BGP Routing Changes: Merging Views from Two ISPs

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

Sharad Agarwal
Microsoft Research
Redmond, WA
sagarwal@microsoft.com

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

ABSTRACT

Large ISPs experience millions of BGP routing changes a day. In this paper, we discuss the impact of BGP routing changes on the flow of traffic, summarizing and reconciling the results from six measurement studies of the Sprint and AT&T backbone networks.

Categories and Subject Descriptors

C.2.2 [**Network Protocols**]: Routing Protocols; C.2.3 [**Network Operations**]: Network Monitoring

General Terms

Measurement, Management, Performance

Keywords

Traffic demands, trafic matrix, hot-potato routing, BGP

1. INTRODUCTION

The delivery of Internet traffic depends on the distributed operation of routing protocols running in and between multiple Autonomous Systems (ASes). Although these routing protocols are described in standards documents and computer-networking books, understanding how the protocols actually operate and interact depends on knowing how they are used in practice. In this paper, we offer a view "from the inside" of two large Internet Service Provider (ISP) backbones—AT&T and Sprint. Our goal is to codify, in one place, how routing protocols interact, and how network events—such as equipment failures, traffic engineering, planned maintenance, and routing changes in other ASes—affect the flow of traffic. We synthesize the results of several measurement studies that jointly analyze routing and traffic data collected inside these ISPs, and we explain why these studies draw seemingly contradictory conclusions about the importance of routing changes on the flow of traffic. Our hope is that this paper aids researchers in designing simulation experiments and analyzing publicly-available measurements of the Internet routing system.

In the next section, we present an overview of how ISPs use the Border Gateway Protocol (BGP) and Interior Gateway Protocols (IGPs), such as OSPF and IS-IS, to compute the forwarding tables that direct traffic through the network. We then delve into several important questions about how BGP routing changes affect the flow of traffic, including:

- Does the large and continuous volume of BGP update messages have a significant impact on the flow of traffic through an AS? Why or why not?
- How often do large traffic shifts? Are they primarily caused by load fluctuations (e.g., due to flash crowds or denial-of-service attacks) or are routing changes a major contributor?
- What kinds of routing changes lead to large traffic shifts? What operational practices reduce the likelihood of large shifts?

To answer these questions, we summarize the key findings of measurement studies of the two ISP networks [1, 2, 3, 4, 5, 6]. We also discuss how differences in the measurement methodology affect the results, and offer guidelines for future studies.

2. OPERATIONAL VIEW OF IP ROUTING

The performance of an IP network or Autonomous System (AS) depends on the distribution of the incoming traffic over the available resources. This distribution is determined by where traffic *enters* and *leaves* the AS, and the forwarding path inside it. We first discuss how ISPs compose routing protocols to control how routers build their forwarding tables, and then we abstract the main aspects of the routing protocols and offered traffic into a simple model that guides our discussion.

2.1 Routing Protocol Interaction

In large ISP networks, the forwarding table at each router depends on the interaction between the routing protocols running in and among thousands of ASes. *Interior Gateway Protocols* (IGPs), such as OSPF and IS-IS, are responsible for determining the paths between routers inside the AS. IGPs compute shortest paths based on link metrics assigned by administrators. In contrast, the *Border Gateway Protocol* (BGP) is responsible for exchanging route information of external destinations with neighboring ASes and propagating reachability information within an AS. A router combines the BGP and IGP information to construct the forwarding table that maps each destination prefix to one or more outgoing links.

A large backbone network typically has multiple BGP-speaking routers, and BGP sessions with multiple neighboring ASes. The AS uses *external BGP* (eBGP) to exchange information in sessions with neighboring ASes. For example, in Figure 1 both routers A and B have eBGP sessions with the neighbor AS. A BGP route has a number of attributes (such as next-hop, AS-path, origin type, and Multiple-Exit-Discriminator) that are conveyed in route advertisements and can be manipulated by local policies. The router applies *import policies* to filter unwanted routes and to manipulate the attributes of the remaining routes. The router then invokes the *BGP decision process* to select exactly one "best" route for each destination prefix among all the routes learned from its neighbors. The decision process consists of a sequence of rules for comparing BGP routes, as summarized in Table 1. If two routes are "equally good" through the first five steps, the IGP distances drive the decision.

