Multi-Path BGP (MBGP): A Solution for Improving Network Bandwidth Utilization and Defense against Link Failures in Inter-Domain Routing

Hiroshi Fujinoki
Department of Computer Science
Southern Illinois University Edwardsville
Edwardsville, Illinois 62026-1656 USA
E-mail: *hfujino@siue.edu*

ABSTRACT
Border Gateway Protocol version 4 (BGP-4) is the routing protocol for inter-domain routing in the Internet. Although BGP-4 is a scalable distributed routing protocol, BGP propagates only the selected best path for a destination to other autonomous systems. This property and long convergence delay in BGP have been known to cause serious inefficiency in network resource utilization and vulnerability to link failures. This paper proposes and describes a new routing protocol, MBGP (Multi-path BGP), to solve the problems by dynamically utilizing concurrent multiple BGP paths in today's Internet without routing loops. MBGP is designed to co-exist with the existing BGP routers. Performance analysis indicates that MBGP has $O(N)$ processing overhead for each MBGP message. Concurrent inter-domain multi-path routing by MBGP with these advantages will enhance the efficiency in the future Internet.

KEY WORDS
BGP, inter-domain routing, multi-path routing, bandwidth utilization, resistance to link failures, routing loop

1. INTRODUECTION

In BGP, each autonomous system (AS) announces a range of IP addresses for the host computers that belong to the AS with its unique AS number using the message called UPDATE. Each AS learns memberships of individual hosts in other ASes and the paths to reach them by UPDATE messages that are broadcasted by the other ASes [1]. Since each AS forwards only the selected best paths to other ASes, ASes in the downstream of an AS will not detect multiple BGP paths beyond their next-hop routers. This implies that although multiple BGP paths exist and they have residual transmission bandwidth, those multiple paths will not be efficiently utilized [2, 3].

CISCO introduced *BGP multipath* that allows BGP speakers to select multiple paths to reach destination ASes from local *Adj-RIB-in* table [4]. However, since *BGP multipath* does not advertise multiple paths to other ASes, routing loops can happen. Figure 1 shows an example. AS1 advertises its local hosts by UPDATE message to AS2 and AS8. AS2 forwards the UPDATE message through AS3, AS4 and AS6. AS8 forwards the UPDATE message through AS7, AS6, AS5 and AS4. AS4 and AS6 implement *BGP multipath*. AS6 gets two paths to AS1: through AS4 and AS7. Similarly, AS4 gets two paths to AS1: through AS3 and AS5. When AS6 sends IP packets to AS1, it can transmit packets through AS4 and AS7. Since AS4 also has two paths to AS1 (through AS3 and AS5), the packets AS6 transmits to AS1 can be forwarded to AS5 by AS4, and AS5 forwards the packets to AS6. This causes a routing loop.

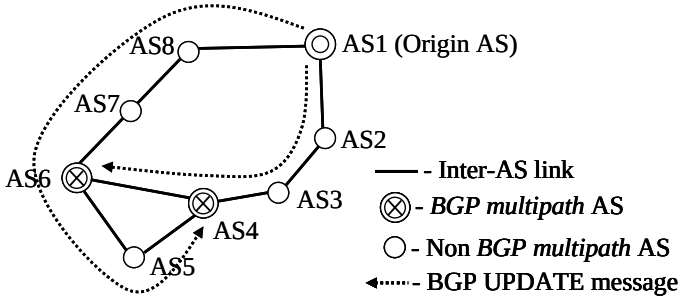

Figure 1 – Possible routing loops due to multipath transmissions

This example implies that the more BGP routers implement *BGP multiplath*, the more likely routing loops can happen. Similarly, since multipath BGP speakers do not communicate with each other, network traffic can be split to a large number of paths even through a few multipath BGP speakers. The third problem is that since *BGP multipath* does not consider current traffic load or delay for each path, efficient load balancing will be difficult.

This paper proposes a new routing protocol that achieves the following innovative routing mechanisms to effectively maximize utilization of residual transmission bandwidth and to improve resistance to unintentional and intentional link failures such as multiple-origin AS conflicts [5] and address hijacking [6] in the Internet.

