Towards Internet Emergency Response via Reconnections at Internet Exchange Points

May 14, 2008 10:24 AM

Abstract

As the Internet becomes a critical infrastructure component of our global information-based society, any interruption to its availability can have significant economical and societal impacts. However, the availability of the Internet can be substantially challenged when a disaster strikes, since in such cases the recovery process as of today is largely manual, slow and inefficient. In this paper, we provide key support to substantially speedup the recovery of connectivity of the Internet after a major disaster or incident. Recognizing that the key issue for a recovering Autonomous Systems (AS) is to discover physical connectivity from the routers of the AS to other routers with desired connectivity and capacity, we explore Internet eXchange Points (IXPs), where multiple ASes exchange their traffic. We propose a systematic solution including the recovery framework, communication channel and protocol, reconnection strategy as well as resource allocation mechanisms. Simulated random Internet disaster cases as well as the recent Taiwan earthquake scenario demonstrate that the proposed recovery scheme can automatically find sufficient AS helpers inside IXPs with available capacities, even for severe disasters.

1 Introduction

As the Internet becomes a critical infrastructural component of our global information-based society, any interruption to its availability can have significant societal impacts. Thus, when listing its requirements on the Internet, the GENI initiative [1] states that: "any future Internet should attain the highest possible level of availability."

However, despite the remarkable availability and responsiveness demonstrated by the Internet in most cases, they still need a substantial improvement when a disaster/emergency strikes [2]. Consider the recent Taiwan earthquake that struck on Dec. 26, 2006 as an example. This earthquake was one of several recent disasters and major incidents that have caused damage to the Internet. Even one week after the Taiwan earthquake, the number of outage Internet networks was still in the order of thousands [3]. This slow recovery process caused substantial global and regional damages.

How to speedup the recovery process? Fast repair of the outage links is clearly necessary. However, the repair of physical equipment is typically a time-consuming and expensive process. For the Taiwan earthquake, the repair of the damaged under-sea cables did not finish until Feb. 14, 2007 [3]. This delay is too long to be acceptable.

To speedup the recovery process, the recovering Autonomous Systems (AS) networks have no choice but to seek out other connectivity options to achieve faster recovery. For the Taiwan earthquake, we conducted a survey of some involved networks. The survey results are quite revealing. After the quake, the options available to the operator of a network using some of the damaged links were quite limited. To recover its routes to important destinations before link repair, the operator had to seek out already existing or easily establish-able physical connectivity from its routers to some other routers with connectivity and capacity. For example, one major ISP indicated that it had to query other networks to recover both connectivity to important destinations (*e.g.*, DNS and MSN) and also capacity to avoid congestion. Obtaining 25 Gbps short-term capacity for temporary use from other networks by Jan. 5, 2007, it was able to carry about 80% of its normal traffic. Although it succeeded in recovering most of the traffic before the repair of the damaged links, the delay was still unacceptable: 9 days after the quake.

Why is the long delay? Although negotiations on contracts and finalization of legal proceedings between two networks take time, the major delay happened before these steps. In particular, an Internet network seeking recovery lacks crucial information: which other networks can effectively help me with this recovery? Without architecture support, in the current Internet, a recovering network has to rely on a manual process. This leads to slow, ineffective Internet disaster response.

In this paper, we propose a first design of Internet disaster response mechanism. Our design objective is not a totally automated disaster response system, as there may always be policy decisions and business procedures involving operators. Instead, our objective is to provide key support to substantially speedup the recovery of the connectivity of the Internet after a major disaster or incident.

To better appreciate the objective of our system, consider BGP, the *de facto* inter-domain routing protocol in the Internet. BGP does not make inter-domain routing decisions. It only carries information from the neighbors of a network to the network so that it can make routing decisions. The decisions are then carried by BGP to the neighbors so that they can make or update their decisions. The decision process (called policy routing) is outside BGP. The limitation of BGP, however, is that it carries information only between networks with pre-established business relationships. During a disaster, these pre-established business relationships may no longer be effective or the pre-established business relationship may prohibit recovering ISPs from using such links for recovery.

Thus, the main objective of our protocol is to facilitate the discovery and communications between networks with existing or easily establish-able physical connectivity, although these networks may not already have pre-established business relationships. As our survey of the Taiwan earthquake shows, this is the key to substantially speedup the disaster response of the Internet.

Specifically, we address the following two issues.

- A key step to fast disaster recovery is the discovery of extral physical connectivity from the routers of a recovering AS. We aim to utilized two kinds of links for recovery: policy-violative link and non-existing but easily connected link. The former is the link that BGP policy may prohibit an AS from advertising, but that policy can be modified in the case of a disaster. The latter is the link that can be easily establish, *e.g.*, two routers are connected to a same switch or are geographically nearby. Since an AS typically connects to other ASes at Internet eXchange Points (IXPs), which are physical infrastructures that allow multiple ASes to exchange Internet traffic between their networks (autonomous systems), it is crucial to discover ASes in IXPs. However, not all IXPs publish their members online; even for the ones who publish, the data could be out-of-date. More importantly, when a disaster happens, the victim AS may be isolated from ASes who can be helpers. How to set up a communication channel with other ASes to check for its route availability without any third party assistance is the first challenge we have to address.

- The second issue is: Are there enough ASes with sufficient capacities in the IXP can provide help? Also, does a disaster affected AS really have to seek help from ASes without pre-established business relationship? One simpler method is to leverage the peering links by relaxing the BGP policies to allow certain traffic of an affected AS to go through. This, in fact, explores a subset of all extra potential connectivity in an IXP/PoP. Is that sufficient?

To address the aforementioned challenges, in this paper, we make the following contributions.

- We present a discovery and communication framework for Internet disaster response. In particular, we design a communication channel within an IXP and a protocol to run over the communication channel, via which, any routers in the IXP can broadcast its query and receive replies for disaster response. We do not assume any third party assistance, including the IXPs.

- We simulate random Internet disasters as well as the Taiwan earthquake scenario. The results demonstrate that by solely relaxing the routing policies on peering links is not sufficient. For example, in the Taiwan earthquake scenario, even when just considering the provider ASes in related IXPs, only 39% can find available helpers with all their peering links in the Internet.

  On the other hand, our communication channel can help find available helpers for all the affected ASes. Even when only 30% of the available helpers are willing to help, with a simple distributed resource allocation scheme, we are able to recover more than 80% of the traffic of all affected ASes in the Taiwan earthquake scenario. This demonstrates that the Internet connectivity redundancy at IXPs is abundant and it can effectively recover Internet from a disaster even for a severe one.

The rest of this paper is organized as follows. In Section 2, we give our problem statement and propose an Internet disaster response framework. Then we present the design of communication channels in Section 3. The results of feasibility evaluation are illustrated in Section 4. After that, we discuss the related work in Section 5 and conclude the paper in Section 6.

2 Overview

2.1 Problem Statement

After a disaster or major incident that causes damage to the network infrastructure of an AS, the network will react and utilize its backup connectivity. The existing backup connectivity, however, may not be sufficient, and thus the AS would need to seek network connectivity from its routers to routers belonging to other ASes, if it does not want to wait for the repair of the damaged equipment.

Consider a router a belonging to the recovering AS, and another router b belonging to another AS. We say that there exists a *candidate disaster recovering connectivity* between a and b if the following four conditions are satisfied.

- *Fast on-demand connection*: Either there already exists a physical link connecting a and b or a physical link is easy and fast to establish.

- *Reachability:* Router b has reachability to at least one IP prefix belonging to the set of recovering destination IP prefixes defined by the recovering AS; but currently b does not advertise such route(s) to a either due to lacking of physical connectivity or prohibition by BGP policy.

- *Capacity:* Router b has available spare capacity;

- *Policy:* Policies at both ASes allow connectivity between a and b; The policy of an AS may be changed due to competition, political, or other concerns.

