Make Best-Effort Forwarding upon Network Abnormality

Meijia Hou (Ph.D. Student)*†, Mingwei Xu*, Dan Wang†, and Qi Li (Ph.D. Student)*
*TNList, Department of Computer Science and Technology, Tsinghua University
†Department of Computing, The Hong Kong Polytechnic University

I. Introduction

In recent years, there are considerable research efforts focusing on abnormality recovery, such as Ghost Flushing, RCN, HLP, Consensus Routing, to name but a few. All these studies usually request for non-trivial changes of the Internet protocols or frameworks. This is a key reason that they have yet led to real deployment, especially for inter-domain. In this proposal, we focus on inter-domain abnormality recovery. While we believe a perfect or even satisfiable solution may require some fundamental upgrade, and a joint force of different schemes, we would like to think whether the current Internet has room to improve by simple yet incremental solutions. From the very first networking course, we learned that the routing of the Internet is best effort. Indeed, the key function of any routing system, by its definition, is to route packets. Nevertheless, as compared to many routing protocols developed in different contexts (wireless, DTN, non-CS), where the routing systems try to deliver packets even when the routing tables are incomplete, the topologies are continuously and dynamically changing, we find that the Internet is the least best effort. Packets are often dropped neither because the network is congested nor because available paths do not exist.

Such Internet routing strategy makes the design simple, the routing fast; and even a packet is dropped, at last, we know we have TCP. Nevertheless, facing the current and future demands on Internet speed/robustness, relying on TCP retransmission is certainly not possible. Here, we rethink this strategy in the BGP convergence period after an abnormality happens. In such period, the network states are inconsistent and the BGP routers may not find outgoing interface for the packets or the packets are trapped in loops. Our key idea is that rather than dropping such packets, the ASes make a best-effort decision to forward the packets, sometimes even by random.

To verify that our scheme can be integrated into the Internet, we carried out a preliminary experiment on real Internet (based on CNGI-6IX, AS 23*** [1]). We forward those packets that deem to be dropped by a simple version of our scheme. We see that a substantial number of packets were delivered (note that this is a pure gain). Also note that such gain is obtained by deploying our scheme on only one AS. This also indicates our scheme is beneficial even when it is incrementally deployed.

We study two simple algorithms, the Simple Random Forwarding (SiRF) and the Stable Random Forwarding (StRF) for packet delivery during the transient abnormal events in BGP. They are hardware implementable and do not request process that is beyond the current switching capability; They request for little modification of the current Internet routing paradigm, i.e., incrementally deployable, and by some preliminary simulations, they delivered 40%-70% of packets that originally should be dropped. We believe our schemes can serve as an intermediate step before the more complex abnormality recovery schemes are fully agreed by the community.

II. A Preliminary Experiment

Intrinsically, we will forward the packets that are to be dropped (caused by abnormality and the current Internet routing paradigm does not know how to forward) to some (partially) randomly chosen neighbors. To verify that, even without upgrade, the neighbor ASes will deliver the packets, we conduct an experiment on CNGI-6IX (AS 23***) [1]. CNGI-6IX is the domestic/international exchange point of the China Next Generation Internet (CNGI) project and connects the whole nationwide next generation Internet backbones and other international research networks.

We attach a source to CNSI-6IX and select 20 reachable destinations attached to remote ASes. Instead of using the stored next hops of CNSI-6IX, we choose two random BGP neighbors AS_1 (AS 7***) and AS_2 (AS 24***) as next hops. We observe that 62.5% and 41.7% of the packets can be correctly delivered if AS_1 and AS_2 are chosen respectively. Note that the results are obtained by a simple neighbor selection and only one AS (CNGI-6IX) is upgraded.

Due to space limitation, the details of this preliminary experiment can be found in [2].

III. Our Approach

In this section, we describe two forwarding algorithms. We first describe the occasions that will trigger our algorithms.

A. Abnormality Identification

We consider two common yet important abnormalities, the blackholes and the loops. These are caused by inconsistent states during the BGP convergence after network failure.

We infer network inconsistence from the data forwarding in the data plane. Let A and B be two neighboring ASes. Assume A receives a packet p from B with destination $p.dest$. A can infer, from the data plane, that its routing state is inconsistent with that of B if p has one of the following characteristics: 1)

Part of this work was done while Meijia Hou was a research assistant in Department of Computing, The Hong Kong Polytechnic University. Contact of the authors: {houmeijia, xmw, liqi}@csnet1.cs.tsinghua.edu.cn, csdwang@comp.polyu.edu.hk.

[1]Due to commercial reasons, in this paper, we cannot list the AS numbers in full.

A has no route to $p.dest$; and 2) from the last control plane update, B told A that B had chosen A as the next hop of B. Yet A is not the next hop of B in the AS path towards $p.dest$. Here 1) reflects a blackhole, and 2) reflects that p is possibly trapped in a loop. We will use a *mark bit* in the packet header upon such abnormalities: a mark bit 0 indicates a *normal packet* and a mark bit 1 indicates an *abnormal packet*.

B. Best-Effort Forwarding Scheme

We propose two forwarding schemes, Simple Random Forwarding (SiRF) and Stable Random Forwarding (StRF). The two random forwarding schemes are designed to be compatible with each other and also with existing BGP.

