PHOSPHORUS: Single-step on-demand services across multi-domain networks for e-science

S. Figuerola[1a], N. Ciulli [b], M. De Leenheer[c], Y. Demchenko[d], W. Ziegler[e], A. Binczewski[f] , on behalf of the PHOSPHORUS consortium

[a] i2CAT Foundation, Gran Capità, 2-4 Nexus I, 08034 Barcelona, Spain
[b]Nextworks, Via Turati 43/45, 56125 Pisa, Italy
[c]Ghent University - IBBT, Gaston Crommenlaan 8, 9050 Gent, Belgium
[d]University of Amsterdam, Kruislaan 403, 1098SJ Amsterdam
[e]Fraunhofer Institut SCAI, 53754 Sankt Augustin, Germany
[f]Poznan Supercomputing and Networking Center, Noskowskiego 12/14 , 61-704 Poznan, Poland

ABSTRACT

The Phosphorus[1,2] project focuses on delivering advanced network services to Grid users and applications interconnected by heterogeneous infrastructures. The project is addressing some of the key technical challenges to enable on-demand end-to-end network services across multiple domains. The Phosphorus network concept makes applications aware of their complete Grid resources environment -computational and networking- and its capabilities. Phosphorus enables and tests dynamic adaptive and optimised use of the heterogeneous network infrastructure interconnecting various high-end resources. The project will demonstrate on-demand service delivery across access-independent multi-domain/multi-vendor research network test-beds on a European and worldwide scope. Phosphorus enhances and demonstrates solutions that facilitate vertical and horizontal communication among applications middleware and the network resources across different domains, managed by existing Network Resource Provisioning Systems (NRPS), or domains that integrate a new Grid-GMPLS (G^2MPLS) Control Plane, both under a new AAA architecture to support policy based on-demand network resource provisioning. This G^2MPLS extends ASON/GMPLS in order to provide part of the functionalities related to the selection, co-allocation and maintenance of both Grid and network resources, by exposing upgraded interfaces at the UNI and E-NNI network reference points -i.e. G.OUNI and G.E-NNI-. The project outcomes are going to be demonstrated in a worldwide test-bed.

Keywords: Grid, GMPLS, optical network management, on-demand, multi-domain, resource provisioning, middleware

1. INTRODUCTION

A new generation of scientific applications is emerging that couples scientific instruments, data and high-end computing resources distributed on a global scale. Many of these applications have requirements such as determinism, shared data spaces and large data transfers, often achievable only through dedicated optical bandwidth. High capacity optical networking can satisfy bandwidth and latency requirements, but software tools and frameworks for end-to-end on-demand provisioning of network services need to be developed in coordination with other resources -CPU and storage-, and need to span multiple administrative and network technology domains.

The Phosphorus project -an IST FP6 project started during the last trimester of 2006- addresses some of the key technical challenges to enable on-demand end-to-end network services provisioning across multi-domain/multi-vendor research network test-beds on a European and worldwide scale. The Phosphorus network concept and test-bed make applications aware of their complete Grid and network resources, since the underlying network will be treated as first class Grid resource. This concept of integrating applications, middleware and transport networks under a new AAA architecture is based on three planes: Service Plane, NRPS Plane and the Control Plane. Therefore, Phosphorus enhances and demonstrates solutions that facilitate vertical and horizontal communication among applications middleware, existing Network Resource Provisioning Systems (NRPS), and the proposed Grid-GMPLS (G^2MPLS) Control Plane.

[1] Sergi.figuerola@i2cat.net; phone +34 93553 2515; fax +34 93 553 2520; i2cat.cat

Phosphorus's key features are to develop integration between application middleware and transport networks, based on the three planes that have been already presented and are with the following characteristics:

Service plane:

- Middleware extensions and APIs to expose network and Grid resources and make advance reservations.
- Policy mechanisms (AAA) for networks participating in a global hybrid network infrastructure, allowing both network resource owners and applications to have apart in the decision to allocate specific network resources.

Network Resource Provisioning plane:

- Adaptation of existing Network Resource Provisioning Systems to support the framework of the project.
- Implementation of interfaces between different NRPS to allow multi-domain interoperability with Phosphorus' resource reservation system.

Control plane:

- Enhancements of the GMPLS CP (G^{2}MPLS) to provide optical network resources as a first-class Grid resource.
- Interworking of GMPLS domains and G^{2}MPLS with NRPS-based domains (UCLP, DRAC and ARGON).

Moreover, the project will provide a set of studies to investigate and evaluate further the project outcomes, by studying resource management and job scheduling algorithms incorporating network-awareness, constraint based routing and advance reservation techniques, and developing a simulation environment supporting the Phosphorus network scenario.

The Phosphorus Grid middleware integrates also network reservation services, services for user-driven or application-driven set-up of execution environments with dedicated capabilities and performance. The MetaScheduling for Network and Grid Resources will be capable of orchestrating all kinds of Grid resources through negotiation with their respective local resource management systems. The Phosphorus architecture (Fig. 1) shows how the project is structured, since during the first phase of the project network resources are provided to the Grid middleware by means of a Network Service Plane (NSP), responsible for advance reservation and end-to-end network provisioning across various administrative domains managed by the different NRPSs involved in the project, such as User Controlled Light Path (UCLP), Dynamic Resource Allocation Controller (DRAC) and Allocation and Reservation in Grid-enabled Optical Networks (ARGON), or the basic GMPLS control plane.

The G^{2}MPLS control plane which will also interface the NRPS systems extends ASON/GMPLS in order to provide part of the functionalities related to the selection, co-allocation and maintenance of both Grid and network resources, by exposing upgraded interfaces at the UNI and E-NNI network reference points. The Grid extensions to GMPLS include the following procedures: Discovery and advertisement of Grid capabilities and resources, Service setup and maintenance and Service monitoring.

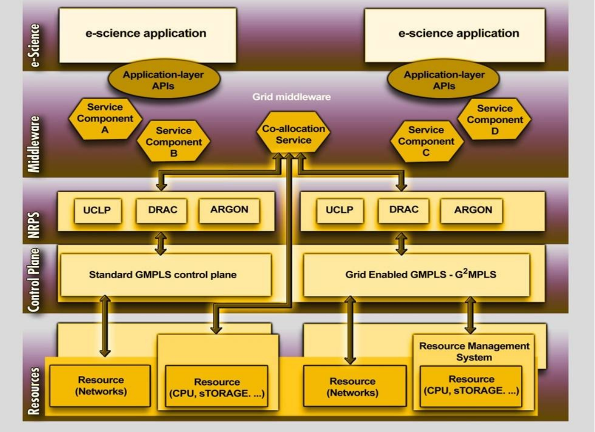


Fig. 1. *Phosphorus's overall Architecture of the two project phases*

The Phosphorus architecture allows the Grid applications to request by means of middleware a set of resources and its needed bandwidth and delay, depending on the application requirements (the project will implement four Grid applications: WISDOM -Wide In Silico Docking On Malasia-, KoDaVis -Collaborative Data Visualisation- and DDSS -Dstributed Data Storage System-), and in a single-step a network infrastructure will be provided for the service delivery.

This paper is organized as follows: Section 2 will describe the architecture developed during the first phase of the project for the NSP and the NRPS. An introduction to the G^{2}MPLS control plane is given in Section 3. In Section 4, the middleware and the AAA architecture developed are introduced. The supporting studies are presented on Section 5, while the whole test-bed of the Phosphorus project is described in Section 6, followed by conclusions in Section 7.

