A WorkFlow Management System for Bioinformatics Grid

Giovanni Aloisio, Massimo Cafaro, Sandro Fiore, Maria Mirto

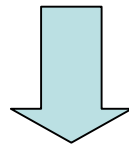CACT/ISUFI SPACI, University of Lecce and NNL/INFM & C N R , Italy

OUTLINE

- Bioinformatics Grid and Web Service

- ProGenGrid Project

- Workflow Management System
  - ✓ Editor

  - ✓ Enactment Service

- Conclusions and future work

Why Bioinformatics Grid?

The need of large scale computational power and a collaborative environment in Bioinformatics

LS-RG (Life Science Research Group) of GGF, defines a BioGrid:

- Deployment, distribution, management system of needed software components;

- Harmonized standard integration of various software layers and services;

- Powerful, flexible policy definition, control and negotiation mechanism for a collaborative grid environment

Service Oriented Architecture

Web/Grid Service

– XML,SOAP,WSDL, UDDI

– Grid

• OGSA & WSRF

(Open Grid Service Architect

& Web Service Resource Framework)


Public Processes

Web Service
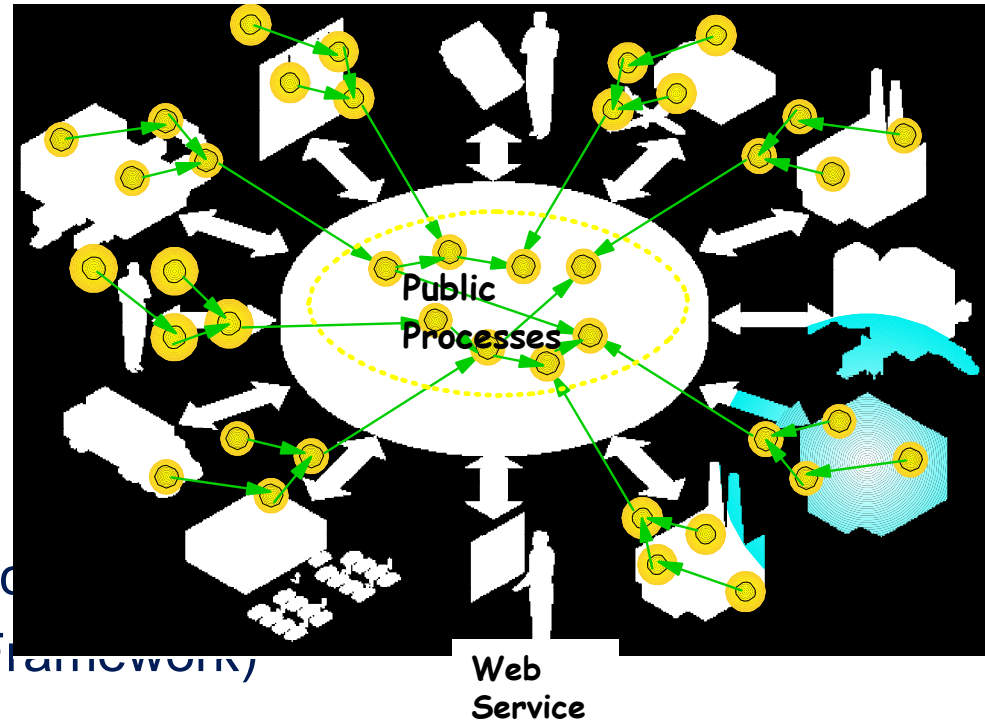
Allows building enhanced services independently of platform, programming language, tools, and network infrastructure.

ProGenGrid Project

The aim of the ProGenGrid project is the creation of a distributed and ubiquitous grid environment (a virtual lab) for supporting "in silico" experiments in bioinformatics.

Using such an environment, that can be considered as a virtual laboratory, the e-scientists will access

- analysis tools (e.g. EMBOSS, Blast),
- biological databases (e.g. GenBank, Protein Data Bank),
- visualization tools (e.g. Rasmol)
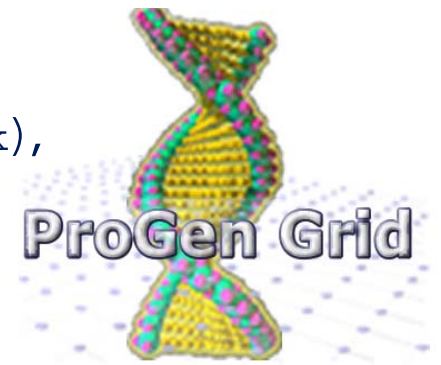
These tools will be available as Web/Grid Service according to a Service Oriented architecture and accessible through a Web Portal.

