EUROfusion

16th IMEG meeting

Implementation and status of the EUROfusion Data Management plan

Pär Strand, Chalmers for the DMP team

M. K. Owsiak, A. Filipczak, B. Pogodziński, K. Niżnik, P. Grabowski, N. Cummings, S. de Witt, A. Parker, J. Hollocombe, T. Farmer, G. Szepesi , J.-F. Artaud, L. Fleury, F. Imbeaux, P. Maini, J. Morales, R. Coelho, D. Borba, P. Strand, D. Yadykin, D. P. Coster, L. Kripner

This work has been carried out within the framework of the EUROfusion Consortium, funded by the European Union via the Euratom Research and Training Programme (Grant Agreement No 101052200 — EUROfusion). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

PSNC cea irfm MAX PLANCK INSTITUTE FOR PLASMA PHYSICS UK Atomic Energy Authority CHALMERS IPP IPFN INSTITUTO DE PLASMAS E FUSÃO NUCLEAR TÉCNICO LISBOA

16th IMEG meeting

Status of the Implementation of the EUROfusion Data Management plan

Pär Strand, Chalmers for the DMP team

M. K. Owsiak, A. Filipczak, B. Pogodziński, K. Niżnik, P. Grabowski, N. Cummings, S. de Witt, A. Parker, J. Hollocombe, T. Farmer, G. Szepesi , J.-F. Artaud, L. Fleury, F. Imbeaux, P. Maini, J. Morales, R. Coelho, D. Borba, P. Strand, D. Yadykin, D. P. Coster, L. Kripner

This work has been carried out within the framework of the EUROfusion Consortium, funded by the European Union via the Euratom Research and Training Programme (Grant Agreement No 101052200 — EUROfusion). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Outline

- EUROfusion Data management plan
  - Strategic approach - FAIR based
- Scenarios and ”Use Cases”
- Implementation of core services
  - IMAS/UDA dependency
  - Additional components
- Site Services (AUG, COMPASS/COMPASS-U, JET, MAST/MAST-U,TCV, WEST)
  - Data mappings
  - Data availability
  - Technology choices and testing

Data Management Plan

Implementation of the Data Managment Project

Goals:

- Long term goal is to provide FAIR based access to EUROfusion based modelling and experimental data in standardized (IMAS) and searchable formats.
- The Implementation of the EUROfusion Data management plan follows the Blueprint architecture developed by Fair4fusion project. Initially, it promotes access to searchable metadata (waveforms) and data access to a subset of data from EUROfusion discharges on the European devices through IMAS structured data. Longer term it should provide FAIR based access to EUROfusion experimental and modelling data.
- The activity separates into core services providing the infrastructure platform and user interfaces/authentication (PSNC) and sites (AUG, COMPASS (-U), JET, MAST(-U), TCV, WEST).

EUROfusion Data management Plan – staged approach

The data management plan defines 4 scenarios/stages of increasing ambition

- Scenario A: making metadata only available and searchable using IMAS data subsets for interoperable definitions of quantities [F,(I)]. Going into production on the new GW Q4 2024 –Q1 2025. Prototype/demo available for testing review now!
- Scenario B: adds to Scenario A by allowing a subset of the data to be accessed using common tools (UDA). Facilities are responsible for the access level and qualification of data through the data mappings [F,A,I,(R)]. ~~Prototype!~~ Start implement! Original scope extended due to additional funding. Continued focus in 2024/25.
- Scenario C: builds on the previous stages and allows for enhanced data provenance and referencing through PID's [F,A,I,R]. Defer. Resource restricted but important! Some interest to pursue BUT available of expertise at experiments a concern.
- Scenario D: adds a lightweight layer for open access to non-embargoed metadata and where allowed by the facilities also data access for export in human readable formats (CSV files) [F,A,I,R] and open. Defer.

Objectives for 2024/25:

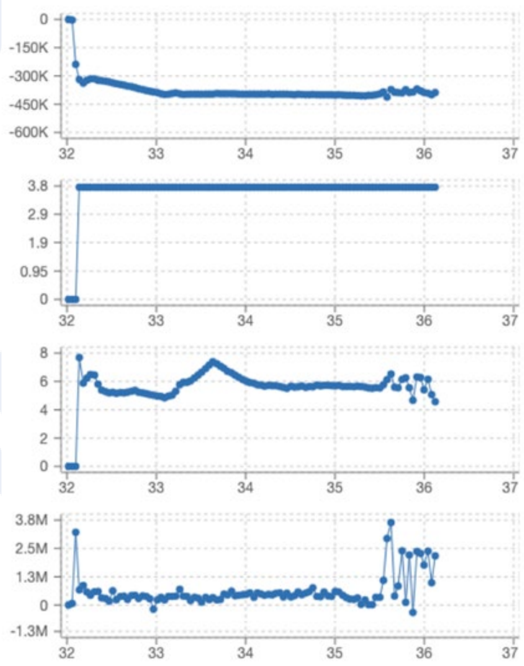
- Provide **searchable metadata** (waveforms) at scale for all participating sites on the new Gateway (scenario A)
- Demonstrate the ability to provide **data access for user driven application needs** (e.g., TSVV-11) for several of the participating sites (scenario B)
- Develop the technology to integrate modelling data through SimDB as a site(facility) of its own –pending the availability of a long term data storage facility!

Longer term vision: A one stop facility for researching, accessing, processing, analysing, and sharing experimental and modelling data.

Scenarios and ”Use Cases”

