Comparative evaluation of tools providing access to different types of data resources exposed on the Grid

A joint test program of INFN, SPACI-UNILE and INAF

Giacinto Donvito

INFN-Bari

www.eu-egee.org

Outline

- Motivations
- The tools under evaluation
- The testbed
- The test plan
- Issues
- Preliminary results
- Conclusions

Motivations

- **Provide access in GRID to Bioinformatics data stored in relational DB's**
  - BioinfoGRID EU project (http://www.bioinfogrid.eu/)
  - LIBI Italian MIUR project (http://www.libi.it/libi)

- **Provide access to multiple astronomical databases (archives, Astronomical Catalogues, etc.) for the Astrophysical community.**
  - The Virtual Observatory (http://www.euro-vo.org/pub/
  - http://www.ivoa.net/)

- **Provide access to population data for porting "public administrations" applications to the GRID**
  - EGG Italian MIUR project

- **Diffuse inside the Italian community the knowledge and the expertise to access relational DB's from Grid.**

The tools under evaluation

- **GRelC:** *Grid Relational Catalog*
  - *Developed by SPACI Consortium and University of Salento, Lecce*
    - Giovanni Aloisio (giovanni.aloisio@unile.it)
    - Sandro Fiore (sandro.fiore@unile.it)
  - Project site: *http://grelc.unile.it*

- **G-DSE (INAF + INFN)**
  - *Developed by INAF and INFN*
    - Edgardo Ambrosi (amborsi@cnaf.infn.it
    - Giuliano Taffoni (taffoni@oats.inaf.it)
    - Andrea Barisani (lcars@infis.units.it)
  - Project site: *http://wwwas.oats.inaf.it/grid/G-DSE*

- **OGSA-DAI**
  - *Developed as part of the Open Middleware Infrastructure Institute UK (OMII-UK) project.*
  - Project site: *http://www.ogsadai.org.uk/index.php*

The GRelC Project

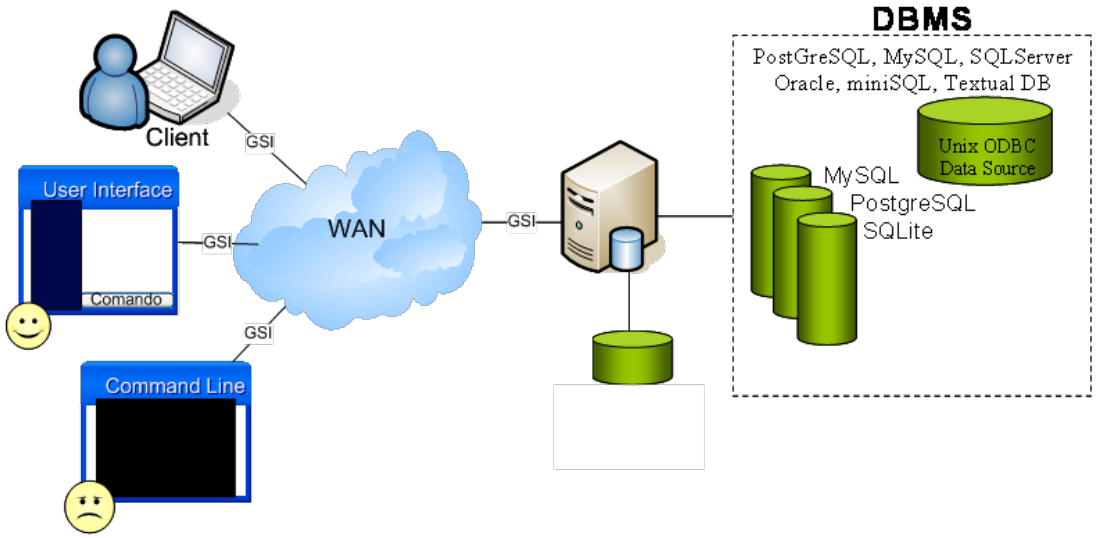
- *Grid Relational Catalogue* is a project which aims at designing and developing a set of efficient, secure and transparent Data Grid Services (Starting date, January 2001).

- *Developed by SPACI Consortium and University of Salento, Lecce*

- *GRelC Data Access Service* aims at providing a large set of functionalities to access both relational and non relational Databases in a grid environment

