Realistic Topology Modeling for the Internet BGP Infrastructure

Veena Raghavan
Department of ECE
Georgia Institute of Technology
raghavan@ece.gatech.edu

George Riley
Department of ECE
Georgia Institute of Technology
riley@ece.gatech.edu

Talal Jaafar
Department of ECE
Georgia Institute of Technology
jaafar@ece.gatech.edu

Abstract

*The complexity and dynamics of Border Gateway Protocol (BGP), the only inter-domain routing protocol available for the Internet, is driving the need for efficient, scalable, realistic and meaningful network simulations. To properly simulate the behavior of BGP as currently deployed, one requires both a realistic topology and a realistic model of BGP. Although a large number of topology generators are available, they have limitations in either their scalability or in the ability to model AS relationships. In this work, we describe the methodology employed to efficiently construct meaningful large scale simulations of the order of a several thousand autonomous systems using parallel and distributed simulations thereby leveraging the power of multiprocessors and clusters of workstations. Using the Georgia Tech Network Simulator (*GTNetS *as the framework, we simulate a realistic topology constructed from the BGP routing table updates collected by RouteViews and further incorporate the policy constraints based on the autonomous systems relationships inferred by the Cooperative Association for Internet Data Analysis (CAIDA). Our BGP model (called BGP++) found in GTNetS is built from the open source Zebra BGP implementation, which is in deployed in a number of existing Internet Autonomous Systems. Using BGP++ and our topology generation methods, we show how the simulations can be used to study the effect of using BGP anycasting at DNS root servers on the several BGP metrics, including convergence time and churn.*

1 Introduction

BGP is the de-facto inter-domain routing protocol of the Internet. In spite of being the dominant protocol used for communication between inter Autonomous Systems (ASs), there has been a significant concern among both Internet operators and researchers about the long term viability of BGP in the Internet. The analysis of BGP is complex and if further complicated by the irregular growth of the Internet. In order to study the performance of BGP in current and future networks, simulation methods are often employed. It is possible to construct a detailed simulation of BGP performance on any arbitrary topology and compare the BGP metrics as the topology size grows. However, most existing simulation studies of BGP fail to account for the inherent AS relationships in the deployed Internet, which are crucial in analyzing the actual behavior of BGP. Further, in many prior studies, the topology used for the study is generated by synthetic topology generation tools. The availability of large repositories of BGP update information, such as that found in RouteViews [8], leads to the ability to construct more representative topologies, resulting in more realistic simulations. This work elaborates on techniques employed in building a realistic topology based on measured BGP update messages, inferred heuristics about AS relationships, and using a realistic model of BGP. The discussed methodology makes

the following contributions to the BGP research community:

1. A generic framework for constructing large scale simulations of BGP.
2. A method of partitioning the network structure to enable distributed simulation.
3. A survey of challenges that are commonly faced in constructing real topologies.
4. A framework that network operators can use to understand the dynamics of BGP, which can help them understand the effect of various configuration parameters on overall performance.

The simulations are constructed and evaluated using BGP++ [3], the BGP model in GTNetS. This model was created by porting the Zebra BGP daemon (bgpd) into GTNetS. The model is a fairly complete implementation of BGP, and has several capabilities and extensions including AS confederations and route reflections.

The remainder of this paper is organized as follows: Section 2 describes in detail our method of constructing realistic and meaningful topologies of the Internet and the memory usage and execution time statistics. Section 3 elaborates on the BGP Anycast Routing Simulation Analysis. Section 4 discusses the limitations of our model. Section 5 describes related work. Finally, Section 6 discusses future research work.

2 Simulation environment

In this section, we describe the simulation environment used to perform our experiments, including the method employed to construct a topology that closely mimics the Internet's BGP infrastructure and the use of parallel and distributed simulations to execute the simulation. This work extends prior work on BGP modeling by Dimitropoulos and Riley [4].

2.1 Simulation of a realistic topology

The reliability and accuracy of simulation results depends on the realism of the underlying models and topologies used in the simulation. In particular, the models should include:

1. Internet topology models at the router and AS levels.
2. The policy relationships between different ASs.
3. Accurate models for BGP behavior with various configuration parameters.

