BGP-Multipath Routing in the Internet

Jie Li, Vasileios Giotsas, Yangyang Wang, and Shi Zhou, Senior Member, IEEE

Abstract—BGP-Multipath (BGP-M) is a multipath routing technique for load balancing. Distinct from other techniques deployed at a router inside an Autonomous System (AS), BGP-M is deployed at a border router that has installed multiple inter-domain border links to a neighbor AS. It uses the equal-cost multi-path (ECMP) function of a border router to share traffic to a destination prefix on different border links. Despite recent research interests in multipath routing, there is little study on BGP-M. Here we provide the first measurement and a comprehensive analysis of BGP-M routing in the Internet. We extracted information on BGP-M from query data collected from Looking Glass (LG) servers. We revealed that BGP-M has already been extensively deployed and used in the Internet. A particular example is Hurricane Electric (AS6939), a Tier-1 network operator, which has implemented >1,000 cases of BGP-M at 69 of its border routers to prefixes in 611 of its neighbor ASes, including many hyper-giant ASes and large content providers, on both IPv4 and IPv6 Internet. We examined the distribution and operation of BGP-M. We also ran traceroute using RIPE Atlas to infer the routing paths, the schemes of traffic allocation, and the delay on border links. This study provided the state-of-the-art knowledge on BGP-M with novel insights into the unique features and the distinct advantages of BGP-M as an effective and readily available technique for load balancing.

Index Terms—Multipath routing, equal-cost multi-path (ECMP), traffic engineering, load balancing, BGP-multipath, Internet routing, looking glass, traceroute, RIPE atlas.

I. INTRODUCTION

The default setting of Border Gateway Protocol (BGP) [1] requires a single “best” path for each prefix. BGP-Multipath (BGP-M) is a technique to enable load balancing on multiple IP-level inter-domain paths of equal cost. Specifically, a network operator can activate the Equal-Cost Multi-Path (ECMP) function at a border router so that when the border router learns from a same neighbor Autonomous System (AS) multiple eBGP paths (via different border links) to a prefix with equal attributes, the border router installs all of these paths in the routing table instead of trying additional tie-breaking attributes. Routers produced by most major vendors support the ECMP function, including Juniper [2], Cisco [3], and Huawei [4]. Although there have been a number of research works on multipath routing, e.g., [5]–[9], BGP-M remains an obscure technique.

In this paper, we present the first measurement and a comprehensive analysis on the BGP-M routing in the Internet. We obtained BGP data from Looking Glass (LG) servers to infer the deployment of BGP-M, and collected traceroute data from RIPE Atlas [10] to extract further details on BGP-M routing paths, the schemes of traffic allocation, and the delay on border links. Our results showed that BGP-M has been deployed extensively in the Internet.

The techniques and results presented in this paper provide the state-of-the-art knowledge on BGP-M. We believe that our work is relevant to industry stakeholders, Internet engineers and researchers interested in Internet routing performance and security.

II. BACKGROUND

A. Border Router and Border Link

Although network operators of ASes can use various intra-domain protocols for routing within boundary of their own networks, BGP [1] is the default inter-domain protocol used universally for routing among ASes throughout the global Internet. BGP is policy-based and allows a lot of flexibility in implementing routing policies.

A border router, also called BGP border router or AS border router, is located at the boundary of an AS with at least one interface connecting to an intra-domain router and at least one interface connecting to a border router in a neighbor AS. A border router is implemented with BGP. It can establish and maintain BGP sessions to exchange routing information with other ASes via BGP messages, and then update its routing table according to the network operator’s policy configurations.

| Notation | Description |
|----------|-------------|
| SrcIP | Source IP address |
| DestIP | Destination IP address |
| DestPrf | Destination prefix |
| NearAS | Near-side AS |
| NearBR | Near-side border router |
| NearP | IP address of ingress interface of NearBR |
| FarAS | Far-side AS |
| FarBR | Far-side border router |
| FarP | IP address of ingress interface of FarBR |
| BL | Border link between ASes |

TABLE I
NOTATIONS AND DESCRIPTIONS

Manuscript received 17 July 2021; revised 24 January 2022; accepted 11 April 2022. Date of publication 24 May 2022; date of current version 12 October 2022. This work was supported by China Scholarship Council (CSC) with grant no. 201406060022. The associate editor coordinating the review of this article and approving it for publication was M. Mellia.

Jie Li was with the Department of Computer Science, University College London, London WC1E 6BT, U.K. (e-mail: jie.li@cs.ucl.ac.uk).

Vasileios Giotsas is with the School of Computing and Communications, Lancaster University, Lancaster LA1 4YW, U.K. (e-mail: v.giotsas@lancaster.ac.uk).

Yangyang Wang is with the Institute for Network Sciences and Cyberspace, Beijing National Research Center for Information Science and Technology, Tsinghua University, Beijing 100190, China (e-mail: wangyy@cernet.edu.cn).

Shi Zhou is with the Department of Computer Science, University College London, London WC1E 6BT, U.K. (e-mail: s.zhou@ucl.ac.uk).

Digital Object Identifier 10.1109/TNSM.2022.3177471
Fig. 1. Examples of Best-path routing, Multipath routing and BGP-Multipath (BGP-M) between two neighboring ASes. See Table I for description of notations. (a) Best-path routing, where a single best path is chosen for routing. (b) Multipath routing, where an intra-domain router divides traffic to a DestIP onto two different paths. If the paths merge within the same AS, they form a so-called intra-domain ‘diamond’; if they cross AS borders, they form an inter-domain diamond. (c) BGP-Multipath (BGP-M) routing, where a border router shares traffic to a DestPrfx on two inter-domain border links.

When two ASes have established BGP sessions via border routers, they peer with each other and are called two peering ASes. Two peering ASes are neighboring ASes if they are connected directly, via physical links or an IXP (for reduced latency and improved routing performance), and they are called neighbor ASes to each other. Otherwise, the two peering ASes are called remote ASes to each other (for monetary savings and increased connectivity [11]). An AS can deploy BGP-M to both neighbor ASes and remote ASes. This paper focuses on BGP-M deployed by an AS to its neighbor ASes which are more common.

A border link is a physical IP-level inter-domain link connecting border routers of two neighboring ASes. As illustrated in Figure 1(a), depending on traffic direction, a border link starts from an egress interface of a border router of the nearside AS, and ends at an ingress interface of a border router of the farside AS. Since the egress interface of a router is invisible in traceroute measurement, a border link is usually denoted by the ingress interfaces of the two border routers, i.e., NearIP on NearBR and FarIP on FarBR, which can be identified as two consecutive IP addresses on a traceroute path that are mapped to two different ASes.

### Table II

| Priority | Attribute | Best path selection rules |
|----------|-----------|---------------------------|
| 1        | LocPref   | Highest local preference  |
| 2        | AS path   | Shortest AS path          |
| 3        | Origin    | Lowest origin type (IGP < EGP < INCOMPLETE) |
| 4        | MED       | Lowest MED (Multi Exit Discriminator) |
| 5        | eBGP/IBGP | Prefer eBGP over IBGP paths |
| 6        | IGP metric| Lowest IGP metric         |
| 7        | Router ID | Lowest router ID           |

#### B. Best-Path Routing

By default, if a border router receives advertisements of different routes to a destination prefix, it should select the best path by considering a series of BGP attributes in order of their priority as shown in Table II, where Router ID is only used as a last-resort tie-breaker if all other attributes have equal values [1].

Until recently, it was expected that there should normally be a single valid IP-level routing path from a source IP address to a destination IP address (see Figure 1(a)). When multiple paths were observed, they were considered as anomalies, possibly due to routing table misconfiguration [12], link failures [13]–[15] or change of routing paths [16]–[20].

#### C. Multipath Routing

In recent years, network operators utilized a traffic engineering technology called the multipath routing, deployed at intra-domain routers within an AS, to enable multiple IP-level routing paths to a destination (see Figure 1(b)). These multiple routes are legitimate, lasting routes. They are used concurrently to balance traffic load in order to achieve improved routing performance and resilience [21], [22].

Researchers identified multipath routing from traceroute data [5]–[8], [23], [24] using Paris traceroute [25] and Multipath Detection Algorithm (MDA) [23]. The research focus was on the link and node discovery and the topological characteristics of diamonds (see Figure 1(b)) or load balancers.

