Improving BGP Protocol to Advertise Multiple Routes for the Same Destination Prefix

Aleksandar Cvjetić and Aleksandra Smiljanić

***Abstract*—An Internet autonomous system (AS) uses BGP policies in order to meet its local objectives (e.g. optimal routing within AS) and peering contracts with neighboring ASes (e.g. advertising some specific routes to a neighbor, forwarding traffic over specific paths, etc.) In this paper we propose an enhancement of BGP route selection and advertisement processes, and demonstrate how the enhanced BGP can be used to flexibly implement diverse policies, regardless of the network topology, that were not feasible before.**

***Index Terms*—BGP, BGP policies, route selection, XORP.**

I. Introduction

BGP policies present a set of rules that define how an AS routes incoming and outgoing traffic to the Internet. In BGP, only one route is selected and advertised for each network destination prefix. However, in common practical cases, advertisement of a single route may not be sufficient for the implementation of basic BGP policies simultaneously [1]. For example, a single selected route that meets the local objectives of an AS might violate peering contracts with neighboring ASes [2]. In addition, large ASes tend to use different scaling techniques for BGP protocol (like BGP route reflectors) that additionally reduce the number of visible routes and hinder the realization of BGP policies.

In [3], [4], extensions of the BGP UPDATE message format are proposed that allow multiple BGP routes for the same destination prefix. In [5], [6], the authors propose different mechanisms to compute a set of alternate routes to a destination prefix that do not share inter-domain links and ASes in order to improve path diversity and network reliability. However, these proposals do not address a failure of the conventional BGP to flexibly implement policies within a given AS, which is the main goal of the BGP extension that we propose.

In practice, we found that only some routers with the latest software release support the advertisement of multiple BGP routes as specified in [3]. There are also few commercial implementations that support the advertisement of the best external BGP (eBGP) route in addition to the routers best local route, if the latter is an internal BGP (iBGP) route. This, however, will not help the realization of BGP policies in problematic cases. Namely, since a router prefers eBGP over iBGP routes in the route selection process, only one route will be advertised. Also, the most popular open-source routers like XORP, Quagga and Vyatta comprise BGP routines that advertise a single BGP route for each destination prefix [7].

In this paper, we propose a novel BGP route selection algorithm, termed BGP with Flexible Routing Policies (BGP-FRP), which advertises multiple routes for a network prefix so that the vital BGP policies can be implemented simultaneously. In BGP-FRP, multiple routes for the same prefix are propagated only within a single AS, while in [5], [6], they are propagated through different domains as well. Unlike in [3]–[6] our solution does not require any extensions of the BGP message format, but only a slight modification of the existing route selection process.

We describe the implementation of our BGP-FRP route selection process within the XORP open-source router and demonstrate policies that were not feasible before. BGP-FRP is designed to require minimal modifications of the conventional BGP implementation, while allowing execution of important policies for traffic routing. In addition, routing can be adjusted much faster in the case of network topology changes if routers store multiple alternative routes. At the end we will analyze additional resources required by BGP-FRP using the data from real networks.

II. Some Common Issues with the Existing BGP

The BGP route selection algorithm consists of several steps in which a router compares BGP route attributes and selects the one with: 1. the highest local preference, 2. the lowest AS path, 3. the lowest origin value, 4. the lowest Multi-Exit Discriminator (MED), 5. the eBGP over the iBGP route, 6. the lowest internal cost to the BGP next hop, 7. the lowest BGP ID and 8. the lowest peer IP address [8]. The selected route is then advertised to other BGP neighbors. However, there are cases where advertising one route may cause violation of peering contracts with neighboring ASes or suboptimal routing within the AS. For example, Figure 1 (a) shows two customer ASes, AS 300 and AS 400, advertising routes for a destination prefix to provider AS 100. Customer AS 300 advertises redundant routes, r_1 and r_2, but assigns higher preference to r_2 (with the lower BGP MED attribute), indicating that AS 100 should use route r_2 to forward traffic to it. With the existing BGP route selection process router R3 in AS 100 will first eliminate route r_1 in the fourth step, because r_1 has higher MED than r_2, but in some of the remaining steps, between r_2 and r_3, R3 may select r_3 as the best route (e.g. if r_3 contains lower value for the router BGP ID). If so, R3 will advertise route r_3 to the other neighbors inside AS 100. Router R2 will select and advertise route r_1, because it prefers eBGP (r_1) over the iBGP route (r_3) in the fifth step. As a result, only route r_1

Manuscript received May 29, 2013. The associate editor coordinating the review of this letter and approving it for publication was H.-P. Schwefel.