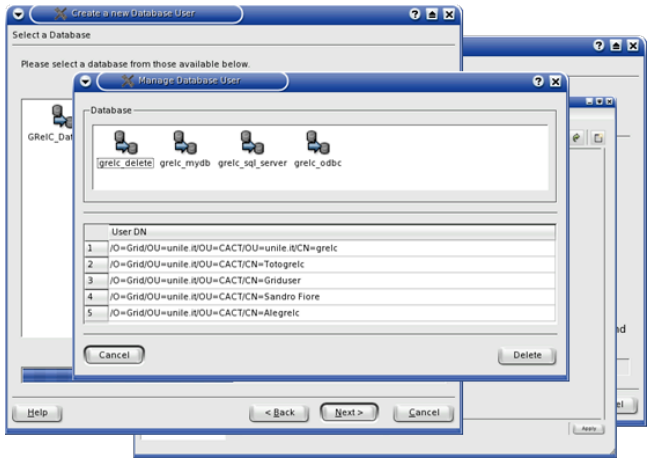
GRelC Data Access Service

- GRelC DAS is a Grid Database Access Service
- Web Service interface GSI enabled and WS-I compliant
- Mutual authentication based on GSI (X.509v3 digital certificates)
- Authorization Framework leveraging local ACL and VOMS



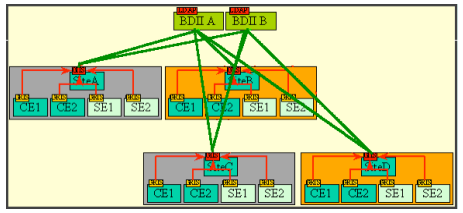
GReIC DAS: Main Features

- Protection against SQL Query Injection
- Single Query: prefetch, dime, memory, etc.
- Dinamic binding to heterogeneous DBMSs:
  - Postgres, MySQL, SQLite
  - DB2, Oracle
  - Microsoft SQL Server
  - UnixODBC
- Graphical User Interface (Qt based)



GRelC DAS and gLite

- Information System Support (BDII compliant)



- Porting on LCG-2-4-0, LCG-2-7-0 and gLite 3.x
- It also runs on the following platforms:
  - Linux
  - MAC OS X
  - FreeBSD
  - Both IA64 and IA32 platforms are supported

- Wide deployment on GILDA t-Infrastructure
  - A tutorial is also available on Grid CT Wiki (GILDA)
    https://grid.ct.infn.it/twiki/bin/view/GILDA/GRelCDataAccessService

