GA No 777413

BC2DC Workshop, eScience 2019, San Diego, CA

DARE: A Reflective Platform Designed to Enable Agile Data-Driven Research on the Cloud

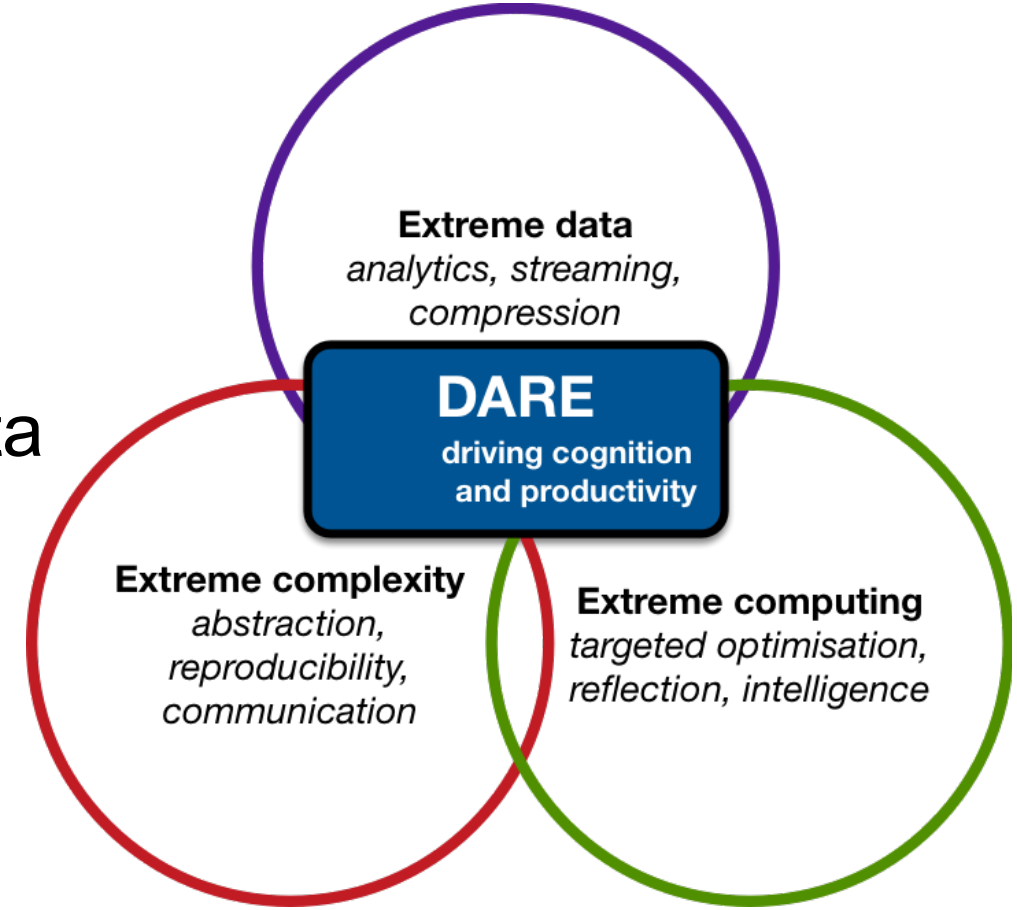
Iraklis Klampanos, NCSR “Demokritos”

Iraklis Klampanos*, Athanasios Davvetas*, André Gemünd†, Malcolm Atkinson‡, Antonis Koukourikos*, Rosa Filgueira§, Amrey Krause§, Alessandro Spinuso¶, Angelos Charalambidis*, Federica Magnoni‖, Emanuele Casarotti‖, Christian Pagé**, Mike Lindner††, Andreas Ikonomopoulos‡‡ and Vangelis Karkaletsis*

Modern data-driven science

- Increasing diversity and distribution of resources
  - Data, computation
- Increasing complexity of methods
- Increasing scale and complexity of data