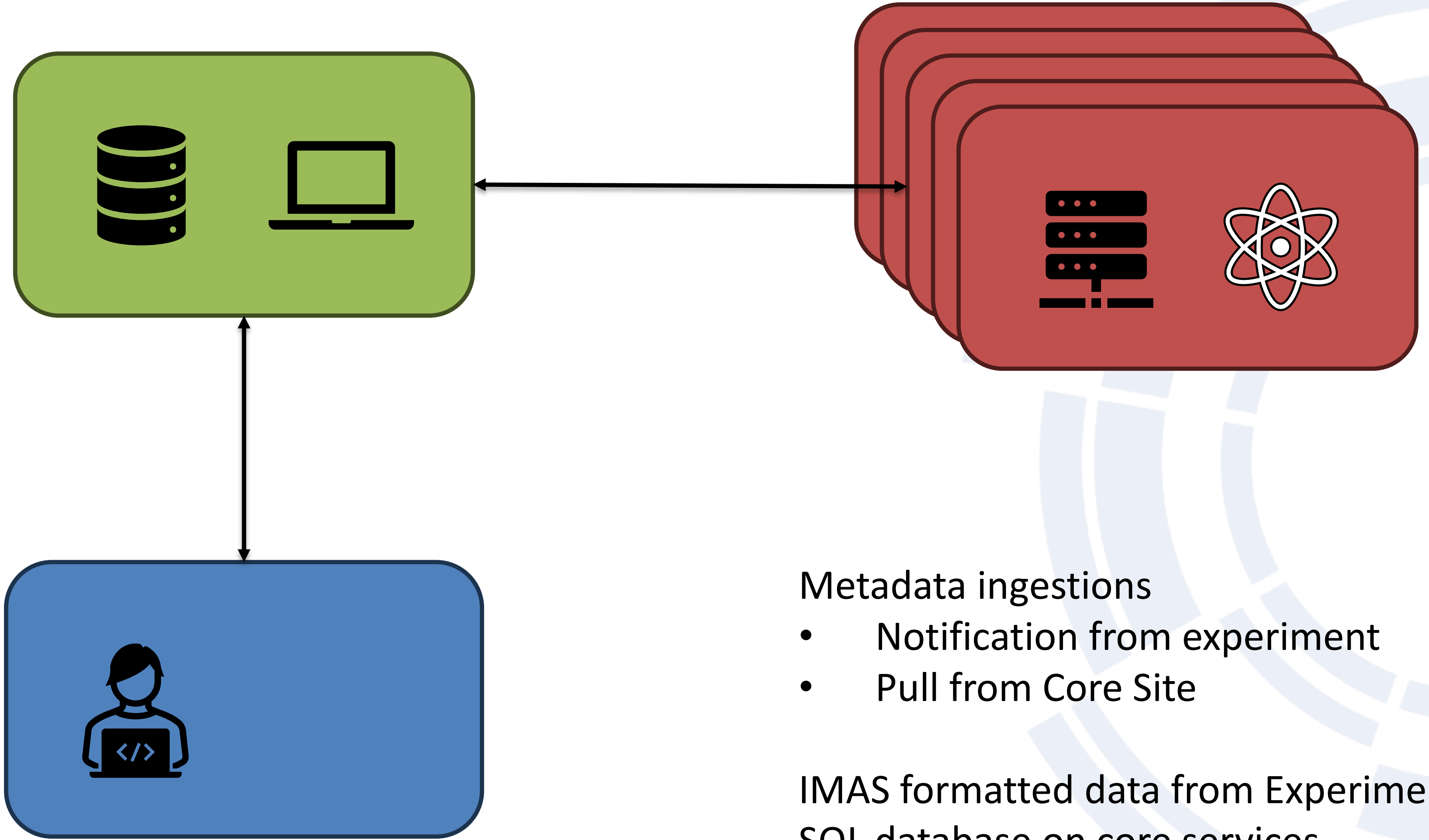
Approach to the ”Scenarios”

- Scenario A: ”Metadata” made available from the fusion devices
  - Metadata is context dependent – someones metadata is someone else’s data and viceversa:
  - Our definition: ”metadata is the waveforms needed for a researcher to not only know that a shot was performed but also to assess the shot for insights and future use”.
  - Also planned as front end for EUROfusion databases on pedestal, disruption, etc
  - Core services delivered and supported by PSNC – provides a front end dashboard providing search option and graphical interfaces
  - Sites provide summary waveforms - (down)sampled to a single time array /discharge which are harvested by core services
  - Provided in IDS format mainly through ”ids_summary”, dataset_description
- Prototyped and ready for larger scale use
  - Waiting for new hardware (EF/gateway) to be made available → Q1 2025.
  - Performance tuning and then production release – several thousands of shots



Dashboard/Metadata server

Experimental sites



User

Metadata ingestions

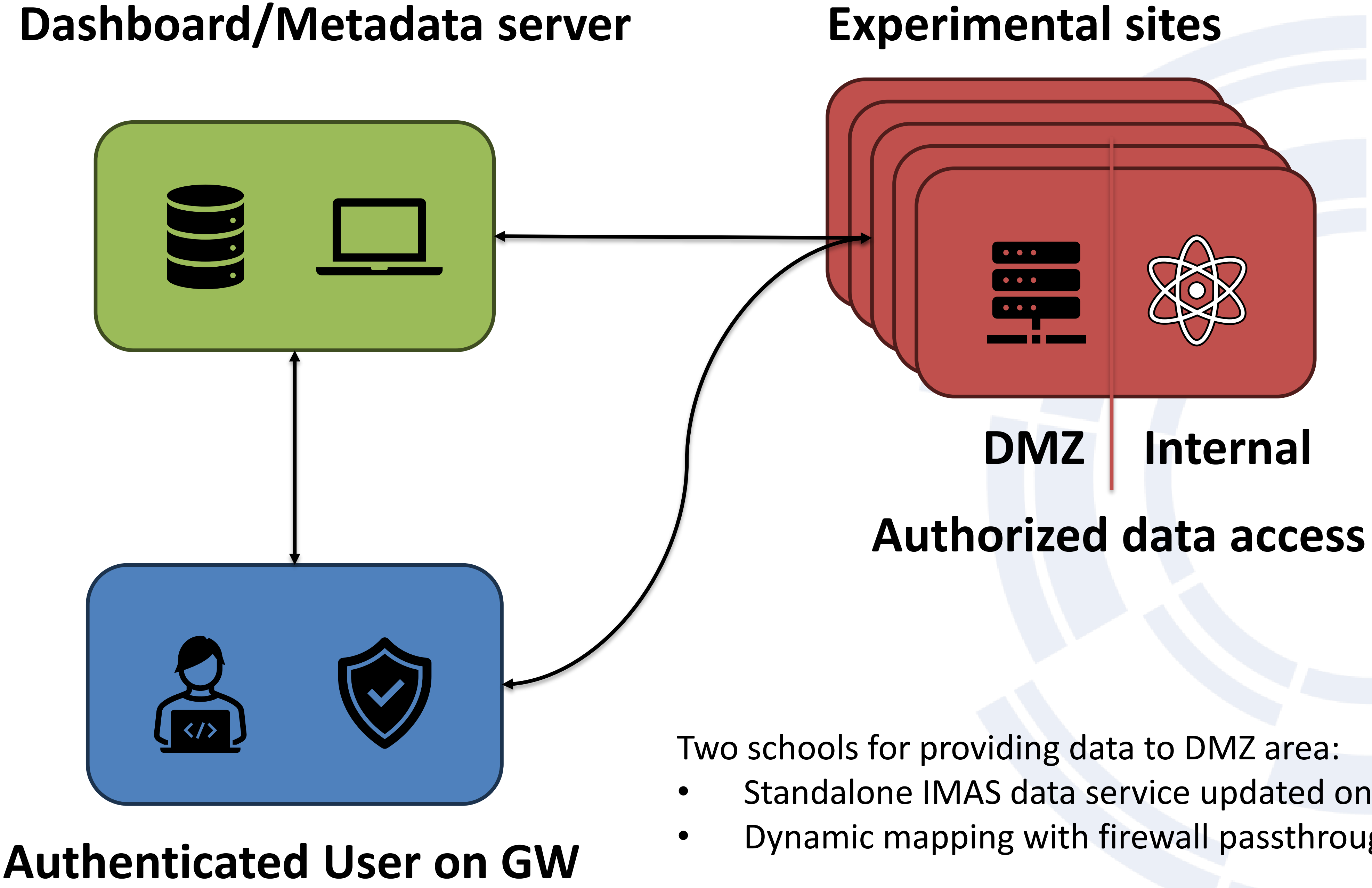
- Notification from experiment
- Pull from Core Site

IMAS formatted data from Experiments sites
SQL database on core services

Approach to the ”Scenarios” – cont’d

- Scenario B: Allow for access to higher (> 0d) dimensionality data for use with common modelling tools.
  - Need to limit data access to data actually used, based on use cases.
    - Use cases → data needed to be available to run a particular code or a ”type” of codes.
    - Need to avoid open ended data mapping requests and provide a clear pathway for validating data mappings.

