International Journal of Future Generation Communication and Networking
Vol.6, No.6 (2013), pp.151-160
http://dx.doi.org/10.14257/ijfgcn.2013.6.6.16

BGP Fast Convergence Based on Message Classification

Wu Xuehui

Modern Education Technology Research Center, Tianjin University of Traditional Chinese Medicine, 88 Yuquan Road, Nankai District, Tianjin, P.R. China, 300193

xiaowu_tianshi@163.com

Abstract

BGP is currently the most popular inter-domain routing protocol used in the Internet, providing stable and secure interconnection schemes for operators, and has a wealth of routing control mechanisms. BGP chooses its route depending on the attribute information contained in the update messages received from its neighbors. It allows each AS to choose their own routing policy, which may result too long convergence time to make packets lost in the application layer. Purely modifying the MRAI timer may reduce the convergence time at some degree, but may also bring some negative impacts like broadcast storms. In this paper, we proposed a BGP fast convergence mechanism by modifying the timer depending on the message type based on the concept that the bad news travels slowly while the good news travels quickly in the Internet. By simulating our mechanism on the SSFnt, we draw the conclusion that it can greatly reduce the convergence time, as well as reducing the redundant packets to some extent.

Keywords: *BGP; routing protocol; fast convergence; SSFnet*

1. Introduction

Border Gateway Protocol (BGP) is currently the standard interdomain routing protocol in the Internet, ant it plays an important role in the network performance. It is a path vector routing protocol, where routers exchange BGP update messages with path informantion for a destination. path vector routing protocols wherein each node advertises the “best” route for each destination to each of it’s neighbors. A BGP node stores all the paths sent to it by its neighbors but uses and advertises only the one that is “ best” according to some criteria. If this primary path fails, BGP selects the next best backup route, which is then advertised to its neighbors. However, there is no guarantee that the backup route is still valid. In case the backup route has also failed, it will be replaced only after a withdrawal is sent by the neighbor which advertised it. At that time, another backup route will be chosen. This absence of information about the validity of a route can cause BGP to go through a number of backup routes before selecting Border Gateway Protocol (BGP) is currently the standard interdomain routing protocol, and it a stable one. Thus, there can be a considerable delay before the cycle of withdrawals/advertisements ends and all BGP nodes have a valid and stable path to the destination. Recent studies have shown that the establishment of stable routes after a node failure can take on the order of 3 to 15 minutes [1]. In [1] it was also shown that in a fully connected network, the lower bound on convergence time is given by Minimum Route Advertisement Interval in an node network where Minimum Route Advertisement Interval is the MRAI interval (usually 30 seconds).

In this paper, we proposed a simple mechanism: by classifying the update message types of the BGP to set the corresponding timer to ease the burden of the

ISSN: 2233-7857 IJFGCN

router and reduce the BGP convergence time in the end. The main contributions of our work are as follows:

Firstly, we divided the BGP update message into different priorities. Further more, we treat different priority messages with different strategies: if a message has a higher priority, we will set its MRAI timer with a small value, while for a low-priority message, because it will not affect the routing decision immediately, we will set its timer with a large value to postpone its processing in order to reduce the router burden. Finally, we simulated our scheme in the SSFnet, and we also compared with the currently typical BGP fast convergence mechanisms. The simulation results showed that our mechanism not only can converge faster, but also can reduce the number of the update messages produced in the convergence, which is a effective BGP fast convergence mechanism.

The rest of the paper is organized as follows: The first section describes the related work, the section is our BGP fast convergence mechanism, in the third part, we simulated our mechanism and analyzed the results, and the last part is the conclusion.

2. Related Work

Previous works[1-4] have concluded that the Minimum Route Advertisement Interval (MRAI) is one of the most important BGP configuration parameters affecting the convergence delay. Research on BGP convergence times has focused on both evaluating the bounds on the convergence times as well as techniques to improve it. In general, BGP does not have any bounds on the convergence times due to its immense flexibility and adaptability achieved from its open-ended structure and policy based routing possibilities [1, 3, 4]. In [5], the authors simulated the BGP convergence process, and it showed how the different BGP MRAI (Minimum Route Advertisement Interval) value will effect the BGP convergence time. It pointed out that, for a given topology, there will be an optimal MRAI value for BGP’s shortest convergence time. This optimal value depends on the network topology, so there is not a unified value for all the topology to converge in the minimum time. In practice, Cisco set 30s as the default value for their BGP routers’ MRAI. Studies have shown that inappropriately shorten the MRAI value may cause longer routing convergence time and more BPG update message.

