CHAPTER 2

INTERNET ROUTING WITH BGP: LITERATURE SURVEY

This chapter presents an in-depth literature survey of the Internet routing using BGP and its issues. The chapter starts with basic routing problems and the manner in which these problems have been addressed earlier, and then the further issues arisen because of the limitations of addressing technique.

2.1 INTRODUCTION

In 1990, the ARPANET services were withdrawn. When NSFNET was decommissioned, and it became possible to carry commercial traffic on the Internet, which led to the commercialization of the Internet during the mid of 1990s. The Internet kept on growing, driven by ever increasing amounts of online information, commerce, entertainment and social networking [Wikipedia (2012)].

In the beginning of 1990s the Internet community started considering the growth rate of the Internet as a problem; it has been realized well that if this problem is not addressed well in time it would become unmanageable later. In December 1990 Internet engineering task force decided to support the idea of restructuring the Internet protocol addressing i.e. ipv4 to increase the lifespan of IP addresses.

Between years 1993 to 1995 notion of eliminating the class full addressing was conceived and it was proposed to introduce a new idea of classless addresses, this was to be done with the varying subnet masks which were fixed until now, based on the class type as for class A it was first eight bits, for class B it was first sixteen bits, and for class C it was first twenty four bits [Li and Fuller (2006)].
The Classless Inter-Domain Routing popularly known as CIDR was to select the number of subnet mask bits based upon the size of the addresses required by the organization. This means that now the subnet masks could range from $n$-bits to 32-bits. For instance if an organization requires addresses for its network whose capacity was to have only 60 computers then subnet masks bits could be first twenty six bits instead of having first 24 bits [Rekhtar and Li (1995)].

The packet forwarding process is the moving packets from one network segment to another by routers in a computer network. The internetworking devices are meant to support some forwarding models. At intermediate nodes which are routers, with multiple outgoing links, forwarding process for a given packet requires a decision making to forward it on the right link. This decision making may vary from the simple one to the complex one. Since a forwarding decision is done for each packet handled by a router, the total time required for this can become a major limiting factor in overall network performance.

The decision making process may be completed in one or two ways, as mentioned below:

- **Routing** – this process uses the information encoded in a device’s address to infer its location on the network.
- **Bridging** – this process depends heavily on broadcasting method for locating unknown addresses, if not known.

The Routing process usually directs packets on the basis of entry found in its routing tables which maintain a record of the routes to destination prefixes. Routing tables are populated by using routing algorithms to compute the best path from this router to the destination. Typically a best path is the one which has the least cost [Routing (2012)]. There are mainly two categories of routing algorithms-

- **Global Routing Algorithm**
- **Decentralized Routing Algorithm**

The Global Routing Algorithm selects the path which offers minimum cost between a source and the destination using topological knowledge of the network. This algorithm takes information from all the nodes present in the network. It calculates the best path and shares this knowledge with all the nodes in the network. This algorithm is known as to as Link State Algorithm because it has the complete view of the network [Kurose and Ross (2009)].
The decentralized or distributed algorithm calculates the least cost path in an iterative manner. None of the node in the network has the complete picture of the topological information; neither it knows about the cost of all links. Each node has information of only links that are directly connected to it. The overall information of the network is gained through an iterative process by a series of updates exchanged with the neighbors. The algorithm is also known as Distance Vector Routing [Kurose and Ross (2009)].

When the routing is performed in the Internet, it is divided in two categories;

- First is the routing within an Autonomous System OR Intradomain Routing.
- Second is the routing between Autonomous Systems OR Interdomain Routing.

Intra-AS or Intradomain routing is used to decide the way to perform routing within an Autonomous System (AS). There are three protocols mainly popular in this category

- The Routing Information Protocol (RIP),
- The Intermediate System-Intermediate System (IS-IS), and
- The OSPF (Open Shortest Path First).

Inter-AS or Interdomain routing is performed between Autonomous Systems (ASes). There is only one protocol available now-a-days:

- The Border Gateway Protocol (BGP)

Autonomous System is a network of routers which is under the control of a single administrative team. It uses the interior gateway protocol to route packets within the Autonomous System. To route packet outside the interior network exterior gateway protocol is used. Companies and organizations might own more than one Autonomous System, but the idea is that each Autonomous System is managed independently with respect to BGP. A unique AS Number (ASN) is allocated to each AS for use in BGP routing. AS Numbers are important because these are used to uniquely identify each network on the Internet [Rekhter et al. (2006)].

The BGP is very complex protocol. The primary role of the BGP is to exchange Network Layer Reachability Information (NLRI) with other BGP speakers. This network layer reachability
information includes the complete path sequence information in the list of Autonomous Systems (ASes) which are traversed in order to reach the destination prefix. The current version of the BGP is 4. The BGP supports a new set of mechanisms for CIDR. The BGP runs over a reliable transport layer protocol TCP. The BGP stores routes at Route Information Base (RIB). When a BGP speaker decides to advertise the route, it may add or modify the path attributes of the route before advertising to a peer. There are four types of messages which show the overall functioning of the BGP: Open messages, Keepalive messages, Notification messages, and Update messages.

