

Network and Grid: the EGEE use-case

Mathieu Goutelle and Jean-Paul Gautier

CNRS UREC

Univ. Pierre & Marie Curie (Tour 66/65 - 4^e étage)
4, place Jussieu – 75252 Paris Cedex 05 – FRANCE
{mathieu.goutelle, jean-paul.gautier}@urec.cnrs.fr

Abstract In this article we propose a description of the networking activities within a grid project: EGEE. These activities are focused on the access to the provided services (resource reservation) and on the operational constraints of such an architecture (operational interface with networks, SLA installation, monitoring). Even if the EGEE project may have some peculiarities, numerous issues are not EGEE specific, particularly the ones about network. They prefigure a possible evolution of future network usage, not only by EGEE or other distributed computing projects, but by any potential user.

Keywords grid, operational interface, Service Level Agreement (SLA), resource reservation, monitoring, end-to-end

Glossary

- BAR** Bandwidth Allocation and Reservation — Software developed by JRA4 to be used as an interface with network services.
- CERN** European Organisation for Nuclear Research (<http://www.cern.ch>)
- ENOC** EGEE Network Operation Centre — Team in charge of the operational issues related to network in EGEE.
- GEANT** The GEANT network — Interconnection network of the European NRENS. See GEANT2.
- GEANT2** European project with aim of building the next generation of interconnection network between European NRENS (<http://www.geant2.net/>).
- GGF** Global Grid Forum — GGF is the community of users, developers, and vendors leading the global standardisation effort for grid computing (<http://www.ggf.org>).
- g-Lite** Grid *middleware* developed within EGEE (<http://www.glite.org/>).
- JRA** Joint Research Activity
- L-NSAP** Local Network Service Access Point — Interface provided by JRA4 to configure end-site network equipments.
- LHC** Large Hadron Collider — The LHC is a very high energy particles accelerator currently being built at CERN (<http://lhc.web.cern.ch>).
- MPLS** Multi-Protocol Label Switching — Provides path-like mechanisms into packet switched network like IP network.

NM-WG Network Measurement Working Group — GGF working group. It identifies network metrics (aka characteristics) useful to grid applications and middleware, and develops standard mechanisms to describe and publish these characteristics to the Grid (<http://nmwg.internet2.edu>).

NOC Network Operation Centre — Team normally in charge of managing and monitoring a network infrastructure.

NPM Network Performance Monitoring — Tool developed by JRA4 providing access to network monitoring data.

NREN National Research and Educational Network

NSAP Network Service Access Point — Interface provided by an NREN to access and use a network service ; speaks with BAR.

PERT Performance Enhancement and Response Team — Networking experts team created as part of GEANT2 to assist users in troubleshooting network performance issues.

SA Service Activity

SLA Service Level Agreement — Agreement “signed” between a customer and a service provider and specifying the quality level of the service.

SLR Service Level Requirement — Description of the user’s requirements.

SLS Service Level Specification — Description of the technical characteristics of a service to meet the requirements put in an SLR.

TNLC Technical Network Liaison Committee — EGEE internal committee serving as an interface between EGEE NRENS for all technical questions.

1 Introduction

The EGEE project (*Enabling Grid for E-sciencE*¹) is a project financed by the European Commission. It brings together experts from 70 industrial and academic partners from over 27 countries to exploit grid expertise generated by many EU, national and international Grid projects to date. The aim of EGEE is to build a production quality grid infrastructure able to provide a consistent, robust and secure service to users. We propose here to show with the help of the EGEE example how operational grids will impact the networking activities with the perspective of end-to-end services.

After a short presentation of EGEE in section 2, section 3 will explain what the networking activities in EGEE are. Then, the consequences on networks will be presented in section 4, especially on networks (local, metropolitan, regional, *etc.*) that are between the National Research and Educational Network (NREN) and the resource centre (computing, storage). Then, we will conclude in section 5.

