Robust Routing Under BGP Reroutes

Anders Gunnar
Swedish Institute of Computer Science
P.O. Box 1263
SE-164 29 Kista, Sweden
Email: anders.gunnar@sics.se

Mikael Johansson
School of Electrical Engineering
KTH
SE 100 44, Stockholm, Sweden
Email: mikaelj@ee.kth.se

Abstract—Configuration of the routing is critical for the quality and reliability of the communication in a large IP backbone. Large traffic shifts can occur due to changes in the Inter-domain routing that are hard to control by the network operator. This paper describes a framework for modeling potential traffic shifts due to BGP reroutes, calculating worst-case traffic scenarios, and finding a single routing configuration that is robust against all possible traffic shifts due to BGP reroutes. The benefit of our approach is illustrated using BGP routing updates and network topology from an operational IP network. Experiments demonstrate that the robust routing is able to obtain a consistently strong performance under large Inter-domain routing changes.

I. INTRODUCTION

An important part of provisioning communication services in an IP network is managing the traffic situation. A thorough understanding of the dynamics of the traffic is necessary in order to optimize utilization of available resources and to meet service level agreements made with customers. In addition, the transfer of critical services such as telephony to IP networks has made it even more important for a network operator to monitor and control the traffic.

However, the traffic situation is highly dependent on the interplay between intra- and inter-operator routing. An operator who acts as a provider for other network operators often receives reachability information for a network from several different places. This reachability information is given in the form of a network prefix which represent the address of the network. Routing is performed by matching the destination address with the prefixes in the routing table and selecting the route with the longest prefix match. When there are several routes available a router has to select one of these routes; *i.e.* the ingress router of the traffic has to select which route to use for forwarding the traffic towards the destination network. How the selection is performed has implications on how the traffic is routed within the network since the ingress router selects an egress router where the traffic leaves the operator's network. This selection of routes may cause large shifts in the load in the network, see *e.g.* [1].

In this paper we introduce a method to control and minimize the implications of load shifts caused by changes in the inter-domain routing. In particular, we model the uncertainty of traffic demands due to BGP reroutes, formulate and solve a convex optimization problem to identify the worst-case scenarios for a given MPLS routing, and sequentially improve the

routing by introducing additional tunnels that allows to hedge against these scenarios. To reduce the number of variables in the problem we devise an algorithm that identifies the prefixes with multiple egress points and large traffic volumes. In addition, worst-case scenarios are generated by considering one link at a time (finding the ingress/egress traffic demands that maximizes the utilization of each individual link, and selecting the traffic scenario that gave the largest link utilization) which allows that part of the algorithm to be highly parallelized.

Our method is applied to traffic data and inter domain routing information from an operational Internet Service Provider. We find that significant improvements are possible under a number of scenarios. For comparison we also include shortest path routing according to the original link weights as well as multi-commodity flow optimization for the nominal traffic situation in our analysis.

Optimization over multiple traffic scenarios has received a lot of attention from researchers (*cf.* [2], [3], [4], [5], [6]). In a pioneering paper by Fortz and Thorup [3] the authors use a search heuristic to optimize the routing over a set of traffic scenarios. Applegate and Cohen [2] calculate an upper bound for the performance of the routing under all possible traffic scenarios. The upper bound on performance is used by Wang *et al.* [5] for comparison with their method which embeds a traffic scenario in a traffic envelope and optimizes the routing for the traffic scenario and limits performance of the routing for every traffic scenario in the envelope. In this paper we follow the approach taken by Ben-Ameur and Kerivin [6] by using column-generation to optimize the routing. However, our approach differs from previous work by incorporating inter domain routing in the solution and thereby make our results directly applicable for large IP networks with several peering points with other operators.

The rest of the paper is organized as follows. In the next section we give a short description of how routing is performed in the Internet. Section III introduces the algorithm, including the generation of worst-case traffic scenarios and the robust routing optimization. The analysis of traffic data from an operational IP network is presented in section IV. Finally we wrap up with conclusions and future work.

Fig. 1. Routing scenario where the prefix 192.168.0.0/16 is announced by two peering points in the network. Router R3 has to select a route using the BGP decision process.

II. BACKGROUND

A. Routing in the Internet

The Internet is a network of independent networks. These networks are referred to as Autonomous Systems (AS) and are administered by separate organizations. The routing inside an AS is managed by an Interior Gateway Protocol (IGP). Typically, IGP is a link state routing protocol like Intermediate System Intermediate System (IS-IS) or Open Shortest Path First (OSPF). In link state routing the network is modeled as a graph where nodes represent routers and arcs represent links connecting the routers. Each node collects information about network topology and calculates the shortest path to each destination node in the network.

