Diligent

A **DI**gital **L**ibrary **I**nfrastructure
on **G**rid **EN**abled **T**echnology

DILIGENT Project

Andrea Manzi
ISTI-CNR, Pisa

Diligent

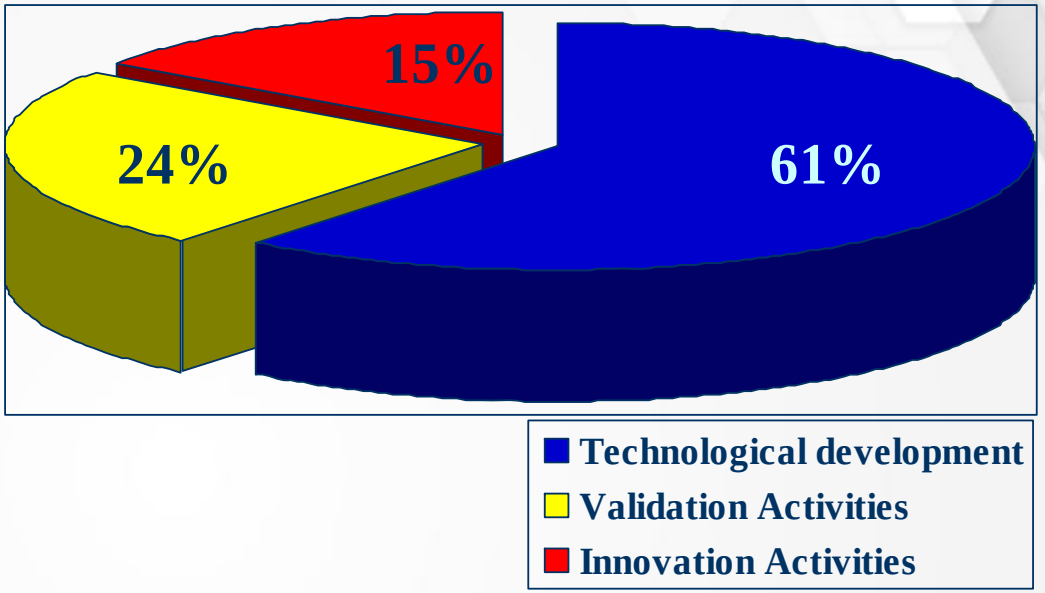
Outline

- Project Description
- Interaction with EGEE
- gLite DILIGENT Infrastructures
- gLite Experimentation
- Problem Using gLite Services
- DILIGENT Requirements
- Future plans

Diligent

Project Description

- Duration: **36** Months
- Start Date: Sept 2004
- Person/Months: **1024**
- Total Costs: **9.5 M €**
  (6.3 M € from EU)



Objective: Create a Digital Library Infrastructure that will allow members of dynamic virtual research organizations to create on-demand transient digital libraries based on shared computing, storage, multimedia, multi-type content, and application resources

Diligent

Participants

- Italian National Research Coucil – ISTI (Italy, Scientific Co-ordinator)
- European Research Consortium for Informatics and Mathematics (France, Administrative Co-ordinator)

- European Organization for Nuclear Research (Switzerland)

- Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. – IPSI (Germany)
- University of Athens (Greece)
- University of Basel (Switzerland)
- University for Health Informatics and Technology Tyrol (Austria)
- University of Strathclyde (United Kingdom)

- Engineering Ingegneria Informatica SpA (Italy)
- Fast Search & Transfer ASA (Norway)
- 4D SOFT Software Development Ltd. (Hungary)

- European Space Agency – ESRIN (Italy)
- Scuola Normale Superiore (Italy)
- RAI Radio Televisione Italiana (Italy)

Diligent

DILIGENT DL infrastructure

Consumers

Implementation of Environmental Conventions

Research on Culture Heritage

Service A

Service B

Service C

DLCreation service

Service D

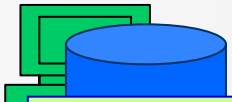
Service E

gLite

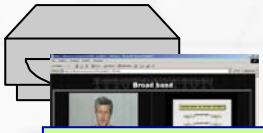
Producers

3D processing

simulation



Feature extraction



Speech recognition

Diligent

Interaction with EGEE

Coordination with EGEE

- Technical interactions
  - 9 technical meetings (mainly with JRA1)
  - gLite mailing lists subscription:
    - glite-discuss@cern.ch
    - project-diligent-glite@cern.ch
  - 1 training on “Grid Technologies for Digital Libraries”
  - 1 tutorial on “gLite Deployment”
- Other interactions
  - 4 EGEE conferences (Cork, The Hague, Athens, Pisa)

