Reliability-aware IBGP Route Reflection Topology Design [*]

Li Xiao, Jun Wang and Klara Nahrstedt

Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL 61801

Abstract

In the Internal Border Gateway Protocol (IBGP), route reflection is widely used as an alternative to full mesh IBGP sessions inside an AS for scalability reason. However, some important issues, such as the impact of route reflection on the reliability of IBGP and the construction of reliable reflection topology with unreliable routers or links, have not been well investigated.

This paper addresses the problem of finding reliable route reflection topologies for IBGP networks, which is of great importance to increase the robustness of IBGP operations. We first present a novel reliability model and two new metrics (IBGP expected lifetime and expected session loss) to evaluate the reliability of reflection topologies, and further to investigate the design problem. After studying the solvability conditions under the router capacity constraints, we prove the NP-hardness of the problem, and then design and implement three heuristic solutions using randomization techniques: heuristic selection, greedy search and simulated annealing. Our extensive computational experiments show that the reliability of IBGP reflection network can be significantly improved by our solutions.

1 Introduction

Border Gateway Protocol (BGP) [13] is the de facto routing protocol for exchanging network reachability information at the inter-domain level. Two BGP routers exchange routing information via *BGP sessions*. A BGP session may cross multiple hops, and it depends on IGP (Internal Gateway Protocol) and TCP for underlying communication support. BGP can be divided into two parts: External BGP (EBGP) and Internal BGP (IBGP). An EBGP session connects two BGP routers which reside in different Autonomous Systems (AS); An IBGP session links two BGP routers which belong to the same AS.

Traditionally, all BGP routers in one AS form a full mesh via IBGP sessions, which is not scalable in large networks [8][18]. Route reflection [2] is a widely used technique to solve the scalability problem in IBGP. The basic idea of route reflection is to divide the BGP routers into multiple clusters. In each cluster, there are one or more Route Reflectors (RR), and other BGP

routers are Route Reflection Clients (RRC). The clients establish IBGP sessions only with the reflectors of the same cluster. All reflectors in the AS establish a full mesh via IBGP sessions. A reflector is responsible for reflecting routing information to its peer reflectors and its clients. A client only communicates with its reflectors. Fig. 1(a) shows an example of a route reflection network, where the solid lines represent physical links; the dotted lines represent IBGP sessions and the shaded nodes represent reflectors. Five routers are divided into two clusters, $\{A, C, E\}$ and $\{B, D\}$. C and D are route reflectors.
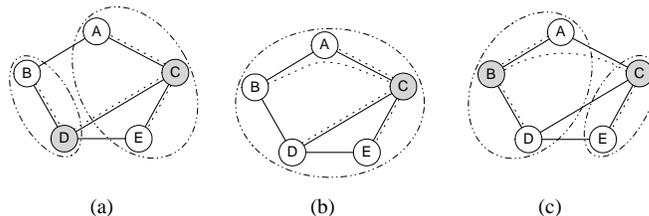

Figure 1. IBGP route reflection networks.

However, so far, the design of route reflection topology has not been well studied. When a domain switches from the full mesh IBGP to route reflection, the guideline for setting up reflection topology is usually to follow the physical topology [9]. In general network topologies, route reflection network design problems have not been well investigated. There are two reasons which make this research necessary: (1) We need to study the impact of reflection topology on IBGP reliability. For example, if IBGP sessions are connected in full mesh, a failure of one session only breaks route exchange between two routers; while in the reflection topology of Fig. 1(a), a session failure between D and C would partition the reflection network. How to model the reliability of IBGP operation under route reflection is a critical problem. (2) It is very helpful to provide a reliability analysis framework for IBGP reflection, and also give some hints on reliable reflection topology design for general networks. IGP hierarchy gives some guidelines for the top level router clustering, but how to design the reflection topology inside an IGP area needs to be carefully considered.

The reliability and robustness of IBGP operation are influenced by the IBGP route reflection topology design, because the topology determines route reflector placement and IGP paths used for transmitting BGP messages. Some powerful routers, possessing a large amount of resource (such as memory and CPU power), are more reliable and thereby more likely to survive in some stressful situations. Also, physical links may

have different failure rates and some links are much more stable than others [10].

Three major issues need to be considered in designing a reliable IBGP route reflection topology: (1) *How many clusters are needed?* The use of too few reflectors may cause single point of failure problem. For example, the reflection topology in Fig. 1(b) may be less robust than the topology in Fig. 1(a), because if C fails, in Fig. 1(b), all IBGP routers will be isolated. On the other hand, increase of cluster numbers leads to more IBGP sessions, which increases router workload and introduces more unreliable components into the IBGP system. We will show later that the optimum number of clusters or reflectors depends on the reliability of links or routers and the redundancy of the physical network topology. (2) *Where to place reflectors?* Intuitively, more reliable routers are more appropriate to be reflector candidates, because reflectors are critical components in reflection networks. For example, in Fig. 1(c), if B has too limited resource to provide reliable route reflection for A and D, then we prefer to use the reflection topology in Fig. 1(a), given D is much more reliable. (3) *How to assign clients to reflectors?* The assignment of clients influences which IGP paths are used by IBGP sessions. Different IGP paths may provide different qualities of IBGP session support. For instance, in Fig. 1(c), A can be assigned to either B or C as client. If link AB is less reliable, it is better to assign A to C.

The purpose of this paper is to find the most reliable route reflection topology for an AS. In order to achieve this goal, we propose two reliability evaluation metrics, based on which our optimization methods will be developed. The first metric is the *expected lifetime of reflection network*. It is defined as the expected time interval between two consecutive IBGP session failures (including IBGP router failures). The second metric is the *expected session loss* in unit time due to a single IBGP failure, where the *loss* is defined as the percentage of broken IBGP sessions. The design target is to maximize the expected lifetime or minimize the expected session loss. On the other hand, the design of reliable route reflection topology is subject to two major constraints. First, each IBGP router has a maximum number of IBGP sessions that can be handled concurrently due to the scalability reason. Second, the reliability optimization is constrained by some human configuration decisions. For example, some IBGP routers are dedicated to be route reflectors.

The benefits of reliable route reflection topology design are twofold: (1) *Increase of network availability:* Since IBGP session failures may cause route withdrawals that make some network addresses unreachable, the reliable route reflection design increases the network availability by reducing the IBGP session failures. (2) *Increase of BGP routing stability and decrease of computing overhead:* Because IBGP session failures result in route flaps (route withdrawal and re-advertising) which may further incur routing oscillation and large computing cost, improving reliability of reflection topology can increase the BGP routing stability and reduce computing overhead at routers.

