Ann. Telecommun.
DOI 10.1007/s12243-008-0065-y

Interdomain quality of service routing: setting the grounds for the way ahead

Alexandre Fonte · Marilia Curado · Edmundo Monteiro

Received: 17 February 2008 / Accepted: 1 July 2008

Abstract A common criticism of the current Internet is the fact that it does not offer quality of service (QoS) guarantees across autonomous system boundaries. The Border Gateway Protocol (BGP) is central to solve this problem, since it enables AS to distribute reachability information. However, BGP is agnostic of any performance or QoS metrics. For this reason, the debate about the requirements for the future interdomain routing architecture and about whether these requirements are best met by an approach of introducing changes into BGP or by replacing BGP is still open. This article provides an insight into the interdomain QoS routing problem. First, the main drawbacks of current interdomain routing with regard to the provision of QoS are identified. Second, a survey of the most relevant interdomain QoS routing approaches are described and discussed. We also give a broad perspective on challenges surrounding the issue of whether to extend or replace BGP to support QoS, with particular emphasis on the technical challenges. However, we also point out some nontechnical unsolved challenges that, in our perspective, are still almost certainly the biggest barrier to the development of interdomain QoS routing.

Keywords QoS · Routing · Interdomain QoS routing · BGP

A. Fonte · M. Curado · E. Monteiro
Departament of Informatics/CISUC, University of Coimbra, Coimbra, Portugal

M. Curado
e-mail: marilia@dei.uc.pt

E. Monteiro
e-mail: edmundo@dei.uc.pt

A. Fonte (✉)
School of Technology, Polytechnic Institute of Castelo Branco, Castelo Branco, Portugal
e-mail: afonte@dei.uc.pt

1 Introduction

The Internet of today has become the global communication system. More and more telecommunication networks and applications are migrating to Internet Protocol (IP). Cost reduction; easier network maintenance; and, above all, productivity improvements resulting from the convergence of applications are the main reasons for employing IP. This increasing interest in IP-based applications, such as IP virtual private networks or voice over IP, also brought an increasing demand for IP services with tighter service-level specifications in terms of end-to-end quality of service (QoS) guarantees. These needs were behind the fact that, in the beginning of the 1990s, a couple of solutions for reservation-based services (i.e., IntServ) and reservation-less services (i.e., DiffServ) started to appear within the Internet Engineering Task Force (IETF) [1].

When paths and links are stable and link congestion is not excessive, IntServ and DiffServ models have been demonstrated to be effective for stringent QoS [2]. However, path and link failures and excessive link congestion are common events in the Internet [3, 4]. This makes it clear that there is a definitive need for extending the QoS concept also to standard Internet routing protocols, such as the Open Shortest Path First

Published online: 25 November 2008

(OSPF) protocol. The main advantage of QoS routing is that it can optimize traffic performance and alleviate reservation-based or reservation-less service outages caused by failures. This has resulted in standards such as QoS routing mechanisms and OSPF extensions proposed for intradomain QoS routing in IP networks [5].

While there has been significant progress to enable the Internet to support QoS, the biggest open problem remains the expansion of QoS across multiple autonomous system (AS) boundaries. In particular, when traffic leaves the end-user network, there is no guarantee that the service quality on the domains crossed meets its QoS requirements. The Internet routing scenario described poses economic and technical challenges for which several approaches have been proposed. However, existing solutions have considerable drawbacks, as described next. First, despite the potential benefits offered by these frameworks, they have not turned out to be sufficiently appealing to be deployed in large scale. In other words, they are still missing the definition of a truly economic framework that might motivate ISPs to deploy QoS within their networks and to offer QoS to end-users that are not their customers. The reality is that, currently, many ISPs still prefer to over-provision their networks to solve QoS problems rather than deliver QoS based on the above mentioned frameworks [6]. Second, there is no effective interdomain routing mechanism available today that supports QoS or enables end-users to have full control over how their traffic is routed throughout multiple AS toward the target destination. For instance, it is not possible to explicitly select paths that circumvent congested or unwanted AS.

In the late 1990s, the IETF recognized that interdomain QoS routing is a critical missing piece for the distribution of information about QoS capabilities supported by each AS (e.g., a class or metaclass of service) [7]. In particular, the present interdomain routing system based on the Border Gateway Protocol (BGP) provides Internet end-users with suboptimal routing, which compromises QoS support. BGP inefficiency stems from the fact that it is agnostic of any performance or QoS metrics, such as end-to-end latency or loss. In effect, BGP was designed only for providing interdomain reachability information. To illustrate this statement, Fig. 1 presents the scenario where an AS, AS1, wants to send traffic to the subnet 12.0.0.0/8 in a remote AS, AS5. In this case, according to the least AS hop count criterion of the BGP decision process, AS1 would select the path AS4–AS5 (learnt from AS4), the default BGP active path, rather than the alternate path AS2–AS3–AS5 (learnt from AS2). However, the problem of this choice is that the packets would experience

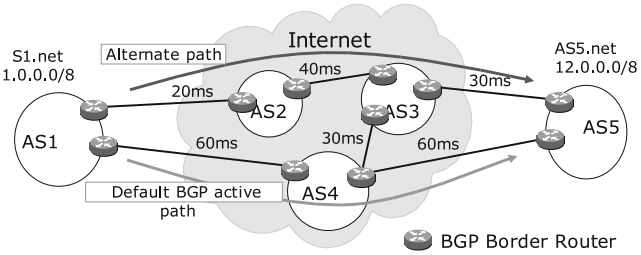


Fig. 1 Illustration of BGP suboptimality

more delay in the path through AS4, i.e., 120 ms, than if they were sent through AS2–AS3–AS5.

This article provides an insight into the interdomain QoS routing problem. First, Section 2 briefly presents the BGP protocol. The main drawbacks of current interdomain routing, regarding the provisions of QoS are then identified in Section 3. Section 4 presents and discusses the most relevant interdomain QoS routing approaches. Then, Section 5 gives a broad perspective on challenges surrounding the issue of whether to extend or replace BGP to support QoS. Finally, Section 6 concludes this article and indicates some research directions for future work in interdomain QoS routing.

2 Interdomain routing: a brief overview

This section presents and introduction to BGP, followed by the description of the main traffic control approaches based on BGP.

2.1 What is BGP?

BGP, at version 4, is the current de-facto interdomain routing protocol [8]. Two key functions of BGP are the distribution of reachability information and the control of traffic exchanges among AS. However, additional capabilities can be easily introduced in BGP due to the flexibility of its architecture. The support of these additional features is captured by the capabilities parameter carried within the initial OPEN messages used to open sessions between BGP speakers [9]. For instance, BGP speakers supporting new extensions to BGP, such as multiprotocol extensions to BGP, should negotiate these capabilities with their peers at the start-up of BGP sessions [10, 11].

