Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Semantic and Risk-Aware Handoff Coordination for Cooperative Perception in Highway V2XNetworks with Strict Tail-Latency Requirements

Roshan Chitranshi

roshanchitranshi@united.ac.in

United College of Engineering and Research, Prayagraj

Mohit Jain

United College of Engineering and Research, Prayagraj

Vineet Kumar Tiwari

United College of Engineering and Research, Prayagraj

Research Article

Keywords: Cooperative perception, V2X, group handoff, handoff agent, semantic communication, conditional value-at-risk, M/G/1 queue, Lundberg inequality, tail latency, resource allocation

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11230849/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Semantic and Risk-Aware Handoff Coordination for Cooperative Perception in Highway V2X Networks with Strict Tail-Latency Requirements

Roshan Chitranshi*, Mohit Jain, Vineet Kumar Tiwari*

Department of Electrical Engineering, United College of Engineering & Research, Prayagraj, India

roshanchitranshi@united.ac.in, mohitjain@united.ac.in, vineettiwarien@gmail.com**

***Abstract*— On highways, cooperative perception depends on connected vehicles sharing urgent descriptions of nearby scenes as they pass through overlapping coverage from neighbouring next-generation base stations (gNBs). During that short interval, perception messages and handoff signalling draw on the same limited pool of radio resources. Current group-handoff methods improve association, balance network load, or reduce average delay, yet usually assume a fixed perception payload and do not control the rare delay spikes that matter most for safety. We introduce *Semantic Risk-Aware Handoff Orchestration* (SRAHO), a cross-layer controller that makes three decisions together: it chooses the smallest semantic payload that meets a required perception-retention level; selects direct vehicle-to-infrastructure (V2I) transmission or handoff-agent (HOA) relaying while accounting for mobility, fading, and the possibility that a helper leaves; and assigns resource blocks (RBs) to minimise the Conditional Value-at-Risk (CVaR) of handoff delay. Our analysis derives closed-form rate moments and high- and low-SNR limits for Rayleigh fading, exact upper and lower bounds for fading-truncated service moments and their quantile-grid scaling, and a Gaussian-margin test for the persistence of a V2V link. It also proves strict convexity of M/G/1 handoff delay with respect to RB allocation, gives a square-root spare-capacity rule for mean-optimal allocation, and obtains a closed-form Lundberg-type upper bound linking queueing-delay CVaR to its mean. For separable discretely convex objectives, the CVaR-greedy allocator is exact; a computable a posteriori certificate bounds any remaining optimality gap. SRAHO delivers the lowest CVaR0.95 handoff delay in all eight stress scenarios. Under interference, it cuts CVaR0.95 from 60.01 ms with Direct V2I to 51.18 ms, a 14.72% reduction, while handoff success rises from 0.955 to 0.983. Semantic adaptation also reduces the payload per step from 100 608 to 69 406 bits, or 31.0%. In model-based tests under interference, CVaR-based allocation reduces deadline violations by 29% against mean-delay allocation. The greedy method matches exhaustive optimization in 1198 of 1200 cases, and the full decision for 12 vehicles takes 8.85 ms.**

***Keywords*—Cooperative perception, V2X, group handoff, handoff agent, semantic communication, conditional value-at-risk, M/G/1 queue, Lundberg inequality, tail latency, resource allocation.**

I. Introduction

Connected and automated vehicles increasingly rely on cooperative perception to extend situational awareness beyond the line of sight of their on-board sensors [1]–[4]. On a highway, however, the exchange of perception data is not a stationary process: a vehicle entering the overlap of two gNBs must simultaneously sustain its perception stream and complete a handoff (HO). Independent V2I handoff of every vehicle generates redundant signalling and contention for resource blocks (RBs), whereas cluster- or relay-assisted handoff concentrates the load on a coordinator whose queue, channel, and residence time in the overlap may all deteriorate abruptly [5], [6]. For perception traffic these effects are amplified, because the payload is one to two orders of magnitude larger than conventional control signalling and its informational value decays with age [7], [8].

Two structural shortcomings of conventional handoff optimisation motivate this work. First, the communication payload is almost universally regarded as fixed. In cooperative perception, however, scene regions differ markedly in hazard relevance, epistemic uncertainty, novelty, and freshness; transmitting the complete scene at every handoff therefore consumes RBs and inflates queueing delay even when only a fraction of the scene is required by the downstream perception task [9], [10]. Second, optimising the *mean* delay conceals the rare but severe excursions that dominate operational risk in safety-critical vehicular services [11], [12]. A controller that is optimal on average may still violate a latency deadline with unacceptable frequency.

SRAHO resolves both shortcomings within a single decision architecture. It first computes the smallest semantic payload that preserves a prescribed perception-retention floor, then evaluates direct and HOA-relayed paths using mobility, fading, queue, and helper-survival information, and finally distributes the residual RB budget according to the marginal reduction in the CVaR of the pooled handoff delay. The controller therefore targets the upper tail of the delay distribution rather than its first moment.

A. Related Work

1) Group handoff and mobility management: Delegating handoff responsibility to selected vehicles can reduce repeated signalling and distribute the control burden. The evolutionary-game mechanism in [5] elects multiple handoff agents (HOAs), models their transmit queues, and allocates RBs to reduce waiting time, while [6] considers power-efficient handoff management with game-theoretic resource allocation. More recent V2N handover work has used learning-based adaptation to cope with changing mobility and radio conditions [32], and

deep reinforcement learning has been used to select vehicular relays jointly with beam management under blockage-induced link outages [40]. These studies motivate mobility-aware handoff control, but they generally treat the payload as fixed and optimize association, signalling load, or average delay rather than the upper tail of perception-transfer latency.

2) Communication-efficient cooperative perception: Cooperative-perception systems such as Cooper [1], V2VNet [2], OPV2V [3], and V2X-ViT [4] established the benefit of exchanging raw or intermediate features. When2com [13] and Where2comm [9] subsequently reduced communication by selecting when, with whom, or where to communicate. Recent IEEE Transactions work has moved further toward task-aware compression: spatial-temporal feature compression jointly reduces traffic while preserving 3D detection performance [33], and point-cloud value assessment has been used to reduce bandwidth consumption without sacrificing perception accuracy [34]. Joint compression, transmission, and computation optimization has also been studied for delay-sensitive vehicular cooperative perception [35]. These approaches demonstrate that perception traffic should be adapted to task utility, but they do not jointly model handoff mobility, HOA survival, queueing, and tail-risk-aware RB allocation.

3) Semantic and task-oriented communication: Semantic communication replaces bit-level reconstruction as the primary objective with task-level utility [10], [14]. Recent systems dynamically drop low-importance features or vary the number of transmitted semantic symbols as channel and task conditions change [36], while vehicular semantic communication has been explicitly studied for stringent xURLLC service provisioning [37]. Semantic feature scheduling has also been combined with rate control to reduce latency subject to a task-performance constraint [38]. These developments support the semantic-retention principle used in SRAHO. However, semantic selection has rarely been embedded into a handoff controller that also reasons about helper mobility, queue stability, and delay-tail risk.

4) Tail-latency and risk-sensitive design: URLLC research has shown that mean latency alone is insufficient when rare delay excursions determine reliability [11], and extreme-value methods have been used to characterize vehicular queue tails [12]. CVaR is a coherent [15], convex, and sample-tractable upper-tail risk measure [16]-[18]. SRAHO uses this perspective at the handoff-orchestration level: the radio-resource decision is driven by the high-delay samples rather than only by the first moment. This distinguishes the proposed controller from allocation schemes that minimise mean delay.

5) Queueing, resource allocation, and fading: The M/G/1 queue and the Pollaczek-Khinchine formula connect traffic intensity and service-time variability to delay [19], while Rayleigh fading governs link-rate variability [20]. Classical risk theory provides exponential waiting-time bounds through the Lundberg adjustment coefficient [21]. Recent V2X work further emphasizes that resource-allocation decisions must be made quickly under rapidly varying channel information; for example, prediction-aided V2X resource allocation has been used to reduce processing time while maintaining near-optimal communication performance [39]. Reliability has also been imposed directly as a constraint in C-V2X radio-resource allocation for multicast automotive messages [41], although as a probability constraint rather than a tail-risk objective. These results motivate the joint treatment of fading, queue stability, computational burden, and handoff risk adopted below.

B. Research Gap

The literature therefore leaves a cross-layer gap rather than an isolated algorithmic gap. During the short overlap interval between neighbouring gNBs, a vehicle must decide (i) how much task-relevant perception information to transmit, (ii) whether the packet should use direct V2I or an HOA-assisted two-hop path, and (iii) how the finite RB budget should be distributed. These decisions are coupled through mobility, fading, helper residence time, queue stability, and the upper tail of handoff delay. Existing handoff studies do not adapt the perception payload; communication-efficient perception studies do not model handoff queues and mobility; and risk-sensitive latency studies do not jointly select semantic payload, helper path, and RBs. Table I therefore compares SRAHO with individual representative works along the six capabilities that the proposed problem jointly requires, namely handoff mobility, semantic payload adaptation, HOA-assisted relaying, queueing, tail-risk (CVaR) optimisation, and resource allocation.

