A First Measurement with BGP Egress Peer Engineering

Ryo Nakamura[1(✉)], Kazuki Shimizu[2], Teppei Kamata[3], and Cristel Pelsser[4]

[1] The University of Tokyo, Tokyo, Japan
upa@nc.u-tokyo.ac.jp
[2] Juniper Networks K.K., Tokyo, Japan
kshimizu@juniper.net
[3] Cisco Systems G.K., Tokyo, Japan
tkamata@cisco.com
[4] University of Strasbourg, Strasbourg, France
pelsser@unistra.fr

Abstract. This paper reports on measuring the effect of engineering egress traffic to peering ASes using Segment Routing, called BGP-EPE. BGP-EPE can send packets destined to arbitrary prefixes to arbitrary eBGP peers regardless of the BGP path selection. This ability enables us to measure external connectivity from a single AS in various perspectives; for example, does the use of paths other than the BGP best paths improve performance? We conducted an experiment to measure latency to the Internet from an event network, Interop Tokyo ShowNet, where SR-MPLS and BGP-EPE were deployed. Our findings from the experiment show BGP-EPE improves latency for 77% of target prefixes, and peering provides shorter latency than transit. We further show factors on which the degree of improvement depends, e.g., the performance-obliviousness of BGP and the presence of remote peering. Also, we find 91% of peer ASes forwarded packets towards prefixes that the peers did not advertise.

Keywords: BGP egress peer engineering · Segment routing · Internet latency

1 Introduction

Since latency over the Internet impacts the quality of experiences (QoE) of users [13,23], reducing the latency is a fundamental challenge on the Internet. A portion of the challenge is to optimize path selection at inter-AS connections. The BGP path selection algorithm does not care about performance, for example, it sometimes chooses paths based on freshness [11] or IP addresses of neighbors [27]. Thus, many studies have developed alternate routing systems to outperform BGP [5,6,19,28,33]. They steer egress traffic from their own ASes to peer ASes based on performance and link capacity in contrast to BGP.

O. Hohlfeld et al. (Eds.): PAM 2022, LNCS 13210, pp. 199–215, 2022.
https://doi.org/10.1007/978-3-030-98785-5_9

Egress traffic engineering is not an emerging topic. The LOCAL_PREF attribute of BGP is a primitive mechanism to choose specific paths for egress traffic [16,28], and a Locator/Identifier Separation Protocol-based method has been proposed [14]. Segment Routing (SR) [18] is also one of the techniques to steer egress traffic to specific peers, called BGP Egress Peer Engineering (BGP-EPE) [17]. SR-based BGP-EPE steers specific traffic to given peers using encapsulation; therefore, it is independent of underlying BGP and IGP policies, unlike LOCAL_PREF. Moreover, SR and BGP-EPE are standardized and have been implemented on well-matured commercial routers [32]. These characteristics enable us to measure the potential benefit of egress traffic engineering without impacting other traffic on a real network where SR is deployed.

To the best of our knowledge, this paper reports the first measurement result using SR-based BGP-EPE deployed on a real network. The aim of this measurement is to clarify the potential benefit of egress traffic engineering from a latency perspective and not to propose a methodology to find better egress paths. The effectiveness of egress traffic engineering is still controversial [7], and this paper provides a case study to this argument. We conducted our experiment at an AS where SR-based BGP-EPE was deployed, via 45 unique ASes (3 transit ASes and 43 peer ASes, of which one peer AS is identical to a transit AS). We performed ping and traceroute to addresses spread on the IPv4 address space. The AS where we conducted the measurement is Interop Tokyo ShowNet [3], an event network built for and operated during a technology exhibition in Japan. Therefore, the measurement period was short—from April 12 to 16 in 2021; however, in summary, this paper shows the following findings:

- BGP-EPE reduces latency for 77% of target prefixes, but the gain depends on the BGP best path selection at the measurement point.
- Peering tends to have shorter latency than transit, and this trend increases on inter-continental paths.
- 91% of ASes peering with ShowNet allowed detouring toward prefixes that the ASes did not advertise.

2 Methodology

2.1 Segment Routing and BGP Egress Peer Engineering

Segment Routing (SR) [18] is an emerging source routing architecture implemented in recent commercial routers. The concept of SR is to represent any topological entities as *segments*, and SR-capable routers perform packet forwarding by segments embedded in packets. A segment can indicate, e.g., an IGP node or adjacency between routers. A list of segments in a packet informs routers where the packet needs to flow. SR currently leverages two data planes: MPLS (SR-MPLS) and IPv6 (SRv6). A segment is identified by an MPLS label in SR-MPLS and by an IPv6 address in SRv6. These identifiers are called Segment Identifiers (SIDs). Since we used SR-MPLS for the measurement, we focus on SR-MPLS in the rest of this paper.

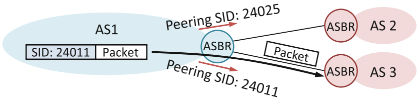


Fig. 1. Overview of BGP-EPE.

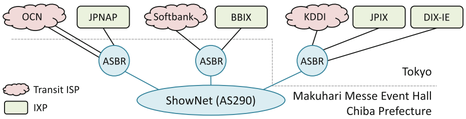


Fig. 2. External diagram of ShowNet.

SR is one of the techniques to steer egress traffic to specific peers. This capability is called BGP Egress Peer Engineering (BGP-EPE) and has been standardized in RFC9087 [17]. BGP-EPE-capable routers represent eBGP peers as segments called BGP Peering Segments or SIDs. Figure 1 illustrates an overview of SR-based BGP-EPE. The AS border router (ASBR) of AS 1 has two eBGP peers and assigns BGP Peering SIDs for them. When the ASBR receives a packet encapsulated with the SID, the router pops the MPLS header and then transmits the IP packet to the peer corresponding to the SID. Because SR-based BGP-EPE is a mechanism to carry packets toward selected peers, a methodology to choose better peers for given destinations is needed to benefit from egress traffic engineering. However, the methodology is not the focus of this paper. Instead, we measured all possible paths to clarify the potential latency benefit.

