Network Working Group
Internet-Draft
Intended status: Informational
Expires: September 22, 2016

P. Lapukhov
Facebook
March 21, 2016

Deploying Identifier-Locator Addressing (ILA) in datacenter
draft-lapukhov-ila-deployment-00

Abstract

Identifier-Locator Addressing defined in [I-D.herbert-nvo3-ila] proposes using locator-identifier split in IPv6 address to realize workload mobility and network virtualization. This document describes how ILA can be implemented in datacenter using BGP as the control-plane protocol. In general, ILA could be built upon different control planes, and BGP is one particular instantiation. BGP is a well-known protocol, sufficient for small to medium size deployments, on scale of few millions of mappings. Defining more generic and scalable control plane is outside of scope of this document.

Table of Contents

1. Introduction

This document provides general guidelines for building an ILA-enabled datacenter using BGP [RFC4271] as the protocol for ILA mapping information dissemination. The reader is assumed to be familiar with the concepts defined in [I-D.herbert-nvo3-ila]. Reading on ILNP architecture defined in [RFC6740] is also recommended, but not needed for understanding of this document. ILA does not implement the full ILNP proposal, but it's based on the same idea, adapting it for datacenter use and employing simpler model for distribution of mapping information.

The full set of ILA benefits is realized in L3 switched (routed) datacenter networks, i.e. networks that do not rely on spanning Layer-2 domains across multiple network devices. Endpoint mobility made possible by ILA is one of the key benefits ILA brings to the datacenter networks. Combining ILA with fully routed network design allows for achieving the robustness of routed network with the flexibility of endpoint mobility. Some practical recommendations for building a fully-routed datacenter network could be found in [I-D.ietf-rtgwg-bgp-routing-large-dc] or [ROUTED-DESIGN].

While workload mobility could also be achieved in L3 switched networks by using "host-route" injection techniques, this has limited applicability, due to high stress put on the underlying routing system. The prefix needs to be removed, re-injected and propagated to all network devices every time an address moves.

ILA offers an alternative to "encapsulation" approaches, such as LISP ([RFC6830]), for realizing the endpoint mobility and network virtualization. Using simple address rewrites significantly reduces the processing overhead on the hosts, and makes various hardware and software network acceleration functions easier to implement. Furthermore, ILA keeps the underlying network fully visible to the applications that use ILA addresses, which makes network troubleshooting easier, as compared to the "encapsulation" approaches.

2. Terminology

This section defines some ILA-specific terminology that will be used through the document.

ILA domain: a collection of ILA hosts and ILA routers that collectively support ILA identifier mobility and network virtualization model. The ILA domain is assigned a single 64-bit IPv6 prefix known as SIR (Standard Identifier Representation, see [I-D.herbert-nvo3-ila]) prefix, which is made known to all hosts

and routers in the domain. This prefix is used to construct the complete 128-bit IPv6 addresses for ILA identifies found in the domain.

ILA host: network endpoint that is capable of accepting and originating traffic for ILA addresses using IPv6 packets. The host maintains its own local version of the ILA mapping table and has at least one ILA locator (64-bit prefix) assigned.

ILA router: network endpoint that is responsible for two main functions:

- Storing and disseminating the ILA mapping information within the ILA domain (NVA role per [I-D.ietf-nvo3-arch]).

- Serving as the gateway between the ILA-domain and non-ILA capable nodes, as well as the gateway for communicating with other ILA domains (NVE role per [I-D.ietf-nvo3-arch]).

ILA mapping table: The table for mapping identifiers to locators present in ILA host or ILA router. This table is updated either via BGP, or ILA redirection messages. ILA routers maintain authoritative copy of the table, while ILA hosts may have their own smaller view of the global mapping state.

Non-ILA host: network endpoint that is not aware of ILA addressing structure and does not participate in ILA address resolution.

Task: the unit of mobility in ILA domain. Each task is assigned an identifier unique within the ILA domain, which follows the task as it changes the hosts and, consequently, the locators. Implementation wise, the task can run within a container or a virtual machine, for example.

Tenant: owner of the tasks executed in the shared environment. All tasks that belongs to the same owner could be grouped and addressed together from the same identifier pool, thus creating simple hierarchy in the ILA address space.

Common Locator Address (CLA): Special ILA address constructed as <locator>::1 and identifying the physical host itself. This address is used to send and receive of the ILA redirect messages.

3. ILA deployment process

The ILA domain consists of the following components:

- L3 switched network that provides reachability among physical hosts, i.e. provides routing within the locator address space.

- ILA hosts, each assigned a unique /64 prefix reachable in the network. Hosts maintains its own local version of ILA mapping table.

- ILA routers, each injecting the domain's SIR prefix in the routed network and maintaining the full mapping table for the ILA domain. The routers could be implemented in software, or using specialized hardware appliances.

- Centralized BGP speaker nodes that peer with all of the ILA hosts and all of the ILA routers within the domain for the purpose of mapping information dissemination. ILA hosts and routers are also assumed to run the BGP processes.

Deploying ILA in datacenter requires multiple logical steps:

- Preparing the network. Assigning locator addressing to the hosts (servers) in the datacenter network and providing routed interconnection among the locator prefixes.

- Configuring ILA hosts and ILA routers. Each ILA domain requires a set of ILA routers to facilitate mapping function and provide connectivity to other ILA domains and the Internet. Each ILA domain is assigned a /64 SIR prefix, which scopes all identifiers in the domain. All ILA hosts and ILA routers within a domain are aware of the SIR prefix of this domain.

- Setting up ILA control plane. Configuring the BGP mesh for mapping information dissemination within the ILA domain and injecting the SIR prefix into routed network from the ILA routers to facilitate communications among the ILA domain and from / to the Internet. See [I-D.lapukhov-bgp-ila-afi] for definition of the corresponding BGP extension.

- Deploying an address management solution to coordinate allocation of ILA identifiers. In simpler cases, the addresses could be generated on each host individually, in ad-hoc fashion.

4. Preparing the network

This section provides overview of the network-related configuration needed for ILA.

4.1. Data-center network topology

For ease of reference, this document adopts the Clos topology described in [I-D.ietf-rtgwg-bgp-routing-large-dc] along with the terminology developed in that document.

