Network Working Group
Internet-Draft
Intended status: Informational
Expires: January 3, 2008

D. Jen
M. Meisel
D. Massey
L. Wang
B. Zhang
L. Zhang

July 2, 2007

APT: A Practical Transit Mapping Service
draft-jen-apt-00.txt

Status of this Memo

Copyright Notice

Abstract

The size of the global routing table is a rapidly growing problem. Several solutions have been proposed. These solutions commonly divide the Internet into two parts, one for customers and one for providers, where only provider addresses are globally routable. Packets destined for customer addresses are tunneled through provider

space. For this process to work, there must be a mapping service that can supply an appropriate provider-edge address for any given customer address. We present a design for such a mapping service. We adhere to a "do no harm" design philosophy: maintain all desirable features of the current architecture without negatively affecting its security or reliability. Our design aims to minimize delay and prevent loss in packet encapsulation, minimize the number of new or modified devices, and keep the level of control traffic manageable.

Table of Contents

1. Requirements Notation

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2. Introduction

The unexpected, explosive growth of the Internet is causing a greater and greater strain on its infrastructure. This problem has been well-documented in [RAWS][AddrAlloc]. Several solutions have been proposed to address this problem [EFIT][CRIO][LISP], most of which involve separating the Internet into two parts -- one for user networks and one for transit providers. Routers in transit space would only need to know how to route to transit prefixes, which are stable and conducive to topological aggregation. When a packet is sent from user address A to destination user address B, A's provider-edge router (the ingress tunnel router, or "ITR", as defined in [LISP]) encapsulates the packet and sends it through transit space to B's provider-edge router (the egress tunnel router, or "ETR"). B's ETR decapsulates the packet and forwards it to the appropriate recipient, B.

When encapsulating a packet, A's ITR must somehow determine B's ETR's transit-space address and include it in the outer header. In general, any ITR must be able to map any given user-space address to a corresponding ETR transit-space address for proper tunneling through transit space. This illustrates the need for a mapping service that can provide this address. The design details of this mapping service will play a large part in determining the effectiveness of any proposed implementation of a user/transit provider address space separation. The mapping service also presents an exciting opportunity to enhance the services currently offered by the Internet, which is further reason to carefully consider how this service should be implemented. Should mapping information be distributed via a push or a pull model? What additional information, if any, should be obtained along with the mapping information? Can we satisfy the mapping requirement without sacrificing any services or packet delivery quality?

Our answers to these questions are rooted in a "do no harm" design philosophy: improve routing scalability without sacrificing any desirable features in the current architecture or negatively affecting its security and reliability. To this end, we present APT, A Practical Transit mapping service designed with the following goals in mind.

- Minimize delay and prevent loss in packet encapsulation.
- Minimize the number of devices that need to be modified to support our new design.
- Minimize the number of devices that will require additional resources or complexity.
- Keep the design modular so that the method used to propagate mapping information is independent from the method used to retrieve mapping information for tunneling.

APT is designed for use with eFIT [efitID][EFIT], one of the major proposals for user/transit provider address space separation. However, APT should be generally applicable to other proposals of the same class.

3. Terminology

User Network (UN) - A network that pays another organization to deliver its packets through the Internet. Each user network is a customer of some Transit Network (see definition below). "User network" holds the same meaning as it does in the eFIT proposal.

Transit Network (TN) - An AS whose business is to provide packet delivery services for its customers. Transit Networks serve as providers for user networks. As a rule of thumb, if the AS appears in the middle of any ASPATH in a BGP route today, it is considered a transit network.

Transit Space (TS) - The address space used by transit networks. Nodes within a transit network are assigned TS addresses. Sometimes the term "transit space" will refer to the non-edge area of the Internet where TS prefixes are routable.

User Space (US) - The address space used by user networks. Nodes within a user AS are assigned user-space addresses. Sometimes the term "user space" will refer to the edges of the Internet whose prefixes are not routable in transit space (though packets to those addresses are deliverable through transit space). We assume that TS and US addresses can be clearly distinguished.

Border Link - A link that crosses the boundary between transit space and user space.

Default Mapper - A new device required by our mapping service. Each transit network MUST have at least one default mapper. A default

mapper carries a complete mapping table. In other words, given any user-space address, default mappers can return the TS address of a provider-edge router corresponding to that address. To support the growing trend towards multihoming, default mapping entries will map a user-space prefix to a non-empty SET of TS destinations, all of which have a direct connection to the destination network in user space.

