INTERNET ROUTING IN A MULTI PROVIDER, MULTI PATH OPEN ENVIRONMENT

Tony Bates, Daniel Karrenberg, Peter Lothberg, Bernhard Stockman and Marten Terpstra

(ripe-82)

Table of Contents

1 Overview of the Architecture 2
2 Background 2
2.1 History 2
3 Global Internet eXchange (GIX) 3
3.1 Establishing consistent routing on the GIX 3
3.2 Distributing the Consistent Routing Information 4
4 Route Server Implementation (RS) 4
4.1 The Basic Idea 4
4.2 How to implement filtering using the RIPE database ? 6
4.3 How to implement multiple paths ? 6
4.4 Management of the route server 6
5 Routing Registry (RR) 6
6 Basic Route Server (RS) 7
7 Scaling Issues 7
8 The Future 9
9 References 9
10 Author's Addresses 10

1. OVERVIEW OF THE ARCHITECTURE

This document proposes an implementation for a general purpose interconnect system to serve a broad range of use providing flexibility and at the same time improving and maintaining the routing stability of the global Internet.

The target is to establish a prototype and start piloting the various components by early 1993. This system is based upon three major components;

(1) **Physical GIX**

The physical GIX is a 'layer-2' type infrastructure. All participating organisations wishing to use this infrastructure bring their router(s) to this media and have the possibility to peer with any other router on a peer-to-peer basis. Any organisation has the right to refuse to peer with any other organisation. This is is a proven concept from the Federal Internet exchange (FIX) architecture and provides needed flexibility and autonomy. The only thing required of media type used is the ability to peer with all routers directly.

(2) **Routing Registry**

Routing stability is implemented by having a Routing Registry (RR) register routes based upon the request from the network owner using similar procedures in place with the current NICs for network number registration. In the GIX context, the RR registers paths preferred from the GIX to the owner of the network. The suggested number of RRs to start with is three: one for routes to North America, one for routes to the Asia-Pacific and one for routes to Europe. This provides a stable model for coordination essentially in line with "Guidelines for Management of IP Address Space [1]" registry procedures. For other domains this could be done either by having a mutual agreement with one of the above RRs or of course coordinating its own RR.

(3) **Route Server**

The Route Server (RS) is pseudo-router running on a host directly connected to the physical GIX. Each RR will operate at least one RS. The RS implements external routing protocols and maintains a routing table for destinations served by the RR operating the RS. The RS does not forward any packets; it then exports its routing table for use by the real routers on the GIX. The RR will receive routing information from all routers within the domain served by the RR. It filters the inbound routing announcements according to the details of the RR routing database and outputs the combined routing information to any router on the GIX wishing to receive it. A basic RS could be any router in use in todays Internet providing it supports BGP and specifically makes use of third-party BGP routes correctly.

2. BACKGROUND

2.1. History

The Internet as known today has evolved from a single top level Administrative domain (AD) using non hierarchical routing protocols such as EGP for inter-AD routing to an extensive collection of several transit ADs with different policies.

The routing protocols have been subsequently improved and the Border Gateway Protocol (BGP) can act as a basic toolbox for this type of environment. However, todays routing technology is based exclusively on destination address which very much limits flexibility. Another limiting factor is the rapid growth of the Internet in terms of the number of unique paths rather than networks. This will only increase as wide-scale deployment of BGP takes place.

Over the years, NSFnet/Merit has done a tremendous job of providing the much needed routing stability within the Internet. However, this is only so for networks that adhere to the terms of the NSFnet AUP. For the rest of the networks, solutions have been based very much on ad-hoc arrangements. Consistent management of these ad-hoc arrangements becomes increasingly more difficult.

We can simplify this problem by looking at the two key needs. Namely, maximal connectivity and maximal flexibility. By this we mean looking at a solution that will give us the greatest connectivity in the most flexible way possible. Currently, there are two proposals examining such issues:

— The "Network Access Point (NAP)", [2].

— The "Global Internet Exchange (GIX)", [3].

Maximal connectivity is where anyone can talk to anyone anywhere at any given time and this could be described as ‘general infrastructure’. Maximal flexibility is where every organisation that participates can easily set their own rules and easily implement them. This could be described as ‘mission oriented’ where the purpose of the network is to perform a specific task.

The GIX proposal has achieved support from various organisations including the IEPG and the CCIRN. This paper outlines the basic characteristics and requirements for a neutral interconnect that could serve a majority of the network operators today. It is recommended that one is familiar with the GIX proposal.

This paper is targeted very much towards maximal connectivity issues and routing stability. It installs the basic ‘hooks’ needed to implement peer-to-peer arrangements to support the specific needs required for maximal flexibility.

