SEVENTH FRAMEWORK PROGRAMME
Research Infrastructure

FP7-INFRASTRUCTURES-2010-2 – INFRA-2010-1.2.3:
Virtual Research Communities

Combination of Collaborative Project and Coordination and Support Actions (CP- CSA)

LinkSCEEM-2
Linking Scientific Computing in Europe and the Eastern Mediterranean – Phase 2

Grant Agreement Number: RI-261600

D8.3
Cyberintegrator scientific workflow application available to computational scientists from the region

Final

Version: 1.1
Author(s): Terry McLaren, NCSA
George Tsouloupas, CaSToRC
Date: 01/03/2012

Project and Deliverable Information Sheet

LinkSCEEM Project	**Project Ref. №: RI-261600**	
	Project Title: LinkSCEEM-2	
	Project Web Site: http://linksceem.eu	
	Deliverable ID: <8.3>	
	Deliverable Nature: <Other>	
	Deliverable Level: RE *	**Contractual Date of Delivery:** 29/02/2012
		Actual Date of Delivery: 05/03/2012
	EC Project Officer: Leonardo Flores Anover	

* - The dissemination level are indicated as follows: **PU** – Public, **PP** – Restricted to other participants (including the Commission Services), **RE** – Restricted to a group specified by the consortium (including the Commission Services). **CO** – Confidential, only for members of the consortium (including the Commission Services).

Document Control Sheet

Document	**Title:** Cyberintegrator scientific workflow application available to computational scientists from the region	
	ID: 8.3	
	Version: 1.1	**Status:** Final
	Available at: http://www.eniac.cyi.ac	
	Software Tool: Microsoft Word 2010	
	File(s): LinkSCEEM-2-D8.3 Cyberintegrator_scientific workflow application available -final.docx	
Authorship	**Written by:**	Terry McLaren, NCSA George Tsouloupas, CaSToRC
	Contributors:	Mohamed Adel, BA
	Reviewed by:	Jens Wiegand, CaSToRC
	Approved by:	PMO

Document Status Sheet

Version	Date	Status	Comments
0.1	27/02/2012	First version	
0.2	28/02/2012	Final	Updates from Mohamed Adel, BA
1.0	02/03/2012	Final	Approved by PMO

Document Keywords

Keywords:	LinkSCEEM-2, Computational Science, HPC, e-Infrastructure, Eastern Mediterranean

Table of Contents

List of Figures

References and Applicable Documents

[1] http://www.linksceem.eu
[2] http://www.prace-project.eu
[3] http://www.fz-juelich.de/portal/DE/Home/home_node.html
[4] https://www.xsede.org/
[5] http://www.unicore.eu/
[6] http://medici.ncsa.illinois.edu
[7] http://medici-ch.cyi.ac.cy/

