The "Global-ISP" Paradigm

Reuven Cohen Amnon Shochot
Department of Computer Science
Technion, Israel

July 26, 2006

Abstract

We present a new paradigm, called "Global ISP" (G-ISP). Its goal is to solve, or at least alleviate, problems of inter-domain routing, such as slow convergence, and lack of QoS and multicast support. One of the most important properties of the proposed paradigm is that it can be gradually deployed on the Internet. A G-ISP can be viewed as an additional ISP that provides transit services to its customers over an *overlay network*. Because a G-ISP differs from a "regular" ISP, some extension to the standard BGP protocol is required. This extension and its effects on the BGP protocol are described in this paper. Algorithms for building a G-ISP overlay network and their applications are also presented.

1 Introduction

The Internet consists of thousands of Autonomous Systems (AS). Each AS is composed of a collection of networks and is administrated by a single authority. There are two types of Internet routing protocols: intra-domain routing protocols and inter-domain routing protocols. An intra-domain routing protocol, like OSPF [21] or RIP [19], is responsible for routing a packet inside an AS. The inter-domain routing protocol is responsible for determining the sequence of ASs through which the packet traverses when the source and destination are located in different ASs.

The standard inter-domain routing protocol in the Internet today is the Border Gateway Protocol (BGP) 4 [27]. Several studies conducted in recent years reveal that BGP suffers from several inherent problems, the most important of which are as follows:

1. Slow convergence and instability: In [16], it is shown that the average delay in Internet inter-domain path failovers is 3 minutes, and that some of the failovers cause oscillations in the

routing tables lasting up to 15 minutes. In [34] and [11] it is shown that conflicting policies result in the BGP protocol not being guaranteed to converge. Furthermore, in [12] it is shown that the problem of exploring potential disputes in the policies of several ASs is NP-hard.

2. Lack of Quality of Service (QoS) support: While intra-domain support of QoS is considered easy today, inter-domain QoS requires bilateral agreements which are very difficult to achieve and may require new signaling protocols. Moreover, unlike link-state intra-domain routing protocols, which can be extended to support QoS [2], BGP has no QoS support, and it is unlikely that it could be extended to have it in the future.

3. Lack of multicast support: While intra-AS multicast is well understood, and protocols like M-OSPF [20], DVMRP [36] and PIM (e.g. PIM-SM [6]) have been employed for many years, the combination of inter-AS routing and multicast is considered difficult. This is because BGP provides only limited information to each router. Building an efficient – or even an inefficient – multicast tree using such information, while still adhering to the policy rules, is exceptionally hard. The multicast backbone (MBone) [5] was built as an overlay multicast network to allow inter-AS multicast routing without relying on BGP multicast. However, there is no standard protocol today for inter-domain multicast[1].

In this paper we present a new paradigm, called "Global ISP" (G-ISP), whose goal is to solve, or at least to alleviate, these problems. A G-ISP can be viewed as an additional ISP that provides transit services to its customers over an *overlay network*. The customers of a G-ISP maintain a BGP connection with the G-ISP in order to learn the routes it advertises. These customers then use their policy rules in order to decide which route is the best for each network prefix. However, a G-ISP customer will usually have no direct connectivity with the G-ISP. This is the main difference between a G-ISP and a "regular ISP" (henceforth referred to simply as an ISP), as an ISP must have a direct link (i.e., layer-2 connectivity) to each of its customers. This difference requires some

[1]The IETF BGMP working group [33] is working towards such a standard.

extension to the standard BGP protocol. Such an extension and its effects on BGP are described in this paper.

The idea behind the new paradigm is to use an ISP for reaching nearby networks and a G-ISP for reaching remote ones. In particular, the ISP will be used to reach the AS of the G-ISP. This idea is borrowed from the telephony business model, where in many countries the customer has a local operator and a global one. As in the telephony world, the customer will use the local operator in order to connect to the global ("long distance") operator. However, while the purpose of a global telephony operator is to provide connectivity to destinations not reachable via the local operator, the purpose of a G-ISP is to provide better connectivity to networks that can also be reached through the regular ISP.

As stated above, the G-ISP and its customers will usually have no direct connectivity. Therefore, the BGP relationship between a G-ISP and its customers is conducted over a virtual link established over multiple intermediate physical links and routers. Moreover, when a customer wants to use a path offered by the G-ISP, it has to forward its packet to the G-ISP. In the absence of a direct link between the customer and the G-ISP, the IP tunneling (IP-in-IP) [32] concept should be used.

As mentioned, the main motivation behind the G-ISP paradigm is to address the BGP problems discussed above. The G-ISP paradigm solves or alleviates these problems by

- Providing customers with short AS-Paths, which have shorter convergence times. According to [17], the longer the AS-Path is, the bigger the risk is for a longer convergence time. Although there might be some exceptions to this rule, it usually reflects the actual situation correctly.

- Migrating customers toward end-to-end QoS support. The G-ISP provides its customers with two forms of QoS: (a) QoS within the G-ISP using well-known schemes for intra-AS QoS [22]; (b) QoS to packets traveling from the customer toward the G-ISP, even if this requires bilateral agreements with the ISP's neighboring AS. This is possible because the

number of ASs sitting between the customer AS and the G-ISP is small. If the destination address of the packet also belongs to a customer of the G-ISP, full end-to-end QoS can be provided.

- Providing inter-AS multicast support: A G-ISP can serve as a core for multicast groups of its customers. The distribution of the multicast packets from the G-ISP to its customers is performed using unicast. The high connectivity of the G-ISP guarantees the efficiency of this method.

The rest of this paper is organized as follows. In Section 2 we discuss related work on BGP. In Section 3 we present the G-ISP paradigm. The export rules for this paradigm are described in Section 4. In Section 5 we present algorithms for building an overlay network whose core is the G-ISP AS. In Section 6 we discuss the feasibility of the G-ISP paradigm. In Section 7 we show how a G-ISP can support inter-AS multicast and end-to-end inter-AS QoS. Finally, Section 8 concludes the paper.

2 Related Work

Much research has recently been conducted on BGP. Some works have tried to evaluate BGP performance and study its properties, while others have pointed out some of its inherent problems. In [14], Huston examined the various longer term trends of the Internet's BGP tables, and how they impact the scaling properties of the inter-domain routing space. In [16] and [17] Labovitz *et al.* examined the time it takes an AS to converge following a path failure, failover and repair. During their two-year study, the average delay in the Internet inter-domain path failover was 3 minutes and some percentage of failovers took up to 15 minutes to converge. They showed that the average time it takes for an AS to learn that network X of another AS is reachable after being down depends linearly on the length of the shortest AS-Path between the two ASs. In addition, they found that packet loss grows by a factor of 30 and latency by a factor of 4 during path restoration. In [10], Griffin *et al.* examined the delay in convergence time caused by the Minimum Route

Advertisement Interval (MRAI) timer. They found that the optimal value for this timer is topology dependent. In [38], Zhang *et al.* examined the packet delivery performance in a network running the BGP routing protocol when a destination might be disconnected from time to time. Their work proposed using packet delivery measures as a parameter for improving routing protocol performance.

