Path Diversity Aware Interdomain Routing

Feng Wang
School of Engineering and Computational Sciences
Liberty University
Lynchburg, VA 24502
fwang@liberty.edu

Lixin Gao
Department of Electrical and Computer Engineering,
University of Massachusetts, Amherst
Amherst, MA 01002
lgao@ecs.umass.edu

***Abstract*—As the Internet becomes the critical information infrastructure for both personal and business applications, fast and reliable routing protocols need to be designed to maintain the performance of those applications in the presence of failures. Today's interdomain routing protocol, BGP, is known to be slow in reacting and recovering from network failures. Increasing path diversity by advertising multiple paths is an effective strategy to achieve reliable interdomain routing. However, designing a scalable and efficient multiple path advertisement to enhance the reliability of interdomain routing is still challenging. In this paper, we propose two path diversity aware routing protocols, D-BGP and B-BGP, to improve the resilience of interdomain routing. The two protocols can establish multiple paths with low routing overhead by exploiting the path diversity existing in the underlying network infrastructure. We evaluate the reliability improvements by simulation.**

I. Introduction

With the continued growth of the Internet, new and large scale mission critical applications, such as Voice-over-IP (VoIP) applications, multiplayer games, and video conferencing, are widely deployed. The reliability of the Internet becomes very important to maintain the performance of those applications. Unfortunately, Internet path failures are fairly common and can impact those applications [11], [10]. To support these applications in the face of failures, it is crucial that the underlying network infrastructure provides high redundancy and that the routing system can fully exploit the redundancy of the underlying network infrastructure to quickly recover from failures.

However, BGP, the current *de facto* interdomain routing protocol, fails to take advantage of this redundancy to provide high degree of path diversity and high end-to-end availability [10], [20], [21]. As a path vector routing protocol, BGP requires routers to advertise only the best path for a destination to its neighbors [18]. Therefore, each BGP speaking router has partial topology information and relies on its neighbors' announcements to select its best route. This can lead to routing disruptions, such as transient routing loops or transient routing outages caused by the delay in discovering alternative routes.

In this paper, we propose a Path Diversity Aware Routing (PDAR) mechanism that combines multiple path advertisement with path diversity to avoid routing disruptions and thus prevent packet losses due to routing failures. To maximize the degree of path diversity existing in the underlying network infrastructure, PDAR allows routers to advertise an alternative path along with the best path. Thus, routers can take advantage of the high degree of path diversity, and are able to find more alternative paths than BGP and existing approaches do. Knowledge of multiple paths leads to quickly recovery from failures. For instance, if a router has multiple paths to the same destination, it could quickly restore connectivity by switching to the backup path when the primary path becomes invalid without waiting for exchanging withdraw and advertisement messages in the network to learn a new best path.

Reliability enhancement through multiple path advertisement is not a new idea. Many efforts have been been made to extend BGP to allow the advertisement of multiple paths [19], [6], [13]. Since the interdomain routing system is one of the largest distributed systems today, designing scalable interdomain routing through multiple path advertisement is challenging. One of those challenges is to understand what degree of path diversity provided by multiple path advertisement is sufficient to overcome network failures. In other words, considering the cost of maintaining multiple paths, we need to understand when to propagate additional paths so that a router receives the alternative paths necessary to perform a local recovery during network failures. In this paper, we propose two conditional alternative path advertisement rules, which determine when an alternative path should be advertised. More specifically, instead of blindly advertising alternative paths, routers intelligently send alternative paths to those neighbors, which will use those paths in the future to recover failures.

We design two routing protocols to implement PDAR: Path Diversity-aware interdomain Routing (D-BGP) and Bloom Filter-based D-BGP (B-BGP). In general, D-BGP extends BGP to allow each router to advertise a most disjoint alternative path along with the best path. In addition, D-BGP incorporates Root Cause Notification (RCN) proposed by previous work [15], [24] to provide fast failure recovery. However, advertising multiple paths in D-BGP would greatly increase the routing table size, rendering it unscalable. The root cause information of failures contained in RCN can reveal the underlying topology, connectivity, and internal configuration and policies of ISPs. The information is preferred to be kept confidential. By aggregating the multiple path information and RCN into Bloom Filters, we extend D-BGP to be a more scalable and practical solution: B-BGP. More important, B-BGP could address the problem of confidential information disclosure of RCN approaches.

To evaluate the effectiveness of D-BGP and B-BGP, we conduct simulation studies to demonstrate their effectiveness

in ensuring connectivity during routing convergence, and the efficiency of the Bloom Filter-based implementation. Our evaluation results show that our protocols can reduce the disconnectivity time significantly, and will produce far fewer routing messages than BGP does. All results confirm that our protocols can achieve fast and reliable interdomain routing. Our work provides insights about incorporating the path diversity existing in the underlying network infrastructure into routing protocol design.

The remainder of the paper is structured as follows. In Section II, we present the main idea of the PDAR mechanism. In Section III, we present our first path diversity-aware routing algorithm D-BGP. In Section IV, we propose a Bloom Filter based D-BGP, called B-BGP. In Section V, we evaluate the effectiveness and efficiency of our approaches. In Section VI, we present the related works, and conclude in Section VII.

II. Main Idea of Path Diversity Aware Routing

As a path vector routing protocol, BGP requires each router to advertise only its best route for a destination to its neighbors [18]. As a result, each BGP speaking router has partial topology information and relies on its neighbors' announcements to select its best route. Ideally, when the network experience failures or router configuration changes, BGP should be able to quickly react to those failures to quickly find alternative paths. However, BGP is known to be slow in reacting and recovering from failure events [9], [10]. In the section, we present the main idea of PDAR mechanism, which combines multiple path advertisement with path diversity to achieve reliable interdomain routing.

