Interdomain Traffic Engineering with BGP

Bruno Quoitin, Cristel Pelsser, and Louis Swinnen, University of Namur
Olivier Bonaventure and Steve Uhlig, Université Catholique de Louvain

ABSTRACT

Traffic engineering is performed by means of a set of techniques that can be used to better control the flow of packets inside an IP network. We discuss the utilization of these techniques across interdomain boundaries in the global Internet. We first analyze the characteristics of interdomain traffic on the basis of measurements from three different Internet service providers and show that a small number of sources are responsible for a large fraction of the traffic. Across interdomain boundaries, traffic engineering relies on a careful tuning of the route advertisements sent via the Border Gateway Protocol. We explain how this tuning can be used to control the flow of incoming and outgoing traffic, and identify its limitations.

INTRODUCTION

Initially developed as a network that connected a small number of research networks, the Internet has become a worldwide data network that is used for mission-critical applications. Supporting such mission-critical applications across the global Internet implies several important challenges. The first challenge is the size of the Internet. The Internet is a large decentralized network that connected about 160 million hosts in June 2002. Furthermore, these hosts are organized in about 13,000 distinct domains, a domain corresponding roughly to one company or one Internet service provider (ISP). All these domains are interconnected to form the global Internet. The Border Gateway Protocol (BGP) is used to route the IP packets exchanged between domains. There are basically two types of domains. *Stub domains* contain hosts that produce or consume IP packets. These domains do not carry IP packets that are not produced by or destined to their hosts. *Transit domains* interconnect different domains, and carry IP packets that are produced by and/or destined to external domains. Additional details on the relationships between domains may be found in [1].

The second challenge is that the research Internet was designed with best effort service in mind where connectivity was the most important issue. Today, connectivity is taken for granted, and the best effort service is used for mission-critical applications with stringent service level agreements (SLAs). To meet these SLAs, several ISPs rely on traffic engineering [2] to better control the flow of IP packets. Large ISPs often need to engineer the flow of packets inside their own domain to reduce congestion by better distributing the traffic on all their links. Several techniques have been developed during the last few years. Some require the utilization of multiprotocol label switching (MPLS) to forward IP packets, while others only require tuning of the traditional IP routing protocols used inside the ISP network. Besides optimizing the flow of packets inside their network, most ISPs also need to better control the flow of their interdomain traffic (i.e., IP packets that cross the boundaries between distinct ISPs). Today, MPLS is not used across interdomain boundaries, and the only solution to engineer the flow of interdomain traffic is to tune the configuration of the BGP routing protocol. This tuning is often done on a trial-and-error basis and suffers from limitations, as will be shown in the rest of this article.

In this article we first introduce the operation of the BGP protocol. We provide recent results about the characteristics of interdomain traffic. Finally, we describe in details several interdomain traffic engineering techniques and show their limitations.

BGP ROUTING IN THE INTERNET

Internet routing is handled by two distinct protocols with different objectives. Inside a single domain, link state intradomain routing protocols distribute the entire network topology to all routers and select the shortest path according to a metric chosen by the network administrator. Across interdomain boundaries, the interdomain routing protocol is used to distribute reachability information and to select the best route to each destination according to the policies specified by each domain administrator. For scalability reasons, the interdomain routing protocol is only aware of the interconnections between distinct domains; it does not know any information about the content of each domain.

BGP [3, 4] is the current de facto standard interdomain routing protocol. In BGP terminology, a domain is called an autonomous system (AS). BGP is a *path-vector protocol* that works by sending *route advertisements*. A route advertisement indicates the reachability of a network (i.e., a network address and a netmask representing a

