

An Incremental QoS Routing Framework for Multihomed Stub AS based on Distributed Overlay Entities and QBGP¹

Marcelo Yannuzzi, Alexandre Fonte, Xavier Masip-Bruin, Edmundo Monteiro,
Sergi Sánchez-López, Marília Curado, Jordi Domingo-Pascual

Contact Author:

Marcelo Yannuzzi
Advanced Broadband Communications Center
Technical University of Catalunya
Av. Victor Balaguer s/n
08800 Vilanova i la Geltru, Barcelona
Catalunya, Spain
Tel.: +34 93 896 77 67
Fax: +34 93 896 77 00
E-mail: yannuzzi@ac.upc.edu

Keywords: Interdomain, Multihoming, QoS Routing, QBGP

Abstract

This paper proposes a novel and incremental approach to interdomain QoS Routing. Our approach is to provide a completely distributed architecture and a routing control layer operating at the edge of the Internet for dynamic QoS provisioning, and to use QoS extensions and Traffic Engineering capabilities of the underlying BGP layer for static QoS provisioning. The proposal is incremental not only in the sense of how static and dynamic QoS is managed, but also in the sense that the routing control layer smartly exploits the capabilities of BGP with the aim of improving end-to-end performance in short timescales. Our focus is mainly on influencing how traffic is exchanged among non-directly connected multihomed stub Autonomous Systems based on specific QoS parameters. We provide evidence supporting the advantages of our approach by means of simulation.

¹ This work was partially funded by the MCyT (Spanish Ministry of Science and Technology) under contract FEDER-TIC2002-04531-C04-02, the CIRIT (Catalan Research Council) under contract 2001-SGR00226 and the European Commission through Network of Excellence E-NEXT under contract FP6-506869.

1. Introduction and Motivations

At present, most emerging network services claim for highly efficient and cost-effective mechanisms to provide different levels of end-to-end Quality of Service (QoS). Such mechanisms should address the issue from two very different perspectives, to be precise, from the intradomain and from the interdomain standpoints. In this latter case, it is widely accepted that the main problem with end-to-end QoS provisioning is on the very foundations of the current interdomain network paradigm. This paradigm is based on a highly scalable and completely distributed network architecture, which relies on the Border Gateway Protocol (BGP) [1] as the glue that keeps the Internet together. The central issue is that BGP has not inbuilt QoS capabilities given that it was designed with very different goals in mind by the early nineties. Although some researchers have proposed to replace BGP, in practice, only incremental solutions are realistic and will have chance to become deployed, especially from the economical perspective given the massive deployment of BGP in today's Internet. Such solutions should then set their focus on the design of innovative interdomain network paradigms, interdomain network architectures, and interdomain routing protocols aiming at complementing the deficiencies of BGP rather than replacing it, and hence, this is the approach we follow in this paper.

A central part of the problem of end-to-end QoS provisioning consists of finding a path that simultaneously satisfies q independent QoS constraints. This Multi-Constrained Path problem is the focus of QoS Routing (QoSR), and it has been proven to be NP-hard when the number of constraints on multiple additive or multiplicative metrics is $q \geq 2$ [2]. Therefore, several researchers have contributed with innovative heuristics to find suboptimal solutions to this problem, but with the advantage of being computable in polynomial time. Most of these heuristics fall into the intradomain routing area, but some years ago the issue also started to become surveyed at the interdomain level. In this latter case, the complexity increases significantly mainly due to the fact that though a suboptimal path may be found, stringent end-to-end QoS guarantees demands for interdomain resource reservation. This reservation requirement essentially reveals the connection oriented nature of QoS, but following such an approach at the interdomain level, imposes at least at present, several tough challenges in practical terms. As an alternative, it is possible to conceive dynamic end-to-end interdomain QoS without any kind of resource reservation, and to follow the IP connectionless paradigm, as long as only soft end-to-end QoS is guaranteed (without resource reservations). This is in fact the approach that we follow in this particular work. The main advantage of this kind of approach is that it is indeed much more cost-effective than any interdomain reservation-based approach, and depending on the frame of the proposal and its implementation, it may also result highly efficient. An additional reason to seek for incremental and cost-effective end-to-end QoS solutions is that most end customers are indeed not willing to invest too much on behalf of getting much better performance from the network than with traditional best-effort traffic.

Above all, multihomed stub Autonomous Systems (ASs) are those which are in need of novel mechanisms that allow them to rapidly manage and distribute their interdomain traffic in order to improve their performance. This particular fraction of ASs crowds together mostly medium and large enterprise customers, Content Service Providers (CSPs), and small Network Service Providers (NSPs). It is worth highlighting that nearly 80% of the total number of ASs that currently compose the Internet are stub, and the majority of this fraction is in fact multihomed. Therefore, the blast of multihomed stub ASs has gained huge interest in both research and commercial fields in the last few years.

Compelling recent studies like [3] demonstrate that the problem of tracking and controlling most of the traffic of multihomed stub ASs turns out to be impracticable. This is because the large variability of the topological characteristics of interdomain traffic, in addition to the limited aggregation of this traffic, indicates that the number of paths to be tracked and controlled is not only highly variable, but also really large. Despite these variability and lack of aggregation issues, several recent studies also show that a very small number of invariant paths are still responsible for a significant fraction of the existing traffic [3, 4]. By invariant we mean that these paths are stable, i.e., they are, typically, permanently present in the BGP tables and hence are not affected by the variability issues mentioned before. For instance, the measurements conducted in [3] reveal that only six invariant AS paths carried about 36% of the one-month total traffic of a real multihomed stub AS, which is indeed a significant fraction of the overall traffic. Thus, a realistic approach is that it is possible to track and manage an important portion of the existing traffic by simply controlling a reduced set of stable paths (typically 6-8 paths), and hence, this is the approach that we follow in this paper. A central issue however, is that multi-connectivity to the Internet does not necessarily guarantee improved end-to-end path diversity. This is due, first, to the fact that BGP only advertises the best path it knows, so BGP considerably prunes the total number of available paths between distant ASs, and second, to the topological characteristics of the Internet at the AS-level. Even though several studies have addressed these scarce path diversity issues [5, 6], recent studies like [7] demonstrate that in practice multihoming in combination with QoS tools are powerful techniques to improve end-to-end performance. A sound explanation for this is that the Internet core is in effect an over-provisioned network.

Another important point is that cooperation among remote ASs is expected for a number of reasons. First, recent studies show that interdomain traffic is mainly exchanged among ASs that are not directly connected. Instead, they are typically 2, 3 and 4 AS hops away [4], and this also applies to the reduced set of paths to be tracked and controlled that we mentioned before. Second, cooperation between remote ASs is expected for both performance and economical reasons. Indeed, such scenario is perfectly suitable, for example, for medium and large enterprises with multi-connected offices in different countries. In such a case, what each multi-connected office primarily needs is to improve the performance but for the relevant traffic exchanged with other enterprise's sites. Once again, the number of relevant paths to be tracked and controlled per-site will be typically small. For this reason, strong manufacturers such as Cisco have recently announced that they will go in this direction [8].

