Macrothink Institute™

Network Protocols and Algorithms
ISSN 1943-3581
2010, Vol. 2, No. 1

A Survey of BGP Session Maintenance Issues and Solutions

Dario Vieira

IUT de Montreuil, Université Paris 8

E-mail: d.vieira@iut-univ-paris8.fr

Abstract

The Border Gateway Protocol (BGP) is the routing protocol that glues together the global Internet. BGP makes use of a managed session to maintain a bidirectional error-free session over which reachability information is exchanged between Autonomous Systems (ASes) in the global Internet. Nevertheless, despite its great importance to BGP, this managed session suffers from some weaknesses such as a slow failure mechanism, which may represent a great deal of lost data, and some security problems. This paper surveys both research and standardization efforts relating to BGP session maintenance.

Keywords: BGP, BGP Session Maintenance, Interdomain Routing, Network, Border Gateway Protocol.

1. Introduction

BGP views the Internet as a collection of autonomous systems[1] (ASes) connected to each other. Over this direct connection, BGP sessions are established. While there are many BGP sessions over each link, there is never BGP session between non-neighboring routers. BGP sessions are used to exchange network reachability information (that is, how to locate other hosts and routers). Each router informs its neighbor what address ranges it knows, how to route to, along with additional information. These are used to take the decision of which router will actually be used to route that part of the address space.

Thereby, one of fundamental functions of routing protocols is to establish and manage a session between two endpoints and keep it alive as much as possible, even though endpoints do not exchange messages. Accordingly, two nodes should be able to communicate as long as there is a path in the underlying network. To ensure connectivity, routing protocols need to, for instance, quickly detect and recover from failures.

As BGP provides information for controlling the flow of packets between ASes, the protocol plays a critical role in the Internet efficiency, reliability, and security. However, BGP suffers from several well-know vulnerabilities: slow convergence, prone to serious anomalies (e.g., persistent oscillation, forwarding loops, and black holes), vulnerable to malicious attack, and difficult to troubleshoot.

Loss of connectivity on the Internet may range from an inconsequential annoyance to a devastating communication failure, which is related to the extension and scope of the lost connections. For instance, today's Internet is home to an increasing number of critical business applications, such as online banking and stock trading, that can cause financial harm to an individual or institution if communication is lost at a critical time (such as during a time-sensitive trading session). As the number of time-sensitive applications on the Internet grows, so will our reliance on the Internet to provide us reliable and secure services.

Researches have created tools for analyzing measurement feeds of BGP update messages in order to both detect and diagnose routing problems (e.g., [1, 2]). These contributions have significantly improved our understanding of BGP and our ability to work around some of its limitations. In addition, many researchers [3, 4, 5] are trying to solve the enigma of BGP root cause analysis. One of the reasons is due to the fact that one event in the Internet may cause multiple messages that need to be analyzed to pinpoint the cause of this event.

In this paper, we intend to survey the state of the art and briefly describe some of the most relevant proposals in interdomain routing, especially concerning BGP session maintenance. The next Section provides an overview of interdomain routing and BGP. Subsequent Sections expose current research addressing BGP session maintenance and interdomain routing issues.

[1] A network under the administrative control of a single organization is called Autonomous System (AS).

2. Border Gateway Protocol

Dynamic routing protocols for IP come in two basic flavors: Interior Gateway Protocol (IGPs) and Exterior Gateway Protocols (EGPs). The boundary of these two levels is defined by Autonomous Systems (ASes). IGPs are used for routing within ASes, whereas EGPs are used for global routing between ASes. This organization reflects the coarse structure of the Internet. There are several IGPs in use, whereas there is only one EGP in use: the Border Gateway Protocol (BGP) [6].

A router running the BGP protocol is known as a BGP speaker. BGP speakers communicate across TCP and become peers or neighbors. Each pair of BGP neighbors maintains a session, over which information is communicated. A BGP speaker's neighbor is one hop away, thus the term per hop refers to the relationship between BGP neighbors.

The classic definition of Autonomous System (AS) is a set of routers under a single technical administration, using an IGP and common metrics to route packets within the same AS, and using an EGP to route packets to other ASes. For instance, the network of an university, corporation, or Internet Service Provider (ISP) would typically be a single AS. Fig. 1 illustrates the definition of AS. Each AS is concerned only with its own intra-domain routing plan and administrators may implement it in any way they choose. However, inter-domain routing is used for the transfer of routing information between ASes. Because routing is collaborative, all routers that participate in the process must make use to the same protocol. It is for this reason that one protocol is used ubiquitously for inter-domain routing throughout the Internet: the Border Gateway Protocol (BGP) [6].

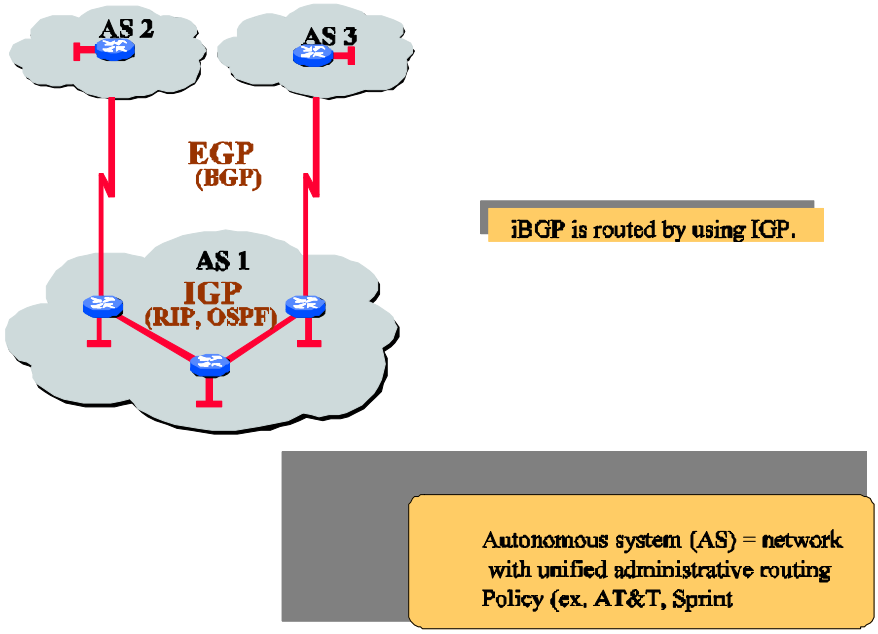

Figure 1: Intra-domain and Inter-domain Routing. IGP is used to route packets within the same AS, while EGP is used to route packets to other ASes.

Border Gateway Protocol (BGP) is the de facto inter-domain routing protocol used to exchange reachability information between ASes in the global Internet. BGP is so-called path vector[2] routing protocol where best route selection for an AS is a function of its routing policies to override distance-based metrics with policy-based metrics.

Within an AS, BGP runs only on those routers chosen by the administrators to perform the duties of inter-domain routing, i.e., it does not need run on all routers in the AS. Although BGP does not concern itself with intra-domain routing, it does interact with the intra-domain routing protocols, known as Interior Gateway Protocols (IGPs), to bridge the gap between the two levels in the routing hierarchy.

An AS can contain many routers which are connected in an arbitrary topology. We can draw a distinction between routers in an AS that are connected only to other routers within the AS, versus those that connect to other ASes. Routers in the former group are usually called internal routers, while routers in the latter group are called border routers in BGP, and similar names in other protocols.

The notion of a border is, of course, the basis for the name of the BGP protocol itself. To actually create the BGP internetwork, the BGP speakers bordering each AS are physically connected to one or more BGP speakers in other ASes, in whatever topology the internetwork designer decrees. When a BGP speaker in one AS is linked to a BGP speaker in another AS, they are deemed neighbors. When we connect ASes together to form an internetwork, the paths between AS border routers make up the conduit by which messages move from one AS to another. These direct connections permit them to exchange information about ASes of which they take part.

It is very important to control the flow of messages[3] between ASes carefully. Depending on circumstances, we may wish to limit or even prohibit certain types of messages from going to or from certain AS. These decisions in turn have a direct impact on BGP route determination.