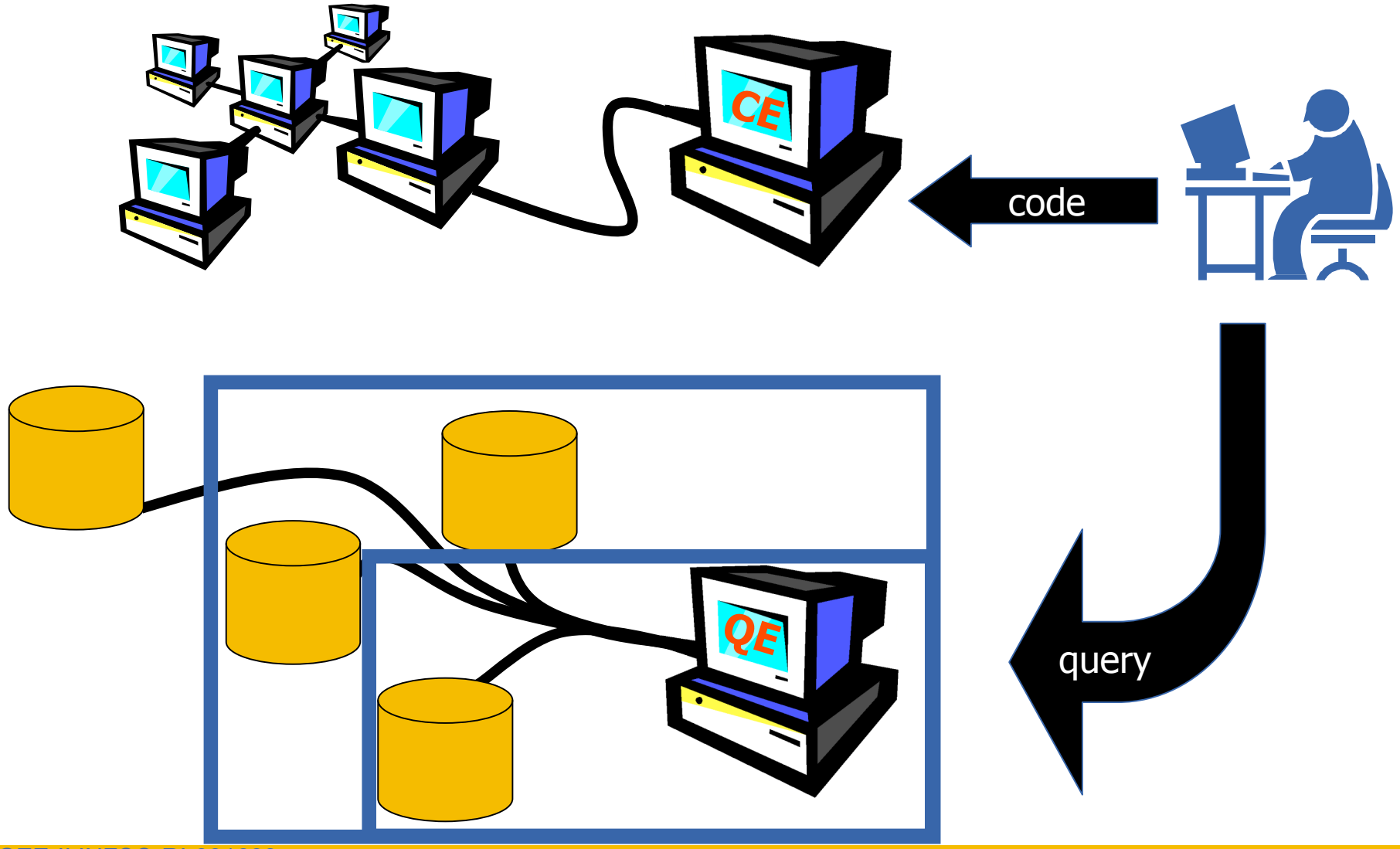
Data Source Engine Concepts

- **The Grid Resource Framework Layer, Information System** and **Data Model** is extended so that a software virtual machine as a Data Source Engine becomes a valid instance for a Grid computing model.

- A **new Grid component** (G- DSE) that enables the access to a Data Source Engine and Data Source, totally integrated with the Grid Monitoring and Discovery System and Resource Broker is defined

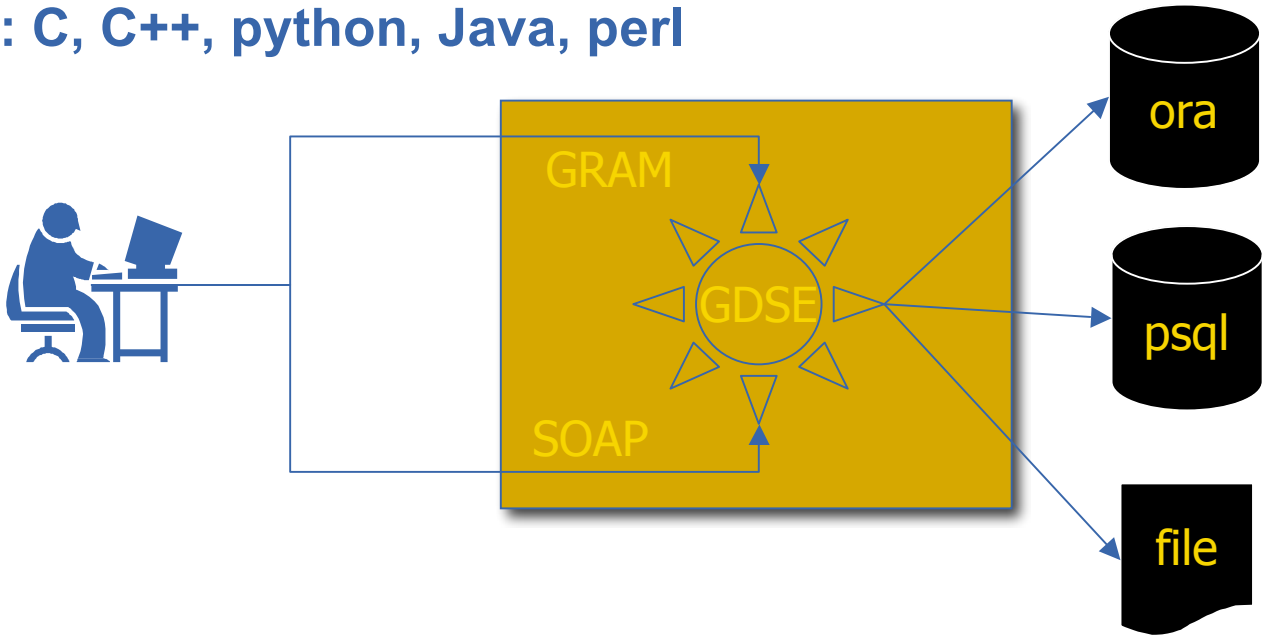
- A new **Grid Element**, **the Query Element,** can be built on top of the G-DSE component.

