eosc | EuroScienceGateway

EuroScienceGateway Kick-Off Meeting Work Package 3

6th-7th October 2022, Freiburg

Funded by
the European Union

Work package 3 - Pulsar Network: Distributed heterogeneous compute

- National HPC and Cloud infrastructures have been established, with differences in
  - Hardware
  - Configuration
  - Software stack
- Access typically targeted at local researchers.
- Different needs for researchers, depending on, for example:
  - Local infrastructure availability and accessibility
  - Sensitivity of the data
  - Experience & skills
- The global pandemic has reshaped the way we look at biological data handling: prompt, straightforward, efficient and structured access to data, tools and workflows supported by suitable IT infrastructures is becoming increasingly critical for researchers.

Work Package 3 - Pulsar Network

Galaxy EUROPE

PULSAR

A User friendly interface to workflows, tools and compute and storage resources:
-> The Galaxy Project and UseGalaxy.eu

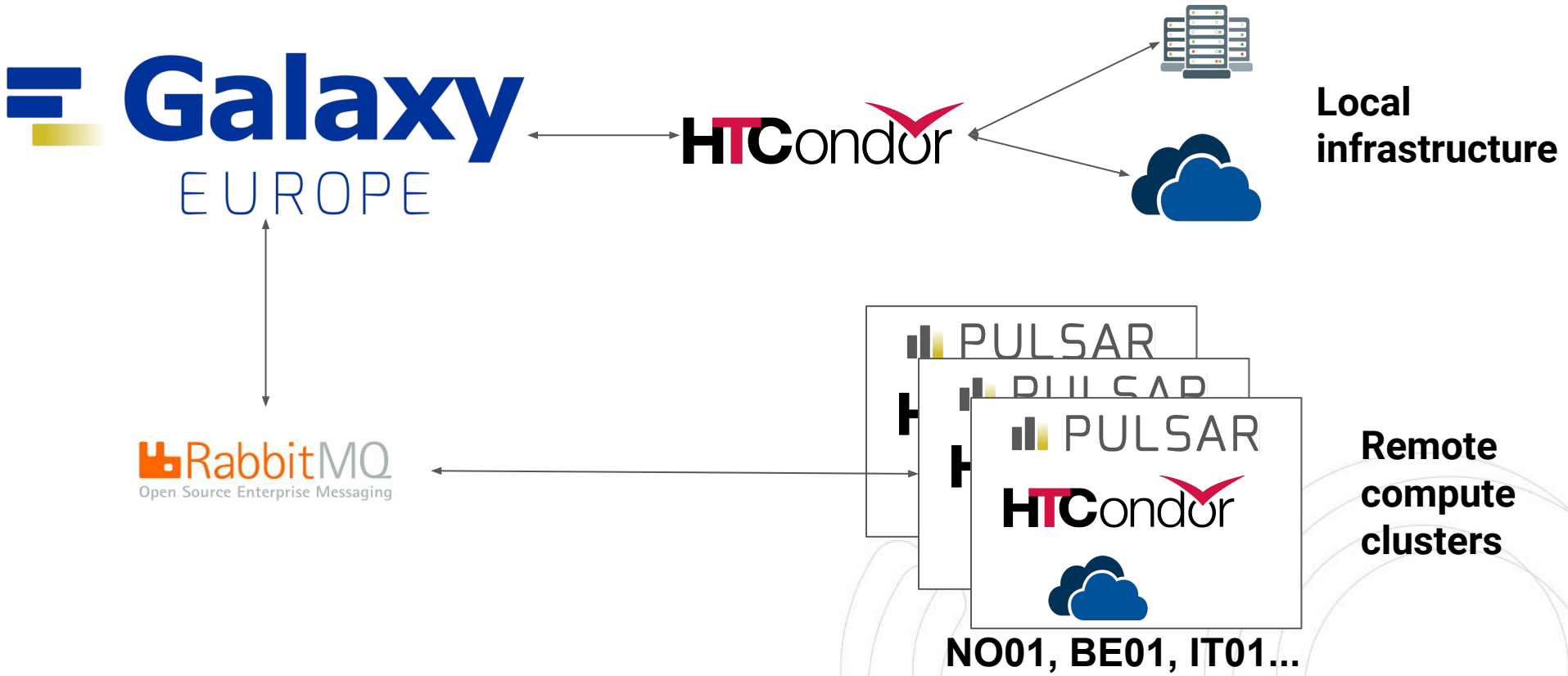
Grant users the access to Compute Infrastructures, regardless of the underlying infrastructure:
-> Pulsar

