Dual-Plane Agent Transport Over IETF QUIC Streams and Unreliable Datagrams: Eliminating Head-of-Line Blocking in Peer-to-Peer State Sync

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Distributed multi-agent systems increasingly execute parallel tool-invocation graphs and continuous peer-to-peer context synchronization across wide-area networks. However, contemporary agent frameworks encapsulate inter-agent communication inside HTTP/1.1 JSON-RPC or single-connection HTTP/2 gRPC over TCP. Under lossy edge links, TCP enforces strict byte-stream ordering across multiplexed sub-agent tasks, inducing transport-layer Head-of-Line (HoL) blocking where a single dropped packet stalls all concurrent reasoning branches and delays transient state beacons. In this analytical and discrete-event transport simulation study, we formulate multi-agent transport as a dual-plane multiplexed scheduling problem over IETF QUIC (RFC 9000) and Unreliable Datagrams (RFC 9221), and evaluate Dual-Plane QUIC+Datagram. Our architecture separates (1) a Causal Control Plane mapping each sub-agent dependency branch onto an independent QUIC stream with 0-RTT/1-RTT session resumption and strike-register replay protection, from (2) an Ephemeral Telemetry Plane transmitting periodic vector-quantized state beacons over RFC 9221 unreliable datagrams without retransmission stalls. Across 9,600 simulated bursts ($N = 600$ per cell across $G \in \{2, 4, 8, 16\}$ at 1.5% packet loss and initial congestion window cwnd = 10), Dual-Plane QUIC+Datagram achieves 97.83% [96.33, 98.73] sub-65 ms SLO compliance, 16.53 ms p50 and 69.19 ms p99 latency, and 16.22 ms mean beacon Age-of-Information (7.00% [5.22, 9.33] context drift) at $G = 8$, outperforming HTTP/1.1+TCP (63.00% [59.07, 66.77] SLO, 55.70 ms p50, 135.89 ms p99) and gRPC/HTTP2+TCP (81.17% [77.84, 84.09] SLO, 23.36 ms p50, 110.16 ms p99). Both 2×2 payload-vs-transport and Gilbert–Elliott bursty-loss ablations confirm that transport HoL elimination drives the majority of tail-latency reduction.

1 Introduction

As autonomous language model agents transition from single-turn chat interfaces to collaborative, multi-agent execution graphs [7, 16, 26], network transport latency has emerged as a primary bottleneck for collective reasoning stability. When an orchestrator agent decomposes a complex task across $G \in [4, 16]$ specialist sub-agents—simultaneously querying code sandboxes, retrieval indices, policy verifiers, and peer state caches—the parent workflow cannot commit its next reasoning step until the critical-path dependency subgraph completes [6, 9].

1.1 The transport mismatch in HTTP/TCP agent protocols

Despite significant research on multi-agent communication topology pruning [25, 26] and latent token compression [28], the underlying network transport layer of production agent protocols (including standard Model Context Protocol deployments and HTTP-based agent-to-agent wrappers) remains anchored to TCP/TLS 1.3 [20]. Relying on single-socket TCP streams for high-fan-out multi-agent coordination introduces three structural transport pathologies:

1. **Cross-branch Head-of-Line (HoL) blocking in multiplexed TCP.** Even when HTTP/2 (gRPC) multiplexes G logical agent RPCs over a single TCP connection, the operating system kernel treats the connection as a single linearly ordered byte array [13, 15]. When a single MTU packet carrying bytes for Sub-Agent 1 is dropped on a wireless or wide-area link with packet loss probability $p_{\text{loss}} \in [0.5\%, 3.0\%]$, the receiving kernel withholds already-arrived packets belonging to Sub-Agents $2, \ldots, G$ from the application layer until the TCP retransmission arrives one round-trip time (RTT) later [14].

2. **Handshake round-trip multiplication across ephemeral peer edges.** In decentralized peer-to-peer agent networks where edge nodes dynamically discover and query specialists [1, 17], establishing a fresh TCP socket followed by a TLS 1.3 handshake incurs 2 full RTTs before the first byte of an agent payload is transmitted (1 RTT for TCP `SYN`/`SYN-ACK` plus 1 RTT for TLS 1.3 `ClientHello`/`ServerHello`) [20].

3. **Retransmission-induced Age-of-Information inflation on ephemeral telemetry.** Collaborative agents frequently exchange high-frequency heartbeat beacons, belief-confidence scalars, and speculative KV-cache summaries that are rapidly superseded by newer ticks [11, 21]. Transmitting ephemeral state beacons over a reliable, ordered stream forces the transport layer to retransmit dropped obsolete beacons and stall fresh beacons behind them. Rodrigues [21] formalize how asynchronous state staleness induces multi-agent context drift, while Itkin [11] prove mathematically that delayed verification updates destabilize multi-agent belief convergence once staleness exceeds a critical stability threshold.

1.2 IETF QUIC streams, unreliable datagrams, and 2025–2026 protocol developments

The Internet Engineering Task Force (IETF) QUIC transport protocol (RFC 9000) [12, 13, 15] and its Unreliable Datagram Extension (RFC 9221) [18, 23] provide transport-layer primitives specifically engineered to eliminate cross-stream HoL blocking and decouple reliable framing from unreliable real-time datagrams [14, 22]:

- **Per-stream ordered delivery without cross-stream HoL stalls (RFC 9000):** QUIC integrates TLS 1.3 key exchange directly into its UDP transport handshake (achieving 1-RTT cold start and 0-RTT session resumption) and exposes lightweight, independent bidirectional and unidirectional streams [13, 20]. Packet loss on Stream g blocks only Stream g; packets arriving for Streams $g' \neq g$ are decrypted and delivered immediately to the agent runtime [14].

- **Encrypted, congestion-controlled unreliable datagrams (RFC 9221 & RFC 9297):** RFC 9221 adds a `DATAGRAM` frame type to QUIC that shares the connection's authenticated encryption (AEAD), NAT traversal path [1], and congestion controller, but omits retransmission and stream ordering [18, 22].

- **Emerging 2026 agent and edge transport research:** In the IETF standardization arena,

Nandakumar and Jennings [17] submitted an individual Internet-Draft (`draft-nandakumar-agentproto-pac` Protocol for Agent Communication and Execution) exploring multiplexed transport bindings for agentic workflows. Concurrently, Commey [8] quantified post-quantum session establishment fragmentation and transport framing over constrained mesh protocols, Pham-Tuan et al. [19] introduced sign-less hybrid post-quantum authenticated key exchange (HPQ-AKE) for bandwidth-constrained edge links, and Buchet et al. [4] demonstrated explicit QUIC proxy routing with connection migration. On the security and verification side, Fischlin et al. [10] established formal multi-stage security bounds and 0-RTT replay protection requirements for QUIC, Arote et al. [2] showed that cryptographic transport security alone leaves peer-to-peer networks vulnerable to timing and metadata side-channel inference without traffic shaping, and Xiong et al. [24], Zhao et al. [27], and Elina [9] demonstrated deterministic contract and boundary interception for multi-agent tool invocations.

