G. Huston
APNIC

February 11, 2005

Architectural Commentary on Site Multi-homing using Level 3 Shim
draft-huston-l3shim-arch-00.txt

Status of this Memo

Copyright Notice

Abstract

This document provides a commentary of the Level 3 Shim approach to site Multi-homing (L3Shim) as described in [ID.L3SHIM], [ID.REFER], [ID.FUNC], and [ID.HBA], using as a framework for this analysis the approach described in [ID.ARCH].

Notes

This initial draft has been prepared as a commentary on the L3 Shim proposal as developed by a Design Team of the Multi6 Working Group. The document attempts to provide a commentary on the proposal according to the framework described in the multi-homing architecture document.

The L3 Shim specification is an initial pass, and there are areas where the documentation is incomplete. This commentary is also incomplete, and will require further revision as the L3 Shim approach is refined.

In addition this initial draft does not analyze the properties of the HBA and CGA address types, and their role in providing some resilience against various forms of third party attacks. This analysis should be included in future revisions of this document.

Table of Contents

1. Introduction

As noted in the general architectural overview of approached to Multi-homing in IPv6 [ID.ARCH] document, there are a number of general approaches to supporting site Multi-homing. These include the use of the routing system, use of mobility mechanisms, modification of existing elements in the protocol stack, or the introduction of a new protocol stack element, and the modification of behaviours of hosts and site-exit routers.

This document provides an commentary of the Level 3 Shim approach to site Multi-homing (L3Shim) as described in [ID.L3SHIM], [ID.REFER], [ID.FUNC], and [ID.HBA], using as a framework for this analysis the approach as described in [ID.ARCH].

2. Summary of L3Shim

The approach used by "Level 3 Shim" (L3Shim) is, as the name suggests, one that is based on the modification of the Internet Protocol stack element within the protocol stack of the endpoint. The modification is in the form of an additional functionality block, as indicated in Figure 1.

```
+-----------------------------------------------------------------+
| Transport Protocols                                             |
+-----------------------------------------------------------------+
+-----------------------------------------------------------------+
| IP Protocol                                                s     |
|                                                                 |
|   IP Endpoint     ------ ------- -------------- -------------   |
|    sub-layer      | AH | | ESP | | Frag/reass | | Dest opts |   |
|                   ------ ------- -------------- -------------   |
|                                   |                             |
|   Multi6                 =====================                  |
|    sub-layer  ========> || multi6 shim layer ||                |
|                          =====================                  |
|                                   |                             |
|   IP Routing                   -------                          |
|   sub-layer                    | IP  |                          |
|                                -------                          |
|                                                                 |
+-----------------------------------------------------------------+
```

L3 Shim Protocol Stack (From [ID.L3SHIM]

Figure 1

Above the L3Shim protocol element the protocol stack uses constant endpoint identities to refer to both itself and to the remote protocol stack. The shim layer provider a set of associations between endpoint identity pairs and locator sets.

As packets are passed from the IP Endpoint sub-layer to the IP Routing sub-layer, the endpoint identities are mapped to a current pair of locators. The reverse mapping is applied to incoming packets, where the incoming locator pair is stripped off the packet, and the associated endpoint identity pair is associated with the packet which is then passed to the IP Endpoint sub-layer. Demultiplexing the IP packet to the appropriate transport session is based on the endpoint identities. In this L3Shim approach the endpoint identities and the locators are both IP addresses. The endpoint identities are the initial addresses used between the two hosts. The locators are the set of IP addresses that are associated with the endpoint.

The intention of this approach is to minimise the amount of change required to support dynamic locator agility in the protocol stack, and support dynamic locator agility as a negotiated endpoint-to-endpoint capability. An application can initiate a session with a remote host by using an entirely conventional lookup of the host's domain name in the DNS, and open up a session with the remote endpoint using one of its addresses as the destination address. The application can continue to exchange packets with this remote host for the duration of the session by continuing to use this destination address. If the local host subsequently opens up a new session with the same remote host, the same destination address may be used, or if the local host passes a reference to a third party as a referral, the same destination address may be used. In terms of semantics and functionality this represented no change to the use of addresses as endpoint identifiers in the IPv6 architecture.

