Simulating Routing Models on Large-Scale Topologies

David Coudert*, Luc Hogie*, Aurélien Lancin*, Dimitri Papadimitriou†, Issam Tahiri*
*INRIA, I3S(CNRS, UNS), 2004 route des Lucioles, 06902 Sophia Antipolis, France
Email: (david.coudert,luc.hogie,aurelien.lancin,issam.tahiri)@inria.fr
†Alcatel-Lucent Bell, Copernicuslaan 50, 2018 Antwerpen, Belgium
Email: dimitri.papadimitriou@alcatel-lucent.com

***Abstract*—The expansion of the Internet routing system resulting from the 10% growth per year of the Autonomous System (AS) topology size (40k AS as of 2012) results in a number of research challenges. Indeed, the Border Gateway Protocol (BGP) starts to show its limits in terms of the number of routing table entries it can dynamically process and control. Dynamic routing protocols showing better scaling properties are thus under investigation. Because deploying newly designed routing protocols on the Internet is not practicable at a large-scale, simulation is an unavoidable step to validate their properties. However, the increasing routing information processing (CPU) and storage (memory) introduces similar challenges for the simulation of state-full routing protocols on large-scale topologies (comprising tens of thousands of nodes). This paper presents the Dynamic Routing Model simulator DRMSim which addresses the specific problem of large-scale simulations of routing models on large networks. The motivation for developing a new simulator lies in the limitation of existing simulation tools in terms of the number of nodes they can handle and in the models they propose.**

I. Introduction

Resulting from its expansion, the Internet routing system, which is based on the Border Gateway Protocol (BGP) [18], needs to accommodate an increasing number of Internet Protocol (IP) routes and an increasing number of Autonomous Systems (AS), each defined as a set of routers under a single technical administration. This situation is exacerbated by i) site multi-homing and AS multi-homing (resulting into an increased meshedness) as well as inter-domain traffic engineering (by means of prefix de-aggregation) and ii) the demand for improving IP connectivity availability from an increasing number of connected hosts. According to the available BGP reports [2], in January 2007, the number of active BGP entries was about 200k and in January 2008 about 250k. Mid-2009, this number reached 300k and 400k early 2012. The currently observed growth rate of active routing table entries ranges between 1.2-1.3 per year. Depending on the extrapolation model, by early 2014, the number of active routing entries of a core router would reach about 500k. Worst-case projections predict that routing engines could have to process and maintain of the order of 1M active routes within the next 5 years. Note that the actual number of BGP routing table entries is higher and depends on the routing/forwarding table ratio (that varies between 2 and a low order of 10). The number of allocated AS numbers is steadily increasing and the number of advertised AS reached about 32k at the end of the third quarter of 2009. As of early 2012, the number of advertised AS confirms the 10% growth rate per year by reaching 41k. As a result of storing an increasingly larger number of (network) states in the routing system, the latter becomes increasingly expensive and places undue cost burdens on network administrative units that do not necessarily benefit from Routing Table (RT) size increases.

Moreover, the impact on the BGP routing protocol dynamics (robustness/stability and convergence) resulting from i) inconsistencies (due to software implementation errors, router misconfigurations, etc.), ii) instabilities (due to routing policies interactions), and iii) topological changes/failures is progressively becoming a key concern for the global Internet community. Indeed, the scalability properties of inter-domain routing do not only depend on the routing algorithm used to compute/select the paths but also on the number of inter-domain routing messages (also known as routing updates) exchanged between routers. Between January 2006 and January 2009, the prefix update and withdrawal rates per day have increased by approximately 2.25 to 2.5 [2]. Routing updates require processing, and result in routing table re-computation/re-selection that, in turn, delay routing tables convergence. As observed, e.g., by [11], uninformed path exploration that characterizes (AS-)path vector routing such as BGP amplifies convergence delay. In this context, a fundamental dimension to take into account is the dynamics of the routing information exchanges between routers (in particular, the routing topology updates that dynamically react to topological structure changes). The Internet routing system architecture is thus facing performance challenges in terms of scalability as well as dynamic properties (convergence, and stability/robustness) that result into major cost concerns for network designers but also routing protocol designers.