2. NETWORK PROVISIONING ARCHITECTURE

One of the aims of the Phosphorus project is related to the architecture (Fig. 2) which allows the provisioning of end-to-end connections across multi-domain and multi-vendor domains. This architecture, in conjunction with a new set of interfaces (North-Bound, South-Bound and East-West interfaces) follows a two-layer centralised approach[3]. It provides the interoperability in a seamless multi-domain environment between the Network Resources Provisioning Systems, and also its interoperability towards the middleware and the GMPLS control Plane.

The architectural basis of this management system is Service Oriented Architecture (SOA) [4]. The two layers integrating the architecture communicate themselves by means of Web Service interfaces, as described before, which have been made compatible with the following directives:

- Common programming model for definition of resource types across the different host platforms.
- Implementation of the Web Service Resource Framework 1.2 (WSRF) port type.
- Aggregation of Java bean classes into a single state model for WSRF.
- Compliance with Web Service Addressing 1.0 (WSA) and Service-Oriented Architecture Protocol 1.2 (SOAP[5]).
- Deployment within Java 2 Platform (J2EE).
- Apache Muse 2.2.0 main frame has been chosen to implement the web services interfaces[6].

The two-layer approach identifies the Network Service Plane, a global broker responsible for creating the end-to-end service connections through different NRPS systems by means of corresponding NRPS adapters, and the Network Resource Provisioning System which is a local domain controller providing services on a single domain. The NSP is responsible for dealing with the NRPSs in order to provide end-to-end paths, manage AAA issues, and keep track of the resource utilization and to coordinate the different actions done. Moreover, the NSP will be enhanced to increase its scalability to provide interoperability towards the G^{2}MPLS architecture developed along the project, and provide easier future interoperation with other international projects like GÉANT2 JRA3, EnLIGHTened or G-Lambda.

The key points and benefits provided by this architecture are:

- Its ability to create point to point, or point to multipoint connections using resources from several domains in a transparent way. The solution proposed speeds up the creation of complex connections with advance reservation features involving several systems by making them interoperable.
- Its simplification of AAA management: once the user is authenticated, he can use any of the services offered by the Network Service Plane. Moreover, his credentials are automatically translated to the local credentials of the systems involved in the service.
- The introduction of the Advance Reservations concept. Users or Grid applications are able to program fixed, deferrable or malleable resource reservations with one or more connections.

The interoperability performed by this architecture in conjunction with the new interfaces developed is understood as the capability to create advance reservations. An advance reservation is defined as a reservation of network resources in the future; this allows users to programme connections in the future that will be set up automatically as scheduled. These reservations are between two end points that can be located within the same or in a different domain at the network level, are defined by a specific bandwidth and a minimum delay.

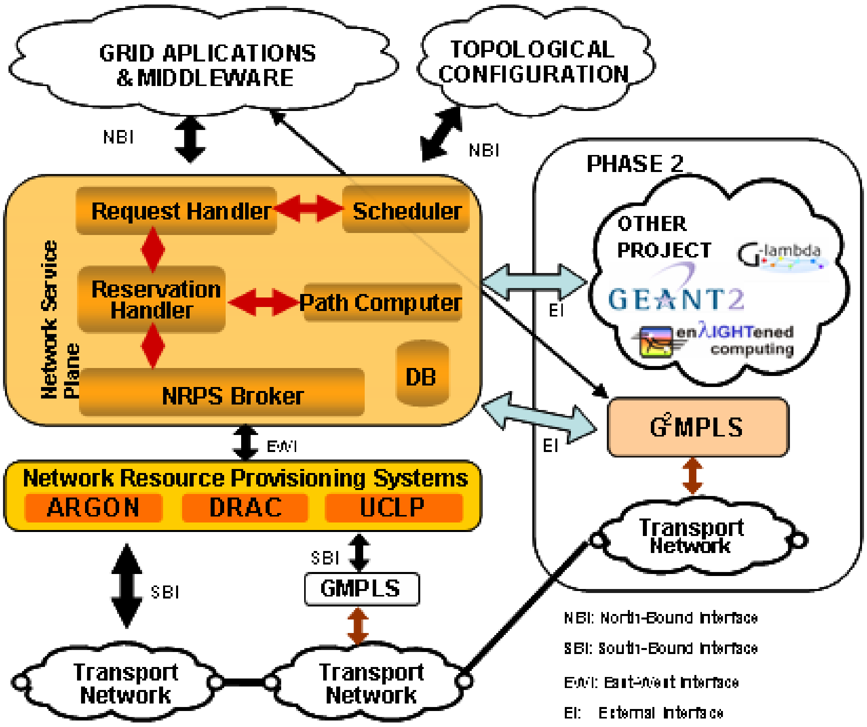
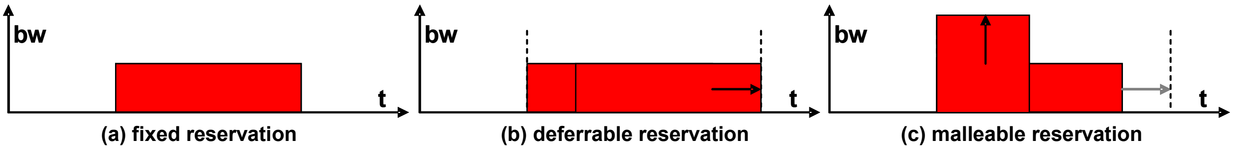


Fig. 2. Architecture of the NSP and NRPS with its interoperability interfaces towards the middleware and external systems

Three different kinds of Advance Reservations are envisaged for the NSP; however the first to be implemented is the Fixed Reservation type. These are:

- *Fixed reservations*: in this type of reservation the user has to specify the bandwidth along with the reservation start and end times.
- *Deferrable Reservations*: in this type of reservation the user has to specify the bandwidth, the duration of the connection and the earliest and the latest point in time when the connection will be useful. This type of reservation helps to find gaps to serve reservations at a fixed bit rate.
- *Malleable reservation*: in this type of reservation the user has to specify the maximum and minimum bandwidth allowed, the amount of information to be transmitted and the earliest and the latest point in time when the connection will be useful. This reservation provides a lot of flexibility to find a slot to serve reservations at a constant bit rate between the minimum and the maximum allowed throughput.



(a) fixed reservation (b) deferrable reservation (c) malleable reservation

Fig. 3. Different types of Advance Reservation to be considered within the scope of Phosphorous

The NRPSs involved in the project are systems already developed and provided by the project partners. These systems have been developed under the scope of European or International projects and a basic description of their aims is presented in Table 1.

The NSP has been built under a centralized approach. The main reason of this election is because the nature of each one of the NRPSs which were already developed, and therefore the project did not cover internal modifications of its architecture, since those systems did not have distributed-control capabilities or interfaces for that purpose.

However the chosen approach provides a centralized intelligence in a single NSP removing some of the problems of distributed systems, where each domain's topology information must be propagated among the different domains involved, what requires a control protocol for its propagation. Therefore, all the domain's knowledge is kept on a domain controller –broker principle– within the NSP. Therefore, and as each single domain does not have knowledge of the topology nor the resource reservation status in the other domains, without a centralized approach advance reservation functionalities would not be able to be performed among them.

Table 1. Basic definition of the NRPSs systems involved within the Phosphorus project

Network Resource Provisioning Systems (NRPS)	
ARGON	The Allocation and Reservation in Grid-enabled Optic Networks system was developed to manage resources of advanced network equipment as it is present in the German VIOLA test-bed. The advance reservation service of ARGON is able to operate on the GMPLS as well as on the MPLS level. It guarantees a certain QoS for applications for the requested time interval. This feature enables a Meta-Scheduling Service to seamlessly integrate the network resources into a Grid environment.
DRAC	The Dynamic Resource Allocation Controller system was developed by NORTEL and it is a commercial-grade network abstraction and mediation middleware platform, acting as an agent for network clients (users, applications, compute resource managers) to negotiate and reserve appropriate network resources on their behalf. DRAC uses client's QoS requirements and pre-defined policies to negotiate end-to-end connectivity across heterogeneous in support of just-in-time or scheduled computing workflows.
UCLP	The User Controlled LightPaths system was developed by CRC, Inocybe, i2CAT and UofO under the CANARIE support. It provides a network virtualization framework upon which communities of users can build their own services or applications. Articulated Private Networks (APN) are presented as the first services. The APN can be considered as a next generation Virtual Private Network where a user can create a complex, multi-domain topology by binding together network resources, time slices, switching nodes and virtual/real routing services.

