Grid for Biomedical Applications

Ana Lucia DA COSTA
HealthGrid, France

ACGRID School

KualaLumpur, 10 Nov 2009

HealthGrid

- Guiding and promoting the grid technologies and their applications into health research and development

- Make successful the regular and worldwide use of healthgrids at the industrial level

Introduction

- **Until 2003: Pre-genomic era with sequencing of 20,000–25,000 genes of the human genome**
- **Since 2003: Post-genomic era to undersand « cell mysteries »**

=> Quantity and heterogeneity characterize data coming from the post-genomic era

=> The life sciences sector collects vast amounts of complex biological data requiring the use of advanced ICT to make sense of it all.

Added Value of Grid for Biologists

- The grid provides the centuries of CPU cycles required on demand
- The grid provides the reliable and secure data management services to store and replicate the biochemical inputs and outputs
- The grid offers a collaborative environment for the sharing of data in the research community on emerging and neglected diseases

Introduction

- **Various aspects of Biomedical Sciences can benefit from a grid-based approach:**

  – Search for new drug targets within the genome and the proteome,
  – Identification of single nucleotide polymorphisms (SNPs) relating to drug sensitivity
  – Protein Folding
  – Drug resistance mechanism elucidations
  – Epidemiological monitoring of disease outbreaks

  And so on…

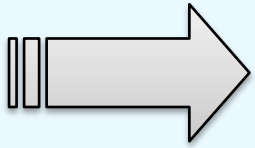
Introduction

- **Biologists need growing capability to handle all the data relevant to their research topics**
  - Design of complex analysis workflows
  - Knowledge management
- **Bioinformaticians who are developing the IT services for the biologists need growing resources**
  - To store, update, curate exponentially growing databases
  - To run increasingly complex algorithms on this growing data set
  - To build new databases exploiting the growing body of knowledge
- **Biologists and bioinformaticians have therefore different needs**
  - Biologists need high level environments and little resources
  - Bioinformaticians need large resources to develop and/or update the services needed by the biologist

Grid enabled drug discovery

- **Pharmaceutical development:**
  - Time-consuming: more than 10 years to develop a new medicine
  - Expensive: hundreds millions of dollars
  - Emergent and neglected diseases need fast and cheap answer

- **Computational tools:**
  - More and more known and registered protein 3D structures
  - More and more libraries of known chemicals
  - More and more computing power available
  - Better quality of prediction for bioinformatics tools, but CPU-consuming



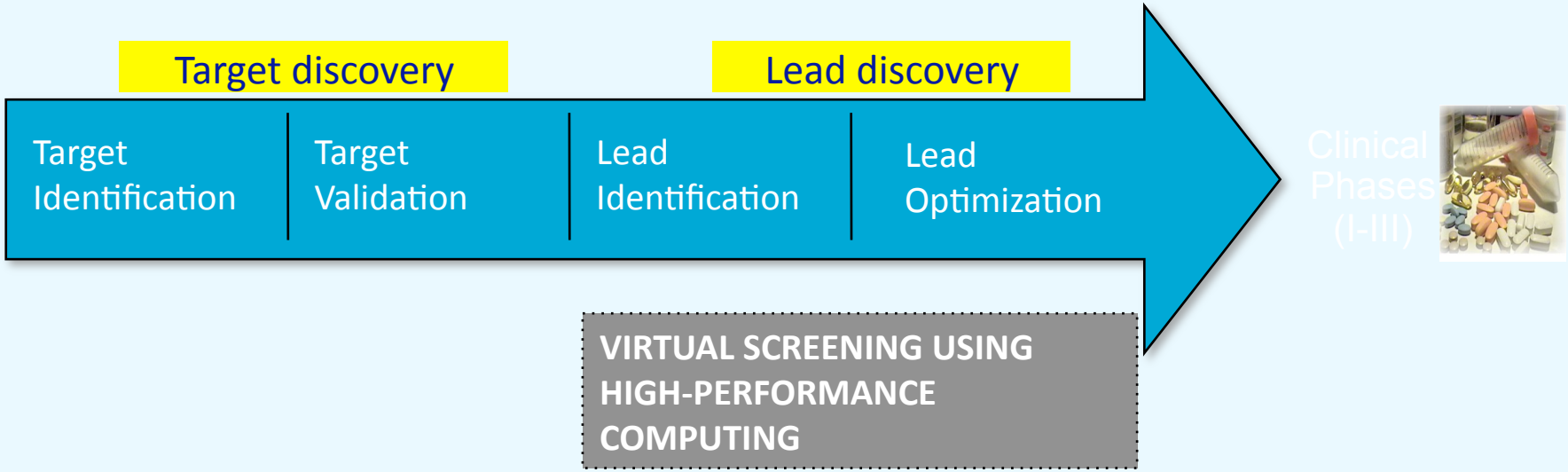
Virtual screening using grid to speed-up the process and minimize the costs

