On Understanding of Transient Interdomain Routing Failures

Feng Wang†, Lixin Gao†, Jia Wang‡ and Jian Qiu†
†Department of ECE, University of Massachusetts, Amherst, MA 01002
‡AT&T Labs-Research, Florham Park, NJ 07932
†{fewang, lgao, jqiu}@ecs.umass.edu, ‡jiawang@research.att.com

***Abstract*— The convergence time of the interdomain routing protocol, BGP, can last as long as 30 minutes [14], [15]. Yet, routing behavior during BGP route convergence is poorly understood. BGP can experience transient loss of reachability during route convergence. We refer to this transient loss of reachability during route convergence as *transient routing failure*. Transient routing failures can lead to end-to-end forwarding failures. Furthermore, the prolonged routing failures can make deploying applications such as voice-over-IP and interactive games infeasible. In this paper, we study the extent to which transient interdomain routing failures occur in the Internet and the duration that these failures can last through both analysis and measurement. We first present a formal model that captures the transient behavior of the interdomain routing protocol. We derive sufficient conditions for and an upper bound for the duration of transient routing failures. Furthermore, we demonstrate the occurrence and duration of transient routing failures in the Internet through measurement. We find that majority of transient failures occur under the commonly applied routing policy setting, and popular and unpopular prefixes can experience transient failures.**

I. Introduction

Routing protocols as the "control plane" of the Internet play a crucial role in the end-to-end performance of the Internet. The Internet is divided into thousands of Autonomous Systems (ASes). Routing information is exchanged using the interdomain routing protocol, Border Gateway Protocol (BGP), and routing within an AS is performed using an intradomain routing protocol such as IS-IS or OSPF. Studies have shown that intradomain routing can achieve convergence time of a few hundred milliseconds [20]. In contrast, the convergence time of BGP can last as long as 30 minutes [14], [15]. Furthermore, BGP routing instability is pervasive and can occur frequently [2], [5], [16], [17]. Yet, routing behavior during BGP route convergence is poorly understood.

BGP can experience transient loss of reachability during route convergence. For example, if router A uses router B to reach a destination, router A does not announce its best route to router B. This can limit the route visibility of router B. Limited route visibility makes it possible for a router to experience transient loss of reachability during the path exploration of the route convergence process. Using the same example as above, suppose that a link failure makes router B's best path infeasible and makes router B reach the destination only via router A. Router B has to withdraw its route from router A before router A can announce a route to router B. That is, router B can temporarily lose reachability to the destination during the route convergence process. We refer to this transient loss of reachability during route convergence as *transient routing failure*. Similarly, routing policies and iBGP configurations can limit the route visibility and therefore can lead to transient routing failures.

Transient routing failures can lead to end-to-end forwarding failures. Furthermore, the prolonged end-to-end forwarding failures can make deploying applications such as voice over IP and interactive games infeasible. Therefore, it is important to understand when transient routing failures can occur and how long these transient routing failures can last. However, analysis and measurement studies of transient routing failures can be challenging. First, existing abstract models for BGP focus on route convergence properties or traffic engineering within an AS [4], [8], [13]. The occurrence and duration of transient failures depend on the timing of propagation of route updates, which can be correlated with timing of various events in the network (e.g., link failures, or network configuration changes). Further, timing of routing updates for one prefix is correlated with timing of routing updates for other prefixes since route update rate limiting timers are typically set for each BGP peering session instead of for each prefix. Second, the measurement of transient routing behavior in the Internet requires to differentiate transient routing failures from failures caused by many network changes (e.g., a network or prefix is temporarily unavailable).

Our major contributions are summarized as follows. (i) Contrast to existing models of BGP [6]–[10], [13], we present an abstract model to capture transient behaviors of BGP, which allows us to scrutinize the detailed interactions between BGP routers. (ii) With the aid of the model, we identify the sufficient conditions for the occurrence of transient routing failures and derive the upper bound of the duration of transient failures. (iii) We derive the upper bound of convergence delay for failover routing changes, which is shown to be much longer than the previous result [16]. (iv) We show that, in a typical BGP system that deploys routing policies conforming to commercial agreements between ASes and applying hierarchical iBGP configurations, any router can experience transient failure. (v) We demonstrate the extent that transient routing failures occur in the Internet by examining a large collection of routing data and configuration files of hundreds of routers. We show that that transient routing failures occur often and can last for a significant period of time. More importantly, we find that

transient failures can have a large impact on data traffic of both popular and unpopular prefixes.

The rest of this paper is organized as follows. Section II describes our model for investigating transient behavior of a BGP system. Section III presents sufficient conditions for transient failures. We study duration of transient failures and convergence delay, and analyze transient failures in a typical BGP system in Sections IV and V. Section VI reports the measurement of transient failures in the Internet. We list related works in Section VII. We conclude the paper with a discussion of future work in Section VIII.

II. Abstract Model for BGP Transient Behavior

In this section, we present an abstract model for investigating transient behavior of a BGP system during a transition triggered by an event. That is, we aim to characterize transient routing states during the route convergence process triggered by an event. Our model extends other existing frameworks [11], [12], [21] that capture the long-term stability of BGP. One of the key challenges in modelling the transient behavior is to capture the timing that routing updates are exchanged. Our model is able to characterize all possible sequences that routing updates are advertised asynchronously along BGP peering sessions, and abstract away the detailed interaction among prefixes and events.

A. Formal Model

A BGP system (G, P) contains topology G and routing policy P. The topology of the BGP system is modelled as a graph $G = (V, E)$, where the node set V consists of all BGP-speaking routers, and the edge set E consists of all BGP peering sessions. We include both iBGP and eBGP sessions in order to capture routing dynamics within an AS. Each BGP speaker belongs to one AS and an AS can have one or more BGP speakers.

In our model, we focus on a single destination prefix d that originates from AS 0. Clearly, prefixes in a BGP system can interact. First, a supernet prefix is used when a subnet prefix is withdrawn. However, we assume that d does not have a supernet. In Section VI, we will see that majority of prefixes that experience transient failures do not have a supernet prefix. Second, BGP routing updates are exchanged at a time triggered by timers. This can lead to correlated routing updates among multiple prefixes. However, our model will capture all possible sequences that routing updates are exchanged. By focusing on a single destination prefix, we will not lose generality.

In order to capture transient behavior of the BGP system, we define the state of the BGP system in terms of the routes stored by each BGP speaker. That is, each speaker remembers the routes received from its neighbors, and also its best route. As such, we define the system state as a vector $S = (s_1, s_2, ..., s_n)$, where s_i denotes the set of routes stored at speaker $i = 1, 2, \ldots, n$, and the first route in s_i is the best route of speaker i.

The route-selection process proceeds in a distributed and asynchronous fashion, triggered by advertisements and withdrawals of routes. Triggered by an update, a BGP speaker applies the BGP selection process to pick the best path to d, after applying import policies to the routing update received along one of its BGP sessions or edges. A BGP speaker sends an update along an edge after its decision process changes its best route and its export routing policy allows such an update. Note that the exact timing that a BGP speaker sends an update along an edge to its neighbor is determined by *Minimum Route Advertisement Interval* (MRAI) Timer.

Formally, we model the BGP route decision process as follows. Once an routing update is triggered by BGP MRAI timers, the routing update is sent to the corresponding BGP speaker and the speaker applies the route selection process. *Triggering* a routing update will cause the corresponding BGP speaking routers (or nodes) to apply the import policy to received routes, to run the BGP path-selection process, and to apply export policy for generating routing updates to the speaker's neighbors.