TABLE I: POSITIONING OF SRAHO RELATIVE TO REPRESENTATIVE PRIOR WORK

Representative work	HO mobility	Semantic payload	HOA / relay	Queue model	Tail/CVaR risk	Joint RB allocation
[5] EGDH / HOA handoff	Yes	No	Yes	Yes	No	Yes
[32] Meta-transfer V2N handover	Yes	No	No	No	No	No
[9] Where2comm	No	Yes	No	No	No	No
[33] Spatial-temporal CP compression	No	Yes	No	No	No	No
[34] Point-cloud value / incentive CP	No	Yes	No	No	No	Partial
[35] Joint compression-transmission-computation	No	Yes	No	Partial	No	Yes
[37] Vehicular semantic xURLLC	Partial	Yes	No	Partial	Yes	Yes
[38] Semantic feature scheduling	No	Yes	No	No	No	Yes

[11], [12] tail-latency / EVT	No	No	No	Yes	Yes	Partial
[40] DRL relay selection, mmWave V2X	Partial	No	Yes	No	No	No
[41] Reliability-constrained C-V2X RA	No	No	No	No	Partial	Yes
SRAHO (this work)	Yes	Yes	Yes	Yes	Yes	Yes

C. Contributions

The principal contributions are as follows.

1. *Semantic payload adaptation with optimality guarantees.* Scene regions are ranked by hazard relevance, uncertainty, novelty, age, and cost. We prove that top-K selection is bit-optimal for any retention floor, that the retained-priority mass is discretely concave in the payload, and we bound the overhead introduced by level quantisation (Proposition 3).
2. *Mobility- and fading-aware link analysis.* We derive a Gaussian-margin certificate for V2V link persistence over the look-ahead horizon (Lemma 1), closed-form rate moments with high- and low-SNR asymptotics (Proposition 1), a geometric criterion for when relaying improves connectivity (Corollary 1), and exact two-sided bounds together with quantile-grid scaling laws for the fading-truncated service moments (Propositions 2 and 7).
3. *Queue-aware delay structure and tail bound.* For the multiclass M/G/1 handoff queue we establish strict convexity and monotonicity of the delay in the RB allocation, a superquadratic payload-scaling law for the waiting time, a square-root spare-capacity rule for the mean-optimal allocation, and a Lundberg-type closed-form upper bound on the CVaR of the waiting time in terms of its mean (Theorem 1).
4. *Risk-aware orchestration with certificates.* We formulate the joint helper, payload, RB, and power problem as a CVaR-minimising mixed-integer programme, prove the convexity of the sampled CVaR objective, justify the survival-adjusted path cost via Wald's identity, prove exactness of the marginal-greedy allocator for separable discretely convex objectives (Theorem 2), and supply a computable a posteriori optimality-gap certificate (Proposition 9).
5. *Comprehensive evaluation.* SRAHO is compared with Direct V2I, Cluster-head, EGDH-original, and EGDH-corrected baselines over eight stress scenarios, complemented by ablation and sensitivity studies that isolate the contributions of semantic adaptation and risk-aware RB allocation, and by a model-based validation that compares CVaR-driven with mean-delay-driven allocation, measures the optimality gap of the greedy allocator against exhaustive search, and reports the execution time of every stage.

Organisation: Section II presents the system, mobility, channel, and semantic-payload models. Section III develops the queue-aware handoff-delay model and its structural properties. Section IV formulates the tail-risk problem and the SRAHO procedure. Section V reports numerical results, and Section VI concludes. Proofs are collected in the Appendix.

Notation: $\mathbb{E}[\cdot]$ and $\mathbb{P}(\cdot)$ denote expectation and probability; $(x)^+ = \max\{x, 0\}$; $\mathbb{1}\{\cdot\}$ is the indicator; $\mathrm{Exp}(1)$ is the unit-mean exponential law; $Q(\cdot)$ is the Gaussian tail function; $E_1(x) = \int_x^\infty s^{-1} e^{-s}\, ds$ is the exponential integral; $\gamma_{\mathrm{E}} \approx 0.5772$ is the Euler–Mascheroni constant; $H_n = \sum_{m=1}^{n} m^{-1}$; and $\mathbf{e}_i$ is the ith canonical basis vector. The principal symbols are summarised in Table II.

TABLE II: PRINCIPAL NOTATION

Symbol	Meaning
$N, \mathcal{N}$	Number and set of vehicles in the HO region
$R_{\mathrm{tot}}, B_{\mathrm{RB}}$	Total RBs; bandwidth of one RB
r_i, P_i	RBs and transmit power of vehicle i
$h_i, \mathcal{J}_i$	Selected helper (-1: direct V2I); candidate path set
$\bar{\gamma}_{u,v}$	Average SINR of link (u, v)
a, μ_1, μ_2	Service factor [s/(bit·RB)] and its truncated moments
$\pi_{i,k}$	Semantic priority of region k of vehicle i
$B_i(\ell), \Phi_i(\ell)$	Payload and retained priority mass at level ℓ
λ_i, ρ_i	Packet arrival rate and queue utilisation
S_i, W_i, T_i	Service, waiting, and sojourn times
D_i	Total handoff delay
β	CVaR confidence level
$\theta^\star$	Lundberg adjustment coefficient
p_{dep}	Helper-departure probability
$d_{\mathrm{V2V}}, \tau, z_{\mathrm{m}}$	V2V range; look-ahead horizon; margin quantile
v_u, σ_a	Vehicle velocity; acceleration standard deviation
$N_0, I, \gamma_{\mathrm{th}}$	Noise and interference power per RB; SINR threshold
$K_0, \ell_i^\star, \Phi_{\min}$	Scene regions; selected semantic level; perception floor
$D_{\mathrm{fix}}, D_{\max}, \kappa$	Protocol delay; handoff deadline; cap factor
M, ξ_m	Fading-grid size; grid points
$H, \mathcal{H}$	Number and set of HOAs

Throughout the remainder of the paper, the subscripts i and j identify vehicles and b identifies a gNB; r_i denotes an integer RB count, B_i the packet payload in bits, and $\bar{\gamma}$ the average SINR after deterministic interference is included. Each symbol is used only with the meaning listed in Table II, and link-specific quantities carry the corresponding (u, v) subscripts.

II. SYSTEM MODEL

A. Network and Handoff Architecture

Consider N vehicles, indexed by $\mathcal{N} = \{1, \ldots, N\}$, travelling on a one-dimensional highway segment of length L served by two overlapping gNBs (Fig. 1). Vehicles in the overlap form HO groups in the sense of [5]: each vehicle either uploads its semantic perception packet directly to the target gNB (V2I) or forwards it over a V2V link to an HOA, which relays it to the gNB. A total of R_{tot} RBs, each of bandwidth B_{RB}, is available for the HO epoch [22].

Figure 1 makes the physical decision process explicit. A vehicle entering the overlap must continue cooperative-perception delivery while transferring from the serving gNB to the target gNB. SRAHO permits direct V2I transmission or an HOA-assisted V2V-to-V2I path. Because both modes draw from the same finite RB budget, the semantic payload, helper

path, and RB allocation are coupled through the resulting service time and handoff deadline.

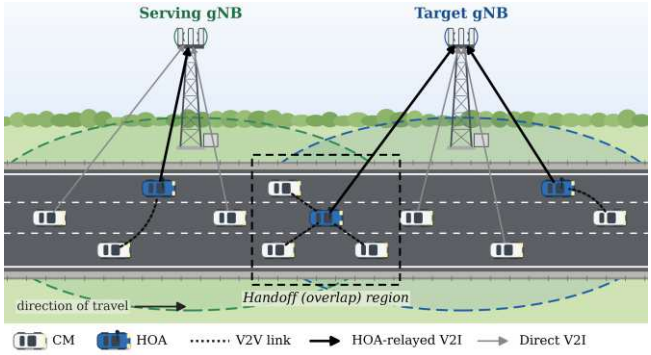


Fig. 1. SRAHO highway V2X system model. Vehicles traverse the overlap between a serving and a target gNB. Cluster members may use direct V2I transmission or a V2V-to-HOA-to-gNB relay path. The dashed box identifies the handoff region in which semantic payload, helper selection, and RB allocation are jointly decided. White vehicles are cluster members (CMs) and blue vehicles are HOAs; dotted lines denote V2V links, black arrows HOA-relayed V2I links, and grey arrows direct V2I links.

B. Mobility Model and Probabilistic V2V Feasibility

Because an HOA is useful only while the V2V link remains available, the first modelling step is to translate vehicle motion into a probabilistic relay-feasibility condition over the handoff look-ahead horizon.

Vehicle u has initial position r_u and velocity $v_u \sim \mathcal{U}[v_{\min}, v_{\max}]$. Over a slot of duration Δt, the inter-vehicle distance and the distance to a gNB located at y_b are

$$\begin{aligned} d_{u,v}(t) &= |(r_u - r_v) + (v_u - v_v)t|, \\ d_{v,b}(t) &= |r_v + v_v t - y_b|, \end{aligned} \tag{1}$$

with positions wrapped modulo L. Accelerations $a_u \sim \mathcal{N}(0, \sigma_a^2)$, independent across vehicles, enter only as a prediction margin. For a look-ahead horizon τ and Gaussian quantile z_{m}, the risk distance and the V2V feasibility event are

$$\begin{aligned} d_{u,v}^{\mathrm{risk}} &= |(r_u - r_v) + (v_u - v_v)\tau| + z_{\mathrm{m}}\tau^2\sigma_a, \\ \mathcal{A}_{u,v} &= \{d_{u,v}^{\mathrm{risk}} \le d_{\mathrm{V2V}}\}. \end{aligned} \tag{2}$$

The following lemma quantifies the reliability implied by $\mathcal{A}_{u,v}$.

