Reliability of Internal BGP Networks: Models and Optimizations

Li Xiao, Klara Nahrstedt, and Jun Wang
Department of Computer Science
University of Illinois at Urbana-Champaign, Urbana, IL 61801

***Abstract*— A reliable routing control plane in Internal Border Gateway Protocol (IBGP) is very important for achieving dependable Internet data communication. However, the reliability modeling of IBGP and the design of reliable IBGP route reflection networks, which are of great importance to increase the robustness of IBGP operations, have not been well investigated.**

The reliability analysis of IBGP networks, which are overlaid on top of IP networks, is challenging, because failures of IBGP sessions may be correlated through the shared IGP routes. In this paper, we first present a model for the reliability analysis in IBGP networks to characterize correlated failures, followed by two metrics to measure the reliability of IBGP. Then, we investigate the optimization problems of finding the most reliable IBGP route reflection topologies. We show that the problems in general are NP-hard and an optimization bound is thus provided. Moreover, we develop efficient algorithms for searching satisfactory near-optimal topologies in general scenarios, as well as the optimal solutions in some special networks. Our study shows that route reflection topologies considerably influence the reliability of IBGP operations. By applying our models and optimization techniques, a route reflection topology can be appropriately configured and the IBGP robustness can be improved significantly.

I. Introduction

Border Gateway Protocol (BGP) [1] is the widely used interdomain routing protocol. Two BGP routers, which communicate with each other directly, are called *BGP peers*. BGP peers exchange routing information via *BGP sessions* which are running over TCP. According to the relation of BGP peers, BGP itself 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), while an IBGP session links two BGP routers which belong to the same AS. In this paper, we focus on reliability modeling for the *IBGP network* which consists of BGP routers in one AS and the IBGP sessions among them.

A. Internal BGP Networks

In a traditional IBGP network, IBGP sessions form a full mesh over all BGP routers in a domain. A hierarchical IBGP structure, called route reflection [2], was proposed to solve the scalability limitation in the full mesh design. Fig. 1 shows an example of two-level IBGP route reflection network. BGP routers are divided into three clusters. In each cluster, at least one router is chosen as a route reflector (e.g., A, B, E and I in Fig. 1), and other routers are route-reflector clients. In cluster I, redundant reflectors are used for higher reliability. All reflectors establish a full mesh via IBGP sessions. A client is only required to share IBGP sessions with the reflectors in its cluster. The sessions between clients of the same cluster are optional. For example, an optional session may be established between K and J. A client is the traditional BGP router and it only needs to communicate with its reflectors. A reflector is responsible for: (1) reflecting routes (routing information) from its client to the peer reflectors and the other clients; (2) reflecting routes from its peer reflectors to its clients. The CLUSTER_LIST loop detection mechanism prevents routes from being reflected back to the clusters where the routes originate.

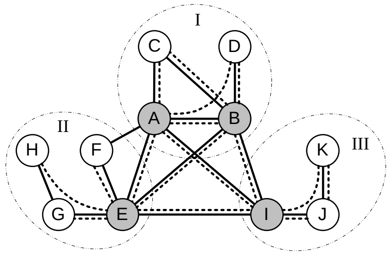


Fig. 1. An example of IBGP route reflection network. Routers are grouped into three clusters. Solid lines stand for IP links; dotted lines stand for IBGP sessions. Shaded nodes represent route reflectors.

Route reflection network and full mesh IBGP network are both overlay networks on top of the underlying IP networks. Each IBGP session is an overlay edge supported by a TCP/IP connection. Some IBGP peers are not adjacent physically. They depend on IGP (Interior Gateway Protocol) to communicate across multiple hops. For example, in Fig. 1, the IBGP session between router A and D is routed along an IGP path through router B.

B. Reliability of IBGP Networks

A reliable IBGP control plane is very important to the quality of Internet routing. When an IBGP session is lost, all related routes in the BGP routing tables have to be withdrawn and thus some IP addresses become unreachable. The route withdrawal messages trigger huge amount of route re-computations and also result in route flaps. It usually takes a long time and lots of network resource to re-establish this lost session.

On the other hand, the reliability and stability of IBGP operations depend on the quality of underlying TCP and IGP routing. It is reported that BGP sessions are sensitive to transport layer stability and routing layer reliability [3][4], especially for the IBGP sessions which cross multiple IP hops. Therefore, understanding the influential factors on IBGP reliability and furthermore finding proper methods to improve IBGP robustness are very crucial for delivering highly available and stable services in Internet routing.

However, so far, issues, such as how to model the reliability of IBGP, which is influenced by IGP routing and transport layer protocols, and how to improve the IBGP reliability within the existing Internet framework, have not been well studied yet. Our previous research [5][6] has shown that TCP provides only a limited support of reliable communication for IBGP sessions and failures in IP networks may make the IBGP sessions fail. In this paper, we focus on analyzing the reliability of the whole IBGP networks given the fact that IBGP sessions are unreliable due to the influence from the underlying IP networks.

The reliability analysis in IBGP networks is complex due to the correlations between different IBGP sessions. For example, in Fig. 1, the IBGP session between H and E is not statistically independent of the session between G and E, because they share one IP link. Cui et al. [7] give an approximate calculation on the probability that two overlay links fail simultaneously. In this paper, we perform a much more extensive study on the reliability of the overlay networks in IBGP. By using the dependent network failure analysis model [8] (based on independent failure-causing events), we investigate the reliability of IBGP networks and propose two novel metrics for network resilience: *IBGP Failure Probability* (IFP) and *Expected Connectivity Loss* (ECL). In a nutshell, IFP defines the probability of *IBGP failures* taking into account conditional failures of IBGP sessions; ECL stands for the average percentage of IBGP router pairs that are isolated in the IBGP control plane.

With the reliability model for IBGP, we can answer many meaningful questions on IBGP reliability and have a quantitative guide in designing reliable IBGP networks. Previously, when we migrated a full mesh IBGP network to a route reflection network, we did not have a precise understanding on the impact of this transition in terms of network reliability. For example, does the network become less reliable because the transition reduces the number of IBGP sessions? How much reliability can we gain by introducing one redundant reflector and how many reflectors are actually needed? How to cluster routers? And, moreover, is it possible to design a route reflection network that is even more reliable than the full mesh IBGP? At present, the only guideline for setting up reflection topology is to follow the physical topology [9]: in large networks, route reflection topology is usually overlaid with IGP hierarchy; routers are clustered according to IGP areas; in each cluster, the core routers are selected as the reflectors, and the others are clients of their core router. But, in general, the reliable IBGP network design problem has not been studied in the literature. In this paper, the IBGP reliability optimization problems are formulated and studied based on the reliability models and metrics we propose.

We aim to increase BGP routing reliability by configuring IBGP route reflection networks appropriately without changing the standard of protocols, so that the rate and impact of IBGP failures are minimized. The goal of design is to minimize IFP or ECL. Our previous work [10] has addressed a simplified design problem, in which redundant reflectors are not considered. In this paper, we study the impact of redundant IBGP components and how to find the reliable IBGP networks in general scenarios. Specifically, we formally prove that this optimization problem is NP-hard, even in some simplified cases; we develop several efficient algorithms to search for satisfactory near-optimal solutions; an optimization lower bound is also given to show the room for improvement. Our extensive computational experiments demonstrate that based on the realistic Internet network topologies, we can find both reliable and efficient solutions for IBGP networks, i.e., the solutions are very close to the global optima and have small number of IBGP sessions involved. In addition, we investigate the optimization problem in a special case – fully mesh IP networks, which has strong application background and has simplified solutions.

Besides the models and the optimization algorithms, we have the following key insights into the design of reliable IBGP route reflection networks. (1) Reliable IBGP network design highly depends on the topology of IP networks, as well as the specific reliability of each IP network component. Even in IP networks where all routers and links have the same reliability, some appropriately designed IBGP networks are much more reliable than other options. (2) It is not always true that the IBGP network with more redundant IBGP reflectors or more IBGP sessions is more reliable. The redundant elements in IBGP should be used properly. (3) Full mesh IBGP is often not the best topology in terms of reliability, even though it has maximum number of IBGP sessions and is traditionally used in small networks. In some cases, we can design a more reliable IBGP network by clustering routers and placing reflectors appropriately.

The rest of the paper is organized as follows. In Section II, we define the network models and describe the failures in IBGP. In Section III, we present a reliability analysis framework for IBGP and define two reliability metrics. In Section IV, the methods for calculating the metrics are discussed, followed by case studies to demonstrate the influential factors on IBGP network reliability. Furthermore, we formulate the reliability optimization problems for IBGP networks in Section V and discuss the techniques for solving these problems. In Section VII, related work is presented. Section VIII concludes the paper.

II. Network Models and IBGP Failures

A. Network Models

We denote a typical IP network as graph $G(V, E)$, where V and E are the sets of routers and IP links, respectively. (u, v) represents the IP link from router u to router v. IGP path from u to v is denoted as P_{uv}, which is the set of routers and links on the path.

IBGP Route reflection network $G_r(V_r, E_r)$ is overlaid on top of the IP network, where V_r is the set of IBGP routers and E_r is the set of IBGP sessions. We consider transitive domains where every router runs BGP, and thus $V_r = V$. IBGP session between u and v is denoted by $\langle u, v \rangle$. IGP paths P_{uv} and P_{vu} are used to support this session. We also use P_e to denote the IGP path used by session e, where $e \in E_r$. In a two-level route reflection network, the nodes are grouped into several clusters. Each cluster contains one or multiple route reflectors. E_r includes the full mesh of IBGP sessions among all reflectors and the sessions between clients and their reflectors. IBGP sessions between the clients in one cluster are optional, which can be used to improve the reliability of IBGP networks. In general, route reflection hierarchy can have an 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 this paper, we only discuss two-level reflection networks.

B. Failures in IBGP Networks

In IBGP networks, both BGP routers and IBGP sessions could fail due to the component failures[1] in the supporting IP networks.

