# Fast Network Failure Recovery Using Multiple BGP Routing Planes

Ning Wang, Yu Guo
University of Surrey
United Kingdom

Kin-Hon Ho
Hong Kong Polytechnic University
Hong Kong, China

Michael Howarth
University of Surrey
United Kingdom

George Pavlou
University College London
United Kingdom

***Abstract*— We present an efficient multi-plane based fast network failure recovery scheme which can be realized using the recently proposed multi-path enabled BGP platforms. We mainly focus on the recovery scheme that takes into account BGP routing disruption avoidance at network boundaries, which can be caused by intra-AS failures due to the hot potato routing effect. On top of this scheme, an intelligent IP crank-back operation is also introduced for further enhancement of network protection capability against failures. Our simulations based on both real operational network topologies and synthetically generated ones suggest that, through our proposed optimized backup egress point selection algorithm, as few as two routing planes are able to achieve high degree of path diversity for fast recovery in any single link failure scenario.**

## I. INTRODUCTION

In plain IGP based routing environments the failure of an intra-AS (Autonomous System) link may trigger the underlying routing protocol to re-converge which may take up to several seconds, while BGP re-convergence following an inter-AS link failure may even take much longer time [1]. Such routing disruptions may lead to transient forwarding loops which result in packet loss. On the other hand, an intra-AS link failure can also disrupt BGP routing due to the hot potato routing effect: the IGP distance from any specific router to individual AS border routers (ASBRs) may change after an intra-AS link failure, hence this router may automatically switch to a new egress point if its IGP distance becomes shorter than the post-failure distance to the original default egress point after IGP re-converges. This type of egress router switching caused by intra-AS link failures is very common, and more importantly, *cannot be as easily handled or even anticipated by the ISP as the inter-AS link failure scenario* [2]. As a result, automatically diverted inter-AS transit traffic after IGP re-convergence may unexpectedly overwhelm the alternate egress point or even downstream ASes.

In this paper, we propose a holistic IP Fast Reroute (FRR) technique that not only protects both intra- and inter-AS link failures, but also enables *controlled* egress point switching that avoids *unexpected* BGP routing disruptions due to the hot potato routing effect. The proposed scheme is based on multi-plane aware BGP routing paradigms that enable multiple concurrent routes towards any specific remote destination prefix. Such multi-path BGP routing protocols have been recently proposed in the literature, including R-BGP [3] and BGP Path Splicing [4]. In both schemes, each BGP speaker maintains multiple inter-AS routes towards remote destination prefixes, with the primary route being used in the normal condition and backup routes (typically enforced through tunnels towards different egress points) used in case the primary ones become unavailable, for instance due to network failures.

In this paper, we define BGP *routing planes* in order to indicate multiple BGP routes maintained at each BGP speaker. All the primary routes used in the normal situation are identified as the paths maintained in the default routing plane (plane 0), and other $K$-1 backup routing planes can be defined if each BGP speaker maintains at most $K$-1 backup routes towards each specific destination prefix. If a remote prefix can be reachable via multiple ASBRs, one of them can be used as the unique primary egress point according to the ISP's normal routing policy (through setting the highest *local_pref* value), while the rest can also be *strategically* selected by the ISP as backups for fast recovery against both intra- and inter-AS failures. Towards this end, we introduce in this paper an efficient backup egress point selection algorithm in the backup routing planes in order to maximize the protection coverage of failures. In case a link failure occurs, the local repairing router may intelligently divert the affected customer traffic onto one of the backup routes (through tunnels based on existing implementation of multi-path BGP platforms [3][4]) with an alternate backup egress point in order to avoid passing through the failed link. In addition to this local repair mechanism in the forwarding plane, we also introduce a complementary *crank-back* technique which allows nearby routers to perform traffic diversion in case the directly attached node of a failed intra-AS link does not have any feasible alternate route. The rationale behind is that, it still takes much shorter time to notify feasible routers a few hops away that are able to perform traffic diversion than directly incurring IGP re-convergence across the entire network, which may further cause unexpected BGP disruptions. Detailed FRR operations on top of the multi-plane BGP platform will be specified in section III.

Our simulations based on both real network topologies and synthetically generated ones suggest that, as few as two routing planes (one primary + one backup) are able to achieve fast recovery for both intra- and inter-AS link failures based on carefully selected backup egress points through the proposed algorithm. On the other hand, the failure of a small proportion of network links cannot be directly handled through local repair, but still only a small number of hops of crank-back operation is sufficient to identify a feasible router for diverting the affected traffic.

