Reducing BGP Convergence Time by Fine Tuning the MRAI Timer on Different Topologies

Pradeep Kirsur and Omar Younis Alani
School of Computing, Science & Engineering
University of Salford
Manchester, M5 4WT, UK
P.Kirsur@edu.salford.ac.uk; o.y.k.alani@ salford.ac.uk

Abstract—**In this paper, we use a discrete event simulator for testing convergence behavior of the only existing inter domain routing protocol BGP. The purpose of this paper is to analyze and understand the convergence behavior of BGP under different network topologies and to fine tune to get better convergence time. OPNET, a discrete event simulator, modular simulation tool that allows simulation of Autonomous System (AS) level topologies is used with varying parameters. The MRAI timer is manipulated and tested on convergence time, and the results are compared with simulations done by similar work in the literature. The effect of different MRAI value was tested on the convergence for different network topologies and we were able to see changes in convergence time. We conclude that MRAI helps in controlling the increase or decrease in the number of updates and the convergence time. However, the tradeoff is between convergence time and end to end traffic loss/delay and BGP traffic.**

Keywords - BGP,Convergence,MRAI, Autonomous System

I. INTRODUCTION

BGP is the core routing protocol of internet i.e., de-facto inter domain routing protocol [3]. BGP-4 is used to establish and maintain connectivity between the autonomous systems in the internet. BGP is that which makes internet work. The convergence property of BGP has been widely studied as BGP has unreasonable behaviour with respect to convergence [1]-[4], [6] and [7]. BGP is different from other preceding routing protocols in that it manages the routes efficiently and more effectively by using routing policies. The route information is made of the network's topology and the different route policies are used. The routing information that BGP uses can be used to improve the network performance, topology design and routing policies [4]. Fine tuning convergence can help in faster re-routing of the packets, lesser packet loss in cases of failure and reduced end to end delay. Network faults such as restarts or flaps have a latency ranging from several seconds to several minutes. Packets are lost during such failures and end to end performance is affected. The delay in convergence is due to the failures and restarts happening in the network. BGP walks through the routing table and calculates all possible paths and decides on best path using the BGP algorithm. The more the number of paths in the network, the more time is taken by BGP as it has to walk through all the routes. The delay observed in path failures can range from 3 minutes to 15 minutes but the reason for this is not known [5]. The reason for the increased number of BGP messages is also not clearly understood [5]. Griffin [8] discusses about the inconsistencies that cause divergence and further introduces the stable paths problem. Convergence optimization can be done by correctly using the MRAI timer w.r.t topology. Since MRAI is one such parameter that affects convergence, we simulate and measure convergence times

OPNET supports discrete event simulation which is a packet based simulation and is best suited for research related to protocol behavior and application performance and hence OPNET was chosen.

We run simulations on simpler topologies such as ring and then move on to partial mesh, mesh and crystal topologies. The results are compared to that of Deshpande [1] and Wang [2].The results confirmed to the convergence time values observed by [1]. In the work by Wang [2], deviation ratios were found for the convergence simulation delay for different topologies such as ring, tree, focus and clique. We found that our results did not hold good for ring and crystal topologies. For down phase, the convergence bounds given by Wang were same for ring and clique topology; however in our simulations, the convergence time for ring is much smaller than that of clique topology and this is because of absence of multiple backup paths in ring topology compared to that of many backup paths in mesh network.

Secondly, we considered different MRAI values in order to test the BGP performance. Different MRAI values used in these tests have caused changes in the routing tables, number of BGP updates and the traffic. The convergence is greatly influenced by different kinds of settings of timers for different kinds of BGP messages. The convergence performance and the BGP message complexity are also greatly influenced by the timer settings. In this paper, we investigate this influence not only on one topology but different topologies. The simulations showed different convergence times which were topology specific. Lastly, the topologies were modified to include more Autonomous Systems (AS) to check if convergence varies with respect to network size. Network sizes of 5 AS were used. One drawback of OPNET is the manual configuration of neighbors which makes it hard to

configure and run simulations for topologies of size 15 or more. We used clique topology, partial mesh, mesh and crystal topologies, results of which are described in coming sections.