The NRPS domains only publish their border endpoints (physical interfaces) to the NSP. Intra-domain connections (connections within domains) are hidden to the NSP and maintained by the corresponding NRPS. The namespace of endpoints is chosen in such a way that the corresponding domain can be identified from the name of the endpoint. Inter-domain connections (connections between domains) are administrated within the NSP.

This centralized implemented approach keeps a virtual image of the different domains and the links / connection points among them in the NSP. It also keeps track of the occupation and resource reservations of the different domains and their interconnections. Moreover, this central entity, the NSP, performs a nexus role between the Grid Middleware and the different NRPS, completing the whole picture of the management architecture from Grid Applications to Transport Networks. Apart from the above mentioned approach, the distributed approach is currently being studied within the Joint Research Activity 3 (JRA3) of the GÉANT2 project. It is therefore expected that Phosphorus will study its outcomes and implement and re-use it if needed, since both projects will collaborate on their research developments. During the second phase of the project, it is expected that a distributed architecture based on distributed instances of the NSP will be built.

As domains are represented only by their endpoints and the intra-domain topology is hidden, the path computer (Fig. 3) of the NSP cannot calculate intra-domain paths. Rather, it will generate a list of endpoints defining an inter-domain path and will receive from the NRPS the following results when an intra-domain path is requested: 0-path available, 1-source occupied, 2-detination occupied, 3-source and destination occupied, 4-no path between source and destination.

Regarding the authorization and accounting mechanisms, the NSP itself does not provide complex authorization or accounting mechanisms. Rather it forwards attributes that are contained in the incoming message from the middleware to the involved NRPS adapters and vice versa. The authorization process is in the scope of each NRPS and the middleware authorization tickets are transparently communicated through the NSP. It is assumed that each domain has its own policy and attribute database. The NRPS adapter may map the global attributes to local ones or local user accounts. In case the global and local attributes are identical this mapping reduces to the identity function.

2.1 System interfaces

One of the main developments carried out under this topic have been the system interfaces, which allow the system interoperability between the different layers. The interfaces developed (Fig. 2) are the:

- *Northbound interface (NBI)*: It receives the reservation requests from the GRID Middleware.

- *East-West interface (EWI)*: It is in charge of the communication between NRPSs by means of the NSP.
- *Southbound interface (SBI)*: It communicates the NRPSs and the lower layers (GMPLS or transport layer).
- *External Interface (EI)*: It provides interoperability between the NSP and the G^{2}MPLS CP or other projects.
- *Topological interface (TI)*: It is used to indicate to the NSP the inter-domain resources (NRPSs endpoints links).

The operation of all these interfaces together allows the request of advance reservation network resources from the middleware to the NRPS systems. Starting from a bottom-up approach, there is the SBI. As mentioned previously, this interface communicates the NRPS with the lower layers and the GMPLS control plane. The NRPS that have their own domain below use proprietary protocols or the well known CLI and SNMP, however when the domain is a standard GMPLS CP, we need a specific set of interfaces that have been developed for the project, these are the so called Thin NRPS and GMPLS-WS (Fig. 4) (also called GMPLS driver). The Thin NRPS is a network resource provisioning system for domains with a GMPLS control plane. It provides a reservation web service, which is used by the NSP to reserve, create and delete network connections via the GMPLS driver. The Thin NRPS does not need a specific NRPS adapter (which is composed by a Reservation Web Service and a Topology Web Service), because it can use the common NRPS adapter interface of all the Phosphorus' NRPS systems.

The GMPLS driver acts as an interface between an NRPS and the GMPLS control plane. It offers a general web service, which is used to create, delete and monitor paths for different GMPLS implementations. It is important to point out that the topic presented deals with a standard GMPLS CP, which can not provides advance reservation functionalities; therefore, this is provided by the Thin NRPS. However, this is under the constraint that resources in a GMPLS domain are always available, unless another reservation is overlapping. Later on we will introduce the G^{2}MPLS CP, which will perform advance reservation functionalities, and therefore a new interface directly towards the NSP will be developed.

The Thin NRPS also provides a notification receiver interface, which enables the GMPLS driver to send update information about endpoints and paths. Moreover, it acts as a client to the topology manager web service of the NSP (e.g. it will add a domain and provide information about endpoints) and as a client to the GMPLS driver web service for creating and deleting network connections, and for obtaining information about endpoints and existing connections.

The GMPLS driver provides the following services and notifications described in Table 2.

Table 2. Basic definition of the GMPLS driver services and notifications

Path creation service	Creates a point-to-point path between two endpoints specified by TNA addresses.
Path termination service	Tears down a point-to-point path which has been set up by an NRPS.
Path monitoring service	Provides status information about the specified path.
Path discovery service	Retrieves any established point-to-point connection in the controlled network. It is needed to refresh the NRPS view on the network in case of losses or reboots.
Endpoint discovery Service	Retrieves information about any endpoint in the controlled network. It is needed to deliver the available endpoints to the NRPS during system initialization and to refresh the NRPS view on the available endpoints in case of memory losses through system failures or reboots.
Registration service	Registers the NRPS to receive messages from the web service in case of path status changes or endpoint changes.
Path delete notification	Informs all registered and authenticated systems, if a path is no longer available.
Endpoint update notification	Informs all registered and authenticated systems when endpoints have been removed or added to the controlled network domain.

The GMPLS services are offered via a web service to the clients. A client could be an NRPS or for test purposes a user, who can access the services via a Java application. All Information about paths, endpoints and devices is kept in a MySQL database, which is only accessed by the core component of the GMPLS driver. It can be administered through any database front-end like phpMyAdmin. GMPLS operations are initiated and controlled by vendor specific modules. Focusing on the Thin NRPS, it provides the following functions: Handles reservation requests from the NSP; checks, if the reservation can be granted and keeps track of the reservations in a database. As the Thin NRPS currently only gets endpoint information from the underlying GMPLS driver, it has no knowledge about internal links and their usage.

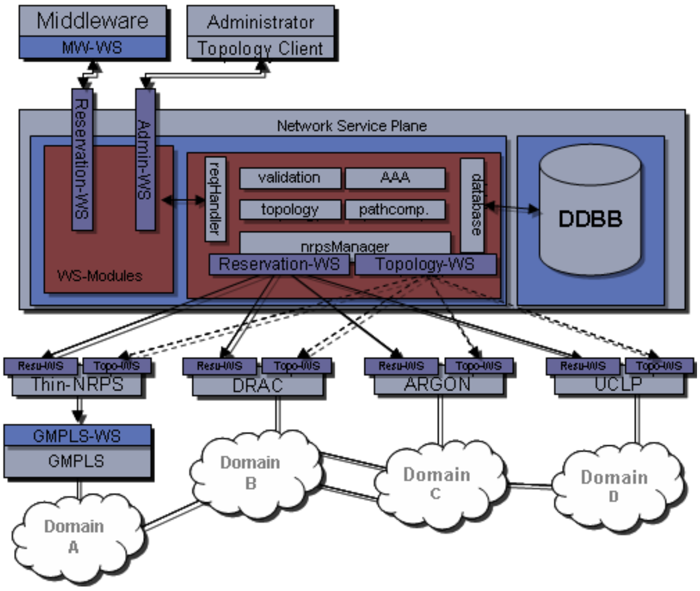


Fig. 4. Detailed internal architecture of the NSP and NRPS layers and their interactions with the different domains