Checking consistency assertions to identify infeasible routes was proposed in [24]. This paper shows that identifying and ignoring infeasible routes substantially reduces both BGP convergence time and the total number of intermediate route changes. Bremler-Barr [3] proposed that convergence time be reduced using improved ghost flushing. In [4], Cobb *et al.* present a solution where a BGP node has the freedom to choose any routing policy. To avoid instabilities, the likelihood for divergence is measured using an efficient cost metric, which is exchanged between nodes. Pei *et al.* propose in [23] another mechanism, called "BGP with root cause notification." It provides an upper bound of $O(d)$ on the routing convergence delay for BGP, where d is the network diameter as measured by the number of AS hops. Their analysis and simulation show that the proposed mechanism can substantially reduce both BGP convergence time and the total number of intermediate route changes.

In [34], Varadhan *et al.* showed that there are collections of BGP routing policies that can cause BGP to diverge. Griffin *et al.* developed this observation in [11] and defined an abstract model of BGP, which they used to demonstrate the conditions on routing policies that guarantee BGP convergence. Later, Griffin *et al.* showed [12] that, given the set of policies of several ASs, identifying whether these policies conflict and cause divergence is NP-hard. In [13], Griffin *et al.* suggested adding a dynamically computed attribute, called *route history*, to the BGP advertisements as a way to dynamically recognize route oscillations caused by policy conflicts. Gao and Rexford proposed [7] some guidelines that, if adopted by all ASs while their routing policies are being set, guarantee BGP convergence. Gao, Griffin and Rexford extended these guidelines [8] to support backup routing.

Several companies have tried to improve Internet performance by offering their customers route control solutions. These solutions monitor performance from the customer to other Internet sites, and detect broken routes, brownouts, and other problems. The collected data is used to optimize the routing of outgoing traffic in order to decrease packet loss rate and latency.

Another relevant work is Resilient Overlay Networks (RON) [1]. This project offers an architecture that allows distributed Internet applications to detect and recover from path outages and periods of degraded performance within several seconds. This architecture is implemented as an overlay network on top of the existing Internet substrate.

Some attempts have been made to apply inter-domain multicast routing to the Internet. Most of these attempts are implemented as an overlay network on top of the existing Internet infrastructure. The most famous attempt is the Multicast Backbone (MBone), described in [5]. Another proposed solution is Overcast [15], designated for content distribution, In [30, 29, 31], Shi investigated a series of issues related to the design of an efficient overlay multicast network. Finally, an inter-domain multicast protocol, called the Border Gateway Multicast Protocol (BGMP) [33], is being developed by the IETF. This new protocol is aimed at supporting "source-specific multicast" as well as "any-source multicast."

Inter-AS QoS is another challenging problem that has been addressed by some recent studies. In [37], Xiao *et al.* define a new metric, the Available Bandwidth Index (ABI), and use it to perform bandwidth advertising and routing. In [18], Levis *et al.* describe how a QoS-enhanced BGP can be used with a notion of Meta-QoS-Classes for building a set of parallel Internet planes with different QoS capabilities. The authors demonstrate that QoS-enhanced services are made possible using their concepts.

3 The BGP Algorithm for Supporting the G-ISP Paradigm

In this section we present the main concepts of the G-ISP paradigm. A G-ISP can be viewed as an additional ISP that provides transit services to its customers over an overlay network. G-ISP

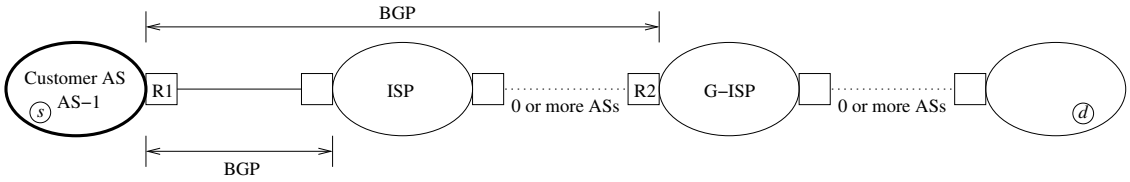


Figure 1: The G-ISP Scheme

customers maintain a BGP connection with the G-ISP in order to learn the routes it advertises. These customers then use their policy rules to decide which route is the best for routing to each network prefix. However, a G-ISP customer will usually have no direct connectivity with the G-ISP. For example, in Figure 1, AS-1 is a G-ISP customer even though it has no direct link with it. This is probably the most important difference between a G-ISP and an ISP, because an ISP must have a direct link to each of its customers. This important difference guarantees the scalability of the G-ISP paradigm and the ability to deploy G-ISP services gradually. However, it requires extension of the BGP protocol.

Consider a customer AS, served by an ISP and a G-ISP. The following rules summarize the main differences between an ISP and a G-ISP:

- While the BGP connection with the ISP is established over a single link, traversing no intermediate routers, the BGP connection with the G-ISP is established over an IP route that goes through the ISP. In Figure 1, AS-1 is a customer of the G-ISP but has no direct link to it, and its BGP connection with the G-ISP is established over a path that goes through the ISP and, possibly, other intermediate ASs.

- Routing through a G-ISP requires that the customer use IP tunneling. When a customer wants to send a packet over a route it learned from its G-ISP, it must use IP tunneling to the G-ISP's AS. Otherwise, when the packet is received by the ISP it will be forwarded over the ISP's route rather than over the G-ISP's route. Consider the source and destination s and d in Figure 1. A packet originated by s has a single IP header with a source IP address IP(s) and a destination IP address IP(d). When the packet is received by the customer's border router

R1, R1 attaches another IP header to this packet, in front of the original header. The latest header indicates that the source is IP(R1) and the destination is IP(R2). When the packet is received by R2, the new header is removed, and the packet continues through the G-ISP toward the destination. This ensures that the packet will be routed over the G-ISP's route and not over the ISP's route. Note that in the reverse direction tunneling is not used unless *d* has its own G-ISP.

- There are special export rules for both the customer and the G-ISP. This issue is discussed in Section 4.

In fact, there are two types of services that can be delivered by the G-ISP to a customer: (a) unidirectional services, and (b) bi-directional services. In the former case the G-ISP serves only packets originated at the customer. This service is good for a large content distributor, who is interested in ensuring good response time, or even some level of QoS, to the content delivered to its customers. In the latter case, the G-ISP serves both packets originated by its customer and packets destined to its customer. Packets originated by the customer are served using IP tunneling as described above. Packets destined to the customer networks are served by advertising attractive routes to these destinations, thus "encouraging" these packets to route to the customer networks through the customer's G-ISP. Note, however, that while our framework ensures that the packets originated by the customer hosts will be routed through the G-ISP, there is no way to force packets in the reverse direction to do the same.

4 Export Rules for the G-ISP Paradigm

In this section we introduce the export rules of all the participants in the extended BGP protocol. In Section 4.1 we discuss the routes advertised by the customer to its G-ISP and the implication of these advertisements on the BGP algorithm. In Section 4.2 we discuss the routes that should be advertised by a G-ISP to its customers. Finally, in Section 4.3 we discuss the routes that are advertised by the G-ISP to its ISP peers.