- Research developers and scientists find it difficult to cope



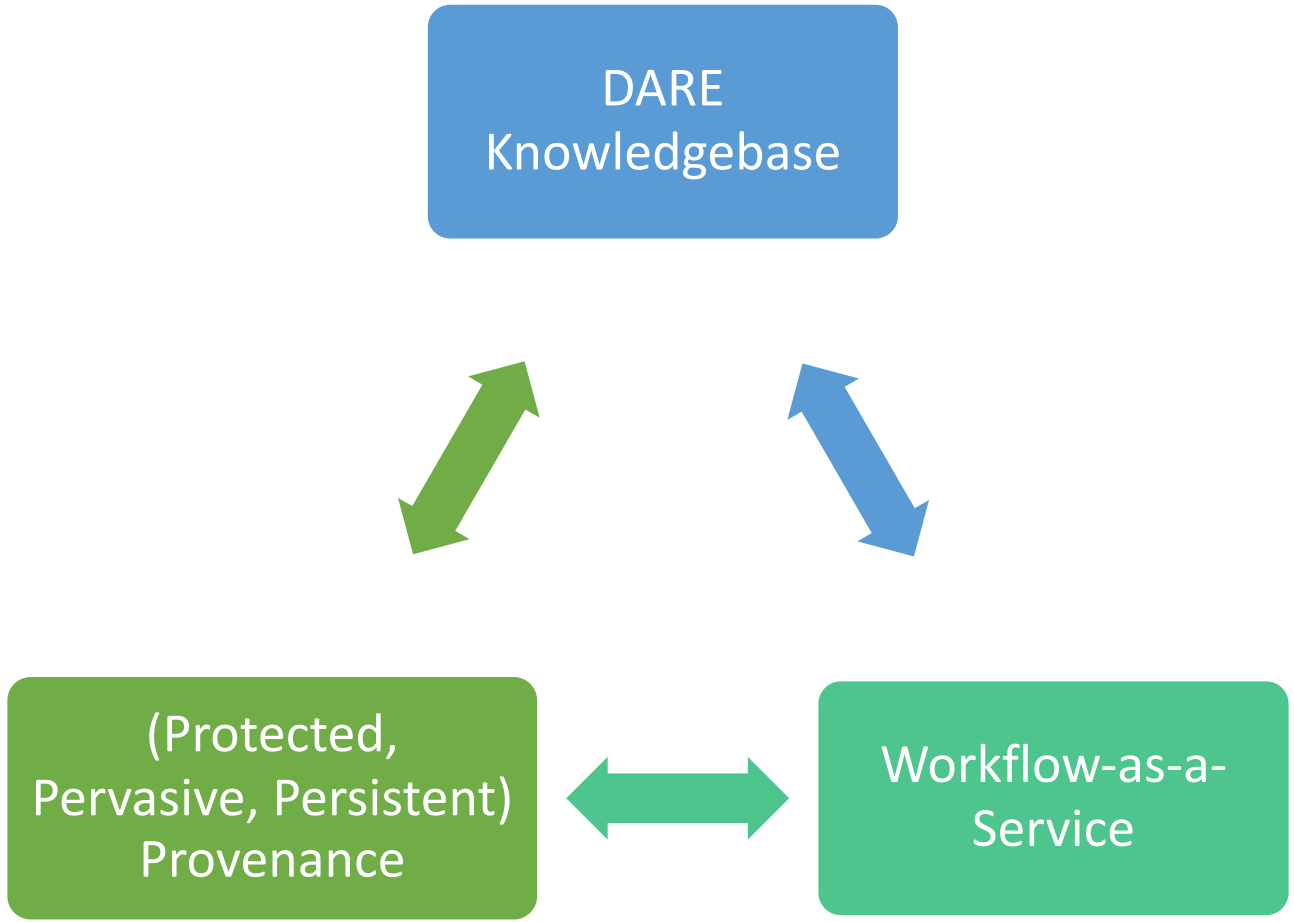
The DARE platform

- A cloud-ready platform for enabling data-driven science
- Provide research developers and scientist with tools to
  - Create powerful workflows in abstract terms
  - Describe and search for semantic properties of methods and data
  - Monitor the execution of parts of workflows
  - Track, query and exchange data provenance records
- Hiding technical detail
- Allow for the use of multiple underlying resources
  - Also of different types

The European Open Science Cloud (EOSC)

- A federation of core cloud-based services
- And infrastructures
- Relationships with RIs, HPC and national stakeholders
- Work-in-progress
- However, consensus appears to start forming:
  - Issuing of PIDs
  - AAI
  - Core services

