A Proposal for Inter-Domain QoS Routing based on Distributed Overlay Entities and QBGP[1]

Marcelo Yannuzzi[1], Alexandre Fonte[2,3], Xavier Masip-Bruin[1], Edmundo Monteiro[2], Sergi Sànchez-López[1], Marilia Curado[2], Jordi Domingo-Pascual[1]

[1] Departament d'Arquitectura de Computadors, Universitat Politècnica de Catalunya (UPC)
Avgda. Víctor Balaguer, s/n – 08800 Vilanova i la Geltrú, Barcelona, Catalunya, Spain
{yannuzzi, xmasip, sergio, jordid}@ac.upc.es
[2] Laboratory of Communications and Telematics, CISUC-DEI, University of Coimbra,
Pólo II, Pinhal de Marrocos, Postal Address 3030-290 Coimbra, Portugal
{afonte, edmundo, marilia}@dei.uc.pt
[3] Polytechnic Institute of Castelo Branco,
Av. Pedro Álvares Cabral, nº12, Postal Address 6000-084, Castelo Branco, Portugal

Abstract. This paper proposes a novel and incremental approach to Inter-Domain QoS Routing. Our approach is to provide a completely distributed Overlay Architecture and a routing layer for dynamic QoS provisioning, and to use QoS extensions and Traffic Engineering capabilities of the underlying BGP layer for static QoS provisioning. As far as we know, no one has tried to combine the best of both worlds into a complementary solution. Our focus is mainly on influencing how traffic is exchanged among non-directly connected multi-homed Autonomous Systems based on specific QoS parameters. We provide evidence supporting the feasibility of our approach by means of simulation.

Keywords**: Inter-Domain QoS Routing, Overlay, BGP**

1 Introduction

At present, nearly 80% of the more than 15000 Autonomous Systems (ASs) that compose the Internet are stub ASs [1], where the majority of this fraction is multi-homed. For these ASs the issue of Quality of Service Routing (QoSR) at the inter-domain level arises as a strong need [2]. Whereas some research groups rely on QoS and Traffic Engineering (TE) extensions to BGP [3-5], others tend to avoid new enhancements to the protocol and propose Overlay networks to address the subject [6-8]. While the former approach provides significant improvements for internets under low routing dynamics, the latter results more effective when routing changes occur more frequently. The main idea behind the overlay concept is to decouple part of the policy control portion of the routing process from BGP devices. In this sense, the two approaches differ in how policies are controlled and signaled. BGP enhancements tend to provide in-band signaling, while the overlay approach provides out-of-band signaling.

[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.

The Overlay Architecture is mostly appropriate when communicating domains are multi-homed, and thus may need some kind of mechanism to rapidly change their traffic behavior depending on network conditions. In fact, multi-homing is the trend that most stub ASs exhibit in nowadays Internet, which mainly try to achieve load balance and fault tolerance on the connection to the network [7]. In addition, present inter-domain traffic characteristics reveal that even though an AS will exchange traffic with most of the Internet, only a small number of ASs is responsible for a large fraction of the existing traffic. Moreover, this traffic is mainly exchanged among ASs that are not directly connected; instead they are generally 2, 3 and 4 hops away [5].

Therefore, the combination of all these features made us focus on QoSR among strategically selected non-peering multi-homed ASs. However, the proposal is certainly not limited to this case and could also be applied if the ASs share EBGP (External BGP) peering links.

The approach to inter-domain QoSR we propose in this paper is to supply a completely distributed Overlay Architecture and a routing layer for dynamic QoS provisioning, while we use QoS extensions and TE capabilities of the underlying BGP layer for static QoS provisioning. In terms of the underlying inter-domain routing structure, two types of BGP routers can operate, namely, non-QoS aware BGP routers and QoS aware BGP (QBGP) routers. Thus, in order to develop highly scalable and stable routing schemes, it is mandatory that QBGP routers only distribute non dynamic QoS information. This is mainly because frequent network changes will translate into frequent BGP updates, which may lead to routing instability. Within the overlay inter-domain routing structure reside special Overlay Entities (OEs), whose main functionalities are the exchange of Service Level Agreements (SLAs), end-to-end monitoring, and examination of compliance with SLAs. These functionalities allow OEs to influence the behavior of the underlying BGP routing layer, to take rapid and accurate decisions to bypass network problems such as link failures, or service degradation for a given Class of Service (CoS). The reactive nature of this overlay structure acts as a complementary layer conceived to enhance the performance of the underlying BGP layer containing both QoS 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 overlay layer, and its coupling with the underlying BGP layer.