BGP uses a fairly simple path-vector algorithm. The advertisement of reachable destinations includes the IP prefixes and information that describes the properties of the paths to these destinations. This specific information is expressed in terms of path attributes,

such as the complete AS PATH sequence (e.g., AS20: AS21:AS22:AS23). By default, a BGP speaker selects the route with the shortest AS PATH sequence. However, other path attributes, such as LOCAL-PREF and multiexit discriminator (MED) attributes, can be used to influence the decision process of BGP routers. UPDATE messages are used to announce (or withdraw) lists of reachable destinations that share common path attributes. Additionally, in order to control the size of routing tables maintained in downstream AS, BGP also supports route aggregation based on classless interdomain routing, where blocks of IP prefixes can be combined into a single IP prefix.

2.2 Path attributes

BGP's standard specifies three well-known mandatory attributes, that is, attributes that all BGP implementations must be able to recognize and process, and that should appear in UPDATE messages: ORIGIN, AS-PATH, and NEXT-HOP. Two additional attributes that are part of the BGP decision process are the MED and the LOCAL-PREF attributes (see Subsection 2.3). These are optional attributes, but often used in traffic control with BGP (see Subsection 2.4). Figure 2 illustrates the use of BGP path attributes. Next, we briefly describe the BGP path attributes in the same order as they are used in the BGP decision process (see Subsection 2.3).

LOCAL-PREF — This attribute expresses the degree of preference for each route toward a given IP prefix. Larger LOCAL-PREF values should be attached to preferred one routes. LOCAL-PREF must not be redistributed between external BGP speakers.

AS-PATH — This attribute records the sequence of AS numbers composing the route so far. It allows BGP speakers to detect loops in the routing.

ORIGIN — This attribute identifies the mechanism that originated the reachability information (0-IGP, 1-EGP, 2-INCOMPLETE).

MED — This is an optional attribute that might be used when an AS has multiple peering links to the same neighboring AS. The peering link (i.e., the exit point) with lowest MED is the preferred by neighboring AS.

NEXT-HOP — This attribute contains the IP address of the next-hop router to the IP prefix announced in the update message.

2.3 Path selection process

This section describes the route selection to the various IP prefixes, referred as BGP path selection process. Figure 3 presents a conceptual model of a BGP speaker. Each BGP speaker maintains three routing information bases (RIBs): the adjacent RIBs-In (Adj-RIBs-In) that store routes learned from peers, the local RIB that stores the best routes selected by the BGP path selection process and that are used to feed the IP forwarding table [i.e., the forwarding information base (FIB)], and Adj-RIBs-Out that store the routes to be advertised to peers, according to the configured policy.

The BGP path selection operates on the acceptable routes, which are the routes stored in the Adj-RIBs-In, after applying incoming filtering. To select the best route, among the existing acceptable routes to the same IP prefix, the BGP speaker identifies the one that has the highest LOCAL-PREF, or in case more than one equally good route exists, it invokes a tie-breaking function. The algorithm follows the criteria present in Fig. 4, which are applied in the order specified. It stops as soon as only one route can be considered.

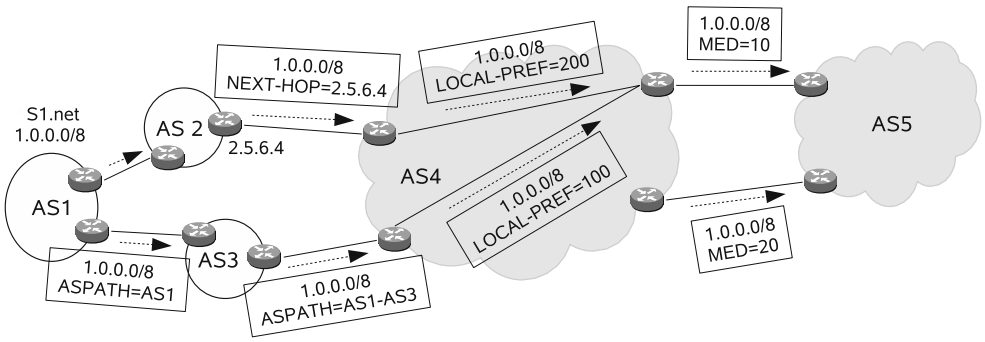


Fig. 2 BGP path attributes examples

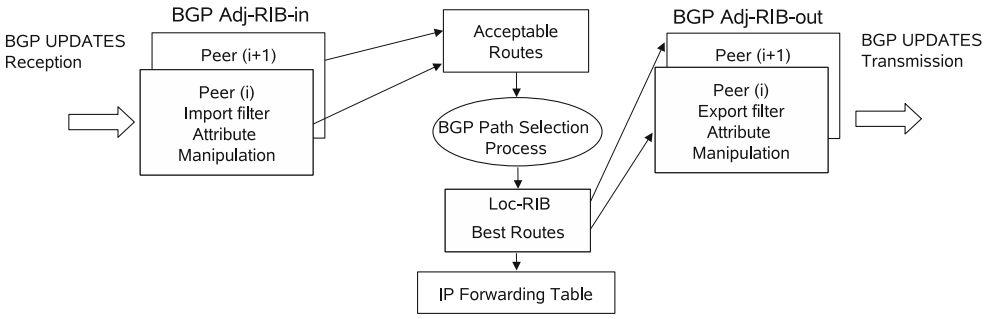


Fig. 3 Conceptual model of a BGP speaker

2.4 Traffic control with BGP

BGP has evolved, and nowadays the routing decision is based on a distributed policy scheme, which has contributed to the increased BGP complexity. However, this routing scheme has become popular, especially in AS of commercial organizations. Much of this popularity can be explained by the fact that BGP enables AS to control the traffic exchanges in the way they want, i.e., reflecting business relationships and traffic engineering (TE) objectives [12].

2.4.1 Business relationships

BGP enables an AS administrator to configure the router filters to import and export routes according to the customer-provider or peer-to-peer relationships. For instance, usually ISPs prefer routes learned from their customers over the routes learned from their providers, to avoid being unnecessarily charged by their providers.

2.4.2 Interdomain TE

TE tools are indispensable to engineer the traffic entering and/or leaving AS, so that a given set of traffic goals (e.g., performance or/and transit costs) are fulfilled. These goals could be achieved by modifying BGP policy filters. The control of outgoing traffic is often done by ranking the equal-good routes using one of the following three techniques to influence the BGP decision process. The first technique consists on modifying the LOCAL-PREF of routes using as criterion, for instance, the transit cost of traffic or information obtained by active and/or passive measurements. The second technique resorts to modifying the cost to attain the next-hops (the egress points of outgoing traffic) provided by the Interior Gateway Protocol (IGP) (e.g, OSPF or IS–IS link weights). One common example is to rely on hot-potato routing, in which the selection of the egress points is decided by comparing IGP costs of intradomain paths. The third, but infrequent, technique, since it is only supported by a few BGP routers, is to insert a MED in the route received.

