COMPBIOMED & DICE Collaboration Update

ALL HANDS MEETING – 23 JUNE 2022

Narges Zarrabi (SURF)

Alastair Smith (UCL)

DICE
Data Infrastructure Capacity for EOSC

CompBioMed

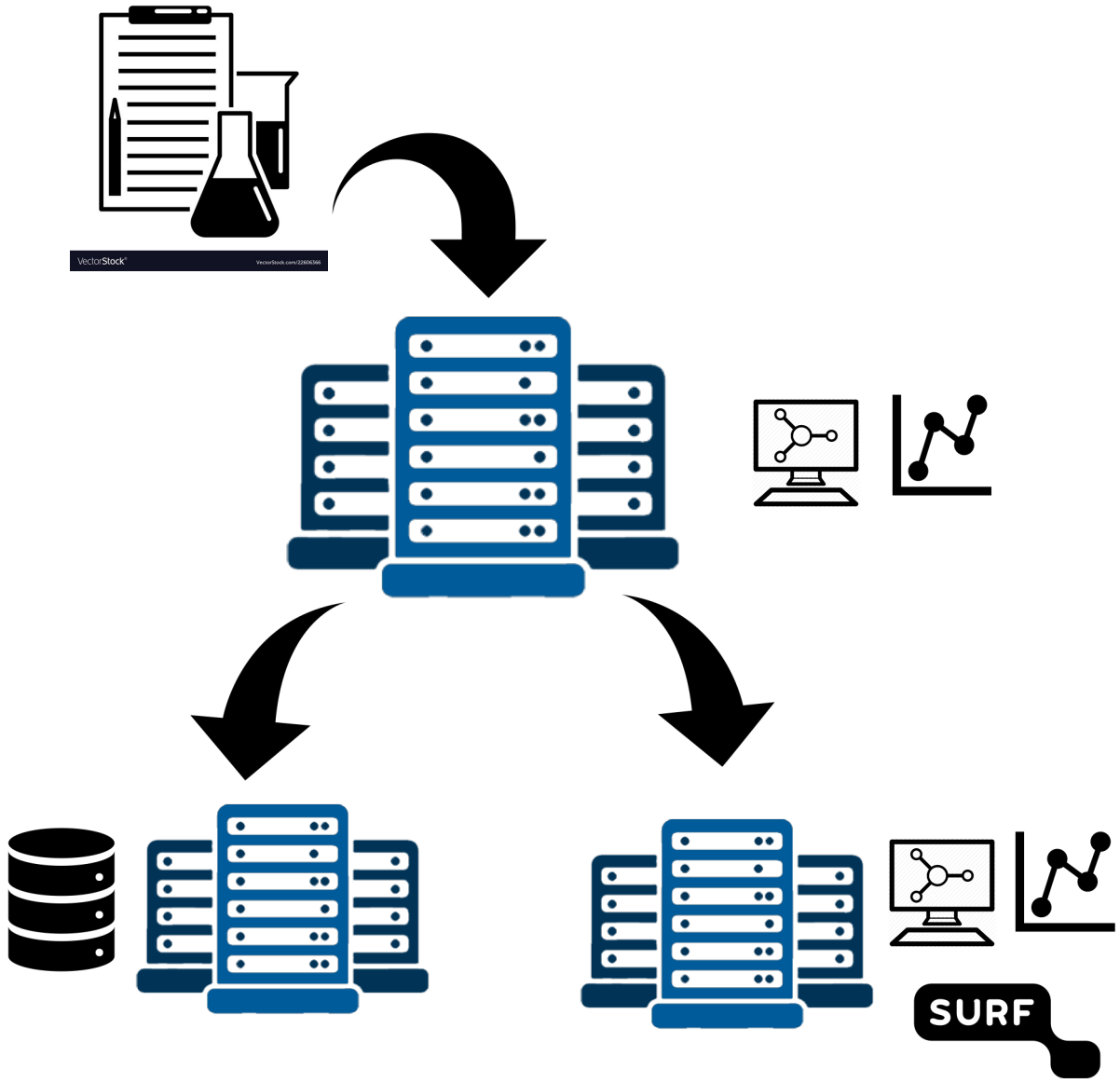

DICE
Data Infrastructure Capacity for EOSC
CompBioMed
Data management and publication –
A DICE & CompBioMed Hackathon
21 June 2022, 12:00 to 18:00
CINECA – Bologna, Italy

