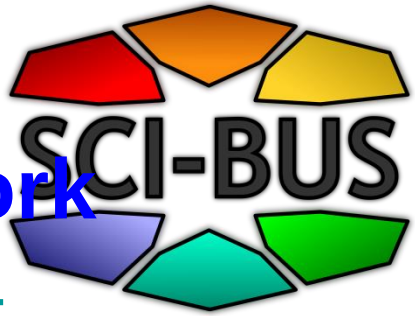
# Workflow applications on EGI with WS-PGRADE

**Peter Kacsuk and Zoltan Farkas**
**MTA SZTAKI**
**kacsuk@sztaki.hu**

# WS-PGRADE/gUSE
# Generic-purpose gateway framework

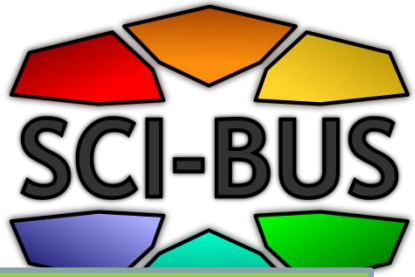



- Based on Liferay
- General purpose
- **Workflow-oriented** gateway framework
- Supports the development and execution of workflow-based applications
- Supports the fast development of domain-specific gateways by a customization technology
- Most important design aspect is **flexibility**

# WS-PGRADE/gUSE Architecture


| | | | |
|---|---|---|---|
| VizIVO gateway | Proteomics Gateway | MoSGrid Gateway | *Application specific gateways (more than 30)* |
| Workflow Editor | Workflow execution Monitor | Data Avenue UI | *Web user interface (WS-PGRADE)* |
| Workflow Management | Workflow Repository | Internal Storages | *Workflow and internal storage services (gUSE)* |
| DCI Bridge | | Data Avenue | *High-level e-infrastructure middleware (gUSE)* |
| HTC Infrastructures | HPC Infrastructures | Large variety of data storages | *Production e-infrastructures* |

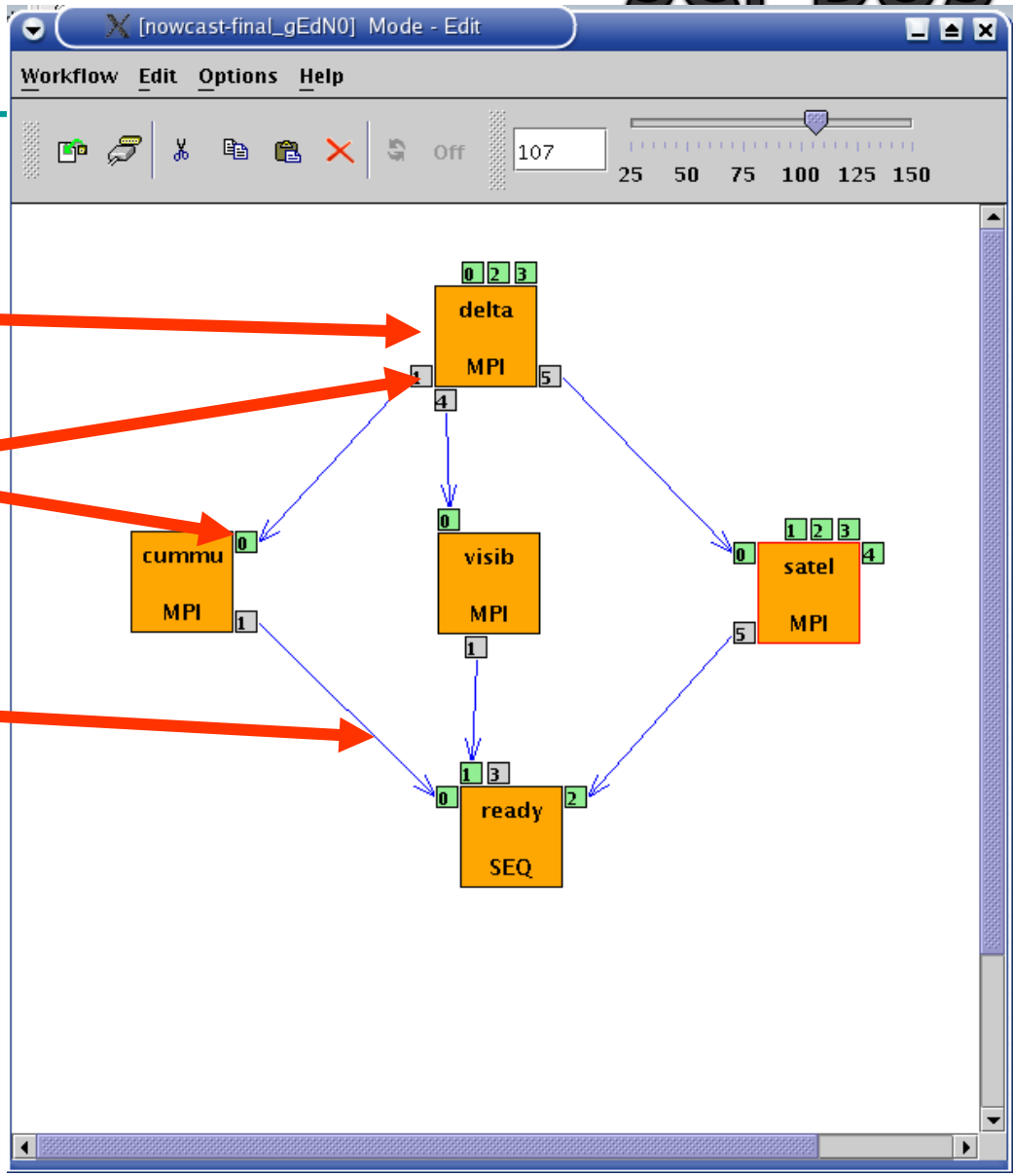
# Flexibility in workflow parallelism



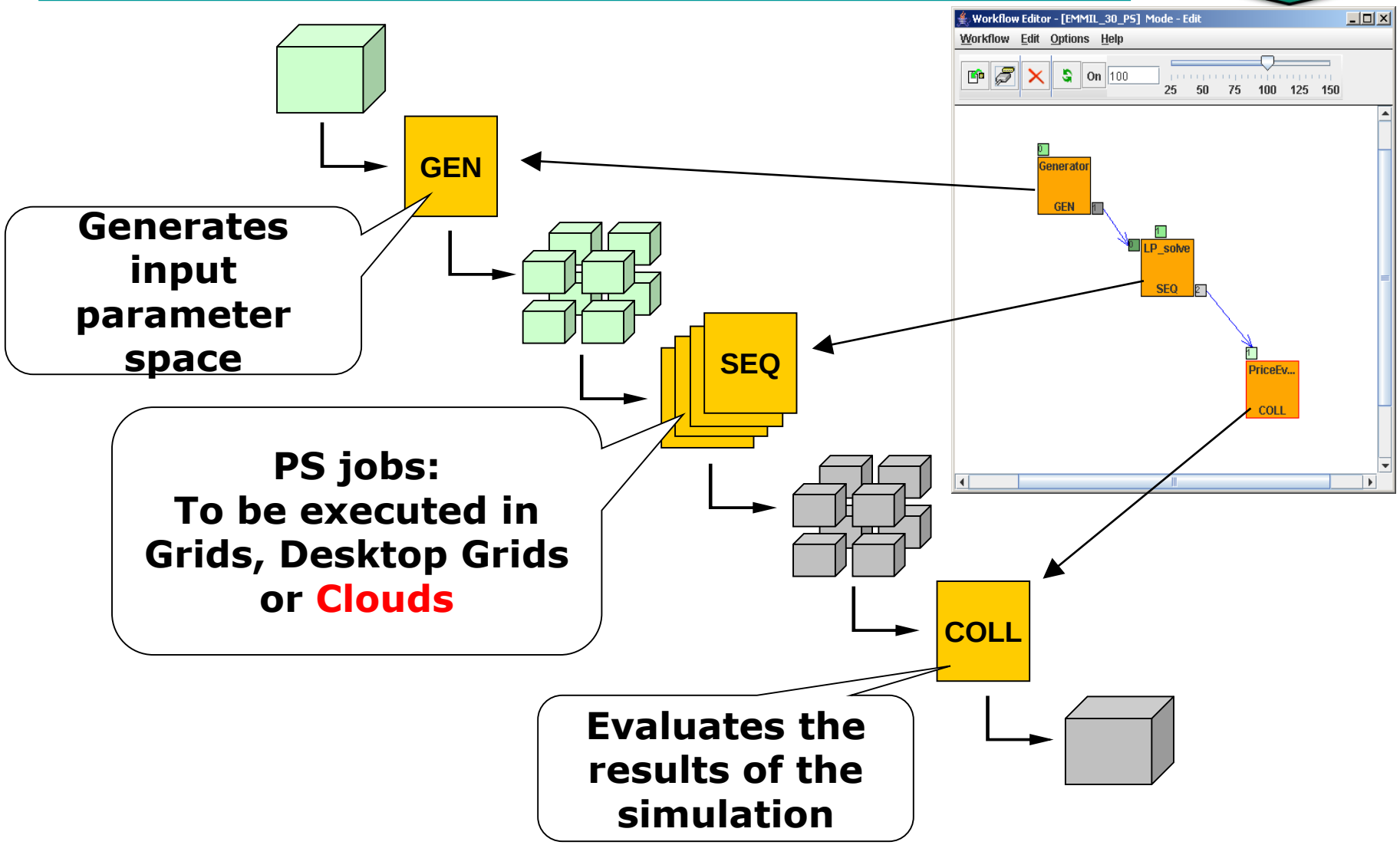

| | | | |
|---|---|---|---|
| VizIVO gateway | Proteomics Gateway | MoSGrid Gateway | *Application specific gateways (more than 30)* |
| Workflow Editor | Workflow execution Monitor | Data Avenue UI | *Web user interface (WS-PGRADE)* |
| Workflow Management | Workflow Repository | Internal Storages | *Workflow and internal storage services (gUSE)* |
| DCI Bridge | | Data Avenue | *High-level e-infrastructure middleware (gUSE)* |
| HTC Infrastructures | HPC Infrastructures | Large variety of data storages | *Production e-infrastructures* |

# Graphical workflow language

- **A directed acyclic graph where**
  - Nodes represent **jobs, services, embedded workflows**
  - Ports represent input/output files the jobs/services expect/produce
  - Arcs represent file transfer operations and job dependencies
- **semantics of the workflow:**
  - A node can fire, i.e. its job can be executed if all of its input files are available