Grid enabled drug discovery

- **To analyse large databases of chemical compounds in order to identify possible drug candidates**

EXPERIMENTAL SCREENING IN WET LABORATORY

Duration: 12 – 15 years, Costs: 500 - 800 million US $



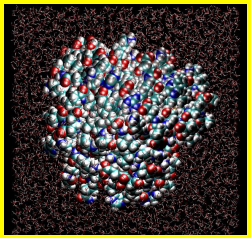
VIRTUAL SCREENING USING HIGH-PERFORMANCE COMPUTING

Duration: few days-months approx., Cost: CPU-time

From Grid to Lab

FLEXX/
AUTODOCK

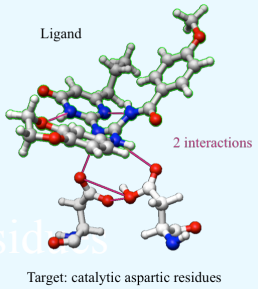
AMBER



Molecular docking

Molecular dynamics

Complex visualization



CHIMERA

Catalytic aspartic residues

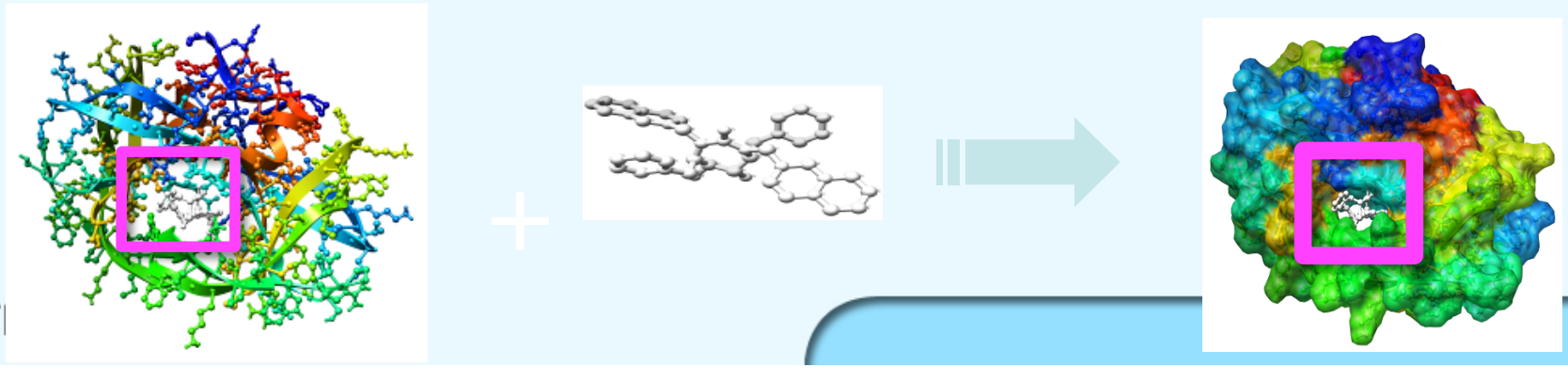
in vitro

WET LABORATORY

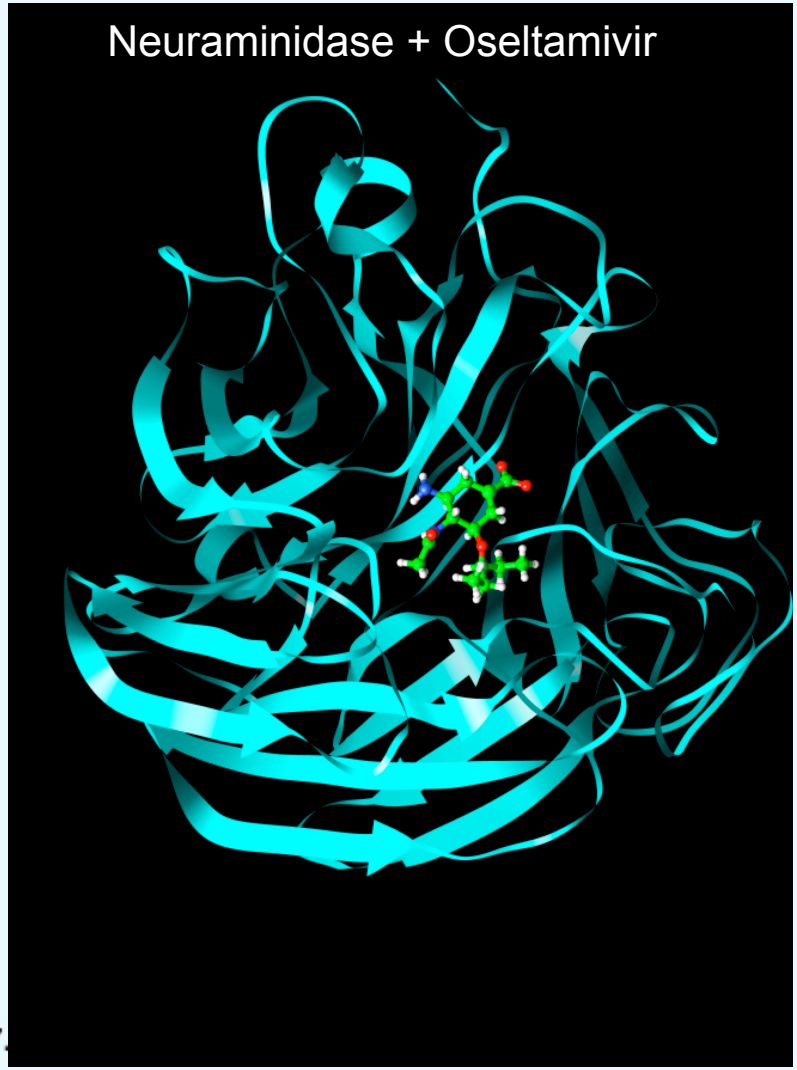
in vivo

Docking

- TARGET: 3D structure for a key protein in a disease
- LIGAND: database of chemical compounds commercially available
- SOFTWARES for virtual screening: docking, molecular dynamics
  - Parallel computations
  - Licences if needed (BioSolveIT and CCDC provided free licences for specific projects)



Flu….



- Docking results between a ligand and a protein based on:
  - Scoring
  - Match information
  - Different parameters settings
  - Knowledge of binding site

Targets

- **The Protein Databank contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies**

  http://www.rcsb.org/

		Proteins	Nucleic Acids	Protein/NA Complexes	Other	Total
Exp. Method	X-ray	48337	1168	2221	17	51743
	NMR	7003	869	150	6	8028
	Electron Microscopy	171	16	65	0	252
	Hybrid	15	1	1	1	18
	Other	115	4	4	9	132
	Total	55641	2058	2441	33	60173

Ligands

- **Comprehensive (> ~500,000 compounds)**
  - search in the dark
- **Diversity-based to cover ‘chemical space’**
  - efficient search in the dark
- **“Focused”or“Targeted”for lead identification**
  - e.g.filtered by 2D or 3D pharmacophores
- **Focused” or “Targeted” for lead optimization**
  - focussing the spotlights
- **Combinatorial Libraries**

