



Project no. FP6-004265

CoreGRID

European Research Network on Foundations, Software Infrastructures and Applications for large scale distributed, GRID and Peer-to-Peer Technologies

Network of Excellence

GRID-based Systems for solving complex problems

D.IRWM.05 – Report about the results on the joint research topics of all research groups, including implementation of prototypes

Due date of deliverable: 30 September 2006

Actual submission date: 15 December 2006

Start date of project: 1 September 2004

Duration: 48 months

Organisation name of lead contractor for this deliverable: PSNC

Revision final

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)		
Dissemination Level		
PU	Public	PU

Keyword List: Progress Report, Prototype, Integration, Information Service, Workflow Analysis, Network Monitoring, Checkpointing, User and Account Management, Web Services, OGSA

Contributors: PSNC, SZTAKI, MU, INFN, UoW,UMUE, FHG

Version	Date	Authors	Sections Affected / Comments
0.0	20.09	PSNC	Framework of the document
0.2	25.09	PSNC, INFN,SZTAKI	Description of prototypes and the current state of work
1.1	6.10		Further improvements
2.1	19.10	PSNC,MU,SZTAKI,UoW,FhG	All sections after the internal IRWM meeting in Kraków
2.2	23.10	PSNC	Summary, minor changes in other sections
2.3	08.11	PSNC	Changes after review

Table of Contents

1. Executive Summary.....	4
2. Introduction.....	6
3. Current Status	7
3.1. Information and Monitoring Services	7
3.2. Checkpointing.....	10
3.3. Workflow Services.....	12
3.4. User and Account Management.....	15
4. Trust & Security Issues.....	17
4.1. Information and Monitoring Services.....	17
4.2. Checkpointing.....	17
4.3. Workflow Services.....	17
4.4. User and Account Management	18
5. Future work.....	19
5.1. Information and Monitoring Services.....	19
5.2. Checkpointing.....	19
5.3. Workflow Services.....	19
5.4. User and Account Management.....	19
6. References.....	20

1. Executive Summary

This document provides an insight into progress in the work done in each of the research groups (*RG*) within the tasks of the Grid Information, Resource and Workflow Monitoring Services (*IRWM*) Institute. After the initial work of research groups, the Institute proposed an *Integrated framework architecture for the Grid Information, Resource and Workflow Monitoring Services* (D.IRWM.04). It became obvious that in case of some collaborations the proof of concept prototypes is a natural consequence of integration work, which was also suggested by group of reviewers (after the first annual review of the CoreGRID project) and the Scientific Advisory Board. In fact, most of the research groups had to do some implementation work in order to integrate products. There is one other reason enforcing some implementation work – the need to validate research results or new algorithms often lead to small “proof of concept” implementations and this is the case of most of the work groups of the IRWM Institute.

According to the D.IRWM.03 (the roadmap) most of the work groups are on schedule with their work. In case of some research groups there are, however, deviations from the assumed plan. A detailed explanation of the problems can be found in the description of each research group in the **Current Status** chapter.

Apart from some exceptions, most of the research groups’ work is advanced enough to enter the stage of proof of concept installations or prototype implementations. The complexity and usefulness of the prototypes varies and is affected mostly by the individual requirements of the particular RG.

The planned activities for the next stage of the project will focus on two, partially overlapping, areas. The first one is strictly connected to the guidelines provided by the roadmap document. This sort of activity encompasses the efforts to prepare a new or extending functionality of the already existing pilot installations. The planned functionality for the pilots is aimed at providing a subset of functionality vital from the point of view of the considered problem or supporting only the selected subset of use cases, rather than providing a standalone product. This means that in most cases the pilot implementations will need some additional effort to get them work with new services or provide functionality that is not strictly necessary at this level of development, so further integration with other services would require additional resources.

Apart from the pilot implementations, there are also plans to extend the currently devised architectures to match the most up-to-date technologies, like virtual machines, which have emerged recently. The updated versions of the architectures may enforce some changes in the pilot implementations and the experience gained from pilots may result in changes in the designed services, so both of the future work paths may provide mutual feedback for each other. Detailed plans for the near future regarding each of the tasks are presented in the **Future work** chapter.

In order to make the results of the research able to be considered valuable enough to deploy in production or scientific environments and compete with commercial solutions, emphasis has to be put on security and scalability of the designed services. These problems are described more precisely in relevant sections of **Current Status** and **Trust & Security Issues**.

The **current work** progress will be presented in detail in the next chapter, respectively for each work group in each of the tasks. There will be a short summary of the aims the group is striving for, followed by a detailed description of the current status of each work group along with remarks regarding the most important achievements and encountered problems. A short summary of work done in each of the four work groups is the following:

The work of the **Network Monitoring** task follows the guidelines indicated in the CoreGRID document D.IRWM.03 (Roadmap), and aims at the operational definition of the Grid Infrastructure Monitoring Services and the Network Monitoring Element. The infrastructure monitoring part is based on the C-GMA (*Capability-based Grid Monitoring Architecture*), which is an extension to the GMA (*Grid Monitoring Architecture*), defined by the Global Grid Forum. The original model was changed in order to increase the scalability by distributing the central mediator service. The current work focuses on further improvement of scalability of this service by applying DHT (*Distributed Hash Table*) algorithms. The other research thread is inspired by the concept of Network Monitoring Element (*NME*) which is the cornerstone of the Network Monitoring architecture and represents the set of features needed to perform the monitoring activity. Due to the guidelines indicating that the monitoring services must be able to scale in a seamless way, without extensive resource consumption, the work is focused on the passive monitoring approach which better adheres to the guidelines. The original passive monitoring toolset based on the MAPI interface has been enriched with the measurement of the packet loss rate using a novel end-to-end technique, whose accuracy has been experimentally validated. The other major activity in Grid monitoring is the GMT (*GEMLCA Monitoring Toolkit*) which monitors the state of the GEMLCA (*Grid Execution Management for Legacy Code Architecture*) services using customized probes. The monitoring results are collected by a portlet that is part of the P-GRADE portal. Several probes were implemented to collect information concerning the state of basic Globus services, local job manager functionality, and GEMLCA services. The probes can immediately be used as standalone tools. For more details regarding the work on Network Monitoring, please consult chapter 3.1

The second task of the IRWM institute strives to define a **checkpointing service** as a standard feature exposed and utilized by the Grid computing environment. Because of the complexity of the subject, the main topic has been divided into several threads, each coping with a different aspect of the problem. The work on each of the threads is performed in cooperation with partners from different institutes. The checkpointing team designed the revised version of the GCA (*Grid Checkpoint Architecture*) prepared in cooperation with partners from the Resource Management and Scheduling institute. The last version of the architecture was prepared with respect to current solutions and functionality provided by the Grid Brokers and the most up-to date mechanisms providing the checkpointing functionality (e.g. Virtual Machines). The most recent version of the architecture was described and presented at the IRWM meeting in Cracow. This version is mature enough to begin proof of concept integration of the key components making up the GCA. The work is now focusing on preparing a proof of concept Grid computing environment featured by the increased level of robustness and fault-tolerance. The fault-tolerance will be achieved by employing the checkpointing services deployed in a way that adheres to the guidelines provided by the GCA. The outcome of the integration will cover all major parts of the production grid – from the user interface to the local job manager running on computing resource. Additional information regarding plans for future work is shown in chapter 5.2. In parallel the different possibilities to achieve a sufficient level of robustness of the application images by using various approaches to scalable Grid storage, in context of both Desktop Grid and classical Grid computing environments, are investigated. The current status and issues encountered during the research are scrutinized in chapter 3.2.