0163-6804/03/$17.00 © 2003 IEEE

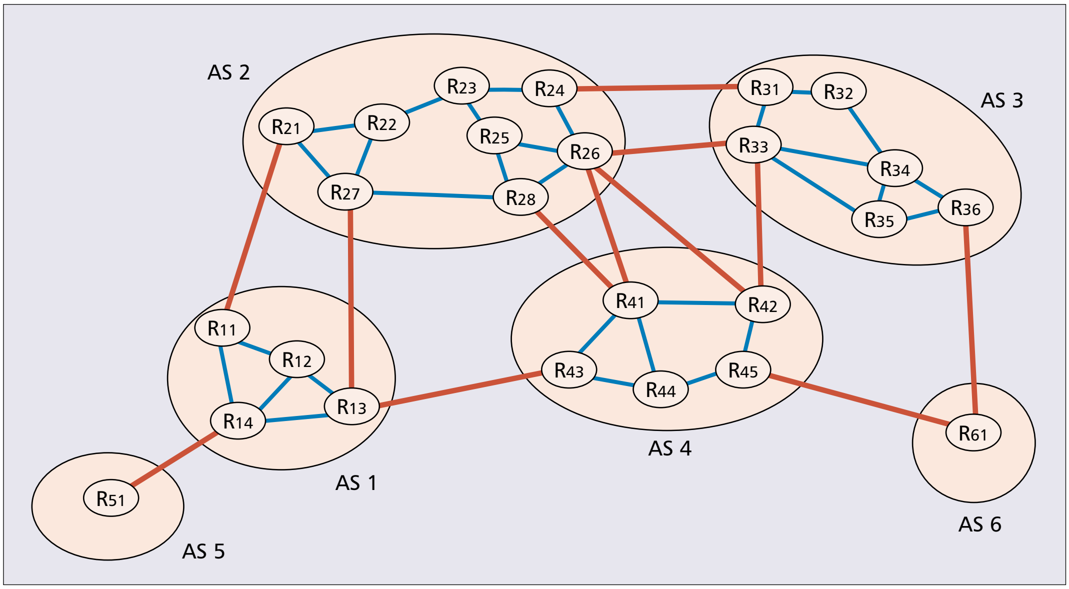

Figure 1. *A simple Internet.*

block of contiguous IP addresses; e.g., `192.168.0.0/24` represents a block of 256 addresses between `192.168.0.0` and `192.168.0.255`) because this network belongs to the same AS as the advertising router or because a route advertisement for this network was received from another AS. Besides the reachable network and the IP address of the router that must be used to reach this network (known as the *next hop*), a route advertisement also contains the AS path attribute, which contains the list of all the transit ASes that must be used to reach the announced network. The length of the AS path can be considered as the route metric. A route advertisement may also contain several optional attributes such as the `local-pref`, multi-exit discriminator (`MED`), or `communities` attributes [3, 4]. An important point to note about BGP is that if a BGP router of ASx sends a route announcement for network *N* to a neighbor BGP router of ASy, this implies that ASx accepts forwarding the IP packets to destination *N* on behalf of ASy.

There are two variants of BGP [3, 4]. The eBGP variant is used to announce the reachable prefixes on a link between routers that are part of distinct ASs (e.g., R_{51} and R_{14} in Fig. 1). The iBGP variant is used to distribute inside an AS the best routes learned from neighboring ASes.

Inside a single domain, all routers are considered equal, and the intradomain routing protocol announces all known paths to all routers. In contrast, in the global Internet, all ASs are not equal, and an AS will rarely agree to provide a transit service for all its connected ASs toward all destinations. Therefore, BGP allows a router to be selective in the route advertisements it sends to neighbor eBGP routers. To better understand the operation of BGP, it is useful to consider a simplified view of a BGP router, as shown in Fig. 2.

A BGP router processes and generates route advertisements as follows. First, the administrator specifies, for each BGP peer, an input filter (Fig. 2, left) that is used to select the acceptable advertisements. For example, a BGP router could only select the advertisements with an AS path containing a set of trusted ASes. Once a route advertisement has been accepted by the input filter, it is placed in the BGP routing table, possibly after updating some of its attributes. The BGP routing table thus contains all the acceptable routes received from the BGP neighbors.

Second, on the basis of the BGP routing table, the BGP decision process (Fig. 2, center) will select the best route toward each known network. Based on the next hop of this best route and the intradomain routing table, the router

Inside a single domain, all routers are considered as "equal" and the intradomain routing protocol announces all known paths to all routers. In contrast, in the global Internet, all ASs are not equal and an AS will rarely agree to provide a transit service for all its connected ASs toward all destinations.