Indeed, business practices in the Internet dictate some types of relationships among the administrations of different autonomous systems. The ASes must establish policy whenever they decide to exchange traffic. There are three basic types of peering relationships: customer-provider, peer-to-peer and sibling-to-sibling relationships. In a customer-provider, one AS (the customer) is provided with access to the rest of the Internet by another AS (the provider). In a sibling-to-sibling relationship, each AS provides the other with access to all of the routes that it know. In a peer-to-peer relationship, two ASes agree to share traffic between their customers.

There is a key distinction between local traffic and through traffic in BGP: (i) Local traffic that is traffic carried within an autonomous system that either originated in that same AS, or is intended to be delivered within that AS; and (ii) Transit traffic, which is traffic that

[2] BGP is categorized as path vector protocol because BGP speakers can use a number of path attributes to select preferred paths. BGP is more flexible than distance vector protocols. Routers executing a distance vector protocol use the path length as the only criterion to select preferred paths.

[3] The flow of messages is sometimes collectively called traffic.

was generated outside an AS *p* and is intended to be delivered outside that AS.

2.1 What Problem is BGP Solving Anyway?

The Goal of the Border Gateway Protocol is to facilitate the exchange of route information between BGP peers, so that each router can determine efficient routes to each of the networks on an IP internetwork. This means that descriptions of routes are the key data that BGP peers work with. Every BGP speaker is responsible for managing route descriptions according to specific guidelines established in the BGP RFC [6].

Conceptually, the overall activity of route information management can be considered to encompass four main tasks:

- **Route Storage**: Each BGP peer stores information about how to reach networks in a set of special databases. It also uses databases to hold routing information received from other devices.
- **Route Update**: When a BGP device receives an update message from one of its peers, it must decide how to use this information. Special techniques are applied to determine when and how to use the information received from peers to properly update the peer's knowledge of routes.
- **Route Selection**: Each BGP uses the information in its route databases to select best routes to each network on the internetwork.
- **Route Advertisement**: Each BGP speaker regularly advertises its peers its knowledge about various networks and how to reach them. This is called route advertisement and is accomplished using BGP update messages.

The routing information management and handling of BGP is relied on a database so-called Routing Information Base (RIB). However, RIB is not a monolithic entity, but it is a fairly complex data structure where BGP peers store considerably information about routes. It is comprised of three separate sub-sets which are used by a BGP peer to cope with the input and output of routing information. These three types of RIB are the mechanism by which information flow is handled in a BGP peer. The different types of RIB are described as following:

- Adj-RIBs-In: A set of input database parts that holds information about routes received from BGP speakers.
- Loc-RIB: The local RIB is the core of RIB. It stores routes that have been selected by a BGP peer and are considered valid to it.
- Adj-RIBs-Out: A set of output database parts which holds information about routes that a BGP speaker has selected to be disseminated to its peers.

The data received from update messages transmitted by BGP speakers is held in the Adj-RIBs-In, with each Adj-RIB-In holding input from one peer. This data is then analyzed and appropriate information selected to update the Loc-RIB, which is the main database of RIB. On a regular basis, information from the Loc-RIB is placed into the Adj-RIBs-Out to be sent

to other peers using update messages. This information flow is accomplished as route update, selection and advertisement known as the BGP Decision Process. The Fig. 2 illustrates the process describe above.

2.1.1 BGP Route Attributes

The information about the path to each route is stored in the RIB of each BGP speaker in the form of BGP path attributes. These attributes are used to advertise routes to networks when BGP devices send out update messages. The storing, processing, sending and receiving of path attributes is the method by which routers decide how to create routes.

There are several different path attributes, each of which describes a particular characteristic of a route. BGP attributes are divided in two set so-called *optional* and *well-known* attributes. The formers are obviously optional and BGP implementations may or not support them. However, every BGP speaker must recognize and process the well-known attributes, but only some are required to be sent with every route. These are further differentiated relied on how they are handled when received by a peer that does not recognize them.

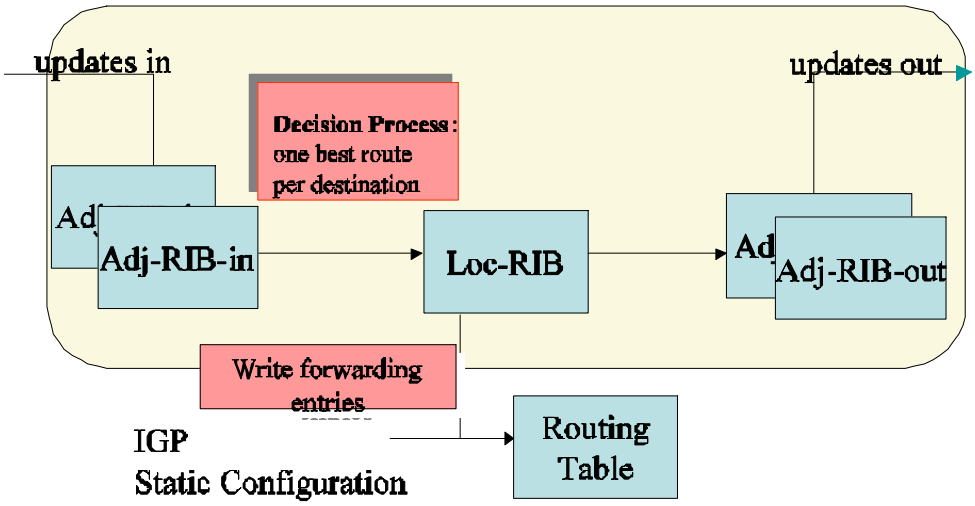

Figure 2: Each BGP router contains a Routing Information Base (RIB) that contains the routing information maintained by that router. The RIB is comprised of three separate sub-sets which are used by a BGP peer to cope with the input and output of routing information: Adj-RIBs-In, Loc-RIB and Adj-RIBs-Out.

The main path attributes defined in BGP may be described as following:

- *Origin* is use to indicate how a particular reachability information was learned. It can assume one of the following values:
  - IGP: The route is learned from the interior to the originating AS. This value is set when the network router configuration command is used to inject the route into BGP.
  - EGP: The route is learned from eBGP.

  - Incomplete: The origin of the route is unknown or learned in some other way. An incomplete origin happens when a route is redistributed into BGP.

- *Next_Hop* is the IP address that is used to reach the advertising router.
- *Multi_Exit Discriminator* (MED): The MED is used for an AS to indicate the best entry point to its neighboring AS in case of multiple connections.
- The *Local Preference*: If there are multiple exit points from the AS, the local preference attribute is used to select the exit point for a specific route. Thereby, the local preference is used within an AS to implement local policies for best exit point. It is a valuable tool for Internet Service Provider (ISP) to influence their costs for outgoing traffic.
- *Weight* is an attribute that is local to a router. The weight attribute is not advertised to neighboring routers. If the router learns about more than one route to the same destination, the route with the highest weight will be preferred.
- The *community* attribute provides a way of grouping destinations, called communities, to which routing decisions (such as acceptance, preference, and redistribution) can be applied. Route maps are used to set the community attribute. Predefined community attributes are as following:
  - No export: do not advertise the route e to eBGP peers.
  - No advertise: do not advertise the route e to any peer.
  - Internet Advertise: advertise the route *e* to the Internet community.
- *AS_Path* is a sequence of ASes through which one can traverse the network. The last AS in this sequence is the originator of a route that manages the subnetworks that contains the prefixes. This is the mechanism that BGP uses to detect routing loops by throwing away any message that contains it own AS number.

The construction of forwarding tables is based on routes that the speaker manages. A BGP route specifies a path to reach the destination. It is used a tuple (NLRI, AS path) to express the function. *Network Level Reachability Information* (NLRI) is the route field that defines the route destination. NLRI contains a list of prefixes. In some case, NLRI is as simple as one prefix to express the IP address range of a subnetwork. More generally, a BGP speaker uses multiple prefixes in a single route to announce the reachability of multiple subnetworks in an AS.

