

The WISDOM experience

*Nicolas jacq
HealthGrid association*

*Credit : the WISDOM collaboration
<http://wisdom.healthgrid.org>*

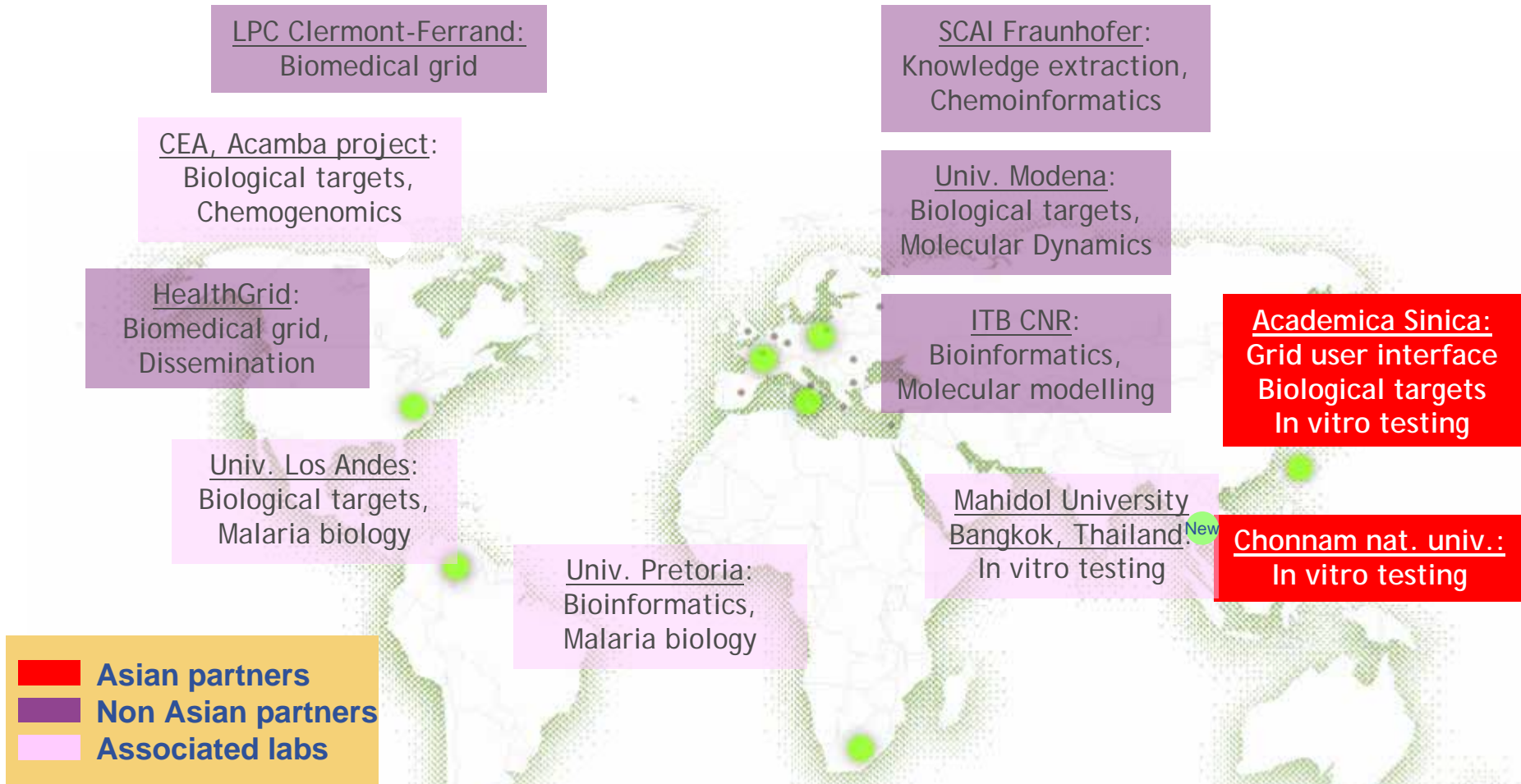
*Biomed Grid School – Varenna (Italy)
16 May 2007*



- **Introduction : goals of the WISDOM application**
- **The production environments**
- **Results**
- **Conclusion and perspectives**

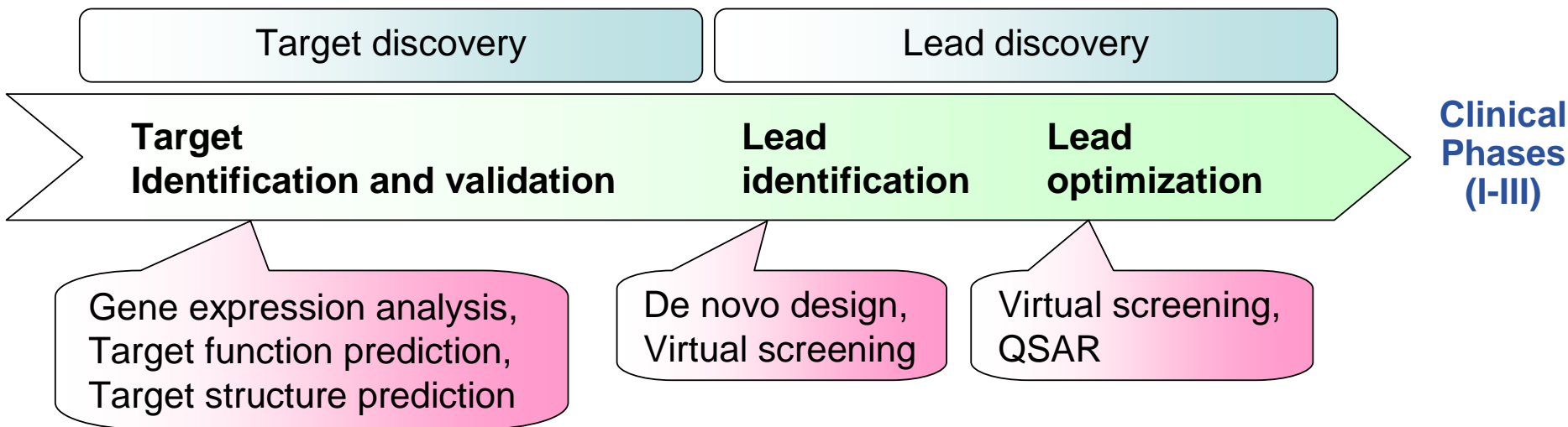
- **WISDOM stands for World-wide In Silico Docking On Malaria**
- **Goal: find new drugs for neglected and emerging diseases**
 - Neglected diseases lack R&D
 - Emerging diseases require very rapid response time
- **Method: grid-enabled virtual docking**
 - Cheaper than in vitro tests
 - Faster than in vitro tests



