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

D9.1
Year 1 Report on Cross-disciplinary Research Activities

Final

Version: 1.0
Author(s): Alan O'Cais, JSC
Date: 17/06/2011

Project and Deliverable Information Sheet

LinkSCEEM Project	**Project Ref. №: RI-261600**	
	Project Title: LinkSCEEM-2	
	Project Web Site: http://linksceem.eu	
	Deliverable ID: < **D9.1**>	
	Deliverable Nature: Report	
	Deliverable Level: PP	**Contractual Date of Delivery:** 31 / August / 2011
		Actual Date of Delivery: DD / Month / YYYY
	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:** <**LinkSCEEM-2** >	
	ID: <**D9.1**>	
	Version: <**1.** >	**Status:** Draft
	Available at: http://www.eniac.cyi.ac.cy	
	Software Tool: Microsoft Word 2007	
	File(s): LinkSCEEM-2-Deliverable-9.1	
Authorship	**Written by:**	Alan O'Cais, JSC
	Contributors:	NCSA, CyI
	Reviewed by:	P.Fitzhenry, CyI
	Approved by:	PMO/SC

Document Status Sheet

Version	Date	Status	Comments
0.1	29/July/2011	Draft	
1.0	31/August/2011	Final version	

Document Keywords

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

Table of Contents

PROJECT AND DELIVERABLE INFORMATION SHEET II
DOCUMENT CONTROL SHEET II
DOCUMENT STATUS SHEET II
DOCUMENT KEYWORDS II
TABLE OF CONTENTS III
LIST OF ACRONYMS AND ABBREVIATIONS III
EXECUTIVE SUMMARY 1

1 INTRODUCTION 2
2 PERFORMANCE ANALYSIS 2
2.1 DEVELOPMENT OF A REGIONAL CROSS-DISCIPLINARY GROUP ON "PERFORMANCE ANALYSIS" 2
2.2 COOPERATION AND COORDINATION OF ACTIVITIES BETWEEN THE GROUP AT JSC AND THE REGIONAL GROUP 2
2.3 SELECTION OF ALGORITHMS FROM THE CLIMATE RESEARCH FIELDS 3
2.4 SCALASCA TRAINING EVENT 3
2.5 PORTING TO SUPERCOMPUTERS 3
2.6 APPLICATION PERFORMANCE ANALYSIS WORK PLAN 4
2.7 HANDS-ON EVENT AT NARSS 5
2.8 PERFORMANCE ANALYSIS OF THE ALGORITHMS 5
3 MATHEMATICAL ANALYSIS AND ALGORITHMS 6
3.1 DEVELOPMENT OF A REGIONAL CROSS-DISCIPLINARY GROUP ON "MATHEMATICAL METHODS AND ALGORITHMS" 6
3.2 ALGORITHMS ANALYSED WITH RESPECT TO ALGORITHMICAL OPTIMISATION 6
4 DATA MANAGEMENT 7
4.1 DEFINE NETWORK LANDSCAPE 7
4.2 RESEARCH AND DESCRIBE ALTERNATIVE ACCESS OPTIONS 8
4.3 TEST OPTIMIZATION OPTIONS 9
4.4 RECOMMEND OPTIMIZATION OPTIONS 10
5 VISUALIZATION 11

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
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
HPL	High Performance LINPACK
HWA	HardWare accelerator
IB	InfiniBand
IBM	Formerly known as International Business Machines
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
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.
NFS	Network File System
NIC	Network Interface Controller
OpenCL	Open Computing Language
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
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

The aims of Work Package 9 are threefold:

- To make the optimization of parallel applications on large-scale applications both more effective and more efficient,
- To adapt and develop tools for the management of data and data repositories in Cultural Heritage, Climate and Synchrotron research,
- To enable the manipulation and visualisation of these complex data sets.

The first year activities of the WP have focussed around creating the groups, knowledge and infrastructure to support these goals.

The established expertise of FZJ-JSC is being used to help establish regional teams in

- Performance analysis,
- Mathematical analysis and algorithms.

Knowledge is being assimilated into the region through training events and the adoption of preliminary protocol for performance analysis. Climate applications have been chosen as a suitable prototype case for this type of work.

Data management is due to become a significant millstone for the project due to the level of connectivity in the region and the projected data creation within the scope of the project. In this first year, we establish the real restrictions that this creates and outline some concepts to limit the impact of this.