Since routing update rate limiting timers operate independently, BGP route updates are triggered asynchronously. For each state, we have a set of routing updates, U, that contains routing updates to be triggered and sent in a future time. At any time, a subset $A \subseteq U$ of edges can be triggered. We call $T = (t_1, t_2, \ldots, t_k)$, an *trigger set*, whereas t_i indicates the trigger of a routing update that will be sent along an edge $u \rightarrow v$. Since a routing update is always sent along an edge, we also use the directed edge to represent the routing update.

Given a state $S = (s_1, s_2, \ldots, s_n)$ and a trigger set $T \subseteq U$, where U is the future trigger set for state S, the next state $S' = (s'_1, s'_2, ..., s'_n)$ and the next future trigger set U' will be derived as follows.

$$S' = \mathrm{D}ecisionProcess(S, T)$$

$$U' = (U - T) \bigcup \mathrm{N}ewT$$

where $\mathrm{D}ecisionProcess(S, T)$ derives the new state for each router by running the import policy and best path selection process for each router that receives a routing update. Other routers remain the same state. $\mathrm{N}ewT$ is a set of routing updates that are triggered by the decision process. In other words, $\mathrm{N}ewT$ contains routing updates generated by routers whose state has changed and such changes are allowed to export to neighbors by the export policies. The union operation performs an union on $(U - T)$ and New$T$ so that newer updates are kept if there is an overlap between the two sets. That is, if an edge is in both $(U - T)$ and New$T$, the update message in $\mathrm{N}ewT$ will be in the future trigger set U'.

A BGP system can go through a series of state transition after the occurrence of an event. These events include link failures, BGP session reset, link addition, router crash, router recovery, and routing configuration changes. In order to capture the transient behavior for a BGP system after the occurrence of an event, we introduce a state graph that describes all possible transient states of the BGP system, and transition between these transient states. A *state graph* is a directed graph, where each node represents a state and its corresponding trigger set, while an edge between states S

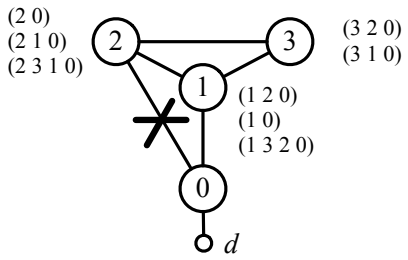


Fig. 1. A BGP system with a link failure. The text around a node lists all paths that are allowed to export to the node, which reflects the export policy. The order of paths reflects the import policy applied by each node. Note that the paths at each node represent potential routes to the destination, and not all of them will show in routing table at the same time.

and S' represents the transition from state S to S' given a subset of the trigger set U in state S. For example, Figure 1 shows a BGP system with a link failure event and Figure 2 shows the state graph for the BGP system. In Figure 1, we show both export and import routing policies in the BGP system. It means, only paths that are allowed to export are shown, and local preference ranking is shown by the order of paths. In Figure 1, the text around a node lists all shows paths that are allowed to export to the node. The order of paths represents their ranks so that the first one has the highest rank. Additionally, node 1 does not forward the path (1 2 0) to node 3 according to its export policies. Each state graph has an initial state that represents the states of BGP speakers initially and the trigger set associated with the event. A directed path in a state graph represents a trigger sequence. For example, in Figure 2, $(T_0T_2T_8T_{18}T_{21})$ is a trigger sequence, where T_0 contains the routing update from node 0 to node 2 indicating withdrawal of the route to node 0.

A state S in a state graph is a *transient state* if $S' \neq S$ for an $T \subset U$. A state S in a state graph is a *stable state* if $S' = S$ for any $T \subset U$. Griffin *et al* have shown in [13] that in a stable state, the best paths to the destination formed from all BGP speakers is a directed tree where the direction of each edge is the same as the direction that packets traverse to reach the destination. We refer to this direct tree as *best path tree* of the stable state. In this paper, we focus on studying the transient behavior of a BGP system that can always reach a stable state for the given event. We also assume that the BGP system is in a stable state before the occurrence of the event. The best paths at the initial routing state form a best path tree. We refer to the best path tree as the *best path tree of the initial state*.

B. Control Plane and Data Plane Failure States

A router is in *control plane failure state* if it has no route to the destination on control plane. For example, state S_2 in Figure 2 is a control plane failure state for router 2. Whether a router goes through a control plane failure state depends on the trigger set sequence it passes. For example, in Figure 2, router 3 does not go through a control plane failure state for trigger set sequence $(T_0T_2T_{10}T_{15})$ while it does for trigger set sequence $(T_0T_2T_9T_{12})$. On the contrary, router 2 will definitely go through a control plane failure state. Given a BGP system and an event, a router experiences *potential control plane transient failure* if there is a path from the initial state to the final stable state such that the path contains a control plane failure state for the router; a router experiences *control plane transient failures for sure* if *any* path from the initial state to the final stable state contains a control plane failure state for the router.

A *forwarding path* on data plane is the path that packets actually pass from a router to a destination. The forwarding path of a router in a state can be constructed by starting from this router and iteratively appending the next hop router of each router to the path. If a router has no complete forwarding path to the destination or the path contains a loop, the router has a *null* path.

A router is in *data plane failure state* if it has a null path to the destination on data plane. It is clear that if a router goes through a control plane failure state, it is sufficient but not necessary for it to experience a data plane failure state. For example, state S_3 of Figure 2 is a data plane failure state for router 3 but not a control plane failure state. Similarly, given a BGP system and an event, a router experiences *potential data plane transient failures* if there is a path from the initial state to the final stable state so that the path contains a data plane failure state for the router; a router experiences *data plane transient failures for sure* if *any* path from the initial state to the final stable state contains a data plane failure state for the router.

III. Sufficient Condition for Transient Failures During Failover

Since a sufficient condition for control plane transient failures is also sufficient for data plane transient failures, we focus on the sufficient condition for the former. Transient failures take place during the failover events that do not disconnected the destination to the network. For simplicity, the events are assumed to be triggered by link failures.

We first introduce *path availability (PA) graph* of a BGP system. A PA graph is a directed graph, consisting of all BGP routers. For two neighboring routers u and v, if the path of one of them is available to the other, there is a directed edge between them. The edge is a solid directed line from u to v if the edge is in the best path tree in the initial state and u installs v's path. Otherwise, the edge is a directed dashed line from u to v if the best path of v is announced to u and installed at u as *backup*. Note that an edge between two nodes can be bi-directional dashed line if both sides announced their best paths to each other as backup paths. But a solid edge cannot be bi-directional. Each node has one and only one outgoing solid edge.