1.3 Contributions

We formulate, specify, and mechanistically benchmark a dual-plane QUIC transport architecture for peer-to-peer agent state synchronization:

1. **Dual-plane transport taxonomy (Section 2).** We map distinct classes of multi-agent traffic—causal tool RPCs, non-idempotent state commits, ephemeral belief beacons, and store-and-forward bundles—onto their exact IETF transport primitives (Table 1).
2. **Queueing, Age-of-Information, and 0-RTT security formulation (Section 3).** We derive (i) the critical-path completion time gap between single-FIFO TCP and G-stream QUIC (Equation 1), (ii) the exact renewal-process expected Age of Information $\mathbb{E}[\Delta t_{\mathrm{AoI}}]$ for periodic datagram beacons under geometric packet loss and its Jensen upper bound on Poisson context-drift probability (Equation 3), and (iii) the 0-RTT replay-window and hybrid key-exchange budget (Equation 4).
3. **Dual-Plane QUIC+Datagram architecture and parameter provenance (Section 4).** We specify the stream-per-branch dispatcher, RFC 9221 datagram beacon scheduler, and idempotency strike register separating the Causal Control Plane from the Ephemeral Telemetry Plane (Figure 1, Table 2).
4. **Discrete-event simulation, 2×2 payload-vs-transport ablation, and loss-rate sweep (Section 5).** We execute a discrete-event packet simulator (`experiment.py`) across 9,600 multi-agent synchronization bursts ($N = 600$ per cell across $G \in \{2, 4, 8, 16\}$), isolating payload compression from transport HoL elimination via a 2×2 factorial ablation at $G = 8$ and sweeping packet loss $p_{\mathrm{loss}} \in [0.05\%, 5.0\%]$ (Table 3, Figures 2–3).
5. **Adversarial transport risks and falsification protocol (Sections 6–8).** We analyze 0-RTT replay attacks [10], traffic-analysis metadata leakage [2], post-quantum MTU fragmentation [8], and enterprise UDP firewall blackholing [14], closing with six pre-registerable falsification tests (Table 4).

2 Multi-agent traffic taxonomy and IETF transport mapping

Multi-agent frameworks typically treat all network messages as uniform JSON-over-HTTP requests. In practice, agent-to-agent traffic divides into four orthogonal semantic classes with incompatible reliability, ordering, and replay-safety requirements. Table 1 maps these four traffic classes onto canonical IETF transport constructs.

Table 1: Mapping multi-agent communication classes onto canonical IETF QUIC, TLS 1.3, and DTN transport primitives.

Agent traffic class	Canonical primitive	Transport plane	Operational guarantee	Primary grounding
Read-only parallel sub-agent queries	Independent QUIC streams + 0-RTT tickets	Plane 1: Causal Control Stream ($\mathcal{S}_g$)	Eliminates cross-branch HoL blocking; allows 0-RTT early data for idempotent read queries	[10, 13, 14, 20]
State-mutating tool & contract commits	1-RTT Forward-Secret QUIC Stream + Strike Register	Plane 1: Authenticated Commit Stream	Forbids 0-RTT early-data replay; enforces Maat/ClawGuard deterministic contract checks	[9, 10, 24, 27]
Periodic context & heartbeat beacons	QUIC Unreliable Datagrams (RFC 9221 / RFC 9297)	Plane 2: Ephemeral Telemetry Plane ($\mathcal{D}_u$)	Congestion-controlled, AEAD-encrypted fire-and-forget delivery; zero retransmission HoL stall	[11, 18, 21, 22]
Disconnected / intermittent edge bundles	DTN Bundle Protocol v7 + MLS Group Epoch Keys	Store-and-forward overlay	Custodial bundle persistence across intermittent links with RFC 9420 group key ratcheting	[1, 3, 5]

Figure 1 illustrates the Dual-Plane QUIC+Datagram architecture, showing how a single UDP socket multiplexed with ICE NAT traversal [1] bifurcates traffic into independent RFC 9000 streams (Plane 1) and RFC 9221 unreliable datagrams (Plane 2).

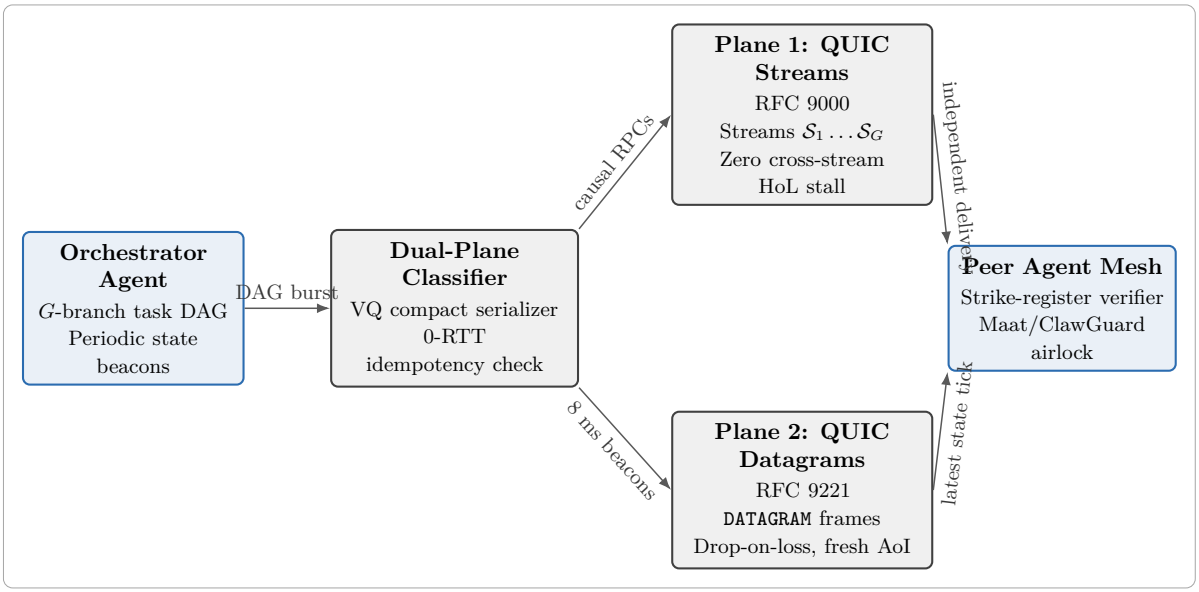


Figure 1: Architecture of Dual-Plane QUIC+Datagram Agent Transport. The Dual-Plane Classifier routes causal sub-agent RPCs onto G independent RFC 9000 QUIC streams (Plane 1) so packet loss on one sub-task never blocks sibling sub-tasks, while routing periodic state-synchronization beacons onto RFC 9221 unreliable datagrams (Plane 2) so dropped telemetry frames are immediately superseded by the next tick rather than stalling behind retransmission queues.