Within this WP is intended that NCSA and CaSToRC collaborate on installation of hardware and software for use initially with cultural heritage applications. The main objective is to provide a tele-collaborative environment between at least two sites where the users can exchange and interact with flight viewpoints of the datasets located at each site. To this end, the neceassary equipment and software for such work has been installed and tested and a training program between the two partners has been agreed upon.

1 Introduction

Within the context of Work Package 9 there are three main goals:

- To make the optimization of parallel applications on large-scale applications both more effective and more efficient,
- To adapt and develop tools for the management of data and data repositories in Cultural Heritage, Climate and Synchrotron research,
- To enable the manipulation and visualisation of these complex data sets.

These goals have been broken down into four tasks within the Work Package. The tasks "Performance Analysis" and "Mathematical Analysis and Algorithms" are derived from the first goal, "Data Management" relates to the second goal and "Visualization" to the third goal. In this document, we describe what has been achieved in the WP on a task-by-task basis and provide an outlook on what is hoped to be achieved in the coming year.

2 Performance Analysis

Within this Task it is intended that CaSToRC, BA and NARSS will collaborate with the existing Cross-Sectional Team at JSC in Performance Analysis to develop a regional Cross-Sectional Team in this area.
Within the first year the goal has been to identify the team members at each site and establish communication between them. Ultimately the Work Package hopes to achieve such goals as:

- Research and development related to parallel application performance,
- Performance consulting services for scientific communities and individual users.

The majority of work of the WP in the first year has gone into establishing the team and communicating knowledge and experience from JSC to that group so that they can ultimately function independently of JSC.
In the following subheadings we describe the steps that have been taken to achieve this goal. In particular, climate research has been identified as a beneficial pilot project for the teams to form a concrete basis with which to achieve these goals.

2.1 Development of a regional cross-disciplinary group on “Performance Analysis”

Team members from the sites have been identified with first efforts towards developing the skill set of the group done through the Scalasca training event as described below.

It has been recommended that each site install the UNITE software package that includes many free performance analysis utilities. Having a common 'toolbox' will facilitate the sharing of expertise between each site.

2.2 Cooperation and coordination of activities between the group at JSC and the regional group

Alan O'Cais at JSC will form the first point-of-contact for the regional groups to the JSC cross disciplinary teams. Inge Gutheil will be the next level of contact for the Mathematical Methods and Algorithms group.

Furthermore, Dr. Lars Hoffmann, who is leading the Climate Science Simulation Laboratory at JSC, will collaborate with teams working on climate related applications.

Prof. Salwa Nassar will form the first point-of-contact to the team at NARSS and Prof. Magdy Nagi will be the counterpart at BA with Patrick Fitzhenry the counterpart at CyI.

The earth system modeling application ECHAM5/MESSy was identified as a useful pilot application for WP9 given the expertise available through WP10 leader MPI Mainz and it's prolific utilisation in Cyprus. A meeting was held on the 26th of November with Hendrik Merx of MPI Mainz who outlined the core calculation steps of the code and highlighted the main computational areas and algorithmical models within the code. Lars Hoffmann also attended this meeting.

ECHAM5/MESSy is known to have high (per core) memory requirements that make it unsuitable for the Blue Gene architecture at NARSS. For this HPC system, NARSS will investigate the Weather Research and Forecasting (WRF) Model which is being actively developed in the US and known to scale well on Blue Gene systems.

2.3 Selection of algorithms from the Climate research fields

To benefit from deliberate overlap in activities, JSC and CaSToRC will both work with ECHAM5/MESSy while BA and NARSS will work with WRF. Furthermore, both these models have a common computational kernel that will facilitate active collaboration between all groups in a common area. This kernel is a software library package called the Kinetic PreProcessor (KPP), which is a software environment for solving chemical kinetics.

2.4 Scalasca training event

Coordinating with WP4, JSC sent Marcus Geimer to lecture on Performance Analysis tools at the LinkSCEEM-2 Winter School. While there, he provided a one day “train the trainers” tutorial day for the technical teams of the 3 sites on Performance Analysis tools. The content of this tutorial formed a basic introduction to the tools that are intended to be used within the WP.