Some recent research addresses the challenges of improving BGP routing reliability and stability. Sangli et al. [14] propose a graceful restart mechanism for BGP to alleviate the impact of BGP session failures. BGP Scalable Protocol [12] uses application level replication and flooding to substitute TCP for reliable and scalable BGP message distribution. However, besides the deployment difficulties, they can not replace the hierarchical IBGP design (e.g. route reflection) because of scalable BGP message processing. Therefore, we still need to consider how to construct a reliable router hierarchy to provide robust IBGP.

Different from these existing approaches, our work aims to increase BGP routing reliability and stability, without modifying any protocol details, by configuring IBGP route reflection infrastructure appropriately, so that the rate and impact of IBGP session failures are minimized. Our previous work in [16] has addressed a reflection topology design problem in a simplified scenario, in which the dependency between different IBGP sessions is ignored. In this paper, we consider the session correlations that come from some shared physical links.

The rest of the paper is organized as follows: In Section 2, we define models for physical networks and route reflection networks. In Section 3, we analyze the reliability of IBGP operation in detail and propose two reliability metrics to evaluate different reflection topologies. In Section 4, the reliability-aware reflection topology design problem is first formulated. We also describe the solvability conditions for this problem, and give the integer programming models to solve this problem. We prove the NP-hardness of the design problem in Section 5. In Section 6, we give three heuristic solutions to the design problem. Section 7 shows the implementation and the result analysis. Section 8 concludes the paper.

2 Network Models

In an AS with route reflection deployed, if we view each IBGP session as a logical link, the IBGP routers actually form an overlay network above the physical network. The overlay network is called the *reflection network*, and its topology is then called the *route reflection graph*. In this section, we define the physical network and reflection network model.

2.1 Physical Network and Reflection Network

A typical network in an AS is represented as graph $G(V, E)$. Node set V represents all the routers. E is the set of physical links. The physical link from u to v is denoted as (u, v). Given any two nodes s and t, IGP routing (e.g., OSPF) provides a path from s to t, written P_{st}.

The overlay route reflection network is based on the physical network. We denote the reflection topology as graph $G_r(V_r, E_r)$. V_r is the set of nodes running IBGP (BGP routers). $V_r \subseteq V$, and some nodes in V may not have BGP deployed. Define $n = |V_r|$ as the number of IBGP routers. E_r represents the set of IBGP sessions, i.e., $E_r = \{\langle u, v \rangle \,|\, u, v \in V_r, u$ and v share an IBGP session. $\}$, where $\langle u, v \rangle$ denotes the IBGP session between u and v. $m = |E_r|$ is the total number of IBGP sessions.

Given a reflection network and the IGP routing, we can map IBGP session $\langle u, v \rangle$ to a sequence of links and routers on the

IGP paths P_{uv} and P_{vu}. Also, for router $s \in V_r$, we denote h_s as the number of IBGP sessions owned by s, i.e., the degree of the node s in graph G_r; denote k_s as the number of IBGP sessions whose IGP paths pass s, but which are not owned by node s. Similarly, for physical link $e \in E$, denote k_e as the number of IBGP sessions which pass link e. For example, in Fig. 1(b), $h_A = 1$, $k_A = 1$, and $k_{AC} = 2$.

In general, route reflection hierarchy can have arbitrary number of reflection levels, i.e., some reflectors are the clients of some higher level reflectors, which are in turn the clients of others, and so on. In a two-level reflection hierarchy, a BGP node is either a reflector or a client. Denote n_r as the number of reflectors. The nodes are grouped into several clusters. Each cluster contains one or multiple route reflectors. Based on the relationship of two nodes, the IBGP session between them falls into three categories: (1) IBGP session between reflectors (recall that a full-mesh of IBGP sessions is formed over all reflectors); (2) IBGP session between a reflector and a client; (3) IBGP session between two clients which are in the same cluster, and this type is optional in a route reflection topology.

2.2 Problem Scope

As stated above, the general route reflection topologies have a few variations, such as different levels of reflection hierarchy and the use of optional IBGP sessions or reflectors. In this section, we define the type of reflection network that will be considered for reliable reflection design in this paper.

First, we only consider a two-level reflection hierarchy. In practice, two or three reflection levels are enough to handle the IBGP scalability problem. The three-level reflection is used in very large networks, in which the first reflection level is usually determined geographically by IGP areas and does not have many options for topology design. Thus, we only investigate the second reflection level which is a two-level reflection network in one IGP area, and has a large design space for reliability optimization. Second, we work on a simple network scenario where the reflection graph does not have redundancy, i.e., there is only one reflector in each cluster and there is no optional IBGP session between clients. We are interested in the fundamental problem – how reliable an IBGP reflection network can achieve without any help from redundancy. The analysis of reflection network with redundant reflectors is left for the future research. Third, for the convenience of problem explanation, we focus on a transit domain where all nodes are BGP routers, i.e., $V = V_r$. Actually, we will find that the problem formulation and optimization methods proposed in this paper can be simply extended to the case where some nodes are not BGP routers.

3 IBGP Reliability Model

The reliability of IBGP reflection network can be analyzed from two aspects: reliability of IBGP routers and reliability of IBGP sessions.

3.1 Reliability of IBGP routers

Different routers, which have a variety of software or hardware platforms, may show different reliability in hosting BGP operation. Running BGP requires a large amount of resources for session maintenance, route selection, handling routing updates and route storage, especially when a router possesses many BGP sessions concurrently. The router becomes less reliable if it is overloaded. For example, memory allocation failures (either by running out of memory or memory segmentation) or extremely high CPU utilization will cause a router to hang[6][5][4]. Chang et al. show that BGP router failures (even cascading failures) are resulted due to large BGP table injection which makes a router run out of memory[3]. Therefore, routers with larger amount of resources (CPU power and memory) are more robust for handling BGP operations.

Based on the above discussion, we can draw two conclusions: (1) It is neither reliable nor scalable for a router to possess too many IBGP sessions, which leads to tremendous resource consumption. For each IBGP router, there is an upper bound on the number of IBGP sessions that the BGP router can have simultaneously. Because different routers may have different capacities or different numbers of existing EBGP sessions, this upper bound of IBGP sessions is specific for each router. For example, if some border router has already established many EBGP sessions with routers in other ASes, it can only handle a small number of IBGP sessions. We use c_i to denote the IBGP session upper bound for BGP router i. (2) Routers with more resources tend to be more reliable in handling BGP operation, especially during route flapping or peer misbehaving. We denote the failure rate of router i by v_i, and model router failure events as a Poisson process. In this paper, we assume c_i and v_i are known, based on which the reliability of reflection will be analyzed.

