Working Group
Internet-Draft
Intended status: Standards Track
Expires: 8 May 2026

X. Li
A. Wang
China Telecom
4 November 2025

Semantic Routing Architecture for AI Agents Communication
draft-li-semantic-routing-architecture-00

Abstract

This document introduces an Semantic Routing (SR) Architecture for enabling intelligent, semantic-driven communication among AI Agents. Unlike traditional IP-based routing or service mesh approaches, SRA leverages application-layer semantics — including service identity, intent vectors, and trust scores — to guide routing decisions dynamically. The architecture supports intent-driven task collaboration, trust-aware policy enforcement, and adaptive routing for multi-agent environments. SRA enables the network to evolve from a passive transport layer to an intelligent collaboration substrate supporting multi-agent coordination and cognitive networking.

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 8 May 2026.

Copyright Notice

Table of Contents

1. Introduction

The emergence of AI-driven ecosystems has transformed communication paradigms across computing and networking infrastructures. Traditional routing systems, designed for host-to-host communication, focus on connectivity, reachability, and link-state optimization. However, in environments where AI agents [AIAgent]—autonomous entities with reasoning and goal-oriented behavior—interact dynamically, such topological routing no longer meets the operational needs. Each agent represents not only a computational endpoint but also a semantic actor that generates intents, expresses capabilities, and negotiates tasks. The network must therefore evolve from a static forwarding fabric into a semantic coordination plane capable of interpreting meaning, context, and trust.

Existing frameworks such as Service Mesh [ServiceMesh] (e.g., Istio [Istio], Linkerd) and Software-Defined Networking (SDN) have improved visibility and control but remain largely syntactic. They route requests based on service names, APIs, or labels, not on why the communication occurs or what semantic goal it represents. For example, in an AI multi-agent system performing distributed reasoning, the decision of which node to contact depends on task semantics—such as "model adaptation," "policy refinement," or "data summarization"—and on dynamic factors like capability, latency, and trustworthiness. None of these can be expressed using IP addresses or conventional service identifiers.

The Semantic Routing (SR) architecture introduced in this draft aims to bridge this semantic gap. It extends routing intelligence from the network layer to the application layer, enabling communication decisions based on intent vectors, policy interpretation, and trust evaluation. Through a semantic control plane, SRA aligns network behavior with business and computational objectives, providing adaptive, secure, and efficient routing among AI agents. This enables networks to support intent-aware task collaboration and to act as intelligent participants in distributed cognition processes.

SRA also addresses emerging challenges of large-scale agent communication, including semantic interoperability, cross-domain trust, and self-optimization. Modern AI ecosystems consist of heterogeneous nodes—cloud agents, edge assistants, embedded inference units—that collaborate under uncertain conditions. Routing must thus adapt to fluctuating workloads, mobility, and trust contexts. Static or location-based approaches cannot efficiently manage such dynamism. By integrating semantic interpretation with continuous telemetry feedback, SRA allows networks to self-optimize: routes are recalculated not only based on network states (e.g., congestion or delay) but also on semantic relevance and agent reliability.

The design of SRA is guided by several fundamental objectives:

* Semantic Awareness Networks should understand and act upon high-level intents derived from AI tasks.

* Trust Integration Routing should consider the reliability and historical behavior of agents.

* Dynamic Adaptation Telemetry-driven feedback loops must continuously refine routing decisions.

* Backward Compatibility SAR should coexist with IP, BGP, and service-mesh infrastructures.

* Distributed Autonomy Each semantic router should make local decisions while aligning with global intent policies.

By embedding intelligence into the control and forwarding planes, SRA transforms the Internet from a data transport medium into a collaborative semantic ecosystem that supports intelligent communication for the next generation of distributed AI systems.

2. Conventions used in this document

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 [RFC2119] .

3. Terminology

The following terms are defined in this draft:

* SRA (Semantic Routing Architecture): The routing framework defined in this document integrating semantic awareness, trust, and policy control.