Data Management Challenges of Research Communninties

- **More efficient data access, sharing and transfer**
  - *Intensive data-sharing and transfer*
  - *Restricted data-sharing and transfer*
- **Preserving research data**
  - *Storage, backup and archiving large data, synchronizing data over distributed places*
  - *data provenance*
- **Accessible research Data**
  - *Making data accessible to research communities, PIDs*
  - *Publishing data with domain specific metadata*
  - *Linking published data to processed and raw data*
- **Findable research data**
  - *A major challenges scientific communities is to discover data from research data collections and repositories*

Workflow using Alya Application

- **Step 1: Data creation and transfer:** The raw data is collected at a lab (ESRF in France). The data is being stored locally on tapes. Currently, a copy of the data is transferred to BSC.
- **Step 2: Data pre-processing:** In BSC, researchers pre-process the data which includes manual and automated steps for image stitching, segmentation and meshing.
- **Step 3: Data replication:** The preprocessed data needs to be replicated from BSC to other HPC centers such as SURF. The replicated data will then be used to run simulations on the supercomputers in these sites.
- **Step 4: Data Processing and analysis:** running simulations and analyze output data



CompBioMed and DICE collaboration

- **Workflows to be implemented:**
  - **Data replication workflow: facilitate large data transfer by making replicas, data preservation, bring data close to compute**
  - **Data publication workflow: A data repository for publishing (large) data and/or metadata, metadata schema for CompBioMed**

- **CompBioMed partners involved: UCL, BSC, SURF**

- **EUDAT and DICE services to be used:**
  - **B2SHARE** – Searchable Data Repository
  - **B2HANDLE** – Persistent Identifier Provider
  - **B2SAFE** – Distributed, Secure Policy Based Data Storage

EUDADT services used in CompBioMed

Service	*Description*	*Resources Needed*	*Provider*
B2SHARE	Data Repository for data publication. Metadata schema can be implemented in this repository. Integration with B2FIND for harvesting data and facilitating findability of the data.	50 TB	UCL
B2HANDLE	Tool required to make persistent identifiers (PIDs) for the data to facilitate findability of the data. The PIDs will potentially be used in B2SAFE and B2SHARE.	1 prefix 10000 PIDs	SURF
B2SAFE	Data staging and safe replication of research data between HPC centers in CompBioMed. The archival storage on tape facilitates long-term preservation of the data.	50 TB 50 TB	SURF BSC

