Impact Analysis of BGP Sessions for Prioritization of Maintenance Operations

Sihyung Lee, Kyriaki Levanti, Hyong S. Kim

October 8, 2010

CMU-CyLab-10-018

CyLab
Carnegie Mellon University
Pittsburgh, PA 15213

Impact Analysis of BGP Sessions for Prioritization of Maintenance Operations

Sihyung Lee
IBM T.J. Watson Research Center, Hawthorne, NY
leesi@us.ibm.com

Kyriaki Levanti, Hyong S. Kim
Carnegie Mellon University, Pittsburgh, PA
klevanti@cmu.edu, kim@ece.cmu.edu

***Abstract*—Network operators in large-scale networks are often faced with long lists of maintenance tasks and find it difficult to track the relative importance of these tasks, without knowing their impact on the network's operation. As a result, operators may react slowly to critical tasks, increasing network downtime and maintenance costs. We present a system that quantifies the impact of maintenance tasks so that operators can prioritize their reaction according to the estimated impact (i.e., spend more time and effort on avoiding the disruption caused by high-impact maintenance tasks). In particular, the proposed system estimates the amount of traffic loss due to maintenance operations on inter-domain routing sessions, one of the most frequently modified aspects of network configurations. We implement the proposed system and apply it to 372 routing sessions in a nation-wide ISP network. The system identifies sessions with a varying degree of impact: sessions with nearly zero data loss, as well as sessions that can result in more than 1,000 GB of data loss if disrupted without any protection mechanism applied. We also show that predicting the amount of data loss is not straightforward since this amount changes over time, often in unexpected ways (e.g., from 50GB to 0 over one-month period). Therefore, the proposed impact analysis system is necessary for network operators to perform periodic audits of the routing sessions' impact and to classify the sessions according to the projected data losses. Operators can then decide the level of protection for each session (e.g., employ more effective and costly methods to protect critical sessions) and thus allocate maintenance costs more efficiently.**

I. INTRODUCTION

When a limited number of operators manage a large set of maintenance tasks, the operators often end up focusing their time on lower-priority issues and pay less attention to more important tasks [1]. As a result, operator errors often lead to critical problems, such as complete loss of connectivity [2][3], and companies spend 80% of IT budgets on maintenance [4]. Knowing the importance of the various network elements (e.g., a routers, interfaces, configurations, and routes) and then identifying the critical parts of a network would allow network operators to prioritize their maintenance tasks. In this manner, operators can create a more efficient plan in terms of (i) the order to perform the tasks, (ii) the amount of time to spend on each task, and (iii) the level of response (i.e., how carefully and with what implementation cost a configuration task should be performed).

To measure the importance of network elements, we propose a system that evaluates the *impact* of a network element; that is the amount of potential negative impact that would be incurred in the event that the element does not operate correctly. Impact can be defined for various network elements and contexts. In this paper, we estimate the impact of shutting down an inter-domain routing session (BGP session). Shutting down a BGP session is a common maintenance activity performed on a daily basis for updating router software, modifying routing policies, and upgrading links [5][6]. This activity causes a transient loss of connectivity until routing converges to alternative routes. Although this loss of connectivity is a temporary event, it can be long-lasting and lead to huge amounts of data loss (e.g., several terabytes in a few seconds). The proposed system estimates the amount of data loss so that network operators can choose cost-effective protection mechanisms to reduce this loss. For example, an expensive but powerful solution should be used to guard a session with a large impact (e.g., setting up a backup MPLS tunnel), while no solution is necessary if a session has nearly zero impact. Thus, this prioritization reduces maintenance costs while improving QoS experienced by end users.

To the best of our knowledge, this paper is the first to consider prioritizing network configuration changes by estimating their impact (Section II). The impact estimation system first simulates the progressive change of routes that follows the shutdown of a BGP session. While tracing these changes, the system accumulates the amount of potential data loss at each router by considering whether the router experiences connectivity loss and how much traffic the router forwards. At the end of the simulation – when routing converges, the accumulated data loss represents the network-wide impact of the session (Section III). The proposed system can be applied to various maintenance scenarios other than shutting down a BGP session, such as link upgrades and router configuration changes (Section III.E). We implement the proposed system and apply it to 372 routing sessions of a nation-wide ISP network. This system identifies a range of routing sessions that can be prioritized for protection: from sessions with nearly zero impact to sessions that may result in more than a terabyte of data loss within a few seconds if not properly protected. Although the impact of several sessions are self-evident (e.g., the large impact of sessions that are known to provide a large number of routes), we show that manually predicting the impact of the rest of the sessions is difficult without the help of the proposed system (Sections IV.A-C). Finally, we provide insights into the causes of the large impact

of routing sessions and suggest routing design guidelines for reducing the criticality of these sessions (Section IV.D).

II. Related Work

A. Impact Estimation and Prioritization

The idea of prioritizing tasks according to estimated impact has appeared in the context of security but not in the context of routing. SecureRank [7] determines the order in which to apply security patches to network nodes by assigning vulnerability scores to these nodes. For example, a highly connected node with a number of high-risk vulnerabilities could be the first to patch.