Docking Programs

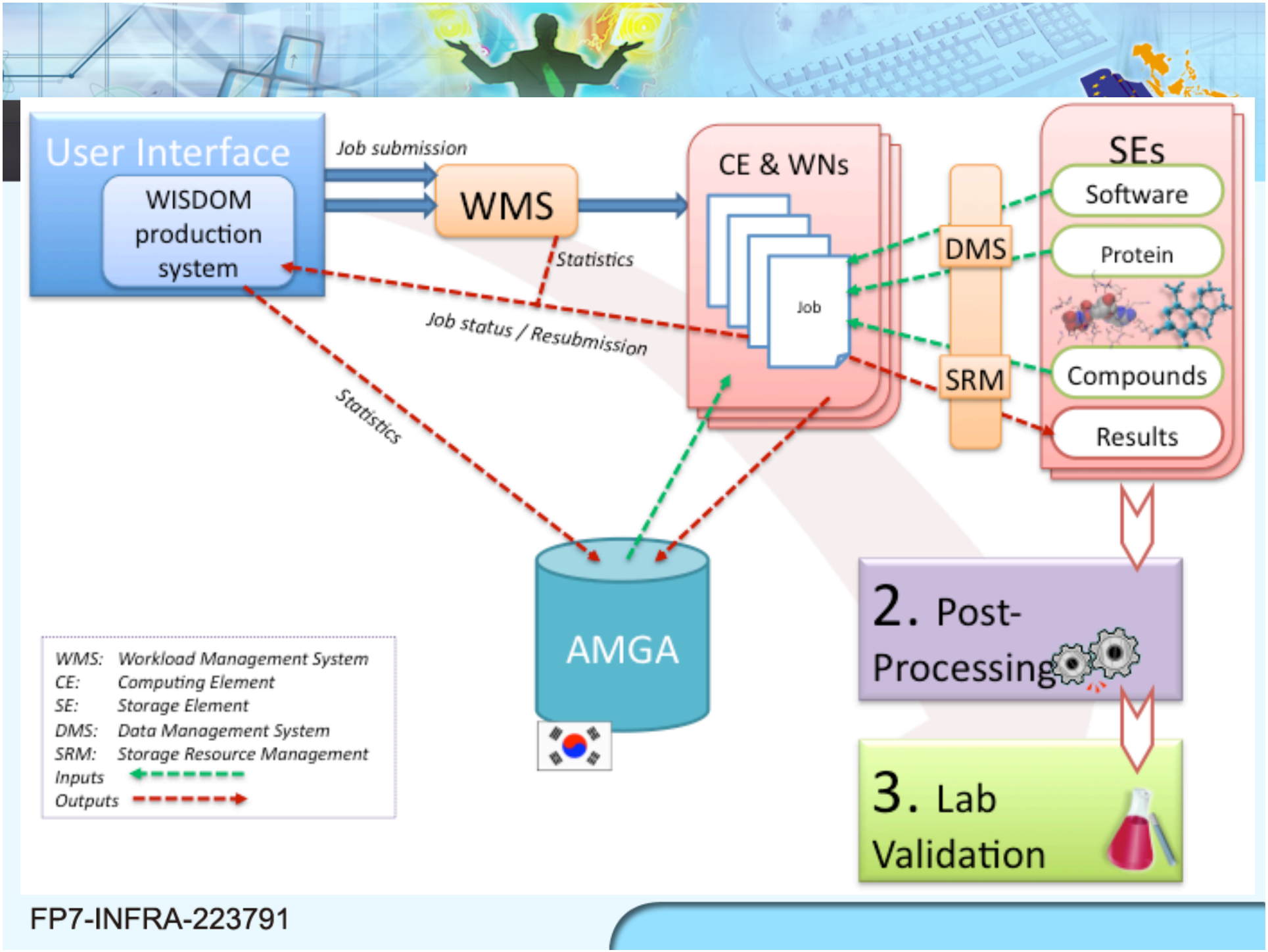
- DOCK: (Kuntz et al. 1982)
- DOCK 4.0 (Ewing & Kuntz 1997)
- AutoDOCK (Morris et al. 1998)
- GOLD (Jones et al. 1997)
- FlexX: (Rarey et al. 1996)
- GLIDE: (Friesner et al. 2004)
- ADAM (Mizutani et al. 1994)
- CDOCKER (Wu et al. 2003)
- CombiDOCK (Sun et al. 1998)
- DIVALI (Clark & Ajay 1995)
- DockVision (Hart & Read 1992)
- FLOG (Miller et al. 1994)
- GEMDOCK (Yang & Chen 2004)
- Hammerhead (Welch et al. 1996)
- LIBDOCK (Diller & Merz 2001)
- MCDOCK (Liu & Wang 1999)
- PRO_LEADS (Baxter et al. 1998)
  SDOCKER (Wu et al. 2004)
- QXP (McMartin & Bohacek 1997)
- Validate (Head et al. 1996)
- ....

