Source-Directed Path Diversity in the Interdomain Routing

Miao Xue, Gang Fu, Ruitao Ma, and Longshe Huo
Network Technology Research Institute of China Unicom, Beijing, China
Email: {xuemiao9, fugang16, mart7, huols}@chinaunicom.cn

Abstract—The Internet has abundant path redundancy, especially in the interdomain routing. However, current routing system cannot exploit the Internet path diversity and utilize the disjoint end-to-end paths efficiently. The unawareness of sources to the path selection and the best paths advertisement mechanism in the interdomain routing make it difficult to use disjoint end-to-end paths. In this paper, we present the Source-Directed Path Diversity (SDPD), leveraging which sources can specify the alternate paths to forward the traffic besides the default path. In SDPD, the packets carry the Source-Directed Tag (SDT) in the packet headers to hint the BGP routers the preference of the sources on the path selection, while the BGP routers forward the packets independently based on the sources’ indication. Moreover, we propose the multipath advertisement of the BGP route reflectors in SDPD to reduce the filtration of the redundant paths in the interdomain routing. We evaluate the SDPD through simulations over a synthetic Internet-like topology. The simulation results show that the SDPD can exploit alternate paths with low similarity and stretch efficiently.

Index Terms—Source-Directed; Path Diversity; Multipath Routing; Interdomain Routing

I. INTRODUCTION

In recent years, path diversity in the Internet has received significant attention. Many previous studies have shown that path diversity has the ability to improve the end-to-end throughput and reliability [1–3]. Driven by the benefits of path diversity, lots of network technologies are deployed in the Internet to improve the path diversity at both the infrastructure level and the protocol level. For example, multihomed stub networks are capable to access multiple ISPs, and multi-interface mobile terminals, such as Laptops and tablet PCs, can access heterogeneous networks simultaneously. While the multipath routing based Enhanced Interior Gateway Protocol (EIGRP) [4] and Open Shortest Path First (OSPF) [5] and the end-to-end multipath transfer [6–8] try to exploit the path diversity at the protocol level.

Although the evolving network infrastructure provides abundant path redundancy, the path diversity in the Internet is still not exploited sufficiently. A measurement study of a large ISP found that almost 90% of Point-of-Presence (PoP) pairs have at least four link-disjoint paths between them [9]. Savage [10] found that although Internet traffic traverses a single path, 30% to 80% of the time, an alternate path with lower loss or smaller delay exists. However in [11], Han shows that a significant portion of the paths from a multihomed site overlaps near the endhosts and in the core of the Internet, and concludes that simply having a stub network connected to multiple ISPs does not necessarily guarantee high levels of path diversity.

The existing multipath routing protocols are mainly used in the intradomain routing, such as, OSPF and EIGRP. Though the interdomain routing has the most abundant path redundancy, the Border Gateway Protocol (BGP) does not support multipath routing, which leads to poor path diversity, at both Autonomous System (AS)-level and route-level. Furthermore, for scaling and ease of management purposes, many ISPs have moved their iBGP architecture from a full mesh of iBGP sessions to route reflection [12]. But the Route Reflector (RR) only advertises its best paths to other RRs as well as to its clients, which hides most of the redundant routes.

Even if there are multiple routing entries for each prefix at the control plane of the BGP routers [18–19], how to utilize the multiple routing entries at the forwarding plane is still a problem. In the routing system using traditional destination-based forwarding, the end system has little knowledge about which path it utilizes. Loose coupling between the data flow and the forwarding path produces a mass of packets reordering, which degrades the throughput of the reliable transport layer protocols drastically [13]. Source routing, which carries routing information in the packet headers, could specify partially or fully the paths taken by the packets. However, it faces the scalability and security problems since each end-system needs a map of the overall network to formulate the end-to-end paths.

Since the path redundancy provided by network infrastructure cannot guarantee the end-to-end path diversity always, it is necessary to study how to exploit the path diversity in the Internet at different levels. To achieve end-to-end path diversity, two issues need to be addressed: 1) setting up multiple independent paths between the end-nodes (multipath routing); 2) utilizing the given independent paths based on the network and/or terminal. Based the principals above, we propose the SDPD. At the control plane, the SDPD improves the RR to advertise multiple routes toward the same prefix over the iBGP sessions, thus the BGP routers have more consistent and comprehensive routing view to provide a...
diverse set of paths. Besides, the BGP routers in SDPD are allowed to install multiple routing entries towards the same IP prefix in the Forwarding Information Base (FIB). At the forwarding plane, the SDPD makes use of the source-directed multipath forwarding. Leveraging the SDPD, the source could specify the end-to-end paths for the data flow, but it is not necessary to strictly restrict which routes to take for the end-system. In SDPD, the source only provides an indication, the SDT, which includes the Connection Identifier (CID) and Path Index (PI), to the BGP routers to hint which path is preferred. The BGP routers in the path choose the next-hops based on the source indication and their own routing policy. The forwarding rules are similar to the mechanisms in [14-15], but the SDPD could recognize the data flow and assign it to a specific path in its lifetime. The source can measure the path characters based on the SDT and specify the paths with better performance.

We evaluate the SDPD in the simulations with a synthetic Internet-like topology. The simulation results show that the SDPD provides more abundant paths at the control plane of the interdomain routing without increasing the update messages overhead and the convergence time significantly. Leveraging the SDPD, the end-systems could exploit more end-to-end paths with low path similarity and stretch. In the simulations, with different PIs, at least 70% of the alternate paths are totally edge disjoint with the best path, and 99% of the alternate paths have the same length as the best path.

The rest of this paper is organized as follows. Related works are reviewed in Section II. In Section III, we present the design of multipath BGP based the multipath advertisement of RR. Section IV details the source-directed multipath forwarding rules. In Section V, we evaluate the source-directed path diversity. Section VI discusses some open issues when deploying the SDPD. Finally, the conclusions are given in section VII.