List of Acronyms and Abbreviations

ACF	Advanced Computing Facility
API	Application Programming Interface
CaSToRC	Computation-based Science and Technology Research Centre of the CyI
CPU	Central Processing Unit
CUDA	Compute Unified Device Architecture (NVIDIA)
CyI	The Cyprus Institute
CyNet	The Cyprus NREN
DEISA	Distributed European Infrastructure for Supercomputing Applications. EU project by leading national HPC centres.
EC	European Community
EoI	Expression of Interest
ESFRI	European Strategy Forum on Research Infrastructures; created roadmap for pan-European Research Infrastructure.
FP	Floating-Point
FPU	Floating-Point Unit
FZJ	Forschungszentrum Jülich (Germany)
GB	Giga (= 2^{30} ~ 10^9) Bytes (= 8 bits), also GByte
Gb/s	Giga (= 10^9) bits per second, also Gbit/s
GB/s	Giga (= 10^9) Bytes (= 8 bits) per second, also GByte/s
GÉANT	Collaboration between National Research and Education Networks to build a multi-gigabit pan-European network, managed by DANTE. GÉANT2 is the follow-up as of 2004.
GFlop/s	Giga (= 10^9) Floating point operations (usually in 64-bit, i.e. DP) per second, also GF/s
GHz	Giga (= 10^9) Hertz, frequency =10^9 periods or clock cycles per second
GigE	Gigabit Ethernet, also GbE
GNU	GNU's not Unix, a free OS
GPGPU	General Purpose GPU
GPU	Graphic Processing Unit
HDD	Hard Disk Drive
HE	High Efficiency
HET	High Performance Computing in Europe Taskforce. Taskforce by representatives from European HPC community to shape the European HPC Research Infrastructure. Produced the scientific case and valuable groundwork for the PRACE project.
HPC	High Performance Computing; Computing at a high performance level at any given time; often used synonym with Supercomputing
HPCC	HPC Challenge benchmark, http://icl.cs.utk.edu/hpcc/
HPL	High Performance LINPACK
HWA	HardWare accelerator
IB	InfiniBand
IBA	IB Architecture
IBM	Formerly known as International Business Machines
IEEE	Institute of Electrical and Electronic Engineers
I/O	Input/Output
ISC	International Supercomputing Conference; European equivalent to the US based SC0x conference. Held annually in Germany.
JSC	Jülich Supercomputing Centre (FZJ, Germany)
KB	Kilo (= 2^{10} ~10^3) Bytes (= 8 bits), also KByte
LQCD	Lattice QCD
LinkSCEEM	Linking Scientific Computing in Europe and the Eastern Mediterranean
LinkSCEEM-2	Linking Scientific Computing in Europe and the Eastern Mediterranean – Phase 2
LS	Local Store memory (in a Cell processor)
MB	Mega (= 2^{20} ~ 10^6) Bytes (= 8 bits), also MByte
MB/s	Mega (= 10^6) Bytes (= 8 bits) per second, also MByte/s
MFlop/s	Mega (= 10^6) Floating point operations (usually in 64-bit, i.e. DP) per second, also MF/s
MHz	Mega (= 10^6) Hertz, frequency =10^6 periods or clock cycles per second
MIPS	Originally Microprocessor without Interlocked Pipeline Stages; a RISC processor architecture developed by MIPS Technology
Mop/s	Mega (= 10^6) operations per second (usually integer or logic operations)
MoU	Memorandum of Understanding.

MPI	Message Passing Interface
MPP	Massively Parallel Processing (or Processor)
NDA	Non-Disclosure Agreement. Typically signed between vendors and customers working together on products prior to their general availability or announcement.
NoC	Network-on-a-Chip
NFS	Network File System
NIC	Network Interface Controller
OpenCL	Open Computing Language
OpenGL	Open Graphic Library
Open MP	Open Multi-Processing
OS	Operating System
pNFS	Parallel Network File System
POSIX	Portable OS Interface for Unix
PRACE	Partnership for Advanced Computing in Europe; Project Acronym
PRACE-1P	Partnership for Advanced Computing in Europe – First Implementation Phase
PRACE	Partnership for Advanced Computing in Europe – Research Infrastructure
RAM	Random Access Memory
SDK	Software Development Kit
SSD	Solid State Disk or Drive
TB	Tera (= 240 ~ 1012) Bytes (= 8 bits), also TByte
TCO	Total Cost of Ownership. Includes the costs (personnel, power, cooling, ...) in addition to the purchase cost of a system.
TFlop/s	Tera (= 1012) Floating-point operations (usually in 64-bit, i.e. DP) per second, also TF/s
Tier-0	Denotes the apex of a conceptual pyramid of HPC systems. In this context the PRACE Supercomputing Research Infrastructure would host the Tier-0 systems; national or topical HPC centres would constitute Tier-1
UNICORE	Uniform Interface to Computing Resources. Software for seamless access to distributed resources.
VO	Virtual Organization
VRC	Virtual Research Community

Executive Summary

Task 8.3 focuses on providing remote access to participating partner resources taking into account the limited bandwidth available in the region. The partners involved in this task will deploy the scientific workflow implemented in Task 8.2 to facilitate the execution of research tasks by the Eastern Mediterranean scientific community via a common interface/API locally and only retrieve the specific data subsets, visualizations and derived products that they need.

These workflows can be run as remote services and co-located with data aggregation services at a site that has better bandwidth or proximity to the data and computing resources. For example, the workflow could be hosted and deployed at the CaSToRC server. Whether this service is a VM cloud or HPC service will be determined based on the workflow's analytic complexity, region of interest, data size or other distinguishing factor. Access to the workflow will be provided based on the tickets awarded to users through the peer-reviewed process described in WP3. Additional resources, e.g. a new data analysis software application installed at a partner site, will be periodically added to the workflow based on regional user requests and available human resources for the preparation of the new software component. A workflow administrator will be responsible for the operation and maintenance of the workflow tool as well as for the testing and validation of new components.

