

Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: January 16, 2009

A. Burness, Ed.
P. Eardley, Ed.
BT
L. Iannone
UC Louvain
July 15, 2008

Locator ID proposal evaluation
draft-burness-locid-evaluate-01

Status of this Memo

By submitting this Internet-Draft, each author represents that any applicable patent or other IPR claims of which he or she is aware have been or will be disclosed, and any of which he or she becomes aware will be disclosed, in accordance with [Section 6 of BCP 79](#).

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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

The list of current Internet-Drafts can be accessed at <http://www.ietf.org/ietf/lid-abstracts.txt>.

The list of Internet-Draft Shadow Directories can be accessed at <http://www.ietf.org/shadow.html>.

This Internet-Draft will expire on January 16, 2009.

Abstract

There are many proposals for improving the Inter-domain routing system, most of which involve a form of locator-identity split. There needs to be a means to reason about the strengths of the different proposals against the design criteria, and without requiring large scale implementations. This document aims to start this process by drawing parallels with existing systems. It identifies a number of questions that need to be more fully thought about whilst we press ahead with system development.

Internet-Draft

Locater ID proposal evaluation

July 2008

Table of Contents

1.	Introduction	4
2.	Design Goals	4
2.1.	Router Scalability	5
2.2.	Traffic Engineering	5
2.3.	Multi-Homing	5
2.4.	Mobility	6
2.5.	Ease of changing providers	6
2.6.	Routing Quality	6
2.7.	Routing Security	6
2.8.	Deployability	7
2.9.	Unclear Requirements	8
2.10.	Address Shortage	8
2.11.	Failure Management	8
3.	Related Working Options	8
3.1.	NAT	9
3.2.	Mobile networks and directory systems	10
3.2.1.	3G Systems	10
3.2.2.	Mobile IP	11
3.2.3.	DNS	11
3.2.4.	Summary	12
3.3.	The routing system	12
4.	Map and Encap Schemes	13
4.1.	Routing System Scalability	13
4.2.	Traffic Engineering	14
4.3.	Multi-Homing	14
4.4.	Mobility	15
4.5.	Changing Provider	15
4.6.	Route Quality	15
4.6.1.	Traffic Volume Overhead	16
4.7.	Routing Security	16
4.8.	Deployability	17
4.9.	Address Shortage	17
4.10.	Failure Handling	17
5.	Translation Schemes	18
5.1.	Routing System Scalability	18
5.2.	Traffic Engineering	18
5.3.	Multi-Homing	18
5.4.	Mobility	18
5.5.	Changing Provider	18
5.6.	Route Quality	19
5.7.	Deployability	19

5.8.	Address Shortage	19
5.9.	Failure Handling	19
6.	Mapping System Design	20
6.1.	Push	20
6.2.	Pull	20

6.2.1.	Data Collection	21
6.2.1.1.	Mapping Cache Size	21
6.2.1.2.	Mapping Cache Efficiency	23
6.2.1.3.	Mapping Lookups	23
6.3.	Route Through	23
7.	conclusions	24
8.	Acknowledgements	25
9.	IANA Considerations	25
10.	Security Considerations	25
11.	References	25
11.1.	Normative References	25
11.2.	Informative References	25
Appendix A.	Additional Stuff	27
	Authors' Addresses	27
	Intellectual Property and Copyright Statements	29

1. Introduction

The Internet routing system has problems with scalability and stability. These problems are made worse by the need to support functionality such as multi-homing and traffic engineering [[IAB](#)]. There have been a multitude of proposals that involve some form of locator-identity split that all aim to solve the problem of routing scalability. However without large scale implementations it is very difficult to assess the relative strengths of these different proposals. On the other hand, it should be possible to characterize the proposals against the requirements. Further, by comparing the proposals against existing systems, we may also be able to start to understand the likely processing, storage and communications requirements.

Whittle [[whittle](#)] has made a study of this type to compare some of the specific locator-ID split proposals. Here, instead of studying specific proposals, we group proposals into simple categories (map and encap schemes which were the focus of the previous study, translation schemes and directory systems) to enable us to understand the likely behaviour of whole groups of proposals at a more generic level.

This paper aims to start a process of evaluation. This document is written not as a truth, but as the perception of the authors that should be challenged.

We begin by reviewing the requirements against which proposals should be assessed. Then we highlight some existing systems which may have

processing, communications or memory requirements similar to those of the proposed schemes. Their behavior might help to guide us in assessing the proposals. This is essentially trying to learn from history. We appeal here in particular to equipment manufacturers who may have a better grasp of equipment capabilities; which are fundamental and which are limits based on market requirements. We then assess the generic schemes against the requirements. There are essentially two main approaches to routing, commonly known as map and encapsulation, and translation. The critique of map and encapsulation is based primarily on an understanding of LISP (draft 5) [[LISP](#)], the apparent current leader in that set of schemes; similarly the translation section is based upon 6/1 [[six-one](#)]. The aim is to be as critical as possible in order to stimulate future activity before making some conclusions.

[2.](#) Design Goals

In order to compare the solutions we need to understand the full

Burness, et al.

Expires January 16, 2009

[Page 4]

Internet-Draft

Locator ID proposal evaluation

July 2008

breadth of requirements for a future routing proposal. The first 7 are direct echoes of the requirements in [[Goals](#)], the later requirements we feel are not sufficiently highlighted in that draft. Although these are the list of requirements for the new routing architecture, there is no need for all these features to be implemented within one protocol. For example, making it easy for networks to change provider may mean that the edge network addresses need to be decoupled from those in the core. However an alternative approach is to develop automated tools that can smoothly manage address changes of hosts, routers and other elements (access control lists for example) within an edge network. Multi-homing may be managed by the routing system, or the routing system might simply(!) expose multiple paths that can be used by another mechanism to support multi-homing. However, we feel that any routing proposal should make clear how well the additional features could be supported in order to assess the whole solution.