II. RELATED WORKS

This section provides an overview of existing solutions to exploit the path diversity for the interdomain routing.

Multipath routing, which provides nodes access to multiple paths for each destination, can increase the reliability and improve the capacity by increasing the number of paths utilized. OSPF explicitly allows equal cost multi-path routing, while the EIRGP provides more aggressive multipath routing by utilizing unequal cost multiple paths. OSPF is also capable to obtain path diversity by doing multi-topology routing [16]. Besides, new intradomain routing architecture [32] is proposed to introduce hierarchical network model and source routing. Since in a specific domain, all the routing devices are under a single management entity, the multipath routing based intradomain routing protocols could be easy to deploy.

The interdomain routing has the most abundant path redundancy, as the ISPs usually design their networks with resiliency in mind, and tend to be multihomed and multiconnected. In order to exploit the path diversity of BGP, many proposals are put forward. MIRO [17] uses BGP by default and can negotiate the use of additional paths between arbitrary pairs of ASes. But it requires establishing additional state at BGP routers for each alternate path and additional out-of-band control-plane signaling. Wang [18] proposes the D-BGP, a path diversity aware routing protocol. The D-BGP extends BGP to allow each BGP router to advertise a most disjoint alternative path along with the best path. But when choosing the alternative path, the D-BGP considers mainly the AS-path length property but less the other properties like Loc_Pref and MED. Add-Paths [19] allows the BGP routers to advertise multiple paths to the same prefix over the iBGP sessions and keeps the eBGP sessions unchanged. However the Add-Paths does not limit the number of the routings to the same prefix in the FIB, which increases the consumption of the routers’ memory significantly. YAMR [21] presents the YPC which constructs a set of policy-complaint interdomain routing paths to tolerate any single interdomain link failure. Besides, the YAMR also puts forward a mechanism to reduce control message overhead imposed by alternative path advertisement by localizing routing updates. BGP-XM [31] allows routers to use multiple paths across different ASes. BGP-XM defines a router architecture tailored to accommodate the information of multiple paths to the same network prefix into a single BGP update message to guarantee path diversity, and proposes an optimal path selection algorithm.

Some novel routing architectures have been proposed to exploit the path diversity in the interdomain routing. Yang presents the NIRA [20], a new interdomain routing system that supports user choice. NIRA allows networks to offer any valley-free path and a user can specify a path using both the source and destination address. The work in [22] proposes the pathlet routing to construct the multipath routing. In pathlet routing, networks advertise fragments of end-to-end paths from which a source can assemble an end-to-end route. Pathlet routing could be a simple generalization of both path vector routing and source routing, depending on the length of each pathlet. However, it needs a significant change to the current Internet to implement the proposed architectures in practice. Reference [33] proposes to establish a logically centralized multi-AS routing control platform leveraging the control plane and forwarding plane separating of SDN(Software Defined Network), for taking efficient routing decisions, detecting policy conflicts, troubleshooting routing problems in a global view.

On the other hand, source-controlled routing reserves some choice of routes for the sources to select on a per-packet basis, which provides more flexible path diversity and more direct control on the path selection. Source routing could specify partially or fully the paths taken by the packets, however, it faces the scalability and security problems. Yang [14] puts forward a tag-based source routing architecture that uses routing deflections to provide path diversity. End-systems tag packets with hints, rather than explicit source routes, and routers use these hints to select among alternate paths. Similar to [14], path splicing [15] makes use of the splicing bits in the
packet header to switch the traffic among multiple routing trees (“slices”) along a single path. By setting different splicing bits, the end-systems can obtain diverse paths both at the AS-level and the route-level. Slick Packets [23] allows the source to embed the routing information within the packet header in the form of a forwarding subgraph. Based on the routing information in the packet header, Slick Packets could quickly switch the packets from a primary path to the alternate paths. Since it is necessary to recognize the source tag or source encode, the routers in [14-15, 23] need to modify the forwarding plane.

Table I shows the comparison of the multipath inter-domain routing protocols mentioned above.

III. MULTI-PATH BGP

A. Path Hiding in Route Reflector

The BGP is the interdomain routing protocol used in the Internet. In order to maintain a consistent routing state, the routers running the BGP instance need to establish sessions to their neighbors to exchange the BGP paths. External BGP (eBGP) sessions are used among adjacent routers belonging to different ASes to exchange paths, while internal BGP (iBGP) sessions are used among the routers belonging to the same AS to exchange the paths learned at the border of their own AS. To guarantee each router in an AS receiving all the usable paths, classical iBGP architecture initially adopted the Full Mesh [24] of iBGP sessions. However, for scaling and ease of management purposes, many ISPs have moved their iBGP architecture from a full mesh of iBGP sessions to route reflection [12]. The RR collects paths received from its neighbors over both the iBGP sessions and eBGP sessions. Based on its own routing view, the RR advertises its best paths to other RRs as well as to its clients and non-client BGP neighbors. Though the utilization of the route reflection improves the scalability of the iBGP and reduces the advertisement of the update messages, it also filters the paths that some routers suppose to receive and leads to the path hiding.

Fig. 1 shows a simple scenario to explain how the RR hides the paths. In this scenario, there are three paths to the prefix D separately via R1, R2 and R3. Providing the path via R1 has higher local preference, thus the RR chooses the path to the prefix D via R1 as the best path and advertises to its clients R2 and R3 and non-client peer R4. As a result, the R4 knows only one path to the prefix D. And the clients of RR also only know the best path and the paths collected by their own, though there exist other paths, the clients are not aware of. For example, the R2 has no knowledge of the path via R3, and vice versa. Although all the routers could receive the best paths and eventually converge, the path diversity is lost seriously comparing to using the full mesh of iBGP sessions. Such a lack of router-local path diversity causes BGP route oscillations, prevents fast recovery and restricts the load balancing on multiple BGP nexthops [19].