Diligent

Interaction with EGEE

Feedback to EGEE

- On EGEE activities
  - gLite bugs submission (JRA1)
- On DILIGENT project
  - status
  - access to EGEE prototype testbeds (JRA1)
  - access to EGEE PPS testbed (SA1)
  - grid related DL requirements (JRA1, NA4)
  - future plans

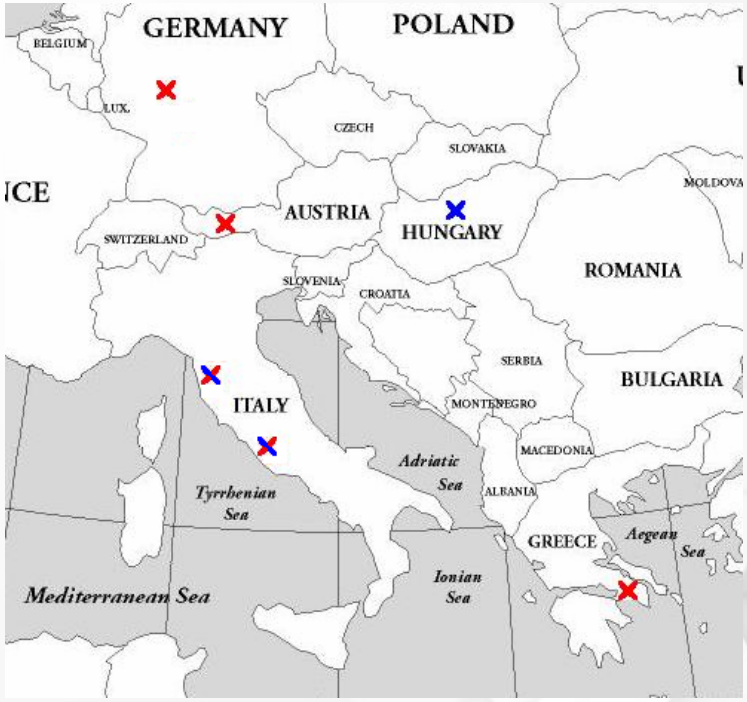
Diligent

gLite DILIGENT Infrastructures

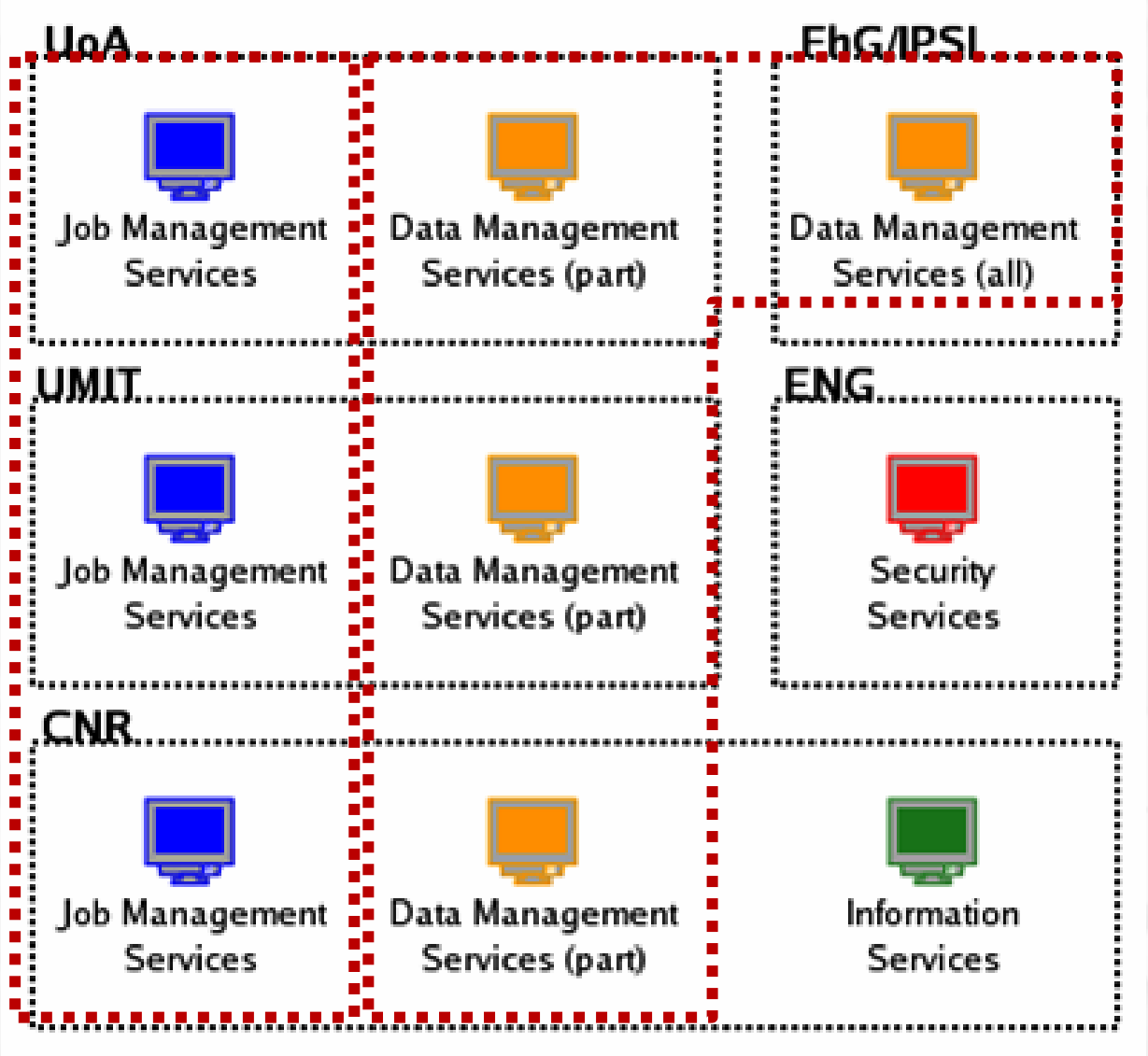
- DILIGENT has 2 independent infrastructures (gLite v1.4)
  - Development infrastructure
  - Testing infrastructure