Lemma 1 (*V2V persistence certificate*)**:** Let $\Delta_{u,v}(t) = (r_u - r_v) + (v_u - v_v)t + 1/2\,(a_u - a_v)t^2$ and $m_{u,v}(t) = \mathbb{E}[\Delta_{u,v}(t)]$. Then, exactly,

$$\begin{aligned} \mathbb{P}\big(|\Delta_{u,v}(\tau)| \le d_{\mathrm{V2V}}\big) &= \Phi_{\mathcal{N}}\left(\frac{d_{\mathrm{V2V}} - m_{u,v}(\tau)}{\varsigma_\tau}\right) \\ &\quad - \Phi_{\mathcal{N}}\left(\frac{-d_{\mathrm{V2V}} - m_{u,v}(\tau)}{\varsigma_\tau}\right), \end{aligned} \tag{3}$$

where $\varsigma_\tau = \tau^2\sigma_a/\sqrt{2}$ and $\Phi_{\mathcal{N}}$ is the standard normal CDF. Moreover, (a) if $\mathcal{A}_{u,v}$ holds, then

$$\mathbb{P}\big(|\Delta_{u,v}(\tau)| > d_{\mathrm{V2V}}\big) \le 2\,Q\big(\sqrt{2}\,z_{\mathrm{m}}\big); \tag{4}$$

(b) if, in addition, $|r_u - r_v| + z_{\mathrm{m}}\tau^2\sigma_a \le d_{\mathrm{V2V}}$, then the bound (4) holds uniformly for the event $\{\exists\, t \in [0,\tau]: |\Delta_{u,v}(t)| > d_{\mathrm{V2V}}\}$.

For $z_{\mathrm{m}} = 1.645$, (4) certifies a V2V persistence probability of at least 0.98; the factor $\sqrt{2}$ arises because the margin $z_{\mathrm{m}}\tau^2\sigma_a$ in (2) is calibrated to a single acceleration, whereas the relative acceleration has variance $2\sigma_a^2$.

C. Channel Model and Rate Statistics

After screening paths for mobility persistence, the physical-layer service rate is determined by path loss, Rayleigh fading, transmit power, interference, and the number of RBs assigned to the transmitting vehicle.

All links experience Rayleigh block fading. The power gain of link (u, v) is

$$G_{u,v} = G_0 \max\big(d_{u,v}, 1\big)^{-\alpha} h_{u,v}, \quad h_{u,v} \sim \mathrm{Exp}(1), \tag{5}$$

where G_0 is the reference gain and α the path-loss exponent. With transmit power P_u, noise power N_0 per RB, and interference I per RB, the average SINR and the achievable rate on r_u RBs are

$$\bar{\gamma}_{u,v} = \frac{P_u G_0 d_{u,v}^{-\alpha}}{N_0 + I}, \tag{6}$$

$$R_{u,v} = r_u B_{\mathrm{RB}} \log_2\big(1 + \bar{\gamma}_{u,v} h_{u,v}\big). \tag{7}$$

Proposition 1 (*Rate moments*)**:** The first two moments of $R_{u,v}$ are

$$\frac{\mathbb{E}[R_{u,v}]}{r_u B_{\mathrm{RB}}} = \frac{e^{1/\bar{\gamma}_{u,v}}}{\ln 2} E_1\left(\frac{1}{\bar{\gamma}_{u,v}}\right) = -\frac{e^{1/\bar{\gamma}_{u,v}}}{\ln 2}\,\mathrm{Ei}\left(-\frac{1}{\bar{\gamma}_{u,v}}\right), \tag{8}$$

$$\frac{\mathbb{E}[R_{u,v}^2]}{(r_u B_{\mathrm{RB}})^2} = \frac{2e^{1/\bar{\gamma}_{u,v}}}{(\ln 2)^2} \int_1^\infty \frac{\ln s}{s}\, e^{-s/\bar{\gamma}_{u,v}}\, ds, \tag{9}$$

and $\mathrm{Var}[R_{u,v}] = \mathbb{E}[R_{u,v}^2] - \mathbb{E}[R_{u,v}]^2$. Furthermore, writing $\bar{\gamma} = \bar{\gamma}_{u,v}$ and $h = h_{u,v}$,

$$\mathbb{E}[\ln(1 + \bar{\gamma} h)] = \begin{cases} \ln\bar{\gamma} - \gamma_{\mathrm{E}} + O(\bar{\gamma}^{-1}\ln\bar{\gamma}), & \bar{\gamma} \to \infty, \\ \bar{\gamma} - \bar{\gamma}^2 + 2\bar{\gamma}^3 + O(\bar{\gamma}^4), & \bar{\gamma} \to 0, \end{cases} \tag{10}$$

$$\lim_{\bar{\gamma}\to\infty} \frac{\mathrm{Var}[R_{u,v}]}{(r_u B_{\mathrm{RB}})^2} = \frac{\pi^2}{6(\ln 2)^2}, \tag{11}$$

and, by Jensen's inequality, $\mathbb{E}[R_{u,v}] \le r_u B_{\mathrm{RB}} \log_2(1 + \bar{\gamma})$.

At high SNR, (10) shows that Rayleigh fading costs a constant $\gamma_{\mathrm{E}}/\ln 2 \approx 0.833$ bit/s/Hz relative to the unfaded link, while (11) shows that the rate variance saturates at $\pi^2/(6\ln^2 2) \approx 3.42$ (bit/s/Hz)^2 irrespective of the SNR; rate *variability*, and hence queueing variability, cannot be removed by power alone. Fig. 2 validates (8).

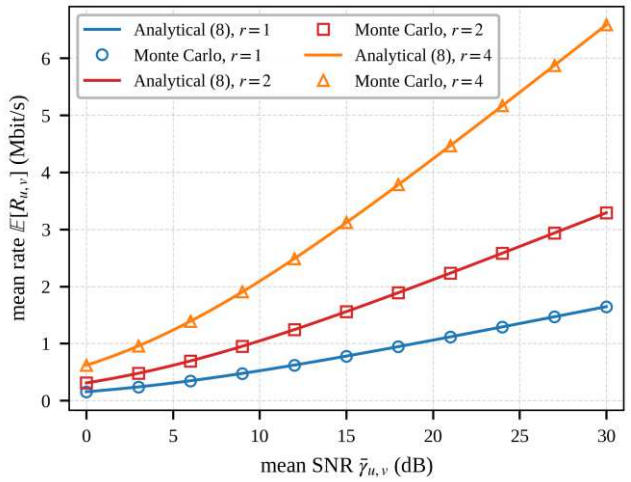


Fig. 2. Mean rate versus average SINR for one, two, and four RBs: analytical expression (8) (lines) against Monte Carlo with 4×10^5 fading draws (markers).

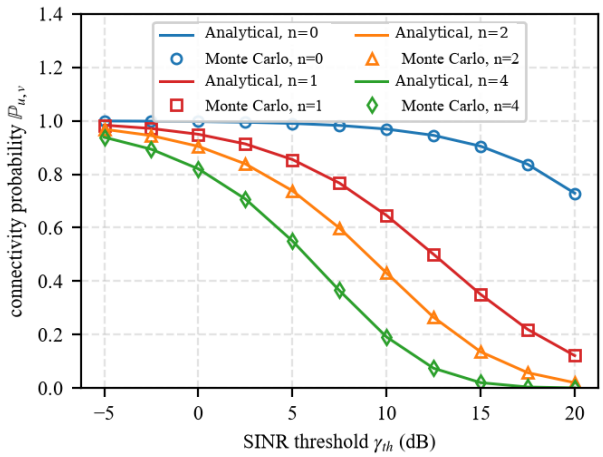


Fig. 3. Connectivity probability versus SINR threshold for $n = 0,1,2,4$ interferers: analytical expression (12) (lines) against Monte Carlo (markers).

D. Connectivity, Outage, and Relaying

The instantaneous rate model is next converted into a link-reliability model so that direct and two-hop relay candidates can be compared using both connectivity and delay.

