Network Working Group
Internet-Draft
Intended status: Experimental
Expires: 2 February 2026

P. Jain
MIPS
S. Hooda
Cisco Systems
D. Srivastava
DataraAI

1 August 2025

LISP-Based Network for AI Infrastructure
draft-jain-lisp-network-ai-infra-01

Abstract

The LISP control plane provides the mechanisms to support both Scale-Up and Scale-Out backend networks within AI infrastructure. This document outlines how LISP can enable a unified control plane architecture that accommodates both scaling technologies, offering flexibility in deployment. This approach allows AI/ML applications—whether focused on training or inference—to operate efficiently on a converged infrastructure, supporting diverse deployment scenarios using the same underlying network fabric.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 RFC 2119 [RFC2119] RFC 8174 [RFC8174].

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

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."

This Internet-Draft will expire on 2 February 2026.

Table of Contents

1. Introduction

This document outlines the architecture and mechanisms for implementing a LISP-based unified control plane to converge both Scale-Up and Scale-Out backend networks within data center AI infrastructure. The proposed architecture leverages LISP protocol capabilities to enable a network infrastructure for AI, utilizing appropriate methods tailored to each scaling technologies. While LISP provides both the data plane and control plane mechanisms, this document focuses specifically on the unified control plane.

The decision to send a flow over either the Scale-Up or Scale-Out network is determined by the traffic's destination. For example, intra-POD traffic (which may be non-IP) is directed to the Scale-Up network, while inter-POD traffic (primarily IP-based) is routed through the Scale-Out network. The segmentation allows both scaling domains to function as distinct network domains, enabling flexible deployment options—including Scale-Up only, Scale-Out only, or hybrid configurations—without requiring changes to the underlying architecture.

The unified solution for Scale-Up and Scale-Out networks enables the extension of either or both domains within data center backend networks, with optimizations tailored for latency, bandwidth, and packet loss. These enhancements help meet the performance demands of diverse AI/ML workloads, including inference and training. A key advantage of this architecture is its flexibility: while most data centers require both Scale-Up and Scale-Out capabilities, some applications may not be agnostic to the underlying network and may assume exclusive use of one model. In such cases, LISP-based segmentation ensures that specific functionalities are confined to either the Scale-Up or Scale-Out domain, preserving architectural integrity while supporting security and application-specific requirements.

2. Definition of Terms

LISP Terminology: All terms related to the Locator/ID Separation Protocol (LISP)—including EID (Endpoint Identifier), RLOC (Routing Locator), xTR (Ingress/Egress Tunnel Router), MS/MR (Mapping System), and publication-subscription etc—are used as defined in the LISP specifications: [RFC9300], [RFC9301], and [RFC9437].

Accelerator (Acc): Specialized hardware designed to accelerate AI workloads. Examples include: GPU (Graphics Processing Unit), TPU (Tensor Processing Unit), NPU (Neural Processing Unit)

PoD (Point of Delivery / Performance-Optimized Datacenter): A modular unit within a data center that integrates compute, storage, and networking resources to deliver localized services. PoDs are designed for scalability and performance optimization.

Scale-Up Network: A segment of the data center backend network optimized for intra-PoD communication, typically among up to 1,024 accelerators. It supports ultra-low latency operations through direct load/store memory access between accelerators.

Scale-Out Network: A segment of the data center backend network designed for inter-PoD communication across clusters of thousands of accelerators. It leverages RDMA (Remote Direct Memory Access) for efficient, high-throughput data transfer between distributed compute nodes.

INCC (In-Network Collective Communication): A technique that offloads collective communication operations to network switches to reduce data movement and improve performance. These operations— such as AllReduce, Broadcast, Gather/Scatter, and AllToAll— are essential for synchronizing data across multiple compute nodes during distributed AI training.

3. AI Infra System Architecture

Figure 1 illustrates the backend network system architecture example of a data center AI infrastructure. The system comprises four PoDs (A, B, C, and D), each equipped with a Scale-Up accelerator fabric, interconnected via a Scale-Out fabric. Within each PoD, accelerators communicate over the Scale-Up network, while inter-PoD communication is handled by the Scale-Out network. Each PoD includes xTRs (Tunnel Routers) that register accelerator EIDs with the LISP Mapping System, enabling seamless connectivity across the AI infrastructure. This architecture supports AI applications for both inference and training.

The recommended deployment model maps intra-PoD traffic (which may be non-IP) to the Scale-Up network, and inter-PoD traffic (typically inter-subnet IP) to the Scale-Out network. This document assumes this traffic mapping model when describing packet flows.