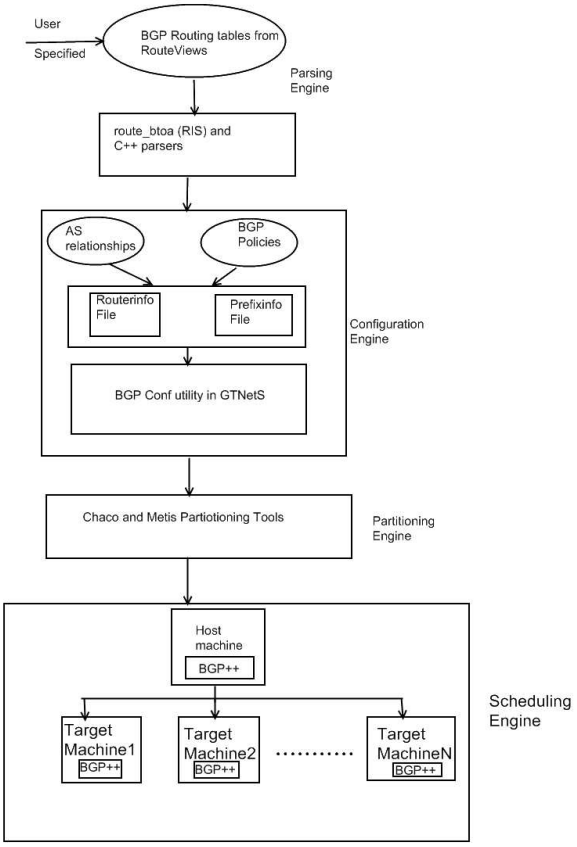

Figure 1. Conceptual model of the simulation toolset in BGP++

Researchers have created many monitoring projects such as the RouteViews project [8]. Routeviews collects various information from existing deployed BGP routers belonging to some of the largest service providers of the Internet. This data provides a fairly accurate representation of the connectivity in the Internet. We constructed our simulations using this information, along with the BGP model in GTNets. Figure 1 shows a conceptual model of the simulation toolset used in setting up the simulation environment. The conceptual model consists of four main components: the parser, the configuration and partitioning engines and the scheduler. Each of these components is explained in detail below.

The parsing component of the model uses the data collected by the routers as seen by RouteViews. In all our simulation studies, we have used the routing table data collected in the first week of January 2008, thus creating a model that closely resembles the current connectivity of the Internet architecture. This data is available in the raw Multi-threaded Root Toolkit (MRT) format which we convert into the ASCII format using the Routing Information Systems (RIS) tool, `route_btoa`. Using parsers we wrote in C++, fields such as the AS node number, its neighbors, its relationships with other ASs, and the different prefixes generated by it are extracted. The extracted data is then used to build the input files which are the router and prefix information files annotated with AS relationships. These are then used as input to the configuration engine as seen in the figure above.

The configuration engine uses the `BGPconf` utility included in GTNetS to parse the AS map to build actual configuration files for each one of the BGP routers considered in the topology. To make our simulations meaningful, we use the AS relationships as inferred using the heuristics described by Gao [6]. Filters and BGP communities are used to translate these relationships into configuration entries. At this stage, features such as logging updates and memory saving schemes may be turned on while building the configuration files for the routers. The BGP model in GTNetS provides the flexibility of specifying the topology size, route update interval, tracing and memory saving schemes in addition to the options available to set the propagation delay and bandwidth of the links which are provided by the general GTNetS model. The setup we used, to construct our simulations uses the tier1 and tier2 topology data that we obtained after parsing the data obtained from RouteViews. A total of 5476 configuration files are generated at this stage, one for each of the BGP routers in the topology we modeled.

