Open Issues in Interdomain Routing: A Survey

Marcelo Yannuzzi and Xavier Masip-Bruin, Technical University of Catalonia, Spain
Olivier Bonaventure, Université catholique de Louvain, Belgium

Abstract

This article surveys several research challenges in interdomain routing. We introduce and describe these challenges in a comprehensible manner, along with a review of the most compelling contributions and ongoing research efforts addressing each of the exposed issues. During this analysis we identify the relation between these research challenges and how they influence each other. We also present our perspectives on why these issues remain largely unsolved, and point out why some of the proposals made so far have not yet been adopted. We hope this can provide some insight on future directions in this complex research area.

At present, interdomain routing is considered a challenging research area [1]. This is mainly rooted in the following two facts.

First, the interdomain routing protocol currently used in the Internet has several limitations, but its replacement is not a realistic option due to its worldwide deployment. These limitations are becoming especially noticeable given the explosive growth the network has experienced in these last few years [2]. This growth refers not only to the size of the network, but also to the amount of and variety of the applications actually available on the Internet. This growth tendency is placing significant stress on both the scalability and capabilities of the interdomain routing protocol.

Second, as its name indicates, interdomain routing denotes routing among distinct domains or networks. These domains are completely autonomous entities, which perform their own routing management based on policies that only have local significance. In this scenario conditions such as business and competition between domains, along with fully independent management using potentially conflicting policies, makes the problem of interdomain routing even harder.

The goals of this article are first to present an up-to-date inspection of some of the main open issues in interdomain routing. Second, we intend to survey the state of the art and briefly describe some of the most relevant proposals in the area. Third, we seek to point out why these issues are so difficult to solve at present, and succinctly explain why most of the existing proposals have never moved into a deployment stage. Our aim is to put things in perspective and summarize the main lessons learned.

This article is addressed to both nonexperts, and researchers and professionals familiar with this particular research area. The rest of the article is organized as follows. We provide a brief introduction to interdomain routing so that nonexpert readers can become acquainted of the framework on which the rest of the article is developed. We present an up-to-date analysis of several of the main research challenges in the area. We also describe the most appealing approaches addressing each of these challenges, and explain why, despite this, they remain largely unsolved. Finally, we conclude the article.[1]

The Basics of Interdomain Routing

The current Internet is a decentralized collection of computer networks from all around the world. Each of these networks is typically known as a domain or an autonomous system (AS). An AS is in fact a network or group of networks under a common routing policy, and managed by a single authority. Today, the Internet is basically the interconnection of more than 20,000 ASes [4]. Every one of these ASes usually uses one or more interior gateway protocols (IGPs), such as Intermediate System to Intermediate System (IS-IS) or Open Shortest Path First (OSPF), to exchange routing information within the AS. This is known as *intradomain routing*. On the other hand, *interdomain routing* focuses on the exchange of routes to allow the transmission of packets between different ASes.

Figure 1 illustrates a simplified (but typical) interdomain scenario depicting the interconnection of several ASes. All the ASes represented in the figure have multiple connections to the network. This is indeed a common practice nowadays, and it is mainly used for resilience and load balancing. When an AS is connected to multiple different ASes, it is referred to as a *multihomed* AS. On the other hand, ASes connected to a single AS are known as *single-homed* ASes. To fix ideas, all the ASes present in Fig. 1 are multihomed except AS3. Even though AS3 is dually connected to the Internet, both connections are with the same AS (AS31).

The Internet is composed of three different types of ASes:

- Single-homed stub ASes such as AS3 in Fig. 1
- Multihomed stub ASes such as AS1 and AS2 in Fig. 1
- Transit ASes, which can be classified into very large transit ASes making up what is usually referred to as the Internet core, and smaller-sized transit ASes such as AS11, AS12, AS21–AS23, and AS31 in Fig. 1.

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

[1] Due to space limitations, we were unable to document all open issues and cite all relevant references. A more detailed version of this article with a longer list of references is available from [3].

0890-8044/95/$05.00 1995 © IEEE

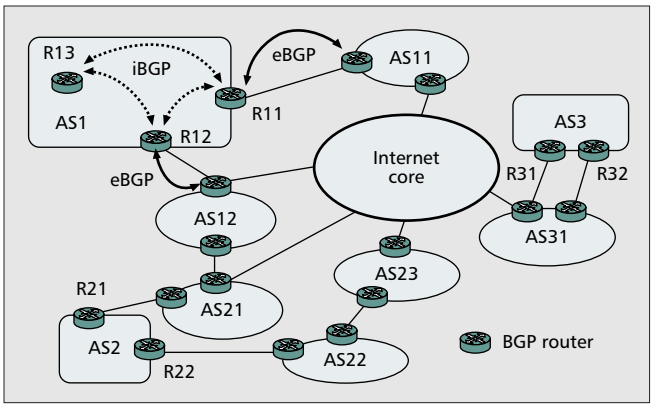


Figure 1. *A simplified interdomain scenario.*

The two types of stub ASes crowd together mostly medium and large enterprise customers, content service providers (CSPs), and small network service providers (NSPs). These two groups correspond to the largest fraction of ASes present in the Internet. The third type includes most Internet service providers (ISPs) in the Internet.

In today's Internet, there is a hierarchy of transit ASes [5]. This hierarchical structure is rooted in the two different types of relationships that could exist between ASes (i.e., customer-provider or peer-to-peer). Thus, for each transit AS any directly connected AS is either a customer or peer. At the top of this hierarchy we found the largest ISPs, which are usually referred to as Tier-1 ISPs. There are about 20 Tier-1s at present [5], which represents less than 0.1 percent of the total number of ASes in the Internet [4]. These Tier-1s are directly interconnected in almost a full mesh and compose the Internet core. In the core all relationships between Tier-1s are peer-to-peer, so a Tier-1 is any ISP lacking an upstream provider. The second level of the hierarchy is composed of Tier-2 ISPs. A Tier-2 is any transit AS that is a customer of one or more Tier-1 ISPs. A representative example of a Tier-2 ISP is a national service provider. Tier-2 ISPs tend to establish peer-to-peer relationships with other neighboring Tier-2s for both economical and performance reasons. This is typically the case for geographically close Tier-2 ISPs that exchange large amounts of traffic. There are also Tier-3 ISPs, which are those transit ASes in the hierarchy that are customers of one or more Tier-2 ISP, such as regional ISPs within a country. Stub ASes are non-transit ASes that are customers of any ISP (Tier-1, Tier-2, or Tier-3). In Fig. 1 ISPs such as AS11, AS12, AS21, AS23, and AS31 would be classified as Tier-2 ISPs, while AS22 represents a Tier-3 ISP. An important corollary of this hierarchical structure is that the diameter of the Internet is very small in terms of AS hops.