1) Simple Random Forwarding Scheme: The intuition of SiRF is to deliver the packets with a mark bit 1 to a random neighbor. We augment the packet header with an additional field, abnormal-hop, indicating the number of times a packet has been delivered inconsistently with corresponding BGP routing table in the AS level topology. If the abnormal-hop is larger than a threshold, the packet will be delivered according to the normal routing table (dropped if no outgoing interface). This threshold balances the network reachability and traffic load. In our simulation, we often found that a 2 or 3 random hop is enough to achieve a high delivery ratio.

2) Stable Random Forwarding Scheme: We have seen in various studies that the Internet failure is far from uniform [4]. Thus, compared to SiRF, we design StRF which sends the abnormal packets on a more stable route.

We add an item to each route to record its arriving time in Adj-RIBs-IN of the BGP border router. The stableness of the routes for a destination can be easily computed each time there are update messages for this destination. We choose to forward the packet to the route that has been stable for longest time. In StRF, if there is no route, the packets will be forwarded to a random selected neighbor (the same as SiRF).

C. Implementation Details

1) Control Plane and Data Plane Separation: Our schemes are mostly carried out on data plane. For SiRF, no change of control plane is necessary. The random forwarding can either be carried out upon receiving an abnormal packet or the RIB can be pre-set randomly. For StRF, in control plane, each time a BGP router receives an update message for a destination, the stablest route to this destination is recomputed. In data plane, the BGP router monitors packets and delivers the abnormal packets to the next hop of the corresponding stablest route, which has already been inserted into the RIB (or a randomly selected next hop if no route to the destination). The control plan and data plane are still totally separated. All these operations can be easily configured in hardware.

2) Cooperation with Intra-domain Routers: In an upgraded AS, a BGP border router adds the next hop AS (computed by SiRF or StRF) as a *transient destination* in the abnormal packet header. The intra-domain routers then deliver the packets to the next hop AS according to the transient destination and their own forwarding tables. With such intra-domain upgrade, SiRF and StRF can work well on the AS-level.

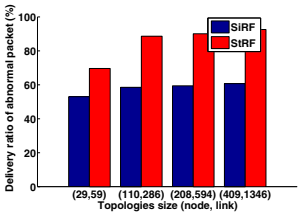

Fig. 1: Delivery ratio as a function of the size of the topology.

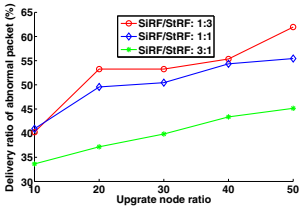

Fig. 2: Delivery ratio as a function of the upgrade ratio.

Clearly, our schemes only need hardware configuration upgrade. They are also incrementally deployable.

IV. Preliminary Simulation

We develop an event-driven simulator to evaluate SiRF and StRF with four realistic multi-AS topologies from SSFNET; all are based on real Internet BGP routing table. As BGP is a policy-based routing protocol, we infer the standard customer-provider, peer-peer and sibling-sibling AS relationships according to the node degree and the ratio of different type of relationships in CAIDA topologies [3]. We design link failures according to negative exponential distribution.

The major evaluation metric is the delivery ratio of the abnormal packets which may be dropped in standard BGP protocol. Fig. 1 shows the performance of SiRF and StRF under different topologies. The performance is surprisingly good. Even for SiRF, more than 40% of the packets that should be dropped are successfully delivered; and StRF is even better. We also evaluate when the network is partially upgraded. Note that when an AS, that has not been configured with our schemes, receives a packet during abnormality, the packet will simply be dropped. Fig. 2 shows the results. We can see that even if there are 10% of the ASes are upgraded, we still see 32% to 40% of packet saved, depending on the ratio of whether they are upgraded with SiRF or StRF. Looking into the details, we find that the packets can be forwarded out of the troubled area very quickly, i.e., in a few AS hops. Clearly, the more nodes are upgraded, the better the performance.

V. Discussion

Our design goal is sharply different from those protection schemes targeting on saving all the packets during failures. We make our design simple and easy to get deployed. It can serve as an intermediate plan before the more complex scheme can be fully agreed. Our scheme may violate the current ISP relationship of the packet delivery. Nevertheless, our schemes will only be initiated upon network failures; thus requests very moderate change. From the bottom line, the packets we try to save are those deem to be dropped; we just ask the Internet routing system to make more best-effort for packet delivery.

References

[1] CNGI-6IX, http://www.cernet2.edu.cn/en/6ix.htm.

[2] http://www.comp.polyu.edu.hk/~csdwang/ForwardingDetails.pdf.

[3] X. Dmitiropoulos, D. Krioukov, M. Fomenkov, B. Huffaker, Y. Hyun, K. Claffy, and G. Riley. "AS relationships: Inference and validation", *ACM SIGCOMM CCR*, 37(1), pp. 29-40, Jan. 2007.

[4] M. Hou, D. Wang, M. Xu, and J. Yang. "Selective protection: a cost-efficient backup scheme for link state routing", In *Proc. IEEE ICDCS*, Montreal, Canada, June. 2009.