GEF status

- May 2016 -

Emanuel Dima, Universität Tübingen

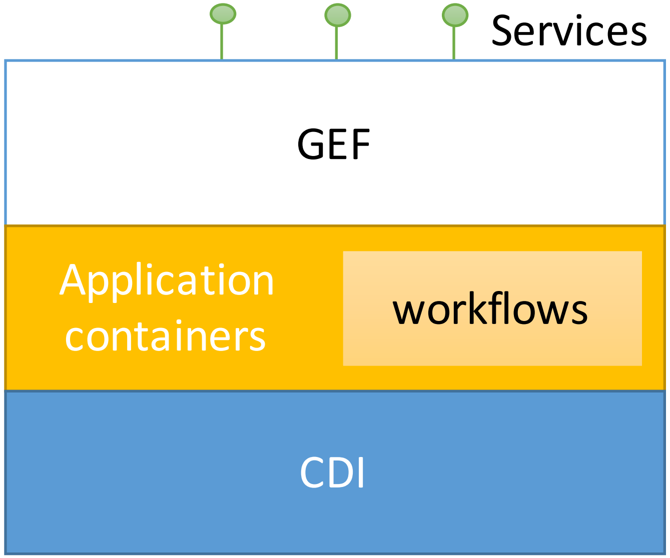
History

- Started in EUDAT/WP7 as research activity
  - “Generic Execution Framework”
  - Initial idea of Christian Pagé: filters for large data sets
  - Extended to the idea of movable computation
  - Results: design document, proof-of-concept
- Continued in EUDAT2020:
  - WP5: service development
  - WP8: research for possible extensions

Argument

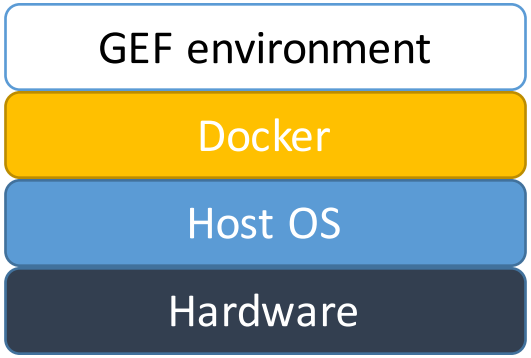
- Datasets have become much larger than the tools
  - It’s more efficient to move the tools than the data, if the tools don’t require intensive processing power
- Some computations are more efficient to be enacted close to the data, thus minimizing data transfers
- Solution: pack computation into movable containers