Experiments

Project	Protein	Function	Ligands
Malaria	Plasmepsin PMII	Hemoglobin degradation	ZINC subset 1 million
	Glutathione-S-Transférase GST	Detoxification	ZINC 4, 3 millions
	Dihydrofolate Reductase DHFR	DNA synthesis	ZINC 4, 3 millions
Avian flu	Neuraminidase	Release of new virus	ZINC subset 300, 000 millions + chemical combinatorial library
Diabetes	Amylase/Gluco-amylase	Carbohydrate cleavage	ZINC subset 300, 000 millions

Grid environment

- **Develop an environment to monitor the deployments on grid: Wisdom Production Environment**
  - Produce a large amount of data during the datachallenge
  - in a limited time and minimal human needs
  - Manage the fact that grid is heterogeneous and dynamic
    - a workflow of grid job handling: automated job submission, status check and report, error recovery
    - push and pull model job scheduling
    - batch mode job handling


User Interface
WISDOM production system
Job submission
WMS
Statistics
Job status / Resubmission
Statistics
CE & WNs
Job
DMS
SRM
SEs
Software
Protein
Compounds
Results
AMGA
2. Post-Processing
3. Lab Validation
WMS: Workload Management System
CE: Computing Element
SE: Storage Element
DMS: Data Management System
SRM: Storage Resource Management
Inputs
Outputs

Results obtained

eGee Enabling Grids for E-sciencE	Number of dockings	CPU years	Real Time	CPUs used	Produced Data size	Crunching Factor	Distribution efficiency	Model
Malaria I	41 millions	80	6 sem	1700	1TB	400	25%	PUSH
Malaria II	142 millions	400	2,5 mois	Jusqu'à 5000	1,6 TB	2000	40%	
Avian Flu	4 millions	100	1,5 mois	1700	800 GB	900	50 % (>80% DIANE)	
Diabetes	300, 000	40	2,5 jours	7000		6000	85 %	PULL

Providing Grid to non experts

WISDOM ENVIRONMENT

- Complex and unflexible
- For people familiar with GRID
- Drug discovery applications

