Speculative Route Invalidation to Improve BGP Convergence Delay under Large-Scale Failures

Amit Sahoo
Dept. of Computer Science
Univ. of Califonia, Davis
Davis, CA 95616
Email:asahoo@ucdavis.edu

Krishna Kant
Intel Corporation
Hillsboro,OR 97124
Email: krishna.kant@intel.com

Prasant Mohapatra
Dept. of Computer Science
Univ. of Califonia, Davis
Davis, CA 95616
Email:pmohapatra@ucdavis.edu

***Abstract*— The Border Gateway Protocol (BGP) has been known to suffer from large convergence delays after failures. We have also found that the impact of a failure rises sharply with the size of the failure. In this paper we present and evaluate a speculative route invalidation scheme aimed at reducing the convergence delays for large-scale failures. Our scheme collects statistics from BGP updates received at a router and identifies Autonomous Systems (ASes) that are likely to be "unstable". Routes that contain these ASes are marked as invalid and not propagated further. This cuts down the number of invalid routes during the convergence process and results in a significant improvement in the convergence delay.**

I. Introduction

The Internet is composed of a large number of independently administered networks (also known as autonomous systems). Packets are routed between different autonomous systems (ASes) by the Border Gateway Protocol [1], [2]. BGP has been used as the primary inter-domain routing protocol in the Internet over the last ten years or so and the excellent scalability of BGP has been a major facilitating factor for the explosive growth of the Internet. However BGP does have its share of shortcomings and the most important one is the long convergence delay (or recovery time) after failures in the network. The failure of network elements can cause BGP to suffer from extended periods of instability during which numerous BGP routers modify their routing tables.

The long convergence delay is characteristic of the routing algorithm (*path-vector*) that is used by BGP. A node, that implements a path vector scheme, sends the best route for each destination to each of its neighbors. A node also stores all the routes that are sent by its neighbors and selects the best route for each destination according to some policy. If the node receives a message saying that the best route for a destination has failed, then it switches to the next best route for the destination and the advertises this route to the neighbors. The backup route is chosen based on the assumption that it is still valid. However BGP doesn't maintain up to date information about the validity of the routes and in case the backup route is later found out to be invalid, the process has to be repeated all over again. This can result in a considerable delay before the cycle of withdrawals/advertisements ends and all BGP nodes have a valid and stable path to each destination.

Numerous studies [3], [4], [5], [6], [7], [8], [9] have been carried out to study the fault tolerance and recovery characteristics of BGP. In particular, it was shown by Labovitz et al. [4] that the convergence delay for isolated route withdrawals can be > 3 min in 30% of the cases. Zhao et al. [9] showed that packet loss rate can increase by 30x and packet delay by 4x during recovery. Furthermore, we have shown in our previous work [10] that multiple simultaneous failures can cause the convergence delay to increase significantly. Though large-scale failures might be rare, they have the potential to cause widespread disruptions to e-commerce as well as critical services such as emergency response, banking, infrastructure monitoring etc. Hence the importance of a short convergence delay after failures (especially large ones) cannot be overstated.

In this paper we present a scheme aimed at decreasing the convergence delay of large-scale failures by reducing the number of invalid routes explored during the convergence process. We do this by collecting statistics about withdrawn/replaced routes and by using this information to estimate whether a route is valid or not. A route that is suspected to be invalid is "avoided" during the convergence process and this suppresses the propagation of invalid routes. Our scheme is designed to run independently at each BGP router/AS. Hence a high degree of coordination between various ASes is not required during deployment and performance gains can be achieved even if the scheme is not implemented at some of the ASes. Our scheme is also compatible with standard BGP implementations because we do not add any new messages. Our experiments show that we can reduce the convergence delay substantially for large-scale failures without causing any significant deterioration for other scenarios.

In this section we have talked about BGP and the need for reducing the convergence delay incurred by BGP after a failure. In the next section, we talk about schemes that try to identify the cause of BGP route updates and whether these schemes can be used to detect invalid routes during the convergence process. We present and discuss our speculative invalidation scheme in Section III. In Section IV we describe our experimental setup followed by the results and analysis in Section V. We wrap up with the conclusion in Section VI, and the references.

II. Related Work

There have been a few proposals that attempt to identify topological changes using BGP updates. Chang et al. [11] collected BGP updates from multiple vantage points and used a two level clustering scheme to group the updates into events. The first clustering scheme groups updates that were received at the same observation point, refer to the same prefix and are closely spaced in time. The authors then identify the *candidate peerings*(links) that might be responsible for the string of updates in the *prefix-based* cluster. The second clustering scheme takes the prefix-based clusters from multiple vantage points and groups those that are closely spaced in time, are of the same *type* and have some common candidate peerings. The authors used the prefix-based clusters and the second level *peering based* clusters to study various characteristics of BGP events, such as the duration of the instability, the number of prefixes affected, etc.

Caesar et al. [12] presented a scheme to identify the cause of "major" BGP events. Whenever a BGP route for a prefix is changed, suspect elements (ASes or links) and the corresponding event at that element are identified. All elements that could be responsible for the route change are included. An AS or link is considered to be of either major or minor significance and marked accordingly; based on the number of prefixes for which the route might have changed because of an event at this element. After all the elements have been marked, a set of rules are used to estimate the location. This procedure is carried out independently for each type of event. Finally for each event *type*, the intersection of the results from various views can be taken to narrow down the set of responsible network elements.