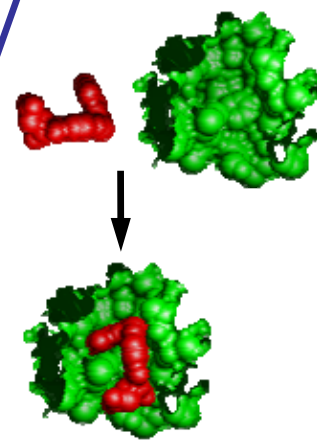
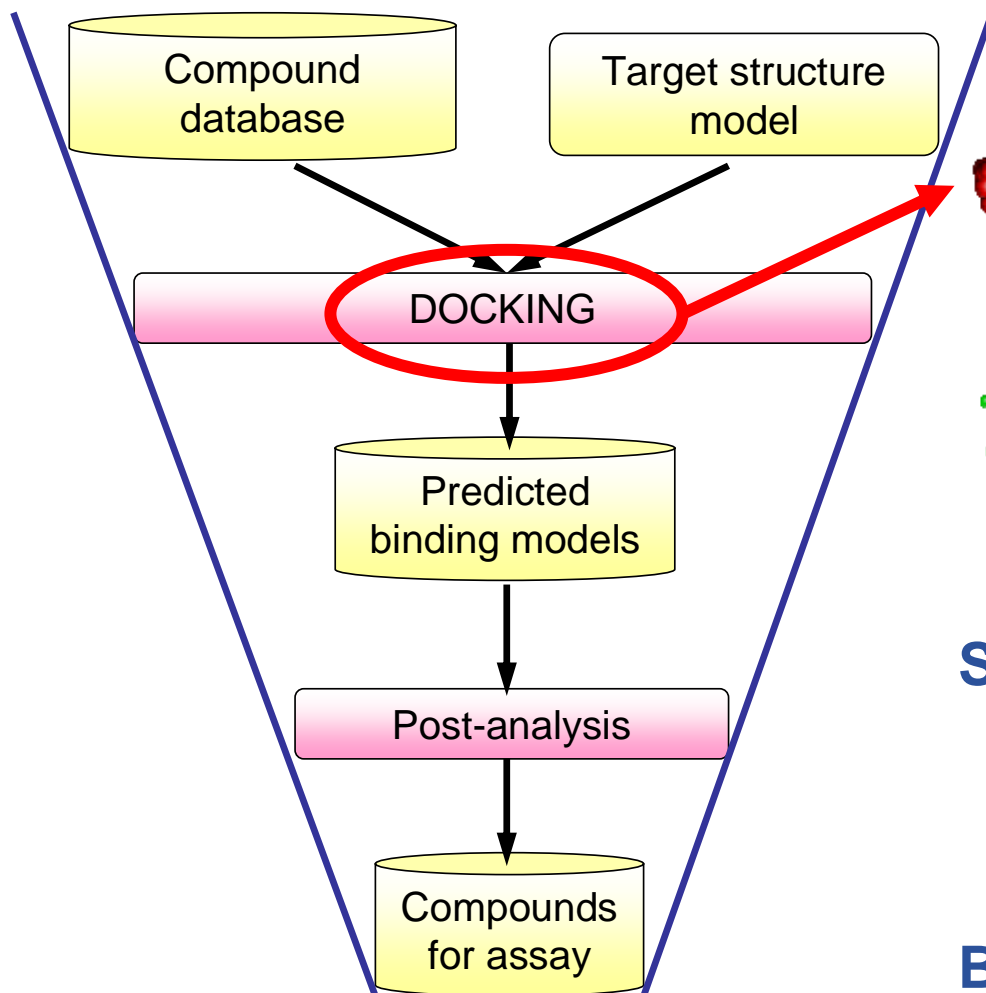


7 partners, 4 associated laboratories providing targets and/or in vitro facilities

DRUG DISCOVERY



IN SILICO DRUG DISCOVERY



Docking: predict how small molecules bind to a receptor of known 3D structure

Successful examples

- rapid,
- cost effective...

But there are limitations

- CPU and data intensive

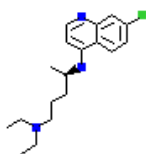
Millions of potential drugs to test against interesting proteins!

~~High Throughput Screening
~10\$/computer per several hours
Too costly for neglected disease!~~

Compounds:

ZINC: 4.3M

Chembridge: 500,000

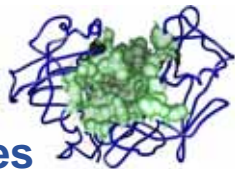


Molecular docking (**FlexX, Autodock**)

~1 to 15 minutes

Targets:

PDB: 3D structures

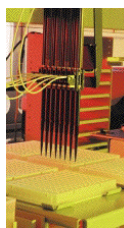


Data challenge on **EGEE**

~ 2 to 30 days on ~5,000 computers

~~Cheap and fast!~~

Selection of the best hits



Hits screening using assays performed on living cells



Leads
Clinical testing
Drug

- **Objective**
 - To dock a whole compound database in a limited time with a minimal human involvement

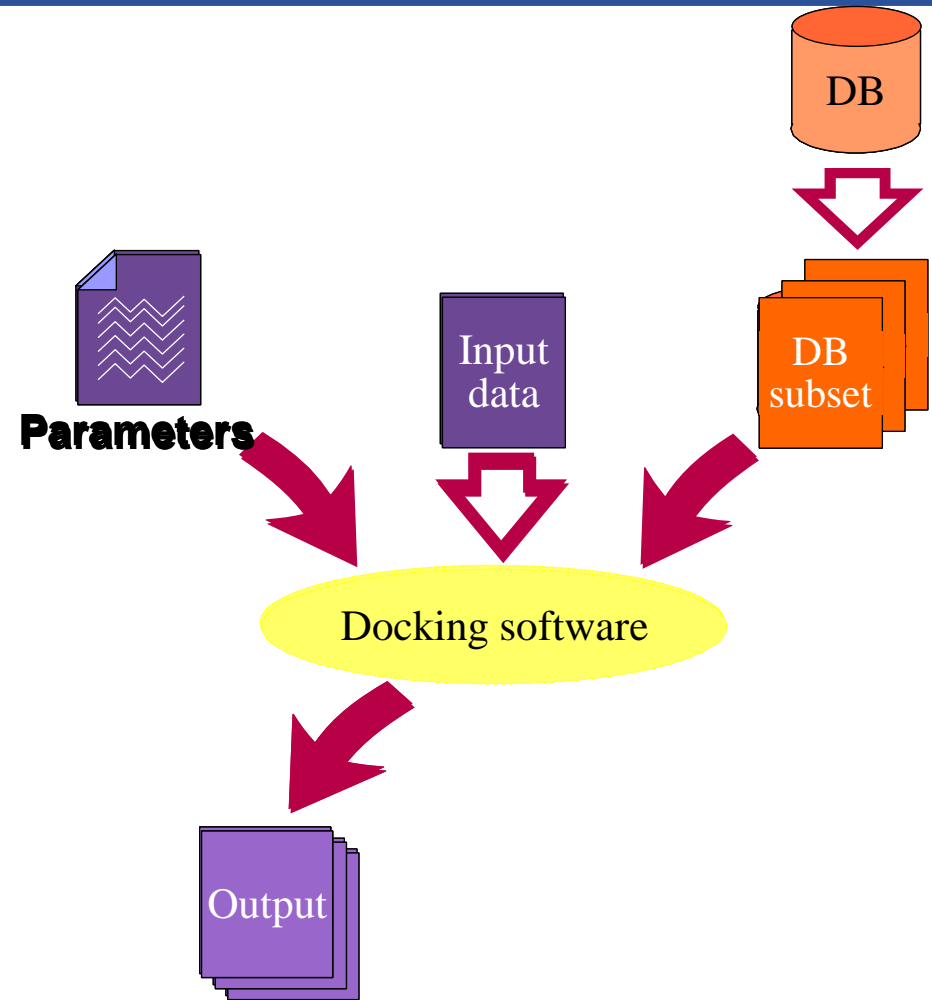
- **Need for an optimized environment**
 - To achieve production in a limited time
 - To optimize performances

