Analyzing Capabilities of Commercial and Open-Source Routers to Implement Atomic BGP

Aleksandar Cvjetić and Aleksandra Smiljanić

Abstract — **The paper analyzes implementations of BGP protocol on commercial and open-source routers and presents how some existing BGP extensions and routing table isolation mechanisms may be used to solve issues found in standard BGP implementation.**

Keywords — **Autonomous system, BGP protocol, BGP polices.**

I. Introduction

INTERNET is a collection of various Autonomous Systems (ASs) and interconnections between them used to exchange IP network prefixes. ASs differ in their ranking on the Internet so an arbitrary AS may have customer, peer or upstream provider ASs as its neighbors. Ranking of an AS is mainly determined by its network scope and interconnections to the rest of the Internet, although some economical factors may influence its position (peering contracts with neighboring ASs, etc.).

ASs exchange network prefix reachability information using BGP protocol. ASs exchange this information according to the local objectives and peering contracts with neighboring ASs. Objectives of an AS are mainly related to load balancing and security aspects of routing, and both can be realized by using BGP policies [1]. If, for example, more than one physical connection exists between two adjacent ASs, AS controls incoming traffic by using BGP policy based on the MED attribute. MED (Multi-Exit Discriminator) is the standard BGP attribute used to announce preferred route to a destination prefix [2]. In this case, AS applies export BGP policy to one of its connections to the neighboring AS, so that all routes (or some of the routes) advertised through this connection get the lower value for MED. Neighboring AS prefers routes with the lower value for MED and forwards the traffic using this preferred route.

Peering contracts between ASs define the routes to be exchanged between them. Usually, all routes should be advertised only to the customer ASs because customers pay for the Internet resources. Peer and provider ASs do not pay for the resources, so they should be prohibited to communicate their routes through the given AS.

In order to carry out requirements imposed by peering contracts, local AS applies export BGP polices to filter routes that should not be advertised to particular ASs. .

Existing implementations of the BGP protocol show some limitations of the route selection and advertisements, influencing BGP applications. Not all AS's requirements (objectives and peering contracts) can be realized at the same time, and some of them may be in conflict also [3].

Section 2 of this paper presents two examples giving a brief introduction to "Atomic Routing Theory" (ART) [3], which proposes some changes to existing BGP protocol implementations in order to provide an AS with capability to realize all its requirements. In Section 3, by using the same examples, we present how certain principles defined by ART may be implemented by mechanisms already available on some commercial routers, particularly Huawei, Cisco and Juniper. In Section 4 we present capabilities of several open-source routers (Quagga, Vyatta and Xorp) to implement the same principles.

II. Atomic Routing Theory (ART)

Upon receiving a set of routes for a destination prefix, an AS border router (ASBR) applies BGP decision process to compare route attributes and choose one route with the best attributes. The best route is then advertised to all neighboring routers except to the router that advertised the route. Before advertising the route, ASBR may change the route attributes or even filter the route for some neighbors using the export BGP policies.

It is a common case for two neighboring ASs to agree on consistency in route advertisement. A consistent route advertisement means that, if more than one peering point exists between neighboring ASs, routes must be advertised equally at all peering points with no route changes applied to one peering point. In addition, customer ASs often require the possibility to control utilization of their route by using standard (e.g. MED) or extended (e.g. communities) BGP attributes, so AS should provide its customers with this possibility. There are examples in which above requirements cannot be realized entirely with standard BGP implementations [3].

Fig. 1 depicts an example of connections between adjacent ASs on the Internet. Suppose AS 100 and AS 200 are peer ASs, while AS 300 and AS 400 are customer ASs of AS 100. AS 500 is the upstream provider for AS 100, i.e. AS that provides connectivity to the rest of the Internet.

Aleksandar Cvjetić, Bilećka 5, Belgrade, Serbia, (e-mail: aleksandar.cvjetić@gmail.com).

Aleksandra Smiljanić, School of Electrical Engineering, Belgrade, Serbia, (e-mail: aleksandra@etf.bg.ac.rs).

