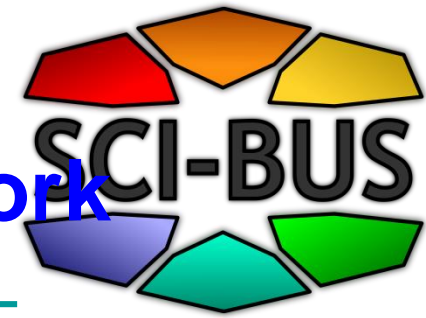


Workflow applications on EGI with WS-PGRADE

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

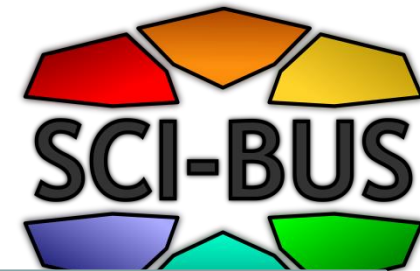
WS-PGRADE/gUSE

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)*

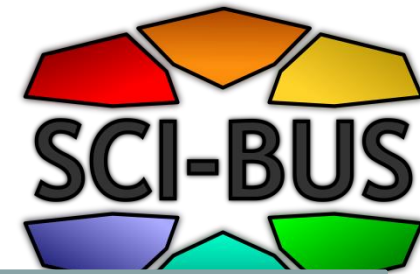
**HTC
Infrastructures**

**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**

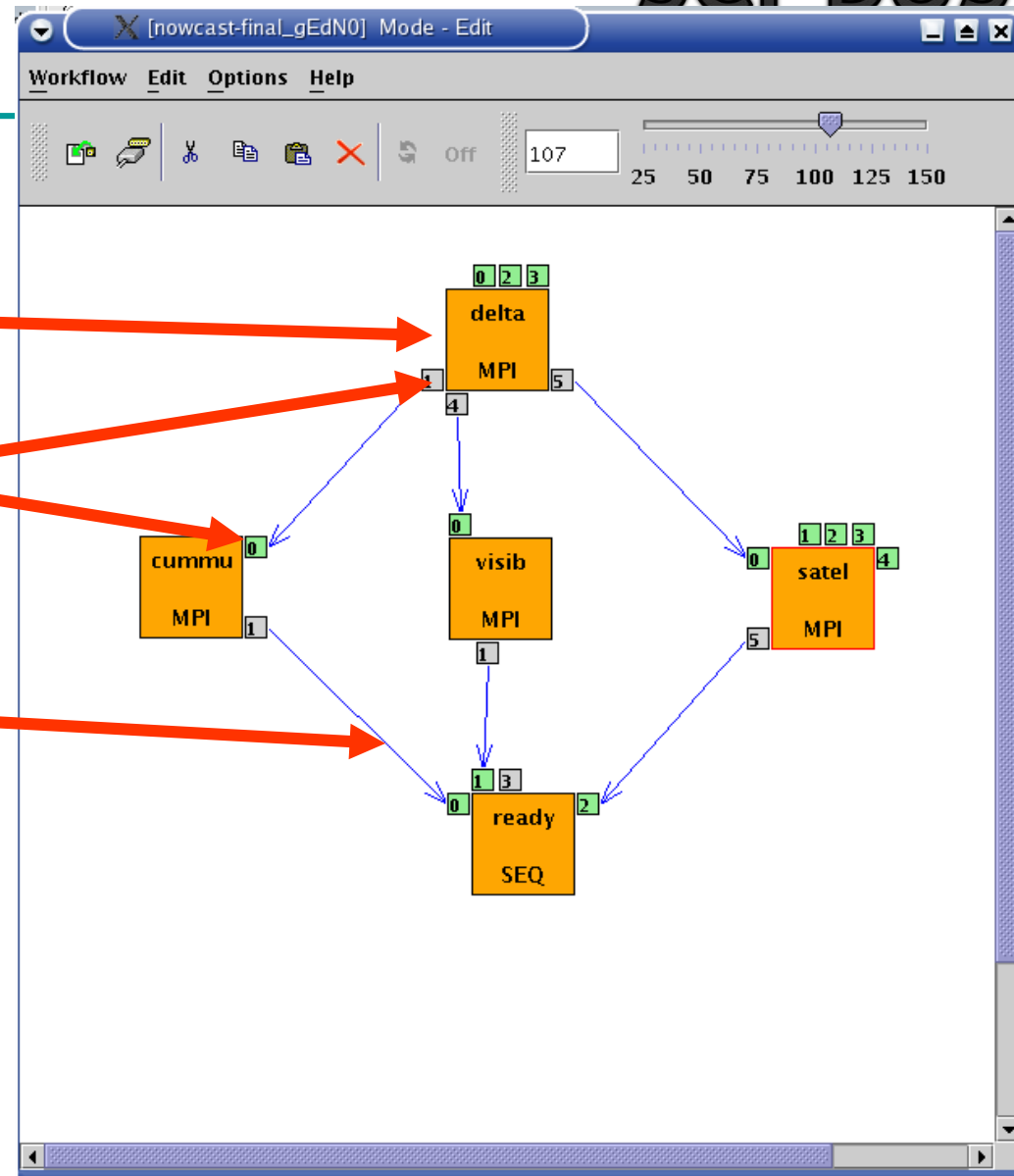
**Large variety
of data
storages**

*Production
e-infrastructures*

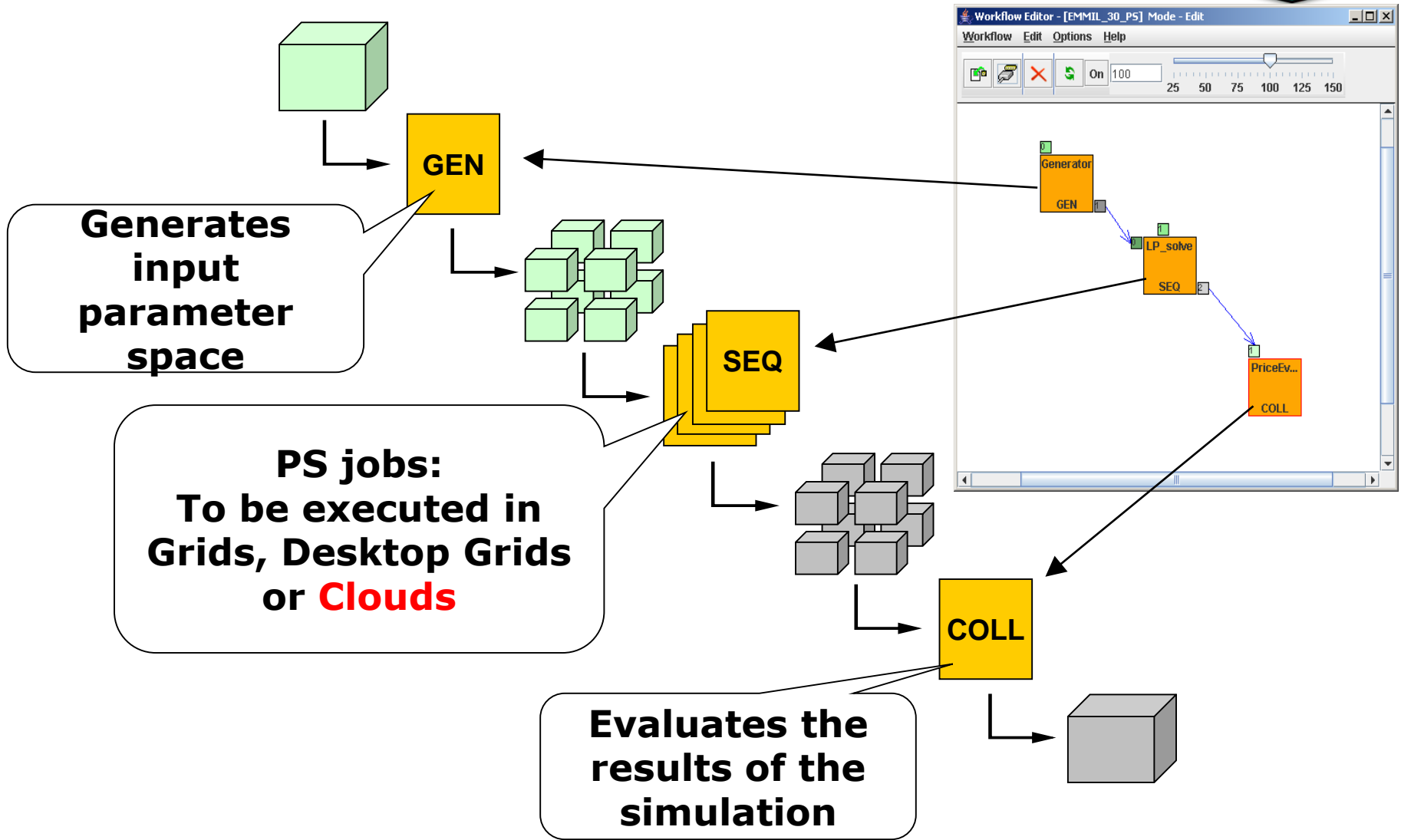
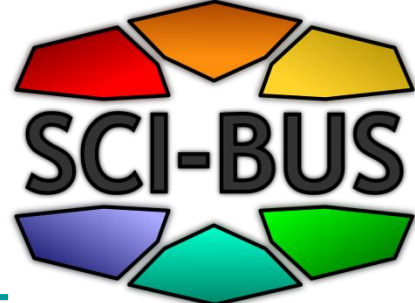
Graphical workflow language



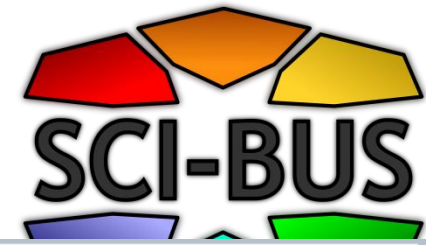
- **A directed acyclic graph where**
 - Nodes represent **jobs, services, embedded workflows**
 - Ports represent input/output files the jobs/services expect/produce
 - Arcs represent file transfer operations and job dependencies
- **semantics of the workflow:**
 - A node can fire, i.e. its job can be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