Several works analyze the impact of changes in routes, route filtering policies, and topologies so that network operators can test what-if scenarios for pending changes. These works focus on the final outcome of the route decision process (e.g., new best routes *after routing converges*), whereas our proposed system estimates transient properties (e.g., the gradual change of routes and data loss *during route convergence*). Netscope [8] and Teixeira's work on hot-potato routing [9] analyze route changes and associated traffic shifts when IGP configurations are modified. Feamster's model of BGP [10] predicts new egress routers when BGP configurations and routes change. All these previous works show route changes after route convergence is finalized. Our system analyzes the transient behavior – how routes progress toward the final outcomes until routing converges – and then computes the amount of data loss that would result from these progressive changes.

B. Solutions to Reducing Data Loss During Route Convergence

Network links and nodes may fail either accidentally or during maintenance, leading to data loss until routing converges to alternative routes. To reduce this data loss, diverse solutions exist that differ in cost and effectiveness. Network operators can choose the most cost-efficient solution, according to the results of our impact analysis system. We classify existing solutions into three groups. The first group is more costly than the other two groups, but it is more effective in reducing data loss. The second and third groups are less effective and less expensive solutions than the first group. These two latter groups differ in the approaches they follow.

The first group of solutions is proactive – they pre-configure bandwidth-guaranteed backup paths so that traffic can be immediately diverted onto these backup paths in event of failures. Although data loss can be reduced to nearly zero, these solutions either require complex configurations of IP/MPLS tunnels and/or modifications to existing hardware and software [11][12][5]. The other two groups aim to minimize data loss while allowing a short period of transient connectivity loss. The second group of solutions gradually migrates routes to alternative routes such that connectivity disruption is minimized [5][6][13][14]. This is done through progressive changes of link weights. The third group reduces route convergence time by (i) optimizing routing-protocol timer values [15], (ii) adding several more routing sessions [16], or (iii) prioritizing the processing of route updates according to their importance [17][18]. Although the last two groups are less costly than the first group, the two groups are not inexpensive, and require tedious scheduling and configurations (e.g., scheduling extra hours of maintenance for when traffic level is low, changing link weights gradually over multiple steps, and waiting after each step to monitor whether routing converges).

III. Impact Estimation Method

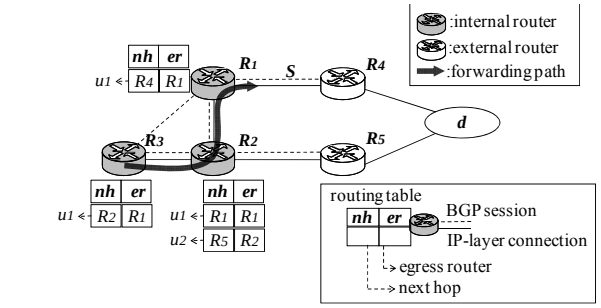
We first illustrate how shutting down a BGP session incurs data loss (Section III.A). We then describe details of our impact-estimation method (Sections III.B-D) and its applications (Section III.E). Our descriptions throughout this section regard a single destination d, and thus the terms "connectivity", "route", and "traffic" refer to those toward d. We explain how to extend the method to multiple destinations in Section III.D. The terms "BGP sessions", "routing sessions," and "sessions" are used interchangeably.

A. Transient Loss of Connectivity Caused by Maintenance Events

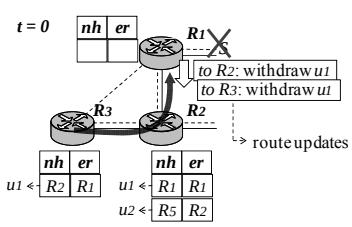
Fig. 1(a) through (f) illustrate gradual changes of routes when a BGP session is shut down. Fig. 1(a) shows how traffic is routed before this session is down. The three shaded routers represent internal routers in our network (R_1, R_2, and R_3), the two unshaded routers represent those in external networks (R_4 and R_5), and d represents the destination network. A solid line between two routers indicates that these two routers are connected at the IP layer. A dashed line represents a BGP session, which runs over the IP layer. For example, R_1 and R_3 maintain a BGP session, and this session runs over the path at the IP layer, (R_1 , R_2 , R_3). The session S between R_1 and R_4 is going to be shut down for maintenance. The tables next to the routers are routing tables; each row represents a route, and *nh* is the next hop in the forwarding path toward the egress router *er*. For example, R_3 will forward traffic to $nh(R_2)$, and this traffic will eventually exit the network through $er(R_1)$. This forwarding path is also shown in the sold black arrow. If multiple routes exist in a routing table, these routes appear in the order of preference. For example, in R_2, the first route (the route via R_1) is the best and the second route (the route via R_5) is an alternative.

Initially, before S is shut down (Fig. 1(a)), the external neighbor R_4 advertises a route u_1 to R_1, and this route is chosen as the best by all three internal routers. Later, the other external neighbor R_5 also advertises a route u_2 to R_2, but R_2 continues to choose u_1 as the best; therefore, u_2 is not advertised further to other routers and remains an alternative route only in R_2.

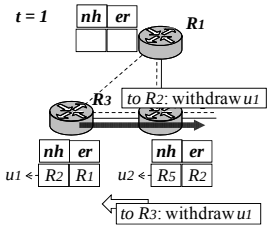
At time t=0 (Fig. 1(b)), session S is shut down. As soon as R_1 detects this failure, it withdraws u_1 from the routing table, and announces this withdrawal to R_2 and R_3. Until (i) all three internal routers are notified of this withdrawal and (ii) begin to use the alternative route u_2 (Fig. 1(f)), packets will not be correctly forwarded to u_2 and will be dropped. For example, in Fig. 1(b), packets from R_3 will continue to be delivered to R_1, which will then be dropped since R_1 has no route.