The Layer 3 shim operates as a per-host header address mapping function. When the shim locator mapping function is activated for a remote endpoint packets passed from the IP endpoint sub-layer to the shim sub-layer have the packet's headers source and destination addresses rewritten with the currently selected locator pair. Incoming packets passed from the IP Routing sub-layer undergo a similar lookup using the locator pair. The packet header is rewritten with the mapped endpoint identifier pair is there is an active mapping entry. This

functionality is indicated in Figure 2. Here the endpoint identities are referred to as Upper Layer Identifiers (ULIDs), and the packet header addresses are referred to as Locators (L). The L3Shim element contains a context state, associating a ULID pair (in this case the pair [ULID(A),ULID(B)] with a set of locators for A and a set of locators for B. The shim elements are synchronised such that complementary mappings are performed at each end of the connection.

```
         ----------------------------          ----------------------------
         | Sender A                 |          | Receiver B               |
         |                          |          |                          |
         |     ULP                  |          |     ULP                  |
         |      |                   |          |      ^                   |
         |      | src ULID(A)       |          |      | src ULID(A)       |
         |      | dst ULID(B)       |          |      | dst ULID(B)       |
         |      v                   |          |      |                   |
         |---L3Shim-----------------|          |---L3Shim-----------------|
         |      |                   |          |      ^                   |
         |      | src L(A)          |          |      | src L(A)
|         |
         |      | dst L(B)          |          |      | dst L(A)          |
         |      v                   |          |      |                   |
         |      IP                  |          |      IP                  |
         ----------------------------          ----------------------------
                |                                     ^
                --------------- network path -- -------
```

Mapping with changed locators.(From [ID.L3SHIM]

Figure 2

The implication of the decision to place the endpoint identity-to-locator mapping protocol element within the IP element is that this mapping function is not implicitly aware of session start and tear down. At this level of the protocol stack there is no information to indicate wither this packet is a single datagram, or the start of an extended packet exchange with a remote entity. Similarly there is no explicit information provided to the shim protocol element to indicate when a session is complete, at which point the mapping state information could be discarded and the associated host resources reclaimed. This is offset by the advantages of this approach in that there is no explicit need to alter the function of any transport protocol, as the shim element continues to present constant endpoint identities to the upper protocol levels, irrespective of the current endpoint/locator mapping being used between the two hosts.

Assuming that the initial choice of a ULID corresponds to a viable network path, the initial state of the L3Shim is a null mapping, as the ULID is also a viable locator. The use of alternate locators by the L3Shim is a triggered response, based on a network path unreachability signal.

3. Endpoint Identity

There are a number of options in the choice of an endpoint identity realm, including the use of existing addresses as an identity tokens, the use of distinguished (possibly non-routeable) addresses as tokens, or the use of tokens drawn from a different realm (such as use of a fully qualified domain name).

L3Shim uses the first of these options, and the endpoint identity for a host is one of the locator addresses that are normally associated with the host. The particular locator address selected to be the endpoint identity (or ULID) is specified in [RFC3484]. L3Shim does not mandate the use of distinguished addresses as identities, although the use non-routeable distinguished addresses in this context is described as an option in this approach.

The L3Shim approach defines the initial selector of the locator addresses pair is to be the same as the ULID pair. Accordingly, the initial state of the multi6 shim element is a null transform. This allows the initial establishment of a transport session without the requirement to perform a multi6 capability negotiation.

The choice of a locator as the endpoint identity for the upper protocol layers implies there is no impact in terms of implied changes to transport protocols or the upper level applications. Applications can continue to resolve fully qualified domain names to a set of addresses, and then open a session with the remote party specifying a selected address as the address of the remote party. The addresses used as source and destination identifiers can continue to be used in the context of pseudo-header checksums, session demultiplexing, packet reassembly contexts following fragmentation, IPSEC security associations, callbacks and referrals, all without alteration. The use in callbacks and referrals can be further generalised to the use of these address in the application payload. Irrespective of any subsequent change in the locator pair, the protocol stack above the Level 3 shim element will continue to

use the original ULID pair, and any use of these values in payloads will continue to match the endpoint identities.

