Inter-Domain Routing Anomalies

Jim Davis, Ski Ilnicki, Valery Kanevsky, Lance Tatman, Alex Tudor

October 15, 2001

1 Introduction

The rapid growth in Internet and Network Service Providers (ISPs, NSPs, SPs) over the past 6 years has changed the scope of Internet routing. The topology of the Internet has evolved from a hierarchy of a few regional service providers interconnected by a centrally controlled backbone to a less structured network of dozens of backbone providers interconnecting thousands of regional service providers. This progression continues today as regional service providers begin to connect not only to multiple backbone providers but to each other as well. While the physical topology of the Internet has expanded, the hardware and software requirements to support these interconnections have also expanded. The routing hardware has gone from supporting maximum line rates of 1544Kb/s to supporting rates of 40Gb/s. Router memory requirements have grown from 128KB to 256MB or more. Processing has been split between forwarding packets and processing routes in an effort to ease CPU load. Evolution on all of these fronts has required the routing protocols in use be capable of exchanging more routing data faster and to select the best paths from a larger set of possible routes. In short, the routing protocol in use today must make decisions that are

many more times complex than they have been historically and it must make them much more quickly.

The early Internet backbone (highly connected infrastructure intended to allow and encourage full connectivity) progressed from a single authority (ARPAnet, NSFnet) to a large community of independently managed domains of networks called Autonomous Systems (ASes). Those ASes connect with each other in order to provide connectivity to end-users. An AS may buy transit from other ASes and/or it may peer with others by having no-fee bilateral agreements. In reality not all ASes are equal; some can provide their peers and direct customers connectivity to vastly more of the Internet than can others. We simplify our discussions by dividing service providers (those offering transit to at least one customer) into three tiers of providers (each provider could be an individual AS or set of ASes). Tier one consists of a few international providers. Tier two consists of the many regional providers, while tier three includes several thousand local providers. There is general disagreement in the industry on which tier a particular service provider is a member of, but the three-tier model is generally accepted.

To provide connectivity across the Internet each AS must manage its own routing (reachability) information as well as exchange routing information with other ASes. Routing defects may arise within ASes (intra-domain routing) as well as from routing across ASes (inter-domain routing). Both types of routing have their own anomalies that are not always easy to detect and isolate. The inter-domain routing problems are particularly difficult to detect, isolate and repair because they are not well understood and their root causes are not easy to discover. Reasons for such difficulties lie in the fact that the Internet is a community of loosely federated independently managed large systems (ASes) where all information needed for analysis may not be always available and such analysis may face a very complex chain of symptoms and causes that are not easily separable from each other.

The de facto inter-domain routing protocol in use to route Internet traffic among various backbone and service providers, is version 4 of the Border Gateway Protocol (BGP4). This routing protocol has been extended over several years to accommodate, among other

things, the above-mentioned fundamental changes in Internet topology and it has done a commendable job in keeping things running. BGP, however, has inherent inadequacies when it comes to avoiding and repairing routing problems. BGP's design goals included the ability to disseminate reachability information in a way that would avoid some well known routing anomalies, but others were not addressed and little attention was paid to fault detection and isolation. Of primary importance here is the very nature of BGP which is to allow the routing of packets among various administrative domains, that is to say between various service providers who need not have any direct relationship with each other. BGP was designed to allow administrative domains to control the routing of data traffic within their networks via an interior gateway protocol (IGP), concealing the details from other ASes, and then to pass the traffic off to a neighboring network if it is destined for an outside network. At issue is how can one determine the root cause and location of inter-domain routing failures. *On who's network did the data fail to pass and what was the cause of the failure?*

Currently, only rudimentary tools are available to determine these causes. Traceroute and ping have proven to be of limited use, but even that usefulness is declining as providers throttle or block ICMP traffic. Determining root causes is currently a manual process involving guesswork and trial and error. Highly experienced network engineers must spend time manually looking at routing tables, remote route views, and public routing registries in an effort to track down why particular routing problems exist. Often, a problem such as sub-optimal routing can exist for long periods without being detected. Problems involving alternate routing during circuit failure may not be discovered until such circuit failures, which is precisely when proper routing is most needed. What makes this particularly disturbing is that a provider with whom one has no relationship can make changes of which one is unaware and create a real effect on the apparent behavior of your network from the point-of-view of your customers. BGP4 is a well-established and studied protocol; however, we will show that a number of routing anomalies can arise from its use.

There have been predictions time and again that the Internet was growing too complex too fast and that it would become impossible to maintain and would collapse. These

predictions of Internet meltdown have not yet occurred. Methodologies and tools for root cause analysis of inter-domain routing anomalies would significantly ease the process of managing the Internet, making it more attractive to various services. Today a lot of poor end-to-end networking performance is a result of misconfigured and overloaded routers, sub-optimal routing, etc. Existing methodologies for determining the root cause of failures are manual and very often take days to detect and isolate routing problems.

Internet growth is also hampered by its unpredictability in performance delivery making certain services unattractive. BGP's traffic engineering controls are insufficient for the deployment of many applications. All jitter dependent applications (e.g. Voice over IP VoIP) have problems in multi-AS environments. As designed BGP4 does not supply the support needed to allow service providers to provide for end-to-end guarantees of service. BGP also does not easily provide for the meshing of local SP interest with global interests; what is optimal for one provider may adversely impact others. Customer support can entail solving cross-AS problems.

In the rest of this paper, wherever specific examples of router commands are provided, these are for Cisco/IOS except where otherwise noted. Equivalent commands generally exist for other manufacturer's routers. Similarly, examples are provided of past and current bugs, and issues in Cisco/IOS; this is not intended to denigrate their products. As a market leader we simply have more information and experience with the Cisco product line and do not wish to burden this paper with a cumbersome industry summary each time reference is made to some command or feature.

The Root Cause Analysis (RoCA) project in SSL/CSD at Agilent Labs is targeted at providing our divisional partner with a prototype system for root cause analysis of one or more routing anomalies. It was felt that the project should narrow its focus by initially targeting a class of anomalies with research potential, value to Agilent's SP customers, and some hope of solution. To this end this document presents a summary of inter-domain routing anomalies that are often referred to in research publications or discussed on the North American Network Operators' Group (NANOG) e-mail distribution [NANOG]. This document is organized as follows: Section 2 of this document provides a

description of BGP. Those who are familiar with BGP may skip this section. Background information describing anomalies can be found in Section 3.

2 Introduction to Internet routing and BGP

An internetwork (or internet) is a collection of local networks interconnected by bridges or routers, while the Internet refers to the principal global internetwork. The Internet is a packet-switched network. Information is sent in relatively small units called packets. These packets are generally addressed with the IP address of their destination and the routers on the Internet forward the individual packets until they reach their destination or are discarded.

Communication on the Internet is achieved by a collection of protocols, some layered upon others. Those protocols are functionally associated with a specific layer in the Open Systems Interconnect model (OSI). The layer responsible for routing packets in a packet-switched network is called the networking layer. The dominant protocol at the networking layer is version 4 -Internetworking Protocol (IPv4), which uses 32 bit numbers (called IP addresses) to describe traffic destinations. IP allows for abstracting out many details of the underlying physical networks. In each network (i.e. connected set of hosts) there are mechanisms to allow the translation of an IP address into those data needed to deliver data on that network (e.g. 48 bit Media Access Control (MAC) address on an Ethernet). What this leaves unsolved is how to route data for an IP address to the appropriate network. For a further introduction to IP and TCP/IP (Transmission Control Protocol / Internetworking Protocol) see, for example, Charles L. Hedrick's tutorial [He87].

An Autonomous System (AS) is an interconnected set of networks under a single administrative domain. For example, the collection of networks that make up Agilent’s corporate network are all in a single AS. An AS should present a simplified view of its network and shield outside ASes from internal routing idiosyncrasies. Within an AS, routers (routing points) may use one or more interior routing protocols, and sometimes

several sets of routing cost metrics. An AS is expected to present to other ASes an appearance of a single connected network, and a consistent picture of the destinations reachable through the AS. A 16-bit Autonomous System number identifies an AS. Agilent's AS number is 14496, while BBN has AS number 1. A single organization may have more than one AS when appropriate. Hewlett-Packard has five AS numbers 71, 151, 1889, 13737, and 19647. These numbers are assigned by the international registries American Registry for Internet Numbers (ARIN), Reseaux IP Europeens (RIPE) & Asia Pacific Network Information Centre (APNIC) though there is a set of private numbers that can be used where global visibility of the AS is not needed.

Originally, the IP v4 protocol classed IP addresses into Classes A through E [RFC0990]. This standard provided for 128 Class A blocks that could contain up to 16,777,214 hosts, 16,384 Class B blocks that could contain up to 65,534 hosts, and 2,097,152 Class C networks that could contain up to 254 hosts (Class D is for multicast and E is a reserved class). This fixed division into Class A, B, and C resulted in a very sparse utilization of IP addresses. In 1993, to overcome this problem and postpone the day when Ipv4 would run out of allocable addresses CIDR (Classless Interdomain Routing) was proposed [RFC1517] [RFC1518] [RFC1519]. IP addresses are often reported in octet notation such as 130.29.240.25. Pre-CIDR networks were referred to with an address/mask pair such as HP's main network 15.0.0.0/0.255.255.255, the CIDR notation (called a prefix) for HP's main network would be 15.0.0.0/8 (or 15/8) to indicate that the first 8 bits are fixed and equal to 15 (00001111). A prefix is usually written in the format {IP address} / {number of significant bits} where the trailing zero octets of the address may be omitted. We say that the prefix 130.29.0.0/21 *matches* any IP address whose first 21 bits match the first 21 bits of the address 130.29.0.0. We refer to a prefix as longer if there are more bits specified, i.e. if it covers a smaller range of addresses; a prefix is shorter if it is more general 130.29.0.0/21 is shorter than 130.29.0.0/24.

Modern routers tend to divide their routing related information into two collections. The Forwarding Information Base (FIB) contains processed data held in a format intended to allow the router to efficiently route packets from incoming to outgoing interfaces with minimal computation. Incoming packets are matched to the most

specific (longest) prefix in the FIB. The Routing Information Base (RIB) holds the information from which the FIB is derived through possibly laborious computation.

