Minutes of the CERN/EFP Meeting

Sept 25th 2025

Agenda: https://indico.cern.ch/event/1586866/

Intro

- Background:
  - The EuroHPC Federated Platform (EFP) project, launched in early 2025, aims to develop a unified federated computing infrastructure for European scientific communities.
  - The Worldwide LHC Computing Grid (WCG) seeks to integrate HPC resources into its workflows.
- Goals:
  - Align technical requirements and interfaces between EFP and WCG teams.
  - Discuss challenges and opportunities for HPC integration.
  - Establish a roadmap for ongoing collaboration and resource sharing.

After the first handshaking in January (meeting at CERN), where experiments were presented, this is intended to be a more technically oriented meeting where services, protocols, tools are described.

EFP Intro and Status

The scope and the status of the EPF project are described.
"The EuroHPC Federation platform provides a one-stop-shop for accessing and utilizing EuroHPC systems and services" The platform will provide:

- Unified solution for users to manage their access and resources across all EuroHPC systems.
- Unified software offering across the systems to reduces the cost and complexity of migrating workloads or utilizing multiple systems in complex workflows
- Advanced graphical interface for: o Creation and management of workflows o Interactive computing using powerful domain specific tools o Novice users
- Features for power users accustomed with direct access to systems and APIs
- Smart scheduling capabilities to optimize compute usage and increase the level of abstractions for the provided compute capacity

Timeline is divided into a first phase (2025-2026) where a Minimum Viable Platform will be deployed, including federated AAI, resource allocations, and complex workflow manager, User

interfaces and helpdesk). The first release is expected Q1/2026, supporting all currently online 9 EuroHPC systems, and will enable user testing; a second release in Q4/2026 will include more systems and features.
The second phase (2027-2029) will include enhancements and support more systems.
A few points arose from the discussion:

- The EFP enables services and features to be levelled among systems. It is for example expected that a single access to the resources obtained via EuroHPC grants will be obtained via the EFP, with the possibility to allocate/deallocate resources.
- The EFP does not impose technical changes / features on the HPC centers, which remain ultimately in control of the resources. Services like CVMFS, access to edge services, remote SLURM API interfaces can be available via the EFP, but only if a center decides to expose the service.

The EFP has the mandate to cover the central (EuroHPC managed) quotas; the national quotas (for the fraction paid by local governments) is not by default included, but in principle the same tools could be used. More specifically, non-EuroHPC allocations on EuroHPC JU systems can utilize the EFP if they are managed in the EFP and sites do the required integration work.
The EFP architecture is described, with its main components. Meeting discussions are focused mainly on the AAI part, which is seen as the first component to synchronize / adapt between WLCG and EFP, in case an integration path is followed. After that, Data Management and Workload distribution would be the next steps.

About the AAI part, third party integration (e.g. for building SaaS) is currently out of the scope, depends on center by center and central decisions from EuroHPC JU. Cross organizational authorization (apart from ssh access) not in scope for the release, but scheduled for the second release. This would enable access via the EFP to web-based services offered at the federated sites.

About Workload management, it should be possible either via Web access and ssh, using keys generated via a portal / an hardware key. The lifetime of ssh keys is expected to be kept short, but this can depend on special arrangements with the centers. Currently only accounts directly associated with physical persons (== “a passport”) are expected to be present, given possible limitation of access for certain nationalities. The case for “group accounts” has been discussed, but eventually would be again a site decision. In the case of large collaborations (say 1000s users) in principle every user accessing the resources would need to sign the TOS (terms of service), which is difficult also given the problem of nationalities.
It is agreed in the discussion that it can be surpassed by using the resources (at least initially) only via the submission of a few users, typically the MC production operators.

About data management, some sites are going to expose a S3 interface towards the global internet; it is a site decision whether this is directly attached to the parallel file system seen by the nodes, or it has some kind of sandboxing. For the current online EuroHPC systems, only a few expose S3 and no systems connect that directly to the parallel filesystem. EOSC, SIMPL, AI factories will need to be integrated (not in the first phase).

EESSI software distribution is explained: while CVMFS will be generally available on the systems, there is no guarantee external repositories will be allowed (like the experiments' software stacks).
In some centers EESSI will be provided to the compute nodes via local areas on the parallel filesystem, synchronized from the CVMFS repo quite often (daily or hourly). The same approach could be used for experiments in the same cases, even if there is a large concern that WLCG CVMFS areas are too big. About the CVMFS local and SQUID caches, it is reported from WLCG sites that site SQUIDs are as small as 100 GB, with local node caches ~ 40 GB.

AAI

The WLCG AAI is described in detail. It is based on JWTs with a special WLCG specific schema. This is discussed to be principle compatible with EFP AAI, modulo a token translation service which should be accepted by the centers (technically as easy as a LDAP, the problem sits in the political part).
The use of tokens in WLCG is described for typical DM and WMS use cases.

