Diligent

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

DILIGENT
(technical overview)

Pasquale Pagano
ISTI-CNR, Pisa

Pedro Andrade
CERN, Geneva

on behalf of the DILIGENT Consortium

Diligent

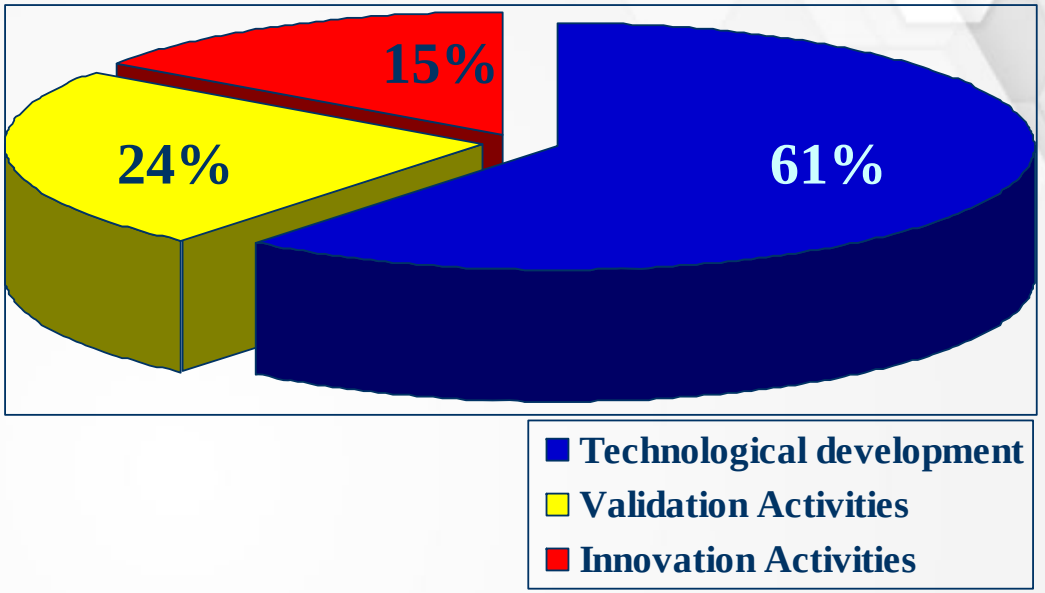
Outline

- Project Description
- Interaction with EGEE
- gLite DILIGENT Infrastructures
- gLite Experimentation
- DILIGENT Services
- Tips/Summary

Diligent

Project Description

- Duration: **36** Months
- Start Date: Sept 2004
- Person/Months: **1024**
- Total Costs: **9.546.561€ (6.300.000€ 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

Project Description

- European Research Consortium for Informatics and Mathematics (ERCIM)
- Institute of Information Science and Technology of the Italian National Research Council (CNR-ISTI)
- Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V. (FhG/IPSI)
- European Organization for Nuclear Research (CERN)
- University of Athens (UoA)
- University for Health Informatics and Technology Tyrol (UMIT)
- University of Strathclyde (USG)
- University of Basel (UNIBAS)
- Engineering Ingegneria Informatica SpA (ENG)
- Fast Search & Transfer ASA (FAST)
- 4D SOFT Software Development (4D SOFT)
- Scuola Normale Superiore di Pisa (SNS)
- European Space Agency (ESA)
- Italian National Broadcaster (RAI)

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)