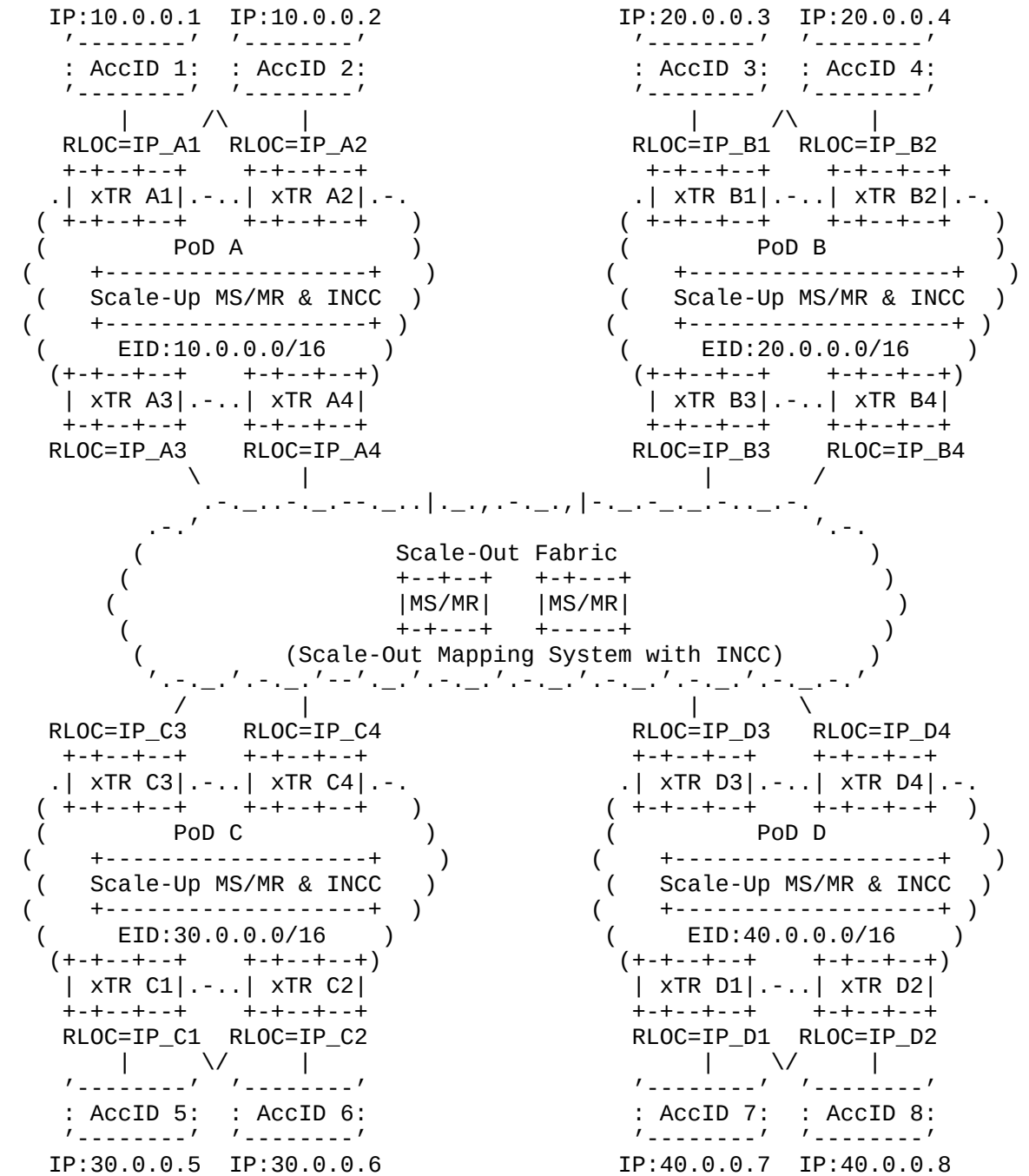

Figure 1: AI Infra System with converged Scale-Up and Scale-Out

To support this unified architecture, the control plane utilizes two primary types of EID-to-RLOC mappings for accelerators:

Scale-Up EID IID, AccID → RLOC IP: Used for the Scale-Up network. The AccID may be a non-IP identifier.

Scale-Out EID IID, AccIP → RLOC IP: The traditional LISP mapping, used for the Scale-Out with IP-based addressing.

4. Scale-Out Network Architecture

The Figure 2 illustrates the Scale-Out backend network architecture of Data Center's AI infrastructure.

