Relay-Based iBGP Multicasting in Software Defined Networks

Ukemeobong Bassey, Amiya Nayak
School of Electrical Engineering & Computer Science
University of Ottawa, Canada

***Abstract*—In the Internet today, learnt prefixes are forwarded within autonomous systems (ASs) over internal Border Gateway Protocol (iBGP) sessions. Existing schemes for iBGP routing include the full-mesh (FM) solution, route reflection (RR) solution and confederation. Optimal prefix routing and route diversity are the main strength of the FM solution. However, it is rarely employed in a large networks due to its poor scalability. The RR scheme solves the scalability challenge at the cost of FM optimality due to Route Reflector's partial view of the network. The concept of Software Defined Networking (SDN) entails decoupling of the control plane from the forwarding plane such that the control plane is logically centralized benefiting from an overall knowledge of the network for decision making. In this work, we propose a solution based on multicasting which employs relay nodes in the iBGP message dissemination. The relay nodes are elected and act as boundaries of multicast groups, relaying and filtering prefixes into other multicast groups. We evaluate the use of single and multiple relay nodes per multicast group. Our solution brings session management scalability and minimization of duplicate prefix announcement through elimination of peer sessions deemed unnecessary. SDN controller is employed to configure and coordinate the multicast tree.**

I. Introduction

Connectivity and message dissemination in the network is handled by routing protocols. Routing protocols in a network such as the Internet enable network devices, e.g routers to connect with each other, send messages across the network and make good decisions based on messages learnt from other routers. In a nutshell, routing among routers involves first the discovery stage. Here, the routers who are active discover and learn about the existence of other routers. After discovery, messages are exchanged and stored in the routers based on the algorithm specified by the network administrators.

Various routing protocols exist, and routers can be seen as multiprotocol elements since they are able to support quite a number of protocols. Some of the protocols include IGRP (Internet Gateway Routing Protocol), Enhanced IGRP, IS-IS (Intermediate System to Intermediate System protocol), BGP (Border Gateway Protocol), etc. Our focus is on BGP, particularly the iBGP. As per BGP specification [1], internal routers are only allowed to reflect routes learnt over external BGP sessions to its iBGP peers and vice versa. BGP deployment revolves around three schemes, namely, BGP full-mesh topology, BGP confederation and BGP route reflectors. To ensure complete message dissemination to each node and to enable all routers to make optimal routing decisions, the full-mesh scheme requires that all nodes peer with each other. The implication of this is that there is a high number of sessions at each router in the real network. For this type of network, the need to configure each router for every network change becomes burdensome. Hence, the full-mesh scheme is rarely employed in a large network. Coupled with the scalability issues, is the great number of duplicates which are mostly unnecessary as stated by the authors in [2]. These duplicates require large RIB size and CPU capacity. The RR scheme brings scalability improvement in that, peers are not required to establish BGP sessions with all other peers.

The RR scheme tackles these challenges by selecting specific nodes to serve as route reflection points. As a result, only the route reflectors are required to have a fully meshed iBGP sessions with each other. The RR nodes have client and non-clients. It reflects route advertised by its clients to both clients and non-client. In this way, the need for full mesh peering is eliminated. Studies in [3] and [4] have shown that the RR while solving full-mesh scheme challenges can also lead to unnecessary filtering of required prefixes by the route reflectors. One of the reasons for this is due to the fact that the route reflectors make their route selection decisions based on their partial knowledge of the network topology. Although problem detection and checking for correctness in route reflector iBGP graphs have been discussed in [5], [6], [7], it is still a point of concern.

A recently proposed scheme in [8] employs multicasting for redistribution of BGP prefixes. Although this work brings an improvement to BGP message dissemination, it faces the challenge of unreliability in BGP prefix distribution and relatively high number of established BGP sessions. Also, the dissemination of unnecessary prefix duplicates may still occur. We target to limit the BGP sessions and BGP prefix duplication which may occur. The author in [2] concludes that duplicate announcement is a significant contributor to network churn. They show that these duplicates exert a lot of burden on the CPU through the generated churn. Investigation carried out by [2], [9], [10] and others indicate the impact of churn in the network. Another study in [11] shows that BGP consumes about or over 60 percent of core routers CPU cycles and high CPU utilization can disrupt other network protocol task. Following the aforementioned cases, the objective of this paper is to minimize prefix duplicates by eliminating unnecessary peering sessions in the multicast tree coupled by filtering performed at selected nodes termed "relay nodes".