- BGP speakers dynamically detect multiple BGP paths to destination ASes and transparently perform multi-BGP path transmissions to the destinations, while the number of multiple BGP paths explored is under control.
- Intermediate BGP speakers dynamically adjust traffic load through multiple BGP paths and detour failed BGP paths without significant delay.

- MBGP achieves the above benefits while backward compatibility is maintained with the existing BGP protocol without routing loops.

The rest of this paper is organized as follows. Section 2 presents our analysis on the availability of multiple BGP paths in today's Internet. Section 3 describes the MBGP routing protocol. Section 4 presents performance analysis. Section 5 summarizes the conclusions and on-going work, followed by a list of the selected references.

2. ANALYSIS OF THE INTERNET STRUCTURE

The number of multiple BGP paths from ASes to the core of the Internet was analyzed. We defined the core of the Internet as the AS that had the largest degree of AS inter-connections, which was AS 701 [7]. Figure 2 shows the results of our analysis based on the *Adj-RIB-in* table owned by AS 65000 [8], but similar results were observed for other ASes. In Figure 2, the numbers of distinct BGP paths from each of 62 ASes to AS 701 are shown for 0 extra AS hop (the minimum-hop BGP paths), 1 extra and 2 extra hops sorted in the ascending order. We recognized each BGP path by a sequence of AS numbers in each path.

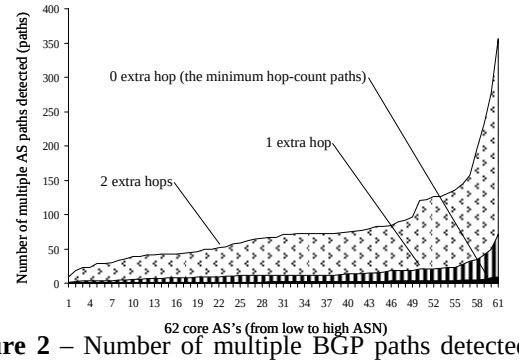

Figure 2 – Number of multiple BGP paths detected at AS 65000

We classified ASes to core or non-core ASes. The core ASes were those that had at least two local links that lead to AS 701, while non-core ASes had only one such a local link. Our analysis detected 62 core ASes in the routing table of AS 65000. We found that 54.8% of the 62 core ASes (34 ASes) had multiple (two or more) shortest BGP paths to AS 701 (the average was 2.2 paths) while the largest number was 9 multiple paths. For multiple paths with up to 1 extra AS hop, the average was 14.3 paths and 64.5% of the 62 ASes had 12 or more paths to AS 701. For up to 2 extra hops, the average was 80.5 paths and 96.7% of the core ASes had 20 or more paths and every AS had at least 9 paths (the largest number was 356 paths). These results suggest that multiple BGP paths are available for most of the cases and multiple-path transmissions can be an effective solution to maximize network resource utilization and reliability.

3. PROTOCOL DESCRIPTIONS

This section describes the implementation of MBGP routing. MBGP uses the following three routers.
MBGP speakers: The BGP speakers that implement MBGP routing. Source MBGP speaker: The MBGP speaker in the AS the source host belongs to. End-peer MBGP speaker: The end-peer MBGP speaker is the one the source MBGP speaker communicates with to dynamically detect multiple BGP paths and adjust traffic load through multiple paths.

Figure 3 shows the source and end-peer MBGP speakers with the default BGP and possible multiple BGP paths between two host computers. The default BGP path is the one existing BGP selected to reach a destination host (*D*) from a source host (*S*). The AS host *S* belongs to (AS1) is the source AS, while AS4 is the destination AS. Router R_1 is the source MBGP speaker and R_{10} is the end-peer MBGP speaker. Routers, R_1, R_2, R_4, R_5, R_7, R_9 and R_{10} are MBGP speakers, while R_3, R_6, R_8 and R_{11} are BGP speakers. Routers that are not either BGP or MBGP speaker are shown as r_x (r_1 through r_{11}). The path shown by the solid links indicates the default BGP path from *S* to *D*. Other links in dotted lines indicate inter-router links that might be used for possible multi BGP paths.