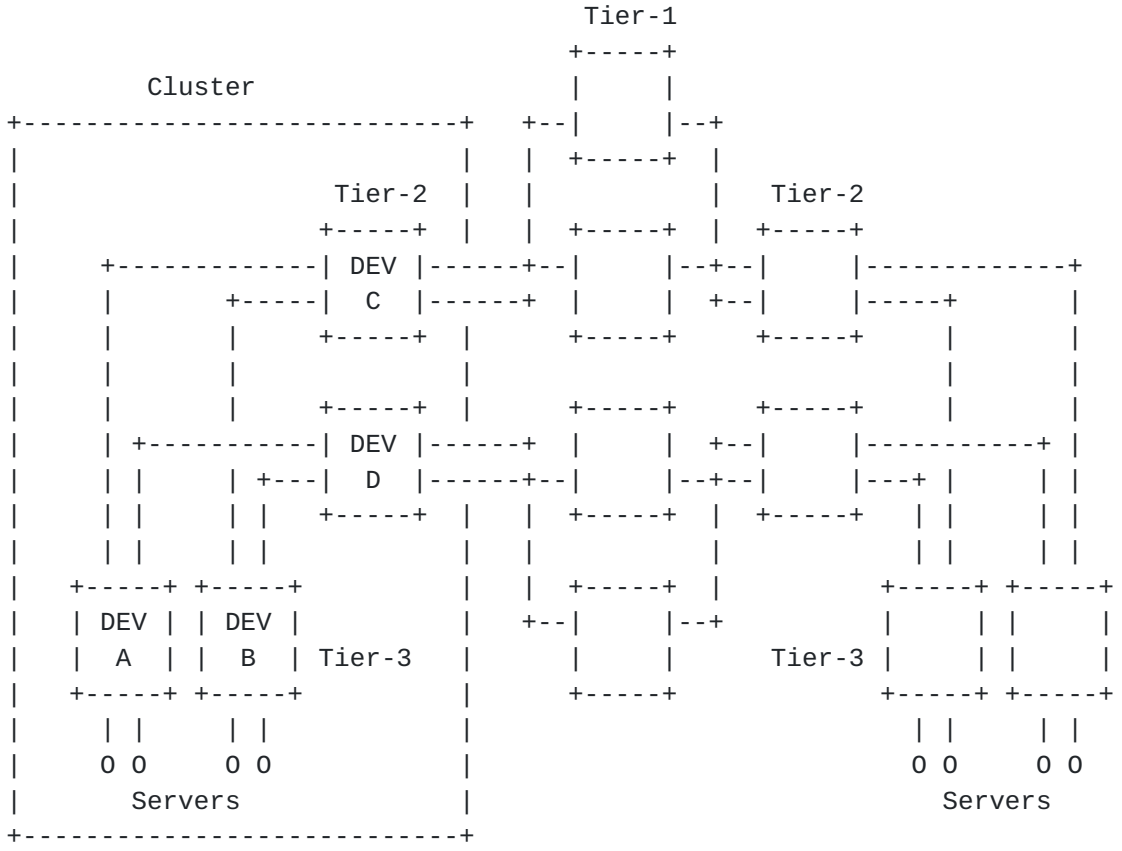


Figure 1: 5-Stage Clos topology

The network is partitioned hierarchically in three tiers, with tier numbering starting at the "middle" stage of the Clos network. The "middle" tier is often called as the "spine" of the network.

A set of directly connected Tier-2 and Tier-3 devices along with their attached servers will be referred to as a "cluster".

Tier-3 switches that connect the servers, and often referred to as "ToR" (Top of Rack) switches or simply "rack switches".

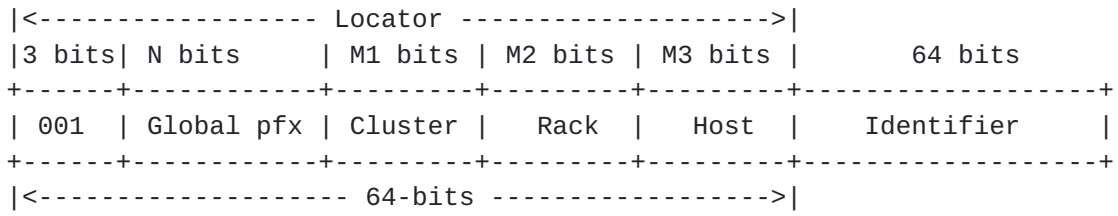
4.2. Configuring locator addressing

A mandatory prerequisite for ILA deployment is enabling IPv6 routing in the network. This could be done using either dual-stack IPv4/IPv6 deployment or IPv6-only deployments. This document assumes the

network has been already configured to forward IPv6 traffic. See [I-D.ietf-v6ops-dc-ipv6] for operational considerations on deploying IPv6 in the datacenter.

ILA requires every ILA host to have at least one 64-bit locator assigned. This means that every host (server) in the datacenter network needs to have at least one /64 IPv6 prefix configured on one of its interfaces (typically the internal loopback). These /64 prefixes could be either globally routable or unique local.

The use of the globally routable addressing scheme allows for deploying highly scalable hierarchical addressing scheme, and make the locators accessible from the Internet. The figure below illustrates the structure of a globally-routable locator:



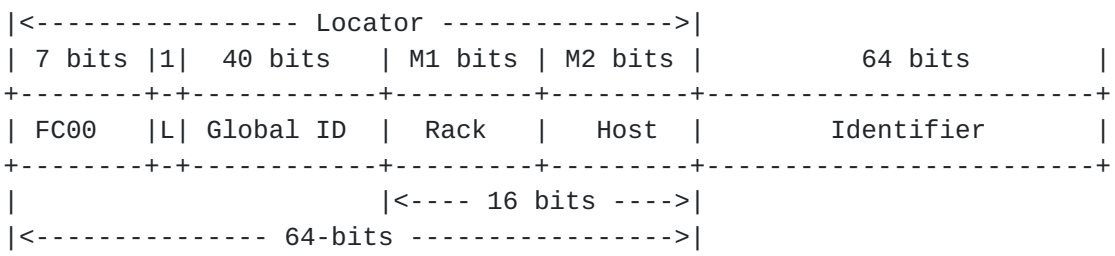
For example, a global /32 prefix (N=29) allows for sub-allocation of 2^32 locators. This sub-allocation could be done hierarchically, mapping to the tiers of network topology. Following the /32 example prefix:

Allocate 256 /64 prefixes per Tier-3 switch (M3 = 8 bits), which allows for up to 256 physical hosts in a rack, with /56 prefix assigned per rack.