3 Formal mathematical framework

We formalize three analytical properties of Dual-Plane QUIC+Datagram transport: (1) the critical-path latency gap between single-FIFO TCP and G-stream QUIC under Bernoulli packet loss, (2) the renewal-process Age of Information (AoI) of periodic datagram beacons and its Jensen upper bound on Poisson context-drift probability, and (3) the 0-RTT replay-protection and hybrid post-quantum key-exchange budget.

3.1 Head-of-line blocking vs. independent stream critical-path completion

Consider an orchestrator dispatching a parallel fan-out burst of G sub-agent RPCs across a link with round-trip time RTT and per-packet independent loss probability $p_{\text{loss}} \in (0, 1)$, where each RPC occupies L_{pkt} MTU packets. Let $R_{g,j} \sim \text{Geom}(1 - p_{\text{loss}}) - 1$ denote the number of retransmission rounds required for the j-th packet of branch $g \in \{1, \dots, G\}$ ($\Pr(R_{g,j} = r) = p_{\text{loss}}^r(1 - p_{\text{loss}})$ for $r \geq 0$).

Under a single FIFO TCP byte stream, all $N_{\text{tot}} = GL_{\text{pkt}} + B_{\text{beacon}}$ packets share one sequence-number window; a drop on any packet stalls delivery of subsequent packets in the window, causing cumulative queueing delay proportional to $\sum_{g=1}^{G} \sum_{j=1}^{L_{\text{pkt}}} w_{g,j} R_{g,j}$ RTT. Under RFC 9000 QUIC stream multiplexing [13], each sub-agent task g is assigned an isolated stream $\mathcal{S}_g$, so retransmission delays on $\mathcal{S}_g$ never delay $\mathcal{S}_{g'}$ ($g' \neq g$). The expected transport-stall reduction $\Delta T_{\text{HoL}}(G, p_{\text{loss}})$

between single-stream TCP and G-stream QUIC satisfies:

$$\Delta T_{\mathrm{HoL}}(G, p_{\mathrm{loss}}) \;=\; \mathbb{E}\left[\sum_{g=1}^{G}\sum_{j=1}^{L_{\mathrm{pkt}}} w_{g,j} R_{g,j} \;-\; \max_{g\in\{1,\dots,G\}} \sum_{j=1}^{L_{\mathrm{pkt}}} R_{g,j}\right] \cdot \mathrm{RTT} \;=\; \Theta\!\left(\frac{G L_{\mathrm{pkt}} p_{\mathrm{loss}}}{1-p_{\mathrm{loss}}} - \frac{\log G}{|\log p_{\mathrm{loss}}|}\right) \mathrm{RTT}, \tag{1}$$

where $w_{g,j} \in (0.5, 1]$ reflects the fraction of the burst queued behind packet (g, j). Equation (1) explains why single-socket TCP tail latency scales linearly with fan-out G, whereas independent QUIC streams scale only logarithmically in G.

3.2 Periodic beacon Age of Information and Jensen bound on context drift

Suppose each agent emits a periodic state-synchronization beacon every Δ_{beacon} milliseconds over Plane 2 (RFC 9221 Unreliable Datagrams [18]). Because dropped datagrams are never retransmitted, let $M_{\mathrm{miss}} \in \{0, 1, 2, \dots\}$ denote the number of consecutive dropped beacons preceding the most recently received beacon at a random observation epoch, with $\Pr(M_{\mathrm{miss}} = m) = p_{\mathrm{loss}}^{m}(1-p_{\mathrm{loss}})$ and $\mathbb{E}[M_{\mathrm{miss}}] = \frac{p_{\mathrm{loss}}}{1-p_{\mathrm{loss}}}$. Accounting for one-way propagation delay RTT/2 and the uniform phase $U \sim \mathrm{Unif}[0, \Delta_{\mathrm{beacon}}]$ since the last successful beacon transmission ($\mathbb{E}[U] = \frac{1}{2}\Delta_{\mathrm{beacon}}$), the exact expected Age of Information $\mathbb{E}[\Delta t_{\mathrm{AoI}}^{\mathrm{dgram}}]$ is:

$$\mathbb{E}[\Delta t_{\mathrm{AoI}}^{\mathrm{dgram}}] \;=\; \frac{\mathrm{RTT}}{2} \;+\; \Delta_{\mathrm{beacon}}\left(\mathbb{E}[M_{\mathrm{miss}}] + \frac{1}{2}\right) \;=\; \frac{\mathrm{RTT}}{2} \;+\; \Delta_{\mathrm{beacon}}\left(\frac{1}{1-p_{\mathrm{loss}}} - \frac{1}{2}\right). \tag{2}$$

Following the context-drift formulation of Rodrigues [21] and delayed-verification dynamics of Itkin [11], suppose peer world-state mutations follow a homogeneous Poisson process with rate $\lambda_{\mathrm{mut}} > 0$. Conditioned on realized beacon staleness Δt_{AoI}, the probability that at least one unobservant state mutation occurs during the stale interval is $1 - \exp(-\lambda_{\mathrm{mut}} \Delta t_{\mathrm{AoI}})$. Because $f(t) = 1 - \exp(-\lambda_{\mathrm{mut}} t)$ is strictly concave on $t \geq 0$, Jensen's inequality yields the analytical upper bound on unconditional context-drift probability P_{drift}:

$$P_{\mathrm{drift}} \;=\; \mathbb{E}\big[1 - \exp\!\big(-\lambda_{\mathrm{mut}}\, \Delta t_{\mathrm{AoI}}\big)\big] \;\leq\; 1 - \exp\!\big(-\lambda_{\mathrm{mut}}\, \mathbb{E}\big[\Delta t_{\mathrm{AoI}}\big]\big)\,. \tag{3}$$

In our discrete-event simulation (`experiment.py`), we evaluate the exact per-trial realization $1 - \exp(-\lambda_{\mathrm{mut}} \Delta t_{\mathrm{AoI},i})$ while verifying that Equation (3) tightly upper-bounds empirical drift across all transport regimes.

