Establishment of a Multiplexed Thredds Installation and a Ramadda Collaboration Environment for Community Access to Climate Change Data

Prof. Giovanni Aloisio
Professor of "Information Processing Systems"
Department of Innovation Engineering, University of Salento, Italy
and
Director, Scientific Computing & Operations Division
Euro Mediterranean Center for Climate Change (CMCC), Italy
office: +39 0832.297221/ fax: +39 0832.297235

Signature: ___________

Sandro Fiore, Ph.D.
Lecturer in "Advanced Data Management"
Department of Innovation Engineering, University of Salento, Italy
and
Scientist, Scientific Computing & Operations Division
Euro Mediterranean Center for Climate Change (CMCC), Italy
office: +39 0832.297371/ fax: +39 0832.297235

Signature: ___________

Osvaldo Marra
Technical Manager – HPC Laboratory
Department of Innovation Engineering, University of Salento, Italy
and
HPC Facilities Manager, Scientific Computing & Operations Division
Euro Mediterranean Center for Climate Change (CMCC), Italy
office: +39 0832.297371/ fax: +39 0832.297235

Signature: ___________

Department Chair: **Prof. Alfonso Maffezzoli**
Phone: +39 0832.297254
Fax: +39 0832.297525
Email: alfonso.maffezzoli@unisalento.it

Signature: ___________

1. Project Summary

Climate Change research is even more becoming a data intensive and oriented scientific activity. Petabytes of climate data, big collections of datasets are continuously produced, delivered, accessed, processed by scientists and researchers at multiple sites at an international level. In this regard the **Euro-Mediterranean Centre for Climate Change – CMCC** (Italy) aims at studying climate change issues both at a global and regional (Mediterranean area) scale, from several points of view: numerical models, information and communication technology, impact studies.
The University of Salento (UNISALENTO in Fig. 1 - Italy) is the ICT-partner of CMCC and in a synergic and strategic way, several courses at the University of Salento, strongly address topics like data management, data mining, parallel computing and distributed systems which are of interest for CMCC and represent the proper ICT foundations to work in the climate change area as a computational scientist. Moreover, several Bachelor, Master and Ph.D. thesis are related to ICT topics in this challenging geoscience field.