2.2 MINIMUM ROUTE ADVERTISEMENT INTERVAL TIMER

When the number of control messages exchanged between peers becomes very high, the need to control this exchange becomes a necessity. The method for the same, suggested in RFC 1771, is to use a special parameter which is known as the Minimum Route Advertisement Interval (MRAI) timer. The parameter MRAI decides the minimum amount of time that must pass between advertisements of routes to a particular destination from a BGP speaker. This MRAI which is a rate limiting procedure applies on a per-destination basis, although the value of MRAI is set on a per BGP peer basis [Rekhter et al. (2006)]. The introduction of Minimum Route Advertisement Interval (MRAI) timer in BGP was to serve the purpose of controlling instabilities by suppressing the multiplication of advertisement messages. Use of the MRAI essentially imposes a restriction on updates for a given address prefix, limiting the frequency of updates changes to a maximum of one per MRAI interval. Once an update has been sent by a BGP speaker to its BGP peers for a given prefix all further updates for that prefix are to be delayed until the previously set MRAI timer expires.

While the MRAI has been deployed widely in the Internet, the common implementation of the MRAI behavior is different in one way or another. Instead of using a per-prefix update suppression timer, most commonly seen implementations of MRAI in BGP uses a per-peer announcement timer. All updates to a BGP peer are put in a queue by the BGP speaker. For the same prefix, the successive BGP updates cause previously queued updates to be flushed from the
queue. At the expiration of the timer set for MRAI the entire output queue state is sent to the BGP peer, and the queue is flushed. The timer is restarted and the queue is reset for another MRAI interval [Mao et al. 2002]).

2.3 BGP ROUTE FLAP DAMPING ALGORITHMS

Route flap dampening is a mechanism to reduce the propagation of relatively unstable routes in the network. The route flap damping is used to suppress the route changes which are caused when link flaps; it distinguishes between unstable routes and stable routes, and gives important information about the stability of the network. Route flap dampening reduces the impact of configuration errors, link failures and software defects. The main goal is to reduce the number of updates in the network so as to decrease the load on the router. It assigns a penalty to the unstable routes then the routes are said to be suppressed and not advertised further. This penalty decays with the rate according to its half life; half life is the penalty to be reduced by half. When the penalty is decreased up to the threshold reuse limit, the route is advertised and reused again. [Villamizar et al. (1998)].

The MRAI timer was designed to limit announcements of route changes during the convergence period, but it cannot control route instabilities which are caused by external factors. Due to these external factors that have potential to cause the flaps on some routes. Route flap damping was designed to control flaps and works as follows [Mao et al. 2002)]:

For each prefix P and for each peer or neighbor N, a BGP router maintains a penalty \( p[ P ; N ] \). The penalty changes according to two simple rules:

1. Whenever a peer N’s route to prefix P changes (either the route transitions from being available to being unavailable, vice-versa, or from one route to a better route, the router increments \( p[ P ; N ] \). This increment is fixed, depending on the type of the change.

2. \( p[ P ; N ] \) decays exponentially with time according to the equation

\[
p[ P ; N ](t') = p[ P ; N ](t) e^{-\lambda(t' - t)}
\]

Where \( \lambda \) is a configurable parameter.
A value of $\lambda$ usually expressed using a half-life parameter $H$ – the time for the penalty to decay to half of its value. We can obtain $\lambda$ from $H$ using the above equation, it is $e^{-\lambda H} = 0.5$. The penalty maintains an exponentially decaying of instability history for a particular route from a particular peer. When a router receives a route from $N$ to prefix $P$, it first updates the penalty $p[P; N]$, as in figure 2.1. A suppression threshold is the value of the penalty above which a route is suppressed. A reuse threshold is the value below which the route is considered reusable.

![Route Flap Damping Penalty Graph](image)

**Figure 2.1: Route Flap Damping Penalty Graph [Mao et al. (2002)]**

### 2.4 ROUTE FLAP DAMPING+

The authors Duan et al. [Duan et al. (2007)] have put their efforts to classify the messages exchanged during a link failure or path withdrawn, shown in table 2.1 after a link, in the network given is figure 2.3, between node 0 and d fails. These authors have categorized the events to be called as path exploration or route flaps based on the degree of preference. Once a route with a higher preference is replaced by a route with a lower preference, the route with a higher preference will not be advertised by the node again. Therefore, the neighbors of the node would
only see the routes with higher preferences once in a BGP path exploration. On the other hand, in a route flap, a route with a higher preference may be seen by the neighbors twice. This observation could be used to distinguish a BGP path exploration from a route flap.

Let $Pr$ denote the preference of a route $r$ and $Pr1 < Pr2$ to indicate that the route $r1$ is of lower preference compared to the route $r2$. It is also assumed that the route withdrawal has the lowest priority among all types of routes advertised to the peer, that is $Pw < Pr$ for a route $r$.