3.2 Reliability of IBGP Sessions

BGP routers detect BGP session failures using a timeout mechanism. Each BGP router expects to receive at least one message from every peer in certain period of time, which is defined by the `Hold Timer`. Accordingly, `KEEPALIVE` messages are sent to peers to keep the sessions alive. Thus, any reasons which cause BGP message delays or losses, such as physical connectivity failures or transport layer instability (severe congestion), may further cause the related BGP sessions to be reset. Iannaccone et al. [10] show that link failures occur as part of everyday operation, and about 50% of the failures last longer than one minute. Although IGP re-routing and TCP retransmission can recover lost packets in case of network failures, BGP messages may be largely delayed due to the re-routing convergence time and the TCP retransmissions. If the BGP `Hold Timer` expires because of the long delay or network partition, the IBGP sessions are broken. Therefore, with some probability, physical link failures may lead to failures of BGP sessions which use this link. Similarly, severe network congestion, which delays the delivery of BGP packets, may

trigger IBGP session failures in some scenarios [15].

In order to model the reliability of IBGP sessions, we assume that the event of failures (including severe congestion) of each physical link follows Poisson process, and denote w_{ij} as the failure rate of link (i,j). The failure of link (i,j) will make the IBGP sessions that pass link (i,j) fail with probability p_{ij}. Likewise, the failure of router s will stop all IBGP sessions that s possesses, and also terminate the IBGP sessions that pass s with probability q_s. The values of p_{ij} and q_s are determined by IP re-routing convergence time, TCP retransmission behavior and BGP timers. By assuming that these parameters are known, our main focus of this paper is to analyze and optimize the reliability of reflection topology. Specifically, we denote $p_e = \overline{p_{ij}}$ and $p_r = \overline{q_s}$, which stands for the average impacts of link failures and router failures on IBGP sessions in the reflection network, respectively. We further assume that the failure events of different physical links or routers are independent.

3.3 Reliability of Route Reflection Topology

We define one *IBGP failure* as the termination of one or several IBGP sessions, which are caused by one router failure or one physical link failure. In real backbone networks, simultaneous physical component failures rarely happen. Thus, we consider different IBGP failures as independent events. Two metrics are proposed to evaluate IBGP reliability: *Expected LifeTime* (ELT) and *Expected Session Loss* (ESL). We can minimize the impact of IBGP failures by maximizing ELT or minimizing ESL.

(1) *Expected Lifetime* $\mathcal{T}$

Given a route reflection network, the expected lifetime $\mathcal{T}$ refers to the remaining time from now to next IBGP failure. Because IBGP failures caused by different physical links or routers are independent, we can calculate the expected lifetime by the failure rates of physical components.

A router failure definitely breaks all the IBGP sessions possessed by that router. In a link failure, if no IBGP session is deployed on that link, the entire reflection network will not be affected. If only one IBGP session is deployed on that link, that IBGP session will fail with probability p_e (because the underlying TCP connection could remain alive as a result of IP re-routing). In general cases, k IBGP sessions are deployed on physical link l. Denote τ as the inter-arrival time of IBGP failure events triggered by link l.

$$
\begin{aligned}
P(\tau > t) &= \sum_{i=0}^{\infty} \frac{(w_l t)^i}{i!} e^{-w_l t} \cdot \left(1 - p_e\right)^{ik} \\
&= e^{-w_l \left[1 - (1-p_e)^k\right] t}
\end{aligned}
\tag{1}
$$

Therefore, we have the following lemma.

Lemma 1 *If k IBGP sessions are deployed on physical link l, which has the failure rate w_l, the IBGP session failure event which is triggered by l is a Poisson process, and the rate is $w_l \left[1 - (1 - p_e)^k\right]$ for $k > 0$; otherwise, the rate is 0.*

For example, in Fig. 1(b), two IBGP sessions are deployed on link (A, C). The IBGP failure, which is triggered by (A, C), follows Poisson distribution and the rate is $w_{AC}(2p_e - p_e^2)$.

Different physical links or routers cause IBGP failures independently, and thus the aggregated IBGP failure event is a Poisson process. For a given route reflection network G_r, the aggregated failure rate of the entire network is

$$
R_f(G_r) = \sum_{i \in V} v_i + \sum_{\substack{l \in E \\ k_l > 0}} w_l \left[1 - (1 - p_e)^{k_l}\right]
\tag{2}
$$

Hence, the expected lifetime of G_r is

$$
\mathcal{T}(G_r) = 1 / R_f(G_r)
\tag{3}
$$

One thing is important to mention: $R_f(G_r)$ and $\mathcal{T}(G_r)$ are irrelevant to p_r, because any router failure definitely leads to at least one IBGP session failure, which means an IBGP failure.

(2) *Expected Session Loss $\mathcal{L}$*

ELT evaluates the time interval between two adjacent IBGP failures. An orthogonal metric is the amount of losses in one IBGP failure. We define the session loss as the ratio of the number of failed IBGP sessions to the total number of IBGP sessions in the reflection network. Because any BGP routing information change is propagated via one or several IBGP sessions jointly in a domain, a higher session loss means a larger damage to IBGP routing.

In a given reflection network G_r, suppose m is the total number of IBGP sessions. Without loss of generality, let us assume that router i possesses h_i IBGP sessions. Meanwhile, another k_i IBGP sessions, owned by other routers, are passing i. Then, the expected session loss caused by router i in unit time is

$$
\mathcal{L}(i) = \frac{h_i + p_r k_i}{m} v_i
\tag{4}
$$

Similarly, if k_l IBGP sessions pass physical link l, the expected session loss caused by l in unit time is

$$
\mathcal{L}(l) = \frac{p_e k_l}{m} w_l
\tag{5}
$$

Therefore, we define the expected session loss of the entire reflection network, $\mathcal{L}(G_r)$, as the maximum expected session loss of all physical links and routers, i.e.,

$$
\mathcal{L}(G_r) = \max_{j \in V \cup E} \mathcal{L}(j)
\tag{6}
$$

4 Reliable Reflection Topology Design

4.1 Problem Formulations

The reliability-aware route reflection topology design problem has multiple flavors, depending on different reliability evaluation metrics. We focus on the two metrics, proposed in previous sections, for topology optimization.

Problem 1 (Reliable Reflection – ELT (RR-ELT))
In network $G(V, E)$, given (1) $\{c_i\}$, the upper bound of the node degree; (2) p_e, the impact of physical link failure on IBGP session; and (3) all IGP paths, find a reflection network $G_r^*(V, E_r)$, such that (1) $h_i \leq c_i$, for $\forall i \in G_r^*.V$; (2) $\mathcal{T}(G_r^*) \geq \mathcal{T}(G_r)$, for any $G_r(V, E_r)$ which satisfies $h_i \leq c_i$ for $\forall i \in G_r.V$.

Problem RR-ELT aims to maximize the expected lifetime of IBGP operation, i.e., the time interval between two adjacent IBGP failures. However, the IBGP routing loss caused by session failures is not considered. Thus, it might result in a topology which has a low IBGP failure rate, but a large number of IBGP sessions would be broken once an IBGP failure indeed occurs (e.g., the single point of failure problem to the extreme).