3. GLOBAL INTERNET EXCHANGE

3.1. Establishing consistent routing on the GIX

To gain maximal connectivity we need to provide routing stability for every AD that has a need to communicate outside its own AD. To do this an accountable neutral routing registry (RR) is needed. The owner of a network registers its preferred path for routing from the GIX with the RR. This would typically happen through the ADs providing service to the network.

The RR does not determine policy in any way. It just acts as a repository for routing information and performs housekeeping and consistency checking on the registered information together with the other RRs. The result is a consistent view of the routing policies towards the ADs served by the RR and to a certain extent between those ADs as well.

The proposal is initially to establish three RRs, one per geographical region, modelled after the traditional IEPG model, Asia-Pacific, North America and Europe.

The motivation for this split is the need to get something implemented and gain experience quickly in order to keep the Internet running in the light of increasing size and topological complexity. We believe that the consensus building and coordination necessary to implement a single global routing database (if possible at all) will take too long and prevent flexible changes caused by implementation experiences. Although this is certainly no research project, it is breaking new ground and quick adjustments may be needed as experience is gained.

Obviously the routing registries will have to coordinate extensively and the respective routing databases will have to be checked for mutual consistency. We believe this is easier to do than to achieve global consensus and maintain it across engineering changes. We certainly hope that once the designs are shaken down and operational experience has been gained the RR databases and RS implementations will converge; maybe even to the point when they are unified.

3.2. Distributing the Consistent Routing Information

Attached to the GIX are route servers, one per RR. The task of the route servers is to disseminate consistent routing information for all ADs of their region to all ADs connected to the GIX. A RS first peers with all routers that serve the region of the RR to acquire the dynamic routing information for its region.

The RS then uses a 'preferred path' database, derived from the routing registry of the RR to filter the dynamic routing information into a consistent routing table for its region. This routing table is then made available via BGP to all routers on the GIX that wish to use it. In order to reduce the number of peering sessions on the GIX one might consider that route-servers also peer with route-servers of other RRs and import their consistent routing information in order to redistribute it unfiltered.

A backup scheme for the route-servers is also needed to provide essential resilience in case of route server failure.

4. ROUTE SERVER IMPLEMENTATION

This implementation proposal is modelled around the European needs and other RRs might (and almost certainly will) require different implementations for their topologies and environments.

4.1. The Basic Idea

Let's imagine the part of the GIX that has relevance to Europe. Figure 1 shows a possible topology.

```
============================================================== GIX-Media
   |          |             |          | | |         |
 EASI       EBONE         PIPEX        OTHER      EUROPEAN ROUTE
 ROUTER     ROUTER        ROUTER       ROUTERS    SERVER (RS)
  |           |             |             |
  |           |             |             |
  |           |             |             |
  |           |             |             |
 Europe     Europe        Europe        Europe
```

Figure 1: Possible European Topology

The router for each of the connected network operators (PIPEX, EBONE, etc) or ADs would provide routing information to the RS. This RS would filter incoming announcements based on some database. The RIPE database comes to mind as one such possible database. The RS allows peering with any connected router on the GIX and supplies them with the best paths it has based on filtering, backup schemes and other appropriate decision criteria. Note that the routers feeding information to the RS will probably also accept information about their own region from the same RS whereas the routers outside the RS's region will just consume information from the RS.

It should be noted that when using BGP the flow of routing information is separate from the general packet flow. Figure 2 outlines the flow of routing versus the flow of traffic. This means that although routing information flows AD1-router <--> RS <--> AD2-router the non-routing packets flows directly AD1-router <--> AD2-router. This way traffic flows across the GIX in only one hop.

```
        WAN link(s)            WAN link(s)            WAN link(s)
           !                      !                      !
     +------------+         +------------+         +------------+
     !            !         !            !         !            !
     ! AD1        !         ! AD2        !         ! RS         !
     !            !         !            !         !            !
     !            !         !            !         !            !
     !            !         !            !         !            !
     !            !         !            !         !            !
     +------------+         +------------+         +------------+
        #  !   *               *  !   #               #  !   #
        #  !   *****************  !   #################  !   #
        #  !                      !                      !   #
        ######################################################
           !                      !                      !
           !                      !                      !
           !                      !                      !
           !                      !                      !
    ==============================================================GIX-LAN

Key:
***  —— Traffic flow.
###  —— BGP peering sessions (Routing flow).
```

Figure 2: Traffic flow versus Routing flow

4.2. How to implement the filtering using the RIPE database ?

This can be done by implementing a method for clearly expressing the routing policy for a given IP network. This method should be easily understandable by todays network operators. One possible method is to describe routing policy in terms of Autonomous Systems (AS). An IP network can only belong to one AS and hence by a process of derivation of routing exchanges between neighboring ASes it should be possible to create routing filter lists for any given European IP network based on what path it takes to and from the GIX.