The update messages exchanged are shown in table 2.1, these messages start with link failure which is between 0 and d. Therefore the node 1 begins experiencing the events of selecting one path after another. And then the priority of paths announced varies from low to high and again to
low as shown in figure 2.2. Hence the authors have successfully distinguished between path exploration and a flap.

### Table 2.1: Update Message Exchange Sequence

<table>
<thead>
<tr>
<th>Stage</th>
<th>Routing Table</th>
<th>New Messages</th>
<th>Preference</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>1(*0d, 30d, 56780d) 3(*0d, 10d, 40d) 4(*0d, 20d, 30d) 2(*0d, 40d) (Steady state)</td>
<td>0→ {1, 2, 3, 4, 8} W</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Link between 0 and d (0d) is down</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1(-, *30d, 56780d) 3(-, *10d, 40d) 4(-, *20d, 30d) 2(-, *40d) 1→ {x,3}[130d], 3→ {1,4}[310d], 4→ {2,3}[420d], 2→ {4} W</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1 (-, -, *56780d) 3 (-, -, *420d) 4 (-, -, *310d) 2 (-, -) 1→ {x} [156780d], 3→ {1,4}[3420d], 4→ {3} W</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1 (-, *3420d, 56780d) 3 (-, -, -) 4 (-, -, -) 2 (-, -) 1→ {x} [13420d], 3→ {1} W</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1 (-, -, *56780d) 3 (-, -, -) 4 (-, -, -) 2 (-, -) 1 → {x} [156780d]</td>
<td>Low</td>
<td></td>
</tr>
</tbody>
</table>

### 2.5 SENDER-SIDE LOOP DETECTION

Labovitz et al. [Labovitz et al. (2001)] have proposed the use of an approach called the Sender-Side Loop Detection (SSLD). It is a method to decrease convergence time when the link or node failure disrupts the reachability to the destination. Sender-Side Loop Detection (SSLD) mechanism is incorporated into BGP where the sender refrains from propagating an announcement to a BGP peer when that peer’s AS number is already in the AS path. This way when it detects the loop then it does not propagate that update, rather than conventional BGP
approach, where the receiver has to perform that role. Using this technique, convergence time become smaller, and the risk of generating a sequence of searching path events is reduced. As per the current existing strategy of having AS number of 4-bytes the SSLD is not compatible, it is unlikely to be actively deployed in the Internet [Huston et al. (2010)].

2.6 CONSISTENCY ASSERTIONS

Pei et al. [Pei et al. (2002)] suggested a method to improve BGP convergence by consistency assertion on both types of updates: the route announcement as well as route withdrawal. In this approach after every received update message, the RIB routes are tested for any conflict with the newly arrived route information, and the result is taken into consideration at the time of path selection. This approach has been shown to reduce the volume of superfluous update messages, and improve average convergence time. The limitation of this approach is additional load imposed, and this would have a negative impact on CPU and memory usage and update processing time.

Using a route’s path information, the authors developed two consistency assertions for path vector routing algorithms that are used to compare similar routes in order to identify whether the routes are infeasible or not. By doing so, infeasible routes are avoided and not installed in Loc-RIB. The authors claim to achieve a significant reduction in convergence time and at the same time total number of intermediate route changes is also minimized.

Simple Path Vector Protocol (SPVP): A node will advertise to its neighbours only one single path to each destination. The latest path received from each neighbour replaces the previous path sent by the same neighbour and is kept as a candidate for path selection. If this new path results in a route change, then the newly selected path is sent to neighbours. When a node loses all the paths to the destination, it sends an empty path to its neighbours to withdraw the path that it sent before.
Valid Path: A path \((N1; D)\) is valid if and only if its view of the path from each \(Pi\) to \(D\) is correct. The empty path, a path \((N1; D) = NULL\), is always valid since it only reflects that \(N1\) does not know a path to \(D\).

Path Consistency: Both \((N1; D)\) and \((N2; D)\) paths are consistent if one of the following
consistency conditions is satisfied:

a. Two empty paths are consistent.
b. Two non-empty paths with no common node (not counting the destination) between them are consistent.
c. Empty path \((N1; D)\) is consistent with non-empty path \((N2; D)\) if \(N1 = Qi\), for all \(i = 1\) to \(m\).
d. Two non-empty paths path \((N1;D)\) and path\((N2;D)\) intersect at nodes \(P = Q\) are consistent if path \(N1 (P;D) = path N2 (Q;D)\).

Infeasible Path: If the path \((N1; d)\) and path \((N2; d)\) are not consistent and \(N1 = Qi\) for some \(i=1\) to \(m\), path \((N2; D)\) is infeasible. An infeasible path is not necessarily an invalid path, but we require that an infeasible path not be selected as the best path to \(D\).

Limitations of this work are summarized as it requires extra computation resources from the router to compute the consistency check, and to send extra information in the BGP messages.