Let the desired received power be exponential with mean $\bar{X}_{u,v} = P_u G_0 d_{u,v}^{-\alpha}$, and let n independent interferers have exponential received powers Y_j with means Λ_j. For an SINR threshold γ_{th}, the connectivity probability is

$$\mathbb{P}_{u,v} = \exp\left(-\frac{\gamma_{\mathrm{th}} N_0}{\bar{X}_{u,v}}\right) \prod_{j=1}^{n} \left(1 + \frac{\gamma_{\mathrm{th}} \Lambda_j}{\bar{X}_{u,v}}\right)^{-1}. \tag{12}$$

With deterministic aggregate interference I, the product is replaced by $\exp\left(-\gamma_{\mathrm{th}} I / \bar{X}_{u,v}\right)$, which is the form used in the simulator. The outage probability is $\mathbb{P}_{u,v}^{\mathrm{out}} = 1 - \mathbb{P}_{u,v}$, and a two-hop path through helper j succeeds with probability

$$\mathbb{P}_{i \to j \to b} = \mathbb{P}_{i,j}\, \mathbb{P}_{j,b}. \tag{13}$$

Corollary 1 (*Diversity and relaying criterion*)**:** (a) The outage probability satisfies

$$\mathbb{P}_{u,v}^{\mathrm{out}} \le \frac{\gamma_{\mathrm{th}}}{\bar{X}_{u,v}} \left(N_0 + \sum_{j=1}^{n} \Lambda_j\right), \tag{14}$$

so each hop has diversity order one, and $1 - \mathbb{P}_{i \to j \to b} \le \mathbb{P}_{i,j}^{\mathrm{out}} + \mathbb{P}_{j,b}^{\mathrm{out}}$. (b) Under deterministic interference I common to all links, relaying through j improves connectivity over direct transmission if and only if

$$\frac{d_{i,j}^{\alpha}}{P_i} + \frac{d_{j,b}^{\alpha}}{P_j} < \frac{d_{i,b}^{\alpha}}{P_i}. \tag{15}$$

For $P_i = P_j$ and a helper located collinearly between i and b, (15) always holds for $\alpha > 1$ by strict superadditivity of $x \mapsto x^{\alpha}$; relaying benefits therefore stem from geometry, and they are eroded when interference inflates both hops, as observed in Section V. Fig. 3 validates (12).

E. Fading-Truncated Service Factor

Queueing requires moments of the packet service time rather than only the mean data rate. Under ideal Shannon transmission and Rayleigh fading, deep fades make the reciprocal-rate moments singular; SRAHO therefore uses a physically interpretable truncation induced by the deterministic fading grid.

Define the per-bit, per-RB service factor $a = 1/\left(B_{\mathrm{RB}} \log_2(1 + \bar{\gamma} h)\right)$ with $h \sim \mathrm{Exp}(1)$, and let $c_{\bar{\gamma}} = \ln 2 / (B_{\mathrm{RB}} \bar{\gamma})$.

Proposition 2 (*Truncated service moments*)**:** $\mathbb{E}[a] = \infty$. For every truncation level $\varepsilon > 0$,

$$c_{\bar{\gamma}} E_1(\varepsilon) \le \mathbb{E}[a\, \mathbb{1}\{h \ge \varepsilon\}] \le c_{\bar{\gamma}} (E_1(\varepsilon) + \bar{\gamma} e^{-\varepsilon}), \tag{16}$$

$$\begin{aligned} c_{\bar{\gamma}}^2 \Psi(\varepsilon) &\le \mathbb{E}[a^2 \mathbb{1}\{h \ge \varepsilon\}] \\ &\le c_{\bar{\gamma}}^2 (\Psi(\varepsilon) + 2\bar{\gamma} E_1(\varepsilon) + \bar{\gamma}^2 e^{-\varepsilon}), \end{aligned} \tag{17}$$

where $\Psi(\varepsilon) = e^{-\varepsilon}/\varepsilon - E_1(\varepsilon)$. Consequently, as $\varepsilon \to 0$,

$$\begin{aligned} \mathbb{E}[a\, \mathbb{1}\{h \ge \varepsilon\}] &= c_{\bar{\gamma}} \left(\ln \varepsilon^{-1} + O(1)\right), \\ \mathbb{E}[a^2 \mathbb{1}\{h \ge \varepsilon\}] &= c_{\bar{\gamma}}^2\, \varepsilon^{-1} \left(1 + o(1)\right). \end{aligned} \tag{18}$$

The exact mean service time of a Shannon-rate link under Rayleigh fading is therefore infinite, and all moments below refer to a fading-truncated service time. Proposition 2 sharpens the qualitative statement into explicit two-sided bounds and reveals an asymmetry of practical consequence: the first moment grows only logarithmically in $1/\varepsilon$, whereas the second moment, which drives the Pollaczek–Khinchine waiting time, grows *linearly* in $1/\varepsilon$ (Fig. 4).

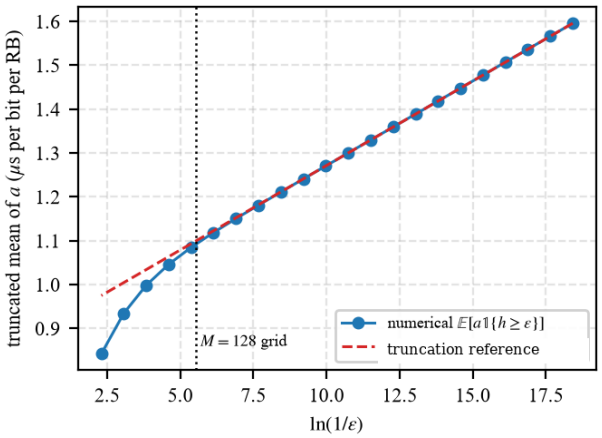


Fig. 4. Truncated mean of the service factor a versus $\ln(1/\varepsilon)$. The curve grows approximately linearly in $\ln(1/\varepsilon)$, in agreement with (18); the dotted line marks the truncation level of the $M = 128$ quantile grid.

F. Semantic Perception Payload Adaptation

The communication load is not treated as fixed. Before path and RB decisions are made, each vehicle ranks perception regions by task relevance and chooses the smallest semantic level that meets the application retention floor.

Vehicle i partitions its scene into K_0 regions. Region k is assigned the priority

$$\pi_{i,k} = \Pi_{[0,1]}(w_h H_{i,k} + w_u U_{i,k} + w_n N_{i,k} + w_a \tilde{A}_i - w_c C_{i,k}), \tag{19}$$

with $(w_h, w_u, w_n, w_a, w_c) = (0.45, 0.20, 0.15, 0.20, 0.10)$, hazard weight $H_{i,k}$, uncertainty $U_{i,k}$, novelty $N_{i,k}$, normalised age $\tilde{A}_i = \min\{1, A_i/D_{\max}\}$ [7], [8], cost $C_{i,k}$, and $\Pi_{[0,1]}$ the Euclidean projection onto $[0,1]$. Level $\ell \in \{0, \dots, \mathcal{L}-1\}$ retains $K_i(\ell) = \max\{1, \lceil K_0 f_\ell \rceil\}$ regions with f_ℓ strictly increasing and $f_{\mathcal{L}-1} = 1$. Each region carries a b_{idx}-bit index and a latent vector of d_z symbols of b_s bits, so that the per-region cost is $c_{\text{reg}} = b_{\text{idx}} + d_z b_s$. With ordered priorities $\pi_i^{(1)} \geq \cdots \geq \pi_i^{(K_0)}$ and $\Pi_i^{\text{tot}} = \sum_{k=1}^{K_0} \pi_i^{(k)} > 0$, the payload, the retained priority mass, and the perception loss are

$$B_i(\ell) = B_{\text{hdr}} + K_i(\ell)\, c_{\text{reg}}, \tag{20}$$

$$\Phi_i(\ell) = \frac{1}{\Pi_i^{\text{tot}}} \sum_{k=1}^{K_i(\ell)} \pi_i^{(k)}, \quad \mathcal{L}_i(\ell) = 1 - \Phi_i(\ell). \tag{21}$$

SRAHO selects the smallest level meeting the perception floor $\Phi_{\min}$,

$$\ell_i^\star = \min\{\ell : \Phi_i(\ell) \geq \Phi_{\min}\}. \tag{22}$$

Because $\Phi_i(\mathcal{L}-1) = 1$, the feasible set in (22) is nonempty.

Proposition 3 (*Semantic selection*)**:** Let $K_i^\star = \min\{K : \sum_{k \leq K} \pi_i^{(k)} \geq \Phi_{\min} \Pi_i^{\text{tot}}\}$. Then: (a) among all region subsets whose retained mass is at least $\Phi_{\min}$, the minimum number of bits is $B_{\text{hdr}} + K_i^\star c_{\text{reg}}$, attained by the top-$K_i^\star$ regions; (b) the retention curve $K \mapsto \Phi_i(K)$ is nondecreasing and discretely concave, i.e., its increments $\pi_i^{(K+1)}/\Pi_i^{\text{tot}}$ are nonincreasing, so the retained mass per transmitted bit is nonincreasing; (c) the quantisation overhead of (22) satisfies

$$0 \leq B_i(\ell_i^\star) - B_{\text{hdr}} - K_i^\star c_{\text{reg}} \leq (K_i(\ell_i^\star) - K_i(\ell_i^\star - 1) - 1) c_{\text{reg}} \tag{23}$$

for $\ell_i^\star \geq 1$, and $\ell_i^\star$ minimises the payload among feasible levels.

Part (b) formalises the diminishing informational return of additional regions and underlies the delay–perception trade-off examined in Fig. 16. Fig. 5 shows the payload and the retained mass versus ℓ for the event mix of Section V.

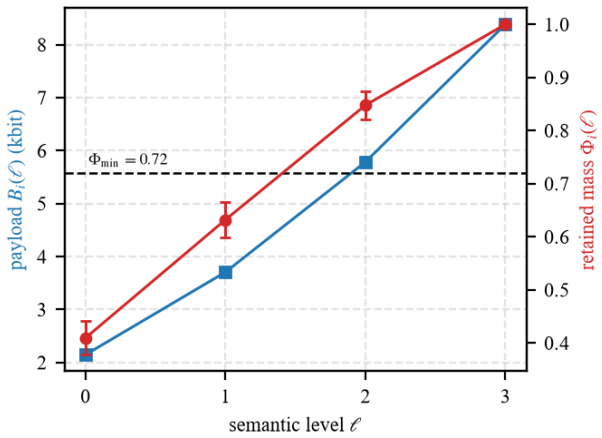


Fig. 5. Payload $B_i(\ell)$ and retained priority mass $\Phi_i(\ell)$ (mean $\pm$ one standard deviation over 2000 random scenes) for the four semantic levels. The dashed line is the perception floor $\Phi_{\min} = 0.72$.