- Infrastructures are geographically distributed, linking 6 sites in Athens, Budapest, Darmstadt, Pisa, Innsbruck and Rome

- Running gLite experimentation tests since July 2005

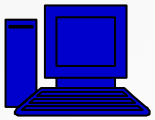


Development Infrastructures



Testing Infrastructure

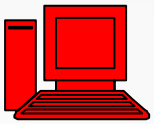
4DSOFT



Job Management Services

Data Management Services

ENG



Security Services

CNR

Information Services

Diligent

gLite Experimentation

- Goal
  - store/manage collections of objects
  - run applications organized in DAGs
  - store the application results for future usage
- Tests plan
  - Data Upload
  - Job Submission
  - Data transfer
- Data

  800K XML files of the Reuters corpus (from Aug96 to Aug97)
- Application

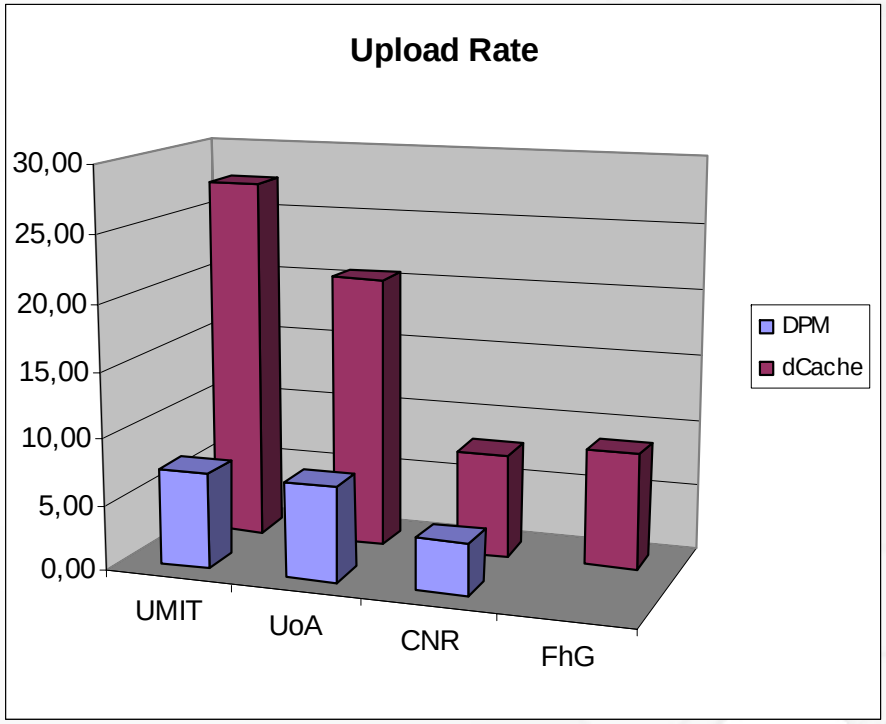
  Feature extraction tool (JIRE Application)