Data & HPC Federation in CompBioMed

BSC

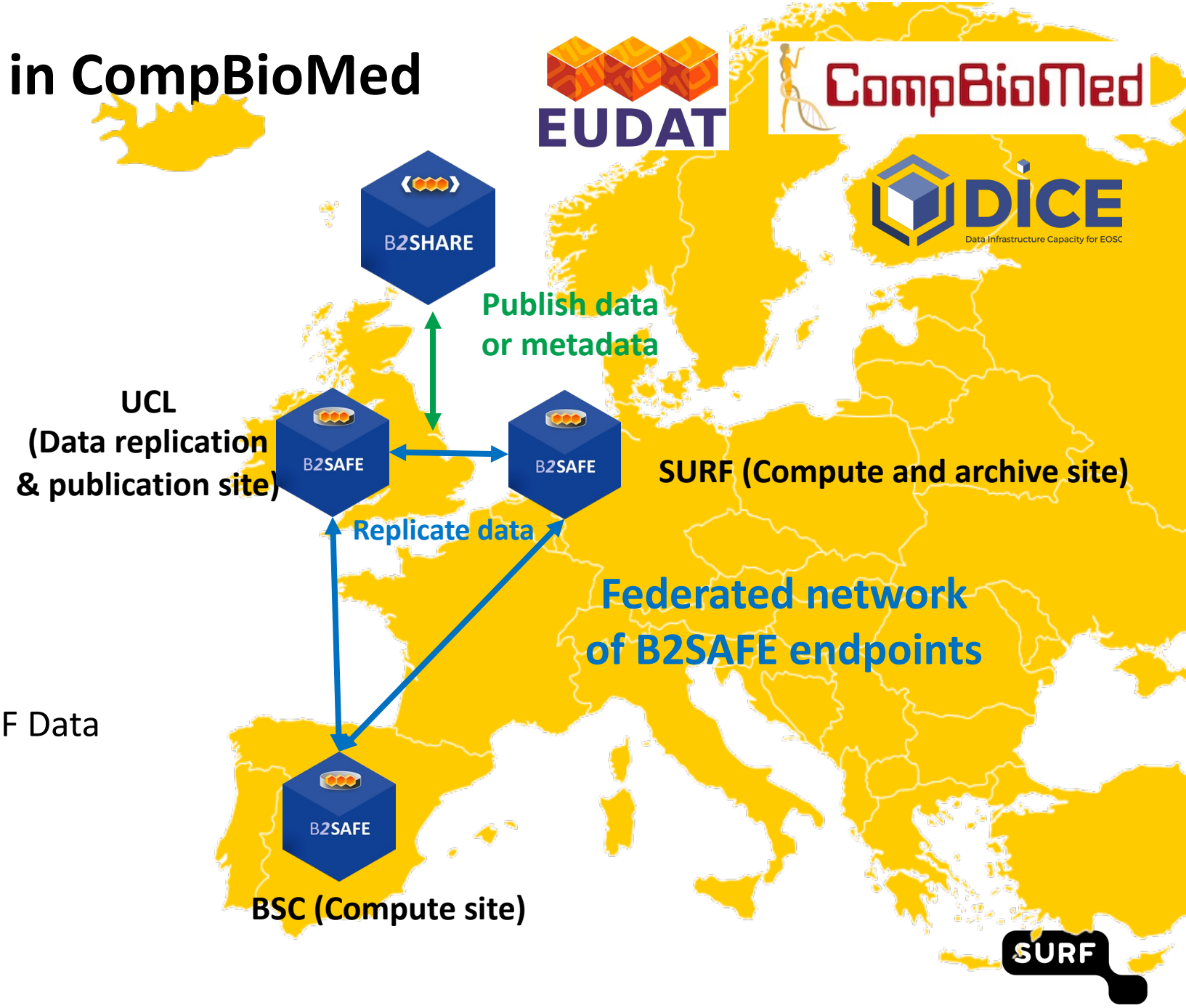
- Compute site
- B2SAFE endpoint

SURF

- Compute site
- B2SAFE endpoint
- Data Archiving site
- Possibility to publish data in SURF Data repository

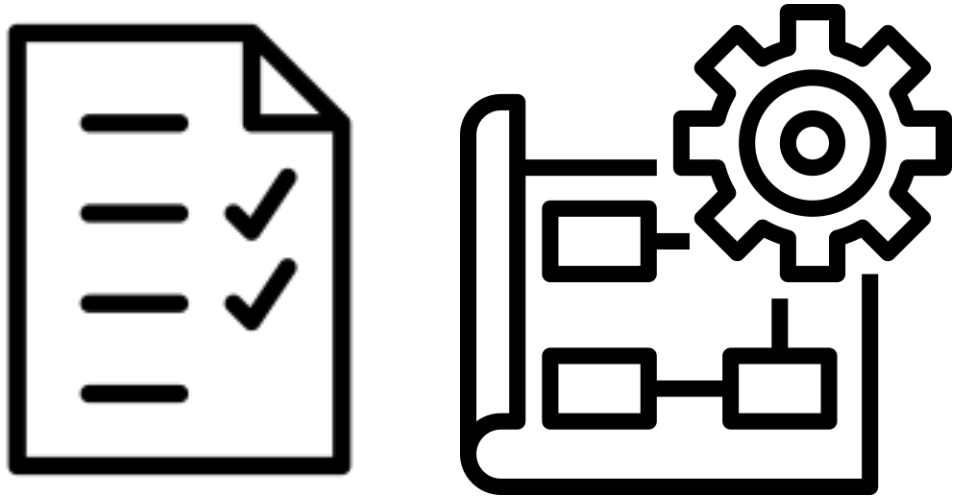
UCL

- Data publication site
- B2SAFE endpoint



Workplan and technical task descriptios

- We have made a technical workplan
- In process of deploying and configuration of services
- Technical support to deploy and using these services is provided through the CompBioMed and DICE collaboration

