Network Working Group
Internet-Draft
Expires: September 2, 2012

M. Xu
J. Wu
S. Yang
Tsinghua University
D. Wang
Hong Kong Polytechnic University
Mar 2012

Two Dimensional IP Routing Architecture
draft-xu-rtgwg-twod-ip-routing-00

Abstract

This document describes Two Dimensional IP (TwoD-IP) routing, a new Internet routing architecture which makes forwarding decisions based on both source address and destination address. This presents a fundamental extension from the current Internet, which makes forwarding decisions based on the destination address, and provides shortest single-path routing towards destination. Such extension provides rooms to solve fundamental problems of the past and foster great innovations in the future.

We present the TwoD-IP routing framework and its two underpinning schemes. The first is a new hardware-based forwarding table structure for TwoD-IP, FIST, which achieves line-speed lookup with acceptable storage space. The second is a policy routing protocol that flexibly diverts traffic.

Status of this Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at http://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on September 2, 2012.

Copyright Notice

Table of Contents

1. Introduction

Since IP routing took place, the current Internet has been making forwarding decisions based on destination addresses. The destination-based routing system provides limited semantics with only a single path towards each destination. Many services, such as multi-homing, multi-path and traffic engineering, face difficulties within the current Internet routing system. Due to the important semantics of source address, recent years see increasing works on adding source addresses into routing controls.

IP source routing [3] carries the routes in packet header. However, IP source routing is disabled in most networks due to security reasons. MPLS [4] uses label switching to manage traffic per-flow. However, MPLS raises scalability issues when the number of label switching paths (LSPs) increases [5]. What’s more, many ISPs prefer pure-IP networks.

In this draft, we describe Two Dimensional IP (TwoD-IP) routing, which makes forwarding decisions based on both source and destination addresses. TwoD-IP routing presents a fundamental extension of the semantics from the current Internet. The network will become more flexible, manageable, reliable, etc. Such extension provides rooms to solve problems of the past and foster innovations in the future.

TwoD-IP routing framework is divided into data plane and control plane. In data plane, packet forwarding needs to check both source and destination addresses. Though current TCAM-based forwarding table can match line speeds with parallel search over the table, with one more dimension in the table, the forwarding table will explode and exceed the maximum storage space of current TCAM. We devise a new forwarding table structure for TwoD-IP, FIB Structure for TwoD-IP (FIST). The new structure makes a separation between TCAM and SRAM, where TCAM contributes to fast lookup speeds and SRAM contributes to a larger memory space. In the control plane, we devise a simple policy based routing protocol. For the traffic of a customer network of an ISP, this policy routing protocol can flexibly divert the traffic from one edge router to another edge router.

This document also presents the deployment issues and objectives of the TwoD-IP routing.

2. Benefits of Introducing TwoD-IP Routing

In this section, we list the use cases that can benefit from TwoD-IP routing.

2.1. Multi-homing

Multi-homing is prevalent among ISPs for better traffic distribution and reliability. Traditionally, Provider Independent (PI) address is used. Because PI address can not be aggregated by higher level ISPs, it will cause explosion of routing table. To solve the problem, Provider Aggregatable (PA) address is proposed. However, PA address complicates network configurations for ISP operators. Besides, due to destination-based routing in traditional networks, PA address has difficulties when facing failures, i.e., the network has to re-compute a new path when failures happen.

