Independent Submission
Request for Comments: 6837
Category: Experimental
ISSN: 2070-1721

E. Lear
Cisco Systems GmbH

January 2013

NERD:
A Not-so-novel Endpoint ID (EID) to Routing Locator (RLOC) Database

Abstract

The Locator/ID Separation Protocol (LISP) is a protocol to encapsulate IP packets in order to allow end sites to route to one another without injecting routes from one end of the Internet to another. This memo presents an experimental database and a discussion of methods to transport the mapping of Endpoint IDs (EIDs) to Routing Locators (RLOCs) to routers in a reliable, scalable, and secure manner. Our analysis concludes that transport of all EID-to-RLOC mappings scales well to at least 10^8 entries.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for examination, experimental implementation, and evaluation.

This document defines an Experimental Protocol for the Internet community. This is a contribution to the RFC Series, independently of any other RFC stream. The RFC Editor has chosen to publish this document at its discretion and makes no statement about its value for implementation or deployment. Documents approved for publication by the RFC Editor 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/rfc6837.

Table of Contents

1. Introduction

The Locator/ID Separation Protocol (LISP) [RFC6830] separates an IP address used by a host and local routing system from the Locators advertised by BGP participants on the Internet in general, and in the Default-Free Zone (DFZ) in particular. It accomplishes this by establishing a mapping between globally unique Endpoint IDs (EIDs) and Routing Locators (RLOCs). This reduces the amount of state change that occurs on routers within the DFZ on the Internet, while enabling end sites to be multihomed.

In some mapping distribution approaches to LISP, the mapping is learned via data-triggered control messages between Ingress Tunnel Routers (ITRs) and Egress Tunnel Routers (ETRs) through an alternate routing topology [RFC6836]. In other approaches of LISP, the mapping from EIDs to RLOCs is instead learned through some other means. This memo addresses different approaches to the problem, and specifies a Not-so-novel EID-to-RLOC Database (NERD) and methods to both receive the database and to receive updates.

NERD is offered primarily as a way to avoid dropping packets, the underlying assumption being that dropping packets is bad for applications and end users. Those who do not agree with this underlying assumption may find that other approaches make more sense.

NERD is specified in such a way that the methods used to distribute or retrieve it may vary over time. Multiple databases are supported in order to allow for multiple data sources. An effort has been made to divorce the database from access methods so that both can evolve independently through experimentation and operational validation.

1.1. Applicability

This memo is based on experiments performed in the 2007-2009 time frame. At the time of its publication, the author is unaware of operational use of NERD. Those wishing to pursue NERD should consider the substantial amount of work left for the future. See Section 10 for more details.

1.2. Base Assumptions

In order to specify a mapping, it is important to understand how it will be used, and the nature of the data being mapped. In the case of LISP, the following assumptions are pertinent:

- The data contained within the mapping changes only on provisioning or configuration operations, and is not intended to change when a link either fails or is restored. Some other mechanism, such as

the use of LISP Reachability Bits with mapping replies, handles healing operations, particularly when a tail circuit within a service provider's aggregate goes down. NERD can be used as a verification method to ensure that whatever operational mapping changes an ITR receives are authorized.

- While weight and priority are defined, these are not hop-by-hop metrics. Hence, the information contained within the mapping does not change based on where one sits within the topology.

- Because a purpose of LISP is to reduce control-plane overhead by reducing "rate X state" complexity, updates to the mapping will be relatively rare.

- Because NERD is designed to ease interdomain routing, its use is intended within the inter-domain environment. That is, NERD is best implemented at either the customer edge or provider edge, and there will be on the order of as many ITRs and EID-Prefixes as there are connections to Internet service providers by end customers.

- As such, NERD cannot be the sole means to implement host mobility, although NERD may be in used in conjunction with other mechanisms.

1.3. What is NERD?

NERD is a Not-so-novel EID-to-RLOC Database. It consists of the following components:

1. a network database format;

2. a change distribution format;

3. a database retrieval/bootstrapping method; and

4. a change distribution method.

The network database format is compressible. However, at this time, we specify no compression method. NERD will make use of potentially several transport methods, but most notably HTTP [RFC2616]. HTTP has restart and compression capabilities. It is also widely deployed.

There exist many methods to show differences between two versions of a database or a file, UNIX's "diff" being the classic example. In this case, because the data is well structured and easily keyed, we can make use of a very simple format for version differences that

simply provides a list of EID-to-RLOC mappings that have changed using the same record format as the database, and a list of EIDs that are to be removed.

1.4. Glossary

The reader is once again referred to [RFC6830] for a general glossary of terms related to LISP. The following terms are specific to this memo.

Base Distribution URI: An Absolute-URI as defined in Section 4.3 of [RFC3986] from which other references are relative. The base distribution URI is used to construct a URI to an EID-to-RLOC mapping database. If more than one NERD is known, then there will be one or more base distribution URIs associated with each (although each such base distribution URI may have the same value).

EID Database Authority: The authority that will sign database files and updates. It is the source of both.

The Authority: Shorthand for the EID Database Authority.

NERD: Not-so-novel EID-to-RLOC Database.

AFI Address Family Identifier.

Pull Model: An architecture where clients pull only the information they need at any given time, such as when a packet arrives for forwarding.

Push Model: An architecture in which clients receive an entire dataset, containing data they may or may not require, such as mappings for EIDs that no host served is attempting to send to.

Hybrid Model: An architecture in which some information is pushed toward the receiver from a source and some information is pulled by the receiver.

2. Theory of Operation

Operational functions are split into two components: database updates and state exchange between ITR and ETR during a communication.

2.1. Database Updates

What follows is a summary of how NERDs are generated and updated. Specifics can be found in Section 3. The general way in which NERD works is as follows:

1. A NERD is generated by an authority that allocates Provider-Independent (PI) addresses (e.g., IANA or a Regional Internet Registry (RIR)) that are used by sites as EIDs. As part of this process, the authority generates a digest for the database and signs it with a private key whose public key is part of an X.509 certificate. [ITU.X509.2000] That signature along with a copy of the authority's public key is included in the NERD.