We propose a multicast relay-based algorithm for BGP

prefix dissemination coordinated by the SDN controller. All border routers are heads of multicast groups while other routers join the group of its closest border router based on IGP cost. Relay nodes are elected per group and routers are further able to join other groups through the elected relay node.

This paper is organized as follows. In Section II, we give an overview of existing routing schemes developed for iBGP prefix dissemination. Our proposed algorithms are explained and analysed through simulation in Sections III and IV. We conclude the paper in Section V.

II. Related Work

Lots of routing schemes have been proposed for iBGP prefix dissemination. The RR scheme is mostly employed in larger ASs. Moreover, to further optimize its prefix dissemination, improvement have been proposed such as the work in [12]. Specifically, authors in [12] increase routing diversity by adding more iBGP sessions to the already existing RR topology. This algorithm relies on the IGP topology of the network, eBGP route received at the border routers, iBGP RR network and the concept of external best route [13]. Some caveat in [12] include convergence issue and the need for network operators to implement tie-break rules to select a next-hop from a pool of available next-hop routers. Another work in [14] thrives to overcome the routing anomalies of RR scheme. The main idea of this work is first to calculate the IGP path cost between each node and ASBRs within an AS and then the optimal next-hop is picked from this output. However, the scheme in [14] is prone to failure due to lack of redundancy scheme in its architecture. Still on routing diversity, authors in [15] propose optimal path selection modes to enhance iBGP prefix dissemination. The Add-All-Paths [15] mode proposes that a peer advertises all received routes to its iBGP peers. The implication of this scheme is that each router has to store all available routes, leading to large memory consumption. The Add-N-Paths [15] mode suggest that a subsection of N paths should be advertised to iBGP peers. The problems with this scheme lies in its lack of optimal routing guarantee and avoidance of multi-exit discriminators. Additionally, the work in [5] aims to optimize iBGP prefix routing in RR networks by selection of optimal egress points. The proposed model in [5] finds the optimal BGP route by checking for the validation conditions.

Other proposed prefix dissemination schemes include the work in [16] which employs shortest path routing. In [16], nodes are separated into black and white. A black node is seen as a node that blocks other nodes from getting to their optimal egress point, while white nodes are always going to learn prefixes advertised by their optimal egress point regardless of quasi-equivalent conditions. Authors in [16] prove that any node belonging to the shortest IGP path is a white node and by extension a non-blocking node, implying that nodes will always learn prefixes from their optimal egress point provided there is a white node between them. Caveats of this scheme includes the computation of path cost from each router to prospective egress point candidate. Cluster-based iBGP prefix dissemination is presented in [17]. This work proposes a scheme leveraging a cluster of routers architecture [17] for processing and storage of routes to increase scalability. Reference [17] employs the balanced neighbor assignment [17] to distribute routers into clusters. Even though, this scheme eliminates single point failure, increases the reliability and processing speed, the potential complexity of implementing the additional peering protocols is a point of concern, given that there is no standard protocol for its implementation yet.

Finally, authors in [8] propose multicast-based routing in iBGP prefix dissemination. It follows that BGP routing aims at redistributing prefixes from the ASBR, a single source to the iBGP peers within the AS, multiple destinations. Hence [8] proposes that ASBRs should be registered as group heads, each with its own unique group address. The internal routers have the responsibility of joining any group of preference.

III. Relay-Based iBGP Multicasting

In this section, we present our proposed relay-based multicast prefix dissemination. To limit the duplication, a basic idea is to establish BGP sessions between roots (ASBR) of all multicast groups in mesh so that each root is able to eliminate already advertised routes from getting to its members. This is inefficient given that there might be a more cost-effective path from other tree members to the adjacent multicast group.

In our scheme, all nodes are permitted to join at most one group at initialization. For our purpose, we define an origin group as the multicast group any node joins at the initialization stage, while target groups refer to other groups apart from the origin group that a node has interest in becoming a member. Our objectives are achieved by electing a group member of the origin group closest to the target group to act as a relay router. This way, prefixes advertised by the target group are disseminated by relay nodes into respective origin groups. Also, already existing group sessions may be utilized. We explain the process in phases.