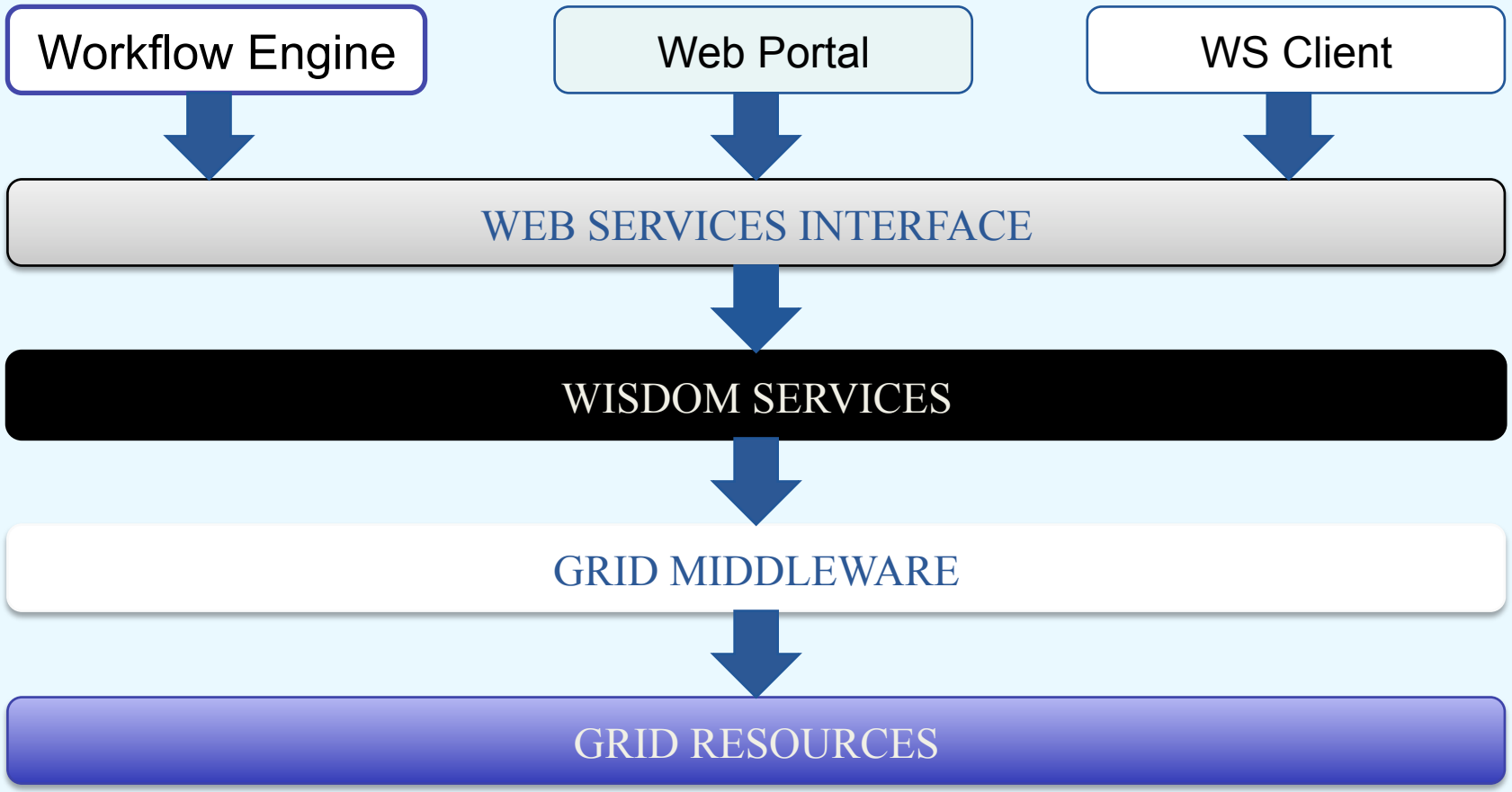
2005-2008

TAKE ADVANTAGE OF THE GRID COMPUTING AND STORAGE RESOURCES

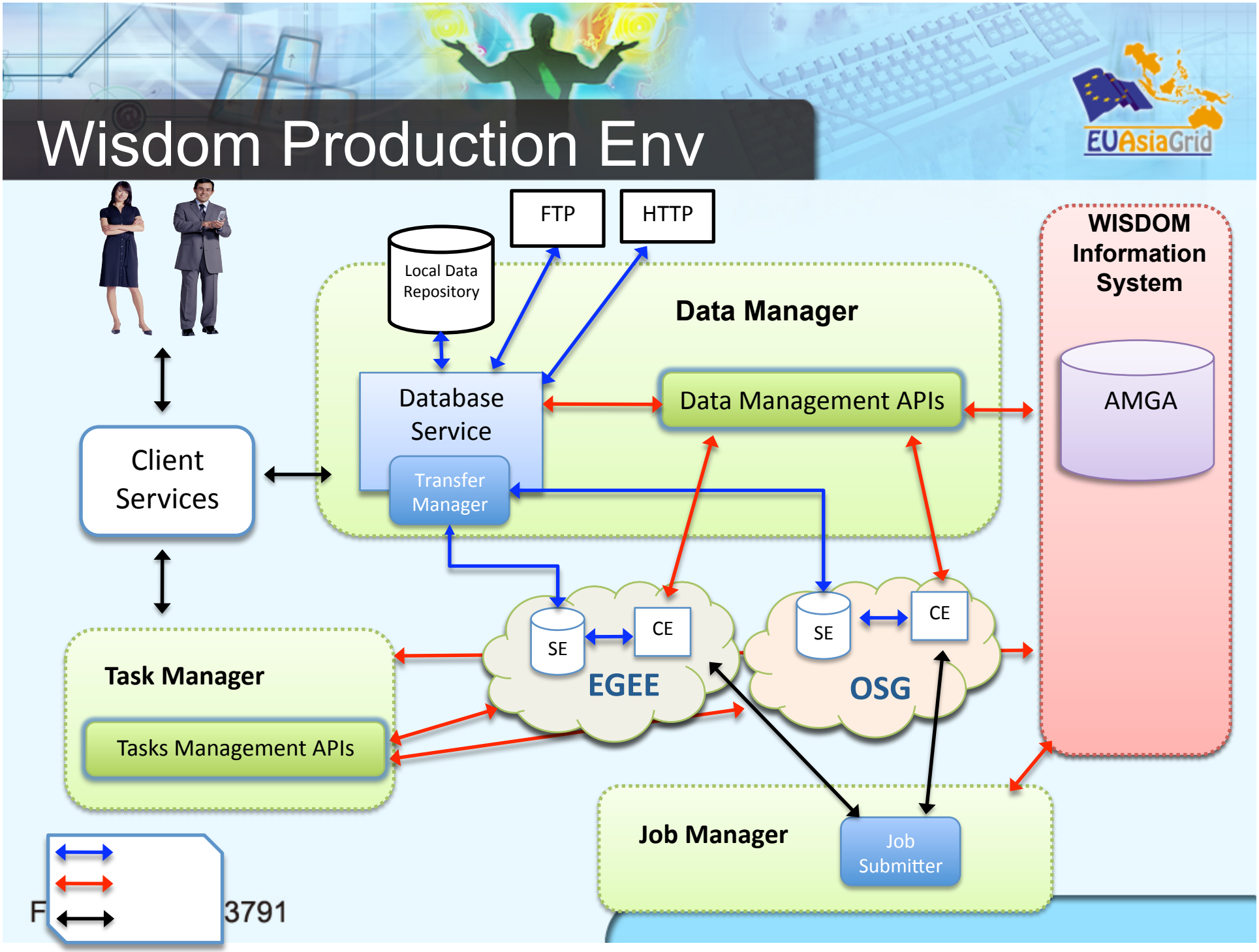
2008-…

- SOA
- Non-expert users