WorkFlow Management Systems
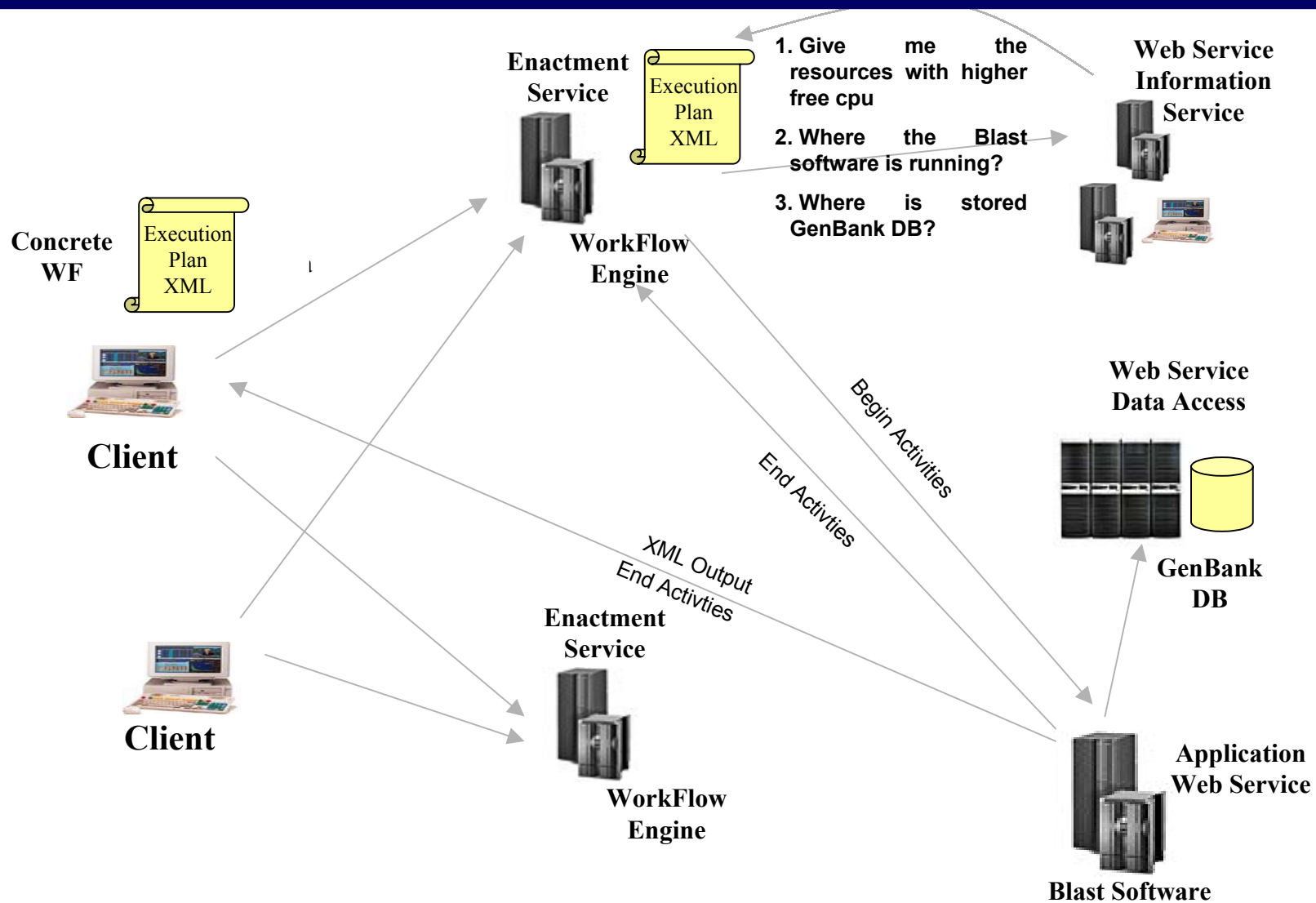
The WorkFlow Management Coalition defines workflow as:

"*The automation of a business process, in whole or part, during which documents, information or tasks are passed from one participant to another for action, according to a set of procedural rules*."