2.5 Porting to supercomputers

The licencing for MESSy also requires a licence for ECHAM and licencing for Juropa was acquired by JSC. Data for testing purposes was provided by Hendrik Merx. This was for a particular T42 model which is known not to scale well beyond 256 cores. The installation scripts for MESSy are quite complicated and have been expanded to include options for many systems. This has bloated the scripts somewhat and makes their modification for new environments somewhat tedious. Modification of the installation scripts for MESSy were required to allow the application to utilise the optimised libraries for BLAS and Lapack calls. A number of compiler optimisations were explored, however the developers employ strict optimisation control that requires bitwise similarity between results from optimised code and the reference model. This excludes aggressive compiler optimisation in a production environment but we allowed it here for this exploratory work. The execution script also required adaptation to the Juropa environment.

EMAC has been ported to Planck and Euclid clusters.. The code currently runs on Planck with the Intel/mvapich2 combination of compilers and MPI. It will only run on a single node on Planck with the Intel/openmpi combination on Planck. Attempts to run on more than a single node on Planck result in a segmentation fault error during the MPI_Bcast() call at the

beginning of the integration. DDT has recently been purchased and installed on Planck to look at this issue. All combinations compiled on Euclid fail with this same segmentation fault.

2.6 Application Performance Analysis Work Plan

It has been decided to focus on three particular performance analysis tools at this early stage of the WP: HPCToolkit, Marmot and Scalasca. These applications provide high levels of functionality for the performance analysis of applications including MPI verification, tracing and profiling. Further applications with more specialised functionality may be incorporated in the future.
HPCToolkit provides a breakdown of the execution time in the code by routine based on statistical timewise sampling and analysis of the call stack. Such an approach requires no modification whatsoever of the executable.
Marmot and Scalasca require the code to be instrumented and this must be done during compilation via modifications to the Makefile. The chosen applications consists of some 100,000+ lines of code and the compilation process is quite complex, requiring some special compiler options to be successfully completed. Marmot merely checks the MPI calls within the code and notes any unusual or potentially incorrect or suspicious behaviour.
Scalasca provides, through instrumentation, comparable output to HPCToolkit with the additional functionality of the detailed description of wall-time in terms of communication and computation. Furthermore, it also provides a tracing facility and extensive possibilities for viewing all derived data.

Using these tools, the following general program of activity has been designed for the performance analysis of applications:

- Definition of a benchmark **based on real life usage of the application.** Interaction with the target user is necessary to define this
- Compile a number of executables for various aspects of testing:
    - Fully optimised
    - Fully optimised with -g
    - Non-optimised with -g
    - Compiled with marmot
    - Optimised and compiled with Scalasca
- Do 6 runs:
    - Fully optimised
    - Non-optimsed with -g
    - HPCToolkit using Fully optimised with -g
    - HPCToolkit using Non optimised with -g
    - Marmot
    - Scalasca optimised
- Analyse marmot output for correctness or warnings
- Check time of runs for hpctoolkit and scalasca. HPCToolkit uses sampling while scalasca uses instrumentation so methods are completely different. Execution times should be comparable. If they are not need to create filter file for scalasca to bring them within ~5%
- Rerun scalasca jobs (if necessary)
- Compare HPCToolkit and Scalasca runs for consistency
- Perform analysis
    - Identify computational kernels
    - Identify associated algorithms

- Identify communication problems (load imbalance...)
- Check whether main kernel algorithms are up to date, check if they are good for GPUs or other specific hardware
- Make recommendations

2.7 Hands-on Event at NARSS

A two day hands-on event was coordinated at NARSS for the 6th and 7th of July. This event was intended to allow the chance for communication and collaboration between all members of the team and to provide the opportunity for coordinated execution of the work plan described above.
Unfortunately, due to rioting and unrest in Cairo during this period, a last-minute decision was made on the 1st of July to postpone the event for a more stable period, expected to be in September of 2011.

2.8 Performance analysis of the algorithms

To date, analysis of MESSy at JSC has been done with the performance analysis tools Scalasca, HPCToolkit and Marmot for a T42 model on between 64 and 128 cores.