The remaining of this paper is organized as follows. Section II presents an overview of our overlay approach. In Section III the main functionalities required from the underlying BGP and overlay layers are analyzed, while Section IV presents our simulation scenario and results. Finally, Section V concludes the paper.

2 Overview of the proposed Overlay approach

As stated in the last Section, we propose in this paper a combined QBGP and Overlay Architecture for inter-domain QoSR. The main ideas behind the Overlay Architecture are:

- The OEs should respond nearly two orders of magnitude faster than the BGP layer in the case of a network failure.
- The OEs should react and try 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 modifications, and remains unaware of the QoSR architecture running on top of it.

The next figure depicts a possible scenario were our proposal could be applied.

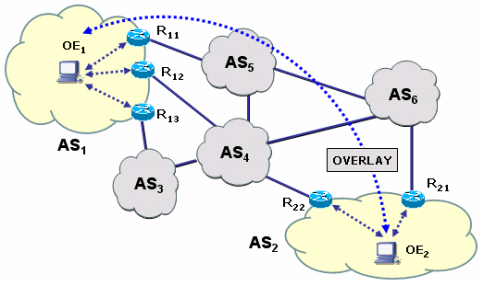

Fig. 1. Inter-Domain QoSR scenario where OEs are used for dynamic QoS provisioning among remote multi-homed ASs.

In this model, two peering OEs belonging to different ASs spanning across several AS hops are able to exchange a SLA and agree upon a set of QoS parameters concerning the traffic among them. The intermediate ASs do not need to participate in the Overlay Architecture, and therefore no OEs are needed within these transit ASs. From our perspective, the real challenge is to develop a completely distributed overlay system, where each OE behaves in a reflective manner. In this sense our overlay approach is like facing a mirror. Instead of proposing a complex scheme to dynamically and accurately manage how traffic enters a target AS, we focus on how traffic should exit from the source AS. Hence, what we seek is that the OE within the source AS behaves like the image in a mirror of the OE in the target AS. This mirroring scheme allows the OE in the source AS to dynamically manage its outgoing traffic to the target AS, depending on the compliance with the previously established SLA for a given set of CoSs. Then, under normal networking conditions, and based on the static QoS provisioning, each OE should measure end-to-end QoS parameters and check for violations to the SLA for every CoS. Within each AS, the OE provides tools for these measurements along every link connecting the multi-homed AS to the Internet. Henceforth, we assume that the topology has at least two different end-to-end paths between any pair of remote ASs participating in our QoSR model. When a violation is detected, the OE in the source AS is capable of reconfiguring 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 time scale [9].

The end-to-end measurements are based on active AS path probing among peering OEs. Hence, each OE within an AS spawns probes targeting the remote AS through every available link connecting the source AS to the Internet. We sustain, and we will show by simulation 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 overlay peering ASs and the number of CoSs remains limited. In fact, the traffic generated between two OEs is negligible, and an efficient provision of resources by means of QBGP implies that the OE will barely work. It is worth noting that a non-complying condition may only occur in a single direction of the traffic, which means that the bottleneck is merely on the upstream or the downstream path. For example, in Fig. 1 the OEs in

AS_1 and AS_2 measure the same parameters, such as One-Way Delay (OWD) [10] or One-Way Loss (OWL) [11], and react in the same manner due to their mirrored behavior. Therefore, either of them is able to independently decide if it should shift its outbound traffic or not.