don’t reinvent the wheel,
use available technology



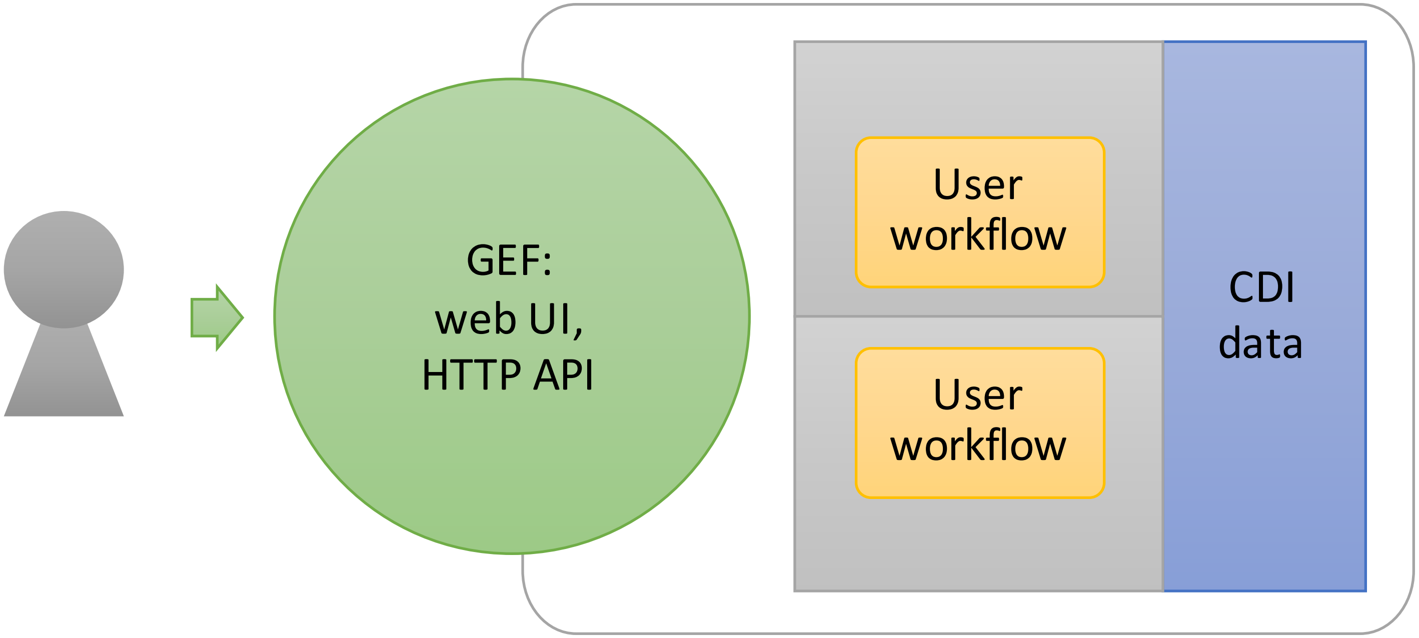
Application containers

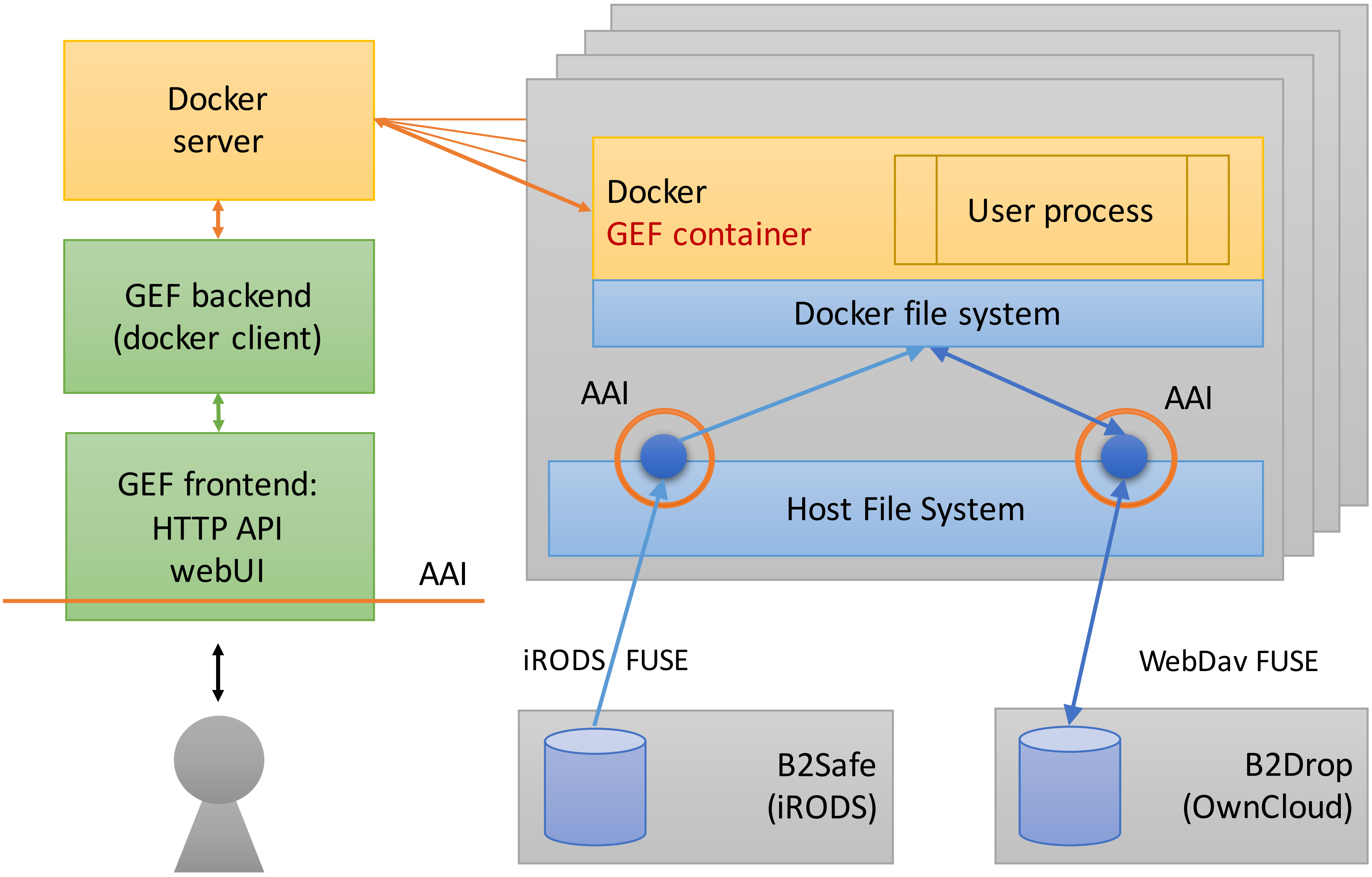
- A virtualization mechanism at the app level
  - Docker: a platform for managing containers
- Very good performance
  - Negligible overhead
  - Very fast to start (low latency)
- Efficient container images
  - Small image size
  - Immutable by design