The Border Gateway Protocol (BGP) is currently the de facto standard interdomain routing protocol in the Internet. Its current official[2] release is BGP-4, which was specified in [6] on March of 1995. BGP is used to exchange reachability information throughout the Internet, and it is mainly an inter-AS routing protocol. However, the reachability information an AS learns from the exterior needs to be distributed within the AS so that every router in the AS could properly reach destinations outside the AS. When reachability information is exchanged between two BGP routers located in different ASes, the protocol is referred to as external BGP (eBGP). On the other hand, when reachability information is exchanged between BGP routers located inside the same AS, the protocol is referred to as internal BGP (iBGP).

For instance, in AS1 the reachability information R11 learns from AS11 is received over eBGP. This information is passed from R11 to the routers inside AS1 (i.e., R12 and R13) so that they are able to reach the routes advertised by AS11. This exchange of reachability information between R11 and the internal routers in AS1 is done by means of iBGP. The same occurs for the external routes R12 learns from AS12.

For scalability reasons BGP does not try to keep track of the entire Internet's topology. Instead, it only manages the end-to-end AS path of one route in the form of an ordered sequence of AS numbers. For this reason BGP is known as a path vector routing protocol, to reflect the fact that it is essentially a modified distance vector protocol. While a typical distance vector protocol like RIP chooses a route according to the least number of routers traversed (router hops), BGP *generally* chooses the route that traverses the least number of ASes (AS hops). For example, the BGP process running in router R21 will typically choose to reach AS1 via the ASes AS21 and AS12. Thus, the AS path chosen by R21 is {AS21, AS12, AS1} (please notice that the Internet core accounts for at least one AS hop more in the AS path if only one Tier-1 ISP is traversed while reaching AS1).

The term generally mentioned before is due to the fact that the AS path length is one of the steps of the BGP decision process, but not the only one. This decision process is used for route selection each time a BGP router has at least two different routes for the same destination. Thus, BGP routing is more complex than simply minimizing the number of AS hops. BGP routers have built-in features to override the AS hop count, and to tiebreak if two or more routes have the same AS path length. The sequence of steps in Fig. 2 represents a simplified version of the BGP decision process.

In this process each subsequent step is used to break ties when the routes being compared were equally good in the previous step. The local preference (LOCAL_PREF) in step 1 and the multi-exit discriminator (MED) in step 3 are two BGP attributes that are used by BGP routers for controlling how traffic flows from and into an AS, respectively. A detailed explanation of this process can be found in [7].

After this short description of the main components and their roles in interdomain routing, we follow with some of the main open issues in this area.

Research Challenges in Interdomain Routing

In the last years the Internet has largely expanded in several ways. First, the number of ASes connected to the Internet has increased enormously [2]. Second, the number of connections per AS to the network has also significantly augmented [8]. Third, the number and diversity of the applications supported in the Internet have remarkably increased as well. This tendency has increased the demands on the scale of the network, and hence is placing significant pressure on the scalability and convergence of BGP.

In addition, the current interdomain routing structure is not precisely prepared to handle the service characteristics several applications are demanding from the network. In effect, the end-to-end performance of these applications is not only affected by the limitations of BGP, but also by the diversity of interests and lack of cooperation between the ASes composing the Internet. Therefore, several issues remain to be solved in the area of interdomain routing. This section analyzes several significant challenges faced by researchers in the area today. The methodology we follow is first to introduce the problem. Next, we survey several proposals addressing the issue, and try to discriminate which are in fact operational palliatives. After that, we discuss why despite these efforts each issue remains largely open.

The order in which the issues are presented is chosen so as to gradually introduce the distinct aspects of BGP and the interdomain routing paradigm, as well as to link how the initial set of issues influences the subsequent ones.

[2] *The IDR working group of the IETF has finalized the revision of [6]. This revision documents the currently deployed code.*

1. Choose the route with the highest local preference (LOCAL_PREF)
2. If the LOCAL_PREFs are equal choose the route with the shortest AS-path
3. If the AS-path lengths are equal choose the route with the lowest MED
4. If the MEDs are equal prefer external routes over internal routes (eBGP over iBGP)
5. If the routes are still equal prefer the one with the lowest IGP metric to the next-hop router
6. If more than one route is still available run tie-breaking rules

Figure 2. *A simplified version of the BGP route selection process.*

Slow Convergence and Chattiness of BGP

In order to exchange reachability information, two BGP routers must establish a BGP session. This session is supported by a TCP connection through which the peers exchange four different types of messages, specifically [6]:

- **OPEN** message: to open a BGP session between peers.
- **UPDATE** message: to transfer reachability information among peers. This message is used to either advertise a feasible route to a peer or withdraw infeasible routes. The UPDATE message is usually referred to as a BGP advertisement.
- **NOTIFICATION** message: sent when an error condition is detected. The BGP session is immediately shut down after this message is sent.
- **KEEPALIVE** message: periodically exchanged to verify that the peer is still reachable.

Each peer is able to determine if the BGP session corresponds to an iBGP or eBGP session from the content of the OPEN message. When a BGP session starts, each peer advertises its entire set of routes. After that, only incremental updates and KEEPALIVE messages are exchanged.

An important performance metric for a routing protocol is its convergence time (i.e., the time required to reroute packets around a failure). The first significant studies of the convergence of BGP were carried out using measurements in the Internet [9]. These studies showed that the convergence of BGP was rather slow, often measured in tens of seconds. This slow convergence is caused by several factors, some of which are inherent to the utilization of path vectors by BGP, while others are due to implementation choices. In short, this slow convergence is mainly rooted in the fact that in the global Internet, a single link failure can force all BGP routers to exchange large amounts of BGP advertisements, while exploring for alternative paths toward the affected destinations. This process is referred to as *path exploration*.

During BGP convergence, routers may need to exchange several advertisements concerning the same prefix. To avoid storms of BGP advertisements, most BGP routers use a timer called minimum route advertisement interval (MRAI), with a recommended default value of 30 s. This timer prevents BGP routers from sending a new advertisement for one prefix if the previous advertisement for the prefix was sent less than 30 s earlier [6]. This reduces the number of BGP advertisements exchanged, but may cause important BGP advertisements to be unnecessarily delayed. Griffin and Presmore showed in [10] that this arbitrary 30 s value has a huge impact on BGP convergence time. They observed that for each network topology and a particular set of experiments, there is an optimal value of the MRAI timer. This optimal value can significantly reduce the convergence time of BGP. Unfortunately, this might be extremely hard to find in practice since it varies from network to network.