Note that our scheme does not change the export rules between a customer and an ISP. Namely, a customer should advertise to its ISP its routes to all its networks and to all its customers' networks. The ISP should send to its customers all the routes that it uses. These routes include the routes the ISP learned from its customers, the routes it learned from its providers and peers, and the routes to its own networks.

4.1 Export rules from the customer to its G-ISP

A G-ISP provides transit services only to packets originated by its customers' networks and optionally to packets destined for its customers' networks. A customer subscribed only to the unidirectional G-ISP services does not have to export any route to the G-ISP. By comparing the source IP address of the packet with the IP address of the border routers of its customers, the G-ISP can infer that a packet received by its border router was indeed originated at a customer and thus should be serviced. Since the G-ISP is not required to service packets destined to this customer, it is not required to advertise to its peers its routes to this customer, nor to receive advertisements from this customer about the customer's networks. Therefore, the rules discussed in the rest of this subsection are only relevant for customers subscribed to the bi-directional services of the G-ISP.

Like an ISP, a G-ISP has to learn the IP prefixes of its customers' networks. Therefore, a G-ISP customer should advertise its networks and its customer's networks (e.g., if a customer is an ISP) to the G-ISP. However, when the G-ISP receives an AS-Path for some customer network prefix from a customer *C*, the received AS-Path is not the full AS-Path from *C* to the G-ISP. Rather, since *C* and the G-ISP are not directly connected, the received AS-Path does not contain the numbers of the ASs in the path between *C* and the G-ISP. Thus, the G-ISP has to append to the received AS-Path the intermediate AS numbers, as learned from its peer ASs.

For an example, consider Figure 2. The G-ISP customers are AS-1 and AS-7. An arrow from A to B indicates that A is a provider of B. A dotted line represents a peering agreement between two ASs. In the figure, AS-2 is the ISP of AS-1 and AS-6 is the ISP of AS-7. Note that AS-1 and AS-7 are not directly connected to the G-ISP. They therefore need to establish a BGP connection with

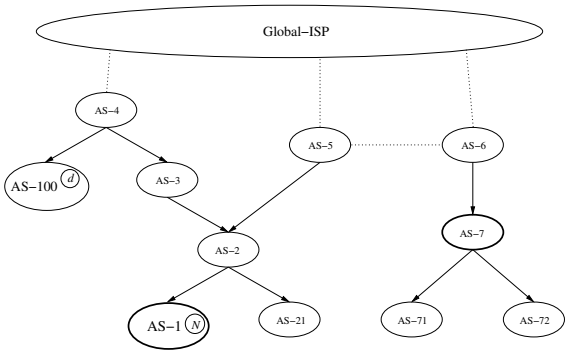


Figure 2: G-ISP as part of the AS Hierarchy Graph

G-ISP using AS-2 and AS-6 respectively. AS-7 sends to G-ISP advertisements for its networks with AS-Path [AS-7] and advertisements for its customers networks with AS-Paths [AS-7 AS-71] and [AS-7 AS-72]. The G-ISP, in its turn, inserts into the received AS-Paths the intermediate AS numbers, i.e., AS-6. Therefore, the routes that are stored in the G-ISP's local database are [AS-6 AS-7] for the networks belonging to AS-7 and [AS-6 AS-7 AS-71] and [AS-6 AS-7 AS-72] for networks belonging to AS-71 and AS-72 respectively.

One may suggest that instead of requiring the customer to advertise its networks to its G-ISP and requiring the G-ISP to modify the received AS-Paths, a G-ISP can learn the relevant information from its peers by examining the AS-Path in the routes it receives. If a customer AS number appears in this AS-Path, then the network to which this AS-Path leads belongs to the customer AS or to a customer's customer AS. (Otherwise, the customer AS would not advertise a route to this network.) However, there are several problems with this approach. The first is that not all the AS-Paths that contain the customer AS number necessarily belong to the customer or to one of its customers. This is the case, for example, when the customer AS provides backup services to some non-customer AS. The second problem is that a G-ISP customer may be interested in bi-directional services for only some of its networks or customers and not for all of them. It is not possible, however, to make this distinction if the customer has no BGP relationship with the G-ISP. The third problem occurs when some intermediate AS in the path from the customer to the G-ISP performs aggregation of network address prefixes. For example, suppose that in Figure 2

AS-1 is a G-ISP customer but AS-21 is not. Suppose also that AS-2 is the single provider for both ASs. Now, assume that AS-1 advertises to AS-2 a network prefix of 192.12.128.0/17 and AS-21 advertises to AS-2 a network prefix of 192.12.0.0/17. AS-2 aggregates these prefixes into a single one, 192.12.0.0/16, and advertises it to its provider, AS-5. AS-5, in its turn, advertises it to the G-ISP. Consequently, the G-ISP learns that the AS-Path for network 192.12.0.0/16 via AS-5 is [AS-5 AS-2 {AS-1 AS-21}], where {AS-1 AS-21} represents the AS-SET of AS-1 and AS-21. However, it has no way to distinguish which part of this address space belongs to AS-1 and which does not. Since the G-ISP has a service agreement with AS-1, it will have to treat this advertisement as a path to its customer, thus serving also AS-21, which is not its customer.

The insertion of the intermediate AS numbers by the G-ISP is a crucial step in preventing routing loops, as shown in the following example. Suppose that AS-1 in Figure 2 is a G-ISP customer. The G-ISP learns two ways to reach AS-1. The route G-ISP learns from AS-4 contains four AS numbers and the route it learns from AS-5 contains only three AS numbers. Suppose that G-ISP prefers the route with the shortest AS-Path and forwards packets destined to AS-1 to AS-5. Now, assume that there is a link failure between the G-ISP and AS-5. In that case, the G-ISP will select AS-Path [AS-4 AS-3 AS-2 AS-1] as its best route to AS-1. As a G-ISP customer, AS-1 advertises to the G-ISP the following route to its network *N*: [AS-1]. If the G-ISP does not perform the AS-Path manipulation as explained before, it may advertise to its peers, and in particular to AS-4, that it can reach network *N* through AS-Path [G-ISP AS-1]. Since AS-Path [G-ISP AS-1] is shorter than [AS-3 AS-2 AS-1], AS-4 may choose this route for routing to network *N*, thus creating a routing loop.

4.2 Export rules from a G-ISP to its customers

As a service provider, a G-ISP should advertise to its customers its best routes to all destinations. These routes are learned from the G-ISP's peers, providers and other customers. As in Section 4.1, since there is no direct link between a G-ISP and its customer, some modifications are required to the AS-Paths that the customer receives from the G-ISP before these paths can be exported further

or used. The modifications in this case are performed on the customer side.