A. PDAR Overview

The basic idea of PDAR is to allow routers to advertise multiple paths for the same destination to increase the path diversity. More specifically, each router advertises a most disjoint alternative path along with the best path when multiple paths for the same destination are available. The most disjoint path is defined as the alternative path most edge (node) disjoint with the best path. By using these alternative paths, routers can perform failure recovery locally. To decrease the cost of maintaining multiple paths for a given prefix, PDAR advertises alternative paths to carefully chosen neighbors according to the degree of path diversity provided by the neighbors.

In addition to advertising multiple paths, location information about the failure is incorporated into PDAR to identify invalid routes quickly when a failure occurs. PDAR adopts the Root Cause Notification (RCN) proposed in previous work [15], [24] to provide fast recovery. When a link failure occurs, the nodes adjacent to the link will detect the change. The node will attach the failed link to the routing update it sends out. The information, called the RCN, is propagated to other routers along each impacted path. Thus, routers can use the RCN to remove all invalid paths at once.

Let us use an example shown in Fig. 1 to illustrate the main idea of PDAR. Note that in this example and all of the following examples, we focus on a single destination prefix originating from AS 0, without loss of generality. First, we consider the case of running standard BGP, shown in Fig. 1(a). When the link between AS 1 and AS 0 fails, AS 1 loses its connectivity to the destination, and sends a withdrawal to AS 2 and AS 3 immediately. AS 2 sends a withdrawal to AS 3 right after. Upon receiving the withdrawal, AS 3 will quickly switch to the path (3 4 0). AS 1 does not have any route to reach the destination until it receives the new path (1 3 4 0) from AS 3. At the same time, when AS 2 receives the withdrawal message, it selects path (2 3 1 0). Note that even though this path is invalid, AS 2 still reroutes traffic to a valid next hop AS, which has a valid path.

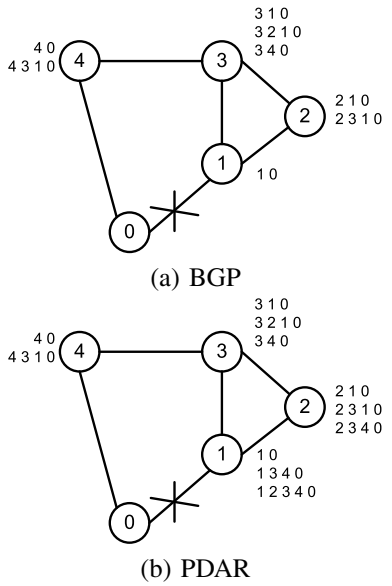


Fig. 1. An example showing the main idea of PDAR protocol. The AS paths around a node represent the available paths in the node's routing table, which are ordered in the descending order of local preference.

Now, we consider the case of running PDAR, shown in Fig. 1(b). Before the link between AS 1 and AS 0 fails, AS 3 selects the disjoint path (3 4 0) as the alternative path, and advertises the path to AS 1 and AS 2. After AS 2 receives the alternative path from AS 3, it selects the path as its alternative path, and advertises it to AS 1. As a result, AS 1 has two more paths, and AS 2 has one more path than they have in BGP. When the link failure occurs, AS 1 sends the RCN along with a withdrawal message to AS 2 and AS 3 respectively. AS 2 and AS 3 can quickly find the backup paths in their routing tables based on the RCN. They can then switch to the backup paths immediately without needing to wait for the backup path advertised by AS 3.

The major difference between PDAR with other works, such as R-BGP and EPIC, is that PDAR advertises multiple paths. PDAR adopts RCN to provide fast recovery. However, in order for RCN to function, path diversity needs to be disclosed. We use the same example shown in Fig. 1(a) to clarify this. We consider the first case of single path advertisement. Before the link between AS 0 and AS 1 fails, every AS advertises only the best path to its neighbor, as shown in Fig. 1(a). When

the link failure occurs, a RCN is sent with the withdrawal by AS 1. After receiving the withdrawal, AS 2 temporarily loses its reachability to AS 0 because all of its paths contain the failed link. The amount of time that AS 2 loses its reachability depends on the delay to get the new path (3 4 0) from AS 3, which in turn is subject to the MRAI timer. On the contrary, without using RCN, the amount of time that AS 2 loses its reachability depends on the propagation delay of the withdrawal from AS 1 to AS 2, which is not subject to the MRAI timer [18]. Thus, the duration of temporary loss of reachability at AS 2 due to using RCN without multiple path advertisement could last longer than that in the case of standard BGP. From the example, we observe the impact of path diversity on the effectiveness of RCN, and the benefit of increasing path diversity through multiple path advertisement in enhancing the reliability of interdomain routing.

Note that PDAR is used to handle single interdomain link failures. Previous studies have shown that IP-level link failures occur with relative frequency due to a variety of causes such as maintenance, equipment malfunction, router crashes, router software bugs, optical fiber cut, protocol implementation errors, misconfiguration, etc. [11], [10], [3], [2]. In addition, recent works have shown that interdomain link failures are common events [22], [2]. We notice that failures do occur simultaneously. Measurement has reported that multiple link failures in IP networks happen much more frequently [3]. However, those multiple link failures are dependent failures. For example, a linecard failure will cause all links connected to it to go down. We believe that independent multiple link failures rarely occur.

B. Conditional Alternative Path Advertisements

Advertising multiple paths for a given prefix can increase path diversity at the cost of increased memory to store additional routing information, and bandwidth consumption to process multiple paths. On the other hand, no matter how many alternative paths a router is allowed to send, the alternative paths are never used until failures happen and the relevant routers cannot find any other path better than the alternative ones. Therefore, an alternative path may never be used for failure recovery.

In PDAR, we specify the following conditions to understand what degree of path diversity provided by advertising multiple paths is needed to effectively provide fast failure recovery. Note that in all previous and the following examples, we consider alternative path advertisement on the AS level. This simplification helps us to focus on the efficiency of multiple path advertisement. In fact, the conditional alternative path advertisements can be applied at the router level and the AS level.