When the effect of MRAI was tested for different clique sizes, we observed that MRAI controlled the number of updates. The more the number of updates, the faster is the convergence but there is an increase in the traffic which might be an overhead in real time networks where the actual traffic conflicts with BGP traffic. Bigger MRAI values led to circulation of false information because the nodes are not updated instantly about the instability but only after the timer expires. This eventually led to increase in number of updates. Until the timer expires, the false information circulates through the network and cause more failures and instabilities. In the rest of the paper, a summary of reference work and motivation for this research is given in section 2. Section 3 describes the methodology. Section 4 includes simulations and results. Section 5 consists of conclusion and lastly the future work.

II. Related Work and Motivation

A. Experiments and Measurements

In this section of the literature review we present studies that have been conducted using experimental and measurement methods. In [6] several un-expected, unreasonable characteristics were found when BGP was studied by collecting real time routing tables of five routers which were placed in the public network. From the data collected, it was found the stability related factors such as variations in the topology and the changes in policies within the autonomous systems make it difficult to explicitly measure the instability and the key reason for this is the sheer size of internet [6]. Instabilities include fluctuations in routes, configuration errors, flapping of links, link restart, etc which in some cases are vendor dependent and in other cases are independent. These dependent and independent behaviors pose a big challenge to internet engineering. Such instabilities lead to problems such as packet loss, increased latency and delayed convergence and eventual degradation of internet efficiency and performance. Three main instabilities called as forwarding instabilities, routing policies and pathological updates, are the reasons for instability across the internet, the authors informally define this overall instability as the rapid change in network reach-ability and the topological information across the internet [6]. The results of the experimental study confirmed that most of the BGP messages consisted of uncontrolled and unreasonable announcements and that the instability was excessively conquered by prefixes of definite lengths [7]. Increased network latency, delayed convergence experienced by core routers on internet are due to increase in internet instabilities [7]. Recent work in this area [6],[7],[8] and [9] has shown that the complexity of the internet is poorly understood, for instance, the efficiency of internet routing is affected by the routing policies and that the routing policies make it difficult to understand and analyze the internet stability [8]. The process of route flap damping is not well understood in large scale network [10]. During the study, when there was lesser number of flaps, the dynamics of routing deviated from the likely behavior and resulted in increased convergence time. Though route flap damping is effective in contributing to internet stability, it sometimes incorrectly dampens stable routes too [11]. The convergence time is badly affected by route flap damping but sometimes, a route which flapped just once might be suppressed for an hour which is not desired [12]. The path vector protocol increases the number of possible route oscillations and the theoretical upper bound of convergence time found does not hold good in real time network [13]. The delay observed in path failures can range from 3 minutes to 15 minutes but the reason for the same is not known. Labovitz also say that the reason for the increased number of BGP messages is also not clearly understood. Labovitz also measured the latency and number of updates that included withdrawals and announcements. BGP will still undergo temporary oscillations which is unavoidable [14]. MRAI can be used to reduce the number of update messages sent [13]. A more practical model of BGP that involves MRAI timer and its influence on different topologies is described in [9],[15]. The paper involves measurement of convergence time and the impact of path selection procedure. The way in which BGP path selection occurs, routers list many alternate paths and replace them every now and then until they are finalized with one stable path and MRAI timer limits and controls the rate of advertisements. These two are the major causes for delayed convergence time [16].

B. Theoretical Work and Simulations

Two models namely SPVP and SPP were proposed by [17] which describe the BGP semantics and dynamics. It is a very hard task to achieve particular convergence time [17]. The convergence behavior determines the network performance in case of network faults. This can be optimized by correctly implementing BGP [18].

A method was proposed by Bremler-Barr [19] to avoid Ghost information circulating in the network. The false information that keeps circulating within the network causing complexities is referred to as Ghost Information .The author claims that ghost information affects the convergence because it remains in the network for some time until it disappears thereby inducing delay in convergence [19]. During routing convergence, BGP spends most of the time in searching for alternate paths and exploring the best paths which results in more number of updates and long convergence [20]. Convergence instability is of two types: fault-agnostic instability and distribution-inherent instability [21]. These instabilities induce and increase the chances of rearranging of the messages which is not desired in case of multimedia applications, this degrades the performance of TCP as well. Other method such as root cause notification is also proposed [22]. Griffin [16] further observed that every topology has a certain MRAI that results in favorable convergence time.