III. Queue-Aware Handoff Delay Model

Section II characterised the service capability of each individual link. This section combines those link statistics with packet arrivals to obtain the handoff delay experienced by each vehicle: it first models the direct and relay transmit queues, then assembles the total handoff delay, and finally establishes the structural properties of the delay that guide the resource allocation in Section IV.

A. Arrivals, Service, and the Multiclass Helper Queue

With mobility, channel, and semantic payload defined, the next step is to map each selected path into a queueing delay. Packet arrivals are modelled at the transmitter, while an HOA aggregates the streams of all members that it serves.

Packets of vehicle i arrive as a Poisson process of rate λ_i. A helper j serving the member set $\mathcal{M}_j$ observes the superposed Poisson stream

$$\lambda_j^{\text{agg}} = \sum_{i \in \mathcal{M}_j \cup \{j\}} \lambda_i = m_j \lambda_0 \ \text{(homogeneous case)}, \tag{24}$$

and every transmit queue is modelled as an M/G/1 FCFS queue [5], [19] (Fig. 6). Let $a_{u,v}$ denote the service factor of link (u, v) evaluated on the fading grid of Section IV-C, and let $\mu_1^{u,v}$ and $\mu_2^{u,v}$ be its two truncated moments. The service times of a B_i-bit packet on the direct path and on the relay path are

$$S_i = \frac{B_i a_{i,b}}{r_i}, \quad S_i^{\mathrm{rel}} = B_i\left(\frac{a_{i,j}}{r_i} + \frac{a_{j,b}}{r_j}\right). \tag{25}$$

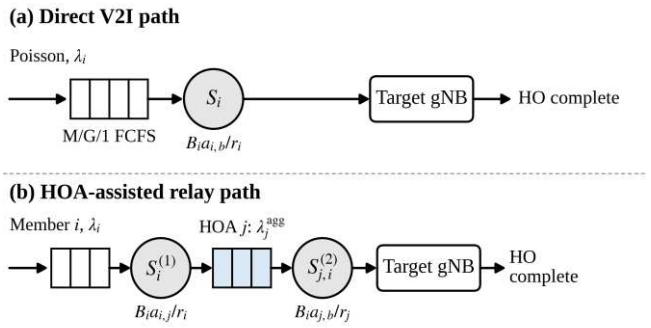


Fig. 6. Queueing representation used by SRAHO. A direct V2I packet experiences one M/G/1 FCFS service stage, whereas an HOA-assisted packet traverses a V2V service stage followed by the HOA aggregate queue and the HOA-to-gNB service stage. Service times depend on semantic payload, fading, and allocated RBs. The label below each server gives the service time of (25) and Section III-A.

On the direct path, the utilisation and the Pollaczek–Khinchine mean waiting time are

$$\rho_i = \lambda_i \mathbb{E}[S_i] = \frac{\lambda_i B_i \mu_1}{r_i}, \qquad W_i = \frac{\lambda_i \mathbb{E}[S_i^2]}{2(1-\rho_i)} = \frac{\lambda_i B_i^2 \mu_2}{2 r_i (r_i - \lambda_i B_i \mu_1)}, \tag{26}$$

and the queue is stable if and only if $\rho_i < 1$. The mean sojourn time and Little's law give

$$T_i = \mathbb{E}[S_i] + W_i, \quad \bar{L}_i = \lambda_i T_i. \tag{27}$$

At a helper, packets of different members form a multiclass stream with class-dependent service $S_{j,i}^{(2)} = B_i a_{j,b}/r_j$. The exact multiclass Pollaczek–Khinchine formula gives

$$\rho_j^{\mathrm{H}} = \frac{\mu_1^{j,b}}{r_j} \sum_{i \in \mathcal{M}_j \cup \{j\}} \lambda_i B_i, \qquad W_j^{\mathrm{H}} = \frac{\mu_2^{j,b} \sum_i \lambda_i B_i^2}{2 r_j \left(r_j - \mu_1^{j,b} \sum_i \lambda_i B_i\right)}, \tag{28}$$

and, under Kleinrock's independence approximation for the store-and-forward tandem [19], the relay sojourn time is $T_i^{\mathrm{rel}} \approx T_i^{(1)} + B_i \mu_1^{j,b}/r_j + W_j^{\mathrm{H}}$, where $T_i^{(1)}$ is the first-hop sojourn obtained from (26)–(27) with $a_{i,j}$. Stability on the direct path is equivalent to the integer requirement

$$r_i > \lambda_i B_i \mu_1 \Leftrightarrow r_i \geq r_i^{\min} = \lfloor \lambda_i B_i \mu_1 \rfloor + 1, \tag{29}$$

and the analogous helper requirement is $r_j \geq \lfloor \mu_1^{j,b} \sum_i \lambda_i B_i \rfloor + 1$. A stable joint allocation exists if and only if $\sum_i r_i^{\min} \leq R_{\mathrm{tot}}$.

B. Handoff Delay

The queueing sojourn time forms only one part of the end-to-end handoff latency. It is therefore combined with the fixed protocol phases to obtain the deadline-relevant total delay used by the risk objective.

The fixed protocol delay D_{fix} aggregates the measurement, preparation, access, execution, and interruption phases (Fig. 7). With delay cap $\kappa D_{\max}$, the total handoff delay is

$$D_i = \begin{cases} \min\{D_{\mathrm{fix}} + T_i,\ \kappa D_{\max}\}, & \rho_i < 1, \\ \kappa D_{\max}, & \rho_i \geq 1. \end{cases} \tag{30}$$

Since T_i in (27) already contains both the service and the waiting time, (30) adds T_i exactly once to the protocol delay and avoids double counting of the service term; unstable queues are assigned the cap.

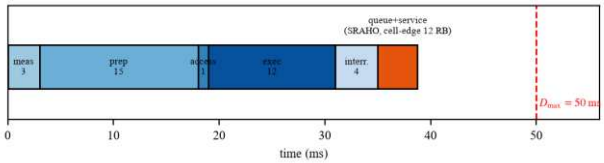


Fig. 7. Composition of the HO delay for SRAHO in the 12-RB cell-edge scenario: five protocol phases totalling 35 ms, plus approximately 3.7 ms of queueing and service, against the 50 ms deadline.

C. Structural Properties of the Delay

The delay expression (30) is now examined as a function of the two quantities that SRAHO controls, namely the RB allocation and the semantic payload. Let $c_i = \lambda_i B_i \mu_1$ denote the stability threshold. A partial-fraction decomposition of (26) yields, on the direct path with $\rho_i < 1$ and inactive cap,

$$D_i(r_i) = D_{\mathrm{fix}} + \frac{B_i \mu_1}{r_i} + \frac{B_i \mu_2}{2\mu_1}\left(\frac{1}{r_i - c_i} - \frac{1}{r_i}\right). \tag{31}$$

Lemma 2 (*Monotonicity and convexity*)**:** For fixed fading samples and other decisions, D_i is nondecreasing in B_i and nonincreasing in r_i (and, on a relay path, in r_j). On the direct path with $r_i > c_i$ and inactive cap,

$$\frac{\partial D_i}{\partial r_i} = -\frac{B_i \mu_1}{r_i^2} - \frac{\lambda_i B_i^2 \mu_2\,(2r_i - c_i)}{2 r_i^2 (r_i - c_i)^2} < 0, \tag{32}$$

$$\frac{\partial^2 D_i}{\partial r_i^2} = \frac{2 B_i \mu_1}{r_i^3} + \frac{B_i \mu_2}{\mu_1}\left(\frac{1}{(r_i - c_i)^3} - \frac{1}{r_i^3}\right) > 0, \tag{33}$$

so D_i is strictly convex and strictly decreasing on (c_i, ∞), with $\partial D_i / \partial r_i \to -\infty$ as $r_i \downarrow c_i$ and $\partial D_i / \partial r_i = O(r_i^{-2})$ as $r_i \to \infty$. On a relay path, T_i^{rel} is jointly convex in (r_i, r_j).

Lemma 2 explains the diminishing returns in Fig. 8 and the convex surface in Fig. 9. Combined with Proposition 3(c), it also shows that $\ell_i^\star$ minimises D_i among all feasible levels, since the delay is nondecreasing in payload. The next result quantifies *how much* semantic compression reduces delay.

Proposition 4 (*Payload scaling law*)**:** Let the payload be scaled as $B_i \mapsto \vartheta B_i$ with $\vartheta \in (0,1]$, $r_i > c_i$, and the cap inactive at B_i. Then the service term scales by ϑ, the stability threshold becomes ϑc_i, and

$$\frac{W_i(\vartheta B_i)}{W_i(B_i)} = \vartheta^2 \frac{r_i - c_i}{r_i - \vartheta c_i} \leq \vartheta^2. \tag{34}$$

Hence $D_i(\vartheta B_i) - D_{\mathrm{fix}} \leq \vartheta (D_i(B_i) - D_{\mathrm{fix}})$, and the waiting-time reduction is at least quadratic in the compression ratio.