- the Galaxy Project's remote job execution system.
- It is a Python server application that accepts jobs from a Galaxy server, submitting them to a local resource and then sending the results back to the originating Galaxy server.
- Support for different resource managers (HTCondor, SLURM, K8s).

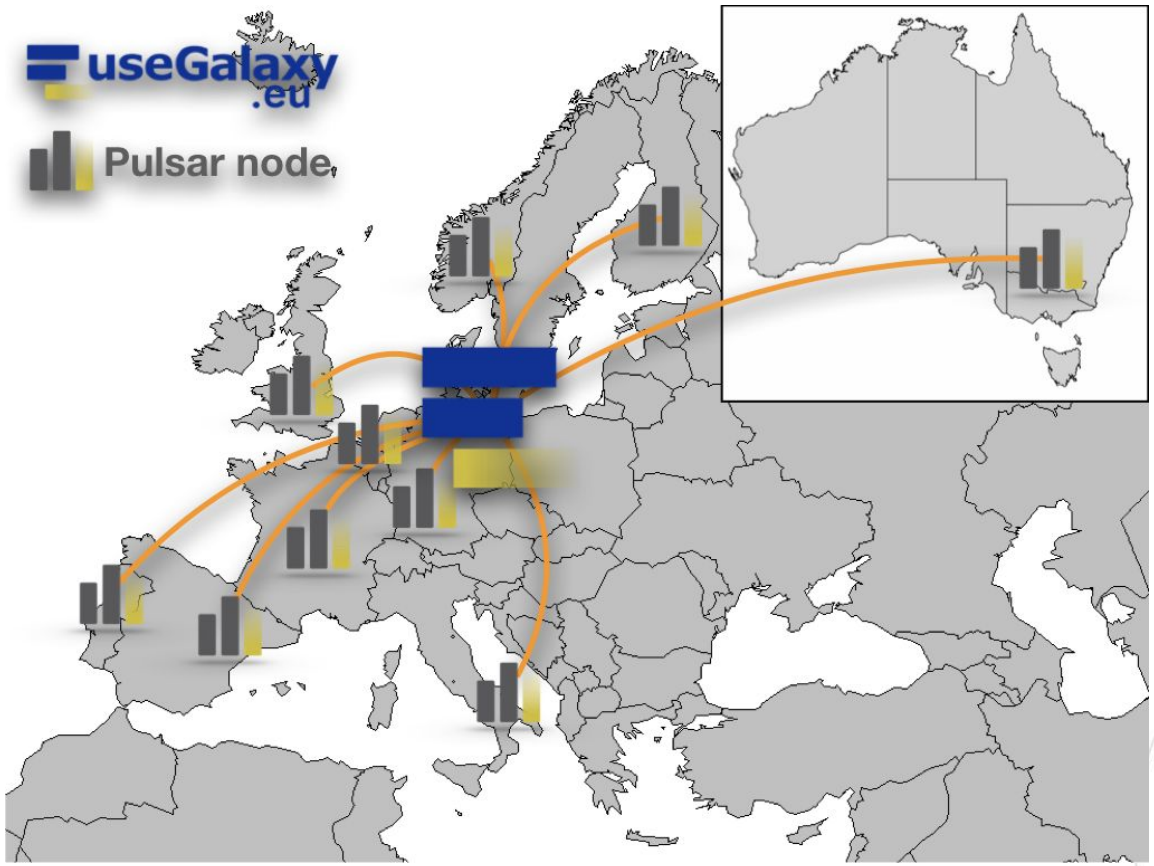
CernVM File system

Grant access to Tools and Reference data:
-> CernVM File System: distributed read-only file system.