In a PA graph, a node u is said *reachable* to the destination if there exists a path from u to the destination such that the edges of the path have the same direction as the path and all of them are solid except the one adjacent to the destination. A *predecessor/successor* of a node means the

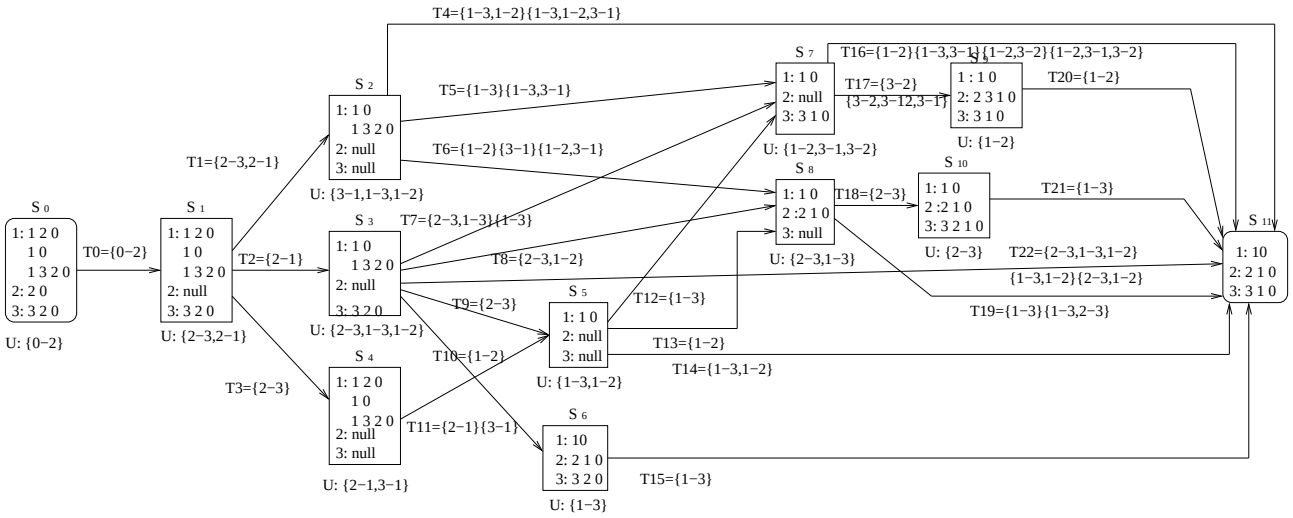
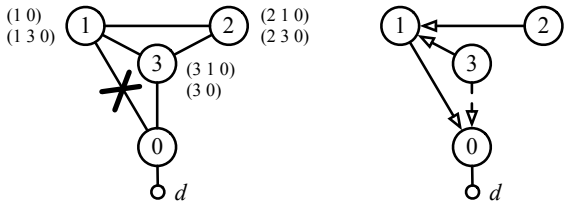


Fig. 2. The state graph for the BGP system described in Figure 1.



(a) BGP system (b) PA graph

Fig. 3. An example of control plane failures.

predecessor/successor of the node in the best path tree in the initial state.

Figure 3(a) gives an example of control plane transient failures. Figure 3(b) shows the corresponding PA graph. Initially, nodes 2 and 3 use node 1 to reach node 0. When the link between nodes 1 and 0 fails, node 2 possibly experience a control plane transient failure if node 1 sends the withdrawal message to node 2 earlier than the arrival of the path (3 0) from node 3. Note that node 3 announces path (3 0) only after it receives the withdrawal of path (1 0) from node 1. Node 1 will temporarily lose all its paths and experiences a control plane transient failure for sure.

With the PA graph, we have the following sufficient conditions for control plane transient failures.

Theorem 1 *A node u in a BGP system will experience potential control plane transient failures when a link l fails if in the corresponding PA graph, (i) l is a solid line; and (ii) if l is removed, any node in node u's best path is not reachable to the destination.*

Proof: We prove the theorem by constructing an trigger set sequence that causes node u have no path to the destination in some state. The second condition implies that all nodes along the best path of node u, including node u, have no alternate path installed in the initial state. Once the node adjacent to link l, which is closer to u, detects the link failure, it will send a withdrawal message to its predecessors, and so on. The withdrawal message will be finally received by u. So we construct an trigger set sequence start at link l. Suppose the best path from u to link l is $(j_1, \ldots, j_m, l)$, we construct an trigger set sequence, $(T_l, T_{j_m}, \ldots, T_{j_1})$ by which node u changes from state S to S'. At state S', node u will have no path to the destination. ■

Theorem 2 *A node u in a BGP system experiences a control plane transient failure for sure when a link l fails if (i) for each predecessor of u, u has at most one path to the destination through it; and (ii) u has only one path to the destination not through u's predecessors but through l.*

Proof: We prove this theorem by considering all possible activation sequences that can cause node u to have no path to the destination. The first condition implies that each of the predecessors of node u may have alternate paths to the destination. But they will not advertise the alternate paths to u until u withdraws its best path since the paths through u are their best paths. The second condition implies that node u receives only one withdrawal message from its successor node after the failure. Therefore, u will temporarily have no path to the destination. Thus u experiences transient failure. Before node u receives the withdrawal from its successor, we already know from the first condition that its predecessors cannot provider an alternate path. Thus, only after node u sends a withdrawal message to its predecessors, its predecessors will provide an alternate path. ■

Theorem 3 *A node u in a BGP system experiences a control plane transient failure for sure when a link l fails if (i) after the removal of l, node u is reachable via u's predecessors only; and (ii) the predecessors of node u prefer the paths through u over any other paths.*

Proof: The first condition implies that node u will finally lose all paths not through its predecessors. The second condition implies that the predecessors of node u cannot advertise their alternate paths to node u unless they receive a withdrawal message from u. After u drops all the paths not through its predecessors, u will temporarily have no path to the destination. Thus, u experiences transient failure. ■

Note that Theorem 2 provides a topological sufficient condition for transient failures for sure, while Theorem 3 relaxes the topological constraint but imposes an additional routing policy constraint.

IV. Duration of Transient Failures and Convergence Delay

Transient failures' duration and convergence delay can be identified with state graph. A control plane transient failure for node u can be precisely represented with a trigger set sequence or a path in the state graph, in which the last state in the path is the only state in which u installs a *route* to the destination. Similarly, a data plane transient failure for node u can be precisely represented with a trigger set sequence or a path, in which the last state in the path is the only state in which u has a *data forwarding path* to the destination.

In order to capture the timing for a state to transit to the next, we annotate each edge of the state graph with a *weight* that represents the time that it takes to activate the trigger set. The *delay of a path led by a trigger set sequence* is defined as the sum of weights of edges on the path. More precisely, given two neighboring nodes u and v in a BGP system, the MRAI timer of u configured for v is denoted by M_{uv} and the weight for the edge in the direction from u to v is d_{uv}. Obviously, $d_{uv} \leq M_{uv}$. Note that an edge in the BGP system can be either an iBGP or an eBGP session. The value of M varies for different edges. Suppose node v_k reaches node v_1 through path $P_{(v_k,v_1)} = (v_k v_{k-1} \cdots v_1)$. The delay of path $P_{(v_k,v_1)}$ is denoted by $d_{v_k v_1} = \sum_{i=k}^{2} d_{v_i v_{i-1}}$. Obviously, $d_{v_k v_1} \leq \sum_{i=k}^{2} M_{v_i v_{i-1}}$.

Given a BGP system and an event, the *duration of transient control plane failure* is the delay of the path led by the trigger set sequence of the transient failure and the *duration of transient data plane failure* is the delay of the path led by the trigger set sequence of the transient failure on the data plane. The *control plane convergence delay* is the delay of the path led by the trigger set sequence from the initial state to the final state and the *data plane convergence delay* is the delay of the path led by a trigger set sequence from the initial state to the final forwarding state, where the final forwarding state is the first state on the path whose corresponding data forwarding state is the same as the final state.

Although [15] has shown that routing convergence delay might be long, their theoretical bound for convergence delay is limited to the events that lead to disconnectivity or bring a route back. However, we argue that the most common events in the Internet are those that lead to failover.

Next, we derive upper bounds of transient failure durations for the case specified by Theorem 2. The upper bounds for the other conditions can be derived in the similar way.

