Origins of Internet Routing Instability

Craig Labovitz, G. Robert Malan, and Farnam Jahanian
University of Michigan
Department of Electrical Engineering and Computer Science
1301 Beal Ave.
Ann Arbor, Michigan 48109-2122
{labovit, rmalan, farnam}@eecs.umich.edu

Abstract— This paper examines the network routing messages exchanged between core Internet backbone routers. Internet routing instability, or the rapid fluctuation of network reachability information, is an important problem currently facing the Internet engineering community. High levels of network instability can lead to packet loss, increased network latency and time to convergence. At the extreme, high levels of routing instability have led to the loss of internal connectivity in wide-area, national networks. In an earlier study of inter-domain routing, we described widespread, significant pathological behaviors in the routing information exchanged between backbone service providers at the major U.S. public Internet exchange points. These pathologies included several orders of magnitude more routing updates in the Internet core than anticipated, large numbers of duplicate routing messages, and unexpected frequency components between routing instability events. The work described in this paper extends our earlier analysis by identifying the origins of several of these observed pathological Internet routing behaviors. We show that as a result of specific router vendor software changes suggested by our earlier analysis, the volume of Internet routing updates has decreased by an order of magnitude. We also describe additional router software changes that can decrease the volume of routing updates exchanged in the Internet core by an additional 30 percent or more. We conclude with a discussion of trends in the evolution of Internet architecture and policy that may lead to a rise in Internet routing instability.

I. Introduction

MUCH to the consternation of the popular press, the imminent "death of the Internet" has yet to materialize [16]. Overall, the Internet has proven remarkably robust. Underlying advances and upgrades in Internet hardware and software infrastructure have forestalled the most serious problems of bandwidth shortages and a periodic lack of router switching capacity.

Though declarations of the Internet's collapse may be premature, managing wide-area networks still remains a significant challenge for network engineers. Although the theoretical properties of routing algorithms have been well studied, the deployed, or actual, behavior of routing protocols has gone virtually without formal analysis. Recent studies of both widely deployed host protocol implementations, and the behavior of routing protocols in operational networks, have shown that deployed behavior of protocols often can differ drastically from expected, theoretical behaviors [12], [5], [19]. Moreover, the scale and complexity of large protocol deployments often introduces side-effects not seen in smaller deployments, such as self-synchronization or orders of magnitude more traffic than anticipated [6].

Routing instability, commonly referred to as route flaps, significantly contributes to poor end-to-end network performance and degrades the overall efficiency of the Internet infrastructure. Routing instability, informally defined as the rapid change of network reachability and topology information, has a number of origins, including router configuration errors, transient physical and data link problems, and software bugs. All of these sources of network instability result in a large number of routing updates that are passed to the core Internet exchange point routers. Network instability can spread from router to router and propagate throughout the network. At the extreme, route flaps have led to the transient loss of connectivity for large portions of the Internet. Overall, instability has three primary effects: increased packet loss to unstable destinations [1] delays in the time for network convergence, and additional overheard (memory, CPU, etc.) within the Internet infrastructure.

In our previous work on routing instability, we observed several orders of magnitude more routing updates messages exchanged between core Internet backbone routers than anticipated [12]. Although the Internet core only maintains reachability information for approximately 55,000 prefixes, we observed between three and six million routing prefix updates each day. On average, this accounted for 125 updates per network on the Internet every day. Our data showed that on at least one occasion, the total number of updates exchanged at the Internet core exceeded 30 million per day. This aggregate rate of instability can place a substantial load on recipient routers as each route may be matched against a potentially extensive list of policy filters and operators. A high level of Internet instability poses a significant problem for all but the most high end of commercial routers. And even high end routers may experience increasing levels of packet loss, delay, and time to reach convergence as instability increases.

As a significant finding of our previous work, we showed that the majority (99 percent) of Internet routing information was pathological and did not reflect real network topological changes. We also described unexpected, specific frequency components in the inter-arrival time distribution of routing update messages. Based on discussions with router vendors and our own analysis, we suggested a number of plausible explanations and solutions for some of these anomalous behaviors. As a result of our interaction with vendors, significant changes have been made to at least one popular commercial router BGP implementation. Internet providers have since widely deployed the updated routing software across core backbone routers. We discuss the impact of these software changes in Section IV.

Since pathological, or redundant, routing information does not affect a router's forwarding tables or cache, the overall impact of this pathological routing information may be relatively benign and may not substantially impact a router's performance. Still, it is critical to understand and characterize the deployed behavior of routing protocols for future protocol design and system architecture evolution, such as the next-generation initiatives of NGI and Internet2 [10]. The level of pathological information in inter-domain routing also adds significant complexity to the analysis of this information for network planning, and debugging.

The work described in this paper extends our earlier analysis by identifying the origins of many of the pathological Internet

Supported by National Science Foundation Grants NCR-9710176 and NCR-9612764, and gifts from both Intel and Hewlett Packard.

[1] Packet loss toward stable destinations occurs only in cache-based forwarding architectures that must periodically refresh routes to a stable destination.

routing behaviors we observed. This paper also discusses the impact of specific commercial router software changes suggested by our earlier work. The analysis in this paper is based on twenty eight months of measurements of the BGP updates generated by service provider backbone routers at the major U.S. public exchange points. Our experimental instrumentation of these exchanges points has provided significant data about the internal routing behavior of the core Internet. This data reflects the stability of inter-domain Internet routing, or changes in topology or policy among autonomous systems. Intra-domain routing instability is not explicitly measured, and is only indirectly observed through BGP information exchanged with a domain's peer.

The major analytical results of our study include:

- The volume of inter-domain routing updates has decreased by an order of magnitude since April 1997. For the first time since the end of the NSFNet, the number of BGP announcements has surpassed the number of withdrawals.
- The majority of BGP messages consists of redundant, pathological announcements.
- A growing proportion of instability stems from specific changes in Internet architecture coupled with limitations in router software and algorithms.
- Instability is not disproportionately dominated by prefixes of specific lengths.
- Persistently oscillating routes dominate the BGP traffic generated by a few Internet providers.
- We experimentally confirmed a number of the origins of pathological routing behavior postulated in our earlier work.

The remainder of this paper is organized as follows: Section II provides background on BGP and Internet routing instability. Section III describes our measurement architecture, including the RouteTracker software we deployed at five of the major US exchange points. Section IV provides analysis of our inter-domain routing data and describes several of the hardware and software pathologies we observed. Finally, Section VI evaluates the impact of some these routing behaviors and provides concluding remarks.