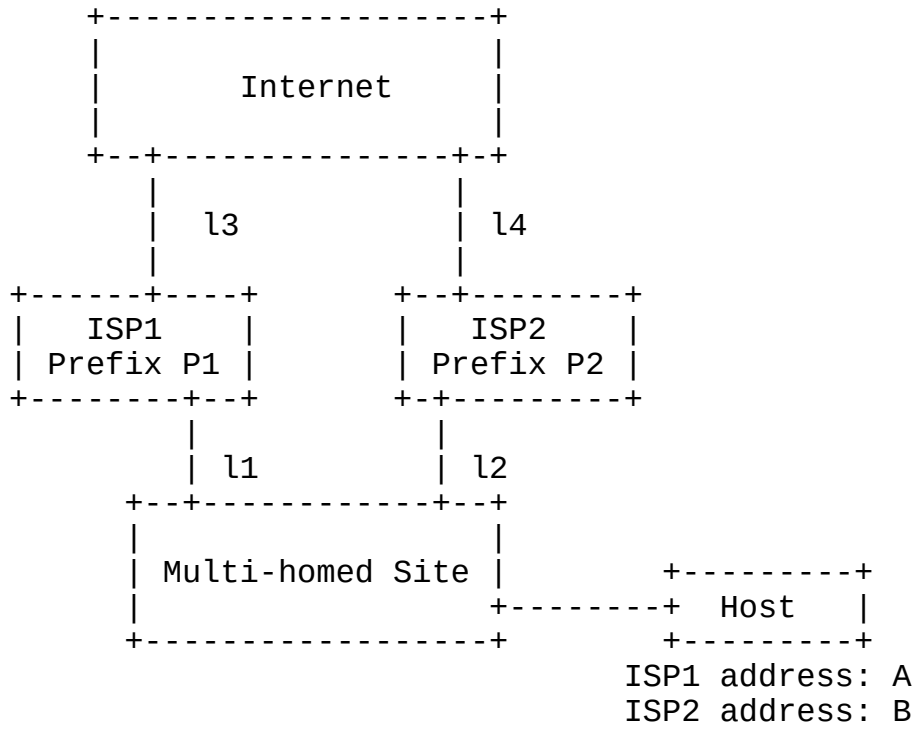

Figure 1: TwoD-IP routing for multi-homing

For example, in Figure 1, assume a multi-homed site is connected to two ISPs: ISP1 and ISP2. ISP1 has a prefix P1, and ISP2 has a prefix P2. A host connect to the multi-homed site has two addresses,

address A that can be aggregated into P1, and address B that can be aggregated into P2. With TwoD-IP routing, the multi-homed site can deliver the traffic from A towards the Internet to ISP1, and deliver the traffic from B towards the Internet to ISP2. If the host is using address A, and the link l1 or l3 fails. Then the host can immediately detect the failure, then switch to address B, and continue to communicate with the Internet via ISP2. With TwoD-IP, the host does not have to wait for routing convergence in the multi-homed site when failures happen.

2.2. Load Balancing

Compared to destination-based routing, TwoD-IP routing can manipulate traffic in a finer-grained granularity. Such that TwoD-IP can achieve better traffic distribution. For example, in Figure 2, assume that there are 5 hosts that are communicating with the same server at 10Mbps. Our goal is to minimize the maximum link utilization over the network. Within destination-based routing, traffic towards the same destination has to travel along the same path in the network. Thus the best traffic distribution is to let all traffic take the north route via router b, and the Min-max link utilization is 83.3%.