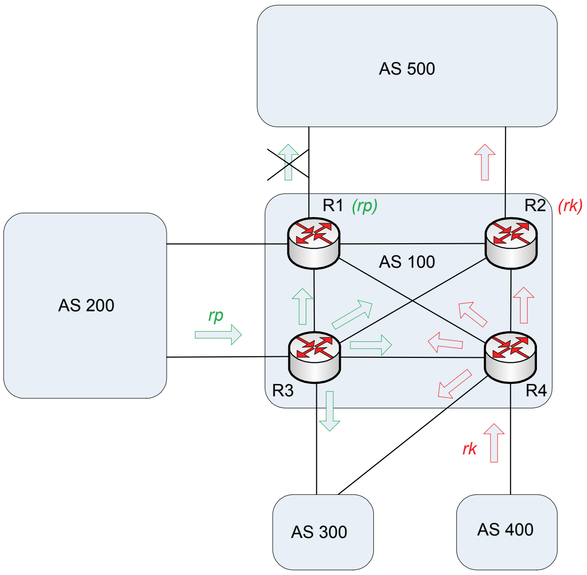

Fig. 1. An example of violating consistent route advertisement.

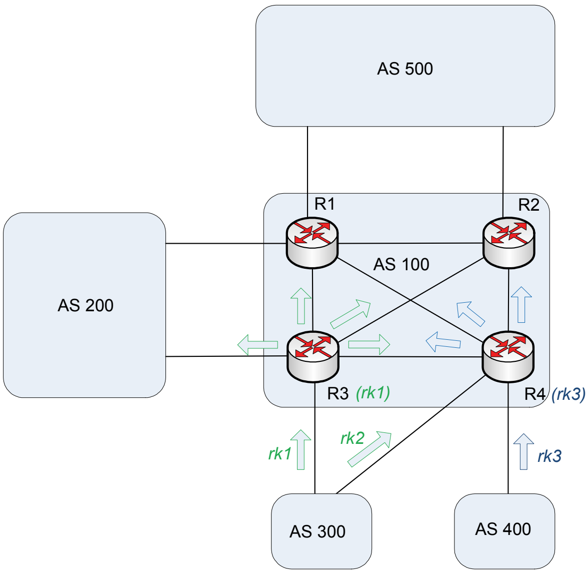

Fig 2. An example of ignoring customer's requirement for preferred route.

Suppose AS 100 and AS 500 agreed to have consistent route advertisements. In the process of route advertisement, ASBRs R1 and R2 receive two routes, *rp* and *rk*, for the same destination prefix *d*. If two routes are the same in the first four steps of BGP decision process (1. highest local preference, 2. shortest AS path, 3. lowest origin type and 4. lowest MED[1]), in the fifth step, routers choose a route with the nearest exit point from the local AS according to the internal link costs. Suppose that the router R1 chooses route *rp* while R2 chooses route *rk*. If the route *rp* was previously advertised by peer AS (AS 200) and route *rk* by customer AS (AS 400), AS 100 should apply BGP export policy so that the router R1 filters route advertisement for *rp* toward provider's AS 500, in order to prevent non-customer ASs to communicate their routes over AS 100 and use the network resources for traffic forwarding.

Applying this export policy, AS 100 realizes its objective to prevent local resources to be used by neighboring ASs which do not pay for them. At the same time, router R2 advertises route *rk* to the provider's AS 500, because this route was previously advertised by the customer AS 300. Thus, routes available for the same destination prefix will not be advertised at all peering points between AS 100 and AS 500, so the consistent route advertising is violated. This happens because the BGP decision process and the BGP export policy are independent activities.

Examine now the case in Fig. 2 presenting two customer ASs advertising routes for the same destination prefix *l* to AS 100. AS 300 advertises route *rk2* with the lower value for MED comparing to *rk1*, so that the incoming traffic would use the route *rk2* (possibly because of the link capacity, load balancing etc.).