As an example and to show the flexibility of our approach, let us assume that C_1 represents a CoS advertised from routers R_{11}, R_{12} and R_{13} in AS_1 and that for local policy reasons AS_1 would prefer to receive compliant traffic of C_1 from AS_2 through R_{11}. However, in case the selected path is not able to provide the SLA constraints previously negotiated between both ASs, AS_1 permits the reception of this traffic through R_{12} or R_{13} as far as the agreement is fulfilled. In this case, instead of using QBGP, the main objective could be satisfied if a static provisioning is done by means of TE-BGP, and the overlay layer provides dynamical QoSR. For instance, based on communities attribute of BGP [5], AS_1 could request AS_4 to prepend its own AS three times before announcing C_1 to AS_2, to prepend it two times before announcing C_1 to AS_6, and to perform no prepending operation at all when announcing this block to any other neighboring AS. Therefore, the advertisements that AS_2 receives under this scenario are: {AS_4, AS_4, AS_4, AS_1}; {AS_6, AS_5, AS_1}. Then AS_2 chooses to forward C_1 through AS_6. Nevertheless, once this is done, the best path chosen by BGP is completely unaware of any kind of QoS requirements or constraints between AS_1 and AS_2. Let us assume now that the link between AS_2 and AS_6 becomes loaded, while the path {AS_4, AS_1} through R_{22} does not. Despite these unequal network conditions, BGP will still prefer the path through R_{21}. Our approach allows the OE within AS_2 to become conscious of these conditions and dynamically reroute its outbound traffic of C_1 through R_{22}. 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.

An important point in our approach is that we try to reduce as much as possible the additional complexity introduced by the overlay layer. Agarwal *et al.* proposed an interesting overlay mechanism to reduce the fail-over time and to achieve load balancing of traffic entering an AS [7]. However, this proposal does not reuse any QoS or TE capabilities from the BGP layer. Moreover, it introduces a centralized and complex server which allows an AS to infer, by means of heuristics, the topology and customer/peer relationships among the multiple ASs that conform all tentative paths known to any given peering AS in the overlay structure. The complexity introduced is mainly due to the fact that accurately controlling how traffic enters an AS is a very intricate task, particularly when this must be done dynamically. As an alternative, our approach deals with the allocation of traffic from the source AS, since we strongly believe that simpler approaches such as this one would result more attractive to become deployed.

3 Main functionalities of the routing layers

In this Section we describe in detail de functionalities of the routing layers of our proposal, namely the top layer, concerning the overlay routing functionalities and the lower layer pertaining to the underlay BGP routing functionalities.

3.1 Top Layer: Overlay Routing Functionalities

This layer is composed by a set of OEs:

3.1.1 Basic set of components:

- At least one OE exists per QoS domain.
- An OE has full access to the border BGP routers within an AS.
- In the simplest scenario, within each QoS domain the OE could operate independently. In this case, the OE checks for local SLA compliance and in case of local violations it tunes BGP on-the-fly to achieve a better traffic distribution.
- An OE has algorithms for both detecting non-conforming conditions for a given CoS, and deciding when and how to reallocate its traffic.

3.1.2 Main components:

An Overlay Protocol: A protocol between remote OEs is needed. This protocol allows OEs to exchange SLAs with each other, and to exchange substantial information for the Overlay Architecture. An OE within a target AS could reallocate some part of its incoming traffic by remotely asking for changes in the outbound traffic to a peering OE within the source AS. This proactive behavior becomes necessary, for example, when the OE within the target AS receives information that some sort of maintenance is going to be performed on a certain link, and that will affect a fraction of the traffic exchanged between both ASs. In this case, instead of waiting for the source OE reaction, a mechanism is provided within the protocol such that the failure is completely bypassed in advance.

Metric Selection: In order to validate our approach, we choose a simple QoS parameter for the dynamical portion of our QoSR model. The parameter we have selected is a smoothed OWD (SOWD), which defines the following metric:

$$\overline{OWD}(m,n) = \frac{1}{N} \sum_{k=n}^{k=n+N-1} OWD(m,k) \qquad (1)$$

This SOWD essentially corresponds to the average OWD through a sliding window of size N. Instead of using instantaneous values of the OWD, we propose to use this low-pass filter, which smoothes the OWD avoiding rapid changes in our metric. Of course a trade-off exists in terms of the size of the window N. 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. In this scenario the SLA exchanged by the OEs is simply the maximum SOWD D_j tolerated for each different CoS C_j.