[2.1.](#) Router Scalability

Memory and processing requirements are growing all the time; already routers need to be upgraded every 2 to 5 years. Many people believe the rate of growth is faster than Moore's law, meaning that cost

could go up significantly or technology could start to fail. The reason behind this growth appears to be a decreasing reliance on address aggregation rather than the absolute growth of the system itself. Also, it is not necessarily the actual memory requirements that is the problem, but the need to be able to read and write those memories quickly, because there is a high rate of churn in the routing system. The churn in the system also adds a processing requirement. Again, churn rates are increasing in line with increasing de-aggregation.

[2.2.](#) Traffic Engineering

Traffic engineering is the ability to direct traffic along non-default path(s). The ability to control the path taken by inbound traffic is as important as the ability to control the outbound path. Both these are non-trivial today: control of the in-bound data path requires manipulation of BGP messages. Control of the outbound path can be made difficult as a result of ingress filtering blocking data which appears to have been spoofed.

[2.3.](#) Multi-Homing

A multi-homed site can connect to the Internet via more than one network provider. Today this is done by injecting multiple, more specific address prefixes into the global routing table, which therefore impacts on BGP's scalability. Therefore any solution

should have a simple and effective means to manage multi-homing. Since one reason for multi-homing is to improve resilience, the multi-homing solution must be clear how failures are detected and repaired. This type of edge network failure management should ideally not impact on the convergence and stability of the global routing system. Availability requirements vary tremendously from a few seconds to as small as possible (ms range) where running sessions should not be affected.

Whilst multi-homing is primarily considered as a means of failure management, where-ever multi-homing appears, the desire for policy controlled routing simultaneously over all potential links soon follows.

[2.4.](#) Mobility

Increasingly nodes and sites will be mobile. An efficient, scalable means is needed to support mobility. When a host moves, hosts and routers that are not in communication with the mobile host should not need to be informed of the mobility. When a network moves, the number of routers informed of the change should be minimized.

[2.5.](#) Ease of changing providers

This is often cited as a key reason behind the increasing use of provider independent (PI) addresses, and hence a key reason behind routing scalability issues. Using PI addresses, end-sites can change providers without renumbering (or at least with much less disruption). Customers may want to change service provider on a yearly basis. A future routing system should make it easy for customers to change provider with minimal configuration requirements on the customer. The process should be as simple as possible, almost certainly automated.

[2.6.](#) Routing Quality

Quality of routes includes convergence time, stability of path, loss, delay and stretch. The first parameters are of interest to the network user. The later parameter gives an indication of efficiency of use of network transmission resources.

[2.7.](#) Routing Security

The new architecture should be at least as secure as the existing system.

[2.8.](#) Deployability

The Internet is stagnating; it is amazingly difficult to get new networking solutions deployed [[Handley](#)]. The solution MUST be:

- o Technically deployable
- o Incrementally deployable

- o All aspects of operation with legacy systems must be well understood. Applications (that have not hard-coded address into themselves) would see no changes. An updated or legacy node in a part of the Internet that uses the new system should be reachable by legacy or updated nodes operating within legacy or updated networks.
- o There must be a motivation for the person or organisation to deploy the system as solving the greater good is not sufficient. This benefit (or a subset) should exist for an isolated deployment. It seems probable that new functionality (rather than faster or even cheaper) is most likely to motivate deployment. This is because any new technology always has hidden costs such as training people to install and manage it for example. Examples of new functionality could be a security improvement, in and out-bound reliable traffic engineering, or visibility of alternative, low delay or highly reliable data paths. However, it is difficult to predict what new features or services will attract users
- o Flexible service models should be supported, in other words a user, edge site or ISP should be able to deploy the service on behalf of others.
- o Key players must not be disadvantaged, or they may try to obstruct standards or restrict deployment. A specific aspect of this to highlight is how network providers today use policy control. Providers are unlikely to support any scheme which make policy management more difficult than today. They are likely to require the ability to check that routes are as diverse as possible, to choose routes based on cost and performance and to avoid routes leaving or entering a specific country or domain.

If the constraints of operation with legacy systems and flexibility in location of functionality are met, then a non-issue is that of host upgradeability. However, host upgradeability is not impossible and recent history suggests this might be easier than network evolution. Recent host upgrades in ECN, IPv6 and RSVP based QoS are not being supported by similar network evolution.

Two other requirements are mentioned in [[Goals](#)].

The first is that mechanisms used must be first class elements within the architecture. I am not totally sure what this means.

The second requirement is that location and identification should be able to be decoupled. It is required that a solution for scalable routing is compatible with (but does not require) a solution that separates the host identification from the host location name-space. This separation should improve the flexibility of the Internet. The significance of this requirement is unclear, perhaps because none of the proposed solutions have failed to meet this requirement, and may only become clearer if assumptions or requirements on the identification such as cryptographic authentication requirements or the need to be able to reverse map from location to identifier are made.

Less often mentioned are two other requirements that we believe are nevertheless critical:

[2.10.](#) Address Shortage

Current predictions are that the unallocated IPv4 address space will soon be used up, with suggestions [[Huston](#)] that IANA will run out of addresses by 2011, with RIR running out by 2012. Routing and addressing are closely related, and the impact of the scheme on the address shortage problem should be considered. It may be easier if one major network overhaul is required rather than two.

[2.11.](#) Failure Management

If the routing system never encountered any changes, then it is likely that there would be no scalability issues. Minimizing connectivity disruption in the presence of failures is critical as failure recovery is one of the drivers behind multi-homing. Some end-sites have a target of no more than 10ms downtime _ although it is not clear that this would ever be achievable! Other sites may be happier with a few seconds disruption. Any scheme should make it clear how failures are handled, and should be no less robust to failure than today's systems.