One of the factors responsible for the behavior observed by Griffin and Premore [6] is the processing overhead of BGP updates. Let’s assume that a node A sends an update to a neighbor B at time t. Let’s also assume that the MRAI is high enough so that all incoming update messages have been processed by the time t+ MRAI. If the MRAI is increased further, it means that the nodes have to wait longer before sending the update messages and this increases the convergence delay. Thus, in this phase, the number of update messages sent remains roughly constant and the delay increases linearly.

In [7], it pointed out that, when all the BGP reached to the steady state, for a given destination prefix, there exists a forwarding tree. The destination node is the forwarding tree root, and its immediate neighbor is as a child in the first layer of this forwarding tree, and so on, constructing a forwarding tree. For every node, there are only two states for the receiving update message: “on tree” or “off-tree”, and every state of the update message is defined a corresponding priority.

[8,9] analyzed BGP routing process, it found that, for the same destination, if the path length is longer than the one announced in the previous time, there will be “ghost information” in the announcement process. In order to delete these “ghost information”, it should tell all the BGP peers to illustrate the path informed before in no longer valid.

4. Classifying BGP Message Priority

In this paper, we classify different BGP update messages into different priorities to set their MARI timer values, which may accelerate the BGP convergence.

In the view of a BGP message receiver, if the destination address of the BGP message is a new one, the message corresponds to a higher priority. Conversely, if it is an old destination address, its priority depends on the update message's state is "on-tree" or "off-tree". If the received update message is sent over along the trunk of the forwarding tree, it is in "on-tree" state, otherwise, it is in "off-tree" state. In the current BGP routing protocol, each sub-tree node knows its parent node, on the contrary, its parent node has no information of its children nodes. In general, we believe that the "on-tree" state update message has a higher priority, as it is good news which may affects routing decision immediately.

When the update message is a withdraw message and in "on-tree" state, which means the path deleted was used before, it implies that the path is very important, which should have a higher priority. On the contrary, it indicates that the deleted path is an alternate path which had not been used before, so it should correspond to a lower priority. If the update message is an "on-tree" state announcement message, it implies that there is a new path to instead the older path announced before, this is a good news which should have higher priority. If the update message is an "off-tree" message, but the path announced is better, that is, it is shorter than that of the older one, it should be in higher priority, otherwise, it should be in lower priority.

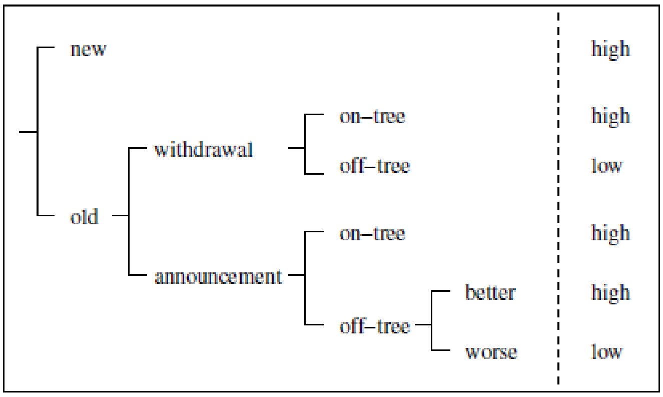


Figure 1. Different Update Messages with Different Priorities

Special treatment for the announcement message: for a same destination, if the announced path is longer that that of the previous, there may be "ghost information" during the convergence time. In order to delete the information in the network to accelerate BGP convergence, it should inform the BGP peers that the path announced previously is no long valid. That is, for announcement message, whether it is in the "on-tree" state or "off-tree" state, as long as calculating a longer path for a same destination address, it implies there are "ghost information" in the network. In this case, the message should be set with higher priority, and the node should send withdraw message for this destination to declare the previous path to this destination is valid. AS a result, Figure 1 shows different messages with different priority.