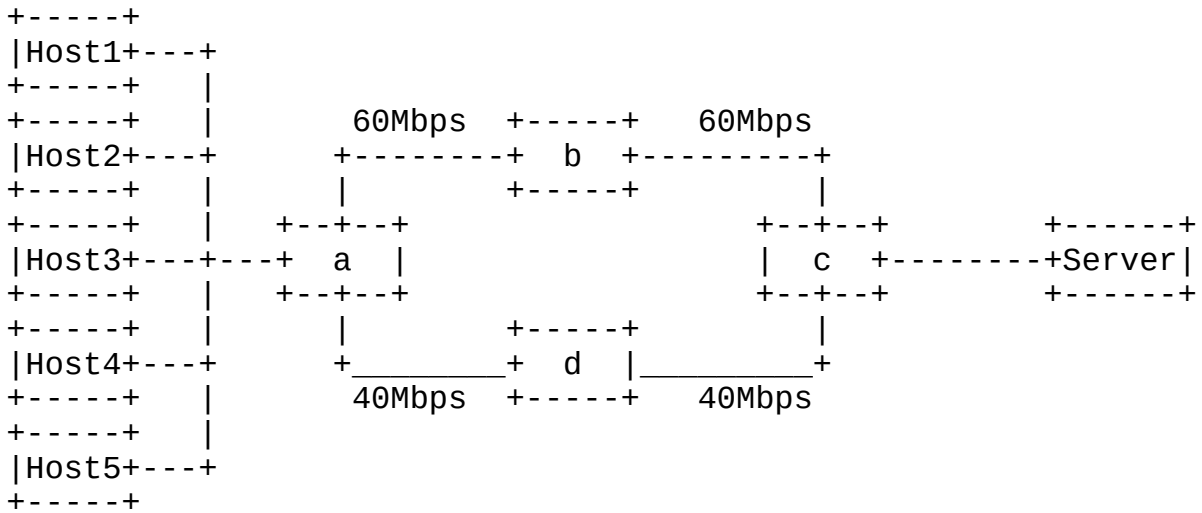

Figure 2: TwoD-IP routing for load balancing

With TwoD-IP routing, we can let the traffic of three hosts (e.g., Host1, Host2 and Host3) take the north route via b, and let the traffic of the other two hsots (e.g., Host4 and Host5) take the south route via d. Thus the Min-max link utilization is only 50.0%.

2.3. Diagnosis

In Figure 3, assume a network has four routers: a, b, c and d. The operator wants to monitor the status of the link between b and d. Thus the operator sets up a monitor at router a, and sends probe packets to d. Theoretically, the probe packets will flow on the shortest path, i.e., a-b-d. However, the network provides traffic engineering capabilities. If there is congestion on the link between b and d, and router b moves the traffic towards d to the north path via c. Thus, the probe packets will flow on the path a-b-c-d, which does not include the link between b and d.

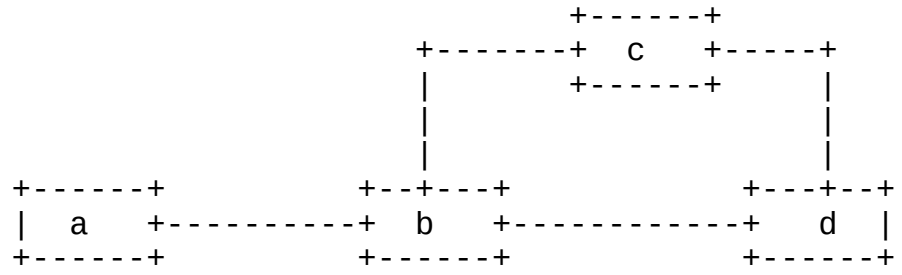

Figure 3: TwoD-IP routing for Diagnosability

With TwoD-IP routing, router b can identify the probe packets from a towards d, and deliver them directly to router d. Thus the link between b and d can be easily monitored.

2.4. Policy Routing

Assume in an ISP network, ISP operator wants that the traffic from source address A towards destination address B passes by router C. With TwoD-IP routing, routers make forwarding decisions based on both destination and source addresses, thus can easily identify the traffic from A towards B, and divert it to the next hop towards C.

2.5. Others

Besides the above-mentioned use cases, TwoD-IP routing is beneficial in many other use-cases. We list the other use-cases briefly.

- Reliability: TwoD-IP provides multiple paths towards destination, rather than the shortest path only. When one path breaks down, routers can immediately switch to another path.

- Multi-path: TwoD-IP can forward packets towards the same destination, and from different sources to different next hops.

If a host has multiple source addresses, the host will have multiple paths towards the same destination.

- Security: Traditional network pushes the security devices to the border routers, the intermediate network just delivers the packets. With TwoD-IP, intermediate routers also have source checking functionality. Thus, the whole network rather than the border network, can defense attacks.

- Measurability: With TwoD-IP, ISP operators can explicitly control the routing paths of probe packets. Thus the number of monitors, and the additional traffic caused by the probe packets, can be reduced [6].

3. Framework

Similar with traditional routing, TwoD-IP routing can be separated into two parts: data plane and control plane.

3.1. Data Plane

Data plane contains the forwarding table, that decides what to do when a packet arrives. Different with traditional destination-based routing, each entry in the TwoD-IP routing forwarding table is a 3-tuple: {destination address, source address, next hop}. When a packet arrives, routers extract both destination and source addresses from the packet, then lookup the forwarding table, and output a matched entry. Finally, routers will forward the packet to the next hop associated with the matched entry.