To cope with flapping routers that regularly advertise and shortly after withdraw their routes, many routers implement BGP route flap damping [11]. This technique works by ignoring routes that change too often. This is necessary to avoid storms of advertisements due to flapping routers, but unfortunately it increases BGP convergence time [12].

Several authors have proposed modifications to reduce the BGP convergence time in case of failures. The ghost-flushing approach proposed in [13] improves the BGP convergence by ensuring that the messages indicating bad news are distributed quickly by the BGP routers, while good news propagates more slowly. The downside of ghost-flushing is that it does not tackle the root of the problem, i.e., path exploration. Instead, it only tries to speed up the convergence of BGP.

Other solutions such as BGP-RCN [14] and EPIC [15] improve the convergence of BGP and also reduce the number of BGP messages exchanged during the convergence by adding to each BGP message an identifier (root-cause) indicating the cause of the BGP message. With this additional information, when a failure occurs on one link, distant routers can avoid to select as their alternate path a path that is also affected by the failure but for which they have not yet received up-to-date information.

The good news is that these proposals significantly limit path exploration. The bad news is that accurately identifying the root cause of a failure still represents a challenging problem. This is first because root cause approaches require modifying BGP to add information to BGP advertisements, but ISPs are cautious about upgrading BGP. Second, they only introduce significant improvements under the assumption of extensive deployment. And most important of all, the additional information needed to identify the root cause of a failure works against the scalability of BGP.

The explanation for this latter is that for scalability reasons the BGP advertisements spawned by ISPs are often aggregated. Two levels of aggregation exist in these advertisements. First, the set of destinations advertised by BGP routers are composed by IP prefixes that aggregate several routes into a single route.[3] Second, the AS paths carried in the BGP advertisements intrinsically represent highly aggregated information, since they do not reveal any clue about the internal details of the ASes in the path (e.g., topology, state of connectivity). While the first level of aggregation reduces the size of BGP routing tables, the second tremendously reduces the amount of detail exchanged between BGP routers. The downside is loss of granularity in the reachability information each BGP router manages. In this framework pinpointing the source of a failure is almost impossible, given that different failures will produce the same BGP UPDATE message [15]. To cope with this, the BGP advertisements from ISPs should be somehow disaggregated, which unfortunately has a direct impact on BGP's scalability.

Clearly, two trade-offs exist:

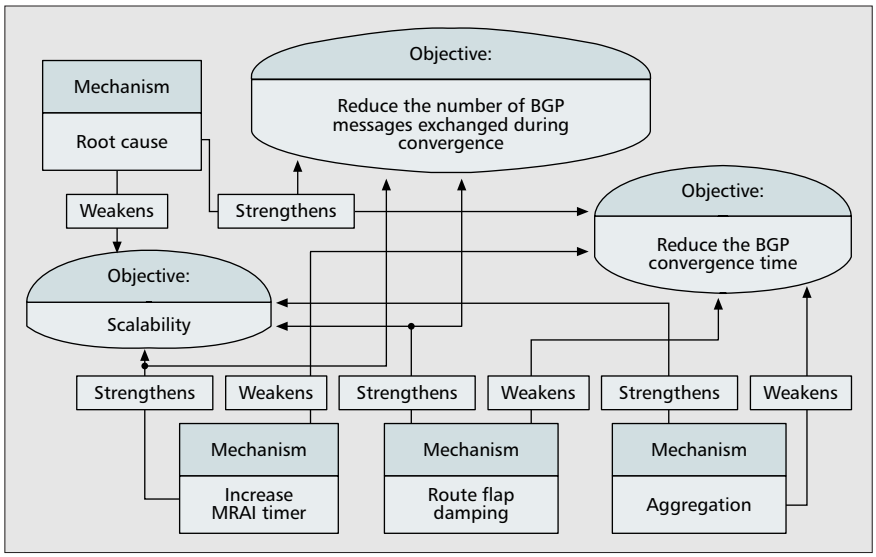
- How to disaggregate, and how much reachability information should be disaggregated in the BGP advertisements so as to accurately identify the source of a failure
- How much BGP convergence time could be reduced while keeping the overall routing system scalable

An interesting alternative to pinpoint the source of a failure without needing to modify BGP was proposed in [16]. Feldmann *et al.* propose inferring the precise location of a failure by analyzing its effects (i.e., observing the flow of BGP UPDATE messages during a convergence process). This is achieved by using multiple observation points (known as *vantage points*) and correlating the data observed along three dimensions: time, vantage point, and prefixes. However, this work proposes an offline methodology to pinpoint the source of a failure, so it was not devised as a mechanism to reduce BGP convergence time.

Figure 3 depicts three major interdomain routing objectives as well as how the set of mechanisms described above strengthen or weaken the accomplishment of these objectives. The fig-

[3] *This aggregation process will be exemplified later.*

ure shows that unfortunately, none of the existing mechanisms is able to strengthen the accomplishment of some of the objectives without weakening the accomplishment of some other. From our perspective the issue remains largely unsolved, and will remain in this state unless we thoroughly understand the intrinsic trade-offs between some of the objectives in Fig. 3 and, based on this understanding, succeed in developing novel mechanisms that could timely balance the accomplishment of all the objectives at the same time.



■ Figure 3. *The complex and still unsolved balance between three interdomain routing objectives.*

Scalability Problems Due to Multihoming

Several studies such as [17] have shown that BGP routing tables are growing significantly fast, which imposes a considerable pressure on the scalability of BGP. In the early 1990s, such a growth resulted in the definition of the CIDR IP address allocation architecture. The main reason for the recent growth lies in the fact that most stub ASes have chosen to increment their connectivity to the Internet for both resilience and load balancing reasons. As stated earlier, this practice of connecting to multiple ISPs is known as multihoming. To explain how multihoming affects the size of BGP routing tables, let us consider the example in Fig. 4. Assume that multihomed stub AS1 originates two IP prefixes, 194.100.80.0/20 (obtained from AS2's block of IP prefixes) and 200.2.160.0/20 (obtained from AS3). In order to load balance its inbound traffic and count with a fault-tolerant routing scheme, AS1 chooses to advertise its prefixes so that:

- Traffic targeting 194.100.80.0/20 should primarily enter the AS via AS2 and use AS3 as a backup path.
- Traffic targeting 200.2.160.0/20 should primarily enter the AS via AS3 and use AS2 as a backup path.