### III. BGP-MULTIPATH

BGP-Multipath (BGP-M) is a load balancing technique deployed at a border router to share traffic load to a destination prefix on different border links using the ECMP function.

#### A. Equal-Cost Multi-Path (ECMP)

Routers produced by most major vendors, such as Juniper, Cisco and Huawei [2]–[4], have already supported the ECMP function. They allow routers to install multiple internal or external BGP paths, called iBGP-Multipath and eBGP-Multipath, respectively. This function is called ‘BGP-Multipath’ by Juniper and Cisco; or ‘BGP Load Balancing’ by Huawei.

#### B. Deployment of BGP-M

As shown in Figure 1(c), in order for a network operator (NearAS) to deploy BGP-M at a border router (NearBR), the
following conditions must be satisfied. (1) NearBR supports the ECMP function; (2) NearBR has multiple border links connecting to border router(s) of a same neighbor AS (FarAS), either directly or via an IXP; (3) NearBR has learned from the neighbor AS multiple routes via different border links, to a given destination prefix (DstPrfx); and (4) the multiple routes have equal values for the first 6 attributes in Table II.

If the above conditions are met (i.e., the routes learned over different paths are considered sufficiently equal), the network operator can deploy BGP-M at the border router by installing the multiple routes in the routing table such that the border router is configured to use these paths concurrently. Because all the relevant BGP attributes for the routes over different paths are the same and the border router still announces one route as the best route, there is no impact on BGP loop detection or other BGP processing [26].

Note that the multiple IP-level paths used in the deployment of BGP-M always follow the same AS-level path. We used the terms ‘Multipath BGP’ or ‘M-BGP’ in our preliminary works [27], [28]. Since then we have changed to the terms ‘BGP-Multipath’ or ‘BGP-M’ to avoid confusion with ‘Multipath BGP’ in [26], [29] and ‘Multipath BGP’ in [30] that are relevant to multiple AS-level paths.

C. Limited Documentation on BGP-M

To a large extent, BGP-M remains an obscure technique because there are only a small number of documents related to BGP-M in literature.

1) Router Vendor Documents: Major router vendors, like Juniper, Cisco and Huawei [2]–[4], provided technical documentations on the ECMP function that underlies the BGP-M.

2) RFC2992 on Analysis of an Equal-Cost Multi-Path Algorithm (2000) [35]: This Request for Comments (RFC) introduced a hash-threshold method for routers to choose a next-hop (path) from equal-cost multiple paths, which is used for the deployment of BGP-M.

3) IETF Draft on Equal-Cost Multipath Considerations for BGP (2019) [36]: This Internet-Draft by the Network Working Group of the Internet Engineering Task Force (IETF) described the application of ECMP in various scenarios. This is perhaps the most relevant document on BGP-M.

4) Research Publications: Valera et al. [26] briefly discussed the concept of BGP router using multiple equal-cost paths concurrently. Mok et al. [37] studied YouTube’s load balancing behavior on inter-domain border links. Augustin et al. [6] and Almeida et al. [9] mentioned the possibility of multipath routing based on ECMP.

D. Challenges in Discovering BGP-M From Traceroute Data

So far there is no dedicated dataset for BGP-M. In the past, traceroute data were used to study multipath routing deployed at intra-domain routers [6]–[9], where specific traceroute tools were designed and deployed and large amounts of data were collected. In theory, traceroute with UDP packets has the potential to discover BGP-M deployed at border routers, but there are a number of challenges.

One challenge is that we will need to design a traceroute probe specially customized for discovering BGP-M. Then, without any prior knowledge, we will have to deploy the traceroute probe in as many ASes as possible; and from each probe, we will have to run traceroute to as many different destination prefixes in as many other ASes as possible.

The largest challenge, however, is the lack of sound tools or datasets for IP-to-AS mapping and AS border mapping – despite more than a decade of research effort. If we use traceroute data to discover BGP-M, we must be able to accurately determine the border of an AS on a traceroute path, so that we can credibly identify border router and border links. A recent study [43] shows that existing efforts on IP-to-AS mapping and AS border mapping [31]–[34], [39]–[42] still cannot avoid erroneous results.

IV. INFERENCE OF THE DEPLOYMENT OF BGP-M FROM LOOKING GLASS (LG) DATA

Here we introduce our effort in inferring the deployment of BGP-M in the Internet based on query data from Looking Glass (LG) servers. This method has a number of advantages: it is relatively easy to conduct and analyse, it reveals a rich set of information, and most importantly, it is accurate and credible such that our result can be considered as ground-truth.

A. Notation of a BGP-M Case

Table I lists the notations used in this paper. We used a 4-tuple, <NearAS, NearBR, FarAS, DstPrfx> to denote a unique case of BGP-M deployment, or BGP-M case. We chose these 4 parameters because (1) these values can be observed and confirmed in the routing information retrieved from LG server, and (2) they defined the owner (NearAS) and location (NearBR) of a BGP-M case as well as the destination (DstPrfx) and the neighbor AS (FarAS) from which the multiple paths were learned from. A BGP-M case is valid for traffic from any source and therefore is irrelevant to SrcIP. Border links used by a BGP-M case are between NearAS and FarAS, and their FarIPs are listed in the response from LG server.1

A NearAS can deploy BGP-M at different NearBRs for the same DstPrfx; or at the same NearBR for different DstPrfxes. All these are considered as different cases of BGP-M deployment as they have different tuples.

For convenience, when we studied a BGP-M case, we only considered traffic routing between two neighboring ASes, i.e., traffic started in NearAS and ended in FarAS. In the real Internet, SrcIP can be outside NearAS and DstPrfx can be outside FarAS – indeed they can be anywhere on the Internet as long as the traffic for the DstPrfx traverses through NearAS and FarAS via NearBR in a BGP-M case. If NearAS and FarAS are indirectly connected at an IXP, the BGP-M tuple does not need to include the IXP because IXPs are ‘transparent’ in BGP routing (and usually considered as a part of the FarAS) [38]. In other words, the existence of IXP does not affect the function and the deployment of BGP-M.

1As we will show later, border links may change from time to time.
TABLE III

LG SERVERS IN THE TOP-200 ASeses AS RANKED BY CAIDA [56]

| # of ASes in each group in Top-200 | 1–10 | 11–50 | 51–200 | Total |
|----------------------------------|------|-------|--------|-------|
| With known LG URL                | 10   | 31    | 82     | 123   |
| With accessible URL              | 8    | 20    | 62     | 90    |
| Support routes command           | 4    | 11    | 36     | 51    |
| With identified BGP-M cases      | 1    | 3     | 0      | 4     |

B. Our Inference Method

1) LG Servers: A Looking Glass (LG) server provides Web-based interfaces at one or more border routers to allow non-privileged execution of network commands (e.g., traceroute, ping, and BGP). These commands provide direct access to the BGP configuration and routing tables of border routers beyond what is propagated through BGP updates collected by RouteViews [44] and RIPE RIS [45]. LG server data have been used to study the Internet topology and path diversity [46]–[50]. Recently the Periscope platform [51] was proposed to unify LG servers with publicly accessible querying API and to support on-demand measurements.

We proposed to infer the deployment of BGP-M by querying LG servers because they can provide non-transitive BGP attributes containing direct and reliable information on the deployment of BGP-M.

2) List of ASes With LG Servers: In our work, we firstly compiled a list of ASes with LG servers from a number of data sources including BGP Looking Glass Database [52], PeeringDB [53], [54], and traceroute.org [55]. The list contained 2,709 AS numbers (ASNs). Table III lists the number of ASes with LG servers in the Top 200 ASes as ranked by CAIDA [56], where in total 51 ASes had accessible LG servers and supported the routes command (e.g., show ip bgp routes detail <IP address>) which is needed for our inference [27].

For clarity, in the following we call a border router to which we send LG enquiry a ‘nearside border router’ (NearBR); and an AS that owns and manages the nearside border router a ‘nearside AS’ (NearAS).

3) List of Neighbour ASes: For each NearBR, we obtained a list of neighbour ASes that were connected to the NearAS at the NearBR.