2. The NERD is distributed to a group of well-known servers.

3. ITRs retrieve an initial copy of the NERD via HTTP when they come into service.

4. ITRs are preconfigured with a group of certificates whose private keys are used by database authorities to sign the NERD. This list of certificates should be configurable by administrators.

5. ITRs next verify both the validity of the public key and the signed digest. If either fail validation, the ITR attempts to retrieve the NERD from a different source. The process iterates until either a valid database is found or the list of sources is exhausted.

6. Once a valid NERD is retrieved, the ITR installs it into both non-volatile and local memory.

7. At some point, the authority updates the NERD and increments the database version counter. At the same time, it generates a list of changes, which it also signs, as it does with the original database.

8. Periodically, ITRs will poll from their list of servers to determine if a new version of the database exists. When a new version is found, an ITR will attempt to retrieve a change file, using its list of preconfigured servers.

9. The ITR validates a change file just as it does the original database. Assuming the change file passes validation, the ITR installs new entries, overwrites existing ones, and removes empty entries, based on the content of the change file.

As time goes on, it is quite possible that an ITR may probe a list of configured peers for a database or change file copy. It is equally possible that peers might advertise to each other the version number of their database. Such methods are not explored in depth in this memo but are mentioned for future consideration.

2.2. Communications between ITR and ETR

[RFC6830] describes the basic approach to what happens when a packet arrives at an ITR, and what communications between the ITR and ETR take place. NERD provides an optimistic approach to establishing communications with an ETR that is responsible for a given EID-Prefix. State must be kept, however, on an ITR to determine whether that ETR is in fact reachable. It is expected that this is a common requirement across LISP mapping systems, and will be handled in the core LISP architecture.

2.3. Who are database authorities?

This memo does not specify who the database authority is. That is because there are several possible operational models. In each case, the number of database authorities is meant to be small so that ITRs need only keep a small list of authorities, similar to the way a name server might cache a list of root servers.

- A single database authority exists. In this case, all entries in the database are registered to a single entity, and that entity distributes the database. Because the EID space is provider-independent address space, there is no architectural requirement that address space be hierarchically distributed to anyone, as there is with provider-assigned address space. Hence, there is a natural affinity between the IANA function and the database authority function.

- Each region runs a database authority. In this case, provider-independent address space is allocated to either RIRs or to affiliates of such organizations of network operations guilds (NOGs). The benefit of this approach is that there is no single organization that controls the database. It allows one database authority to back up another. One could envision as many as ten database authorities in this scenario. One drawback to this

approach, however, is that any reference to a region imposes a notion of locality, thus potentially diminishing the split between Locator and identifier.

- Each country runs a database authority. This could occur should countries decide to regulate this function. While limiting the scope of any single database authority as the previous scenario describes, this approach would introduce some overhead as the list of database authorities would grow to as many as 200, and possibly more if jurisdictions within countries attempted to regulate the function. There are two drawbacks to this approach. First, as distribution of EIDs is driven to more local jurisdictions, an EID-Prefix is tied even more tightly to a location. Second, a large number of database authorities will demand some sort of discovery mechanism.

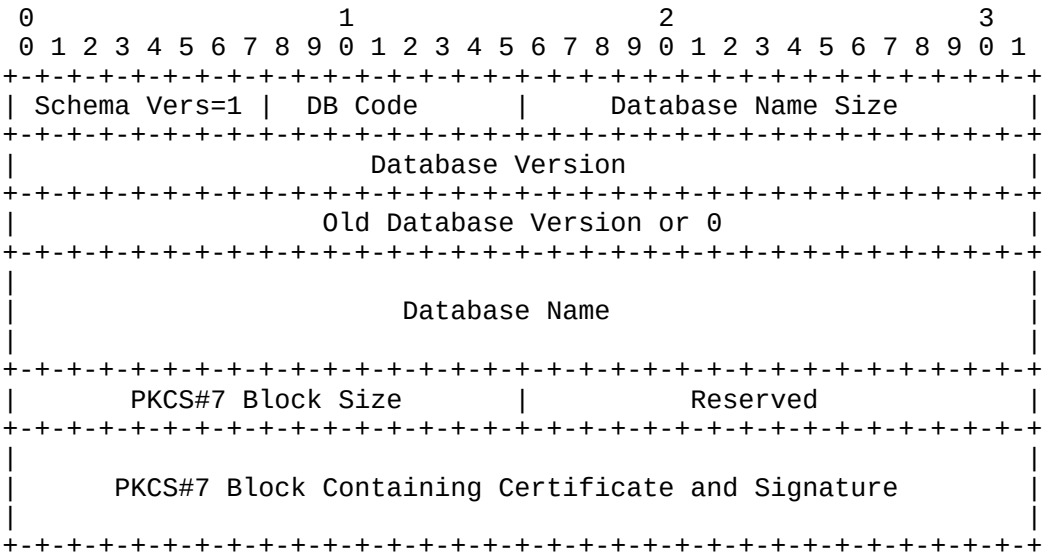
- Independent operators manage database authorities. This has the appeals of being location independent and enabling competition for good performance. This method has the drawback of potentially requiring a discovery mechanism.

The latter two approaches are not mutually exclusive. While this specification allows for multiple databases, discovery mechanisms are left as future work.

3. NERD Format

The NERD consists of a header that contains a database version and a signature that is generated by ignoring the signature field and setting the authentication block length to 0 (NULL). The authentication block itself consists of a signature and a certificate whose private-key counterpart was used to generate the signature.

Records are kept sorted in numeric order with AFI plus EID as primary key and prefix length as secondary. This is so that after a database update it should be possible to reconstruct the database to verify the digest signature, which may be retrieved separately from the database for verification purposes.



Database Header

The 'DB Code' field indicates 0 if what follows is an entire database or 1 if what follows is an update. The 'Database Version' field holds the database file version, which is incremented each time the complete database is generated by the authority. In the case of an update, the field indicates the new database file version, and the old database file version is indicated in the 'Old Database Version' field. The database file version is used by routers to determine whether or not they have the most current database.

