Internet Research Task Force (IRTF)
Request for Comments: 6115
Category: Informational
ISSN: 2070-1721

T. Li, Ed.
Cisco Systems

February 2011

Recommendation for a Routing Architecture

Abstract

It is commonly recognized that the Internet routing and addressing architecture is facing challenges in scalability, multihoming, and inter-domain traffic engineering. This document presents, as a recommendation of future directions for the IETF, solutions that could aid the future scalability of the Internet. To this end, this document surveys many of the proposals that were brought forward for discussion in this activity, as well as some of the subsequent analysis and the architectural recommendation of the chairs. This document is a product of the Routing Research Group.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Research Task Force (IRTF). The IRTF publishes the results of Internet-related research and development activities. These results might not be suitable for deployment. This RFC represents the individual opinion(s) of one or more members of the Routing Research Group of the Internet Research Task Force (IRTF). Documents approved for publication by the IRSG are not a candidate for any level of Internet Standard; see Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc6115.

Copyright Notice

Copyright (c) 2011 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (http://trustee.ietf.org/license-info) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document.

Table of Contents

1. Introduction

It is commonly recognized that the Internet routing and addressing architecture is facing challenges in scalability, multihoming, and inter-domain traffic engineering. The problem being addressed has been documented in [Scalability_PS], and the design goals that we have discussed can be found in [RRG_Design_Goals].

This document surveys many of the proposals that were brought forward for discussion in this activity. For some of the proposals, this document also includes additional analysis showing some of the concerns with specific proposals, and how some of those concerns may be addressed. Readers are cautioned not to draw any conclusions about the degree of interest or endorsement by the Routing Research Group (RRG) from the presence of any proposals in this document, or the amount of analysis devoted to specific proposals.

1.1. Background to This Document

The RRG was chartered to research and recommend a new routing architecture for the Internet. The goal was to explore many alternatives and build consensus around a single proposal. The only constraint on the group's process was that the process be open and the group set forth with the usual discussion of proposals and trying to build consensus around them. There were no explicit contingencies in the group's process for the eventuality that the group did not reach consensus.

The group met at every IETF meeting from March 2007 to March 2010 and discussed many proposals, both in person and via its mailing list. Unfortunately, the group did not reach consensus. Rather than lose the contributions and progress that had been made, the chairs (Lixia Zhang and Tony Li) elected to collect the proposals of the group and some of the debate concerning the proposals and make a recommendation from those proposals. Thus, the recommendation reflects the opinions of the chairs and not necessarily the consensus of the group.

The group was able to reach consensus on a number of items that are

included below. The proposals included here were collected in an open call amongst the group. Once the proposals were collected, the group was solicited to submit critiques of each proposal. The group was asked to self-organize to produce a single critique for each proposal. In cases where there were several critiques submitted, the editor selected one. The proponents of each proposal then were given the opportunity to write a rebuttal of the critique. Finally, the group again had the opportunity to write a counterpoint of the rebuttal. No counterpoints were submitted. For pragmatic reasons, each submission was severely constrained in length.

All of the proposals were given the opportunity to progress their documents to RFC status; however, not all of them have chosen to pursue this path. As a result, some of the references in this document may become inaccessible. This is unfortunately unavoidable.

The group did reach consensus that the overall document should be published. The document has been reviewed by many of the active members of the Research Group.

1.2. Areas of Group Consensus

The group was also able to reach broad and clear consensus on some terminology and several important technical points. For the sake of posterity, these are recorded here:

1. A "node" is either a host or a router.

2. A "router" is any device that forwards packets at the network layer (e.g., IPv4, IPv6) of the Internet architecture.

3. A "host" is a device that can send/receive packets to/from the network, but does not forward packets.

4. A "bridge" is a device that forwards packets at the link layer (e.g., Ethernet) of the Internet architecture. An Ethernet switch or Ethernet hub are examples of bridges.

5. An "address" is an object that combines aspects of identity with topological location. IPv4 and IPv6 addresses are current examples.

6. A "locator" is a structured topology-dependent name that is not used for node identification and is not a path. Two related meanings are current, depending on the class of things being named:

   1. The topology-dependent name of a node's interface.

2. The topology-dependent name of a single subnetwork OR topology-dependent name of a group of related subnetworks that share a single aggregate. An IP routing prefix is a current example of the latter.

7. An "identifier" is a topology-independent name for a logical node. Depending upon instantiation, a "logical node" might be a single physical device, a cluster of devices acting as a single node, or a single virtual partition of a single physical device. An OSI End System Identifier (ESID) is an example of an identifier. A Fully Qualified Domain Name (FQDN) that precisely names one logical node is another example. (Note well that not all FQDNs meet this definition.)

8. Various other names (i.e., other than addresses, locators, or identifiers), each of which has the sole purpose of identifying a component of a logical system or physical device, might exist at various protocol layers in the Internet architecture.

9. The Research Group has rough consensus that separating identity from location is desirable and technically feasible. However, the Research Group does NOT have consensus on the best engineering approach to such an identity/location split.

10. The Research Group has consensus that the Internet needs to support multihoming in a manner that scales well and does not have prohibitive costs.

11. Any IETF solution to Internet scaling has to not only support multihoming, but address the real-world constraints of the end customers (large and small).

1.3. Abbreviations

This section lists some of the most common abbreviations used in the remainder of this document.

DFZ Default-Free Zone

EID Endpoint IDentifier or Endpoint Interface iDentifier: The precise definition varies depending on the proposal.

ETR Egress Tunnel Router: In a system that tunnels traffic across the existing infrastructure by encapsulating it, the device close to the actual ultimate destination that decapsulates the traffic before forwarding it to the ultimate destination.

FIB Forwarding Information Base: The forwarding table, used in the

data plane of routers to select the next hop for each packet.

ITR Ingress Tunnel Router: In a system that tunnels traffic across the existing infrastructure by encapsulating it, the device close to the actual original source that encapsulates the traffic before using the tunnel to send it to the appropriate ETR.

PA Provider-Aggregatable: Address space that can be aggregated as part of a service provider's routing advertisements.

PI Provider-Independent: Address space assigned by an Internet registry independent of any service provider.

PMTUD Path Maximum Transmission Unit Discovery: The process or mechanism that determines the largest packet that can be sent between a given source and destination without being either i) fragmented (IPv4 only), or ii) discarded (if not fragmentable) because it is too large to be sent down one link in the path from the source to the destination.

RIB Routing Information Base. The routing table, used in the control plane of routers to exchange routing information and construct the FIB.

RIR Regional Internet Registry.

RLOC Routing LOCator: The precise definition varies depending on the proposal.

xTR Tunnel Router: In some systems, the term used to describe a device which can function as both an ITR and an ETR.

2. Locator/ID Separation Protocol (LISP)

2.1. Summary

2.1.1. Key Idea

Implements a locator/identifier separation mechanism using encapsulation between routers at the "edge" of the Internet. Such a separation allows topological aggregation of the routable addresses (locators) while providing stable and portable numbering of end systems (identifiers).

2.1.2. Gains

- topological aggregation of locator space (RLOCs) used for routing, which greatly reduces both the overall size and the "churn rate" of the information needed to operate the Internet global routing system
- separate identifier space (EIDs) for end systems, effectively allowing "PI for all" (no renumbering cost for connectivity changes) without adding state to the global routing system
- improved traffic engineering capabilities that explicitly do not add state to the global routing system and whose deployment will allow active removal of the more-specific state that is currently used
- no changes required to end systems
- no changes to Internet "core" routers
- minimal and straightforward changes to "edge" routers
- day-one advantages for early adopters
- defined router-to-router protocol
- defined database mapping system
- defined deployment plan
- defined interoperability/interworking mechanisms
- defined scalable end-host mobility mechanisms
- prototype implementation already exists and is undergoing testing
- production implementations in progress

2.1.3. Costs

- mapping system infrastructure (map servers, map resolvers, Alternative Logical Topology (ALT) routers). This is considered a new potential business opportunity.
- interworking infrastructure (proxy ITRs). This is considered a new potential business opportunity.

- o overhead for determining/maintaining locator/path liveness. This is a common issue for all identifier/locator separation proposals.

2.1.4. References

[LISP] [LISP+ALT] [LISP-MS] [LISP-Interworking] [LISP-MN] [LIG] [LOC_ID_Implications]

2.2. Critique

LISP+ALT distributes mapping information to ITRs via (optional, local, potentially caching) Map Resolvers and with globally distributed query servers: ETRs and optional Map Servers (MSes).

A fundamental problem with any global query server network is that the frequently long paths and greater risk of packet loss may cause ITRs to drop or significantly delay the initial packets of many new sessions. ITRs drop the packet(s) they have no mapping for. After the mapping arrives, the ITR waits for a re-sent packet and will tunnel that packet correctly. These "initial-packet delays" reduce performance and so create a major barrier to voluntary adoption on a wide enough basis to solve the routing scaling problem.

ALT's delays are compounded by its structure being "aggressively aggregated", without regard to the geographic location of the routers. Tunnels between ALT routers will often span intercontinental distances and traverse many Internet routers.

The many levels to which a query typically ascends in the ALT hierarchy before descending towards its destination will often involve excessively long geographic paths and so worsen initial-packet delays.

No solution has been proposed for these problems or for the contradiction between the need for high aggregation while making the ALT structure robust against single points of failure.

LISP's ITRs' multihoming service restoration depends on their determining the reachability of end-user networks via two or more ETRs. Large numbers of ITRs doing this is inefficient and may overburden ETRs.

Testing reachability of the ETRs is complex and costly -- and insufficient. ITRs cannot test network reachability via each ETR, since the ITRs do not have the address of a device in each ETR's network. So, ETRs must report network unreachability to ITRs.

LISP involves complex communication between ITRs and ETRs, with UDP and 64-bit LISP headers in all traffic packets.

The advantage of LISP+ALT is that its ability to handle billions of EIDs is not constrained by the need to transmit or store the mapping to any one location. Such numbers, beyond a few tens of millions of EIDs, will only result if the system is used for mobility. Yet the concerns just mentioned about ALT's structure arise from the millions of ETRs that would be needed just for non-mobile networks.

In LISP's mobility approach, each Mobile Node (MN) needs an RLOC address to be its own ETR, meaning the MN cannot be behind a NAT. Mapping changes must be sent instantly to all relevant ITRs every time the MN gets a new address -- LISP cannot achieve this.

In order to enforce ISP filtering of incoming packets by source address, LISP ITRs would have to implement the same filtering on each decapsulated packet. This may be prohibitively expensive.

LISP monolithically integrates multihoming failure detection and restoration decision-making processes into the Core-Edge Separation (CES) scheme itself. End-user networks must rely on the necessarily limited capabilities that are built into every ITR.

LISP+ALT may be able to solve the routing scaling problem, but alternative approaches would be superior because they eliminate the initial-packet delay problem and give end-user networks real-time control over ITR tunneling.

2.3. Rebuttal

Initial-packet loss/delays turn out not to be a deep issue. Mechanisms for interoperation with the legacy part of the network are needed in any viably deployable design, and LISP has such mechanisms. If needed, initial packets can be sent via those legacy mechanisms until the ITR has a mapping. (Field experience has shown that the caches on those interoperation devices are guaranteed to be populated, as 'crackers' doing address-space sweeps periodically send packets to every available mapping.)

On ALT issues, it is not at all mandatory that ALT be the mapping system used in the long term. LISP has a standardized mapping system interface, in part to allow reasonably smooth deployment of whatever new mapping system(s) experience might show are required. At least one other mapping system (LISP-TREE) [LISP-TREE], which avoids ALT's problems (such as query load concentration at high-level nodes), has already been laid out and extensively simulated. Exactly what mixture of mapping system(s) is optimal is not really answerable

without more extensive experience, but LISP is designed to allow evolutionary changes to other mapping system(s).

As far as ETR reachability goes, a potential problem to which there is a solution with an adequate level of efficiency, complexity, and robustness is not really a problem. LISP has a number of overlapping mechanisms that it is believed will provide adequate reachability detection (along the three axes above), and in field testing to date, they have behaved as expected.

Operation of LISP devices behind a NAT has already been demonstrated. A number of mechanisms to update correspondent nodes when a mapping is updated have been designed (some are already in use).

3. Routing Architecture for the Next Generation Internet (RANGI)

3.1. Summary

3.1.1. Key Idea

Similar to Host Identity Protocol (HIP) [RFC4423], RANGI introduces a host identifier layer between the network layer and the transport layer, and the transport-layer associations (i.e., TCP connections) are no longer bound to IP addresses, but to host identifiers. The major difference from HIP is that the host identifier in RANGI is a 128-bit hierarchical and cryptographic identifier that has organizational structure. As a result, the corresponding ID->locator mapping system for such identifiers has a reasonable business model and clear trust boundaries. In addition, RANGI uses IPv4-embedded IPv6 addresses as locators. The Locator Domain Identifier (LD ID) (i.e., the leftmost 96 bits) of this locator is a provider-assigned /96 IPv6 prefix, while the last four octets of this locator are a local IPv4 address (either public or private). This special locator could be used to realize 6over4 automatic tunneling (borrowing ideas from the Intra-Site Automatic Tunnel Addressing Protocol (ISATAP) [RFC5214]), which will reduce the deployment cost of this new routing architecture. Within RANGI, the mappings from FQDN to host identifiers are stored in the DNS system, while the mappings from host identifiers to locators are stored in a distributed ID/locator mapping system (e.g., a hierarchical Distributed Hash Table (DHT) system, or a reverse DNS system).

3.1.2. Gains

RANGI achieves almost all of the goals set forth by RRG as follows:

1. Routing Scalability: Scalability is achieved by decoupling identifiers from locators.

2. Traffic Engineering: Hosts located in a multihomed site can suggest the upstream ISP for outbound and inbound traffic, while the first-hop Locator Domain Border Router (LDBR; i.e., site border router) has the final decision on the upstream ISP selection.

3. Mobility and Multihoming: Sessions will not be interrupted due to locator change in cases of mobility or multihoming.

4. Simplified Renumbering: When changing providers, the local IPv4 addresses of the site do not need to change. Hence, the internal routers within the site don't need renumbering.

5. Decoupling Location and Identifier: Obvious.

6. Routing Stability: Since the locators are topologically aggregatable and the internal topology within the LD will not be disclosed outside, routing stability could be improved greatly.

7. Routing Security: RANGI reuses the current routing system and does not introduce any new security risks into the routing system.

8. Incremental Deployability: RANGI allows an easy transition from IPv4 networks to IPv6 networks. In addition, RANGI proxy allows RANGI-aware hosts to communicate to legacy IPv4 or IPv6 hosts, and vice versa.