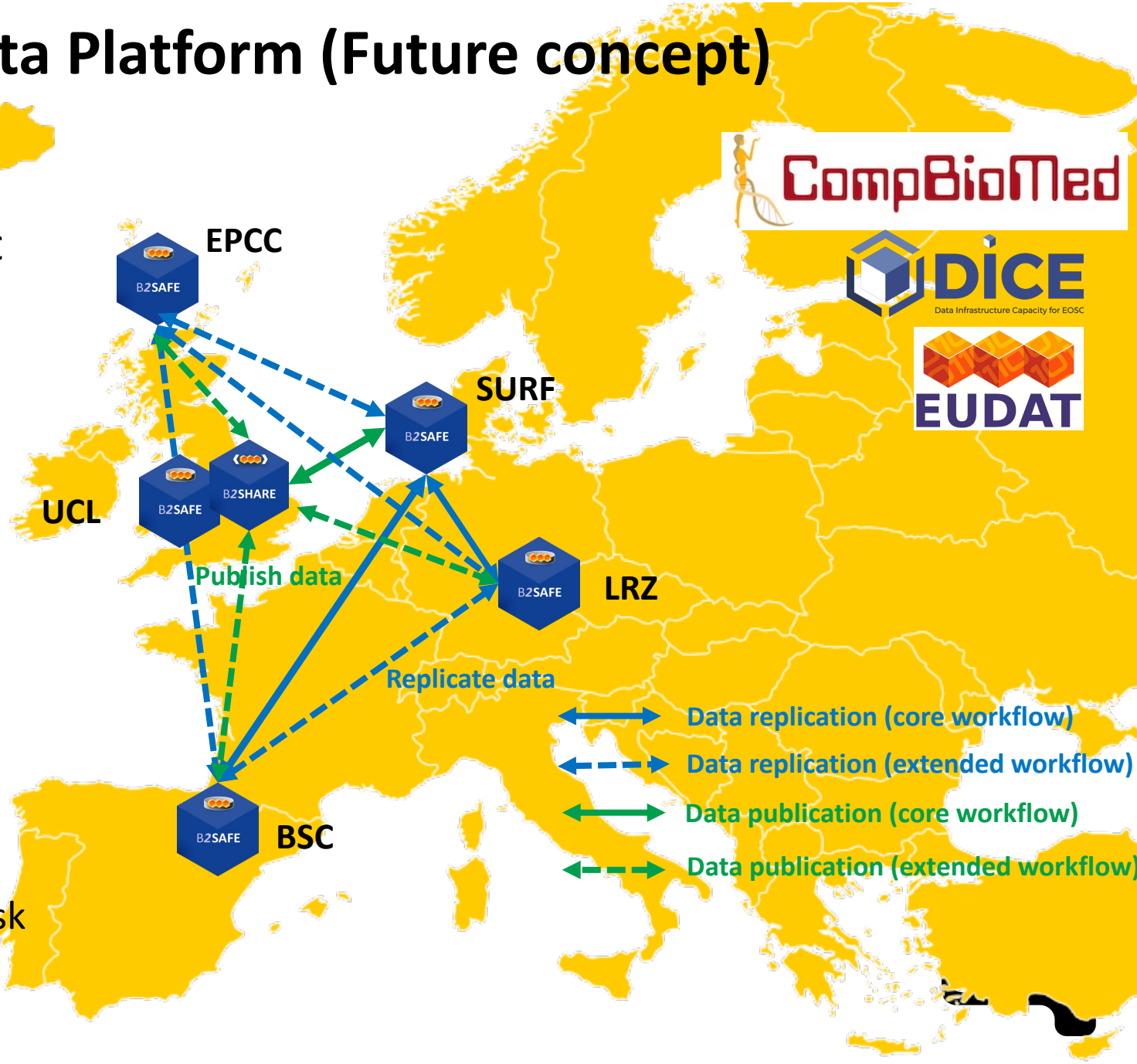


Detailed technical tasks

- BSC (Compute site)
  - Deployment of B2SAFE tool
  - B2Handle or handle prefix (for making PIDs)
  - Federation with SURF B2SAFE endpoints
    - Allocation of storage in B2SAFE
- SURF (Compute and archive site)
  - Deployment of B2SAFE tool
  - B2Handle or handle prefix (for making PIDs)
  - Federation with other B2SAFE endpoints
    - Allocation of storage in B2SAFE and tape storage
    - Monitor integration of B2SAFE-B2SHARE
- UCL (Data publication site)
  - B2Handle or handle prefix (for making PIDs)
  - Deployment of B2SAFE tool
  - Federation with other B2SAFE endpoints
    - Deployment of B2SHARE data repository
    - Integration B2SHARE-B2FIND

CompBioMed Federated Data Platform (Future concept)

- Extend access to the platform to other HPC centers (e.g. LRZ, EPCC), research and medical centers in the community
- Safe data replication and data preservation
- Allocation of PIDs to replicated data
- Facilitate large data transfer
- Bring data close to compute
- Scale-up compute power
- B2SAFE-B2SHARE integration
- Metadata schema for CompBioMed community (addressed in CompBioMed Task 3.4)



UCL Advanced Computing Centre (ARC) Involvement in DICE

- B2Share – a web based service for storing and publishing data sets, intended for European scientists.