Tunnel Router (TR) - These devices will replace all current provider-edge routers, located at the provider end of border links. Like ITRs and ETRs in LISP [LISP], TRs provide the encapsulation and decapsulation services required for tunneling user packets through transit space. A TR has both ITR and ETR functionality, meaning that any TR can perform both encapsulation and decapsulation of packets. To properly encapsulate any given user-space packet, TRs can query the default mappers for mapping information. TRs also cache commonly used mapping entries locally. Note that TR cache entries are NOT identical to the mappings stored at default mappers (see the definitions of "mapping" and "mapping entry" below). TRs are designed to be as simple and as fast as possible, adding only what is necessary for proper tunneling functionality.

APT Node - A general term referring to any new device type introduced by APT. This includes both default mappers and TRs.

Router - These are ISP-owned non-border routers that exist today. Other than minor configuration changes, these routers need no alteration or replacement, and can be used just as they are used currently.

Mapping - A mapping contains a user-space prefix and a non-empty SET of ETR TS addresses associated with the prefix. Mappings also include related information such as the user's public key and priority rankings for each of the ETRs in the set. Default mappers store mappings.

Mapping Entry - A mapping entry contains a user-space prefix and any SINGLE ETR TS address associated with the prefix. Any mapping entry is a subset of the complete mapping for its user-space prefix. TRs store mapping entries along with an associated TTL. A mapping entry is removed once its TTL expires.

4. The Mapping Service

To minimize the latency introduced by encapsulation, APT seeks to store mapping information as close to the ITR as possible. However, the global mapping table is likely to grow very large over time. To avoid undue memory requirements for ITRs while still keeping mapping

information within reach, we introduce the concept of default mappers.

A TR does not need to store the entire global mapping table. Instead, it queries a default mapper for mapping information and caches recently used mapping entries.

Default mappers are the only devices in the network that need to store the complete global mapping table. As we will see in the following example, TRs only use default mappers in the event of a cache miss. This means that, given large enough caches at the TRs, network latency will not heavily depend upon default-mapper performance. Additionally, we propose the use of anycast to reach default mappers within an AS. Each TN AS need only have a single default mapper, but the use of anycast makes it easy for a TN to deploy more. The result is a robust, scalable default mapping system.

4.1. A Mapping Example

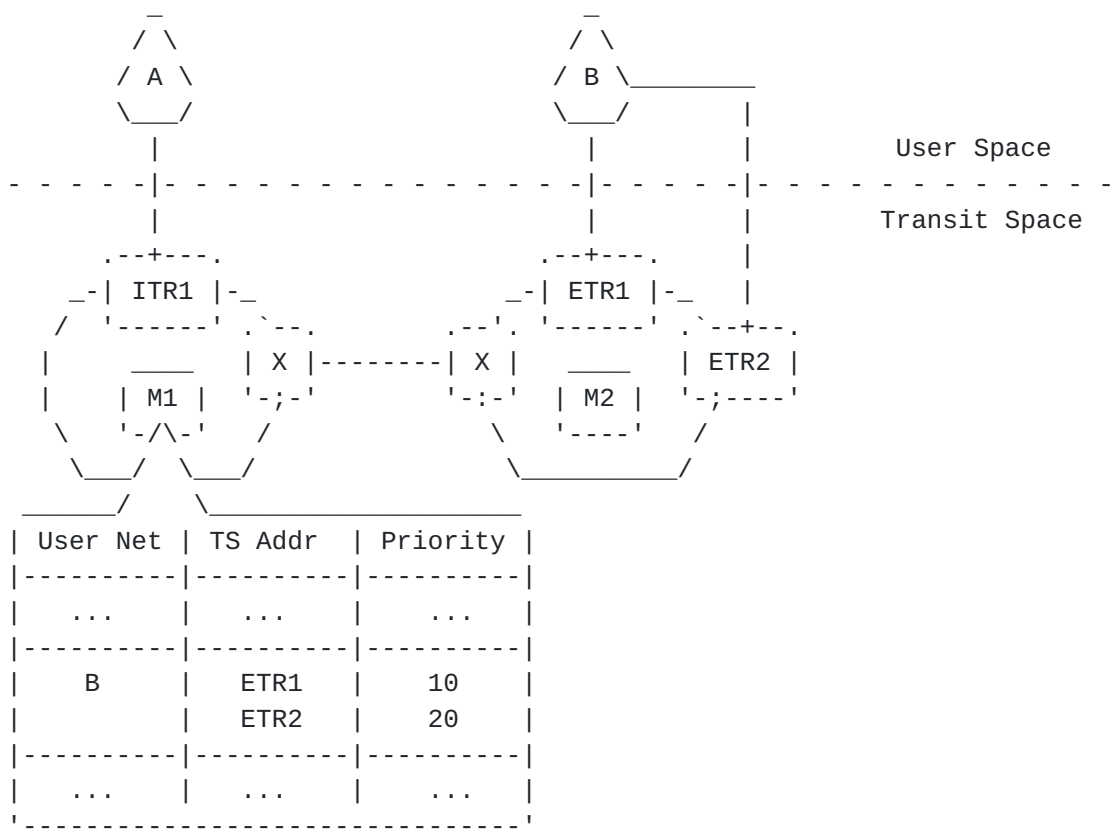

Figure 1. This is a simple topology for demonstrative purposes. A and B are user networks addressable via user-space prefixes, ITR1,

ETR1, and ETR2 are TRs, any node labeled "X" is a router, and M1 and M2 are default mappers. A portion of the mapping table for M1 is shown.

