Impact of Hot-Potato Routing Changes in IP Networks

Renata Teixeira
CNRS and Univ. Pierre et Marie Curie
Paris, France
renata.teixeira@lip6.fr

Aman Shaikh
AT&T Labs–Research
Florham Park, NJ
ashaikh@research.att.com

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

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

***Abstract*— Despite the architectural separation between intradomain and interdomain routing in the Internet, intradomain protocols do influence the path-selection process in the Border Gateway Protocol (BGP). When choosing between multiple equally-good BGP routes, a router selects the one with the *closest* egress point, based on the intradomain path cost. Under such *hot-potato* routing, an intradomain event can trigger BGP routing changes. To characterize the influence of hot-potato routing, we propose a technique for associating BGP routing changes with events visible in the intradomain protocol, and apply our algorithm to a tier-1 ISP backbone network. We show that (i) BGP updates can lag 60 seconds or more behind the intradomain event, (ii) the number of BGP path changes triggered by hot-potato routing has a nearly uniform distribution across destination prefixes, and (iii) the fraction of BGP messages triggered by intradomain changes varies significantly across time and router locations. We show that hot-potato routing changes lead to longer delays in forwarding-plane convergence, shifts in the flow of traffic to neighboring domains, extra externally-visible BGP update messages, and inaccuracies in Internet performance measurements.**

I. Introduction

End-to-end Internet performance depends on the stability and efficiency of the underlying routing protocols. A large portion of Internet traffic traverses multiple Autonomous Systems (ASes), making performance dependent on the routing behavior in multiple domains. In the large ASes in the core of the Internet, routers forward packets based on information from both the *intra*domain and *inter*domain routing protocols. These networks use the Border Gateway Protocol (BGP) [1] to exchange route advertisements with neighboring domains and propagate reachability information within the AS. The routers inside the AS use an Interior Gateway Protocol (IGP) to learn how to reach each other. In large IP networks, the two most common IGPs are OSPF [2] and IS-IS [3], which compute shortest paths based on configurable link weights. A router combines the BGP and IGP information to construct a forwarding table that maps destination prefixes to outgoing links.

The two-tiered routing architecture should isolate the global Internet from routing changes within an individual AS. However, in practice, the interaction between intradomain and interdomain routing is more complicated than this. The example in Figure 1 shows an AS with two external BGP (eBGP) sessions with a neighboring AS that advertises routes to a destination prefix. The two routers B and C propagate their eBGP-learned routes via internal BGP (iBGP) sessions with router A. This leaves A with the dilemma of choosing between two BGP routes that look "equally good" (e.g., with the same number of AS hops). The BGP decision logic dictates that A directs traffic to the closest *egress point*—the router with the smallest intradomain distance (i.e., router C). This decision is commonly called *early-exit* or *hot-potato* routing. Hot-potato routing tends to limit the bandwidth resources consumed by the traffic by moving packets to a next-hop AS at the nearest opportunity. However, suppose the IGP distance to C changes from 9 to 11, in response to a link failure along the original path or an intentional change in a link weight for traffic engineering or planned maintenance. Although the BGP route through C is still available, the IGP distance change would cause A to select the route through egress point B. We refer to this as a *hot-potato routing change*.

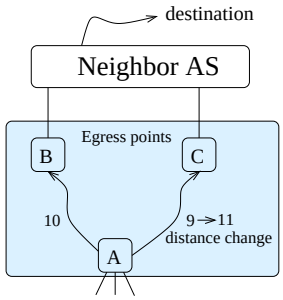

Fig. 1. Hot-potato routing change from egress C to B

Hot-potato routing changes can have a significant performance impact: (i) transient packet delay and loss while the routers recompute their forwarding tables, (ii) BGP routing changes visible to neighboring domains, and (iii) shifts in traffic that may cause congestion on the new paths through the network. Traffic shifts impact both ISPs and end-users. End-user applications will need to adapt to new end-to-end path characteristics (new RTT and available bandwidth). ISPs need to adapt to a (possibly large) variation in the *traffic matrix*, or where traffic enters and leaves the network.

The frequency and importance of these effects depend on a variety of factors. A tier-1 ISP network connects to many neighboring domains in many geographic locations. In practice, an ISP typically learns "equally good" BGP routes at each peering point with a neighboring AS, which increases the likelihood that routing decisions depend on the IGP distance to the egress points. In addition, the routers have BGP routes for more than 150,000 prefixes[1], and a single IGP distance change may cause many of these routes to change at the same

[1]To obtain up-to-date information about the size of routing tables, please reference http://www.cidr-report.org.

time. If these prefixes receive a large volume of traffic, the influence on the traffic matrix of the AS and on its downstream neighbors can be quite significant. In this paper, we quantify these effects by analyzing the IGP-triggered BGP updates in a tier-1 ISP network.

On the surface, we should be able to study hot-potato routing changes in an analytical or simulation model based on the protocol specifications. However, the interaction between the protocols depends on details not captured in the IETF standards documents, as discussed in more detail in Section II. Vendor implementation decisions have a significant impact on the timing of messages within each protocol. The design of the network, such as the number and location of BGP sessions, may also play an important role. In addition, the behavior of the routing protocols depends on the kinds of low-level events—failures, traffic engineering, and planned maintenance—that trigger IGP path changes, and the properties of these events are not well-understood. In light of these issues, our study takes an empirical approach of a joint analysis of IGP, BGP, and traffic measurements collected from a large ISP network.

Although previous studies have characterized IGP link-state advertisements [4], [5], [6], [7] or BGP update messages [7], [8], [9], [10] in isolation, this is the first study to present a joint analysis of the IGP and BGP data. The work in [9] evaluates how BGP routing changes affect the flow of traffic inside the Sprint backbone but does not differentiate between routing changes caused by internal and external events. The main contributions of this paper are:

- **Identifying hot-potato BGP routing changes:** Our algorithm for correlating the IGP and BGP data (i) generates a sequence of distance changes that may affect BGP routing decisions, (ii) classifies BGP routing changes in terms of possible IGP causes, and (iii) matches BGP routing changes with related distance changes that occur close in time.
- **Evaluation in an operational network:** We apply our algorithm to routing data collected from a large ISP network, and identify suitable values for the parameters of the algorithm. Our study demonstrates that hot-potato routing is sometimes a significant source of BGP update messages and can cause relatively large delays in forwarding-plane convergence.
- **Exploring the performance implications:** We join our stream of hot-potato routing changes with traffic measurements to evaluate the impact on the traffic matrix. We discuss how hot-potato routing changes can lead to (i) packet loss due to forwarding loops, (ii) significant shifts in routes and the corresponding traffic, and (iii) inaccuracies in measurements of the forwarding system.

These contributions are presented in Sections III–V, followed by a summary of our results in Section VI.

II. Modeling Hot-Potato Routing

In this section, we present a precise definition of a "hot potato routing change" and explain why identifying these routing changes in an operational network is surprisingly difficult.

0. Ignore if egress point unreachable
1. Highest local preference
2. Lowest AS path length
3. Lowest origin type
4. Lowest MED (with same next-hop AS)
5. eBGP-learned over iBGP-learned
6. Lowest IGP distance to egress point ("Hot potato")
7. Vendor-dependent tie break

TABLE I
STEPS IN THE BGP DECISION PROCESS

A. Hot-Potato BGP Routing Changes

The BGP decision process [1] on a router selects a single best route for each destination prefix by comparing attribute values as shown in Table I. Two of the steps depend on the IGP information. First, a route is excluded if the BGP next-hop address (i.e., the egress point) is not reachable. For example, in Figure 1, the router A does not consider the BGP route from C if A's forwarding table does not have an entry that matches C's IP address. Then, after the next five steps in the decision process, the router compares IGP distances associated with the BGP next-hop addresses and selects the route with the smallest distance—the *closest* egress point. If multiple routes have the same IGP distance, the router applies additional steps to break the tie. When the BGP decision process comes down to the IGP distance, we refer to the BGP decision as *hot potato* routing. When a change in an IGP distance leads a router to select a different best BGP route, we refer to this as a *hot potato routing change*.

To guide our characterization of hot-potato routing, we propose a simple model that captures the path selection process at a single router (which we denote as a *vantage point*):

- **Distance vector (per vantage point)**: The vantage point has a distance vector that represents the cost of the shortest IGP path to every router in the AS. The distance vector, which changes in response to link failures and modifications to the link weights, is a concise representation of the aspects of the IGP that can affect BGP routing decisions.
- **Egress set (per prefix):** Routers usually receive routes to a prefix p from multiple neighbors. When routes are tied through step 4 in the BGP decision process, step 5 says that the router selects the route learned from an external neighbor to reach p. We call each router that selects an external route to p an egress point to p (for example, routers B and C in Figure 1). Each egress point propagates the external route to p to other routers in the AS via iBGP. For routers that learn the route to p via iBGP (as router A), the routes announced by all the egress points are equally good. We call the set of all egress points to p the egress set of p.