# Parameter Sweep (PS) application workflow template



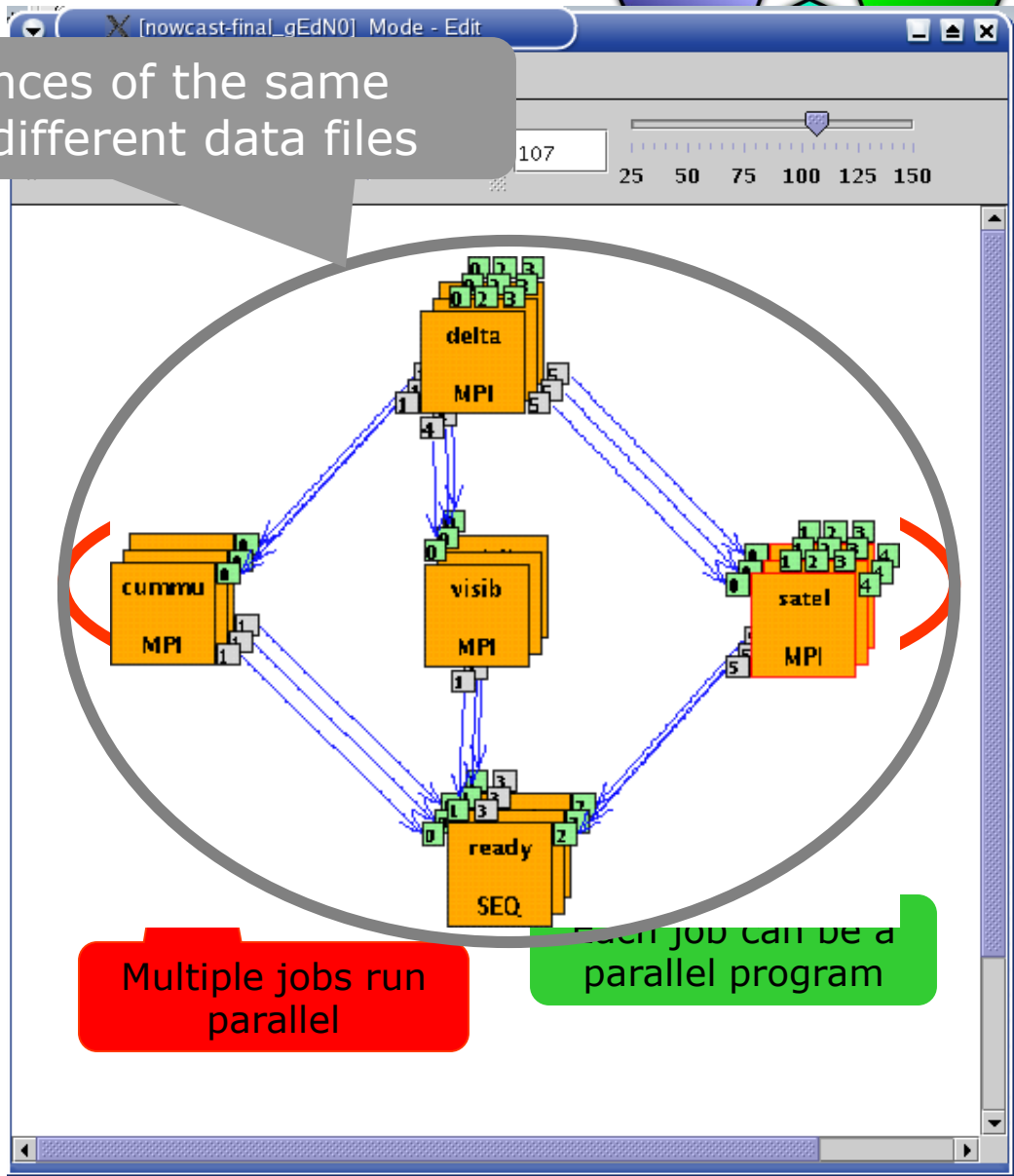
# Flexibility in exploiting parallelism

Multiple instances of the same workflow with different data files

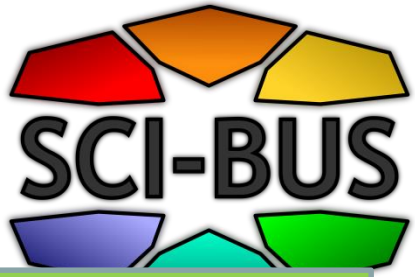
– Parallel execution inside a workflow node

– Parallel execution among workflow nodes

– Parameter study execution of the workflow



# Flexibility in using various DCIs




| | | | |
|---|---|---|---|
| VizIVO gateway | Proteomics Gateway | MoSGrid Gateway | *Application specific gateways (more than 30)* |
| Workflow Editor | Workflow execution Monitor | Data Avenue UI | *Web user interface (WS-PGRADE)* |
| Workflow Management | Workflow Repository | Internal Storages | *Workflow and internal storage services (gUSE)* |
| DCI Bridge | | Data Avenue | *High-level e-infrastructure middleware (gUSE)* |
| HTC Infrastructures | HPC Infrastructures | Large variety of data storages | *Production e-infrastructures* |

# Flexibility in using various DCIs

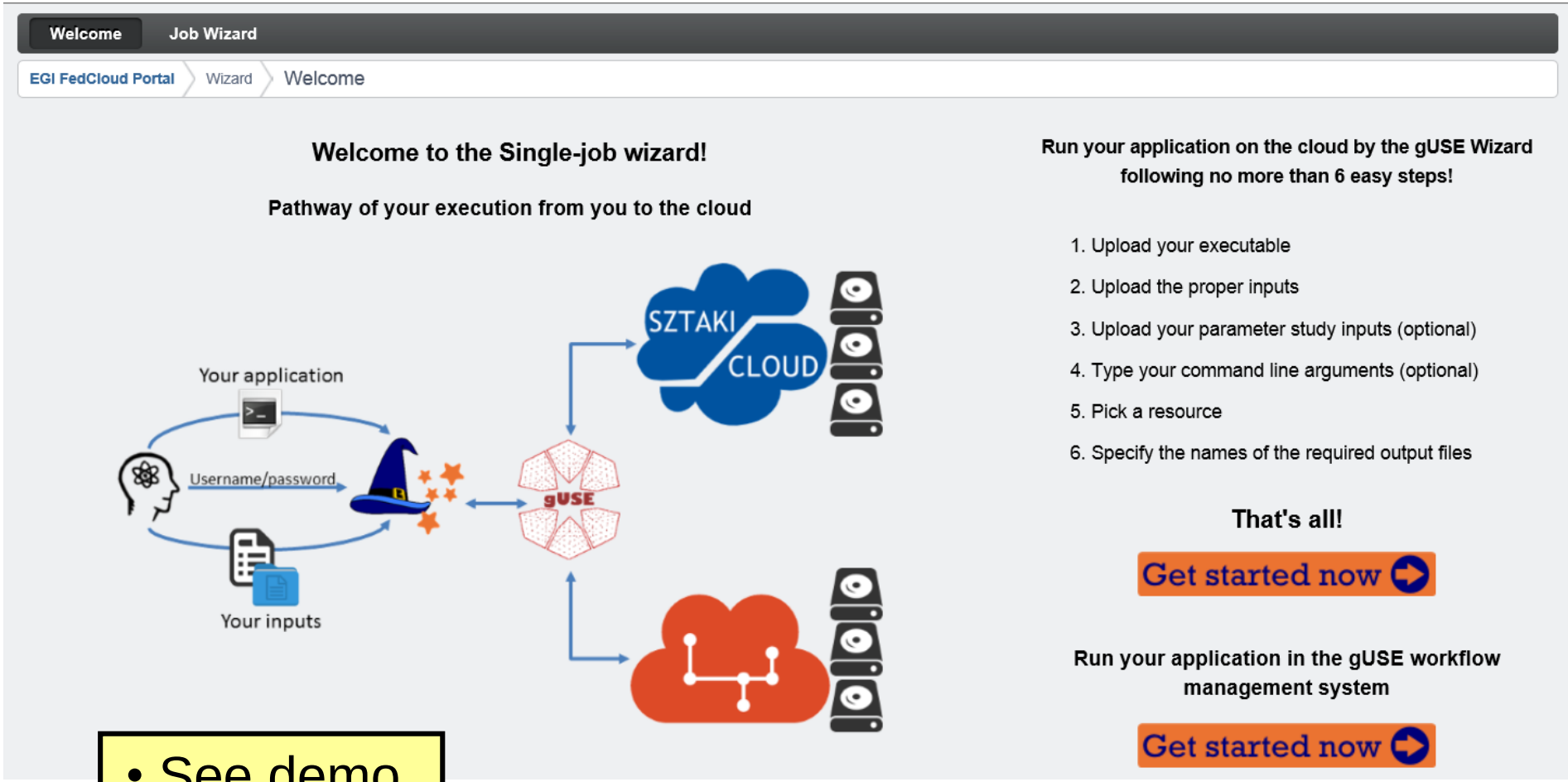
- Flexible management of **Security:**
  - Individual users' certificate
  - Robot certificates
- Flexible access to **various types of DCIs**:
  - Clusters (PBS, LSF, MOAB, SGE)
  - Cluster grids (ARC, gLite, GT2, GT4, GT5, UNICORE)
  - Supercomputers (UNICORE, XSEDE)
  - Desktop grids (BOINC)
  - Cloud access to EGI FedCloud

# Flexibility in using cloud

- Any workflow node can be executed as **submitted job** to the clouds of the EGI FedCloud (see demo)
- Any workflow node can be used to deploy a **service** in the cloud (and the next nodes can use this service)
- Any workflow node can be used to deploy a **virtual infrastructure** in the cloud (and the next nodes can use this VI)

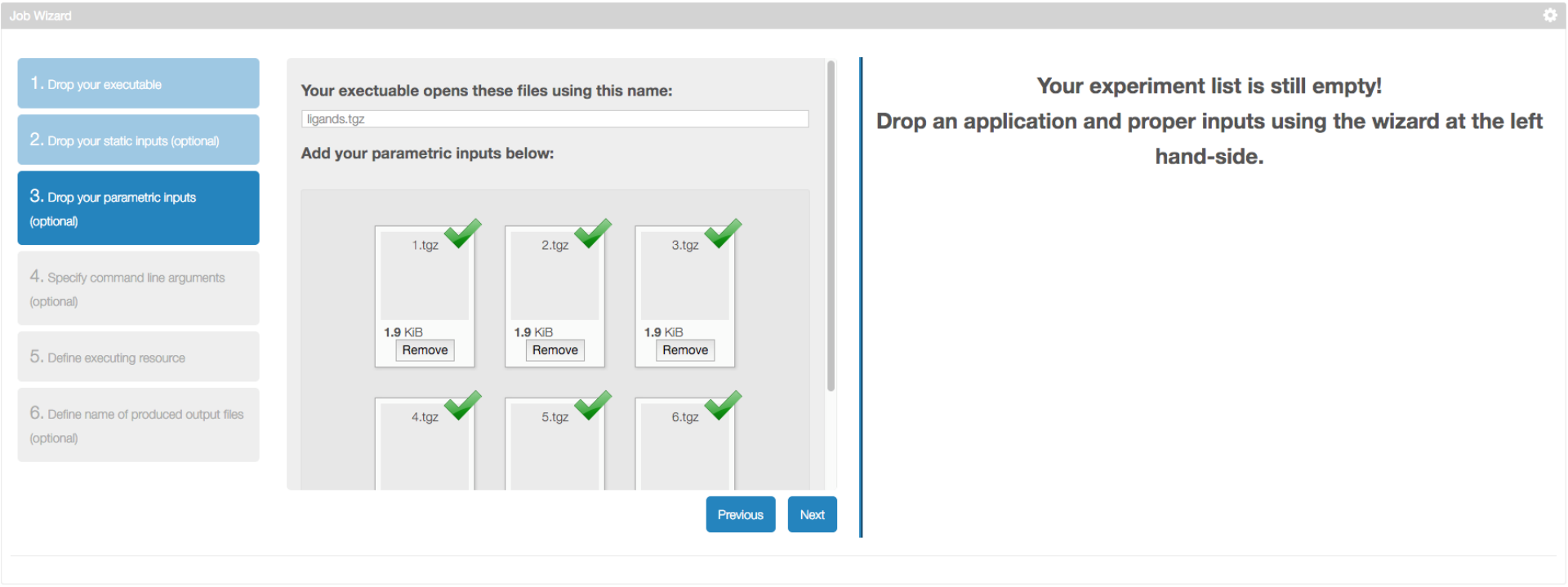
- **Infrastructure-aware workflow concept**