Work Package 3 - Pulsar Network



Work Package 3 - Pulsar Network



https://pulsar-network.readthedocs.io

Objectives

What the WP is planning to achieve?

O3.1 - Build an European wide job-scheduling network

- T3.1, T3.3, T3.4 and T3.5

O3.2 - Make Pulsar endpoints conform to GA4GH Task Execution service standard

- T3.2

O3.3 - Deploying a TRL-9 web service to access the Pulsar Network

- T3.2 and T3.5

Objectives

What the WP is planning to achieve?

- At least 10 Pulsar endpoints, routing the incoming jobs from Galaxy and other workflow management systems to local compute resources.
- 6 national Galaxy instances that will make use of the Pulsar Network

Objectives - Task 3.1

How are we planning to achieve the objectives?

Develop and maintain an Open Infrastructure based deployment model for Pulsar endpoints **(M1-M36)**

Task Lead: INFN

Task Members: ALU-FR, CESNET, CNR, IISAS

Goals:

- **Extend the Open Infrastructure for the Pulsar Network deployment.**
- Further extend to AWS, Azure and Google cloud and container orchestrator (k8s).
- Include EOSC-compliant AAI to facilitate integration with other services.

Status:

- documentation: https://pulsar-network.readthedocs.io
- github: https://github.com/usegalaxy-eu/pulsar-infrastructure
- Ansible roles, terraform recipes and documentation already available.

Objectives - Task 3.1

How are we planning to achieve the objectives?

Open infrastructure:

- set of tools to have a ready-to-go Pulsar environment easily deployable into a cloud infrastructure;
- enable consortium partners (and beyond) to deploy new pulsar nodes to further extend the computing capacity of the network.

Objectives - Task 3.1

How are we planning to achieve the objectives?

- **A virtual machine image, named Virtual Galaxy Compute Nodes (VGCN), that provides everything is needed to run Galaxy jobs.**
- Terraform scripts that take care of the infrastructure deployment over the Cloud resources
- Ansible scripts to complete the Pulsar's configuration and have then an easy mechanism for its update.

Objectives - Task 3.1

How are we planning to achieve the objectives?

- A virtual machine image, named Virtual Galaxy Compute Nodes (**VGCN**), that provides everything is needed to run Galaxy jobs.
- **Terraform scripts that take care of the infrastructure deployment over the Cloud resources**
- Ansible scripts to complete the Pulsar's configuration and have then an easy mechanism for its update.

Terraform is a software for creating and managing virtual infrastructures by exploiting machine-readable configuration files.

Objectives - Task 3.1

How are we planning to achieve the objectives?

- A virtual machine image, named Virtual Galaxy Compute Nodes (**VGCN**), that provides everything is needed to run Galaxy jobs.
- Terraform scripts that take care of the infrastructure deployment over the Cloud resources
- **Ansible scripts to complete the Pulsar's configuration and have then an easy mechanism for its update.**

Ansible is an open-source software that automates cloud configuration management, application deployment and service orchestration.