2.2 Experimental Environment

Interop Tokyo ShowNet. We conducted the measurement experiment with the SR-based BGP-EPE at Interop Tokyo ShowNet [3]. Interop Tokyo [2] is a large annual exhibition of network technologies in Japan, and ShowNet (AS290) is an event network built at Interop Tokyo. In 2021, Interop Tokyo was held from April 14 to 16, and ShowNet was built in the event hall. The network provided Internet connectivity for exhibitors and visitors at the event, and simultaneously demonstrated new technologies, and conducted several inter-operability tests. The ShowNet backbone in 2021 was composed of SRv6 L3VPN and SR-MPLS, on which we put the measurement experiment.

Figure 2 illustrates a simplified diagram of ShowNet focusing on BGP-EPE. ShowNet had three ASBRs, ASR9902 from Cisco Systems, MX204 from Juniper Networks, and NE8000-X4 from Huawei, connected to one major Japanese Internet Service Provider (ISP) and one or two Internet eXchange Points (IXPs) each. The ASBRs assigned BGP Peering SIDs to transit providers and peer ASes who agreed to join the experiment. ShowNet started peering at the IXPs on the evening of April 12. Eventually, we performed the measurement over 101 eBGP peers of 45 unique ASes listed in Table 1.

Table 1. Number of eBGP peers and ASes involved in the measurement.

	JPNAP	BBIX	JPIX	DIX-IE	Transit	Total
# of eBGP peers	25	32	34	6	4	101
# of ASes	24	26	30	6	3	45 (unique)

Measurement Procedure. The high-level measurement procedure is simple: a Linux server deployed at ShowNet performed ping and traceroute to targeted IPv4 addresses via each eBGP peer by SR-based BGP-EPE. We used scamper [24] for ping and traceroute; ping is performed over UDP with a probe count of five, and traceroute using UDP-paris [8].

We leveraged two techniques in addition to SR-based BGP-EPE: BGP Link State (BGP-LS) [29] to automate detecting new peer ASes and Linux network namespace to measure multiple egress ASes in parallel. Peering with ShowNet was automated and ASBRs dynamically assigned BGP Peering SIDs for the peers. Therefore, the measurement server also dynamically learned the SIDs through BGP-LS using cRPD [22] from Juniper Networks. To accomplish extensive measurements at the ephemeral event network, we ran 8 scamper processes parallel at the server. Each scamper ran on a Linux network namespace, which has a separated routing table, with a hardware-assisted virtual interface by Single Root I/O Virtualization [26]. In this setup, we executed the following procedure in succession: configuring a namespace to encapsulate packets with appropriate SIDs for a peer, and spawning scamper on the namespace.

The target IP addresses of the measurement are 2,638,382 IPv4 addresses. These addresses were extracted from the source IPv4 addresses in ICMP Echo Replies from a packet trace of MAWI [12] on April 8 and 9, 2020[1]. We kept the IPv4 addresses that responded to ping at the end of March 2021. Finally, we extracted one address per /24 prefix. As a result, we obtained 2.6M target addresses. Note that the trace contains reply packets for probes sent from a host of USC's ANT Project [30][2]. Hence, we indirectly used a partial list of theirs.

2.3 Ethical Considerations

As presented in the previous section, our traffic analysis did not consider user traffic. Our experiment involves sending packets destined to the target addresses even when a peer does not advertise their prefixes. It is obviously not an intended behavior on peering links. Therefore, we arranged an agreement with ASes before peering. The agreement states that the purpose of this experiment is measurement and demonstration of SR-based BGP-EPE, the experiment sends UDP or ICMP probe packets having arbitrary destinations to ShowNet peers, the amount of measurement traffic to a peer is up to 15 Mbps (actual throughput was about 3.5 Mbps), and the measurement results are used only for sharing information with network communities and research purposes.

3 Data

Our data was collected at the event network for five days; therefore, the analysis in this paper is a point-in-time, not a longitudinal, analysis. On the other hand,

[1] We used an unanonymized version of the trace with a responsible person's consent.

[2] MAWI Traffic Archive FAQ, https://mawi.wide.ad.jp/mawi/faq.html.

the number of samples we obtain is reasonable with respect to the length of the experiment. This section describes the data we collected and its preprocessing.

In our experiments, a single scamper process performs ping or traceroute to 2.6M target addresses via an eBGP peer chosen from the 101 eBGP peers. Eventually, 688 ping iterations and 176 traceroute iterations finished during the period. Thus, we have results of six or seven ping and one or two traceroute iterations for each eBGP peer.

The probe count for ping is five; therefore, at most we obtain 35 RTTs for a destination address via an egress eBGP peer. Because small samples cause statistical errors, we omit ping results with less than 10 RTTs for a pair from further analysis. Figure 3 shows the histogram of numbers of measured RTTs to a destination address via an eBGP peer. About 25% of pings failed, and 28.6% have less than 10 RTTs. After the exclusion of pairs with less that 10 RTTs, the number of remaining target IP addresses is 2,303,253. Through all the ping measurements, we sent 9,062,842,170 probes and received 6,027,615,205 replies. Thus, its success rate is 66.51%. A reason for the low success rate is that it includes ping probes to non-advertised prefixes via peering ASes. The ping success rate via the transit ASes is 77.58%, and via peering ASes is 66.18%. Further analysis on filtering packets to unintended destinations at peering ASes is described in Sect. 4.3.

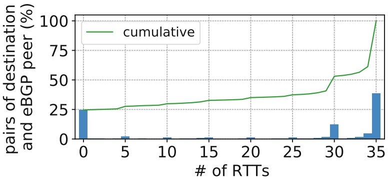


Fig. 3. Histogram of numbers of measured RTTs to a destination address via an egress eBGP peer.