Among the above four conditions, the fast on-demand connection condition is probably the most challenging, and we define the objective of our Internet disaster response system to be that of discovery and exchange of candidate disaster recovering connectivity among ASes to speed up the recovery of Internet networks. It is worth mentioning at this point that there does not exist any such system or protocol in the Internet. A number of recent related pre-work [4, 5] is different from our work in the problem and solutions and we will compare them in Section 5. In the current Internet, the main inter-domain protocol is BGP. However, BGP is not designed for our stated objective. Specifically, BGP carries information only between networks with pre-established business relationships with existing links. During a disaster, these pre-established links may no longer exist or the pre-established business relationship may prohibit recovering ASes from using such links for recovery.

In particular, we study the following two problems in this paper. First, given an AS affected by a disaster, how to set up communication channels to the ASes which can potentially provide candidate disaster recovering connectivity without any third party intervention? The second question of the paper is: how effective is such a recovery at the IXPs, especially when facing a severe disaster? How is it compared with a simpler solution of BGP policy relaxation on peering links? Extensive analysis are presented in Section 4.

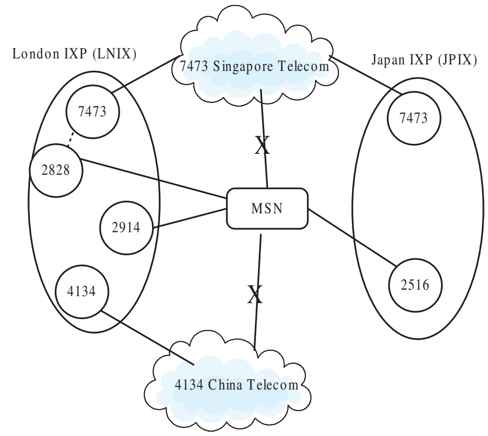


Figure 1: Illustration of exploring potential helpers.

Since vast majority of ASes interconnect at Internet eXchange Points (IXP) which is ideal for the fast on-demand connection, we design a communication channel in an IXP, via which, any router in the IXP can broadcast its query and receive replies for diaster response. Because a major AS should have routers in multiple IXPs, even it loses connections to an IXP, queries for recovery can be sent in other available IXPs under control. Note that, we do not assume any support from third parties, including IXPs. The mechanism is general and can be applied in other router aggregation points such as Point of Presences (PoPs) as well. We show an example with the recent Taiwan Earthquake, and present the framework in the rest of this section. More detailed design is described in Section 3.

2.2 Example for Fast On-demand Connection at IXPs

To illustrate the idea, we consider the recent Taiwan earthquake as an example.

Figure 1 shows an example scenario in Japan IXP (JPIX) and London IXP (LINX) during the Taiwan earthquake incident. Please note that, we just select some ASes in these two IXPs for indication and more detailed results will be shown in Section 4. In Figure 1, the big ovals stand for IXPs and small cycles are ASes in IXPs (the number in each cycle indicates its AS number). In the disaster, Singapore Telecom (7473) and China Telecom (4143) lost their connections to MSN, but AS 2828,

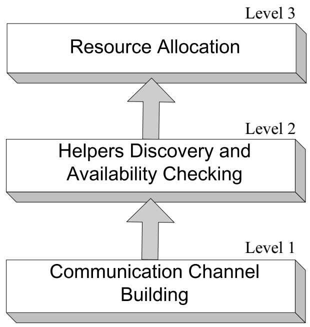


Figure 2: System framework for Internet disaster recovery.

AS 2914 in LINX and AS 2516 in JPIX are still able to access MSN. Please note that, there is a peer relationship between AS 7473 and AS 2828. To recover the traffic from AS 4134 to MSN, we could create a new provider-customer between AS 4134 and AS 2914 (or AS 2828) in LINX and then the traffic would traverse AS 2914 (or AS 2828) to reach MSN. To recover the traffic from AS7473 to MSN, we could either upgrade the peer link 7473-2828 to a provider-customer link or creating a new provider-customer link between AS 7473 and AS 2914 in LINX (or between AS 7473 and AS 2516 in JPIX).

We define the failed network an *disaster affected AS* or in short *affected AS* (*e.g.*, AS 7473 and AS 4134 in this example). We also call it a *buyer* and the assistant network *potential helper* if satisfies "fast on-demand connection" and "reachability" before the disaster happens. A potential helper can be an *available helper* only if both conditions still satisfy after the disaster happens (*e.g.*, AS 2828, AS 2914 and AS 2516 in this example). Then an available helper can possibly connects to the buyer, and becomes an *active helper*.

In the scenario we described above, AS 7473 is a potential helper of AS 4134, but since it also lost its connectivity to MSN, only AS 2828 and AS 2914 are available helpers of AS 7473. Generally, since one network may have routers in more than one IXPs, a buyer could have available helpers in multiple IXPs, *e.g.*, AS 7473 in this example can recover its traffic either in JPIX or in LINX or both.

2.3 System Overview

We now give an overview of our IER system framework as shown in Figure 2. The system framework is divided into three levels: communication channel building, potential helper discovery and availability checking, and resource allocation. Each router in an IXP will be installed a lightweight ER (Emergency Response) module if the owner AS is willing to participate in the emergency response system.

First, when an emergency happens, the ER modules establish "communication channels" among the routers in an IXP, by which any router can broadcast a message to all the other routers inside the IXP as long as they are all connected. Such broadcast is implemented by recursively sending message to all neighbors, very similar to the spanning tree construction. Such a communication channel is simple, robust and incentive yielding as described in Section 3.

Then in level 2, we design a lightweight signaling protocol to inquire and check if the potential helpers are really available for the desired destination networks, such as MSN, DNS, and popular mail servers. For example, the affected AS can send "show ip bgp" script to potential helpers to check the reachability to desired destinations. The replies will be collected and sent to the Network Operation Center (NOC) or local management of affected AS for the task in level 3. There are more details on the protocol in Appendix A.

The final step, depicted as level 3, is to allocate resources which includes helpers pre-selection, business models, and the link reconnections. This step is mainly a human-driven process. For example, the helper pre-selection totally depends on buyers' own decision when receiving multiple helpers' replies on their available bandwidth and the per-unit bandwidth price. Three possible business models are outlined in Appendix B. Besides, two types of reconnection modes, adding a new physical link and using the existing links, are discussed in Section 2.4.

When a disaster occurs, the routers of an affected AS A in some IXPs will be notified to start ER search by the NOC of A, or its local operator when it is cut off from the NOC. Note that A will start the ER search in multiple routers of different IXPs, not limited to those IXPs near the disaster area. For example, we found the IXPs in Europe are very helpful for Asia ASes in the Taiwan Earthquake recovery scenario. Therefore, even an affected AS loses connectivity to one of the IXPs, it can still use the connectivity in other IXPs. Here, we do not have the ER start working automatically because of some temporarily traffic fluctuations or small scale failures which can be recovered by other existing automatic mechanisms such as backup links.

2.4 Interconnection Between Buyers and Helpers

There are two models for AS interconnection: the direct circuit interconnection and the exchange-based interconnection [6]. The former involves a point-to-point circuit between the exchange parties, while the latter involves a meet point (*i.e.*, a switch box) where ASes exchange traffic.

For the direct interconnection model, one possible solution is to add a direct connector between the two routers physically if they did not have one before and modify their routing tables and local policy to connect them logically. It requires sending crews to add a new connector which takes much longer time than the automatic configuration. However, we may not need to create a new physical link to detour the buyer's traffic to the helper. Suppose a topology that A connects to B, and B connects to C. If A would like to involve C as its active helper, A could connect C through B, instead of setting a new link directly between A and B. The advantage of this method is obviously that we avoid setting a new physical link at the cost of involving more ASes.

In the setup of Exchange-based interconnection model, the data path setting is relatively simple. An automatic cross-connector configuration by a switch is possible. The switch exports the addresses of the border routers with helper AS and buyer AS in a control message and then connects the two ports related to the addresses. Meanwhile, the buyer router and the helper router are required to modify their routing tables and local policy to connect them logically.