To avoid needing routing entries for every possible Internet destination, most hosts and routers use a *default route* (some routing tables contain nothing but a single default route). A default route has a prefix of 0/0. In other words, it matches every IP address, but since there are no fixed bits, any other prefix matching would be selected because it would be a more specific prefix for that address. The default route will only be used if there are no other matches in the routing table, thus its name. Default routes are quite common, and are put to best use on networks with only a single link connecting to the global Internet. On such a network, routing tables will have entries for local nets and subnets, as well as a single default route leading to the outbound link. Large SPs use what is known as a *default free* routing table that lists all reachable prefixes in the Internet.

Service providers classify their relationship with other service providers into three disjoint classes: customers, suppliers (a.k.a. transit providers), and peers. Customers are the direct customers, or customers of customers. Peers (a symmetric, non-transitive, non-reflexive relation) agree to allow each other to route traffic among them and their customers. Direct customers of a service provider P typically pay P for transit. Buying transit is an asymmetric arrangement in which one party (A) pays another for the right to send traffic to and receive traffic from A, customers of A, the peers of A, and the customers of those peers. P is a customer to its own suppliers; it pays them for such service (customer and provider are not merely asymmetric, but anti-symmetric relations). Customers may also be referred to as downstreams while suppliers are sometimes called upstreams. Another relationship is the paid peer relationship wherein two suppliers become mutual customers of each other with billing based on the traffic differential. Note that other relationships are possible, but they are not commonly found in practice.

Routing and routing protocols can be categorized by several different criteria. *Exterior routing* occurs between autonomous systems. *Interior routing* occurs within an autonomous system. RIP and OSPF, are interior routing protocols while BGP4 is an exterior routing protocol. *Static* protocols configure their routing *ab initio* while *dynamic*

protocols adjust the FIB as routing information arrives. *Link state* protocols use information on the detailed connectivity to compute the best path to each prefix while *distance-vector* protocols abstract out detailed topology from the purview of distant routers.

There are several approaches to routing on the Internet, each appropriate in some situations.

- Static routes are manually configured routes. These can be simple or complex, but they are unresponsive to changing needs (e.g. if a link goes down then traffic may be black holed until someone notices and manually repairs the static routes). Static routes are appropriate only in very simple topologies.

- Default routing allows a router to forward all (or some) off network (non-local) traffic to a specific router. This allows a network with a single connection to avoid the complexities of being aware of Internet routing. Default routing can occur with static or dynamic routing.

- Link-state protocols such as Intermediate System to Intermediate System Intra-Domain Routeing Exchange Protocol *sic* (IS-IS), and Open Shortest Path First (OSPF) are generally used as interior routing protocols. Each router maintains at least a partial map of the network. A link state protocol communicates link information to all participants, which make routing decisions based on the detailed state of the network. Link-state protocols do not scale to allow for exterior routing on the Internet.

- Distance-Vector Routing Protocols require that each router simply inform its neighbors of its routing table. For each network path, the receiving routers pick the neighbor advertising the lowest cost, and then add this entry into its routing table for re-advertisement. DCN-Hello [RFC0891] and RIP [RFC1388] are common D-V routing protocols. Common enhancements to D-V algorithms include *split horizon, poison reverse, triggered updates,* and *holddown*. You will find a good discussion of distance-vector, or *Bellman-Ford* algorithms in RIP's

protocol specification, RFC 1058. Because they abstract away all path information distance-vector algorithms require a global understanding of the "cost" of a path, and if negative costs are allowed can create loops.

- Path vector protocols such as BGP are essentially a distance-vector algorithm, but with several added twists. BGP ignores interior details of other ASes allowing it to scale better than a link-state protocol. By passing the vector of ASes through which each route passes, BGP claims to guarantee that no route will loop between multiple Autonomous Systems. In section 3.11 we show that this claim is not always upheld.

2.1 Overview of BGP 4 protocol and implementation

BGP routers exchange routing information during BGP sessions. Each session involves exactly two endpoint routers that are called BGP peers. A participant in a BGP session is also sometimes called a BGP speaker. BGP does not have its own reliability mechanisms, but instead uses a reliable transport protocol TCP for end-to-end reliability. This ensures orderly delivery of messages, detects duplicates, and recognizes when information has been lost. Four kinds of messages are used in BGP sessions:

- An **Open** message is sent immediately after the TCP connection has been established. Its purpose is identification and mutual agreement about values of some protocol parameters, like timers.
- **Keep Alive** messages are exchanged periodically to indicate that the peer on the other side is still up and running.
- **Update** messages contain routing information.
- **Notification** messages are used to indicate errors that have occurred during the BGP session.

When both routers commit to a BGP session, they may start exchanging routing information. To improve the efficiency of routing, a single autonomous system may deploy several BGP routers. If that is the case, those routers must use BGP to synchronize the edge routers to avoid leaving some destinations temporarily unreachable (called a “black hole”). External usage of BGP is usually called E-BGP and the internal usage is called I-BGP. Routes to other ASes are first learned from some external router, via E-BGP. After that, I-BGP can be used to disseminate them to other BGP speakers in the same AS. To solve the scaling problem of requiring a full mesh of I-BGP speakers, route reflectors are used [RFC2796]. In this approach a dedicated BGP speaker advertises learned routes in a star topology to a set of other I-BGP speakers. An important property of BGP is that once advertised, routes do not need to be refreshed. They stay active until they are explicitly revoked or until the TCP connection breaks. In the latter case, each router must stop using the information it heard from the other one.

One of the most important messages is the **Update** message. The **Update** message carries withdrawals as well as announcements of routes. Route withdrawals are immediately propagated to other peers while we shall see that announcements of new routes are delayed by a configurable amount. An **Update** message consists of three parts, as shown on Figure 2.1.

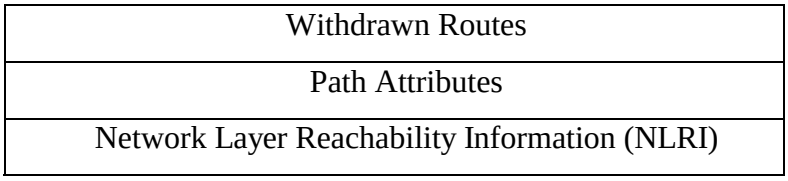


Figure 2.1: **Update** message

The Withdrawn Routes field lists the destinations prefixes for which the sending router is no longer ready to forward packets. The NLRI field lists destinations (IP prefixes) that the sender of the **Update** message can reach using some route. The route, together with additional attributes is given in the Path Attributes field. It is assumed that those attributes apply to all the destinations from the NLRI field. The

exact set of path attributes that an **Update** message contains may vary. Generally, attributes can be well known or optional. Well-known attributes must be recognized by all BGP implementations. A path may also contain a number of optional attributes. They are not specified by the BGP standard and each router uses them at its own convenience to communicate additional information about the path. Well-known attributes can be further subdivided into mandatory and discretionary attributes. Mandatory attributes must be included in every **Update** message, while discretionary attributes may or may not be included. Below are examples of some of the well-known attributes:

AS Path is a well-known mandatory attribute that describes a route. It is encoded as a list of path segments, where each segment is either an **AS Set**, which contains an unordered collection of ASes, or an **AS Sequence**, which contains an ordered collection of ASes. In practice, most routes are advertised as single **AS Sequence**s. The purpose of **AS Set**s is to allow aggregation of prefixes with different **AS Path**s.

Local Pref is a well-known discretionary attribute that describes the sender's degree of preference for the advertised route. We will see later that each router has a locally configured policy for determining this level of preference. It plays a crucial role in the route selection process. An AS may have several different exit points, and hence, several active BGP routers at a given time. In addition to exchanging information with the outside world, those routers also need to communicate among themselves. The **Local Pref** attribute will be included only in **Update** messages sent to other BGP peers in the same AS (I-BGP).

Next Hop is a well-known discretionary attribute that contains the IP address of the router that should be used as the next hop to the destinations listed in the **Update** message. This is usually the advertising router itself, but in some cases it can be some other (possibly a non-BGP) router from the advertising router's AS.

Multi Exit Discriminator (MED) is a well-known discretionary attribute. Its purpose is to discriminate among entry points to the same neighboring AS. The value of this attribute describes a certain metric, so that the routes with smaller **MED**s will be preferred. Below we give one example of the usage of **MED**.

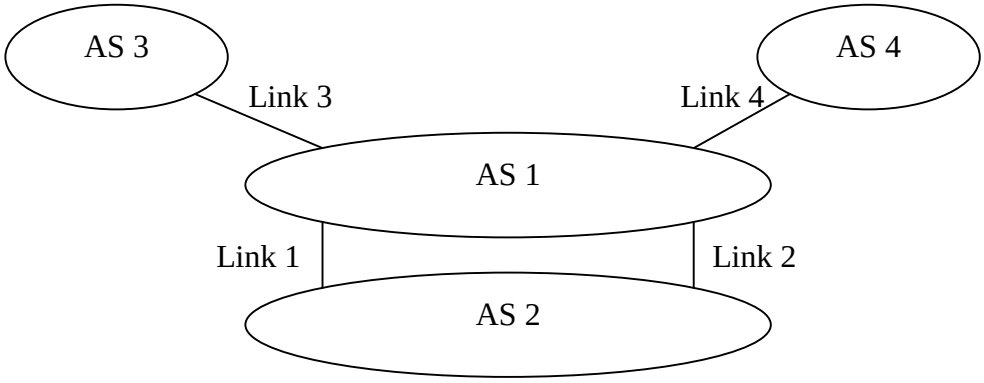


Figure 2.2

Consider the situation shown in Figure 2.2. There are four AS's and four BGP sessions (depicted as links). AS1 would like to inform AS2 that it should favor Link 1 for traffic towards AS3 and Link 2 for traffic towards AS4. Notice that AS1 advertises to AS2 routes for AS3 and AS4 on both links (1 and 2). In order to signal the preference to AS2, AS1 should advertise a route for AS3 with higher **MED** value on Link 2 and lower **MED** value on Link 1. The situation is opposite for routes to AS4: higher **MED** on Link 1 and lower **MED** on Link 2. Contrasting **MED** with **Local Pref** we see that **MED** is used in choosing where traffic will enter an AS while **Local Pref** is used in determining where traffic will exit an AS.

After an **Update** message is received with all its attributes, the receiving router needs to process it. Conceptually, every BGP router has its routing table split into three parts, usually called RIBs (Routing Information Bases):

- **Adj-RIB-In**. There is one Adj-RIB-In per peer -- it stores routing information that has been learned from that peer.

