ES-TMP: Inter-domain Egress Selection based on Traffic Migration Prediction

Dan Zhao, Xiaofeng Hu, Chunqing Wu
School of computer
National University of Defense Technology
Changsha, China
danzhao.nudt@gmail.com

***Abstract*—Hot-potato routing is commonly used to break tie among multiple equally-good exit points associating with inter-domain BGP routes. However, hot-potato routing only takes the network control plane into consideration, where it provides the routers the possibility of enabling early exit of traffic using barely protocol-related information of IGP distance. In this paper, we argue that egress selection of inter-domain routing should pay more attention to traffic forwarding, because the large traffic migration caused by egress change, although not quite often, can degrade the network performance or even make the network crash. We propose Egress Selection based on Traffic Migration Prediction (ES-TMP). We use traffic demand to predict the traffic migration, which is used as important criteria for egress selection. If the volume of traffic migration is large, ES-TMP keeps the egress unchanged. Otherwise, the small traffic migration enables the routers use the closest egress without apparent influence on network performance. ES-TMP can either be implemented with standard BGP protocol or by dedicated servers to perform global routing optimization.**

Keywords-BGP; egress selection; hot-potato; traffic engineering

I. INTRODUCTION

The Internet routing is composed of two major blocks, i.e. intra- and inter-domain routing which emerge as a two-tiered architecture. Inside the Autonomous System (AS), intra-domain routing follows the path that has the shortest Interior Gateway Protocols (IGP) distance, which is the sum of each link weight (or cost) along that path. Inter-domain routing protocol allows border routers to exchange reachable information to external destinations. The ultimate forwarding path is concatenated of inter-domain links between ASes and the path segments from ingress to appropriate egress towards the destination prefix inside each intermediate AS.

Internet service providers (ISPs) usually manage the network to peer with each other at multiple locations for fault tolerance and load balancing. This translates into several egress points available for a single destination from inner-AS router's perspective. Since selecting egress among multiple available exit points is a fundamental part of the Internet routing architecture [3], it deserves better design and implementation considering routing and forwarding performance.

It is well known that BGP [5] is the de-facto inter-domain routing protocol which is stateful and featured with policy-based path vector behavior. BGP routers at the periphery of AS learn how to reach a destination prefix by eBGP session and notify interior BGP routers using iBGP session. The route is also diffused by IGP so that each router can select appropriate egress point to reach the destination. When there are multiple "equally-good" paths in terms of BGP attributes to a prefix, each router in the AS normally selects the closest egress in context of minimal IGP distance. This policy of hot-potato routing directly expresses that AS should deliver the traffic out of the network as early as possible in order to minimize resource consumption.

However, hot-potato routing cannot guarantee minimum resource usage. As the link weight is partially a configurable parameter, it cannot directly interpret the link capability. Besides, network operators often reconfigure link weights for purposes like traffic engineering [6], [7], disruption avoidance [8] and so on. Thus the IGP distance may not have necessary relationship with hop count, propagation delay or link capacity, and selecting the closest egress point does not definitely improve performance [3]. Moreover, hot-potato routing is argued to be too sensitive to the changes of topology. Minor IGP changes can cause hot-potato to switch the egress, resulting in BGP updates and convergence which may take several minutes. Also, there is possibility of large traffic migration [1], [2]. The abrupt traffic increase can potentially cause network congestion of current and neighboring AS. For most of the network failures are transient [13], the instant traffic adaptation to IGP changes could make the network more unstable.

Previous work attempts to make the hot-potato routing more robust to network changes and flexible for performance objectives [3], [9]. The work emphasizes on routing stability and resilience in presence of topology changes and protocol dynamics, which is treated as the major goal of network optimization. However, as most of the parameters associating with hot-potato routing are human-configured variables, the extended hot-potato routing may not work on real network status. Besides, whether these proposals can deal with traffic migration still remains questionable. There are also researches that aim to handle the traffic dynamics induced problems in inter-domain routing [15], [16], [17]. But these proposals mainly concentrate on inter-domain traffic engineering goals such as load balance among multiple neighboring ASes, equal traffic splitting among multiple equal-cost paths or traffic demand regulation for reducing resource consumption. However, the traffic migration caused by egress switching hasn't been addressed in these works.