3.1.3. Costs

1. A host change is required.

2. The first-hop LDBR change is required to support site-controlled traffic-engineering capability.

3. The ID->locator mapping system is a new infrastructure to be deployed.

4. RANGI proxy needs to be deployed for communication between RANGI-aware hosts and legacy hosts.

3.1.4. References

[RFC3007] [RFC4423] [RANGI] [RANGI-PROXY] [RANGI-SLIDES]

3.2. Critique

RANGI is an ID/locator split protocol that, like HIP, places a cryptographically signed ID between the network layer (IPv6) and transport. Unlike the HIP ID, the RANGI ID has a hierarchical structure that allows it to support ID->locator lookups. This hierarchical structure addresses two weaknesses of the flat HIP ID: the difficulty of doing the ID->locator lookup, and the administrative scalability of doing firewall filtering on flat IDs. The usage of this hierarchy is overloaded: it serves to make the ID unique, to drive the lookup process, and possibly other things like firewall filtering. More thought is needed as to what constitutes these levels with respect to these various roles.

The RANGI document [RANGI] suggests FQDN->ID lookup through DNS, and separately an ID->locator lookup that may be DNS or may be something else (a hierarchy of DHTs). It would be more efficient if the FQDN lookup produces both ID and locators (as does the Identifier-Locator Network Protocol (ILNP)). Probably DNS alone is sufficient for the ID->locator lookup since individual DNS servers can hold very large numbers of mappings.

RANGI provides strong sender identification, but at the cost of computing crypto. Many hosts (public web servers) may prefer to forgo the crypto at the expense of losing some functionality (receiver mobility or dynamic multihoming load balancing). While RANGI doesn't require that the receiver validate the sender, it may be good to have a mechanism whereby the receiver can signal to the sender that it is not validating, so that the sender can avoid locator changes.

Architecturally, there are many advantages to putting the mapping function at the end host (versus at the edge). This simplifies the problems of neighbor aliveness and delayed first packet, and avoids stateful middleboxes. Unfortunately, the early-adopter incentive for host upgrade may not be adequate (HIP's lack of uptake being an example).

RANGI does not have an explicit solution for the mobility race condition (there is no mention of a home-agent-like device). However, host-to-host notification combined with fallback on the ID->locators lookup (assuming adequate dynamic update of the lookup system) may be good enough for the vast majority of mobility situations.

RANGI uses proxies to deal with both legacy IPv6 and IPv4 sites. RANGI proxies have no mechanisms to deal with the edge-to-edge aliveness problem. The edge-to-edge proxy approach dirties up an otherwise clean end-to-end model.

RANGI exploits existing IPv6 transition technologies (ISATAP and softwire). These transition technologies are in any event being pursued outside of RRG and do not need to be specified in RANGI drafts per se. RANGI only needs to address how it interoperates with IPv4 and legacy IPv6, which it appears to do adequately well through proxies.

3.3. Rebuttal

The reason why the ID->locator lookup is separated from the FQDN->ID lookup is: 1) not all applications are tied to FQDNs, and 2) it seems unnecessary to require all devices to possess a FQDN of their own. Basically, RANGI uses DNS to realize the ID->locator mapping system. If there are too many entries to be maintained by the authoritative servers of a given Administrative Domain (AD), Distributed Hash Table (DHT) technology can be used to make these authoritative servers scale better, e.g., the mappings maintained by a given AD will be distributed among a group of authoritative servers in a DHT fashion. As a result, the robustness feature of DHT is inherited naturally into the ID->locator mapping system. Meanwhile, there is no trust issue since each AD authority runs its own DHT ring, which maintains only the mappings for those identifiers that are administrated by that AD authority.

For host mobility, if communicating entities are RANGI nodes, the mobile node will notify the correspondent node of its new locator once its locator changes due to a mobility or re-homing event. Meanwhile, it should also update its locator information in the ID->locator mapping system in a timely fashion by using the Secure DNS Dynamic Update mechanism defined in [RFC3007]. In case of simultaneous mobility, at least one of the nodes has to resort to the ID->locator mapping system for resolving the correspondent node's new locator so as to continue their communication. If the correspondent node is a legacy host, Transit Proxies, which fulfill a similar function as the home agents in Mobile IP, will relay the packets between the communicating parties.

RANGI uses proxies (e.g., Site Proxy and Transit Proxy) to deal with both legacy IPv6 and IPv4 sites. Since proxies function as RANGI hosts, they can handle Locator Update Notification messages sent from remote RANGI hosts (or even from remote RANGI proxies) correctly. Hence, there is no edge-to-edge aliveness problem. Details will be specified in a later version of RANGI-PROXY.

The intention behind RANGI using IPv4-embedded IPv6 addresses as locators is to reduce the total deployment cost of this new Internet architecture and to avoid renumbering the site's internal routers when such a site changes ISPs.

4. Internet Vastly Improved Plumbing (Ivip)

4.1. Summary

4.1.1. Key Ideas

Ivip (pronounced eye-vip, est. 2007-06-15) is a Core-Edge Separation scheme for IPv4 and IPv6. It provides multihoming, portability of address space, and inbound traffic engineering for end-user networks of all sizes and types, including those of corporations, SOHO (Small Office, Home Office), and mobile devices.

Ivip meets all the constraints imposed by the need for widespread voluntary adoption [Ivip_Constraints].

Ivip's global fast-push mapping distribution network is structured like a cross-linked multicast tree. This pushes all mapping changes to full-database query servers (QSDs) within ISPs and end-user networks that have ITRs. Each mapping change is sent to all QSDs within a few seconds. (Note: "QSD" is from Query Server with full Database.)

ITRs gain mapping information from these local QSDs within a few tens of milliseconds. QSDs notify ITRs of changed mappings with similarly low latency. ITRs tunnel all traffic packets to the correct ETR without significant delay.

Ivip's mapping consists of a single ETR address for each range of mapped address space. Ivip ITRs do not need to test reachability to ETRs because the mapping is changed in real-time to that of the desired ETR.

End-user networks control the mapping, typically by contracting a specialized company to monitor the reachability of their ETRs, and change the mapping to achieve multihoming and/or traffic engineering (TE). So, the mechanisms that control ITR tunneling are controlled by the end-user networks in real-time and are completely separate from the Core-Edge Separation scheme itself.

ITRs can be implemented in dedicated servers or hardware-based routers. The ITR function can also be integrated into sending hosts. ETRs are relatively simple and only communicate with ITRs rarely -- for Path MTU management with longer packets.

Ivip-mapped ranges of end-user address space need not be subnets. They can be of any length, in units of IPv4 addresses or IPv6 /64s.

Compared to conventional unscalable BGP techniques, and to the use of Core-Edge Separation architectures with non-real-time mapping systems, end-user networks will be able to achieve more flexible and responsive inbound TE. If inbound traffic is split into several streams, each to addresses in different mapped ranges, then real-time mapping changes can be used to steer the streams between multiple ETRs at multiple ISPs.

Default ITRs in the DFZ (DITRs; similar to LISP's Proxy Tunnel Routers) tunnel packets sent by hosts in networks that lack ITRs. So multihoming, portability, and TE benefits apply to all traffic.

ITRs request mappings either directly from a local QSD or via one or more layers of caching query servers (QSCs), which in turn request it from a local QSD. QSCs are optional but generally desirable since they reduce the query load on QSDs. (Note: "QSC" is from Query Server with Cache.)

ETRs may be in ISP or end-user networks. IP-in-IP encapsulation is used, so there is no UDP or any other header. PMTUD (Path MTU Discovery) management with minimal complexity and overhead will handle the problems caused by encapsulation, and adapt smoothly to jumbo frame paths becoming available in the DFZ. The outer header's source address is that of the sending host -- this enables existing ISP Border Router (BR) filtering of source addresses to be extended to encapsulated traffic packets by the simple mechanism of the ETR dropping packets whose inner and outer source address do not match.

4.1.2. Extensions

4.1.2.1. TTR Mobility

The Translating Tunnel Router (TTR) approach to mobility [Ivip_Mobility] is applicable to all Core-Edge Separation techniques and provides scalable IPv4 and IPv6 mobility in which the MN keeps its own mapped IP address(es) no matter how or where it is physically connected, including behind one or more layers of NAT.

Path lengths are typically optimal or close to optimal, and the MN communicates normally with all other non-mobile hosts (no stack or application changes), and of course other MNs. Mapping changes are only needed when the MN uses a new TTR, which would typically occur if the MN moved more than 1000 km. Mapping changes are not required when the MN changes its physical address(es).

4.1.2.2. Modified Header Forwarding

Separate schemes for IPv4 and IPv6 enable tunneling from ITR to ETR without encapsulation. This will remove the encapsulation overhead and PMTUD problems. Both approaches involve modifying all routers between the ITR and ETR to accept a modified form of the IP header. These schemes require new FIB/RIB functionality in DFZ and some other routers but do not alter the BGP functions of DFZ routers.

4.1.3. Gains

- Amenable to widespread voluntary adoption due to no need for host changes, complete support for packets sent from non-upgraded networks and no significant degradation in performance.

- Modular separation of the control of ITR tunneling behavior from the ITRs and the Core-Edge Separation scheme itself: end-user networks control mapping in any way they like, in real-time.

- A small fee per mapping change deters frivolous changes and helps pay for pushing the mapping data to all QSDs. End-user networks that make frequent mapping changes for inbound TE should find these fees attractive considering how it improves their ability to utilize the bandwidth of multiple ISP links.

- End-user networks will typically pay the cost of Open ITR in the DFZ (OITRD) forwarding to their networks. This provides a business model for OITRD deployment and avoids unfair distribution of costs.

- Existing source address filtering arrangements at BRs of ISPs and end-user networks are prohibitively expensive to implement directly in ETRs, but with the outer header's source address being the same as the sending host's address, Ivip ETRs inexpensively enforce BR filtering on decapsulated packets.

4.1.4. Costs

QSDs receive all mapping changes and store a complete copy of the mapping database. However, a worst-case scenario is 10 billion IPv6 mappings, each of 32 bytes, which fits on a consumer hard drive today and should fit in server DRAM by the time such adoption is reached.

The maximum number of non-mobile networks requiring multihoming, etc., is likely to be ~10 million, so most of the 10 billion mappings would be for mobile devices. However, TTR mobility does not involve frequent mapping changes since most MNs only rarely move more than 1000 km.

4.1.5. References

[Ivip_EAF] [Ivip_PMTUD] [Ivip_PLF] [Ivip_Constraints] [Ivip_Mobility] [Ivip_DRTM] [Ivip_Glossary]

4.2. Critique

Looked at from the thousand-foot level, Ivip shares the basic design approaches with LISP and a number of other map-and-encap designs based on the Core-Edge Separation. However, the details differ substantially. Ivip's design makes a bold assumption that, with technology advances, one could afford to maintain a real-time distributed global mapping database for all networks and hosts. Ivip proposes that multiple parties collaborate to build a mapping distribution system that pushes all mapping information and updates to local, full-database query servers located in all ISPs within a few seconds. The system has no single point of failure and uses end-to-end authentication.

A "real time, globally synchronized mapping database" is a critical assumption in Ivip. Using that as a foundation, Ivip design avoids several challenging design issues that others have studied extensively, that include

1. special considerations of mobility support that add additional complexity to the overall system;

2. prompt detection of ETR failures and notification to all relevant ITRs, which turns out to be a rather difficult problem; and

3. development of a partial-mapping lookup sub-system. Ivip assumes the existence of local query servers with a full database with the latest mapping information changes.

To be considered as a viable solution to the Internet routing scalability problem, Ivip faces two fundamental questions. First, whether a global-scale system can achieve real-time synchronized operations as assumed by Ivip is an entirely open question. Past experiences suggest otherwise.

The second question concerns incremental rollout. Ivip represents an ambitious approach, with real-time mapping and local full-database query servers -- which many people regard as impossible. Developing and implementing Ivip may take a fair amount of resources, yet there is an open question regarding how to quantify the gains by first movers -- both those who will provide the Ivip infrastructure and

those that will use it. Significant global routing table reduction only happens when a large enough number of parties have adopted Ivip. The same question arises for most other proposals as well.

One belief is that Ivip's more ambitious mapping system makes a good design tradeoff for the greater benefits for end-user networks and for those that develop the infrastructure. Another belief is that this ambitious design is not viable.

4.3. Rebuttal

Since the Summary and Critique were written, Ivip's mapping system has been significantly redesigned: DRTM - Distributed Real Time Mapping [Ivip_DRTM].

DRTM makes it easier for ISPs to install their own ITRs. It also facilitates Mapped Address Block (MAB) operating companies -- which need not be ISPs -- leasing Scalable Provider-Independent (SPI) address space to end-user networks with almost no ISP involvement. ISPs need not install ITRs or ETRs. For an ISP to support its customers using SPI space, they need only allow the forwarding of outgoing packets whose source addresses are from SPI space. End-user networks can implement their own ETRs on their existing PA address(es) -- and MAB operating companies make all the initial investments.

Once SPI adoption becomes widespread, ISPs will be motivated to install their own ITRs to locally tunnel packets that are sent from customer networks and that must be tunneled to SPI-using customers of the same ISP -- rather than letting these packets exit the ISP's network and return in tunnels to ETRs in the network.

There is no need for full-database query servers in ISPs or for any device that stores the full mapping information for all Mapped Address Blocks (MABs). ISPs that want ITRs will install two or more Map Resolver (MR) servers. These are caching query servers which query multiple (typically nearby) query servers that are full-database for the subset of MABs they serve. These "nearby" query servers will be at DITR sites, which will be run by, or for, MAB operating companies who lease MAB space to large numbers of end-user networks. These DITR-site servers will usually be close enough to the MRs to generate replies with sufficiently low delay and risk of packet loss for ITRs to buffer initial packets for a few tens of milliseconds while the mapping arrives.

DRTM will scale to billions of micronets, tens of thousands of MABs, and potentially hundreds of MAB operating companies, without single points of failure or central coordination.

The critique implies a threshold of adoption is required before significant routing scaling benefits occur. This is untrue of any Core-Edge Separation proposal, including LISP and Ivip. Both can achieve scalable routing benefits in direct proportion to their level of adoption by providing portability, multihoming, and inbound TE to large numbers of end-user networks.

Core-Edge Elimination (CEE) architectures require all Internet communications to change to IPv6 with a new locator/identifier separation naming model. This would impose burdens of extra management effort, packets, and session establishment delays on all hosts -- which is a particularly unacceptable burden on battery-operated mobile hosts that rely on wireless links.

Core-Edge Separation architectures retain the current, efficient, naming model, require no changes to hosts, and support both IPv4 and IPv6. Ivip is the most promising architecture for future development because its scalable, distributed, real-time mapping system best supports TTR mobility, enables ITRs to be simpler, and gives real-time control of ITR tunneling to the end-user network or to organizations they appoint to control the mapping of their micronets.