We first locate the nodes where node u can obtain a path, i.e., the node that can be activated to provide u failover path information. After link failure occurs, node u sends a withdrawal to its predecessors in the best path tree at the initial state. The predecessors will forward the withdrawal message to their predecessors if they do not have alternate paths, and so on. If any predecessor can reach the destination via a path other than the best path, it has an alternate path in its routing table in the initial state. For each predecessor of u, there is at most one such node. We denote these nodes with a set $\mathrm{B} = \{\beta_1, \beta_2, \ldots, \beta_\mathrm{m}\}$. Once one of them switches to the alternate path after being activated, the new path information will be advertised to its neighbors. One of the new paths will be finally installed at u.

With the above notions, we have the following theorems.

Theorem 4 *The control plane transient failure duration at node u is bounded by*

$$\min_{\beta \in B}(d_{u\beta} + d_{\beta u})$$

Proof: The transient failure on control plane of node u starts when u loses its only path that goes through link l and ends when it obtains the ***first*** path from one of its predecessors. Node u obtains the failover path from m possible β nodes and the failover path is triggered by the withdrawal announced by u. So after u announces the withdrawal, the delay of obtaining the first path is upper bounded by $\min_{\beta \in B}(d_{(u,\beta)} + d_{(\beta,u)})$. ■

Theorem 5 *The data plane transient failure duration at node u is bounded by*

$$d_{lu} + \min_{\beta \in B}(d_{u\beta} + d_{\beta u}) = \min_{\beta \in B}(d_{l\beta} + d_{\beta u})$$

where l is the failed link.

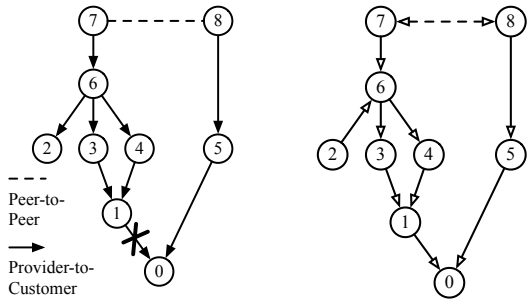
Proof: Suppose v is adjacent to link l and on the same side of l as u. The duration of the data plane transient failure can be derived directly from the duration of the control plane failure in Theorem 4. The only difference is the delay of the withdrawal message from the node v to u, which is a fixed value $d_{(l,u)}$. ■

Theorem 6 *The convergence delay on the control plane and the data plane at node u is bounded by*

$$d_{l,u} + \max_{\beta \in B}(d_{u\beta} + d_{\beta u}) = \max_{\beta \in B}(d_{l\beta} + d_{\beta u})$$

where l is the failed link.

Proof: To derive the upper bound for convergence delay on the control plane and data plane, we assume that the path in the final state is the last failover path arriving at node u. The convergence time is related to the longest delay for a route advertisement from node $\beta \in B$ to node u, that is,



(a) BGP system (b) PA graph

Fig. 4. Transient failures take place in a typical BGP system

$\max_{\beta \in B}(d_{(\beta,u)})$. The route announcement is activated by the withdrawal from v which is adjacent to link l and on the same side of l as u. So the longest delay for the withdrawal message is $d_{(l.u)} + \max_{\beta \in B}(d_{u,\beta}) = \max_{\beta \in B}(d_{(l,\beta)})$. Therefore, the upper bound of the convergence delay is $\max_{\beta \in B}(d_{(l,\beta)} + d_{(\beta,u)})$. ■

V. Transient Failures in a Typical BGP System

A typical BGP system means that every router in the system applies *common routing policies*. Routing policies are typically guided by commercial relationships between ASes. In a *provider-to-customer* relationship, the customer pays the provider for access to the rest of the Internet. Two ASes with a *peer-to-peer* relationship exchange traffic between their respective customers. In a typical BGP system, the provider-to-customer relationships form a hierarchy of ASes, that is, the directed graph formed by provider-to-customer relationships does not contain a cycle. The export routing policies are typically guided by the *no-valley* routing policy, in which an AS does not export its provider or peer routes to its providers or peers. An AS always announces its customer routes to all its neighbors, and its peer or provider routes to its customers. The import routing policies are guided by the *prefer-customer* routing policy, in which each AS prefers its customer routes over its peer or provider routes.

A. Transient Failures at AS Level

Assume that each AS has a single router. We will relax this assumption in Section V-B and consider transient failures at all routers inside an AS (i.e., at router level).

Transient failures can be prevalent in a typical BGP system. For instance, Figure 4(a) shows a typical BGP system, in which we assume that each AS contains only one router and the destination is in AS 0. Figure 4(b) illustrates the corresponding PA graph. Suppose that the link between nodes 1 and 0 breaks, then node 6 satisfies the sufficient conditions in Theorem 3, nodes 1, 3, and 4 satisfy the sufficient conditions in Theorem 2, and node 2 satisfies the conditions in Theorem 1. Therefore, nodes 1, 3, 4, and 6 will definitely experience transient failure and node 2 may experience the failure.

Now, we describe necessary conditions for transient failures occurred in tier-1 and tier-2 ASes since our measurement studies in Section VI focus on tier-1 and tier-2 ASes only due to the limited public available routing data. Note that a tier-2 AS is an AS that has at least one tier-1 AS as its provider. Before showing the property of tier-1 and tier-2 ASes that experience transient failures, we have the following two lemmas.

Lemma 1 *In a typical BGP system, if AS u converges to a customer route after a link failure, then AS u does not experience transient failure during the failover convergence.*

Proof: Suppose after the link failure, AS v_k uses path $P = (v_k v_{k-1} \dots v_1)$ to reach AS v_1, who originates the destination, and $(v_k v_{k-1})$ is a customer link. According to the no-valley policy, $(v_i v_{i-1})$ must be a customer link for any $i = k, k-1, \dots, 2$. We prove the lemma by induction on i.
Base step: $i = 2$. Since after a transient failure, AS v_2 has a path $(v_2 v_1)$ to AS v_1. AS v_2 must install this path before the failure because v_1 advertises the destination to all providers. Meanwhile, due to prefer-customer routing policies, v_2's best path to v_1 should also use a customer link. When the link failure occurs, v_2 has at least one path $(v_2 v_1)$ to v_1. Therefore, v_2 does not experience transient failure.
Inductive step: Assume that before the transient failure, AS v_{k-1}'s best path to AS v_1 goes through a customer link. Ater the failure, its best path to v_1 still goes through a customer link and v_{k-1} does not experience transient failure. Then v_k always has at least one path to v_1 through v_{k-1}. Therefore, v_k does not experience transient failure. ■

Lemma 2 *In a typical BGP system, if AS u converges to a peer route after a link failure, AS u does not experience transient failure.*

Proof: Suppose after the link failure, AS u uses a peer link through its peer AS v to reach the destination. Due to no-valley policies, v must use a customer link to reach the destination after the failure. According to Lemma 1, v will not experience transient failure, which implies that v always has a customer path during the failure. The paths must be advertised and installed at u. Thus, during the failure, AS u will always has at least one path to the destination through v. So, AS u will not experience transient failure. ■

With the aid of Lemmas 1 and 2, we show the property of tier-1 and tier-2 ASes that experience transient failures in the following two theorems.