1. Highest local preference
2. Lowest AS path length
3. Lowest origin type
4. Lowest Multiple-Exit Discriminator (with same next-hop AS)
5. eBGP-learned over iBGP-learned
6. Lowest IGP distance to egress point ("Hot potato")
7. Lowest router-id of BGP speaker

Table 1: Main steps in the BGP decision process.

Different routers in an AS apply the BGP decision process independently and might select different "best" routes for the given prefix, depending on their locations in the network. Both routers A and B select the route learned from the neighbor AS to the destination prefix p, and then propagate it to routers C, D, and E via *internal BGP* (iBGP) sessions. For routers inside the AS, both routes to reach the outside destination p (learned from A and B) look "equally good" through step 5 of the BGP decision process. This leaves C, D, and E with the dilemma of choosing between egress points A and B to forward packets to p. Step 6 of the BGP decision process represents what is called *early-exit* or *hot-potato* routing. Routers direct traffic to the *closest* egress point—the egress with the smallest IGP distance (e.g., C selects router A with IGP distance 9 from C, whereas E selects router B). The IGP tie-break plays a crucial role in many BGP routing decisions, since an ISP often has multiple eBGP sessions with each neighboring AS, and may learn routes to the same destination prefix from multiple neighbors.

2.2 Model of Network-wide Traffic Flow

We now introduce a simple model that captures the properties of the routing system that determine the flow of traffic[1]. We summarize this notation in Table 2. We define the *IGP distance* $d(i, e)$ as the sum of the metrics of the links in the shortest path between two routers i and e in an AS. In Figure 1, the IGP distance from router C to A is $d(C, A) = 9$.

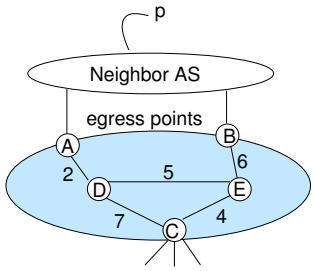

Figure 1: Example of BGP/IGP interaction to select a route to an external destination prefix p.

We use $r(i, p)$ to represent the route selected by the BGP decision process at router i to reach prefix p; $r(i, p)$ contains all BGP attributes of the route, including the egress point. Each router that learns a route from an eBGP session to a destination prefix p is a potential *egress router* for packets destined to p. We define the *egress set* E_p as the set of all egress routers that have "equally-good" eBGP-learned routes for p. We consider that two routes $r(e, p)$ and $r(e', p)$ are equally good if they are tied up to the IGP comparison step in the BGP decision process. In Figure 1 both routers A and B learn equally-good BGP routes to p, so the egress set of p is $E_p = \{A, B\}$. All other routers in the AS need to select one of the egress points in E_p to forward traffic to p. We use $b(i, p)$ to represent the the egress point that router i selects to forward traffic to p, i.e., $b(i, p) = argmin_e\{d(i, e) \mid e \in E_p\}$. In the example, $b(C, p) = A$ (which means that $r(C, p) = r(A, p)$), because C directs traffic to egress point A, which is closer than B.

[1]For scalability reasons, some ISPs introduce hierarchy through IS-IS/OSPF areas and BGP route reflectors or confederations. For simplicity, our illustrative model does not capture these details.

$d(i, e)$	IGP distance between routers i and e
p	destination prefix
$r(i, p)$	BGP route selection
E_p	egress set of p
$b(i, p)$	egress router selected by i to forward traffic to p
$V(i, p)$	traffic demand from i to p
$\mathcal{T}M(i, e)$	traffic from ingress i to egress e

Table 2: Summary of notation.

Although this model captures the outcome of the path-selection process, instead of the dynamics of how routers select these paths, it is useful for describing the impact of routing changes after routing convergence. The BGP route to a destination prefix $r(i, p)$ may change because of a variety of events (such as equipment failures or reconfiguration of BGP policies) that happen inside or between a number of ASes. We classify BGP route selection changes as:

- A **BGP routing change** (Δr) is a change in any of the attributes of the route. For example, the network administrators of prefix p may decide to buy service directly from the main AS in Figure 1 and connect to router A. A will then change its best route to use the shorter route to p.
- An **egress-point change** (Δb) involves a change in the BGP next-hop attribute. In the previous example, when A changes to the direct route to p, it also experiences an egress point change, because the egress link is now different. Router C, however, only experiences a BGP routing change and not an egress point change. C continues to use egress router A to reach p, though using a different route.