Feldmann et al. [13] and Lad et al. [14] also proposed schemes similar to the ones above. The objective of all these schemes is to identify the cause of the BGP updates. Once the cause has been identified, we can just avoid the invalid routes and thus the BGP convergence delays can be reduced greatly. However, none of these schemes are suitable for real time application because of the high processing overhead associated with them. As BGP updates keep coming in, the algorithms would have to be run repeatedly and that is not feasible in real time. Furthermore, all the schemes except the one proposed by Caesar et al. require data to be collated from multiple views; and that complicates things further. Our scheme employs a approach similar to the ones above, but it has a much lower overhead and operates independently at each BGP router.

III. Speculative Invalidation Scheme

As we mentioned earlier, the primary cause of long convergence delay is the absence of information in BGP about the validity of routes. Our goal is to expedite the BGP convergence process for large-scale failures by identifying invalid routes and reducing their effect on the convergence process. We identified the following constraints that must be satisfied by any candidate scheme:

- All the analysis has to be done independently at a BGP router. Communication between routers will lead to greater complexity as well as result in delays while inputs are collected from various sources. Furthermore some messages are liable to be lost during periods of instability, which will cause additional delays. The performance of a distributed scheme in a particular area would be heavily dependent on the number of ASes that implement it, and hence deployment would have to be coordinated between multiple ASes/organizations. Communication between routers/ASes might also affect the compatibility with routers/ASes running standard BGP.
- The scheme should have a low processing overhead. Routers are becoming more and more powerful; however it is advisable to keep the overhead low so that the timely processing of BGP updates in not affected.
- The convergence delay for situations other than large-scale failures should not be increased.

Our scheme leverages the fact that large-scale failures generate a large number of updates that can be used to identify the failed/affected ASes. Our scheme is also based on the assumption that failed/affected ASes will be present in a large number of the generated updates. Later on in the section, we discuss the possible consequences of the second assumption.

We maintain a *failCount* for each AS at each BGP router. This value is incremented when a route containing that AS is withdrawn or replaced. It is decremented when we receive a new route containing that AS. When an AS suffers total or partial failure, a large number of routes containing that AS are likely to be withdrawn. Hence, we consider the *failCount* to be a measure of the likelihood that an AS has failed. We then select the ASes that have the largest *failCount* and consider all routes that contain those ASes to be "invalid". However the overhead of going through all the stored routes and checking them to see if they are invalid, will be too high. Therefore, we install routes filters that will mark new route advertisements as invalid if they contain any of the "suspect" ASes. Another approach would have been to check each route every time we have to select the best route for a destination, but this would again lead to higher overhead.

We only want to consider the recent history for selecting the "suspect" ASes. Therefore we divide the time into slots and use a configurable parameter, *history*, to determine how many slots of data we want to consider. BGP updates are sent periodically, and the period is determined by the MRAI [1]. We select the length of the time slot to be equal to the MRAI (after adjusting for jitter [1]), so that we receive one cycle of update messages in each slot. A larger slot length will only increase the response time for our scheme. A smaller value will decrease the response time but there could be a lot of variation in the number of messages received in two contiguous slots and that might lead to instability.

The *failCount*s for all the ASes are stored in a two dimensional array *failCounts[numASes][history]*. At the beginning of each time slot, we add up the *failCount*s for the last *history* slots for each AS, and consider the ASes with the largest cumulative *failCount* to be "suspect" for the duration of the time slot. We use the *avgFailedPathLength* parameter to determine the number of ASes that will be "invalidated". *avgFailedPathLength* is the average path length of the routes replaced/withdrawn during the last *history* time slots and is measured by our scheme. We use this parameter to compute another value, *failThreshold*. Starting from the front (highest *failCountSums*) we identify the smallest set of ASes for which the sum of *failCountSums* is greater than the *failThreshold*, and consider all the ASes in this set to be "suspect". The pseudocode for our scheme is listed in Algorithm 1.

Algorithm 1 Invalidation Scheme

```
1: if (Number of destinations for which the routes has changed in the last history slots) $<$ largeFailureThresh then
2:     Stop
3: end if
4: sumFailCounts $\leftarrow 0$
5: for $i = 1$ to $numASes$ do
6:     failCountSums[i] $\leftarrow \sum_{j=1}^{history} failCounts[i][j]$
7: end for
8: sumFailCounts $\leftarrow \sum_{i=1}^{numASes} failCountSums[i]$
9: Sort failCountSums in descending order
10: failThreshold $\leftarrow$ sumFailCounts / avgFailedPathLength
11: currentSum $\leftarrow 0$
12: for $i = 0$ to $numASes$ do
13:     if currentSum $<$ failThreshold then
14:         Add a filter to deny routes containing the AS at position i
15:         currentSum $\leftarrow$ currentSum+ failCountSums[i]
16:     end if
17: end for
```

If we assume conservatively that each replaced/withdrawn route contains one failed/unstable AS; then after the replacement/withdrawal of a route we increment the *failCount* of one impaired AS, and the *failCounts* of (*avgFailedPathLength*-1) stable ASes on average. If the sum of the *failCounts* for all ASes at an observation point is *X*, then the sum of the *failCounts* of the failed/unstable ASes can be expected to be close to *X/avgFailedPathLength*. This idea is used by the our scheme to identify the "suspect" ASes.