With a new dimension, the size of forwarding table will increase to be O(N^2) (where N is the size of source/destination address space), which is too large for current TCAM-based storage to accomodate. To avoid forwarding table explosion, we design a new forwarding table structure in Section 4.

3.2. Control Plane

In traditional routing, the control plane is concerned with the network status, e.g., network topology. Within TwoD-IP routing, the control plane is concerned with both network status and user demands. TwoD-IP routing not only provides basic connectivity service, but also satisfies kinds of user demands, e.g., policy routing, multi-path and traffic engineering. TwoD-IP routing protocol has two components:

- Destination-based routing protocol: To be compatible with traditional routing (especially when most networks only support destination-based routing), TwoD-IP routing protocol should support destination-based routing. Such that ISPs can provide the same connectivity service, while upgrading routers with TwoD-IP functionality. To provide better connectivity services, destination-based routing protocol should respond instantly to the changes of network topology.

- Source-related routing protocol: Combined with source addresses, TwoD-IP routing can make better forwarding decisions for users. Source-related routing protocols focus on providing services that are related with source addresses. They may need to collect demands from users, and compute the routing table to satisfy these demands. Depending on the specific user demands, some source-related routing protocols need real-time updates, while others do

not. The newly designed source-related routing protocols should be:

* Consistent, they should be consistent with other routing protocols, including the destination-based routing protocol and other new source-related routing protocols;

* Efficient, they should not bring lots of additional overheads to the network.

4. Forwarding Table Design

4.1. Design Goals

The forwarding table stores a set of 3-tuple rules, {pd, ps, nh}, where pd is a destination prefix, ps is a source prefix, and nh indicates the next hop. When a packet arrives, if its destination address matches pd according to LMF (longest match first) rule among all rules, and its source address matches ps according to LMF rule among all rules that are associated with pd. Then the router will forward the packet to the next hop nh.

The new forwarding table should satisfy the following requirements.

- Storage requirement: The new forwarding table should not cause forwarding table explosion problem. Current storage technology should be able to accomodate the table.

- Speed requirement: The new forwarding table should match line-speeds.

4.2. Forwarding Table Structure

We design a new TwoD-IP forwarding table structure, called FIST. As shown in Figure 4, FIST consists of four parts.