This work is financed in part by the Serbian Ministry of Science.

A. Cvjetić is with Vip Mobile, Belgrade, Serbia (e-mail: aleksandar.cvjetic@gmail.com).

A. Smiljanić is with the School of Electrical Engineering, Belgrade University, Belgrade, Serbia (e-mail: aleksandra@etf.rs).

Digital Object Identifier 10.1109/LCOMM.2013.111513.131250

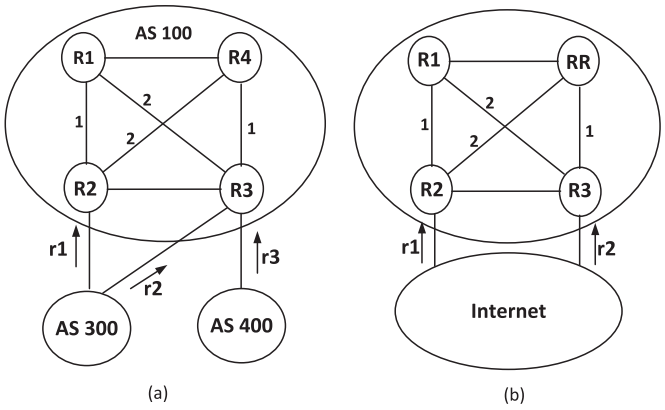

Fig. 1: (a) Violation of peering contracts with customer; (b) suboptimal routing within AS.

of customer AS 300 is known within provider AS 100, and this route will be used for sending traffic to this customer, although the customer assigned the lower preference to it.

For the transit traffic provider ASes usually implement policy based on hot-potato routing, i.e. routing where each provider router selects the route with the lowest metric to the BGP next hop that corresponds to the highest capacity link. In addition, large ASes usually deploy route reflectors (RRs) so that other routers within the AS maintain iBGP connections only with the RR, and RR advertises routes between iBGP neighbors. For example, in Figure 1 (b) if RR receives two routes, r_1 and r_2, from its neighbors R2 and R3 respectively, and if those routes are equal in the first five steps of the BGP route selection process, in the sixth step RR will select and advertise r2 because it has a lower metric (1) to the BGP next hop than route r1 (with the metric 2). Since R1 receives only routes from RR, it will also select r_2 for which the BGP next hop is still router R3 (RR does not change the BGP next hop while advertising routes). However, from the perspective of R1, r_2 does not meet the rules of hot-potato routing, because the metric to the BGP next hop is higher (2) than the one that would be provided by route r_1 (1).

These particular examples show how the advertising of a single route in BGP may influence both the peering contracts and the internal routing objectives of an AS. In addition, single route advertising limits some other aspects of Internet routing, like robustness to routing failures, fast convergence and load balancing. In the following section we will describe our novel BGP route selection algorithm, BGP-FRP, which allows the advertisement of multiple BGP routes with minimal modification of the existing route selection process.

III. BGP WITH FLEXIBLE ROUTING POLICIES

Previously proposed protocols do not resolve the problem of inconsistent policies described in the previous section. We divide all BGP route selection steps into two groups: 1. AS-specific steps, in which comparison of the attributes provides the same result for the best route on all routers within the AS, and 2. router-specific steps, in which comparison of the attributes provides different results for the best route on routers within the AS. Attributes like the local preference, AS path and MED are usually changed only by AS boundary routers (ASBRs) that apply BGP policy before the route selection process, and they stay unchanged throughout the AS. On the other side, attributes like the internal cost of a route, BGP ID and peer address change as a route traverses the routers of an AS, and different routers may select different routes based on this set of attributes. Since the origin attribute stays unchanged from the beginning of the route advertisement, we classify steps from 1 to 5 as AS-specific while the remaining steps are router-specific, and formulate the following rules for multiple route advertisement:

1) If a router selects the best route in some of the AS-specific steps, there is no need to advertise other routes within AS, because all routers will select the same route anyway.

2) If after the completion of AS-specific steps some routes remain in the route selection process, they should be advertised to other routers within the AS. In this way, routers in the AS will store a complete set of viable routes and can comply with flexible routing policies.