This means, that in case of advance reservation, no checks for availability of bandwidth on the internal links can be performed. Therefore, as already commented, the Thin NRPS will only check if endpoints are available and if there are conflicts concerning the usage of endpoints within overlapping reservations. It also schedules creation and termination of network connections using the GMPLS driver, and registers its domain at the NSP domain manager, retrieves user and border endpoints from the GMPLS driver, and forwards border endpoints to the NSP domain manager. Finally, it handles endpoint updates from the GMPLS driver and informs the NSP domain manager. As presented, the Thin NRPS has a limited functionality: only fixed reservations are currently supported, because no topology information about links and their usage is available

The normal NRPS Adapter is the adaptation layer between the NSP and the NRPS itself. It contains a common communication part for all the NRPSs that consist of the required Web Services to receive the calls to the operations at the NRPS level. Each adapter translates the incoming calls to the invocations implemented by the NRPS, and redirects them to the underlying system.

The Web Service present at the adapter consists mainly of the operations implemented by the Reservation-WS, which offers the following functions: Availability request, Reservation request, Cancel reservation, Status request, Bind request, Activation request, Complete Job, Cancel Job and Retrieve features

Moreover, the adapter implements a registration service in order to register the Domain and its endpoints automatically in NRPS start time. This service provides the NSP with all the required information about the Domain, the NRPS and the physical resources. It updates periodically the NSP with the information about the topology controlled by the NRPS.

The same Reservation-WS is present at the NSP towards the application layer, forming the Northbound Interface. It contains the same operations and communication functionalities, but it is implemented from the NSP point of view.

The NRPS adapters, jointly with the lower part of the NSP, constitute the so called East-West (E-W) interface. It implements the communication between NRPSs and the NSP-NRPS, enabling their interoperability, since the system is centralized and the NSP is the one in charge of the communication with all the entities.

2.2 System workflow for topology and advance reservation establishment

The system workflows for topology and reservation services are based on a set of actions over the NSP and NRPS architecture (Fig. 4). The inter-domain topologies between domains –between endpoints of the domains– and their characteristics will be entered in the system by means of an API (topology client) developed for that purpose. Once the domain resources are included within the system, it is updated automatically by each one of the NRPS Adapters as explained before. This topology information is completed with the inter-domain links, introduced manually through the topology client. This process can not be done automatically since NRPSs act autonomously and do not have knowledge of the resources of the other NRPSs, since the interconnection issues are driven by the NSP in a centralized way.

Regarding the reservation flow, the NSP receives a reservation request and acts as a broker for that request. Therefore, the Middleware-WS (MW-WS, Fig. 4) sends to the Reservation-WS of the NSP an end-to-end reservation request, once it is validated, the system identifies the user's rights towards the different network resources available. Afterward, the Topology and the Path Computer service compute the path by cheeking its availability to the data base system, and define the inter-domain path. Once the path is selected, and the domains to be passed through are identified, the nrpsManaged manages to complete the reservation interaction between the domains, by splitting the reservation with all the domains involved on the process, and therefore, depending on the availability of the intra-domain links it will be successful or cancelled.

3. THE GRID-ENABLED GMPLS CONTROL PLANE: G2MPLS

One of the main goals and technical domains of the Phosphorus project concerns the architectural definition, software design and prototypal implementation of the Grid-enabled GMPLS (G^2MPLS) Network Control Plane, as a enhancement of the ASON/GMPLS[7] Control Plane architecture that implements the concept of Grid Network Services (GNS). In the PHOSPHORUS framework, GNS is a service that allows the provisioning of network and Grid resources in a single-step, through a set of seamlessly integrated procedures.

G^2MPLS[8] results in a more powerful Control Plane solution than the standard ASON/GMPLS, because it will comply with the needs for enhanced network and Grid services required by network "power" users/applications (i.e. the Grids). Nevertheless, G^2MPLS is not conceived to be an application-specific architecture, and it will support any kind of endpoint applications by providing network "legacy" ASON/GMPLS transport services and procedures. This compliance fosters the possible integration of Grids in operational/commercial networks, by overcoming the limitation of Grids operating on dedicated, stand-alone network infrastructures.

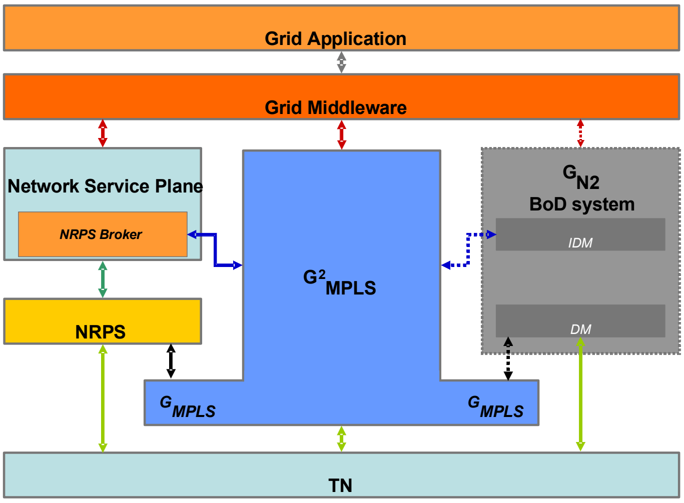


Fig. 5. Positioning of the G^2MPLS Control Plane in the Phosphorus framework

The main rationale behind this new G2MPLS architecture is made of different points: firstly, it is a dual approach with regards to grid brokers working with and configuring network resources. Secondly, grid nodes can be modelled as network nodes with node-level Grid resources to be advertised and configured, and this is a native task for GMPLS. Thirdly, it can also inherit "useful" GMPLS native features, e.g. crank back and recovery

- The G2MPLS NCP can bring to an innovation in this field, because of its

- Faster dynamics for service setup in the same time-scale of the NCP ones
- Availability of well-established procedures for traffic engineering, resiliency and crankback
- Uniform interface (G.OUNI) for the Grid-user to trigger Grid & network transactions not natively dependent on a specific Grid middleware.

Basically, G2MPLS extends the ASON/GMPLS architectures in order to provide part of the functionalities related to the selection, co-allocation and maintenance of both Grid and network resources, by exposing upgraded interfaces at the UNI and E-NNI network reference points (i.e. G.OUNI and G.E-NNI). The Grid extensions to GMPLS include the following procedures:

- *Discovery and advertisement* of Grid capabilities and resources of the participating Grid sites (Vsites);
- *Service setup*
    - *Coordination* with those parts of the Grid middleware needed for the local configuration of the Grid job (i.e. the local job scheduler in particular)
    - *Configuration* of the deriving network connections among the Vsites participating to the Grid job. Depending on the Grid application capabilities and requirements, the network attachment endpoints could be specified or not: e.g. in case of distributed computing and visualization, the network attachment endpoints could be declared, in case of distributed storage they could not. Moreover, depending on the administrative partitioning of the network, the configured network service might span multiple domains.
    - *Management of resiliency* for the installed network services and, depending on the specification of the network attachment endpoints, possible escalation to the Grid middleware components that could be responsible for check-pointing and recovering the whole job (e.g. by changing the involved Vsite or pausing/postponing the job, etc.)
    - *Advanced reservations*[2] of Grid and network resources, aimed to guarantee connection availability at job execution time by providing users with a priori information about start, wait and completion times.
- *Service monitoring*
    - Retrieving of the status of a Grid job and the related network connections.