Assuming 256 Tier-3 switches per cluster, one would allocate /48 per cluster (M2 = 8 bits).

This leaves room for 16-bits (64K) cluster per datacenter (M1 = 16 bits). This space could be further sub-divided if multiple network fabrics have been deployed.

The use of unique-local addressing for locators is more limiting in terms of available space, as it only offers 16-bits for sub-allocation. It does, however, have the benefit of ad-hoc allocation. This could work better for smaller deployment, e.g. allocating 10-bits to enumerate Tier-3 switches (physical racks of servers) and 6 bits to enumerate hosts within a rack. For instance, the address structure may look as following, here M1 = 10 bits and M2 = 6 bits.



In either case, the addressing scheme is hierarchical, allowing for simple route summarization logic and better routing system scaling (see [RFC2791]). This is especially important in case of IPv6, since contemporary datacenter network switches have smaller IPv6 lookup tables as compared to IPv4. Route summarization also requires certain network design changes to avoid packet black-holing under link failures. This problem gets more complicated in Clos topologies, and analyzed in more details in [I-D.ietf-rtgwg-bgp-routing-large-dc].

In greenfield deployments, each ILA host could be assigned the /64 locator prefix prefix during provisioning phase. There are multiple options to accomplish this:

- Assigning static link-local addresses to servers and statically routing /64 prefixes from Tier-3 switches to the servers over those link-local addresses. In this model, the operator would plan and pre-allocate per ILA-host prefixes beforehand, and configure the Tier-3 switches accordingly. From operational risk perspective, persistent routing loops may form due to static routing, if a server is not properly configured. Additionally, if the server is not present while the static route is configured on Tier-3 switch, packets destined to the corresponding /64 prefix will cause the switch to continuously generate IPv6 NDP packets ("gleaning"), which puts extra stress on the device's CPU.

- The servers may request the /64 prefix using IPv6 Prefix Delegation mechanism as defined in [RFC3633]. This allocation could be made "permanent" by proper DHCPv6 server configuration and ensuring the same prefix is always being delegated to the same server. The Tier-3 switch would act as DHCPv6 relay and will install the corresponding /64 IPv6 route dynamically. This approach addresses both the allocation and the routing problem, but makes the setup potentially more fragile operationally (reliance on additional protocol) and harder to debug (additional process involved).

- The server may run a routing daemon (e.g. BGP process) and inject the allocated /64 prefix into Tier-3 switch. The address

allocation in this case needs to happen by some other means. This is more suitable for ad-hoc ILA testing and small, rapid deployments.

The server itself may use one of the IPv6 addresses in /64 prefix for its own addressing, e.g. for remote access or management purposes. Alternatively, the server may obtain another IPv6 address from a different (non-locator) IPv6 address range allocated for the datacenter. This document proposes using <locator>::1 as the special identifier, naming it as "Common Locator Address" (CLA). Such choice of identifier make it easy to differentiate from regular identifiers. This identifier will be used as the source and destination identifier for the ILA redirect messages.

Route summarization for the locator prefixes is highly desirable to reduce the stress on the network switches forwarding tables and improve control-plane stability, and need to be implemented at least on Tier-3 switches. In simplest case, the switches could be statically preconfigured with the summary routes. These routes need to agree with the prefixes that are assigned to the servers, especially in the case when dynamic prefix injection is used. As a possible alternative, simple virtual aggregation could be employed, where hosts inject both the specific and the summary route, and installation of corresponding FIB entries is suppressed as per the rules defined in [RFC6769]. The latter approach does not improve the control plane scalability, but solves the issues with packet black-holing in presence of network summarization. It also requires the network hardware support, which may not be present.

In retrofitting scenarios, the servers are likely to already have 128-bit IPv6 addresses assigned, allocated from the datacenter address space, e.g. by using a single /64 prefix per Tier-3 switch. In this case, the additional locator prefix needs to be assigned in the same way as described above for greenfield deployments. The only difference is that the new prefix and the old server address may be allocated from different IPv6 address ranges.

5. Deploying ILA routers

ILA routers perform multiple functions within the ILA domain:

- Serve as the centralized store of the identifier-to-mapper information in the domain. The mappings are delivered to the ILA routers as described in Section 7.

- Act as the gateway between the ILA hosts and non-ILA capable hosts, e.g. the Internet.

The ILA hosts will send the packets destined to identifiers they don't have mappings for to the ILA routers initially to perform the ILA mapping resolution, and the hosts outside of the ILA domain will use the ILA routers for all communications with the domain. The ILA routers do not host any ILA identifiers themselves.

5.1. Configuration parameters

The ILA routers need the following configured for their operation:

- Regular, non-anycast 128-bit IPv6 address to connect the ILA router to the datacenter network.

- The /64 SIR prefix for the ILA domain, shared by all ILA routers. This prefix is advertised into the network in anycast fashion and "intercepts" all traffic destined from hosts outside of ILA domains to the identifiers in the domain. The prefix could be injected in "always-on" fashion, e.g. by using BGP injectors on ILA routers. This couples the ILA router's life-cycle with the prefix injection cycle. Other, more sophisticated schemes are possible, e.g. stopping injecting the prefix based if ILA router's resource utilization gets too high, but discussing their implementation is outside the scope of this document.

- Control-plane configuration, i.e. the IPv6 addresses of BGP route reflectors, and possibly some configuration for the local BGP process. This is discussed in more details in Section 7.

- Management settings, such as maximum rate of ILA redirect messages, and associated security attributes (e.g. the key pair used for message signing).

- A configuration flag that instructs the router whether the ILA redirect messages needs to be sent out. The ILA router does not receive ILA redirect messages, since it does not host any identifiers.

5.2. ILA router operation

Upon booting, the ILA router is first required to join the control plane mesh and learn of the mappings that exist in the ILA domain. It is also aware of the SIR prefix that is used within its domain. After the router has learned of the mappings, it may inject the anycast SIR prefix in the datacenter network and join the operational group of ILA routers.