There are of course many ways to implement route filtering to gain the needed consistency for the route server. However, the major point to note is that the current RIPE database has many of the essential objects needed to implement such a route filter mechanism today. The actual details of the implementation are beyond the scope of this paper.

4.3. How to implement multiple paths ?

In the case of access to Europe where one network operator provides backup for another, or a network operator is multihomed on the GIX. Routing information for that network must be accepted from more than one router. In this case additional policy information is needed. This could be based on some form of static ordering of ADs or various dynamic criteria based on the information provided by the routing protocol. Equally if you there was more than one connection to the same service provider at the GIX some sort of additional policy information would be needed. In the light of current implementations there may be some scaling problems with this in terms of manageability but initially this is not perceived as a problem as it left for further study.

4.4. Management of the route server

It should be clear that the management of the RS should be carried out by a neutral organisation. The RIPE NCC is an organisation that is in a position to provide this function. The NCC could decide to delegate the operational management to another organisation providing a neutral service.

The RIPE NCC would act as the RR for European based networks. It would then generate RS configuration files derived from the RIPE database.

5. ROUTING REGISTRY

The task of the RR is to register preferred routes. This has to be performed on a neutral basis with respect to the different IP network providers. The official administrative contact for an IP network registered with the NIC is the single source of authority that has the privilege to register routing attributes with the RR. For smooth operation, this may well be done by the network operator on behalf of its ‘customer’ network. If there is any conflict then the administrative contact will be contacted.

The RR is responsible for a domain or area which in this proposal is based on a continent like Europe. When dealing with multi-national organisations the guideline is that the networks are regarded to be under the responsibility of the RR in which area the network is connected to the global Internet. If there are network operators whose connections span more than one continent then their internal topology should be taken into account.

In Europe, the RIPE database already contains most of the database objects needed to store the information to implement an RR.

6. BASIC ROUTE SERVER

A basic RS can be viewed as a virtual router making policy based routing decisions which it then communicates to the real routers. The real routers base their forwarding decisions on that routing information. The RS imports routes from all routers that connect to the domain that the corresponding RR has responsibility for and exports routes to anyone that starts a BGP peering session with the RS. For redundancy and resilience it is assumed that eventually there will be at least two route servers per domain, peering with all AD routers and using IBGP between themselves to maintain consistency with each other.

Figure 3 attempts to show the basic set up of the European route servers on the GIX. European routers EU-AD[1..5] are routers belonging to the different European ADs. Each European AD router peers using BGP with the two European route server, EU-RS1 and EU-RS2. The route servers in turn peer with each other using IBGP.

```
      !              !              !              !              !
   +------+       +------+       +------+       +------+       +------+
   ! EU   !       ! EU   !       ! EU   !       ! EU   !       ! EU   !
   ! AD 1 !       ! AD 2 !       ! AD 3 !       ! AD 4 !       ! AD 4 !
   +------+       +------+       +------+       +------+       +------+
      !              !              !              !              !
      !              !              !              !              !
*************************************************************************
                             !             !
                             !             !
                          +------+      +------+
                          ! EU   !-IBGP-! EU   !
                          ! RS 1 !      ! RS 2 !
                          +------+      +------+
```

Figure 3: Basic European Route Server Set-up

Once this is seen to operate reliably and general procedures are in place then the next step is to add another RR domain. Taking North America as an example, figure 4 shows how we deal with two route-servers serving two different domains.

A European router that wants to determine how to deliver packets to a North American destination, peers with the two North American route servers NA-RS1 and NA-RS2 which in turn point to the preferred North American AD router (e.g. NA-AD1). If for some reason the same network is announced by two RRs the forwarding decision is made based on policies implemented in the source router rather than a particular RR.

The RS could (and should) be enhanced to provide multiple routing tables or databases to different sets of AD routers. How to specify those features and how to guarantee overall consistency is beyond the scope of this paper and is left for further study. Any implementation of an RS should be designed to allow this kind of future enhancement.