For each prefix, the vantage point selects the egress point (from the egress set) with the smallest distance (from the distance vector). A hot-potato routing change occurs when a vantage point changes the selection of egress points for a prefix because of a change in the distance vector (i.e., that makes the new egress point closer than the old one). For example, initially router A in Figure 1 has an egress set of p of $\{B, C\}$,

distances of 10 and 9 to B and C, respectively, and a best egress point to p of C; then, when the distance to C changes to 11, A selects egress point B. Our goal in this paper is to *determine what fraction of the BGP routing changes are hot-potato routing changes in an operational network.*

B. Challenges of Characterizing Hot-Potato Routing

Given the egress set for each destination prefix and the distance vector for each vantage point at any point in time, it is relatively simple to determine which BGP messages were triggered by IGP routing changes. However, several factors conspire to make the problem extremely challenging:

Incomplete measurement data: In a large operational network, fully instrumenting all of the routers is not possible; instead, we must work with data from a limited number of vantage points. In addition, commercial routers offer limited opportunities for collecting detailed routing data—we can only collect measurements of the routing protocol messages that the routers exchange among themselves. Collecting IGP measurements is difficult, because it often requires a physical connection to a router located in a secure facility. Fortunately, in link-state protocols like OSPF and IS-IS, the routers *flood* the link-state advertisements (LSAs) throughout the network, allowing us to use data collected at one location to reconstruct the distance changes seen at other routers in the network. However, this reconstruction is not perfect because of delays in propagating the LSA from the point of a link failure or weight change to other routers in the network. Collecting BGP data from multiple routers is easier because BGP sessions run over TCP connections that do not require a physical adjacency. However, BGP messages from the operational router must traverse the network to reach the collection machine, which introduces latency; these delays may increase precisely when the IGP routes are changing. In addition, since BGP is a path-vector protocol, we only have *the best route of the monitored router*, making it difficult to know the complete set of routing choices that are available at any given time. [2]

Complex routing protocol dynamics: IGP routing changes stem from topology changes (i.e., equipment going up or down) and configuration changes (i.e., adjustments to the link weights). Monitoring the IGP messages shows only the *effects* of these events. In practice, multiple LSAs may occur close together in time (e.g., the failure of a single router or an optical amplifier could cause several IP links to fail). If one LSA follows close on the heels of another, the routing system does not have time to converge after the first LSA before the next one occurs. Similarly, a prefix may experience multiple BGP routing changes in a short period of time (e.g., a neighboring AS may send multiple updates as part of exploring alternate paths [12], [8]); or a hot-potato routing change might trigger multiple iBGP routing changes as the network converges. In addition, the global routing system generates a constant churn of BGP updates, due to failures, policy changes, and (perhaps) persistent oscillations. Continuously receiving several updates

[2] A proposal such as the IETF's BGP Monitoring Protocol [11] would allow a router to send all its routes to a prefix.

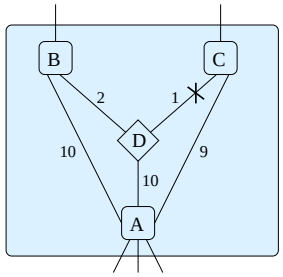

Fig. 2. Router A changes best route without distance change

a second is not uncommon. This makes it difficult to identify which BGP routing changes are caused by hot-potato routing inside the AS. The Multiple Exit Discriminator (MED) attribute introduces additional complexity because the BGP decision process compares MED values only across routes learned from the same next-hop AS, resulting in scenarios where a router's local ranking of two routes may depend on the presence or absence of a third route [13].

Hierarchy of iBGP sessions inside the AS: Large networks often employ *route reflectors* to reduce the overhead of distributing BGP information throughout the AS [14]. However, route reflectors make the dynamics of network-wide routing changes extremely complicated, because they hide some routes from the other routers in the network. In the example in Figure 2, router D is a route reflector with clients A, B, and C. Both A and D have shorter IGP paths to C than B. When the C–D link fails, router D shifts its routes from egress C to egress B. However, since A is a client of D, it too would change its routes to use egress B even though its own distance vector has not changed! Determining which BGP routes from A are caused by IGP changes requires knowing the route-reflector configuration of the network and which BGP routing changes from D were caused by the IGP. Some *under-counting* of hot-potato routing changes is inevitable, though focusing the analysis on vantage points that are "top-level" route reflectors helps limit these effects.

Vendor implementation details: Although the routing protocols have been standardized by the IETF, many low-level details depend on implementation decisions and configuration choices. The vendor implementations have numerous timers that control when the router recomputes the IGP paths, reruns the BGP decision process, and sends update messages to BGP neighbors. The router operating system may have complex techniques for scheduling and preempting tasks when multiple events occur close together in time. These router-level details can have a first-order impact on the network-wide dynamics of hot-potato routing.

Together, these issues suggest that computing an exact measure of hot-potato routing changes is extremely difficult, and that we should seek approximate numbers based on reasonable heuristics.

III. Measurement Methodology

In this section, we present our methodology for measuring hot-potato changes experienced by operational routers. Fig-

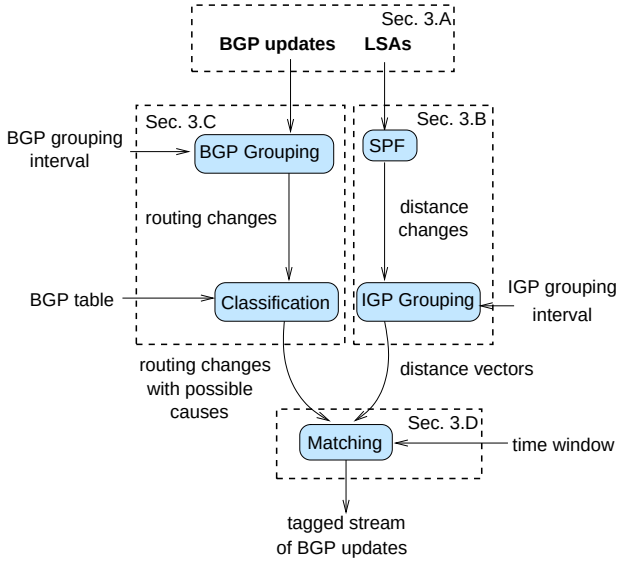

Fig. 3. Identifying hot-potato routing changes. Dotted boxes are labeled with the number of the subsection that describes it

ure 3 presents the steps to correlate BGP updates from a vantage point with OSPF LSAs. (Each dotted box represents steps described in a particular subsection.) Section III-A presents the measurement infrastructure used to collect BGP updates and OSPF LSAs. We describe how to compute the distance vector from the OSPF LSAs in Section III-B. Section III-C explains the classification of BGP routing changes in terms of the possible causes. This sets the stage for the discussion in Section III-D about how to associate BGP routing changes with related distance changes that occur close in time.

A. Measurement Infrastructure

In this paper, we study the backbone network of a tier-1 ISP. This network has hundreds of routers that are located in a few dozen Points of Presence (PoPs) spread throughout the US. The ISP uses OSPF with areas as intradomain routing protocol. OSPF weights are set to reflect the positioning in geographic regions and traffic engineering constraints. The ISP uses equal-cost paths when there are parallel paths for redundancy and better network utilization. The network also deploys an iBGP hierarchy with two route reflectors per PoP. Top-level route reflectors are connected in a full mesh.

We have deployed an OSPF and a BGP monitor in this network. Figure 4 depicts our measurement infrastructure. The OSPF monitor [15] is located at a PoP and has a direct physical connection to a router in the network. We connect our monitor to a router in area 0 to have complete view of the distances to reach each router. The monitor timestamps and archives all LSAs. The BGP monitor has iBGP sessions (running over TCP) to at least one top-level route reflector at each PoP and is in the same peer group as the clients of the route reflector. Using an *iBGP* session allows the monitor to see changes in the "egress point" of BGP routes. The BGP monitor also dumps a snapshot of its routes four times a day to provide an initial view of the best route for each prefix for each vantage point, so that we can later classify the type of BGP change as discussed in Section III-C. The OSPF and BGP monitors

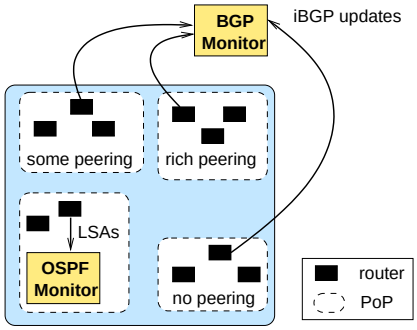

Fig. 4. Measurement infrastructure in a large tier-1 ISP backbone

run on two distinct servers and timestamp the routing messages with their own local clocks; to minimize timing discrepancies, both monitors are NTP synchronized.

