Applying the BGRP Concept for a Scalable Inter-Domain Resource Provisioning in IP Networks

Peter Sampatakos, Eugenia Nikolouzou and Iakovos Venieris
National Technical University of Athens, 9 Heroon Polytechniou 15773 Athens, GREECE

Abstract: Providing end-to-end QoS guarantees to mission critical applications comprises a main challenge for the today's Internet infrastructure. The Differentiated Services architecture (DiffServ) enhanced by the Bandwidth Broker (BB) approach is a first step towards resource management. In this context, the architecture proposed in this paper realises a distributed BB architecture, in order to provide a more efficient way for managing the resources of a single domain.

Moreover, in order to provide a way for inter-domain resource control, the BGRP framework is applied and enhanced. The BGRP protocol provides sink-tree based aggregation of resource reservations for the delivery of end-to-end QoS to applications across multiple separately administered domains. Our discussion extends to quiet grafting mechanisms, which succeed in limiting the signaling load and efficiently handling the reserved resources between domains.

Key words: BGRP, quiet grafting, Inter-domain resource reservation, DiffServ

1. INTRODUCTION

Internet is the technology that has become part of our every-day life over the past years and gains significant momentum day by day. Although it started as an experiment, nowadays, it is a serious business and it aims to be the integrated infrastructure that will concentrate most or even all of the services, existing and future ones. However, the protocols and mechanisms of the current Internet technology seem to be insufficient for delivering the traffic of the arising and demanding multimedia applications with the

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35673-0_28
C. McDonald (ed.), *Converged Networking*

appropriate Quality of Service (QoS) characteristics, and thus enhanced mechanisms have to be deployed to provide a QoS-enabled Internet infrastructure. Towards to this effort, two distinct approaches have been defined: the *Integrated Services* (IntServ) [1] and the *Differentiated Services* (DiffServ) [2,3]. The first approach was the first significant step for the introduction of QoS in the Internet. *IntServ* uses the Resource Reservation Protocol (RSVP) for the explicit set-up of reservation state on each network node along the path from the sender to the receiver. However, the constant exchange of RSVP messages, as well as the need for separate reservation establishment for each flow raised scalability concerns. In contrast to the per-flow orientation of RSVP, *DiffServ* architecture classify packets into one of a small number of aggregated flows, based on the DiffServ codepoint (DSCP) in the packet's IP header. The primary benefit of *DiffServ* is scalability, since it eliminates the need for per-flow state and pre-flow processing and therefore scales well to large scale networks. Although these are two independent models where the DiffServ model was introduced as a rather simple and easily deployable model that came to replace the IntServ model and overcome its scalability issues, it is finally realized that they are not competitive but rather complementary in the pursuit of end-to-end QoS [4,5,6].

The concept of the Bandwidth Broker (BB) that has been introduced from the early stages of the DiffServ model [3] is responsible for performing policy-based admission control, managing network resources, configuring specific network nodes, among others. Nowadays, the Internet community directs its efforts towards the specification and standardization of the mechanisms of the BB, as well as the development of a prototype [7,8]. However, a protocol for communication between the BBs must be specified and standardized.

The QBone bandwidth broker architecture defines a model for the BB and specifies an inter-domain interface between peering BBs [8]. Resource allocation requests from end-systems are sent to the BB of a domain, which performs admission control for that domain and forwards the request to the next hop domain on the way to the sink domain. For communication between the BBs, the Simple Inter-domain Bandwidth Broker Signalling (SIBBS) [8] protocol is proposed, forming a single layer of bandwidth brokers, which control the resources within each domain. The SIBBS architecture has several drawbacks, since intra-domain and inter-domain resource control are not clearly separated. Instead, the BB is responsible both for controlling resources within the domain and for inter-domain resource allocation. Moreover, in this way the BB has to cope with each single request, which impedes the deployment of a scalable solution. In fact, SIBBS assumes that each end-system request is forwarded along the path to the destination domain. To reduce the number of signaling messages, core

tunnels are proposed, which provide a pre-reserved pipe for further bandwidth allocations.

