WP6: Data Transfer, AAI, VISA deployment

J-F. Perrin

ERSF

29th Nov 2022

PaNOSC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 823852

Data Transfer

Three initially identified use cases:

An RI wants to archive its experimental data in a remote data centre. Cold backup.

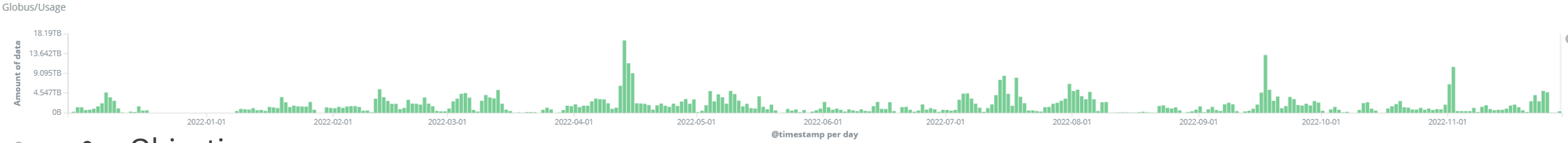
A facility user wants to transfer a large dataset from an RI's archive to a remote compute center or her/his home PC.

A user wants to access a data analysis service, data has to be available "transparently".

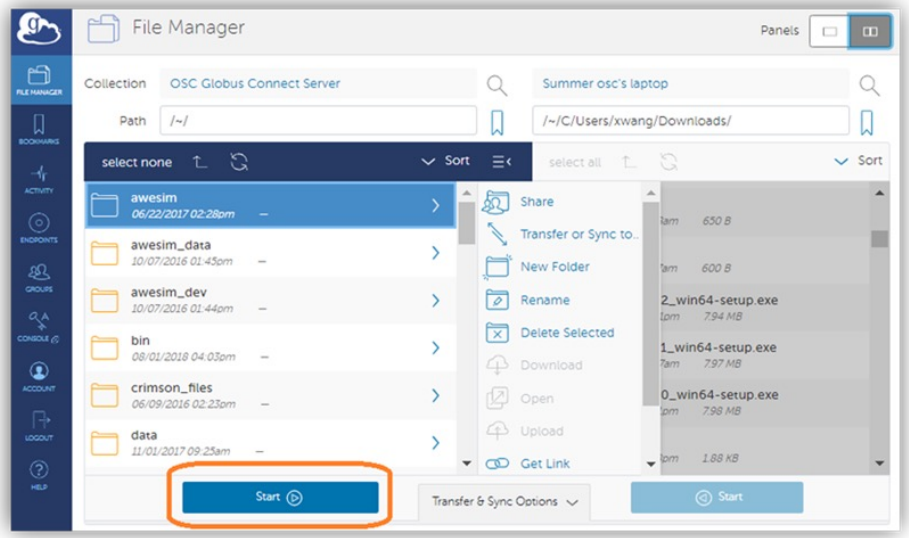
1st Use case : Archiving

- Aims
  - Automate cold backup of raw data after experiment
  - Restore data
- Pilot
  - ILL as RI (data producer) (GPFS + NFS gateway)
  - UKRI-STFC as archive centre (Ceph S3 interface)
- Solutions evaluated
  - FTS3 / GridFTP (https://fts.web.cern.ch/fts/) – Multiprotocol data transfer scheduler
  - Rucio (https://rucio.cern.ch/) – declarative data transfer
  - **Rclone** (https://rclone.org/)
- Future
  - Pilot has been completed.
  - Rclone as a simple and versatile solution that could be used in various scenarios.