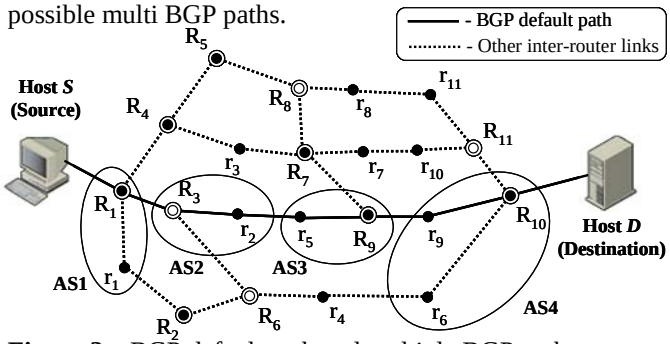

Figure 3 – BGP default path and multiple BGP paths

MBGP routing is implemented by five phases of MBGP initiation, end-peer MBGP speaker discovery, multi BGP-path discovery, MBGP transmissions and termination.

Phase 1: Initiating MBGP routing

MBGP routing is initiated by the source MBGP speaker. MBGP performs multi-path routing to manually specified destination hosts. VPN and VoIP to specific destination hosts are examples for such applications, although MBGP can dynamically activate multi-path routing by monitoring on-going traffic load for other applications.

Phase 2: Discover the end-peer MBGP router

The source MBGP speaker identifies the end-peer MBGP speaker. The source MBGP speaker transmits a DISCOVER _ROUTERS (DR) message as a UDP packet that has the source and destination IP addresses for this source MBGP speaker and a destination host (IP_{SS} and IP_{DH} respectively). MBGP uses UDP packets for all its messages on a particular port. The DR message consists of the two fields. The source MBGP speaker initializes M_{LABEL} as described below and R_{LIST} as an empty list.

- M_{LABEL}: "DISCOVER_ROUTERS" character string
- R_{LIST}: List of MBGP speakers on the default BGP path

Each MBGP speaker that sees a DR message replies with ECHO message, which consists of the following two fields:

- M_{LABEL}: Constant "ECHO" character string
- R_{LIST}: The copy of R_{LIST} field from a DR message

1. If $P.M_{LABEL}$ = "DISCOVER_ROUTERS", then, go to step 2. Otherwise forwards this packet to the next router on the default BGP path and terminates.
2. $P.R_{LIST} \leftarrow P.R_{LIST} + <ASN_{THIS}, IP_{THIS}>$
3. $E.R_{LIST} \leftarrow P.R_{LIST}$ and transmit E to IP_{SS}.
4. If this router belongs to the destination AS, then, terminate. Otherwise proceed to step 5.
5. Forward P to the next router on the default BGP path and terminate.

Figure 4 – Procedure at an intermediate MBGP router

Each MBGP speaker that sees a packet on the UDP port used by MBGP performs the procedure in Figure 4. Assume that P represents a DR message while E represents an ECHO message. ASN_{THIS} and IP_{THIS} hold the AS number and the IP address of this intermediate MBGP speaker. The "+" operator indicates concatenation at the end of R_{LIST}, while "←" indicates assignment. The intermediate MBGP speaker sends this echo message to the source MBGP speaker using the source and destination IP addresses of this MBGP speaker and the source MBGP speaker (IP_{THIS} and IP_{SS}).

In Figure 3, the source MBGP speaker, R_1, transmits a DR message to D through its next router, R_3. Since R_3 is not a MBGP speaker, R_3 ignores the message and forwards it to router r_2 simply as a UDP packet. Since r_2 and r_5 are not a MBGP speaker, they forward the message to r_5 and R_9. When R_9 sees the DR message, R_9 adds its AS number ("AS 3") and IP address to the message. Then, R_9 forwards the message to its next router, r_9. After that, R_9 sends an ECHO to R_1, using IP_{SS}. Similar to r_2 and r_5, r_9 forwards the message to its next router, R_{10}. Since R_{10} is another MBGP speaker, it performs the procedure in Figure 4. Step 4 in Figure 4 prevents R_{10} from forwarding the DR message any farther because R_{10} belongs to the destination AS.