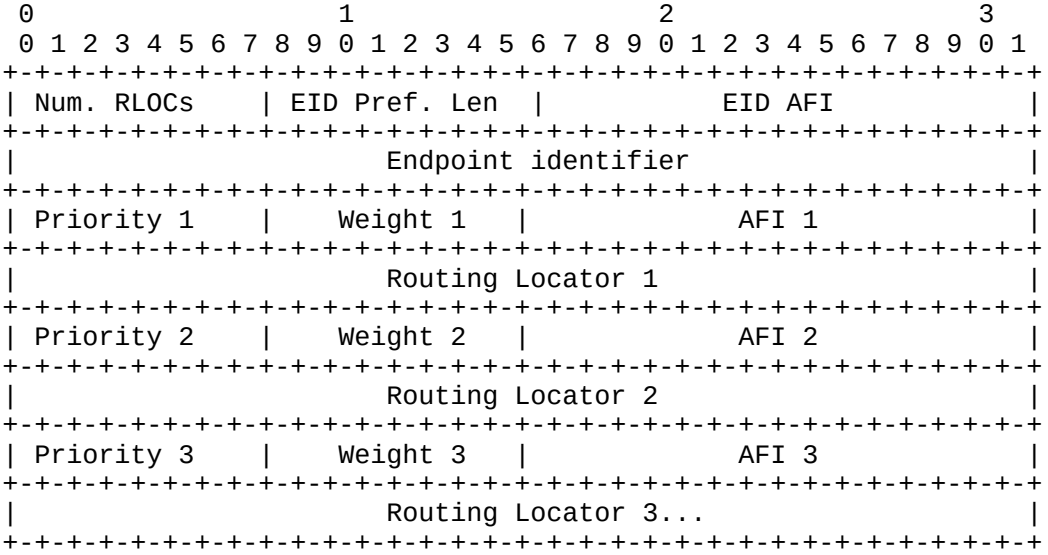
The 'Database Name' field holds a DNS-ID, as specified in [RFC6125]. This is the name that will appear in the Subject field of the certificate used to verify the database. The purpose of the database name is to allow for more than one database. Such databases would be merged by the router. It is important that an EID-to-RLOC mapping be listed in no more than one database, lest inconsistencies arise. However, it may be possible to transition a mapping from one database to another. During the transition period, the mappings would be identical. When they are not, the resultant behavior will be undefined. The database name is padded with NULLs to the nearest fourth byte.

The PKCS#7 [RFC2315] authentication block contains a DER-encoded [ITU.X509.2000] signature and associated public key. For the purposes of this experiment, all implementations will support the RSA encryption signature algorithm and SHA1 digest algorithm, and the standard attributes are expected to be present.

N.B., it has been suggested that the Cryptographic Message Syntax (CMS) [RFC5652] be used instead of PKCS#7. At the time this experiment was performed, CMS was not yet widely deployed. However, it is certainly the correct direction and should be strongly considered in future related work.

3.1. NERD Record Format

As distributed over the network, NERD records appear as follows:



EID AFI is the AFI of the EID. Priority N, Weight N, and AFI N are associated with Routing Locator N. There will always be at least one RLOC. The minimum record size for IPv4 is 16 bytes. Each additional IPv4 RLOC increases the record size by 8 bytes. The purpose of this format is to keep the database compact, but somewhat easily read. The meaning of weight and priority are described in [RFC6830]. The format of the AFI is specified by IANA in the "Address Family Numbers" registry, with the exception of how IPv6 EID-Prefixes are stored.

NERD assumes that EIDs stored in the database are prefixes, and therefore are accompanied with prefix lengths. In order to reduce storage and transmission amounts for IPv6, only the necessary number of bytes of an EID as specified by the prefix length are kept in the record, rounded to the nearest 4-byte (word) boundary. For instance, if the prefix length is /49, the nearest 4-byte word boundary would require that 8 bytes are stored. IPv6 RLOCs are represented as normal 128-bit IPv6 addresses.

3.2. Database Update Format

A database update contains a set of changes to an existing database. Each {AFI, EID, mask-length} tuple may have zero or more RLOCs associated with it. In the case where there are no RLOCs, the EID entry is removed from the database. Records that contain EIDs and prefix lengths that were not previously listed are simply added. Otherwise, the old record for the EID and prefix length is replaced by the more current information. The record format used by the database update is the same as described in Section 3.1.

4. NERD Distribution Mechanism

4.1. Initial Bootstrap

Bootstrap occurs when a router needs to retrieve the entire database. It knows it needs to retrieve the entire database because either it has none or it has an update too substantial to process, as might be the case if a router has been out of service for a substantially lengthy period of time.

To bootstrap, the ITR appends the database name plus "/current/entiredb" to a base distribution URI and retrieves the file via HTTP. More formally (using ABNF from [RFC5234]):

```
entire-db =    base-uri dbname "/current/entiredb"
base-uri  =    uri ; from RFC 3986
dbname    =    DNS-ID ; from RFC 6125
```

For example, if the base distribution URI is "http://www.example.com/eiddb/", and assuming a database name of "nerd.arin.net", the ITR would request "http://www.example.com/eiddb/nerd.arin.net/current/entiredb". Routers check the signature on the database prior to installing it, and they check that the database schema matches a schema they understand. Once a router has a valid database, it stores that database in some sort of non-volatile memory (e.g., disk, flash memory, etc).

N.B., the host component for such URIs should not resolve to a LISP EID, lest a circular dependency be created.

4.2. Retrieving Changes

In order to retrieve a set of database changes, an ITR will have previously retrieved the entire database. Hence, it knows the current version of the database it has. Its first step for retrieving changes is to retrieve the current version number of the

database. It does so by appending "/current/version" to the base distribution URI and database name and retrieving the file. Its format is text, and it contains the integer value of the current database version.

Once an ITR has retrieved the current version, it compares the version of its local copy. If there is no difference, then the router is up to date and need take no further actions until it next checks.