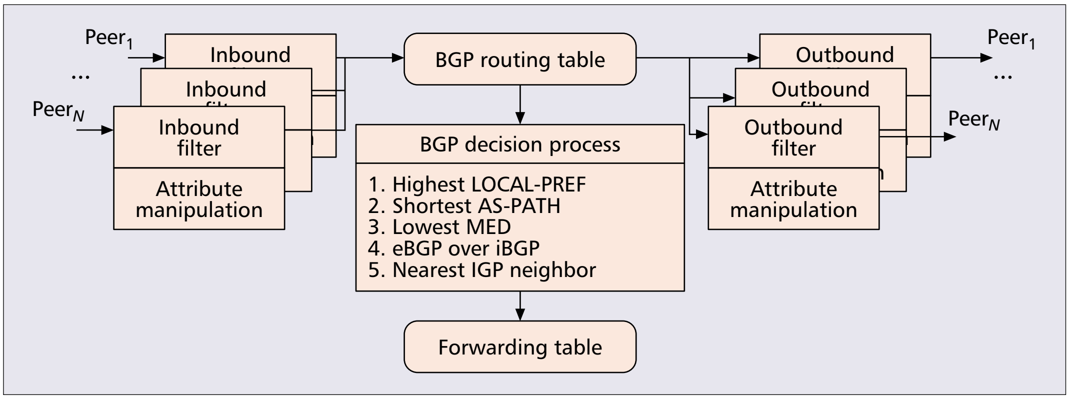

Figure 2. *Simplified operation of a BGP router.*

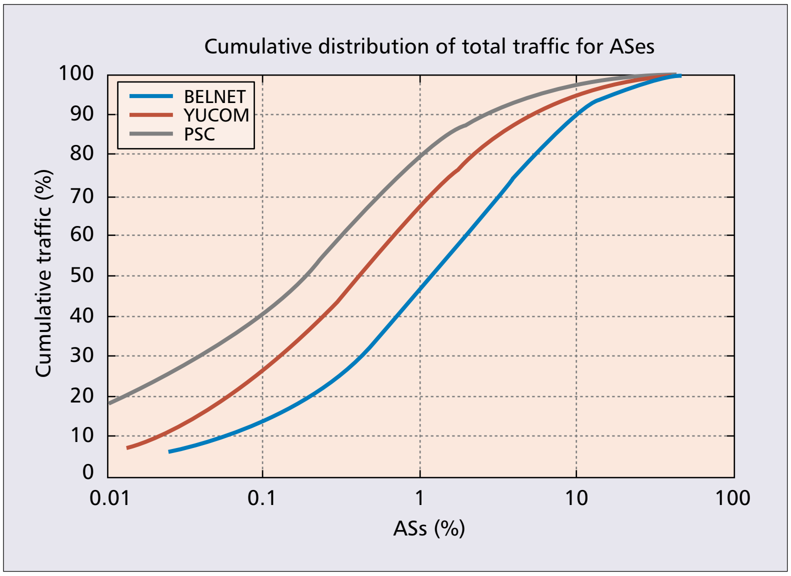

Figure 3. *Cumulative distribution of the traffic for each studied ISP.*

will install a route toward this network inside its forwarding table. This table is then looked up for each received packet and indicates the outgoing interface that must be used to reach the packet's destination.

Third, the BGP router will use its output filters (Fig. 2, right) to select, among the best routes in the BGP routing table, the routes that will be advertised to each BGP peer. At most one route will be advertised for each reachable destination. The BGP router will assemble and send the corresponding route advertisement messages after a possible update of some of their attributes.

The input and output filters used in combination with the BGP decision process are the key mechanisms that allow a network administrator to support within BGP the business relationships between two ASes. Many types of business relationships can be supported by BGP. Two of the most common are the customer-to-provider and peer-to-peer relationships [1]. To understand how these two relationships are supported by BGP, consider Fig. 1. If AS5 is AS1's customer, AS5 will configure its BGP router to announce its routes to AS1. AS1 will accept these routes and announce them to its peer (AS4) and upstream provider (AS2). AS1 will also announce to AS5 all the routes it receives from AS2 and AS4. If AS1 and AS4 have a peer-to-peer relationship on the link between R_{13} and R_{43}, router R_{13} will only announce on this link the internal routes of AS1 and the routes received from AS1's customer (i.e., AS5). The routes received from AS2 will be filtered and thus not announced on the R_{13}–R_{43} link by router R_{13}. Due to this filtering, AS1 will not carry traffic from AS4 toward AS2.

Characteristics of Interdomain Traffic

An important element to consider when engineering the interdomain traffic of an AS are the characteristics of this traffic. Informal discussions with network operators on this topic indicate that often a small number of ASs are responsible for a large fraction of the total traffic received or sent by a given ISP. However, while there are many studies on the topology of the Internet [1, references therein] or the evolution of the BGP routing tables [5, references therein] as well as many studies on the packet-level characteristics of the traffic (e.g., [6] among many other papers), few papers analyze the traffic and its topological distribution together [7–9].

In the framework of a detailed analysis of interdomain traffic, we have collected several traces of all the traffic received or sent through the border routers of three stub ISPs. Due to practical reasons, it was unfortunately not possible to collect a trace at the same time with three ISPs. The first trace was collected during one entire week in December 1999. This trace covers all the interdomain links of BELNET. The trace contains all the interdomain traffic received by BELNET, the Belgian Internet provider for universities and research laboratories. During this period, BELNET received 2.1 Tbytes of data from 4243 distinct ASs. The second trace was collected during five consecutive days in April 2001 at the border routers of YUCOM, an ISP based in Belgium that provides dialup access to home users. Again, this trace covers all the interdomain links of this ISP. During this five-day period, YUCOM received IP packets corresponding to 1.1 Tbytes of data from 7669 distinct ASs. The last trace was collected during one day at the border routers of the Pittsburgh Supercomputing Center (PSC) in March 2002. PSC provides access to Internet and Internet2 for organizations in western Pennsylvania. The trace captured all the interdomain traffic sent by PSC through its border routers. During the studied day, the border routers of PSC sent IP packets corresponding to 574 Gbytes of data to 11,791 distinct ASs. The difference in the number of ASs for each studied ISP is mainly due to the number of ASs in the Internet at the time of the measurement. In December 1999, BELNET had 6298 distinct ASs in its routing table, while YUCOM knew 10,560 ASs in May 2001, and PSC knew almost 12,000 ASes in March 2002.