The usage of Coloured Petri Nets for describing complex Grid Workflows is the main research topic for the partners collaborating in the **Workflow Service** task which is the next part of the IRWM institute. The subject was investigated resulting in a definition of the Grid workflow description language (*GWorkflowDL*). The language was subject to further extension in order to encompass hierarchical workflows by expressing subworkflows as single transitions. A prototype execution engine, the Grid Workflow Execution Service (*GWES*) is currently being implemented.

The fault tolerance is the subject of research not only in the checkpointing task, but in scope of interest of Workflow Services as well. The Grid Workflow Execution research group developed a monitoring architecture which is capable of providing real-time and historical data about the availability of computing, storage and legacy code resources. The resource status monitoring system has been integrated into the P-GRADE/GEMLCA Portal and graphical toolset thus made it interoperable with the EGEE (*Enabling Grids for E-science*) and UK NGS (*National Grid Service*) production Grids as well as with the GIN VO (*Grid Interoperation Now Virtual Organization*) of OGF (*Open Grid Forum*). The P-GRADE/GEMLCA Portal became the official resource testing portal of GIN VO due to the fact that this is the only portal that can monitor all the different Grid resources that are integrated within the GIN VO. The details are presented in chapter 4.3.

The efforts of **User and Account Management Architecture** task members focused on studying the state of the art on the area, along with a discussion on the existing solutions, need for combination and improvements especially in context of Virtual Organization specific requirements and confines. The research focused on extending the scope of resource virtualization which resulted in the introduction of Virtual Environment together with mechanisms required to access and manage these environments. Possibilities of transparent access and automatic management of Virtual Environments like Virtual Accounts and Virtual Machines were researched and compared, which resulted in a joint paper. The details can be found in chapter 4.4 for current status and 5.4 for the description of the ongoing work.

The original major key activities of the CoreGRID project were to prepare joint papers documenting the effects of collaboration between the participants, and it is very important that all the progress efforts are well documented and presented either as internal CoreGRID technical reports, or as joint papers presented at various conferences.

The overall conclusion of the current status of the work progress is good – the IRWM institute has already yielded major technical results, co-operation is good and the results, for most of the work groups, are being submitted in a timely manner.

2. Introduction.

The primary objective of the Institute on Grid Information, Resource and Workflow Monitoring Services is the development of general and scalable approaches to an information and monitoring infrastructure for large scale heterogeneous Grids. Current Grid information and monitoring frameworks have identifiable drawbacks, as they are either too focused on specific aspects or do not scale enough. The performance of the infrastructures is also not satisfactory, especially when security and reliability are required. The institute focuses its research to better understand the reasons and to find models and frameworks to overcome these limitations. A possibility for convergence of currently distinct approaches to information services and monitoring services is also being taken into account, with the aim to identify a unified framework.

The information provided by the monitoring services will be used to get a better understanding of Grid behavior. The current lack of understanding of grid performance in general, and the non-existence of generally accepted set of metrics to evaluate Grid performance, makes the task of Grid evaluation and performance comparison not possible. The institute focuses on the development of new Grid performance models that will provide the means and tools for the evaluation of services deployed on the Grid.

Complex job workflows represent another challenge, as the monitoring information must be synchronously gathered from many different sources and appropriately processed to provide a coherent view (state information) of the whole workflow and its components. The job workflow itself must be extracted from programming models and the monitoring and information services must be tightly coupled with job checkpointing and migration support to provide an environment where even complex job workflows could be easily deployed, executed, and monitored. Models and methods to provide a virtualized end user account system are a specific part of the combined job flow support and information services.

The full list of IRWM Institute partners is following:

- FHG – Fraunhofer Gesellschaft (Germany)
- FORTH – Institute of Computer Science, Foundation
- INFN – National Institute for Research in Nuclear
- MU – Masaryk University Brno (Czech R.)
- PSNC – Poznan Supercomputing and Networking
- SZTAKI – Computer and Automation Research
- UMUE – University of Muenster (Germany)
- UNICAL – University of Calabria (Italy)
- UCAM – University of Cambridge (UK)
- UNI DO – University of Dortmund (Germany)
- UoW – University of Westminster (UK).

Four major tasks are making up the IRWM Institute are: Information and Monitoring Services, Checkpointing, Workflow Services and User and Account Management. Due to the complexity of the subject matter each task has to cope with, the tasks were divided into several sub-tasks. For each of the sub-tasks a research group is established, attracting partners interested in this specific area. The aims of each research group, along with a list of partners, are presented in the **Current Status** chapter.

3. Current Status

3.1. Information and Monitoring Services

Contributing partners: FORTH

Grid Infrastructure Monitoring Services

This particular work is based on the former research done mainly at Masaryk University in the field of Grid infrastructure monitoring complemented by the work on Content-based Publish Subscribe Systems developed at INFN. The infrastructure monitoring part is based on the C-GMA (*Capability-based Grid Monitoring Architecture*), a Grid monitoring architecture which endorses interoperability and coexistence of various monitoring tools in a single Grid. The work is motivated by the heterogeneity imposed by connecting various virtual organizations into large-scale Grids (such as when connecting multiple NGIs into a multinational Grid project).

C-GMA is an extension to the GMA (*Grid Monitoring Architecture*), defined by Global Grid Forum. There are two key additions to the GMA: metadata layers and extension of the GMA component model by adding a new component: the mediator. In C-GMA, the original model is extended with several metadata layers – components provide metadata describing their behavior and requirements (their *capabilities*) and also metadata describing data they are willing to process (*attributes*). There are also different metadata layers, which describe the required data in terms of some query language (such as SQL, OQL, X-Query and others) The schematic behavior is shown in figure below.

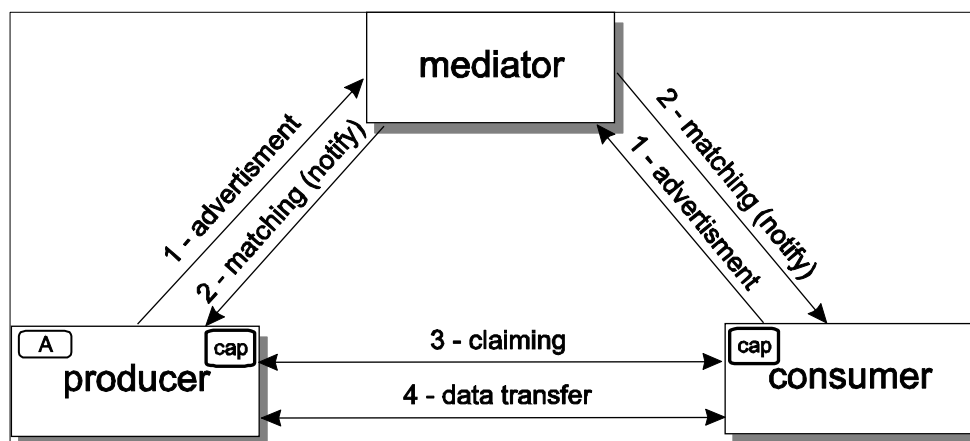


Figure 1. C-GMA Components Behavior.