Regarding the control of incoming traffic, the approach followed aims at influencing BGP peers to prefer certain routes over others. There is a plethora of procedures to perform such tuning of BGP, such as MED assignments, COMMUNITIES distribution, AS prepending/padding, and selective announcements [13, 14]. One common aspect of these procedures is that they require external BGP updates and support from downstream AS.

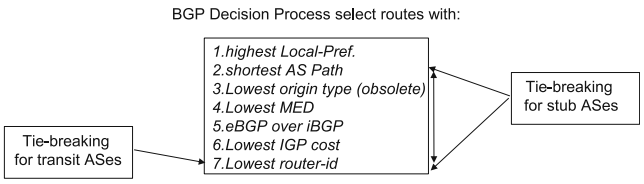


Fig. 4 BGP decision process criteria

3 Open issues in interdomain routing

Interdomain routing is based on BGP. Unfortunately, BGP provides end-users with suboptimal routing in terms of performance and reliability. At the root of interdomain routing inefficiency, there are structural short-comings that affect BGP, such as the coupling between policy filtering and the route discovery and selection mechanisms. The goal of this section is thus to diagnose the actual role of BGP, identifying its main problems and limitations which potentially create or exacerbate the interdomain routing inefficiency.

3.1 QoS support

The BGP standard only specifies the means that enable BGP speakers to exchange reachability information. At present, the proposals of BGP route attributes to carry QoS or congestion information within BGP

advertisements and the modifications to the BGP decision process to handle this type of data have not yet been standardized. Therefore, unless attribute manipulation is exercised, BGP speakers are currently constrained to adapt route selection to the least sequence of AS the routes transverse, as announced within the BGP AS PATH attribute [17]. The problem with this approach is that the AS PATH length is a metric which does not reflect the real end-to-end packet latency. The correlation between path length and round-trip time (RTT) has been shown to be rather poor, suggesting that BGP path selection might be similar to a purely random choice. In effect, experimental results, obtained using a real Internet topology and RTT data, showed that the AS PATH length metric achieved only a 50% success rate of the trials performed to identify the destinations with smaller RTT [18].

3.2 Route convergence

Route convergence time is a metric commonly used to measure the "speed" of a routing protocol (i.e., the time it requires) to adapt routing to topology changes (e.g., a fail-down or a new route). In case of BGP, this time is dependent on Internet topological aspects and routing policies. In particular, these aspects affect the length of available backup paths for a certain prefix and, thus, the convergence time. Therefore, at Internet scale, the BGP failover process may take several minutes [19, 20]. More specifically, the BGP failover may take up to MINROUTEADVER.$\max_{p \in P}|p|$ seconds, where MINROUTEADVER (currently 30 s) is a timer that states the minimum time that must expire between two consecutive advertisements, P is the set of all available paths between two remote AS, and $|p|$ is the length of a path $p \in P$. In short, considering that BGP could be extended to support QoS, this slow convergence problem of BGP could create an additional burden due to the need to achieve shorter response times while choosing paths that are able to fulfill traffic QoS requirements. However, the burden would depend on the difference in the time-scale of the traffic and the time-scale of BGP convergence.

3.3 Protocol configuration and path control

BGP enables AS to control how traffic enters or leaves their networks, reflecting TE goals and business relationships. However, it is notoriously difficult to find beforehand the proper configurations of BGP routers (i.e., the right import and export route filters). After applying them, the outcome might not be the one expected [21, 22].

Even if each AS is able to configure BGP routers properly, BGP provides little control over the end-to-end path selection and, thus, how the traffic generated by each AS is routed to the target destinations. In fact, the existing BGP techniques for controlling outbound traffic, such as the LOCAL-PREF attribute, only enable control over the first AS hop. Not to mention the issue, AS usually choose the most preferred neighbor to forward traffic, according to the economic transit cost of using such neighbor. Thereby, it becomes difficult to guarantee end-to-end QoS or to select paths that circumvent congested or unwanted AS. Furthermore, the existing techniques for inbound traffic control, such as MED tweaking, exhibit, in general, poor effectiveness because they need support from other AS; and above all, they only enable AS to achieve coarse-grained control of incoming traffic [14].

3.4 Path diversity

Path switching technique enables multihoming AS to protect traffic from network service outages or QoS degradation by searching for alternative paths able to bypass the failure or congested AS. However, the effectiveness of path switching is dependent on the paths available. Unfortunately, the BGP RIB does not contain all available paths for a given destination prefix, which can not only constrain the set of feasible options, but it can also lead suboptimal choices in terms of QoS or stability.

There are two features of the BGP routing model which are at the origin of the reduced degree of path diversity. First, BGP is a single-path routing protocol, and a BGP speaker only advertises the best paths to downstream peers. Second, the BGP decision process is deterministic. Although this characteristic confers predictability to path selection, it can reduce the path diversity if the tie-breaking rule—*to prefer the routes learned from the lowest router ID*—is often used. Unfortunately, recent observations have shown that between 40% and 50% of route selections are made using this rule [25].

3.5 Routing oscillations

One major cause of BGP instability is protocol oscillations due to policy disputes [26]. Policy disputes among AS can occur when the AS paths are selected using the LOCAL-PREF rule. This practice is relatively common due to the need to reflect business relationships among AS or preference in using transit traffic services from certain AS. The reason why BGP might not converge is that, after any AS selects and advertises its best

paths, an AS in the system might switch to a better path, causing the stable paths problem (SPP). These interactions among policies may lead to persistent route oscillations [23]. The SPP problem has been shown to be NP-complete. A heuristic based on the notion of a "dispute wheel" (i.e., a circular set of conflicting policies) has been proposed and it has been shown that "if no dispute wheel can be constructed, then there exists an unique solution for the SPP" [24]. In other words, this result implies that the set of policies in the system does not oscillate.

4 Research efforts on improving interdomain routing QoS

This section presents and discusses relevant work that aims at facing the problem of improving QoS support of current interdomain routing. The work analyzed includes existing approaches to the problem, which can be classified into two main broad classes (and thus two themes of researching), namely, QoS extensions to BGP and traffic control schemes.