The above description of each ISP reveals an important fact. Each ISP exchanges IP packets with a large fraction of the Internet over a period of a few days. Based on this sole information, interdomain traffic engineering would appear difficult since an AS would need to influence most of the Internet to control its traffic. Fortunately, a closer look at the traffic exchanged with each AS reveals several interesting points.

In Fig. 3, we show the cumulative distribution of the interdomain traffic received by BELNET and YUCOM and sent by PSC. To plot this figure, we classified the traffic in each trace on the basis of its source and destination AS. A similar result would have been found at the prefix level (see [9] for such an analysis with BELNET).

A first point to note about Fig. 3 is that the studied ISPs do not exchange the same amount of traffic with each remote AS. The 10 (resp. 100) largest sources of traffic for YUCOM contribute to more than 30 percent (resp. 72 percent) of the traffic received by this ISP. Similarly, the 10 (resp. 100) largest sources of traffic for BELNET contribute to 22 percent (resp. 64 percent) of the traffic it receives during one week. For PSC, the concentration of the traffic sinks is even more important as the 10 (resp. 100) largest destinations receive 38 percent (resp. 78 percent) of the total traffic sent by PSC. [8] mentions a similar distribution for the interdomain traffic of a large *tier 1* ISP.

Another important point to mention about the interdomain traffic exchanged by the studied ISPs is the distance (measured in AS hops) between the remote ASs and each studied ISP. Figure 4 shows, for each ISP, the percentage of its interdomain traffic that was produced by or sent to remote ASes as a function of their distance measured in AS hops. This analysis shows that the studied ISPs only exchange a small fraction of their traffic with their direct peers (AS hop distance on 1). Most of the packets are exchanged with ASs that are only a few AS hops away. For the BELNET trace, most of the traffic is produced by sources located three and four AS hops away, while YUCOM mainly receives traffic from sources that are two and three AS hops away. PSC, on the other hand, sends traffic to ASs located up to four AS hops away.

This analysis has two important implications for interdomain traffic engineering. First, although an AS will exchange packets with most of the Internet, only a small number of ASs are responsible for a large fraction of interdomain traffic. This implies that an AS willing to engineer its interdomain traffic could move a large amount of traffic by influencing a small number of distant ASs. Second, the sources or destinations of interdomain traffic are not direct peers, but they are only a few AS hops away. This implies that interdomain traffic engineering solutions should be able to influence ASs a few hops beyond their upstream providers or direct peers.

Interdomain Traffic Engineering

Interdomain traffic engineering requirements are diverse and often motivated by the need to balance the traffic on links with other ASs and to reduce the cost of carrying traffic on these links. These requirements depend on the connectivity of an AS with others, but also on the type of business handled by this AS.

Connectivity between ASs is mainly composed of two types of relationships. The most frequent relationship between ASs is the *customer-to-provider* relationship, where a customer AS pays to use a link connected to its provider. This relationship is the origin of most of the interdomain cost of an AS. A stub AS usually tries to maintain at least two of these links for performance and redundancy reasons [1]. In addition, larger ASs typically try to obtain *peer-to-peer* relationships with other ASs and then share the cost of the link with the other AS. Negotiating the establishment of those *peer-to-peer* relationships is often a complicated process since technical and economical factors, as exposed in [10], need to be taken into account.

Moreover, an AS will want to optimize the way traffic enters or leaves its network, based on its business interests. Content providers that host a lot of Web or streaming servers and usually have several customer-to-provider relationships with transit ASs will try to optimize the way traffic leaves their networks. On the contrary, access providers that serve small and medium enterprises, and dialup or xDSL users typically wish to optimize how Internet traffic enters their networks. Finally, a transit AS will try to balance the traffic on the multiple links it has with its peers.