5. Hierarchical IPv4 Framework (hIPv4)

5.1. Summary

5.1.1. Key Idea

The Hierarchical IPv4 Framework (hIPv4) adds scalability to the routing architecture by introducing additional hierarchy in the IPv4 address space. The IPv4 addressing scheme is divided into two parts, the Area Locator (ALOC) address space, which is globally unique, and the Endpoint Locator (ELOC) address space, which is only regionally unique. The ALOC and ELOC prefixes are added as a shim header between the IP header and transport protocol header; the shim header is identified with a new protocol number in the IP header. Instead of creating a tunneling (i.e., overlay) solution, a new routing element is needed in the service provider's routing domain (called ALOC realm) -- a Locator Swap Router. The current IPv4 forwarding plane remains intact, and no new routing protocols, mapping systems, or caching solutions are required. The control plane of the ALOC realm routers needs some modification in order for ICMP to be compatible with the hIPv4 framework. When an area (one or several autonomous systems (ASes)) of an ISP has transformed into an ALOC realm, only ALOC prefixes are exchanged with other ALOC realms. Directly attached ELOC prefixes are only inserted to the RIB of the local ALOC realm; ELOC prefixes are not distributed to the DFZ. Multihoming can be achieved in two ways, either the enterprise

requests an ALOC prefix from the RIR (this is not recommended) or the enterprise receives the ALOC prefixes from their upstream ISPs. ELOC prefixes are PI addresses and remain intact when an upstream ISP is changed; only the ALOC prefix is replaced. When the RIB of the DFZ is compressed (containing only ALOC prefixes), ingress routers will no longer know the availability of the destination prefix; thus, the endpoints must take more responsibility for their sessions. This can be achieved by using multipath enabled transport protocols, such as SCTP [RFC4960] and Multipath TCP (MPTCP) [MPTCP_Arch], at the endpoints. The multipath transport protocols also provide a session identifier, i.e., verification tag or token; thus, the location and identifier split is carried out -- site mobility, endpoint mobility, and mobile site mobility are achieved. DNS needs to be upgraded: in order to resolve the location of an endpoint, the endpoint must have one ELOC value (current A-record) and at least one ALOC value in DNS (in multihoming solutions there will be several ALOC values for an endpoint).

5.1.2. Gains

1. Improved routing scalability: Adding additional hierarchy to the address space enables more hierarchy in the routing architecture. Early adapters of an ALOC realm will no longer carry the current RIB of the DFZ -- only ELOC prefixes of their directly attached networks and ALOC prefixes from other service providers that have migrated are installed in the ALOC realm's RIB.

2. Scalable support for traffic engineering: Multipath enabled transport protocols are recommended to achieve dynamic load-balancing of a session. Support for Valiant Load-balancing (VLB) [Valiant] schemes has been added to the framework; more research work is required around VLB switching.

3. Scalable support for multihoming: Only attachment points of a multihomed site are advertised (using the ALOC prefix) in the DFZ. DNS will inform the requester on how many attachment points the destination endpoint has. It is the initiating endpoint's choice/responsibility to choose which attachment point is used for the session; endpoints using multipath-enabled transport protocols can make use of several attachment points for a session.

4. Simplified Renumbering: When changing provider, the local ELOC prefixes remains intact; only the ALOC prefix is changed at the endpoints. The ALOC prefix is not used for routing or forwarding decisions in the local network.

5. Decoupling Location and Identifier: The verification tag (SCTP) and token (MPTCP) can be considered to have the characteristics of a session identifier, and thus a session layer is created between the transport and application layers in the TCP/IP model.

6. Routing quality: The hIPv4 framework introduces no tunneling or caching mechanisms. Only a swap of the content in the IPv4 header and locator header at the destination ALOC realm is required; thus, current routing and forwarding algorithms are preserved as such. Valiant Load-balancing might be used as a new forwarding mechanism.

7. Routing Security: Similar as with today's DFZ, except that ELOC prefixes cannot be hijacked (by injecting a longest match prefix) outside an ALOC realm.

8. Deployability: The hIPv4 framework is an evolution of the current IPv4 framework and is backwards compatible with the current IPv4 framework. Sessions in a local network and inside an ALOC realm might in the future still use the current IPv4 framework.

5.1.3. Costs and Issues

1. Upgrade of the stack at an endpoint that is establishing sessions outside the local ALOC realm.

2. In a multihoming solution, the border routers should be able to apply policy-based routing upon the ALOC value in the locator header.

3. New IP allocation policies must be set by the RIRs.

4. There is a short timeframe before the expected depletion of the IPv4 address space occurs.

5. Will enterprises give up their current globally unique IPv4 address block allocation they have gained?

6. Coordination with MPTCP is highly desirable.

5.1.4. References

[hIPv4] [Valiant]

5.2. Critique

hIPv4 is an innovative approach to expanding the IPv4 addressing system in order to resolve the scalable routing problem. This critique does not attempt a full assessment of hIPv4's architecture and mechanisms. The only question addressed here is whether hIPv4 should be chosen for IETF development in preference to, or together with, the only two proposals which appear to be practical solutions for IPv4: Ivip and LISP.

Ivip and LISP appear to have a major advantage over hIPv4 in terms of support for packets sent from non-upgraded hosts/networks. Ivip's DITRs (Default ITRs in the DFZ) and LISP's PTRs (Proxy Tunnel Routers) both accept packets sent by any non-upgraded host/network and tunnel them to the correct ETR -- thus providing the full benefits of portability, multihoming, and inbound TE for these packets as well as those sent by hosts in networks with ITRs. hIPv4 appears to have no such mechanism, so these benefits are only available for communications between two upgraded hosts in upgraded networks.

This means that significant benefits for adopters -- the ability to rely on the new system to provide the portability, multihoming, and inbound TE benefits for all, or almost all, their communications -- will only arise after all, or almost all, networks upgrade their networks, hosts, and addressing arrangements. hIPv4's relationship between adoption levels and benefits to any adopter therefore are far less favorable to widespread adoption than those of Core-Edge Separation (CES) architectures such as Ivip and LISP.

This results in hIPv4 also being at a disadvantage regarding the achievement of significant routing scaling benefits, which likewise will only result once adoption is close to ubiquitous. Ivip and LISP can provide routing scaling benefits in direct proportion to their level of adoption, since all adopters gain full benefits for all their communications, in a highly scalable manner.

hIPv4 requires stack upgrades, which are not required by any CES architecture. Furthermore, a large number of existing IPv4 application protocols convey IP addresses between hosts in a manner that will not work with hIPv4: "There are several applications that are inserting IP address information in the payload of a packet. Some applications use the IP address information to create new sessions or for identification purposes. This section is trying to list the applications that need to be enhanced; however, this is by no means a comprehensive list" [hIPv4].

If even a few widely used applications would need to be rewritten to operate successfully with hIPv4, then this would be such a disincentive to adoption to rule out hIPv4 ever being adopted widely enough to solve the routing scaling problem, especially since CES architectures fully support all existing protocols, without the need for altering host stacks.

It appears that hIPv4 involves major practical difficulties, which mean that in its current form it is not suitable for IETF development.

5.3. Rebuttal

No rebuttal was submitted for this proposal.

6. Name Overlay (NOL) Service for Scalable Internet Routing

6.1. Summary

6.1.1. Key Idea

The basic idea is to add a name overlay (NOL) onto the existing TCP/IP stack.

Its functions include:

1. Managing host name configuration, registration, and authentication;

2. Initiating and managing transport connection channels (i.e., TCP/IP connections) by name;

3. Keeping application data transport continuity for mobility.

At the edge network, we introduce a new type of gateway, a Name Transfer Relay (NTR), which blocks the PI addresses of edge networks into upstream transit networks. NTRs perform address and/or port translation between blocked PI addresses and globally routable addresses, which seem like today's widely used NAT / Network Address Port Translation (NAPT) devices. Both legacy and NOL applications behind a NTR can access the outside as usual. To access the hosts behind a NTR from outside, we need to use NOL to traverse the NTR by name and initiate connections to the hosts behind it.

Different from proposed host-based ID/locator split solutions, such as HIP, Shim6, and name-oriented stack, NOL doesn't need to change the existing TCP/IP stack, sockets, or their packet formats. NOL can coexist with the legacy infrastructure, and the Core-Edge Separation solutions (e.g., APT, LISP, Six/One, Ivip, etc.).

6.1.2. Gains

1. Reduce routing table size: Prevent edge network PI address from leaking into the transit network by deploying gateway NTRs.

2. Traffic Engineering: For legacy and NOL application sessions, the incoming traffic can be directed to a specific NTR by DNS. In addition, for NOL applications, initial sessions can be redirected from one NTR to other appropriate NTRs. These mechanisms provide some support for traffic engineering.

3. Multihoming: When a PI addressed network connects to the Internet by multihoming with several providers, it can deploy NTRs to prevent the PI addresses from leaking into provider networks.

4. Transparency: NTRs can be allocated PA addresses from the upstream providers and store them in NTRs' address pool. By DNS query or NOL session, any session that wants to access the hosts behind the NTR can be delegated to a specific PA address in the NTR address pool.

5. Mobility: The NOL layer manages the traditional TCP/IP transport connections, and provides application data transport continuity by checkpointing the transport connection at sequence number boundaries.

6. No need to change TCP/IP stack, sockets, or DNS system.

7. No need for extra mapping system.

8. NTR can be deployed unilaterally, just like NATs.

9. NOL applications can communicate with legacy applications.

10. NOL can be compatible with existing solutions, such as APT, LISP, Ivip, etc.

11. End-user-controlled multipath indirect routing based on distributed NTRs. This will give benefits to the performance-aware applications, such as video streaming, applications on MSN.com, etc.

6.1.3. Costs

1. Legacy applications have trouble with initiating access to the servers behind NTR. Such trouble can be resolved by deploying the NOL proxy for legacy hosts, or delegating globally routable PA addresses in the NTR address pool for these servers, or deploying a proxy server outside the NTR.

2. NOL may increase the number of entries in DNS, but it is not drastic because it only increases the number of DNS records at domain granularity not the number of hosts. The name used in NOL, for example, is similar to an email address hostname@example.net. The needed DNS entries and query are just for "example.net", and the NTR knows the "hostnames". Not only will the number of DNS records be increased, but the dynamics of DNS might be agitated as well. However, the scalability and performance of DNS are guaranteed by its naming hierarchy and caching mechanisms.

3. Address translating/rewriting costs on NTRs.

6.1.4. References

No references were submitted.

6.2. Critique

1. Applications on hosts need to be rebuilt based on a name overlay library to be NOL-enabled. The legacy software that is not maintained will not be able to benefit from NOL in the Core-Edge Elimination situation. In the Core-Edge Separation scheme, a new gateway NTR is deployed to prevent edge-specific PI prefixes from leaking into the transit core. NOL doesn't impede the legacy endpoints behind the NTR from accessing the outside Internet, but the legacy endpoints cannot access or will have difficultly accessing the endpoints behind a NTR without the help of NOL.

2. In the case of Core-Edge Elimination, the end site will be assigned multiple PA address spaces, which leads to renumbering troubles when switching to other upstream providers. Upgrading endpoints to support NOL doesn't give any benefits to edge networks. Endpoints have little incentive to use NOL in a Core-Edge Elimination scenario, and the same is true with other host-based ID/locator split proposals. Whether they are IPv4 or IPv6 networks, edge networks prefer PI address space to PA address space.

3. In the Core-Edge Separation scenario, the additional gateway NTR is to prevent the specific prefixes from the edge networks, just like a NAT or the ITR/ETR of LISP. A NTR gateway can be seen as an extension of NAT (Network Address Translation). Although NATs are deployed widely, upgrading them to support NOL extension or deploying additional new gateway NTRs at the edge networks is on a voluntary basis and has few economic incentives.

4. The stateful or stateless translation for each packet traversing a NTR will require the cost of the CPU and memory of NTRs, and increase forwarding delay. Thus, it is not appropriate to deploy NTRs at the high-level transit networks where aggregated traffic may cause congestion at the NTRs.

5. In the Core-Edge Separation scenario, the requirement for multihoming and inter-domain traffic engineering will make end sites accessible via multiple different NTRs. For reliability, all of the associations between multiple NTRs and the end site name will be kept in DNS, which may increase the load on DNS.

6. To support mobility, it is necessary for DNS to update the corresponding name-NTR mapping records when an end system moves from behind one NTR to another NTR. The NOL-enabled end relies on the NOL layer to preserve the continuity of the transport layer, since the underlying TCP/UDP transport session would be broken when the IP address changed.

6.3. Rebuttal

NOL resembles neither CEE nor CES as a solution. By supporting application-level sessions through the name overlay layer, NOL can support some solutions in the CEE style. However, NOL is in general closer to CES solutions, i.e., preventing PI prefixes of edge networks from entering into the upstream transit networks. This is done by the NTR, like the ITRs/ETRs in CES solutions, but NOL has no need to define the clear boundary between core and edge networks. NOL is designed to try to provide end users or networks a service that facilitates the adoption of multihoming, multipath routing, and traffic engineering by the indirect routing through NTRs, and that, in the mean time, doesn't accelerate or decelerate the growth of global routing table size.

Some problems are described in the NOL critique. In the original NOL proposal document, the DNS query for a host that is behind a NTR will induce the return of the actual IP addresses of the host and the address of the NTR. This arrangement might cause some difficulties for legacy applications due to the non-standard response from DNS. To resolve this problem, we instead have the NOL service use a new

namespace, and have DNS not return NTR IP addresses for the legacy hosts. The names used for NOL are formatted like email addresses, such as "des@example.net". The mapping between "example.net" and the IP address of the corresponding NTR will be registered in DNS. The NOL layer will understand the meaning of the name "des@example.net" , and it will send a query to DNS only for "example.net". DNS will then return IP addresses of the corresponding NTRs. Legacy applications will still use the traditional FQDN name, and DNS will return the actual IP address of the host. However, if the host is behind a NTR, the legacy applications may be unable to access the host.

The stateless address translation or stateful address and port translation may cause a scaling problem with the number of table entries NTR must maintain, and legacy applications cannot initiate sessions with hosts inside the NOL-adopting End User Network (EUN). However, these problems may not be a big barrier for the deployment of NOL or other similar approaches. Many NAT-like boxes, proxy, and firewall devices are widely used at the ingress/egress points of enterprise networks, campus networks, or other stub EUNs. The hosts running as servers can be deployed outside NTRs or can be assigned PA addresses in an NTR-adopting EUN.

7. Compact Routing in a Locator Identifier Mapping System (CRM)

7.1. Summary

7.1.1. Key Idea

This proposal (referred to here as "CRM") is to build a highly scalable locator identity mapping system using compact routing principles. This provides the means for dynamic topology adaption to facilitate efficient aggregation [CRM]. Map servers are assigned as cluster heads or landmarks based on their capability to aggregate EID announcements.

7.1.2. Gains

- Minimizes the routing table sizes at the system level (i.e., map servers). Provides clear upper bounds for routing stretch that define the packet delivery delay of the map request / first packet.

- Organizes the mapping system based on the EID numbering space, minimizes the administrative overhead of managing the EID space. No need for administratively planned hierarchical address allocation as the system will find convergence into a set of EID allocations.

- Availability and robustness of the overall routing system (including xTRs and map servers) are improved because of the potential to use multiple map servers and direct routes without the involvement of map servers.

7.1.3. Costs

The scalability gains will materialize only in large deployments. If the stretch is bounded to those of compact routing (worst-case stretch less or equal to 3, on average, 1+epsilon), then each xTR needs to have memory/cache for the mappings of its cluster.