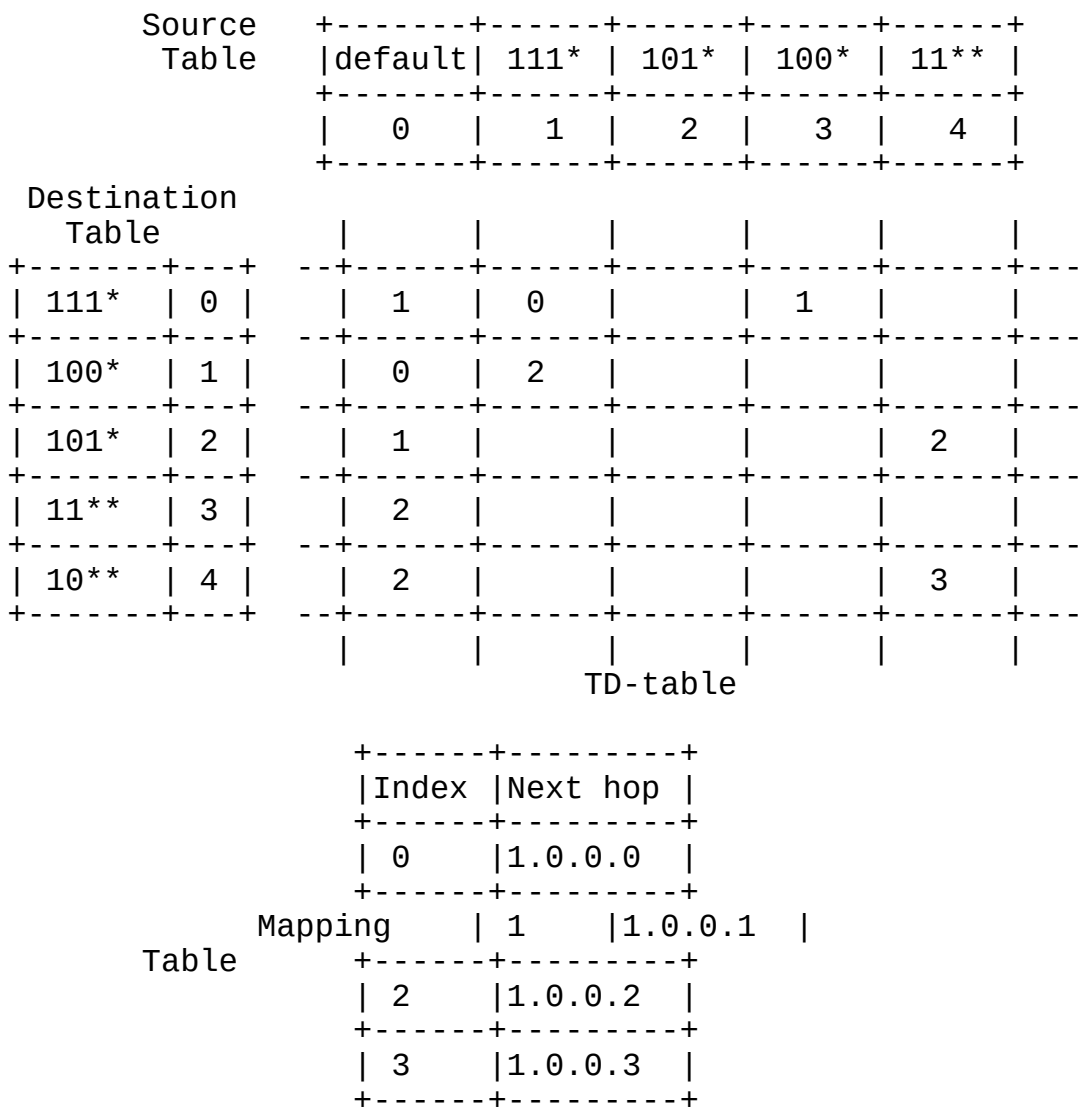

Figure 4: Forwarding Table for TwoD-IP

- Destination table: It resides in TCAM, and stores the destination prefixes. Each destination prefix in destination table corresponds to a row number.

- Source table: It resides in TCAM, and stores the source prefixes. Each source prefix in source table corresponds to a column number.

- Two Dimensional Table (TD-table): It is a two dimensional array that resides in SRAM. Given a row and column numbers, we can find a cell in TD-table. Each cell in TD-table stores an index value, that can be mapped to a next hop.

- Mapping table: It resides in SRAM, and maps index values to next hops.

For example, in Figure 4, the destination table contains 5 destination prefixes, and the destination prefix 101* corresponds to a row number 2. The source table contains 4 source prefixes and a default one (can be see as "****"), the source prefix 11** corresponds to a column number 4. The TD-table has 5 rows and 5 columns, the cell that is in the 2nd row and the 4th column has index value 2. In the mapping table, we can see that the index value 2 is related with the next hop 1.0.0.2.

If destination prefix pd outputs row number n, and source prefix ps outputs column number m, we use (pd, ps) to denote a cell in the nth row and mth column of the TD-table.

4.3. Lookup Action

When a packet arrives at a router, the lookup action is as follows.

1. Extract the destination address d and source address s from the packet;

2. Perform the following two operations in parallel:

   * Lookup the destination address d in the destination table using the LMF rule, and output the row number n;

   * Lookup the source address s in the source table using the LMF rule, and output the column number m;

3. Lookup the cell that is in the nth row and mth column of the TD-table, and output the index value v;

4. Lookup v in the mapping table, and ouput the corresponding next hop;

5. Forward the packet to the next hop.

The 2nd step takes one TCAM clock cycle to match both d and s, and one SRAM clock cycle to get the row/column number. The 3rd step takes one SRAM clock cycle to get the index value, the 4th step takes one SRAM clock cycle to get the next hop. Thus, the lookup speed is one TCAM clock cycle plus three SRAM clock cycles. Beside, the lookup process can be pipelined to achieve higher speed.