  - Use cases in the EUROfusion environment:
    1. Equilibrium and MHD stability
    2. Predictive transport modelling (TSVV-11),
    3. Turbulence modelling (TSVV-1) - similar data needs as TSVV-11
    4. Energetic particle workflows (TSVV-10) – as above but a further need for distribution IDS.



Two schools for providing data to DMZ area:

- Standalone IMAS data service updated on need/request
- Dynamic mapping with firewall passthrough

Approach to the ”Scenarios” – cont’d

- Scenario B: Allow for access to higher (> 0d) dimensionality data for use with common modelling tools.

  - The use cases have the benefit of having mature codes and expert code users involved, with codes that are or are in the process of being fully IMASified.
  - A lot of existing work from the EUROfusion WP CD activity can be (at least partially) reused.
  - A drawback is that some use cases need to have ”processed data” - core_profiles and/or core_sources etc. Typically not generated in experiments automatic pipelines.
    - Consider moving more postprocessing at the user end (e.g., IMAS based IDA systems)
      - Issues with data qualification and provenance trails
  - Data mappings effort ongoing, at all participating devices, that are now started to be tested.
  - Data access is through remote access with UDA/AL5 tools with (in the longer term) AAI control.

Implementation of core services

Core services – amalgation of EUROfusion and ITER software components

The ITER Integrated Modelling & Analysis Suite (IMAS)

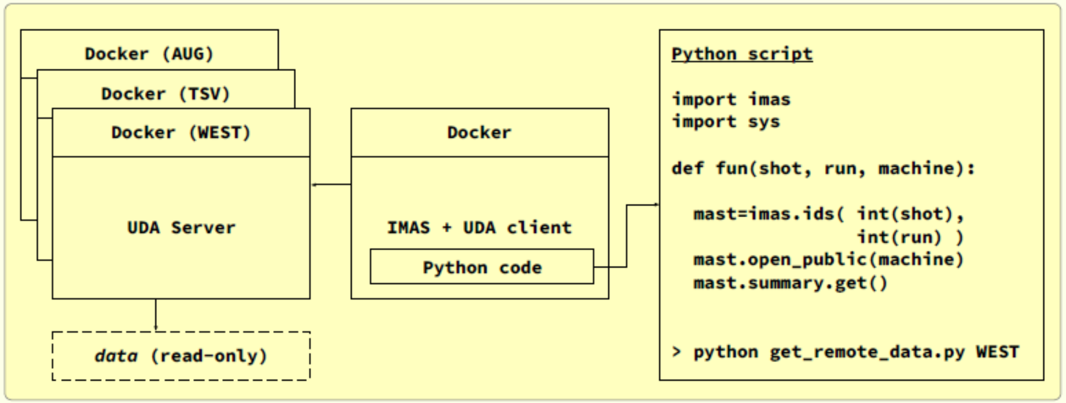
- using **common APIs** for data storage and data access makes it easier to integrate information from various data sources
- machine-generic data structures (**Data Dictionary**) makes it possible to unify the information and store it inside **Interface Data Structures** (IDS)

Simulation code / Data source
High Level API (*Fortran/C++/Python/MATLAB/Java*) + ***Data Dictionary***
common and ***core APIs***
common metadata format - well defined data structure (**Data Dictionary**)
data storage / data access (*MDS+/HDF5/UDA*)

From upcoming SOFT poster:
“Gathering and exposing experimental meta data through a dedicated catalog System”, M. Owsiak et al

Universal Data Access (UDA)

UDA project enables remote sites with the ability of sharing and accessing **IMAS** compatible data remotely. This way it is possible to unify the way data is shared between experiments and clients.

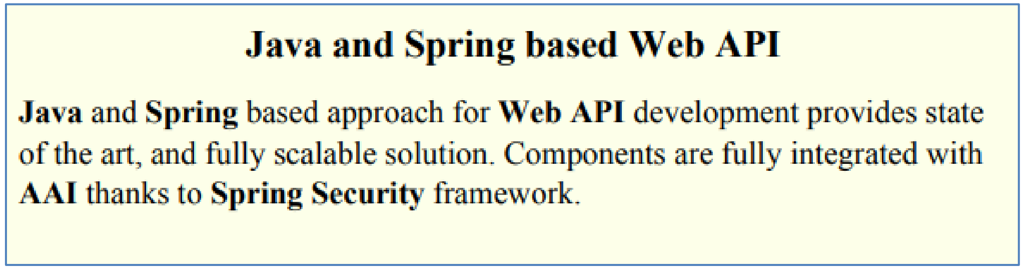


```
import imas
import sys

def fun(shot, run, machine):

  mast=imas.ids( int(shot),
                 int(run) )
  mast.open_public(machine)
  mast.summary.get()


> python get_remote_data.py WEST
```

Containerised infrastructure-agnostic components

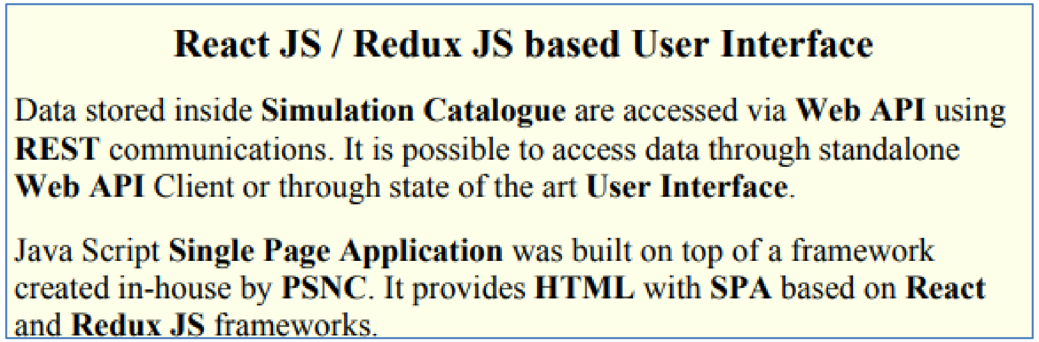
- Using **Docker** and container oriented approach enables us to provide flexible development and delivery of all software components.
- All software components that are part of **DMP** are cloud ready.
- We have combined **Docker Bake** and **GitLab CI/CD** to ensure full automation of containers deployment.
- **Docker** images are available for partners directly from the ***container registry***. Containers can be installed at local sites. This way, experimental site can easily setup **UDA Server** and host experimental data.

User interfaces



Java and Spring based Web API

Java and **Spring** based approach for **Web API** development provides state of the art, and fully scalable solution. Components are fully integrated with **AAI** thanks to **Spring Security** framework.