Once the messages are classified, the next step is how to deal with them. There are two methods to realize it.

First, in the view of the receiver –setting a priority queue: the update message classification is done at the receiver, so the easiest way is to set the messages with different priorities. The advantage of doing so is easy to know the message corresponds to which priority by simply checking the forwarding table. However, the receiver may do a lot of processing work previously, thereby increasing the burden of the receiver.

Secondly, in the view of the sender: set different priority messages with different delay timers. The method is to set different MRAI timer values for different messages with different priorities in the receiver. In this paper we set lower priority message timers with the default timer value as in the BGP protocol, while the high-priority messages use a smaller value of the timer. The advantage of the method can reduce the BGP convergence time, because of the shorter time of the update message elapsed in every hop. To do this, one need only to modify the value of MRAL, making smaller modifications for the protocol, and on the other hand, using this method can fasten the BGP convergence in the same time.

5. BGP Fast Convergence Algorithm

Based on the BGP update message classification and the corresponding timer value, we give the BGP fast convergence algorithm pseudo-code as in Figure 2. The mechanism is realized in the BGP message sender. Messages with higher priority during the transmission are sent directly, while that with lower priority will be postponed. In order to make BGP converge fast, we set the MRAI value of the message with lower priority as the default value in the BGP protocol standard, and the value of the message with higher as half of the default value.

When sending an update message, first it will check the message type. If it is a withdraw message and the path is in the “on-tree” state, which indicating that the deletion of the path is a branch of the forwarding tree, the message is sending directly with the smaller MRAI timer value to accelerate the BGP convergence. If the path is in the “off-tree” state, which means it has not been used before, and postpone his withdraw-path will not affect the path used currently, so the withdraw can be send later to reduce the message produced in the BGP convergence process.

When the message is an announcement message, it first checks the destination address announced is new one or old. If it is a new destination address, it means there is a new path to that message, which is good news. So the message should be set with higher priority. If it is an old destination address, and the announced path is better than the previously one, that is shorter than the previous one, then the announcement message has a higher priority. But if the announced path is longer than the previous one, it means that there may be “ghost information” in the network. In order to remove these message quickly, the message should be sent immediately regardless of whether the timer expires, and the sender will set the path to that destination as Null. Thus, when receiving a new path to this destination, because there is no path longer than the previous null path, it can send announcement immediately. In this way, reducing the time of the “ghost information” in the network, and reduce the BGP convergence time.

```
Upon sending message (type, Peer AS path, destination) to peers of router in AS
If (type==withdraw)
  {
      If (AS path == on tree)
      {
      Send message (withdraw, {}, destination) to each peer at time $LastAnouncementTime_{dst}$+1/2MRAI
      else
      Send message (withdraw, {}, destination) to each peer at time $LastAnnouncementTime_{dst}$+MRAI
   }
If (type==announcement)
{
   If (new AS pathdst!=pre AS path$_{dst}$)
      Send message (announcement, AS pathset, destination) at time $LastAnnouncementTime_{dst}t$+1/2MRAI
 else
{
      If (new AS pathdst is short than previous AS pathdst)
      Send message (announcement, AS pathdst, destination) at time $LastannouncementTime_{dst}$+1/2MRAI
      else
      {
      An empty path ({}) is considered longer than any other paths
      Send message (withdraw, {}, destination) to each peer immediately
      }
  }
}
```

Figure 2. BGP Fast Convergence Algorithm Pseudo-code

6. Simulation

The simulation is in SSFnet, in which two scenarios were tested: generating a new route and link failure e(deleting a route). In each of the scenario, three parameters were analyzed.

(1) ENCT (Effective network convergence time): the interval from a source router sending an update message to all the nodes generating the route to that source node.

(2) AENCT (Average effective network convergence time): the interval that all the nodes had generated the routes to all the other nodes in the network, which is the average of valid network convergence time of all nodes in the network.

