Anatomy of Multipath BGP deployment in a large ISP network

Jie Li
University College London
London, United Kingdom
jie.li@cs.ucl.ac.uk

Vasileios Giotsas
Lancaster University
Lancaster, United Kingdom
v.giotsas@lancaster.ac.uk

Shi Zhou
University College London
London, United Kingdom
s.zhou@ucl.ac.uk

***Abstract*—Multipath routing is useful for networks to achieve load sharing among multiple routing paths. Multipath BGP (M-BGP) is a technique to realize *inter-domain* multipath routing by enabling a BGP router to install multiple equally-good routes to a destination *prefix*. Most of previous works did not distinguish between intra-domain and inter-domain multipath routing. In this paper, we present a measurement study on the deployment of M-BGP in a large Internet service provider (ISP) network. Our method combines control-plane BGP measurements using Looking Glasses (LG), and data-plane traceroute measurements using RIPE Atlas. We focus on Hurricane Electric (AS6939) because it is a global ISP that connects with hundreds of major exchange points and exchanges IP traffic with thousands of different networks. And more importantly, we find that this ISP has by far the largest number of M-BGP deployments among autonomous systems with LG servers. Specifically, Hurricane Electric has deployed M-BGP with 512 of its peering ASes at 58 PoPs around the world, including many top ASes and content providers. We also observe that most of its M-BGP deployments involve IXP interconnections. Our work provides insights into the latest deployment of M-BGP in a major ISP network and it highlights the characteristics and effectiveness of M-BGP as a means to realize load sharing.**

***Index Terms*—Multipath BGP, Internet, Looking Glass, traceroute, multipath routing, RIPE Atlas, IXP**

I. Introduction

Multipath routing helps a network obtain higher capacity and performance through load balancing, improve the timeliness of their response to path changes, and enhance their resilience and security in the face of failures and attacks [1]. Various approaches have been proposed both to enable multipath routing [2], [3], and measure the deployment of multipath routes in the Internet [4]–[6]. However, most of the existing studies [4]–[6], [23], [25] have either focused on intra-domain routing, or did not distinguish between intra-domain and inter-domain links.

A key challenge with multipath inter-domain routing is to make the technique compatible with existing BGP semantics and BGP routers [1]. Today, most major router vendors, including Juniper, Cisco [31], and Huawei, support Multipath BGP (M-BGP) to enable *load sharing* between inter-domain paths of equal cost. Specifically, when a BGP router learns multiple eBGP paths from the same peering AS to a prefix with equal preference metrics (e.g. Weight and LocPref), length and MED values, it installs all of these paths together in the routing table instead of trying additional tie-breaking metrics. Load sharing can be realized per-destination using a hash of the IP headers. M-BGP differs from the other multipath routing techniques in that the multiple equally-good paths are learnt from the same peering AS; and they are for the same destination prefix, not for the same destination IP.

Aside from limited M-BGP approaches supported by the existing router deployments, the fact that the feature is optional and its application happens only for paths with no tie-breakers in the BGP path selection process, means that its actual deployment and impact on the inter-domain paths is obscure. The difficulty in measuring M-BGP paths has been exacerbated by the difficulties in pinpointing the inter-domain borders in traceroute paths. Despite over a decade of research in IP-to-AS mapping, accurate border mapping is still a challenge [7]–[12]. As a result, to the best of our knowledge there has not been measurement studies on the deployment of M-BGP.

In this paper, we present a first step toward this direction by implementing a measurement methodology that combines control-plane BGP measurements using Looking Glasses, and data-plane traceroute measurements over RIPE Atlas.

We focus on AS6939 (Hurricane Electric) because its LG server provides access to border routers across hundreds of Points-of-Presence where it establishes inter-domain connectivity and it also hosts active RIPE Atlas probes at overlapping locations. To avoid false positives, we apply strict rules on identifying the deployment of M-BGP. Hence our results provide a lower bound on AS6939's deployment of M-BGP.

Our findings reveal a wide deployment of M-BGP in AS6939. Overall it deploys M-BGP with 512 peering ASes at 58 PoPs – more than half of queried PoPs. We discover that most of its M-BGP deployments involve IXPs interconnections. 82.8% of the M-BGP deployments involve 2 inter-domain, alternative routes, 8.7% involve 3 routes, and 8.4% involve 4 routes. We have not observed any deployments involves more than 4 routes. Those with more than two routes typically involve large Content Provider Networks (CDNs), such as Apple, Cloudflare, or Microsoft.

We then execute a traceroute campaign to study the data-plane behavior of M-BGP load sharing for paths with overlap-

Jie Li thanks China Scholarship Council for the funding with No. 201406060022.

978-3-903176-27-0

ping locations of RIPE Atlas probes and LG vantage points. The traceroute data show that when M-BGP is deployed, the use of multiple inter-domain links is split almost equally between the number of IPs in the destination prefix. The egress link selected for each destination IP remains stable across our measurement period of 4 days indicating that the same per-flow load sharing algorithm is used across all border routers.

The techniques and results we present in this paper provide a first step toward developing a more thorough understanding of M-BGP deployment. We believe that our contributions are relevant to industry stakeholders, Internet engineers and researchers who can apply our techniques to assess the impact of M-BGP on performance, BGP dynamics and the routing behavior under conditions of stress.

II. Multipath Routing

Detection of different IP-level routing paths between a pair of hosts has been the basis for the study of 'anomalies' and 'routing dynamics [14]'. There was a considerable effort to characterise [15]–[18] and predict changing patterns of routing paths [19], [20]. While some of the observed different routing paths could indeed be due to anomalies or routing dynamics, it is now understood that many of them could be legitimate routes due to multipath routing [21], [22]. Indeed, in recent years network operators and service providers increasingly utilize multipath routing for traffic load balancing and load sharing to improve performance and resilience [2]. Multipath routing has attracted significant research attention, with proposals that span different layers, protocols, and techniques [3].