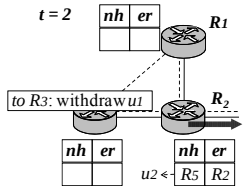
(a) Initial configuration and routing table status. u_1 and u_2 are routes from R_4 and R_5, respectively.



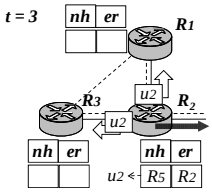
(b) Session S – (R_1, R_4) – is down.



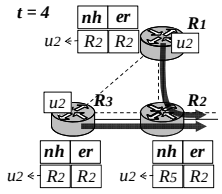
(c) R_2 receives a withdrawal.



(d) R_3 receives a withdrawal.



(e) R_2 sends updates.



(f) R_1 and R_3 receive updates from R_2.

Figure 1. Transient Loss of Connectivity Caused by a Session Down

B. *Complexity of Estimating Data Loss in Transient Loss of Connectivity*

Estimating the amount of data loss is not immediately apparent since it requires consideration of many different factors: (i) the failure location; (ii) the existence and initial point of advertisement of backup routes; (iii) the network topologies at both the IP and BGP layers; and (iv) the amount of traffic that each router is forwarding to the affected destination. To illustrate this complexity, we analyze the router states, as shown in Fig. 1(b) to (f).

- t=0 (Fig. 1(b)): S goes down, and R_1 sends two withdrawals of u_1, one to R_2 and the other to R_3. Until these two withdrawals are received by R_2 and R_3, packets from these two routers continue to be forwarded to R_1, which then drops these packets since it has no route.
- t=1 (Fig. 1(c)): the withdrawal message arrives at R_2, and R_2 regains connectivity by switching to its alternative route, u_2. The withdrawal for R3 is still in transit because the IP-layer path $R_1 \rightarrow R_3$ is longer than the path $R_1 \rightarrow R_2$. Note that R_3 still uses the invalidated route u_1, but its packets are successfully forwarded via R_2; this is because (i) R_3's next hop is set to R_2, and (ii) R_2 happens to have withdrawn u_1 and correctly chosen u_2.
- t=2 (Fig. 1(d)): the withdrawal message arrives at R_3, and R_3 loses connectivity since it has no route. Note that R_3 initially loses its connectivity at t=0, temporarily gains connectivity at t=1, and then loses its connectivity again at t=2. This is due to the topologies and locations alternative routes.
- t=3 (Fig. 1(e)): R_2 announces its new best route u_2 to R_1 and R_3. This announcement may have been sent at t=1 or t=2 – a BGP-speaking router maintains a timer, and when this timer expires, the router sends route updates. Therefore, depending on the current setting of the expiration time, the updates could have been sent earlier.
- t=4 (Fig. 1(f)): R_1 and R_3 finally regain connectivity as they receive the alternative route u_2. Although R_1 and R_3 gain connectivity later than R_2, R_1 and R_3 do not necessarily lose more data. The amount of data loss at router R_i depends on the amount of traffic that R_i has to forward.

This example illustrates how the multiple factors complicate the process of impact estimation. The proposed system estimates impact by accurately considering all of these complicated factors, as shown in the following sections. In the evaluation (Section IV.C), we show that the combination of the different factors changes over time, further complicating the impact estimation. For example, the distribution of alternative routes heavily depends on the set of routes advertised by external neighbors, which may change as new peers are added, routing policies are modified, and neighbors advertise different sets of routes.

C. *Impact Estimation System*

The previous section demonstrates the complexity and dynamicity of BGP's transient behavior. In order to precisely trace this complex behavior, and thus to estimate impact, we simulate the network-wide changes in routing tables. Fig. 2 illustrates the components of our impact estimation system.

The estimation requires a set of input data: a snapshot of the network's topology, the link metrics, the routing states of each router, and traffic demands. We describe how to obtain these data in Section IV.A. We first simulate shutting down a target session S, and then we generate a sequence of routing table snapshots. This sequence represents the subsequent route changes toward alternative routes (e.g., Fig. 1(a) through (f)). A new snapshot is generated each time any routing entry (toward

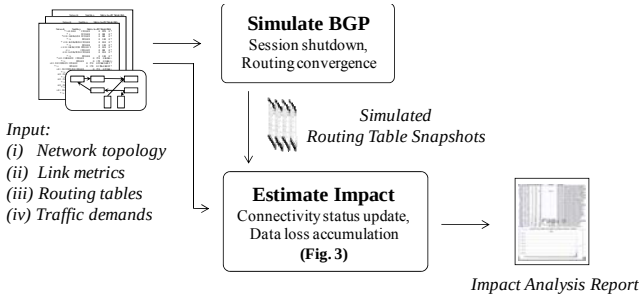


Figure 2. Overview of Impact Estimation System

the destination) changes in routing tables. According to each of these snapshots, we continuously update the connectivity status of each router (i.e., *connected*: a forwarding path to *d* exists, or *not_connected*) and also accumulate the amount of data loss if a router is *not_connected*. At the end of the simulation (when the routing converges), the accumulated data loss represents the impact of S, $I(S)$.

