ARGON: Reservation in Grid-enabled Networks

Christoph Barz, Uli Bornhauser, Peter Martini,
Markus Pilz, Christian de Waal, Alexander Willner
Institute for Computer Science IV, University of Bonn, Germany
{barz,ub,martini,pilz,dewaal,willner}@cs.uni-bonn.de

Abstract: Grid computing offers heterogeneous and distributed resources to scientific communities. Apparently, networks connecting these resources can also be considered as Grid resources. This paper presents ARGON, a system that integrates metro and wide area networks into Grid environments by providing advance reservations and guaranteed network services. Here, single-domain as well as multidomain network environments are considered. A major objective is to support metaschedulers in the planning of workflows for e-science applications with demanding network requirements.

1 Introduction

One objective of Grid computing [F+03] is to offer a standardized interface to heterogeneous resources such as computational clusters, data storage sites, and scientific instruments. These resources are distributed and typically interconnected via the Internet. Jobs within a Grid are often specified as workflows. As part of the workflow planning and execution, different types of network services can be used. This includes the transfer of data as well as network connectivity for synchronous streaming and coupling of parallel jobs between different Grid sites.

According to [Fer07], the analysis of local and remote data becomes more and more important in the field of e-science, medicine, engineering, and digital art. These data sets or streams may be as large as several terabytes, and may be real-time or preprocessed. It is important to guarantee that the data to be processed are delivered timely. It is envisioned that requests for multiple Gbps need to be fulfilled in order to leverage these applications. Since currently this requested quality of service (QoS) cannot be ensured in the Internet, many Grid sites are additionally interconnected by high-speed networks. These networks usually just provide manually arranged connections with a given QoS.

However, even in a high-speed network environment it might not be feasible to overprovision the network links, and the available bandwidth must be shared and allocated in an efficient manner. Furthermore, a static network setup restricts the selection of available Grid resources. In addition to high bandwidth requirements, interactive and collaborative environments may require specific round trip times, a low packet loss, and small jitter or a combination of those.

In order to allow for a flexible assortment, a system is needed that allocates network paths

on demand and offers this service to higher layers. This system needs to align the requirements of the Grid applications with the traffic engineering capabilities. As different Grid resources may have to be coallocated with the network resources, it is also necessary to make reservations in the network in advance. Usually, this coordination of different Grid resources is done by a metascheduling service (MSS). As a consequence, a network management system is an additional system to be queried by such a scheduler.

Both the demand of applications on the network and the above mentioned challenges lead to the design and implementation of ARGON (Allocation and Reservation of Grid-enabled Optical Networks). ARGON has the ability to abstract the network resources and makes them available by providing a reservation interface. On the one hand, ARGON integrates into an existing MSS [B+07a] that orchestrates and reserves all available Grid resources including a single network domain. On the other hand, it may be used by an interdomain system like the Network Service Plane (NSP) [F+07] to allow for a network allocation in multidomain environments.

This paper is structured as follows. In Section 2, the related work in this field is presented. We then proceed to present ARGON, our proposal for such a network reservation system in Section 3. In the following Section 4 we extend the considered objective to interdomain environments. Finally, future work is discussed and our conclusions are summarized in Section 5.

2 Related Work

A starting point for ARGON (Allocation and Reservation of Grid-enabled Optical Networks) was the German research project VIOLA (Vertically Integrated Optical Testbed for Large Applications in DFN) [VIO07]. The project established a vertically integrated approach to combine scientific applications, Grid middleware, and advanced networking equipment. In this section, we mainly focus on related projects and architectures that allow for an automated configuration of network resources for Grid applications. While there is no common terminology for such a system, we adopt the term **Network Resource Manager (NRM)** following the terminology of the Grid community. Other projects or systems might use terms related to network management.