A. Phase 1- Initialization

1) For each internal node, construct the shortest path to all ASBRs, that is multicast roots. Nodes become members of one and only one ASBR group which is the minimum of the shortest path in the constructed graph. A multicast tree is formed by ASBR as root and nodes of which it qualified as the minimum shortest path cost. Peer sessions are established between all directly connected nodes in the multicast tree.
2) At the end of this phase, all nodes belong to only one multicast group. This group is the node's "origin group", while all other groups except the origin belong to the set of "target groups".

B. Phase 2 - Relay-node Selection

Relay-node selection process is performed per group since we consider only one relay node per multicast group. For each group, say G_1, a node is elected to act as relay node between

the origin and target groups. In this case, G_1 is the origin group, while $G_2...G_n$ refers to the target groups.

1) Elect a group member of G_1 which has the minimum cumulative shortest path to the roots of the target groups by iterating through all the group members of G_1. In our algorithm, the cumulative shortest path `cum_sp` is the sum of the cost from each node to all ASBRs which belong to the set of the target group, it is given in the equation below, where G_i is a target group.

$$\texttt{cum\_sp} = \sum_{i=1}^{K} sp(node, G_i)$$

2) The elected node acts as a relay router between the group pairs, transmitting BGP prefixes from target groups into its original group G_1, and filtering duplicate prefixes already advertised by its own ASBR or from another border router with a preferred cost.

C. Relay-Based Multicast Algorithm

TABLE I
SYMBOLS AND MEANING USED IN ALGORITHMS AND EQUATION

Table of Notation	
Symbol	Meaning
$A_{routers}$	Set of all ASBRs
N_{nodes}	Set of all network nodes excluding ASBRs
$sp(A_j, N_i)$	Shortest path between ASBR, A_j, and node, N_i
M_{A_j}	Multicast group of ASBR A_j
R_{A_j}	Relay node of multicast group of A_j
`cum_sp`	Cumulative shortest path
$r_{potential}$	Set of nodes eligible for relay node election

Following the explained algorithm as in Algorithm 1, the network sessions scenario is depicted in Figures 1 and 2 with nodes C, D and B, E as relay nodes respectively. The full-mesh sessions for the same network is depicted in Figure 3, the RR scheme sessions in Figure 4 where rr1 and rr2 are the route reflectors.

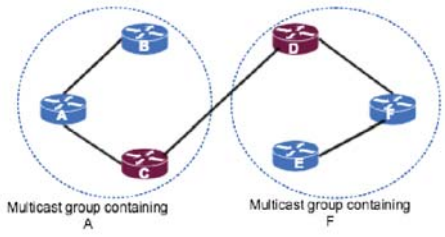

Fig. 1. iBGP sessions in relay-based Multicast scheme with nodes C and D as relay nodes

D. Multi-Node Relay-Based Multicast Scenario

In the event of rapid increase in BGP prefixes advertised by the target groups, without any increase in the bandwidth and memory of elected group relay node, overloading and ultimately crashing may occur. This algorithm proposes the use of a set of relay routers per multicast group. We give the

Algorithm 1 SDN Controller Multicast Computation and Relay Node Selection

Input: $A_{routers} \cup N_{nodes}$: Set of all network nodes
Output: Relay node of each multicast group

```
Procedure_initialization():
Init multicast group for all A_routers
for N_i in N_nodes do:
    for A_j in A_routers do:
        Calculate sp(A_j, N_i)
        //Selection of minimum shortest path
        if sp(A_j, N_i) < sp(A_{j-1}, N_i) then
            add N_i to group M_{A_j}
        end if
    end for
end for
Procedure_RelayNode_selection():
A'_j refers to set of all ASBR excluding A_j
Set cum_sp(A_j, A'_j) = ∞
for node in M_{A_j} do
    Calculate cum_sp = sp(node, A'_j)
    if cum_sp < cum_sp(A_j, A'_j) then
        cum_sp(A_j, A'_j) = cum_sp
        Select node as relay node
    end if
end for
```

