Adapting ROMS to execute on GRID using a hybrid parallelization model

Carmen Cotelo Queijo, Andrés Gómez Tato and Ignacio López Cabido

Supercomputing Center of Galicia – CESGA

José Manuel Cotos Yañez

University of Santiago de Compostela

Spanish e-science

ADVCOMP 2008 September 29 - October 4, Valencia

RETELAB

A Virtual Laboratory for the National Network of Oceanographic Remote Sensing (RETEMAR).

RETELAB is a research project funded by:

Spanish Education and Science Ministry

(National Plan 2004-07: R&D Projects)

URL: http//www.retelab.org

FEDER FONDO EUROPEO DE DESENVOLVEMENTO REXIONAL

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN

CSIC CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS

XUNTA DE GALICIA

CESGA

RETELAB Partners

Keys Issues

Oceanographic research community has strong requirements for storage and satellite images processing and also for numeric simulation.

Data storage:

More data at a higher resolution every day.

Large amount of raw data.

Access to distributed data sets

Data processing:

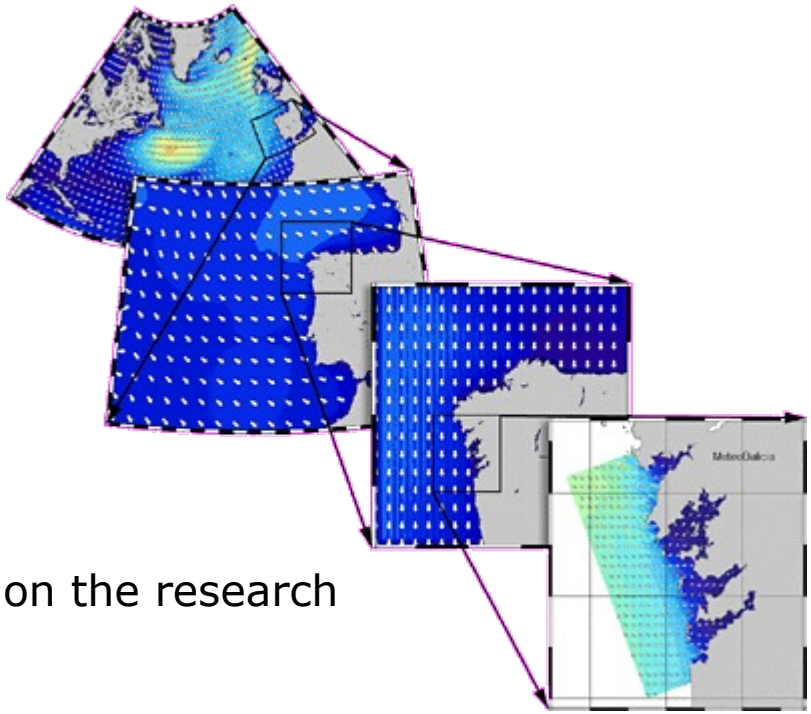
Data analysis is tedious.

Ocean study is an interdisciplinary task.

Heterogeneous data processing (it depends on the research group).

Simulation:

Currents and other oceanographic variables forecast.



Source : MeteoGalicia

Virtual laboratory

Grid technology enables oceanographic researchers to run algorithms that require high computational power and create a collaborative framework between centres, users, models and distributed data.

Main requirements:

Scientific applications used in ocean study.