- B2Safe – a service for the long-term preservation of research data. Data is replicated between different sites.

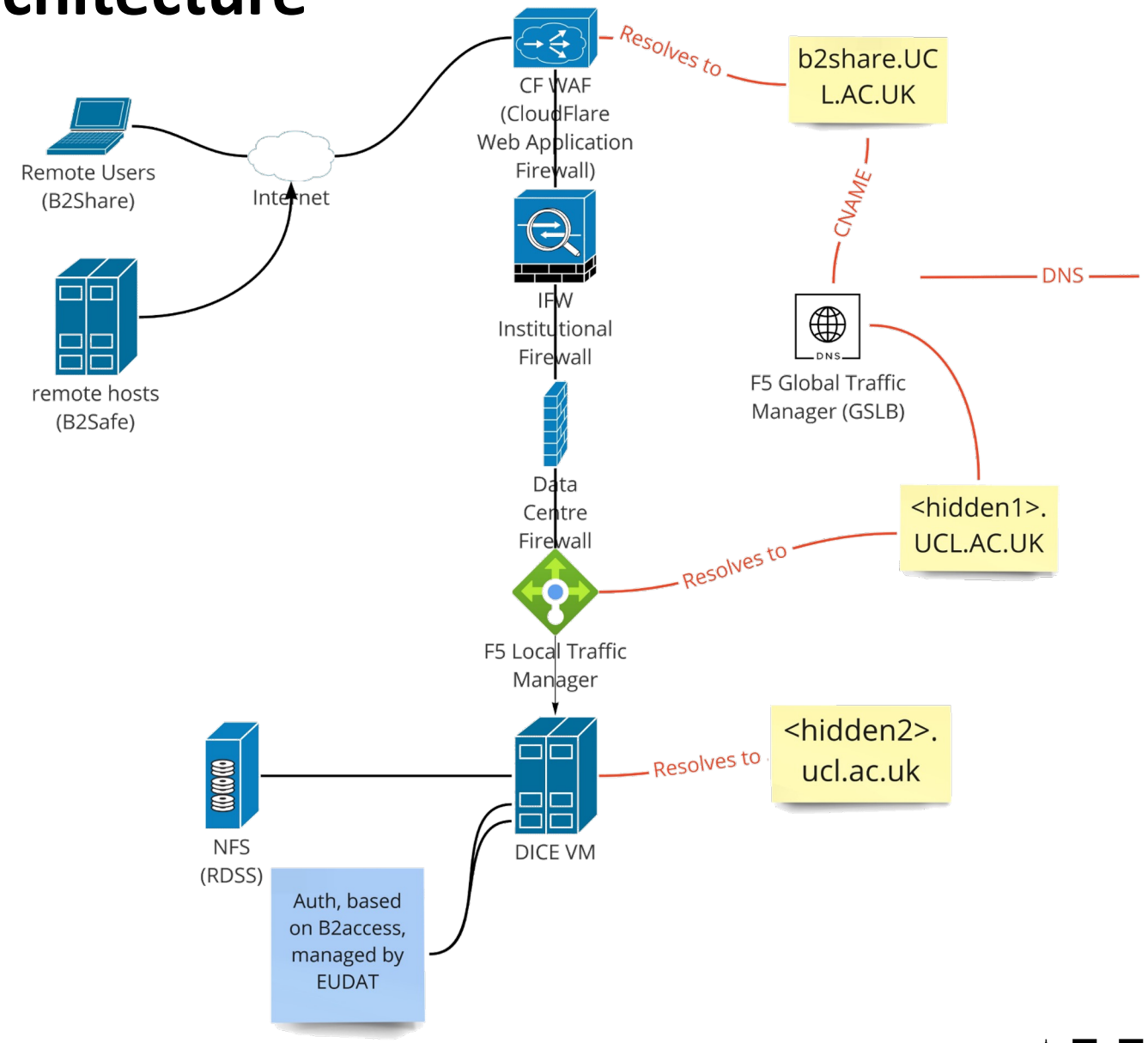
UCL

UCL interest in DICE

- Support the immediate needs of our colleagues in the CompBioMed 2 project
- Evaluate in practice benefits of EUDAT tools as part of the UCL data management ecosystem
  - Current system includes managed Research Data Storage Service (which enables private sharing) and Research Data Repository (Figshare for Institutions) - interested whether EUDAT can complement this longer term