- **Need for a fault tolerant environment**
 - To handle Grid heterogeneity and dynamics
 - To collect and store critical data

- **Need for user-friendly high-level interfaces**
 - To ease the execution
 - To offer a service to the non grid experts

- Introduction : goals of the WISDOM application
- **The production environments**
- Results
- Conclusion and perspectives

- The application code can not be modified.
- The applications are **not designed for grid computing.**
- A resource estimation is needed before the deployment
- A common strategy is **to split the application into shorter tasks**
- The application package requires **installation and testing**
- **License management** for commercial software is not adapted for large infrastructure



Embarrassingly parallel application

Real Time Monitor (Imperial College London)

- **Large number of CPUs available**
- **Reliable and secured Data Management Services**
 - Sharing of results
 - Replication of the data
 - ACLs
- **Availability of the resources**

Statistics:

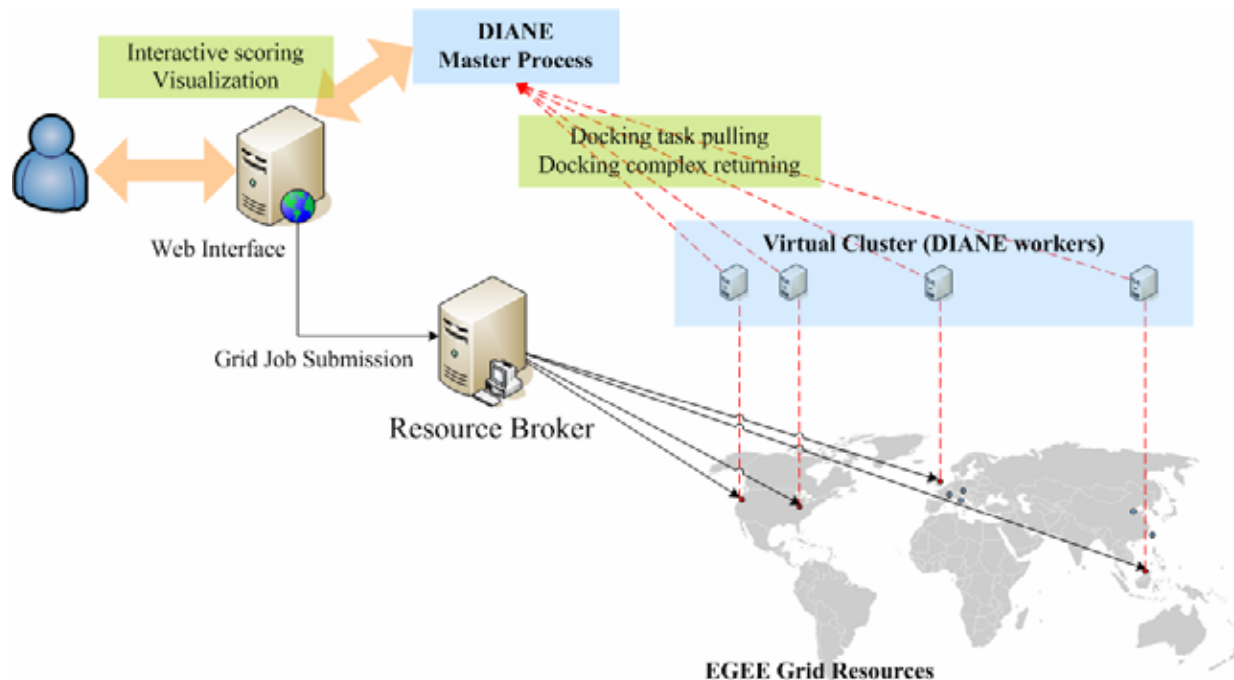
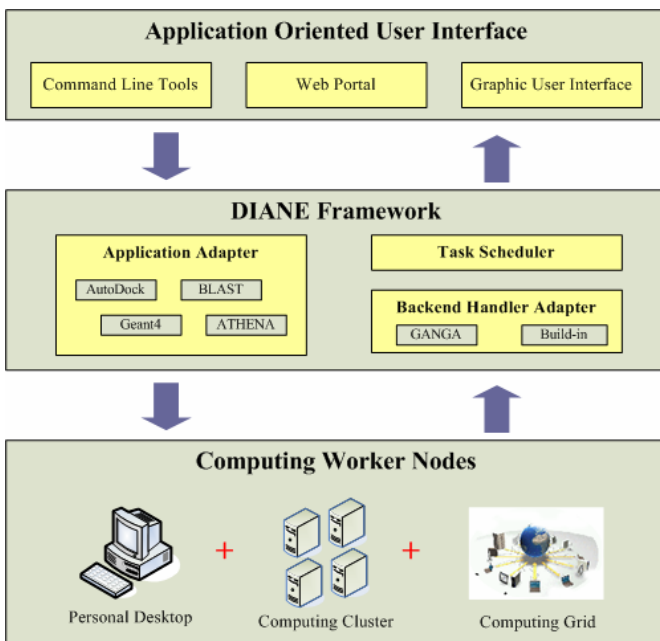
Submitted:	977	■
Waiting:	1903	■
Ready:	3218	■
Scheduled:	18482	■
Running:	8670	■

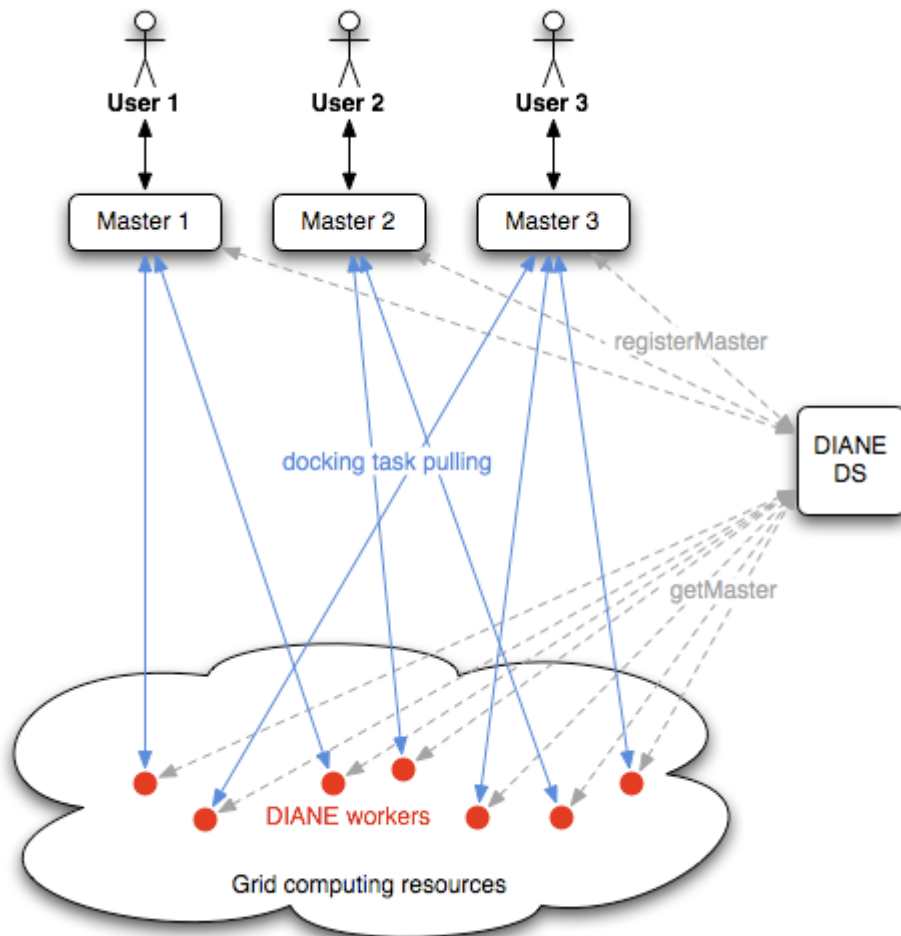
- **Biomed VO leader : V. Breton**
- **~80 participants, see <http://egeena4.lal.in2p3.fr>**
- **Three active subgroups**
 - **Medical imaging** (J. Montagnat)
 - **Bioinformatics** (C. Blanchet)
 - **Drug discovery** (V.Breton)
- **Biomedical VO manager: Y. Legré, legre@clermont.in2p3.fr**
- **See <http://cic.in2p3.fr> (VO information, publication of data challenge...)**
- **1 VOMS server, 1 LFC, +20 RBs**
- **+100 CEs, +8,000 CPUs (but many users)**
- **+110 SEs, ~Tens of TB available on disk**
- **27 countries**