The G^{2}MPLS Network Control Plane brings an innovation in the field of co-allocation of Grid and network resources, because of its faster dynamics for service setup, adoption of well-established procedures for traffic engineering, resiliency and crankback and uniform interfaces for the Grid user to trigger Grid & network transactions.

From a *user's perspective*, G^{2}MPLS enables a real node-to-node deployment of on-demand Grid services. Part of the middleware functionalities related to selection, co-allocation and maintenance of both Grid and network resources are provided through the new G.OUNI. From a *network operator perspective*, G^{2}MPLS enables the integration of Grids and automated network control plane technologies in real operational and commercial networks. In Phosphorus, the G^{2}MPLS will interface to a number of existing Network Resources Provisioning Systems (UCLP, ARGON, DRAC), and to the Inter-Domain Management system defined by the GÉANT2 project, in order to provide as seamless as possible operations for Grid resources provisioning.

According to its architectural positioning in the Phosphorus framework (Fig. 5), the G^{2}MPLS has a number of service interfaces: when it deals with network and Grid resources seamlessly (i.e. in the "G^{2}MPLS Integrated Model"), it needs to directly interact with the grid middleware; when it deals with non-G^{2}MPLS domains (i.e. connecting a requested set of Grid resources under an NRPS domain), it might need to interact with the NSP (e.g. in the "G^{2}MPLS Overlay Model"). As already commented, the G^{2}MPLS provides advance reservation functionalities. The interface used for its interoperability with the NSP will be based on an enhancement of the Thin NRPS adapter. Once the core developments are completed, a new interface will be developed in order to provide full interoperability of G^{2}MPLS and the GÉANT2 Bandwidth-on-Demand system.

[2] As defined in OGF GRAAP-WG, advance reservation is a possibly limited or restricted delegation of a particular resource capability over a defined time interval, obtained by the requester from the resource owner through a negotiation process.

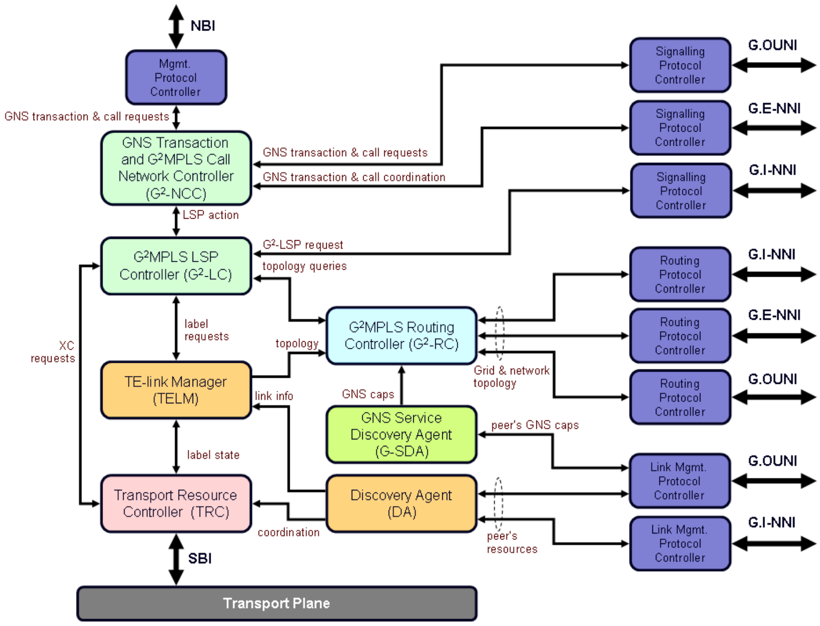

NBI
Mgmt. Protocol Controller
GNS transaction & call requests
GNS Transaction and G²MPLS Call Network Controller (G²-NCC)
GNS transaction & call requests
GNS transaction & call coordination
LSP action
G²MPLS LSP Controller (G²-LC)
G²-LSP request
topology queries
XC requests
label requests
topology
G²MPLS Routing Controller (G²-RC)
Grid & network topology
TE-link Manager (TELM)
link info
GNS caps
GNS Service Discovery Agent (G-SDA)
peer's GNS caps
label state
Transport Resource Controller (TRC)
coordination
Discovery Agent (DA)
peer's resources
SBI
Transport Plane
Signalling Protocol Controller
G.OUNI
Signalling Protocol Controller
G.E-NNI
Signalling Protocol Controller
G.I-NNI
Routing Protocol Controller
G.I-NNI
Routing Protocol Controller
G.E-NNI
Routing Protocol Controller
G.OUNI
Link Mgmt. Protocol Controller
G.OUNI
Link Mgmt. Protocol Controller
G.I-NNI

Fig. 6. Functional architecture of a G²MPLS controller

The list of modules and their roles/functionalities are as described on table 3.

Table 3. Basic definition of the GMPLS driver services and notifications

GNS Transaction and G²MPLS Call Controller (**G2-CC**)	- control and management of both GNS Transactions and the related G²MPLS Calls - 2 types: calling/called party G2-CC (**G2-CCC**), network G2-CC (**G2-NCC**)
G²MPLS LSP Controller (**G2-LC**)	- management of each Label Switched Path (LSP) that is part of a G²MPLS call
G²MPLS Routing Controller (**G2-RC**)	- stores an updated topology view of Grid and network resource - uses the topology for the computation of paths upon a request from G2-LC
TE-link Manager (**TEM**)	- selection and allocation/de-allocation of resources (<Data-link, label>) in TE-link - management of the TE-link status and bundling information for topology purposes
GNS Service Discovery Agent (**G-SDA**)	- discovers Grid and network capabilities between a Vsite attached to the G.OUNI (client side) and the G²MPLS NCP (network side)
Discovery Agent (**DA**)	- discovers network resources (i.e. by correlation and verification) in the G²MPLS NCP
Transport Resource Controller (**TRC**)	- abstracting the technology specific details of the switching resources for NCP
Protocol Controllers (**PC**)	- maps contents and actions of interfaces into objects and messages of the G²MPLS protocol instances. 3 types: Signaling (**SPC**) G².RSVP-TE; Routing (**RPC**) G².OSPF-TE; Link management (**LPC**) LMP (G.LMP?)

4. THE MIDDLEWARE AND THE AAA ARCHITECTURE

Distributed Grid applications and workflows often involve transfers of huge amounts of data. In state-of-the-art best-effort networks, this results in unpredictable performance of the application or workflow. To resolve this issue and deliver predictable performance, it is required that all Grid resources including the network can be reserved and allocated in a coordinated way. The MetaScheduling Service uses WS-Agreement[9] to negotiate and agree upon the reservations with the underlying local resource management systems, e.q. schedulers or batch queuing systems, WS-Agreement is a proposed standard for the creation and monitoring of Service Level Agreements developed in the Open Grid Forum. Reservations thus are Service Level Agreement which may include besides the guaranties on a service, i.e. the availability of a certain resource with the required QoS properties, also penalties that might be applied if a Service Level Agreement is violated by either of the parties. Based on this technology, the MetaScheduling Service will be capable of orchestrating all kinds of Grid resources by negotiating with their respective local resource management systems[10]. These are typically batch-systems for compute resources and network resource provisioning systems as developed in Phosphorus for the network. A main goal of the project is to offer reservation and allocation functionality to network and other Grid resources via the G^{2}MPLS. Driven by the demand of the Phosphorus demonstration applications, the MetaScheduling Service will interface to UNICORE 6. However, support for Globus Toolkit 4 is also envisaged.