* AI Agent: An autonomous software entity capable of making context-based decisions, performing actions, and communicating with other agents.

* Intent Vector: A structured representation of the communication goal, expressed semantically (e.g., task type, priority, resource needs).

* Semantic Router (SR): Entity interpreting intent metadata and enforcing semantic forwarding policies.

* Semantic Forwarding Table (SFT): Forwarding table mapping intent categories to next hops and constraints.

4. Architecture Overview

The Semantic Router (SR) architecture introduces a layered design that bridges application semantics and network operation. It defines a semantic control framework capable of understanding agent-generated intents, evaluating contextual trust, and translating these into actionable routing policies. At its core, SRA consists of four interacting planes—Application, Control, Data, and Feedback—each responsible for distinct yet interdependent functions.

The Application Plane hosts AI agents that issue Intent Vectors (IVs) representing goals such as “request model inference” or “synchronize state.” The Semantic Control Plane collects these intents, authenticates identities, and maps them to routing policies via the Policy Engine (PE). These policies are then propagated to Semantic Routers (SRs) in the Data Plane, which execute the forwarding logic using Semantic Forwarding Tables (SFTs) that link intent types to paths and constraints. Finally, the Feedback Plane, driven by Telemetry Agents (TAs), monitors latency, trust, and service quality, feeding the results back into the control plane for continuous optimization.

This closed-loop system ensures that SAR continuously aligns network operation with evolving task goals. The architecture is designed to integrate seamlessly with existing IP and SDN environments, relying on overlays or extended routing attributes (e.g., BGP communities or SRv6 tags) to express semantic metadata.

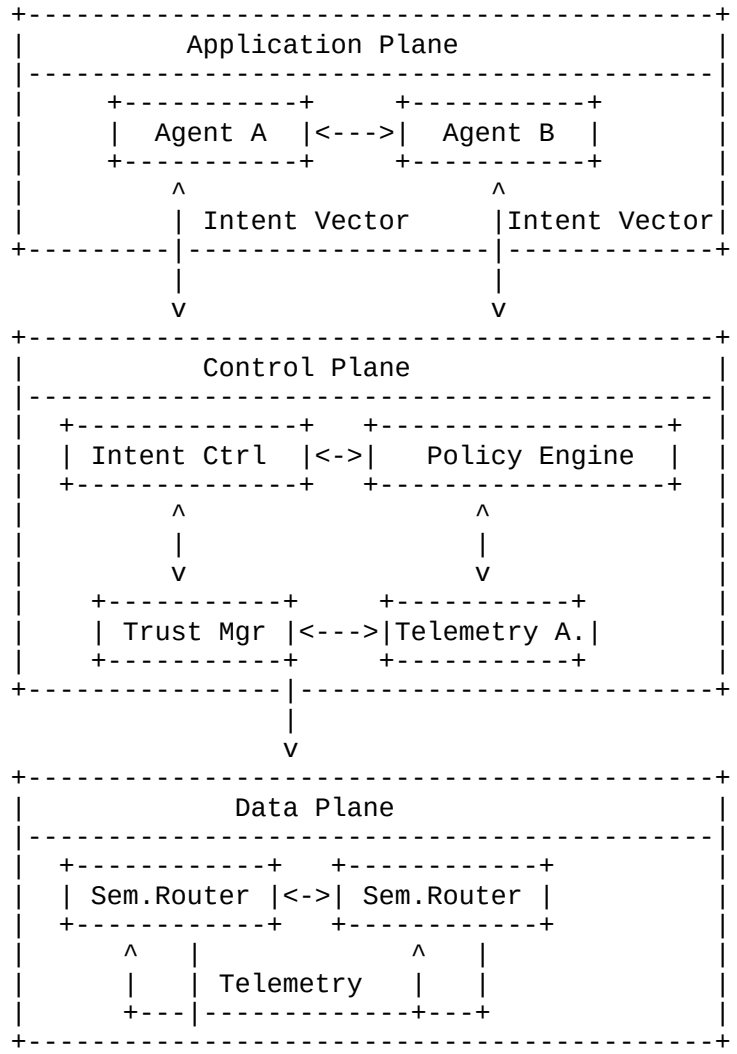

Figure 1 The overall architecture for SAR

5. Functional Layers and Design Principles

SAR's design is organized into five functional layers, each aligned with a core principle that ensures scalability, intelligence, and interoperability.