The Marmot report did not indicate any MPI problems within the code (in terms of the correctness of the statements). Instrumentation with Scalasca introduced an enormous overhead that caused the walltime for the calculation to increase three-fold. This happens due to the huge amount of times that functions are called. To reduce this overhead, a function filtering file is used to ignore the instrumentation of particular functions. This can be done based on a report derived from the fully instrumented output using another Scalasca utility. With filtering, the execution time is similar to that using HPCToolkit and provides a similar distribution of execution time (as one should expect). Scalasca provides additional information in the form of being able to separate out MPI and execution time. Using this analysis, the conclusion to date is that (for this model) ~40% of walltime is spent waiting in MPI_Broadcast statements. Looking at the Scalasca analysis (and from talking to Hendrik Merx) this probably comes from the computational load imbalance derived from the use of a spherical coordinate system. Of the actual computation, the vast majority (>70%) is spent in the KPP components.

Benchmark runs on Planck were performed with the T42 benchmark. Multiple 3 day simulation runs were done to produce standard deviation in runtimes on a loaded system (to be able to test for actual improvements). A 4 week simulation run done to provide reference outputs for restart files (to test for binary equivalence of outputs). Compilation under Marmot 2.4.0 fails. This is apparently due to a bug in this version of Marmot. Marmot 2.3.1 will be installed on Planck and retested.

3 Mathematical Analysis and Algorithms

Progress in this task is dependent on the results of the performance analysis. Since progress in that area has been stunted somewhat by unforeseen events in Egypt, optimal progress in algorithms has not been made. Nevertheless, a number of important topics have been addressed and some initial results are available for the climate applications. These are described below.

3.1 Development of a regional cross-disciplinary group on "Mathematical Methods and Algorithms"

Team members from the sites have been identified and the LinkSCEEM Winter Training School was leveraged to initiate contact between team members.
Alan O'Cais at JSC will form the first point-of-contact for the regional groups to the JSC cross disciplinary teams. Markus Geimer will be the next level of contact for the Performance Analysis group at JSC. Prof. Salwa Nassar will form the first point-of-contact to the team at NARSS and Prof. Magdy Nagi will be the counterpart at BA and Patrick Fitzhenry the counterpart at CyI.

3.2 Algorithms analysed with respect to algorithmical optimisation

Discovered within the scope of this task, it was found that MESSy and WRF appear to have a common computational kernel in KPP that will facilitate active collaboration between all groups in a common area.

There are development efforts in the US to get the KPP package to run on GPUs. Also, Intel released a (stiff) ODE solver library in recent years that may allow for algorithm substitution within KPP.

Initial performance analysis results have shown that >70% of actual computation time is spent in the MECCA submodel of MESSy in our implementation. The memory access patterns are very clear within MESSy and given that the code has a large per process memory footprint, it would appear that some relatively naive OpenMP implementation within the code should benefit performance. This could be coupled with using the auto-parallelisation features of the Intel compiler so that one can focus on the Mecca submodel for explicit use of OpenMP without burning cores elsewhere during execution.

The load imbalance inherent in the application is connected to the same submodel. The load imbalance is due to the position of sun relative to the earth, with regions during dusk and dawn having more chemical interactions. The gird distribution of points therefore causes an inherent load imbalance within the algorithm.

Attempting to emulate the WRF effort and do a first GPU implementation of the KPP solver used by the code may ameliorate this situation. This is due to the fact that the amount of computation for GPUs is minor and the majority of time is spent simply transferring data from and to the accelerator. Such an approach could dramatically reduce the time spent waiting for the CPUs cores to synchronise. For MESSy, this coupled with the OpenMP could be very powerful on accelerated nodes. For WRF, porting a significant computational kernel to GPUs would also have obvious benefit.

4 Data Management

The goal of this task was to understand the current network landscape between the partner sites and propose data management and workflow strategies that would be applicable to the current environment. The task will also describe how the NCSA Tupelo middleware could be leveraged and recommend strategies and approaches to creating a shared distributed data storage capability among the partner sites.

4.1 Define Network Landscape

Each partner was asked to define a server on their site where the other partners could run tests to help determine network access and latency between the sites. The tests included ping, trace route, and data transfers. With the sites defined, each partner was asked to initiate the tests to each of the partner sites and record their data on the wiki. Additionally, the Cyprus Institute ran http response tests to the partner sites. All data aggregated from the partners is available at:

- http://eniac.cyi.ac.cy/display/LS2/The+Network+Landscape
- http://eniac.cyi.ac.cy/display/LS2/CyI+Network+performance+between+partner+sites

The data from the contributing partners is aggregated in the following graphs. This first graph shows the minimum, maximum and average time in milliseconds for a round trip request for ping and http responses from the Cyprus Institute to the partner sites.