We implemented BGP-FRP using the XORP open-source platform. We modified the class DecisionTable in the xorp/bgp directory of the XORP router that contains methods for executing the BGP route selection process so that these methods return multiple best routes instead of one route. We used the C++ STL (Standard Template Library) list container to store the best routes and our method follows the existing rules up to the fifth step, after which all the remaining routes are assigned to the list of new winners. We also added a new command to the command line interface of the XORP router, so that a new process can be explicitly activated by users. It is added to the BGP template file and to the XRL (Xorp Resource Locator) interface file as a new method of the BGP protocol. The syntax of the command that we implemented to activate the new BGP route selection process is: *set protocols bgp use-multi-ibgp true*, where *use-multi-ibgp* represents the parameter that activates the new route selection process.

IV. PERFORMANCE OF BGP-FRP

We will demonstrate how BGP-FRP implements basic sets of BGP policies that are needed in practice, but were not feasible before. We create a virtual network consisting of five XORP routers installed on virtual nodes with Linux OS. BGP implementation on these virtual routers deploys a modified version of the route selection algorithm, as described in the previous section. Additional resources (such as bandwidth, CPU and memory) required by our BGP-FRP will be estimated using data from the real networks.

A. Advertising preferred route of customer AS 300

In order to meet the peering contract with customer AS 300 router R3 in provider AS 100 must advertise routes r_2 and r_3 to the other routers within AS 100 (Figure 1 (a)). Only in this case the preferred route of customer AS 300 (r_2) will be known within AS 100 and will be used by other routers to forward traffic. Figure 2 (a) depicts a virtual network that we created to demonstrate a routing scenario using BGP-FRP. We configured the IP addresses and BGP IDs of each router as shown in the figure. BGP is configured to exchange routes

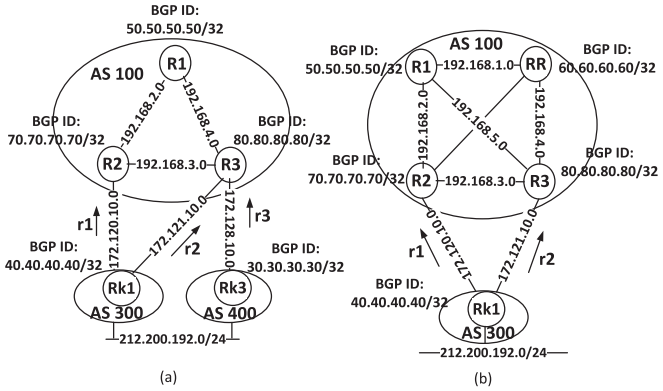

Fig. 2: (a) Meeting peering contract with customer AS 300; (b) optimal routing within AS 100.

```
root@aleksR1> show route table ipv4 unicast final
212.200.192.0/24 [ibgp(200)/0]
                > to 192.168.2.2 via eth1/eth1
root@aleksR2> show route table ipv4 unicast final
212.200.192.0/24 [ebgp(20)/0]
                > to 172.120.10.2 via eth0/eth0
root@aleksR3> show route table ipv4 unicast final
212.200.192.0/24 [ebgp(20)/0]
                > 172.128.10.2 via eth3/eth3
```

```
root@aleksR1> show route table ipv4 unicast final
212.200.192.0/24 [ibgp(200)/0]
                > to 192.168.4.2 via eth0/eth0
root@aleksR2> show route table ipv4 unicast final
212.200.192.0/24 [ibgp(200)/0]
                > to 192.168.3.2 via eth1/eth1
```

Fig. 3: Routing table outputs from the routers in AS 100 before and after activation of the new BGP route selection algorithm on R3.

of neighboring ASes and OSPF is configured to exchange internal routes between routers in AS 100. Both customers AS 300 and AS 400 advertise the common network prefix $212.200.192.0/24$ to the provider AS 100 via BGP.

With the current BGP implementation, router R3 will select and advertise only route r_3, while router R2 will select route r_1 as can be seen from the outputs of routing tables in the upper box of Figure 3. However, if we activate BGP-FRP on R3, it will advertise both r_2 and r_3 because both routes are equal in AS-specific steps. As shown in the lower box of Figure 3, R1 and R2 choose to forward traffic for destination prefix $212.200.192.0/24$ through the next hops $192.168.4.2$ and $192.168.3.2$ respectively, which are IP addresses of the corresponding links to router R3 (links R1-R3 and R2-R3), so those routers do not use route r_1 anymore for this destination.