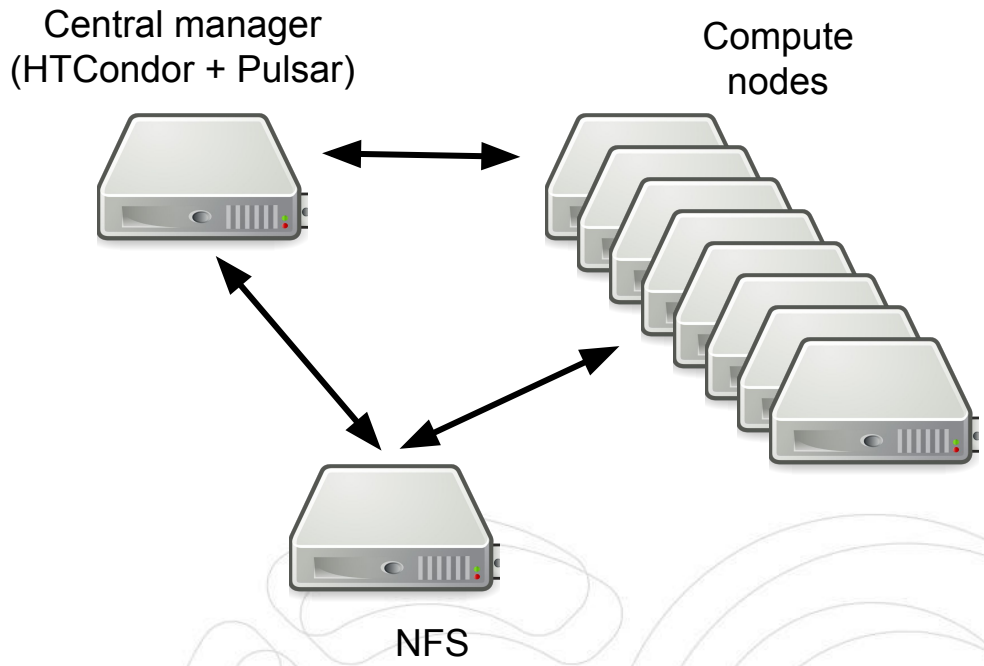
Objectives - Task 3.1

How are we planning to achieve the objectives?

For a prototype setup, the minimum requirements are:

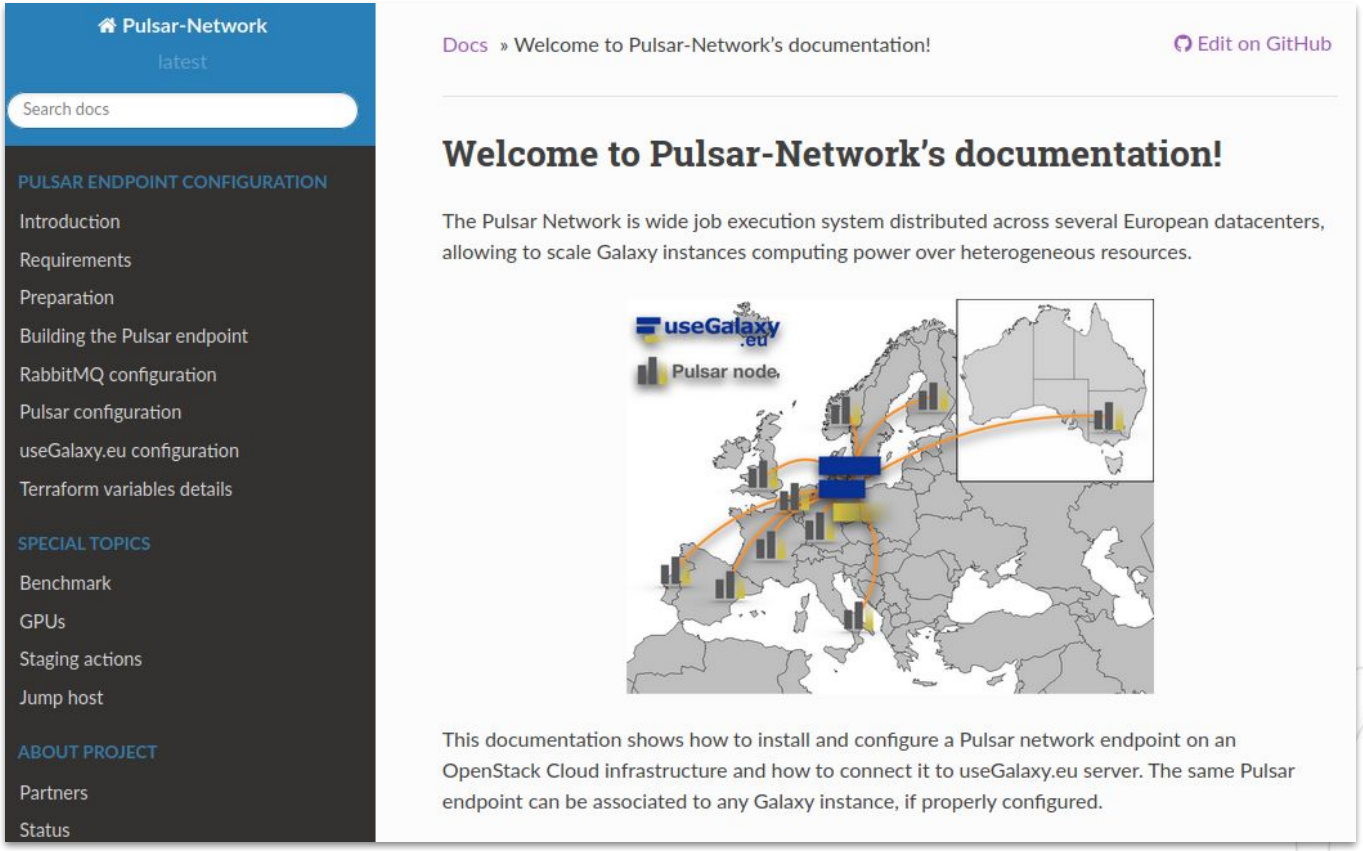
- **Central manager and NFS server** each with 4 cores, 8 GB
- **Computational workers** each with 4-8 cores, 16 GB
- **>200 GB** volume