We assume that each OE uses one logical address for each different CoS, and also that specific local policies are applied to IBGP (Internal BGP). With this approach, an OE is able to probe a remote OE for any given CoS, in a round-robin fashion, through all available external links in the local AS. Hence, m and n correspond to the n_{th} probe generated by a source OE and sent towards the m_{th} external link of the AS. Then, the OEs are computing a per-class of service cost to reach the remote AS over every external link m based on the previous metric. Furthermore, packets probing a specific CoS belong to that CoS. For instance, in a QBGP framework based on Differentiated Services (DiffServ), when probing a particular CoS which is locally mapped to an As-

sured Forwarding (AF) class in each intermediate AS, the probes are tagged under the same AF class [12].

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.

Piggy-Backing mechanism: An important issue is that an active probing technique developed to measure the OWD requires feedback from the remote OE. However, the mirroring scheme implies that the remote OE is already probing the local OE and expects feedback from this latter as well. Thus, the easiest way to avoid unnecessary messages traversing the network is to endow the protocol between the OEs with a piggy-backing technique. Then, feedback for the OWD is carried on the probes itself.

Stability: Another central issue is that the traffic reallocation process should never generate network instability. For instance, it is likely to occur that an OE operates completely unaware that it may be sharing a group of paths with other unknown OEs belonging to non-peering ASs. Furthermore, these OEs could force conflicting constraints over those shared paths, and no centralized OE exists to break the tie in this case. These unrelated conflicting situations could trigger several uncoordinated fast reroute actions from different OEs which may lead to oscillating network conditions. 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 their selves 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 $[T_h\text{-}t,t]$ 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 [13], 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:

$$R(t) = R(T)e^{-\left(\frac{t-T}{\tau}\right)} \quad \textbf{(2)}$$

Where T_h, 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 SLAs. In order to prevent simultaneous reallocations for all the affected CoSs, we endow the OE with a contention mechanism which prioritizes the relevance of the different CoS. Then, more relevant CoSs are reallocated faster than less priority classes. The contention algorithm operates as follows:

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

$\Rightarrow$ *Then, i.e. 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 link 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 link. Instead of tuning the contention algorithm to efficiently cope with both problems at the same time, we rely on the probing technique since a link failure will cause the complete loss of probes for all CoS within the link. Our proposal is based on incrementing the frequency of the probes per-CoS as soon as losses are detected. We maintain that this rising in the frequency does not exacerbate the load on the network, firstly because the fraction of traffic generated by the OE that detects the problem is negligible in terms of the overall traffic exchanged between both ASs. Secondly, this is 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 decreases back to its normal value.

3.2 Bottom Layer: Underlay BGP Routing Functionalities

The set of routes to be tested by the OE using the probing techniques described in the previous sub-section, are predetermined by the underlying BGP-based layer. In this layer two types of devices can operate; legacy BGP routers and QBGP routers. On the one hand, a legacy BGP router distributes Network Layer Reachability Information (NLRI) to its peers according to local routing policies. On the other hand, a QBGP router is also able to distribute QoS information and take QoS routing decisions.

Scalability is one of most important requirements in inter-domain QoSR. Therefore, providing scalable internets and keeping overhead in acceptable levels, requires a trade-off between the frequency of advertisements and the inaccuracy of routing information. Consequently, in our approach QBGP routers should only handle non-dynamic QoS information [14], and take routing decisions per-CoS constrained to the previously established SLA between different peering domains. In our model, QBGP routers can be seen as the practical tool to establish the overall inter-domain QoSR infrastructure composed by several sub-routing layers, one for each CoS, which in addition could be dynamically influenced by the overlay layer. Interesting approaches and further information on the subject of QBGP could be found in [3, 4, 14].

3.3 Combined QoSR Algorithm

The next scheme (Fig.2) depicts our combined QoSR algorithm. Let *m* be the external link currently allocating traffic of class C_j. It is important to remark that the approach we follow is that a traffic reallocation process could only occur when a SLA is violated. In this sense, 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 SLA is detected. Then, two distinct threads of events occur upon the reception of a probe for class C_j. Initially, the probe *(k,l)* is separated from the piggy-backed feedback *OWD(m,n)*. In order to accurately reply back to the sender, the first to be

processed is the *OWD(k,l)* which is shown as (I) in Fig. 2. On the other hand, the piggy-backed *OWD(m,n)* is processed, which is depicted as (II) in the figure.