II. Background

This paper builds on the the analysis and background discussion provided in our earlier work. We assume the reader is familiar with the Internet architecture and BGP routing concepts discussed in [12], [11], [8].

The Internet is dominated by a handful of large Internet service providers. These national and international providers, often referred to as tier one providers, account for the majority of routes and bandwidth that comprise the public Internet. Approximately four to six thousand smaller regional networks, or tier two providers peer with the tier one providers at one or more private or public exchange points. At the end of the NSFNet in 1995, the National Science Foundation established five Network Access Points (NAPs) in the continental U.S. These large public exchange points were considered the core of the Internet where providers peered, or exchanged routing information and traffic. In the last several years, most tier one providers have migrated a significant portion of their traffic from the NAPs to a large number of topologically diverse private peering points. The tier one providers believe the new interconnections provide improved scalability, dependability and performance for their inter-domain traffic exchanged with other tier one providers. The Internet in the U.S. now includes more than 60 geographically diverse regional exchange points [1]. A significant level of traffic still transits the original NAPs, but the overall trend in Internet architecture is towards a more distributed model of inter-provider connectivity.

In an optimal, stable wide-area network, routers only should generate routing updates for relatively infrequent policy changes, the addition of new physical networks and network failures. Routes exchanged in the inter-domain protocol used by most ISPs, the Border Gateway Protocol (BGP4), have two primary origins: transit routes and internally originated routes. A transit route is a prefix learned via an external BGP peering session and re-advertised to one or more external BGP peers. An internal route is a prefix learned via an internal, intra-domain routing protocol running within the autonomous system. Internal routes may also be statically configured on internal, or border routers.

Most backbone providers use an internal variant of BGP, called internal BGP, or IBGP, to distribute externally learned BGP routes amongst all the BGP routers in an autonomous system [8]. In these IBGP backbone architectures, internal routers commonly default to the closest BGP border router for all prefixes not reachable via the autonomous system's intra-domain routing (IGP). Use of IBGP allows backbones to limit the amount of reachability information maintained on internal, or IGP, routers.

Unlike external BGP, also referred to as EBGP, all routers participating in IBGP share the same autonomous system number and participate in a complete mesh of IBGP-speaking routers. Every route distributed via IBGP may be tagged with a BGP local preference, or LOCAL_PREF, attribute to specify the preference of the route per the autonomous system's policy. IBGP peering sessions depend on IGP, or alternatively static information, to maintain reachability information between peer border routers.

In most backbones, intra-domain routing serves as the basis for much of the information exchanged in inter-domain routing with external backbones. The interaction between internal and external gateway protocols varies based on network topology and backbone provider policy. In the case where a customer network is single-homed, or only has a single path to the Internet core, providers may choose to statically route the customer. In this configuration, the route to the customer network always will remain static, or constant in the BGP inter-domain information. At the other extreme, providers may choose to inject all intra-domain information directly into BGP. In this configuration, BGP will re-announce all changes to the intra-domain path affecting the customer network. Directly injecting IGP information into IBGP, and some EGP configurations, is generally discouraged, as minor configuration errors can readily result in routing loops. As an intermediate solution, most providers aggregate intra-domain information at their backbone boundary. Multiple intra-domain routes will fall under a larger, aggregate prefix. A high level of aggregation will result in a small number of globally visible prefixes, and theoretically a greater stability in prefixes that are announced. The backbone border router will maintain a path to an aggregate super-net prefix as long as a path to one or more of the component prefixes is available. This effectively limits the visibility of instability stemming from unstable customer circuits or routers to the scope of a single autonomous system.

In addition to providing support for aggregation, most routing protocol implementations include a minimum per-peer advertisement timer on out-bound protocol announcements. This timer serves two purposes: dampen extremely high frequency oscillation, and improve the efficiency of protocol processing. Several BGP implementations, including [4], [13], use this jittered timer to coalesce multiple outbound routing updates with

shared attributes into a single BGP update message. The BGP specification recommends a 30 second timer interval for generation of transit routes advertisements, and a 10 second interval for internal route advertisements [11].

In an effort to limit the level of routing instability, a number of vendors also have implemented route dampening algorithms in their routers [21]. These algorithms "hold-down", or refuse to believe, updates about routes that exceed certain parameters of instability, such as exceeding a certain number of updates in an hour. A router will not process additional updates for a dampened route until a preset, administratively configurable period of time has elapsed. By default, the router applies the dampening algorithm solely to route announcement/withdrawal oscillations. Most dampening algorithm implementations also provide mechanisms for dampening oscillations involving specific BGP path attributes, including ASPath and MED.

Route dampening algorithms, however, are not a panacea. Dampening algorithms can introduce artificial connectivity problems, as routes dampened due to earlier instability may delay "legitimate" announcements about network topological changes. Moreover, the applicability of current inter-domain dampening algorithms is limited to path oscillations that traverse a single border router. Because of the requirement for consistency amongst BGP border routers, dampening only occurs on individual EBGP routers before routes are injected into IGP or redistributed via IBGP. Specifically, although an AS with multiple peerings will independently dampen oscillations from on each border from adjacent border routers, the routers do not communicate information about the penalty nor reach distributed consensus. We discuss dampening behaviors more in Section IV. With the exception of minimum advertisement timers, BGP implementations do not provide dampening of IGP or IBGP routing information.

Overall, our research in this paper and [12], [14] has shown that the Internet continues to exhibit high levels of routing instability despite the increased emphasis on aggregation and the aggressive deployment of route dampening technology. Further, a recent study has shown that the Internet topology is becoming even less hierarchical with the rapid addition of new exchange points and peering relationships [7]. As the topological complexity grows, the quality of Internet address aggregation will likely decrease, and the potential for instability will increase as the number of globally visible routes expands. Overall, scalability is a critical aspect of the design and implementation of the Internet routing infrastructure. The autonomous system concept provides a layer of abstraction that limits the level of globally visible policy and reachability information in the Internet. Ideally, the internal routing behavior of an autonomous systems should never be propagated beyond the autonomous system's border.