Our analysis focuses on 176 days of data collected from January 2003 to July 2003. Because details of network topology, peering connectivity, and the absolute number of routing messages are proprietary, we omit router locations and normalize most of our numerical results. We study data collected from one route reflector per PoP (all Cisco routers) and, for simplicity, present the results for three of these vantage points. To explore the effects of router location and connectivity, we select three vantage points in PoPs with different properties. *Rich peering* is a router in a PoP that connects to a large number of peers, including most major ISPs. *Some peering* is a router in a PoP that connects to some but not all major peers. *No peering* is a router in a PoP with no peering connections. Most traffic is directed to egress points in two nearby PoPs. The three PoPs are located in the eastern part of the United States, relatively close to the location of the two route monitors.

Resets of the monitoring session would affect the accuracy of our results, especially if IGP routing changes are correlated with iBGP session resets. Each of the BGP monitoring sessions experienced at most five resets per month, perhaps due to temporary disruption of the monitor's connection to the rest of the network. These results suggest that IGP events were not a significant contributor to iBGP session resets in the network. In fact, the default keep-alive and hold timers for BGP sessions (60 and 180 seconds, respectively) make it unlikely that transient disruptions during IGP convergence would affect iBGP reachability. Before conducting our analysis, we eliminate all destination prefixes where the BGP routing decisions depend on MEDs; to be conservative, we exclude any prefix that had *any* BGP update with a non-zero MED attribute during the period of the data collection, which represent approximately 13% of the total number of prefixes.

B. Computing Distance Vector Changes

OSPF is a link-state routing protocol where each unidirectional link is assigned an administrative weight that is flooded throughout the network in a reliable fashion [2]. Our algorithm processes the LSAs as they arrive to continuously track the OSPF topology and compute the distance vector changes from each vantage point. First, our algorithm disregards any LSAs that do not reflect a change in the OSPF topology;

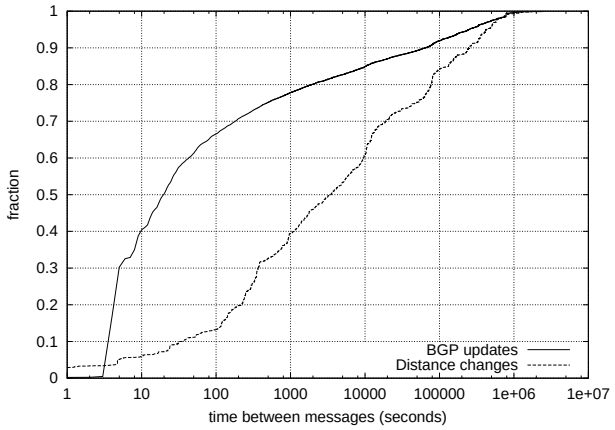

Fig. 5. CDF of message interarrivals for each protocol

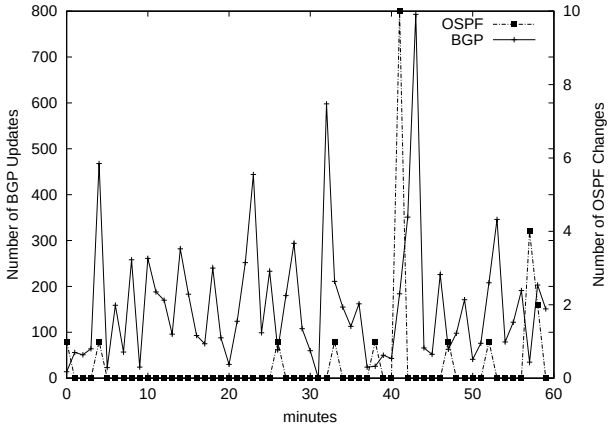

Fig. 6. One-hour time series of BGP updates and distance changes

this process excludes OSPF's periodic refresh LSAs as well as any duplicate LSAs sent in the reliable flooding process. For the remaining LSAs, we emulate the OSPF shortest-path computation [2] to determine the distance from each vantage point to every other router at the boundary of the network (i.e., any router that could serve as an egress point for one or more prefixes).

Some OSPF topology changes do not trigger distance changes. For example, some links with high OSPF weights do not appear on any shortest path (e.g., links under maintenance or provisioning); an increase in the weight or the failure of the link would not affect any of the shortest paths. Also, some links always appear as part of multiple shortest paths along with other links (e.g., parallel links between two routers). A failure of one of these links will not trigger a distance change. Other LSAs may change the distances for one vantage point but not another. Whenever an LSA changes one or more distances for a given vantage point, our algorithm produces a new distance vector for that vantage point. If the vantage point cannot reach another router (e.g., due to a failure or network partition), we represent the distance as ∞. Our study focuses on the common case of distance changes from one finite value to another.

In practice, multiple LSAs may occur close together in time. Even if these LSAs stem from different events (e.g., two independent failures), the delays in propagating the LSAs and in converging to new routes make it impossible to analyze these LSAs separately. Instead, we group distance changes that occur within a small time window into a single distance vector change. We select the interval duration based on analysis of our OSPF measurements, shown by the "distance changes" curve in Figure 5. To generate the curve, we consider the interarrival times of the distance changes between each (vantage point, egress router) pairs and plot the resulting cumulative distribution across all pairs. About 5% of the distance changes occur within ten seconds of each other. These may correspond to LSAs caused by a single physical event, such as rebooting a router. Otherwise, the curve increases gradually over the range of values. Half of the distance changes have an interarrival time of more than 3400 seconds, and 10% are more than 252,000 seconds (almost a month). In the next Section, we apply a time interval of 10 seconds for grouping distance changes; additional experiments showed that the results were not sensitive to small changes in the size of the interval.

C. Classifying BGP Routing Changes

The global BGP routing system generates a continuous stream of update messages, as shown by the example in Figure 6. This graph plots the number of BGP updates (left y-axis) and distance changes (right y-axis) seen by the "rich peering" router over one hour, with one-minute bins. In this example, the router sometimes makes several hundred BGP routing changes in a minute. In contrast, very few intervals have more than a handful of distance changes, and these changes do not necessarily cause the router to switch from one egress point to another for any prefix. The large volume of BGP updates stems, in part, from the exploration of multiple alternate routes when a router switches from one best path to another [12], [8]. These short-lived BGP routes do not correspond to stable path changes but rather the *transition* from one stable route to another. The details of path exploration depend on timing details at routers throughout the Internet. Instead, in our study, we are interested in how IGP distance changes cause a router inside the AS to switch from one stable route to another with a different egress point.

To focus on changes from one stable route to another, we group BGP updates at the same router for the same prefix that occur close together in time, based on the "BGP updates" curve in Figure 5. To generate the curve, we consider the interarrival times of the BGP updates from each vantage point for each prefix and plot the resulting cumulative distribution. More than 30% of the BGP updates have an interarrival time of five seconds or less. This stems from the 5-second minimum-route advertisement timer used by Cisco routers to pace the update messages on iBGP sessions. Previous studies have shown that interarrival times of around 30 seconds are quite common for external routing changes, since Cisco routers use a 30-second minimum-route advertisement timer for eBGP sessions [12]. In Figure 5 about two-thirds of the BGP updates have a spacing of less than 70 seconds. In the next Section, we apply a time interval of 70 seconds for grouping BGP messages to combine many of the transient updates occurring during path exploration. Additional experiments showed that

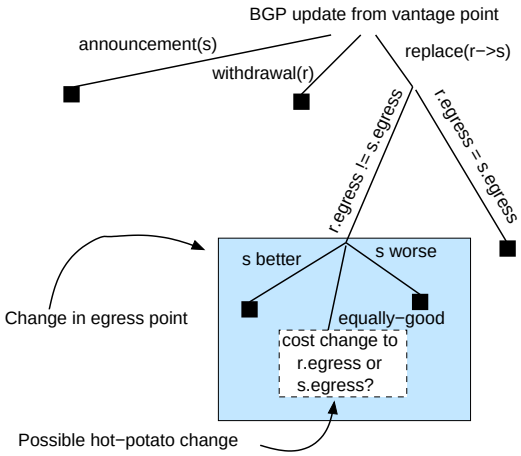

Fig. 7. Classifying BGP routing changes at a vantage point

the results were not sensitive to small changes in the size of the grouping interval.

Many BGP routing changes have no relationship to the distance vector changes in the interior of the network. Drawing on the BGP decision process, our algorithm classifies BGP routing changes in terms of their possible causes. Starting with an initial BGP routing table, we consider a stream of changes in the best route for each prefix. Figure 7 illustrates how we classify a BGP routing change from route r to route s for a prefix at a particular vantage point. Hot-potato routing changes cause a router to switch from one BGP route to another. As such, changing from or to a null route does not represent a hot-potato routing change. However, hot-potato routing changes can cause s to *replace* r. In this case, further analysis helps narrow down the possible causes. If r and s have the same egress point, a change in the distance vector cannot be responsible.