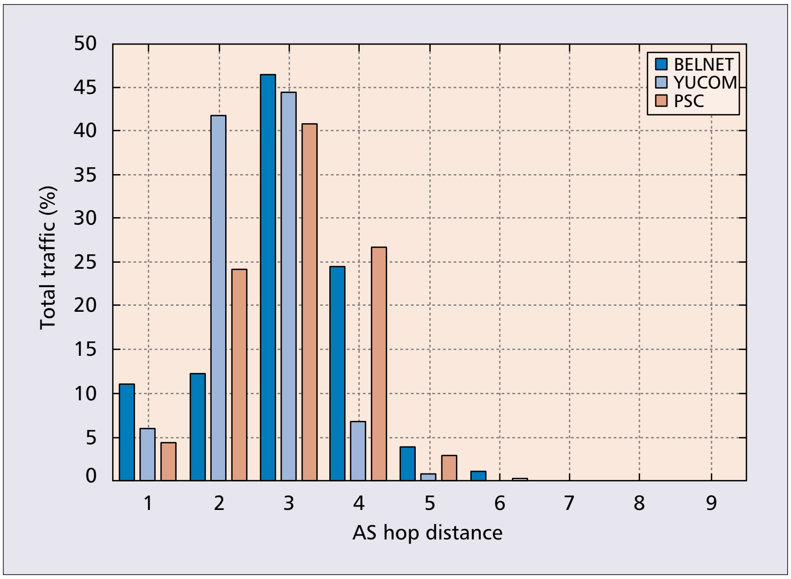

Figure 4. *Per-AS hop distribution of traffic.*

Optimizing the way traffic enters or leaves a network means to favor one link over another to reach a given destination or receive traffic from a given source. This type of interdomain traffic engineering can be performed by tweaking the BGP routers of the AS. In order to understand how BGP can be used to control the way traffic enters, leaves or crosses an AS, a better understanding of the BGP decision process is required. A BGP router receives one route toward each destination from each of its peers. To select the best route among this set of routes, a BGP router relies on a set of criteria called the decision process. Most BGP routers apply a decision process similar in principle to the one shown in Fig. 2. The set of routes with the same destination are analyzed by the criteria in the sequence indicated in Fig. 2. These criteria act as filters, and the *N*th criterion is only evaluated if more than one route has passed the *N* – 1th criterion. It should be noted that most BGP implementations allow the network administrator to optionally disable some of the criteria of the BGP decision process.

Control of the Outgoing Traffic

To control how the traffic leaves its network, an AS must be able to choose which route will be used to reach a particular destination through its peers. Since an AS controls the decision process on its BGP routes, it can easily influence the selection of the best path. Two techniques are frequently used.

A first technique is to rely on the `local-pref` attribute. This optional attribute is only distributed inside an AS. It can be used to rank routes and is the first criterion of the BGP decision process (Fig. 2). For example, consider a stub AS with two links toward one upstream provider: a high bandwidth and a low bandwidth link. In this case, the BGP router of this AS could be configured to insert a low `local-pref` to routes learned via the low bandwidth link and a higher value to routes learned via the high bandwidth link. A similar situation can occur for a stub AS connected to a cheap and a more expensive upstream provider.

In practice, the manipulation of the `local-pref` attribute can also be based on passive or active measurements. Recently, a few companies have implemented solutions [11] that allow multihomed stub ASs and content providers to engineer their interdomain traffic. These solutions usually measure the load on each interdomain link, and some rely on active measurements to evaluate the performance of interdomain paths. Based on these measurements and some knowledge of the Internet topology (obtained either through a central server or from the BGP router to which they are attached), they attach appropriate values of the `local-pref` attribute to indicate which route should be considered as the best route by the BGP routers.

To control how the traffic leaves its network an AS must be able to choose which route will be used to reach a particular destination through its peers. Since an AS controls the decision process on its BGP routes, it can easily influence the selection of the best path.

A second technique, often used by large transit ISPs, is to rely on the intradomain routing protocol to influence how a packet crosses the transit ISP. As shown in Fig. 2, the BGP decision process will select the nearest IGP neighbor when comparing several equivalent routes received via iBGP. For example, consider in Fig. 1 that router R_{27} receives one packet whose destination is R_{45}. The BGP decision process of router R_{27} will compare two routes toward R_{45}, one received via R_{28} and the other received via R_{26}. By selecting router R_{28} as the exit border router for this packet, AS2 will ensure that this packet will consume as few resources as possible inside its own network. If a transit AS relies on a tuning of the weights of its intradomain routing protocol as described in [12], this tuning will indirectly influence its outgoing traffic.

Control of the Incoming Traffic

The first method that can be used to control the traffic that enters an AS is to rely on selective advertisements and announce different route advertisements on different links.[1] For example, in Fig. 1, if AS1 wanted to balance the traffic coming from AS2 over the links $R_{11} – R_{21}$ and $R_{13} – R_{27}$, it could announce only its internal routes on the R_{11}–R_{21} link and only the routes learned from AS5 on the R_{13}–R_{27} link. Since AS2 would only learn about AS5 through router R_{27}, it would be forced to send the packets whose destination belongs to AS5 via router R_{27}. However, a drawback of this solution is that if link R_{13}–R_{27} fails, AS2 would not be able to reach AS5 through AS1. This is not desirable, and it should be possible to utilize link R_{11}–R_{21} for the packets toward AS5 at that time without being forced to change the routes that are advertised on this link.

A variant of the selective advertisements is the advertisement of more specific prefixes. This advertisement relies on the fact that an IP router will always select in its forwarding table the most specific route for each packet (i.e., the matching route with the longest prefix). For example, if a forwarding table contains both a route toward `16.0.0.0/8` and a route toward `16.1.2.0/24`, a packet whose destination is `16.1.2.200` would be forwarded along the second route. This fact can also be used to control the incoming traffic. In the following example, we assume that prefix `16.0.0.0/8` belongs to AS3 and that several important servers are part of the `16.1.2.0/24` subnet. If AS3 prefers to receive the packets toward its servers on the R_{24}–R_{31} link, it would advertise both `16.0.0.0/8` and `16.1.2.0/24` on this link and only `16.0.0.0/8` on its other external links. An advantage of this solution is that if link $R_{24} – R_{31}$ fails, then subnet `16.1.2.0/24` would still be reachable through the other links.