Augustin *et al.* [6] presented one of the first measurement studies of multipath routing by developing the Multipath Detection Algorithm (MDA) to identify diamond-shaped IP-level routing paths in traceroute data. Their work focused mostly within the boundary of a domain and as they remarked "*the traditional concept of a single network path between hosts no longer holds*". MDA was first proposed in [26] to detect multipath routing from a single source and a single destination. It uses Paris traceroute and adapts the number of probes to send hop by hop, in order to find as many load balancing behaviors as possible. It was then improved with a number of follow-up modifications [5], [27].

Recently, a number of research works have extended MDA to improve the completeness of load balancing identification in traceroute paths and reduce the measurement cost. Vermeulen *et al.* [23] introduced D-Miner to discover load-balanced paths at scale by utilizing the high-speed probing techniques of Yarrp [24]. Almeida *et al.* [25] proposed Multipath Classification Algorithm (MCA) to identify and classify load balancing in the Internet. Specifically, it extended the existing formalism and router model of [27], and the discovery techniques of [6] to capture paths that rely on arbitrary per-packet load balancing.

The multipath routing or load balancing behavior studied in the above-mentioned studies did not distinguish intra-domain from inter-domain links. Our work studies the load balancing on inter-domain links and in particular the deployment of Multipath-BGP, which differs from these multipath routing in installing multiple equally good paths to the same destination prefix learnt from the same peering AS.

In terms of inter-domain routing, Giotsas *et al.* introduced the Constrained Facility Search (CFS) algorithm, which relies on topology data from different levels of abstractions to map IP connectivity to PoPs [28]. Motamedi *et al.* [29] presented the mi^2 (mapping Internet interconnections) algorithm that improved PoP mapping through more accurate identification of inter-domain borders. Nur and Tozal [21] presented the cross-AS topology maps and defined the cross-border interfaces to study relevant topological properties. However, the above works focused on Internet mapping or topological properties, lacking of knowledge on how the diverse inter-domain connectivity is used in multipath routing. The closest work to ours is by Mok *et al.* [22], which studied the load-balancing behavior on inter-domain links by YouTube with data-plane data (i.e., traceroute data). Ours work focuses on studying the deployment of M-BGP with control-plane data provided by LG server for identifying M-BGP deployment and data-plane measurements for revealing more details.

III. Multipath BGP (M-BGP)

By default, BGP requires that for each prefix a single "best" path should be installed in the routing table to be used for traffic forwarding and be advertised to the BGP sessions [30]. To rank all the available paths BGP uses a multi-step decision process that examines a series of attributes in strict order. While the actual metrics may differ across different vendors, almost all major deployments consider the Local Preference, the AS path length and the Multi-Exit Discriminator (MED) values as part of their path selection process. Local Preference (`LocPref`) is a numerical value that can be set arbitrarily for each path to denote the preference of a route. The path with the highest `LocPref` value will be selected as the most preferable and will be installed in the routing table. `LocPref` is assigned locally to a router and is not propagated through BGP updates. If two routes have equal `LocPref` values, the path with the shortest path, namely the smallest number of AS hops, will be preferred. If both `LocPref` and path length cannot determine the best path, BGP continues the path selection process by checking the protocol through which a value is received, and then prefers paths with the lowest MED value if they are received from the same AS neighbor.

Although not defined in the original standard, most major router vendors, including Juniper, Cisco [31], and Huawei, have added optional support for multipath BGP in the case of Equal Cost Multipath Routing (ECMP). If multipath BGP is activated, when there are multiple equally good eBGP paths learnt from the *same peering AS*, and all the first six attributes of the BGP decision process (i.e., `LocPref`, AS Path, Origin, MED, eBGP/iBGP, and Metric) have the same value, instead of comparing the Router ID as a last-resort tie-breaker, multipath BGP allows the router to install more than one paths learnt from different border routers. The `maximum-paths` configuration controls the number of paths to be used.

core1.tor1.he.net> show ip bgp routes detail 142.46.150.1

Matching Routes	4
Status Codes	**A** - Aggregate **B** - Best **b** - Not Install Best **C** - Confederation eBGP **D** - Damped **E** - eBGP **H** - History **I** - iBGP **L** - Local **M** - Multipath **m** - Not Installed Multipath **S** - Suppressed **F** - Filtered **s** - Stale **x** - Best-External

Status	Network	Next Hop	Metric	LocPrf	Weight	Path	Origin	ROA
BMEx	142.46.150.0/24	198.32.181.46	0	100	0	19752	IGP	?
ME	142.46.150.0/24	206.108.34.48	0	100	0	19752	IGP	?
I	142.46.150.0/24	198.179.18.29	80	100	0	19752	IGP	?
E	142.46.150.0/24	198.32.181.50	0	100	0	6327, 19752	IGP	?
Last Update	26d21h41m39s ago (2 paths installed)							

Fig. 1. Example of LG response to the command of `show ip bgp route detail`

Load sharing can then happen per-destination using a hash of the IP headers, or per-packet using balanced or weighted round robin [32]. While per-packet load balancers have been found to be less frequent [6], [23], [25], their deployment may have been underestimated in the past [25]. The default M-BGP deployment uses a per-destination hash function, therefore M-BGP provides per-flow load sharing among different IP destinations in the same IP prefix. The amount of traffic or the available link capacities are not considered in the default load sharing functionality of M-BGP. Nonetheless, operators are able to override the default M-BGP behavior and implement either weighted load-sharing to reflect link capacities, or per-packet load balancing.

The studies on M-BGP are limited in literature. Valera *et al.* [1] explained the motivations to apply M-BGP and discussed some alternatives to M-BGP for achieving multipath routing. Therefore, while M-BGP is the de-facto technique to achieve load balancing between ASes, we lack insights with regards to the level of its deployment in the Internet.

IV. Looking Glass Announcements of M-BGP

Despite the extensive work in the enumeration of multipath routes in traceroute paths, distinguishing inter-domain from intra-domain multipath routing can be particularly challenging due to the difficulties in identifying the border routers between ASes. Traditionally, IP-to-AS mapping has been used to detect pairs of consecutive IP hops that belong in different ASes, and infer the AS border at these IPs. In recent years a number of border mapping techniques have found that such border identification can lead to inaccurate mapping since neighboring ASes may number their interfaces with IPs of neighboring ASes [11], [12]. Accordingly, novel border mapping techniques have been introduced to address these issues with bdrmapIT [10] considered as the state-of-the-art. However, recent works have found that even bdrmapIT can lead to erroneous border identification [33]. Therefore, identifying M-BGP through traceroutes alone can lead to a non-trivial amount of false-positives.

