The BGRP Plus Architecture for Dynamic Inter-Domain IP QoS

Stefano Salsano[1], Martin Winter[2], Natalia Miettinen[3]
[1] DIE, University of Rome "Tor Vergata" - [2] SiemensAG – [3] AG ELISA communications
stefano.salsano@uniroma2.it, martin.winter@siemens.com, natalia.miettinen@elisa.fi

Abstract

This paper describes a scalable inter-domain resource control architecture for DiffServ networks. The architecture is called BGRP Plus, as it extends the previously proposed BGRP framework. The reference network scenario for inter-domain QoS is first presented, highlighting the requirements for an inter-domain resource reservation mechanisms. The key aspects of the proposed solution are described, followed by the messages and the procedures of the BGRPP. Finally a trial implementation and some performance measurements are presented.

1. Introduction

The DiffServ architecture specifies a set of "user plane" mechanisms, which can be used to provide QoS in IP networks. Intentionally, the IETF DiffServ WG has not covered any "signaling plane" aspect. QoS signaling capabilities are indeed needed to extend the provisioning of QoS in IP networks from a static model towards a dynamic one. The IETF WG NSIS (Next Steps In Signaling) [1] has been specifically chartered to address the signaling aspects of QoS in IP networks. The NSIS WG is currently defining the requirements for the QoS signaling mechanisms, and is considering a set of reference scenarios. One of these scenarios is the QoS reservation/ negotiation over administrative boundaries or "inter-domain" QoS.

In order to realize true end-to-end QoS services in the Internet, spanning multiple administrative domains, efficient and scalable signaling and resource control mechanisms are needed. In particular the scalability is a fundamental issue in the definition of an inter-domain QoS model, because the ambitious goal is to support QoS services on the scale of the global Internet.

This paper describes an architecture that originates from the BGRP protocol framework proposed in [2], providing a mechanism for aggregation of resource reservations spanning multiple (DiffServ) domains. The aggregated reservations are negotiated between so-called BGRP agents, which are deployed at each BGP-capable border router of each DiffServ domain. In this way, each domain can perform some kind of admission control, taking into account the available resources within that domain, the available resources on the inter-domain links and the inter-domain Service Level Agreement. By aggregating the reservations according to the "sink trees" created by the BGP routing protocol [3], the number of reservations and thus the amount of state information stored in the network can be reduced.

However, aggregation of reservations is just the first step towards scalability. To limit the signaling load and the processing power required in the BGRP agents, it is also necessary to reduce the number of signaling messages. We propose mechanisms for the early response to reservation messages, in [2] called "quiet grafting", so that not each message has to travel edge-to-edge through the DiffServ network region. The architecture proposed in this paper is called BGRP Plus (BGRPP or BGRP+). The BGRPP architecture has been implemented in the context of the AQUILA IST project [4]. The implemented trial is described in section 6.

2. Reference Network Scenario

The architecture described in this document assumes a DiffServ region consisting of several connected, but administratively separated domains. The traffic can enter and leave the domains at two different types of routers:

– An edge router (ER) connects a domain to a network, which is not taking part in the BGRPP resource allocation mechanism, e.g. an access network.
– A border router (BR) connects a domain to another domain, which also takes part in the BGRPP resource allocation mechanism.

A source or destination domain is a domain with at least one edge router. A transit domain is a domain with at least two border routers, which forwards traffic received at one border router to another border router. All edge routers and border routers are required to run BGP. Corresponding to each border router, a BGRP+ agent is instantiated. Fig. 1 shows the reference architecture.

We assume that the source domain is capable of performing some kind of per flow admission control, taking into account the available resources up to BR1. The source domain has however no information about the availability of resources along the further path of the reservation through other domains. The intra-domain architecture is not relevant for BGRPP, therefore the intra-domain resource control elements are intentionally not specified in Fig. 1. To perform inter-domain admission control, the source domain determines the egress border router and contacts the corresponding BGRPP agent. Through the BGRPP protocol this agent is able to determine, whether resources are available in each domain along the path and on the links connecting the domains. Each BGRPP agent cares for the resources on the next path segment from its associated border router towards the destination. Resource reservation in the source and destination domain is not the task of the BGRPP protocol. However, as we describe later, BGRPP has to provide mechanisms to enable resource reservation in the destination domain.

