CATALIST
Internet-Draft
Intended status: Informational
Expires: 13 September 2026

T. Eckert, Ed.
Futurewei Technologies USA

12 March 2026

Framework for Agent Communications Internet Protocol (ACIP) based Agent Aware Networks
draft-eckert-catalist-acip-framework-00

Abstract

This document outlines the use case scenario, problems with existing solutions and framework for a proposed new protocol solution for (AI) Agent to (AI) Agent based communications infrastructure using a new protocol tentatively called "Agent Communications Internet Protocol" (ACIP).

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 13 September 2026.

Copyright Notice

Table of Contents

1. Internal review considerations (to be removed)

This document tries to highlight aspects that hopefully will make the solution sound better to the IETF community. We want to be able for the solutions network forwarding nodes to have as much insightinto agent to agent traffic as today only done with IETF blessing in a variety of

2. Reference scenario

This document discusses the framework for ACIP based on the reference scenario outlined in this section. Further details on requirement can be found in [I-D.liu-rtgwg-agent-gateway-requirements].

A set of fintech enterprises wants to move to Agent to Agent communications for their next generation of fintech services, such as (for example) payment services involving credit/debit cards.

2.1. Today

Today (pre AtA), communication for similar services, such as HTTP based service, may be based on a hodge-podge of one-off B2B protocols with lots of operational and security challenges, running often over (a hodgepodge) or "secure" controlled networks, such as L2VPN or L3VPN.

Nevertheless, even with agents replacing such protocols, the fundamental complexities of setting up, managing and securing the AtA communications remains challenging, unless they are enhanced and simplified. In some respects they even increasing in complexity, because AtA communication can be intrinsically much less constrained to only approved/desired transactions (with typical HTTP B2B protocols being very well specified and hence easily parseable for security devices), making compliance with regulatory constraints and observability much more complex.

Figure 1 outlines the abstract architecture used today for HTTP B2B communications across the Internet, replacing just HTTP with "Agent". Enterprises "simply" connect to the Internet, but they need to ensure at their enterprise edge some degree of security embodied via enterprise edge HTTP proxy servers in A1 and A2 as well as Firewalls in the edge routers "FW-A1", "FW-A2".

With all those components running in the enterprises, it still requires several cloud service with various degrees of trust/security challenges all the way from Certificate Authorities, Naming services, coordination servers and so on. And then it is still extremely difficult to avoid data leakage because there is no trusted third party in the data plane guaranteeing an isolate communication domain. This too would need to be managed as a self-maintained overlay network.

Therefore, and while this "Internet model" for Agent to Agent communication is important, this memo proposes a complementary model which reflects an evolution of the concepts of private underlay/ overlay network designs as well as that of HTTP intermediaries.

