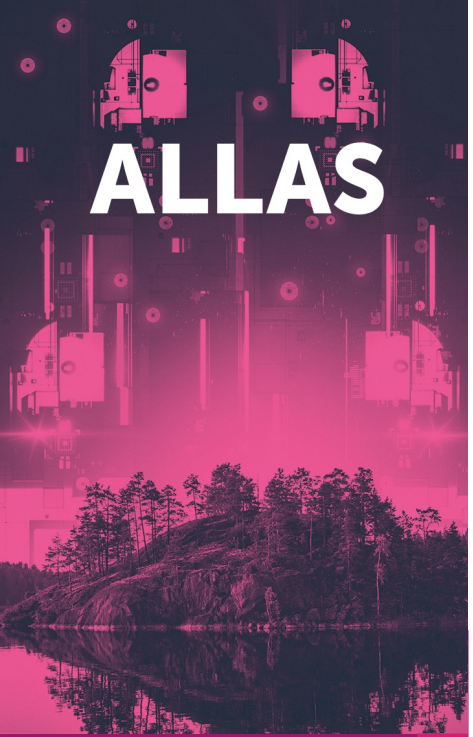
CSC

ICT Solutions for Brilliant Minds

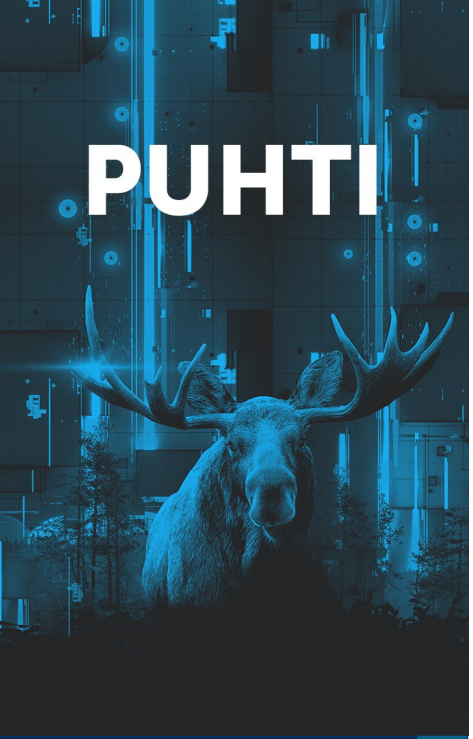
Storage in the Puhti



Allas – How to get access to Puhti and Allas

Use **https://my.csc.fi** to
1) Register to CSC (haka)
2) Set up a project at CSC (Principal Investigator)
3) Apply for Puhti and Allas service, quota and billing units for your project
4) Add other registered users to your project
5) Members have to register and accept the services in https://my.csc.fi

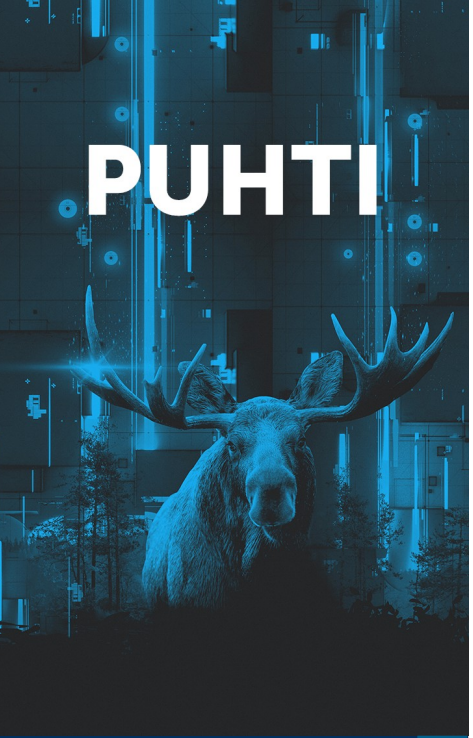
All project members have equal access to the data in Puhti and Allas.



Disk areas in Puhti

Directory	Default size	Default number of files	Accessibility
Home ($HOME)	10 GB	100 000	User only
/projappl/project_name	50 GB	100 000	Project members
/scratch/project_name	1 TB	1 000 000	Project members

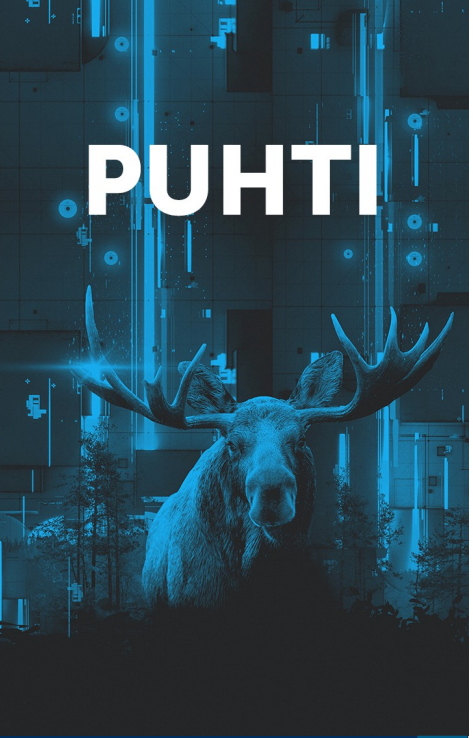
- No more personal $WRKDIR
- No more persistent project directories
- No more DONOTREMOVE-dir