III. Methodology and Architecture

Most Internet service providers monitor routing instability through facilities provided on their backbone routers. Commercial routers, including those from Cisco Systems and Bay Networks, provide commands to log protocol processing, packet tracing and various protocol statistics to the console. For archival of routing data, routers commonly rely on UDP-based syslog or TCP vtty terminal connections to offload trace information to a remote system. Limited memory and processing power on routers generally constrain the use of router debugging facilities to limited, selective traces of protocol activity. Enabling more than a minimal level of protocol logging on production routers can significantly degrade the router's switching performance, or even render the router unreachable. The high cost of commercial routers, often tens of thousands of dollars, also limits their use as dedicated data collection or probe machines.

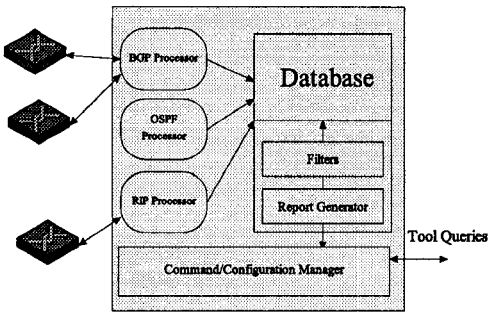


Fig. 1. Overview of the RouteTracker architecture.

In response to both the needs of network operators and researchers for software to monitor inter and intra-domain routing protocols, we developed a light-weight, distributed architecture for monitoring wide-area network protocols, called RouteTracker. We deployed the RouteTracker software on Sun Microsystem workstations at Internet exchange points and at a number of internal regional backbone nodes. Our software architecture, shown in Figure 1(a), includes two core components: routing protocol processors, and a database agent.

The protocol processing units provide light-weight implementations of the most widely deployed routing protocols, including RIPv2, BGP-4, and OSPF. The protocol processors passively participate in TCP or UDP peering sessions with neighbor routers over serial links, or local area networks. By "passively participate," we mean that RouteTracker participates in the peering session, but does not originate any new routing advertisements. The main role of the protocol agents is to record all routing protocol packets, and events, such as the loss of a peer, to local disk.

The database agent provides a means for externalizing the collected routing protocol data. Listening on a well-known TCP socket, the database agent responds to external queries for both unprocessed and processed (i.e, "cooked") reports on protocol performance. We used several tools from the MRT and IPMA toolkits [13], [9] to query the database and analyze the BGP updates collected from the exchange point backbone routers. The database agent also provides a means for replaying streams of recorded routing instability events to simulation and modeling tools. We used data recorded from Mae-East in conjunction with modeling tools to experimentally verify the possible origins of many of the anomalous routing behaviors we observed [15]. We used the IPMA and MRT software tools in conjunction with a local test-bed of more than 40 commercial routers, switches and Unix-based PC routers.

Over the course of twenty eight months, we logged BGP routing messages exchanged with RouteTracker probe machines at five of the major U.S. exchange points: AADS, Mae-East, Mae-West, PacBell, and Sprint. At these geographically diverse exchange points, network service providers peer by exchanging both traffic and routing information. The largest public exchange, Mae-East located near Washington D.C., currently hosts over 60 service providers, including ANS, BBN, MCI, Sprint, and UUNet. We also collected and analyzed IBGP routing information the state of Michigan's public Internet backbone, MichNet.

Although we analyzed data from all of the major exchange

points, we simplify the discussion in much of this paper by concentrating on the logs of the largest exchange, Mae-East. Since autonomous system border routers synchronize via IBGP, BGP information collected from Mae-East border routers should reflect the routing behavior of each autonomous system pending local router policies, and local hardware or software failures. We analyze the BGP data in an attempt to characterize and understand both the origins and operational impact of routing instability. For the purposes of data verification, we have also analyzed sample BGP backbone logs from a number of large service providers [2].

IV. Analysis

In this section, we first describe several long term trends in the overall level of Internet routing stability. We then summarize some of our previous findings and analyze the impact of specific router vendor software changes on observed Internet routing performance. Next, we describe additional pathological behavior we observed in inter-domain Internet routing and suggest additional software architectural changes. We then describe persistent routing oscillation observed in the BGP traffic from a number of providers. Finally, we show that instability remains well distributed across prefix and autonomous system space, and that instability is not related to prefix length.

A. Analysis of Gross Trends

Since our last analysis of Internet routing [12], the volume of inter-domain routing messages in the Internet core has decreased by an order of magnitude. The graph in Figure 2 shows the number of BGP announcements and withdrawals exchanged between backbone routers at the Mae-East exchange point during a 28 month period from March 1996 to June 1998. Overall, the graph depicts a dramatic decline in the aggregate volume of routing updates. The abrupt dip in trend data for the first two months of 1998 is due to a temporary data collection error (disk failure). Throughout 1996, the Mae-East exchange point averaged 3 to 5 million BGP updates every day. By summer 1998, the aggregate number of BGP updates had dropped to several hundred thousand per day. This decline is due to a significant drop in the number of pathological BGP withdrawals. In June 1996, Mae-East routers generated in excess of two million pathological withdrawals per day. In marked contrast, two years later the volume of pathological withdrawals at Mae-East consistently remained below ten thousand.

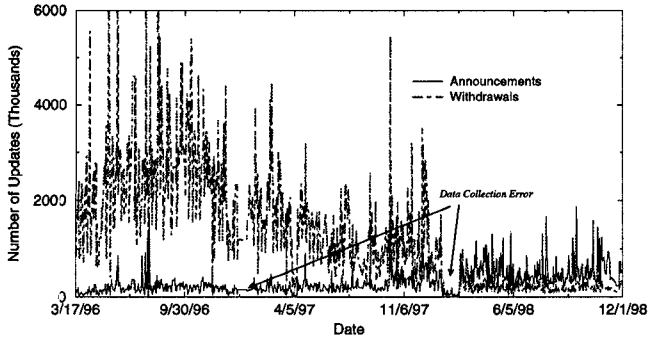


Fig. 2. Number of BGP updates at the Mae-East exchange point between March 1996 and December 1998.