7.1.4. References

[CRM]

7.2. Critique

The CRM proposal is not a complete proposal and therefore cannot be considered for further development by the IETF as a scalable routing solution.

While Compact Routing principles may be able to improve a mapping overlay structure such as LISP+ALT, there are several objections to this approach.

Firstly, a CRM-modified ALT structure would still be a global query server system. No matter how ALT's path lengths and delays are optimized, there is a problem with a querier -- which could be anywhere in the world -- relying on mapping information from one or ideally two or more authoritative query servers, which could also be anywhere in the world. The delays and risks of packet loss that are inherent in such a system constitute a fundamental problem. This is especially true when multiple, potentially long, traffic streams are received by ITRs and forwarded over the CRM networks for delivery to the destination network. ITRs must use the CRM infrastructure while they are awaiting a map reply. The traffic forwarded on the CRM infrastructure functions as map requests and can present a scalability and performance issue to the infrastructure.

Secondly, the alterations contemplated in this proposal involve the roles of particular nodes in the network being dynamically assigned as part of the network's self-organizing nature.

The discussion of clustering in the middle of page 4 of [CRM] also indicates that particular nodes are responsible for registering EIDs from typically far-distant ETRs, all of which are handling closely related EIDs that this node can aggregate. Since MSes are apparently

nodes within the compact routing system, and the process of an MS deciding whether to accept EID registrations is determined as part of the self-organizing properties of the system, there are concerns about how EID registration can be performed securely, when no particular physical node is responsible for it.

Thirdly, there are concerns about individually owned nodes performing work for other organizations. Such problems of trust and of responsibilities and costs being placed on those who do not directly benefit already exist in the inter-domain routing system and are a challenge for any scalable routing solution.

There are simpler solutions to the mapping problem than having an elaborate network of routers. If a global-scale query system is still preferred, then it would be better to have ITRs use local MRs, each of which is dynamically configured to know the IP address of the million or so authoritative Map Server (MS) query servers -- or two million or so assuming they exist in pairs for redundancy.

It appears that the inherently greater delays and risks of packet loss of global query server systems make them unsuitable mapping solutions for Core-Edge Elimination or Core-Edge Separation architectures. The solution to these problems appears to involve a greater number of widely distributed authoritative query servers, one or more of which will therefore be close enough to each querier that delays and risk of packet loss are reduced to acceptable levels. Such a structure would be suitable for map requests, but perhaps not for handling traffic packets to be delivered to the destination networks.

7.3. Rebuttal

CRM is most easily understood as an alteration to the routing structure of the LISP+ALT mapping overlay system, by altering or adding to the network's BGP control plane.

CRM's aims include the delivery of initial traffic packets to their destination networks where they also function as map requests. These packet streams may be long and numerous in the fractions of a second to perhaps several seconds that may elapse before the ITR receives the map reply.

Compact Routing principles are used to optimize the path length taken by these query or traffic packets through a significantly modified version of the ALT (or similar) network, while also generally reducing typical or maximum paths taken by the query packets.

An overlay network is a diversion from the shortest path. However, CMR limits this diversion and provides an upper bound. Landmark routers/servers could deliver more than just the first traffic packet, subject to their CPU capabilities and their network connectivity bandwidths.

The trust between the landmarks (mapping servers) can be built based on the current BGP relationships. Registration to the landmark nodes needs to be authenticated mutually between the MS and the system that is registering. This part is not documented in the proposal text.

8. Layered Mapping System (LMS)

8.1. Summary

8.1.1. Key Ideas

The layered mapping system proposal builds a hierarchical mapping system to support scalability, analyzes the design constraints, presents an explicit system structure, designs a two-cache mechanism on ingress tunneling router (ITR) to gain low request delay, and facilitates data validation. Tunneling and mapping are done at the core, and no change is needed on edge networks. The mapping system is run by interest groups independent of any ISP, which conforms to an economical model and can be voluntarily adopted by various networks. Mapping systems can also be constructed stepwise, especially in the IPv6 scenario.

8.1.2. Gains

1. Scalability

    A. Distributed storage of mapping data avoids central storage of massive amounts of data and restricts updates within local areas.

    B. The cache mechanism in an ITR reasonably reduces the request loads on the mapping system.

2. Deployability

    A. No change on edge systems, only tunneling in core routers, and new devices in core networks.

    B. The mapping system can be constructed stepwise: a mapping node needn't be constructed if none of its responsible ELOCs is allocated. This makes sense especially for IPv6.

C. Conforms to a viable economic model: the mapping system operators can profit from their services; core routers and edge networks are willing to join the circle either to avoid router upgrades or realize traffic engineering. Benefits from joining are independent of the scheme's implementation scale.

3. Low request delay: The low number of layers in the mapping structure and the two-stage cache help achieve low request delay.

4. Data consistency: The two-stage cache enables an ITR to update data in the map cache conveniently.

5. Traffic engineering support: Edge networks inform the mapping system of their prioritized mappings with all upstream routers, thus giving the edge networks control over their ingress flows.

8.1.3. Costs

1. Deployment of LMS needs to be further discussed.

2. The structure of the mapping system needs to be refined according to practical circumstances.

8.1.4. References

[LMS_Summary] [LMS]

8.2. Critique

LMS is a mapping mechanism based on Core-Edge Separation. In fact, any proposal that needs a global mapping system with keys with similar properties to that of an "edge address" in a Core-Edge Separation scenario can use such a mechanism. This means that those keys are globally unique (by authorization or just statistically), at the disposal of edge users, and may have several satisfied mappings (with possibly different weights). A proposal to address routing scalability that needs mapping but doesn't specify the mapping mechanism can use LMS to strengthen its infrastructure.

The key idea of LMS is similar to that of LISP+ALT: that the mapping system should be hierarchically organized to gain scalability for storage and updates and to achieve quick indexing for lookups. However, LMS advocates an ISP-independent mapping system, and ETRs are not the authorities of mapping data. ETRs or edge-sites report their mapping data to related mapping servers.

LMS assumes that mapping servers can be incrementally deployed in that a server may not be constructed if none of its administered edge addresses are allocated, and that mapping servers can charge for their services, which provides the economic incentive for their existence. How this brand-new system can be constructed is still not clear. Explicit layering is only an ideal state, and the proposal analyzes the layering limits and feasibility, rather than provide a practical way for deployment.

The drawbacks of LMS's feasibility analysis also include that it 1) is based on current PC power and may not represent future circumstances (especially for IPv6), and 2) does not consider the variability of address utilization. Some IP address spaces may be effectively allocated and used while some may not, causing some mapping servers to be overloaded while others are poorly utilized. More thoughts are needed as to the flexibility of the layer design.

LMS doesn't fit well for mobility. It does not solve the problem when hosts move faster than the mapping updates and propagation between relative mapping servers. On the other hand, mobile hosts' moving across ASes and changing their attachment points (core addresses) is less frequent than hosts' moving within an AS.

Separation needs two planes: Core-Edge Separation (which is to gain routing table scalability) and identity/location separation (which is to achieve mobility). The Global Locator, Local Locator, and Identifier (GLI) scheme does a good clarification of this, and in that case, LMS can be used to provide identity-to-core address mapping. Of course, other schemes may be competent, and LMS can be incorporated with them if the scheme has global keys and needs to map them to other namespaces.

8.3. Rebuttal

No rebuttal was submitted for this proposal.

9. Two-Phased Mapping

9.1. Summary

9.1.1. Considerations

1. A mapping from prefixes to ETRs is an M:M mapping. Any change of a (prefix, ETR) pair should be updated in a timely manner, which can be a heavy burden to any mapping system if the relation changes frequently.

2. A prefix<->ETR mapping system cannot be deployed efficiently if it is overwhelmed by worldwide dynamics. Therefore, the mapping itself is not scalable with this direct mapping scheme.

9.1.2. Basics of a Two-Phased Mapping

1. Introduce an AS number in the middle of the mapping, the phase I mapping is prefix<->AS#, phase II mapping is AS#<->ETRs. This creates a M:1:M mapping model.

2. It is fair to assume that all ASes know their local prefixes (in the IGP) better than other ASes and that it is most likely that local prefixes can be aggregated when they can be mapped to the AS number, which will reduce the number of mapping entries. Also, ASes also know clearly their ETRs on the border between core and edge. So, all mapping information can be collected locally.

3. A registry system will take care of the phase I mapping information. Each AS should have a registration agent to notify the registry of the local range of IP address space. This system can be organized as a hierarchical infrastructure like DNS, or alternatively, as a centralized registry like "whois" in each RIR. Phase II mapping information can be distributed between xTRs as a BGP extension.

4. The basic forwarding procedure is that the ITR first gets the destination AS number from the phase I mapper (or from cache) when the packet is entering the "core". Then, it will extract the closest ETR for the destination AS number. This is local, since phase II mapping information has been "pushed" to the ITR through BGP updates. Finally, the ITR tunnels the packet to the corresponding ETR.

9.1.3. Gains

1. Any prefix reconfiguration (aggregation/deaggregation) within an AS will not be reflected in the mapping system.

2. Local prefixes can be aggregated with a high degree of efficiency.

3. Both phase I and phase II mappings can be stable.

4. A stable mapping system will reduce the update overhead introduced by topology changes and/or routing policy dynamics.

9.1.4. Summary

1. The two-phased mapping scheme introduces an AS number between the mapping prefixes and ETRs.

2. The decoupling of direct mapping makes highly dynamic updates stable; therefore, it can be more scalable than any direct mapping designs.

3. The two-phased mapping scheme is adaptable to any proposals based on the core/edge split.

9.1.5. References

No references were submitted.

9.2. Critique

This is a simple idea on how to scale mapping. However, this design is too incomplete to be considered a serious input to RRG. Take the following two issues as example:

First, in this two-phase scheme, an AS is essentially the unit of destinations (i.e., sending ITRs find out destination AS D, then send data to one of D's ETRs). This does not offer much choice for traffic engineering.

Second, there is no consideration whatsoever on failure detection and handling.

9.3. Rebuttal

No rebuttal was submitted for this proposal.

10. Global Locator, Local Locator, and Identifier Split (GLI-Split)

10.1. Summary

10.1.1. Key Idea

GLI-Split implements a separation between global routing (in the global Internet outside edge networks) and local routing (inside edge networks) using global and local locators (GLs and LLs). In addition, a separate static identifier (ID) is used to identify communication endpoints (e.g., nodes or services) independently of any routing information. Locators and IDs are encoded in IPv6 addresses to enable backwards-compatibility with the IPv6 Internet. The higher-order bits store either a GL or a LL, while the lower-

order bits contain the ID. A local mapping system maps IDs to LLs, and a global mapping system maps IDs to GLs. The full GLI-mode requires nodes with upgraded networking stacks and special GLI-gateways. The GLI-gateways perform stateless locator rewriting in IPv6 addresses with the help of the local and global mapping system. Non-upgraded IPv6 nodes can also be accommodated in GLI-domains since an enhanced DHCP service and GLI-gateways compensate for their missing GLI-functionality. This is an important feature for incremental deployability.

10.1.2. Gains

The benefits of GLI-Split are:

- Hierarchical aggregation of routing information in the global Internet through separation of edge and core routing
- Provider changes not visible to nodes inside GLI-domains (renumbering not needed)
- Rearrangement of subnetworks within edge networks not visible to the outside world (better support of large edge networks)
- Transport connections survive both types of changes
- Multihoming
- Improved traffic engineering for incoming and outgoing traffic
- Multipath routing and load balancing for hosts
- Improved resilience
- Improved mobility support without home agents and triangle routing
- Interworking with the classic Internet
  - without triangle routing over proxy routers
  - without stateful NAT

These benefits are available for upgraded GLI-nodes, but non-upgraded nodes in GLI-domains partially benefit from these advanced features, too. This offers multiple incentives for early adopters, and they have the option to migrate their nodes gradually from non-GLI-stacks to GLI-stacks.

10.1.3. Costs

- Local and global mapping system
- Modified DHCP or similar mechanism
- GLI-gateways with stateless locator rewriting in IPv6 addresses
- Upgraded stacks (only for full GLI-mode)

10.1.4. References

[GLI]

10.2. Critique

GLI-Split makes a clear distinction between two separation planes: the separation between identifier and locator (which is to meet end-users' needs including mobility) and the separation between local and global locator (which makes the global routing table scalable). The distinction is needed since ISPs and hosts have different requirements, with both needing to make the changes inside and outside GLI-domains invisible to their opposites.

A main drawback of GLI-Split is that it puts a burden on hosts. Before routing a packet received from upper layers, network stacks in hosts first need to resolve the DNS name to an IP address; if the IP address is GLI-formed, it may look up the map from the identifier extracted from the IP address to the local locator. If the communication is between different GLI-domains, hosts may further look up the mapping from the identifier to the global locator. Having the local mapping system forward requests to the global mapping system for hosts is just an option. Though host lookup may ease the burden of intermediate nodes, which would otherwise to perform the mapping lookup, the three lookups by hosts in the worst case may lead to large delays unless a very efficient mapping mechanism is devised. The work may also become impractical for low-powered hosts. On one hand, GLI-Split can provide backward compatibility where classic and upgraded IPv6 hosts can communicate. This is its big virtue. On the other hand, the need to upgrade may work against hosts' enthusiasm to change. This is offset against the benefits they would gain.

GLI-Split provides additional features to improve TE and to improve resilience, e.g., exerting multipath routing. However, the cost is that more burdens are placed on hosts, e.g., they may need more lookup actions and route selections. However, these kinds of tradeoffs between costs and gains exist in most proposals.

One improvement of GLI-Split is its support for mobility by updating DNS data as GLI-hosts move across GLI-domains. Through this, the GLI-corresponding-node can query DNS to get a valid global locator of the GLI-mobile-node and need not query the global mapping system (unless it wants to do multipath routing), giving more incentives for nodes to become GLI-enabled. The merits of GLI-Split, including simplified-mobility-handover provision, compensate for the costs of this improvement.

GLI-Split claims to use rewriting instead of tunneling for conversions between local and global locators when packets span GLI-domains. The major advantage is that this kind of rewriting needs no extra state, since local and global locators need not map to each other. Many other rewriting mechanisms instead need to maintain extra state. It also avoids the MTU problem faced by the tunneling methods. However, GLI-Split achieves this only by compressing the namespace size of each attribute (identifier and local/global locator). GLI-Split encodes two namespaces (identifier and local/global locator) into an IPv6 address (each has a size of 2^64 or less), while map-and-encap proposals assume that identifier and locator each occupy a 128-bit space.

10.3. Rebuttal

The arguments in the GLI-Split critique are correct. There are only two points that should be clarified here. First, it is not a drawback that hosts perform the mapping lookups. Second, the critique proposed an improvement to the mobility mechanism, which is of a general nature and not specific to GLI-Split.

