Simulating Convergence Properties of BGP

Johan Nykvist and Lenka Carr-Motyčková

{johan, lenka}@sm.luth.se

Division of Computer Science and Networking

Department of Computer Science and Electrical Engineering

Luleå University of Technology

S-97187 Luleå, Sweden

Abstract— We present a simple BGP simulator that is used for testing convergence properties of inter-domain routing. The discrete-event driven simulator contains a high-level abstraction model of the BGP protocol. The purpose is to perform simulations to examine the complexity of routing information exchange and to test different possible reasons for delayed convergence of BGP routing. We have simulated effects of BGP timers on convergence and compared the results with simulations done by Labovitz and Griffin. Our simulator displayed the theoretically predicted characteristics. We then tested the effects of skewed timers, split horizon and the sender side loop detection (SSLD) techniques on convergence for different network topologies. We conclude that SSLD helps to decrease the number of announcements and the convergence time. However, SSLD increases the number of withdrawals.

I. INTRODUCTION

The Border Gateway Protocol version 4 (BGP) [1] is the de facto inter-domain routing protocol used in the Internet today. BGP is a policy-based, path-vector protocol that distributes route information between Autonomous Systems (AS:s). Since the introduction of BGP, its convergence properties have been investigated as it does not perform well in this sense.

The rather loose BGP specification [1] allows the application of different techniques to tune its performance. The specification also suggests using different parameters (e.g., timer settings), but their implementation and the settings are left unspecified. We are interested in interaction of different BGP parameter settings for dynamic topologies and effects of different techniques on BGP's convergence properties.

Characteristics of routing convergence and stability have a high impact on the network's ability to perform quick route repairs and provide good performance in the presence of network faults. Currently inter-domain route repair convergence in the Internet may have a latency on the order of several minutes. During the latency period end-user traffic quality is highly degraded [2]. *Delayed convergence* of BGP is primarily caused by the bouncing problem: the protocol is designed so that it inspects all possible routes between the source and the destination. As the number of paths between autonomous systems grows exponentially with the number of systems, the problem can result in temporary routing table fluctuations (route flaps) after a fault.

In addition to delaying the route selection process, route flapping can cause rapid changes in network reachability. Such a situation is called *routing instability*. Note that, a route is considered stable if it lasts at least tens of minutes. In a more severe case, we can observe *persistent oscillations* that are probably caused by inconsistent setting of routing policies in autonomous systems. Persistent oscillations are high frequency oscillations lasting for several hours or more. They constitute about 10% of observed route flapping. Persistent route oscillations were not only proved as theoretically possible, but recently observed in the Internet and reported at the NANOG 21 conference [3].

Convergence latency and its communication complexity can be optimized by correct implementation of BGP. As many issues such as usage of timers and application of loop detection or split horizon (SH) techniques are left open in the BGP specification, we simulated different mechanisms and measured their convergence properties.

We developed a discrete-event driven simulator with a high-level abstraction model of the BGP protocol and used it in investigations of BGP's convergence properties. We first ran simulations to verify the correctness of the BGP implementation in our simulator as well as the correctness of the simulator itself. We compared our results to observations made by Labovitz [4] and simulation results of Griffin [5]. While our results confirmed Labovitz's observations, we found discrepancies from Griffin's. These slight differences were explained by the use of the SH technique in Griffin's model [6], which is not specifically required by BGP specification.

Second, we tested different settings of Minimum Route Advertisement Interval (MRAI) timers. Settings of timers for different kinds of BGP messages, interaction between timers, and using different timers for different topologies are issues that influence convergence complexity of BGP. We are especially investigating MRAI interaction, for example skewed randomly or evenly. Some experiments show a substantial difference between convergence properties of a model with synchronized timers and a model with skewed timers. As the model with skewed timers is closer to a real network we believe that BGP simulators should be set up with skewed timers in order to get realistic results.

In the final phase, we looked at the convergence properties as they change with the size of the network. We used a clique topology, representing a mesh network, for our experiments, as it is the most complex topology where we get worst case re-

sults. We tested the effect of SH and sender-side loop detection (SSLD) techniques for cliques of different sizes. Both techniques decrease the number of announcements and the convergence time. The effect of SSLD was greater. On the other hand, introducing SSLD or SH increases the number of withdrawals.

In the remainder of the paper, an overview of related work and motivation for our work is given in Section 2. Section 3 describes the simulator and the BGP simulation model. Section 4 presents some experiments performed with the simulator. Finally, in section 5 our conclusions and direction for future work are presented.

