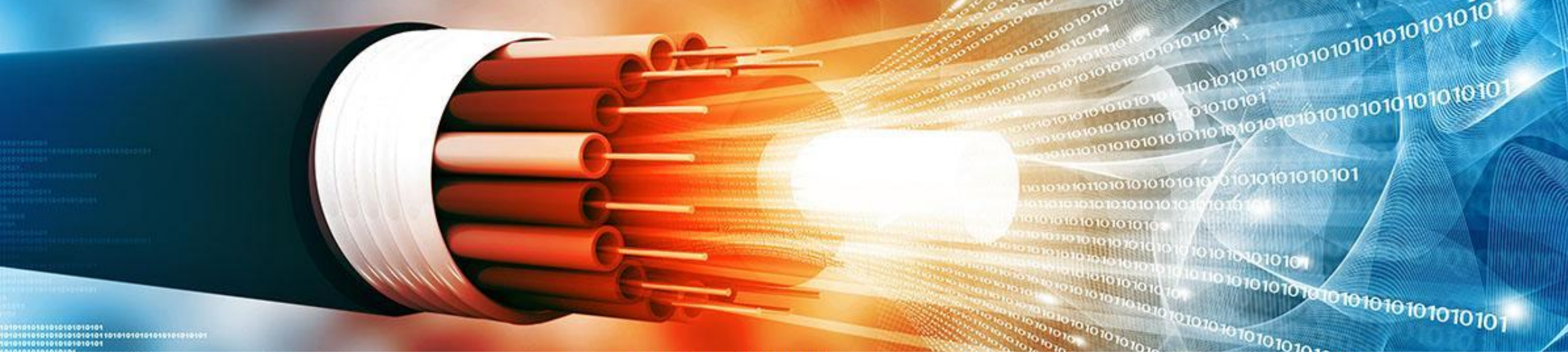


Cross-border analysis on Climate, Biodiversity and Nordic Language Processing Laboratory data (NLPL)

Abdulrahman Azab, UiO

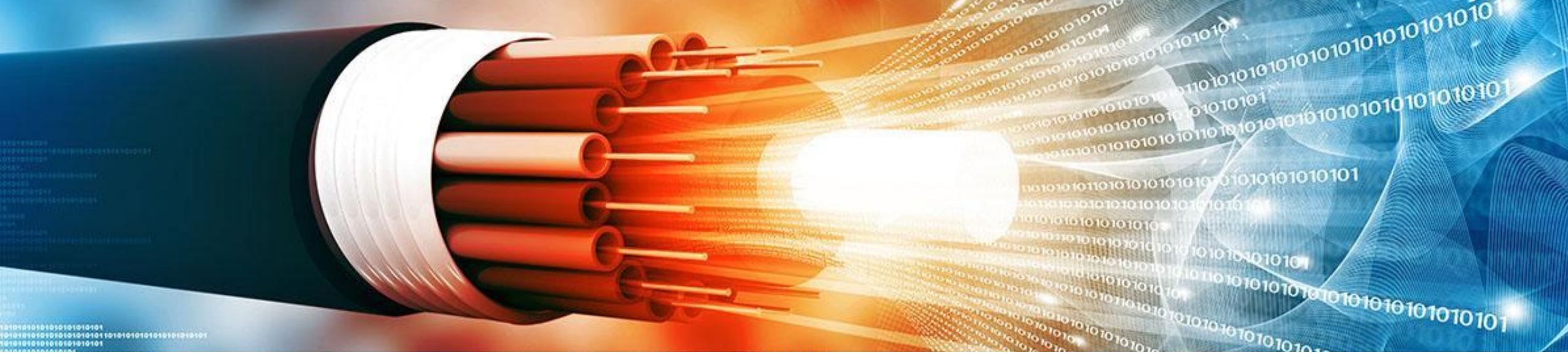
EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

T5.2 - Analysis and Post-processing across

- **T5.2.1:** Cross-border data processing workflows
- **T5.2.2:** Code Repositories, Containerization and "virtual laboratories"
- **T5.2.3:** Platform as a Service for Scientific Cloud Computing and Cloud Native Execution mechanisms

The aim of T5.2.1 is to facilitate data pre- and post-processing workflows on distributed data and computing resources across borders by

1) Enabling community portals to schedule jobs on multiple remote resources (Nordic HPC lusters) instead of one local cluster
2) Developing generic modules and solutions to: package analysis services and tools according to the best practices, and allow running and deploying these services on different infrastructures independent of the software available on target platforms



EOSC NORDIC

Machine Learning for Climate Modelling

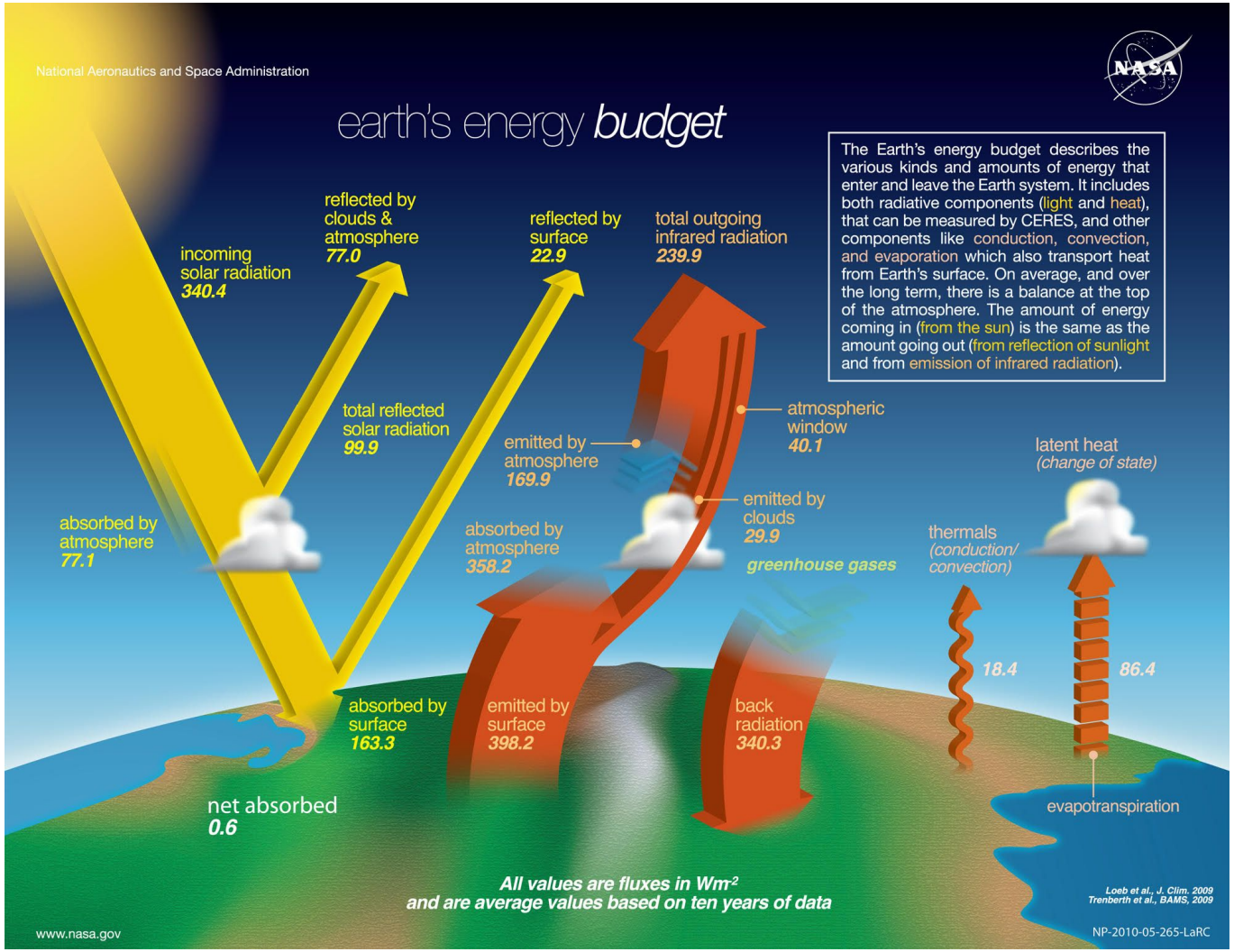
Ernir Erlingsson (UIce),
Jean Iaquinta (UiO),
Helmut Neukirchen (UIce)

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

The Earth's Energy Budget

Earth's energy budget: accounts for the balance between the energy that Earth receives from the Sun, and the energy the Earth radiates back into outer space.