## II. RELATED WORKS

Various IP FRR techniques have been proposed in the literature for seamless network recovery in order to avoid disruptions to real-time services. Next-hop deflection is a commonly adopted technique that allows local repairing routers to intelligently deflect the affected traffic onto

978-1-4244-4148-8/09/$25.00 ©2009

alternate next-hops that are not necessarily in the default IGP paths towards the destination [5][6]. Despite its simplicity, basic packet deflection is not able to guarantee failure recovery for *every* single link failure scenario. It should be also noted that, "careless" packet deflections may also cause unexpected BGP routing disruption, as the alternate next-hop router may use a different egress point according to its own IGP distance towards individual ASBRs. The Notvia scheme [7], currently being standardized in the IETF, makes use of IP tunnels that are able to automatically bypass network failures with guaranteed 100% failure coverage within a single AS. In [8] A. Kvalbein *et al* proposed to use multi-topology IGPs such as MT-OSPF [9] for achieving fast failure recovery where the affected traffic can be locally remarked to backup routing topologies in case a failure occurs in the default topology. To enable fast recovery in case of inter-AS link failures, O. Bonaventure *et al* proposed an intelligent FRR mechanism that allows the default egress router to immediately divert customer traffic through pre-established IP tunnels towards the secondary egress point once the primary route via its directly attached inter-AS link becomes unavailable [1]. It should be noted that existing IP FRR solutions deal with intra- and inter-AS failures separately, in which case dedicated mechanisms need to be applied against different types of failures. In contrast, we propose a holistic solution that is able to protect against both types of failures, and more importantly, to enable predictable and controlled egress point switching against the hot potato routing effect.

## III. MULTI-PLANE BGP FRR OVERVIEW

Before introducing the proposed scheme, we first review the basic procedure of conventional IGP re-convergence and its potential impact on BGP routing decisions. For simplicity we assume a full-mesh i-BGP connection within the network. Once an intra-AS link fails, its directly attached router will send updated link state advertisements (LSAs) to notify other nodes about the failure. After all routers have re-computed new IGP routes (which may take up to seconds for RIB/FIB updating), the IGP distance from some routers towards individual ASBRs may change, in which case these routers may further change their egress point selection decisions for some remote destination prefix accordingly. Again, this procedure takes additional time, and meanwhile the original BGP routing configuration can be disrupted due to the unexpected egress point switching.

In order to (1) achieve seamless failure recovery and (2) avoid potential BGP routing disruptions due to unexpected egress point switching, we introduce our proposed FRR scheme for carefully pre-provisioning backup BGP routes that can be completely controlled by the ISP. The main idea is that both the primary and backup egress points are strategically selected *a priori* in the default and backup routing planes respectively for each remote prefix (*one single* egress per plane for each prefix). In case an intra-AS link fails, the local repairing router immediately diverts the affected transit traffic away from the failure and sends it towards a desirable backup egress (through pre-installed MPLS/L2TP tunnels) that does not involve the failed link. Similarly, in case of an inter-AS link failure, the directly attached ASBR may also forward the affected traffic through tunnels towards pre-selected backup egress points, which is similar to what has been proposed in [1]. Compared to this more straightforward operation, we mainly focus on how to achieve fast failure recovery against intra-AS failures that may potentially cause unexpected BGP routing disruptions.