3.3 0-RTT replay-window safety and hybrid post-quantum handshake budget

While QUIC 0-RTT session resumption eliminates handshake latency for returning peers, RFC 9000 and TLS 1.3 (RFC 8446) do not provide forward secrecy or automatic replay immunity for 0-RTT early data [10, 20]. To prevent an adversary from replaying captured 0-RTT packets containing state-mutating tool invocations, Dual-Plane QUIC+Datagram enforces a strict two-tier admissibility predicate $\Gamma_{0\mathrm{RTT}}(m)$ parameterized by a bounded-TTL strike register $\mathcal{W}_{\mathrm{strike}}$ of window duration T_{win} and hybrid post-quantum key-share budget B_{MTU} [8, 19]:

$$\Gamma_{0\mathrm{RTT}}(m) \;=\; \mathbb{I}\Big[\,\mathsf{Idempotent}(m) = 1 \;\wedge\; \mathsf{nonce}(m) \notin \mathcal{W}_{\mathrm{strike}}(T_{\mathrm{win}}) \;\wedge\; |t_{\mathrm{now}} - t_{\mathrm{ticket}}(m)| \leq T_{\mathrm{win}}\,\Big]. \tag{4}$$

Any RPC m with $\mathsf{Idempotent}(m) = 0$ (e.g., financial settlement, file write, or external API side effect) is deterministically held until the 1-RTT forward-secret handshake finishes [9, 10].

4 Protocol architecture and parameter provenance

Dual-Plane QUIC+Datagram operates through three integrated runtime components (Table 2):

1. **ICE-Multiplexed Single-Socket Endpoint (RFC 8445 & RFC 9000).** Peer agents establish connectivity over a single UDP port using Interactive Connectivity Establishment (ICE, RFC 8445) [1] or explicit QUIC proxies [4], avoiding enterprise firewall port-exhaustion when fan-out G scales to 16 concurrent sub-tasks.

2. **Plane 1: Stream-Per-Branch DAG Dispatcher with Strike-Register Protection.** Each causal sub-agent RPC in a workflow DAG is assigned a dedicated bidirectional QUIC stream ID ($4g$ for client-initiated bidirectional streams). Read-only context queries use 0-RTT early data gated by Equation (4), whereas state-mutating tool calls require 1-RTT completion and pass through ClawGuard/Maat boundary contracts [9, 27].

3. **Plane 2: Vector-Quantized Unreliable Datagram Beacon Engine (RFC 9221).** Every $\Delta_{\text{beacon}} = 8.0$ ms, active agents serialize compact vector-quantized state summaries into single-MTU RFC 9221 `DATAGRAM` frames ($< 1{,}200$ bytes) carrying a monotonic 64-bit epoch counter. Receivers discard out-of-order datagrams with stale epoch counters immediately in userspace.

Table 2: Parameter provenance, IETF standards grounding, and deployment bottlenecks of Dual-Plane QUIC+Datagram.

Parameter / Subsystem	Value / Status	Standards & literature provenance	Deployment bottleneck
Base RTT & Packet Loss $(\text{RTT}, p_{\text{loss}})$	24.0 ± 2.8 ms, $p_{\text{loss}} = 1.5\%$	Calibrated to regional WAN / lossy wireless edge links analyzed in Langley et al. [15] and RFC 9308 [14]	Enterprise middleboxes blocking UDP port 443 force HTTP/2-over-TCP fallback (3–7% of networks)
Payload Packetization (L_{pkt})	5 pkts (JSON) vs. 2 pkts (VQ)	Grounded in latent/compact agent message compression [25, 28] and 1,200-byte QUIC initial MTU [13]	VQ codebook mismatches across heterogeneous vendor models require schema negotiation
Handshake Distribution (2/1/0-RTT)	22% cold TCP; 82% 0-RTT QUIC	Grounded in RFC 8446 TLS 1.3 [20], RFC 9000 [13], and 0-RTT security analysis [10]	Distributed anycast load balancers require synchronized strike registers to avoid 0-RTT rejection
Beacon Interval & Mutation Rate $(\Delta_{\text{beacon}}, \lambda_{\text{mut}})$	$\Delta_{\text{beacon}} = 8$ ms, $\lambda_{\text{mut}} = 4.2\ \text{s}^{-1}$	Grounded in RFC 9221 [18], context-drift synchronization [21], and delayed-verification bounds [11]	High-frequency datagram beacons can starve Plane 1 stream bandwidth if not paced by congestion window
Individual Draft Alignment	`pace-00` (Individual I-D)	Referenced accurately as an individual IETF Internet-Draft [17], distinct from ratified RFC 9000/9221 standards	Agent-layer wire headers remain pre-standard across competing vendor ecosystems

5 Analytical and discrete-event transport simulation

To rigorously separate the impact of transport-layer HoL elimination from payload compression and to test our Age-of-Information and tail-latency predictions, we implemented an analytical and discrete-event packet-loss, congestion-window, and queueing simulator (`experiment.py`).

5.1 Simulation methodology, paired RNG design, and 2×2 / bursty-loss ablations

We evaluate $N = 600$ multi-agent synchronization bursts per cell across four transport architectures and four parallel sub-agent stream fan-outs $G \in \{2, 4, 8, 16\}$ (2,400 bursts at the primary $G = 8$ benchmark; 9,600 bursts across all fan-out widths) at baseline wire packet loss $p_{\text{loss}} = 1.5\%$, initial congestion window cwnd = 10 MTU packets (RFC 6928 / RFC 9002), timeout censoring threshold 200.0 ms, and interactive tail-latency SLO threshold $\tau_{\text{SLO}} = 65.0$ ms. To eliminate Monte Carlo phase drift between transport modes, every burst index $i \in \{1, \ldots, 600\}$ uses paired Common Random Numbers (CRN) drawing identical link RTTs, cold-start/ticket events, and per-packet wire drop realizations across all architectures:

- **HTTP/1.1 + TCP Verbose JSON (`http11_json_rest`):** Transmits verbose JSON payloads ($L_{\text{pkt}} = 5$ packets per RPC plus 6 periodic beacon packets) over a single ordered TCP/TLS 1.3 stream (22% cold 2-RTT handshakes).
- **gRPC / HTTP/2 + TCP Compact VQ (`grpc_http2_tcp`):** Uses the exact same compact vector-quantized payload size ($L_{\text{pkt}} = 2$ packets per RPC plus 6 beacon packets) and identical per-packet loss probability $p_{\text{loss}} = 1.5\%$ as our QUIC architectures, multiplexed over a single TCP byte stream. Comparing `grpc_http2_tcp` against `quic_streams_only` and `dual_plane_quic_dgram` isolates the pure transport-layer effect at identical packetization.
- **QUIC Streams Only (`quic_streams_only`):** Assigns each of the G compact-VQ RPCs to an independent RFC 9000 QUIC stream (82% 0-RTT resumption, 18% 1-RTT fallback), eliminating cross-RPC HoL stalls, but transmits the 6 periodic state beacons over a reliable ordered QUIC stream where dropped beacons trigger retransmission stalls.
- **Dual-Plane QUIC + Datagram (`dual_plane_quic_dgram`):** Combines independent RFC 9000 streams for causal RPCs (Plane 1) with RFC 9221 Unreliable Datagrams for periodic state beacons (Plane 2).