Having different egress points $r.egress$ and $s.egress$ does not necessarily imply that hot-potato routing is responsible. The new route s might be "better" than the old one at some earlier stage in the decision process; for example, s might have a shorter AS path or a larger local-preference. Alternatively, the route r might have been withdrawn; because our monitor sees only the best route at each vantage point, we can only infer that r was withdrawn if s is "worse" than r. Hence, if r and s are not "equally good" through steps 0–5 of the BGP decision process, we can dismiss hot-potato routing as a possible cause. If the routes are equally good, hot-potato routing *might* be responsible if the relative "closeness" of the two egress points has changed—making the egress point s closer than egress point r.

D. Matching Distance Changes with BGP

To further refine our inference that an IGP routing change caused the vantage point to select s, we inspect the stream of distance vectors for this vantage point to see if $s.egress$ became closer than $r.egress$ within some small time interval. We verified the correctness of this algorithm using the router testbed presented in [16]. In this scenario, all BGP routes are stable and the only changes are related to distance changes; our algorithm correctly identified the OSPF LSA that caused each BGP update. However, BGP routes are *not* stable in the operational network. Hence, our algorithm might mistakenly match a BGP routing change with an *unrelated* distance vector change. The BGP routing change might have been triggered by an external event, such as a policy change or a failure in another AS, that caused r to be withdrawn or replaced by a less attractive route. Yet, a seemingly-related distance vector change could occur nearby in time that is consistent with the vantage point's decision to switch to route s. In this situation, our algorithm would mistakenly associate the replacement of r by s with the distance change. (In practice, the IGP event might have caused a similar BGP routing change anyway if the external event had not happened first!)

Although these kinds of mismatches are difficult to avoid completely, three aspects of our algorithm reduce the likelihood of false matches: (i) preprocessing the distance vector changes and BGP update messages as discussed in Section III-B and III-C, (ii) the fine-grained classification in Figure 7 which eliminates many of the external BGP routing changes, and (iii) the careful selection of the time window for correlating the two datasets. To find the appropriate time window, we first consider distance vector changes that occur within ten minutes before or after the BGP routing change. Although our algorithm did find occasional matches over the entire 20-minute interval, the vast majority of hot-potato BGP routing changes occurred within *three minutes* of the distance vector change, for reasons we explore in more detail in the next section. In experiments where we did *not* pre-process the OSPF and BGP data, we tended to see a larger number of (presumably false) matches in the large time intervals, suggesting that our preprocessing is helpful for reducing the likelihood of false matches.

Our algorithm finds some matches where the BGP routing change appears to happen 1–2 seconds *before* the distance vector change. Although this seems counter-intuitive, this can occur in practice for two reasons. First, the OSPF LSA may take longer to reach our OSPF monitor than for the related BGP update to reach the BGP monitor. The reliable flooding of OSPF LSAs is typically implemented in software on the router, which may subject these messages to higher delays. In contrast, BGP update messages are sent via a TCP connection between two routers; the IP packets carrying these messages traverse the hardware forwarding path through the routers. Second, the BGP monitor has a coarser timestamp resolution than the OSPF monitor. To account for these two issues, we allow a small *negative* time difference between the distance vector change and the BGP change. Therefore, we believe a time window of $(-2, 180)$ is a reasonable way to avoid false matches while still capturing the bulk of the real hot-potato routing changes. We use this window for the analysis in the rest of the paper.

IV. Characterizing Hot-Potato Routing

This section presents a case study of hot-potato routing changes in an operational network. Our goal is to identify and understand the main properties of hot-potato routing changes, rather than to highlight specific numerical values that might

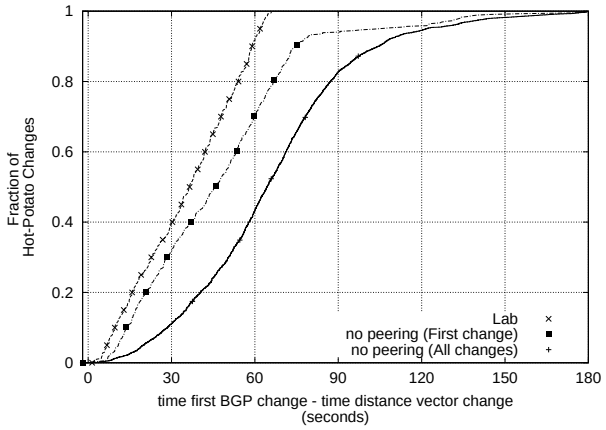

Fig. 8. CDF of time lag between the distance vector change and related BGP routing changes, using a 10-second window to group OSPF LSAs, a 70-second window to group the BGP update messages, and a $(-2, 180)$ window to correlate the distance vector changes with BGP routing changes.

vary from one network to another. Although most hot-potato routing changes occur within 60 seconds, extra delays of 1–2 minutes sometimes arise due to the iBGP hierarchy and the transfer of update messages. The frequency of hot-potato routing changes varies significantly across time and router location. Interestingly, the hot-potato BGP updates have a much more even spread across the destination prefixes than the remaining update messages.

A. BGP Reaction Time to Distance Changes

Figure 8 presents the cumulative distribution of the delay between a distance vector change and a correlated BGP routing change for the "no peering" router from January 2003 to July 2003. For comparison purposes, this graph also presents the lab results from [16], which were generated from a controlled experiment with a router in a lab, using the same router model deployed in the operational network. For this experiment all BGP routing changes were caused by an OSPF routing change. The lab curve is almost perfectly linear in the range of 5 to 65 seconds due to the influence of two timers. First, the router imposes a 5-second delay after receiving an LSA before performing the shortest-path computation to avoid multiple computations when several LSAs arrive in a short period of time [17]. A second LSA that arrives during this interval does not incur the entire five-second delay, as evidenced by the small fraction of LSAs that experienced less than five seconds of delay. Second, the router has a 60-second scan timer that runs periodically to sequence through the BGP routing table and run the BGP decision process for each prefix [18]. The BGP change does not occur until the scan process runs and revisits the BGP routing decision for this prefix. As such, the delay in the BGP routing change is uniform in $[5, 65]$, as evidenced by the straight line in the graph. A router also imposes a 10-second interval between two consecutive shortest-path calculations, which explains delays in the $[65, 70]$ range.

The graph shows a significant gap between the results for the lab experiments and the curve for *all* hot-potato changes sent by the "no peering" router. Upon receiving a new LSA, the router must (i) rerun the IGP shortest-path computation, (ii) apply the BGP decision process to select the best route for each prefix, and (iii) send update messages to BGP neighbors for the routes that have changed. The first two steps represent the time required to react to a distance vector change, and the third step depends on the number of BGP routing changes. The lab experiments evaluated only the first two steps. To have a fair comparison, we also measure and report the delay between the distance vector change and the *first* prefix experiencing a hot-potato routing change.

The graph shows that most hot-potato routing changes occur within 80 seconds of the distance vector change, which is closer to the 70 seconds upper limit of the lab results. The extra 10 seconds are explained by the rate of LSA arrivals and the number of routes in an operational router. When the rate of LSAs is higher, the likelihood of incurring the 10-second delay between consecutive shortest-path calculations increases. The scan process may require several seconds in an operational router because of the large number of BGP routes. The 60-second timer restarts after the *completion* of the previous scan; hence, the BGP reaction time also includes the time for running the scan process. These two factors contribute to longer reaction times in the operational router. We discuss the reaction times longer than 80 seconds in the next subsection.

B. Transfer Delay for Multiple Prefixes

The difference between the curve for *all* hot-potato changes and the one for the *first* change corresponds to the delay to transfer BGP updates for multiple prefixes. When a distance vector change affects a large number of prefixes, the transmission of the BGP update messages to iBGP and eBGP neighbors introduces additional delay. We study the effect of this additional delay by inspecting some large hot-potato changes. For example, for the "no peering" router one distance change affected more than $80,000$ prefixes. Although the BGP change for the first prefix occurs 66 seconds after the distance vector change, the routing change for the last prefix occurred 83 seconds later, 149 seconds after the OSPF change. This delay is determined by the volume of updates to be transferred and the TCP transmission rate between the vantage point and the BGP monitor.

In our experiments, the BGP monitor is within a few hundred miles of the "no peering" router and the update packets travel just a few hops through the network. Longer transfer delays might be possible over iBGP sessions between pairs of routers with longer round-trip times, which may also contribute to longer delays in reacting to hot-potato routing changes. We should expect lower (but still significant) delays with higher-speed routers. For instance, a test of a 40Gbps router in a controlled environment found that the recovery from a BGP session failure, for a session with 500 thousand prefixes, can take up to 18 seconds [19].