Hence, the Internet routing system is facing performance limitations in terms of scalability (resulting from the growth rate of the number of active BGP routing table entries) as well as in terms of dynamics (resulting from the architectural properties of BGP). Both limitations lead to major concerns for both network and system designers. Therefore, new routing schemes have been recently proposed, such as *compact routing* [1], [20] and *geometric routing* [15], [19]. A compact routing scheme decreases the size of the routing tables by omitting some network topology details such that resulting path length increase stays relatively small. A geometric routing operates by assigning to nodes (virtual) coordinates in a metric space used as addresses to perform point-to-point routing in this space. By routing scheme, we refer to a set/class of routing models that are based on the same principles and thereby sharing the same essential and global characteristics as well as structuring/cohesive elements. On the other hand, a routing algorithm is defined as a distributed algorithm that, for any node's network attachment identifier (e.g., IP address), computes and/or selects a loop-free routing path so that incoming messages (e.g., IP datagrams) directed to a given destination can reach it.

Because deploying newly designed routing protocols on the Internet is not practicable at a large-scale, simulation is an unavoidable step to validate their behavioral and performance properties. Unfortunately, the simulation of inter-domain routing protocols over large-scale networks (comprising tens of thousands of nodes) becomes a real issue [23] due to the increasing routing information processing (CPU) and storage (memory) they require. To the best of our knowledge, no simulator provides the capability to investigate in-

depth the behavior and performance properties of routing schemes when applied to large-scale topologies (comprising tens of thousands of nodes).

This paper presents the Dynamic Routing Model simulator DRMSim which addresses the specific problem of large-scale simulations of (inter-domain) routing models on large networks. The motivation for developing a new simulator lies in the limitation of existing simulation tools in terms of the number of nodes they can handle and in the models they propose.

II. State of the art

We have to distinguish three classes of simulators when it comes to routing: (routing) protocol simulators, routing configuration simulators, and (routing) model simulators.

Simulators dedicated to the performance measurement and analysis of the routing protocol (procedures and format) at the microscopic level. These can be further subdivided between simulators specialized for BGP protocol specifics, simulators dedicated to routing protocols and general protocol simulators. The ns [14] discrete-event simulator that relies on the BGP daemon from Zebra [24] belongs to the second sub-category. This daemon can be used to build realistic inter-domain routing scenarios but not on large-scale networks due to the low level execution of the protocol procedures. On the other hand, the SSFNet [22] discrete event simulator, relies on the implementation of the BGP protocol that was tailored and validated for the needs of a BGP-specific simulators. In SSFNet, a simulated router running BGP maintains its own forwarding table. It is thus possible to perform simulation with both TCP/IP traffic and routing protocols to evaluate the impact of a change in routing on the performance of TCP as seen by the end systems (hosts, terminals, etc.).

Simulators dedicated to simulation of BGP protocol operations including the computation of the outcome of the BGP route selection process by taking into account the routers' configuration, the externally received BGP routing information and the network topology but without any time dynamics. These simulators can be used by researchers and ISP network operators to evaluate the impact of modified decision processes, additional BGP route attributes, as well as logical and topological changes on the routing tables computed on individual routers assuming that each event can be entirely characterized. Topological changes usually comprise predetermined links and routers failures whereas logical changes include changes in the configuration of the routers such as input/output routing policies or IGP link weights. These simulators are thus specialized and optimized (in terms, e.g., of data structures and procedures) to execute BGP on large topologies with sizes of the same order of magnitude than the Internet since these simulators are not designed to support real-time execution. These simulators usually support complete BGP decision process, import and export filters, route-reflectors, processing of AS_path attributes and even custom route selection rules for traffic engineering purposes, and BGP policies. Simulators like SimBGP [21] or C-BGP [17] belong to this category. These simulators are gradually updated to incorporate new BGP features but are complex to extend out of the context of BGP.

Simulators dedicated to the simulation of routing models, category to which DRMSim [6] belongs. Designed for the investigation of the performance of dynamic routing models on large-scale networks, these simulators allow execution of different routing models and enable comparison of their resulting performance. Simulators in this category consider models instead of protocols, meaning they do not execute the low level procedures of the protocol that process exact protocol formats but their abstraction. Thus these simulators require specification of an abstract procedural model, data model, and state model sufficiently simple to be effective on large-scale networks but still representative of the actual protocol execution. However, incorporating (and maintaining up to date) routing state information is also becoming technically challenging because of the amount of memory required to store such data. In practice, processing of individual routing states impedes the execution of large-scale simulations. DRMSim addresses this issue by means of efficient graph-based data structures. Moreover, by using advanced data structures to represent routing tables, DRMSim can still run simulation whose number of nodes exceeds ten thousands.