Another method would be to allow an AS to indicate a ranking among the various route advertisements that it sends. Based on the utilization of the length of the `AS-path` as the third criteria in the BGP decision process, a possible way to influence the selection of routes by a distant AS is to artificially increase the length of the AS path attribute. Coming back to Fig. 1, assume that AS3's primary interdomain link is link R_{61}–R_{45}, while link R_{61}–R_{36} is only used as a backup link. In this case, AS6 would announce its routes normally on the primary link (i.e., with an AS path of AS6) but would attach its own AS number several times instead of once in the AS path attribute (e.g., `AS6 AS6 AS6`) on the R_{61}–R_{36} link. The route advertised on the primary link will be considered as the best route by all routers that do not rely on manually configured settings for the `weight` and `local-pref` attributes. This technique can be combined with selective advertisements. For example, an AS could divide its address space in two prefixes $p1$ and $p2$ and advertise prefix $p1$ without prepending and prefix $p2$ with prepending on its first link and the opposite of its second link.

The last method to allow an AS to control its incoming traffic is to rely on the MED attribute. This optional attribute can only be used by an AS multiconnected to another AS to influence the link that should be used by the other AS to send packets toward a specific destination. It should, however, be noted that the utilization of the MED attribute is usually subject to a negotiation between the two peering ASs, and some ASs do not take the MED attribute into account in their decision process.

Community-Based Traffic Engineering

In addition to these techniques, several ISPs have been using the `communities` attribute to give their customers finer control over the redistribution of their routes. The `communities` attribute is an optional attribute that can be attached to routes. This attribute can contain several 32 bits wide community values. Community values are often used to attach optional information to routes such as a code representing the city where the route was received or a code indicating whether the route was received from a peer or a customer. The community values can also be used for traffic engineering purposes. In this case, predefined `community` values can be attached to routes in order to request actions such as not announcing the route to a specified set of peers, prepending the `AS-path` when announcing the route to a specified set of peers or setting the

[1] *It should be noted that such behavior is considered a wrong behavior in peer-to-peer relationships by some ISPs.*

local-pref. However, this technique relies on an ad hoc definition of community values and manual configurations of BGP filters, which makes it difficult to use and subject to errors.

The Internet Engineering Task Force (IETF) is currently considering the definition of a new standard type of extended communities called *redistribution communities* [13] to solve the drawbacks of the utilization of classical communities to do traffic engineering. These redistribution communities can be attached to routes to influence the redistribution of those routes by the upstream AS. The redistribution communities attached to a route contain both the traffic engineering action to be performed and the BGP peers affected by this action. One of the supported actions allows an AS to indicate to its upstream peer that it should not announce the attached route to some of its BGP peers.

Another type of action allows an AS to request its upstream to perform AS-path prepending when redistributing a route to a specified peer. To understand the usefulness of such redistribution communities, let us consider again Fig. 1, and assume that AS6 receives a lot of traffic from AS1 and AS2 and that it would like to receive the packets from AS1 (resp. AS2) on the R_{45}–R_{61} (resp. R_{36}–R_{61}) link. AS6 cannot achieve such a traffic distribution by performing AS-path prepending itself. However, this becomes possible with the redistribution communities by requesting AS4 to perform the prepending when announcing the AS6 routes to external peers. AS6 could thus advertise to AS4 its routes with a redistribution community that indicates that this route should be prepended two times when announced to AS2. With this redistribution communities, AS4 would advertise path AS4:AS4:AS6 to AS2 and path AS4:AS6 to AS1. AS2 would thus receive two routes toward AS6: AS4:AS4:AS6 and AS3:AS6, and would select the route via AS3. AS1, on the other hand, would select the AS4:AS6 route that is shorter than the AS2:AS3:AS6 route.

Discussion

The sections above describe several techniques that can be used by ISPs to engineer their interdomain traffic. However, there are some limitations to be considered when deploying those techniques.

A first point to note is that the control of the outgoing traffic with BGP is based on the selection of the best route among the available routes. This selection can be performed on the basis of various parameters, but is limited by the diversity of routes received from upstream providers, which depends on the connectivity and the policy of these ASs.