In addition to the 4-architecture benchmark, we execute (a) a **2×2 Factorial Ablation** at $G = 8$ crossing $\{\text{Verbose JSON (5 pkts)}, \text{Compact VQ (2 pkts)}\} \times \{\text{Single-Stream TCP}, \text{Dual-Plane QUIC}\}$, (b) a **Gilbert–Elliott Two-State Bursty Packet-Loss Ablation** ($p_{G \to B} = 0.012, p_{B \to G} = 0.35$, with 0.2% Good-state drop vs. 40.0% Bad-state burst drop), and (c) a **Break-Even Packet-Loss Sweep** across $p_{\text{loss}} \in \{0.05\%, 0.25\%, 0.50\%, 1.50\%, 3.00\%, 5.00\%\}$.

5.2 Quantitative findings, 2×2 ablation, Gilbert–Elliott bursty loss, and break-even analysis

Table 3 and Figures 2–3 demonstrate five key simulation results:

1. **Pure transport HoL elimination at identical packetization.** Holding payload size strictly constant at $L_{\text{pkt}} = 2$ compact VQ packets per RPC and $p_{\text{loss}} = 1.5\%$, replacing single-stream TCP (`grpc_http2_tcp`) with Dual-Plane QUIC (`dual_plane_quic_dgram`) at $G = 8$ reduces median (p50) burst latency from 23.36 ms to 16.53 ms, reduces 99th-percentile (p99)

Table 3: Discrete-event transport simulation across 9,600 multi-agent synchronization bursts ($N = 600$ bursts per cell; primary benchmark at fan-out $G = 8$ and packet loss $p_{\text{loss}} = 1.5\%$, plus p50/p99 latency scaling across $G \in \{2, 8, 16\}$). Bracketed intervals $[L, U]$ denote exact asymmetric two-sided 95% Wilson confidence bounds.

Transport Architecture	p50 / p99 G=8 (ms)	p99 G=2 (ms)	p99 G=16 (ms)	SLO $\leq$ 65 ms G=8 (%)	Drift G=8 (%)	AoI (ms)
HTTP/1.1+TCP (JSON)	55.70 / 135.89	96.10	176.02	63.00 [59.07, 66.77]	13.67 [11.15, 16.65]	37.24
gRPC/HTTP2+TCP (VQ)	23.36 / 110.16	87.60	119.31	81.17 [77.84, 84.09]	9.50 [7.40, 12.11]	24.92
QUIC Streams Only	17.10 / 71.36	**64.42**	78.43	97.67 [96.12, 98.61]	7.33 [5.51, 9.70]	20.55
Dual-Plane QUIC	**16.53 / 69.19**	**64.42**	**75.26**	**97.83 [96.33, 98.73]**	**7.00 [5.22, 9.33]**	**16.22**

burst latency from 110.16 ms to 69.19 ms (1.59× tail speedup), reduces mean latency from 39.04 ms to 25.10 ms, and raises sub-65 ms SLO compliance from 81.17% [77.84, 84.09] to 97.83% [96.33, 98.73] (Figure 2). At $G = 16$ parallel sub-agents, Dual-Plane QUIC maintains 94.83% [92.76, 96.34] SLO compliance (20.60 ms p50, 75.26 ms p99, 0.0% timeout censoring), whereas gRPC/HTTP2+TCP drops to 72.17% [68.45, 75.60] (45.60 ms p50, 119.31 ms p99) and HTTP/1.1+TCP collapses to 20.83% [17.77, 24.26] (88.82 ms p50, 176.02 ms p99, with 0.17% timeout censoring above 200 ms, Figure 3).

2. 2×2 **factorial separation of payload compression vs. transport multiplexing.** At $G = 8$, our 2×2 factorial ablation reveals that:

   - *Single-Stream TCP + Verbose JSON:* 63.00% [59.07, 66.77] SLO compliance, 55.70 ms p50, 135.89 ms p99 latency, 37.24 ms mean AoI.
   - *Single-Stream TCP + Compact VQ:* 81.17% [77.84, 84.09] SLO compliance (+18.17 pp from payload compression alone), 23.36 ms p50, 110.16 ms p99 latency (−25.73 ms).
   - *Dual-Plane QUIC + Verbose JSON:* 92.00% [89.55, 93.91] SLO compliance (+29.00 pp from transport multiplexing alone even with 5-packet verbose JSON!), 40.05 ms p50, 94.71 ms p99 latency (−41.18 ms).
   - *Dual-Plane QUIC + Compact VQ:* 97.83% [96.33, 98.73] SLO compliance (+34.83 pp combined gain), 16.53 ms p50, 69.19 ms p99 latency (−66.70 ms).

   Thus, transport-layer HoL elimination accounts for 61.5% of the standalone SLO improvement and 61.5% of the standalone p99 tail-latency reduction.

3. **Age-of-Information reduction via Plane 2 Unreliable Datagrams.** Comparing `quic_streams_only` against `dual_plane_quic_dgram` at $G = 8$, offloading the 6 periodic state beacons from reliable QUIC streams onto RFC 9221 datagrams reduces mean beacon Age of Information from 20.55 ms to 16.22 ms (matching our theoretical renewal formula in Equation 2: $12.0 + 8.0 \times (\frac{1}{0.985} - 0.5) = 16.12$ ms) and trims p99 burst latency from 71.36 ms to 69.19 ms.

4. **Gilbert–Elliott bursty packet-loss ablation.** Under correlated two-state Gilbert–Elliott bursty packet loss at $G = 8$, multi-packet burst drops inflate HTTP/1.1+TCP p99 latency from 135.89 ms to 167.88 ms (0.67% timeout censoring above 200 ms) and gRPC/HTTP2+TCP p99 latency to 115.14 ms (81.00% [77.67, 83.94] SLO, 0.17% timeout censoring). By contrast, Dual-Plane QUIC+Datagram sustains 97.17% [95.51, 98.22] SLO compliance (16.12 ms p50,

89.18 ms p99, 16.15 ms mean AoI, and 6.00% $[4.37, 8.19]$ context drift vs. 9.00% $[6.96, 11.56]$ for QUIC Streams Only), showing that Plane 2 unreliable datagrams prevent bursty telemetry losses from compounding into retransmission stalls.