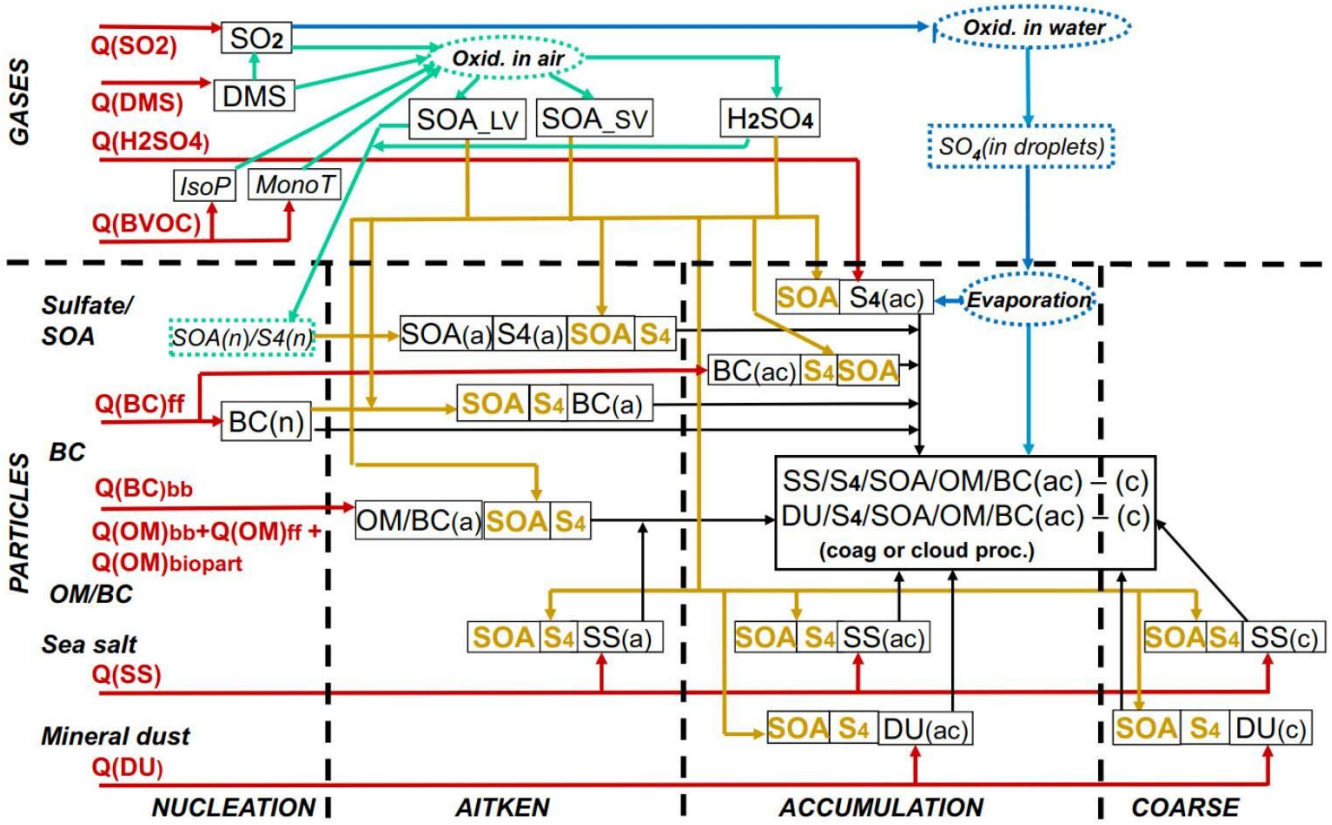
Here, we focus on modelling the radiative transfer of aerosol, using deep learning.



Aerosol in Climate Studies

The reflectivity properties of different aerosol compositions in the atmosphere is an important part of calculating the Earth's energy budget.

The number of aerosol combinations, however, are vast and therefore complex.



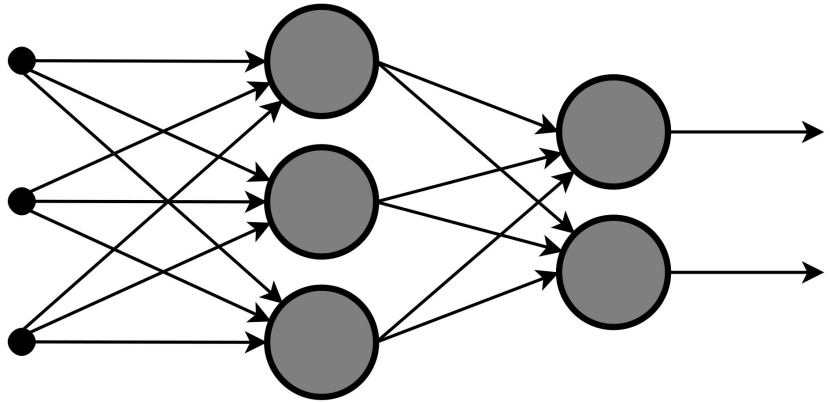
⇒ Extremely time-consuming to compute a good approximation.
Traditionally: using specialised software which is difficult to scale and port.

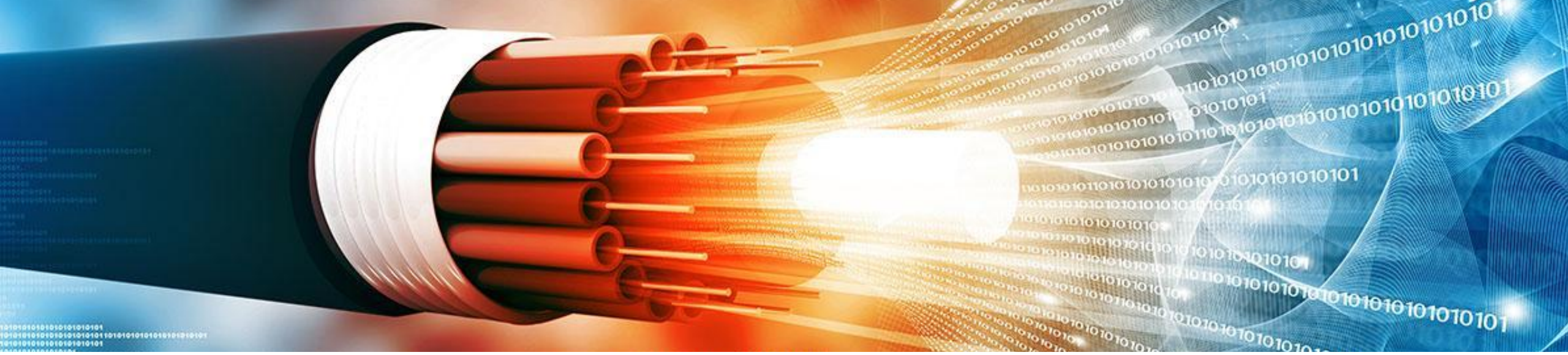
Predictions using Deep Learning

With deep learning (using Tensorflow and Keras), we explore the potential of replacing the specialized software with neural networks.

Goals:

- Increase its performance and portability,
- Examine possible non-linear correlations between different types of aerosol.