Interaction with EGEE

Feedback to EGEE

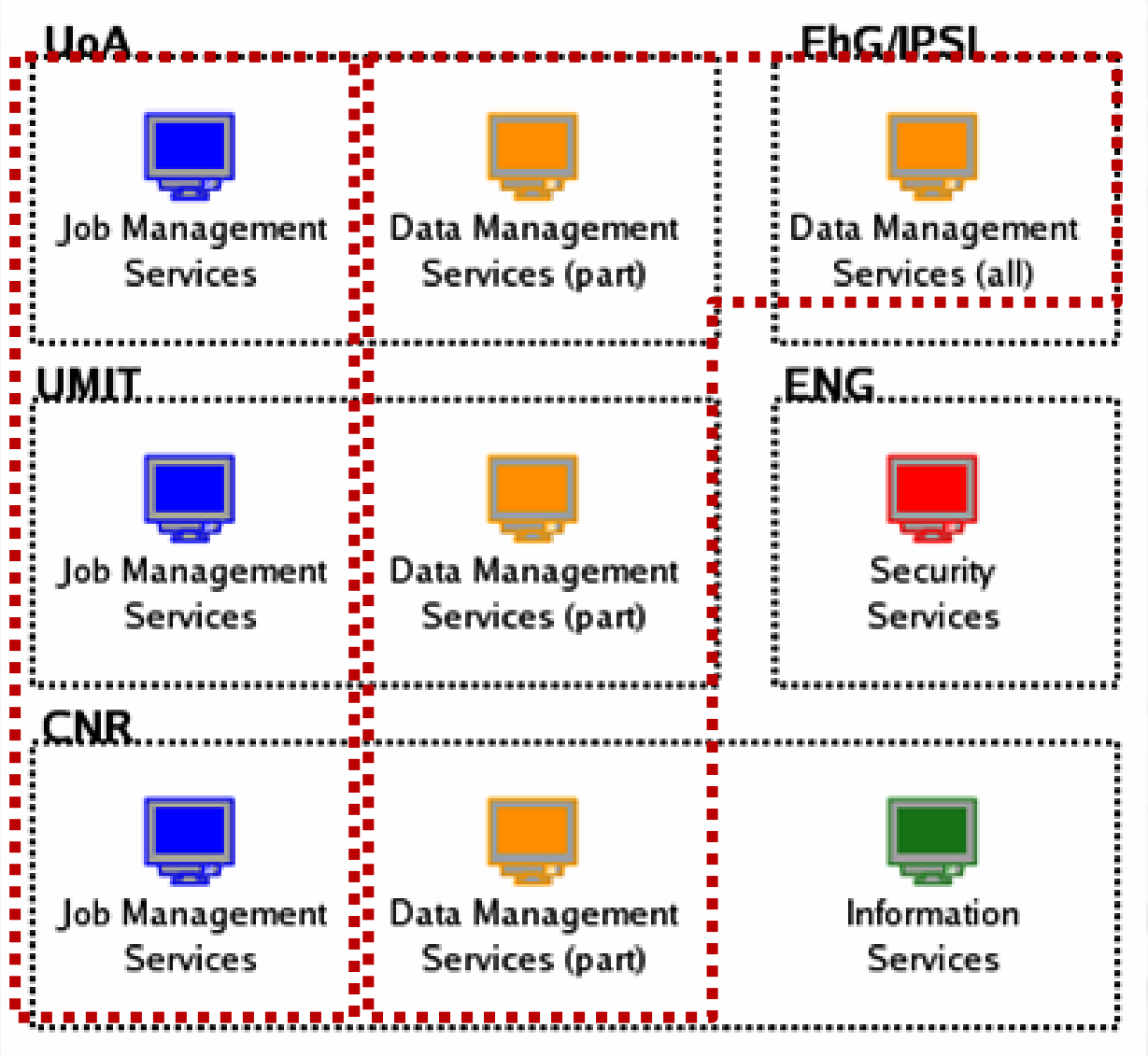
- On EGEE activities
  - gLite bugs submission (JRA1)
- On DILIGENT project
  - status
  - access to EGEE prototype testbeds (JRA1)
  - access to EGEE PPS testbed (SA1)
  - grid related DL requirements (JRA1, NA4)
  - future plans

Diligent

gLite DILIGENT Infrastructures

- DILIGENT has 2 independent infrastructures
  - 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

gLite DILIGENT Infrastructures



gLite Experimentation

- Understand how DILIGENT can profit from the gLite middleware produce useful material for DILIGENT developments
- A DL framework has standard requirements:
  - Store/manage collections of objects
  - Run applications (sometimes organized in dags or workflows) that produce some output
  - Store the result of these applications (from application to application) for future usage
- The Reuters experimentation is an example of this process