Now we consider when a router needs to advertise its alternative paths to neighbors based on the paths learned from its neighbors. Suppose that a router u has neighbors, and those neighbors have advertised router u their best paths and alternative paths if the alternative paths are available. We use $\mathcal{P}_i$ to represent the best path learned from a neighbor i, and $\mathcal{Q}_i$ to denote the alternative path learned from neighbor i. Among all these paths, u selects its best path, denoted as $\mathcal{P}_{best}$, and an alternative path, denoted as $\mathcal{Q}_{alt}$, which is the most edge/node-disjoint path with its best path. We use $|\mathcal{P}_a \cap \mathcal{P}_b|$ to represent the number of shared common links/nodes between path $\mathcal{P}_a$ and path $\mathcal{P}_b$.

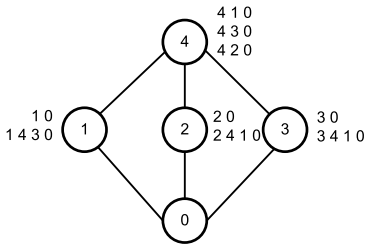


Fig. 2. An example that AS 4 applies Condition 1 to avoid advertising an unnecessary alternative path (4 3 0) to AS 2. The AS paths around a node represent the available paths in the node's routing table, which are ordered in the descending order of local preference.

Condition 1 *A router u does not advertise its alternative path $\mathcal{Q}_{alt}$ to a neighbor v if*

(i) *u has v's best path $\mathcal{P}_v$, an alternative path $\mathcal{Q}_v$, or both; and*

(ii) $|\mathcal{P}_{best} \cap \mathcal{P}_v| = 0$ *or* $|\mathcal{P}_{best} \cap \mathcal{Q}_v| = 0$

This condition implies that if a neighbor already has a disjoint path, there is no need to send an alternative path to the neighbor. Thus, only the best path is advertised to the neighbor. We demonstrate this using an example. In Fig. 2, AS 4 learns two disjoint alternate paths from AS 2 and AS 3 respectively. Suppose that AS 4 decides to send the alternative path (4 3 0) to AS 2. However, at AS 2, path (4 3 0) is redundant with the path (2 4 1 0) sent by AS 4. Having path (2 4 1 0) from AS 4, AS 2 may never use path (4 3 0) as a recovery path unless both the link between AS 0 and AS 2 and the path (4 1 0) (either the link between AS 1 and AS 0 or the link between AS 1 and AS 4) fail simultaneously, which rarely occurs.

Condition 2 *A router u does not advertise its alternative path $\mathcal{Q}_{alt}$ to a neighbor v if* $|\mathcal{P}_{best} \cap \mathcal{Q}_{alt}| < |\mathcal{Q}_{alt} \cap \mathcal{P}_v|$.

In other words, u advertises an alternative path to v if the number of shared common links/nodes between the alternative path and v's best path is less than that between u's best path and the alternative path. This condition implies that a router always sends a neighbor an alternative path, which should increase the neighbor's path disjointness. An AS selects the most disjoint alternative path from its own perspective, i.e., the one most disjoint from its *own* best path. Due to this selfish selection, the alternative path may share some common links with a neighbor's best path. For example, in Fig. 3, suppose that AS 4 advertises the alternative path (4 3 5 0) to AS 2, which is totally disjoint with its best path (4 1 0). However, when AS 2 receives this path, the path will share one common link (5 0) with AS 2's best path (2 5 0). As a result, the

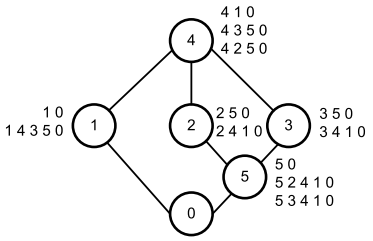


Fig. 3. An example that AS 4 applies Condition 2 to avoid advertising the alternative path (4 3 5 0), which will cause poor path disjointness at AS 2. The AS paths around a node represent the available paths in the node's routing table, which are ordered in the descending order of local preference.

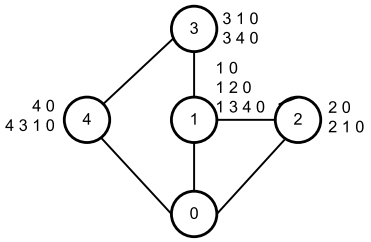


Fig. 4. An example that an AS withdraws its previously announced alternative path. The AS paths around a node represent the available paths in the node's routing table, which are ordered in the descending order of local preference.

alternative path cannot be used to recover the link failure between AS 0 and AS 5. On the other hand, AS 2 already has a disjoint path (2 4 1 0) from AS 4. In this case, path (4 3 5 0) can be used only when both the link between AS 2 and AS 5 and path (4 1 0) fail simultaneously, which rarely occurs. Therefore, in this example, AS 4 does not need to advertise the alternative path (4 3 5 0) to AS 2.

In PDAR, routers employ the two conditions before advertising alternative paths. If either or both of the conditions are met, the alternative paths will not be advertised. However, a router may not know the best path of its neighbors, which could be due to BGP loop avoidance feature or routing policies. In this case, the router just advertises the alternative path to those neighbors.

The two conditions can effectively reduce the number of routing updates arising from multiple path advertisements as seen in our simulation result, which is shown in Section V. More important, they do not impact the redundancy provided by underlying network infrastructure.

Based on PDAR, we design two routing protocols D-BGP and B-BGP. D-BGP simply incorporates the conditional alternative path advertisement and the root cause information. D-BGP is used to illustrate the basic idea of efficiently exploiting the path diversity to enhance the reliability of interdomain routing. Based on D-BGP, we present a more scalable and confidential routing solution: B-BGP. In the following sections, we will discuss the two protocols in more detail.