If the versions differ, the router next sends a request for the appropriate change file by appending "current/changes/" and the textual representation of the version of its local copy of the database to the base distribution URI. More formally:

```
db-version    =    base-uri dbname "/current/version"
db-curupdate  =    base-uri dbname "/current/changes/" old-version
old-version   =    1*DIGIT
```

For example, if the current version of the database is 1105503, the router's version is 1105500, and the base URI and database name are the same as above, the router would first request "http://www.example.com/eiddb/nerd.arin.net/current/version" to determine that it is out of date, and also to learn the current version. It would then attempt to retrieve "http://www.example.com/eiddb/nerd.arin.net/current/changes/1105500".

The server may not have that change file, either because there are too many versions between what the router has and what is current or because no such change file was generated. If the server has changes from the router's version to any later version, the server issues an HTTP redirect to that change file, and the router retrieves and processes it. More formally:

```
db-incupdate    =    base-uri dbname "/" newer-version
                     "/changes/" old-version
newer-version   =    1*DIGIT
```

For example:

"http://www.example.com/eiddb/nerd.arin.net/1105450/changes/1105401" would update a router from version 1105401 to 1105450. Once it has done so, the router should then repeat the process until it has brought itself up to date.

This begs the question: how does a router know to retrieve version 1105450 in our example above? It cannot. A redirect must be given by the server to that URI when the router attempts to retrieve differences from the current version, say, 1105503.

While it is unlikely that database versions would wrap, as they consist of 32-bit integers, should the event occur, ITRs should attempt first to retrieve a change file when their current version number is within 10,000 of 2^32 and they see a version available that is less than 10,000. Barring the availability of a change file, the ITR can still assume that the database version has wrapped and retrieve a new copy. It may be safer in future work to include additional wrap information or a larger field to avoid having to use any heuristics.

5. Analysis

We will start our analysis by looking at how much data will be transferred to a router during bootstrap conditions. We will then look at the bandwidth required. Next, we will turn our concerns to servers. Finally, we will ponder the effect of providing only changes.

In the analysis below, we treat the overhead of the database header as insignificant (because it is). The analysis should be similar, whether a single database or multiple databases are employed, as we would assume that no entry would appear more than once.

5.1. Database Size

By its very nature, the information to be transported is relatively static and is specifically designed to be topologically insensitive. That is, every ITR is intended to have the same set of RLOCs for a given EID. While some processing power will be necessary to install a table, the amount required should be far less than that of a routing information database because the level of entropy is intended to be lower.

For purposes of this analysis, we will assume that the world has migrated to IPv6, as this increases the size of the database, which would be our primary concern. However, to mitigate the size increase, we have limited the size of the prefix transmitted. For purposes of this analysis, we shall assume an average prefix length of 64 bits.

Based on that assumption, Section 3.1 states that mapping information for each EID/prefix includes a group of RLOCs, each with an associated priority and weight, and that a minimum record size with IPv6 EIDs with at least one RLOC is 30 bytes uncompressed. Each additional IPv6 RLOC costs 20 bytes.

10^n EIDs	2 RLOC	4 RLOC	8 RLOC
4	500 KB	900 KB	1.70 MB
5	5.0 MB	9.0 MB	17.0 MB
6	50 MB	90 MB	170 MB
7	500 MB	900 MB	1.70 GB
8	5.0 GB	9.0 GB	17.0 GB

Table 1: Database size for IPv6 routes with average prefix length of 64 bits

Entries in the above table are derived as follows:

$$E * (30 + 20 * (R - 1))$$

where E = number of EIDs (10^n), R = number of RLOCs per EID.

Our scaling target is to accommodate 10^8 multihomed systems, which is one order of magnitude greater than what is discussed in [CARP07]. At 10^8 entries, a device could be expected to use between 5 and 17 GB of RAM for the mapping. No matter the method of distribution, any router that sits in the core of the Internet would require near this amount of memory in order to perform the ITR function. Large-enterprise ETRs would be similarly strained, simply due to the diversity of sites that communicate with one another. The good news is that this is not our starting point, but rather our scaling target, a number that we intend to reach by the year 2050. Our starting point is more likely in the neighborhood of 10^4 or 10^5 EIDs, thus requiring between 500 KB and 17 MB.

5.2. Router Throughput versus Time

Table Size (10^n)	1 MB/s	10 MB/s	100 MB/s	1 GB/s
6	8	0.8	0.08	0.008
7	80	8	0.8	0.08
8	800	80	8	0.8
9	8,000	800	80	8
10	80,000	8,000	800	80
11	800,000	80,000	8,000	800

Table 2: Number of seconds to process NERD

The length of time it takes to receive the database is significant in models where the device acquires the entire table. During this period of time, either the router will be unable to route packets using LISP or it must use some sort of query mechanism for specific EIDs as it populates the rest of its table through the transfer. Table 2 shows us that at our scaling target, the length of time it would take for a router using 1 MB/s of bandwidth is about 80 seconds. We can measure the processing rate in small numbers of hours for any transfer speed greater than that. The fastest processing time shows us as taking 8 seconds to process an entire table of 10^9 bytes and 80 seconds for 10^10 bytes.

5.3. Number of Servers Required

As easy as it may be for a router to retrieve, the aggregate information may be difficult for servers to transmit, assuming the information is transmitted in aggregate (we'll revisit that assumption later).

# Simultaneous Requests	10 Servers	100 Servers	1,000 Servers	10,000 Servers
100	720	72	72	72
1,000	7,200	720	72	72
10,000	72,000	7,200	720	72
100,000	720,000	72,000	7,200	720
1,000,000	7,200,000	720,000	72,000	7,200
10,000,000	72,000,000	7,200,000	720,000	72,000

Table 3: Retrieval time per number of servers in seconds

This assumes an average of 10^8 entries with 4 RLOCs per EID and that each server has access to 1 GB/s, 100% efficient use of that bandwidth, and no compression.

Entries in the above table were generated using the following method:

For 10^8 entries with four RLOCs per EID, the table size is 9.0 GB, per our previous table. Assume 1 GB/s transfer rates and 100% utilization. Protocol overhead is ignored for this exercise. Hence, a single transfer X takes 48 seconds and can get no faster.