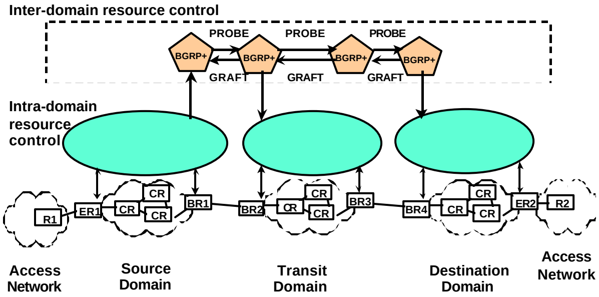

Fig. 1. Reference network configuration

As shown in Fig. 1, there are mainly two types of messages involved in the reservation set-up:

- PROBE messages are initiated at the source domain and are forwarded between BGRPP agents along the BGP path. They check the conformance of the request with the SLAs between neighbored domains. The path is recorded in the message, to enable the response to take the same way in the backward direction.
- GRAFT messages indicate the availability of resources towards the destination domain. During message processing, the resource availability is checked and the resources are actually reserved in each path segment.

3. Requirements for inter-domain resource reservation

Let us consider the set of requirements for the architecture and protocol for inter-domain resource reservation leading to the definition of BGRP plus architecture.

The most important requirement is probably related to scalability: the architecture has to scale well with the current Internet size and growth. The aspect of scalability is specifically dealt with later in the following subsections.

Each domain should be autonomous in the handling of its inter-domain resources and it can take autonomous admission control decisions.

Another aspect is the independence of inter-domain architecture with respect to inter-domain resource control mechanisms. Each network operator can independently administer his network and configure the behavior of intra-domain resource control independently. The inter-domain protocol should be independent of the intra-domain resource control architecture. A domain may use static provisioning as well as dynamic resource allocation. However, a common interface between intra-domain and inter-domain resource control has to be defined.

As for the routing, we think that inter-domain resource reservation shall rely on existing inter-domain routing (e.g. BGP) for routing decisions. An interface has to be defined between inter-domain routing and inter-domain resource reservation to allow the latter to retrieve routing information (e.g. the NEXT_HOP). We assume that BGP remains unaffected by the inter-domain resource reservation architecture.

Finally, we note that in the inter-domain scope, the most important constraint is not an optimization of network resources but the conformity to the Service Level Agreement (SLA) with the neighbor.

3.1 Scalability aspects: the quiet grafting mechanism

BGRPP is an inter-domain protocol to allow resource reservation. The main issue that BGRPP should tackle is the scalability problem, related to the handling of state information for each reserved flow and to the rate of reservation messages. In particular the handling of state information impacts on the memory needs for each router that runs reservation protocol and the rate of reservation messages impacts on the CPU usage for message processing and on the bandwidth utilization for signaling messages.

BGRP as described in [2] mainly addresses the scalability in terms of the amount of state information kept in each BGRP agent. It aggregates reservations along the BGP sink trees and thus achieves a scalability behavior, where the memory requirements are proportional to the number of sink trees with simultaneous active reservations.

However, in order to check the resource availability and the compliance to inter-domain SLA, BGRP messages have still to travel along the full path to the destination domain, asking each BGRP agent to check for that part of the network it is responsible for. So the message processing load may be rather large, especially in big backbone networks.

In order to solve this problem, a hint is given in [2]: resources may be kept in advance at a BGRP agent, so that further requests may be already terminated at an earlier stage. In this document we provide the complete functional specification of this mechanism, called “quiet grafting”. Together with the sink-tree-based aggregations, this mechanism provides a scalable solution for inter-domain resource reservations.

When an inter-domain reservation is initiated at a source domain, the first BGRPP agent constructs a PROBE message indicating the amount of bandwidth required and the destination address. It is not evident, to which sink tree this reservation will belong. So the PROBE message is forwarded hop-by-hop between BGRPP agents along the BGP route to the destination. Obviously, the last BGRPP agent, which corresponds to the root of the sink tree, can assign a sink tree id to the reservation. The GRAFT message sent back will contain this sink tree id. As this message travels back to the source domain, it installs the necessary sink tree reservations in the path segments.

To enable an intermediate BGRPP agent to answer a PROBE message successfully with a positive response, the following conditions have to be met:

1. The BGRPP agent must be able to determine the sink tree, to which the reservation belongs.
2. The BGRPP agent must have pre-reserved resources for this sink tree, so that he can guarantee, that the resources are available on the path segment from the current point to the destination domain.
3. As the last BGRPP agent may no longer be informed about a new reservation, the BGRPP agent must provide means to contact the destination domain, so that resources can also be reserved on the not-BGRPP-controlled path segment from BR4 to ED2 (see Fig. 1).