2.1.2 BGP Decision Process

As we have pointed out, the RIB contains sections for holding input information received from BGP peers, and for holding output information each BGP peer. The functions of route update, selection and advertisement are concerned with analyzing this input information, deciding which one to include in the local database, updating the database, and then choosing which routes to send. The mechanism so-called the *Decision Process* is responsible in BGP to accomplish these tasks. It is made up of three overall phases:

- *Phase 1*: Each route received from a BGP speaker in a neighboring AS is analyzed and assigned a preference level. These routes are then ranked according to preference. Afterward, the best one of them is advertised to other BGP peers within the AS.
- *Phase 2*: The best route for each destination is selected from the incoming data relied on preference levels, and used to update the local routing information base (the Loc-RIB).
- *Phase 3*: Routes in the Loc-RIB are selected to be sent to neighboring BGP speakers in other ASes.

The assigning of preferences amongst routes only becomes important when more than one route has been received by a BGP speaker for a particular network.

In the case where a set of routes to the same network are all calculated to have the same preference, a tie-breaking scheme is used to select from among them. Additional logic is used to handle special circumstances, such as the case of overlapping networks. The Fig. 3 illustrates the process describe above.

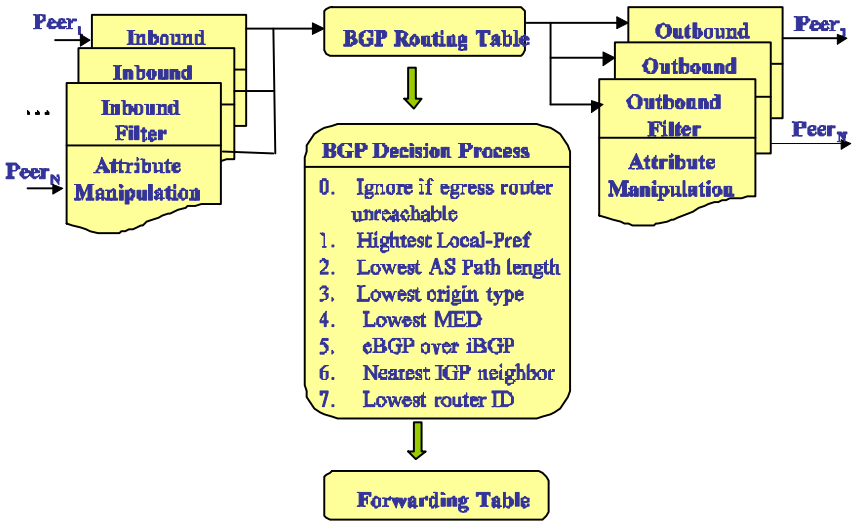

Figure 3: The BGP Decision Process is responsible to analyze the input information, and afterward deciding which one to include in the local database, updating the database, and then choosing which routes to send.

2.1.3 Limitations on Route Selection Process

When considering the route selection process, it is very important to remember that BGP is a routing protocol that operates at the inter-autonomous-system level. Thus, routes are chosen between ASes and not at the level of individual routers within an AS. So, for example, when BGP stores information about the path to a network, it stores it as a sequence of autonomous systems, not a sequence of specific routers. BGP cannot deal with individual routers in an AS because by definition, the details of what happens within an AS are supposed to be hidden from the outside world. It does not know the structure of ASes outside its own.

However, this has an important implication for how BGP selects routes: BGP cannot guarantee that it will select the fastest, lowest-cost route to every network. It can select a route that minimizes the number of ASes that lie between itself and a particular network, but of course ASes are not all the same. Some ASes are large and consist of many slow links, while others are small and fast. This situation is illustrated by the Fig. 4 extracted from [7]. Choosing a route through two of the latter type of AS will be better than choosing a route through one of the former, but BGP cannot know that. Policies can be used to influence AS selection to some extent, but in general, since BGP does not know what happens in an AS, it cannot guarantee the efficiency of a route overall.

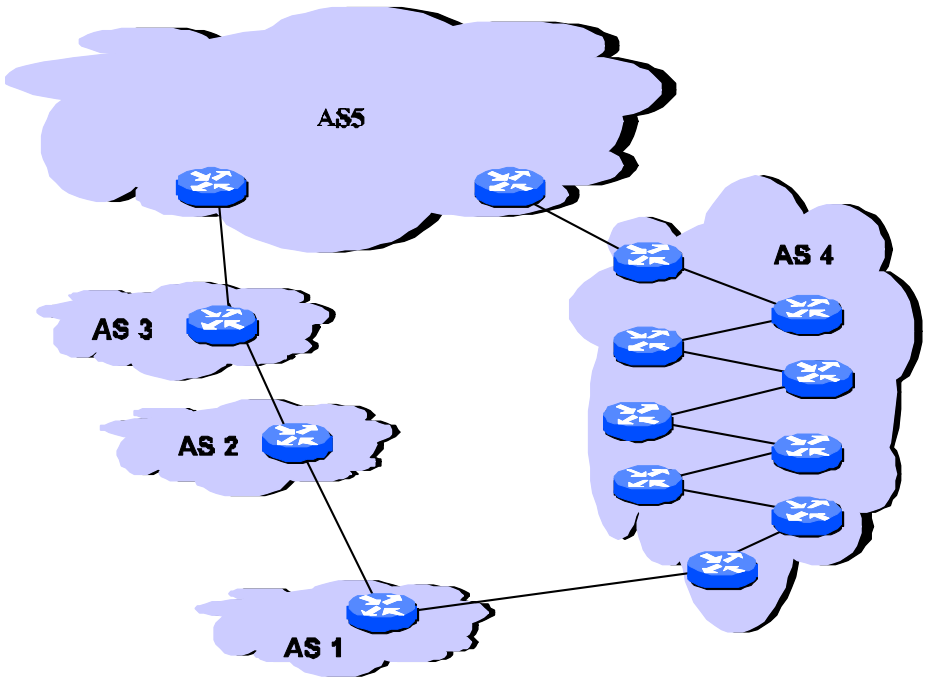

Figure 4: Limitations on BGP's ability to select efficient routes. In this example, BGP selects the path 4 1 as better than 3 2 1.

3. BGP Session Maintenance

BGP views the Internet as a collection of ASes connects to each other. Over this direct connection, it is established BGP sessions. While there are many BGP session over each link, there are never BGP session between non-neighboring routers. BGP session is used to exchange network reachability information. Each router informs its neighbor what address ranges it knows how to route to, along with ancillary information that is used to make the decision of whether this router will actually be used to route that part of the address space. In this subsection, we will describe how BGP session maintenance work and we will point out some of its problem.

3.1. Running over TCP

A major design choice of BGP is that the protocol runs over TCP. Exchanging connectivity information over reliable transport protocol has a number of advantages and possibly a couple of drawbacks. The Fig. 5 illustrates the Architecture of BGP run over TCP.

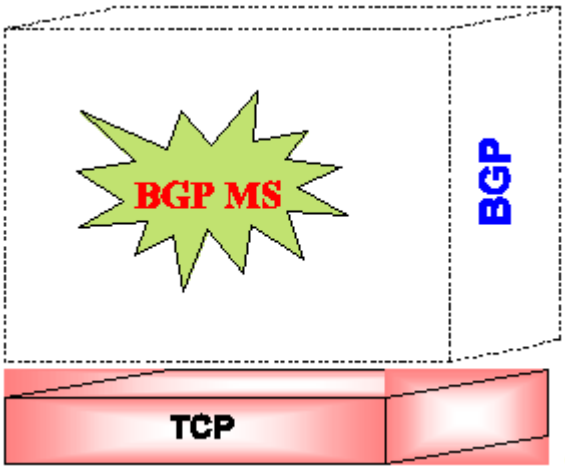

Figure 5: BGP Session Maintenance (BGP MS) run over TCP

Delegating all error control functions to TCP makes the protocol much simpler. There is no need to design complex error recovery mechanisms as well as no need to couple the size of BGP messages with the size of IP datagrams. Using a TCP connection provides only an indication that either the link is broken, or it is functional. Indeed, TCP will give an error indication only if one tries to send data over TCP connection and if these data are no acknowledged by the recipient after several retransmission attempts.