- Implementation of prototypes to test the feasibility of the proposed solutions

Diligent

gLite Experimentation – Data Upload

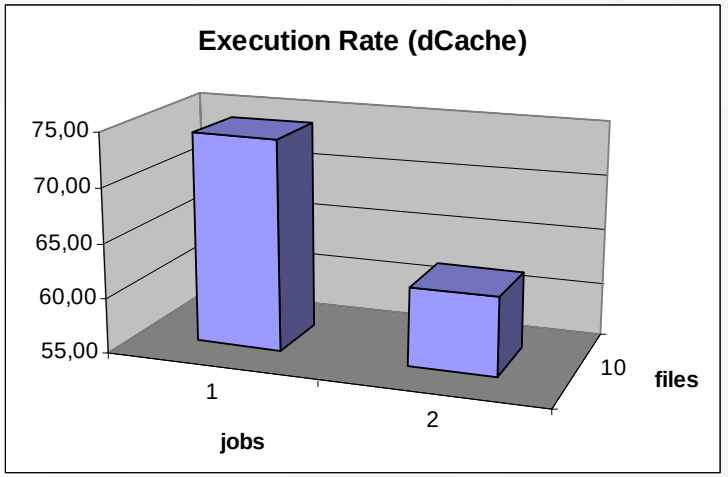
- Two Mass Storage Systems (MSS) were tested: dCache and DPM
- dCache:
  - success rate: 69,06 %
  - avg. rate: 16,18 s/file
  - several problems!
- DPM:
  - success rate: 97,26 %
  - avg. rate: 6,10 s/file

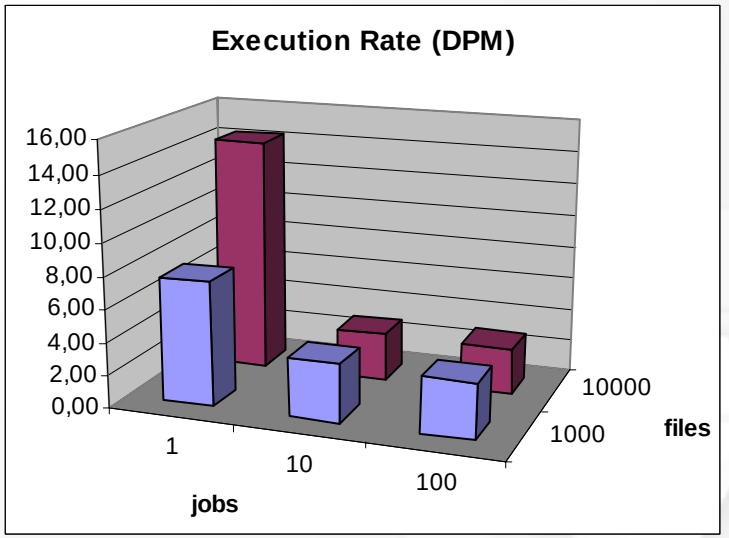


Diligent

gLite Experimentation – Job Submission

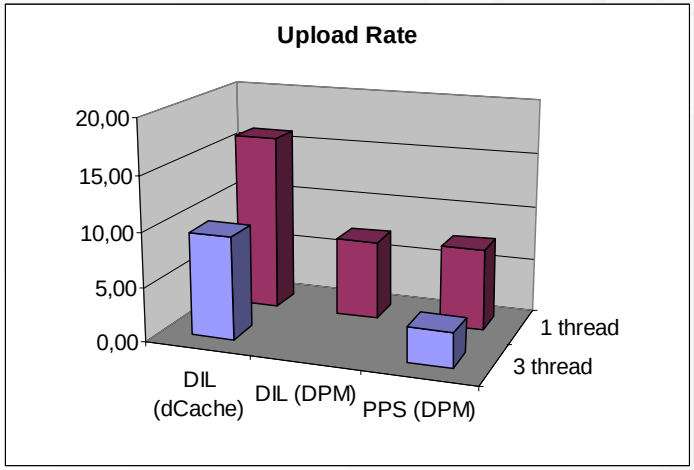
- Jobs using dCache data MSS:
  - several problems!
- Jobs using DPM data MSS:
  - success rate: 100%
  - avg. rate: 5,77 s/file
  - comparable performance using 10 and 100 jobs due to the small number of available worker nodes

