Contribution ID: **61** Type: **Oral (Regular)**

Global Data Sharing: A Comparative Study of Pelican and CVMFS for remote access to MASTU and DIII-D data for UFO detection workloads

Friday 12 September 2025 10:15 (25 minutes)

Data aggregation across multiple fusion devices has enormous value for improving machine learning models and for validating simulation tools. One challenge in forming and using such datasets can simply be the latency caused by the distance between experimental sites and the computational facilities where data is used. Other challenges arise from the different data access interfaces exposed by each data provider, and the formats and representations of that data.

Pelican Platform[1] and CVMFS[2] are both examples of data distribution services which provide consolidated access interfaces and use caches and data mirrors to reduce latency when accessing multiple, globally distributed data sources. Pelican is a data federation platform which aims to unify access to different kinds of storage backends (S3, Posix, HTTP) through adaptor services, it uses XRootD for data transfer and includes features for user authorisation and authentication. CVMFS is a CERN-developed data distribution technology widely used in High Energy Physics, initially focused on sharing software, it uses the HTTP protocol for data transfer with a convenient virtual filesystem interface and aggressive use of local caching to improve performance for certain data access patterns.

In this presentation we will cover our experiences to date comparing these two potential infrastructure frameworks and how we made use of them to share fusion data between data sources and HPC facilities at both UKAEA and DIII-D. This data sharing involves the use of SciTokens[3] (a federated authentication and authorisation infrastructure) that has been used at DIII-D to enable remote connection to their dataset via the US DOE-funded Fusion Data Platform [5], an initiative led by General Atomics.

We use a UFO detection tool, ENEJETIC[4], as a demonstration HPC workload for processing remote image data from both the MAST and DII-D tokamaks. UFOs are a class of impurity within the plasma that can lead to damaging disruption events and are often formed from debris coming off the wall of the vessel. They are difficult to monitor during operation without human intervention, but quick responses are required to avoid disruption. Originally trained from JET image data, ENEJETIC (Enhanced Neural Engine for JET Image Classification) uses a convolutional Neural Network to automate the process of the detection and logging of these events.

[1] Pelican Platform, https://pelicanplatform.org/
[2] CernVM File System, https://cvmfs.readthedocs.io/en/stable/
[3] SciTokens, https://scitokens.org/
[4] Phys. Plasmas 32, 042508 (2025) https://doi.org/10.1063/5.0261120
[5] Fusion Data Platform, https://ga-fdp.github.io/

Speaker's email address

stephen.dixon@ukaea.uk

Speaker's Affiliation

UKAEA

Member State or International Organizations

United Kingdom

Authors: Dr PARKER, Adam (UKAEA); Dr SAMMULI, Brian (General Atomics); Mr CLARK, C.M. (General Atomics); Mr HOLLOCOMBE, Jonathan (UKAEA); Mr FIELD, Matthew (UKAEA); Dr DEWITT, Shaun (UKAEA); DIXON, Stephen (UKAEA)

Presenter: DIXON, Stephen (UKAEA)

Session Classification: Information Retrieval and Visualisation

Track Classification: Information Retrieval and Visualisation