1) Failures of BGP routers: Different routers, which have a variety of software or hardware platforms, may show different reliability in hosting BGP routing function. 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 by memory segmentation) or extremely high CPU utilization will cause a router to hang[15][16][17]. Thus, routers with larger amount of resources (CPU power and memory) are more robust for handling BGP operations.

2) Failures of IBGP Sessions: BGP routers detect BGP session failures using a heart-beat mechanism. Each BGP router expects to receive at least one message from every peer in a 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 make BGP messages delayed or lost, such as IP packet forwarding failures, may further cause the related BGP sessions to be reset. Although TCP retransmission can recover the lost packets in case of network failures and congestion, BGP messages may be severely delayed due to the routing recovery time and the TCP retransmissions. If the BGP `Hold Timer` expires because of the long delay, the IBGP sessions are broken. Therefore, with some probability, the failures in IP networks lead to the failures of the BGP sessions.

III. Reliability Analysis for IBGP Networks

A. Analysis Framework

In route reflection network $G_r(V_r, E_r)$, the failures of different edges, i.e., IBGP sessions, are not necessarily independent. Two overlaid sessions may share the same IP routers or links in their IGP paths. There are several approaches to studying the network reliability with dependent component failures. In this paper, we make use of the cause-based reliability analysis model [8] in IBGP networks.

We identify all major *failure scenarios* in IP networks. For example, a failure scenario can be the case in which a single router or a single IP link fails. Different failure scenarios happen independently. Let $\mathcal{S}$ denote the set of all failure scenarios we are interested in plus one special scenario, s_0, where no failure exists. The probability that a scenario happens can be derived from the historical network operation information, and let r_s denote the probability that scenario s happens. If $\mathcal{S}$ includes all network states, $\sum_{s \in \mathcal{S}} r_s = 1$. Moreover, we use F_s to denote the set of components in IP networks that fail in scenario s ($s \in \mathcal{S}$), and F_s is a subset of $V \cup E$. Other components in IP networks, which are not in F_s, work properly.

In scenario s, the IBGP sessions that belong to the routers in F_s fail. If F_s partitions the network, the IBGP sessions crossing two partitioned sub-networks also fail. For example, in Fig. 1, the failure of router G causes sessions $\langle G, E \rangle$ and $\langle H, E \rangle$ to fail. On the other hand, if the IP network is not partitioned, the affected IBGP sessions may or may not be broken, depending on the IGP routing recovery time. We thus model the impact of IP network on IBGP sessions as a probability variable. For example, in Fig. 1, if IP link (A, B) fails, the IBGP sessions $\langle A, D \rangle$ and $\langle A, B \rangle$ fail with a certain probability. In general, we denote q_s as the conditional failure probability of the IBGP sessions that are affected by the IP network failures in scenario s, i.e.,

$$q_s = Pr\left[\text{session } e \text{ fails} \mid F_s \text{ fails and other components are up}\right], \tag{1}$$

where $e \in E_r$, $s \in \mathcal{S}$, and $P_e \cap F_s \neq \emptyset$. q_s is related to the IGP routing recovery time and the configuration parameters in BGP and TCP retransmissions. In [5], we showed that q_s is not significantly affected by the round trip time between the two BGP routers of an IBGP session. Thus, to simplify the following discussion on IBGP route reflection topology optimization problem, we assume that q_s is the same for all influenced IBGP sessions in scenario s. Moreover, in a given failure scenario, all components in IP network $G(V, E)$ are in deterministic states. Each BGP router detects the session failure independently. The BGP timers at different routers are set and shifted independently, too. Thus, the conditional failures of the affected IBGP sessions (with probability q_s) are independent, and we can apply reliability analysis techniques in networks with independent failures.

Generally speaking, the size of $\mathcal{S}$ can be very large if we want to cover all failure scenarios. However, in practice, we can get a satisfying statistical coverage (i.e., $\sum_{s \in \mathcal{S}} r_s$ is very close to 1) by only analyzing the failure scenarios that could most likely happen. According to [18], the possibility that multiple physical components fail simultaneously in one administrative domain is extremely small. It is also shown that, in IP networks, most of the failures only involve single IP links or routers[19]. Therefore, in this paper, we assume that at most one IP component fails at any time in one domain of IP networks. Under this assumption, the number of network failure states is $|V| + |E|$, and this gives us enough precision for the purpose of designing route reflection topology.

Note that several IP links may share one segment of fiber, and thus they may fail coincidentally when the fiber is cut. In this scenario, we can include the event of single fiber-cut as a failure scenario in $\mathcal{S}$, and analyze the simultaneous failures of the related IP links. Similar analytic models can thus be applied.

Table I summaries the major notations used in this paper.

[1] In this paper, the failure means fail-stop, i.e., the failed components stop functioning. We do not consider Byzantine failures, such as misbehaving or adversaries in routing.

TABLE I

Table of Notations

(u,v)	Physical link from router u to router v
$\langle u,v\rangle$	IBGP session between router u and router v
$G(V,E)$	Physical network with router set V and link set E
$G_r(V_r,E_r)$	IBGP network with router set V_r and IBGP session set E_r
P_{uv}	IGP path from router u to router v
$\mathcal{S}$	Set of all network failure scenarios
F_s	Set of failed components in state s, $F_s \subseteq V \cup E$
r_s	Probability that network failure scenario s occurs
q_s	Conditional session failure probability in scenario s
n	Number of IBGP routers in a domain
m	Number of IBGP sessions in a domain
θ_s	Number of IBGP sessions owned by the routers in F_s
ϕ_s	Number of IBGP sessions passing F_s but not owned by F_s
$\mathcal{P}$	IBGP failure probability (IFP)
$\mathcal{L}_c$	Expected connectivity loss (ECL)

B. Reliability Metrics

In order to characterize the reliability of IBGP route reflection networks, we propose two metrics in this section: *IBGP Failure Probability* (IFP) and *Expected Connectivity Loss* (ECL). Though our methods focus on route reflection IBGP networks (and the full mesh IBGP), they can also be applied to the confederation IBGP networks with very few modifications.

We define the *IBGP failure* as the termination of one or several non-optional IBGP sessions, which is caused by router failures or IP link failures[2]. Moreover, we define θ_s as the number of IBGP sessions that are owned by the routers in F_s; let ϕ_s denote the the number of IBGP sessions that pass F_s, but are not owned by F_s. θ_s and ϕ_s are determined by the route reflection topology and its relation to the IP network topology. For example, $\theta_{s_0} = 0$ and $\phi_{s_0} = 0$. If B fails in Fig. 1, sessions $\langle A,B\rangle$, $\langle B,C\rangle$ and $\langle B,D\rangle$ fail, because they are owned by B. However, session $\langle A,D\rangle$ only passes B, and thus it fails with probability q_s. Thus, $\theta_s = 3$ and $\phi_s = 1$.

1) IBGP Failure Probability (IFP) $\mathcal{P}$: IFP means the probability that IBGP failure happens. Given a route reflection network and the failure scenario set $\mathcal{S}$, the probability of IBGP failure $\mathcal{P}$ can be calculated from the probability of each scenario and the topologies of IBGP networks:

$$\mathcal{P} = 1 - \prod_{s\in\mathcal{S}} \left[1 - r_s + r_s(1-q_s)^{\phi_s}\mathbf{1}_{\{\theta_s=0\}}\right]. \tag{2}$$

At the right hand side of Equation 2, we calculate the probability of no IBGP failure for each scenario, which contains two part: either the scenario does not happen or no sessions actually fail.

In practice, if we only consider failures of single IP links or routers, i.e., $\mathcal{S} = V \cup E$, $\mathcal{P}$ in Equation 2 can be simplified as

$$\mathcal{P} = 1 - \underbrace{\prod_{s\in V}(1-r_s)}_{\bar{\mathcal{P}}_V}\underbrace{\prod_{s\in E}\left[1-r_s+r_s(1-q_s)^{\phi_s}\right]}_{\bar{\mathcal{P}}_E}. \tag{3}$$

$\bar{\mathcal{P}}_V$ is the same for any IBGP route reflection topology, because every router failure definitely breaks all the IBGP sessions possessed by it and causes IBGP failure. $\bar{\mathcal{P}}_E$ is contributed by IP link failures. If no IBGP sessions are deployed on a link (i.e., $\phi_s = 0$), its failure does not influence IBGP; otherwise, the sessions passing it are influenced and fail with probability q_s. Thus, the probability of no IBGP failure is $(1-q_s)^{\phi_s}$.

Because $\bar{\mathcal{P}}_V$ is not related to IBGP route reflection topologies, the optimization problem with respect to IFP only needs to consider $\bar{\mathcal{P}}_E$. Then, minimizing $\mathcal{P}$ is equivalent to maximizing $\bar{\mathcal{P}}_E$.

2) Expected Connectivity Loss (ECL) $\mathcal{L}_c$: IFP reflects how likely an IBGP failure will happen. However, it does not characterize how severe the IBGP failure is. Another dimension in the design space is the amount of losses in IBGP failures. We define an elaborate loss evaluation metric, Expected Connectivity Loss (also called *Expected Resilience Loss*), to characterize the detailed function loss of IBGP networks, and denote it as $\mathcal{L}_c$.

The function of an IBGP network is to distribute external routing information, which is learned from outside of the AS, to all other IBGP routers in this domain after necessary processing. In a healthy network, there is a valid IBGP signaling path between any two IBGP routers. BGP update information originated from one router can be advertised to all other IBGP routers in the domain by these IBGP signaling paths. Some of these signaling paths are just direct IBGP sessions in the IBGP networks; some of the paths involve multiple IBGP sessions and reflectors. In network failure scenarios, if some IBGP sessions fail, the IBGP signaling paths between some routers become unavailable, and thus the function of IBGP network is partially defunct. Therefore, we can use the number of failed IBGP signaling paths to measure the function loss of IBGP networks.