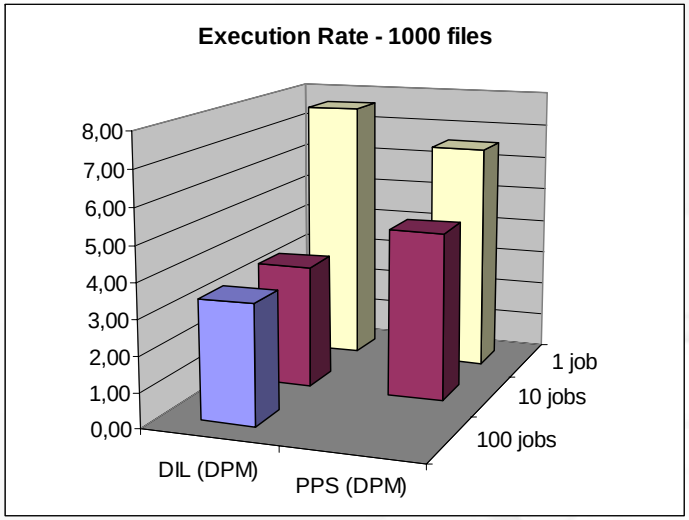




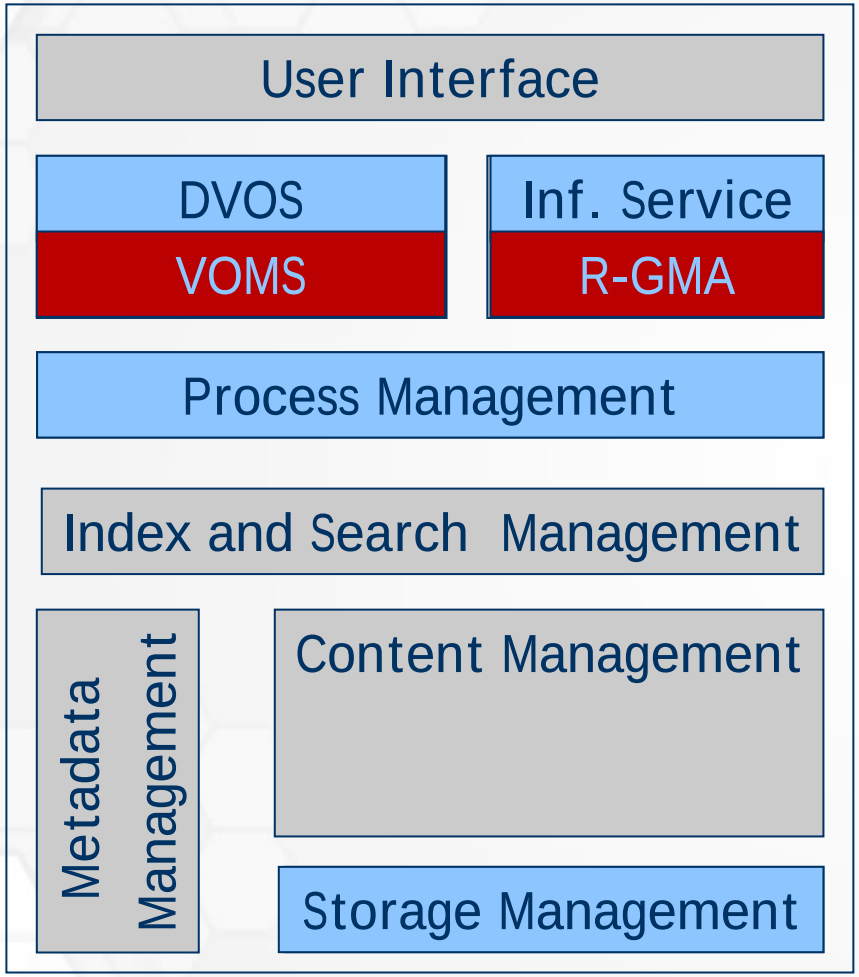
gLite Experimentation

- DILIGENT Vs PPS infras.
- Data upload
  - similar results (for DPM)
- Job submission
  - similar results
  - DILIGENT dCache not considered (didn't work with 1000 files)