Published by Atlantis Press, Paris, France.

In this paper we argue that egress selection should consider more about traffic forwarding. The traffic migration caused by egress switching could be very large [1], many of which are related to by hot-potato routing changes [2]. The traffic bursts may have impact on downstream AS neighbors, which results in unpredictable inbound traffic [10]. Although this situation happens infrequently, we must cope with it because the network can seem to be running smoothly only to crash unexpectedly after some seemingly "small" event [4]. Thus, we propose Egress Selection based on Traffic Migration Prediction (ES-TMP), which can keep the traffic flow considerably stable when selecting the suitable exit point. Our proposal is a further step of the researches on egress selection optimization in the context of inter-domain traffic engineering, which has main features:

- The traffic migration prediction is based on demand of inbound traffic of the network with respect to destination prefixes. This captures the high-level behavior of network traffic flow thus it is meaningful for Internet service optimization.
- The characteristics of the predicted traffic flow changes are used to directly guide the egress selection. The major objective is to avoid egress switching in case of large traffic migration for routing and forwarding stability, whereas the small traffic fluctuation can make the network select the closest egress to minimize performance degradation.

The rest of the paper is organized as follows. In section II we generally outline the related work. Section III formulates the problem addressed in this paper, and describes the fundamental details of ES-TMP algorithm. We discussed several issues related to ES-TMP, including implementation considerations in section IV. At last, we give the conclusions in section V.

II. RELATED WORK

Our work relates to several aspects of internet routing as follows.

Hot-potato disruptions. Early measurement researches have shown that hot-potato routing change can lead to long convergence delay, large traffic migration and eBGP update [1], [2]. Subsequent work proposed metrics of network sensitivity to help minimizing hot-potato disruptions [4].

Traffic engineering. Various studies have investigated the traffic engineering approaches for multiple objectives. In intra-domain routing, tuning link weight configuration can achieve low link utility [6], [7]. PEFT [14] is proposed to split traffic over multiple paths with an exponential penalty on longer paths. COPE [15] is a class of traffic engineering algorithms that optimize for the expected scenarios while providing a worst-case guarantee for unexpected scenarios. REPLEX [16] solves the problem of dynamically distributing load in the presence of bursty and hard to predict changes in traffic demands. The work of [17] proposes to change the traffic demand by rehoming edge links for better traffic engineering.

Egress selection optimization. The problem of egress selection relates to adaptation to topology change while stabilizing the network routing. TIE [3] proposes a mixed metric of dynamic sensitivity to topology and fixed ranking of egress point. BGP-ESOM [9] is a two-phase, feedback control model for egress selection. It is implemented on a new framework named BGP-RCS which is quite different from current BGP decision logic.

TABLE I. NOTATIONS

Notation	Description
G	Undirected graph to represent topology
ΔG	Network change
$d(a,b)$	IGP distance between node *a* and *b*
$TD(i,p,t)$	Traffic demand of ingress *i* towards prefix *p* during a time interval *t*
ε	Mapping from tuple *(i,p,t)* to egress router *e*, i.e. $\varepsilon(i,p,t)=e$
$E(p)$	Egress set for prefix *p*
$TM(i,p,t)$	Element of Traffic matrix during a time interval t
$f(i,e,t)$	Traffic flow between tuple *(i,e)* towards prefix *p*

III. EGRESS SELECTION BASED ON TRAFFIC MIGRATION PREDICTION (ES-TMP)

A. Problem formulation

Consider a network that is stable in routing and forwarding during a time interval *t*, i.e. network topology and eBGP routes do not change, and the traffic demand is considerably constant with only small fluctuations. Given the notations in Table 1, the traffic matrix element can be defined as:

$$TM(i,e,t)=\sum_{p\in P,\varepsilon(i,p,t)=e} TD(i,p,t) \tag{1}$$

where P is the set of all destination prefixes.