The flow of traffic in the network also depends on the incoming traffic. We describe the traffic load that enters the network for each destination prefix as the *traffic demand*, which is a matrix V, where each element $V(i, p)$ represents the volume of traffic entering at ingress router i and headed toward a destination prefix p. Operators usually represent the traffic as a *traffic matrix*, which captures the load from each ingress point i to each egress point e ($\mathcal{T}M(i, e)$). The traffic matrix is a useful abstraction, because by combining it with the intradomain forwarding paths, operators can determine the load in each link in the network. The traffic matrix is essentially the composition of the *traffic demands* and the *egress-point selection*. In particular, each traffic-matrix element $\mathcal{T}M(i, e)$ represents the traffic from ingress point i aggregated over all destinations p reached through egress point e; that is, $\mathcal{T}M(i, e) = \sum_p\{V(i, p) \mid b(i, p) = e\}$. In recent years, numerous studies have proposed techniques for measuring the traffic demands [7] and the traffic matrix [8, 9, 10, 11].

Fluctuations in the traffic demands impact the traffic entering the network and egress-point changes impact where traffic leaves. Therefore, both kinds of changes have a direct impact on the traffic matrix. The next section discusses which changes are responsible for the largest variations in the traffic matrix.

3. MEASURING ROUTING CHANGES

Several measurement studies have explored the effects of BGP routing changes on the flow of traffic through the Sprint and AT&T backbones [1, 2, 3, 4, 5, 6]. The researchers joined continuous

feeds of BGP update messages from the operational routers with traffic measurements collected using Cisco's Netflow feature (for both AT&T and Sprint) or a custom packet monitor (for Sprint). Although the results of each study in isolation may lead to contradictory conclusions, collectively, the papers show that: while most BGP routing changes have little influence on the traffic, a small fraction of routing changes have a large impact. The studies show that local events, such as IGP topology changes and eBGP session resets, are responsible for the largest traffic shifts. This section presents an overview of these results and a summary of the lessons learned about measurement methodology.

3.1 Impact of BGP Changes on Traffic

Each border router in a large ISP backbone changes its routing decisions around $200,000$ times a day, due to equipment failures, policy changes, and BGP updates received from neighboring ASes. However, the vast majority of these routing changes have little, if any, impact on the flow of traffic through the ISP backbone; that is, most Δr do not imply a significant change in $\Delta \mathcal{T} M$. First, most BGP routing changes (Δr) affect a small number of destination prefixes that typically receive very little traffic [1, 2]. In contrast, popular destination prefixes have stable BGP routes for days or weeks at a time [1]. Second, the majority of BGP routing changes occur in remote ASes and do not impact the egress point for most traffic [2, 6]. For example, the border router may receive a BGP advertisement that reflects a change several AS hops away. These kinds of changes typically do not cause any routers in the ISP backbone to change how they forward traffic; that is, most Δr do not imply a Δb.

In addition, large fluctuations in the traffic matrix are relatively rare. This is not surprising because large ISP backbones carry significant volumes of highly aggregated traffic. However, some traffic matrix elements vary by a significant amount (e.g., more than four times their normal variations) several times a week [4]. These traffic variations can have many causes, including flash crowds, denial-of-service attacks, and routing changes in other ASes. Interestingly, BGP routing changes seen by the ISP are responsible for the *largest* of these variations in the traffic matrix [4]. When large BGP egress shifts happen (Δb), they cause a correspondingly large traffic shift ($\Delta \mathcal{T} M$) [4, 6]. So, a small fraction of the BGP routing changes have a very significant impact on the traffic matrix, even though the vast majority of BGP routing changes do not.

3.2 Causes of Large Traffic Shifts

The BGP egress selection $b(i, p)$ may change because of routing changes in other ASes or local events in the ISP network. The measurement studies found that the largest traffic shifts stemmed from events occurring at the peering points with neighboring ASes or inside the ISP [4, 5, 6]. First, the failure of an eBGP session, due to link failure or planned maintenance, can cause routers i throughout the AS to change egress points $b(i, p)$ for many destination prefixes p; every router must pick a new BGP route and direct traffic to the next closest egress point. The failure and recovery of eBGP sessions to large peers, such as other tier-1 providers, tend to cause very large shifts in traffic [6]. Other external BGP routing changes, such as routing changes in downstream ASes tend not to have as much influence, since they usually affect the egress set E_p for much fewer destination prefixes.

