

The Interoperability between Geospatial Services and Grid Infrastructure

Denisa Rodila, Dorian Gorgan
Computer Science Department
Technical University of Cluj-Napoca
[denisa.rodila](mailto:denisa.rodila@cs.utcluj.ro), [dorian.gorgan](mailto:dorian.gorgan@cs.utcluj.ro)@cs.utcluj.ro

Outline

- Scope & Objectives
- Interoperability Context
- enviroGRIDS Project
- Interoperability Challenges
- OGC Web Services
- Advantages of Grid technology
- Solutions and Approaches
- Experimental Tests
- Conclusions

Scope & Objectives

□ Scope

- Design and implement an architecture to support integration of geospatial domain (represented by OWS standards) in the Grid environment

□ Objectives

- Analyze and propose different solutions and approaches to bridge the gap between the Geospatial and the Grid infrastructures
- Propose different “gridification” levels of the OGC Web services and solve the interoperability problem in the context of enviroGRIDS project
- Present a proof of concept for the proposed solutions
- Propose new standards in the OGC and OGF collaboration

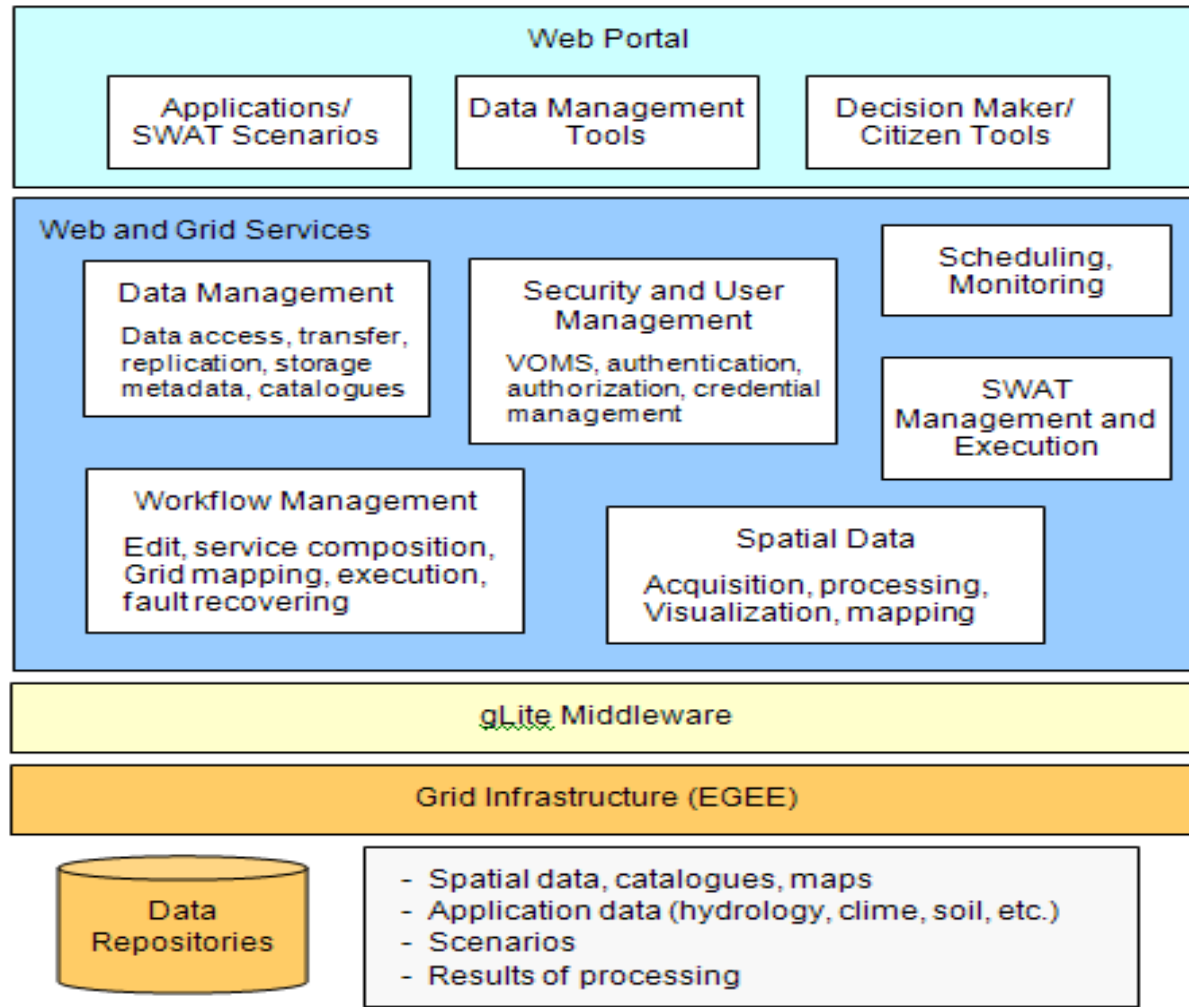
Interoperability Context

- OGC – Open Geospatial Consortium
 - represented by OGC Web services
- OGF – Open Grid Forum – OGF
 - represented by the gLite middleware
- G-OWS – gLite OWS
- Collaboration between OGC and OGF should provide the necessary infrastructure for developing tools, software and services for multiple communities
- Research projects working on the interoperability between the two platforms: SAW-GEO, CYCLOPS, GDI-Grid, GEO-Grid, DEEGREE

enviroGRIDS Project

- The FP7 **enviroGRIDS** Project (Black Sea Catchment Observation and Assessment System supporting Sustainable Development) aims to:
 - Develop a SDI (Spatial Data Infrastructure) targeting the Black Sea catchment region
 - Use new international standards to store, analyze, process, and visualize important information regarding this area
 - Perform distributed spatially-explicit simulations of environmental changes
- Project objectives:
 - Modeling large scale and high resolution distributed hydrological processes
 - Access real time data coming from sensors and satellites
 - Organize and standardize geospatial data to improve its interoperability due to the large number of heterogeneous data sources used in the project

enviroGRIDS Project – Functional Layers



Challenges

- ❑ Geospatial data come from multiple heterogeneous sources
- ❑ Geospatial data have to be accessed, integrated, analyzed and presented across a distributed computing environment
- ❑ Processing and storing resources in different formats
- ❑ Security and digital rights management
- ❑ User authentication and authorization

OGC Web Service

- Layered on top of Internet standards: HTTP, URLs, MIME, and XML World Wide Web standards
- OWS standards includes:
 - Web Map Service – WMS