For this reason, we can minimize the ESL of the reflection network to optimize the impact of single IBGP failure, and we call this optimization problem *RR-ESL*. However, RR-ESL is based on the min-max optimization, i.e., it optimizes only for the worst component in the network. Situations at other components are not reflected explicitly in this metric at all. Therefore, we combine these two metrics together to define a new optimization problem: minimize ESL first and use ELT to break ties. The problem is stated as follows.

Problem 2 (Reliable Reflection – ESL/ELT (RR-SLL)) *In network $G(V, E)$, given (1) $\{c_i\}$, the upper bound of the node degree; (2) p_e and p_r, the impact of physical link failure and router failure on IBGP session; and (3) all IGP paths, find a reflection network $G_r^*(V, E_r)$, such that (1) $h_i \leq c_i$, for $\forall i \in G_r^*.V$; (2) $\mathcal{L}(G_r^*) \leq \mathcal{L}(G_r)$, for any $G_r(V, E_r)$ which satisfies $h_i \leq c_i$ for $\forall i \in G_r.V$; (3) $\mathcal{T}(G_r^*) \geq \mathcal{T}(G_r)$, for any $G_r(V, E_r)$ which satisfies $\mathcal{L}(G_r^*) = \mathcal{L}(G_r)$.*

In problems of RR-ELT and RR-SLL, p_e and p_r are two tunable parameters. They are decided by domain administrators to reflect the reliability of an IBGP session during the failure of a physical link or a router on which the session is deployed. $\{c_i\}$ is decided based on the reliability and work load of each router.

To solve the RR-ELT and RR-SSL problems, many design issues have to be considered, such as the number of clusters. In [17], we investigate two special cases, full mesh networks and circle networks, and show that the optimum reflector number is influenced by the redundancy of physical network. In a network with large redundancy, RR-ELT problem favors few reflectors; RR-SLL problem tends to use a large number of reflectors. Increasing the number of reflectors does not necessarily improve IBGP reliability. Besides network redundancy, reflector number is also affected by the maximum number of IBGP sessions a router can have ($\{c_i\}$ constraints) and the reliability of network components, which will be shown in the following sections.

4.2 Solvability Conditions

The node degree upper bound $\{c_i\}$ constrains the possible reflection topology that can be constructed. Thus, we need to investigate the solvability conditions of the reflection topology design problems, i.e., the range of c_i to guarantee the existence of a reflection topology for a given physical network.

Let us assume that the node degree constraint set $\{c_i\}$ is sorted into an ordered sequence $c_1' \geq c_2' \ldots \geq c_n'$. The solvability conditions for the reflection topology design problems are stated in the following theorem.

Theorem 1 *The sufficient and necessary conditions for the reflection topology design problems to be solvable are*

(a) $\sum_{i=1}^{k} c_i' \geq k^2 - 2k + n$,
    where $k = \max\{i \mid c_i' \geq 2i - 3, i = 1, 2, \ldots, n\}$.

(b) $c_n' \geq 1$.

Proof: Please refer to [17]. ∎

Discussions: (1) Theorem 1 provides a convenient method to check if the given degree bounds are feasible. First, each degree bound should be no less than 1. Second, sort the bounds decreasingly, find the smallest c_i' which is greater than or equal to $2i - 3$, and then check if condition (a) is satisfied. The time complexity to do so is $\Theta(n \log n)$.

(2) A special case is that all nodes have a uniform degree bound c. Applying Theorem 1, we have $k = \lfloor (c + 3)/2 \rfloor$, and $kc \geq k^2 - 2k + n$ should be satisfied. By solving this inequity, we have the following corollary. (In our previous work [16], we have similar results.)

Corollary 1 *If all nodes have a uniform node degree constraint, i.e., $c_i = c$, the necessary and sufficient conditions for the reflection topology design problems to be solvable are: if c is an even number, $c \geq \lceil 2\sqrt{n} - 2 \rceil$; otherwise, $c \geq \lceil \sqrt{4n + 1} - 2 \rceil$.*

(3) Without the node degree constraints, the number of reflectors, denoted as n_r, can be any number from 1 to n. However, the range of n_r may be limited to a smaller interval if we are given an ordered degree constraints, say, $c_1', c_2', \ldots, c_n'$. Theorem 1 can be further extended to calculate the range of n_r. Basically, in order to find the range of n_r, we only need to check the inequity condition (a) and $c_k' \geq k - 1$ for every possible k from 1 to n. Fig. 4.2 shows an example of finding the range of n_r based on certain degree constraints. The condition (a) requires the curve $\sum_{i=1}^{k} c_i'$ be above the curve $k^2 - 2k + n$, which in turn gives the feasible interval $n_r \in [3, 33]$. Meanwhile, the curve c_k' should be above the curve $k - 1$, resulting in another interval $n_r \in [1, 30]$. Combining these two intervals, we have that the range of n_r should be $[3, 30]$. This result is very helpful for solving the reflection topology design problems, because it provides the searching range of n_r to heuristic algorithms. In the special case where $c_i = c$ for all i, we have

$$n_r \in \left[\frac{c + 2 - Q}{2}, \frac{c + 2 + Q}{2} \right] \qquad (7)$$

where $Q = \sqrt{c^2 + 4c - 4n + 4}$ and $c \leq n - 1$. If c is smaller than the lower bound given in Corollary 1, the above interval is empty.
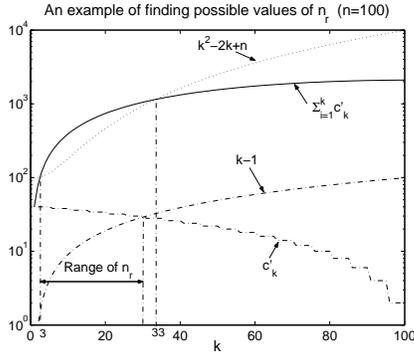
Figure 2. On finding the possible values of n_r .

4.3 Integer Programming Models

In this section, we give the Integer Programming (IP) models for RR-ELT and RR-SLL problems. Suppose $i, j \in V$, and $l \in V \cup E$. Some $0 - 1$ variables are defined as follows to formulate the problem. $f_{ijl} = 1$ indicates IGP path P_{ij} passes l ($l \neq i$ and $l \neq j$); $x_i = 1$ indicates that node i is a route reflector; $r_{ij} = 1$ indicates that node i is the route reflector of node j; $s_{ij} = 1$ indicates that i and j are connected by an IBGP session; otherwise those binary variables are 0.