When ILA router receives a packet with the upper 64-bits of the destination IPv6 address matching its configured SIR prefix, it performs the following:

- Checks if the source IPv6 address matches the local SIR prefix. If it does, the packet is coming from the ILA hosts in this router's ILA domain, and the ILA router should check if the source identifier has a matching locator, discarding the packet if there is none found, to prevent possible identifier spoofing attacks. This operation should be logged, with rate-limit applied to logging messages.

- Attempts to find the locator matching for the destination identifier found in the bottom 64-bits of the destination IPv6 address. If the mapping for destination identifier is not found, the original packet is dropped, and an ICMPv6 "Destination Unreachable" message, type "3" is sent back to the message originator. Otherwise, the router does the following:

  * Rewrites the SIR prefix in the destination IPv6 address with the new locator and forwards the packet back to the datacenter network.

  * If sending of ILA messages is permitted, the router sends the ILA redirect message back to the originator of the packet, by looking up the source identifier and finding the corresponding locator. The redirect informs the source of the actual destination locator. The redirect messages will be rate-limited to avoid sending ILA redirect for every incoming IPv6 packet.

For transit packets who's destination does not match the SIR prefix, the ILA router should discard the packets, as those are not supposed to be received by the ILA router.

If the source IPv6 address check reveals that the packet is not coming from the ILA domain the router belongs to (i.e. it does not match the local SIR prefix), the ILA router does not need to send back the ILA redirection message, but instead simply continue to forward the packet as if the locator for the destination identifier could be found. The ILA router will still send the ICMPv6 "Destinationa Unreachable" message for unknown mappings.

5.3. Scaling considerations

Due to high load and reliability concerns, the ILA domain needs multiple ILA routers. The simplest way to provide redundancy is by letting the ILA routers inject the /64 SIR IPv6 prefix into the

datacenter network in anycast fashion ([RFC4786]). This will allow to naturally use the datacenter network's Equal-Cost Multipath (ECMP) capabilities to distribute traffic among the ILA routers.

For redundancy purposes, the ILA routers would need to be spread across multiple physical racks in the datacenter. More ILA routers could be added incrementally to reduce the load and scale capacity horizontally, and join the operational ILA group in non-disruptive fashion, after they have learned the full mapping table for the ILA domain.

Use of anycast method does have some routing implications. For example, using the network described in Section 4.1 will result in ILA hosts preferring to use the ILA routers in the same cluster, since those are closer based on the routing metric. Thus, the network may not evenly spread their packets across all ILA routers in the datacenter. It is therefore possible that some ILA routers will receive more traffic than the others. This issue is specific to anycast routing, and not ILA in general.

6. Deploying ILA hosts

This section reviews the deployment considerations for the ILA hosts.

6.1. Configuration parameters

The ILA hosts need to be configured with the following:

- SIR prefix of the ILA domain.
- IPv6 addresses of the BGP route reflectors.
- The routable /64 locator assigned to the host.
- ILA mapping entries expiration time, to time out unused entries.
- Whether ILA redirection messages sending / receiving is enabled.

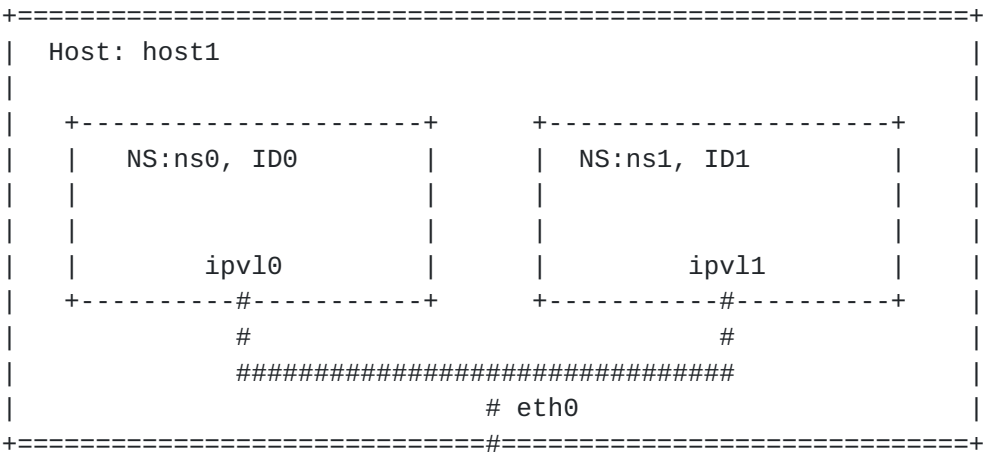
By disabling both the ILA mapping expiration time and sending of ILA redirect messages the host is effectively configured for the "push" ILA mapping distribution distribution mode (see Section 8). In this mode, the BGP (control plane) is assumed to populate all of the ILA mapping entries in response to the identifier move events.

6.2. Providing task isolation

In simplest case, the host only needs to implement the ILA address rewrite function and inform the tasks starting on the host of the ILA addresses they can use. However, it might be desirable to provide the tasks with strong networking isolation guarantees, i.e. making sure tasks are only allowed to use the IPv6 ILA address they have been allocated. For instance, with Linux operating system, this is possible by using the [LINUX-NAMESPACES] and [IPVLAN] techniques together.

Each task running on the host will be contained to its own networking namespace, and has the allocated ILA address bound to an interface that belongs to this namespace. The task would then only be able to bind to the single IPv6 ILA addresses delegated to the namespace.

With "ipvlan" technique, the packets arriving on physical host's NIC need to have their locator field adjusted before delivering to the task (the locator field is set to the /64 prefix assigned to the host). No additional routing lookups need to be performed on the physical host. On the egress path, all IPv6 lookups and rewrites happen in the default namespace, in Linux terminology. The figure below demonstrates a host with two tasks running, each in its own networking namespace. The namespace names are "ns0" and "ns1", and the corresponding task ILA identifiers are ID0 and ID1.



Tasks running in Linux namespaces with ipvlan

The use of "ipvlan"-like techniques is not strictly necessary. An alternative would be use the ILA host as a proper IPv6 router and treating the attached namespaces as hosts. This, however, has much higher performance overhead, due to multiple forwarding lookups that need to be done in the kernel.