- **Loc-RIB** contains the local routing information that survived input filtering. This includes both selected and unselected routes, though not suppressed routes or those that were filtered out.

- **Adj-RIB-Out**. There is one Adj-RIB-Out per peer -- it stores the routing information that is to be advertised to that peer.

Adj-RIB-Ins contain unprocessed routing information that has been advertised to the router. A special decision process selects from Adj-RIB-Ins the routes that will be put in the router's own Loc-RIB and eventually disseminated to other peers. The process starts by computing the degree of preference for each route in Adj-RIB-In. The degree of preference is computed using the local router's preconfigured policy.

A policy can be any non-negative integer valued function that depends only on the attributes present in the Adj-RIB-In. Routes are subjected to input filtering before being entered into the Loc-RIB. It is possible that several routes to the same destination in the Loc-RIB have the same degree of preference, so the router may need to look at other attributes to break the ties. Those additional attributes are considered in the following order:

1. Routes with shorter **AS Path** attributes are preferred.

2. Routes with smaller values of the **MED** attribute are preferred.

3. Routes with shorter internal distance to the **Next Hop** are preferred. This distance can, for instance, be learned from the interior routing protocol.

4. BGP can run in the external (E-BGP) and the internal (I-BGP) mode. In this step, the preference is given to the routes learned through E-BGP.

5. The final tie is broken by selecting the route advertised by the peer with the lowest BGP identifier number.

Whenever the router's Loc-RIB changes, the changes may need to be propagated to all its peers. However, if the received **Update** message caused no changes in the Loc-RIB or output filtering eliminates the visibility of the change, no further propagation is needed. Also, whenever a new BGP router joins by establishing a BGP session, the other peer needs to advertise to the new router all the information from its Local-RIB. BGP version 4 employs five timer constants. Below are their names, interpretations and values suggested by the standard.

ConnectRetry is used during the initial phase of a BGP session, when the router is trying to establish a peering connection. It represents the maximum amount of time that a transport connection request to the other router can ”hang" without being answered. The suggested value is 120 seconds.

HoldTime denotes the maximum amount of time that can elapse between two consecutive messages (of the type **Keep Alive** or **Update**) from a given peer. If the peer has been silent for longer than HoldTime, an error **Notification** message is sent and the connection is reset. The suggested value for HoldTime is 90 seconds.

KeepAlive determines the frequency of KEEPALIVE messages sent to the peer. The value of KeepAlive is the time between two successive messages. KEEPALIVE messages must not be sent more frequently than one per second. The suggested value is 30 seconds.

MinRouteAdvertisementInterval denotes the minimum amount of time that must elapse between route advertisements for a particular destination from a single peer. The suggested value is 30 seconds.

MinASOriginationInterval denotes the minimum amount of time that must elapse between successive advertisements that report changes within the advertising router's own AS. The suggested value is 15 seconds.

3 Symptoms & Causes of Inter-domain routing irregularities

The following describes a set of inter-domain routing anomalies, in particular anomalies that result from BGP operation in the Internet. For the last several years researchers have tried to understand how BGP or inter-domain routing anomalies affect the performance and efficiency of the Internet. Some of the inter-domain routing behavior is fairly well understood; however, more research is required to understand BGP operation in a large and complex environment such as the Internet.

We have separated our discussion into fifteen areas. Some of these are clearly causes of inter-domain routing anomalies while some might be considered symptoms. In at least one case one of the symptoms listed can be caused by the methods on another section. This is neither a taxonomy nor a clean partition; it is intended to loosely categorize areas for examination and discussion.

Index of Anomalies

3.1 Mismatch between router configuration and observed behavior

This anomaly occurs when one or more routers are misconfigured relative to their announced policy; this can be an error in the routers policy, an error in the announced policy or an intentional variance intended for some corporate purpose.

Policy may be divided into several areas: route aggregation, route damping, traffic engineering, and “pure policy” (prefix and AS filtering). Configuration of a router’s policy is a complex, error-prone manual process. It is not uncommon for network engineers to exchange configuration files, which are then edited to taste and deployed. The lack of configuration validation tools (a difficult task given that configuration commands may change on every router OS release) coupled with the inability to simulate conditions that trigger configured policy, make policy maintenance slow and risky.

In this section we are going to define the concept of routing policy, outline its basic attributes, explain its purpose and present an example, which illustrates the complexity of policy configuration.

A service provider’s routing policy is a set of rules, implemented as an algorithm to reflect the reality of business needs, traffic engineering and good citizenship (i.e. considerate use of shared Internet resources). This set of rules evolves along with the change in network environment: size of the network, traffic’s volume and pattern, software and hardware upgrades, business concerns, etc. Today, changing routing policies is a complex error-prone manual process.

Business needs include, but are not limited to, custom service (the ability to serve different customers according to specific needs and criteria), efficient use of network assets, and mutually beneficial relationships with competitors. A service provider uses traffic engineering to make its network reliable, to balance volume load and to make the customers’ connections symmetrical. Symmetry is a provider’s ability to receive and send traffic to a *multi-homed* subscriber via the same link (a customer that is connected to the Service Provider (SP) with at least two links or to two or more SPs). This is an important service goal because it reduces traffic load on the customer’s network. Proper prefix

aggregation and judicious propagation of updates are examples of good citizenship (aggregation conserves the size of the default free zone routing table and controlled propagation conserves a router's BGP processing time.) Aggregation is proper if it does not defeat the intent of the inclusion of the prefix lost in the aggregation. Prefixes introduced in support of a specific policy, such as a multi-homed customer should not be aggregated. Update propagation is judicious if it does not propagate information that should remain local. For example, some versions of Cisco IOS always propagate the Community attribute, contrary to the BGP 4 specification. Policies are implemented in terms of *inbound* and *outbound filters. Inbound filters* are a mechanism to check any **Update** messages coming from peers, IGPs and static routing. It is essential to mention that filtering means both elimination of certain information regarding routing and its alteration as well. Local selections are combined with IGP selections and static information to finally produce the router's forwarding table (FIB). Every time a routing change occurs, the FIB is recomputed. Cisco IOS implements three distinct categories of filters:

- *prefix-lists* (Cisco specific configuration language) to filter *prefixes*,
- *filter-lists* to filter AS numbers (or *access-list* for both ASes and prefixes) and
- *route-maps* to apply the policy *and* make post-inbound and pre-outbound filtering modifications.

Outbound filters are applied in processing **Update**s prior to announcing updates to peers. Periodically (according to protocol timers), data from IGPs, static configuration and Adj-RIB-In is again filtered and modified and finally announced to a particular peer. Both inbound and outbound filters can be applied on a per peer basis; mechanisms exist (e.g. community attribute) to specify policy by group.

As stated previously, policy includes route aggregation, route damping, traffic engineering, and "pure policy" (prefix and AS filtering). Routing policy controls both inbound and outbound traffic, by prefix, by AS, by egress/ingress point (in the case where the AS in question is multi-homed to a single provider) and by provider (in the

case where the AS in question is multi-homed to multiple providers. The policy differs depending on the location of the router (e.g. at an exchange point, at the customer edge), whether the AS served by the router is a stub (does not provide transit, and is a customer of one or more service providers) or transit AS, and other considerations.

To illustrate the complexity of the process of policy configuration let's look at what "TelecomSP" needs to do in order to configure a single peering router, taking into consideration just one of many requirements.

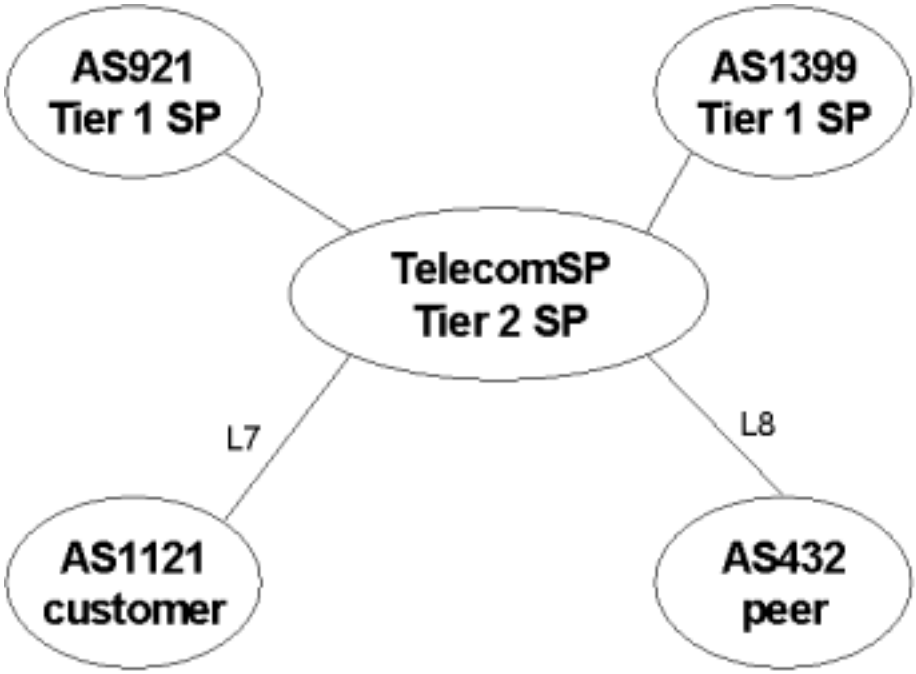


Figure 3.1

An example requirement for Telecom SP (AS7341 & AS7167) may read as follows: "**AS432** will be permitted to send packets to destinations within and receive packets originated from sources within AS7341 and AS7167. This connectivity will utilize link L8 with AS7341 and link L8 and L7 with AS1121. AS7341's primary link to AS432 is considered to be link L8. Secondary connectivity to AS432 can be provided through AS921 and AS1399." [PDG98] (Secondary connectivity not shown)

This may be translated into *easier* English to read as follows: Telecom SP is a transit provider. It is comprised of two ASes, 7341 and 7167. Telecom SP is multi-homed to

AS921 and AS1399, its upstream service providers. AS432 is a peer of AS7341. AS432 cannot use AS7341 or AS7167 for transit traffic. L8 links AS7341 to AS432. AS1121 is a BGP speaking customer of Telecom SP. L7 links AS7341 to AS1121.

Expressing this policy fragment requires configuration on multiple routers. We will show just the partial configuration of AS7341's peer-facing router (with AS432.)