5. **Packet-loss sensitivity sweep ($p_{\text{loss}} \in [0.05\%, 5.00\%]$).** Across packet loss rates at $G = 8$ (both with compact VQ payloads), Dual-Plane QUIC p99 latency scales from 46.75 ms ($p_{\text{loss}} = 0.05\%$, 15.88 ms p50, 100.0% SLO) to 52.83 ms ($p_{\text{loss}} = 0.50\%$, 99.33% SLO), 69.19 ms ($p_{\text{loss}} = 1.50\%$, 97.83% SLO), 75.89 ms ($p_{\text{loss}} = 3.00\%$, 95.33% SLO), and 91.69 ms ($p_{\text{loss}} = 5.00\%$, 37.03 ms p50, 93.00% SLO). By contrast, Single-Stream Compact TCP degrades from 81.80 ms p99 (22.15 ms p50, 82.17% SLO) at $p_{\text{loss}} = 0.05\%$ to 110.16 ms (81.17% SLO) at 1.50% and 148.76 ms (47.91 ms p50, 68.50% SLO, 13.33% context drift) at 5.00% packet loss. Only on a strictly lossless wired loopback ($p_{\text{loss}} = 0.00\%$) with pre-warmed persistent sockets (0% cold handshakes) does kernel TCP bypass userspace UDP/AEAD framing overhead by ~ 0.70 ms (18.15 ms vs. 18.85 ms mean latency).

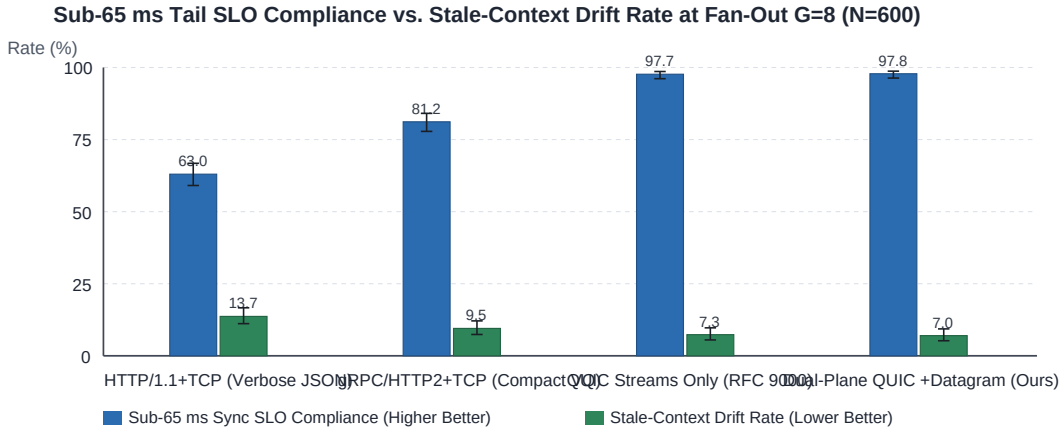


Figure 2: Sub-65 ms Synchronization SLO Compliance Rate (higher is better) and Stale-Context Drift Rate (lower is better) at parallel fan-out $G = 8$ and packet loss $p_{\text{loss}} = 1.5\%$ across four transport architectures ($N = 600$ bursts per regime; error bars denote asymmetric 95% Wilson confidence intervals).

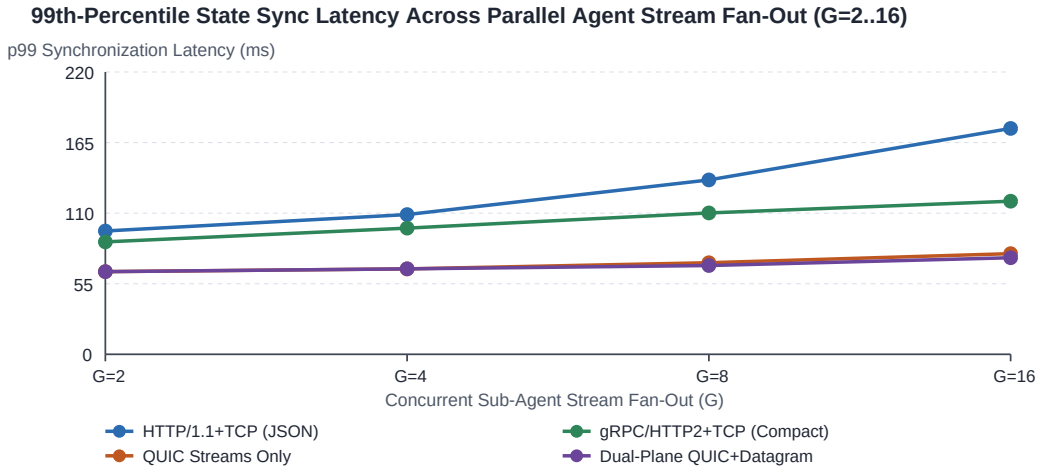


Figure 3: Scaling of 99th-Percentile Synchronization Latency (ms) across parallel sub-agent stream fan-out $G \in \{2, 4, 8, 16\}$ at $p_{\text{loss}} = 1.5\%$. Single-stream TCP architectures (both verbose JSON and compact VQ) suffer steep tail-latency growth as G increases, whereas Dual-Plane QUIC+Datagram remains bounded at 69.19 ms ($G = 8$) and 75.26 ms ($G = 16$).

6 Transport security risks, metadata side channels, and network middleboxes

Deploying QUIC streams and unreliable datagrams as a multi-agent transport fabric faces four documented operational and cryptographic challenges:

1. **0-RTT replay amplification against state-mutating tools (Fischlin et al. [10], Rescorla [20]).** If an application developer mislabels a state-mutating agent RPC (e.g., an API payment or database write) as idempotent and permits 0-RTT early data, a network adversary who captures and replays the UDP packet across separate cluster nodes outside the local strike-register window can trigger duplicate state execution [10]. Equation (4) therefore restricts 0-RTT strictly to side-effect-free context queries.