6.3. ILA host operation

When ILA host boots up, it joins the control-plane mesh by peering with the BGP route-reflectors. It may learn the active ILA mappings from the BGP route reflectors, or may initially keep the ILA mapping table empty, depending whether "push" or "pull" distribution model has been selected.

When a tasks starts it will have an ILA identifier allocated, and the corresponding IPv6 address (built out of SIR prefix + the allocated identifier) bound to an interface within the networking namespace created for the task. The mapping is then propagated over BGP peering sessions to all ILA routers.

For outgoing packets, the ILA host performs the following:

- Matches the destination IPv6 address against the SIR prefix.
- If prefix matches, attempts to look-up the identifier portion of the address in the local ILA mapping table.
- If a match is found in ILA mapping table, rewrite the destination address and replace the SIR prefix with the actual locator.

For packets with destination IPv6 addresses no matching the SIR prefix, the usual forwarding rules apply. If no match is found for the destination, the packet is sent as is, and is expected to be delivered to the ILA routers, since those advertise the SIR prefix into the routing domain (without getting the locator portion rewritten - the packet has the SIR prefix for the locator).

For incoming packets, the ILA host should perform the following:

- Match their destination IPv6 addresses against the locator prefix (64 bits) of the host.
- If the destination address matches, deliver the packet to the corresponding namespace, based on the identifier portion.
- If the destination identifier in the incoming packet does not match any of the ILA mappings, and sending of ILA redirect message is enabled, the host sends an ILA redirect message back to the originator of the packet. The message will have an empty locator value, and informs the sender that the mapping it has for the identifier is no longer valid, erasing the corresponding entry in the sender's ILA mapping table.

Sending an ILA redirect message by the ILA host requires the host to translate the source identifier of the original message. Assuming that flow was likely bi-directional, the entry should be readily available in the local ILA mapping table. If not, the ILA redirect message will be routed toward the originator via the ILA routers, i.e. sent back with locator equal to the SIR prefix. It is possible that both source and destination identifiers of the flow have moved, resulting in mutual sending of ILA redirect messages, and temporarily falling back to using the ILA routers.

If the ILA mapping entry expiration time is set to non-zero, the unused ILA mapping entries will eventually be deleted. The entry expiration needs to be disabled if the mappings are learned in event-driven fashion via the BGP mesh ("push" distribution mode).

7. Using BGP as the ILA control plane

This section discusses the use of BGP for ILA mapping information dissemination. The choice of BGP is made to allow for easier integration of hardware appliance, e.g. network switches with extended functionality, where BGP is commonly used as the control plane. Furthermore, BGP itself offers a simple way of disseminating data and converging on a key-value mapping across multiple nodes in eventually consistent fashion, and has proven track record of use in the industry. Furthermore, use of BGP allows for leveraging the monitoring extensions developed for the protocol. For example, [I-D.ietf-grow-bmp] could be used to observe ILA mapping changes in the network using existing tooling.

7.1. BGP topology

Per the common practice, a group of BGP route-reflectors (see [RFC4456]) should be deployed and peered over IBGP with all hosts and routers in the ILA domain. The reflectors themselves would also be peered in "full-mesh" fashion to provide backup paths for mapping information distribution, e.g. in case if one of reflectors loses a session to a host. Those reflectors do not need to be in the data-path, but merely serve for the purpose of information distribution. The number of route-reflectors should be at least two, to allow for redundancy. See below sections for discussion of route-reflection settings.

It is possible to co-locate the BGP route-reflectors with the ILA routers. This saves on having additional nodes for the purpose of just BGP route-reflection, but puts extra memory and CPU stress on the ILA routers, and therefore is less desirable. Furthermore, it makes capacity-planning more difficult, and therefore is not recommended.

The route-reflectors are required to peer with potentially a very large number of ILA hosts, which may put scaling limits on the size of the ILA domain due to the overhead of maintaining large amount of BGP peering sessions. To alleviate this problem, the pool of ILA hosts may be split into "shards" and each shard would peer with a different group of route-reflectors. For example, the ILA domain may have four groups of route reflectors, each with four route-reflectors inside. The sixteen route-reflectors may then peer in a full-mesh fashion, to exchange the mappings they have received from the corresponding "shard" of the ILA domain. This method avoid the issues related to maintaining large amount of TCP sessions, but every BGP route-reflector is still required to maintain the full ILA mapping table.

In addition to ILA AFI/SAFI's, other AFI/SAFIs could be configured on BGP speakers, e.g. using [I-D.lapukhov-bgp-opaque-signaling] for opaque information dissemination in the ILA domain, e.g. to facilitate in distributed address allocation.

7.2. Any-to-any mapping distribution

In this mode, the ILA routers could act as IBGP route-reflectors [RFC4456] for all of the IBGP sessions they have, and relay the mapping information among the ILA hosts. This would allow the hosts to avoid initially sending packets to the ILA routers, at the expense of maintaining the ILA mapping table. Additionally, this allows for completely disabling the ILA redirect messages and using only the mapping information propagated by BGP.

7.3. Hub-and-spoke mapping distribution

Alternatively, BGP could be used to deliver the mappings from ILA hosts to ILA routers only. The hosts and the routers would establish IBGP peering sessions with the route-reflectors in hub-and-spoke fashion, with BGP reflectors being the hubs. The ILA router sessions will be configured as the "route-reflector clients" on the route-reflectors, while the ILA hosts sessions will be left as ordinary IBGP sessions. This will propagate all needed mappings to the ILA routers and allow them to properly redirect the hosts. The ILA hosts are responsible for withdrawing and announcing the mappings as they change.

8. Push vs pull mapping distribution modes

The default mode of operations in ILA is "pull" mode, where mappings are learned by the ILA hosts via ILA redirect messages. Effectively, the ILA mapping table fill process is reactive and driven by data-plane events. In some case, e.g. upon identifier move, this may

result in short periods of packet loss, while the sender receives the ILA redirect message and switches back to forwarding via the ILA routers. Furthermore, the use of ILA redirect messages requires security configuration to avoid message spoofing and cache poisoning attacks.