In network failure scenario s, if router i can not exchange routing information with router j directly or indirectly due to IBGP session failures, i and j are *isolated* from each other and we denote this relation as $i \overset{s}{\nleftrightarrow} j$. Likewise, if i and j are reachable to each other in failure state s, it is denoted as $i \overset{s}{\leftrightarrow} j$. Please note that two routers may be isolated even if they are connected in a route reflection graph. For example, if session $\langle A,D\rangle$ fails in Fig. 1, A and D are isolated. Though A and D both share an IBGP session with B, B does not reflect routes between A and D, because B is in the same cluster as A. CLUSTER_LIST loop detection policy [2] prevents this type of route reflection. In Section IV, we will discuss the calculation of the isolation probability by considering valid route reflection.

In a full mesh IBGP network, any two routers communicate directly using a dedicated IBGP session. If session $\langle i,j\rangle$ fails, only i and j are isolated from each other. In a route reflection network, the routers are organized hierarchically. One IBGP session may on the signaling paths of multiple router pairs.

[2] We will show later that any non-optional IBGP session failure leads to function loss of IBGP routing in the existing BGP protocol, no matter whether redundant reflectors or sessions are employed (Lemma 4).

Thus, the termination of one IBGP session may invalidate several signaling paths. For example, in Fig. 1, the external routing information learned by router H has to be reflected by E, by I, and then it can be received by J. If IBGP session $\langle H, E\rangle$ fails, H is isolated from the other BGP routers in the domain. In addition, failures of different IBGP sessions or routers have different impacts on IBGP operations. For instance, the termination of session $\langle A, B\rangle$ only makes A and B isolated from each other. However, if session $\langle E, I\rangle$ breaks, the routers in cluster II and III are isolated.

In a general IBGP network, based on the above descriptions about IBGP function loss, the connectivity loss in scenario s, denoted as $\mathcal{L}_c(s)$, is

$$\mathcal{L}_c(s) = \frac{2\sum_{i,j\in V_r, i\neq j} Pr[i \overset{s}{\nleftrightarrow} j]}{n^2 - n}, \tag{4}$$

where n is the number of IBGP routers. On the other hand, the IBGP network connectivity $\bar{\mathcal{L}}_s$ is

$$\bar{\mathcal{L}}_c(s) = 1 - \mathcal{L}_c(s) = \frac{2\sum_{i,j\in V_r, i\neq j} Pr[i \overset{s}{\leftrightarrow} j]}{n^2 - n}. \tag{5}$$

Therefore, the ECL over the entire state space is $\mathcal{L}_c = \sum_{s\in\mathcal{S}} r_s \mathcal{L}_c(s)$ and $\bar{\mathcal{L}}_c = \sum_{s\in\mathcal{S}} r_s \bar{\mathcal{L}}_c(s)$, where r_s is the probability that the network is in state s. It is easy to verify the following facts: $0 \leq \mathcal{L}_c, \bar{\mathcal{L}}_c \leq 1$ and $\mathcal{L}_c + \bar{\mathcal{L}}_c = 1$.

In addition, the ECL metric reflects the routing service availability to end users. From users' point of view, routing reliability means the reachability of network addresses which further depends on the reliability of IBGP signaling routes.

IV. Reliability Calculation of IBGP Networks

In this section, we discuss how to calculate the reliability metrics based on the topology of IBGP networks. According to Equation 2, we first need to calculate $\{\theta_s\}$ and $\{\phi_s\}$ based on the relation between IP network topology and IBGP network topology, and then compute IFP. The challenging part is to calculate $\mathcal{L}_c$, which can be reduced to computing network connectedness [20] in IBGP signaling graphs. We will discuss this in detail next, followed by a case study to show the intuitions behind the metric definitions and the reliable IBGP network design.

A. Calculation of IBGP Network Connectivity Loss $\mathcal{L}_c$

In order to calculate the connectivity loss $\mathcal{L}_c$, we need to obtain the isolation probability $Pr[i \overset{s}{\nleftrightarrow} j]$ for any pair of nodes, i and j, in any failure scenario s. Because not all paths in IBGP route reflection graph are valid signaling paths, the reflection graph can not be used directly to calculate the isolation probability, and we will define auxiliary graphs for this purpose instead.

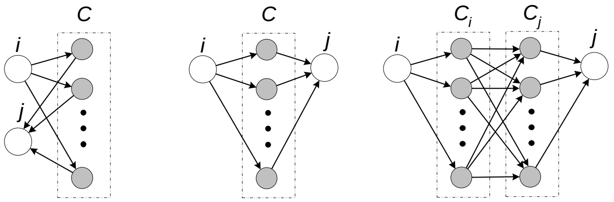
Let us first explain what is the valid IBGP signaling path. In a route reflection network, if routing information is sent from a client to its reflector, we define this advertising relationship as C-R; similarly, R-C and R-R stand for sending routing information from a reflector to its client and from a reflector to its peer reflector in different clusters, respectively. Thus, according to IETF RFC [2], the valid route advertising path is the subsequence or the whole of the following sequence: C-R ⇒ . . . C-R ⇒ R-R ⇒ R-C. . . ⇒ R-C[3].

The calculation of isolation probability can be reduced to the (s,t)-*connectedness* problem[4] [20] in the directed acyclic graph G^s_{ij} in which the edges may fail independently. The auxiliary graph G^s_{ij} is generated based on the route reflection graph $G_r(V_r, E_r)$ with the following three modifications: (1) The IBGP routers that fail in network failure scenario s and the IBGP sessions they own are removed from V_r and E_r. (2) In G^s_{ij}, the edges, which pass F_s, have failure probability q_s; other edges have zero failure probability. (3) The nodes and edges that are not on the valid IBGP signaling paths between i and j in $G_r(V_r, E_r)$ are removed. In the resulted graph after the above operations, the directions of the edges are determined, such that the edge direction conforms to the shortest path from i to j in terms of hop-count in graph G_r.

Lemma 1: A signaling path is valid, if and only if it is a path in graph G^s_{ij} from i to j.

Proof Sketch: First, it is easy to see that the direction of every edge in G^s_{ij} can be uniquely determined. In reflection graph G_r, according to the rule of IBGP route reflection, the set of all valid signaling paths from i to j is equivalent to the set of shortest paths between i and j in terms of hop-count. Because only the edges in G_r that are not used by signaling paths from i to j are removed in G^s_{ij} and the edge directions in G^s_{ij} are determined to conform to shortest paths from i to j, G_r and G^s_{ij} share the same set of shortest paths from i to j. Thus, the set of signaling paths is equivalent to the set of the shortest paths in G^s_{ij}. Moreover, we can show that all paths from i to j in G^s_{ij} have the same distance. Therefore, we conclude that signaling path set is equivalent to the path set in G^s_{ij}. ■

Based on the above lemma, the problem of computing the isolation probability $Pr[i \overset{s}{\nleftrightarrow} j]$ is equivalent to calculate the probability that all paths from i to j fail in graph G^s_{ij}. Next, we will focus on the two-level route reflection topology to discuss the detailed calculation.



(a) Two clients in same cluster.

(b) Client vs. reflector in other cluster.

(c) Two clients in different clusters.

Fig. 2. G^s_{ij} for calculating $Pr[i \overset{s}{\nleftrightarrow} j]$ in two-level route reflection networks.

In a two-level route reflection network, the routing in-

[3] If optional IBGP sessions between clients in the same cluster are used, the signaling paths can also be C-C. Because signaling paths of this type are independent of others and the related isolation probability can be simply calculated, we ignore the optional sessions in this section for clarity.

[4] The (s,t)-connectedness problem aims to compute the probability that at least one path from i to j does not fail in the directed probabilistic graph.

formation can be reflected at most twice. We thus divide the calculations into the following three cases based on the relationship between router i and router j. For convenience of explanation, we denote the failure probability of IBGP session $\langle u, v\rangle$ in failure scenario s simply as p_{uv}.

First, if both i and j are reflectors or i is a client of the reflector j, because of route reflection rules and CLUSTER_LIST loop detection, the routes from i can not be reflected to j by any other reflectors and vice versa. Thus, only one IBGP signaling path exists and $Pr[i \overset{s}{\nleftrightarrow} j] = p_{ij}$.

Second, if i and j are clients in the same cluster or i is a client and j is a reflector in other cluster, the graphs G^s_{ij} of these two scenarios are shown in Fig. 2(a) and Fig. 2(b), respectively. There are $|C|$ independent paths from i to j, where C is the set of reflectors in the cluster of i. Thus,

$$Pr[i \overset{s}{\nleftrightarrow} j] = \prod_{c \in C} (p_{ic} + p_{cj} - p_{ic}p_{cj}) \tag{6}$$

Third, if i and j are clients in different clusters, the graph G^s_{ij} of this scenario is shown in Fig. 2(c). C_i and C_j are the sets of reflectors in the cluster of i and j, respectively. There are $|C_i||C_j|$ different paths from i to j. Thus, $Pr[i \overset{s}{\nleftrightarrow} j]$ is the probability that all these paths fail. However, because these paths are not independent, the probability calculation could be a difficult problem. In general, the following lemma shows that it is unlikely to find efficient solutions to calculate the isolation probability in this scenario if $|C_i|$ and $|C_j|$ are large.

Lemma 2: If i and j are clients in different clusters, the problem of computing $Pr[i \overset{s}{\nleftrightarrow} j]$ is #**P**-complete.

Proof: The proof sketch is in Appendix I. ■

In practice, if the sessions with zero failure probabilities cover a path from i to j in G^s_{ij}, then $Pr[i \overset{s}{\nleftrightarrow} j] = 0$; otherwise, the isolation probability can be computed fast enough by some existing network reliability analysis methods. The reasons are that the number of IBGP sessions that have nonzero failure probabilities in a failure scenario is small and the number of redundant reflectors ($|C_i|$ and $|C_j|$) is also quite limited. In this paper, we use the factoring algorithms [21] to calculate the isolation probability.