B. Multipath Advertisement of RR

In order to recover the path diversity in an AS, we convert the RR as a routing relay and allow the RR to disseminate multiple BGP nexthop-disjoint paths towards the same IP prefix. Advertising multiple paths increases the overhead of the update messages, thus it is necessary to select the paths advertised carefully to guarantee the path diversity as well as to reduce the control plane overhead.

<table>
<thead>
<tr>
<th>Protocols</th>
<th>Path Diversity</th>
<th>Control Plane Overhead</th>
<th>Data Plane Overhead</th>
<th>Forward Table (FT)</th>
<th>Loop- free</th>
<th>Scalability</th>
<th>Source-Directed</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIRO</td>
<td>High</td>
<td>Low</td>
<td>Tunnel ID</td>
<td>Tuned ID based FT</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Add-Paths</td>
<td>High</td>
<td>High</td>
<td>Local Path ID</td>
<td>Destination and Path ID based forwarding</td>
<td>--</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>D-BGP</td>
<td>Medium</td>
<td>Medium</td>
<td>No</td>
<td>BGP FT</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>YAMR</td>
<td>High</td>
<td>High</td>
<td>Path ID</td>
<td>Destination and Path ID based forwarding</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>BGP-XM</td>
<td>High</td>
<td>Medium</td>
<td>No</td>
<td>BGP FT</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>NIRA</td>
<td>High</td>
<td>Medium</td>
<td>No</td>
<td>Multiple FT</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Pathlet</td>
<td>Medium</td>
<td>High</td>
<td>Forwarding ID</td>
<td>Forwarding ID based</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SDN based</td>
<td>High</td>
<td>High</td>
<td>No</td>
<td>Multiple FT</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Source directed based</td>
<td>Low</td>
<td>Low</td>
<td>Deflect tag</td>
<td>Multiple next-hop</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Path slicing</td>
<td>Low</td>
<td>Low</td>
<td>Splicing bits</td>
<td>Multiple FT</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Slick packet</td>
<td>Medium</td>
<td>Low</td>
<td>FS bits</td>
<td>Multiple next-hop</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Table I. Multipath Inter-domain Routing Protocols Comparison
The RR plays two roles in SDPD. The first one is to reduce the number of BGP sessions, as the role it plays normally. The second one is to collect the paths from its clients and non-client peers and forward the paths to other clients and non-client peers. Leveraging the multipath advertisement, the RR disseminates not only the best paths, but also other next-hop-disjoint paths. Actually, by utilizing the full mesh of iBGP sessions, though the iBGP router knows all the paths of the other iBGP routers in the same AS and has high path redundancy, many paths are worthless. For example, assuming there are multiple next-hop-disjoint paths with relatively higher local preference, the paths with low local preference do not need to be disseminated over the iBGP sessions, as these paths with little chance are chosen to forward packets for the factors of routing policies and economic relationship. Therefore, the RR in SDPD only selects the next-hop-disjoint paths with better condition to advertise.

BGP uses incremental updates. After having exchanged all their routing entries among the BGP speakers, BGP routers only need to send BGP Updates to each other if a path changes. The new BGP Update implicitly replaces the previous BGP message for the same prefix. Herein, we modify the BGP routers to recognize a path not only based on the destination prefix but also the Next-hop in the path. If a path in the Adj-Rib-Ins has the same destination prefix and Next-hop as the Network Layer Reachability Information (NLRI) in the Update messages, the path is replaced by the new path. Since the Withdrawn routes have no Next-hop information in the Update messages, we extend the Withdrawn routes encodings to <length, prefix, Next-hop> to avoid withdrawing one prefix but multiple paths.

To support the multipath BGP, the FIB needs to store multiple next-hop-disjoint paths for each destination prefix. In [19], the BGP routers import all the paths in the Routing Information Base (RIB) to FIB, while in [25], the FIB only reserves 4 best paths for each destination prefix. In order to reduce the consumption of the FIB memory, we take a similar method as in [25] to keep 4 paths for each destination prefix, one best path and three alternate paths. It is worthwhile to note that the paths installed in the FIB should be next-hop-disjoint, which provides the path diversity at the best effort.

The multipath advertisement is only implemented over the iBGP sessions, between the RRs and the clients or the RRs and non-client peers. The eBGP sessions still follow the standard BGP and only advertise the best path towards a prefix. Providing a best path to a prefix in one of the eBGP peers becomes unfeasible, the border router does not need to propagate the Withdrawn route over the eBGP session, as long as there is an alternative path to that prefix existing in the FIB of the border router.

The multipath advertisement of RR increases the path diversity, at the cost of reflecting Update messages and re-triggering the BGP decision process more often. The additional Update messages impact the control plane convergence, but since the FIB has multiple paths to the same prefix, it is not necessary for the BGP routers to forward packets after the convergence of the control plane, as long as there is at least one path left in the FIB.

C. Path Selection of Multipath Advertisement

To reduce the control plane overhead, the RR only advertises the next-hop-disjoint paths towards the same destination prefix, which provides path diversity as well as the capability for load balancing. The other BGP routers still disseminate the best paths. We propose four modes in this section to select the paths for advertising.

**BGP-RR-all-Paths:** The RR advertises all the next-hop-disjoint paths in the RIB to its clients and non-client peers. Intuititionally, this mode brings the most path diversity and also the most Update messages. As no path is filtered, this mode has nearly the equivalent effect as using the full mesh of iBGP sessions on supplying the path diversity.