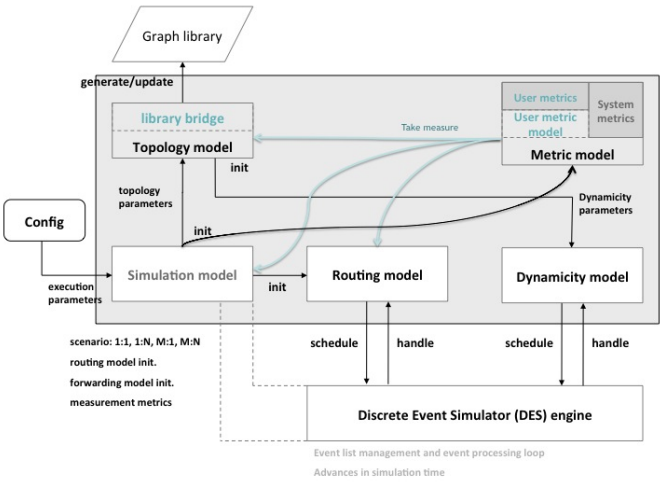

Fig. 1. DRMSim architecture

All simulators previously cited here above share many properties. Like DRMSim, they all rely on Discrete-Event Simulation (DES). However, on one hand, BGP simulators, in order to keep an acceptable level of performance, optimize their procedures and data structures for BGP protocol executions; thus, they can not be easily extended to accommodate other routing protocol models. On the other hand, general routing protocol simulators designed to investigate the effects of routing protocol dynamics are usually limited to networks of few hundred nodes; thus, preventing large-scale simulations of state-full routing protocols over networks comprising of the order of ten thousands nodes.

III. Simulator

To measure BGP performances, we rely on the DRMSim [6], a JAVA-based software providing a routing model simulator. DRMSim enables the construction of routing tables by means of routing path computation and/or selection procedures and, in turn, the evaluation of the behavior and performances of various distributed routing models. The main performance metrics supported include the stretch of routing paths produced, the size of routing tables, the number of messages, and the adaptivity to topological modifications.

A. DRMSim architecture

DRMSim implements the Discrete-Event Simulation (DES) approach. In DES, the operation of a system is represented as a chronological sequence of events (associated with any change in the state of the system). A DES typically implements three data structures: a set of state variables, an event list, and a global clock variable (that denotes the instant on the simulation time axis at which the simulation resides). In DRMSim, an event is a data structure comprising the event's timestamp, the event type, and the event code (a routine which implements what the event consists of). DRMSim comprises a simulation model, a system model, a dynamics model, a metric model and a set of routing models. Figure 1 details the architecture and relationships between these models.

1) *Simulation model*: initializes the system model, the metric model and the routing model. It also defines the simulation scenario. A scenario could be, for example, the simulation of

BGP until convergence upon failure of a set of routers or links initiated at specific times during the simulation.

2) *System model*: controls the network topology. It relies on a graph library to create the network topology, to compute information - like the shortest path matrix - and to perform structural modifications. To avoid dependence on a single graph library and to allow the routing model designer to choose its own graph library, the topology model uses graph library bridges. For each different graph library, a specific bridge must be developed and integrated to DRMSim. If the graph library allows graph partitioning, the system model keeps information about node/link's partition.
3) *Dynamics model*: performs maintenance operations on the network infrastructure as well as router failures. It schedules at a given time - according to the simulation scenario - dynamics events which are router or link failure/repair.
4) *Routing models*: each model comprises the routing procedure(s), the data model and the communication model to be simulated. DRMSim proposes a set of basic routing models (source routing, random schemes, broadcasting, etc). These models allow to verify the correctness of the simulation engine and serve as reference to compare performance with respect to advanced routing protocols. The set of models currently provided by DRMSim includes the Border Gateway Protocol (BGP) [18], the Routing Information Protocol (RIP), and compact routing schemes such as NSR [13] and AGMNT [1].
5) *Metric model*: listens to the simulation and topology models. It allows to monitor a selected set of routers/links. This model has been also extended to support measure in case of partitioned network on boundary routers/links and partitions. The memory and CPU usage mainly depend on the metrics, on the set of routers/links onto which they are applied, on the measurement interval, and their respective computational complexity. This dependence can leads to extensive use of memory/CPU. To simplify the development of new specific metrics, the metric model is composed of a *system metric model* and a *user metric model*. The former defines a set of default performance related metrics, including the additive routing stretch, the multiplicative routing stretch, the number of routing table entries, and the size of routing tables. The latter provides to the routing model designer, an API to extend the system metric model to perform routing model-specific measures.

