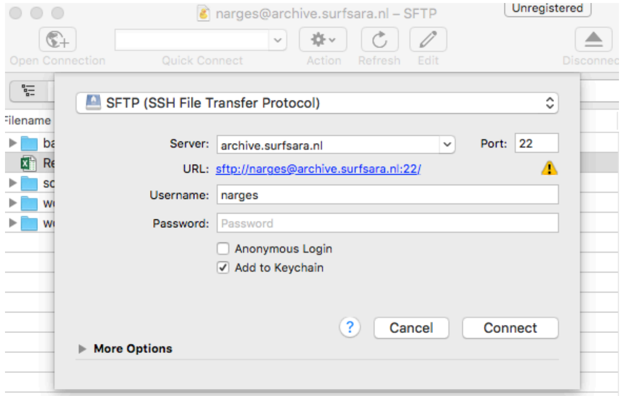
URL: https://narges@beehub.nl:443/

Username: narges

Password: Password

☐ Anonymous Login

☑ Add to Keychain



SFTP (SSH File Transfer Protocol)

Server: archive.surfsara.nl Port: 22

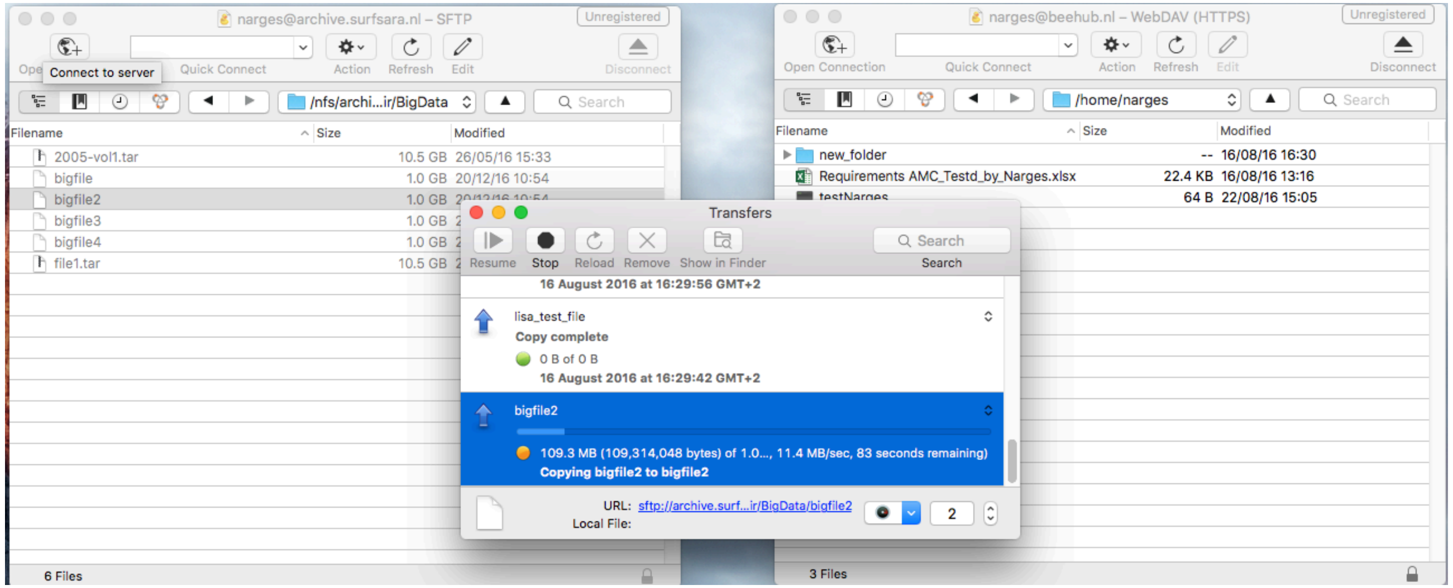
URL: sftp://narges@archive.surfsara.nl:22/

Username: narges

Password: Password

☐ Anonymous Login

☑ Add to Keychain



/nfs/archi...ir/BigData

Filename	Size	Modified
2005-vol1.tar	10.5 GB	26/05/16 15:33
bigfile	1.0 GB	20/12/16 10:54
bigfile2	1.0 GB	20/12/16 10:54
bigfile3	1.0 GB	2
bigfile4	1.0 GB	2
file1.tar	10.5 GB	2

6 Files

/home/narges

Filename	Size	Modified
new_folder	--	16/08/16 16:30
Requirements AMC_Testd_by_Narges.xlsx	22.4 KB	16/08/16 13:16
testNarges	64 B	22/08/16 15:05