- **Managing thousands of jobs and files is a manually labor-intensive task**
 - Job preparation, submission and monitoring, output retrieval, failure identification and resolution, job resubmission...
- **The rate of submitted jobs must be carefully monitored**
 - In order to avoid Resource Brokers overload
 - In order to efficiently use the resources
- **The amount of transferred data impacts on grid performance**
 - The data must be installed on the grid
 - Storing subsets of the database instead of large unique compound files
- **Grid process introduces significant delays**
 - The submitted jobs must be sufficiently long in order to reduce the impact of this middleware overhead

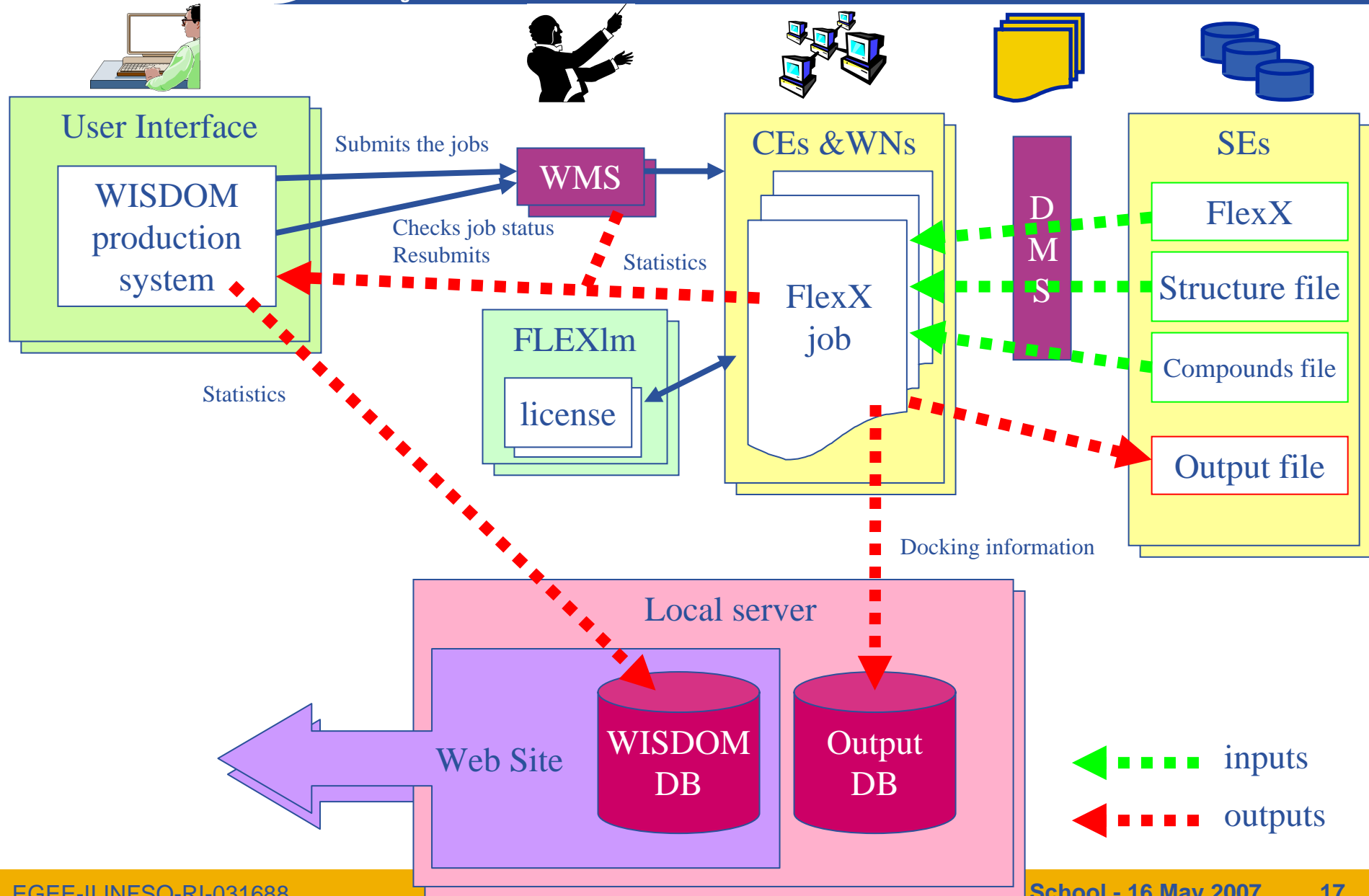
- **The ATLAS production system - The ATLAS experiment**
- **BOSS and CRAB - The CMS experiment**
- **Alien - The Alice experiment**
- **DIRAC - The LHCb experiment**
- **DIANE - CERN**
- **Ganga, a user interface**
- **GridICE, Monalisa and LHC Dashboard, three monitoring services for users**

- DIANE: Distributed Analysis Environment
- An overlay system on top of a variety of distributed computing environment, taking care of all synchronization, communication and workflow management details on behalf of application
- A lightweight framework for parallel scientific applications in master-worker model
- Pull model job scheduling + interactive mode job handling with flexible failure recovery mechanism