Parameter Sweep (PS) application workflow template



Flexibility in exploiting parallelism

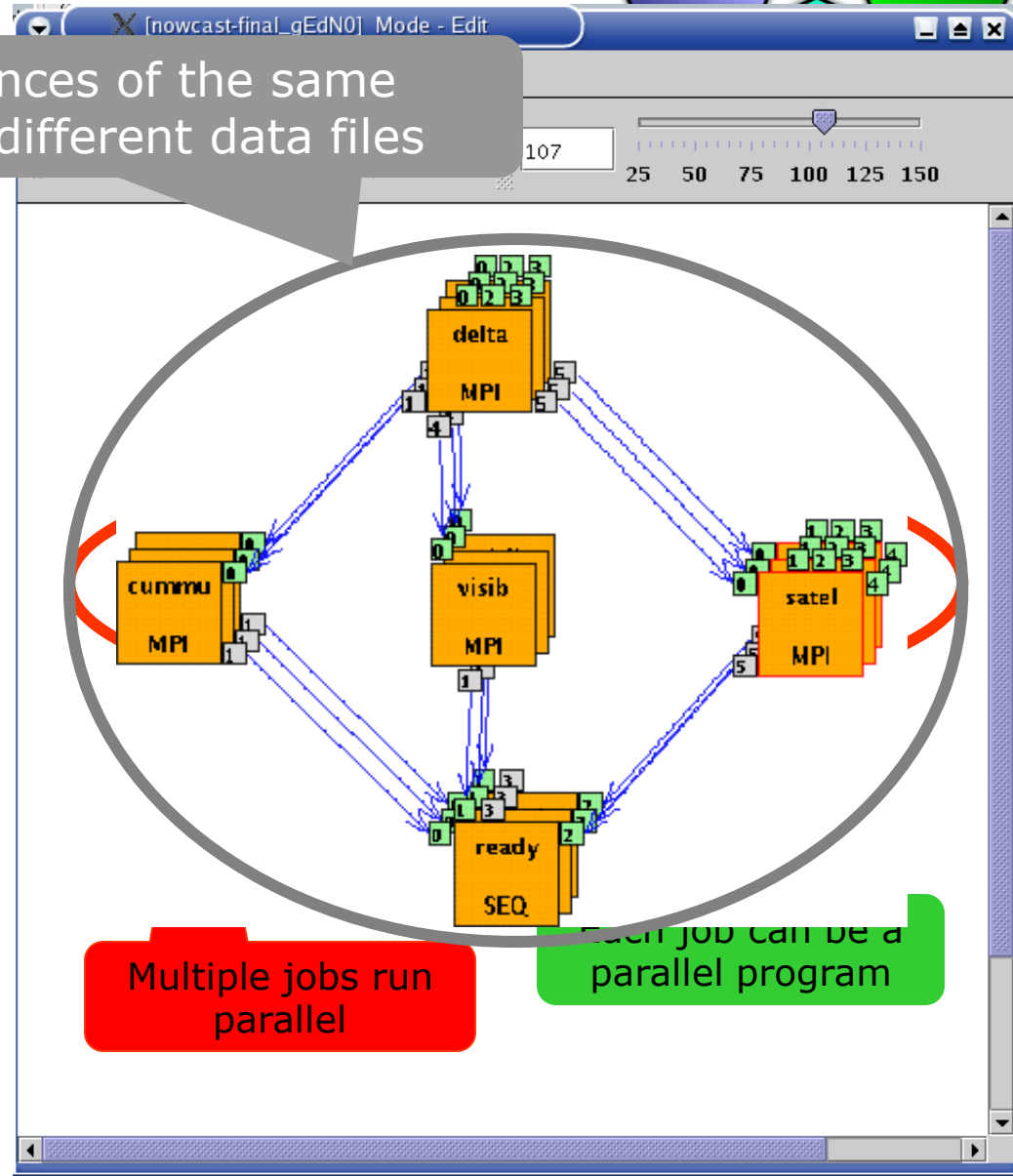


Multiple instances of the same workflow with different data files

– Parallel execution inside a workflow node

– Parallel execution among workflow nodes

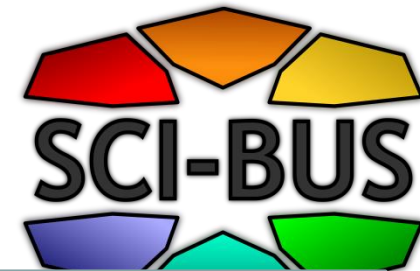
– Parameter study execution of the workflow



Multiple jobs run parallel

each job can be a parallel program

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)*

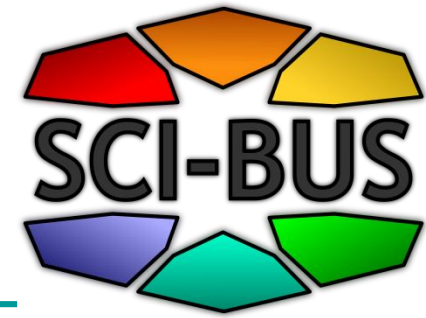
**HTC
Infrastructures**

**HPC
Infrastructures**

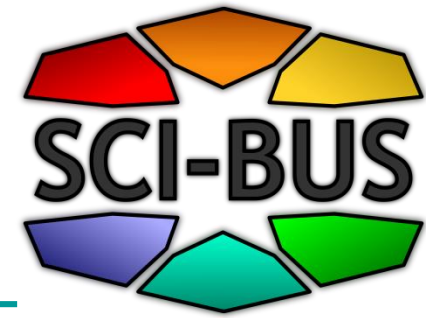
**Large variety
of data
storages**

*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



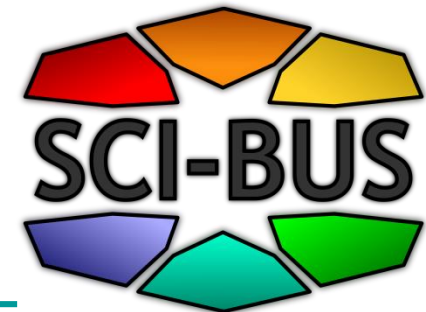
Flexibility in using cloud

- 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

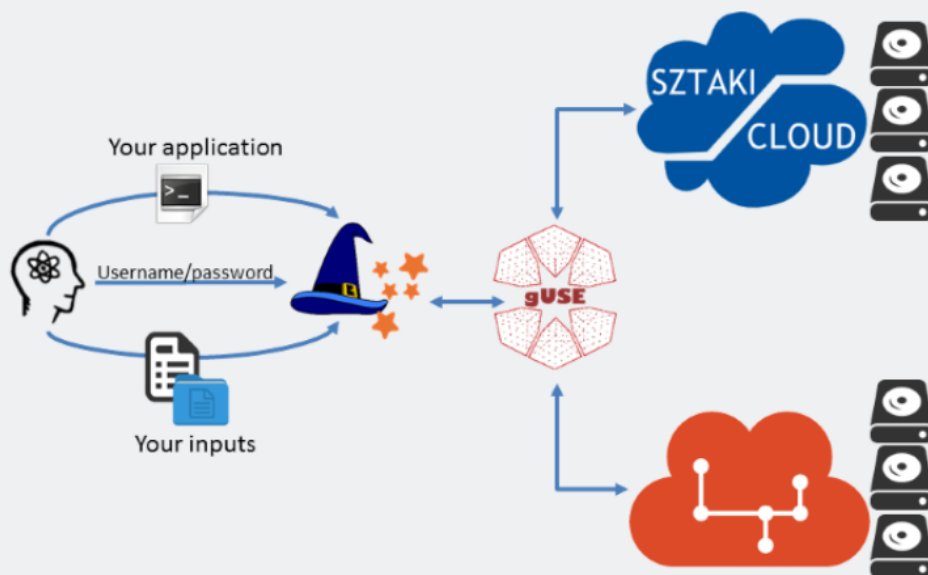


Welcome Job Wizard

EGI FedCloud Portal > Wizard > Welcome

Welcome to the Single-job wizard!

Pathway of your execution from you to the cloud



Run your application on the cloud by the gUSE Wizard following no more than 6 easy steps!

1. Upload your executable
2. Upload the proper inputs
3. Upload your parameter study inputs (optional)
4. Type your command line arguments (optional)
5. Pick a resource
6. Specify the names of the required output files

That's all!

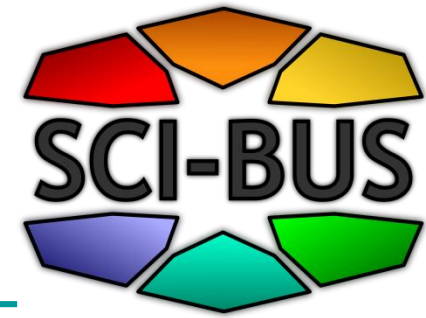
[Get started now](#)

Run your application in the gUSE workflow management system

[Get started now](#)

• See demo

Step 3 – Define parametric inputs



Job Wizard

1. Drop your executable
2. Drop your static inputs (optional)
- 3. Drop your parametric inputs (optional)**
4. Specify command line arguments (optional)
5. Define executing resource
6. Define name of produced output files (optional)

Your executable opens these files using this name:
ligands.tgz

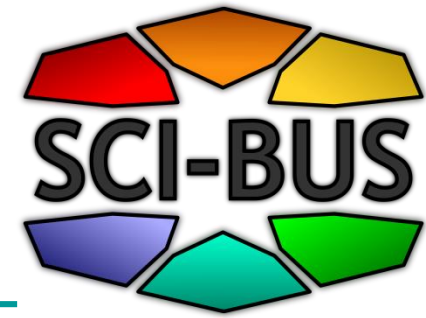
Add your parametric inputs below:

1.tgz ✓ 1.9 KIB Remove	2.tgz ✓ 1.9 KIB Remove	3.tgz ✓ 1.9 KIB Remove
4.tgz ✓	5.tgz ✓	6.tgz ✓

Previous Next

Your experiment list is still empty!
Drop an application and proper inputs using the wizard at the left hand-side.

Step 5 – Select resource and instance size



Job Wizard

1. Drop your executable
2. Drop your static inputs (optional)
3. Drop your parametric inputs (optional)
4. Specify command line arguments (optional)
5. Define executing resource
6. Define name of produced output files (optional)

Choose a resource where the experiment will be executed:

INFN-CATANIA-STACK (20)

Choose an instance size to run your executable:

Medium (At least 2 vCPUs, 4 GB RAM)

Previous

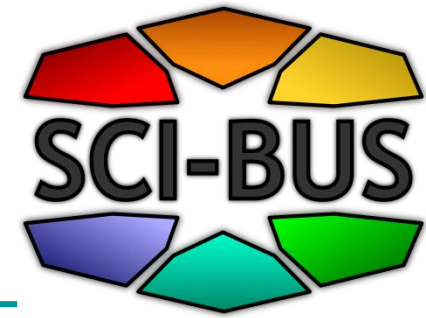
Next

Your experiment list is still empty!
Drop an application and proper inputs using the wizard at the left hand-side.