In order to simulation an AS topology of this size, we utilized parallel and distributed simulation methods as described by Riley and Jaafar [13], which utilizes *Ghost Nodes* to facilitate inter–simulator routing without excessive memory usage. The space–parallel partitioning for the distributed simulations is done by the partitioner. To effectively load balance a network, the Chaco and METIS graph partitioning packages [7] are used. These packages provide substantial control to the user in terms of the number of target machines and the different graph partitioning algorithms that may be used to partition the overall topology. Using the output of the graph partitioning utilities and the partial simulation script generated by the BGP conf utility in GTNetS, a final simulation script is created. This is then augmented to include DNS query rates for the end applications, background traffic and other information that the user specifies. In our experiments, the topology was divided using the multilevel graph partitioning algorithm available in Chaco to divide the topology of the Internet into 16 graphs. The simulation script modeled each one of the ASs as nodes and the links connecting these ASs had a propagation delay of 10 ms and a link bandwidth of 10Mbps. While these link delay and bandwidth assumptions are clearly not accurate, we used the same values throughout the comparative experiments.

The final component to the model is the execution engine which uses the simulation script written in the previous stage, to assign the execution to various target machines depending on the available memory and other important criteria. This consists of a set of scripts that distribute the various processes to a set of target platforms specified by the user.

In the above model, by changing just the routing update table file obtained from RouteViews, any BGP simulation may be constructed and different parameters of interest may be studied which is shown in Section 4. Thus, we provide a framework for studying BGP simulations.

2.2 Memory usage and execution time statistics

One of the main problems often faced when running large–scale simulations is the excessive memory consumption and the extremely long execution times taken. Memory consumption depends on the size of the topology, the number of neighbors each router has, and the size of the routing table at each router among others. The GTNetS BGP++ implementation uses a method where the AdjRIBIn, LocalRIB and AdjRIBOut share the same memory where possible. Further, it utilized the memory-efficient NIx–Vector routing method [12]. With these optimizations, the memory consumption of the entire simulation is manageable. The memory usage for all the routers used in the distributed simulation of the topology consisting of 5476 ASs is shown in Table 1.

Long execution times are almost always a significant concern when running large–scale simulations. In a re-

Table 1. Memory usage statistics

Total memory usage for all the routers	1.6GB

Table 2. Execution Time statistics

Initialization Time(in seconds)	1200
Total Execution Time(in seconds)	4894

cent work, Nicol [10] reported that it takes a substantial amount of time, some times many hours for even moderate scale simulations. This is partially due to the length of time to perform the initial configuration setup for the simulation . Using techniques designed by Dimitropoulos [4], we save the state of the simulation after initial configuration setup, and later reuse the saved information as needed. In our simulation study of 5476 nodes, the execution time observed to run the entire parallel distributed simulation is very moderate, as seen in Table 2.

3 BGP anycast routing simulation analysis

In this section, we describe our motivation to perform an analysis of the performance of BGP when the *Anycast* approach is used to replicate the DNS root servers. Out anycast simulation environment includes providers for the C-,F-,K-,M- and J-Root servers. We also describe the various simulations performed to evaluate the effect of Anycast on DNS and BGP performance. Finally, we describe some interesting evaluations we have incorporated in our simulation studies.

3.1 Motivation

While certainly the long term viability of BGP as the Internet routing infrastructure is of critical importance, the future of the Internet is equally dependent on the performance of the Domain Name System (DNS) [9]. DNS is a hierarchical database mapping names and addresses on the Internet. The root of the hierarchy tree is referred to as a DNS *root server*. Currently, there are 13 anycast root servers deployed worldwide. To ensure high availability of the DNS service, some of the root servers are replicated or mirrored in various locations. This is achieved through *IP anycasting* [11]. The main goals of IP anycasting are to facilitate robust distributed system operation, ensure availability, and to reduce latency perceived by the user.

Since the deployment of IP Anycasting, a number of studies have been conducted to characterize the advantages of IP Anycasting the IP prefix of Root servers [1, 14]. The studies have shown that the availability of the Anycast prefixes is improved, and the end-to-end latency as perceived by the user is decreased. However these studies have shown a few failures which can make the Anycast prefixes unavailable for several minutes. In these studies, the increased down time is attributed to slow BGP convergence time.

The studies mentioned above are among the detailed analysis case studies which were performed on Anycasting. They provide a wide range of results about real-time behavior of the system. However, these studies are limited to data extracted through probes from a few locations. It is also important to be able to evaluate the performance of Anycast under certain failure cases which may not occur during the observation period of study. We use our simulation environment in order to characterize the performance of Anycast in a controlled environment. The simulation framework that we constructed as described in the above section provides a generic model to construct any topology of interest.