According to our vision, a Grid Workflow is a Workflow which the tasks are Web/Grid Service components and the chose of the invocation of them is made in order to obtain a given level of performance (cpu load, memory, etc.) supplying also data, resources and applications needed for the execution of the experiment.
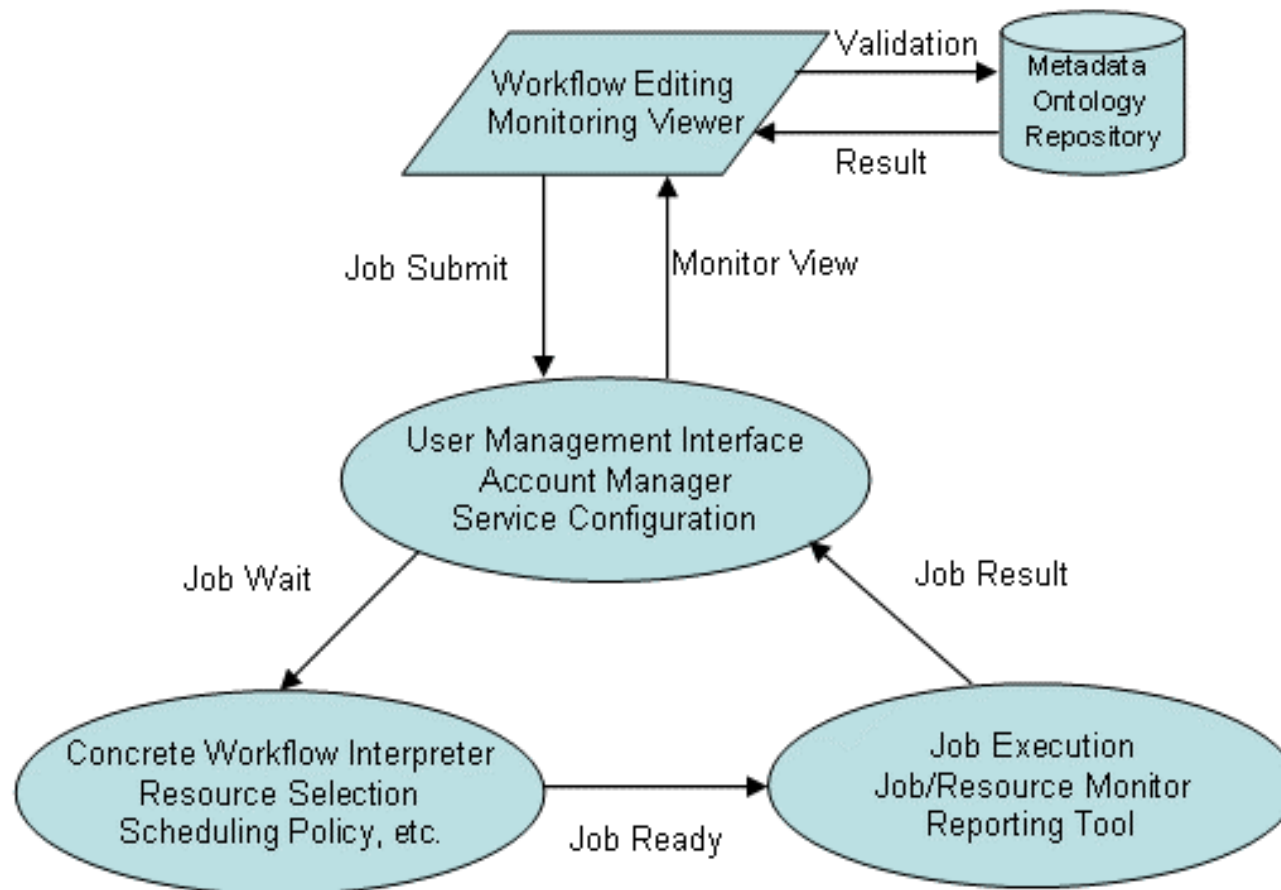
WorkFlow Management Coalition (WFMC)
- founded in 1993, 24 countries, 170 members
- terminology, standard interfaces, promotion