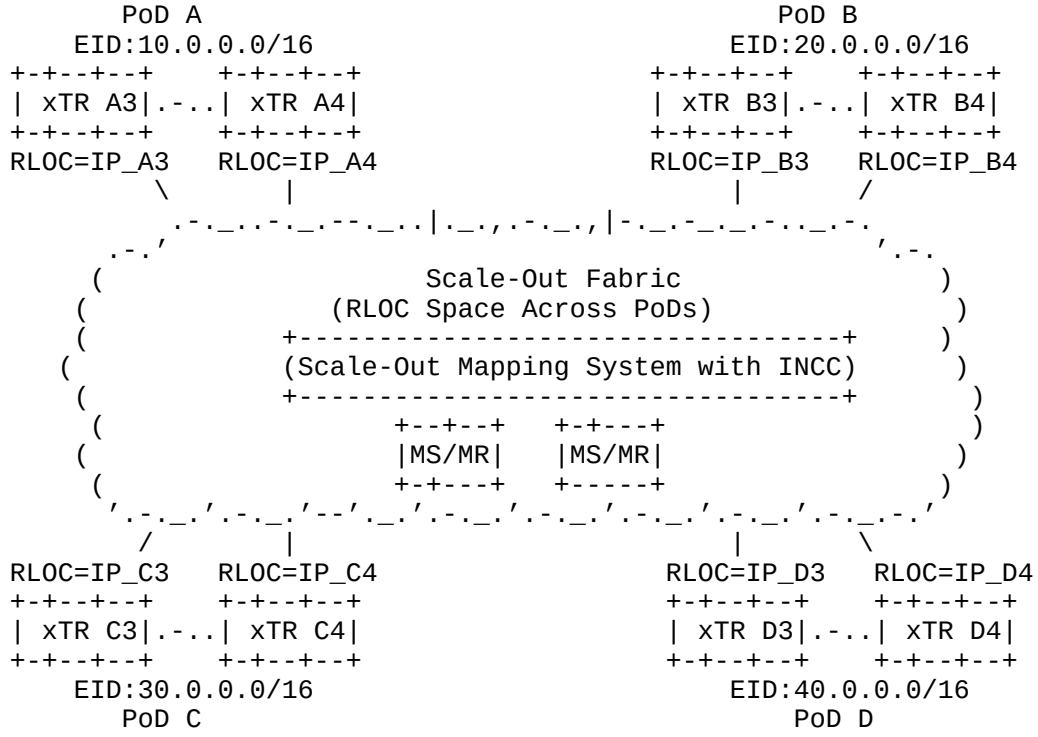

Figure 2: Scale-Out AI Infra Network Architecture

4.1. Scale-Out Registration and Fabric Paths

Each PoD is assigned one (or more) unique IP subnet, which is provisioned and registered with both the Scale-Out and Scale-Up mapping systems by the xTRs/pxTRs participating in the Scale-Out network, as illustrated in Figure 2. MAP-Register message format is as specified in [RFC9301] and [I-D.ietf-lisp-site-external-connectivity]. Additional Accelerator metadata MAY be encoded using vendor-specific LCAF types as defined in [RFC9306]

Once registered or de-registered, these subnets changes are published to all remote xTRs/pxTRs participating in the Scale-Out network and to all local xTRs of Scale-Up network within the PoD, following the publication mechanisms outlined in [RFC9437] and [I-D.ietf-lisp-site-external-connectivity]. MAP-Notify/Publication message format is as specified in [RFC9301] and [I-D.ietf-lisp-site-external-connectivity]. Additional Accelerator metadata MAY be encoded in the MAP-Notify or Publication message formats using vendor-specific LCAF types [RFC9306].

Upon receipt, the published subnets are added to the map-cache (routing table) of each xTR or pxTR of both Scale-Out and Scale-Up domains, establishing forwarding paths toward the xTRs/pxTRs responsible for the respective subnet.

Each accelerator’s IP Address within a PoD is also registered with the local mapping system (Scale-Up MS/MR in Figure 1) by its associated xTR (RLOC). These registrations are published to all xTRs within the PoD including remote xTRs/pxTRs, where they are added to the map-cache or routing table as the path to the registering xTR.

4.2. Scale-Out Subscription and Publication

When an accelerator connected via an ITR initiates communication with a remote accelerator, the ITR sends a Map-Request for the remote EID. The request includes the N-bit set in the EID-Record, indicating that the ITR wishes to be notified of any changes to the RLOC-set associated with that EID, as defined by the publish-subscribe mechanism [RFC9437]. MAP-Request/Subscription message format is as specified in [RFC9301] and [I-D.ietf-lisp-site-external-connectivity]. Additional Accelerator metadata MAY be encoded using vendor-specific LCAF types as defined in [RFC9306].

Any xTR or pxTR in PoD that had subscribed to updates for the Scale-Out EID— via the Scale-Out Mapping System (e.g., Scale-Out MS/MR) or Scale-Up Local Mapping System (e.g., Scale-Up MS/MR)— receives a Map-Notify from the mapping systems. This notification includes the updated RLOC-set (e.g., addition or removal of locators), as

specified in the Mapping Notification Publish Procedures in [RFC9437]. MAP-Notify/Publication message format is as specified in [RFC9301] and [I-D.ietf-lisp-site-external-connectivity]. Additional Accelerator metadata MAY be encoded in the MAP-Notify or Publication message formats using vendor-specific LCAF types [RFC9306].

Upon receipt, the published EID is added to the map-cache (routing table) of each xTR or pxTR, establishing forwarding paths toward the xTR(s) responsible for the respective EID.

When a destination accelerator is either undiscovered or deregistered in the Mapping System, it is treated as an Unknown Accelerator. In such cases, the Map-Server SHOULD respond to a Map-Request or subscription targeting the unknown accelerator with a Negative Map-Reply specifying the action "Drop" as in [RFC9301] and [I-D.ietf-lisp-site-external-connectivity].

4.3. Scale-Out Packet Flow

Inter-PoD traffic is encapsulated using Layer 3 (L3) encapsulation, following the procedures defined in [RFC9300], [RFC9301], and [RFC9437]. It is assumed that the accelerator in PoD A is aware of the EID (IP address) of the destination accelerator in PoD D— this may be obtained via packet inspection, address resolution, or provisioning. The following example illustrates a unicast packet flow and the associated control plane operations for the topology shown in Figure 1, where an accelerator in PoD A communicates with an accelerator in PoD D:

Fabric paths across PoDs are established as described in Section 4.1.