B. Case Studies - functional reliability analysis

We perform a functional reliability analysis on eight IBGP networks which are overlaid on top of the same IP network. The functional reliability analysis means to analyze the reliability of the IBGP network in which the failure probabilities of all components (including IBGP sessions) are the same. Let us denote r as the happening probability of each failure scenario and denote q as the conditional failure probability of the influenced IBGP sessions in all network failure scenarios. We only consider failures of single IP link or single router in the following analysis.

Table II shows eight IBGP reflection networks and their reliability metrics $\mathcal{P}$ and $\mathcal{L}_c$. Except for Case (b) that has two clusters, all cases have only one cluster. β_k denotes $1 - r + r(1-q)^k$ to simplify the representation of IFP $\mathcal{P}$. We use Case (c) as an example to show the computation. If E fails, $\mathcal{L}_c(s) = \frac{10}{10}$, because all routers are definitely isolated; if A fails, A is isolated and B loses contact with others with probability q, so $\mathcal{L}_c(s) = \frac{4+3q}{10}$; etc. By combining $\mathcal{L}_c(s)$ of all network failure scenarios, we obtain ECL of Case (c): $\mathcal{L}_c = \frac{r}{10}(26+23q-q^2)$. The calculation of $\mathcal{P}$ follows Equation 2.

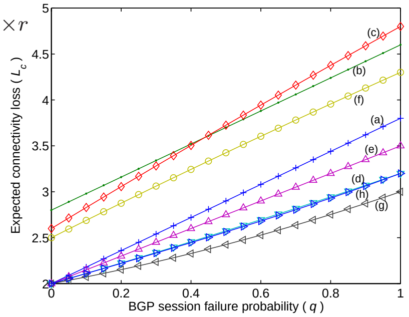


Fig. 3. The comparison of expected connectivity loss.

It is straightforward to see that $1 - r \le \beta_k \le 1$ and $\beta_k \ge \beta_{k+1}$. Thus, we have the order of IFP for the eight cases:

$$\text{IFP } \mathcal{P} : (b) \le (c) \le (f) \le (d) \le (e) \le (a) = (g) = (h).$$

The above order holds for any q. However, the order of $\mathcal{L}_c$ is slightly influenced by the specific value of q and the order is shown in Fig. 3.

With respect to IFP, Case (b) is the best option, because it has the minimum number of IBGP sessions and covers the minimum number of IP links. Optimizing IFP tends to use small number of IBGP sessions and IP links, which may lead to large function loss in IBGP failures. On the other hand, ECL takes into account the loss of IBGP failures. Case (b) and (c) are less reliable than other cases with respect to ECL, such as the traditional design in Case (a), due to the single point of failure problem. For example, if E fails in (c), all routers are isolated.

There are two ways to increase IBGP network resilience: using redundant reflectors and adding redundant IBGP sessions between clients. Case (d) employs two reflectors. It is much more resilient than Case (c), due to the redundant reflectors and three additional sessions. It is even more reliable than Case (a) which has the maximum number of sessions. The reason is that there is only one signaling path between any two routers in Case (a), while multiple IBGP signaling paths may exist in Case (d). Thus, the route reflection can avoid some cases of router isolation. For example, if link (C, D) fails, in Case (a), $Pr[B \overset{s}{\nleftrightarrow} D] = q$, because other routers do not reflect routes between B and D. But, in Case (d), the redundant reflectors, C and E, both reflect routes between B and D, i.e., there are two independent paths from B to D in graph G^s_{BD}. Therefore, the communication between B and D is not affected by the failure of (C, D).

However, using more redundant reflectors does not necessarily provide higher reliability. Case (e) uses one more reflector and two more sessions than Case (d), but it still performs worse. This is because a reflector can not reflect

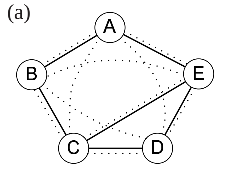
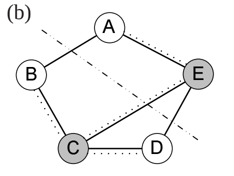
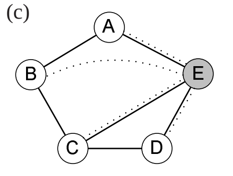
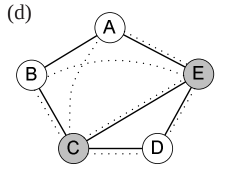
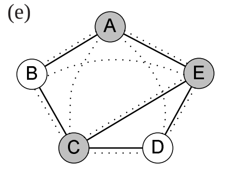
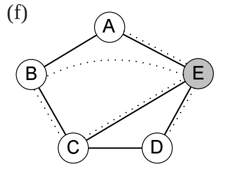
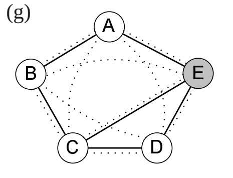
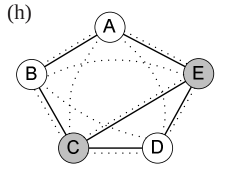
G, G_r				
$IFP:\ \mathcal{P}$	$1-(1-r)^5\beta_3^3\beta_2^2\beta_1$	$1-(1-r)^5\beta_1^4$	$1-(1-r)^5\beta_2\beta_1^3$	$1-(1-r)^5\beta_2^3\beta_1^3$
$ECL:\ \mathcal{L}_c$	$\frac{r}{10}(20+18q)$	$\frac{r}{10}(28+18q)$	$\frac{r}{10}(26+23q-q^2)$	$\frac{r}{10}(20+11q+q^2)$
G, G_r				
$IFP:\ \mathcal{P}$	$1-(1-r)^5\beta_3^2\beta_2^2\beta_1^2$	$1-(1-r)^5\beta_2\beta_1^4$	$1-(1-r)^5\beta_3^3\beta_2^2\beta_1$	$1-(1-r)^5\beta_3^3\beta_2^2\beta_1$
$ECL:\ \mathcal{L}_c$	$\frac{r}{10}(20+15q)$	$\frac{r}{10}(25+19q-q^2)$	$\frac{r}{10}(20+7q+3q^2)$	$\frac{r}{10}(20+11q+q^3)$

TABLE II

CASE STUDIES OF ROUTE REFLECTION NETWORKS. $r = r_s$, THE PROBABILITY OF A FAILURE SCENARIO. $q = q_s$, THE FAILURE PROBABILITY OF IBGP SESSIONS IN FAILURE SCENARIOS. $\beta_k = 1 - r + r(1-q)^k$.

routes between its redundant reflectors and their clients (due to CLUSTER_LIST loop detection), i.e., too many reflectors may make the IBGP signaling paths less redundant. In Case (e), there is only one path in graph G^s_{AD} from A to D. If link (D,E) fails, $Pr[A \overset{s}{\nleftrightarrow} D] = q$. But, in Case (d), two independent paths exist, because both A and D are clients and they can exchange routes via reflector C and E. Therefore, if link (D,E) fails, $Pr[A \overset{s}{\nleftrightarrow} D] = 0$.

Using redundant sessions between clients of the same cluster can also improve reliability. Based on Case (c), we introduce one more session between node B and node C in Case (f). This improves ECL, because the number of independent signaling paths between B and C increases. Case (g) even constructs a full mesh among all clients, and it is most reliable among all these IBGP networks. In addition, in some scenario, using many redundant sessions among clients can not improve ECL significantly. For example, Case (h) only obtains very slightly smaller ECL than Case (d), thus these three additional sessions are not worthwhile.

In summary, this case study gives some intuitions about optimizing IBGP networks for reliability: (1) The traditional full mesh IBGP network is not the most reliable solution, and we can make IBGP networks more reliable by introducing redundant reflectors and sessions appropriately, without incurring much additional overhead; (2) Redundant reflectors can improve BGP network reliability, but they have to be used appropriately, because too many redundant reflectors may decrease IBGP robustness.

V. RELIABILITY OPTIMIZATION FOR IBGP NETWORKS

In previous sections, we defined reliability metrics for IBGP networks and showed that the route reflection topology affects the reliability of an IBGP network. We observed that the traditional IBGP network with the fully meshed IBGP sessions is not the most reliable IBGP configuration, which motivates us to optimize route reflection topologies to increase the IBGP robustness. In this section, we discuss several categories of optimization problems and their solutions. In next section, we will present the experimental studies on these problems.

A. Optimization Problem Description and Complexity

The optimization objective is to minimize one of the three metrics: IFP, ESL and ELT. We use $\mathcal{M}$ to denote any of the metrics, and describe the general form of the optimization problem as follows:

Problem 1 (Reliable RR Network Design): In IP network $G(V,E)$, given (1) the probabilities of failure scenarios $\{r_s | s \in \mathcal{S}\}$, where $\mathcal{S} = V \cup E$, (2) IBGP session failure probability $\{q_s | s \in \mathcal{S}\}$, and (3) all IGP paths, find the route reflection network G^*_r based on G, such that $\mathcal{M}(G^*_r) \leq \mathcal{M}(G_r)$, for any route reflection network G_r based on G.

The design problem is called *RR-IFP* and *RR-ECL*, when the metric to be minimized is IFP and ECL, respectively. The lemma below shows the complexity of finding the optimum solution for the general reliable IBGP route reflection design problem.

Lemma 3: The reliable route reflection design problems, RR-IFP and RR-ECL, are NP-hard.

Proof: We will show (in Lemma 7) that a special case of the problems, where the IP network is a complete graph, is NP-hard. Thus, RR-IFP and RR-ECL in general are NP-hard. ■

Since the general optimization problem is NP-hard, we need heuristic algorithms to find satisfying near-optimal solutions in practice. In order to have more insights into the design problem, we separate the optimization problem into two categories according to the structure of IP networks, that is, optimization in fully meshed IP networks and in general IP networks. There are two reasons for this separation: (1) In the fully meshed IP networks, the reliability metric computation and IBGP topology optimization can be substantially simplified, from which we can obtain helpful intuitions on designing the reliable IBGP network; (2) Investigating the fully meshed IP networks is also useful in real applications. Small AS's may have the fully meshed IP links deployed to increase the communication robustness. In large AS's, the backbone routers in one PoP usually are fully connected by IP links. Thus, the design of IBGP route reflection inside fully meshed IP networks has strong application background.