DARE core components

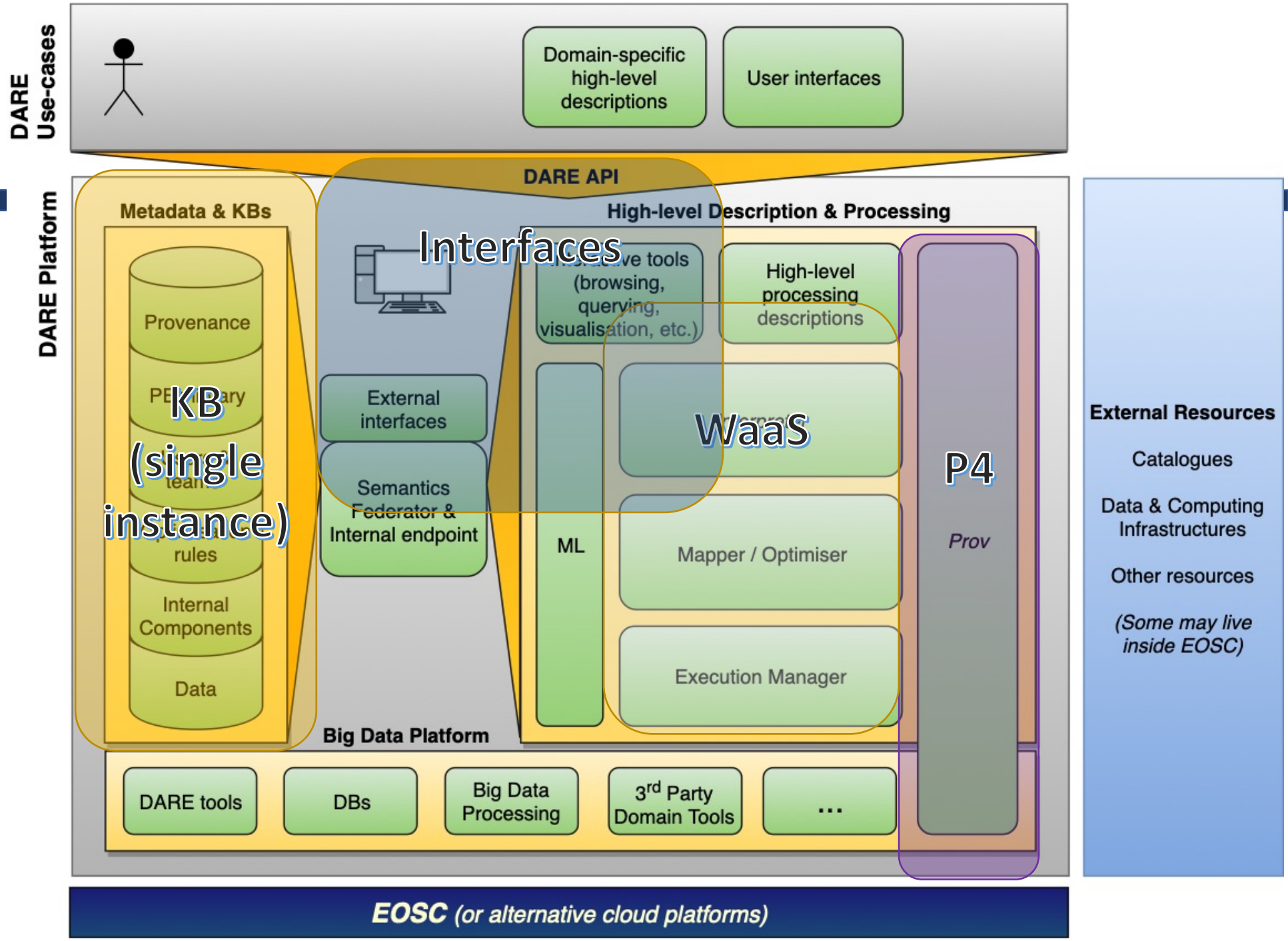


- Interfaces with the world via:
  - Tools and UIs
  - APIs
- Potentially serving
  - Humans
    - Research developers
    - Scientists
    - Data scientists
    - Policy makers
    - …
  - Systems
    - Optimisation
    - Federation
    - …

M. Atkinson et. al. "Comprehensible Control for Researchers and Developers facing Data Challenges", eScience 2019

Platform overview

- DARE KB
  - Catalogues, metadata and semantics
- Optimising WaaS
- Data provenance
- Integrator of big data tools
- RESTful API