Fig. 3 sketches our algorithm for estimating $I(S)$ given a sequence of routing table snapshots. For two consecutive snapshots at t_0 and t_1, if a router R_i loses connectivity at t_0, then we increase $I(S)$ by the amount of data that R_i has to forward during the period between t_0 and t_1 (lines 09-15). We determine the connectivity of R_i as follows. R_i obtains connectivity if (i) R_i chooses a route that is received directly from one of R_i's external neighbors (lines 23-24), or (ii) R_i chooses a route leading to an egress router that satisfies condition (i) (lines 25-26). Otherwise, R_i loses connectivity. The condition (ii) can be evaluated by examining whether the *next hop* of the selected route obtains connectivity, rather than examining the connectivity of every hop en-route to the egress router. This is because the *do* loop (lines 19-29) and the definitions of (i) and (ii) make sure that the next hop is *connected* only if every hop on the forwarding path to the egress router is *connected*. For example, in Fig. 1(a), R_1 is set to *connected* according to the condition (i); then, R_2 is set to *connected* according to the condition (ii) since the next hop R_1 is *connected*; finally, R_3 is set to *connected* according to (ii) since the next hop R_2 is connected, which implies that R_2's next hop R_1 is also connected. Note that the connectivity is determined starting from the *connected* egress router, backward on a forwarding path, i.e., R_1, R_2, and then R_3.

Table I presents iterations of the algorithm in Fig. 3 given a series of routing table snapshots in Fig. 1. Each row represents one snapshot. The snapshot at t=3 (Fig. 1(e)) is not included because the routing tables are equivalent to those in the previous snapshot at t=2. The first column lists the time when each snapshot is generated. The next three columns show the status of the three internal routers – whether the routers have connectivity or not. The last column shows the amount of data loss accumulated at each snapshot. We assume that the neighbors of R_1, R_2, and R_3 send traffic at constant rates of 100Mbps, 500Mbps, and 200Mbps, respectively. Upon each iteration, if a router is *not_connected*, traffic from this router will be dropped and thus is added to $I(S)$.

Traffic(Nij→Ri): average traffic amount that neighbor *Nij* sends to *Ri* (in bps)
Ri.best_route: route that *Ri* selects as best

```
00 I(S) = 0;                                  // impact of S: sum of data loss (in bits)
01 prev_time = 0;
02
03 // This function is executed for each routing table snapshot at t.
04 Estimate_Loss_of_Data (time t) {
05     Calculate_Loss_of_Data(t);      // calculate data loss for previous period
06     Update_Connectivity_Status(); // update routers' connectivity status
07 }
08
09 Calculate_Loss_of_Data (time t) {
10     for each router Ri
11       if (Ri.state == not_connected)
12           for each neighbor Nij of Ri
13               I(S) += Traffic(Nij→Ri) × (t – prev_time);
14     prev_time = t;
15 }
16
17 Update_Connectivity_Status() {
18     initialize Ri.state = not_connected, ∀i;
19     do
20       for each router Ri
21         if (Ri.best_route == ∅) Ri.state = not_connected; // Ri has no route
22           else
23             if (Ri.best_route.egress == Ri)                  // egress is itself
24                   Ri.state = connected;
25             else if (Ri.best_route.next_hop.state == connected)
26                   Ri.state = connected;
27             else
28                   Ri.state = not_connected;
29     until (no changes in Ri.state, ∀i)
30 }
```

Figure 3. Estimation of Data Loss When a Session *S* is Down

D. Additional Considerations

Multiple Destinations: The algorithm in Fig. 3 estimates $I(S)$ for a single destination d. When we receive routes toward multiple destinations over S, the algorithm runs for each of these destinations. Note that the impact estimation considers the withdrawal of only the routes that are selected as the best; a transient loss of connectivity happens only when the best routes change. When a number of best routes exist, running the algorithm for every route may become computationally expensive. In this case, we selectively run the algorithm for destinations that draw significant amounts of traffic, since a small fraction of destinations account for nearly 90% of all traffic [19][10].

Weighted Impact Estimation: Operators can assign different weights to different types of traffic when the protection of particular types of traffic is more significant according to SLAs (e.g., assign heavy weights to real-time traffic, such as VoIP and video-conferencing). This can be accomplished by formulating the function $Traffic(N_{ij} \rightarrow R_i)$ in the algorithm such that a traffic type determines the weight to apply.

Identification of Termination Time: The estimation method in Section III.C requires that the simulation automatically terminates when routing converges to alternative routes. To identify this termination time, we examine whether every router has selected alternative routes after the routes from S are withdrawn, and we evaluate this condition for each snapshot. These alternative routes can be determined according to the

TABLE I. APPLICATION OF ALGORITHM IN FIG. 3 ON EXAMPLE IN FIG. 1

Time (sec)	*R1* (100 Mbps)	*R2* (500 Mbps)	*R3* (200 Mbps)	$I(S)$: Impact of Session S (bits)
t=0	*not_ connected*	*not_ connected*	*not_ connected*	$I(S) = 0$
t=1	*not_ connected*	*connected*	*connected*	$I(S)$+= (100 + 500 + 200) × (1 - 0)
t=2	*not_ connected*	*connected*	*not_ connected*	$I(S)$ += (100) × (2 - 1)
t=4	*connected*	*connected*	*connected*	$I(S)$ += (100+200)×(4 - 2)
				$I(S) = 1500$

BGP decision process [20] (i.e., by comparing routes across a set of route attributes and by then choosing the one with the most desirable attributes). We can also determine alternative routes by using other prediction methods [10].

E. Applications of Impact Estimation

In this section, we investigate various maintenance scenarios and network elements in which the proposed impact estimation can be used.