Data management

Typical data management fundamentals as used by WLCG are presented, with features like third party copies etc.
In the case of HPCs, a few modalities are discussed:

- The use of site services like S3. Technically could work but it is not seen as a performant solution, and the fact that S3 and parallel file systems can be out of sync is problematic
- The request of edge nodes which see the parallel file system, on which to deploy WLCG services. It would be optimal, bypassing the EFP/center data management, but hard to obtain apart from some specific centers.
- Again, one could start by deploying workloads at HPC centers which do not cover all the spectrum. The use of HPCs to process MC from scratch (or from generator files) could reduce by a lot the I/O needs, and suggest the deployment of a caching system instead of a fully managed data management. That would need the "how to stage the output" to be solved in a different way.

Workload management

The workload the experiment uses is a different object with respect to the workload management EFP deploys.
The point of contact is the need to have a mechanism to provision pilots, which in turn then join the "experiment side" workload management, which is a different object.
This is true at least in ideal terms; there are possible solutions to drop the pilot model in case of needs, relying on the SLURM at the center and to bulk submission – but not all the experiments seem to be ready for this and certainly the model introduces large limitations.

In an ideal pilot model, the work done by the experiments in the large 10 years largely decouples the HPC and the WLCG environments:

- Pilots need to be started on compute nodes; the pilot mechanism is the preferred but there are other options (vacuum models, pilots sent from inside the site using pressure mechanisms …).
- The availability of virtualization (apptainer/singularity is the preferred since it lives in userland) and software on-demand via CVMFS reduces by a lot the needs of the intervention of sysadmins and the customization of the nodes.
    - It has been shown that userland solutions like cvmfsexec are a possible solutions, but their applicability at scale is a concern
- Access (outgoing) to at least a few external subnets is of high importance. In the PIC talk (see later) it was shown how some processing can be executed without access, but it is very limiting.
- The experiments have mechanisms to correctly steer payloads with the correct architecture, and to identify / use accelerators if on board.
- The capability to deploy light site-oriented services at the edge of a HPC has been shown as very handy when allowed (for example, to implement direct submission to SLURM or to implement the pressure model from within the site).

Examples of past integration efforts

PIC-BSC

The integration effort PIC (WLCG Tier-1) and BSC (MareNostrum 4 and 5) has been presented. The major obstacle is the absence of (outgoing) networking from compute nodes.
The attempt has been triggered by a request by the Spanish Funding Agency to pledge at least 50% of LHC pledges using the ErutoHPC center.
Workarounds has been put in place, most notably:

- Services as CEs and custom gateways have been deployed in close-by WLCG centers
- Apptainer containers containing the full sw stack have been pre-placed on BSC storage. This limits the processing capabilities to specific versions of the experiments' software stacks (needed since no CVMFS available)
- Communications compute node - central experiment services are not possible. They have been substituted (for example in the case of CMS HTCondor connections) with file systems calls. This required specific developments deployed together with the HTCondor team.
- Access to conditions data has been operated via ssh tunnels (scaling??)
- Access to input data is limited to those preplaced on BSC storage, pre-filled via a specially deployed transfer service from/to PIC.
- So far the exploitation since 2020 is compatible with the 50% target (wrt to Spanish LHC CPU pledges) for exploitation set for 2024 on
    - Mostly MC simulations. Supporting other workflows with input data foreseen
- The work done to integrate this resource took many FTE efforts from the WLCG Spanish community and from international teams (HTCondor and experiment frameworks)

- As a generic idea, experiments' pilots could be make available in EESSI, while the real payloads could depend on experiment level CVMFS repositories.

CERN-LUMI

LUMI integration is reported by CERN IT (we are aware separate attempts were carried on by the experiments).
Main challenges / solutions reported are:

- No CVMFS – various unprivileged workarounds with caveats:
    - CVMFSexec mode1 (requires fusemount)
        - Must unmountrepo on job kill
        - Attention to cache size and file descriptor limits when sharing
    - CVMFSexec mode3 (requires userns)
        - Similar cache/FD needs
    - singCVMFS / CMVMFSexec mode 4
        - requires setuid singularity or userns
        - Each container creates own cache (by default)
        - Extra steps for cache sharing (alien or ext3 image) • "Fat" containers images
    - Diskless nodes unexpected for some workloads
        - Ramdisk on LUMI is 'RAMSIZE – 32GB' (for system image)
        - above CMVFS caches are all on NFS or ephemeral…
    - File transfer
        - Only SSH-based tooling to standard storage (single socket)
        - Object storage via s3cmd, rclone, among others
    - CERN (old) SSO Incompatible with eduGAIN (resolved 5/2023)
        - Web portal login; SSH requires registering key with MyAccessID IAM

  The last point is particularly interesting since it shows an integrtation path between CERN SSO (and hence WLCG AAI) and LUMI.

Conclusions and how to go on

- It is quite clear there is no space for new requirements for the first phase; still discussions should continue to have eventually some of the features added for the second phase
- HEP would like to be part of the beta testing already in phase 1. While a beta testing group has not been defined yet, there is agreement on the inclusion asap
- It is clear the time from now to the first release is short, and on the EFP side there is no margin for extended discussions on future phases; still there is interest in establishing now "no guaranteed response" communication channels
    - We propose a mailing list HEP+EFP for this purpose
    - Cern egroup?
- There is no user-ready documentation on EFP side, but it will be there by the time of the release. HEP is asking for access to it as soon as available, together with the inclusion with the beta test users

- On the HEP side, we need to report to the WLCG MB and see experiments / nations / sites interest in the collaboration