The following constraints come from the structure of a two-level reflection topology:

$$\sum_{i \in V, \, i \neq j} r_{ij} = 1 - x_j \qquad \forall j \in V \tag{8}$$

$$x_i \geq r_{ij} \qquad \forall i \in V \tag{9}$$

$$s_{ij} \geq r_{ij} + r_{ji} \quad \forall i, j \in V, i \neq j \tag{10}$$

$$s_{ij} \geq x_i + x_j - 1 \quad \forall i, j \in V, i \neq j \tag{11}$$

$$\sum_{j \in V, \, i \neq j} s_{ij} \leq (c_i - 1)x_i + 1 \qquad \forall i \in V \tag{12}$$

Equation (8) guarantees that a client node has exactly one reflector, and a reflector can not be a client of other nodes. Equation (9) ensures that a client can not be a reflector for any other nodes. Equation (10) and (11) states that IBGP sessions must be established between a client and its reflector and also between any two reflectors; the minimization of objective function, which will be discussed later, guarantees no session exists in other cases. Equation (12) ensures that the largest number of IBGP sessions that a node establishes can not exceed node degree upper bound.

The number of IBGP sessions that are deployed on a router or a physical link can be computed from Equation (13).

$$u_l = \sum_{i, j \in V} s_{ij} f_{ijl} \quad \forall l \in V \cup E \tag{13}$$

In the RR-ELT problem, the target is to maximize $\mathcal{T}(G_r)$. According to Formula (3), this is equivalent to minimize $R_f(G_r)$. Because $\sum_{i \in V} v_i$ is a constant, we only need to minimize the second term of Formula (2), and let us denote it as R_f^e. Thus, the objective function is

$$\min_{G_r} R_f^e = \min_{G_r} \sum_{l \in E} w_l \left[1 - (1 - p_e)^{u_l} \right] \tag{14}$$

This is a nonlinear function. In a special scenario, where $p_e = 1$, any physical link failure will terminate all related IBGP sessions. Based on Formula (2), the optimization target is

$$\min_{G_r} \sum_{l \in E} w_l u_l' \tag{15}$$

where u_1' is a binary variable and satisfies the following equation:

$$u_l' \geq \sum_{i, j \in V} s_{ij} f_{ijl} / n^2 \quad \forall l \in V \cup E \tag{16}$$

That is, u_l' equals 1 if any IBGP session passes l; otherwise it is 0. Thus, under condition $p_e = 1$, RR-ELT optimization problem is converted into a linear integer programming model (Formula 8-12, 16 and 15), and many off-the-shelf integer programming optimizers, such as CPLEX, can be applied.

In the RR-SLL problem, besides the constraints from Formula (8)-(13), we have the following additional constraints:

$$n' = \sum_{i \in V} x_i \tag{17}$$

$$m = n + n'(n' - 3)/2 \tag{18}$$

$$g \geq v_k \left(p_r u_k + \sum_{j \in V, k \neq j} s_{kj} \right) / m \quad \forall k \in V \tag{19}$$

$$g \geq p_e v_l u_l / m \quad \forall l \in E \tag{20}$$

In Equation (17), n' is the number of reflectors. Given n', the number of IBGP sessions is $m = \binom{n}{2} + n - n'$, and thus we have Equation (18). Equations (19) and (20) are used to define the expected session loss. In RR-SLL problem, minimizing ESL has higher priority than minimizing R_f^e. The optimization target is thus defined as follows:

$$\min_{G_r} \left(\eta g + \epsilon R_f^e \right) \tag{21}$$

where η should be a big number and ϵ is a small number. From the computational experiments in Section 7.3, we will show that the method of weighted aggregation of R_f^e and ESL is effective in finding topologies with better ELT metric while keeping the optimality of ESL in the meantime.

In the Integer Programming models for RR-ELT and RR-SLL, some of the constraints and objective functions are nonlinear. Intuitively, we can not find an optimum solution efficiently. Actually, we will prove in Section 5 that even in the simplest case of RR-ELT, where $p_e = 1$, the optimization problem is still NP-hard. Therefore, to solve the reliable reflection topology design problem relies on finding some efficient heuristic algorithms, which will be presented in Section 6.

5 Complexity Analysis

In this section, we formally prove that the RR-ELT problem is NP-hard, even in a very simplified case. The NP-hardness of the RR-SLL problem directly follows. The following is the proof sketch: first, we define a special case of the RR-ELT problem, called the *sELT* problem; second, we prove that the sELT problem is NP-hard by reducing it from a known NP-hard problem – the Facility Location Problem; finally, we conclude that the RR-ELT problem is NP-Hard, too.

The sELT problem is a special case of the RR-ELT problem that satisfies the following three additional conditions: (1) $p_e = 1$; (2) $c_i = n$, for $\forall i \in V$, i.e., there is no node degree constraint; and (3) the physical links form a full mesh over all routers, and no IGP paths share common physical links. In the sELT problem, each IBGP session is deployed on only one physical link, and the weight of an IBGP session is defined as the link's failure rate. Thus, the objective function can be converted to $\min \sum_{l \in E_r} w_l$.

Uncapacitated Facility Location Problem (UFL): $\mathcal{F}$ is a set of n_f potential facilities, and $\mathcal{D}$ is a set of n_d clients. For any $i \in \mathcal{F}$, a fixed nonnegative cost f_i is given as the opening cost of facility i. For every client $i \in \mathcal{D}$ and facility $j \in \mathcal{F}$, there is a connection cost c_{ij} between client i and facility j. The problem is to open a subset of the facilities of $\mathcal{F}$, and assign every client to an open facility such that the total cost, including the opening cost and the connection cost, is minimized. That is $\min_{F' \subseteq \mathcal{F}} \left[\sum_{i \in F'} f_i + \sum_{i \in \mathcal{D}} \min_{j \in F'} c_{ij} \right]$. Cornuéjols, Nemhauser and Wolsey [7] prove the UFL problem to be NP-hard by reducing it from the node cover problem .

If we require that exactly k IBGP routers be selected as the reflectors in the sELT problem, we get the k-sELT problem, where $0 < k \le n$. Similarly, if the number of opened facilities is required to be k, we have the k-UFL problem, where $0 < k \le n_f$.

Lemma 2 *k-UFL problem is NP-hard.*

Proof: By contradiction. Let us assume that there is an algorithm that can solve k-UFL in polynomial time. Because the number of facilities n_f is finite, we can also find the optimal solution for UFL in polynomial time by solving k-UFL with $k = 1, 2, \ldots, n_f$, respectively. While, it is unable to find a solution for UFL in polynomial time, unless P=NP. Therefore, k-UFL has to be NP-hard. ∎

Lemma 3 *k-sELT problem is NP-hard.*

Proof: We prove the NP-hardness of k-sELT by reducing k-UFL problem to k-sELT problem. From the k-UFL problem, we construct a graph, as shown in Fig. 3, to form the k-sELT problem. We define sets $\mathbf{F} = \{F_i \,|\, 1 \le i \le n_f\}$ and $\mathbf{D} = \{D_i \,|\, 1 \le i \le n_d\}$ to be set of facilities and set of clients respectively. An auxiliary node T is also introduced. The link weights of the graph are set as follows: (1) For any $F_i \in \mathbf{F}$, the weight between T and F_i is $\frac{f_i}{M}$, where M is a very large positive number; (2) The weight between T and any $D_i \in \mathbf{D}$ is infinity; (3) For any link between $F_i, F_j \in \mathbf{F}$, the weight is $\frac{f_i + f_j}{k-1}$; (4) The weight of a link between any two nodes in $\mathbf{D}$ is infinity; (5) For any $F_j \in \mathbf{F}$ and $D_i \in \mathbf{D}$, the weight is c_{ji}.