In our earlier analysis of inter-domain routing [12], we showed that the majority of BGP updates (99 percent) consisted of pathological, duplicate BGP withdrawal messages. We postulated that the majority of these extraneous, pathological withdrawals stemmed from specific BGP software implementation decisions made on at least one widely deployed commercial router. In particular, we described an Internet router vendor who made a time-space tradeoff implementation decision in their routers: not to maintain state regarding information advertised to the router's BGP peers. We referred to this implementation as stateless BGP withdrawals. Upon receipt of any topology change, these stateless BGP routers transmitted withdrawals to all BGP peers regardless of whether they had previously sent the peer an announcement for the route. Withdrawals were sent for every explicitly and implicitly withdrawn prefix. At each public exchange point, this stateless BGP implementation contributed an additional $O(N * U)$ updates for each legitimate change in topology, where N is the number of peer routers and U is the number of updates. It is important to note that the stateless BGP implementation was compliant with the current IETF BGP standard [11]. After the publication of findings [12], the vendor responsible for the stateless BGP implementation updated their router operating system software. Most commercial routers today now maintain at least partial state on advertisements to their BGP peers, and will only transmit updates when topology changes affect a route between the local and peer routers.

The router vendor described above first released the updated, stateful BGP withdrawal software as a limited beta release in late 1996. Analysis of BGP data from individual ISP routers at Mae-East and private communication with theses ISPs regarding their upgrade schedules shows a direct correlation between router software upgrades and the volume of withdrawals generated by that router. The router vendor committed the software changes to their mainline product distribution in March of 1997. By summer of 1998, service providers had widely deployed the new software across Internet backbone routers. Analysis of current data from the five major U.S. exchange points show only two ISPs routers continue to exhibit stateless BGP withdrawal behavior.

Although not as dramatic as the order of magnitude change in the number of withdrawals, the number of BGP announcements per day rose by more than 60 percent over the 28 month period of our study. Throughout most of 1996, the Mae-East routers generated an average of 275,000 announcements per day. At the end of 1997, we see the start of a gradual rise in the number of announcements, ending with a mean of 427,000 announcements per day in May and June 1998. Overall, the growth in the number of BGP announcements appears to be disproportional to any corresponding increase in the number of default-free routing table entries, or the number of autonomous system paths. Graphs of the corresponding ASPath and routing table, omitted here for brevity, are available at [9]. We show later in Section IV-C that much of the increase in announcements may stem from specific Internet provider policy changes and ongoing changes in the Internet's topology.

As of February 1998, the number of announcements per day at Mae-East surpassed the number of withdrawals for the first time since the end of the NSFNet. Although we omit the analysis for brevity, this trend holds true across all the exchange points we monitored during our study. On average, exchange point routers in 1998 generated only half (54 percent) the number of withdrawals as the number of announcements.

[2] Additional data was supplied by Verio, Inc., ANS Communications, and the statewide networking division of Merit Network, Inc.

B. Analysis of Routing Update Categories

In this section, we turn our attention to the specific elements of topological and policy information conveyed in BGP routing updates. We show that pathological information still dominates Internet routing and that some of the pathologies derive from specific aspects of router vendor software implementations. We first review our taxonomy for discussing the different categories of BGP update information, and then posit a number of explanations for the trends and anomalous routing behavior we observed.

In this paper, we analyze sequences of BGP updates for each (prefix, peer) tuple over the duration of our twenty eight month study. As we describe later, the majority of BGP updates from a peer for a given prefix exhibit a high temporal locality of reference, usually occurring within several minutes of each other. In these sequences of updates for a given (prefix, peer) tuple, we identify five types of successive events:

AADiff: A route is implicitly withdrawn and replaced by an alternative route as the original route becomes unreachable, or a preferred alternative path becomes available.

AADup: A route is implicitly withdrawn and replaced with a duplicate of the original route. We define a duplicate route as a subsequent route announcement that does not differ in any BGP path attribute information. AADup is pathological behavior.

WWDup: The repeated transmission of BGP withdrawals for a prefix that is currently unreachable. WWDup is pathological behavior.

Tup and Tdown: Fluctuation in the reachability for a given prefix. An announced route is withdrawn and transitions down (Tdown), or a currently unreachable prefix is announced as reachable and transitions up (Tup).

As we described in the previous section, WWDup dominated Internet BGP routing information from 1996 until the beginning of 1998. For clarity, we omit WWDup withdrawals from the discussion in the remaining sections. In Figure 3, we show BGP announcements at Mae-East broken down into the four remaining categories by their percentage of the overall number of updates generated each day at the exchange. Errors in our data collection architecture account for the linear slopes and plateaus in the data, including the days between 10/97 and 11/97, and between 12/97 and 1/7/98.

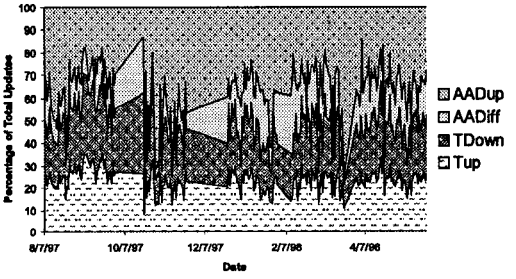


Fig. 3. Breakdown of BGP updates at the Mae-East exchange point.

Over the course of our study, fluctuation in prefix reachability information (Tup and Tdown) accounted for over forty percent of all non WWDup BGP traffic. Moreover, the percentage of Tdown transitions was roughly equal to the percentage of Tup transitions. The correspondence between Tdown and Tup is somewhat reassuring as it suggests that the majority of occasions when a prefix became unreachable, it was later readvertised as reachable. An average of ten to fifteen percent of BGP updates were AADiff. We discuss the probable origins of AADiff behavior in the next section.

After January 1998, AADup comprised the single largest category of BGP update information. In particular, duplicate BGP announcements accounted for 30 to 40 percent of the total number of BGP updates generated each day. Analysis of our data indicates that the AADup behavior was well-distributed across Internet service provider routers. Through our analysis and ongoing discussions with vendors, we have found that the the majority of AADup behavior may stem from two specific elements of the BGP software implementation on a widely deployed router operating system: non-transitive attribute filtering and the combination of a BGP minimum advertisement timer with stateless BGP.

All BGP route announcements include a number of associated path and policy attributes. Some of these attributes, such as community and ASPath are transitive, while attributes including MED and LOCAL_PREF are non-transitive. In general, non-transitive BGP attributes only have meaning within the context of the a single autonomous system's policy. The BGP specification requires that autonomous systems never reannounce, or propagate, non-transitive attributes to other autonomous systems [11]. Internet backbone architectures generally use statically configured policy filters to set the LOCAL_PREF attribute value on received routes, and the MED value on routes sent to another autonomous systems. Although both the MED and LOCAL_PREF attributes may impact the route selection process of a local autonomous system, changes in LOCAL_PREF or MED only indirectly affect the attributes associated with the route announced to external BGP peers. For example, based on the local preference of a route, an ISP may decided to announce a different nexthop for a given network destination.