Assume that a G-ISP advertises to a customer C an AS-Path π to a destination prefix δ. Assume also that there is no direct link between C and the G-ISP. Hence the AS-Path C receives from the G-ISP does not contain the numbers of the intermediate ASs between C and the G-ISP. Therefore, in order for π to reflect the actual AS-Path to δ, C needs to add the list of intermediate AS numbers to the beginning of π. Note that this list appears in the AS-Path C uses for routing to the G-ISP. The customer should not change the next-hop attribute received for π as this is the IP address to which C should send packets via the IP tunneling mechanism, as described in Section 3. Note that this next-hop is not the real next-hop the customer uses in order to send the packets through the G-ISP, because the real next-hop actually belongs to the ISP of C. When the packet is sent to the G-ISP through the tunnel, the path used by customer C is the one learned from its ISP. In this path the next-hop attribute is the next-hop in ISP. C uses this next-hop to forward packets to its G-ISP, which in turn routes these packets to δ.

We will now demonstrate why it is important that the customer C has the full AS-Path from a G-ISP to the destination prefix δ, including the numbers of the ASs between its ISP and its G-ISP. In Figure 2 AS-1 is a customer of AS-2 (an ISP) and is also a customer of a G-ISP. Suppose that the policy of AS-1 is to choose the route with the shortest AS-Path. Consider the routing from AS-1 to a destination d in AS-100 whose prefix belongs to δ. AS-1 learns two ways to reach δ. The first way, learned from AS-2, uses AS-Path [AS-2 AS-3 AS-4 AS-100]. The second, learned from G-ISP, is [G-ISP AS-4 AS-100]. If AS-1 does not perform the AS-Path manipulation as explained above, AS-1 will prefer routing to δ through G-ISP because [AS-2 AS-3 AS-4 AS-100] is longer than [G-ISP AS-4 AS-100]. But this is of course a wrong decision.

4.3 Export rules from G-ISP to its peers

A G-ISP is supposed to provide transit services only to packets originated by its customer's networks and, optionally, to packets destined for these networks. To achieve the former goal, packets originated by a G-ISP's customer are forwarded to the G-ISP through the IP tunneling mechanism

as described in Section 3. In order to serve packets that are destined for a customer's host, the G-ISP should advertise to its peers the best routes to those customers that are subscribed to a bi-directional service. Using this export policy, packets destined to these customers would be able to traverse the G-ISP on the way to their destinations. When a G-ISP peer receives such an advertisement for the G-ISP, it can adopt this route or ignore it. Since the G-ISP does not advertise its routes to destination addresses that do not belong to one of its customers, it will not receive packets for these destinations from its peers. Note that a G-ISP has no way to force its peers to use the routes it advertises. Therefore, these routes should be attractive enough (e.g., have a short AS-Path), to "encourage" the other ASs to route packets toward these destinations through the G-ISP.

5 Algorithms for Building a G-ISP Overlay Network

As already stated, the main motivation behind the proposed G-ISP extension for BGP is improving the routing of the G-ISP customers' packets, by reducing the end-to-end delay of delay-sensitive applications, reducing the loss rate of loss-sensitive applications, and providing "a better than best-effort" service when necessary. This goal could have been achieved through a direct link between the G-ISP and every AS in the Internet, in which case the G-ISP scheme would not be necessary at all. However, since the number of ASs in the Internet is more than 15,000, such an approach is of course impractical. Using the proposed extension for BGP, a G-ISP can be connected to a much smaller group of ASs, while still limiting the maximum number of ASs in a path. The idea is that the G-ISP will have a PoP (Point of Presence) in strategic Internet locations, such that packets transmitted to or from the G-ISP's customers will have to traverse only the G-ISP's AS and a limited number of non-G-ISP ASs before reaching their destinations. G-ISP customer's packets may encounter loss or excessive delay only in a non-G-ISP transient AS. By reducing the number of such ASs, we decrease the expected end-to-end delay and loss rate, without an unreasonable increase in the G-ISP size.

In this section we study the correlation between the number of ASs to which a G-ISP should

have direct connectivity and the maximum number of non-G-ISP transient ASs in a path between any two ASs. We present an algorithm that minimizes the number of ASs directly connected to the G-ISP. As a result, the number of intermediate ASs on the path between the G-ISP and some other AS is guaranteed not to be larger than r. This guarantees that the number of ASs between any source-destination pair is not larger than $2r + 1$. The values $r = 1$ and $r = 2$ are of special interest to us. We also show that by reducing the number of ASs that should be "covered" to 80% or 90% of the total ASs, the number of ASs with which the G-ISP must have direct connectivity can be substantially reduced. (The ASs that should be covered are those within a distance of r transient ASs from the G-ISP.)

5.1 The Minimum Dominating Set and the Minimum r-Dominating Set Problems

Suppose we are interested in limiting the maximum length of the shortest AS-Path from the G-ISP to another AS, AS_0 for example, to be no longer than 2. This implies that the G-ISP has to be directly connected to AS_0 or to a neighbor of AS_0. If r is defined to be the number of transient ASs between the G-ISP and AS_0, then $r = 1$. This gives rise to an optimization problem, where we seek a minimal subset of nodes in the AS graph for which a direct G-ISP connection will guarantee that every node not in this subset will be directly connected to another node in the subset. This problem is known as the Minimum Dominating Set Problem (MDS). It can be generalized, of course, from $r = 1$ to an arbitrary r. In the following, this generalization is referred to as the *Minimum r-Dominating Set* Problem (MrDS).

A formal definition of these problems is as follows:

Problem 1 (Minimum Dominating Set Problem) *Given an undirected graph $G = (V, E)$, find a minimum sized subset $V' \subseteq V$, such that for every $u \in V \setminus V'$ there exists $v \in V'$ such that $(v, u) \in E$.* □

Problem 2 (Minimum r-Dominating Set Problem) *Given an undirected graph $G = (V, E)$ and*

a positive integer $r \leq |V|$, find a minimum sized subset $V' \subseteq V$, such that every $u \in V \setminus V'$ has a shortest path of length r or less to some $v \in V'$. □

The Minimum r-Dominating Set Problem (MrDS) can be viewed as a Minimum Dominating Set Problem (MDS) with a domination *r*. In MrDS, a node dominates not only itself and its neighbors but also the nodes that are at a distance of *r* or less.

The MDS problem is known to be NP–complete [9]. Therefore, MrDS is NP-hard. In what follows we present an approximation algorithm for solving MrDS. This algorithm is based upon another algorithm for a similar problem called MSC (Minimum Set Cover) and is defined as follows:

Problem 3 (Minimum Set Cover Problem) *Given a set C of subsets of a finite set S, find a minimum sized subset $C' \subseteq C$ such that the union of all subsets in C' is equal to S.* □

MSC is also NP-complete. However, the following greedy algorithm guarantees an approximation ratio of $1 + log|C|$ (see [35]):

Algorithm 1 *(A greedy algorithm for MSC)*

Let $C' \leftarrow \varnothing$;

While there are elements in S which are not in any set a in the cover C' do:

- *Find the set $a \in C$ with the greatest number of elements that are not yet covered by any other set in C';*
- *Add a to the cover (i.e. $C' \leftarrow C' \cup \{a\}$);*

Output C' as the solution. □

Algorithm 2, presented in the following, uses Algorithm 1 as a building block in order to solve MrDS.