III. Our Methodology

This section includes a description of the methodology used in this paper. As we know, simulation is a way of modeling a complex system into a lesser complex system while retaining the original behavior of the original system, the simulations here, are simplified. To make things easier, some assumptions are made: a node represents a router which belongs to one AS. Hence one node is one AS i.e., there is only one router in each of the AS. The RIP is the chosen IGP running within the AS. The complex behavior of IBGP and routing policies are ignored. Multiple BGP sessions on one node are ignored as they bring in complexity. Whenever there is a convergent event taking place, the route update messages propagate and are delivered to the corresponding router in the AS. Events such as shutting down of a physical link during simulations are carried out to note the number of update messages and convergence. The propagation of update messages is observed in the form of routing updates. The number of update messages also lead to number of routing table entries, OPNET allows to export the routing tables and also lets us plot graphs indicating the number of routing table entries in relation to the time the convergent event in the network.

Update propagation and observation method: Update messages are crucial for understanding the routing changes and convergence. But it is not possible to answer all the questions by looking at the updates. Therefore researchers such as Labovitz and Mao [12] induced faults into the network. These faults led to the variation in the updates and other changes such as packet loss rate, delay in traffic received at one end, and routing changes. This was done using beacons and monitoring points. In this methodology, a similar approach is used to observe the routing changes and convergence behavior. Many of the previous researchers have used Up-Phase and Down-Phase. In this paper, the convergence time in Down Phase is analyzed. The down phase is when the destination prefix origin BGP node announces to its peers, the un-reachability of the prefix, and upon this notification, all the peers converge again.

In this methodology, not all BGP updates are considered as not all updates are important. There is more importance to traffic flow and a stable state is desired after a routing change takes place. Therefore, a set of updates are observed at specific time intervals. By using a timeout period during simulation, group of updates are observed. The convergence time was considered to be the total amount of time elapsed from the time the link goes down till the time the router shows up the last update in the routing table. At any given point T1, number of BGP updates is observed which are caused due to the failure of the link. Similarly, the last set of updates can be observed at time T2 which we call as convergence point. In between these two times, path exploration takes place and some intermediate routes are generated. The previous best path and new best path are observed in the routing table. Faults are injected at predefined amount of time , i.e., the time at which failure happens and time at which the last set of routing updates (i.e, convergence point) are observed. The instability events are well separated by using an appropriate timeout value so as to calculate the correct convergence time, the events should be well separated; this leads to a distinctly visible set of routing updates.

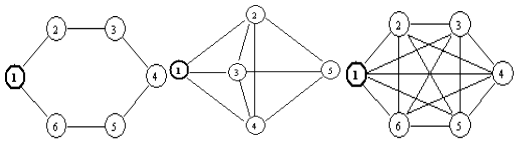
The test setup is explained as below:

1. Load the network topology by generating traffic using a server.
2. Induce a failure into the network (eg., Link flap or failure)
3. BGP should then detect the failure and forward the traffic over alternative link, observe the route changes in route tables.
4. Measure the occurrence of updates on all the routers, observe the first update after the instability event takes place and observe the last update on any of the router concerning the failure. Calculate the time difference using the first update T1 and last update T2.

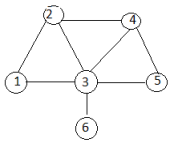
Alternatively, packet loss or frame loss can also be measured using external monitoring software agent.

IV. Experiments and Results

Whenever a withdrawal event takes place, BGP explores new routes which leads to exchange of update messages which inturn increase the convergence time. The MRAI timer limits these updates and this comes in the form of number of MRAI rounds [1]. The convergence lower bound is at-least (n-3) rounds of MRAI where 'n' is the number of ASes [18]. MRAI when applied reduced update messages and hence the complexity. In the simulations conducted, the MRAI timers are synchronized. In all the scenarios, we use the 'update method' to calculate the convergence time and test the same on different topologies with varying MRAI values. The effect of MRAI is tested on four different clique topologies: Partial mesh, Ring topology, Full mesh and Crystal topology as shown below:



a. Ring b. Crystal c. Mesh



d. Partial Mesh

Figure 1- Various network topologies

A. Simulation Setup

We carried out 40 simulations per scenario to measure accurate values for convergence time and to measure the number of updates, to study the effect MRAI timers on the convergence time in a clique topology once a failure happens. The simulation consisted of 5-6 ASes. Each router is belongs to one AS. The experimental system consisted of 5-6 routers, forming a clique topology. To test the impact of MRAI on topology behavior, 40 simulations per scenario were run for MRAI values 10,20,30, and 60.The simulation run time was set to 20 minutes. Multiple link failure scenarios were also simulated to observe the change in convergence time. This was done by including a failure/recovery specification module in OPNET which allows us to fail/recover specific links in the topology. The failure recovery module fails a link at specific time T1 and recovers at time T2 thereby causing advertising/withdrawal messages during the instability event. The time difference between the failure and recovery was kept long enough to see different set of updates in the routing table. Keep-alive interval was kept to be 30, and hold time interval was kept 90.The routing decisions were based on shortest path first selection method. The total number of updates after the failure event was recorded and convergence was measured using the first set of updates and the last update (concerning failure) observed in the router in the topology. The FIB table and routing tables were checked to see the insertion times for the first and last update concerning the failure event. The number of MRAI rounds generated per topology per scenario was recorded by generating the MRAI performance report in OPNET.

B. Results and discussion

The graphs in Figure 2-5 show the performance of different topologies with respect to MRAI timer, number of updates. From Fig-5, it can be clearly seen that the lower MRAI values cause more BGP traffic. This is mainly because at low MRAI values, the routes are explored soon once the failure occurs. The neighboring routers are notified through BGP updates. BGP walks through the route table and finds the next best path. More number of MRAI rounds lead to more number of updates which is shown in Fig 2&3. For complex topologies, more number of MRAI rounds generate more number of BGP updates, this is undesired in real time network where BGP packets flood the network. The excessive path enumeration takes place in case of low MRAI values. When we increase the MRAI value, a limit is put on the path selection, which causes a router to suppress the updates. For MRAI of 10, more number of routing updates were generated and convergence was quicker and more BGP traffic was observed. The convergence is quicker because rate limiting is at its least, neighbors are notified about the failure quickly, this avoids circulation of wrong path information throughout the network. For partial mesh and full mesh topologies, for MRAI 10, huge amount of BGP messages was generated and traffic reduced as MRAI increased. This is because, the MRAI timer imposes a gap between two successive rounds of MRAI. This is shown in Table-1 and Fig-2 where we can see that more complex topology such as full mesh goes through more MRAI rounds and smaller and simpler topologies have very few MRAI rounds. This is because of the presence of multiple backup paths to the same destination. This was reflected in terms of MRAI rounds. If links fail or take a flap, the total number of updates sent and received totally depended on the MRAI value. In case of single failure on partial mesh topology, convergence was found to be slightly more than a minute. In most of the topologies, the routers converged under 2 minutes which hold good with the previous work by [1]. Our results for the convergence times hold good with previous research work by [1].

Simulations were performed in all topologies keeping the MRAI constant at 30 which is the recommended length for MRAI[23]. In all the topologies, MRAI 30 generated ideal amount of traffic with a reasonable convergence time. From Fig-1, we can see that MRAI 30 generated 2,4,2,3 rounds of MRAI for partial topology, mesh and ring and crystal topology respectively. For MRAI of 30, the convergence was found to be 60,31,42 and 71 for partial, ring, mesh and crystal topologies respectively. Compared to MRAI of 10, lesser BGP updates were generated in case of MRAI of 30, but the convergence was found to be more than the latter because for each round of 30s, the convergence delay increases. From the Fig-3&4, we observe that after MRAI 30, the number of updates remain constant for Ring and Crystal topology which is because of absence of invocation of MRAI timer and quick convergence (31 & 42 seconds)which is shown in Table-1. From Fig 4&5, number of updates/BGP traffic decreases as MRAI increases. Thereafter the BGP updates remain constant for ring and crystal topology. More traffic is observed in mesh topology after MRAI 30 because suppression of updates can cause spreading of false information across the network. This false information also referred to as Ghost information can remain in the network and increase convergence time until it disappears [19] .The convergence is then observed to increase linearly while the updates remain almost constant.

MRAI	Partial	Ring	Mesh	Crystal
10	51	22	65	24
20	54	24	70	32
30	60	31	71	42
60	68	45	84	52

Table 1 – Convergence times for each topology for different MRAI values.