2 The EGEE project

The EGEE project has started in April 2004 for two two-year periods and followed a number of computer grids projects, especially DataGrid². The aim is

¹ <http://www.eu-egee.org>

² <http://eu-datagrid.web.cern.ch/eu-datagrid/>

to propose to users a huge computing and storage resource through an easy interface and services for authentication, jobs submissions, data management (upload, replication, download, *etc.*). The resources are provided at the national level, by the project partners involved. It is up to the project to aggregate these resources into a seamless infrastructure. Up to now, EGEE puts together more than 160 sites from 40 countries in Europe and around the world, providing more than 17,000 CPU and more than 5 petabytes³ storage. The middleware used in EGEE, g-Lite, integrates many components from existing middlewares (LCG, Alien, EDG, *etc.*) and proposes many improvements and new functionalities.

First foreseen to meet the requirements of high energy physics experiments at CERN and analysis of the huge amount of data produced by the Large Hadron Collider (LHC), the grid now accommodates many other applications: biomedical applications (tomographic emissions simulation, protein sequences analysis, *etc.*), Earth Science (seismic data analysis, earth observation, climatology, *etc.*), computational chemistry...

To interconnect the numerous resources centres, the grid depends a lot on the underlying network, its quality and the provided services. The grid uses the NRENS and GEANT, the pan-european network interconnecting the NRENS, to build the “EGEE network”, the overlay network interconnecting the EGEE nodes. This network is the one tackled in EGEE, involving more than 40 NRENS in Europe and around the world, with their heterogeneity in structure and connectivity. The “EGEE network” will take advantages of the services provided and developed in the context of these networks. To achieve this, two networking activities are contributing to EGEE. The first one, *Network Resource Provision* (SA2), deals with operational issues. Development of internal network services for EGEE are insured by the JRA4 activity, *Development of Network Services*.

3 Networking activities in EGEE

From a general point of view, our aim is to make grid people aware of the network as a grid **resource**, just like computing and storage. Currently, the network remains relatively “transparent” and is therefore not really considered a part of what should be managed in such an infrastructure. There are two main reasons to explain this:

- Firstly, the network is an **external resource** for the grid since it is provided outside of EGEE and more generally of the grid, by the NRENS to the resource centres;
- Secondly, means are missing to **automatically manage this resource** through software interfaces, particularly services that are provided by the network apart from the simple IP connectivity. It prevents the building of grid resource management architectures (including the network of course) from happening.

³ One petabyte is equal to 1024 terabyte (about 1×10^{12} bytes).

Nevertheless, the grid is very dependent on the networks and their services, without a real control on this resource. That is why we are working in close collaboration with GEANT and NRENS (see. figure 1) to anticipate on the future network services, *e.g.* the ones developed within GEANT2, and to be ready to use them as soon as they become available. For example, the collaboration takes place for the definition of common interfaces and operational models. As an interface with network providers, the networking activities should gather the requirements of the other EGEE activities, which are network users, such as the middleware or applications.

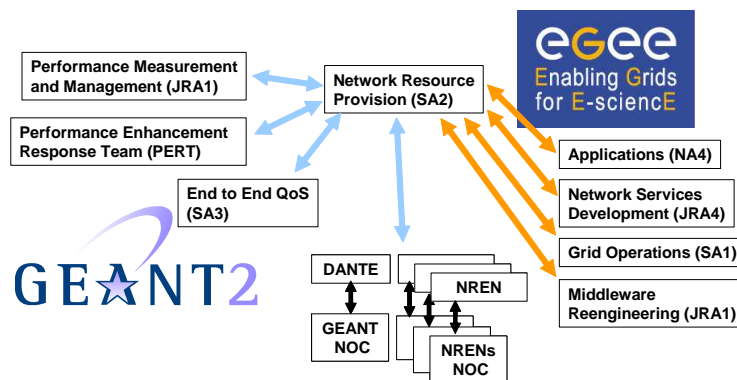


Figure 1. Interactions between EGEE and the networks

3.1 JRA4 activity: “Network Service Development”

The JRA4 activity is a software development activity to be integrated in the grid middleware. This activity is subdivided into two sub-activities, Bandwidth Allocation and Reservation (BAR) and Network Performance Monitoring (NPM).