To accomplish these goals, AS1 selectively *prepends* its own AS number in its BGP advertisements with the aim of increasing the AS path length for the specific prefixes, and hence influence selection of the best route in upstream ASes.[4] Figure 4 shows the BGP advertisements sent by AS1. In this figure we assume that AS2 and AS3 are configured differently. AS3 propagates the two BGP advertisements received from AS1. AS2, on the other hand, sends an aggregate advertisement for 194.100.0.0/16. As this prefix includes 194.100.80.0/20, the advertisement received from AS1 is not propagated. This is typically so when a customer advertises a prefix that belongs to one of its ISP's block of prefixes. In such a case the ISP could aggregate the customer advertisement into a shorter prefix when advertising the prefix to other customers or peers.

As shown in Fig. 4, even though AS1 originates only two prefixes, AS4 receives four routes for three different prefixes: 194.100.80.0/20 (from AS3), 194.100.0.0/16 (from AS2), and 200.2.160.0/20 (from both AS3 and AS2). This increases the size of the BGP routing table of R41 since it receives more than one route for the same prefix.

Despite the prepending operation, all traffic from AS4 toward AS1 will be routed via AS3. This is because:

- The shortest AS-path for 200.2.160.0/20 is via AS3.
- The traffic for 194.100.80.0/20 will also be sent via AS3 because a BGP router always prefers the most specific (i.e., longest) prefix when forwarding packets.

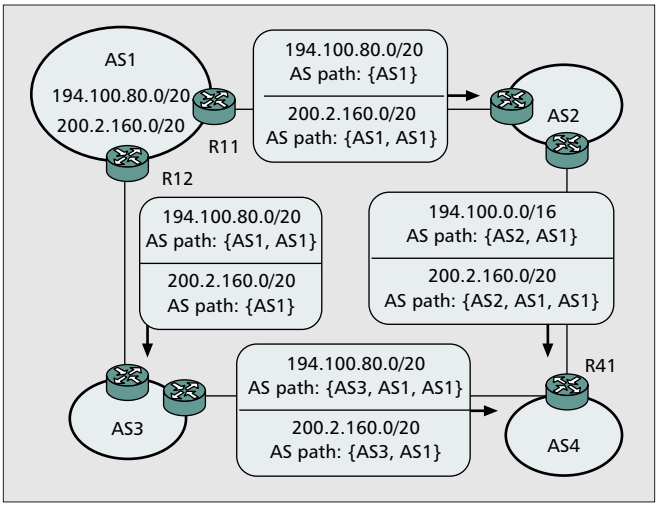
In such conditions AS2 will usually stop aggregating AS1's prefixes so that AS1 can start receiving traffic for 194.100.80.0/20 via AS2. This disaggregation causes AS2 to advertise two prefixes to AS4: the customer's prefix, 194.100.80.0/20, and the aggregate 194.100.0.0/16 with an additional increment in size of the BGP routing tables.

In the example, prefix 194.100.10.0/20 belongs to AS2, so this prefix cannot be aggregated by another ISP (AS3). As a general rule, a multihomed AS has several providers, and its prefixes cannot be aggregated by all of its providers. In fact, when a multihomed stub AS has allocated its own IP address space, the usual situation is that none of the providers is able to aggregate the prefixes of this AS.

In sum, load balancing and poor aggregation are the main reasons BGP tables are growing so fast.[5] The application of these practices makes the overall BGP routing tables nearly

[4] *It is worth mentioning that even though prepending is widely used in operational networks to influence how traffic enters an AS, for several reasons it does not always work. One of these reasons is addressed in the rest of the example. Other reasons are addressed later.*

[5] *While the maximum number of entries in a BGP router was around 1×10^5 in 2001, at present this number is larger than 1.63×10^5 [4, 17].*



■ Figure 4. *Growth of BGP routing tables: lack of aggregation and load balancing.*

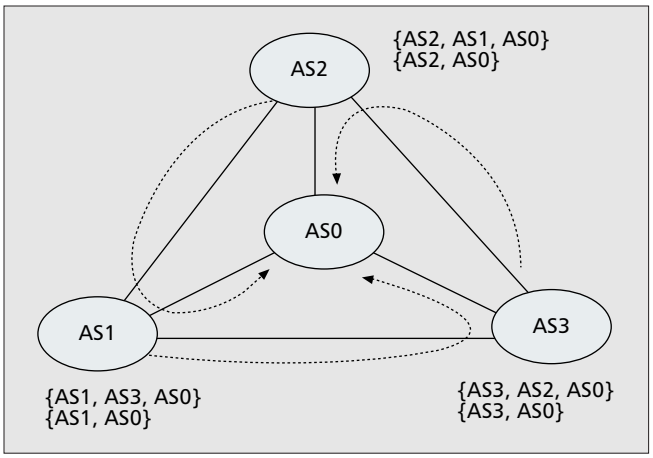


Figure 5. *The bad gadget example.*

50 percent larger than their optimal size (i.e., if aggregation was perfectly used) [4].

To cope with the problem and leverage aggregation, most ISPs filter the advertisements of long prefixes. Typically, several ISPs will not allow advertising to the global Internet prefixes longer than /22, or even longer than /20 [18]. This filtering process is an operationally palliative, but its downsides are considerable. A first consequence is that some routes are not distributed to the rest of the network. Furthermore, filtering does not tackle the root of the problem; it only works around it. The real challenge is to devise novel proposals that endow multihomed stub ASes with load balancing and fault tolerance mechanisms, while diminishing (or avoiding) the impact on the BGP routing tables. This is indeed a complex and open problem at present.

An alternative in the long term could be to define a better multihoming architecture. Several efforts are being carried out in order to deal with this issue in IPv6. Some appealing solutions are currently being developed in the Site Multihoming in IPv6 working group of the Internet Engineering Task Force (IETF) [19]. However, the problem remains largely unsolved for IPv4.

Expressiveness and Safety of Policies

Each AS in the Internet administrates its traffic in a completely autonomous way based on a set of policies that have only local significance to the AS. In other words, the way in which BGP routes are advertised through the global Internet and the way in which routing is finally performed are the result of the application of several independently configured policies. This lack of global coordination between the policies used in different domains is a major weakness of the current interdomain routing paradigm.

Several studies such as [20, 21] have demonstrated that without coordination, the interaction between independent policies may lead to global routing anomalies, such as inconsistent recovery from link failures or even route oscillations. Figure 5 depicts one of these routing anomalies. This particular configuration is known as "the bad gadget" [20], and illustrates how the policy-based nature of BGP may lead to configurations that are guaranteed to diverge (i.e., BGP does not converge). In this configuration the routing policies are such that each AS prefers the counterclockwise route to reach AS0 instead of the direct route. For example, AS2 prefers the route {AS2, AS1, AS0} over the route {AS2, AS0}. Given that AS1 and AS3 have analogous preferences, this configuration clearly causes divergence of the BGP protocol.