Within a PoD, each accelerator (Scale-Out EID eg 40.0.07) is also registered with the local mapping system (Scale-Up MS/MR in Figure 1) by its associated xTRs (RLOCs IP_D1 & IP_D2) as described in Section 4.1. These registrations (mapping 40.0.0.7->IPD1 & IPD2) are published to all xTRs/pxTRs within the PoD D, where they are added to the map-cache or routing table as the path towards the registering xTRs D! & D2 for Accelerator 40.0.0.7. Accelerators, unknown or external to PoD are registered as pETRs to local mapping system, as defined in [I-D.ietf-lisp-site-external-connectivity].

Accelerator 1 in PoD A sends an IP packet with source address 10.0.0.1 and destination address 40.0.0.7. Since the destination lies in a different subnet, the local xTR in PoD A forwards the packet based on its map-cache/routing table to pxTR A3, which acts as the default gateway for the PoD. pxTR A3 then forwards the packet to xTR D3, using its map-cache entry that contains the subnet information for PoD D, as published by the Scale-Out mapping system described in section 4.1.

xTR/pxTR D3 performs a Layer 3 lookup in its local map-cache for the destination IP 40.0.0.7. Since Accelerator 7 is registered with the local mapping system and is published in PoD D, xTR D3 has a valid map-cache entry pointing to xTR D1 and D2 (with RLOCs IP_D1 and IP_D2).

xTR D3 encapsulates the packet using LISP, setting the destination RLOC to either IP_D1 or IP_D2, depending on the load-balancing or redundancy policy.

4.4. Scale-Out Path Change

When the Publish/Subscribe mechanism [RFC9437] is used, the signaling flow to manage accelerator path changes (due to any failure, congestion etc) proceeds as follows:

Registration: Upon attachment of Accelerator 7 to PoD D, the local ETR D1/D2 updates its local database with the mapping for the EID <IID1, 40.0.0.7>. The ETR then sends a Map-Register message to the local mapping system, registering its RLOCs (e.g., IP_D1, IP_D2) as locators for the EID. The mapping system is updated with this EID-to-RLOC association.

First Communication Request/Subscription: When Accelerator 1 connected via an ITR A1 initiates communication with a remote Accelerator 7, the ITR sends a Map-Request for the remote EID. The request includes the N-bit set in the EID-Record, indicating that the ITR wishes to be notified of any changes to the RLOC-set associated with that EID, as described in section 4.2.

Deregistration: When Accelerator 7 is detached, the mapping system receives a deregistration message for the EID <IID1, 40.0.0.7> from ETR D1. It then sends a Map-Notify to ETR D1 to confirm the deregistration. ETR D1 subsequently removes the local mapping entry and ceases to advertise the EID.

Notification/Publication: Any xTR or pxTR in PoD D (or elsewhere) that had subscribed to updates for the EID <IID1, 40.0.0.7>— typically via the local mapping system (e.g., Scale-Up MS/MR)— receives a Map-

Notify from the mapping system. This notification includes the updated RLOC-set (e.g., addition or removal of locators), as specified in the Mapping Notification Publish Procedures in [RFC9437].

Map-Cache Update: Upon receiving the Map-Notify, the subscribing ITR (e.g., xTR D3) updates its local map-cache or routing table to reflect the new RLOC-set for the EID (Acccelerator 7/40.0.0.7). For example, if IP_D1 is removed for Accelerator 7 (40.0.0.7), xTR D3 stops forwarding traffic to that locator (IP_D1), ensuring accurate and up-to-date routing behavior.

4.5. Deployment Considerations

4.5.1. Scale-Out Segmentation and INCC

LISP Scale-Out segmentation is based on the use and propagation of Instance-IDs (IIDs), which are treated as part of the EID in control plane operations. The encoding format for IIDs is defined in [RFC8060]. Instance-IDs are unique within a given Mapping System and MAY be used to distinguish between Scale-Up and Scale-Out domains.

A key aspect of Scale-Out segmentation is the ability to associate In-Network Collective Communication (INCC) groups with specific IIDs. In this model, an INCC domain— functionally equivalent to a Virtual Forwarding (VRF) instance— can be mapped to a corresponding IID representing a Scale-Out domain. Alternatively, each INCC group may be mapped in a 1:1 relationship with a unique Scale-Out segment instance.

This use of Instance-IDs enables support for multiple Scale-Out segments, similar to extended VRFs or multi-VPN, as described in [I-D.ietf-lisp-vpn].

4.5.2. Scale-Out Mappings

When an accelerator is attached or detected in an ETR that provides Scale-Out services and path change, Scale-Out Mappings are registered to the mapping system with the following structures:

* The EID 2-tuple (IID, Acc-IP Address) with its binding to a corresponding ETR locator set (RLOC IP Address).

* The EID 2-tuple (IID, Acc-IP Subnet) with its binding to a corresponding pETR locator set (RLOC IP Address).

The registration of these Accelerator/Subnet EIDs MUST follow the LCAF format as defined in [RFC8060] with the specific EID record as specified in [RFC9301] and can be used with pETR registration [I-D.ietf-lisp-site-external-connectivity].

4.5.3. Scale-Out Mapping System (MS/MR)

Scale-Out (across PoDs) Mapping System also uses services from Scale-Up Mapping Systems (within PoDs) to establish end to end Scale-Out network of Accelerators.