B. IBGP Reflection over Fully Meshed IP Networks

In a fully meshed IP network, all IGP routes are independent of each other, except for the shared source nodes or destination nodes. Therefore, any two IBGP sessions are independent of each other unless they are owned by the same router. Due to this property, the design problem of IBGP route reflection topology can be significantly simplified.

In RR-IFP problem, because every IP link has at most one IBGP session deployed, $\theta_{ij} = 1$, if router i and router j are IBGP peers; otherwise, $\theta_{ij} = 0$. Based on Equation 3, we can simplify the optimization objective as follows:

$$\begin{aligned} \arg\min_{G_r} \mathcal{P} &= \arg\max_{G_r} \bar{\mathcal{P}}_E \\ &= \arg\max_{G_r} \prod_{\langle i,j\rangle \in E_r} (1 - r_{ij}q_{ij}) \\ &= \arg\min_{G_r} \sum_{\langle i,j\rangle \in E_r} \log \frac{1}{1 - r_{ij}q_{ij}} \qquad (7) \end{aligned}$$

In RR-ECL problem, the optimization can also be simplified. In general, if a cluster has only one reflector, the reflector becomes a bottleneck and the IBGP network will be severally isolated when this critical router fails. (1) In IBGP networks having only one cluster, optional sessions between clients can solve this single point of failure problem, such as Case (g) in Table II. However, a large number of optional IBGP sessions have to be used, which is not scalable to large networks. (2) In IBGP networks with multiple clusters, the failure of the single reflector can isolate the cluster that the single reflector resides in. In this case, the use of optional sessions does not help[5] and significant amount of connectivity loss is thus caused. Therefore, we set up more than one reflector in each cluster as a general rule in this paper to provide a clear explanation and we will briefly discuss the single reflector case after this general case.

We first present a lemma to show the direct impact of a BGP session failure on IBGP network connectivity.

Lemma 4: If IBGP session $\langle u, v\rangle$ fails and it is a non-optional session in IBGP route reflection networks, router u and router v are isolated from each other.

Proof: Since u and v share a required IBGP session, there are three possible relationship between them: (reflector, reflector), (client, single reflector), and (client, one of the multiple reflectors). We use Fig. 1 as an example. If both u and v are reflectors, such as A and E, because other reflectors can not reflect routing information between two other reflectors, u and v are isolated. If u is the only reflector of client v, v loses the single connection with the other IBGP routers. If v has multiple reflectors, for example, C uses A and B as its reflectors, because the CLUSTER_LIST loop detection is enforced, A (or B) does not reflect information from C to B (or A). Therefore, by summarizing all scenarios, u and v are isolated from each other. ■

Please recall that n denotes the number of routers. Due to the multiple reflectors in the fully meshed IP networks, we have the following lemma concerning the connectivity loss caused by network failures.

Lemma 5: If each cluster has more than one reflector in the fully meshed IP network $G(V, E)$, then the optional IBGP sessions between clients in the same cluster do not influence ECL, and moreover,

a) ECL caused by the failure of routers, written $\mathcal{L}_c(V)$, is irrelevant to reflection topology design, and $\mathcal{L}_c(V) = \frac{2}{n}\sum_{i\in V} r_i$, where r_i is the probability that router i fails.
b) ECL caused by the failure of link (i, j), written $\mathcal{R}_c(i, j)$, is nonzero only if router i and j share an IBGP session $\langle i, j\rangle$, and $\mathcal{L}_c(i, j) = \frac{2}{n^2-n} r_{ij}q_{ij}$.

Proof: Because all IGP routes are one-hop in the fully meshed IP networks and the number of reflectors are more than one in each cluster, there are at least two independent IBGP signaling routes between any two clients in the same cluster through the reflectors. Since, we only consider the single IP component failure, the use of optional IBGP sessions, adding another signaling path, does not influence ECL.

Single router failure only resets the IBGP sessions it possessed, and it does not influence the IBGP sessions between other BGP routers, because all IBGP sessions cross only one IP hop. To be more specifically, if the failed router is a client, obviously the reflection structure is not changed; if it is a reflector, the other redundant reflectors in the cluster ensure the connectivity between the clients and other part of the IBGP network. Thus, the failed router is the only isolated IBGP router. No matter what reflection topology is used, ECL is the same, and $\mathcal{L}_c(V) = \sum_{i\in V} \frac{2(n-1)}{n^2-n} r_i = \frac{2}{n}\sum_{i\in V} r_i$.

In fully meshed IP network, session $\langle i, j\rangle$ only crosses link (i, j) and link (i, j) can only be used by session $\langle i, j\rangle$. Thus, route reflection network is influenced by link (i, j) if and only if i and j share an IBGP session. We further argue that if session $\langle i, j\rangle$ breaks, i and j is the only isolated pair as follows. By lemma 4, we know that i and j are isolated from each other, if $\langle i, j\rangle$ fails. Moreover, due to the redundant reflectors in each cluster, the failure of $\langle i, j\rangle$ does not influence signaling paths between routers except i and j. Therefore, the calculation of ECL due to link (i, j) failure can be simplified as $\mathcal{L}_c(i, j) = \frac{2}{n^2-n} q_{ij} r_{ij}$. ■

From the lemma above, the solution to RR-ECL problem in the fully meshed IP network is

$$\arg\min_{G_r} \mathcal{L}_c = \arg\min_{G_r} \mathcal{L}_c(E) = \arg\min_{G_r} \sum_{\langle i,j\rangle\in E_r} r_{ij}q_{ij}, \quad (8)$$

and the optional sessions between clients in the same cluster are not necessary.

By combining the results from Equations 7 and 8 together, we note that the problems of reliable IBGP route reflection network are equivalent to minimizing the weight summation

[5] Optional IBGP sessions are only used between clients in the same cluster.

of the whole route reflection graph. In a formal manner, let $\mathcal{W}(G_r)$ denote the weight summation of all links in route reflection graph G_r, i.e., $\mathcal{W}(G_r) = \sum_{\langle i,j\rangle \in E_r} w_{ij}$. We have the following problem formulation.

Problem 2 (Simplified Reliable RR Network Design): In the complete graph $G(V,E)$, given the link weight $\{w_{ij} | i,j \in V\}$, find a subgraph of G to be the route reflection network G_r^*, such that $\mathcal{W}(G_r^*) \leq \mathcal{W}(G_r)$, for any route reflection network G_r based on G.

Specifically, in RR-IFP problem, $w_{ij} = -\log(1 - r_{ij}q_{ij})$; in RR-ECL problem $w_{ij} = r_{ij}q_{ij}$. It is interesting to note that these two weight definitions are approximately the same when $r_{ij}q_{ij}$ is small, because $\log(1+x) \simeq x$ for small x. In the existing Internet, the network failure probabilities are indeed very small, and therefore, RR-IFP and RR-ECL have approximately the same optimization objective in fully meshed IP networks, i.e., if we find the optimum solution for one, we get an approximate solution for the other.

Intuitively, in order to minimize $\mathcal{W}(G_r)$, we would like to make use of reliable IP links and have a small number of IP links involved in route reflection networks. The traditional full mesh IBGP network has every IP link involved, and thus it has worse reliability than other properly designed IBGP networks. Quantitatively, in the fully meshed IP networks $G(V,E)$, if the traditional full mesh IBGP is used, $\mathcal{P} = 1 - \prod_{i \in V}(1 - q_i)\prod_{e \in E}(1 - q_e r_e)$, and $\mathcal{L}_c = \frac{2}{n}\sum_{i \in V} r_i + \sum_{e \in E} \frac{2}{n^2 - n} r_e q_e$. In later sections, we will see how much reliability improvement can be made by route reflection network optimization.

1) Single-cluster Case: We impose one more constraint on the optimization problem: only one cluster is allowed. This restriction is reasonable in small IP networks. Two important questions need to be answered: how many redundant reflectors are needed, and where to place reflectors. The following lemma gives the optimum number of reflectors.

Lemma 6: In fully meshed IP networks, if a single cluster is used in the overlaid IBGP route reflection network, the optimum number of reflectors is 2.

Proof: Because there are at least two reflectors in the cluster, we only need to show that any reflection network with $k+1$ reflectors ($k \geq 2$), written G_r^{k+1}, has larger $\mathcal{W}(G_r)$ than a reflection network with k reflectors, written G_r^k. In network G_r^{k+1}, we choose arbitrarily one reflector i and change it into client to obtain a new reflection network G_r^k. The sessions between i and other reflectors in G_r^{k+1} remain in G_r^k, but they change from R-R relationship to R-C relationship. The sessions between i and the clients in G_r^{k+1} are missing in G_r^k, because sessions between clients are not required. Other sessions are the same in the two networks. Thus, $G_r^k.E_r \subset G_r^{k+1}.E_r$. It follows naturally that $\mathcal{W}(G_r^k) < \mathcal{W}(G_r^{k+1})$. Therefore, the optimum number of reflectors is 2. ■

Lemma 6 shows that the route reflection network is most reliable if two reflectors are used. Using more reflectors actually decreases the reliability of the IBGP network, even though more resources are consumed. The intuitive reason is that more redundant reflectors lead to less number of IBGP signaling paths in route reflection networks. As an extreme example, in the traditional full mesh IBGP network where every router can be viewed as a reflector, only one signaling path exists between any two routers and it is less reliable than the IBGP networks with fewer route reflectors.