* Intent Layer: Generates and encodes Intent Vectors. Agents describe their goals in structured form, including task types, urgency, and context. The network uses these to infer optimal paths and collaborators.

* Identity and Trust Layer: Manages authentication, authorization, and reputation. Each agent is bound to a unique identity certificate, and trust scores are computed from telemetry.

* Policy Layer: The Policy Engine maps intents and trust data into enforceable rules, determining which paths, nodes, or bandwidth allocations are permitted.

* Semantic Routing Layer: Semantic Routers interpret policy rules and update SFT entries dynamically based on trust or performance metrics.

* Feedback Layer: Collects telemetry (e.g., latency, success rate, anomaly detection) and continuously refines both trust and policies.

SAR adheres to the following design principles:

* Semantic Composability: Each intent can be decomposed and recombined, enabling fine-grained routing for multi-step agent workflows.

* Trust Anchoring: Decisions are always contextualized by dynamic trust values, preventing compromised agents from influencing routing unfairly.

* Closed-Loop Adaptation: Every policy or path update is verified through telemetry feedback, ensuring stable yet flexible routing evolution.

* Interoperability: SAR MAY extend BGP, IS-IS, or gRPC metadata to distribute semantic and trust information while maintaining backward compatibility.

6. Control and Forwarding Procedures

SAR operates through coordinated procedures that integrate semantic interpretation, trust evaluation, and routing execution. These processes are logically divided between the control plane and forwarding plane, yet are interconnected via telemetry and feedback.

1. Agent Registration: When an agent joins the network, it authenticates with the Intent Controller (IC) and registers its capabilities (e.g., compute type, model domain). The IC issues credentials and a unique semantic prefix for the agent.

2. Intent Submission: The agent generates an Intent Vector and submits it to the IC. The Policy Engine (PE) parses the intent, referencing domain policies to determine allowed routing strategies.

3. Policy Translation: Based on the agent's trust score and system objectives, the PE compiles an executable rule set for the Semantic Router (SR). These rules specify target domains, quality preferences, and security constraints.

4. Routing Execution: SR uses its Semantic Forwarding Table (SFT) to determine next hops. Forwarding is influenced by trust, latency, and semantic relevance rather than just IP reachability.

5. Telemetry Feedback: The Telemetry Agent (TA) reports performance data back to the Trust Manager (TM). Trust scores are recalculated periodically, triggering policy adjustments when thresholds are exceeded.