NPM has been developed to access seamlessly monitoring and performance data from the “EGEE network”. This data is provided by different domains and monitoring frameworks [2]. The JRA4 approach was to adopt a standard interface to access and give access to the data (see figure 2). The choice has been made to use the work of the Global Grid Forum (GGF) and particularly of the Network Measurement Working Group (NM-WG).

A standardised interface allows you to access any existing measurement tools, provided that it uses the same interface to publish data. It is already the case for end-to-end tools developed within DataGrid⁴, but also for the PerfSonar framework⁵ for backbone measurements inside GEANT and NRENS. On the other hand, many tools (diagnostic tools, middleware components, user) can have access to

⁴ <http://ccwp7.in2p3.fr/>

⁵ <http://www.geant2.net/server/show/nav.754>

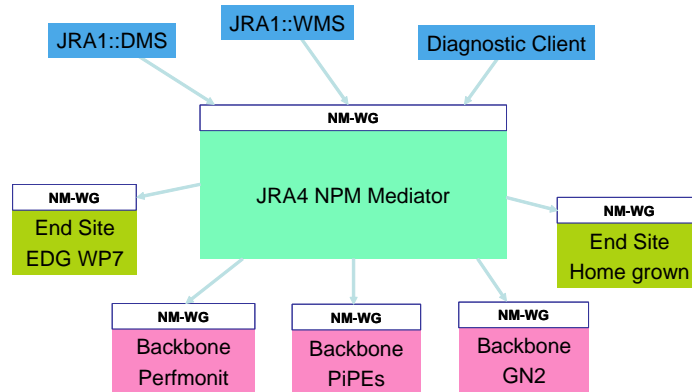


Figure 2. The NPM architecture

this data through that interface in an unified manner (*i.e.* without choosing which is the right tool for its request) because NPM aggregates the gathered data and publishes it always through the same interface.

BAR, the other JRA4 activity, should allow grid applications to reserve network services and to use them [1]. BAR should be seen as an interface with mechanisms being put in the networks. In other words, BAR will never configure any network equipment, but will contact a service, the Network Service Access Point (NSAP), under the total control of the NRENS. It is then crucial for BAR to strengthen the collaboration with networks and to obtain a strong cooperation from NRENS, especially from the JRA3 and SA3 activities of the GEANT2 project. To solve this configuration issue at the extremities of the path inside the local network of the sites, it is foreseen to deploy a Local Network Service Access Point (L-NSAP) in the resource centres.

The complete architecture of the mechanisms involved in BAR are depicted in figure 3: a request coming to BAR is first validated (parameters check). It is then send to the other components, firstly the NSAP of the first adjacent network which will check the feasibility of the requested reservation at the network level. To achieve this, it will speak with all the NSAP in the chain of domains involved, and with the remote BAR instance. The two BAR instances ask also their respective L-NSAP to check the feasibility of the request at the level of the local networks of the two sites. If the request is rejected by either one of the NSAPs or the L-NSAPs, the request is rejected by BAR which will ask the components which have accepted the request to cancel it. Otherwise, the network configuration is applied in each domain by the associated NSAP and L-NSAP.

At the moment, the only network service available to BAR is the *Premium* IP service because it is the most widely deployed within NRENS⁶. The EGEE applica-

⁶ About 50% of NRENS have deployed *Premium* IP according to a survey done at the beginning of 2005 (<https://edms.cern.ch/document/477861>).

