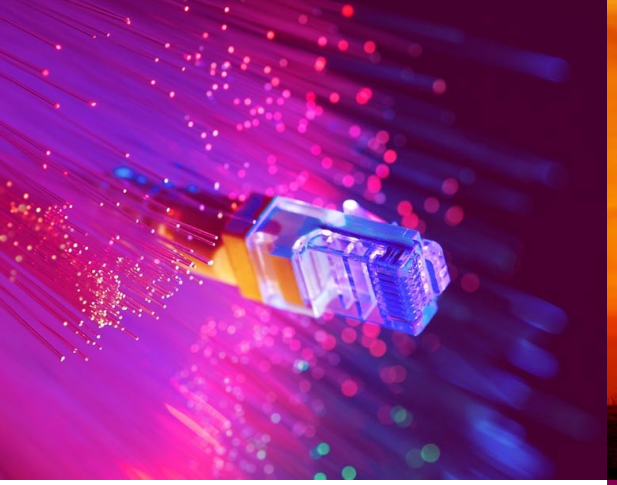


Bringing users' data to the (super)computers at CSC

CS3 11.3.2024 – Kalle Happonen

CSC – Finnish research, education, culture and public administration ICT knowledge center

Our special expertise includes for instance research infrastructures, interoperability, and digital transformation

Turnover in 2022

64 M€

One of the world's most eco-efficient datacenter in Kajaani

Non-profit state enterprise with special tasks owned by the state of Finland 70 % and Finnish higher education institutions 30 %

Approx.

567

employees in 2022

Who is CSC (- IT Center for Science Ltd.)?

- We've been around for ~50 years
- Our history is supercomputers (now LUMI) and FUNET (Finnish NREN)
- We do much more nowadays, but what's relevant for today:

We offer IT services for academic R&E in Finland. Our services include supercomputing, infrastructure clouds, container clouds, object storage, data management services, sensitive data services, etc. etc.

How users traditionally brought data to our supercomputers

```
scp -r dataset/  username@supercomputer.csc.fi:/projdir/data
```

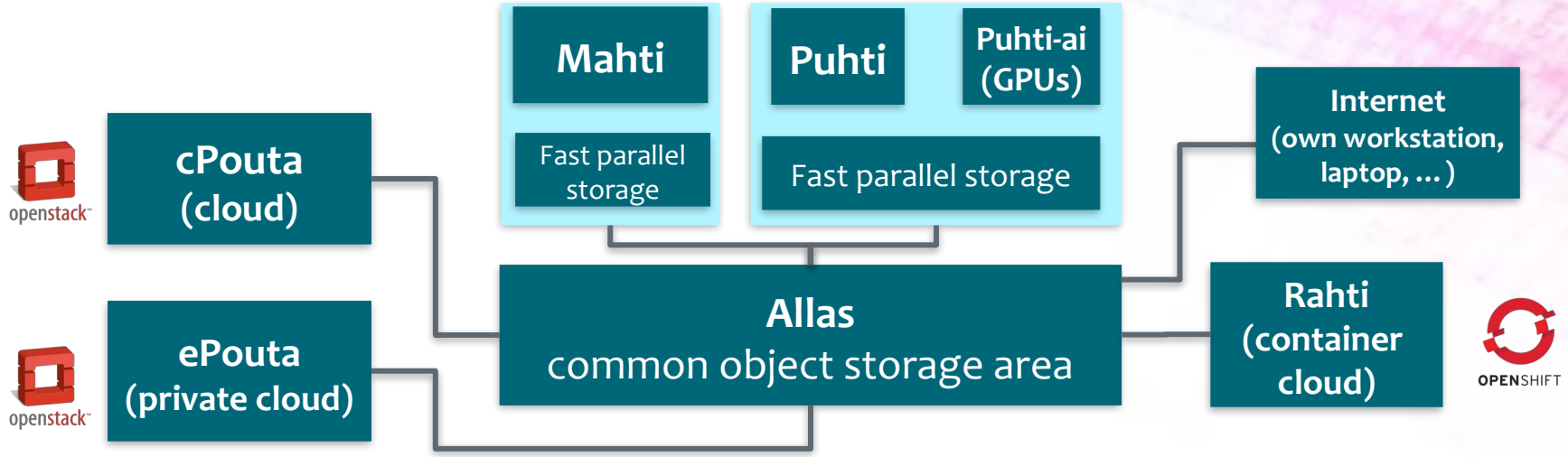
Challenges with this approach

- SCP not THAT bad – but
- It's basically a command line tool
- It relies on SSH with all that that brings
- The data is not easily accessible nor shareable
- It's hard to develop higher level services around this