[3.](#) Related Working Options

What can we learn from running systems today? This section is by no means complete, but rather presented as a starting point. In

particular, we have not got hard data on systems that exactly match anything than any new systems are trying to do. We are simply placing a stake in the ground at a loosely justifiable point; and asking people to move it. Note however that at this stage, we are looking for order of magnitude figures and hence OM movement of the stake!

3.1. NAT

NAT solves the problems of address shortage and provider independence. Hence, whatever we may feel about the architectural violations of NAT, we could imagine simply promoting the greater use of NAT to reduce the scalability problem as provider-dependent addresses can then be more easily promoted. It is after all clearly deployable. Many edge sites, mobile operators and even some ISPs are already using NAT, often claiming increased security in addition to the other benefits. For example, by hiding the addresses of servers and routers inside the network, it makes it a bit harder for an attacker to try and establish a session with these devices.

A NAT box can control traffic flows over different links if it is multi-homed, thus providing some traffic engineering capabilities. In particular, for sessions that are started behind the NAT, then the in and out-bound data path can be controlled by the choice of address that the NAT box uses for the session. One could imagine enhancements to NAT that would enable widely separated NAT boxes to communicate to support different multi-homing architectures.

Of course, there are issues with NAT which is why it has never been proposed as a solution to the routing system scalability problem; most significantly it breaks the end to end semantics of the Internet.

However, it is interesting to note that NAT is typically not used by the larger sites, and it appears to be the performance rather than any purist objections that lead to this.

The performance limitations come from the fact that NAT requires a high level of per-flow tracking and per-packet modifications. Because there are so many flavours of NAT, it is hard to get quantifiable information on the performance. For NAT-PT, we should expect to map one IP address to 65,000 different sessions using the port identifier. CISCO's web site [[Cisco](#)] suggest that a typical NAT router would not need to support more than 10,000 translations, and based on the same source, 128,000 such sessions would take 40MBytes DRAM.

Assuming we can easily support 128,000 NAT sessions, we can then

estimate then how many users this corresponds to. Each TCP flow is mapped to a different NAT session. A peer to peer application may run 100 concurrent sessions. Perhaps only 10% of an ISP customers are peer-peer users; the remaining 90% will typically have a low number of concurrent connections, say 5. So on average a customer has 14.5 active TCP sessions, meaning that the NAT as described can handle 8,827 users. This might mean that universities and medium enterprises could all be placed behind NAT devices, but larger corporate bodies and large ISPs would need something a little different, or very many co-ordinated NAT boxes. If we assume that the mappings maintained are between pairs of IP addresses rather than each individual TCP sessions, then we may be able to handle 3 times that number of users behind a single device.

Netflow is another networking tool that supports per flow packet processing. Cisco [[Cisco](#)] claim that their NETFLOW accounting tool can support 128,000 simultaneous connections - similar in scale to our NAT estimations.

In summary, per flow processing of each packet is likely to lead to limitations on how fast edge devices can operate, putting a limit on how many users could be behind such a box. Routers work fast today because they are highly optimised towards a single simple forwarding duty.

[3.2.](#) Mobile networks and directory systems

[3.2.1.](#) 3G Systems

GSM and 3G cellular systems already have a locater-identity split. For the voice system, the phone number acts as an identity. The Home Location Database (HLR) contains a mapping of the phone number to its current location as identified as a routing area. The HLR system will typically have, without known scalability concerns, up to tens of millions of users in this centralized database. A routing area will contain 10's or even 100's of cells which range in size from the few metres (pico cells in buildings in cities) to several kilometres in the countryside. To find a user, the HLR is used to discover which routing area last knowingly contained the phone. All nodes in

that routing area then receive a paging message in an attempt to discover the actual location of the user. This temporary location mapping is then held by the router responsible for that routing area.

The HLR mapping system does not know if the end node is reachable (and indeed for many data services the end node may not be reachable). This is discovered during the paging process, which means that it can take 5 to 10 seconds in order to make initial contact with a mobile device. When a user moves around a routing

area, it is not necessary to update the HLR of the location unless the node changes routing area. The size of the routing area depends not on how fast the HLR can be updated but on how much paging is expected as paging wastes the resources (battery power) of all phones in the area.

Handover (without session disruption) is only possible within one service provider network, as much as anything due to the time taken to manage security associations and general business concerns. Handover is managed locally with co-ordination between the different base stations. A make before break system is used to minimize service disruption.

Roaming occurs when a node changes service provider network. Here, the HLR will be updated to point to a Visitor Locator Database in the visited network which is updated with the routing area associated with the node. The (hand-crafted) peering arrangements to allow roaming are sorted by management processes.

[3.2.2.](#) Mobile IP

Mobile IP (MIP) uses a different scheme. Data is directed via a home agent which is updated with the current location of the mobile device. (In one sense this is similar to schemes where the data is re-directed via the mapping system). Mobile IP is much less widely deployed. Reasons for this could include the performance implications of the tunnelling process and the amount of per-node state management at the home agent. Designing adequate security mechanisms has also troubled MIP development.

[3.2.3.](#) DNS

Within the Internet, we already have experience with a large distributed database for mapping from name to address. DNS works well, so well in fact that people are loathe to change it in case it gets disrupted; after all it is a critical piece of the communications infrastructure and a user is unlikely to care if it is a routing or DNS problem that disrupts their on-line shopping trip - it will be equally broken.

It is usually stated that DNS works well because of the hierarchy in the name space (although the structure is relatively flat at about 3 levels) and the aggressive use of caching. Time to live (TTL) values are typically set at about an hour. However recent studies [[DNS](#)] are beginning to suggest that caching is not as vital as previously thought and that much shorter TTL of the order a few hundred seconds would not noticeably degrade DNS performance. This is in part because a DNS update message is only processed locally, there is no

attempt to keep all DNS servers with up to date information.