However, we need send probe messages at regular intervals in order to get reachability information, i.e., check that the link to the neighbour is still operational. Whether the link has gone down, TCP will not be able to transmit the probe messages and will signal an error. As TCP is a very resilient protocol, one can be sure that some data will continue to pass even if the link is barely functional and the error rate quite high.

The choice of running over TCP has an effect on the volume of data exchanged between the routers. As messages are reliably transmitted, one can use incremental update instead of retransmitting the whole table at regular intervals. Thereby, BGP requires that only a fraction

of reachability information that changed be transmitted.

TCP provides for a "reliable byte stream" between the connected programs, while routing protocols like BGP actually exchange routing messages. The BGP protocol must thus include a delimitation function that will separate the byte stream in a set of independent messages. This is done very simply by attaching before all BGP messages a fixed-length header that includes the length of the routing message.

3.2 BGP Session

A BGP speaker maintains connections with neighboring speakers through a kind of Managed Session (MS) - commonly referred as BGP sessions. A BGP speaker uses BGP MS to distributed network reachability information to those neighbors and to select its own best or preferred paths to destinations based on the reachability information learned from them - the Fig. 6 illustrates the scenarios described above. Essentially, these path vectors are exchanged (by using a BGP massage called UPDATE) between adjacent routers via TCP connections. Each BGP speaker not only makes its own routing information, but also calculates its best routes towards all reachable destination networks based on routes advertised by neighboring routers. BGP routing information includes the complete route to each destination, and it is used in order to keep a database of network reachability information, which is exchanged with other BGP peers. Besides, network reachability information is used so as to create a graph of AS connectivity. This allows BGP to take away routing loops and enforce policy decisions at the AS level.

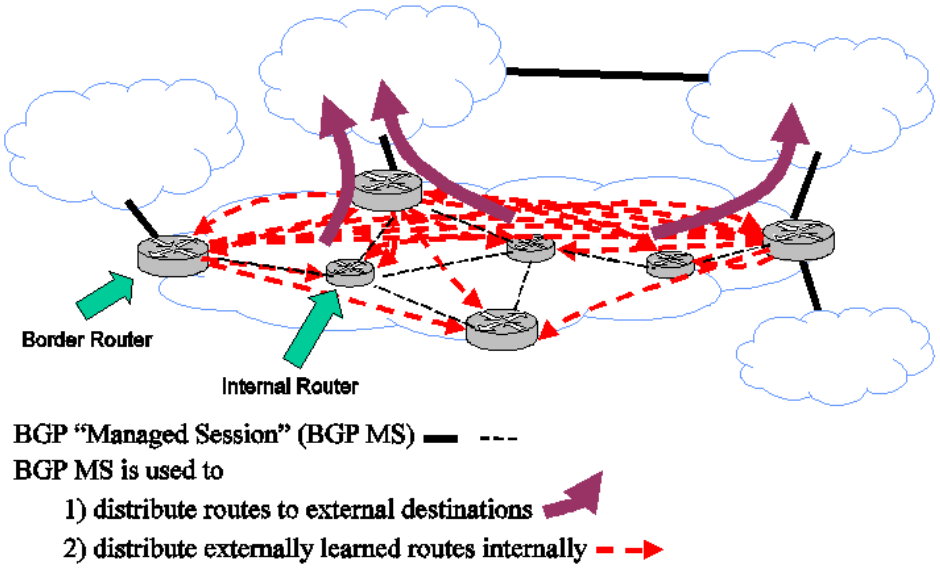

Figure 6: A BGP speaker maintains connections with neighbouring speakers through a kind of Managed Session (MS)

If no UPDATE messages are exchanged, BGP MS uses a packet called KEEPALIVE, which is exchanged periodically, in order to hold open a connection between two peers. However, merely sending regular probes does not provide a complete guarantee that the connection is functional. It is need to check that messages are arriving regularly from the peer. For that, BGP MS uses a mechanism called Hold Time, i.e., the maximum delay during

which the peer should have to wait between successive messages. Failure to receive a message during this delay will indicate that the peer has ceased to function properly, even though the TCP connection may remain operational. However, BGP failure mechanism cannot detect failures before at least a few tens of seconds.

Reachability information is exchanged by establishing a session between two BGP speaking router, relying on an underlying TCP connection on port 179. After a session is set up, first all best routes are announced to the neighbors. Afterwards, update messages are only sent whenever the current best route changes.

There are two flavors of BGP session: eBGP session and iBGP session. eBGP session is used on inter-domain links in order to connect routers of different ASes, i.e., eBGP session is used to announce reachable prefixes. On the other hand, iBGP session is used to either propagate reachability information learned from external router or announce modification inside an AS to external router via border router.

A BGP speaker uses BGP sessions as well to select its own best or preferred paths to destinations based on the reachability information learned from neighboring speaker. Such a decision is based on the distributed shortest-path computations using the Bellman-Ford algorithm. BGP is categorized as path vector protocol because BGP speakers can use a number of path attributes to select preferred paths.

Essentially, these path vectors are exchanged between adjacent routers via TCP connections. Each BGP speaker not only makes its own routing information, but also calculates its best routes towards all reachable destination networks based on routes advertised by neighboring routers. BGP routing information includes the complete route to each destination, and it is used in order to keep a database of network reachability information, which is exchanged with other BGP peers. Besides, network reachability information is used to create a graph of AS connectivity. This allows BGP to take away routing loops and enforce policy decisions at the AS level. These reachability information is exchanged by using a BGP message so-called UPDATE.

3.2.1 BGP Messages

There are four types of messages that can be exchanged between two BGP peers:

Open Message:

Before a BGP session can be used to exchange routing information, a connection must first be established between BGP peers. This process begins with the creation of a TCP connection between the peers. Once this is done, the BGP peers will attempt to create a BGP session by exchanging BGP Open messages.

The OPEN message has two main purposes. The first is identification and initiation of a link between the two devices; it allows one peer to tell the other ``I am a BGP speaker named X on autonomous system Y, and I want to start exchanging BGP information with you". The second

is negotiation of session parameters. These are the terms by which the BGP session will be conducted. One important parameter negotiated using Open message is the method that each device wants to use for authentication. The importance of BGP means that authentication is essential, to avoid bad information or a malicious person from disrupting routes.

Each BGP receiving an OPEN message should process it. If its contents are acceptable, including the parameters the other device wants to use, it responds with a KEEPALIVE message as an acknowledgment. Each peer must send an OPEN and receive a KEEPALIVE acknowledgment for the BGP session to be initialized. If either is not willing to accept the terms of the OPEN, the session is not established. In that case, a Notification message may be sent to convey the nature of the problem.

Update Message:

Once BGP speakers have made contact and a session has been established using OPEN messages, the peers begin the actual process of exchanging routing information. Each BGP router uses its BGP Decision Process to select certain routes to be advertised to its peer. This information is then placed into BGP UPDATE messages, which are sent to every BGP peer for which a session has been established. These messages are the way that network reachability knowledge is propagated around the internetwork.

Each Update message contains one or both of either Route Advertisement (the characteristics of a single route) or Route Withdrawal (a list of networks that are no longer reachable). Only one route can be advertised in an UPDATE message, but several can be withdrawn. This is because withdrawing a route is simple; it requires simply the address of the network for which the route is being removed. In contrast, a route advertisement requires a fairly complex set of path attributes to be described, which takes up a significant amount of space. Besides, it is possible for an UPDATE message to only specify withdrawn routes and not advertise a route at all.

Because of the amount of information it contains, and the complexity of that information, BGP UPDATE messages use one of the most complicated structures in all of TCP/IP. Since a single route may be associated with more than one network, there can be more than one prefix in the NLRI field of one UPDATE message. In that case, specifying multiple network prefixes in the same UPDATE is more efficient than generating a new one for each network.

Notification Message:

Once established, a BGP session will remain open for a considerable period of time, allowing routing information to be exchanged between peers on a regular basis. During the course of operation, certain error conditions may crop up that may interfere with normal communication between BGP peers. Some of these are serious enough that the BGP session must be terminated. When this occurs, the peer detecting the error will inform its peer of the nature of the problem by sending it a BGP NOTIFICATION message, and then close the connection.