Projects mainly focusing on intradomain configuration: The *UCLP* (User Controlled Lightpath Provisioning) system [UCL07] (the commercial Version is called ARGIA) is designed to allow end users to provision and dynamically reconfigure end-to-end lightpaths in optical networks across a single domain or multiple independently managed domains. UCLP supports the delegation of the control of subnetworks to other users. It is designed to allow for the creation of application-specific IP networks for high-end e-science and Grid applications. Nortel developed a proof-of-concept middleware called *DRAC* [T+05] (Dynamic Resource Allocation Controller) that allows for an application initiated configuration of network resources on an end-to-end basis. DRAC is able to coordinate optical and packet switched networks on demand or based on reservations. The main goal of the *G-lambda* project [T+06a] is to establish a standardized Web service interface between Grid resource

management systems and network resource management systems that also support advance reservations. The main focus of the project is on optical network technologies. The *EnLIGHTened Computing* project [B+07b] focuses on dynamic optical lightpaths between supercomputing sites which are created and torn down in advance or on demand based upon application needs. A domain manager allocates network resources by dynamically setting up and deleting dedicated circuits using GMPLS control plane signalling.

Projects mainly focusing on interdomain configuration: Originating from the GÉANT2 project, the *AutoBAHN* (Automated Bandwidth Allocation across Heterogeneous Networks) architecture targets at the needs of a multidomain, multitechnology research and communication community. Based on the InterDomain Manager (IDM) of JRA3 [C+06], the AutoBAHN architecture defines an interdomain network reservation mechanism based on a decentralized architecture for interdomain signalling. In addition, an interface to Domain Managers (DM) is defined. The DMs are responsible for the management and configuration of the local domains. The *DRAGON* project (Dynamic Resource Allocation in GMPLS Optical Networks) [T+06b] aims at both the dynamic intra- and interdomain provisioning of packet and circuit switched network resources in response to user requests for high-performance e-science applications. Its architecture is based on the GMPLS control plane stack and allows for deterministic services on an interdomain basis and across heterogeneous network topologies. DRAGON uses a peer or augmented interdomain model supporting the exchange of abstracted topology information for signalling and interdomain path computation. Its model is similar to the IETF Path Computation Element Architecture [FVA06]. As an achievement of the *DICE* (DANTE-Internet2-CANARIE-ESnet) collaboration [DIC07] of CANARIE, ESnet, GÉANT2 and Internet2, a web-based interdomain control plane was developed where IDCs (InterDomain Controller) communicate in a decentralized way to provision end-to-end multidomain network paths. Every IDC advertises an abstracted topology. As part of the admission procedure, an interdomain path is computed and signalled. The internal configuration of the domains involved is explicitly out of scope of the collaboration.

3 ARGON – An approach to intradomain network reservations

The following section describes the way ARGON factors network resources into a Grid environment. Core components of and services provided by the NRM are presented. This section concludes with a discussion on processing advance reservation requests.

3.1 ARGON's Core Components

Authentication and authorization (AA) are a cornerstone in the context of Grid computing. Every request processed by ARGON's **AA Management** is obliged to contain corresponding information about the requester. Immediately upon the receipt of a new request the contained AA information is being analyzed. Successfully authenticated and authorized

requests are passed to the Request Handler.

The **Request Handler** dispatches messages depending on their type (cf. Section 3.2) to corresponding components and handles replies which contain processing results and the status of reservations.

The **Resource Management** keeps track of available and allocated network resources. Note that while reservations are being made via the resource reservation system, spare network capacity can still be used for best effort traffic if traffic shaping is supported by the network components.

Using the information provided by the Resource Management, the **Path Computer** determines feasible resource allocations (cf. Section 3.3). While availability requests are only processed by the path computer, allocated resources are reserved if a reservation request is handled. After a successful allocation of resources, the Scheduler is informed about the corresponding starting and ending time of the reservations.

Once a reservation is accepted, the **Scheduler** is used to trigger the network configuration. This is necessary because network components used in today's systems are not aware of time and do not support scheduled configuration.

The **Signalling Subsystem** is triggered by the Scheduler and is responsible for the setup and tear down of paths in conjunction with traffic shaping configurations in the network. Considering the processing of an entire request, the configuration of the network components can be a time consuming process: Command line interfaces still dominate the configuration process of routers as comprehensive machine-machine interfaces are rare. Currently, MPLS and GMPLS entities from different vendors are supported.

3.2 ARGON's Supported Services

ARGON supports six different services to factor network resources into the Grid by supporting advance reservations. These services and the corresponding interface support a metascheduler in planning Grid application workflows with network specific requirements [B+07a].