A host typically begins each transport layer session with a DNS lookup. This can take up to 2 seconds to resolve, although it is usually much quicker.

The DNS system is held together by IP addresses that are hand-coded into the system. A question to answer is what happens if the IP address is replaced by an intransient identifier and a transient locater. If the DNS servers need to be identified and their current location found before a DNS query could be resolved, then the performance of the identifier resolution system will have a big impact here as several DNS servers often need to be found to achieve a single name resolution. Further, if the DNS system is the identifier resolution system, we would have a nasty circular dependency.

[3.2.4](#). Summary

We can make some observations about the systems that work well. They seem to have extremely low functionality with low rates of change of the data. These changes are effectively confined (localized). Data changes are not propagated around the system. Hard-wiring of directory associations is commonplace. Perhaps an automated discovery and topology building protocol may give more problems than

its worth for this type of system? It is possible that automation is only required for systems with large amounts of change.

[3.3.](#) The routing system

So having considered systems that work well, what are the characteristics of the routing system? A mid-tier isp network may contain double the number of prefixes as the core of the Internet - thus we must be careful of designs that move complexity from the core to the periphery of the network.

how many prefixes; how many AS; how many nodes; how many end sites; how many transits; how big is the DFZ;

The churn rate is very high and very variable. If a site receives on average 400 BGP messages a minute, it may easily expect to have 8000 or 80,000 updates at peak periods of intense instability.

Many of these messages are not really indicating true physical problems. A site may rapidly flap its links in an effort to manipulate the flow of data between different multi-paths. A site may perform mild policy updates.

Burness, et al.

Expires January 16, 2009

[Page 12]

Internet-Draft

Locater ID proposal evaluation

July 2008

Churn is typically slowed by the introduction of timers to delay sending of messages. Often however these timers are turned as low as possible to try and maximise network availability. Ideally, local repair mechanisms should be used to recover from failure without involving the entire routing system.

[4.](#) Map and Encap Schemes

[4.1.](#) Routing System Scalability

These schemes aim to encourage use of provider dependent addresses thus removing the load from the core routing system. This is achieved by making addresses in the edge network independent from those in the core transit system, so that provider lock-in is avoided.

All these schemes require a mapping system to translate between edge

and core network locators. The scalability of mapping system is uncertain. We shall assume that the mapping system holds essentially static information. We further assume that (using LISP terminology) End Point Identifiers (EID) are aggregatable so a system of required size could be built. It is probable that this system could be built to store and return all locators associated with an end point identifier prefix range. Issues that would impact the probable scalability of this system are

- o if the system needs to propagate this information globally, in which case it would become very sensitive to churn rate and bandwidth. In this case, it could not sensibly be used for mobility management for example
- o if the system was used to propagate policy or traffic engineering information, as all the evidence is that this information is very rapidly changing

The third item to be considered are the edge routers which may need to do per-flow packet processing. This processing may be required to manage reachability information (is it sufficient to hold a mapping to the core locator of the edge router and to know that the lower layer routing system thinks this address is still valid, or do we need to know that the higher layer functionality is alive?) Further, the LISP description of multi-homing management seems to imply per-flow packet processing for example reading of headers on return packets of a flow to discover which of the possible edge routers are prepared to handle this session). If per-flow packet processing is required, we may run into scalability problems as in NAT routers today. Is the per-flow assumption fair? If we were considering all

flows to a specific tunnel end-point, perhaps there may be some way to aggregate information? This would depend on the location of the tunnel end points. If they are near to the network edge it is quite likely that there will be a limited number of flows heading towards a specific the tunnel router. If the edge routers are near the core, we then introduce a scaling problem behind the edge routers, where all networks now have provider independent address spaces. Since the absolute size of the mid-tier networks is greater than that of the DFZ, adding scaling pressure here is unlikely to be a good idea.

Of course, another way to consider these schemes is to assume that

they do nothing apart from append a new packet header at the edge router: in this case, a better simile would be with MPLS; where the primary scalability worry to date comes from lack of labels (only 20 bytes available). The main issues with MPLS are the ability to verify reachability, rather than processing and memory requirements. Certainly MPLS has yet to be implemented inter-domain and is not suggested as a solution itself.

In summary, the main scalability questions may arise only when a clearer understanding of how multi-homing with traffic engineering are to be managed.

[4.2.](#) Traffic Engineering

Traffic engineering and policy controls may require co-ordination between two layers. It requires the ITR to respect ETR instructions. It is probable that some policy opaqueness is lost. One interesting question for example, is how peering relationships are managed, as to be reachable by any node, the ETR must be advertised openly in the mapping system, and once this is done, how is it ensured that only networks with distinct peering relationships use the more expensive links?

[4.3.](#) Multi-Homing

By separating the routing system into two parts, it is expected that it will be possible to implement multi-homing management separately from the routing system.

The mapping system may return many possible locaters. The edge routers using edge to edge communications manage multi-homing. In LISP it is described how an ITR will spray packets from a flow across the different possible ETRs, according to the weights associated with the ETR devices. The ETRs communicate back to the ITR(s) which addresses they would prefer to see used. This is used for traffic engineering as well as simply reachability purposes. If this information is piggybacked onto a data session , which may raise

security questions [[Bagnulo](#)] , how is this managed for UDP applications which may have the return control channel as a different session to the data channel? This also breaks the model that TCP has, of packets typically following a single path which may have

unfortunate implications both for congestion control and for TCP performance. If we assume that a TCP flow is kept together, but that packets destined to the same end site are spread amongst the edge routers, we now definitely have per-flow state, and unlike ECMP, associated packet processing (adding the correct outer header).