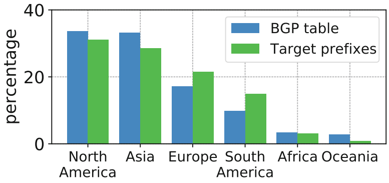


Fig. 4. Percentages of prefixes per region in the BGP table versus in our the target prefixes.

Besides, because the selection process of the target IPv4 addresses does not consider the distribution of the addresses on the IPv4 address space, we summarize the measured RTTs by prefixes in the BGP table. We treat RTTs to a target IP address as RTTs to a prefix that includes the address. The remaining 2.3M target addresses cover 38.06% of prefixes in the BGP table of an ASBR at ShowNet as of April 16, 2021, which is the last day of the measurement. We call these prefixes having measured RTTs target prefixes.

To verify the diversity in the target prefixes and assess a potential bias, we compare these with all prefixes in the BGP table from a regional perspective. Mapping of prefixes and regions is generated from AS number allocation: (1) obtaining prefix-to-ASN mapping from the BGP table, and (2) obtaining ASN-to-region mapping from Geoff Huston's page [20]. Figure 4 shows the percentages of prefixes in each region. As shown, the distribution of target prefixes is similar to the BGP table on a regional allocation basis. The differences are not significant (less than 6% in all regions). Thus, we conclude that the target prefixes properly represent the diversity of prefixes in the BGP table.

4 Analysis

In this section we analyze the data collected during our experiment. Section 4.1 provides a case study of how BGP-EPE improves latency against the BGP best paths. Section 4.2 compares latency through peering and transit. Finally, Sect. 4.3 reveals behaviors of ASes when they receive packets having destinations that the ASes do not advertise. Note that information that can reveal specific ASes, e.g., AS numbers and AS names, is anonymized because of a non-disclosure agreement at ShowNet. ShowNet is sponsored by equipment vendors, transit service providers and connections at IXPs; therefore, we cannot expose matters that are disadvantageous to the contributors.

4.1 How Best are the Best Paths

Reducing latency by steering egress traffic away from the BGP best path is a fundamental ability of BGP-EPE. We first clarify the room for latency improvement by comparing RTTs via the best paths and alternative paths. For this comparison, we extracted the minimum RTTs to each target prefix via the best paths versus all received paths. We call paths that achieved the minimum RTTs among all received paths *alternate paths* in accordance with the previous literature [7,28]. We choose the minimum, not median or average, because of the small number of samples. Appendix A shows the results with the median values.

Figure 5 shows CDF of RTTs to the target prefixes via the best paths (hereafter called $RTT(best)$) and alternate paths (hereafter called $RTT(alternate)$). The graph demonstrates alternate paths achieve better latency than the best paths, as expected. The difference increases when RTT is over about 100 ms.

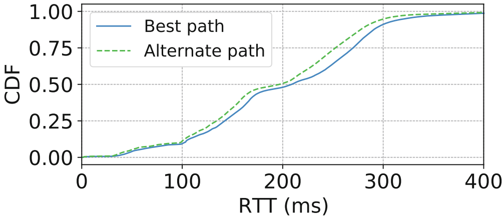
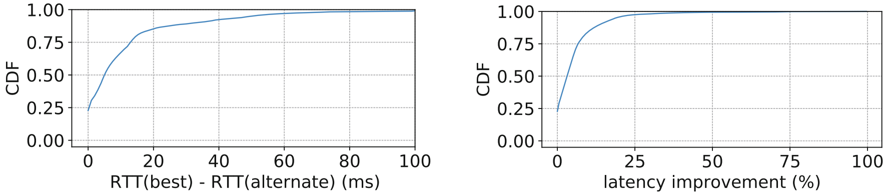


Fig. 5. CDF of RTTs to the target prefixes via the best or alternate paths.

Figure 6 shows the potential latency improvement with the alternate paths. Figure 6a shows improvement calculated from $RTT(best)$ minus $RTT(alternate)$ for each prefix. 23% of the target prefixes have no improvement, which means



(a) Differences between $RTT(best)$ and $RTT(alternate)$ for each prefix

(b) Rate of latency improvement to $RTT(best)$ for each prefix

Fig. 6. CDF of latency improvement by the alternate paths.

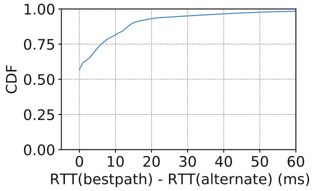


Fig. 7. CDF of improved latency to prefixes accommodating Alexa top sites.

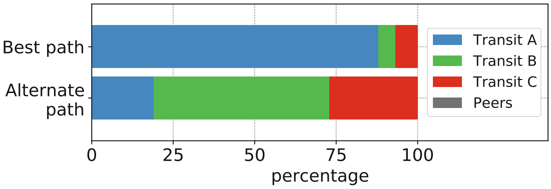


Fig. 8. Percentage of egress ASes on the best paths and alternate paths.

that the best paths are the best. Meanwhile, BGP-EPE improves latency for 77% of the prefixes, e.g., 44% of the prefixes benefit from up to 10 ms latency improvement, and 18% of the prefixes get 10–20 ms latency improvement. Figure 6b shows a relative version of Fig. 6a: rate of latency improvement by $RTT(alternate)$ to $RTT(best)$. The latency improvement for 97% of the target prefixes is within 25%, and the prefixes between 75% and 97% benefit from 6.6% to 25% improvement.