III. D-BGP

In this section, we first introduce how D-BGP selects and advertises the alternative paths, and reacts to a single link or node failure. And then, we discuss the correctness and complexity of D-BGP.

A. Multiple Path Advertisement

In general, D-BGP allows each router to advertise the most disjoint alternative path along with the best path to its neighbors. D-BGP adopts the same idea proposed in EPIC [24] to replace the AS path attribute with a *path list*, which is used to identify the BGP sessions between successive BGP routers. Note that D-BGP routers could advertise multiple paths by using BGP add-path extensions [19], which implements the capability of multiple paths advertisement for the same prefix. D-BGP uses path identifiers to identify the best path and the alternative path respectively.

Before an alternative path is sent out, it should pass the relevant route export operations, for example, the AS number should be appended to the AS path. Furthermore, the alternative path should pass the export routing policies and the two alternative path advertisement rules. If the best path is the only path that can be sent to a neighbor, the router only advertises the best path. On the other hand, if the best path is not permitted, for example, a router cannot advertise its best path to the neighbor from which the best path is learned, but the alternative path can be sent, the router will generate a new update message with an empty best path, attach the alternative path, and then send it to the neighbor. Finally, if both the best path and the alternative path cannot be sent, the router will send a withdrawal.

A router may withdraw its previously announced best paths or alternative paths because those paths are not available. In this case, the router sends a RCN along with the triggered withdrawal to inform every neighbor the change. Path identifiers can be used in the withdrawals to indicate which path is unavailable. On the other hand, a router may withdraw its previously announced alternative paths even though those paths are still available. According to Condition 1, a router does not advertise its alternative path to a neighbor who can provide a path disjoint with the router's best path. However, it is possible that the router sends out the alternative path before learning the disjoint path, for example, due to the path propagation delay. In this case, after the router receives the disjoint path, it realizes that it should not send the alternative path so that it sends a withdrawal with the alternative path identifier to withdraw the path.

For example, in Fig. 4, AS 3 chooses path (3 1 0) as its best path so that AS 1 will not know AS 3's best path. Suppose that AS 3 has not advertised its alternative path (3 4 0) to AS 1. Since AS 1 does not have any path coming from AS 3, AS 1 advertises its alternative path (1 2 0) to AS 3. When AS 3 prepares to advertise path (3 4 0), AS 3 finds that its best path (3 1 0) and the path (3 1 2 0) from AS 1 share the link (3 1) (does not satisfy Condition 1), and the alternative path (3 4 0) is totally disjoint with all AS 1's paths (does not satisfy Condition 2). Thus, AS 3 advertise the alternative path to AS 1. After AS 1 receives the alternate path, it realizes

that AS 3 can provide a disjoint alternative path so that AS 1 does not need to advertise its alternative path (1 2 0) to AS 3. Finally, AS 1 sends a withdrawal to withdraw the path.

B. The Best and Alternative Path Selection

D-BGP revises BGP's best path selection procedure. In particular, an alternative path is always inferior to the best paths. As a result, the alternative paths should not be used unless the best path is unavailable. The reason is that using alternative routes requires control plane signaling to avoid possible forwarding loops. If only alternative routes are left in a router's routing table, the one with the shortest path to the destination is installed as the best path. When an alternative path is selected as the best path, the router should send a withdrawal back to the sender of this alternative path as poison reverse. In D-BGP, each router selects an alternative path among all available paths according to the path disjointness. That means, the alternative path should be the path that is the most disjoint with the router's best path.

C. Reaction to Single Link/Node Failures and Policy Changes

When a link/node failure occurs, if the failure only impacts alternative paths, withdrawals without RCN are propagated to neighbors. On the contrary, if the failure affects the availability of the best path, a RCN corresponding to the location of the failure is attached along with the triggered withdrawal messages. In other words, the RCN is propagated about a change of the best path but not an alternative path. Suppose that router u and v are adjacent to the failed link, and u, which is far away from the destination, detects the failure. Router u will use its router ID and the adjacent router v's ID to generate a RCN, and send the RCN with a withdrawal. Once a router receives the withdrawal with the RCN, it will first remove the withdrawn route from its routing table. The following operation of the router depends on whether the incoming withdrawal affects the router's best path or the alternative path:

- **The best path is deleted**. In this case, the router needs to select a new best path. First, the router uses the RCN to examine each neighbor to see if the neighbor's best path and the alternative path traverse the failed link. If all paths traverse the failed link, the neighbor cannot be chosen as the next hop to reach the destination. Second, after identifying invalid paths, the router will select a new best path among all valid paths. If the new best path comes from a neighbor's best path, an ordinary announcement without the RCN is sent. Note that this feature is useful when an intradomain link fails and an iBGP router can provide the new best path. In this case, the RCN about the intradomain link failure will not be propagated outside this AS, which can help reduce the impact of routing changes in one AS on neighboring ASes. On the contrary, if the new path comes from an alternative path, an announcement with the RCN will be advertised. The purpose of attaching the RCN is to help other neighbors to quickly identify valid paths. However, if the router does not have any path, it will forward the withdrawal message with the RCN to its neighbors.
- **An alternative path is deleted**. In this case, the router needs to examine if the removed alternative path has been advertised. If this is true, it will find a new alternative path and use an announcement without the RCN to inform its neighbor of the change.

Next, we consider the reaction to a routing policy change. Suppose that a router a implements a policy change and no longer advertises a route for a destination to a neighbor b. In this case, router a can still reach the destination, but the link between a and b will no longer be used to reach the destination. Router a will send a withdrawal with the RCN to router b. After receiving the withdrawal, router b will remove all routes previously received from router a, and then select the best path as we described before.