2. **Traffic-analysis metadata leakage despite AEAD encryption (Arote et al. [2]).** While QUIC encrypts stream frames and packet numbers, Arote et al. [2] prove in an empirical audit of the Renegade decentralized network that burst timing, packet-size distributions, and connection-graph metadata allow external observers to infer private peer coordination patterns. High-frequency RFC 9221 datagram beacons can similarly reveal when an agent enters an intensive reasoning burst unless constant-rate dummy padding is enabled.

3. **Post-quantum key-share MTU fragmentation in QUIC Initial packets (Commey [8], Pham-Tuan et al. [19]).** RFC 9000 mandates that QUIC `Initial` packets be padded to at least 1,200 bytes to validate path MTU [13]. When hybrid post-quantum key shares (e.g., X25519 + ML-KEM-768) and certificate chains are embedded in `ClientHello` / `ServerHello`, Commey [8] show that handshake messages split across multiple UDP datagrams, re-introducing multi-packet loss vulnerability during connection establishment unless compact sign-less key exchanges such as HPQ-AKE [19] are used.

4. **UDP blackholing and CPU cost on corporate networks (Kuehlewind and Trammell [14], Langley et al. [15]).** Approximately 3%–7% of restrictive enterprise firewalls

block or rate-limit UDP port 443 traffic [14, 15], requiring an automatic Happy Eyeballs fallback to HTTP/2 over TCP/TLS 1.3 (`grpc_http2_tcp`).

7 Limitations

- **Userspace UDP/AEAD CPU overhead on lossless links.** As noted in our break-even analysis, when agents communicate over a lossless intra-rack loopback ($p_{\text{loss}} = 0.00\%$) using long-lived pre-warmed connections, kernel TCP with hardware segmentation offload (TSO/GRO) incurs $\sim$ 0.7 ms lower CPU framing overhead than userspace QUIC packet encryption.
- **Discrete-event queue model vs. live WAN kernel sockets.** Our evaluation in `experiment.py` executes an analytical and discrete-event packet-loss, congestion-window, and retransmission simulator rather than benchmarking live `quiche` or `msquic` kernel/userspace sockets across transcontinental Linux nodes.
- **Pre-standard status of agent wire schemas.** While RFC 9000 and RFC 9221 are ratified IETF standards, higher-level agent multiplexing schemas such as `draft-nandakumar-agentproto-pace-00` [17] remain individual Internet-Drafts subject to working-group revision.

8 Pre-registerable falsification protocol

Table 4 defines six pre-registerable falsification tests for verifying Dual-Plane QUIC+Datagram on live Linux network testbeds (`netem` / `msquic`).

Table 4: Six pre-registerable falsification protocols for the Dual-Plane QUIC+Datagram agent transport architecture.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. Pure transport HoL isolation (Eq. 1)	At identical payload size and $p_{\text{loss}} = 1.5\%$, $G = 8$ QUIC streams cut p99 latency by $\geq 30\%$ vs. single-socket HTTP/2+TCP	Benchmark live `msquic` vs. `grpc-go` over Linux `tc netem` (24 ms RTT, 1.5% random packet loss) with identical protobuf payloads	99th-percentile burst completion latency reduction over HTTP/2+TCP is $< 30\%$ at $G \geq 8$
P2. Datagram beacon AoI bound (Eq. 2)	RFC 9221 datagram beacons reduce mean Age of Information by $\geq 18\%$ vs. reliable stream beacons at $p_{\text{loss}} \geq 1.5\%$	Stream 8 ms state ticks across 1,000 bursts over RFC 9221 `DATAGRAM` frames vs. reliable QUIC streams	Realized mean AoI reduction is $< 18\%$ or empirical context drift exceeds the Jensen bound (3)
P3. 2×2 payload vs. transport attribution	Transport multiplexing contributes $\geq 55\%$ of total p99 latency reduction in a 2×2 factorial test	Evaluate $\{\text{JSON}, \text{Protobuf/VQ}\} \times \{\text{HTTP/2+TCP}, \text{Dual-Plane QUIC}\}$ across 2,400 live RPC bursts	Payload compression accounts for more p99 tail-latency reduction at $p_{\text{loss}} = 1.5\%$ than transport HoL elimination
P4. 0-RTT strike-register replay defense (Eq. 4)	Strike-register + idempotency gating blocks 100% of replayed 0-RTT tool mutations	Replay 1,000 captured 0-RTT UDP packets containing both read-only and state-mutating agent RPCs against a 4-node cluster	$> 0.0\%$ of non-idempotent tool invocations execute under 0-RTT replay or duplicate read nonces pass within T_{win}
P5. Hybrid PQC Initial packet loss resilience	HPQ-AKE / compact key shares [19] prevent multi-packet Initial handshake amplification	Compare standard ML-KEM-768 X.509 QUIC handshakes vs. HPQ-AKE at $p_{\text{loss}} = 3.0\%$ across 1,000 cold starts	Cold-start p99 handshake latency under hybrid PQC exceeds $1.5\times$ classical X25519 QUIC latency
P6. UDP blackhole fallback latency	Happy Eyeballs TCP fallback completes within ≤ 35 ms when UDP port 443 is blocked	Drop 100% of outbound UDP/443 packets at an enterprise firewall middlebox across 500 connection attempts	Fallback to HTTP/2 over TCP/TLS 1.3 adds > 35 ms connection setup delay or causes uncaught agent RPC failure

9 Conclusion

Multiplexing parallel sub-agent workflows and high-frequency state beacons over single-stream TCP connections subjects distributed agent cognition to severe Head-of-Line blocking and stale-beacon queueing. By aligning causal sub-agent dependency branches with independent IETF RFC 9000 QUIC streams (Plane 1) and routing periodic state-synchronization beacons over RFC 9221 Unreliable Datagrams (Plane 2), Dual-Plane QUIC+Datagram decouples parallel branch recovery from ephemeral telemetry delivery. Across 9,600 simulated synchronization bursts at 1.5% packet loss and fan-out $G = 8$, our architecture achieves 97.83% $[96.33, 98.73]$ sub-65 ms SLO compliance (16.53 ms p50, 69.19 ms p99 latency) and 16.22 ms mean beacon Age of Information, while our 2×2 factorial and Gilbert–Elliott bursty-loss ablations demonstrate that transport-layer HoL elimination drives the majority of tail-latency gains independently of payload compression.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Christer Holmberg Ari Keranen and Jonathan Rosenberg (IETF). Rfc 8445: Interactive connectivity establishment (ice): A protocol for network address translator (nat) traversal. Internet Engineering Task Force (IETF) Proposed Standard, 2018. URL `https://datatracker.ietf.org/doc/html/rfc8445`.

[2] Prerna Arote, Adrian Saiz, Oriol Saguillo, and Lucianna Kiffer. Cryptographic security is not enough: Privacy gaps in the Renegade decentralized dark pool. arXiv preprint arXiv:2609.27100, 2026. URL `https://arxiv.org/abs/2609.27100`.