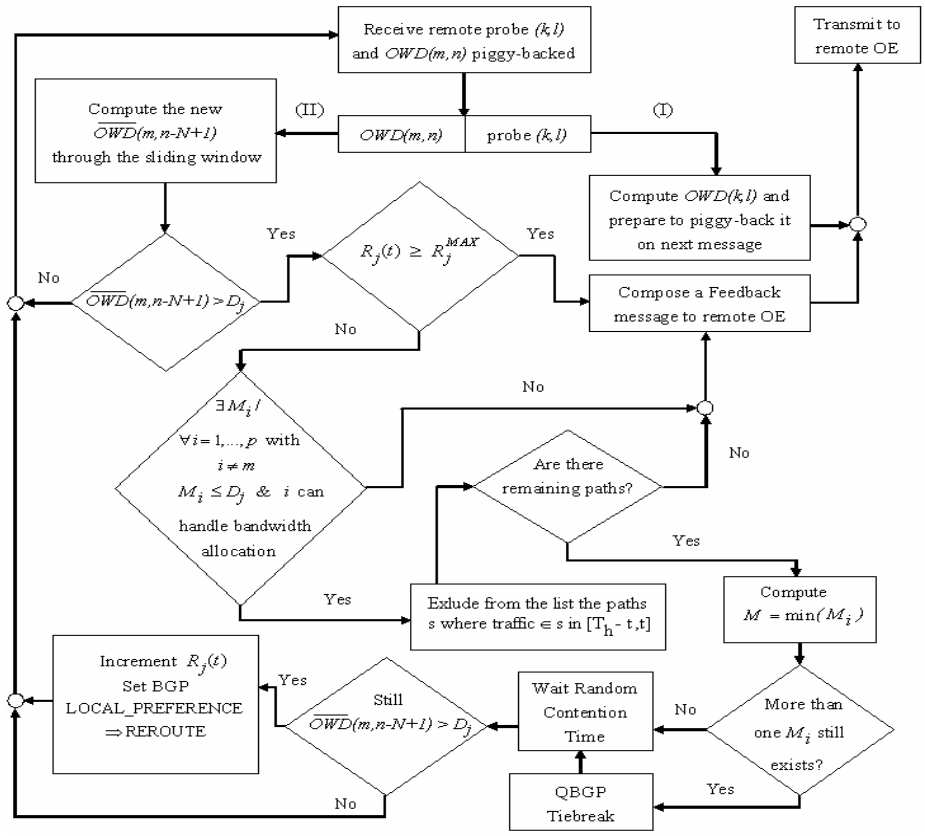

Fig. 2. Combined QoSR Algorithm

Once the SOWD is computed, the algorithm checks for violations to the maximum SOWD tolerable, that is D_j. If no violations have occurred the algorithm simply waits for the next incoming probe. 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. 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 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 $[T_h\text{-}t,t]$. 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 exe-

cuted 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.

4 Simulation Results

The Overlay Architecture proposed in this paper is being evaluated and validated by simulation. In this section some preliminary results are presented to allow a first evaluation of the overall architecture and its capability to support QoS traffic classes in a dynamic way. We are using the J-Sim simulator [15] with the BGP Infonet suite [16] which is compliant with BGP specification RFC 1771 [17]. A set of Java components with the functionalities of the overlay layer was developed. In order to allow the Overlay Entities to have full access to the Adj-RIBs-In and the Loc-RIB of a BGP speaker, and to have control over the BGP decision process, it was necessary to add some extensions to the Infonet suite. A precise definition of the terms Adj-RIBs-In and Loc-RIB can be found in [17]. Furthermore, other extensions were needed to allow the Loc-RIB to be updated with the changes in the Adj-RIB-In. Finally, and because no standardized QBGP protocol exists, we have also included the following QoS BGP extensions to the Infonet suite:

- An optional transitive attribute to distribute the identifications (IDs) of the different CoSs, and a set of modifications to BGP tables to allow the storage of this additional information, following a similar approach to the one described in [3].
- A set of mechanisms to: i) allow BGP speakers to load local supported CoS; ii) allow each local IP prefix to be announced within a given CoS; iii) allow BGP speakers to set the permissibility of routes by using filters to deny/allow a given route based on local QoS supported capabilities and policies.