For the aggregate compression observed in Section V, $\vartheta = 69\,406/100\,608 \approx 0.690$, so (34) guarantees that the mean

waiting time falls to at most 0.476 of its uncompressed value, strictly more than the 31% payload reduction itself suggests.

Lemma 3 (*Optimality of full power*)**:** Under the fixed-interference model of (6), $\mu_1^{i,\cdot}$, $\mu_2^{i,\cdot}$, and D_i are nonincreasing in P_i, and $\mathbb{P}_{i,\cdot}$ is nondecreasing in P_i. Therefore $P_i = P_{\max}$ is optimal for every vehicle in the problem of Section IV.

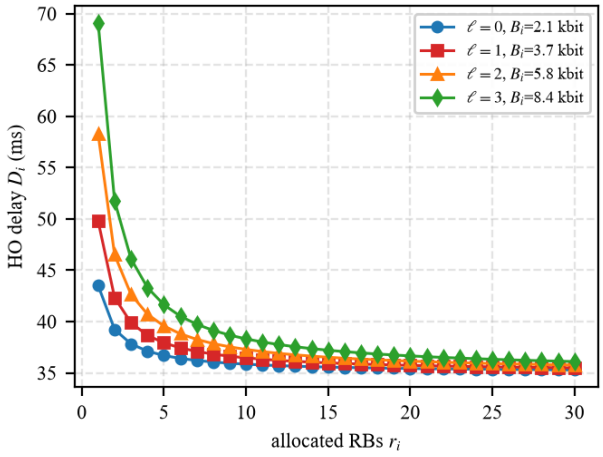


Fig. 8. Handoff delay of a vehicle on an interference-limited direct link (1500 m from the gNB, $I = 2 \times 10^{-12}$ W) versus allocated RBs for the four semantic payloads $B_i(\ell)$. The curves decrease and exhibit the diminishing returns predicted by Lemma 2.

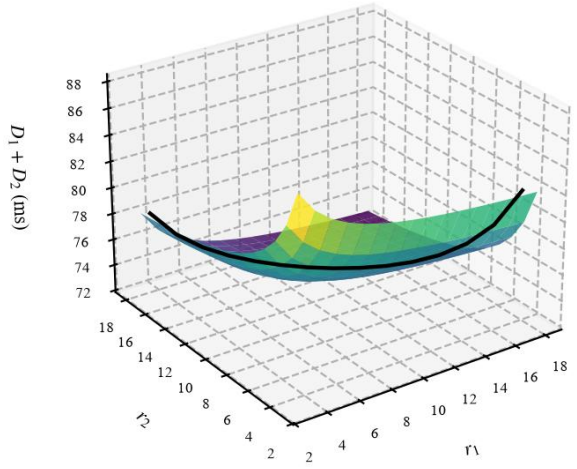


Fig. 9. Sum of the HO delays of two vehicles at 900 m and 1500 m from their gNB versus their RB allocations r_1 and r_2. The black curve is the constraint $r_1 + r_2 = 20$ and the red marker its minimiser, which allocates more RBs to the weaker link, as predicted by (35).

D. Mean-Optimal Allocation: A Square-Root Rule

These structural properties allow the mean-optimal allocation to be characterised in closed form, which serves as a reference for the tail-aware allocation of Section IV. Consider the continuous relaxation $\min \sum_i D_i(r_i)$ subject to $\sum_i r_i = R_{\text{tot}}$ and $r_i > c_i$, with all vehicles on direct paths and caps inactive.

Proposition 5 (*KKT structure*)**:** The relaxed problem has a unique solution if and only if $\sum_i c_i < R_{\text{tot}}$. It is characterised by $\partial D_i / \partial r_i\,(r_i^\star) = -\nu^\star$ for all i, where $\nu^\star > 0$ is the unique root of $\sum_i (-\partial D_i / \partial r_i)^{-1}(\nu) = R_{\text{tot}}$. In the resource-scarce regime in which the waiting term dominates,

$$r_i^\star \simeq c_i + \left(R_{\text{tot}} - \sum_k c_k\right)\frac{\omega_i}{\sum_k \omega_k}, \quad \omega_i = \sqrt{\frac{B_i \mu_2^{i,b}}{\mu_1^{i,b}}}. \tag{35}$$

Rule (35) is the handoff analogue of square-root staffing: every vehicle first receives its stability threshold, and the spare capacity is then shared in proportion to $\sqrt{B_i \mu_2 / \mu_1}$. Because weak links exhibit a larger ratio μ_2/μ_1 (Proposition 2), they receive more spare RBs, which is exactly the behaviour of the constrained minimiser in Fig. 9.

E. Exponential Tail Bound and CVaR of the Waiting Time

The allocation of Section III-D targets the mean delay. The following result links that mean to the tail of the waiting time, which is the quantity SRAHO ultimately controls. Truncation renders the service time bounded, $S_i \le s_i^{\max}$, so its moment generating function $M_{S_i}(\theta) = \mathbb{E}\left[e^{\theta S_i}\right]$ is finite for every θ. Define the Lundberg function

$$\varphi_i(\theta) = \lambda_i\left(M_{S_i}(\theta) - 1\right) - \theta, \quad \theta \ge 0. \tag{36}$$

Theorem 1 (*Closed-form tail and CVaR bound*)**:** If $\rho_i < 1$, then φ_i has a unique positive root $\theta_i^\star$ (the adjustment coefficient), and the stationary waiting time satisfies, for all $x \ge 0$ and $\beta \in (0,1)$,

$$\mathbb{P}(W_i > x) \le e^{-\theta_i^\star x}, \quad \text{CVaR}_\beta(W_i) \le \frac{1 + \ln\frac{1}{1-\beta}}{\theta_i^\star}. \tag{37}$$

Moreover, $\theta_i^\star \ge \underline{\theta}_i = \min\{1/s_i^{\max},\, 2(1-\rho_i)/(e\lambda_i \mathbb{E}[S_i^2])\}$, whence

$$\text{CVaR}_\beta(W_i) \le \left(1 + \ln\frac{1}{1-\beta}\right)\max\{s_i^{\max},\, e\,\mathbb{E}[W_i]\}. \tag{38}$$

Corollary 2 (*Tail of the handoff delay*)**:** Let $\widetilde{D}_i = \min\{D_{\text{fix}} + S_i + W_i, \kappa D_{\max}\}$ be the random handoff delay whose mean-sojourn surrogate is (30). Then

$$\text{CVaR}_\beta\left(\widetilde{D}_i\right) \le \min\{\kappa D_{\max},\, D_{\text{fix}} + \text{CVaR}_\beta(S_i) + \frac{1 + \ln\frac{1}{1-\beta}}{\theta_i^\star}\}. \tag{39}$$

Theorem 1 establishes that, for the bounded-service M/G/1 handoff queue, the tail risk is controlled by the mean waiting time up to the explicit factor $e\left(1 + \ln\frac{1}{1-\beta}\right)$, which equals 10.86 for $\beta = 0.95$. Every mechanism that shrinks $\mathbb{E}[S_i^2]$, notably semantic compression through Proposition 4, therefore also shrinks the *tail*, whereas mechanisms that only shift $\mathbb{E}[S_i]$ at the expense of $\mathbb{E}[S_i^2]$ can inflate it. This provides the analytical rationale for pairing payload adaptation with a CVaR objective.

IV. Semantic Risk-Aware Handoff Orchestration

The delay model of Section III is now embedded in an optimisation problem. This section defines the tail-risk objective, formulates the joint helper, payload, RB, and power

problem, and derives the SRAHO procedure, which exploits the structural results of Sections II and III to avoid solving that problem monolithically.

A. Tail-Risk Objective

Because a handoff deadline is violated by rare long delays rather than by the average, SRAHO measures delay risk by the CVaR, which is defined first. For a delay D and confidence level $\beta \in (0,1)$, the Conditional Value-at-Risk is [16]

$$\mathrm{CVaR}_\beta(D) = \min_{\eta \in \mathbb{R}} \left\{ \eta + \frac{1}{1-\beta}\, \mathbb{E}[(D-\eta)^+] \right\}, \tag{40}$$

whose minimiser is $\eta^\star = \mathrm{VaR}_\beta(D)$ and which admits the Acerbi–Tasche representation [18]

$$\mathrm{CVaR}_\beta(D) = \frac{1}{1-\beta} \int_\beta^1 \mathrm{VaR}_u\,(D)\, du. \tag{41}$$

For samples $D_1, \dots, D_n$ with order statistics $D_{(1)} \le \cdots \le D_{(n)}$ and $k = n - \lceil \beta n \rceil + 1$, the estimator is

$$\hat{C}_\beta = \frac{1}{k} \sum_{j=n-k+1}^{n} D_{(j)} = \max_{\mathcal{S} \subseteq [n],\, |\mathcal{S}|=k} \frac{1}{k} \sum_{j \in \mathcal{S}} D_j. \tag{42}$$

Inside the controller, (42) is applied to the pooled fading-grid delay samples of all vehicles. For reporting, the per-step mean delay and its tail risk over T steps are

$$\bar{D}_t = \frac{1}{N} \sum_{i=1}^{N} D_i^{(t)}, \quad \hat{C}_{0.95} = \mathrm{CVaR}_{0.95}(\bar{D}_1, \dots, \bar{D}_T). \tag{43}$$

By subadditivity and positive homogeneity, $\mathrm{CVaR}_\beta(\bar{D}_t) \le N^{-1} \sum_i \mathrm{CVaR}_\beta\left(D_i^{(t)}\right)$, so the reported metric never exceeds the average per-vehicle tail risk.