1. The additional burden on the hosts is actually a benefit, compared to having the same burden on the gateways. If the gateway would perform the lookups and packets addressed to uncached EIDs arrive, a lookup in the mapping system must be initiated. Until the mapping reply returns, packets must be either dropped, cached, or sent over the mapping system to the destination. All these options are not optimal and have their drawbacks. To avoid these problems in GLI-Split, the hosts perform the lookup. The short additional delay is not a big issue in the hosts because it happens before the first packets are sent. So, no packets are lost or have to be cached. GLI-Split could also easily be adapted to special GLI-hosts (e.g., low-power sensor nodes) that do not have to do any lookup and simply let the gateway do all the work. This functionality is included anyway for backward compatibility with regular IPv6 hosts inside the GLI-domain.

2. The critique proposes a DNS-based mobility mechanism as an improvement to GLI-Split. However, this improvement is an alternative mobility approach that can be applied to any routing architecture (including GLI-Split) and also raises some concerns, e.g., the update speed of DNS. Therefore, we prefer to keep this issue out of the discussion.

11. Tunneled Inter-Domain Routing (TIDR)

11.1. Summary

11.1.1. Key Idea

Provides a method for locator/identifier separation using tunnels between routers on the edge of the Internet transit infrastructure. It enriches the BGP protocol for distributing the identifier-to-locator mapping. Using new BGP attributes, "identifier prefixes" are assigned inter-domain routing locators so that they will not be installed in the RIB and will be moved to a new table called the Tunnel Information Base (TIB). Afterwards, when routing a packet to an "identifier prefix", first the TIB will be searched to perform tunneling, and secondly the RIB will be searched for actual routing. After the edge router performs tunneling, all routers in the middle will route this packet until the packet reaches the router at the tail-end of the tunnel.

11.1.2. Gains

- Smooth deployment
- Size reduction of the global RIB
- Deterministic customer traffic engineering for incoming traffic
- Numerous forwarding decisions for a particular address prefix
- Stops AS number space depletion
- Improved BGP convergence
- Protection of the inter-domain routing infrastructure
- Easy separation of control traffic and transit traffic
- Different layer-2 protocol IDs for transit and non-transit traffic
- Multihoming resilience

- New address families and tunneling techniques
- Support for IPv4 or IPv6, and migration to IPv6
- Scalability, stability, and reliability
- Faster inter-domain routing

11.1.3. Costs

- Routers on the edge of the inter-domain infrastructure will need to be upgraded to hold the mapping database (i.e., the TIB).
- "Mapping updates" will need to be treated differently from usual BGP "routing updates".

11.1.4. References

[TIDR] [TIDR_identifiers] [TIDR_and_LISP] [TIDR_AS_forwarding]

11.2. Critique

TIDR is a Core-Edge Separation architecture from late 2006 that distributes its mapping information via BGP messages that are passed between DFZ routers.

This means that TIDR cannot solve the most important goal of scalable routing -- to accommodate much larger numbers of end-user network prefixes (millions or billions) without each such prefix directly burdening every DFZ router. Messages advertising routes for TIDR-managed prefixes may be handled with lower priority, but this would only marginally reduce the workload for each DFZ router compared to handling an advertisement of a conventional PI prefix.

Therefore, TIDR cannot be considered for RRG recommendation as a solution to the routing scaling problem.

For a TIDR-using network to receive packets sent from any host, every BR of all ISPs must be upgraded to have the new ITR-like functionality. Furthermore, all DFZ routers would need to be altered so they accepted and correctly propagated the routes for end-user network address space, with the new LOCATOR attribute, which contains the ETR address and a REMOTE-PREFERENCE value. Firstly, if they received two such advertisements with different LOCATORs, they would advertise a single route to this prefix containing both. Secondly, for end-user address space (for IPv4) to be more finely divided, the DFZ routers must propagate LOCATOR-containing advertisements for prefixes longer than /24.

TIDR's ITR-like routers store the full mapping database -- so there would be no delay in obtaining mapping, and therefore no significant delay in tunneling traffic packets.

[TIDR] is written as if traffic packets are classified by reference to the RIB, but routers use the FIB for this purpose, and "FIB" does not appear in [TIDR].

TIDR does not specify a tunneling technique, leaving this to be chosen by the ETR-like function of BRs and specified as part of a second kind of new BGP route advertised by that ETR-like BR. There is no provision for solving the PMTUD problems inherent in encapsulation-based tunneling.

ITR functions must be performed by already busy routers of ISPs, rather than being distributed to other routers or to sending hosts. There is no practical support for mobility. The mapping in each end-user route advertisement includes a REMOTE-PREFERENCE for each ETR-like BR, but this is used by the ITR-like functions of BRs to always select the LOCATOR with the highest value. As currently described, TIDR does not provide inbound load-splitting TE.

Multihoming service restoration is achieved initially by the ETR-like function of the BR at the ISP (whose link to the end-user network has just failed). It looks up the mapping to find the next preferred ETR-like BR's address. The first ETR-like router tunnels the packets to the second ETR-like router in the other ISP. However, if the failure was caused by the first ISP itself being unreachable, then connectivity would not be restored until a revised mapping (with higher REMOTE-PREFERENCE) from the reachable ETR-like BR of the second ISP propagated across the DFZ to all ITR-like routers, or the withdrawn advertisement for the first one reaches the ITR-like router.

11.3. Rebuttal

No rebuttal was submitted for this proposal.

12. Identifier-Locator Network Protocol (ILNP)

12.1. Summary

12.1.1. Key Ideas

- Provides crisp separation of Identifiers from Locators.
- Identifiers name nodes, not interfaces.

- Locators name subnetworks, rather than interfaces, so they are equivalent to an IP routing prefix.

- Identifiers are never used for network-layer routing, whilst Locators are never used for Node Identity.

- Transport-layer sessions (e.g., TCP session state) use only Identifiers, never Locators, meaning that changes in location have no adverse impact on an IP session.

12.1.2. Benefits

- The underlying protocol mechanisms support fully scalable site multihoming, node multihoming, site mobility, and node mobility.

- ILNP enables topological aggregation of location information while providing stable and topology-independent identities for nodes.

- In turn, this topological aggregation reduces both the routing prefix "churn" rate and the overall size of the Internet's global routing table, by eliminating the value and need for more-specific routing state currently carried throughout the global (default-free) zone of the routing system.

- ILNP enables improved traffic engineering capabilities without adding any state to the global routing system. TE capabilities include both provider-driven TE and also end-site-controlled TE.

- ILNP's mobility approach:

  * eliminates the need for special-purpose routers (e.g., home agent and/or foreign agent now required by Mobile IP and NEMO).

  * eliminates "triangle routing" in all cases.

  * supports both "make before break" and "break before make" layer-3 handoffs.

- ILNP improves resilience and network availability while reducing the global routing state (as compared with the currently deployed Internet).

- ILNP is incrementally deployable:

  * No changes are required to existing IPv6 (or IPv4) routers.

  * Upgraded nodes gain benefits immediately ("day one"); those benefits gain in value as more nodes are upgraded (this follows Metcalfe's Law).

  * The incremental deployment approach is documented.

- ILNP is backwards compatible:

  * ILNPv6 is fully backwards compatible with IPv6 (ILNPv4 is fully backwards compatible with IPv4).

  * Reuses existing known-to-scale DNS mechanisms to provide identifier/locator mapping.

  * Existing DNS security mechanisms are reused without change.

  * Existing IP Security mechanisms are reused with one minor change (IPsec Security Associations replace the current use of IP addresses with the use of Identifier values). NB: IPsec is also backwards compatible.

  * The backwards compatibility approach is documented.

- No new or additional overhead is required to determine or to maintain locator/path liveness.

- ILNP does not require locator rewriting (NAT); ILNP permits and tolerates NAT, should that be desirable in some deployment(s).

- Changes to upstream network providers do not require node or subnetwork renumbering within end-sites.

- ILNP is compatible with and can facilitate the transition from current single-path TCP to multipath TCP.

- ILNP can be implemented such that existing applications (e.g., applications using the BSD Sockets API) do NOT need any changes or modifications to use ILNP.

12.1.3. Costs

- End systems need to be enhanced incrementally to support ILNP in addition to IPv6 (or IPv4 or both).

- DNS servers supporting upgraded end systems also should be upgraded to support new DNS resource records for ILNP. (The DNS protocol and DNS security do not need any changes.)

12.1.4. References

[ILNP_Site] [MobiArch1] [MobiArch2] [MILCOM1] [MILCOM2] [DNSnBIND] [Referral_Obj] [ILNP_Intro] [ILNP_Nonce] [ILNP_DNS] [ILNP_ICMP] [JSAC_Arch] [RFC4033] [RFC4034] [RFC4035] [RFC5534] [RFC5902]

12.2. Critique

The primary issue for ILNP is how the deployment incentives and benefits line up with the RRG goal of reducing the rate of growth of entries and churn in the core routing table. If a site is currently using PI space, it can only stop advertising that space when the entire site is ILNP capable. This needs (at least) clear elucidation of the incentives for ILNP which are not related to routing scaling, in order for there to be a path for this to address the RRG needs. Similarly, the incentives for upgrading hosts need to align with the value for those hosts.

A closely related question is whether this mechanism actually addresses the sites need for PI addresses. Assuming ILNP is deployed, the site does achieve flexible, resilient, communication using all of its Internet connections. While the proposal addresses the host updates when the host learns of provider changes, there are other aspects of provider change that are not addressed. This includes renumbering routers, subnets, and certain servers. (It is presumed that most servers, once the entire site has moved to ILNP, will not be concerned if their locator changes. However, some servers must have known locators, such as the DNS server.) The issues described in [RFC5887] will be ameliorated, but not resolved. To be able to adopt this proposal, and have sites use it, we need to address these issues. When a site changes points of attachment, only a small amount of DNS provisioning should be required. The LP resource record type is apparently intended to help with this. It is also likely that the use of dynamic DNS will help this.

The ILNP mechanism is described as being suitable for use in conjunction with mobility. This raises the question of race conditions. To the degree that mobility concerns are valid at this time, it is worth asking how communication can be established if a node is sufficiently mobile that it is moving faster than the DNS update and DNS fetch cycle can effectively propagate changes.

This proposal does presume that all communication using this mechanism is tied to DNS names. While it is true that most communication does start from a DNS name, it is not the case that all exchanges have this property. Some communication initiation and referral can be done with an explicit identifier/locator pair. This does appear to require some extensions to the existing mechanism (for

both sides to add locators). In general, some additional clarity on the assumptions regarding DNS, particularly for low-end devices, would seem appropriate.

One issue that this proposal shares with many others is the question of how to determine which locator pairs (local and remote) are actually functional. This is an issue both for initial communications establishment and for robustly maintaining communication. It is likely that a combination of monitoring of traffic (in the host, where this is tractable), coupled with other active measures, can address this. ICMP is clearly insufficient.

12.3. Rebuttal

ILNP eliminates the perceived need for PI addressing and encourages increased DFZ aggregation. Many enterprise users view DFZ scaling issues as too abstruse, so ILNP creates more user-visible incentives to upgrade deployed systems.

ILNP mobility eliminates Duplicate Address Detection (DAD), reducing the layer-3 handoff time significantly when compared to IETF standard Mobile IP, as shown in [MobiArch1] and [MobiArch2]. ICMP location updates separately reduce the layer-3 handoff latency.

Also, ILNP enables both host multihoming and site multihoming. Current BGP approaches cannot support host multihoming. Host multihoming is valuable in reducing the site's set of externally visible nodes.

Improved mobility support is very important. This is shown by the research literature and also appears in discussions with vendors of mobile devices (smartphones, MP3 players). Several operating system vendors push "updates" with major networking software changes in maintenance releases today. Security concerns mean most hosts receive vendor updates more quickly these days.

ILNP enables a site to hide exterior connectivity changes from interior nodes, using various approaches. One approach deploys unique local address (ULA) prefixes within the site, and has the site border router(s) rewrite the Locator values. The usual NAT issues don't arise because the Locator value is not used above the network-layer. [MILCOM1] [MILCOM2]

[RFC5902] makes clear that many users desire IPv6 NAT, with site interior obfuscation as a major driver. This makes global-scope PI addressing much less desirable for end sites than formerly.

ILNP-capable nodes can talk existing IP with legacy IP-only nodes, with no loss of current IP capability. So, ILNP-capable nodes will never be worse off.

Secure Dynamic DNS Update is standard and widely supported in deployed hosts and DNS servers. [DNSnBIND] says many sites have deployed this technology without realizing it (e.g., by enabling both the DHCP server and Active Directory of the MS-Windows Server).

If a node is as mobile as the critique says, then existing IETF Mobile IP standards also will fail. They also use location updates (e.g., MN -> home agent, MN -> foreign agent).

ILNP also enables new approaches to security that eliminate dependence upon location-dependent Access Control Lists (ACLs) without packet authentication. Instead, security appliances track flows using Identifier values and validate the identifier/locator relationship cryptographically [RFC4033] [RFC4034] [RFC4035] or non-cryptographically by reading the nonce [ILNP_Nonce].

The DNS LP record has a more detailed explanation now. LP records enable a site to change its upstream connectivity by changing the L resource records of a single FQDN covering the whole site, thereby providing scalability.

DNS-based server load balancing works well with ILNP by using DNS SRV records. DNS SRV records are not new, are widely available in DNS clients and servers, and are widely used today in the IPv4 Internet for server load balancing.

Recent ILNP documents discuss referrals in more detail. A node with a binary referral can find the FQDN using DNS PTR records, which can be authenticated [RFC4033] [RFC4034] [RFC4035]. Approaches such as [Referral_Obj] improve user experience and user capability, so are likely to self-deploy.

Selection from multiple Locators is identical to an IPv4 system selecting from multiple A records for its correspondent. Deployed IP nodes can track reachability via existing host mechanisms or by using the SHIM6 method. [RFC5534]

13. Enhanced Efficiency of Mapping Distribution Protocols in Map-and-Encap Schemes (EEMDP)

13.1. Summary

13.1.1. Introduction

We present some architectural principles pertaining to the mapping distribution protocols, especially applicable to the map-and-encap (e.g., LISP) type of protocols. These principles enhance the efficiency of the map-and-encap protocols in terms of (1) better utilization of resources (e.g., processing and memory) at Ingress Tunnel Routers (ITRs) and mapping servers, and consequently, (2) reduction of response time (e.g., first-packet delay). We consider how Egress Tunnel Routers (ETRs) can perform aggregation of endpoint ID (EID) address space belonging to their downstream delivery networks, in spite of migration/re-homing of some subprefixes to other ETRs. This aggregation may be useful for reducing the processing load and memory consumption associated with map messages, especially at some resource-constrained ITRs and subsystems of the mapping distribution system. We also consider another architectural concept where the ETRs are organized in a hierarchical manner for the potential benefit of aggregation of their EID address spaces. The two key architectural ideas are discussed in some more detail below. A more complete description can be found in [EEMDP_Considerations] and [EEMDP_Presentation].

It will be helpful to refer to Figures 1, 2, and 3 in [EEMDP_Considerations] for some of the discussions that follow here below.