4.1 QoS extensions to BGP

A straightforward approach to enable BGP to support QoS aware routing is to add new BGP attributes that carry the QoS information within UPDATE messages. QoS enhancements to BGP based on the definition of two new BGP QoS attributes have been proposed [27, 28]. The first proposal defines a variable length, optional, and nontransitive BGP QoS attribute that allows a domain to decide which type of QoS information a BGP border router redistributes. The nontransitivity property implies that, if the attribute is not supported by a BGP speaker, it must not forward to its peers. This attribute associates with the announced prefix, the DiffServ per-hop behavior (PHB), namely, best effort, assured forwarding and expedited forwarding, supported by BGP QoS-aware routers, type and value of QoS parameters (e.g., maximum bandwidth associated to PHB, or available bandwidth, or even maximum or minimum delay), or the required QoS signaling (e.g., indication that the border router supports RSVP).

The second proposal defines an optional and transitive BGP QoS attribute, named QoS_NLRI (network layer reachability information). This attribute allows carrying three pieces of QoS information: the type of QoS information (e.g., packet rate, delay, PHB), the subtype of QoS information (reserved rate, available rate, loss rate, min/max/average delay), and the value of the QoS information identified in the previous fields. Figure 5 illustrates a possible scenario of using this attribute.

Both proposals of QoS extensions to BGP have drawbacks. First, the use of in-band signaling results in low convergence and instability problems. Second, none of these proposals suggest any solution to represent dynamic changes on network state. Another important issue not addressed is the potential problems raised by the nonuniform semantic of QoS information, i.e., AS might use different types of QoS information, with different meanings, and precision. One interesting possibility to solve this problem is to use a common external representation to keep a uniform semantic of distributed QoS information, such as the IST Mescal Meta-Class concept, to characterize a domain transfer capability [15].

A new interdomain QoS metric, called index available bandwidth index (ABI), has been designed to face two main challenges in extending BGP, namely, scalability and heterogeneity of interdomain links [29, 48]. Therefore, the ABI index is conceived as a

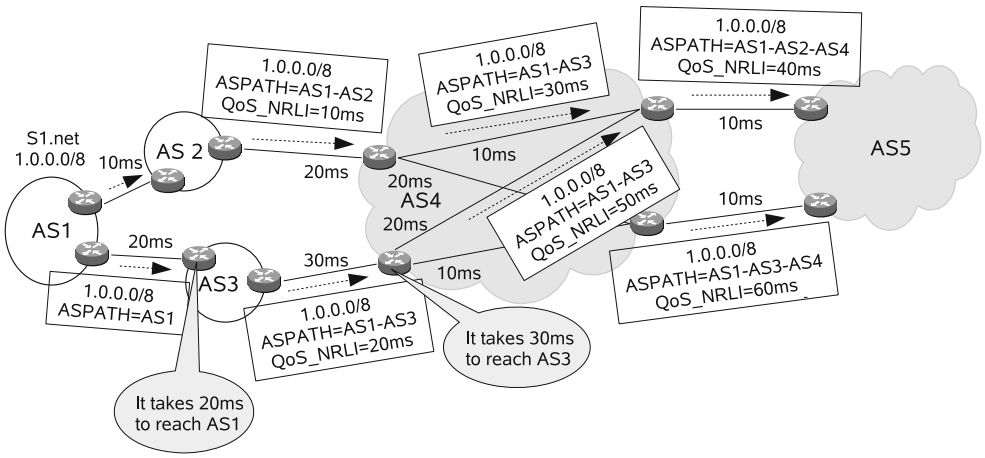


Fig. 5 QoS_NLRI capable BGP speakers with path delay information

semidynamic metric, defined as the probability that the available bandwidth belongs to a given interval. For path ABI propagation between BGP-peers, a similar approach is used to the one proposed in [28]. At the decision process level, the decision criterion used is the maximum weight, W, of a QoS route. The idea of using weights representing ABI indices is to facilitate the comparison between the ABIs of the available path options. Finally, two thresholds are proposed as stability mechanisms: route update threshold, a new route is installed into the local routing table only if its W is significantly better than the W of the actual route, and the link state threshold, to control UPDATE generation, such that only important variations in the available bandwidth of a link are propagated.

The ABI proposal solves the problem of the negrained notification of dynamic changes in network state. It also makes the precision of QoS information uniform. However, the proposal is based on the controversial assumption of allocating reasonable CPU processing resources to the execution of mathematical operations during the computation of ABI indices of links and paths. This feature, at the current Internet scale, could affect the scalability of the solution.

4.2 Traffic control schemes

There are two main approaches for traffic control at the interdomain level, namely, Internet-wide approaches resorting to overlay network-based mechanisms and end-point approaches using multihoming and smart routing-based mechanisms. This section details the most relevant solutions developed according to both paradigms.

4.2.1 Internet-wide approach: overlay network-based mechanisms

Overlay networks solutions were developed to overcome the disadvantages of plain BGP extensions. With this approach, a large number of overlay entities is strategically placed across several AS. In general, the role of these nodes is to periodically monitor the performance and availability of paths between them. Then, once an Internet route fails or does not perform as expected, the overlay user shifts its traffic to an alternate route.

There are two main groups of solutions that resort to overlay networks, according to the level of interaction with the underlying routing layer. The first group of solutions decouples part of the policy control portion of the routing process from BGP devices. In this line, the Overlay Policy Control Architecture (OPCA) has been proposed to enhance the BGP's fail-over and its limitation on the control of incoming traffic [30]. In this architecture, overlay entities called policy agents (PA), communicating via a new protocol, overlay policy protocol, process incoming policies or route changes constrained to local AS policies (e.g., pricing constraints or SLA). The key requirement of OPCA is the knowledge of AS relationships. In this sense, it includes a centralized AS topology and relationship mapper (RMAP) to deduce AS relationships from BGP routing table dumps. Then, when a PA detects a failure, it queries the RMAP to discover which remote PA should be contacted in order to bypass the difficulty. The main drawbacks of this scheme are low scalability and inaccuracy. In fact, all PA actions are dependent on the RMAP component and on its ability to deduce inter-AS topology.

In the second group of solutions, known as pure-overlays, an additional routing layer independent from the underlying routing is used. The major advantage of this approach is an easier implementation, since it enables QoS support and resilience capability, without requiring that physical links along logical paths employ QoS mechanisms (e.g., scheduling or buffer management). Two examples of this approach are presented in [31, 32]. The first proposal uses a mechanism called controlled-loss virtual link to provide per-flow bandwidth differentiation, rate assurance, and congestion control. The second proposal presents a complete set of mechanisms for QoS routing, including hierarchical aggregation for overlay networks. Similar to OPCA, overlay brokers are strategically placed across domains to form an overlay service network, which provides a unified access by QoS applications to routing and resource allocation.

The main difference between OPCA and pure overlays is that the latter circumvents BGP to route packets, which can lead to violations of the commercial policies between AS and undesired interactions with the underlying infrastructure [40]. In addition, although well-known studies have shown the validity of pure-overlays to offer better performance (e.g., throughputs, RTTs, loss rates, and path availability), it might not be enough to ensure the QoS levels required due to the absence of control of the underlying infrastructure behavior [33].