A **Reservation Request** is used to make an obligatory reservation of network resources in advance. The request itself defines a workflow containing a hierarchy of three levels. The top level, denoted as *reservation*, aggregates several different *services* to a single workflow. Such a workflow is accepted or rejected as a whole by the reservation system. A service in turn comprises multiple *connections* between different endpoints specifying the same service type with the same time constraints. Connection parameters comprise a set of constraints like bandwidth or delay constraints as well as binding information (cf. Binding Request).

The **Availability Request** specifies the same parameter as a Reservation Request. This allows a user or metascheduler to check whether resources are currently available. If a requested workflow is unfeasible, the reservation system replies with the first starting time at which the requested workflow is feasible. This information facilitates the planning of

alternative configurations for temporary unfeasible grid jobs.

A **Cancel Request** is used to revoke an accepted reservation request. Reservations which are not yet in their usage phase are trivial to cancel, as only resource allocation information has to be adjusted. In contrast, cancelling a reservation during its usage phase requires a reconfiguration of routers in the network.

The **Modification Request** is used to change accepted reservations. The processing of a modification is handled in such a way that resources reserved by the original request cannot be allocated by other users. So, if a modification fails, the original reservation remains.

A reservation made in advance can be specified in such a way that either the configuration of network devices is triggered automatically by the Scheduler entity or explicitly triggered by an additional request. The **Activation Request** is used in the latter case to initiate the signalling process. Hence, allocated but unused resources might be temporarily be used by best effort traffic or other reservations.

The **Binding Request** can be used to specify additional information necessary for provisioning purposes. As an example, a binding is required if multiple cluster nodes are connected to the same endpoint (e.g. a router) of a connection and only a subset of the connected nodes participate. The parameters define the traffic to be mapped on a tunnel at the endpoint by using IP address and port information. A separate request type is required for specifying this information, because these parameters may not be available at the point in time when the reservation is made. This allows a reservation to be made although the involved cluster nodes are still unknown.

3.3 Processing Advance Reservations for Network Resources

Advance reservation requests contain time as well as resource related parameters that have to be mapped on available network resources. The task of the path computer is to compute an optimal schedule for a single request as well as to reschedule accepted requests to optimize the current resource usage. After introducing the resource management utilized in ARGON, the processing of advance reservations is discussed.

3.3.1 Resource Information Management

Information about the allocation of network resources for previously accepted requests is required to determine the feasibility of a new request. The resource information management keeps track of residual capacities (or usage profiles) in the network topology, i.e. the topological structure as well as allocated resources are regarded with respect to time. This information is temporally limited by a so-called *book-ahead* interval. The timeslot based management of allocated resources is an established way [GO00] to manage this information as it allows to decouple the computational complexity of the admission decision from the number of already accepted requests.

ARGON uses an enhanced timeslot model [BBMP08]: Timeslots have a dynamic length with a fixed granularity. In brief, for every new reservation, existing intervals may be split at the start and the end time of a requested connection. So, every time a connection of a reservation is accepted, no more than two new timeslots are created in addition to the already existing ones. This results in a timeline being segmented in at most $2k + 1$ non-overlapping subintervals, where k is the number of accepted requests. It should be noted that even if the number of timeslots depends on the number of reservations, an upper bound is given by the length of the book-ahead and the granularity.

3.3.2 Fixed and Deferrable Advance Reservations

A basic form of an Advance Reservation (AR) request is defined as follows: The request is received at t_{arrival}, is admitted and starts at t_{start}. Furthermore, the usage phase (duration) is limited by t_{end}. This life cycle is depicted on the left hand side in Figure 1.

A fixed advance reservation request for a single connection is defined as tuple $(t_{\text{start}}, t_{\text{end}}, s, d, C)$ where $t_{\text{start}} < t_{\text{end}}$. The reservation starts at t_{start} and ends at t_{end}. The endpoints of the connections are specified by s and d. C represents additional resource constraints which are usually the required capacity and delay constraints. The resource allocation information of the timeslots involved in the request is accessed to determine whether a request is feasible or not. The time complexity of the path selection (or path routing) is dependent on the constraint set. If only one connection with link constraints (e.g. minimum capacity) is requested, a shortest path algorithm with a polynomial running time identifies a candidate using a constrained topology. If multiple paths are requested at the same time or other path constraints (e.g. loss rate, delay) are included, the complexity of the path selection process can increase to a super-polynomial time complexity. In these cases, heuristics with potentially suboptimal results can be applied to keep the processing time low, e.g. multiple paths within a request are processed independently one after another.