Some ASes provided both the routes command and the summary command (e.g., show ip bgp summary). The summary command allowed us to not only find the neighbour ASes, but also identify those neighbour ASes connected to the NearBR via multiple border links. Figure 2 is an example table returned by the summary command from core1.tor1.he.net (tor1), a border router of Hurricane Electric. The table lists the ASNs of the BGP neighbours and the IP addresses of the interfaces through which the BGP sessions are established. Those neighbour ASes highlighted in red boxes were connected to tor1 via multiple neighbour addresses, i.e., via multiple border links, and therefore were potential candidates for the deployment of BGP-M.

For ASes that did not provide the summary command, we are unable to directly obtain their neighbour ASes connected to border routers via multiple border links. Therefore, we firstly obtained the AS paths from BGP RIB entries provided by RouteViews [44]. Then for each NearAS, we extracted its neighbour ASes as those next to it in any RIB AS path. Afterwards, we query all the extracted neighbour ASes, which requires more analysis in our inference.

4) Retrieving Routing Table: For each NearBR, we retrieved its routing table information using the routes command, e.g., show ip bgp routes detail <IP address>. Queries to any IP address in a prefix should return the same routing table. Hence, we only queried one IP address in each prefix. Thus we set the parameter IP address as X.Y.Z.1 for IPv4 (or X::Y::Z::1 for IPv6) for each prefix in a neighbour AS. We obtained the full list of prefixes in each neighbour AS from data provided by RouteViews [44]. For simplicity, we only considered /24 prefixes for IPv4 and /48 prefixes for IPv6, which accounted for 57.5% and 45.8% in the RouteViews data for IPv4 and IPv6, respectively.

5) Identifying BGP-M Cases: Figure 3 shows an example response to the command from the border router tor1 of Hurricane Electric. The router has learned and installed two paths, via different next hops 198.32.181.46 and 206.108.34.48 (i.e., two border links), towards the same destination prefix (142.46.150.0/24) in the neighbour AS (AS19752).

Both paths are labeled with status codes of “M” and “E”, meaning they are multipath learned via eBGP. The two paths all have the same values for attributes of LocPref, AS Path, Origin, Metric ‘0’ (for IGP) and MED (not shown in the figure), suggesting they are equal-cost multiple paths. This is the ground-truth evidence that Hurricane Electric (AS6939) has deployed BGP-M at tor1 to a destination prefix in the neighbour AS19752. This BGP-M case is denoted as <AS6939, tor1, AS19752, 142.46.150.0/24>.

In this study, we did not query all prefixes in a neighbour AS because we aimed to reduce the total number of queries to LG servers, which often set a cap on the number or frequency of queries from a host. If a prefix in a neighbour AS was identified as having BGP-M at the border router, we stopped querying prefixes in that neighbour AS and we continued with another neighbour AS. Thus, we may not uncover all BGP-M...
Fig. 3. An example of LG response to the routes command (show ip bgp routes detail <IP address>), from the border router core1.tor1.he.net of Hurricane Electric.

### TABLE IV

| AS number | AS name         | AS rank | # of BGP-M cases | # of neighbour ASes | # of nearside border routers |
|-----------|-----------------|---------|------------------|---------------------|-----------------------------|
|           |                 |         | total / IXP / Direct / Hybrid | total / BGP-M / ratio | total / BGP-M / ratio       |
| 6939      | Hurricane Electric | 7       | 1,088 / 1,006/68/14 / 5,868 | 611 / 10.4% / 112 / 69 | 61.0% / 2 / 51 / 22.2%       |
| 9002      | RETN            | 13      | 155 / 87/65/3 / 1,547 | 108 / 7.0% / 130 / 51 | 39.2% / 2 / 22.2%           |
| 3216      | PJSC VimpelCom  | 25      | 2 / 20 / 2 / 59 | 1 / 7.5% / 7 / 1 | 50.0% / 1 / 14.3%           |
| 20764     | CJSC RASCOM     | 30      | 27 / 23/4 / 858 | 2 / 7.5% / 27 / 6 | 22.2% / 1 / 50.0%           |
| 12303     | ISZT            | stub    | 2 / 20 / 2 / 59 | 1 / 7.5% / 7 / 1 | 50.0% / 1 / 14.3%           |
| 22691     | ISPnet          | stub    | 3 / 0 / 24 / 3 / 24 | 3 / 6.5% / 7 / 1 | 14.3% / 1 / 2.0%            |
| 48972     | BetterBe        | stub    | 2 / 20 / 2 / 59 | 1 / 7.5% / 7 / 1 | 50.0% / 1 / 14.3%           |
| 52201     | TCTEL           | stub    | 1 / 0 / 2 / 11 | 1 / 7.5% / 7 / 1 | 50.0% / 1 / 14.3%           |
| 131713    | IDNIC           | stub    | 1 / 0 / 2 / 11 | 1 / 7.5% / 7 / 1 | 50.0% / 1 / 14.3%           |
| 196965    | TechCom         | stub    | 24 / 2 / 40 / 36 | 15 / 41.7% / 2 / 1 | 100.0% / 1 / 20.0%          |
| 328112    | LBSD            | stub    | 13 / 0 / 2 / 29 | 13 / 44.9% / 2 / 1 | 50.0% / 1 / 14.3%           |

| IDNIC | IDNIC-SPICE/Link-AS-ID | LBSD: Linux-Based-Systems-Design-AS |
|-------|------------------------|-------------------------------------|

| AS number | AS name         | AS rank | # of BGP-M cases | # of neighbour ASes | # of nearside border routers |
|-----------|-----------------|---------|------------------|---------------------|-----------------------------|
|           |                 |         | total / IXP / Direct / Hybrid | total / BGP-M / ratio | total / BGP-M / ratio       |
| 6939      | Hurricane Electric | 7       | 390 / 266/14/20 / 3,880 | 146 / 3.8% / 112 / 35 | 31.3% / 2 / 18.5%           |
| 9002      | RETN            | 13      | 45 / 25/18 / 926 | 23 / 2.5% / 130 / 24 | 18.5% / 1 / 100.0%          |
| 8647      | AS-T2012        | stub    | 2 / 20 / 46 | 2 / 4.3% / 1 / 1 | 100.0% / 1 / 20.0%          |
| 48972     | BetterBe        | stub    | 2 / 20 / 2 / 6 | 1 / 16.7% / 4 / 2 | 50.0% / 1 / 20.0%           |
| 131713    | IDNIC           | stub    | 1 / 0 / 5 | 1 / 20.0% / 5 / 1 | 20.0% / 1 / 20.0%           |
| 328112    | LBSD            | stub    | 6 / 0 / 28 | 6 / 21.4% / 2 / 1 | 50.0% / 1 / 20.0%           |

C. Our Inference Results

Table IV summarizes our inference results obtained by applying the above method to all 2,709 ASes with an LG server. It shows that BGP-M has been widely deployed not only by large transit ASes, such as Hurricane Electric, RETN, PJSC VimpelCom and CJSC RASCOM, but also by many stub ASes. BGP-M is deployed on both IPv4 and IPv6 Internet.

Note that although we tried to discover as many BGP-M cases as possible, our inference result was far from a complete measurement. In fact, only a small portion of ASes provide a LG server, of which only a small portion are publicly accessible and support the routes command. In addition, as explained above, we only had limited resource and time to uncover at most one BGP-M case (for one of many destination prefixes) in each neighbour AS. Thus, it is highly likely that there are a lot more BGP-M cases already deployed in the Internet. Our inference result provided a lower bound estimation.

V. Analysis of BGP-M Cases Deployed by Hurricane Electric

The most notable AS in our inference result is Hurricane Electric (HE, AS6939). It is a Tier-1 network, ranked 7th in the Internet. As a major Internet service provider, it had 112 border routers neighboring with 5,868 ASes on IPv4 in January 2020. It is remarkable that it has already extensively implemented at least 1,088 BGP-M cases to (at least one) prefixes in 611 of its neighbour ASes at 69 border routers. Note that there could
be many more cases of BGP-M to be discovered, i.e., those deployed to other prefixes in these neighbour ASes and those deployed to prefixes in remote ASes.