(3) NEUM (The number of exchanged update messages in the network): the number of the exchanged update messages during the period that the network begins to work to all the nodes reaching steady states.

In order to simulate the actual operation of the network function, we tested multiple update messages generated by multi-AS simultaneously in seven different topologies. These seven topologies contained node number varied from 29 to 830. in each topology, three nodes were chosen as the routing announcers, and each node sends update messages in the network containing its network prefix. In each topology, all the nodes with its neighbors less than four.

Simulation results

1 generating a new route: in this scenario ENCT, AENCT, NEUM were tested. Parameters at each node is configured as follows in Table 1 where the meanings of the parameters are as bellows:

Table 1. Simualtion Parameters

Parameters	Values
Link delay	0.01-0.1(sec)
MRAIhigh(for high priority)	15 (sec)
MRAIlow(for low priority)	30 (sec)
Number of advertisers	3

Link delay: the processing delay of a message in each node.

MRAIhigh: the MRAI of the message with higher priority, which is 15 second in our simulation.

MRAIlow: the MRAI of the message with lower priority, which is 30 second in our simulation.

Number of advertisers: the number of the advertiser in the same time.

First the performance of the proposed algorithm is tested, and is compared with the standard BGP protocol in which the default MRAI timer were set as 15 second and 30 second respectively. In order to anayse the announcement messages, the three announcers advertise new routing information in different time intervals. When a announcer advertises new routing information, the other information of the two announcers were treat as cross traffic.