With this in mind, each entry is as follows:

```
         max(1X,N*X/S)

  where N = number of transfers,
        X = 72 seconds, and
        S = number of servers.
```

If we have a distribution model in which every device must retrieve the mapping information upon start, Table 3 shows the length of time in seconds it will take for a given number of servers to complete a transfer to a given number of devices. This table says, as an example, that it would take 72,000 seconds (20 hours) for 1,000,000 ITRs to simultaneously retrieve the database from 1,000 servers, assuming equal load distribution. Should a cold-start scenario occur, this number should be of some concern. Hence, it is important to take some measures both to avoid such a scenario and to ease the load should it occur. The primary defense should be for ITRs to first attempt to retrieve their databases from their peers or upstream providers. Secondary defenses could include data sanity checks within ITRs, with agreed norms for how much the database should change in any given update or over any given period of time. As we will see below, dissemination of changes is considerably less volume.

% Daily Change	100 Servers	1,000 Servers	10,000 Servers
0.1%	300	30	3
0.5%	1,500	150	15
1%	3,000	300	30
5%	15,000	1,500	150
10%	30,000	3,000	300

Table 4: Transfer times for hourly updates, shown in seconds

Assuming 10 million routers and a database size of 9 GB, resulting transfer times for hourly updates are shown in seconds, given number of servers and daily rate of change. Note that when insufficient resources are devoted to servers, an unsustainable situation arises where updates for the next batch would begin prior to the completion of the current batch.

This table shows us that with 10,000 servers the average transfer time with 1 GB/s links for 10,000,000 routers will be 300 seconds with 10% daily change spread over 24 hourly updates. For a 0.1% daily change, that number is 3 seconds for a database of size 9.0 GB.

The amount of change goes to the purpose of LISP. If its purpose is to provide effective multihoming support to end customers, then we might anticipate relatively few changes. If, on the other hand, service providers attempt to make use of LISP to provide some form of traffic engineering, we can expect the same data to change more often. We cannot conclude much in this regard without additional operational experience. The one thing we can say is that different applications of LISP may require new and different distribution mechanisms. Such optimization is left for another day.

5.4. Security Considerations

If an attacker can forge an update or tamper with the database, he can in effect redirect traffic to end sites. Hence, integrity and authenticity of the NERD is critical. In addition, a means is required to determine whether a source is authorized to modify a given database. No data privacy is required. Quite to the contrary, this information will be necessary for any ITR.

The first question one must ask is who to trust to provide the ITR a mapping. Ultimately, the owner of the EID-Prefix is most authoritative for the mapping to RLOCs. However, were all owners to sign all such mappings, ITRs would need to know which owner is authorized to modify which mapping, creating a problem of O(N^2) complexity.

We can reduce this problem substantially by investing some trust in a small number of entities that are allowed to sign entries. If an authority manages EIDs much the same way a domain name registrar handles domains, then the owner of the EID would choose a database authority she or he trusts, and ITRs must trust each such authority in order to map the EIDs listed by that authority to RLOCs. This reduces the amount of management complexity on the ETR to retaining knowledge of O(# authorities), but does require that each authority establish procedures for authenticating the owner of an EID. Those procedures needn’t be the same.

There are two classic methods to ensure integrity of data:

- secure transport of the source of the data to the consumer, such as Transport Layer Security (TLS) [RFC5246]; and
- provide object-level security.

These methods are not mutually exclusive, although one can argue about the need for the former, given the latter.

In the case of TLS, when it is properly implemented, the objects being transported cannot easily be modified by interlopers or so-called men in the middle. When data objects are distributed to multiple servers, each of those servers must be trusted. As we have seen above, we could have quite a large number of servers, thus providing an attacker a large number of targets. We conclude that some form of object-level security is required.

Object-level security involves an authority signing an object in a way that can easily be verified by a consumer, e.g., a router. In this case, we would want the mapping table and any incremental update to be signed by the originator of the update. This implies that we cannot simply make use of a tool like CVS [CVS]. Instead, the originator will want to generate diffs, sign them, and make them available either directly or through some sort of content distribution or peer to peer network.

5.4.1. Use of Public Key Infrastructures (PKIs)

X.509 provides a certificate hierarchy that has scaled to the size of the Internet. The system is most manageable when there are few certificates to manage. The model proposed in this memo makes use of one current certificate per database authority. The two pieces of information necessary to verify a signature, therefore, are as follows:

- the certificate of the database authority, which can be provided along with the database; and
- the certificate authority’s certificate.

The latter two pieces of information must be very well known and must be configured on each ITR. It is expected that both would change very rarely, and it would not be unreasonable for such updates to occur as part of a normal OS release process.

The tools for both signing and verifying are readily available. OpenSSL (http://www.openssl.org) provides tools and libraries for both signing and verifying. Other tools commonly exist.

Use of PKIs is not without implementation complexity, operational complexity, or risk. The following risks and mitigations are identified with NERD's use of PKIs:

The private key of a NERD authority is exposed:
In this case, an attacker could sign a false database update, either redirecting traffic or otherwise causing havoc. The NERD administrator must revoke its existing key and issue a new one. The certificate is added to a certificate revocation list (CRL), which may be distributed with both this and other databases, as well as through other channels. Because this event is expected to be rare, and the number of database authorities is expected to be small, a CRL will be small. When a router receives a revocation, it checks it against its existing databases, and attempts to update the one that is revoked. This implies that prior to issuing the revocation, the database authority would sign an update with the new key. Routers would discard updates they have already received that were signed after the revocation was generated. If a router cannot confirm whether the authority's certificate was revoked before or after a particular update, it will retrieve a fresh new copy of the database with a valid signature.

The private key associated with a CA in the chain of trust of the Authority's certificate is compromised:
In this case, it becomes possible for an attacker to masquerade as the database authority. To ameliorate damage, the database authority revokes its certificate and get a new certificate issued from a CA that is not compromised. Once it has done so, the previous procedure is followed. The compromised certificate can be removed during the normal OS upgrade cycle. In the case of the root authority, the situation could be more serious. Updates to the OS in the ITR need to be validated prior to installation. One possible method of doing this is provided in [RFC4108]. Trust anchors are assumed to be updated as part of an OS update; implementors should consider using a key other than the trust anchor for validating OS updates.