The combination of all the aforementioned features made us focus in this paper on a cost-effective and incremental QoS framework among a reduced set of strategically selected and non-directly connected multihomed stub ASs. The proposal is certainly not limited to this case and it could also be applied if the ASs share EBGP (External BGP) peering links. However, as described before this is typically not the case.

The remaining of the paper is organized as follows. Section 2 cites some relevant related work and introduces our QoS framework. After that, Section 3 overviews the different elements composing this framework, such as the architecture and the routing control layer. Then, Section 4 surveys the main functionalities of each element in the proposed framework, and Section 5 presents our simulation scenario and evaluation results. Finally, Section 6 concludes the paper.

2. An Incremental QoS Framework

Several research efforts are being carried out in order to address the issue of QoS provisioning at an inter-domain level. These efforts have contributed with solutions that could be primarily divided into two different groups. The first group gathers together solutions trying to enhance BGP with new capabilities, such as Traffic Engineering (TE) and QoS extensions. Some compelling proposals in this area can be found in [3, 4] [9]. All these in-band solutions, i.e., solutions intrinsically supported and signaled using BGP, are able to supply significant improvements in terms of rather static QoS or TE provisioning. However, they are inadequate to handle the rearrangement of interdomain traffic in short timescales. The reasons for this are that: (a) BGP is a slow reacting protocol [10]; (b) such approaches would significantly increase the number of messages exchanged between BGP peers, which may lead to network instabilities [11]. Thus, these solutions are not able to cope with the current demands of multi-connected stub ASs.

The second group is conformed by solutions tending to decouple the complexity of QoS and TE provisioning from BGP devices. These out-of-band solutions, i.e., solutions not intrinsically supported and signaled using BGP are able to operate in short, and even very short timescales. These out-of-band solutions can be in turn divided into two different types of approaches.

On the one hand, we find overlay networks which are capable of improving end-to-end QoS by circumventing BGP routing. Some strong proposals in this area can be found in [12-18]. However, even though overlay networks enhance in several ways end-to-end performance, none of the current proposals is capable of tackling down the central issues of end-to-end QoS provisioning at the AS-level. While some of the already cited proposals are centralized and rely on highly complex heuristics [12], others are distributed but definitively not scalable [15], or need massive deployment in order to be able to operate (at least one overlay node per-AS in every AS). In effect, none of the cited proposals has been implemented in real practice at the AS-level. Furthermore, Akella et al [7] have recently compared pure overlay routing to multihoming route control, and their results clearly show that it is not necessary to circumvent BGP routing to achieve good end-to-end performance. This is perfectly aligned with our first motivation in the previous section (to follow an incremental approach).

On the other hand, we can find route optimizing tools. Two types of route optimizers are currently emerging, namely, DNS-based optimizers [19], and BGP-based optimizers [8, 20]. The DNS-based solutions are addressed to small organizations which are not willing to deal with the difficulties of BGP peering and management, and hence are out of the scope of this work. Current BGP-based route optimizers are quite limited. The available proposals consist of standalone selfish devices, so no cooperation between ASs could be granted (we have already presented the motivations for this cooperation in Section 1). Another important issue is that the implications of rearranging interdomain traffic in very short timescales, but magnified by the number of sources simultaneously injecting these selfish perturbations to the network are completely unpredictable in terms of global stability [21].

Therefore, none of the state-of-the-art proposals is able to supply a cost-effective, incremental, and collaborative approach able to improve end-to-end QoS in short timescales at the AS-level. In order to address these open issues, we propose in this paper a novel and incremental QoS framework for multihomed stub ASs. This framework is based on a distributed architecture and a routing control layer operating at the edge of the Internet for dynamic QoS provisioning, but taking advantage of and reusing QoS extensions and Traffic Engineering

(TE) capabilities of the underlying BGP layer for static QoS provisioning. The proposal is incremental in two ways. On the one hand, because of the way we control how static and dynamic QoS is managed. On the other hand, because the distributed control layer we propose smartly exploits the capabilities of BGP with the aim of improving end-to-end performance in shorter timescales. In other words, our proposal does not circumvent BGP routing with the intention of dynamically improving end-to-end QoS such as a legacy overlay network will do.

Our focus in this work is mainly on influencing how traffic is exchanged among remote multihomed stub ASs in order to improve the performance for delay-sensitive applications such as Voice over IP, or video streaming. We took this approach: (a) in order to validate our ideas; (b) because these are major applications which are perfectly suitable for collaborative edge routing schemes (such as medium or large enterprise customers).

The distributed architecture which handles how dynamic end-to-end QoS is provisioned consists of special Overlay Entities (OEs), which in essence implement all the functionalities of the QoS control layer. At least one OE is needed in any multi-connected stub AS participating of our QoS framework. The Figure 1 may help to understand the architecture we are proposing. These OEs within two non-peering multihomed ASs are able to: (a) exchange a Service Level Specification (SLS) regarding the traffic among them; (b) examine the compliance with the SLS; (c) and configure on-the-fly BGP in order to improve end-to-end QoS for a given set of Classes of Service (CoSs).

These functionalities allow the OEs to influence the behavior of the underlying BGP routing layer, by taking rapid and accurate decisions to bypass network problems such as link failures, or service degradation for a particular CoS. The reactive nature of this control structure acts as a complementary layer conceived to enhance the performance of the underlying BGP layer containing both QoS aware BGP (QBGP) routers and non-QoS aware routers. Our research goal and so the scope of this paper is on the most general framework with focus on the routing control layer, and its control over the underlying BGP layer.

3. Overview of the architecture and the routing control layer

As stated in Section 2, we propose in this paper a combined QBGP and control architecture for interdomain QoS. The main challenges behind the routing control architecture are:

- The OEs should respond much faster than the BGP layer in the case of a network failure.
- The OEs should react and handle how to tweak BGP in order to reroute traffic when non-compliant conditions concerning QoS parameters previously negotiated for a given CoS are detected.
- The underlying BGP structure does not need any modification, and remains unaware of the distributed routing control architecture conformed by the OEs.

The Figure 1 depicts a possible scenario were our proposal could be applied. In this model, two peering OEs belonging to different ASs spanning across several AS hops are able to exchange a SLS and agree upon a set of soft QoS parameters concerning the traffic among them. We assume that cooperation between these ASs is desired (for monetary and/or performance reasons).

This scheme has several advantages:

- First, with only a very small number of OEs, but located at advantageously selected remote multi-homed sites is enough to control a significant portion of the traffic of an AS. As we have shown in Section 1 typically only 6-8 OEs will be needed.
- Second, the intermediate ASs do not need to participate of the control architecture, and hence no OEs or any kind of modifications are needed in any transit ASs connecting the remote ASs in our scheme. Therefore, the complexity of dynamic QoS provisioning is pushed to the edge of the network by means of a completely distributed architecture.
- Third, our approach is that an OE within a source AS dynamically manages BGP in order to control the allocation of its outbound traffic towards a remote AS in our scheme, depending on the network conditions and QoS constraints for each CoS. This allows tweaking BGP even in very short time-scales given that no BGP messages will be ever spawned. Thus, instead of proposing a complex scheme to dynamically and accurately manage how traffic enters a target AS, we focus on a collaborative scheme which handles how the traffic exits from the source AS.