IP network owners have clear incentives to manage these failures well, because unmanaged failures can cause severe service disruptions and lead to significant financial and reputation damages [4].

Building dedicated redundant networks or setting up backup arrangements ahead of the failure requires significant investment. In addition, since it is impossible to predict when and where a failure will occur, the coverage and gain of such investment on redundancy has essential limitations. In IER framework, an affected AS only borrow connectivity ability from others on demand during the failure, therefore, IER pools the resources of multiple Ases together for mutual backup to improve network reliability at low cost.

3 Communications Channels for Emergency Response

When implementing a fast disaster recovery mechanism, the first important thing is to have an unimpeded underlying communications between a failed AS and other ASes inside related IXPs.

As mentioned in Section 2.4, there are two models for AS interconnection: the exchange based interconnection and the direct circuit interconnection [6]. In the former case, we can use the MAC layer broadcast for efficient communications. But in the latter case, there exists some POS (Packet over SONET) interfaces which do not support the MAC layer communications, so we need to elaborate designing the underlying communication protocol. Here in this section we propose an IP layer ad hoc communication mechanism.

3.1 Communication Channel Architecture

We create a neighboring communication entity in each router and configure the ASes/routers in an IXP making them forming a virtual communication group running over the control path. Fig. 3 illustrates the peer-entity connection for neighboring communications.

We add two modifications for each router as following:

1. Adding an Emergency Recovery (ER) module in a router's control plane, which is a software patch in charge of inquiring/replying the neighbors when failures detected.

2. Setting up the communications between ER and the Intra-TE Resource Management modules. The latter serve for managing the network resources such as a router's interfaces, each link's usage and the sharing of a router by different ASes. Generally, it exists a one-to-one mapping between an Intra-TE Resource Management module and an AS, meaning that if a router connects with K ASes, it will have K such modules each serving for an AS, and all the interfaces connecting to one same AS share a same Intra-TE module. The corresponding module will help evaluate how much bandwidth is lost for the failed AS and how much additional bandwidth at the time can be sold out safely for an available helper.

Below we list the prerequisites for setting up the communication channel.

- An IXP has an unique ID and each router in the IXP knows its collocated IXP ID.
- Each router could get its physically-linked neighbors' IP addresses.
- Each AS has a management center(or called a central controller), which is in charge of managing the whole AS network, just as most of the ASes do currently.
- Each router in an IXP is equipped with an ER module and the Intra-TE Resource Management modules accordingly.

An ER module consists of the following four parts:

1. Neighbor List: it consists of each neighbor's IXP ID, Node ID, active status, IP address as well as the router's own interface No., through this physical interface a direct one-hop communication is enabled.

2. Communication Channels: used for request/reply signaling between the buyers and the potential helpers.

3. Availability Check Sub-module: sending inquiries for checking the availability and replying the requests.

4. Resource Allocation Assistant Sub-module: help to deal with the resource allocation related tasks between the buyers and the sellers.

By exchanging the information about the IXP ID between neighbors, a router can know if its neighbor routers share the same IXP with it. In this way, an ER will build a neighboring list for helping provide the underlying communications inside an IXP.

3.2 Algorithms for Communication Channel Construction

We modify the classical spanning tree to build a communication channel to efficiently spread requests and replies inside an IXP in an ad-hoc manner. The building process of a spanning tree will be initialized by a root node sending an *INIT* message. The root node can be any router in the IXP, *e.g.*, the one with the smallest AS number. The root node starts sending the message to all its neighbors, and the neighbors should forward the message to their neighbors in the IXP except the node who sends the message, and so on. Each router in the IXP marks the transmitter of the first *INIT* message as their parent and all the subsequent message receptions which present the same semantic indications are ignored. The routers should only forward the message once and should send an *ACK* message back to the parent. In this way, each node will know its parent and its children and as a result a bi-directional tree is built.

The pseudo-code that a root node initializes the communication channel building is outlined in Algorithm 1. Neighbors in the algorithm are all the routers which a router connects with. The function *Is_collocation(neighbor)* returns $true$ if a neighbor is in the same IXP.

Algorithm 1 On initializing channel building procedure

```
while (neighbor) do
  if (Is_collocation(neighbor)) then
    Send INIT message(MessID, NodeID);
    neighbor++;
  end if
end while
```

Algorithm 2 describes the actions when a router receives an *INIT* message, where the function *FirstRecINIT()* returns $true$ if the an *INIT* message is received for the first time. If *FirstRecINIT()* is true, update the *ER_table* and then send back an ACK message to the upstream neighboring router (or called parent node). Upon receiving the *INIT* message, each intermediate router will also spread the *INIT* message to its downstream routers except the parent node, where the function *ParentNode()* is $true$ for the message-incoming router. All the following *INIT* messages received later (if any) will be ignored but an explicit *ACK* message is also sent back to the sender (Valid-bit=0) indicating that the machine will not be its child.

Algorithm 2 On receiving *INIT* messages

```
if (FirstRecINIT(MessID)) then
  ER_table ⇐ Entry(NodeID, IP_Address, Interface No.);
  Send back ACK message(MessID, NodeID, Valid-bit(=1));
  /*begin to forward the INIT messages*/
  while (neighbor) do
    if (Is_collocation(neighbor) && NOT ParentNode(neighbor)) then
      Send INIT message(MessID, NodeID);
      neighbor++;
    end if
  end while
else
  Drop INIT message;
  Send back ACK message(MessID, NodeID, Valid-bit(=0));
end if
```

Algorithm 3 is a pseudo-code about the actions when a router receives an *ACK* message, where the function *RecACK()* returns $true$ if a valid *ACK* message is received.

If a node receives no $INIT$ message or no ACK message, it meets problems and reports the exceptions.

Algorithm 3 On receiving *ACK* messages

```
if (RecACK(MessID)) then
  ER_table ⇐ Entry(NodeID, IP_Address, Interface No.);
end if
```

Fig. 4 shows an example for the building process of the communication channel. The hollow nodes inside the large dashed cycle are the routers in the IXP. Suppose router $a1$ initializes the channel building process. Fig. 4(b) draws the *INIT* message flows where a dashed arrow line means the *INIT* message is neglected by the node since it is not the first received one. The solid nodes outside the IXP will not receive the *INIT* messages. By sending *ACK* messages back from receivers to senders, a communication channel is finally established as illus-

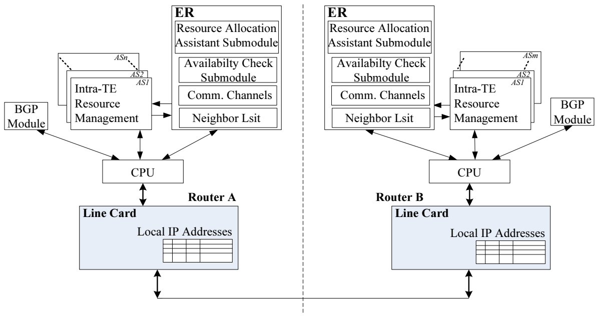
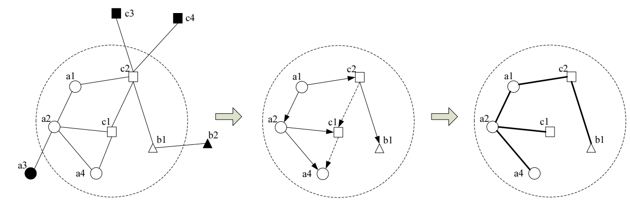


Figure 3: Peer-entity communications between routers in an IXP.



(a) Connection topology. The hollow nodes inside the large dashed cycle are the nodes in the IXP. Different shapes stands for routers of different ASes.

(b) Sending *INIT* messages. A dashed arrow line means the *INIT* message is neglected by the node since it is not the first received one. The solid nodes outside the IXP will not receive any *INIT* message.