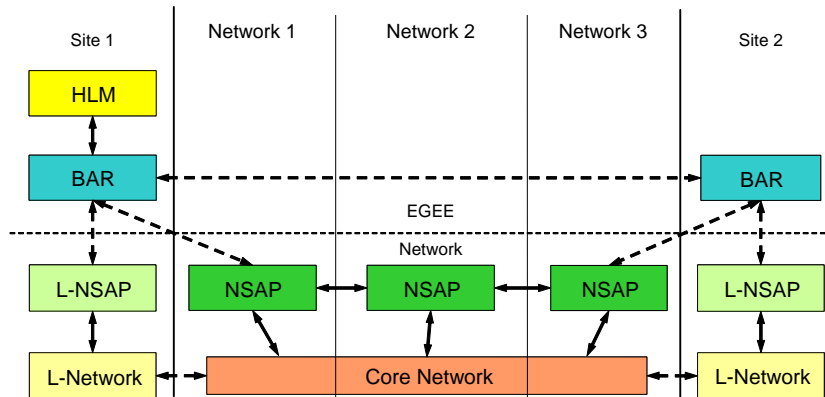


Figure 3. Architecture of the BAR-network interface. HLM means *Higher Layer Middleware*.

tions will be able to use other network services, as soon as they become available at the NSAP level.

JRA4 is also in charge of assessing the impact of the likely IPv6 deployment on a grid like EGEE [3]. This work has identified the advantages that IPv6 could bring to the infrastructure and to the applications, but also the issues, the peculiarities in the migration and coexistence process between IPv4 and IPv6 and the integration of IPv6 in the middleware.

3.2 SA2 activity: “Network Resource Provision”

If JRA4 could be seen as the “software interface” between the grid and the network, SA2 provides the “operational interface”. This relationship between NRENS and GEANT has been for example formalised by the creation of the Technical Network Liaison Committee (TNLC). This committee fosters technical collaboration and sharing of information between the two communities.

The operational interface is also the place where information about incidents, maintenance operations and troubles faced by users on the EGEE network can be exchanged. To ease this workflow, procedures have been defined [6] to exchange trouble tickets between the trouble ticket management systems of the NRENS Network Operation Centres (NOCs) and the one of EGEE⁷. These information flows, essential to build a network-aware operational grid, are bi-directional: one grid user can report a network related problem to the EGEE user support. This ticket is then assigned to the EGEE Network Operation Centre (ENOC), which will then try to pinpoint where the problem can be (site, machine, *etc.*), if necessary identify the concerned network and contact it. It will then follow the evolution of the problem resolution, taking in charge the questions from both the contacted NOC and the user. Because of the huge number of users and involved networks,

⁷ <http://www.ggus.org/>

it seems unrealistic to make them talk directly. That is why the creation of the ENOC seems necessary, as a single point of contact for all the parties involved (resource centres, NOCs and users).

As an operational interface between the grid and the network, SA2 should also examine as often as possible if the network requirements of the applications are satisfied. This means first that the specific needs of the applications should be collected, derived from applications use cases. From these requirements, a small number of user-oriented service classes has been identified [7], aggregating similar usage (real-time interactions or transfers, “bulk” transfer and virtual private network). These technology-independent service classes are then associated to underlying network services [4]: for example a service like *Premium* IP⁸ could be suitable for real-time interactions messages. This translation could be done either in the middleware (the request to the NSAP is done in terms of service) or in the NSAP (the request to the NSAP is done in terms of characteristics). The second one is the preferred one because it evolves naturally with the arrival of new network services.

If the usage of network services spreads, the careful management of the resources by the providers becomes mandatory. To achieve this, to agree on a SLA between the customer and the provider seems necessary. To be able to provide a service to the EGEE users, there is indeed a need to describe its characteristics technically and administratively. The SLA is the outcome of a joint work between the service provider and the user. The process starts with the definition of the user requirements in a Service Level Requirement (SLR). It goes on with the technical description in a Service Level Specification (SLS) of a service which meets these needs. The SLA is after all only the formalisation of an agreement between the two parties on the use of the service described in the SLR.