In addition to analyzing all the target prefixes, we pick prefixes of popular services to clarify the benefit derived from BGP-EPE. From the viewpoint of an ISP accommodating end-users, improving latency for such prefixes brings better QoE of the users. As popular services, we used Alexa top 1 million sites [1] obtained on July 21, 2021. We resolved the domains' IP addresses and picked $RTT(best)$ and $RTT(alternate)$ for prefixes that contain the IP addresses. Eventually, we obtained 781,501 IP addresses of 548,680 domains in the Alexa list, corresponding to 25,096 prefixes out of the target prefixes. The result focusing on the Alexa top sites is shown in Fig. 7. As with Fig. 6, latency to the prefixes accommodating Alexa top sites is also improved, e.g., latency for 43% of the prefixes is improved, and 36% of prefixes get up to 20 ms latency improvement.

Even though egress traffic engineering is a promising approach for improving latency, the result where latency for approximately 77% of target prefixes can

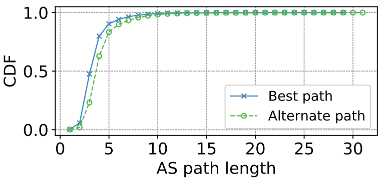


Fig. 9. AS path length of the best and alternate paths.

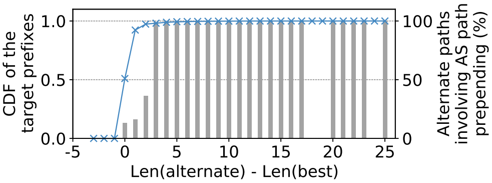


Fig. 10. CDF of the differences in AS path length of the best and alternate paths for each prefix, and percentages of alternative paths involving AS path prepending.

be improved seems a significantly higher ratio than the previous report [7]. We found that a cause of the high ratio is the imbalance of egress ASes, first hop ASes on the paths, on the best paths. Figure 8 shows the percentage of egress ASes on the best paths and alternate paths. Transit A occupies 88% of egress ASes on the best paths. However, its percentage is 19% in the alternate paths, and the other two transit ASes occupy approximately 54% and 27%, respectively. Paths via the peer ASes also appear in Fig. 8; however, their presence is relatively small (0.06% of the best paths and 0.05% of the alternate paths). Thus, peers are not visible in the figure.

The imbalance of egress ASes on the best paths arose from the order in which ShowNet established eBGP sessions with the transit providers. ShowNet established the eBGP session with transit A before others. As a result, the oldest paths survived in accordance with the BGP path selection algorithm [11]. If ShowNet established eBGP sessions in a different order, the result might change.

We next focus on AS path length of the best paths (hereafter called $Len(best)$) and alternate paths (hereafter called $Len(alternate)$) to the target prefixes. Figure 9 shows that the alternate paths are slightly longer than the best paths, as expected. The solid line in Fig. 10 shows CDF of differences in AS path length calculated from $Len(alternate)$ minus $Len(best)$ for each prefix. We can see several prefixes have longer best paths than their alternate paths, where $x < 0$. This is because ShowNet prefers peers over transit providers by configuration using LOCAL_PREF. The best paths to these 9 prefixes are advertised from peers to ShowNet, and these paths involve AS path prepending by origin ASes. On the other hand, the alternate paths to the prefixes through transit ASes are not prepended. As a result, ShowNet routers choose the longer paths via peers as best by a higher LOCAL_PREF. The shortest paths would be used if ShowNet did not prefer peers over transit.

About half of the prefixes have best and alternate paths of the same length ($x = 0$). We found that both best and alternate paths to 99.9% of those prefixes are advertised from the transit ASes. Further, the egress AS of 88% of the best paths to the prefixes is transit A, and 55% of the prefixes derive the latency improvement from their alternate paths through transit B or C. This result also demonstrates the performance-obliviousness of BGP, which sometimes does not choose better paths when AS path lengths are the same. Moreover, some prefixes have very long alternate paths compared with their best paths, e.g., $Len(alternate)$ is 25 hops longer than $Len(best)$ for two prefixes. These significant differences arise from AS path prepending. The bars in Fig. 10 show percentages of alternate paths involving AS path prepending, and 100% of alternate paths are prepended where $x > 5$. For the length difference above 5, the length is artificially increased and no longer reflects the latency of the path.

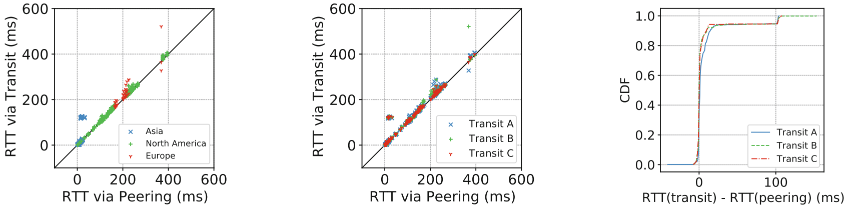
This section provides two findings: (1) BGP-EPE can reduce latency as the previous studies reported, even in a leaf of the Internet; (2) however, the degree of the latency improvement highly depends on the situation. 77% of the target prefixes got latency improvement. Furthermore, latency to prefixes of Alexa top sites is also reduced. In addition to large content providers benefiting from egress traffic engineering [28,33], the result demonstrates that a single ISP also can benefit from BGP-EPE to bring better QoE for their end-users. On the other hand, as Fig. 8 shows, the relatively high degree of improvement arises from the imbalance of the BGP best paths, which is due to the deployment process of ShowNet. A similar situation could happen anywhere, and the degree of improvement by BGP-EPE depends on the quality of the main best egress AS.

4.2 Peering Versus Transit

SR-based BGP-EPE can selectively send packets to a destination via peer *or* transit ASes. This characteristic enables us to measure and compare latency to a destination through peers or transit providers. Network operators prefer peering over transit from a cost perspective. However, clarifying the difference in performance is still challenging because it fundamentally requires egress traffic engineering or layer-7 techniques and the help of a large CDN [4]. In this study, we filled the former requirement.