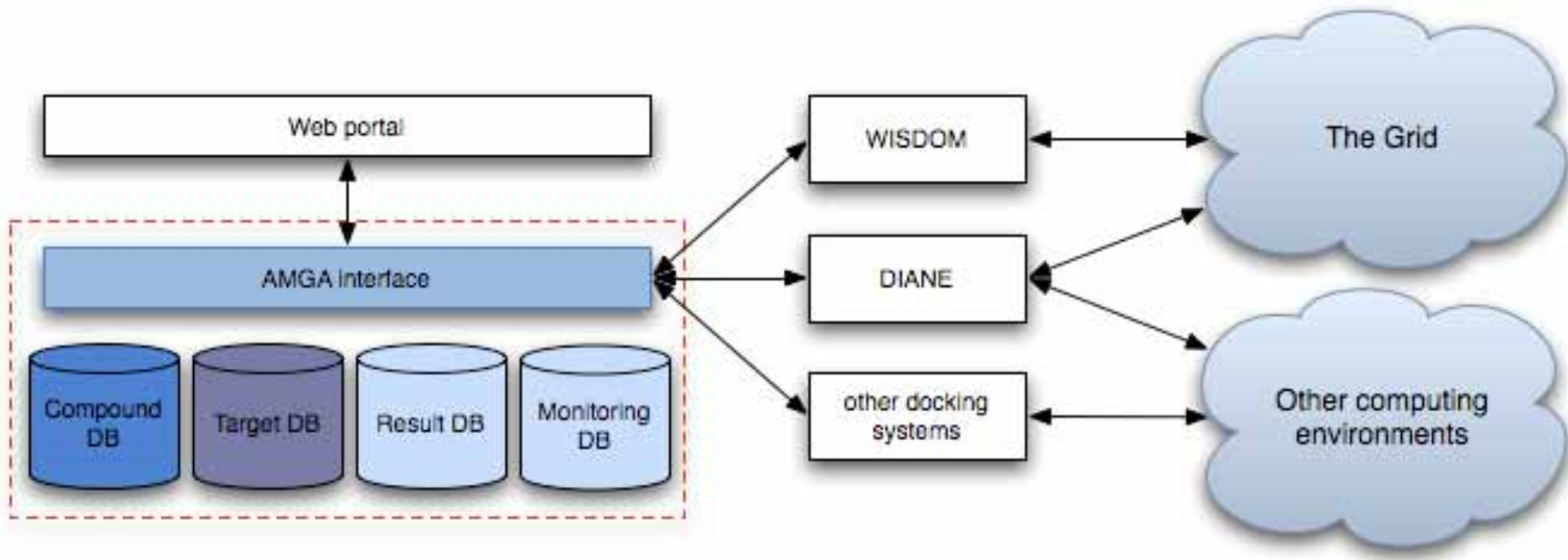


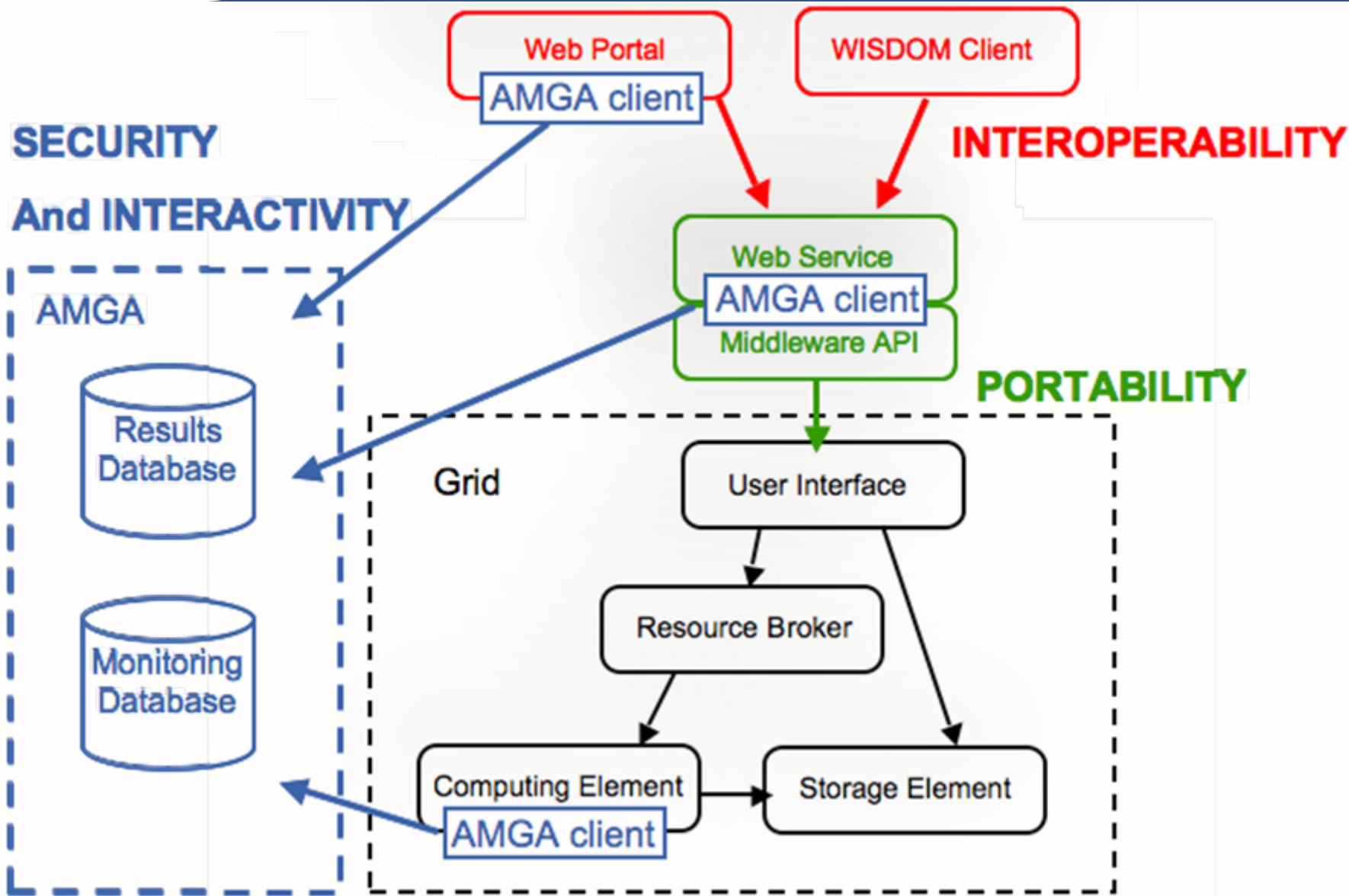


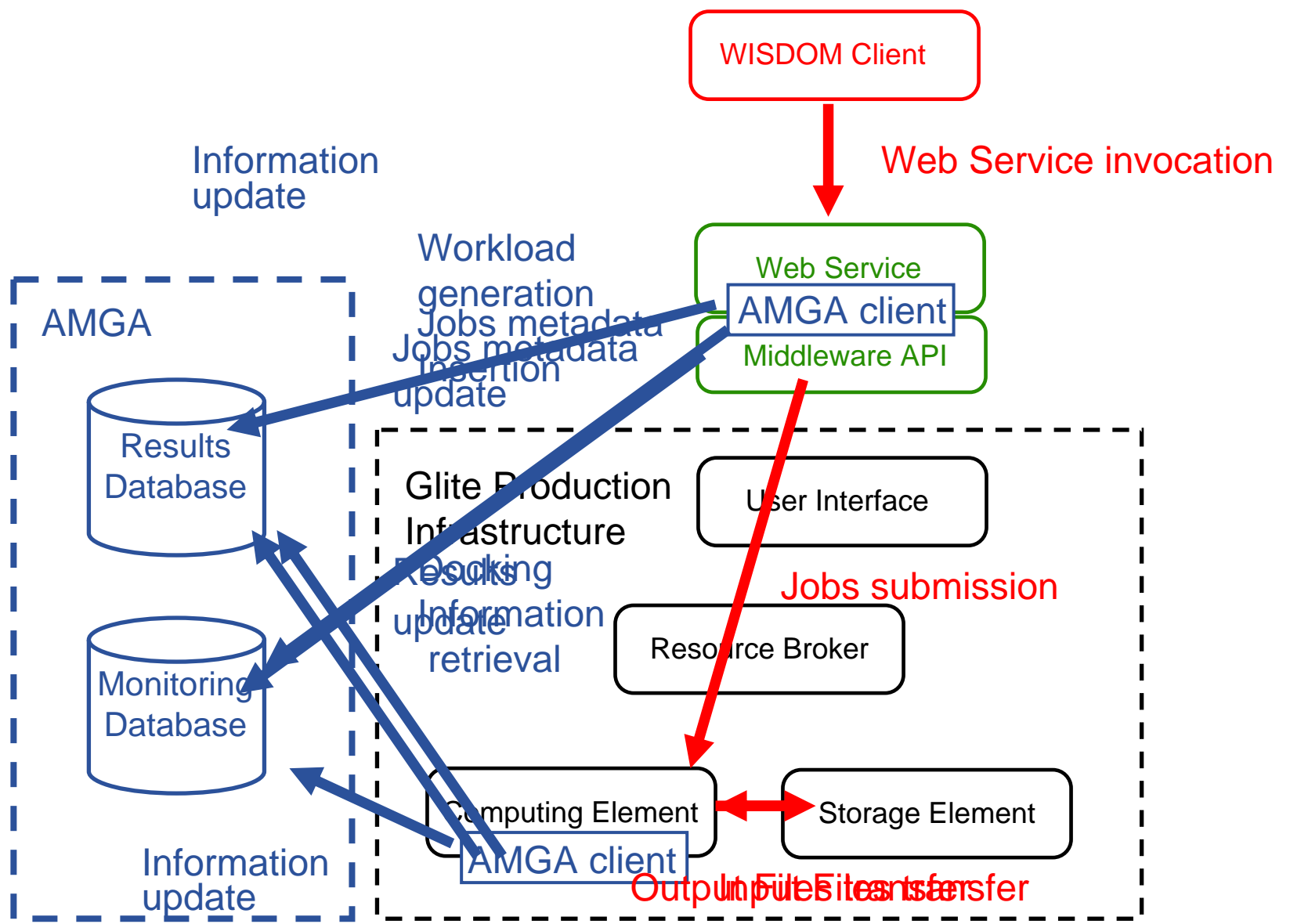
- Improving the scalability of the DIANE framework
- The Directory Service is a server containing a list of all the masters
- The Master register itself to the Directory Service
- The Workers obtain a Master through the Directory Service
- Directory Service has an algorithm for the load balancing of the workers and prioritization of the masters



- Chemical properties to better annotate the compounds
- Results essential for further analysis are extracted and stored in a result database
- Database access through AMGA
 - for access control
 - for data replication







Available at <http://wisdom-demo.healthgrid.org>

- Real-Time monitoring of the Grid
- Customizable interface
- Drag and drop components



DATA CHALLENGE: AVIAN FLU