4. Functional Decomposition

4.1 Establishing Session State

What form of token is passed to the IP layer from the upper level protocol element as an identification of the local protocol stack?

There is no requirement to change the conventional behaviour of the protocol stack. The upper protocol level may use a specified address as a source address, or the upper level may explicitly defer the selection of a source address to the IP level. Conventionally, the selected source address is the IP address of the outbound interface that the IP protocol will use to send the packet towards the destination address. In the case of an L3Shim-enabled stack, the source address selection function would need to consult a local state as to whether the destination address is associated with a currently active M6 state (interpreting the destination address as a ULID). In this case the selected source address, as seen by the upper level protocol stack element is the ULID of the stored state associated with the destination ULID. Otherwise the selected source address is a selected IP address from the set of addresses associated with the particular host interface that will be used to send the packet, as happens in a conventional IPv6 protocol stack.

What form of token is passed to the IP layer from the upper level protocol element as an identification of the remote session target?

The token passed to the IP layer as the ULID of the destination is the address of the destination host. If the initial identification of the remote host is via a domain name, then this approach assumes that there are a one or more locators held in the DNS. The local host to performs a name-to-address DNS lookup to obtain a set of locators (recorded in the DNS using AAAA resource records). The host then performs a selection from this set of locators and uses the selected address as the identification of the remote host. This implies no change to the conventional behaviour of the IP protocol stack element.

What form of token is used by the upper level protocol element as a endpoint identification mechanism for use within the application payload?

> There is no change to the existing behaviour in this approach. The upper level protocol element may use a domain name, or an IP address as an identification token.

Does the identity protocol element need to create a mapping from the upper level protocol's local and remote identity tokens into an identity token that identifies the session? If so, then is this translation performed before or after the initial session packet exchange handshake?

> In looking at the interface between the application level and the transport level of the protocol stack, there is no requirement to create a mapping between the upper level identifiers and the session identifiers, as the session identifiers are the same upper level identifiers. In looking at the interface between the transport and internet protocol stack elements, then the L3 Shim element has to check if there is an already established L3 Shim state that is associated with the ULIDs of the packet being sent. If so, then the translation from the ULID pair to the currently active locator pair is performed by the L3Shim protocol element. If not, then no state is created and no mapping is performed. This infers that an initial session packet exchange handshake is supported without the requirement to establish an identity to locator mapping state.

How does the session initiator establish that the remote end of the session can support the multi-homing capabilities in its protocol stack? If not, does the multi-homing capable protocol element report a session establishment failure to the upper level protocol, or silently fall back to a non-multi-homed protocol operation?

> The session initiator determines the ability of the remote end to support the L3Shim protocol via explicit negotiation. The L3Shim protocol will continue to operate in a conventional mode if the capability negotiation fails for L3Shim support. The nature of the communication exchange to determine the capability to use L3Shim support is not described in [ID.L3SHIM].

How do the endpoints discover the locator set available for each other endpoint (locator discovery)?

The mechanism is by explicit exchange of locator sets between the hosts. The L3Shim description does not describe the precise mechanism. Section 6 of [ID.L3SHIM] notes that once the initial capability exchange has completed "both ends know a set of locators for the peer that are acceptable as the source in received packets." This explicit exchange of locators is not necessarily aligned to multiple AAAA Resource records in the DNS.

What mechanisms are used to perform locator selection at each end for the local selection of source and destination locators?

The initial choice of source and destination locators matches the initial choice of upper level identifiers, namely the initial addresses used as the upper level identifiers. The remote address is obtained using conventional DNS lookup. The local address is based on an address selection from the addresses associated with the outbound interface, using the procedure described in [RFC3484].

What form of mechanism is used to ensure that the selected site exit path matches the selected packet source locator?

This is not described in the current L3Shim description.

4.2 Rehoming Triggers

What triggers are used to identify that a switch of locators is desirable?