13.1.2. Management of Mapping Distribution of Subprefixes Spread across Multiple ETRs

To assist in this discussion, we start with the high level architecture of a map-and-encap approach (it would be helpful to see Figure 1 in [EEMDP_Considerations]). In this architecture, we have the usual ITRs, ETRs, delivery networks, etc. In addition, we have the ID-Locator Mapping (ILM) servers, which are repositories for complete mapping information, while the ILM-Regional (ILM-R) servers can contain partial and/or regionally relevant mapping information.

While a large endpoint address space contained in a prefix may be mostly associated with the delivery networks served by one ETR, some fragments (subprefixes) of that address space may be located elsewhere at other ETRs. Let a/20 denote a prefix that is conceptually viewed as composed of 16 subnets of /24 size that are denoted as a1/24, a2/24, ..., a16/24. For example, a/20 is mostly at

ETR1, while only two of its subprefixes a8/24 and a15/24 are elsewhere at ETR3 and ETR2, respectively (see Figure 2 [EEMDP_Considerations]). From the point of view of efficiency of the mapping distribution protocol, it may be beneficial for ETR1 to announce a map for the entire space a/20 (rather than fragment it into a multitude of more-specific prefixes), and provide the necessary exceptions in the map information. Thus, the map message could be in the form of Map:(a/20, ETR1; Exceptions: a8/24, a15/24). In addition, ETR2 and ETR3 announce the maps for a15/24 and a8/24, respectively, and so the ILMs know where the exception EID addresses are located. Now consider a host associated with ITR1 initiating a packet destined for an address a7(1), which is in a7/24 that is not in the exception portion of a/20. Now a question arises as to which of the following approaches would be the best choice:

1. ILM-R provides the complete mapping information for a/20 to ITR1 including all maps for relevant exception subprefixes.

2. ILM-R provides only the directly relevant map to ITR1, which in this case is (a/20, ETR1).

In the first approach, the advantage is that ITR1 would have the complete mapping for a/20 (including exception subnets), and it would not have to generate queries for subsequent first packets that are destined to any address in a/20, including a8/24 and a15/24. However, the disadvantage is that if there is a significant number of exception subprefixes, then the very first packet destined for a/20 will experience a long delay, and also the processors at ITR1 and ILM-R can experience overload. In addition, the memory usage at ITR1 can be very inefficient. The advantage of the second approach above is that the ILM-R does not overload resources at ITR1, neither in terms of processing or memory usage, but it needs an enhanced map response in of the form Map:(a/20, ETR1, MS=1), where the MS (More Specific) indicator is set to 1 to indicate to ITR1 that not all subnets in a/20 map to ETR1. The key idea is that aggregation is beneficial, and subnet exceptions must be handled with additional messages or indicators in the maps.

13.1.3. Management of Mapping Distribution for Scenarios with Hierarchy of ETRs and Multihoming

Now we highlight another architectural concept related to mapping management (please refer to Figure 3 in [EEMDP_Considerations]). Here we consider the possibility that ETRs may be organized in a hierarchical manner. For instance, ETR7 is higher in the hierarchy relative to ETR1, ETR2, and ETR3, and like-wise ETR8 is higher relative to ETR4, ETR5, and ETR6. For instance, ETRs 1 through 3 can relegate the locator role to ETR7 for their EID address space. In

essence, they can allow ETR7 to act as the locator for the delivery networks in their purview. ETR7 keeps a local mapping table for mapping the appropriate EID address space to specific ETRs that are hierarchically associated with it in the level below. In this situation, ETR7 can perform EID address space aggregation across ETRs 1 through 3 and can also include its own immediate EID address space for the purpose of that aggregation. The many details related to this approach and special circumstances involving multihoming of subnets are discussed in detail in [EEMDP_Considerations]. The hierarchical organization of ETRs and delivery networks should help in the future growth and scalability of ETRs and mapping distribution networks. This is essentially recursive map-and-encap, and some of the mapping distribution and management functionality will remain local to topologically neighboring delivery networks that are hierarchically underneath ETRs.

13.1.4. References

[EEMDP_Considerations] [EEMDP_Presentation] [FIBAggregatability]

13.2. Critique

The scheme described in [EEMDP_Considerations] represents one approach to mapping overhead reduction, and it is a general idea that is applicable to any proposal that includes prefix or EID aggregation. A somewhat similar idea is also used in Level-3 aggregation in the FIB aggregation proposal [FIBAggregatability]. There can be cases where deaggregation of EID prefixes occur in such a way that the bulk of an EID prefix P would be attached to one locator (say, ETR1) while a few subprefixes under P would be attached to other locators elsewhere (say, ETR2, ETR3, etc.). Ideally, such cases should not happen; however, in reality it can happen as the RIR's address allocations are imperfect. In addition, as new IP address allocations become harder to get, an IPv4 prefix owner might split previously unused subprefixes of that prefix and allocate them to remote sites (homed to other ETRs). Assuming these situations could arise in practice, the nature of the solution would be that the response from the mapping server for the coarser site would include information about the more specifics. The solution as presented seems correct.

The proposal mentions that in Approach 1, the ID-Locator Mapping (ILM) system provides the complete mapping information for an aggregate EID prefix to a querying ITR, including all the maps for the relevant exception subprefixes. The sheer number of such more-specifics can be worrisome, for example, in LISP. What if a company's mobile-node EIDs came out of their corporate EID prefix? Approach 2 is far better but still there may be too many entries for

a regional ILM to store. In Approach 2, the ILM communicates that there are more specifics but does not communicate their mask-length. A suggested improvement would be that rather than saying that there are more specifics, indicate what their mask-lengths are. There can be multiple mask lengths. This number should be pretty small for IPv4 but can be large for IPv6.

Later in the proposal, a different problem is addressed, involving a hierarchy of ETRs and how aggregation of EID prefixes from lower-level ETRs can be performed at a higher-level ETR. The various scenarios here are well illustrated and described. This seems like a good idea, and a solution like LISP can support this as specified. As any optimization scheme would inevitably add some complexity; the proposed scheme for enhancing mapping efficiency comes with some of its own overhead. The gain depends on the details of specific EID blocks, i.e., how frequently the situations (such as an ETR that has a bigger EID block with a few holes) arise.

13.3. Rebuttal

There are two main points in the critique that are addressed here: (1) The gain depends on the details of specific EID blocks, i.e., how frequently the situations arise such as an ETR having a bigger EID block with a few holes, and (2) Approach 2 is lacking an added feature of conveying just the mask-length of the more specifics that exist as part of the current map response.

Regarding comment (1) above, there are multiple possibilities regarding how situations can arise, resulting in allocations having holes in them. An example of one of these possibilities is as follows. Org-A has historically received multiple /20s, /22s, and /24s over the course of time that are adjacent to each other. At the present time, these prefixes would all aggregate to a /16 but for the fact that just a few of the underlying /24s have been allocated elsewhere historically to other organizations by an RIR or ISPs. An example of a second possibility is that Org-A has an allocation of a /16. It has suballocated a /22 to one of its subsidiaries, and subsequently sold the subsidiary to another Org-B. For ease of keeping the /22 subnet up and running without service disruption, the /22 subprefix is allowed to be transferred in the acquisition process. Now the /22 subprefix originates from a different AS and is serviced by a different ETR (as compared to the parent \16 prefix). We are in the process of performing an analysis of RIR allocation data and are aware of other studies (notably at UCLA) that are also performing similar analysis to quantify the frequency of occurrence of the holes. We feel that the problem that has been addressed is a realistic one, and the proposed scheme would help reduce the overheads associated with the mapping distribution system.

Regarding comment (2) above, the suggested modification to Approach 2 would be definitely beneficial. In fact, we feel that it would be fairly straightforward to dynamically use Approach 1 or Approach 2 (with the suggested modification), depending on whether there are only a few (e.g., <=5) or many (e.g., >5) more specifics, respectively. The suggested modification of notifying the mask-length of the more specifics in the map response is indeed very helpful because then the ITR would not have to resend a map-query for EID addresses that match the EID address in the previous query up to at least mask-length bit positions. There can be a two-bit field in the map response that would be interpreted as follows.

(a) value 00: there are no more specifics

(b) value 01: there are more specifics and their exact information follows in additional map-responses

(c) value 10: there are more-specifics and the mask-length of the next more-specific is indicated in the current map-response.

An additional field will be included that will be used to specify the mask-length of the next more-specific in the case of value 10 (case (c) above).

14. Evolution

14.1. Summary

As the Internet continues its rapid growth, router memory size and CPU cycle requirements are outpacing feasible hardware upgrade schedules. We propose to solve this problem by applying aggregation with increasing scopes to gradually evolve the routing system towards a scalable structure. At each evolutionary step, our solution is able to interoperate with the existing system and provide immediate benefits to adopters to enable deployment. This document summarizes the need for an evolutionary design, the relationship between our proposal and other revolutionary proposals, and the steps of aggregation with increasing scopes. Our detailed proposal can be found in [Evolution].

14.1.1. Need for Evolution

Multiple different views exist regarding the routing scalability problem. Networks differ vastly in goals, behavior, and resources, giving each a different view of the severity and imminence of the scalability problem. Therefore, we believe that, for any solution to be adopted, it will start with one or a few early adopters and may not ever reach the entire Internet. The evolutionary approach

recognizes that changes to the Internet can only be a gradual process with multiple stages. At each stage, adopters are driven by and rewarded with solving an immediate problem. Each solution must be deployable by individual networks who deem it necessary at a time they deem it necessary, without requiring coordination from other networks, and the solution has to bring immediate relief to a single first-mover.

14.1.2. Relation to Other RRG Proposals

Most proposals take a revolutionary approach that expects the entire Internet to eventually move to some new design whose main benefits would not materialize until the vast majority of the system has been upgraded; their incremental deployment plan simply ensures interoperation between upgraded and legacy parts of the system. In contrast, the evolutionary approach depicts a system where changes may happen here and there as needed, but there is no dependency on the system as a whole making a change. Whoever takes a step forward gains the benefit by solving his own problem, without depending on others to take actions. Thus, deployability includes not only interoperability, but also the alignment of costs and gains.

The main differences between our approach and more revolutionary map-and-encap proposals are: (a) we do not start with a pre-defined boundary between edge and core; and (b) each step brings immediate benefits to individual first-movers. Note that our proposal neither interferes nor prevents any revolutionary host-based solutions such as ILNP from being rolled out. However, host-based solutions do not bring useful impact until a large portion of hosts have been upgraded. Thus, even if a host-based solution is rolled out in the long run, an evolutionary solution is still needed for the near term.

14.1.3. Aggregation with Increasing Scopes

Aggregating many routing entries to a fewer number is a basic approach to improving routing scalability. Aggregation can take different forms and be done within different scopes. In our design, the aggregation scope starts from a single router, then expands to a single network and neighbor networks. The order of the following steps is not fixed but is merely a suggestion; it is under each individual network's discretion which steps they choose to take based on their evaluation of the severity of the problems and the affordability of the solutions.

1. FIB Aggregation (FA) in a single router. A router algorithmically aggregates its FIB entries without changing its RIB or its routing announcements. No coordination among routers

is needed, nor any change to existing protocols. This brings scalability relief to individual routers with only a software upgrade.

2. Enabling 'best external' on Provider Edge routers (PEs), Autonomous System Border Routers (ASBRs), and Route Reflectors (RRs), and turning on next-hop-self on RRs. For hierarchical networks, the RRs in each Point of Presence (PoP) can serve as a default gateway for nodes in the PoP, thus allowing the non-RR nodes in each PoP to maintain smaller routing tables that only include paths that egress that PoP. This is known as 'topology-based mode' Virtual Aggregation, and can be done with existing hardware and configuration changes only. Please see [Evolution_Grow_Presentation] for details.

3. Virtual Aggregation (VA) in a single network. Within an AS, some fraction of existing routers are designated as Aggregation Point Routers (APRs). These routers are either individually or collectively maintain the full FIB table. Other routers may suppress entries from their FIBs, instead forwarding packets to APRs, which will then tunnel the packets to the correct egress routers. VA can be viewed as an intra-domain map-and-encap system to provide the operators with a control mechanism for the FIB size in their routers.

4. VA across neighbor networks. When adjacent networks have VA deployed, they can go one step further by piggybacking egress router information on existing BGP announcements, so that packets can be tunneled directly to a neighbor network's egress router. This improves packet delivery performance by performing the encapsulation/decapsulation only once across these neighbor networks, as well as reducing the stretch of the path.

5. Reducing RIB Size by separating the control plane from the data plane. Although a router's FIB can be reduced by FA or VA, it usually still needs to maintain the full RIB to produce complete routing announcements to its neighbors. To reduce the RIB size, a network can set up special boxes, which we call controllers, to take over the External BGP (eBGP) sessions from border routers. The controllers receive eBGP announcements, make routing decisions, and then inform other routers in the same network of how to forward packets, while the regular routers just focus on the job of forwarding packets. The controllers, not being part of the data path, can be scaled using commodity hardware.

6. Insulating forwarding routers from routing churn. For routers with a smaller RIB, the rate of routing churn is naturally reduced. Further reduction can be achieved by not announcing

failures of customer prefixes into the core, but handling these failures in a data-driven fashion, e.g., a link failure to an edge network is not reported unless and until there are data packets that are heading towards the failed link.

14.1.4. References

[Evolution] [Evolution_Grow_Presentation]

14.2. Critique

All of the RRG proposals that scale the routing architecture share one fundamental approach, route aggregation, in different forms, e.g., LISP removes "edge prefixes" using encapsulation at ITRs, and ILNP achieves the goal by locator rewrite. In this evolutionary path proposal, each stage of the evolution applies aggregation with increasing scopes to solve a specific scalability problem, and eventually the path leads towards global routing scalability. For example, it uses FIB aggregation at the single router level, virtual aggregation at the network level, and then between neighboring networks at the inter-domain level.

Compared to other proposals, this proposal has the lowest hurdle to deployment, because it does not require that all networks move to use a global mapping system or upgrade all hosts, and it is designed for each individual network to get immediate benefits after its own deployment.

Criticisms of this proposal fall into two types. The first type concerns several potential issues in the technical design as listed below:

1. FIB aggregation, at level-3 and level-4, may introduce extra routable space. Concerns have been raised about the potential routing loops resulting from forwarding otherwise non-routable packets, and the potential impact on Reverse Path Forwarding (RPF) checking. These concerns can be addressed by choosing a lower level of aggregation and by adding null routes to minimize the extra space, at the cost of reduced aggregation gain.

2. Virtual Aggregation changes the traffic paths in an ISP network, thereby introducing stretch. Changing the traffic path may also impact the reverse path checking practice used to filter out packets from spoofed sources. More analysis is need to identify the potential side-effects of VA and to address these issues.

3. The current Virtual Aggregation description is difficult to understand, due to its multiple options for encapsulation and popular prefix configurations, which makes the mechanism look overly complicated. More thought is needed to simplify the design and description.

4. FIB Aggregation and Virtual Aggregation may require additional operational cost. There may be new design trade-offs that the operators need to understand in order to select the best option for their networks. More analysis is needed to identify and quantify all potential operational costs.