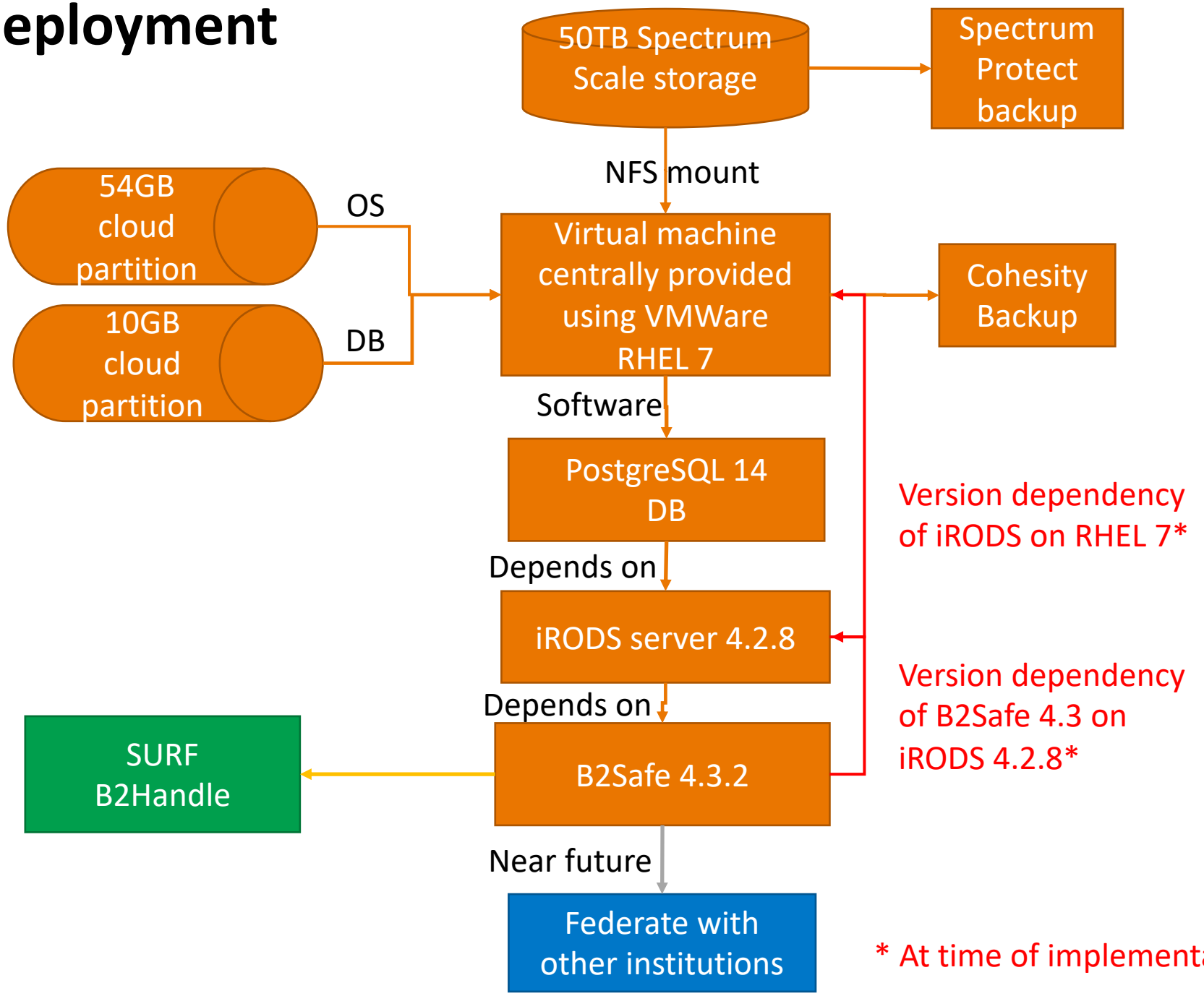
UCL

UCL Planned Network Architecture



UCL

B2Safe deployment



Version dependency of iRODS on RHEL 7*

Version dependency of B2Safe 4.3 on iRODS 4.2.8*

* At time of implementation

UCL

Difficulties

- Usage and installation information spread over different sites.
- Documentation is incomplete
- Yum repo for B2Safe has broken dependency
- Yum repo for B2Safe appears to be down for several weeks.
- Limited support
- Local UCL policies on using latest enterprise OS for security and maintenance purposes (currently RHEL 8).
- iRODS is not compatible with latest OSs (roughly three years lag).
- B2Safe lags behind latest iRODS

UCL

Thank you!