Based on the construction in Fig. 3, we can easily map the k-UFL problem into the $(k+1)$-sELT problem. The following verifies that this mapping is valid. Because the weight between any two nodes in $\{T\} \cup \mathbf{D}$ is infinity, at most one node from $\{T\} \cup \mathbf{D}$ can be chosen as the reflector. The other k or $k+1$ reflectors are from $\mathbf{F}$. We choose M large enough, i.e., $M \gg$
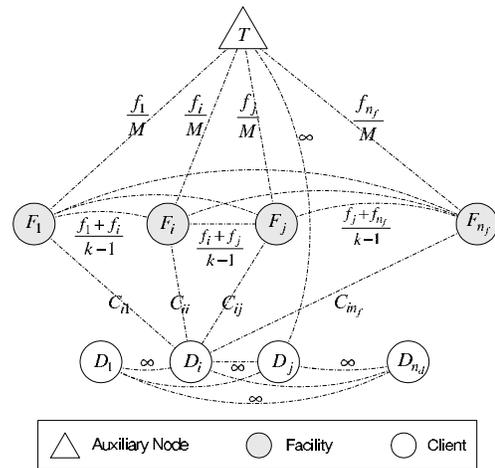

Figure 3. Reduction from k-UFL to k-sELT.

$\max \left(\max_{i,j,l} \frac{f_l}{c_{ij}}, \ \max_{i,j,l} \frac{(k-1)f_l}{f_i + f_j} \right)$, such that T is guaranteed to be a reflector. The other k reflectors are chosen from $\mathbf{F}$.

Because the weight between T and any node in $\mathbf{D}$ is infinity, the nodes in $\mathbf{D}$ can only be assigned to the reflectors in $\mathbf{F}$. Likewise, due to the large M, the nodes in $\mathbf{F}$ that are not reflectors are assigned to the reflector T. Therefore, the total weight of the reflection graph, γ, is

$$
\begin{aligned}
\gamma &= \text{weight of reflector mesh} + \text{weight of client connections} \\
&= \sum_{i,j \in \mathcal{R}} w_{ij} + \sum_{i \in \mathcal{C}} \min_{j \in \mathcal{R}} c_{ij} \\
&= \min_{\substack{F' \subseteq \mathbf{F} \\ |F'| = k}} \left(\sum_{i \in \mathbf{F}} f_i / M + \sum_{i \in F'} f_i + \sum_{i \in \mathbf{D}} \min_{j \in F'} c_{ij} \right) \\
&= \min_{\substack{F' \subseteq \mathbf{F} \\ |F'| = k}} \left(\sum_{i \in F'} f_i + \sum_{i \in \mathbf{D}} \min_{j \in F'} c_{ij} \right) + C \quad (22)
\end{aligned}
$$

where $C = \sum_{i \in \mathbf{F}} f_i / M$ and C is a constant. From equation 22, the $(k+1)$-sELT problem has the same optimization function as the k-UFL problem. Therefore, we have reduced the k-UFL problem to the k-sELT problem. By lemma 2, we know that the k-sELT problem is indeed NP-hard. ∎

Theorem 2 *sELT problem is NP-hard.*

Proof: Let us assume that we know the number of the reflectors of the optimal solution for a sELT problem. According to lemma 3, we still can not solve the sELT problem in polynomial time based on this additional information, unless P=NP. Thus, the sELT problem is also NP-hard. ∎

Finally, since the sELT problem is a special case of the original RR-ELT problem, we know the the RR-ELT problem is NP-hard.

6 Heuristic Solutions

The RR-ELT problem and the RR-SLL problem both aim to minimize certain cost functions by designing the reflection

topology appropriately. The cost functions are defined in Equation 14 and Equation 21, respectively. As has been shown in Section 5, the classic facility location problem can be reduced to a special case of RR-ELT problem. However, the techniques and results from the facility location problems and their variants can not be easily applied to our reflection topology optimizations. There are several obstacles: The costs of different sessions are not independent and they also do not necessarily satisfy triangle inequality condition; reflector (facility) setup cost is not constant but related to other selected reflectors; The optimization target is nonlinear in the general case of RR-ELT and RR-SLL problems. In this paper, we develop and implement three heuristic solutions to these non-linear optimization problems.

6.1 Heuristic Selection

Heuristic Selection (HS) algorithm imitates the manual reflection topology design, but it repeats for many iterations to improve quality of the solutions. Before optimization, we generate three ordered lists of routers: R_r, R_h and R_d. R_r is ranked increasingly based on the router failure rates; R_h is ranked increasingly according to the sum of the hop-counts from a router to all other routers; R_d is ranked decreasingly according to node degree upper bounds $\{c_i\}$. In optimization process, first, the number of reflectors, n_r, is randomly selected from the possible range (see Section 4.2). Second, from the first $1.3 \cdot n_r$ elements of R_r, n_r elements which have enough c_i are randomly selected as reflectors. Third, we connect each of the remaining routers (clients) to the reflector which has the most reliable IGP path to it. If a topology can not be generated due to the $\{c_i\}$ constraints, we try R_h and R_d in sequence. The above optimization process is repeated for n^2 iterations and returns the best result, where n is the number of routers.

6.2 Greedy Search

Greedy Search (GS) is a randomization algorithm which explores the neighborhood structures in a greedy manner to find better solutions. Based on an initial topology, GS enumerates its neighboring solutions until a new topology with lower cost is found. Then, the new solution is accepted as the base for neighborhood searching of the next iteration. This process is executed iteratively for many times.

The neighborhood structure of reflection topology is defined as follows. From a route reflection topology, a neighbor solution can be generated by changing reflectors or changing clients:

(1) Changing Clients: The clients of a reflection topology can be changed in two ways: (a) A randomly selected client node is moved from its previous reflector to another reflector whose IBGP session number does not exceed the node degree constraint $\{c_i\}$; (b) Two clients i and j, which are previously connected to different reflectors r_i and r_j, are chosen to swap their reflectors, i.e., i is now assigned as r_j's client and j is assigned to r_i.