After receiving ECHO messages, the source MBGP speaker identifies the MBGP speaker that replied with the longest R_{LIST} as the end-peer MBGP speaker and notifies the router to be the end-peer speaker by sending an EPS_ NOTIFY (EN) to it, which has the following fields.

- M_{LABEL}: "EPS_NOTIFY" character string
- IP_{SH}: The source host IP address in CIDR prefix
- IP_{DH}: The destination host IP address in CIDR prefix

Phase 3: Discover multiple BGP paths

After the source MBGP speaker notifies the end-peer MBGP speaker, it broadcasts a DISCOVER_PATHS (DP) message towards the end-peer MBGP speaker. The purpose of the DP message is to detect multiple BGP paths to reach the end-peer MBGP speaker. Each DP message consists of the following six fields. Q_{NUM} is assigned a number that uniquely identifies each DP message for the pair of IP_{SH} and IP_{DH}.

- M_{LABEL}: "DISCOVER_PATHS" character string
- IP_{SH}: Same as in EN message
- IP_{DH}: Same as in EN message
- Q_{NUM}: The message sequence number
- L_{ASL}: List of inter-AS links that forms a BGP path
- C_{SPRIT}: Counter of BGP path splits (described later)

The source MBGP speaker initializes each field of the message. The source MBGP speaker adds the identification (AS number and IP address) of itself and next-hop router with its current time stamp at the beginning of L_{ASL}. Then, the source MBGP speaker transmits the message to each of its direct next-hop routers using an UDP packet that has the source and destination IP addresses for the source and end-peer MBGP speakers (IP_{SS} and IP_{EPS} respectively).

1. If $K.M_{LABEL}$ = "DISCOVER_PATHS", go to step 2. Otherwise forwards this packet to the next router on the default BGP path and terminate.
2. Test if $IP_{THIS} \notin K.L_{ASL}$. If the test holds true, proceed to step 3. Otherwise terminate.
3. If $K.IP_{EPS} = IP_{THIS}$, perform the following and terminate. Otherwise go to step 4.
   - $Y.L_{ASL} \leftarrow K.L_{ASL} + ID_{THIS}$
   - $Y.Q_{NUM} \leftarrow K.Q_{NUM}$
   - Transmit Y to IP_{SS}
4. Check the local routing policy to find out if this AS is willing to transfer network traffic from $K.IP_{SH}$ to $K.IP_{DH}$. If not, terminate. Otherwise proceed to step 5.
5. For each next-hop router available at this router, perform following:
   - Find an outgoing link through which $K.IP_{EPS}$ can be reached.
   - Duplicate K to G
   - $G.L_{ASL} \leftarrow K.L_{ASL} + <ID_{THIS}, ID_{NHR}, t_x>$
   - Forward G to the next-hop router.
6. Delete G and repeat step 5.

Figure 5 – Procedure executed for DISCOVER_PATHS message at each intermediate MBGP router

Each MBGP speaker that sees a DP message performs the procedure in Figure 5, where ID_{THIS} is the identification of this MBGP speaker, ID_{NHR} is the identification of the next-hop router and t_x is the timestamp at this MBGP speaker. Any non-MBGP router ignores DP messages (represented by K in Figure 5) and it forwards the messages to its next router without processing them. When the end-peer MBGP speaker receives a DP message, it creates a REPLY (Y in Figure 5) and transmits it to IP_{SS}. Each REPLY consists of the following fields.

- M_{LABEL}: "REPLY" character string
- L_{ASL}: List of inter-AS links that forms a BGP path
- Q_{NUM}: The value used in DP message

The source MBGP speaker collects replies from the end-peer MBGP speaker. L_{ASL} in each REPLY message contains a list of the visited MBGP speakers with the last entry in the list being the one for the end-peer MBGP speaker.

The following scenario visualizes the above procedure in Phase 3. First, the source MBGP speaker (R_1) broadcasts a DP message to its neighbor routers, r_1, R_3 and R_4 towards the destination host D (① in Figure 6). Since r_1 is not a MBGP speaker, it forwards the message to its next-hop router without any modification to the message (②). Router r_1 has only one outgoing link and the message will be forwarded to R_2. If there is more than one outgoing link at a non-MBGP router, the message will be forwarded only to the next-hop router on the default path. Since R_3 is not a MBGP speaker, R_3 forwards the message only to r_2 (③).