II. RELATED WORK AND MOTIVATION

A. Experiments and Measurements

Labovitz, et al. ran an instability measurement project during 1996 – 1998. They were interested in effects of route withdrawals and updates on routing table changes and end-to-end traffic performance. They measured the number of BGP announcements and withdrawals between backbone routers. They observed that the percentage of route flaps grows linearly with the size of routing table [7]. Router misconfigurations, transient link problems and poor implementations of BGP were found to be the most common reasons for instability. Routing instability represented by route flaps resulted in sending redundant messages. Instability caused packet loss to unstable destinations, delayed convergence of routes, and additional overhead (processing routing messages). Instability was distributed across both autonomous systems and prefix space, which suggests that there is a general problem that is not caused by a small set of routers.

In [2] Labovitz, et al. examine the latency of BGP failover, both actively measuring and developing a model that gave theoretical upper and lower bounds for the convergence latencies. Measurements show that failover latency averaged around 3 minutes but can also reach tens of minutes due to routing table oscillations. They measured latency and the number of messages that resulted from route announcements, withdrawals, and replacements of a shorter or longer route. Reasons for delayed convergence stemmed from the ambiguous specification of BGP that causes the *bouncing problem*. BGP may examine all possible paths $O(n!)$ of all possible lengths among n autonomous systems.

A more realistic model of BGP that considers MRAI timer influence, different topologies, and policy settings is reported in [4]. The paper reports results on fail-over latency, convergence time, and path selection. Fail-over latency stems from temporary routing table oscillations. Convergence time of BGP is dominated by an implicit MRAI timer setting (30 seconds), the length of the longest alternative AS path between the source and destination, and the set of all possible paths between source and destination. Path selection depends not only on individual timer settings but also on the interaction of timers on multiple alternative paths in a topology. The authors present a plot that shows a linear dependence of average convergence latency and the longest AS path length. To improve convergence, they pose the interesting idea of introducing adaptive MRAI timers and adding routing information to the withdrawal messages.

B. Theoretical Work and Simulations

Griffin, et al. published several papers [8], [9] that analyze BGP as a distributed graph algorithm. BGP is not based on pure shortest path routing. For algorithms that find a shortest path between routers, it is easy to prove that the computation will stabilize in finite time to a state that will not change until the network itself changes. Shortest path protocols such as RIP converge in $O(n^3)$ [10] but these protocols do not allow for setting of policies. Since BGP allows network operators to prefer a path without considering its length (policy decisions), the proof of convergence fails. Looking closer into BGP routing, not only can the convergence not be proved but in practice changes in routing tables, going on for several minutes even if the topology itself does not change, can be observed. This implies that the routing policies of autonomous systems conflict. With a different setting of the policies, no oscillations would occur. The same topology may behave differently with different policy settings; it may converge but need not do so. This ambiguity stems from timing of routing message exchange and the particular policy settings used. A theoretical proof of BGP routing divergence was published in [11], where Varadhan, et al. show that independent route selection protocols like BGP may exhibit persistent oscillations and never converge to a single route due to conflicting policies.

Griffin presents a formal specification [12] of inter-domain routing protocol, which adds new attributes to BGP. The attributes are used to identify cycles in path histories. Once the oscillating paths are detected, they can be suppressed. In [9], the authors present a sufficient condition for a convergent routing system, and they propose BGP-like Simplified Path-Vector Protocol [12] based on it that is proven to be convergent.

A simulator described in [5] was used to verify optimal values for MRAI, and effects of Sender Side Loop Detection (SSLD) and Withdrawal Rate Limiting (WRATE) on convergence. Even though WRATE is in violation of the BGP standard RFC 1771 [1], Griffin [5] observes that it may have a positive effect. It seems that every topology has a specific value of MRAI that is optimal for convergence and communication complexity. Experiments with WRATE surprisingly show that setting up a timer for withdrawals may have a positive effect if MRAI is not optimized. SSLD has little influence once MRAI is optimized. Because it is not realistic to expect an optimal MRAI setting in real conditions, experiments with WRATE and SSLD are still valuable. For non-optimized MRAI settings and a clique topology, setting up WRATE substantially decreases convergence time and is more effective than SSLD.

III. SIMULATOR

Several network simulators already exist, such as ns-2, JavaSim, SSFNet and MaRS. With the exception of MaRS, these simulators have a model granularity that includes the IP protocol. The model used in ns-2 even comprises link layer protocols. Because we wanted to avoid interaction with underlying protocols and stay on a control plane in first stage of our work, we built our own simulator.