To compare latency via peering and transit, we extracted the minimum RTT for each prefix via the peer ASes advertising the prefix (hereafter called $RTT(peering)$) and, also, the minimum RTT via each transit (hereafter called $RTT(transit)$). Note that SR-based BGP-EPE only influences the egress traffic. Return paths from destinations to ShowNet were the same regardless of egress ASes; hence differences between peering and transit reflect only outbound paths.

Figure 11 shows a first comparison between $RTT(peering)$ and $RTT(transit)$. Figure 11a and 11b present the two types of RTTs to each prefix in scatter plots. The x-axis indicates RTT via peering, and the y-axis indicates that of transit. Therefore, when a dot is above $x = y$, peering is better than the transit for the prefix. Since ShowNet had three transit providers, a prefix is represented by up to three dots in the figures if probes to the prefix succeeded. Both Fig. 11a coloring dots per region of the prefixes and Fig. 11b coloring dots per transit show more dots above the line: 1360 dots above $x = y$ versus 832 dots under $x = y$. Figure 11c shows the CDF of $RTT(transit)$ minus $RTT(peering)$ for each prefix. The figure also demonstrates that peering provides shorter latency than transit. The ratio of prefixes with better latency via peering is 39%, 59%, and 52% for transit A, B, and C, respectively.



(a) Categorized by regions (b) Categorized by transits (c) Latency improvement with peering

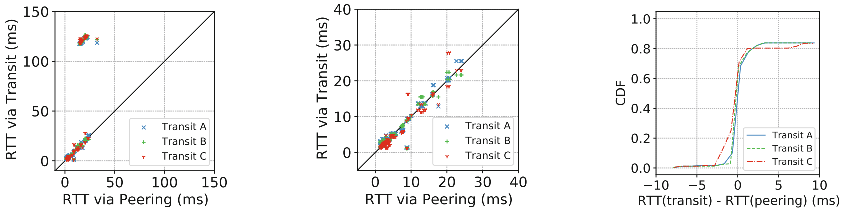
Fig. 11. Peering vs. transit. Each dot in a and b indicates a prefix; the x-axis and y-axis values are RTT to the prefix via peering and transit, respectively. (Color figure online)

To clarify the difference in more detail, we analyze the results from a regional perspective in the following sections. To determine regions of prefixes, we used the procedure described in Sect. 3.

Asia. Figure 12 shows results for the subset of prefixes from Fig. 11 allocated in Asia. Almost all ASes peering with ShowNet were Japanese ASes; hence most prefixes in Fig. 12 are located in Japan. In other words, this result compares domestic peering versus transit paths. In Fig. 12a, we observe a group of prefixes with over 100 ms RTTs via the transit providers but less than 50 ms RTTs via peering. We found that these prefixes belong to a Japanese local ISP in a region far from major cities. In Japan, cores of major ISP backbone networks converge on two urban cities: Tokyo and Osaka. A previous experiment reports that regional ISPs outside of Tokyo and Osaka often have long latency via transit ISPs due to economic reasons, so that IXPs extended to regional areas can provide significant improvement in RTT for such regional ISPs [31]. The group that appeared in Fig. 12a implies such a condition in Japan.

Figure 12b zooms on the 0–40 ms range of Fig. 12a to avoid the outliers described in the previous paragraph. The figure shows there is no significant difference in trends between peering and transit in the domestic connection. Moreover, latency improvement with peering is also not significant except for the outlier group as shown in Fig. 12c.

North America. Figure 13a focuses on prefixes allocated in North America. It shows the same trend—no significant difference between peering and transit. However, peering certainly improves latency, as shown in Fig. 13b. This result indicates peering improves latency on inter-continental connections rather than domestic as shown in Fig. 12. Paths from Japan to North America involve long distances over submarine and regional fiber-optic cables; therefore, underlying cable routes of intermediate providers have a relatively larger impact on latency than domestic connections. Besides, we can see that transit B and C achieved

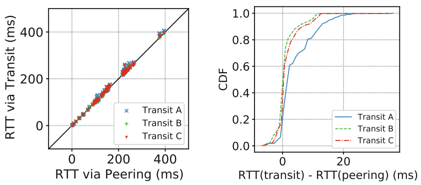


(a) RTTs via peering and transit

(b) Focusing on 0–40 ms in Figure 12a

(c) Latency improvement with peering

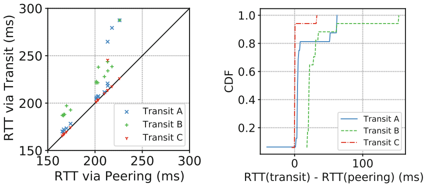
Fig. 12. Peering vs. transit for prefixes in Asia.



(a) RTTs via peering and transit

(b) Latency improvement with peering

Fig. 13. Peering vs. transit for prefixes in North America.



(a) RTTs via peering and transit

(b) Latency improvement with peering

Fig. 14. Peering vs. transit for prefixes in Europe.

more comparable latency with peering than transit A. We also attribute the difference to their cable routes.

Europe. Figure 14, which focuses on prefixes in Europe, clearly shows differences between the transit providers. The notable point is that transit C achieves comparable latency to peering. AS paths to the prefixes in Fig. 14 are advertised from a peer AS in Europe via remote peering [10] at an IXP. In addition, we found in the traceroute data that transit C also peers with the AS in Europe at the same IXP. Namely, the paths are, `ShowNet-the IXP-the AS in Europe`, and `ShowNet-Transit C-the IXP-the AS in Europe`. Therefore, ping via the peering and transit C flow through the same inter-continental cable route from the IXP to the remote peer AS in Europe.

The result indicates that although peering improves latency on an inter-continental connection, the improvement depends not only on peering but also on other factors, e.g., underlying cable routes have an impact on latency in long-distance connections. The IXP provides a better path from Japan to Europe than others thanks to remote peering. ShowNet peered with the European AS directly at the IXP; however, it is not necessarily so. BGP's performance obliviousness can lead to long RTTs when shorter delay paths are possible.