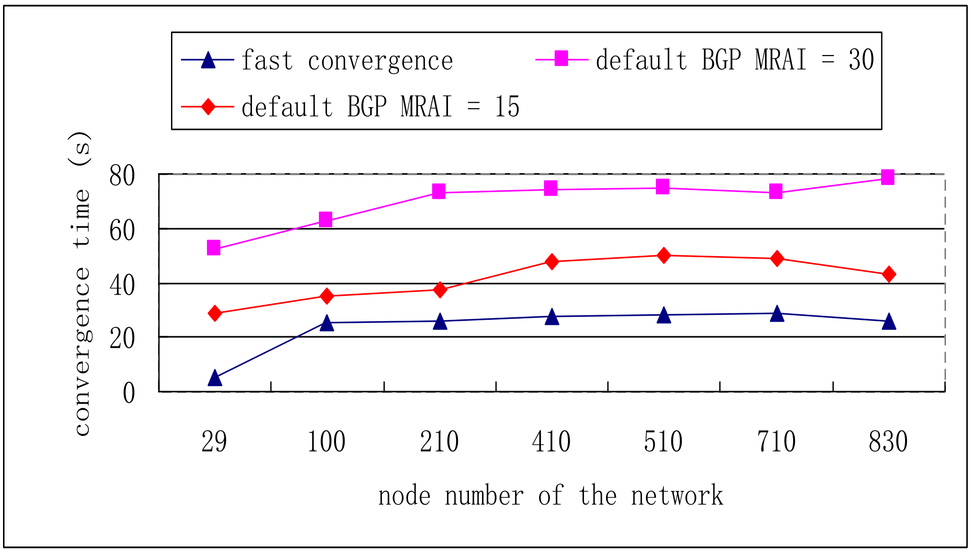


Figure 3. ENCT in Generating a New Route Scenario

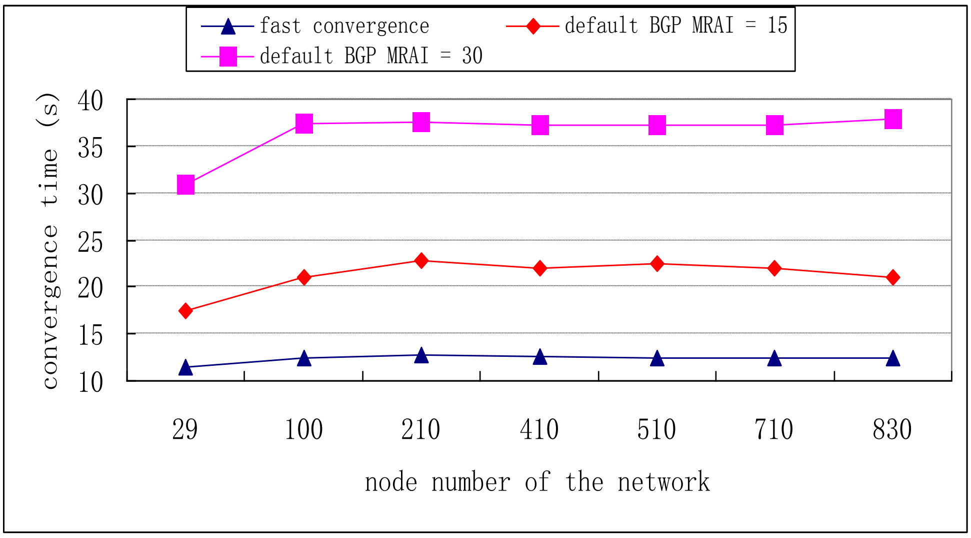


Figure 4. AENCT in Generating a New Route Scenario

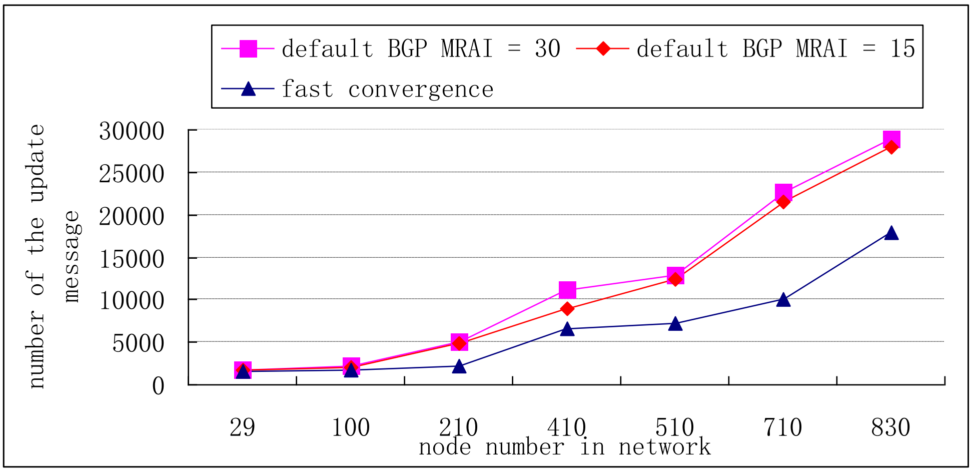
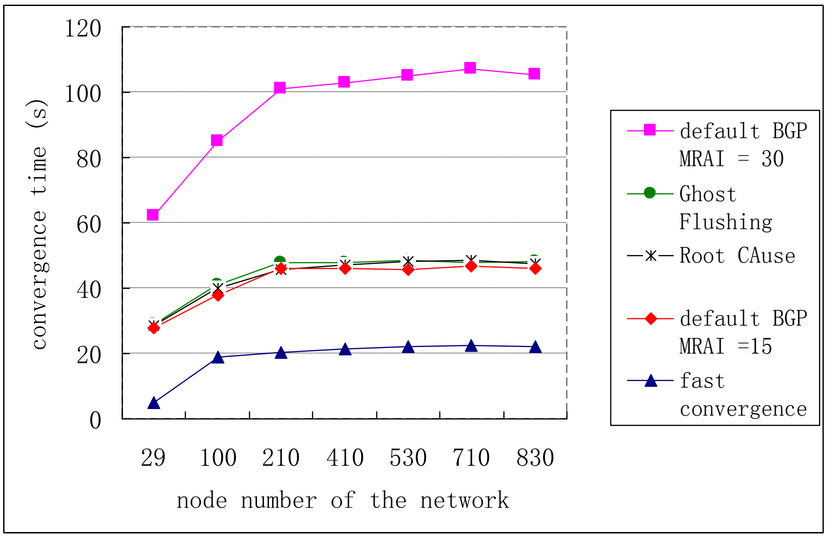


Figure 5. NEUM in Generating a New Route

Figures 3, 4, 5 show the simulation results of the ENCT, AENCT, NEUM respectively. AS shown in Figure 2 and Figure 3, the ENCT and AENCT are smaller than that of the default BGP and default BGP with MRAI is 15 respectively. The update message is also less than the other two.

