TIE Breaking: Tunable Interdomain Egress Selection

Renata Teixeira
UC San Diego
La Jolla, CA
teixeira@cs.ucsd.edu

Timothy G. Griffin
University of Cambridge
Cambridge, UK
Timothy.Griffin@cl.cam.ac.uk

Mauricio G. C. Resende
AT&T Labs–Research
Florham Park, NJ
mgcr@research.att.com

Jennifer Rexford
Princeton University
Princeton, NJ
jrex@cs.princeton.edu

1. INTRODUCTION

The Internet's two-tiered routing architecture was designed to have a clean separation between the intradomain and interdomain routing protocols. However, the appropriate roles of the two protocols becomes unclear when an Autonomous System (AS) has interdomain routes to a destination prefix through multiple border routers—a situation that is extremely common today because neighboring domains often connect in several locations. Selecting among multiple egress points is now a fundamental part of the Internet routing architecture, independent of the current set of routing protocols.

In the Internet today, border routers learn routes to destination prefixes via the Border Gateway Protocol (BGP). When multiple border routers have routes that are "equally good" in the BGP sense (e.g., local preference, AS path length, etc.), each router in the AS directs traffic to the *closest* border router, in terms of Interior Gateway Protocol (IGP) distances (as shown in the example in figure 1). This policy of *early-exit* or *hot-potato* routing is hard-coded in the BGP decision process implemented on each router.

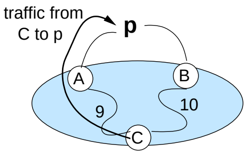

Figure 1: Router C has egress points A and B to reach destination prefix p, and selects the closest egress (i.e., egress A with distance 9).

We believe that the decision to select egress points based on IGP distances should be revisited, in light of the growing pressure to provide good, predictable communication performance for applications such as voice-over-IP, online gaming, and business transactions. We argue that hot-potato routing is:

- **Too restrictive:** The underlying mechanism dictates a particular policy rather than supporting the diverse performance objectives important to network administrators.
- **Too disruptive:** Small changes in IGP distances can sometimes lead to large shifts in traffic, long convergence delays, and BGP updates to neighboring domains [1, 2].
- **Too convoluted:** Network administrators are forced to select IGP metrics that make "BGP sense," rather than viewing the two parts of the routing system separately.

Selecting the egress point and computing the forwarding path to the egress point are two very distinct functions, and we believe that they should be decoupled. Paths inside the network should be selected based on some meaningful performance objective, whereas the egress selection should be flexible to support a broader set of traffic-engineering goals.

In this work, we propose a new way for each router to select an egress point for a destination, by comparing the candidate egress points based on a weighted sum of the IGP distance and a constant term. The configurable weights provide flexibility in deciding whether (and how much) to base BGP decisions on the IGP metrics. Network management systems can apply optimization techniques to automatically set these weights to satisfy network-level objectives, such as balancing load and minimizing propagation delays. To ensure consistent forwarding through the network, we advocate the use of lightweight tunnels to direct traffic from the ingress router to the chosen egress point. Our new mechanism, called TIE (Tunable Interdomain Egress) because it controls how routers break ties between multiple equally-good BGP routes, is both simple (for the routers) and expressive (for the network administrators). Our solution does not introduce any new protocols or any changes to today's routing protocols, making it possible to deploy our ideas at one AS at a time and with only minimal changes to the BGP decision logic on IP routers. More details on TIE can be found in [3].

2. TIE: TUNABLE INTERDOMAIN EGRESS SELECTION

Ideally, an optimization routine could compute the egress points directly based on the current topology, egress sets, and traffic, subject to a network-wide performance objective. However, the routers must adapt in real time to events such as changes in the underlying topology and egress sets, leading us to design a simple mechanism that allows a separation of timescales—enabling both rapid adaptation to unforeseen events and longer-term optimization of network-wide objectives. In addition, the design of our mechanism places an emphasis on generality to allow us to support a wide variety of network objectives, rather than tailoring our solution to one particular scenario.

Our mechanism allows each router to have a ranking of the egress points for each destination prefix. That is, router i has a metric $m(i, p, e)$, across all prefixes p and egress points e. For each prefix, the router considers the set of possible egress points and selects the one with the smallest rank, and then forwards packets over the shortest path through the network to that egress point. TIE relies on lightweight tunneling to guarantee that traffic is forwarded to the selected egress point.