Earlier we assumed convergence of BGP as a fact, and based on it we exposed that the speed of this convergence is affected not only by the intrinsic properties of path vector routing protocols, but also by implementation decisions of BGP. The previous example shows that the convergence of BGP is indeed a much more complex and open problem, since managing routing based on independent policies means that convergence cannot be assumed as a fact.

The main reasons for the absence of cooperation between domains are:

- The characteristics of the BGP policy's expressiveness.
- The ASes are not willing to disclose details about their internal configuration and policies.

The expressiveness of policies is particularly tricky. On one hand, this expressiveness is rich enough to construct intricate local routing policies. Unfortunately, these policies may conflict with policies from other domains, leading to the global routing problems described before. On the other hand, this expressiveness is not enough to attach information to a route so that it could be straightforwardly shared and used throughout the network.

It should become clear that both the expressiveness of policies and the basis for autonomic management of policies able to guarantee robust convergence of the interdomain routing protocol are in very early stages of development. We need to thoroughly understand these two central aspects of distributed policies in order to balance the complex trade-off between allowing the ASes to disclose only the set of details they are willing to disclose, and guaranteeing robust convergence of BGP. Further discussion of these issues can be found in [22].

Robustness of BGP Sessions

The exchange of messages among two BGP routers is supported by a TCP connection, which supplies a reliable transport layer for communication between routers. Despite this reliability, some previous studies showed that the resilience of BGP sessions was formerly affected by congestion. In 1999 Labovitz *et al.* observed that KEEPALIVE messages were delayed during periods of peak network usage [23]. This led BGP sessions to fail when KEEPALIVE messages were delayed beyond the BGP hold timer [6]. Another previous study concerning the resilience of BGP sessions to congestion was presented in [24]. This study showed that increased queuing and delays had negative effects on the resilience of BGP. One of the main conclusions of [24] was the need to differentiate somehow routing protocol messages from normal data traffic. For this reason, an operational palliative several operators use at present is to prioritize BGP messages by setting their IP precedence to 7.

More recent work such as [25] shows that the conservative behavior of TCP retransmissions actually aggravates the instability of BGP sessions when network failures occur. The authors analyze the case of iBGP sessions, and propose a simple modification of TCP to increase the robustness of these sessions. Unfortunately, the community remains cautious about upgrading TCP.

Furthermore, the robustness of BGP sessions is an important issue at present for security reasons. This is because a BGP session will fail if the TCP connection fails due to an attack. This is addressed in the next section.

Security Issues

Among the issues presented in this article, the one probably capturing more attention at present is security. The reason for this is the concern of many operators that the vulnerabilities of BGP may cause large disruptions of service under possible attacks [26, 27]. Two main types of security issues exist with the current interdomain routing architecture and BGP protocol.

The first type of security issues are possible attacks on the transmission of BGP messages by legitimate routers. Given that two BGP peers maintain a BGP session over a TCP connection between themselves, the endpoints of this TCP connection (IP addresses and port numbers) can often easily be determined by a distant attacker. Furthermore, for a BGP router, a BGP session (and the corresponding interdomain link) remains up as long as BGP messages can be exchanged over the TCP connection. This implies that if the TCP connection fails for any rea-

son, the BGP session fails as well. An attacker could exploit this weakness by sending spoofed TCP RST segments to cause a TCP connection supporting a BGP session to fail.

To address this problem, some operational solutions are possible. A first solution is to authenticate the TCP segments carrying BGP messages by relying on MD5 [28]. This forces BGP peers to maintain a shared password. A second solution, mainly applicable to protect iBGP sessions, is to use filters on the border routers to ensure that spoofed packets using local addresses as sources cannot reach the network. Those solutions are also applicable to ensure that a distant attacker is not able to send spoofed BGP messages inside an existing BGP session. Clearly, these are operational palliatives but do not tackle the root of the problem (i.e., how to devise robust BGP sessions among BGP routers).

The second type of security issues are related to the lack of authentication in BGP. A BGP router can be configured to advertise any IP prefix, and most routers support powerful filters that can be used to completely change the content of received BGP messages.[6] Besides exploiting these vulnerabilities to conduct attacks, measurement studies have shown that misconfigurations of BGP routers are common events [29]. In any case, a BGP router should only be allowed to advertise IP prefixes that have been either allocated to its ASes, or learned from legitimate peer or customer ASes.

A first solution to improve the security of BGP has been proposed in S-BGP [30]. S-BGP relies on a public key infrastructure (PKI) to allow routers to include a route attestation with each advertisement. A route attestation is a cryptographic signature confirming that the S-BGP speaker is allowed to advertise this path. The main concerns about S-BGP compared to BGP are the cost (CPU, memory, and bandwidth) of producing, storing, and distributing attestations, and the need to bootstrap the PKI. Therefore, several alternate solutions have been proposed to lower the cost of securing BGP [31–35].

As of this writing, none of the solutions cited in the previous paragraph have actually been implemented in operational networks. From our perspective, this is mainly for two reasons. First, some are too heavyweight to be deployed, which makes them unappealing to ISPs. Second, standardization work is not well advanced. Only solutions that efficiently balance the trade-off between their effectiveness and cost to implement will have a chance at deployment.

Lack of Multipath Routing

A BGP router could receive multiple advertisements for the same route from multiple sources. For instance, in Fig. 4 router R41 receives two advertisements for the prefix 200.2.160.0/20, and hence will need to run its BGP decision process (Fig. 2) to select the best path to reach this destination. In its current release BGP selects only one path as the best path, and this is the path it places in the forwarding table. In addition, each BGP router only advertises to its peers the best route it knows to any given destination. Thus, R41 will typically install in its forwarding table the path via AS3 for the prefix 200.2.160.0/20 (choosing the shortest AS path), and this is the path it will advertise to its own peers.

This behavior introduces two important limitations. First, since the routing protocol only uses one best route, load balancing is not feasible even between paths presenting the same AS path length. For this reason some vendors have implemented and actually support multipath extensions in their BGP implementations. Despite this fact, only the best route is still advertised to other peers in both implementations. This is the second and most important limitation. Given that a BGP router only advertises the best route it knows many alternative paths that could have been potentially used by any source of traffic will be unknown. For example, a peer of R41 will receive an advertisement that the network 200.2.160.0/20 is reachable via {AS4, AS3, AS1}, but it will not know that the prefix is also reachable via {AS4, AS2, AS1, AS1}. This causes the BGP messages received in an AS to contain only a subset of all the available paths to a destination. This pruning behavior inherent in BGP introduces several limitations to the current interdomain routing paradigm, especially from the end-to-end quality of service (QoS) and traffic engineering (TE) viewpoints.[7]

At present, efforts are being carried out so that a BGP router will be able to advertise multiple routes for the same destination to its peers. One of the most recent proposals can be found in [37].