2.7 AN EXPERIMENTAL ANALYSIS OF THE BGP CONVERGENCE TIME

In this work [Griffin et al. (2001)], the convergence time of BGP can be reduced by determining the impact of

- MRAI (Minimum Route Advertisement Interval)
- SSLD (Sender Side Loop Detection)
- WRATE (Withdrawal Rate Limiting)
The SSLD is the server side loop detection, it checks for the loops that are created in any AS path. It can be checked by simply avoiding the AS number from where the path is being started. The WRATE is the withdrawal rate limiting. It is the simple MRAI timer.

Many observations have been performed to reduce this delay in BGP.

Observation 1: For each network topology and each kind of experiment (UP or DOWN), there is an optimal value for MRAI, beyond which the average total number of updates required for convergence is stable.

Observation 2: For each network topology and each kind of experiment (UP or DOWN), there is an optimal value for MRAI, where average convergence time is minimized, average convergence time increases linearly.

Observation 3: The optimal MRAI value $Mt$ increases with the average router workload while $U_{min}$ remains stable.

Observation 4: An optimal value for MRAI can dramatically decrease the convergence time. However, this optimal value varies from network to network, and may be difficult to approximate in practice.

Observation 5: In terms of convergence time, WRATE can result in either a gain or a loss, depending on the network and the experiment type (UP or DOWN). If MRAI is close to $Mt$, then WRATE has little effect.

Observation 6: SSLD never increases convergence time, and may decrease it by a small amount. Limitations of this work: This approach is only useful for networks which are connected to one or two networks. If the network is connected to many networks then this approach is not effective.
2.8 IMPROVED BGP CONVERGENCE VIA GHOST FLUSHING

The authors Bremler-Barr et al. in their work [Bremler-Barr et al. (2003)] suggested a minor modification to the BGP that eliminates the problem pointed out and substantially reduces the convergence time and communication complexity of the BGP. The reduction in convergence time is done by the messages which are acknowledging failure should propagate fast while the messages which acknowledging new routes should propagate as they do in the BGP, after one Minimum Route Advertisement Interval (MRAI).

In this work two rules were introduced for the withdrawal messages and for the announcement messages. These rules are:

i. Ghost Flushing Rule

ii. Ghost Bursting Rule

Ghost Flushing Rule: In this rule, a router sends a withdrawal to its neighbours as soon as it learns (with no delay) that the last AS path it has announced for that route has been changed and became longer or not valid. This solution reduces the convergence latency to $d \cdot h$ and $d$ is the longest AS path a router has in the network to the destination and $h$ is the average delay between two neighbouring BGP routers.

Ghost Buster Rule: As per this rule not only the withdrawals are propagated as fast as possible, but it makes sure that announcements are guaranteed to be delayed. In the original BGP algorithm, in most cases, announcements are delayed due to the Minimum Route Advertisement Timer, especially when the Minimum Route Advertisement Interval is implemented per peer and not per destination (i.e., for each announcement sent on the corresponding interface, and not for each announcement that corresponds to the same destination). The Ghost Buster Rule requires that any new announcement be delayed, even an announcement regarding a more preferable AS path.
Limitations of this work:

- The ghost flushing rule is only useful when the AS path is not having any alternate long path then the previous one, not for shorter paths.

- The ghost bursting rule is not useful in case of shorter paths because they are also delayed by one MRAI.

2.9 DIFFERENTIATED UPDATE PROCESSING

Sun et al. [Sun et al. (2006)] introduced the technique which they named as Differentiated Update Processing (DUP), as per their claims this DUP succeeds in reducing almost 30 percent BGP updates and also improves convergence time by nearly 80%. This technique works by dividing updates in different classes and then it treats each in a different manner. For sending announcements it checks for the class a particular announcement falls in, depending on the class they are in, it may be delayed for a specific duration. The classification of updates depends on the novelty of an update and on the traversed topology. The idea can be incrementally deployed. This technique does not require any changes to the original BGP protocol. The authors also admit that limitation of their approach is that it increases overhead in terms of CPU and memory usage.

2.10 LIMITING PATH EXPLORATION IN BGP

The AS path attribute is insufficient to correctly distinguish invalid routes from those that are valid. Since a router does not makes any distinction between AS paths it exports to different neighbours. That is, the outgoing or forwarding edge is not embedded in the announced AS path. The forward edge can captures the missing information that will enable a router to identify if paths suffer a failure. The use of Forward Edge Sequence Numbers (FESN) to capture the state of a forwarding edge [Chandrashekar et al. (2005)]. There are two different types of forward edge sequence numbers used here the major, and the minor. The first number, which is major,
for a pair of adjacent ASes and is shared across all the minor edges between them. The second one, which is minor, is used to differentiate between routes learned over different minor edges of the same neighbour AS.

At any AS, say X, corresponding to each neighbour, say Y, associate an edge sequence number of class major, which may have different value to each of its neighbours. The notation (X:Y,n) is used to describe the major edge sequence number along with link X - Y. The value of n is incremented when link X-Y is restored after a failure. Importantly, the value of n is not incremented when the edge fails. The (X:Y,n) is managed by AS X, i.e. AS X is responsible for incrementing the value of edge sequence number n.