The control of incoming traffic is based on careful tuning of the advertisements sent by an AS. This tuning can cause several problems. First, an AS that advertises more specific prefixes or has divided its address space in distinct prefixes to announce them selectively will advertise a number of prefixes larger than required. All these prefixes will be propagated throughout the global Internet and will increase the size of the BGP routing tables of potentially all ASs in the Internet. Reference [5] reports that more specific routes constitute more than half of the entries in a BGP table. Faced with this increase of their BGP routing tables, several large ISPs have started to install filters to ignore the BGP advertisements corresponding to more specific prefixes. The deployment of those filters implies that the more specific prefixes will not be announced by those large ISPs, and thus the technique will become much less effective.

When considering the manipulation of the AS-path attribute, we have mentioned that it can be used on backup links. It is sometimes also used to better balance the traffic ([5] reports that AS-path prepending affected 6.5 percent of the BGP routes in November 2001). However, in practice it can be difficult to predict the outcome of performing AS-path prepending on a given interdomain link. Usually, ISPs that rely on AS-path prepending select the amount of prepending on a trial and error basis.

The redistribution communities can provide a finer granularity than AS-path prepending or selective announcements. In practice, it can be expected that those communities will be used to influence the redistribution of routes toward large transit ISPs with a large number of customers. For example, consider as an example YUCOM, discussed earlier. This ISP has two major upstream providers that allow it to reach the entire Internet. These two providers are then each connected to several tier 1 ISPs that provide most of their connectivity. Figure 5 provides a subset of the Internet topology as seen by YUCOM on the basis of the BGP advertisements it received from its two providers. In this figure, we show the three largest tier 1 ISPs that were connected to YUCOM's providers. Based on the BGP advertisements received by YUCOM, it appears that both providers sent advertisements for routes reachable via one of those tier 1 ISPs (tier 1B in Fig. 5), while only one of those providers sent advertisements for routes reachable via each of the two other tier 1 ISPs. In addition to this topological information, Fig. 5 also reports the number of distinct ASs reachable via each tier 1 ISP via the two providers of YUCOM. For tier 1C, this number indicates that provider 2 sent to YUCOM BGP advertisements toward 2470 distinct ASs that are reachable via tier 1C.

Figure 5 reveals two interesting pieces of information. First, each tier 1 ISP provides connectivity and thus announces routes toward a large number of ASs. In total, the three largest tier 1 providers announce more than 8500 ASs. Second, the studied ISP learns routes toward more than 2000 different ASs reachable via tier 1B via its two upstream providers. By using redistribution communities targeted at those large tier 1 ISPs, our ISP could influence the redistribution of its routes to a large number of ASs with only a few communities. For example, the studied ISP could utilize a single redistribution community to request its first upstream provider to announce its local routes with AS-path prepending only toward tier 1B. The result of this modified advertisement by the first provider will be that the traffic coming from ASs attached to tier 1B would be received through the other provider. Another point to mention concerning the usefulness of the redistribution communities is that, as shown in Fig. 4, most sources of traffic are located at only a few AS hops away. The redistribu-

The redistribution communities attached to a route contain both the traffic engineering action to be performed and the BGP peers affected by this action. One of the supported actions allows an AS to indicate to its upstream peer that it should not announce the attached route to some of its BGP peers.

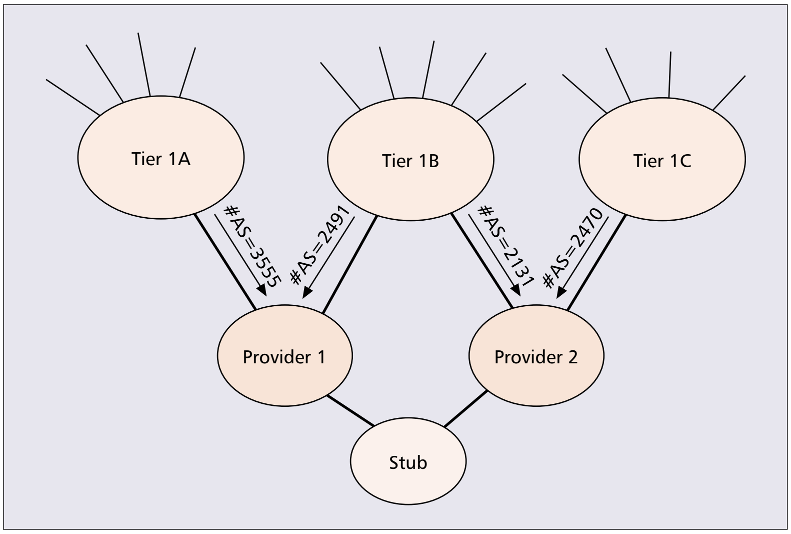

Figure 5. *Subset of the interdomain topology seen from the studied ISP and number of different ASes advertised by each tier-1 ISP.*

tion communities can directly influence sources located two AS hops away and indirectly sources three or four AS hops away.

A last point to note concerning the techniques that require changes to the attributes of BGP advertisements is that any (small) change to an attribute will force the route advertisement to be redistributed to potentially the entire Internet. Although it would be possible to define techniques relying on measurements to dynamically change the BGP advertisements of an AS for traffic engineering purposes, a widespread deployment of such techniques would increase the number of BGP messages exchanged and could lead to BGP instabilities. Any dynamic interdomain traffic engineering technique that involves frequent changes to the values of BGP attributes should be studied carefully before being deployed.