Of the 1,088 BGP-M cases on IPv4, 911 cases used 2 border links, 92 cases used 3 links, and 85 cases used 4 links. Of the 300 BGP-M cases on IPv6, 248 cases used 2 links, 33 cases used 3 links, and 19 cases used 4 links. There are much less BGP-M cases on IPv6 than on IPv4, possibly because there is less demand for load balancing on IPv6 than on IPv4.

A. BGP-M Cases via IXP

We relied on data from PeeringDB [54] to identify whether a BGP-M case was deployed via IXP or not. We compiled a list of IXPs and the prefixes belonging to them. If all of the FarIPs in a case belonged to IXPs, this case was identified as via IXP; if none of the FarIPs in a case belonged to IXPs, this case was via Direct links; otherwise, this case was via Hybrid links.

It is notable that IXPs played a vital role in HE’s the deployment of BGP-M as they were involved in 92.5% (i.e., 1,006 in 1,088) of the IPv4 cases, and 88.9% (i.e., 266 in 300) of the IPv6 cases.

Because we only considered one data source on IXP data, some cases via IXP might be mis-classified as via direct links or via hybrid links. Thus, IXPs might be more important in the deployment of BGP-M than we observed.

B. Analysis on Neighbour ASes

Figure 4(a) plots the 611 neighbour ASes of HE with at least one BGP-M case on IPv4, ordered by each AS’ customer cone size [56], [57]. The customer cone of an AS X is a set of ASes including (1) X’s customer ASes, and (2) X’s customer ASes’ customer ASes, and so on. The customer cone size of AS X is the number of ASes in X’ customer cone [57]. The plot also shows the total number of HE border routers deployed with BGP-M to a neighbour AS in a large circle, and the number of HE border routers deployed with BGP-M to a neighbour AS via IXP in a small circle.

There are three interesting observations. Firstly, Yahoo! (AS10310), a content provider network with customer cone size of 41 and AS rank of 747, was deployed with BGP-M by HE at as many as 32 border routers. Secondly, small & medium ASes (with customer core size < 100) were more likely to be deployed with BGP-M at multiple border routers, suggesting Hurricane Electric has deployed richer and more complex connections to small & medium ASes than to top-rank ASes. Third, IXPs were widely involved in Hurricane Electric’s the deployment of BGP-M. For many neighbour ASes, all of their BGP-M cases were connected through IXP. It is very likely that HE’s heavy reliance on IXP is a reason why we observed so many BGP-M cases with small & medium ASes.

Table V lists the 10 highest ranked neighbour ASes deployed with BGP-M by HE on IPv4 and IPv6. Although these ASes are not highly ranked, they are all well-known content provider networks or content delivery networks, and most of them are among the list of 15 hyper-giant ASes recognised by Böttger et al. [58], where hyper-giant ASes are defined as ASes having wide geographical coverage, large port capacity, and large traffic volumes [58].

A comparison between Table V and Table VI highlights the difference between top-rank ASes and hyper-giant ASes (with low ranks) in terms of the requirement for BGP-M. BGP-M was more needed and useful for routing with content providers, where load balancing can be crucial for delivery of large traffic volume.

Although the amount of BGP-M cases on IPv6 was much lower than IPv4, they exhibited similar properties, suggesting HE has applied similar BGP-M policies on IPv4 and IPv6.

C. Analysis on Border Routers

Hurricane Electric’s LG server [59] covered 112 border routers distributed around the world. Table VII shows that most of its routers were located in North America and Europe and many of them have been implemented with BGP-M. The geo-locations of the border routers were directly obtained from their names as given by the LG server. Although there were
only a few border routers located in Asia and other parts of the world, a large portion of them have been implemented with BGP-M.

Figure 5 plots the number of neighbour ASes in triangle and the number of neighbour ASes with BGP-M in square

Fig. 5. List of 112 border routers of Hurricane Electric (AS6939). The border routers are ordered by the number of IPv4 neighbour ASes.

Fig. 6. Relation between the number of border routers and the number of neighbour ASes deployed with BGP-M in Hurricane Electric.

D. Relation Between Neighbour ASes and Border Routers

Figure 6 shows the relations between the number of neighbour ASes and the number of border routers deployed at each of HE’s 112 border routers on IPv4 and IPv6. On IPv4, HE has deployed BGP-M to the largest number (78) of neighbour ASes at the border router par2; and on IPv6, the border router ams1 had BGP-M cases to 38 neighbour ASes.
with BGP-M by HE. Figure 6(a) shows that HE has deployed BGP-M on IPv4 to 493 neighbour ASes at only one border router, and BGP-M to the other 118 neighbour ASes at least 2 border routers. HE has deployed BGP-M with Yahoo! (AS10310) at 32, the largest number of, border routers on IPv4, accounting for 87% of the border routers connected to Yahoo! Similar observation on IPv6.

Figure 6(b) shows that on IPv4, 9 of HE’s border routers had BGP-M to only one neighbour AS; while other border routers had BGP-M to at least two neighbour ASes. The border router $\text{par2}$ was deployed with BGP-M to the largest number (78) of neighbour ASes. It is evident that BGP-M can be deployed in a flexible way to suit a network’s needs.

E. Summary

Our inference results from LG data showed that BGP-M has been deployed by both large transit ASes and stub ASes. IXPs were widely involved in the deployment of BGP-M, suggesting the important role IXPs play in facilitating the deployment of BGP-M. Moreover, our results revealed that the small & medium ASes, especially those content provider networks, were more likely to be deployed with BGP-M at multiple border routers, indicating these ASes’ heavy reliance on load balancing to improve their inter-domain traffic delivery.

VI. STUDY ON BGP-M ROUTING PATHS

For a known BGP-M case, we can use traceroute to reveal exact details on how traffic is shared on border links.

A. Our Traceroute Probing

Among existing traceroute projects, including RIPE Atlas [10], CAIDA Archipelago (Ark) [60] and iPlane [61], we found that RIPE Atlas installed publicly accessible traceroute probes in 5 of the 12 ASes where we identified BGP-M cases (see Table IV). These 5 ASes were Hurricane Electric, PSIC VimpelCom, CJSC RASCOM, ISZT and BetterBe, and they had 3, 3, 4, 2, and 1 RIPE Atlas probes, respectively. They had in total >1,400 BGP-M cases on IPv4 and IPv6.

For each BGP-M case identified as $\langle \text{NearAS}, \text{NearBR}, \text{FarAS}, \text{DstPrfx}\rangle$, we sent traceroute probes from available RIPE Atlas probes in the NearAS to all IP addresses between $X.Y.Z.1$ and $X.Y.Z.254$ of DstPrfx on IPv4, or the first 254 IP addresses between $X:Y:Z::1$ and $X:Y:Z::fe$ on IPv6. We used ICMP packets and UDP packets, Paris traceroute variation 16 [25] with default settings on RIPE Atlas, e.g., 3 packets for probing to each destination IP.

For each BGP-M case, we check whether the traceroute paths sent from a probe to IP addresses in the DstPrfx actually traversed the NearBR where the BGP-M case was deployed. If the traceroute paths traverse elsewhere, we discard them. Below is our procedure.

1) Obtain the list of ending points of border links, i.e., FarIPs, which are given in the routing table returned by the routes command (see ‘Next Hop IPs’ in Fig. 3).

2) For each traceroute path, check if any of the FarIPs appears in the traceroute path. If yes, go to (3); otherwise, discard this traceroute path.

3) Use the DNS Chain service provided by RIPEstat Data API [62] to obtain the router name of the predecessor IP address of the FarIP by using the link of https://stat.ripe.net/data/dns-chain/data.json?resource=<IP address>. If the router name is NearBR, finish the process; otherwise, discard this traceroute path.

Step (3) is necessary because (1) it confirms that the traceroute paths traversed the NearBR of the BGP-M case under study and (2) it also locates the NearBR when different BGP-M cases with different NearBRs share the same FarIPs of the same FarAS.

In this study, we set a standard for traceroute measurement. That is, we will only consider traceroute measurement of a BGP-M case if we are able to obtain traceroute data to at least 250 of the 254 IP addresses in the destination prefix and they traverse the relevant NearBR and BLs.

B. Load Balancing Algorithms of Cisco Routers

For all BGP-M cases that we were able to run traceroute measurement, the LG commands (routes and summary) and the returned information were all in the style and format of Cisco, indicating the NearBRs where these BGP-M cases were implemented were all Cisco routers.