When X sends a route announcement to its neighbour Y, it attaches number n to the route X-Y, and it look like (X:Y,n) which is the major FESN to the route X-Y. Similarly every router performs the same while announcing its path to the neighbour along the way and consequently, a path contains an ordered list of major FESNs, called the FESNList of the path, as shown in table 2.2.

Table 2.2: FESN List at Router in AS 5 of Figure 2.4.

<table>
<thead>
<tr>
<th>AS PATH</th>
<th>EDGE SEQ NO. ( FESN ) LIST</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 2 1 0</td>
<td>{(3:5,1) (2:3,1) (1:2,2) (0:1,3)}</td>
</tr>
<tr>
<td>4 2 1 0</td>
<td>{(4:5,1) (2:4,2) (1:2,2) (0:1,3)}</td>
</tr>
<tr>
<td>7 6 1 0</td>
<td>{(7:5,2) (6:7,1) (1:6,1) (0:1,3)}</td>
</tr>
</tbody>
</table>

The path 3 2 1 0 appeared in table 2.2 is through the routers which now understand the link status well with the help of edge sequence number. The link between the AS 0 and AS 1 has experienced three failures while as the link connecting AS 3 to AS 5 has experienced failure only once. The table 2.2 has been obtained from the network topology of 8 ASes, which is shown in figure 2.4.
In figure 2.4 the AS 0 connected with AS 1 with the link which has sequence number 3, the link originating at AS 1 connecting AS 3 has sequence number 2, whereas the same AS 1 connected with AS 6 with different sequence number which is 1, this way stability is considered based on the link rather than the AS itself. Similar case can be seen with the AS 2 connected with AS 3 and AS 4. This way the AS paths carried in the route updates are identical, but the corresponding FESNList's are different. More generally, the FESNLists sent to different neighbors are different because of different FESNList. Another scene is of the connectivity between AS 6 and AS 7 is different from other ASes in the network. There are multiple minor edges between neighboring ASes, they are all associated with the same major FESN. To distinguish between routes learned from different routers in the same AS neighbor, a minor FESN, specific to each router level peering session, is used.

When a link failure event occurs, for instance link between AS 1 and AS 2 fails, an attached router AS 2 generates a route withdrawal to AS 3 and AS 4, it will insert the FESNList of the invalid route into the withdrawal. The neighbor AS router when receives this withdrawal message will invalidate those routes matching with the fensList, and then forwards the update to
the next neighbor AS router if required. When a neighbor receives this withdrawal and generates a subsequent routing update, it attaches the original withdrawal.

In case of the link failed between AS 1 and AS 2 is repaired and available for use, the attached AS 1 will increment the FESN of the link from (1:2, 2) to (1:2,3) and announce this to the neighbour router 2.1 of AS 2, the router 2.1 after receiving the connected route, will install it in its routing table and sends the update to upstream routers.

### 2.11 A TECHNIQUE FOR REDUCING BGP UPDATE ANNOUNCEMENTS THROUGH PATH EXPLORATION DAMPING

The Authors [Huston et al. (2010)] have defined and evaluated Path Exploration Damping (PED). PED is a router-level mechanism for reducing the volume of likely transient update messages within a BGP network. It also decreases the average time to restore reachability compared to current BGP update damping practices. PED delays and suppresses the propagation of BGP updates. It either increases the length of an existing AS Path or varies an existing AS path without shortening its length. Huston et al. have also described how PED can be incrementally deployed in the internet, as it interacts well with Minimum Route Advertisement Interval (MRAI) deployment. It also enables restoration of reachability quickly.

The authors have classified the updates into seven different categories. These categories are mentioned in table 2.3.

<table>
<thead>
<tr>
<th>S.No.</th>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>AA+</td>
<td>Update contains path announcement for previously announced prefix with longer path length.</td>
</tr>
<tr>
<td>2</td>
<td>AA-</td>
<td>Update contains path announcement for previously announced prefix with smaller path length.</td>
</tr>
<tr>
<td>3</td>
<td>AA0</td>
<td>Update contains path announcement for previously announced prefix with the same path length but the</td>
</tr>
</tbody>
</table>
This approach delays those updates that contain either the longer path or the same length path than previously announced path for a particular destination prefix, these updates are delayed for a period called as PED interval.

Figure 2.5: Effect of single withdrawal on path exploration [Huston et al. (2010)]
Figure 2.5 shows the effect of a withdrawal on the other routers in neighboring ASes, when the link connecting AS 1 with AS 2 fails. AS 2 detects the failure and sends a withdrawal W to AS 3, and AS 5. The AS 5 thinks that AS 1 is still reachable through other paths i.e. through AS 3, and it switches its path to path 5, 3, 2, 1, but soon after it receives withdrawal from AS 3 for the same, then it (AS 5) again switches its previously selected path 5, 3, 2 1, to the path 5, 4, 3, 2, 1 and then again after receiving withdrawal from AS 4, it finally withdraws its reachability to AS 1 and sends withdrawal to further upstream routers in ASes. During this time after the link failure event, AS 5 has announced a rapid sequence of longer paths one after another to the upstream routers. Based on the classification mentioned in table 2.3, the announce update can be put in this order “ NA, AA+, AA+, AW ” with unnecessary sequence of “ AA+, AA+ ”, which indicates the event of path exploration, and the network converges only after the withdrawal i.e. AW. The Algorithm for PED, which is on the basis of per-peer, and per-prefix, follows:

Table: 2.4 Algorithm for Path Exploration

| Step 1: | Take a new timer PEDI |
| Step 2: | Take a new Queue Q [ this Q is to hold updates ] |
| When Router is Ready to send an update for the prefix P |
| Step 3: | If Sending an update, Then |
| Step 4: | If update category is AA+ || AA|| AA*|| AA0 |
| Then |
| 4.1 If timer PEDI not active |
| Then |
4.1.1 Q←Update
4.1.2 PEDI Start

4.2 Else

4.2.1 Delete any previously queued update for the same prefix

4.2.2 Q←new update

4.2.3 PEDI Restart

Step 5: Else If update category is AW || AA-

Then

5.1 Delete previously queued update for this prefix

5.2 Send the new update immediately

Step 6: Else

6.1 Send immediately (it means update is now NA)

Step 7: Stop

2.12 BGP CONVERGENCE: OPTIMALITY VS. REACHABILITY

T. Li and G Huston in their work [Li and Huston (2007)] had divided the convergence in routing system into two parts. First part is reachability and the second part is optimality.

The optimality ensures that each BGP router knows the best path to reach the destination prefix. While the reachability ensures that there is always a possible path to reach the destination, that possible may be considered suboptimal. Though the reachability and optimality are sometimes accomplished at the same time, but there are situations where the reachability and the optimality are accomplished at different times, the reachability is achieved much before the optimality.
Figure 2.6 (a): Converged Network for Prefix 1.0.0.0/8 [Li and Huston (2007)]

Figure 2.6 (b): Network Maintaining Reachability After Link Failure for Prefix 1.0.0.0/8 [Li and Huston (2007)]
In the duration of path exploration event, reachability might be achieved many times, while the optimality might not be achieved at all. The similar situation are observed in the figure 2.6 (a), (b), and (c).

Feng Wang et al. [Wang et al. (2009)] have captured the transient behavior of the Interdomain routing protocol with the help of a formal BGP model. They derived sufficient conditions for the occurrence of transient routing failures. They had also analyzed transient routing failures in typical BGP systems. Here, commonly used routing policies were applied. Their analysis to improve their network performance and stability could be applied by Network administrators.

Wei Zhang et al. [Zhang et al. (2010)] have introduced a new Interdomain routing scheme based on a centralized routing service called the Global Path Service (GPS). Their new scheme provides alternate inter-AS paths different from ordinary BGP routes, facilitates diversified inter-AS forwarding paths through GPS-enabled ASs, offers their Interdomain transit tunnels (Multi-Protocol Label Switching), and facilitates diversified source routing at the AS-level. These authors have also illustrated a framework to implement the infrastructure of diversified
routing. The infrastructure allows for novel Interdomain routing services, such as Interdomain Quality of Service (QoS) or multipath routing.

K. Krishna Chaitanya and Ch. Ravi Kishore [Krishna et al. (2010)] have discussed an approach to the BGP for shortest path and lowest cost routing based on the Open Shortest Path First (OSPF) technique. This approach is useful in mobile Ad-hoc networks for efficient power management. There may be an increase in the BGP routing table size but the cost of transferring data to the destination is decreased.

Yi Wang et al. [Wang et al. (2009)] have applied export filtering policies in their proposed work to address correctness problems. To provide customizable route selection, they have proposed Neighbor-Specific BGP before the route selection process. Neighbor-specific BGP selects routes on a per-neighbor basis. Existing mechanisms such as Virtual Routing and Forwarding (VRF), encapsulation, BGP add-paths, etc. can be used by a single ISP to deploy both solutions incrementally, requiring only changes in software to its routers.

Kanyapat Watcharasitthiwat and Paramote Wardkein [Kanyapat and Wardkein (2009)] have provided an algorithm to achieve economic and reliable network design in relatively small computational time. The proposed algorithm is based on improved ant colony optimization by introducing two techniques to improve neighborhood search and re-initialization process.

Yunsheng Liu and Zheng Wang [Yunsheng and Zheng (2012)] have developed a new routing protocol that exploits a local flooding mechanism to contact multiple neighboring nodes. Their proposed protocol alleviates the flooding overhead in the gradient setup phase with the help of a back-off waiting scheme. The proposed protocol uses a multiple route selection mechanism to forward packets, in case of rapid topological changes; a compact gradient reconfiguration is performed.