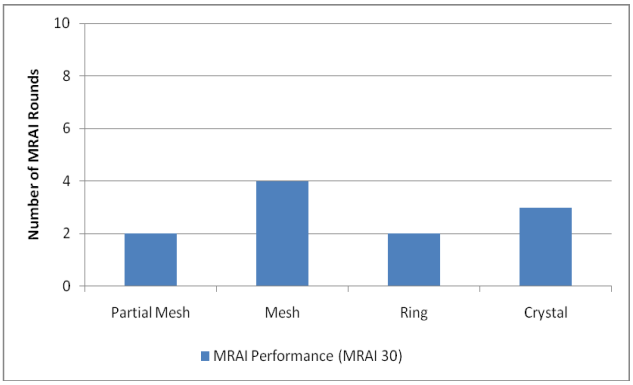


Figure 2 –MRAI performance- MRAI 30

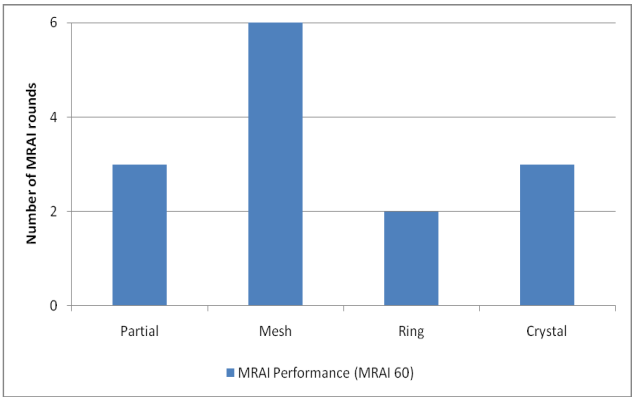


FIGURE 3 –MRAI PERFORMANCE- MRAI 60

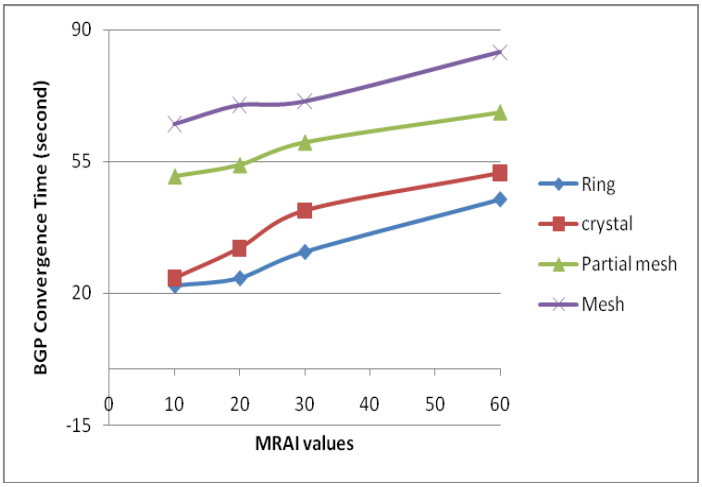


Figure 4 – MRAI Vs Convergence Time

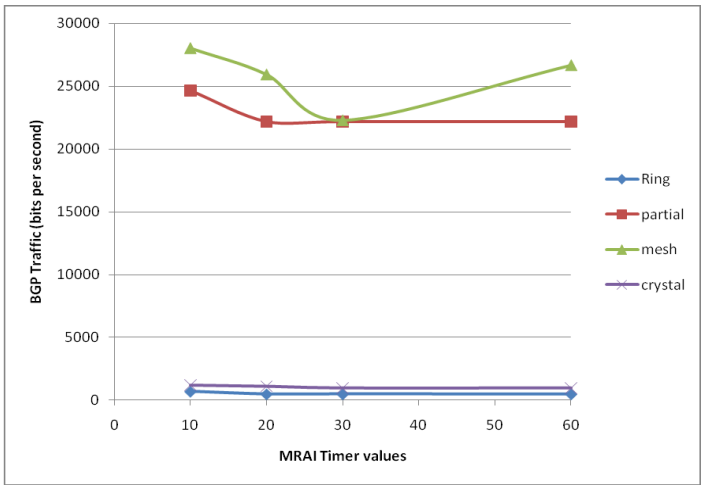


Figure 5- MRAI Vs BGP Traffic (in bits per second)

V. CONCLUSION

To verify the impact of MRAI on convergence, we carried out several simulations under varying parameters and compared the results to the one obtained by others. We performed simulations with the influence of MRAI timers on convergence time and BGP traffic and we found that for complex topologies, MRAI 30 serves ideal with reasonable convergence time, where as MRAI has lesser influence on simpler topologies. Our graphs match with that of Griffin [16] .We saw that number of messages decrease with increase in MRAI. There is one point at which the convergence and MRAI value are ideal for a better performance and this varies for each topology.