(2) link failure scenario

In this scenario, in addition to compare three parameters of our algorithm with default BGP and default BGP with MRAI 15, we also compared them with current BGP route convergence algorithms such as ghost flushing and root-cause. We compared the performance with that in the default BGP with MRAI time is 30 and 15 respectively, and with that in root-cause with MRAI time is 15. in each topology, we choose a node with neighbors randomly, and the node advertises routing with different network prefixes as destination addresses to its neighbors. For every network prefix, the node advertised the reachability information to its neighbors. When the network convergence to a steady state, we cut off a certain link between the node to its neighbor, and observe the performance of our algorithm



.

Figure 6. ENCT in Link Failure Scenario

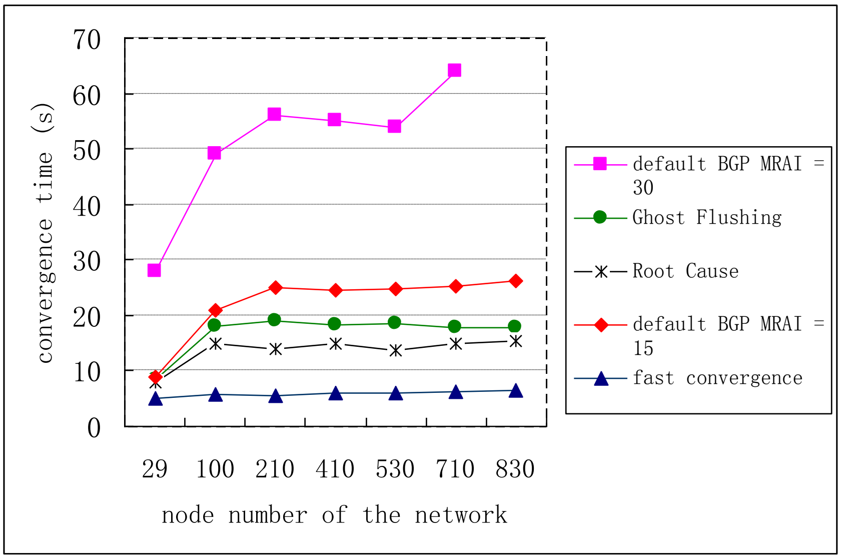


Figure 7. AENCT in Link Failure Scenario

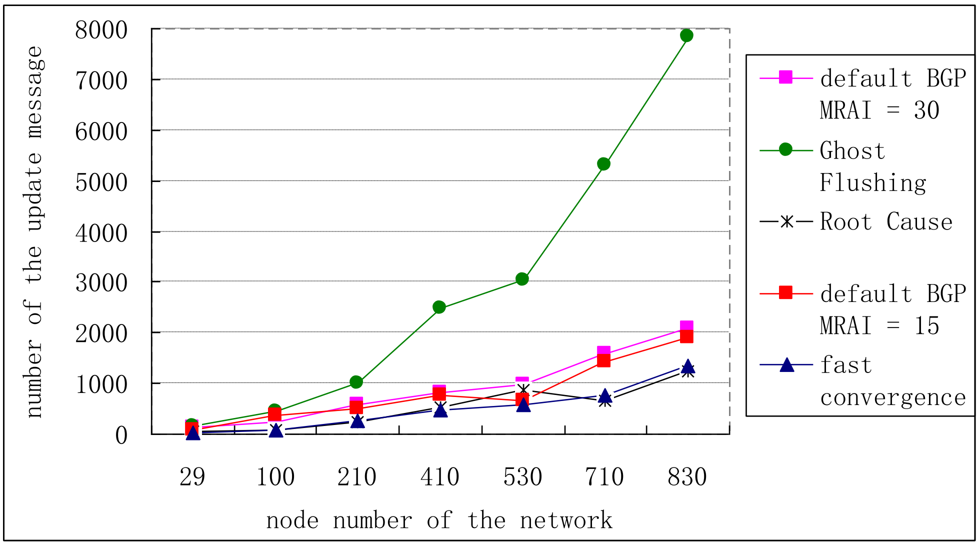


Figure 8. The Number of Update Message in Link Failure Scenario

Figures 6, 7, 8 show the simulation results of the ENCT, AENCT, NEUM when deleting a route in different algorithms respectively. The simulation results showed that not only the BGP convergence time of our proposed algorithm is shorter than that of the default BGP and Root cause when deleting a route, but also the number of the update message number is also less than the other schemes. Although the convergence time of our proposed algorithm is larger than that of the ghost flushing, ghost flushing generated more update messages than that in our proposed algorithm.

6. Conclusion

In this paper, we have proposed a BGP fast convergence mechanism based on the message classification. The message were classified by their states in the forwarding, in which the "on-tree" message has higher priority and smaller MRAI timer value, and the "off-tree" state message has lower priority and larger MRAI timer value. Based on the message classification, we proposed the BGP fast convergence algorithm, and we simulated it in the SSFnet. The results show that our algorithm has smaller BGP convergence time that that of the default BGP protocol, Ghost flushing. It has the similar performance like root cause mechanism, but generating less update message than other mechanisms. In the future work, we will continue to look for ways of improving the proposed schemes.