April 14th - 27th

+ ADD COUNTER

+ ADD COMPOUNDS GRAPH

+ ADD JOBS GRAPH

VIEW ALL MENUS

CLOSE MENU

? HELP

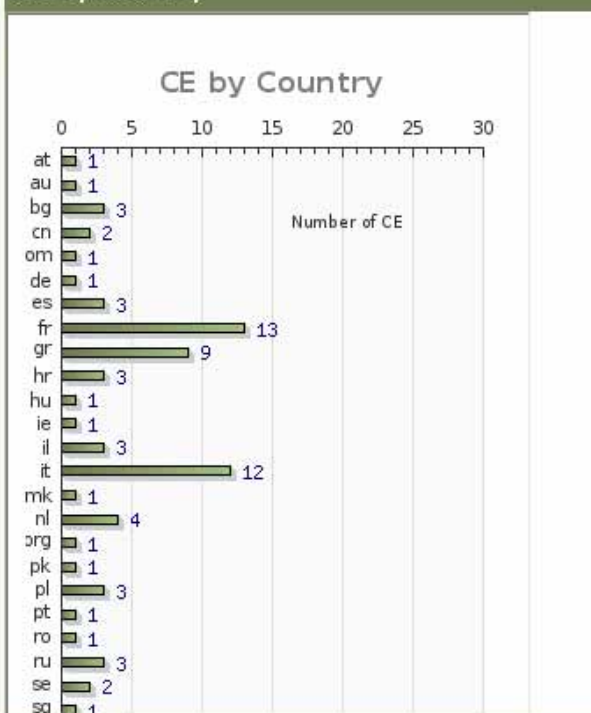


Compounds vs Time

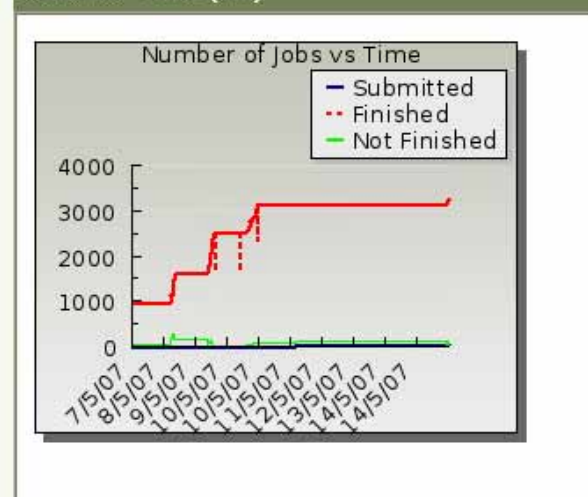
All figures

Docked compounds:	3250
Cost in silico (est.):	864 €
Cost in vitro (est.):	32,500 €
CPU days consumed:	36 Days
Size of data produced:	1134.219 MB

Ce by country



Jobs vs Time (all)



- **User Friendly Interface for biologists**
- **Real Time output of the results**
 - 3D views of the docking poses and structures
- **Resubmission and monitoring of docking jobs**

Grid Application Portal

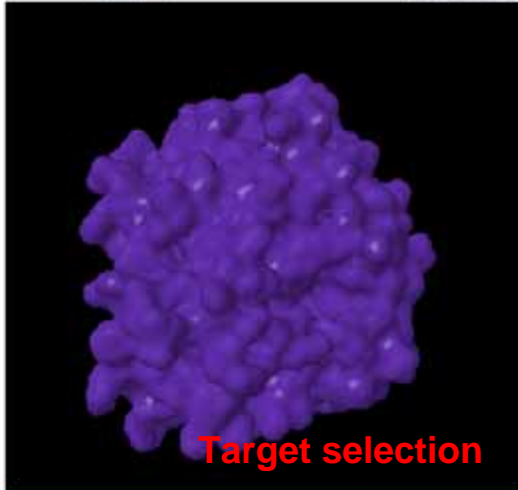
Logged in user: gs

VirtualScreening
Job Management
User Information
Document

Home
Help
Logout

Description

MacroFile: [view paramFile](#)



Target selection

Compound selection

Select Library:

No filter rule

select	file_name	script
<input type="checkbox"/>	100.sdf	
<input type="checkbox"/>	101.sdf	

Page: 1 of 259

Initial Translation, Quaternion and Torsion Step Sizes and Reduction Factors

Translation step / Å:

Quaternion step/deg:

Torsion step/deg:

Translation reduction factor / per cycle:

Quaternion reduction factor / per cycle:

Torsion reduction factor / per cycle:

Docked Conformation Clustering Parameters for "analysis" command

Cluster tolerance (Angstroms):

External grid energy:

Maximum allowable initial energy:

maximum number of retries:

Genetic Algorithm (GA) and Lamarckian Genetic Algorithm Parameters

Number of individuals in population:

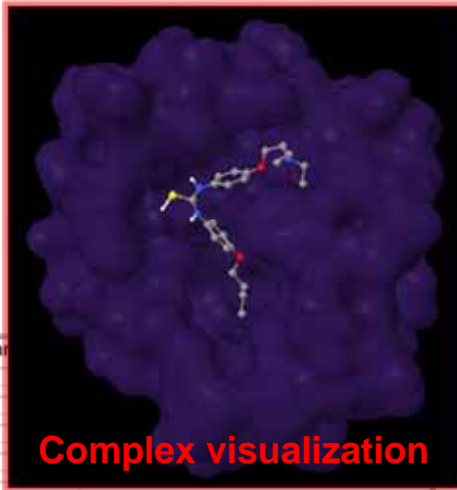
Maximum number of energy evaluations:

Maximum number of generations:

Number of top individuals that automatically survive:

Rate of Gene mutation:

Rate of Crossover:



Complex visualization

Job Details										
id	submitTime	startTime	finishTime	computing element	status	view results	output sandbox	resubmit	energy	pdb
de2ee3cb:1115494a807	2007-03-15 07:51:47 GMT	2007-03-15 07:52:28 GMT	2007-03-15 07:59:03 GMT	quanta.grid.sinica.edu.tw	DONE	Please drop down	download	resubmit	-9.44	view
de2ee3cb:1115494a805	2007-03-15 07:51:47 GMT	2007-03-15 07:52:29 GMT	2007-03-15 07:58:30 GMT	quanta.grid.sinica.edu.tw	DONE	Please drop down	download	resubmit	-10.89	view
de2ee3cb:1115494a806	2007-03-15 07:51:46 GMT	2007-03-15 07:52:08 GMT	2007-03-15 07:58:38 GMT	quanta.grid.sinica.edu.tw	DONE	Please drop down	download	resubmit	-11.19	view
de2ee3cb:1115494a804	2007-03-15 07:51:45 GMT	2007-03-15 07:52:09 GMT	2007-03-15 07:58:31 GMT	quanta.grid.sinica.edu.tw	DONE	Please drop down	download	resubmit	-7.41	view

Energy table

Probability of performing local search on an individual:

GA or LGA runs:

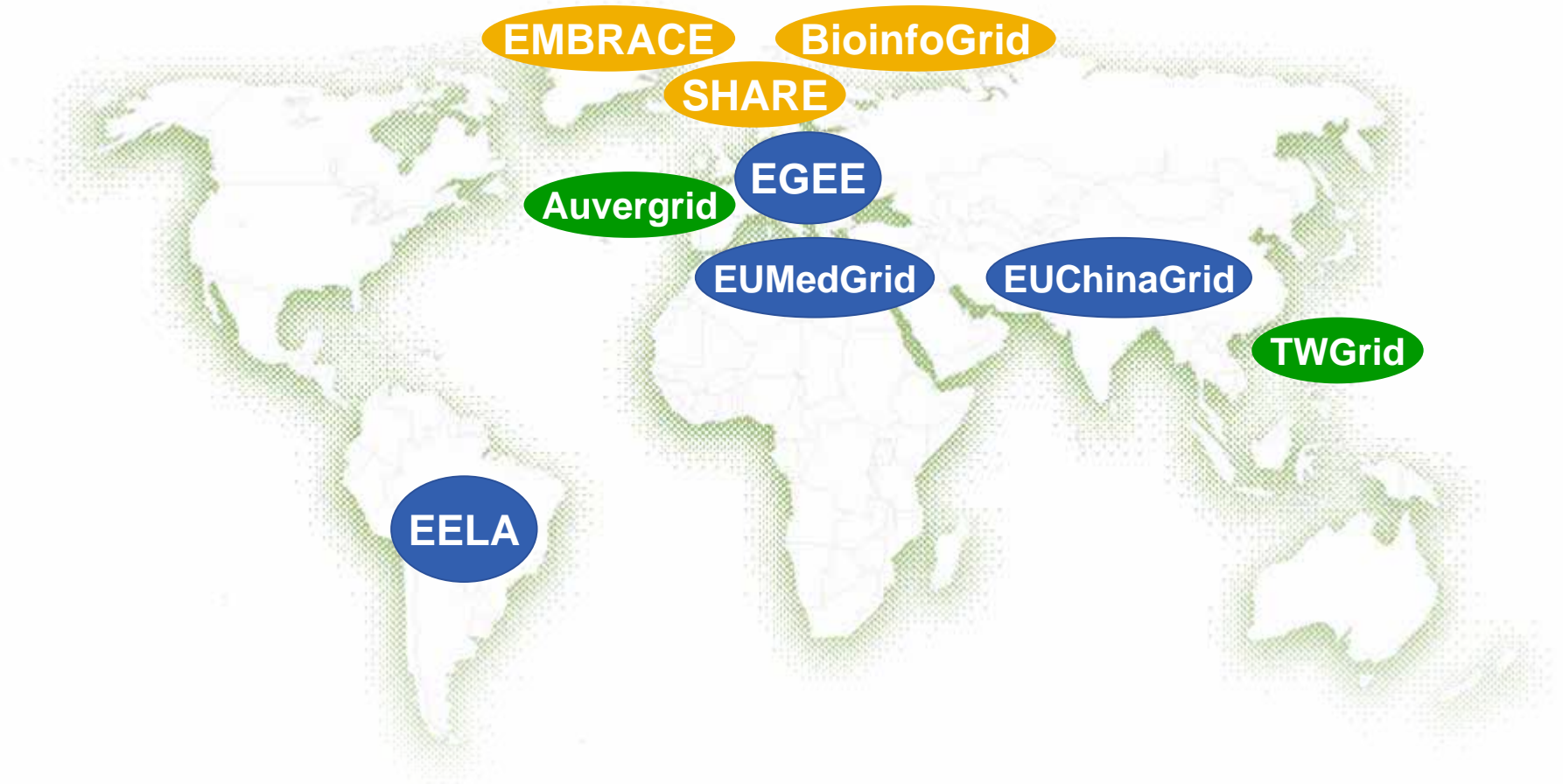
Docking parameter setter

- Introduction : goals of the WISDOM application
- The production environments
- **Results**
- Conclusion and perspectives

- **First Data Challenge: July 1st - August 15th 2005**
 - Target: malaria
 - 80 CPU years
 - 1 TB of data produced
 - 1,700 CPUs used in parallel
 - 1st large scale docking deployment world-wide on a e-infrastructure