The following subsections describe the mechanisms defined in BGRPP to meet the above conditions.

3.2 NLRI Labeling for Sink Tree Identification

According to [2], a sink tree is identified by the destination AS number and a border router id. The network layer reachability information (NLRI) associated with the root of the tree can be used to identify the tree at a distant point in the network. As BGP may aggregate this information into aggregated routes, it is not directly derivable from BGP routing information. Instead, we propose to propagate the NLRI back from the root of the sink tree in GRAFT messages and to store this information in each BGRPP agent, which processes the message.

When a new PROBE message arrives, the BGRPP agent will compare the destination address with the already stored NLRI thus trying to identify the sink tree. Successful sink tree identification is a prerequisite to the following steps, to perform successful quiet grafting.

To reduce memory requirements, BGRPP agents will only store NLRI information for those sink trees, where actual reservations exist.

3.3 Pre-reservation

When a BGRPP agent processing a PROBE message was able to determine the sink tree, he will now check, whether he has pre-reserved resources on this sink tree. If this is the case, he can generate a GRAFT message and terminate the PROBE at this stage. If there are not sufficient resources available, the BGRPP agent will further forward the PROBE message to the next BGRPP agent, as determined by the NEXT_HOP attribute of the BGP path to the destination.

The amount of pre-reserved resources that is available at a given time for a sink tree is called resource cushion. A possible solution to build this resource cushion at each BGRPP agent is the “delayed resource release” mechanism. This means, that a BGRPP agent will not immediately release unused resources, but instead keep them in an attempt to satisfy further requests. Resources are however released, when they are unused for some time. The exact specification of this algorithm and the analysis of its performance is provided in [7].

3.4 Signaling in the last domain

When a PROBE message is answered “in advance”, the last domain may not be informed about the new request and cannot reserve resources within that domain. To include the resource availability check in the last domain in the overall admission control, we propose to back-propagate a reference to a signaling interface in the destination domain interface, that can be used by the originating domain to directly send a reservation request. This information is also stored within each BGRPP agent that handles this message. The originating BGRPP agent can then use this reference to directly request the required resources in the destination domain. This reference is also necessary for the release of a reservation: since resources are not immediately released along the whole path when the original end-user request is removed, possibly the destination domain will not be informed at all about this event. For this purpose it is necessary that the originator of the request can directly inform the destination domain.

In our current specification and implementation of BGRPP we use a reference to the intra-domain resource control in the destination domain. This is not the optimal solution (and it will be fixed in the next version of the specification) as it introduces a dependence on the specific intra-domain mechanism used. The correct solution is to use a reference to the destination domain

BGRPP agent and let it contact its own intra-domain resource control.

4. BGRPP messages

In the following description of messages and procedures we assume that "reservations" which are originated by a source domain (and propagated up to the destination domain) are characterized by a bandwidth parameter. Actually, a reservation should be also be characterized by the required service and other parameters besides the bandwidth could be needed. We assume that a set of "Globally Well Known Services" is defined to characterize the required service. The assumption that the bandwidth is the only needed parameter has been used in the AQUILA trial implementation of BGRPP. Depending on the level of aggregation that can be reached in the inter-domain link and on the relative amount of QoS services on a link, this assumption could also be used in the real world.

The **PROBE** message consists of a reservation request and destination network information. It travels from origin AS towards destination AS and collects routing information. It contains the information of the requested amount of resources.

Table 1. The fields of PROBE message

Sender	It is a BGRPP agent identifier, it is composed of AS id and the IP address of BR that has sent the PROBE.
ProbeId	The origin AS chooses the ProbeId. The ProbeID scope is local to the originating BGRPP agent and only used there to match the GRAFT (or ERROR) message. In other words, only the originating AS stores a "PROBE state" while Intermediate ASs nodes does not need correlate PROBE and GRAFT.
GwksId	GWKS id
Destination	IP address prefix of the host or network that is destination of request
Required BW	Requested bw [bit/s]
TreeId	Id of sink tree, it is NULL if the sink tree is unknown. This information is actually redundant, as each node could check weather the Destination matches an existing sink tree. By avoiding this check, it enhances the performances.
Path	Record of the route in terms of BGRPP agents from origin AS

When the destination AS (or a transit AS if it can do quiet grafting) receives a PROBE and it can accept the request, it returns a **GRAFT** message towards the origin AS, using the Path information collected by PROBE message. This message carries sink tree information. It also carries the amount of resources that are reserved by the node sending the GRAFT.