An alternative to "pull" mapping distribution on the hosts, is "push" mode, where all ILA hosts receive exactly the same mapping information as the ILA routers. In this case, the ILA message sending could be disabled in the ILA domain altogether. The "push" mode allows for proactive creation of the ILA mappings, and avoiding the packet loss, provided that the new mapping reaches the sending host before the destination identifier has moved. The trade-off here is the overhead of maintaining full mapping set on all ILA hosts.

For simplicity, this document recommends that all ILA hosts in the domain operate either in "push" or "pull" modes. In "push" mode the ILA mapping entries expiration needs to be turned off, along with sending of ILA messages. If an ILA host receives a packet for the ILA address it cannot map to locally, it is expected to send an ILA redirect message. If sending the ILA messages is disabled, the host must at least send an ICMPv6 "Destination Unreachable" message with code "3" - "Address Unreachable" to aid in debugging of missing mapping message. Notice that the ILA routers always operate in "push" mode, i.e. they only learn of mappings via the control plane exchange.

9. ILA address management

The ILA control plane and redirect messages perform mapping information dissemination, but the identifier allocation needs to be done separately. The address management process also depends on whether there is some hierarchy desired in the ILA namespace, e.g. if allocating a prefix per-tenant is needed.

9.1. Decentralized address management

In simplest case, each ILA host may independently allocate unique identifier per task when it first starts, and the task will retain it for the duration of its lifetime (see Appendix A of [I-D.herbert-nvo3-ila]). The chances of collision are very low given the 60-bit value of the identifier. The scheduler is responsible for starting and moving the task in the ILA domain. The tasks belonging to the same tenant may discover each other's addresses by some out-of-band signaling mechanism, e.g. a key-value store such as ([MEMCACHED]) or [ETCD] or use BGP for the same purpose as described in [I-D.lapukhov-bgp-opaque-signaling]. For instance, the task may

publish its own identifier, consisting of the tenant name and task name, mapped to the SIR address of the task.

Decentralized allocation is still possible even if the unit of address allocation is prefix, e.g. when multiple tenants are sharing the infrastructure, and unique VNID (see [I-D.herbert-nvo3-ila] for definition) is needed per tenant to build the 96-bit prefixes allocated to tenants from the /64 SIR prefix. Since the size of VNID space is rather small, generating random VNIDs becomes more prone to collision. In this case, decentralized address allocation schemes, such as one described in [RFC7695] could be used. These techniques require the ILA nodes to have some shared communication medium for nodes to "claim" the prefixes and avoid collisions. Once again, various distributed key-value stores could be used to accomplish this.

9.2. Centralized address management

In the case where high level of control is needed to allocate the addresses, e.g. per-tenant prefixes, centralized address management schemes could be used in the ILA domain. This could be either proprietary address allocation system, or system built on top of protocols such as DHCPv6.

9.3. Role of Task scheduler

The ILA domain needs a tasks scheduler responsible for resource allocation and starting of tenant's tasks on the ILA nodes. Defining functions of such scheduler is outside of scope of this document. At the very minimum, the scheduler would need agents running on every ILA host, participating in ILA address allocation, and communicating with the ILA control plane to publish and remove the mappings. Since it's the scheduler that is responsible for task movements, it makes sense for the scheduler to update the mappings in the domain.

The schedule needs some kind of API to interact with the BGP process on the box. Defining the exact API is outside of scope of this document, but as an option the scheduler may use a BGP session to inject prefixes into the BGP process running on the box.

10. ILA domain federation

In default operation mode, the ILA domains act as if the other domain is unaware of mappings that exist in another. It is possible to let the two domains exchange the mapping information and honor the ILA redirect messages from another domain by "joining" full or partial mapping tables of the two domains. For example, one can envision multiple compute clusters, each being its own ILA domain. In

standard ILA model, those clusters would need to communicate via the ILA routers only, increasing stress on the data-plane. To allow traffic flowing directly between the hosts in each cluster and bypassing the ILA routers, the ILA domains may exchange the mapping information, and program the ILA mappings in ILA hosts to facilitate direct paths.

Since each domain may re-use the 64-bit identifier space on its own, the use of SIR prefix is requires to make the identifiers globally unique. This requirement is easily fulfilled since the SIR prefix is required to be globally routable in the Internet.

To enable ILA domain federation, the BGP route-reflectors in each domain need need to be fully meshed and configured to use the "VPN-ILA" SAFI with "ILA AFI" (see [I-D.lapukhov-bgp-ila-afi]). This will propagate the mappings known to each route-reflector scoped with the SIR prefix of the local domain. If multiple domains are federated in this way, intermediate route-reflectors could be used, and filtering techniques such as described in [RFC5291] and [RFC4684] could be employed. The filtering may be further used to allow leaking of only select mappings, e.g. for the identifiers or tenants that carry lots of traffic.

If "push" distribution model is chosen with ILA domain federation, the ILA hosts will need to be configured to use "VPN-ILA" SAFI on their peering sessions with the BGP route reflectors. The ILA mapping entries lookup then need to be keyed both on the SIR prefix and the identifier to be resolved. Given the large volume of mappings that may exist in federated model, the "pull" model might become more preferable.

11. Operational Considerations

ILA introduces additional step in packet routing and thus adds more complexity to network troubleshooting process. At the same time, relative to the virtualization techniques that employ encapsulation and tunneling, ILA makes the underlying physical network fully visible to the tasks, and thus make tenant-driven troubleshooting simpler. This section discusses some operational procedures specific to ILA and the additional fault models that are possible in presence of ILA.

11.1. Operational procedures for ILA routers

ILA routers may be added/removed from the network at any time. Adding a router is commonly needed to scale the capacity of the ILA router group when peak loads increases. Adding an ILA router is non-disruptive procedure. It starts by configuring the ILA router to

peer with the BGP mesh to learn of all mappings in the domain. The use of BGP graceful restart (see [RFC4724]) would allow the new router to learn when all mappings have been advertised. At this time, the router may inject the SIR prefix, joining the operational group of ILA routers and start forwarding ILA traffic.

To gracefully take the ILA router out of service, it may be instructed to stop announcing the SIR prefix, or, in case of BGP, announce it with less preferable path attributes. This will allow the router to still accept and forward all in-flight packets, but will redirect the remaining packets toward the remaining ILA routers.