B. Border Gateway Protocol (BGP)

The Border Gateway Protocol (BGP) is the inter-Autonomous System routing protocol of the Internet. The primary function of a BGP speaking routers is to exchange routing information (including network reachability information, and list of Autonomous Systems (ASes) that reachability information traverses). This information is sufficient for constructing a graph of AS connectivity for this reachability from which routing loops may be pruned, and, at the AS level, some policy decisions may be enforced. Routing information exchanged by means of BGP supports only the destination-based forwarding paradigm, which assumes that intermediate routers forward a packet based solely on the destination address.

1) Model description: the network model used in DRMSim considers AS-level topology, meaning that every node represents an Autonomous System (AS). This implies that DRMSim focuses on inter-domain routing, which is implemented by External BGP (eBGP), i.e., BGP sessions are established between routers belonging to different ASes. Note that eBGP together with Internal BGP (iBGP), i.e., BGP sessions are established between routers belonging to the same AS, constitute the core of the BGP protocol. A router running BGP segments its Routing Information Bases (RIB) into three logical structures. First, the Loc-RIB which contains all the routes (i.e., a destination prefix, an AS-Path and its associated set of attributes) locally selected following the rules of the node-based decision process; these routes are those populating the local routing table and subsequently used by the forwarding process. Second, the Adj-RIB-In and the Adj-RIB-Out enable the router to provide a neighbor-based filtering for, respectively, incoming and outgoing advertised routes. To simplify processing at each node, a single RIB, the Loc-RIB, is implemented in DRMSim. The Loc-RIB stores routes by taking into consideration the most important attributes (e.g., the AS-path) per destination prefix while leaving the flexibility for adding new attributes. DRMSim features three implementations of the BGP routing protocol. The first one implements the full BGP state machine. However, because of the large amount of computational resources required to simulate the complete procedures as specified by BGP, we decided to implement several optimizations.

2) Optimizations: in order to reduce the computational resources required for the simulation of BGP, DRMSim implements the following enhancements:

- *Event reduction*: reduction of the number of events by assuming that each BGP session has only two possible states: IDLE or ESTABLISHED. This reduction impacts the establishment time of the BGP sessions. In term of performance, the initialization phase of BGP sessions will thus complete faster.
- *Data structures*: when modeling a router, two main data structures have to be considered. First, a routing table which contains all the computed/selected routes derived from the routing information received from its peers. This table is usually implemented in software. Secondly, a router contains also forwarding table which only stores the necessary information to forward packets; it is usually implemented in hardware and, therefore, makes the forwarding process very fast. When simulating routing models, both data structures are coded in software. In order to compare the efficiency of maintaining both data structures or only the routing table, we performed the same simulations using both approaches. We found that maintaining only the routing table data structure and using the "compute on demand" method for the forwarding table entries was the best solution.
- *Database lookups*: code profiling showed that database lookup operations took the largest part of the simulation execution time. Therefore, we investigated many alternatives to overcome this problem. The best solution we found was to assign to each router an identifier from 1 to n (where n is the number of routers in the network) and to index the routing table entries accordingly. The index value for a given routing table entry is the identifier of the destination corresponding to that entry.
- *Update processing*: for every router, we include a bit-vector whose size is the number of routers in our network, and for which the bit at the i^{th} position indicates whether this router has or not a route for the destination with the identifier i. Using this information, efficient logical operations on bit-vector pairs (each composed by the local and the peering router bit- vector) can be performed to determine the useful/useless entries of an update message exchanged between these two routers.