Theorem 7 *In a typical BGP system, a tier-1 AS cannot experience transient failures.*

Proof: Since a tier-1 AS does not have a provider, according to Lemma 1 and 2, the tier-1 AS can not experience transient failure. ■

Theorem 8 *In a typical BGP system, if a tier-2 AS experiences transient failures after a link failure, it must use a customer link to reach the destination before the link failure.*

Proof: A tier-2 AS u has a tier-1 AS w as a provider. First, w does not experience transient failures with Theorem 7. Second, w announces its best route to u except when w's best route goes through u. Therefore, u can experience transient failures only when w uses u to reach the destination. However, this is only possible when u uses a customer route to reach the destination according to no-valley policy. We prove by contradiction. Assume that a tier-2 AS u experiences a transient failure during a link failure. Suppose that before the failure,

1) AS u uses a provider v to reach the destination. According to Theorem 7, v is tier-1 and will never experience transient failure. Because its best path is always known to u, AS u would not experience transient failure, which is a contradiction.

2) AS u use a peer v to reach the destination. According to Lemma 1 and 2, after the failure, AS u will go through a tier-1 provider w to the destination. However, w will never experience transient failure. Because u always has a path through w, it would not experience transient failure, which is a contradiction.

According to 1) and 2), u must go through a customer link to reach the destination before the failure. ■

B. Transient Failures at Router Level

Our previous analysis shows that tier-2 ASes will experience transient failures, and tier-1 ASes will not experience transient failures under the assumption that those ASes contain only one router. Here, we relax this assumption, i.e., we consider transient failures for an AS containing a set of routers. We focus on transient failures occurring at routers in tier-1 ASes.

Although a tier-1 AS cannot experience transient failures at AS level, routers in a tier-1 AS can. Tier-1 ASes typically have a fully meshed or a hierarchical iBGP structure. Here we focus on describing a hierarchical iBGP structure. Similarly, we can describe scenarios that routers in a fully meshed iBGP structure experience transient routing failures. In a hierarchical iBGP structure, route reflectors are deployed and it consists of a set of *Backbone Routers* (BRs), which are BGP route reflectors, and a set of *Edge Routers* (ERs), which are route reflectors' clients. An edge router could be an *access router* that connects to customer network, or a *peer router* that connects to peer network. For example, Figure 5 shows four backbone routers, which are fully meshed. Each BR connects to a set of edge routers. In terms of route export and import policies, BRs have a peer-to-peer like relationship between each other, i.e., a BR does not transfer the routes between two other BR routers. A BR and its ERs have a provider-to-customer like relationship, i.e., they import and export routes from and to each other without discrimination.

Let's consider the case that all routers use the same egress point to reach the destination. For example, in Figure 5, if

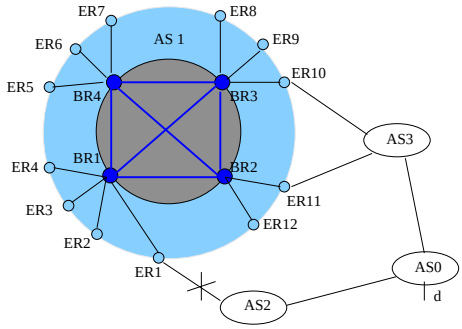


Fig. 5. A tier-1 AS with a hierarchical iBGP structure.

AS 2 is a customer of AS 1, and AS 3 is a peer of AS 1, according to prefer-customer routing policy, the path via ER1 is assigned higher local preference value than those via ER10 and ER11. As a result, all routers inside AS 1 will use the path via ER1 to reach the destination. Once the link between ER1 and AS 2 fails, all routers inside the AS experience transient failures. According to Theorem 2, all BR1, BR2, BR3, and ER1 will experience transient failure for sure because they have only one path through ER1 to reach the destination before the failure, and all of their predecessors have only one path to reach the destination. All ERs except ER1, ER10, and ER11, and BR4 will experience potential transient failures according to Theorem 1. Therefore, those routers cannot reach the destination along the path through ER11 as soon as the failure occurring. The above example shows if all of routers inside a tier-1 AS use the same egress point to reach a destination, they will experience transient failures if the egress point cannot reach the destination.

C. Transient Failure Duration and Convergence Delay

In Section IV, we derive the upper bound of transition failure duration and convergence delay for the case specified by the sufficient conditions in Theorem 2. But for the case in Theorem 3, it is difficult to precisely identify a set B who provides failover path to node u because node u has a set of paths to the destination through the failure link l in a general BGP system. However, in a typical BGP system, we are able to define a similar node set B_{T1} that consists of *routers* 1) in the (direct or indirect) providers of u, 2) in tier-1 providers, 3) use their direct customer or peer links to reach the destination. Note that the nodes from whom u obtains a failover path should be the nodes in the (direct or indirect) providers of u that do not experience transient failures if u is not in a tier-1 AS, the delay that u obtains a backup path from its provides will be no worse than the delay that u obtains a backup path from a node in B_{T1}. Therefore, with the notion of B_{T1}, we are able to formulate the upper bounds of transient failure duration and convergence delay similar to Theorems 4∼6.

Theorem 9 *The control plane transient failure duration at node u is bounded by*

$$\min_{\beta \in B_{T1}} (d_{(u,\beta)} + d_{(\beta,u)}).$$

Proof: If u is in a tier-1 AS, it is trivial to show that the theorem provides a loose upper bound of the transient failure duration. Suppose u is not in a tier-1 AS, the node that provides u the first failover path must be in the same AS as u or in the direct or indirect providers of u, which is denoted by v. We will show that $d_{(u,v)} + d_{(v,u)} \leq \min_{\beta \in B_{T1}}(d_{(u,\beta)} + d_{(\beta,u)})$ by contradiction. Suppose w is such a node in B_{T1} that has the minimal round trip propagation delay to u and $d_{(u,v)} + d_{(v,u)} > d_{(u,w)} + d_{(w,u)}$. Along the path from u to w, there must exist a node v' which provides u a failover path. However, $d_{(u,w)} + d_{(w,u)} > d_{(u,v')} + d_{(v',u)}$, which contradicts with the assumption that v is the first among the nodes that provide u failover paths. Therefore, the upper bound is $\min_{\beta \in B_{T1}}(d_{(u,\beta)} + d_{(\beta,u)})$. ■

Theorem 10 *The data plane transient failure duration at node u is bounded by*

$$\max_{P_j \in P_{(u,l)}}(d_{(l,u)}) + \min_{\beta \in B_{T1}}(d_{(u,\beta)} + d_{(\beta,u)}).$$

where l is the failed link.

Proof: In the case specified by the sufficient condition in Theorem 3, u has a number of paths to the destination through the failure link l. The transient failure on the control plane begins when all these paths are withdrawn. Therefore, before u experiences transient failure on control plane, the duration that node u should take to wait for all the paths in $P_{(u,l)}$ to be withdrawn is upper bounded by $\max_{p_j \in P_{(u,l)}}(d_{(l,u)})$. We know that the transient failure on data plane begins when the link failure happens and ends when u gets the first failover path. According to Theorem 10, the upper bound of the data plane transient failure duration is $\max_{p_j \in P_{(u,l)}}(d_{(l,u)}) + \min_{\beta \in B_{T1}}(d_{(u,\beta)} + d_{(\beta,u)})$. ■

Similarly, we get the upper bound on the convergence delay as follows.

Theorem 11 *The convergence delay on the control plane and the data plane at node u is bounded by*

$$\max_{P_j \in P_{(u,l)}}(d_{l,u}) + \max_{\beta \in B_{T1}}(d_{(u,\beta)} + d_{(\beta,u)})].$$

where l is the failed link.