In this section, we describe how the simulation studies we constructed from the topology consisting of 5476 nodes described above may be used to analyze important metrics of performance for BGP. These will provide valuable insight about several aspects of BGP to the research community. This framework can also be used by network administrators to help them better configure the BGP routers. Further, these simulations also shed light on whether the internal AS topology model bears an influence on the performance of BGP, measured in terms of the convergence time and BGP churn.

3.2 Measurement methodology and BGP anycast environment

In this section, we describe the methodology we employed to construct the simulation. We also describe in detail the various experiments we performed and the analysis of the results. All our simulation studies using the above topology consist of the following steps:

1. The simulation of the topology is started and BGP is allowed to converge
2. Multiple Anycast servers are started in the ASs which advertise the Anycast prefix (these are the DNS root servers corresponding to the C-, F-, K-, J- and M- Root servers)
3. Multiple clients are started in different ASs (these are the DNS clients)
4. The DNS clients send UDP packets representing DNS requests to the server at specific time intervals.

5. Failures are introduced into the system using either explicit prefix withdrawal, silent link failure, or complete router failures

6. Simulation is stopped at a pre-determined time of 6000 simulation seconds.

7. The various log files are parsed to are parsed to obtain metrics of interest.

The simulation topology consisting of 5476 nodes is interconnected through 14,468 links. Each of the 16 individual simulators modeled about 350 nodes using space–parallel distributed simulation. In our topology, 10 ASs provide service to the F-root server, 6 ASs provide service to the J-Root, 17 ASs provide service to the K-root, 4 ASs provide support to the M-Root. Initially we had each BGP speaker advertising one prefix, but experiments showed that the memory requirements for such a network was very large. Thus we changed the configuration files such that only the Anycast prefixes are advertised. We verified using a smaller topology that advertising the non-Anycast prefixes does not have any measurable effects on the metrics of our study. We expect that this is because the failures we induced were only on the Anycast server nodes.

In order to make our simulations experiments more realistic, we used the DNS statistics data for the C-, M-, F-, J- and K- Root servers collected by CAIDA. The CAIDA data showed an average request rate of 6691.8 queries per second from all clients. We divided this by the number of clients accessing each root server, and used an exponential distribution to model the inter–request time intervals.

3.3 Metrics of interest

The following metrics of interest were used in the evaluation of our simulation studies.

1. BGP Convergence: After inducing failures in the topology, we measure how long it takes for BGP to reach a steady state both in the Anycast and the non-Anycast deployment.

2. BGP Churn: The failures will cause an exchange of update messages, and we quantify this exchange both in the Anycast and in the non-Anycast deployment.

3.4 Experimental results

In all the simulations described below, we used the hierarchical mode as described by Sarat [14], where both the local and global nodes advertised a /24 prefix. This closely follows what is used in the current Internet infrastructure. All experiments have a simulation time of 6000 seconds. The first set of experiments provide the baseline case, with no failures included. In the second experiment, we take down one of the interfaces of an F-root server (AS27319) at 1300 seconds of simulation time. Next, in the third experiment we further take down interfaces of the J-,K-,M- and C- Root server interfaces at 1300 seconds of simulation time. As already stated, we are interested in analyzing the robustness of the Anycast deployment in terms of the parameters mentioned above. The experimental results show a loss of 16,321 DNS requests (0.24 percent) destined to AS27319 due to bringing down one of its interfaces. As predicted, taking the J-, K-, C-, and M-Root interfaces down did not have any effect on the loss rate of DNS request destined to the F-Root Anycast address. However the convergence time in the latter was longer, 214 seconds compared to 152 seconds.

Table 3. BGP Performance

	Anycast	Unicast
BGP Convergence Time	214s	152s
BGP Churn(Update messages)	13044	31276

The last two experiments were all run with non-Anycast deployment, with traditional Unicast deployment using one-to-one mapping between DNS servers and IP addresses. The only difference between these two experiments is that F-,J-,K-,C- and M-Root Servers all go down at 1300 seconds in the second experiment, as was done in the earlier experiment. The results of this experimental setup enable us to compare the performance of both the Anycast and the non-Anycast deployment, as well as their effect on BGP in terms of the parameters described above.