Finally, we describe how to use an alternative path. When a router needs to use an alternative path, the router needs to inform the other routers along the path of the change. Otherwise, re-directing traffic to the protection path could cause forwarding loops. There are several technologies that can be used to avoid forwarding loops. Alternate paths can be implemented by using encapsulation schemes such as MPLS over IP or IP tunnels.

D. Correctness and Complexity of D-BGP

First, we model the network as a graph $G = (V, E)$, where V is the set of routers running D-BGP and E is the set of peering sessions between routers. In this section, we show that D-BGP can help every node in G to discover multiple paths before failures. To achieve this goal, we assume that the graph is 2-edge-connected. A graph G is 2-edge connected if it remains connected after the removal of any one of its edges. Ideally, if a graph G is 2-edge-connected, each node must have two edge-disjoint paths to reach other nodes according to Whitneys theorem. That is, the routing protocol running at each node should be able to discover the two disjoint paths. However, BGP may fail to find such paths. Thus, we will show that D-BGP can ensure that every node in a 2-edge-connected graph has installed the two edge-disjoint paths in its routing table.

Our assumption about 2-edge-connected graph is a realistic assumption. First, the availability of these two edge or node disjoint paths ensures that one of them remains operational in the presence of any single link or node failure. According to Menger's theorem, 2-edge-connected graph is the necessary and sufficient condition for a graph to tolerate any single link failure. Second, in recent years the number of multi-homed prefixes has been rapidly increasing. A prefix multi-homed directly or indirectly to multiple tier-1 ASes, can be described as a 2-edge-connected graph [12]. In addition, we also notice that full path disjointness is not necessary to realize reliable interdomain routing. D-BGP still can be applied to other networks to provide reliable routing.

Before proving the correctness of D-BGP, we introduce several definitions needed to characterize best paths and al-

ternative paths. For a 2-edge-connected graph, we divide it into two components: (1) the *best path tree* to the destination, and (2) *bridges* between branches of the best path tree. Griffin, *et al*. have shown in [5] that the best paths to the destination formed from all BGP speakers is a directed tree. We refer to this direct tree as its best path tree. We define the *parent* of a node at the best path tree as its parent, and its *children* as the children of the node at the best path tree. We define a *leaf node* as a node of degree 1 in a best path tree.

Based on Whitneys theorem, the following property of a 2-edge-connected graph is easily proven:

Property 1 *Given a 2-edge-connected graph, all leaf nodes at the best path tree must have one or more bridges, and their disjoint alternative paths must be via those bridges.*

Based on this property, we can easily prove the following lemma, which can be used to identify if a graph is a 2-edge-connected graph or not.

Lemma 1 *If every leaf node at the best path tree of a graph G has a path disjoint with its best path via a bridge, G is a 2-edge-connected graph.*

Therefore, based on the lemma, we have the following theorem about D-BGP.

Theorem 1 *Given a 2-edge-connected graph G, D-BGP ensures that every node in G has two edge-disjoint paths in its routing table after the routing system converges.*

Proof: Suppose the best path tree in G contains N branches originating from the root node. We prove the theorem by induction on N.

Base step: $N = 2$. According to Lemma 1, each leaf node in G must have a bridge connecting to another branch of the best path tree. When the best path tree has two branches, the leaf nodes advertise their best paths to each other so that their routing tables must have a disjoint path via the other. In addition, D-BGP requires the leaf nodes to advertise the alternative paths to their parents, or all nodes along their own best path branch. Thus, every node in G must have a disjoint alternative path.

Induction step: assume that the best path tree in G has $N-1$ branches and all nodes have a disjoint alternative path. Suppose that the leaf node at the Nth branch of the best path tree is u. The leaf node u must have a bridge connecting to node v at another branch of the best path tree. Since every node in $N-1$ branches already has an alternative path disjoint with its best path. Thus, v must have a disjoint alternative path. By advertising the alternative path via the bridge, u must have the disjoint alternative path. Since D-BGP also requires a node to advertise a disjoint path to its parent, all nodes along the Nth branch finally will have the disjoint path from the leaf node u.

■

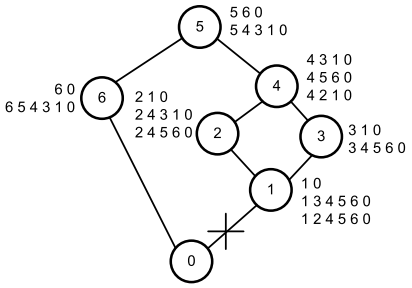


Fig. 5. An example of the resultant alternate paths after a link failure between AS 0 and AS 1. The AS paths around a node represent the available paths in the node's routing table, which are ordered in the descending order of local preference.

D-BGP can pre-compute the best recovery path when the link connecting to the root of the best path tree fails. In other words, all routers already have the best alternative paths in their routing tables before the link failure occurs. We use an example shown in Fig. 5 to show this. Before the link between router 1 and router 0 fails, all routers already have the best alternate paths in their routing tables, which are represented by the bold lines. Therefore, when the link between router 1 and router 0 fails, other routers can quickly find the best recovery path by examining the RCN.

Now we analyze the number of messages transferred by D-BGP during failover events, in which the current route to the destination becomes unavailable and is replaced by a less preferred alternative path with different next-hop or AS path to the destination. The message overhead for a failover event in D-BGP is bounded by $O(|E|)$, where $|E|$ is the number of links. After a node receives its first withdrawal/announcement, it will send out at most one withdrawal or announcement message to each neighbor. Thus, the message overhead is bounded by $O(|E|)$.

IV. B-BGP

The major limitation of D-BGP is that the fine-granularity information containing in RCN, such as router level connections, can reveal underlying the topology, connectivity, and internal configuration and policies of ISPs. The information is preferred to be kept confidential. In particular, the root cause information of failures is typically considered confidential and will not likely share this information with their customers and other ASes. In addition, advertising multiple paths would greatly increase the routing table size. To address these issues, we present Bloom Filter-based D-BGP, called B-BGP, which uses Bloom Filters to aggregate the multiple path information, and avoid disclosure of the confidential information containing in RCN.