4.4. Update Action

Although FIST can reduce TCAM storage space, and achieve fast lookup speed, it also faces new challenges. The challenges are caused by performing LMF rule on source addresses. Assume a packet should match destination prefix pd, and source prefix ps. However, if the source table contains a source prefix ps' that also mathces the packet and is longer than ps, then the packet will match (pd, ps') within FIST.

For example, if the forwarding table on a router is shown in Figure 4, and a packet with destination address of 1011 and source address of 1111 arrives on the router. According to the matching rule, destination prefix 101* is matched first, and source prefix 11** should be matched. However, within FIST, destination prefix 101* and source prefix 111* are matched. But the cell (101*, 111*) is empty.

To resolve the confliction, we should pre-compute and fill the empty cell with appropriate index value. For example, in Figure 4, we should fill the cell (101*, 111*) with the index value 2, that is the index value of cell (101*, 11**). We will discuss the update action in the next version of this document.

4.5. Scalability Improvements

In Section 4.2, we design the FIST structure, where each destination prefix corresponds to a row, and each source prefix corresponds to a column. Considering the large number of address prefixes, we can make improvements in the following two aspects:

- Not every destination prefix need to be mapped to a row, because ISPs only need to divert traffic for part of the destination prefixes. The destination table of FIST should be divided into two parts, each destination prefixes in the first part points to a row and each destination prefix in the second part points directly to an index value.

- Different destination/source prefixes can be mapped to the same row/column, because ISPs may implement the same policy on different prefixes. For example, ISPs wants to divert the traffic of some customer network, which has multiple prefixes, to another path.

5. Routing Protocol Design

5.1. Protocol Overview

In this section, to illustrate TwoD-IP routing protocol, we design a simple policy routing protocol. The routing protocol provides a flexible tool for ISPs to divert traffic (that is from some customer networks towards the foreign Internet) to another path.

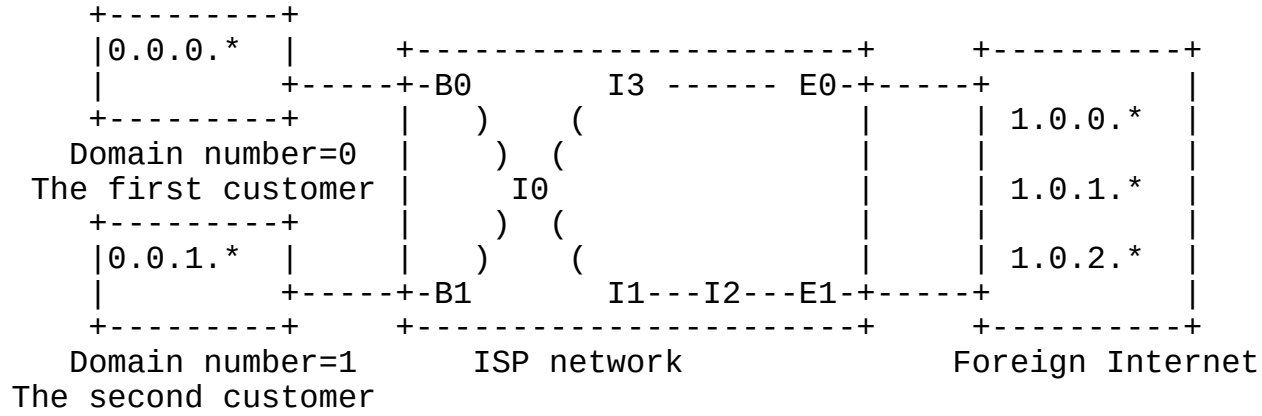

Figure 5: A simple policy routing protocol

For example, in Figure 5, the ISP has two customer networks, the first customer network has domain number of 0 and one prefix of 0.0.0.*, the second customer network has domain number of 1 and one prefix of 0.0.1.*. The first customer network is conneted to provider edge router (PE router) B0 and the second customer network is connected to PE router B1. The ISP is connected to the foreign Internet through two edge routers, E0 and E1, besides, it has four intermediate routers (P router), I0, I1, I2 and I3. The shortest paths from the customer networks to the foreign Internet are B0-I0-I3-E0 and B1-I0-I3-E0. However, due to congestion on E0, the ISP operator wants to divert the traffic of the second customer network (behind B1) to the path through E1, i.e., B1-I0-I1-I2-E1.