**BGP-RR-Highest-Locpref-Paths:** The RR advertises the next-hop-disjoint paths with the highest local preference. This mode takes the routing policy and economic relationship of the ISPs into consideration. Since the paths with low local preference are unlikely to be selected by the BGP path decision process to install in the FIB, it is not necessary to disseminate these paths. Thus, this mode filters a fraction of paths useless.

**BGP-RR-Shortest-ASlen-Paths:** The RR advertises the next-hop-disjoint paths with the shortest AS length after the selection of the BGP-RR-Highest-Locpref-Paths mode. This mode considers the path condition and filters out more paths than using BGP-RR-Highest-Locpref-Paths.

**BGP-RR-Lowest-MED-Paths:** After using the BGP-RR-Shortest-ASlen-Paths mode, the RR selects the paths with the lowest MED and advertises them. If there are multiple paths to the same prefix via the same nexthop AS, only the paths with the lowest MED are chosen.

Here it needs to note that the selection modes are only implemented on the paths having the same destination prefix. The path selection of the multipath advertisement also follows the best path decision process of BGP, but the process may break before reaching the end in order to get multiple candidate paths. In addition, the path selection of the multipath advertisement in the RR strictly complies the export policies and import policies as the standard BGP. The computational cost to run the selection modes still remains low, as comparing to standard BGP, the selection modes do not go through the whole sequence of the decision rules. The multipath advertisement introduces additional control plane overhead, but also provides more path diversity in the RIB of the BGP routers. Therefore, it needs a tradeoff between the control plane overhead and the path diversity.

IV. SOURCE-DIRECTED MULTIPATH FORWARDING IN SDPD

Previous multipath routing selects the forwarding path based on the local decisions of the router, for example, the OSPF makes use of the Round-Robin algorithm to send data packets over the multiple paths. Although the utilization of the multipath routing improves the path diversity, the sources can not tell which paths they are
using. The source only knows the endpoint interfaces of the end-to-end path, but is unaware how the paths go through the network. SDPD reserves some routes choice for the sources on per-packet basis. For example, sources can perceive the link failure in the end-to-end path and switch to a working path within a few Round-Trip Times (RTTs), also the sources can pick better paths based on the observed performance. Thus, SDPD is a promising approach to improve the utilization of path diversity.

In order to make the source have a more comprehensive knowledge about the end-to-end path it utilizes, we propose the Connection Identifier (CID) which is used to identify a data flow in the network, and Path Index (PI) which indicates the path preference of the source. The CID and PI both are encoded into the Source-Directed Tag (SDT), and we detail the generation of the SDT in the next section. The BGP routers need to refer the indications of the CID and PI to forward packets. The forwarding goal of the SDPD is twofold, the fast failure reaction by switching the path to alternate one, and the flexibility of routes chosen by sources at the edge of the network.

A. Source-Directed Tag Generation

We design the Source-Directed Tag to be an IP option in the IP header. But it is important to note that the location of the SDT should not be limited by our design, it also may be at a shim header between the IP and MAC layer, or in the header of a next-generation Internet protocol. There are two principles in designing the encoding format of SDT. Firstly, the size of the encoding format should be minimized; secondly, the processing at forwarding plane should be simple.

The CID is used to identify a data flow in the network. When an end system begins to communicate with a peer, <source IP, destination IP, source port, destination port> is utilized to generate the CID. Here we apply the hash algorithm, CRC16, which is considered having more computational efficiency[28], to calculate the CID.

\[
CID = CRC16(srcIP, dstIP, srcport, dstport)
\]

The source could get the 4-tuple from the application service or the negotiation signals for initiating the connection, for example, the TCP SYN. Although there are transport protocols using multiple addresses in the communication, such as the Stream Control Transfer Protocol (SCTP) and Datagram Congestion Control Protocol (DCCP), the CID is generated by utilizing the 4-tuple which is carried in the packet header in the initialization of the connection. Since the data flow is unidirectional, the CIDs may be different in the peers of the same communication connection. The length of the CID is 2 bytes. When the packet with CID going through a BGP router, the router records the CID in its flow list. If there is no packet with the CID passing the router in a specific period, the router removes the CID from the flow list. We set the specific period to 2s, as the default timeout of TCP and SCTP is 1s. Even if the packets with the CIDs removed reach the router again and are deflected along alternative path, there is little chance to cause packet reordering. We detail the flow list and the packet forwarding rules in the next section.

The PI is the index of the path the source selects. In this paper, the length of the PI is 1 byte, that is, the source could specify 256 paths at most. The PI does not need to specify exactly each hop’s selection as in [14-15, 22-23], since it is difficult to predict how many hops the path goes through and it is also insecure to provide the network map to the source. The PI only gives the routers a hint on which path is preferred, however it does not require the routers to forward the packets strictly following the source’s indication. The routers have completely independent forwarding decision, they could take the PI as a forward option, but they may also ignore the PI if necessary, for example, the source indication conflicts with the router’s local policy. The Fig. 2 shows the encoding format of the SDT option.

![Figure 2. Encoding format of SDT](image)

B. Forwarding Principles

In this section, we elaborate the forwarding principles of the packets with SDT. When sending packets, the source embeds the SDT in the IP headers. Upon receiving a packet, the BGP router checks the IP option. If there is no IP option of SDT in the packet header, for the backward compatibility, the router forwards the packet along the best path. Otherwise, the router checks the value of CID in SDT. If the CID is not in the router’s flow list, which means the data flow did not pass through this router, the router adds the SDT.CID to its flow list, forwards the packets over the best path and records the SDT.PI going through the best path. If the router records a CID which equals to the SDT.CID, which means the flow has passed through the router, the router chooses the forwarding path based the SDT.PI. If the SDT.PI equals the CID:PI in the flow list, the packet is forwarded along the best path. If the SDT.PI is 0, but the CID:PI in the router’s flow list does not equal to 0, set the CID:PI to CID:0. When the SDT.CID is in the router's flow list and SDT.PI does not equal to 0 and in the flow list the CID:PI is 0, the forwarding path index is calculated by FPI=PI%MAXPI, where the FPI is the index of the forwarding path, MAXPI is the number of the paths towards the same destination as the received packet. And the packet is forwarded along the FPIth path in the router. The forwarding algorithm is described in the Algorithm 1.