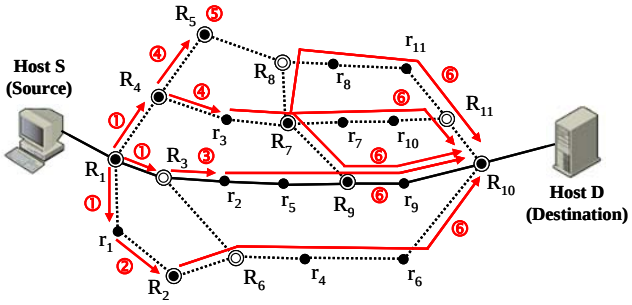

Figure 6 – Detection of five available multiple BGP paths between R_1 and R_{10}

When R_4 receives a DP message, it performs the procedure in Figure 5 (④). Assume that R_4 accepts the traffic from R_1 to R_{10}. R_4 finds an outgoing link to r_3, through which R_{10} can be reached. R_4 adds R_4's and r_3's identification with its timestamp in the DP message. Then, R_4 forwards the DP message to r_3. R_4 finds another outgoing link to R_5 through which R_{10} can be reached. R_4 repeats step 5 in Figure 5 for R_5. Since there is no other outgoing link, R_4 terminates its procedure.

Assume that R_5 is not willing to forward network traffic from R_1 to R_{10}, R_5 drops the message and it terminates the procedure when R_5 receives the DP message from R_4 (⑤). Also assume that R_7 and R_9 are willing to forward the network traffic from R_1 to R_{10}, there will be five different ways the DP message from R_1 can reach R_{10} (⑥). The end-peer MBGP speaker, R_{10}, will receive five DP messages.

When the end-peer MBGP speaker receives a DP message, it sends a REPLY message to the source MBGP speaker. Each time the source MBGP speaker receives a REPLY message from the end-peer MBGP speaker, the source MBGP speaker detects a distinct BGP path between the source and the end-peer MBGP speakers.

In the above example, R_1 will receive five REPLY messages from R_{10}, each of which contains L_{ASL} as listed below, with each representing a BGP path between R_1 and R_{10} (the items between "(" and ")" are the one added by a MBGP speaker).

(a) (R_1, R_4, t_1)-(R_4, r_3, t_2)-(R_7, R_8, t_5)-(R_{10})
(b) (R_1, R_4, t_1)-(R_4, r_3, t_2)-(R_7, r_7, t_6)-(R_{10})
(c) (R_1, R_4, t_1)-(R_4, r_3, t_2)-(R_7, R_9, t_7)-(R_9, r_9, t_0)-(R_{10})
(d) (R_1, R_3, t_1)-(R_3, r_2, t_3)-(R_9, r_9, t_8)-(R_{10})
(e) (R_1, r_1, t_1)-(R_2, R_6, t_4)-(R_{10})

When an intermediate MBGP speaker accepts a DP message at step 4 in Figure 5, it constructs the MBGP routing table (Table 1). Router R_7 constructs its MBGP routing table with "Next-hop Routers" field initially empty.

Source Network Address (IP_{SH})	Destination Host Address (IP_{DH})	Next-hop Routers
146.163.0.0/16	24.181.0.0/16	(blank)

Table 1 – MBGP routing table at router R_7

When R_1 receives a REPLY from R_{10}, it notifies each intermediate MBGP speaker a link to a next-hop router, using IMS_NOTIFY (MN). Each MN message consists of the following fields.