It might be difficult to identify the failed/unstable ASes correctly if the amount of data available for analysis is small. Hence we execute the invalidation scheme only if the number of destinations for which the route has changed is greater than *largeFailureThresh*. We will be looking at the effect of this parameter in the results section.

The success of our invalidation scheme is dependent on the assumption that the failed ASes are present in the withdrawn replaced routes. There can be scenarios where this assumption does not hold true [15]. However, these scenarios are known to be rare [16], and its very unlikely that a large number of updates will be generated due to these events. Similarly, addition of new links or peerings, and policy changes lead to BGP updates that will increment the *count* for some ASes, but again the volume of updates generated by these events can be expected to be low.

Our scheme manages to keep both the storage and processing overhead low. The overhead of incrementing/decrementing the *counts* and of running the new routes through the filters can be expected to be much lower than the overall processing overhead for a BGP update. The overhead of sorting the *counts* and installing the filters will also have little effect because these activities are only carried out once every time slot. The storage overhead will only be about 100 KB, with a *history* of two, as the *failCount* for each timeslot can be safely stored in two bytes per AS. The storage requirements can be lowered if we only store the *failCount*s for the ASes that have appeared in the withdrawn routes. However in that case the time to access the *failCount*s will be increased as we would have to use dynamic storage.

Our scheme does have a few issues that we need to work around. A new route is marked invalid if it contains a "suspect" AS. We need a way to remove the "invalid" flag if the AS in question is no longer considered to be "suspect". For this purpose, we need to maintain a list of routes that have been invalidated because of a particular AS. If this AS is not "suspect" anymore, then we can retrieve the corresponding list and validate the routes. We believe that the overhead for this process can be restricted to a manageable level, as we only need to do this once every time slot. Another issue is the possibility that all the routes for a destination are marked invalid. In such a scenario, we store the destination in a list and check for the existence of a valid route at the beginning of the next time slot.

As the scheme runs independently at each router, the set of "suspect" ASes at two BGP routers in the same AS might be different, which in turn might cause them to select different routes for the same destination. In order to solve this problem, we consider the routes received from iBGP peers to be always valid. This means that all the routers in an AS will have the same set of candidate route for each destination, and hence will select the same "best" route as long as they use the same scheme to calculate the degree of preference of a route.

IV. Evaluation Methodology

We used a number of synthesized topologies for our studies. A modified version of BRITE [17] was used for topology generation and BGP simulations were carried out using SSFNet [18].

A. Topology Generation

BRITE can generate topologies with a configurable number of ASes and with multiple routers in each AS. BRITE supports a number of AS topology generation schemes such as Waxman [19], Albert-Barabasi [20], and GLP [21]. The Albert-Barabasi and GLP models try to generate a power-law degree distribution, however the results are generally not satisfactory if the number of nodes (ASes) is less than a thousand. We modified BRITE to allow more flexible degree distributions so that we do not have the aforementioned problem and it is possible to assess the impact of degree on recovery time in

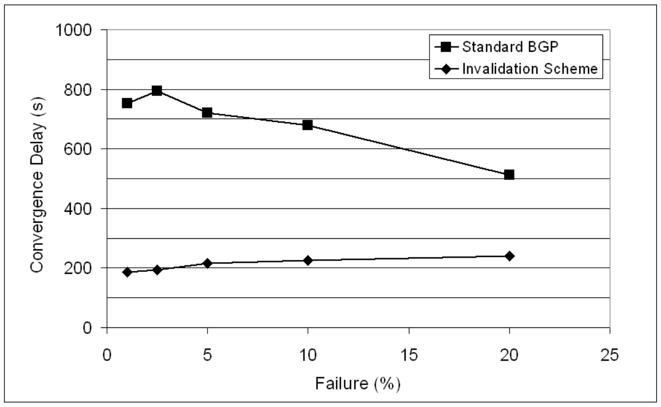

Fig. 1. Convergence delay with invalidation scheme

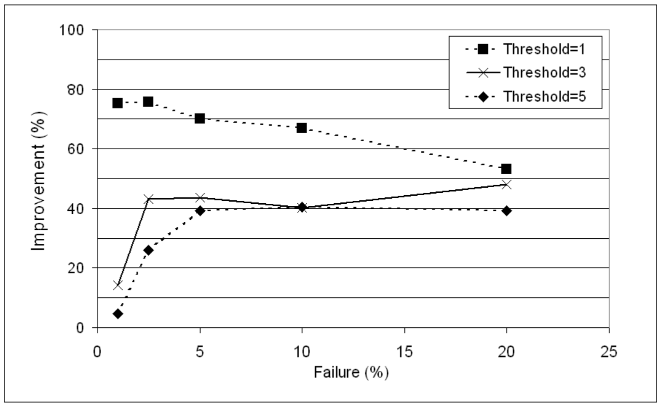

Fig. 2. Effect of *history* on convergence delay

more controlled settings (e.g., uniform degree, mixture of high and low degree, etc.). We also modified the code to generate variable number of routers for the ASes.