Fig. 3 demonstrates a simple forwarding example using SDT. In the Fig. 3, the Router A has two disjoint iBGP nexthops, Router B and Router C, and the path passing Router C is the best to reach the prefix D. There are four packets with SDT option reaching the Router A. The SDT of the first packet is \(CID_1\) : 1. Since the Router A does not record the \(CID_1\), it adds the \(CID_1\) to its flow.
list, forwards the packet along the best path, and records the PI using the best path in the flow list, here the PI is 1. Then the second packet arrives with SDT CID = 3. Checking the flow list of Router A, there exists CID. As the PI using the best path does not equal to 0, and the SDT.PI is 3, following the Algorithm 1 the Router A get the FPI 1 for the second packet. Thus, the second packet is sent over the alternative path via Router B. The flow list of Router A has no CID, therefore the third packet is transmitted along the best path. As the fourth packet’s SDT.PI is 1, we obtain the FPI is 1, thus the Router A propagates the packet along the alternative path.

Algorithm 1: BGP Router Forwarding Algorithm
Forwarding Procedure:
\[\text{if } \text{SDT.CID} \text{ not in the Router.flow_list then} \]
\[\text{add } \text{SDT.CID} \text{ to Router.flow_list}\]
\[\text{Router.flow_list}[\text{SDT.CID}] = \text{SDT.PI}\]
\[\text{forwarding packet along the best path}\]
\[\text{else}\]
\[\text{if } \text{Router.flow_list}[\text{SDT.CID}]! = 0 \text{ then}\]
\[\text{forwarding packets along the best path}\]
\[\text{Router.flow_list}[\text{SDT.CID}] = 0\]
\[\text{end if}\]
\[\text{else}\]
\[\text{FPI = SDT.PI}\%\text{MAXPI}\]
\[\text{forwarding packet along the FPIth path}\]
\[\text{end if}\]
\[\text{end if}\]

![Figure 3. A simple forwarding scenario using SDT](image)

The forwarding algorithm is designed with two principles in mind. Firstly, the forwarding algorithm should exploit the path diversity at the best effort. Secondly, when exploiting the path diversity, the best path has the highest priority to be selected to forward packets, as the best path in BGP is usually considered as the shortest path to the destination and the most consistent path in policy.

C. Loop-free Alternate Paths

The multipath BGP improves the path redundancy and installs multiple nexthops in the FIB for the same prefix. According to the standard BGP, the packets are forwarded along the best path in each router, which can guarantee the end-to-end path is loop-free. While the SDT could make the BGP routers deflect the packets over different BGP routing trees and change the AS egress point (and hence next ingress point) comparing to the best routes, which may cause loops in the end-to-end forwarding paths.

Fig. 4 shows an example of routing loop caused by SDPD. In Fig. 4, suppose that AS 1 receives three path advertisements to prefix D from its neighbors. Following the design in Section 3, all the three paths are installed in the routers’ FIB in AS 1. And the AS 2 also has three paths to the prefix D in the routers’ FIB. Here we assume the network has converged. If a packet whose destination is prefix D reaches the AS 1, the standard BGP forwards the packet along the best path (1 3 5). However, the multipath BGP may introduce paths not the best, which may violate the “prefer-customer” routing policy. For example by using SDMF, the AS 1 forwards a packet with SDT.PI = 3 along the path (1 2 4 5) to the peer AS 2. When the AS 2 receiving the packet with SDT.PI = 3, it may also select the path (2 1 3 5) to send the packet back to AS 1. Thus, the packet is forwarded back and forth between the AS 1 and AS 2 and falls into a loop.

![Figure 4. An example of routing loop caused by SDPD. The AS paths around a node represent the available paths in the nodes routing table, which are ordered in the descending order of local preference](image)

The Local Preference Attribute represents the AS’s preference in forwarding the traffic. Furthermore, the Local Preference Attribute plays an important role in implementing the “prefer-customer” and “valley-free” routing policies and avoiding the routing loop. When ASes use “prefer-customer” and “valley-free” routing policies, it means that any router of the AS will only choose an egress point that advertises the most preferred path, barring inter-AS loops as long as there are no
customer-provider loops[14]. In the multipath BGP, the multipath advertisement is only implemented in the iBGP, while the eBGP still disseminates the most preferred path. That is, the relationship of the ASes in the multipath BGP is just as in the standard BGP. Therefore, though the routers running multipath BGP receive multiple paths, if they follow the “prefer-customer” and “valley-free” routing policies, the loop can be avoided. Formally, we have the following theorem.

**Theorem 4.1:** If all the AS paths follow the “prefer-customer” and “valley-free” routing policies, the end-to-end path using SDPD is loop-free.

Proof: Suppose the forwarding path is \( s, \ldots, u_i, [u_{i+1}, u_{i+2}, \ldots, u_{i+n}] \) where \( u_i, [u_{i+1}, u_{i+2}, \ldots, u_{i+n}] \) means the multiple forwarding edges in the FIB of the BGP routers, \( s \) is the source and \( d \) is the destination. We prove the theorem using the proof by contradiction. Assuming the AS path \((a, b, c, a)\), a portion of the SDPD forwarding path, is a loop. There are 3 loop scenarios as shown in Fig. 5.