In this section, we illustrate how TRs and default mappers interact within an AS to properly tunnel user-space packets through transit space.

In Figure 1, assume a node in network A sends a packet to a user-space address in network B. When this packet arrives at ITR1, ITR1 looks up the destination user-space address in its mapping cache. If a matching prefix is present in its cache, ITR1 simply encapsulates the packet with the corresponding TS destination address and send it across transit space. If a matching prefix is not present, ITR1 will send the packet through its default mapper. It does this by encapsulating the packet with the anycast address for default mappers in its AS as the destination.

This packet will arrive at M1, the only default mapper in ITR1's AS. When M1 receives the packet, it decapsulates it and examines the user-space destination address. Since default mappers store the full, global mapping table, a default mapper will always be able to encapsulate the packet with a valid TS destination address. All packets encapsulated by a default mapper MUST contain the default mapper's TS address as the source address.

In addition to forwarding the packet to an appropriate ETR (ETR1, in this case), M1 also treats the incoming packet as an implicit request from ITR1 for mapping information. M1 responds to ITR1 with an ICMP packet containing a mapping entry that maps B to ETR1. This allows ITR1 to add this mapping entry to its cache so that ITR1 can tunnel further packets destined for B directly to ETR1. The mapping entry also contains a time to live (TTL) that is set by M1. The TTL ensures that ITR1 will occasionally re-request this mapping information from M1. At this time, if the mapping information changed in any way since ITR1's prior request, M1 can respond with an updated mapping entry. Without this TTL, ITR1's cached information may become inaccurate over time.

5. Multihoming Support

In the example above, the observant reader may have noted that B is multihomed. That is, B can be reached through both ETR1 and ETR2. Multihoming provides B with both enhanced reliability in case of a connectivity failure and the flexibility to split incoming traffic across different ETRs.

In accordance with our design goals, all of the logic for selecting a destination for a multihomed user is contained within default mappers. Default mappers will store mappings containing all of the ETRs for a given user-space prefix, and ITRs will only store a single mapping entry per user-space prefix. When an ITR requests a mapping entry for a multihomed user, it is up to the default mapper to decide which one to return.

Many users will want to have some control over which ETR is used for incoming traffic. To allow this, we let users assign a priority value to each of the mapping entry for their prefixes, making it available to all default mappers throughout the transit space (see Section Section 6). The number is to be treated like a ranking -- an ETR with a lower priority value is more preferable.

At the same time, a transit network may also has its own preference regarding which of the ETRs to use for a given user-space prefix. Default mappers can use a combination of locally configured routing policies and the user priority information to choose from a set of valid ETR addresses. Going back to Figure 1, assume that ITR1 does not have a mapping entry for B in its cache. When A sends B a packet, ITR1 will send the packet to M1. If M1 has no preference between ETR1 and ETR2, it will examine the priority values in B's mapping and select ETR1, B's most preferred ETR. M1 forwards the packet to ETR1 and returns the appropriate mapping entry to ITR1, which stores the mapping entry in its cache.

In the case of a priority value tie, the default mapper can break the tie by picking the ETR to which it has the shortest path. If some ETRs are tied in terms of both lowest priority value and shortest path, the default mapper is free to break the tie arbitrarily. The address of the selected ETR will be used as the destination address when encapsulating the packet.

We envision that users will be able to manipulate their incoming traffic load by setting appropriate priority values in their mapping. A user who wants load balancing can assign the same priority value to all of his mapping entries. A user who wants to have one TN as a primary provider and another only as a backup can simply assign a higher priority value to his ETR at his backup provider.

5.1. Using Alternate ETRs During Failures

When a network failure has caused an ETR to become unreachable, an affected multihomed user will expect his traffic to be temporarily routed through alternate ETRs. There are three general types of failures that would require an ITR to use an alternate ETR: (1) an ITR may discover via BGP that it can no longer reach the TS prefix

containing the address of the intended ETR, (2) an ITR may learn via ICMP Destination Unreachable packets that its intended ETR is unreachable, and (3) the link between a user network and its TR may be down, a new problem introduced by the tunneling architecture. We will explain how to handle each of these failure types below, using Figure 1 as a reference. We assume that, at the time of failure, all TN ASes are using ETR1 to reach B.