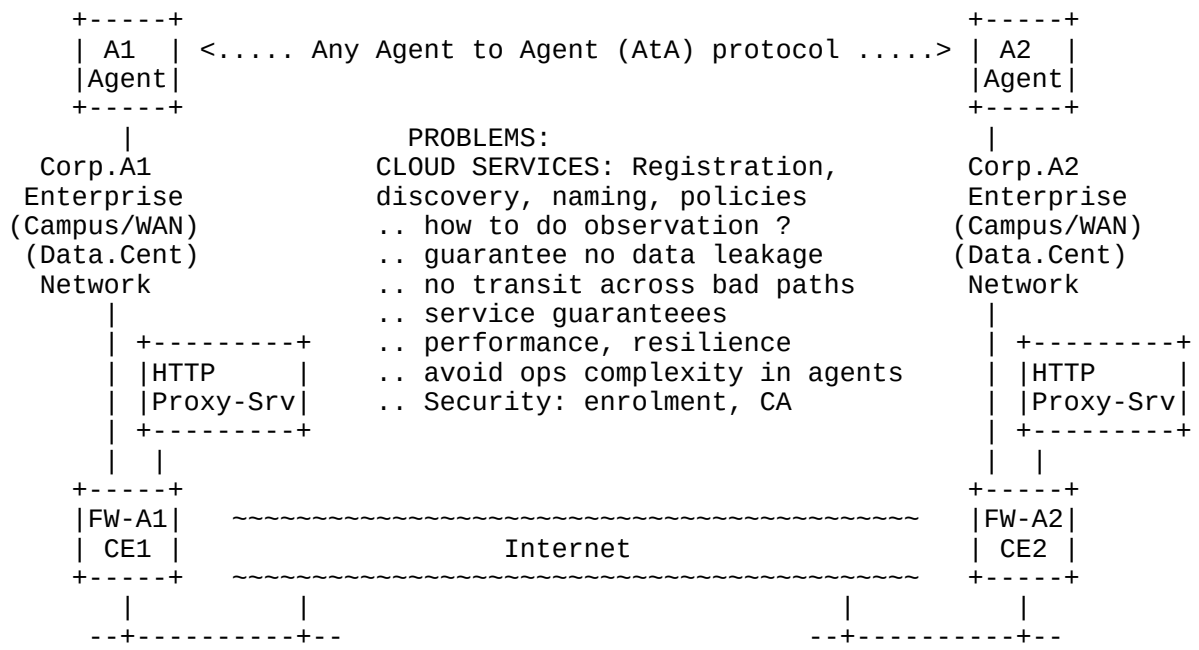


Figure 1: Internet Scenario Reference Topology

2.2. With ACIP

Ideally, such group of fintech enterprises would like for the communication model to also allow outsourcing such communication related problems to a trusted and legally for the purpose accredited third party, such as a network service provider and allow for the actual AI agents to require as little as possible additional complexity to support such complex scenarios and their security, network and regulatory requirements. Nevertheless, operators also want solutions to support self-managed/onprem equipment that supports all the necessary communication attachment functions to protect their own agents and backends, including the ability to apply policy, observability and security (access control, filtering) to AtA traffic.

A solution deployment for such a group of enterprises may include on or more providers of credit/debit cards, Issuing Banks, Point-of-Sales Terminals of Banks and Enterprises providing retail payment systems, Enterprises offering electronic payment systems linking to the credit/debit cards, co-brand enterprises and compliance entities. Each single debit/credit transaction may involve a series of AtA exchanged between two of these collaborating entities, and the solultion context too may impact policies and services (such as a national vs. international setup).

The communication aspects that such a communication infrastructure needs to support on a per-agent basis include attachment to service, registration of capabilities that the agent os permitted to provide / consume. Communications needs to be able to enforce desired or compliance besed access control, routing / load-balancing to agents providing group of agents, resilience/failover - and potentially transaction accounting or billing.

The following picture, Figure 2 outlines a simplified example for the above described use-case setup. A corporation A1 running some Agent A1 wants to have that Agent be able to discover and automatically connect in a secure, redundant and network path service optimized fashion to Agents from other corporations such as A2 so that those agents together can fulfil a distributed AI task such as aforementioned fintech services.