gLite Experimentation

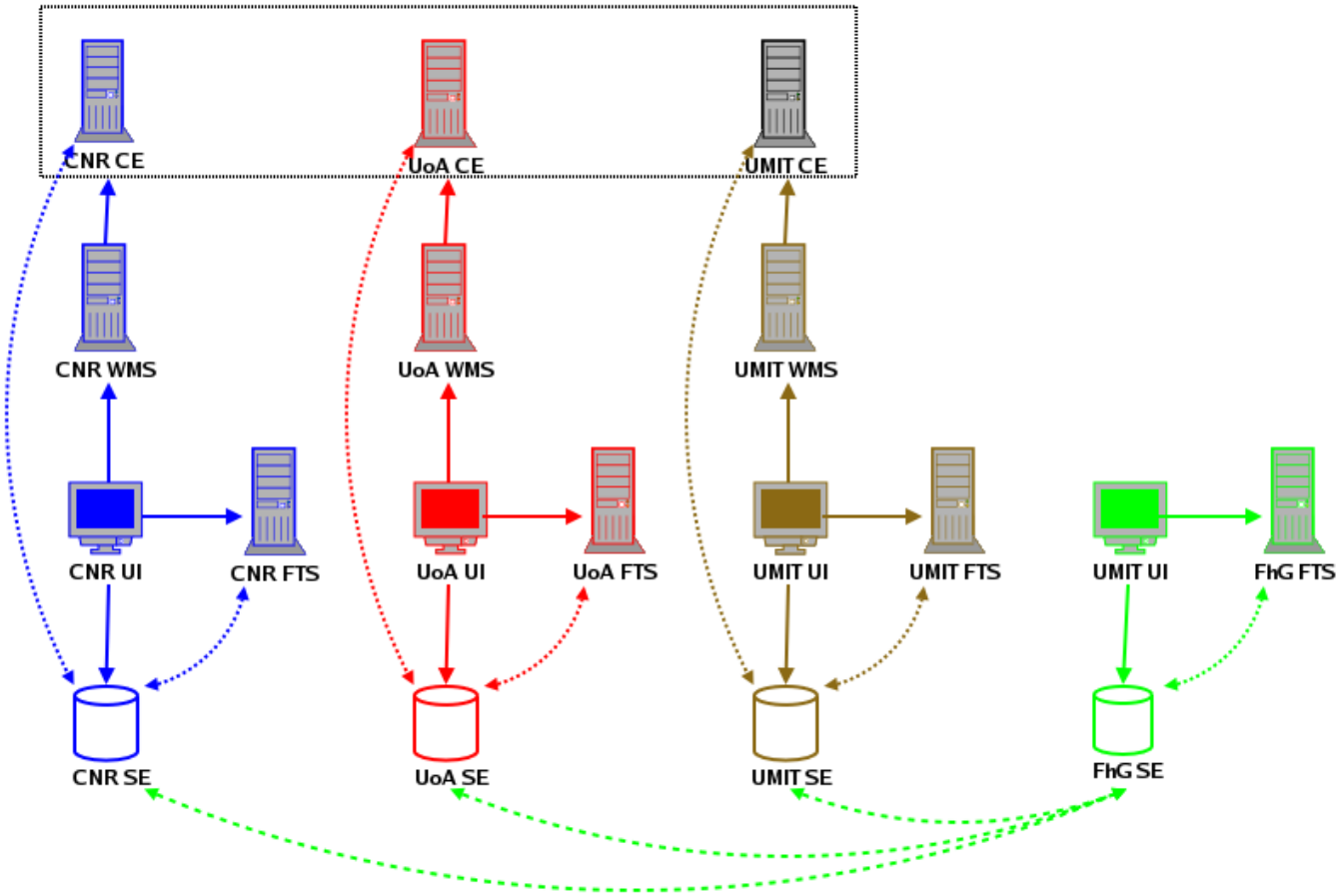
- 3 main tests:
  - Data Upload, Job Submission and Data transfer
- Data:
  - Part of the Reuters corpus (from Aug96 to Aug97). This corpus is a collection of 800.000 xml files distributed in several directories.
- Application
  - JIRE application: information retrieval framework which allows to extract features vectors in order to
    - build indexes of collections or train classifiers (such as Rocchio, SVM, Bayes, etc.)
    - compute performance measures (such as Precision, Recall and F1) using classifiers and a test set of documents