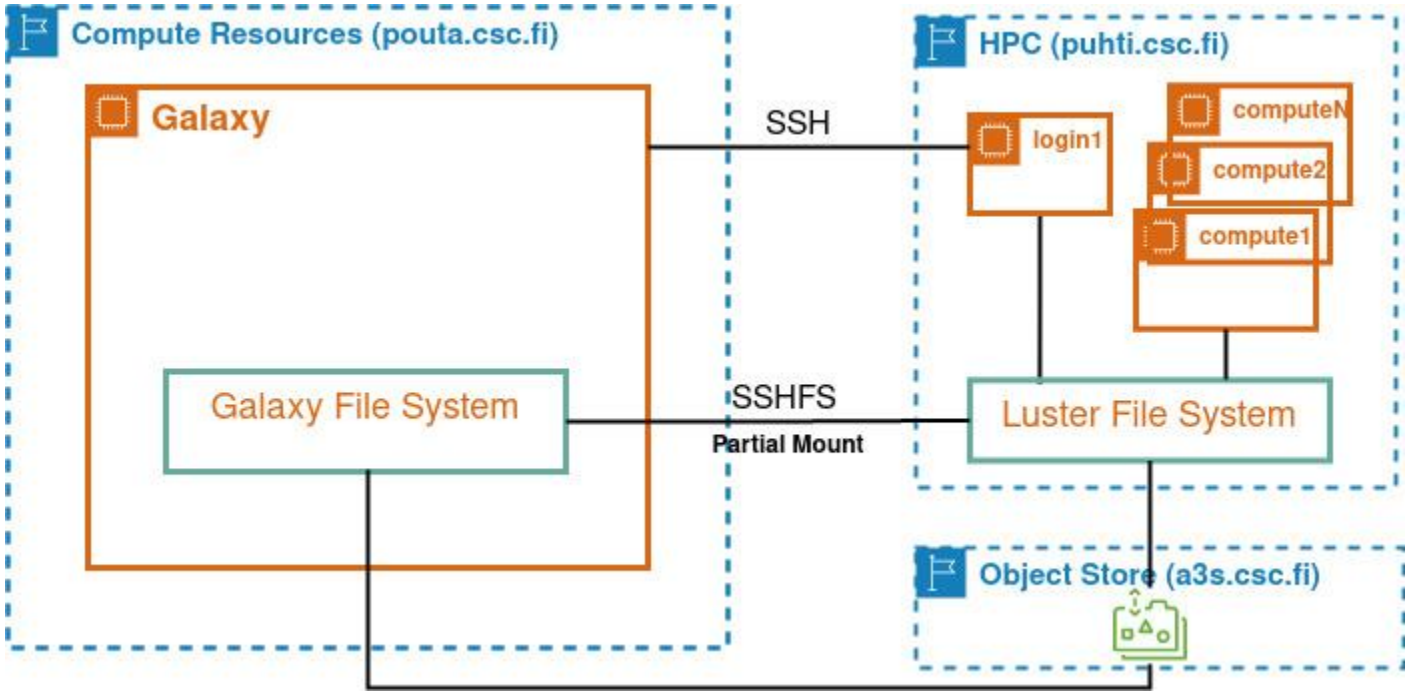
EOSC NORDIC

Cross-borders computing through portals

Tewodros Deneke, CSC-IT
Anne Fouilloux, UiO
Kessy Abarenkov, UT

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

Controlling Climate bigdata analysis with Galaxy



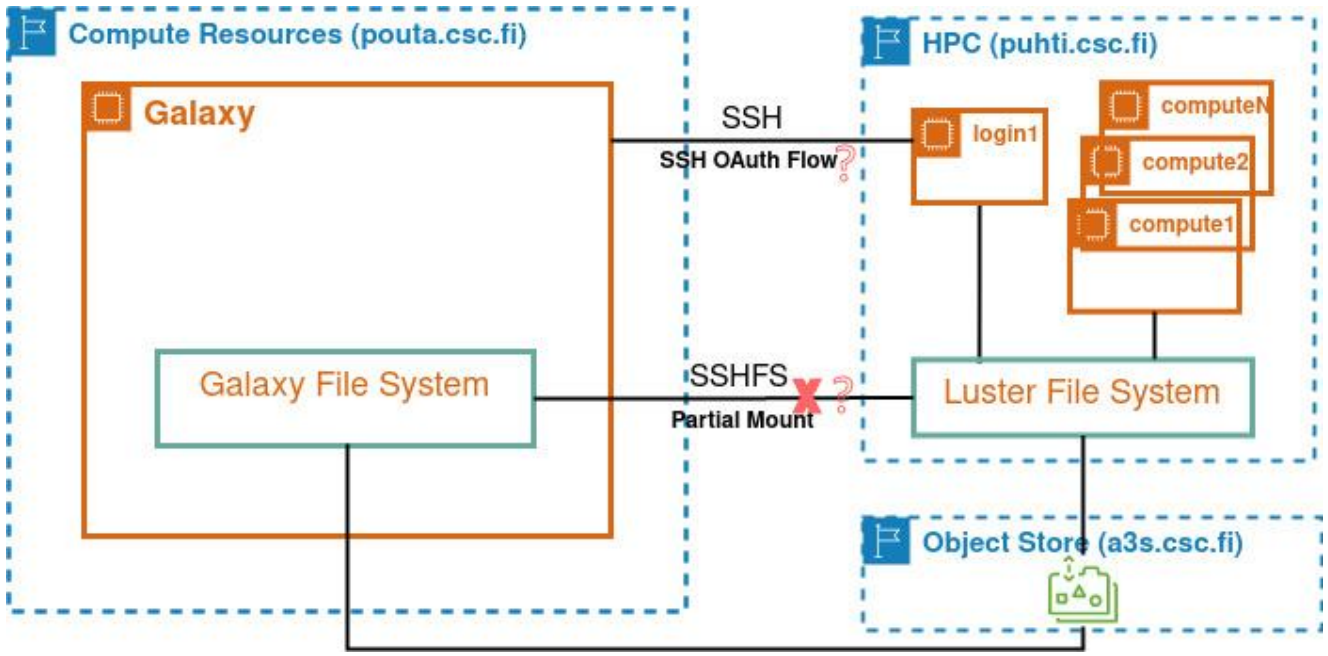
- Galaxy running on a cloud VM
- Jobs submitted to HPC via a robot account
- SSH based CLI job runner
- Slurm job plugin

- SSHFS mount Galaxy's job script and job history directories

- Object storage as galaxy secondary storage and remote data source

- Sample tools that fetch data directly from storage

Controlling Climate bigdata analysis with Galaxy



- A way to avoid transfer of huge data
  - Use DOIs or other data identifiers (that respect FAIR principles) as references
  - Use external storages for big data and bypass Galaxy local storage when sensable
  - On galaxy, store references only
- A way to allow users to set and use own HPC cluster credentials and under user preferences
  - Could help removes the need for file system mounts and shared file assumptions
  - SSH OAuth flow and token based SSH access
- Rehink cross-border accounting
  - Standardize interface? Puhuri?