There are two main issues to investigate regarding the development of these solutions. First, the degree of cooperation between AS required by OPCA-like solutions is unclear. Therefore, future research should answer the following questions: *Do we need high levels of cooperation to achieve good levels of resilience and performance? How will the cooperation be achieved*?

Second, because pure overlays might violate the commercial policies between ISPs, *can we, with much less routing flexibility, still achieve the same levels of resilience and performance*?

4.2.2 End-point approach: multihoming and smart routing-based mechanisms

Multihoming consists of the increasing of Internet connectivity by contracting multiple broadband lines (e.g., Business DSL, E1, E2, or E3) from two or three different ISPs. Studies have shown that multihomed stub AS experience a potential performance benefit compared to single homed AS of at least 40%, as well as significant reliability benefits [34]. However, the use of the multihoming technique by itself is not enough to obtain such improvements because interdomain routing of IP packets still relies on BGP.

Routing mechanisms, referred as smart route controller (SRC) systems are, thus, being increasingly used by multihomed stub AS, as they provide a holistic way to solve local end-to-end traffic challenges (e.g., latency, or loss rate bounds) through shifting some traffic between ISPs in short timescales. One key function of SRCs is, thus, to capture the performance of paths. To address this issue, an SRC usually employs active path monitoring methods [45].

Figure 6 illustrates a simple scenario of two AS employing SRC, where the SRC of AS2 might improve the performance of the outbound traffic toward the remote stub AS, AS3, through switching among the paths AS3–ISP1–ISP3 and AS3–ISP1–ISP4–ISP5 across ISP3 and ISP5, respectively.

In contrast to pure overlays, SRCs never circumvent BGP to meet the requirements of traffic. This way, the additional complexity needed to cope with BGP inefficiency is set apart from BGP. In other words, although SRCs interact with BGP, they do not require any changes in BGP routers nor support from ISPs or the cooperation with AS along the paths. SRCs run as a local or remote process, which only require access to the RIBs to collect available paths and to issue command scripts to routers in a shorter timescale than BGPs timescale (to indicate the ranks of paths, typically, by tuning their LOCAL-PREFs). Furthermore, in contrast to OPCA, it does not require any further support from AS along the paths.

To conclude the presentation of SRCs, it is worthy to notice that SRCs, as they were introduced, are not a new concept, since many companies have been devoting efforts to research and develop SRC products [35, 36]. However, little is known about the technical details of commercial SRCs. On its turn, the research community has produced some publications devoted to the design and stability issues of SRCs [37, 38].

Besides the research topics described, there are two additional important issues to address. First, *is it unclear if the levels of route control, performance, and reliability offered by the Smart Routing are enough?* If not, would it be beneficial to combine smart routing with overlays in a hybrid mechanism to deal this issue, taking advantage of the best of both worlds: the simplicity of smart routing and the routing flexibility of overlays?

The second issue is motivated by the fact that smart routing, as well as overlays, is typically selfish in nature. That is, smart routing greedily select paths, observing only local traffic goals. Unfortunately, this behavior does not necessarily lead to the best routing in the Internet [39]. Regarding this issue, inter-AS cooperative routing seems a promising approach to address selfishness; however, again, it could demand undesirable levels of cooperation between AS [41].

5 Extend or replace BGP?

This section gives a broad perspective on challenges surrounding the issue of whether to extend or replace BGP to support QoS.

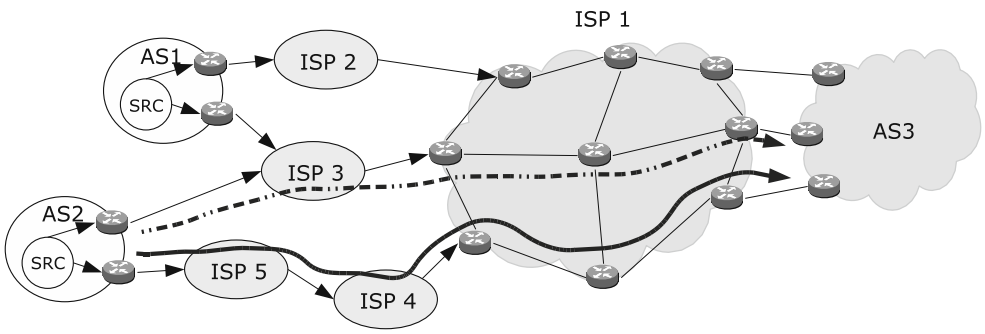


Fig. 6 A simple scenario of two multihomed stub AS employing smart routing control

5.1 Challenges to interdomain QoS routing deployment

The issue of whether to extend or replace BGP to support additional features, such as QoS, is still under discussion from both economic and technical perspectives. One significant capability added recently to BGP, the multiprotocol extensions to BGP, has intensified this discussion inside the IETF. From this discussion, two alternate perspectives for the use of the BGP infrastructure were produced. In the first, BGP is used as a general purpose transport (GPT) infrastructure and, in the second, BGP is used as a special purpose transport (SPT) infrastructure [42]. The key idea of GPT is to use the BGP data distribution mechanism as a generic application transport mechanism. On the one hand, the main concern of GPT is to observe whether the data distribution application requirements can be satisfied by the BGP data distribution mechanism. On the other hand, SPT assumes that the BGP data distribution mechanism has been designed to carry routing information. One of main concerns is to ensure that additional complexity added to BGP is bounded so that it does not cause BGP instability.

Risk, interference, and application fit are important concepts that might be used to describe the trade-offs involved when extending BGP [42]. Risk is focused on robustness trade-offs, modeling the impact of the addition of a new application on an existing implementation. Interference is focused on how a new application affects the behavior of existing applications. In other words, interference relates to the coupling or interdependence among applications. Application fit refers to how the requirements of the data to be conveyed match the BGP data distribution mechanism. As a result, given the concerns of SPT and GPT models, it implies that SPT is sensitive to risk and interference, and GPT is focused on application fit.

Following this, using this terminology and these concepts, we discuss significant challenges associated with interdomain QoS routing deployment. However, in this discussion, we consider the SPT approach, assuming that the QoS information requirements match the BGP data distribution mechanism. More specifically, when adding QoS to BGP as the protocol implementation is modified, we consider there is an intrinsic risk to degrade and destabilize the BGP behavior. In particular, this might happen in the case when multiple classes of service are added. This behavior is similar to the case when BGP carries multiple application data types, which may cause interference among the multiple applications, destabilizing also the BGP routing system.