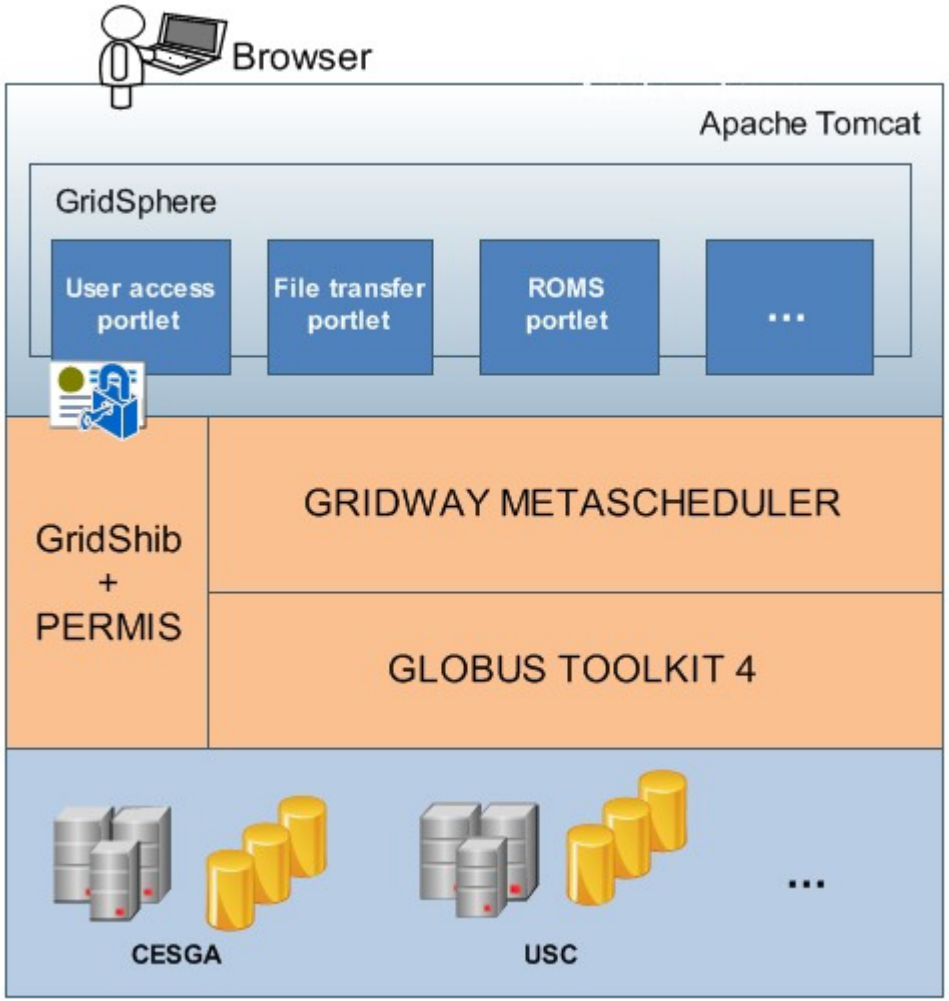
Management of security covering researchers access.

Distributed storage capacity for data exchange and management.

Processing capacity needed for data processing.

Visualisation of processed/stored data.

RETELAB Architecture



Web portal:

Developed using Java portlets (based on GridSphere Portal Framework)

Role Base Access Control using X.509 attribute certificates

GRID-enable workload management:

GridWay metascheduler

Allocates the best resources available and allows the execution of jobs.