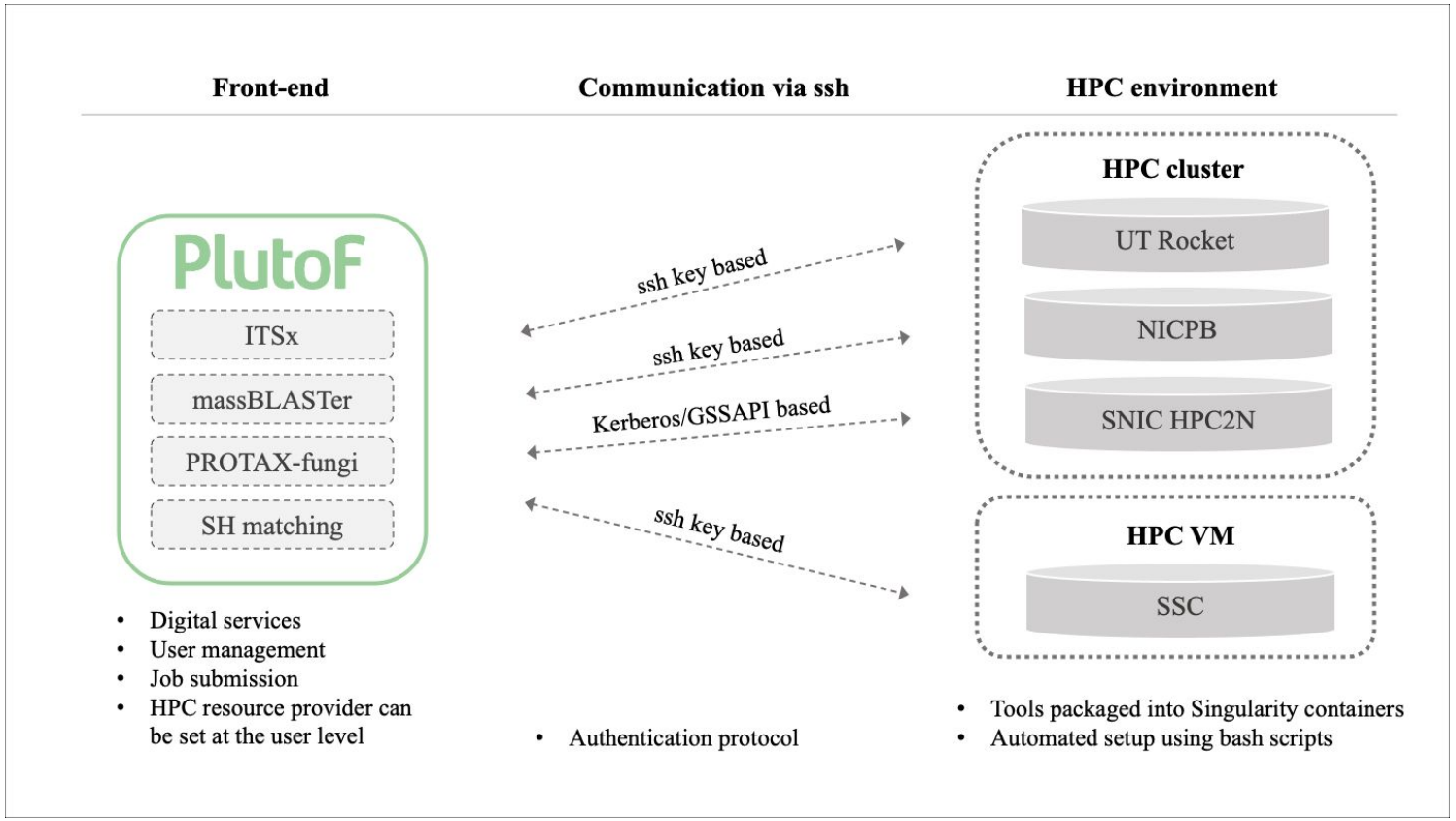
[Demo]

Biodiversity Pilot I

Biodiversity Pilot – digital services to support **global species discovery from environmental DNA (eDNA) are provided by the PlutoF platform** (https://plutof.ut.ee).

PlutoF is an online workbench and computing service provider for biology and related disciplines. In EOSC-Nordic project, PlutoF acted as a a test bed to provide data management and analysis services for the marine eDNA collected by the European Artificial Reef Monitoring (ARMS) program.

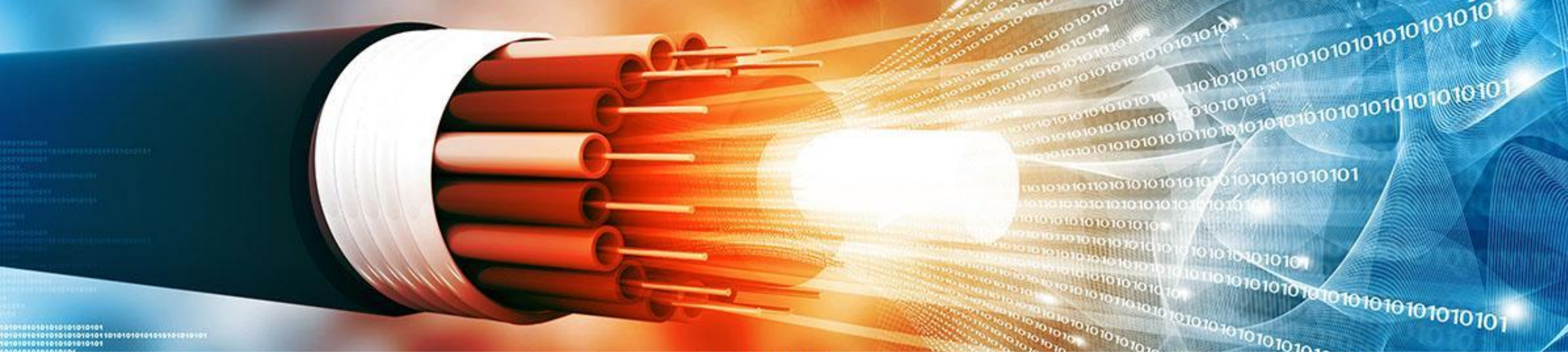
Biodiversity Pilot II



1) Services were packaged into Singularity containers to easily build, transfer and run the services independent of the software available in remote HPC resources

2) Support for sending analysis jobs to different HPC clusters and HPC VMs was added

3) Worked out recommended procedures on how users can apply for HPC resources and how to set up access to EOSC HPC resources from PlutoF



NLPL Virtual laboratories

Andrey Kutuzov, UiO
Sabry Razick, UiO
Stephan Oepen, UiO
Abdulrahman Azab, UiO

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

Introduction

The **Nordic Language Processing Laboratory (NLPL)** is a collaboration of university research groups in Natural Language Processing (NLP) in Northern Europe with a vision to **implement a virtual laboratory for large-scale NLP research**. In this project a number of **software packages** and **software pipelines** are used. As the participants come from different institutes and use disparate infrastructure, the possibility to **install these software in a uniform way** has been a requirement.

Target: provisioning of software for NLP research in a manner that makes it possible and cost-efficient to **maintain the exact same software stack on multiple systems**. Here, systems initially mean different High Performance Computer (HPC) systems: **6 HPC systems located in Norway, Sweden, and Finland**

The case

Software with their dependencies are compiled from source code on the system it will be running. This is in contrast to using pre-compiled binary installations (including standard Python wheels and conda packages). this process is cumbersome as:

1. need to **locate the sources** code of the software
2. discover **all dependencies** so the software can be compiled and run on any system, regardless of its current configuration.
3. need to **test the installation**
4. need to place the software in a consistent manner need to construct a method for **different versions of the software to live side-by-side**

Easybuild

EasyBuild is a software build and installation framework that allows you to manage (scientific) software on HPC systems in an efficient way.

The **NLPL virtual lab is technically a set of so called easyconfigs**: description files (with the *.eb extension) which EasyBuild uses to actually build and deploy the corresponding software as loadable modules. **Modules are dependent on each other and accompanied by a set of convenience scripts and instructions**.

Our experience shows its suitability for the deployment of **reproducible software environments for complex NLP tasks across different HPC systems**, including multi-GPU and multi-node setups. It has also been successfully used in teaching deep learning for NLP in 2021 by the Language Technology Group at the University of Oslo

Easybuild

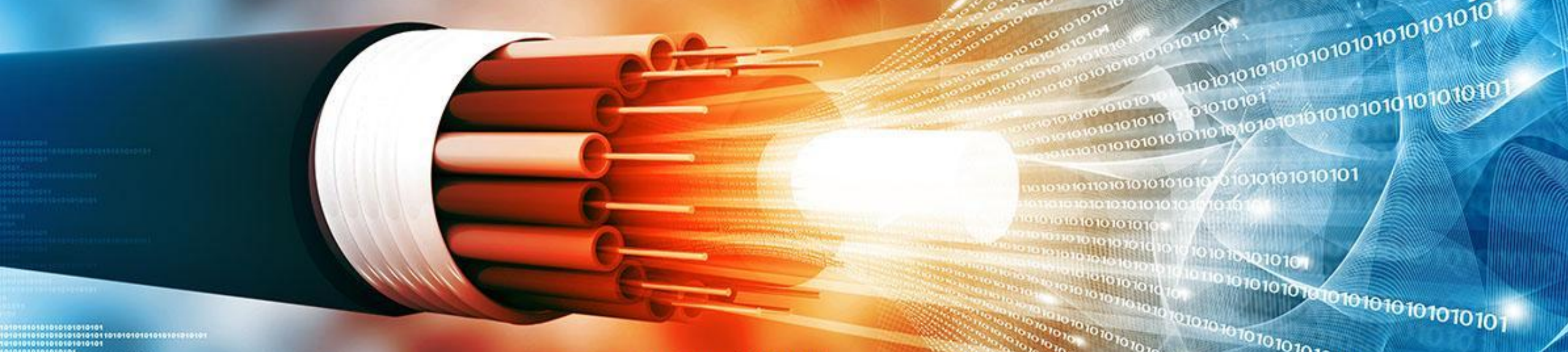
EasyBuild is a software build and installation framework that allows you to manage (scientific) software on HPC systems in an efficient way.