The producer and consumer components are Grid monitoring elements which process the monitoring data. They rely on mediator component which is responsible on processing the advertisements and—in the process called *matchmaking*—locating the potential producers and consumers which are compatible and expressed interest in communicating with each others (in terms of data queries and definitions). This matchmaking process is metadata-driven. Our current experiments are using Classified Advertisements from Condor as a metadata language. It is obvious, that the mediator component may impose significant bottlenecks in the infrastructure, if designed improperly. Thus, the goal is to propose a scalable design of this component.

The subject of the join work in the context of the first task of IRWM is the development of the distributed mediator component, using the concept of the Content-based Publish Subscribe Systems. The involved partners have been MU and INFN. The overall design of the distributed mediator has been proposed and current state has been demonstrated at the DAPSYS 2006 conference in the paper entitled: *Designing a Distributed Mediator for the C-GMA Monitoring Architecture* [IM1]. The major added value of the combined work has been a scalability improvement of the original C-GMA framework, where the mediator has been a single element.

We have evaluated several approaches based on CBPS and proposed basic solution using this approach. We have also identified some drawbacks which are subject to future research to further eliminate overhead of the currently adopted solution and thus to further endorse scalability and robustness of the architecture. We are also planning to explore different approaches, such as DHT peer-to-peer networks, as alternate network infrastructure for the distributed mediator, with the potential to further increase the scalability of the whole C-GMA framework.

Currently, the research group is seeking new partner to continue in the joint research in this field, since INFN has left this group.

Network Monitoring Element

The work of the project is inspired by the concept of the Network Monitoring Element, as described at the Grid Integration Workshop in 2005.

The Network Monitoring Element is the cornerstone of the Network Monitoring architecture. It represents the capabilities needed to perform the network monitoring activity (presently we focus on passive monitoring tools for scalability reasons), managing this activity in coordination with other Network Monitoring Elements and brokers, submitting such information to an appropriate dissemination engine, which falls outside the scope of our activity. Network Monitoring Elements are located near the egress points of the Domains (not necessarily related with DNS domains), where the local infrastructure, whose monitoring is managed locally, meets the backbone, which cannot be inspected and needs to be monitored from the outside. The measurements reported by a Network Monitoring Element are Domain to Domain.

The guidelines for the design of the Network Monitoring Element are that the provided service must seamlessly scale with the system size, with a light footprint on Grid resources (here included the network itself), and high security guarantees, since the insertion of malicious agents can severely damage the overall Grid operation.

The work of the Network Monitoring project follows the guidelines indicated in the CoreGRID document D.IRWM.03 (Roadmap), and aims at the operational definition of the Network Monitoring Element. The passive monitoring toolset has been enriched with the measurement of the packet loss rate using a novel end-to-end technique, whose accuracy has been experimentally validated. A paper on this subject has been accepted by the CoreGRID Integration Workshop, in Krakow. The interface with the Grid environment relies on session descriptions submitted to the Network Element: while the descriptor of such sessions is quite clear, and is largely based on the API interface offered by MAPI, the passive monitoring API used for passive monitoring, the dissemination of such results is still an open issue, and relies on the either the design of an appropriate infrastructure, or on an existing GIS (one candidate is GridICE). The issue is presently considered opened, and the design strategy tries to leave any option open to that.

The Management of the Network Monitoring infrastructure requires the coordination of the Network Monitoring Elements that are distributed in the Grid. Such coordination must be flexible, allowing join and leave operations, scalable, ranging from tens to thousands of Network Monitoring elements, and secure, avoiding the insertion of malicious agents. An original solution to the problem has been found, and its soundness has been verified by simulation and by a small scale experiments. The basics of the proposed solution have been presented in September at DAPSYS, supported by simulation results. The solution is based on a randomized token passing protocol, and is appropriate for systems of hundreds of Network Monitoring Elements. A running prototype has been implemented in Perl, basically to test its practical feasibility.

The other major activity in Grid monitoring is the GMT (*GEMMLCA Monitoring Toolkit*) developed in the framework of the P-GRADE/GEMMLCA portal. In order to offer GEMMLCA legacy code services for production Grid systems, automatic testing of these services is inevitable. Production Grids run thorough tests on their available resources on a regular basis to offer a high quality of service. The UK NGS, for example, runs GITS (*Grid Integration Test Script*) tests to ensure the setup of Grid services on the infrastructure. Other, more advanced monitoring toolkits, like Inca, GRASP (*Grid Assessment Probes*), or MonaLISA (Monitoring Agents in a Large Integrated Services Architecture) are also available to test, monitor and verify the functionality of Grid

resources by running a set of probes. However, none of these solutions are integrated with GT4 (*Globus Toolkit 4*) at the moment, and there are no probes that can directly test GT4-based services. In order to overcome this problem the GEMMLCA Monitoring Toolkit (GMT) was developed to provide monitoring information based on probes concerning the status of GEMMLCA resources. Using the GMT, system administrators are automatically alarmed when a test fails and can also request the execution of any test on-demand. The GMT also assists P-GRADE portal users when mapping the execution of workflow components to resources by offering only verified Grid resources when creating a new workflow or when rescuing a failed one.

The main objectives of the GEMMLCA Monitoring Toolkit are the following:

- Test GEMMLCA resources in pre-defined regular intervals, and alarm and support system administrators in identifying any problems with GEMMLCA resources.
- Automate the validation of GEMMLCA Grid services, and provide a user-friendly interface for this task by integrating it into the P-GRADE portal.
- Provide reliable and dependable environment for GEMMLCA end-users by assuring them that the GEMMLCA resources where the tasks are mapped working properly.
- Support the further development of GEMMLCA by collecting information regarding resource availability.

In order to fulfill these generic objectives the following types of resources have to be tested:

- The basic network connectivity verifying that the remote sites are accessible.
- Services of the underlying Grid middleware. The current GEMMLCA implementation is based on GT4 utilizing the MyProxy, WS-GRAM and GridFtp services.
- Functionality of the local job manager. GEMMLCA is submitting the legacy code as a batch job to a local job manager. Current GEMMLCA implementation is capable to submit to the Fork and Condor schedulers.
- The three Grid services, GLCAdmin, GLCList and GLCProcess providing the GEMMLCA functionality.
- The implementation of the GMT is based on MDS4 (Monitoring and Discovery System) that is part of the Globus distribution.

Site administrators can configure the MDS4 service to run the various probes at pre-defined intervals. The results are collected by a portlet that is integrated into the P-GRADE portal. Administrators can also select a specific probe from a drop-down list displayed by a portlet and run it to verify the state of a specific service at a specific site on demand (Figure 2). GMT probes can also be integrated into the workflow editor of the portal to assist end-users when mapping a new workflow execution onto available Grid resources, or when rescuing and re-mapping a failed workflow.