Disk areas in Puhti

Directory	Default size	Default number of files	Accessibility
Home ($HOME)	10 GB	100 000	User only
/projappl/project_name	50 GB	100 000	Project members
/scratch/project_name	1 TB	1 000 000	Project members

- Default quotas are small → ask more quota if needed
- There is a limit for the number of files
- All project members share the **scratch** and **projappl** directories, but you can make private sub directories
- You can belong to several projects and thus have access to several project and scratch directories



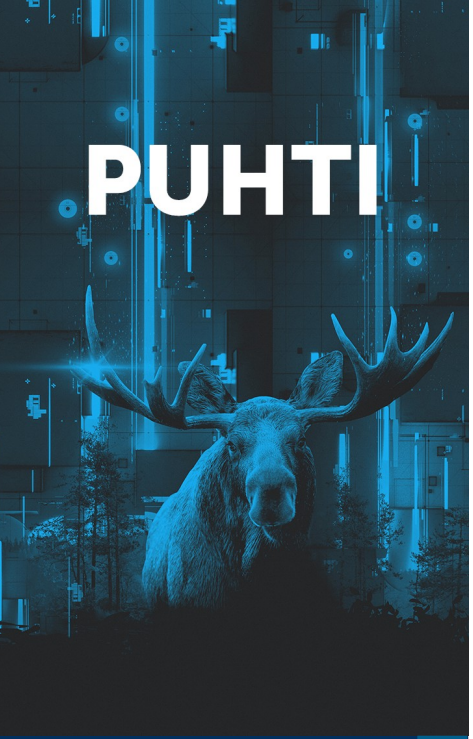
Disk areas in Puhti

Use **csc-workspaces** to check available projects and quotas

```
[kkmattil@puhti-login1 ~]$ csc-workspaces
Personal home folder        Quota
--------------------------------------------------------------------------------
/users/kkmattil             Capacity:   1.945G/10G    Files:    22.29k/100k

Project applications        Quota
--------------------------------------------------------------------------------
/projappl/project_2000178   Capacity:   3.054G/50G    Files:    23.87k/100k
/projappl/project_2000828   Capacity:     128k/50G    Files:        0k/100k
/projappl/project_2001659   Capacity:   48.63G/100G   Files:   239.00k/250k
/projappl/project_2001740   Capacity:       4k/50G    Files:        0k/100k
/projappl/project_2002078   Capacity:   4.271G/50G    Files:    76.61k/250k
/projappl/project_2002389   Capacity:       4k/50G    Files:        0k/100k

Project scratch             Quota
--------------------------------------------------------------------------------
/scratch/project_2000178    Capacity:   534.6G/1T     Files:     3.52k/1000k
/scratch/project_2000828    Capacity:   9.699M/1T     Files:     1.33k/1000k
/scratch/project_2001659    Capacity:   1.795T/8T     Files:  2496.07k/5000k
/scratch/project_2001740    Capacity:      40k/1T     Files:        0k/1000k
/scratch/project_2002078    Capacity:   11.84G/1T     Files:   213.02k/1000k
/scratch/project_2002389    Capacity:   3.876G/1T     Files:      .34k/1000k
```



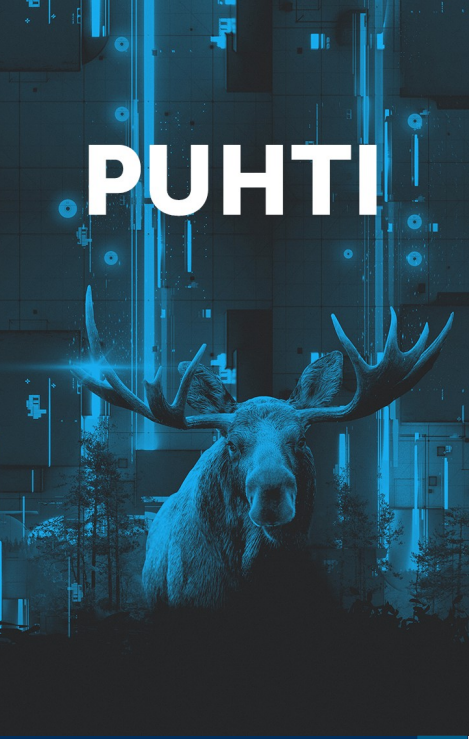
Importing data to Puhti

Your own data:

- **Allas** is preferred (https://docs.csc.fi/data/Allas/)
- Direct tasport with tools like **scp**, **WinSCP**, **Cyberduck** and **Filezilla** can be used too (https://docs.csc.fi/ → data → Moving data)

From the internet:

- **wget** and **curl** are general purpose tools for downloading public data directly to Puhti