- Filter inbound RFC1918 private address space etc.

```
access-list 2 deny 10.0.0.0 0.255.255.255
access-list 2 deny 172.6.0.0.0 0.15.255.255
access-list 2 deny 192.168.0.0 0.0.255.255
```

- Filter outbound routes to disallow AS432 from using AS7341 for transit (do not announce their prefixes); do not announce any prefix whose path starts with 432.

```
ip as-path access-list 20 deny ^432_
```

- Filter outbound to announce to AS921 and AS1399, AS1121 prefixes and allow AS1121 to provide transit to other ASes; this line applies to the router facing AS1121 only; it is shown here for clarity only. If AS1121 would *not* be permitted transit the last token would be _1121$ (meaning the path ends with 1121.)

```
ip as-path access-list 21 permit _1121_
```

- Filter inbound routes to allow prefixes originating in AS432; allow prefixes whose **AS Path** ends in 432 (i.e. are originated by AS432).

```
ip as-path access-list 22 permit _432$
```

- Filter inbound routes to disallow client's (AS1121) prefixes coming through peer AS432.

```
ip as-path access-list 20 deny _1121_
```

The examples above show a number of inbound and outbound filtering criteria. Here are a few of the *good practice* (aggregation, traffic engineering & damping not included) rules that are often overlooked:

- Do not announce **default** (i.e. prefix 0/0) to your peers, and do not accept **default** from your peers as this would encourage abuse of the peering relationship;

- Do not accept peers' prefixes unless their originating AS is that of your peers or their downstream to avoid allowing leaking routes to propagate;

- Do not accept your own customers' prefixes (or their customers') from your peers (Accepting prefixes belonging to other SPs from your peers means that you are using them for transit; this violates the peering agreement. Additionally, since your customer has a direct connection to you, you would service them better.);

- Do not announce your upstreams' prefixes to your peers (this includes either partial or full routes received from your upstreams);

- Do not announce prefixes with private AS numbers (you would have private AS numbers – 64512 through 65535- if your customer required BGP but did not have his own AS number);

- Keep track of your peers' and customers' Regional Internet Registry (RIR) allocations and filter on prefixes, in addition to **AS Path**;

- Keep track of new *golden prefixes* (Internet infrastructure such as DNS servers, etc.) to ensure you are not filtering them; and

- Keep track of ARIN, RIPE & APNIC and emerging RIRs' allocation policies (if you are filtering according to their rules).

Customers are gained and lost, require more or less addressing space, become multi-homed, make traffic engineering changes, the list goes on and on; every instance requires configuration changes on multiple routers, belonging to customers, service providers and peers. Many of the changes are not or cannot be localized to the adjacent neighbor only. Their effects may be far reaching.

A better way of managing policy change is needed. Errors such as simple typos, transcription errors, and failure to follow good practices can cause ASes to become unreachable, and links to be over or under utilized.

3.2 Circuitous Routing

[Pa97] defines circuitous routing as a sub-optimal (in a geographic sense) forward or reverse path that a packet takes. An example of such a situation is when a packet originating from California with an Oregon destination, passes thru Japan. There are several possible reasons for such behavior:

- Incorrect router configuration
  For example, given two possible paths for a given prefix, the wrong preference value is assigned. The resulting FIB is incorrect.

- Inadequate regional/local peering infrastructure
  Any operating SP will try to minimize transit traffic and maximize peering traffic, because peering is inexpensive, while transit has higher costs. But peering, either privately or at an exchange, does have its costs (e.g. circuit provisioning and upkeep). Before reaching the point where it is cheaper to peer than to pay transit, local/regional SPs achieve inter-connectivity via a third party transit provider. If the SPs peered locally, traffic would remain geographically local. When SPs do not or cannot peer, traffic may travel a long way to, hopefully, the nearest point where the SP's respective upstream providers peer. The proliferation of local ISPs, which may buy from different upstream providers, only makes the situation worse.

- Telephone company tariff arbitrage
  A tariff is the set of rates of a given telecommunications carrier. While the FCC Communications Act of 1996 has increased competition and liberalized tariffs, the same is not true for the rest of the world, where telephone companies (or large chunks thereof) are government monopolies (and as such a large source of revenue). Government ownership of telecommunications infrastructure can lead

to unusual pricing situations. For example, a phone call from France to Germany direct, is usually more costly than routing it through New York. As a result of this discrepancy in pricing European ISPs may buy transit thru the US (a check of the Band-X bandwidth index on 8/17/01 indicates that 2Mbps from London to Frankfurt is cheaper than London to New-York; checking pricing a year back, the New-York circuit was only somewhat higher - ~20%).

The net result of arbitrage is that packets don't take the best routes (meaning link quality, bandwidth, number of hops, directionality, etc.), but the ones that are most cost effective to the intervening NSPs. This is the intended behavior, but can appear anomalous to end-users.

- Hot potato routing policy between tier one peers

  It is generally not in the interest of tier one providers to carry traffic across their backbone for customers, which are not paying for QoS. In other words, a provider's backbone bandwidth is reserved for premium customers. So if a provider receives a packet in San Francisco with the destination New York, it will try to unload the packet at a nearby local exchange point, such as MAE-West or the PAIX, rather than transport it across its national network and unloading it at an exchange in New York. This type of routing is called hot potato (or early exit). In the interest of pursuing this policy carriers will send a packet in the opposite geographical direction from where it is going. Again this sub-optimality is the desired behavior from the SP's view, though customers may view it as an anomaly.

All of the above are long-term routing behaviors. There is another instance of circuitous routing that was observed in [Pa97]. Paxson refers to this behavior as short-term flutter. This behavior is observed when packets take alternating paths to the same destination due to load balancing. In these cases, the presence of more than a single router on a single hop may be displayed when using traceroute. This trait of routing behavior is typically designed into the network purposefully and does not introduce routing problems.

3.3 Unannounced or Improperly Announced Prefixes by Upstreams of a Provider's New Customers

This anomaly occurs when a service provider, other than a customer's paid service provider, improperly announces the customer's routes or does not announce them at all. This results in an inability for customers of that service provider to reach the improperly announced customer. This is best illustrated by example. In the case below, Agilent is a typical customer. Agilent buys Internet access services from SP1. The key in this example, and in this anomaly, is that Agilent has a portable network address. This means that Agilent was not assigned their network numbers by their service provider, SP1, but had obtained their own network numbers, which are not associated with any service provider. For Agilent, this has the advantage of allowing them to change service providers without having to renumber their network. It means they are not so tightly tied to a single service provider. However, this also means that the address is not part of their providers normal address space. Typically, a provider allocates address space out of a block of addresses that are already advertised. For example, if a provider had the class B address space 128.102.0.0 and you purchased services from this provider, you may be assigned address 128.102.18/24. In this case, the provider is likely already advertising the entire class B address space to their peers, so there is no action required by peers of your provider. However, if you went to that same provider with your own address space, there may need to be configuration changes made to your provider's *peer* routers, over which your provider has no access or control. To make matters worse, Agilent may not know there is a problem until a customer from outside Agilent and SP1, tries to access an Agilent web server. This customer then would need to know to contact Agilent to let them know their web servers were unreachable and Agilent would need to let SP1 know, who would then need to look at the routing table as viewed from outside of their own network in order to determine exactly what the problem is. One can see the potential for a long elapsed time between the creation of the anomaly and the resolution. The example below walks step by step through how this anomaly might occur, though it is simplified to cover only a single peer.

Step 1

Agilent is a customer of Service Provider SP1. Agilent advertises out to SP1 Network 192.203.230.x (A portable network address it brought with them). Assume this network is used for Agilent's External Web Servers

Step 2

SP1 fulfills its obligation to Agilent by accepting and re-advertising net 192.203.230.x to its peers and customers and routing traffic from its network and the Internet to Agilent

Step 3

SP2 has a peering arrangement with SP1, but, as a matter of security, and routing stability, they filter all routes advertised by SP1 that they have not specifically agreed to accept. Therefore, even though the route to Agilent is being advertised by SP1, SP2 is not accepting the route either because SP1 did not inform them of the change or because SP2 has not updated their filters and is therefore not re-advertising Agilent's network to its customers or its peers.

Step 4

Customers of SP2 and peers of SP2 that are not receiving the route elsewhere, are unable to reach Agilent's web server

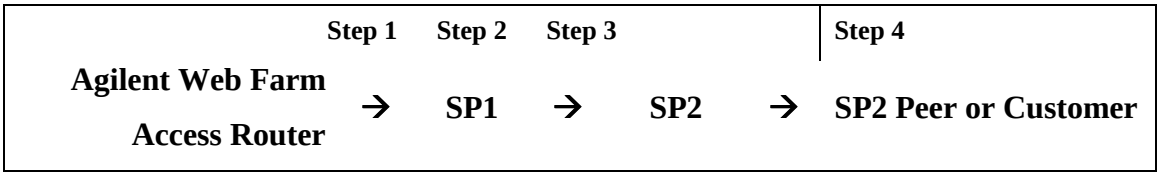


Figure 3.2

3.4 Withdrawals of Dead Prefixes and/or Duplicate Announcements of Existing Prefixes With No Attribute Change

Withdrawing a previously withdrawn route or announcing a route identical to an existing route is a waste of router resources and network bandwidth. Although the protocol

described in Section 2 provided for BGP tracking the state of the information it had communicated with each peer, some implementations failed to do this or did it improperly. This behavior was studied by Craig Labovitz and is explained in detail in [La99]. Two primary causes were found.

The first cause identified was that a particular router vendor did not maintain state regarding information advertised to the router's BGP peers. While this was within the BGP specification, implementation of BGP in this form was identified as the root cause of this problem. Labovitz explains the behavior of a stateless BGP implementation, "upon receipt of any topology change, these routers will transmit announcements or withdrawals to all BGP peers regardless of whether they had previously sent the peer an announcement for the route. Withdrawals are sent for every explicitly and implicitly withdrawn prefix" [La99, p43]. In other words, the router making the announcement or withdrawal does not keep track of whether the announcement or withdrawal is a duplicate. This leads to a significant number of duplicates [La99, p45].