Testing Activities

Job Submission

Data Transfer

Data Upload

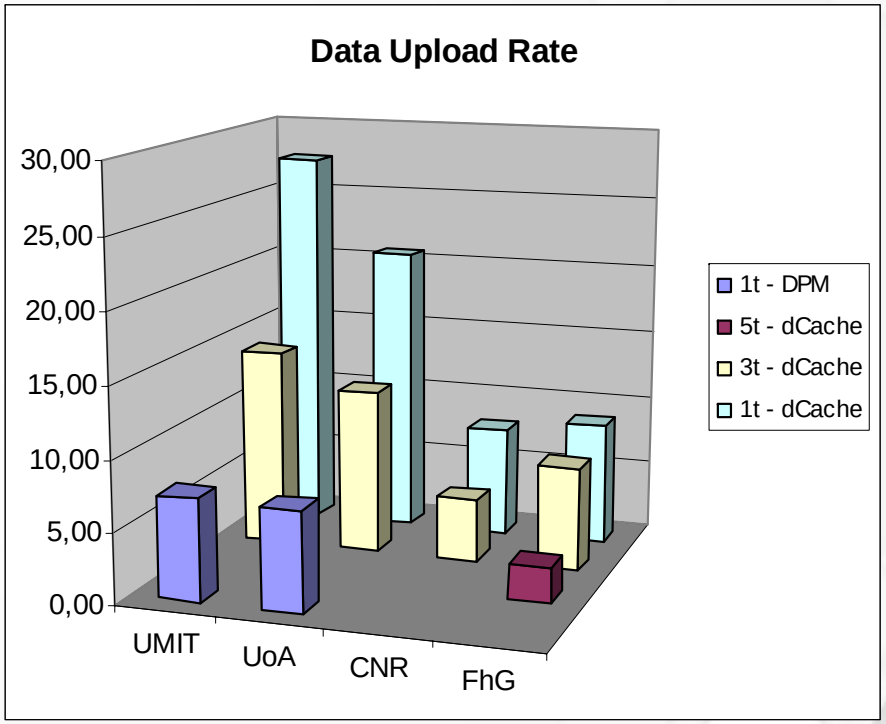


gLite Experimentation

- Data upload results

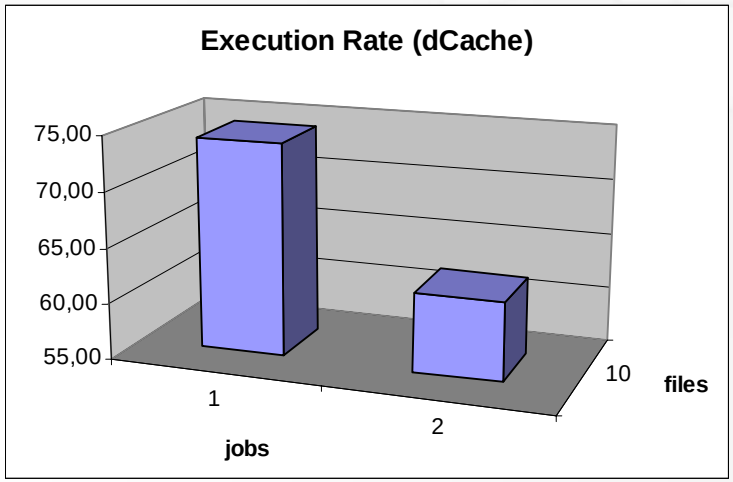
- To DILIGENT dCache SEs:
  - 1,77 GB / 368 989 files
  - success rate: 69,06 %
  - avg. rate: 11,69 s/file
  - several problems!

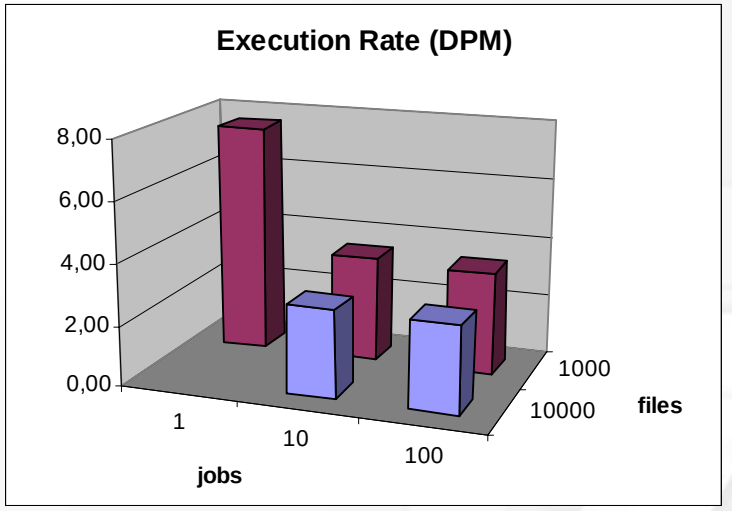
- To DILIGENT DPM SEs:
  - 198 MB / 39 418 files
  - success rate: 97,97 %
  - avg. rate: 7,19 s/file



gLite Experimentation

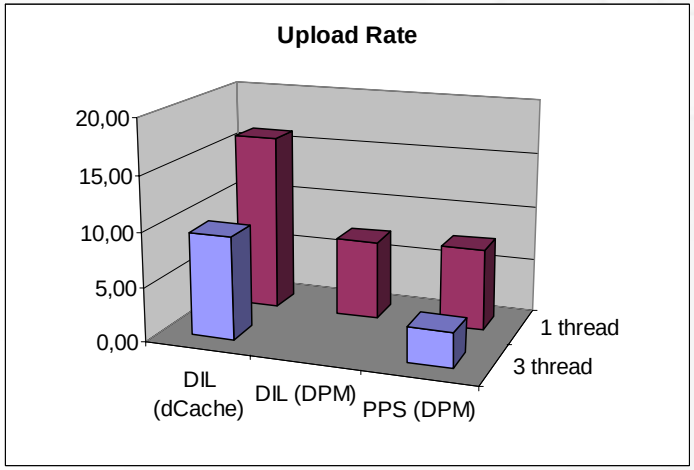
- Job submission results
- Using data from dCache:
  - 40 files/200KB processed
  - max 100 files per exec.
  - several problems!
- Using data from DPM:
  - 23k files/110MB processed
  - 10000 files per exec. ok
  - avg. rate: 4,06 s/file