Upon deployment of the workflow the administrator will report on quarterly basis to the Resource Allocation Committee monitoring access to project resources and be available for additional information on an as needed basis.

1 Introduction

Workflow use cases have been identified and range from local analytics to remote HPC access. The tools to facilitate the workflow needs of a science community also differ. For example, teams currently use simple and complex batch scripts (Unix shell, Perl, Python, etc...) to run their data through local filtering, translation and simple analytic processes. Other systems like the Medici data store use predefined workflows that are automatically executed on file upload to extract information from the file, generate derived data for preview purposes and can automate static workflows for ease of use purposes. Additionally, teams have use cases that require HPC resources to run complex analytics or to divide and conquer large data sets across multiple compute nodes. In this case, UNICORE (Uniform Interface to Computing Resources) has been chosen as the common tool to access HPC resources. UNICORE provides workflow definition and job management system and is in use by many of the major HPC centers such as the Juelich Supercomputing Centre[3] and the XSEDE[4] HPC collaboration based in the USA. It should be noted that the use of Cyberintegrator, as originally intended, was not possible within LinkSCEEM-2 due to technical issues. For this reason, the title of the deliverable was modified by removing the reference to Cyberintegrator.

The training for both UNICORE and Medici occurred during the 2012 Winter School that was held at the Cyprus Institute in Feb 2012. UNICORE training was performed on Friday Feb 10th while Medici training occurred on Saturday Feb 11th followed by Medici developer training on Monday Feb 13th and Tuesday Feb 14th. This provided the baseline for integrating and using the software packages across all HPC sites, which is a significant step towards integrated resources as key task within LinkSCEEM-2.

2 Workflow Approaches

2.1 Scripted Workflows

Each science team has existing workflow scripts and it is fully recognized they will persist for their local purposes. We fully expect the teams to continue using and building local workflows using a scripting approach. At some point the team may want to automate these process with data upload or need to redefine them for scalability purposes and migrate the workflow to one of the other available enviornments.

2.2 Medici Workflows

Medici[6] provides an option by which a user can create, or customize existing workflows, to create automated workflows known to Medici as extraction services. Extending existing or creating new extraction services was discussed at the Medici developer training sessions held during the 2012 Winter School. This process is documented on the Medici wiki located at: https://opensource.ncsa.illinois.edu/confluence/display/MMDB/Creating+New+Extractors.

During the training event, extending the automated workflows were discussed with the Digital Cultural Heritage (DCH) and Climate teams and identified the following needs.

1. DCH: the need to scale down the polygon count for large objects for preview purposes.
    a. Add a new step in the workflow to reduce the number of polygons before a preview object is created.
2. DCH: generage high resolution images (tif, bitmaps) from video and 3D formats (ply, obj, stl, x3dom).
    a. Note: The ability to generate high resolution images from video already exists.
    b. Extend the current extraction service to include 3D fromats.
3. Climate: ability to display the contents of a NetCDF file.
    a. Create new NetCDF extraction service and identify a previewer.
4. Integration with existing workflows. Allow generic scripts to handle data staging (in and out of Medici via an API)
    a. Extend current web service API to facilitate scripted queries.

Since activities were recently discussed, they are candidate activities for future work to extend the existing workflow capabilties of the data repository. These capabilties are NOT part of the current environment and will be considered for future development activities.

2.3 UNICORE (Uniform Interface to Computing Resources) Workflows

UNICORE [5] provides a standard mechanism for job submission, monitoring and defining workflows for HPC. It provides a Grid enabled environment with multiple clients (application and command line) and a server interfaces to create and access remote computing environments. UNICORE was developed at Juelich and was recently adopted by XSEDE, the successor to Teragrid. Leveraging these fully supported, mature environments will not only provide the LinkSCEEM community a state of the art environment, it will also provide the community members a common environment that's used by other research communities and provides a common gateway to other EU and USA research projects. UNICORE is an active project in the European Middleware Initiative and is engaging its user community with new software capabilities, updates, training and along with manuals and client video tutorials. UNICORE also provides an API that interested communities can build their own custom tools.

3 Workflow Deployment Status

3.1 Medici

Medici has been deployed at Cyprus Institute [7] and discussions about updating the workflows for the DCH team are currently work in progress.