In order to connect AS'es and exchange connectivity information, an External Gateway Protocol is used. The protocol currently in use is called Border Gateway Protocol version 4 (BGP4) [7]. BGP is a path vector protocol where an AS announces to its neighboring AS'es which networks it has a route to. In order to avoid routing loops the path of AS'es is included in the routing messages. In addition, the routing decision is also based on policies reflecting the relation the AS has with other AS'es, e.g. peering, customer or provider relations. When an AS has more than one route to a prefix, BGP has to select one route from the set of available routes. This is performed according to a decision process. The first step is to determine if there is a route to the egress point of the AS. Next BGP examines a number of BGP specific attributes.

If BGP still is unable to select one route, the shortest distance according to IGP is considered. This is sometimes referred to as hot-potato routing [8]. The final step is to use a vendor-specific tie-breaking. Figure 1 illustrates a simple example of a situation where a prefix is announced by two routers. In the example router R3 selects the route announced by R2 since it has the shortest IGP distance to R3. However, if the route announced by R2 is withdrawn the traffic towards network 192.168.0.0/16 injected in the network by R3 is shifted from the route announced by R2 to the route announced by R1, causing a potentially massive change of the load on the links in the network.

III. ROBUST ROUTING UNDER BGP REROUTES

A. Robust routing under uncertain traffic demands

Robustness, referring to the ability to cope with variations from the nominal operating conditions, is a key property of any engineering system. In this spirit, a robust network should be able to sustain acceptable performance despite foreseeable traffic variations and component failures. A common optimization objective in robust networking is to minimize the worst-case link loads, where worst-case should be understood as over all potential load variations or component failures. Our focus is on demand variations due to BGP reroutes.

Several methods for robust routing have been proposed recently [4], [6], [9], [10]. We will base our developments on the approach by Ben-Ameur and Kerivin [6] as we find it the most transparent. The method starts out from a standard arc-path formulation of multi commodity network flows

$$\begin{aligned} & \text{minimize} && u_{\max} \\ & \text{subject to} && \sum_k \sum_{\pi \in \Pi_k} r_{l\pi} \alpha_{\pi k} s_k \leq c_l u_{\max} \forall l \\ & && \sum_{\pi \in \Pi_k} \alpha_{\pi k} = 1, \quad \alpha_{\pi k} \geq 0 \end{aligned} \quad (1)$$

Here, s_k is the aggregate traffic between source-destination pair k , Π_k is the set of all paths between source-destination pair k and $r_{l\pi}$ is an indicator variable taking the value one if path π traverses link l and zero otherwise. The optimization variables $\alpha_{\pi k}$ determine what fraction of the traffic between source destination pair k that is routed across path π . The first set of constraints state that the total traffic across each link l is bounded by the link capacity times the maximal link utilization, while the second constraint states that all traffic must be routed across some path. The classical way of solving (1) is by column generation. Rather than explicitly enumerating all paths in the network, one starts out with a small subset of paths (e.g., the shortest-hop routing) and then sequentially adds new paths to the problem to improve the optimization objective, see e.g., [11] for details.

The robust multi commodity network flow problem is to find the routing that guarantees the smallest link utilization for all feasible traffic scenarios. We can formulate the problem as

$$\begin{aligned} & \text{minimize} && u_{\max} \\ & \text{subject to} && \sum_k \sum_{\pi \in \Pi_k} r_{l\pi} \alpha_{\pi k} s_k \leq c_l u_{\max} \forall l, \forall s \in \mathcal{S} \\ & && \sum_{\pi \in \Pi_k} \alpha_{\pi k} = 1, \quad \alpha_{\pi k} \geq 0 \end{aligned} \quad (2)$$

Depending on the nature of the traffic uncertainty set $\mathcal{S}$, this problem may or may not admit an efficient solution. If the traffic uncertainty is polyhedral $\mathcal{S} = \text{co}\{s^{(1)}, \dots, s^{(V)}\}$, then (2) can be equivalently expressed as

$$\begin{aligned} & \text{minimize} && u_{\max} \\ & \text{subject to} && \sum_k \sum_{\pi \in \Pi_k} r_{l\pi} \alpha_{\pi k} s_k^{(v)} \leq c_l u_{\max} \forall l, v \\ & && \sum_{\pi \in \Pi_k} \alpha_{\pi k} = 1, \quad \alpha_{\pi k} \geq 0 \end{aligned} \quad (3)$$