# gUSE wizard for novice cloud users

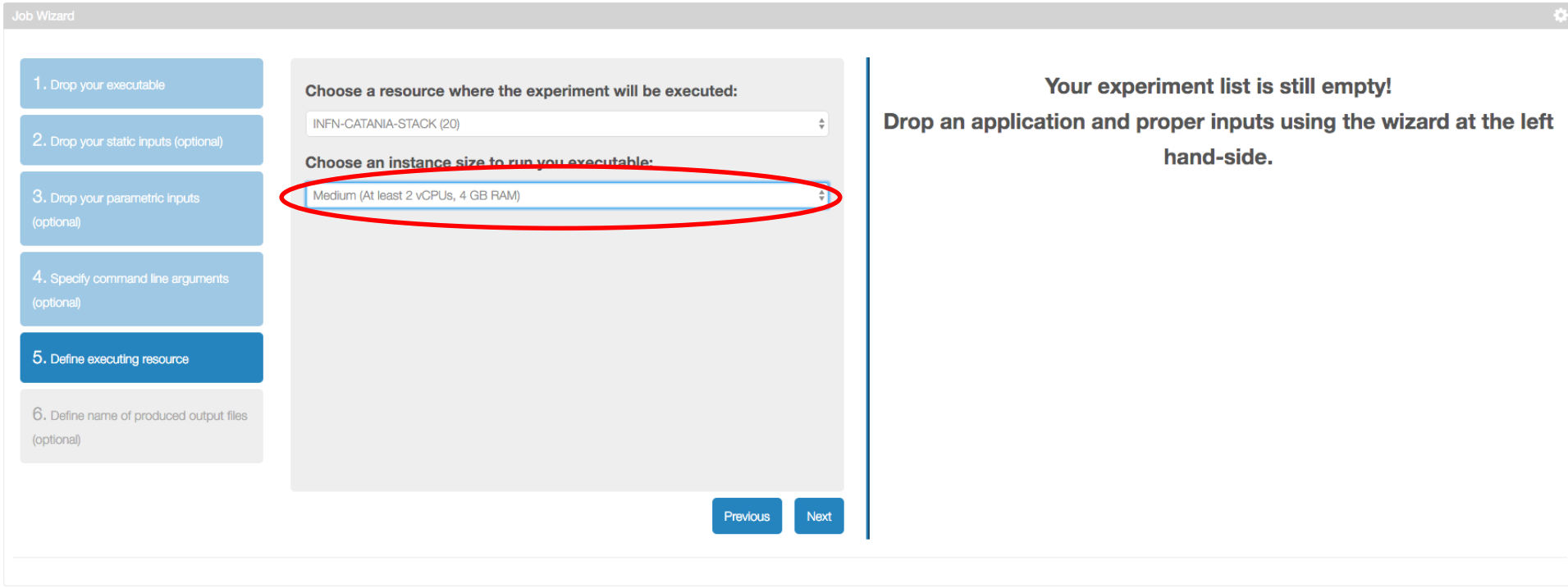



- See demo

# Step 3 – Define parametric inputs



# Step 5 – Select resource and instance size

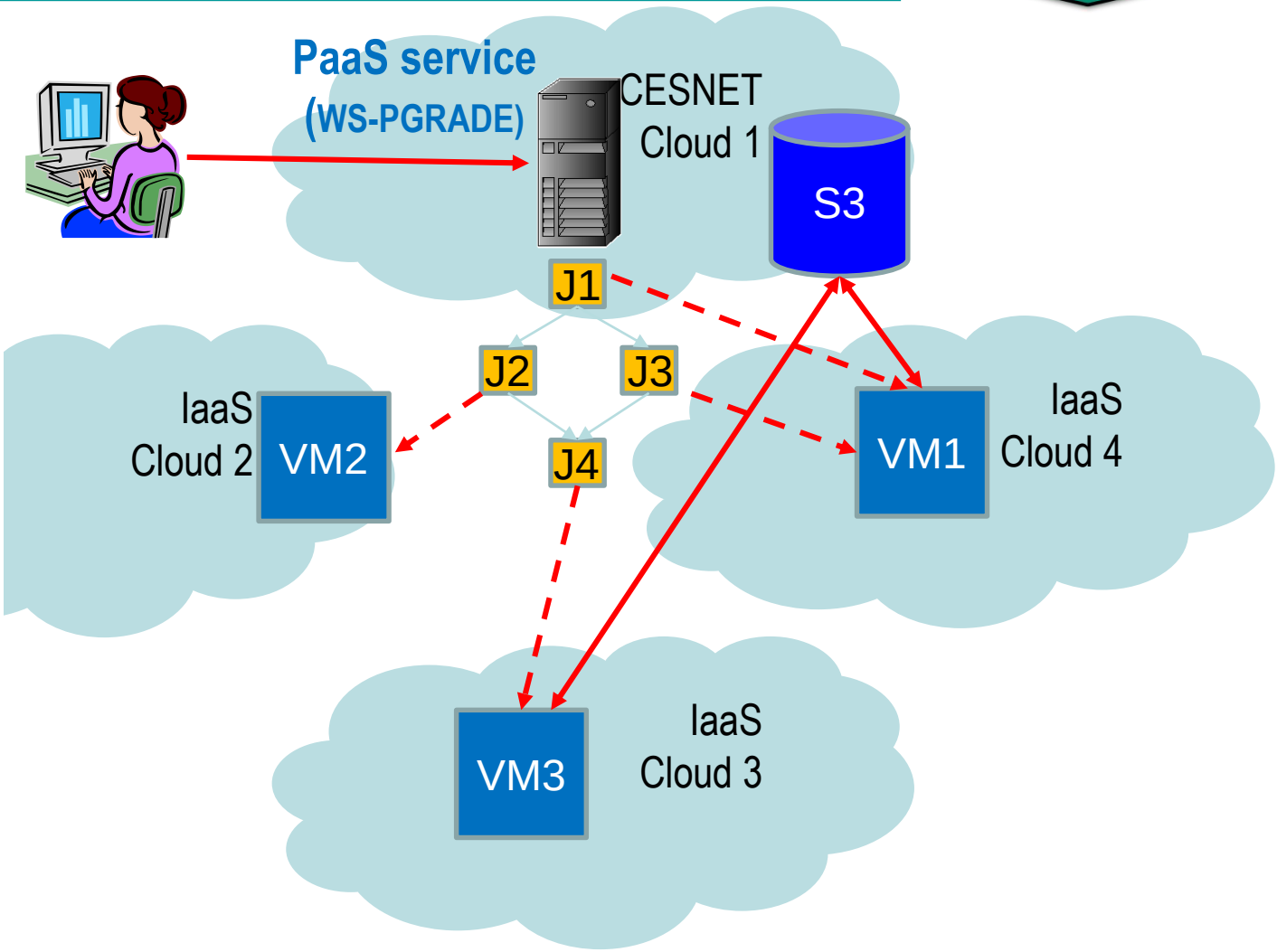



- Small instance: at least 1 vCPU, 2 GB RAM
- Medium instance: at least 2 vCPUs, 4 GB RAM
- Large instance: at least 4 vCPUs, 8 GB RAM

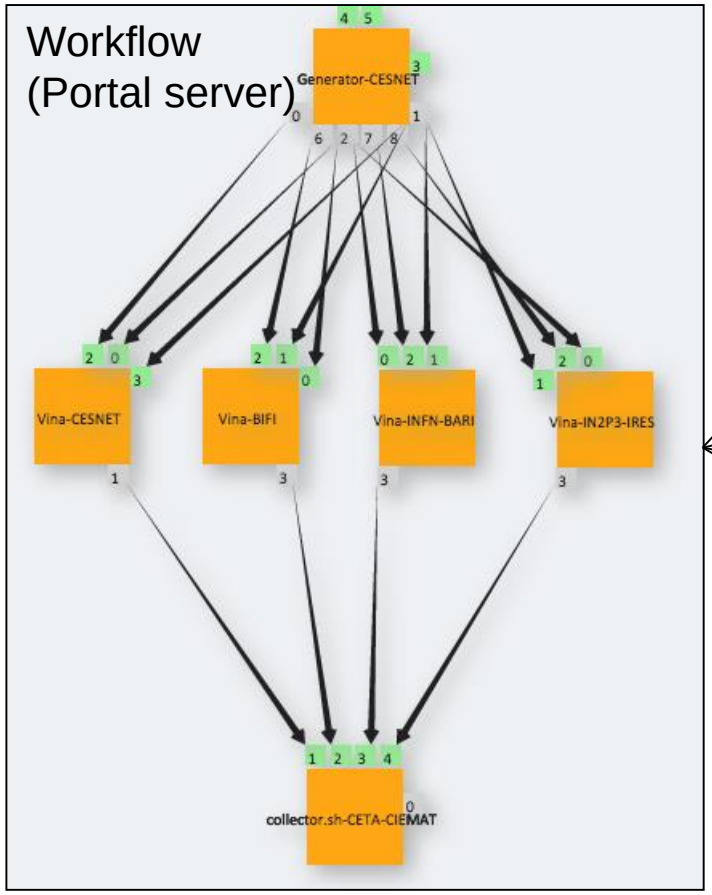
• Only at the long tail of science portal

# WS-PGRADE/gUSE for EGI FedCloud

- User creates complex WF application that runs in the clouds of EGI FedCloud
- WF execution uses new VMs on-demand from different clouds