Despite the limitations described above, it is very unclear how to endow BGP with multipath routing capabilities without deeply impacting its scalability. If more routes are selected and advertised by BGP routers, more entries will exist in the BGP routing tables, increasing the problem exposed earlier.

Transit through an AS: iBGP Issues

BGP is an interdomain routing protocol and, as such, is mainly concerned with transmission of routes and packets between ASes. However, as an AS may contain thousands of routers, it is necessary to specify how the interdomain routes and packets can transit an AS. When a border router learns a new interdomain route, it needs to distribute this route to other routers inside its AS. This is done by sending the interdomain routes over iBGP sessions inside the AS. If the AS is small, a full mesh of iBGP sessions is established between the BGP routers. If the AS is larger, route reflectors [38] or confederations [39] are used to replace this unscalable iBGP full mesh.

When a border router of a transit AS receives a packet whose destination is not local, it will consult its BGP routing table to determine the BGP next hop (i.e., the egress border router) inside its own AS. However, there can be several intermediate routers between the ingress router and egress router. To ensure that an interdomain packet will reach the BGP next hop selected by the ingress border router, the transit AS must ensure that all intermediate routers also select this next hop.

This problem was discussed early in the development of BGP [40], and two techniques have emerged. The first solution, proposed in 1990, is to use encapsulation: the ingress border router encapsulates the interdomain packets inside a tunnel toward the egress border router chosen by its BGP decision process. At that time, encapsulation suffered from a major performance drawback given the difficulty of performing encapsulation on the available routers. Today, high-end routers are capable of performing encapsulation or decapsulation at line rate when using multiprotocol label switching (MPLS) or IP-based tunnels. The main advantage of using encapsulation is that the BGP forwarding table is consulted only once (by the ingress border router) per interdomain packet inside each AS.

Unfortunately, this is not often common practice in pure IP-based transit networks. This type of network typically uses another technique, called Pervasive BGP, in which BGP is run on all (border and non-border) routers inside the transit AS. As all intermediate routers must consult their BGP forwarding table for each interdomain packet, there is a risk of deflection or worse routing loops when the forwarding tables are not perfectly synchronized, such as during BGP convergence [41], or when route reflectors or confederations are used [42].

The main issue at present is that route reflectors and/or con-

[6] *These filters are precisely those that support the construction of the intricate and autonomic routing policies described earlier.*

[7] *It is important to highlight that the shortest AS path does not necessarily supply the best end-to-end traffic performance [36].*

federations have become absolutely necessary given the tremendous scalability they have supplied to large transit ASes. However, anomalies such as those described before can occur, especially in the event of a link or router failure with Pervasive BGP. The question that arises, then, is how can it be guaranteed that iBGP configurations remain highly scalable and anomaly-free at the same time?

Limited Traffic Engineering Capabilities

The current interdomain routing model offers scarce TE capabilities for a number of reasons. First, BGP was designed as a protocol to distribute reachability information. Second, as exposed earlier, the inability of BGP to advertise multiple routes for the same destination limits the number and quality of the alternative paths that could be used to reroute packets around a failure. In addition, the limitation of BGP in terms of multipath routing restricts the possibilities of balancing traffic across domains to certain setups and vendor-specific implementations.

Second, as shown previously, the autonomic management of policies and limitations in the expressiveness of these policies impose strong restrictions on how ASes are able to control and manage the flow of their interdomain traffic. For instance, even though BGP allows an AS to flexibly manage its outbound traffic, it exhibits a scarce degree of control in managing and balancing how traffic enters an AS across multiple paths. In other words, accurately controlling inbound traffic with BGP is a very complex task, and it is still unclear how it can be optimally accomplished. The reason for this lies in the lack of global coordination between the policies used in different domains. Thus, each AS in any given path may apply its own local policies and route its outbound traffic as desired, overriding any routing advertisement and requirement from downstream ASes.

To cope with the problem of controlling the inbound traffic of an AS, several operational palliatives are possible. Some techniques rely on the utilization of AS path prepending [43], some on BGP communities [44], and others on network address translation (NAT) [45].

However, all have several limitations. A corollary of what we exposed earlier is that AS path prepending might not always work. BGP communities provide more control than AS path prepending, but are neither perfect nor always supported. Finally, controlling traffic by using NAT is simply infeasible for medium and large ASes. As a result, the common practice in the global Internet is that inbound traffic is manually configured and tuned on a trial and error basis, and hence remains an open problem in terms of interdomain TE. An incrementally deployable solution called Virtual Peering was proposed recently in [46]. The approach is that a pair of ASes cooperate and set up a unidirectional IP tunnel between their border routers to manage the traffic between them.

Another important topic is that the objectives of interdomain TE drastically vary depending on the type of AS. The classification of three types of ASes made earlier is pertinent since the requirements of and problems faced by each type are quite different. For instance, the current trend for multihomed stub ASes is to deploy selfish TE techniques able to operate on short timescales [47]. These techniques typically try to exploit the multiconnectivity of the AS, with the aim of improving the performance and reduce the monetary costs. The main problem behind this is that if more and more ASes keep on using such selfish techniques, it could place significant stress on the scalability and reliability of the entire interdomain routing system.

On the other hand, TE mechanisms developed for transit multihomed ASes such as large ISPs are designed to operate on large timescales (typically on the order of weeks or months). These ASes usually use a routing practice known as *hot potato routing* [48]. In this practice a BGP router within an AS will be able to reach a certain destination by multiple exit points of the AS, so the router needs to run the BGP decision process in Fig. 2. Typically, a subset of those multiple exit points will supply the same AS path length toward the destination, so the decision BGP process usually reaches step 4 or 5 in Fig. 2. These two steps basically mean that the routing criterion is to try to get rid of the packets from the AS as fast as possible. This is typically determined by the intradomain routing protocol running on the AS (step 5, Fig. 2).

One of the main problems these transit ASes face in terms of TE is that the attempt to improve their hot potato routing has a profound impact on interdomain traffic (and reciprocally) [48]. This causes traffic patterns to change across the boundaries of the AS, affecting other ASes. These ASes may now run their own TE policies, which in turn may negatively impact back on the original AS. This brings back the problem of routing instabilities due to poor or no coordination between the policies used in different domains.