11.2. Multicast routing

Defining multicast routing and group membership dissemination is outside of scope of this document.

11.3. ILA mappint table complications

Every packet egressing from an ILA host and matching the SIR prefix is subject to lookup and translation in the local ILA mapping table. If entry is not found, the packet is forwarded to the ILA routers by the virtue of SIR prefix injected in the datacenter network. If the ILA router does not have the mapping, the ICMPv6 "Destination Unreachable" message will be sent back. There are few observations to make here:

- Packets egressing the ILA host and not matching the SIR prefix are routed as usual.

- ILA destinations that are not yet present in the ILA mapping table will be initially routed toward the ILA routers (e.g. the ILA routers will show up in the initial "traceroute" command output).

- In case of missing identifier mapping, it's the ILA router that informs the sender of this event via an ICMPv6 "Destination Unreachable" message.

Thus, the case of missing mapping is easily debuggable, though the "transition period" when the mapping is not yet in the ILA mapping table might confuse the operator using the "traceroute" command.

Worst kind of ILA mapping table malfunction would be presence of incorrect mapping, i.e mappings pointing to a non-existent or incorrect locator.

- Non-existent locator. This will route the packet through the network, and eventually result either in packet getting discarded

due to missing route or IPv6 NDP entry, or packet dropped due to routing loop and hop-limit expiration. In either case, the original sender may detect this condition either via reception of ICMPv6 "Destination Unreachable" messages, or by observing the output of the "traceroute" command. The ILA host may also be configured to make sure the identifiers fall within the known prefix range.

- Incorrect locator. In this case, the packet will be delivered to the wrong ILA host, that does not have the mapping for the identifier. Depending on whether the sending of ILA redirect messages is enabled on the host, two scenarios are possible:

  * The destination ILA host sends back an ILA redirect message with empty locator, informing the sender that mapping is invalid. The sender will invalidate the ILA mapping entry and switch over to forwarding via the ILA routers. The latter will either inform if of the new mapping, or send an ICMPv6 "Destination Unreachable" message back.

  * The destination ILA host is not configured to send the ILA redirect messages back. In this case, it simply responds with the ICMPv6 "Destination Unreachable" messages for the duration of time the sender keeps sending the packets using the incorrect mapping. The mapping needs to be flushed our updated by some external mean.

Next possible failure is dropped ILA redirect messages. However, given that the ILA redirect message sending process has no memory, the recipient will eventually receive one of them, or at least finish the communication via an ILA router.

11.4. ILA routers complications

The ILA routers serve as proxies for traffic entering the ILA domain, as well as temporary transit hops for traffic between the ILA hosts when they don't have matching mappings, in case if "pull" distribution model is utilized. The following operational observations apply:

- Traffic between the ILA domain and external world will necessarily flow asymmetrically. The packets toward the ILA hosts sent from the outside will always cross the ILA routers (see Section 10 for exceptions from this case) and traffic returning from the ILA hosts to the external world will flow directly, bypassing the ILA routers. This will show up in the outputs of the "traceroute" command running from sender and destination and showing asymmetric

paths. This being said, asymmetric traffic flows are very common in modern networks, and thus it should be a problem on its own.

- A failure of ILA router should be handled by re-balancing the load automatically by means of ECMP re-hashing in the network, and therefore should be mostly transparent to the ILA hosts, unless the load increases significantly after the failure. It is possible to have cascading failure and lose all ILA routers, or have them over-utilized. This event should be detected by external monitoring system, and be acted upon by adding more ILA routers to the domain - either automatically or manually. From troubleshooting perspective, the event will manifest itself via massive packet loss toward all hosts in the ILA domain.

- A malfunction of single ILA router (e.g. network interface card issue) would manifest itself in somewhat increased packet drop ratios for flows crossing the ILA routers, mostly traffic from external nodes. The more ILA routers the domain has, the harder to notice this ratio would be, since ECMP mostly spreads traffic evenly over all the ILA routers. This problem is more specific to ECMP behavior, and tooling exists to deal with it in datacenter networks.

To sum the above up - the health of ILA router is critical to the ILA domain functions, even if "push" model is employed and the ILA routers are used mostly for external communications. The ILA routers should be monitored closely for vital parameters, such as CPU and memory utilization, traffic rates on their network interfaces, and packet loss toward the ILA routers themselves.

12. Deployment Scenario Primer

Building upon the concepts presented above, this section provides a simple ILA deployment scenario.

- For locator addressing, unique-local addresses should be allocated, with 16-bit available for sub-allocation. This allows, for example, supporting 1024 (2^10) Tier-3 switches with 64 (2^4) servers under each Tier-3 switch. Using the Clos topology from section Section 4.1 one can build 32 clusters with 32 Tier-3 switches each.

- The hosts in the network could use BGP to peer with Tier-3 switches and inject their locator prefixes. It's desirable, but not necessary to configure the route summarization on the network switches, depending on the size of the deployment.

- Given the small to moderate scale of deployment, four IBGP route-reflectors could be deployed in the ILA domain, without the need for extra level of aggregation hierarchy. Each route-reflector will need to be configured to accept the BGP sessions from all of ILA hosts validated to be able to maintain thousands of peering sessions.

- The ILA hosts and routers should be configured with a single SIR prefix, and set up for "push" mapping distribution model, by disabling sending the ILA redirection messages. This will result in all ILA mappings propagated to all hosts and ILA routers via BGP. Each ILA host and router will need to be running a BGP process and peer with all four route-reflectors.

- The ILA routers will inject the SIR prefix using BGP into the datacenter network.

- For tasks running on ILA hosts, the globally unique 60-bit ILA identifiers should be allocated independently in pseudo-random fashion by the host that first starts the task.

- As task is moved, the task scheduler will update the mapping and publish it via BGP, forcing the ILA routers and ILA hosts to update their ILA mapping tables.

- ILA domain federation is not used, making every ILA domain communicate to each other via the ILA routers only.

13. IANA Considerations

None

14. Manageability Considerations

TBD

15. Security Considerations

The ILA introduces new security considerations described below.