Rodrigo Pantoni and Brandao Dennis [Rodrigo and Brandao (2011)] have presented a new network routing mechanism with acknowledgment to provide a new means to achieve high delivery rates in urban networks, with tolerated end-to-end delay values for a street lighting network. The mechanism consists of selecting nodes from neighbors other than those previously selected to find a path to the destination.
Wang Ling et al. [Wang et al. (2009)] have presented a routing algorithm that combines the shortest path and adaptive routing schemes for Networks on Chips. Their scheme stores routing information in a series of routing tables created at the routers along the routing path from the source to the destination. To minimize space and timing cost, the routing table for each node is created off-line and updated on-line to reflect dynamic changes of network status to avoid network congestion.

Dana Arash et al. [Dana et al. (2008)] have proposed a protocol that uses multiple backup paths to eliminate recurrent path failures affecting the quality of service in ad-hoc networks. The proposed protocol also converges into a reliable path set with no message overhead.

### 2.13 AGGREGATING PREFIXES

The Aggregation is the process of combining one or more address prefixes into a single address prefix, the resultant prefix is called aggregated prefix. The aggregation is used to reduce numbers of route entries in routing table of a router.

Rekhtar Y. and Li T [Rekhtar and Li (1995)] explained the routing table structure of BGP and also different blocks of the routing tables which explain how updates received in AdjRib-In are transformed into FIB, and to AdjRib-Out.

Uzami et al [Uzami et al. (2010)] mentioned in their technical report on aggregation that as the information comes from different sources which are having different attributes aggregation process causes loss of information. As a sign for the loss of information a well known attribute Atomic_Aggregation which is discretionary attribute may be observed. So when any router sends an aggregated entry which is causing information loss, it is necessary for that router to attach attribute Atomic_Aggregation along with it.

Sam Halabi [Halabi (2000)] has pointed out that in route aggregation a range of network prefixes are summarized into few aggregated network prefixes by varying the CIDR of the aggregated network prefixes which is relatively smaller and efforts are to keep it as smaller as possible but this approach has a severe limitation of losing the granularity of more specific routes which exists prior to aggregation.
When two or more routes are aggregated into one route and this aggregated route is then sent to peers as a single advertisement then the path information that was there in many routes is lost. This has potential to lead to routing loops.

Draves R et al. [Draves and J. (1999)] proposed the optimal routing table constructor for route aggregation and emphasized that it may be a good scheme to limit routing table’s growth if applied with proper care but it must not be forgotten that it may be hazardous and may lead to loops in routing or black holes in routing if it is applied carelessly which has potential to bring the whole network down. The situation of black hole occurs when unintended destination for a particular traffic receives that traffic and the receiving router cannot forward it further. These routing challenges are more evident when multiple address allocation patterns are observed and also when interaction with aggregation is learnt.

Ahmed Elmokashfi et al [Elmokashfi and K (2008)] the growth of the routing table size and the rate at which BGP updates are growing are two separate issues of BGP scalability. The number of networks which either may fail or trigger a change in the route increases the size of the routing table along with the increase in the number of updates.

Vogt, C. [Vogt (2008), Jen and M (2008)] have given a routing design to address the issue of scalability in the network of Internet edge. The problem addressed by making use of provider independent edge addresses inside edge network. Transit addresses allocated by provider do not move to edge networks as all inbound packets are translated into edge addresses.

Yaoqing Liu et al [Liu and Z (2010)] described the weak and strong behaviors of forwarding correctness on the basis of prefixes being forwarded after aggregation which are not forwarded if aggregation is not done. Side effects of aggregation in the form of extra routable space have also been highlighted.

Francis P. [Francis (2011)] in his proposed work FIB suppression does on demand aggregation of routing information base to forwarding information base. This approach is to give only a temporary solution. It may be used for aggregation when number of updates received are limited to only one but it also degrades the state of the aggregated forwarding table. It uses virtual aggregation to aggregate RIB into FIB by partitioning the address space into large
prefixes which are named as virtual prefixes and those virtual prefixes need not be of the same size. Virtual aggregation makes virtual prefix aggregation with help of the tunnel which is created from the point of aggregation router to the next hop router. The Author suggested the tunneling to be done in the first case when the ingress router has no prefix for aggregation then it is created up to the point of aggregation, and second case when an ingress router has prefixes for aggregation then the tunnel is created up to egress router [Jen and M (2008)].

Challal et al [Challal et al. (2011)] have considered node failures and network intrusions as unavoidable events and presented fault tolerant routing scheme with the help of simulations done on a small operating system including the capability to protect itself from intrusions to make their scheme reliable.

Mary Wu and Kim ChongGun [Mary and ChongGun, (2010)] have given a routing method based on the minimum-cost matrix and the next-node matrices calculated from the adjacency-cost matrix, to express the link costs of the network. Their method uses link costs instead of hop counts as a routing metric to make routing efficient.

To address to the problem of slow convergence in BGP several efforts have been made in the literature and the majority of those efforts emphasizes on keeping the BGP design unmodified and making temporary changes to the protocol.

2.14 CONVERGENCE TIME

The network when become unstable it takes time to come back into stable state again. The period, for which the routers in the network are not able to get complete updated knowledge of the changed scenario, is called convergence time.