Suppose that the route *rk2* advertised by AS 300 and route *rk3* advertised by AS 400 are the same in the first three steps of the BGP decision process at the router R4. In the fourth step, R4 does not compare MED attribute for two routes because they came from different ASs, so the decision process continues by selecting the route with the nearest exit point. Finally, router R4 chooses the route *rk3* because it has the nearest exit point and advertises it within AS 100. At the same time, router R3 chooses to advertise route *rk1* as its best route to destination *l*. As a result, route *rk2*, which is the preferred route for the incoming traffic of customer AS 300, will not be advertised within AS 100 and used for subsequent traffic forwarding.

Considering issues found in the previous examples, ART suggests to improve BGP protocol in such a way that: 1. ASBRs have possibility to choose and advertise different routes to different types of neighboring ASs, which may require several BGP processes; 2. route filtering and modification should be done before route selection; 3. among the routes that are left after MED comparison (step four), ASBR needs to disseminate at least one route within AS, and at least one route for each next-hop AS that uses MED [3].

Following the first ART principle, routers R1 and R2 will have possibility to use a separate decision process for provider AS, so that routes which should not be advertised to it (like route *rp*) are not considered in the decision process. Route *rp*, however, must be filtered before the route selection (according to the second ART principle), so that the decision process would consider only the routes that are suitable for advertisement. Now, if there is a possibility to use a separate decision process per neighbor type, and to filter and modify all routes before the route selection, export policies are not needed [3].

Following the third ART principle, router R4 will have the possibility to advertise both routes (*rk2* and *rk3*) for the same destination prefix *l* within AS 100, i.e. one route for each next-hop AS which uses MED attribute. In this

[1] According to BGP standard (RFC 1771), MED attribute is compared only among the routes advertised by the same neighboring AS. If routes come from different ASs, router skips this step.

way, AS 100 fulfills customer's requirement to advertise its preferred route for the incoming traffic, so that this route should be used to forward traffic to AS 300.

Improved BGP protocol, which provides more flexible BGP policy implementations, has the possibility to fulfill all peering requirements, and it is called atomic [3].

III. Analyzing Capabilities of Commercial Routers to Implement ART

In this Section we present capabilities of the commercial routers provided by Huawei, Cisco and Juniper to implement principles defined by ART. For this purpose, we use commands available at command line interfaces (CLI) of three vendors. We use the topology in Fig. 1, with an assumption that ASBR routers R1, R2, R3 and R4 are not directly connected to each other, but through other routers that use Internal Gateway Protocol (IGP) like OSPF or IS-IS to learn the AS topology. We assume that the internal BGP connections (iBGP) are established between ASBRs of AS 100[2], and external BGP connections (eBGP) are established between ASBRs of different Ass.

A. Huawei CLI

In order to present capabilities of Huawei routers to implement ART principles, we use CLI of *Quidway* NE80E router with software version V300R003 [4]. From the first two principles, BGP process on Huawei routers R1 and R2 should advertise only customer routes to provider's AS 500, where all other routes must be filtered before the route selection. For this purpose, we use mechanisms available for BGP/MPLS VPN implementations [5]. On each ASBR we have configured one VPN (or VRF) instance (Virtual Routing and Forwarding instance) for each type of neighboring AS to which ASBR has an external BGP connection. The VPN instances should contain only received routes and routes that should be advertised to the corresponding neighboring ASs. VPN instance is a virtual routing instance of a router with its own routing and forwarding table, corresponding routing processes and interfaces. Hence, on router R1 we have configured two VPN instances, one for peer AS 200 and another for provider's AS 500, on router R2 one for provider's AS 500, etc. The sample configuration of VPN instance for provider's AS 500 (we call it "*up_ as*") on router R1 is:

```
[R1]ip vpn-instance up_as
[R1-vpn-instance-up_as]route-distinguisher 500:1
[R1-vpn-instance-up_as]vpn-target 500:1 export-
extcommunity
[R1-vpn-instance-up_as]vpn-target 300:1 import-
extcommunity
[R1-vpn-instance-up_as]vpn-target 400:1 import-
extcommunity
```