15.1. ILA host security

If unsecured ILA redirect messages are used, the ILA hosts could be exposed to cache poisoning attacks. This calls for ILA redirect message authentication, e.g. by use of digital signatures, such as [ED25519]. This will also require to use some mechanism for propagation of public keys associated with the SIR prefix (the ILA

routers) and every locator in the domain, since the ILA redirect message could be sent by either.

To prevent tasks from every being able to sent packets directly bypassing the mapping layer, the ILA hosts should prohibit the task from sending packets toward the address space associated with the locators. Given that all locators will likely to belong to one large prefix, this could be accomplished by installing a single filtering rule on the ILA host.

15.2. ILA router security

TBD

15.3. Tenant security

ILA does not natively isolate the tenant traffic from each other, nor from the underlying physical infrastructure. In fact, this is seen as one benefit that makes many troubleshooting processes easier. The access control then become responsibility of the tenant itself, by employing traffic filtering rules. To this point, implementing filtering rules gets simpler if the tenant is allocated single prefix, as opposed to each task getting an unique identifier.

16. Acknowledgements

TBD

17. Informative References

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <http://www.rfc-editor.org/info/rfc4271>.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <http://www.rfc-editor.org/info/rfc4456>.

[RFC4684] Marques, P., Bonica, R., Fang, L., Martini, L., Raszuk, R., Patel, K., and J. Guichard, "Constrained Route Distribution for Border Gateway Protocol/MultiProtocol Label Switching (BGP/MPLS) Internet Protocol (IP) Virtual Private Networks (VPNs)", RFC 4684, DOI 10.17487/RFC4684, November 2006, <http://www.rfc-editor.org/info/rfc4684>.

[RFC5291] Chen, E. and Y. Rekhter, "Outbound Route Filtering Capability for BGP-4", RFC 5291, DOI 10.17487/RFC5291, August 2008, <http://www.rfc-editor.org/info/rfc5291>.

[RFC6740] Atkinson, RJ. and SN. Bhatti, "Identifier-Locator Network Protocol (ILNP) Architectural Description", RFC 6740, DOI 10.17487/RFC6740, November 2012, <http://www.rfc-editor.org/info/rfc6740>.

[RFC2791] Yu, J., "Scalable Routing Design Principles", RFC 2791, DOI 10.17487/RFC2791, July 2000, <http://www.rfc-editor.org/info/rfc2791>.

[RFC3633] Troan, O. and R. Droms, "IPv6 Prefix Options for Dynamic Host Configuration Protocol (DHCP) version 6", RFC 3633, DOI 10.17487/RFC3633, December 2003, <http://www.rfc-editor.org/info/rfc3633>.

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <http://www.rfc-editor.org/info/rfc4724>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <http://www.rfc-editor.org/info/rfc4760>.

[RFC4786] Abley, J. and K. Lindqvist, "Operation of Anycast Services", BCP 126, RFC 4786, DOI 10.17487/RFC4786, December 2006, <http://www.rfc-editor.org/info/rfc4786>.

[RFC6769] Raszuk, R., Heitz, J., Lo, A., Zhang, L., and X. Xu, "Simple Virtual Aggregation (S-VA)", RFC 6769, DOI 10.17487/RFC6769, October 2012, <http://www.rfc-editor.org/info/rfc6769>.

[RFC6830] Farinacci, D., Fuller, V., Meyer, D., and D. Lewis, "The Locator/ID Separation Protocol (LISP)", RFC 6830, DOI 10.17487/RFC6830, January 2013, <http://www.rfc-editor.org/info/rfc6830>.

[RFC7695] Pfister, P., Paterson, B., and J. Arkko, "Distributed Prefix Assignment Algorithm", RFC 7695, DOI 10.17487/RFC7695, November 2015, <http://www.rfc-editor.org/info/rfc7695>.

[I-D.herbert-nvo3-ila]
Herbert, T., "Identifier-locator addressing for network virtualization", draft-herbert-nvo3-ila-02 (work in progress), March 2016.

[I-D.ietf-rtgwg-bgp-routing-large-dc]
Lapukhov, P., Premji, A., and J. Mitchell, "Use of BGP for routing in large-scale data centers", draft-ietf-rtgwg-bgp-routing-large-dc-09 (work in progress), March 2016.

[I-D.lapukhov-bgp-opaque-signaling]
Lapukhov, P., Marques, P., and E. Nkposong, "Use of BGP for Opaque Signaling", draft-lapukhov-bgp-opaque-signaling-01 (work in progress), February 2016.

[I-D.ietf-v6ops-dc-ipv6]
Lopez, D., Chen, Z., Tsou, T., Zhou, C., and A. Servin, "IPv6 Operational Guidelines for Datacenters", draft-ietf-v6ops-dc-ipv6-01 (work in progress), February 2014.

[I-D.lapukhov-bgp-ila-afi]
Lapukhov, P., "Use of BGP for dissemination of ILA mapping information", draft-lapukhov-bgp-ila-afi-00 (work in progress), March 2016.

[I-D.ietf-grow-bmp]
Scudder, J., Fernando, R., and S. Stuart, "BGP Monitoring Protocol", draft-ietf-grow-bmp-17 (work in progress), January 2016.

[I-D.ietf-nvo3-arch]
Black, D., Hudson, J., Kreeger, L., Lasserre, M., and T. Narten, "An Architecture for Overlay Networks (NVO3)", draft-ietf-nvo3-arch-05 (work in progress), March 2016.

[ED25519] "Ed25519: high-speed high-security signatures", <https://ed25519.cr.yp.to>.

[ETCD] "coreos/etcd", <https://github.com/coreos/etcd>.

[MEMCACHED]
"Memcached", <https://memcached.org/>.

[ROUTED-DESIGN]
"High Availability Campus Network Design", 2008, <http://www.cisco.com/c/en/us/td/docs/solutions/Enterprise/Campus/routed-ex.html>.

[LINUX-NAMESPACES]
"Namespaces in operation, part 1: namespaces overview", 2013, <https://lwn.net/Articles/531114/>.

[IPVLAN] "IPVLAN Driver HOWTO", 2013, <https://github.com/torvalds/linux/blob/master/Documentation/networking/ipvlan.txt>.

Author's Address

Petr Lapukhov
Facebook
1 Hacker Way
Menlo Park, CA 94025
US

Email: petr@fb.com