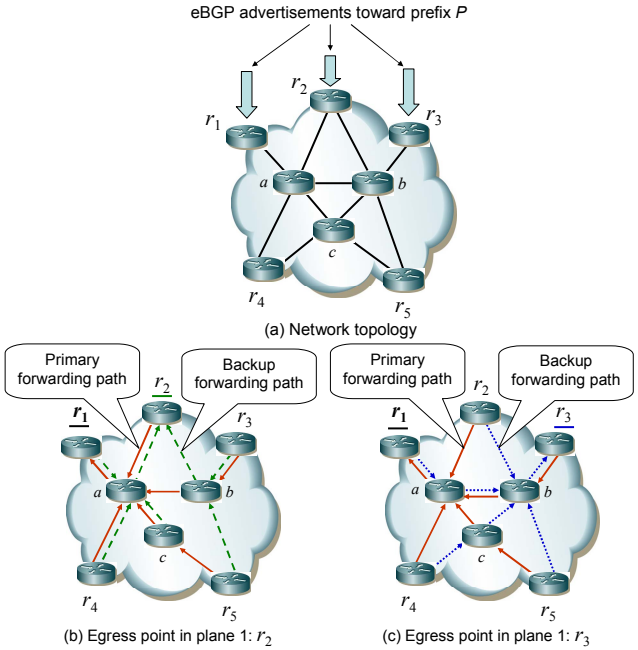


Figure 1. Multi-plane BGP reroute examples

Consider Figure 1 as an example where the IGP weight for each network link is set to 1. Inter-AS traffic is injected into the network via ASBRs *r1*-*r5*, and among them *r1*, *r2* and *r3* have BGP routes towards a remote prefix *P* (Figure 1(a)). These routes, which are advertised by neighboring ASes, can be obtained from the BGP *Adj-RIBs-In*. By provisioning two BGP routing planes (i.e. one primary and one backup route for each prefix), we first assume that the egress points selected for that prefix in plane 0 (primary) and plane 1 (backup) are *r1* and *r2* respectively (Figure 1(b)). Under such configuration, the next-hop toward *P* at router *b* points to router *a* in plane 0, as router *a* is the next-hop on the shortest IGP path to reach the selected egress point *r1* (shown in the solid line). Meanwhile, the backup forwarding path from router *b* towards prefix *P* leads to the alternate egress point *r2*. Under the normal condition where no link failure occurs, the default forwarding table populated based on the BGP RIB in plane 0 is used at all routers. As shown in Figure 1(b), all packets towards the destination prefix *P* are sent towards the primary egress point in plane 0 (i.e. router *r1*) from where they are delivered out of the local AS. The actual forwarding paths from ingress routers towards *r1* are indicated with solid lines in the figure.

### A. Local repair

Once a router has detected the failure of its directly attached link, it needs to immediately divert the affected traffic which originally uses that link to reach the corresponding destination prefixes in plane 0. In doing so, this router, which is known as the local repairing router, switches to the backup path (effectively an MPLS/L2TP tunnel) in plane 1 leading to the backup egress point. Consider again Figure 1(b) as an example. Once router *b* detects the failure of link *b*→*a*, it immediately diverts the

978-1-4244-4148-8/09/$25.00 ©2009

affected traffic towards prefix *P* to the backup egress point *r2* through the pre-established tunnel. In this case the actual forwarding path from ingress router *r3* towards the backup egress point is *r3*→(solid line)→*b*→(dash line)→*r2*. Since this tunnel is MPLS/L2TP based (instead of IP-in-IP), *r2* will not consult its own BGP table and return the diverted packets back to the original primary egress point *r1* which is assigned with the highest *local_pref* value [1].

In case of an inter-AS link failure (i.e. the loss of eBGP sessions), the directly attached default egress router may immediately divert the affected traffic to a pre-selected alternate egress point. For instance, in Figure 1(b) if *r1* detects the unavailability of the primary route through its inter-AS link, it activates the pre-established tunnel as the backup forwarding path (the dash path) in plane 1 towards the backup egress *r2*, and from there the packets are delivered out of the local AS.