A. Overview of the Simulator

Our simulation model includes some fundamental mechanisms of the BGP protocol, but excludes underlying mechanisms such as TCP/IP. The discrete-event driven simulator, *RouteSim*, is implemented in Java and is composed of a simulation engine, model entities, and events. The entities model components such as routers, links, and messages. The events represent changes of the system state (such as message arrivals). The simulation engine keeps all pending events chronologically ordered and dispatches the next event. The entities, and thus the system model, are declared in a text file and loaded before the simulation commences. Events are either created implicitly by an entity at some system state change, or explicitly injected. Injected events are typically events from outside the simulation model, for example a change in the routing policy for a given router.

The system model forms a graph, describing a network topology of BGP *routers* as vertices and communication *links* between the routers as edges. A link entity models a full duplex point-to-point communication link, delivering *packet* entities between a pair of routers. The transmission latency does not depend on the packet's size or type. The value of the transmission latency and its uniformly distributed, random maximum deviation (jitter) is configurable on each link. Since the latencies and the jitter are bounded, packet delivery is guaranteed and no packet loss is modeled. The characteristics of a link entity are comparable to a reliable transport protocol.

A BGP router entity models a *BGP speaker* that exchanges External BGP (EBGP) routing information with peer routers in neighboring AS:s. Every router represents one particular AS, and Internal BGP (IBGP) is not modeled.

The simulation is performed on the *routing control-plane*, the only modeled communication is the exchange of routing information between peering routers. Forwarding of user traffic is not supported. But, the effect of user traffic can be modeled by adding link delay jitter to every packet.

The model only supports the advertisement of one single network destination from a single router in any given topology. Therefore there is no need to explicitly include any network addressing scheme (i.e., IP). The only addressing scheme present in the model is the enumeration of ASes with AS numbers (ASN). CPU workload latency for every router is modeled so that every packet received by a router is imposed with a uniformly distributed random latency before it is processed. The only BGP message type modeled is the Update type. A consequence of this is that peering between the routers is assumed to be in *established* state at all times during a simulation and no type of failure that would normally cause a Notify message and terminate the connection can occur. In accordance to the BGP specification, Update messages containing a route advertisement are temporally constrained by MRAI timers. Update messages containing route withdrawals are not. A router has one configurable timer per peer. A router can be configured to apply routing policies. These policies are defined by an enumerated set of preferred AS-paths which override shorter, unpreferred paths. A router can be instructed to advertise and later withdraw a route either by configuration in the system file, or by the injection of a BGP Announce Event.

IV. EXPERIMENTS AND RESULTS

When facing a withdrawal, BGP enumerates routes with increasing lengths, which results in an exchange of a large number of routing messages and delayed convergence time. The usage of MRAI timers play an important role in limiting excessive route enumeration, as it forces synchronization of path selection into "computational rounds" where the AS path length increases monotonically for every round. In a clique topology of n AS:s, the usage of MRAI timers imposes a convergence time with a linear lower bound of $(n - 3) * MRAI$ [2]. The bound holds for a routing system with synchronized MRAI timers, but not necessarily when the timers are skewed.

In our experiments, we applied different techniques that are supposed to improve the convergence. We studied their effect on clique topologies of different sizes. We tested three different scenarios: Basic, Split Horizon (SH), and Sender-Side Loop Detection (SSLD).

Basic: in line with the BGP specification, all routers perform receiver-side loop detection upon the receipt of update messages carrying route announcements. It is called Basic because routers do not apply any additional technique, not even Split Horizon.

Split Horizon: the routers do not announce a route back to the peer from which they learned it.

Sender-Side Loop Detection: before a route is sent, the routers perform loop detection on behalf of the receiving peer.

A. Simulation setup

We have performed 3000 simulations to study what effect MRAI timers have on the convergence complexity in a clique topology after a withdrawal. This was done by configuring a designated router to advertise, and later withdraw a route to the destination. The time between the advertisement and the withdrawal was kept long enough for the system to converge into an idle state where all routers had learned and selected a feasible route to the destination. During simulations, we set the CPU processing latency between 0.01 and 1.0 seconds. MRAI timer jitter was not used, and routing decisions were based on shortest-path route selection. The link latency was set to 0.01 seconds with no deviation. To our best knowledge, no known documentation about CPU workload latency in routers is available. However, the used values for CPU workload latency and link latency were considered plausible in [5].

The total number of update messages exchanged after the initial withdrawal was recorded. Also, the convergence time was measured from the initial withdrawal until no more update messages were exchanged. The routers did not apply any route filtering.

