Scalable Resilient BGP – Fast Recovery from Transient Inter-Domain Link Failures

Hailong Ma[1], Jianwei Zhang[2,1], Yunfei Guo[1], Lei He[1]
[1]National Digital Switching System Engineering and Technology Research Center, Zhengzhou, Henan, 450002, China
[2]School of Computer Science and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou, Henan, 450002, China
{mhl,zjw,gyf,hl}@mail.ndsc.com.cn

Abstract

As more and more mission critical services emerge on the Internet, there is a growing demand for the Internet to provide stringent service availability and reliability. Many studies show that inter-domain links fail as frequently as intra-domain links. The performance of those services in the present of failures would be greatly deteriorated. In this paper, we propose a scalable and resilient inter-domain routing protocol, called SR-BGP, to dynamically establish a protected tunnel. Once the inter-domain link failure is detected, routers could react quickly to perform non-stop routing relying on the pre-established tunnel. Through simulation we demonstrate that SR-BGP is more resilient to link failures than BGP and the inflation of routing table size due to the deployment of SR-BGP is acceptable.

1. Introduction

The Internet has changed its roles from the special network for data transmission to the critical information infrastructure for personal and business applications. More and more mission critical services such Voice-over-IP (VoIP) applications and online games and other real-time applications have been deployed over an IP-based infrastructure, there is a growing demand for the Internet to provide reliable services. To support those mission critical applications, networks should be able to guarantee very stringent Service Level Agreements (SLA). Under the situation where the network is stable, ISP could easily provide the performance guarantees required by their customers. Unfortunately, the links failures are fairly common in the Internet [2][3][4][17][18][11][14]. Wang et al. show that a single inter-domain link failure can produce hundreds of loss bursts which may last for up to 20 seconds [9]. In today's high speed networks, even a short time can cause huge packet losses. For example, if an OC-48 link is down for ten seconds, close to 3 million packets could be lost [15]. Furthermore, those link failures incur BGP re-convergence during which some networks may suffer transient disconnectivity and those mission critical services are interrupted. Kushman et al. report that half of VoIP outages occur within 15 minutes of a BGP update [13].

It spends a few seconds for BGP to re-converge a new stable state, because the advertisement time of link failure event is restricted by a rate-limiting timer, called Minimum Route Advertisement Interval (MRAI) timer, which determines the minimum amount of time that must elapse between routing updates to a particular destination for the same neighbor [12]. Before the Internet re-converges to a new stable state, the routing tables of different routers may exhibits short-term inconsistencies due to asynchronous route computation. The inconsistencies cause some networks not to be reached. Hence, *the objective of this work is to design a scalable and resilient routing which ensures that Internet domains are continuously connected as long as policy compliant paths exist in the underlying network.*

Traditional approaches deal with link failures in a reactive manner. They have focused on shrinking convergence times [6][7][8][16], however, such methods are limited by the size of the Internet and the complexity of the BGP. As an alternative, proactive approaches are brought out. Bonaventure et al [5] propose a solution using pre-established tunnels to reroute traffic during link failures. This approach is appropriate for resolving the problem of transient routing failures occurring on the eBGP peering links. R-BGP [1] precomputes a few strategically chosen failover paths and maintains enough state consistency across domain to ensure continuous path availability. A more recent method, BRAP [4] can locally provide alternate routes upon the occurrence of failures. All these proactive methods guarantee Internet domains quickly recovery from transient link failures. However, the limitation is that the required forwarding memory has to be doubled.

In this paper, we propose a new fast recovery technique, called Scalable Resilient BGP (SR-BGP), which could provide fast restoration in milliseconds order upon inter-

978-0-7695-3497-8/08 $25.00 © 2008 IEEE
DOI 10.1109/IITA.2008.96

IEEE computer society

domain link failures. The most important is that SR-BGP would not cause the inflation of BGP routing table size. The experiments results manifest that the routing table size increase 1.2% of current at most. At the same time, the resilience to link failures of SR-BGP is similar to other approaches.

The remainder of this paper is structured as follows. In Section 2 we describe packet loss due to inter-domain link failure. In Section 3, we present SR-BGP in more detail. In Section 4, we evaluate the resilience and scalability of SR-BGP. Lastly, we conclude in Section 5.

2. Packet Loss During Transient Inter-Domain Link Failure

In this section, some terminologies are defined through examples. And then, we illustrate how burst packet loss can happen during transient inter-domain link failure.