 - produces maps of spatially referenced data dynamically from geographic information. It standardizes the display of information that comes simultaneously from multiple remote and heterogeneous sources;
 - Web Feature Services
 - provide data access functionality and operations on geographic features. They can be used for accessing vector data and expose it in GML-format. It standardizes the retrieval and update of digital representations of real-world entities references to the Earth surface;

OGC Web Service

- OWS standards includes:
 - Web Coverage Service – WCS
 - specified to describe and provide multidimensional coverage data. It standardizes the access to spatially extended coverages, usually encoded in a binary format and offered by a server;
 - Web Processing Service – WPS
 - Define basic request-response interactions for remote execution of a service
 - The only service able to store intermediate results at an external resource and use it as input data in a later service call;
 - Catalog Service for Web – CSW
 - standardizes interfaces to publish, discover, browse and query metadata about data, services and other resources;

Advantages of Grid technology

□ **Data Management**

- Handling massive data
- Catalogs and discovery services (find and access data)
- Event monitoring and management services

□ **Computational Power**

- Need for real time data processing
- Data, execution and workflow services (process data)
- High performance computing – HPC – and High throughput computing – HTC
- Significant gain in performance
- Workflow management

□ **Security**

- Certificates based access

Advantages of Grid technology

- Execute computational intensive calculations on very large amounts of data by using web services, on standard approach
- Provide a distribution of calculations and datasets on grid nodes with possibility of distributed high-speed data transfer
- The advantages introduced by the Grid infrastructure are visible only in certain cases:
 - the time required to execute a request is considerable larger than the overhead introduced by job creation and management on the Grid
 - several requests are made in parallel
 - a request can be split into several parallel sub-request

Advantages of Grid technology

- For simple requests, the overhead introduced by the Grid (job creation and management) is not compensated and the execution time is greater
- One approach : introduce the Grid only for the requests for which the overhead is compensated and the execution time is improved (**Mediator Component**)
- To differentiate the requests for which the Grid can bring an improvement from those for which it introduces additional overhead, some analysis have to be made regarding:
 - the type of requested service
 - the request parameters
 - the type of functionality executed inside the service