In addition to estimating the impact of one particular session, which needs to be shut down for a maintenance activity, we can also perform the impact analysis for all sessions in a network. We can then identify a subset of sessions with a large impact and protect this subset in case the sessions are accidentally dropped. Accidental link failures happen as frequently as those due to planned maintenance. For example, [12] found more than 9K failures across 47 BGP peering links over a three-month period.

The proposed impact estimation can also be applied to network elements other than routing sessions, as long as maintenance actions on these elements cause route changes and thus data loss (e.g., routers, links, routes, and routing policy configurations). For example, operators often modify routing policies, and modification of a routing policy may withdraw a small set of routes U from a session S. The impact of this activity can be measured by withdrawing U in the simulation, rather than withdrawing all routes from S. We can then trace subsequent route changes and estimate data loss. Network configurations related to routing are shown to comprise a major portion (i.e., up to 70%) of network configurations, and they are modified the most frequently [21].

IV. APPLYING IMPACT ESTIMATION TO AN ISP NETWORK

To demonstrate the impact estimation method, we implement the proposed estimation system and analyze the impact of 372 BGP sessions of a nation-wide ISP network. First, we provide a brief explanation of the ISP network and how we obtain the input parameters for impact estimation. Then, we describe the experimental setup and the implementation of the impact estimation system. Lastly, we present the estimated impact and suggest various methods that reduce the impact of shutting down a routing session on data forwarding.

A. Obtaining Inputs for Impact Analysis

The ISP network is a major carrier in Europe, and it maintains 372 sessions with carriers in Africa, Asia, Europe, and South America. This network's average traffic demands per session range from a few Mbps to tens of Gbps. We focus on one snapshot of the network's routing tables and traffic demands collected on February 2, 2008. We also analyze daily snapshots over a one-month period (February 2 to March 2, 2008) to measure the degree of changes.

The ISP routinely collects route and traffic measurements for network management purposes. We leverage this data to extract the input parameters for our impact estimation system.

- **BGP routing tables** refer to the Loc-RIB as defined in [20], which includes all the working routes received from BGP-speaking neighbors. This is collected once a day in the ISP network. This network maintains several BGP monitors that establish internal BGP sessions to the network's routers and archive all BGP updates received from these routers. The monitors then dump their routing tables once a day in order to provide a periodic snapshot of the best and alternative routes.
- Daily snapshots of **traffic demands** are estimated from load statistics. These statistics are readily available to network operators [22][23]. The traffic demands represent the rate of outgoing traffic over each session, averaged over a 24-hour period. Using these average values, this analysis assumes a constant rate of traffic for each session. We also tested the system with a time-varying rate (e.g., by assuming a normal distribution and inferring instant traffic demands for each snapshot), but we did not see significant differences in the results.
- **The IP and BGP network topologies** and **link metrics** are derived from a snapshot of routers' IGP and BGP configurations. The ISP network collects configuration snapshots once a day in order to keep reference of the configuration changes.

B. Experimental Setup and Implementation

Fig. 4 illustrates our implementation of the impact estimation system. We parse input data and store parsed data in a MySQL database. The configuration parser is developed for Cisco IOS and Juniper JUNOS commands [24]. We then simulate shutting down BGP sessions and estimate impact values according to the algorithm in Fig. 3. This simulation is based on SSFNet [25], a network simulator widely used for studying transient behavior of BGP (e.g., [26][15][17]). We run a total of 30 simulations (for each daily snapshot) and compute the average amount of data loss over these 30 runs.

To precisely simulate the behavior of BGP, the proposed impact estimation system requires configuration of several parameters, as shown in Table II. The first two parameters, MRAI and WRATE, can be extracted from network configurations. The last two parameters, BGP update processing time and link delay, cannot be extracted from the configurations. For these two parameters, we present a range of

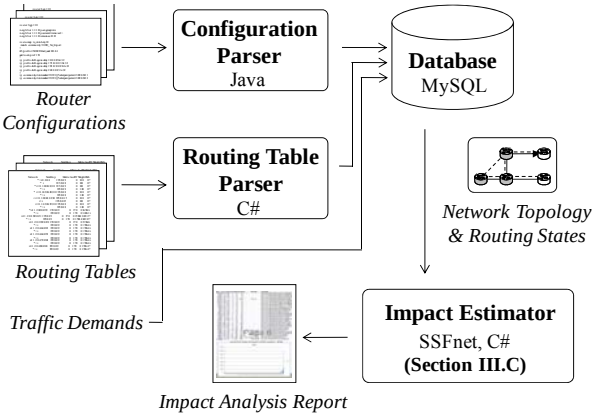


Figure 4. Implementation of Impact Estimation System

values that are shown to be valid in empirical studies [15][27] and thus are commonly used in evaluations [17][26][28].

We assume that routes do not change dramatically within the short period of time of route convergence – from when a session is shut down ($t = 0$) until the routing converges ($t = t_c$). This assumption is necessary since input to the impact analysis system is a single snapshot of routing tables at t=0; the subsequent snapshots ($0< t \leq t_c$) are generated by our system as we simulate the routing convergence, assuming the snapshot at t=0 has not changed. In the evaluation, we found that the routing typically converges in less than 20 seconds in this network, and it is unlikely that routes change significantly within 20 seconds, invalidating the estimated impact values. [29] shows that routes that draw the most traffic (and thus contribute the most to the estimated impact values) are stable for days or weeks.