Community builds the easy build scripts
Slurm job: builds 40+ modules, lasts 24 hrs with resolving dependencies
Effort: 5 PMs: search for solution + build scripts + automation

Why not containers?

1. Concern for reduced transparency from the user point of view.
2. Containerizing individual software modules severely challenges modularization: There is no straightforward way to **'mix and match'** multiple containers into **a uniform process environment**.

Option to go: provisioning the *full NLPL* software in a container

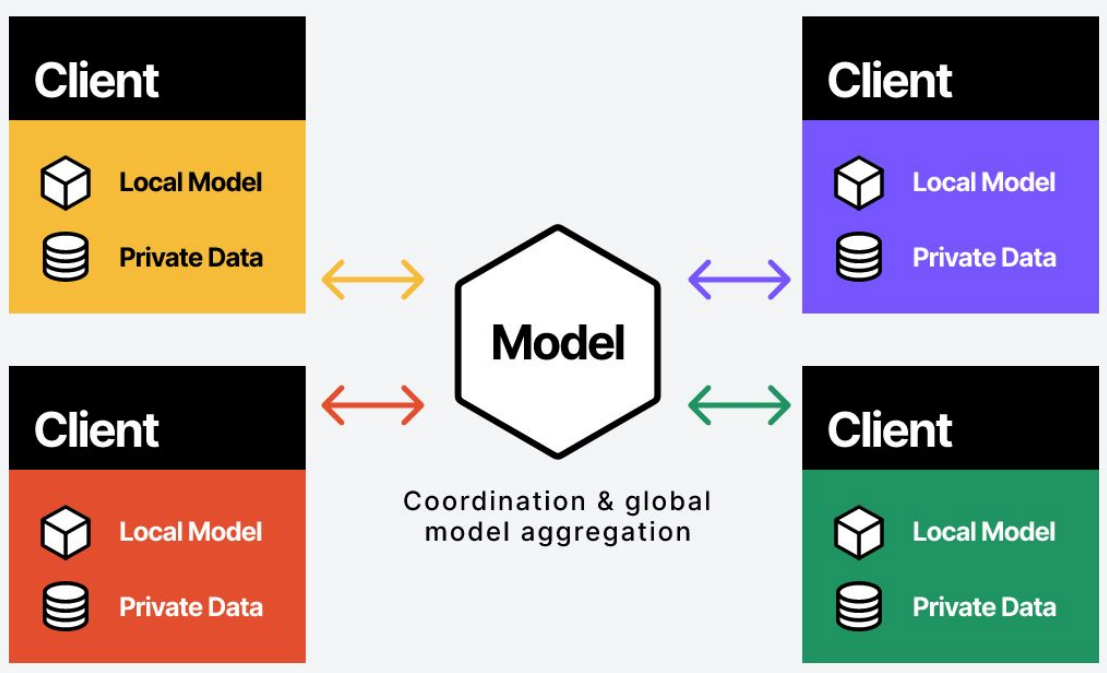


EOSC Nordic T5.2.3: Cross border computing on distributed cloud resources

Lorand Szentannai, UNINETT Sigma2

EOSC-Nordic project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857652

EOSC Nordic T5.2.3
Cross border computing on distributed cloud resources

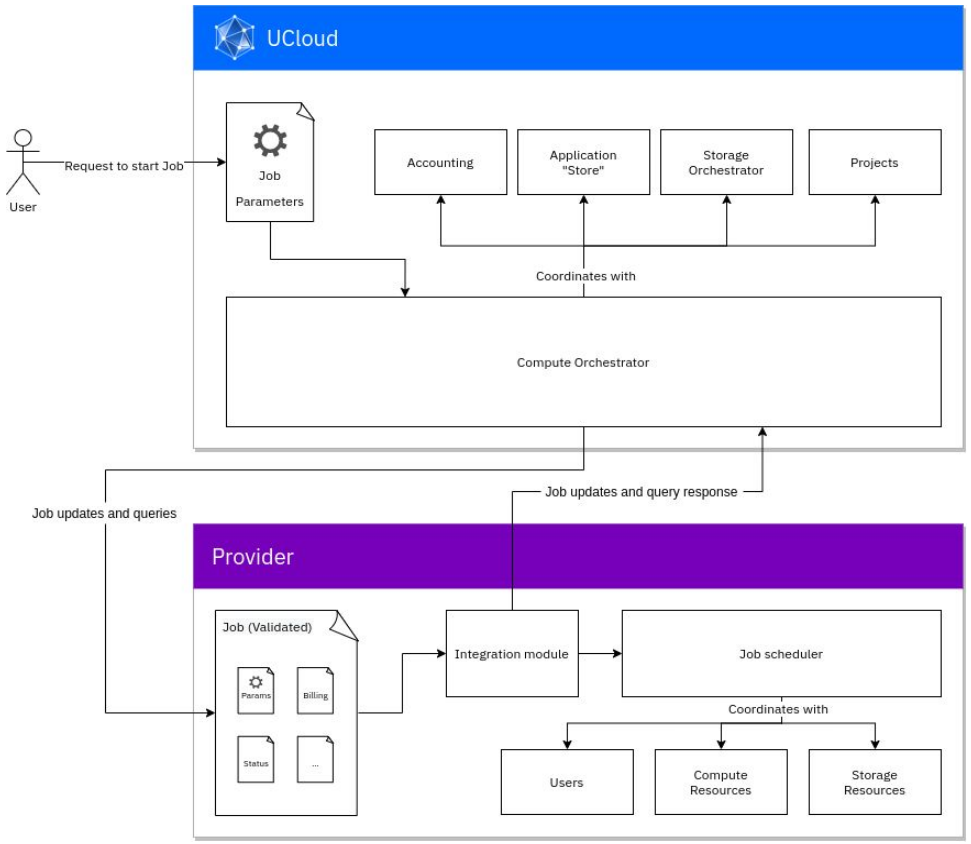


Biodiversity use case

- in dialog with the SNIC and UGOT / KSO / SGU
- subsea image analysis using distributed cloud resources
- technical challenges related to data security, limited licensing, and transfer of big data
- **federated machine learning** facilitating alignment with licensing and cross border policies
- focuses on **moving computation to data** as an alternative
- the task showcases well the potential of cross border cloud computing, and highlights the challenges due to policy limitations
- held workshop to understand the needs in the community
- demonstrated capability of Scaleout's FEDn
- aims to set up a FEDn test instance with FEDn clients at service providers in Sweden, Finland (at CSC) and Norway (on the NIRD infrastructure)

https://www.scaleoutsystems.com/federated-machine-learning

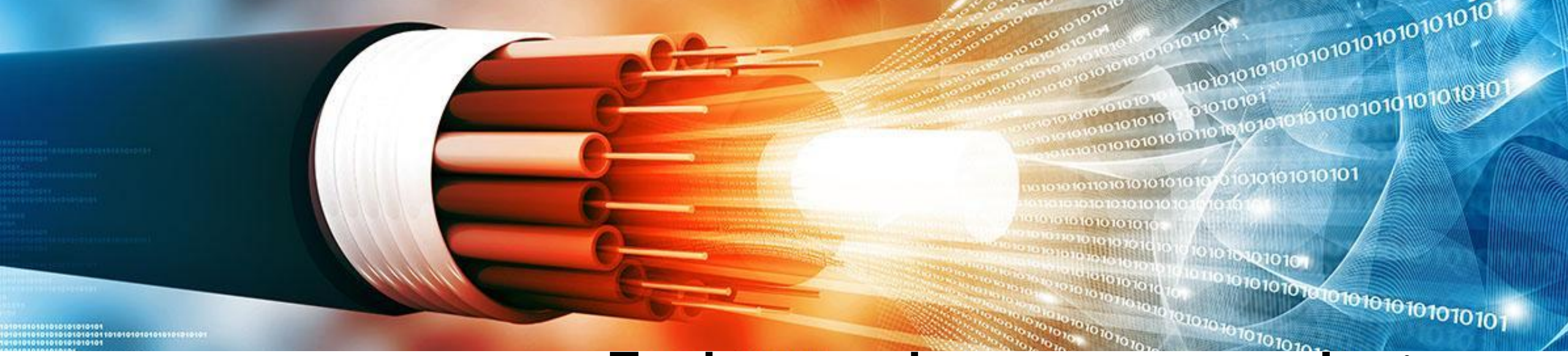
EOSC Nordic T5.2.3
Cross border computing on distributed cloud resources



https://docs.cloud.sdu.dk/dev/backend/app-orchestrator-service/README.html

Digital humanities and natural language processing use case

- SDU / UCloud in dialog with NEIC NDHL - Nordic Digital Humanities Laboratory
- aiming to demonstrate **orchestration of workloads** over existing cloud solutions for research in Nordic countries
- use UCloud to **deploying jobs on cloud resources geographically located in different countries**
- the goal is to deploy the UCloud orchestrator's Integration Module, i.e. the provider, in test environments on cPouta at CSC, and on NIRD at Sigma2