# Mapping WFs to EGI FedCloud by WS-PGRADE/gUSE




Acts as a proxy for the portal

DCIB VM — IaaS Cloud 1

DCIB VM — IaaS Cloud 2

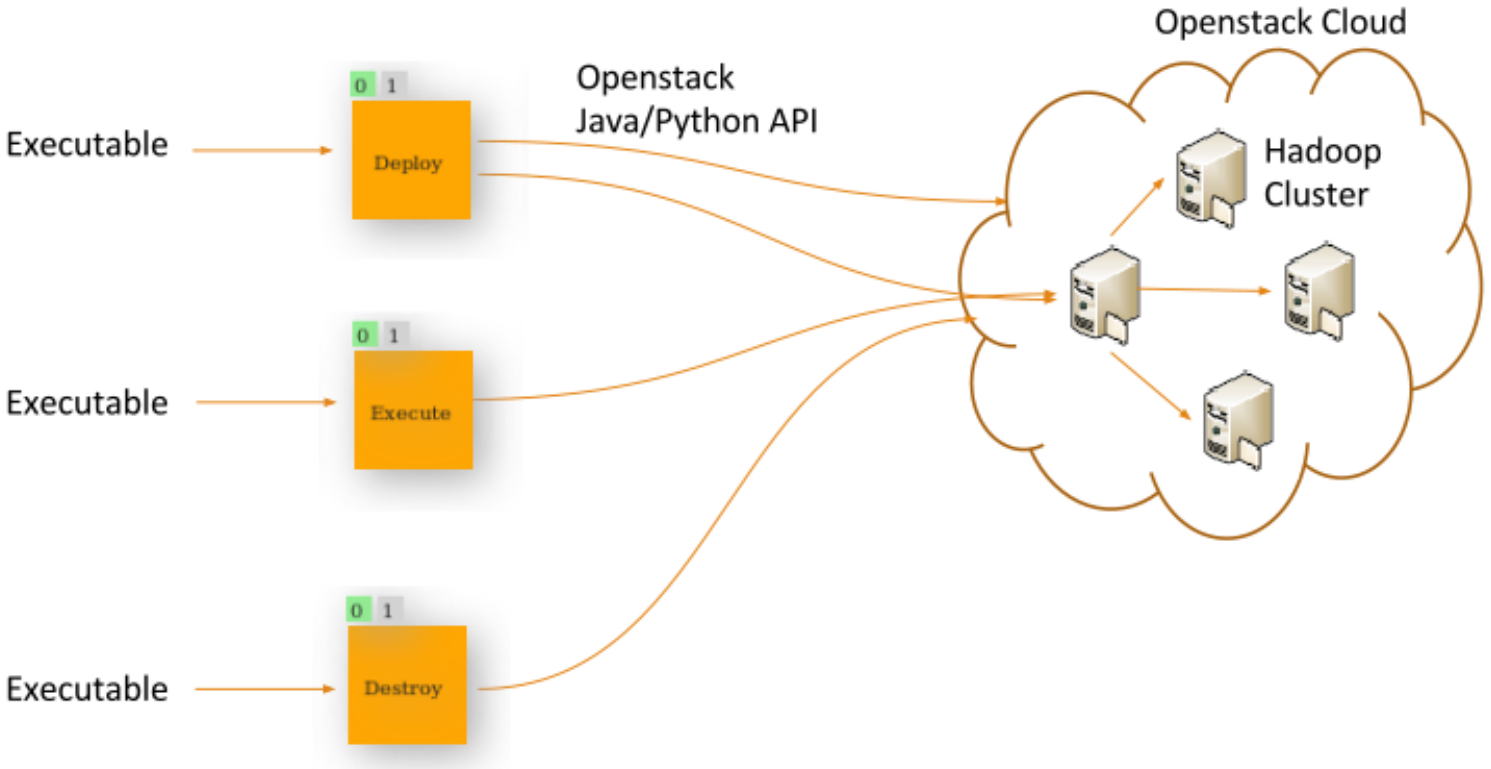
DCIB VM — IaaS Cloud 3

. . .

DCIB VM — IaaS Cloud 20

# Support for hadoop/mapreduce applications

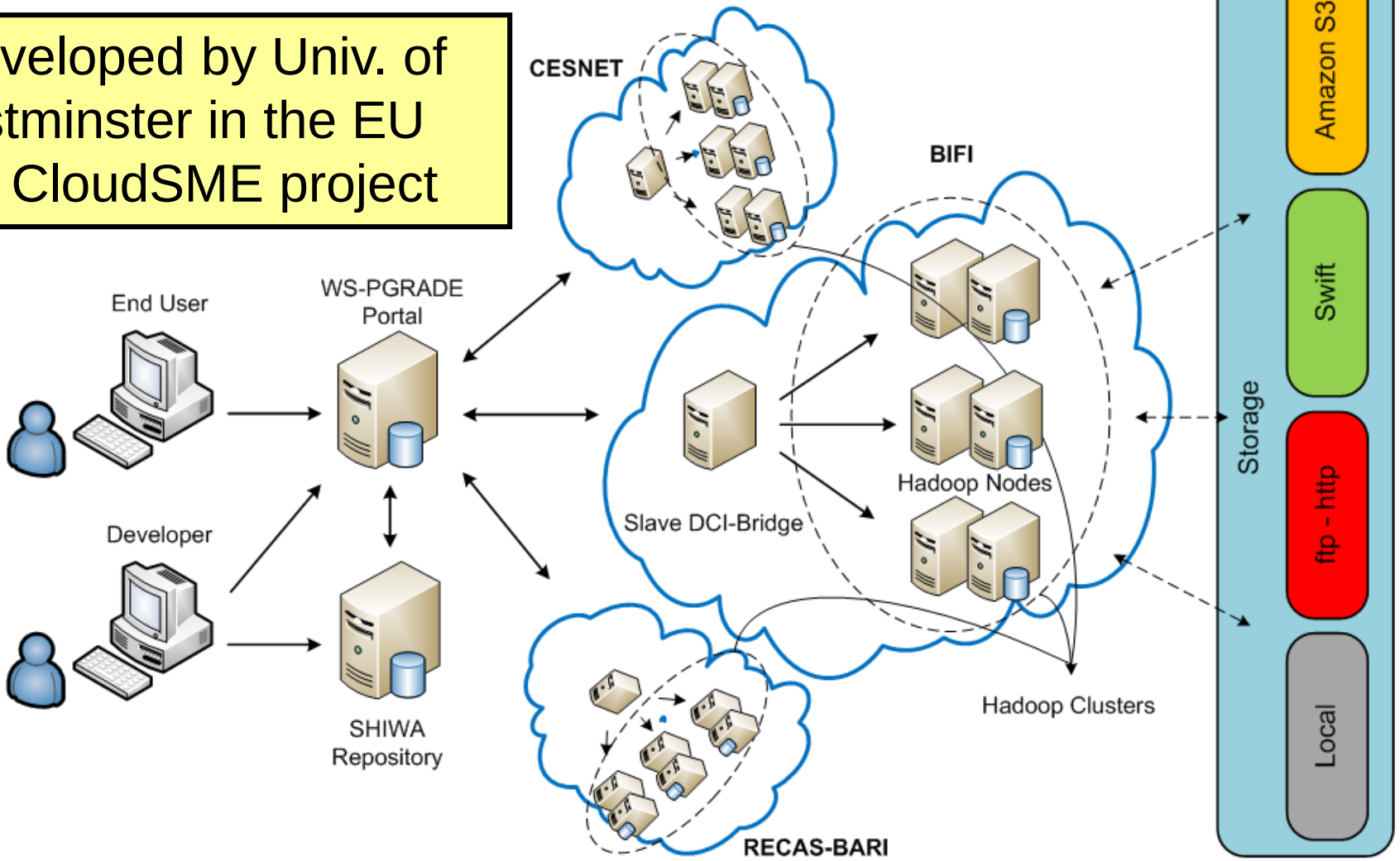
- Stage 1 (Deploy Hadoop Node): Launch servers in cloud, connect to master node, setup Hadoop cluster and save Hadoop cluster configuration
- Stage 2 (Execute Mapreduce Node): Upload input files and job executable to master node, execute job and get result back
- Stage 3 (Destroy Hadoop Node): Destroy cluster to free up resources



# EGI FedCloud Implementation

SCI-BUS

- Developed by Univ. of Westminster in the EU FP7 CloudSME project



# How to use it?

1. Create an account on the EGI FedCloud WS-PGRADE Gateway:

   https://guse-fedcloud-gateway.sztaki.hu/

2. Import the Hadoop workflow(s) to your account from the SHIWA Workflow Repository

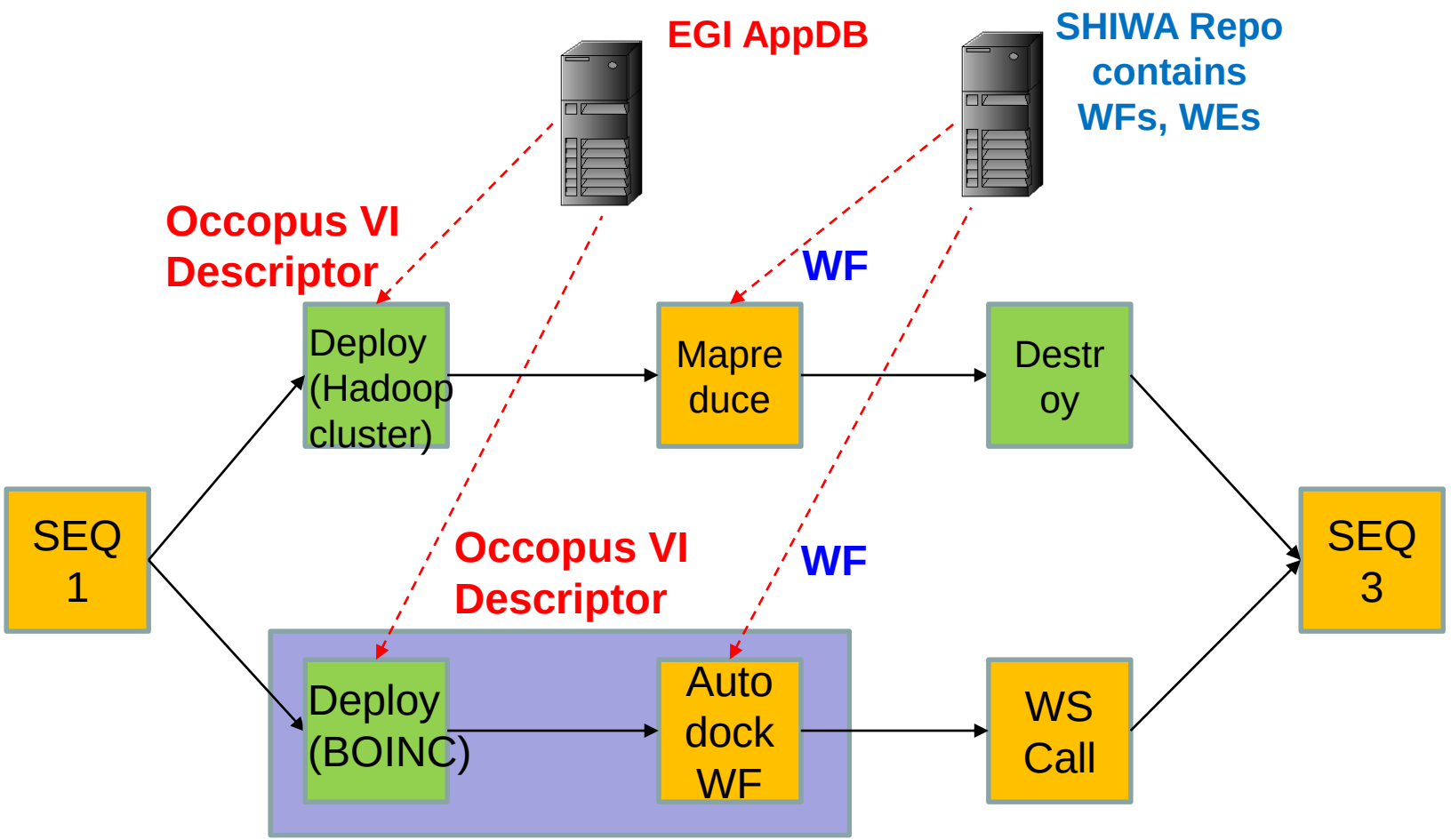
3. Download and customise sample configuration files