The interface between xTRs/pxTRs and the Mapping System follows the procedures defined in [RFC9301] and [I-D.ietf-lisp-site-external-connectivity]. The addition and removal of subnets are handled through pxTR registration/deregistration and publication processes, as described in Section 4.1 and Section 4.5.2. Mapping System MAY be implemented as a distributed mapping system to avoid single point of failure.

To support system convergence following an accelerator or subnet path change, the local Mapping System (Scale Up MS/MR) MUST also send a Map-Notify message to the full RLOC set—including all relevant pxTRs—within the PoD where the affected EID was last registered. This notification is triggered upon receiving a registration update for that specific accelerator or subnet EID. The Map-Notify serves to indicate the unavailability or change in the accelerator’ s path, as detailed in Section 4.3.

4.5.4. Scale-Out Unknown Accelerators

When a destination accelerator is either undiscovered or deregistered in the Mapping System, it is treated as an Unknown Accelerator. In such cases, the Map-Server SHOULD respond to a Map-Request or subscription targeting the unknown accelerator with a Negative Map-Reply specifying the action "Drop" as per [RFC9301] and [I-D.ietf-lisp-site-external-connectivity].

Alternatively, the forwarding plane may be configured to default to the "Drop" action for Unknown Accelerators, thereby suppressing any forwarding attempts toward unregistered or unreachable destinations

5. Scale-Up Network Architecture

Scale Up network architecture is as shown in Figure 3. This section uses PoD C & PoD D to describe the details.

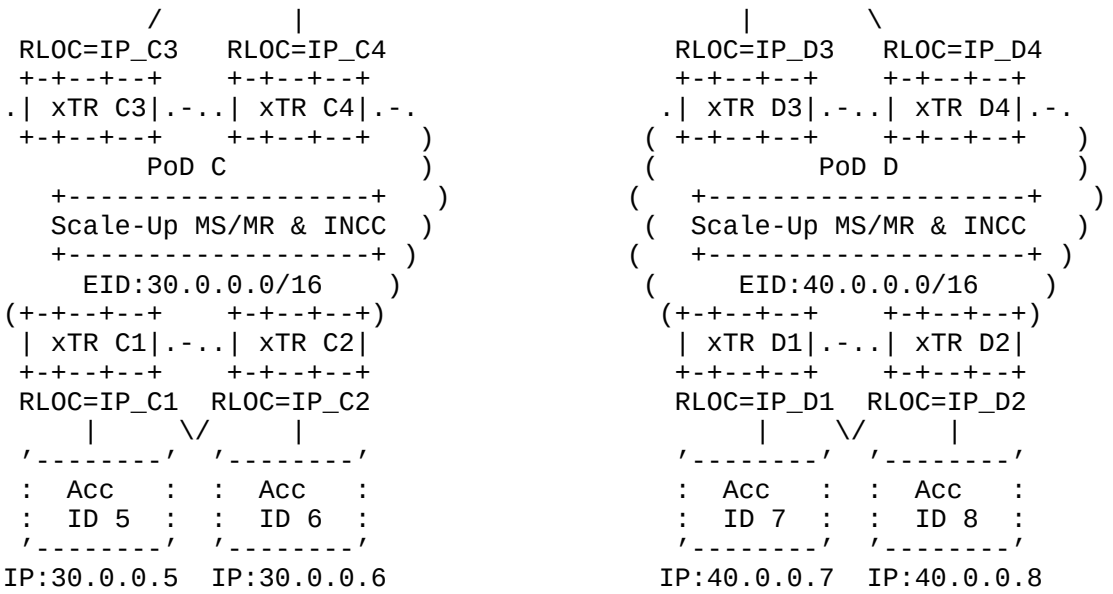

Figure 3: Scale-Up AI Infra Network Architecture

5.1. Scale-Up Registration

Each accelerator within a PoD is registered by its associated xTR(s) (RLOCs) using its AccID as the Scale-Up EID, with the registration sent to the local mapping system (Scale-Up MS/MR shown in Figure 3). The AccID MAY be a non-IP identifier and MAY be encoded using the LISP Name Encoding format defined in [RFC9735]. MAP-Register message format is as specified in [RFC9301]. Additional Accelerator metadata MAY be encoded using vendor-specific LCAF types as defined in [RFC9306].

The local mapping system then publishes the EID-to-RLOC mappings (i.e., AccID to xTR(s)) to all subscribed and authorized xTRs within the PoD, in accordance with [RFC9437]. MAP-Notify/Publication message format is as specified in [RFC9301]. Additional Accelerator metadata MAY be encoded in the MAP-Notify or Publication message formats using vendor-specific LCAF types [RFC9306]. These published mappings are subsequently added to the map-cache or routing table of remote xTRs/pxTRs within the same PoD, establishing the forwarding path to the registering xTR(s).

5.2. Scale-Up Subscription and Publication

When an accelerator connected via an ITR initiates communication with a remote accelerator, the ITR sends a Map-Request for the remote EID. The request includes the N-bit set in the EID-Record, indicating that the ITR wishes to be notified of any changes to the RLOC-set associated with that EID, as defined by the publish-subscribe mechanism [RFC9437]. MAP-Request/Subscription message format is as specified in [RFC9301] and [I-D.ietf-lisp-site-external-connectivity]. Additional Accelerator metadata MAY be encoded using vendor-specific LCAF types as defined in [RFC9306].