To alleviate this issue we utilize Looking Glasses which can provide a direct and reliable source of information on M-BGP deployment, since they allow to query directly the BGP configuration and routing table of border routers, and obtain BGP information beyond what is propagated through BGP updates in RouteViews and RIPE RIS collectors [34].

A. Looking Glass (LG) servers

Many network operators host LG servers, which provide Web-based interfaces to allow non-privileged execution of network commands (e.g., traceroute, ping, and BGP) at one or more border routers for network measurement and diagnosis [35]. LG servers enable researchers and network operators to study a network's performance from the perspective within the network. Different LG servers may provide different sets of commands.

In January 2020, more than 1,200 ASes have LG servers distributed across the world, including many top-ranked ASes [36], [37].

LG routing data, along with other data sources like RouteViews, have been widely used in studies on the Internet topology and path diversity [35], [38]–[40]. More recently the Periscope platform was proposed [41] to unify LG servers with publicly accessible querying API and to support on-demand measurements.

B. Identifying M-BGP in LG announcement

Some LG servers provide information on whether and how an AS implements M-BGP with its peering AS in responses to the command `show ip bgp detail <IP address>`.

Figure 1 shows an example response from `tor1`, a border router of AS6939. There are two different routes (the two 'Next Hops' 198.32.181.46 and 206.108.34.48) towards the same destination prefix (142.46.150.0/24). The two routes are labelled with status codes of "M" and "E", meaning these are multipath routes learnt via external BGP. Both routes have the same values for all routing metrics including LocPref, Weight and Path. Therefore, the LG output provides ground-truth on the routes that have M-BGP installed.

V. Case Study on AS6939 (Hurricane Electric)

To reveal more details on the M-BGP deployment, we conducted a thorough analysis of the AS6939 connectivity.

AS6939's LG lg.he.net covers border routers in 112 PoPs. As shown in Table I, these PoPs are located in 43 countries around the world. They support ping, traceroute, BGP route, BGP summary (IPv4) and BGP summary (IPv6). As a first step, we only study M-BGP on IPv4.

TABLE I
AS6939'S BORDER ROUTERS' GEOGRAPHICAL DISTRIBUTION.

	Number of border routers	
	overall	with M-BGP deployment
North America	**55**	**24**
United States	47	19
Canada	8	5
Europe	**40**	**26**
Germany	5	4
United Kingdom	3	2
France	2	2
Other	30	18
Asia	**6**	**4**
Other	**11**	**4**
Total	**112**	**58**

core1.tor1.he.net> show ip bgp summary

Local AS Number	6939
Number of Neighbors Configured	247, 229 up
Number of Routes Installed	3298347 (326536353 bytes)
Number of Routes Advertised	68081406 (4658509 entries) (223608432 bytes)
Number of Attribute Entries	680924 (61283160 bytes)

Neighbor Address	ASN	State	Time	Rt:Accepted
198.32.181.61	19551	ESTAB	483d20h26m	67
198.32.181.46	19752	ESTAB	570d 6h 5m	72
206.108.34.48	19752	ESTAB	47d 9h 8m	72
206.108.34.73	20161	ESTAB	92d22h 0m	105
206.108.34.164	20365	CONN	250d 5h53m	0
206.108.35.117	20473	ESTAB	92d21h56m	55
206.108.34.24	20940	ESTAB	74d12h54m	45
216.66.14.42	21513	ESTAB	260d 6h17m	17
206.108.34.102	21724	ESTAB	92d22h 0m	58
206.108.34.184	21834	CONN	185d21h56m	0
206.108.34.233	21834	CONN	305d22h40m	0
206.108.34.31	21949	ESTAB	56d 0h47m	147
209.51.168.70	22264	ESTAB	329d 4h 9m	1
206.108.35.7	22616	ESTAB	92d21h59m	3
206.108.35.8	22616	ESTAB	9d17h 3m	3
198.32.181.38	22634	ESTAB	403d 1h 4m	8

Fig. 2. Example of LG response to the command `show ip bgp summary`. Each red circle shows a peering AS with multiple neighbor addresses.

A. Identifying M-BGP

The LG command of `show ip bgp detail <IP address>` requires a target IP address as a parameter, so we need to firstly compile a list of potential targets.

We first query each of AS6939's border routers with the command `show ip bgp summary` to obtain the BGP connectivity of AS6939 at each of the corresponding locations. The command returns a summary table with the ASNs of the BGP neighbors and the addresses of the remote IP interface through which the BGP session is established. Figure 2 is an example table from the border router `tor1`, listing the information about each peering AS at this router. Cross-checking the neighbor address with peering AS with PeeringDB data can help us to know if the peering is deployed via IXP.

Figure 3 shows the number of peering ASes of AS6939's 112 border routers, which are ordered by their number of peering ASes. In total, AS6939 is peering with 5,868 unique ASes, of which 4,622 ASes are peered at 97 border routers via IXPs. This result highlights the role of IXPs in providing interconnection between AS peers. Note that a peering AS may be counted by a set of border routers.

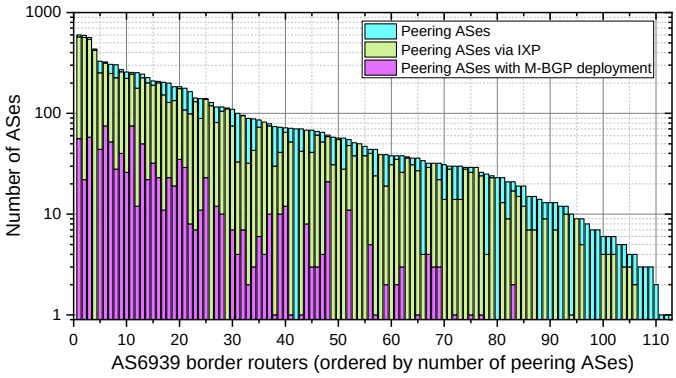


Fig. 3. AS6939's border routers ranked by the number of their peering ASes.