The L3Shim documentation covers a number of options, but does not provide definitive answers to this question. The [ID.FAIL] notes four approaches: namely positive feedback from the upper level sessions, negative feedback from the upper level sessions, explicit reachability tests and ICMP error messages.
From the discussion in this draft it appears that negative feedback from upper layer transport sessions in the form of ACK timeouts is the preferred locator change trigger mechanism.

Are the triggers based on the end-to-end transport session and/or on notification of state changes within the network path from the network?

[ID.FAIL] argues that network path-based triggers, in the form of received ICMP errors messages are prone to spoofing, and should only be used "as a hint to perform an explicit reachability test". Triggers are based on explicit negative information being passed from an active transport session (ACK timeouts). There is also the possibility of using positive feedback from the transport sessions, where a timeout of positive indication is an indication of a reachability problem. In this case, as with ICMP, an explicit reachability test is required to confirm the indication of locator failure.

What triggers can be used to indicate the direction of the failed path in order to trigger the appropriate locator repair function?

The [ID.FAIL] description does not provide a description of detection of the failed path. The L3Shim approach attempts to treat path failure as a failure of the locator pair, rather than failure of a single locator, so the direction of the failure is not necessarily critical information in the search for a new functional pair.

4.3 Rehoming Locator Pair Selection

What parameters are used to determine the selection of a locator to use to reference the local endpoint?

The selection of a locator is based on the application of the tables as described in RFC 3484 [RFC3484]. The approach also allows local policy settings to place a preference for particular locator pairs. Selection of a specific locator pair is based on the successful outcome of a return reachability test between the two endpoints.

If the remote endpoint is multi-homed, what parameters are used to determine the selection of a locator to use to reference the remote endpoint?

Same as the previous response.

Must a change of an egress site exit router be accompanied by a change in source and / or destination locators?

This appears to be an area for further study. The situation is not explicitly addressed in the L3Shim documentation.

How can new locators be added to the locator pool of an existing session?

The explicit L3Shim capability negotiation allows the two endpoints to exchange a set of locators as part of the initial setup. This set is then tested, as required, using explicitly reachability tests when the endpoints are searching for a viable locator pair. The outcome of locator pair reachability tests are stored in an ageing local cache. This allows recently tested pairs that passed the reachability test to be used in preference to untested locator pairs.
[ID.FAIL] describes a set of abstract message exchanges for L3Shim locator set maintenance that includes explicit "add" and Delete" commands to allow a host to instruct the remote end to add or remove locators from its locator set.

4.4 Locator Change

What are the preconditions that are necessary for a locator change?

The preconditions necessary is that there has been a successful establishment of packets between the two hosts, L3Shim capabilities have been successfully negotiated and locator sets have been exchanged, and there is an explicit trigger for a locator change that has been generated by an active transport session. IN addition reception of a packet where the locator par is a member of the locator set for this host pair implies a remotely-triggered locator change.

How can the locator change be confirmed by both ends?

The approach proposed here is by using a return reachability test, where a host searches through locator pair selections until it receives an explicit acknowledgement of a poll.

What interactions are necessary for synchronisation of locator change and transport session behaviour?

As noted in [ID.FAIL], there is consideration that any locator change in the Layer 3 shim should trigger a notification to the transport layer protocol. In the case of TCP this notification would be used to trigger a resetting of the TCP congestion state to slow start, corresponding to the selection of a new network path.

4.5 Removal of Session State

How is identity / locator binding state removal synchronised with session closure?

As this is a layer 3 function there is no explicit concept of sessions. A L3 Shim mapping state needs to be maintained for as long as there is packet activity in either direction. The removal of state would most likely be associated with a removal eligibility condition associated with a packet activity timeout, and an eligible state removal pass being undertaken by the L3 Shim statement management module.

What binding information is cached for possible future use?

The L3 Shim state information is the association of a ULID pair with a set of local and remote locators. This information may require periodic refreshing with the exchange of locator sets with the remote host in order to ensure that the remote host is also maintaining a L3 Shim state, and the locator sets are synchronised.