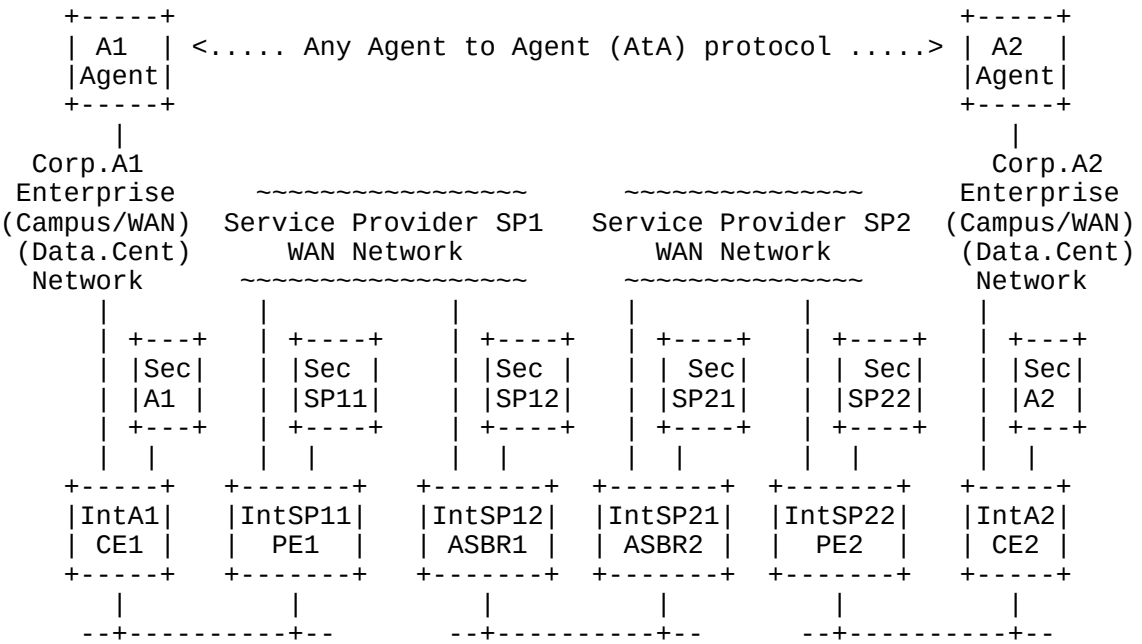


Figure 2: Scenario Reference Topology

2.2.1. Network topology

Corp.A connects via a CE router to a Service Provider SP1 via that service providers PE1 router. Likewise, Corp.A2 has logically the same setup, connecting its Agent A2 to Service Provider SP2 via C2 -> PE2. SP1 and SP2 interconnect via ASBR1/ASBR2. The overall network setup should considered to be primarily a private network setup, which may in some places pass across the Internet and/or have agents explicity integrated even though they are only reachable across the public Internet, but foremost, any reachability of agents to each other is meant to be explicitly created/allowed.

2.2.2. Cryptographic overlay

When, as often is the case, cryptographic security is required, the type of networks do include elements able to support cryptographic traffic processing. Because of historic limitations and/or inflexibility of traditional network devices such as CE1, CE2, PE1, PE2, cryptographically secure network paths are often created by co-placing software/CPU-only based processing, and routing is accordingly set up to pass the desired data plane traffic through those security devices. In the picture, the ’Sec’ devices provide exactly this role, and their co-location with each of the ’edge’ routers of one of the interconnecting organizations (Corporation or Service Provider) shows a typical setup of currently evolving solutions such as [SCION]. Intermediate nodes are part of such a secure cryptographic interdomain network design so that there can be trusted policy paths for example (such as forcing traffic inside specific countries for regulatory purposes).

2.2.3. ACIP inband processing

Finally, the novel ACIP networking technology proposed by this document is indicated via the IntA1, IntSP11, .. IntA2 modules. It is indicated as a new layer or service of forwarding functionality inside the pre-existing routers because the ability to support that option is of specific interest to maximize performance, lower cost (capex and opex) and to also increase security (device not being on-path do always have the problem of ensuring that they will not be bypassed.

In one example from past experiences, dedicated add-on firewalls in (high speed trading) fintech deployments where not able to keep up with the performance required from them whereas the actual network equipment was able to cope with the necessary speeds, but the application protocols where not designed such that the network equipment was able to perform the desired policy filtering.

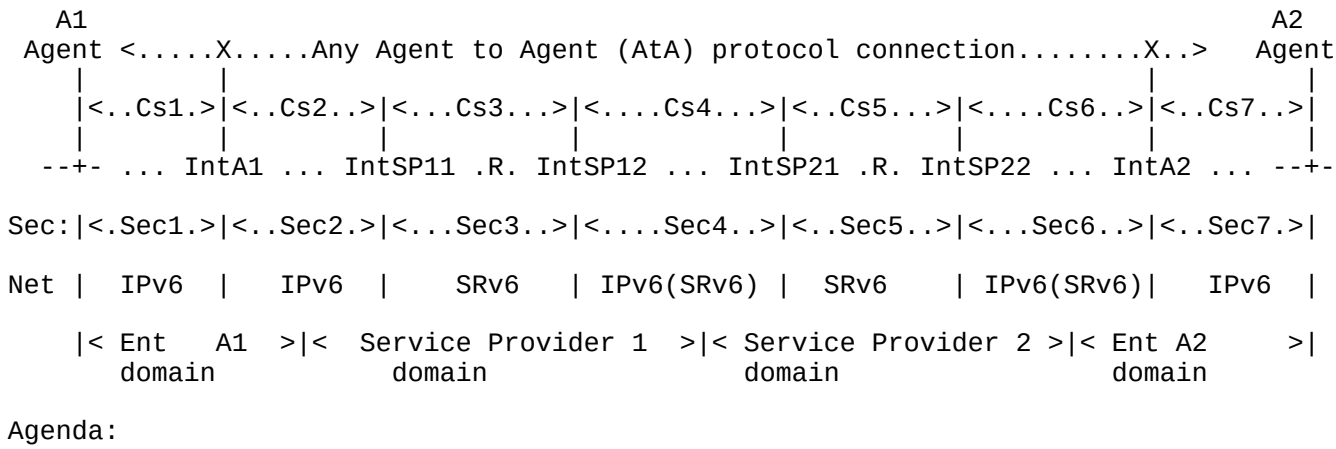
While the technology proposed by this document is intended to be designed that it is able to as much as possible run inside the actual routers, adding this functionality via additional devices such as those 'Sec' shown is also very important to support, especially for initial service agility. Also hybrid methods are of interest, where the