2nd use case: transfer of data by RI users



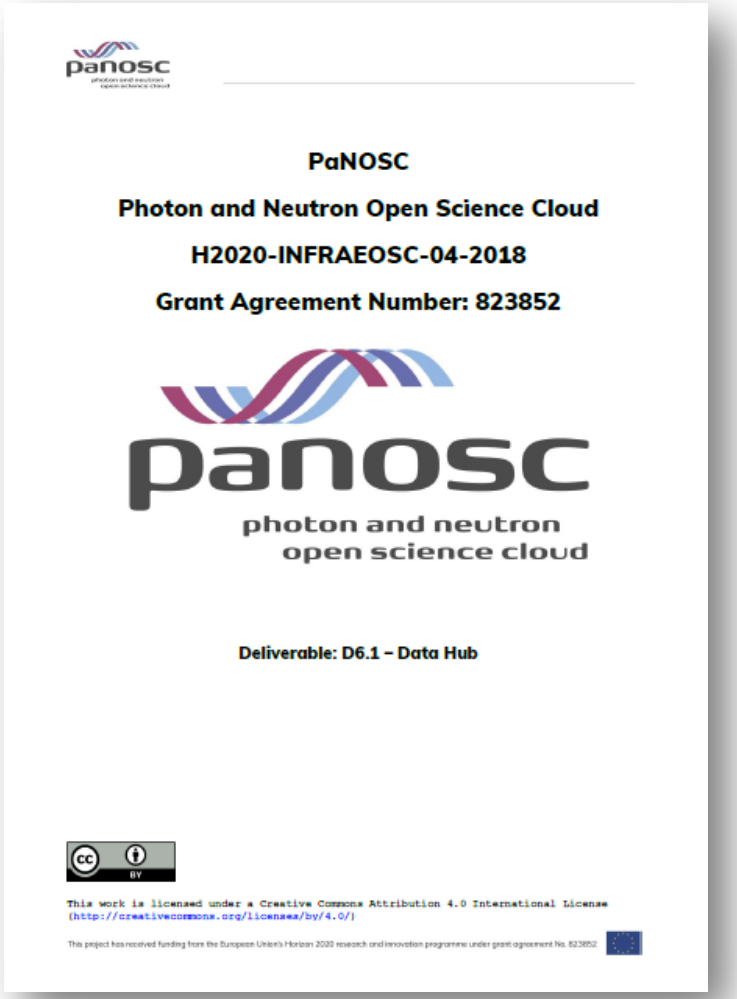
- Objective
  - Provide a (super easy, reliable, able to resume transfer, without tricky configuration on the client side …) solution for RI **users** (40 000) to transfer large datasets (10s of GB to 10s of TB) to their home lab/company
- Pilot
  - ESRF as RI (data producer) (GPFS + NFS gateway)
  - ESRF users as receivers
- Solution envisaged
  - IBM Aspera
  - **Globus**
- Current state – Future
  - Globus as become a community standard for such use case.
  - IBM Aspera also fulfils the data transfer needs, often used in industry.