Keepalive Message:

If no UPDATE messages are exchanged, BGP uses a packet called KEEPALIVE, which is exchanged periodically, in order to hold open a connection between two peers. However, merely sending regular probes does not provide a complete guarantee that the connection is functional. It is need to check that messages are arriving regularly from the peer. For that, BGP uses a mechanism called Hold Time, i.e., the maximum delay during which the peer should have to wait between successive messages. Failure to receive a message during this delay will indicate that the peer has ceased to function properly, even though the TCP connection may remain operational. The BGP specification [6] recommends a maximum spacing of one third of the Hold Time between two KEEPALIVE messages to indicate that a peer is still operating normally and, consequently, to keep the BGP session alive. Besides, a Hold Timer must be no shorter than three seconds. The BGP specification recommends a value of Hold Timer of ninety seconds, which is used in the most BGP implementations. However, BGP failure mechanism cannot detect failures before at least a few tens of seconds.

3.2.2 BGP Finite State Machine

There are seven states in which a BGP connection can be established. The rules on how to go between the states of BGP FSM is given by the state transition diagram that is depicted in Fig. 7.

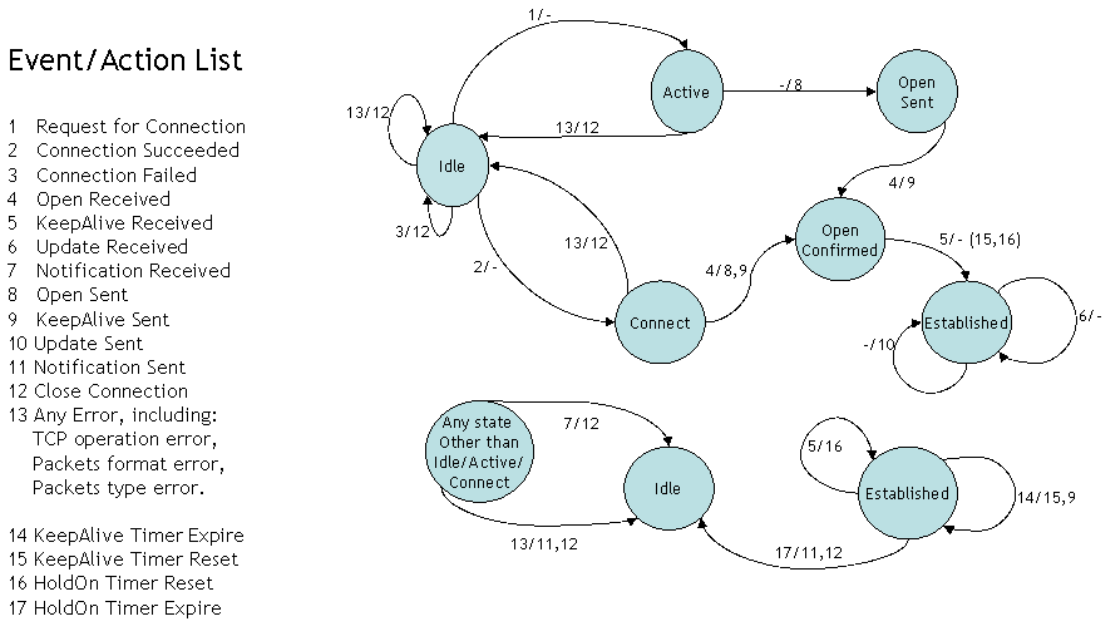

Figure 7: The BGP FSM is a set of rules that is applied to a BGP speaker's set of configured peers for the BGP operation. The BGP FSM must be initiated and maintained for each new incoming and outgoing peer connection. However, in steady state operation, there will be only one BGP FSM per connection per peer. Besides, there are seven states in which a BGP connection can be established.

A peer starts a BGP session by initiating a TCP connection to other BGP peer. If the TCP connection succeeds, than it takes place the BGP establish process.

After sending the OPEN packet, a BGP peer needs to receive an OPEN packet from its peers. Upon receipt of this packet, the BGP peer carries on validation fields in order to check up on whether this is a valid packet. If the validation is succeeds, the BGP peer send a KEEPALIVE packet to its peers. This packet is used to confirm that the BGP peer is still alive and the association can be established. After both sides receiving the second KEEPALIVE packet of the four-way handshake, the session is established and the peers can exchange UPDATE packets.

Aiming to check up on the presence of connection collision, upon receipt of an OPEN message, the local system must check up all of its connections which are in the OPENCONFIRM state in order to detect the presence of association collision. If an association collision takes place, one of them must be closed. The convention is to compare the BGP Identifier of the peers involved in the collision and to retain only the connection initiated by the BGP speaker with the higher-valued BGP Identifier.

4. Robustness of BGP Sessions

The TCP design choice has also been criticized sometimes for its sensitivity to network congestion. Most implementations of the TCP include the slow start and congestion avoidance algorithm. When losses are observed, TCP immediately reduces its rate of emission by shirk the congestion control window, which is the number of bytes that can be sent on a connection before an acknowledgement is received. However, consider the BGP connection; it may well be carrying the routing updates that are needed to cure the congestion. Slowing down this transmission is very counterproductive, because routing protocol will thus converge very slowly. Labovitz et al. [8] noticed that the KEEPALIVE messages were delayed during periods of peak network usage beyond the BGP hold timer, which led the BGP session to fail. The work presented by Shaikh et al. [9] shows the need of differentiate somehow routing protocol messages from normal data traffic. In addition, other researches (e.g., [10]) have demonstrated that the conservative behavior of TCP retransmissions actually aggravates the instability of BGP session when network failures takes place.

4.1 BGP graceful shutdown

When a BGP session of the router under maintenance is shut down, the router removes the routes and then triggers the BGP convergence on its BGP peers. Dubois et al [11] proposed an approach to graceful shutdown a BGP session, which they refer to as *BGP graceful shutdown*. The goal of BGP graceful shutdown is to initiate the BGP convergence to find the alternate paths before the nominal paths are removed. Accordingly, all routers learn and use the alternate paths (if one is provided) and fewer packets are lost during the BGP convergence process since at any time, all routers have a valid path. Afterward, the nominal BGP session can be shut down. As a result, it may be possible to minimize packet loss when the BGP session is re-established following the maintenance.

4.2 Graceful restart mechanism

Xian et al. [12] proposed a modification of TCP so as to improve the robusteness of iBGP sessions. *Gracefu Restart mechanism* is a mechanism that was proposed to allow a router to continue using the routes learned from its neighbor even when its session with the neighbor is down [13]. However, we cannot assume that the physical link is functioning during a BGP session failure. Wang et al. [14] proposed a Bloom-filter based approach that can speed up the table exchange after a session recovers from a failure. Bonaventure et al. [15] proposed an approach to forward packets to an alternative egress point via a protection tunnel, when the underlying link of an eBGP session fails.

The robustness of BGP sessions is an important issue at present for security reasons. This is because a BGP session will fail if the TCP connection fails due to an attack. To address this problem, some operational solutions are possible. This is addressed in the next section.

5. BGP Security Problem

As BGP runs over TCP, it is protected against misordered, lost, or replayed packets to the extent that the TCP sequence number management facility is secure. All BGP protocol exchanges can be authenticated to guarantee that only trusted routers participate in the AS's routing. BGP provides an authentication security mechanism through *Message Digest 5* (MD5) [16] signatures for each TCP connection. However, there is no prescribed key management scheme and there is no facility for sequence numbering of BGP messages. Therefore, BGP is highly vulnerable to a variety of malicious attacks due to the lack of a mechanism to verify the authenticity and legitimacy of BGP control traffic.

5.1 Protecting the BGP Session

This is the main concern of many operators that the vulnerabilities of BGP may cause large disruptions of service under possible attacks [17, 18]. Besides, exploiting these vulnerabilities to conduct attacks, measurement studies have shown that miscofigurations of BGP routers are common events [19]. The first solution to improve the security of BGP has been proposed in S-BGP [20]. However, the main problems of S-BGP are the cost (CPU, memory and bandwidth) of producing, storing and distributing attestations, and the need to bootstrap the public key infrastructure (PKI) [21]. Several alternate solutions have been proposed to get around this problem such as [22, 23, 24].