There are at least two problems with this formulation. First, the traffic uncertainty sets are typically not given in vertex form, but as the set of solutions to a system of linear inequalities (cf. the demand uncertainty set $\mathcal{S}$ in Johansson and Gunnar [4]). Secondly, the uncertainty set may have many vertices, so that explicit enumeration is computationally unattractive. These two issues can be addressed similarly to the way column generation is used to avoid explicit enumeration of all paths in the nominal formulation: one starts out with a single traffic scenario in the uncertainty set, solves the routing problem, and then verifies whether the computed routing satisfies the link constraints for all feasible traffic loads. If this is not the case, one adds the traffic matrix that violates the constraints the most to the vertex description of the uncertainty set and repeats. The resulting method is a combined column- and constraint generation scheme, and is readily shown to have finite convergence (e.g. [6]).

B. A model for traffic uncertainty due to BGP reroutes

To describe traffic uncertainty under BGP reroutes, it will be convenient to be explicit about the source and destination node for each demand. Thus, rather than using the notation s_k for the traffic between source-destination pair k , we will write s_{oe} to emphasize that the traffic originates at nodes o and is destined for egress point e . Let $E(p)$ be the set of egress points for prefix p (i.e., the set of peering points that could potentially announce prefix p) and, conversely, let $P(e)$ be the set of prefixes that can be announced by peers connected to egress node e . The total demand from node s exiting the system at node e can then be described as

$$s_{oe} = \sum_{p \in P(e)} d_{op} \delta_{pe}$$

with

$$\sum_{e \in E(p)} \delta_{pe} = 1, \quad \delta_{pe} \geq 0 \text{ and } \delta_{pe} = 0 \text{ for } e \notin E(p)$$

In this formulation δ_{pe} can be interpreted as the relative amount of traffic demand for prefix p that can be served via egress point e . At first, this model might seem counter-intuitive as the peering autonomous systems can only decide whether or not to announce a certain prefix and not influence the relative amount of demand for a specific prefix that it will allow to transit. However, as we will see shortly, the model serves its purpose. Now, assume that the internal routing is fixed. The utilization of link l can then be written as

$$u_l = c_l^{-1} \sum_o \sum_e \alpha_{loe} s_{oe}$$

where α_{loe} is the fraction of the traffic between nodes o and e that traverses link l . In terms of the notation in the previous section, if (o, e) is source-destination pair k , then

$$\alpha_{loe} = \sum_{\pi \in \Pi_k} r_{l\pi} \alpha_{\pi k}$$

Combining this with the expression above, we find

$$u_l = c_l^{-1} \sum_o \sum_e \alpha_{loe} \sum_{p \in P(e)} d_{op} \delta_{pe} \quad (4)$$

The worst-case traffic scenario is when prefixes are announced at peering points in a way that maximizes the maximum link utilization. From the expression above, we see that the worst-case situation is when prefix p is only announced at the egress e with largest value of $\alpha_{loe} d_{op}$ (i.e. when $\delta_{pe} = 1$ for this egress and zero for the others). Thus, in worst-case traffic scenarios generated by adjusting the prefix distributions to maximize the worst-case link utilization will be such that each prefix is announced by a single peer only, and thus compatible with realistic (and admissible) BGP configurations.

C. Optimizing routing for BGP reroute uncertainty

We are now ready to summarize our procedure for finding a routing that is robust to BGP re-routes.

- 1) Generate a nominal traffic scenario set $\mathcal{S}$ by picking a single peering point for each prefix and computing the associated traffic matrix.
- 2) Compute the robust routing for the traffic scenario $\mathcal{S}$ by solving (3).
- 3) Fix the current routing and determine the prefix distribution that maximizes the utilization of the most loaded link by solving (4) for each link l . If the worst-case utilization is higher than predicted when optimizing the routing, add the corresponding traffic matrix to the scenario set $\mathcal{S}$ and return to step 2), otherwise terminate the algorithm.

Since the complete scenario set is finite, the algorithm has finite convergence. However, our computational experience, reported next, indicates that only a handful of iterations need to be carried out before the worst-case traffic scenarios are found and the optimal routing can be determined.

IV. ANALYSIS ON TRAFFIC DATA FROM AN OPERATIONAL IP NETWORK

In this section we evaluate our approach using traffic data from an IP network operator. We start with describing the network and highlight some properties of the routing and traffic data.