B. Optimal routing within AS 100

In order for all routers in AS 100 to meet the policy of hot-potato routing RR must advertise r_1 and r_2 to the neighbors in AS 100 (Figure 1 (b)). In Figure 2 (b), router Rk1 in AS 300 advertises the network prefix $212.200.192.0/24$ to AS 100 border routers R2 and R3 via BGP. Each border router then advertises its route for this destination to RR. The upper box of Figure 4 shows that the route for $212.200.192.0/24$

```
root@aleksR1> show route table ipv4 unicast final
212.200.192.0/24 [ibgp(200)/0]
                > to 192.168.5.2 via eth1/eth1
```

```
root@aleksR1> show route table ipv4 unicast final
212.200.192.0/24 [ibgp(200)/0]
                > to 192.168.2.2 via eth0/eth0
```

Fig. 4: Routing table outputs on R1 before and after activation of new BGP route selection process on RR.

on R1 points to the next hop $192.168.5.2$ when using the standard BGP algorithm, which is the IP address of the link R1-R3 with the higher metric (see Figure 1 (b)). However, when we activate BGP-FRP on RR, it will advertise both r_1 and r_2 to the neighbors in AS 100. Now the route for the $212.200.192.0/24$ on R1 points to the next hop $192.168.2.2$ which is the IP address of the link R1-R2, i.e. the link with the lower metric (the lower box of Figure 4).

C. Performance considerations

Main advantage of BGP-FRP is that it supports more comprehensive BGP policies compared to the existing solutions. Multiple routes for a given destination prefix should only be advertised when the standard BGP algorithm cannot meet the specific requirements of a BGP policy, or if the network reliability needs to be improved. In addition, multiple routes are advertised only within an AS, and are not propagated to the other ASes.

In order to estimate how many routes will be advertised within a real AS, we first collected the outputs of BGP Routing Information Base (RIB) tables from six public route servers connected to the Internet Exchange Points (IXPs) in Amsterdam, London, New York, Frankfurt, Moscow and Sao Paulo. We plotted the probability density functions (PDFs) and cumulative density functions (CDFs) of the number of twin routes with the same AS-specific attributes. Figure 5 shows these two functions based on the sample of 8M routes that were collected by the Amsterdam route server. The Amsterdam router was chosen as an adequate representative since it exchanges 95% of all routes on the Internet. It can be seen from Figure 5 that the maximum number of routes with the same AS-specific attributes is 17, and for 95% of the prefixes only 11 routes need to be advertised. Similar, in fact better, curves are obtained for other routers. In the typical (stable) network conditions around 600 routing table entries change every 5 minutes on average [10], which means that a new route arrives approximately every half-second. Since at most 17 routes arrive every half-second when multiple routes are advertised in BGP-FRP in the worst case and the maximum size of the BGP UPDATE message is 4096 bytes, then the worst-case maximum link bandwidth consumed by BGP-FRP can be calculated as follows:

$$BW = \frac{17 \times 4096 \times 8\text{bits}}{0,5\text{s}} = 1,1\frac{\text{Mb}}{\text{s}} \qquad (1)$$

This worst-case bandwidth consumption is still acceptable, while its probability is below 0.01 permille (Figure 5 (a)).

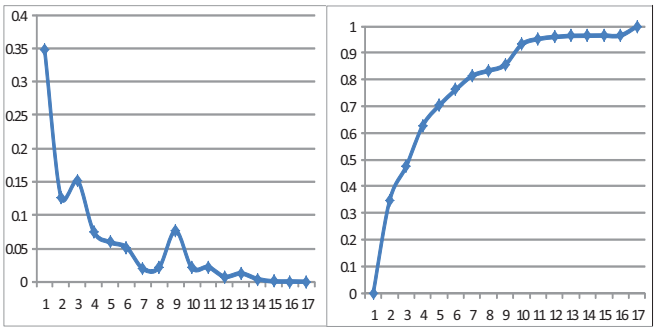

Fig. 5: (a) PDF and (b) CDF of the number of occurrences of the routes with the same AS-specific attributes.

BGP-FRP increases only the transmission overhead of the iBGP communication, while it maintains overhead of eBGP the same unlike protocols proposed in [5], [6].

The BGP protocol on Internet routers usually runs on several processes, each handling specific tasks: BGP connection establishment (BGP Open process), BGP message handling (BGP I/O), route announcement and best-path calculation (BGP Router) and routing table scanning (BGP Scanner). BGP processes directly affected by our modification are BGP Router and BGP I/O. The complexity of the BGP Router process increases linearly with the number of BGP routes for 100 routes it consumes approximately 100ms of CPU run-time and for 1000 routes it consumes1000ms of CPU run-time [11]. In stable network conditions where a new route arrives every half-second (500ms), the BGP Router process will have to process 17 routes for 17ms every 500 milliseconds, which is still a reasonably low workload. The BGP I/O process shows a little impact on the average CPU utilization as the number of routes increases. It consumes less than 100ms of the CPU run-time if there are 100 or 1000 routes, so we do not expect a significant increase in CPU utilization when BGP-FRP is applied.