To assist in handling these failures, we include a time till retry (TTR) for each mapping entry in every mapping stored in default mappers. Normally, the TTR for each mapping entry is set to zero, indicating that it is usable. Any mapping entry with a non-zero TTR value is considered invalid. We will refer to the action of setting a mapping entry's TTR as "invalidating the entry." Mapping entries that map to unroutable destinations are also considered invalid. So long as a mapping entry is invalid, default mappers will not use this entry as a destination address or include it in mapping responses. The role of the TTR for handling failures will become clear in the explanations below.

5.1.1. Handling TS Prefix Failure

For failures of type (1), ITR1 has no route to ETR1. Assume a host in network A attempts to send a packet to a host in network B. If ITR1 does not have B's mapping in its cache, it will forward the packet to M1 (see Section Section 4.1). If ITR1 does have B's mapping in its cache, it will see that it has no path to ETR1, and send the packet to M1 instead. M1 will also see that it has no route to ETR1, and thus select the next-most-preferred ETR for B, ETR2. If it has a route to ETR2, it sends the packet with ETR2 as the TS destination address and replies to ITR1 with the corresponding mapping entry. M1 can assign a relatively short TTL to the mapping entry in its response. Once this TTL expires, ITR1 will forward the next packet for B to the default mapper, which will respond with the most-preferred mapping entry that is routable at that time. This allows ITRs to quickly go back to using ETR1 once it becomes routable again.

5.1.2. Handling Single TS Address Failure

In the second case, the TS prefix containing ETR1 is still routable from ITR1, but ETR1 is unreachable from ITR1. Thus, ITR1 will receive an ICMP Destination Unreachable message in response to any packet sent to ETR1. ITR1 will need to turn to its default mapper for an alternate TS destination address for B. M1 will send an alternate valid mapping entry (if available) to ITR1. For this to work, TRs MUST forward all received ICMP Destination Unreachable messages to their default mappers. Default mappers MUST then

invalidate ALL mapping entries that map to the unreachable TS destination address. To allow this, default mappers will have a reverse-mapping table to go along with their mapping table. These reverse-mapping tables map TS addresses to their corresponding user-space prefixes. Now default mappers can look up the unreachable TS address in their reverse-mapping tables, and temporarily invalidate all entries that map to that TS address.

5.1.3. Handling User-to-TR Link Failure

The final case involves a failure of the link connecting ETR1 to B. In the previous two cases, current Internet standards were in place to allow ITR1 to know that a failure occurred. This case, on the other hand, is a new type of failure that does not exist in today's infrastructure. Therefore, it will require a new type of failure message. These messages will take the form of a new ICMP message type, which will include the user-space prefix that was not reachable. TRs MUST be configured to forward all border link failure ICMP messages to their default mappers, in the same fashion that TRs forward all destination unreachable ICMP messages to their default mappers.

Going back to our example, when ETR1 discovers it cannot forward the packet to B due to a border link failure, it will send ITR1 an ICMP packet of our new type stating that B's prefix is currently unreachable. ITR1 will forward the border link failure ICMP message to its default mapper, which will invalidate that mapping entry. If the mapping entry is already invalid, it will reset the entry's TTR. If the prefix has an alternate valid mapping entry, M1 will send this mapping entry to ITR1.

Furthermore, to minimize packet losses, ETR1 should not simply drop the packet addressed to the unreachable user network. Instead, ETR1 should send this packet to M2 in hopes of finding an alternate ETR that can reach the user network. However, M2 will then look up a TS destination address for B and choose ETR1 again. This is undesirable, since we are seeking an alternative destination. Therefore, when encapsulating packets for forwarding, default mappers MUST check if the chosen TS destination address is the same as the TS sender address in the packet's original TS header. If so, this indicates that the TS-to-user link is down at this ETR. In such cases, default mappers MUST invalidate the corresponding mapping entry and seek an alternative.

To complete our example, ETR1 sends an ICMP message to ITR1 and also sends the data packet to M2. M2 looks up a destination TS address for the packet and finds ETR1. M2 then compares this TS address with the TS address of the original sender of the packet, which is also

ETR1. Since they are the same, M2 invalidates this mapping entry and finds an alternate destination, ETR2. M2 then forwards the packet to ETR2.

5.2. Summary of Requirements for Multihoming Support

TR cache entries MUST include a TTL value, which will be provided by their default mapper.

In default mappers, every TS destination address in a mapping MUST include a time until retry (TTR). Usable mapping entries have a TTR of zero. When a mapping entry becomes unreachable due to failures, the TTR MUST be set to a pre-configured value. An alternate entry in the same mapping MUST be used in place of an invalid mapping entry if available.

Default mappers MUST be able to invalidate all mapping entries that map to a particular TS destination address that has become unreachable. This can be implemented using a reverse-mapping table.