On providing a more scalable and easy deployable solution for the inter-domain resource control mechanism, the Border Gateway Resource Protocol (BGRP) [9] framework is adopted in our architecture. The basic idea of the BGRP is the aggregation of reservations along the sink-trees formed by the BGP routing protocol [10,11]. Nevertheless, a number of extensions and enhancements are added to the current BGRP approach, in order to provide a more scalable and promising solution.

The paper is organized as follows: Section 2 specifies the overall architecture, describing the intra-domain architecture realized by the introduction of the Resource Control Layer (RCL). Moreover, the requirements for defining the inter-resource control protocol are specified. Section 3 introduces the BGRP framework, determining its limitations, and providing a number of solutions, named quiet grafting mechanism. Finally, conclusions and future work are delineated in Section 4.

2. ARCHITECTURE

The aim of this section is to give a very brief introduction of the Resource Control Layer (RCL) architecture[1], which is a prototypical next generation Internet architecture. The RCL architecture introduces a control layer that acts as distributed bandwidth broker and controls the resources of the underlying DiffServ-aware network. Although the classical BB architecture proposes a centralized approach where one BB is responsible for an administrative domain, RCL is designed as a hierarchical and distributed BB, to overcome potential scalability problems.

Synoptically, the RCL interacts with host applications and end-users for QoS requests, performs per flow policy-based admission control, configures edge routers to accommodate the admitted flows, and monitors and manages the overall resources in the network. As depicted in Figure 1, the three key components of the RCL are: the Resource Control Agent (RCA), the Admission Control Agent (ACA) and the End-User Application Toolkit (EAT).

The *Resource Control Agent (RCA)* represents the ultimate principle of the domain concerning the management of network resources. In order to simplify the efficient handling of resources, the RCA makes use of the concept of Resource Pools (RPs), each one of which concentrates on a particular sub-area of the network. Therefore, the Resource Pools build a tree

[1] The reader is referred to [12] for a thorough presentation

hierarchy, following the network topology, in the sense that the root of the tree is in charge of the available resources in the whole network, while each leaf is associated to one edge router (ER) of the network. The Resource Pools are initially assigned with a specific portion of resources, according to traffic load forecasts and results retrieved by measurements, and they also deploy an intelligent load-balancing algorithm for the re-distribution of resources among them, in case the initial distribution does not reflect the actual traffic load [13].

The *Admission Control Agent* mainly performs user authentication and authorization, reservation handling, and admission control. Policing and admission control are performed only at the edges of the network. Therefore, in the context of the core functionality of the ACA, the corresponding ingress and egress points (ingress-egress ACAs) of the flow are identified, and the resources assigned in the respective RPs are checked to ensure that the new flow can be accepted. Upon a successful reservation request, the corresponding ACAs consequently configure the edge routers appropriately to accommodate the new flow.

Reservation requests are forwarded to the ACAs from the *End-User Application Toolkit*, which mediates between end-users or applications and the network. The EAT interacts with the ACA to be aware of the available network services. A reservation request specifies the flow identifiers, the selected network service and the expected traffic profile for it. Special support is foreseen for legacy applications as well as for end users that are not aware of traffic description details, through the concept of Application Profiles.

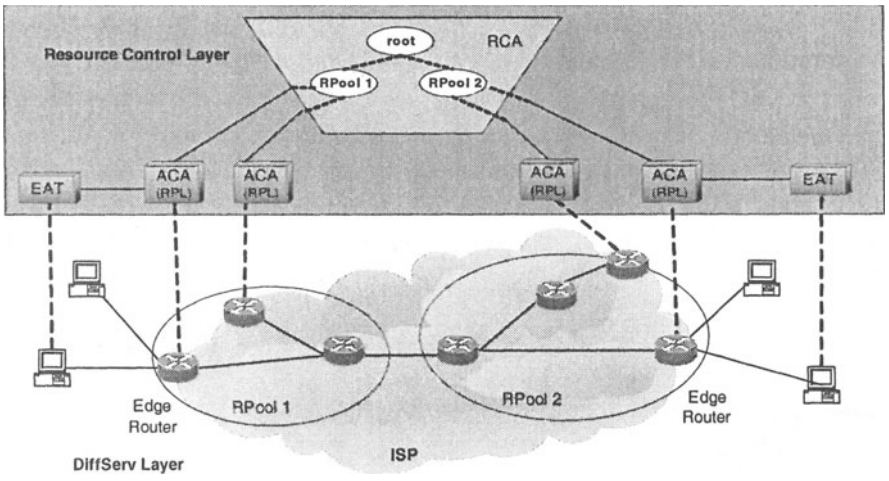

Figure1. The inter-domain architecture: RCL Layer

2.1 Requirements for the inter-domain resource control entity

The RCL architecture is responsible for handling the resources of a single domain. Resource control between domains has some major differences to resource control within a domain. This enforces the need to establish new mechanisms to handle the inter-domain case.