An algorithm used if either the certificate or the signature is cracked:
This is a catastrophic failure and the above forms of attack become possible. The only mitigation is to make use of a new algorithm. In theory, this should be possible, but in practice it has proved very difficult. For this reason, additional work is recommended to make alternative algorithms available.

The NERD authority loses its key or disappears:
In this case, nobody can update the existing database. There are few programmatic mitigations. If the database authority places its private keys and suitable amounts of information in escrow, under agreed upon circumstances (for example, no updates for three days), the escrow agent would release the information to a party competent of generating a database update.

5.4.2. Other Risks

Because this specification does not require secure transport, if an attacker prevents updates to an ITR for the purposes of having that ITR continue to use a compromised ETR, the ITR could continue to use an old version of the database without realizing a new version has been made available. If one is worried about such an attack, a secure channel (such as SSL) to a secure chain back to the database authority should be used. It is possible that, after some operational experience, later versions of this format will contain additional semantics to address this attack. SSL would also prevent attempts to spoof false database versions on the server.

As discussed above, substantial risk would be a cold-start scenario. If an attacker found a bug in a common OS that allowed it to erase an ITR’s database, and was able to disseminate that bug, the collective ability of ITRs to retrieve new copies of the database could be taxed by collective demand. The remedy to this is for devices to share copies of the database with their peers, thus making each potential requester a potential service.

6. Why not use XML?

Many objects these days are distributed as either XML pages or something derived as XML [W3C.REC-xml11-20040204], such as SOAP [W3C.REC-soap12-part1-20070427] [W3C.REC-soap12-part2-20070427]. Use of such well-known standards allows for high-level tools and library reuse. XML’s strength is extensibility. Without a doubt XML would be more extensible than a fixed field database. Why not, then, use these standards in this case? The greatest concern the author had was compactness of the data stream. In as much as this mechanism is

used at all in the future, so long as that concern could be addressed, and so long as signatures of the database can be verified, XML probably should be considered.

7. Other Distribution Mechanisms

We now consider various different mechanisms. The problem of distributing changes in various databases is as old as databases. The author is aware of two obvious approaches that have been well used in the past. One approach would be the wide distribution of CVS repositories. However, for reasons mentioned in Section 5.4, CVS is insufficient to the task.

The other tried and true approach is the use of periodic updates in the form of messages. The good old Network News Transfer Protocol (NNTP) [RFC3977] itself provides two separate mechanisms (one push and another pull) to provide a coherent update process. This was in fact used to update molecular biology databases [gb91] in the early 1990s. Netnews offers a way to determine whether articles with specified Article-Ids have been received. In the case where the mapping file source of authority wishes to transmit updates, it can sign a change file and then post it into the network. Routers merely need to keep a record of article ids that it has received. Netnews systems have years ago handled far greater volume of traffic than we envision [Usenet]. Initially this is probably overkill, but it may not be so later in this process. Some consideration should be given to a mechanism known to widely distribute vast amounts of data, as instantaneously as either the sender or the receiver wishes.

To attain an additional level of hierarchy in the distribution network, service providers could retrieve information to their own local servers and configure their routers with the host portion of the above URI.

Another possibility would be for providers to establish an agreement on a small set of anycast addresses for use for this purpose. There are limitations to the use of anycast, particularly with TCP. In the midst of a routing flap, an anycast address can become all but unusable. Careful study of such a use as well as appropriate use of HTTP redirects is expected.

7.1. What about DNS as a mapping retrieval model?

It has been proposed that a query/response mechanism be used for this information and specifically that the Domain Name System (DNS) [RFC1034] be used. The previous models do not preclude DNS. DNS has the advantage that the administrative lines are well drawn, and that the ID-to-RLOC mapping is likely to appear very close to these

boundaries. DNS also has the added benefit that an entire distribution infrastructure already exists. There are, however, some problems that could impact end hosts when intermediate routers make queries, some of which were first pointed out in [RFC1383]:

- Any query mechanism offers an opportunity for a resource attack if an attacker can force the ITR to query for information. In this case, all that would be necessary would be for a "botnet" (a group of computers that have been compromised and used as vehicles to attack others) to ping or otherwise contact via some normal service hosts that sit behind the ETR. If the botnet hosts themselves are behind ETRs, the victim's ITR will need to query for each and every one of them, thus becoming part of a classic reflector attack.

- Packets will be delayed at the very least, and probably dropped in the process of a mapping query. This could be at the beginning of a communication, but it will be impossible for a router to conclude with certainty that this is the case.

- The DNS has a backoff algorithm that presumes that applications are making queries prior to the beginning of a communication. This is appropriate for end hosts who know in fact when a communication begins. An end user may not enjoy that a router is waiting seconds for a retry.

- While the administrative lines may appear to be correct, the location of name servers may not be. If name servers sit within PI address space, thus requiring LISP to reach, a circular dependency is created. This is precisely where many enterprise name servers sit. The LISP experiment should not predicate its success on relocation of such name servers.

Nevertheless, DNS may be able to play a role in providing the enterprise control over the mapping of its EIDs to RLOCs. Posit a new DNS record "EID2RLOC". This record is used by the authority to collect and aggregate mapping information so that it may be distributed through one of the other mechanisms. As an example:

```
$ORIGIN 0.10.PI-SPACE.
 128   EID2RLOC   mask 23 priority 10 weight 5 172.16.5.60
       EID2RLOC   mask 23 priority 15 weight 5 192.168.1.5
```

In the above figure, network 10.0.128/23 would delegated to some end system, say, EXAMPLE.COM. They would manage the above zone information. This would allow a DNS mechanism to work, but it would also allow someone to aggregate the information and distribute a table.