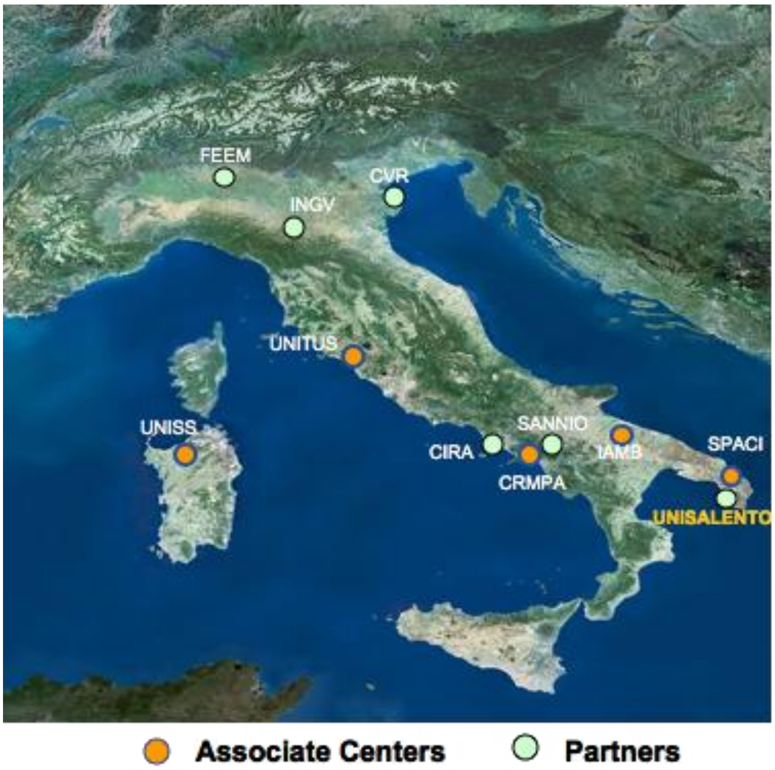


Fig 1. CMCC partners and associate centers

This proposal is intended to improve the link between the HPC laboratory at the University of Salento and the Euro-Mediterranean Centre for Climate Change by providing the proper funding to purchase a powerful IBM machine hosting both a multiplexed THREDDS installation and a RAMADDA collaboration environment for Climate Change. The infrastructure will be available to all of the students attending the "**Advanced Data Management**" course at the University of Salento to carry out data access, analysis, visualization and mining exercises on a comprehensive archive with atmospheric, oceanographic and climate change impacts data.

2. Project Description

Climate change represents an important and critical challenge for several scientists and researchers. Increasingly complex simulation models, management of petabytes of datasets (which are already too massive for current storage devices) are issues that must be faced up in the related centres. Key elements that must be taken into account are strongly connected data and metadata management.

In 2005, the Italian government, through the Ministry of the Environment and Protection (MATT), the Ministry of Education, University and Research (MIUR), and the Ministry of Economy and Finance (MEF) started a scientific initiative (namely the Euro-Mediterranean Centre for Climate Change, CMCC) aimed at establishing a national research centre devoted to climate change research. This Centre is distributed in nature among several sites at a geographical scale and it is comprised of several research divisions which provide support for computing and operations activities, numerical modeling, impact studies (on health, energy, economy, coastal zone, Mediterranean sea, agriculture, soil, etc.) training and dissemination. This Centre represents the most ambitious initiative undertaken in Italy, within the framework of the National Research Plan, and specifically the National Research Plan on Climate. One of the basic idea behind CMCC is to create a unified environment able to concentrate into the same place numerical models, simulations, big amount of data as well as metadata, post-processing, visualization and analysis tools, community services, etc. exploiting and joining knowledge and skills in the field of climate modeling, impact studies and information technology.

The SuperComputing facility consists of both scalar parallel (IBM Power6) and vector parallel (NEC SX9) machines providing about 30TeraFlops and a storage of about 1,5 PetaByte (1PetaByte for the tape library and approximately 500TByte of disk for online data).

In the last three years several datasets related to atmospheric, oceanographic and climate change impacts studies have been produced by the Centre. A selection of these heterogeneous datasets will represent our multifaceted contribution to the large Unidata Community. Moreover, the CMCC researchers and scientists will produce during 2011 CMIP5 data that could be of interest too for the purposes of this proposal. Downscaled data related to the Mediterranean area will be taken into consideration as well. NetCDF, GRIB, plain text (CSV, ASCII Grid Data), JPG, GIF are just some of the main formats that will be related to this archive.

UNIDATA THREDDS and RAMADDA represent two middleware tools bridging the gap between data providers and data users. While the goal of the first is to simplify the discovery and use of scientific data and to allow scientific publications and educational materials to reference scientific data, the second one represent a comprehensive content repository, publishing platform and collaboration environment for the Earth Sciences. THREDDS provides the infrastructure needed for publishing and accessing scientific data in a similarly convenient fashion and it is already adopted at CMCC (along with the Integrated Data Viewer) for scientific purposes. RAMADDA is in the CMCC plans and this proposal could represent a useful exercise to build the needed skills and competences about this tool.

A good point to fund this proposal is that **climate change data** related to the **Mediterranean area** could be available to the community for scientific and academic purposes. The students attending the ”**Advanced Data Management**” course at the University of Salento will also learn about the Common Data Model and the NetCDF Java library, the ISO19115 and the ISO19139 standards through a series of seminars titled “**eScience data management**”. The seminars topics will provide them the proper foundations to run data mining algorithms on climate change NetCDF data, to better understand the multidimensional nature of these datasets and to manage THREDDS and RAMADDA services. They will also create a student-contribution folder in RAMADDA where they will put the output (for instance images or software code) of their work. This will represent a concrete output of this experience and useful material for other classes and future courses.