The transfer delay may also be responsible for the instances in Figure 8 in which the reaction time exceeds 80 seconds for the "first change" curve. These kinds of delays may be caused by the propagation of hot-potato BGP routing changes from

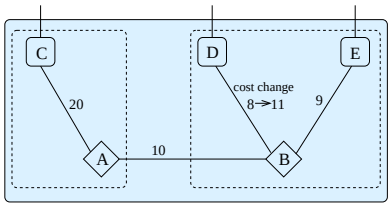

Fig. 9. Router A waits for B's decision.

one router to another, as shown in Figure 9. In the example, routers A and B are route reflectors and routers C, D, and E are egress points; C is a client of A, and D and E are clients of B. Initially, A and B select egress point D, with distances of 18 and 8, respectively. A is unaware of the route via E because B only advertises its best route to A. When the B-D distance increases to 11:

1. The LSA is flooded throughout the network and each router computes new distances to D. For example, A and B compute new distances of 21 and 11, respectively.
2. After their scan timers elapse, A and B rerun the BGP decision process. If A runs first, A selects the egress point C with a distance of 20, since this is smaller than 21. Sometime afterwards, B selects egress point E.
3. B sends the new route (with egress point E) to A, and A selects egress point E with a distance of 19.

Suppose a distance vector change triggers a large number of BGP updates from B, but some of these updates do not trigger hot-potato changes in A. Then, A may have to wait for the transfer of a number of BGP updates before experiencing a hot-potato change. This explains some of the reaction times longer than 80 seconds in Figure 8. Other instances with longer reaction times may also be due to false matches in associating a BGP routing change with a distance vector changes.

Combining the results of the "first change" curve in Figure 8 and the transfer delays, a router's reaction to distance vector changes may take 0–80 seconds for the first prefix and an additional 80 seconds (in extreme cases) for the remaining prefixes. Combining these effects, the vast majority of hot-potato changes take place within three minutes of the distance vector change, as shown by the "all changes" curve in Figure 8.

C. Temporal and Spatial Variability

The influence of hot-potato routing varies significantly across time. Figure 10 presents the number of hot-potato updates. For ease of presentation, the graph plots the days in increasing order of the number of hot-potato BGP routing changes and we only show the 46 days with higher number of hot-potato changes. The plot shows that on most days the routers did not experience *any* hot-potato routing changes. Still, on a few days the number was much higher. For the "no peering" router, one day had 400 thousand hot-potato routing changes, this unusually large number represented 82% of the BGP routing changes on that day. The variability across the days may stem from natural differences in the time and location of IGP weight changes and maintenance activity. The large variation across days makes it difficult to define a

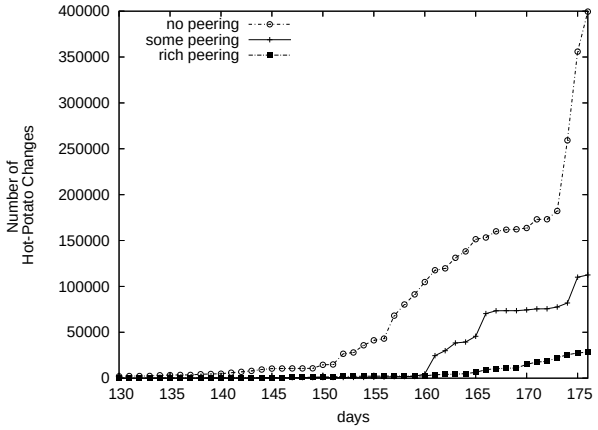

Fig. 10. Hot-potato changes across days and locations

representative statistic for the frequency of hot-potato routing changes.

Comparing the three curves in Figure 10 highlights the influence of the location of the router on the likelihood of hot-potato routing changes. Over the period of our study, the "rich peering" router was always the least affected by distance changes, as seen by the bottom curve lying very close to the x-axis in Figure 10. The likelihood that a distance change affects the selection of the BGP best route depends on the proximity of the router to each of its nearby egress points.

To better understand the impact of router location, we study the connectivity of each vantage point using one typical BGP table snapshot during our measurement period. Figure 11 presents in the x-axis the egress ID normalized by the total number of egress points and in the y-axis the number of prefixes per egress point for each vantage point in log scale. Although there are hundreds of egress points in the network, each vantage point routes most prefixes using very few of them (the top-five egress points cover approximately 65% of prefixes for each vantage point). This plot shows that the top-three egress points for the "no peering" and the "rich peering" routers are exactly the same. We inspected these egress points and found that all of them are located at the "rich peering" PoP. Given that all top-five egress points for the "no peering" router are in other PoPs, this vantage point is more sensitive to internal routing changes than the others (as seen in Figure 10). In fact, distance changes between the "no peering" router and the "rich peering" PoP have the potential to impact the BGP decision of over 90 thousand prefixes.

The fact that the top-three egress points for the "rich peering" router are located in its PoP explains why very few distance changes cause the router to select a different egress point. This suggests that a natural way to reduce the number of hot-potato routing changes would be to have rich peering at *every* PoP. However, having rich peering at all locations is infeasible in practice, due to distance and geographic constraints. A service provider is bound to have routers in some remote locations that are not close to PoPs of the other large providers. A study of sensitivity of the same network to single link and router failures confirms this variation across routers [20]. In fact, this study pinpoints the

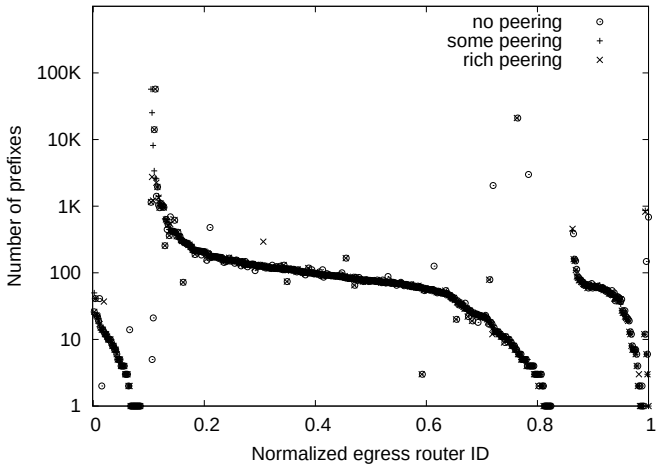

Fig. 11. Number of prefixes per egress router

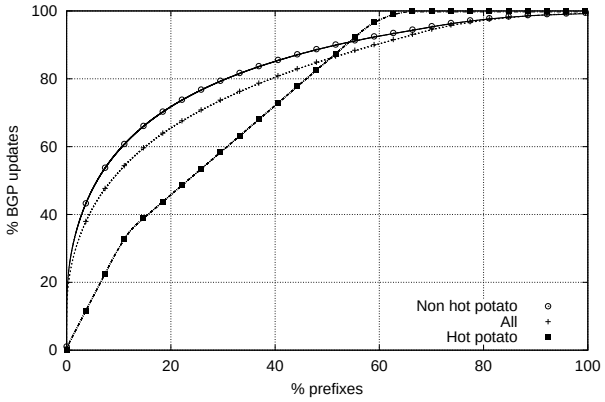

Fig. 12. CDF of BGP updates across destination prefixes

"no peering" router as one of the most at risk of hot-potato routing changes in this network. Both the model presented in [20] and the tool presented in [21] can be used to determine the most sensitive routers in a network.

To summarize, the impact of hot-potato routing varies across days and locations; our measurements show that the fraction of BGP routing changes due to hot-potato routing varied from 0% to over 80% of the BGP changes of a router during a day. Individually, the three routers do not experience hot-potato routing changes all that often. The "no peering" router experiences hot-potato changes once a week on average. However, for an entire network, across all prefixes, these changes occur much more often than per-router results suggest. There are hundreds of routers in the network and they experience hot-potato routing changes at different times. The impact of internal changes depends on both the location and the internal events that happened in a day. In the same day a router in one location of the network may experience a large number of hot-potato changes, whereas another has none.

D. Hot-Potato Variation Across Prefixes

Previous studies have shown that a small fraction of unstable prefixes are responsible for most of the BGP route updates [7], [8] and that the popular prefixes responsible for the bulk of the traffic have very few BGP updates [8], [9]. The BGP routes for the remaining prefixes stay the same for days or weeks at a time. An interesting question is to what extent these results translate to hot-potato changes. Figure 12 plots the cumulative distribution of BGP update messages across the destination prefixes for the "no peering" router for June 2003, which was a typical month in terms of hot-potato routing changes. To compare our results with previous work, the graph plots the number of *BGP update messages* rather than the number of BGP routing changes. The prefixes are sorted according to their contribution to the number of BGP messages. The middle curve corresponds to all of the BGP messages. About 20% of the prefixes contribute 65% of the BGP updates, consistent with previous findings [7], [8]. However, the bottom curve shows that the distribution of BGP updates caused by *hot-potato* routing changes has a much more even spread across the prefixes.