[4.4.](#) Mobility

As in multi-homing, by separating the routing system into two parts, it is expected that it will be possible to implement mobility management as an overlay to the routing system.

It may be possible to manage simple portability by updating the mapping system so that new sessions would start correctly. This assumes that the mapping system operates like DNS today, without the information needing to be distributed globally. In session mobility however requires the updating of the mappings dynamically; Discussions on the mailing lists [[MailList](#)] to date imply that this is difficult, with suggestions that it should be an application specific signalling. It is likely that should source and destination simultaneously move, the session will be dropped, unless the edge routers offer a forwarding functionality.

[4.5.](#) Changing Provider

This is by design extremely simple as only the mapping system needs updating. However there may still be issues ensuring packet filters and firewalls are correctly configured. These have been covered to some extent for Ipv6 in [RFC 4192](#) [[RFC4192](#)] where make before break techniques have been described, but this may not be suitable on the whole for IPv4.

[4.6.](#) Route Quality

Since multiple edge routers can be associated with a name, the network system may have a greater choice of routes to use to reach a specific device (although it is not clear that this control could be passed back to the data sources).

If the mapping replies take a long time, a TCP session start up may be disrupted. Similarities with ARP are not necessarily relevant: ARP is an extremely local process that can resolve very quickly, and ARP entries are normally within a cache because they are used frequently.

Since multi-homing requires a flow to be sent along diverse paths TCP may see lots of out of sequence packets and congestion control mechanisms may not work as expected .

[4.6.1.](#) Traffic Volume Overhead

It is not clear how easy it is to solve the problem of tunnel overheads and packet fragmentation, or if indeed that is a major issue. During the study of locator-ID cache performance, described below in [Section 6.2.1](#) an analysis was also made of the overhead in terms of traffic volume. Table 1 and Table 2 compares the original traffic volume (expressed in Mbit/sec) with the volume obtained encapsulating all packets, respectively for incoming traffic and outgoing traffic. As can be observed, the overhead introduced by the tunneling approach consists in few Mbit/sec. For outgoing traffic this means an overhead that ranges from 4.15% up to 11.15% . For incoming traffic this means an overhead that ranges from 3.8% up to 5.75%. What the table also show is that even if in terms of absolute bandwidth the overhead is more important during the high traffic load period (i.e., day), in terms of percentage points it is more important during the low traffic load period (i.e., night).

Traffic	Min	Max	Avg Night	Avg Day
Original	13.22	108.10	18.70	85.62
Encapsulated	13.98	112.21	19.72	89.17
	(+5.75%)	(+3.8%)	(+5.45%)	(+4.15%)

Table 1: Incoming Traffic Volume (in Mbit/sec)

Traffic	Min	Max	Avg Night	Avg Day
Original	6.28	48.25	9.75	32.58
Encapsulated	6.98	51.67	10.68	35.63
	(+11.15%)	(+7.09%)	(+9.54%)	(+4.15%)

Table 2: Outgoing Traffic Volume (in Mbit/sec)

[4.7.](#) Routing Security

Pending further thought. The security analysis so far performed [[Bagnulo](#)] was on LISP version 1.

[4.8.](#) Deployability

- o Technically deployable
- o It is not clear how incrementally deployable this is. If it is required that (PI) EID space is advertised in the legacy routing system to enable communication with legacy nodes, then the scaling pressures on the routing system will shoot up dramatically during the early stages of deployment.
- o Operation with legacy systems is not well understood
- o There is no clear motivation why an edge system should deploy this scheme. Since provider lock-in can be avoided today using existing well known techniques, there is no motivation for a end site to chose LISP over the familiar technology. Traffic engineering and multi-homing control have been mentioned as possibilities to motivate a deployment, but to date are too poorly described to be able to judge if they meet all requirements well.
- o There may be opposition as traffic engineering and policy control requires communications between ITR and ETR devices, which may reduce the opaqueness of the policy control over existing techniques. Policy control may become more complicated

[4.9.](#) Address Shortage

Although described for IPv4, which is seen as an advantage, these schemes are essentially IP version agnostic. Unlike the NAT solutions of today, the EIDs in any domain must have global uniqueness for the mapping system, thus potentially making the problem worse. Although better allocations of addresses may become possible, it is unlikely that addresses can be easily recovered.

[4.10.](#) Failure Handling

These schemes always require an additional global database infrastructure. This is therefore as critical a resource as the current DNS system is. All things being equal, the addition of this would decrease the resilience of the overall Internet. Further,

fault tracing would become yet more complex. The underlying routing system takes care of path failures between the tunnel routers. However tunnel routers become critical points of failure if they hold state.

[5.](#) Translation Schemes

[5.1.](#) Routing System Scalability

It aims to encourage use of provider dependent addresses removing the load from the core routing system. It does this by providing a different way to manage multi-homing. Since the edge routers are not state holding, and only need to tamper with the first few packets of a flow, the scalability of these edge routers should be better than that of current NAT devices.

[5.2.](#) Traffic Engineering

In and outbound traffic engineering is managed through either the node or egress router setting the routing portion of the locater. For in-bound sessions, this only works when both ends are translation aware. Existing policy control is possible, although there is motivation to move to alternative ways to achieve same goal. For example AS pre-pending to indicate that a route should be avoided could be replaced with a translation to the preferred route. Since this could work more reliably than AS pre-pending there is a driver for change.

[5.3.](#) Multi-Homing

For multi-homed edge networks (as opposed to multi-homed hosts) this can be controlled by edge networks but is visible to end hosts. Applications bind only to the identifier part of the address.

[5.4.](#) Mobility

Since applications can tolerate the address changing, mobility should be simplified. Many of the functions to support multi-homing are

like those to support mobility but it is not clear that the details and overlaps have been fully identified, especially with regard to security.