At the end of the course, groups of 2-3 students will present their work during the final course examination. The course will start at the end of March and will last till the end of June. If this proposal will get the final approval by UNIDATA it is our intention to ask the UNIDATA team to take in the last weeks of the course, one Web-seminar about topics like THREDDS, RAMADDA and the NetCDF Java library to our class of students to give them a deeper insight (i.e. about security, internal architecture and APIs) related to these two tools and the Java library.

It is important to note that the **proposal** and the course on **Advanced Data Management** are strongly synchronized. The proposal should get the final notification by May 2011, which means at least one month before the end of the course. On the other hand, students are expected to do the course examination in the timeframe July 2011 – April 2012, that means one month before May 2012 - the expected end date of the project.

2.1. Motivations

A Virtual Machine based environment to host climate change datasets will represent, for the HPC Laboratory at the University of Salento, the proper infrastructure where students will be able to test and learn more about Unidata software like THREDDS, RAMADDA, IDV, etc. The proposal has strong **educational** and **research benefits**:
- laboratory exercises will rely on a real environment with data coming from climate change models running at CMCC;
- a **VM based environment** will help in creating and managing the proper environment for each groups of students and to transparently and easily change or adapt CPU, disk and memory settings.
- A **multiplexed configuration** for the THREDDS installation will be tested too against the standard configuration with a single server to evaluate performance and load balancing issues. The experience will be reported and will represent part of the documentation in output of this project.
- **data mining algorithms** will run on climate change datasets by exploiting the C and Java NetCDF libraries. These libraries will be part of the software installed on the VMs devoted to the laboratory exercises. Also in this case, the code will be published on specific folders of the RAMADDA community platform.
- the provided **data** will be really **heterogeneous**, thus contributing in a concrete way to the understanding of different aspects of climate change, in particular to those related to the impacts on economy, soil and coast.
- the **archive** that will be created for these purposes will be **available to other students and users**. The results will be stored in the RAMADDA service too and will contribute to create a knowledge base titled **“*Teaching experiences on climate change data management and learning by doing through the UNIDATA software*”**.

2.2. Serving the Unidata Community

Our installation will allow users to access to **climate change data** at the global and **Mediterranean** scale. Users will be able to join the infrastructure in different ways:
- accessing to the data via THREDDS;
- visualizing existing data via IDV;
- contributing to the RAMADDA content repository, publishing platform and collaboration environment by adding new interesting data.

2.3. Importance of this project

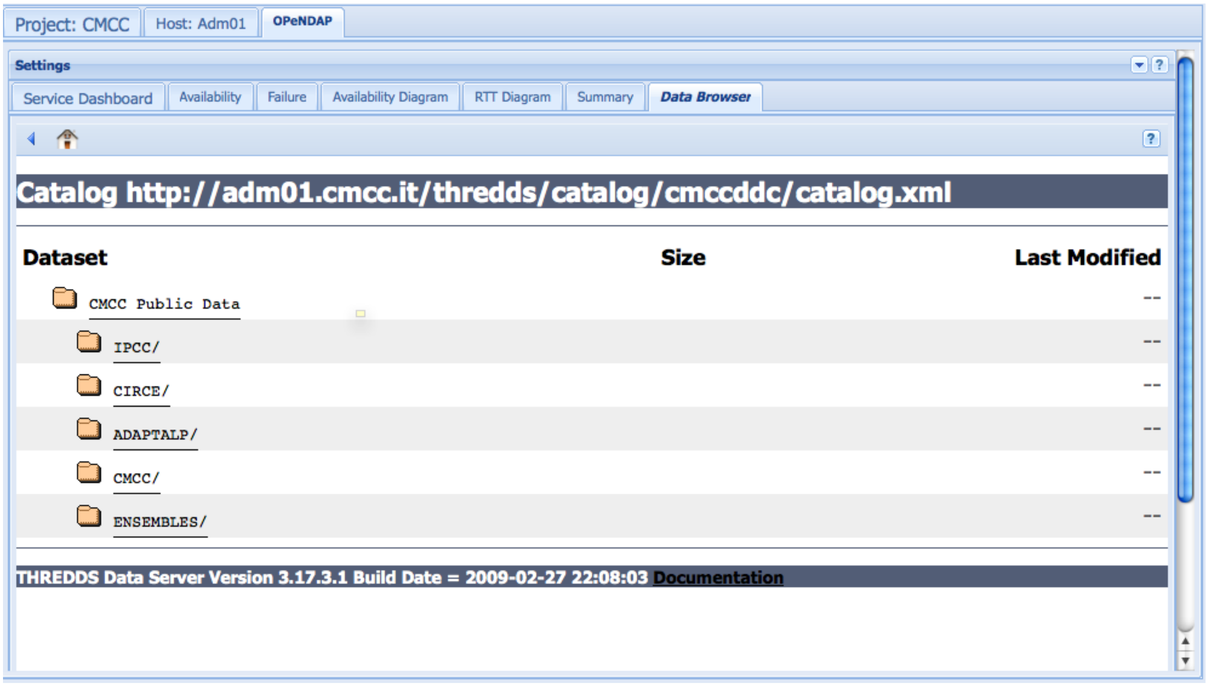
This proposal strongly addresses the Unidata mission and vision “*to provide the data services, tools, and cyberinfrastructure leadership that advance Earth system science, enhance educational opportunities, and broaden participation*”. In particular we focus on **climate change** datasets.
The educational aspects are also relevant and notable as pointed out in Section 2.1.