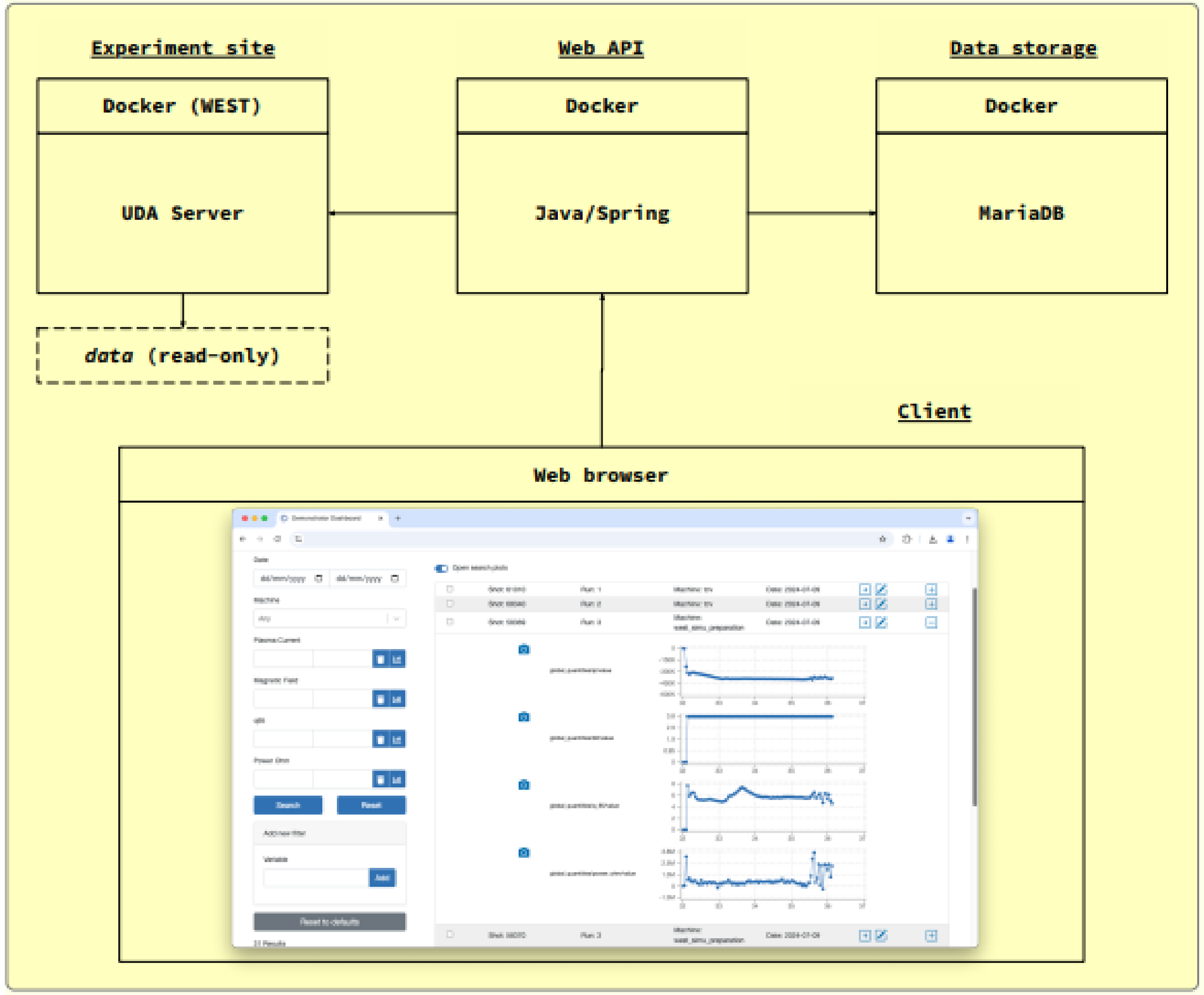
React JS / Redux JS based User Interface

Data stored inside **Simulation Catalogue** are accessed via **Web API** using **REST** communications. It is possible to access data through standalone **Web API** Client or through state of the art **User Interface**.

Java Script **Single Page Application** was built on top of a framework created in-house by **PSNC**. It provides **HTML** with **SPA** based on **React** and **Redux JS** frameworks.

Demonstration planned at the SOFT conference.

Additional tools have been developed for the data ingestion services



Experiment site

Docker (WEST)

UDA Server

data (read-only)

Web API

Docker

Java/Spring

Data storage

Docker

MariaDB

Client

Web browser

Backend infrastructure has been presented at IMEG before
→ now being put in production.

Site services

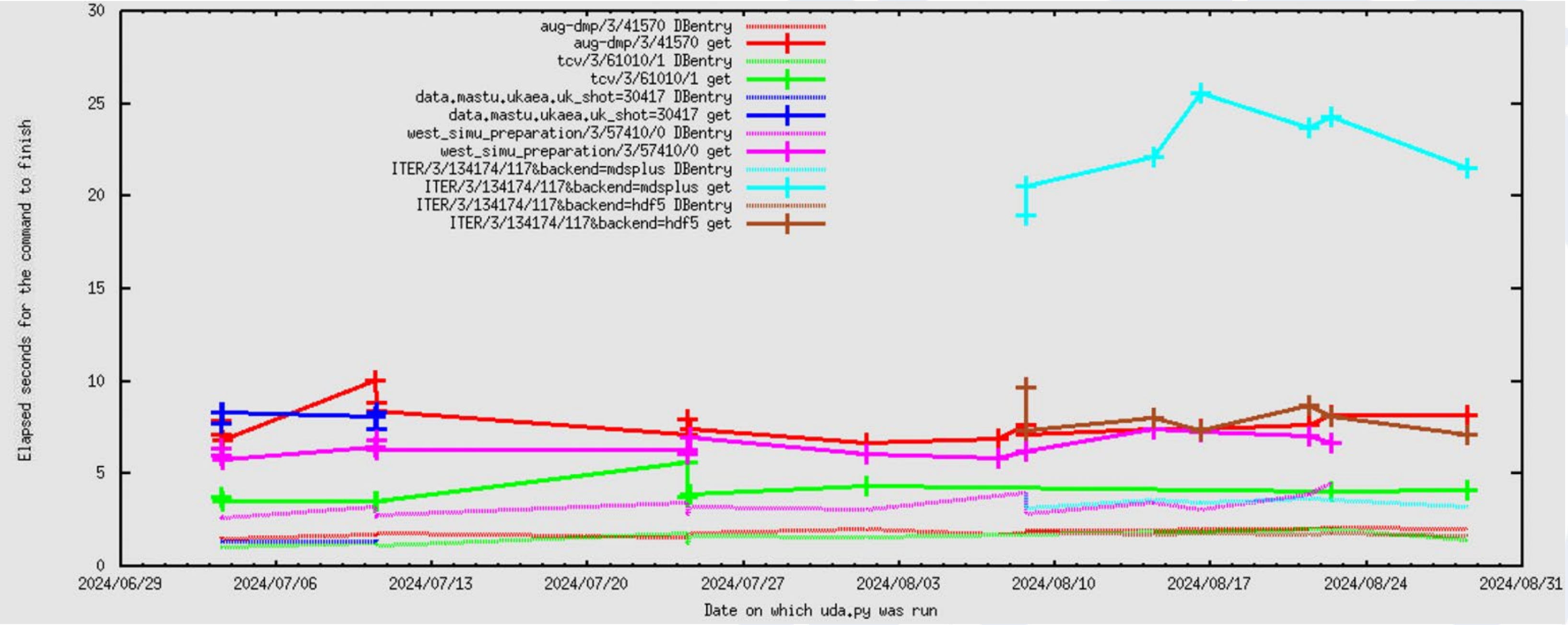
- A “helicopter” view of activites at different sites