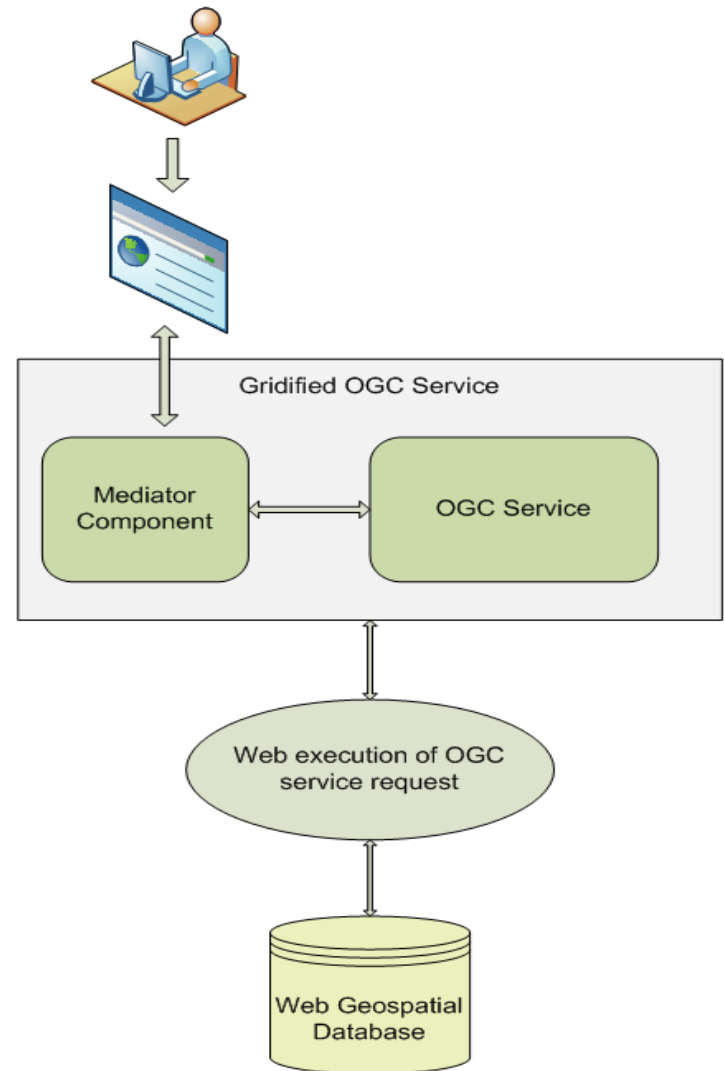
Advantages of Grid technology

- An estimation should be made regarding the boundary beyond which the advantages of the Grid are visible in the execution time – **Mediator Component**
- A modified OGC service will be adapted to support different execution flows depending on the decision made regarding the complexity of the request and the necessity of the Grid as execution environment – **Mediator Component**
 - execute the service directly
 - on Web databases
 - on Grid database
 - split the initial request into several jobs and send them to execution on the Grid
 - the workers connect to Web databases
 - the workers connect to Grid databases

Solutions and Approaches

- Case 1: The service uses Web Geospatial database.