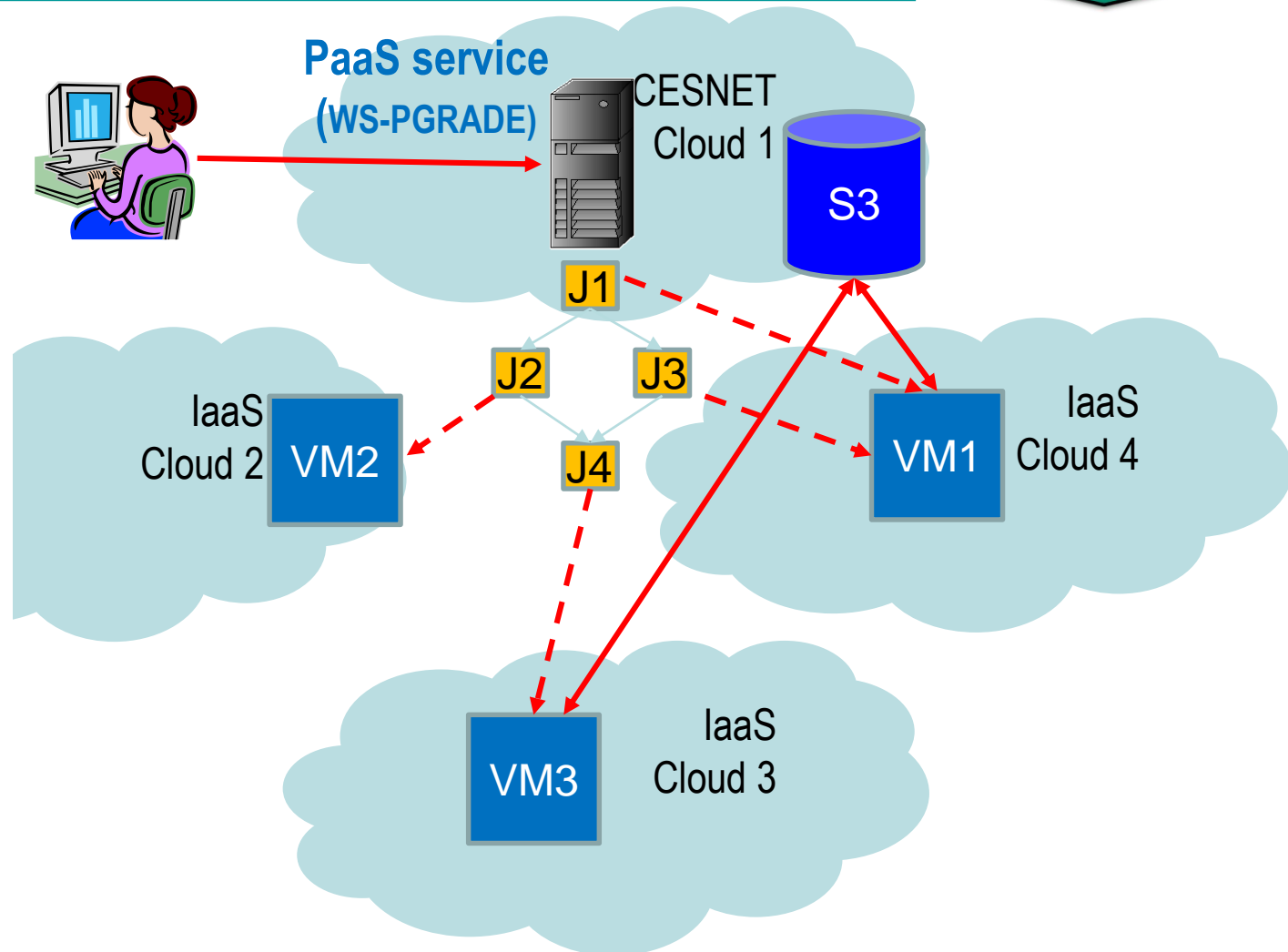
- Small instance: at least 1 vCPU, 2 GB RAM
- Medium instance: at least 2 vCPUs, 4 GB RAM
- Large instance: at least 4 vCPUs, 8 GB RAM

• Only at the long tail of science portal

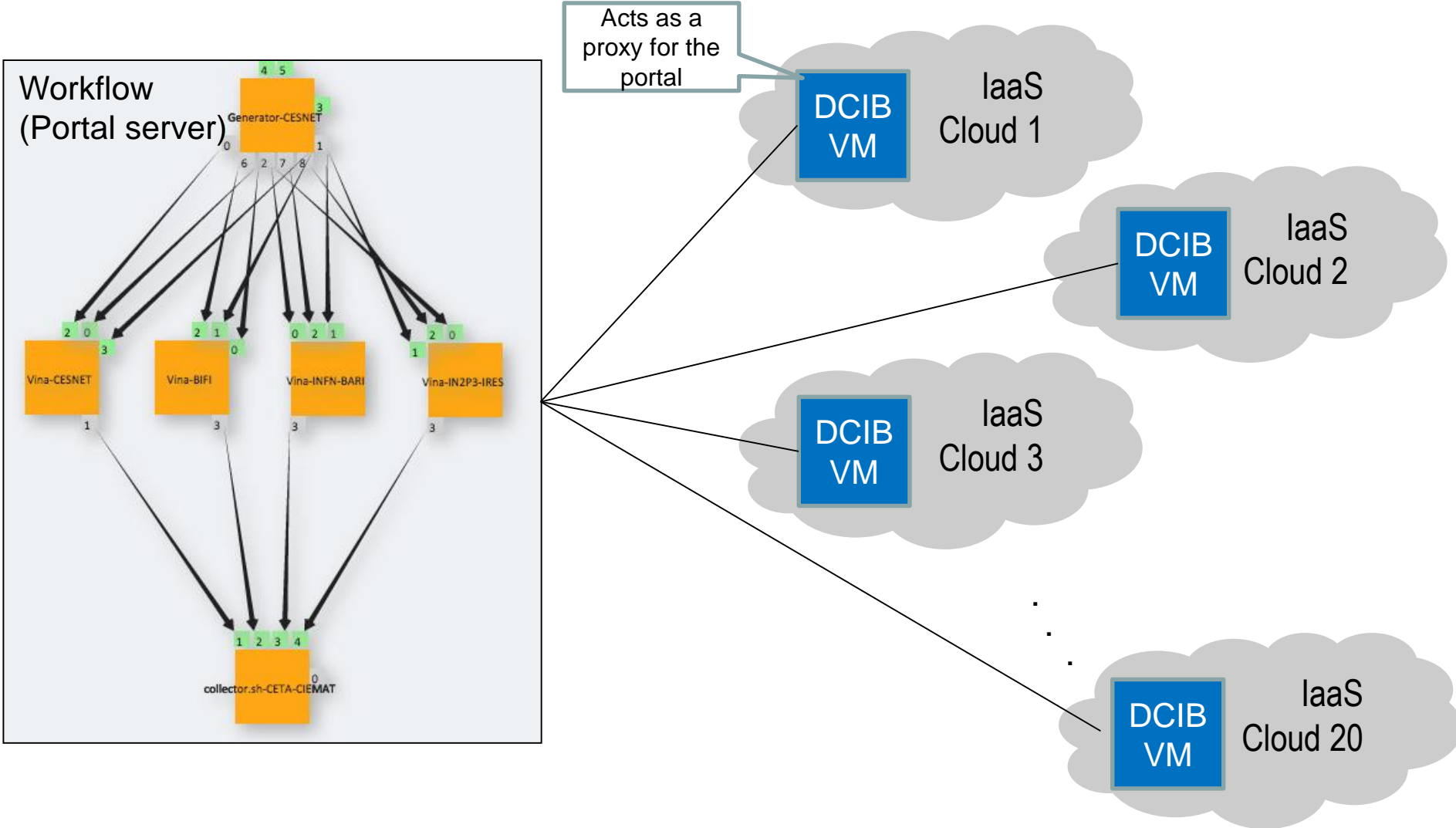
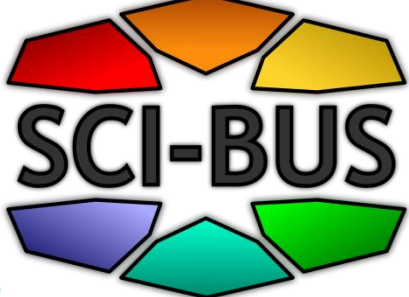
WS-PGRADE/gUSE for EGI FedCloud



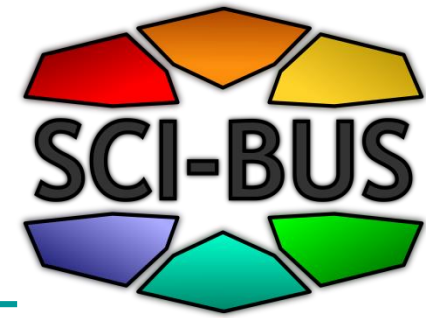
- User creates complex WF application that runs in the clouds of EGI FedCloud
- **WF execution uses new VMs on-demand from different clouds**



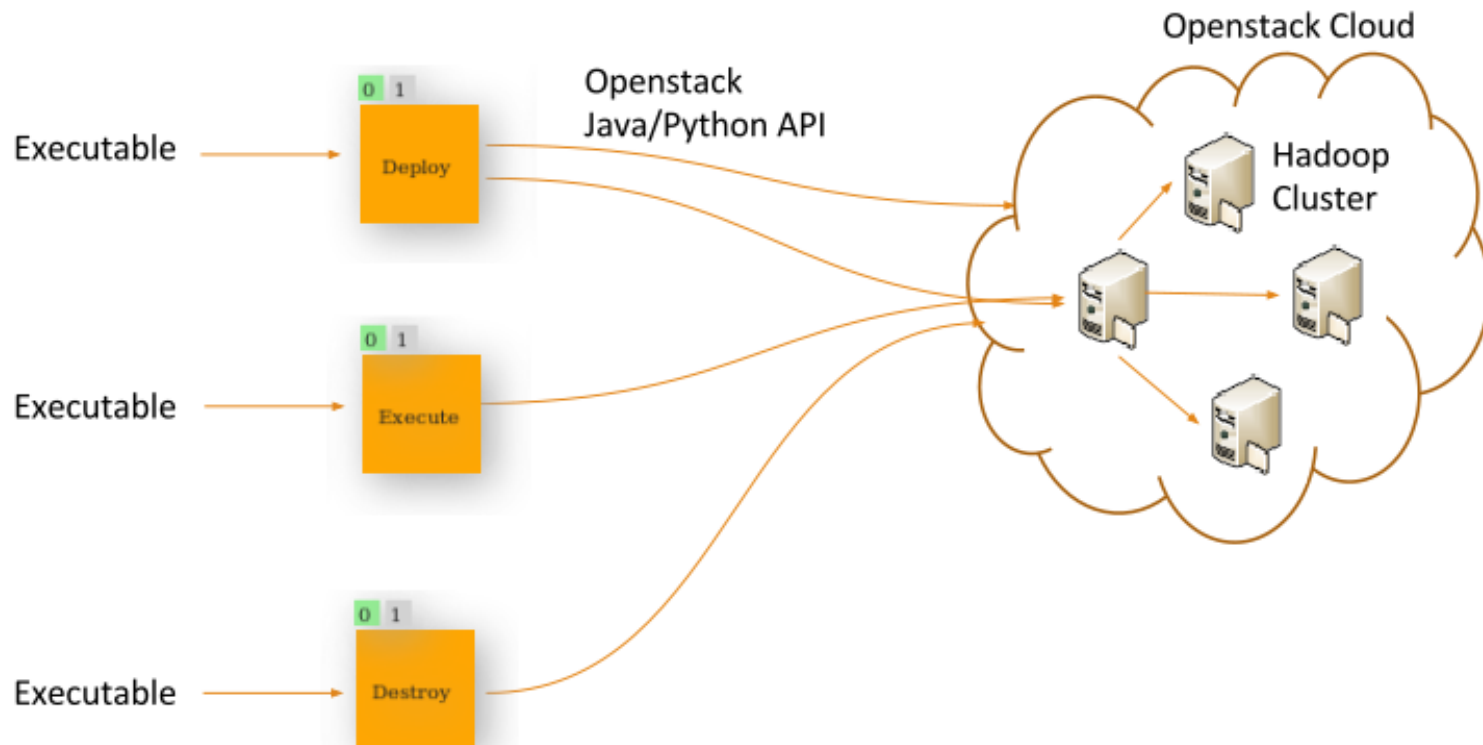
Mapping WFs to EGI FedCloud by WS-PGRADE/gUSE