At last, we compare our result with [14], in which the convergence delay of a fail down event is upper bounded by $n \times MRAI$, where n is the length of the longest path to the destination. Our result shows that convergence time can be longer for failover events since the longest path between failure link l and the β node could overlap the longest path from node β to node u. For instance, suppose the link between nodes 0 and 1 shown in Figure 6 fails, the longest path between node 4 and failure node 1 is 2. However, the convergence delay can be as long as $4 \times MRAI$ because the length of the longest path between nodes 1 and 2 (a β node) is 2, and the length of the longest path between nodes 2 and 4 is also 2.

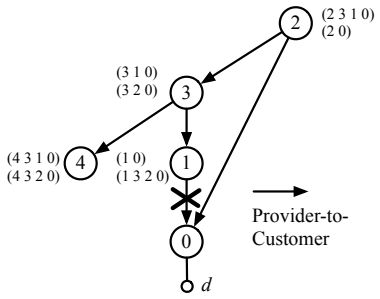


Fig. 6. A simple topology with commonly applied routing policy.

VI. Measuring Transient Failures

In this section, we measure transient routing failures in the Internet. We study the extent to which transient routing failures occur in the Internet by examining routing data collected from a number of tier-1 and tier-2 ISPs. In addition, we investigate the impact of routing policies on transient failures and the popularity of prefixes experiencing transient failures.

A. Measurement Infrastructure

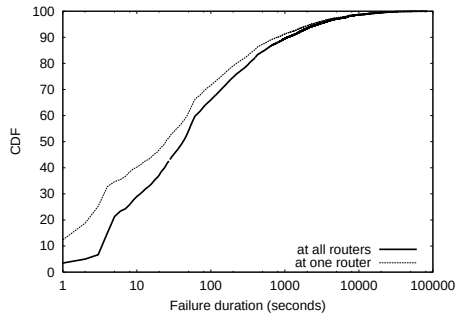
Our measurement uses both public and proprietary data. In particular, we collect the following two sets of data. The first set of data contains the BGP updates, routing table snapshots, and configuration files from a large tier-1 commercial IP backbone with hundreds of edge routers connecting to customer and peer networks. The routing updates are collected by using a BGP monitor that has iBGP sessions (running over TCP) to some top-level backbone routers and to edge routers connecting to peer networks. A snapshot of BGP routing table and the configuration file from each router are collected on a daily basis. These data were collected over the period of 20 days in July 2004.The second set of data contains BGP updates from Oregon RouteViews collected over the entire month of July 2004. We select 4 tier-2 ASes, which are inferred based on the inferred AS relationships [6].

B. Routing Failures in the Internet

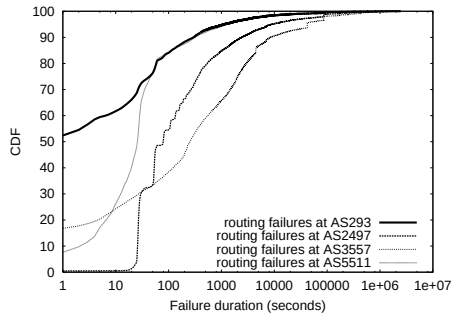
We consider the router experiencing routing failure to reach the prefix if we observe a withdrawal message for a prefix at that router. We compute the duration of a routing failure as the time period between the withdrawal message and the first announcement of the prefix after the withdrawal.

We analyze routing updates collected from the tier-1 AS and observe that the monitored routers (52 routers) experience a large number of routing failures. Figure 7(a) shows the cumulative distributions of routing failure duration at all monitored routers and one backbone router (the x-axis is plotted in log-scale). We observe that most of the routing failures are short-lived. More than 60% of routing failures last less than 100 seconds, and about 50% of failures last less than 30 seconds. This observation holds for both backbone and edge routers.

Similar observation also holds on tier-2 ASes. Figure 7(b) shows the cumulative distribution of the failure duration of routing failures for four tier-2 ASes. We observe that majority of routing failures last less than 100 seconds.



(a) Tier-1 AS



(b) Tier-2 ASes

Fig. 7. Cumulative distribution of routing failure duration.

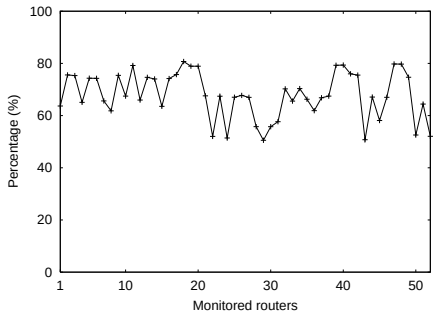


Fig. 8. Percentage of candidate transient failures among all routing failures that last less than 30 seconds.

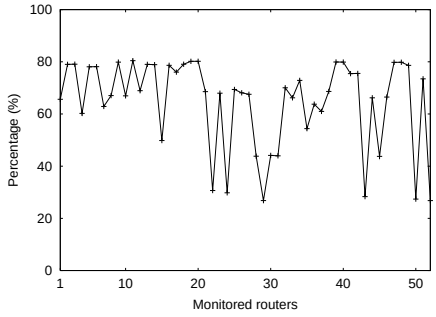


Fig. 9. Percentage of candidate transient failures that are verified.

C. Transient Failures in Tier-1 Network

From Figure 7, we find that the Internet has experienced a large number of path failures. Now we analyze transient failures in tier-1 network as following steps.

1) Identifying Candidate Transient Failures: Among all the observed routing failures, we first identify candidate transient failures. If the paths used before and after a short-lived routing failures (i.e., failures that last less than 30 seconds) are learned from different edge routers, we consider the failure as a *candidate* transient failure. Otherwise, if the paths before and after a routing failure are learned from the same edge router, we consider the failure as *faildown* because there is no alternate path to reach the destination during the failure. Here, we focus on short-lived routing failures. The reason is that an edge router can take as long as $4 \times MRAI$ time to obtain an alternate path from another edge router, and the MRAI timer among iBGP session is 5 seconds. The duration is bounded by Theorem 4. If we consider additional MRAI time delay between the edge router and the monitor, at the worst case, the control plane transient failure duration could last about 30 seconds.

For each routing failure, we define the *best path before failure* as the path in the last route announcement before the withdrawal. The path has the highest local preference or shortest AS path, and should be stable for certain period of time (e.g., 5 minutes). If we are not able to find the best path before a failure, we consider that failure is related to a previous failure and ignore it in our analysis. According to our definition of transient failure, we consider the *path after failure* as the first path after the failure.

Figure 8 shows the percentage of candidate transient failures among all routing failures that last less than 30 seconds. We find that 55% to 85% of the routing failures observed at each monitored router are candidate transient failures. The result indicates that majority of short-lived routing failures can be transient failures.

2) Verifying Candidate Transient Failures: We examine if candidate transient failures are due to failover event. For each candidate transient failure, we check if there is an edge router that still has available path to the destination at the time the failure occurs. We consider the time period from 70 seconds before the failure to 70 seconds after it. If there is an edge router that still has an available path to the corresponding destination, we call the candidate transient failure as a *verified transient failure*. Figure 9 shows the percentage of verified transient failures out of all the candidate transient failures. We find that at most of routers, about 60% of candidate transient failures can be verified. In the remaining of the paper, the analysis is conducted on the verified transient failures.

Figure 10 shows the cumulative distribution of the failure durations for the verified transient failures occurring at all routers and at one backbone router. We observe that for both cases, transient failures are short-lived. More than 95% of transient failures last less than 10 seconds, and more than 85% of them last less than 5 seconds.