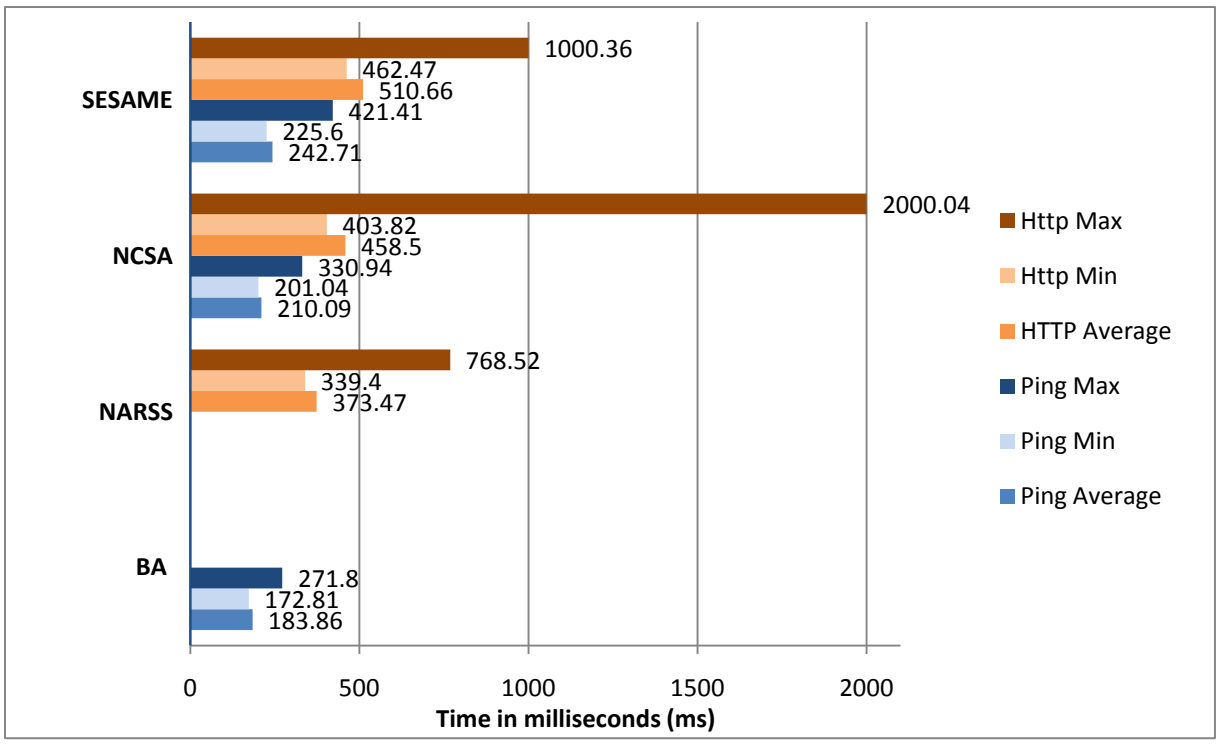

Figure 1: Network Connectivity from CyI to Partner Sites

The next graphic displays the data transfer time of a 480MB png file from a partner site to the Cyprus Institute in terms of minutes and seconds.