Any xTR or pxTR in PoD that had subscribed to updates for the EID —via the Scale-Up Mapping System (e.g., Scale-Up MS/MR)— receives a Map-Notify from the mapping system. This notification includes the updated RLOC-set (e.g., addition or removal of locators), as specified in the Mapping Notification Publish Procedures in [RFC9437]. MAP-Notify/Publication message format is as specified in [RFC9301] and [I-D.ietf-lisp-site-external-connectivity]. Additional Accelerator metadata MAY be encoded in the MAP-Notify or Publication message formats using vendor-specific LCAF types [RFC9306].

Upon receipt, the published EID is added to the map-cache (routing table) of each xTR or pxTR, establishing forwarding paths toward the xTR(s) responsible for the respective EID.

When a destination accelerator is either undiscovered or deregistered in the Mapping System, it is treated as an Unknown Accelerator. In such cases, the Map-Server SHOULD respond to a Map-Request or subscription targeting the unknown accelerator with a Negative Map-Reply specifying the action "Drop" as per [RFC9301] and [I-D.ietf-lisp-site-external-connectivity]..

5.3. Scale-Up Packet Flow

Scale-Up packet (could be non-IP) utilizes Scale-Up technology (e.g., PCIe or Ethernet) and is encapsulated accordingly. This section illustrates an example of Scale-Up unicast packet flow and the associated control plane operations, based on the topology shown in Figure 3. In this scenario, Accelerator 7 in PoD D communicates with Accelerator 8, also located in PoD D. It is assumed that Accelerator 7 is aware of Accelerator 8’ s AccID (e.g., learned via packet exchange, dynamic resolution, or a management interface).

Each Accelerator within a PoD is registered by its xTRs (RLOCs) using its AccID (e.g., AccID 7, AccID 8) as the EID, as described in Section 5.1.

If a path is not pre-established (e.g. default or during provisioning), when Accelerator 7 (connected via xTR D1) initiates communication with Accelerator 8 (connected via xTR D2), ITR D1 issues a Map-Request for Accelerator 8. Following the Mapping Request Subscribe Procedures defined in [RFC9437], the Map-Request includes the N-bit set on the EID-Record to ensure the ITR is notified of the mapping and any RLOC-set changes for the Accelerator.

The local mapping system publishes all the AccID-to-xTR(s) mappings (EID-to-RLOC(s)) to subscribing xTRs/pxTRs. These subscribing xTRs/ pxTRs then establish a path to the registering xTR(s) within the PoD, as outlined in Section 5.1.

When Accelerator 7 sends a Scale-Up packet, it includes the destination AccID 8 and source AccID 7, in accordance with [RFC9300] and [RFC9301].

ITR D1 performs a lookup in its local map-cache for the destination AccID 8.

Since ITR D1 already has the EID-to-RLOC mapping for AccID 8 (pointing to xTR D2), it encapsulates all subsequent packets to AccID 8 using destination RLOC IP_D2 and source RLOC IP_D1.

5.4. Scale-Up Path Change

This section describes the mechanism for handling path changes of an accelerator within a PoD due to failure, upgrade, congestion etc, while maintaining uninterrupted communication among accelerators connected via multipath in the same Scale-Up domain. The mechanism ensures fast convergence of the Scale-Up network when an accelerator' s path changes. Updates to ITR map-caches are managed using the Publish/Subscribe mechanisms defined in [RFC9437]. The following steps outline the signaling and packet flow when the link between Accelerator 8 and xTR D2 (as shown in Figure 3) becomes unavailable due to congestion, link failure, or other issues:

Initially, when Accelerator 7 (connected via ITR D1) establishes communication with Accelerator 8, ITR D1 issues a Map-Request for Accelerator 8. In accordance with the Mapping Request/Subscribe Procedures defined in [RFC9437], the Map-Request includes the N-bit set on the EID-Record to enable notification of any RLOC-set changes for Accelerator 8.

When Accelerator 8 experiences a path change within PoD D (e.g., the link to xTR D2 becomes unavailable or congested), ETR D2 removes the local mapping for Accelerator 8's EID IID, AccID 8. It then sends a Map-Deregister message to the local mapping system, deregistering RLOC IP_D2 as a locator for that EID.

Upon receiving the deregistration, the mapping system updates the locator set for Accelerator 8's EID by removing IP_D2 and sends a Map-Notify back to ETR D2. ETR D2 then deletes the mapping from its local database and ceases registration for IID, AccID 8.

Any ITR or PiTR participating in the same Scale-Up domain (associated with IID) that was previously encapsulating traffic to AccID 8 would have subscribed to receive updates on RLOC-set changes. The local mapping system publishes the updated locator set to these subscribers by sending Map-Notify messages, as defined in the Mapping Notification Publish Procedures in [RFC9437].

Upon receiving the Map-Notify, the ITR updates its local map-cache for EID IID, AccID 8. Once the cache is updated, traffic is redirected and tunneled to the new xTRs (e.g., xTR D1), and traffic via xTR D2 is halted.