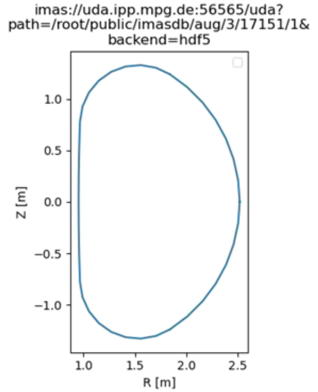
Site contributions from IPP to the DMP implementation

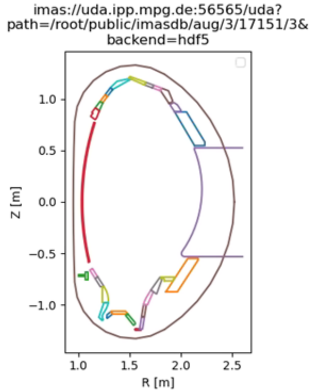
- Installation of IMAS on two of the computing environments at IPP
  - “TOK” cluster: IMAS/3.39.0-4.11.5, IMAS/3.39.0-5.0.0, IMAS/3.41.0-4.11.10
  - “Citrix Virtual Desktops”: IMAS/3.38.1-4.11.4, IMAS/3.39.0-4.11.5
- Provision of uda.ipp.mpg.de utilising docker images from PSNC to provide IMAS files containing
  - SUMMARY IDS information (11150 AUG shots)
  - Data for one shot mapped by trview ('core_profiles', 'dataset_description', 'equilibrium', 'ic_antennas', 'nbi', 'pulse_schedule', 'summary', 'tf', 'wall')
  - Data for one shot with mappings of 'equilibrium', 'magnetics', 'pf_active', 'tf', 'wall' (two variants differing on the wall data)
  - All without any security model
    - Preliminary work on implementing a security model just started
- Preliminary work on using UDA mapping in the very early stages

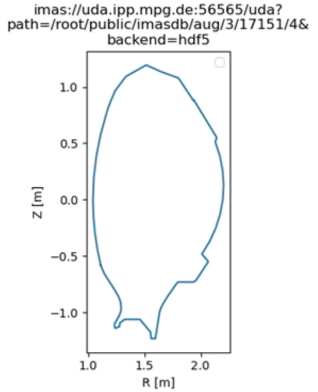
Timing tests for UDA access and 3 AUG wall representations

- Program to time the access to UDA provided data written
  - HDF5 seems faster as a back-end
- Also provides representative plots for some of the available IDSes
  - Here 3 different examples of the AUG wall







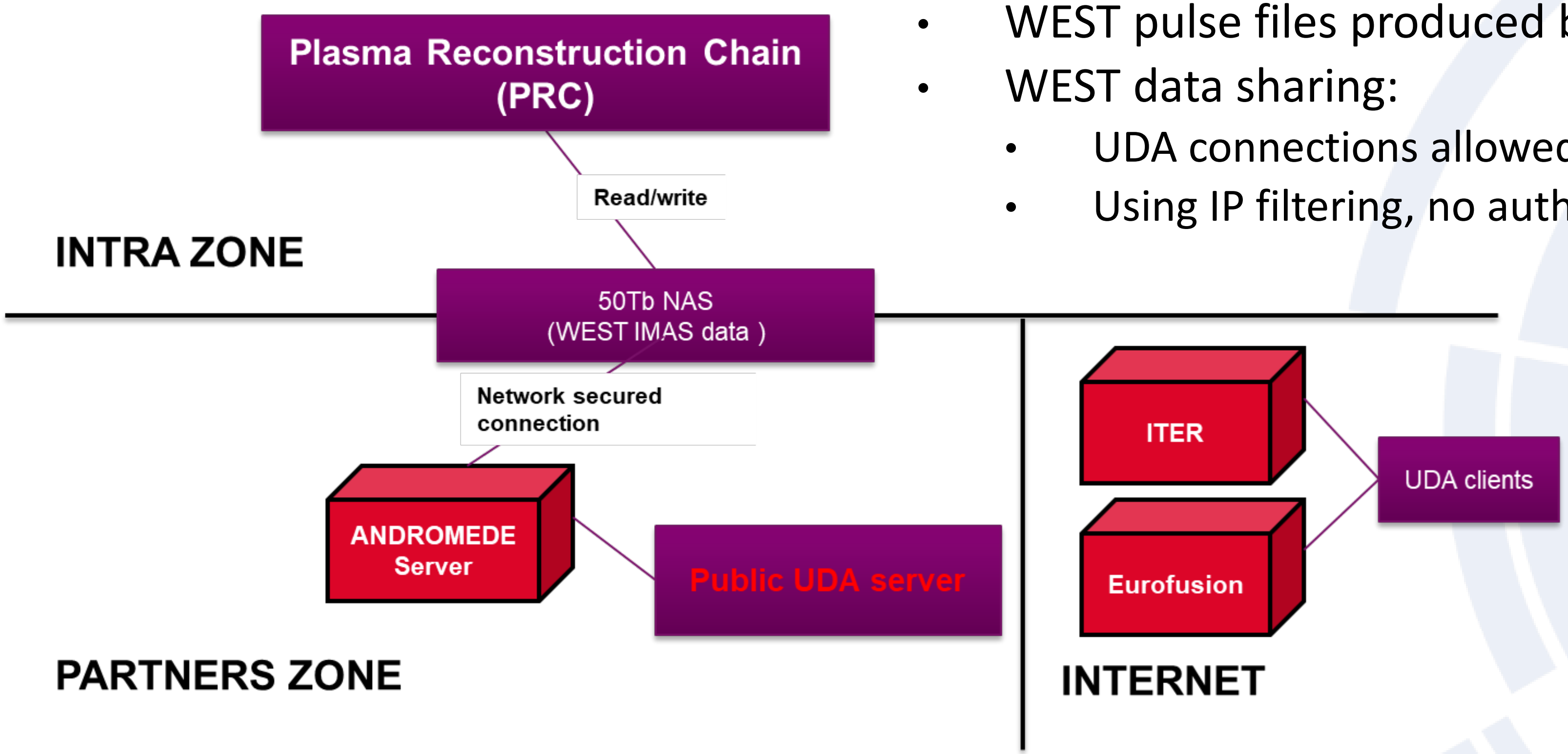


Site contributions from WEST to the DMP implementation

- Installation of a public UDA server on WEST Partners Zone
  - Installation of IMAS Access Layer v5 for using the latest IMAS plugin for UDA (WEST production still uses v4)
  - Installed with plugin 'IMAS_PLUGIN' which reads directly HDF5 pulses files stored on 50TB NAS
  - CEA IT WEST data access agreement using IP addresses filtering
  - Firewall updated accordingly (ITER, Gateway, PSNC)
  - SSL certificates installed