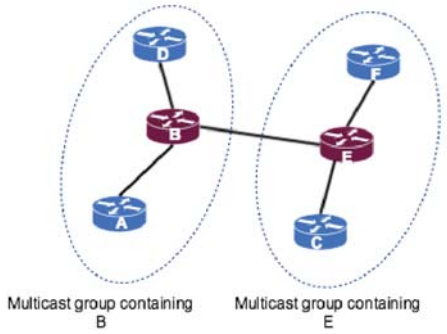

Fig. 2. iBGP sessions in relay-based Multicast scheme with nodes B and E as relay nodes

conditions for nodes eligibility for relay node election. The algorithm introduces a minimal number of iBGP sessions to be added to the AS. We achieve route diversity, evade single-point failure, decrease the risk of vulnerability to single-point attack, and increase reliability.

The working principle of the algorithm is divided into two parts as in the previous algorithm, the initialization phase remains same, where every node joining the network calculates its cost to all available ASBR and then joins the multicast group of which the cost is minimum. However, in the relay node selection phase, the first round involves the selection of a node with minimum cumulative cost to target groups as a

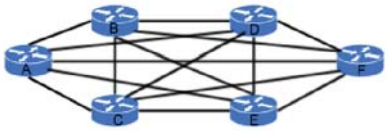

Fig. 3. iBGP Sessions in Full-Mesh Scheme

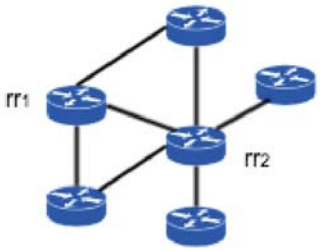

Fig. 4. iBGP Sessions in Route Reflector Scheme

member of the relay node set. The second round of selection compares the cumulative cost of other nodes in the group to that of the selected relay node. A node is selected to join the relay node set, if and only if, it minimizes some of the highest cost of the already selected relay node. Algorithm 2 shows the process.

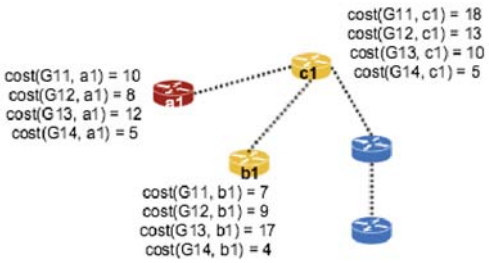

Fig. 5. Applying the relay node algorithm, node set = 2

Take the case of Figure 5 for example, where the cumulative cost of a_1 is 35, b_1 is 37 and c_1 is 46. In the first round of relay node election, node a_1 was selected because it had the least cumulative cost, 35. In the second round, both nodes b_1 and c_1 are eligible for relay node positions. However, node b_1 is selected since it has the next minimum cost to some of the target groups. After the selection, target group re-assignment occurs, and a_1 will be assigned target group G12 and G13 while b_1 is assigned G11 and G14 since it minimizes the cost for these groups.

Algorithm 2 Algorithm for Selection of 2 Relay-Nodes per Multicast Group - SDN Controller

// A'_j refers to all target multicast group
`Compare_Cost_Minimization`($r_{potential}$, r_{relay}):
for A_j* in A'_j **do**:
  if $sp(r_{potential}, A_j*) < sp(r_{relay}, A_j*)$ **then**
    calculate minimization cost
  end if
end for
`Procedure_RelayNode_Selection()`:
Init relay node, R_{A_j} for all $A_{routers}$
Set `cum_sp`$(A_j, A'_j) = \infty$
for node in M_{A_j} **do**
  //cumulative cost between each node and target group
  `cum_sp` $= sp(node, A'_j)$
  if `cum_sp` $<$ `cum_sp`(A_j, A'_j) **then**
    `cum_sp`(A_j, A'_j) = `cum_sp`
    select *node* as relay node
  end if
end for
// r_{relay} is relay node selected at first round
R_{A_j} = [r_{relay}] //relay node set
//minimization cost is mc
mc = 0
for r_n in $r_{potential}$ **do**:
  ccm=`Compare_Cost_Minimization`(r_n, r_{relay})
  if ccm>mc of all other $r_{potential}$ **then**
    mc = ccm
    add r_n to R_{A_j}
  end if
end for