References

[1] C. Labovitz, A. Ahuja, A. Bose and F. Jahanian, "Delayed Internet routing convergence", IEEE/ACM Transactions On Networking, vol. 9, no. 3, **(2001)** June, pp. 293-306.

[2] Y. Rekhter and T. Li, "Border Gateway Protocol 4," RFC 1771, **(1995)** March.

[3] B. Halabi, "Internet Routing Architectures", Cisco Press, **(1997)**.

[4] D. Pei and B. Zhang, "An analysis of convergence delay in path vector routing protocols," Computer Networks, vol. 30, no. 3, **(2006)** February, pp. 398-421.

[5] C. Villamizar, R. Chandra and R. Govindan, "BGP route flap damping", RFC 2439, **(1998)** November.

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

[7] L. Gao and J. Rexford, "Stable Internet routing without global coordination[J], IEE E /A CM Transactions on Networking, vol. 9, **(2001)**, no. 6, pp. 681-692.

[8] T. Griffin and G. Wilfong, "An analysis of BGP convergence properties", Proceedings of ACM SIGCOMM, Boston, MA, September, pp. 277-288.

[9] Z. Mao, G. Govindan Varghese and R. Katz, "Route Flap Damping Exacerbates Internet Routing Convergence", A CM SIGCOMM 2002 Conference, New York, U SA: A CM press, **(2002)**, pp. 221-233.

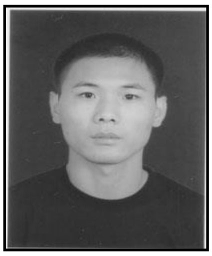
[10] D. Gupta and A. K. Sharma, "Comparative Investigations on Performance of Routing Protocols in Presence of Realistic Radio Models for WSNs", IJAST, vol. 29, **(2010)** April, pp. 101-112.

[11] P. Nand and S. C. Sharma, "Performance study of Broadcast based Mobile Adhoc Routing Protocols AODV, DSR and DYMO", IJSIA, vol. 5, no. 1, **(2011)** January, pp. 53-64.

[12] M. Bazarganigilani, "Web Service Selection Using Quality Criteria and Trust Based Routing Protocol", IJSIA, vol. 6, no. 4, **(2012)** October, pp. 109-118.

[13] D. Gupta and A. K Sharma, "Investigations on Energy Efficiency for WSN Routing Protocols for Realistic Radio Models", IJFGCN, vol. 4, no. 3, **(2011)** September, pp. 61-72.

Author



Wu Xuehui, received his B.Sci degree in Education Technology from Tianjin University of Technology and Education (TUTE) and M.Eng degree in Computer science and technology from Nankai University,PR china in 2003 and 2008 respectively. He is currently researching on Business Intelligence (BI) and Network Computing.