- Full WEST 'C7' campaign available for DMP scenario A
  - SUMMARY IDS information available for ~1400 shots
- DMP scenario B already enabled for accessing data of all IDSs from WEST pulse files (all campaigns)
- Plan is to use AL5 for the next WEST campaign in october 2024

Providing WEST data



- WEST pulse files produced by PRC components
- WEST data sharing:
  - UDA connections allowed from: ITER, Gateway and PSNC
  - Using IP filtering, no authentication

- 2 IMAS databases are available:
  - One containing diagnostics processed data
  - One including METIS interpretive simulations carried out systematically for each pulse

TCV: IMAS Infrastructure

- 2 Dedicated servers
  - **spcimas**: Server for local IMAS processing/IDS generation
  - **spcimasdata**: Server open to EF GW/ITER for accessing TCV data via UDA
  - Red Hat Entreprise Linux 9 compatible distribution (Rocky Linux 9)

- IMAS Environment
  - Server configuration, software building is fully automated using Ansible
  - Built with most recent CMake based installer
  - As of August 2024: IMAS/3.41.0-5.3.0 , idstools/1.14.2 , uda/2.7.5

TCV: Available Data

- IDS Generation done with "tcv2ids" matlab package in **spcimas**
- Summary IDS generated for ~6000 shots (Will be extended to all TCV shots)
- summary, equilibrium, pf_active, core_profiles, nbi, tf, wall, thomson_scattering ids generated for one example shot
- For now, HDF5 backend selected for lower disk use
- For future shots, summary IDS generation is automated

Site contributions from MAST(-U) and JET to the DMP implementation

Current setup

MAST(-U):
UDA server installed on externally facing hardware at UKAEA
'On-the-fly' mappings of data to IDSs available using UDA, JSON mapping plugin, and IMAS plugin
Authentication currently relies on IP whitelisting for connection to *data.mastu.ukaea.uk*
SDCC and GW are whitelisted and can remotely connect to the mapping server

Preliminary summary IDS mapped for both MAST and MAST-U, **however interpolation method to be finalised**
"imas://data.mastu.ukaea.uk:56560/uda?mapping=MAST&path=/&shot=30417"

JET:
UDA server installed on the JDC, similar setup the mappings for MAST(-U)
Externally accessible via *uda.jetdata.eu* from the GW
Currently no authentication restrictions in place as no PPFs or JPFs are available via UDA server
Authentication will need to be controlled in the future

Preliminary summary IDS mapped for JET using the CPF output
Example uri: ***"imas://uda.jetdata.eu:56565/uda?mapping=JET&path=/&pulse=80000"***

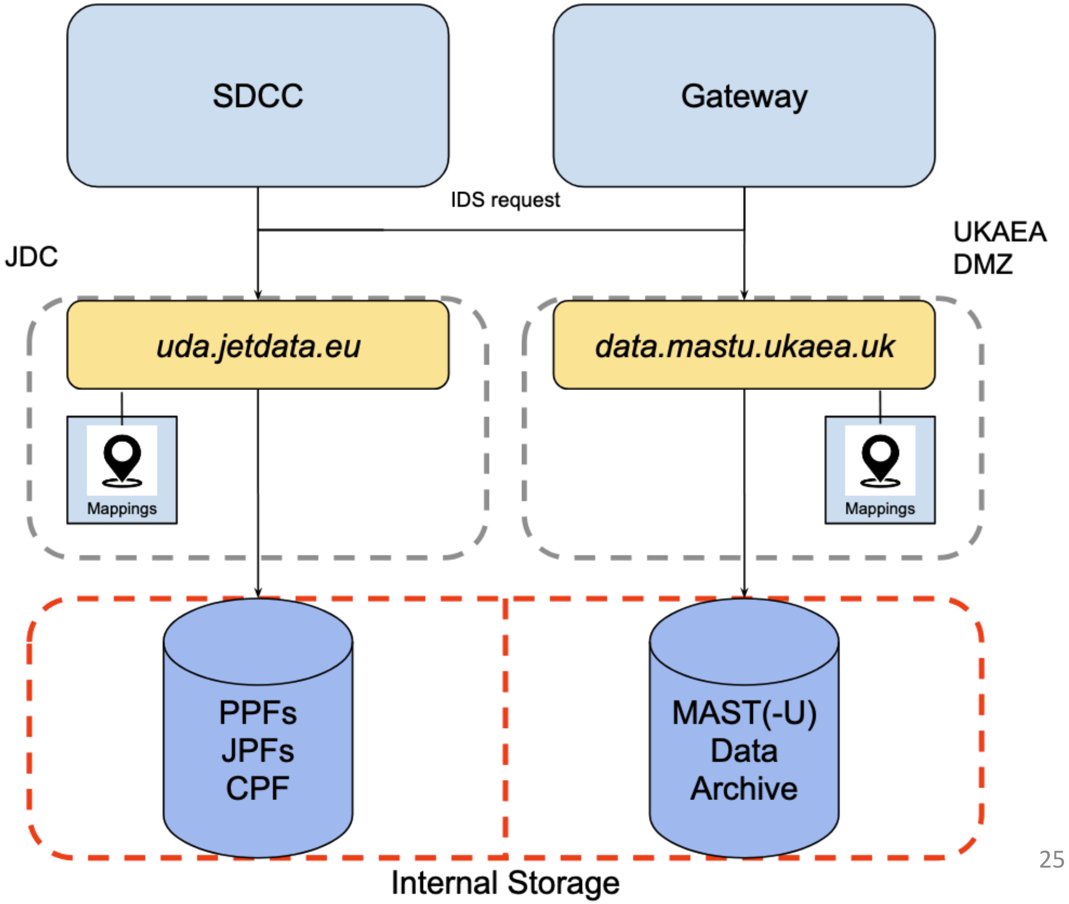
To provide dynamic mappings only al-cpp (al-core) is needed.
Current mappings target DD3.39.0 and tests have been performed only with AL5.0.0+

Future Plans at JET

- Harmonise authentication and authorisation options for UDA with DMP/SICO groups and experiments (- currently MAST(-U) required SSL certificates)