To support flexible policy while adapting automatically to network changes, the metric $m(i, p, e)$ must include both configurable parameters and values computed directly from a real-time view of the topology. We represent the IGP distance between an ingress

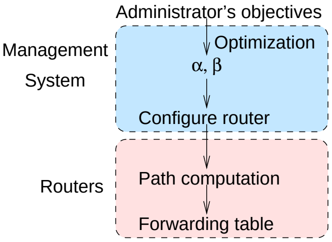

Figure 2: A management system optimizes α and β for a high-level policy and configure routers. Routing adapts the egress selection at real time in reaction to network events.

router i and an egress e as $d(i,e)$. The metric is computed as a weighted sum of the IGP distance and a constant term:

$$m(i,p,e) = \alpha(i,p,e) \cdot d(i,e) + \beta(i,p,e),$$

where α and β are configurable values. The first component of the equation supports automatic adaptation to topology changes, whereas the second represents a static ranking of routes for that prefix. Together, these two parameters can balance the trade-off between adaptation and robustness. This simple metric satisfies our three main goals:

- **Flexible policies:** By tuning the values of α and β, network administrators can cover the entire spectrum of egress-selection policies from hot-potato routing to static rankings of egress points.

- **Simple computation:** The metric is computationally simple—one multiplication and one addition—based on information readily available to the routers. This allows routers to compute the appropriate egress point for all destination prefixes immediately after a change in the network topology or egress set.

- **Ease of optimization:** The mechanism offers two knobs (α and β) that can be easily optimized by a management system based on diverse network objectives. We explore the expressive power of TIE next.

3. USING TIE

We do not envision that network administrators will configure α and β by hand. Instead, they define a network-wide goal, and an automated management system selects the appropriate values to achieve this high-level goal and configures the routers as presented in Figure 2. In this section, we give two examples of high-level goals that can be easily achieved using TIE.

Minimizing Sensitivity to Failures with Bounded Delay

Suppose a network administrator wants to minimize the sensitivity of egress-point selection to equipment failures, subject to restrictions on increasing the propagation delay. By setting the IGP link weights according to geographic distance, the shortest IGP path between two nodes would correspond to the smallest delay. The policy of minimizing sensitivity can be implemented as follows. At design time, select the closest egress point (say b). If an internal failure occurs, the administrators want node i to continue directing traffic to b unless the delay to b exceeds $T \cdot d(i,b)$ for some threshold $T > 1$. If the delay to reach the egress point exceeds the threshold, then node i should switch to using the (new) closest egress point to minimize delay.

We developed a prototype management system to solve this problem using TIE. We need to find values of $\alpha(i,p,e)$ and $\beta(i,p,e)$, for each i, e and p, that lead to the desired egress-point selections over a set of failures. Our solution has two main steps. First, a *simulation phase* determines the desired egress selection both at design time and after each failure. The output of this phase is a set of constraints on the α and β values for each (i,p) pair. Then, an *optimization phase* determines the values of α and β that satisfy these constraints by applying integer programming. We evaluate the resulting solution using topology and routing data from two backbone networks (Abilene, the US research network, and a tier-1 ISP network). TIE was able to achieved the desired behavior of avoiding unnecessary egress changes. For example, for the Abilene network TIE had 15% less egress changes than hot-potato routing without exceeding the specified delay ratio threshold $T = 2$.

Traffic Engineering

Traffic engineering—adapting the flow of traffic to the prevailing network conditions—is a common task. We propose an optimization problem that balances link utilization on the network only by selecting the appropriate egress point for each pair (i,p) (i.e., by setting the values of $\beta(i,p,e)$). This is in contrast with the common practice of optimizing link utilization by either tweaking IGP link weights or BGP policies. We formulate the egress-selection problem as a *path-based* multicommodity-flow problem that accounts for the constraints that the intradomain routing imposes on the flow of traffic. We evaluate our solution by comparing the link utilizations achieved using TIE to that using the current network configuration, and found that TIE had lower overall link utilization for both networks.

4. CONTRIBUTIONS

This work makes the following research contributions:

- **Flexible mechanism for egress-point selection:** TIE is simple enough for routers to adapt in real time to network events, and yet is much more amenable to optimization than today's routing protocols.

- **Optimization of network-wide objectives:** We present two example problems that can be solved easily using TIE. First, we show how to minimize sensitivity to internal topology changes, subject to a bound on propagation delay, using integer programming to tune the weights in our mechanism. Second, we show how to balance load in the network by using multicommodity-flow techniques to move some traffic to different egress points.

- **Evaluation on two backbone networks:** Our experiments for two network-management problems demonstrate the effectiveness of our new mechanism and the ease of applying conventional optimization techniques to determine the best settings for the tunable parameters.

5. REFERENCES

[1] R. Teixeira, A. Shaikh, T. Griffin, and J. Rexford, "Dynamics of Hot-Potato Routing in IP Networks," in *Proc. ACM SIGMETRICS*, June 2004.

[2] R. Teixeira, N. Duffield, J. Rexford, and M. Roughan, "Traffic matrix reloaded: Impact of routing changes," in *Proc. Passive and Active Measurement Workshop*, March/April 2005.

[3] R. Teixeira, T. Griffin, M. Resende, and J. Rexford, "TIE Breaking: Tunable Interdomain Egress Selection," Tech. Rep. TD-69EJBE, AT&T Labs – Research, February 2005.