(c) The built communication channel

Figure 4: Illustration of the building process of a communication channel.

trated in Fig. 4(c).

3.3 Properties of Communication Channels

Robustness In case routers in an IXP are divided into isolated islands (*i.e.*, the routers in an IXP are partitioned into several disconnected sub-groups) or there is some failure inside an IXP, we also leverage Internet for delivering "help" messages to the potential helpers of an affected AS, as much as the Internet can help. Each AS conducts measurement in advance to get participants in the IXPs where it has presence, via techniques such as those published in [7, 8]. Then once in disaster, it will try to deliver the messages to other ASes in the IXP in addition to using the communication channels within the IXP.

The communication channels can be established before emergency or on-demand. When set up on demand, it can also help cope with failures within an IXP because the spanning tree construction can bypass failed nodes or links.

Small Overhead As for the memory overhead, each router is only required to store its parent and children list in the communication channel (at most N nodes where N is the number of routers in an IXP). When failure happens, each router needs to forward the request to its children and to send the reply to its parent. So it requires to send $O(N)$ messages to broadcast request and receive reply in total, where N is the number of routers in an IXP. So this overhead of building the communication channel is quite lightweight.

Providing Incentive There are several reasons why ASes in an IXP are willing to participate in constructing the communication channel. First, any AS may face the threat from disasters because such disasters happen randomly. In order to protect each AS own benefit, establishing an urgent cooperation alliance could reflect everybody's wish. If an AS fails to participate in such an alliance, it will lose the opportunity of getting help from other AS(es) when emergency happens. Secondly, when fulfilling the required responsibilities, if an AS does not follow the rules (for example stopping forwarding the received message to its neighbors), it is easy to be detected, and this will seriously affect the bad players's credit. Finally, the affected AS can pay the other ASes in the same IXPs for forwarding the communication messages after it finds some available helper(s). Therefore, this communication channel mechanism is of incentives.

Security For exchanging sensitive information as bandwidth in supply/demand, price negotiation *etc*, the buyer and potential helper need to have shared keys for encryption. There are existing measurement techniques such as [7, 8] that an AS can infer other ASes in an IXP, and thus set up public/private keys in advance for future disasters. Note that the key setup does not require a third party, *e.g.*, a public key infrastructure (PKI).

4 Feasibility Evaluation

This section is focused towards answering two main questions: 1) How many potential, available, and actual helpers are there when a large-scale failure like the one caused by Taiwan earthquake happens? 2) Are they sufficient to carry the traffic for the failures?

The experiments are based on 115 IXPs we got from LordofLinks [8] and the Packet Clearing House (PCH) [9]. First, we investigate the efficiency of solely utilizing peering links in the IXPs and the potential of using non-existing but easily establish-able links. Next, we perform a simulated experiment to construct random failures in real IXPs and check how many potential and available helpers we could have in such failures. At last, we check the recovery ability in the Taiwan Earthquake incident and show that even under such a critical disaster, we can still find enough resources by the proposed disaster response system to recover the affected traffic.

4.1 Evaluation Methodology

As described in Section 2.2, the potential helpers are the ASes that belong to the same IXP as the affected AS and might possibly help them reach specific destinations. According to the feedback from an ISP, the following destinations are the most important ones which the ISP would like to maintain a connection to: MSN, DNS root, Google and Yahoo. To check whether the collocated ASes have routes to a specific destination, we study the BGP routing table from Routeview [10] and RIPE/RIS [11] of Dec. 9, 2007, which contains 44 vantage points that are widely distributed in Asia, America, Europe and Africa.

Our reachability checking is based on whether there exists an AS path to the desired destination. For example, we would like to check whether an AS X has route to MSN whose AS number is 8075. If there is an AS path in the routing table like "...X, ..., 8075, ...", AS X is a helper to MSN, otherwise, it is not. It is worthwhile noting that when we check the AS paths of Google and Yahoo, even in normal cases, their appearances are very low. The reason for this should be address aggregation. To tackle this problem we found out the providers of these destinations and assumed that if an AS can reach these providers, it can also reach the destination. Furthermore, we regard an AS as a helper to DNS root, if it can reach any of the 13 DNS root severs. Taking some representative IXPs as an example, Table 1 depicts the number of potential helpers (in these IXPs) to the above four destinations. The most of the participants in a IXP can be potential helpers in normal cases. The number of available helpers is smaller in an outage which will be shown later.

Note that different helpers can provide different bandwidth to a specific destination. If a path from a helper to the desired destination traverses provider-customer links or tier-1 peering links, it is considered as a *high bandwidth path* (HBP). Otherwise, if there is at lease one peering link but not tier-1 peering link, the path from helper to desired destination is recognized as a *low bandwidth path* (LBP). We have obtained tier-1 ASes list from [12] which contains 21 ASes. We further use inferred AS relationship data available in [13, 14] to identify the AS paths indicated in BGP table as HBP or LBP. The data contains 119K relationship entries related to more than 20K ASes. The number of HBP and LBP is also shown in Table 1.

4.2 Are Peering Links Sufficient for Disaster Recovery?

The proposed ER system is able to find both peering ASes and ASes with non-existing business relationship in a same IXP to be a helper.

Usually, an AS knows what networks it peers with. Thus, utilizing the peering links should be much easier than exploring other potential helpers with which it may not have physical links in the same IXP. In this section, we aim to answer the question of how much help peering ASes can provide and whether this help is sufficient for disaster recovery.

4.2.1 How many ASes have peering ASes?

Based on the published data in [13, 14], we investigate the total number of peering links in whole Internet. We observe that the number of peering relationships for each AS is 2.1 in average and the distribution is quit uneven. As shown in Figure 5, 88.24% of the ASes have no peer links and only 6.78% of the ASes have more than 3 peer ASes. Even if we exclude the small ASes which have no customer networks, there are still more than 60% (or 45% after excluding ASes with less than three customer networks) of the ASes without peers as illustrated in Figure 6.

4.2.2 How many ASes without peers have presence in IXPs?

The proposed system can help in the recovery process for the ASes that have presence in the IXPs (or PoPs). It is quite difficult to get the complete list of IXPs existing in the whole Internet. We fetch data from [9, 13], and use the inference method described in [8] to collect information for 115 IXPs which contains 2357 unique ASes [13]. As shown in Figure 7, 48% of the ASes in the IXPs have more than 3 peer ASes, 22% of the ASes have 1 to 3 peer ASes, and 30% of the ASes have no peering ASes at all. Apparently, two thirds of the ASes have the possibility to upgrade the peering links into customer-provider links to find helpers in the disaster. However, are these peers really helpful? We will look into this question in the following section.

4.2.3 How much help can peering ASes provide?

To answer this question, let us take a look at the Taiwan earthquake incident. As we shall reveal at a later stage of this section, we inferred 85 affected ASes from 25 IXPs which need direct recovery. Based solely on these ASes, we evaluate the amount of help offered by peering ASes and then subsequently we compare it with the help offered by the ASes in the same IXPs.

As shown in Table 2, among these 85 affected ASes, 48(56%) of them have peering ASes so that they can find 480 potential helpers (PH) with 303 low bandwidth paths (LBP) and 177 high bandwidth paths (HBP). For the remaining 37(44%) ASes, they cannot be recovered by upgrading the existing peering relationship. While

Name of IXP	# of ASes	#PHs to MSN	#PHs to Google	#PHs to DNS root	#PHs to Yahoo	#PHs of All
HKIX	100	97 (47, 50)	96 (67, 29)	96 (64, 32)	96 (63, 33)	96
JPIX	112	110 (17, 93)	110 (86, 24)	110 (42, 68)	110 (84, 26)	110
AMSIX	289	275 (75, 200)	276 (194, 82)	276 (119, 157)	276 (186, 89)	275
DECIX	220	217 (70, 147)	217 (146, 71)	217 (101, 116)	217 (141, 76)	217
LINX	262	256 (83, 173)	257 (192, 65)	256 (136, 120)	256 (181, 75)	256
PARIX	44	41 (15, 26)	42 (33, 9)	41 (23, 18)	41 (31, 10)	41
AIX	57	54 (16, 38)	54 (40, 14)	54 (25, 29)	54 (35, 18)	53
NYIIX	101	100 (33, 67)	100 (83, 17)	100 (49, 51)	100 (80, 20)	100
SIX	105	101 (23, 78)	101 (71, 30)	101 (40, 61)	101 (61, 40)	101

Table 1: The number of potential helpers and path property they can provide to important destinations. In the parentheses, the first number is the HBPs, the second one is the LBPs.