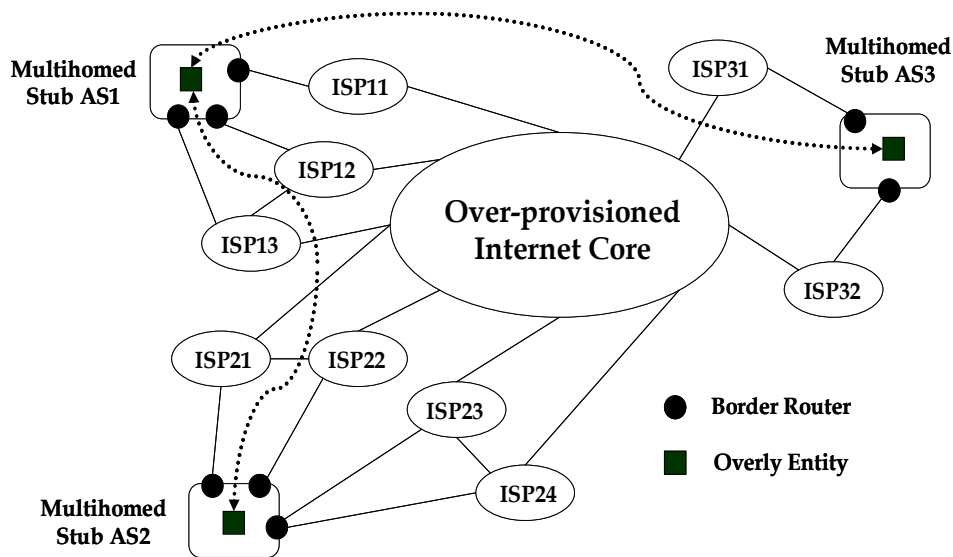


Fig.1. Illustrative example of the proposed interdomain QoS framework

Then, within each edge AS in our scheme, the OE should measure end-to-end QoS parameters along every link connecting the multihomed AS to the Internet, and check for violations to the SLSs. Henceforth, we assume that the topology has at least some diversity between any pair of remote ASs participating in our QoSR model. Furthermore, when a violation to the SLS is detected for a given CoS, the OE in the source AS is capable of re-configuring on-the-fly its traffic pattern to the remote AS for the affected CoS. Here, the time scale needed to detect and react to a certain problem is very small when compared with the BGP timescale [10].

The foremost motivations for influencing traffic in this way are rooted on what we have presented in Section 1. The essence in this approach is that the QoS perception between a pair of remote ASs is basically the one that the OEs have of each other.

In order to gather the QoS information and exploit multihoming the OEs are endowed with mechanisms to spawn small probes targeting the reduced set of remote ASs in our scheme through every available link connecting the source AS to the Internet. We sustain that the AS-AS probing practice is not demanding neither in terms of traffic nor in terms of processing, as long as the number of remote peering ASs and the number of CoSs remains limited. To fix ideas, typically each OE will only need to probe 6-8 remote OEs (see Section 1), for 2-4 CoSs, and through 2-3 egress links. It is worth mentioning that inside each AS in our scheme, the OEs and their respective egress BGP routers are set and configured to bypass the hot potato routing issues for the OE-to-OE communication. Thus, each OE could be configured to probe a remote OE for the relevant CoSs between each other, through all the available egress links of its AS.

Following the recommendations in the literature, we used a Pseudo-Random Poisson Process to generate the probes [22, 23]. In this process N_u random sampling times uniformly distributed are generated over consecutive intervals of duration dT_u . The parameters N_u and dT_u are selected so that $N_u / dT_u = \lambda$ where λ^{-1} is the average sampling time of a classical Poisson process. This approach is typically chosen to avoid the occasional lengthy spaces between sampling times that a classical Poisson process could create, since this may be unacceptable for many real-time applications [22, 23]. Then, the probability distribution function for the sampling times is given by:

$$F(t) = \text{Uniform} \{N_u, t \in [ndT_u, (n+1)dT_u]\} \quad \forall t \geq 0 \wedge \forall n \geq 0 / n \in Z \quad (1)$$

We made the following decisions while designing the probe packets. Firstly, each probe belongs to the same CoS that it is probing. For instance, in a QBGP framework based on Differentiated Services (DiffServ), when probing a particular CoS which is mapped to an Assured Forwarding (AF) class in the intermediate AS, the probes are tagged under the same AF class [24]. Second, as recommended in [22] we used a random padding technique to generate fixed sized probes avoiding that the measurements were influenced by the size of the samples. In addition, we set the size and frequency of the probes to correlate the measurements with the class of traffic (application) being controlled.

It is worth noting that based on these QoS measurements, the detection of a non-complying condition with a previously established SLS usually is due to a bottleneck occurring only in a single direction of the traffic. This means that the bottleneck is merely on the upstream or the downstream path between the remote ASs. For instance, in Fig. 1 if the OEs in AS1 and AS2 measure the same one-way QoS parameter, such as One-Way Delay (OWD) [22], and react in the same manner, then either of them is able to independently decide if it should shift its outbound traffic or not. An advantage of this approach is that BGP updates could be completely avoided if, for example, the LOCAL PREFERENCE (LOCAL_PREF) is used when reallocating this traffic. Indeed, this is the approach we will use for the dynamical part of our QoSR framework.

4. Main functionalities of the routing control layer and the QBGP layer

In this Section we describe in detail the main functionalities of both, the routing control layer and the BGP routing layer.

A. Functionalities of the routing control layer

This layer is composed by a set of OEs:

A.1) Basic set of components of the routing control layer:

- At least one OE exists per domain taking part in our QoSR framework
- An OE has full access and management capabilities on the border BGP routers within its AS
- An OE has algorithms for both detecting non-conforming conditions for a given CoS, and deciding when and how to reallocate its traffic

A.2) Main components of the routing control layer:

A Routing Control Protocol. A protocol between the OEs is required. This protocol allows OEs not only to exchange the SLSs with each other, but also to exchange substantial information for the routing control layer. This latter part of the information basically includes the transmission and treatment of the probes for the different CoSs, and the exchange of messages in order to open, sustain, and release connections between the OEs.

QoS Metric Selection. With the aim of improving end-to-end performance for delay-sensitive applications, and also in order to validate our approach we choose a simple QoS metric for the dynamical portion of our QoSR model. The metric we have selected is a Smoothed OWD (SOWD), which is given by:

$$SOWD_j(m, n) = Median\{OWD_j(m, k)\}, \quad k \in [n - N + 1, n] \quad (2)$$

In (2) n and m correspond to the n_{th} probe generated by a source OE and sent through the m_{th} egress link of the AS, and the sub-index j represents the CoS C_j . Then, the OEs compute a per-CoS cost to reach a remote ASs in our QoSR framework over every egress link m based on the previous metric. This SOWD corresponds to the median OWD through a sliding window of size N . Instead of using instantaneous values of the OWD, we propose to use this filter, which smoothes the OWD avoiding rapid changes in our metric. A trade-off exists in terms of the size of the window. A large value of N implies a slow reaction when network conditions change and maybe the reallocation of traffic is needed. On the other hand, small values of N could translate into frequent traffic reallocations since it is likely to occur that non-compliant conditions are more frequently met.