Since we simulate the behavior of BGP, the impact values estimated by the proposed system may not be exactly equal to those in the real network. The real values also depend on the peculiarities of routers' hardware and software. However, the purpose of this estimation is to measure an approximate amount of data loss and then to evaluate the relative impact of shutting down different BGP sessions. We can then identify a subset of sessions with a relatively large impact and give priority to protecting these sessions.

C. Impact Analysis Results of an ISP Network

Fig. 5 presents the distribution of impact values that we estimated over the 372 BGP sessions in the network on a single snapshot. The bars on the horizontal axis represent the 372 sessions, and the height of each bar represents the amount of data loss in GB during route convergence. These 372 sessions are ordered according to their impact values. The bar on the far right represents the session with the highest impact, which is greater than 2,000 GB.

First, we observe an opportunity to classify sessions according to their estimated impact such that we can prioritize the level of responses to each class. More than half of the sessions (~55.9%) have zero impact, but a substantial fraction of the sessions, i.e., nearly 45 sessions (~12.1%), have an impact larger than 10 GB of data loss. If we take a closer look at these 45 large-impact sessions, multiple sharp inclines exist (e.g., near the 360th session, the impact of a session goes beyond 50GB), and the peak reaches even above 2 TB; if this session is shut down without any protection mechanism, more than 2TB of data can be lost within a few seconds. This result indicates that we can classify the sessions according to their impact values, and then apply different levels of protection according to this classification. For example, we can employ expensive but powerful protection mechanisms for sessions with a large impact and employ no protection for sessions with zero impact. This kind of prioritization can significantly reduce maintenance costs, compared to extreme protection policies (i.e., protection of all sessions, or protection of no sessions). These two extremes in session protection result in either excessive maintenance costs or a large amount of data loss.

TABLE II. PARAMETERS REQUIRED FOR IMPACT ESTIMATION

Parameter	Description	Value
Minimum Route Advertisement Interval (MRAI)	Minimum interval to be elapsed before a new update is sent. One MRAI timer is maintained for each destination.	Found in configurations. Otherwise, default values are used [15]: Cisco routers – 30 sec for eBGP sessions 5 sec for iBGP sessions Juniper routers – 0 sec.
Withdrawal Rate Limiting (WRATE)	If WRATE is turned off, withdrawals are sent immediately without waiting for MRAI timer to expire [20].	Found in configurations. Default values depend on particular implementations.
BGP update processing time	Reflects workload in a router's CPU, which delays the processing of a BGP update, such as other BGP updates, OSPF calculations, and SNMP requests.	[0.01s, 0.5s] under different background load according to the empirical studies in [27][26]
Link delay	Link delays between routers	[0.01s, 0.1s] [15][17][28]

Second, we find that it is difficult for network operators to predict impact values based on their prior knowledge of the network, such as routing policies, and neighbor types (e.g., customer, providers, and peers). (i) Among the top 45 largest-impact sessions, four sessions are those with provider networks that are known to announce a large number of routes to the ISP network. Two sessions are with customer networks that draw a large amount of traffic. However, 39 other sessions are with peer networks, and these sessions have identical routing policies with sessions that have a much lower impact. Thus, impact values have little correlation with common knowledge, such as neighbor types and routing policies. (ii) We also analyze daily snapshots of the network's routing tables for a month-long period to see how the impact values evolve over time. In the top 45 list, thirteen sessions in the first snapshot were replaced by the end of the month period, experiencing dramatic changes in their impact (e.g., impact reduced from 51GB to nearly zero). These results imply that impact values cannot be accurately predicted based on previous impact values. To summarize, the results (i) and (ii) show that network operators cannot easily estimate impact on their own.

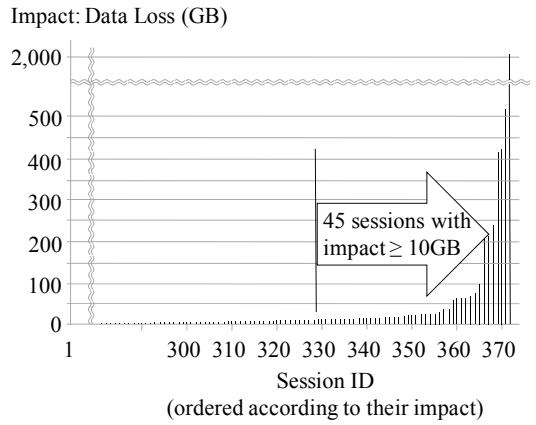


Figure 5. Impact of 372 BGP sessions

Therefore, to more precisely estimate impact values, we need to perform the proposed impact estimation on a regular basis.

D. Guidelines for Reducing Impact of Routing Sessions

We perform an in-depth analysis of the estimated impact values and discover a set of conditions that contribute to these impact values (e.g., backup routes are densely located only in a few routers). By decreasing the influence of these conditions, network operators can reduce the impact of a session if this session turns out to have a large impact (e.g., backup routes are distributed throughout more routers by rearranging sessions).

We first identify several important factors that determine the impact of a BGP session S, $I(S)$. Let $U(S)$ denote all routes that are heard over S, and $U_{best}(S)$ denote the subset of $U(S)$ that are selected as the best routes in the network.