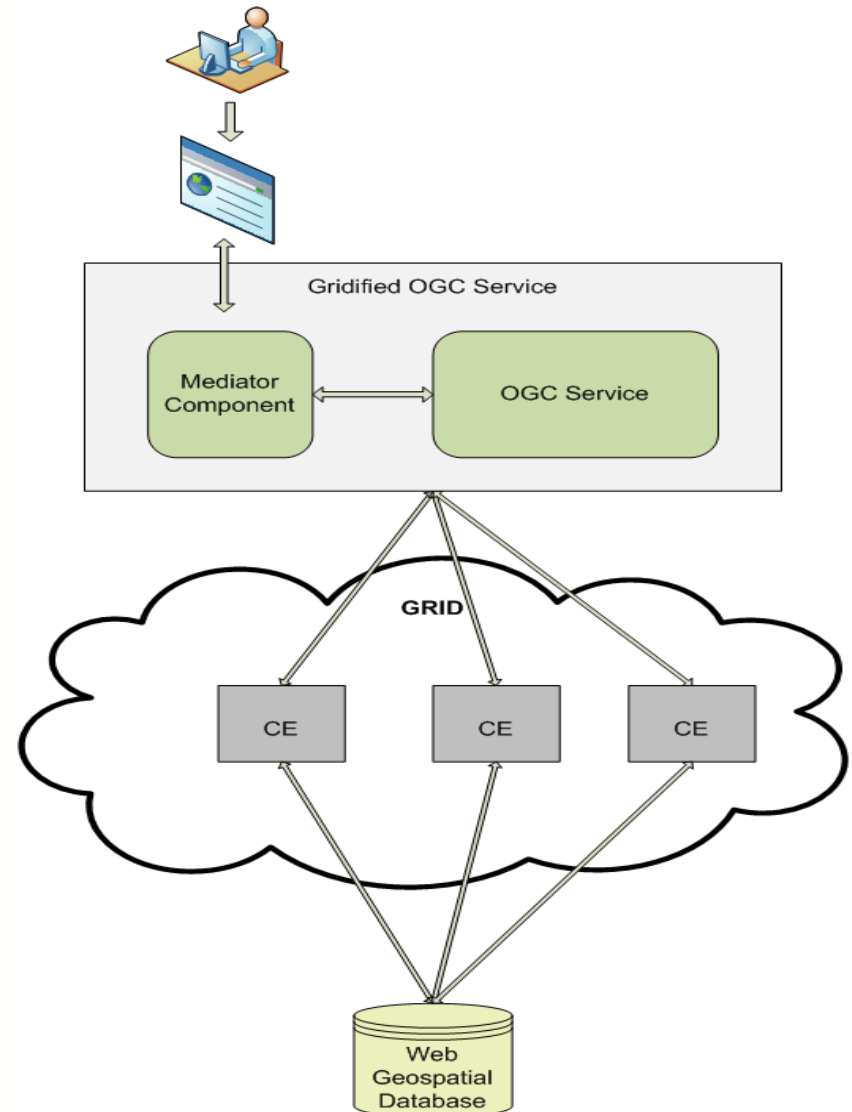
The simple requests (those for which the execution time does not exceed the overhead introduced by Grid) are executed directly on the Web server and the Grid environment is no longer used



Solutions and Approaches

- Case2: The service uses Web Geospatial database but it is using the Grid environment for the execution.

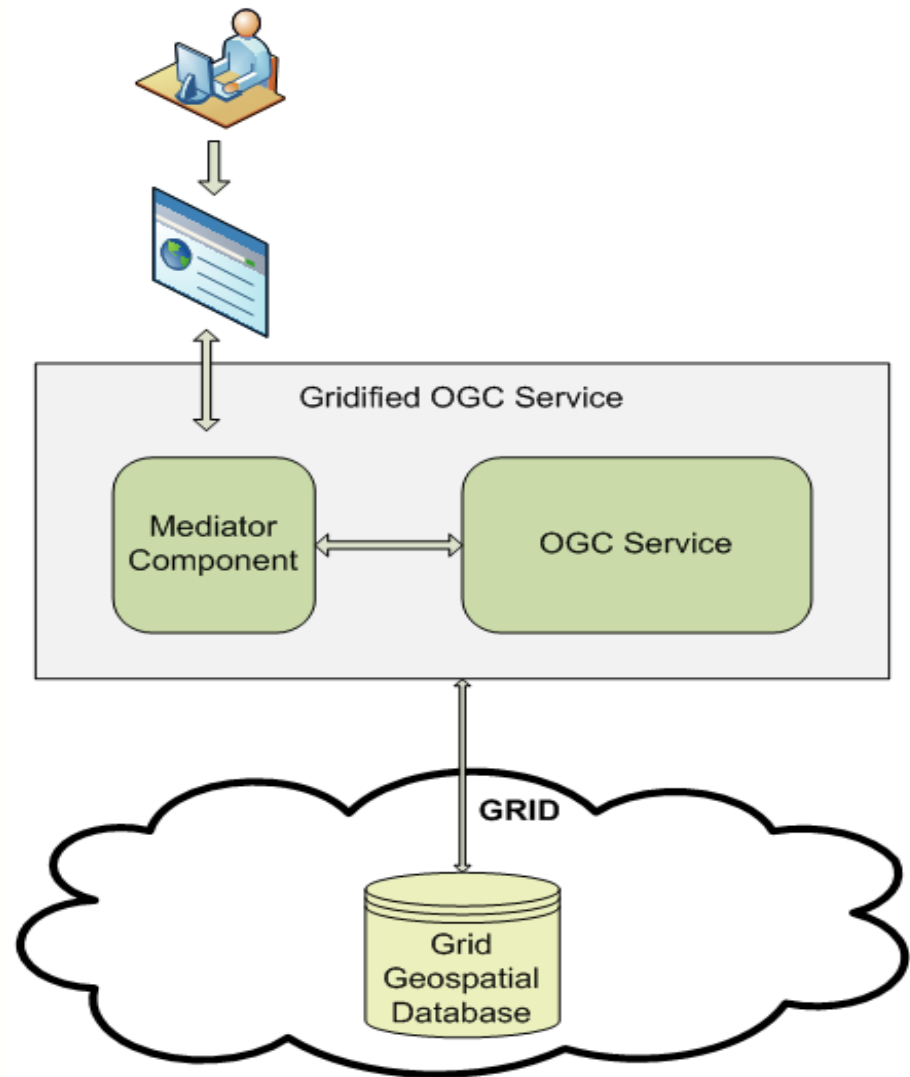
The request is split into several sub-requests which are executed on individual workers. The workers connect to the Web Geospatial database and retrieve the necessary data.