The motivation for choosing the median OWD is that this is an excellent estimator of the OWD that the user's applications are actually experiencing on the network [22]. Another estimator quite frequently used is the mean. However, the median has two important advantages when compared to the mean. First, the mean is much

more biased by a small number of outliers than the median. Second, computing a single QoS metric like the mean OWD through a sliding window needs an heuristic or a special treatment in case of losses. On the other hand, lost samples of OWD may be set to infinite without problem while computing the median through the sliding window.

Based on the cost metric selection in (2), the SLS exchanged by the OEs basically consists of the maximum SOWD tolerable for each different CoS C_j , which we call D_j .

Furthermore, given that the QoS measurements require the accurate computation of the OWD, we assume that the OEs are properly synchronized (e.g., by means of GPS) and the details concerning synchronization are out of the scope of this work.

Feedback Information. An important issue is that QoS measurements based on OWD imply that those measurements are performed by the OE on the destination AS. However, the dynamic reallocation of traffic is performed from the source AS. Thus, the source OE requires feedback from the remote OE. In order to avoid unnecessary messages traversing the network, we use the OE-to-OE protocol to provide the necessary feedback to the source. Our approach is to compute the SOWD in (2) always on the destination, and to send back feedback only when the SOWD changes with respect the previously advertised value. This scheme saves lots of messages back and forth the network, especially under steady network conditions. In such conditions, the instantaneous values of OWD will have a low dispersion. Thus, the median computed in (2) will be mostly the same during long time periods, and hence no feedback information will be sent. Even tough this saves signalling information and overhead for the routing control layer, it is not reliable to lack of feedback information for long periods of time. Therefore, the SOWDs are periodically refreshed by means of Hello/Keepalive messages which are exchanged among the OEs. This is done in order assure that any remote OEs are working and computing the median delays as expected.

Stability: Another central issue is that the traffic reallocation process should never generate network instability. In order to prevent this from happening, but keeping in mind that we follow a completely distributed architecture design where the OEs should rely on themselves to cope with these problems, we impose the following restriction:

“Traffic targeting a certain CoS C_j should never be reallocated over a link s , if and only if the primary link to reach C_j was s in $[t - T_h, t]$ or C_j has exceeded its maximum number of possible reallocations $\Rightarrow R_j(t) \geq R_j^{MAX}$ ”

In this way the parameter T_h avoids short-term bounces, while the parameter R_j^{MAX} avoids the long-term ones. Then, each time a traffic reallocation process takes place for a given CoS C_j the variable $R_j(t)$ is incremented. Our approach is to provide a sort of soft penalization similar to BGP damping [11], where the penalty is incremented by a fixed value P with each new allocation, but it decays exponentially with time when no reallocations occur according to:

$$R_j(t) = R_j(T) e^{-\left(\frac{t-T}{\tau}\right)} \quad (3)$$

The parameters T_b , R_j^{MAX} , P and τ are configurable parameters, whose values depend on the degrees of freedom in the number of short and long-term reallocations we allow for a given CoS C_j . An additional challenge in terms of stability arises when a path becomes heavily loaded, since several CoSs within the path could experience non-compliant conditions with their respective SLSs. In order to prevent simultaneous reallocations for all the affected CoSs, we endow the OEs with a contention mechanism which prioritizes the relevance of the different CoSs. Then, more relevant CoSs are reallocated faster than less priority classes. The contention algorithm operates as follows:

$$\begin{cases} \text{Let } C_j \text{ be one of the } q \text{ affected CoS within link } m, \text{ where } j = 1, \dots, q \\ C_j \text{ will be reallocated in } T_j, \text{ where } T_j \in [K_{j-1}, K_j) \text{ and } T_j \text{ is randomly selected, with } K_0 = 0 \end{cases}$$

\Rightarrow Then, the highest priority classes C_1 within link m will be reallocated in a random time $T_1 \in [0, K_1)$, classes C_2 will be reallocated in a random time $T_2 \in [K_1, K_2)$, and so on.

Clearly, our contention mechanism allows an OE to iteratively reallocate traffic from a loaded path, and to dynamically check if the remaining classes continue under non-compliant conditions. It is likely that as soon as we begin to extract traffic from the path, the remaining classes will start to experience better end-to-end performance. However, a different situation is generated when a path failure occurs. In this case, an OE should react as fast as possible to reallocate all traffic from the affected path. Then, a trade-off exists in terms of both the contention mechanism and the ability to rapidly redistribute all traffic from any given path. An alternative way to avoid tuning the contention algorithm to efficiently cope with both problems at the same time is to rely on the probing technique given that a path failure will cause the complete loss of probes for all the CoSs using the path. An approach could be to selectively increment the frequency of the probes when losses are detected for all CoSs. In such a case the increment in the frequency could be done for a short period of time and only with the aim of speeding up the re-routing process. Once a CoS is reallocated, the frequency of the probes should decrease back to its normal value.

B. BGP Routing Functionalities

The set of routes to be tested by the OEs using the probing techniques described in the previous subsection, are predetermined by the BGP layer. In this layer two types of devices can operate; legacy BGP routers and QBGP routers. A QBGP router is able to distribute QoS information and take routing decisions per-CoS constrained to the previously established SLS between different peering domains. In our model, QBGP routers can be seen as the practical tool to establish the overall static or pseudo-static interdomain QoSR infrastructure. Interesting approaches and further information on the subject of QBGP could be found in [9, 25].

C. Combined QoS Algorithm

The Figure 2 depicts our combined QoS algorithm. Let m be the egress link currently allocating traffic of class C_j . It is important to remark that the approach we follow is that even though an alternative path could have a better cost in terms of SOWD, we avoid reallocating traffic of class C_j from link m until a violation to the SLS is detected. Then, two distinct threads of events may occur depending if the arrival is of a probe (k,l) , or if the arrival is of feedback information. If a probe (k,l) is received, then (2) is computed locally for the corresponding CoS, and if the median changed then feedback needs to be sent to the remote OE. This is shown as (I) in Figure 2. On the other hand, if feedback information is received then the algorithm checks for violations to the maximum SOWD tolerable, that is D_j . This is shown as (II) in Figure 2. If no violations have occurred the algorithm simply waits for the next incoming feedback message. However, if a violation is detected in link m the algorithm checks if the maximum number of allowed reallocations R_j^{MAX} is exceeded. In case this is true, the local OE is able to compose a feedback message and warn the remote OE about this situation.