When the network change $\delta\in\Delta G$ causes ingress i to select a new egress e' towards prefix p, i.e. $\varepsilon(i,p,t)=e'$, the traffic flow shifts correspondingly. The volume of shifted traffic between i and e can be represented as:

$$\Delta V(i,e,t)=\sum_{\substack{p\in P,\varepsilon(i,p,t)\neq e,\\ \varepsilon(i,p,t-1)=e}} TD(i,p,t) \tag{2}$$

Note that there may be many prefixes performing the egress switching, thus ΔV represents the total traffic migration of egress change. However, the egress is selected from the perspective of single prefix in BGP decision procedure. Fortunately, the traffic migration of individual prefix can be easily obtained by traffic demand, i.e. $\Delta V_{(i,p,t)}=TD(i,p,t)$. Hence, we only consider $\Delta V_{(i,p,t)}$ in the egress selection procedure.

Egress switching can cause large traffic migration, which may incur the congestion in current and neighboring AS. Moreover, this kind of traffic burst could trigger downstream AS to adjust its routing policy for traffic engineering reasons. Thus the entire network behaves more complicated than the situation where only routing is involved. To our point of view, the egress selection should keep the traffic flow considerably stable, especially for large traffic migration. In this paper we propose to use $\Delta V_{(i,p,t)}$ as the criteria to break tie among multiple equally-good egresses towards the same prefix. When $\Delta V_{(i,p,t)}$ is small, we use the closest exit point as the egress; otherwise, the former egress is still used by the prefix. This approach has notable advantages for single AS and entire Internet. Small traffic fluctuation can hardly bring about congestion and should be delivered out of the network as soon as possible. In contrast, AS should avoid large traffic migration to improve network performance. Moreover, the stability of traffic flow is meaningful to intra- and inter-domain traffic engineering.

B. Definition of predicted traffic migration

The effectiveness of ES-TMP mainly depends on the predicted traffic migration behavior. We define the predicted traffic migration as $\Delta V_{(i,p,t)}$ if ingress *i* changes egress for prefix *p* according to hot-potato routing. Although $\Delta V_{(i,p,t)}$ doesn't necessarily reflect the real-time traffic volume, it represents the total traffic of a period that the network must forward. Thus the egress selection based on $\Delta V_{(i,p,t)}$ is helpful to improve the network forwarding performance.

When using $\Delta V_{(i,p,t)}$ to select the egress, it is crucial to determine what extent of traffic change is "large". We adopt the method proposed in [2] to capture the properties of the traffic itself. The variation of traffic demand *TD* over time is approximated as a stationary process with mean value and variance $2\sigma^2$. The variance value can be measured by traffic changes of *TD* for a period of time, and is used to normalize the traffic variations, i.e. $\Delta \tilde{TM} = \Delta TM / \sqrt{2}\sigma$. Similarly, we can use this variance value to normalize the traffic migration, i.e. $\Delta \tilde{V}_{(i,p,t)} = \Delta V_{(i,p,t)} / \sqrt{2}\sigma$. Besides, the results of [2] also show that most of the normalized traffic changes $\Delta \tilde{TM}$ vary in the $[-4,4]$ range. Note that the variation of $\Delta \tilde{TM}$ is the cumulative effect on traffic matrix element of all prefixes associated with specific egress router. For individual prefix, the traffic migration is considered to be large if the normalized variation exceeds $4/|P_i|$, where $|P_i|$ denotes the number of prefixes that ingress *i* has learned from BGP. If all prefixes avoid the migration more than $4/|P_i|$, the overall variation of $\Delta \tilde{TM}$ can never go beyond the $[-4,4]$ range. Hence the threshold of large traffic migration can be defined as $4/|P_i|$, which is used as condition for selecting suitable exit point in the ES-TMP algorithm.