Firstly, we find out those peering ASes with multiple neighbor addresses in the bgp summary, such as AS19752, AS21834 and AS22616 shown in Figure 2, as this is a condition for two paths to have tied cost. We notice that all these ASes are peered via IXP.

For each of these peering ASes, we obtain the list of announced prefixes from BGP announcement provided by RouteViews in March 2020. For simplicity, we only study /24 prefixes because (1) /24 prefixes are the most common prefixes installed in BGP routing, e.g. around 60% of prefixes in the RouteViews data are /24 prefixes; and (2) more importantly, our purpose is to find any evidence of M-BGP deployment with a peering AS, where any of the peering AS's prefixes can provide sufficient evidence, regardless of its size.

Finally, we query each of AS6939's border routers using command `show ip bgp detail <IP address>` where `IP address` is set as `x.x.x.1` for each of the obtained prefixes. From each response we identify whether M-BGP is deployed with the peering AS at the border router towards the destination prefix as explained in Section IV-B. Our queries to a border router will stop if any prefix of a peering AS is identified as having M-BGP, since M-BGP would be activated for every prefix learnt through the same set of neighbor interfaces.

B. Results

Querying the LG server takes much longer time than we expected because we should avoid violating the querying rate limitation set by AS6939. By the time we write this paper, we have identified 950 M-BGP deployments by AS6939, at 58 border routers with 512 (around 9% in 5,868) peering ASes. Figure 3 plots in purple bar the number of peering ASes with M-BGP deployment at each border router. Note that a peering AS may be deployed with M-BGP for different prefixes at a set of border routers. Table I shows the 58 border routers are deployed with M-BGP around the world.

Figure 4 plots 509 of the peering ASes with M-BGP deployment, ranked by their customer cone sizes (in red). These ASes are in four groups by their AS ranks in CAIDA's AS Rank data [13], with the numbers of ASes in each group 22, 75, 52, and 360, suggesting AS6939 deploys M-BGP widely

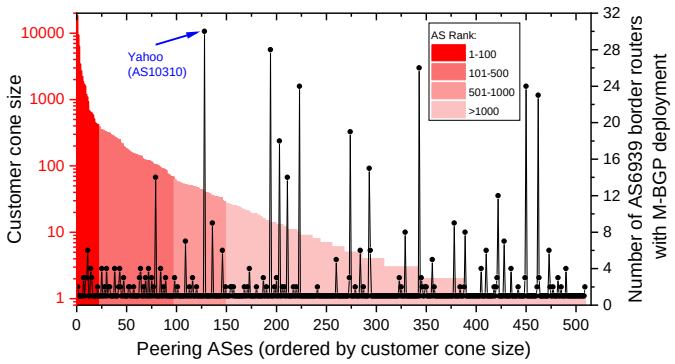


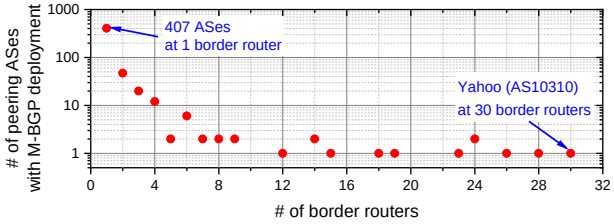
Fig. 4. The numbers of peering ASes, peering ASes via IXP, and peering ASes with M-BGP deployment at AS6939's border routers.

TABLE II
AS6939'S 10 HIGHEST RANKED PEERING ASES WITH M-BGP DEPLOYMENT ANNOUNCED BY AS6939'S BORDER ROUTERS.

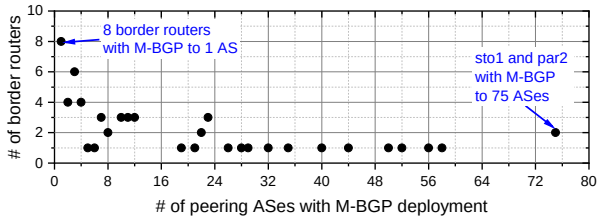
AS Rank	AS Number	Customer cone size	AS Name	Number of border routers
8	13101	18,372	ennit server GmbH	2
12	6461	9,368	Zayo Bandwidth	1
16	9002	6,366	RETN Limited	1
22	12389	3,415	PJSC Rostelecom	1
28	3216	2,691	PJSC "Vimpelcom"	1
33	6830	2,204	Liberty Global B.V.	1
40	8359	1,867	MTS PJSC	3
41	286	1,705	KPN B. V.	1
42	58453	1,601	China Mobile	3
51	41095	1,198	IPTP LTD	3

with its peering ASes among different rank groups. Note that there are 3 ($= 512 - 509$) ASes missing in the plot because the data snapshot [13] we used did not provide the information for them. The plot also shows the number of border routers where each peering AS is deployed with M-BGP (in black). We can observe from the figure that the low rank ASes are more likely to be deployed with M-BGP at multiple border routers, suggesting AS6939's richer connection to low-rank ASes than to top-rank ASes. Table II lists the 10 peering M-BGP ASes with the highest ranks.

Figure 5 shows the relation between the number of border routers and the number of peering ASes with M-BGP deployment. Figure 5(a) shows that 407 ASes are deployed with M-BGP at one border router, and 105 ($= 512 - 407$) ASes are deployed with M-BGP at multiple border routers. Among the peering ASes, Yahoo (AS10310) is deployed with M-BGP at the most (30) border routers, which is also labelled in Figure 4. Figure 5(b) shows that 8 border routers are deployed with M-BGP to only one peering AS, while 50 ($= 58 - 8$) border routers are deployed with M-BGP to multiple peering ASes. Among the border routers, `sto1` and `par2` are both deployed with M-BGP to the most (75) peering ASes.



(a) Number of peering ASes with M-BGP deployment as a function of number of border routers.



(b) Number of border routers as a function of number of peering ASes with M-BGP deployment.

Fig. 5. Relation between the number of border routers and the number of peering ASes with M-BGP deployment.