The second source of large traffic shifts is BGP routing changes induced by changes in the underlying IGP topology, due to hot-potato routing [3, 4, 5]. We call these changes *hot-potato routing changes*. That is, a change in the IGP distances $d(i, j)$ can cause multiple routers i to switch to different egress points $b(i, p)$ for reaching one or more destination prefixes p. For example, the failure of the link C-D in Figure 1 would increase the IGP distance $d(C, A)$ from 9 to 11. Even though the BGP route through A is still available, the IGP change would lead C to select the route through egress point B, with a distance $d(C, B)$ of 10. These kinds of IGP topology changes can occur for several reasons, including equipment failures, planned maintenance, and traffic engineering. For example, the network operator may change the metric of the C-D link to 9 to reduce load on this link, triggering an inadvertent change in C's choice of egress point.

The likelihood of hot-potato routing changes varies significantly, depending on the ISP's peering policies and IGP topology. The placement of peering points with neighboring ASes plays a significant role [5]. For example, the network in Figure 1 would not experience hot-potato routing changes if C had its own direct connection to the neighbor AS. The network topology and IGP metric settings matter as well [3, 5]. In Figure 1, C is nearly equi-distant from two egress points, allowing small changes in the IGP topology to have a large influence on the choice of egress point. Operational practices, such as traffic engineering and planned maintenance, also influence the likelihood of hot-potato routing changes. If network operators tune link metrics based only on the intradomain topology and the traffic matrix, they may inadvertently select metric settings that cause large perturbations in the traffic matrix [5]. Instead, network operators can apply network modeling tools that consider the traffic demands $V(i, p)$ and the egress sets E_p to predict how IGP topology changes would affect the flow of traffic [5, 12, 13, 14]. Ultimately, the likelihood of large hot-potato routing changes depends on whether an AS is designed and operated with these issues in mind, and whether the AS has multiple egress points for reaching a large number of external destination prefixes.

3.3 Measurement Lessons for Future Work

Across all six studies which span two major backbone networks measured at different times in different ways, we can contrast and learn how different measurement methodologies impacted the scope of the findings. This in turn allows us to provide guidelines for future work in this area.

Measure multiple networks with different designs and policies: There is clearly a large variation in the design of networks. For instance, the studies of the AT&T network considered AS 7018, which primarily covers the U.S., while the Sprint studies considered AS 1239, which includes North America, Europe, and parts of Asia and South America. Thus the internal structure of the two ASes is quite different—one includes many more inter-continental links than the other. The difference in the range of IGP metrics impacts the extent of hot-potato routing changes—for example, link metric changes in the European part of the Sprint topology only caused hot-potato changes in traffic to egress points in Europe and the east coast of the U.S. [5]. Similarly, we have found that the path diversity of the network topology, the locations of peering points, the setting of local preferences for certain peering points, and export policies in neighboring ASes all impact the extent of hot-potato changes. While we did not consider "tier-2" and "tier-3" ISPs, we speculate that they would have fewer neighbors, fewer peering points, and less aggregation of traffic, all of which might reduce the significance of hot-potato routing changes relative to the natural statistical fluctuations in the traffic. Any work that measures routing and traffic dynamics should consider how the local AS design and policies impact the results, and any work that emulates these findings should also model different network designs.

Measure at multiple vantage points: Some routers in a network experience very different behavior than others. In particular,

some ingress points may be much more susceptible to internal BGP routing changes than others [3], making analysis of routing stability very dependent on where data are collected. The studies in [1, 2] came to the conclusion that BGP routing changes do not have a significant impact on the flow of traffic. However, these studies focused on BGP data collected from few routers. For instance, the analysis in [1] studies routers in two large cities with rich connectivity to other large ISPs. An analysis over a wider range of routers in the same network showed that some locations experience hot-potato routing changes that affect the BGP routing decisions for many popular destinations [3]. The wide variation across vantage points makes it difficult to rely on publicly-available BGP update logs, such as the RouteViews and RIPE-NCC data sets, since they typically include data collected from just one, or at most two, routers in each AS. A BGP feed from a different router in the same AS might look quite different [15].

Measure for long periods of time over different network conditions: The traffic demands and network topology vary over time, due to diurnal changes in load and operational activities such as traffic engineering and planned maintenance. Collecting and analyzing measurement data over a long period of time is important for capturing the full range of network conditions. For example, many changes in the IGP topology occur during planned maintenance [16], when the operators add, remove, and upgrade equipment in the network. In addition, operators sometimes need to tune the link metrics in response to heavy traffic loads or in anticipation of large maintenance activities, but these do not happen very frequently. The study in [2] analyzed time periods that did not include maintenance activities and, as such, did not see many hot-potato routing changes. In contrast, the study in [5] of the same network illustrates that IGP topology changes would cause large shifts in traffic. Similarly, in contrast to [1], the studies in [3, 4]—covering a period of several months—show that IGP topology changes can trigger significant changes in the traffic matrix.