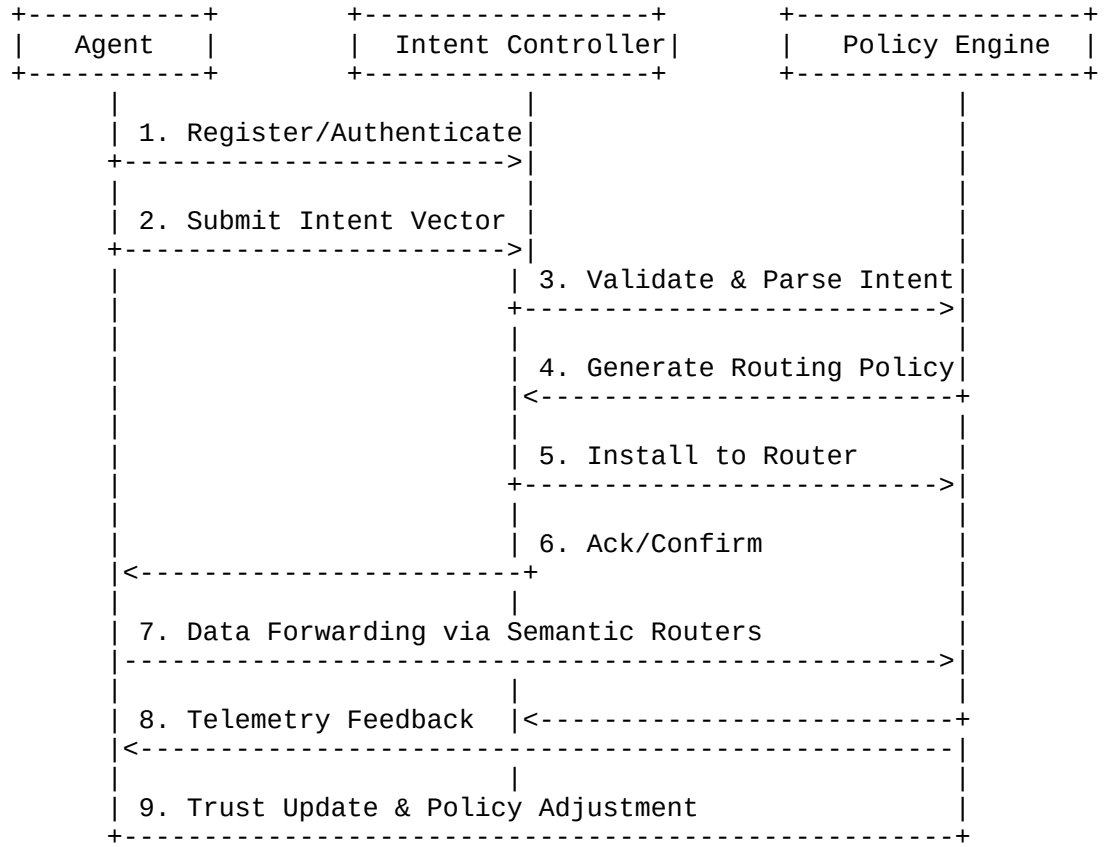

Figure 2 The Workflow Overview for SAR

7. Conclusion

The SRA (Semantic Routing architecture) redefines how intelligent systems communicate by integrating semantic intent, trust evaluation, and adaptive policy control directly into the routing process. It extends the traditional Internet model beyond topology and content toward a truly intent-driven communication fabric that aligns network behavior with the goals of autonomous AI agents. Through its layered design—including intent processing, trust management, semantic routing, and telemetry-driven feedback—SAR provides a coherent framework capable of supporting large-scale, cross-domain AI ecosystems with dynamic, secure, and efficient coordination.

Looking forward, several research and standardization opportunities remain. First, common intent representation languages must be defined to ensure interoperability among heterogeneous agents and vendors. Second, mechanisms for distributed trust computation

require standard metrics and synchronization protocols across administrative domains. Third, integration of SAR with existing Internet routing protocols such as BGP, IS-IS, or SRv6 will need careful consideration to balance scalability with semantic expressiveness. Finally, future work should investigate AI-assisted optimization within the SAR control plane, enabling predictive policy adjustments based on contextual learning.

In conclusion, SAR offers a foundational step toward an autonomous, cognition-aware Internet, where the network itself participates in decision-making, ensuring that communication among AI agents becomes purposeful, trustworthy, and adaptive.

8. IANA Considerations

TBD

9. Acknowledgement

TBD

10. Normative References

[AIAgent] N, D., "Framework for AI Agent Networks draft-zyyhl-agent-networks-framework-01", 2017.

[Istio] L, Larsson., "Impact of etcd deployment on kubernetes, istio, and application performance", 2020.

[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>.

[ServiceMesh] Li, W., "Service mesh: Challenges, state of the art, and future research opportunities", 2019.

Authors' Addresses

Xueting Li
China Telecom
Beiqijia Town, Changping District
Beijing
Beijing, 102209
China
Email: lixt2@foxmail.com

Aijun Wang
China Telecom
Beiqijia Town, Changping District
Beijing
Beijing, 102209
China
Email: wangaj3@chinatelecom.cn