As part of the GMT, several probes were implemented that collect information concerning the state of basic Globus services, local job manager functionality, and GEMMLCA services. The probes can immediately be used as standalone tools executed automatically from the MDS by implement these solutions and put them into production level operation on the UK National Grid Service. Based on the success of the GMT in the UK NGS representatives of the GIN VO asked us to provide the P-GRADE/GEMMLCA portal for the GIN VO that integrates resources from UK NGS, TeraGrid, OSG, EGEE and NorduGrid. For their request we extended the probes to monitor all these resources except for the NorduGrid. As a result the GIN VO uses the P-GRADE/GEMMLCA portal as their official production level resource test monitoring portal.

The screenshot shows the P-Grade Portal interface in a Mozilla Firefox browser. The page title is "P-Grade Portal" and the URL is "http://pakko:9080/gridsphere/gridsphere?cid=MDS4Monitor&JavaScript=enabled". The page features a navigation menu with items like "Workflow", "Certificates", "Settings", "Information System", "Help", and "GEMLCA Administration Tool". The main content area is titled "Monitor" and displays a "ServiceGroup Overview".

The "ServiceGroup Overview" section provides a brief overview of Web Services and/or WS-Resources that are members of a WS-ServiceGroup. It states: "This WS-ServiceGroup has 10 direct entries, 10 in whole hierarchy." Below this, a table lists the resources:

Resource Type	ID	Information	
gmtpingtest	pakko GMT Probe "gmtpingtest" for https://pakko:8443/wsrf/services/WidgetService		detail
gmtgemlcalistcodes	pakko GMT Probe "gmtgemlcalistcodes" for http://pakko:8080/wsrf/services/uk/ac/wmin/cpc/genlca/frontend		detail
GRAM	pakko 1 queues, submitting to 0 cluster(s) of 0 host(s).		detail
GRAM	pakko 0 queues, submitting to 0 cluster(s) of 0 host(s).		detail
gmtgridftpstest	pakko GMT Probe "gmtgridftpstest" for https://pakko:8443/wsrf/services/WidgetService		detail
gmtwsgramtest	pakko GMT Probe "gmtwsgramtest" for https://pakko:8443/wsrf/services/ManagedJobFactoryService		detail
gmtgemlcaur1stest	pakko GMT Probe "gmtgemlcaur1stest" for http://pakko:8080/wsrf/services/uk/ac/wmin/cpc/genlca/frontend		detail
gmtwsgramfttest	pakko GMT Probe "gmtwsgramfttest" for https://pakko:8443/wsrf/services/ManagedJobFactoryService		detail
RFT	pakko 0 active transfer resources, transferring 0 files. 2.47 MB transferred in 293 files since start of database.		detail
gmtgemlcagetdtd	pakko GMT Probe "gmtgemlcagetdtd" for http://pakko:8080/wsrf/services/uk/ac/wmin/cpc/genlca/frontend		detail

At the bottom of the page, it notes: "This Portlet uses code written by the Globus Alliance." and "XSLT transformation provided by servicegroupable.xsl version 1.5.4.1."

Figure 2. GMT Probe Results in the P-Grade portal – automatic execution.

3.2. Checkpointing

Contributing partners: PSNC, SZTAKI, UCO

The main long-term objective of task 5.2 is to devise and provide the concept of an architecture that would make the checkpointing technology available in the Grid environment. To fulfill that objective we proposed the Grid Checkpointing Architecture (GCA) that is a specification of the Grid Services, design patterns, rules, principles and tools that make it possible to integrate the already existing (legacy, old-fashioned, not Grid-aware) and future checkpointers with the Grid oriented computing environment. The undertaken task is complex and difficult as such. The legacy checkpointing packages do not fit the Grid environment at all and differ greatly in functionality, requirements and interfaces. There are not any standards or specifications, neither commercial nor open implementation, concerning the low-level checkpointing techniques. Each checkpointing package can be completely different in its semantics. Additionally, as regards the future checkpointing packages we are not able to anticipate all possible scenarios and semantics of those packages. It all implies that the GCA should be as flexible and as adaptable as possible.

The GCA as the specification is quite broad and touches numerous areas and aspects of the Grid technology. One of important elements of the GCA is set of so called Core Services (that is the actual low-level checkpointers). Within the confines of the GoreGRID we are going to provide the implementation of one Core Service that will be able to deal with PVM applications on the Altix platform.

Another aspect of the GCA is the issue of cooperation between the Core Services and the Local Resource Managers. To gain experience in that area we are working on integrating the AltixC/R with the TORQUE

Resource Manager 2. The AltixC/R is the checkpointer developed at PSNC and the TORQUE is an open source resource manager.

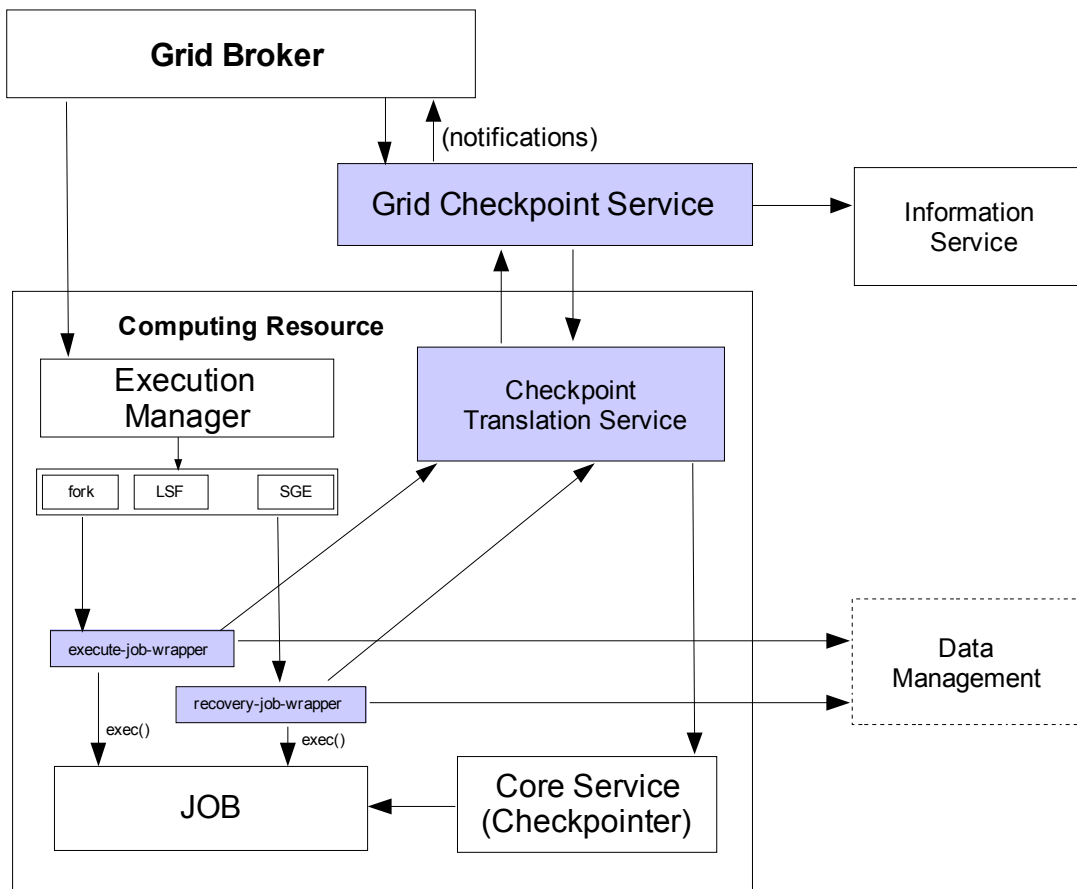


Figure 3. Revised general architecture of GCA.