### *B. Crank-back operations*

Under certain circumstances it is not possible for the head router of a failed intra-AS link to directly find any feasible alternate path that can successfully bypass the failure. For instance, in Figure 1(b) link $c \rightarrow a$ constitutes both the primary path (towards the default egress point *r1*) and the backup tunnel (towards backup egress point *r2*) for prefix *P*. In this case, link *c*→*a* is regarded as a *critical link* for prefix *P* which means this link is fully shared by both primary and backup planes and hence the head node *c* does not have any alternate route to bypass this link once it fails. In such a situation, we introduce a simple IP *crank-back* mechanism at core routers that allows previous-hop nodes to perform rerouting without forcing the entire network to re-converge. In the literature, crank-back operations have been proposed for MPLS-based failure recovery [10], but how this can be achieved in hop-by-hop based IP rerouting has not been investigated. As previously mentioned, updating IP forwarding tables at individual routers accounts for most of the time spent in IGP re-convergence. In contrast, crank-back operations only introduce very short time in notifying nearby routers (not necessarily back to ingress routers!) to switch to *pre-installed* backup paths in case of failures, which is significantly quicker. Now we continue with the previous example where *c* is about to deal with its local link failure. Since *c* itself does not have any alternate route available, it broadcasts a *route-failure notification* message *Rt_FAIL(P)* for prefix *P* on all the other network interfaces except the failed one (i.e. to routers *b*, *r4* and *r5*). For scalability purpose, one such *route-failure notification* message may contain multiple affected prefixes in order to avoid broadcasting excessive dedicated messages in case a large number of prefixes are affected due to the same link failure. Nevertheless for simplicity we only illustrate with one prefix in our example. Detailed design of packet structure for *route-failure notification* messages is not specified in this paper. For each of those neighbors that receive a *route-failure notification*, if its next-hop towards prefix *P* in the default forwarding table is not the interface that received this message, it does not need to take any action. For instance in Figure 1(b) routers *r4* and *b* simply drop this message as their default next-hops to reach *P* do not point to router *c* in normal forwarding. As router *r5* finds the interface that receives the *route-failure notification* is exactly the one used as the default next-hop towards *P*, it will find a feasible backup tunnel to deliver the affected packets via an alternate egress point (*r2* in this example). As a result the actual backup path is *r5*→(dash line)→*b*→(dash line)→*r2*. In case still no alternate routes can be found in the backup planes, the intermediate router will further forward the *route-failure notification* to all its interfaces except the one that has received it in order to continue the crank-back procedure. The entire procedure terminates either until a router that is able to successfully find a feasible alternate route for traffic diverting, or when an ingress border router (like r5) is reached. If the ingress node still does not have any feasible route, IGP re-convergence has to be performed in order to regain connectivity, as it is the case for most existing IP FRR schemes that are not able to guarantee 100% protection coverage. Figure 2 shows the basic operations for an intermediate router *r* that has received a route-failure notification message from one of its neighbors. In the figure, $NH_r^0(P)$ indicates router *r*'s *default* next-hop for forwarding traffic destined to prefix *P* in the normal condition, and $Path^k(u \rightarrow v)$ represents the path from router *u* to *v* in plane *k*.

**Operations for Router *r* to perform crank-back**

In the event of receiving message *Rt_FAIL(P)* from interface *x*:

**If** $NH_r^0(P) \neq x$ **Then** no action is taken

**Else** exam each of the selected backup egress points $j_k(P)$ for *P*, $(0 < k < K)$

**If** find $\exists k^* \rightarrow x \notin Path^{k^*}(r \rightarrow j_{k^*}(P))$

**Then** divert the affected traffic towards *P* using plane $k^*$ through the corresponding tunnel to $j_{k^*}(P)$

**Else** forward message *Rt_FAIL(P)* to *Neighbor(r)\{x}* to continue crank-back

Figure 2. Crank-back operation

## IV. OPTIMIZING BACKUP EGRESS POINT SELECTION

Although the crank-back mechanism at core routers provides additional capability for fast failure recovery, the procedure inevitably takes longer time compared to pure *local* repairs. Moreover, for some failure scenarios crank-back may still not be able to identify any feasible diverting router at all. In order to improve this situation, how to optimally pre-determine backup egress points needs to be carefully considered in the routing plane. In the previous example, if router *r3* is selected as the backup egress point in plane 1, as indicated in Figure 1(c), we can see that the repairing router *c* is able to perform local repair by using *r3* as the backup egress point for diverting the affected traffic towards *P* (shown in dot lines). In effect, in this specific example local repair is sufficient to deal with any single link failure that happens to all the nodes with more than one outgoing links, without resorting to crank-back operations. From this example we can clearly see the benefit of intelligent egress point selection in backup planes in order to achieve high degree of *path diversity* from each potential repairing router to the selected egress points. In this paper, we consider *single* link failure since it is the predominant form of failure in communication networks [15].

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

We formulate the task of back egress point selection into an optimization problem that can be solved with a greedy algorithm. The physical network topology of an AS can be modeled as a directed graph ($V$, $E$) with node set $V$ and link set $E$. Each AS has a border router set $J \subset V$, through which (1) eBGP reachability advertisements on remote prefixes are received from neighboring ASes, and (2) customers inject traffic into the network. In addition, an AS may contain some core routers that are not directly connected to local customers or other ASes. We consider each remote destination prefix $P$ that can be reached through a set of ASBRs. Let *Out(P)* denote the set of ASBRs at which an advertisement for prefix $P$ has been received. In BGP multi-plane routing, we consider $K$ logical planes to be pre-provisioned in the local AS so that one dedicated egress router can be selected for each destination prefix $P$ within each plane $k$ ($0 \le k < K$). More specifically, one primary egress point is selected in the default routing plane 0, while up to $K$-1 egress points are selected in backup planes $k$ ($0 < k < K$). The total number of backup planes (i.e. the maximum allowable backup egress points for any specific prefix) can be determined by the ISP's policies. Nevertheless, later in the paper we will show that one single backup routing plane will normally be sufficient for comprehensive failure protection, with very short distance of crank-back for a small proportion of critical links.