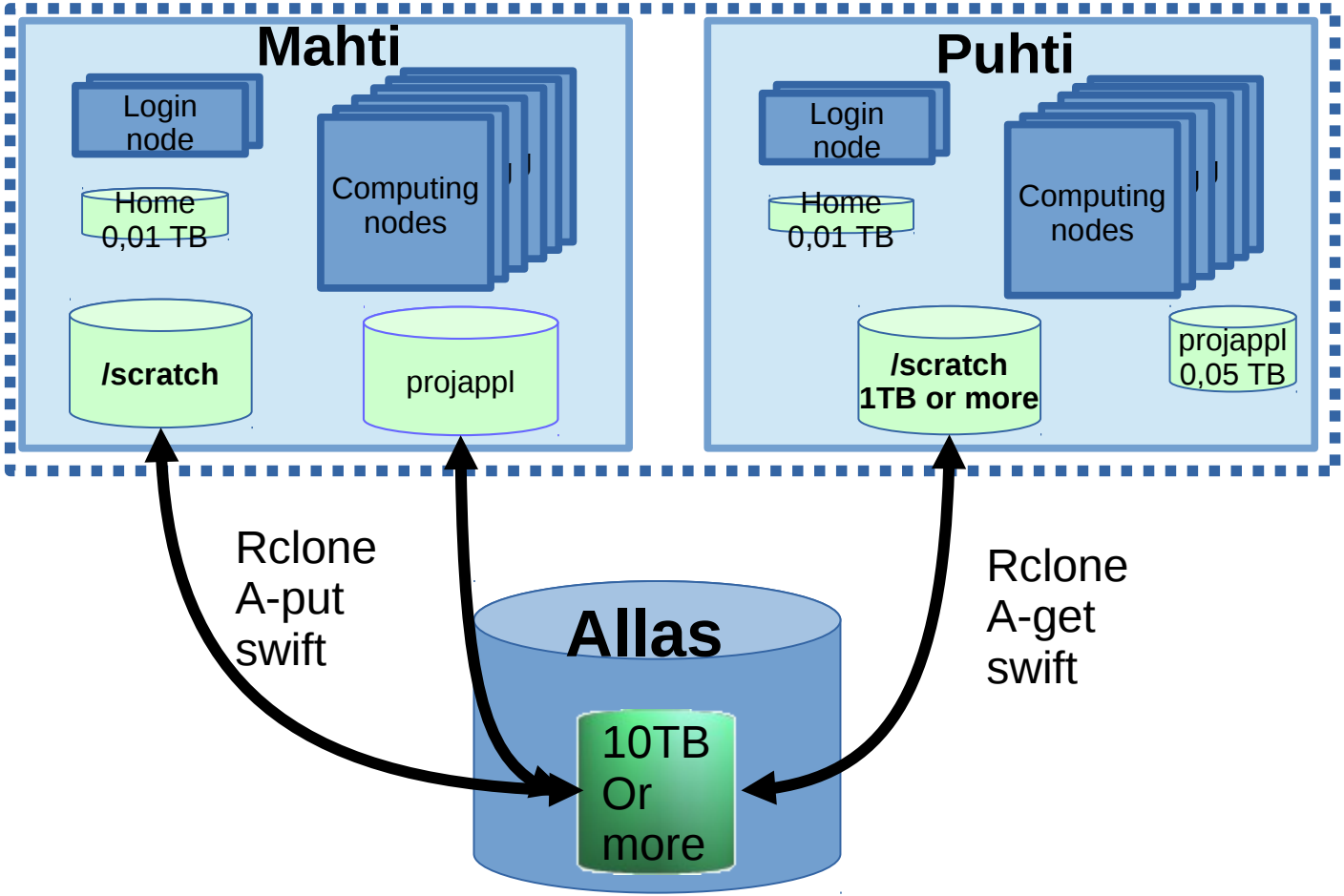


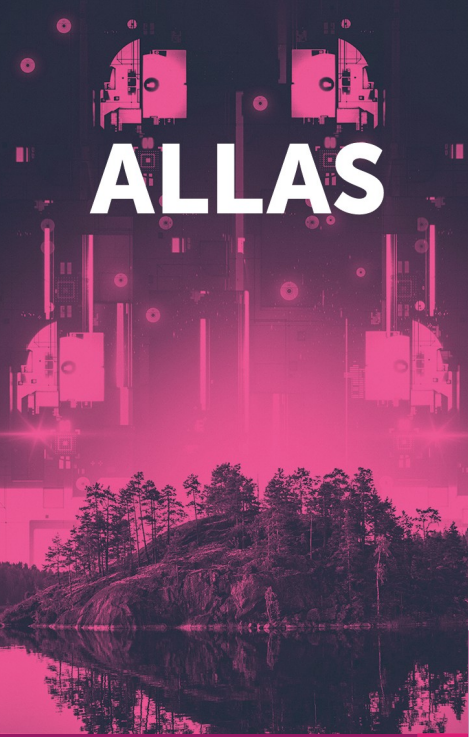
Importing data to Puhti

Bioscience specific tools:

- **Entrez edirect** (https://docs.csc.fi/apps/edirect/)
- **ENA browser tools** (enaDataGet, enaGroupGet)
- **SRAtoolkit** (fasterq-dump)
- **Illumina BS client** (https://docs.csc.fi/apps/bs/)

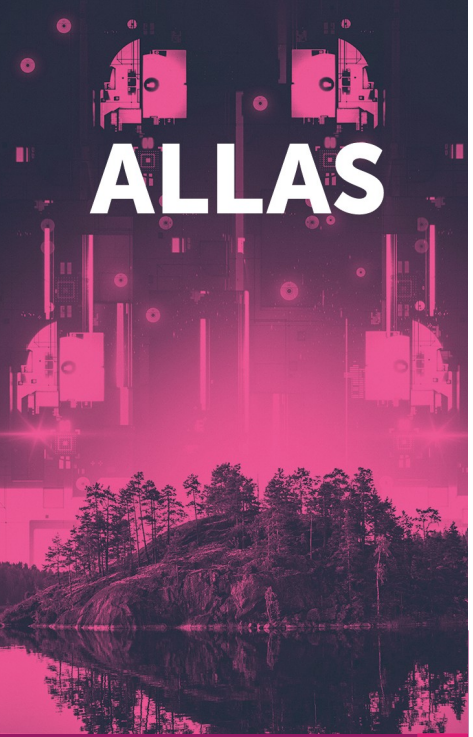
ALLAS


Mahti
Login node
Home 0,01 TB
Computing nodes
/scratch
projappl
Puhti
Login node
Home 0,01 TB
Computing nodes
/scratch 1TB or more
projappl 0,05 TB
Rclone A-put swift
Rclone A-get swift
Allas
10TB Or more



Allas – object storage: what it is for?