We can thus directly obtain an algorithm, from Lemma 6, to find the optimum route reflection network by enumerating all reflector pairs. The optimum placement of reflectors is:

$$\arg\min_{i,j \in V} M_i + M_j - w_{ij}, \tag{9}$$

where $M_i = \sum_{k \neq i, k \in V} w_{ik}$, $\forall i \in V$. In the resulted IBGP network, $2n-3$ IBGP sessions are used.

Discussions: Besides the IBGP design with redundant reflectors, it is also possible to create the redundancy by using optional IBGP sessions between clients. That is, we choose one router to be the single reflector, the remaining routers are clients, and the clients form a full mesh of IBGP sessions. In this design, between any two clients there are two independent IBGP signaling paths; however, between the reflector and a client there is still only one signaling path. Thus, the optimum reflector is determined by $\arg\min_{i \in V} M_i$. The optimized ECL is smaller than the previous redundant reflector design shown in Equation 9, but the number of IBGP sessions is $\binom{n}{2}$, which is much larger than the redundant reflector design ($2n-3$).

2) Multi-cluster Case: In this section, we consider the case of multiple clusters and the problems of both router clustering and reflector placement in fully meshed IP networks. Because using only one reflector in a cluster may isolate the whole cluster in the multi-cluster case, we focus on the design with redundant reflectors in a cluster. We have the following lemma concerning the complexity of the problem.

Lemma 7: In fully meshed IP networks, the simplified reliable RR network design problem, RR-IFP and RR-ECL, is NP-hard.

Proof: The proof is in appendix II. ■

Router clustering makes the problem much harder than the single-cluster case. In order to solve the design problem, we need heuristic algorithms, which will be discussed in next section. Moreover, we introduce an Integer Linear Programming (ILP) model for solving this problem, especially for fully meshed IP networks. The ILP model can be used to find the optimum solution when the network size is not large by using some powerful mathematical programming solvers, such as CPLEX [22].

Let us consider the problem in which the number of clusters is fixed to be n_c. We define the following binary variables: $g_{ij} = 1$ means router i is a reflector in cluster j; $h_{ij} = 1$ indicates router i is a client in cluster j; $s_{ij} = 1$ means router i and router j share an IBGP session; otherwise, these variables are zero. Below is the ILP formulation.

$$\sum_{i \in V} g_{ij} = n_r \qquad 1 \leq j \leq n_c \tag{10}$$

$$\sum_{1 \leq j \leq n_c} g_{ij} + h_{ij} = 1 \qquad \forall i \in V \tag{11}$$

$$s_{ij} \geq \sum_{1 \leq k \leq n_c} (g_{ik} + g_{jk}) - 1 \quad \forall i,j \in V, i < j \tag{12}$$

$$s_{ij} \geq g_{ik} + h_{jk} - 1 \quad \forall i,j \in V, i < j, 1 \leq k \leq n_c \tag{13}$$

$$s_{ij} \geq g_{jk} + h_{ik} - 1 \quad \forall i,j \in V, i < j, 1 \leq k \leq n_c \tag{14}$$

The optimization objective is as follows.

$$\min \sum_{i,j\in V, i<j} w_{ij} s_{ij} \tag{15}$$

Formula 10 ensures that the number of reflectors in one cluster is n_r ($n_r = 2$ in our discussion). Formula 11 guarantees that any router can be either a reflector or a client in just one cluster. Formula 12 ensures that two reflectors share one IBGP session. Formulas 13 and 14 guarantee that a reflector and any of its clients share one IBGP session. Moreover, we can avoid the unnecessary searching in clustering the routers by the following constraints, which are not required for defining a valid route reflection network but can improve the speed of solving the optimization problem.

$$\sum_{i\in V} (h_{i\,j} + g_{i\,j}) \le \sum_{i\in V} (h_{i\,j+1} + g_{i\,j+1}) \quad \forall i \in V, i \le n-1 \tag{16}$$

Because at least two reflectors are required in one cluster, the range of n_c is from 1 to $\frac{n}{2}$, we can use CPLEX to solve the optimization problem by enumerating all possible values of n_c, and choose the optimal n_c and the reflection topology.

Furthermore, as a special case, if all links and routers are uniform in terms of reliability, the following lemma shows that the single-cluster design is the solution.

Lemma 8: In a fully meshed IP network, if each failure scenario probability (r_s) and each conditional IBGP session failure probability (q_s) are the same, the single-cluster design is the optimum solution for the simplified reliable RR network design problem.

Proof: Because all links and routers are uniform, w_{ij} is the same for any i-j pair. Thus, $\mathcal{W}(G_r)$ is proportional to the number of IBGP sessions used in the reflection network. Given that there are two reflectors in one cluster, the number of sessions is $\binom{2n_c}{2} + 2(n - 2n_c) = 2n + 2n_c^2 - 5n_c$. Therefore, if $n_c = 1$, $\mathcal{W}(G_r)$ is minimized. ■

Therefore, in this special case, this problem is reduced to the problem in Section V-B.1.

C. IBGP Reflection over General IP Networks

If the IP network topology is a general graph, the optimization problems are much more difficult to solve than in the fully meshed IP network. Moreover, optimizing RR-IFP and RR-ECL may lead to different results, even if the network failure probabilities are small, as has been demonstrated in Section IV-B.

The mathematical programming model for RR-IFP problem includes the Formulas 10-14, and the following formula:

$$u_l = \sum_{i\neq j, i,j\in V} s_{ij} f_{ijl}, \qquad \forall l \in E, \tag{17}$$

where binary variable f_{ijl} is one if IGP path P_{ij} passes IP link l; otherwise f_{ijl} is zero. Thus, u_l stands for the number of IBGP sessions passing link l. f_{ijl} can be calculated from IGP routing results. The optimization objective is to maximize $\bar{\mathcal{P}}_E$ in Equation 3, i.e.,

$$\max \sum_{l\in E} \log\left(1 - r_l + r_l\,(1 - q_l)^{u_l}\right). \tag{18}$$

This model contains a nonlinear component. In a special case, where $q_s = 1$ for any failure scenario s, the above model becomes linear, and can be solved using tools such as CPLEX.

In RR-ECL problem, there even does not exist a closed-form model to formulate the problem because of the difficulty in calculating the ECL metric. Therefore, in order to solve the general form of RR-ECL and RR-IFP problems, we apply iterative search techniques. The basic ideas are: probing and calculating various configurations of IBGP networks, choosing the next searching target according to some rules, and optimizing the reliability iteratively. In the following text, we focus on RR-ECL problem to explain the optimization techniques.

1) Greedy Select (GS): Greedy Select (GS) is an intuitive way to design reliable reflection topologies. We first randomly pick up several routers to be reflectors from a candidate reflector set, each reflector is designated to be in an independent cluster, and each of the remaining routers (clients) is connected to a reflector which has the most reliable IGP path to it. In each cluster, the most reliable router other than the existing reflector is chosen as the second reflector. GS repeats this iteration for many times and returns the best topology it finds. In this paper, for a network with n BGP routers, GS iterates for $2n^2$ times. In each iteration, the candidate reflector set consists of the top 60% of the most reliable routers.

2) Tabu Search (TS): Tabu Search (TS) is an efficient meta-heuristic algorithm which can find satisfactory near-optimal solutions in large combinatorial optimization problems. Due to the space limit, we skip the details on tabu search itself, but briefly describe the specific settings we use for solving RR-ECL problem. Interested readers are referred to [23] on the basics of tabu search.

TS optimization is performed based on a fixed number of clusters. We search through all possible values of cluster numbers (from 1 to n) by TS to find the smallest value of ECL. In each TS optimization, we take a two-level approach to optimize RR topologies as follows. At the higher level, the *placement* of reflectors is optimized; at the lower lever, with the fixed reflectors structure, we optimize the *assignment* of the clients to reflectors. The neighborhood structure of the reflector placement is defined by the following procedures: swapping two reflectors in different clusters, swapping a reflector with a client, changing a reflector to be a client, and changing a client to be a reflector. The neighborhood structure of client assignment is defined by moving a client from one cluster to another cluster. The tabu list contains the new clients, the new reflectors, and the clients just being moved to avoid loops in the searching process. However, if a neighboring solution is better than the best result found so far, the tabu condition is disabled. The initial RR topology is generated by using the GS algorithm.

Previously, we has shown that the robustness of IBGP can be increased by using optional IBGP sessions between clients in the same cluster. However, using a large number of optional sessions increases the overhead of BGP routers, especially in large networks. Thus, in practice, we have to make a tradeoff between the scalability and robustness. This tradeoff can also be implemented in our heuristic searching algorithms. In addition, our experimental results, which are based on the realistic Internet network topologies and will be presented later, show that by using optional IBGP sessions, the robustness of IBGP is increased only slightly, compared with the appropriately designed IBGP route reflection networks with no optional sessions, which means that we can obtain both reliable and efficient IBGP networks without using many IBGP sessions.

Using TS, we can find the optimum RR topology for the IP network in the case study previously discussed in Section IV-B. The best RR topology without using optional BGP sessions is: forming one cluster and router B and E being reflectors. The ECL is $\frac{20+11q}{10}r$. The result is better than Case (h) which uses the maximum number of sessions. This example demonstrates the benefits of designing the RR topology properly. When the optional session is allowed, Case (g) is the best design.

3) Lower Bound of ECL optimization: We develop a lower bound for the minimum ECL in RR-ECL problem to evaluate the performance of our optimization algorithms. For convenience, we rank all links in IP network $G(V, E)$ based on the product of the scenario occurrence probability and the corresponding conditional session failure probability, i.e., $\{r_s q_s | s \in E\}$; let $\tilde{r}(i)$ denote the i^{th} smallest value.