3) MRAI Impact on BGP Convergence Time: the dynamics and convergence properties of BGP play an important role in determining network performance as BGP (indirectly) controls the forwarding of inter-domain traffic. Proeminent studies have shown that upon occurrence of a node failure, the re-convergence of routing states can take on the order of 3 to 15 minutes [11]. In a fully connected network, [11] demonstrated that the lower bound on BGP convergence time is given by $(N-3)\cdot MRAI$, where N is the number of AS in the network, and $MRAI$ is the MinRouteAdvertisementInterval (MRAI) time [18]. The MRAI time, by default set to 30 seconds on eBGP sessions, determines the minimum amount of time that must elapse between an advertisement and/or

withdrawal of routes to a given destination by a BGP speaker to a peer. Thus, two BGP update messages sent to a given peer by a BGP speaker (that advertises feasible routes and/or withdrawal of infeasible routes to some common set of destinations) are separated by at least one MRAI. This rate limiting mechanism, applied on a per-destination prefix basis, results in suppressing the advertisement of successive updates to a peer for a given prefix until the MRAI timer expires (as it is intended to prevent exchange of transient states between BGP routers). However, the MRAI-based rate limitation results also in routing state coupling between topologically correlated BGP updates for the same destination prefix: the MRAI introduces time synchronization. As a consequence, even if one may think that decreasing the MRAI value would result in decreasing the convergence time, in practice, decreasing the MRAI time value below a certain threshold leads to adversary effects in terms of number of BGP updates (communication cost) and BGP convergence time [16].

4) BGP Metrics: to enable computation of the communication cost, we have extended the DRMSim metric model in order to measure the number of BGP update messages, the number of entries they comprise and their size during the total execution time of a simulation but also per router/link.

IV. DRMSim Software

The DRMSim sofware is the result of a joint research project jointly conducted by Alcatel-Lucent Bell, Universite de Bordeaux (LaBri) and INRIA Sophia Antipolis (Mascotte project). It is now conducted and funded by the European Commission under the EULER STREP project (Grant No.258307) part of the Future Internet Research and Experimentation (FIRE) objective of the Seventh Framework Programme (FP7).

The DRMSim software, distributed under version 3 of the GNU General Public License (GPLv3) - see http://www.gnu.org/licenses/gpl.html, is available at:

- DRMSim website: http://drmsim.gforge.inria.fr
- Gforge website: https://gforge.inria.fr/projects/drmsim
- SVN repository: https://scm.gforge.inria.fr/svn/drmsim

The DRMSim software manual which describes its installation procedure, and its configuration for routing model simulation, is available at https://gforge.inria.fr/docman/view.php/2807/7719/UserManual.pdf.

V. Conclusion

The expansion of the Internet results in a number of challenges at the routing system level: the Border Gateway Protocol (BGP) starts to show its limits in terms of the number of routing tables entries it can dynamically process and control with satisfying performance and stability. More scalable routing protocols have to be proposed that overcome these limitations. Because experimenting newly designed routing protocols directly on the Internet is not practicable (partly due to the size of the Internet topology but also for obvious operational reasons), research and development have to make use of large-scale simulation.

This paper presents DRMSim, a simulator enabling the simulation of large-scale simulations of routing protocols. The motivation for developing a new simulator lies in the limitation of existing simulation tools in terms of the number of nodes they can handle but also in the routing models they can execute. For this purpose, DRMSim proposes a general routing model which accommodates any network configuration. Aside to this, it includes specific models for GLP [3], and K-chordal network topologies, as well as implementations of routing protocols, including the NSR routing protocol and lightweight versions of BGP. Substantial development work has been already performed to realize a first release of the DRMSim software. The validation of the DRMSim BGP implementation has been performed by means of systematic verification experiments. Several BGP New features are being developed and will be incorporated into the simulator. In particular, to address the challenge of simulation of larger networks when running BGP, the next step is to enhance the code as well as to go further with parallel/distributed simulation [8]. Indeed, as documented in [4], moving the DRMSim routing model simulator to Distributed Parallel Discrete Event seems to provide a promising technique in order to make abstraction of the size of the topologies provided the induced communication overhead between partitions remains acceptable.

References

[1] I. Abraham, C. Gavoille, D. Malkhi, N. Nisan, and M. Thorup, *Compact name-independent routing with minimum stretch*, ACM Transactions on Algorithms, vol.4, no.3, Jun.2008.