The broader distribution across prefixes occurs because distance vector changes can affect the distances to reach the egress points for a wide variety of prefixes. Still, some prefixes do not experience *any* hot-potato BGP updates, as seen in the flat portion in the upper-right part of the graph. This part of the curve corresponds to prefixes with a very small number of egress points, including the prefixes that have a *single* egress point. Every router in the network would always pick this single egress point as the best egress point for the prefix. The relatively uniform distribution across the remaining prefixes may have important implications. For prefixes that generally have stable *eBGP*-learned routes, internal distance changes could be a primary cause of the BGP routing changes observed inside an AS. Since some of these prefixes may be responsible for a large volume of traffic, limiting the frequency of hot-potato routing changes may be useful to avoid large traffic shifts and transient performance disruptions.

E. Cause and Duration of Distance Vector Changes

A manual inspection of the distance vector changes responsible for the hot-potato changes that affect the largest fraction of prefixes indicates that their main cause is link maintenance. This result is consistent with a study of the Sprint network [22]. The network of study has the policy of increasing the link weight to a very high value before taking down the link, a procedure called "cost out" a link [23]. Most of the large hot-potato shifts are triggered by distance changes caused by cost-in and cost-out procedures, and link failure or recovery. Router failures or reboots are rarer.

Many of these changes last for more than an hour. Figure 13 shows the complementary cumulative distribution function of the duration of link down events that triggered hot-potato changes at least at one of the three vantage points. Although around 25% of hot-potato changes recover within ten minutes, the majority of them (approximately 60%) last for more than one hour. This result might seem in contrast with a previous study that shows that only 10% of link failures are longer than 20 minutes [6]. However, we only focus on the subset of link failures that trigger hot-potato changes and *very short* link

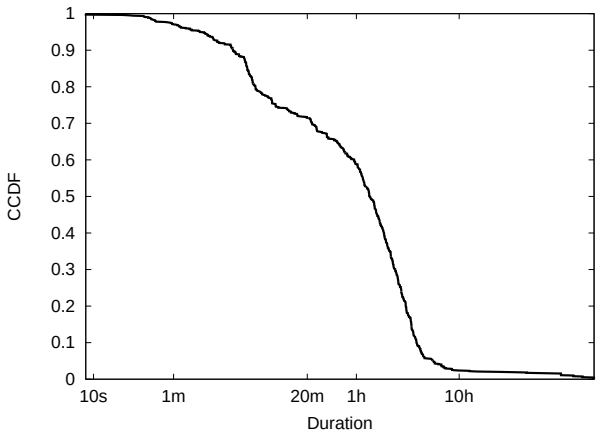

Fig. 13. CCDF of the duration of hot-potato changes

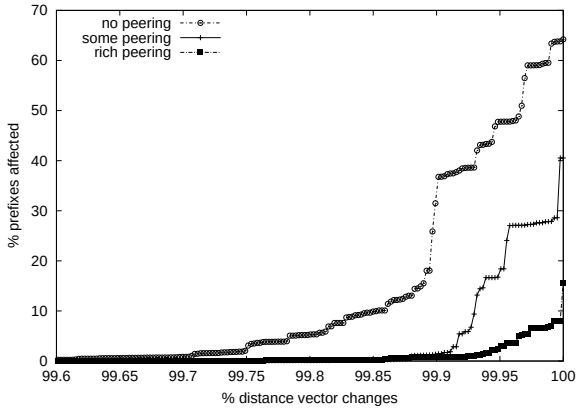

Fig. 14. Fraction of prefixes affected by distance vector change

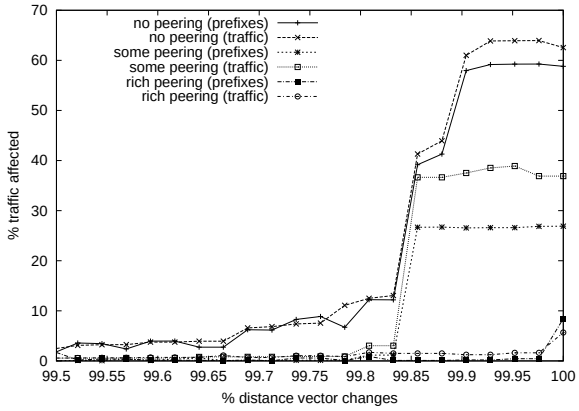

Fig. 15. Fraction of traffic affected by internal changes.

failures rarely last long enough to impact the BGP selection process. Given that the BGP scan process runs only once a minute, links may fail and recover without ever triggering BGP routing changes.

V. Implications of Hot Potatoes

Hot-potato changes in BGP routing influence network performance by causing shifts in the flow of traffic to neighboring domains and extra delays in the convergence of the forwarding plane. In addition, hot-potato changes can introduce inaccuracy in active measurements of the forwarding plane and external monitoring of BGP update messages.

A. Performance Degradation

1) Routing and Traffic Shifts: Hot-potato routing can sometimes cause a router to change the egress points for multiple destination prefixes, which will lead to a large number of BGP update messages at the same time. Even if these destination prefixes carry no traffic, this burst of updates may disrupt the control plane by temporarily overloading the CPU of the routers. In Figure 14, we explore how many destination prefixes are affected at a single router when a distance change occurs. More than 99% of the distance changes do not affect the egress point for any prefix. The vast majority of intradomain events occur far away from the router, and as result do not affect the path distances for nearby egress points. Even when changes occur closer to the router, they might not affect the router's local ranking of the two closest egress points for a given prefix or might not last long enough to impact the BGP decision. However, when hot-potato routing changes *do* occur, the effects can be dramatic. For the "no peering" router in the top curve in Figure 14, 0.1% of the distance changes affect the BGP route for more than 40% of the prefixes.

These kinds of routing changes can lead to sudden increases in traffic at the new egress points and along the downstream paths. To estimate these effects we used Cisco's Netflow [24] measurements at each vantage point as explained in [25]. These measurements report the number of bytes in ten-minute bins that enter a vantage point toward each destination prefix. In Figure 15, we replot the graph from Figure 14 together with the fraction of traffic affected by distance vector changes for March and April 2004, which is the timeframe for which we also have traffic data available. Although the timeframe is different, the "prefixes" curves in Figure 15 are qualitatively similar to that in Figure 14.

Figure 15 shows that the percent of traffic affected by a distance vector change is roughly the same as the percent of prefixes. In fact, for the very large hot-potato routing changes, the traffic shift is even larger than the fraction of prefixes would suggest. This occurs because the large shifts affect nearly every prefix that has multiple egress points. The remaining prefixes include a large number of smaller customers that connect to the backbone at a single location. On average, these customers do not receive as much traffic as the other prefixes that are reachable via multiple egress points. As a result, it is precisely the more *popular* destination prefixes that are most affected by the hot-potato routing changes, leading to even larger shifts in traffic than expected.

In fact, our detailed analysis of the Netflow data [25] shows that:

1. **Although the likelihood of large traffic fluctuations is small, big changes do sometimes occur.** In any given ten-minute time interval, less than 0.02% of the traffic matrix elements studied have a traffic variation of more than 4 times the normal traffic variations. However, some elements vary by more than 4 times the normal variations several times a week.
2. **Most routing changes do not cause much variation in**

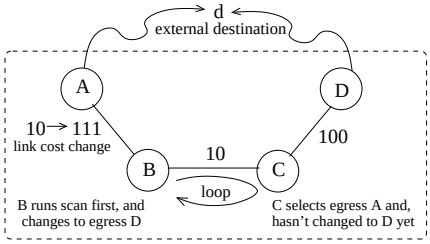

Fig. 16. Transient forwarding loop for packets destined to d

the traffic matrix. Previous studies [9], [8] have shown that routing changes typically do not cause large traffic shifts; most BGP routing changes affect destination prefixes that receive very little traffic.

3. **Routing changes are responsible for many of the large traffic shifts:** In nearly 60% of the instances where a traffic-matrix element fluctuated by more than 10 times the normal variation for the traffic change could be explained by a BGP routing change. In particular, the largest variations of the traffic matrix were caused by hot-potato routing changes.