4. CONCLUSIONS

Network design and operational practices in ISP backbones have a significant influence on the flow of traffic through the Internet. Understanding how routing protocols are used in practice, and how network events lead to routing changes and traffic shifts, is crucial for creating accurate models of Internet routing and interpreting measurement data. In this paper, we consolidate, and reconcile, the findings of several measurement studies that had an "inside view" of commercial ISPs, with the goal of identifying key phenomena affecting the flow of traffic. The studies collectively show that, while most BGP routing changes have little influence on the flow of traffic, a small number of routing changes have very significant impact. Hot-potato routing changes and eBGP session resets are responsible for the most significant traffic shifts.

It may be surprising that with our privileged access to data from two major backbone networks, some of our previous studies in isolation led to seemingly contradictory conclusions. These contradictions illustrate the complexity of interpreting the large volume of data collected at each one of these networks to create a clear picture of network-wide behavior. Before we, as a community, can understand the behavior of a system with global scope like the Internet, we need both access to more detailed data from each individual network and more efficient tools for mining this data.

Ultimately, future research studies should evaluate the complex interplay between network topology, routing configuration, offered traffic, and network events in a controlled fashion. The creation of accurate models of the underlying network events (due to failures, planned maintenance, and traffic engineering) would be an important step in that direction. Measurements of operational networks are very useful for creating these models. Once in place, these models can be used by a much wider community to create better routing protocols, network architectures, and operational practices. Also, future measurement studies should consider the performance impact of routing changes on user applications, both due to transient disruptions during routing-protocol convergence and the longer-term changes in path properties.

5. REFERENCES

[1] J. Rexford, J. Wang, Z. Xiao, and Y. Zhang, "BGP routing stability of popular destinations," in *Proc. Internet Measurement Workshop*, November 2002.

[2] S. Agarwal, C.-N. Chuah, S. Bhattacharyya, and C. Diot, "Impact of BGP dynamics on intra-domain traffic," in *Proc. ACM SIGMETRICS*, June 2004.

[3] R. Teixeira, A. Shaikh, T. Griffin, and J. Rexford, "Dynamics of hot-potato routing in IP networks," in *Proc. ACM SIGMETRICS*, June 2004.

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

[5] 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.

[6] 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. Networked System Design and Implementation*, May 2005.

[7] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True, "Deriving traffic demands for operational IP networks: Methodology and experience," *IEEE/ACM Trans. on Networking*, June 2001.

[8] J. Cao, D. Davis, S. V. Wiel, and B. Yu, "Time-varying network tomography," *J. American Statistical Association*, December 2000.

[9] A. Medina, N. Taft, K. Salamatian, S. Bhattacharyya, and C. Diot, "Traffic matrix estimation: Existing techniques and new directions," in *Proc. ACM SIGCOMM*, August 2002.

[10] Y. Zhang, M. Roughan, N. Duffield, and A. Greenberg, "Fast, accurate computation of large-scale IP traffic matrices from link loads," in *Proc. ACM SIGMETRICS*, June 2003.

[11] Y. Zhang, M. Roughan, C. Lund, and D. Donoho, "An information-theoretic approach to traffic matrix estimation," in *Proc. ACM SIGCOMM*, August 2003.

[12] R. Teixeira, T. Griffin, A. Shaikh, and G. Voelker, "Network sensitivity to hot-potato disruptions," in *Proc. ACM SIGCOMM*, August 2004.

[13] N. Feamster, J. Winick, and J. Rexford, "A model of BGP routing for network engineering," in *Proc. ACM SIGMETRICS*, June 2004.

[14] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, and J. Rexford, "NetScope: Traffic engineering for IP networks," *IEEE Network Magazine*, pp. 11–19, March 2000.

[15] R. Teixeira and J. Rexford, "A measurement framework for pin-pointing routing changes," in *Proc. ACM SIGCOMM Network Troubleshooting Workshop*, September 2004.

[16] A. Markopoulou, G.Iannaccone, S.Bhattacharyya, C.N.Chuah, and C.Diot, "Characterization of failures in an IP backbone network," in *Proc. IEEE INFOCOM*, March 2004.