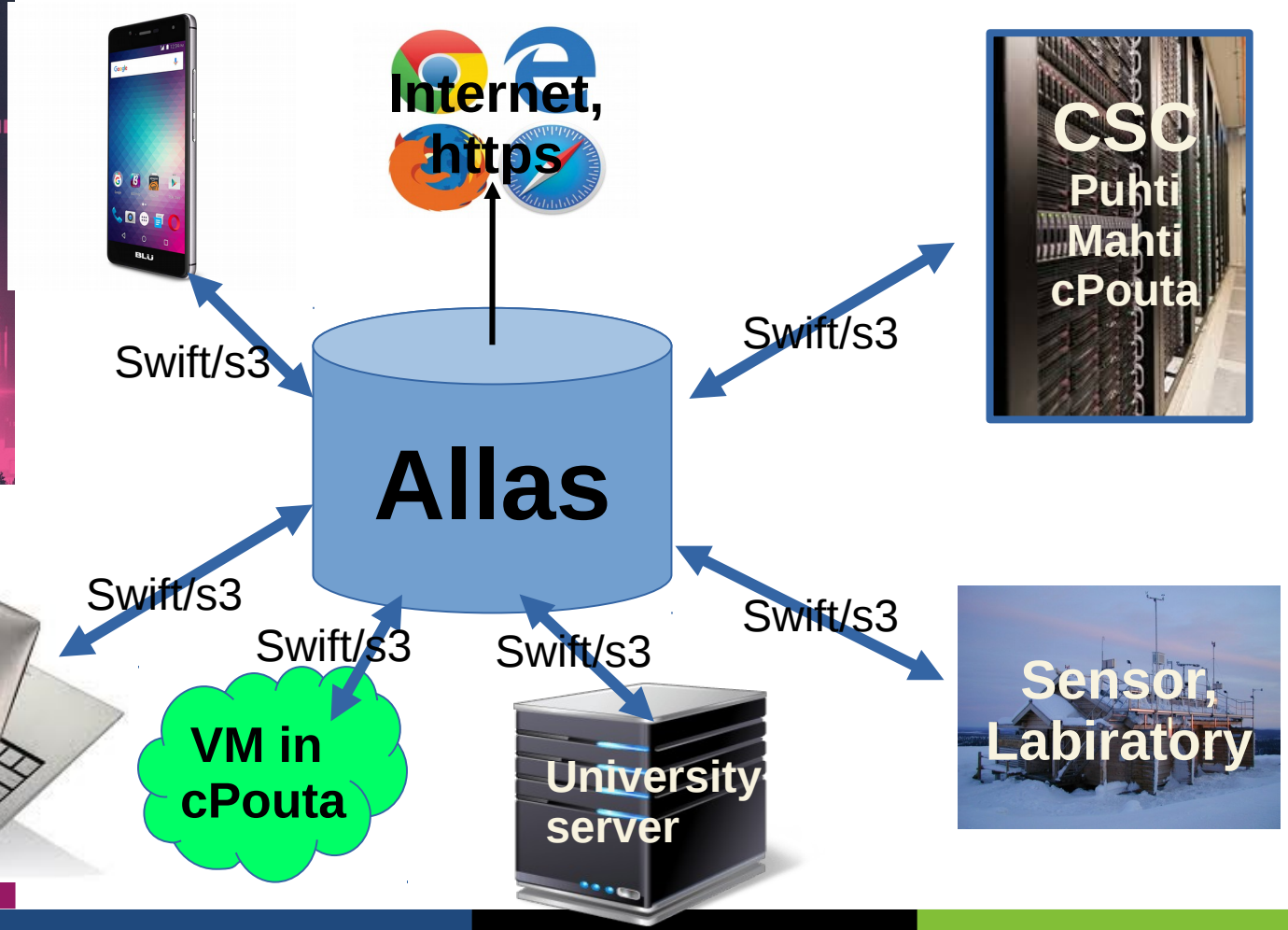
- Allas is new storage service for all computing and cloud services
- CEPH based object storage
- Meant for data during project lifetime
- Default quota 10 TB / Project.
- Possible to upload data from personal laptops or organizational storage systems into Allas
- Available in Puhti and Mahti
- Data can also be shared via Internet

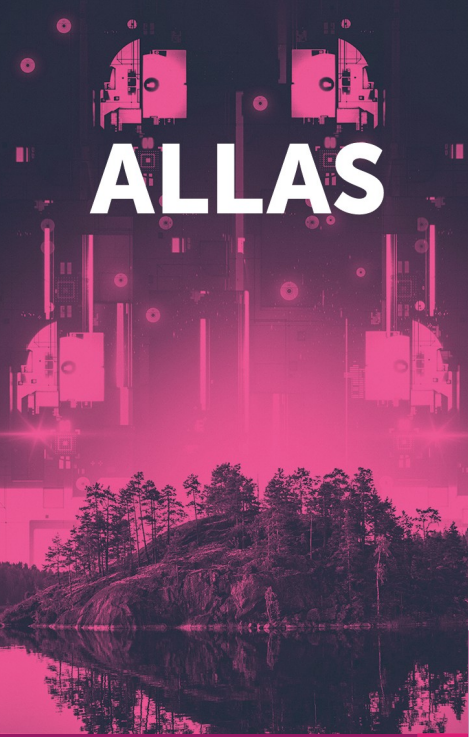


Allas – object storage: what it is for?