Recent studies reveal that the topological characteristics of interdomain traffic show large variations over time. Indeed, large fractions of AS paths are only present in the BGP routing tables for a few minutes. This behavior increases the number of BGP messages traversing the network. Despite this variability, three important results demonstrate that TE at the interdomain level is in fact feasible [49, 50]:

- Measurement studies show that in one AS, a small fraction of the destination prefixes are responsible for a large fraction of the interdomain traffic.
- Regardless of the large number of BGP update messages, popular prefixes represent stable entries in the BGP routing tables for weeks or even months.
- The majority of update events correspond to prefixes that do not receive much traffic.

These results have important TE repercussions since researchers can focus on devising novel TE mechanisms network operators could apply to the majority of their traffic, whose routes are typically stable.

Lack of QoS Support

Applications such as voice over IP or virtual private net-works have strong requirements in terms of QoS. To fulfill those requirements, many ISPs have deployed mechanisms to provide differentiated services in their networks. The customers of those ISPs are now requiring similar levels of QoS across interdomain boundaries [51]. BGP has no built-in QoS capabilities since it was designed only as a protocol to distribute reachability information. This inability of BGP to supply and distribute QoS information was recognized as a missing piece by the IETF in mid-1998 [52].

This issue has received attention during the past few years. Due to space limitations we cannot review the entire literature, but an appealing proposal can be found in [53].

Despite these efforts and over a decade of work, the astonishing outcome is that none the proposals has turned out to be sufficiently appealing to become deployed in practice. This is because ISPs have preferred to overprovision their networks rather than deliver and manage QoS. The debate about overprovision vs. QoS is still open. Leaving aside issues like the monetary cost to deploy and maintain QoS, or the development of possible businesses leading to tangible sources of profit for ISPs, from our perspective the issue remains unsolved mainly because *all* the issues presented so far are actually strong limitations on QoS at the interdomain level. The interdomain routing paradigm itself is in fact a major cause for this lack of QoS support.

An alternative could be to change the paradigm, but at present only incrementally deployable approaches seem realistic and hence have a chance to be adopted. We believe that efficient mechanisms allowing network operators to improve their end-to-end performance while demanding nearly no effort to support and maintain are still missing.

Conclusions and Lessons Learned

Interdomain routing is still a challenging research area. The main challenge resides in the intricate relationships and coupled trade-offs between the open issues presented in this article. Rather than tackling these issues one by one and in an isolated manner, we need to thoroughly understand their relationships and dependencies if we expect to make any real progress in the area. This is in part why several of the valuable proposals made so far have never reached the deployment stage. Unfortunately, this is not the only reason. It is also due to the fact that ISPs are reluctant to introduce changes and test them if there is no clear source of revenue. Clearly, this makes the problem of making real progress even harder.

We emphasize that while some of the issues exposed in this article are rooted in the intrinsic limitations of BGP and the current interdomain routing architecture, others derive from the intricate interactions and dependencies between domains. As we have described, routing management is performed in an autonomous manner by each domain, and the fact of greatest concern is that this is done based on potentially conflicting policies. Thus, more social or collaborative policies may need to be developed in the mid-term.

An alternative in the long term could be to gradually replace BGP or even the whole interdomain routing paradigm. However, this might be infeasible for IPv4-based networks given the large installed base. From our perspective, future MPLS-based and optical networks offer a neat path to address from scratch several of the issues exposed in this article. We should take advantage of the lessons learned and avoid incurring the same mistakes of the past.