The Query Element



QE implementation

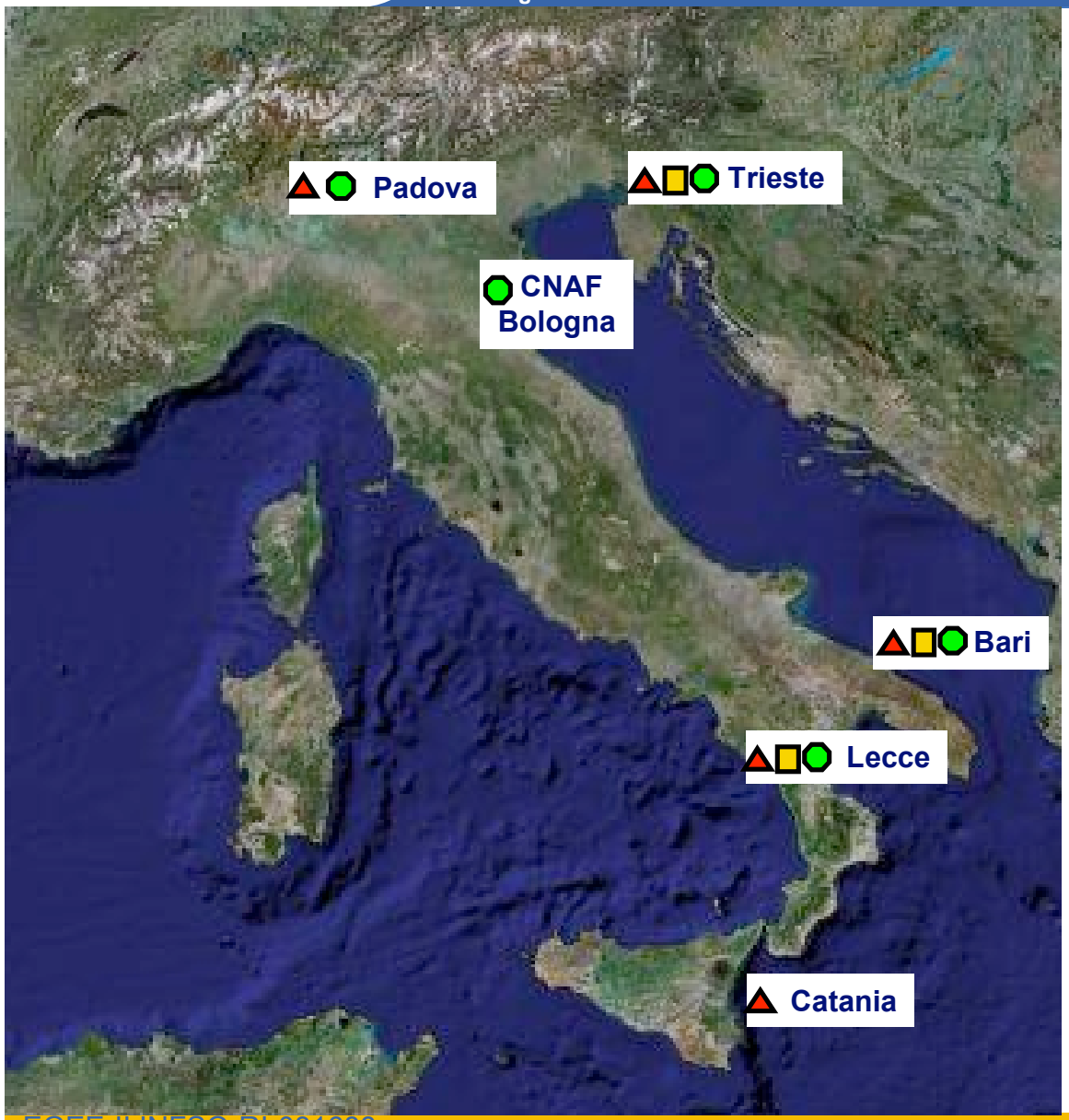
- Runs on any linux/unix flavor: GT>=2.4.3
- Backends: any DB vendor (MySQL, Oracle, PostgreSQL, etc…) + flat files
- Two protocols: GRAM or WS
- Authentication based on GSI
- Authorization based on VOMS
- API: C, C++, python, Java, perl



OGSA-DAI

- OGSA-DAI can support the following:
  - Different types of data resources - including relational, XML and files - can be exposed via web services. A number of popular data resource products are supported.
  - Data within each of these types of resource can be queried and updated.
  - Data can be transformed (using XSLT), compressed and decompressed (using ZIP and GZIP compression).
  - Data can be delivered to clients, other OGSA-DAI web services, URLs, FTP servers, GridFTP servers, or files.
  - Requests to OGSA-DAI web services have a uniform format irrespective of the data resource exposed by the service. (though the actions specified within each request may be data resource-specific).
  - Information on the data resources exposed by an OGSA-DAI web service and the functionality supported by the service can be accessed by clients.
  - OGSA-DAI users can extend OGSA-DAI web services to expose their own data resources and to support application-specific functionality, in addition to that provided by OGSA-DAI.

TESTBED



GRelC ▲
G-DSE ■
OGSA-DAI ●

Test Database: Bioinformatics database containing just a "molecule table" with about 500.000 tuples (350MB, PostgreSQL).