- Data can be moved to and from Allas directly without using supercomputer.
- For the computation the data has to be typically copied to a file system in some computer
- Data can be shared publicly to Internet, which is otherwise not easily possible at CSC.

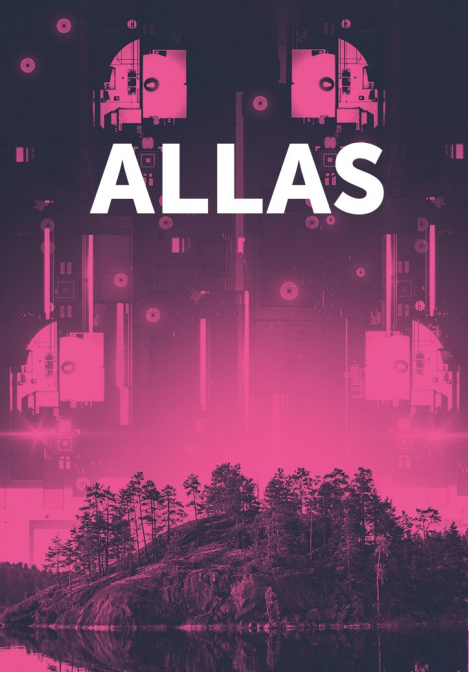
ALLAS

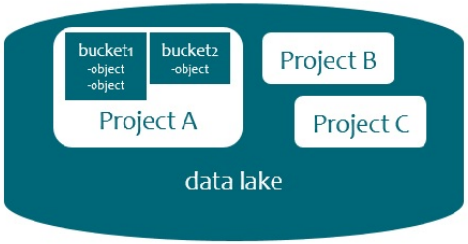

Internet, https
Swift/s3
CSC
Puhti
Mahti
cPouta
Allas
Your Laptop
VM in cPouta
University server
Sensor, Labiratory



Allas - storage

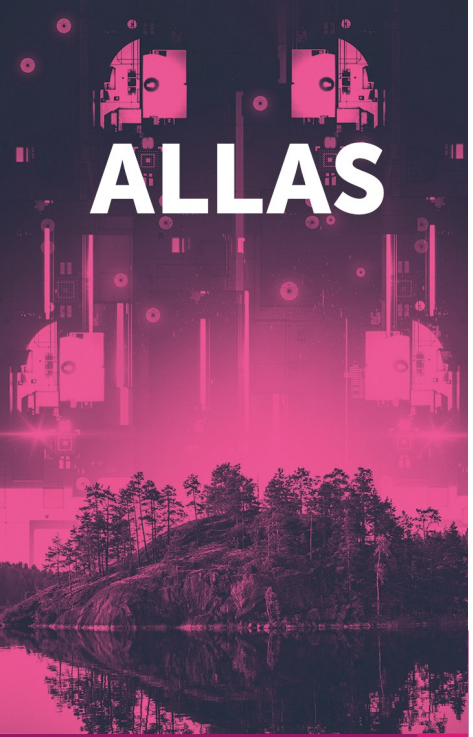
- An object is stored in multiple servers so a disk or server break does not cause data loss.
  - **There is no backup** i.e. if a file is deleted, it cannot be recovered
  - Data cannot be modified while it is in the object storage – data is immutable.
- Rich set of data management features to be built on top of it, initially S3 and Swift APIs supported





Allas – object storage: terminology

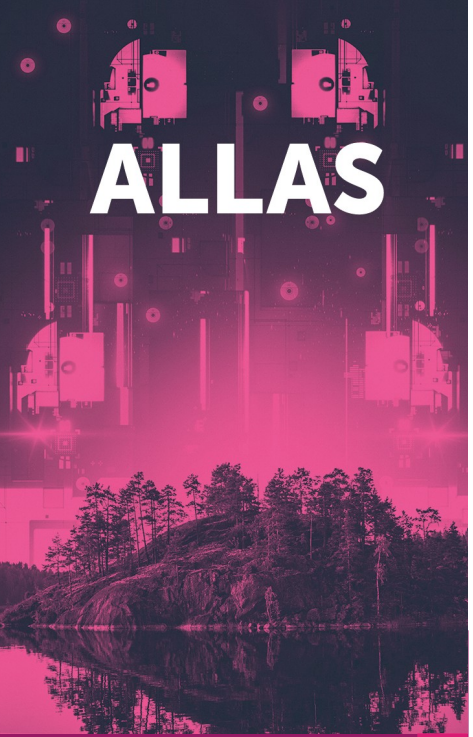
- Storage space in Allas is provided **per CSC project**
- Project space can have multiple **buckets** (up to 1000)
  - Only one level of hierarchy of buckets (no **buckets** within **buckets**)
- Data is stored as **objects** within a **bucket**
  - Blobs of data, can be anything (generally, **object** = file)
  - 500 000 objetcs / bucket
- Name of the **bucket** must be unique within Allas
- **Objects** can have metadata
- In reality, there is no hierarcical dirctory structure, even tough it sometimes looks like that.