The experimental results clearly show the advantages of IP Anycasting. As seen in Table 3, the failures induced in the topology (taking root server interfaces down) cause BGP churn of 31,276 update messages in the non-Anycast deployment compared to 13,044 update messages with the Anycast deployment. This is due to the fact that with the Anycast deployment, the updates only propagate to the affected routers and the best path of other routers remains the same. However, the convergence time due to failures in the case of Anycasting is slightly longer than the normal case.

The DNS response time is measured by noting the time of each DNS request and the time each response is delivered back to the client. Figure 2 and Figure 3 show the comparison of DNS response time between both the Anycast and the non-Anycast deployment without and

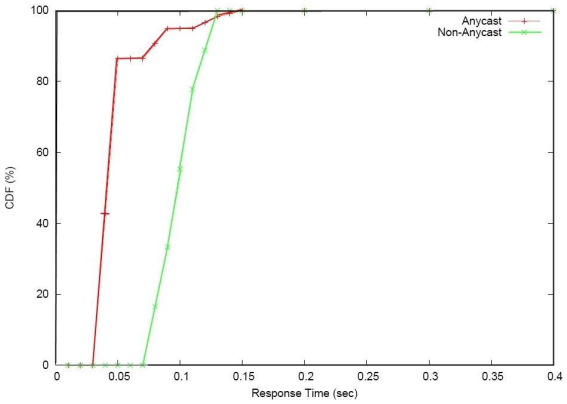

Figure 2. DNS Response Times Without Failures(5476 nodes)

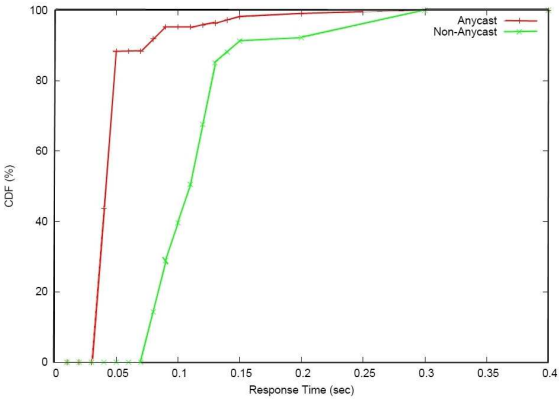

Figure 3. DNS Response Times with Topology Failures (5476 nodes)

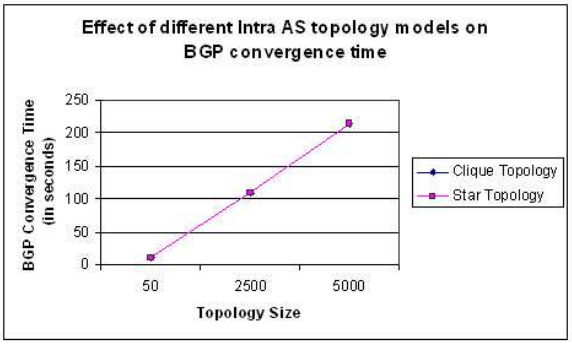

Figure 4. Effect of different intra AS topology models on the BGP convergence time with failures

with topology failures respectively. It is easy to see from the CDF of Figure 2 that 85 percent of the DNS requests were answered at 0.05 seconds with Anycast compared to 0.115 seconds using Unicast prefixes for the F-Root servers. Furthermore, introducing topology failures has a direct effect on the response time in normal case compared to the Anycast case. The response time for 85 percent of the requests was around 0.13 seconds, an increase of 13 percent. With Anycast, the probability to find a best path after a topology change is higher than that in the normal case.

Up till now, no analysis has been done on studying the effect of the Intra–AS topology having multiple BGP speakers on the Internet BGP infrastructure. The internal topology within an AS does not have a common well defined structure, as each network provided will design an internal topology to meet their specific needs. Measurement data that is available for a small number of ASs is not sufficient to generalize this to all ASs.