2) Slow Forwarding-Plane Convergence: Compared to other kinds of routing changes, hot-potato routing changes cause longer delays in forwarding-plane convergence, since each router must recompute its IGP routes *and* rerun the BGP decision process before updating the forwarding table. Differences in when the routers revisit their BGP decisions can lead to transient forwarding loops, as illustrated in Figure 16. In this example, the AS has four routers and two egress points to prefix d. The numbers on the edges represent the IGP link weights, and we omit the full-mesh of iBGP sessions for simplicity. At first, routers B and C both identify router A as the closest egress point, causing C to direct traffic to d through B. When the weight of the B–A link increases to 111, both routers eventually switch to the route learned at D. However, if B runs its BGP decision process first and updates its forwarding table, B starts forwarding traffic destined to d toward D while C continues to forward the traffic toward A—resulting in a forwarding loop.

During the interval before C runs its decision process and updates its forwarding-table entry for d, all packets destined to d are caught in a forwarding loop between B and C. The packets would repeatedly traverse the loop until the IP Time-to-Live (TTL) field expires, causing one of the routers to discard the packet. The forwarding loop causes packet loss for the hosts communicating with d, and increased congestion for other traffic traversing the B–C link. Depending on the alignment of the BGP scan timers on the two routers, this problem can persist for up to 60 seconds, even though the intradomain routing protocol has converged[3]. If TCP transfer latency or the iBGP hierarchy cause larger delays in forwarding-plane convergence, the loops can persist even longer. Such loops would not happen in a network running MPLS, because packets would be tunneled from the ingress router directly to the egress.

According to a previous study of packet-level measurements in a large ISP backbone [26], most forwarding loops last for less than 10 seconds. This is consistent with typical delays for IGP convergence [6], [27]. However, the study also found that, for one of the links, about 35% of the loops persisted for 10–60 seconds. Based on our results, we speculate that these forwarding loops can be explained by hot-potato routing changes.

The convergence problem, while serious, can be addressed by changing router implementation. There are already many changes of router implementation and configuration to reduce IGP convergence [28], router vendors can also change the interaction between OSPF and BGP inside routers. Routers could have an event-driven implementation that immediately revisits the BGP routing decisions after a change in the intradomain topology. For instance, Juniper routers and newer versions of Cisco's IOS no longer have a scan timer.

B. Measurement Inaccuracies

Active measurement analysis of the performance of IP networks or passive measurement analysis of routing and traffic that ignore the interaction between IGPs and BGP may lead to inaccurate conclusions.

1) Active Probes of the Forwarding Plane: The effects of slow forwarding-plane convergence may be difficult to capture using traditional active measurement techniques. Service providers and third-party measurement companies deploy probe machines in various parts of the network in order to exercise the paths between pairs of hosts. Referring to Figure 16, suppose the provider connected one probe machine to router A and another to router D. Probe packets sent from A to D would traverse the path A–B–C–D. When the IGP weight of the B–A link changes, these probes might experience temporary loss while the IGP reconverges. However, the forwarding path of the probe packets would *not* be affected by the 60-second scan timer since there would be no change in the egress point used to reach the destination address of the probe packets; both B and C continue to use the egress point D to reach the destination probe machine. This is true, in general, for probe machines that connect to a single location inside an AS. As such, measurements between these kinds of probe machines would only capture the transient effects of *IGP* convergence, and not the combined *IGP-BGP* convergence process. Accurately capturing the performance impact of hot-potato routing changes would require a more complex active measurement infrastructure with probe machines reachable through multiple egress points.

2) External Analysis of BGP Updates: A hot-potato routing change does not necessarily cause an AS to advertise new BGP routes to neighboring ASes. First, the export policy for the eBGP session might filter the route. This decision depends on the commercial relationship with the neighbor (e.g., a route learned from one peer would not be exported to another) and on whether route aggregation is performed. Second, the router

[3]Note that the extra convergence delay for hot-potato routing changes does *not* affect the stability of the forwarding path for the iBGP sessions themselves. The IP packets sent over iBGP sessions travel between routers within the backbone and the forwarding of traffic between these routers depends only on the IGP! The delivery of BGP updates to our route monitor is not affected either, since the network has a single egress point to reach the monitor.

might decline to forward the new route if it does not differ significantly from the old route. For example, routers typically perform *non-transitive attribute filtering* to avoid propagating routes that differ only in local attributes (like BGP next-hop or local-preference) rather than global ones (such as AS path). Third, the router might not propagate the route due to BGP timers, such as the minimum-route advertisement timer, that are used to pace the rate of updates to neighboring ASes. If the router changes its best BGP route for the prefix multiple times during the interval, the intermediate BGP routes would not be visible to the neighbor.

For a rough estimate of the externally-visible updates, we look at BGP routing changes that affect the *AS path attribute*, since these would be propagated to neighboring domains subject to the export policy and the BGP timers. In Figure 17, if A is using the route via AS 3, when C switches to egress point B; we would not classify this routing change as externally visible. However, if router A's best route was the one learned from AS 2, the AS path would change; router C would propagate the new route to its eBGP neighbors. Looking over the month of June, we estimate that around 14% of the hot-potato routing changes seen at the "no peering" router would be sent to its neighbors; this would account for 5% of the externally-visible BGP routing changes. For the "some peering" router, these two numbers are 5% and 2%, respectively—about 60% smaller than for the "no peering" router. Although these average numbers are relatively small, the values vary substantially from day to day; on one day *all* hot-potato updates at all three routers had changes in the AS path.

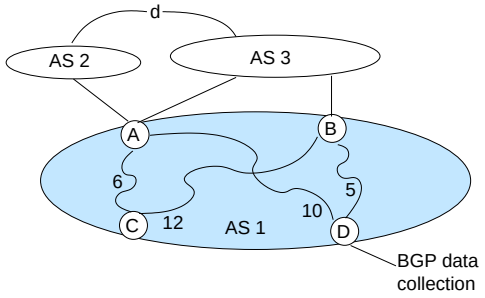

Fig. 17. BGP changes are not detected at data collection point.

These externally-visible BGP updates may affect the results of research studies based on public BGP routing data [29], [30] collected from eBGP sessions with routers in large ASes throughout the Internet. Depending on which router in an ISP network connects to these public servers, the contribution of hot-potato routing changes to the data may vary significantly! Hot-potato routing implies that different routers in an AS may pick different BGP-level routes. Referring again to Figure 17, suppose that router A chooses the route via AS 2 based on an arbitrary tie break, such as the router ID. Based on hot-potato routing, router D selects the route through B and router C selects the route through A. As such, BGP data collected from D would only reveal the route via AS 3. Now suppose that a failure occurs on the link connecting router A to AS 2. Then, both A and C would switch to the route via AS 3, which may lead to a change in the properties of the end-to-end paths for traffic entering AS 1 at router C. However, the link failure does *not* cause a change in the BGP route at D and, as such, the change is not visible to the measurement system. When viewed from outside the AS, a hot-potato routing change that affects a large number of prefixes in one network may also be indistinguishable from a BGP session reset at another nearby location. This lack of visibility is a challenge for tools for locating the origin of BGP instability such as [31], [32], as discussed in more detail in [33].

VI. Summary

The interplay between intradomain and interdomain routing has important implications on the stability and efficiency of Internet routing and, in turn, on end-to-end performance. In this paper, we presented a methodology for joint analysis of OSPF and BGP measurement data and a characterization of the interplay between the protocols in an operational network. Our results show that hot-potato routing plays an important role in BGP routing changes, and that BGP updates can lag 60 seconds (or more!) behind the related IGP events. This delay can lead to surprisingly high latency for forwarding-plane convergence that greatly exceeds the typical delays for IGP convergence [6], [27].

The frequency and impact of hot-potato routing depends on the topology and configuration of the network under study. Indeed, even routers in the same network can be more or less impacted by hot-potato routing changes depending on their location and on the intradomain routing changes that happened during the measurement period. This dependency on vantage point has important implications for network performance and routing measurements. We also show that large traffic variations, while rare, do sometimes happen. Although most routing changes typically do not affect much traffic, routing is usually a major contributor to large traffic variations. In particular, hot-potato routing changes are responsible for the largest shifts in the traffic matrix.

After the publication of our initial results on hot-potato routing [16], two follow-up studies have confirmed and extended our results. First, the work in [34] analyzed the effects of BGP routing changes in the same backbone network we studied, using newly-available BGP feeds from each of the border routers. Capitalizing on the additional data, the study confirmed that hot-potato routing changes were responsible for most of the large routing and traffic shifts. Comparing directly with our results, the study confirmed that 95% of the large hot-potato shifts were also detected by our algorithm. This demonstrates the accuracy of our algorithm. Second, researchers at Sprint showed that network operators need to account for the influence of the IGP weights on the selection of egress points when doing traffic engineering [35]. Setting IGP link weights without accounting for possible changes in the egress points can lead to routing configurations that cause unnecessary congestion.