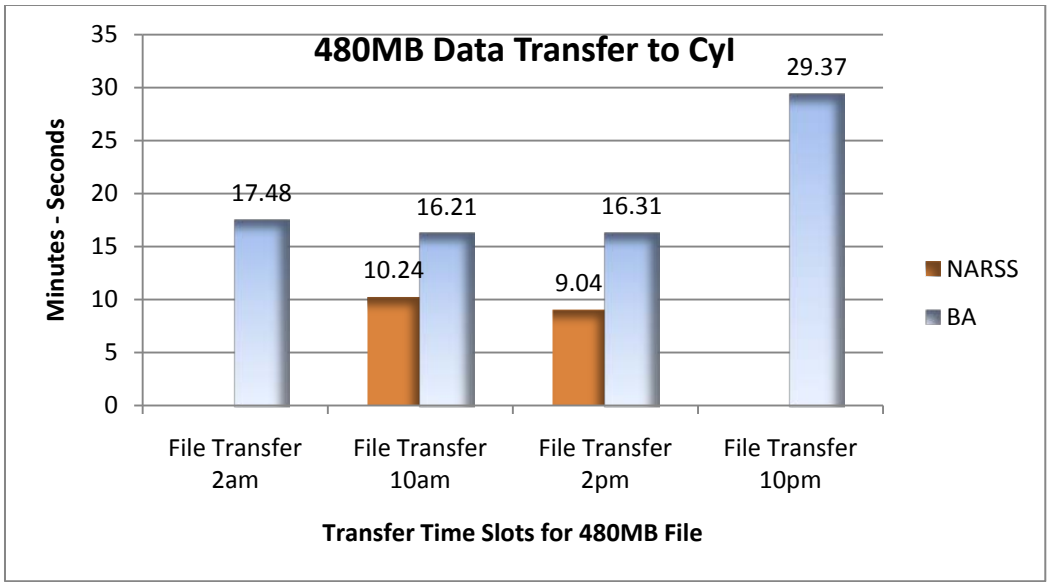

Figure 2: Data Transfer Time (min.sec) from Partner Sites to CyI

4.2 Research and Describe Alternative Access Options

There are a number of strategies for creating a shared distributed data space among the LinkSCEEM partner sites. This overview is focused on a few of the common open source strategies and technologies:

- Secure FTP (SFTP), secure copy (SCP),
- OpenAFS, http://www.openafs.org/,
- Globus tools like Reliable File Transfer (RFT, http://www.globus.org/toolkit/data/rft/), Replica Location Services (RLS, http://www.globus.org/rls/), GridFTP, http://www.globus.org/toolkit/data/gridftp/,
- Lustre WAN, http://wiki.lustre.org,
- BigTable data model, http://labs.google.com/papers/bigtable-osdi06.pdf.

Secure FTP and SCP use OpenSSH to create a secure tunnel from site A to site B. SCP is a basic file transfer mechanism while SFTP provides file system operations such as directory listings, remote file removal and the ability to resume interrupted transfers. A basic shared file system could be implemented with these tools with files being synchronized via a script running at some time interval or initiated by community members in an on demand fashion. While this approach provides a basic file shared file environment it does not provide a shared metadata solution.

OpenAFS is based on the Andrew File System (AFS) that was developed at Carnegie Mellon University. This is a trusted server model and provides a location transparent file system with an increased focus on security and scalability. AFS provides automated file synchronization, access control and with concurrent write access implemented by a file locking strategy.

The Globus Toolkit provides a variety of tools for building distributed data spaces. The primary tool set includes GridFTP, Reliable File Transfer (RFT) and Replica Location Services (RLS). GridFTP provides a secure, robust, fast and efficient mechanism for transferring data and is specifically suited for bulk data. GridFTP provides a server, a scriptable command line client and an API for building custom applications. The Reliable File Transfer (RFT) service provides an interface for controlling and monitoring file transfers

between GridFTP servers. Data is synchronized between sites using a job scheduler approach with simple source and destination URLs to represent directories or file blobs. Once scheduled, the system moves the files and provides status updates and notification of state changes. The Replica Location Services (RLS) provides a way to keep track of one or more copies of a file in a grid environment. RLS provides a server registry that is distributed to the different sites to provide redundancy in the system by avoiding a single point of failure. RLS creates a catalog of that contains the 'logical file name' (global) with mappings to the 'physical file name' (local location) along with other basic attributes like size and checksum.

The Lustre WAN provides a distributed environment where data can be aggregated and shared among different communities at different locations. Lustre is a Linux based distributed file system that can support thousands of nodes and petabytes of storage. Its design provides a single metadata server that stores filename, directory structure and layout information about its object storage servers or nodes. While Oracle has announced the decision to discontinue support of Luster on open Linux platforms and only support Oracle based hardware solutions there are a number of Lustre derived file systems that will keep the technology viable in the future. A few of these efforts include Open Scalable File Systems, OpenSFS.org, the European Open Source File System (OSF) SCE consortium and Whamcloud, whamcloud.com.