2.1. Definitions

BGP is a policy-based protocol, in which each BGP router maintains routing information learned from neighbors, selects the best route based on local policy rather than on the shortest AS-path and advertises the best route to its proper neighbors following export policy.

Definition 1: If router U's best path toward a destination D is via router V_1, V_1's best path toward a destination D is via router V_2, and in turn, V_{n-1}'s best path toward a destination D is via router V_n. So, the path $(U,V_1,V_2,\ldots,V_n,D)$ is called the *primary path* of router U. V_1 is said to be the *primary neighbor* of U, while U is defined as the *reverse neighbor* of V_1.

Definition 2: Assuming V is the primary neighbor of router U, simultaneously, U has another policy-compliant path $(U, W_1, W_2,\ldots, W_m,D)$. If $W_1 \neq V$, the path $(U, W_1, W_2,\ldots, W_m,D)$ is called *backup path* of router U, and the router W_1 is said to be *backup neighbor* of router U.

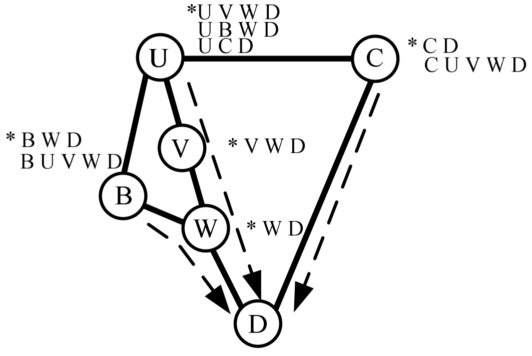


Figure 1 Example used for illustrating concepts throughout the paper. The note around a node represents its routing table and the asterisk indicates that it is a selected route. The dashed line represents the primary path of the corresponding routers.

For example, in Fig.1, router V and path (U,V,W,D) is the primary neighbor and primary path of router U, respectively. Router U is the reverse neighbor of router V. The path (U,C,D) is the backup path of router U, and the path (C,U,V,W,D) is the backup path of router C. Router U and C are the backup neighbor of each other.

2.2. Packet Loss During Inter-Domain Link Failure

When the inter-domain link goes down, BGP takes a long time to find another usable route as described in Section 1. We use an example shown in Fig. 1 to demonstrate packet loss during inter-domain link failure. Router B and V consider router W as next hop to access the destination. Since router W is the primary neighbor of router B and V, BGP's poison-reverse policy forbid router B and W to propagate their best paths back to W. Thus, router W does not realize the existence of the backup path (W,V,U,C,D). If the link between W and D fail, router W loses its route to the destination. In order to re-converge to a new state, router W had to send a withdrawal message to activate U to advertise its backup path (U,C,D). Before router W receives this new path, it will discard all packet destined to D.

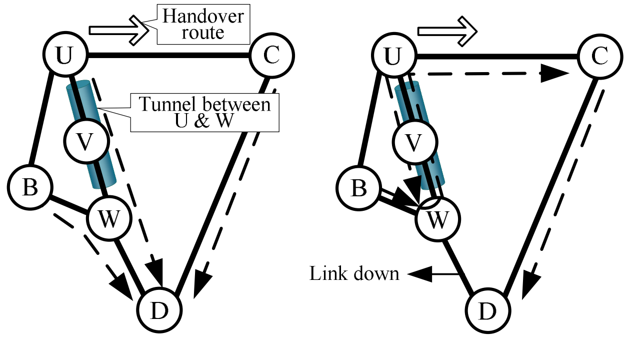
3. Fast Reroute with Protected Tunnel

We present SR-BGP in detail in this section. Firstly, we provide an overview of our method. Then, the setup procedure of protected tunnel is described. Lastly, the route information carried by BGP Update message is presented.

3.1. Overview

Our objective is to design a scalable and resilient routing that can recover quickly so as to guarantee that no packets are discarded upon the occurrence of link failures.

The primary paths of routers to destination compose of a tree whose root is the destination. Furthermore, BGP adopts incremental update to advertise the change of route. This means that link failure influences part of routers whose primary path includes the failed link. Simultaneously, there exist many routers which routes will not change because their primary paths don't pass the failed link. The path which dose not include the failed link is called *safe path*, otherwise, is called *failed path*. This motivates us to design SR-BGP that dynamically establishes a *protected tunnel*, relying on which the traffic arriving at the upstream router of the failed link can be rerouted to another router which has safe path so as to ensure continuous connectivity. For example, in Fig. 2, SR-BGP deems router U as a protected tunnel egress and installs handover route whose next hop is router C. Then, router U announces to its primary neighbor V tunnel route which is propagated to router W. After that, a protected tunnel T(W,U) is established between router W and U. When the link W-D goes down, W immediately redirects the traffic with T(W,U) to router U. Depending on the installed handover route, U diverts the traffic to router C which succeeds forwarding it along its primary path (C,D) to the destination.