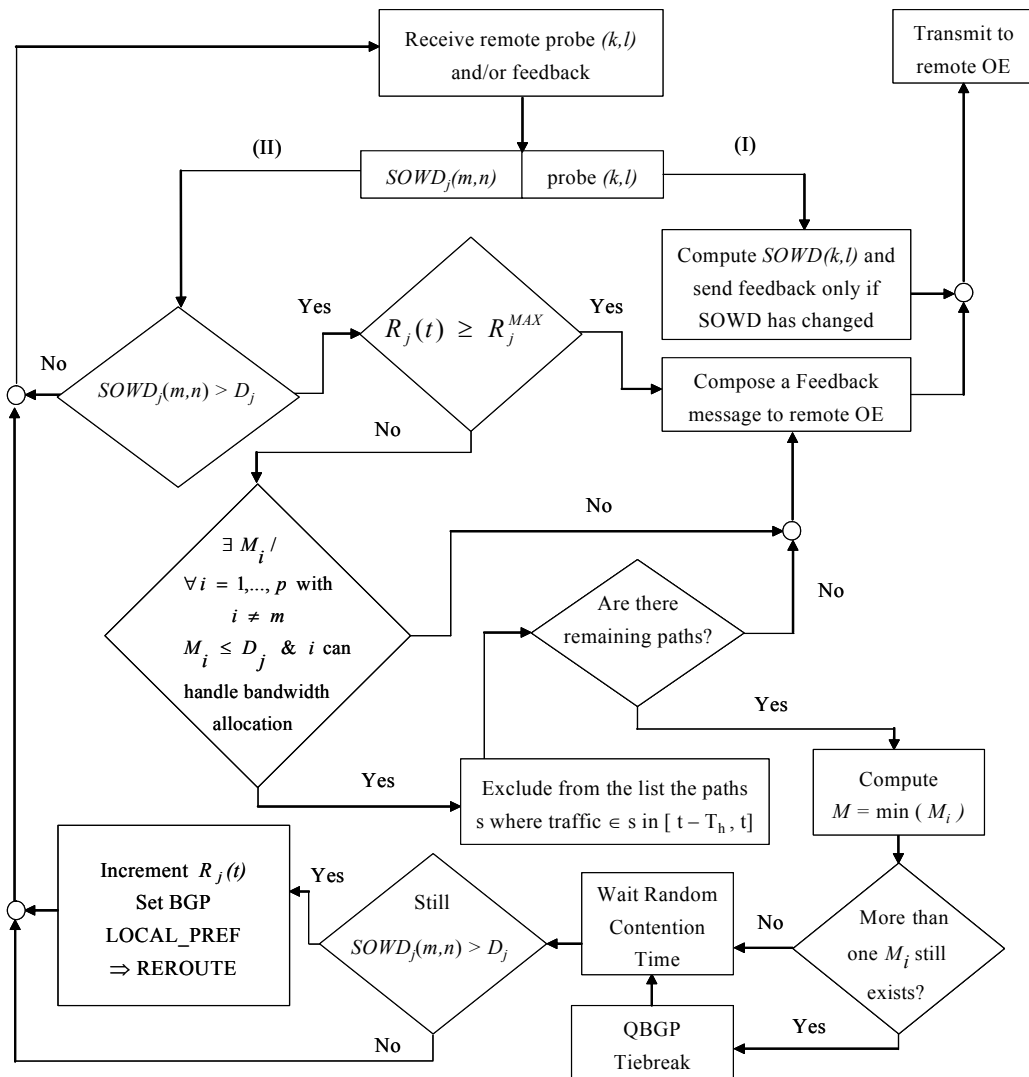


Fig.2. Combined QoS Algorithm

The main idea is that the feedback process provides information to the remote OE, and thus it can try to handle the problem by tuning its static QoS provisioning using either QBGP or TE-BGP. If R_j^{MAX} is not exceeded, then the OE needs to check, within all the external available links p , excepting m , if there exists at least one link i whose cost M_i satisfies the constraint for the class C_j . Moreover, it also needs to check if the link has enough room to handle the class reallocation. Subsequently, and in order to avoid any short term bounce, the OE excludes from the set of capable links those who had allocated traffic of C_j in $[t - Th, t]$. Once this is done, we rely on QBGP to tiebreak in case two or more links show the same cost in terms of the SOWD. At this step a single link is left as the target for the reallocation of the class. Then, the contention algorithm is executed and T_j seconds later the OE checks if the class still remains in a violating condition. If this is the case, the OE increments $R_j(t)$ by P and reroutes the traffic of C_j .

5. Simulation Results and Implementation Cost

We have conducted by simulation two sets of experiments to evaluate and validate the proposed architecture. The first goal is the validation of the initial assumption that our incremental approach, which uses a complementary routing control layer speeds up the reaction of the routing infrastructure in the case of link failures. Thus, as a performance indicator we chose to compare the response time to a link failure of QBGP to the one obtained using our complementary proposal. The second major goal is to study under variable QoS dynamics, how the OEs contribute to meet the SLS constraints for each CoS between two remote domains in our QoS framework. This is shown through the ability of the OEs to exploit the available paths in order to reduce the OWD for packets composing the QoS traffic aggregates. In addition, we survey the traffic transfer efficiency and also the number of path shifts needed to accomplish with the corresponding QoS constraints. As an indicator of the traffic transfer efficiency we use the parameter $Eff_{dj} = C_{dj} / C_{sj}$. This parameter allows to assess the traffic performance for each QoS class, where C_{dj} is the throughput at the destination d , and C_{sj} is the throughput at the source domain s for the class of service j .

Our simulations were performed using the J-Sim [26] simulator with the BGP Infonet suite [27] which is compliant with BGP specification RFC 1771 [1]. In this environment, we have implemented the functionalities of the QoS control layer. In order to allow the OEs to have full access to the Adj-RIBs-In and the Loc-RIB of a BGP router [1], and to have control over the BGP decision process, it was necessary to add some extensions to the Infonet suite. Furthermore, we have also included the following QoS BGP extensions:

- An optional transitive attribute to distribute the CoS identification (ID)
- A set of modifications to the BGP tables to allow the storage of this additional information, following a similar approach to the one described in [9]
- A set of mechanisms to allow BGP speakers to load the supported CoSs
- A mechanism to allow each local IP prefix to be announced within a given CoS
- A mechanism to allow BGP speakers to set the permissibility based on local QoS policies and supported capabilities.

As network topology for our simulations we use part of the GÉANT European Academic Backbone [28], as it is illustrated in Figure 3. This topology represents a typical scenario where the remote multihomed domains used are AS1 and AS2. The output links of the QoS aware BGP (QBGP) routers have the same capacity $C = 8.0\text{Mbps}$ and propagation delay $Pd = 3\text{ms}$, with the exception of AS2 links where, in order to have some bottleneck, the capacity chosen was $C/8$. The simulated network aims at being a multi-service network. Thus, all QBGP routers implemented the standard Per-Hop Behaviors (PHB): EF, AF21, AF11 and Best-Effort. On the ingress domain we used edge DiffServ capabilities to mark packets with a specific DSCP (DiffServ Code Point) depending on its corresponding service. These marks were applied both to regular IP packets, and to the probes generated. For complexity concerns, we modeled each AS as a single QBGP router.