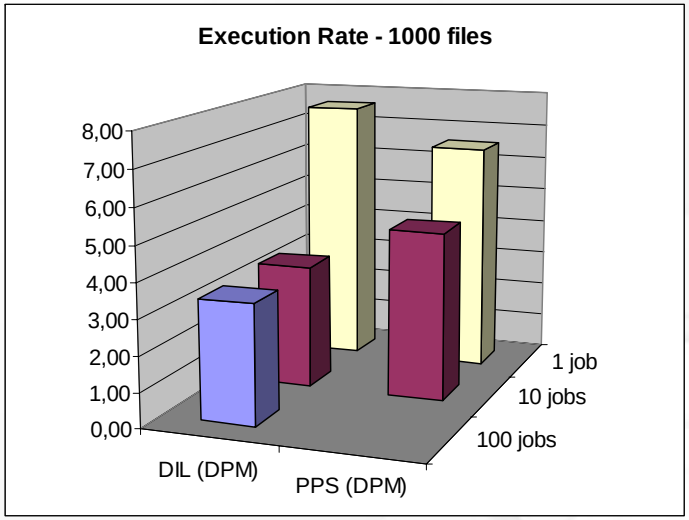




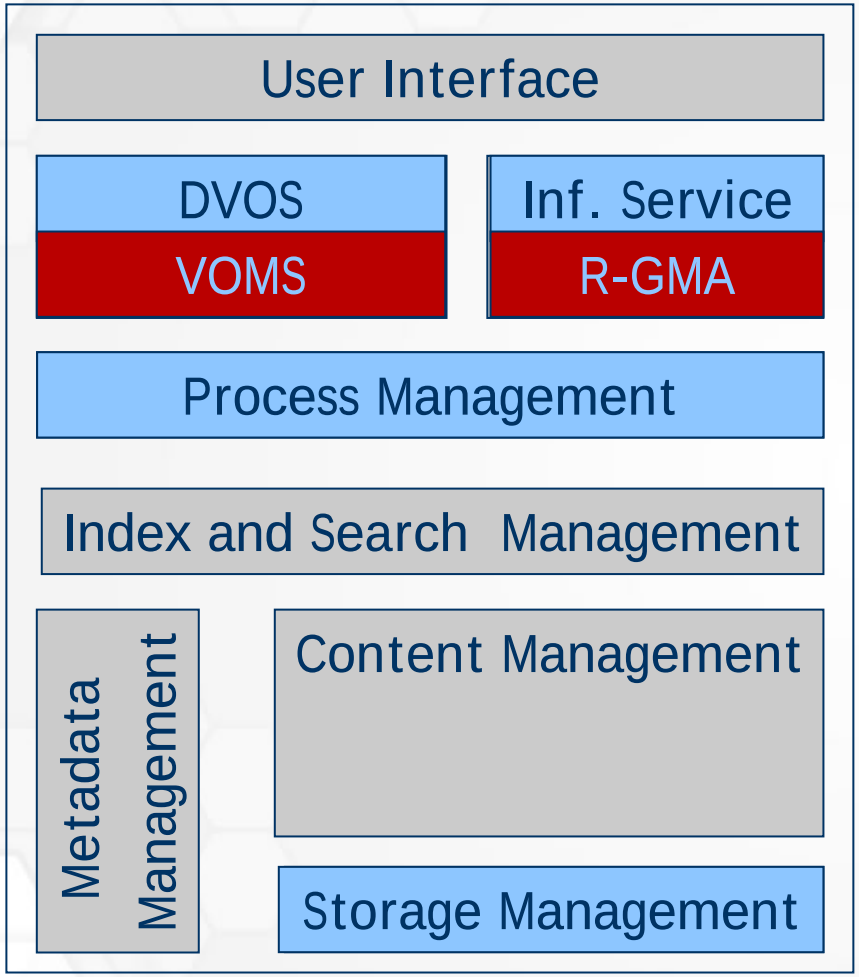
gLite Experimentation

- 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.

gLite JM

gLite WMS

gLite StorageBroker

gLite SE

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

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

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

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

Pasquale Pagano pasquale.pagano@isti.cnr.it
Pedro Andrade pedro.andrade@cern.ch

Thank you