We used 200 node topologies for our experiments. This was dictated partly by the fact that the Java Virtual Machine could allocate a maximum of 1.5 GB of memory on the 32 bit machines that we used and hence we could simulate at most ∼250 nodes. For our experiments, we predominantly used topologies with a "realistic" degree distribution, derived from the actual degree distribution for inter-AS links [22]. The average measured inter-AS degree from the Internet AS-level topology is about 8.0 [22]. However, the Internet has over 22000 ASes and the maximum inter-AS degree is in the thousands. For our 200 AS network we decided to restrict the maximum degree to 50 and used the degree distribution in the range 1-50. This gave us a degree distribution which decays as a power law with an exponent of about -1.85. The average degree was very close to 4.

Although large-scale failures could be scattered throughout the network, many failure scenarios (e.g., those caused by natural and man-made disasters) are expected to be geographically concentrated. We randomly placed all the routers on a 1000x1000 grid and then considered failures in contiguous areas of the grid (usually the center of the grid to avoid edge effects). For all links, we used a one way delay of 25 ms (transmission, propagation and reception delays).

B. BGP Simulation

We used the SSFNet simulator for our experiments because it has been used extensively in the research community for large-scale BGP simulations and BRITE can export topologies in the format supported by SSFNet. In the simulations, the *path length* (i.e., number of hops along the route) was the only criterion used for selecting the routes and there were no policy based restrictions on route advertisements. All the timers were jittered as specified in RFC 4271 [1] resulting in a reduction of up to 25%. In our experiments the MRAI timer was applied on a per-peer basis rather than a per-destination basis, as is commonly done in the Internet. The MRAI timer was set to 30 seconds for eBGP peers and 0 for iBGP peers. We did simulate processing delays for BGP updates, but the delays were were much lower than the eBGP MRAI and hence can be expected to have little effect on the recovery time.

We failed a set of routers for our simulations and measured parameters such as the recovery time, the connectivity and the average path length after convergence is complete, and the loss in connectivity during the convergence process. As stated earlier, we assumed a contiguous failure area for large-scale failures. We further assumed that all routers and links in the failed area become inoperative. The scenarios where only the links (but not the routers) fail are unlikely for large-scale failures and are not considered here.

V. Results

In this section we present and analyze the results of our experiments. We used 200 node/AS (1 router per AS) topologies with "realistic degree" distributions (average degree 4) and a *history* of one time slot for our experiments, unless stated otherwise. We found that we were able to get good results with our scheme implemented at "high" degree nodes only. Thus, we can achieve improved convergence delays by installing this scheme at just a small subset of nodes/ASes. By default, the results that we present here were obtained with the scheme implemented at the nodes/ASes with a degree of greater than 4. We study the effect of the partial deployment of our scheme later on in this section.

We first compare the convergence delay for standard BGP and our invalidation scheme in Fig. 1. We set *largeFailureThresh* to 1 for these experiments. We plot the magnitude of the failure (in terms of the fraction of routers failed) on the x-axis and the convergence delay (in seconds) on the y-axis. We can see that there is a huge improvement in the convergence delay for each failure magnitude when we use the invalidation scheme.

We have plotted the improvement in the convergence delay over the base case (in %) for different values of *largeFailureThresh* in Fig. 2. We see that the performance is improved as *largeFailureThresh* is decreased. A lower *largeFailureThresh* means that our algorithm will be executed more often and more ASes will be marked as "suspect", which in