Conclusion

In this article, we describe several techniques that are used today to control the flow of packets in the global Internet. We first describe the current organization of the Internet and the key role played by BGP.

We then discuss the characteristics of interdomain traffic based on long traces covering all the interdomain links of three distinct ISPs. Two common characteristics appear in those traces. First, although the Internet is composed of about 13,000 ASs today, a small percentage of those ASs contribute to a large fraction of the traffic received or sent by those ISPs. Second, those highly active sources or destinations are located only a few AS hops away, although the adjacent ASs are only responsible for a small fraction of the total traffic.

We explain how BGP is tuned today for interdomain traffic engineering purposes. We have shown that an AS has more control over its outgoing than over its incoming traffic. Several techniques can be used to control the incoming traffic, but they have limitations. The selective advertisements and the more specific prefixes have the drawback of increasing the size of the BGP routing tables. With AS path prepending, it can be difficult to select the appropriate value of prepending to achieve a given goal. Finally, we show how redistribution communities could allow an AS to flexibly influence the redistribution of its routes toward indirectly connected ISPs.

Acknowledgments

This work was supported by the European Commission within the IST ATRIUM project. We would like to thank C. Rapier (PSC), B. Piret (YUCOM), and M. Roger (BELNET) for their traffic traces. We also thank S. De Cnodder, C. Filsfils, A. Danthine, and the anonymous reviewers for their useful comments.

References

[1] L. Subramanian *et al.*, "Characterizing the Internet Hierarchy from Multiple Vantage Points," *INFOCOM 2002*, June 2002.
[2] D. Awduche *et al.*, "Overview and Principles of Internet Traffic Engineering," IETF RFC3272, May 2002.
[3] Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)," Internet draft, draft-ietf-idr-bgp4-17.txt, work in progress, May 2002.
[4] J. Stewart, *BGP4 : Interdomain Routing in the Internet*, Addison Wesley, 1999.
[5] A. Broido, E. Nemeth, and K. Claffy, "Internet Expansion, Refinement and Churn," *Euro. Trans. Telecommun.*, Jan. 2002.
[6] K. Thompson, G. Miller, and R. Wilder, "Wide-Area Internet Traffic Patterns and Characteristics," *IEEE Network*, vol. 11, no. 6, Nov./Dec. 1997.
[7] W. Fang and L. Peterson, "Inter-AS Traffic Patterns and Their Implications," *IEEE Global Internet Symp.*, Dec. 1999.
[8] N. Feamster, J. Borkenhagen, and J. Rexford, "Controlling the Impact of BGP Policy Changes on IP Traffic," AT&T tech. memo, 011106-02, Nov. 2001.
[9] S. Uhlig and O. Bonaventure, "Implications of Interdomain Traffic Characteristics on Traffic Engineering," *Euro. Trans. Telecommun.*, Jan. 2002.
[10] S. Bartholomew, "The Art of Peering," *BT Tech. J.*, vol. 18, no. 3, July 2000.
[11] S. Borthick, "Will Route Control Change the Internet?," *Bus. Commun. Rev.*, Sept. 2002.
[12] B. Fortz, J. Rexford, and M. Thorup, "Traffic Engineering with Traditional IP Routing Protocols," *IEEE Commun. Mag.*, Oct. 2002.
[13] O. Bonaventure *et al.*, "Controlling the Redistribution of BGP Routes," Internet draft, draft-ietf-ptomaine-bgp-redistribution-01.txt, work in progress, Aug. 2002.

Biographies

OLIVIER BONAVENTURE (Bonaventure@info.ucl.ac.be) received his Ph.D. from the University of Liege, Belgium. His research interests include interdomain routing and traffic engineering, IP-based mobile networks, and network measurements. He was with the University of Namur, Belgium, and is currently a professor at the Université Catholique de Louvain, Belgium.

CRISTEL PELSSER (cpe@info.fundp.ac.be) received her M.S. degree in computer science from the University of Namur in 2001. She currently works as a researcher at the university. Her research interests include interdomain traffic engineering and routing.

BRUNO QUOITIN (bqu@info.fundp.ac.be) received his M.S. degree in computer science from the University of Namur in 1999. He worked in industry for two years and is currently a researcher at the university. His research interests include interdomain routing and traffic engineering.

LOUIS SWINNEN (lsw@info.fundp.ac.be) received his M.S. degree in computer science from the University of Namur in 2001. His research interests include interdomain routing, network simulations, and simulator development. He also gives some courses at the Haute Ecole Mosane d'Enseignement Superieur, Belgium.

STEVE UHLIG (suh@info.ucl.ac.be) received his M.S. degree in computer science from the University of Namur in 2000.