3. ACIP Architecture

This chapter summarizes proposed ACIP terminology and functionality, condensed from the reference scenario described above and explains it in contrast to existing technologies.

The following picture outlines the reference scenario again with focus on more details of layering and functionality. The focus still on data-plane components. Components carrying the actual data traffic of an AtA protocol connection. Management and Control plane components or protocols (if not co-located to data plane elements) are not shown but referred to by example, when needed.

3.1. Architecture layers



Agenda:

AtA - end-to-end Agent to Agent protocol
Google A2A and ANP are example instances of an AtA

Cs - Communication segment
AtATP connections run across a communication segment

... .R. Network path of a communication segment ... could be
a direct link. .R. means that there will be one or more
network routers in the path. Typically between an ingress
IntSP and egress IntSP (.R. represents all SP 'P' nodes).

Intxx - Intermediate node - forwarder of the AtATP.
HTTP proxy, Slim Node and ACIP router are examples of Intxx

AtATP - Agent to Agent Transport Protocol, across individ
HTTP, Slim and ACIP are example instances of AtATP

Figure 3: Logical Reference Topology

Two (of the multiple) agents, A1 and A2 are using some AtA (agent to agent) communication protocol.

In the network, A1 belongs to some Enterprise A1 and A2 belongs to some Enterprise A2. The network domain of A1 (EntA1) connects to the network domain of a Service Provider 1 (SP1), the network domain of A2 (EntA2) to the network domain of a Service Provider 2 (SP2). SP1 and SP2 also connect to each other.

EntA1 and EntA2 use IPv6, SP1 and SP2 use IPv6/SRv6. EntA1 and EntA2 may connect into L3VPN services in SP1 and SP2 as one example of private network functionality leveraged.

The traffic for AtA connection is carried across a series of intermediaries for ACIP shown as Intermediary A1 (IntA1), Intermediary SP1 1 (IntSP11) ... IntA2.

The path between each adjacent intermediaries and between an agent and an intermediary is called a Communication Segment (Cs1, ... Cs7). For each Cs, a specific protocol is used to encapsulate/carry the AtA protocol traffic, this protocol is called the "AtA transport protocol" (AtATP).

For simplicity, we consider that all Cs use the same AtATP. The reference for existing AtATP is HTTP. The name for the AtATP proposed by this framework document is tentatively "Agent Communication Internet Protocol" (ACIP).

3.2. HTTP proxies as precursors

To better understand the idea of ACIP forwarders in the solution, we refer to the most widely deployed existing technology that served as the inspiration, HTTP ’proxies’.

For HTTP, all intermediaries are HTTP "proxies". A1 opens an HTTP connection to IntA using an HTTP POST where the URL indicates A2. This causes a sequence of HTTP connections to be established from IntA1->IntSP11 up to IntA2->A2. Determining for each Cs the destination of the HTTP connection us subject to the control plane used.

3.3. ACIP layer functionality

On each intermediary, the desirable operations are performed: routing the call to the next appropriate intermediary, access control (filtering, policing), accounting. Routing may include complex policies such as assigning specific path resources including using only a subset of possible routes. For example, in fintech, connections may need to stay within specific permissible areas (such as within country). If HTTP is used, desired path and resource policies can be indicated through appropriate HTTP headers. Enabling the policies can be achieved by creating appropriate SRv6 policies for the connection traffic.

Routing information for the URLs is derived from appropriate routing information including controller based configuration, (private) DNS, or routing protocols such as BGP with a new address family the URLs used.