photon and neutron open science cloud

3rd use case: data transfer from RIs' repositories to compute resources

a.k.a service composability



PaNOSC

Photon and Neutron Open Science Cloud

H2020-INFRAEOSC-04-2018

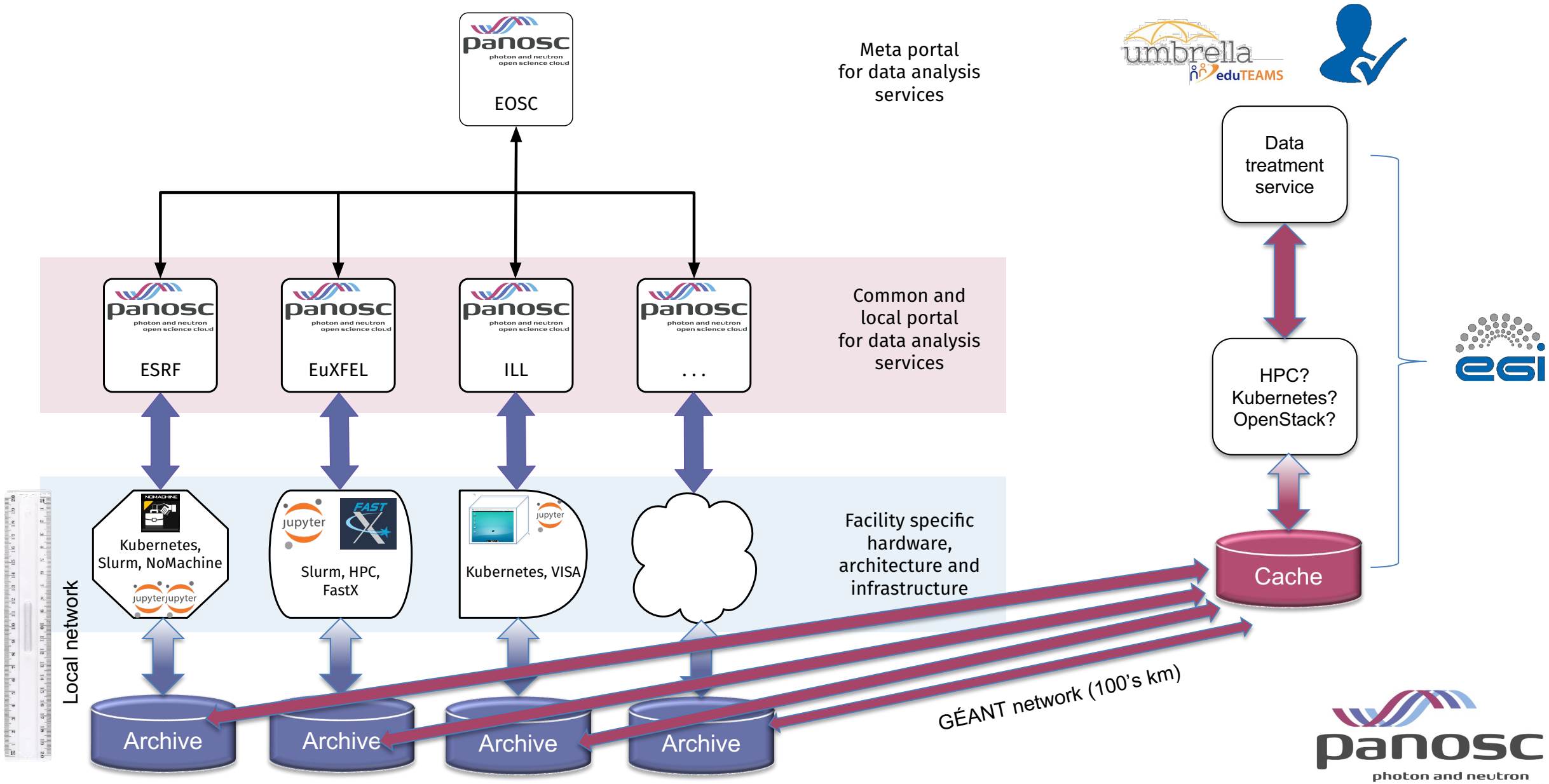
Grant Agreement Number: 823852

Deliverable: D6.1 – Data Hub

Objectives:

- Transfer data to "EOSC" (i.e. not necessarily from the PaN community) data processing services
- Transfer data "transparently" from the point of view of the users
- Archive the results back to the originating RI
- Authenticate users using EOSC ready AAI (UmbrellaID)
- Authorise data access (open or embargoed data)

3rd use case: data transfer from RIs' repositories to compute resources



3rd use case: data transfer from RIs' repositories to compute resources

Conclusion for 2022

- Different solutions have been tried (FTS, OneData, …):
  - Currently not the main use case for the communities
  - Scalability (billions of files generated per RI, maintaining a central catalogue is questionable)
  - Integration with the different RI infrastructures could be complex (Access Authorisation, support and monitoring, knowledge of the teams, …)
- In 2022, the best solution is to work with existing community tools (users have be trained, many scientific communities commonalities, …) and let users proceed with the transfer manually.

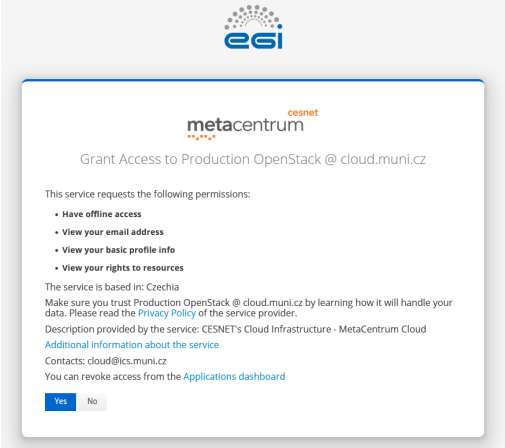
Future and perspectives

- Need for a global approach with cross community specifications (i.e. like done for the AARC BPA model)
- Machine actionable data transfer using DOI metadata (see DataCite metadata Schema 4.5 RFC)

AAI

Where do we stand in Nov 2022

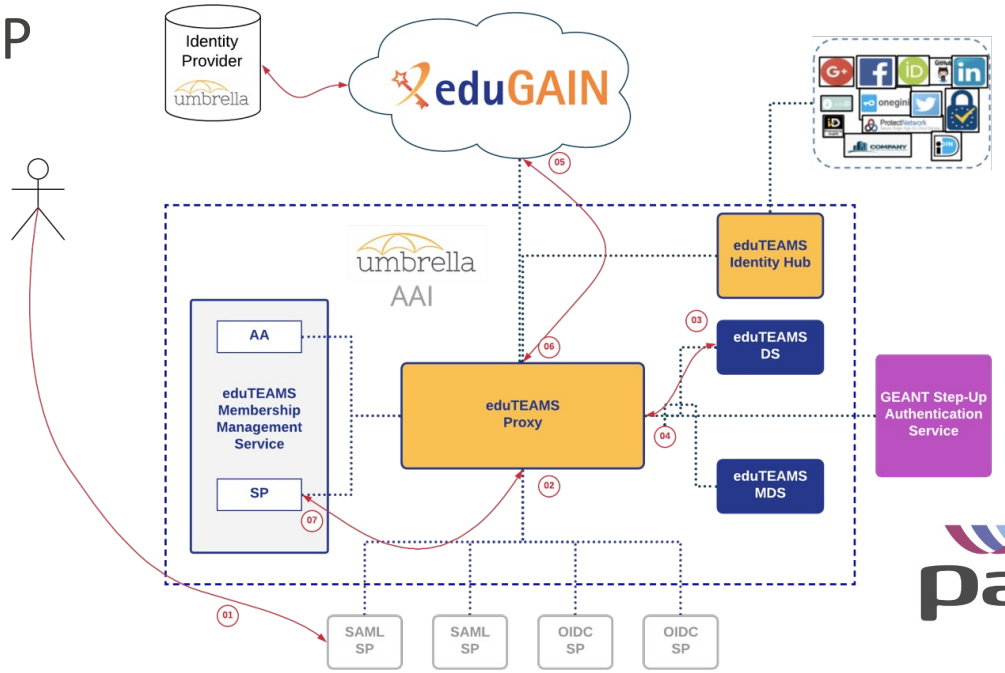
- UmbrellaID proxy is supported by eduTEAMS (AARC BPA)
  - Protocols: SAMLv2 and OpenID Connect
  - Metadata management (user attributes)
  - VO and Group management
  - Technical documentation platform (https://wiki.geant.org/display/UmbrellaID)
- UmbrellaID is now part of the EOSC AAI federation (Oct 2022) like ESCAPE AAI, EOSC life, EGI Check-in and others.
  - Technical interoperability at the proxy level
  - Policy conformance
- In practice: when a community/provider opens a service to EOSC users, your UmbrellaID account should be directly usable.