Among the 950 M-BGP deployments, 787 (82.8%) are with 2 inter-domain links, 83 (8.7%) are with 3 inter-domain and 80 (8.4%) are with 4 inter-domain links. Moreover, M-BGP paths with more than 2 inter-domain links are predominantly through large CDNs who have elevated capacity requirements. Our results confirm previous studies that found that the so-called Internet hyper-giants rely increasingly on IXPs as part of their content delivery backbone [53], [54].

In summary, our result suggests that AS6939 has deployed M-BGP widely, at more than half of its border routers with around 9% of its peering ASes distributed around the world. We confirm the vital role IXPs play in AS6939's peering fabric and deployment of M-BGP. Note that we only consider prefixes of length /24 provided by RouteViews and peering ASes with multiple neighbor addresses via IXP. Thus, our result provides a lower bound of AS6939's M-BGP deployment.

VI. TRACEROUTE MEASUREMENT OF AS6939'S M-BGP

This section introduces our traceroute measurement for revealing more details on AS6939's deployment of M-BGP.

A. RIPE Atlas traceroute measurement

Among the existing traceroute data or projects (e.g., RIPE Atlas [42], CAIDA Archipelago (Ark) [43] and iPlane [44]), we use RIPE Atlas for our traceroute measurement. RIPE Atlas has deployed probes within AS6939, which enables us to probe from AS6939 and ensures the traceroute paths will traverse the border routers.

At the time of our work, AS6939 had three RIPE Atlas probes (IPv4) being actively connected in the United States, Canada and Iran. Because of the hot-potato rule, we expected traceroute sources to be geographically close to some border routers with M-BGP deployment. We chose the probes in the

United States (in Milpitas, CA, near border router `sjc2` in San Jose, CA) and Canada (in Hamilton, near border router `tor1` in Toronto). We did not use the probe in Iran because it is geographically far away from any border routers with M-BGP deployment.

The destination prefixes are those identified with M-BGP deployment in the peering ASes to `sjc2` and `tor1`. We ran traceroute to each IP address (from `x.x.x.1` to `x.x.x.254`) in the destination prefix. Each source-destination pair is probed 50 times with 7-minute interval in February 2020. We use RIPE Atlas default settings, namely ICMP and Paris traceroute [45] variation 16.

B. IP-to-AS mapping

From traceroute raw data, we need to do IP-to-AS mapping to locate border routers. There are many existing methods (e.g. [10]–[12], [46]) and public data sets (e.g., [47]–[49]). For high confidence, we choose to use both bdrmapIT [10] and RIPEstat Data API (RIPEstat for short) [49]; and we only use results that are agreed by them.

bdrmapIT takes traceroute raw data as input, and integrates other data sets to conduct IP-to-AS mapping. The required data sets include prefixes announced by ASes (from RouteViews), IXP data (from PeeringDB [50]), AS relationship data, customer cone data and AS-to-organization data (from CAIDA [13], [51], [52]). The output of bdrmapIT is the AS that manages the router that an IP address (of an ingress interface) belongs to. bdrmapIT maps an IXP's IP to the AS of the next hop IP on the traceroute.

RIPEstat returns the AS that an IP belongs to. Normally, when an IP belongs to an IXP, RIPEstat does not provide an AS number for it. For example, according to RIPEstat, we obtain the IP-to-AS mapping result for an IP segment $IP1-IP2-IP3$ as $IP1(AS1)-IP2(?)-IP3(AS2)$. Therefore, it is highly likely that IP2 belongs to an IXP. We rely on IXP data from PeeringDB and the following process for confirmation.

1) If IP2 belongs to a member AS of an IXP, it is mapped to the member AS (normally AS2); otherwise, go to 2).
2) If IP2 belongs in an IXP's own prefix, it is mapped to the AS of the next hop IP (in this example IP2 is mapped to AS2). Otherwise, the mapping is failed and this traceroute data is discarded. Note that in our study the traceroute is carried out between AS1 and AS2 (i.e., AS6939 and one of its peering ASes), so it is impossible for IP2 to belong to a third member AS of the IXP.

Both bdrmapIT and RIPEstat can not map all IPs to ASes. In our study, they reach an agreement for 98% of the overlapped IPs that both can successfully map. We only use the overlapped and agreed IP-to-AS mapping result for our analysis.

If two consecutive IPs on a traceroute path are mapped to different ASes, we use these IPs to represent (logically) an inter-domain border link, connecting between two border routers of two peering ASes, which are called nearside AS and farside AS.

C. Two types of M-BGP deployment

All the M-BGP deployment studied in this paper are established through IXPs. Thus, a relevant traceroute path traverses firstly the nearside IP in AS6939, then an IP in an IXP and finally a farside IP in the farside AS, where the IXP's IP is mapped to the farside AS as the result of IP-to-AS mapping used in this paper (see Section VI-B). This means the Next Hops or the Neighbor Addresses in the response of LG commands are actually IP addresses of interfaces of the IXP sitting between the nearside AS and the farside AS.

Based on the traceroute data, we classify the identified M-BGP deployment into two types: parallel and divergent. Table III lists the details of four cases, two in each type. Cases 1-3 are all from the same source IP (i.e. the same RIPE Atlas probe), hence the same border router and the same nearside IPs. Case 4 is from a different source IP (i.e. another RIPE Atlas probe), thus a different border router and a different nearside IP. We illustrate the topology and traffic for Cases 1 and 3 in Figures 6-7, separately.

Table III shows two cases of the parallel type M-BGP, in which each of the two IXP IPs is followed by a single farside IP. In Case 1, traffic enters the border router `sjc2` via two nearside IPs; and then exits the border router and enters a geographically nearby IXP (Equnix San Jose) via two IXP IPs with equal probabilities. The traffic from each IXP IP is forwarded to one of the two links between IXP and farside AS. There is no cross traffic, i.e. there are only two unique paths between the nearside and the farside and traffic does not mix in the IXP.

Figure 6 illustrates the topology and routing of Case 1. The figure shows that the traffic is already split before entering `sjc2`. We believe this is caused by intra-domain load sharing and has no impact on M-BGP deployment because the traffic from either ingress interface of `sjc2` is forwarded to the two IXP IPs equally, indicating a full mesh between ingress interfaces and egress interfaces of `sjc2`.