4.3 Detouring

So far, previous sections have analyzed paths received from the eBGP peers. Yet, SR-based BGP-EPE can send packets destined to arbitrary destinations to any SID-assigned eBGP peers regardless of received prefixes. Using this ability, we now observe the behavior of ASes when they receive packets having destinations that the ASes do not advertise. Such an experiment can cause unintended or abusive detouring. We however had the consent of the peers, under the agreement mentioned in Sect. 2.3.

First, we determine ASes who did not forward packets to destinations they did not advertise, using the traceroute data. As a result, we found that only four peer ASes blocked packets to non-advertised prefixes. The remaining 39 peers carried packets to the Internet; the rate of ASes not blocking such packets is 91%. Blocking packets to inappropriate destinations requires packet filtering in the data plane, e.g., access control lists, at ASBRs in addition to typical AS path- or prefix-based filtering on the BGP control plane. As a case study in New Zealand reported [21], not every AS spends effort for such filtering. However, commoditization of BGP-EPE might increase the possibility that unintended detouring will happen.

Next, the result raises a question: can BGP-EPE improve latency by detouring to such ASes? Figure 15 gives an answer. Figure 15 shows the CDF of $RTT(best)$, $RTT(alternate)$, and the minimum RTTs to the target prefixes via all possible paths including detouring (hereafter called $RTT(detouring)$). The lines of best paths and alternate paths are identical to Fig. 5. As shown, detouring improves latency, especially in the area of RTTs between 200–300 ms.

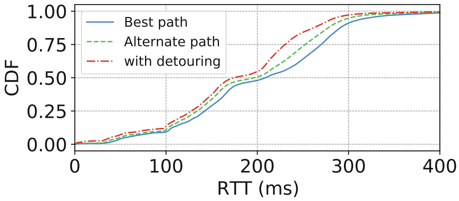


Fig. 15. CDF of the minimum RTTs to the target prefixes with detouring in addition to Fig. 5.

To clarify factors involved in the latency improvement with detouring, we compare the improvement from a regional perspective. Figure 16 shows histograms of latency differences between the best or alternate paths versus detouring for prefixes in each region. The x-axis is $RTT(best)$ or $RTT(alternate)$ minus $RTT(detouring)$ for each prefix, and the y-axis is the number of target prefixes. In contrast to Asia and North America, where reduced latencies are mainly of less than 25 ms, RTTs to prefixes in Europe are significantly improved (25–50 ms). Since RTTs from Japan to Europe are approximately around 200 ms,

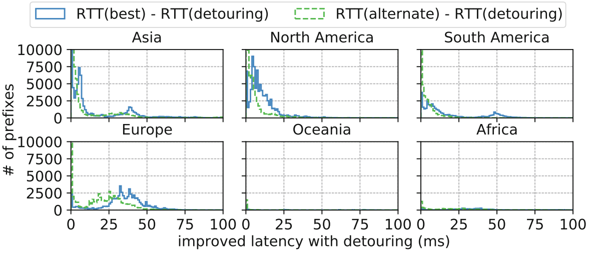


Fig. 16. Histograms of latency improvement with detouring in each region.

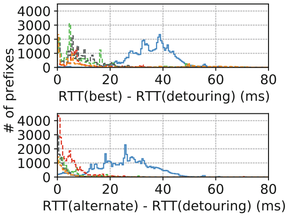


Fig. 17. Latency improvement by the top 5 ASes.

the prefixes in Europe considerably contribute to the latency improvement with detouring shown in Fig. 15.

We found that one peer AS mainly contributes to improving latency to Europe. Figure 17 shows the histogram of improved latency for the five most beneficial peer ASes. The gain of a peer AS is calculated by totaling up $RTT(best)$ or $RTT(alternate)$ minus $RTT(detouring)$ for each prefix that the peer AS provides the minimum RTT. The x-axis and the y-axis are identical to Fig. 16. As shown, the peer AS represented as the solid blue line reduces latency by 20 to 50 ms. This AS is a large AS in Europe (ranked within the top 100 in CAIDA AS Rank [9]). Thus, detouring via the AS brings shorter latency to broad networks in Europe.

These results provide two findings: (1) a security issue on peerings and (2) the potential of partial transit [15,25]. Although detouring is fundamentally abusive, current deployments often do not prevent malicious ASes from detouring. A mechanism to filter packets along with the BGP control plane is needed. On the other hand, the significant latency improvement from Japan to Europe indicates that an AS can provide peer ASes with shorter latency by selectively advertising prefixes. It is similar to partial transit, which advertises a partial BGP table to customers. If the European AS advertises the prefixes to customers under an agreement, the customers can derive the latency benefit from the partial transit service of the AS without abusive detouring.

5 Related Work

Performance-aware routing, in contrast to performance-agnostic BGP, is a longstanding issue that many studies have tackled. Early work aimed for multihomed end networks to effectively choose upstream based on performance and cost [5,19]. Those implementations involving the monitoring aspect have reached programmable switches [6]. On the other hand, [28,33] propose inter-domain performance-aware routing for large content providers. They establish BGP sessions with peers and dynamically shift egress traffic from a peering link to another to achieve better performance and avoid congestion. However, the degree of improvement reachable via performance-aware routing is still controversial;

BGP mostly chooses suitable routes, so that latency gains are small [7]. In addition to the previous studies based on large content providers, this paper provides a new case study at a single AS with a standardized technique.

Measuring the Internet while resisting routing, e.g., longest prefix matching and BGP best paths, is difficult. Measuring transit and peering selectively is a typical example. Ahmed et al. [4] accomplished this feat using client-side JavaScript and with the help of a commercial CDN. They showed peering improves end-to-end latency by at least 5% for 91% ASes. Sect. 4.2 reproduces similar results, but the methodology is different. We demonstrated that SR-based BGP-EPE enables such measurement from a single AS without any help.