A. Data collection and evaluation data set

For the evaluation we have access to traffic data obtained from Netflow measurements as well as BGP routing information base and network topology. The data set was obtained from the Geant network [12] connecting European national research and university networks and consists of 23 nodes and 74 links. The measurements were conducted during a four month period and consist 15 minute flow export of sampled Netflow measurements with sampling rate of 1/1000; i. e. one packet of one thousand is sampled. In addition, a dump of the BGP routing information base from each day of the measurement period was conducted. The analysis in this paper was performed on data from one 15 minute measurement.

Fig. 2. Cumulative traffic distribution

Fig. 3. Number of prefixes with multiple exit points in the network

More details about the network and traffic data can be found in Uhlig *et al.* [13].

B. Evaluation

1) *Preliminary data analysis:* Figure 2 shows the cumulative distribution of traffic in the Geant network classified by prefix. The prefixes are ranked by the amount of traffic sent towards them during the measurement period. The figure reveals that only around seven percent of the prefixes have traffic routed towards them. The distribution of exit points for the prefixes is shown in the histogram in Figure 3. One can see that more than 60% of the prefixes are announced by five different locations and only three percent are announced by a single location. This could lead to a disruptive behavior in the traffic distribution since BGP might select another egress router for the traffic, with a potentially large impact on the load on internal network links. Figure 4 reveals that while most of the traffic is routed towards networks with only one exit point announced, 40% of the total traffic has multiple exit points and can thus be shifted around due to BGP reroutes.

2) *Reducing the number of variables:* Solving the optimization problem in Equation (4) for every prefix in the network would create a huge optimization problem since a typical backbone router has in the order of 160000 prefixes in its routing table. However, from Figure 2 we learn that only a small fraction of the prefixes account for the traffic in the network. Hence, by filtering out the prefixes with negligible

Fig. 4. Number of bytes destined to prefixes with multiple exit points

traffic we are able to reduce the number of variables substantially. In our experiments we selected the prefixes that account for 90% of the traffic in the Geant network. Thus reducing the number of prefixes in our equations to 3600. In addition, since we consider the worst-case link utilization as optimization metric, we can treat links one-by-one, reducing the number of variables even further. With these tricks we are able to reduce the number of variables to the order 60000. Although this still constitutes a large optimization problem, most of the variables are uniquely determined by the constraints and the problem is readily solved on a regular desktop computer.

3) *Experimental results:* The nominal traffic situation in our experiments is the traffic demands where all possible routes are announced in the network and link weights are set to the original values. In our experiments we have calculated the link loads for the following routing principles:

- **ROBUST:** the approach described in Section 3 where the worst case traffic scenarios from repeated optimization of Eqn.(4) are used to form the polyhedral $\mathcal{S}$.
- **MCNF_NL:** Multi commodity network flow routing using node-link formulation to minimize the maximum link utilization under nominal traffic.
- **MCNF_LP:** Multi commodity flow using a link-path formulation, i.e. solving problem (1), under nominal traffic.
- **SPF:** Shortest path first routing using the original link weights from the Geant network.

Figure 5 shows the utilization for the links in the Geant network under ROBUST, MCNF_NL and SPF routing for the nominal traffic scenario (in which the robust routing coincides with MCNF_LP). We can see that although the node-link and link-path formulations achieve the same maximum link utilization, the robust routing achieves a better balance in the overall link utilization. This is due to that new paths are calculated using the dual variables of the link constraints in (3), which discourages routing across highly loaded links.

In Figure 6 we have plotted maximum link utilization under feasible traffic shifts for three routing configurations (SPF, MCNF_LP and ROBUST) and four scenarios (nominal traffic and three worst-case scenarios generated during the robust optimization). The robust routing is able to route efficiently

Fig. 5. Link utilizations in the Geant network for robust, optimal and shortest path routing using the real link weights in the nominal traffic scenario.

Fig. 6. Maximum link load for SPF, MCNF_LP and ROBUST evaluated for the critical traffic scenarios generated by the robust routing algorithm. The maximum link utilization for the non-robust routings is around 0.6 (in the Nominal case for SPF, and in Scenario 1 for MCNF_LP) while it never exceeds 0.33 for the robust routing.

in all three scenarios whereas the multi-commodity network flow routing optimized for the nominal traffic scenario suffers a substantial performance losses under BGP-reroutes, and performs on par with the original shortest-path routing.