(a)Install protected tunnel (b) Forward with protected tunnel

Figure 2 Procted Tunnel in Action: A procteced tunnel is established between U and W by SR-BGP and is used to react quickly to link failure.

The conceptual simplicity of the protected tunnel idea hides three significant challenges.

- Which router has capability to be a tunnel egress?
- To whom should tunnel route be advertised?
- Which neighbor could be deemed as the next hop of the handover route?
- What contents are contained in the tunnel route?

In the following three sections, we address these challenges and corresponding solutions are presented.

3.2. Identifying tunnel egress router

In order to prevent forwarding loop, tunnel egress router must have capability to divert tunnel packets to a peer different from its primary and reverse neighbor. That is to say, if a router has capability to be tunnel egress, it must have at least a backup neighbor. However, how does a router identify whether it has backup neighbors? Actually, once a router obtains a route from a peer different from its primary, the peer must be a backup neighbor. Thus, we have such rule:

Rule 1: A router has capability to be tunnel egress, once it has at least one backup path.

If a router finds it has capability to be a tunnel egress, it should select a path from its backup ones and install a handover route. Then, the corresponding tunnel route information is created and propagated to proper neighbors. A router which receives the tunnel route could be tunnel ingress and establish protected tunnel with tunnel egress.

3.3. Tunnel route propagation

Following the BGP rules, routers advertise only the best path to all policy compliant neighbors and don’t tell any routes to its primary neighbor. If a router has ever advertised a route to a peer which now becomes primary neighbor, BGP’s poison-reverse policy ensures that a withdrawal be sent in this case. The rule induces some routers have only one path, e.g. router W and V in Fig. 1. When the only path of this type of router is invalid, it would incur packets loss. So, tunnel egress routers should announce the tunnel route to its primary neighbor so that once the primary path is invalid, protected tunnel could be used to forward traffic. Furthermore, a tunnel egress router has an incentive to advertise tunnel route to its primary neighbor because if it does not, the primary neighbor may be left without a path and drop all packets coming from its reverse neighbors. For example, in Fig. 2.a, router U would like to advertise tunnel route to the next hop routers V and W along its primary path. Otherwise, W would drop all the packets coming from router U because the unique path of router W has been invalid when the link between router W and D goes down. However, router U is less incented to offer tunnel route to router B, because that router U never delivers its packets to router W. Thus, we have such rule:

Rule 2: A router only advertises tunnel route per destination to the next hop router along its primary path.

Advertising tunnel route does not obviously add update message overhead because each router advertises at most one message to its neighbor, just like current BGP. To see this, recognize that the poison-reverse policy require routers to withdraw advertised route from its primary neighbor. SR-BGP replaces the withdrawal message with an advertisement of the tunnel route, keeping the overhead at a minimum.

3.4. Handover route selection

The above rules specify the router which has capability to be a tunnel egress and the neighbor to which a router advertises tunnel route. However, the question of how to select and install handover route on the tunnel egress is not answered. So, this section states the rule of handover route selection.

There may be several backup paths on tunnel egress. The selection among these paths determines not only the next hop of the handover route but also routers to which the tunnel route can be propagated. In Fig. 2.a, router U has three paths including a primary path and two backup paths. Suppose that router U selects the backup path (U,B,W,D) as its *handover path*, the tunnel information is only able to be advertised to router V. Router W would never receive this tunnel route due to the Send Side Loop Detection (SSLD) [4]. Thus, protected tunnel is only established between router U and V, and only one link, e.g. link V-W, can be protected. On the other hand, if the backup path (U,C,D) is selected as the handover path of router U, the tunnel route can be propagated to router V and W. At this time, there are two links, e.g. link V-W and W-D can be protected. So, we have the conclusion: the more disjoint path from the primary is selected, the more links can be protected. Thus, we have such rule:

Rule 3: Tunnel egress router installs handover route which next hop is the one from which the most disjoint path from primary is received.

Note that it may be more costly to require tunnel egress router to select the most disjoint path as the handover path and redirect traffic to the peer. However, protected tunnel is only used for a short period during transient link failures and it is used to guarantee connectivity to the tunnel egress router. Thus, we believe most routers are willing to propagate such tunnel route.