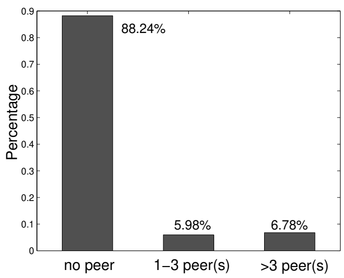


Figure 5: Distribution of the peers for ASes in the Internet.

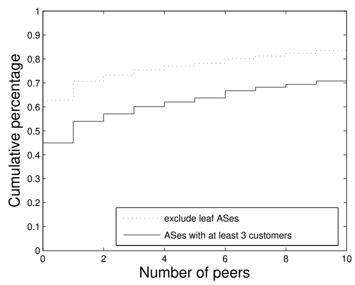


Figure 6: CDF of the peers for ASes with customers in the Internet.

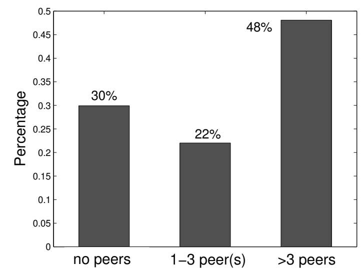


Figure 7: Distribution of the peers for ASes in the our 115 IXPs.

Type	Coverage	#PH	#HBP	#LBP
Peers	56%	480	177	303
IXP ASes	100%	1050	264	786

Table 2: PHs comparison for 85 affected ASes

Type	Coverage	#AH	#HBP	#LBP
Peers	39%	68	26	42
IXP ASes	100%	77	38	39

Table 3: AHs comparison for 85 affected ASes

in the IXPs, we can find potential helper(s) for all of the affected ASes (coverage 100%) if non-existing links are allowed. The total number of potential helpers for these 85 ASes is 1050 (with 786 low bandwidth paths and 264 high bandwidth paths). From this comparison, it is easy to see that the peering ASes provide much less resources compared to the ASes in the same IXPs. Table 3 shows the available helpers (AH) comparison during Taiwan earthquake incident. The main observation from this table is that while all the 85 affected ASes can still find available helpers in the IXPs, only 33 (39%) of them can actually find available helpers from their peering ASes. In other words, the rest 52 ASes cannot find any peering AS as its available helper at all. This implies that when the disaster is severe, the peering AS resource is far from sufficient.

4.2.4 What did the real ISPs do?

We know that during the Taiwan earthquake, several ASes used newly built links for traffic recovery. For example, Bharti (AS9498) built a new link to buy transit from France Telecom (AS5511); China Telecom (AS4134) connected to Chunghwa Telecom (AS3462) to reach Sprint (AS1239) [3]. To strengthen this fact, we checked the BGP paths for 9 heavily affected ASes from the BGP Tables on Dec. 19,2006 (before Taiwan quake) and Jan.1, 2007 (after Taiwan quake). The aim of this part of the experiment is to see if these affected ASes use newly connected neighbors (previously non-existing links) for traffic recovery. Table 4 summarizes the result. It is obvious that, during the earthquake, all these heavily affected ASes have tried to use previously non-existing links for their traffic recovery. Given an option where these ASes could easily find enough helpers simply by upgrading their peering links into customer-provider links for traffic offloading, they probably wouldn't choose to build new links. Thus, with this argument, we believe that using non-existing links is necessary in practice. And with the proposed framework, the seeking process of such links is much faster than manually checking which is the current case for

AS	China Telecom 4134	VSNL 4755	INDOSAT 4761	INDOSAT2 4795	China Netcom 4837	Bharti 9498	Singapore Telecom 7473	CNCNET-CN 9929	TMHK 24077
# of NN	19	10	11	5	3	17	10	5	4

Table 4: Statistics of new neighbor (NN) ASes for 9 heavily affected ASes during the Taiwan earthquake.

ASes.

4.3 Simulation for Disasters with Random Failures

In the simulated disaster scenario, we include 6 big representative IXPs (*i.e.*, AMSIX, DECIX, HKIX, JPIX, LINX, and PARIX), where the topologies and potential helpers are mostly geographically distributed crossing Asia and Europe.

Totally, they have 751 unique ASes. MSN, Google, DNS root and Yahoo are considered as popular destinations according to the feedback from an ISP. We randomly select a number of unique ASes to be incrementally affected in such a simulated disaster. In each round, we choose more ASes to be affected, and we keep a history of the previously chosen affected ASes. Note that the failed ASes should be dependent. For example, if AS1 can access MSN only through AS2, then AS1 will fail if the link between AS2 and MSN is down. If an AS exists in more than one IXPs, it can find potential helpers in all the IXPs it locates.

4.3.1 Results on potential helpers

Starting from 100 affected ASes, we incrementally select an extra 100 in each iteration. In each step, we noted down the number of potential helpers that can be found by the affected ASes in each IXP. In each iteration, We repeat the experiment 100 times and record confidence bars for maximum, media and minimum number of potential helpers. The upper top of the confidence bar is with 95% percentile, the star in the confidence bar is the media value and the lower top of the confidence bar is with 5% percentile. We incrementally select a few ASes to be affected by the disaster. Note that, if an AS exists in more than one IXP, it can find available helpers in all the IXPs it locates. From the left figure in Figure 8, we can observe that when selected number of ASes are affected, there are still lots of other participants in the same IXPs can be potential helpers.

4.3.2 Results on available helpers

We further check for the available helpers in the simulated disaster. In a similar way, we select affected ASes incrementally. When checking for the available helpers, we make sure that they can reach the destination using a safe path. We do not involve those ASes that reach the popular destinations by transiting the other affected AS(es). Suppose that AS1 and AS2 are selected as affected ASes in our experiment. If AS3 is presented in the same IXP as AS1 and has only one BGP path to MSN through AS3, it cannot be an available helper. The right figure of Figure 8 gives the result. The figure demonstrates that during each iteration of choosing ASes to be affected, they can still find sufficient available helpers in the IXP(s) who can help them recover their lost connections. Furthermore, by comparing the two figures, we can find that the safe ASes are not affected much by the simulated disaster.

4.4 Simulation for Taiwan Earthquake Recovery

We try to mimic the Taiwan quake incident to check the efficiency of the proposed method in recovering this big Internet disaster. We collect 299 unique ASes associated with the IXPs in China, Japan, Taiwan, Singapore, India, Philippines, which may possibly be affected by the earthquake. Please note that a number of ASes listed in Table 4, which were recognized as big victims in this quake incident, are included in our disaster dataset.

Considering the fact that 7 out of 9 under-sea cables were severely damaged in Luzon Strait [3], it will not be possible for us to use North American IXPs for recovery mainly due to the following reasons: 1) The links from Asia to North America may have been damaged, 2) The surviving links should be highly congested (note that only 2 survived). For this recovery, we include 25 IXPs from the IXP dataset we had, in which the affected ASes had their presence among which 11 are from Asia and 14 are from Europe. For some affected ASes, we inferred that they did use European ASes as helpers for recovery, *e.g.*, Singapore Telecom (AS7473) used Italy Telecom (AS6762) as a helper, Bharti (AS9498) used France Telecom (AS5511) as a helper [3]. For those are not aware of, we check the connectivity information between the affected ASes and the IXPs where its available helpers were located. For example, China Telecom (AS4134) can find potential helpers in LINX (London), so we first check that the traffic can reach LINX during the disaster and then add available helpers in LINX into the available helpers list of China Telecom. Note that the above 25 IXPs are reachable based on our checking.