Algorithm 2 *(A greedy algorithm for MrDS)*

Given a graph $G(V, E)$ and a radius r do:

- *For every $v_i \in V$, let $\alpha_r(v_i)$ be the set containing v_i and all the nodes that are within a distance of r edges or less from v_i;*
- *Let $C = \{\alpha_r(v_i) : v_i \in V\}$;*
- *Run Algorithm 1 with $S = V$ and $C = \{\alpha_r(v_i)\}$ and return the collection of nodes represented by the elements of C' as an output;* □

Lemma 1 *Algorithm 2 solves the MrDS problem; namely, every node $v \in V$ is within a distance of r or less from a node in C'.*

Proof: Assume there is a node u that is not dominated by any of the nodes in C'. Hence, for every $v \in C'$, the distance between u and v is larger than *r*. This implies that there is no $v \in V$ such that $u \in \alpha_r(v)$. Hence, node u is not covered by Algorithm 1, thus contradicting the condition for the termination of this algorithm. ■

Lemma 2 *The approximation ratio of Algorithm 2 is $1 + log|V|$. Namely, if the size of the optimal solution is $|Opt|$, then the size of the solution C' returned by Algorithm 2 is at most $(1 + log|V|) \cdot |Opt|$.*

Proof: Follows directly from the fact that the approximation ratio of Algorithm 1 is $1 + log|C|$. ■

Note that from [25] it follows that MDS is not approximable within $C \cdot log|V|$, for some $C > 0$. Therefore, the approximation given above is the best possible up to a constant factor, and it is also the best known.

6 The Feasibility of the G-ISP Paradigm

In order to evaluate the feasibility of the G-ISP paradigm in the context of the proposed algorithms, we address three issues:

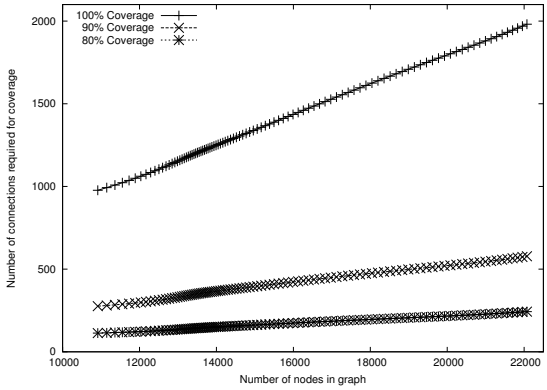


Figure 3: The number of ASs with which the G-ISP should have direct connectivity in order to achieve 80%, 90% and 100% coverage, as a function of the total number of ASs

- What is the number of ASs to which the G-ISP needs to connect in order to ensure that the length of the AS-Paths it advertises to its customers is at most $r + 1$, (i.e., that there are at most r non-G-ISP ASs between the G-ISP and every AS)? Specifically, we are interested in the case where $r = 1$ and $r = 2$. Note that this guarantees that the longest AS-Path from a G-ISP customer to any AS in the Internet will traverse at most $2r$ non-G-ISP ASs.

- If we are willing to compromise the number of ASs in the coverage, e.g., that the G-ISP would have a path with at most r transient ASs to only 80% or 90% of the total ASs, what is the number of ASs to which the G-ISP needs to have direct connectivity?

- What is the number of ASs the G-ISP needs to connect to in order to achieve 80% or 90% coverage for $r = 1$, when certain ASs (e.g., customer ASs) must be included in the coverage?

To answer these questions, we implemented Algorithm 2 and built 14 AS graphs. These graphs are based on more than five years of RIB table data, from April 2001 to May 2006. This data was obtained from the servers of the Route Views project [28]. We first checked the number of ASs the G-ISP would have had to connect in order to guarantee that the number of transient ASs between the G-ISP and any other node would not exceed r. To this end, we executed Algorithm 2 on our AS graph instances with $r = 1$ and $r = 2$. We found that for $r = 1$, this number ranged from 977 in April 2001, when the considered AS graph contained 10,915 nodes and 23,757 links, to 1981 in

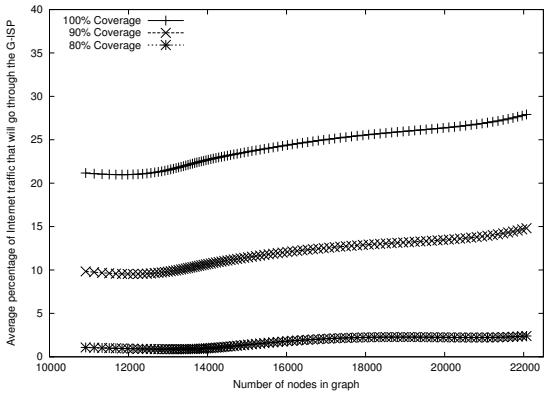


Figure 4: The percentage of improved AS-Paths, as a function of the total number of ASs

May 2006, when the AS graph contained 22,072 nodes and 44,780 links. The percentage of nodes to which the G-ISP would have had to be directly connected ranged between 8.6% to 9.1% for all the AS graph instances we checked over 5-year period – evidence for a linear dependency between these two numbers (see Figure 3). The percentage of AS-Paths that could have been improved as a result of adding the G-ISP node to the AS graphs is presented in Figure 4. When r is increased to 2, connecting to 155 ASs in April 2001 and to 311 ASs in May 2006 was enough.

As mentioned earlier, the G-ISP needs to have direct connectivity with almost 1500 ASs in order to advertise to its customer AS-Paths with no more than one intermediate non-G-ISP AS. Since such a requirement is impractical, at least during initial deployment, we also checked the percentage of ASs to which the G-ISP will have an AS-Path of length r or less as a function of the number of ASs with which the G-ISP has direct connectivity. The results of this study are presented in Figure 5. Two cases are considered: the case where $r = 1$, namely, there is at most one non-G-ISP AS between the G-ISP and every AS, and the case where $r = 2$.

The number of G-ISP connections required to cover only 80% or 90% of the ASs in the path – that is, the number of connections guaranteeing at most r intermediate non-G-ISP ASs – is surprisingly smaller than the number of connections required to cover all the ASs in the graph. Specifically, only 243 connections are required to guarantee that the G-ISP will have no more than 1 transient AS between itself and 80% of the ASs in the graph. Achieving the same connectivity for