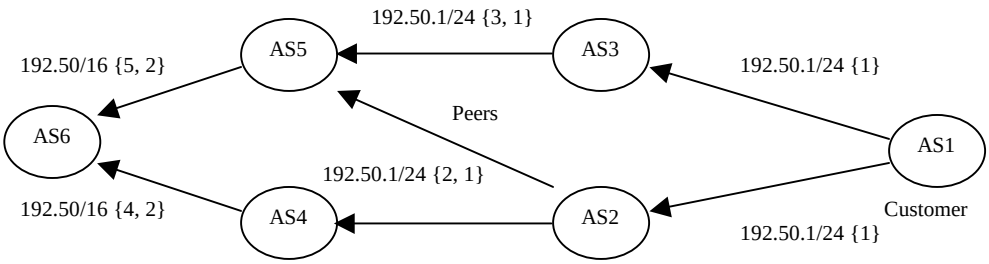
Labovitz designated the second cause identified as non-transitive attribute filtering. This again is specific to vendor BGP software implementations. This bug causes duplicate announcements of prefixes when there is no apparent change in the path between the local router and the peer. Duplicate announcements occurred when changes in BGP attributes that were non-transitive, in other words, attributes that should not propagate beyond the nearest peer, changed. The bug causes any change in a non-transitive attribute to mark the prefix as changed and causes the router to re-announce the route even though there is no difference in the route.

After both of these implementation errors were corrected (e.g. in Cisco/IOS version 12) the incidence of such duplicates dropped by about a factor of ten. However, some duplicate announcements and withdrawals still occur. There is speculation that they may be caused by improper policy, improper IGP->EGP interaction, or more software bugs.

3.5 Loss of Routing Information Due to Aggregation

Aggregation of prefixes reduces router table sizes and BGP protocol traffic, but it can result in loss of connectivity. Introduction of redundant routing paths for improving network reliability and load balancing increases significantly the size of routing tables, i.e., number of prefixes in the RIB. More prefixes means that the routers must perform more computation and need more memory and as a result significantly stress the routers. Reducing the number of routes/prefixes by aggregation is a solution but with its own problems. The way prefixes can be aggregated depends on a few factors [AB00]. One of these factors is the relationship between the receiver of an announcement and the neighboring AS: it may be a peer, customer, or supplier. In a peer relationship the AS often may aggregate a more specific announcement unless it comes from the customer. In other words, the AS cannot aggregate a more specific prefix with a prefix that "contains" it (e.g., 192.50/16 and 192.50.1/24) if such prefixes come from the customer; however it could do it if such prefix announcements come from a peer. The AS takes more liberties aggregating peer routes than customer routes because customers are paying for the service. These liberties can cause problems. By doing aggregation we may suffer from loss of routing information necessary to provide optimal routing, e.g., for load balancing.

The following example demonstrates how load balancing could be unintentionally prevented upstream by aggregation. Lets assume that AS1 (a customer of AS2) received a prefix 192.50.1/24 from AS2 (AS2 manages the full prefix 192.50/16 and it assigns 192.50.1/24 to AS1). AS1 for load balancing and redundancy is multi-homed with AS2 and AS3. AS2 is peering with AS4 and AS5. Also AS3 is peering with AS5.



As peers to AS6, AS4 and AS5 may choose to aggregate more specific AS1 announcements and send only /16 prefix announcements to AS6 (192.50/16 {5,2} and 192.50/16 {4,2}). Then AS6 may choose AS4 as the next hop for 192.50/16 prefix and also announce only 192.50/16 {6, 4, 2} up the stream. If that happens, traffic to AS1 from AS6 will always go through AS4 and AS2 and never through AS3, defeating the load balancing AS1 was looking for. But even if AS6 chooses load balancing with AS4 and AS5, traffic reaching AS1 may not be well balanced.

In an attempt to achieve "reasonable" routing table sizes reliability and load balancing are often unintentionally sacrificed. This is not a trade off because the aggregator is unaware of the damage incurred. It would useful to have a tool that could assist network operators in showing the lost capabilities resulting from prefix aggregation and filtering. Some of the aggregation could be done automatically and some may require human interaction. The BGP protocol currently provides no well known mechanism to allow for the sufficient information on multi-homing and other reasons to suppress aggregation to allow routers to make properly informed decisions.

3.6 **Inconsistent Prefix Filtering Policy**

The IANA allocates IP address blocks to the three principal regional Internet registries. Those are RIPE, ARIN, and APNIC and shortly, two new registries for Latin America and Africa. Registries in turn allocate AS numbers and IP address blocks according to the needs and qualifications of the requesting organizations. The longest prefix (most specific, smallest set of included addresses) allocated by a registry is a /24, which is typically granted only to Internet infrastructure elements (DNS servers, exchange points, etc.). Otherwise the longest allocation is /20. Registries also further restrict allocations in a number of class A and C blocks [RIR].

Prefix filtering may be applied as follows:

- "Martians", i.e. impossible prefixes (e.g. 127/8 which is loopback, 192.41.177/24 which is an exchange point block, etc.) (This is not specifically prefix *length* filtering, but invalid prefix filtering)

- Class B space advertised with a mask longer than 16 (other class or prefix specific rules exist)

- According to RIRs' allocation guidelines (e.g. a mask longer than 20 on the 24.0.0.0/8 supernet)

- According to business considerations (e.g. accept prefixes known to belong to a given peer)

The purpose of prefix filtering is the reduction of the default free routing table size, which increased dramatically due to both Internet growth and proliferation of multi-homing [Ho01].

Let us look at how multi-homing affects the routing table size.

When an AS connects to the Internet for the first time, it receives its original prefix as a sub-allocation of its SP's RIR allocation; the SP can announce all of its customers (aggregating their network prefixes) to the upstream or peer using the covering (i.e. that includes the customer's sub-prefix) RIR allocation. When the customer decides to multi-home, the sub-allocation now needs to be announced by the second SP. However, given that BGP prefers a more specific (longer prefix length) to a less specific prefix, the second SP's route would always be preferred; this is why the first, original SP, is forced to not only announce the covering prefix but also the more specific prefix that is also announced by the second SP. In this way the preference for one route or another (i.e. choosing SP1's path vs. SP2's path) would now be subject to **AS Path** length and other BGP selection criteria on an equal basis. Thus multihoming a prefix, which used to take a single announcement (the covering, or aggregated prefix of the customer's original SP) now takes three (the covering prefix, the specific prefix from SP1 and the specific prefix from SP2).

Analysis of BGP table trends has established that a 30% reduction in size is achievable with aggressive (i.e. applying RIRs allocation rules) prefix length filtering [Be01]. The downside of aggressive filtering is that about 0.3% of the Internet would become unreachable, but more importantly, it would disrupt multi-homed ASes' traffic

engineering. However, today's lack of a uniform filtering policy makes it difficult to isolate the very same problems.

3.7 Robustness and Security

Network operators [NANOG] observed that some pathological oscillations of routing information and failures of the Internet backbone were results of malformed or incorrect BGP routing control messages. BGP messages could be authenticated [He98] using signatures based on MD5 hash and a shared secret. If BGP traffic is not authenticated, any malicious attacker could create malformed or incorrect BGP messages to confuse BGP routers. Some incorrect messages come from AS border routers that improperly announce ("leak") routes and contaminate other ASes. Applying filters at the BGP routers cures most of the effects of undesired BGP routing control information. Routers should be running robust operating systems; this is not always the case. Consider Cisco IOS as an example. In a 2-month period there have been three security advisories [CI01a], [CI01b], [CI01c]. This is a high rate; the IOS is not a general-purpose operating system running user written code, it is a special purpose OS running manufacturer provided code. Although these data are from Cisco, these problems are not limited to Cisco but industry-wide. There has been little emphasis up to now in the cracker community on router attacks, but they represent an opportunity to any agency desiring to damage the Internet as a whole.

3.8 Link Down

Link down refers to an interface on a router that does not have connectivity to the equipment on the far end of the link. This can be a local area network (LAN) connection or a wide area network (WAN) connection. There can be many causes. The most common are cables that are mistakenly unplugged from the router or a patch panel, a wide area circuit (T1, DS3, OC-3 etc.) that fails or is disconnected unintentionally, unintentional cable cuts from construction projects etc, and hardware failures on interface cards. The particular link becomes unavailable and all communication through the link ceases until a repair is made. At times, when attempting to repair and test a downed link,

a technician may bring the line up and then take it down several times while testing to see if the link has been fully repaired. This oscillation may further aggravate the problem as peers may damp the routes to this link for some extended period of time due to the oscillations. The operator will observe that the link is up, but customers from outside the local domain will still see the link as down until the damping timer has expired.

A particular instance of link down problems has been speculated [La99] to be the cause of unnecessary BGP route flaps. This instance is the misconfiguration of CSU/DSU devices used for many synchronous serial data circuits. The equipment on each end of a synchronous serial data circuit is configured independently. At times, this equipment may be configured so that each piece of equipment is set to use its own internal clock. The correct configuration would be to slave one end's clock from the other end's clock, take the clock timing from the line or to use a common external clock source such as GPS at both ends. Configuring the equipment so that each end independently uses its own internal clock will bring the circuit up and it will appear to work fine for periods of time, however, there will be points where the clocks slip enough to cause the line to go down for a few milliseconds. Router interfaces are sensitive to outages of this time scale and so the router would mark the interface as down and advertise to its peers that the line was no longer available. Milliseconds later the line would come back up and the router would advertise that the line was again available. [La99]

3.9 Routing Information Self Synchronization

Floyd and Jacobson [FJ94] studied the synchronization of periodic routing messages and their impact on network performance and overall efficiency. Self-Synchronization is a phenomenon (SSP) of emerging spontaneous "long range" dependencies in time or/and space in complex (non-linear) systems [PS84]. SSP is ubiquitous in biology, chemistry, collective behavior of humans (e.g., stock market) and other species (e.g., flocking in birds and insects), etc.. In routing it manifests itself by the unintended synergy between distant routers in time of broadcasting routing information. They observed that unsynchronized routing information traffic abruptly turned into synchronized routing information traffic. Even a single router could cause traffic to synchronize. Paxson also

observed [Pa97a] the self-synchronization of traffic done by routers. This self-synchronization inadvertently affects Internet performance. Labovitz, Malan, and Jahanian [LM97] pointed out that such synchronization would result in a large number of BGP routers transmitting routing information **Update**s simultaneously. Of course any increase of BGP routing **Update**s stresses even more border (BGP) routers and increases delays of inter-domain routing convergence.