- Several bioinformatic's
- tools

Implementation



Wisdom Production Env

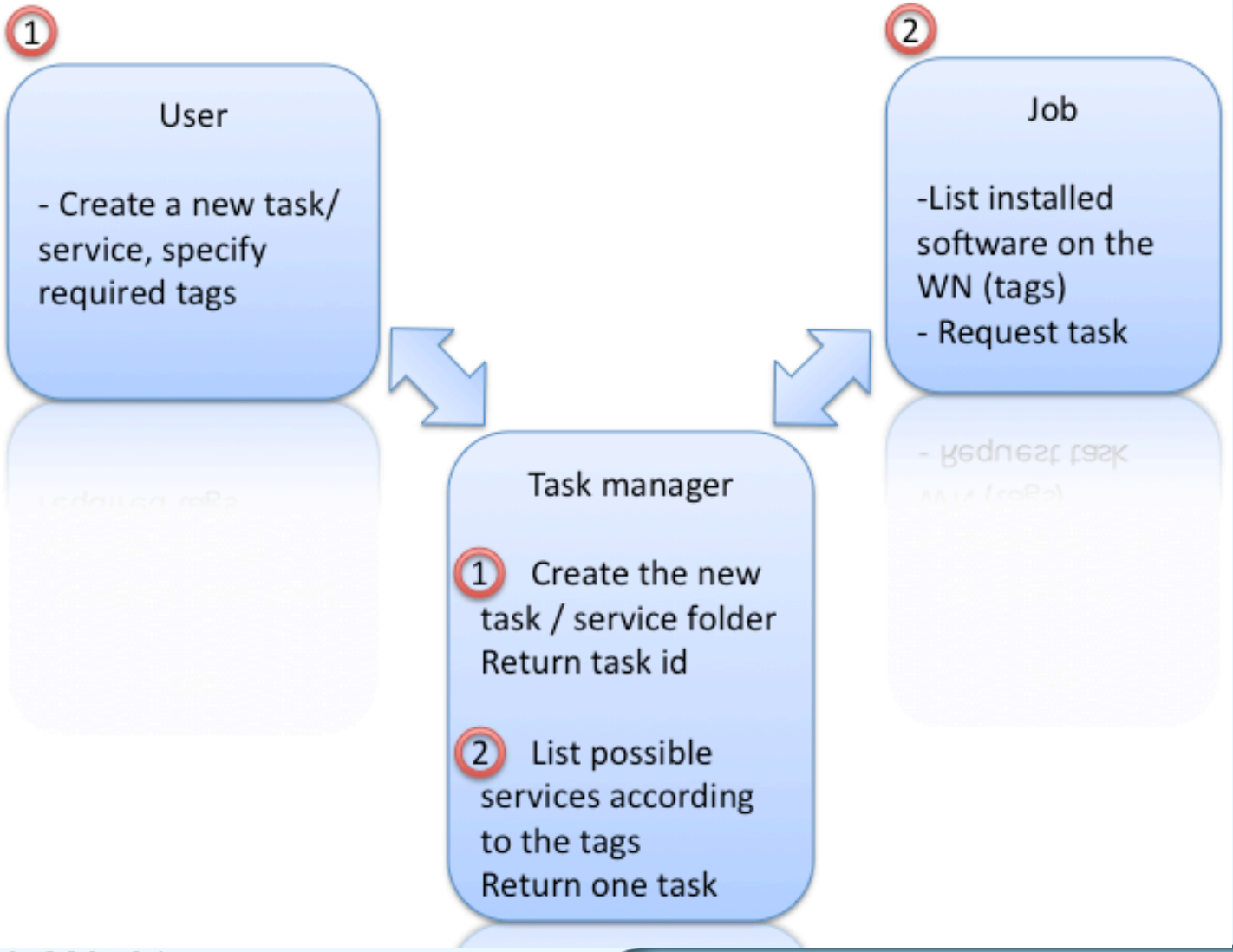


Wisdom Production Env

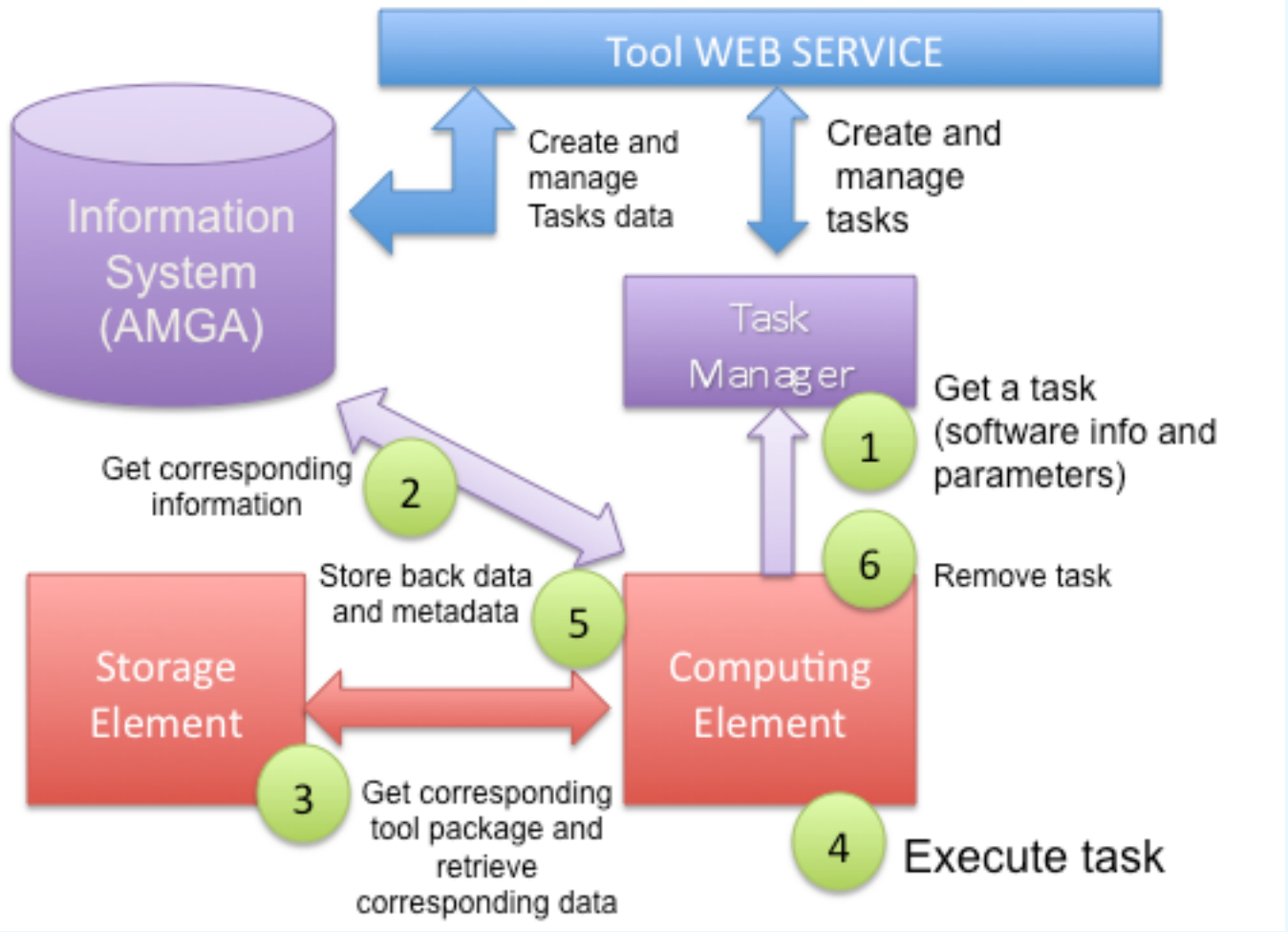
- **WISDOM data manager**
  - high-level services to manage the data and metadata related to the applications and tasks
  - service that can be used to automatically deploy and synchronize data on the grid (including databases)
  - set of APIs to access and query data

- **WISDOM information system**
  - based on AMGA metadata catalogue

Task Manager Interactions



Task submission process



Now it’s time for hands-on…