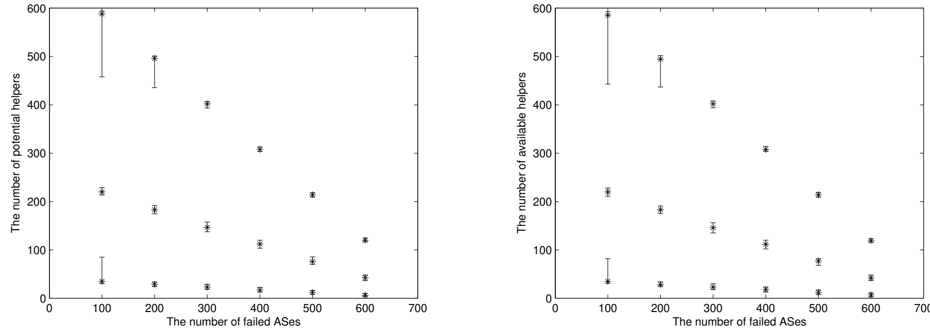


Figure 8: The number of potential helpers (*left*) and available helpers (*right*) in simulated disaster. The figure shows the confidence bars for maximum, media and minimum number of helpers. The upper top of the confidence bar is with 95% percentile, the star point in the confidence bar is the media value and the lower top of the confidence bar is with 5% percentile.

Destination	#affected ASes	#IXP	#PH
MSN	286	25	1050
Google	14	5	245
DNS root	12	4	273
Yahoo	15	5	244

Table 5: Statistical results of the potential helpers (PH) to MSN, Google, DNS root, and Yahoo for the affected ASes.

4.4.1 Results on potential helpers

We checked the BGP table on Sep 12, 2007 for the potential helpers. For each affected AS inferred above, we first find the IXPs it belongs to and then checked the number of potential helpers in each IXP. Table 5 shows the results of potential helpers for different popular destinations. Using the BGP routing tables from RIPE/RIS [11] and Routeview [10], we inferred 286 ASes lost connectivity to MSN, 14 ASes lost connectivity to Google, 12 ASes lost connectivity to all the 13 root DNS severs, and 15 ASes lost connectivity to Yahoo. The reason of disproportionate losses of connectivity to MSN may be MSN does not use distributed servers/mirrors. Since Affected ASes are located in 25 different IXPs, so the total number of PHs for MSN is larger than the number of PHs for other destinations. However, since 286 ASes lost connection to MSN (20 times of other destination), the average number potential helpers for each affected AS is a little bit smaller.

4.4.2 Results on available helpers

For inferring the available helpers, we make the following assumptions about the affected ASes: 1) An affected AS always prefers those helpers that can help it recover connectivity to all its lost destinations simultaneously. For example, if AS A can find its helper list $\{B, C\}$ to MSN and $\{C, D\}$ to Google, it will select AS C as its helper. 2) Recovery can be dependent. Suppose AS A and AS B are victims in the disaster and AS A is the provider of AS B, if we can recover AS A, AS B can also be recovered indirectly through its provider A. We classify these affected ASes hierarchically according to their relationships and try to recover the provider ASes at the top of this hierarchy. Based on these two assumptions, we get 85 affected ASes which need direct recovery. For available helpers, in a similar way, we classify them hierarchically. If both AS C and AS D can provide help to AS A, and D is C's only provider to A's desired destination, the total bandwidth provided by both C and D to A cannot exceed the bandwidth capacity D could provide. Table 6 gives the available helpers we found in the 25 IXPs. Please note that *AHs in the last column means the subset of AHs in the previous column that can reach all of the four destinations simultaneously. For example, 14 out of the 235 available helpers for the 15 affected ASes of Yahoo can also reach MSN, Google, and DNS root. Since the routes to MSN are more heavily affected, the number of AH for MSN is smaller and the number of *AH is larger.

Destination	#affected ASes	#IXP	#AH	#*AHs
MSN	286	25	77	77
Google	14	5	233	14
DNS root	12	4	239	13
Yahoo	15	5	235	14

Table 6: Statistical results of the available helpers (AH) to MSN, Google, DNS root, and Yahoo for the affected ASes. *AHs means the subset of AHs in the forth column that can reach all of the four destinations (MSN, Google, DNS root and Yahoo) simultaneously.

We further distinguish the helper path into high bandwidth path and low bandwidth path. For high bandwidth path, we assume that it can provide 2.5 Gbps, while for low bandwidth path, we assume that it can provide 1 Gbps. The assumption is based on our survey from a real ISP. During the disaster, it bought 5 Gbps and 2.5 Gbps from two different ASes to recover its traffic. Here, we conservatively use the minimal value as the bandwidth of our high bandwidth path. Among the 25 IXPs we find 77 unique available helper ASes who can help to recover traffic to MSN, Google, DNS root, and Yahoo simultaneously. The total available bandwidth from these 77 ASes is 134 Gbps (38 high bandwidth paths and 39 low bandwidth paths). Recall that 7 out of 9 links have been damaged during the Taiwan earthquake. Suppose each link has the physical capacity of 10 Gbps, the total impacted bandwidth cannot exceed 70 Gbps. Intuitively, we have sufficient bandwidth.

Our result shows that, for each of the affected ASes, there is at least one available helper. We also find that the available helpers for different ASes were distributed quite unevenly. Figure 9 shows the cumulative distribution of available helpers for these affected ASes. We observe three significant projections, 19% of the affected ASes can have only 1 available helper, 15% of the affected ASes can have 4 available helpers, and 31% of the affected ASes can have 8 available helpers. We can also see that 18% of the affected ASes can find more than 10 helpers. Figure 10 depicts the relation between the number of available helpers of an AS and the number of IXPs it belongs to. From the figure, we can see that a lot of affected ASes are only present in one IXP. But we observed that the number of helpers that they can find varies quite significantly. More interestingly, we can roughly observe two lines (top and bottom) in the figure. The reason for this is that there are three IXPs which can provide relatively large number of available helpers. For affected ASes, if they have presence in three IXPs, then they can find more AHs and move upwards towards the top line otherwise, they stay along the bottom line. Note that, in this experiment, we only include a limited number of IXPs (25) for recovery. The more the IXPs that are included in the recovery, the better the result should be.

4.4.3 Results on Disaster Recovery

The optimal bandwidth assignment can be formulated as shown in (1-4) which describes the upper bound of the bandwidth assignment. The descriptions of the notations are illustrated in Table 7.

$$Obj. \qquad \max \sum_i \frac{\sum_{j,\,k} g^i_{jk}}{R_i} \qquad (1)$$

$$st. \qquad \sum_{j,\,k} g^i_{jk} \le R_i, \qquad \forall i \in \{1, 2, \cdots, n\} \quad (2)$$

Notations	Descriptions
P	The set of available helpers $P = \{P_1, P_2, \cdots, P_j, \cdots, P_m\}$).
C	The set of buyer networks $C = \{C_1, C_2, \cdots, C_i, \cdots, C_n\}$).
B_j	The maximum bandwidth the provider j can provide.
R_i	The required bandwidth by C_i.
g^i_{jk}	The bandwidth provided by P_j for C_i in IXP k.
x_{ik}	$x_{ik} = 1$ if buyer C_i has at least one router in IXP k; otherwise, $x_{ik} = 0$.
y_{jk}	$y_{jk} = 1$ if provider P_i has at least one router in IXP k; otherwise, $y_{jk} = 0$.