Proposition 6 (*Properties of the sampled objective*)**:** (a) $\hat{C}_\beta$ is convex, nondecreasing, and positively homogeneous in the sample vector. (b) If every sample $D_j(\mathbf{r})$ is convex in $\mathbf{r}$, then $\hat{C}_\beta(\mathbf{r})$ is convex, and a subgradient is

$$\mathbf{g}(\mathbf{r}) = \frac{1}{k} \sum_{j \in \mathcal{T}(\mathbf{r})} \nabla_{\mathbf{r}}\, D_j(\mathbf{r}), \tag{44}$$

where $\mathcal{T}(\mathbf{r})$ is any index set of k largest samples. (c) For i.i.d. samples, $\hat{C}_\beta \to \mathrm{CVaR}_\beta(D)$ almost surely; if the distribution function of D is continuous and strictly increasing at $\mathrm{VaR}_\beta(D)$, then

$$\sqrt{n}\left(\hat{C}_\beta - \mathrm{CVaR}_\beta(D)\right) \Rightarrow \mathcal{N}\left(0, \frac{\mathrm{Var}\left[\left(D - \mathrm{VaR}_\beta(D)\right)^+\right]}{(1-\beta)^2}\right). \tag{45}$$

B. Joint Decision Problem

With the delay model of Section III and the tail-risk objective of Section IV-A in place, the decisions of all vehicles can be stated as a single optimisation problem. Each vehicle selects a helper $h_i \in \{-1\} \cup \mathcal{N} \backslash \{i\}$ (-1 denoting direct V2I), a semantic level ℓ_i, an RB count r_i, and a transmit power P_i. The joint problem is

$$\text{(P1)} \quad \min_{\mathbf{h}, \boldsymbol{\ell}, \mathbf{r}, \mathbf{P}} \hat{C}_\beta(D_1, \dots, D_N) \tag{46}$$

subject to

$$\begin{gathered} \sum_{i=1}^{N} r_i \le R_{\mathrm{tot}},\ r_i \in \mathbb{N},\ r_i \ge r_i^{\min}, \\ P_{\min} \le P_i \le P_{\max},\ \Phi_i(\ell_i) \ge \Phi_{\min}, \\ h_i \in \mathcal{J}_i,\ \forall i \in \mathcal{N}. \end{gathered} \tag{47}$$

(P1) is a mixed-integer sampled programme: helper selection, semantic level, and RB allocation are discrete, power is continuous, and the delay cap makes the objective nonconvex. Its RB subspace alone contains $\binom{R_{\mathrm{tot}}-1}{N-1}$ strictly positive allocations, e.g., $\binom{47}{11} \approx 1.74 \times 10^{10}$ for $N = 12$ and $R_{\mathrm{tot}} = 48$, before helper and level choices multiply the search space. Rather than solving (P1) monolithically, SRAHO exploits the structure established above: Proposition 3 and Lemma 2 fix $\ell_i = \ell_i^\star$, Lemma 3 fixes $P_i = P_{\max}$, and the remaining helper and RB decisions are resolved by risk-adjusted screening and CVaR-greedy allocation.

C. SRAHO Decision Procedure

1) Semantic Stage and HOA Candidates

Set $\ell_i = \ell_i^\star$. Under log-normal CSI error, the controller observes $\hat{G}_c = \bar{G}_c \exp(e_c - \sigma_e^2/2)$ with $e_c \sim \mathcal{N}(0, \sigma_e^2)$, which is an unbiased observation, $\mathbb{E}\left[\hat{G}_c\right] = \bar{G}_c$. Each vehicle c is scored by uplink quality, link stability, and queue load:

$$\begin{gathered} s_c = 0.50 \ln\left(1 + \frac{P_{\max} \hat{G}_c}{N_0 + I}\right) + 0.35\, \zeta_c - 0.15\, \rho_c, \\ \zeta_c = \frac{1}{N-1} \sum_{i \ne c} \mathbb{1}\left\{ d_{i,c}^{\mathrm{risk}} \le d_{\mathrm{V2V}} \right\}. \end{gathered} \tag{48}$$

The HOA set $\mathcal{H}$ contains the H highest scores, and the candidate paths of vehicle i are $\mathcal{J}_i = \{-1\} \cup \{j \in \mathcal{H} \backslash \{i\} : \mathcal{A}_{i,j}\}$. By Lemma 1, every relay candidate retains its V2V link over the horizon with probability at least $1 - 2Q\left(\sqrt{2} z_{\mathrm{m}}\right)$.

2) Deterministic Fading Evaluation

Expectations over fading are evaluated on the midpoint quantile grid, with the reversed grid on the second hop of a relay:

$$\xi_m = -\ln\left(1 - \frac{m - 1/2}{M}\right), \quad m = 1, \dots, M. \tag{49}$$

The first grid point $\xi_1 \in \left[\frac{1}{2M}, \frac{1}{2M-1}\right]$ acts as an implicit truncation level; for $M = 128$, $\xi_1 \approx 3.9 \times 10^{-3}$, i.e., fades deeper than about 24 dB are excluded. The grid moments $\mu_k^{(M)} = M^{-1} \sum_m a\,(\xi_m)^k$ obey the following scaling laws.

Proposition 7 (*Grid scaling laws*)**:** Let $\mathrm{S}_M = 2H_{2M} - H_M = \ln M + 2\ln 2 + \gamma_{\mathrm{E}} + O(M^{-1})$. Then

$$c_{\bar{\gamma}}(\mathrm{S}_M - 1) \le \mu_1^{(M)} \le c_{\bar{\gamma}}(\mathrm{S}_M + \bar{\gamma}), \tag{50}$$

$$\lim_{M \to \infty} \frac{\mu_2^{(M)}}{M} = \frac{\pi^2}{2}\, c_{\bar{\gamma}}^2. \tag{51}$$

Moreover, for any integrand $f(u)$ of bounded variation $V(f)$ on $[0,1]$,

$$\left|\frac{1}{M}\sum_{m=1}^{M} f\left(\frac{m-1/2}{M}\right) - \int_0^1 f(u)\,du\right| \le \frac{V(f)}{2M}. \tag{52}$$

Two consequences follow. First, since the capped delay is monotone in the fade and confined to $[D_{\mathrm{fix}}, \kappa D_{\max}]$, its grid average errs by at most $(\kappa D_{\max} - D_{\mathrm{fix}})/(2M)$ by (52) [23]; when $(1-\beta)M$ is an integer, the top-k grid average in (42) is precisely a midpoint quadrature of the Acerbi integral (41). Second, (51) shows that the second service moment, and hence W_i, grows linearly with M: the grid resolution is not a mere numerical parameter but fixes an effective outage cutoff, consistent with Proposition 2. The same deterministic grid is reused for all candidate paths, yielding a repeatable, low-variance evaluation.

Remark 1 (*Coupling of relay hops*)**:** Pairing the grid with its reversal couples the two hop fades countermonotonically. For two summands, countermonotonic coupling is minimal in the convex order [24], and CVaR is consistent with that order; the relay CVaR estimate is therefore a lower approximation of its value under independent hops. The survival penalty in (53) and the helper-queue term of (28) counteract this optimism, and a product grid of M^2 points removes it at quadratic cost.

3) Risk-Adjusted Path Choice

Let $\hat{C}_i(j)$ be (42) evaluated on the M delay samples of vehicle i on path j with $r_i = 1$ and $P_i = P_{\max}$. With helper-departure probability p_{dep},

$$h_i = \operatorname*{argmin}_{j \in \mathcal{J}_i} \frac{\hat{C}_i(j)}{1 - p_{\mathrm{dep}}\,\mathbb{1}\{j \ge 0\}}. \tag{53}$$

Proposition 8 (*Survival-adjusted cost*)**:** Suppose that a relay attempt fails because of helper departure with probability p_{dep}, independently across attempts, that failed attempts are repeated on the same path, and that each attempt incurs a random cost with mean $\bar{c}(j)$ independent of the departure events. Then the expected total cost is $\bar{c}(j)/\left(1 - p_{\mathrm{dep}}\right)$ by Wald's identity. Consequently, (53) replaces $\bar{c}(j)$ by the tail statistic $\hat{C}_i(j)$, and a relay is preferred only if $\hat{C}_i(j) \le \left(1 - p_{\mathrm{dep}}\right)\hat{C}_i(-1)$.

Since $-1 \in \mathcal{J}_i$ carries a unit denominator, the selected path never exceeds the direct path in the surrogate cost.

4) CVaR-Greedy RB Allocation

Let $\hat{C}_{\mathrm{sys}}(\mathbf{r})$ denote (42) on the pooled NM samples. Starting from $\mathbf{r}^{(0)} = \mathbf{1}$ with $\Delta = R_{\mathrm{tot}} - N$,

$$\begin{aligned} i_t^\star &= \operatorname*{argmax}_{i \in \mathcal{N}} \left[\hat{C}_{\mathrm{sys}}\left(\mathbf{r}^{(t)}\right) - \hat{C}_{\mathrm{sys}}\left(\mathbf{r}^{(t)} + \mathbf{e}_i\right)\right], \\ \mathbf{r}^{(t+1)} &= \mathbf{r}^{(t)} + \mathbf{e}_{i_t^\star}, \quad t = 0, \dots, \Delta - 1. \end{aligned} \tag{54}$$

Each update assigns exactly one RB, so the final allocation exhausts the budget and preserves at least one RB per vehicle.