It is clear that SSP in routing is due to the mutual connectivity between routers: changes in the state of one router produce messages that affect the states of others. If routing related changes did not affect the timing of routing protocol messages information SSP would disappear. Two remedies to avoid self-synchronization [FJ94] are either to randomize routing timer intervals or implement routing timers that are independent from external events. It was observed in 1998 [LMJ99] that router vendors implemented unjittered timers and by summer of 1998 providers deployed such updates. It is still an open question if the self-synchronization problem is gone. If not, detection and elimination of this behavior could be accomplished and would be useful.

3.10 Slow BGP Convergence

One of the inter-domain routing anomalies that in several ways impacts networking performance and efficiency is BGP convergence. BGP convergence is the process by which stable inter-domain routes are established. BGP routing **Update**s (announcements and withdrawal of prefixes/routes) have to propagate across the network and this takes time.

BGP convergence can be defined from the point of view of a destination, a collection of routers, or the Internet as a whole. BGP routes to a destination can be said to have converged if no router changes its route to that destination during a “full update cycle” (typically MinRouteAdvertisementInterval). All BGP routers on the Internet can be said to have converged if no router advertises any change for a “full cycle”. In practice, the

Internet does not converge to this degree as the frequency of changes exceeds the speed of convergence.

For practical purposes defining a speed of BGP convergences should be confined to a specific set of routers (ASes) rather than to the Internet at large. We could define BGP convergence as follows. The BGP routes converge if all routers within a specific set of ASes reach a stable state with respect to a specific topology change (announcement or withdrawal). If the routing table does not stabilize and keeps changing forever, resulting from a specific prefix announcement or withdrawal we will say that BGP diverges.

It is essential that BGP converge as fast as possible because while this process is in progress traveling packets need to be routed across large areas, between different ASes (autonomous systems). For example, if the invalidation of a route (withdrawal of prefix) is not acted upon then sent packets would be lost because they will be routed to a "black hole" and eventually dropped. Upper layer protocols will try to recover those lost packets by resending them, creating even more traffic and wasting network bandwidth. The processing of BGP routing information puts a lot of stress on routers. Default free routers are already dealing with 100,000+ prefixes today. Frequent announcements and withdrawal of the same routes may create route oscillations and even routing loops causing even greater delays of packet transmissions. The routing tables could be in constant state of flux due to independent route announcements or withdrawals [LM97] [La99]. According to research [LA00], BGP may take several minutes to stabilize in the event of a route change or link failure. There are many causes of slow BGP convergence. Those that are described in research literature are mentioned in this document. For example, conflicting BGP routing policies between ASes may slow down convergence, creating multi-homed ASes for redundancy will increase the time of convergence (non-aggregatable prefixes), network self synchronization, lack of prefix aggregation, network topology, etc.

3.11 Routing Loops

Routing loops are instances when packets continue to cycle through the same set of routers/interfaces until they are finally dropped due to the expiration of the TTL (time to

live) counter. Routing loops are especially dangerous as the traffic is magnified by the loop; a small amount of originated traffic can stress a router as many copies of each packet are routed around the loop.

Paxson [Pa97a] and others [NANOG] used simple tools like *traceroute* to discover routing loops, instances where sent messages were wandering between a set of routers until TTL expired or sometimes forwarded to the destination. TTL in an IP header packet is a counter and anytime the IP packet passes through a router it is decremented by one. When this counter reaches zero the IP packet is dropped. *Traceroute* uses UDP packets with a controlled value for TTL in the sent packet header. When TTL goes to zero the router echoes back an ICMP message and traceroute prints the router name or IP address as well as the round trip delay. Traceroute sends a series of UDP messages at close intervals and increments TTL for each consecutive packet. By default *traceroute* actually sends three packets with the same TTL and then increments by one for the next three packets and so on. Of course the UDP packet's destination IP address is always the same target IP address.

Using traceroute, Paxson [Pa97] discovered a couple of types of routing loops and some erroneous routing. He classified routing loops as "persistent" that lasted longer than a single traceroute measurement and "temporary" that were resolved before the measurement completed. He observed that some of the persistent routing loops lasted for hours and sometimes tens of hours suggesting that connectivity outages caused by those routing loops were not repaired or otherwise detected for considerable period of time. Paxson [Pa97a] speculated that some of these routes are transient results from slow BGP convergence.

The BGP protocol has a built-in mechanism to prevent loops. It uses the **AS Path** (sequence of intermediate AS numbers between source and destination) to detect loops. Upon receiving BGP routing **Update**s a router invalidates any advertised route that already contains its own AS number. But even with this mechanism in place, Paxson [Pa97a] observed inter-domain loops that he attributed to slow BGP convergence. Some of the persistent routing inter-domain loops may come from static routes misconfigured

into some BGP participants. Routing loops result in poor performance of the involved links including an increase in traffic, latency and the probability of dropped packets.

One specific example of routing loops observed by Paxson was apparently caused by corrupted prefix attributes. There are a number of documented cases [Pa97], where the reachability information (next-hop attribute) for a prefix was seemingly correct, but caused incorrect routing behavior. Exactly how such a situation can happen is subject to speculation. It may result from corruption of information that either comes from the announcer and is not checked or is due to a software defect that installs an otherwise valid interface address next to the wrong prefix. When a prefix is announced, the receiving BGP software can check the next-hop attribute value against the neighbor's interface address, as configured in the receiving router's configuration file.

[Pa97] classifies this problem as an erroneous routing problem that causes a temporary routing loop. According to the same study the problem was seen relatively infrequently.

3.12 Congestion Caused Loss of Routing Control Packets

Network providers [NANOG] observed that congested routers may drop BGP routing information. BGP peers that exchange routing information are using TCP/IP connections. They periodically check the state of links with its peers or even if their peers are alive by sending periodic KeepAlive test messages. If **Keep Alive** or **Update** messages are dropped because of congestion at a particular router this may cause withdrawal of neighboring routes making a part of the network unreachable. Lebowitz et al. [LA98] observed and concluded that such BGP pathological oscillations resulting from congestion leads to slow convergence. When the required level of KeepAlive message exchanges with a particular router is not met that router is marked as "down". When the "down" router recovers from a transient congestion problem it will try to re-initiate the BGP session with its neighbors. This process requires exchanging the entire routing information and not just deltas. It creates a "storm" of routing information that may propagate to other sections of Internet. Such additional load on the network may cause failures of other routers. By now the problem of congestion affecting routing instability should be fixed because newer routers have implemented methods to give higher priority

to BGP traffic when routers become congested. In newer Cisco routers when an option SPD (Selective Packet Discard) is configured, the non-routing packets are dropped instead of routing packets when a link is overloaded.

3.13 Mismatch between registered policy and observed behavior

Routes observed by peers are sometimes inconsistent with publicly registered policies. This may leave some of the networks unreachable that, from the data in the registry, should be reachable. The IRR (Internet Routing Registry) is comprised of a number of private and public registries. RADB (Merit), ARIN, APNIC and RIPE provide public registries open to all service providers who obtained address space and AS numbers from them. In North America, ANS, CW and CA*Net are private registries that record the routing policies of their respective customers. Service providers maintain all public registries on a voluntary basis. Private registries (which are run by large networks) are more forceful in *encouraging* customers to maintain their respective data up to date. Tier one SPs also require peers to register their routing policy. Some of the larger exchange points encourage peers to register as well. Peering information, which becomes more easily accessible if published in the IRR [IRR], is regarded as economically sensitive, particularly in the US. A service provider need only appear in a single registry; registration with multiple IRRs can cause confusion and should be avoided. All registries coordinate entries at least daily [RFC2650].

The IRR's purpose is to aid in debugging operational routing problems by providing partial visibility to SP policies. The registries are useful only in so far as the data they contain is kept up to date and correct. Since a routing policy change may require modifications on multiple routers a centrally coordinated router configuration update mechanism is highly desirable. The Routing Arbiter Toolset [RAT] and the Routing Policy Specification Language (RPSL) [RFC2622], on which the toolset is based, are a critical step in the right direction. The major RIRs have already deployed RPSL databases (the previous registry language, RIPE-181 [RIPE181], on which RPSL is based, is being slowly phased out.)

Even though RPSL is extensible and supports all that is needed to configure a router, (route registration, traffic engineering, aggregation, damping, distribution, etc.) its use is currently limited to the functions listed below:

- Edge filtering (customers update the IRR, and SP filters according to IRR)
- Limited peering (i.e. two SPs exchange partial customer routes)
- Isolate the router (and its maintainer) responsible for faulty advertisements
- Create filters automatically (ready to be used in a router's configuration file)

When there is a mismatch between the registered and observed policies, users of the SP observe anomalous behavior – e.g. connectivity they expect may be absent. A higher participation rate, more monitoring infrastructure (at exchange points and edges), more information (i.e. recording traffic engineering, aggregation, etc.) and better tools are needed to realize the potential the IRR holds for a smoother operating Internet.

3.14 Inconsistent Flap Damping Policy

A flap is defined as an announce/withdrawal pair. The Internet's growth resulted in more instability, with the result that in the mid 90's flaps were becoming a significant processing overhead on default free routers. A mechanism to control the flapping, best done as closely as possible to the source of the problem, was needed. With the appearance of BGP Route Flap Damping [RFC2439] router manufacturers began incorporating flap damping into their BGP implementations. Some router manufacturers refer to this as dampening (although no liquid is actually added).

If a threshold number of flaps (the actual number of flaps/minute required to trigger suppression varies with configuration parameters; the default Cisco IOS setting causes damping when there are 3 flaps within 5 minutes [BGP4 Case Studies]) is exceeded in a given period (cutoff threshold) the prefix is not announced (held down) for a calculated period (penalty); the penalty is increased on each subsequent flap; the penalty is

decremented using a half-life parameter until it reaches a threshold (reuse) at which time the prefix can be re-advertised.

Flapping is due to line-flaps (circuits), unstable IGP data that is distributed into BGP by injecting and withdrawing routes or hardware failures. The classic remedies on the flapping router's side are to either switch from IGP distribution into BGP to static distribution, and/or to aggregate the announcement to the point of masking the unstable IGP data. BGP sessions (in Cisco routers) are quickly terminated upon circuit failure because a parameter known as "fast-external-fallover" is on by default. If this parameter is turned off, BGP will use longer keepalive and hold timers (the default is 60/180 seconds); this solution does not work well in an environment where the same line has multiple BGP sessions, as this knob cannot be applied per peer.