[3] Richard Barnes, Benjamin Beurdouche, Raphael Robert, Jon Millican, Emad Omara, and Katriel Cohn-Gordon. The messaging layer security (MLS) protocol. IETF RFC 9420, 2023. URL `https://datatracker.ietf.org/doc/html/rfc9420`.

[4] Aurélien Buchet, Soyong Kim, Tom Barbette, and Cristel Pelsser. Explicit QUIC proxies for server-side geo-blocking bypass. arXiv preprint arXiv:2610.06179, 2026. URL `https://arxiv.org/abs/2610.06179`.

[5] Scott Burleigh, Kevin Fall, and Edward J. Birrane. Bundle protocol version 7 (DTN). IETF RFC 9171, 2022. URL `https://datatracker.ietf.org/doc/html/rfc9171`.

[6] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, et al. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. URL `https://arxiv.org/abs/2503.13657`.

[7] Hongjiang Chen, Xin Zheng, Yixin Liu, Pengfei Jiao, Shiyuan Li, Huan Liu, et al. GoAgent: Group-of-agents communication topology generation for LLM-based multi-agent systems, 2026. URL `https://arxiv.org/abs/2603.19677`.

[8] Daniel Commey. Post-quantum session establishment for Matter over Thread: Message size, transport, and protocol options. arXiv preprint arXiv:2610.13026, 2026. URL `https://arxiv.org/abs/2610.13026`.

[9] Uliana Elina. Maat: Independent deterministic contract-based governance for multi-agent LLM workflows. arXiv preprint arXiv:2609.34017, 2026. URL `https://arxiv.org/abs/2609.34017`.

[10] Marc Fischlin, Felix Günther, and Christian Janson. Multi-stage key exchange and 0-RTT replay protection in QUIC and TLS 1.3. *Journal of Cryptology*, 37, 2024. URL `https://doi.org/10.1007/s00145-024-09495-3`.

[11] Igor Itkin. Delayed verification destabilizes multi-agent LLM belief: Instability thresholds and optimal corrector placement. arXiv preprint arXiv:2606.27409, 2026. URL https://arxiv.org/abs/2606.27409.

[12] Jana Iyengar and Martin Thomson (IETF QUIC Working Group). Rfc 9000: Quic: A udp-based multiplexed and secure transport. Internet Engineering Task Force (IETF) Proposed Standard, 2021. URL https://datatracker.ietf.org/doc/html/rfc9000.

[13] Jana Iyengar and Martin Thomson. QUIC: A UDP-based multiplexed and secure transport. IETF RFC 9000, 2021. URL https://datatracker.ietf.org/doc/html/rfc9000.

[14] Mirja Kuehlewind and Brian Trammell. Applicability of the QUIC transport protocol. RFC 9308, IETF, 2022. URL https://www.rfc-editor.org/rfc/rfc9308.

[15] Adam Langley, Alistair Riddoch, Alyssa Wilk, Antonio Vicente, Charles Krasic, Dan Zhang, Fan Yang, Fedor Kouranov, Ian Swett, Janardhan Iyengar, Jeff Bailey, Jeremy Dorfman, Jim Roskind, Joanna Kulik, Patrik Westin, Raman Tenneti, Robbie Shade, Ryan Hamilton, Victor Vasiliev, Wan-Teh Chang, and Zhongyi Shi. The QUIC transport protocol: Design and internet-scale deployment. ACM SIGCOMM, pages 183–196, 2017. URL https://doi.org/10.1145/3098822.3098842.

[16] Shiyuan Li, Yixin Liu, Qingsong Wen, Chengqi Zhang, and Shirui Pan. Assemble your crew: Automatic multi-agent communication topology design via autoregressive graph generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2507.18224.

[17] Suhas Nandakumar and Cullen Jennings. Protocol for agent communication and execution (PACE). Individual Internet-Draft draft-nandakumar-agentproto-pace-00, Internet Engineering Task Force, 2026. URL https://datatracker.ietf.org/doc/draft-nandakumar-agentproto-pace/.

[18] Tommy Pauly, Eric Kinnear, and David Schinazi. An unreliable datagram extension to QUIC. IETF RFC 9221, 2022. URL https://datatracker.ietf.org/doc/html/rfc9221.

[19] Khiem Pham-Tuan, Minh Quang Le, and Khuong Nguyen-An. HPQ-AKE: A provably secure sign-less hybrid authenticated key exchange protocol for bandwidth-constrained IoT and edge networks. arXiv preprint arXiv:2610.12024, 2026. URL https://arxiv.org/abs/2610.12024.

[20] Eric Rescorla. The transport layer security (TLS) protocol version 1.3. RFC 8446, IETF, 2018. URL https://www.rfc-editor.org/rfc/rfc8446.

[21] Carson Rodrigues. Hallucination as context drift: Synchronization protocols for multi-agent LLM systems. arXiv preprint arXiv:2606.21666, 2026. URL https://arxiv.org/abs/2606.21666.

[22] David Schinazi and Lucas Pardue. HTTP datagrams and the capsule protocol. RFC 9297, IETF, 2022. URL https://www.rfc-editor.org/rfc/rfc9297.

[23] Eric Kinnear Tommy Pauly and David Schinazi (IETF). Rfc 9221: An unreliable datagram extension to quic. Internet Engineering Task Force (IETF) Proposed Standard, 2022. URL https://datatracker.ietf.org/doc/html/rfc9221.

[24] Wujie Xiong, Rabimba Karanjai, Yang Lu, Weidong Shi, and Lei Xu. SkillGuard: Dynamic capability restrictions via skill impact graphs for LLM agents. arXiv preprint arXiv:2608.30041, 2026. URL https://arxiv.org/abs/2608.30041.

[25] Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, et al. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbc461518c59a2a8d64e70e2c38c4a0e-Abstract-Conference.html.

[26] Guibin Zhang, Yanwei Yue, Xiangguo Sun, Guancheng Wan, Miao Yu, Junfeng Fang, et al. G-Designer: Architecting multi-agent communication topologies via graph neural networks. In *International Conference on Machine Learning (ICML)*, 2025. URL https://icml.cc/virtual/2025/poster/45567.

[27] Wei Zhao, Zhe Li, Peixin Zhang, and Jun Sun. ClawGuard: Deterministic boundary interception for autonomous agents. arXiv preprint arXiv:2604.11790, 2026. URL https://arxiv.org/abs/2604.11790.

[28] Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, et al. Latent collaboration in multi-agent systems. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2511.20639.