Note that prefixes can be nested and both the supernet and subnet can exist in the routing table. For example, 10.1.16.0/20 is a subnet of 10.1.0.0/16, and 10.1.0.0/16 is the *supernet* of 10.1.16.0/20. If there is a withdrawal on the route corresponding to the subnet, the destination may still be reachable via the routes corresponding to the supernets. Therefore, the subnet will not experience routing failure. The subnet will experience

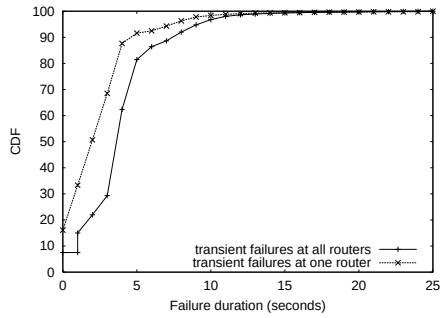


Fig. 10. Cumulative distribution of failure duration for verified transient failures.

transient failure only when all the supernets are not available. We examine whether those prefixes that experience transient failures have supernets in the routing table. We find that more than 78% of prefixes that experience transient failures do not have supernet in the routing table. For those prefixes that have supernets (about 22%), we find that all of their supernets will experience at least one routing failures during our measurement. It means that those prefixes are still likely to experience transient failures.

D. Candidate Transient Failures in Tier-2 ASes

According to Theorem 8 in Section V, transient failures at tier-2 ASes under commonly applied routing policy have the following property. Before the failures, tier-2 ASes use the paths through their customer networks to reach the destination, and after the failures, they use the paths through their provider networks (i.e., tier-1 ASes) to reach the destination. If a routing failure has such path change pattern, we consider that failure as *candidate* transient failure in tier-2 ASes.

Table I illustrates the number of candidate transient failures, i.e., the prefixes which change their best paths from paths learned from customer networks to paths learned from provider networks after routing failures, in 4 tier-2 ASes. Figure 11 shows the cumulative distribution of the failure duration of routing failures and candidate transient failures in AS D. The curves for the other tier-2 ASes are similar and are not shown here. We observe that majority of the routing failures and candidate transient failures last less than 100 seconds. Since we do not verify if a routing failure is failover or permanent failure, we observe that some routing failures last more than several days. Almost all candidate transient failures last less than 100 seconds. This duration is bounded by Theorem 4, i.e., there are 2 $\times$ MRAI time (30 seconds) applied at the eBGP session between the tier-2 AS and the tier-1 AS, and 4 $\times$ MRAI (5 seconds) applied at the iBGP session within each AS. Note that we are not able to verify these candidate transient failures due to the limited routing information available from the provider networks. In the rest of the paper, we will focus on those verified transient failures occurring at tier-1 routers.

E. Impact of Routing Policies on Transient Failures

In Section V, we study how commonly applied routing policies affect transient failures. However, a tier-1 AS may

TABLE I

CANDIDATE TRANSIENT FAILURES IN TIER-2 ASES.

AS	Routing failures	Candidate transient Failures (%)
A	1354781	169945 (12.5%)
B	431897	14313 (3.3%)
C	289439	113316 (39%)
D	1379213	355877 (25.8%)

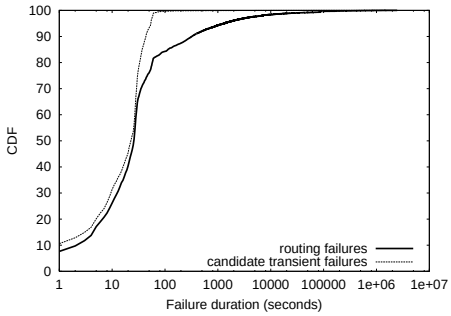


Fig. 11. Cumulative distribution of the failure duration of routing failures and candidate transient failures in AS5511.

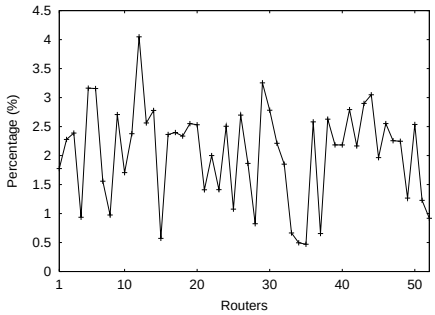


Fig. 12. Percentage of transient failures that violates *prefer-customer* routing policy.

be configured to prefer a path from its peer over the path from its customer (e.g., backup routing policy). For all verified transient failures, we examine if those transient failures are caused by violation of commonly applied routing policies. We examine the local preference and AS path length of the best route before the failure with the route after the failure, and the type of routers (i.e., peer routers or access routers) from which the paths are learned. We consider a router is configured with prefer-customer routing policy if it always assigns higher local preference to the access router than to the peering router. Otherwise, if the router assigns higher local preference to the peering router, we consider it as violation of prefer-customer routing policy. Figure 12 shows the percentage of transient failures that violate prefer-customer routing policy at each router. We find that only a small fraction of transient failures is due to violation of prefer-customer routing policy.

F. Popularity of Prefixes Experiencing Transient Failures

We measure the popularity of a prefix by the traffic volume that prefix carries at the tier-1 AS. We aggregate the NetFlow data collected in the tier-1 AS during the week (1/2/2005 $\sim$ 1/8/2005) and compute prefix-level traffic statistics. For

each destination prefix, we compute the percentage of the total traffic destined to that prefix. We find that more than 30% of prefixes experience routing failures, and about 10% of prefixes experience transient failures. We also observe that the prefixes experiencing transient failures carry about 4% of the total traffic.

We further rank prefixes in the reverse order of the traffic volume they carry. Figure 13(a) shows the scatter plot of the rank of prefixes experiencing transient failures (y axis) versus the number of transient failures those prefixes experience (x axis). Here, we normalize the number of transient failures. Each point corresponds to a prefix that experiences a transient failure. We observed that both popular and unpopular prefixes can experience transient failures. Compared with unpopular prefixes, popular prefixes are likely to experience fewer transient failures. Figure 13(b) shows the scatter plot of traffic volume of prefixes experiencing transient failures (y axis) versus the number of transient failures those prefixes experience (x axis). The number of transient failures and traffic volume for each prefix are also normalized. We observe that there is no correlation between the number of transient failures and the traffic volume of a prefix. For example, a prefix that carries more than 1% of the total traffic volume may also experience a large number of transient failures. Figure 13(c) shows the scatter plot of traffic volume of prefixes experiencing transient failures (y axis) versus the total duration of transient failures those prefixes experience (x axis). Similarly, the total duration of transient failures for each prefix is normalized. Again, there is no correlation between the total duration of transient failures and the traffic volume of a prefix. Popular prefixes may also experience long duration of transient failures.

VII. Related Work

Previous studies focus on understanding of stability of inter-domain routing and measuring end-to-end path performance. Several abstract models [6]–[10], [13] for routing convergence properties aims to capture the long-term routing stability, and ignore the details of the transient behavior. Griffin et al. show that routing policy conflicts could lead to protocol divergence and characterize sufficient conditions for BGP route convergence [9], [10], [13]. Gao and Rexford [6], [8] exploit AS commercial relationships to ensure the convergence of the BGP system. Our model differs from these existing models in the sense that we aim to capture the transient behavior of BGP, and identify the potential of transient routing failures.

Labovitz *et al.* analyze the convergence delay of BGP and derive theoretical upper and lower bounds for the convergence delay [14]–[16]. However, their works focus on the convergence delay when a network prefix becomes available or unavailable. Obradovic developed a real-time BGP model to analyze convergence delay under the hierarchical AS relationships [18]. Our work focuses on the convergence delay during the path exploration process of a link failure event.