We will use a new type of ICMP message to indicate border link failure.

TRs MUST forward all ICMP destination unreachable and border link failure messages to their default mapper.

If an ETR cannot send a packet due to a border link failure, it MUST send this packet to its default mapper. This ETR MUST use its own TS address as the source TS address of the packet.

Upon receipt of any data packet, default mappers MUST check if the chosen TS destination address is the same as the TS source address in the packet's original TS header. If so, default mappers MUST invalidate the corresponding mapping entry and look for an alternate ETR for the packet.

6. Exchanging Mappings Between ASes

In order for default mappers to store a full, global mapping table, there must be some way for them to receive mappings from other ASes. To avoid introducing latency or packet loss when encapsulating packets, the default mappers must have a full set of mappings available locally. To accomplish this, we distribute mappings using a push method. Default mappers MUST regularly announce the mappings for all of their customers to the rest of the network.

When a default mapper receives new mappings, it stores them in its

mapping table, replacing any existing mappings. When a TR receives new mappings, it simply deletes any matching cache entries. Any further communication with the formerly cached host will require the use of a default mapper. This ensures that only the default mappers need to validate mapping announcements and enforce policy.

Mapping messages will be flooded throughout the network via BGP. A new BGP attribute will be required for this purpose.

We have selected BGP initially in order to ease incremental deployment and minimize the changes required to existing routers. However, mapping announcements could easily be distributed via a different reliable broadcast protocol at a later date. Transitioning mapping distribution to a different protocol will not affect any other aspect of APT.

6.1. In Defense of BGP

Despite the use of BGP, mapping announcements will not cause the same problems that BGP routing announcements do in the Internet today for the following reasons.

First, for routing announcements, the path taken to reach each router is a crucial piece of information. For mapping announcements, the path taken to reach each APT node is not meaningful. This means that only a single copy of each mapping announcement needs to reach each APT node, providing an opportunity to prune duplicates, or even to make use of a spanning tree. This also means that path exploration and its repercussions do not exist for mapping announcements.

Second, mapping announcements only require processing at default mappers and, to a lesser extent, TRs. Other routers in the network need only pass these announcements along to their peers. Thus, the processing burden placed on other routers by excessive routing updates is completely avoided.

Finally, there will be far fewer mapping announcements than there are routing announcements. TNs rarely change the addresses of their equipment, and customers are generally under a monthly contract with their provider TNs. Therefore, permanent mapping changes are unlikely to occur more than once per month per customer. Furthermore, transient failures do not cause mapping announcements in APT. The most common cause of mapping announcements will be regular refresh announcements, which should never need to be sent more than every other day in most cases.

7. Security and Robustness

Using BGP to distribute mapping announcements guarantees that they are only accepted from manually configured BGP peers. This ensures that mapping announcements are no less secure than routing announcements today. When applied to the eFIT architecture, however, the security of this scheme is greatly increased. This is due to the fact that eFIT TS addresses are not addressable from user space [efitID][EFIT]. This turns out to be a major boon for the BGP trust model, since only other TS nodes are valid BGP peers.

The complete separation of the eFIT TS address space provides another security benefit: malicious users cannot attack equipment that they cannot address. End users simply cannot affect the TS nodes that their packets travel through within transit space.

Despite these benefits, there are some additional issues introduced by APT. Manually configured mappings provide an opportunity for human error, our reliance on ICMP packets provides an opportunity for spoofing and cache poisoning, and storing the entire global mapping table at default mappers poses a threat to long-term scalability. The remainder of this section will address each of these issues in turn.

7.1. Detecting Misconfigurations

Due to the fact that only TNs will have access to transit space, false mapping updates are far more likely to be the result of accidental misconfigurations than malicious attacks. With this in mind, we present a simple, extensible authentication scheme that can detect and, in some cases, prevent accidental misconfigurations.

The types of misconfigurations that could potentially be harmful are those that result in one provider accidentally interfering with the mapping for another provider's customer. This can happen whenever a provider accidentally announces a mapping for the wrong user-space prefix. These types of accidental conflicts fall into three categories: (1) a provider announces a mapping for a prefix owned by another provider's customer, (2) a provider announces a mapping for a shorter user-space prefix that contains a longer user-space prefix owned by owned by another provider's customer, and (3) a provider announces a mapping for a longer user-space prefix that is a subset of a shorter user-space prefix owned by another provider's customer.