Scenario

Enactment Service

Execution Plan XML

Concrete WF

Execution Plan XML

Client

Web Service Information Service

1. Give me the resources with higher free cpu

2. Where the Blast software is running?

3. Where is stored GenBank DB?

WorkFlow Engine

Begin Activities

End Activties

Web Service Data Access

GenBank DB

XML Output
End Activties

Enactment Service

WorkFlow Engine

Client

Application Web Service

Blast Software

Workflow Architecture

Workflow Editor: features

✓UML as a graphical formalism for modelling the experiment and describing it in detail:

wait	A	A ▼	❘ —	◇	●	◉
Wait node	Atomic node	Compound node	Fork/join	decision/merge	start	final
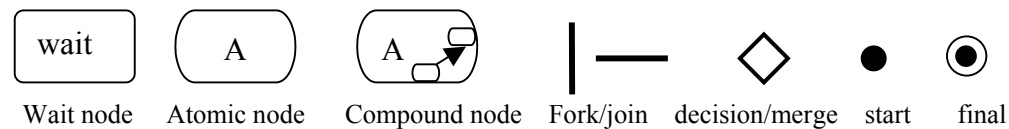
✓An XML-based language for specifying the job flow for biological applications (abstract workflow) which are not developed for grid environments;

✓An ontology of software for the bioinformatics domain, in order to validate the experiment during its preparation.

Workflow Enactment

- Applications consist of a number of components linked together in a dataflow manner
- User specifies work as abstract workflow
- The abstract workflow needs to be mapped down to a set of component implementations which will run on resources (concrete workflow)
 - Mapping Workflow Graph over Resource Graph

Issues

- We have:
  - Multiple resources where components can run
  - Multiple implementations of components
  - The choice of one resource component mapping can affect the others

- How to choose the "best" mapping of workflow over resources?

Solution

- We need to take into account:
    - Some parameters such as CPU load, memory, etc., by leveraging iGrid and NWS information systems
    - Inter-component effects of workflows
        - Workflow Scheduler by using GRB (Grid Resource Broker) for launching the applications
    - Workload on resources, for optimizing the execution of an application