The initial middleware layer in Phosphorus is based on the infrastructure and developments used in the German VIOLA project: UNICORE and the MetaScheduling Service[11]. In the first phase of the project UNICORE will be used as Grid middleware stack, in the second phase GT4 will be included as additional middleware stack. The major effort changing the middleware is the modification of the MetaScheduling Service to become a UNICORE Server. Furthermore, extensions to the protocol between the UNICORE Server and the Target Systems are needed, in order to allow the negotiation protocol of the MetaScheduling Service flow along with the other UNICORE protocols. This implies changes in the Target System Interface allowing the negotiation protocol for the Target Systems and the adapters respectively. The extension of the existing protocol between the different UNICORE Servers located in different administrative domains (Usites) will be one of the project outcomes. In order to make the new reservation and co-allocation capabilities available to the Grid users, the UNICORE client will also be extended. Along with the job requirements for compute resources, the user can specify the communication requirements. In the second phase of the project, resource selection will be further automated by using resource annotation.

Another of the main objectives within the project is the definition of the architecture and subsequently design, implement and test an AAA[12] authorisation to support policy based on-demand network resource provisioning across different administrative domains, which are governed by Network Resource Provisioning Systems (NRPS) representing these domains. The research and development in this topic will address AAA related issues at all functional network planes: the Network Provisioning Plane, the Control Plane and Data-Forwarding Plane.

The project will respond to the practical need for policy based and customer/project centric complex resource provisioning[4]. The development is based on the experience of developing AuthZ services for Network resources provisioning and Grid applications[13] . The mechanisms implemented contribute towards the standardisation activity at IETF and OGF on developing standard AuthZ architecture and components for Grid based applications, and token-based policy enforcement mechanisms in network resource provisioning. This will ensure future compatibility and interoperability of the proposed solutions by using a wide use case base, and by coordinating development with different Grid and network related projects such as Internet2, GÉANT2, EGEE, and Globus.

4.1 AAA Operational models

The Generic AAA Authorization Framework[14] (GAAA-AuthZ) as described in RFC2904 and GFD.038 defines three basic authorization sequences: the push sequence where the user first requests an authorization decision from an authorization service, obtains some proof which s/he will next present to the resource in a service request; the pull sequence where the user directly submits a request to the resource, and then request an authorisation decision from the authorization service, (the agent sequence where the user and the resource delegate handling a service request to an agent). After the service request is authorized, the agent will provision the service.

The AAA operational model for multi-domain network resource allocation and provisioning can be described using one or more basic AAA sequences. These sequences may incorporate interactions with resource reservation and/or scheduling systems, and (virtual) organizations providing user attributes among others.

The basic provisioning sequences (Fig.7) are:

- *Polling sequence (P).* The user client polls all individual network domains to make a reservation.
- *Relay (R) or hop-by-hop reservation*. The user client contacts only the local network domain/provider and each consecutive domain provides a path to the next domain.
- *Agent (A)*. Sequence in which the user delegates network provisioning negotiation to an Agent that will perform all necessary negotiations with all involved domains.

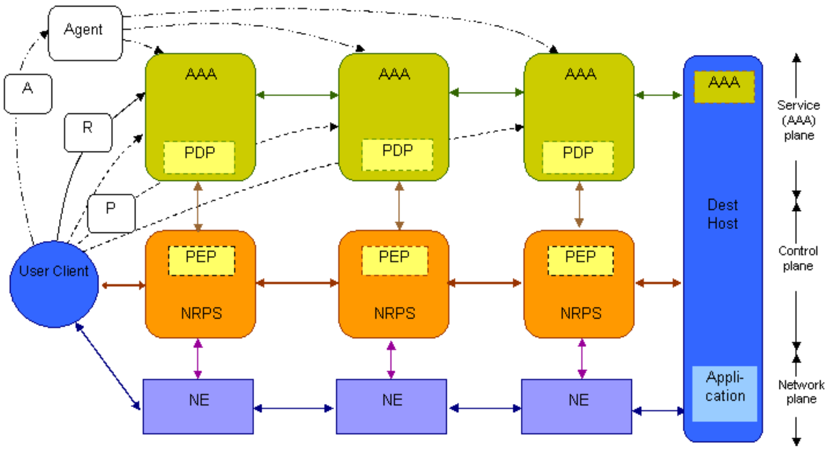


Fig. 7. Components involved in multi-domain network resource provisioning and the basic provisioning sequences

Figure 7 illustrates major interacting components in the multi-domain Optical Network Resource Provisioning:

- A User/Requestor and a Destination/target service or application.
- Multiple Network elements (NE) (related to the Network plane).
- Network Resource Provisioning Systems (NRPS) (typically related to the Control plane).
- AAA service controlling access to the domain- related resources that can also operate own communication infrastructure, including components Policy Enforcement Point (PEP), Policy Decision Point (PDP), Policy Authority Point (PAP).
- Token Validation Service (TVS) that allows efficient authorisation decision enforcement when accessing reserved resources.

Access to the resource or service is controlled by the NRPS and protected by the AAA service that enforces resource access control policy by placing a Policy Enforcement Point (PEP) gateway at the NRPS. Depending on the basic GAAA-AuthZ sequence (push, pull or agent) [2, 3], the requestor can send a resource access request to the resource or service (which in our case are represented by NRPS) or an AuthZ decision request to the designated AAA server which in this case will act as a Policy Decision Point (PDP). The PDP identifies the applicable policy or policy set and retrieves them from the Policy Authority Point (PAP), collects the required context information and evaluates the request against the policy.

The user can provide as much (or as little) information about the subject/requestor, the resource or the action as it decides necessary according to the implemented authorisation model and resource access control policies. The Policy Decision Point (PDP), which is the part of the AAA AuthZ service, evaluates the request and makes the decision whether to grant access or not. Based on a positive AuthZ decision (in one domain) the AuthZ ticket (AuthzTicket) can be generated by the PDP or PEP and communicated to the next domain where it may be processed as a security context or policy evaluation environment. In order to get access to the reserved resources the requestor needs to present the reservation credentials that can be in a form of an AuthZ ticket or a token (AuthzTicket or AuthzToken) which will be evaluated by the PEP to grant access to the reserved network elements or the resource. In more complex provisioning scenarios the token or credential validation function may be outsourced to the TVS service that can additionally support a interdomain trust management infrastructure for off-band token and key distribution between the PEP-NRPS and AAA services. The TVS as special GAAA-AuthZ component to support token-based enforcement is described below.

Using AuthZ tickets during the reservation stage for communicating the interdomain AuthZ context is essential to ensure effective decision making. At the service access/consumption stage the reserved resource may be simply identified by the reservation ID created as a result of the successful reservation process. To avoid significant policy enforcement overhead when handing service reservation context, the ticket can be cached by an NRPS or a TVS in each domain and referred to with the AuthzToken that can be much smaller and even communicated in-band. At the resource PEP it can be compared with the cached AuthzTicket, AuthZ session context or reservation context and will allow local PEP/resource access control decisions. Such an access control enforcement model is being implemented in the Token Based Network (TBN) [15].

4.2 AAA Enforcement mechanisms

The TBN network resource provisioning model and token based policy enforcement require specific mechanisms to cryptographically bind data-flows to their origin or target applications and corresponding enforcement mechanisms when data flows are sent over heterogeneous network infrastructure. Important component of such mechanisms is the distribution of the key material between domains that allows creating an overlay virtual infrastructure of the provisioned network. This work is also undertaken within an IETF ForCES based router extension that acts as a gateway between a general purpose IP network (e.g. campus network) and a dedicated GMPLS network, enforcing and providing tokens at both the IP and GMPLS side. This allows the integration of token mechanisms inside GMPLS control plane layer using RFC2750 policy data object as a base. The development of the GAAA-AuthZ middleware related components is being coordinated with other European projects such as EGEE and GEANT2 JRA3 and JRA5. This activity will also continuously cooperate with the Internet2 OSCARS and DRAGON projects. The goal of such cooperation is to achieve compatibility of the Phosphorus AAA infrastructure and those developed and currently used in other projects, namely GN2, eduGAIN, Internet2 Shibboleth, and EGEE VOMS attribute/federation services and access control infrastructure.