Currently the task 5.2 strives to fulfil the goals defined in the D.IRWM.03. There are several research groups focusing on different aspects of the appliance of checkpointing service in the Grid environment. Work on architecture of the Checkpoint Grid Service is progressing. The refined and revised version of the GCA was presented in the technical paper [CKPT1]. The architecture was redone with special attention to the cooperation with existing brokers and with regard to real implementations of Grid services. Emergence of VMM (*Virtual Machine Manager*) technologies allowing to divide physical machine to many virtual machines forced us to rethink the approach to the checkpointing subject because of the unique capabilities of the virtual machines. One of such capabilities is the feature that allows to freeze the state of the virtual machine. This consequently enables the migration of the entire virtual machine to another node without losing the intermediate computation results. The state of the frozen virtual machine may be considered as a checkpoint, hence it should be included in the GCA. The initial ideas of incorporating the VMM technology was mentioned in the current version of GCA however we have to scrutinize this technology more closely in order to verify whether the way we want it to cooperate with Grid services is a good one. As the flexibility and scalability were primary concerns during the first stages of work on the GCA, the current version does not need any further major changes to enable deployment in large scale test-beds. Support for the plethora of different core checkpointers, free and commercial ones, was one of the major assumptions, therefore the GCA may be perceived as a heterogeneous-environment-ready solution. Currently, due to before mentioned reasons, the checkpointing functionality is rarely used in commercial products and the existing integration solutions we are aware of are sharing neither the same motivation nor the ultimate goals. The components making up the architecture of the GCA are depicted in the Figure 3.

The other thread of the research – namely the integration of the Core Service with PVM applications is being prepared. Support for the PVM standard, rather than MPI, is caused by the fact there is no project hosted by any of the CoreGRID partners that would allow for such integration. The general approach, however, allows for integration with the MPI library if it supports cooperation with external checkpointers. Currently, the integration encompasses kernel level checkpointing and the modified PVM library that allows for disconnection of the applications while taking checkpoint and further re-establishing the network connections when restarting or resuming the computations. The integration of Total Checkpoint (TCKPT) developed in SZTAKI with the

AltixC/R kernel checkpointer prepared in PSNC is almost completed. The reason of the delay is an additional amount of work introduced by the Itanium architecture specific problems that emerged during the work on customizing the kernel checkpointer to match the requirements of the Total Checkpoint package. The significance of applying TCKPT in Grid can be understood if we want to checkpoint and migrate parallel (PVM) applications among cluster resources of a Grid system. Since clusters can have different software environments installed, the relevant design goals or requirements of a parallel checkpoint tool are *compatibility* (with the surrounding software components) and *integrity* (of the checkpoint information of the application). While the first goal ensures the seamless operation of checkpointer on clusters with various middleware, the second one is a basis for application migration among clusters.

In order to fulfill the compatibility requirement the following conditions must be accomplished:

- Operating system cannot provide checkpointing facility
- Solution cannot rely on checkpoint support of the job manager
- Solution must rely on the native version of message-passing system
- Dependence from external auxiliary process cannot exist

These 4 conditions correspond to compatible operation of checkpointing frameworks. Following these conditions an application can be checkpointed in a way which enables the application to be checkpointed in software heterogeneous ClusterGrid environment i.e. under the control of any kind of execution environment. Moreover, the application will not be limited to be resumed under the same execution environment where it was checkpointed i.e. it is compatible with the different software environments installed on the clusters. The TCKPT concept and architecture has been published at the DAPSYS'06 international workshop under the title: „Checkpointing with TCKPT on ClusterGrids – A novel checkpointing approach” [TCKPT].

To gain experience in the area of utilizing the Core Services on nodes managed by Local Resource Managers the AltixC/R checkpointer has been integrated with TORQUE. After integration we are able to force the TORQUE to take checkpoints of jobs in a periodical way. Later the jobs can be submitted to be executed from the point where the last checkpoint has been taken. The TORQUE system along with the integrated checkpointer will be used in the GCA proof of concept demo that will present operational, checkpoint-aware Grid computing environment.

One of the subjects related to checkpointing in the Grid environment that we scrutinized more in depth is the problem of storing the images of applications in a highly distributed and heterogeneous environment. The aim of the study was to ensure that the application image data will be available even in case of local storage failure. The results of the study are described in the paper entitled ”BackupGRID: Using Desktop Nodes to Provide a Grid Storage Service” [CKPT2] that will be presented on the Cracow Grid Workshop 2006.

3.3. Workflow Services

Contributing partners: UMUE, FHG, UNICAL, INFN, UoW, MU

The main goal of workflow services is the management of complex jobs and service-level agreements. Those jobs are often described and automated as workflows in a distributed environment with co-allocation constraints and dependencies that must be considered for a wide range of diverse resources. Present systems have architectural and design limitations that make them usable in a productive manner only for simple and static workflows. This is not sufficient for highly complex GRID applications to be expected in important application domains such as industrial design, engineering, drug design and bioinformatics. The current limitations must be overcome by a common set of workflow management and execution services based on a powerful model. The goal of this task is to define and provide services able to coordinate the execution of vastly complex compound and dynamic GRID jobs represented by workflows without a need for user supervision.

To establish a job workflow service on a GRID, there are two major requirements: an adequate description of GRID jobs and a service that takes care about the reliable execution of the GRID jobs.

Workflow description languages using high-level Petri nets for GRID workflows

The objective of this research group is to develop and promote a common Grid workflow description language which is based on the Petri Net formalism.

The research group specified a common workflow description language, and coordinates the implementation of tools for creating, parsing, analyzing, modifying and monitoring description documents based on this language.

The following specific results have been achieved:

- The usage of Coloured Petri Nets for describing complex Grid Workflows was investigated and the GWorkflowDL was defined. The language makes use of available Grid middleware but is extensible and not bound to a certain one. Different underlying middlewares are addressed by using different extensions to the platform-independent GWorkflowDL. A first draft of the GWorkflowDL was presented at PPAM 2005 in

Poznan [WFL1] and published as the CoreGRID Technical Report Number TR-0032. The next step was to adapt it to our needs so the language was extended to encompass hierarchical workflows by expressing subworkflows as single transitions. The results were published at the CoreGRID Integration Workshop 2005 in Pisa [WFL2]. The GWorkflowDL is being validated, by using it to describe the workflow for non-trivial scientific case studies. The first results have been accepted for publication at the 2006 e-Science Conference to be held in Amsterdam [WFL3]. The paper is accepted for publication. The overall results and experiences with using Petri Nets for Grid workflows are summarized in a chapter of a book that will be published by Springer-Verlag [WFL4].