5. Additional Comments

The approach of using a IP layer mapping between upper level endpoint identity and lower level locators has a number of specific issues that have yet to be fully specified in the L3Shim documentation. Some of these are listed here.

The signalling interface between the L3 Shim and the upper layers pf the protocol stack requires further consideration. The decision to initiate a L3Shim capability negotiation with a remote host may benefit from an explicit upper layer signal to the shim protocol element. In turn this could allow applications to signal a desire to initiate this capability negotiation at the start of an extended communication session. Equally, it may be of benefit for the upper level protocol to be ab;e to query the L3 state for a particular remote host, to

establish whether there has been a capability negotiation performed, and if successful, the current active locator and the full locator set.

It may also be useful to allow the upper level protocol to explicitly indicate that any form of L3 functionality should not be applied to this session. The implication of this functionality is that incoming packets need to provide some form of positive indication that the incoming locator pair should be mapped to an equivalent ULID pair, while packets without this indication should be processed in a conventional fashion with any L3 Shim packet header mapping. The L3 documentation suggests that some form of explicit tagging should be performed in the IPv6 Flow Id field, but further details have not been provided.

The potential use of unreachable ULIDs as the initial choice of ULIDs and the consequent requirement to undertake a reachable locator search, capability negotiation and establishment of a L3 Shim mapping state is mentioned in the L3 Shim documents, but at a relatively abstract level. This requires further consideration in terms of the potential failures, and the appropriate signalling to be passed back to the ULP in such cases.

The issue of ambiguity of demultiplexing may require further consideration. If there are multiple AAAA resource records in the DNS, or the resource records change over the lifetime of active communication, it is possible to have multiple L3 Shim states set up for the same remote host, with distinct ULIDs for the remote host. An incoming packet with a given locator pair will, according to the L3Shim documentation, need to use the locator pair as a lookup key into the L3 Shim state information to establish the associated ULID pair. In the case of multiple active ULIDs for the same remote host this lookup will result in multiple ULIDs.

The treatment of trigger conditions for locator change also requires further consideration. As noted in [ID.ARCH], different upper level transports may have different sensitivity requirements to locator triggers. When the mapping is performed on a host=-by-host basis rather than per transport session, there is a consequent requirement to balance the relative levels of sensitivity to locator change across all concurrently active transport session. It may be necessary to explore the concept of associating a L3 Shim state to particular transport sessions, allowing each session to establish its preferred level of sensitivity to network events

that may trigger a locator change.

The interaction between locator pair selection, local forwarding decision, site exit routers and packet ingress filters on the immediately adjacent upstream provider routers does not appear to be considered in the current l3 Shim documentation.

6. References

6.1 Normative References

[ID.ARCH] Huston, G., "Architectural Approaches to Multi-Homing for IPv6", Work in progress: Internet Drafts draft-ietf-multi6-architecture-03.txt, January 2005.

[ID.FAIL] , J., "", Work in progress: Internet Drafts draft-arkko-multi6dt-failure-detection-00.txt, 2004.

[ID.FUNC] , M. and J. , "Functional Decomposition of the M6 protocol", Work in progress: Internet Drafts draft-ietf-multi6-functional-dec-00.txt, 2005.

[ID.HBA] , M., "Hash Based Addresses (HBA)", Work in progress: Internet Drafts draft-ietf-multi6-hba-00.txt, 2004.

[ID.L3SHIM] , E. and M. , "Multihoming L3 Shim Approach", Work in progress: Internet Drafts draft-ietf-multi6-l3shim-00.txt, 2005.

[ID.REFER] , E., "Multi6 Application Referral Issues", Work in progress: Internet Drafts draft-ietf-multi6-app-refer-00.txt, 2005.

6.2 Informative References

[RFC3484] Draves, R., "Default Address Selection for Internet Protocol version 6 (IPv6)", RFC 3484, February 2003.

Author's Address

Geoff Huston
APNIC

Intellectual Property Statement

Disclaimer of Validity

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 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.

Copyright Statement

Acknowledgment

Funding for the RFC Editor function is currently provided by the Internet Society.