Solutions and Approaches

- Case 3: The service connects directly to the Grid database, using grid certificates and some special libraries.

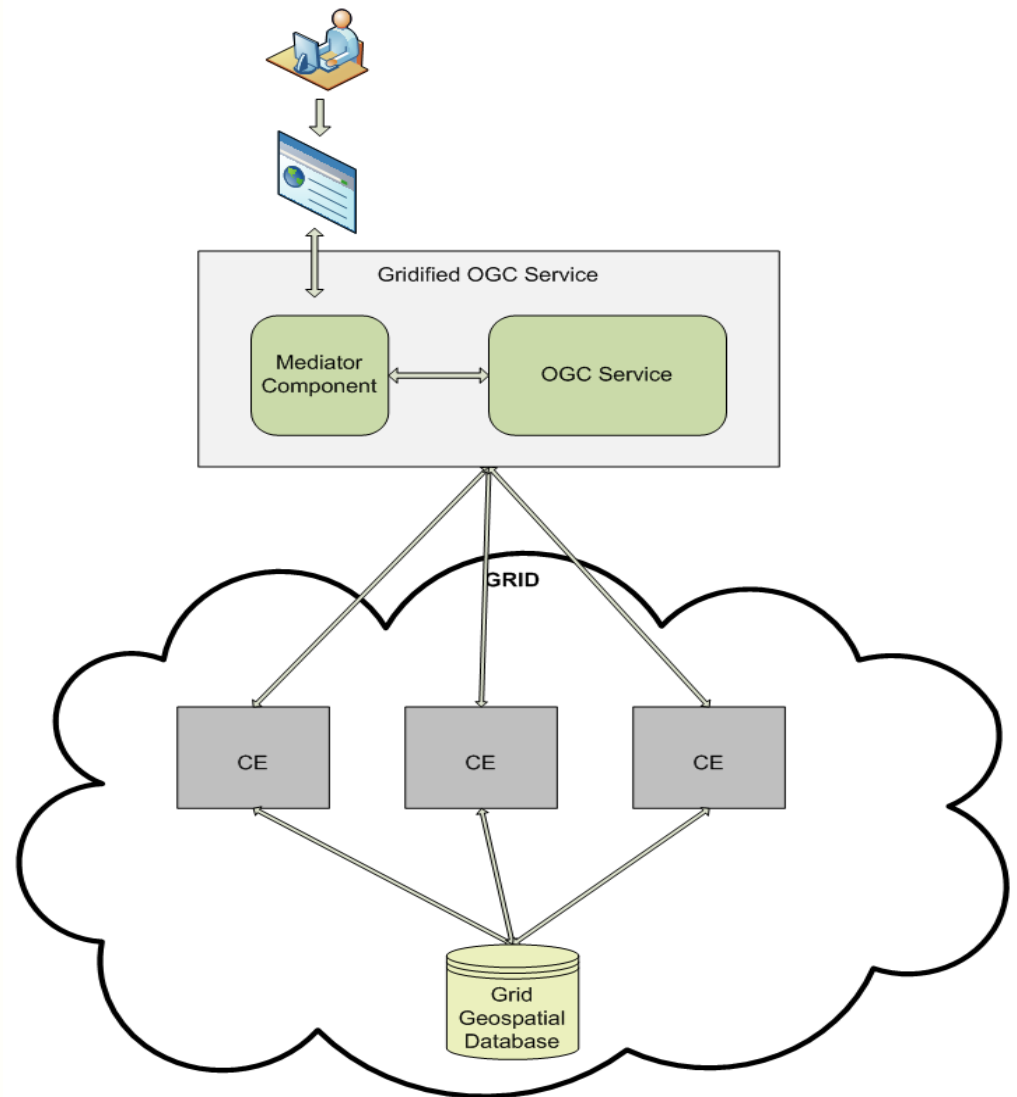
The connection is established using HTTPS and the data is copied using dedicated scripts.