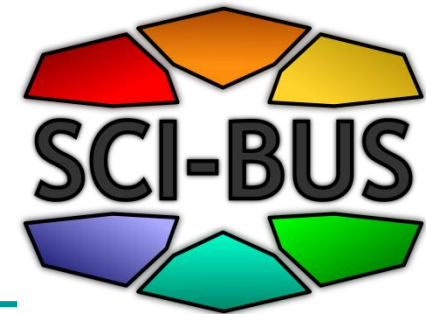
Support for hadoop/mapreduce applications



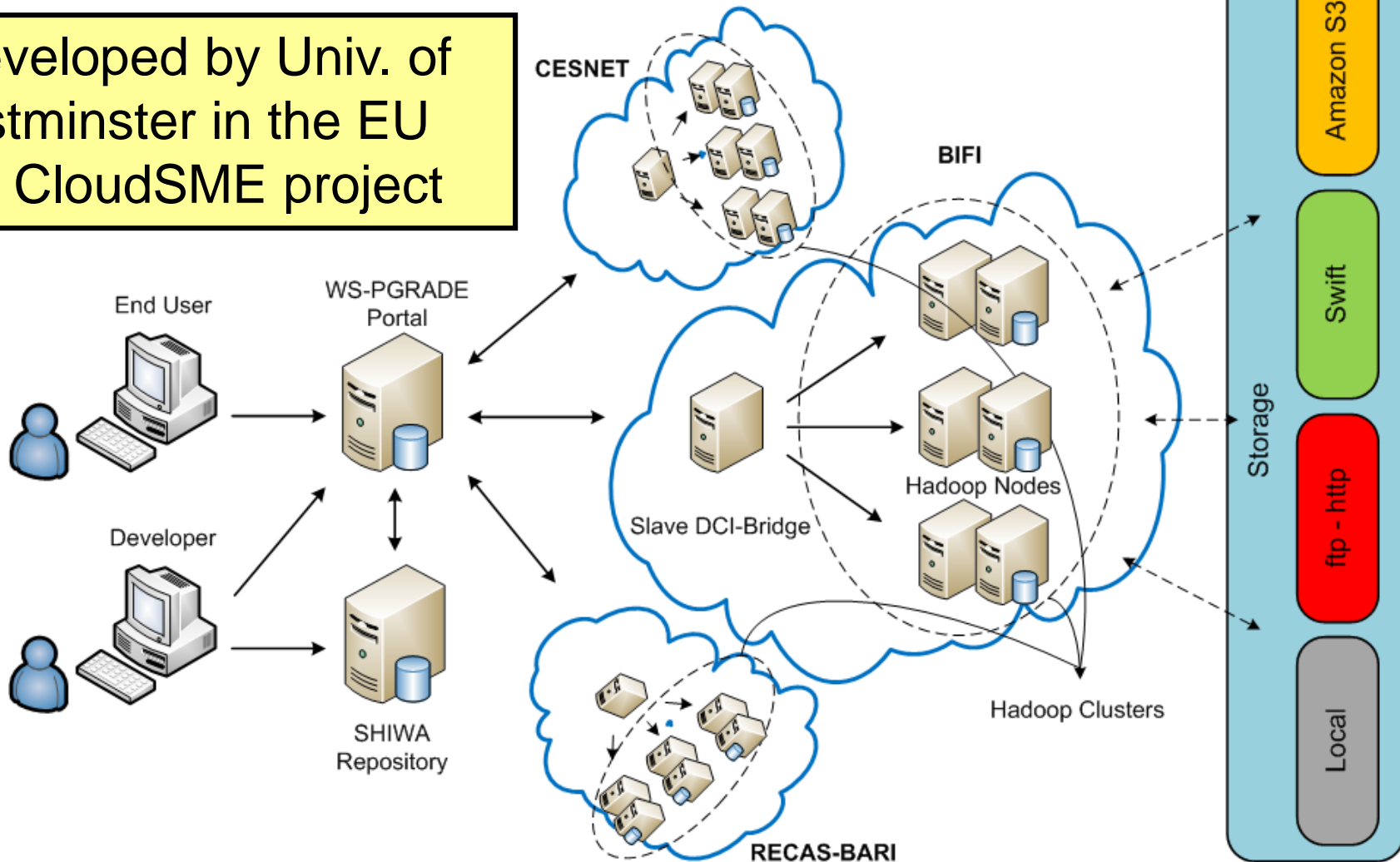
- Stage 1 (Deploy Hadoop Node): Launch servers in cloud, connect to master node, setup Hadoop cluster and save Hadoop cluster configuration
- Stage 2 (Execute Mapreduce Node): Upload input files and job executable to master node, execute job and get result back
- Stage 3 (Destroy Hadoop Node): Destroy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



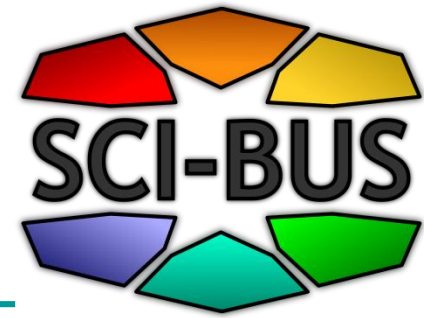
EGI FedCloud Implementation



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



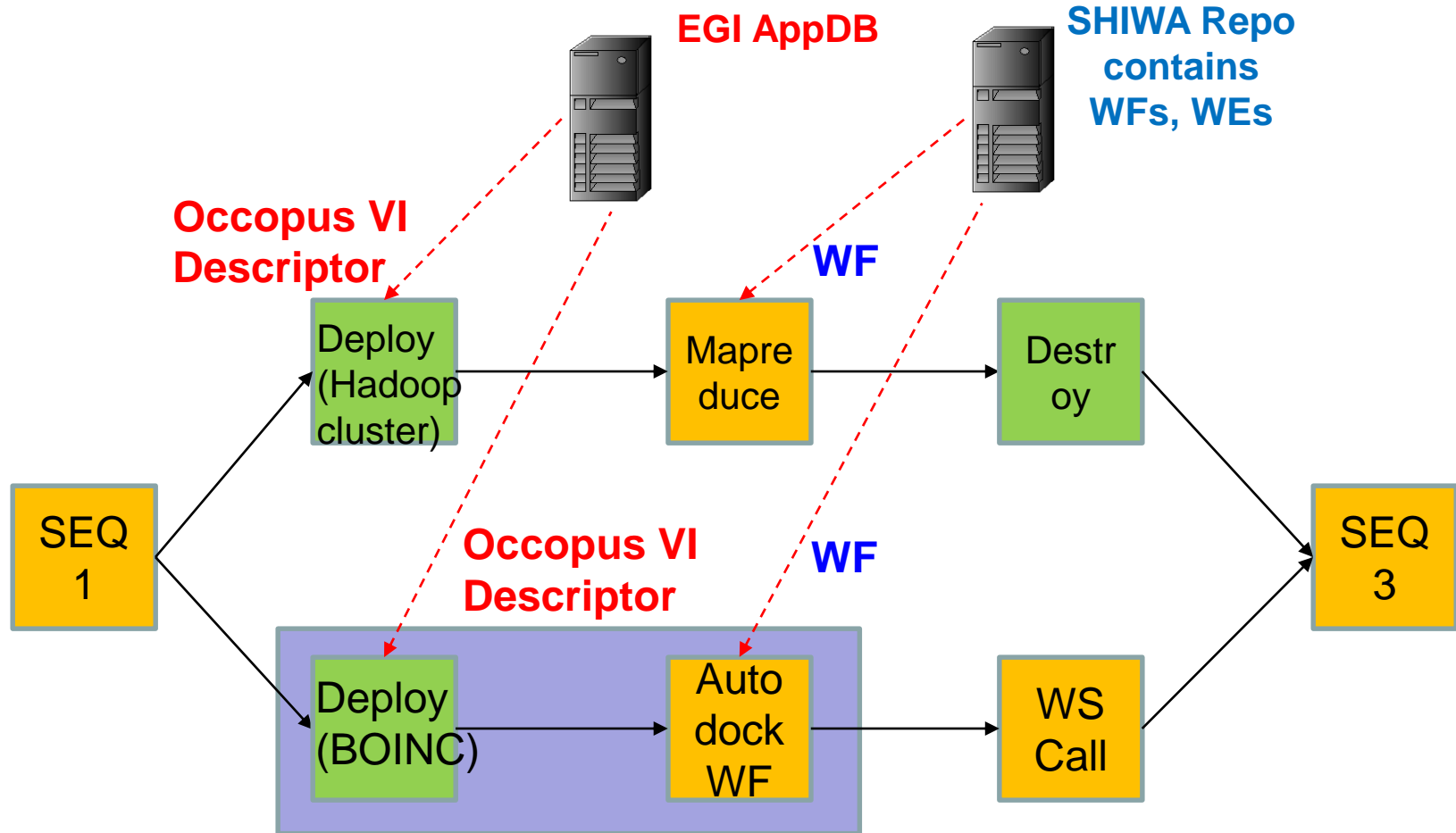
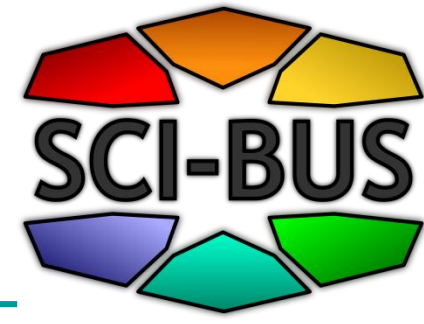
How to use it?