Authentication, confidentiality and integrity for AtATP traffic can be provided by either per-segment connection security, such as TLS and hence HTTPs or underlying network security such as IPsec - or a combination of both.

Authentication, confidentiality and integrity for the actual AtA payload is typically separate from that for AtATP confidentiality meaning that in the case of HTTP as AtATP, that the HTTP transaction payload is encrypted as well.

In some cases, intermediaries must be enabled to inspect the AtA payload to observe and potentially filter or account it. This is typically a core purpose of intermediaries belonging to the Enterprises operating either Agent (IntA1, IntA2). This is shown in the picture as an "X" in the AtA protocol connection and needs to be achieved by an appropriate protocol mechanism to share the AtA encryption key with the intermediary. Intermediaries representing regulatory mandatory observation points (often required for fintech transactions) may introduce the same requirement (not shown).

3.4. ACIP packetization

This section proposes just some core high-level concept ideas without presuming that any specific details are set in stone.

Consider that ACIP is fundamentally an internet protocol similar to IPv6 in that it is meant to allow forwariding and policy routing of datagrams across ACIP routers/forwarders. Nevertheless, by using application specific addressing via ’skills’, but allowing a more policy rich decision making and by supporting datagram and connection oriented transmissions, it is clearly targeted as an interworking layer on top of IPv6 or other existing network layers.

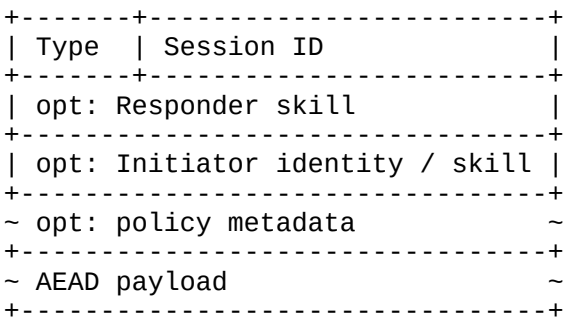


Figure 4: ACIP Header

The goal of ACIP is to allow the direct use of a ’skill’ based addressing space in which the functionality of the responder agent can directly be expressed, in the same way as e.g.: HTTP proxies would route on the URL.

The responder skill would in most cases be a logical address instead of an individual node identity (which it typically is in IPv6).

To enable rich policy based routing, a flexible metadata field is included, which could for example be a a set of TLV.

A Type field allows to distinguish different packet types.

ACIP can support datagram forwarding. This is indicated through an appropriate type. When datagram forwarding is used, the Session ID serves as a guaranteed hash to ensure that the ACIP packets with the same combination of Session-ID, Responder Skill and Initiator identity will always be delivered to the same responder. This ensures that datagram service can be lossy but that initiating agent and responder can perform retransmission or other forms of multi-packet exchange interactions to allow for a reliable end-to-end connection.

ACIP can support connection oriented transmissions. Initial connection setup would rely on specific ACIP packet types and would establish a session ID hop-by-hop from ACIP forwarder to the next ACIP forwarder. Connection setup may be complex at it involves routing decision which may take policy metadata into account and likely more complex cryptographic authentication. Therefore it may happen at control plane level of the ACIP forwarders.

Once an ACIP sesseion (connection) is established, the optional header fields may be left out. Instead, ACIP packets are forwarded by the Session-ID, which may be swapped ACIP-hop by ACIP-hop to avoid the need of coordinated end-to-end session-IDs. This is one example option how to allow fast hardware forwarding after session setup.

The payload is cryptographically authenticated, for example by AEAD, which may also include the necessary parts of the header.

3.5. Control Plane

Candidate new protocols for control plane operations are currently described in the following two draft.

The protocol which is tentatively called "Agent Attachment Protocol" [AAP] describes the control plane aspects between Agents and "first-hop" ACIP routers.

* The protocol operating across the first communication segment (Agent <-> IntA)

* Responsible for authentication, authorization, and skill registration

* Establishing the trust anchor for policy metadata used within ACIP

* Providing the attachment context from which ACIP session establishment proceeds

The protocol to support exchange and management of information between ACIP routers that can not better be supported by extending existing protocols is called tentatively "Agent Metadata Synchronization Protocol" [AMSP].