Allas supports Two Protocols

- **S3** (used by: **s3cmd**)
- **Swift** (used by: **swift**, **rclone**, **a-tools**, **cyberduck**)

- Authentication is different
  - **S3:** permanent key based authentication – nice, easy and unsecure
  - **Swift:** authentication based on temporary tokens – more secure, requires authentication every 8 hours
- Metadata is handled in different ways
- Over 5G files are split in different ways
- → **Don’t cross-use Swift and S3 based objects!**



Allas Clients: read, write, delete

- Puhti, Mahti, Linux servers:

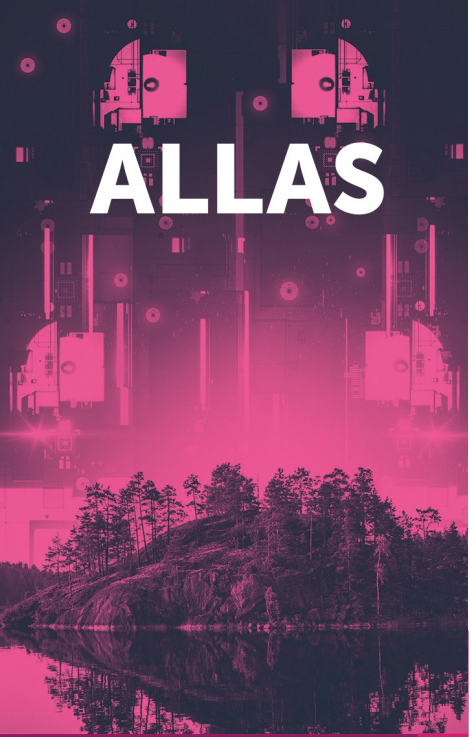
- rclone, switft, s3cdm, a-tools

- Virtual machines, small servers:

- in addition FUSE based virtual mounts

- Laptops:

- Cyberduck, FileZilla(pro), pouta-www interface



Allas – first steps for Puhti

1) Use https://my.csc.fi to apply Allas access for your project – Allas is not automatically available

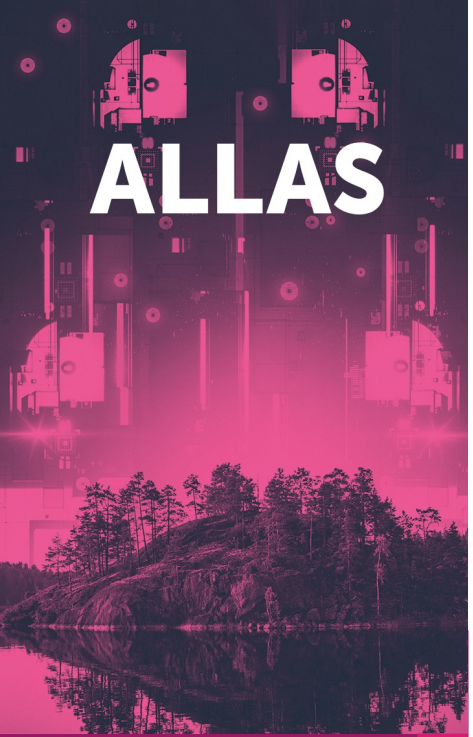
2) In Puhti and Taito, setup connection to Allas with commands:

```
module load allas

allas-conf
```

Study the manual and Start using Allas with rclone or a-tools:

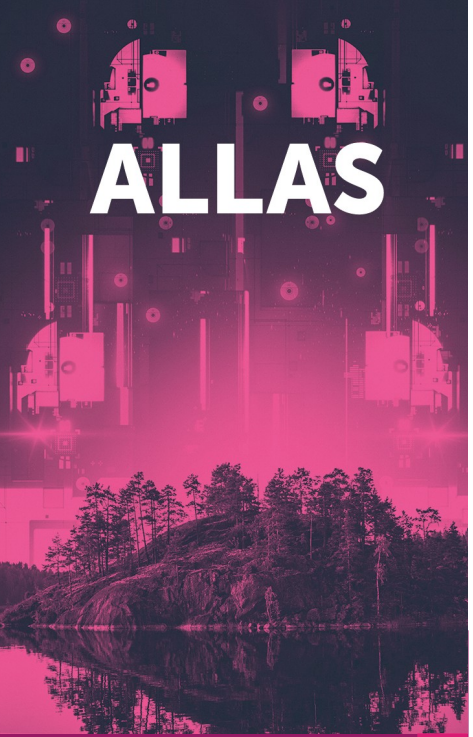
https://docs.csc.fi/#data/Allas/



Allas – rclone

- Straight-forward power-user tool with wide range of features
- Fast and effective.
- Available for Linux, mac and windows.
- Overwrites and removes data without asking!
- The default configuration at CSC uses swift-protocol but S3 can be used too.
- Use with care:

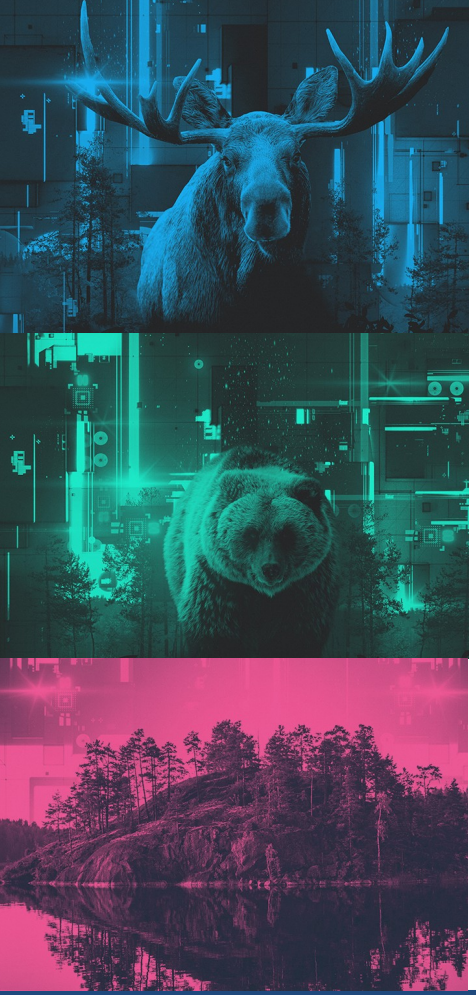
https://docs.csc.fi/#data/Allas/using_allas/rclone/

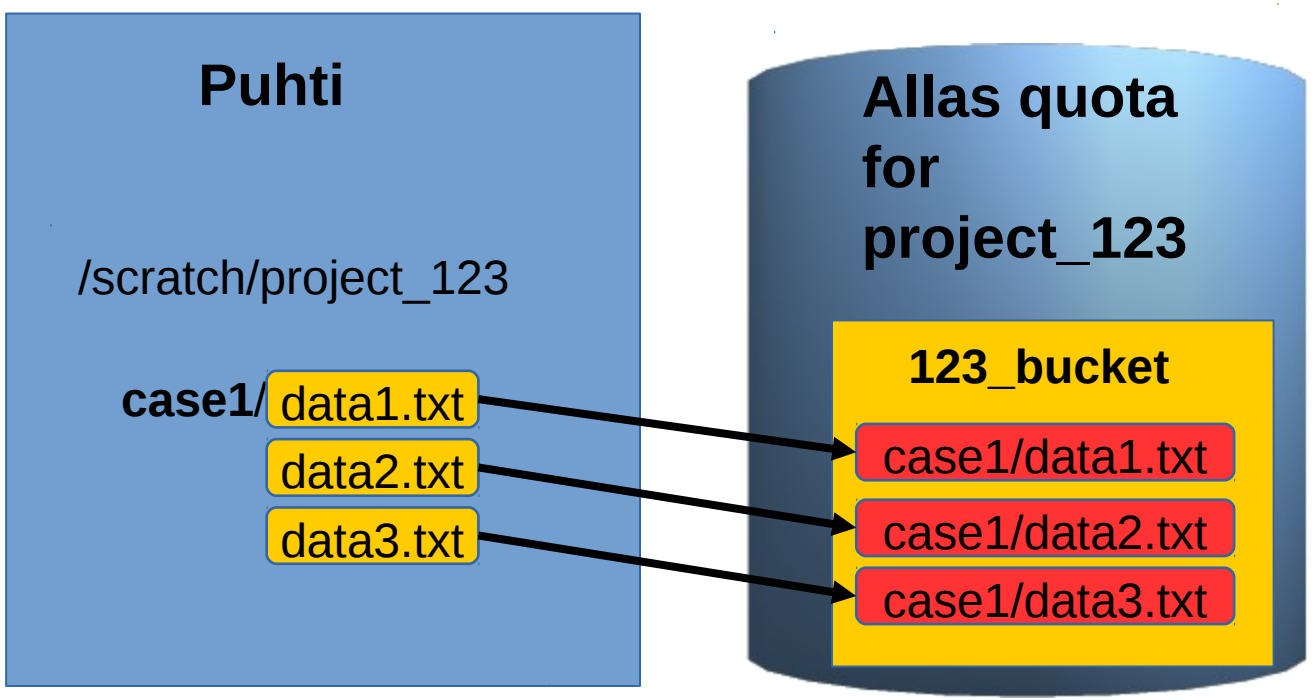


Allas – a-tools

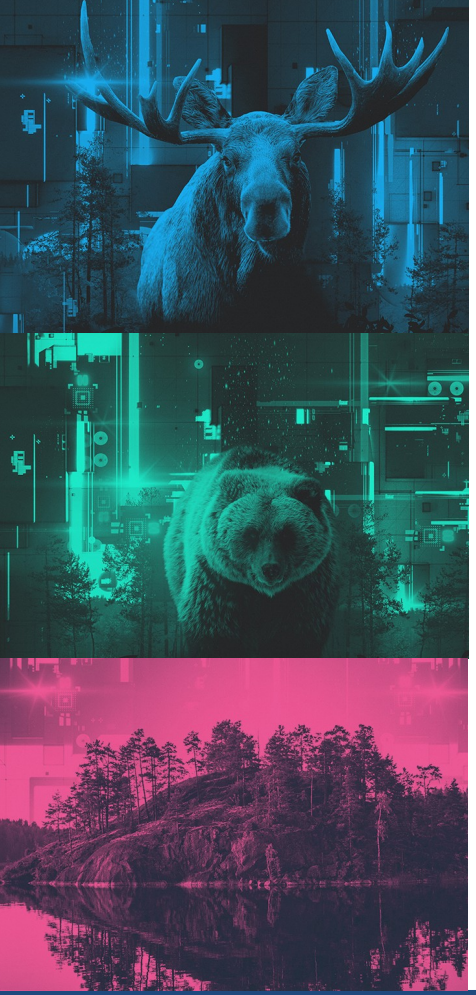
- Rclone based scripts for using Allas in Puhti
- A-tools try to provide easier and safer way to use Allas for occasional Allas user users.
- Developed for Taito and Puhti but you can install the tools in other linux and mac machines too.
- Do not overwrite and remove data without asking!
- Automatic packing and compression.
- Default bucket names based on directories of Puhti