GEF overview

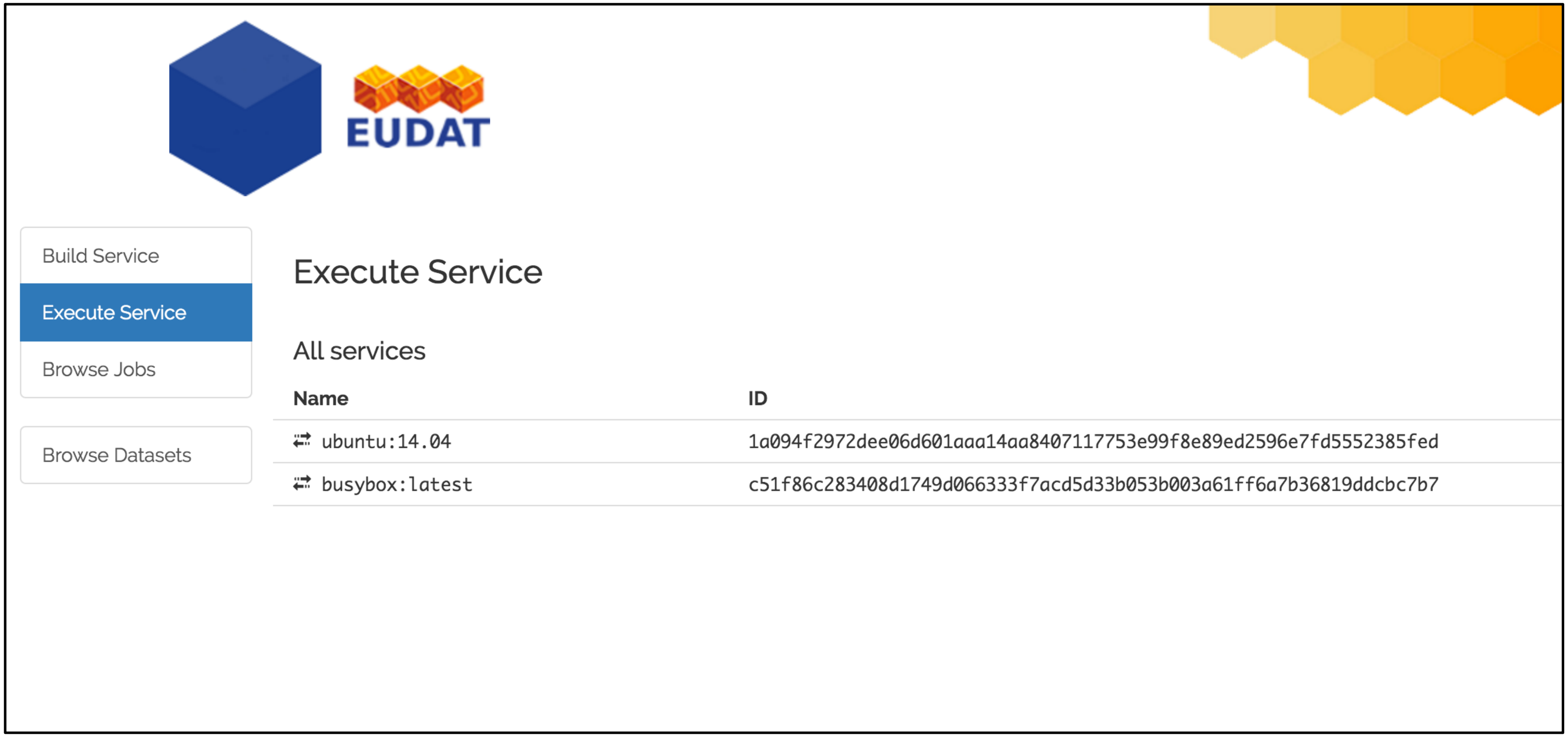
- Create-Your-Own-Service: upload-able services, as container images, containing off-the-shelf tools
- A container image becomes a GEF service and can be invoked on any EUDAT CDI data set





GEF overview

Demo



Execute Service

All services

Name	ID
⇄ ubuntu:14.04	1a094f2972dee06d601aaa14aa8407117753e99f8e89ed2596e7fd5552385fed
⇄ busybox:latest	c51f86c283408d1749d066333f7acd5d33b053b003a61ff6a7b36819ddcbc7b7

The purpose of the demo is to show what the user experience can look like.
Some things are hardcoded for this purpose, e.g. user access to irods/b2drop.

Community use cases

- Extracting subsets of data from large data sets
  - climate science
- Filtering out data from highly structured datasets
  - computational linguistics: searching in treebanks
- Annotating data
  - running WebLicht tools
- Reproducibility of scientific results:
  - Immutable, archived, documented, processing tasks
  - Automatic creation of provenance data

Extending CDI services

- Plugin system for some CDI services (B2SHARE)
- Browsing into datasets
    - Compressed files, archives
    - HDF5 files
- Automatic metadata extraction
    - Community specific file formats
- PIDs for dynamically generated data?

Future

- WP5 development
  - New team!
  - UI, container introspection, HTTP APIs, WPS?, etc.
  - Docker volume plugins
- WP8 research:
  - EGI, workflows (myExperiment, ...), provenance, etc.
- Roadmap?
  - fragmented time spread among several people
  - new, relatively unknown, technologies

Summary

GEF

- executing off-the-shelf tools
- close to the data
- via Docker containers

Usefulness

- community maintained services integrated into the CDI
- plugin framework
- efficiency, reproducibility, provenance