3. The Monitoring and Registry Facility

The VM-based environment foreseen in this proposal, will be equipped with a Dashboard system developed at CMCC and providing both **monitoring** and **registry** capabilities.

3.1. THREDDS Registry

The **registry** interface (see Fig. 2) allows to manage into the same web-page all of the (multiplexed) THREDDS installation deployed during the project. It will provide web browsing capabilities across the THREDDS catalogues.



Catalog http://adm01.cmcc.it/thredds/catalog/cmccddc/catalog.xml

Dataset	Size	Last Modified
CMCC Public Data		--
IPCC/		--
CIRCE/		--
ADAPTALP/		--
CMCC/		--
ENSEMBLES/		--

THREDDS Data Server Version 3.17.3.1 Build Date = 2009-02-27 22:08:03 Documentation

Fig 2. Registry service

3.2. System Monitoring

From a **monitoring** point of view, the Dashboard system will provide a “*service dashboard*” (see Fig. 3) with several charts and reports about the service availability, round trip time from a central CMCC location, other useful statistics and data summaries.

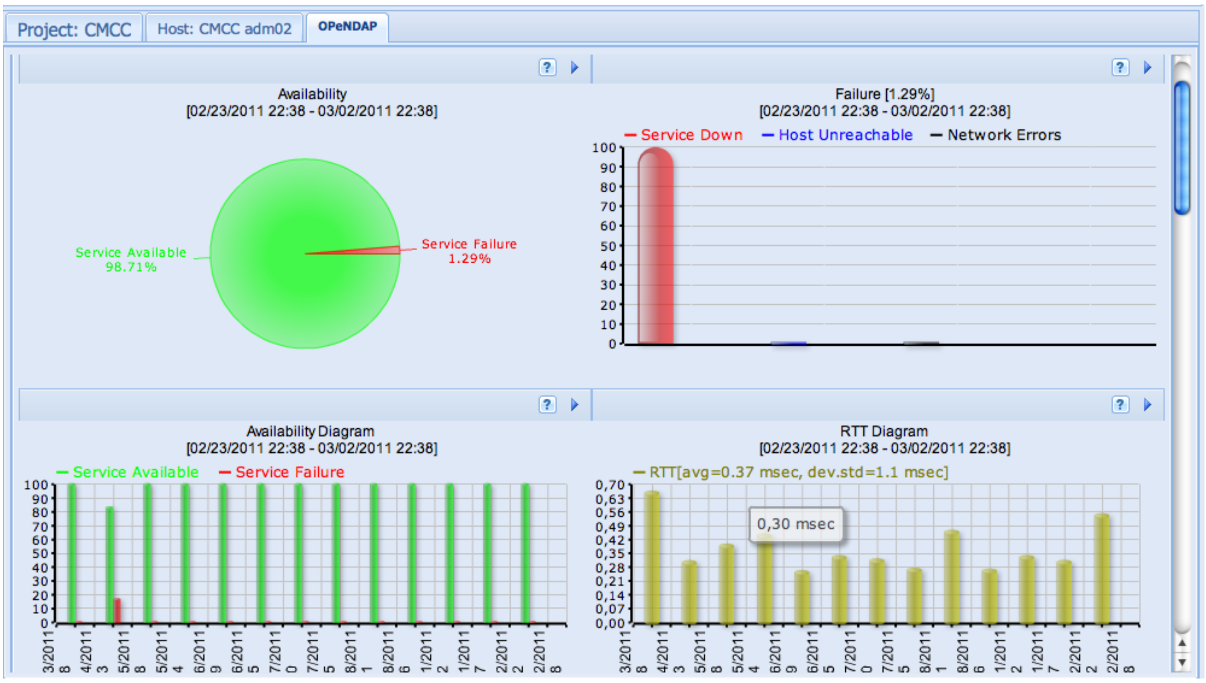


Fig 3. Monitoring system

4. Computing Facility Overview

The requested hardware (an IBM server equipped with 12TByte of storage, 2 CPU quadcore, 48Gbyte of RAM memory) would fit well into the computing facility of our University of Salento Campus. The computing room provides high speeed connectivity, cooling system and a direct link to the Italian high speed University GARR network. The budget foreseen in the "Budget" section should allow us to purchase the needed infrastructure satisfying the needed memory, CPU and storage disk requirements. At CMCC we have a lot of experience on Virtual Machine environments (ESXI software) as well as THREDDS software, so we should set up the environment plus the THREDDS service very quickly. The same cannot be stated for the RAMADDA platform, but we hope that this experience could help us in learning more about it from an end-user and administration point of views.

5. Budget