- $I(S)$ is primarily determined by $|U_{best}(S)|$, the **number of best routes** that are advertised over S; only when the best routes are withdrawn, routes change and data get lost. For example, 116 sessions advertise more than one route ($|U(S)|>0$), but none of these routes are selected as best ($|U_{best}(S)|=0$). These 116 sessions thus have zero impact.
- The **location of alternative routes** to $U_{best}(S)$ is another factor that determines $I(S)$. When S is shut down at router R_i and a best route $u_1 \in U_{best}(S)$ is withdrawn, R_i will drop packets toward u_1 until an alternative route u_2 is advertised from other routers in the network. The farther u_2 is from R_i, the more data will R_i lose. Otherwise, if R_i already has u_2 in the routing table, R_i can immediately switch to u_2 as soon as u_1 is withdrawn. This situation is illustrated in Fig. 1 and Table I. R_2 initially has an alternative route u_2, but R_1 and R_3 do not. Thus, R_2 recovers from the loss of connectivity more quickly than the other two routers.
- The **number of neighbors** at each router and their **traffic demands** toward $U_{best}(S)$ also affect $I(S)$. A router will lose more data if its neighbors forward a large amount of traffic toward $U_{best}(S)$.
- The analyzed ISP (and many large networks) uses a **route-reflector hierarchy** in order to scale the number of intra-network sessions to a large number of routers [30]. In this hierarchy, routers maintain sessions with other routers such that the intra-domain session topology forms a tree – router r has a routing session only with its parents and children; r can thus avoid the overhead of a full mesh topology – maintaining a routing session with every other BGP-speaking router in the network. However, the route-reflector hierarchy slows down the propagation of route updates because route updates need to travel multiple BGP hops. As a result, the hierarchy increases route convergence time, and consequently, the amount of data loss. We estimated $I(S)$ by modifying the network's topology to the full-mesh topology, and we found that $I(S)$ at each router is reduced up to 82%, compared to the route reflector hierarchy.

According to the identified factors, we suggest the following in order to reduce $I(S)$. Most of these suggestions target better arrangement of alternative routes. This can be performed by strategically adding a few BGP sessions or by relocating existing sessions [16]. Note that we do not need to rearrange sessions for all routes in S. Rearranging sessions only for a few popular routes (those that carry the most traffic) is sufficient to reduce $I(S)$ significantly. The proposed suggestions complement the existing methods in Section I.B; the suggestions could be more effective and less costly, but they are not always feasible.

- If S belongs to router R_i, we can arrange BGP sessions such that alternative routes to $U_{best}(S)$ reside in R_i, rather than in other routers. If S is shut down, R_i will immediately switch to alternative routes; therefore, all other routers will continue to see R_i as the egress router and forward packets over the same path without losing any data. For example, if R_1 in Fig. 1 has an alternative route, R_2 and R_3 will continue to forward their data to R_1, and no data loss will occur. One way to implement this arrangement is to add a session (R_i, R_j) if an external neighbor R_j announces alternative routes. This allows R_i to receive the alternative routes over the new session.
- If arranging alternative routes in R_i according to the previous guideline is not practical (e.g., adding a session to R_i overloads the router), then we can distribute alternative routes in as many other routers as possible; more routers can regain connectivity quickly when the best route is withdrawn. When placing a limited number of alternative routes, we can reduce $I(S)$ more efficiently by giving priority to routers with higher traffic demands. We re-estimated $I(S)$ for the analyzed network by distributing alternative routes in all other routers except R_i, where S is down. This reduced $I(S)$ by 60-70% on average.
- If the route-reflector hierarchy is used, place alternative routes close to route reflectors – the hubs of the hierarchy. In this way, alternative routes will be propagated through the network more quickly. This

rearrangement can also be done by adding or relocating intra-network sessions. For example, the operator can add a session (R_r, R_k), where R_r is a route reflector, and R_k is a router (either internal or external) that has alternative routes. These routes will then be directly announced to R_r. We re-estimated $I(S)$ by relocating alternative routes on route reflectors, which reduced $I(S)$ by up to 14%.

V. CONCLUSIONS

We present a method that estimates the impact of shutting down BGP sessions. According to the estimated impact, network operators can prioritize different configuration tasks (e.g., by selectively protecting sessions with a high impact) and thus have considerable savings in maintenance costs. We implemented the proposed method and evaluated it with the routing sessions of a large ISP network. We found sessions that can lead to more than one thousand GB of data loss if not properly protected. We also found that the impact values fluctuate over time, often in an unpredictable way, as new peers are added and these peers advertise different sets of routes. Based on these results, we suggest that network operators estimate impact values on a regular basis in order to maintain their networks more carefully.

We are currently extending the impact estimation system to include various network configuration elements, such as packet filters, access control policies and routing policies. Once we estimate the impact of a network element, we can also apply a diverse set of prioritized actions according to the selected element. For example, we identify a route where very complex packet filters are applied in tandem, such that these filters are likely to confuse operators and cause configuration errors in the future; we can then reduce the complexity of the packet filters by streamlining policies, as shown in [31]. This streamlining prevents errors and saves the amount of time operators spend analyzing the complex filters when these filters cause problems.

In addition to *impact*, we are also looking into the *risk* of network elements, another component that measures the importance of network elements. As opposed to *impact*, which measures the amount of potential impact when negative events happen, *risk* represents the possibility that these negative events happen to a network element. For example, the *impact* of misconfiguring a packet filtering policy could prevent a thousand users from accessing a database; the *risk* of misconfiguring this policy could be the number of devices where the policy is configured. *Risk* can also incorporate the frequency and extent of misconfigurations in the past. We believe that a combination of the *impact* and *risk* metrics can more precisely measure the importance of a network element for the uninterrupted operation of the network.