References

[1] R. Atkinson and S. Floyd, Eds., "IAB Concerns and Recommendations Regarding Internet Research and Evolution," RFC 3869, Aug. 2004.
[2] G. Huston, "Analyzing the Internet's BGP Routing Table," *IP J.*, vol. 4, no. 1, 2001.
[3] https://cba.ccaba.upc.es/research/bgp/
[4] CIDR report, July 2005: http://www.cidr-report.org/
[5] L. Subramanian *et al.*, "Characterizing the Internet Hierarchy from Multiple Vantage Points," *INFOCOM 2002*.
[6] Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)," RFC 1771, Mar. 1995.
[7] S. Halabi and D. McPherson, *Internet Routing Architectures*, 2nd ed., Cisco Press, 2001.
[8] S. Agarwal, C. Chuah, and R. Katz, "OPCA: Robust Interdomain Policy Routing and Traffic Control," *IEEE OPENARCH*, Apr. 2003.
[9] C. Labovitz *et al.*, "Delayed Internet Routing Convergence," *Proc. ACM SIGCOMM*, 2000.
[10] T. Griffin and B. Presmore, "An Experimental Analysis of BGP Convergence Time," *Proc. IEEE ICNP*, Nov. 2001.
[11] C. Villamizar, R. Chandra, and R. Govindan, "BGP Route Flap Damping," RFC 2439, Nov. 1998.
[12] Z. M. Mao *et al.*, "Route Flap Damping Exacerbates Internet Routing Convergence," *Proc.ACM SIGCOMM*, 2002.
[13] A. Bremler-Barr, Y. Afek, and S. Schwarz, "Improved BGP Convergence via Ghost Flushing," *Proc. IEEE INFOCOM*, 2003.
[14] D. Pei *et al.*, "BGP-RCN: Improving BGP Convergence through Root Cause Notification," *Comp. Net.*, vol. 48, no. 2, 2005, pp. 175-94.
[15] J. Chandrashekar *et al.*, "Limiting Path Exploration in BGP," *Proc. INFCOM*, Miami, FL, 2005.
[16] A. Feldmann *et al.*, "Locating Internet Routing Instabilities," *Proc. ACM SIGCOMM*, Portland, OR, Sep. 2004.
[17] T. Bu, L. Gao, and D. Towsley, "On Routing Table Growth," *Proc. IEEE Global Internet Symp.*, 2002.
[18] S. Bellovin *et al.*, "Slowing Routing Table Growth by Filtering Based on Address Allocation Policies," unpublished manuscript, June 2001.
[19] IETF Site Multihoming in IPv6 Working Group, http://www.ietf.org/html.charters/multi6-charter.html.
[20] T. G. Griffin and G. T. Wilfong, "An Analysis of BGP Convergence Properties," *Proc. SIGCOMM*, Cambridge, MA, Aug. 1999, pp. 277-88.
[21] T. G. Griffin, F. B. Shepherd, and G. Wilfong, "The Stable Paths Problem and Interdomain Routing," *IEEE/ACM Trans. Net.*, vol. 10, no. 2, Apr. 2002, pp. 232–43.
[22] A. D. Jaggard and V. Ramachandran, "Towards the Design of Robust Interdomain Routing Protocols," *IEEE Network*, Special Issue on Interdomain Routing, Nov./Dec. 2005.
[23] C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental Study of Internet Stability and Backbone Failures," *Proc. FTCS-29, 29th Int'l. Symp. Fault-Tolerant Comp.*, Madison, WI, June 1999, pp. 278-85.
[24] A. Shaikh *et al.*, "Routing Stability in Congested Networks: Experimentation and Analysis," *Proc. ACM SIGCOMM*, Stockholm, Sweden, Aug. 2000.
[25] L. Xiao and K. Nahrstedt, "Reliability Models and Evaluation of Internal BGP Networks," *Proc. IEEE INFOCOM 2004*, Hong Kong, China, Mar. 2004.
[26] O. Nordstrom and C. Dovrolis, "Beware of BGP attacks," *ACM SIGCOMM Comp. Commun.* 2004.
[27] S. Murphy, "BGP Security Vulnerabilities Analysis," Internet draft, draft-ietf-idr-bgp-vuln-01.txt, Oct. 2004, work in progress.
[28] A. Heffernan, "Protection of BGP Sessions via the TCP MD5 Signature Option," RFC 2385, Aug. 1998.
[29] R. Mahajan, D. Wetherall and T. Anderson," Understanding BGP Misconfigurations," *ACM SIGCOMM 2002*, Aug. 2002.
[30] S. Kent, C. Lynn, and K. Seo, "Secure Border Gateway Protocol (S-BGP)," *IEEE JSAC*, Apr. 2000.
[31] G. Goodell *et al.*, "Working Around BGP: An Incremental Approach to Improving Security and Accuracy of Interdomain Routing," *NDSS*, Feb. 2003.
[32] W. Aiello, J. Ioannidis, and P. McDaniel, "Origin Authentication in Interdomain Routing," *Proc. 10th ACM Conf. Comp. and Commun. Sec.*, 2003.
[33] R. White, "Securing BGP through Secure Origin BGP," *IP J.*, Sept. 2003.
[34] Y.-C. Hu, A. Perrig, and M. Sirbu, "SPV: Secure Path Vector Routing for Securing BGP," *ACM SIGCOMM 2004*, Sept. 2004.
[35] M. Zhao, S. Smith, and D. Nicol, "The Performance Impact of BGP Security," *IEEE Network*, special issue on Interdomain Routing, Nov./Dec. 2005.
[36] B. Huffaker *et al.*, "Distance Metrics in the Internet," *IEEE Int'l. Telecommun. Symp.*, 2002.
[37] D. Walton, A. Retana, and E. Chen, "Advertisement of Multiple Paths in BGP," Internet draft, draft-walton-bgp-add-paths-04.txt, Aug. 2005, work in progress.
[38] T. Bates, R. Chandra, and E. Chen, "Route Reflection — An Alternative to Full Mesh iBGP," RFC 2796, IETF, Apr. 2000.
[39] P. Traina, "Autonomous System Confederations for BGP," RFC 1965, June 1996.
[40] Y. Rekhter, "Constructing Intra-AS Path Segments for an Inter-AS Path," *ACM SIGCOMM Comp. Commun.*, 1991.
[41] A. Sridharan, S. B. Moon, and C. Diot, "On the Correlation between Route Dynamics and Routing Loops," *Proc. IMC*, Miami, FL, Oct. 2003.
[42] T. G. Griffin and G. Wilfong, "On the Correctness of iBGP Configuration," *Proc. ACM SIGCOMM 2002*, Aug. 2002.
[43] R. K. C. Chang and M. Lo, "Inbound Traffic Engineering for Multihomed ASes Using AS Path Prepending," *IEEE Network*, Mar. 2005.
[44] B. Quoitin *et al.*, "Interdomain Traffic Engineering with Redistribution Communities," *Comp. Commun.*, vol. 27, no. 4, 2004.
[45] F. Guo *et al.*, "Experiences in Building a Multihoming Load Balancing System," *INFOCOM 2004*, 2004.
[46] B. Quoitin and O. Bonaventure, "A Cooperative Approach to Interdomain Traffic Engineering," *1st Conf. Next Gen. Internet Networks TE (NGI 2005)*, Rome, Italy, 2005.
[47] A. Akella, S Seshan, and A. Shaikh, "Multihoming Performance Benefits: An Experimental Evaluation of Practical Enterprise Strategies," *USENIX Annual Tech. Conf. 2004*, Boston, MA.
[48] S. Agarwal, A. Nucci, and S. Bhattacharyya, "Controlling Hot Potatoes in Intradomain Traffic Engineering," SPRINT ATL res. rep. RR04-ATL-070677, July 2004.
[49] S. Uhlig *et al.*, "Implications of the Topological Properties of Internet Traffic on Traffic Engineering," *Proc. 19th ACM Symp. Applied Comp., Special Track on Comp. Networks*, Nicosia, Cyprus, Mar. 2004.
[50] J. Rexford *et al.*, "BGP routing Stability of Popular Destinations," *Proc. Internet Measurement Wksp.*, Nov. 2002.
[51] M. Morrow *et al.*, "Challenges in Enabling Interprovider Service Quality in the Internet," *IEEE Commun. Mag.*, vol. 43, no. 6, June 2005.
[52] E. Crawley *et al.*, "A Framework for QoS-based Routing in the Internet," RFC 2386, Aug. 1998.
[53] L. Xiao *et al.*, "QoS extensions to BGP," ICNP2002, Nov. 2002.

Biographies

MARCELO YANNUZZI (yannuzzi@ac.upc.edu) received a degree in electrical engineering from the University of the Republic, Uruguay, in 2001, and a D.E.A. from the Technical University of Catalonia (UPC), Spain, in 2005. He is currently a Ph.D. student at the Department of Computer Architecture, UPC. His current research interests are in the area of interdomain routing and traffic engineering in both IP and optical networks.

XAVIER MASIP-BRUIN (xmasip@ac.upc.edu) received M.S. and Ph.D. degrees from UPC, both in telecommunications engineering, in 1997 and 2003, respectively. He is currently an associate professor of computer science at UPC. His current research interests lie in broadband communications, QoS management and provision, and traffic engineering. His publications include around 40 papers in national and international refereed journals and conferences. Since 2000 he has participated in many research projects: IST projects E-NEXT, NOBEL, and EuQoS; and Spanish research projects SABA, SABA2, SAM, and TRIPODE.

OLIVIER BONAVENTURE (http://www.info.ucl.ac.be/people/OBO) leads the network research group at Université catholique de Louvain (UCL), Belgium. His current research interests include intra- and interdomain routing, traffic engineering, and network security.