IV. Simulation and Evaluation of Proposed iBGP Multicast Algorithm

We present the evaluation of our proposed algorithm in this section. In the simulation, we assume that every BGP speaker is able to re-advertise BGP prefixes. Also, the decision to advertise a prefix or not solely lies with the BGP speaker. Our network is implemented using pyretic [18] SDN programming language with POX controller. Computation of multicast tree and assignment of dynamic multicast addresses are handled by the POX controller, which inturn writes policies to our network nodes ensuring that prefixes are disseminated following the computed multicast tree. In the case of network changes resulting from node leaving or joining the network, the POX controller notifies other network nodes and appropriate actions follow, whether it is deleting of multicast group, withdrawing of advertised routes or re-computation of multicast trees.

In our simulation, we use MiniNExT to integrate Quagga [19] into the virtual environment of mininet. We employ ExaBGP in injecting BGP routes into the network. First, we study the number of BGP sessions generated by the full-mesh, route reflector scheme and the proposed relay-based scheme.

Figures 6 and 7 depict the number of sessions in small and medium sized network respectively. We observed that both the route reflector and the proposed relay-based multicast schemes have lesser number of sessions established in comparison to the full-mesh topology. This is due to the fact that in full mesh topology all BGP peers created a session with all other nodes in the network. However, our topology established iBGP sessions following the created multicast tree.

We further considered a base network of 10 BGP speakers including 2 ASBRs, therefore 2 multicast groups according to our algorithm. Each BGP peer is a running Quagga instance. When a router comes alive in the network, if it is an ASBR then the controller assigns to it a multicast address, else the peer joins one multicast group of which it is closest to the ASBR given the IGP cost. Every multicast group has one relay router. Routers are elected as relay router if the previous relay router fails or when a new group is formed as explained in Algorithm 1. Once a peer fails or leaves the network, all advertised routes are withdrawn and if the peer was an ASBR, the multicast group is removed. The simulation was run for 30 times and data retrieved from a BGP peer ("p3") of the multicast group. The number of BGP prefixes advertised range from 22 to 65 BGP routes. As shown in Figure 8, we achieve about 60% improvement with respect to the total number of messages sent within the AS compared to the FM scheme. The decrease in total BGP messages is due to appointment of relay nodes and the reduced number of sessions at the BGP peer. Moreover, with respect to the RIB size of "p3" as in Figure 9, we achieved a 10% decrease compared to FM scheme coupled with a relatively smaller improvement compared to the RR scheme. This decrease is significant when viewed cumulatively from the perspective of the entire network.

Additionally, Figure 10 depicts the memory consumption of the three schemes. The decrease in memory consumption in our relay-based multicast scheme is due to the restriction of unnecessary duplicate prefix dissemination through employment of elected relay nodes to filter and store the duplicate route collectively.

Finally, we note that one of the cost associated with our approach is the convergence time and computation as in Figure 11. Our scheme requires that a multicast tree is built and that advertised routes travel through the multicast tree to group members. Network changes have to travel down the multicast path, resulting in about 30% convergence time increase compared to FM scheme and 17% increase compared to the RR scheme.