A corresponding interface on router R1, used for connection to provider's AS 500, should be added to this VPN instance. Assume that this interface is GigabitEthernet 1/0/1 with IP address 212.200.198.161/30, then the required configuration is:

```
[R1]interface GigabitEthernet 1/0/1
[R1-GigabitEthernet1/0/0]ip binding vpn-instance up_as
[R1-GigabitEthernet1/0/0]ip address 212.200.198.161 30
```

In the above configuration of VPN instance "*up_ as*", we have configured two parameters, RD and RT, which are appended to all routes of the VPN instance before they are advertised within AS 100. Route Distinguisher (RD) is the 64-bit parameter used to distinguish between routes of different VPN instances if overlapping address space is used [5]. Here, we assume globally unique Internet addresses, so the RD parameters do not have to be unique between VPN instances. Upon adding RD to all IPv4 routes of a VPN instance they become VPN-IPv4 routes [5]. Route Target (RT), in Huawei's implementation VPN target, is a BGP extended community attribute appended to VPN-IPv4 routes to control route redistribution between different VPN instances [6]. *Export-extcommunity* is the export RT, i.e. RT used to export routes from VPN instances to all other VPN instances within the same AS. For "*up_as*" we have configured export RT with value 500:1. *Import-extcommunity* (i.e. import RT) is used to control which routes are allowed to be imported in the VPN instance. In order to enable route redistribution between two VPN instances, export RT of a VPN instance must correspond to at least one of the import RTs of another VPN instance and vice versa [5]. For "*up_as*" we have configured two import RTs, 300:1 and 400:1, to enable importing routes of both customer ASs.

A similar configuration to the one above should be applied to other VPN instances on all ASBRs of AS 100. We have configured two VPN instances on R4, one for AS 300 and another for AS 400, with export RTs 300:1 and 400:1 respectively, while import RTs must provide ability to import all routes (because all routes should be advertised to customer ASs). A VPN instance on router R3 for customer AS 300 is configured with export RT 300:1. In addition, a VPN instance for peer AS 200 is configured on both R1 and R3 with export RT 200:1, while import RTs are 300:1 and 400:1 to enable importing customer routes.

As we mentioned before, routes exchanged between VPN instances are VPN-IPv4 routes. These routes are to be exchanged using Multiprotocol BGP (MP-BGP), a BGP protocol extension for advertising routes of address families other than IPv4 (VPN-IPv4, IPv6, etc.) [7]. For example, VPN instance "*up_as*" on router R1 is configured using MP-BGP to exchange IPv4 routes with the ASBR of provider's AS 500, and VPN-IPv4 routes

[2] Generally, iBGP connections are established between all routers of an AS or between client routers and route reflectors. For more details, refer to RFC 1771.

with all other ASBRs within AS 100[3]:

```
[R1]bgp 100
[R1-bgp]ipv4-family vpn-instance up_as
[R1-bgp-up_as]peer <ip address of ASBR in AS 500> as-number 500
[R1-bgp]ipv4-family vpnv4
[R1-bgp-af-vpnv4]peer <ip address of ASBR in AS 100> enable
```

Number 100 in the above configuration is the number of the BGP process, and it must correspond to the number of local AS. The last command must be repeated for all ASBRs of AS 100, with their corresponding IP addresses used for BGP connections. MP-BGP should also be configured on other three ASBRs of AS 100 in a similar way. Finally, to enable routes of different VPN instances to be exchanged directly between ASBRs of AS 100, we must configure VPN tunnels between them. There are many ways to configure VPN tunnels and we used the most popular MPLS (Multiprotocol Label Switching) tunnels to enable BGP/MPLS VPNs within AS 100. For the sake of simplicity, we omit this part of configuration, but it should be noted that MPLS must be enabled on all routers within AS 100, as well as LDP protocol (Label Distribution Protocol) used for exchange MPLS label mapping information [8].