The separation of the points in time at which a reservation request is received and at which it is supposed to begin results in a new degree of freedom for the reservation system. It may not be necessary to reply an advance reservation request directly. However, it is assumed that the reservation system operates like an online system, meaning that the negotiation phase is as short as possible. This is very important for coallocating resources, as following reservations of non-network resources may depend on the admission control decision made by the NRM. Nevertheless, if the resource usage in a certain time interval exceeds a threshold, a rescheduling process can be started to optimize the currently accepted reservations. The time complexity of this process is again dependent on the accepted reservations and corresponding constraints. Heuristics for this offline optimization problem are currently under development.

An additional type is a deferrable advance reservation which has a certain degree of freedom in the time domain. In particular, time related parameters define a range of possible values to establish the reservation. The life cycle of a deferrable advance reservation is given on the right hand side in Figure 1 and defined as tuple $(t_{\text{release}}, t_{\text{deadline}}, \Delta t, s, d, C)$ where $t_{\text{release}} + \Delta t < t_{\text{deadline}}$. The reservation can start at t_{release} and must end before t_{deadline}. The length of the usage phase is specified by the duration $\Delta t > 0$. Compared

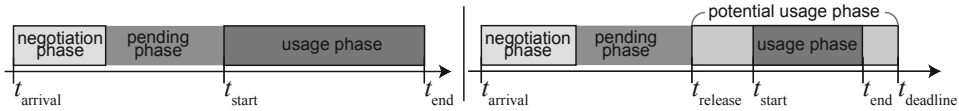

Figure 1: Life cycle of a fixed advance reservation (left) and a deferrable advance reservation (right)

to a fixed advance reservation, the parameters t_{start} and t_{end} ($\Delta t = t_{\mathrm{end}} - t_{\mathrm{start}}$) can be determined by the NRM. An option is to specify the minimum and maximum length of a usage phase as an interval. Again, various strategies can be introduced to handle this parameter range, e.g. to reduce the usage phase down to the minimum value, if this allows for the admission of additional requests.

3.3.3 Malleable Advance Reservation

A specification of the exact transmission rate can be omitted when a fixed amount of data has to be transmitted. Only general capabilities of the sender and the receiver, such as the maximal transfer rate and timing constraints for the transmission like a fixed deadline or an earliest starting time (release time), have to be regarded. By joining the time and resource constraints, the reservation system can find the most efficient solution for the requested transmission. This kind of reservation is denoted as malleable advance reservation [Bur04] (or advance cumulative reservation [GO00]). A motivating example regarding efficiency is to fill gaps between allocated resources which are caused by accepted reservations.

A malleable reservation request is defined as tuple $(t_{\mathrm{release}}, t_{\mathrm{deadline}}, s, d, S, C)$ where $t_{\mathrm{release}} < t_{\mathrm{deadline}}$. The endpoints of the connections are specified by s and d. S determines the data size (transmission rate and time product) and C represents additional constraints. Typical constraints for such a reservation are a lower and an upper boundary for the transmission rate. Currently, malleable advance reservations are realized by computing a deferrable advance reservation which satisfies the requested data size. Further processing strategies are under consideration [BMPP07]. Eventually, it is up to the capabilities of the systems involved which strategies can be used.

4 Integration into multidomain environments

The ARGON system introduced in the previous section allows for an integration of network resources into a Grid environment. To achieve this, ARGON requires detailed knowledge of the underlying network's topology and must be able to directly control the network. This is only feasible if the network belongs to a single administrative authority.