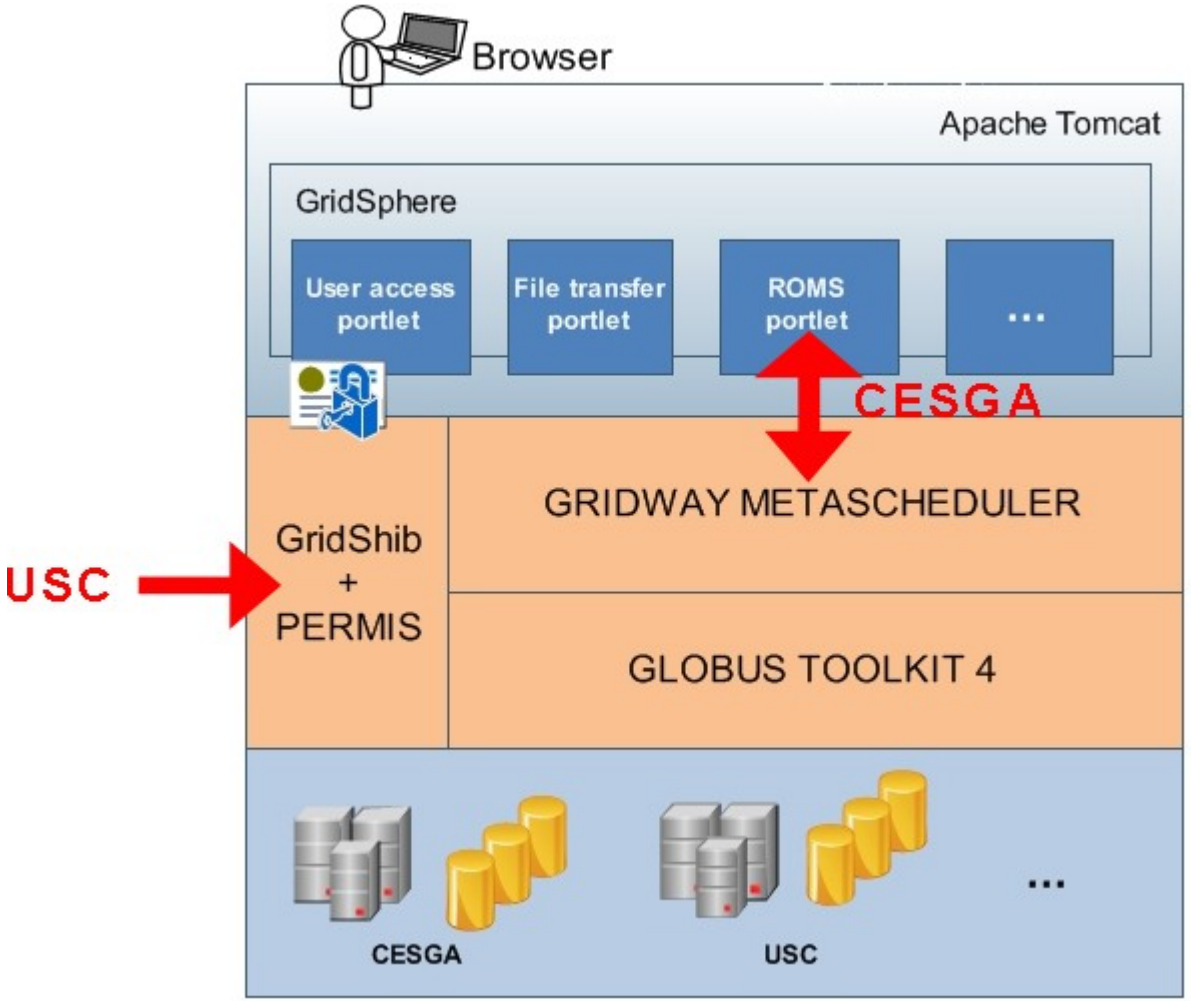
Supports Job Submission Description Language (JSDL) and DRMAA OGF standard

Grid middleware:

Globus Toolkit 4.

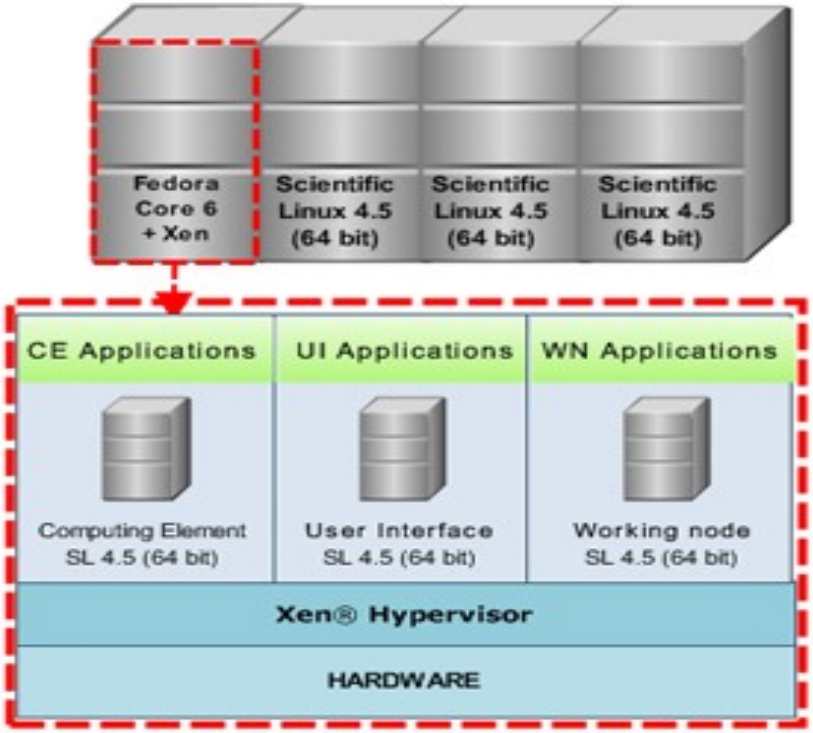
Enables the sharing of many different types of resources.

RETELAB Architecture



Computational infrastructure

Initial pilot is based on a cluster composed of 4 HP blade systems with Intel quad-core CPUs. One of the systems runs 3 Xen virtual machines.