Note that these two drafts use the term "gateway" instead of ACIP router because they where trying to not finalize a specific terminoloy for the forwarding plane. The reason for doing such a tentative name for the forwarding plane in this memo with ACIP is that the term "gateway" is highly overloaded between many solutions and hence difficult to use when trying to distinguish different proposals. ACIP (routers) was tentatively choosen to empasize on the forwarding plane aspect between subnets.

Note that these two control plane protocols are not meant to capture the totality of necessary or beneficial control plane options to be used in conjunction with ACIP: Existing control plane protocols can equally be used and should be preferred for all aspects where especially operational well working practices make them preferrable.

For example, Skills could easily be a new address family in BGP to allow using the well established operational practices of routing services in Service Provider with BGP.

3.6. Scope extensions

ACIP is not meant to only be used to carry Agent 2 Agent protocols, but equally the traffic of any other protocol where be policy benefits of ACIP are beneficial - and where the naming of payload with ACIP "skills" is appropriate.

On easily very well applicable set of communications are those between an Agent and an "Model Context Protocol" (MCP) Gateway. One simply needs to define an appropriate set of Skill names and then ACIP can be used to route, manage/filter, load-balance traffic between Agents and MCP-Gateways - and forego the need for Agents to individually discovery and select MCP-gateway instances when instead a fitting instance can simply be directly reached by using the right "skill" name as the address of the MCP-gateway in ACIP.

Likewise, the topologies of interest for ACIP are not only those where enterprise hosted agents connect to other enterprise hosted Agents (or MCP-gateways). The following example "topology" options are possibly of interest too.

In Data Centers, simple default routing for agent to agent traffic may not be ideal. In these cases using ACIP for Skill based routing (foregoing discovery) may be beneficial. In one simple instance, Top-of-Rack (ToR) switch may be set up as ACIP routers performing this action. This may specifically be interesting case of multiple paths also in inter-DC connections available.

Internet/Cloud-hosted Agents will be very likely candidates to be brought into an ACIP solution through a "first-hop" ACIP router connecting to such Agents via the Internet. In this case, the Agent itself may not want or can-not speak ACIP itself, but the first-hop ACIP-router itself will "proxy-encapsulate" traffic from/to that Agent into ACIP. As well as mapping the control-plane mechanism used with ACIP to those used with "Internet Agent-to-Agent" communications.

4. Problems with current approaches

Today, the type of communication scenarios that ACIP attempts to improve use variety of B2B protocols, often HTTP based. HTTP intermediaries are used to solve communication scenario requirements such as access control/ filtering or routing (to an origin server or next intermediary/proxy).

4.1. PKI issues

Making all the necessary intermediaries trusted to pass on HTTP transaction and being ablde to inspect them at least on the HTTP level is one of the main complexities, often involving varieties of private PKI. Or when using WebPKI having to be extremely cautious with the minimum necessary subset of trusted Root CA. A solution wide unified and well managed PKI with participant permission management (such as intermediaries having only ability to be part of the HTTP communication but not the encrypted payloads) for example is

already a significant enhancement whereever possible.

4.2. Filtering for "extranet" private (L3VPN) networks

As mentioned in the reference scenario chapter, the network layer setup for such scenarios often involves private networks such as L2VPN or L3VPN, but these area already very difficult to manage in multi-enterprise scenarios because typically connectivity between the collaborating enterprises is typically intended to not be completely open at the network layer. Therefore, the network services may already need to provide strict filtering for known TCP server/ports. In other industries, this problem has already been recognized to be so error prone that automations such as MUD have been developed. These are not well applicable for ACIP target scenarios because the communication requirements are per-scenario and not per-device (such as required for MUD). Therefore, often HTTP intermediaries are used as URL "name" based access control enforcment nodes.

4.3. Naming on the DNS layer

Likewise, DNS based naming of entities is complex when setting up private scenarios because it requires private DNS trees, that should not leak to the Internet. With those scenarios evolving over time, multiple roots of such private DNS domains may need to be stitched together, which further increases complexity/insecurity.

4.4. Policy at the IPv6 layer

Attempting to provide routing/load-sharing/resource-allocation and access-control policy for application layer granularity directly on the layer of IPv6 addresses has often been attempted and is actually used in large scale, but complexity limited scenarios such as single-operator data-centers or constrained ISP sources. For example encoding allication context into some bits of the interface identifier. And then assigning to hosts (agents) a large number of IPv6 addresses, each one representing a specific application layer semantic. These approaches typically fall apart when the application layer purpose can be a combination of few independent aspects.