In this field, the SA2 task has been to propose SLA templates [5] sufficiently generic to accommodate the wide variety of networks involved in EGEE⁹ but also precise enough to be usable. These templates consist of two parts. The first one contains administrative information: technical contact, duration of the reservation, troubleshooting procedures, *etc.* The second one describe the technical characteristics of the reservation: involved source and destination, guarantees, traffic limits, excess treatment policies, *etc.* The templates take into account the various situations likely to come up in the EGEE network in terms of deployment and service availability. Many scenarios have been elaborated to overcome the network heterogeneity (partial implementation or lack of a service in one domain involved in the path). Each SLA is defined for a particular domain (a NREN *e.g.*). The per-domain SLAs are then combined in order to provide an end-to-end SLA describing the service offered to the application (see figure 4).

SLAs are new for both the NRENS and the grid community. For this reason and to acquire experience in this domain and validate the SLA establishment

⁸ Service providing marked IP packets a strict priority on the other ones, and so giving them the minimum queueing delay.

⁹ We have estimated to about 40 the number of networks involved in the interconnection of EGEE sites.

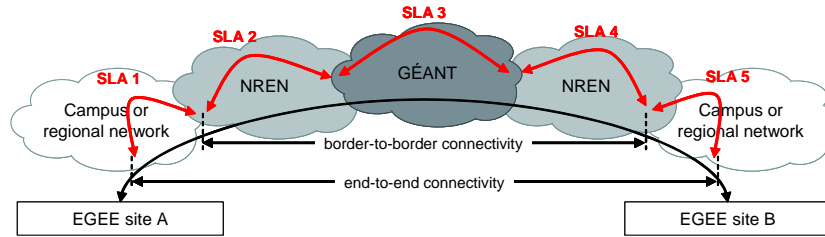


Figure 4. Per-domain SLAs combined to build the end-to-end SLA

procedures, SA2 has conducted an experiment (cf. figure 5). The aim was to use the *Premium* IP service between pairs of sites in France, Greece and Russia. The sites have been chosen to tackle the various possible situations: a site directly connected to its NREN or the presence of some intermediate campus or metropolitan networks. This experiment has led to the definition of SLAs describing the provided service in each domain participating in the experiment and then to the definition of an end-to-end SLA as described above.

Two main lessons concerning service usages can be drawn from these tests. The first one is the difficulty to set up the service. Once set up, it works as we expected. But, prior to use the service, the preliminary work, both technical and administrative, remains a long and heavy task. The technical part setup needs a lot of synchronisation, configuration and testing to be sure that everything works. The administrative part, mainly authorisation and SLAs, needs also a lot of effort, mainly because people are not used to the process of making such agreements. This is possible for seldom or long-term use of a service, but this does not scale for a deployment in an infrastructure the size of EGEE. The second conclusion is that an application can not use yet a service, like we did in these tests. There are still some missing functionalities in the middleware to take completely advantages of BAR. We did circumvent this issue but our solution doesn't use BAR and is therefore not generic enough and sufficiently abstracted from a user's point of view.

The SLA monitoring issue remains also unsolved. The solution being set up will depend on the policies chosen regarding the respect of the SLA guarantees and the available monitoring granularity: it can be a **live** monitoring of the characteristics during the whole duration of the reservation or an **a posteriori** verification to assess if an application received the requested guarantees **on average** on the use duration.

3.3 Network service use case

Let us now explain what could be the mechanism used by an application to use a network service between two sites. We consider that we are still in the current situation where the manual configuration of the backbone equipments