The first category of conflicts is the only one that we intend to actively prevent. Clearly, the user that owns a particular user-space prefix should be the ultimate authority for his mapping information. However, user networks do not announce their mappings

to the network directly, but rather through their providers. In order to ensure a mapping update for a user-space prefix is approved by its rightful owner, we must include some sort of user authorization string in each announcement. To this end, we introduce a public-key field into each mapping. This field SHOULD contain a cryptographically valid public key, but it will only rarely need to be used as such. In the normal case, when a default mapper receives a new mapping announcement that would replace an existing one, it only needs to ensure that the public key has not changed. (This scheme is similar in spirit to the way that OpenSSH uses its 'known_hosts' file.) However, as long as all of a UN's providers store the corresponding private key, the distribution of public keys also introduces the possibility of using cryptographic signatures for any number of purposes within transit space.

For the other two categories, it is less clear that such an announcement is the result of a misconfiguration. It is possible, for example, that the owner of a /16 user-space prefix has resold some of the contained /24 prefixes to other UNs. In such cases, only the administrators will know if the announcement is valid. It is for this reason that (in the spirit of PHAS [PHAS]) we do not attempt to prevent such changes, but only detect and notify interested parties. Since legitimate mapping changes are infrequent, notifying interested parties of mapping changes via e-mail is a perfectly viable option. These notifications could also prove useful in debugging the mapping service, or a particular provider's configuration.

7.2. ICMP Mapping Packets

ICMP mapping packets are used exclusively by default mappers to send mapping entries to the TRs within their AS. Therefore, there is no reason that these ICMP packets should ever need to travel between ASes. In order to prevent cache poisoning through spoofing, these ICMP packets simply MUST be dropped at all border routers within transit space.

7.3. Other ICMP Packets

Our mapping service also depends on two other types of ICMP packets: existing ICMP Destination Unreachable messages, and our new ICMP Border Link Failure messages. Both of these packet types must traverse AS boundaries. Again note that, under the eFIT architecture, these packets are already more trustworthy than ICMP packets in the current infrastructure -- they can only be generated by hosts in transit space. However, if this level of security is deemed insufficient, the keys used for detecting misconfigurations could be used to cryptographically sign such packets, ensuring that they are coming from the appropriate sender.

7.4. Default Mapper Scalability

Theoretically, the global mapping table could grow to contain a separate mapping for every user-space prefix. In the case of IPv6 prefixes, the total number of mappings would be on the order of 10^18, far more than we can expect to be able to store on a single device. If the global mapping table were to approach such gargantuan proportions, a few simple changes to the default-mapper model would allow APT to scale gracefully.

Instead of each default mapper storing the full, global mapping table, each default mapper would store only a subset of the table. This subset would be aggregatable by user-space prefix. For addresses outside of this subset, a default mapper would store mappings that mapped short, artificially aggregated prefixes to the TS addresses of other default mappers. Like virtual prefixes in CRIO [CRIO], the user-space prefixes in these mappings would not necessarily correspond to actual user prefixes.

Each virtual prefix would be announced by the default mapper responsible for the corresponding subset of the global mapping table. In order to ensure complete coverage of the user address space, some central authority would need to assign these virtual prefixes to individual transit networks.

This scheme allows for a tradeoff between latency and default mapper storage requirements. (For more discussion of the characteristics of such a tradeoff, see [CRIO].) However, this scheme also requires some providers to become authoritative sources for mappings owned by other providers' customers. Both this requirement and the need to involve a central authority could prove problematic for deployment. Therefore, we do not recommend using this scheme unless the size of the global mapping table demands it.

8. Incremental Deployment

Clearly, the deployment of APT will coincide with the deployment of eFIT (or a similar architecture). Though incremental deployment of the eFIT architecture itself is beyond the scope of this document, we must at least show that APT will behave properly under partial deployment.

Under the eFIT architecture, addresses outside of transit space will not change. This means that user-space prefixes will initially share the existing IP address space. This fact provides us with a simple method for delivering packets to addresses for which no mapping is available. Presumably, the only such addresses will be those

connected to providers who have not yet adopted the new architecture. In order to deliver such packets, APT nodes can simply return them to the old infrastructure, and they will be routed as they are today. In order to support this feature, default mappers will respond to TRs with an ICMP mapping packet that indicates that no entry exists for the given user-space prefix. TRs will keep a negative cache entry for such prefixes so that they can forward such packets directly to a non-TS router.

In discussing incremental deployment, we must also address the issue of how new default mappers will acquire the complete mapping table when they are first connected to transit space. Since our mapping service design requires that all ASes re-announce all of their mappings at a regular interval, commissioning a new default mapper only requires connecting it to the network and waiting for all other TS ASes to re-announce their mappings. Yet, this introduces a potential problem -- if there is no upper bound on the regular refresh interval, there will be no upper bound on how long a new default mapper needs to wait until its mapping table is complete. Therefore, there needs to be an upper bound on the refresh interval for mappings. An appropriate value would be once a week. This would mean that a newly deployed default mapper would be able to reach the entire transit space one week later (with the exception of any ASes that failed to follow protocol).