but the more the better



Objectives - Task 3.1

How are we planning to achieve the objectives?



https://pulsar-network.readthedocs.io/en/latest/

Objectives - Task 3.2

How are we planning to achieve the objectives?

Add the GA4GH Task-Execution-Service (TES) API to Pulsar **(M1-M12)**

Task Lead: CESNET

Task Members: ALU-FR, CNR

Goals:

- Implement support for the GA4GH Task Execution Service, allowing other services to submits jobs via TES to Pulsar and to the European Pulsar Network.

Status:

- TES spec: https://github.com/ga4gh/task-execution-schemas

Objectives - Task 3.2

How are we planning to achieve the objectives?

- The Task Execution Service (TES) API is an effort to define a standardized schema and API for describing batch execution tasks. A task defines a set of input files, a set of (Docker) containers and commands to run, a set of output files, and some other logging and metadata.
  https://ga4gh.github.io/task-execution-schemas/docs/

- Proof-of-concept TESP API: https://github.com/ndopj/tesp-api
  A separate microservice, decoupled from the Pulsar, implementing the TES standard and distributing TES tasks to Pulsar applications (currently using Pulsar REST API).

Objectives - Task 3.3

How are we planning to achieve the objectives?

Build an European-wide network of Pulsar sites **(M7-M36)**

Task Lead: CESNET

Task Members: ALU-FR, VIB, EPFL, CESNET, BSC, CNRS, CNR, INFN, UiO, AGH / AGH-UST. IIAS, TUBITAK

Goals:

- Deploy and maintain pulsar endpoints

Status:

- documentation: https://pulsar-network.readthedocs.io
- github: https://github.com/usegalaxy-eu/pulsar-infrastructure
- Several Pulsar endpoints are already online.

Objectives - Task 3.4

How are we planning to achieve the objectives?

Add TES support to WfExS (Workflow Execution Service) **(M18-M36)**

Task Lead: BSC

Task Members: UNIMAN

Goals:

- Extend WfExS to support ESG as compute platform
- execute task on the Pulsar Network using TES API developed in T3.2

Status:

- Github: https://github.com/inab/WfExS-backend

Objectives - Task 3.4

How are we planning to achieve the objectives?

WfExS is a high-level workflow execution service backend, developed within EOSC-Life as part of Demonstrator 7 (D7), which can manage workflows across different domains.

It has a strong focus on reproducible and replicable analysis by using digital objects like RO-Crate.

- Fetches workflows from WorkflowHub.
- identifies the workflow type and run it using its native workflow execution engine (currently CWL and NextFlow).
- Identifies the containers needed by the workflow and fetches them.
- Optionally describes the results with a RO-Crate and makes them available to users.

Objectives - Task 3.5

How are we planning to achieve the objectives?

Developing and maintaining national or domain-driven Galaxy servers **(M1-M36)**

Task Lead: VIB

Task Members: ALU-FR, UiO, UB, CNRS, CNR

Goals:

- Develop and maintain an Open Infrastructure for deploying National Galaxy instances.
- Deploy National Galaxy instances to access local infrastructure and the Pulsar Network.
- User support

Status:

- Github: https://github.com/usegalaxy-eu
- Ansible roles and terraform recipes available. Some useGalaxy national instances (Belgium, France) already up and running.