[5.5.](#) Changing Provider

This will be complicated, and additional protocol support will be required. As well as DHCP re-configuration of hosts, there will be DNS updates and firewall and filter settings. Also the intra-domain routing system may be affected. This later problem may be made more manageable if internal routers can mask out the network address portion within the internal routing system. This may make it harder to do efficient routing inside the network or to manage edge node failures. An architectural viewpoint would suggest that this problem will remain unsolved because identifiers are only allocated to end

Burness, et al.

Expires January 16, 2009

[Page 18]

Internet-Draft

Locater ID proposal evaluation

July 2008

systems, making IP address the primary identifier used in all management systems.

[5.6.](#) Route Quality

The scheme adds minimal additional delays. All data translations are based only on locally held, locally visible material. Alternative routes, as indicated by different address pairings, are visible to the end devices.

[5.7.](#) Deployability

- o Technically deployable
- o Proxy support, to avoid upgrading of hosts, may look very like NAT with a break in the end to end semantics
- o Some of the new benefits over the existing system (specifically in-bound TE) are only evident when there is a large deployed base.
- o Operation with legacy hosts is possible provided all 6/1 elements can identify it as a legacy host
- o Motivation is based on additional feature of in-bound TE. The ability to see and use different routes, as identified through different addresses may also be valuable.

- o Hosts, edge devices and possibly internal networks all need to be upgraded.

[5.8.](#) Address Shortage

Forces upgrade to IPv6

[5.9.](#) Failure Handling

Since the edge devices are expected only to translate on the first packets of a flow (relying on the end host to use the correct address once it is made aware), the edge devices become less critical as they are not state holding. It has been suggested that Should the edge router or access link fail, a local mechanism (similar to handover in a cellular system) can be used to achieve fast recovery.

Relies on DNS system to provide the locator mapping. Currently DNS servers are found through the hard-coding of related DNS server addresses. If addresses become transient what does this mean for the DNS system? Thus although a separate resolution system is not required, some consideration on DNS use would still be needed. Would

DNS servers need to be logically within the transit (provider independent address) zone?

[6.](#) Mapping System Design

The concept of tunnelling IP data packets across a large scale network is not new. Many years ago there was much activity put into the design of networks that could run IP over ATM clouds. This activity failed because of the difficulty of managing the mapping process - hence the design of MPLS which uses a single IP control plane across the entire network. Are there any lessons to be learnt from this experience?

There appear to be three basic options: push, pull or route through.

[6.1.](#) Push

If the full database is pushed to all tunnel routers, these devices

may end up with larger storage requirements than current routers because all end sites now have provider independent addresses and so no aggregation is possible here. There is also the problem of keeping the database securely up to date. This is the way that name to address mappings were originally managed, before DNS was introduced. This new database could however be smaller than DNS because you have a locator associated with an EID prefix (ie roughly equivalent to having a locator associated with bt.com, not one for www.bt.com, mail.bt.com etc). There have been claims that this mapping system would be easier to manage than the current routing system because it can be the same everywhere, whereas a routing table varies according to the router. However, link state protocols actually distribute a topology database which is the same everywhere, and they are not used for very large scale networks is this because there is no localisation of changes and they are considered un-scalable, or is this because they do not provide suitable hooks for policy routing? It is possible that the scalability concerns are out-dated [[OSPF-LITE](#)]

[6.2.](#) Pull

DNS is an example of a pull system. It enables localisation of changes so could be used to carry more dynamically varying information, although the rate of updates should be slower than the cache lifetimes. The disadvantage of this scheme used mid-flight is the additional delays that will be introduced. These, as well as being annoying, may also upset protocols such as TCP. Further, as the query is performed by a network element this opens up the potential for a DOS attack where a source simply sends initial

packets to unknowable destinations. However, the results of a study, summarised below suggest that the size of the cache mapping and be made small, with a relatively small timeout, whilst still achieving a high hit rate. This could mean that the negative impacts will be constrained. More results can be found in [[MAPCOST](#)].

[6.2.1.](#) Data Collection

During the end of May and the beginning of June 2007, NetFlow [[NETFLOW](#)] traces have been collected from UCL [[UCL](#)] Campus network, from the border router, a Cisco Catalyst 6509. The UCL network has almost ten thousand active users per day. It uses a class B (i.e.,

/16) prefix block and is connected to the Internet through a border router that has a 1 Gigabit link toward the Belgian National Research Network (Belnet). These traces have been used to try and emulate the behavior of the Loc/ID separation cache, as if a protocol such as LISP ([\[I-D.farinacci-lisp\]](#)) was deployed on the border router of UCL campus network.

The analysis performed assumes that the ID-to-Locator mapping has the same granularity as the prefix blocks assigned by RIRs. A first analysis of the traffic of UCL network shows that the number BGP prefixes contacted per minute ranges from 3,618 up to 11,074, with a clear night/day cycle. In particular, during the night the average number of prefixes contacted per minute is 4,000, while during the day the average raises to slightly more than 8,000, thus doubling the load.

[6.2.1.1](#). Mapping Cache Size

The cache emulator uses a timeout policy in order to flush unused cache entries. In particular, the analysis on the traces has been performed three times with three different timeout values, respectively three (3), thirty (30) and three hundred (300) minutes.

Table 3 shows the summary of the size of the Mapping Cache, expressed in number of entries, for a daylong observation period. The table also shows the average size during night period and day period. The night period is the average of the Mapping Cache size between 0 am and 6 am, which is the period with the lowest traffic load, while the day period is the average between 10 am and 4 pm, which is the period with the highest traffic load.