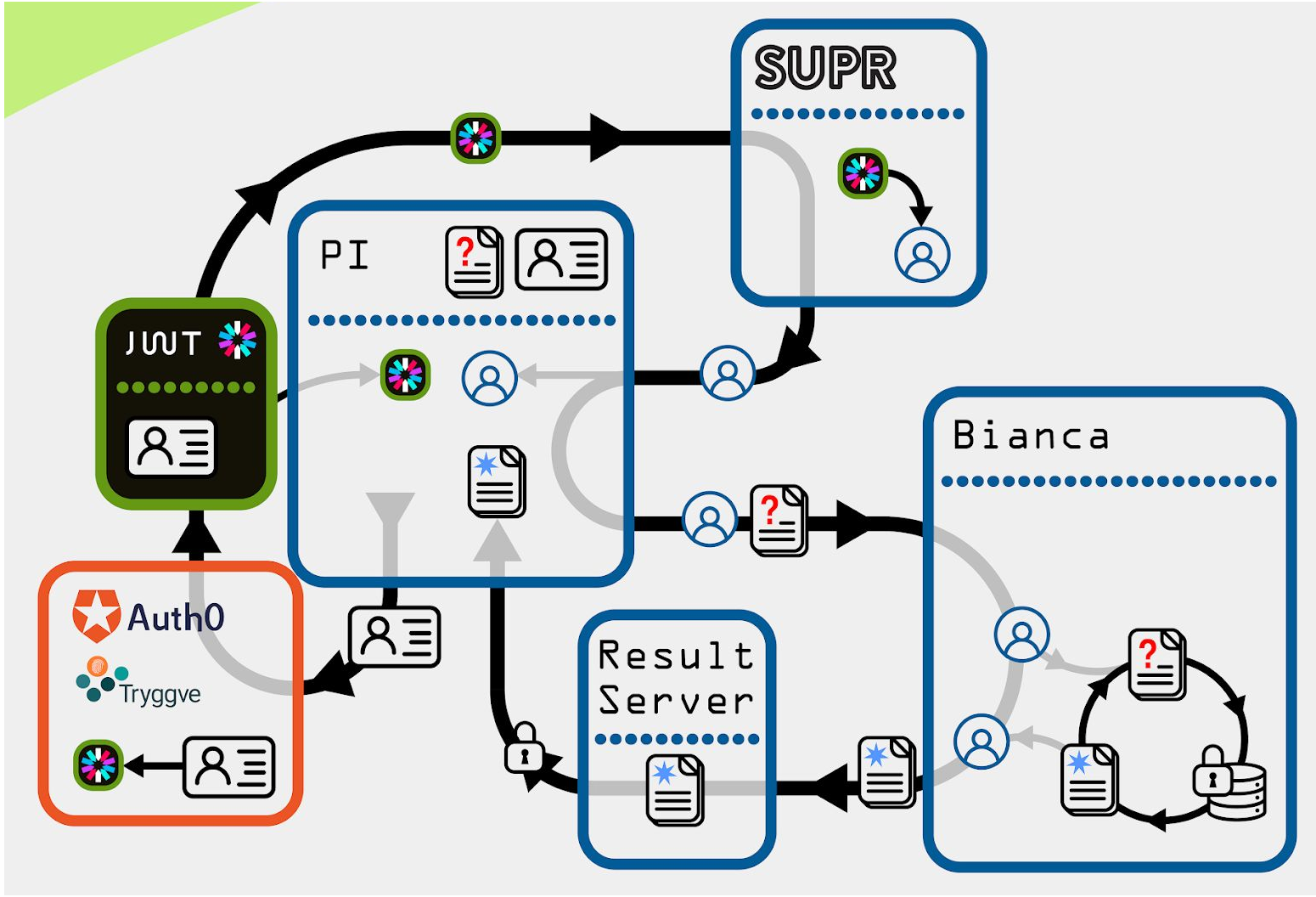


Federated storage, analysis platform and sharing options for sensitive data

EOSC-Nordic WP5.4 “Sensitive data”
Puhuri project & ELIXIR AAI
LUMI User Support Team (LUST) & LUMI Sensitive data task-force