Route damping can be tuned, by AS and prefix, if desired, thus permitting the router operator to apply different damping policies to peers' ASes and their prefixes.

Damping has been largely successful. However, a number of problems have appeared as a result of lack of coordination of damping policy [RIPE210]:

- Damping of /16 prefixes is less aggressive than that of /24, as /24 prefixes are considered less stable than shorter prefixes; uncoordinated damping can cause additional flapping or inconsistent routing; for example a shorter prefix may be held down while a longer one is already available elsewhere; this would cause a prefix's routing to change in unpredictable ways.

- Uncoordinated hold-down and reuse-threshold timers can cause partial outages; for example, if an SP's upstream damps the SP more aggressively than it damps its own customer; there may be reachability from the other SPs customers but not from elsewhere.

3.15 Constrained Policy

Griffin and Wilfong [GW00] pointed out that the BGP protocol is the only widely accepted IP routing protocol that is not *safe* for convergence. Safe means that the protocol will always converge. BGP is a distance-vector routing protocol with a twist. It is called a path-vector protocol because unlike typical distance-vector protocols where only the next hop is defined, in BGP the route announcements include the path of ASes involved (or at least an unordered set of them). In addition, BGP is a routing protocol that allows selecting best routes based on distance-vector metrics (shortest path first policy) and other routing policies, such locally defined policies or **MED** (multi-exit discriminator) policies. This BGP multi-policy capability allows overwriting the distance-vector metrics by other policy metrics. In a broader sense BGP policies reflect commercial relationships between neighboring ASes. AS local policies specify which route to propagate to its own neighbors and which not to propagate without exposing those policies and its network topology to others.

Varadhan, Govidan and Estrin [VGE96] pointed out that such independence in defining BGP policies by different ASes (Autonomous System) might lead to BGP protocol oscillation in the absence of topology changes or even to a situation in which BGP routing may diverge. In order to prove that this is possible they demonstrated route selecting preference functions/policies that will cause route “feedbacks”, manifesting in route oscillations. However, this has not been observed in practice. Griffin and Wilfong [GW99] built a formal BGP model and they argued that static analysis of BGP convergence is either NP-complete, or NP-hard even if knowledge of routing policies of various ASes is available. According to them static analysis may detect policy conflicts that may lead to slow convergence but will not allow one to determine a cure. Only dynamic analysis, they argued, could provide a dynamic solution to BGP convergence (e.g., based on heuristics). This solution may not be practical because it may require keeping track of past traces and it may require modifying implementations of BGP. On other hand, Gao and Rexford [GR00] suggested that a static solution is possible if we

could constrain AS configurations to some hierarchical structure. For example, by establishing a creation of customer-provider and peer-to-peer relationships, they argued, BGP will be safe and will not diverge. Also Griffin at el [GW00] [GSW99] pointed out that there is also a possibility to define BGP policies in such a way that the BGP routing will not diverge. According to them eliminating certain problematic routes and making tradeoffs between stability and route preference could accomplish this.

Researchers [LW01][LA00] studying BGP instability observed that even under constrained BGP policies, BGP takes an order of magnitude longer to converge than was previously assumed, i.e., minutes instead of seconds. BGP allows routing path selection based on various policy attributes including local preferences and values. However, it is observed that most of deployed policies are based on a shortest path first policy. But even such a constrained policy may lead to delayed convergence under certain circumstances. For redundancy purposes Internet providers are more and more often orchestrating their networks as multi-homed systems. They are trying to establish connectivity to the same destination (prefix) through multiple connections using different service providers (SPs). It is a common practice to managed multi-homed failover by announcing to downstream SPs different paths for the same prefix of upstream SPs. The downstream SP could "see" the same prefix announcement with a short path through one SP and another announcement for the same prefix with a longer path through another SP. To be sure that the longer path is actually longer the **AS Path** is populated with redundant AS numbers. When the shortest path fails (BGP withdrawal of shortest path announcement), the downstream provider will select the longer path. As Lebovitz et al [LA00] showed such multi-homed failover is similar to route failures that affects BGP convergence and the BGP convergence delay will depend on the longest alternative **AS Path**. Others [GGR01] presented a general model for safe routing failover (safe convergence) that increases network reliability through properly defined local BGP routing policies.

4 Conclusions and Future Direction

These fifteen issues are not an exhaustive list of inter-domain routing issues, nor are they cleanly partitioned, however, they do fall into the three broad categories of equipment failure and performance; human error; and lack of communication and cooperation between operators. These fifteen issues do, however, show the scope of the problem. Internet routing is approximately 2 orders of magnitude less reliable than the routing of the public switched telephone network (PSTN) [La99]. Service providers struggle daily to detect, identify, isolate, and repair problems; today they address each of these issues by manual, often painful, *ad hoc* processes. Still the Internet is amazingly robust, functioning in the near continual presence of inter-domain routing problems.

In examining the anomalies for this paper, it was apparent that very often it is difficult to determine whether an anomaly exists independently or is the result of some other anomaly. Indeed, informal conversations with network operators point to router misconfiguration (Section 3.1 “Mismatch Between Router Configuration and Observed Behavior”) as the primary cause for disruption.

Due to the lack of tools to assist service providers in detecting and isolating anomalous behavior in routing between domains, and the prevalence of problems in this area, we believe our contribution lies in providing network operators tools, measurements and methodologies to detect and isolate anomalies more quickly. Agilent is primarily a test and measurement company and work in the area of addressing inter-domain routing issues is appropriate and on target to our divisional partner’s business needs.

Our next step will be to select one or more of the fifteen listed anomalies, determine current methods of isolation and detection and then to explore ways in which these may be improved or automated so that new tools may be developed to better enable operators and engineers to find inter-domain routing anomalies more quickly.

5 Acronyms

For acronyms not found below you might try http://home.iae.nl/users/jrm/ or http://www.cisco.com/univercd/cc/td/doc/cisintwk/ita/, http://www.cknow.com/index.htm, or http://www.mentortech.com/learn/welcher/misc/iosacronym.htm.

APNIC	Asia Pacific Network Information Centre: Internet numbering registry for the Asian / Pacific region
ARIN	American Registry for Internet Numbers: Internet numbering registry for North America
ARPA	Advanced Research Projects Agency
AS	Autonomous System: Administrative boundary within which routing is typically done with full knowledge; ASes abstract away some details of network topology in communicating outside themselves.
BGP, BGP4	Border Gateway Protocol version 4, A layer-3 routing protocol of IP, for use between multiple autonomous networks. See [RFC1771]
CIDR	Classless Inter-Domain Routing (CIDR), see [RFC1517]
Cisco/IOS	Internetwork Operating System (Cisco term for router Operating System and Command Line Interface)
CPU	Central Processing Unit, also SPU
CSU	Channel Service Unit: Digital interface device that connects end-user equipment to the local digital telephone loop. Often referred to together with DSU, as *CSU/DSU*. See also DSU.
DARPA	Defense Advanced Research Projects Agency
DNS	Domain Naming Service: Mapping from names to IP addresses, the route names are controlled by ICANN
DS3, DSn	Digital Signaling 3 DS0 = 64 Kbps Out of band signaled, 56 Kbps in band signaled DS1 = 1.544 Mbps = another name for a T1 DS3 = 44.768 Mbps = 28 T1 = 672 DS0 circuits

DSU	Data Service Unit: Device used in digital transmission that adapts the physical interface on a DTE device to a transmission facility, such as T1 or E1. The DSU also is responsible for such functions as signal timing.
E-BGP. eBGP	External BGP, use of BGP in communicating with other ASes
FIB	Forwarding information Base: Information used in real time by router in processing incoming packets, see also RIB
GPS	Global Positioning System; a satellite based navigation system allowing reveivers to determine their position latitude, longitude, altitude, time with high accuracy.
IAB	Internet Architecture Board. Board of internetwork researchers who discuss issues pertinent to Internet architecture. Responsible for appointing a variety of Internet-related groups, such as the IANA, IESG, and IRSG. The trustees of the ISOC appoint the IAB. See also IANA, IESG, IRSG, and ISOC.
IANA	Internet Assigned Numbers Authority:
I-BGP, iBGP	Internal BGP, use of BGP in synchronizing multiple interfaces between an AS and the outside
ICANN	Internet Corporation for Assigned Names and Numbers. Non-profit, private corporation that assumed responsibility for IP address space allocation, protocol parameter assignment, domain name system management, and root server system management functions that formerly were performed under U.S. Government contract by IANA and other entities.
IEEE	Institute of Electrical and Electronics Engineers. Professional organization whose activities include the development of communications and network standards. IEEE LAN standards are the predominant LAN standards today.
IESG	Internet Engineering Steering Group. An organization appointed by the IAB that manages the operation of the IETF.
IETF	Internet Engineering Task Force, see http://www.ietf.org/
IGP	Interior Gateway Protocol: Internet protocol used to exchange routing information within an autonomous system. Examples of common Internet IGPs include IGRP, OSPF, and RIP.
IP, IPv4	Internet Protocol
IRR	Internet Routing Registry, see http://www.irr.net/
IRSG	Internet Research Steering Group. Group that is part of the IAB and oversees the activities of the IRTF.
IS-IS	Intermediate System to Intermediate System Intra-Domain Routeing *sic* Exchange Protocol

ISOC	Internet Society. International nonprofit organization, founded in 1992, that coordinates the evolution and use of the Internet. In addition, ISOC delegates authority to other groups related to the Internet, such as the IAB. ISOC is headquartered in Reston, Virginia (United States).
ISP	Internet Service Provider
IX	Internet Exchange: A component of the Internet backbone where ISPs can connect together.
LAN	Local Area Network
MAC	Media Access Control
MAE, MAE-West, MAE-East	Metropolitan Area Ethernet: A MAE is a large Network Access Point (NAP). A company known as MFS constructed the first MAE in Washington, D.C. and, later, a second in Silicon Valley (known generally as MAE-East and MAE-West). MAE is a term specific to the MFS facilities; but NAP and MAE are starting to mean the same thing.
MAN	Metropolitan Area Network
MED	Multi-Exit Discriminator: a well known BGP attribute used to suggest to other ASes the appropriate entrance to an AS to be used for a particular prefix.
NANOG	The North American Network Operators' Group mailing archive. http://www.cctec.com/maillists/nanog/index.html
NAP	Network Access Point: A component of the Internet backbone where ISPs can connect together. Like hub airports, NAPs are also points of congestion for Internet traffic. The new name for a NAP is an Internet Exchange (IX).
NSF	National Science Foundation
NSP	Network Service Provider