3 Files

16 August 2016 at 16:29:56 GMT+2

lisa_test_file
Copy complete
0 B of 0 B
16 August 2016 at 16:29:42 GMT+2

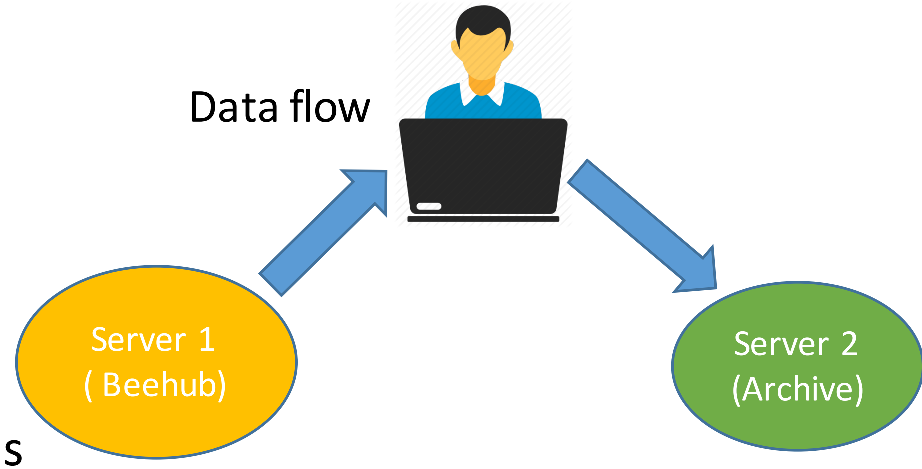
bigfile2
109.3 MB (109,314,048 bytes) of 1.0..., 11.4 MB/sec, 83 seconds remaining)
Copying bigfile2 to bigfile2

URL: sftp://archive.surf...ir/BigData/bigfile2
Local File:

2

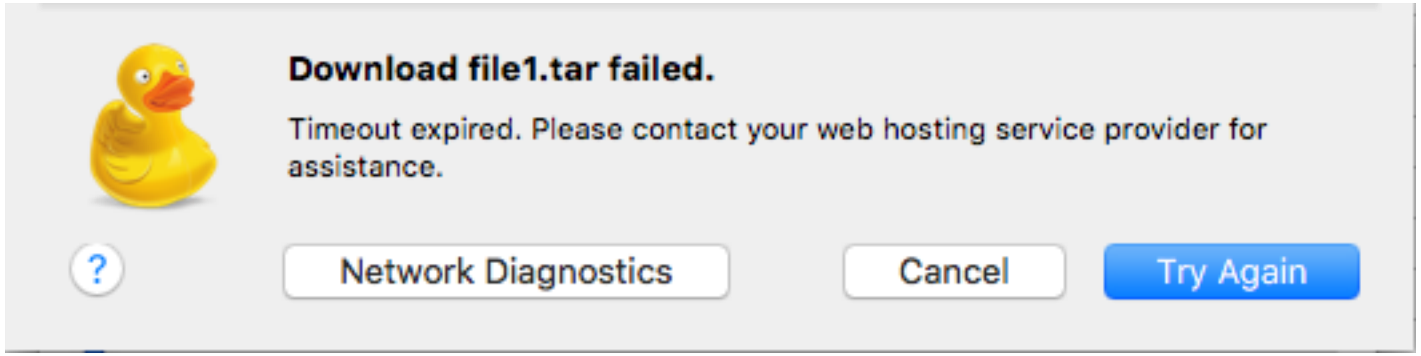
Advantages & Limitations

- Advantages:
  - Easy data transfer (to the archive)
  - Good for dumping data to the archive, and not fetching data
  - Transfer data between services (Only possible for small data)
  - Can be accessed from Windows, Mac and Linux machines
- Limitations
  - 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).



Transfer Data using Cyberduck

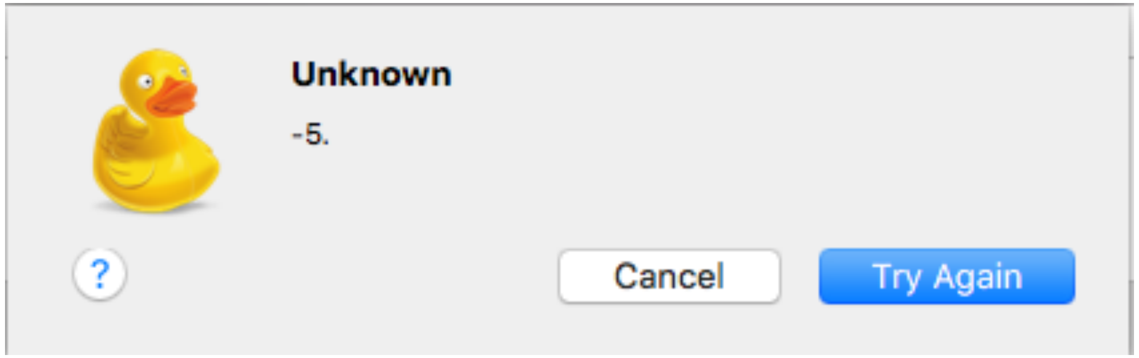
- Error: If the file is on tape, and not on disk. The files needs to be stages first.



Download file1.tar failed.

Timeout expired. Please contact your web hosting service provider for assistance.

- Error: If the internet connection is lost.



Unknown

-5.