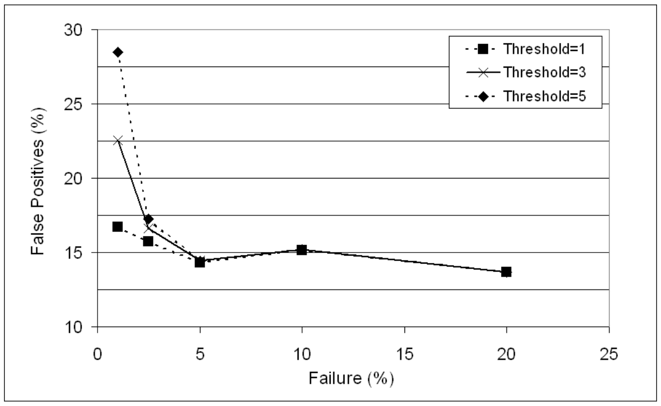

Fig. 3. Number of valid routes marked as invalid

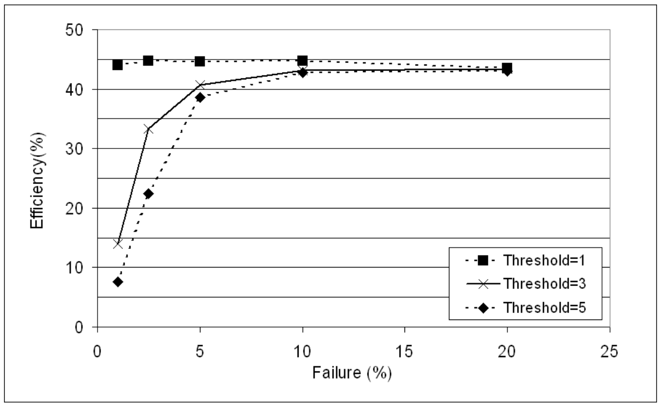

Fig. 4. Number of correctly identified invalid routes

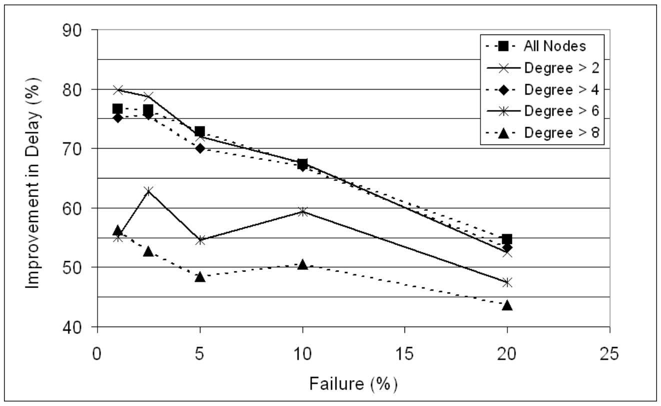

Fig. 5. Effect of deploying the invalidation scheme on a subset of nodes

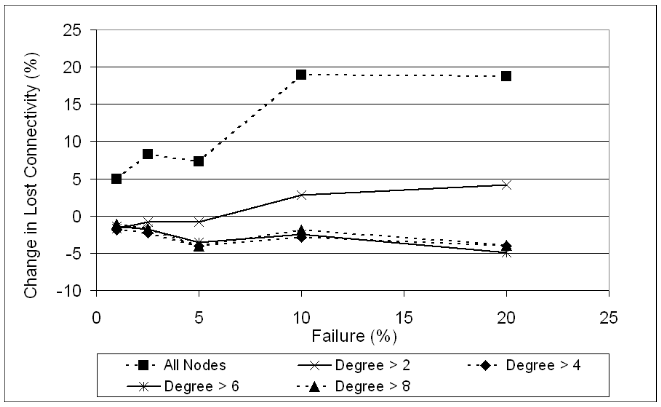

Fig. 6. Effect of *highDegreeThreshold* on "Lost Connectivity"

turn causes a higher fraction of invalid routes to be removed. This leads to a quicker convergence process. For large failures, routes are changed for a large number of destinations, and hence the difference in the number of invocations of our scheme (for different values of *largeFailureThresh*) is decreased. Thus the convergence delays start to converge as the size of the failures is increased.

There is no guarantee that our scheme will identify the failed/unstable ASes correctly. In case that we mark a stable AS as "suspect", we will deny any routes that contain that AS (for that time slot). In order to study the correctness of the decisions made by our scheme, we kept track of three different quantities during the convergence process: the number of routes that we deny, the number of routes which are actually valid but are denied by our scheme, and the total number of invalid routes. We plot the fraction of false positive identifications (error rate) made by our scheme in Fig. 3. For the lowest possible value of *largeFailureThresh* (=1), we see that the error rate declines slowly. This happens because the amount of data available for analysis increases with the size of the failure and thus the validity of the deductions is improved. As we increase the *largeFailureThresh*, we were surprised to see that the error rate went up for the smaller sized failures. The reason for this behavior is clear from Fig. 4, in which we plot the fraction of invalid routes that we correctly identify (efficiency). As we can see, the "efficiency" for "threshold=1" is pretty much constant for different failure sizes. If we increase the threshold, the amount of data analyzed per invocation (of our scheme) is increased; however the total amount of data analyzed across the network goes down because our scheme is executed less often. Thus for smaller failures, the effectivity of our scheme is decreased leading to lower "efficiency" and higher "error rate". From these results it is clear that we get the best performance when we set *largeFailureThresh* and that is what we used for the rest of the results in this section.

We take a look at the effect of partial deployment on the performance of our scheme in Fig. 5. We implement our scheme only at those nodes whose degree exceeds a certain value (*highDegreeThreshold*) and compare the improvement in the convergence delay to the case where we implement our scheme at all the nodes. We see that the convergence delay doesn't increase significantly when *highDegreeThreshold* is increased from 0 to 4. However there is a noticeable difference when *highDegreeThreshold* is increased to 6 or 8. With our current degree distribution, the fraction of nodes with degree greater than 4 is roughly 20%. Thus we are able to improve the convergence delay by up to 75% with our scheme running

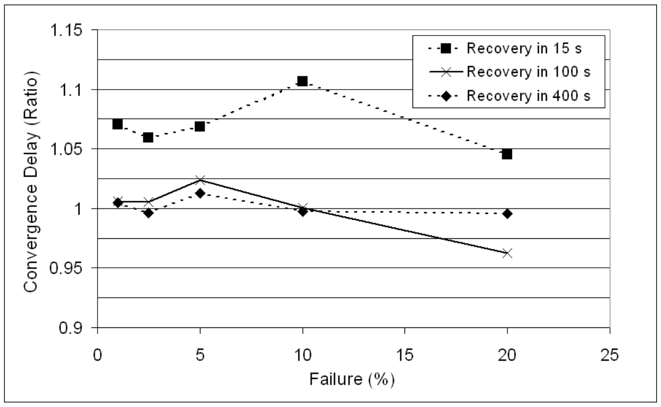

Fig. 7. Convergence delay for recovery scenario

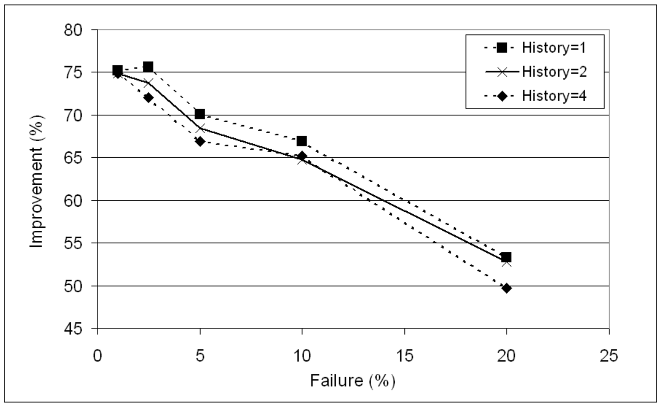

Fig. 8. Effect of *history* on convergence delay

at just 1 out of 5 nodes. We are also able to decrease the delay by half with the scheme running at only 1 in every 10 nodes (*highDegreeThreshold*=8). This is significant because we can get sizeable benefits without implementing our scheme at smaller ASes that typically have limited connectivity.