5. In contrast to a number of other proposals, this solution does not provide mobility support. It remains an open question as to whether the routing system should handle mobility.

The second criticism is whether deploying quick fixes like FIB aggregation would alleviate scalability problems in the short term and reduce the incentives for deploying a new architecture; and whether an evolutionary approach would end up with adding more and more patches to the old architecture, and not lead to a fundamentally new architecture as the proposal had expected. Though this solution may get rolled out more easily and quickly, a new architecture, if/once deployed, could solve more problems with cleaner solutions.

14.3. Rebuttal

No rebuttal was submitted for this proposal.

15. Name-Based Sockets

15.1. Summary

Name-based sockets are an evolution of the existing address-based sockets, enabling applications to initiate and receive communication sessions based on the use of domain names in lieu of IP addresses. Name-based sockets move the existing indirection from domain names to IP addresses from its current position in applications down to the IP layer. As a result, applications communicate exclusively based on domain names, while the discovery, selection, and potentially in-session re-selection of IP addresses is centrally performed by the IP stack itself.

Name-based sockets help mitigate the Internet routing scalability problem by separating naming and addressing more consistently than what is possible with the existing address-based sockets. This supports IP address aggregation because it simplifies the use of IP

addresses with high topological significance, as well as the dynamic replacement of IP addresses during network-topological and host-attachment changes.

A particularly positive effect of name-based sockets on Internet routing scalability is the new incentives for edge network operators to use provider-assigned IP addresses, which are more aggregatable than the typically preferred provider-independent IP addresses. Even though provider-independent IP addresses are harder to get and more expensive than provider-assigned IP addresses, many operators desire provider-independent addresses due to the high indirect cost of provider-assigned IP addresses. This indirect cost is comprised of both difficulties in multihoming, and tedious and largely manual renumbering upon provider changes.

Name-based sockets reduce the indirect cost of provider-assigned IP addresses in three ways, and hence make the use of provider-assigned IP addresses more acceptable: (1) They enable fine-grained and responsive multihoming. (2) They simplify renumbering by offering an easy means to replace IP addresses in referrals with domain names. This helps avoiding updates to application and operating system configurations, scripts, and databases during renumbering. (3) They facilitate low-cost solutions that eliminate renumbering altogether. One such low-cost solution is IP address translation, which in combination with name-based sockets loses its adverse impact on applications.

The prerequisite for a positive effect of name-based sockets on Internet routing scalability is their adoption in operating systems and applications. Operating systems should be augmented to offer name-based sockets as a new alternative to the existing address-based sockets, and applications should use name-based sockets for their communications. Neither an instantaneous, nor an eventually complete transition to name-based sockets is required, yet the positive effect on Internet routing scalability will grow with the extent of this transition.

Name-based sockets were hence designed with a focus on deployment incentives, comprising both immediate deployment benefits as well as low deployment costs. Name-based sockets provide a benefit to application developers because the alleviation of applications from IP address management responsibilities simplifies and expedites application development. This benefit is immediate owing to the backwards compatibility of name-based sockets with legacy applications and legacy peers. The appeal to application developers, in turn, is an immediate benefit for operating system vendors who adopt name-based sockets.

Name-based sockets furthermore minimize deployment costs: Alternative techniques to separate naming and addressing provide applications with "surrogate IP addresses" that dynamically map onto regular IP addresses. A surrogate IP address is indistinguishable from a regular IP address for applications, but does not have the topological significance of a regular IP address. Mobile IP and the Host Identity Protocol are examples of such separation techniques. Mobile IP uses "home IP addresses" as surrogate IP addresses with reduced topological significance. The Host Identity Protocol uses "host identifiers" as surrogate IP addresses without topological significance. A disadvantage of surrogate IP addresses is their incurred cost in terms of extra administrative overhead and, for some techniques, extra infrastructure. Since surrogate IP addresses must be resolvable to the corresponding regular IP addresses, they must be provisioned in the DNS or similar infrastructure. Mobile IP uses a new infrastructure of home agents for this purpose, while the Host Identity Protocol populates DNS servers with host identities. Name-based sockets avoid this cost because they function without surrogate IP addresses, and hence without the provisioning and infrastructure requirements that accompany surrogate addresses.

Certainly, some edge networks will continue to use provider-independent addresses despite name-based sockets, perhaps simply due to inertia. But name-based sockets will help reduce the number of those networks, and thus have a positive impact on Internet routing scalability.

A more comprehensive description of name-based sockets can be found in [Name_Based_Sockets].

15.1.1. References

[Name_Based_Sockets]

15.2. Critique

Name-based sockets contribution to the routing scalability problem is to decrease the reliance on PI addresses, allowing a greater use of PA addresses, and thus a less fragmented routing table. It provides end hosts with an API which makes the applications address-agnostic. The name abstraction allows the hosts to use any type of locator, independent of format or provider. This increases the motivation and usability of PA addresses. Some applications, in particular bootstrapping applications, may still require hard coded IP addresses, and as such will still motivate the use of PI addresses.

15.2.1. Deployment

The main incentives and drivers are geared towards the transition of applications to the name-based sockets. Adoption by applications will be driven by benefits in terms of reduced application development cost. Legacy applications are expected to migrate to the new API at a slower pace, as the name-based sockets are backwards compatible, this can happen in a per-host fashion. Also, not all applications can be ported to a FQDN dependent infrastructure, e.g., DNS functions. This hurdle is manageable, and may not be a definite obstacle for the transition of a whole domain, but it needs to be taken into account when striving for mobility/multihoming of an entire site. The transition of functions on individual hosts may be trivial, either through upgrades/changes to the OS or as linked libraries. This can still happen incrementally and independently, as compatibility is not affected by the use of name-based sockets.

15.2.2. Edge-networks

Name-based sockets rely on the transition of individual applications and are backwards compatible, so they do not require bilateral upgrades. This allows each host to migrate its applications independently. Name-based sockets may make an individual client agnostic to the networking medium, be it PA/PI IP-addresses or in a the future an entirely different networking medium. However, an entire edge-network, with internal and external services will not be able to make a complete transition in the near future. Hence, even if a substantial fraction of the hosts in an edge-network use name-based sockets, PI addresses may still be required by the edge-network. In short, new services may be implemented using name-based sockets, old services may be ported. Name-based sockets provide an increased motivation to move to PA-addresses as actual provider independence relies less and less on PI-addressing.

15.3. Rebuttal

No rebuttal was submitted for this proposal.

16. Routing and Addressing in Networks with Global Enterprise Recursion (IRON-RANGER)

16.1. Summary

RANGER is a locator/identifier separation approach that uses IP-in-IP encapsulation to connect edge networks across transit networks such as the global Internet. End systems use endpoint interface identifier (EID) addresses that may be routable within edge networks but do not appear in transit network routing tables. EID to Routing

Locator (RLOC) address bindings are instead maintained in mapping tables and also cached in default router FIBs (i.e., very much the same as for the global DNS and its associated caching resolvers). RANGER enterprise networks are organized in a recursive hierarchy with default mappers connecting lower layers to the next higher layer in the hierarchy. Default mappers forward initial packets and push mapping information to lower-tier routers and end systems through secure redirection.

RANGER is an architectural framework derived from the Intra-Site Automatic Tunnel Addressing Protocol (ISATAP).

16.1.1. Gains

- provides a scalable routing system alternative in instances where dynamic routing protocols are impractical
- naturally supports a recursively-nested "network-of-networks" (or, "enterprise-within-enterprise") hierarchy
- uses asymmetric security mechanisms (i.e., secure neighbor discovery) to secure router discovery and the redirection mechanism
- can quickly detect path failures and pick alternate routes
- naturally supports provider-independent addressing
- support for site multihoming and traffic engineering
- ingress filtering for multihomed sites
- mobility-agile through explicit cache invalidation (much more reactive than dynamic DNS)
- supports neighbor discovery and neighbor unreachability detection over tunnels
- no changes to end systems
- no changes to most routers
- supports IPv6 transition

- compatible with true identity/locator split mechanisms such as HIP (i.e., packets contain a HIP Host Identity Tag (HIT) as an end system identifier, IPv6 address as endpoint interface identifier (EID) in the inner IP header and IPv4 address as Routing LOCator (RLOC) in the outer IP header)

- prototype code available

16.1.2. Costs

- new code needed in enterprise border routers

- locator/path liveness detection using RFC 4861 neighbor unreachability detection (i.e., extra control messages, but data-driven) [RFC4861]

16.1.3. References

[IRON] [RANGER_Scen] [VET] [SEAL] [RFC5201] [RFC5214] [RFC5720]

16.2. Critique

The RANGER architectural framework is intended to be applicable for a Core-Edge Separation (CES) architecture for scalable routing, using either IPv4 or IPv6 -- or using both in an integrated system which may carry one protocol over the other.

However, despite [IRON] being readied for publication as an experimental RFC, the framework falls well short of the level of detail required to envisage how it could be used to implement a practical scalable routing solution. For instance, the document contains no specification for a mapping protocol, or how the mapping lookup system would work on a global scale.

There is no provision for RANGER's ITR-like routers being able to probe the reachability of end-user networks via multiple ETR-like routers -- nor for any other approach to multihoming service restoration.

Nor is there any provision for inbound TE or support of mobile devices which frequently change their point of attachment.

Therefore, in its current form, RANGER cannot be contemplated as a superior scalable routing solution to some other proposals which are specified in sufficient detail and which appear to be feasible.

RANGER uses its own tunneling and PMTUD management protocol: SEAL. Adoption of SEAL in its current form would prevent the proper utilization of jumbo frame paths in the DFZ, which will become the norm in the future. SEAL uses "Packet Too Big" [RFC4443] and "Fragmentation Needed" [RFC0792] messages to the sending host only to fix a preset maximum packet length. To avoid the need for the SEAL layer to fragment packets of this length, this MTU value (for the input of the tunnel) needs to be set significantly below 1500 bytes, assuming the typically ~1500 byte MTU values for paths across the DFZ today. In order to avoid this excessive fragmentation, this value could only be raised to a ~9k byte value at some time in the future where essentially all paths between ITRs and ETRs were jumbo frame capable.

16.3. Rebuttal

The Internet Routing Overlay Network (IRON) [IRON] is a scalable Internet routing architecture that builds on the RANGER recursive enterprise network hierarchy [RFC5720]. IRON bonds together participating RANGER networks using VET [VET] and SEAL [SEAL] to enable secure and scalable routing through automatic tunneling within the Internet core. The IRON-RANGER automatic tunneling abstraction views the entire global Internet DFZ as a virtual Non-Broadcast Multi-Access (NBMA) link similar to ISATAP [RFC5214].

IRON-RANGER is an example of a Core-Edge Separation (CES) system. Instead of a classical mapping database, however, IRON-RANGER uses a hybrid combination of a proactive dynamic routing protocol for distributing highly aggregated Virtual Prefixes (VPs) and an on-demand data driven protocol for distributing more-specific Provider-Independent (PI) prefixes derived from the VPs.

The IRON-RANGER hierarchy consists of recursively-nested RANGER enterprise networks joined together by IRON routers that participate in a global BGP instance. The IRON BGP instance is maintained separately from the current Internet BGP Routing LOCator (RLOC) address space (i.e., the set of all public IPv4 prefixes in the Internet). Instead, the IRON BGP instance maintains VPs taken from Endpoint Interface iDentifier (EID) address space, e.g., the IPv6 global unicast address space. To accommodate scaling, only O(10k) -- O(100k) VPs are allocated e.g., using /20 or shorter IPv6 prefixes.

IRON routers lease portions of their VPs as Provider-Independent (PI) prefixes for customer equipment (CEs), thereby creating a sustainable business model. CEs that lease PI prefixes propagate address mapping(s) throughout their attached RANGER networks and up to VP-owning IRON router(s) through periodic transmission of "bubbles" with authentication and PI prefix information. Routers in RANGER networks

and IRON routers that receive and forward the bubbles securely install PI prefixes in their FIBs, but do not inject them into the RIB. IRON routers therefore keep track of only their customer base via the FIB entries and keep track of only the Internet-wide VP database in the RIB.

IRON routers propagate more-specific prefixes using secure redirection to update router FIBs. Prefix redirection is driven by the data plane and does not affect the control plane. Redirected prefixes are not injected into the RIB, but rather are maintained as FIB soft state that is purged after expiration or route failure. Neighbor unreachability detection is used to detect failure.

Secure prefix registrations and redirections are accommodated through the mechanisms of SEAL. Tunnel endpoints using SEAL synchronize sequence numbers, and can therefore discard any packets they receive that are outside of the current sequence number window. Hence, off-path attacks are defeated. These synchronized tunnel endpoints can therefore exchange prefixes with signed certificates that prove prefix ownership in such a way that DoS vectors that attack crypto calculation overhead are eliminated due to the prevention of off-path attacks.

CEs can move from old RANGER networks and re-inject their PI prefixes into new RANGER networks. This would be accommodated by IRON-RANGER as a site multihoming event while host mobility and true locator-ID separation is accommodated via HIP [RFC5201].

17. Recommendation

As can be seen from the extensive list of proposals above, the group explored a number of possible solutions. Unfortunately, the group did not reach rough consensus on a single best approach. Accordingly, the recommendation has been left to the co-chairs. The remainder of this section describes the rationale and decision of the co-chairs.

As a reminder, the goal of the research group was to develop a recommendation for an approach to a routing and addressing architecture for the Internet. The primary goal of the architecture is to provide improved scalability for the routing subsystem. Specifically, this implies that we should be able to continue to grow the routing subsystem to meet the needs of the Internet without requiring drastic and continuous increases in the amount of state or processing requirements for routers.

17.1. Motivation

There is a general concern that the cost and structure of the routing and addressing architecture as we know it today may become prohibitively expensive with continued growth, with repercussions to the health of the Internet. As such, there is an urgent need to examine and evaluate potential scalability enhancements.

For the long term future of the Internet, it has become apparent that IPv6 is going to play a significant role. It has taken more than a decade, but IPv6 is starting to see some non-trivial amount of deployment. This is in part due to the depletion of IPv4 addresses. It therefore seems apparent that the new architecture must be applicable to IPv6. It may or may not be applicable to IPv4, but not addressing the IPv6 portion of the network would simply lead to recreating the routing scalability problem in the IPv6 domain, because the two share a common routing architecture.

Whatever change we make, we should expect that this is a very long-lived change. The routing architecture of the entire Internet is a loosely coordinated, complex, expensive subsystem, and permanent, pervasive changes to it will require difficult choices during deployment and integration. These cannot be undertaken lightly.

By extension, if we are going to the trouble, pain, and expense of making major architectural changes, it follows that we want to make the best changes possible. We should regard any such changes as permanent and we should therefore aim for long term solutions that place the network in the best possible position for ongoing growth. These changes should be cleanly integrated, first-class citizens within the architecture. That is to say that any new elements that are integrated into the architecture should be fundamental primitives, on par with the other existing legacy primitives in the architecture, that interact naturally and logically when in combination with other elements of the architecture.