C Labovitz et al. [Labovitz (2000)] have proposed a method to improve the BGP’s convergence by including additional synchronization, and diffusing updates in the protocol but all the proposed changes increases complexity and also increases router overhead.

Anat Bremler-Barr et al. [Anat Bremler-Barr et al. (2003)] have attributed a piece of information that is outdated but still keeps floating in the network as ghost information.
Because of the existence of the ghost information, nodes start relying on the wrong information and therefore AS paths become longer until a loop is detected. They have suggested that the improvement of convergence time is related to the early removal of the ghost information.

Dan Pei et al. [Pei et al. (2005)] have proposed a method of improving the BGP’s convergence time by finding the location of the fault and isolating all those routers from the forwarding paths which are in the fault zone and then piggyback information about the main cause on each routing update message caused by the main cause. The authors have made a provision by which the fault location is identified and then the router/routers which are connected to the fault advertised as root cause notification so that every router can avoid using paths that those routers exist.

Amit Sahoo et al. [Sahoo et al. (2006)] have proposed a method for processing updates in groups that they called batch update processing and the number of updates exchanged is reduced. The authors claimed to effectively maintain a separate logical queue for each destination. When an update arrives, the destination address is extracted from the update, and it is queued appropriately. As a result queuing of all the updates to a destination are processed together.

S. Deshpande et al. [Deshpande and Sikdar (2004)] in their work on the route processing and MRAI timers have proposed an approach to minimize convergence time by identifying redundant timers and then cancelling timers for those peers for which routes are no longer in use. They have also used adaptive MRAI with the help of new variables for a counter and a threshold.

Abdelshakour Abuzneid et al. [Abuzneid and Stark (2010)] in their work on improving convergence time using MRAI timer have shown that the withdrawal rate limiting in the BGP may help in detecting routing loops but its role in reducing convergence time is not significant but the MRAI timer is effective for the same purpose.

Kun-Ming Yu et al. [Kun-Ming et al. (2011)] have improved the response time of their ad-hoc protocol by introducing a concept of backup routes along with the main route, the authors have made a provision for maintaining multiple backup routes which may be used in case main route failure. Their modified protocol then constructs the main route while the data
transmission which was broken earlier may be resumed immediately. The information of the backup routes is stored in a backup routing table while the information of the main routes is stored in the main routing table.

Tobias Heer et al. [Heer et al. (2011)] have addressed the issue of security in their work that emphasizes on resource constrained networks. The authors have emphasized on the characteristics of the protocol that complicate the designing issues while managing vital resources like communication channel bandwidth, the CPU power etc.

### 2.15 ROUTING POLICIES

The routers in the Internet can be configured to handle a particular traffic flow in a different manner, these may be seen as the tools to circumvent the normal routing behavior, this configuration is called as routing policies.

Bates T., Chen E., and Chandra R. [Bates et al. (2006)] have given emphasis on anomalies caused by dependency among interior BGP and IGP. Large networks’ implications are addressed using the route reflection approach, but this approach reduces the visibility of EBGP routes which subsequently affects the performance of the forwarding process and user traffic may experience suboptimal services. All nodes may not know the external BGP routes received by egress points in the network. The criterion for the selection of the best route to any given destination is the distance of interior gateway protocol to the egress router.

Peter W. Thai, and Jaudelice C. De Oliveira [Thai and Oliveira (2012)] has proposed software defined networking based approach to control Interdomain policies. The authors have given methods of integrating Autonomous Systems, which consists of policy language, policy controller, and Autonomous System bridge. Policy Language was used for coding configurations related to common relationships between Autonomous Systems. Policy controller is used to forward packets to destinations that are outside Autonomous Systems and authenticated as well. Autonomous System Bridge is a piece of code lies in an external Autonomous System which interface with software defined networking stack of the external router.
Hongsuda Tangmunarunkit, Ramesh Govindan, Scott Shenker, and Deborah Estrin [Tangmunarunkit et al. (2001)] have analyzed the impact of policy routing. In the absence of the knowledge of the inter-ISP relationships, they have used a model in that the policy path is determined by the shortest path. They have shown in their results that the use of policy worsens the routing and inflate the path length in the network. Their findings discourage the use of policy in internet.

Chia-Wei Chang, Han Liu Guanyao Huang, Bill Lin, and Chen-Nee Chuah [Chang et al. (2012)] have used alternative paths in the network to take advantage of multiple equal cost path routing. Their approach gives advantage in traffic measurement by dividing the traffic among the multiple paths. They have used game theory based rerouting policy for exchanging dialogue and decision making among routers. The authors have designed a cost function for links in such a way that better links attract traffic towards them and routers compete to minimize their cost of the paths.

After going through various research papers, documents, RFCs etc. in this section, findings may be briefly summarized as follows:

- Path failure has negative impact on the BGP, which is further aggravated if the new path selected is appropriate.
- The Internet growth has contributed to the huge routing information, which is cause of problem when stored on cards.
- The longer convergence time put the network vulnerable for longer duration which needs to be minimized.
- Routing policies may play very important role by managing the growing size of the internal network of an AS.