4.5. HTTP issues

Can not well be hardware forwarded - too complex to parse. No good separation control / data-plane, therefore not well suited to have CPU based control plane vs. hardware-forwarding based data-plane connctions..

Details TBD

5. Informative References

[AAP] Dunbar, L. and K. Majumdar, "Agent Attachment Protocol", Work in Progress, Internet-Draft, draft-dunbar-agent-attachment-00, 2 March 2026, <https://datatracker.ietf.org/doc/html/draft-dunbar-agent-attachment-00>.

[AgentCard] "Agent2Agent (A2A) Protocol Specification (Release Candidate v1.0)", 5 March 2026, <https://github.com/a2aproject/A2A/blob/main/docs/specification.md#5-agent-discovery-the-agent-card>.

[AMSP] Liu, B., "Agent Metadata Synchronization Protocol (AMSP)", 2016. draft-liu-agent-metadata-sync-protocol-00

[I-D.liu-rtgwg-agent-gateway-requirements] Liu, B., Geng, N., Shang, X., Gao, Q., Li, Z., and J. Gao, "Requirements for Agent Gateway", Work in Progress, Internet-Draft, draft-liu-rtgwg-agent-gateway-requirements-01, 27 November 2025, <https://datatracker.ietf.org/doc/html/draft-liu-rtgwg-agent-gateway-requirements-01>.

[SCION] de Kater, C., Rustignoli, N., and A. Perrig, "SCION Overview", Work in Progress, Internet-Draft, draft-dekater-panrg-scion-overview-06, 7 May 2024, <https://datatracker.ietf.org/doc/html/draft-dekater-panrg-scion-overview-06>.

Appendix A. Further technical considerations

A.1. Connection full or less operations

ACIP may be designed operate either/or in a connection based or connection less model. At this point in time it is not clear if one of the models alone is sufficient or if both options may be desirable.

In a connection less model, such as (by intention, not necessary by actual use) IPv6, each packet carries all (header) information for intermediary nodes (IPv6 routers) to perform all intended operations such as forwarding, (re-)routing, filtering/policing or assigning of processing resources.

In a connection oriented model such as HTTP/TCP binaries, a connection setup phase, such as an initial "HTTP GET" command towards the first intermediary causes the desired decisions, routing the connection to the next intermediary or origin server. Such connections may persist across transactions (as in some HTTP intermediaries) or across some "end-to-end connection".

As HTTP was not designed for network-device type proxies, it does not allow to easily distinguish between connection management data (such as the GET/PUT command) and actually transferred data. In connection oriented (internet) protocols designed to support network type devices, there is a clear separation between the connection management traffic (often called control plane) and the data transfer (called data plane). In typical network devices, the data plane can be million times faster than the control plane, hence the need to distinguish them.

In the same way that HTTP was not designed for intermediaries in the first place, but is now heavily used with them, it could equally be extended to support the establishment of such network device compatible data paths. Whether to do this or do a clean re-design, such as was done for constrained use cases with CoAP is an open design question.

Connection oriented solutions scale less well in the face of large number of connections because of the per-connection state in intermediate nodes. There is very limited experience with connection oriented internetwork technologies in TCP/IP networks. The most widely and largest scaling one is likely IP Multicast with PIM (up to 100,000 connections in deployments) followed by RSVP for Traffic engineering (RSVP-TE) in ISPs (likely up to 40,000 connections in deployments). Both are known to work fairly well within their scalability limited but suffer from long recovery times under failures and recovery. In contrast, connectionless operations suffers higher per-data-packet header overhead and per-packet processing cost in network devices (based on how complex the header is). In contrast, firewalls that provide security up to HTTP (or higher) layers have been known to not scale up to desired amount of supportable traffic/connections requiring often significant redesigns and reduction in actual security, siuch as in fintech use-cases.

Appendix B. Changelog

00 - initial version

Author’s Address

Toerless Eckert (editor)
Futurewei Technologies USA
2220 Central Expressway
Santa Clara, CA 95050
United States of America
Email: tte@cs.fau.de