Timeout	Min Size	Max Size	Avg Night Size	Avg Day Size
3 Min.	7,530	17,804	8,056	14,093
30 Min.	22,588	43,529	24,161	38,405

300 Min.	61,939	103,283	65,600	81,060
+-----+	+-----+	+-----+	+-----+	+-----+

Table 3: Mapping Cache Size (in number of entries)

From the previous table is easy to compute the size of the Mapping Cache in terms of bytes by using the following equation:

$$S = E \times (5 + N \times 6 + C) \quad (1)$$

Where S is the size of the cache expressed in bytes and E is the number of entries. N represents the number of RLOCs per EID. Note that due to multihoming, an EID can be associated to more than one RLOC. The number 6 represents the size of an RLOC assuming four bytes for the address and two other bytes for traffic engineering purposes (e.g. priority and weight like in [[I-D.farinacci-lisp](#)]). C represents the overhead in terms of bytes necessary to build the cache data structure. Assuming the cache is organized as a tree, C can be set to 8 bytes, just the size of a pair of pointers. The number 5 represents the size of an EID. Since we are using mappings with a granularity of BGP prefixes, five bytes are necessary, four for the IP prefix address and one for the prefix length. Table 4 shows the maximum size (in Kbytes) for the Mapping Cache assuming respectively one, two, or three RLOCs for each EID. Depending on the timeout used, the size of the cache can range from few hundreds KBytes, up to few MBytes.

+-----+	+-----+	+-----+	+-----+
Timeout	1 RLOC	2 RLOCs	3 RLOCs
+-----+	+-----+	+-----+	+-----+
3 Min.	334	440	545
30 Min.	807	1062	1317
300 Min.	1917	2522	3127
+-----+	+-----+	+-----+	+-----+

Table 4: Mapping Cache Maximum Size (expressed in Kbytes)

[6.2.1.2.](#) Mapping Cache Efficiency

The cache hit rate does not increase proportionally with the cache size. Table 5 shows the summary of the analysis of the hit ratio for a daylong observation period. The averages present in the table are calculated in the same time period as for the size ([Section 6.2.1.1](#)).

Timeout	Min	Max	Avg Night	Avg Day
3 Min.	91.4%	97.5%	93.5%	96.4%
30 Min.	96.8%	99.5%	98.5%	99.2%
300 Min.	98.9%	99.9%	99.7%	99.8%

Table 5: Mapping Cache Efficiency (Hit Ratio)

[6.2.1.3.](#) Mapping Lookups

The previous sections have shown some analysis related to the Mapping Cache. In order to build the cache a lookup operation is needed each time the correct mapping is not present in the cache. This means that the lookup operation, which consists in a query to a mapping distribution system, can be triggered by both a new outgoing flow as well as a new incoming flow. Table 6 shows a summary of the lookup operations for a daylong observation time.

Timeout	Min	Max	Avg Night	Avg Day
3 Min.	1,301	4,046	1502.5	2381.4
30 Min.	257	1,211	357.1	540.7
300 Min.	19	328	78.7	161.7

Table 6: Lookups queries per Minute

[6.3.](#) Route Through

Routing through systems will increase the work expected from name resolution servers. It may lead to inefficient routing. If this is only used for the start of a data flow (and for all short sessions of course), then TCP flow rates will frequently be incorrect (too fast or slow for the path they have been changed to). Applications such as voice also seem to struggle to cope with large path changes because of the delay variation seen. This might also make fault tracing much more complex.

Internet-Draft

Locator ID proposal evaluation

July 2008

[7.](#) conclusions

1. There is no obvious correct solution. The two classes of solution both aim to increase the use of aggregatable addresses and essentially differ in the driver they assume is the more critical, ie provider lock-in or multi-homing support. The working assumption should be that both problems must be adequately solved by any solution, unless one requirement can be proven to be irrelevant
2. We are not really sure if there is a problem, although it could be major and if we leave it until we are certain it is likely to be too late to solve it. More importantly, the exact nature of the problem (FIB size, RIB size, processing churn, writing FIB updates etc) has escaped definition. A simpler solution may be possible.
3. Each of the different approaches deserves further research.
4. the area that has received least real attention is legacy inter-working and partial deployment.
5. the mapping system is a real crunch point and needs some serious analysis
6. We are focusing on the locator-ID split, but have in reality two types of split, one which is recognizable as a locator-identity split and the other which could be termed a locator-locator split which involves splitting the addressing regions into core and edge. The addition of an identifier has been proposed in other quarters for security and authentication reasons. What are the wider implications of a locator space split?
7. Compact routing is a completely different routing algorithm that essentially trades path stretch for router state. At present there is no way to implement a distributed dynamic version of compact routing so this particular protocol may be very far out. Nevertheless, there is no apparent study of the potential of different routing algorithms
8. Schemes such as HRA which simply look at how we organize the routing system are not included.

9. ROFL assumes that there is really no need for any locator at all and it may be correct. It assumes that using modern techniques (based on DHTs) we could build an adequate system based on semantic-free identifiers. It may be that the problems we face are caused by things other than scalability (eg lack of

accountability means that we get endless pointless update messages, and means that there is no back-pressure on de-aggregation).

10. We are looking at the simple schemes; complex schemes such as NODE-ID and HRA are not considered. However, in considering small scale changes, are we missing the point that we should first have a long term target architecture that any point solution should be compliant with?

[8.](#) Acknowledgements

An preliminary version of this document was prepared for Chinacom with help from Sheng Jiang and Xiaohu Xu.

We are grateful to help from Olivier Bonaventure, and to the mailling list discussions, especially Robin Whittle and Joel M. Halpern for very useful comments

[9.](#) IANA Considerations

This memo includes no request to IANA.

[10.](#) Security Considerations

[11.](#) References

[11.1.](#) Normative References

[min_ref] authSurName, authInitials., "Minimal Reference", 2006.