9. Future Work

Optimally, any design paper should include an evaluation section. In the future, we will examine traces of Internet activity to determine the characteristics of the tradeoff between TR cache size and default mapper workload, the amount of traffic overhead that would be incurred by our push-based design, and any other results that the community deems useful.

We are also considering automating user mapping updates. Under our current design, whenever a user needs to update his mapping information (he may add, subtract, or change providers, or change his priority values), the user must contact his providers offline and request that they announce the updated mapping information. It is then up to the providers to update the mapping information. As we have seen with DNS updates, human involvement introduces the possibility of human error and delay. We hope to provide UNs with an automated way to manage their mapping information.

10. IANA Considerations

This memo includes no request to IANA.

11. Security Considerations

Security considerations for APT are discussed in Section Section 7.

12. References

12.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

12.2. Informative References

[efitID] Massey, D., Wang, L., Zhang, B., and L. Zhang, "A Proposal for Scalable Internet Routing and Addressing", Internet Draft, http://www.ietf.org/internet-drafts/draft-wang-ietf-efit-00.txt, 2 2007.

[EFIT] Massey, D., Wang, L., Zhang, B., and L. Zhang, "A Scalable Routing System Design for Future Internet", SIGCOMM IPv6 Workshop , 8 2007.

[LISP] Farinacci, D., Fuller, V., and D. Oran, "Locator/ID Separation Protocol (LISP)", Internet Draft, http://www.ietf.org/internet-drafts/draft-farinacci-lisp-00.txt, 2007.

[PHAS] Lad, M., Massey, D., Pei, D., Wu, Y., Zhang, B., and L. Zhang, "PHAS: A Prefix Hijack Alert System", USENIX Security .

[AddrAlloc] Meng, X., Xu, Z., Zhang, B., Huston, G., Lu, S., and L. Zhang, "IPv4 Address Allocation and BGP Routing Table Evolution", ACM SIGCOMM Computer Communication Review (CCR) special issue on Internet Vital Statistics, Volume 35, Issue 1, p71-80.

[RAWS] Meyer, D., Zhang, L., and K. Fall, "Report from the IAB Workshop on Routing and Addressing", Internet Draft, http://www.ietf.org/internet-drafts/draft-iab-raws-report-02.txt, 2007.

[CRIO] Zhang, X., Francis, P., Wang, J., and K. Yoshida, "Scaling IP Routing with the Core Router-Integrated Overlay", Proc. International Conference on Network Protocols , 11 2005.

Appendix A. BGP Mapping Announcement Fields

Address Type - This field specifies the type of user-space addresses used for the user-space prefixes in the announcement. All user-space prefixes in a single mapping announcement MUST be of the same address type. Currently, this is expected to be either IPv4 or IPv6, but any other address type is possible provided that it is supported by the APT nodes in the ASes that wish to use it. APT nodes MUST ignore mapping announcements for address types that they do not understand.

Total Length - This field specifies the total number of bytes used by all mappings in the announcement. Each mapping announcement can contain mappings for multiple prefixes, each with multiple mapping entries.

Each mapping in the announcement is described by the following fields:

User-space Prefix - This is the user prefix for the mapping.

Public Key - This is a public key that can be used to verify signatures, decrypt data, and prevent misconfigurations for the corresponding user-space prefix. See Section Section 7.1 for more information.

Time To Live (TTL) - This is the amount of time in hours that this mapping should persist in default mappers before being considered obsolete and erased. This value MUST be set to at least three times the regular refresh interval lest the corresponding user-space prefix become unreachable. The TTL is specified in hours to prevent misconfigurations from causing excessive mapping updates.

TS Address Count - This is the total number of TS addresses that the corresponding user-space prefix maps to.

TS Address Set - This is a set of TS addresses, each with a priority. The total number of addresses is specified by the previous field. Priorities are arbitrary integers that only have meaning in reference to each other. Addresses with lower priority values are considered more preferable.

Appendix B. ICMP Mapping Message Fields

User-space Prefix - This prefix is used to match the input address for mapping cache lookups.

TS Address - This is the destination that the user-space prefix maps to.

TTL - This is the time that the entry stays in the cache. Its value is determined by the default mapper.

Appendix C. ICMP Border Link Failure Fields

Prefix - This field contains the user-space prefix that cannot be reached as a result of the border link failure.

Signature - This field can optionally contain a signature generated using the UN's private key. It can then be used to verify the legitimacy of the message.

Appendix D. Hidden Backup Mappings

As mentioned in our mapping section, our design allows users to assign backup providers and perform traffic engineering through appropriate assignment of their TN priority values. Of course, this method will only prove effective if all transit networks generally respect these priority values. This may not be the case in practice.