AAI

Cont.

- What still need to be done:
  - Enable registration for all services (name, email) to be compliant with REFEDS R&S
  - Introduce eduGAIN and ORCID Identity Providers (IdPs)
  - Users will have their account linked with an institutional account
  - Discuss the need to maintain our IdP
  - Refresh our public web site



panosc
photon and neutron open science cloud

VISA deployment support

Questions addressed during WP6 work

- Deployment on OpenStack?
- Extending the infrastructure to cover specific needs (HPC, GPU, …)
- For RIs' IT teams, VISA represents one additional infrastructure to set up and maintain. How WP6 could help to reduce workload impact and contribute to sustainability?
- Could we reinforce community collaboration?
- Could we have VISA hosted by non PaN RI providers?

VISA deployment support

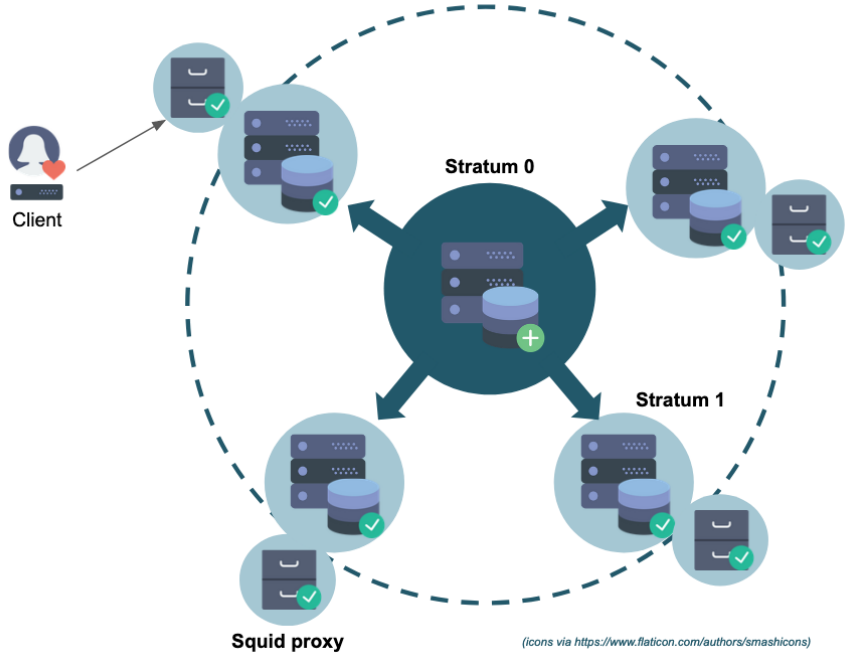
Software provisioning

- Initially all software installed on VISA VM images
  - Heavy work
  - Long list (150+) of community software (https://software.pan-data.eu)
  - End up with very large images (unpractical)
  - Libraries dependencies headache

- Benefited from interaction with ESCAPE (in particular GSI)
  - Use of lightweight containers to package software
  - CVMFS for distribution of software

panosc
photon and neutron
open science cloud

VISA deployment support



Software provisioning

- Software packaged in lightweight containers (Singularity/Apptain
  - Independency from the operating system
  - No need to locally “install” software
  - Allow to preserve and offer release history
- CVMFS for the distribution of software
  - Same repository could be used by different compute infrastructure (VISA, HPC, workstations …)
  - HTTP protocol
  - Software repo could be exposed to the internet

Future and perspective

- VISA could be deployed on 3rd party infrastructure (VISA@EGI).
- With an appropriate trust framework we could imagine consuming SW prepared by other RIs or communities out of the box.

panosc
photon and neutron open science cloud

Thank you

wp6@panosc.eu

PaNOSC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 823852