- Finalise **summary** and **dataset_description** for all machines

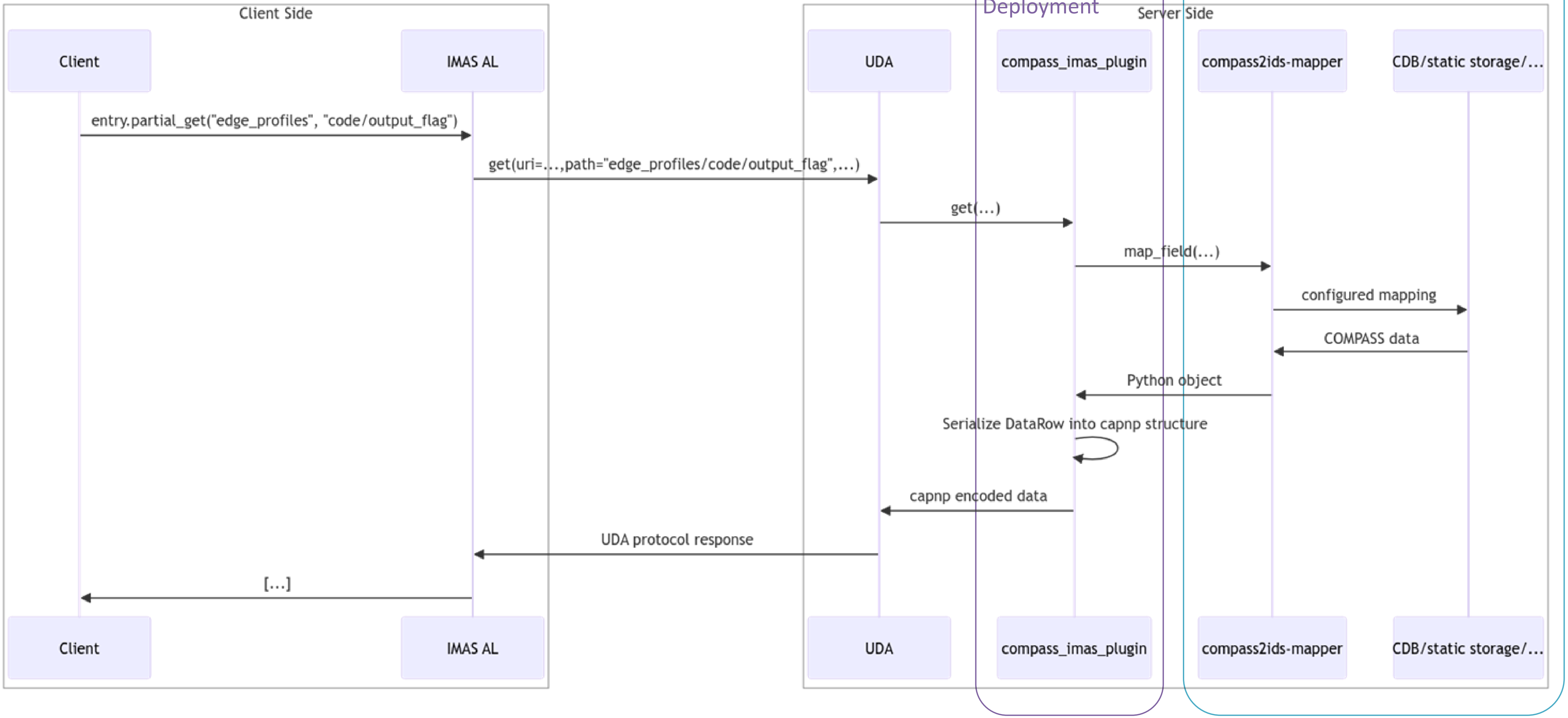
- Planned IDSs to support via mappings:
  - magnetics,
  - pf_active,
  - pf_passive,
  - wall,
  - tf,
  - Equilibrium



Initial mappings for many of these IDSs already exists

COMPASS data to IMAS IDS mapping

Team: Lukáš Kripner, Ivo Hanák, Stanislav Tokoš



UDA plugin

Py →C++ serializer

Deployment

COMPASS to IDS mapper

Python library

Perform actual data mapping

Disclaimer: The proposed architecture is in its early stages and may be subject to change.

COMPASS IDS plugin

- Developed with best practices:
  - continuous testing
  - continuous deployment as:
    - docker image or deb package
- Deployment version:
  - Depends only on open-source packages (e.g. UDA server)
- Testing version:
  - Utilizes Python IMAS AL for end-to-end testing
- Python to C++ serialization
  - Python facilitates easier development of future mappings
- Mapping process
  - Actual mapping performed by the `COMPASS to IDS mapper` library.
- First deployment of testing COMPASS UDA server:
  - Limited to selected IP addresses
  - Scheduled for the end of September 2024
- Next steps:
  - complete COMPASS mapping schema
  - Secure UDA server:
    - Consider using only SSL and/or OAuth2 for authentication.
  - Deploy a second instance with COMPASS-U synthetic data, followed by experimental data when they become available.

COMPASS to IDS mapper

- Standalone Python Library
  - Relies solely on open-source dependencies (e.g., OMAS)
  - Optionally supports closed-source backend libraries (e.g., IMAS)
- Uses json/yaml configuration files for mapping specification
- Mapping Capabilities:
  - Supports mapping to various data sources:
    - COMPASS database (CDB) [1]
    - Static files (e.g., HDF5, netCDF, CSV)
  - Future enhancement: machine description will be sourced from the machine description server
  - Aggregations:
    - For example, summary IDS fields require resampling of the time axis to unify data from various sources
- Direct mapping of both: single IDS field and whole IDS entry
- Core `Abstract mapper` provide with functions: `map_entry`, `map_field` for flexible mapping

[1] J. Urban, et al.: *Integrated data acquisition, storage, retrieval and processing using the COMPASS DataBase (CDB)*, Fus. Eng. and Design, **89** (5), 2014

```
{
   "entry": "summary",
   "time_axis": "time",
   "time_factor": 1e-3,
   "resampling_method": "interp1d",
   "resampling_args": {
       "freq": 1e4
   },
   "fields_or_datasets": [
       {
           "dataset": "EFIT",
           "fields": [
               {
                   "source": "li",
                   "target": "global_quantities.li",
               },...]},
```

```
{
           "dataset": "MAGNETICS_RAW",
           "fields": [
               {
                   "source": "diamagnetic_loop_1_1_RAW",
                   "target": "global_quantities.v_loop",
               },
               ...
            ]}}]
```