- A prototype execution engine, the Grid Workflow Execution Service (GWES) is currently being implemented. It coordinates the creation and execution process of Grid workflows specified using the GWorkflowDL. It provides interfaces to a Web Portal for user interaction and to the Low-Level Grid Middleware for the invocation of application operations.
- In a theoretical part, the task of modeling different aspects (e.g. cost information) of Grid services or components using Coloured Petri Nets are studied in order to analyze complex Grid workflows and allow for efficient execution. An experimental analysis tool is currently being developed in the context of a master thesis at UMUE
- The research group also maintains the Grid Workflow Forum (<http://www.gridworkflow.org/>) as a collaborative platform for public information exchange and discussions about scientific and commercial approaches in the domain of Grid Workflows

Compatibility and conversion of different GRID workflow description languages

The objective of this research group is a joint survey about compatibility and conversion (mapping) issues between commonly available workflow description languages. Although the group already has surveyed many different approaches and accumulated information in the Grid Workflow Forum (<http://www.gridworkflow.org/>), the research group is currently “on hold”, because one partner (INFN) has left the group.

Fault tolerance in Grid workflow execution:

The research group aims at extending the most widely used production Grids by a fault tolerant workflow manager layer. Globus, LCG and gLite middleware-based systems are taken into consideration. Following the concepts presented in the joint publication of the Research Group:

- the partners developed a monitoring architecture which is capable of providing real-time and historical data about the availability and computing, storage and legacy code resources (see Chapter on Information and Monitoring Systems). The resource status monitoring system has been integrated into the P-GRADE Portal and GEMLCA graphical toolset thus made it interoperable with the EGEE and UK NGS production Grids.
- the prototype system currently provides information for the end users of the GEMLCA – P-GRADE Portal. Integrating the system with a broker in order to support it with feedback information about the actual and historical status of Grid resources is the topic of ongoing and future research. The GTBroker is a broker developed in SZTAKI in order to support Globus based Grids by a fault-tolerant broker mechanism. GTBroker selects the best resource based on the user provided criteria and manages the job submission and necessary file transfers to the selected Grid resource. The unique fault-tolerant feature of GTBroker is that it periodically checks the status of the submitted job and if the job is not started on the Grid resource for a certain amount of time GTBroker selects another Grid site based on the user requirements and transfer the job and its files to the new selected site. GTBroker carries on this migration of the job until finally a resource can execute the job. We made a comparison measurement between the LCG broker and GTBroker from the point of view of reliability and speed in two different Grids. Figure 4 shows the measured results on VOCE (Virtual Organization Central Europe of EGEE). The figure clearly shows that GTBroker significantly better performs (except one job) than the LCG broker. Figure 5 shows the measured values on SEE-GRID (South-East European Grid). Here 60 jobs are compared and it can be seen that GTBroker was always able to manage the 10 minute jobs within half an hour while 25% of LCG Broker manages jobs could not be finished even in an hour. Based on these very promising results we have decided to integrate GTBroker into the P-GRADE portal in order to provide it as a service for the UK NGS and for the EGEE VOs. This work has been published at the Austrian Grid Symposium’06 under the title: “Multi-Grid Brokering in the P-GRADE portal” [GTBroker].

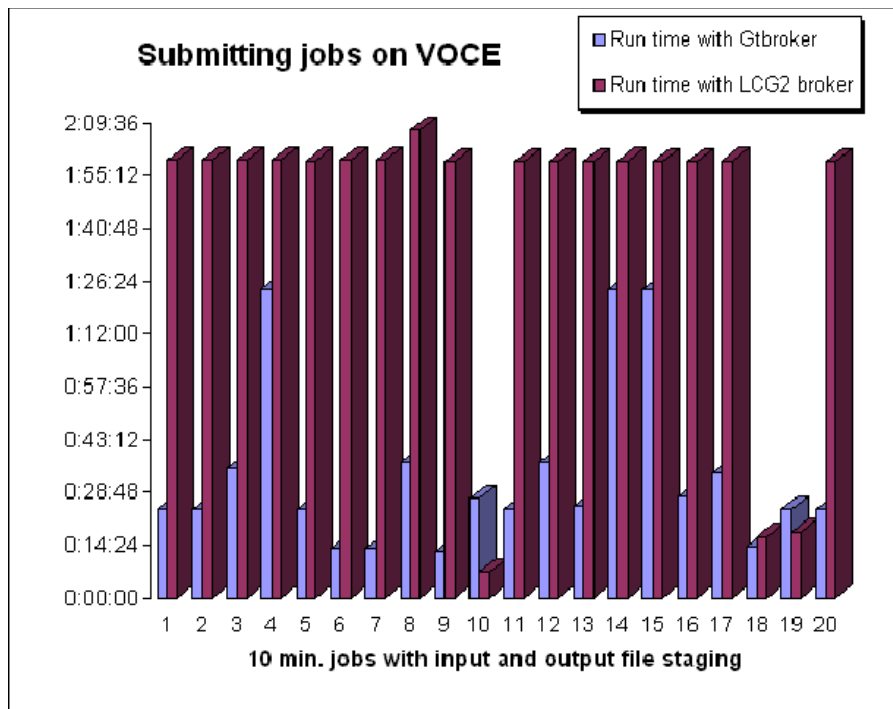


Figure 4. Measured results on VOCE.

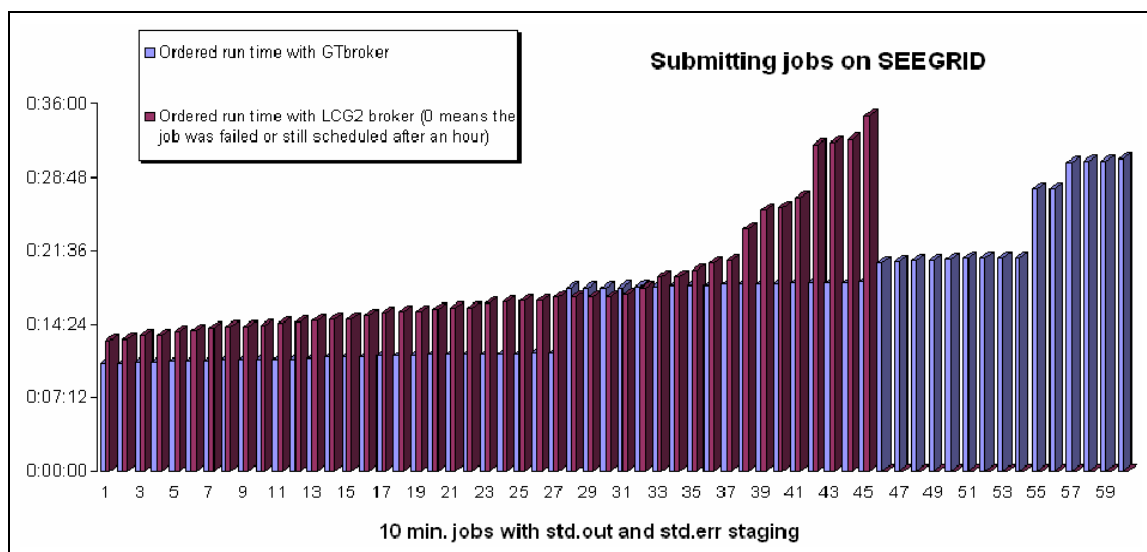


Figure 5. Measured values on SEE-GRID.