Table 2. The fields of GRAFT message

Sender	It is a BGRPP agent identifier, it is composed of AS id and the IP address of BR that has sent the GRAFT.
Id	Msg id to match GRAFT with PROBE, echoed from PROBE
GwksId	GWKS id
Destination	Ip address prefix of the host or network that is destination of request, echoed from PROBE
ReservedBW	Reserved bw [bit/s]
TreeId	Id of sink tree
DestResMgr	Reference to destination domain reservation manager for reservation in the last domain
Address PrefixList	NLRI
Path	Record of the route in terms of BGRPP agents to be followed back

The **REFRESH** message contains the indication of the current amount of needed BW. It is sent by a previous (upstream) node to reduce the amount of requested resource. A REFRESH with zero bandwidth is used to tear down a reservation. REFRESH messages must never be used by a previous (upstream) node to increase the amount of requested resources.

Table 3. The fields of REFRESH message

Sender:	It is a BGRPP agent identifier, it is composed of AS id and the IP address of BR that has sent the REFRESH.
GwksId:	GWKS id
ActualBw:	actual reserved bandwidth
TreeId:	id of sink tree

The **ERROR** message indicates that some error has occurred. It contains a description of the specific error case. For example if the resources are not available, an ERROR message is sent and propagated backwards up to the originator of the request. Finally, the **TEAR** message was defined in the BGRP architecture. Currently, it has no use in the BGRPP architecture, because it is replaced by the use of refresh messages.

5. BGRPP procedures

5.1 State information in BGRPP agents

In order to process the BGRPP protocol messages, the BGRPP agent stores the following sink tree status information. For each sink tree (with at least one reservation) and for each service, the next ("outgoing") hop and a list of previous ("incoming") hops is stored. The value of reserved resources for the outgoing hop and for each incoming hop is stored (this information is replicated per each GWKS). For each sink tree, the NLRI is stored, to enable sink tree identification for quiet grafting (see 3.2). Additionally, a reference to the intra-domain resource control of the destination domain is stored, to enable signaling to the last domain (see 3.4). The resource cushion is the difference between the Outgoing.res and the sum of the Incoming[i].res.

Table 4. State information for a sink tree

Next hop	IP address of next hop BGRPP agent
Outgoing.res[g]	Reserved BW for the sink tree x towards the Next hop for the GWKS g
Previous hop[i]	IP addresses of previous hop BGRPP agents
Incoming[i]. res[g]	Reserved BW for the sing tree assigned to each Previous hop for the GWKS g
NLRI	NLRI for the sink tree
Intra-dom. RC Ref.	Reference to the intra-domain resource control of the destination domain

5.2 Message handling

(In the following the dependence from the GWKS g will always be omitted).

The message handling for a transit AS is described, the procedures for Originating AS and Terminating AS can be derived. In [6] a complete pseudo-code description of the message handling is provided.

The **PROBE** messages can be:

- rejected because SLA/SLS does not match or there are no resources on outgoing link or on intra-domain links
- accepted with Quiet Grafting
- forwarded to downstream node with no change in RequiredBW

When a PROBE is received a preliminary admission control is performed. If the BGRPP agent is at an ingress Border Router, it checks for SLA between the requester AS and its own AS and decides if the request can be accepted. The BGRPP agent could also check if intra-domain resource manager could accept the new request. If the BGRPP agent is associated to an egress BR it has to check the SLA with the AS of next hop BR. It can also check for the resources on the outgoing link towards the next hop BR. If any of these checks fail, the request cannot be accepted and an ERROR message is sent to previous hop. If the checks are OK, the node (both the ingress and the egress BR) first verifies if the Destination of the PROBE matches an existing sink tree. In this case it is first checked if the resource cushion can fit the RequiredBW. In case of positive answer, Quiet Grafting is performed: the Incoming[PH].res related to the Previous Hop is increased by the RequiredBW value and a GRAFT message is sent to the Previous hop. If the Destination of the PROBE does not match an existing sink tree, or the resource cushion is not enough, the PROBE is forwarded and the state information of the node is not touched in any way.

The resource reservation is performed when a **GRAFT** arrives. A GRAFT tells the receiving node the downstream node can accept a given amount of BW for a given sink tree. Now the receiving node should decide if it has the resources to send this amount of BW towards the next hop BR and if the SLA matches. In particular, an Ingress BR should check the Incoming SLS with upstream AS and should consider the availability of intra-domain resources towards egress BR. An Egress BR should check the Outgoing SLS towards the downstream AS and should consider the resources on the outgoing link towards the Ingress BR of the downstream AS. (The nodes could have already performed these checks during PROBE phase, but the situation could have changed).