Over the history of the Internet, we have been very good about creating temporary, ad-hoc changes, both to the routing architecture and other aspects of the network layer. However, many of these band-aid solutions have come with a significant overhead in terms of long-term maintenance and architectural complexity. This is to be avoided and short-term improvements should eventually be replaced by long-term, permanent solutions.

In the particular instance of the routing and addressing architecture today, we feel that the situation requires that we pursue both short-term improvements and long-term solutions. These are not incompatible because we truly intend for the short-term improvements

to be completely localized and temporary. The short-term improvements are necessary to give us the time necessary to develop, test, and deploy the long-term solution. As the long-term solution is rolled out and gains traction, the short-term improvements should be of less benefit and can subsequently be withdrawn.

17.2. Recommendation to the IETF

The group explored a number of proposed solutions but did not reach consensus on a single best approach. Therefore, in fulfillment of the routing research group's charter, the co-chairs recommend that the IETF pursue work in the following areas:

Evolution [Evolution]

Identifier-Locator Network Protocol (ILNP) [ILNP_Site]

Renumbering [RFC5887]

17.3. Rationale

We selected Evolution because it is a short-term improvement. It can be applied on a per-domain basis, under local administration and has immediate effect. While there is some complexity involved, we feel that this option is constructive for service providers who find the additional complexity to be less painful than upgrading hardware. This improvement can be deployed by domains that feel it necessary, for as long as they feel it is necessary. If this deployment lasts longer than expected, then the implications of that decision are wholly local to the domain.

We recommended ILNP because we find it to be a clean solution for the architecture. It separates location from identity in a clear, straightforward way that is consistent with the remainder of the Internet architecture and makes both first-class citizens. Unlike the many map-and-encap proposals, there are no complications due to tunneling, indirection, or semantics that shift over the lifetime of a packet's delivery.

We recommend further work on automating renumbering because even with ILNP, the ability of a domain to change its locators at minimal cost is fundamentally necessary. No routing architecture will be able to scale without some form of abstraction, and domains that change their point of attachment must fundamentally be prepared to change their locators in line with this abstraction. We recognize that [RFC5887] is not a solution so much as a problem statement, and we are simply recommending that the IETF create effective and convenient mechanisms for site renumbering.

18. Acknowledgments

This document presents a small portion of the overall work product of the Routing Research Group, who have developed all of these architectural approaches and many specific proposals within this solution space.

19. Security Considerations

Space precludes a full treatment of security considerations for all proposals summarized herein. [RFC3552] However, it was a requirement of the research group to provide security that is at least as strong as the existing Internet routing and addressing architecture. Each technical proposal has slightly different security considerations, the details of which are in many of the references cited.

20. Informative References

[CRM] Flinck, H., "Compact routing in locator identifier mapping system", <http://www.tschofenig.priv.at/rrg/CR_mapping_system_0.1.pdf>.

[DNSnBIND] Liu, C. and P. Albitz, "DNS & BIND", 2006, 5th Edition, O'Reilly & Associates, Sebastopol, CA, USA. ISBN 0-596-10057-4.

[EEMDP_Considerations] Sriram, K., Kim, Y., and D. Montgomery, "Enhanced Efficiency of Mapping Distribution Protocols in Scalable Routing and Addressing Architectures", Proceedings of the ICCCN, Zurich, Switzerland, August 2010, <http://www.antd.nist.gov/~ksriram/EEMDP_ICCCN2010.pdf>.

[EEMDP_Presentation] Sriram, K., Gleichmann, P., Kim, Y., and D. Montgomery, "Enhanced Efficiency of Mapping Distribution Protocols in Scalable Routing and Addressing Architectures", Presented at the LISP WG meeting, IETF 78, July 2010. Originally presented at the RRG meeting at IETF 72, <http://www.ietf.org/proceedings/78/slides/lisp-6.pdf>.

[Evolution] Zhang, B. and L. Zhang, "Evolution Towards Global Routing Scalability", Work in Progress, October 2009.

[Evolution_Grow_Presentation]
Francis, P., Xu, X., Ballani, H., Jen, D., Raszuk, R., and L. Zhang, "Virtual Aggregation (VA)", November 2009, <http://www.ietf.org/proceedings/76/slides/grow-5.pdf>.

[FIBAggregatability]
Zhang, B., Wang, L., Zhao, X., Liu, Y., and L. Zhang, "An Evaluation Study of Router FIB Aggregatability", November 2009, <http://www.ietf.org/proceedings/76/slides/grow-2.pdf>.

[GLI] Menth, M., Hartmann, M., and D. Klein, "Global Locator, Local Locator, and Identifier Split (GLI-Split)", April 2010, <http://www3.informatik.uni-wuerzburg.de/TR/tr470.pdf>.

[ILNP_DNS]
Atkinson, R. and S. Rose, "DNS Resource Records for ILNP", Work in Progress, February 2011.

[ILNP_ICMP]
Atkinson, R., "ICMP Locator Update message", Work in Progress, February 2011.

[ILNP_Intro]
Atkinson, R., "ILNP Concept of Operations", Work in Progress, February 2011.

[ILNP_Nonce]
Atkinson, R., "ILNP Nonce Destination Option", Work in Progress, February 2011.

[ILNP_Site]
Atkinson, R., Bhatti, S., Hailes, S., Rehunathan, D., and M. Lad, "ILNP - Identifier-Locator Network Protocol", updated 06 January 2011, <http://ilnp.cs.st-andrews.ac.uk>.

[IRON] Templin, F., "The Internet Routing Overlay Network (IRON)", Work in Progress, January 2011.

[Ivip_Constraints]
Whittle, R., "List of constraints on a successful scalable routing solution which result from the need for widespread voluntary adoption", April 2009, <http://www.firstpr.com.au/ip/ivip/RRG-2009/constraints/>.

[Ivip_DRTM]
Whittle, R., "DRTM - Distributed Real Time Mapping for Ivip and LISP", Work in Progress, March 2010.

[Ivip_EAF]
Whittle, R., "Ivip4 ETR Address Forwarding", Work in Progress, January 2010.

[Ivip_Glossary]
Whittle, R., "Glossary of some Ivip and scalable routing terms", Work in Progress, March 2010.

[Ivip_Mobility]
Whittle, R., "TTR Mobility Extensions for Core-Edge Separation Solutions to the Internet's Routing Scaling Problem", August 2008, <http://www.firstpr.com.au/ip/ivip/TTR-Mobility.pdf>.

[Ivip_PLF]
Whittle, R., "Prefix Label Forwarding (PLF) - Modified Header Forwarding for IPv6", <http://www.firstpr.com.au/ip/ivip/PLF-for-IPv6/>.

[Ivip_PMTUD]
Whittle, R., "IPTM - Ivip's approach to solving the problems with encapsulation overhead, MTU, fragmentation and Path MTU Discovery", January 2010, <http://www.firstpr.com.au/ip/ivip/pmtud-frag/>.

[JSAC_Arch]
Atkinson, R., Bhatti, S., and S. Hailes, "Evolving the Internet Architecture Through Naming", IEEE Journal on Selected Areas in Communication (JSAC) 28(8), October 2010.

[LIG] Farinacci, D. and D. Meyer, "LISP Internet Groper (LIG)", Work in Progress, February 2010.

[LISP] Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "Locator/ID Separation Protocol (LISP)", Work in Progress, October 2010.

[LISP+ALT]
Fuller, V., Farinacci, D., Meyer, D., and D. Lewis, "LISP Alternative Topology (LISP+ALT)", Work in Progress, October 2010.

[LISP-Interworking]
Lewis, D., Meyer, D., Farinacci, D., and V. Fuller, "Interworking LISP with IPv4 and IPv6", Work in Progress, August 2010.

[LISP-MN] Meyer, D., Lewis, D., and D. Farinacci, "LISP Mobile Node", Work in Progress, October 2010.

[LISP-MS] Fuller, V. and D. Farinacci, "LISP Map Server", Work in Progress, October 2010.

[LISP-TREE]
Jakab, L., Cabellos-Aparicio, A., Coras, F., Saucez, D., and O. Bonaventure, "LISP-TREE: A DNS Hierarchy to Support the LISP Mapping System", IEEE Journal on Selected Areas in Communications, Volume 28, Issue 8, October 2010, <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5586446>.

[LMS] Letong, S., Xia, Y., ZhiLiang, W., and W. Jianping, "A Layered Mapping System For Scalable Routing", <http://docs.google.com/fileview?id=0BwsJc7A4NTgeOTYzMjFlOGEtYzA4OC00NTM0LTg5ZjktNmFkYzBhNWJhMWEy&hl=en>.

[LMS_Summary]
Sun, C., "A Layered Mapping System (Summary)", <http://docs.google.com/Doc?docid=0AQsJc7A4NTgeZGM3Y3o1NzVfNmd3eGRzNGhi&hl=en>.

[LOC_ID_Implications]
Meyer, D. and D. Lewis, "Architectural Implications of Locator/ID Separation", Work in Progress, January 2009.

[MILCOM1] Atkinson, R. and S. Bhatti, "Site-Controlled Secure Multi-homing and Traffic Engineering for IP", IEEE Military Communications Conference (MILCOM) 28, Boston, MA, USA, October 2009.

[MILCOM2] Atkinson, R., Bhatti, S., and S. Hailes, "Harmonised Resilience, Multi-homing and Mobility Capability for IP", IEEE Military Communications Conference (MILCOM) 27, San Diego, CA, USA, November 2008.

[MPTCP_Arch]
Ford, A., Raiciu, C., Barre, S., Iyengar, J., and B. Ford, "Architectural Guidelines for Multipath TCP Development", Work in Progress, February 2010.

[MobiArch1] Atkinson, R., Bhatti, S., and S. Hailes, "Mobility as an Integrated Service through the Use of Naming", ACM International Workshop on Mobility in the Evolving Internet (MobiArch) 2, Kyoto, Japan, August 2007.

[MobiArch2] Atkinson, R., Bhatti, S., and S. Hailes, "Mobility Through Naming: Impact on DNS", ACM International Workshop on Mobility in the Evolving Internet (MobiArch) 3, Seattle, USA, August 2008.

[Name_Based_Sockets] Vogt, C., "Simplifying Internet Applications Development With A Name-Based Sockets Interface", December 2009, <http://christianvogt.mailup.net/pub/vogt-2009-name-based-sockets.pdf>.

[RANGER_Scen] Russert, S., Fleischman, E., and F. Templin, "RANGER Scenarios", Work in Progress, July 2010.

[RANGI] Xu, X., "Routing Architecture for the Next Generation Internet (RANGI)", Work in Progress, August 2010.

[RANGI-PROXY] Xu, X., "Transition Mechanisms for Routing Architecture for the Next Generation Internet (RANGI)", Work in Progress, July 2009.

[RANGI-SLIDES] Xu, X., "Routing Architecture for the Next-Generation Internet (RANGI)", <http://www.ietf.org/proceedings/76/slides/RRG-1/RRG-1.htm>.

[RFC0792] Postel, J., "Internet Control Message Protocol", STD 5, RFC 792, September 1981.

[RFC3007] Wellington, B., "Secure Domain Name System (DNS) Dynamic Update", RFC 3007, November 2000.

[RFC3552] Rescorla, E. and B. Korver, "Guidelines for Writing RFC Text on Security Considerations", BCP 72, RFC 3552, July 2003.

[RFC4033] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "DNS Security Introduction and Requirements", RFC 4033, March 2005.

[RFC4034] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "Resource Records for the DNS Security Extensions", RFC 4034, March 2005.

[RFC4035] Arends, R., Austein, R., Larson, M., Massey, D., and S. Rose, "Protocol Modifications for the DNS Security Extensions", RFC 4035, March 2005.

[RFC4423] Moskowitz, R. and P. Nikander, "Host Identity Protocol (HIP) Architecture", RFC 4423, May 2006.

[RFC4443] Conta, A., Deering, S., and M. Gupta, "Internet Control Message Protocol (ICMPv6) for the Internet Protocol Version 6 (IPv6) Specification", RFC 4443, March 2006.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, September 2007.

[RFC4960] Stewart, R., "Stream Control Transmission Protocol", RFC 4960, September 2007.

[RFC5201] Moskowitz, R., Nikander, P., Jokela, P., and T. Henderson, "Host Identity Protocol", RFC 5201, April 2008.

[RFC5214] Templin, F., Gleeson, T., and D. Thaler, "Intra-Site Automatic Tunnel Addressing Protocol (ISATAP)", RFC 5214, March 2008.

[RFC5534] Arkko, J. and I. van Beijnum, "Failure Detection and Locator Pair Exploration Protocol for IPv6 Multihoming", RFC 5534, June 2009.

[RFC5720] Templin, F., "Routing and Addressing in Networks with Global Enterprise Recursion (RANGER)", RFC 5720, February 2010.

[RFC5887] Carpenter, B., Atkinson, R., and H. Flinck, "Renumbering Still Needs Work", RFC 5887, May 2010.

[RFC5902] Thaler, D., Zhang, L., and G. Lebovitz, "IAB Thoughts on IPv6 Network Address Translation", RFC 5902, July 2010.

[RRG_Design_Goals] Li, T., "Design Goals for Scalable Internet Routing", Work in Progress, January 2011.

[Referral_Obj]
Carpenter, B., Boucadair, M., Halpern, J., Jiang, S., and K. Moore, "A Generic Referral Object for Internet Entities", Work in Progress, October 2009.

[SEAL] Templin, F., "The Subnetwork Encapsulation and Adaptation Layer (SEAL)", Work in Progress, January 2011.

[Scalability_PS]
Narten, T., "On the Scalability of Internet Routing", Work in Progress, February 2010.

[TIDR] Adan, J., "Tunneled Inter-domain Routing (TIDR)", Work in Progress, December 2006.

[TIDR_AS_forwarding]
Adan, J., "yetAnotherProposal: AS-number forwarding", March 2008, <http://www.ops.ietf.org/lists/rrg/2008/msg00716.html>.

[TIDR_and_LISP]
Adan, J., "LISP etc architecture", December 2007, <http://www.ops.ietf.org/lists/rrg/2007/msg00902.html>.

[TIDR_identifiers]
Adan, J., "TIDR using the IDENTIFIERS attribute", April 2007, <http://www.ietf.org/mail-archive/web/ram/current/msg01308.html>.

[VET] Templin, F., "Virtual Enterprise Traversal (VET)", Work in Progress, January 2011.

[Valiant] Zhang-Shen, R. and N. McKeown, "Designing a Predictable Internet Backbone Network", November 2004, <http://conferences.sigcomm.org/hotnets/2004/HotNets-III%20Proceedings/zhang-shen.pdf>.

[hIPv4] Frejborg, P., "Hierarchical IPv4 Framework", Work in Progress, October 2010.

Author's Address

Tony Li (editor)
Cisco Systems
170 West Tasman Dr.
San Jose, CA 95134
USA

Phone: +1 408 853 9317
EMail: tony.li@tony.li