7.2. Use of BGP and LISP+ALT

The Border Gateway Protocol (BGP) [RFC4271] is currently used to distribute inter-domain routing throughout the Internet. Why not, then, use BGP to distribute mapping entries, or provide a rendezvous mechanism to initialize mapping entries? In fact, this is precisely what LISP Alternative Topology (LISP+ALT) [RFC6836] accomplishes, using a completely separate topology from the normal DFZ. It does so using existing code paths and expertise. The alternative topology also provides an extremely accurate control path from ITRs to ETRs, whereas NERD's operational model requires an optimistic assumption and control-plane functionality to cycle through unresponsive ETRs in an EID-Prefix's mapping entry. The memory-scaling characteristics of LISP+ALT are extremely attractive because of expected strong aggregation, whereas NERD makes almost no attempt at aggregation.

A number of key deployment issues are left open. The principle issue is whether it is deemed acceptable for routers to drop packets occasionally while mapping information is being gathered. This should be the subject of future research for ALT, as it was a key design goal of NERD to avoid such a situation.

7.3. Perhaps use a hybrid model?

Perhaps it would be useful to use both a prepopulated database such as NERD and a query mechanism (perhaps LISP+ALT, LISP-CONS [LISP-CONS], or DNS) to determine an EID-to-RLOC mapping. One idea would be to receive a subset of the mappings, say, by taking only the NERD for certain regions. This alleviates the need to drop packets for some subset of destinations under the assumption that one's business is localized to a particular region. If one did not have a local entry for a particular EID, one would then make a query.

One approach to using DNS to query live would be to periodically walk "interesting" portions of the network, in search of relevant records, and to cache them to non-volatile storage. While preventing resource attacks, the walk itself could be viewed as an attack, if the algorithm was not selective enough about what it thought was interesting. A similar approach could be applied to LISP+ALT or LISP-CONS by forcing a data-driven Map Reply for certain sites.

8. Deployment Issues

While LISP and NERD are intended as experiments at this point, it is already obvious one must give serious consideration to circular dependencies with regard to the protocols used and the elements within them.

8.1. HTTP

In as much as HTTP depends on DNS, either due to the authority section of a URI or to the configured base distribution URI, these same concerns apply. In addition, any HTTP server that itself makes use of Provider-Independent addresses would be a poor choice to distribute the database for these exact same reasons.

One issue with using HTTP is that it is possible that a middlebox of some form, such as a cache, may intercept and process requests. In some cases, this might be a good thing. For instance, if a cache correctly returns a database, some amount of bandwidth is conserved. On the other hand, if the cache itself fails to function properly for whatever reason, end-to-end connectivity could be impaired. For example, if the cache itself depended on the mapping being in place and functional, a cold-start scenario might leave the cache functioning improperly, in turn providing routers no means to update their databases. Some care must be given to avoid such circumstances.

9. Open Questions

Do we need to discuss reachability in more detail? This was clearly an issue at the IST-RING (Information Science Technologies - Routing in Next Generation) workshop. There are two key issues. First, what is the appropriate architectural separation between the data plane and the control plane? Second, is there some specific way in which NERD impacts the data plane?

Should we specify a (perhaps compressed) tarball that treads a middle ground for the last question, where each update tarball contains both a signature for the update and for the entire database, once the update is applied?

Should we compress? In some initial testing of databases with 1, 5, and 10 million IPv4 EIDs and a random distribution of IPv4 RLOCs, the current format in this document compresses down by a factor of between 35% and 36%, using Burrows-Wheeler block sorting text compression algorithm (bzip2). The NERD used random EIDs with prefix lengths varying from 19-29 bits, with probability weighted toward the smaller masks. This only very roughly reflects reality. A better test would be to start with the existing prefixes found in the DFZ.

10. Conclusions

This memo has specified a database format, an update format, a URI convention, an update method, and a validation method for EID-to-RLOC mappings. We have shown that beyond the predictions of 10^8 EID-prefix entries, the aggregate database size would likely be at most 17 GB. We have considered the amount of servers to distribute that information, and we have demonstrated the limitations of a simple content distribution network and other well-known mechanisms. The effort required to retrieve a database change amounts to between 3 and 30 seconds of processing time per hour at today's gigabit speeds. We conclude that there is no need for an off-box query mechanism today and that there are distinct disadvantages for having such a mechanism in the control plane.

Beyond this, we have examined alternatives that allow for hybrid models that do use query mechanisms, should our operating assumptions prove overly optimistic. Use of NERD today does not foreclose use of such models in the future, and in fact both models can happily coexist.

Since the first draft of this document in 2007, portions of this work have been implemented. Future work should consider the size of fields, such as the version field, as well as key roll-over and revocation issues. As previously noted, CMS is now widely deployed. Current work on DNS-based authentication of named entities [RFC6698] may provide a means to test authorization of a NERD provider to carry a specific prefix.

We leave to future work how the list of databases is distributed, how BGP can play a role in distributing knowledge of the databases, and how DNS can play a role in aggregating information into these databases.

We also leave to future work whether HTTP is the best protocol for the job, and whether the scheme described in this document is the most efficient. One could easily envision that when applied in high-delay or high-loss environments, a broadcast or multicast method may prove more effective.

Speaking of multicast, we also leave to future work how multicast is implemented, if at all, either in conjunction or as an extension to this model.

Finally, perhaps the most interesting future work would be to understand if and how NERD could be integrated with the LISP mapping server [RFC6833].

11. Acknowledgments

Dino Farinacci, Patrik Faltstrom, Dave Meyer, Joel Halpern, Jim Schaad, Dave Thaler, Mohamed Boucadair, Robin Whittle, Max Pritikin, Scott Brim, S. Moonesamy, and Stephen Farrel were very helpful with their reviews of this work. Thanks also to the participants of the Routing Research Group and the IST-RING workshop held in Madrid in December of 2007 for their incisive comments. The astute will notice a lengthy References section. This work stands on the shoulders of many others' efforts.

12. References

12.1. Normative References

[ITU.X509.2000] International Telecommunications Union, "Information technology - Open Systems Interconnection - The Directory: Public-key and attribute certificate frameworks", ITU-T Recommendation X.509, ISO Standard 9594-8, March 2000.

[RFC3986] Berners-Lee, T., Fielding, R., and L. Masinter, "Uniform Resource Identifier (URI): Generic Syntax", STD 66, RFC 3986, January 2005.