3.5. Packet forwarding

In R-BGP or BRAP, packet forwarding during link failures requires detecting whether a packet is on the failover path or the primary and storing the next hop for both primary and failover paths. Interface-specific forwarding technique and MPLS-based solution are used to achieve the objective. Those techniques are required on all routers of the failover path.

Comparing with those approaches, packet forwarding with SR-BGP is more simple and easy to be implemented. SR-BGP provides fast recovery during link failures with pre-established protected tunnel. Several types of tunnels exist: IP over IP, GRE, IPSec, L2TP, and MPLS over IP, etc. The typical IP over IP tunnel may be preferred. Among the routers included by protected tunnel, only the ingress and egress routers of protected tunnel are aware of the difference of packet forwarding. The other routers still adopt destination-based forwarding as usual. The operations on the ingress and egress routers are as follows:

- Fast link failure detection and packet encapsulation on ingress router: Tunnel ingress router detects the link failure by using a trigger from the physical layer such as a SONET loss of signal [19] or a protocol such as BFD [10]. Once link failure is detected, the tunnel ingress router would encapsulate IP packets with a new IP header which including the IP address of tunnel egress and corresponding forwarding directive (FD). The optional parameter of IP header can be used to carry FD information.
- Packet decapsulation on egress router: When the tunnel egress router receives a packet whose destination IP address is exactly the router's IP address, it would search the pre-defined optimal parameter to ensure whether it is a tunnel packet. If it is, the router would decapsulate the packet and forward it based on the handover route indexed with FD.

In summary, we can conclude the information contained by tunnel route:

- The NLRI is the local IP address of tunnel egress router.
- The AS-PATH attribute contains the AS list through which the route passing. It is used to aid the execution of SSLD.
- The Forwarding Directive attribute which is a new attribute is optional transitive. It is used to index the handover route.

4. Evaluation

In this section, we use simulation to measure the performance of SR-BGP in term of percentage of ASes seeing loss during link failures and the routing table size. We implement our simulation on the real AS graph of the Internet generated from BGP RIB at Routeviews vantage points [20]. Additionally, we use the best know policy inference algorithms [21] to annotate inter-AS links, e.g. customer-provider, provider-customer, and peer-peer. The sibling relationship is treated as peer-peer, since treating them as anything else leads to provider-customer loops. The data used in the simulation is obtained at Jan 1, 2007. At that time, the number of active ASes is 23,971 and links is 49,275.

4.1. Dual-homed Domains Resilience to Link Failures

At present, multi-homing which is treated as a more popular access approach is deployed in many domains to improve the reliability of networks. However, does the increment of inter-domain links improve the reliability during links failures? Fig. 3 reports 23.5% of ASes experience transient disconnectivity when one of the links to a dual-homed destination domain fails for all 9,547 dual-homed ASes.

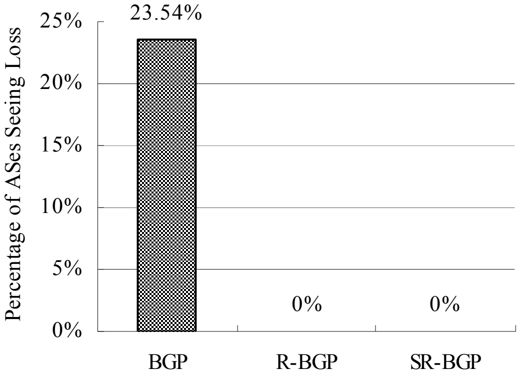


Figure 3 Percentage of AS that experience transient disconnectivity to a dual-homed edge domain when one of its links goes down.

If SR-BGP is deployed in the Internet, we can see that when the one link of dual-homed ASes goes down, the percentage of dual-homed ASes which experience the transient disconnectivity is zero. The effect of SR-BGP is comparable with R-BGP.

4.2. Route Table Size

The routing scaling problem is an important challenge that the routing protocol faces. So, the approaches of improving the reliability of routing should not dramatically inflate the routing table size. Simulation results manifest that some approaches, e.g. R-BGP or BRAP, double the routing table size of current. However, the route table size of SR-BGP is comparable with current. Fig. 4 manifests that the routing table size increase 1.2% of current at most.

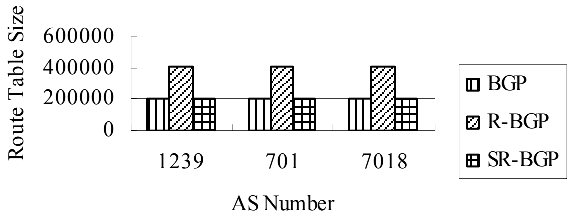


Figure 4 The route table size of AS corresponding three different routing protocol.