B. The length of MRAI

MRAI reduces the number of messages and the computational load on routers. On the other hand it delays end-to-end fail-over. Loop detection, as required by the BGP specification, is usually implemented by vendors on the receiver side. Labovitz shows that sender-side loop detection speeds up convergence time and suggests that loop detection should be implemented on both sides. Our experiments confirm that implementation of SSLD considerably improves the convergence time.

Fig. 1. Graphs showing average and maximum deviation for traffic complexity and convergence time as a function of the MRAI value. The MRAI timers are synchronized. The topology is a clique of size 15.

The experimental system consisted of 15 routers, forming a clique topology. SH or SSLD was not applied. To examine the impact of different MRAI timer values, 50 simulations with different random seeds were run for every MRAI integer value between 0 and 30 seconds. The graphs in Fig. 1 shows the result from the 1500 simulation runs. In Fig. 1(a) and 1(b), the symptom of excessive route enumeration for low values of MRAI is clearly shown in the form of large number of routing messages and long convergence delays. When the MRAI increases, this behavior is dampened as the MRAI timers impose synchronization of the route selections.

At a MRAI value of approximately 8 seconds, the system experiences the lowest number of routing messages and the shortest convergence time. At this point the MRAI is long enough to impose the route selection synchronization. For MRAI values larger than 8 seconds, the number of routing messages is constant (Fig. 1(c)), but the convergence time increases linearly (Fig. 1(d)). This follows from the fact that the MRAI timers impose excessive delay between the computational rounds in the route selection. Fig. 1(d) shows how the convergence time is confined by the lower bound, $(n - 3) * MRAI$. This result should however not be interpreted such that 8 seconds is the absolute optimal value for the MRAI, this is only true for this particular system, other configurations give different results [5].

C. Interaction of timers

In a real BGP routing network, it is unlikely that all MRAI timers are synchronized. The BGP protocol specification sug-

gests the usage of timer jitter to avoid peaks in the distribution of routing messages. We have performed simulations where we elaborated with skewed timers. Simulations were performed in the Basic scenario (without SH or SSLD), in a clique topology with 15 routers. The *length* of MRAI is kept constant at 30 seconds. Instead of using synchronized MRAI timers, we applied *skewed* timers after the initial withdrawal. In the BGP specification [1], 30 seconds is the recommended length for the MRAI.

The basic functionality of MRAI timer follows: In router i , an MRAI timer $T_{i,j}$ is used to impose the MRAI delay between advertisements to a peer router j . The timer keeps a record of the time of the last sent advertisement, $T_{i,j}.last$. An advertisement is postponed until the MRAI time has passed since the last advertisement, $T_{i,j}.delay = \max(0, MRAI - (now - T_{i,j}.last))$.

Randomized skewing is achieved by setting the timer offset like this: $T_{i,j}.last = now - U(0, MRAI)$ where $U(a, b)$ is an uniformly distributed pseudo-random number between a and b . Applying of randomized skewing gives more variable results than the synchronized timers because the order of events plays a more important role (messages are not synchronized in rounds). For *even skewing*, we apply the following timer offset: $T_{i,j}.last = now - \frac{i}{n} * MRAI$. Even skewing keeps a certain level of regularity in timing but still suppresses the effect of rounds.

The effect of randomly skewed timers is shown in Fig. 2(a) and 2(b). The corresponding characteristics for evenly skewed

timers are on Fig. 2(c) and 2(d). With evenly skewed timers we observe slower increase of convergence time with the growing value of MRAI, than in the case of synchronized or randomly skewed timers. Optimal value of MRAI for the convergence time is not influenced by skewing timers. The number of messages stays constant once MRAI reaches the optimal value. The number of announcements is lowest in the case of synchronized timers and largest for randomly skewed timers. The optimal value of MRAI for the messages is not influenced by skewing.

D. Experiments with SH and SSLD for different topologies

We tested effects of SH and SSLD for clique topologies with sizes ranging from 2 to 40. MRAI was set to 30 seconds, without jitter. After the initial withdrawal, all MRAI timers were skewed randomly.

Fig. 3(a) shows that the number of announcements grows as a polynomial function of the clique size in all three scenarios. The Basic scenario has the fastest growth, tightly followed by SH. SSLD has the slowest growth. For a clique with size 40, the Basic scenario exchanges approximately 15000 additional announcements compared to the SSLD scenario. Fig. 3(b) shows that the effect of SH and SSLD on the number of withdrawals was opposite: the number of withdrawals increased. It grew polynomially with the size of a clique in all three scenarios but with SSLD the growth is fastest. The growth of the total number of messages (announcements and withdrawals) is polynomial and the values are about the same in all three scenarios.