[RFC5234] Crocker, D. and P. Overell, "Augmented BNF for Syntax Specifications: ABNF", STD 68, RFC 5234, January 2008.

[RFC6125] Saint-Andre, P. and J. Hodges, "Representation and Verification of Domain-Based Application Service Identity within Internet Public Key Infrastructure Using X.509 (PKIX) Certificates in the Context of Transport Layer Security (TLS)", RFC 6125, March 2011.

[RFC6830] Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "The Locator/ID Separation Protocol (LISP)", RFC 6830, January 2013.

12.2. Informative References

[CARP07] Carpenter, B., "IETF Plenary Presentation: Routing and Addressing: Where we are today", March 2007.

[CVS] Grune, R., Baalbergen, E., Waage, M., Berliner, B., and J. Polk, "CVS: Concurrent Versions System", November 1985.

[LISP-CONS] Farinacci, D., Fuller, V., and D. Meyer, "LISP-CONS: A Content distribution Overlay Network Service for LISP", Work in Progress, April 2008.

[RFC1034] Mockapetris, P., "Domain names - concepts and facilities", STD 13, RFC 1034, November 1987.

[RFC1383] Huitema, C., "An Experiment in DNS Based IP Routing", RFC 1383, December 1992.

[RFC2315] Kaliski, B., "PKCS #7: Cryptographic Message Syntax Version 1.5", RFC 2315, March 1998.

[RFC2616] Fielding, R., Gettys, J., Mogul, J., Frystyk, H., Masinter, L., Leach, P., and T. Berners-Lee, "Hypertext Transfer Protocol -- HTTP/1.1", RFC 2616, June 1999.

[RFC3977] Feather, C., "Network News Transfer Protocol (NNTP)", RFC 3977, October 2006.

[RFC4108] Housley, R., "Using Cryptographic Message Syntax (CMS) to Protect Firmware Packages", RFC 4108, August 2005.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC5246] Dierks, T. and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.2", RFC 5246, August 2008.

[RFC5652] Housley, R., "Cryptographic Message Syntax (CMS)", STD 70, RFC 5652, September 2009.

[RFC6698] Hoffman, P. and J. Schlyter, "The DNS-Based Authentication of Named Entities (DANE) Transport Layer Security (TLS) Protocol: TLSA", RFC 6698, August 2012.

[RFC6833] Farinacci, D. and V. Fuller, "Locator/ID Separation Protocol (LISP) Map-Server Interface", RFC 6833, January 2013.

[RFC6836] Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "Locator/ID Separation Protocol Alternative Logical Topology (LISP+ALT)", RFC 6836, January 2013.

[Usenet] Wikipedia, "Usenet", January 2013, <http://en.wikipedia.org/w/index.php?title=Usenet&oldid=531545312>.

[W3C.REC-soap12-part1-20070427]
Gudgin, M., Lafon, Y., Moreau, J., Hadley, M., Karmarkar, A., Mendelsohn, N., and H. Nielsen, "SOAP Version 1.2 Part 1: Messaging Framework (Second Edition)", World Wide Web Consortium Recommendation REC-soap12-part1-20070427, April 2007, <http://www.w3.org/TR/2007/REC-soap12-part1-20070427>.

[W3C.REC-soap12-part2-20070427]
Karmarkar, A., Hadley, M., Mendelsohn, N., Nielsen, H., Lafon, Y., Gudgin, M., and J. Moreau, "SOAP Version 1.2 Part 2: Adjuncts (Second Edition)", World Wide Web Consortium Recommendation REC-soap12-part2-20070427, April 2007, <http://www.w3.org/TR/2007/REC-soap12-part2-20070427>.

[W3C.REC-xml11-20040204]
Cowan, J., Maler, E., Sperberg-McQueen, C., Paoli, J., Bray, T., and F. Yergeau, "Extensible Markup Language (XML) 1.1", World Wide Web Consortium First Edition REC-xml11-20040204, February 2004, <http://www.w3.org/TR/2004/REC-xml11-20040204>.

[gb91] Smith, R., Gottesman, Y., Hobbs, B., Lear, E., Kristofferson, D., Benton, D., and P. Smith, "A mechanism for maintaining an up-to-date GenBank database via Usenet", Computer Applications in the Biosciences (CABIOS), April 1991.

Appendix A. Generating and Verifying the Database Signature with OpenSSL

As previously mentioned, one goal of NERD was to use off-the-shelf tools to both generate and retrieve the database. To many, PKI is magic. This section is meant to provide at least some clarification as to both the generation and verification process, complete with command-line examples. Not included is how you get the entries themselves. We'll assume they exist and that you're just trying to sign the database.

To sign the database, to start with, you need a database file that has a database header described in Section 3. Block size should be zero, and there should be no PKCS#7 block at this point. You also need a certificate and its private key with which you will sign the database.

The OpenSSL "smime" command contains all the functions we need from this point forth. To sign the database, issue the following command:

```
openssl smime -binary -sign -outform DER -signer yourcert.crt \
        -inkey yourcert.key -in database-file -out signature
```

-binary states that no MIME canonicalization should be performed. -sign indicates that you are signing the file that was given as the argument to -in. The output format (-outform) is binary DER, and your public certificate is provided with -signer along with your key with -inkey. The signature itself is specified with -out.

The resulting file "signature" is then copied into to PKCS#7 block in the database header, its size in bytes is recorded in the PKCS#7 block size field, and the resulting file is ready for distribution to ITRs.

To verify a database file, first retrieve the PKCS#7 block from the file by copying the appropriate number of bytes into another file, say, "signature". Next, zero this field, and set the block size field to 0. Next use the "smime" command to verify the signature as follows:

```
openssl smime -binary -verify -inform DER -content database-file
        -out /dev/null -in signature
```

OpenSSL will return "Verification OK" if the signature is correct. OpenSSL provides sufficiently rich libraries to accomplish the above within the C programming language with a single pass.

Author's Address

Eliot Lear
Cisco Systems GmbH
Richtistrasse 7
Wallisellen CH-8304
Switzerland

Phone: +41 44 878 9200
EMail: lear@cisco.com