When router R1 receives VPN-IPv4 routes from other ASBRs of AS 100, it checks whether export RTs of those routes match at least one of the import RTs configured for its VPN instances. If so, corresponding routes are imported into VPN instance for which matching applies.

According to the described configuration, we allow only customer routes to be imported into the VPN instance "*up_as*" on router R1. After importing customer routes, the BGP decision process is applied and the best routes are advertised to the provider's AS 500. Similarly, customer routes on router R2 are imported into the corresponding VPN instance for the provider's AS 500 and the best routes are advertised to this neighbor. In this way, both R1 and R2 choose to advertise only customer routes to the provider's AS 500, so the consistent route advertisement is achieved.

Following the third ART principle, the Huawei's router R4 should advertise two routes (*rk2* and *rk3*) for the same destination prefix *l* within AS 100, i.e. one for each neighboring AS that uses MED. Existing BGP implementation on Huawei routers does not support multiple BGP route advertisements for the same destination prefix, so this principle cannot be implemented [4].

B. Cisco CLI

We have analyzed the capabilities of Cisco routers to implement ART principles by using *Dynamips* software[4]. In order to simulate the network environment, we use the architecture of Cisco 7200 series routers for ASBRs with software version IOS 12.4 (13b) [9].

As in the Huawei example, we use mechanisms available for BGP/MPLS VPN implementations to implement the first two principles defined by ART. Because the configuration steps are similar to those for Huawei's routers (except for some difference in command line syntax), we omit this part of configuration. Cisco routers support IP extended community lists to filter routes based on RT matching [10]. These extended community lists may be configured and applied to each BGP session separately, for inbound and outbound routes. As an example, on ASBR R1, we have configured IP extended community list 1 to filter all route advertisements with export RT 200:1 (routes from peer AS 200), and applied this community list to iBGP connection between R1 and R2, for outbound updates:

```
R1(config)# ip extcommunity-list 1 deny rt 200:1
R1(config)#router bgp 100
R1(config-router)#neighbor <ip address of R2> filter-list 1 out
```

This configuration prevents peer routes to be advertised to ASBR R2 which does not need them, because its neighbor is the provider's AS 500. Remember, in Huawei implementations unwanted routes are filtered on the receiver side based on RTs, so here we reduce the number of unnecessary advertisements through AS 100.

BGP implementation on Cisco routers does not have the capability to advertise multiple routes for the same destination prefix, so the third ART's principle cannot be implemented [11].

C. Juniper CLI

The third commercial router vendor whose equipment we have analyzed is Juniper. In particular, we have focused on the BGP implementation of T-series routers with software version JUNOS 9.6 [12]. According to the available documentation [13], the same mechanisms are provided for the implementation of ART's first two principles.

As before, the BGP protocol on Juniper routers does not support multiple route advertisements for the same destination prefix, so the third ART's principle cannot be implemented [13].

IV. Analyzing Capabilities of Open-Source Routers to Implement ART

Apart from commercial routers, we have analyzed the capabilities of several open-source routers to implement principles defined by ART. In particular, we have

[3] VPN-IPv4 routes are usually exchanged between routers of the same AS. IPv4 routes are exchanged between routers of different ASs. For more details, refer to RFC 2547.

[4] *Dynamips* is a simulator of Cisco hardware which uses real Cisco IOS operating system. For more details, refer to http://dynagen.org.

considered Quagga, Vyatta and Xorp.

Quagga supports the configuration of multiple BGP instances (i.e. BGP processes) on the same router [14]. For example, we may configure two different BGP processes on router R1:

```
bgp multiple-instance
router bgp 100
neighbor <ip address of ASBR in AS 200> remote-as 200
router bgp 1000
neighbor <ip address of ASBR in AS 500> remote-as 500
```

The winning routes are placed into the kernel's routing table, i.e. into the global routing table of the underlying operating system. However, only one route is chosen for each destination prefix, and the route is advertised to all BGP neighbors. Quagga does not support VRF instances and MP-BGP, so the first two ART principles cannot be implemented [14]. Because Quagga's BGP implementation advertises only one route to all neighbors, the third principle cannot be implemented.