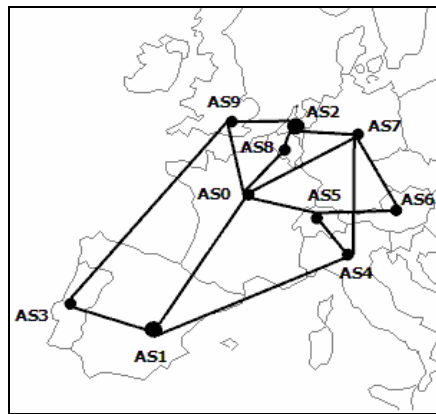


Fig. 3. Simulation topology

Our traffic mix between AS1 and AS2 consisted of up-to forty VoIP calls, up-to thirty video calls, prioritized data downloading, web browsing, and email downloading. New voice and video connections arrived at AS1 border router and the corresponding durations are uniformly distributed between $[300, 800]$, and $[0, 100]$ seconds respectively. Both kinds of connection calls are active along all the simulation run time, which is $[300, 900]$ seconds. The corresponding marking scheme and source models are the following [29]:

- i) EF Voice traffic is marked with EF value. EF source is an ON-OFF VoIP generator. The On and OFF states are characterized by an exponential random variable with mean 400ms, and 600ms respectively. In the on state, EF voice source generates traffic at a peak rate of 64kbps.
- ii) Video traffic is marked with AF21 value. AF21 source is a video traffic generator characterized by Pareto ON-OFF. The average of ON state of video traffic is 360 ms, and the average of OFF state of this source is set to 400 ms. In the on state AF21 source generates video traffic at a peak rate of 200kbps. All ON-OFF packets have size equivalent to 567 bytes.
- iii) Finally, the data prioritized traffic is marked with AF11 value. Both AF11 traffic and BE connections are characterized by a Poisson process. The size of packets generated by the data connections in these simulations is 1000 bytes.

Simulated time was always 900 seconds of which the first 300 seconds were discarded as “warm-up period”. This is the time needed for QBGP routers to advertise all the network reachability information so that paths become available to our OE at ingress domain. During the evaluations we configured the following default set of parameters: $R_j^{MAX} = \infty \forall j$, the sliding window parameter to $N = 8$, and the hold and contention timers to [EF: 2, AF21: 4, AF11: 6] seconds, and [EF: 5, AF21: 9, AF11: 13] seconds, respectively. The Pseudo-Random Sampling parameters were aggressively set to $N_u = 8$ and $dT_u = 36s$. The size of each probe packet is equivalent to the corresponding size of traffic packets.

In both experiments we run simulations separately for QBGP and OEs coupled with QBGP. We have considered a SLS exchanged between the remote domains based on maximum packet OWD of [35, 75 and 90] ms, for voice, video, and data prioritized traffic transferences respectively. These maximum OWDs tolerated per-service were heuristically chosen to be representative values of the kinds of traffic sources considered (voice, video and data prioritized).

In the first experiment, we compare the response time to a link failure between QBGP, and QBGP coupled with OEs. Fig. 4 depicts a set of plots for traffic of class AF11 showing the throughput measured at the destination, and the path shifts determined by changes in the next-hop for the source AS, namely AS1. From these plots, we can observe that a pure QBGP framework (without OEs running on AS1 and AS2) and configured with small keepalive and hold timer values in order to improve the reaction of QBGP (30 and 90 seconds respectively), needs about 90 seconds to overcome a link failure, but in the worst case only 45 seconds are needed when the OEs are running. This result validates our initial assumption. It is worth mentioning that this last value includes not only the implicit link failure detection condition based on a violation to the maximum SOWD tolerated, but also includes a random contention interval of 13 seconds before re-routing.

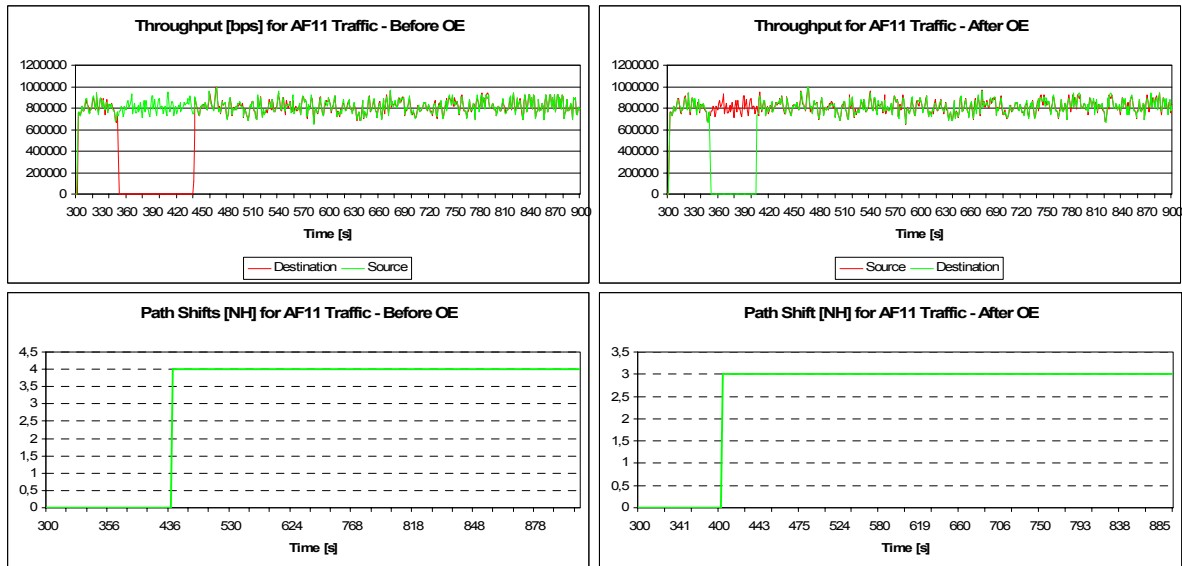


Fig.4. Link failure reaction with and without OEs

In the second experiment, under the SLSs constrains and equal network load conditions, we compare the behavior of both QBGP to OEs coupled with QBGP. As depicted in the Fig. 5, Fig. 6 and Fig. 7 the comparison of

both shows that while QBGP is unable to react to SLS violation events, the OEs coupled with QBGP are able to react to those events. QBGP is only able to react when the hold timer expires due to loss of consecutive keepalive messages during link congestion episodes. On the other hand, the OEs exploit multihoming due to their ability to react and choose the best path depending on the QoS dynamics. As expected some traffic reallocations are carried out, during which a transitory period is needed in order to accommodate the traffic aggregates into the best paths among the available ones. From the Figure 5 it is evident that even for less important traffic classes the delays needed to reach the egress domain are lower after a transitory period needed to accommodate these aggregates into the best paths among the available ones.