If the resources are available and SLA/SLS matches, the graft is accepted and the node:

1. Increases the Outgoing.res by the ReservedBW value.
2. Increases the Incoming[PH].res of the Previous hop PH by the ReservedBW value.
3. Propagates the GRAFT to the Previous hop.

If the resources are not available or (incoming/outgoing) SLA/SLS not matching, the GRAFT cannot be accepted. The node will immediately send an ERROR towards the first originator of the PROBE. If the reason of refusal is outgoing SLA not matching or unavailability of resources, the node has also to release reserved BW to next hop because it can not utilize it. To this purpose the node sends a REFRESH message to its next hop. If the reason of refusal is incoming SLA not matching the node can maintain the additional reserved BW towards its next hop to enlarge resource cushion.

6. Trial

The AQUILA project has implemented the BGRPP architecture in a trial, interacting with commercial routers running the BGP protocol. The BGRPP message exchange

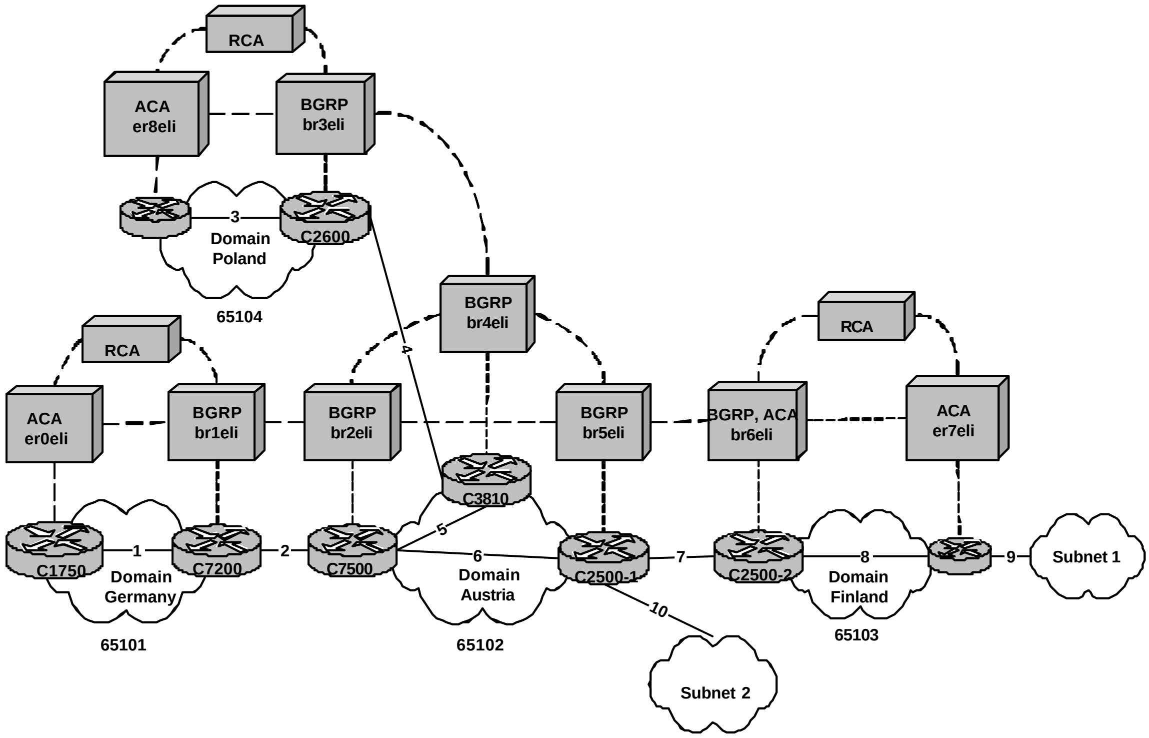

Fig. 2. Inter-domain scenario test network

has been realized as invocation to remote objects using the CORBA distributed processing environment, hereafter we will present some measurements taken on the testbed.

In the context of the AQUILA project, the BGRPP performances have also been analysed by simulation. Results are available in [7] about the scalability of the proposed solution.