Workflow-oriented Grid infrastructure for biomedical purposes

The long-term goal of the research group is to build a biomedical Grid for collaboration of biomedical experts and knowledge publication and sharing. This research group consists of three partners: MU, UoW and MTA SZTAKI, the research is done as a direct three partner cooperation.

Members of the research group have already

- identified use cases for the medical architecture. The biomedical community targeted by the architecture is quite diverse, with many scattered research laboratories, university departments and hospitals included. In such environment knowledge required for a medical experiment is very often distributed among several parties. The knowledge is represented by software services, executable jobs or possessed by medical experts. High-level tools are required to support the integration of the scattered knowledge typically into

workflows, to execute and supervise the progress of these processes. Very often real time collaboration among several parties is also required as a supportive task for the workflows.

- specified the general architecture. The backbone of the architecture would be a computational Grid built with one of the standard Grid middleware solutions. Since Grid services provide relatively low-level functionalities – such as job execution, storage facility and resource brokering – additional layers must be built on top of that to realize the required high-level specialized services:
 - Central module library – logically centralized (in practice may be distributed) library of medical applications
 - Central ontological library – logically centralized (in practice may be distributed) library of data interfaces of the medical modules
 - Module repository – repository of existing modules and references to resources on which they are available
 - User interface – practically a Web portal which provides a virtual desktop for medical experts.
- implemented the first prototype using the P-GRADE Portal GEMICA toolset. The first implementation was realized on the Central European VO of the EGEE Grid. The partners have developed and executed example applications on the infrastructure in order to demonstrate the capabilities and to collect requests for improvements from the medical parties. This work has been presented on the 6th EGEE Conference in Geneva.

3.4. User and Account Management

Contributing partners: **PSNC, MU**

The main aim of the user management system is controlled, secure access to Grid resources. The considerations we try to introduce must take into account the fact that we are dealing with a production GRID environment which has ability to change its configuration dynamically. Security requires authentication of the user and authorization based on combined security policy from the resource provider and virtual organization of the user. The second important issue is a possibility of logging user activities for accounting and security reasons and then gathering these data both by the resource provider and virtual organization of the user. From the user's point of view, an important feature is single sign-on.

The problem of user management is a non-trivial one, especially in an environment that includes a bulk number of computing resources, data, and hundreds or even thousands of users participating in lots of virtual organizations. The complexity rises from the point of view of time required for administration tasks and automation of these tasks. There are many solutions that attempt to fulfil these basic requirements and solve the mentioned problem, but none of them, solve the problem in a complex and satisfactory way.

The research topics include virtualization of user resource access on the heterogeneous Grid, investigation of approaches for real-time on-demand virtual environment (user account or virtual machine) creation and management, support for hierarchical VOs, user and job separation, correct data protection (including failure recovery) and accountability, neutrality with respect to the actual job submission and authorization service used.

Virtualization of resources can be scalable depending of the method used for virtualization. Virtual account method is fully scalable, it does not introduce any significant overhead on resources. With virtual machines situation is different. Each virtual machine runs its own instance of operating system, using some amount of memory and CPU. Using virtual machines make sense if there are only few virtual machines per physical machines. In this context it is not scalable. but in our package we work not on the virtual machine itself but rather on using them in Grid resources, manage and account. Our architecture of management of virtual environment is fully scalable and does not depend on number of virtual environments managed.

There are some commercial products for vitalizing resources. But our research focus on general concept of virtualisation and integration of virtual environment with Grid middleware. Regarding to our knowledge there are currently no commercial solutions for Grid accounting and Grid resource virtualisation.

In our implementation we use Usage Record format prepared by GGF Usage Record Working Group. The schema is already finished.

Currently there are no standardization group for virtualization, if such group will be formed we will joint it.

The current status of the work on the User and Account Management Architecture is the following:

- State of the art on the area was researched and different approaches and concepts were explained. Requirements concerning User and Account Management were collected and discussed. The results were presented during the PPAM 2005 conference and the paper [MGMT1] printed in the conference proceedings

- An existing solutions, need for their combination and improvements were discussed. Special attention was put on enabling accounting and audit features with the full context of user identity, role and Virtual Organization. The concept of Virtual Environment was introduced in order to gain a higher level of virtualization of the resource access that fulfill the identified requirements. The result of this work is the paper [UMVO] , presented during the CoreGRID Integration Workshop 2005 and to be printed in the conference proceedings. The work is also described in more detail in the CoreGRID Technical Report TR-0012.
- The further work was focused on the architecture of framework for virtualized access to the computational resources in the Grid. Possibilities of transparent access and automatic management of Virtual Environments like Virtual Accounts and Virtual Machines were researched and compared. A detailed architecture of such a system was described in the paper [VE1], presented during the CoreGRID Workshop on Grid Middleware 2006 and to be printed in the conference proceedings.
- The research group contributed to deliverable D.IRWM.04: [VE2] and IRWM JPA3.

4. Trust & Security Issues

4.1. Information and Monitoring Services

The C-GMA concept requires two levels of security and trust management. The interaction between producers and consumers and the mediator requires an authentication scheme based on trusted third parties (such as X.509 certificates). As the mediator could not restrict new producers or consumers to join and publish data, the certificate based authentication is used just to check the appropriateness (the element does possess a valid certificate verifying its authenticity with respect to trusted certification authorities) but not to impose restrictions in terms of access control (i.e. no strict authorization in a general framework is required). The second level deals with the actual interaction between selected producer and consumer, where the C-GMA concept on purpose does not dictate any explicit security mechanism and deliberately leaves on the partners (the producer and consumer) willing to communicate to select the authentication and authorization method best suited to their need (low to none for non sensitive data, very high and complex authentication and authorization for highly sensitive data). However, the C-GMA metadata layers offer means for specifying authentication and authorization mechanisms (using capabilities and attributes) so that these non-functional requirements may also be a subject to the described matchmaking process.

The accent is on the protection of the membership of Network Monitoring Elements from intrusion. A malicious agent could perform arises of damages, among others:

- publish inexistent resources,
- overload existing resources,
- obtain traffic statistics for malicious purposes
- compromise the coordination between Network Monitoring Elements.

For the above reasons, we consider the secure join/leave operation as a primary concern.

To this purpose, we consider that each NME owns a certificate (released by a Grid Certification Authority), and that such a certificate is used in order to join the membership of Network Monitoring Elements. Successive interactions between NMEs are encrypted using public keys. The lightweight dissemination of public keys is part of the coordination protocol which is in development.

4.2. Checkpointing

The services making up the GCA are encountering the security issues, as there is need to access to the user's job metadata and ensuring security while handling the images of the applications. As the images might be considered as a special files belonging to owner of the application, the GCA components will rely on standard Grid mechanisms managing access to user-owned resources. Encryption of the files might be a feature increasing the security of the files, but by default the GCA trusts the storage services to be secure.

4.3. Workflow Services

Workflow description languages using high-level Petri nets for GRID workflows

The goal of this workgroup is to develop a platform-independent workflow language and platform-specific tools. The tools developed in this research group have to be able to take into account and make use of the trust and security mechanisms provided by the underlying Grid platforms, such as e.g. the widely used Grid Security Infrastructure. In order to achieve this goal, workflow managers must implement the security requirements of these infrastructures.