These factors pose a challenge in performing studies. Since we don't have reliable information about internal topologies within the ASs, we need to determine what effect this might have on our measured results using simulation. In order to determine if the intra AS topology does bear an influence on the BGP anycast structure in our simulation studies, we created two extreme structures, the *clique* and *star* topologies, connecting all BGP speakers within an AS. We ran a number of experiments and and observed the effect on the BGP convergence time and the BGP churn in the two cases. The BGP churn and convergence time were measured after the failures discussed above were induced. It was seen that that the BGP churn in both cases was 13044 messages. This is expected, since the internal BGP topology should have no effect on the number of updates forwarded to peers. Figure 4 shows the BGP convergence

Table 4. Number of update messages after injection of prefixes

BGP Churn (Anycast mode)	1341
BGP Churn (Non-anycast mode)	4580

time after failures are introduced in the system in both the clique and the star topology models. Since the convergence time remains nearly the same in the two cases, we are confident at this point that internal AS topology will have negligible effect on observed inter–AS metrics. This requires more study however to draw firm conclusions.

The robustness of the inter-domain routing infrastructure depends on whether the routers in the ASs are able to filter *bogon* prefixes (private addresses and address space that has not been allocated by the Internet Assigned Number Authority (IANA)). Also, sometimes valid prefixes are filtered by the routers. Feamster [5] reports that about 40 percent of the time new prefixes are filtered even though they are valid. This can cause serious implications in the Internet: the failure to advertise valid prefixes can prevent legitimate routes from being publicly visible and the advertisement of bogon prefixes can cause excessive route flapping and thus affecting the BGP convergence time and the number of update messages sent. In our simulation studies, we randomly inject prefixes into the network belonging to the private address space 192.168.0.0 - 192.168.255.255 after 1300 seconds of simulation time. Routers advertising this prefix belonged to the ASs (AS 10071 and AS174) that had the most BGP neighbors and the effect of this was measured in terms of BGP churn. From Table 4, it can be seen that the number of update messages sent after a new private address prefix was injected is 1,341 messages in the Anycast deployment as against 4,580 messages in the non-Anycast deployment. This demonstrates the need for the proper deployment of bogon filters at the routers by the network operators to prevent such mis–configured BGP advertisements to leak across the Internet.

4 Limitations of the model

Although we use real topology data to build our simulations, this information is based on observed metrics from an incomplete set of the entire Internet topology, specifically the set of RouteViews peers. While we are confident that the model is sufficiently complete to study the behavior of BGP in detail, we do not claim that the topology model is a 100% accurate model of actual Internet topology. Further the set of peering relationships and BGP configurations are based on inferences, rather than exact known relationships and configurations.

Little is known about the internal topology of the tier 1 and tier 2 Internet Service Providers (ISPs). Indeed, such information is generally a closely guarded secret since the ISPs consider this information to be part of their economic competitive advantage which would be detrimental to their business relationships should it become public knowledge. Lacking information regarding internal topologies, we have assumed two extremes in the work here, namely a simple star topology and a fully connected clique. Results presented here show little difference in the measured performance of BGP in these two extremes. Similarly, there is no publicly available information about link speeds and speed of light delays on links connecting Autonomous Systems together, for similar reasons. Here we have assumed reasonable values for these, and have consistently applied them for all experiments.

As in almost all simulation–based experiments, our topology models and workload models are not a completely accurate representation of the deployed Internet. However, we have based the models on observable information, and have consistently used the same assumptions for comparative experiments. This leads to a realistic comparison of the effects of BGP anycasting as reported here.

5 Related work

The SSFNet [2] simulator is one of the most detailed and widely used BGP simulators. However one of the main problems with this simulator is the large memory demand when very large topologies are constructed. Simulations run on this can scale up to only a few hundred nodes.

The recent work by Yu Liu and Boleslaw [15] also uses a distributed packet-level simulation model for BGP networks under Genesis. Their model uses several memory saving schemes to address the memory usage problem. However their simulation is confined to a small topology of the campus network.