Other Databases:

Sakila (MySQL 23 tables)
World (MySQL 6 tables)
Dellstore (PostgreSQL 8 tables)
uniutrdb_test (MySQL 35 tables)
go_5_06 (MySQL 18 tables)
homo_sapiens (MySQL 74 tables)
2MASS (PostgreSQL)
Population db (PostgreSQL)

The test plan

- **Sequential tests** (preliminary results)
  - extraction of zero tuples - Estimate GSI services overhead
  - extraction of 10, 100, 1000, 10000, 100000 simple tuples
  - extraction of result sets of increasing dimensions: O(kb), O(MB), O(100MB)
  - extraction of 10, 100, 1000, 10000, 100000 with increasing table complexity
  - Submission of complex queries (join, multiple queries, etc)
  - Submission of **INSERT, UPDATE,** and **DELETE** queries
    - Performance evaluation for a single action
  - Evaluate the differences between LAN and WAN queries

The test plan

- **Concurrent tests** (planned, not executed)
  - Whit 10 concurrent clients extract
    - **Zero, 10, 100, 1000 tuples**
  - Repeat the extractions with O(100) concurrent clients

- **Use a common working environment for the three tools**
  - Only GSI authentication
    - VOMS authentication not supported by some of them
  - Same type of output format
    - No date post or pre processing allowed (avoid translation in a more user friendly format (XML,….))

Issues

- No OGSA-DAI expert (developer) available in the group
- While for GRelC or G-DSE we can ask developer to implement the best working conditions in order to optimize the test results
- This was not possible for OGSA-DAI
- The OGSA-DAI results presented here, have therefore not been optimized
- They represent what a generic user can get just following the published instructions and recommendations