C. ES-TMP algorithm

In this section we present the main algorithm of ES-TMP. As described above, ES-TMP requires the predicted volume of traffic migration when the IGP distance to former egress becomes longer than other routers in egress set $E(p)$. ES-TMP generally consists of two phases to perform egress selection. The static phase is for measurement of traffic dynamics, including the traffic demand and its normal variation. Besides, the number of prefixes can be dumped from router's BGP table during this phase. These parameters are the base ground of ES-TMP, but also can be used as long-term statistical data for network management. The decision phase takes effect when the BGP decision process is invoked. When the former egress is considered to be sub-optimal in the context of hot-potato routing, ES-TMP maintains the exit to this former egress if large traffic migration is predicted to happen. Otherwise, the small volume of traffic is forwarded to the early exit point.

1. static phase
   measure the traffic demand $TD(i,p,t)$ at each ingress and the variance value $2\sigma^2$, monitor the number of prefix $|P_i|$
2. decision phase
   for $\delta \in \Delta G$, ingress i and former egress *e* towards *p*
   if $e \neq \arg\min_o \{d(i,o) \mid o \in E(p)\}$
   calculate the predicted traffic migration $\Delta \tilde{V}_{(i,p,t)} = TD(i,p,t) / \sqrt{2}\sigma$
   if $\Delta \tilde{V}_{(i,p,t)} \geq 4/|P_i|$
   $\varepsilon(i,p,t) = e$
   otherwise
   $\varepsilon(i,p,t) = e' = \arg\min_o \{d(i,o) \mid o \in E(p)\}$

Figure 1. ES-TMP algorithm

One may notice that the initial egress selection of ES-TMP still uses the border routers that have the shortest IGP distance. This is because we consider ES-TMP to be an optimization approach in egress selection process. Before invoking ES-TMP, the operator generally optimizes the network performance, especially applying traffic engineering of configuring the topology for prevailing traffic [6], [7]. This preliminary optimization, which is beyond the scope of this paper, can be carried out in the static phase whenever needed, using the network measurement data other than traffic demand. In such a well-provisioned network, ES-TMP further optimizes network routing and forwarding by stabilizing the traffic flows in context of egress selection, since it reduces the possibility of network reconfiguration towards adaptation to traffic changes which may lead to globally unpredictable behavior.

IV. DISCUSSION

In this paper we propose to use the predicted traffic migration to select appropriate egress. The traffic migration

represents the continuous traffic flow deviation between fixed ingress and changeable egress. Both border and intra-domain routers on the forwarding path of traffic flow can measure the traffic volume. According to the uniqueness of traffic flow, ES-TMP can guarantee forwarding consistency in a network with full-meshed iBGP configuration.

However, large networks usually employ route reflectors to overcome the scaling problems of iBGP sessions. If the route reflector applies ES-TMP according to its own traffic measurement, then its client routers are unlikely to avoid large traffic migration. We recommend adding a new mechanism for traffic information communication between route reflector and client routers, which can be achieved by extending the BGP session. To reduce the complexity, route reflector can accumulate the overall traffic demand of its client routers, and apply ES-TMP to choose common exit point for all clients using the cumulative traffic demand.

Rather than modifying the BGP decision process implemented on the routers, AS could employ dedicated servers to perform path selection on behalf of the intra-domain routers, as proposed in [11] and [12]. The servers build up a functionality gap between routing and forwarding that can minimize unexpected or unwanted interactions among multiple protocols. As these servers can gather the traffic measurement data of the entire network, we can easily implement ES-TMP on these servers. Thus only border routers need to monitor the traffic demand and normal variation, which help to reduce the network cost. Besides, since the servers have global and complete view of the network, they can verify the consistency of routing and forwarding produced by ES-TMP. Correspondingly, ES-TMP could extend the path selection based on routing and forwarding separation, which makes centralized servers for network routing more promising

V. CONCLUSION

In this paper, we propose ES-TMP which takes traffic dynamics into considering when selecting appropriate exit point inside AS. The objective of ES-TMP is to avoid large traffic migration which can potentially degrade network performance or even break the network down, and may incur complex network dynamics of routing and traffic forwarding. ES-TMP can be implemented with standard BGP protocol, either in a full-meshed iBGP network or with route reflector configuration. Moreover, this approach can be integrated to dedicated servers to perform global optimization of routing and forwarding. In the future work, we will implement ES-TMP in a network with routing and forwarding separation and evaluate the performance.