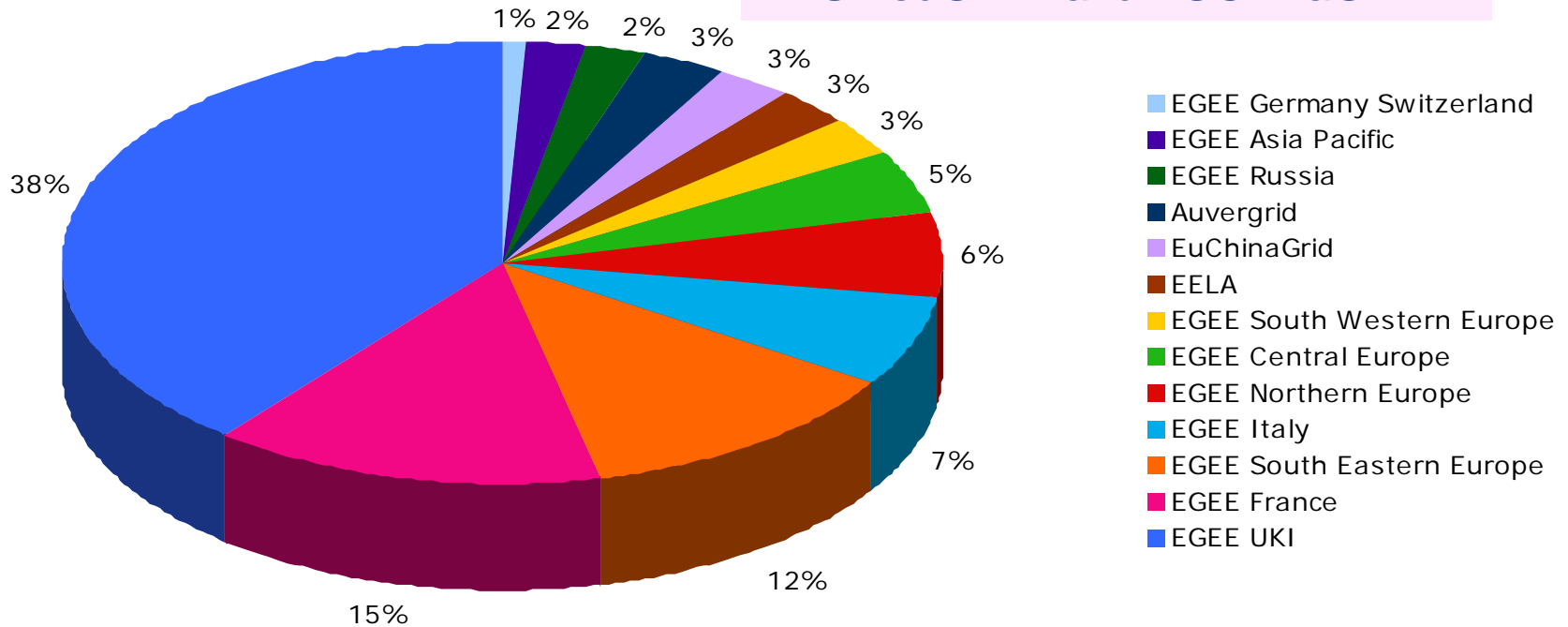
- **Second Data Challenge: April 15th - June 30th 2006**
 - Target: avian flu
 - 100 CPU years
 - 800 GB of data produced
 - 1,700 CPUs used in parallel
 - Collaboration initiated on March 1st: deployment preparation achieved in 45 days

- **Third Data Challenge: October 1st - 15th December 2006**
 - Target: malaria
 - 400 CPU years
 - 1.6 TB of data produced
 - Up to 5,000 CPUs used in parallel
 - Very high docking throughput: > 100,000 compounds per hour



- : European grid infrastructure
- : European grid project
- : Regional/national grid infrastructure

Significant contributions from EELA, EUMedGRID and EUChinaGRID



Over 420 CPU years in 10 weeks

A record throughput of 100,000 docked compounds per hour

WISDOM calculations used FlexX from BioSolveIT
(6k free, floating licenses)

	Rate	Reasons
Success rate after checking output data	46 %	
Grid success rate	63%	After subtracting license server and WISDOM failures
Workload Management failure	10 %	Overload, disk failure Mis-configuration, disk space problem Air-conditioning, power cut
Data Management failure	4 %	Network / connection Power cut Other unknown causes
Sites failure	9 %	Mis-configuration, tar command, disk space Information system update Job number limitation in the waiting queue Air-conditioning, electrical cut
Unclassified	4 %	Lost jobs Other unknown causes

- **Grid success rate: 80%**
 - Constant and slower job submission flow
 - Manual control of resubmission process
 - WISDOM fault-tolerance improved
 - Grid reliability improved (Workload Management System)

- Avian flu data challenge: in the selection of 2,250 compounds out of initial 308,585 compounds:**
 - 5 out of 6 known effective inhibitors were found.
 - enrichment factor of 111 was observed. (<1 in most cases)
- Experimental assay confirms 7 active out of 123 purchased “potential hits”.**
- Data challenges on malaria: the 25 most promising compounds out of 500,000 are now being purchased and will be tested in vitro at Chonnam National University, South Korea**

Global effectiveness:

$$\frac{(\text{Hits}_{\text{sampled}}/N_{\text{sampled}})}{(\text{Hits}_{\text{total}}/N_{\text{total}})}$$

Pearlman & Charifson, JMC, 2001

Pre-sceneing (AUTODOCK)
over collection and sample first 15%
EF¹
= (5/6)/15% = 5.5

Re-ranking (SDDB) first 15% and
sample first 5%
EF² = (5/6)/(5%*15%) = 111



- Introduction : goals of the WISDOM application
- The production environments
- Results
- **Conclusion and perspectives**

- **WISDOM proposes a new approach to drug discovery thanks to the grid**
 - Rapid deployment of very large scale virtual screening
 - Collaborative environment for the sharing of data in the research community

- **WISDOM fully exploits EGEE services, APIs and resources.**
 - AMGA allows to store securely results and statistics immediately
 - Web Service Interface using WS-I profile guarantees interoperability

- **First biochemical results demonstrate grid relevance to the drug discovery community**
 - Grid is a superior tool to discover new drugs

- **Molecular dynamics application on grid**
 - Successfully deployment of Amber software on grid infrastructure
 - Larger deployment in the next months
 - Collaboration with Univ of Modena and SCAI Fraunhofer

- **2nd data challenge against avian flu**
 - Testing phase: May, 2007, Official launch: June, 2007
 - Targets: the old targets used in 1st data challenge and the open conformation suggested by a Nature paper
 - Compounds: 500,000 laboratory owned compounds
 - 300 CPU years is required
 - The new environment will be used and tested
 - More Asian partners and collaborations
 - EUChinaGrid, CNGrid (China)
 - Dr. Domain Kim (S. Korea), Dr. Kun-Qian Yu (Shanghai, China)
 - PRAGMA (Pacific Rim Applications and Grid Middleware Assembly)

- To all **members** of the WISDOM collaboration for their contribution to the project (CNRS-IN2P3, ASGC, ITB-CNR, SCAI Fraunhofer...)
- To all **grid nodes** which committed resources and allowed the success of the initiative
- To all **projects** which supported the initiative by providing either computing resources or manpower to develop the WISDOM environment (EGEE, BioinfoGRID, Embrace...)
- To **BioSolveIT** by offering up to 6000 free licenses of FlexX