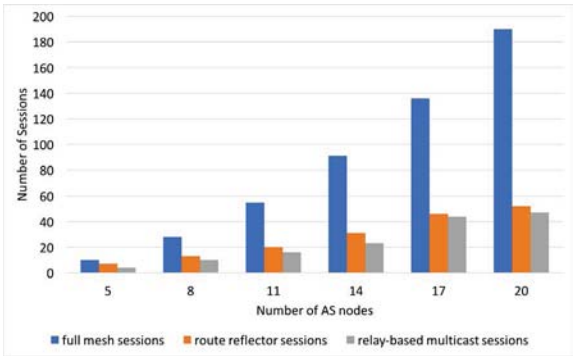

Fig. 6. Small Topology BGP Sessions

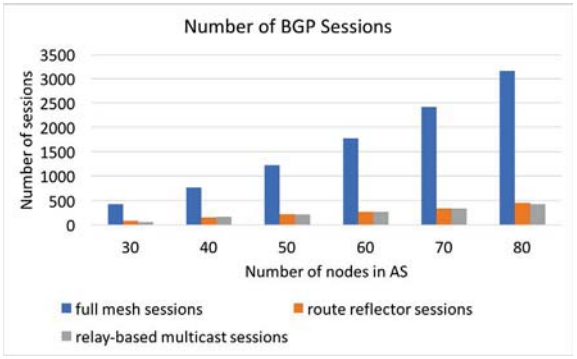

Fig. 7. Number of BGP sessions

V. Conclusion

In this work, we started by pointing out the duplicate problem in multicast iBGP prefix advertisement in the Internet. First, we proposed a relay-based multicast BGP message dissemination involving a two-phase process. The first phase describes the initialization process and the second phase describes the relay node selection process. In this scheme, we employ one relay node per multicast group. SDN controller is employed to coordinate the multicast tree and policies of the network.

Furthermore, we show by simulation that compared to FM and RR schemes, our algorithm is capable of achieving a decrease in the total number of iBGP sessions established in the AS. We also achieve a 60% decrease in the total messages compared to FM scheme, and a reasonable improvement in memory consumption and RIB size of the internal BGP peer. Generally, our performance is very close to that of RR scheme.

References

[1] Y. Rekhter, T. Li, and S. Hares, "IETF RFC 4271 a Border Gateway Protocol (BGP-4)," in *Reston: Internet Society*, 2006.

[2] A. Elmokashfi, A. Kvalbein, and C. Dovrolis, "BGP Churn Evolution: A Perspective from the Core," in *IEEE INFOCOM*, 2010, pp. 1–9.

[3] L. Xiao, J. Wang, and K. Nahrstedt, "Optimizing IBGP Route Reflection Network," in *7th ACM Conference on Embedded Networked Sensor Systems*, 2003, pp. 1765–1769.

[4] J. Park, R. Oliveira, A. Shane, M. Danny, and Z. Lixia, "BGP Route Reflection Revisited," *IEEE Communications Magazine, Vol. 50, No. 7*, pp. 70–75, 2012.

[5] M. Buob, M. Meulle, and S. Uhlig, "Checking for Optimal Egress Points in iBGP Routing," in *6th Int. Workshop on Design and Reliable Communication Networks*, 2007, pp. 1–8.