From the one hand, a domain's topology is built and controlled by a single network operator. Therefore, abstraction mechanisms like the resource pools can successfully be used to provide a coarse view of the topology. The Internet topology however, is based on a set of bilateral agreements between network operators. Moreover, scalability is an issue both for intra-domain and inter-domain resource control. One can handle single flows within a domain with an approach using distributed processing (one ACA per edge device). For a scalable inter-domain resource control however it is necessary to aggregate flows [9].

Taking into account these requirements, a different resource control mechanism used between domains is developed, based on the proposal made in [9]. In particular, the *BGRP Agent* is responsible to communicate reservation requests between domains that can be locally controlled by RCAs. A BGRP agent is associated with each border router and interacts with the egress and ingress ACAs, respectively, of the neighbor domains. The basic idea of BGRP is the aggregation of reservations along the sink-trees formed by the BGP routing protocol. The BGP routing protocol is characterized by the property of forwarding all packets for the same destination Autonomous System (AS) to the same next hop AS, which guarantees the formation of a sink-tree for each destination AS. All traffic destined for the same AS travels along the branches of this tree towards the root, while reservations for the same destination AS are aggregated at each BGRP agent. Therefore, a hop-by-hop reservation, based on the BGP routes, is used to follow the bilateral agreements of neighbored domains.

Apart from the sink-tree based aggregation of reservations, which is the base for scalability, the signaling load should also be controlled and reduced. Early reservation responses are used to control the amount of signaling traffic, while open interfaces allow this reservation mechanism to be used in conjunction with any kind of intra-domain resource control, which can provide edge-to-edge QoS.

The two-leveled logical overlay network above the underlying core network is depicted in Figure 3. This illustrates the main components of the overall architecture and their communication interfaces.

3. THE BGRP PROTOCOL

3.1 Overview

When resource reservation mechanisms will be deployed in a large-scale environment, it is imperative that scalability issues are taken into utter consideration. The problem particularly arises in large transit domains where the number of simultaneous reservations processed at the domains' routers may become extremely high and thus prohibitive with regard to the CPU processing, memory and link bandwidth requirements. More precisely, the number of reservations grows like $O(N^2)$ with N being the number of autonomous systems (AS) in the internet. It is therefore evident that individual handling of each flow is considered not to be a viable solution when applicable to inter-domain QoS reservation schemes. In essence, the need for aggregation is made apparent, no matter the granularity, for addressing effectively the aforementioned constraints.

For Internet routing, BGP provides a mechanism, which is independent of the routing protocol used within domains. However, it co-operates with intra-domain routing, so that loop-free end-to-end routing paths can be guaranteed. A similar approach is necessary for resource reservation: BGRP is a mechanism for Internet-wide resource reservations, which is independent, but co-operating with resource control mechanisms within each domain.

The BGRP approach [9] proposes the aggregation of reservations on the basis of the destination AS. This functionality is closely related to the property of the BGP routing protocol that enables the creation of sink-trees (see Figure 2) while domains trace their route towards a particular AS. Consequently, reservations are aggregated along the sink-trees created by the BGP protocol [10,11], thus limiting to a great extent the number of active reservations maintained at the routers to a factor of O(N). The BGRP protocol operates between the border routers of an AS or between entities associated to them.