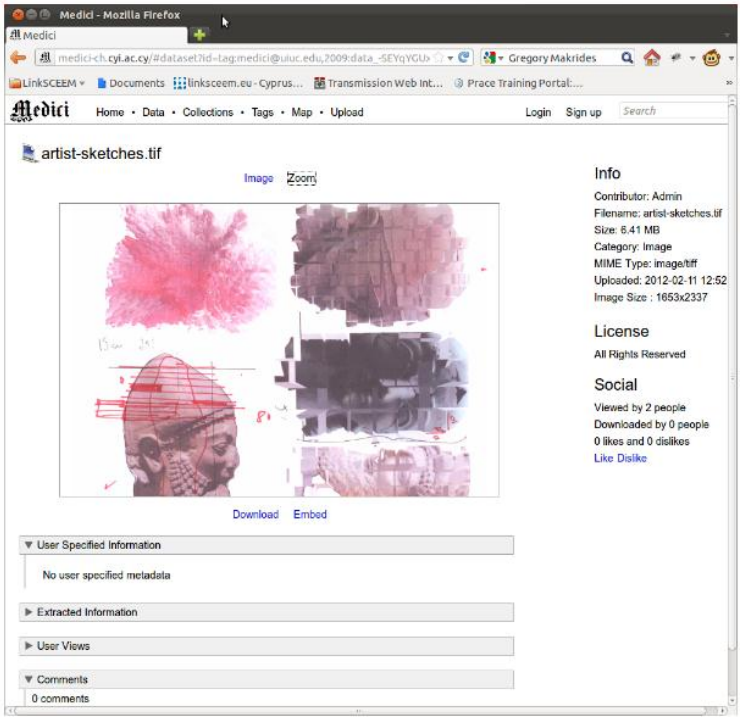

Figure 1: Medici hosting DCH Data

3.2 Unicore Deployment Status

The main UNICORE components have been installed at the three planned sites, CaSToRC, BA and NARSS. The overall plan that was followed is shown in Figure 2.

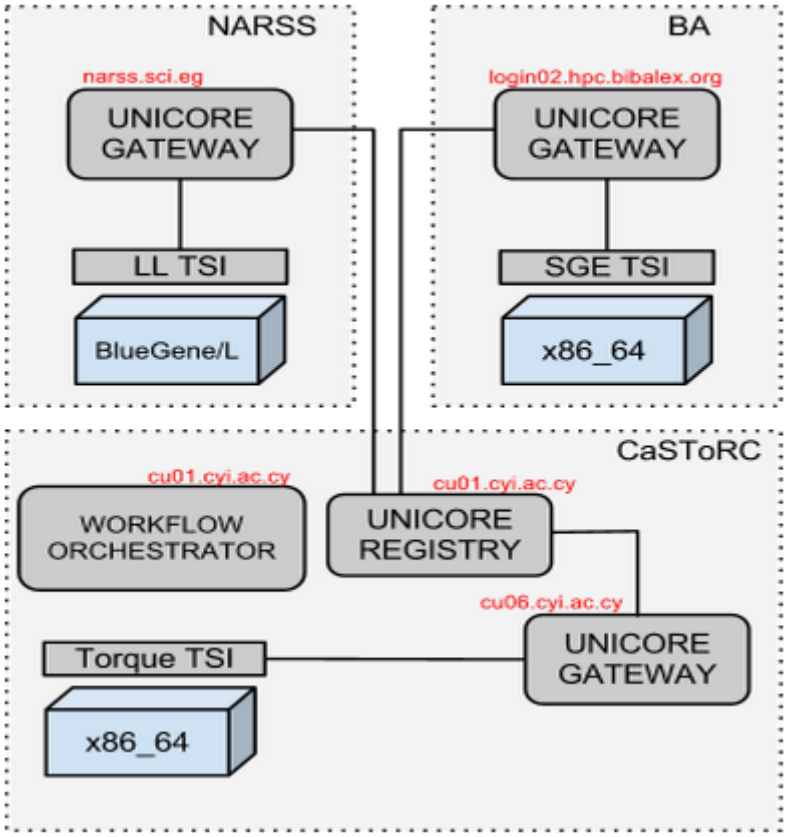

Figure 2: Unicore Deployment Overview

In particular, we installed the following components:

- Gateway, UNICOREx and the TSI at BA;
- Gateway, UNICOREx and the TSI at NARSS;
- Gateway, UNICOREx and the TSI at CaSToRC;
- Registry at CaSToRC;
- Orchestrator at CaSToRC;

UNICORE User Interfaces

UNICORE, in addition to the command-line client (ucc), offers a GUI client based on the Eclipse Rich Client Platform. Figure 3 shows a screenshot of the client connected to the LinkSCEEM Infrastructure.