Figure 6(b) also shows the destination IPs. This suggests two important observations. Firstly, each IXP IP is used for traffic to half of destination IPs. And secondly, the choice of IXP IP for each destination IP is permanent. That is, if an IXP IP is chosen for traffic to a particular destination IP, this IXP IP will always be used for all future traffic to that destination. This is exactly the kind of routing property expected from M-BGP. The same can be observed for the other cases.

Cases 3 and 4 in Table III are both divergent type, in which each IXP IP is followed by multiple farside IPs. We take Case 3 as an instance with its topology and traffic shown in Figure 7. In this case, traffic again exits `sjc2` and enters Equinix San Jose via two IPs. Each IXP IP is used for traffic to half of IP addresses in the destination prefix. Traffic from each IXP IP is then split onto 3 different links between the IXP and the farside AS with similar proportions.

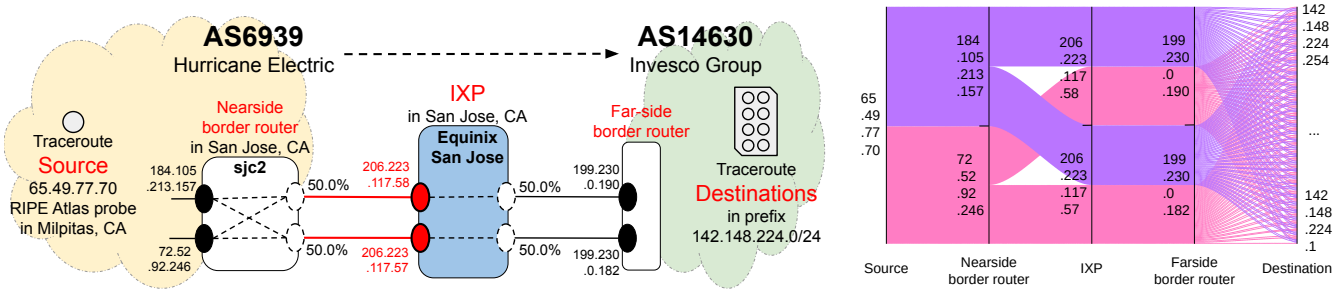
VII. Discussion

This paper reports our study on the deployment of M-BGP in a large ISP network of AS6939 or Hurricane Electric.

TABLE III

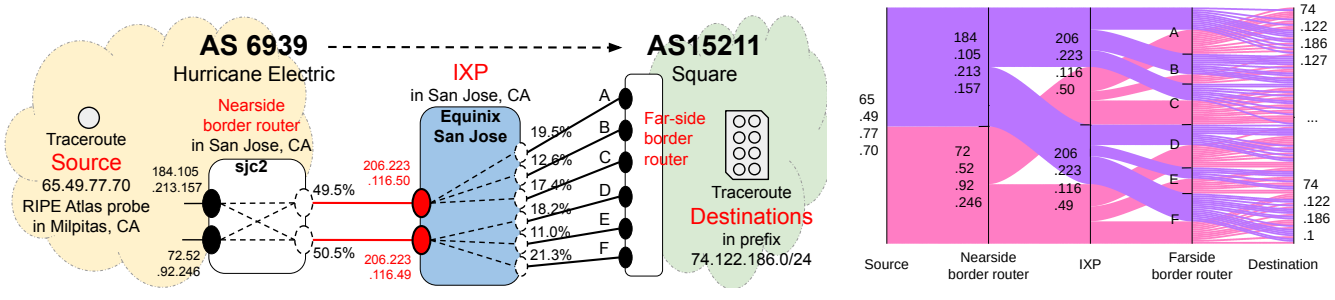
CASES OF THE TWO TYPES OF M-BGP DEPLOYED IN AS6939. IN EACH CASE, WE COLLECTED TRACEROUTE MEASUREMENT FROM A RIPE ATLAS PROBE WITHIN AS6939 (SOURCE) TO EACH IP IN A /24 PREFIX IN A PEERING AS (DESTINATION). THE TRACEROUTE DATA REVEALED THE NEARSIDE IPS (I.E. INGRESS INTERFACES OF NEARSIDE BORDER ROUTER), IXP IPS, FARSIDE IPS AND THE ALLOCATION OF ROUTES (%) TO ALL DESTINATION IPS ON EACH INTER-DOMAIN LINKS. FIGS. 6 AND 7 ILLUSTRATE THE TOPOLOGY AND ROUTING FOR CASES 1 AND 3.

M-BGP Type	Case No.	Traceroute source	Border router	Nearside IPs	IXP name	IXP IPs: routes%	Farside IPs: routes%	Destination: ASN & prefix
Parallel	1	65.49.77.70	sjc2	184.105.213.157	Equinix San Jose	206.223.117.58: 50.0%	199.230.0.190: 50.0%	AS14630 142.148.224.0/24
				72.52.92.246		206.223.117.57: 50.0%	199.230.0.182: 50.0%	
	2	65.49.77.70	sjc2	184.105.213.157	Equinix San Jose	206.223.117.18: 50.0%	64.16.254.8: 48.4%	AS63440 192.76.120.0/24
				72.52.92.246		206.223.116.110: 50.0%	64.16.254.2: 50.0%	
Divergent	3	65.49.77.70	sjc2	184.105.213.157	Equinix San Jose	206.223.116.50: 49.5%	A: 74.122.191.5 : 19.5%	AS15211 74.122.186.0/24
							B: 74.122.191.25: 12.6%	
							C: 74.122.191.35: 17.4%	
				72.52.92.246		206.223.116.49: 50.5%	D: 74.122.191.7 : 18.2%	
							E: 74.122.191.27: 11.0%	
							F: 74.122.191.37: 21.3%	
	4	209.51.186.5	tor1	209.51.161.49	Equinix Toronto	198.32.181.46: 50.3%	142.47.202.50: 25.1%	AS19752 142.46.150.0/24
							142.47.203.14: 25.2%	
					TorIX	206.108.34.48: 49.7%	142.47.202.50: 24.0%	
							142.47.203.14: 25.7%	



(a) Topology (b) Routing

Fig. 6. Illustrations of topology and routing of a Parallel-type M-BGP deployment between AS6939 and AS14630 (Case 1 in Table III).