5. SUPPORTING STUDIES UNDER THE PHOSPHORUS PROJECT

The purpose of this work is the design and evaluation of innovative architectures and algorithms to efficiently manage optical Grid infrastructures. This work is essential because it allows to effectively evaluating interesting alternatives with short turn-around times. The main effort has been allocated to building a facilitating simulation toolkit for these experiments. The software makes abstraction of the control plane and management plane implementation details, allowing assessment of optical Grid infrastructure dimensioning and resource management algorithms. During the project, novel management and control details will be incorporated, to refine the modelling and to guide control plane and service plane design.

An initial study focused on obtaining realistic models describing typical Grid job arrival patterns and execution times. In a first part we *successfully gathered data at different aggregation stages* (Grid level vs single cluster sites) from existing, real life Grid infrastructure. The second portion of this study focused on proposing *candidate models for synthetic job generation* (i.e. job inter-arrival times and execution times). We subsequently judged their usefulness by verifying how well the models could be fit to produce traces similar to the measurements discussed before, and came to two main conclusions[16]. First, job *inter-arrival times* on the observed Grid level can be successfully modelled by a Poisson process, but on the Grid site level the long range dependency needs to be taken into account and Hyper-Exponential, Markov-modulated Poisson process or Pareto-Exponential models need to be used. Second, for the *job execution times*, we achieved the most satisfactory results with a (3 phase) hyper-exponential process.

A number of routing approaches have been proposed that take into consideration physical layer characteristics (e.g. chromatic and polarization mode dispersion, crosstalk), to provide optimum resource utilization and offer improved QoS. Accurate analytical models that evaluate physical layer degradations have been developed and integrated into the routing procedure to allow optimized routing performance. Simulation results (see Fig. 8) revealed a noteworthy improvement in the network performance for a wide range of design parameters indicating the need to upgrade current routing approaches to include optical constraints.

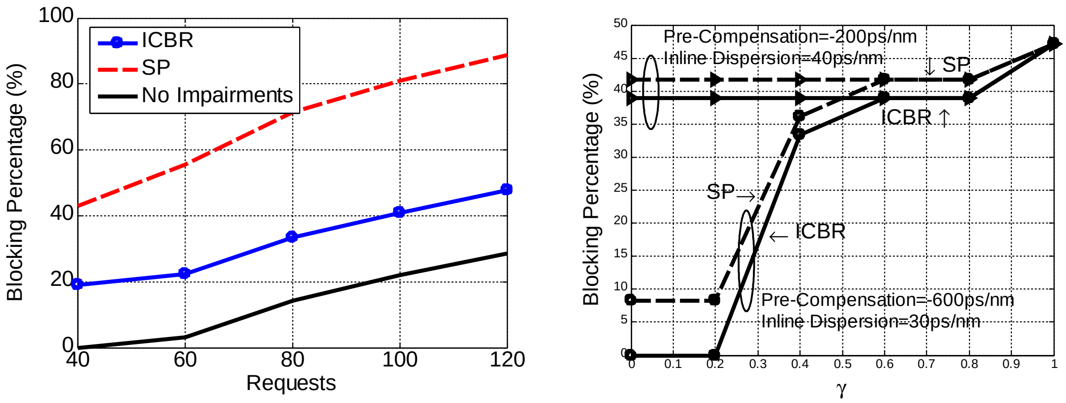


Figure 8: Impairment Constraint Based Routing (ICBR) reduces blocking (SP = shortest path)

Furthermore, we developed enhanced *anycast routing* algorithms to provide coordination between the submitted jobs and the optical network components and resources capable of processing the jobs (Fig. 9). Research has also focused on the concept of advance reservations to orchestrate a set of resources and services (ultimately allowing job workflows in the Grid), in which we have proposed and analyzed both a meta-scheduling environment and a protocol suitable for advance reservations. Additional effort was focused on defining a framework and algorithms for providing fairness and QoS to Grid end-users (Fig. 9)

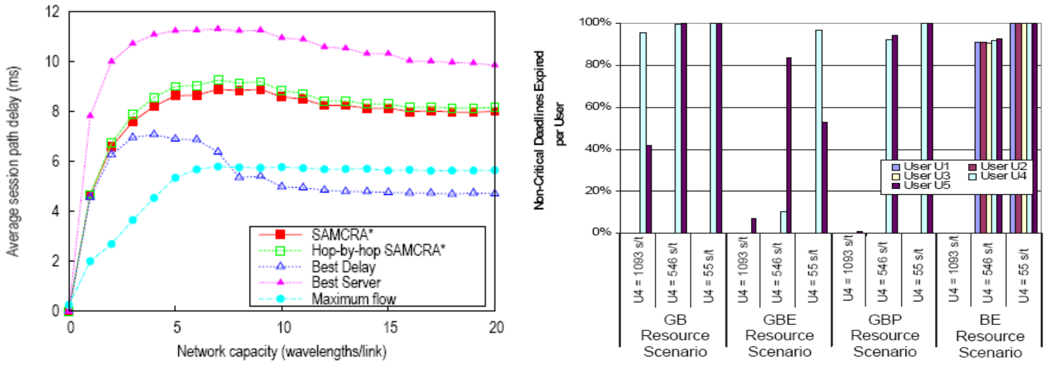


Figure 9: Performance of anycast routing and results on fairness scheme (users 1-3 have guaranteed service, 4-5 best effort)

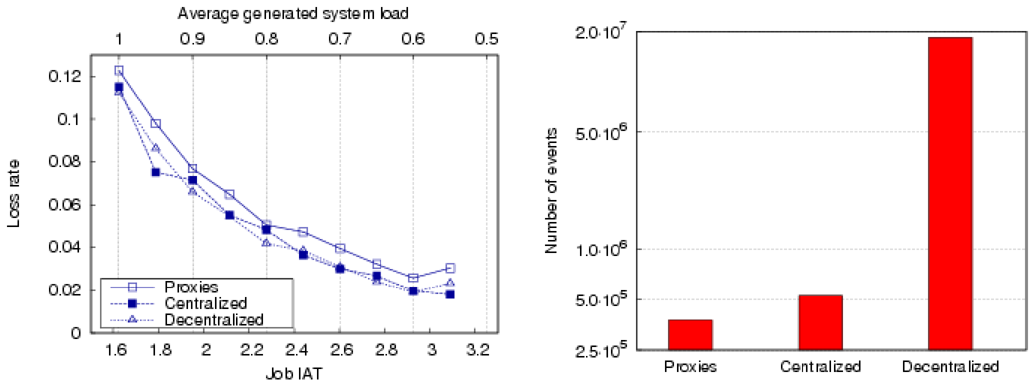


Figure 10: Anycast proxy architecture reduces control traffic and maintains loss performance

Finally, multi-domain issues deal with the heterogeneity of the local networks that compose the Grid environment, and generally causing scalability issues for the control plane. Our proposal for an anycast proxy infrastructure for inter-domain job allocation can significantly reduce the control plane overhead, while keeping job loss rates comparable with those associated to the other strategies (Fig. 10). As such, it offers a scalable solution for a growing network with an increasing number of clients and computational resources.