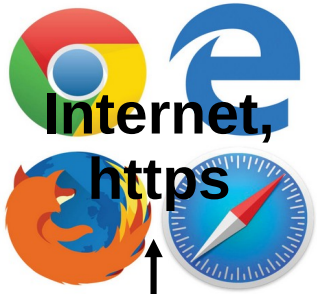
Our current architecture

- In late 2010's when we were planning new supercomputers, we thought we may need a change
- Imaginatively, we came up with a service called "Lake"
- The idea was a central data repository, accessible from all CSC services and externally interactively and programmatically
  - It needs to scale to many petabytes
  - It needs to be multi-tenant
  - Performance must scale

This is what we came up with

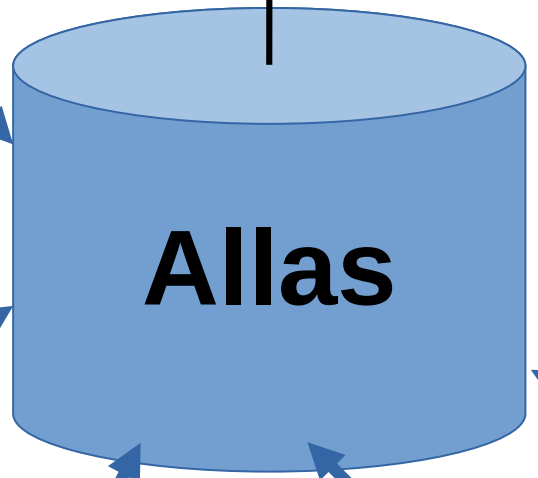


ALLAS



Swift/s3

Swift/s3

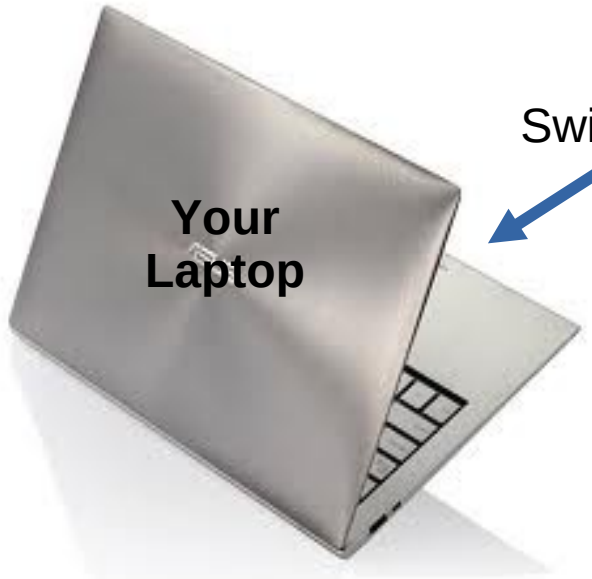


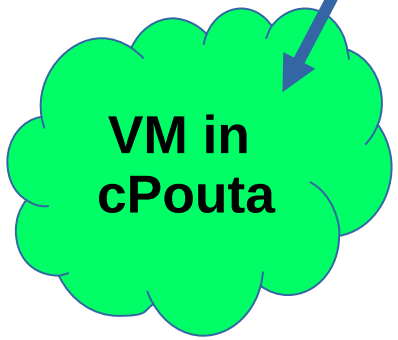
Swift/s3

Swift/s3

Swift/s3

Swift/s3





What the users see

- Their projects can request access to Allas – and be granted quota based on need
- The object storage is accessible by S3 / SWIFT and built on Ceph
- Command line tools on our supercomputers to easily access the data
- These tools can also easily be downloaded from github and used from elsewhere
- Any other S3/SWIFT tools/integrations also generally just work
- Instructions and documentation

Some statistics

- Built on Ceph
- 17 PiB of usable storage
- ~3900 different projects
- ~2 billion objects stored
- Average reads between 200 MiB/s – 1 GiB/s, peaks at 6 GiB/s
  (loosely correct, based on a graph I checked today to show you cool numbers)

What users would like to see

- We don’t have a good web interface for Allas
- The use of Allas requires an account at CSC. How do external collaborators bring in data to Allas?

Conclusions

- In general this has been a well received service
- Many additional services are already built upon Allas
- This has also brought in new users to CSC who only use the Allas service
- S3/Swift is a good start – not the final product for everyone
  - Most users do use the S3/SWIFT interfaces directly via API or low-level tools
  - We need better interfaces for the data for other users
- This importance of the service will likely grow constantly

https://www.facebook.com/CSCfi

https://twitter.com/CSCfi

https://www.youtube.com/c/CSCfi

https://www.linkedin.com/company/csc---it-center-for-science