(a) Topology (b) Routing

Fig. 7. Illustrations of topology and routing of a Divergent-type M-BGP deployment between AS6939 and AS15211 (Case 3 in Table III)

We show that M-BGP is widely deployed by AS6939 with hundreds of its peering ASes at more than half of its border routers around the globe. We observe that most of its M-BGP deployments involve IXP interconnections. All of our datasets are freely available at a GitHub repository [55].

Since we only make queries to a limited number of prefixes that belong to each of AS6939's peering ASes, our result provides a lower bound of AS6939's deployment of M-BGP with its peering ASes. We focused on AS6939 due to its very high centrality in the Internet routing system. According to CAIDA's AS Rank [13], AS6939 is the 7th largest AS in terms of customer cone size and provides transit between more than 8k ASes (12% of ASNs in the global routing table). Additionally, AS6939 is a network of very high peering affinity, with more than 6k peers and presence in 236 IXPs (most than any other AS). Therefore, the extent of AS6939 connectivity combined with its data transparency through the provided LG make it an ideal vantage point for understanding the deployment of M-BGP in the Internet and evaluating the proposed measurement techniques.

As part of our ongoing work, we are applying our technique to a much wider list of ASes hosting LG servers that support the required commands in order to provide a more extensive view of M-BGP deployments in the wild. Note that the M-BGP deployments presented in this paper are all via IXP and multiple inter-domain links. Our preliminary findings from a wider set of vantage points revealed M-BGP deployments using single inter-domain links or direct private peerings, and cases of multipath BGP routes with paths of unequal lengths. In addition we are expanding our traceroute measurements to evaluate the efficacy of MDA in discovering these M-BGP paths and to reveal potential non-canonical M-BGP deployments that use per-packet load balancing.

We believe that the measurement, characterization and analysis of M-BGP is of particular interest both for network practitioners and Internet researchers. The potential of M-BGP in improving the performance, stability and resilience of inter-domain paths, has not been yet thoroughly studied and understood. Therefore, our work can inform and enable the necessary measurement studies to illuminate this crucial aspect of BGP.

ACKNOWLEDGMENT

The authors would like to give special thanks to the anonymous reviewers for their constructing comments on the improvement of this paper.

REFERENCES

[1] F. Valera, I. Van Beijnum, A. Garcia-Martinez, and M. Bagnulo, "Multipath BGP: Motivations and solutions," in Next-Generation Intenet Architectures and Protocols, B. Ramamurthy, G. N. Rouskas, and K. M. Sivalingam, Ed. Cambrige, UK: Cambridge Univ. Press, 2011.

[2] S. K. Singh, T. Das, and A. Jukan, "A survey on Internet multipath routing and provisioning," IEEE Commun. Surv. Tutor. vol. 17, no. 4, pp. 2157—2175, fourthquarter 2015.

[3] 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. Surv. Tutor. vol. 17, no. 4, pp. 2176-–2213, fourthquarter 2015.

[4] B. Augustin, T. Friedman, and R. Teixeira, "Measuring load-balanced paths in the Internet," in Proc. ACM IMC'07, pp. 149–160.

[5] K. Vermeulen, Stephen D. S., O. Fourmaux, and T. Friedman, "Multi-level MDA-Lite Paris traceroute," in Proc. ACM IMC'18, 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, June 2011.

[7] R. Motamedi, R. Rejaie, and W. Willinger, "A survey of techniques for Internet topology discovery," IEEE Communications Surveys & Tutorials 17, no. 2 (2014): 1044-1065.

[8] B. Huffaker, A. Dhamdhere, M. Fomenkov, and k claffy, "Toward topology dualism: Improving the accuracy of AS annotations for routers," in PAM'10, A. Krishnamurthy and B. Plattner (Eds.). Springer International Publishing, pp. 101–110.

[9] J.-J. Pansiot, P. Mérindol, B. Donnet, and O. Bonaventure, "Extracting intra-domain topology from mrinfo probing," in PAM'10, A. Krishnamurthy and B. Plattner (Eds.). Springer International Publishing, pp. 81–90.

[10] A. Marder, M. Luckie, A. Dhamdhere, B. Huffaker, kc claffy, and J. M. Smith, "Pushing the boundaries with bdrmapIT: Mapping router ownership at Internet scale," in Proc. ACM IMC'18, pp. 56-–69.

[11] M. Luckie, A. Dhamdhere, B. Huffaker, D. Clark, and kc claffy, "Bdrmap: Inference of borders between IP networks," in Proc. ACM IMC'16, pp. 381—396.

[12] A. Marder and J. M. Smith. "MAP-IT: Multipass accurate passive inferences from traceroute," in Proc. ACM IMC'16, pp. 397-–411.

[13] CAIDA AS Rank, http://as-rank.caida.org/. (February 2020).

[14] N. Ahmed and K. Sarac, "An experimental study on inter-domain routing dynamics using IP-level path traces," in Proc. IEEE ICN'15, pp. 510—517.

[15] G. Comarela, G. Gürsun, and M. Crovella, "Studying interdomain routing over long timescales," in Proc. ACM IMC'13, pp. 227—234.

[16] R. Fanou, P. Francois, and E. Aben, "On the diversity of interdomain routing in Africa," in PAM'15, J. Mirkovic and Y. Liu (Eds.). Springer International Publishing, pp. 41-–54.

[17] A. Medem, C. Magnien, and F. Tarissan, "Impact of power-law topology on IP-level routing dynamics: Simulation results," in Proc. IEEE INFOCOM'12, pp. 220—225.

[18] M. Rimondini, C. Squarcella, and G. Di Battista, "Towards an automated investigation of the impact of BGP routing changes on network delay variations," in Proc. PAM'14, pp. 193-–203.

[19] I. Cunha, R. Teixeira, D. Veitch, and C. Diot, "DTRACK: A system to predict and track Internet path changes," IEEE/ACM Trans. Netw. vol. 22, no. 4 pp. 1025—1038, Aug. 2014.