The University of Salento will not provide direct funds to purchase the proposed equipment. Anyway it will provide the needed support to successfully deploy the UNIDATA software. The University of Salento will contribute in terms of personnel for the installation and administration of the puchased systems. CMCC will also provide support in terms of personnel and know how about the UNIDATA systems and the provided data.

5.1. Hardware Justification

To host our climate change datasets we strongly believe that we need a Virtual Machine-based environment for a (multiplexed) THREDDS and RAMADDA installations with a large amount of RAM, fast disk storage and processing power.
We are strongly confident that our proposed system will work well for the proposal needs. Our request relies on similar infrastructural experiences at CMCC, leveraging on Virtual Machine based environments to host climate change data for internal CMCC scopes and purposes.

Server model	**IBM x3650 M3**
CPUs	2 x 4-core Xeon 3,6GHz X5687
RAM	12 x 4GB 1333MHz DIMM dual-rank (2 dimm for channel)
Hard-disks	12 x 1TB SAS 7,2K (12TB)
Optical drive	DVD-RW drive
Ethernets ports	1 dual-port integrated Gb ethernet card + 1 additional dual-port Gb ethernet optional daughter card
Power supply	2 Hot-Swap Power Supplies
Warranty	3 year onsite limited warranty
Price	$ 15.400,00 US Dollar

6. Project Milestones

Assuming that the project will be funded on June 1, 2011:

Date	Task
July 15, 2011	Purchase and delivery of the IBM server
August 10, 2011	Testing and deployment of IBM server completed including the VM environment plus the THREDDS server. Preliminary NetCDF datasets available for test and laboratory activities.
September 1, 2011	Installation and test of RAMADDA completed
October 31, 2011	New datasets available through THREDDS and RAMADDA
November 30, 2011	New datasets added to the catalogue
December 31, 2011	New datasets added to the catalogue
From August 1 to the end of the project	Data available to the students and to the UNIDATA community. Laboratory activities on the available datasets. Students reports added to RAMADDA platform.
May 31, 2011	Final report about the project activity