Vyatta router supports the following BGP standards [15]: RFC 4271, RFC 4273, RFC 1997, RFC 3065, and RFC 2796. The list of supported standards does not include MP-BGP (RFC 2858). Also, Vyatta does not have support for the VRF configuration so that the first two principles of ART cannot be implemented. Quagga BGP's implementation advertises only one route for each destination prefix which is not sufficient for the implementation of the third ART's principle [15].

Xorp is the third open-source router that we have considered in this paper. The following BGP standards are supported by Xorp [16]: RFC 4271, RFC 3392, RFC 2545, RFC 1997, RFC 2796, RFC 3065, RFC 2439, and RFC 4893. Xorp BGP implementation does not support VPN-IPv4 route advertisements, which we used to exchange routes between different VPN instances (although routes of some other address families are supported, like IPv6). The configuration of the VRF instances is not supported, and in the Xorp BGP implementation, only one route can be advertised for each destination prefix, so none of the ART's principles can be implemented.

V. Conclusion

BGP protocol cannot always fulfill all the desired policies. ART is a theory that sets the principles of the improved BGP in which chosen policies are always consistent. In this paper, we presented how some of the ART's principles can be implemented using the existing commercial routers of three vendors. The cost of the improved BGP is the increased memory requirements of the routers, because VRF instances may contain a huge number of public Internet routes compared to the number of routes in VPN implementations, where the customers typically use the private IP address range.

In addition, we have analyzed BGP implementations of several open-source routers and recognized that they are not capable of implementing the ART's principles. We concluded that no BGP implementation considered in this paper supports multiple BGP route advertisements for the same destination prefix and this is a major obstacle towards the atomic BGP. The third ART's principle could be implemented using the BGP ADD-PATH as suggested in [17], or using some other mechanism to advertise more than one route for a destination prefix.

References

[1] M. Caesar, J. Rexford, "BGP Routing Policies in ISP networks", *IEEE Network Magazine*, November/December 2005. Available: http://cs.princeton.edu

[2] *RFC 1771 – A Border Gateway Protocol 4 (BGP-4)*, March 1995. Available: http://www.ietf.org

[3] R. Z. Shen, Y. Wang, J. Rexford, "Atomic routing theory", Technical Report, Department of Computer Science, Princeton University, July 2008. Available: http://www.cs.princeton.edu

[4] *Quidway NetEngine5000E&80E&40E Router Feature Description - IP Routing*, Product Manual, Available: http://support.huawei.com

[5] *RFC 2547 – BGP/MPLS VPNs*, March 1999. Available: http://www.ietf.org

[6] *RFC 4360 – BGP Extended Communities Attribute*, February 2006. Available: http://www.ietf.org

[7] *RFC 2858 – Multiprotocol Extensions for BGP-4*, June 2000. Available: http://www.ietf.org

[8] *RFC 3036 – LDP Specification*, January 2001. Available: http://www.ietf.org

[9] *Cisco IOS Software Major Release 12.4 Features and Hardware Support*, Product Literature, July 2006. Available: http://www.cisco.com

[10] *BGP Support for Named Extended Community Lists*, Feature Guide, August 2007. Available: http://www.cisco.com

[11] *Cisco BGP Overview*, Cisco IOS IP Routing Protocols Configuration Guides, May 2009. Available: http://www.cisco.com

[12] *Feature Guide*, Junos 9.6: Software Documentation, May 2009. Available: http://www.juniper.net

[13] *Routing Protocols Configuration Guide*, Junos 9.6: Software Documentation, May 2009. Available: http://www.juniper.net

[14] http://www.quagga.net/docs.php

[15] *BGP Reference Guide*, Reference Guides, March 2009. Available: http://www.vyatta.com

[16] *XORP User Manual*, Version 1.6, January 2009. Available: http://www.xorp.org

[17] *draft-walton-bgp-add-paths-03*, January 2009. Available: http://www.ietf.org