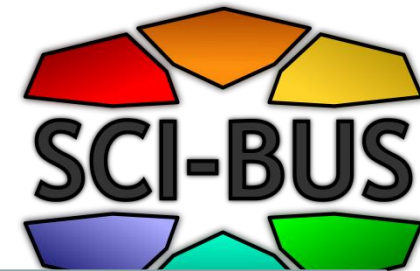
1. Create an account on the EGI FedCloud WS-PGRADE Gateway:
<https://guse-fedcloud-gateway.sztaki.hu/>
2. Import the Hadoop workflow(s) to your account from the SHIWA Workflow Repository
3. Download and customise sample configuration files
4. Configure workflow by uploading configuration files and Hadoop 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)*

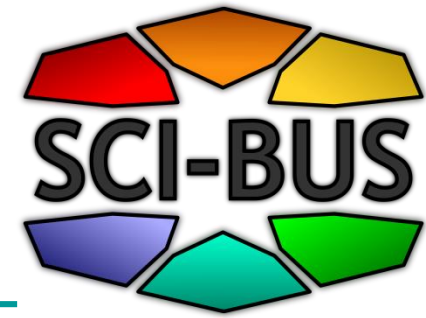
HTC
Infrastructures

HPC
Infrastructures

Large variety
of data
storages

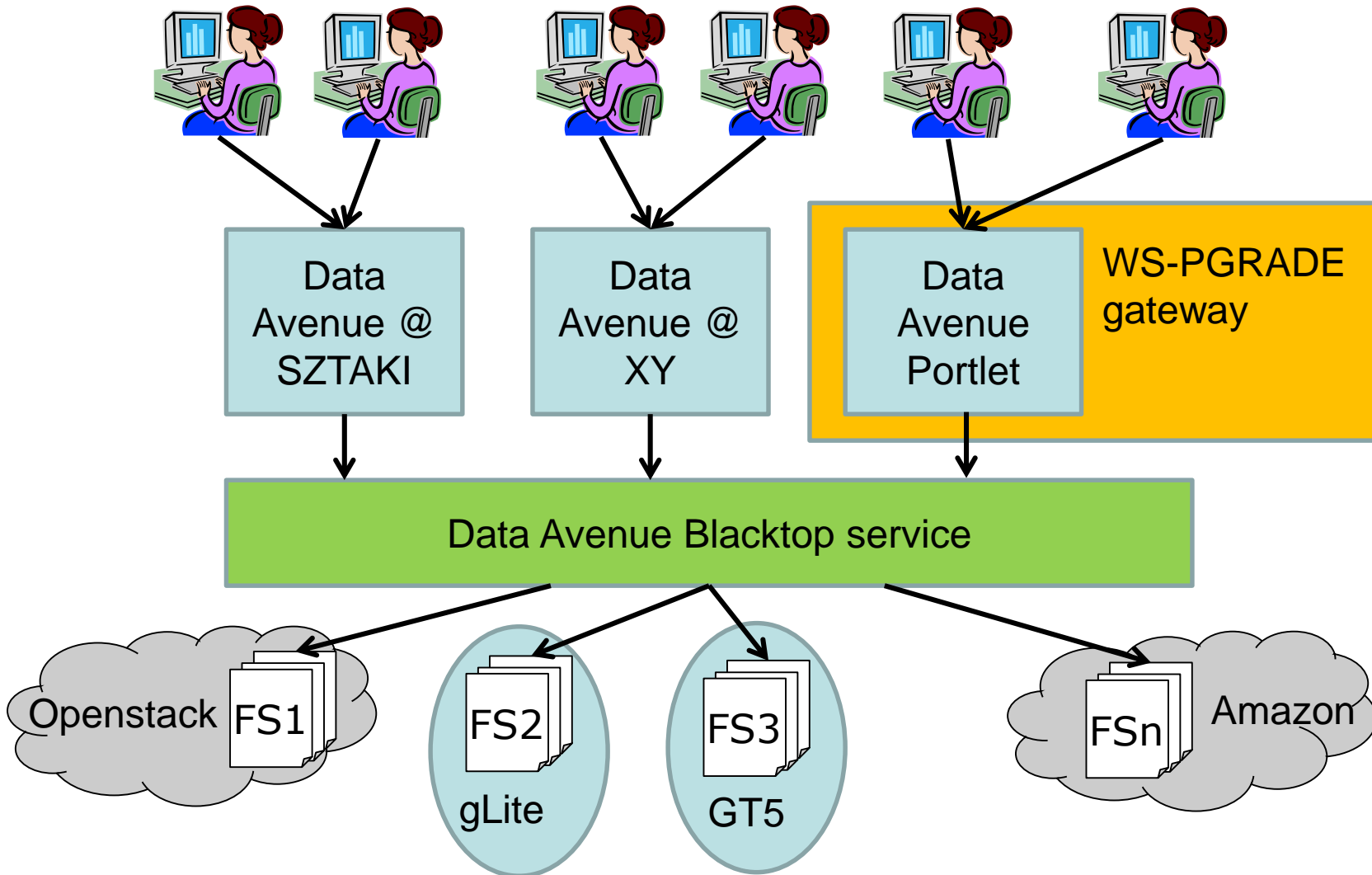
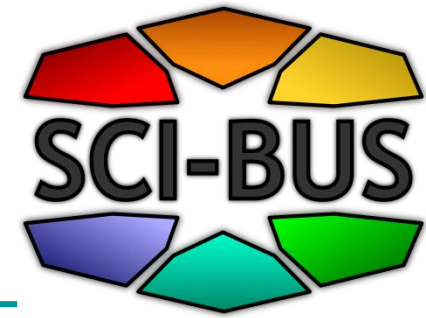
*Production
e-infrastructures*

Flexibility in data storage access

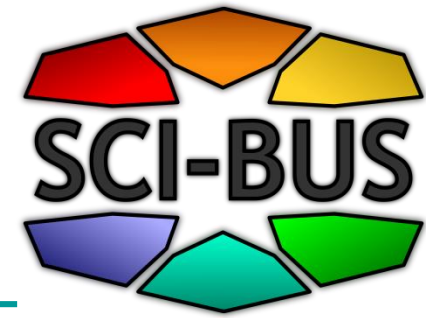


- **Use Data Avenue Blacktop service**
 - To access data storages in different DCIs
 - To transfer files among the storages of different DCIs
 - To upload/download files to/from the storages of different DCIs
- **Data Avenue Liferay portlet** to access the data transfer services of Data Avenue Blacktop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

Data Avenue services



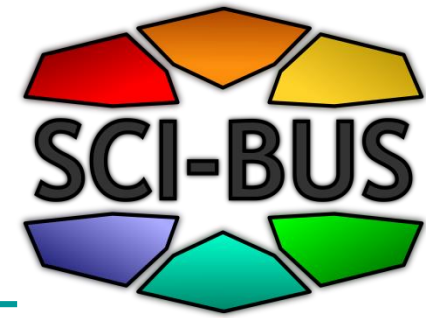
Data Avenue demo video



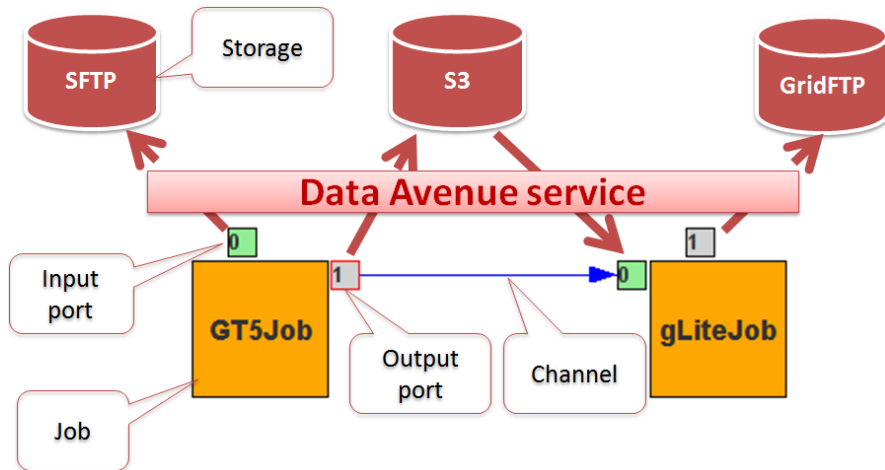
https://data-avenue.eu/en_GB/data-transfer-manuals

A screenshot of a web browser window. The browser's address bar shows the URL "https://data-avenue.eu/en_GB/data-transfer-manuals". The page content includes a menu with "File", "Edit", "View", "Favorites", "Tools", and "Help". Below the menu is a toolbar with icons for home, RSS, print, and other functions. The main content area displays a list of files, with "Upload (a file from your local hard drive to the current remote directory)" highlighted. Below this, a paragraph states: "Upload, copy and move operations are executed asynchronously. These operations may require Refresh command to update the directory contents in the corresponding side." A section titled "Data Avenue tutorial video" contains a video player. The video player shows a video titled "Data Avenue data transfer guide for beginners" with a thumbnail image of a hand pointing at a device. The video player interface includes a progress bar, a volume icon, and a "2:37 / 2:37" timestamp. The video player is set to play on YouTube. On the right side of the browser window, there are social media icons for Facebook, Twitter, and a globe icon. At the bottom of the browser window, a footer reads: "Data Avenue Blacktop v2.0.1 (b23/07/2014-10:07), MTA SZTAKI 2014 - Powered by Liferay".

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, *bypassing* the Blacktop service if the middleware is capable of handling the protocol



The screenshot shows the "Configure" window for a job named "Job0". The "Optional note" field contains "Description of Job". Below this are icons for "Job Executable", "Job ID", "IDLRSL", and "History". The "Port Number 0" section is expanded, showing "Port Name: PORT0" and "Description of Port". The "Input Port's Internal File Name" is "PORT0". The "Port dependent condition allowing the run of the job:" section contains a "View Hide" button and a "Data Avenue" icon. The "Source of input directed to this port:" section shows a "URL" field with a masked address and an "Authentication id: 1000-1334". A "Browse" button is next to the URL. The "Parametric Input details:" section has a "View Hide" button. A red arrow points from the "Data Avenue" icon in the screenshot to the "Data Avenue service" box in the diagram above.

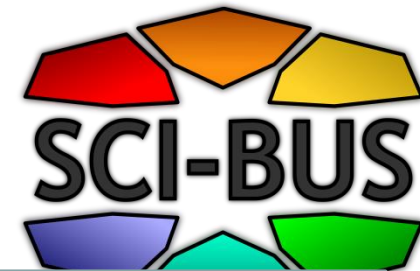
The "Port Number 1" section is collapsed, showing "Port Name: PORT1" and "Description of Port".

Below the configuration window is a "Configure" window showing a "Job's name" field with "Job0" and an "Optional note" field with "Description of Job". Below this are icons for "Job Executable", "Job ID", "IDLRSL", and "History". The "Port" section is expanded, showing a "First Step", "Second Step", and "Third Step". Below this is a table with columns "Name", "Size", and "Last modified".