For our simulations, we used the topology presented in Fig. 7. The topology is based in the GÈANT European Academic Backbone with some simplifications to reduce the complexity of the simulation model. In this topology we considered as remote multi-homed AS domains, AS_1 and AS_2. All links were assumed to be bidirectional with the same capacity C (C=2Mbps) and delay propagation P_d (P_d=10ms), with the exception of AS_2 links where, in order to have some bottleneck, the capacity chosen was C/2. According to this and just for complexity concerns, we modeled each AS as a single QBGP router with core DiffServ capabilities configured to support four different IP packet treatments (EF, AF11, AF21 and Best-effort) allowing four different classes of traffic, namely CoS1, CoS2, CoS3 and CoS4. To complete the scenario, on the domain where traffic was injected we used edge DiffServ capabilities to mark packets with a specific DSCP (DiffServ Code Point) depending on its corresponding CoS. These marks were applied to IP packets from the traffic sources and to the probes generated by the OE. The test conditions are summarized in Table 1. The results obtained are presented in Fig. 3 to Fig. 6. The maximum SOWD tolerated per-CoS (D_j) was heuristically chosen to allow the OEs to take advantage of alternative paths. The SWOD computed when probes were lost was also heuristically chosen. The criteria selected was that 3 consecutive losses imply nearly a rise of 25% in the SWOD. For the tests presented we set $R_j^{MAX} = \infty \ \forall j$. Moreover, no probes were generated for Best-effort traffic (NA=Not Available), and a sliding window of 3 seconds was used in all tests, which is shown as Mov.Average in Table 1.

Table 1. Test conditions

CoS	CBR (Mbps)	Pkt. Size (KB)	PHB	Max. SOWD (ms)	Probing Freq.	Hold (Contention) &T_h (s)	Mov. Average
CoS1	0,4	1	EF	85	1 s, 1KB	3 & 8	3 s
CoS2	0,8	1	AF11	100	1 s, 1KB	6 & 12	3 s
CoS3	1,0	1	AF21	120	1 s, 1KB	9 & 20	3 s
CoS4	1,6	N.A	BE	N.A	N.A	N.A	N.A

The first objective of the simulation was the validation of the initial assumption that our approach, based on a complementary routing layer, enhances the reaction of the overall routing infrastructure. Then, as a performance indicator, we chose to compare the response time to a link failure. Fig. 3 depicts a set of plots for traffic of CoS1 showing the throughput measured at the destination, the SOWD experienced by probes for all available paths, and the path shifts determined by changes in the next-hop for the source AS, namely AS_1. From these plots, we can observe that a pure QBGP framework (without OEs running on AS_1 and AS_2) needs about 80 seconds to overcome a link failure, but only 5 seconds are needed when 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 3 seconds before re-routing.

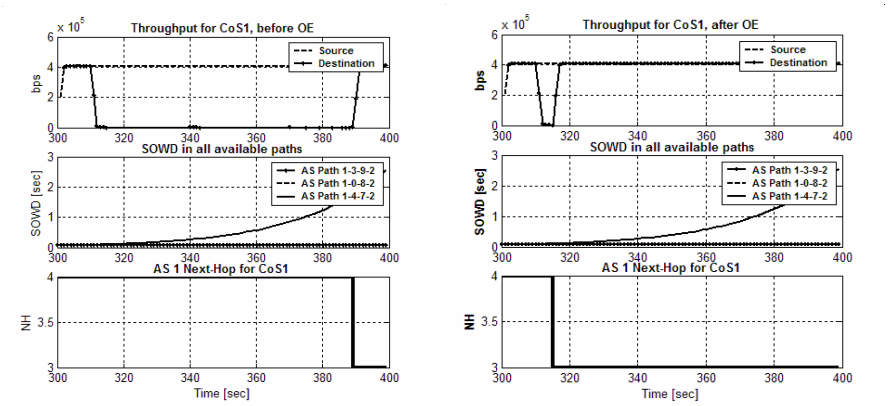

Fig. 3. Link failure reaction with and without OE.