It can be easily inferred that if a critical link fails, there are no alternate paths in any plane for its head node to directly divert the affected traffic. In this case, backup egress router selection that incurs minimum number of critical links is desirable. Towards this end, we define a binary variable $Q^l(P)$ to indicate whether intra-AS link $l$ is a critical link with regard to its head node and remote destination prefix $P$. More specifically:

$$Q^l(P) = \begin{cases} 1 & if \sum_{k=0}^{K-1} Y^{l,k}(P) = K \\ 0 & otherwise \end{cases}$$

where

$$Y^{l,k}(P) = \begin{cases} 1 & if\ l \text{ constitues the path in plane } k \text{ from its head node to prefix } P \\ 0 & otherwise \end{cases}$$

As we have mentioned, the probability of having critical links can be influenced by egress router selection across individual backup planes. We define another binary variable $X^{j,k}(P)$ to indicate the actual egress point selection for prefix $P$ in each plane $k$. Single egress point selection is adopted in our scheme, which means within each plane one single egress point is selected for each prefix across *all* BGP speakers. That is:

$$X^{j,k}(P) = \begin{cases} 1 & if\ j \text{ is selected for prefix } P \text{ as the primary egress router in plane } k \\ 0 & otherwise \end{cases}$$

In summary, the overall objective is to determine the value of a set of $X^{j,k}(P)$ for each independently advertised prefix $P$ in each routing plane $k$ in order to:

$$\text{Minimize} \sum_{l \in E} Q^l(P)$$

subject to the following constraints:

$$\text{If } X^{j,k}(P) = 1, \text{ then } j \in Out(P) \quad \forall j \in J,\ 0 \le k < K \qquad (1)$$

$$X^{j,k}(P) \in \{0,1\}, Y^{l,k}(P) \in \{0,1\} \quad \forall j \in J,\ 0 \le k < K \qquad (2)$$

Constraint (1) means the selected egress router $j$ must be able to reach the destination prefix $P$ in the first place. Constraint (2) makes sure that variables $X$ and $Y$ are binary.

A simple greedy algorithm for solving the backup egress point selection problem is briefly described as follows. First of all, the default egress point for a prefix $P$ under the normal condition is selected in plane 0 according to the ISP's operational objectives such as conventional traffic engineering (TE) policies. Backup egress point selection is then performed plane by plane with the objective of maximizing the diversity as compared to those trees that have already been previously determined. Each of these trees can be described as the IGP path set from each router towards the selected egress router in each plane. When the egress point is to be selected in plane $k$, all candidate border routers in *Out(P)* that have not been selected in planes 0 to $k$-1 are examined. The one that incurs the least number of critical links with the trees that have already been determined is selected. If multiple candidate egress points exist with equal number of critical links, then the one with the lowest anticipated post-failure traffic load will be selected as tie-breaking. Of course, in order to achieve this type of load balancing after traffic diversion, the ISP needs to estimate in advance the inter-AS traffic matrix. The egress point selection algorithm for each prefix $P$ is shown in Figure 3.