ACKNOWLEDGMENT

The work described in this paper is supported by the NSFC under Grant No.61103189 and No.61070199, Program for Changjiang Scholars and Innovative Research Team in University (No.IRT 1012), Program for Science and Technology Innovative Research Team in Higher Educational Institutions of Hunan Province: “network technology”, and Hunan Province Natural Science Foundation of China (11JJ7003).

REFERENCES

[1] R. Teixeira, A. Shaikh, T. Griffin, and J. Rexford, “Dynamics of hot-potato routing in IP networks,” in Proceedings of ACM SIGMETRICS, pp. 307–319, Jun. 2004.

[2] R. Teixeira, N. Duffield, J. Rexford and M. Roughan, “Traffic Matrix Reloaded: Impact of Routing Changes,” In Proceedings of Passive and Active Measurement, March/April 2005.

[3] R. Teixeira, T. Griffin, M. Resende, and J. Rexford, “TIE Breaking: Tunable Interdomain Egress Selection,” In Proceedings of ACM CoNext, August 2005.

[4] R. Teixeira, T. Griffin, A. Shaikh, and G. Voelker, “Network sensitivity to hot-potato disruptions,” In Proceedings of ACM SIGCOMM, September 2004.

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

[6] B. Fortz and M. Thorup, “Internet traffic engineering by optimizing OSPF weights,” In Proceedings of IEEE INFOCOM, pp.519-528, March 2000.

[7] B. Fortz and M. Thorup, “Optimizing OSPF/IS-IS Weights in a Changing World,” IEEE Journal on Selected Areas in Communications, Vol. 20, No. 4, pp. 756-767, May 2002.

[8] P. Francois, M. Shand and O. Bonaventure, “Disruption free topology reconfiguration in OSPF networks,” In Proceedings of IEEE INFOCOM, pp. 89-97, 2007.

[9] Yaping Liu, Junfeng He, Zhenghu Gong, “BGP-ESOM: BGP Egress Selection Optimization Model Based on Traffic Demand,” Future Generation Communication and Networking, Vol. 1, pp. 118-123, 2007.

[10] Nick Feamster, Jay Borkenhagen and Jennifer Rexford, “Guidelines for Interdomain Traffic Engineering,” ACM SIGCOMM Computer Communications Review, Vol. 33, Num. 5, pp. 19-30, 2003.

[11] N. Feamster, H. Balakrishnan, J. Rexford, A. Shaikh, and J. van der Merwe, “The case for separating routing from routers,” ACM SIGCOMM Workshop Future Directions Network Architecture, pp. 5–12, 2004.

[12] M. Caesar, D. Caldwell, N. Feamster, J. Rexford, et al., “Design and Implementation of a Routing Control Platform,” In Proceedings of USENIX 2nd Symposium on Networked Systems Design & Implementation, pp. 15-28, 2005.

[13] A. Markopoulou, G. Iannaccone, S. Bhattacharyya, C.-N. Chuah, and C. Diot, “Characterization of failures in an IP backbone network,” in Proceedings of INFOCOM 2004, vol. 4, Mar. 2004, pp. 2307–2317.

[14] Dahai Xu, Mung Chiang and Jennifer Rexford, “Link-State Routing With Hop-by-Hop Forwarding Can Achieve Optimal Traffic Engineering,” IEEE/ACM Transactions On Networking, Vol. 19, No. 6, 2011.

[15] Hao Wang, Haiyong Xie, Lili Qiu, Yang Richard Yang, et al., “COPE: Traffic Engineering in Dynamic Networks,” In Proceedings of ACM SIGCOMM, 2006

[16] Simon Fischer, Nils Kammenhuber and Anja Feldmann. REPLEX-Dynamic Traffic Engineering Based on Wardrop Routing Policies. In Proceedings of ACM CoNext, 2006

[17] Eric Keller, Michael Schapira and Jennifer Rexford, “Rehoming Edge Links for Better Traffic Engineering,” ACM Computer Communication Review, Vol. 42, pp. 65-71, 2012