Table 7: Table of notations

$$\sum_{i,\,k} g^i_{jk} \le B_j, \qquad \forall j \in \{1, 2, \cdots, m\} (3)$$

$$0 \le g^i_{jk} \le x_{ik} y_{jk} R_i, \quad \forall i, j, k \qquad (4)$$

We define the welfare of a buyer network as $\frac{\sum_{j,\,k} g^i_{jk}}{R_i}$, and define the social welfare as $\sum_i \frac{\sum_{j,\,k} g^i_{jk}}{R_i}$. We propose the following model to assign the bandwidth with the object of maximizing the social welfare. The object function (1) represents the total welfare of buyers. The constraint (2) ensures that the bandwidth allocations for a buyer AS is no more than its demand; (3) is the constraint of available resource; (4) states that g^i_{jk} cannot be negative, and the buyer can only buy the bandwidth from the providers collocated in the same IXP. The objective function and the constraints are all linear and can be easily transform to the standard form of linear programming.

In our implementation, we got the real traffic volume from a big ISP that was affected by the disaster. For those ASes for which the traffic requirement was not known, we generated the volumes based on the number of destinations they wanted to reach, the number of customer ASes they have, and the traffic demand of the real ISP we have already known. Specifically, we assume each AS i has a basic requirement of $100Mbps$ for each of the four popular destinations. If it has more customers, we incrementally increase R_i by $100Mbps$ for each customer. For each helper AS j, as discussed before, if it can provide HBP we assume its B_j as 2.5Gbps, otherwise $1Gbps$.

With the bandwidth resources we found in the IXPs, we can easily solve the Taiwan quake incident even without computing the optimal result. Note that the main contribution of this paper is not on how to efficiently allocate the bandwidth we inferred from IXPs, instead, we just randomly select one heuristic solution satisfying the constraints(2-4). Specifically, in our implementation, each buyer network will have a list of available helpers which can provide varying bandwidth, we allow

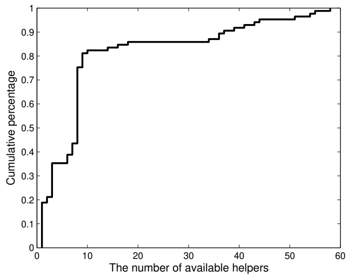


Figure 9: CDF – the number of AHs we found in the Taiwan earthquake.

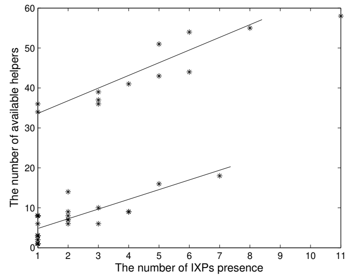


Figure 10: Scatter plot – the correlation of the number of AHs for an AS with the number of IXPs it locates.

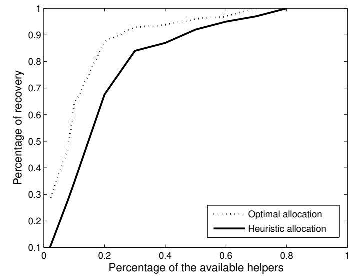


Figure 11: The Taiwan earthquake recovery curves.

Algorithm 4 The heuristic allocation

/* Each buyer b_k has access to a subset of sellers s_k*/
for ($i = 0$ to k) **do**
  while (TRUE) **do**
    randomly select a seller ($rand_seller$) from s_k;
    helperlist($rand_seller$);
    /* update helper list*/
    update seller list;
    if (bandwidth requirement of buyer is satisfied or no sellers) **then**
      break;
    else
      continue;
    end if
  end while
end for

each buyer to greedily buy bandwidth from any of its helper list. The algorithm is illustrated in Algorithm 4. Obviously, if we could implement optimal mechanism, the result should be not worse than what we achieved by heuristic method. Here, we evaluate the traffic offloaded percentage, which is defined as the ratio of recovered traffic and the demand traffic. We randomly select a certain percentage of available helpers that is willing to be actual helpers, and calculate how many requests can be satisfied.

Figure 11 shows the result we get from the heuristic allocation and optimal allocation by solving the linear programming in (1-4). Optimal curve describes the recourses we found to recover the traffic to the popular destinations. It is obviously that only 20% of the available helpers can have the ability to recover 90% of the traffic requirement. And 70% of them can have the ability to recover all the traffic. The reason why we have this gap is that some disaster ASes find less helpers than the others, only most of the these helpers would like to help can the traffic be well offloaded. This is a limitation of our experiment here, however, with more IXPs involved, we think the situation can be greatly improved. Heuristic curve just depicts one random allocation we used as described above. It can be more realistic for practical operation. Although this random algorithm is not the best effort, we can see that 50% of the available helpers can help to recover 90% of the traffic, while 80% of them can recover all the traffic.

5 Related Work

Although the routing on today's internet demonstrated remarkable availability and responsiveness on average, there is a serious lack of resilience to significant emergencies [15]. Several previous work [4, 5, 16] started to develop techniques which can be employed to improve the network resilience recently, but they are quite different from our work. Before illustrating the detailed design space of each failure recovery method, we first present failure types in realistic and extra connectivity can be utilized for recovery.

Based on the type of damaged links, we classified failures into the following three categories.

1. **Intra-domain failure.** An AS breaks into a few isolated parts and only intra-domain links are damaged.

2. **Inter-domain failure.** An AS disconnects from its providers or customers or peers. Only Inter-domain links are down.

3. **large-scale hybrid failure.** A regional failure causes several ASes suffering reachability problem, which damages both intra-domain and inter-domain links.

When a path between two ASes is down, there are roughly four kinds of connectivity can be utilized to reconnect them.

1. If only a few but not all of the policy-compliant connections between two ASes fail, another policy-

compliant path can be used. BGP itself has the ability to discover and utilize such kind of connectivity, but suffers packet loss during the discovery and switching to another policy-compatible path.

2. When all the policy-compliant connections between two ASes fail, a few policy-violative connections are still useful to reconnect two ASes. For example, we can relax BGP constraint by upgrading peer links to provider-customer links to keep the connection for Affected ASes.

3. Non-existing but easily connected links can be also employed to recover traffic. For example, to connect to another helper ASes in a same IXP.

4. Build dedicated backup infrastructure.

The comparison of design space is shown in Table 8, where failure number and connectivity number are described as above.

methods	REIN	R-BGP	ERN	IER
failure #1	✓	×	✓	✓
failure #2	×	✓	✓	✓
failure #3	×	×	✓	✓
connectivity #1	×	✓	×	×
connectivity #2	✓	×	×	✓
connectivity #3	×	×	×	✓
connectivity #4	×	×	✓	×

Table 8: Design spaces of different methods.

In paper [4], the authors propose a solution framework called reliability as an inter-domain service (REIN) to improve the redundancy of IP networks across multiple networks. The idea is to recover the intra-domain failure through inter-domain link with neighbors. This framework is efficient for intra-domain failures (Failure #1), but it can not recover inter-domain traffic, and large scale disaster like Taiwan earthquake incident. Furthermore, REIN only uses the existing inter-domain links (connectivity #2) for traffic recovery.

R-BGP is proposed in [5] to ensure that Internet domains stay connected as long as the underlying network is connected. R-BGP works by pre-computing a few strategically chosen fail over paths. R-BGP provably guarantees that a domain will not become disconnected from any destination as long as it will have a policy-compliant path to that destination after convergence. Since R-BGP is based on BGP, it utilize policy-compliant paths (Connectivity #1) and works for inter-domain failures (failure #2). In fact, R-BGP aims to eliminate the packet loss caused by BGP dynamics and does not improve the ability to failure resilience.

Dedicated Emergency Response Network (ERN) is introduced in [16] to sustain critical communications in emergencies. The idea is to use Advanced Metering Infrastructure (AMI), surveillance camera mesh networks, and enhanced cell phones for emergency communication when disaster happens. ERN aims to solve large scale failures(failure #3), however, to build such n dedicated backup infrastructures (connectivity #4) is expensive and there are still many unsolved challenges in this proposal, for example, the detection of an emergency, the platform support and the topology coverage.