Therefore, protecting the connection between two BGP speakers is based on both protecting the underlying TCP session and implementation defenses that protect the BGP session itself.

5.1.1 Protecting the underlying TCP session

Some other projects have explored the idea of extending TCP protocol in order to overcome connection failover. One such system [25] extends the TCP with an option that enables migration of connections from one host to another, whereas [26] proposes an architecture that allows a replicated service to survive crashes without breaking its TCP connections. Zandy et al. [27] proposed two user-level techniques, called rocks and racks, which provide transparent network connection mobility, including mechanisms for failure detection and connection recovery. However, this system makes use of the same slow standard TCP keep-alive in order to detect failures on an established connection. The TCP-Authentication Only (TCP-AO) [28], a proposal by the IETF, replaces TCP-MD5. This approach provides replay protection and allows for rekeying during a TCP connection without any packet loss.

5.1.2 Mechanism for securing BGP session

Smith et al. [29, 30] proposes five countermeasures by altering both BGP session environment and the BGP protocol message attributes. The goal of two of these countermeasures is to protect BGP control messages by encrypting all BGP data between peers and including two sequence numbers to impose a total ordering on the messages. The goal of the other countermeasures is to provide both protection for UPDATE messages and to provide an UPDATE sequence number or timestamp, a new path attribute (called PREDECESSOR), which identifies the last AS before the destination AS, and an approach

for peers have digital signatures of all fields in the UPDATE message whose values are fixed.

Many proposals have recommend the use if IPsec [31, 32] as mechanism for securing BGP session. The IPsec is a suite of protocols provides security at the network layer. The IPsec Authentication Header protocol (AH) [33] and Encapsulating Security Payload (ESP) [34] protocol provide packet-level security with differing guarantees. All of these services are used so as to guarantee the confidentiality and authenticity of BGP messages passed over IP between two peers.

The Generalized TTL Security Mechanism (GTSM) provides a method for protecting peers from remote attacks [35]. This approach relies on the fact that in many of BGP peering session, the two peers are adjacent to each other. So, this approach makes use of the notion of *Multihop BGP sessions* (that is possible but not common in practice) where peers are more than one hop away from each other). The Time-to-Live (TTL) attribute in an IP packet is set to a value that is decremented at every hop.

Gouda et al. [36] propose a suite of protocols (callled Hop Integrity Protocols) so as to provide security at the IP layer. In this approach, *hop integrity* is the property that peers can detect any modification or replay of exchanged information.

All these solutions are operational palliatives. Accordingly, they do not tackle the root of the problem, i.e., how to conceive robust BGP sessions among BGP routers.

5.1.3 Analyzing BGP session failure

A session failure could introduce a significant amount of routing instability due to the large number of routes withdrawn after the failure and re-exchanged after the session reestablishment.

An approach to analyze BGP session failure is to use analytical models and testbed experiments. Shaikh et al. [37] demonstrated that the probability of session failure depends on the congestion level. Xiao et al. [38] proposed a probability model for iBGP session failures and analyzed the effects of BGP timers and TCP retransmission behaviors on session failures. Bonaventure et al. [15] demonstrated that eBGP peering link failures were common in one transit ISP, whereas Wang et al. [39] studied whether these link failures actually caused BGP session failures or what other factors could cause BGP session failures.

Inferring the failures of remote BGP sessions remains an open question.

6. Converge Problem

One of the primary metrics used to evaluate the effectiveness of a routing protocol is convergence time. Generally, the convergence time of a distributed system is the time required for state of the system to become stable after a change occurs within the system. The BGP is considered to be in a stable state when no UPDATE messages are actively generated or sent and all speakers have stable RIBs. A variety of changes that can trigger the routing activities, such as: failure or repair of a physical link; a down BGP session; a router crash; a

link failure; policy changes in originating or transiting ASes, or addition/deletion of network prefixes.

Accordingly, BGP speakers will propagate such a change via UPDATE messages through the network. The convergence time measures the length of time for the system to return to a stable state. Longer convergence times reflect increased network instability and decreased network reliability, and may cause severe network performance problems such as delayed packet forwarding, prolonged latencies, increase of packet loss rate, and increase of network congestions.

During BGP convergence, routers may need to exchange several advertisements concerning the same prefix. To avoid storms of BGP advertisements, BGP includes a minimum per-prefix advertisement timer to limit the BGP UPDATE rate. This timer is so-called Minimum Route-Advertisement interval timer (MRAI timer), with a recommended default value of 30 seconds. This reduces the number of BGP advertisements exchanged, but may cause important BGP advertisements to be unnecessary delayed.

The TCP design choice has also been criticized sometimes for its sensitivity to network congestion. Most implementations of the TCP include the slow start and congestion avoidance algorithm. When losses are observed, TCP immediately reduces its rate of emission by shirk the congestion control window, which is the number of bytes that can be sent on a connection before an acknowledgement is received. However, consider the BGP connection; it may well be carrying the routing updates that are needed to cure the congestion. Slowing down this transmission is very counterproductive, because routing protocol will thus converge very slowly. Labovitz et al. [8] noticed that the KEEPALIVE messages were delayed during periods of peak network usage beyond the BGP hold timer, which led the BGP session to fail. The work presented in [9] shows the need of differentiate somehow routing protocol messages from normal data traffic. In addition, other researches (e.g., [10]) have demonstrated that the conservative behavior of TCP retransmissions actually aggravates the instability of BGP session when network failures takes place.

Griffin and Presmore showed in [40] that the arbitrary 30s value of MRAI has a huge impact on BGP convergence time. To deal with flapping routers that regularly advertise and shortly after withdraw their routes, many routers implement BGP route flap damping [41]. However, this technique can increase BGP convergence time [42]. Other authors (such as [43]) have proposed modifications to reduce the BGP convergence time in case of failure. Other solutions such as BGP-RCN [44] and EPIC [45] improve the convergence of BGP and also reduce the number of BGP messages exchanged during the convergence.

7. Towards a more scalable interdomain routing

The growth of the Internet has introduced considerable complexity into interdomain routing, as features have been put into BGP to cope with more flexibility and large scale. Routing protocol behavior has become increasingly unpredictable and error prone due to this complexity.

It is useful, therefore, to pose the question as to whether we can continue to make incremental changes to the BGP protocol and routing platforms, or whether the pace of growth will, at some point in time, mandate the adoption of a routing architecture that is better attuned to the evolving requirements of the Internet.

Huston [46] argues that BGP may already be too monolithic a protocol in that it simultaneously performs multiple distinct functions, such as exchanging reachable prefixes, learning about (local) topology, binding prefixes to paths, and implementing routing policy. He argues that interdomain routing might be more scalable if these functions where performed by separate protocols.

An alternative approach to inter-domain routing is to separate the different functions in well-defined modules such as connectivity maintenance, address reachability, and policy negotiation. For instance, a connectivity protocol can be employed to identify all viable paths between a source and a destination domain. A policy negotiation protocol can be applied to guarantee that there are consistent sequences of per-domain forwarding decisions that will pass traffic from the source domain to the destination domain. An address reachability protocol can be exploited to associate a collection of address prefixes with each destination domains. The Fig. 8 depicted the framework described above.

