Access : Authentication and Autorization

Michel Jouvin, IJCLab

Open Science Practices in Nuclear Physics

6 December 2022

michel.jouvin@ijclab.in2p3.fr

Who Am I

- I am probably not well known in the NP community…!
- I have been working at LAL since 1992 with a strong involvement in HEP computing, in particular LHC experiment computing (WLCG)
  - WLCG GDB chair during 4 years (GDB is a forum of world-wide WLCG sites and experiments, with a monthly meeting, a key structure in WLCG technical coordination)
- Currently the head of IJCLab Computing Department (~50 people)
  - Technical coordinator of GRIF, a major WLCG T2 (site), since its inception in 2004
  - Scientific coordinator of VirtualData, a large academic cloud (10000+ cores), operated by IJCLab
- Involved in several projects around computing resource federation
  - MesoNET: French Equipex+ project to build a federation of university HPC centres (mesocentres)
- IJCLab covering NP, HEP, Astro&Cosmo and more: motivated by cross-fertilizing the computing expertise built in the various communities

Open Science/Data Challenges

- Data must be easily accessible by anyone entitled to do it
    - Requires some sort of data federation using standard protocols
    - Need to be able to identify any user unambiguously
- Controlled access: not every data is open the same way
    - Communities must be confident that only the relevant data are open, even if stored at the same facilities as more restricted data
    - Changing who has the right to use a dataset should be easy and controlled by the community
- Long-term data and analysis preservation
    - One of the goal of open data is allow to reuse data from previous experiments in current analysis
    - Long-term (open) data preservation is not only about preserving data but also the capacity to use/analyse it: analysis preservation is somewhat more challenging, even though technologies like containers may help a lot
    - Controlled access is less a problem: on the long term, data generally becomes public

Data Federation Challenges

- Data federation is a requirement
  - No single site has the capacity and funding to host data for a whole community: even in HEP, CERN is not able to do that
  - Data ownership: not always possible/acceptable to delegate to others the responsibility for storing and controlling access to the data
  - Resilience: multiple copies in different sites, possibly in different countries, is the best guarantee against data loss due to a major site incident
  - Cost: allow to consolidate the different funding sources attached to each participating site
- Challenges
  - Large-scale, coordinated, intensive, data transfers: HEP developed an advance expertise and several open-source tools to deal with the data avalanche
  - Shared data facilities: leverage existing infrastructures and expertise at some sites to build multi-disciplinary facilities and optimise the management cost
  - **Access control: a robust and global authentication/authorization infrastructure is needed**

Authentication and Authorization Service

- Foundation for any distributed/federated infrastructure: a global/federated authentication and authorization service
  - Global: from the service, should appear as a global service describing all the community members and their roles that will be used to decide what they are authorized to do
  - Federated: should not be yet-another-user-database with other credentials that users must maintain, instead must rely on standard user identity/credentials from their home institute
  - To allow cross-community relationship, must rely on standard protocols
- Initial attempt was using user certificates but they are too heavy to manage
  - Users generally hate them! Not used for normal activities and home institution authz
  - Creation/Renewal process generally requires some sort of identity vetting that duplicates internal processes of a community
  - Used heavily by High Energy Physics community, in particular LHC experiments

SAML / Shibboleth: federated authentication

- A robust protocol (SAML) used to build identity federations for web applications
  - Not supporting non web applications, in particular CLI-based services, as an http redirect is used to redirect authentication to the home institute
  - Most well-known (academic) federation: eduGAIN.
  - A vast majority of academic sites connected to the eduGAIN federation: potential for web applications around the world to use the same identity/authentication source
- Shibboleth does not help with authorization which is the required feature in a federated infrastructure
  - Enforce consistent behaviour across the infrastructure when the namespace is spread over multiple sites or when data is replicated
  - Shibboleth doesn’t allow to carry any useful information (groups, roles, scopes…) that can be used as input to autorisation decisions
  - SAML allow to propagate this information: used by VOMS service with certificates but not by Shibboleth

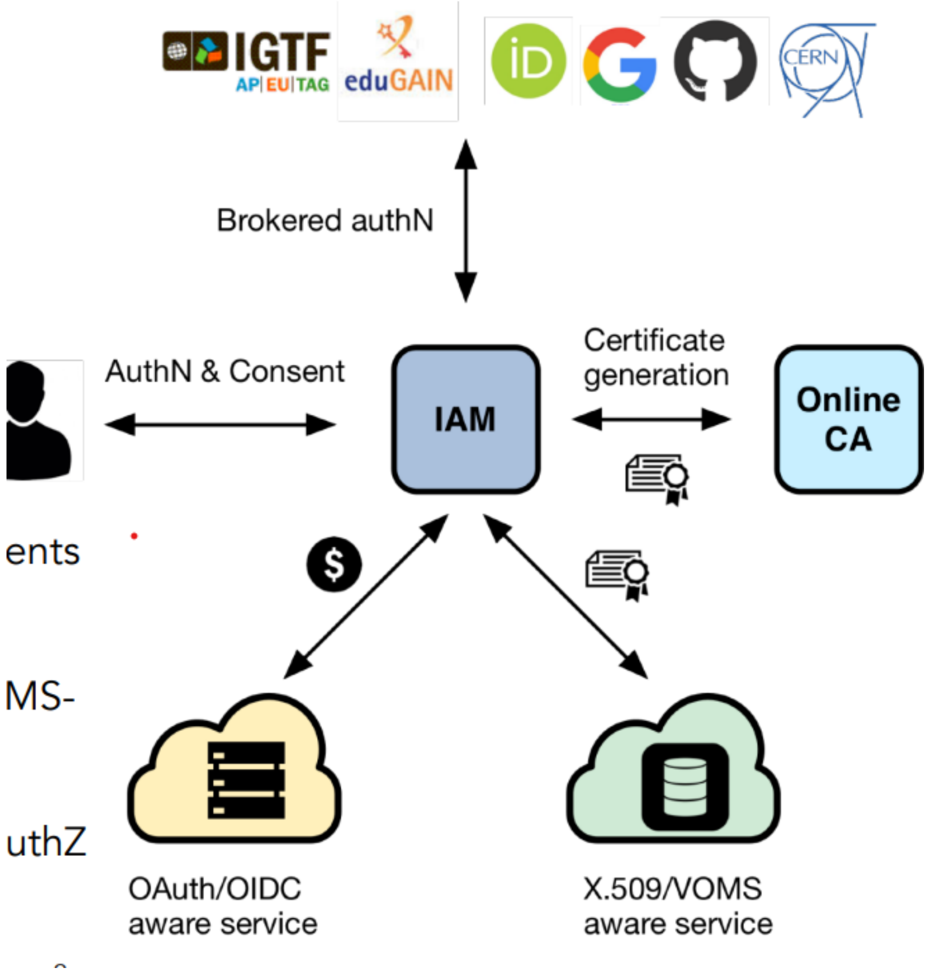
OpenID Connect (OIDC) : Identity and Access