Name	Size	Last modified
extras		
apache-tomcat-6.0.39-deployer.tar.gz	987.0 KB	28.01.2014 00:01:33
apache-tomcat-6.0.39-deployer.zip	989.9 KB	28.01.2014 00:01:33
apache-tomcat-6.0.39-fulldocs.tar.gz	3.5 MB	28.01.2014 00:01:33
apache-tomcat-6.0.39-windows-x64.zip	8.3 MB	28.01.2014 00:01:33
apache-tomcat-6.0.39-windows-x64.zip	7.8 MB	28.01.2014 00:01:33
apache-tomcat-6.0.39-windows-x86.zip	7.7 MB	28.01.2014 00:01:33
apache-tomcat-6.0.39.exe	7.7 MB	28.01.2014 00:01:33
apache-tomcat-6.0.39.tar.gz	6.7 MB	28.01.2014 00:01:33
apache-tomcat-6.0.39.zip	7.1 MB	28.01.2014 00:01:33

At the bottom of the window are "Previous" and "Finish" buttons.

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)*

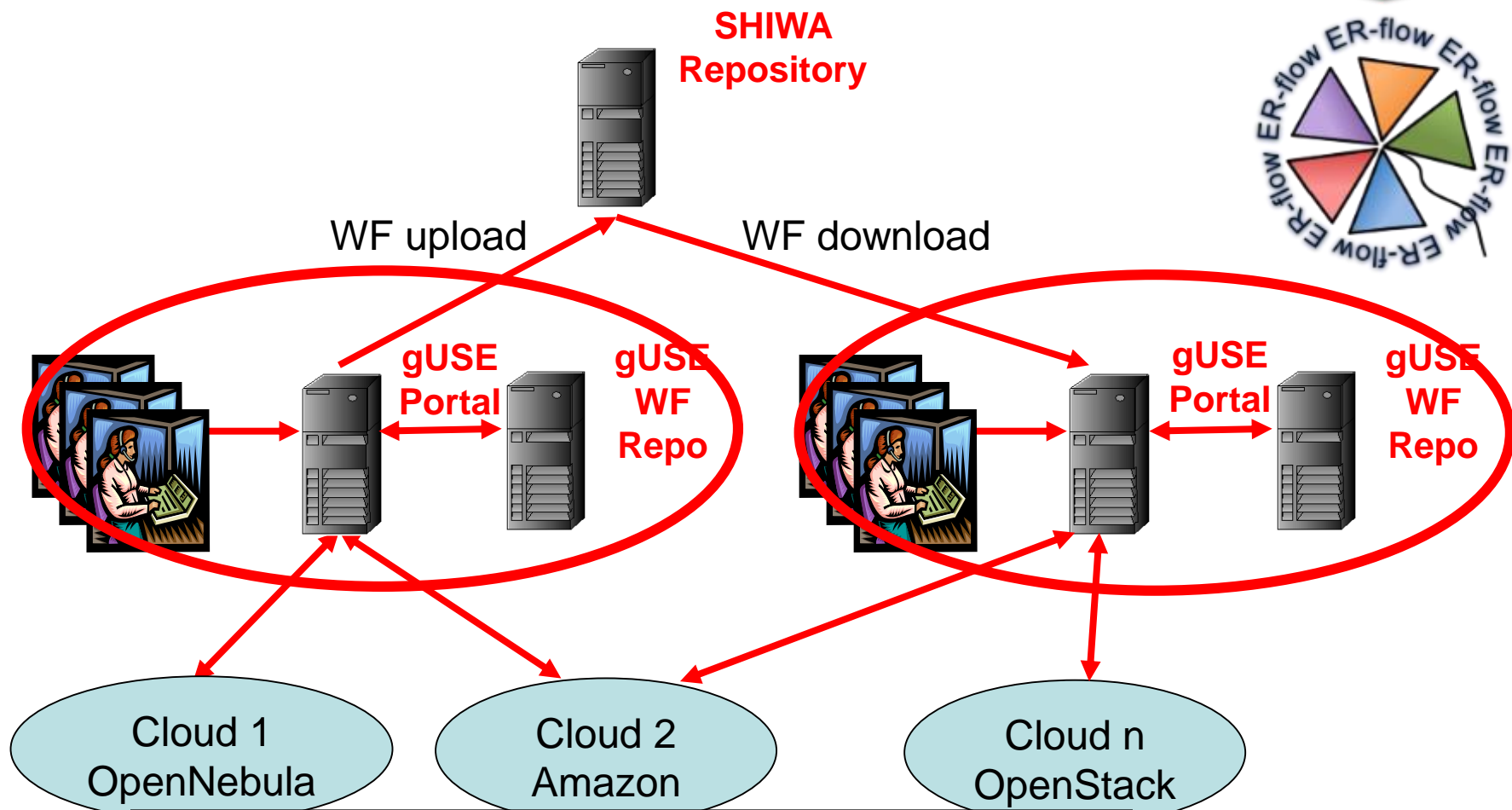
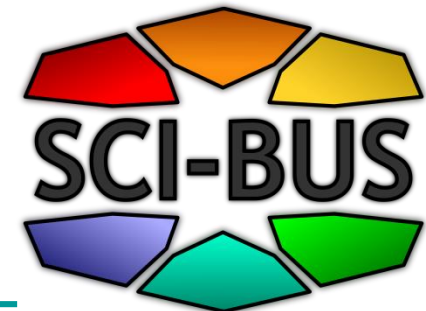
HTC
Infrastructures

HPC
Infrastructures

Large variety
of data
storages

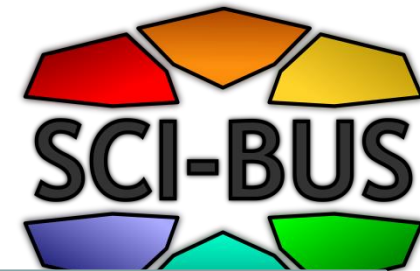
*Production
e-infrastructures*

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**

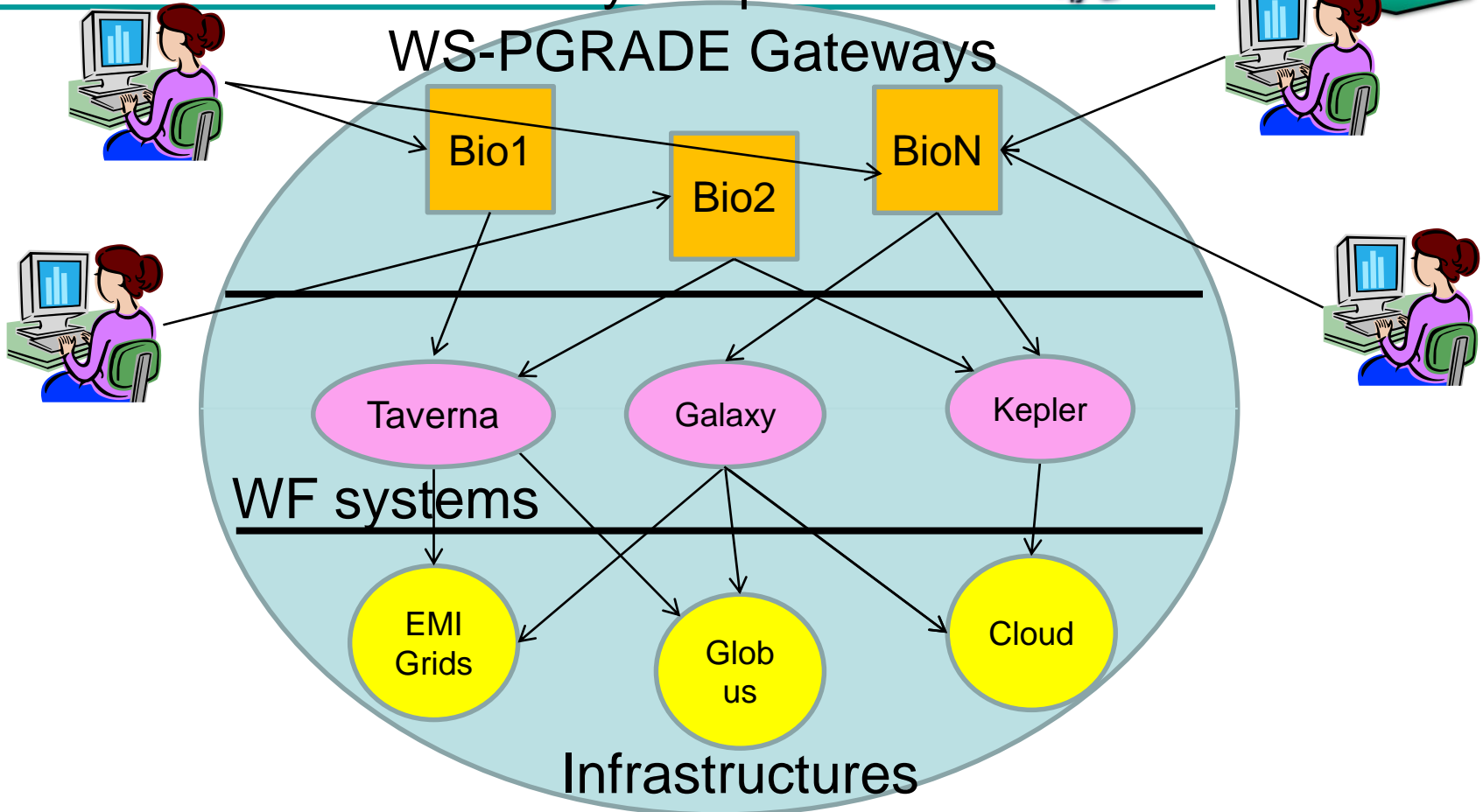
*Production
e-infrastructures*

Flexibility in using different workflow systems



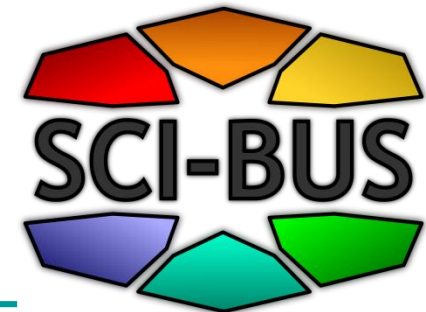
Cyberspace

WS-PGRADE Gateways



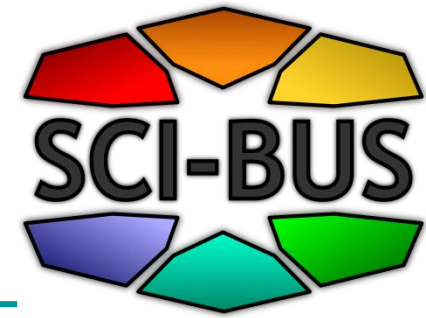
Combining SCI-BUS and SHIWA technologies (supported by ER-Flow) users can access and use many WFs and many infrastructures in an interoperable way no matter which is their home WF system