Table I summarizes performance for each iteration of the algorithm in section III-C. After four iterations the algorithm terminates with 758 paths. The algorithm has added 252 paths to be set up by MPLS in addition to the 506 shortest paths from link state routing.

V. CONCLUSIONS AND FUTURE WORK

In this paper we have introduced a novel method to find critical traffic scenarios that can be used to find a routing setting that can route efficiently under all realistic traffic scenarios that can occur in a network due to inter domain rerouting. The scenarios are identified by finding the worst

Iteration	1	2	3	4
$u_{\max}$	0.6	0.39	0.37	0.33
Paths	506	705	730	758

TABLE I
MAXIMUM LINK UTILIZATION AND NUMBER OF PATHS IN NETWORK FOR EACH ITERATION OF THE ALGORITHM

case setting of the Inter-domain routing by solving a convex optimization problem. We show that the robust routing is able to minimize link load under a number of plausible traffic scenarios.

Our approach only considers changes in the external routing. Many occurrences of massive traffic shifts in a network stems from changes in the internal topology. To devise an algorithm that take these changes into account is a much more challenging problem and is one avenue of future work. Further, our results has only been tested on one sample of traffic and routing data from one network. A more interesting scenario is to test our algorithms on a time series of data and for data from other networks. For instance Figure 4 reveals that only 20 percent of the traffic is routed to prefixes announced in five places. A network with a larger fraction of traffic routed to prefixes announced in multiple places would have illustrated the benefit of our approach clearer. Another property of the Geant network that caused some problems in our experiments was that the links in the network have highly diverse capacity, indicating that it could be relevant to study other performance measures than worst-case link utilization.

ACKNOWLEDGMENT

This work was supported by SICS Center for Networked Systems, the Linnaeus center ACCESS and the Swedish Research Council.

REFERENCES

- [1] R. Teixeira, N. Duffield, J. Rexford, and M. Roughan, "Traffic matrix reloaded: The impact of routing changes," in *Proceedings of PAM*, 2005.
- [2] D. Applegate and E. Cohen, "Making intra-domain routing robust to changing and uncertain traffic demands: Understanding fundamental tradeoffs," in *Proc. ACM SIGCOMM*, Karlsruhe, Germany, August 2003.
- [3] B. Fortz and M. Thorup, "Optimizing OSPF/IS-IS weights in a changing world," *IEEE Journal on Selected Areas in Communications*, vol. 20, no. 4, pp. 756–767, 2002.
- [4] Mikael Johansson and Anders Gunnar, "Data-driven traffic engineering: techniques, experiences and challenges," in *Broadnets 2006*, San Jose, California, Oct. 2006.
- [5] Hao Wang, Haiyong Xie, Lili Qiu, Yang Richard Yang, Yin Zhang, and Albert Greenberg, "Cope: traffic engineering in dynamic networks," in *SIGCOMM '06: Proceedings of the 2006 conference on Applications, technologies, architectures, and protocols for computer communications*, New York, NY, USA, 2006, pp. 99–110, ACM Press.
- [6] W. Ben-Ameur and H. Kerivin, "Routing of uncertain demands," *Optimization and Engineering*, vol. 6, no. 3, pp. 283–313, 2005.
- [7] S. Halabi and D. McPherson, *Internet Routing Architectures*, Cisco Press, 2001.
- [8] R. Teixeira, A. Shaikh, T. Griffin, and J. Rexford, "Dynamics of hot-potato routing in ip networks," 2004.
- [9] D. Applegate and E. Cohen, "Making intra-domain routing robust to changing and uncertain traffic demands: Understanding fundamental issues," in *ACM SIGCOMM*, Karlsruhe, Germany, August 2003.
- [10] A. Sridharan, R. Guérin, C. Diot, and S. Bhattacharyya, "The impact of traffic granularity of robustness of traffic aware routing," Technical report, University of Pennsylvania, 2004, Available via <http://einstein.seas.upenn.edu/mnlab>.
- [11] M. Pioro and D. Medhi, *Routing, Flow and Capacity Design in Communication and Computer Networks*, Morgan Kaufmann Publishers, 2004.
- [12] *Geant*, <http://www.geant.net>.
- [13] S. Uhlig, B. Quoitin, S. Balon, and J. Lepropre, "Providing public intradomain traffic matrices to the research community," *ACM SIGCOMM Computer Communication Review*, vol. 36, no. 1, January 2006.