5.5. Deployment Considerations

5.5.1. Scale-Up Segmentation and INCC

Similar to Scale-Out segmentation, LISP Scale-Up segmentation is based on the propagation and use of Instance-IDs (IIDs), which are treated as part of the EID in control plane operations. The encoding of Instance-IDs is defined in [RFC8060]. These IIDs are unique within a Mapping System and may be used to distinguish between Scale-Up and Scale-Out domains.

A key aspect of Scale-Up segmentation is the potential mapping of INCC groups to Instance-IDs. In this context, an INCC Domain—functionally equivalent to a VRF as a forwarding context—can be mapped to an IID representing a Scale-Up domain. Alternatively, an INC group may be mapped directly in a one-to-one relationship with a Scale-Up segment instance.

Instance-IDs enable support for multiple Scale-Up segments, similar to extended VRFs or multi-VPN, as described in [I-D.ietf-lisp-vpn].

5.5.2. Scale-Up Mappings

When an accelerator is attached to or detached from an ETR providing Scale-Up services, a corresponding Scale-Up EID is registered or deregistered with the mapping system. The Scale-Up mapping follows this structure:

EID Tuple: The Endpoint Identifier is represented as a 2-tuple (IID, AccID), where:

AccID may be a non-IP identifier. If the AccID is non-IP based, it may be encoded using the mechanisms described in [RFC9735].

The structure of the Accelerator EID record adheres to the format defined in [RFC9301].

The AccID is bound to a locator set consisting of one or more IP RLOCs.

5.5.3. Scale-Up Mapping System (MS/MR)

The interface between xTRs and the Mapping System is defined in [RFC9301]. All accelerators are registered with the local Scale-Up Mapping System. Mapping System MAY be implemented as a distributed mapping system to avoid single point of failure.

To support rapid system convergence following a path change, the Map-Server MUST send a Map-Notify to the entire RLOC set within the PoD that last registered the same EID, as well as to any xTRs in the PoD that have subscribed to that EID. This Map-Notify serves to track changes in the path of Accelerator EIDs, as described in Section 5.3.

5.5.4. Scale-Up Unknown Accelerators

When a destination accelerator is either undiscovered or deregistered in the Mapping System, it is treated as an Unknown Accelerator. In such cases, the Map-Server SHOULD respond to a Map-Request or subscription targeting the unknown accelerator with a Negative Map-Reply specifying the action "Drop".

Alternatively, the forwarding plane may be configured to default to the "Drop" action for Unknown Accelerators, thereby suppressing any forwarding attempts toward unregistered or unreachable destinations.

5.5.5. IP Forwarding of Scale-Up Traffic

Providing non-IP extensions to cloud platforms is not always feasible. As a result, ip/subnets might need to be used and extended using Layer 3 (L3) to support intra PoD traffic as well.

6. Multihoming & Multipaths

Multihoming support relies on the mechanisms defined in [RFC9300] and [RFC9301] to enable LISP-based multihoming for accelerators within a backend network. To illustrate the multihoming packet flow, this section references Figure 3. For example, in Figure 3, xTRs D1 and D2 within PoD D provide multihoming services for Accelerators 7 and 8.

6.1. Multihomed Accelerators Registration

The Site-ID, as defined in [RFC9301], serves as an identifier for logically grouping multiple xTRs that provide multihoming within a Scale-Up domain (e.g., a PoD). All EID-to-RLOC mappings from ETRs in a multihomed Scale-Up PoD MUST be registered with the corresponding Site-ID (e.g., PoD ID) by setting the 'I' bit in the Map-Register message.

6.2. Multihoming xTRs/RLOCs Merging

Supporting multihoming requires that participating xTRs discover one another and implement multipath forwarding procedures. This is achieved through the registration of a common accelerator EID by all participating xTRs. Each registration includes the PoD ID as the Site-ID, indicating the PoD in which multihoming is being provided. The Mapping System merges these registrations and notifies all participating xTRs with the aggregated locator set.
Using Figure 3 as a reference, the xTR discovery process in a multihomed Scale-Up group proceeds as follows:

xTR D1 registers the EID "POD-D-AccID-7" with its locator set containing IP_D1.

The Map-Server creates a mapping entry: EID ("POD-D-AccID-7") → RLOC (IP_D1), and sends a Map-Notify to xTR D1 with this mapping.

xTR D2 then registers the same EID "POD-D-AccID-7" with its locator set containing IP_D2.

The Map-Server merges this new registration with the existing one, resulting in: EID ("POD-D--AccID-7") → RLOC {IP_D1, IP_D2}. It then sends a Map-Notify to both xTR D1 and xTR D2 with the updated locator set.

Whenever an xTR joins or leaves a multihoming group, the Map-Server MUST send an updated Map-Notify to all remaining participating xTRs to ensure they maintain an accurate and synchronized view of the locator set. As a result, all participating xTRs maintain an up-to-date view of the multihomed group, enabling coordinated multipath forwarding.