Secondly, we examined the behavior of our overall inter-domain QoS routing infrastructure. For each traffic class, we used as performance indicators the throughput at the destination and the SOWD experienced by the probes through all available paths. From figures 4 and 5, we can observe that without OEs there are clear violations to the SLAs established between the end-to-end domains. However with OEs, it becomes clear that the network is able to react to SLA violations, and find the best paths to reallocate traffic for the affected traffic classes. Consequently, after a transitory interval of approximately 13 seconds, needed to accommodate the traffic for each CoS, it is visible that a steady state is reached and the SLAs are satisfied for all affected classes.

Furthermore, and in order to evaluate overall link utilization, we measured the throughput over all available links at the destination AS (AS_2). Fig. 6 shows that with

OEs, in addition to the compliance with the established SLAs a better distribution of inter-domain traffic is obtained, and thus, resources are more efficiently used. In contrast, without OEs, traffic distribution depends on the accuracy of the static provisioning, which does not take into account the real dynamics on the network. The extra cost in these cases was merely an increment of 8 Kbps, per-CoS, on each link in the remote AS-AS traffic, when oversized probes of 1 KB are spawned every one second.

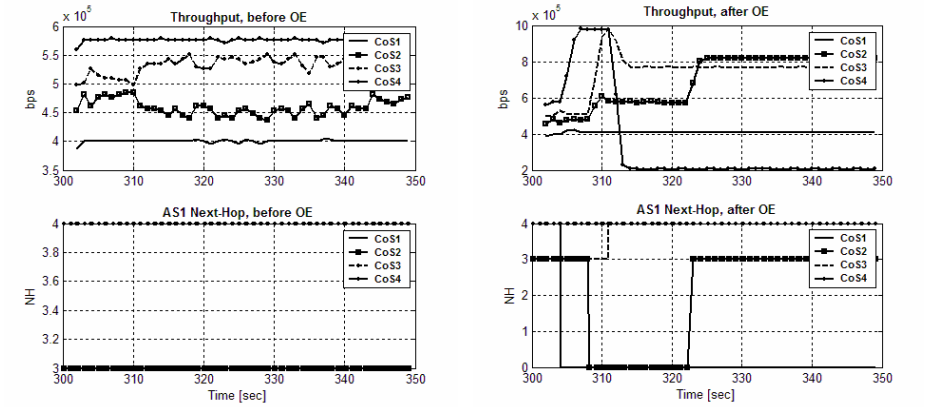

Fig. 4. Throughput for traffic of CoS1-CoS4, with and without OE

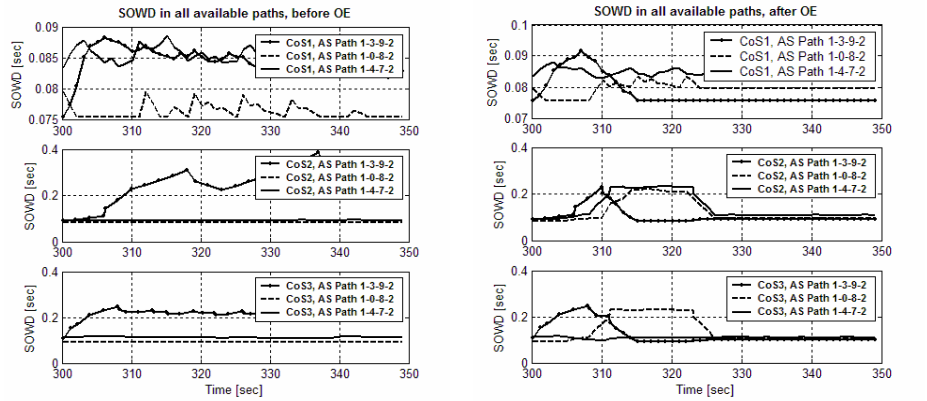

Fig. 5. OWD in all available paths for CoS1-CoS4, with and without OE

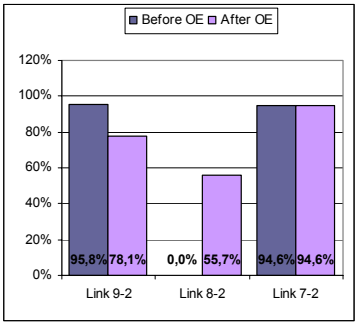

Fig. 6. Remote AS link utilization

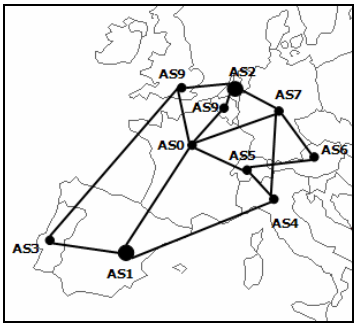

Fig. 7. Topology based on the GÉANT Network [18]

5 Conclusions

This paper depicts the framework for a combined inter-domain QoSR paradigm based on a completely distributed Overlay Architecture 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 overlay with a DiffServ QBGP underlying layer. The results obtained show that our distributed Overlay 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, the overlay structure arises as a strong candidate to provide flexible and value-added out-of-band inter-domain QoSR. In particular, this becomes perfectly suitable when inter-domain traffic patterns need to dynamically adapt and rapidly react to medium or high network changing conditions, where the former solutions seem impracticable at the present time.