6 Conclusion

We provide the first latency measurement with SR-based BGP-EPE that enables engineering egress traffic to peering ASes. We conducted the experiment at an ephemeral event network, Interop Tokyo ShowNet, in which SR-MPLS and BGP-EPE were deployed for five days in April 2021. Despite the short measurement period, the collected data brings three findings: (1) using paths other than the BGP best paths certainly improves latency; however, the gain depends on the performance-obliviousness of the BGP configuration, (2) peering provides shorter latency than transit, especially in inter-continental connections depending on underlying cable routes, and (3) 91% of peer ASes forwards packets to the Internet, although they do not advertise a full BGP table. To conclude, egress traffic engineering is effective in improving latency, but the gain depends on various factors. Meanwhile, the experiment demonstrated the potential of BGP-EPE for measuring external connectivity. A stable network where BGP-EPE is deployed would bring an opportunity for a more comprehensive and accurate measurement of the long-standing issue of performance-agnostic BGP.

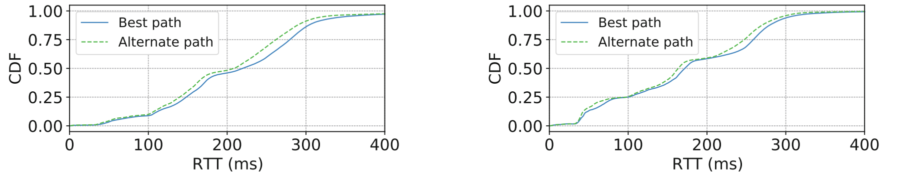
Acknowledgments. We would like to thank our shepherd Dr. Marcel Flores and the anonymous reviewers for their insightful feedback. We also thank all the people involved in Interop Tokyo ShowNet 2021.

Appendices

A Other Metrics for Representative RTTs

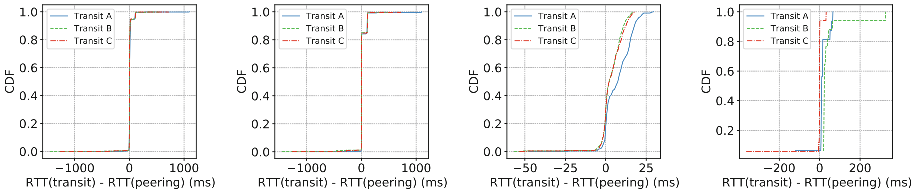
We chose the minimum, not median, RTTs for each prefix to avoid statistical errors. Figure 18a shows the median version of Fig. 5. As shown, there is no significant difference between minimum and median in a broad view. However, Fig. 19, which shows the median version of improved latency with peering (Fig. 11c, 12c, 13b, and 14b), includes such errors. Accidentally overestimated RTTs cause inaccurate latency differences between peering and transit.

Throughout the paper, we summarized RTTs by prefixes in the BGP table. Figure 18b shows the not-summarized version, which means per-address RTT, of



(a) Median RTT on a per-prefix basis (b) Minimum RTT on a per-address basis

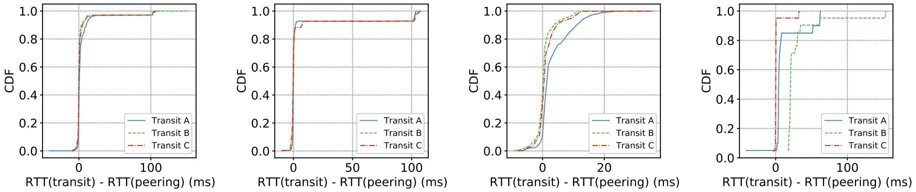
Fig. 18. CDF of RTTs to target prefixes or addresses.



(a) All (b) Asia (c) North America (d) Europe

Fig. 19. Peering vs. transit with median RTTs on a per-prefix basis.

Fig. 5. Differences between the figures, e.g., the per-address version has a higher rate with RTTs under 100 ms, arises from a bias on the distribution of target addresses. In addition, Fig. 20 shows the per-address version of peering versus transit (Fig. 11c, 12c, 13b, and 14b). The figure shows results similar to the per-prefix versions. This is because the number of prefixes advertised from the peers was small (3526 unique prefixes). As a result, there were no especially large prefixes that accommodated many target addresses.



(a) All (b) Asia (c) North America (d) Europe

Fig. 20. Peering vs. transit with minimum RTTs on per-address basis.

References

1. Alexa top 1 million sites. http://s3.amazonaws.com/alexa-static/top-1m.csv.zip
2. Interop Tokyo 2021 (2021). https://interop.jp/en/