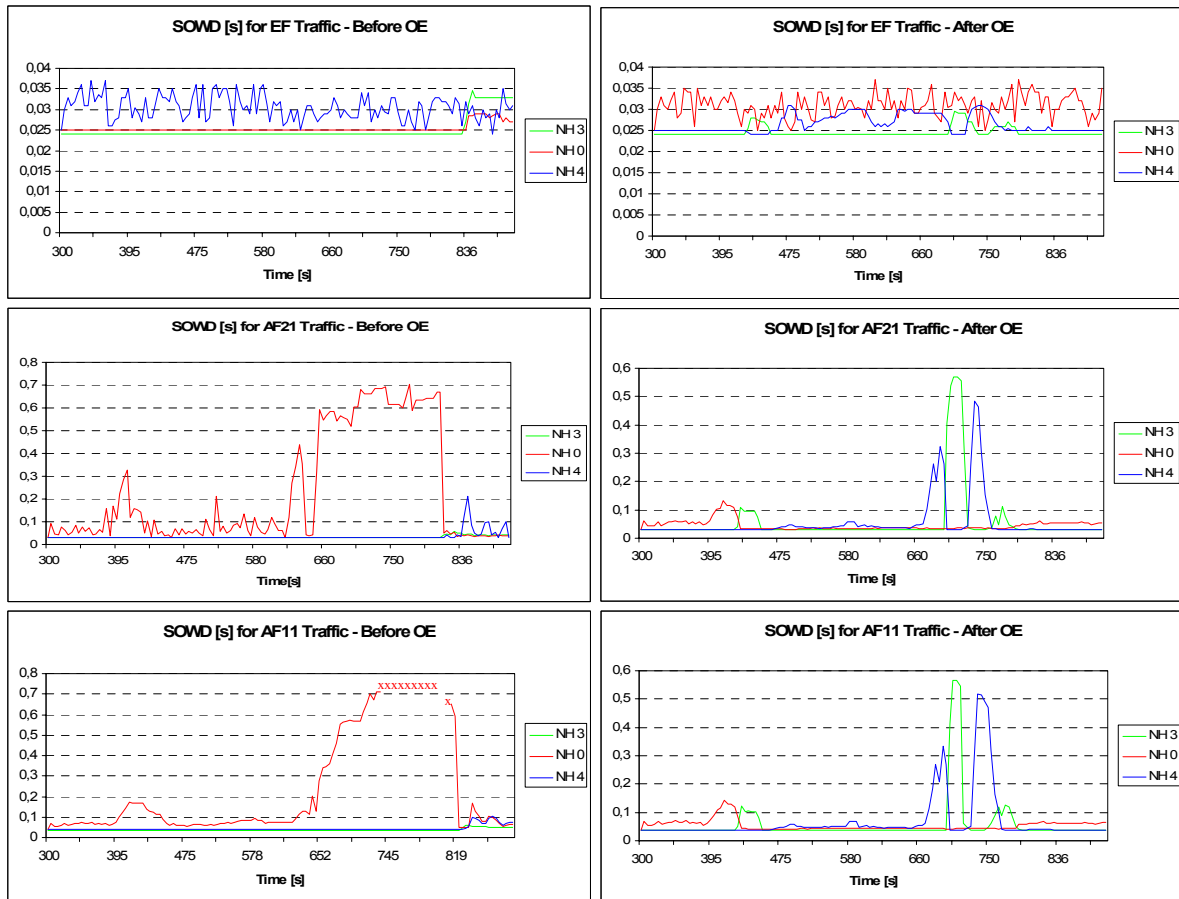


Fig.5. OWDs for EF, AF21, and AF11 over a 900-sec period, with and without OE.

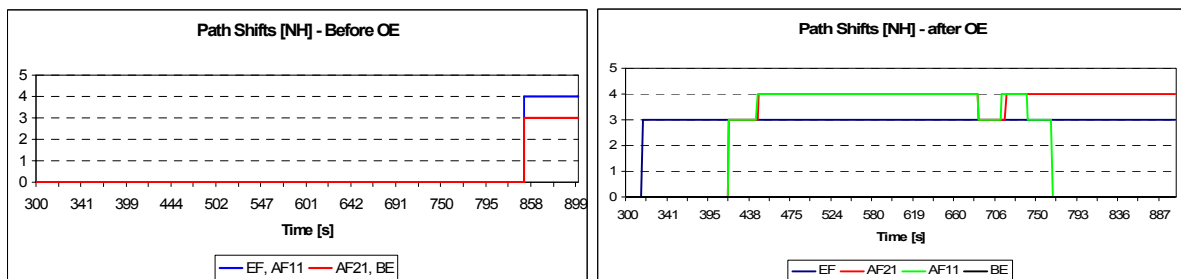


Fig.6. Path Shifts for EF, AF21, and AF11 over a 900-sec period, with and without OE

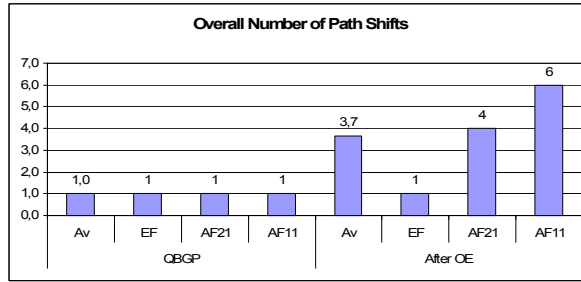


Fig.7. Path Shifts for EF, AF21, and AF11 over a 900-sec period, with and without OE

Furthermore, as it is presented in Fig. 8 and Fig. 9 the overall throughputs for all classes of traffic and the corresponding efficiency for background traffic is highly improved by our routing control layer.

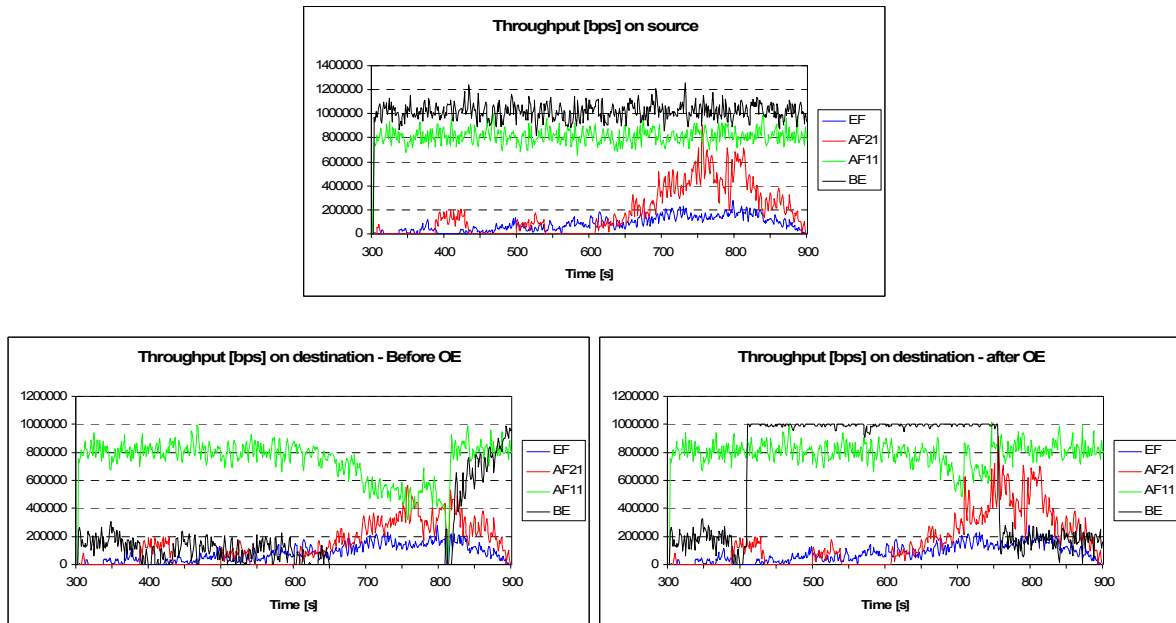


Fig.8. Throughput for EF, AF21, AF11 and BE over a 900-sec period, with and without OE