In Fig. 5 (a), the loop \( a \rightarrow b \rightarrow c \rightarrow a \) corresponds to Downhill path, Peer-to-Peer edge, Uphill path. According to the “valley-free” principles, node \( c \) can be followed by only provider-to-customer edge. While the provider-to-customer edge can not appear in an Uphill path. Thus the AS path loop shown in Fig. 5 (a) is untenable.

In Fig. 5 (b), the loop \( a \rightarrow b \rightarrow c \rightarrow a \) corresponds to Peer-to-Peer edge, Uphill path, Downhill path. According the “valley-free” principles, node \( b \) can only export paths of itself and paths learned from its customers to node \( a \). Therefore, node \( b \) should be followed by a provider-to-customer edge. However in Fig. 5 (b), node \( b \) is followed by a customer-to-provider edge, which apparently violates the “valley-free” principle. Thus the AS path loop shown in Fig. 5 (b) is untenable.

In Fig. 5 (c), the loop \( a \rightarrow b \rightarrow c \rightarrow a \) corresponds to Uphill path, Peer-to-Peer edge, Downhill path. Apparently, \((a, b, c, a)\) is a “provider-customer” loop. Following the “valley-free” principles, node \( a \) should not export the paths learned from its provider \( b \) to its another provider \( c \). Therefore the AS path loop shown in Fig. 5 (c) is also untenable.

In order to avoid the routing loop, we improve the path decision process by only allowing the paths with the highest Local Preference to be installed in the FIB. Thus, the routers no longer violate the “prefer-customer” and “valley-free” routing policies as they are in the single path case. For example in Fig. 4, the routers in AS 1 only select the paths with the highest Local Preference, \((1 \ 3 \ 5)\) and \((1 \ 4 \ 5)\), to install in the FIB. The path \((1 \ 2 \ 4 \ 5)\) is not allowed to install in the FIB, as the AS 1 and AS 2 are “Peer-to-Peer” relationship and have relatively low preference. However, the maximum number of the paths towards each prefix in the FIB is still 4.

**D. Path Diversity Metrics**

Since the SDT could hint the BGP routers to forward the packets along the alternate paths, it is necessary to measure how much these paths differ from the best path. Here we propose two metrics to evaluate the alternate paths.

Suppose \( P_s = \{A, N_1, N_2, \ldots, N_m, B\} \) is the best path between node A and B, \( P_e = \{A, M_1, M_2, \ldots, M_n, B\} \) is an alternative path between node A and B. If the edge \((M_i, M_{i+1})\) in \( P_e \) does not appear in \( P_s \), we consider it as an edge difference. In this paper, we mainly measure the edge difference of the multiple paths towards the same destination.

1) **Path Similarity:** Given the same (source, destination) pair, the path similarity is denoted

\[
\text{Path Similarity} = \frac{|P_s \cap P_e|}{|P_s|} \quad (1)
\]

where \(|P_s|\) denote s the edge length of path \( P_s \).

Path similarity provides a diversity metric of \( P_s \) and \( P_e \) between the same source-destination pair. From the definition of the path similarity, we can see that the path similarity decreases as the increasing of the disjointness of \( P_s \) and \( P_e \). And two paths that are completely edge-disjoint have path similarity 0.

2) **Path Stretch:** Given the same (source, destination) pair, the path stretch is defined as the edge length ratio of \( P_s \) and \( P_e \). The path stretch is denoted

\[
\text{Path Stretch} = \frac{|P_s|}{|P_e|} \quad (2)
\]

When the SDT.FI is not 0, the forwarding path is not the best path between the (source, destination) pair, and it is necessary to quantify the additional length that is incurred by the alternate paths. There are different metrics, such as, the actual end-to-end latency of the alternate paths [15] and the number of hops that the paths traverse [14], while here we count the edges that the alternative paths go through.

The path similarity and the path stretch are somewhat conflict goals in quantifying the path diversity. On one hand, the path diversity prefers small path similarity, thus the paths could be more disjoint. While on the other hand, the small path similarity implies the alternative path deflects the default shortest path substantially, which may result a large path stretch. Therefore, the well-designed forwarding principles which can select the paths with low path similarity and acceptable path stretch are very important.
E. SDPD Analysis

SDPD not only provides path diversity in the control plane but also guarantees the end host using the disjoint end-to-end paths in the forwarding plane. On the one hand, through the multipath advertisement of RR, the control plane of the interdomain routing could have multiple paths to the same destination. On the other hand, the source leverages CID and PI to hint the path selection, that is, the sources can exploit the diverse paths usage when the BGP router support the forwarding based the CID and PI. Therefore, the SDPD exploits the path diversity both the control plane and the forwarding plane.

V. SIMULATION AND EVALUATION

This section performs a set of simulations to evaluate the SDPD. We first evaluate the impact of multipath advertisement with different modes to the BGP convergence time, the control plane overhead and the path redundancy in the RIB and the FIB. Then we measure the path similarity and the path stretch of different source-destination pairs in the whole network when using SDPD forwarding.

A. Simulation Configuration

In this paper, we use the SimBGP [26], a BGP simulator written in Python, to simulate the multipath advertisement of BGP RR and measure SDPD. SimBGP is well suited for dynamic BGP simulations, as it takes the propagation and processing delays of the messages into consideration. In addition, it is an event-driven simulator that relies on an ordered queue to successively process simulation events. The original SimBGP supports the classical BGP. We extended it to support the multipath advertisement of the RRs, the receipt of multiple paths for the same prefix and installing multiple paths to the same destination in the FIB.