4. Configure workflow by uploading configuration files and Hadoop source/executables

5. Submit

See demonstration and user manual for further details

# Generic solution (future work): Infrastructure-aware workflow



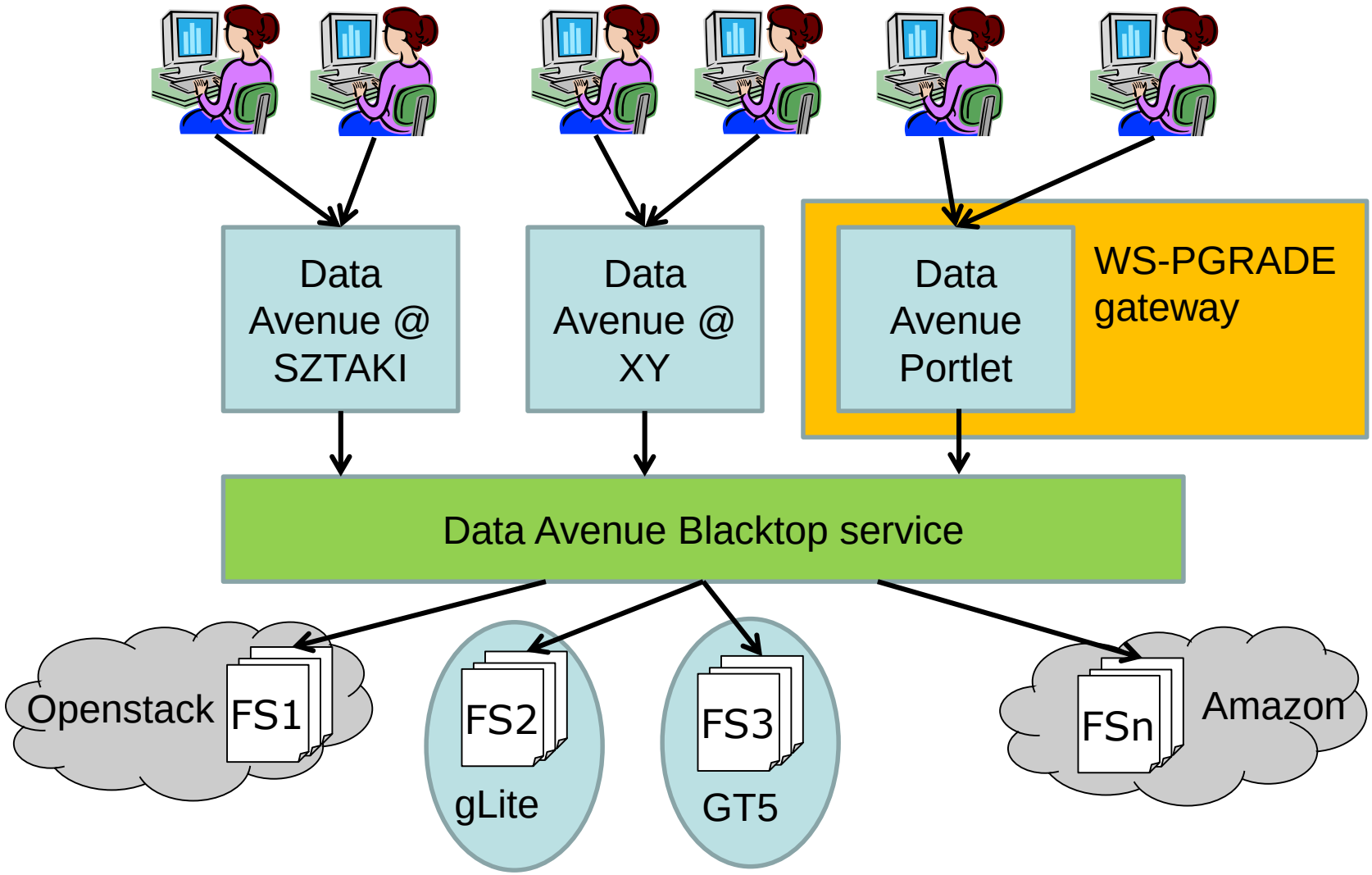
# Flexibility in data storage access


| | | | |
|---|---|---|---|
| VizIVO gateway | Proteomics Gateway | MoSGrid Gateway | *Application specific gateways (more than 30)* |
| Workflow Editor | Workflow execution Monitor | Data Avenue UI | *Web user interface (WS-PGRADE)* |
| Workflow Management | Workflow Repository | Internal Storages | *Workflow and internal storage services (gUSE)* |
| DCI Bridge | | Data Avenue | *High-level e-infrastructure middleware (gUSE)* |
| HTC Infrastructures | HPC Infrastructures | Large variety of data storages | *Production e-infrastructures* |

# Flexibility in data storage access

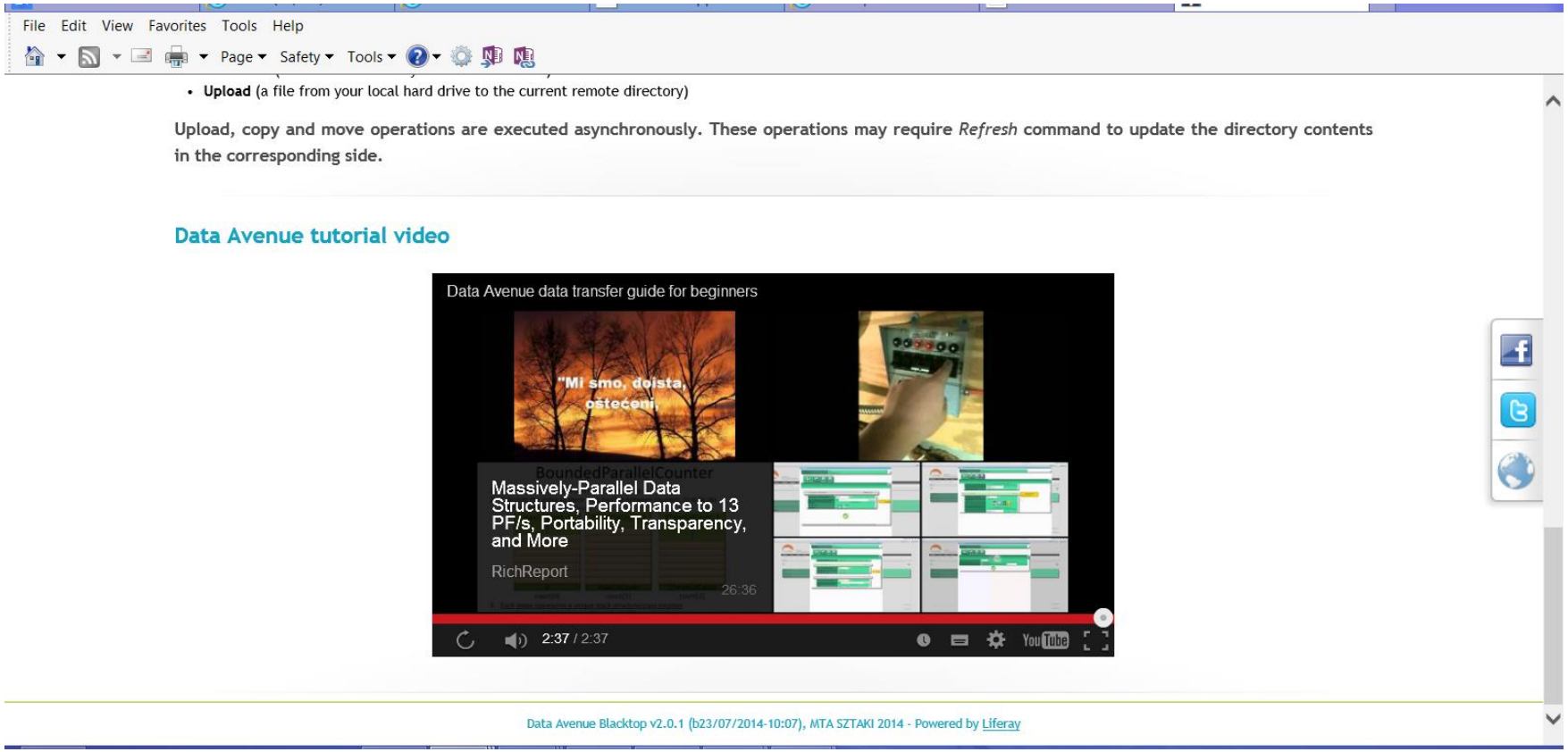
- **Use Data Avenue Blacktop service**
  - To access data storages in different DCIs
  - To transfer files among the storages of different DCIs
  - To upload/download files to/from the storages of different DCIs
- **Data Avenue Liferay portlet** to access the data transfer services of Data Avenue Blacktop
- See details: http://data-avenue.eu/home
- Currently supported protocols:
  - http, https, ftp, gsiftp, srm, iRODS, LFC, S3