According to Cisco’s configuration documentation [63], by default, Cisco routers can configure BGP-M as per-session load balancing, where traffic allocation decisions are made by hash algorithm for each pair of source and destination IP addresses. Network operators can also configure their routers as per-packet round robin load balancing, or per-flow load balancing where the hash algorithm considers source and destination IP addresses and port numbers.

Figure 7(a) shows the topology map of the BGP-M case $\langle\text{AS6939}, \text{tyo1}, \text{AS2907}, 160.18.2.0/24\rangle$, where the NearAS connects with the FarAS via two border links. We sent traceroute packets from two RIPE Atlas probes located inside AS6939, i.e., SrcIP-1 and SrcIP-2. Traffic from the two sources arrived at the NearBR (tyo1) at two different ingress interfaces, i.e., NearIP-1 and NearIP-2. Traffic from each source was shared on the two border links, i.e., BL-1 and BL-2. According to the IXP data from PeeringDB [54], FarIP-1 and FarIP-2 belong to two IXPs, named as JPIX TOKYO and JPNAP Tokyo, respectively.

Figures 7(b)–7(e) show the routing maps observed from traceroute probes with different settings. Figures 7(b) and 7(c) are both based on UDP packets sent from SrcIP-1 (209.51.186.5), but at two time points 15 minutes apart. Figures 7(d) and 7(e) are both based on ICMP packets at the same time point, but sent from different sources: SrcIP-1 (209.51.186.5) and SrcIP-2 (65.19.151.10), respectively. Here are some observations.

1) Firstly: All four routing maps show that probes to the IP addresses in the destination prefix were always equally shared on the two border links, which, as expected, showed...
BGP-M provides load balancing at the level of destination prefix.

2) Secondly: On the routing maps in Figures 7(b) and 7(c) based on UDP packets, the packets to different destination IPs were randomly allocated on the two border links, and the allocations varied at different time points. This is the hallmark of load balancing based on the so-called include-ports algorithm [63], which takes into account IP addresses and port numbers of source and destination. While hash function is sensitive to any change of bits in the identifiers, the UDP packets have not only different destination IPs, but also different port numbers when sent at different time points.

3) Thirdly: on the routing maps in 7(d) and 7(e), the ICMP packets were allocated on the two border links in a regular way: packets to 4 consecutive destination IPs were allocated on one border link, and the next 4 on the other border link; then the pattern repeated alternately. This suggests (1) the Cisco router was configured to conduct per-session load balancing for ICMP traffic using the so-called universal algorithm [63] which considers only source and destination addresses; and (2) only a part of the destination IP address was considered [8].

Closer inspection revealed that the BGP-M allocation patterns in the two routing maps were exactly opposite to each other, i.e., destination IPs allocated to BL-1 in 7(d) were allocated to BL-2 in 7(e), and vice versa. This is because the routing maps were based on packets sent from different source addresses. Indeed, due to the universal algorithm, there are only two possible allocation patterns for ICMP packets from any sources to IP addresses in a destination prefix.

Cisco routers implement load balancing for UDP and ICMP traffic in vastly different ways. Since most real traffic flows are TCP or UDP, we should conduct traceroute measurements with UDP packets to reveal the true picture of BGP-M load balancing.

C. Diverse Routing Patterns in FarAS

The topology map in Figure 7(a) shows that in this BGP-M case, the two border links entered the FarAS via two different
IXPs, each of which further split traffic onto two internal links within the FarAS. The routing maps in Figure 7 show that the two IXPs conducted load balancing in a random way so that the internal links received similar portions of traffic, whether probed by UDP or ICMP from the same or different sources. Due to limited availability of RIPE Atlas probes located in relevant nearside ASes, we were only able to run traceroute measurements with ICMP packets for 89 of all BGP-M cases uncovered in this study.

We observed the followings: (1) in 33 cases, traffic on each border link was further split onto different links in the FarAS; (2) in 22 cases, traffic on different border links were forwarded separately to destination prefix via parallel routing paths; (3) in 14 cases, traffic on all border links were later merged into a single path; (4) in 15 cases, there were complex routing in the FarAS, and (5) in the other 5 cases, traceroute contained unresponsive hops. We conducted traceroute with UDP packets and observed similar results.

The above observation on diverse intra-domain routing in FarAS suggests that the deployment of BGP-M at border routers of an AS is independent, or transparent, to intra-domain load balancing in other ASes.

VII. MEASUREMENT OF DELAYS ON BORDER LINKS
To understand the effectiveness and performance of load balancing by BGP-M, here we study traffic delay on BGP-M border links based on traceroute measurements using ICMP and UDP packets.3

A. Measuring Link Delay Based on Traceroute RTT
Figure 8 shows the link delays measured on the two BLs for the BGP-M case <AS6939, hkg1, AS20940, 23.67.36.0/24>, which was deployed by HE (AS6939) at its border router core1.hkg1.he.net (with NearIP 184.105.64.129) to the DestPrfx 23.67.36.0/24 in a neighbour AS called Akamai (AS20940). We sent traceroute using ICMP packets and UDP packets, with one-minute separation, from a RIPE Atlas probe to each of 254 IP addresses in the destination prefix at 15-minute intervals for 3 days from 00:00am Hong Kong local time (i.e., GMT + 08:00) on 16 June 2021. We used default RIPE Atlas traceroute settings.

From each traceroute measurement to a destination IP, we obtained the Round Trip Time (RTT) value at each IP hop; and then we calculated the delay on a border link, which was the difference between the RTT values of NearIP and FarIP of the border link.

The link delay includes the following sources: (1) processing delays at NearBR and FarBR, which are negligible because of border routers’ high performance; (2) serialization delay at NearBR, which is negligible because of border links’ high bandwidths; (3) transmission delay on border links, which is negligible because of small distance (< 30km) between the relevant facilities, all located in Hong Kong; and (4) queuing delay at NearBR, which accounts for most of the link delay.

Since the link delay mainly measures the queuing delay at the NearBR, it reflects the level of traffic congestion for each of the border links and therefore can be considered as an indicator of routing performance.

As explained in the previous section, when using ICMP packets, traceroute paths to a given destination IP at different time points always go through the same border link due to per-session load balancing that considers only source and destination addresses; whereas for UDP packets, traceroute paths to a given destination IP at different time points are allocated to any of the two border links at random due to per-flow load balancing that considers IP addresses and port numbers of source and destination.

B. Congestion-Free Transit on Border Links
Figures 8(a) and 8(b) show the distributions of link delays on the two border links to the destination IPs measured by 15-minute interval during three days based on ICMP and UDP packets, respectively. In general, the link delay values were mostly small, mainly between 20ms and 40ms. Although there were delays more than a few seconds, they were very rare. This indicates that transit on these border links were mostly free of congestion.

The congestion-free transit can be explained by the large bandwidths and relatively small traffic volumes on the two border links. The bandwidths of the two border links were: 10G for BL-1 (with FarIP-1 103.247.139.17) and 100G for BL-2 (with FarIP-2 123.255.91.169). We obtained the bandwidth information from PeeringDB [54], where the public peering data for Akamai (AS20940) was last updated on 10 July 2021. We also obtained from Akamai’s technical report4 that Akamai at Hong Kong (where the FarBRs were located) had average traffic volume of 21.9 Mbps and peak volume of 129.5 Mbps in Q1 2017. Although the report was four years ago, today’s traffic volume is likely to remain well below the bandwidths of the border links.

C. Load Balancing on the Border Links
Figures 8(c) and 8(d) show the change of link delay on the two border links at different time points in 3 days. For each time point, we show the median as well as the 25th and 75th percentiles of calculated link delay values to all destination IPs.