![AS-level simulation topology](image)

Following the [19], we get an Internet-like synthetic topology to provide an indicative evaluation of the SDPD. At the AS-level, the topology includes simple business relationships between ASes. In the AS, we choose a cluster size of 10 routers as in [19]. Two routers in each cluster are chosen as the RR and the backup RR. The other routers in the cluster connect with both the RR and backup RR. All the RRs are connected in a full-mesh way in the AS. To put more focus on the impact of multipath advertisement to the Tier-1 AS, the topology includes only a few Tier-2 ASes and stub ASes using to trigger the routing update event. In our simulation topology, we set that the Tier-1 ASes include 100-130 routers, the Tier-2 ASes include 30-50 routers, and the stub ASes include 10-20 routers. The Fig. 6 shows the AS-level simulation topology.

In the simulations, each router has a random processing delay with uniform distribution between 1 and 10 milliseconds. The bandwidth of each link is set to 100MB, and a queuing delay is uniformly distributed between 10 and 100 milliseconds. In addition, the Minimum Route Advertisement Interval (MRAI) timer for the iBGP sessions is set to 15 seconds and for eBGP sessions 30 seconds.

B. Control Plane Overhead of Multipath Advertisement

In this section, we simulate to have an Update advertising in a Tier-1 AS and measure the Update messages overhead and control plane convergence time. In order to compare with the standard BGP intuitively, we compute the ratio between the values of each advertisement mode and the standard BGP. Thus, the ratio value for the standard BGP is always 1. We take 10 Tier-1 ASes in the topology to measure and compare the average BGP Update message overhead and convergence time in this simulation.

![BGP Update messages ratio in the Tier-1 ASes](image)

Since the RRs disseminate the redundant paths towards the same prefix to their clients and non-client peers, the proposed four advertisement modes introduce more Update messages overhead intuitinally. Fig. 7 shows the ratio of the BGP Update messages in the selected Tier-1 ASes. From the Fig. 7, we can see that the BGP-RR-All-Paths mode produces the most Update messages. As the path selection of the other three modes is progressively strict, the advertised backup paths keep decreasing. Though the 10 Tier-1 ASes chosen have different internal topologies, the trend that the Update message ratio decreases as the progressively strict of the path selection modes is the same. In the simulation, we observe that, the BGP-RR-Shortest-ASLen-Paths and BGP-RR-Lowest-MED-Paths have approximately the same ratio. And we...
consider that there are little paths with different MEDs but equivalent AS length.

Figure 8. BGP control plane convergence time ratio in the Tier-1 ASes

Once there is no Update message about the advertising prefix in an AS, we start to compute the convergence time in this AS. The Fig. 8 demonstrates the control plane convergence time ratio. The modes of BGP-RR-All-Paths and BGP-RR-Highest-Locpref-Paths have larger convergence time ratio. While the convergence time of the other two modes is very close to the standard BGP, though the Update message overhead of this two modes is higher comparing to the standard BGP. Besides, the BGP-RR-Shortest-ASLen-Paths and BGP-RR-Lowest-MED-Paths have an overlapped convergence time ratio, though there is a little disparity in the Update messages ratio between the two modes. The Fig. 8 shows that filtering the advertised paths carefully do have effect to reduce the control plane convergence time. Besides, when backup paths in the FIB are available, the BGP forwarding plane convergence time could be cut down to nearly 0, as the data forwarding could continue by using an alternative path and it is not necessary to wait the control plane convergence.

C. Path Redundancy in RIB

We evaluate the path redundancy caused by the multipath advertisement of RRs in this section. We take 10 Tier-1 ASes and 10 Stub ASes in the synthetic topology, and calculate the average number of the routing entries towards the advertised prefix in each BGP router.

Fig. 9 demonstrates the path redundancy of RIBs in the routers of the Tier-1 ASes. The results show that the BGP-RR-All-Paths mode produces the most path redundancy, the average number of the paths in the RIB is nearly 12 times of the standard BGP. Too much path redundancy is not the goal we pursue, as the larger the number of the paths in the RIB is, the more memory is needed. Here, we seek to exploit the path redundancy with the most diversity. The other three advertisement modes also improve the path redundancy, though not so good as the BGP-RR-All-Paths. In the Fig. 10 we can observe the average number of the paths in the RIBs in the routers of the Stub ASes. The path redundancy in the Stub ASes is relatively scarce comparing to that in the Tier-1 ASes.

Roughly, the increase in terms of path redundancy is proportional to the number of paths advertised. In additional, the number of available paths to the same prefix also depends on the level of the AS. Generally, the large and highly connected ASes have more path redundancy than the ASes with a few peering/provider links. In our observation, the BGP-RR-Lowest-MED-Paths mode introduces about average 2-3 times of routing entries in RIB comparing to the standard BGP, in both the Tier-1 ASes and the Stub ASes. We consider that the BGP-RR-Lowest-MED-Paths mode can provide sufficient path redundancy and the additional memory requirement is acceptable. Such a result may encourage the operators to configure the RR with the BGP-RR-Lowest-MED-Paths as the default multipath advertisement mode.

D. Path Redundancy in FIB

The number of the paths in the RIB has a directly impact on the path diversity in the FIB, as the BGP path decision process selects the best paths from the paths in the RIB. In this section, we measure the average number of paths in the FIB using the same Tier-1 ASes and Stub ASes above. Table II and Table III separately show the average number of paths in the FIB.
From the tables we can observe that the Tier-1 ASes have more path diversity in the FIB than the Stub ASes. Nevertheless, the Stub ASes still have average 2 paths in the FIB at least, which could guarantee the fast convergence of the forwarding plane. Besides, since the maximum path number that is allowed to install in the FIB is 4, many ASes reach that limit value, which implies more candidate paths existing.