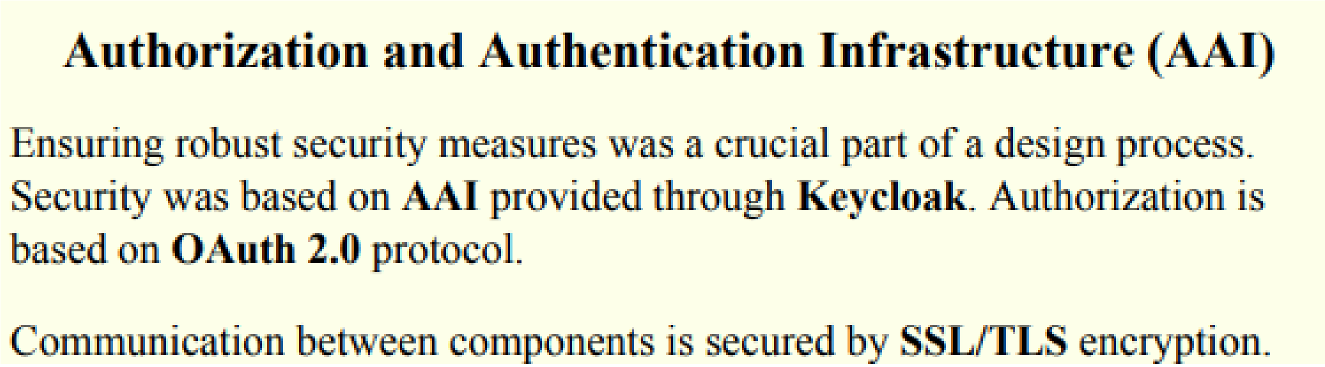
Some further considerations

Exploring support for IMAS EasyBuild option

- IPP are using EasyBuild to try to build IMAS
  - On the Citrix Virtual desktops (running Ubuntu)
    - Success with both FOSS and (eventually) intel versions
      - IMAS/3.41.0-2024.07-foss-2023b
      - IMAS/3.42.0-2024.08-foss-2023b
      - IMAS/3.42.0-2024.08.1-foss-2023b
      - IMAS/3.42.0-2024.08.1-intel-2023b
  - With TOK cluster (running SLES 15)
    - EasyBuild does not seem to support SLES that well
    - But eventually succeeded
      - IMAS/3.42.0-2024.08.1-foss-2023b
      - IMAS/3.42.0-2024.08.1-intel-2023b
- TCV are also starting to look into EasyBuild for IMAS to have similar environment with ITER servers
- Evaluation/testing at EUROfusion gateway to do the same

AAI selection/decision to be taken

- Harmonise authentication and authorisation options for UDA over the different sites
  - Some sites (e.g., MAST(-U) required SSL certificates)
  - Issues possibly more policy oriented than restricted by technology.
- Proposal



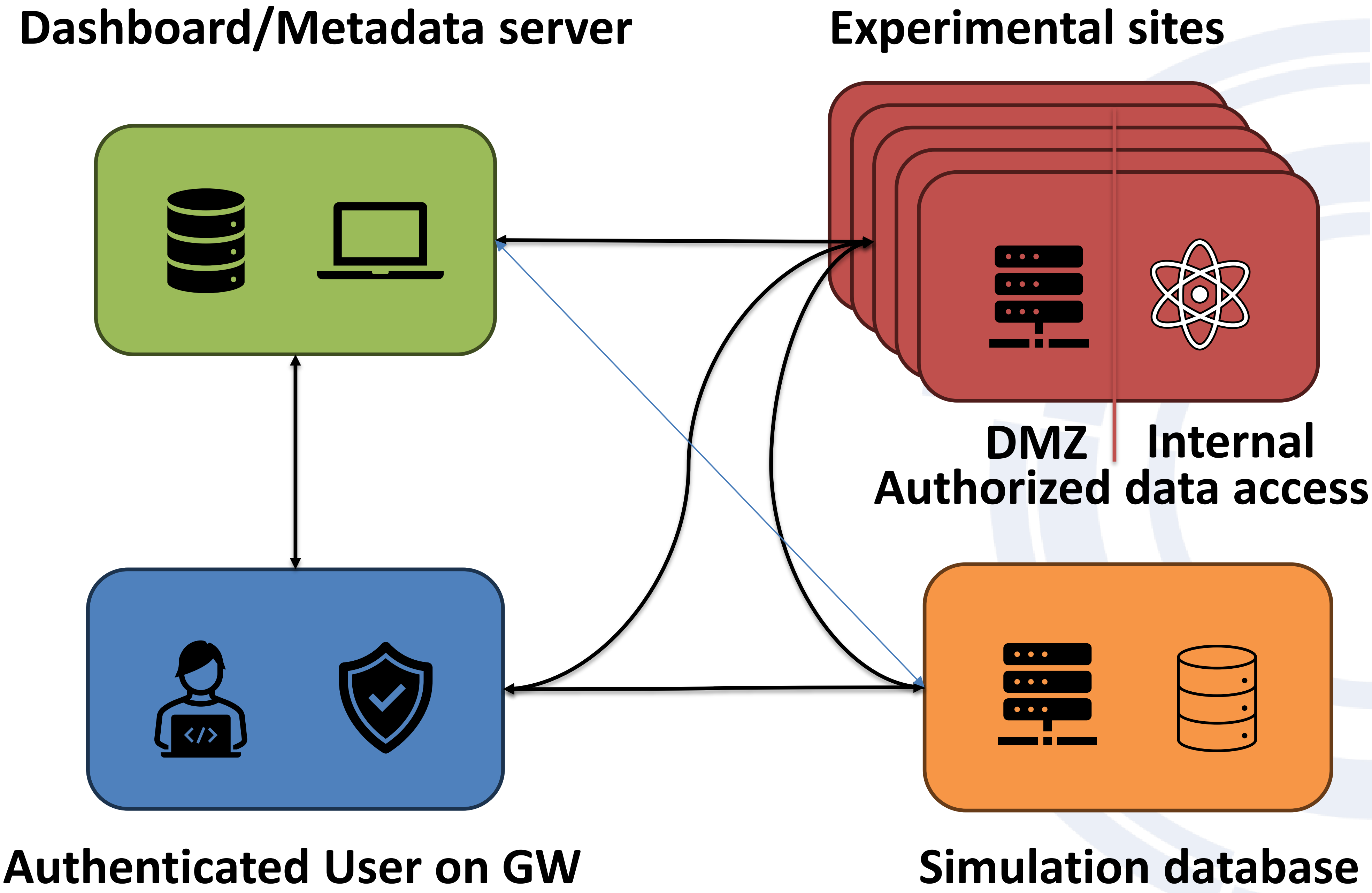
Authorization and Authentication Infrastructure (AAI)

Ensuring robust security measures was a crucial part of a design process. Security was based on **AAI** provided through **Keycloak**. Authorization is based on **OAuth 2.0** protocol.

Communication between components is secured by **SSL/TLS** encryption.

Promote wider use of UDA (JSON) mappings

- Already the chosen option for MAST(-U), Compass-U and JET
- Will implement/test some UDA mappings to AUG data
  - Rather than writing IMAS files that has been the approach takens so far
- TCV is considering using json files (used by UDA plugin) as single point of entry for tcv2ids and any future solution
- (WEST is native IMAS)



Dashboard/Metadata server

Experimental sites

DMZ | Internal

Authorized data access

Authenticated User on GW

Simulation database

Summary

- Implementation of the EUROfusion Data management plan is underway
- Several sites: AUG, TCV, WEST, JET, MAST(-U), COMPASS(-U), different stages of dev.
- Scenario A: “wave form metadata” mostly ready.
  - AUG (1150shot), TCV (~6000 shots) and WEST (1400 shots) ready to scale up to more complete coverage (including automatic updates of new discharges in some devices)
  - Held back by need to move to new Gateway platform for production level hardware
- Scenario B: > 0D data for simulations
  - Benefits from previous work and existing routines (Trview, tcv2ids,…) and machine descriptions from WP CD
  - First test for users and use cases: ETS, HFPS, GENE, …
  - Need AAI implementation/agreements before it can move to broader production
- Hardening infrastructure towards production
- Exploring different technologies
  - AAI
  - Easybuild
  - JSON plugins etc.