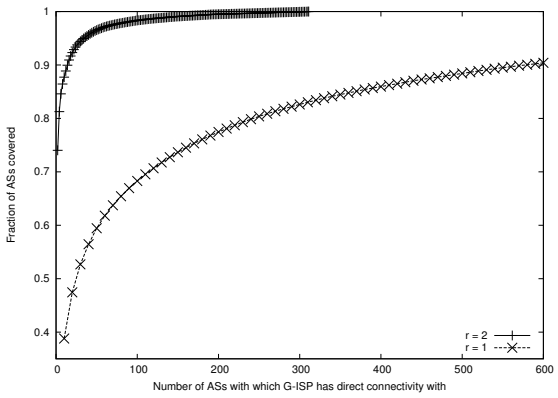
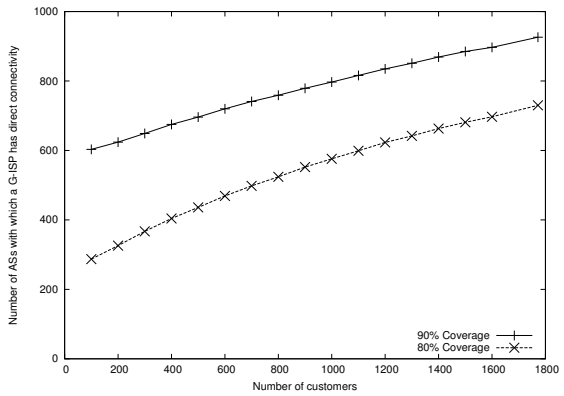


Figure 5: The fraction of "covered ASs," i.e., those with r or less transient ASs from the G-ISP vs. the number of ASs with which the G-ISP had to be directly connected (for the May 2006 data)

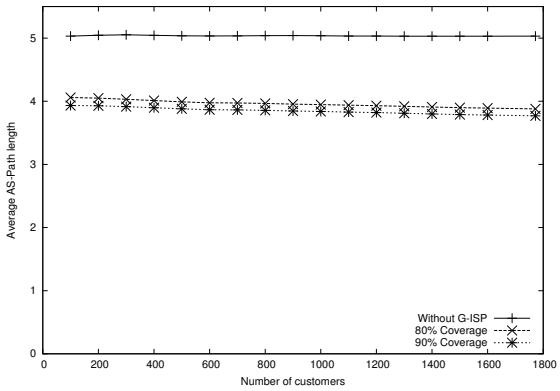
90% of the ASs requires 578 connections. From Figure 3 it follows that the same linear dependency found for 100% coverage holds for 80% and 90% coverages as well. However, from Figure 4 it follows that in these cases the effect of the G-ISP on the shortest AS-Paths is also reduced. When we increase r to 2, only 4 connections are needed for 80% coverage and only 15 connections are needed for 90% coverage.

The number of AS connections varies greatly from one AS to another and so the average AS-Path length also varies. A highly connected AS with an average AS-Path of length 3 does not need the services of the G-ISP as it can already reach most ASs quite efficiently. Therefore, selecting random ASs as potential G-ISP customers does not yield the best results. Consequently, we define a *potential customer* of the G-ISP to be an AS with an average AS-Path of length 4 or longer, for which a G-ISP connection improves its average AS-Path by at least 1 AS.

The AS graph for May 2006 contained 7984 ASs with an average AS-Path of length 4 or longer, 1772 of which could have reduced this average by 1 or more by becoming a customer of a G-ISP for $r = 1$. The average AS-Path lengths were computed without applying any restriction on the routing paths in the AS graph. Such restrictions are likely to reduce the number of feasible paths, and can, therefore, only increase the number of ASs with an average AS-Path longer than 4.



(a) Number of connections required for 80% and 90% coverage as a function of the number of customers



(b) Average AS-Path length with a G-ISP, for 80% and 90% coverage, and AS-Path without a G-ISP

Figure 6: Number of connections required for covering the potential customers, and the reduction of the average AS-Path length

We first checked how many ASs the G-ISP needs to connect to in order to ensure that it has no more than $r = 1$ transient ASs to all its customers and to a total of 80% or 90% of the ASs in the AS graph, as a function of the number of customers the G-ISP has. The customers were randomly chosen from the group of 1772 *potential customers* in the May 2006 AS graph. The results are presented in Figure 6(a). When the number of potential customers is 100, and the required coverage is 80%, the G-ISP would have had to connect to 287 ASs. If the number of potential customers is 1772 (the maximum number), then 730 G-ISP connections are required to guarantee 80% coverage. For 90% coverage, the number of required G-ISP connections increases to 603 and 926 respectively. The reduction in the average AS-Path length of these customers is presented in Figure 6(b). For 80% coverage, the average reduction for a G-ISP customer varies from 0.97 ASs for 100 customers to 1.15 ASs for 1772 potential customers. For 90% coverage, the average reduction varies from 1.1 ASs for 100 potential customers to 1.26 for 1772 potential customers. Note that although the reduction in the average AS-Path is around 20%, the actual benefit for a customer, in terms of reduced loss and latency, is greater. This is because one of the AS numbers in the AS-Path usually belongs to the G-ISP itself. Since the G-ISP employs an intra-AS QoS mechanism, the number of ASs that do not guarantee QoS but are traversed by a

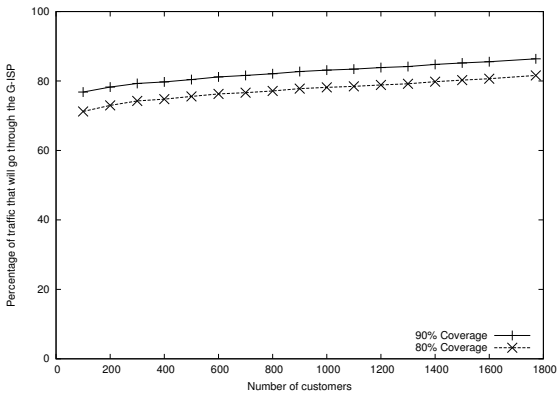


Figure 7: The percentage of improved AS-Paths as function of the number of customers (for May 2006)

packet is reduced by around 40%.

The percentage of improved AS-Paths for potential customers is presented in Figure 7. In May 2006, for 80% coverage, about 75% of the AS-Paths of the potential customers are improved, while for 90% coverage, about 80% are improved. The reason for these high numbers is that when there is no G-ISP connection, the intermediate AS-Path is long in most cases. When these potential customers have the option to route through the G-ISP, however, most of these long AS-Paths are significantly shortened.

In order to examine the applicability of our results, we examined two Internet databases: the Route Views project and the IRR. The Route Views project [28] is a BGP based database that collects a snapshot of the Internet AS level topology on a daily basis. The Internet Routing Registry [26] is a union of world-wide routing policy databases that use the Routing Policy Specification Language. These databases contain, among other things, the local connectivity of the registered ASs. Using up-to-date data from April 2006, we discovered that over 120 ASs have more than 150 peering relationships with other ASs. Moreover, we found that 18 ASs are connected to more than 500 ASs. We therefore conclude that our requirement of a G-ISP with 150-200 connected peers is very reasonable.

7 Applications of the G-ISP Paradigm

As we will show in this section, gradual G-ISP deployment is possible and the investment required is reasonable. This is one of the most important properties of the proposed scheme.