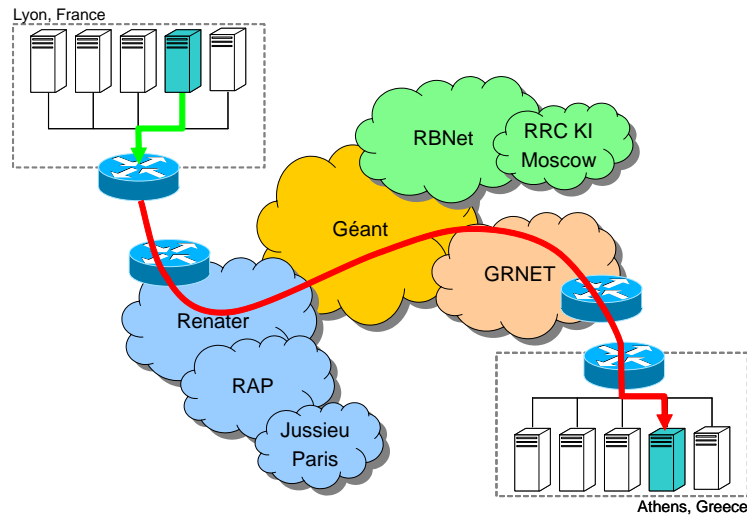


Figure 5. SLA setup experiment

imposes a lead time of at least one working day. In a near future¹⁰, even if this configuration becomes automatic, the user could still have to take into account that a new configuration is applied to backbone equipments only periodically (twice a day for example).

The service usage needs a SLA established in advance between the different actors: between the sites and all the involved networks on the data path. These SLAs could be established in advance, **for a long-term duration** (one or two year *e.g.*). They contain among other things much technical information on the quality level the user can expect from the service.

To use the service, the request initiator¹¹ describes his needs in qualitative terms, by choosing one service class (real-time interaction, real-time flow, “bulk” transfer, *etc.*), and in quantitative terms (duration, bandwidth, *etc.*). Upon the submission of this *Service Request*, the middleware will check it (authentication, authorisation and resource management). If the request is valid, the middleware will speak to the network through BAR to obtain a **mid-term** reservation (a couple of weeks or months). By combining the per-domain SLAs, it is possible to build an end-to-end SLA, describing the expected characteristics from the application’s point of view. These characteristics (bandwidth, delay, *etc.*) should be monitored by the EGEE middleware to verify that the service complies with the request or not. Of course, the provider does the same thing, but only on its domain. If it happens that the guarantees were not met, the ENOC might have to get involved to troubleshoot the problem and, if needed, ask for an action from one of the NOCs. Another solution for the ENOC is to contact the GEANT

¹⁰ before the end of GEANT2, *i.e.* August 2008.

¹¹ a user, an application or a middleware component.

Performance Enhancement and Response Team (PERT): this team of experts has been constituted within GEANT2 precisely to solve end-to-end performance problems.

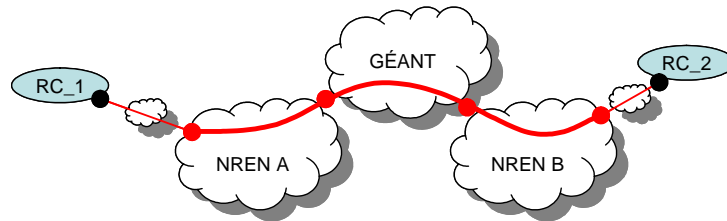


Figure 6. First step of a resource reservation (*Service Request*)

At this time (see figure 6), only the backbone network is configured to accept the traffic based on the possible senders and receivers (IP subnet *e.g.*). It is still not possible to know (in the general case) the exact source and destination because they will be chosen later on at job resource allocation time. To avoid intentional or unintentional unauthorised use of the service, the site equipments are still configured to block the utilisation of the service (path extremities on figure 6).