7. SCALING ISSUES

Using this first simple approach to the RS still needs to take into account some scaling issues in that the GIX implementation must be able to scale in terms of the number of data paths and traffic present from the GIX and although this paper concentrates very much on the RS itself as opposed

```
        !              !              !              !              !
     +------+       +------+       +------+       +------+       +------+
     ! EU   !       ! EU   !       ! EU   !       ! EU   !       ! EU   !
     ! AD 1 !       ! AD 2 !       ! AD 3 !       ! AD 4 !       ! AD 4 !
     +------+       +------+       +------+       +------+       +------+
        !              !              !              !              !
        !              !              !              !              !
********************************************************************** GIX-Media
*                              !             !
*                              !             !
*                          +------+      +------+
*                          ! EU   !-IBGP-! EU   !
*                          ! RS 1 !      ! RS 2 !
*                          +------+      +------+
*
*                          +------+      +------+
*                          ! NA   !-IBGP-! NA   !
*                          ! RS 1 !      ! RS 2 !
*                          +------+      +------+
*                             !             !
*                             !             !
**********************************************************************
        !              !              !              !              !
        !              !              !              !              !
     +------+       +------+       +------+       +------+       +------+
     ! NA   !       ! NA   !       ! NA   !       ! NA   !       ! NA   !
     ! AD 1 !       ! AD 2 !       ! AD 3 !       ! AD 4 !       ! AD 5 !
     +------+       +------+       +------+       +------+       +------+
        !              !              !              !              !
```

Figure 4: Two Route Servers serving different RRs

to the transport of traffic it is worth spending some time on how to deal with this problem. The general issue is how to make efficient use of ’private’ interconnections among different routers at the GIX. This can be used to improve throughput and also give added flexibility and possible resilience between various providers as well.

The approach used is that the RS itself does not need to know anything about the topology of any private interconnections. Figure 5 shows a possible topology with some interconnections between various European AD routers.

L1, L2, L3 and L4 are logical interfaces in each AD router respectively. The ADs use their own network and assign a host address on the router for this effective subnet. All BGP peering are then sourced/received using this address and hence the RS will use this address as the next hop.

To establish the needed BGP peering, the RS now has to know how to route to these logical interface addresses. This can be done simply be having static based routes pointing to the respective GIX interface address G1, G2, etc. The AD routers also need to know how to get to the logical interfaces. This could be done by either running an IGP or using static routes.

So as an example, a network connected to and announced by the AD3 router will have its next hop announced by the RS as L3. If the source is AD1, then it can make its own path selection based on some metric to prefer the private connection over the GIX. Traffic between AD routers that does not have private links will still use the GIX for data packets in this scheme.

```
                   Long links to each AD
    !                !                !                !
    !                !                !                !
    !                !                !                !
    !****************!****************!****************!******
    ! !              !                !                !    !
    ! ! *************!****************!******          !    !
    ! !  !           !    !           !    !           !    !
    ! !  !           !    !           !    !           !    !
   +------+         +------+         +------+         +------+
   ! EU   !--L1     ! EU   !--L2     ! EU   !--L3     ! EU   !--L4
   ! AD 1 !         ! AD 2 !         ! AD 3 !         ! AD 4 !
   +------+         +------+         +------+         +------+
      ! G1             ! G2             ! G3             ! G4
      !                !                !                !
*****************************************************************
                             !             !
                             !             !
                          +------+      +------+
                          ! EU   !-IBGP-! EU   !
                          ! RS 1 !      ! RS 2 !
                          +------+      +------+
```

Figure 5: Possible topology with some European AD interconnections

8. THE FUTURE

This paper has introduced some concepts for an implementation and advancement of the ideas muted in the original IEPG GIX paper [3]. It is, as stated, very much focused on the European requirements. The outcome of this is a basic framework for a project in which a prototype implementation of the above ideas could take place.

It is hoped that a more formal and structured implementation plan can be built from the ideas outlined. However, this is outside the context and scope of this paper.

9. REFERENCES

1 — "Guidelines for Management of IP Address Space", E. Gerich, October 1992.

2 — "NSF Solicitation Concept", B. Aiken, H-W. Braun, P. Ford, S. Wolff, July 1992.

3 — "Global Internet Exchange (GIX)", G. Almes, P. Ford, P. Lothberg, June 1992.

10. AUTHOR'S ADDRESSES

Tony Bates†
RIPE Network Coordination Centre
Kruislaan 409
NL-1098 SJ Amsterdam
+31 20 592 5065
T.Bates@ripe.net

Daniel Karrenberg
RIPE Network Coordination Centre
Kruislaan 409
NL-1098 SJ Amsterdam
+31 20 592 5065
D.Karrenberg@ripe.net

Peter Lothberg
STUPI
Box 9129
102 72 Stockholm
+46 8 6699720
roll@stupi.se

Bernhard Stockman
SUNET/NORDUnet
Royal Institute of Technology
Drottning Kristinas Vag 37B
S-100 44 Stockholm
+46 8 7906519
boss@sunet.se

Marten Terpstra
RIPE Network Coordination Centre
Kruislaan 409
NL-1098 SJ Amsterdam
+31 20 592 5065
M.Terpstra@ripe.net

† Part of this work was done whilst Tony Bates was at the University of London Computer Centre, 20 Guilford Street, London, WC1N 1DZ, United Kingdom.