6. THE INTERNATIONAL PHOSPHORUS TEST-BED

One of the main objectives of Phosphorus is to demonstrate an application development environment that will validate the new services and the infrastructure developed in the project. To achieve this objective a distributed test-bed is being built, in which the project developments will be incrementally introduced. The test-bed constitutes a real environment in which the project outcome is demonstrated with a set of real scientific applications in a set of real-life scenarios. The test-bed is constructed from communications equipment (optical switches, TDM switches, Gigabit Ethernet switches, transmission equipment) and advanced GRID resources (like computing nodes, visualisation and storage resources) provided by the partners. The communications equipment is the platform of the project's developments and implementations in order to allow applications to utilise the functionality of an advanced optical network and assure seamless cooperation of various test-bed equipment and technologies.

The Phosphorus test-bed consists of multiple local test-beds located in several sites in Europe and outside Europe. The local test-beds will be interconnected with other local test-beds using multiple international optical networks (including GÉANT2 and others) to create a global, heterogeneous test-bed comprising several technologies and administrative domains. The structure of the test-bed is based on requirements of all the Phosphorus activities.

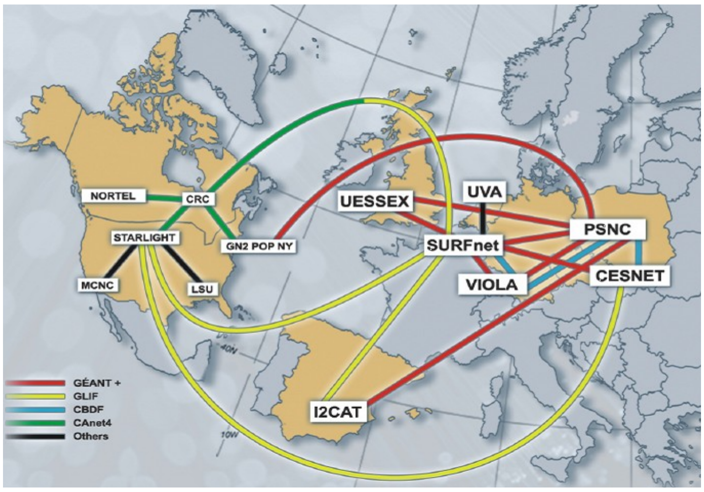


Fig. 11. Phosphorus test-bed interconnection data links.

7. CONCLUSIONS

In this paper we presented an overview of the whole research that is currently being done under the FP6 EU funded Integrated Project Phosphorus that is addressing some of the key technical challenges to enable on-demand, end-to-end network services provisioning across multiple domains in a seamless and efficient way for e-science applications. In these infrastructures to be build, both the Grid and optical network resources will coexist under the same control plane for provisioning and recovery purposes, spanning seamlessly across different control and management domains. This paper explains an overview of the architecture built to allow different NRPS systems provision e2e paths across different domains (included a standard GMPLS CP) with advance reservation functionalities, some procedures of the new architecture for a Grid-enabled GMPLS control plane (G^{2}MPLS), an overview of the AAA architecture to be developed and the Phosphorus middleware used for integration purposes. Also some of the simulation studies performed to validate current developments. However, one of the main challenges the project will face during 2008 is the integration of all of these different research topics in order to make them interoperable under the same Phosphorus umbrella. The first project prototype results are going to be publicly shown at the Supercomputing 2007 workshop, to be held on November in

Reno, US. During this workshop it will be demonstrated the establishment of an e2e connection in a multi-domain, multi-vendor scenario, where different NRPS are involved.

AKNOWLEDGEMENTS

This work is supported by the FP6 EU funded Integrated project PHOSPHORUS (Lambda User Controlled Infrastructure for European Research) IST-034115.We would like to specially acknowledge the contributions provided by Thomas Eickermann (Forschungszentrum Jülich GmbH, Germany), Leon Gommans, Cees de Laat (Universiteit van Amsterdam), Marcin Garstka (Poznan supercomputing and networking Center, Poland) and Dimitra Simeonidou (University of Essex, England), and also Christian de Waal, Alexander Willner, Ferdinand Hommes Bram Peeters, Antoon Prins, Bert Andree, Michel Savoie, Hanxi Zhang, Scott Campbell, John MacAuley, Joan Antoni Garcia and Angel Sanchez for their inputs on the NSP architecture, and the long list of all the other members of the Phosphorus project for their great commitment on the project research and developments

REFERENCES

[1]S. Figuerola, *et al:* Sample manuscript showing specifications and style. IST-PHOSPHORUS Deliverable 1.1, February 2007.

[2]N.Ciulli, *et al*.: The Grid-GMPLS Control Plane architecture, IST-PHOSPHORUS Deliverable 2.1, 2.2. Feb. 2007.

[3]D. Gavalas, D. Greenwood, M. Ghanbari, M. O'Mahony, *A Hybrid Centralised - Distributed Network Management Architecture*, *Proceedings of the EEE International Symposium on Computers and Communications, 1999.*

[4]F.L. Verdi, R. Duarte, F.C. de Lacerda, E. Cardozo, M. Magalhaes and E. Madeira. “Provisioning and Management of Inter-domain Connections in Optical Networks: A Service Oriented Architecture-Based Approach”. 10th IEEE/IFIP Network Operations and Management Symposium (NOMS 2006). April 3-7, 2006, Vancouver, Canada

[5]W3C Recommendations: SOAP, *http://www.w3.org/TR/soap*

[6]Muse, *http://ws.apache.org/muse*

[7]ITU-T G.8080/Y.1304 Recommendations, Architecture for the Automatic Switched Optical Network (ASON), 2001

[8]G. Markidis, A. Tzanakaki, N. Ciulli, G. Carrozzo, D. Simeonidou, R. Nejabati, G. Zervas. EU Integrated Project PHOSPHORUS: Grid-GMPLS Control Plane for the Support of Grid Network Services. ICTON 2007

[9]Heiko Ludwig; Toshiyuki Nakata; Oliver Wäldrich; Philipp Wieder; Wolfgang Ziegler: Reliable Orchestration of Resources using WS-Agreement. The 2006 International Conference on High Performance Computing and Communications , (HPCC06), Munich Germany, September 2006, Springer LNCS 4208, S. 753 – 762

[10]Christoph Barz, Thomas Eickermann, Markus Pilz, Oliver Wäldrich, Lidia Westphal, Wolfgang Ziegler: Co-Allocating Compute and Network Resources - Bandwidth on Demand in the VIOLA Testbed. CoreGRID Symposium, Rennes, Frankreich, August 2007, CoreGRID Series, Towards Next Generation Grids, S. 193 - 202.

[11]Web Service Agreement (WS-Agreement). GFD.107 proposed recommendation, available at <http://www.ogf.org/documents/GFD.107.pdf>.

[12]C. de Laat, G. Gross, L. Gommans, J. Vollbrecht and D. Spence, *Generic AAA Architecture*, RFC2903, 2000.

[13]Demchenko Y., L. Gommans, C. de Laat, “Using SAML and XACML for Complex Authorisation Scenarios in Dynamic Resource Provisioning”, in Proc. The Second International Conference on Availability, Reliability and Security (ARES 2007), Vienna, Austria, April 10-13, 2007. IEEE Computer Society, ISBN: 0-7695-2775-2, pp. 254.

[14]J. Vollbrecht, P. Calhoun, S. Farrell, L. Gommans, G. Gross, B. de Bruijn, C. de Laat, M. Holdrege and D. Spence, *AAA Authorization Framework*, RFC2904, 2000.

[15]"Token-based authorization of connection oriented network resources", by Leon Gommans, Franco Travostino, John Vollbrecht, Cees de Laat, and Robert Meijer, in Proceedings of GRIDNETS, San Jose, CA, USA, Oct 2004.

[16]K. Christodoulopoulos, M. Varvarigos, C. Develder, M. De Leenheer, B. Dhoedt, Job Demand Models for Optical Grid Research, Proc. of the 11th Conference on Optical Network Design and Modelling (ONDM), Athens, Greece, May 2007