We design the protocol based on the extension of OSPF [2], which can disseminate the information within the network. To illustrate the protocol, we first clarify the following aspects.

- Through e-BGP, edge routers know the prefixes of foreign Internet, e.g., both E0 and E1 know that there are three foreign Internet prefixes, 1.0.0.*, 1.0.1.*, 1.0.2.*;

- Through OSPF, PE routers know the prefixes of the customer networks behind them, e.g., B0 knows that prefix 0.0.0.* belong to the first customer network in Figure 5. Besides, PE routers know the customer domain number of the customer networks behind them, e.g, B0 knows that the customer domain number of the first customer network is 0. Through manual configuration or automatic selection (e.g., selecting the router that has lower utilization), edge routers know the preferences of customer networks on edge routers, e.g., B1 knows that the second customer network in Figure 5 prefers to pass by E1.

With these preconditions, each edge router can announce the foreign Internet prefixes combined with its own router identification to the network, each PE router can announce the customer prefixes combined with the corresponding customer domain number, PE routers are also responsible for announcing the preference of customer networks on edge routers. When receiving all necessary information, both PE and P routers will construct the routing table, which can be used to generate the forwarding table.

5.2. Router Actions

We first define three types of messages.

Announce(Prefixes, Router_ID): Edge routers send this message, to announce the binding relations between foreign IP perfixes and the edge router identification (can be represented by the IP address of the edge router). This message indicates that traffic can reach the foreign Internet through the edge router.

Bind(Prefixes, Domain_Number): PE routers send this message, to announce the binding relations between customer network IP prefixes and customer domain number. This message indicates that the customer network IP prefixes belong to the cusomter network that owns the Domain_Number.

Pref(Domain_Number, Router_ID): PE routers send this message, to announces the preference of a customer network on an edge router. This message indicates that the customer network that owns the Domain_Number prefers to pass by the edge router that owns the Router_ID.

Then the actions on different types of routers are as follows.

Edge Routers: Edge routers have to send Announce(Prefixes, Router_ID) to announce the foreign Internet prefixes to the network. For example, in Figure 5, E0 will send Announce(1.0.0.*, E0), Announce(1.0.1.*, EO) and Announce(1.0.2.*, EO). E1 will

send Announce(1.0.0.*, E1), Announce(1.0.1.*, E1) and Announce(1.0.2.*, E1).

PE Routers:

1. PE routers have to send Bind(Prefixes, Domain_Number) to announce the customer network prefixes to the network. For example, B0 will send Bind(0.0.0.*, 0), B1 will send Bind(0.0.1.*, 1).

2. PE routers have to send Pref(Domain_Number, Router_ID) to announce the preference of the cusomter network on an edge routers. For example, B1 will send Pref(1, E1).

3. After receiving Announce(Prefixes, Router_ID) from edge routers, PE routers should construct the routing table.

Intermediate Routers: After receiving Announce(Prefixes, Router_ID) from edge routers, Bind(Prefixes, Domain_Number) and Pref(Domain_Number, Router_ID) from PE routers, P routers should construct the routing table.

5.3. TwoD-IP Routing Table Construction

Receiving the necessary information (including customer network prefixes, foreign Internet prefixes and preferences of customer networks), both PE and P routers should construct the routing table. Edge routers do not need to construct the routing table, unless they also belong to PE/P routers.

The routing table consists of two parts, the first part (traditional routing table) is constructed based on OSPF, the second part (TwoD-IP routing table) is construted based on our TwoD-IP policy routing protocol. When forwarding a packet to the destination, routers first lookup the TwoD-IP routing table, if there does not exist a matched entry, routers will lookup the traditional routing table. We focus on the construction of TwoD-IP routing table in this document. For simplicity, we assume that there are only threee fields in each entry of TwoD-IP routing table, i.e., (Destination, Source, Next hop). Both the destination and source fields represent an IP prefix, the next hop field denotes the outgoing router interface to use (see Section 11 of [1] for more details).