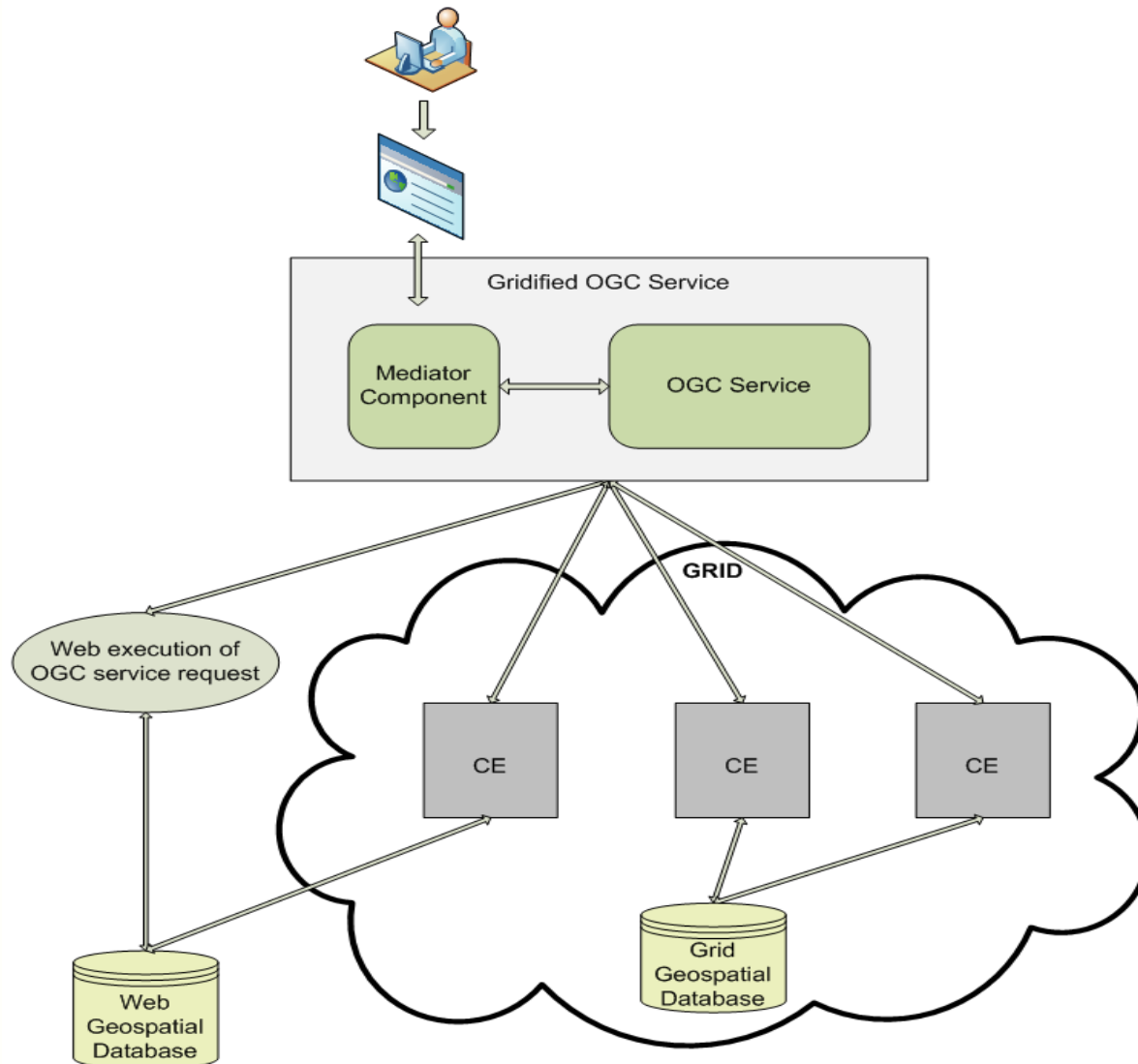
Solutions and Approaches

- Case 4: The service request is split into several jobs which will be executed on different workers (Computation Elements).

The workers are each responsible for connecting to the SEs (Storage Elements) and obtaining the necessary data.



Solutions and Approaches



Gridification of OGC services

- Gridification = casting of existing applications and services into the framework of a grid environment
- The gridification of OGC services must be done while:
 - Maintaining the functionality and the interface in the geospatial context
 - Take advantage of the Grid architecture in executing OGC workflow services
- The parallelism offered by the Grid technology can be applied at different levels:
 - Data parallelism
 - Computing parallelism

Partitioning of Spatial Data

- Three methods for data partitioning:
 - Object type
 - Processing of data can be parallelized by object types, e.g. buildings, transportation areas, water bodies, etc.
 - Computation nodes
 - Different operations can be executed on the whole dataset, distributed on different computation nodes
 - Data tiling
 - The whole dataset of a territory will be divided into tiles, which can be processed in parallel
- Computing parallelism:
 - Hard to achieve
 - Strongly related to the computation process

Experimental Tests

- The development of the gridified services started from the standard implementation of some of the most common OGC Web services developed inside the **deegree** project.
- The implementation of the gridified OGC services is included in the enviroGRIDS project.
- The tests were performed on WFS and WCS services using PostgreSQL database enabled with the PostGIS functionalities on small test data

Experimental Tests

- We have tested the first level of gridification for OGC services - the deployment of an OGC service instance to different CE nodes inside a Tomcat container.
 - After the submitted request has passed the complexity analysis, it is split into several sub-requests and for each sub-request a new job is created which deploys an instance of the respective service to a Tomcat container and submits the sub-request to be executed on the Grid