Lemma 9: In IP network $G(V, E)$, the lower bound for the solution of RR-ECL problem is $\frac{L}{\binom{|V|}{2}}$, where

$$L = \sum_{i \in V} (|V| - 1) r_i + \max\left((2|V| - 3)\tilde{r}(1), \sum_{i=1}^{|V|-1} \tilde{r}(i)\right). \quad (19)$$

Proof: We analyze the bound by considering router failures and IP link failures separately as follows. (1) On each router failure, we underestimate the ECL by only considering the router isolation related to the failed router. Thus, at least $|V| - 1$ pairs of routers are isolated due to the single router failure, which contributes to the first term in Equation 19. (2) Because the minimum number of the non-optional sessions is $2|V| - 3$ (see the proof of Lemma 8) and each session is influenced by at least one IP link, the minimum amount contributed by these links is $(2|V| - 3)\tilde{r}(1)$ [6]. On the other hand, at least $|V| - 1$ IP links are used by IBGP sessions; otherwise, the IBGP network is not connected at the routing layer. These links contribute at least $\sum_{i=1}^{|V|-1} \tilde{r}(i)$ to ECL. Moreover, by finding the maximum of the above two values, we get the second term in Equation 19. ■

In IP networks that are not strongly connected, the failures of the critical links or nodes can partition the network. For example, in Fig. 1 node G is a critical node. The above lemma can be extended to give a tighter lower bound in this case. We denote $N(i)$ to be the number of node pairs (including the failed nodes) that are partitioned due to the failure of i, where $i \in V \cup E$. $\tilde{r}(i)$ is determined only based on the non-critical links (i.e., the links that can not partition the network). Then, similar to the reasoning in Lemma 9, the lower bound of ECL is $\frac{L}{\binom{|V|}{2}}$, where

$$L = \sum_{i \in V \cup E} N(i) r_i + \sum_{i=1}^{|V|-1-\beta} \tilde{r}(i), \quad (20)$$

where β is the number of the critical IP links which can partition the network.

D. Summary of the Problems and the Solutions

Based on the above discussions on optimizing IBGP route reflection networks, we summarize the problem categories and the time complexities of the solutions in Fig. 4. In some special cases, we can find the optimum design by polynomial algorithms. In general cases, we have to use heuristics based on iterative searching to obtain the near-optimal solutions.

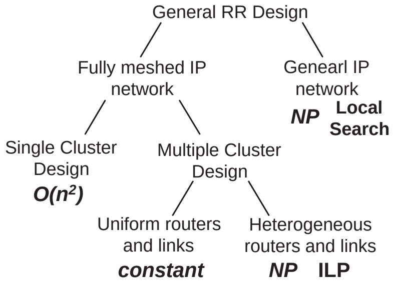


Fig. 4. Reliable IBGP RR network design problems.

VI. Experimental Studies of IBGP Network Optimization

A. Fully meshed IP Networks

In fully meshed IP networks, the RR network design is irrelevant to the reliability of BGP routers. Thus, in this subsection, we assume that all BGP routers are perfect. The failure probabilities of IP links p_f is generated randomly from the interval $[0.02 - \delta, 0.02 + \delta]$, where δ equals 0.005, 0.01, and 0.02, respectively, in the three cases to be studied. The failure scenario occurrence probability is r_e and $r_e = p_f(e) \prod_{l \neq e} [1 - p_f(l)]$. The IBGP session conditional failure probability, q_e, is 0.31, which corresponds to 36 seconds failure recovery time under the default setting of BGP and TCP timers [5]. In addition, because $r_e q_e$ is small, the results of the RR-IFP problem are approximately the same as the RR-ECL problem, and thus we only describe the results of RR-ECL as follows.

In the single-cluster design problem in a network of 20 nodes, we enumerate all possible route reflection networks with two and three reflectors, calculate the ECL for each configuration, sort by ECL, and draw the results in Fig. 5. In the figure, each point stands for a feasible solution to the RR network design problem. The optimum solutions for the three cases are marked in the figure. The IBGP networks with three reflectors are worse than the IBGP networks with two reflectors, which confirms the statement in lemma 6. We also notice that the range of ECL is larger when the range of the link failure probability is higher, which means that we can take advantage of more reliable links and avoid using unreliable links to make RR network more robust.

In the case of multi-cluster design, we use CPLEX to solve the ILP model discussed in Section V-B.2. In a fully meshed IP network with 14 nodes, the link failure probabilities are

[6] Please note that if a non-optional IBGP session fails, the related two IBGP routers are isolated (see Lemma 4).

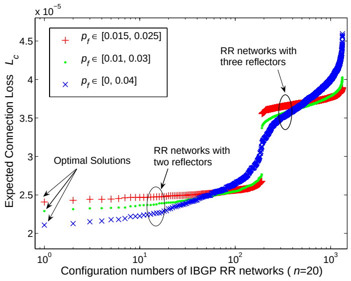


Fig. 5. ECLs of all RR networks with two or three reflectors in a single cluster in a fully meshed IP network.

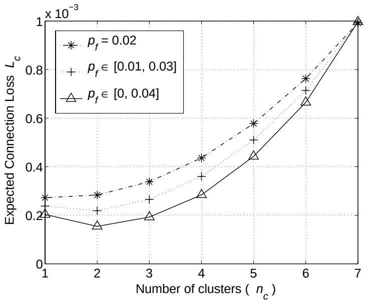


Fig. 6. The optimum ECLs wrt. different number of clusters in a fully meshed IP network.

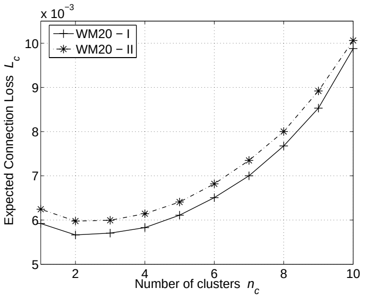


Fig. 7. The optimum ECLs wrt. different number of clusters in WM20 by tabu search.

TABLE III
IP NETWORKS USED IN EXPERIMENTS.

Name	Router Number	Link Number	Description
WM20	20	40	Waxman Model, Brite
WM40	40	80	Waxman Model, Brite
AS6461	19	34	RocketFuel
AS3251	41	87	RocketFuel

generated randomly with δ set to 0, 0.01 or 0.02. Fig. 6 shows the optimization results of when the number of clusters n_c are 1 through 7. If the link failure probabilities are the same for all IP links, ECL increases monotonically as n_c, and the single-cluster design has the smallest ECL, which confirms Lemma 8. On the other hand, if the link failure probabilities are not the same, as shown in the figure, the optimal cluster number is two in these two cases. The reason is that IBGP sessions can be arranged to avoid passing unreliable links by grouping routers into multiple clusters appropriately.

B. General IP Networks

The general network topologies we used in our experiment are summarized in Table III. The first two topologies are generated by Brite topology generator [24] at the router level with Waxman model. The other two are PoP level topologies taken from the rocketfuel project [25], which stands for building a third level route reflection on top of the PoP graphs.

The failure probability of each router is randomly generated from interval $[0.02-a, 0.02+a]$ and the link failure probability is generated from $[0.04-b, 0.04+b]$. In our experiment, we test two settings of a and b: one is $(a, b) = (0.02, 0.04)$; the other is $(a, b) = (0.005, 0.005)$. We call them I and II, which stand for large and small variances of network component failure probabilities, respectively. Thus, we denote the experiment configuration with network WM20 and the first failure setting as 'WM20-I' and other configurations are similarly denoted.

We develop and implement an efficient TS algorithm to solve the RR-ECL problem. Fig. 7 demonstrates the optimization results of TS with respect to different numbers of clusters. In this network (WM20), the optimal cluster number is 2. It is interesting to note that, in the scenario of larger variance of network component failure probability, we can find RR topology with smaller ECL by deploying reflectors and IBGP sessions on the more reliable router and links.

In Fig. 8 and 9, we compare the optimization results of GS and TS with the traditional fully meshed IBGP and the lower bound of ECL in all eight scenarios. It is obvious to observe that by performing RR topology optimization, we can find much more reliable IBGP RR network than the traditional fully meshed IBGP. Even the simple GS algorithm can achieve significant improvement in terms of reliability. By using TS, ECL can be further minimized. Comparing the results with the lower bound of ECL, we conclude that the TS algorithm can find satisfying results which are close to the global optimum solution.

Previously, we have discussed that the reliability of IBGP networks can be improved by using optional sessions between clients of the same cluster. In Fig. 10, we compare the results if the optional sessions are allowed with the results when no optional sessions are used in the algorithm of TS. The figure shows that in the network topologies we studied, the use of optional sessions only improves IBGP reliability very slightly. The reason is that, in these networks, the impact of single component failures on IBGP is already minimized to a small level by placing appropriately redundant reflectors and clustering routers. Thus, the optional sessions are not necessary in these cases and the additional cost in managing the optional sessions in saved. Therefore, we can design both reliable and efficient IBGP route reflection networks.

VII. RELATED WORK

Some recent research addresses the challenges of improving BGP reliability. Sangli et al. [11] propose a graceful restart mechanism for BGP to alleviate the impact of BGP session failures. BGP Scalable Transport (BST) [12] uses application level replication and flooding to substitute TCP for reliable and scalable BGP message distribution. However, besides the deployment difficulties, BST can not replace the hierarchical IBGP design (e.g. route reflection) because a BGP router can not have a large number of BGP peers. 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, without modifying any protocol details, by configuring IBGP route

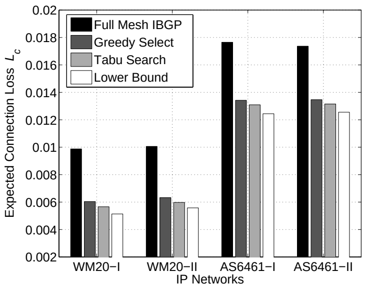


Fig. 8. Results of RR-ECL in WM20 and AS6461.

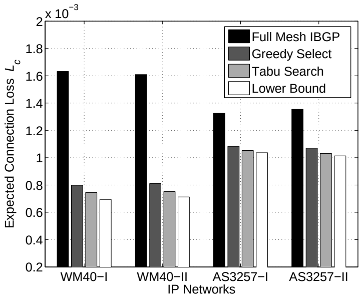


Fig. 9. Results of RR-ECL in WM40 and AS3257.