Analysis of our data and ongoing discussion with vendors, indicates that under certain conditions at least one commercial router's software implementation generates duplicate EBGP announcements, or AADups, as a result of changes to non-transitive attributes. In this implementation, a border router participating in IBGP or IGP peering detects a change in one or more non-transitive route attributes. The router then marks the route as changed in its internal routing information base, and subsequently re-advertises the route to external peers – filtering the non-transitive attribute. The border router does not detect that the filtered route – absent the changed attributes – is a duplicate of the previously advertised route. We will subsequently refer to this implementation as filtered non-transitive attribute announcements. We experimentally verified the origins and impact of the non-transitive attribute software implementation on a number of commercial routers in our testbed.

In addition to non-transitive attribute filtering, a second probable source of AADup may the combination of BGP minimum advertisement timers with stateless BGP implementations. As described in Section II, most BGP implementations include a minimum 30 second advertisement timer on out-bound BGP advertisements to each external peer. Under certain conditions, multiple changes in policy, or reachability for a given prefix during the timer interval may result in the generation of only a single BGP update at the end of the interval. For example, a ASPath route oscillation with a five second periodicity will share the same final ASPath attribute as the route at the beginning of the interval. A stateless BGP implementation will not detect the duplication of state and will re-advertise the duplicate route at the end of the interval. Using BGP data recorded at the exchange points in conjunction with the routing

software tools describe in Section III, we experimentally verified this AADup behavior with a number of commercial routers in our testbed.

The router vendor responsible for both the non-transitive attribute filtering and stateless BGP implementation has developed software updates to correct the problems. Experimental deployment of the updated software at an exchange point suggests the new code successful limits the generation of AADups. The vendor reports reductions in the volume of BGP routing updates generated by their routers by as much as 30 percent. This number agrees with our findings that AADups comprise approximately 30 percent of BGP update information.

C. Analysis of AADiffs

In this section, we concentrate on a single category of BGP update information: oscillation in BGP attribute information, or AADiff. A complete description of the different BGP path attributes is provided in [8]. Although the scale and complexity of the Internet make exact determination of the origins of routing behaviors difficult, we show that a growing proportion of AADiffs may derive from specific changes in provider policy and Internet topology.

Figure 4 shows a breakdown of AADiff attribute changes over the period of our study. For clarity of presentation, we cap the vertical axis at 40,000 updates. As before, errors in our data collection architecture account for the linear slopes and plateaus in the data, including the dates between 10/97 and 11/97, and between 12/97 and 1/7/98.

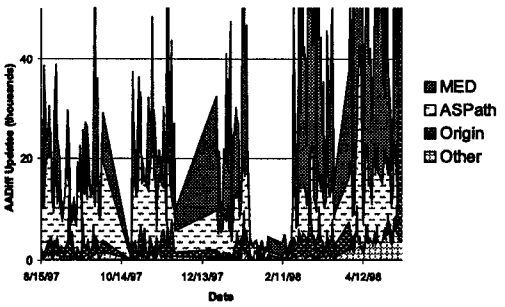


Fig. 4. Breakdown of BGP updates at the Mae-East exchange point.

The graph in Figure 4(b) shows an average ASPath AADiffs accounted for an average of 13,586 AADiffs per day, or between 20 and 30 percent of all AADiffs generated at the Mae-East exchange point between August 1997 and July 1998. The relatively low percentage of ASPath AADiffs per day was surprising. Since ASPath oscillation explicitly reflects changes in the inter-domain forwarding path of a route, we would expect ASPath AADiffs to comprise a significant proportion of all AADiffs. In contrast, oscillation in LOCAL_PREF, MED and other attributes may reflect IGP instability or changes in policy. Internet providers generally set these policy attributes using static router configuration rules, or less commonly, dynamically map the attributes using policy filters.

The "Other" category in Figure 4(b) includes oscillations in the community, aggregator, nexthop, and origin attributes. Oscillations in these attributes accounted for a combined average of approximately ten percent of all AADiffs over the course of our study. Five to ten percent of AADiffs, or an average of 2,882 per day, included the aggregator attribute. Since aggregator and atomic aggregate primarily serve as debugging and bookkeeping information, and only implicitly reflect routing changes, the volume of instability involving these attributes was surprising. Aggregator and atomic aggregator are unique amongst BGP attributes in that they do not explicitly provide forwarding nor policy information. Generally, aggregator and atomicagg AADiffs only record which router performed aggregation per an autonomous system's local policy. Analysis of our data showed that the majority of AADiffs involving aggregator and atomicagg were persistently oscillating routes. We discuss the origin and impact of the persistent route oscillations in Section VI.

Our analysis showed that the level of ASPath AADiffs remained mostly constant over the course of our study. In contrast, the number of origin, aggregator and community AADiffs per day all showed significant increases. Specifically, origin AADiffs rose from an average of 2641 per day in April 1997, to 3620 per day in April 1998. The growth in the number AADiffs for these policy attributes may stem from the architecture and policy issues discussed later in this section. The rise in community AADiffs most likely reflects the recent adoption of the attribute in several provider's policies.

Oscillations in MED constituted the single largest category of AADiffs, averaging between 25 and 40 percent of all AADiffs over the last eleven months of our study. The magnitude of MED oscillation was somewhat unexpected – service providers commonly use static configuration rules on their border routers to set the MED attribute value on route advertisements. As described earlier, providers use the MED as a hint to their preferred entry point, or border router, for routes that may transit multiple links between two adjoining autonomous systems.

The significant majority (90 percent) of MED oscillations were generated by only two large, Internet service providers. Through analysis of our data and ongoing discussions with these two providers, we traced these oscillations to specific routing policies. Specifically, the two providers dynamically map the MED value on EBGP routes based on the IGP metric for the IBGP path to the preferred border router for each route. Figure 5 shows an example of this network configuration. We will subsequently refer to this policy as IBGP mapped MED.

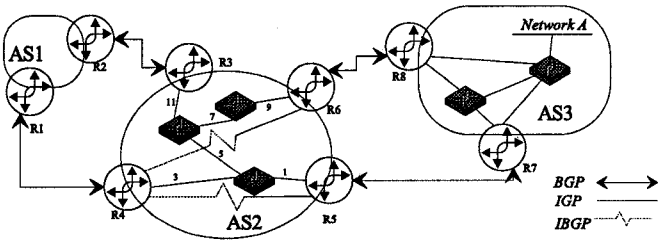


Fig. 5. Example of dynamically mapping MED values based on IGP metric.