Objectives - Task 3.5

How are we planning to achieve the objectives?

GitHub

- Continuous testing
- Continuous Deployment

Jenkins

5 contributors

219 lines (207 sloc) | 13.7 KB Raw Blame

```yaml
---
- name: UseGalaxy.eu
  hosts: sn06
  become: true
  become_user: root
  vars:
    # The full internal name.
    hostname: sn06.galaxyproject.eu
    # The nginx user needed into the galaxyproject.nginx role
    nginx_conf_user: galaxy
    # This server has multiple CNAMEs that are important. Additionally it
    # provides proxying for many of the other services run by Galaxy Europe.
    # These server_names are passed to certbot. They generally should not need
    # to be updated unless you add a new domain. They *only* work with the
    # route53 provider, so if we want to do usegalaxy.xy, it may require
    # refactoring / multiple certbot runs.
    #
    #
    # The best way to expand them is to run the playbook, it will leave a message with the command it would have run (look for `skipped, since /etc/letsencrypt/
    #
    # Then take this command to the command line (root@sn04) and run it with `--expand`. E.g. (DO NOT COPY PASTE (in case the config changes))
    #
    # $ /opt/certbot/bin/certbot certonly --non-interactive --dns-route53 \
    #     -m security@usegalaxy.eu --agree-tos -d 'usegalaxy.eu,*.usegalaxy.eu,galaxyproject.eu,*.galaxyproject.eu,*.interactivetoolentrypoint.interactivetool.u
    # Saving debug log to /var/log/letsencrypt/letsencrypt.log
    # Credentials found in config file: ~/.aws/config
    # ....
    # IMPORTANT NOTES:
    #  - Congratulations! Your certificate and chain have been saved at:
```

- usegalaxy-eu / **vgcn**
- usegalaxy-eu / **vgcn-infrastructure**
- usegalaxy-eu / **infrastructure-playbook**
- usegalaxy-eu / **cvmfs-example**
- usegalaxy-eu / **usegalaxy-eu-tools**
- usegalaxy-eu / **infrastructure**
- usegalaxy-eu / **workflow-testing**