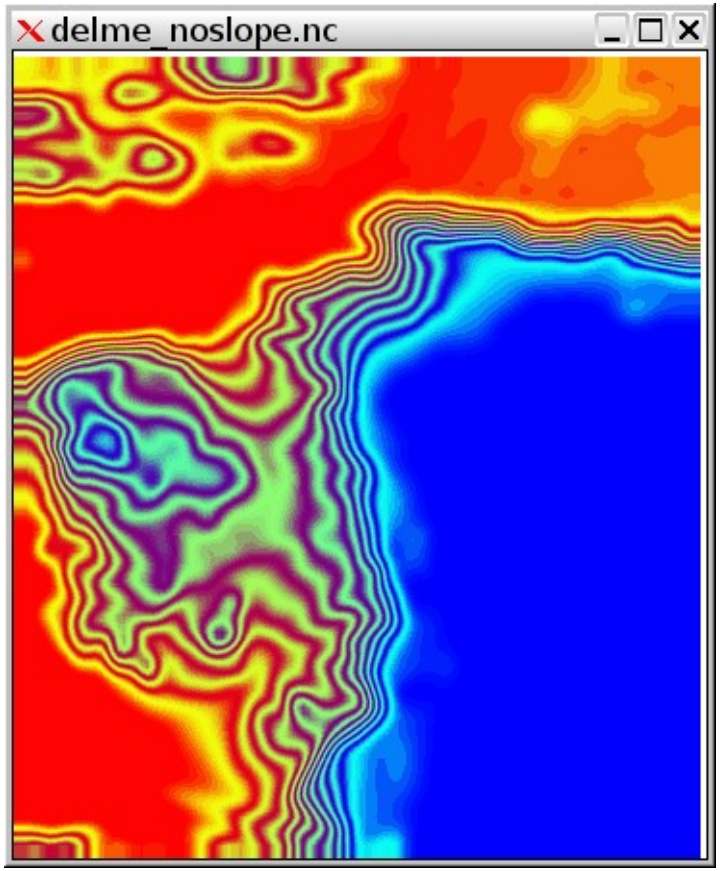
Demonstration application: ROMS

The Regional Ocean Modeling System (ROMS) is a a new-generation model for oceanic circulation:

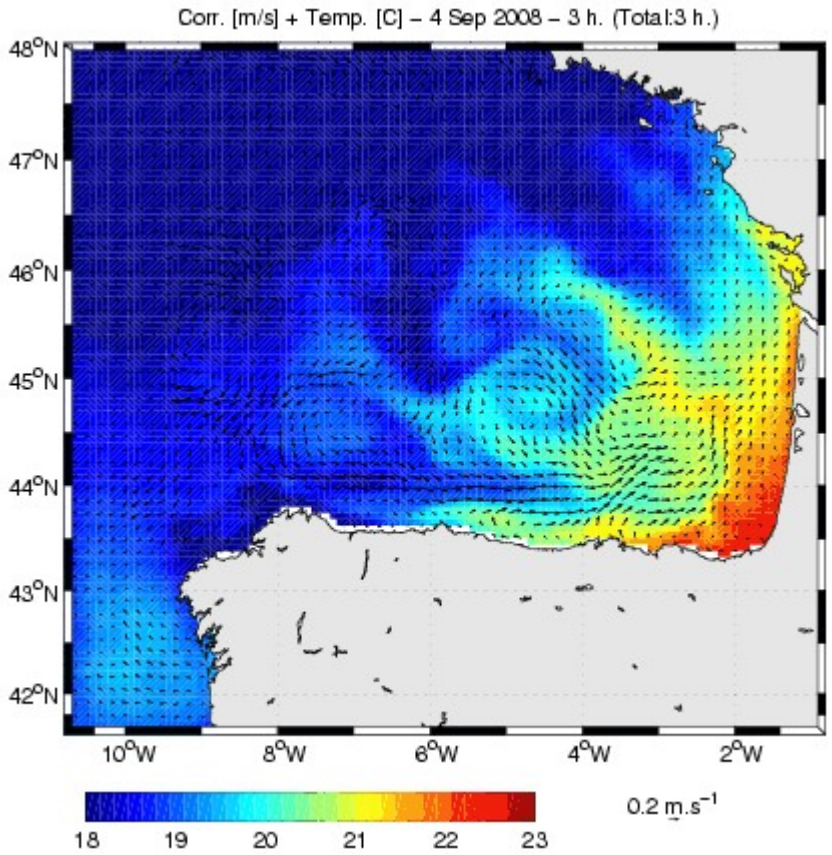
- Specially designed to simulate regional ocean systems accurately.
- Can be run in either serial or parallel computers using MPI and OpenMP programming models but not concurrently.
- Extensive pre and post-processing software for data preparation, analysis, plotting and visualization.