In Figure 5, we depict an example involving three interconnected autonomous systems: AS1, AS2 and AS3. All three autonomous systems use an IGP, such as ISIS [8], to maintain reachability amongst their internal and border routers. Every IGP router has an administratively assigned metric associated with each of its interfaces. For clarity, we only show these metrics in AS2. IGP routers use the metrics in a shortest path first algorithm to select the best path to all routers and networks within the AS. For example, based on IGP link metrics, the shortest path between R4 and R5 in AS2 has a distance of 4; the path between R3 and R6 a distance of 27. In figure 5, we assume AS3 announces Network A to AS2 at only one exchange point via its border router R7. AS2 learns the route from AS3 via its border router R5. In turn, AS2 announces Network A to

AS1 through its two border routers R3 and R4. In an effort to limit the burden of transit traffic on its infrastructure, AS2 always wants traffic flowing from AS3 to AS1 to take the shortest path through its network. So, instead of setting the MED value via static configuration rules, AS2 dynamically maps the IGP distance between R5 and R3, and between R5 and R4 to the MED attribute value associated with route advertisements from routers R3 and R4 to AS1. As described earlier, the BGP MED provides a means for AS2 to influence AS1's selection of AS2's border routers for reaching Network A. In the above example, AS2 will prefer the route via R4.

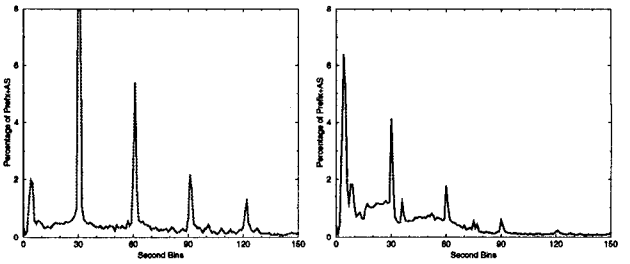
Overall, the IBGP mapped MED policy provides a powerful mechanism for reducing the traffic load on an autonomous system's infrastructure. However, the marriage of IGP metrics and BGP policy attribute information has significant implications. Specifically, previously hidden changes in IGP topology or policy now may be globally visible. For example, in Figure 5, a change in AS1's IGP reachability may affect the MED value on routes to AS3. In addition, instability in the physical links between autonomous systems, or changes to an AS's inter-domain policies also may have visibility beyond the scope of the autonomous system's direct neighbors. Again using the example in Figure 5, assume AS3 announces Network A to AS2 via its two border routers R7 and R8. If the link between R5 and R7 oscillates, R4 in AS2 will alternate between selection of the IBGP route from R6 and R5. This oscillation will result in AS2 announcing routes with different MED values to AS1 via R4. Similarly, an oscillation in the MED values associated with AS3's route advertisements to AS1 may result in an oscillation of MED values in the advertisements from AS2 to AS1. If all autonomous systems in a path use an IBGP mapped MED policy, previously localized instability events now have the potential to instigate a global chain reaction of BGP routing update changes.

More importantly, the currently deployed router dampening algorithms are not effective in suppressing these types of inter-domain oscillations. As explained in Section I, autonomous systems dampen routes only on border EBGP routers. Each border router in an AS maintains local, independent state on the instability of EBGP routes. Since autonomous system routers do not coordinate the penalties they assign to flapping routes, a round-robin oscillation of MED values amongst border routers with the same peer AS at twenty different exchange points will count only as a single oscillation on each border router. The prevalence of these types of oscillations will grow as the number of public and private exchanges increases. As an illustration of this trend, our data shows that in March 1996 Mae-East routers generated in average of 25,498 MED AADiffs per day. In May 1998, this number of MED AADiffs almost doubled, with an average of 40,834 per day.

D. Frequency

In our analysis in [12], we examined the frequency components of several categories of routing instability. To review, we defined a routing update's frequency as the inverse of the inter-arrival time between routing updates; a high frequency corresponds to a short inter-arrival time. As a significant finding of our earlier analysis, we found that the predominant frequencies of Internet routing instability have a 30 second and one minute periodicity. The fact that these frequencies accounted for half of the measured statistics was surprising. Normally one would expect an exponential distribution for the inter-arrival time of routing updates, as they might reflect exogenous events, such as power outages, fiber cuts and other natural and human occurrences. We offered a number of plausible explanations for this phenomena, including: self-synchronization, misconfiguration of IGP/BGP interactions, router software problems, CSU link oscillation and events of a higher frequency than the fixed timer interval used in at least one widely deployed commercial router.

Through additional analysis and ongoing discussions with vendors, we found that the frequency components most probably stem from a fixed minimum BGP advertisement timer used by at least one router vendor. Most Internet standards require that all protocol timers include a random jitter. This requirement stems from the danger of self synchronization. As described in [6], under certain conditions, an initially unsynchronized system of apparently independent routers may inadvertently synchronize and generate unexpectedly high aggregate traffic loads. After the publication of results [12], the vendor responsible for the unjittered timer implementation developed a software update. The vendor first released the new code in late January 1997. By summer of 1998, Internet providers had deployed the updated software on a significant number of core backbone routers.



(a) October 1997 (b) May 1998

Fig. 6. Distribution of update inter-arrival times for Prefix+AS instability measured at the Mae-East exchange point during October 1997 and May 1998.

The graphs in Figure 6 represent a histogram distribution of the inter-arrival time difference between all routing instability events for Prefix+AS pairs. We also repeated the analysis for each category of routing update sequences (AADup, AADif, Tup and Town) with similar results. The vertical axis represents the percentage of updates contained in each histogram bin; the horizontal axis represents one second bins. The graph includes data for frequency components of both routing instability and pathological updates.

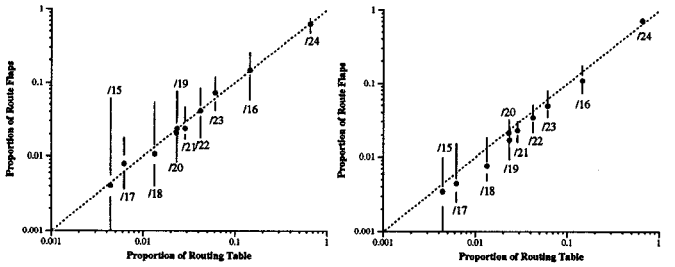
As illustrated in Figure 6, the predominant frequencies in each of the graphs are captured by the 30 second and one minute bins. The graph for October 1997 shows significantly more binning at 30 second intervals than the graph for May 1998. In addition, the graph for 1998 shows significantly less binning for all frequencies. Overall, our analysis shows that the difference in frequency strengths between the two graphs corresponds to previously described software upgrades deployed on backbone routers.