It mainly uses three messages, i.e. PROBE, GRAFT and REFRESH messages. A PROBE message is the one initiated by the source domain and keeps being forwarded between BGRP speaking entities (BGRP agents) hop-by-hop until it reaches the destination domain (root of the sink-tree). On its way towards the root domain, the path of AS domains traversed by the PROBE message is recorded within it and resource availability within each domain is checked upon. When the PROBE message reaches its destination, a GRAFT message is generated containing an identification of the destination domain (sink-tree id). The GRAFT message travels back to the source along the recorded path. Each BGRP agent belonging to this path,

after processing the GRAFT message, aggregates the reservation with the existing ones pertaining to the same sink-tree and reserves the requested resources. Finally, REFRESH messages are exchanged regularly between BGRP agents with the aim of preserving the reservation state established at the corresponding routers.

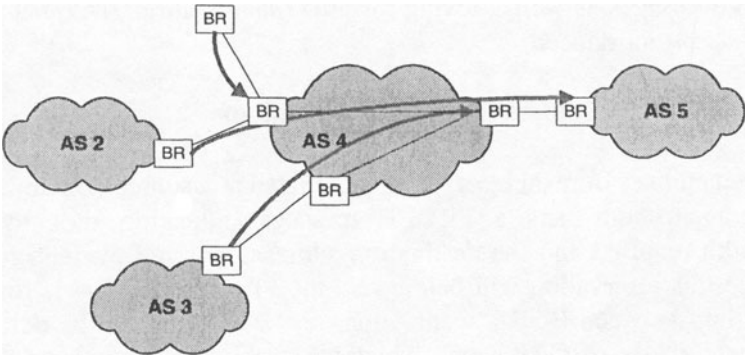

Figure2. Sink-tree creation

In this way, BGRP mainly addresses the path length that each message has to travel, alleviating to a greater extent the scalability problem. In Figure 3 the exchange of BGRP messages is illustrated over the intra-domain RCL architecture.

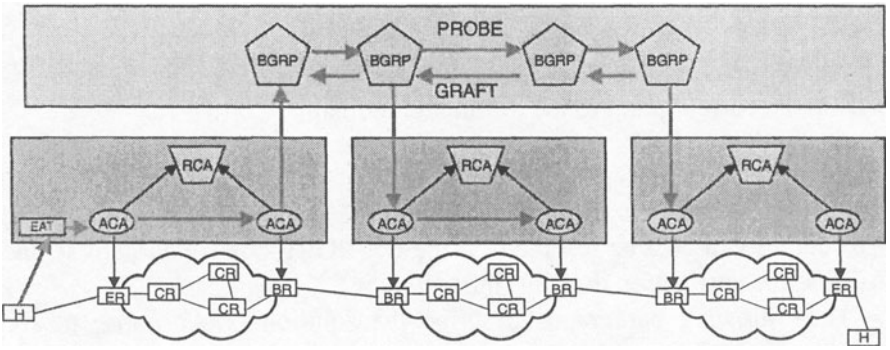

Figure 3. The end-to-end architecture

3.2 Quiet Grafting Mechanism

As described in the previous chapter, the fundamental concept of BGRP is the aggregation it performs on destination-based inter-domain reservations and thus resulting to the formation of sink-trees. It is therefore evident that

pure BGRP addresses the problem of scalability to a certain extent, i.e. it significantly reduces the amount of control state information kept at routers in conjunction with the volume of REFRESH messages needed for the maintenance of the corresponding state.

Of interest to us, however, is enhancing BGRP with additional mechanisms so as to achieve a reduction in the number of PROBE and GRAFT messages as well. Moving towards this direction, the quiet grafting mechanism is introduced.

3.2.1 Raised Problems & Proposed Solutions

When an inter-domain reservation is initiated at a source domain, the first BGRP 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 BGRP agents along the BGP route to the destination. Obviously, the last BGRP 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 BGRP agent to answer a PROBE message successfully with a positive response, the following conditions have to be met:

- The BGRP agent must be able to identify the sink-tree, to which the reservation belongs.
- The BGRP 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.
- As the last BGRP agent may no longer be informed about a new reservation (since the PROBE message is terminated earlier), the BGRP agent must provide means to contact the destination domain, so that resources can also be reserved on the not-BGRP-controlled path segment inside the destination domain (intra-domain).

The following paragraphs describe the solutions we propose to solve these requirements.

3.2.2 Early sink-tree identification

In order to terminate a PROBE message earlier on the way to the destination domain, an intermediate BGRP agent must be able to determine the sink-tree, to which this reservation belongs. Sink-trees collect all traffic identified by the network layer reachability information (NLRI) announced

by that root. They are identified by the AS number of the destination domain and an identification of the border router in the destination AS (entry point).