To the best of our knowledge, our simulation is one of the first simulation studies to analyze BGP using real data of Internet infrastructure connectivity and realistic peering relationships between the ASs.

6 Conclusion and future work

This work provides a generic framework that can help construct large scale models of the Internet efficiently.

We have shown how this model helps construct a topology that may be used as a simulation environment by researchers to perform simulations and analyze various features and parameters of interest. The anycast experiments that were used to quantify the effect of the current anycast deployment of the DNS root servers provided valuable information regarding both the performance of BGP and the performance of the DNS deployment. To the best of our knowledge, this is the first detailed BGP simulator coupled with IP anycast service. In the future, this work can be used by the BGP research community as a basis for the study of IP anycast and used to analyze possible future anycast deployments in terms of BGP stability. For example, are there possible deployments of anycasting that would cause BGP to never converge or cause excessive route flaps in the presence of simple link or router failures. This model may also be used to understand if there are intelligent ways to place replicas that would mitigate any detrimental effects on BGP and if there are any limits on the number of anycast replicas for a given topology.

For future work, we will perform a more detailed study of the effect of both internal AS topology and inter–AS link characteristics on the overall measured results. Clearly, very high speed links connecting all BGP speakers will result in smaller convergence time that slower links. However, it is not clear whether this would results in significant or relatively minor differences. Similarly, differing internal AS topologies will naturally affect in some way the overall flow of information within an AS; however we expect that this will likely be small relative to the overall performance of BGP globally.

7 Acknowledgements

The authors would like to thank K. C. Claffy and Marina Fomenkov from CAIDA for helpful discussions and suggestions about the availability of measurement data used in our studies. Further, we are grateful to Sebastian Castro from CAIDA for providing the DNS query rate statistics. Finally, we acknowledge the significant effort by Xenofontas Dimitropoulos in creating the BGP++ simulation models and the various scripts and other tools we used to construct the actual simulations.

References

[1] L. Colitti. Effects of anycast on k-root, some early results., 2005.

[2] J. Cowie, A. Ogielski, and D. Nicol. The SSFNet network simulator. Software on-line: http://www.ssfnet.org/homePage.html, 2002. Renesys Corporation.

[3] X. Dimitropoulos and G. Riley. Creating realistic bgp models. In *Proceedings of Eleventh International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS'03)*, pages 64 – 69, Oct 2003.

[4] X. A. Dimitropoulos and G. F. Riley. Large–scale simulation models of BGP. In *Proceedings of the Twelvth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS'04)*, 2004.

[5] N. Feamster, J. Jung, and H. Balakrishnan. An empirical study of bogon route advertisen=ments. *ACM SIGCOMM Computer Communications Review*, January 2005.

[6] L. Gao. On inferring automonous system relationships in the internet. *IEEE Transactions on Networking*, 9(6):733–745, December 2001.

[7] B. Hendrickson and R. Leland. The Chaco user's guide, 1994.

[8] D. Meyer. Oregon routeviews database. http://www.antc.uoregon.edu/route-views/. University of Oregon Advanced Network Technology Center.

[9] P. Mockapetris. Internet RFC1035: Domain names - implementation and specification. Network Working Group, Nov 1987.

[10] D. Nicol. The scalability of networks revisited. In *Proceedings of the Communication Networks and Distributed Systems Modelling and Simulation Conference*, February 2003.

[11] C. Partridge, T. Mendez, and W. Milliken. Internet RFC1546: Host anycasting service. Network Working Group, November 1993.

[12] G. F. Riley, M. H. Ammar, and R. M. Fujimoto. Stateless routing in network simulations. In *Proceedings of the Eighth International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems*, August 2000.

[13] G. F. Riley and T. Jaafar. Space–parallel network simulations using ghosts. In *18th Workshop on Parallel and Distributed Simulation*, May 2004.

[14] S. Sarat, V. Pappas, and A. Terzis. On the use of anycast in DNS. In *Proceedings of ACM SIGMETRICS*, 2005.

[15] B. K. Szymanski, Y. Liu, and R. Gupta. Parallel network simulation using distributed genesis. In *17th Workshop on Parallel and Distributed Simulation*, June 2003.