E. Path Similarity and Stretch in SDPD

In this section, we evaluate the path similarity and path stretch of SDPD. Each end host in the synthetic topology starts a traceroute to the new advertised prefix with different PIs. Thus, there are totally 2456 source-destination pairs and 2456*(maximum PI) traceroutes in each simulation. The BGP-RR-Lowest-MED-Paths mode is utilized in RRs to collect the paths at the control plane. We compare each alternative path with the best path between the same source and destination and calculate the path similarity and path stretch. We set diverse maximum PIs in the simulations and set the PI from 1 to the maximum PI in each traceroute.

Fig. 11 shows the path similarity with different maximum PIs. From the Fig. 11 we can see that, approximately 70% of the paths are totally edge disjoint. There are two reasons for the low path similarity: firstly, the control plane has the path redundancy; secondly the FI indicates the routers to forward the packets along the diverse path at the best effort. As the increasing of the used maximum FI, the fraction of the total edge-disjoint paths is cut down.

Fig. 12 demonstrates the path stretch with different maximum PIs. As the SDPD forwarding algorithm prefers the best path when forwarding packets, thus even deflected from the best path, the packets are still sent along the best paths in alternative BGP routing trees. Therefore, though the packets may be deviated from the best path repeatedly, the path stretch is still low and the length of most paths is very close to the best path. In the Fig. 12, 99% of the paths have a path stretch 1. Furthermore, the maximum path stretch in our simulation is only 1.6.

VI. DISCUSSIONS

This section investigates the challenges to deploy the SDPD in practice and discusses the open issues on the implementation.

Changes to end systems. To support the source-directed path diversity, the end systems are required to have the ability to generate and process the SDT option. Though the SDPD provides multiple choices for the forwarding path, the end systems still need to be compatible backward. If the end systems just want the packets to be forwarded along the best path, they do not insert the SDT option to the packet header. The SDT option also could be generated on the gateways of the edge networks, such as the Ingress Tunnel Routers of the LISP [27].

Changes to BGP routers. The multipath advertisement requires changes to the control plane of the BGP routers in order to support receiving multiple routings towards the same destination. In addition, since the RR is considered as a routing relay between the iBGP peers, it is required to have the ability to advertise multiple routings towards the same destination. To reduce the Update message overhead, the RR has to implement methods to filter the paths for advertising, for example, the four modes proposed in section 3.3. Moreover,
routers need to improve the forwarding plane to recognize the CIDs and forward packets depending on the PIs. Note that the changes in the forwarding plane do not introduce implementing burden to the route system, as the design is similar to the hash-based multipath routing[28] which has deployed in the Internet[29].

Routing scalability. As the BGP routing table increases rapidly in recent years, the BGP routing scalability is a major concern of the Internet routing system. The multipath advertisement disseminates multiple paths with the same destination prefix, which does not produce new prefix. Therefore, the total number of the reachable prefix in the multipath BGP is the same with the standard BGP. The multipath advertisement does not bring new routing scalability problem.

Memory consumption. The RIB of BGP routers will contain more paths and thus consume more memory because of the multipath advertisement. It is important to note that the actual memory increase due to the reception of multiple paths towards the same IP prefix is rendered sub-linear with the number of paths thanks to attribute-sharing [30]. Thus, to a certain extent, data structure optimization counteracts the increment of the memory consumption caused by the multiple path advertisement. Besides, the RIB is part of the control plane, the speed of memory access to the RIB does not need to be as high as the FIB, therefore the RIB memory can be extended easily by adding RAM to the routers. To reduce the memory consumption of the FIB, we set that the maximum number of the paths allowed to install in the FIB is 4. The operator can adjust the maximum value based on each router.

CID space. In this paper we set the length of CID to be 16 bits in the IP option. It is noted that, the CID space is about $10^3$, which is not enough to identify all flows in the Internet, but the CIDs are mainly used to identify the flows in a BGP router, and $10^7$ is enough to identify the flows in most of the routers[28]. Even if there are CID conflicts in BGP router, the router still can forward the packets over the right paths based on the destination prefix. The SDPD just aims to exploit the path diversity at the best effort, the CID conflicts could be tolerated.

Transport layer performance. Different SDT.FIs cause the packets transmitting along edge-disjoint paths. Provided a path has good quality, such as low delay or high available bandwidth, the end systems could guide the application data to that path by specifying the CID and FI. Utilizing the SDPD in conjunction with the end-to-end multipath transfer, such as CMP-SCTP[6] and MPTCP[7], in the end systems with multiple network access interfaces, it is promising to achieve better throughput performance as well as reliability.

VII. CONCLUSIONS

This paper has presented the design of SDPD to exploit the path diversity in the interdomain routing. The main contributions are as below: firstly, we have proposed the multipath advertisement of RR in BGP to exploit the path diversity in the control plane. Secondly, a source-directed multipath forwarding scheme in SDPD, which leverages the SDT to allow the end systems to find alternate paths, has been proposed. The end systems tag packets with CIDs and PIs, and the BGP routers use the CIDs and PIs to select alternate paths. Lastly, we performed simulations over an Internet-like topology to evaluate the proposed SDPD. The simulation results have demonstrated that the multipath advertisement of RR increases the path redundancy at the control plane significantly. Furthermore, the results have also shown that even with a random PI in the SDT, the source could exploit alternative paths with low similarity and small stretch. Since the source have the ability to impact the forwarding operation of the BGP routers, the SDPD is promising to be applied with end-to-end multiple paths transfer and source-based traffic engineering.

As this paper mainly focuses on the interdomain path diversity exploited by the sources, the efficiency of the forwarding plane needs to be further investigated, and we plan to systematically study on the CID lookup and fast forwarding of the packets of SDPD in our next step work.

ACKNOWLEDGMENT

This work was supported by the National Key Technology R&D Program of China, under grant No.2012BAH06B01.

REFERENCES