Figures 8(e) and 8(f) show the statistics of link delay to each IP address in the destination prefix. As explained in previous section, ICMP packets to destination IPs are equally allocated to the two border links in exactly the same way at every time point. That is, in Figure 8(e), the statistics for BL-1 show link delays to only 128 destination IPs, each of which was calculated from 288 measurements (= 3 days × 24 hours × 4 times/h); whereas the statistics for BL-2 show link delays to 125 different destination IPs.5 By comparison, UDP packets to destination IPs are equally, but randomly, allocated to the two border links, and allocation changes randomly at every

---

2With the measurement standard we set in Section VI-A.
3Our recent work reported in [28] used only ICMP packets.
4https://www.akamai.com/us/en/multimedia/documents/state-of-the-internet/q1-2017-state-of-the-internet-connectivity-report.pdf
5No traceroute data to 23.67.36.1.
Fig. 8. Delays on the two border links of BGP-M case <AS6939, hkg1, AS20940, 23.67.36.0/24>. Delays were measured by sending traceroute ICMP and UDP packets (with one minute separation) from a RIPE Atlas probe in AS6939 to each IP address in the destination prefix 23.67.36.0/24 at 15-minute intervals starting from 00:00am GMT on 16 June 2021 for 3 days. Y-axis is plotted on log scale. (a) and (b) show the distribution of link delay values (in 20ms bins) to all IP destination at all time points, where the inset shows the distribution (in linear scale) of delays between 0ms and 340ms that account for >95% of all values. (c) and (d) show statistics of link delay calculated over link delay values to all destination IPs at a given time point; (e) and (f) show statistics calculated over values to a given destination IP at all time points.

The two border links always had similar delay values, throughout the duration of measurement, to all destination IPs, and for both ICMP and UDP packets. This is as expected. It vividly illustrates the desired result of BGP-M load balancing.
where traffic is equally shared between border links to fully utilize routing capacity and diversity available at borders of ASes, with the purpose to reduce congestion and improve routing flexibility and resilience.

**VIII. Extraordinary Effort of HE in BGP-M**

According to the CAIDA AS Rank data [56], Hurricane Electric (HE, AS6939) is the 7th largest AS in terms of customer cone size, and it provides transit between more than 8k ASes, which account for 12% of all ASes in global routing tables. HE is also a hyper-giant AS of very high peering affinity and port capacity [58], with more than 6k peers and presence at 236 IXPs – more than all other ASes. Hence, HE has a particular need for the best practice of traffic engineering in order to achieve stable, reliable and high routing performance with its enormous number of neighbors.

**A. Extensive Deployment of BGP-M Cases**

Our results show that HE has extensively deployed BGP-M across its entire network. It has deployed BGP-M to 611 (i.e., >10%) of its neighbour ASes at 69 (i.e., >60%) of its border routers on the IPv4 Internet. HE has also widely deployed BGP-M on the IPv6 Internet. Notably, HE deployed large numbers of BGP-M cases to other hyper-giant ASes, especially large content providers, on both IPv4 and IPv6 Internet (see Section V-B). We have confirmed with HE on their wide deployment of BGP-M through our consultation with their network operators.

The fact that a large, traffic-intensive, transit network like HE has devoted such extraordinary effort in wide deployment of BGP-M across its world-wide network is an evident indication of the significant benefit and advantage that this load balancing technique can provide.

**B. Active Maintenance of BGP-M Cases**

As shown in Figure 3, LG server’s response to the routes command (show ip bgp routes detail <IP address>) contains rich details on the deployment of BGP-M, including the time since the routing table has been last updated. This allows us to track the changes of a BGP-M case by re-querying the relevant border router at a later time. As shown in Table VIII, at the beginning of July 2021, we revisited each of BGP-M cases of HE that we observed in our 2020 measurement.

For the 1,088 cases deployed by HE on the IPv4 Internet, 632 (or 58%) of the cases remained exactly the same; and 60 cases had additional or replaced border links, which, according to our definition, were still of the same BGP-M cases as they were deployed at the same border routers for the same destinations. We also observed that 396 (or 36%) cases were either disappeared or changed since our 2020 measurement. For example, LG queries suggested some nearside border routers were ‘Not existing’ anymore, and some routes were not labeled as ‘M’ (i.e., multipath) anymore. A small number of cases were still there but with changed attributes making them different or new BGP-M cases, for example with a changed farside AS. We observed similar results for the BGP-M cases on IPv6.

The above observations suggest that HE has been actively maintaining and rearranging its BGP-M cases. Some of the changes might occur in reaction to network changes while others were likely for the purpose of achieving better configuration to gain more benefit from BGP-M load balancing.

**IX. Advantages and Benefits of BGP-M**

As a load balancing technique, BGP-Multipath not only provides balanced traffic and enhanced routing performance, but also offers a number of unique advantages and benefits in terms of deployment and operation.

**A. Wide Availability and Readiness for Implementation**

Both hardware and software requirements for the deployment of BGP-M are already widely available in the Internet.

Firstly, there is a wide presence of multiple border links between ASes in the Internet, where more than one border links are connecting from a border router of an AS to border router(s) of a neighbour AS. Such multiple border links commonly exist, especially among core ASes or between core and peripheral ASes.

Secondly, most border routers provided by major router vendors, such as Cisco, Juniper and Huawei, already support BGP-M load balancing, which is an integral part of their design and function.

This means BGP-M can be readily implemented by network operators with many of their neighbors without changing or upgrading their infrastructure or agreements.

**B. Easy Implementation**

The implementation of BGP-M is rather simple and straightforward. For example, the minimum action required on a Cisco border router is to activate BGP-M by changing a single parameter maximum-paths from its default value 1 to the number of (different) paths for a given DstPrfx. There is literally no additional cost to implement BGP-M.

---

TABLE VIII

| Changes of BGP-M Cases Deployed by HE on IPv4 and IPv6 Internet |
|---------------------------------------------------------------|
|                  | IPv4 | IPv6 |
| 2020 measurement dates | Jan-May 2020 | July-Oct 2020 |
| Total # of BGP-M cases | 1,088 | 300 |
| 2021 measurement dates | July 2021 | July 2021 |
| Total # of cases | 692 | 218 |
| Exactly same as before | 652 | 204 |
| Increased # of BLs | 33 | 7 |
| Same # but different BLs | 27 | 7 |
| # of remaining cases | 692 | 218 |
| Total # of cases | 396 | 82 |
| ‘NearBR ‘Not existing’ | 13 | 0 |
| ‘No routes’ for DstPrfx | 143 | 12 |
| Status without ‘M’ (multipath) | 109 | 25 |
| Status without ‘E’ (eBGP) | 102 | 32 |
| Other changes | 29 | 13 |
C. Independent, Flexible and Transparent Deployment

Although we call it BGP-M and the technique follows BGP’s best path selection process, network operators do not need to alter their BGP process to deploy BGP-M as the load balancing will still follow exactly the same AS-level path as before. As such, network operators can freely and independently implement or remove BGP-M without informing or obtaining new agreement from their neighbour ASes. Network operators can deploy, revise and cancel BGP-M for any selections of destination prefixes in any neighbour ASes.

There is no interference between the deployment of BGP-M, any other multipath routing techniques implemented within or outside of the AS, and any traffic engineering configurations elsewhere. For example, as we showed in previous sections, there is no impact on BGP-M load balancing whether the border links connect to the neighbour AS directly or via IXPs, or whether and how IXPs further apply their own load balancing arrangement.

Basically, BGP-M deployed at a border router is transparent to other parties participating in the relevant traffic routing, which gives network operators flexibility and convenience.

D. Benefits of BGP-M Load Balancing

The benefits of load balancing gained from the deployment of BGP-M is no less than any other multipath routing techniques. It can increase more balanced use of border links and reduce risk of congestion in face of traffic surges. It can also improve routing path diversity, which can be useful for network resilience and security.

In addition, the border links are common, already there, and shared by networks. Many border links have high bandwidths. It makes sense to fully utilize these resources that are readily available, especially when it is easy and convenient to do so. A network operator benefits from the deployment of BGP-M regardless of whether or how many other networks have implemented the technique. The more deployment, the more benefit. And such benefits are likely to be mutually beneficial to not only the AS that deploys BGP-M but also its neighbour ASes.

E. Immense Potential for Future Deployment of BGP-M

Although our work shows BGP-M has already been widely deployed, as explained in previous sections, our inference is a conservative, lower bound estimate of the scale of the deployment of BGP-M in the Internet and there could be many more BGP-M cases. Indeed we have recently discovered a few hundreds more BGP-M cases by querying LG servers all prefixes (of all sizes) announced by each neighbour AS. Many of them are BGP-M cases implemented by HE to other hyper-giant ASes. Considering only a relatively small portion of ASes provide LG services, there should be more ASes that have implemented BGP-M in the Internet.