The routers memory is most affected by BGP-FRP. A BGP router maintains three distinct data structures for storing BGP routes: the RIB-IN data structure, one per BGP neighbor, used to store received BGP routes from neighbors; the LOC-RIB data structure, used to store local best routes; and the RIB-OUT data structure, one per BGP neighbor, used for storing routes to be advertised to BGP neighbors. BGP-FRP influences RIB-IN and RIB-OUT data structures, as each router still keeps one local best route in LOC-RIB for traffic forwarding. Let us assume a router with an extremely large number of neighbors, e.g. n = 100 BGP neighbors. The memory required in BGP-FRP equals the memory required in BGP multiplied by the mean value of the number of twin routes that should be advertised within AS, which is maximally m=5 in the case of the Frankfurt route server that receives 95% of the Internet routes. If we assume that each neighbor advertises a full BGP routing table (close to p = 0.5M prefixes according to recent data [12]), and that each routing entry is c = 100B, the total memory required by BGP-FRP can be calculated as:

$$BW = m(cp2n) = 50\text{GB} \quad (2)$$

However, we believe this memory requirement will not be an issue with the recent development of router hardware. BGP-FRP can incorporate different tradeoffs between complexity and performance similarly as the protocols in [5], [6]. Namely, both the complexity and the diversity will increase with the number of paths per prefix, and the particular tradeoff can be selected by varying the maximum number of paths per prefix. BGP-FRP will improve the network reliability, since it will react faster to node or link failures. BGP-FRP is expected to be responsive to failures similarly as the protocol proposed in [6], while supporting more comprehensive BGP policies within a single domain.

V. Conclusion

We modified the existing BGP route selection algorithm so that flexible BGP policies can be supported. We implemented the new BGP algorithm, named BGP-FRP, in the XORP open-source router and demonstrated how it can be used to realize policies that were not feasible before. BGP-FRP reconciles internal AS policies with the requirements of neighboring ASes and allows more efficient routing in large ASes that use route reflectors. With BGP-FRP, MED oscillations are prevented thanks to the visibility of the AS-wide preferred routes [9]. Reliability of routing is improved, since routers store multiple routes to the given destination, and they can quickly choose the next best route when the best route fails.

References

[1] [Y. Wang, M. Schapira, and J. Rexford, "Neighbor-specific BGP: more flexible routing policies while improving global stability," *2009 SIGMETRICS.*
[2] R. Zhang-Shen, Y. Wang, and J. Rexford, "Atomic routing theory: making an AS route like a single node," Dept. of Computer Science, Princeton University, NJ, Tech. Rep., July 2008.
[3] D. Walton *et al.*, "Advertisement of multiple paths in BGP," IETF draft, Dec. 2012.
[4] M. Bhatia, J. M. Halpern, and P. Jakma, "Advertising multiple NextHop routes in BGP," IETF draft, Aug. 2006.
[5] A. Manolova *et al.*, "Enhancing network performance under single link failure with AS-disjoint BGP extension," in *Proc. 2010 WESEAS.*
[6] I. Ganichev *et al.*, "YAMR: yet another multipath routing protocol," *SIGCOMM Comput. Commun. Rev.*, vol. 40, pp. 13–19, Oct. 2010.
[7] A. Cvjetić and A. Smiljanić, "Analyzing capabilities of commercial and open-source routers to implement atomic BGP," *Telfor J.*, vol. 2, 2010.
[8] Y. Rekhter, T. Li, and S. Hares, RFC 4271 A Border Gateway Protocol 4 (BGP-4), IETF, Jan. 2006.
[9] J. Uttaro *et al.*, "Best practices for advertisement of multiple paths in BGP," IETF draft, Jul. 2010.
[10] S. Agarwal, C. Chuah, S. Bhattacharyya, and C. Diot, "Impact of BGP dynamics on router CPU utilization," *PAM 2004*, vol. 3015, pp. 278–288, Apr. 2004.
[11] B. Freeman, "Observations on BGP and OSPF CPU utilization," *2009 UK Network Operators Forum.*
[12] Hurricane Electric Internet Services, http://bgp.he.net.