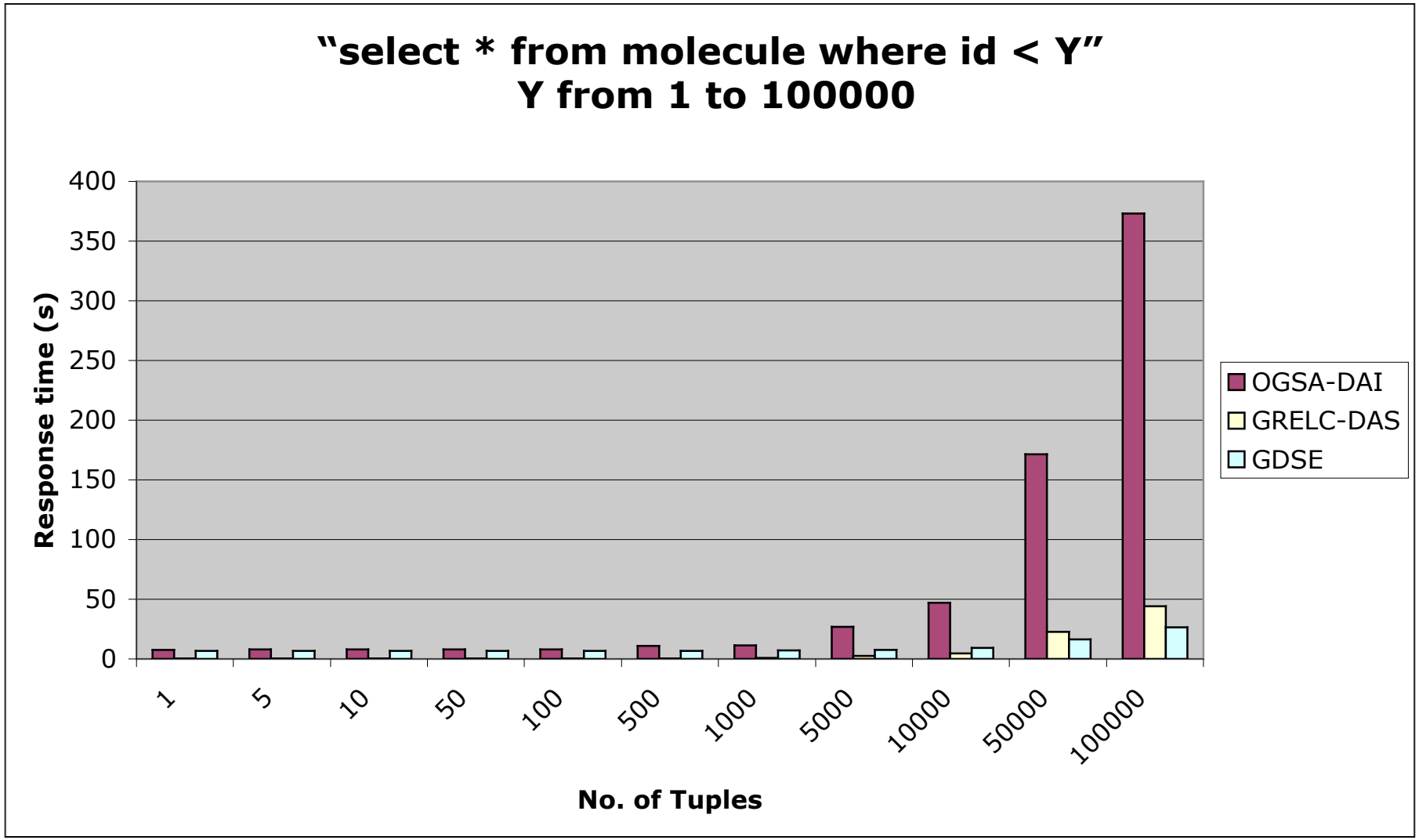
Tools installation

- **GReIC:**
  - Easy to install and configure
  - Postgresql and Mysql drivers currently available from the website. New drivers related to Oracle, DB2, etc. are in a preproduction phase.

- **G-DSE:**
  - The installation and configuration procedure was not straightforward.
  - Installation based on yaim (glite tool)
  - Found problems in almost all the sites

- **OGSA-DAI:**
  - The installation and configuration procedure was not straightforward.
  - Found problems in almost all the sites
  - Greater variety of data sources accessed