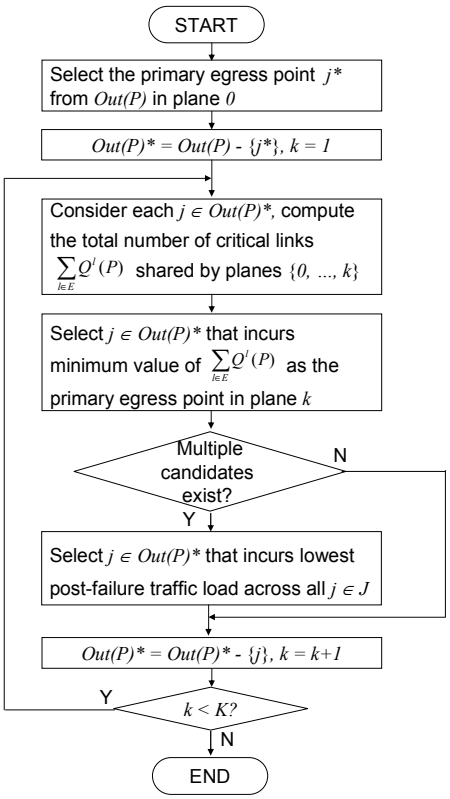

Figure 3. Egress point selection algorithm

## V. PERFORMANCE EVALUATION

In order to evaluate the performance of our proposed fast failure recovery scheme, we use the topologies of two operational networks, namely the Abilene network (AS11537) [11] and the GÉANT network (AS20965) [12]. Information on inter-AS connections of these two networks is obtained from Rocketfuel [13] and [12] respectively. In

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

addition, we also conducted experiments based on synthetically generated topologies by the BRITE topology generator [14]. The topologies contain 50 nodes with border routers being randomly selected each time.

Figure 4 shows the average proportion of critical links across all the independently advertised prefixes in both the GÉANT and Abilene topologies. We can see that with optimized backup egress point selection for each prefix (indicated by *Opt*. in the figure), only 26.5% and 12.2% of the network links are critical ones with two routing planes (i.e. one primary and one backup) in the two network topologies respectively. Three planes are sufficient to eliminate any critical link in the Abilene topology, while the corresponding proportion is reduced to 18.4% in the GÉANT network. However the situation is not significantly further improved with additional backup routing planes. On the other hand, we also implemented the non-optimized scheme with *random* selection of backup egress points (indicated by *Ran*. in the figure). As we can also see, the proportion of critical links in this case becomes much higher, especially when a small number of planes are used. Overall, five routing planes are needed in order to eliminate all critical links in the Abilene topology. This observation clearly indicates that the proposed failure recovery mechanism needs to be accompanied with careful backup egress point selection for achieving maximum efficiency.

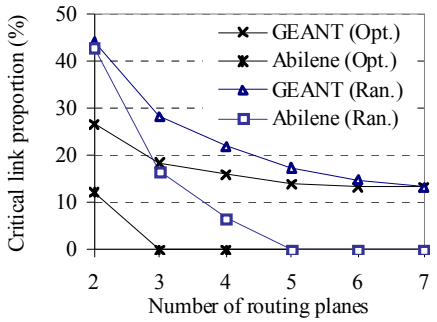


Figure 4. Proportion of critical links vs. number of planes (Real topologies)

TABLE I. DISTRIBUTION OF CRANK-BACK PATH LENGTH IN GÉANT AND ABILENE (OPTIMIZED EGRESS SELECTION)

| | Crank-back path length (No. of hops) | | | |
|---|---|---|---|---|
| | **1** | **2** | **3** | **4** |
| **2 planes** | 57.14% | 14.29% | 28.57% | 0% |
| **3 planes** | 0% | 0% | 0% | 0% |

(a) The Abilene topology

| | Crank-back path length (No. of hops) | | |
|---|---|---|---|
| | **1** | **2** | **3** |
| **2 planes** | 80.26% | 19.74% | 0% |
| **3 planes** | 86.36% | 13.64% | 0% |
| **4 planes** | 84.48% | 15.52% | 0% |
| **5 planes** | 82.69% | 17.31% | 0% |
| **6 planes** | 82.00% | 18.00% | 0% |
| **7 planes** | 82.00% | 18.00% | 0% |

(b) The GÉANT topology

We further evaluate the performance of crank-back operations on those critical links that cannot be eliminated through optimized backup egress point selection. Table I shows the crank-back path length distributions (average values across all examined prefixes) in both network topologies with *optimally* selected backup egress points. As far as the Abilene topology is concerned, feasible diverting routers can be found by cranking-back with one single hop for 57.14% of the critical links if two planes (i.e. one backup path for each prefix) are provisioned. The worst situation is that three hops of crank-back is needed for 28.57% of critical links in the topology. For the GÉANT scenario, although critical links cannot be fully eliminated using as many as seven planes (see Figure 4), the good news is that only two hops of crack-back is sufficient to identify a feasible diverting router for any critical link with as few as two routing planes. In effect, only less than 20% of the critical links need two hops of crank-back in such a situation. For comparison purpose, we also evaluated the performance with *random* selection of backup egress points. As it can be inferred from Table II, the average crank-path length becomes significantly higher than that used by the optimized algorithm with the same number of routing planes. By comparing between the results in Table I and II, we can see that *optimized backup egress selection minimizes not only the proportion of critical links, but also the crank-back path length for more efficient routing failure recovery*.