6.1 Test Environment

The main goal of these trials is to evaluate the set-up time in the AQUILA inter-domain architecture. The test environment consists of four individual domains. Poland and Finland domains have one virtual edge router and one border router each. Austria domain consists of three border routers. Germany domain consists of one border router and one edge router. The reservations are started either from Germany domain or Poland domain and the reservations end point is in Finland domain. In each domain there are AQUILA RCL and BGRP corresponding to border routers.

6.2 Reservation Processing Delay

In order to measure the reservation processing delay, a set of reservations from Germany domain (er0eli) to Finland domain (er7eli) were setup and released. All possible tracing traffic was switched off to minimize additional delays caused by debugging. The first reservation takes a very long time (in the order of 20 seconds), because the communications between the elements of the Distributed Processing Environment (CORBA) must be initialized and because the BGRP agent have to setup telnet connections with the routers. Then the following reservation are much faster, because the CORBA communications are active and because the request will fit into an existing sink tree. Each following request will find available the resources that were released from the previous reservation.

The test was repeated twenty times. Of course only the setup time for subsequent reservation have been considered. The average times and deviations calculated from the test results are presented in the following tables.

Table 5. Signalling processing delay when request fits into sink tree

Setup Delay [s]		Release Delay [s]	
Delay	Deviation	Delay	Deviation
1.452	0.1	0.506	0.03

The analysis of the setup delay must take into account that the setup time includes intradomain operations and the procedure for signaling in the last domain. From other measurements, we can estimate that these operations account for about 1s. Therefore the setup time directly related to the BGRPP procedures is in the order of 0.4 s in case the request fits the existing sink tree. Note that in this case the BGRPP messages do not propagate in the following domains because the request can be fulfilled by the first BGRPP agent in an existing sink tree.

We have analyzed the case where there is an actual propagation of BGRPP messages in the following domains because the reservations do not fit into existing sink tree. In this scenario the following reservations are added without releasing the previous ones, so that no resources are available in the sink-tree. The setup time for this case is in the order of 2 s. Considering that the interdomain operations and the procedure for signaling in the last domain always account for about 1 s, we can estimate that about 1 s is needed in this case for the propagation and processing of BGRP messages.

It was also evident that the number of the ongoing reservations has no impact on reservation set-up time, because the setup delay was not increasing with the subsequent requests.

7. Conclusions

Drawing the conclusions, it is worth to recall the extensions of our work with respect to the original BGRP proposal. The architecture now includes the aspects related to the interactions with intra-domain resource reservation mechanisms. The quiet grafting mechanisms simply mentioned in the BGRP proposal has been fully specified. The mechanisms to match a flow into the sink-tree by means of the NLRI information and to distribute this NLRI information where needed have been specified. The issue of the missing signaling in the last domain due to quiet grafting has been addressed and resolved. The semantic content of the messages has been described. The procedures for message handling have been given.

The BGRPP architecture is largely compliant to the requirements for the QoS signaling under definition by NSIS WG.

The AQUILA project has implemented the BGRPP architecture in a trial, interacting with commercial routers running the BGP protocol. The BGRPP message exchange has been realized as invocation to remote objects using the CORBA distributed processing environment. Performance measurement have been taken in the testbed.

8. Acknowledgements

Part of this work has been funded under the European Commission 5th framework IST program. The authors would like to acknowledge all their colleagues in the AQUILA project for their important contribution to this work.

References

1. M. Brunner (Editor), "Requirements for QoS Signaling Protocols", draft-ietf-nsis-req-06.txt, December 2002, Work in Progress
2. P. Pan, E. Hahne, H. Schulzrinne: "BGRP: Sink-Tree-Based Aggregation for Inter-Domain Reservations", Journal of Communications and Networks, Vol. 2, No. 2, June 2000, pp. 157-167, http://www.cs.columbia.edu/~pingpan/papers/bgrp.pdf
3. Y. Rekhter, T.J. Watson, T. Li: "A Border Gateway Protocol 4 (BGP-4)", RFC 1771, March 1995
4. AQUILA IST Project http://www.ist-aquila.org/
5. Y. Bernet et al., "Integrated Services Over Diffserv Networks", RFC 2998, November 2000
6. S. Salsano (Editor) “Inter-domain QoS Signaling: the BGRP Plus Architecture”, draft-salsano-bgrpp-arch-00.txt, May 2002, Work in Progress http://www.ist-aquila.org/aquila/files/standardization-activities.htm
7. E. Nikolozou et al. “BGRPP: Performance evaluation of the proposed Quiet Grafting mechanisms”, <draft-nikolouzou-bgrpp-sim-00.txt>, http://www.ist-aquila.org/aquila/files/standardization-activities.htm