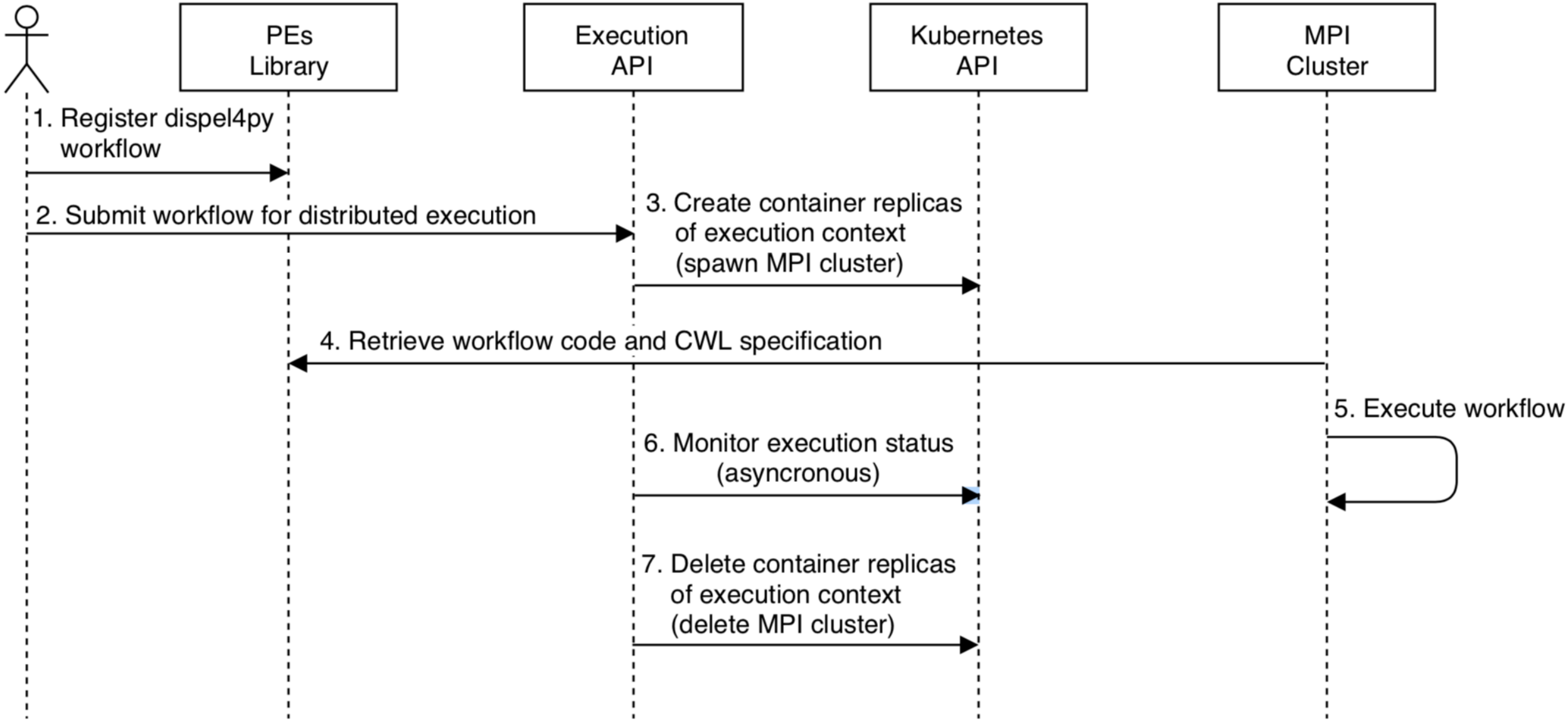


Platform components as microservices

- Each component has its own RESTful API
- Some resources are for use by other DARE components only
- Some are exposed to the outside
- The set of all resources exposed to the outside world forms the DARE API

- Decouple individual components
- Good fit for k8s deployment, scaling and management of containerized applications

Workflow execution on MPI cluster on the cloud



Concepts and semantics

- Concepts re methods, data, past runs, etc. are described in the DARE KB
- The KB is the point of reference within the platform
- PEs and workflows can be identified and described
- Data provenance can be used to draw insights about the use of methods and e-infrastructures
- The data and components catalogues will allow users to abstract away from data and software details in the near future
- A graph of concepts pertaining to an experiment can be extracted and communicated (soon)

Next steps

- DARE platform to be packaged in a release within the week.

- By April 2020:
  - Implement EOSC AAI
  - Improve data and components catalogues
  - Link KB components not currently linked
  - Improve optimization of WaaS
  - Deploy DARE on EOSC-ready resources

Related work

- Several platforms and ongoing projects to add value to EOSC
- Currently ongoing platform-driven projects:
  - *PROCESS*: services and tools to enable extreme scale data processing (ogranised the EINFRA workshop this morning)
  - *DEEP-Hybrid-DataCloud*: ML and DL as-a-service
  - *eXtreme-DataCloud*: unifying, lower level data acquisition and transformation services
  - *EUXDAT*: tools for man- aging extremely large datasets
- Often significant overlap
  - Moves towards reusing technologies and services

More information

- M. Atkinson et. al. “Comprehensible Control for Researchers and Developers facing Data Challenges”, eScience 2019
- http://project-dare.eu
- https://www.eosc-portal.eu

Thank you!

Questions?