Figure 7 summarizes the main trade-offs between the addition of QoS and the stability, scalability, and accuracy aspects of interdomain routing, which allow to analyse the risk and interference profiles. Next, the trade-offs between QoS and scalability and between QoS and stability are analyzed.

Regarding the QoS vs scalability trade-off, it is important to understand the effects of adding QoS on memory requirements and on CPU load at BGP speakers. The main factors that impact memory requirements at legacy BGP speakers, include the number of IP prefixes/networks (N), the mean AS distance in terms of hop count (M), the total number of unique AS paths (A), the mean number of BGP peers per BGP speaker (S), and the lengths of the binary words required to store a network (R) and to store an AS number (P). Then, an estimate of memory requirement (MR) at legacy BGP speakers is given by $\mathrm{MR} = (((N * R) + (M * A) * P) * S)$ [43]. However, when adding a QoS attribute to BGP (e.g.,

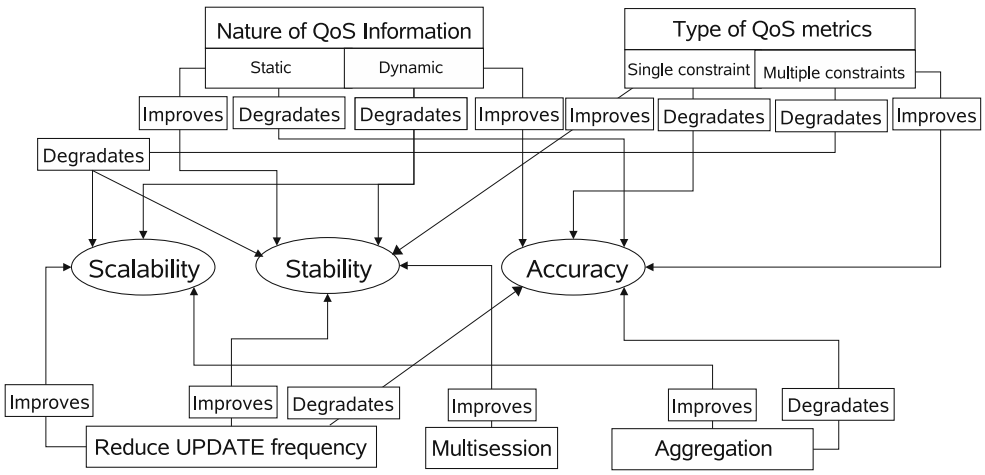
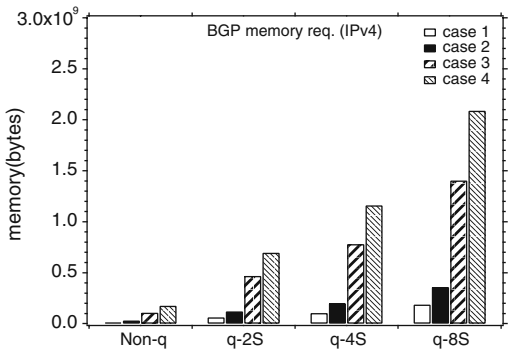


Fig. 7 q-BGP design trade-offs



	Networks (N)	Mean AS Distance (M)	Unique Paths (A)	BGP peers (S)
Case #1	100000	20	3000	20
Case #2	100000	20	15000	20
Case #3	120000	10	15000	100
Case #4	140000	15	20000	100

Fig. 8 BGP memory requirements (IPv4)

QoS_NLRI attribute) an estimate of MR (ignoring implementation details) at QoS-enabled BGP (q-BGP) speakers can be given by $\mathrm{MR}_q = ((((N * (R + L)) + (M * A) * P) * S)) * C$. In the previous equation, L represents the minimum length of the binary word to store the QoS information carried on the QoS attribute and C represents the number of supported services, including the best-effort service. Figure 8 illustrates the growth of q-BGP MR for the cases presented in [43], considering $L = 4$ bytes and up to eight services. As we can observe, it is clear that even for just two or four additional services QoS extensions will undoubtedly require tighter memory requirements to store the additional amounts of routing state information.

The second significant scalability concern, as mentioned before, is the BGP CPU load. One the one hand, QoS extensions will require more paths to be advertised per prefix destination, and thus, more processing. One the other hand, there is also some correlation between the BGP dynamics and the CPU utilization. That is, the number of BGP UPDATE message announcements received on a given period of time might increase the routers' CPU load, as every update demands some processing for route in-filtering, route selection, RIB updates, FIB updates, and route out-filtering. Measurements have confirmed this correlation especially when BGP routing tables are unstable [44]. These have also shown that high router CPU loads can increase the convergence times and Internet outages. Because the state information of the network needs to be distributed more often, QoS extensions have a potential to consume idle CPU cycles, and thus, they can exacerbate this problem.

To illustrate this point, Fig. 9 presents a rough estimation of the additional CPU load at QoS-enabled BGP routers, as a function of the fraction of the number of routes that have just changed, the number QoS services supported and the ratio between the timescale of standard BGP data distribution and the corresponding timescale of q-BGP. In this figure, to infer the additional CPU load experimented by a q-BGP router, we considered as reference the worse-case scenario experimented by a legacy BGP router, that is when a BGP router needs to process all BGP UPDATES to all destinations (i.e, N) from all BGP peers (i.e., P). As we observe, the potential additional CPU load increases very rapidly as we increase the number of services and the frequency of BGP UPDATE messages. For instance, when only about 6% of the routes have changed (which is common in ISP), q-BGP routers supporting four services and eight services would need, respectively, about 50% and 100% of the total number of CPU cycles

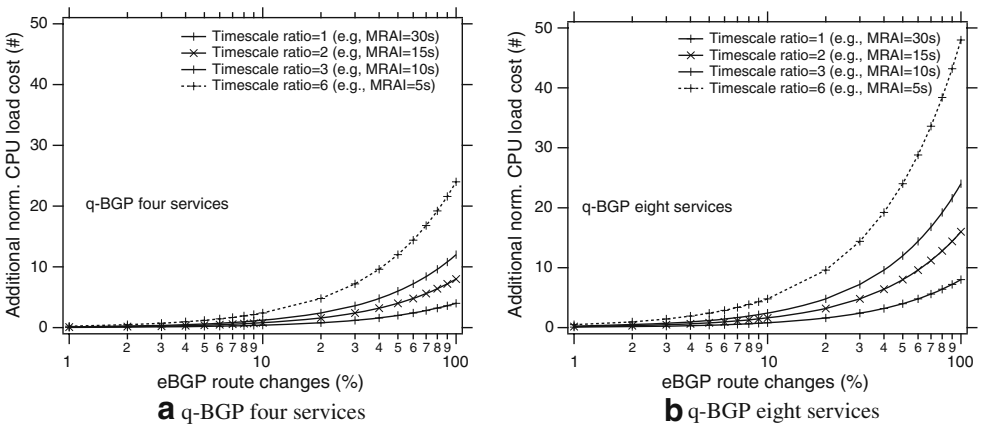


Fig. 9 Rough estimation of the additional CPU load (**a**, **b**)

needed to handle the worse-case scenario of regular BGP for a MRAI timer equal to half of the original value. Notice that we did not attempted to model in detail the BGP's router load because it depends of several factors that are really hard to quantify, such as the real reaction of the CPU load to different kinds and volumes of BGP messages and the influence of the prefix compression capability of BGP, not to mention the significant influence of the router operating system.