 - The results of each job are merged to obtain the final result which is passed to the user

Experimental Tests

- At the moment the complexity analysis takes into considerations only the number of queries a request contains
- A more optimal complexity analysis will take into account, aside from the number of queries, the amount of requested data and the type of operations performed on the requested data
- Based on the results of this analysis the request will be divided more accurately and the parallelization of the corresponding jobs will be done more efficiently

Interoperability achievements

- We have created the first version of the Mediator components
 - Analyze the complexity of service requests based on the number of queries present in the request;
 - Splits the complex requests into sub-requests;
 - Maps the sub-requests to Grid jobs;
 - Sends the created jobs into execution:
 - On each worker node, deploy an instance of the service in a copied container and execute the sub-request on the deployed service;
 - Collects the results from the jobs and create the final result
- We have started experimenting Grid services with Globus Toolkit
 - Idea: Use Globus Toolkit (components from Globus Toolkit) as Grid service management above the gLite middleware

Conclusions

- The integration of OGC Web services into the GRID environment is a complex process and has the following target points:
 - Data parallelism management;
 - Data security;
 - Complex execution parallelism;
- We have identified the main challenges and we proposed different gridifications levels of OGC Web services inside the enviroGRIDS platform at different architectural levels
- Some of the mention challenges have already been implemented and the others have been planned to be achieved

Thank you!

Questions?

Denisa Rodila, Dorian Gorgan
Computer Science Department
Technical University of Cluj-Napoca
<mailto:{denisa.rodila, dorian.gorgan}@cs.utcluj.ro>