The Big Table data model was created and is used by Google to provide distributed storage for structured data across many commodity servers. BigTable leverages a basic row and column structure of a table with a time stamp to create an associated byte array that scales to petabytes across hundreds of thousands machines. Key implementations include the Google file system (GFS, http://labs.google.com/papers/gfs.html), Apache's Cassandra database, http://cassandra.apache.org/, or Hbase, http://hbase.apache.org/, which leverages the Hadoop Distributed Filesystem (HDFS, http://hadoop.apache.org/hdfs/). This technology provides very large scale, distributed access to structured data.

4.3 Test Optimization Options

The Tupelo middleware was designed as an abstraction layer that sits above a data store and enables a semantic data store. A semantic data store stores both the data blobs and metadata as a single entity and the middleware provides the abstraction layer and APIs that clients use to reference the data and metadata as single objects. The Medici semantic data store, http://medici.ncsa.illinois.edu/, provides both a backend and both a web client and desktop client interfaces to teh data store. It has been tested and optimized to work with a couple of community-drive configurations with the default configuration being installed on a local machine that hosts the data on local file system and metadata in RDF format in a MySQL server from a single server. An alternative configuration is to install Medici on a server that is configured with a Luster WAN. This configuration supports a local Medici server and local meta data store while hosting the data blobs on a remote Lustre system. We've tested this configuration on a Lustre service called the Data Capacitor which is hosted at Indiana University, http://pti.iu.edu/dc. This NSF funded multi-petabyte storage system provides users access to very large file system as mount point on the local system. While Medici has the ability to leverage a distibuted file system such as Lustre, its current implementation only provides single server access and it does not currently support a distributed front-end or metadata sharing capability. Additional design and software development would be required to enable Medici to work with multiple servers and share metadata between the servers and clients.

4.4 Recommend Optimization Options

There a number of distributed file system options with many common and unique features. Determining which backend data store and tool set will best serve the community needs to be discussed. It is recommended a team of scientific and technical members assembled to discuss and clearly articulate appropriate use cases and define system requirements. Afterwards, a smaller technical team could research the available options and make a recommendation about which backend technology and front-end client tools will best serve the community needs and requirements.

5 Visualization

Within this Task is intended that NCSA and CaSToRC CyI collaborate on installation of hardware and software for use with cultural heritage applications. The main objective, is to provide a tele-collaborative environment between at least two sites were the users can exchange and interact with flight viewpoints of the datasets located at each site.
The following steps have been taken toward this objective:

- Hardware preparation and Installation on site (Cyprus)
  A rear-projected stereoscopic display constituted by 2 JVC RS-25 full HD (1920x1080 pixel) projectors, polarizing filters and corresponding stereo glasses, large film screen, and a PC-based workstation with a spaceball 3-D input device were acquired and tested at NCSA in collaboration with a CaSToRC member. The equipment was then shipped to Cyprus and installed and tested on site (Cyprus). This has involved a strict collaboration between NCSA and CyI to solve upcoming installation problems. The hardware is fully functional and has been used for some preliminary dataset testing between NCSA and CyI.
- Software description and installation
  Installed software includes *Virtual Director*, an interactive graphical system supporting choreography and networked remote collaboration with time-varying 3-D datasets; *vmaya*, a package (integrated with Virtual Director) for interacting with 3-D scenes created with the commercial animation package, Maya; and *partiview*, desktop interactive software for displaying point- and surface-based 3D models. Also provided is software for playing stereoscopic HD animations, and creating those animations from image sequences. The preceding NCSA-created software is Linux-based; Windows 7 is also installed on the workstation.
- Training workplan
  Conference calls between NCSA and CyI have been initiated and a preliminary working plan of the training has been decided. In particular a person at the CyI, in the field of cultural heritage and with some capabilities in visualisation, must be identified for the training session. Before starting the training, data that relate to a particular case study will be identified and exchanged with NCSA for further examination and adaptation to their visualization tools. This data set will be used as case study for the training session. The training session is suppose to start not before than October-November 2011.
- Future Collaborations
  NCSA has also had some preliminary discussions with Wayne Pitard at the University of Illinois who is leading the current efforts in WP11 to determine if there are any opportunities for collaboration, or the use of visualization system that has been installed at CyI. Discussions will continue through the remainder of 2011and any opportunities that emerge will be discussed for possible inclusion in future work package deliverables