A. Bloom Filter

Since a Bloom Filter can be used to compress a set and check for membership, we use Bloom Filters to compress the path lists used in D-BGP, which include the best paths and the alternative paths. A Bloom Filter is an array of m bits, with

all bits initialized to zero. A Bloom Filter uses k independent hash functions $h_1, h_2, \ldots, h_k$, and each hash function maps inputs to random number distributed uniformly over the range $\{1, \ldots, m\}$. To check if a record y is in S, we compute $h_1(y)$, $h_2(y)$,..., $h_k(y)$ and check the bits at positions $h_1(y)$, $h_2(y)$,..., $h_k(y)$. If any of them is 0, then y is not in the set S. If all positions are set to 1, y is considered to be in the set. However, there is a certain probability that y is not in the set, which is called a false positive.

B. Incorporating Bloom Filter

Each router initially constructs a Bloom Filter that represents its router ID, and propagates the Bloom Filter to its neighbors. Following the reception of the Bloom Filter, each router maps its router ID to the Bloom Filter and advertises it. Once a link failure occurs, the router adjacent to the failure will advertise a withdrawal message along with a Bloom Filter, which is used to encode the failed link. After receiving the withdrawal, a router performs a membership test. That is, if the failed link is found in a path's Bloom filter, that path is declared to be impacted by the failure.

Since B-BGP aggregates path information into Bloom Filters, we need to modify some of the mechanisms used in D-BGP. We notice that Bloom Filters can cause false positives, which can impact the reliability of B-BGP. First, we focus on how B-BGP works, and assume we have perfect Bloom Filters. Finally, we discuss the false positive in B-BGP and present solutions.

1) Link Signature and Path Signature: Here, we formally describe the Bloom Filter used to encode a link and a path. B-BGP uses a *link signature* to represent a BGP session between two routers. For a pair of BGP routers, the link signature is defined as:

$$L_{ab} = B_a \cup B_b$$

where B_a represents the Bloom Filter value at router a and $\cup$ represents the bitwise OR operation.

Given a path $P_i = (i, i-1, \ldots, 1)$, the path signature is defined as

$$S_i = B_i \cup S_{i-1} = B_1 \cup B_2 \cup \ldots \cup B_i.$$

In order to avoid forwarding loops, a router can detect if a path signature contains a loop by checking itself with the signature.

2) Most Disjoint alternative path: In B-BGP, we design a metric, called *path similarity*, to identify a most disjoint alternative path. Given a pair of path signatures, S_1 and S_2, we define *path similarity* between S_1 and S_2 as following:

$$sim(S_1, S_2) = \frac{|S_1 \cap S_2|}{|\max(S_1, S_2)|}$$

where $|\cdot|$ represents the number of 1s in the filter, and $\cap$ represents the bitwise AND operation. Among a set of path signatures, we select the one that has the smallest path similarity value as the most disjoint alternative path. Furthermore, in order to identify a disjoint neighbor in B-BGP, we define a threshold SIM to identify a disjoint alternative path. That is, if $sim(S_1, S_2) \leq SIM$, then we consider that the two paths are disjoint. In the next section, we will show the threshold value.

C. Reducing The Negative Impact of False Positive

False positive errors can cause long-term route losses. For example, in Fig.2, AS 4 chooses path (4 1 0) as its best path, and has two alternative paths (4 2 0) and (4 3 0). Suppose that the link between AS 0 and AS 1 fails. If AS 4 incorrectly considers path (4 2 0) as an invalid route due to false positive, it will lose the path until AS 2's best path has been changed and AS 2 re-announces the new path to it.

To reduce the impact of false positive on the reliability of B-BGP, we can optimize the Bloom Filter design so that the false positive rate is the minimum, or we can improve B-BGP to reduce the impact. We present our solutions as follows.

- **Incremental Bloom Filter.** Motivated by current work [7], we propose to use an incremental Bloom Filter to control the Bloom Filter size. The basic idea is that in the beginning, we construct a Bloom Filter with a small number of bits, and ensure that the false positive rate requirement is met. A small Bloom Filter is propagated to the next hop router. After certain hops, the subsequent routers will add the second Bloom Filter to handle the additional links. They will map their router ID to the second Bloom Filter. If a router wants to perform a failed link test, it checks if the link signature is found in any of the filters. If the link signature belongs to any of the filters, it is declared to be in the set. Moreover, we can use the incremental Bloom Filter to estimate a path's distance, which can help B-BGP to select the shortest disjoint path among several available disjoint paths. For example, the order of those filters can be used to estimate the distance.
- **Duplicated Announcement.** Another way to reduce the impact of false positive on B-BGP is to change BGP's routing update trigger mechanism. Standard BGP advertises a routing update only when its best path has been changed. In B-BGP, a router is allowed to send its best path to a neighbor whenever the router receives an announcement or withdrawal *with a failed link signature*. In other words, even though an incoming update message does not impact the router's best path, the router still needs to advertise its best path to the neighbor. Note that only a *update message with a failed link signature* can trigger the duplicated announcement. That is, a router sends a update message only when its best path is affected by an incoming routing update with a failed link signature. The duplicated announcement is sent only to the neighbor who just sent a routing update. In addition, the duplicated announcement does not forward the failed link signature so that the announcement will not trigger any more duplicated announcements.

V. Performance Evaluation

A. Simulation Environment

To evaluate the performance improvements made by our multipath routing, we compare the simulation results of D-BGP without the conditional alternative path advertisements (DW-BGP), D-BGP, and B-BGP with standard BGP.