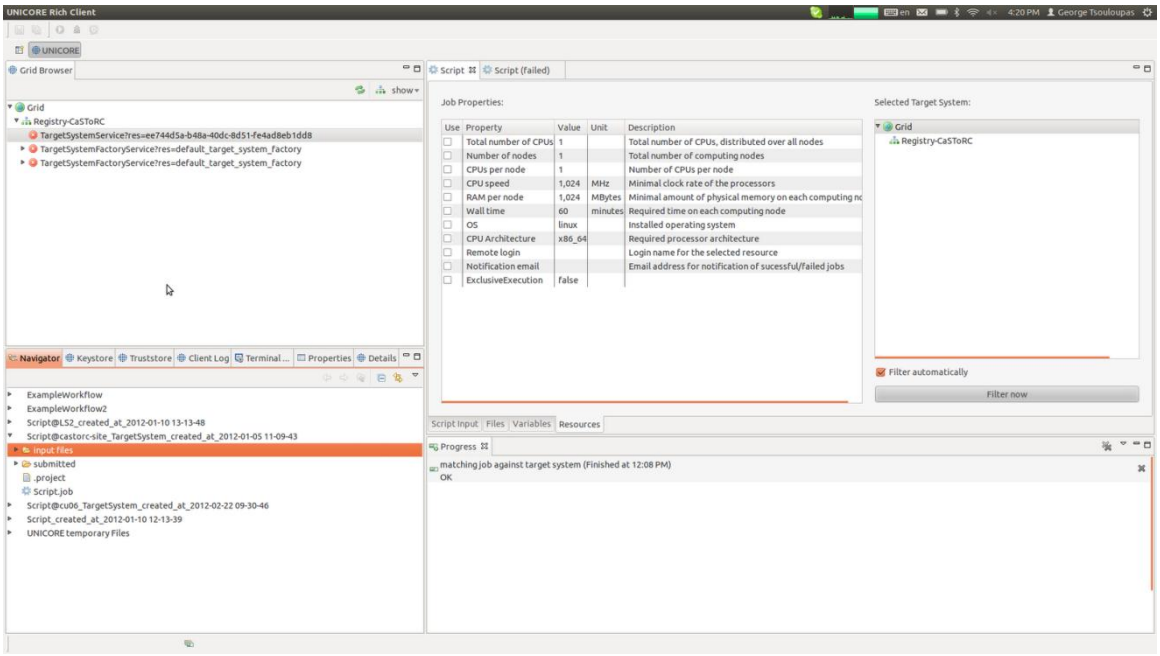

Figure 3: The UNICORE Rich Client Application

Certificates

Due to the advanced security capabilities and requirements of the PKI-based authentication infrastructure of UNICORE, each host that runs UNICORE needs a Certification Authority (CA) signed certificate. At this point testing has been performed using test certificates. All participating sites are in the process of obtaining host certificates signed by their country's CA. The system will be considered in production immediately after the test certificates are removed.

4 Conclusion

LinkSCEEm-II workflow approaches have been defined and initial implementations have been deployed. Over time additional scientific communities need to be engaged to understand their needs and unique requirements and then the appropriate enviornment can be deployed to meet their needs.

Next steps

- **Integration with actual HPC queue management systems.** Once all the UNICORE components have undergone final testing (namely WORKFLOW ORCHESTRATOR, the SGE and LoadLeveller Target System Interfaces) and all machine certificates are obtained from valid Certification Authorities, then we can proceed with the final connection to the actual HPC machines rather than test-systems.
- **Working with applications** (DCH, Climate) to build application-specific workflows.
- **Integration with Medici.** (query for relevant files to stage, upload results such as 3D models)
- **Deployment of UNICORE at SESAME to utilize UFTPd**
- While not initially planned, it will be beneficial to install UNICORE at SESAME so that we can take advantage of UFTPd to transfer large data files (in addition to GlobusOnline)

- **Common user database (xuudb)** During the discussions on the architecture and roll-out strategy it was also decided that user databases (i.e. the mechanism by which UNICORE authorizes users and maps them to local system users on the HPC systems) will be replicated across the three sites (while sites will still maintain control of exaclty what is replicated). Users that are granted access to specific systems through preparatory and production calls (WP3) will be replicated accordingly. This will not replace the creation of system accounts that still remains the responsibility of the local system administrators.