Once there are several domains, each controlled by its own NRM, naturally the desire will arise to provide support for services crossing the domain borders via interdomain links. Integrating all these domains under a single NRM is rarely a viable solution, since the possibility to use an existing, customized system will have to remain. Also, an administra-

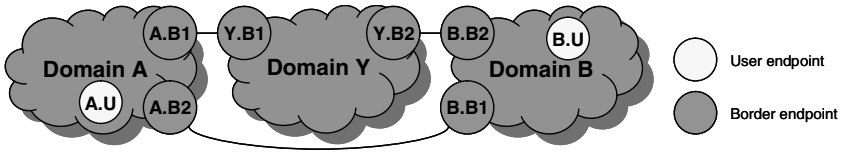

Figure 2: Illustration of an interdomain reservation scenario

tive authority will most probably insist that only its NRM is allowed to directly configure the network. In many cases, it will not even want to disclose its internal network topology.

Therefore, an interdomain connection has to be composed by multiple single domain connections that terminate at endpoints connected by interdomain links. Since this introduces additional complexity, a new system is needed that offers an NRM-like interface for a metascheduler (cf. Section 3.2) or a human user and that hides the multidomain aspects of the underlying network.

Such a system is currently being developed in the Phosphorus project funded by the European Union. In the first project stage, the *Network Service Plane* (NSP) [F+07] that unifies and abstracts from the underlying domains is implemented as a single, central entity. It is planned to extend the current solution to allow for a distributed operation in the future.

To integrate several NRMs with different interfaces in the NSP, the concept of an *adapter* is used. In general, the NSP does not communicate with a NRM directly, but through a NRM adapter whose task is to "translate" between the unified NSP interface and the NRM-specific interface. Conceptually, the adapter code does not belong to the NSP, but is an extension of the NRM that allows it to be part of a multidomain network. Currently, NRM adapters exist for ARGON, DRAC, and UCLP.

The interface specification of the NSP interface towards a metascheduler or towards a human user has been extended from the ARGON interface. It is the same as that of the NRM adapter interface towards the NSP. The reason is that the services offered by the NSP are the same services as those offered by a NRM.

The topology information managed in the NSP is comprised of domains, endpoints, and interdomain links (cf. Figure 2). Endpoints are divided into two different categories: *Border endpoints* are used for interdomain links and must therefore be known within the NSP, but are not of further interest for a metascheduler or for a human user. *User endpoints* in contrast are actually used by applications. Therefore, they need not be known within the NSP, they must merely be locatable, i.e. there must be a mechanism to look up which domain a user endpoint is located in. In the current NSP implementation, this is achieved by assigning separate user endpoint address spaces to the domains.

Topology information currently enters the NSP through an administrative interface, the *topology interface*. This interface is accessed by the NRM adapters running inside the domains. They are in charge of keeping the domain-specific information up to date. Currently, interdomain links are entered manually through an interactive management client. A protocol to exchange topology information between multiple NSP instances is under

development.

An example of an interdomain reservation is now illustrated for the scenario depicted in Figure 2. When the NSP receives a request for user endpoints A.U and B.U to be connected, it is not aware of the resource usage within the domains. It therefore must split the requested interdomain connection to intradomain connections, each of which fulfils the constraints specified in the original request. The path computation module would first return the path A.U–A.B2, B.B1–B.U. The NSP queries the availability of each of these connections at the corresponding domain's NRM adapter. In case all connections are available, the corresponding reservations are made and a positive reply is sent back to the requester. If, however, resources turn out to be unavailable (e.g. the intradomain path A.U–A.B2), they are marked accordingly and a new path is calculated, in this case A.U–A.B1, Y.B1–Y.B2, B.B2–B.U. If no available path can be found, a negative reply is returned.

5 Conclusions and further work

The ARGON system presented in this paper enables the integration of metro and wide area network resources into a Grid environment. This is achieved by offering guaranteed network services via advance reservations. Different types of network connectivity are supported according to the needs of the applications. The corresponding interface towards the Grid allows applications or metaschedulers to efficiently plan complex workflows that require coallocation of heterogeneous Grid resources comprising the network.

ARGON has been designed, developed, and successfully demonstrated in the scope of the VIOLA and Phosphorus project. In this context ARGON acts as a single-domain NRM and interfaces to MPLS and GMPLS equipment from different vendors. Furthermore, ARGON has been successfully integrated into the Phosphorus testbed containing multiple NRMs which originate from various research projects. The newly developed NSP (cf. Section 4) connects different NRMs and provides a heterogeneous multidomain solution. In both cases, the practical experience gathered was and is used to enhance ARGON.