preliminary results

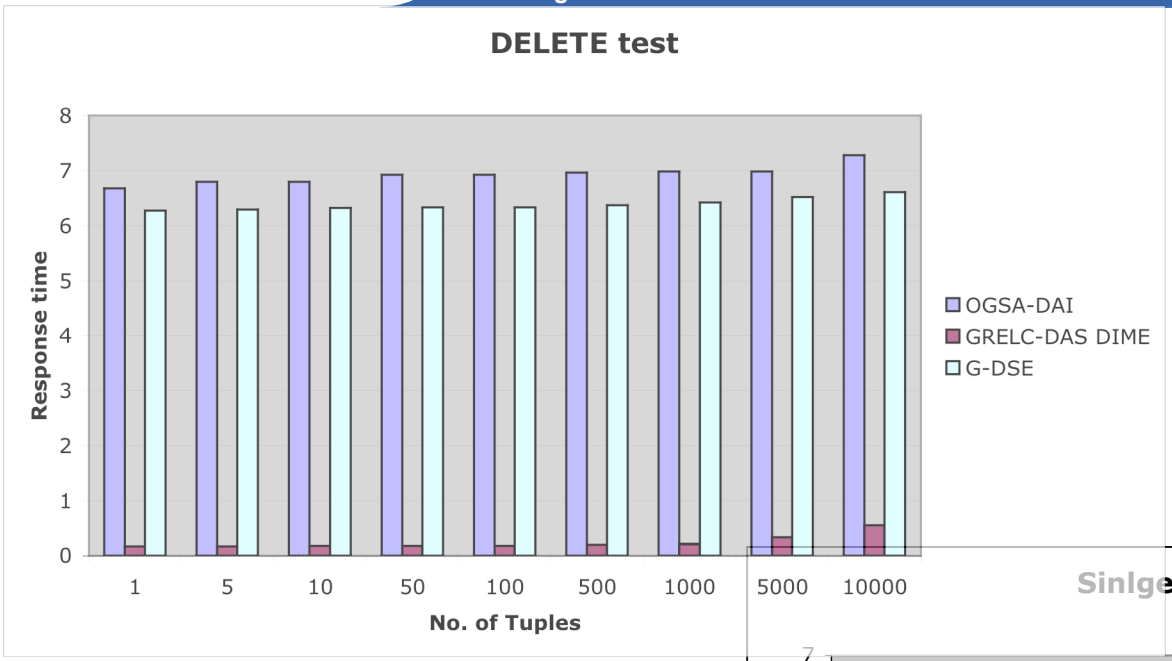


- Response time as a function of the No. of Tuples

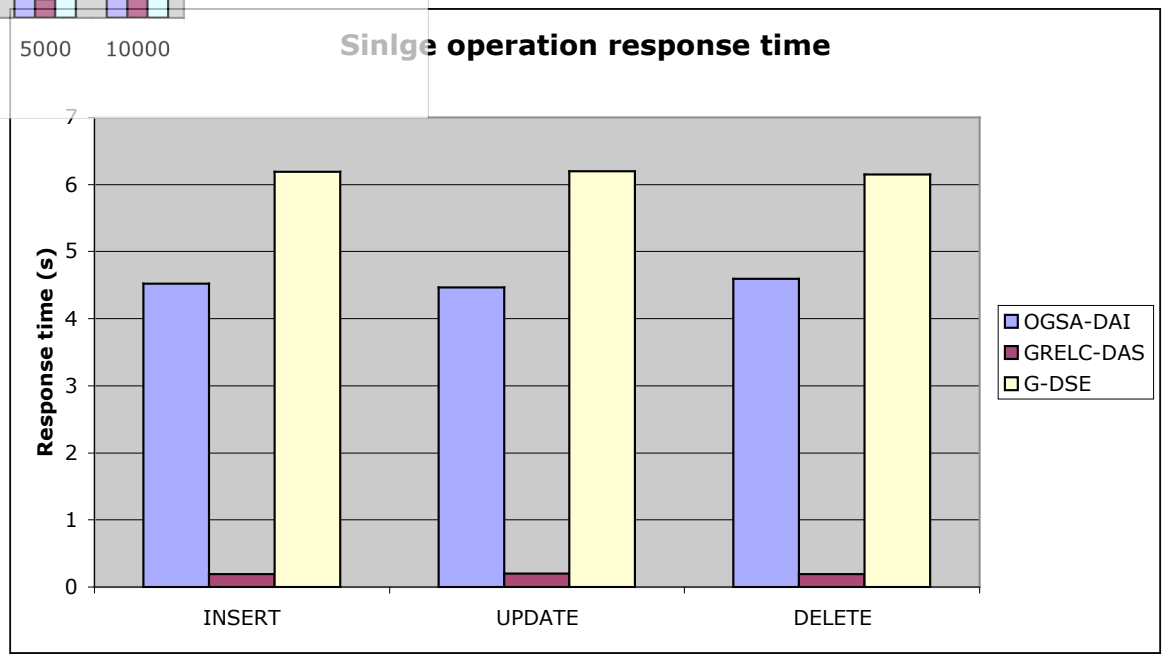
preliminary results

No.oftuples	Size (kB)	OGSA-DAI	GRELC-DAS	GDSE
1	0.022	7.70s	0.23s	6.757s
5	2	7.82s	0.23s	6.709s
10	4.5	7.88s	0.24s	6.717s
50	22	8.06s	0.25s	6.699s
100	42	8.18s	0.26s	6.838s
500	256	10.88s	0.47s	6.731s
1000	542	11.42s	0.65s	6.985s
5000	2900	26.73s	2.40s	7.430s
10000	5900	47.24s	4.80s	9.125s
50000	29000	171.384s	22.78s	16.391s
100000	57000	372.947s	44.24s	26.282s