[20] S. Wassermann, P. Casas, T. Cuvelier, and B. Donnet, "NETPerfTrace: Predicting Internet path dynamics and performance with machine learning," in Proc. ACM Big-DAMA'17, pp. 31-–36.

[21] A. Y. Nur, and M. E. Tozal, "Cross-AS (X-AS) Internet topology mapping," Comput. Netw. vol. 132, pp. 53-–67, 2018.

[22] R. K. P. Mok, V. Bajpai, A. Dhamdhere, and K. C. Claffy, "Revealing the load-balancing behavior of YouTube traffic on interdomain links," in PAM'18, R. Beverly, G. Smaragdakis, and A. Feldmann (Eds.), pp. 228—240.

[23] K. Vermeulen, J. P. Rohrer, R. Beverly, O. Fourmaux and T. Friedman, "Diamond-Miner: Comprehensive discovery of the Internet's topology diamonds," in Proc. USENIX NSDI'20, to appear.

[24] R Beverly,"Yarrp'ing the Internet: Randomized high-speed active topology discovery," In Proceedings of the 2016 Internet Measurement Conference, pp. 413-420. 2016.

[25] R. Almeida, I. Cunha, R. Teixeira, D. Weithc and C. Diot, "Classification of load balancing in the Internet," in Proc. IEEE INFOCOM'20, to appear.

[26] B. Augustin, T. Friedman, and R. Teixeira, "Multipath tracing with Paris traceroute," in Proc. IEEE E2EMON'07, pp. 1–8.

[27] D. Veitch, B. Augustin, R. Teixeira, and T. Friedman, "Failure control in multipath route trace," in Proc. IEEE INFOCOM'10, pp. 1395–1403.

[28] V. Giotsas, G. Smaragdakis, B. Huffaker, M. Luckie and kc, claffy, "Mapping peering interconnections to a facility," in Proc. ACM CoNEXT'15, Article No. 37.

[29] R, Motamedi, B. Yeganeh, B. Chandrasekaran, R. Rejaie, BM. Maggs, W. Willinger. "On mapping the interconnections in today's Internet," IEEE/ACM Trans. Netw., vol. 27, no. 5, pp. 2056–2070, 2019.

[30] Y. Rekhter, T. Li, and S. Hares, "A border gateway protocol 4 (BGP-4)," RFC 4271, January 2006.

[31] BGP Best Path Selection Algorithm – CISCO, https://www.cisco.com/c/en/us/support/docs/ip/border-gateway-protocol-bgp/13753-25.html#anc5

[32] Halabi B, Halabi S, McPherson D. "Internet routing architectures," Cisco Press, 2000.

[33] 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'19, pp. 202–216.

[34] University of Oregon Route Views Project, http://www.routeviews.org/. (February 2020).

[35] A. Khan, T. T. Kwon, H.-C. Kim, and Y. Choi, "AS-level topology collection through looking glass servers," in Proc. ACM IMC'13, pp. 235–241.

[36] BGP Looking Glass Databases, http://www.bgplookingglass.com/. (January 2020).

[37] PeeringdB API Documentation, https://www.peeringdb.com/apidocs/. (January 2020).

[38] H. Chang, R. Govindan, S. jamin, S. J. Shenker, and W. Willinger, "Towards capturing representative AS-level Internet topologies," Computer Networks, vol. 44, pp. 737–755, 2004.

[39] B. Zhang, R. Liu, D. Massey, and L. Zhang, "Collecting the Internet AS-level Topology," ACM SIGCOMM CCR, vol. 35, no. 1, pp. 53–62, 2005.

[40] J. Han, D. Watson, and F. Jahanian, "An experimental study of Internet path diversity," IEEE Trans. Dependable and Secure Computing, vol. 3, no. 4, pp. 273–288, 2006.

[41] V. Giotsas, A. Dhamdhere, and kc claffy, "Periscope: Unifying looking glass querying," in Proc. PAM'16, pp. 177–189.

[42] RIPE NCC Staff, "RIPE Atlas: A global Internet measurement network," The Internet Protocol Journal. vol. 18, no. 3 pp. 2—26, 2015.

[43] CAIDA: Archipelago (Ark) Measurement Infrastructure, http://www.caida.org/projects/ark/. (December 2018).

[44] H. V. Madhyastha, T. Isdal, M. Piatek, and C. Dixon, "iPlane: An information plane for distributed services," in Proc. USENIX OSDI'06, pp. 367-–380.

[45] B. Augustin, X. Cuvellier, B. Orgogozo, F. Viger, T. Friedman, M. Latapy, C. Magnien, and R. Teixeira, "Avoiding traceroute anomalies with Paris traceroute," in Proc. ACM IMC'06, pp. 153—158.

[46] A. Faggiani, E. Gregori, A. Improta, L. Lenzini, V. Luconi, and L. Sani, "A study on traceroute potentiality in revealing the Internet AS-level topology," in 2014 IFIP Networking Conference, pp. 1-–9.

[47] MaxMind: IP Geolocation and Online Fraud Prevention, https://www.maxmind.com

[48] Team Cymru, http://www.team-cymru.com

[49] RIPEstat Data API, https://stat.ripe.net/docs/data_api#whois

[50] PeeringDB, https://www.peeringdb.com/

[51] M. Luckie, B. Huffaker, A. Dhamdhere, V. Giotsas, and kc claffy, "AS relationships, customer cones, and validation," in Proc. ACM IMC'13, pp. 243—256.

[52] The CAIDA UCSD AS to Organization Mapping Dataset, <20200101> http://www.caida.org/data/as_organizations.xml

[53] T. Böttger, C. Felix, and U. Steve, "Looking for hypergiants in peeringDB," ACM SIGCOMM Computer Communication Review 48, no. 3 (2018): 13-19.

[54] T. Böttger, C. Félix, T. Gareth, C. Ignacio, and U. Steve. "A Hypergiant's View of the Internet," ACM SIGCOMM CCR 47, no. 1 (2017).

[55] jieliucl/M-BGPDeployment, https://github.com/jieliucl/M-BGPDeployment.