# Data Avenue services



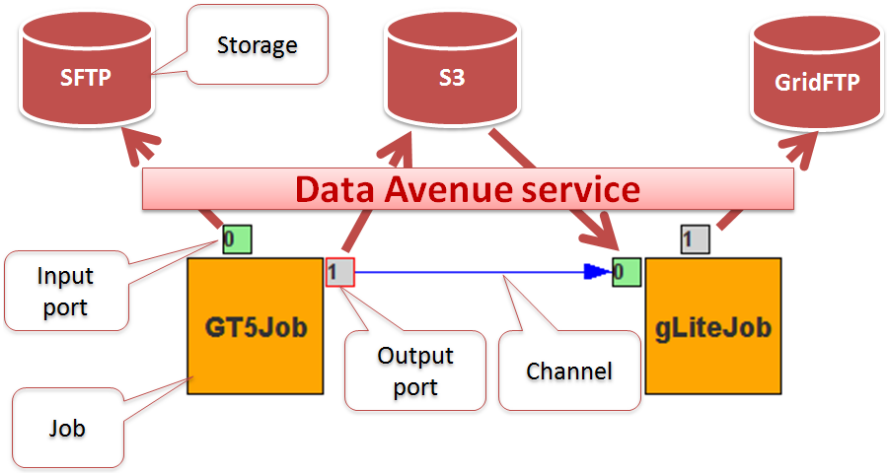
# Data Avenue demo video

https://data-avenue.eu/en_GB/data-transfer-manuals

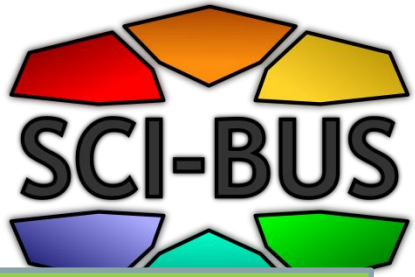


# Data Avenue in WS-PGRADE/gUSE

- Data sources and destinations of jobs can be selected from major storage types
- gUSE automatically manages data transfers using Data Avenue Blacktop
- Actual transfer is delegated up to the worker node wherever possible, *by*passing the Blacktop service if the middleware is capable of handling the protocol

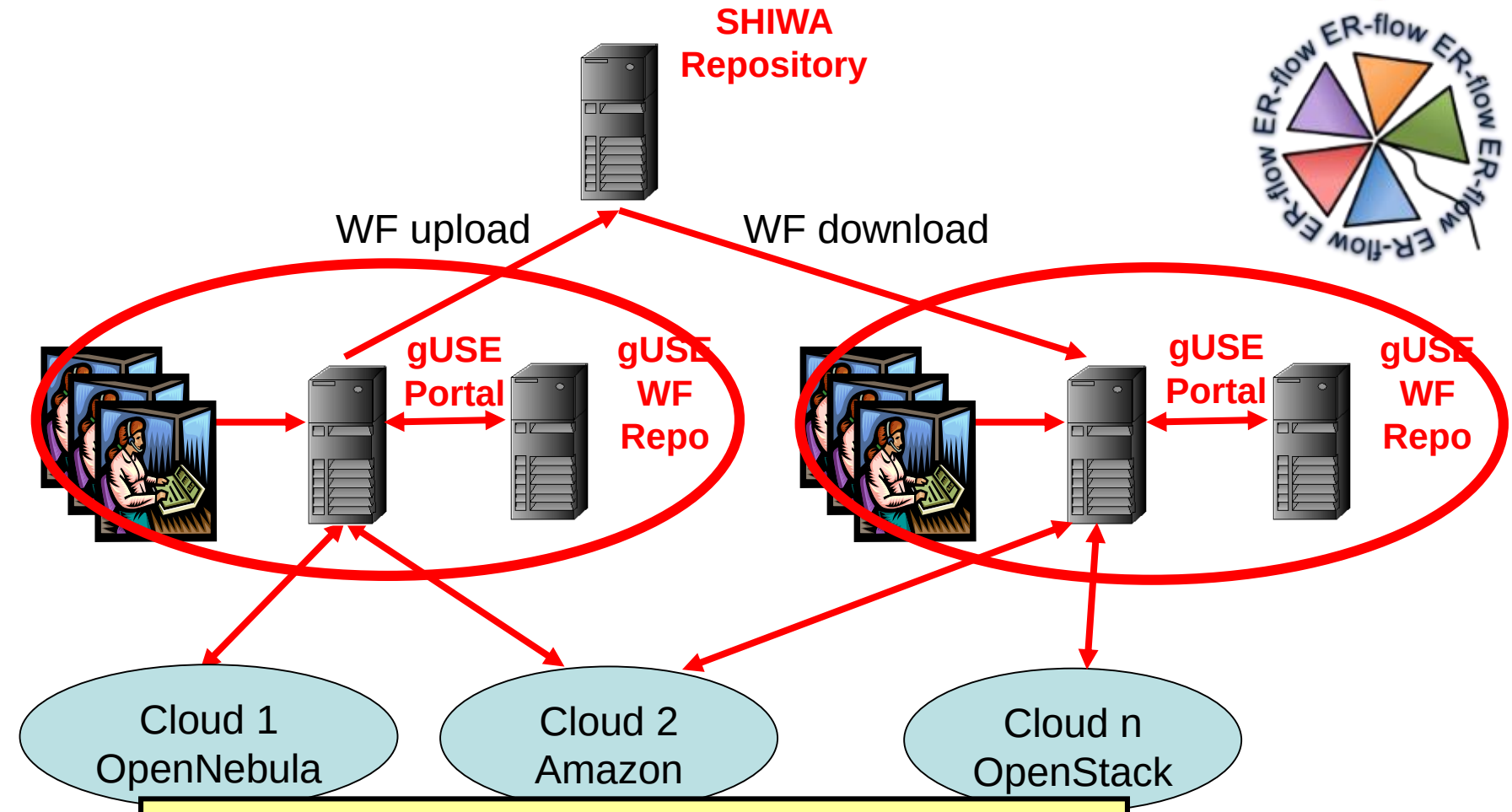



# Flexibility in collaboration among community members




| | | | |
|---|---|---|---|
| VizIVO gateway | Proteomics Gateway | MoSGrid Gateway | *Application specific gateways (more than 30)* |
| Workflow Editor | Workflow execution Monitor | Data Avenue UI | *Web user interface (WS-PGRADE)* |
| Workflow Management | Workflow Repository | Internal Storages | *Workflow and internal storage services (gUSE)* |
| DCI Bridge | | Data Avenue | *High-level e-infrastructure middleware (gUSE)* |
| HTC Infrastructures | HPC Infrastructures | Large variety of data storages | *Production e-infrastructures* |

# Flexibility in collaboration among community members

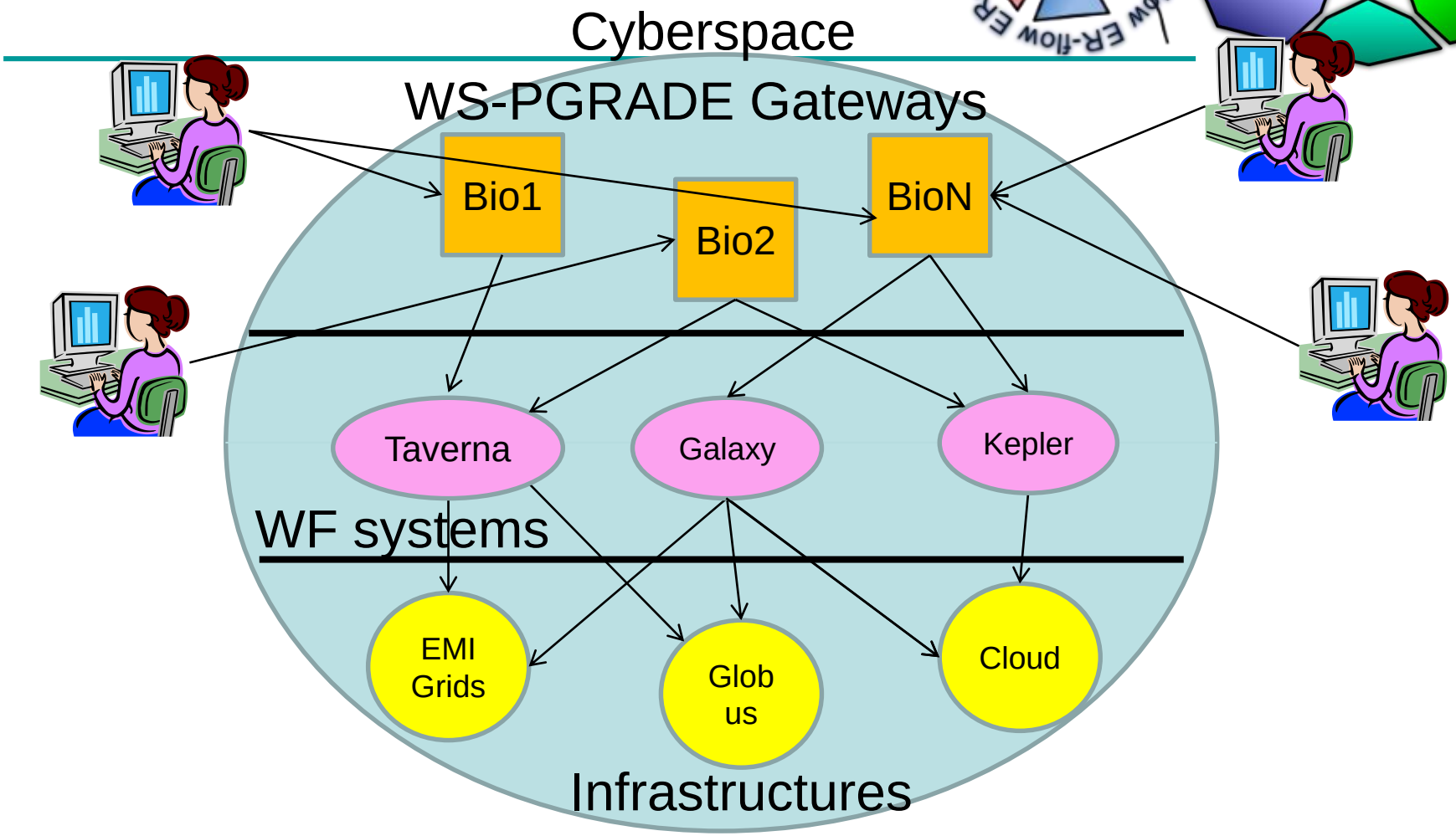



- Developed in EU FP7 SHIWA project
- Exploited in EU FP7 ErFlow project

# WS-PGRADE/gUSE Architecture


| | | | |
|---|---|---|---|
| VizIVO gateway | Proteomics Gateway | MoSGrid Gateway | *Application specific gateways (more than 30)* |
| Workflow Editor | Workflow execution Monitor | Data Avenue UI | *Web user interface (WS-PGRADE)* |
| Workflow Management | Workflow Repository | Internal Storages | *Workflow and internal storage services (gUSE)* |
| DCI Bridge | | Data Avenue | *High-level e-infrastructure middleware (gUSE)* |
| HTC Infrastructures | HPC Infrastructures | Large variety of data storages | *Production e-infrastructures* |