3. Interop Tokyo 2021 ShowNet (2021). https://www.interop.jp/shownet/en/
4. Ahmed, A., Shafiq, Z., Bedi, H., Khakpour, A.: Peering vs. transit: performance comparison of peering and transit interconnections. In: 2017 IEEE 25th International Conference on Network Protocols (ICNP), pp. 1–10 (2017). https://doi.org/10.1109/ICNP.2017.8117549
5. Akella, A., Maggs, B., Seshan, S., Shaikh, A.: On the performance benefits of multihoming route control. IEEE/ACM Trans. Netw. **16**(1), 91–104 (2008)
6. Apostolaki, M., Singla, A., Vanbever, L.: Performance-driven internet path selection. In: Proceedings of the Symposium on SDN Research, SOSR '21. Association for Computing Machinery, New York (2021). https://doi.org/10.1145/3482898.3483357
7. Arnold, T., et al.: Beating BGP is harder than we thought. In: Proceedings of the 18th ACM Workshop on Hot Topics in Networks, pp. 9b–16. HotNets '19. Association for Computing Machinery, New York (2019). https://doi.org/10.1145/3365609.3365865
8. Augustin, B., et al.: Avoiding traceroute anomalies with Paris traceroute. In: Proceedings of the 6th ACM SIGCOMM Conference on Internet Measurement, IMC '06, pp. 153–158. Association for Computing Machinery, New York (2006). https://doi.org/10.1145/1177080.1177100
9. CAIDA: As rank: A ranking of the largest autonomous systems (as) in the internet. https://asrank.caida.org/
10. Castro, I., Cardona, J.C., Gorinsky, S., Francois, P.: Remote peering: more peering without internet flattening. In: Proceedings of the 10th ACM International on Conference on Emerging Networking Experiments and Technologies, CoNEXT '14, pp. 185–198. Association for Computing Machinery, New York (2014). https://doi.org/10.1145/2674005.2675013
11. Chen, E., Sangli, R.S.: Avoid BGP best path transitions from one external to another. RFC 5004 (2007). https://doi.org/10.17487/RFC5004. https://rfc-editor.org/rfc/rfc5004.txt
12. Cho, K., Mitsuya, K., Kato, A.: Traffic data repository at the wide project. In: Proceedings of the Annual Conference on USENIX Annual Technical Conference, ATEC '00, p. 51. USENIX Association, USA (2000)
13. Claypool, M., Claypool, K.: Latency can kill: precision and deadline in online games, pp. 215–222. Association for Computing Machinery, New York (2010). https://doi.org/10.1145/1730836.1730863
14. Dac Duy Nguyen, H., Secci, S.: LISP-EC: enhancing lisp with egress control. In: 2016 IEEE Conference on Standards for Communications and Networking (CSCN), pp. 1–7 (2016). https://doi.org/10.1109/CSCN.2016.7785189
15. Faratin, P., Clark, D., Bauer, S., Lehr, W., Gilmore, P., Berger, A.: The growing complexity of internet interconnection. Commun. Strat. **1**, 51–72 (2008)
16. Feamster, N., Borkenhagen, J., Rexford, J.: Guidelines for interdomain traffic engineering. SIGCOMM Comput. Commun. Rev. **33**(5), 19–30 (2003). https://doi.org/10.1145/963985.963988
17. Filsfils, C., Previdi, S., Dawra, G., Aries, E., Afanasiev, D.: Segment routing centralized BGP egress peer engineering. RFC 9087 (2021). https://doi.org/10.17487/RFC9087. https://rfc-editor.org/rfc/rfc9087.txt
18. Filsfils, C., Previdi, S., Ginsberg, L., Decraene, B., Litkowski, S., Shakir, R.: Segment routing architecture. RFC 8402 (2018). https://doi.org/10.17487/RFC8402. https://rfc-editor.org/rfc/rfc8402.txt

19. Goldenberg, D.K., Qiuy, L., Xie, H., Yang, Y.R., Zhang, Y.: Optimizing cost and performance for multihoming. SIGCOMM Comput. Commun. Rev. **34**(4), 79–92 (2004)
20. Huston, G.: AS Names. https://bgp.potaroo.net/cidr/autnums.html
21. Jager, M.: Securing ixp connectivity. APINIC 34 (2012). https://conference.apnic.net/34/pdf/apnic34-mike-jager-securing-ixp-connectivity_1346119861.pdf
22. Juniper Networks: Containerized routing protocol daemon (CRPD) (2021). https://www.juniper.net/us/en/products/routers/containerized-routing-protocol-daemon-crpd.html
23. Khan, F.: The cost of latency—digital realty (2015). https://www.digitalrealty.com/blog/the-cost-of-latency
24. Luckie, M.: Scamper: a scalable and extensible packet prober for active measurement of the internet. In: Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement, IMC '10, pp. 239–245. Association for Computing Machinery, New York (2010). https://doi.org/10.1145/1879141.1879171
25. Norton, W.: DrPeering white paper - the art of peering: the peering playbook, 7. partial transit (regional) (2010). http://drpeering.net/white-papers/Art-Of-Peering-The-Peering-Playbook.html#7
26. PCI-SIG: Single root i/o virtualization and sharing specification revision 1.1 (2010). https://pcisig.com/single-root-io-virtualization-and-sharing-specification-revision-11
27. Rekhter, Y., Hares, S., Li, T.: A Border Gateway Protocol 4 (BGP-4). RFC 4271 (2006). https://doi.org/10.17487/RFC4271. https://rfc-editor.org/rfc/rfc4271.txt
28. Schlinker, B., et al.: Engineering egress with edge fabric: steering oceans of content to the world. In: Proceedings of the Conference of the ACM Special Interest Group on Data Communication, SIGCOMM '17, pp. 418–431. Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3098822.3098853
29. StefPrevidi, S., Talaulikar, K., Filsfils, Clarence Filand Patel, K., Ray, S., Dong, J.: Border Gateway Protocol - Link State (BGP-LS) Extensions for Segment Routing BGP Egress Peer Engineering. RFC 9086 (2021). https://doi.org/10.17487/RFC9086. https://rfc-editor.org/rfc/rfc9086.txt
30. The ANT Lab: IP Address Space Hitlists. https://ant.isi.edu/datasets/ip_hitlists/format.html
31. Tsurumaki, S.: How do we improve internet connectivity outside major cities? A Japanese approach. APNIC Blog (2021). https://blog.apnic.net/2021/09/02/how-do-we-improve-internet-connectivity-outside-major-cities-a-japanese-approach/
32. Ventre, P.L., et al.: Segment routing: a comprehensive survey of research activities, standardization efforts, and implementation results. IEEE Commun. Surv. Tutor. **23**(1), 182–221 (2021). https://doi.org/10.1109/COMST.2020.3036826
33. Yap, K.K., et al.: Taking the edge off with espresso: scale, reliability and programmability for global internet peering. In: Proceedings of the Conference of the ACM Special Interest Group on Data Communication, SIGCOMM '17, pp. 432–445. Association for Computing Machinery, New York (2017). https://doi.org/10.1145/3098822.3098854