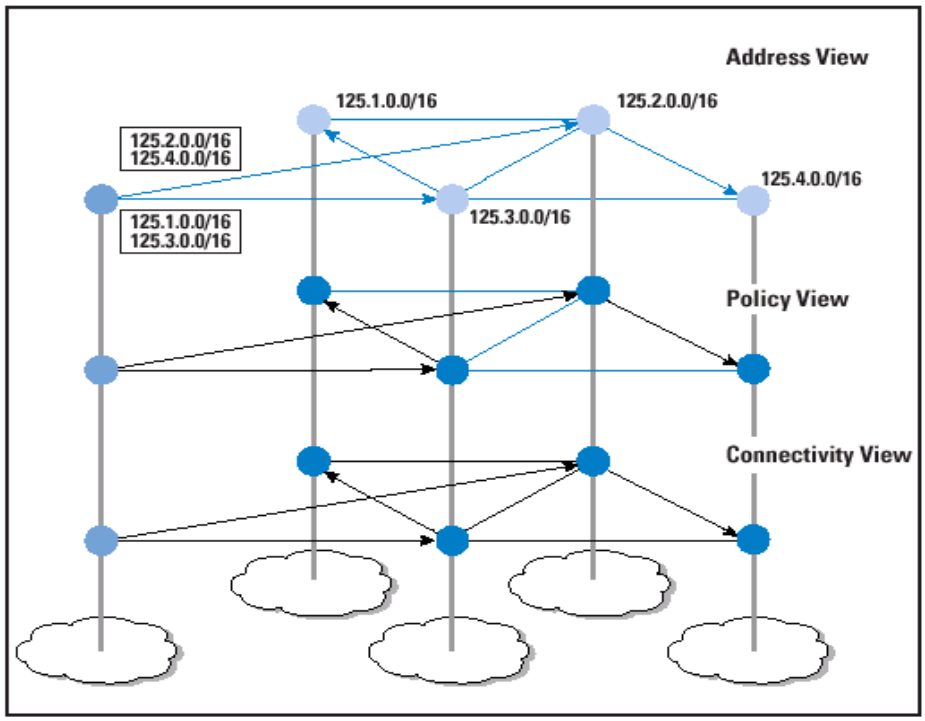

Figure 8: A Multi-Tiered Approach to Interdomain Routing

Based in this insight, several authors have proposed decoupling BGP in different protocols, such as [47], [48], [49] and [50]. Feamster et al. [47] proposed removed Inter-AS routing decision process from routers into a separate logically-centralized control function refer to as the Route Control Platform (RCP). Yang [51] proposed a design of a new Internet routing architecture (NIRA). NIRA is an architecture that is designed to provide a use the

ability to choose the sequence of Internet service providers a packet traverses. Snoeren et al. [48] Proposed an approach to separate forwarding policy from route discovery which can allow users to select among the possibly many inter-AS paths available to them and also enable ISPs to more effectively manage the end-to-end behavior of their customers' traffic.

7.1 Decoupling BGP from Maintenance Session

Convergence after a BGP process failure usually takes longer than with other routing protocols, resulting in an outage of greater duration. Moreover, the effect of failed BGP process can propagate across multiple networks, instead of being restricted to just one domain. Besides, BGP "managed session" cannot detect failures before at least a few tens of seconds.

Cavalli et al. [50] proposed an approach to decouple BGP from session, called Managed Session Protocol (MSP). Compared to routing information exchange and routing table computation, MSP can be kept rather simple and would typically not be subject to advanced configuration. As a result it should be less prone to fail, which is advantageous considering that the MSP is the module that exposes the application end-point to an external peer. Moreover, the session manager provides an elegant way to hide the internal structure of multiple information and database management. In addition, MSP provides a certain degree of freedom when it comes to how it is implemented. The session manager may either be mapped onto a processing element, or different placement strategies can be applied. Accordingly, sophisticated routing protocols, tuned to various specific needs, can be layered on top of MSP. In this way, designers of routing protocols can focus on the more complex demands while being assured that basic connectivity is still being provided by MSP. In addition, the authors proposed the use of multi-session capability. As a result, routing protocol would make use of this capability so as to avoid routing flap. The Fig. 9 depicted the overall architecture of MSP.

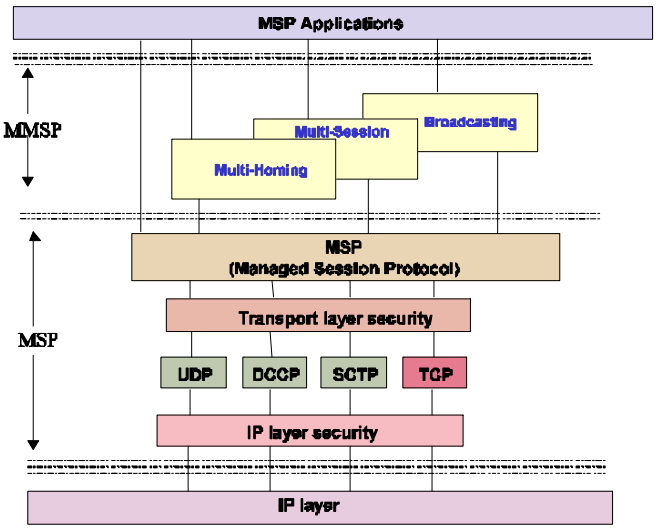

Figure 9: In order to meet the requirements for extensible, architecturally MSP is made up two protocol layers: MSP itself, and Multiple Managed Session Protocol (MMSP) that copes with services such as multisession, multihoming and broadcasting.

We can give a step further in order to reduce more the complexity of routing system, by using some suggestions presented in [49]. Relying on this approach, we can enforce a clean separation of protocol mechanisms (link-state, path-vector, and so on) from routing policy (how routes are described and compared). This may help to reduce more the complexity of routing system.

8. Conclusion

The rapid growth of the Internet in terms of traffic volumes and number of users, combined with an increasing demand for new services, pose new requirements on routers and other network systems when it comes to scalability, flexibility, and robustness. However, due to the monolithic architecture of today router architecture, network management is so complex and a tedious task (it is still arguably a black art – largely a manual process with little systematic methodology and architectural support), and therefore with little support to extensibility.

One of the reasons why network management is so difficult is the complexity of network elements. Routing protocol behavior has become increasingly unpredictable and error prone due to this complexity. For instance, as the Border Gateway Protocol (BGP) provides information for controlling the flow of packets between Autonomous Systems, the protocol plays a critical role in Internet efficiency, reliability, and security. Nonetheless, BGP suffers from several well-know vulnerabilities: it is hard to configure, slow to converge, prone to serious anomalies, vulnerable to malicious attack, difficult to troubleshoot, and overly sensitive to small topology changes. Even though the research community has proposed many solutions, such as static-analysis tools, or tools to detect and diagnose routing problems, we have few meaningful guidelines for solve this problem, i.e., the complexity of the individual network elements.

This survey has looked at the BGP session maintenance and presented some problems and solutions to ensure its robustness. In addition, it exposes areas where it is commonly believed that BGP still needs improvements.

Huston [46] argues that we need to rethink the traditional, monolithic architecture of network system in the direction of coming up with a level of abstraction to design and manage at the network level. We need to come up with a solution that supports flexibility and extensibility while always maintaining backwards compatibility without jeopardizing availability.

The question is: how should we address this issue? How to get a model for decomposing a network system design in order to be used for a general modularization approach?

Recent studies (e.g., [2, 3, 5, [49], [50]) have shown that moving interdomain routing functionalities into a small collection of servers is a promising way to provide less convoluted and more flexible support for interdomain routing control. So, a way to meet the

abovementioned requirements is to investigate distributed and modular designs, where network system is composed of multiple modules (or elements), which communicate through open well-defined interfaces over an internal network.

This approach has several advantages:

1) Scalability is improved because modules can be added as capacity requirements increase.
2) Flexibility comes from the ability to dynamically add, remove and modify modules.
3) Robustness is obtained mainly due to two factors: first, the modularity makes it possible to use redundancy and replication of critical functionality over multiple modules. Second, the modular structure in itself tends to limit the impact of faults in individual modules, and encourages sound engineering design principles.
4) Performance is improved because modules can be replicated and added as capacity requirements increase.

The approach proposed by Sobrinho et al. [49] can by a step in the direction of such model. However, this question continues an important open issue.

Pinpointing the source of a failure of BGP (especially concerning session maintenance) without needing to modify BGP is another important open issue. Besides, the security solutions of BGP session is still an open issue.

References

[1] T. G. Griffin and G. Wilfong, "A Safe Path Vector Protocol," *In. Proc IEEE INFOCOM*, 2000.

[2] L. Gao, T. G. Griffin, and J. Rexford, "Inherently Safe Backup Routing with BGP," *In Proc. IEEE INFOCOM*, April 2001.

[3] A. Feldmann, O. Maennel, Z. M. Mao, A. Berger, and B. Maggs, "Locating Internet Routing Instabilities," *ACM SIGCOMM*, 2004.