The proposed solution is to back-propagate all information necessary to identify the sink-tree with the GRAFT message. All BGRP agents store this information and use it to identify the sink-tree for subsequent reservations before the PROBE message reaches the final destination domain. BGP routes cannot be used for this purpose, because BGP may aggregate several routes. Sink-trees, however, may not be aggregated.

It is evident that storing this information causes additional memory usage. For each sink-tree with active reservations, the NLRI must be stored in addition to any other data. Provided, that the number of reachable subnets per AS can be statistically described by a fixed factor, this is however a linear increase in memory usage and therefore does not affect scalability. As with all reservation information, also the NLRI information is periodically updated using the REFRESH message. In this way, changes in the NLRI are propagated to all affected BGRP agents.

3.2.3 Hysteresis for the creation of Resource Cushions

Irrespective of the identification of a sink-tree, the quiet grafting of a new reservation will not be feasible if there are not enough resources so as to accommodate the new request. Therefore, it is required that BGRP nodes are assigned with more bandwidth than currently reserved. To accomplish that, over-reservation, quantisation and hysteresis techniques are likely to be employed.

The first two approaches are likely to introduce some limitations with respect to their impact on the Quiet Grafting probability and can inadvertently produce the opposite effects to the ones envisioned. Over-reservation, when performed without control, can lead to a reduction of the Quiet Grafting probability. This is due to the fact that future requests, coming from BGRP nodes other than the ones with over-reserved resources, are likely not to be grafted onto the tree or not to be accommodated at all. Another impact of over-reservation requests can be PROBE messages that travel unnecessarily towards the root of the tree even if the requested resources are available, operating as before at the expense of Quiet Grafting.

Quantization of the requested resources, i.e. rounding them up to a multiple of a chosen quantum, can lead to undesirable results as well. This is particularly true in the case of sink-trees that consist of many leaf nodes (N) and are moreover incapable, due to lack of future reservation requests, of using the remaining amount of the reserved bandwidth. As a result, the root of the sink-tree will end up having reserved at least N*quantum resources, far surpassing the actual needs. Thus, the adoption of the quantization

mechanism can only be justified for sink-trees with a limited amount of leaves and a much greater amount of reservation requests.

In essence, given the aforementioned concerns coupled with the complexity induced by the adoption of the corresponding techniques, it is proposed that the BGRP nodes do not perform over-reservation or quantization upon receiving a new reservation request. Instead, hysteresis on the release of resources is more appropriate for the formation of resource cushions at the BGRP nodes. The resource cushion mechanism that will be employed bases its resource release policy on the existence of the release period and thresholds.

When a BRPG agent receives a RELEASE message and it does not forward this message further downstream towards the root of the sink-tree, then it allocates resources downstream, which are not in use upstream. These resources, which are reserved downstream but not upstream of a BGRP agent (and which are collected through) due to retained RELEASE messages, are called resource cushions. A resource cushion is tied to a sink-tree. A BGRP agent may build resource cushions for all of its sink-trees. Building resource cushions has as an impact the reduction of the signaling load, because retained RELEASE messages reduce the signaling load of downstream domains. The use of resource cushions for arriving reservation requests further reduces the signaling load of downstream domains, when reservation requests are not forwarded but served from resource cushion immediately.

3.2.4 Signalling in the last domain

The quiet grafting mechanism inhibits the forwarding of signaling messages to the destination domain, since these are already answered at some intermediate stage. This has as impact that the last domain is unaware of a new reservation or release request.

Specifically, resource reservations are carried out or pre-reserved resources are used up to the ingress border router of the destination domain. Therefore, only the ingress border router of the last domain has reserved resources, while no resource reservation is performed on the path from the ingress border router to the egress edge router, which is connected to the destination host. In order to enable though edge-to-edge QoS-aware services, it is necessary to reserve resources within the last domain, which is achieved by allowing for a direct communication between the originating and the last domain.