Fault tolerance in Grid workflow execution

The research group is using the EGEE and UK NGS production Grid environments for its work. Both of them are using the Grid Security Infrastructure (GSI). The research group does not foresee any demand for stronger security or trust management mechanisms.

Workflow-oriented Grid infrastructure for biomedical purposes:

Medical communities raise serious concerns about the security solutions of current production Grids. As it has already been expressed by the biomedical user community of the EGEE Grid, the Grid Security Infrastructure –

the security infrastructure used by all the main production Grids worldwide – does not provide an acceptable level of security and flexibility for several application areas. Consequently, the EGEE developer community is already working on a set of new solutions to fulfil the needs of such communities. Because these systems are in a prototype (or even more initial) stage, our research group is only following their status but has not started their adoption. As soon as they are available, our group will adapt and integrate them into the targeted medical workflow environment. The efforts of this task has been published in [BW1].

4.4. User and Account Management

The security of virtual account management is crucial and comprises the following elements:

- identification (authentication) of users
- authorized access to computational resources
- control of user actions in the terms of audit (logging) and accounting
- proper level of isolation of user jobs
- authorized access to accounting and audit data
- confidentiality of the data sent over the network

In the proposed framework for Virtual Environments authentication and confidentiality issues are addressed by the existing Globus Grid Security Infrastructure. Fine-grained and flexible authorization is achieved by the Globus Toolkit v.4 authorization framework and a set of authorization plugins, implementing different authorization methods and reusing existing services like e.g. VOMS. Authorization enforcement and job isolation is assured by mapping the user to properly configured VE. Control-over-user actions (audit and accounting) are possible thanks to the VE Database which contains data from the local logs stored in the context of the global user identity and his/her Virtual Organization.

5. Future work

5.1. Information and Monitoring Services

The goal of the infrastructure monitoring research group is to extend the work on mediator scalability using the peer to peer approaches. A new partner is actively sought for to fill the gap created by the leave of the INFN researchers that worked in this area.

Efforts will be made to integrate both monitoring efforts (network monitoring and infrastructure monitoring) into a single multi-partner research taskforce.

The next step in the Roadmap D.IRWM.03 indicates a prototyping activity in October. Such activity will focus on the integration between the membership management activity, and the creation of the monitoring layout. A short visit is planned in order to coordinate the partners towards this result.

The support for workflow monitoring, based on the LB service developed by MU team members. This will be done in a collaboration with partners working on workflow description and submission systems. The work will focus on extensions of workflow description and processing with a monitoring component.

5.2. Checkpointing

The outlined shape of the GCA is considered as the intermediate one. Above all, in the nearest future we plan to extend the architecture to fully support the up to date virtualization technology. When we reach an agreement on the general form of the GCA, the next phase of the work on GCA will include a more detailed and formal description of the interfaces, the WS-Resource Properties and the additional tools.

5.3. Workflow Services

Workflow description languages using high-level Petri nets for GRID workflows

The research group is currently implementing prototypes for workflow analysis and execution. The prototypes will be validated by implementing nontrivial case studies and the GWorkflowDL will be extended and augmented if necessary. Also, the current workflows documents have to be created manually. We plan to implement a tool for the graphical creation of GWorkflowDL workflows.

Fault tolerance in Grid workflow execution:

The currently available prototype system provides information on computing resources and on application services for end users through a Web-based portal interface. Based on resource availability users must manually modify or reallocate the components of their workflows. Ongoing and future work is focused on the integration of the monitoring data with a workflow manager system in order to support the automated predictive and fault tolerant scheduling. Because the broker components of the Grids that are used by the research group (EGEE and UK NGS) do not support workflow level optimization and fault tolerance, the result of the research group can extend the capabilities of these production infrastructures.

Workflow-oriented Grid infrastructure for biomedical purposes:

The research group set up the prototype version of the Grid based workflow-oriented medical desktop environment. The partners will demonstrate its capabilities to the medical experts they are in contact with (due to other national projects.) Based on the feedback coming from these medical partners, the research group will extend the system with support for collaborative team work.

5.4. User and Account Management

A proposal of paper on virtualization techniques in resource access [MGMT2] was prepared and the abstract was accepted for the Cracow Grid Workshop, to be held in November 2006. The CoreGRID technical report on the Virtual Environments framework is work in progress. The next step will be a proof of concept implementation and deployment of such a system.

6. References

- [CKPT1] Gracjan Jankowski, Radoslaw Januszewski, Rafal Mikołajczak, Jozsef Kovacs: Towards checkpoint aware GRID
- [CKPT2] BackupGRID: Using Desktop Nodes to Provide a Grid Storage Service
- [TCKPT] J. Kovacs: Checkpointing with TCKPT on ClusterGrids, DAPSYS'06 Int. Workshop, Innsbruck, 2006
- [WFL1] Martin Alt, Andreas Hoheisel, Hans-Werner Pohl, Sergei Gorlatch: A Grid Workflow Language Using High-Level Petri Nets. In: Proceedings of the PPAM05, Poznan, 2005
- [WFL2] Martin Alt, Andreas Hoheisel, Hans-Werner Pohl, Sergei Gorlatch: Using High Level Petri-Nets for Describing and Analysing Hierarchical Grid Workflows. In: Proceedings of the CoreGRID Integration Workshop 2005, Pisa, 2005
- [WFL3] Martin Alt, Sergei Gorlatch, Andreas Hoheisel and Hans-Werner Pohl: Using High-Level Petri Nets for Hierarchical Grid Workflows.
- [WFL4] Andreas Hoheisel and Martin Alt: Petri Nets. In: Workflows for eScience, Ian J. Taylor, Dennis Gannon, Ewa Deelman, and Matthew S. Shields (Eds.), Springer, 2006
- [MGMT1] Best Practices of User Account Management with Virtual Organization Based Access to Grid
- [MGMT2] Virtualized Access to the Grid Computational Resources
- [GTBroker] A. Kertesz, G. Sipos and P. Kacsuk: Multi-Grid Brokering in the P-GRADE Portal, Austrian Grid Symposium'06, Innsbruck, 2006
- [IM1] Ondřej Krajíček et al. Designing a Distributed Mediator for the C-GMA Monitoring Architecture. In: Proceedings of the DAPSYS 2006 Conference. Innsbruck, 2006.
- [BW1] Ondřej Krajíček et al. Building Biomedical Grid Infrastructure using P-GRADE Portal and GEMLCA. Poster at the CoreGRID Integration Workshop, Cracow, 2006.
- [UMVO] Jiri Denemark , Michał Jankowski, Ludek Matyska, Norbert Meyer, Miroslav Ruda , Pawel Wolniewicz : User Management for Virtual Organizations
- [VE1] Michal Jankowski, Jiri Denemark, Pawel Wolniewicz, Norbert Meyer and Ludek Matyska: Virtual Environments - Framework for Virtualized Resource Access in the Grid. In Proceedings of the CoreGrid Workshop on Grid Middleware in Conjunction with Euro-Par Conference, Dresden, Germany (to appear), August 2006.
- [VE2] Integrated framework architecture for the Grid Information, Resource and Workflow Monitoring Services