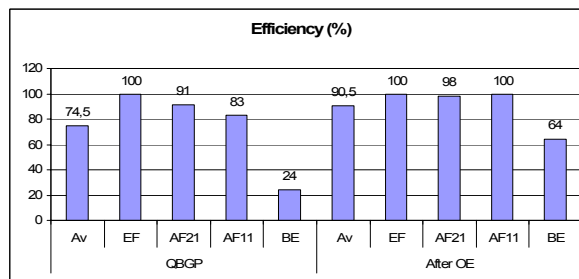


Fig.9. Efficiency computed over a 900-sec period. Data measured for all traffic classes and under default SLS configuration (Av = Average)

6. Conclusions

This paper depicts the framework for a combined interdomain QoS paradigm based on a completely distributed architecture supporting an edge routing control layer coupled with a QBGP or TE-BGP routing layer. As a first step in our research, and in order to validate our approach we have focused on the coupling of the routing control layer with a DiffServ QBGP underlying layer. The results obtained show that our distributed control architecture substantially enhances end-to-end QoS when compared with a pure QBGP model. We believe that whereas significant extensions and enhancements to BGP are certainly going to be seen, decoupled but incremental routing control architectures like the one proposed in this paper arise as strong candidates to provide flexible and value-added out-of-band interdomain QoS. In particular, this becomes perfectly suitable when interdomain traffic patterns need to dynamically adapt and rapidly react to medium or high network changing conditions, where the former in-band solutions seem impracticable at the present time.

References

- [1] Y. Rekhter, T. Li, "A Border Gateway Protocol 4 (BGP-4)," Internet Engineering Task Force, Request for Comments 1771, March 1995.
- [2] M. R. Garey, and D. S. Johnson, "Computers and Intractability: a guide to the theory of NP-completeness," Freeman San Francisco, 1979.
- [3] S. Uhlig, V. Magnin, O. Bonaventure, C. Rapiere and L. Deri, "Implications of the Topological Properties of Internet Traffic on Traffic Engineering," Proceedings of the 19th ACM Symposium on Applied Computing, Special Track on Computer Networks, Nicosia, Cyprus, March 2004.
- [4] B. Quoitin, S. Uhlig, C. Pelsser, L. Swinnen, O. Bonaventure, "Interdomain Traffic Engineering with BGP," IEEE Communications Magazine, May 2003.
- [5] A. Akella, B. Maggs, S. Seshan, A. Shaikh, R. Sitaraman, "A Measurement-Based Analysis of Multihoming," in Proceedings of ACM SIGCOMM 2003, Karlsruhe, Germany.
- [6] C. de Launois, B. Quoitin, and O. Bonaventure, "Leveraging network performances with IPv6 multihoming and multiple provider-dependent aggregatable prefixes," 3rd International Workshop on QoS in Multiservice IP Networks (QoSIP 2005), Catania, Italy, February 2005.
- [7] A. Akella, J. Pang, B. Maggs, S. Seshan and A. Shaikh, "A Comparison of Overlay Routing and Multihoming Route Control," in Proceedings of ACM SIGCOMM04, Portland, USA, August 2004.
- [8] Cisco Optimized Edge Routing, <http://www.cisco.com/>
- [9] Cristallo, G., C. Jacquenet, "An Approach to Interdomain Traffic Engineering," Proceedings of XVIII World Telecommunications Congress (WTC2002), France, September 2002.
- [10] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet routing convergence," in Proc. ACM SIGCOMM, 2000.
- [11] C. Villamizar, R. Chandra, R. Govindan, "BGP Route Flap Damping," Internet Engineering Task Force, Request for Comments 2439, November 1998.

- [12] S. Agarwal, C. Chuah, R. Katz “OPCA: Robust interdomain policy routing and traffic control,” IEEE Openarch, April 2003.
- [13] S. Savage, A. Collins, A. Aggarwal, et al, “Detour: a Case for Informed Internet Routing and Transport”, IEEE Micro, January, 1999
- [14] A. Collins, “The Detour Framework for Packet Rerouting”, PhD Qualifying Examination, October 1998.
- [15] D. G. Andresen, H. Balakrishnan, M. F. Kaashoek, R. Morris, “Resilient Overlay Networks”, in Proceedings, 18th of ACM SOSP, 2001.
- [16] Zhi Li, Prasant Mohapatra, “QRON: QoS-aware Routing in Overlay Networks”, IEEE Journal on Selected Areas in Communications, June, 2003.
- [17] Z. Duan, Z. L. Zhang, Y. T. Hou, “Service Overlay Networks: SLAs, QoS, and Bandwidth Provisioning”, IEEE/ACM Transactions on Networking, Vol. 11, number 6, December 2003.
- [18] L. Subramanian, I. Stoica, H. Balakrishnan, R. H. Katz, “OverQoS: Offering Internet QoS using overlays,” ACM SIGCOMM Computer Communications Review, vol. 33-1, January, 2003.
- [19] A. Akella, S Seshan, and A. Shaikh, “Multihoming Performance Benefits: An Experimental Evaluation of Practical Enterprise Strategies,” USENIX Annual Technical Conference 2004, Boston, MA, USA
- [20] Internap Network Services Corp., Flow Control Platform, <http://www.internap.com/>
- [21] M. Yannuzzi, X. Masip-Bruin, E. Monteiro, " Towards Self-Adaptive Inter-Domain Edge Routing," accepted for publication in the IEEE Infocom Student Workshop, Miami, USA, March 2005.
- [22] G. Almes, S. Kalidindi, M. Zekauskas, “A One-way Delay Metric for IPPM”, Internet Engineering Task Force, Request for Comments 2679, September 1999.
- [23] G. Almes, S. Kalidindi, M. Zekauskas, “A One-way Packet Loss Metric for IPPM”, Internet Engineering Task Force, Request for Comments 2680, September 1999.
- [24] J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, “Assured Forwarding PHB Group”, Internet Engineering Task Force, Request for Comments 2597, June 1999.
- [25] L. Xiao, K.S. Lui, J. Wang, K. Nahrstedt, “QoS Extension to BGP,” IEEE ICNP, November 2002
- [26] J-Sim Homepage, <http://www.j-sim.org>.
- [27] Infonet Suite Homepage, <http://www.info.ucl.ac.be/~bqu/jsim/>
- [28] GÉANT Website, <http://www.dante.net/server/show/nav.007>
- [29] Hamada Alshaer and Eric Horlait, Expedited Forwarding Delay Budget Through a Novel Call Admission Control, 3rd European Conference on Universal Multiservice Network (ECUMN’2004), Porto, Portugal.