EOSC-Nordic project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 857652

Federated token based access – in progress

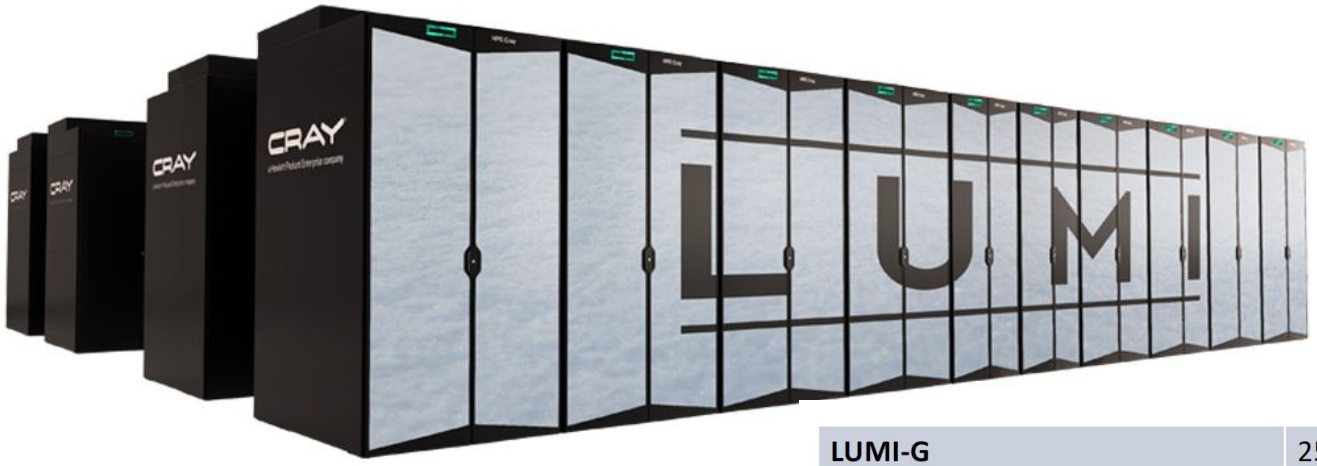


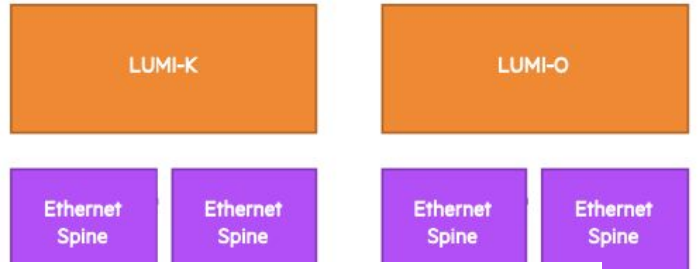
Unresolved issues

- Level of assurance
- Identity verification
- Secure data transfer

LUMI for sensitive data

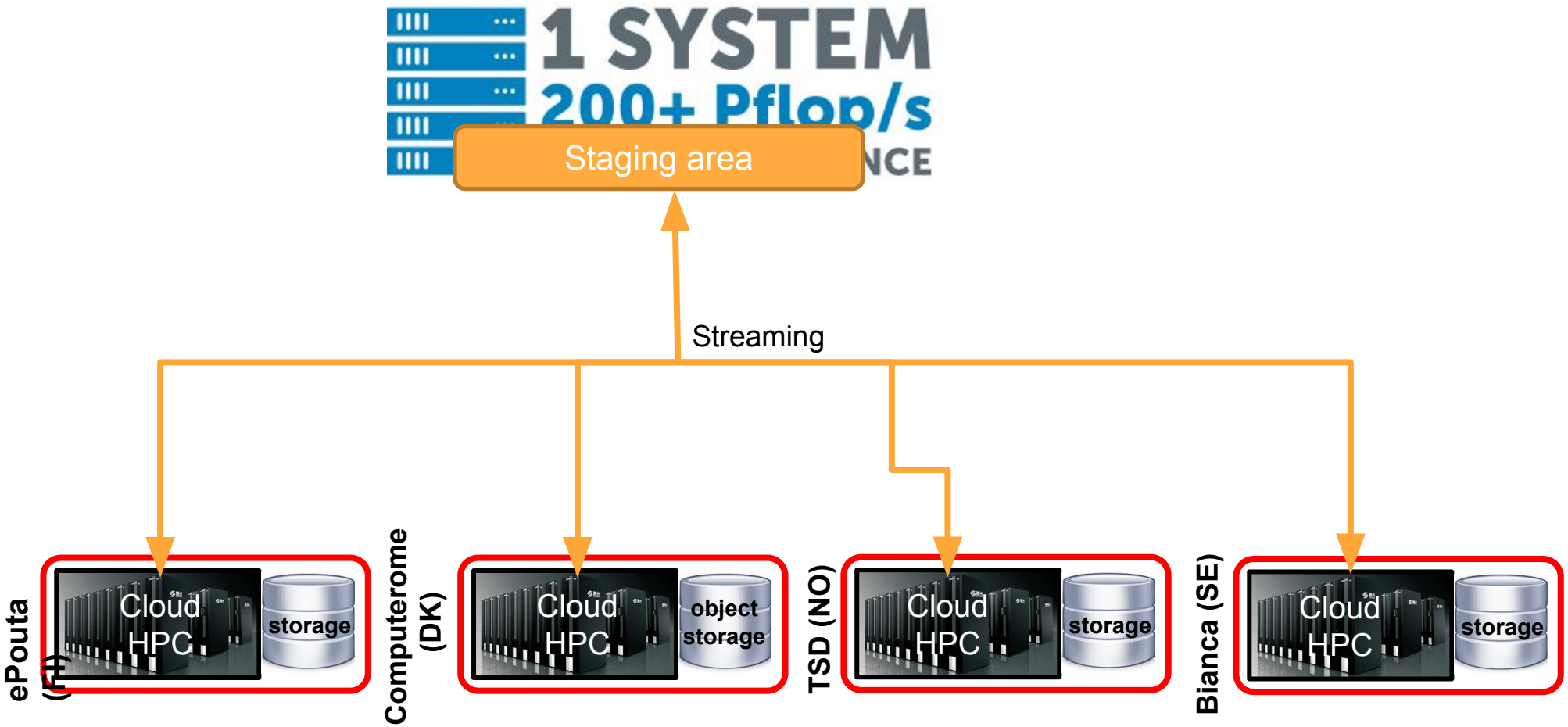
- 5 use cases



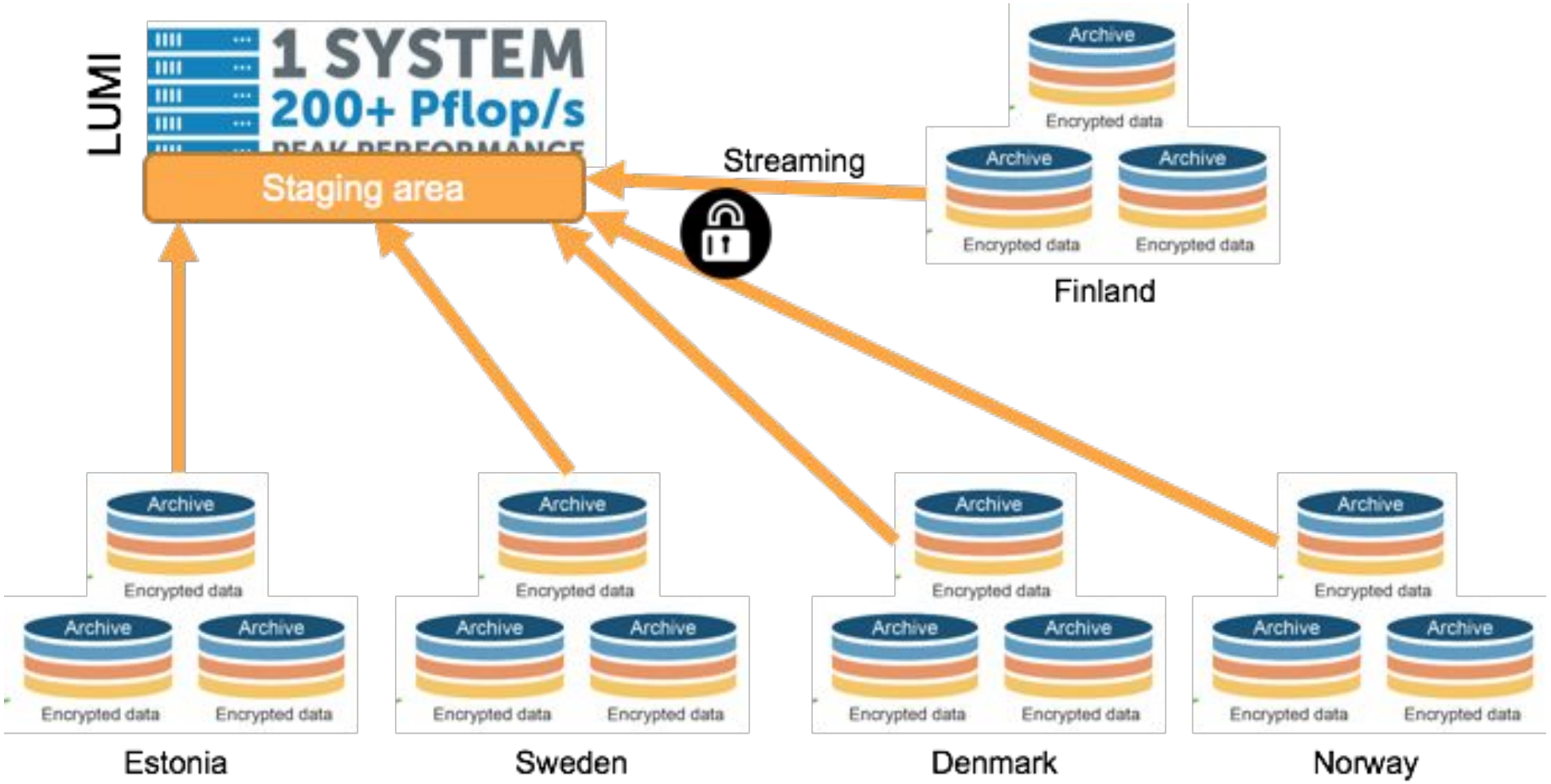


LUMI-G	2560 nodes with 4xAMD MI200 (128 GB HBM2e) + 1 AMD Trento, 4x200 Gbit/s Slingshot
LUMI-C	1536 nodes with 2 x AMD Milan, 200 Gbit Slingshot, of these 128 L and 32 XL nodes
LUMI-D	32 TB shared memory, 64 x Nvidia Quadro
Peak Performance (Pflop/s)	552 (LUMI-G), 8 (LUMI-C)
HPL Performance (Pflop/s)	375 (LUMI-G), 5 (LUMI-C)
Storage (PB)	7 (LUMI-F), 80 (LUMI-P), 30 (LUMI-O)
Storage bandwidth (GB/s)	1760 (LUMI-F), 960 (LUMI-P)
Total Power	8.1 MW
Targeted Acceptance / GA	July 2021 (everything besides LUMI-G), December 2021 (LUMI-G)

LUMI for sensitive data (proposed)