Regarding the QoS vs stability trade-off, one first issue to understand is how to manage the problem of adding the class-based routing feature to the BGP implementation without introducing instability. This issue could be solved by adding two mechanisms. First, the BGP data distribution mechanism should be able to advertise multiple classes of routing information. Part of this issue is already solved with the introduction of the multiprotocol extensions, which enable BGP to transport information for multiple address families and subfamilies, distinguished by distinct address family identifier (AFI)/subsequent address family identifier (SAFI). However, the actual point of multiplexing is located at the BGP layer, which constitutes a serious BGP robustness problem. That is, one single corrupt message for a given AFI/SAFI might terminate the BGP session and compromise other AFI/SAFI. One solution for this problem is to move the point of multiplexing of this data into the transport layer and, thus, to allow multiple sessions between two BGP peers [16]. However, it still lacks a mechanism that enables BGP to advertise multiple classes of information for the same address.

The second issue to understand is the nature of the QoS information. On the one hand, when using static QoS information, such as the distribution of the ID of the supported classes of service or maximum available bandwidth toward a prefix, it can improve the stability of routing at the cost of adding some inaccuracy into the routing state. On the other hand, when using dynamic QoS information, such as available bandwidth toward a prefix, it can improve the accuracy of the routing information. However, it can almost certainly to introduce some additional instability in to the interdomain routing system. In addition, the choice of dynamic QoS information demands important modifications in the BGP decision process as, in this case, the selection of the best routes depends on the QoS information.

5.2 External challenges to BGP

Often protocol designers and engineers focus only on the technical facet of engineering problems, and do not take into account their economic context. For instance, the modest success of QoS architectures (e.g., IntServ and DiffServ) can be partially justified by this decoupling between the technical and economic aspects of the various problems to solve. This is an important lesson to learn. From our perspective, besides a clear demonstration of the potential QoS benefits, the economic features should also be taken into account by proposals which aim at adding QoS to BGP. Recognizing this need, two recent proposals include economic frameworks to motivate ISPs to provide good QoS at predictable costs to end-users as the prices charged by ISPs are public [46, 47].

From an operational standpoint, one important issue that should be addressed is the objection of network administrators to complexity and their reluctance to changes. Moreover, the provision of additional bandwidth to links is an attractive alternative to QoS; it is simple, it works, and it is becoming cheaper. In short, a concrete proposal will not become compelling for the majority of AS, and therefore, it will not be widely deployed, if it cannot answer the following questions:

1. *What does the network of my AS have to gain if q-BGP is adopted?*
2. *Does the additional complexity introduced by q-BGP makes the configuration of BGP routers or the debugging of network problems more difficult?*
3. *Can the q-BGP solution be incrementally deployed?*

Understanding the granularity of the routing problem and common operational networking practices and tools in an AS are also important aspects to consider. For instance, even though the AS is the base unit of interdomain routing, previous proposals of adding QoS to BGP still consider quite unrealistic models of AS. More specifically, these proposals modeled each AS as just single nodes with 2–3 peers. In contrast, present measurements clearly show that there are a significant number of AS that are composed of hundreds of BGP speakers, and have some hundreds of peers (e.g., AS Numbers 701, 7018 or 1239). Another important limitation of previous schemes is the lack of the definition of interfaces with other auxiliary or fundamental layers or mechanisms on Inter-AS QoS provisioning, such as a SLA management layer or TE.

These issues lead to an important question that must be clearly answered: *what would the role of BGP be on providing inter-AS QoS*? Probably, considering these two additional driving forces, QoS-managed AS would prefer to model BGP as a GPT infrastructure that would assist AS in SLA trading or to control inter-AS QoS interconnections, instead of considering BGP as a routing protocol. In other words, this perspective is appealing for the development of BGP-based

mechanisms to control inter-AS QoS interconnections. The combination of these mechanisms with, for instance, DiffServ bandwidth brokers (BB) can be used to exchange and negotiate the conditions about QoS connectivity services (e.g., bandwidth and latency bounds, routing, pricing, and penalties) among peers. The role of the BB-like entities would be to allocate and control shared resources (e.g., bandwidth), as well as to make decisions about QoS interconnection policies.

6 Summary

In this article, we have surveyed research work that aims at facing the problem of interdomain QoS routing, presenting also the main short-comings of each proposal. However, the discussions about the requirements for the future interdomain routing architecture and about whether these requirements are best met by an approach of introducing changes into BGP or by replacing BGP is still open. In particular, we emphasize the fact that, while some challenging issues reside in the deployment of q-BGP, others are derived from external challenges issues to BGP. In summary, our aim in this paper was basically to:

- Support the necessity of tackling interdomain QoS routing.
- Clearly expose the most important open issues in the area of interdomain QoS routing.
- Briefly present an up-to-date set of proposals that address some of the challenges in interdomain QoS routing.
- Argue that, unless the role of BGP would be rethought to include also the culture of operational networking, interdomain routing will continue to suffer from a chronic failure, that is, the lack of QoS support.