- M_{LABEL}: "IMS_NOTIFY" character string
- IP_{SH}: Same as EN message
- IP_{DH}: Same as EN message
- Q_{NUM}: The value used in DP message
- L_{NHR}: A link to a next-hop router with a time stamp

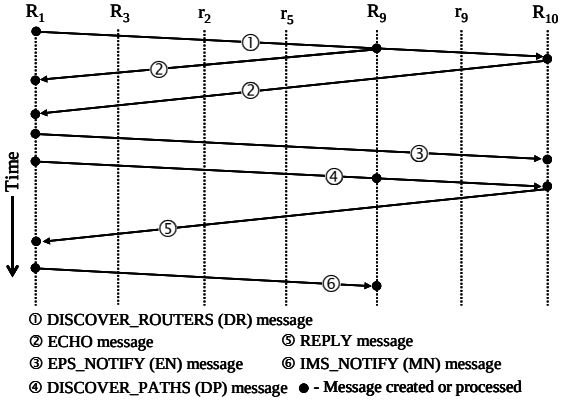

Figure 7 – MBGP messages for multi-path detection and traffic load allocation

For example, when R_1 receives a reply for a DP message from R_{10} that contains BGP path (a) above, R_1 sends (R_8, t_5) to R_7. Similarly, R_1 sends (r_7, t_6) and (R_9, t_7) to R_7 for path (b) and (c). When R_7 receives three MN messages from the source MBGP speaker, R_7 completes "Next-hop Routers" by setting "R_8, r_7, R_9" in Table 1. The source MBGP router does the same for each of the other intermediate MBGP routers, R_2, R_4, and R_9. Figure 7 visualizes MBGP messages exchanged between MBGP speakers. The figure shows the MBGP messages from the routers only on the default BGP path in Figure 6.

For each multiple BGP path, allocation of network traffic load will be computed based on the observed round trip delay. When intermediate MBGP speakers receive multiple MN messages from the source MBGP speaker, they calculate the difference between a timestamp in a MN message and its current time. For example, since R_7 receives three MN messages (one for each multiple BGP path),

MBGP calculates $\Delta t_5 = T_a - t_5$, $\Delta t_6 = T_b - t_6$, and $\Delta t_7 = T_c - t_7$ where T_a, T_b *and* T_c represent the timestamps at R_7 on the arrival of the MN messages for (a), (b) and (c) respectively.

The traffic load weight for a BGP path k specifies the percentage of the traffic load assigned to path k to the total traffic load from IP_{SH} to IP_{DH} at a MBGP router. The traffic load weight is calculated using the following four steps:

Step 1: In m multiple BGP paths, the path with the shortest round trip delay is called the primary BGP path. If there is only one BGP path, it is the primary BGP path and 100% of the traffic load is assigned to the path. Then the procedure is terminated. Otherwise proceed to Step 2.

Step 2: For each of the (m-1) non-primary BGP paths, the ratio of their round trip delay to that of the primary BGP path (designated as λ_1 through $\lambda_{(m-1)}$) is calculated by dividing the round trip delay of a non-primary path, Δt_k, by that of the primary path, Δt_i, (thus $1.0 \le \lambda_k$ for any non-primary path k: $1 \le k \le (m\text{-}1)$).

Step 3: An index to determine the traffic load weight, called the weight score, is calculated for each path. For the primary path, a weight score of 100 is always assigned. For other paths, the weight score is calculated as: if λ_k is at or less than the lower threshold, λ_{Low}, a weight score of 100 is assigned to path k. If it is higher than the upper threshold, λ_{High}, then a score of 0 is assigned. If λ_k is between λ_{Low} and λ_{High}, the weight score is assigned inversely proportional to the distance between λ_{Low} and λ_{High}, using formula (1). For example if λ_k is 1.5 and λ_{Low} = 1.2 and λ_{High} = 2.0, the score for path k will be 62.5.

$$100 \times (((\lambda_{High} - \lambda_{Low}) - (\lambda_k - \lambda_{Low})) / (\lambda_{High} - \lambda_{Low})) \quad (1)$$

Step 4: Add the weight scores for all the paths and the traffic load weight is determined as a ratio of a score for a path to the total score. For example, if there are four paths and the weight scores for the three non-primary paths are 20, 40 and 60, the total weight score will be 220 (20+40+60+100). The traffic load weights for the four paths will be 9.1% (20/220), 18.2%, 27.3%, and 45.5% of the traffic load from IP_{SH} to IP_{DH} at this MBGP speaker.