Enactment Pipeline
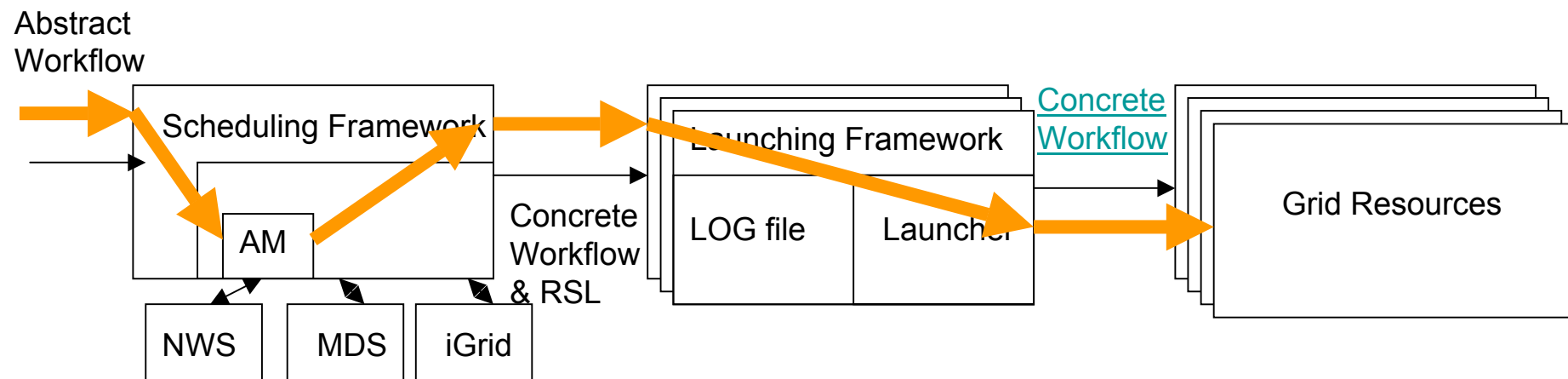
iGrid Architecture
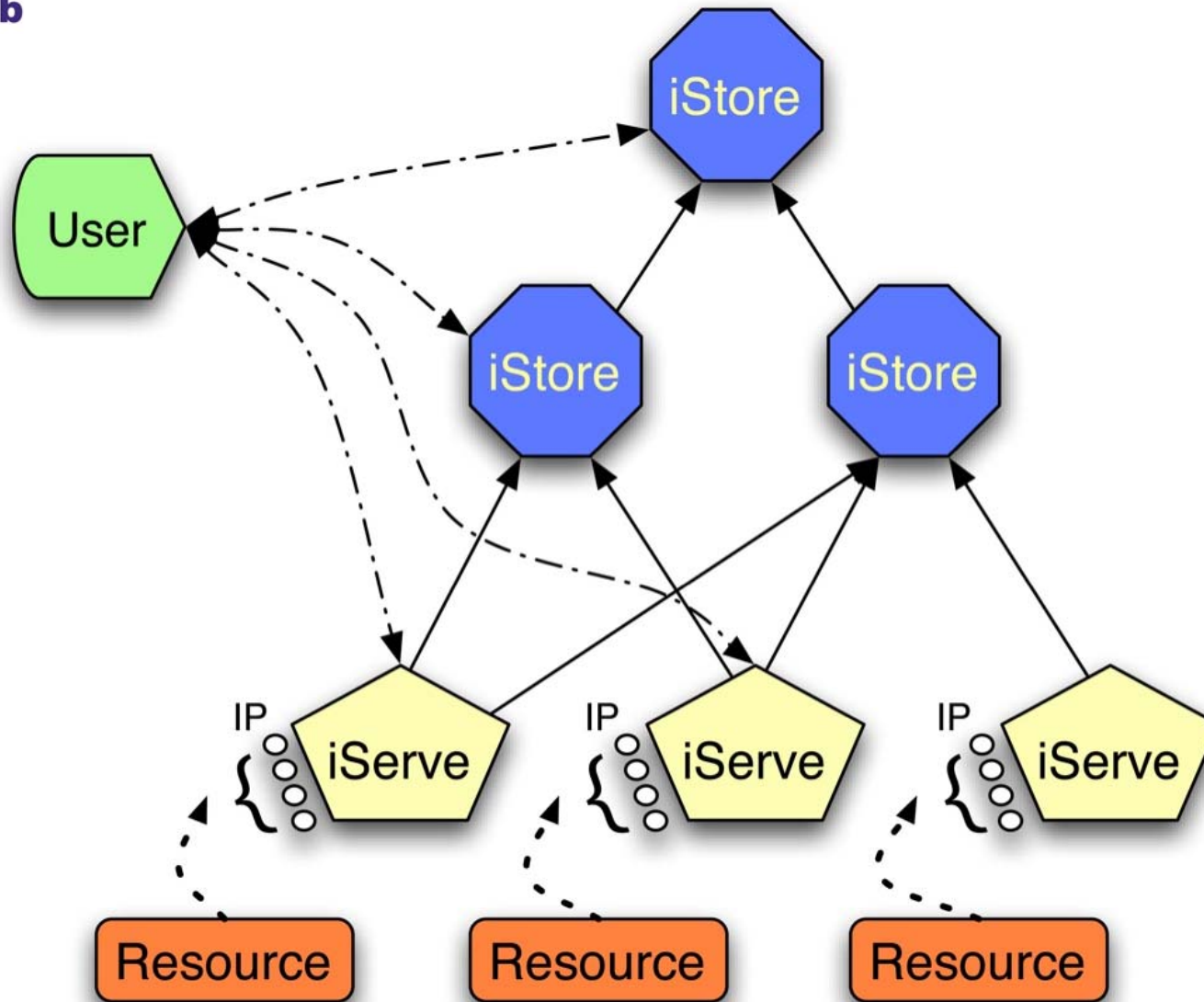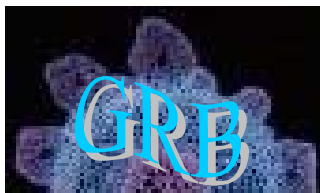
The G R B Architecture