We implement D-BGP and B-BGP based on the event-driven BGP simulator, simBGP [1]. Each router has a random processing delay with uniform distribution between [0.001, 0.01] millisecond. Each link is assigned bandwidth 100MB and a queuing delay uniformly distributed between [0.01, 0.1] millisecond. The MRAI timer for the BGP sessions is set to 30 seconds. In our simulation, when a node is ready to send a routing message to one of its peer and if the MRAI timer for this peer is not set, we will artificially block the message with a duration uniformly distributed between [0,MRAI]. We also assign 200 msec to each node as the FIB updating delay.

In our B-BGP simulation, we implement standard Bloom Filters, and select $m = 64$ bits as the size of the path signatures. In order to determine a disjoint path, we select 0.2 as the similarity threshold.

To derive a simulation topology that resembles the Internet topology, we use topologies provided by previous work [17]. In each simulation, we pick one router to originate a destination prefix. When every router reaches stable states, we break one of the egress links of the origin router. Because the routers in the topologies are densely connected, the router is still connected to the network, which triggers a failover event. We repeat the operation for every egress link of a destination for 5 times.

B. Simulation Results

We measure the performance of D-BGP and B-BGP in terms of the number of messages and convergence delay when a destination is advertised, defined as an "Up" event, and when one of link egress links of the origin router fails, defined as a "failover" event. We also measure the duration of failover events.

Fig. 6(a) and (b) show the convergence delay during the two events. From Fig. 6(a), we observe that D-BGP and B-BGP experience longer convergence delay than BGP when a destination prefix is announced. The reason is that the alternative paths need to be propagated along the longest path. However, our two algorithms converge more quickly than DW-BGP because of the conditional alternative path advertisements. Fig. 6(b) shows that D-BGP and B-BGP can reduce convergence delay significantly during failover events. Compared with BGP, we find that our schemes can reduce the convergence delay during failover events by more than 60%.

Fig. 6(c) and (d) show the number of messages exchanged during two events. From Fig. 6(c), we observe that D-BGP and B-BGP produce fewer number of routing updates than DW-BGP. We also observe that our schemes produce more messages than BGP due to alternative routes. However, from Fig. 6(d), we find that our schemes generate the least number of routing updates during failover events. The reason is that our algorithms can find more alternative routes than BGP, and can quickly identify the valid alternative routes by using RCN.

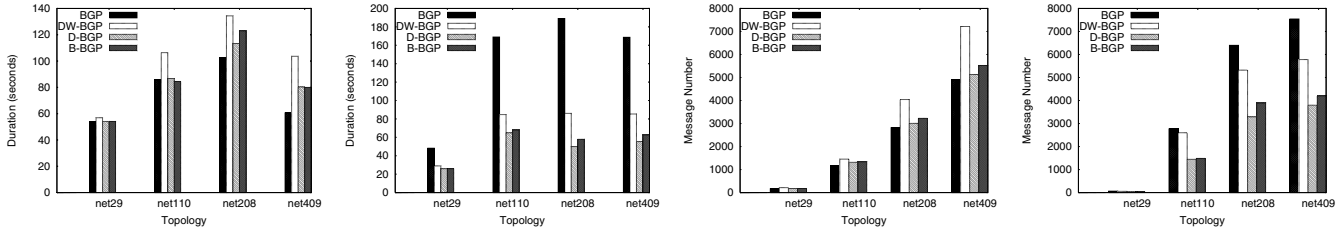
D-BGP and B-BGP also improve reachability significantly. We measure network reachability in terms of the duration of lack of connection. Note that we measure the duration of lack of connectivity on data plane. That means, we not only measure if a router has a route to the destination, but also examine if the route is valid or not. For example, a router may have a path to a destination in its routing table. But some intermediate routers along the path at that time lose their routes to the destination. As a result, we consider the router does not have connectivity to the destination. Our simulation results show that D-BGP and B-BGP can totally avoid transient failures.

VI. Related Works

Several solutions have been proposed to rapidly indicate and remove invalid routes to suppress the exploration of obsoleted paths [16], [4], [15], [24]. Consistency Assertions (CA) [16] tries to achieve this goal by examining path consistency based solely on the AS path information carried in BGP announcements. However, the AS path consistency might not contain sufficient information about invalid paths. It is hard to accurately detect invalid routes based solely on the AS path information. Ghost Flushing [4] reduces convergence delay by aggressively sending explicit withdrawals to quickly remove invalid paths. BGP-RCN and EPIC [15], [24] propose to use location information about failures, or root cause information, to identify invalid routes. EPIC [24] further extends the idea of root cause notification so that it can be applied at a router rather than an AS. Those approaches are efficient in reducing convergence delay. However, simply applying those methods might not lead to reliable routing. Our paper focuses on increasing path diversity to implement reliable routing, rather than removing invalid routes to accelerate BGP convergence process.

Techniques have been proposed [2], [8] to find local pre-planned protection paths before failures. Bonaventure et al. have proposed a fast re-route technique to protect interdomain links. The basic idea is that each router pre-computes a restoration path for each of its BGP peering links. The recovery path is used to re-route traffic when the protected BGP link fails. R-BGP aims to solve the transient routing failures problem for any interdomain link failure, not just for the failure of a direct neighboring interdomain link [8]. Our work extends these previous studies by designing path diversity-aware routing protocols to improving the reliability of interdomain routing.

Extensive works [19], [6], [13] have been done to allow BGP to advertise multiple paths or multiple nexthops for a given prefix in UPDATE messages. In addition, there are two proposals for multiple path interdomain routing. The first one is MIRO [23] that allows routers to inform their neighbors multiple routes instead of only the best one. Thus, MIRO can allow ASes to have more control over the flow of traffic in their



(a) Convergence time (Up) (b) Convergence time (Failover) (c) Message Number (Up) (d) Message Number (Failover)

Fig. 6. Simulation results. An"Up" event is defined as the time when the destination is advertised by the original router, while a "failover" event is defined as the time when one of link egress links of the origin router fails.

networks, as well as enable quick reaction to path failures. The second one is Path Splicing [14], which aims to take advantage of alternate paths in BGP routing table to discover multiple paths. Instead of using only the best path in the BGP routing table, a packet can select any path in the BGP routing table by indicating which one to use in its header. Our work focuses on improving the route reliability by advertising multiple paths for the same prefix.