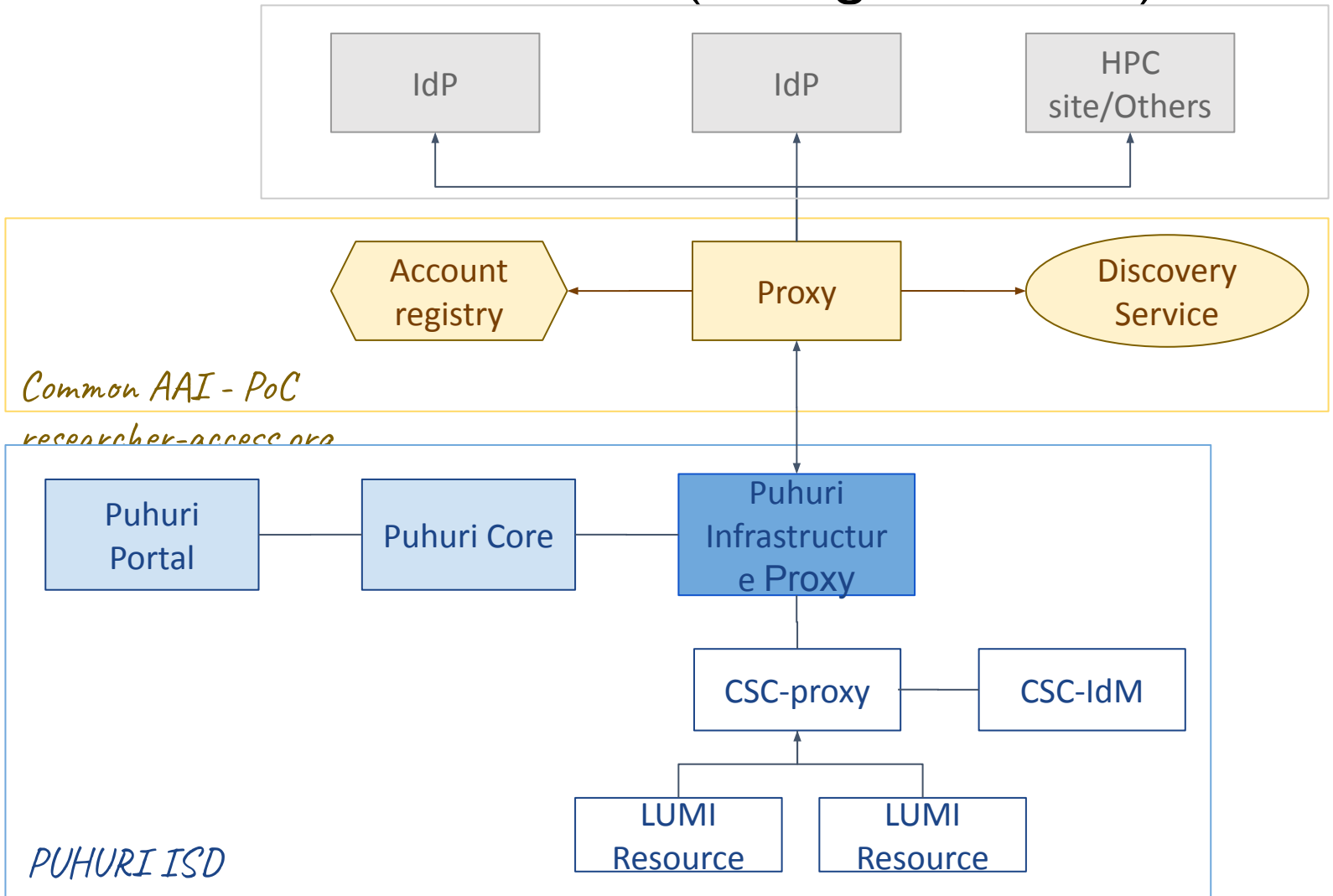


LUMI for sensitive data (proposed)



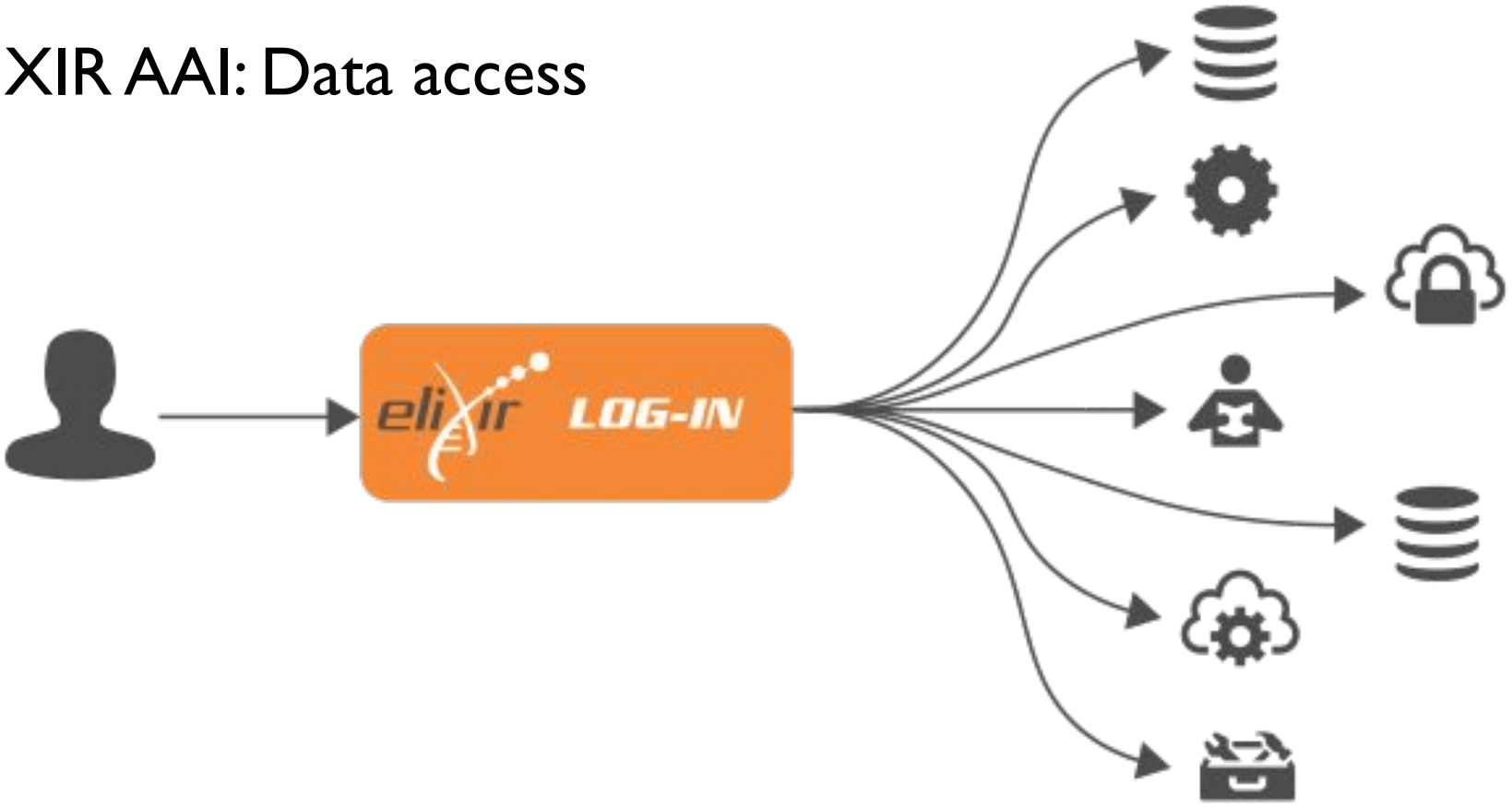
AAI (under discussion)

- Puhuri AAI: Resource allocation (Storage and HPC)



AAI (under discussion)

- ELIXIR AAI: Data access



Outstanding issues

- Coordination with the local service providers
- Data processor agreement for LUMI
- Secure data upload/download

Q&A

Andulrahman Azab

azab@uio.no

EOSC NORDIC

www.eosc-nordic.eu

https://twitter.com/EOSC_Nordic

https://www.linkedin.com/groups/13756550/

https://www.linkedin.com/groups/13756550/