6.3. Multihoming/Multipath forwarding

In a PoD, both Scale-Out and Scale-Up xTRs can be used to provide multihomed access and forward traffic to and from remote PoDs or accelerators. Unicast traffic is typically load-balanced or sprayed across the multiple xTRs that have registered the accelerator's EID. In multicast scenarios, only the designated Scale-Up xTR may join the multicast group or replication list. If a Scale-Out xTR chooses to join the multicast group, it MUST implement split-horizon filtering and ensure that traffic from PoD is not forwarded back into the PoD, in order to prevent duplication. xTRs providing active multihoming access to a PoD' s accelerators MUST support the following:

Registration: All active xTRs must register the PoD' s Scale-Up mappings with the Mapping System. Each registration must include the 'I' bit set and carry both the Site-ID (PoD ID) and the corresponding xTR-ID.

Multicast forwarding: Only the selected Scale-Up xTR joins the Scale-Up multicast group or replication list.

Broadcast Forwarding: Only the designated Scale-Out xTR is permitted to forward broadcast traffic to and from remote PoDs.

7. Data Plane Encapsulation Options

The LISP control plane is decoupled from the data plane encapsulation, allowing flexibility in the choice of encapsulation formats for Scale-Up and Scale-Out. Common encapsulation formats include VXLAN-GPE, LISP, and VXLAN:

VXLAN-GPE Encapsulation: Defined in [RFC9305], VXLAN-GPE supports encapsulation of both Scale-Up and Scale-Out packets. The VNI field directly maps to the Instance-ID used in the LISP control plane. For unified deployments, the P-bit is set, and the Next-Protocol field is used to indicate the payload type.

LISP Encapsulation: As specified in [RFC9300], this format also supports encapsulation of both Scale-Up and Scale-Out packets. The Instance-ID embedded in the EID maps directly to the Instance-ID in the LISP header. Upon decapsulation at the ETR, the IID may be used to determine whether the packet should be processed as part of a Scale-Up or Scale-Out flow.

Any alternative encapsulation format optimized for backend networks, capable of supporting a 24-bit Instance-ID, MAY be used to deploy Scale-Up, Scale-Out, or unified network data planes.".

8. IANA Considerations

No IANA considerations apply to this document.

9. Security Considerations

There are no additional security considerations except what already discussed in [RFC9301].

10. Acknowledgements

This draft builds on top of many LISP RFCs and drafts. Many thanks to the combined authors of those RFC and drafts.

11. Normative References

[I-D.ietf-lisp-site-external-connectivity]
Jain, P., Moreno, V., and S. Hooda, "LISP Site External Connectivity", Work in Progress, Internet-Draft, draft-ietf-lisp-site-external-connectivity-02, 28 March 2025, <https://datatracker.ietf.org/doc/html/draft-ietf-lisp-site-external-connectivity-02>.

[I-D.ietf-lisp-vpn]
Moreno, V. and D. Farinacci, "LISP Virtual Private Networks (VPNs)", Work in Progress, Internet-Draft, draft-ietf-lisp-vpn-12, 19 September 2023, <https://datatracker.ietf.org/doc/html/draft-ietf-lisp-vpn-12>.

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/info/rfc2119>.

[RFC8060] Farinacci, D., Meyer, D., and J. Snijders, "LISP Canonical Address Format (LCAF)", RFC 8060, DOI 10.17487/RFC8060, February 2017, <https://www.rfc-editor.org/info/rfc8060>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/info/rfc8174>.

[RFC9300] Farinacci, D., Fuller, V., Meyer, D., Lewis, D., and A. Cabellos, Ed., "The Locator/ID Separation Protocol (LISP)", RFC 9300, DOI 10.17487/RFC9300, October 2022, <https://www.rfc-editor.org/info/rfc9300>.

[RFC9301] Farinacci, D., Maino, F., Fuller, V., and A. Cabellos, Ed., "Locator/ID Separation Protocol (LISP) Control Plane", RFC 9301, DOI 10.17487/RFC9301, October 2022, <https://www.rfc-editor.org/info/rfc9301>.

[RFC9305] Maino, F., Ed., Lemon, J., Agarwal, P., Lewis, D., and M. Smith, "Locator/ID Separation Protocol (LISP) Generic Protocol Extension", RFC 9305, DOI 10.17487/RFC9305, October 2022, <https://www.rfc-editor.org/info/rfc9305>.

[RFC9306] Rodriguez-Natal, A., Ermagan, V., Smirnov, A., Ashtaputre, V., and D. Farinacci, "Vendor-Specific LISP Canonical Address Format (LCAF)", RFC 9306, DOI 10.17487/RFC9306, October 2022, <https://www.rfc-editor.org/info/rfc9306>.

[RFC9437] Rodriguez-Natal, A., Ermagan, V., Cabellos, A., Barkai, S., and M. Boucadair, "Publish/Subscribe Functionality for the Locator/ID Separation Protocol (LISP)", RFC 9437, DOI 10.17487/RFC9437, August 2023, <https://www.rfc-editor.org/info/rfc9437>.

[RFC9735] Farinacci, D. and L. Iannone, Ed., "Locator/ID Separation Protocol (LISP) Distinguished Name Encoding", RFC 9735, DOI 10.17487/RFC9735, February 2025, <https://www.rfc-editor.org/info/rfc9735>.

Authors' Addresses

Prakash Jain
MIPS
San Jose, CA
United States of America
Email: prjain@mips.com

Sanjay Hooda
Cisco Systems
San Jose
Email: shooda@cisco.com

Durgesh Srivastava
DataraAI
Cupertino, CA
United States of America
Email: durgesh@DataraAI.ai