In particular, each domain should provide a standardized interface to its intra-domain resource control entity. Therefore, enhancing the information carried by the GRAFT message can solve the problem of signaling in the last

domain. It is actually proposed to back–propagate a reference to this interface as well as the IP address of the ingress border router of the last domain within the GRAFT message so that this information is stored at each intermediate BGRP agent. In this way, the source intra-domain resource control entity retrieves the reference of the destination's intra-domain entity, which is responsible for reserving resources within the last domain. Consequently, a direct communication between the two domains is achieved and then, the initiating domain explicitly requests the resources in the destination domain, if necessary.

This reference is also necessary for the release of a reservation: since resources are not immediately released, 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.

4. CONCLUSIONS

The overall architecture presented in this paper addresses the problem of QoS provisioning in IP networks, providing a scalable approach for inter-domain resource reservation. However, the efficiency of an inter-domain protocol requires also an intra-domain architecture capable of providing QoS guarantees. The introduced RCL architecture is considered as a reference intra-domain architecture, which is responsible for the handling of the reservation requests, performing policy-based admission control, configuring the network in a top-down approach, managing the network resources and dynamically redistributing them among the network elements.

The main focus of the paper thought, is the description of the BGRP protocol framework, providing sink-tree based aggregation for resource reservations over a number of different AS. However, BGRP is just the first step towards the scalability. Aiming at reducing the number of signaling messages, the quiet grafting mechanism has been introduced. A set of problems has been identified and solutions were proposed and analyzed. More specifically, a proposal for the identification of the sink tree was investigated, a resource management scheme was presented and a solution concerning the signaling in the last domain was provided.

5. ACKNOWLEDGMENT

This work was performed in the framework of IST Project AQUILA (Adaptive Resource Control for QoS Using an IP-based Layered

Architecture - IST-1999-10077) [14] funded in part by the EU. The authors wish to express their gratitude to the other members of the Aquila Consortium for valuable discussions.

6. REFERENCES

[1] R. Braden, D. Clark, S. Shenker, "Integrated services in the Internet architecture: an overview", RFC 1633, 1994.

[2] D. Black, S. Blake, M. Carlson, E. Davies, Z. Wang, W. Weiss, "An Architecture for Differentiated Services", RFC 2475, 1998.

[3] K. Nichols, V. Jacobson, L. Zhang, "A Two-bit Differentiated Services Architecture for the Internet", RFC 2638, 1999.

[4] Y. Bernet, "The Complementary Roles of RSVP and Differentiated Services in the Full-Service QoS Network", IEEE Communications Mag., Vol. 38, No. 2, Feb. 2000.

[5] G. Eichler, H. Hussmann, G. Mamais, I. Venieris, C. Prehofer, S. Salsano, "Implementing Integrated and Differentiated Services for the Internet with ATM Networks: A Practical Approach", IEEE Com. Mag., January 2000.

[6] Y. Bernet, R. Yavatkar, P. Ford, F. Baker, L. Zhang,. "A Framework for Integrated Services Operation over Diffserv Networks", RFC 2998, 2000.

[7] Neilson, R.; Wheeler, J.; Reichmeyer, F.; Hares, S.: A Discussion of Bandwidth Broker Requirements for Internet2 Qbone Deployment. Internet2 Qbone BB advisory Council, August 1999.

[8] QBone Bandwidth Broker Architecture, Work in Progress, http://sss.advanced.org/bb/bboutline2.html.

[9] BGRP: Sink-Tree-Based Aggregation for Inter-Domain Reservations Ping P. Pan, Ellen L. Hahne, and Henning G. Schulzrinne, KICS 2000.

[10] Y. Rekhter and T. Li, "A border gateway protocol 4 (BGP-4)," Request for Comments (Draft Standard) 1771, IETF, March. 1995.

[11] P. Traina , "BGP-4 Protocol Analysis", Request for Comments (Draft Standard) 1774, Internet Engineering Task Force, March 1995.

[12] M. Winter et al., "System architecture and specification for the second trial", AQUILA deliverable D1202, September 2001.

[13] E. Nikolouzou, G. Politis, P. Sampatakos, I. Venieris, "An adaptive algorithm for resource management in a differentiated services network," Proc. of *IEEE ICC2001*, Helsinki, Finland, June 2001.

[14] Aquila project: http://www.ist-aquila.org.