OC-3, OCn	Optical Carrier Speed classification: OC-1 = transmission rate of 51 Mbps OC-3 = transmission rate of 155 Mbps (STM-1 in SDH) OC-9 = transmission rate of 466 Mbps OC-12 = transmission rate of 622 Mbps (STM-4 in SDH) OC-18 = transmission rate of 933 Mbps OC-24 = transmission rate of 1244 Mbps OC-36 = transmission rate of 1866 Mbps OC-48 = transmission rate of 2488 Mbps (STM-16 in SDH) OC-96 = transmission rate of 4976 Mbps OC-192 = transmission rate of 9953 Mbps (STM-64 in SDH)
OSI	Open Systems Interconnect model: ISO's reference model for enabling multivendor systems to intercommunicate as described in ISO 7498-1984.
OSPF	Open Shortest Path First
PAIX	Palo Alto Internet Exchange
PSTN	Public Switched Telephone Network
RADB	Routing Arbiter Database: A routing registry run by MERIT, see http://www.radb.net/.
RAT	Routing Arbiter Toolset, see http://www.isi.edu/ra/RAToolSet/
RFC	Request for Comment: the mechanism used to propose IETF standards, IETF RFCs can be obtained from http://www.ietf.org/rfc/
RIB	Data used by router in computing appropriate paths for packets to be handled later, used to populate the FIB
RIP	Routing Information Protocol, see RFC1388
RIPE	Reseaux IP Europeens: Internet numbering registry for Europe
RIR	Regional Internet Registry: see ARIN, RIPE, APNIC
RPSL	Routing Policy Specification Language (RPSL), see [RCS2622]
SP	Service Provider (NSP or ISP)

T1, Tn	See DS1 T1 = transmission rate of 1.5 Mbps T2 = transmission rate of 6.3 Mbps T3 = transmission rate of 46 Mbps T4 = transmission rate of 281 Mbps
TCP/IP	Transmission Control Protocol / Internetworking Protocol
TTL	Time To Live: a header in an IP packet, the TTL is decremented each time the packet is processed by a router. When the count reaches zero the packet is discarded and an ICMP error may or may not be sent to the packet originator.
WAN	Wide Area Network

6 References

[AB00] A. Ahuja, R. Bush, "Internet Draft, Effects of Aggregation and Filtering on Routing Table Growth", draft-ptomaine-taxonomy-00.txt

[AU99] A. Antony, H. Uijterwaal, "Routing Information Service Design Note", Ripe-200, 1999, http://ssl/docushare/dscgi/ds.py/Get/File-2987/RIPE_Routing_Information_Service_Design_Note.pdf

[Be01] S. Bellovin *et al.*, "Slowing Routing Table Growth by Filtering Based on Address Allocation Policies," August 2001 http://www.research.att.com/~jrex/papers/filter.pdf

[CI01] Cisco BGP Case Studies: Route Flap Damping, http://www.cisco.com/warp/public/459/16.html

[CI01a] "Cisco Security Advisory: IOS HTTP Authorization Vulnerability" June 27, 2001, http://www.cisco.com/warp/public/707/IOS-httplevel-pub.html

[CI01b] "Security Advisory: IOS Reload after Scanning Vulnerability", May 24, 2001 http://www.cisco.com/warp/public/707/ios-tcp-scanner-reload-pub.shtml

[CI01c] "Cisco Security Advisory: Cisco IOS BGP Attribute Corruption Vulnerability" May 10, 2001. http://www.cisco.com/warp/public/707/ios-bgp-attr-corruption-pub.shtml

[Do98] P. Donner, “TELECOM-SP ISP Network Design Issues”, http://www.cisco.com/public/cons/workshops/isp-workshop/WhitePapers/Policy_Implementation_CS1.31.PDF

[FJ94] S. Floyd and V. Jacobson, "The Synchronization of Periodic Routing Messages", ACM Trans. on Networking, Vol. 2, No. 2 (April 1994), pp. 122 – 136

[GGR01] Lixin Gao, Timothy G. Griffin, and Jennifer Rexford, Inherently Safe Backup Routing with BGP, Proc. IEEE INFOCOM, April 2001

[GH] G. Houston, “Internet Draft, Architectural Requirements for Inter-Domain Routing in the Internet”, draft-iab-bgparch-01.txt

[GSW99] Timothy G. Griffin, F. Bruce Shepherd, and Gordon Wilfong. Policy disputes in path-vector protocols. In Proceedings of ICNP '99 Conference, Toronto, Canada, October 1999

[GW00] T. Griffin and G. Wilfong. A Safe Path Vector Protocol. Proc. of IEEE Infocom, Mar. 2000.

[GW99] T. G. Griffin and G. Wilfong. An Analysis of BGP Convergence Properties. In Proceedings of the ACM SIGCOMM '99, Sept. 1999.

[Ha97] B. Halabi. “Internet Routing Architectures (1st edition)”. Addison Wesley Longman, 1997.

[He87] “Introduction to the Internet Protocols” by Charles L. Hedrick http://oac3.hsc.uth.tmc.edu/staff/snewton/tcp-tutorial/

[He98] A. Heffernan. Protection of BGP Sessions via TCP MD5 Signature, RFC 2385, August 1998.

[Ho01] G. Houston, “Analyzing the Internet BGP routing table,” *Internet Protocol Journal*, March 2001. http://www.cisco.com/warp/public/759/ipj_4-1/ipj_4-1_bgp.html

[IRR] http://www.irr.net

[LA00] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, “Delayed internet routing convergence”, In Proc. ACM SIGCOMM '00 (Stockholm, Sweden, 2000), pp. 175--187.

[LA98] C. Labovitz, A. Ahuja, “Experimental Study of Internet Stability and Wide-Area Backbone Failures”, CSE-TR 382-98, Department of Electrical Engineering and Computer Science, University of Michigan, 1998.

[La99] C. Labovitz, “Scalability of the Internet Backbone Routing Infrastructure”, Ph.D. thesis, University of Michigan, 1999, http://ssl.labs.agilent.com/docushare/dscgi/ds.py/Get/File-2978/SCALABILITY_OF_THE_INTERNET_BACKBONE_ROUTING_INFRASTRUCTURE.pdf

[LM97] C. Labovitz, G. R. Malan, and F. Jahanian. Internet Routing Instability. In Proceedings of the ACM SIGCOMM '97, Sept. 1997.

[LMJ99] C. Labovitz, G. R. Malan, and F. Jahanian. Origins of Internet Routing Instability. In Proceedings of the IEEE INFOCOM '99, 1999.

[LR00] Lixin Gao and Jennifer Rexford. Stable internet routing without global coordination. In ACM SIGMETRICS, 2000.

[LW01] C. Labovitz, R. Wattenhofer, S. Venkatachary, and A. Ahuja, "The impact of Internet policy and topology on delayed routing convergence," in Proc. IEEE INFOCOM, April 2001

[NANOG] The North American Network Operators' Group mailing archive. http://www.cctec.com/maillists/nanog/index.html

[Ob00] Davor Obradovic, "Formal Analysis of Convergence of Routing Protocols" PhD Thesis Proposal, Department of Computer and information Science, university of Pennsylvania, November 2000.

[Pa97] V. Paxson, "Measurements and Analysis of End-to-End Internet Dynamics", Ph.D. thesis, University of California, Berkeley, 1997, http://ssl.labs.agilent.com/docushare/dscgi/ds.py/Get/File-2958/measurmentspaxsonphdthesis.pdf

[Pa97a] V. Paxson, "End-to-End Routing Behavior in the Internet", IEEE/ACM Transactions on Networking, 5(5): 601--615, Oct. 1997

[PS84] Ilya Prigogine and Isabelle Stengers, Order out of Chaos. Heinemann, London, 1984

[RAT] http://www.isi.edu/ra/RAToolSet/

[RFC0891] D. L. Mills, "DCN Local-Network Protocols", http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc0891.html

[RFC0990] Reynolds & Postel, "Network Numbers", http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc0990.html

[RFC1388] G. Malkin, "RIP Version 2 Carrying Additional Information", http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1388.html

[RFC1517] R. Hinden, Editor, "Applicability Statement for the Implementation of Classless Inter-Domain Routing (CIDR)", http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1517.html

[RFC1518] Y. Rekhter "An Architecture for IP Address Allocation with CIDR", http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1518.html

[RFC1519] V. Fuller "CIDR: an Address Assignment and Aggregation Strategy", http://www.cis.ohio-state.edu/cgi-bin/rfc/rfc1519.html

[RFC1771] Y. Rekhter *et al.*, "A Border Gateway Protocol 4" http://www.ietf.org/rfc/rfc1771.txt

[RFC1772] Y. Rekhter, P. Gross, “Application of the Border Gateway Protocol in the Internet” http://www.freesoft.org/CIE/RFC/1772/index.htm

[RFC2260] T. Bates *et al.*, “Scalable support for Multi-homed Multi-provider Connectivity” http://www.ietf.org/rfc/rfc2260.txt

[RFC2439] C. Villamizar *et al*, “BGP Route Flap Damping”, RFC 2439 http://www.ietf.org/rfc/rfc2439.txt

[RFC2622] C. Alaettinoglu *et al.*, “Routing Policy Specification Language (RPSL)” http://www.ietf.org/rfc/rfc2622.txt

[RFC2650] D. Mayer *et al.*, “Using RPSL in Practice” http://www.ietf.org/rfc/rfc2650.txt

[RIPE181] T. Bates *et al.*, “Representation of Routing Policies in a Routing Registry”

[RIPE210] T. Barber *et al.*, “RIPE Routing-WG Recommendation for coordinated route-flap damping parameters”, RIPE 210

[RIR] http://www.apnic.net/db/min-alloc.html
http://www.ripe.net/ripe/docs/smallest-alloc-sizes.html
http://www.arin.net/regserv.html

[SK00] A. Shaikh, L Kalampoukas, A. Varma, and R. Dube, “Routing stability in congested networks: Experimentation and analysis”, In Proc. ACM SIGCOMM '00, pages 163--174, Stockholm, Sweden, 2000

[VGE96] K. Varadhan, R. Govindan, and D. Estrin. Persistent Route Oscillations in Inter-Domain Routing. ISI technical report 96-631, USC/Information Sciences Institute, 1996.