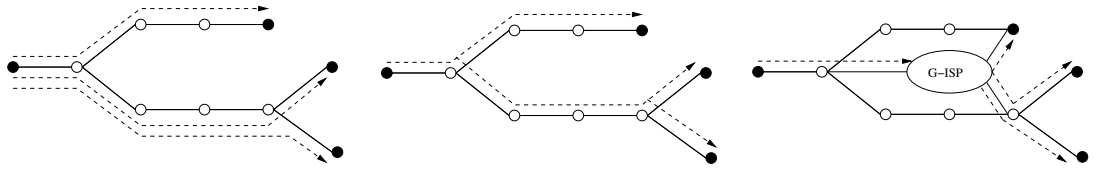
7.1 Inter-AS multicast using the G-ISP paradigm

As noted earlier, the G-ISP paradigm can facilitate the deployment of multicast services. Inter-AS multicast is a tough problem because it combines the difficulties of policy-based inter-AS unicast routing with those of building efficient multicast trees while accommodating dynamic changes in the multicast groups. Therefore, there is no standard protocol for inter-domain multicast, and multicast services are not delivered across the AS boundaries. Efforts toward such a protocol are being pursued by the IETF BGMP working group [33].

We propose to use the G-ISP as a core for multicast services. The G-ISP will use a standard intra-AS multicast protocol like PIM [6] for distributing the multicast packets over its overlay AS. When a customer AS wants to send a multicast packet to some multicast group, it will first send it to the G-ISP like a unicast packet. The G-ISP will then distribute this packet to its border routers over the internal multicast tree, and from these routers to its other customer ASs using unicast. Finally, at each customer AS, the packet will be distributed using the intra-domain multicast protocol deployed at this AS. The advantages of this approach are: (a) the standard BGP is used for policy-based routing; (b) standard internal multicast protocols are used for building the multicast tree in the G-ISP AS.

This approach offers a novel way to provide inter-AS multicast. While other approaches for providing this facility, like MBone [5] and BGMP, rely on the existent Internet topology, our approach expands Internet connectivity for the G-ISP customers. The high connectivity of the G-ISP, as well as the fact that it is designed to be located in proximity to its customers, enables it to provide them with efficient inter-AS multicast services.

In order to evaluate the effectiveness of the proposed scheme, we compared the performance



(a) unicast-based approach (b) inter-AS SPT approach (c) G-ISP-based approach

Figure 8: The three different approaches for Inter-AS multicast

of three approaches for providing inter-AS multicast services over the Internet:

1. Using unicast: A dedicated copy of each packet is sent from the sender AS to the AS of every group member. The advantage of this approach is that BGP is used for inter-AS routing. The disadvantage is that many copies of the same packet will be forwarded over the same routes. Note, however, that an AS with multiple group members receives a single copy.

2. Using a hypothetical inter-AS shortest-paths multicast tree (SPT) with the sender as the root. This is the best approach one could imagine, where multicast trees are allowed to cross AS boundaries. This approach is very similar to the proposed BGMP standard [33], with one main difference: while in BGMP a core AS serves as the root of the tree for each multicast group, here we assume that the sender AS is the root for the multicast. Hence, the results of this approach are expected to be better than those of pure BGMP.

3. Using the proposed G-ISP-based approach as described above.

These approaches are described in Figure 8, which presents an AS graph. ASs with members in the multicast group are represented by black nodes, while non-member ASs are represented by white nodes. The leftmost node is the multicast originator node. The figures show, using dashed lines, the AS routes traversed by every copy of every multicast packet.

We computed the multicast cost for the different approaches for the case where the destination ASs are customers of the G-ISP. These customer ASs were randomly selected from the group of *potential customers* as discussed earlier. The first two approaches were tested using the regular AS graph, without considering a G-ISP. For the third approach we added a G-ISP to the AS graph.

This G-ISP was connected such that each of its routes to 80% of the ASs in the Internet has at most 1 transient AS. Recall that achieving this 80% coverage requires the G-ISP to have direct connectivity with 287 ASs for 100 customers and 730 ASs for 1772 customers. The cost C of the multicast is computed as $\sum C_i$, where C_i is the number of multicast copies received by AS-i as a result of a single multicast packet. In Figure 8, for example, $C = 14$ for the unicast-based approach, $C = 9$ for the inter-AS shortest-path multicast approach, and $C = 7$ for the G-ISP-based approach. Note that the cost of the G-ISP-based approach can be lower than that of the inter-AS SPT approach, as happens to be the case in Figure 8. The reason is that, although we do add a G-ISP, we likewise increase the connectivity of the AS graph. This in turn increases the possible routes available to the customers.

Figure 9 shows the cost C for each approach as a function of the number of ASs in the multicast group. The main conclusion from this graph is that the performance of the G-ISP-based multicast is almost equal to the performance of the best inter-AS multicast protocol one could ever implement over the Internet. If there are less than 500 ASs in the multicast group, then the cost of multicast using the G-ISP-based approach is equal to or even smaller than the cost of the optimal approach. When the number of AS members increases, the G-ISP-based approach is only slightly inferior to the optimal approach. This is because the cost of adding a new member to the multicast group in the G-ISP-based approach is constant, whereas in the inter-AS SPT approach the cost decreases when the size of the multicast tree increases.

7.2 End-to-end QoS for virtual private networks

Supporting end-to-end inter-AS QoS in the Internet is another intractable problem. Intra-domain QoS is relatively easy because an AS is administrated by a single authority that usually has sufficient knowledge of the domain topology, the available resources, and the actual needs. However, since an AS administrator does not have a global view of the Internet, intra-AS QoS mechanisms cannot be extended to inter-AS. For this reason, although some ISPs offer their customers "a better than best-effort" Service Level Agreement (SLA), inter-AS end-to-end QoS cannot be guaranteed.

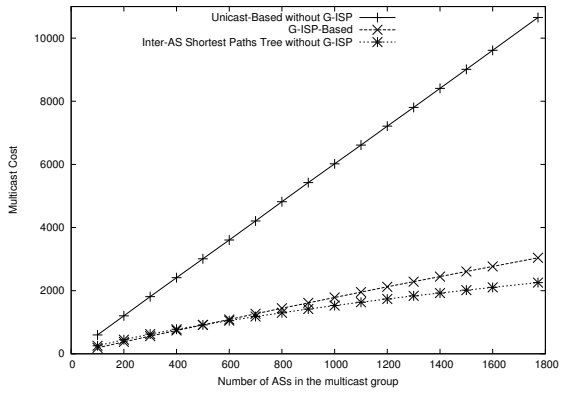


Figure 9: The cost of multicast for the different approaches

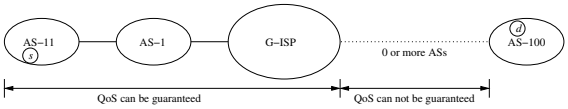


Figure 10: QoS with the G-ISP

The main demand for inter-AS QoS arises today in the context of Intranet/Extranet VPN. A VPN is a private data network established over a public data network like the Internet. An Intranet/Extranet VPN establishes layer-3 connectivity between remote sites of one (in the case of Intranet) or more (in the case of Extranet) organizations. If the VPN is established over the Internet, the lack of inter-AS QoS mechanisms precludes the provision of VPN QoS. The G-ISP paradigm can facilitate inter-AS QoS support for Intranet/Extranet VPNs in the following way. Consider the situation described in Figure 10. In this figure AS-11 is a customer of the G-ISP. As such, it is separated from the G-ISP only by its ISP, AS-1. Assuming that this ISP has some mechanism for ensuring intra-AS QoS, AS-11 can obtain an appropriate QoS on the tunnel to the G-ISP by negotiating an appropriate SLA with this ISP. If every other AS in the VPN of AS-11 is also a customer of the G-ISP, end-to-end QoS is guaranteed within this VPN.