Some existing ASes have a couple hundreds of links to other ASes. This implies that unlimited flooding of DP messages can cause serious message overhead. To control the message overhead, each DP message contains a field called "counter of BGP path splits". This field (C_{SPRIT}) is initialized by a positive integer before a DP message is transmitted by the source MBGP speaker. Each intermediate MBGP speaker decreases the counter, if the DP message is sent to more than one next-hop router. For example, at R_7 in Figure 6, since the message is duplicated to three next-hop routers, the counter is decreased by 2 (= 3–1).

If a MBGP speaker has more outgoing BGP paths than C_{SPRIT} in a DP message, some of the BGP paths will be selected and C_{SPRIT} is set to 0. Once the counter reaches 0, any MBGP speaker in the downstream will not duplicate the DP message any more. Although the split counter is used as an approximate control over the message overhead, it will control the number of multiple paths.

Phase 4: Start MBGP transmission

After the MBGP routing is started, the source MBGP speaker periodically transmits REFRESH messages. The message consists of the same contents as DP messages except that "counter of BGP path splits" field is not included. The message is propagated in the same way as DP messages. When a REFRESH message reaches a MBGP speaker, the router finds the next-hop routers from the "Next-Hop Routers" field in its MBGP routing table and forwards the refresh message only to those routers. The end-peer MBGP speaker processes each REFRESH message in the same way as a DP message. When the source MBGP speaker receives a REPLY message for REFRESH message, it updates each intermediate MBGP speaker using a MN message.

Using MN messages for refresh, intermediate MBGP speakers dynamically recalculate the traffic load weights based on the latest round trip delay. If a path fails, the source MBGP speaker detects the failure by lack of REPLY message and will notify the failure to every affected intermediate MBGP speaker. The intermediate MBGP speakers then drops the failed path and forwards the next REFRESH message to another link to the destination (if any), which will dynamically activate a new BGP path.

REFRESH messages will be broadcasted with a certain interval (e.g., 1 second). It is the interval changes in the round trip delay and link failures are detected, which allows MBGP speakers to adjust traffic load or detour a failed path.

Phase 5: Termination of MBGP routing

Termination of a MBGP routing is performed by TERMINATE message. The source MBGP speaker broadcasts TERMINATE messages in the same way as REFRESH messages. When an intermediate MBGP speaker sees a TERMINATE message, it deletes the corresponding entry from its MBGP routing table.

4. PERFORMANCE ANALYSIS

We evaluated the performance of MBGP on, resistance to link failures, the network resource utilization, and protocol processing overhead for each MBGP message. Assume that there are m multiple paths P_1, P_2 through P_m, between two MBGP speakers R_2 and R_6 (Figure 8). Path P_i through R_4 (R_2-R_4-R_6) is the default BGP path.

For resistance to link failures, let us assume that the probability of failure for each link in a particular time interval is p ($0 < p < 1.0$). Each BGP path consists of links from the source host to the source MBGP speaker, from the source to the end-peer MBGP speakers, and from the end-peer MBGP speaker to the destination host.

For a BGP path, all the links must be working for successful transmissions of network traffic. Thus, the expected probability for successful transmission will be $(1-p)^{\delta}$, where δ represents the total number of AS links between

the source and destination hosts. For MBGP, let us assume that the number of links between the source host, *S*, and R_2 is α links. The average number of links each of the *m* multiple paths has is β links and the one between R_6 and the destination host is χ links. The improvement in the probability for successful transmission is predicted by formula (2).

$$(1-p)^{\alpha} \times (1-(1-(1-p)^{\beta})^{m}) \times (1-p)^{\chi} - (1-p)^{(\alpha+\beta+\chi)} \qquad (2)$$

The major delay in MBGP routing is for end-peer MBGP router discovery and multi path discovery. However the delay will be transparent from the end hosts since those discovery operations are all performed while their network traffic is going through the default path. As multiple paths are being detected, the network traffic is transparently distributed to dynamically detected multiple paths.