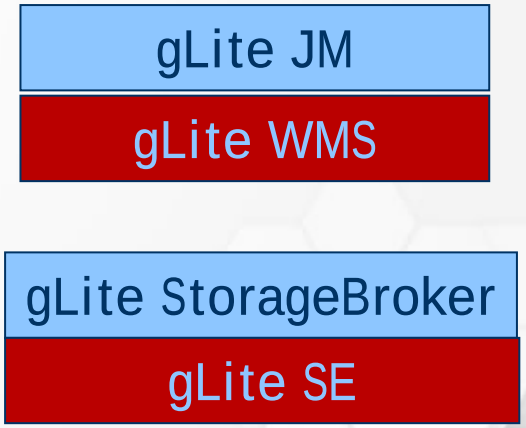




gLite Experimentation



The experimental DILIGENT DL exploits gLite storing and processing on demand the stored products on the GRID. This allows to produce usable end-user manifestations upon requests.



Diligent

Problem Using gLite Services

- gLite deployment
  - gLite architecture and configuration are complex
  - gLite 1.0 was released in April 2005 (since then four new releases were made available)
  - limited information available (it has been made available gradually)
  - several bugs were found in deploying and using gLite (many are solved)
- Software porting to 64 bit is not complete. Some gLite services (WMS, CE) can't be deployed on 64 bit machines.

Diligent

Problem Using gLite Services [cont]

- Job submission:
  - Slow Job execution phase
  - Anyway gLite job management system showed to be reliable:
    - more jobs
    - same performance
- Data upload:
  - A lot of performance issues using DCache backend
    - gLite-put/gLite-get/gLite-rm simultaneous
    - large amount of small files
  - DILIGENT needs 100% successful upload rate-> DPM
  - dead-links on Fireman when glite-put ends with errors

Diligent

DILIGENT Requirements

- DILIGENT aims to run executables that repeat the same operations for each input files belonging to a given collection.
- Each single execution takes few minutes (or less) but it must be repeated for hundreds of thousands times (even millions).
- These executables usually are organised in a DAG to deliver a more complex functionality

Diligent

DILIGENT Requirements [cont]

In order to support this framework, it should be possible:

- *To query for the maximum number of CPUs concurrently available*
  - in order to allow to a DILIGENT high level service to automatically prepare a DAG where each node will be entitled to process a partition of the data collection
- *To use parametric jobs/automatic partitioning on data*
  - Submission of a same computation on a set of n input data should be more efficient than the submission of n jobs
- *To use Condor as LRMS (Local resource management System)*

Diligent

DILIGENT Requirements [cont]

- *To support service certificate*
  - it should be possible to obtain a service certificate for a high level service
- *To specify a job specific priority*
  - the same user/service should be able to specify priorities for his/its own jobs
- *To specify a priority for a user or for a service*
  - it is required to prioritize the DILIGENT infrastructural services jobs with respect to the end-user services requests

Diligent

DILIGENT Requirements [cont]

- *To ask for on-disk encryption of data*
  - It should be possible to ask for encryption of the data on disk to prevent data leaks at the storage site level
- *To dynamically manage VO creation*
  - The creation of a new VO should be supported without deploying and configuration of services by hand
- *To dynamically support user/service affiliation to a VO*
  - The user/service affiliation to a VO should be automathized as much as possible

Diligent

Future Plans

- Monitor gLite developments and continue the current work of deploying gLite in DILIGENT infrastructures
- Continue the ongoing gLite experimentation using DILIGENT and EGEE PPS infrastructures
- Continue gridifying the following services needed in the DILIGENT DL experimentation.
  - Metadata Management
  - Content Management
  - Index and Search Management
  - Process (workflow) Management

Diligent

Tips / Summary

- DILIGENT has successfully installed and now maintains its own gLite infrastructures. DILIGENT development infrastructure can join the EGEE infrastructure
- An active EGEE-DILIGENT collaboration has been established and this has been key for the achievement of our first goals
- DILIGENT has identified a concrete set of open issues that we need to address. The gLite and DL experimentation activities have shown that we are on the right track

Diligent

DILIGENT Web Site http://www.diligentproject.org

DILIGENT Training DL http://diligent-training.isti.cnr.it

Experimental DL http://diligent-dl1.isti.cnr.it

Andrea Manzi andrea.manzi@isti.cnr.it

Thank you