HashiCorp Packer

HashiCorp Terraform

ANSIBLE

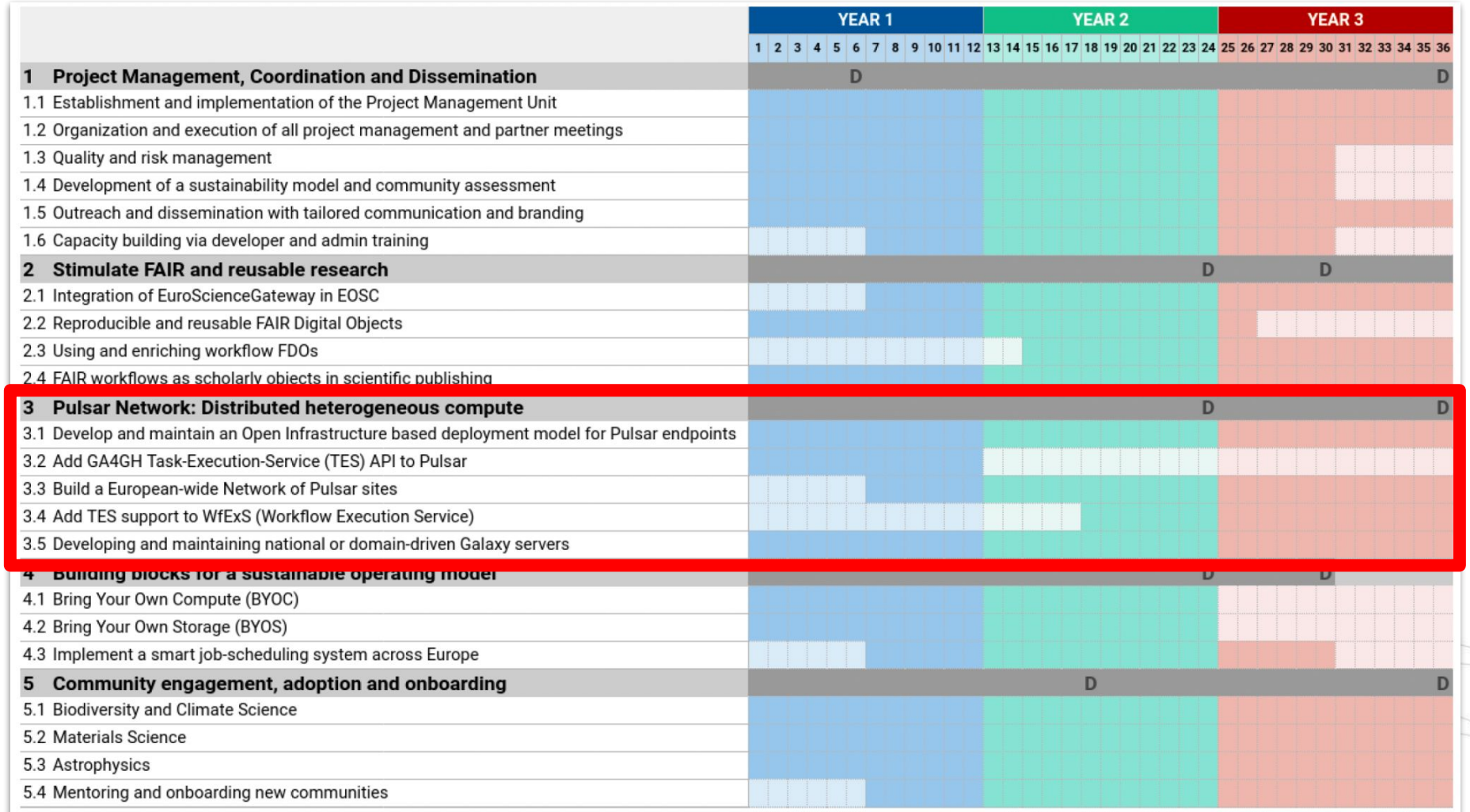


Figure 3.1a Gantt chart with overview of work packages, task durations and main deliverable deadlines.

Deliverables and Milestones

D3.1	Operations documentation on the Open Infrastructure deployment	WP3	INFN	R	PU	M24
D3.2	Publication on the Pulsar Network, integrated in workflow management systems	WP3	CNR	R	PU	M36

M3.1	Pulsar network is TRL-9: operational in environment	WP3	M36	Service available
M3.2	Demonstrated job submission via the WfExS to the Pulsar Network	WP3	M36	Service available
M3.3	National Galaxy servers reaching TRL-9 (operational in environment)	WP3	M36	Service available

Connection to the other Work Packages

How can we work together?

Work Package 4

- BYOC Development -> usage of the Open Infrastructure to deploy new pulsar endpoint.
- BYOS Development -> mechanism for data locality development based on a caching layer, tracking which Pulsar endpoint has a specific dataset already available.
- Smart job-scheduling system development.

Work Package 5 -Use cases work package.

Conclusions & next steps

Open questions

- We plan to move the Pulsar Network from TRL-7 to TRL-9 by expanding the APIs, hardening the deployments already available and deploying new ones.
- The Pulsar Network will become a production-ready interface to European computing resources.
- National Galaxy instances across Europe and other workflow management systems will be enabled to submit jobs to this distributed compute network.

Work Package 3 planning meeting held on 6th of September.
Planning 1 WP3 monthly meeting.
Kick start Task meeting this month for T3.1, T3.2 and T3.3.

This project was funded by the European Union's HORIZON-INFRA-2021-EOSC-01, under the Grant Agreement number 101057388.

Backup

Pulsar Network - current contributors

Pulsar endpoints:

- DE, de.NBI cloud
- IT, ReCaS-Bari
- BE, Vlaams Supercomputer Centrum (VSC)
- PT, Tecnico Lisboa
- ES, Barcelona Supercomp. Center (INB-BSC)
- NO, University of Bergen
- CZ, CESNET
- FI, CSC
- UK, Diamond Light Source
- FR, GenOuest