As shown in Fig. 3(c), the convergence time grows linearly with the respect to clique size in all three scenarios. SH and SSLD have shorter convergence time. For a clique of size 10, the Basic scenario converged after an average of 202 seconds, SH after 185 seconds, and SSLD after 124 seconds (see Fig. 3(d)).

V. CONCLUSIONS AND FUTURE WORK

Our model implements a simplified version of BGP. To verify the correctness of our simulator, we have run several tests and compared them to results obtained by others from BGP traffic measurements or simulations. The model exhibits properties corresponding to the behavior of Internet implementations of BGP. We ran experiments with the influence of timer setting and their mutual synchronization on the convergence time and number of update messages. Our simulator showed the characteristic predicted theoretically by Labovitz, while Griffin's simulation graph shows a diagram with a flatter slope for message growth with the increasing value of MRAI. The most significant difference between the simulations is the convergence time above the time-optimal value of MRAI.

Our work shows new results on the effects of timer skewing and the effects of split horizon and server-side loop detection. By simulation, we have shown that randomly skewed timers increase the variance in convergence time and number of messages. In addition, our data shows that the total number of Update messages is not greatly affected by the usage of split horizon or sender-side loop detection in a clique topology. Sender-side loop detection typically decreases the number

of announcement messages and increases the number of withdrawal messages for a route withdrawal. A single Update message can contain a set of route announcements together with a set of route withdrawals. Multiple route announcements can be sent in a single Update message as long as all the announced routes share the same AS path. If several diverse route destinations are converging to a withdrawal at the same time, it might be an advantage if much of the information exchange is in terms of withdrawals as these can more easily be sent in fewer Update messages.

Our plan for the future includes an extension of the simulation model with more BGP mechanisms such as route flap dampening and inbound/outbound route filtering. Furthermore, we intend to perform simulations with more realistic topologies and routing policies. We will study effects of router position in a topology (relative to the backbone) and its timer setting on convergence. We also intend to apply routing policies in our model and to test different mechanisms for suppressing oscillations, e.g. recording a path history.

REFERENCES

- [1] Y. Rekhter and T. Li. A Border Gateway Protocol 4 (BGP-4). Request For Comments RFC 1771, Internet Engineering Task Force, March 1995.
- [2] Craig Labovitz, Abha Ahuja, Abhijit Bose, and Farnam Jahanian. Delayed Internet Routing Convergence. In *Proc. of ACM SIGCOMM*, volume 30, pages 175–187, October 2000.
- [3] NANOG - The North American Network Operators' Group, <http://www.nanog.org/>.
- [4] Craig Labovitz, Abha Ahuja, Roger Wattenhofer, and Srinivasan Venkatasachary. The Impact of Internet Policy and Topology on Delayed Routing Convergence. In *Proc. of INFOCOM*, pages 537–546, IEEE, April 2001.
- [5] Timothy G. Griffin and Brian J. Premore. An Experimental Analysis of BGP Convergence Time. In *Proc. of ICNP*, November 2001.
- [6] Timothy G. Griffin and Brian J. Premore. Personal communication, April 2002.
- [7] Craig Labovitz, G. Robert Malan, and Farnam Jahanian. Origins of Internet Routing Instability. In *Proc. of INFOCOM*, IEEE, June 1999.
- [8] Timothy G. Griffin and Gordon Wilfong. An Analysis of BGP Convergence Properties. In *Proc. of ACM SIGCOMM*, pages 277–288, September 1999.
- [9] Timothy G. Griffin, F. Bruce Shepherd, and Gordon Wilfong. Policy Disputes in Path Vector Protocols. In *Proc. of ICNP*, November 1999.
- [10] Karthikeyan Bhargavan, Davor Obradovic, and Carl A. Gunter. Formal Verification of Distance-Vector Routing Protocols. In *International Conference on Theorem Proving and Higher Order Logics*, August 2000.
- [11] Kannan Varadhan, Ramesh Govindan, and Deborah Estrin. Persistent Route Oscillations in Inter-Domain Routing. *Computer Networks*, 32(1):1–16, 2000.
- [12] Timothy G. Griffin and Gordon Wilfong. A Safe Path Vector Protocol. In *Proc. of INFOCOM*, pages 490–499, IEEE, March 2000.

Fig. 2. Graphs showing average and maximum deviation for traffic complexity and convergence time as a function of the MRAI value. The MRAI timers are skewed randomly in (a), (b), and evenly in (c), (d). The topology is a clique of size 15.

Fig. 3. Graphs showing the number of messages and convergence time as a function of the clique size. The MRAI value is set to 30 seconds, and the MRAI timers are randomly skewed.