Flexibility of gateway types and user views

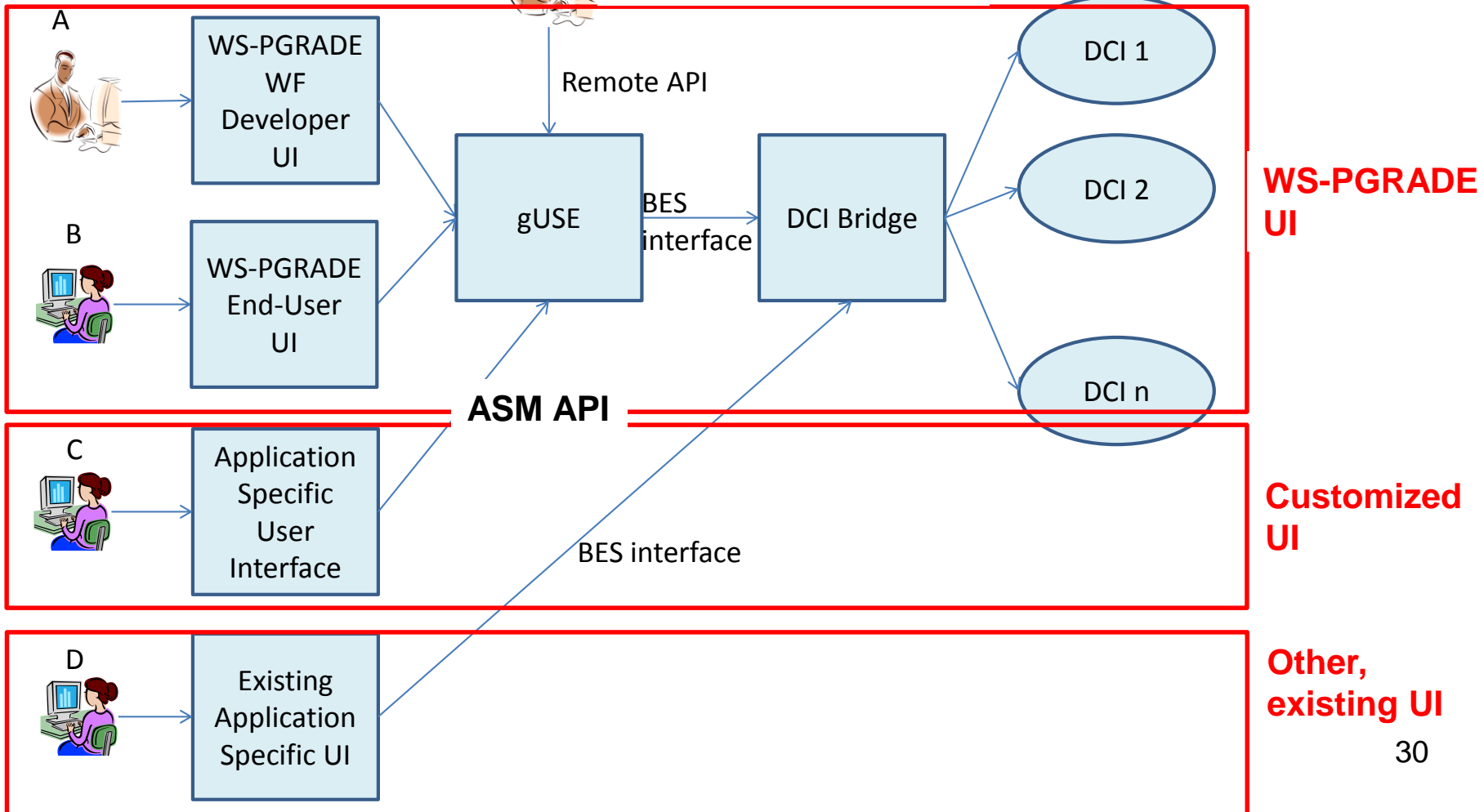


1. **Generic purpose gateways for clouds (workflow view)**
 - Core WS-PGRADE/gUSE (e.g. Greek NGI)
2. **Generic purpose gateway for specific technologies (workflow view)**
 - SHIWA gateway for workflow sharing and interoperation
3. **Domain-specific science gateway instance**
 - Autodock gateway (end-user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

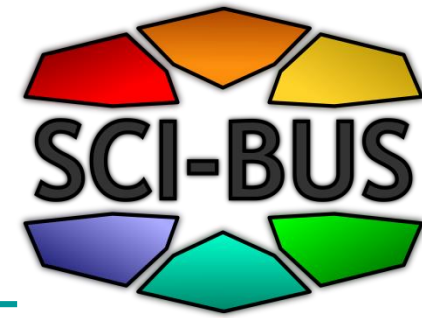
Flexibility in user access modes



WS-PGRADE workflow UI

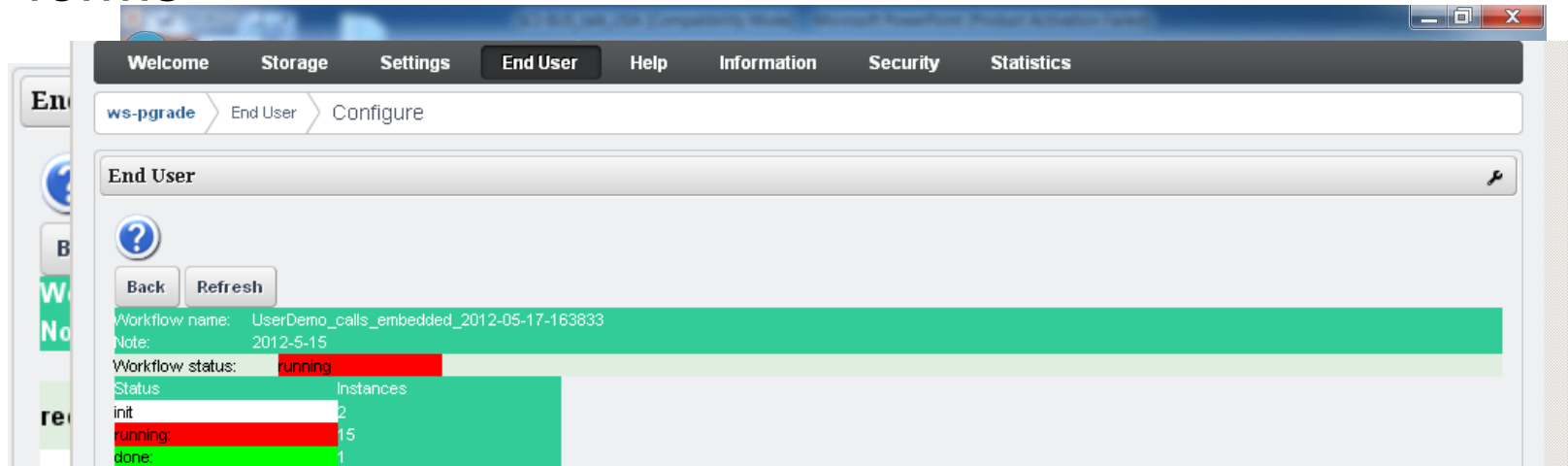


End-user view based gateways



What is required from the end-user?

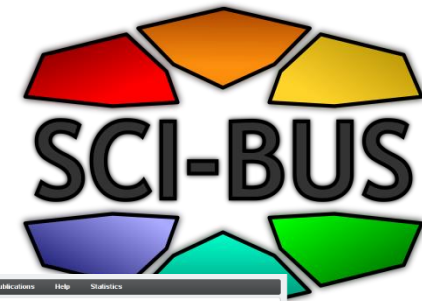
- Import workflow from repository
- Customise, execute and monitor application using simple web forms



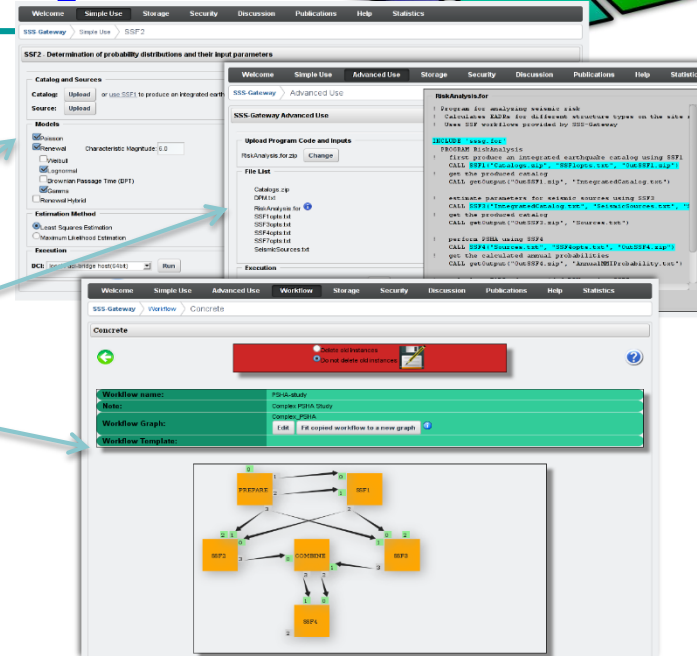
What needs to be done by the workflow developer?

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

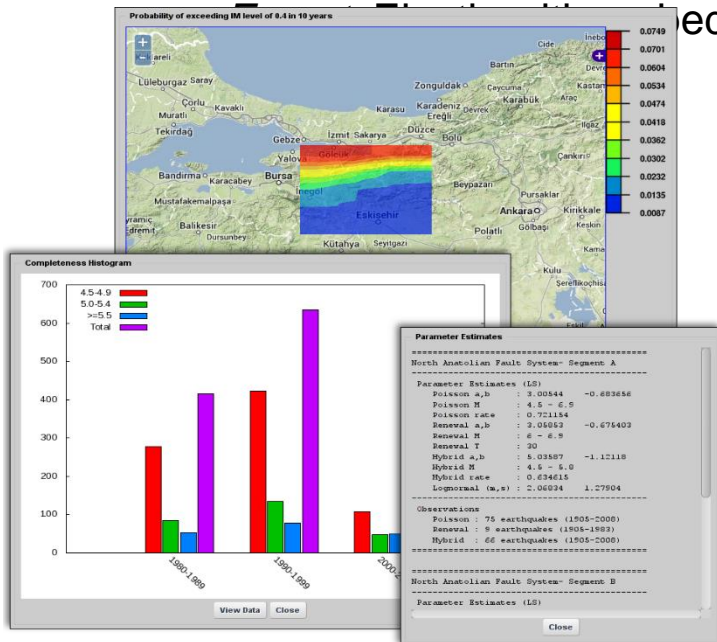
Using ASM API: Statistical Seismology Science Gateway