User networks may be negatively affected if priorities are not respected. For example, imagine that a user has a cheap primary provider and an expensive backup provider. If enough transit networks ignore the UN's preference and send his traffic through the backup provider, the financial impact on the user could be significant. For this reason, users may not want to depend on other ASes to respect their priority values.

In today's Internet, multihomed user networks can use BGP trickery to hide their backup providers unless they are needed. The backup provider simply does not announce a route to the UN's prefix unless it receives a withdrawal for that prefix from the UN's primary provider. At this point, the backup provider will announce its path to the UN's prefix. Once it receives a new announcement for the prefix from the primary provider, the backup provider withdraws its path to the UN's prefix, putting it back into hiding.

In accordance with our "do no harm" design philosophy, we present a

method for including a hidden backup feature into APT. Hidden backup support introduces new ICMP packets, mapping tables, and state into APT. We leave it as an open question whether this feature should be included at all. If transit networks are willing to respect the priority values included with mapping entries, hidden backup support (and its complexity) can be omitted entirely from APT.

Appendix D.1. Hidden Backup Mapping Protocol

A user would want to activate his hidden backup provider in the same three failure situations that require switching to an alternate provider (see Section Section 5.1). We will explain how to handle each of these failure types.

Situation (1) is detectable by the backup provider via BGP. When the backup provider learns that there are no routes to the UN's primary provider, he MUST announce his own backup mapping and begin servicing the user network. If the UN's primary provider later becomes reachable, the backup provider MUST re-announce the original mapping. The responsibility to re-announce the original mapping lies with the backup provider in order to prevent route flapping from causing mapping flapping. The backup provider SHOULD wait until the primary provider has been stable for a set period of time before re-announcing the original mapping. Also note that these mapping announcements are indistinguishable from those generated by permanent mapping changes, leaving default mappers throughout transit space no choice but to respect them.

Situation (2) is detectable by the primary provider via IGP. When the primary provider learns that one of his TRs is down, he MUST inform the backup providers for the affected user networks. This could be done via BGP flooding, but it seems excessive to flood the entire core with a message that is only relevant to a handful of providers. Instead of flooding, the primary provider needs to inform the relevant backup providers directly. To support this, primary providers MUST store a "backup-mapping table" that maps each of their customers to their corresponding backup providers. This table should not be very large, since each provider will only store entries for his own customers. Furthermore, customers who do not have a hidden backup can be excluded from the backup-mapping table.

When a TR goes down, one of the provider's default mappers can use its reverse-mapping table (see Section Section 5.1) to determine which user prefixes are affected. It can then use its backup-mapping table to determine which backup providers need to be notified. The rest of the communication will be implemented using two new ICMP message types, "Primary Provider Failure" and "Primary Provider Recovery". Each of these types will require an acknowledgment (ACK)

flag to ensure delivery.

Primary providers MUST send an ICMP "Primary Provider Failure" message to each of the appropriate backup providers. These messages MUST contain the relevant mapping entry. Upon receipt of such a message, a backup provider MUST respond with an identical packet, except that it MUST set the ACK flag. Then, it MUST announce a backup mapping entry. When the customer's primary provider detects a recovery, it MUST send an ICMP "Primary Provider Recovery" message to the appropriate backup providers. The backup providers MUST acknowledge the message, and re-announce the original mapping. As in situation (1), re-announcing of the original mapping is left to the backup providers to prevent mapping flapping.

Situation (3) is detectable by the TR whose link to a user has gone down. The TR MUST inform his default mapper of this failure via the new ICMP type described in Section Section 5.1.3. At this point, the primary provider can lookup the affected user in his backup-mapping table, and proceed as in situation (2).

The ICMP communication described above is essential to hidden backup functionality. Thus, these messages must be secure and reliable. Security can be achieved with public-private key cryptography (see Section Section 7.3). For reliability, the primary provider MUST continue to send "Primary Provider Failure" and "Primary Provider Recovery" ICMP packets periodically until it receives an acknowledgment from the backup provider. Backup providers MUST always acknowledge these types of ICMP messages, regardless of the state of the corresponding mapping.

Mapping announcements and ICMP communication will be carried out by default mappers unless otherwise specified. Backup-mapping tables are also stored in the default mappers.

Authors' Addresses

Dan Jen

Email: jenster@cs.ucla.edu

Michael Meisel

Email: meisel@cs.ucla.edu

Dan Massey

Email: massey@cs.colostate.edu

Lan Wang

Email: lanwang@memphis.edu

Beichuan Zhang

Email: bzhang@cs.arizona.edu

Lixia Zhang

Email: lixia@cs.ucla.edu

Full Copyright Statement

Intellectual Property

Acknowledgment

Funding for the RFC Editor function is provided by the IETF Administrative Support Activity (IASA).