References

1. Blake S et al (1998) An architecture for differentiated services. IETF RFC 2475, IETF
2. Evans J, Filsfils C (2004) Deploying Diffserv at the network edge for tight SLAs, part I. IEEE Internet Comput 8(1): 61–65
3. Akella A, Seshan S, Shaikh A (2003) An empirical evaluation of wide-area internet bottlenecks. In: Proc of ACM SIGCOMM/USENIX internet measurement conference 2003, Miami, 27–29 October 2003
4. Markopoulou A, Iannaccone G, Bhattacharyya S, Chuah C, Diot C (2004) Characterization of failures in an IP backbone. In: IEEE Infocom2004, Hong Kong, 7–11 March 2004
5. Apostolopoulos G, Kama S, Williams D, Guerin R, Orda A, Przygienda T (1999) QoS routing mechanisms and OSPF extensions. IETF RFC 2676
6. Filsfils C, Evans J (2002) Engineering a multiservice IP backbone to support tigh SLA. IEEE Comput Netw 40(1):131–148
7. Crawley E, Nair R, Rajagopalan B, Sandick H (1998) A framework for QoS-based routing in the internet. IETF RFC 2386
8. Rekhter Y, Li T, Hares S (2006) A border gateway protocol 4 (BGP-4). IETF RFC 4271, IETF
9. Chandra R, Scudder J (2002) Capabilities advertisement with BGP-4. IETF RFC 3392, IETF
10. Chandra R, Traina P, Li T (1996) BGP communities attribute. IETF RFC 1997, IETF
11. Bates T, Chandra R, Katz D, Rekhter Y (2007) Multiprotocol extensions for BGP-4. IETF RFC 4760, IETF
12. Caesar M, Rexford J (2005) BGP routing policies in ISP networks. IEEE Netw
13. Quoitin B, Tandel S, Uhlig S, Bonaventure O (2003) Interdomain traffic engineering with redistribution communities. Comput Commun J 27(4):355–363
14. Chang R, Lo M (2005) Inbound traffic engineering for multihomed ASs using as path prepending. IEEE Netw 19(2): 18–25
15. Management of End-to-end Quality of Service Across the Internet at Large (2003) Specification of business models and a functional architecture for inter-domain QoS delivery. Deliverable D1.1 of the IST MESCAL project interdomain QoS for the Internet. www.mescal.org
16. Scudder J, Appanna C (2007) Multisession BGP, draft-ietf-idr-bgp-multisession-03.txt, IETF draft, IETF
17. Savage S et al (1999) The end-to-end effects of internet path selection. In: Proc of ACM SIGCOMM 99, Cambridge, 31 August–3 September 1999
18. Huffaker B, Fomenkov M, Plummer D, Moore D, Claffy K (2002) Distance metrics in the internet. In: Proc of IEEE international telecommunications symposium (ITS), Minneapolis, August 2002
19. Labovitz C, Wattenhofer R, Venkatachary S, Ahuja A (2001) The impact of internet policy and topology on delayed routing convergence. In: Proc of IEEE INFOCOM 2001, Anchorage, 22–26 April 2001
20. Labovitz C, Ahuja A, Bose A, Jahanianitz F (2001) Delayed internet routing convergence. In: IEEE/ACM transactions on networking, vol 9, no 3. IEEE, Piscataway
21. Mahajan R, Wetherall D, Anderson T (2002) Understanding BGP misconfiguration. In: Proc of the ACM SIGCOMM 2002, Pittsburgh, 19–23 August 2002
22. Feamster N, Rexford J (2002) Network-wide BGP route prediction for traffic engineering. In: Proc workshop on scalability and traffic control in ip networks, SPIE ITCOM conference, Boston, 29–31 July 2002
23. Varadhan K, Govindan R, Estrin D (1996) Persistent route oscillations in inter-domain routing. ISI technical report 96-631, USC/Information Sciences Institute
24. Griffin T, Shepherd B, Wilfong G (2002) The stable paths problem and interdomain routing. IEEE/ACM Trans Netw 10:232–243
25. Launois C (2004) Leveraging internet path diversity and network performances with IPv6 multihoming. http://www.info.ucl.ac.be/people/delaunoi/diversity/
26. Labovitz C, Malan GR, Jahanian F (1997) Internet routing instability. SIGCOMM Comput Commun Rev 27(4):115–126

27. Bonaventure O (2001) Using BGP to distribute flexible QoS information. draft-bonaventure-bgp-qos-00, IETF draft, IETF
28. Cristallo G, Jacquenet C (2002) Providing quality of service indication by BGP-4 protocol: the QOS_NLRI attribute. IETF draft, IETF
29. Xiao L et al (2002) QoS extension to BGP. In: 10th IEEE international conference on network protocols (ICNP'02), Paris, 12–15 November 2002
30. Agarwal S, Chuah C, Katz R (2003) OPCA: robust interdomain policy routing and traffic control. In: Proceeedings of the 6th international conference on open architectures and network programming (Openarch). IEEE Communications Society, San Francisco
31. Li Z, Mohapatra P (2004) QRON: QoS-aware routing in overlay networks. IEEE J Sel Areas Commun 22:29–40
32. Subramanian L, Stoica I, Balakrishnan H, Katz R (2002) OverQoS: offering QoS using overlays. In: Proc of first workshop on hop topics in networks (HotNets-I), Princeton, October 2002
33. Andersen D et al (2001) Resilient overlay networks. In: Proc of the 18th ACM symposium on operating system principles, Banff, 21–24 October 2001
34. Akella A et al (2003) A measurement-based analysis of multihoming. In: The proc of ACM SIGCOMM 2003, Karlsruhe, 25–29 August 2003
35. Cisco Systems (2008) Optimized edge routing. Cisco, San Jose
36. Internap Networks (2008) Flow control platform. Internap Networks, Atlanta
37. Guo F et al (2004) Experiences in building a multihoming load balancing system. In: Proc of IEEE INFOCOM 2004, Hong Kong, 7–11 March 2004
38. Gao R et al (2006) Avoiding oscillations due to intelligent route control systems. In: The Proc of IEEE INFOCOM 2006, Barcelona, 23–29 April 2006
39. Qiu L et al (2003) On selfish routing in internet-like environments. In: Proc of ACM SIGCOMM 2003. Kalsruhe, 25–29 August 2003
40. Liu Y et al (2005) On the interaction between overlay routing and underlay routing. In: Proc of IEEE INFOCOM 2005, Miami, 19–17 March 2005
41. Shrimali G et al (2007) Cooperative interdomain traffic engineeing using nash bargaining and decomposition. In: Proc of IEEE INFOCOM 2007, Anchorage, 6–12 May 2007
42. Meyer D (2004) Operational concerns and considerations for routing protocol design—risk, interference, and fit (RIFT). draft-ietf-grow-rift-01.txt, IETF draft, IETF
43. Meyer D, Patel K (2006) BGP-4 protocol analysis. IETF RFC 4274, IETF
44. Agarwal S, Chuah C, Bhattacharyya S, Diot C (2004) Impact of BGP dynamics on router CPU utilization. In: Proc of the passive and active measurement workshop, Antibes Juanles-Pins, 19–20 April 2004
45. Paxson V, Almes G, Mahdavi J, Mathis M (1998) Framework for IP performance metrics. IETF RFC 2330, IETF
46. Yahaya A, Suda T (2006) iREX: inter-domain QoS automation using economics. In: Proc of IEEE CCNC, Las Vegas, 7–10 January 2006
47. Estan C, Akella A, Banerjee S (2007) A la carte: an economic framework for multi-isp service quality. Tech Report ID 1591, University of Wisconsin-Madison
48. Xiao L, Wang J, Lui K-S, Nahrstedt K (2004) Advertising interdomain QoS routing information. IEEE J Sel Areas Commun 22(10):1949–1964