[4] J. Wu, Z. M. Mao, J. Rexford, and J. Wang, "Finding a Needle in a Haystack: Pinpointing Significant BGP Routing Changes in an IP Network," *In Proc. Networked Systems Design and Implementation*, 2005.

[5] R. Teixeira and J. Rexford, "A Measurement Framework for Pinpointing Routing Changes," *in Proc. ACM SIGCOMM Network Troubleshooting Workshop*, 2004.

[6] Y. Rekhter and T. Li, "A Border Gateway Protocol 4 (BGP-4)," http://www.ietf.org/rfc/rfc1771.txt, March 1995.

[7] T. G. Griffin, "An introduction to interdomain routing and bgp," SIGCOMM 2001 Tutorial Session, 2001.

[8] C. Labovitz, A. Ahuja, and F. Jahanian, "Experimental Study of Internet Stability and Backbone Failures," *Proc. FTCS-29, 29th International Symposium Fault-Tolerant Comp.*,

pp. 278–85, June 1999.

[9] A. Shaikh, A. Varma, L. Kalampoukas, and R. Dube, "Routing Stability in Congested Networks: Experimentation and Analysis," *Proc. ACM SIGCOMM*, August 2000.

[10] Xiao and K. Nahrstedt, "Reliability Models and Evaluation of Internal BGP Networks," *Proc. IEEE INFOCOM* , March 2004.

[11] N. Dubois, B. Decraene, B. Fondeviole, and Z. Ahmad, "Requirements for planned maintenance of bgp sessions," Internet Draft, July 2005.

[12] L. Xiao and K. Nahrstedt, "Reliability models and evaluation of internal bgp networks," in *INFOCOM*, 2004.

[13] S. Sangli, Y. Rekhter, R. Fernando, J. Scudder, and E. Chen, "Graceful restart mechanism for bgp," Work in Progress, July 2006.

[14] L. Wang, D. Massey, K. Patel, and L. Zhang, "FRTR: A scalable mechanism for global routing table consistency," in *DSN '04: Proceedings of the 2004 International Conference on Dependable Systems and Networks*. Washington, DC, USA: IEEE Computer Society, 2004, p. 465.

[15] O. Bonaventure, C. Filsfils, and P. Francois, "Achieving sub-50 milliseconds recovery upon bgp peering link failures," *IEEE/ACM Trans. Netw.*, vol. 15, no. 5, pp. 1123–1135, 2007.

[16] R. Rivest, "The MD5 Message-Digest Algorithm," Internet Draft IETF - RFC 1321 , April 1992.

[17] O. Nordstrom and C. Dovrolis, "Beware of BGP Attacks," *ACM SIGCOMM Comp. Communication*, 2004.

[18] S. Murphy, "BGP Security Vulnerabilities Analysis," Internet Draft, draft-ietf-idr-bgp-vuln-01.txt, October 2004, work in progress.

[19] R. Mahajan, D. Wetherall, and T. Anderson, "Understanding BGP Misconfigurations," *ACM SIGCOMM* , August 2002.

[20] S. Kent, C. Lynn, and K. Seo, "Secure Border Gatewey Protocol (S-BGP)," *IEEE JSAC*, April 2000.

[21] M. Yannuzzi, X. Masip-Bruin, and O. Bonaventure, "Open Issues in Interdomain Routing: A Survey," *IEEE Network, Special Issue on Interdomain Routing*, December 2005.

[22] G. G. et al., "Working around BGP: An Incremental Approach to Improving Security and Accuracy of Interdomain Routing," *NDSS*, February 2003.

[23] Y.-C. Hu, A. Pering, and M. Sirbu, "SPV: Secure Path Vector Routing for Securing BGP," *ACM SIGCOMM 2004*, Sept. 2004.

[24] M. Zhao, S. Smith, and D. Nicol, "The Performance Impact of BGP Security," *IEEE Network, Special Issue on Interdomain Routing*, December 2005.

[25] A. Snoeren, D. Andersen, and H. Balakrishnan, "Fine-grained Failover Using Connection Migration," *In Proc. 3rd USENIX symp. on Internet Technologies and Systems (USITS)*, pp. 97–108, 2001.

[26] D. Zagorodnov, K. Marzullo, L. Alvisi, and T. C. Bressoud, "Engineering Fault-Tolerant TCP/IP Servers Using FT-TCP ," *In Proc. IEEE Intl. Conf. on Dependable Systems and Networks (DSN)*, pp. 393–402, 2003.

[27] V. Zandy and B. Miller, "Reliable Network Connections," *In Proc. ACM MobiCom*, 2002.

[28] J. Touch, A. Mankin, and R. Bonica, The TCP Authentication Option, Internet draft, July 2009.

[29] B. R. Smith and J. Garcia-Luna-Aceves, "Securing the border gateway routing protocol," in *Proceedings of Global Internet*, 1996.

[30] B. R. Smith and J. J. Garcia-Luna-Aceves, "Efficient security mechanisms for the border gateway routing protocol," *Computer Communications*, 1998.

[31] S. Kent and K. Seo, "Security architecture for internet protocol," RFC 4301, December 2005.

[32] R. Thayer, N. Doraswamy, and R. Glenn, "Ip security document roadmap," RFC 2411, November 1998.

[33] S. Kent, "Ip encapsulating security payload (esp)," RFC 4303, Dec. 2005.

[34] IP Authetication Header, RFC 4302, Dec. 2005.

[35] V. Gill, J. Heasley, and D. Meyer, "The generalized ttl security mechanism (gtsm)," RFC 3682, Feb. 2004.

[36] M. G. Gouda, E. N. Elnozahy, C. t. Huang, and T. M. McGuire, "Hop integrity," *Computer Networks*, 2000.

[37] A. Shaikh, A. Varma, L. Kalampoukas, and R. Dube, "Routing stability in congested networks: Experimentation and analysis," in *Proc. ACM SIGCOMM*, 2000, pp. 163–174.

[38] L. Xiao and K. Nahrstedt, Reliability models and evaluation of internal bgp networks, in *Proceedings of IEEE INFOCOM*, 2004.

[39] L. Wang and M. Saranu, Understanding BGP Session Failures in a Large ISP, in *Proceedings of IEEE INFOCOM*, 2007, pp. 348 – 356.

[40] T. Griffin and B. Presmore, An Experimental Analysis of BGP Convergence Time, *Proc. IEEE ICNP*, November 2001.

[41] C. Villamizar, R. Chandra, and R. Govindan, BGP Route Flap Damping, RFC 2439, November 1998.

[42] Z. M. et al, Route Flap Damping Exacerbates Internet Routing Convergence Time, *Proc.*

ACM SIGCOMM, 2002.

[43] A. Bremler-barr, Y. Afek, and S. Schwarz, Improve BGP Convergence via Ghost Flushing, *Proc. IEEE INFOCOM*, 2002.

[44] D. P. et al, BGP-RCN: Improving BGP Convergence through Root Cause Notification, *Computer Network*, vol. 48, no. 2, pp. 175–94, 2005.

[45] J. C. et al, Limiting Path Exploration in BGP, *Proc INFOCOM*, 2005.

[46] G. Huston, Scaling interdomain routing, Internet Protocol Journal, Dec. 2001.

[47] N. Feamster, H. Balakrishnan, J. Rexford, A. Shaikh, and J. van der Merwe, The Case for Separating Routing from Routers, *ACM SIGCOMM Workshop on Future Directions in Network Architecture (FDNA)*, September 2004.

[48] A. C. Snoeren and B. Raghavan, "Decoupling policy from mechanism in internet routing," *SIGCOMM Comput. Commun. Rev.*, vol. 34, no. 1, pp. 81–86, 2004.

[49] T. G. Griffin and J. L. Sobrinho, "Metarouting," *SIGCOMM 2005*, 2005.

[50] A. Cavalli, T. G. Griffin, and D. Vieira, "MSP: A Novel Maintenance Session Protocol," 14th IEEE International Conference on Networks ICON 2006, July 2006.

[51] X. Yang, "Nira: a new internet routing architecture," in *FDNA '03:* Proceedings of the *ACM SIGCOMM Workshop on Future Directions in Network Architecture*. New York, NY, USA: ACM, 2003, pp. 301–312.