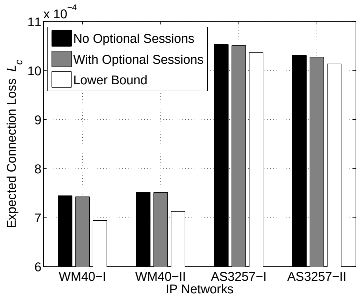


Fig. 10. Performance comparison of RR design with/without optional BGP sessions in WM40 and AS3257.

reflection networks appropriately, so that the rate and impact of IBGP failures are minimized.

Besides reliability, routing convergence is also an important issue in IBGP. The convergent routing can be ensured by either changing route advertising in IBGP or restricting the configuration of IBGP networks. Basu et al. [13] extend the route reflection policies to prevent route oscillations. The basic idea is to advertise not only the best route but also other received routes. Griffin et al. [14] present sufficient conditions to guarantee deterministic and unique IBGP routing and to avoid forwarding deflections. We view the convergence and the reliability issues as two parallel objectives in the design of IBGP networks. In practice, multiple factors should be considered in determining the favorable route reflection topologies. In this paper, we focus on the reliability issue.

VIII. Conclusion

The reliability of IBGP networks has a remarkable impact on the dependability of Internet routing. Based on the existing Internet framework, how to model and how to improve the resilience of IBGP networks are of significant importance. In this paper, we propose a reliability model and two novel reliability metrics for IBGP networks, which take into account the dependent failures of IBGP sessions and quantify the resilience of IBGP networks in various network failure scenarios. Then, we formulate the optimization problems for finding the most reliable route reflection networks and analyze their properties in the fully meshed IP networks and general networks. Moreover, several efficient algorithms are developed to solve the problems efficiently. Through extensive experiments, we show that our model is effective in characterizing the resilience of IBGP networks and we can find satisfying designs which are close to the global optimum solutions.

In our future research, we will take into account other important properties of IBGP route reflection networks, such as convergence. We would like to study the co-design of IBGP networks by considering reliability and other factors, which can lead us to a rational approach to network protocol design and configuration.

Appendix I
Proof Sketch of Lemma 2

The problem of computing $Pr[i \overset{s}{\nleftrightarrow} j]$ is equivalent to finding the probability that all paths from i to j fail in graph G_{ij}^s (shown in Fig. 2(c)). Though C_i and C_j form a complete bipartite, it is a more general case than a general bipartite, because if one edge does not exist, the corresponding edge in G_{ij}^s can have 1 as the failure probability. Thus, the reduced result of Corollary 3.4 in chapter 3.2 of [20] by using the proof technique of Theorem 3.2 in [20] is a special case of G_{ij}^s. Bipartite Independent Set problem, which is $\#\mathbf{P}$-complete, can be reduced to the problem of computing $Pr[i \overset{s}{\nleftrightarrow} j]$, and therefore the result in the lemma follows.

Appendix II
Complexity Analysis

In this section, we show that the simplified route reflection network design described in Problem 2 is NP-hard. We investigate a special case of Problem 2 in which the number of clusters is k and the number of reflectors in each cluster is one, and we call this special route reflection problem k-SRR. This simplification is also useful in the case of IFP minimization.

Uncapacitated Facility Location Problem (UFL) is proved to be NP-hard in [26]. In 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 $\min\limits_{F' \subseteq \mathcal{F}} \left[\sum\limits_{i \in F'} f_i + \sum\limits_{i \in \mathcal{D}} \min\limits_{j \in F'} c_{ij} \right]$. It is easy to see that the problem is still NP-hard even if the number of the opened facilities is fixed to k, because by enumerating k from 1 to n_f, we can find the solution for UFL in polynomial number of iterations. We call the UFL problem with k opened facilities k-UFL problem. Next, we will reduce k-SRR from k-UFL problem.

Lemma 10: k-SRR problem is NP-hard.

Proof: From the k-UFL problem, we construct a graph, as shown in Fig. 11, to form the k-sRRD problem. We define sets $\mathbf{F} = \{F_i \,|\, 1 \leq i \leq n_f\}$ and $\mathbf{D} = \{D_i \,|\, 1 \leq i \leq n_d\}$ 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}.

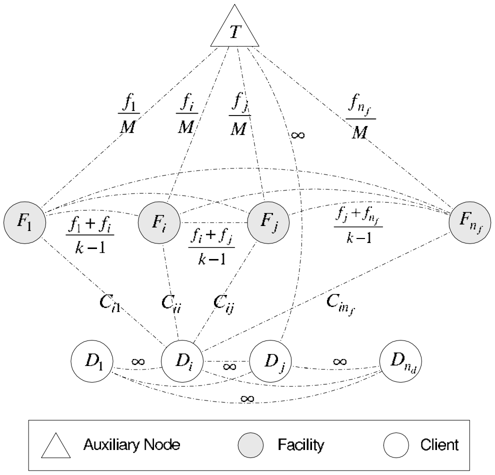


Fig. 11. Reduction from k-UFL to k-SRR.

By constructing the graph in Fig. 11, we map the k-UFL problem into the $(k+1)$-SRR 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. We choose M large enough, i.e., $M \gg \max\left(\max_{i,j,l} \frac{f_l}{c_{ij}}, \max_{i,j,l} \frac{(k-1)f_l}{f_i+f_j}\right)$, 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, $\mathcal{W}(G_r)$, is

$$\begin{aligned} \mathcal{W}(G_r) &= \text{weight of reflector mesh} + \text{weight of client connections} \\ &= \sum_{i,j \in \mathcal{R}} w_{ij} + \sum_{i \in \mathcal{C}} \min_{j \in \mathcal{R}} c_{ij} \\ &= \min_{\substack{F' \subseteq \mathbf{F} \\ |F'|=k}} \left(\sum_{i \in \mathbf{F}} f_i/M + \sum_{i \in F'} f_i + \sum_{i \in \mathbf{D}} \min_{j \in F'} c_{ij} \right) \\ &= \min_{\substack{F' \subseteq \mathbf{F} \\ |F'|=k}} \left(\sum_{i \in F'} f_i + \sum_{i \in \mathbf{D}} \min_{j \in F'} c_{ij} \right) + C \qquad (21) \end{aligned}$$

where $C = \sum_{i \in \mathbf{F}} f_i/M$ and C is a constant. From Equation 21, the $(k+1)$-SRR problem has the same optimization function as the k-UFL problem. Therefore, we proved that k-SRR problem is NP-hard. ■

Finally, since the k-SRR problem is a special case of Problem 2, we know that Problem 2 is NP-hard.

REFERENCES

[1] Y. Rekhter and T. Li, *A Border Gateway Protocol 4 (BGP-4). IETF RFC 1771.*, March 1995.

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

[3] L. Wang, X. Zhao, D. Pei, R. Bush, D. Massey, A. Mankin, S. F. Wu, and L. Zhang, "Observation and analysis of BGP behavior under stress," in *Proceedings of ACM SIGCOMM Internet Measurement Workshop*, 2002.

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

[5] L. Xiao and K. Nahrstedt, "Reliability models and evaluation of internal BGP networks," in *Proceedings of IEEE INFOCOM*, 2004.

[6] L. Xiao, G. He, and K. Nahrstedt, "Understanding bgp session robustness in bandwidth saturation regime," tech report, Department of Computer Science, University of Illinois, UIUCDCS-R-2004-2483, October 2004.

[7] W. Cui, I. Stoica, and R. H. Katz, "Backup path allocation based on a correlated link failure probability model in overlay networks," in *Proceedings of IEEE ICNP*, 2002.

[8] K. V. Le and V. O. Li, "Modeling and analysis of systems with multimode components and dependent failures," *IEEE Transaction on Reliability*, vol. 38, April 1989.

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

[10] L. Xiao, J. Wang, and K. Nahrstedt, "Reliability-aware ibgp route reflection topology design," in *Proceedings of IEEE ICNP*, 2003.

[11] 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.

[12] Packet Design, Inc., "BGP scalable transport," in *http://www.packetdesign.com/technology/bst.htm*.

[13] A. Basu, C.-H. L. Ong, A. Rasala, F. B. Shepherd, and G. Wilfong, "Route oscillations in I-BGP with route reflection," in *Proceedings of ACM SIGCOMM*, 2002.

[14] T. G. Griffin and G. Wilfong, "On the correctness of IBGP configuration," in *Proceedings of ACM SIGCOMM*, 2002.

[15] Cisco Systems Inc., "Troubleshooting router hangs," in *http://www.cisco.com/warp/public/63/why_hang.html*.

[16] Cisco Systems Inc., "Troubleshooting memory problems," in *http://www.cisco.com/warp/public/63/mallocfail.shtml*.

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

[18] D. Zhou and S. Subramaniam, "Survivability in optical networks," *IEEE Network*, vol. 14, December 2000.

[19] A. Markopoulou, G. Iannaccone, S. Bhattacharyya, C.-N. Chuah, and C. Diot, "Characterization of failures in an IP backbone network," in *Proceedings of IEEE INFOCOM*, 2004.

[20] C. J. Colbourn, *The Combinatorics of Network Reliability*. Oxford University Press, 1987.

[21] L. B. Page and J. E. Perry, "Reliability of directed networks using the factoring theroem," *IEEE Transaction on Reliability*, vol. 38, December 1989.

[22] ILOG Inc., "Ilog cplex," in *http://www.ilog.com/products/cplex/*.

[23] F. Glover, "Tabu search: A tutorial," *Interfaces*, vol. 20, pp. 74–94, August 1990.

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

[25] N. Spring, R. Mahajan, and T. Anderson, "Quantifying the causes of path inflation," in *Proceedings of ACM SIGCOMM*, 2003.

[26] G. P. Cornuéjols, G. L. Nemhauser, and L. A. Wolsey, "The uncapacitated facility location problem," in *Discrete Location Theory*, pp. 119–171, 1990.