(2) Changing Reflectors: The methods of changing reflectors include three types: (a) A former client i is selected to be a new reflector, a previous reflector t is changed into a client, and t's previous clients are assigned to i. If i's node degree limit prohibits accommodating all previous clients of t together, the extra clients are assigned to other possible reflectors which have the most reliable IGP paths between them; (b) A randomly selected reflector i is changed into a client. i and the previous clients of i are redistributed to the remaining reflectors based on the reliability of IGP paths; (c) A client i becomes a new reflector. Suppose the previous reflector of i is r_i. The clients of r_i are redistributed among i and r_i based on the reliability of IGP paths. Notice that deleting reflectors or adding reflectors (the second type or the third type) must not violate the condition on the reflector number, which is described in Fig. 4.2 of Section 4.2.

Changing reflectors has larger impact on reflection topologies than changing clients. Let δ denote the probability of changing reflectors when generating a neighbor. In our experiments (Section 7), we choose δ from $[0.05, 0.4]$ with respect to different optimization problems.

6.3 Simulated Annealing

Simulated annealing [1] algorithm (SA) is based on GS algorithm, but it can find high quality solutions for large combinatorial optimization problems.

Because GS only accepts a neighbor solution that has lower cost than the current solution, GS can easily get stuck into local minima, and its performance is highly sensitive to the initial topologies. On the other hand, in addition to accepting better solutions, SA can also accept deteriorated solutions (i.e., the neighbors with higher cost) intentionally with probability $\exp\left(-|\Delta C|/D\right)$, where ΔC is the cost difference and D is a control parameter. We decrease this probability by decreasing D as the number of searching iterations increases. This controlled randomization searching strategy has better opportunity to search larger solution space without getting trapped into some local minima, which makes SA not sensitive to the choice of initial solutions. Due to space limitations, we skip detailed descriptions on SA (interested readers can refer to [1]). We briefly present the configurations of SA algorithms as follows for our optimization problems.

The neighborhood structure of SA is the same as in GS. After enough neighborhood structures has been probed (n^2 iterations, where n is the number of nodes), we decrease D exponentially, i.e., $D_{\text{new}} = \alpha D_{\text{old}}$, where α is picked from $[0.75, 0.95]$. The initial value D_0 should be large enough, so that almost any neighbor solution is acceptable. We define P_0 as the acceptance probability in the initial n^2 iterations, which can be used to calculate the required D_0. Suppose the cost difference between the best solution and the worst solution, in the initial n^2 iterations, is ΔC_0. Then D_0 can be computed by $\frac{-|\Delta C_0|}{\ln P_0}$. Notice that P_0 should be a value close to 1. In our experiments, P_0 is 0.999.

7 Computational Experiments

In this section, we evaluate the optimization results of RR-ELT, RR-SLL and RR-ESL problems by computational experiments. Three optimization methods, HS, GS and SA, are implemented.

Network topologies are generated using BRITE network topology generator[11]. The Waxman model is used and nodes are placed according to the heavy-tail distribution. The failure rates of physical links, $\{w_{ij}\}$, are generated randomly from interval $[0.1, 2.0]$. The failure rates of routers, $\{v_i\}$, are generated randomly from $[0.01, 1.0]$. Four network topologies are generated, which have 50, 80, 120 and 150 nodes respectively. In each topology, 20% of the nodes are selected as border routers. If node i is a border router, c_i is randomly selected from $[1, 10]$; otherwise c_i is generated randomly from $[2\sqrt{n}+2, 2\sqrt{n}+17]$. We assume that the minimum hop-count routing is used in IGP. The impacts of link failures and router failures on IBGP sessions are: $p_e = 0.1$ and $p_r = 0.2$, respectively.

The initial reflection topology for GS and SA is generated by using one iteration of heuristic selection algorithm. For each topology, SA and GS are executed for 10 times with different randomization seeds. From the results of these 10 executions, the minimum, maximum and mean values are calculated for performance comparison in the best, worst and average cases, respectively. Furthermore, GS runs the same number of iterations as SA does in each case, making their optimization results comparable.

7.1 Optimization Results of RR-ELT Problem

RR-ELT aims at minimizing the IBGP failure rate. The cost function of RR-ELT is R_f^e (in Equation 14), which is the rate of IBGP failures that are caused by link failures. Table 1 shows the optimization results obtained by SA, GS and HS algorithms. The probability of changing reflectors, δ, is within $[0.2, 0.4]$ in SA and GS algorithms.

n	SA			GS			HS
	Min	Mean	Max	Min	Mean	Max	
50	12.91	13.10	13.42	13.10	13.93	15.54	13.50
80	22.93	23.95	24.47	23.48	24.54	26.72	24.17
120	38.42	39.84	40.40	38.88	41.69	43.80	42.99
150	46.13	46.58	47.44	46.51	48.78	52.75	50.70

Table 1. R_f^e optimized by SA, GS and HS.

The data shows that SA has the best optimization result among the three. The costs of SA in the worst case are lower than the results of HS (except one case) and even lower the average cost of GS. HS, which stands for the manual heuristic solution, has the worst performance. When the node number n is large, the result of HS is about 10% worse than the results of SA. This number seems not so significant. The reason is that, in RR-ELT problem, R_f^e is largely determined by the reflectors selected. HS searches the reflector structure intensively and connects a client to a reflector using the most reliable IGP path. However, HS could result in a reflection topology whose expected session loss is large. We will show next that, in terms of session loss metric, HS performs much worse than other algorithms.

7.2 Optimization Results of RR-SLL Problem

The RR-SLL problem combines both ESL and ELT to design the reflection topology. The cost function is $\eta\mathcal{L}(G_r) + \epsilon R_f(G_r)$. In our experiments, $\eta = 100$ and $\epsilon = 0.00001$. The optimization results are shown in Table 2. In the experiments, the probability of changing reflectors, δ, is $[0.05, 0.2]$.

n	SA			GS			HS
	Min	Mean	Max	Min	Mean	Max	
50	2.461	3.021	3.249	2.990	5.119	7.165	10.85
80	1.854	2.176	2.746	3.466	5.901	17.66	7.412
120	1.220	1.363	1.759	1.886	3.440	10.56	5.908
150	0.977	1.230	1.643	1.626	3.075	4.625	5.510

Table 2. RR-SLL solved by SA, GS and HS.

SA performs much better than GS and HS. The mean costs of SA are smaller than the corresponding results of HS, and even smaller than the minimum costs of GS, except the case of 50 nodes. The mean cost achieved by SA is about 50%-60% smaller than that of GS. The optimization results from HS, which imitates the manual heuristic configurations, are $340\% - 460\%$ worse than the best results of SA, especially when the network is large. These numbers demonstrate the significant benefits from designing route reflection network appropriately.