11.2. Informative References

- [Bagnulo] Bagnulo, M., "Preliminary LISP Threat Analysis", 2007, <<https://datatracker.ietf.org/drafts/draft-bagnulo-lisp-threat/>>.
- [Cisco] Cisco, "NAT FAQ", 2008, <<http://www.cisco.com/warp/public/556/nat-faq>>.
- [DNS] Jung, J., "DNS performance and the effectiveness of caching", 2001, <SIGCOMM workshop on Internet Measurement>.

Burness, et al. Expires January 16, 2009 [Page 25]

Internet-Draft Locator ID proposal evaluation July 2008

- [FLOWTOOLS] Fullmer, M., "Flow-tools - tool set for working with netflow data.", Available Online at: <http://www.splintered.net/sw/flow-tools/docs/flow-tools.html>.
- [Goals] Li, T., "Design Goals for Scalable Internet Routing", 2007, <Internet draft [draft-irtf-rrg-design-goals-01](#)>.
- [Handley] Handley, M., "Why the Internet only just works", 200, <BT Technology Journal>.
- [Huston] Huston, G., "IPv4 address report", 2007, <<http://www.potaroo.net/tools.ipv4/index.html>>.
- [I-D.farinacci-lisp] Farinacci, D., Fuller, V., Oran, D., Meyer, D., and S. Brim, "Locator/ID Separation Protocol (LISP)", [draft-farinacci-lisp-08](#) (work in progress), July 2008.
- [I-D.vogt-rrg-six-one] Vogt, C., "Six/One: A Solution for Routing and Addressing in IPv6", [draft-vogt-rrg-six-one-01](#) (work in progress), November 2007.
- [IAB] Meyer, D., "Report from the IAB workshop on Routing and Addressing", 2007, <Internet draft [draft-iab-raws-report-02.txt](#)>.

- [LISP] Farinacci, D., "Locator/ID separation Protocol (LISP)", 2007, <Internet draft [draft-farinacci-lisp-05.txt](#)>.
- [MAPCOST] Iannone, L. and O. Bonaventure, "On the Cost of Caching Locator/ID Mappings.", 3rd Annual CoNEXT Conference, 2007.
- [MailList] Farinacci, D., "e-mail thread", 2007, <<http://www.ops.ietf.org/lists/rrg/2008/msg00232.html>>.
- [NETFLOW] Cisco Sytems, "Introduction to cisco ios netflow - a technical overview.", Available Online at: http://www.cisco.com/en/US/products/ps6601/products_white_paper0900aecd80406232.shtml.
- [OSPF-LITE] Thomas, M., "OSPF-Lite", 2007, <<https://datatracker.ietf.org/drafts/draft-thomas-hunter-reed-ospf-lite/>>.

Burness, et al.

Expires January 16, 2009

[Page 26]

Internet-Draft

Locater ID proposal evaluation

July 2008

- [RFC4192] Baker, F., Lear, E., and R. Droms, "Procedures for Renumbering an IPv6 Network without a Flag Day", [RFC 4192](#), September 2005.
- [UCL] "Universite Catholique de Louvain.", <http://www.uclouvain.be>.
- [six-one] Vogt, C., "Six/one: A solution for routing and addressing in IPv6", 2007, <Internet draft [draft-vogt-rrg-six-one-01](#)>.
- [whittle] Whittle, R., "Comparing LISP-NERD/CONS, eFIT-APT and Ivip", 2007, <<http://www.firstpr.com.au/ip/ivip/comp/>>.

[Appendix A](#). Additional Stuff

This becomes an Appendix.

Authors' Addresses

Louise Burness (editor)
BT
BT Adastral Park
Martlesham Heath, Suffolk
UK

Phone: +44 1473 646504
Email: louise.burness@bt.com

Philip Eardley (editor)
BT
BT Adastral Park
Martlesham Heath, Suffolk
UK

Phone:
Email: philip.eardley@bt.com

Burness, et al.

Expires January 16, 2009

[Page 27]

Internet-Draft

Locater ID proposal evaluation

July 2008

Luigi Iannone
UC Louvain
Place St. Barbe 2
Louvain la Neuve, B-1348
Belgium

Phone: +32 10 47 87 18
Email: luigi.iannone@uclouvain.be
URI: <http://inl.info.ucl.ac.be>

Full Copyright Statement

Copyright (C) The IETF Trust (2008).

This document is subject to the rights, licenses and restrictions contained in [BCP 78](#), and except as set forth therein, the authors retain all their rights.

This document and the information contained herein are provided on an "AS IS" basis and THE CONTRIBUTOR, THE ORGANIZATION HE/SHE REPRESENTS OR IS SPONSORED BY (IF ANY), THE INTERNET SOCIETY, THE IETF TRUST AND THE INTERNET ENGINEERING TASK FORCE DISCLAIM ALL WARRANTIES, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO ANY WARRANTY THAT THE USE OF THE INFORMATION HEREIN WILL NOT INFRINGE ANY RIGHTS OR ANY IMPLIED WARRANTIES OF MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE.

Intellectual Property

The IETF takes no position regarding the validity or scope of any Intellectual Property Rights or other rights that might be claimed to pertain to the implementation or use of the technology described in this document or the extent to which any license under such rights might or might not be available; nor does it represent that it has made any independent effort to identify any such rights. Information on the procedures with respect to rights in RFC documents can be found in [BCP 78](#) and [BCP 79](#).

Copies of IPR disclosures made to the IETF Secretariat and any assurances of licenses to be made available, or the result of an attempt made to obtain a general license or permission for the use of such proprietary rights by implementers or users of this specification can be obtained from the IETF on-line IPR repository at <http://www.ietf.org/ipr>.

The IETF invites any interested party to bring to its attention any copyrights, patents or patent applications, or other proprietary rights that may cover technology that may be required to implement this standard. Please address the information to the IETF at ietf-ipr@ietf.org.