We have seen that the convergence delay increases as *highDegreeThreshold* is increased. This happens because the fraction of the invalid routes that we are able to identify (efficiency) decreases (in general) when we run it at fewer nodes. For example, efficiency is equal to about 45% when *highDegreeThreshold* is equal to 4, but the efficiency drops to the 30 to 35% range when *highDegreeThreshold* is increased to 8. However, increasing the *highDegreeThreshold* does have a benefit, namely, decreasing the “error rate” (introduced earlier in this section). For example, increasing the *highDegreeThreshold* from 0 to 8, dropped the maximum “error rate” from 20 to 10%. Thus, a lower *highDegreeThreshold* increases the probability that we mistakenly consider some routes as invalid. In order to study that effect, we measure another parameter that we call “lost connectivity”.

After a failure, connectivity to some destinations is truly lost. However other destinations might not be accessible for some time even though there might exist a valid path from the source. For example, standard BGP might prefer shorter invalid routes over longer but valid routes. Similarly in our scheme, a router might not have a path to an AS because all the candidate routes are mistakenly considered to be invalid. We say that a router “lost” connectivity to a “connected” AS for *x* seconds if the router did not have a valid route to that AS for a total of *x* seconds during the convergence process. Let us assume that a router did not have a valid route to AS 1 for 2 seconds and it did not have a valid route to AS 2 for 3 seconds. If the router had a valid route to all other connected ASes at all times during the convergence process, then the total “lost connectivity” for that router is 5 seconds. We measure the “lost connectivity” for all routers in the network and compute the average. We show the difference in the “lost connectivity”, as compared to standard BGP in Fig. 6. We observe that with a *highDegreeThreshold* of 4 or higher the “lost connectivity” is improved marginally. However, for *highDegreeThreshold* equal to 0, there is a significant increase from the standard case.

Thus we see that the convergence delay can be improved significantly without increasing the “lost connectivity” if we run our invalidation scheme at the high degree nodes. It is simple to see the reason why running our scheme at low degree nodes does not improve the performance. Let us consider an extreme case in which a node *A* is connected only to one other node *B*, which has a large number of neighbors. *B* would receive a large number of updates during a large failure, only a subset of which will be forwarded to *A*. Running the invalidation scheme at *A* does not provide any additional benefit, since *B* can do a better job. Furthermore, *A* might consider *B* to have failed if there are a large number of route withdrawals. For the remaining results in this section, we use a *highDegreeThreshold* of 4.

We also evaluated the performance of our scheme on some other topologies. Numerous studies have found that the degree distribution in the Internet decays as a power law. In fact the “realistic” degree distribution that we used had a decay rate of about -1.85. We generated topologies with decay rates of -1.5 and -1.6 (both with average degree 4) and measured the difference in the convergence delay and “lost connectivity” for standard BGP and our invalidation scheme. We again found that our scheme provided big improvements in convergence delay without any significant increase in the “lost connectivity”. We would like to verify the effectiveness of our scheme for a larger number of topologies, and that work is ongoing.

We have seen that our scheme works well for large scale failure scenarios. However this improvement in convergence delay should not come at the cost of worse performance in recovery scenarios. We therefore carried out experiments in which we recovered the failed routers after some delay. After the recovery all the routes are actually valid, however our scheme might still consider some routes to be invalid, because of the length of the time slot or because a large number of updates are still flying around. We show the ratio of the convergence delays for our scheme and standard BGP, for different recovery scenarios in Fig. 7. Markopolou [23] et al.