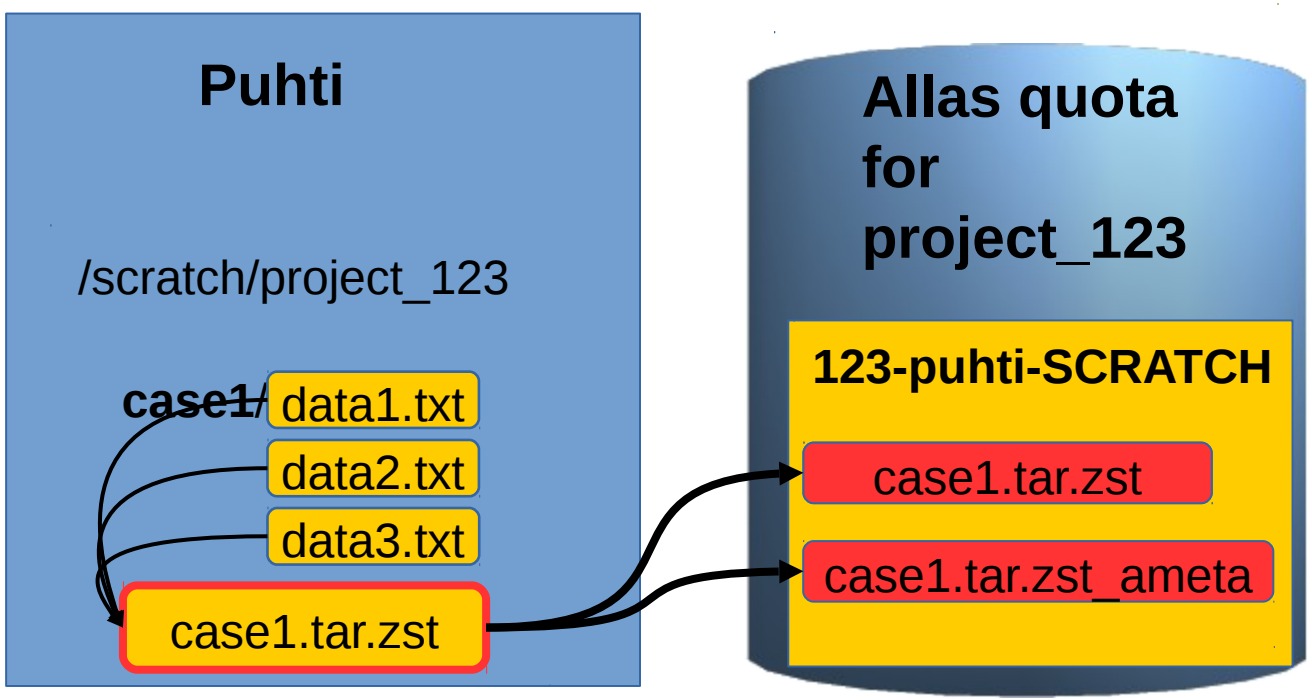
https://docs.csc.fi/#data/Allas/using_allas/a_commands/



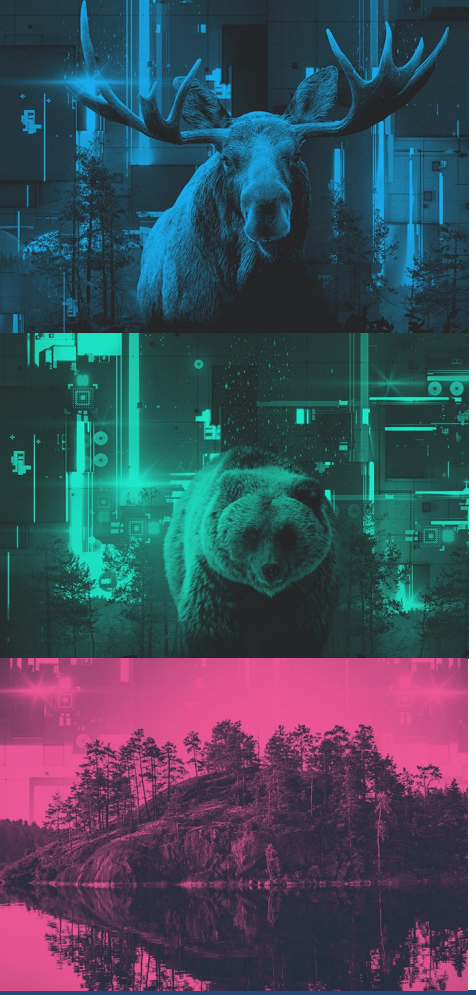



```
rclone copyto case1/ allas:123_bucket/case1/
```




Puhti
/scratch/project_123
~~case1/~~
data1.txt
data2.txt
data3.txt
case1.tar.zst
Allas quota for project_123
123-puhti-SCRATCH
case1.tar.zst
case1.tar.zst_ameta

```
a-put case1
```



Things that users should consider

- Should I store files as one object or as bigger chunks?
- Should I use compression?
- Who can use the data: Projects and accession permissions ?
- What will happen to my data later on?
- How to keep track of all the data I have in Allas?