TABLE II. DISTRIBUTION OF CRANK-BACK PATH LENGTH IN GÉANT AND ABILENE (RANDOM EGRESS SELECTION)

| | Crank-back path length (No. of hops) | | | |
|---|---|---|---|---|
| | **1** | **2** | **3** | **4** |
| **2 planes** | 21.74% | 17.39% | 26.09% | 34.78% |
| **3 planes** | 65.22% | 21.74% | 13.04% | 0% |
| **4 planes** | 92.31% | 7.69% | 0% | 0% |

(a) The Abilene topology

| | Crank-back path length (No. of hops) | | | |
|---|---|---|---|---|
| | **1** | **2** | **3** | **4** |
| **2 planes** | 59.38% | 33.33% | 6.25% | 1.04% |
| **3 planes** | 77.22% | 21.52% | 1.27% | 0% |
| **4 planes** | 80.60% | 19.40% | 0% | 0% |
| **5 planes** | 83.61% | 16.39% | 0% | 0% |
| **6 planes** | 81.82% | 18.18% | 0% | 0% |
| **7 planes** | 79.17% | 20.83% | 0% | 0% |

(b) The GÉANT topology

We also evaluated the same performance metrics with synthetically generated network topologies that contain 50 nodes, with border routers that have inter-AS connections (i.e. egress point candidates) varying from 5 to 25. An important objective is to investigate the performance of failure protection coverage with various richness of inter-AS routes that can be reflected by the total number of egress point candidates that may receive advertised BGP reachability messages via eBGP. As shown in Figure 5, the provisioning of two routing planes leads to 14.18% of critical links with 5 egress point candidates. The corresponding value decreases to 8.37% and 7.55% respectively if the total number of egress point candidates increases to 10 and 15. Further increase of the richness in inter-AS routes almost does not further improve the situation and hence is not shown in the figure. On the other hand, the proportion of critical links reduces as the number of routing planes increases up to 4, but any additional routing planes will not be able to improve the performance beyond that point, which is similar to the GÉANT and Abilene scenarios. Figure 6 indicates the proportion of critical links with random egress point selection. By comparing Figures 5 and 6, once again we can clearly see that non-optimized backup egress point selection leads to much higher proportion of critical links, which results in poor failure recovery performance.

978-1-4244-4148-8/09/$25.00 ©2009
This full text paper was peer reviewed at the direction of IEEE Communications Society subject matter experts for publication in the IEEE "GLOBECOM" 2009 proceedings.

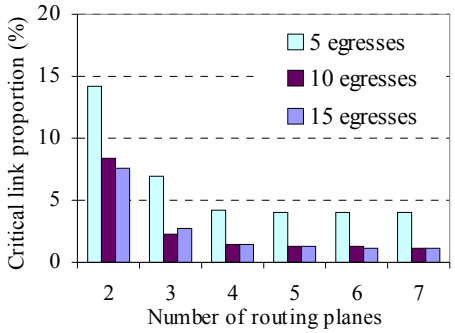


Figure 5. Proportion of critical links vs. number of planes (synthetically generated topologies with *optimally* selected backup egress points)

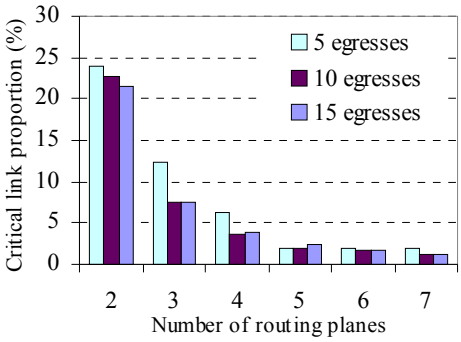


Figure 6. Proportion of critical links vs. number of planes (synthetically generated topologies with *randomly* selected backup egress points)