References

1. Olivier Bonaventure, Bruno Quotin, Steve Uhlig, “Beyond Interdomain Reachability”, Workshop on Internet Routing Evolution and Design (WIRED), October 2003.
2. E. Crawley, R. Nair, B. Rajagopalan, H. Sandick, “A Framework for QoS-based Routing in the Internet”, Internet Engineering Task Force, Request for Comments 2386, August 1998.
3. Cristallo, G., C. Jacquenet, “An Approach to Inter-domain Traffic Engineering”, Proceedings of XVIII World Telecommunications Congress (WTC2002), France, September 2002.
4. Li Xiao, King-Shan Lui, Jun Wang, Klara Nahrstedt, “QoS Extension to BGP”, IEEE ICNP, November 2002.
5. Olivier Bonaventure, Steve Uhlig, Bruno Quotin, et al, “Interdomain Traffic Engineering with BGP”, IEEE Communications Magazine, May 2003.
6. L. Subramanian, Ion Stoica, Hari Balakrishnan, R. Katz, “OverQoS: Offering Internet QoS Using Overlays”, ACM SIGCOMM, Computer Communications Review, January 2003
7. S. Agarwal, C. Chuah, R. Katz “OPCA: Robust Interdomain Policy Routing and Traffic Control”, IEEE Openarch, April 2003.
8. Zhi Li, Prasant Mohapatra, “QRON: QoS-aware Routing in Overlay Networks”, IEEE Journal on Selected Areas in Communications, June, 2003
9. C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet routing convergence," in Proc. ACM SIGCOMM, 2000.
10. G. Almes, S. Kalidindi, M. Zekauskas, “A One-way Delay Metric for IPPM”, Internet Engineering Task Force, Request for Comments 2679, September 1999.
11. G. Almes, S. Kalidindi, M. Zekauskas, “A One-way Packet Loss Metric for IPPM”, Internet Engineering Task Force, Request for Comments 2680, September 1999.
12. J. Heinanen, F. Baker, W. Weiss, J. Wroclawski, “Assured Forwarding PHB Group”, Internet Engineering Task Force, Request for Comments 2597, June 1999.
13. C. Villamizar, R. Chandra, R. Govindan, “BGP Route Flap Damping”, Internet Engineering Task Force, Request for Comments 2439, November1998.
14. IST MESCAL project, “Specification of Business Models and a Functional Architecture for Inter-domain QoS Delivery”, Deliverable D1.1, June 2003.
15. J-Sim Homepage, http://www.j-sim.org.
16. Infonet Suite Homepage, http://www.info.ucl.ac.be/~bqu/jsim/
17. Y. Rekhter, T. Li , “A Border Gateway Protocol 4 (BGP-4)”, Internet Engineering Task Force, Request for Comments 1771, March 1995.
18. GÉANT Website, http://www.dante.net/server/show/nav.007