# Flexibility in using different workflow systems

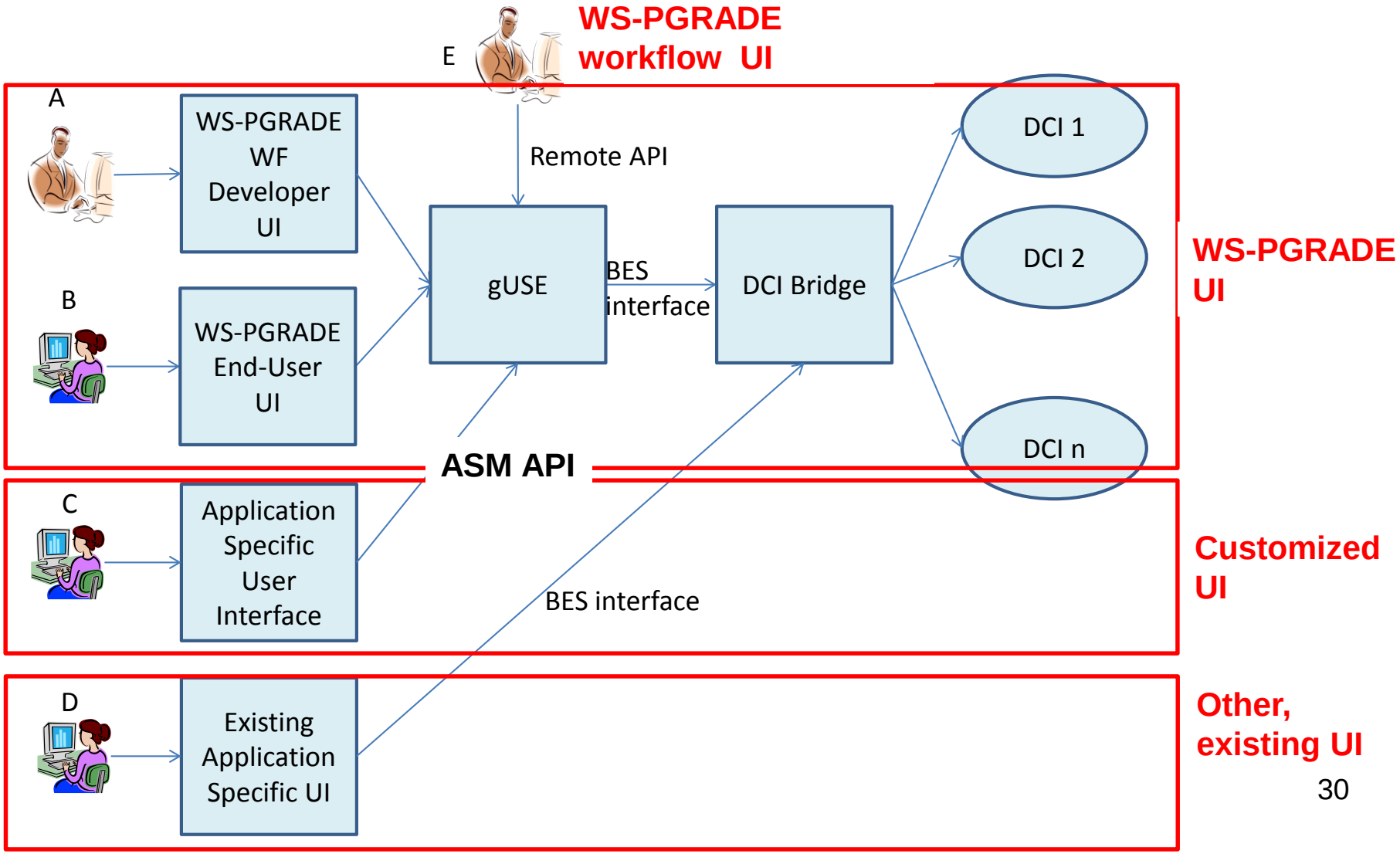



**Combining SCI-BUS and SHIWA technologies (supported by ER-Flow) users can access and use many WFs and many infrastructures in an interoperable way no matter which is their home WF system**

# Flexibility of gateway types and user views

1. Generic purpose gateways for clouds (workflow view)
   - Core WS-PGRADE/gUSE (e.g. Greek NGI)
2. Generic purpose gateway for specific technologies (workflow view)
   - SHIWA gateway for workflow sharing and interoperation
3. Domain-specific science gateway instance
   - Autodock gateway (end-user view)
   - Swiss proteomics portal (customized GUI using ASM API)
   - VisIVO Mobile (use of Remote API)

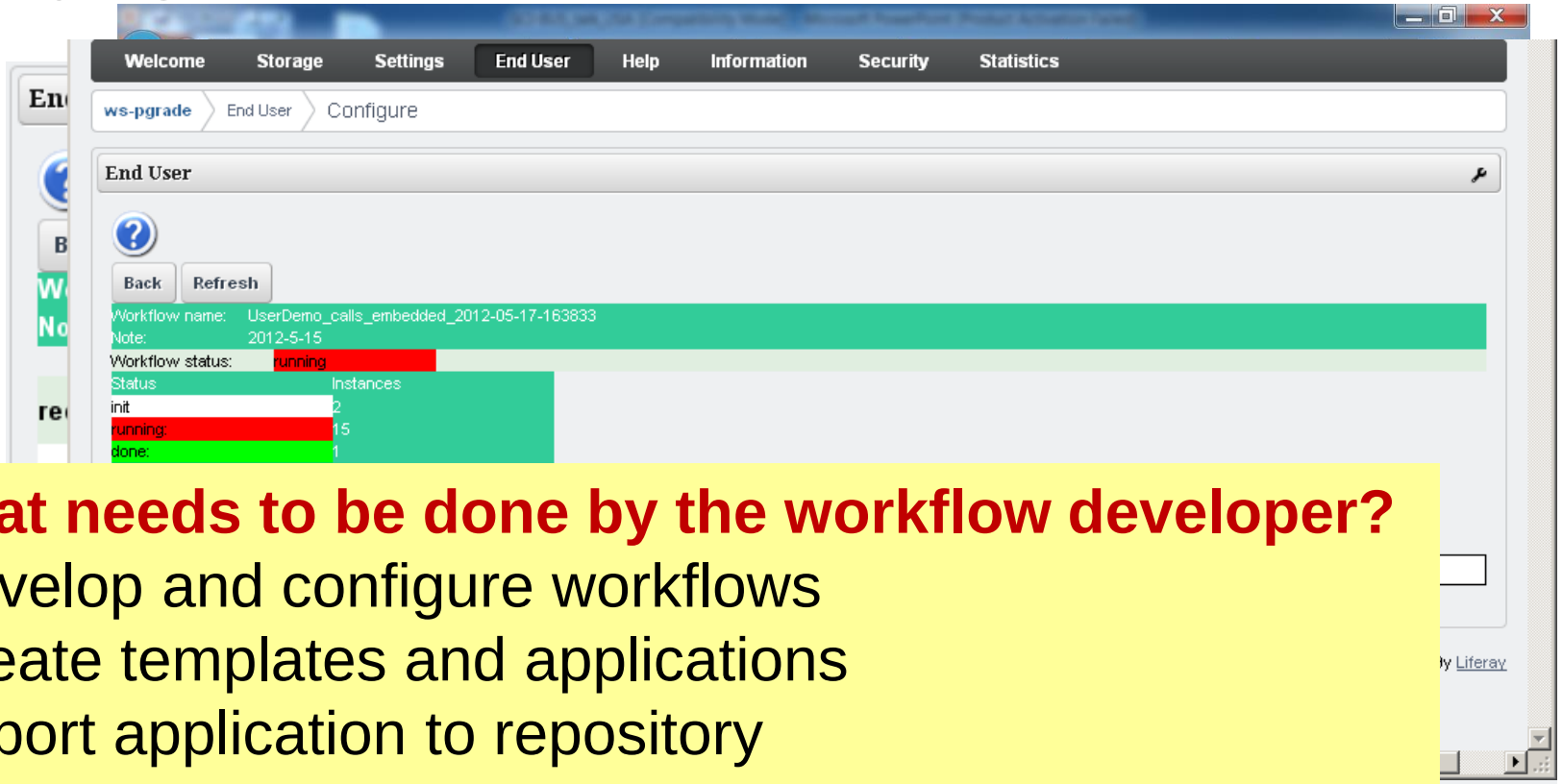
# Flexibility in user access modes



# End-user view based gateways

## What is required from the end-user?

- Import workflow from repository
- Customise, execute and monitor application using simple web forms

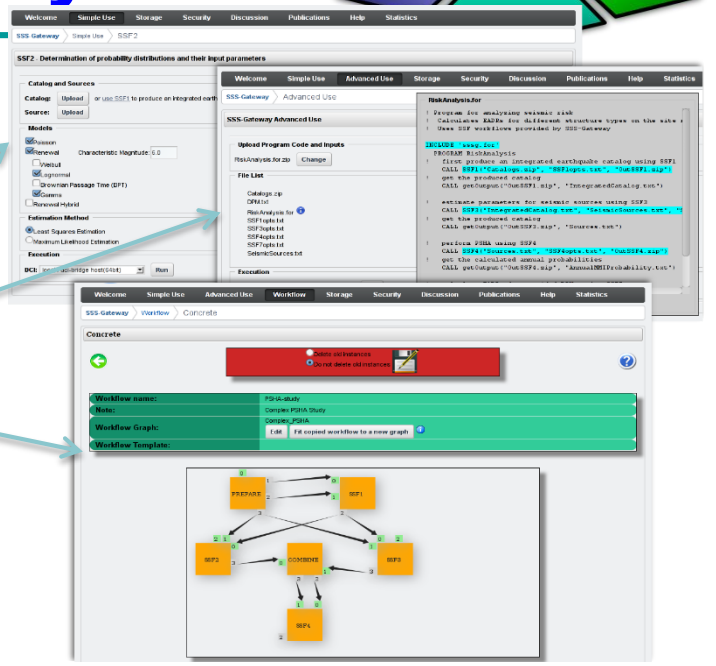


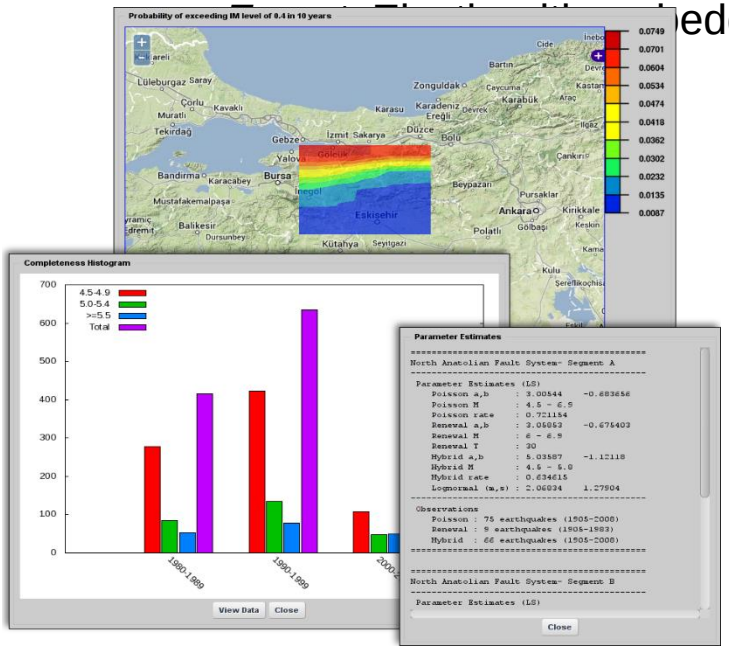

## What needs to be done by the workflow developer?

- Develop and configure workflows
- Create templates and applications
- Export application to repository