VI. FUTURE WORK

Our plan for the future work includes measuring the convergence time specific to ports on the routers. This can also be performed using a third party software which helps in measuring convergence on specific ports on the router. Though complex to implement, this would provide an understanding of convergence behavior. Furthermore, we intend to carry out simulations with different routing policies. We will study the influence of MRAI on specific routers in topology. We also like to study the impact of link failures on queuing delay incurred in routers.

REFERENCES

[1] Deshpande, S., & Sikdar, B , “On the impact of route processing and MRAI timers on BGP convergence times ” Paper presented at the Global Telecommunications Conference, GLOBECOM '04. IEEE , 2004.

[2] Wang, W., Shen, Q., & Zhong, Q. , “On the relationship between BGP convergence delay and network topology.” Communication Technology, 2008. ICCT 2008

[3] Laskovic, N., & Trajkovic, L,”). BGP with an adaptive minimal route advertisement interval”, Computing, and Communications Conference, IPCCC 2006,2006.

[4] Ke, X., Wang, A.-p., Jian ping, W., & Bin, W., “Research on routing policy and routing information propagation of boarder gateway protocol

version 4 (BGP-4)" the TENCON '02. Proceedings,IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, Oct, 2002.

[5] Labovitz, C., Malan, G. R., & Jahanian, F., " Internet routing instability ", IEEE/ACM Transactions on, 6(5), 515-528,1998

[6] Labovitz, C., Malan, G. R., & Jahanian, F, "Origins of Internet routing instability Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE ,1999.

[7] Labovitz, C., Malan, G. R., & Jahanian, F. , "Origins of Internet routing instability" INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies,1999.

[8] Griffin, T. G., Shepherd, F. B., & Wilfong, G .," The stable paths problem and interdomain routing"EEE/ACM Transactions on Networking (TON), 10(2),232-243,2002

[9] Labovitz, C., Ahuja, A., Bose, A., & Jahanian, F., "Delayed Internet routing convergence. Networking"IEEE/ACM Transactions on, 9(3), 293-306,2001

[10] Beichuan, Z., Pei, D., Massey, D., & Zhang, L.," Timer Interaction in Route Flap Damping", Paper presented at the Distributed Computing Systems, June 2005.

[11] Zhenhai, D., Chandrashekar, J., Krasky, J., Kuai, X., & Zhi-Li, Z.," Damping BGP route flaps". Performance, Computing, and Communications, 2004.

[12] Mao, Z. M., Govindan, R., Varghese, G., & Katz, R. H.,"Route flap damping exacerbates Internet routing convergence", 2002.

[13] Labovitz, C., Ahuja, A., Bose, A., & Jahanian, F. "Delayed Internet routing convergence." *ACM SIGCOMM Computer Communication Review, 30*(4), 175-187,2000.

[14] Griffin, T. G., & Wilfong, G. ,"An analysis of BGP convergence properties",1999.

[15] Labovitz, C., Ahuja, A., Wattenhofer, R., & Venkatachary, S,"The impact of Internet policy and topology on delayed routing convergence",2001

[16] Griffin, T. G., & Premore, B. J., "An experimental analysis of BGP convergence time. "Ninth International Conference , 2001.

[17] Obradovic, D. , "Real-time model and convergence time of BGP". Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE,2002

[18] Nykvist, J., & Carr-Motykova, L. ,"*Simulating convergence properties of BGP",* Computer Communications and Networks,Proceedings. Eleventh International Conference on,Oct 2002

[19] Bremler-Barr, A., Afek, Y., & Schwarz, S.," Improved BGP convergence via ghost flushing.",Paper presented at the INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies, April 2003

[20] Beichuan, Z., Massey, D., & Lixia, Z.," *Destination reachability and BGP convergence time [border gateway routing protocol",* Paper presented at the Global Telecommunications Conference, GLOBECOM '04. IEEE. 29 Nov.-3 , Dec. 2004

[21] Hongwei, Z., Arora, A., & Zhijun, L.,"A stability-oriented approach to improving BGP convergence. " presented at the Reliable Distributed Systems, Proceedings of the 23rd IEEE International Symposium, Oct 2004.

[22] Pei, D., Azuma, M., Massey, D., & Zhang, L.," BGP-RCN: improving BGP convergence through root cause notification". Computer Networks and Isdn Systems, 48(2), 175-194,2005.

[23] Rekhter, Y., & Li, T.," A border gateway protocol 4 (BGP-4)" 1995.