Although network designers and operators can try to prevent hot-potato routing changes, we believe that the Internet's routing architecture should evolve to have less coupling between

interdomain and intradomain routing. We are exploring new approaches for egress-point selection as part of our ongoing work [36].

ACKNOWLEDGMENTS

We would like to thank Jay Borkenhagen, Nick Feamster, Flavio Junqueira, Rich Kwapniewski, Zhuoqing Morley Mao, Dan Pei, Geoffrey M. Voelker, Jia Wang, and the anonymous reviewers for their comments. Thanks also to Alex Gerber for his help with the Netflow data.

REFERENCES

[1] Y. Rekhter, T. Li, and S. Hares, "A Border Gateway Protocol 4 (BGP-4)." RFC 4271, January 2006.
[2] J. Moy, "OSPF Version 2." RFC 2328, April 1998.
[3] R. Callon, "Use of OSI IS–IS for Routing in TCP/IP and Dual Environments." RFC1195, December 1990.
[4] D. Watson, C. Labovitz, and F. Jahanian, "Experiences with Monitoring OSPF on a Regional Service Provider Network," in *Proc. International Conference on Distributed Computing Systems*, pp. 204–213, May 2003.
[5] A. Shaikh, C. Isett, A. Greenberg, M. Roughan, and J. Gottlieb, "A Case Study of OSPF Behavior in a Large Enterprise Network," in *Proc. Internet Measurement Workshop*, November 2002.
[6] G. Iannaccone, C.-N. Chuah, R. Mortier, S. Bhattacharyya, and C. Diot, "Analysis of link failures in an IP backbone," in *Proc. Internet Measurement Workshop*, November 2002.
[7] C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental Study of Internet Stability and Wide-Area Network Failures," in *Proc. International Symposium on Fault-Tolerant Computing*, June 1999.
[8] J. Rexford, J. Wang, Z. Xiao, and Y. Zhang, "BGP Routing Stability of Popular Destinations," in *Proc. Internet Measurement Workshop*, November 2002.
[9] S. Agarwal, C.-N. Chuah, S. Bhattacharyya, and C. Diot, "Impact of BGP Dynamics on Intra-Domain Traffic," in *Proc. ACM SIGMETRICS*, June 2004.
[10] C. Labovitz, R. Malan, and F. Jahanian, "Internet Routing Instability," *IEEE/ACM Trans. Networking*, vol. 6, pp. 515–558, October 1998.
[11] J. Scudder, "BGP Monitoring Protocol," August 2005. Expired Internet Draft, draft-scudder-bmp00.txt.
[12] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, "Delayed Internet Routing Convergence," *IEEE/ACM Trans. Networking*, vol. 9, pp. 293–306, June 2001.
[13] T. Griffin and G. Wilfong, "An Analysis of the MED Oscillation Problem in BGP," in *Proc. International Conference on Network Protocols*, 2002.
[14] S. Halabi and D. McPherson, *Internet Routing Architectures*. Cisco Press, second ed., 2001.
[15] A. Shaikh and A. Greenberg, "OSPF Monitoring: Architecture, Design and Deployment Experience," in *Proc. USENIX Symposium on Networked Systems Design and Implementation*, March 2004.
[16] R. Teixeira, A. Shaikh, T. Griffin, and J. Rexford, "Dynamics of Hot-Potato Routing in IP Networks," in *Proc. ACM SIGMETRICS*, June 2004.
[17] Cisco, "Configure Router Calculation Timers." `http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/12cgcr/np1_c/1cprt1/1cospf.html#xtocid2712621`.
[18] Cisco, "Understanding BGP Processes on Cisco." `http://www.cisco.com/warp/public/459/highcpu-bgp.html#topic1`.
[19] C. Rossenhovel, "40-Gig Router Test Results." Light Reading, November 2004. Available from `http://www.lightreading.com/document.asp?site=testing&doc_id=63606&page%_number=6.`
[20] R. Teixeira, T. Griffin, A. Shaikh, and G. Voelker, "Network sensitivity to hot-potato disruptions," in *Proc. ACM SIGCOMM*, (Portland,OR), September 2004.
[21] B. Quoitin and S. Uhlig, "Modeling the routing of an autonomous system with C-BGP," *IEEE Network Magazine*, vol. 19, November 2005.
[22] G. Iannaccone, C.-N. Chuah, S. Bhattacharyya, and C. Diot, "Feasibility of IP Restoration in a Tier-1 Backbone," *IEEE Network Magazine*, March 2004.
[23] R. Teixeira and J. Rexford, "Managing routing disruptions in internet service provider networks," *IEEE Communication Magazine*, March 2006.
[24] Cisco, "Sampled Netflow." `http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/120newft/120limit/120s/120s11/12s_sanf.htm`.
[25] R. Teixeira, N. Duffield, J. Rexford, and M. Roughan, "Traffic Matrix Reloaded: Impact of Routing Changes," in *Proc. of Passive and Active Measurement Workshop*, vol. 3431 of *Lecture Notes in Computer Science*, (Boston, MA, USA), pp. 251–264, Springer, March 2005.
[26] U. Hengartner, S. B. Moon, R. Mortier, and C. Diot, "Detection and Analysis of Routing Loops in Packet Traces," in *Proc. Internet Measurement Workshop*, November 2002.
[27] C. Alaettinoglu, V. Jacobson, and H. Yu, "Toward Milli-Second IGP Convergence," November 2000. Expired Internet Draft, draft-alaettinoglu-isis-convergence-00.txt.
[28] P. Francois, C. Filsfils, J. Evans, and O. Bonaventure, "Achieving sub-second IGP convergence in large IP networks," *ACM SIGCOMM Computer Communication Review*, vol. 35, July 2005.
[29] "Route Views Project." `http://www.routeviews.org`.
[30] "RIPE NCC RIS." `http://www.ripe.net/ripencc/pub-services/np/ris-index.html`.
[31] M. Caesar, L. Subramanian, and R. H. Katz, "Towards localizing root causes of BGP dynamics," Tech. Rep. CSD-03-1292, UC Berkeley, November 2003.
[32] A. Feldmann, O. Maennel, Z. M. Mao, A. Berger, and B. Maggs, "Locating Internet Routing Instabilities," in *Proc. ACM SIGCOMM*, September 2004.
[33] R. Teixeira and J. Rexford, "A Measurement Framework for Pin-Pointing Routing Changes," in *Proc. ACM SIGCOMM Network Troubleshooting Workshop*, September 2004.
[34] J. Wu, Z. M. Mao, J. Rexford, and J. Wang, "Finding a needle in a haystack: Pinpointing significant BGP routing changes in an IP network," in *Proc. USENIX Symposium on Networked Systems Design and Implementation*, May 2005.
[35] S. Agarwal, A. Nucci, and S. Bhattacharyya, "Measuring the shared fate of IGP engineering and interdomain traffic," in *Proc. International Conference on Network Protocols*, November 2005.
[36] R. Teixeira, T. Griffin, M. Resende, and J. Rexford, "TIE Breaking: Tunable Interdomain Egress Selection," *IEEE/ACM Trans. Networking*, vol. 15, August 2007.

Renata Teixeira received the B.Sc. degree in computer science and the M.Sc. degree in electrical engineering from Universidade Federal do Rio de Janeiro, Brazil, in 1997 and 1999, respectively, and the Ph.D. degree in computer science from the University of California, San Diego, in 2005. During her Ph.D. studies, she worked at the AT&T Research. She is currently a Researcher with the Centre National de la Recherche Scientifique (CNRS) at LIP6, Université Pierre et Marie Curie-Paris 6, Paris, France.

Aman Shaikh is a Technical Specialist at AT&T Labs – Research. He obtained his Ph.D. and M.S. in Computer Engineering from the University of California, Santa Cruz in 2003 and 2000 respectively. He also holds a B.E. (HONS) in Computer Science and an M.Sc. (HONS) in Mathematics from the Birla Institute of Technology and Science, Pilani, India. His current research interests include IP routing, and network management and operations. He has published several research and technical papers in these areas.

PLACE
PHOTO
HERE

Timothy G. Griffin received the B.S. degree in mathematics from the University of Wisconsin, Madison, in 1979 and the Ph.D. degree in computer science from Cornell University, Ithaca, NY, in 1988. He has worked at Bell Laboratories, AT&T Research, and Intel Research. He is currently on the faculty of the Computer Laboratory at the University of Cambridge, Cambridge, U.K.

PLACE
PHOTO
HERE

Jennifer Rexford received the B.S.E. degree in electrical engineering from Princeton University, Princeton, NJ, in 1991 and the M.S.E. and Ph.D. degrees in computer science and electrical engineering from the University of Michigan, Ann Arbor, in 1993 and 1996, respectively. From 1996 to 2004, she was a member of the Network Management and Performance Department at AT&T Research. She is currently a Professor in the Computer Science Department at Princeton University.