8 Conclusions

We presented an extension to BGP referred to as the G-ISP paradigm. Its goal is to solve inherent problems associated with inter-domain routing. The main idea behind the proposed scheme is to distinguish between a local ISP and a global ISP, as is done in the telephony world. An AS will use the physical connectivity with the local ISP in order to set up a logical BGP connectivity with the global ISP. We described the BGP extension required to support the proposed paradigm. Only the BGP of ASs that desire G-ISP services is affected by the proposed extension.

A G-ISP can offer its customers a set of improved services, including efficient inter-AS multicast support and end-to-end inter-AS QoS for VPNs. Such services are not available on the Internet today. These services are in addition to the improved routing services the G-ISP can provide to its customers. We believe that with the ongoing increase in demand for better Internet services, will come an increased willingness in many organizations and corporations to pay more for such services. The current limitations of Internet inter-domain routing make the G-ISP paradigm a promising solution with a strong business case.

References

[1] D. Andersen. Resilient overlay networks, master thesis, MIT, 2001.

[2] G. Apostolopoulos, D. Williams, S. Kamat, R. Guerin, A. Orda, and T. Przygienda. QoS routing mechanisms and OSPF extensions. RFC-2676, August 1999.

[3] A. Bremler-Barr, Y. Afek, and S. Schwarz. Improved BGP convergence via ghost flushing. In *INFOCOM 2003*, 2003.

[4] J. Cobb and R. Musunuri. Enforcing convergence in inter-domain routing. In *IEEE Global Communications (GLOBECOM) Conference*, volume 2, pages 511–517, 2004.

[5] Hans Eriksson. MBONE: the multicast backbone. *Communications of the ACM*, 37(8):54–60, 1994.

[6] D. Estrin, D. Farinacci, A. Helmy, D. Thaler, S. Deering, M. Handley, V. Jacobson, C. Liu, P. Sharma, and L. Wei. Protocol independent multicast-sparse mode (PIM-SM): protocol specification. RFC-2362, June 1998.

[7] L. Gao and J. Rexford. Stable internet routing without global coordination. In *Measurement and Modeling of Computer Systems*, pages 307–317, 2000.

[8] Lixin Gao, Timothy Griffin, and Jennifer Rexford. Inherently safe backup routing with BGP. In *INFOCOM*, pages 547–556, 2001.

[9] M.R. Garey and D.S. Johnson. *Computers and Intractability – A guide to the theory of NP-completeness*. W.H. Freeman and Company, 1979.

[10] T. Griffin and B. Premore. An experimental analysis of BGP convergence time. In *ICNP 2001*, 2001.

[11] T. Griffin, F. Shepherd, and G. Wilfong. Policy disputes in path-vector protocols. In *ICNP*, pages 21–30, 1999.

[12] T. Griffin and G. Wilfong. An analysis of BGP convergence properties. In *SIGCOMM*, pages 277–288, 1999.

[13] T. Griffin and G. Wilfong. A safe path vector protocol. In *INFOCOM (2)*, pages 490–499, 2000.

[14] G. Huston. Commentary on inter-domain routing in the Internet. RFC-3221, December 2001.

[15] J. Jannotti, D. Gifford, K. Johnson, M. Kaashoek, and J. O'Toole. Overcast: Reliable multicasting with an overlay network. In *Fourth Symposium on Operating System Design and Implementation (OSDI)*, pages 197–212, 2000.

[16] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian. Delayed internet routing convergence. In *SIGCOMM*, pages 175–187, 2000.

[17] C. Labovitz, A. Ahuja, R. Wattenhofer, and V. Srinivasan. The impact of Internet: Policy and topology on delayed routing convergence. In *INFOCOM*, pages 537–546, 2001.

[18] P. Levis, M. Boucadair, P. Morand, J. Spencer, D. Griffin, G. Pavlou, and P. Trmintzios. A new perspective for a global QoS-based internet. *Journal of Communications Software and Systems*, 2005.

[19] G. Malkin. RIP version 2. RFC-2453, November 1998.

[20] J. Moy. Multicast extensions to OSPF. RFC-1584, March 1994.

[21] J. Moy. OSPF version 2. RFC-2328, April 1998.

[22] K. Nichols, S. Blake, F. Baker, and D. Black. Definition of the differentiated services field in the IPv4 and IPv6 headers. RFC-2474, December 1998.

[23] D. Pei, M. Azuma, D. Massey, and L. Zhang. BGP-RCN: improving BGP convergence through root cause notification. *Comput. Netw. ISDN Syst.*, 48(2):175–194, 2005.

[24] D. Pei, X. Zhao, L. Wang, D. Massey, A. Mankin, S. Wu, and L. Zhang. Improving BGP convergence through consistency assertions. In *INFOCOM 2002*, New York, NY, Jun 2002.

[25] R. Raz and S. Safra. A sub-constant error-probability low-degree test, and sub-constant error-probability PCP characterization of NP. In *Proc. 29th Ann. ACM Symp. on Theory of Comp.*, 1997.

[26] The Internet Routing Registry. http://www.irr.net.

[27] Y. Rekhter and T. Li. A border gateway protocol 4 (bgp-4). RFC-4271, January 2006.

[28] Routeviews.org. The university of oregon route views archive project.

[29] S. Shi and J. Turner. Placing servers in overlay networks. In *International Symposium of Performance Evaluation of Computer and Telecommunication Systems (SPECTS)*, 2002.

[30] S. Shi and J. Turner. Routing in overlay multicast networks, 2002.

[31] S. Shi, J. Turner, and M. Waldvogel. Dimensioning Server Access Bandwidth and Multicast Routing in Overlay Networks. In *Proceedings of NOSSDAV 2001*, pages 83–92, June 2001.

[32] W. Simpson. IP in IP tunneling. RFC-1853, October 1995.

[33] D. Thaler. Border gateway multicast protocol (BGMP): Protocol specification. draft-ietf-bgmp-spec-05.txt, June 2003.

[34] K. Varadhan, R. Govindan, and D. Estrin. Persistent route oscillations in inter-domain routing. *Computer Networks*, 32(1):1–16, 2000.

[35] V. Vazirani. *Approximation Algorithms*. Springer, 2001.

[36] D. Waitzman, C. Partridge, and S. Deering. Distance vector multicast routing protocol. RFC-1075, November 1988.

[37] L. Xiao, K. Lui, J. Wang, and K. Nahrstedt. QoS extension to BGP. In *The 10th IEEE International Conference on Network Protocols*, November 2002.

[38] B. Zhang, D. Massey, and L. Zhang. Destination reachability and BGP convergence time. In *Global Telecommunications Conference, 2004. GLOBECOM '04. IEEE*, volume 3, pages 1383–1389, 2004.