Correlation between end-to-end path failures and routing instability have been studies through measurement. Paxson identified Internet failures and discovered that routing instability can disrupt end-to-end connectivity [19]. Feamster *et al.* studied the location and duration of end-to-end path failures and correlated end-to-end path failures with BGP routing instability [3]. Their results show that most path failures last less than 15 minutes and most failures that coincide with BGP instability appear in the network core.

Teixeira *et al* [22] found that routing changes are responsible for the majority of the large traffic variations within a large ISP network. Routing failure within an AS has been studied in [1] by characterizing failures that are correlated with IS-IS routing updates. Our work complements this study by focusing on interdomain routing failures.

VIII. Conclusions

In this paper, with the aid of a formal BGP model, we investigate the nature of transient behavior of the inter-domain routing protocol. We find that network changes that do not cause prefixes unreachable might still cause transient loss of reachability for these prefixes. Both the analytical and measured results show the existence of such transient routing failures in today's Internet routing system. Our measurement results demonstrate that more than half of the observed failures are transient routing failures. These routing failures can last up to 100 seconds and affect both the unpopular and popular prefixes. Therefore, transient routing failures can have a significant impact on the end-to-end performance in the Internet.

Acknowledgements

We are grateful to Jay Borkenhagen, Jennifer Rexford, Aman Shaikh, Oliver Spatscheck, and Zhi-Li Zhang for their valuable comments and suggestions. The work is partially supported by NSF grants CNS-0325868, ANI-0208116, ANI-0085848, and the Alfred P. Sloan Fellowship.

References

[1] Boutremans, C., Iannaccone, G., Bhattacharyya, S., Chuah, C., and Diot, C. Characterization of Failures in an IP Backbone. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop* (November 2002).

[2] Chang, D. F., Govindan, R., and Heidemann, J. The temporal and topological characteristics of BGP path changes. In *Proceedings of IEEE ICNP* (November 2003).

[3] Feamster, N., Andersen, D., Balakrishnan, H., and Kaashoek, M. Measuring the Effects of Internet Path Faults on Reactive Routing. In *Proceedings of ACM SIGMETRICS* (San Diego, CA, June 2003).

[4] Feamster, N., Winick, J., and Rexford, J. A Model of BGP Routing for Network Engineering. In *Proceedings of ACM SIGMERTRICS* (2004).

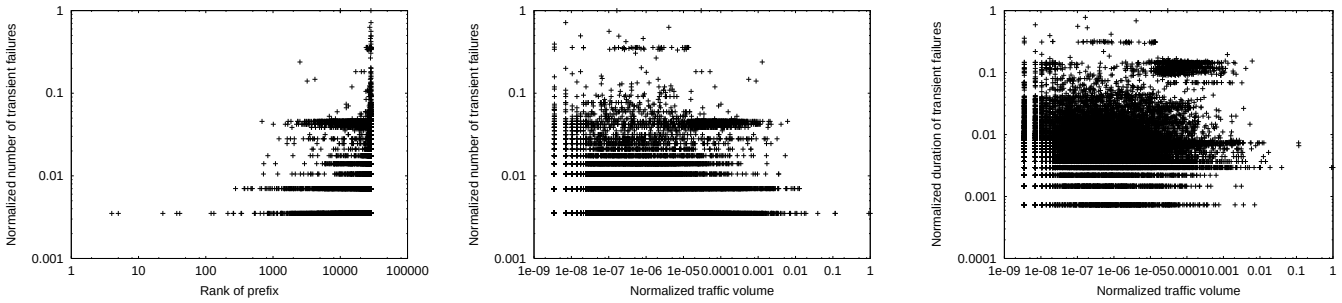
[5] Feldmann, A., Maennel, O., Mao, Z. M., Berger, A., and Maggs, B. Locating Internet Routing Instabilities. In *Proceedings of ACM SIGCOMM* (2004).

[6] Gao, L. On Inferring Autonomous System Relationships in the Internet. *IEEE/ACM Transactions On Networking 9*, 6 (December 2001).

[7] Gao, L., Griffin, T., and Rexford, J. Inherently Safe Backup Routing with BGP. In *Proceedings of IEEE INFOCOM* (2001).

[8] Gao, L., and Rexford, J. A Stable Internet Routing without Global Coordination. *IEEE/ACM Transactions On Networking 9*, 6 (December 2001), 681–692.

[9] Griffin, T., Shepherd, F. B., and Wilfong, G. T. Policy Disputes in Path-Vector Protocols. In *Proceedings of the IEEE ICNP* (Toronto, Canada, November 1999).



(a) (b) (c)

Fig. 13. Scatter plots of (a) the rank of prefixes experiencing transient failures versus the number of transient failures experienced by those prefixes, (b) the traffic volume of prefixes experiencing transient failures versus the number of transient failures experienced by those prefixes, and (c) the traffic volume of prefixes experiencing transient failures versus the total duration of transient failures experienced by those prefixes.

[10] GRIFFIN, T., AND WILFONG, G. T. A Safe Path Vector Protocol. In *Proceedings of IEEE INFOCOM* (2000), pp. 490–499.

[11] GRIFFIN, T. G., JAGGARD, A. D., AND RAMACHANDRAN, V. Design Principles of Policy Languages for Path-vector Protocols. In *Proceedings of ACM SIGCOMM* (August 2003).

[12] GRIFFIN, T. G., SHEPHERD, F. B., AND WILFONG, G. The Stable Paths Problem and Interdoman Routing. *IEEE/ACM Transactions on Networking 10*, 2 (April 2002), 232–243.

[13] GRIFFIN, T. G., AND WILFONG, G. T. An Analysis of BGP Convergence Properties. In *Proceedings of ACM SIGCOMM* (Cambridge, MA, August 1999).

[14] LABOVITZ, C., AND AHUJA, A. The Impact of Internet Policy and Topology on Delayed Routing Convergence. In *Proceedings of IEEE INFOCOM* (Anchorage, Alasks, April 2001).

[15] LABOVITZ, C., AHUJA, A., BOSE, A., AND JAHANIAN, F. Delayed Internet routing convergence. *IEEE/ACM Transactions on Networking 9*, 3 (June 2001), 293–306.

[16] LABOVITZ, C., AHUJA, A., AND JAHANIAN, F. Experimental Study of Internet Stability and Backbone Failures. In *Proceedings of FTCS* (1999), pp. 278–285.

[17] LABOVITZ, C., MALAN, G. R., AND JAHANIAN, F. Internet Routing Instability. *IEEE/ACM Transactions on Networking 6*, 5 (1998), 515–528.

[18] OBRADOVIC, D. Real-time Model and Convergence Time of BGP. In *Proceedings of IEEE INFOCOM* (2002).

[19] PAXSON, V. End-to-end routing Behavior in the Internet. *IEEE/ACM Transactions on Network 5*, 5 (1997), 601–615.

[20] SHAIKH, A., AND GREENBERG, A. OSPF Monitoring: Architecture, Design and Deployment Experience. In *Proceedings of USENIX Symposium on Networked Systems Design and Implementation* (March 2004).

[21] SOBRINHO, J. Network Routing with Path Vector Protocols: Theory and Applications. In *Proceedings of ACM SIGCOMM* (August 2003).

[22] TEIXEIRA, R., DUFFIELD, N., REXFORD, J., AND ROUGHAN, M. Traffic Matrix Reloaded: Impact of Routing Changes. In *Proceedings of PAM'05* (BOSTON, MA, March 2005).