REFERENCES

[1] "Network lifecycle management: a solution approach to managing networks", Hewlett-Packard, White Paper, Oct. 2007.

[2] R. Mahajan, D. Wetherall, and T. Anderson, "Understanding BGP misconfigurations", in *Proc. ACM SIGCOMM*, pp. 3-16, Aug. 2002.

[3] D. Oppenheimer, A. Ganapathi, and D. Patterson, "Why do Internet services fail, and what can be done about it?" in *Proc. USITS*, pp. 1-16, 2003.

[4] Z. Kerravala, "As the value of enterprise networks escalates, so does the need for configuration management", Enterprise Computing and Networking, Yankee Group, 2004.

[5] P. Francois, O. Bonaventure, B. Decraene, and P-A. Coste, "Avoiding disruptions during maintenance operations on BGP sessions," in *IEEE TNSM*, vol. 4, no. 3, Dec. 2007.

[6] S. Raza, Y. Zhu, and C. N. Chua, "Gracefule network operations," in *Proc. IEEE INFOCOM*, 2009.

[7] R. A. Miura-Ko and N. Bambos, "SecureRank: A risk-based vulnerability management scheme for computing infrastructures," in *Proc. IEEE ICC*, 2007.

[8] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, and J. Rexford, "NetScope: traffic engineering for IP networks," in *IEEE Network*, vol. 14, no. 2, pp. 11-19, Mar/Apr 2000.

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

[10] N. Feamster, J. Winick, and J. Rexford, "A model of BGP routing for network engineering," in *Proc. ACM SIGMETRICS*, pp. 331-342, 2004.

[11] S. Lee, S. Nelakuditi, Y. Yu, Z-L. Zhang, and C. N. Chua, "Proactive vs. reactive approaches to failure resilient routing," in *Proc. IEEE INFOCOM*, Mar. 2004.

[12] O. Bonaventure, C. Filsfils, and P. Francois, "Achieving sub-50 milliseconds recovery upon BGP peering link failures," in *Proc. ACM CONEXT*, Oct. 2005.

[13] P. Francois, M. Shand, and O. Bonaventure, "Disruption-free topology reconfigurations in OSPF networks," in *Proc. IEEE INFOCOM*, May 2007.

[14] R. Teixeira and J. Rexford, "Managing routing disruptions in internet service provider networks," in *IEEE Communications*, Mar. 2006.

[15] T. G. Griffin and B. J. Premore, "An experimental analysis of BGP convergence time," in *Proc. IEEE ICNP*, pp. 53-61, Nov. 2001.

[16] C. Pelsser, T. Takeda, E. Oki, and K. Shiomoto, "Improving route diversity through the design of iBGP tolopogies," in *Proc. IEEE ICC*, pp. 5732-5738, 2008.

[17] W. Sun, Z. M. Mao, and K. G. Shin, "Differentiated BGP update processing for improved routing convergence," in *Proc. IEEE ICNP*, Nov. 2006.

[18] Y. Afek, A. Bremler-Barr, and S. Schwarz, "Improved BGP convergence via ghost flushing," in *IEEE JSAC*, vol. 22, no. 10, 2004.

[19] N. Feamster, J. Borkenhagen, and J. Rexford, "Guidelines for interdomain traffic engineering," in *ACM CCR*, vol. 33, no. 5, Oct. 2003.

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

[21] S. Lee, T. Wong, and H. S. Kim, "To automate or not to automate: on the complexity of network configuration," in *Proc. IEEE ICC*, May 2008.

[22] A. Feldmann, A. Greenberg, C. Lund, N. Reingold, J. Rexford, and F. True, "Deriving traffic demands for operational IP networks: methodology and experience," in *IEEE Transactions on Networking*, vol. 9, no. 3, pp. 265-280, 2001.

[23] 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*, pp. 206-217, 2003.

[24] S. Lee, T. Wong, and H. S. Kim, "NetPiler: Detection of ineffective router configurations", in *IEEE Journal on Selected Areas in Communications (JSAC) Special Issue on Network Infrastructure Configuration*, vol. 27, no. 3, pp. 291-301, Apr. 2009.

[25] "Scalable Simulation Framework (SSF)," http://www.ssfnet.org/

[26] Z. M. Mao, R. Govindan, G. Varghese, and R. H. Katz, "Route flap damping exacerbates Internet routing convergence," in *Proc. ACM SIGCOMM*, pp. 221-233, Aug. 2002.

[27] A. Feldmann, H. Kong, O. Maennel, and A. Tudor, "Measuring BGP pass-through times," in *Proc. Passive and Active Measurement Workshop (PAM)*, pp. 267-277, 2004.

[28] I. Stoica, D. Adkins, S. Zhang, S. Shenker, and S. Surana, "Internet indirection infrastructure," in *Proc. ACM SIGCOMM*, pp. 205-218, 2002.

[29] J. Rexford, J. Wang, Z. Xiao, and Y. Zhang, "BGP routing stability of popular destinations," in *Proc. ACM IMC*, 2002.

[30] T. Bates, E. Chen, and R. Chandra, BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP), RFC-4456, 2006.

[31] S. Lee, T. Wong, and H. S. Kim, "Improving dependability of network configuration through policy classification," in *Proc. IEEE DSN*, 2008.