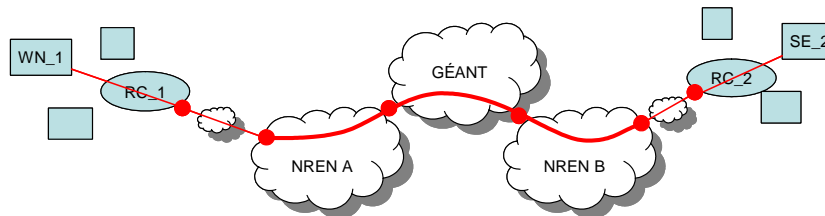


Figure 7. Second step of a resource reservation (*Service Activation*)

At the arrival time of the job on the cluster node chosen by the middleware, the reservation still needs to be activated during the execution of the job (see figure 7). This **short-term** activation step (*Service Activation*) then allows to configure the border equipments of the site to authorise the sender (the machine chosen for this particular grid job) to inject traffic in the reserved service. The resource reserved can be used by another job either at the same time provided that there is enough available capacity within the associated Service Request, or after the end of the first job as long as the reservation is still valid.

This two-step mechanism seems necessary for two reasons:

- It allows to get rid of the difference in reservation duration granularity between the network (reservation of many days) and the grid (reservation of a few hours). The scalability is also improved because many “grid” reservations can be aggregated in one “network” reservation and further managed by the middleware ;
- It allows to solve the issue of the flow authorisation at the network equipments level. The configuration of these equipments can only be done when the end nodes are known and ready to use the reservation.

4 Consequences on networks

Building an infrastructure of distributed computing like a grid has consequences on the networks. The most obvious (and the most expected by administrators) is an increase in the usage of the links, even if not necessarily automatic or significant. However, not all grid applications imply huge data transfers. In this case¹² anyway, private infrastructure are envisaged [9] first to guarantee the service to the experiments and also not to disturb the normal network usage. The public Internet will only be used to access the grid, to consult data and to execute specific applications.

Among the unexpected consequences, there is first the direct ones on the intermediate networks up to the NRENS : local, metropolitan, regional networks depending on the adopted hierarchy in each country (some may be not present). The first consequence is the need for operational management. The critical point is not really the end-to-end high availability of the networks: the distributed nature of the grid accommodates quite well with the unreachability of a site. The critical point is more about the incident escalation and resolution procedures. The availability of the information in a centralised manner is necessary to the grid. This has a great impact on networks if they do not already have an operational system in place.

The second significant consequence of the arrival of a grid node inside a network is the need for an end-to-end service, either in terms of provisioning (available bandwidth), configuration (special traffic authorisation) or quality. Reaching the “end-to-end” is a goal but a true challenge if we want to go beyond IP best effort and provide quality of service using IP or Multi-Protocol Label Switching (MPLS) tunnels for example. To evaluate this difficulty, it is only necessary to limit the request to the availability of a non-guaranteed 1 Gb/s bandwidth between two storage nodes! If this is certainly possible between two major resource centres in Europe, it becomes in most of the cases a real challenge (multiple contacts, configuration synchronisation, SLAs establishment, *etc.*), without mentioning all the possible technical dead-ends (inappropriate equipment, incompatibility, resource unavailable, *etc.*). Anyway, the generalisation of such service on a large scale seems unrealistic if most of the preliminary work still requires manual interventions.

¹² This is among others the case on the LHC project.

This is why a long term vision on the network evolution seems to foster the setup of mechanisms to offer the user more control over the network resource. It does not concern only the grid user, who may only be considered as a pioneer, but all the users. A part of the network resource management should be put progressively in the user's hands, who will be able to request something to his provider. It could accept or reject the request based on agreed criteria between the two parties, on the request's characteristics and on the network's conditions. This should be dynamic, on-demand and above all, should provide a multi-domain end-to-end service. Some progress is expected, either from the GEANT2 project which is preparing this kind of services for the NRENS in the GEANT2-SA3¹³ and GEANT2-JRA3¹⁴ activities, or through initiatives of vendors, telecommunication carriers and access providers like the IPSphere forum¹⁵ [10].