Although not shown in Figure 6, BGP routing also continues to exhibit daily and weekly cyclic trends. More discussion of these trends is provided in [12], [15], [14].

E. Prefix Length Statistics

We analyzed the data to evaluate the relationship between the prefix length of route announcements and routing instability. As described earlier, a prefix represents a set of destination

IP address blocks. The length, or network mask, of a prefix represents the number of possible subnet addresses reachable via that network address. The introduction of classless inter-domain routing (CIDR) [20] has allowed backbone operators to group large numbers of customer network IP addresses into one or more large "supernet" route advertisements at their autonomous system's boundaries. A prefix may be as specific as a single machine (32 bits), or as general as a default route (0 bits); however, in practice, aggregate prefix lengths from 8 to 24 bits are commonly used in inter-domain routing.



(a) AADiff (b) Tup

Fig. 7. This figure shows the proportion of routing updates versus prefix length for both AADiff and Tup events. The vertical axis represents the proportion of routing updates generated by a set of prefixes with identical length. For example, a /16 represents the proportion of updates generated by all prefixes of length 16. These data represent August 1996 at the Mae-East exchange.

Figure 7 shows the breakdown for two categories of instability during August 1996. The horizontal axes represent the proportion of the default-free routing table with a network mask of each length; the vertical axes represent the proportion of a day's routing updates stemming from networks of each given prefix length (both axes are logarithmically scaled). The data shown in these graphs are modified box plots: the black dot represents the median proportion of routing updates corresponding to a given prefix length for all the days in August; the vertical line below the dot contains the first quartile of daily proportions; and the line above the dot represents the fourth quartile. The dotted diagonal line represents unity. Although it is difficult to show linear fit on a log-log plot, the figure shows the rough distribution of the instability data.

Since less specific prefix masks represent larger sections of the Internet's address space, we would expect the length of a route announcement to bear some relationship to the stability of that announcement. That is, we would expect longer announcements (e.g. a /24 for a campus LAN) to be less stable than less specific announcements (e.g. a /8 covering all of a nationwide ISPs customers). Our data, however, show that instability is evenly distributed across entries in the default-free routing table. Graphically, this means that the quartile graphs in Figure 7a and 7b are bisected by the unity line. However, the data in Figure 7b shows that 24 bit aggregates represent a slightly disproportionate amount of Tup instability.

V. Previous Work

The deployed, as opposed to the theoretical, behavior of network protocols is an area of active research. A number of studies, including [12], [5], [15], [18], [19] have developed monitoring software and deployed probe machines across wide area and local networks. In [19], Paxson describes an architecture that categorizes traces of active TCP sessions to validate protocol correctness. Similarly, the Windmill tool [15] measures high-level protocols, such as BGP, in deployed settings.

In an earlier analysis of routing instability [12], we used data collected from BGP probe machines to identify a number of gross trends and pathologies in inter-domain routing information. Other studies of routing instability include the work of Chinoy, Paxson, and Govindan [3], [18], [7]. Chinoy measured the instability of the NSFNet backbone [3] in 1993. Unlike the current commercial Internet, the now decommissioned NSFNet had a relatively simple topology, different and less complex routing protocols, and heterogeneous routing technology. Chinoy's analysis did not uncover any of the pathological behaviors or trends we describe in this paper or in our earlier work. Paxson studied routing stability from the standpoint of end-to-end performance [18]. We approach the analysis from a complementary direction – by analyzing the internal routing information that will give rise to end-to-end paths. The analysis of this paper is based on data collected at Internet routing exchange points. Govidian examined similar data, but focused primarily on gross topological characterizations, such as the growth and topological rate of change of the Internet [7].

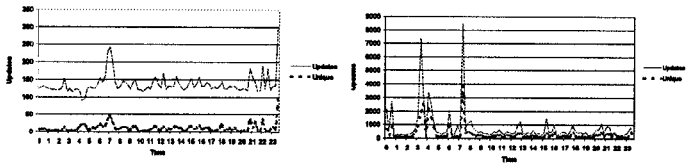
VI. Impact of Routing Instability and Conclusion

At the time of our previous analysis [12], the volume of routing update messages in the Internet core posed a significant problem for backbone providers. As described earlier, forwarding instability can have a significant deleterious impact on the Internet infrastructure. Instability that reflects real topological changes can lead to increased packet loss, delay in network convergence, and additional memory/CPU overhead on routers. Throughout 1995 and 1996, network operators routinely reported backbone outages and other significant network problems directly related to the occurrence of route flaps [17].

As a major finding of this work, we showed that volume of routing update messages decreased by an order of magnitude a between 1997 and 1998 at the five largest U.S. exchange points. Further, we demonstrated that this decrease stemmed from specific recent software changes deployed on the majority of core Internet backbone routers. Through analysis of our data and experimentation in our testbed, we found that these software changes successfully suppressed the generation of pathological withdrawals. Further, we described new router software changes currently undergoing testing that may reduce instability levels by an additional 30 percent. Overall, our data showed that both providers and vendors had addressed the most systemic problems described in [12]. Measurement and evaluation of these newly deployed routers remains an area for future research.

If we ignore the remnants of pathological information in Internet routing, our analysis showed that instability is well distributed across both autonomous system and prefix space [12]. More succinctly, no single service provider or set of network destinations appears to be at fault. Only a small percentage of the total number of BGP updates generated each day are due to persistent oscillations. We define a persistent oscillation as a sustained (usually on the order of several hours or more), high frequency oscillation in the reachability or attribute information associated with a prefix. The top line of both graphs in Figure 8 reflects the number of BGP announcements during fifteen minute intervals from two providers at Mae-East throughout the day on July 10, 1998. The bottom line shows the number of unique prefixes involved in each fifteen minute period. The graph in Figure 8(b) is typical of most Internet service providers: the majority (80 percent) of updates in each fifteen minute interval involved largely different prefixes. In other words, only a

minority of the prefixes in each interval were advertised multiple times. The graph in Figure 8(b) shows a number of exceptions to this distribution, including peaks at 3:30, 4:15 and 8:30. These peaks usually reflect brief periods of network oscillation after significant link or interface failures.