ROMS: http://www.myroms.org

Demonstration application: ROMS



Source: USC-TELSIG



Source: AZTI

Adapting ROMS

RETELAB architecture and the domain decomposition of ROMS MPI version will allow executing it in a hybrid model (OpenMP+MPI).

Tasks to execute ROMS using a hybrid model:

- Adapting ROMS code in order to use OpenMP and MPI simultaneously.
- Defining a new JSDL extension including multilevel parallelization features.
- Developing a specialized portlet able to translate job requirements into a JSDL document.
- Adapting Gridway components in order to use the new JSDL extension.

Adapting ROMS

Defining a new JSDL extension:

- JSDL is the Open Grid Forum standard.
- There are standard and non-standard JSDL extensions but none of them covers exactly this scenario.
- New extension includes multilevel parallelization features.
- This extension describes the aplication requiriments, not the technical details of the infrastructure.

JSDL-HTC example for NAS-MZ BT Benchmark

```
<jsdl:Application>
<jsdl:ApplicationName>NAS</jsdl:ApplicationName>
<jsdl:ApplicationVersion>3.0</jsdl:ApplicationVersion>
<jsdl-htc:HTCApplication>
   <jsdl-posix:Executable>BT.D.OpenMP</jsdl-posix:Executable>
   <jsdl-posix:Input>input.dat</jsdl-posix:Input>
   <jsdl-posix:Output>output.dat</jsdl-posix:Output>
   <jsdl-spmd:NumberOfProcesses>1</jsdl-spmd:NumberOfProcesses>
   <jsdl-htc:MinNumThreads>2</jsdl-htc:MinNumThreads>
   <jsdl-htc:MaxNumThreads>4</jsdl-htc:MaxNumThreads>
   <jsdl-htc:MemProcess>1000000</jsdl-htc:MemProcess>
   <jsdl-htc:DiskProcess>1000000</jsdl-htc:DiskProcess>
</jsdl-htc:HTCApplication>
<jsdl-htc:HTCApplication>
   <jsdl-posix:Executable>BT.D.4</jsdl-posix:Executable>
   <jsdl-posix:Input>input.dat</jsdl-posix:Input>
   <jsdl-posix:Output>output.dat</jsdl-posix:Output>
   <jsdl-spmd:NumberOfProcesses>4</jsdl-spmd:NumberOfProcesses>
   <jsdl-htc:MaxNumThreads>1</jsdl-htc:MaxNumThreads>
   <jsdl-htc:MemProcess>300000</jsdl-htc:MemProcess>
    <jsdl-htc:DiskProcess>300000</jsdl-htc:DiskProcess>
    <jsdl-spmd:SPMDVariation>
    http://www.ogf.org/jsdl/2007/02/jsd-spmd/MPI
   </jsdl-spmd:SPMDVariation>
</jsdl-htc:HTCApplication>
</jsdl:Application>
```

JSDL-HTC example for NAS-MZ BT Benchmark

Adapting ROMS

Developing a specialized portlet:

- The researcher defines the experiment details in an easy web interface.
- The portlet translates the problem into job requirements and then into a JSDL document.
- This JSDL document describes the job requirements for all the feasible ways to execute the problem (parallel and serial).
- The JSDL template is submitted to Gridway metascheduler.

Adapting ROMS

Adapting Gridway components:

- Support the new JSDL extension.
- Obtain detailed information about available resources.
- Select the best resources available and the best parallelization mode to start the asap job execution.

Questions?

THANKS!

Ignacio López Cabido
(nlopez@cesga.es)

FEDER
FONDO EUROPEO DE DESENVOLVEMENTO REXIONAL
GOBIERNO DE ESPAÑA
MINISTERIO DE CIENCIA E INNOVACIÓN
CSIC
CONSEJO SUPERIOR DE INVESTIGACIONES CIENTÍFICAS
XUNTA DE GALICIA
CESGA