5. Conclusion

This paper presents a solution to provide fast recovery during inter-domain link failures. The key of our approach is to redirect traffic to a pre-established protected tunnel that is not influenced by the link failure. By using protected tunnel and corresponding handover route, our approach can maximize the number of available paths and minimize the size of routing table so that it can provide more resilience and more scalability.

Acknowledgment

The authors would like to thank Dongnian Cheng, Na Wang for valuable comments and discussion on earlier versions of this document. This work is supported by Major State Basic Research Development Program (No. 2007CB307102, No. 2007CB307100).

References

[1] N. Kushman, S. Kandula, D. Katabi, B. M. Maggs. "R-BGP: Staying connected in a connected world". In Proc. NSDI, 2007, pp.341-354.

[2] C. Labovitz, A. Ahuja, F. Jahanian. "Experimental study of Internet stability and backbone failures". In Proc. FTCS 1999, pp.278 - 285

[3] C. Labovitz, G. R. Malan, F. Jahanian. "Internet routing instability". IEEE/ACM Transactions on Networking, vol 6, issue 5, 1998, pp.515–528.

[4] F. Wang, L. Gao. "A backup route aware routing protocol – Fast recovery from transient routing failures". In Proc. INFOCOM, 2008.

[5] O. Bonaventure, C. Filsfils, and P. Francois. "Achieving sub-50ms recovery upon BGP peering link failures". In Proc. Co-Next, 2005.

[6] A. Bremler-Barr, Y. Afek, and S. Schwarz. "Improved BGP convergence via ghost flushing". In Proc. INFOCOM, vol.2 2003, pp.927-937.

[7] D. Pei et al. "Improving BGP convergence through consistency assertions". In Proc. INFOCOM, vol. 2, 2002, pp.902-911.

[8] D. Pei et al. "BGP-RCN: Improving BGP convergence through root cause notification". Computer Networks Journal, Vol. 48, Issue 2, 2005, pp. 175-194.

[9] F. Wang, Z. M. Mao, J. W. L. Gao, and R. Bush. "A measurement study on the impact of routing events on end-to-end Internet path performance". In Proc. SIGCOMM, 2006, pp.375-387.

[10] D. Katz, D. Ward. "Bidirectional forwarding detection". Internet Draft, draft-ietf-bfd-base-03.txt, 2005.

[11] A. Markopoulou, G. Iannaccone, S. Bhattacharyya, C. Chuah, C. Diot. "Characterization of failures in an IP backbone". In Proc. INFOCOM, vol. 4, 2004, pp.2307-2317.

[12] Y. Rekhter, T. Li. "A border gateway protocol 4 (BGP-4)". RFC 1771, 1995.

[13] N. Kushman, S. Kandula, D. Katabi. Can you hear me now?! It must be BGP. In ACM CCR, Vol. 37, No. 2, 2007, pp.75-84.

[14] G. Iannaccone, C. Chuah, R. Mortier, S.Bhattacharyya, C. Diot. "Analysis of link failures in an IP backbone". In Proc. ACM SIGCOMM IMW, 2002, pp.237-242.

[15] S. Lee, Y. Yu, S. Nelakuditi, Z. Zhang, C. Chuah. "Proactive vs reactive approaches to failure resilient routing". In Proc. INFOCOM, 2004, pp.176-186.

[16] J. Lou, J. Xie, R. Hao, X. Li. "An approach to accelerate convergence for path vector protocol". In Proc. Globecom, vol. 3, 2002, pp. 2390-2394.

[17] N. Feamster, D. Andersen, H. Balakrishnan, M. Kaashoek. "Measuring the effects of Internet path faults on reactive routing". In ACM SIGMETRICS, 2003, pp.126-137.

[18] A. Feldmann, O. Maennel, M. Mao, A. Berger, B. Maggs. "Locating Internet routing instabilities". In ACM SIGCOMM CCR, vol. 34, Issue 4, 2004, pp.205-218.

[19] J. –P. Vasseur, M. Pickavet, P. Demeester. "Nework recovery: Protection and restoration of optical, SONET-SDH, IP and MPLS", Morgan Kaufmann publisher, 2004.

[20] University of Oregon Route Views. www.routeviews.org.

[21] X. Dimitropoulos, D. Krioukov, M. Fomendov, B. Huffaker, Y. Hyun, kc claffy. "AS relationships: Inference and validation". ACM SIGCOMM CCR, vol. 37, issue 1, 2007, pp.29-40.