(a) Provider A (b) Provider B

Fig. 8. Number of BGP updates and unique prefixes announced over the course July 10, 1998 at the Mae-East exchange point by two large service providers.

In contrast, the graph of provider A in figure 8(a) shows a small number of persistently oscillating prefixes generated the majority of BGP update traffic for day from that ISP. Our analysis showed that Provider A's BGP traffic included 10-12 prefixes consistently oscillating between different aggregate blocks and attribute information. Overall, we found that during a four month period provider A's Mae-East router consistently generated alternating announcements for these prefixes every 30 to 75 seconds. Over the course of the day graphed in Figure 8(a), Provider A advertised several prefixes more that 7000 times each. Discussions with both these ISPs and router vendors revealed that a portion of the persistent oscillations were due to specific bugs in their router software implementations. The manufacturer of Provider A's Mae-East router indicated they were aware of the problem and are developing a software update. In general, our analysis showed that approximately 10 percent or less of BGP updates were due to persistently oscillating routes. During the course of the last week of June 1998, we traced the origins of 36 persistently oscillating routes. Of these 36 routes, approximately half of the oscillations stemmed from IGP/BGP configuration errors. The remainder of the persistent oscillations derived from a number of origins, including: policy configuration errors, additional router software bugs and oscillating hardware failures.

Our analysis suggests that the majority of BGP updates, however, likely represent "legitimate" changes in topology or policy. As a significant finding of [12], we demonstrated instability was closely tied to network usage patterns. This relationship suggests that the Internet experiences widespread network congestion or failures as bandwidth demands increase during the course of each day. A more detailed analysis of this relationship is addressed in [15], [14].

Although the number of BGP withdrawals dropped by an order of magnitude during our 28 month study, we showed that the number of announcements almost doubled during the same period. As a significant finding of our work, we demonstrated that the increase in AADiff instability stemmed in part from specific topological and policy changes in the Internet core. Specifically, we described two large providers' implementation of a MED-IGP mapping policy. Although the current level of routing instability does not pose an immediate danger of overloading the Internet core backbone routers, the trend towards coupling elements of inter and intra-domain routing hierarchy may lead to additional levels of globally visible routing instability. Overall, the loss of abstraction and hierarchy in the Internet's architecture may pose a significant risk to the future scalability of the network.

By combining simulation with direct measurements of the BGP information shared by Internet Service Providers at several major exchange points, this paper identified the probable origins of several important trends and anomalies in Internet routing instability. This work in conjunction with several other research efforts has begun to examine inter-domain routing through experimental measurements [2], [7]. These research efforts help characterize the effect of added topological complexity in the Internet since the end of of the NSFNet backbone. Further studies are crucial for gaining insight into routing behavior and network performance so that a rational growth of the Internet can be sustained.

Acknowledgments

We wish to thank Abha Ahuja, Sean Doran, Susan Hares, John Hawkinson, Masaki Hirabaru, Burt Rossi, Eric Sobosinski, and Mark Turner for their helpful comments and insight. We also thank the INFOCOM '99 anonymous referees for their feedback and constructive criticism.

References

[1] Exchange point information web page hosted by Information Science Institute (ISI), http://www.isi.edu/div7/ra/NAPs.

[2] Cooperative Association for Internet Data Analysis, home page: http://www.caida.org

[3] B. Chinoy, "Dynamics of Internet Routing Information," in Proceedings of ACM SIGCOMM '93, pp. 45-52, September 1993.

[4] Cisco Systems, Inc., Home page http://www.cisco.com.

[5] S.Dawson, F. Jahanian, and T. Mitton, "Experiments on Six Commercial TCP Implementations Using a Software Fault Injection Tool," Software Practice and Experience, vol. 27, no. 12, pp. 1385-1410, December 1997.

[6] S. Floyd, and V. Jacobson, "The Synchronization of Periodic Routing Messages," IEEE/ACM Transactions on Networking, V.2 N.2, p. 122-136, April 1994.

[7] R. Govindan and A. Reddy, "An Analysis of Inter-Domain Topology and Route Stability," in Proceedings of the IEEE INFOCOM '97, Kobe, Japan. April 1997.

[8] B. Halabi, "Internet Routing Architectures." New Riders Publishing, Indianapolis, 1997.

[9] Internet Performance Measurement and Analysis project (IPMA), http://www.merit.edu/ipma.

[10] University Corporation for Advanced Internet Development (UCAID) home page, http://www.ucaid.org.

[11] K. Lougheed and Y. Rekhter, "A Border Gateway Protocol (BGP)," RFC-1163 June 1990.

[12] C. Labovitz, G.R. Malan, and F. Jahaniam, "Internet Routing Instability," in Proceedings of the ACM SIGCOMM '97, Nice, France, MD, August, 1997.

[13] C. Labovitz, "Multithreaded Routing Toolkit – Final Report to the National Science Foundation," Merit Network, Inc. Technical Report (MERIT-960501).

[14] C. Labovitz, A, Ahuja, F. Jahanian, "Experimental Study of Internet Stability and Wide-Area Backbone Failure," University of Michigan Technical Report (number not yet assigned), December, 1998.

[15] G.R. Malan, and F. Jahanian, "An Extensible Probe Architecture for Network Protocol Performance Measurement," in Proceedings of the ACM SIGCOMM '98, Vancouver, Canada, September 1998.

[16] B. Metcalf, "Predicting the Internet's Catastrophic Collapse and Ghost Sites Galore in 1996," InfoWorld, December 4, 1995.

[17] Merit Joint Technical Staff mail archives, http://www.merit.edu/mjts/msg00078.html.

[18] V. Paxson, "End-to-End Routing Behavior in the Internet," in Proceedings of the ACM SIGCOMM '96, Stanford, C.A., August 1996.

[19] V. Paxson, "Automated Packet Trace Analysis of TCP Implementations", in Proceedings of the ACM SIGCOMM '97, Cannes, France, September 1997,

[20] Y. Reckhter, and C. Topolcic, "Exchanging Routing Information Across Provider Boundaries in the CIDR Environment," RFC 1520, September 1993.

[21] C. Villamizer, R. Chandra, and R. Govindan, "BGP Route Flap Damping," RFC-2439, November, 1998.