- With the development of internet, the need for global authentication and authorization became a major requirement
  - In particular for interoperation/cooperation between services
  - A technology emerged in the web world : Jason Web Tokens (JWT). JWT are signed pieces of information that can be securely exchanged with a proof of authenticity, on behalf of a user.
- OIDC leverages OAuth2 protocol dedicated to authorization to manage identity information: Identity and Access Management (IAM)
  - Both protocols use JWT
  - OIDC allows to use external sources to verify the identity of a user
    - E.g. ability to authenticate on a service using social logins like Google, Facebook…
  - OAuth2 allows an application to act on behalf of a user without forwarding his credentials
  - Adopted/developed by the main Internet players (e.g. GAFAM)
  - Has the potential to be used with command line tools, not only web applications
  - Much easier to integrate into applications than Shibboleth
- See https://indico.cern.ch/event/1185598/ for more details

Federated IAM: INDIGO IAM service

- Several Identity and Access Management services based on OIDC emerged in the last years
    - Keycloak (Red Hat), EGI Checkin, CILogon OIDC service…
- INDIGO IAM is one of them, adopted by WLCG as its next generation Authentication and Authorization Infrastructure
    - Developed by a European project, INDIGO, about clouds and cloud federations (2014-2018)
    - Support multiple user authentication protocols: SAML, OIDC, X509, user/password
    - Identity brokering: a user can have several identities linked to its IAM account. A single user seen by applications.
    - Group management: users can be assigned to groups that are passed to applications in the JWT for authorization decisions
    - Allow to define scopes for capacity-based, fine-grained, authorization decisions (ex: read right for a specific file namespace/directory)
    - Talk OIDC/OAuth2 with applications (OIDC clients)

Indigo IAM Workflow



IAM : integrating applications and services

- Applications (OIDC clients) must be registered with the IAM server (OIDC issuer)
  - Required to establish the trust as it cannot be derivated from a certificate
- OIDC has been designed with a focus on easy integration with web applications
  - Most web frameworks allow a seamless integration of OIDC authentication
  - Extraction of authorization related information of tickets is the responsibility of the application/services
  - When using an Apache-based application, it is just a matter of adding « Require » statements
- Non-web experiments, in particular CLI-based applications can be integrated too
  - Generally rely on a third-party tool (eg. oidc-agent, Hashicorp Vault/htgettoken, mytoken…) to allow the user to acquire a token (a string) and pass it to the application
- Data management/storage must provide endpoints supporting OIDC
  - Now available with all storage implementations from HEP and storage exposing a S3 endpoint (e.g. Ceph)

INDIGO IAM adoption outside HEP

- Increasing number of projects requiring a federated authorization and authentication and choosing OIDC technology to implement it
  - USA OSG infrastructure moved completely out of certificates/GSI to JWT/OIDC (SciTokens) in Spring 2022
  - EGI, an infrastructure part of EOSC, has also a token-based authentication infrastructure (EGI Checkin)
- A few projects adopted INDIGO IAM as their OIDC solutions in the last years, in addition to WLCG
  - UK IRIS: INDIGO IAM used for supporting various communities of different sizes, including SKA
  - Italy: used for authentication in INFN cloud used by various communities
  - France: MesoNET, a federation of regional mesocentres (HPC facilities mainly)
  - ESCAPE: European project

EURO-LABS Program of Work in this Area

- EURO-LABS is the startup of a medium/long-term effort in NP to build a federated data management infrastructure for open data and open science
    - Main deliverable: provide an Authentication and Authorization Infrastructure (AAI) for NP community/communities enabling to use existing federated data management infrastructure deployed in the context of EOSC
- Will be based on Indigo IAM
- EURO-LABS contribution is about service deployment and operation
    - No service/software development done by the project
    - Service operation is more than software deployment: in particular understand how to reflect the structure of the NP community in Indigo IAM: 1 global IAM with groups, several federated IAM instances…
- IJCLab will be the driver of this work
    - Well connected with the work around Indigo IAM in HEP
    - Already has some experience in running this service for some small communities