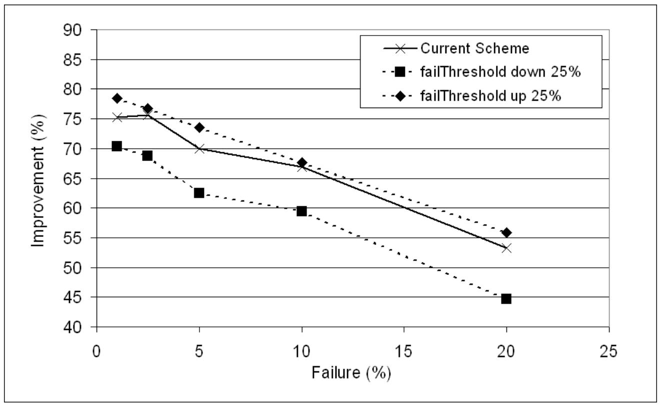

Fig. 9. Effect of changing the value of the *failThreshold*

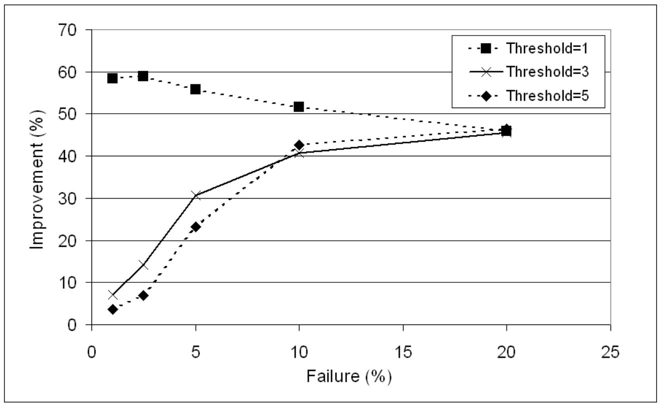

Fig. 10. Convergence delay for multi-router ASes

measured the "time to repair" for router failures and found the median value to be about 400 seconds. If we recover the failed nodes/ASes after 400 seconds, the convergence delay for our scheme is effectively equal to that for standard BGP. We ran simulations with a lower recovery times (15 and 100 seconds) to check if the performance changes drastically. For a recovery time of 100 seconds, we see that the difference is a few percent at most. For a recovery time of 15 seconds (highly unlikely for large-scale failures), the convergence delays are moderately greater ($\sim$10% at the higher end) than the base case. However even in that scenario, the deterioration in the convergence delay is significantly less than the improvement that we observed in Fig. 2.

All the results that we have looked at till now, used a *history* of 1 time slot. It makes the most sense to have a history of 1 time slot, because that way will be basing our decisions on the most recent data that we have. A longer *history* does provide us more data, but that data could be stale. We look at the effect of *history* on the convergence delay in Fig. 8 (*largeFailureThresh*=1). We see that the effect is not very strong, but the performance does go down as *history* is increased. We also found that the "error rates", that we talked about previously, went up with the *history*. We also explored the effect of changing the way *failThreshold* (Algorithm 1) is calculated. We scaled the value of the *failThreshold* up and down by 25% and measured the performance. In both the cases, there was not much change in the "lost connectivity". The effect on the convergence delay is shown in Fig. 9. As we can see, the convergence delay increases moderately when we decreased the *failThreshold*. When we increase the *failThreshold*, the convergence delay goes down a few percent but there was a big jump in the "error rate" as well. Hence we believe that our current method of calculating the *failThreshold* is a good compromise.

Finally we evaluated the performance of our invalidation scheme for topologies with multiple routers per AS. We briefly describe the procedure for generating such topologies here. We selected the number (1-100) of routers in an AS from a heavy tailed distribution. Studies of the real Internet topology have found that the geographical extent of an AS is strongly correlated to the AS size (i.e., number of routers in the AS) [24]. We assumed a perfect correlation and made the geographical area (the region over which the routers of an AS are placed) of an AS proportional to its size (number of routers). The routers of an AS are distributed randomly over the geographical area assigned to it. Internet studies also show that larger ASes are better connected [25]. Therefore, we generated a set of inter-AS degrees using the "realistic" degree distribution and assigned the highest degrees to the largest ASes. The ASes are linked together using a pseudo-preferential connectivity scheme in which one of the ends of an edge is selected randomly but the other end is selected according to the degree of the node. Once an inter-AS edge has been created, we randomly select a router from one of the ASes and preferentially connected it to a nearby router in the other AS. We used the default Waxman scheme for creating the intra-AS edges.

In Fig. 10 we plot the improvement in the convergence delay for the multirouter-AS topology. The trends are similar to what we observed in Fig. 2, with the lowest *largeFailureThresh* providing the best results. We also see that the improvement in the delay is somewhat less than what we observed in Fig. 2. This can be attributed in part to the possibility of partial AS failures. A failure could affect the capability of an AS to provide conduit to one destination but not to another. However, our scheme is not sophisticated enough to make such distinctions. Despite this issue, we are able to reduce the convergence delays by more than 50%.

We have seen in this section that our invalidation scheme can greatly improve the convergence delay for large-scale failures without either degrading the performance for recovery scenarios or increasing the "lost connectivity" during the convergence process. Furthermore, we have also observed that, for both the *largeFailureThresh* and the *history* parameters, the lowest values provided the best results, thus effectively eliminating the need for the parameters.

VI. Conclusion

In this paper we proposed and evaluated a speculative invalidation strategy, designed to reduce the convergence delay for large-scale failures. We identify failed/unstable ASes by analyzing the BGP updates received at a router. Routes containing these "suspect" ASes are then removed from contention. This action goes a long way in reducing the uncertainty, about the validity of routes, faced by BGP during the convergence process. As a result the convergence delays were reduced greatly. At the same time, there was no significant increase in the convergence delay for realistic recovery scenarios. We found that all the benefits of our scheme can be obtained even if we implement it at a subset of "high degree" nodes.

Currently we are looking at ways to improve the performance of our invalidation scheme further. The algorithm that we used to identify the "suspect" ASes is fairly lightweight. It is also very general as we do not make any assumptions about the preference functions used to select the best routes. We are considering more complex (both in terms of time and space) schemes that might be able to provide even better results, possibly by independently analyzing updates sent by a neighbor or updates sent for a group of prefixes. However the better performance will have to be traded off against the increased complexity.