[2] BGP Reports. http://bgp.potaroo.net/index-bgp.html.

[3] T. Bu, and T. Don, *On distinguishing between Internet power law topology generators*, Proc. 21th Annual IEEE International Conference on Computer Communications (INFOCOM), vol.2, 2002.

[4] D. Coudert, L. Hogie, A. Lancin, D. Papadimitriouy, S. Perennes, I. Tahiri, *Feasibility study on distributed simulations of BGP*, To be published, 26th ACM/IEEE/SCS Workshop on Principles of Advanced and Distributed Simulation, Zhangjiajie, China, Jul.2012.

[5] The Cooperative Association for Internet Data Analysis (CAIDA), *Archipelago (Ark)*. http://www.caida.org/projects/ark/.

[6] *DRMSim simulator*. http://drmsim.gforge.inria.fr.

[7] A. Elmokashfi, A. Kvalbein, and C. Dovrolis, *SIMROT: A Scalable Inter-domain Routing Toolbox*, SIGMETRICS Perform. Eval. Rev., 2011.

[8] R. Fujimoto, *Parallel Discrete Event Simulation*, Communications of the ACM, 1990.

[9] L. Hogie, D. Papadimitriou, I. Tahiri, and F. Majorczyk, *Simulating routing schemes on large-scale topologies*, Proc. 24th ACM/IEEE/SCS Workshop on Principles of Advanced and Distributed Simulation (PADS), Atlanta (GA), United States, May.2010.

[10] G. Huston, M. Rossi, and G. Armitage, *A Technique for Reducing BGP Update Announcements through Path Exploration Damping*, IEEE Journal on Selected Areas in Communications (JSAC), vol.28, no.8, Oct.2010.

[11] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, *Delayed Internet Routing Convergence*, Proc. ACM SIGCOMM 2000 Conference, pp.175–187, Stockholm, Sweden, Sep.2000.

[12] *METIS - Serial Graph Partitioning and Fill-reducing Matrix Ordering*. http://glaros.dtc.umn.edu/gkhome/metis/metis/overview.

[13] N. Nisse, K. Suchan, and I. Rapaport, *Distributed computing of efficient routing schemes in generalized chordal graphs*, Proc. International Colloquium on Structural Information and Communication Complexity (SIROCCO), 2009, Lecture Notes in Computer Science (LNCS) 5869, pp.252–265, 2010.

[14] *The Network Simulator - ns-2*. http://www.isi.edu/nsnam/ns/.

[15] C.H. Papadimitriou, D. Ratajczak, *On a Conjecture Related to Geometric Routing*, Theoretical Computer Science, vol.344, no.1, pp.3–14, 2005.

[16] B.J. Premore and T.G. Griffin, *An Experimental Analysis of BGP Convergence Time*, Proc. 9th IEEE International Conference on Network Protocol (ICNP'01), pp.53–61, Riverside (CA), USA, Nov.2001.

[17] B. Quoitin, S. Uhlig, *Modeling the routing of an autonomous system with C-BGP*, IEEE Networks, vol.19, no.6, pp.12–19, Nov.2005.

[18] Y. Rekhter, T. Li, S. Hares (Ed's), *A Border Gateway Protocol 4 (BGP-4)*, RFC 4271, Internet Engineering Task Force (IETF), Jan.2006.

[19] M.A. Serrano, D. Krioukov, M. Boguna, *Self-Similarity of Complex Networks and Hidden Metric Spaces*, Phys.Rev.Lett., vol.100, no.7, 2008.

[20] M. Thorup and U. Zwick, *Compact routing schemes*, Proc. 13th Annual ACM Symposium on Parallel Algorithms and Architectures (SPAA'01), pp.1–10, Heraklion (Crete), Greece, Jul.2001.

[21] *SimBGP, a simple BGP simulator*. http://www.bgpvista.com/simbgp.php.

[22] *Scalable Simulation Framework (SSF)*. http://www.ssfnet.org/

[23] G. Yaun, et al., *Large scale network simulation techniques: Examples of TCP and OSPF models*, ACM SIGCOMM Computer Communication Review (CCR), 2003.

[24] *Zebra Routing software* (Distributed under GNU GPL). http://www.zebra.org/