Nevertheless, based on our data and analysis so far, we estimate that the scale of existing deployment of BGP-M is still far smaller than the intra-domain multipath routing, of which millions of cases have been uncovered throughout the Internet. This means there is an immense scope for future deployment of BGP-M by more ASes to more destinations.

In recent years, a significant amount of investment and effort have been devoted in coping with the rapid increase of traffic volume in the Internet. This study provides a technical and economic case for more deployment of BGP-M as an option for load balancing, which is compatible and complementary to other traffic engineering techniques.

X. Conclusion and Future Works

A. Conclusion

This paper reports the first measurement of the deployment of BGP-M in the Internet. Our measurement was based on Looking Glass server data, which provides not only the ground truth, but also rich information on various aspects of BGP-M.

We provided the state-of-the-art knowledge on the BGP-M. We ran traceroute from RIPE Atlas probes to reveal the exact patterns of traffic allocation as well as delays on border links. We provided in-depth analysis on a series of BGP-M cases deployed by Hurricane Electric, as a capital example of large-scale deployment of BGP-M by a major transit network.

Our work is valuable to network operators interested in load balancing. For example, our work and the example of HE may inspire more network operators to consider deploying BGP-M and steering their traffic with neighbour ASes via multiple border links.

Our work is also valuable to Internet researchers, who can use our measurement data to obtain a fuller and more detailed picture of the inter-domain routing between ASes. This can help the study on inter-domain congestion [64] and the study on Internet traffic map [65].

We have shared on GitHub [66] the LG data and the traceroute datasets for all discovered BGP-M cases reported in this paper.

B. Future Works

As an ongoing research, our work can be improved and extended in several ways.

Firstly, we plan to discover more BGP-M cases. One method is to study more ASes with LG servers, which requires access to more LG servers without invoking ASes’ policies. Another approach is to conduct large-scale traceroute measurement similar to recent studies on multipath routing [7], [8], where the challenge is to achieve accurate AS border mapping. The aim is to provide a more complete and accurate picture of the deployment of BGP-M on the global Internet. It would be interesting to find out whether BGP-M has been more extensively deployed than we present in this paper.

Secondly, with more measurement data, we will study the reasons and motivations for network operators to deploy, or not deploy, BGP-M when multiple border links are available. For example, we will investigate whether the deployment of BGP-M is related to traffic volume and AS relationships. Such study may provide new knowledge on BGP-M including its utility as well as drawbacks and limitations.

Thirdly, we will further analyse the routing performance of BGP-M using comparative measurements. For example, we will assess performance of two border routers with the same topological connectivity (in terms of border links with
neighbour ASes) but only one of them has deployed BGP-M. We will also monitor traffic performance of a border router before and after it activates BGP-M, which requires collaboration from network operators. Such study will enrich our understanding of BGP-M.

REFERENCES

[1] Y. Rechter, T. Li, and S. Hares, “A border gateway protocol 4 (BGP-4),” Internet Eng. Task Force, RFC 4271, Jan. 2006.
[2] (Juniper Networks, Sunnyvale, CA, USA). Understanding BGP-Multipath, [Online]. Available: https://www.juniper.net/documentation/en_US/junos/topics/topic-map/bgp-multipath.html (Accessed: May 2022).
[3] (Cisco, San Jose, CA, USA). BGP Best Path Selection Algorithm. [Online]. Available: https://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/13753-25.html#anc5 (Accessed: May 2022).
[4] (Huawei, Shenzhen, China). Example for Configuring BGP Load Balancing. [Online]. Available: https://support.huawei.com/enterprise/en/doc/EDOC1000178324/dd0029a9/example-for-configuring-bgp-load-balancing (Accessed: May 2022).
[5] K. Vermeulen, D. S. Stephen, O. Fournaux, and T. Friedman, “Multi-level MDA-Lite Paris traceroute,” in Proc. ACM IMC, 2018, pp. 29–42.
[6] B. Augustin, T. Friedman, and R. Teixeira, “Measuring multipath routing in the Internet,” IEEE/ACM Trans. Netw., vol. 19, no. 3, pp. 830–840, Jun. 2011.
[7] K. Vermeulen, J. P. Rohrer, R. Beverly, O. Fournaux, and T. Friedman, “Dsrusiner: Comprehensive discovery of the Internet’s topology diamonds,” in Proc. USENIX NSDI, 2020, pp. 479–493.
[8] R. Almeida, I. Cunha, R. Teixeira, D. Veitch, and C. Diot, “Classification of load balancing in the Internet,” in Proc. IEEE INFOCOM, 2020, pp. 1987–1996.
[9] R. Almeida, O. Morais, E. Fazzion, D. Guedes, W. Meira, Jr., and I. Cunha, “A characterization of load balancing on the IPv6 Internet,” in Proc. PAM, 2017, pp. 242–254.
[10] RIPE NCC Staff, “RIPE Atlas: A global Internet measurement network,” Internet Protocol J., vol. 18, no. 3, pp. 2–26, 2015.
[11] V. Giotsas et al., “O peer, where art thou? Uncovering remote peer-interconnections at IXPs,” IEEE/ACM Trans. Netw., vol. 29, no. 1, pp. 1–16, Feb. 2021.
[12] U. Javed, I. Cunha, D. R. Choffnes, T. Anderson, and K. Vermeulen, “A characterization of load balancing in the Internet,” in Proc. IEEE INFOCOM, 2020, pp. 1025–1038, Aug. 1997.
[13] J. Qadir, A. Ali, K. A. Yau, A. Sathiaseelan, and J. Crowcroft, “Exploiting the power of multiplicity: A holistic survey of network-layer multipath,” IEEE Commun. Surveys Tuts., vol. 17, no. 4, pp. 2176–2123, 4th Quart., 2015.
[14] S. K. Singh, T. Das, and A. Jukan, “A survey on Internet multipath routing and provisioning,” IEEE Commun. Surveys Tuts., vol. 17, no. 4, pp. 2157–2175, 4th Quart., 2015.
[15] J. Qadir, A. Ali, K. A. Yau, A. Sathiaseelan, and J. Crowcroft, “Exploiting the power of multiplicity: A holistic survey of network-layer multipath,” IEEE Commun. Surveys Tuts., vol. 17, no. 4, pp. 2176–2123, 4th Quart., 2015.
[16] B. Augustin, T. Friedman, and R. Teixeira, “Multipath tracing with Paris traceroute,” in Proc. IEEE E2EMON, 2007, pp. 1–8.
[17] D. Veitch, B. Augustin, R. Teixeira, and T. Friedman, “Failure control in multipath route trace,” in Proc. IEEE INFOCOM, 2010, pp. 1395–1403.
[18] B. Augustin et al., “Avoiding traceroute anomalies with Paris traceroute,” in Proc. ACM IMC, 2006, pp. 153–158.
[19] F. Valera, I. Van Beijnum, A. Garcia-Martinez, and M. Bagnulo, “Multi-path BGP: Motivations and solutions,” in Next-Generation Internet Architectures and Protocols, B. Ramamurthy, G. N. Rouskas, and K. M. Sivalingam, Eds. Cambridge, U.K.: Cambridge Univ. Press, 2011.
[20] J. Li, V. Giotsas, and S. Zhou, “Anatomy of BGP-multipath deployment in a large ISP network,” in Proc. TMA, 2020, p. 9.
[21] J. Li, S. Zhou, and V. Giotsas, “Performance analysis of BGP-multipath,” in Proc. IEEE GI, 2021, pp. 1–6.
[22] H. Fujinoki, “Multi-Path BGP (MBGP): A solution for improving network bandwidth utilization and defense against link failures in inter-domain routing,” in Proc. IEEE ICON, 2008, pp. 1–6.
[23] I. van Beijnum, J. Crowcroft, F. Valera, and M. Bagnulo. “Loop-freeness in BGP-multipath through propagating the longest path,” in Proc. IEEE INFOCOM, 2009, p. 6.
[24] B. Huffaker, A. Dhamdhere, M. Fomenkov, and K. Claffy, “Toward topology dualism: Improving the accuracy of AS annotations for routers,” in Passive and Active Measurement, A. Krishnamurthy and B. Plattner, Eds. Springer Int., 2010, pp. 101–110.
[25] A. Marder, M. Luckie, A. Dhamdhere, B. Huffaker, K. C. Claffy, and J. M. Smith, “Pushing the boundaries with bdrmapIT: Mapping router ownership at Internet scale,” in Proc. ACM IMC, pp. 56–69.
[26] A. Y. Nur and M. E. Tozal, “Cross-AS (X-AS) Internet topology mapping,” Comput. Netw., vol. 132, pp. 53–67, Feb. 2018.
[27] A. Marder and J. M. Smith, “MAP-IT: Multicast accurate passive inferences from traceroute,” in Proc. ACM IMC, pp. 397–411.
[28] C. Hoops, “Analysis of an equal-cost multi-path algorithm,” Internet Eng. Task Force, RFC 2992, Nov. 2000.
[29] P. Lapukhov, “Equal-cost multipath considerations for BGP,” Network Working Group Internet Draft draft-lapukhov-bgp-ecmp-considerations-02, Internet Eng. Task Force, Fremont, CA, USA, Jul. 2019. [Online]. Available: https://tools.ietf.org/id/draft-lapukhov-bgp-ecmp-considerations-02.html.
[30] R. K. P. Mok, V. Bajpai, A. Dhamdhere, and K. C. Claffy, “Revealing the load-balancing behavior of YouTube traffic on interdomain links,” in Passive and Active Measurement, R. Beverly, G. Smaragdakis, and A. Feldmann, Eds. Cham, Switzerland: Springer, 2018, pp. 228–240.
[31] E. Jasinska, N. Hilliard, R. Raszuk, and N. Bakker, “Internet exchange BGP route server,” Internet Eng. Task Force, RFC 7947, Sep. 2016.
[32] J.-J. Pansiot, P. Mérindol, B. Donnet, and O. Bonaventure, “Extracting intra-domain topology from mrinfo probing,” in Passive and Active Measurement, A. Krishnamurthy and B. Plattner, Eds. Heidelberg, Germany: Springer Int., pp. 81–90, 2010.
[33] V. Giotsas, G. Smaragdakis, B. Huffaker, M. Luckie, and K. C. Claffy, “Mapping peer interconnections to a facility,” in Proc. ACM CoNEXT, 2015, p. 37.
[34] R. Motamedi, B. Yeganeh, B. Chandrasekaran, R. Rejaie, B. M. Maggs, and W. Willinger, “On mapping the interconnections in today’s Internet,” IBM Journal of Research and Development, vol. 44, no. 6, pp. 1325–1325, Oct. 2010.
[35] M. Luckie, A. Dhamdhere, B. Huffaker, D. Clark, and K. C. Claffy, “Bdrmap: Inference of borders between IP networks,” in Proc. ACM IMC, 2016, pp. 381–396.
[36] B. Yeganeh, R. Durairajan, R. Rejaie, and W. Willinger, “How cloud traffic goes hiding: A study of Amazon’s peering fabric,” in Proc. ACM IMC, 2019, pp. 202–216.
[37] “University of Oregon Route Views Project.” Jan. 2020. [Online]. Available: http://www.routeviews.org/
[38] “RIPE Routing Information Service.” [Online]. Available: http://www.ripe.net/ris
[39] A. Khan, T. T. Kwon, H.-C. Kim, and Y. Choi, “AS-level topology collection through looking glass servers,” in Proc. ACM IMC, 2013, pp. 235–241.
[40] H. Chang, R. Govindan, S. Jamin, S. J. Shenker, and W. Willinger, “Towards capturing representative AS-level Internet topologies,” Comput. Netw., vol. 44, no. 6, pp. 737–755, 2004.
[41] J. Han, D. Watson, and F. Jahanian, “An experimental study of Internet path diversity,” IEEE Trans. Dependable Secure Comput., vol. 3, no. 4, pp. 273–288, Oct.–Dec. 2006.
[42] B. Zhang, R. Liu, D. Massey, and L. Zhang, “Collecting the Internet AS-level topology,” ACM SIGCOMM Comput. Commun. Rev., vol. 35, no. 1, pp. 53–62, 2005.
B. Holbert, S. Tati, S. Silvestri, T. F. La Prota, and A. Swami, “Network topology inference with partial information,” IEEE Trans. Netw. Service Manage., vol. 12, no. 3, pp. 406–419, Sep. 2015.