Our scheme is aimed at reducing the convergence delay after a failure. However the delay tells us about only one aspect of the convergence process. Another important aspect is the "lost connectivity" parameter that we introduced. This parameter is a measure of the total loss of connectivity (between all possible source destination pairs) as a result of the failure. Any reduction in "lost connectivity" will definitely reduce the impact of a network failure. We are working towards schemes designed for that purpose.

References

[1] Y. Rekhter, T. Li, and S. Hares, "Border Gateway Protocol 4," RFC 4271, Jan. 2006.

[2] "The Border Gateway Protocol". [Online]. Available: http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/bgp.htm

[3] C. Labovitz, G. R. Malan, and F. Jahanian, "Internet Routing Instability," *IEEE/ACM Transactions on Networking*, vol. 6, no. 5, pp. 515–528, Oct. 1998.

[4] Labovitz, C., Ahuja, et al., "Delayed internet routing convergence," in *Proc. ACM SIGCOMM 2000*, Stockholm, Sweden, Aug. 28–Sep. 1, 2000, pp. 175–187.

[5] Dan Pei, B. Zhang, et al., "An analysis of convergence delay in path vector routing protocols," *Computer Networks*, vol. 30, no. 3, Feb. 2006, pp. 398–421.

[6] T.G. Griffin and B.J. Premore, "An experimental analysis of BGP convergence time," in *Proc. ICNP 2001*, Riverside, California, Nov. 11–14, 2001, pp. 53–61.

[7] D. Obradovic, "Real-time Model and Convergence Time of BGP," in *Proc. IEEE INFOCOM 2002*, vol. 2, New York, Jun. 23–27, 2002, pp. 893–901.

[8] G. Siganos and M. Faloutsos, "Analyzing BGP Policies: Methodology and Tool," in *Proc. IEEE INFOCOM 2004*, vol. 3, Hong Kong, Mar. 7–11, 2004, pp. 1640-1651.

[9] X. Zhao, D. Pei, D. Massey, and L. Zhang, "A study on the routing convergence of Latin American networks," in *Proc. LANC 2003*, La Paz, Bolivia, Oct. 4–5, 2003, pp. 35–43.

[10] A. Sahoo, K. Kant, and P. Mohapatra, "Characterization of BGP recovery under Large-scale Failures," to be presented at ICC 2006.

[11] Di-Fa Chang, R. Govindan, and J. Heidemann, "The Temporal and Topological Characteristics of BGP Path Changes," in *Proc. ICNP 2003*, Atlanta, Georgia, Nov. 4–7, 2003, pp. 190-199.

[12] M. Caesar, L. Subramanian, R. Katz, "Towards localizing root causes of BGP dynamics," U.C. Berkeley Technical Report UCB/CSD-04-1302, November 2003

[13] A. Feldmann, O. Maennel, Z. M. Mao, A. Berger, and B. Maggs, "Locating Internet Routing Instabilities," in *Proc. ACM SIGCOMM 2004*, Portland, Oregon, Aug. 30–Sep. 3, 2004, pp. 205-218..

[14] M. Lad, A. Nanavati, D. Massey, and L. Zhang, "An algorithmic approach to identifying link failures," in *Proc. IEEE PRDC*, Papeete, Tahiti, Mar. 3–5, 2004, pp. 25-34.

[15] R. Teixeira and J. Rexford, "A measurement framework for pinpointing routing changes," in *Proc. ACM SIGCOMM workshop on Network troubleshooting*, Portland, Oregon, Sep. 3, 2004, pp. 313–318.

[16] F. Wang, L. Gao, "On inferring and characterizing internet routing policies," in *Proc. IMC 2003*, Miami, Florida, Oct. 27–29 2003, pp. 15–26.

[17] A. Medina, A. Lakhina, et al., "Brite: Universal topology generation from a user's perspective," in *Proc. MASCOTS 2001*, Cincinnati, Ohio, August 15–18, 2001, pp. 346-353.

[18] "SSFNet: Scalable Simulation Framework". [Online]. Available: http://www.ssfnet.org/

[19] B. Waxman, "Routing of Multipoint Connections," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 9, pp. 1617–1622, Dec. 1988.

[20] A.L. Barabasi and R. Albert, "Emergence of Scaling in Random Networks," *Science*, pp. 509–512, Oct. 1999.

[21] T. Bu and D. Towsley, "On Distinguishing between Internet Power Law Topology Generators," in *Proc. IEEE INFOCOM 2002*, vol. 2, New York, Jun. 23–27, 2002, pp. 638–647.

[22] B. Zhang, R. Liu, et al., "Measuring the internet's vital statistics: Collecting the internet AS-level topology ," *ACM SIGCOMM Computer Communication Review*, vol. 35, issue 1, pp. 53–61, Jan. 2005.

[23] A. Markopoulou, G. Iannaconne, S. Bhattacharrya, C-N. Chuah, and C. Diot, "Characterization of Failures in an IP Backbone," in *Proc. IEEE INFOCOM 2004*, vol. 4, Hong Kong, Mar. 7–11, 2004, pp. 2307-2317.

[24] A. Lakhina, J.W. Byers, et al., "On the Geographic Location of Internet Resources," *IEEE Journal on Selected Areas in Communications*, vol. 21 , no. 6, pp. 934–948, Aug. 2003.

[25] H. Tangmunarunkit, J. Doyle, et al, "Does Size Determine Degree in AS Topology?," *ACM SIGCOMM Computer Communication Review*, vol. 31, issue 5, pp. 7–10, Oct. 2001.