The future work on ARGON and the NSP comprises technical aspects, enhancements of the Grid services and the corresponding interface, as well as associated theoretical issues. Furthermore, emerging standards like WS-Agreement and WS-Notification are under consideration, e.g. WS-Notification can be used to distribute information on changing conditions and especially network failure situations in a standardized way. In these cases, an additional negotiation phase can be triggered in order to modify, postpone, or cancel reservations.

Finally, also distributed control plane solutions are evolving parallel to the centralized Network Resource Managers. In the Phosphorus project, not only the Network Service Plane described in Section 4 is being developed, but also an extended GMPLS control plane that (among other Grid features) is capable of making advance reservations. Integrating such a distributed approach with the centralized approaches of Network Resource Managers and the Network Service Plane is a challenging future goal of Phosphorus.

References

[B+07a] C. Barz et al. Co-Allocating Compute and Network Resources - Bandwidth on Demand in the VIOLA Testbed. In *Proceedings of the CoreGRID Symposium*, CoreGRID Series. Springer, September 2007.

[B+07b] Lina Battestilli et al. EnLIGHTened Computing: An Architecture for Co-scheduling and Co-allocating Network, Compute, and other Grid Resources for High-End Applications. In *Proceedings of 4th International Symposium on High Capacity Optical Networks and Enabling Technologies (HONET)*, 2007.

[BBMP08] C. Barz, U. Bornhauser, P. Martini, and M. Pilz. Timeslot based resource management in grid environments. In *Proceedings of the IASTED International Conference on Parallel and Distributed Computing and Systems (PDCN)*, February 2008.

[BMPP07] C. Barz, P. Martini, M. Pilz, and F. Purnhagen. Experiments on Network Services for the Grid. In *Proceedings of the 32nd IEEE Conference on Local Computer Networks (LCN'07)*, pages 45–54, 2007.

[Bur04] Lars-Olof Burchard. *Advance Reservations of Bandwidth in Computer Networks*. PhD thesis, Berlin University of Technology, August 2004.

[C+06] Mauro Campanella et al. Deliverable DJ3.3.2: Functional Specification of GÉANT2 Inter-domain Manager (IDM) Prototype. Technical report, GÉANT2, 2006.

[DIC07] DICE (DANTE-Internet2-CANARIE-ESnet) Collaboration. `http://www.geant2.net/server/show/conWebDoc.1308`, accessed 31 Dec. 2007.

[F+03] I. Foster et al. *The Grid 2: Blueprint for a New Computing Infrastructure-Application Tuning and Adaptation*. Morgan Kaufman, San Francisco, CA, 2003.

[F+07] S. Figuerola et al. PHOSPHORUS: single-step on-demand services across multi-domain networks for e-science. *Proceedings of SPIE*, 6784, 2007.

[Fer07] Tiziana Ferrari. Grid Network Services Use Cases from the e-Science Community. Technical report, Open Grid Forum (OGF), December 2007. OGF Reference: GFD.122.

[FVA06] A. Farrel, J. P. Vasseur, and J. Ash. A Path Computation Element (PCE)-Based Architecture. IETF RFC 4655 (Proposed Standard), August 2006.

[GO00] Roch Guérin and Ariel Orda. Networks with Advance Reservations: The Routing Perspective. In *Proceedings of the IEEE INFOCOM*, pages 118–127, 2000.

[T+05] F. Travostino et al. Project DRAC: Creating an applications-aware network. *Nortel Technical Journal*, February 2005.

[T+06a] Atsuko Takefusa et al. G-lambda: coordination of a grid scheduler and lambda path service over GMPLS. *Future Gener. Comput. Syst.*, 22(8):868–875, 2006.

[T+06b] T.Lehman et al. DRAGON: a framework for service provisioning in heterogeneous grid networks. *IEEE Communications Magazine*, 44:84–90, March 2006.

[UCL07] UCLP. `http://www.uclp.ca`, accessed 31 Dec. 2007.

[VIO07] VIOLA. `http://www.viola-testbed.de`, accessed 31 Dec. 2007.