The gathering of most of the Internet market actors leads us to believe in the future of such mechanisms. They are all the more promising because they rely on well-known network technologies, such as *Premium* IP for GEANT2-SA3 or MPLS¹⁶ for GEANT2-JRA3 and IPSphere while remaining technologically neutral. Finally, they "simply" propose first to add reservation interfaces for the user on top of existing and well-proven mechanisms and secondly to tackle the end-to-end issue, *i.e.* to make different domains speak together to finalise the reservation whatever the data path could be. This is here where the choice of well-proven technologies and concept takes all its meaning: it is far easier to build an end-to-end solution if the underlying mechanisms are already deployed in the involved domains and if you don't rely on a particular network technology but can accommodate many existing ones provided they offer the same higher level characteristics.

The network interoperability to provide a service seems actually to represent progress not only for the users but the networks too. It indeed eases and speeds up the processes of service setup and allows to provide users (or customers) with advanced services without any future complication or large scale deployment impossibility.

5 Conclusion

The EGEE project is an ambitious one: its aim is to provide the grid users the ability to use one of the biggest distributed computing infrastructure with a production quality. It is however realistic: the will to rely strongly on existing software and to concentrate efforts on the necessary functionalities to achieve its goal contributed to the successful first two years of the project.

¹³ <http://www.geant2.net/server/show/nav.893>

¹⁴ <http://www.geant2.net/server/show/nav.756>

¹⁵ <http://www.ipsphereforum.org/>

¹⁶ More exactly GMPLS, its extension to new transport protocols at the link layer, like lightpath. See [8].

In this context of production quality, the networking activities bring more in terms of services and operational workflows than in terms of quality and high availability which are already common properties of today's networks. This is why these activities concentrate their efforts primarily on the operational interface with networks and on the monitoring and secondly on the end-to-end services (SLA and reservation). They work for the future, simply because some necessary components are still missing, particularly in the field of on-demand service reservation. Nevertheless, the work of present-day projects or initiatives for the network evolution leads us to think that these missing parts will soon be available and not only in the world of NRENS but in the commodity Internet too.

References

1. JRA4 activity. Specification of interfaces for bandwidth reservation service. Technical Report DJRA4.1, EGEE – Enabling Grid for E-sciencE, November 2004. <https://edms.cern.ch/document/501154>.
2. JRA4 activity. Definition of standardized network measurement query/response interfaces. Technical Report DJRA4.2, EGEE – Enabling Grid for E-sciencE, January 2005. <https://edms.cern.ch/document/533215>.
3. JRA4 activity. Report on implications of IPv6 usage for EGEE Grid. Technical Report DJRA4.3, EGEE – Enabling Grid for E-sciencE, September 2005. <https://edms.cern.ch/document/603955>.
4. SA2 activity. Initial requirements aggregation model, specification of services as SLSs on the networks. Technical Report MSA2.2, EGEE – Enabling Grid for E-sciencE, January 2005. <https://edms.cern.ch/document/509455>.
5. SA2 activity. Institution of SLAs and appropriate policies. Technical Report DSA2.2, EGEE – Enabling Grid for E-sciencE, March 2005. <https://edms.cern.ch/document/565447>.
6. SA2 activity. Operational interface between EGEE and GEANT/NRENS. Technical Report MSA2.3, EGEE – Enabling Grid for E-sciencE, March 2005. <https://edms.cern.ch/document/565449>.
7. SA2 and JRA4 activities. Survey of pilot applications requirements on networks, initial SLRs and service classes. Technical Report DSA2.1, EGEE – Enabling Grid for E-sciencE, October 2004. <https://edms.cern.ch/document/495204>.
8. L. Berger. Generalized Multi-Protocol Label Switching (GMPLS) Signaling Functional Description, January 2003. RFC3471.
9. David Foster, Erik-Jan Bos, Edoardo Martelli, and Paolo Moroni. LHC high-level network architecture. Technical report, LHC Computing Grid project, June 2005. <http://lcg.web.cern.ch/LCG/PEB/gdb/nw-grp.htm>.
10. Tom Nolle. A New Business Layer For IP Networks. *Business Communications Review*, pages 24–29, July 2005. <http://www.ipsphereforum.org/newsevents/07nollereprint.pdf>.