V. Giotsas, A. Dhamdhere, and K. C. Clayf, “Periscope: Unifying looking glass querying,” in Proc. PAM, 2016, pp. 177–189.

“BGP Looking Glass Databases.” Jan. 2020. [Online]. Available: http://www.bgplookingglass.com/

“PeeringDB API Documentation.” Jan. 2020. [Online]. Available: https://www.peeringdb.com/apidocs/

“PeeringDB.” Jul. 2021. [Online]. Available: https://www.peeringdb.com/

“traceroute.” Jan. 2020. [Online]. Available: http://traceroute.org/

“CAIDA AS Rank.” Jan. 2020. [Online]. Available: http://as-rank.caida.org/

M. Luckie, B. Huffaker, A. Dhamdhere, V. Giotsas, and K. C. Clayf, “AS relationships, customer cones, and validation,” in Proc. ACM IMC, 2013, pp. 243–256.

T. Böttger, C. Felix, and U. Steve, “Looking for hypergiants in peeringDB,” ACM SIGCOMM Comput. Commun. Rev., vol. 48, no. 3, pp. 13–19, 2018.

“Looking Glass—Hurricane Electric (AS6939).” [Online]. Available: https://lg.he.net/ (Accessed: May 2022).

“CAIDA: Archipelago (Ark) Measurement Infrastructure.” [Online]. Available: http://www.caida.org/projects/ark (Accessed: May 2022).

H. V. Madhyastha, T. Isdal, M. Piatek, and C. Dixon, “iPlane: An information plane for distributed services,” in Proc. USENIX OSDI, 2006, pp. 367–380.

“RIPStat Data API.” [Online]. Available: https://stat.ripe.net/docs/data_api#whois (Accessed: May 2022).

(Cisco, San Jose, CA, USA). IP Switching Cisco Express Forwarding Configuration Guide, Cisco IOS XE Release 3S. [Online]. Available: https://www.cisco.com/c/en/us/td/docs/ios-xml/ios/ipswitch_cef/configuration/xe-3s/isw-cef-xe-3s-book/isw-cef-load-balancing.html#GUID-D8A86BB9-FCA8-48CA-881D-153F4383728D (Accessed: May 2022).

A. Dhamdhere et al., “Inferring persistent interdomain congestion,” in Proc. ACM SIGCOMM, 2018, pp. 1–15.

“GitHub-jieliucl/BGP-M.” [Online]. Available: https://github.com/jieliucl/BGP-M (Accessed: May 2022).

Jie Li received the B.S. and M.S. degrees from the Dalian University of Technology, China, in 2011 and 2014, respectively, and the Ph.D. degree from the University College London in November 2021, where this research was conducted. He is currently a Senior Engineer with Huawei, China, with a focus on data center network. His research interests include network measurement, BGP routing, mobile social networks, socially aware networking, and ad hoc networks.

Vasileios Giotsas received the Ph.D. degree from University College London. He is a Lecturer with Lancaster University, where he leads the Networks area research of the Security Institute. His research focuses on network measurements and the analysis of the routing system.

Yangyang Wang received the B.S. degree in computer science and technology from Shandong University, China, in 2002, the M.S. degree from Capital Normal University, China, in 2005, and the Ph.D. degree from the Department of Computer Science, Tsinghua University, China, in 2013. He is currently a Research Assistant Professor, Institute for Network Sciences and Cyberspace, Tsinghua University. His current research interests include Internet routing architecture, SDN/NFV, and network measurement.

Shi Zhou (Senior Member, IEEE) received the B.Sc. and M.Sc. degrees in electrical and electronics engineering from Zhejiang University, China, and the Ph.D. degree in telecommunications from the Queen Mary University of London, University of London, U.K., in 2004. He is an Associate Professor with the Department of Computer Science, University College London, London, U.K. His research interests span the Internet routing and topology, complex networks, cybersecurity, cloud computing, and online social media analysis.