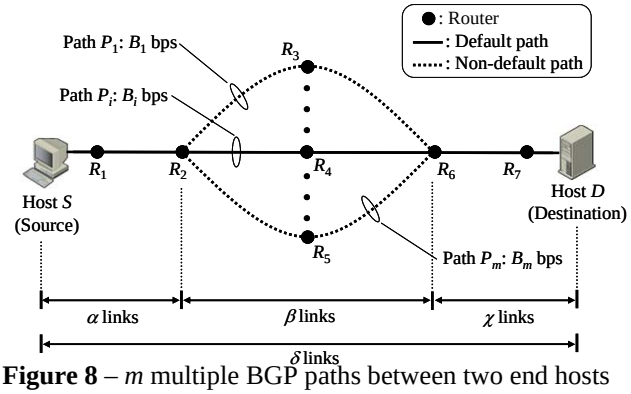

Figure 8 – *m* multiple BGP paths between two end hosts

Routing loops are prevented by Step 2 in Figure 5. The second step makes sure that a MBGP speaker never appears more than once in L_{ASL} in each DP message, which guarantees no routing loop. The complexity for processing each MBGP message at a MBGP speaker will be all $O(1)$ except DP and REFRESH messages. DP and REFRESH messages require $O(N)$, where *N* represents the number of links at a MBGP router.

Regarding bandwidth utilization for the core ASes, assume that the transmission bandwidth utilized by path *k* is B_k. Since only B_i is available for BGP, its utilization will be:

$$U_{BGP} = B_i / \sum_{j=1}^{m} B_j \qquad (3)$$

MBGP can utilize the available bandwidth for *m* multiple paths. Deducting the MBGP message overhead (B_{OH}), the improvement from BGP is predicted by (4), which implies that we will benefit as long as the message overhead is lower than the residual bandwidth in (*m*-1) multiple paths.

$$\left(\sum_{j=1}^{m} (B_j) - B_i - B_{OH} \right) / \sum_{j=1}^{m} B_j \qquad (4)$$

5. CONCLUSIONS AND FUTURE WORK

This paper proposes innovative multi-path inter-domain routing, called MBGP. MBGP will maximize network resource utilization by utilizing residual transmission bandwidth in multiple paths, as well as resistance to unintentional and intentional link failures. Our analysis on the existing multiple BGP paths in today's Internet suggest that multi-path routing can be an effective solution for the inefficiency and reliability problems.

MBGP dynamically and transparently detects multiple paths and adjusts traffic load while the network traffic is going through the default path, which hides the delay for path detection and traffic load balancing. MBGP is designed with backward compatibility to BGP-4 and will work with limited number of MBGP routers deployed in the Internet without routing loops.

Performance analysis suggests that MBGP will improve the network bandwidth resource as long as its message overhead is lower than the total residual transmission bandwidth of multiple paths. Performance analysis predicts that MBGP will improve resistance to link failures as a factor of the number of multiple paths. MBGP has $\boldsymbol{O}(1)$ or $\boldsymbol{O}(\boldsymbol{N})$ complexity for processing a MBGP message. Currently, simulation experiments are conducted by SSFNet simulator [9] using the topology information reconstructed from the actual BGP routing table at existing ASes.

References

[1] Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)," *RFC 1771*March 1995.

[2] M. Yannuzzi and et al., "Open Issues in Interdomain Routing: A Survey," *IEEE Network*, vol. 19, no. 6, pp. 49-56, December 2005.

[3] D. Walton, A. Retuna, and E. Chen, "Advertisement of Multiple Paths in BGP," *Internet Draft*: draft-walton-bgp-paths-04.txt, August 2005.

[4] M. Tufail, "IPv6 - An Opportunity for New Service and Network Features," *Proceedings of the International Conference on Networking and Services*, pp. 11, 2006.

[5] X. Zhao and et al., "An Analysis of BGP Multiple Origin AS (MOAS) Conflicts," *Proceedings of the 1st ACM SIGCOMM Workshop on Internet Measurement*, pp. 31-35, 2001.

[6] M. Lad, R. Oliveira and B. Zhang, "Understanding the Impact of BGP Prefix Hijacks," *Proceedings of ACM SIGCOMM*, 2006.

[7] *CIDR Report*, May 2008: http://www.cidr-report.org/

[8] BGP Table Statistics, http://bgp.potaroo.net/as2.0/bgp-active.html

[9] "SSFNet: Scalable Simulation Framework," http://www.ssfnet.org/