First Tier

- -

Second Tier   G R B Web Server   My Proxy Server

GRB Libraries	GRB Web Services

Security	Info	Jobs	File/Data

GSI	M D S	G R A M	GridFTP

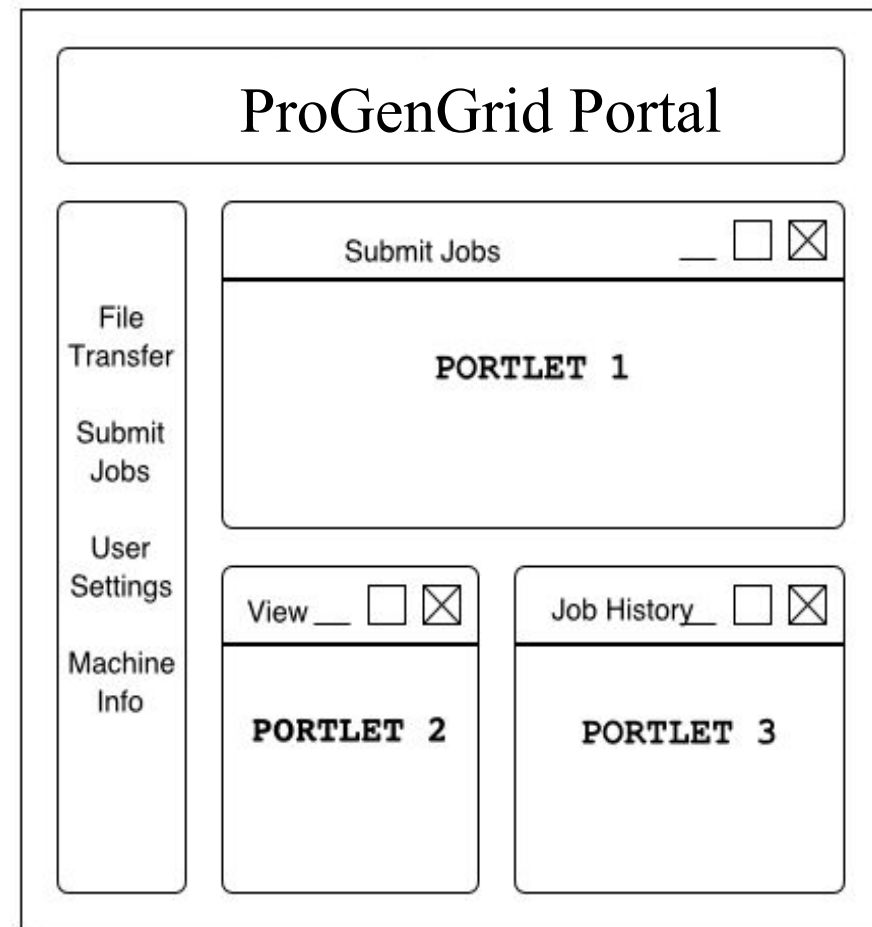- -

Third Tier   a user's grid

GUI Portal

- Portlets define how to construct and deliver Web content as modular components within a Web page.
- Portlets can be "maximized" or "minimized" within a Web page.
- Portlets support various modes
  – View, Edit, Help, Configure
- Users can choose to which portlets they want to be "subscribed".

Conclusions and future work

- A Workflow Management Systems ontology driven has been developed
- Further research is required in scheduling algorithms and reservation techniques
- Implementation of the Web interface both applications and workflow through the Portlet technology
- Provides a framework for experimentation with
  - Different scheduling algorithms
  - A performance model
  - Different reservation policies
- Comparison with other WFMS (Taverna, Pegasys etc.)
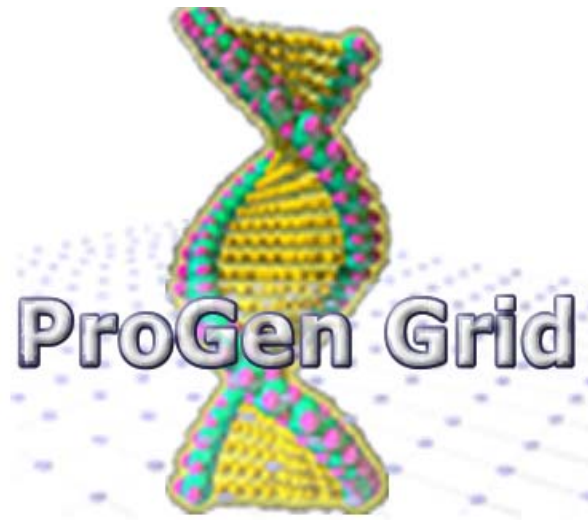
For any information

To contact authors:
Giovanni Aloisio {giovanni.aloisio@unile.it}
Massimo Cafaro {massimo.cafaro@unile.it}
Sandro Fiore {sandro.fiore@unile.it}
Maria Mirto {maria.mirto@unile.it}

About ProGenGrid Project
Director: Prof. Giovanni Aloisio
Project P. I. : Maria Mirto
WebSite : **http://www.spaci.it**