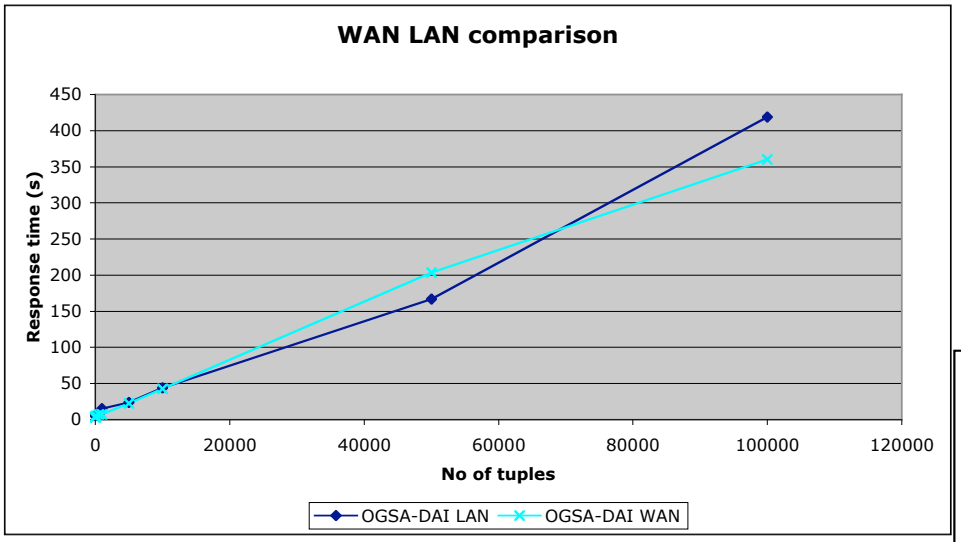
preliminary results



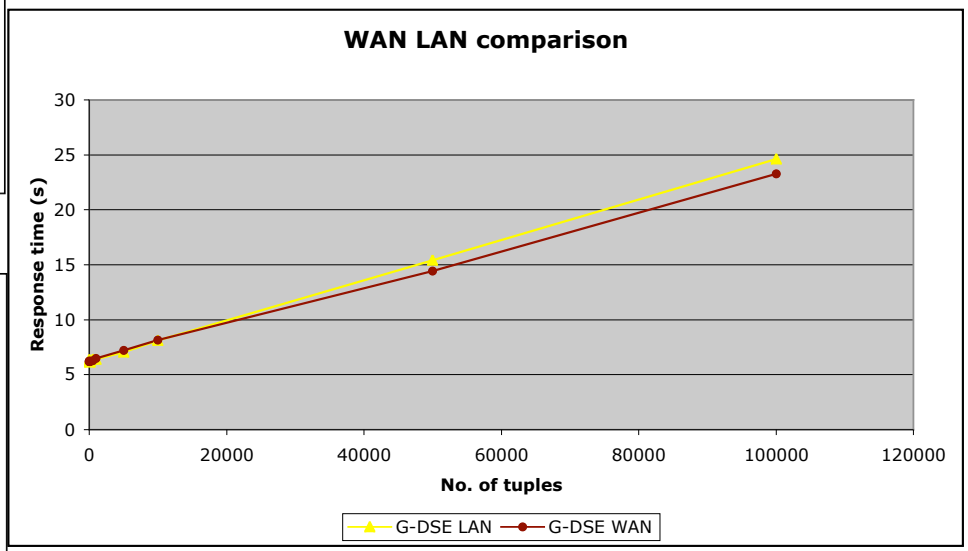
Test Database: Sakila Database (MySQL)

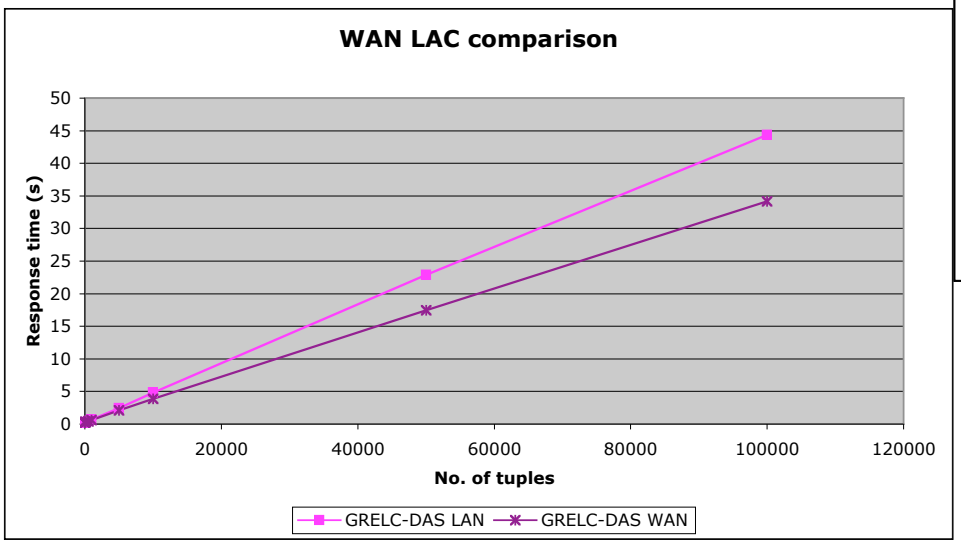


preliminary results



Test Database:
Bioinformatics database containing just a "molecule table" with about 500.000 tuples (350MB, post).





Conclusions

- **The test are still on going**
  - Some input from OGSA-DAI developers could be very appreciated
- **From the first results it seems that each middleware has some specific field of excellence:**
  - OGSA-DAI:
    - has lot of advanced features
    - Widely distributed
  - GRelC:
    - Very fast for small/medium datasets. Good performance for large datasets
    - Fast DML query submission (very low GSI latency).
    - Easy to install. Already ported on gLite (VOMS enabled)
  - GDSE
    - Very efficient for huge query
    - Easy to manage
    - Good integration with gLite environment (BDII and VOMS enabled)

Acknowledgments

- TAFFONI, Giuliano (INAF)
- VUERLI, Claudio (INAF)
- BARISANI, andrea (INAF)
- PASIAN, Fabio (INAF)
- MANNA, Valeria (INAF)
- GISEL, Andreas (CNR-ITB)
- GIORGIO, Emidio (INFN)
- AIFTIMIEI, Cristina (INFN)
- ATUL, Jain (INFN+Politecnico Bari)
- BARBERA, Roberto (INFN+Università Catania)
- CAROTA, Luciana (INFN)
- DONVITO, Giacinto (INFN)
- GHISELLI, Antonia (INFN)
- LA ROCCA, Giuseppe (INFN)
- MAZZUCATO, Mirco (INFN)
- PIERRO, Antonio (INFN)
- VERLATO, Marco (INFN)
- FIORE, Sandro (Univ. del Salento, Lecce)
- ALOISIO, Giovanni (Univ. del Salento, Lecce)
- CAFARO, Massimo (Univ. del Salento, Lecce)
- VADACCA, Salvatore (Univ. del Salento, Lecce)
- NEGRO, Alessandro (Univ. del Salento, Lecce)
- DEL FREO, Federico (EGG project)
- RICCI, Giovanni (EGG project)

BioinfoGRID
Bioinformatics Grid Application for life science

- Work supported in part by BioinfoGRID and LIBI projects

LIBi