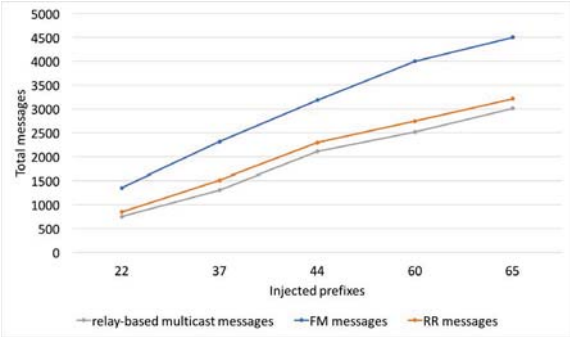

Fig. 8. BGP received messages

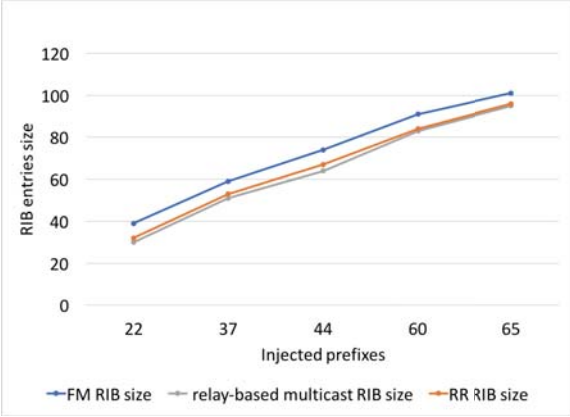

Fig. 9. Routing Infomation Base Size

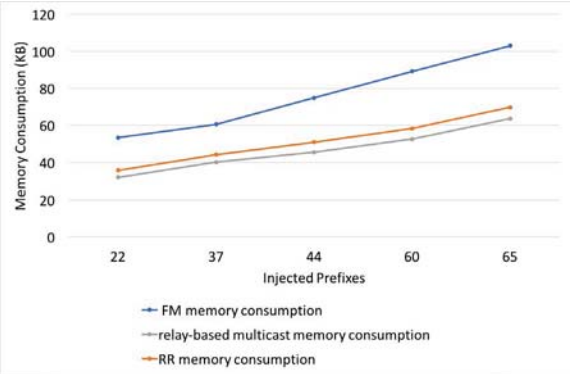

Fig. 10. BGP Peer Memory Consumption

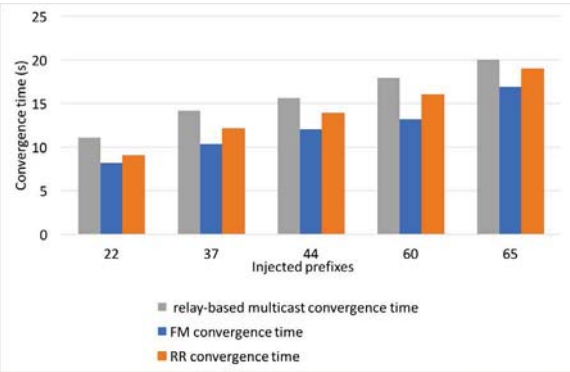

Fig. 11. BGP convergence time

[6] I. Juniper Networks, "Configuring BGP routing - Advertising Routes: BGP advertise–best–external–to–internal," 2007. http://www.juniper.net/techpubs/software/erx/junose71/swconfig-bgp-mpls/html/bgp-config10.html

[7] P. Gransson and C. Black, "Software Defined Networks: A Comprehensive Approach," 2014.

[8] F. Godn, S. Colman, and E. Grampn, "Multicast BGP with SDN Control Plane," in *7th Int. Conference on the Network of the Future (NOF)s*, 2016, pp. 1–5.

[9] J. H. Park, D. Jen, M. Lad, S. Amante, D. McPherson, and L. Zhang, "Investigating Occurrence of Duplicate Updates in BGP Announcement," in *Krishnamurthy A., Plattner B. (eds) Passive and Active Measurement*, vol. 6032, 2009.

[10] G. Huston and G. Armitage, "Projecting future IPv4 Router Requirements from Trends in Dynamic BGP Behaviour," in *Australian Telecommunication Networks and Applications Conference (ATNAC)*, 2006.

[11] S. Agarwal, C. Chuah, and S. Bhattacharyya, "Impact of BGP Dynamics on Router CPU Utilization," in *5th Int. Workshop of Passive and Active Network Measurement*, 2004, pp. 278–288.

[12] C. Pelser, T. Takeda, E. Oki, and K. Shiomoto, " Improving Route Diversity through the Design of iBGP Topologies," in *IEEE Int. Conference on Communications*, 2008, pp. 5732–5738.

[13] T. G. Griffin and G. Wilfong, " On the Correctness of iBGP Configuration," in *ACM SIGCOMM*, 2002.

[14] B. Sarakbi and S. Maag, "BGP skeleton An Alternative to iBGP Route Reflection," *Int. Journal of Distributed Sensor Networks*, vol. 12, pp. 1–5, 2010.

[15] V. Schrieck, P. Francois, and O. Bonaventure, "BGP Add-Paths: The Scaling/Performance Tradeoffs," *IEEE Journal on Selected Areas in Communications, Vol. 28, No. 8*, pp. 1299–1307, 2010.

[16] M. Buob, A. Lambert, and S. Uhlig, "iBGP2: A Scalable iBGP Redistribution Mechanism Leading to Optimal Routing," in *IEEE INFOCOM*, 2016, pp. 1–9.

[17] X. Zhang, X. Lu, J. Su, B. Wang, and Z. Lu, "A Scalable,Distributed BGP Routing Protocol Implementation," in *IEEE 12th Int. Conference on High Performance Switching and Routing*, 2011, pp. 191–196.

[18] J. Reich, C. Monsanto, N. Foster, J. Rexford, and D. Walker, "Modular SDN Programming with Pyretic Proposal," in *Usenix, The Advanced Computing Systems Association*, vol. 38, 2013.

[19] K. Holter, "Wireless Extensions to OSPF: Implementation of the Overlapping Relays Proposal," in *Master Thesis, University of Oslo*, 2006.