Different from above three methods, IER proposed in this paper can make use of policy-violative links (upgrade peer link into provider-customer links) as well as non-existing but easily connected links (set up new links) to solve all kinds of failures. Without building new network infrastructure, IER is only method which can utilizes all the available connectivity for failure recovery.

6 Conclusion

In this paper, we have proposed a framework for addressing Internet disasters. The design philosophy behind our framework was not to achieve a completely automated disaster response system but to speedup the recovery of the connectivity of the Internet after a major disaster. The idea is to upgrade existing links or to connect non-existing but easily connected links to help recover the traffic in affected ASes. We have designed a communication channel within an IXP, via which affected ASes in disaster can find all the available helper ASes for traffic recovery. In the paper, we check the efficiency of using existing links and non-existing but easily connected links by performing simulated disaster experiments and mimicing Taiwan earthquake scenario. The experiments show that even under critical situations like this Taiwan earthquake we can still find enough resources to recover the affected traffic through proposed emergency response system.

References

[1] GENI Planning Group, "GENI: Conceptual design, project execution plan," in *GENI Design Document*.

[2] NANOG, "North American Network Operators Group Mailing List Archive," http://www.merit.edu/mail.archives/nanog/.

[3] S. Wilcox, "Quaking Tables: The Taiwan Earthquakes and the Internet Routing Table," http://www.thedogsbollocks.co.uk/tech/0705quakes/AMSIXMay07-Quakes.ppt, 2007.

[4] H. Wang, Y. R. Yang, P. H. Liu, J. Wang, A. Gerber, and A. Greenberg, "Reliability as an interdomain service," in *SIGCOMM'07*, 2007.

[5] N. Kushman, S. Kandula, D. Katabi, and B. M. Maggs, "R-BGP: staying connected in a connected world," in *USENIX/SIGCOMM NSDI'07*, 2007.

[6] W. B. Norton, "Interconnection Strategies for ISPs," http://www.equinix.com/pdf/whitepapers/ISPInterconnectionStrategies2.pdf, 1999.

[7] K. Xu, Z. Duan, Z.-L. Zhang, and J. Chandrashekar, "On Properties of Internet Exchange Points and Their Impact on AS Topology and Relationship," in *IFIP Networking'04*, 2004.

[8] Y. He, G. Siganos, M. Faloutsos, and S. V. Krishnamurthy, "A systematic framework for unearthing the missing links: Measurements and Impact," in *USENIX/SIGCOMM NSDI'07*, 2007.

[9] Packet Clearing House, "Internet exchange directory," http://www.pch.net/resources/data.php?dir=/exchange-points.

[10] Routeviews, http://archive.routeviews.org/oix-route-views/2007.09/.

[11] RIPE, "Routing Information Service," http://www.ripe.net/projects/ris/.

[12] N. Spring, R. Mahajan, and T. Anderson, "Quantifying the causes of path inflation," in *SIGCOMM'03*, 2003.

[13] Y. He, G. Siganos, M. Faloutsos, and S. Krishnamurthy, "The Lord of the Links," http://www.cs.ucr.edu/~yhe/LordOfLinks/.

[14] CAIDA, "AS ranking," http://as-rank.caida.org.

[15] J. Wu, Y. Zhang, Z. M. Mao, and K. G. Shin, "Internet routing resilience to failures: Analysis and implications," in *CoNEXT'07*, 2007.

[16] M. LeMay and C. A. Gunter, "Supporting Emergency-Response by Retasking Network Infrastructures," in *HotNets-VI*, 2007.

A Signaling protocol

We define five categories of messages for the signaling operations:

(1) Helper_Request: This message originates from a buyer. It includes the message ID, ASN, Nodes ID and public key, LOG and list of the specific networks to which it lost connectivity. Optional parameters include the demand bandwidth and the price enquiring.

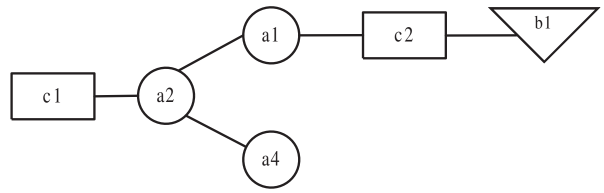


Figure 12: Example networks used in the description of signaling protocol.

(2)Helper_Request_Reply: Dispatched by a receiver. It contains the message ID, the ASNs of receivers and the buyer, the receivers ID and the LOG. Optional parameters include the provided bandwidth and the price.

(3) Reply_ACK: Acknowledged by a buyer to a receiver indicating that the buyer has safely received the message of Helper_Request_Reply, (or the Helper_Accept_Reply, in the following item (5)). It contains the message ID, the ASNs of both the receiver and the buyer, and the LOG.

(4) Helper_Accept: It originates from a buyer. It includes the message ID, the ASNs, public-key and the IP address(es) of the buyer AS, the LOG, and the bandwidth the buyer wants to buy as well as the bugging-price (if any) as options.

(5) Helper_Accept_Reply: Dispatched from an available helper to a buyer, it includes the message ID, the ASN, public-key and the IP address(es) of helper's AS, the LOG, the adjusted(if any) bandwidth and the changed-bugging-price.

Please note that due to the possible initial inconsistence between a buyer and a seller, the above messages 4 and 5 may be negotiated iteratively. For simplicity, we describe the signaling procedure with the network topology indicated in Fig. 12.

Consider the scenario where a buyer network c is suffering in an emergency, and its border router $c1$ spreads a Helper_Request message to all the routers(ASes) in the same IXP through the established communication channel. Fig. 12 shows the communication spanning tree for buyer $c1$. First $c1$ sends a Helper_Request message to its neighbor $a2$, $a2$ will first run an integrity check to see whether the message is tampered by the preceding node (to maintain the public-key unchanged). If the message passes the integrity test, the node will parse the message and relay this message to its neighbors $a1$, $a4$ and so on. Then $a2$ will do the accessability check to the specific networks indicated in the request.

After $a2$ has finished the available check, it will send a Helper_Request_Reply message back to the buyer. This operation is also acknowledged by a buyer sending back a Reply_ACK message. If $a2$ connects to the specific networks it will become an available helper. The remaining routers(ASes) in the spanning tree will do the same with $a2$. Finally, $c1$ will receive all the nodes' replies. After $c1$ makes a decision, it indicates the acceptance of a helper (for example here $b1$), by replying to a Helper_Accept message with possibly a charged bandwidth and price attached. Then $b1$ will send back $c1$ a Helper_Accept_Reply, for further negotiations.

B Business models for bandwidth demand/supply

We illustrate three possible mechanisms among providers and buyers.

Free bid mechanism

When one plans to buy something, he/she first collects the information about who can provide it. If there is more than one provider, he/she selects with his/her own preference. The free bid mechanism in the Internet emergency response context stems from a similar idea. The failed AS runs the ASes/routers discovery in the IXPs and gets a list of available helpers, then it negotiates with the selected providers for bandwidth and price agreement.

Broker based mechanism.

If we can set up a broker in IXPs who buys a certain amount of bandwidth from the potential helpers and then resells it to the failed ASes in a more flexible way, the broker, the buyers and the providers could all be benefited. The broker earns money from it, the buyers can save money by avoiding accidental selection of larger bandwidth levels, and the providers save cost of possibly building agreements with a large number of buyers.

Double auction mechanism.

Providers bid the per-unit bandwidth price and the amount of the bandwidth they want to sell in each IXP, and buyers do the same for what they want to buy. When the price and bandwidth amount meet, a deal is struck. The bandwidth in an IXP is just like the different stocks in a stock market which have different prices. Buyers and providers can withdraw and modify their bids. Furthermore, the bandwidth amount can be divided and a deal can be struck entirely or partly only when a buyer's price and a seller's price meet.