The routing table construction process is as follows.

1. For each received Pref(Domain_Number, Router_ID), lookup the traditional table, and obtain the next hop towards the edge router that owns Router_ID. We use Next_Hop to denote the

obtained next hop.

2. For each foreign Internet prefix (Foreign_Prefix), lookup the traditional table, and obtain the next hop towards the Foreign_Prefix. We use Next_Hop' to denote the obtained next hop.

3. If Next_Hop!=Next_Hop', for each customer network prefix (Customer_Prefix) that belongs to the customer network that own Domain_Number, we add a new entry (Foreign_Prefix, Customer_Prefix, Next_Hop) to the TwoD-IP routing table.

For example, we continue the example in Figure 5, the TwoD-IP routing table on the P router I0 is shown in Figure 6.

Destination	Source	Next hop
1.0.0.*	0.0.1.*	I1
1.0.1.*	0.0.1.*	I1
1.0.2.*	0.0.1.*	I1

Figure 6: TwoD-IP routing table on the P router I0

6. Deployment

TwoD-IP should support incremental deployment, and during deployment, the following requirements should be satisfied.

Backward compatibility: During deployment, reachability should be guaranteed, and loops should be avoided.

Incentive: After deploying partial routers, ISPs should be able to see visible gains, e.g., their policies are implemented, traffic distribution is improved or security level is enhanced.

Effectivity: The deployment should maximize the benefits for ISPs, e.g., the deployment sequence should be carefully scheduled, such that ISPs can obtain maximum benefits in each step.

7. Implementation Status

We have developed a prototype of the TwoD-IP policy routing protocol (see Section 5) on a commercial router, and set up small scale tests under VegaNet [7], a high performance virtualized testbed.

Currently, we are developing the prototype of TwoD-IP router, that uses the FIST forwarding table strucute (see Section 4.2).

8. Security Considerations

TwoD-IP routing will enhance the security level of the networks, because routers will check source addresses, which is an important identity of the senders. Distributed attack defenses will be an important topic of TwoD-IP routing, because source checking functionality is deployed deeper in the network.

However, TwoD-IP routing protocols must be carefully designed, to avoid to be used by hackers.

9. IANA Considerations

Some newly designed TwoD-IP routing protocols may need new protocol numbers assigned by IANA.

10. References

10.1. Normative References

[1] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

[2] Zinin, A., Roy, A., Nguyen, L., Friedman, B., and D. Yeung, "OSPF Link-Local Signaling", RFC 5613, August 2009.

[3] Estrin, D., Li, T., Rekhter, Y., Varadhan, K., and D. Zappala, "Source Demand Routing: Packet Format and Forwarding Specification (Version 1)", RFC 1940, May 1996.

[4] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.

[5] Yasukawa, S., Farrel, A., and O. Komolafe, "An Analysis of Scaling Issues in MPLS-TE Core Networks", RFC 5439, February 2009.

10.2. Informative References

[6] Breitbart, Y., Chan, Chee-Yong., Garofalakis, M., Rastogi, R., and A. Silberschatz, "Efficiently monitoring bandwidth and latency in IP networks", INFOCOM 2001, Apr 2001.

[7] Chen, Wenlong., Xu, Mingwei., Yang, Yang., Li, Qi., and Dongchao. Ma, "Virtual Network with High Performance: VegaNet", Chinese Journal of Computers vol. 33, no. 1, 2010.

Authors' Addresses

Mingwei Xu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R. China

Phone: +86-10-6278-5822
Email: xmw@csnet1.cs.tsinghua.edu.cn

Jianping Wu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R. China

Phone: +86-10-6278-5983
Email: jianping@cernet.edu.cn

Shu Yang
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R. China

Phone: +86-10-6278-5822
Email: yangshu@csnet1.cs.tsinghua.edu.cn

Dan Wang
Hong Kong Polytechnic University
Department of Computing, Hong Kong Polytechnic University
Hong Kong
P.R. China

Phone: +852-2766-7267
Email: csdwang@comp.polyu.edu.hk