# Using ASM API: Statistical Seismology Science Gateway

- Provides seven statistical seismology functions
  to international seismology community with three service levels:
  - *Simple:* Simplified GUIs
  - *Advanced:* Powerful programming interface
  - [illegible]edded workflows



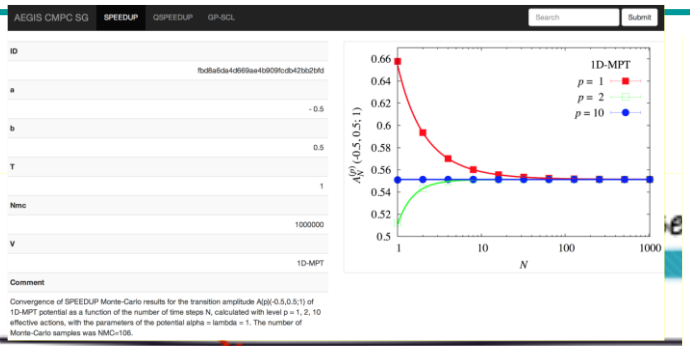


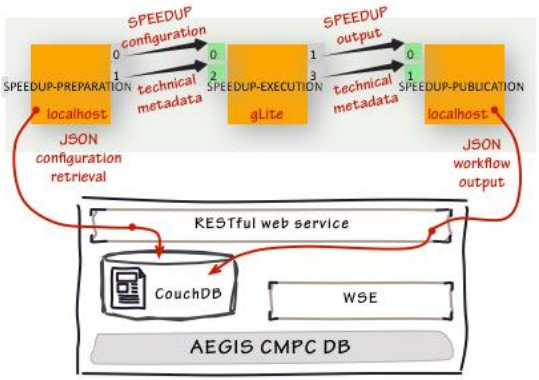

- Based on gUSE with ASM and Remote API
- Various methods/models from "*integration of multi-source data*" to "*generation of hazard maps*"

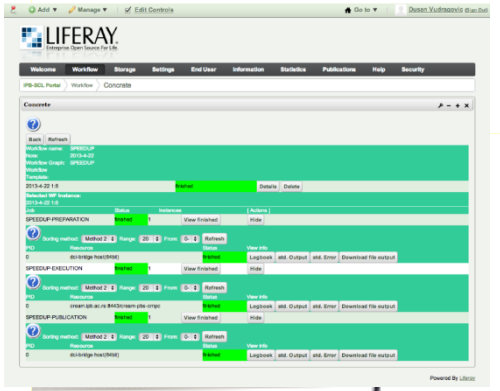
***seismo.ceng.metu.edu.tr***
***sss-gateway@sci-bus.eu***

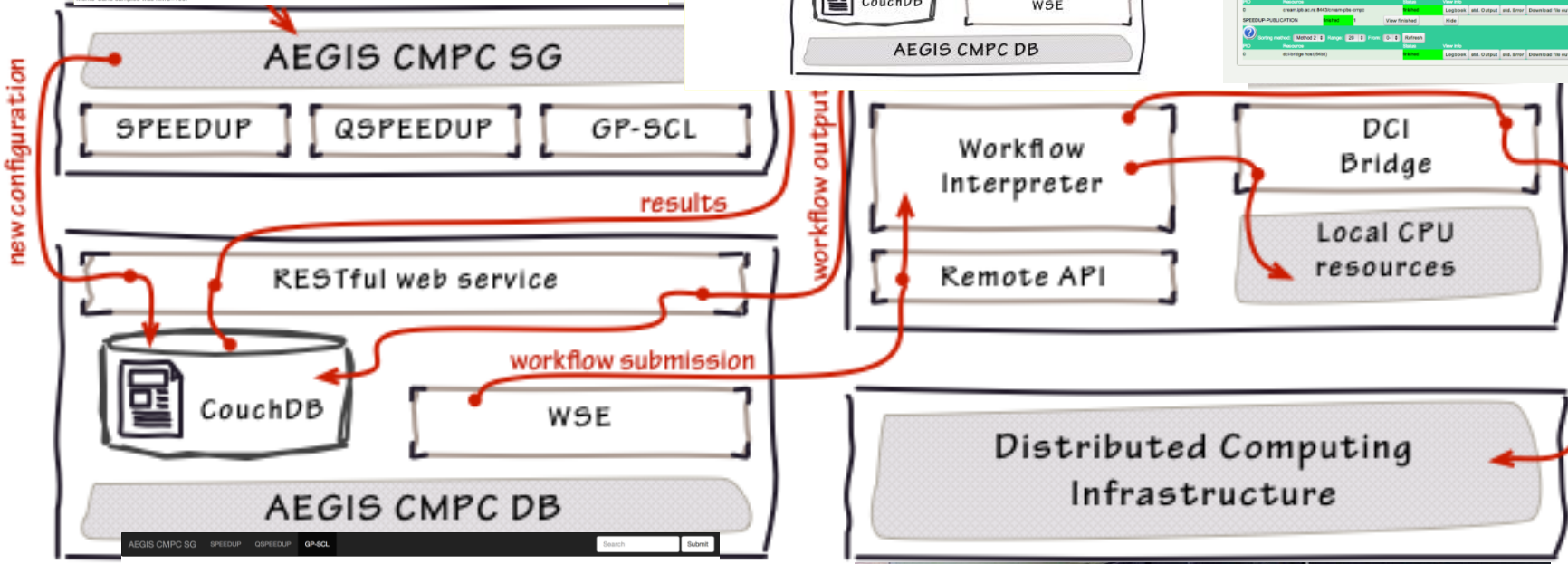
METU

# Using Remote API: AEGIS CMPC SG Institute of Physics Belgrade



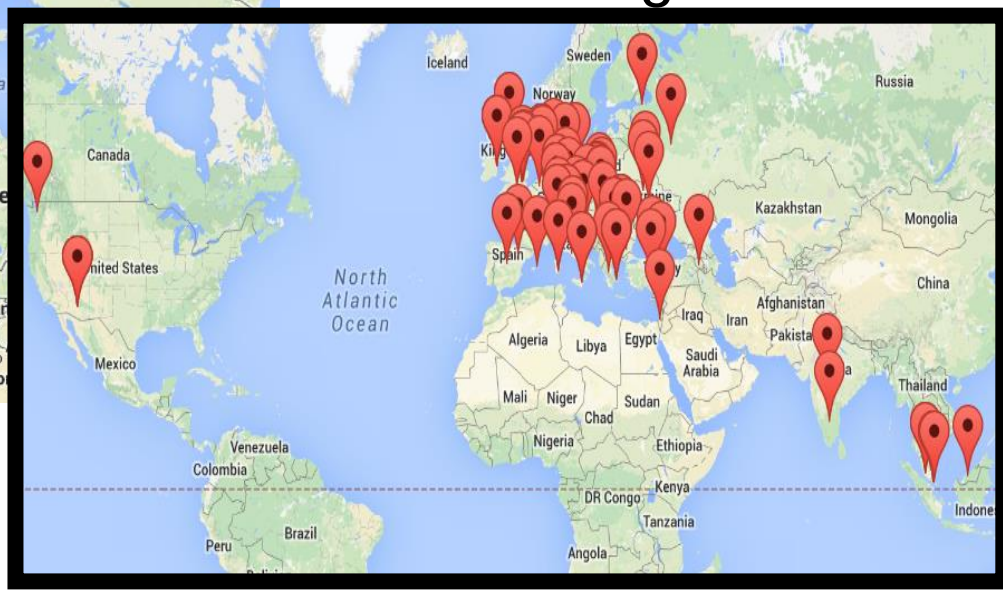







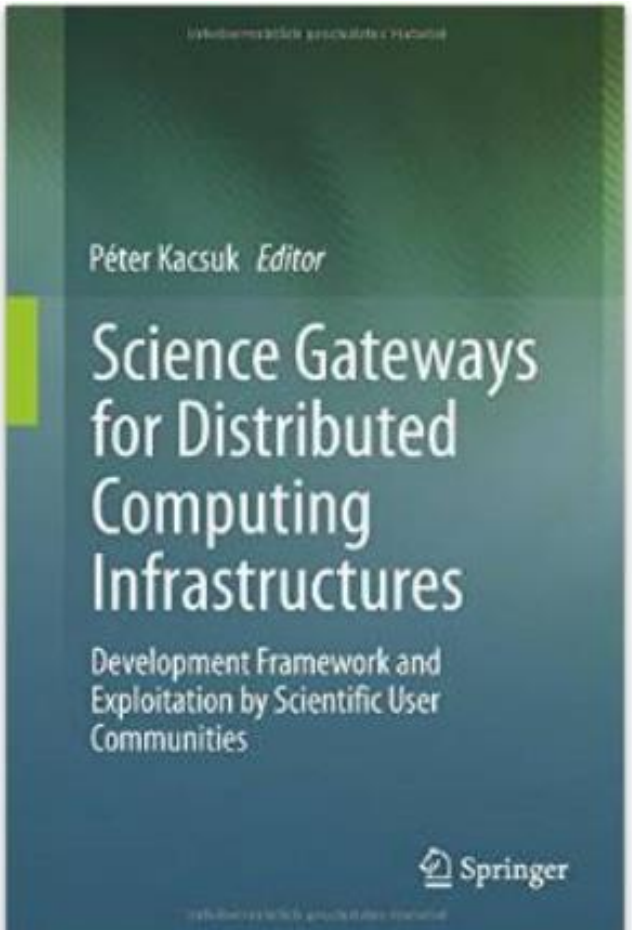

# Technology adaptation


- More than 100 deployments world-wide
- Nearly 22.000 downloads from 80 countries on sourceforge



# Further readings

- SCI-BUS web page:
  - http://www.sci-bus.eu/
- gUSE on sourceforge
  - http://sourceforge.net/projects/guse/
  - http://sourceforge.net/projects/guse/develop
- To try the gateway go to FedCloud gateway:
  - https://guse-fedcloud-gateway.sztaki.hu/



# Conclusions

Why to use WS-PGRADE/gUSE?

1. Robustness
   - Already large number of gateways used in production
2. Sustainability
   - Within the DARIAH CC of the EGI ENGAGE project
3. Functionalities
   - Rich functionalities that are growing according to the various community needs
4. How easy to adapt for the needs of the new user community?
   - Already large number of gateways customized from gUSE/WS-PGRADE
5. You can influence the progress of WS-PGRADE/gUSE