GS has better results than HS does. However, in some cases ($n = 80, 120$), the maximum cost of GS is worse than HS. This is because GS may get trapped into some local minima due to the greedy search strategy. Not surprisingly, the experimental results show that GS is sensitive to the initial conditions. In four cases, the differences between the minimum and maximum costs in GS are 4 to 15 times larger those that in SA.

7.3 Comparison of RR-SLL and RR-ESL

RR-ESL problem purely optimizes the ESL metric. RR-SLL, keeping the optimal ESL, aims at finding better reflection topology with as small R_f as possible. We apply SA algorithm to these two problems. The results show that SA finds approximately the same ESL values in the RR-SLL problem as in the RR-ESL problem. The difference of the average costs between the two problems is below 5%.

Also, it is very meaningful to optimize ELT without degrading ESL. Table 3 compares the ELT results between RR-SLL problem and RR-ESL problem. In RR-SLL, SA achieves better ELT than in RR-ESL. For example, in the case of $n = 120$, the mean R_f, achieved in RR-SLL, is about 12% smaller than the mean R_f of RR-ESL. The RR-ESL problem only optimizes the worst component in the reflection network, and other parts may not be well designed in terms of ELT metric. Thus, RR-SLL has enough solution space for searching for topologies with better ELT metric and simultaneously keeping optimal ESL.

n	RR-SLL: $R_f(G_r)$			RR-ESL: $R_f(G_r)$		
	Min	Mean	Max	Min	Mean	Max
50	42.90	44.68	46.47	44.20	46.73	50.65
80	74.63	79.20	85.82	86.25	88.60	91.91
120	125.07	140.13	150.58	151.55	158.89	167.40
150	143.21	174.05	195.19	167.45	193.83	214.24

Table 3. Comparisons of RR-SLL and RR-ESL.

7.4 Optimum Reflector Numbers vs. Node Degree Constraints

This experiment uses the network with 80 nodes to show the impact of node degree constraints $\{c_i\}$ on the optimum number of reflectors. For convenience, all nodes have a uniform node degree upper bound c_i. SA is used to solve the optimization problems.
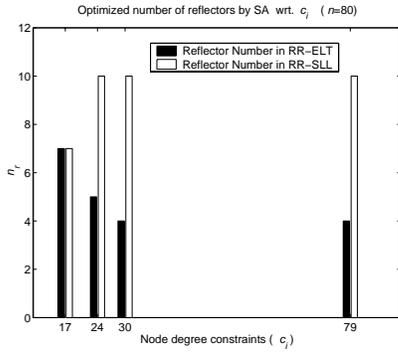

Figure 4. Optimized number of reflectors wrt. node degree constraints.

Fig. 4 displays the relationship between the optimized n_r and c_i in the RR-ELT problem and the RR-SLL problem. Because the network has 80 nodes, the minimum value of c_i is 16 (refer to Corollary 1 in Section 4.2). When c_i is small (e.g. $c_i = 17$), it tightly determines the possible values of n_r. When c_i increases, the solution space becomes larger. RR-ELT optimization tends to use smaller number of reflectors, while RR-SLL optimization favors larger number of reflectors. When c_i is large enough, in RR-ELT, n_r converges to 4; in RR-SLL, n_r converges to 10. The optimum reflector number does not converge to 1 or 80, because of the influence from network redundancy and specific reliability of routers and links. This shows that too large or too small reflector numbers do not necessarily lead to a reliable reflection topology.

8 Conclusion

In this paper, we address the reliable reflection topology design problem. This problem is of great importance to increase the reliability of IBGP operation and decrease the route flaps and packet forwarding errors caused by IBGP session failures. Our experiments show that by optimizing IBGP reliability, we can find much better reflection topologies than the results from manual heuristic design, especially in terms of the session loss metric. There are three major contributions of this paper: First, we present a reliability model for IBGP and formulate the reliable reflection topology design problem; Second, we prove the design problem to be NP-hard, and give the solvability conditions for the problem; Third, we implement heuristic solutions based on randomization algorithms, and also give hints on how to configure a reliable reflection topology.

References

[1] E. Aarts and J. Korst. *Simulated Annealing and Boltzmann Machines*. John Wiley and Sons Ltd., 1989.

[2] T. Bates, R. Chandra, and E. Chen. *BGP Route Reflection - An Alternative to Full Mesh IBGP. RFC 2796*. Network Working Group, April 2000.

[3] D.-F. Chang, R. Govindan, and J. Heidemann. An empirical study of router response to large BGP routing table load. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, 2002.

[4] Cisco Systems Inc. Troubleshooting high cpu utilization on cisco routers. In *http://www.cisco.com/warp/public/63/highcpu.html*.

[5] Cisco Systems Inc. Troubleshooting memory problems. In *http://www.cisco.com/warp/public/63/mallocfail.shtml*.

[6] Cisco Systems Inc. Troubleshooting router hangs. In *http://www.cisco.com/warp/public/63/why_hang.html*.

[7] G. P. Cornuéjols, G. L. Nemhauser, and L. A. Wolsey. The uncapacitated facility location problem. In *Discrete Location Theory*, pages 119–171, 1990.

[8] R. Dude. A comparison of scaling techniques for BGP. *ACM Computer Communication Review*, 29(3), Jul. 1999.

[9] S. Halabi and D. McPherson. *Internet Routing Architectures*. Cisco Press, 2000.

[10] G. Iannaccone, C. nee Chuah, R. Mortier, S. Bhattacharyya, and C. Diot. Analysis of link failures in an IP backbone. In *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, 2002.

[11] A. Medina, A. Lakhina, I. Matta, and J. Byers. Boston university representative internet topology generator. In *http://cs-www.bu.edu/brite/*.

[12] Packet Design, Inc. BGP scalable transport. In *http://www.packetdesign.com/company/bst.html*.

[13] Y. Rekhter and T. Li. *A Border Gateway Protocol 4 (BGP-4). RFC 1771*. Network Working Group, March 1995.

[14] S. R. Sangli, Y. Rekhter, R. Fernando, J. G. Scudder, and E. Chen. *Graceful restart mechanism for BGP. Internet Draft draft-ietf-idr-restart-05.txt*. Network Working Group, June 2002.

[15] A. Shaikh, A. Varma, L. Kalampoukas, and R. Dube. Routing stability in congested networks: Experimentation and analysis. In *Proceedings of ACM SIGCOMM*, 2000.

[16] L. Xiao, J. Wang, and K. Nahrstedt. Optimizing IBGP route reflection network. In *Proceedings of IEEE International Conference on Communications*, 2003.

[17] L. Xiao, J. Wang, and K. Nahrstedt. Reliability-aware IBGP route reflection topology design. Tech report, Department of Computer Science, University of Illinois, UIUCDCS-R-2003-2375, August 2003.

[18] J. Yu. *Scalable Routing Design Principles. RFC 2791*. Network Working Group, July 2000.