TABLE III. DISTRIBUTION OF CRANK-BACK PATH LENGTH IN SYNTHETICALLY GENERATED TOPOLOGIES (OPTIMIZED)

| | | Crank-back path length (No. of hops) | | |
|---|---|---|---|---|
| | | 1 | 2 | 3 |
| 5 egress Routers | 2 planes | 89.09% | 10.91% | 0% |
| | 3 planes | 87.88% | 12.12% | 0% |
| | 4 planes | 96% | 4% | 0% |
| | 5 planes | 100% | 0% | 0% |
| 10 egress Routers | 2 planes | 91.43% | 8.57% | 0% |
| | 3 planes | 100% | 0% | 0% |
| 15 egress Routers | 2 planes | 90.62% | 9.38% | 0% |
| | 3 planes | 100% | 0% | 0% |
| 20 egress Routers | 2 planes | 86.11% | 13.89% | 0% |
| | 3 planes | 100% | 0% | 0% |
| 25 egress routers | 2 planes | 91.18% | 8.82% | 0% |
| | 3 planes | 100% | 0% | 0% |

Finally, Table III shows the crank-back path length performance by the optimized backup egress point selection algorithm, based on the synthetically generated topologies. In case of relatively scarce inter-AS routes, for instance with only five egress point candidates, five routing planes are needed to guarantee maximum 1-hop crank-back for all critical links. When the number of ASBRs becomes as high as ten, three planes are sufficient to achieve the same effect. On the other hand, we notice that in every single scenario the maximum number of crank-back hops is two even if two routing planes are provisioned. Moreover, in most of the cases around 90% of critical links can be tackled with a single hop of crank-back. Once again the efficiency of our proposed scheme is indicated with optimized selection of backup egress points in the routing plane.

## VI. CONCLUSION

In this paper we introduce a novel fast failure recovery scheme based on multi-plane BGP reroute. Once an intra- or inter-AS link fails, the directly attached repairing router may immediately divert the affected traffic towards optimally selected alternate tunnels that are pre-installed in backup routing planes. A distinct benefit from the proposed scheme is that routing disruptions caused by intra-AS link failures (due to the hot potato routing effect) can be avoided, as the affected traffic will be always diverted to the egress points that are pre-determined by the ISP, rather than unexpectedly switching to undesired ones which may therefore suffer from post-failure congestions. In addition, we also proposed routing optimizations in terms of backup egress point selection for enhancing the failure recovery capability. Our simulations based on both real and synthetically generated network topologies show that only a small number of routing planes will lead to high degree of path diversity for fast reroute based on carefully selected backup egress points. These results indicate that the proposed paradigm can be regarded as an efficient and scalable solution for supporting high reliability in real-time multimedia communications.

## REFERENCES

[1] O. Bonaventure et al, "Achieving Sub-50 Milliseconds Recovery upon BGP Peering Link Failures", IEEE/ACM Trans. on Netw. Vol 15, No 5, October 2007

[2] R. Teixeira et al, "TIE Breaking: Tunable Inter-domain Egress Selection", IEEE/ACM Trans. on Netw. Vol. 15, No. 4, 2007

[3] N. Kushman et al, "R-BGP: Staying Connected in a Connected World", Proc. USENIX NSDI, April 2007

[4] M. Motiwala et al, "Path Splicing", Proc. ACM SIGCOMM 2008

[5] A. Atlas, "Basic Specification for IP Fast-Reroute: Loop-free Alternates", IETF RFC 5286, September 2008

[6] S. Nelakuditi et al, "Fast Local Rerouting for Handling Transient Link Failures", IEEE/ACM Trans. on Netw. Vol. 15, No. 2, 2007

[7] M. Shand et al, "IP Fast Reroute with Notvia Addresses", IETF Internet draft, February 2008, work in progress

[8] A. Kvalbein et al, "Fast IP Network Recovery using Multiple Routing Configurations", Proc. IEEE INFOCOM 2006

[9] P. Psenak et al., "Multi-Topology (MT) Routing in OSPF", RFC 4915, June 2007

[10] A. Farrel et al, "Crank-back Signaling Extensions for MPLS and GMPLS RSVP-TE", RFC 4920, July 2007

[11] http://www.stanford.edu/services/internet2/abilene.html

[12] http://www.geant.net/upload/pdf/GEANT_Topology_12-2004.pdf

[13] N. Spring et al, "Measuring ISP Topologies with Rocketfuel", IEEE/ACM Trans. on Netw. Vol. 12, No. 1, 2004

[14] The BRITE topology generator, http://www.cs.bu.edu/brite/

[15] A. Markopulou et al, "Characterization of Failures in an Operational IP Backbone Network", IEEE/ACM Trans. on Netw., Vol. 16, No. 4, August 2008

## ACKNOWLEDGEMENT

This work is partially funded by the EU IST FP6 EMANICS Project (http://www.emanics.org).

978-1-4244-4148-8/09/$25.00 ©2009