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



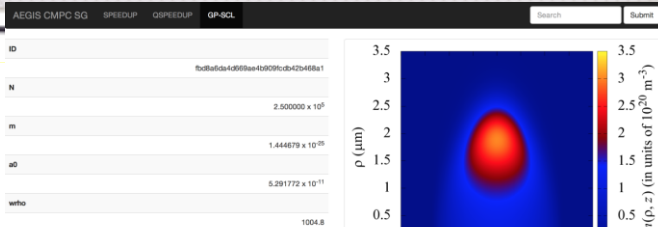
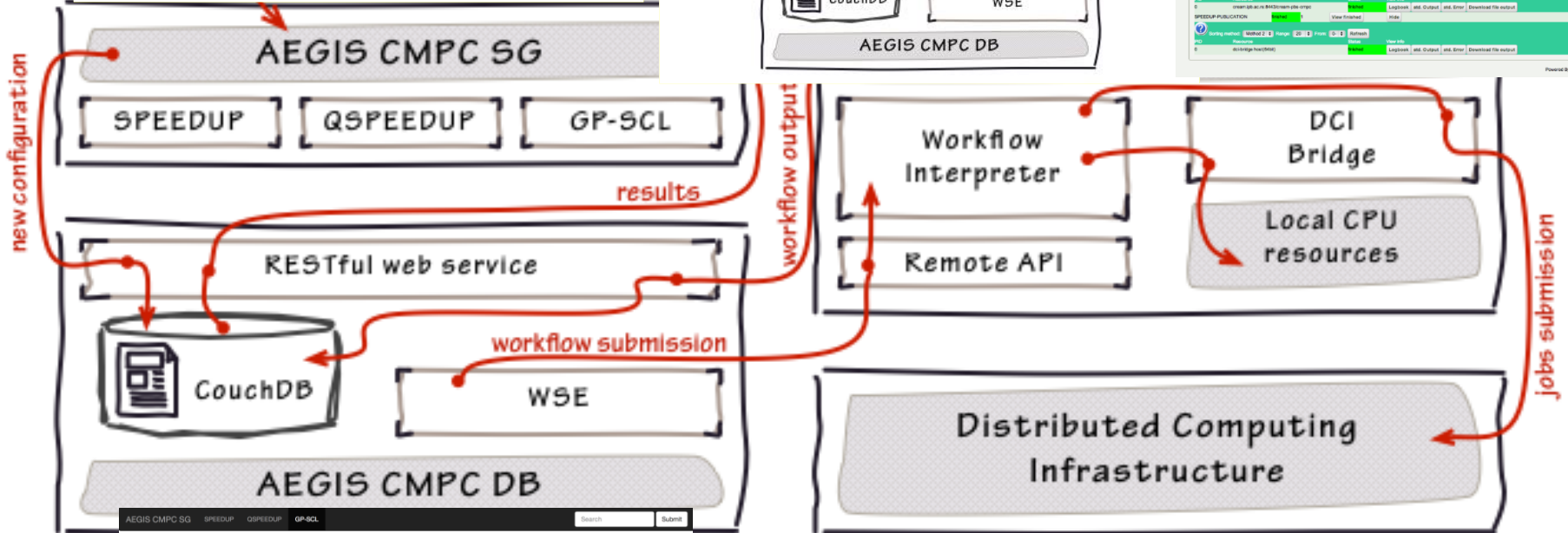
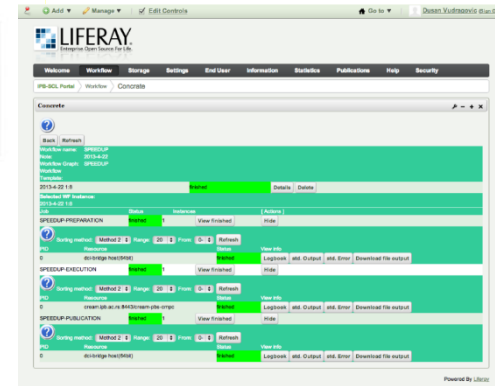
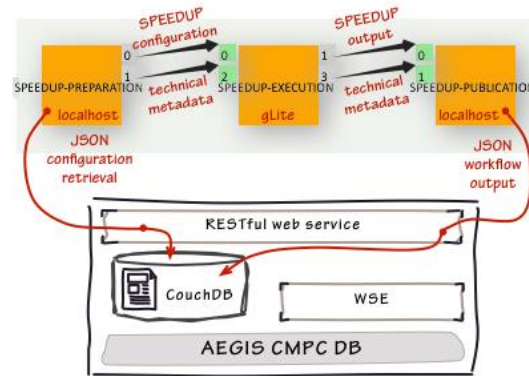
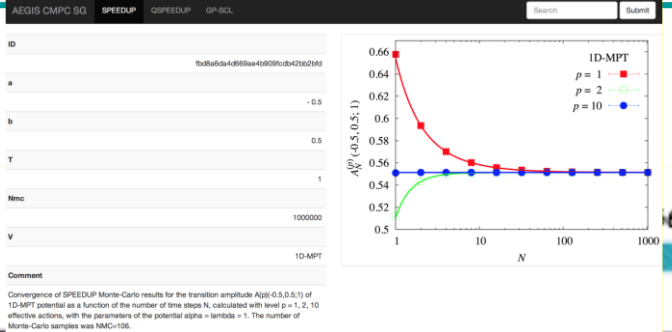
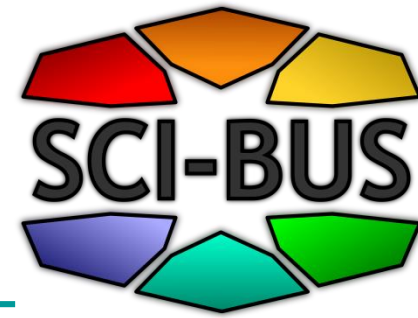
Embedded workflows



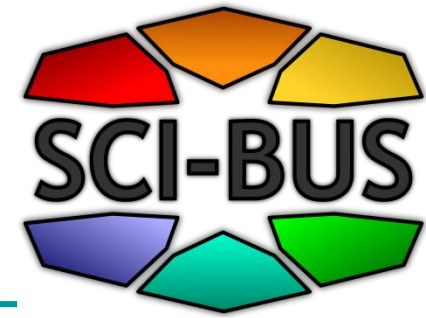
- Based on gUSE with ASM and Remote API
- Various methods/models from “*integration of multi-source data*” to “*generation of hazard maps*”

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

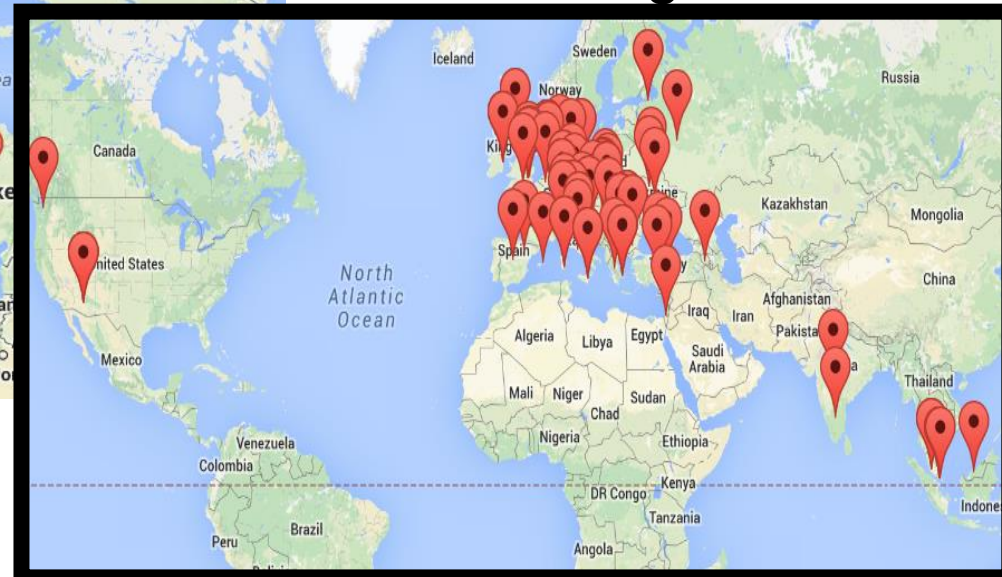
Using Remote API: AEGIS CMPC SG Institute of Physics Belgrade



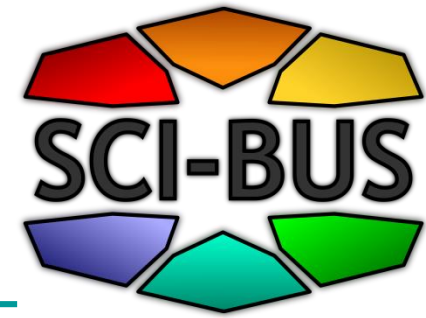
Technology adaptation



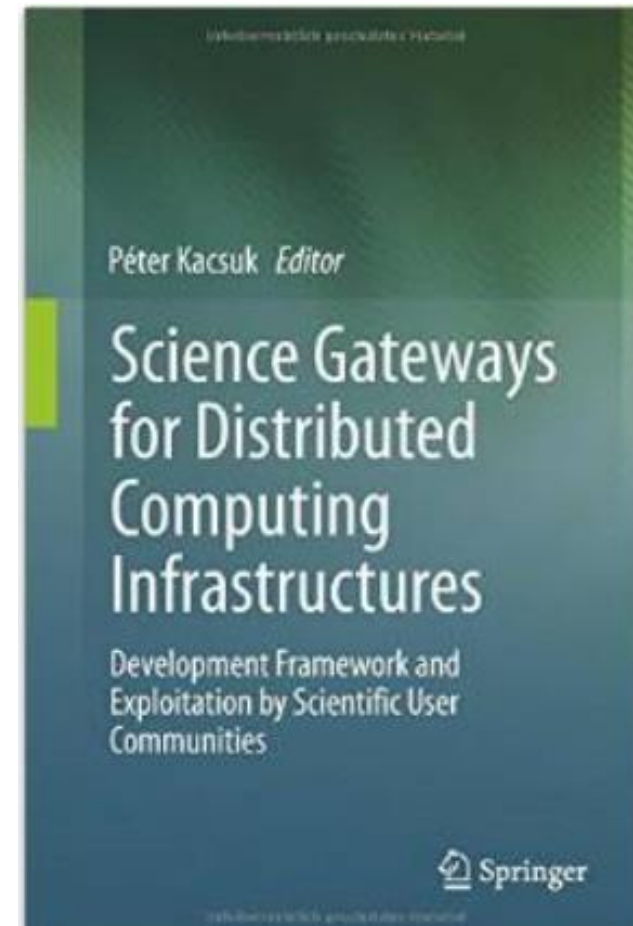
- More than 100 deployments world-wide
- Nearly 22.000 downloads from 80 countries on sourceforge



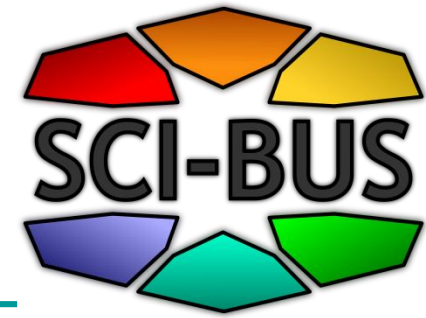
Further readings



- 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