VII. Conclusion and Future Work

In this paper, we proposed path diversity aware routing protocols to enhance the reliability of interdomain routing. Our proposed protocols can improve path diversity through the use of alternative path propagation. To keep the low overhead of multiple path advertisements, we investigate what degree of path diversity provided by multiple paths is needed to mitigate the effect of path failures. Based on our analysis, we propose D-BGP, which allows routers to discover more useful routing information. Moreover, we extend D-BGP by aggregating the multiple path information into Bloom Filters to design a more scalable and practical solution. We use Bloom Filters to face the problem of confidential information disclosure of RCN approaches. Our evaluation results show that our design can improve the reliability of interdomain routing significantly. At the same time, our approaches can converge faster and produce far fewer routing messages than BGP does. Our work provides insights about incorporating multiple paths that exist in the underlying network infrastructure into diverse routing design. Our future research will focus on 1) improving the multiple path discovery protocol to find the optimal alternative routes; and 2) considering multiple independent failures recovery.

Acknowledgments

We would like to thank anonymous reviewers for their constructive comments. The work is partially supported by NSF grants CNS-0626617, CNS-0626618, and CNS-0325868.

References

[1] Simple BGP Simulator. http://www.bgpvista.com/simbgp.shtml/.

[2] Bonaventure, O., Filsfils, C., and Francois, P. Achieving sub-50 milliseconds recovery upon bgp peering link failures. In *In Proceedings of ACM CoNEXT* (Toulouse, France, 2005).

[3] Boutremans, C., Iannaccone, G., Bhattacharyya, S., Chuah, C., and Diot, C. Analysis of Link Failures in an IP Backbone. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop* (November 2002).

[4] Bremler-Barr, A., Afek, Y., and Schwarz, S. Improved bgp convergence via ghost flushing. In Proc. IEEE INFOCOM. IEEE, Mar 2003., 2003.

[5] Griffin, T. G., and Wilfong, G. T. An Analysis of BGP Convergence Properties. In *Proceedings of ACM SIGCOMM* (Cambridge, MA, August 1999).

[6] Halpern, J. M., Bhatia, M., and Jakma, P. Advertising Equal Cost Multipath routes in BGP. draft-walton-bgp-add-paths-06, March 2006.

[7] Hao, F., and and; T. V. Lakshman, M. K. Incremental Bloom Filters. In *INFOCOM* (2008).

[8] Kushman, N., Kandula, S., Katabi, D., and Maggs, B. R-BGP: Staying Connected In a Connected World. In *4th USENIX Symposium on Networked Systems Design & Implementation* (2007).

[9] Labovitz, C., Ahuja, A., Bose, A., and Jahanian, F. Delayed Internet routing convergence. *IEEE/ACM Transactions on Networking 9*, 3 (June 2001), 293–306.

[10] Labovitz, C., Ahuja, A., and Jahanian, F. Experimental Study of Internet Stability and Backbone Failures. In *Proceedings of FTCS* (1999), pp. 278–285.

[11] Labovitz, C., Malan, G. R., and Jahanian, F. Internet Routing Instability. *IEEE/ACM Transactions on Networking 6*, 5 (1998), 515–528.

[12] Liao, Y., Gao, L., Guerin, R., and Zhang, Z.-L. Reliable Interdomain Routing Through Multiple Complementary Routing Processes. In *Proceeding of ReArch'08 - Re-Architecting the Internet* (December 2008).

[13] Mohapatra, P., Fernando, R., Filsfils, C., and Raszuk, R. Fast Connectivity Restoration Using BGP Add-path. draft-pmohapat-idr-fast-conn-restore-00, September 2008.

[14] Motiwala, M., Feamster, N., and Vempala, S. Path Splicing. In *SIGCOMM* (SEATTLE, WA, AUGUST 2008).

[15] Pei, D., Azuma, M., Massey, D., and Zhang, L. Bgp-rcn: improving bgp convergence through root cause notification. *Computer Network and ISDN Systems 48*, 2 (205), 175–194.

[16] Pei, D., Zhao, X., Wang, L., Massey, D., Mankin, A., Wu, S. F., and Zhang, L. Improving BGP Convergence Through Consistency Assertions. In *INFOCOM* (2002).

[17] Premore, B. http://www.ssfnet.org/Exchange/gallery/asgraph/index.html.

[18] Rekhter, Y., Li, T., and Hares, S. A Border Gateway Protocol 4 (BGP-4). RFC 4271, Jan. 2006.

[19] Walton, D. A. R., Chen, E., and Scudder, J. Advertisement of Multiple Paths in BGP. draft-walton-bgp-add-paths-06, March 2006.

[20] Wang, F., Gao, L., Wang, J., and Qiu, J. On Understanding of Transient Interdomain Routing Failures. In *Proceedings of IEEE ICNP* (2005).

[21] Wang, F., Mao, Z. M., Wang, J., Gao, L., and Bush, R. A measurement study on the impact of routing events on end-to-end internet path performance. In *Proceedings of ACM SIGCOMM* (2006).

[22] Wu, J., Mao, Z. M., Rexford, J., and Wang, J. Finding a needle in a haystack: Pinpointing significant BGP routing changes in an IP network. In *NSDI* (2005).

[23] Xu, W., and Rexford, J. Miro: multi-path interdomain routing. In *SIGCOMM '06* (2006), pp. 171–182.

[24] Zhenhai, J. C. Limiting path exploration in bgp. Proceedings of IEEE INFOCOM, 2005.