Theorem 2 (*Exactness for separable objectives*)**:** Let $F(\mathbf{r}) = \sum_i f_i(r_i)$, where each $f_i: \mathbb{N} \to \mathbb{R}$ has nonincreasing marginal decrements $\delta_i(t) = f_i(t) - f_i(t+1)$. Then (54) applied to F returns a global minimiser of F over $\{\mathbf{r} \in \mathbb{N}^N : \sum_i r_i = R_{\mathrm{tot}},\ r_i \ge 1\}$ [25], [26].

By Lemma 2, the uncapped mean delay satisfies the hypothesis on $\{r_i > c_i\}$, so the greedy stage is exact for the separable mean-delay objective. For the pooled CVaR, which couples vehicles through the top-k set, (54) is a marginal-descent rule along the subgradient (44); its quality is quantified a posteriori as follows.

Proposition 9 (*Optimality-gap certificate*)**:** Let $\bar{r}_i$ be the smallest allocation for which vehicle i's cap is inactive for all samples, assume $\sum_i \bar{r}_i \le R_{\mathrm{tot}}$, and let $\mathbf{r}^{\mathrm{g}} \ge \bar{\mathbf{r}}$ be the greedy output with subgradient $\mathbf{g}$ from (44) and $g_{\min} = \min_i g_i$. Then, for every feasible $\mathbf{r}$ with $\mathbf{r} \ge \bar{\mathbf{r}}$,

$$\hat{C}_{\mathrm{sys}}(\mathbf{r}^{\mathrm{g}}) - \hat{C}_{\mathrm{sys}}(\mathbf{r}) \le \sum_{i=1}^{N} (g_i - g_{\min})\left(r_i^{\mathrm{g}} - \bar{r}_i\right). \tag{55}$$

The right-hand side of (55) is computable at the cost of one additional objective evaluation and vanishes exactly when all vehicles that received spare RBs share a common marginal tail reduction, i.e., at a Karush–Kuhn–Tucker point of the relaxation.

Remark 2 (*Plateau of capped samples*)**:** When a vehicle is unstable at $r_i = 1$, all of its samples equal the cap, and a single additional RB may leave $\hat{C}_{\mathrm{sys}}$ unchanged until stability is reached. Initialising (54) at $\mathbf{r}^{(0)} = \max\{\mathbf{1}, \mathbf{r}^{\min}\}$ removes this plateau whenever $\sum_i r_i^{\min} \le R_{\mathrm{tot}}$; the results reported here use $\mathbf{r}^{(0)} = \mathbf{1}$.

5) Feasibility Projection

Variants without the greedy stage use urgencies $u_i = \Pi_{[0.05,1]}\left(\max_k \pi_{i,k}\right)$ and the largest-remainder projection

$$\begin{aligned} \tilde{s}_i &= \frac{\Delta\, u_i}{\sum_k u_k}, \quad r_i = 1 + \lfloor \tilde{s}_i \rfloor + \mathbb{1}\{i \in \mathcal{S}\}, \\ P_i &= P_{\min} + (P_{\max} - P_{\min})\frac{u_i}{\max_k u_k}, \end{aligned} \tag{56}$$

where $\mathcal{S}$ contains the $\Delta - \sum_i \lfloor \tilde{s}_i \rfloor$ indices with the largest fractional parts, so that the integer allocation exactly exhausts the budget. The projection enforces $r_i \ge 1$ and $P_i \in [P_{\min}, P_{\max}]$, and the rounding error satisfies $|r_i - 1 - \tilde{s}_i| < 1$ for every i.

6) Computational Complexity

One evaluation of $\hat{C}_{\mathrm{sys}}$ requires $O(NM)$ operations using linear-time selection of the k largest samples. HOA scoring costs $O(N^2)$, path choice $O(NHM)$, and the greedy stage ΔN objective evaluations. Hence

$$\begin{aligned} T_{\mathrm{SRAHO}} &= O\left((R_{\mathrm{tot}} - N)N^2 M + NHM + N^2\right), \\ T_{\mathrm{proj}} &= O(N \log N). \end{aligned} \tag{57}$$

Maintaining the pooled samples in an order-statistic tree reduces each greedy step to $O\left(NM\log(NM)\right)$, since a candidate update replaces only the M samples of one vehicle. Measured execution times of all stages are reported in Section V-F (Table VII).

ALGORITHM 1: SRAHO DECISION AT ONE SLOT

Input: positions, velocities, observed gains $\hat{G}_c$, priorities $\pi_{i,k}$, parameters of Table III.

1: **for all** $i \in \mathcal{N}$ **do**
2: $\ell_i \leftarrow \ell_i^\star$ by (22); compute $B_i(\ell_i)$ by (20)
3: **end for**

4: Score all vehicles by (48); select $\mathcal{H}$; build $\mathcal{J}_i$
5: Build the fading grid (49)
6: **for all** $i \in \mathcal{N}$ **do** $h_i \leftarrow$ (53)
7: $\mathbf{r}^{(0)} \leftarrow \mathbf{1}$; $P_i \leftarrow P_{\max}$ (Lemma 3)
8: **for** $t = 0$ **to** $\Delta - 1$ **do**
9: choose $i_t^\star$ by (54); $\mathbf{r}^{(t+1)} \leftarrow \mathbf{r}^{(t)} + \mathbf{e}_{i_t^\star}$
10: **end for**
Output: h_i, ℓ_i, r_i, P_i for all $i \in \mathcal{N}$

V. Numerical Results and Discussion

This section evaluates SRAHO numerically. After describing the simulation setup, it compares SRAHO with the benchmarks, examines reliability and tail behaviour, isolates the contribution of each component through ablation, studies the sensitivity to the main parameters, and finally validates the allocation stage in a model-based study.

A. Simulation Setup, Metrics, and Benchmarks

All methods share identical scene, event, and channel realisations for each random seed. Packet delivery is a Bernoulli event with the link success probability (12); for a relayed path, the two hop-success probabilities are combined as in (13). A handoff is counted as successful only if delivery occurs, the selected path remains available, the queue is stable, and the handoff delay does not exceed $D_{\max}$:

$$s_i^{\mathrm{succ}} = \mathbb{1}\{\text{delivered} \wedge \text{path available} \wedge \rho_i < 1 \wedge D_i \le D_{\max}\}. \quad (58)$$

Deadline compliance is thus captured by (58), whereas the delay tail is captured by the $\mathrm{CVaR}_{0.95}$ of the per-step mean delay in (43). Unlike the mean or a single percentile, CVaR accounts for the magnitude of every excursion beyond the value-at-risk, and therefore distinguishes an allocation that merely improves the average from one that suppresses rare high-delay events.

The benchmarks are: Direct V2I, with equal RB allocation, full power, and full perception payload; Cluster-head, in which a single head relays packets of in-range members; EGDH-original, which selects three HOAs by uplink quality and centrality in the spirit of [5]; and EGDH-corrected, which uses the stability-aware HOA score of (48) with load-proportional RB allocation. Unless otherwise stated, each reported point averages five seeds of 200 time steps. The parameters are listed in Table III.

Table III: Simulation Parameters

Parameter	Value	Parameter	Value
N / R_{tot}	12 / 48	L	4000 m
B_{RB}	180 kHz	gNB positions	1800, 2200 m
G_0 / α	10^{-3} / 2.5	N_0 (per RB)	7.2×10^{-16} W
P range	0.1–5 W	γ_{th}	5 dB
λ	2 Hz	$D_{\max}$	50 ms
D_{fix}	35 ms	κ	10
M	128	β	0.95
H / τ / σ_a	3 / 1 s / 1 m/s^2	$\Phi_{\min}$	0.72
$b_{\mathrm{hdr}}, b_{\mathrm{idx}}, b_s, d_z$	64, 8, 8, 64	f_ℓ	0.20, 0.40, 0.65, 1.00

Eight scenarios stress the methods: cell-edge geometry (gNBs at 600 and 3400 m) with 12, 24, and 48 RBs; interference of 2×10^{-12} W per RB; bursty arrivals with probability 0.20 and multiplier 12; a dense 24-vehicle case with CSI error $\sigma_e = 0.35$; a narrow-overlap case with $d_{\mathrm{V2V}} = 140$ m, $\tau = 1.8$ s, and $\sigma_a = 2$ m/s^2; and an HOA-departure case with $p_{\mathrm{dep}} = 0.12$.

B. Performance Comparison with Benchmarks

Fig. 10 reports the $\mathrm{CVaR}_{0.95}$ HO delay versus the RB budget in the cell-edge scenarios. SRAHO is lowest at every budget, with reductions of 4.16%, 2.22%, and 1.14% relative to Direct V2I at 12, 24, and 48 RBs. The narrowing gap is a direct manifestation of Lemma 2: the marginal delay reduction decays as $O(r_i^{-2})$, so with abundant RBs the service and waiting terms become small for every method and the fixed 35 ms protocol delay dominates.

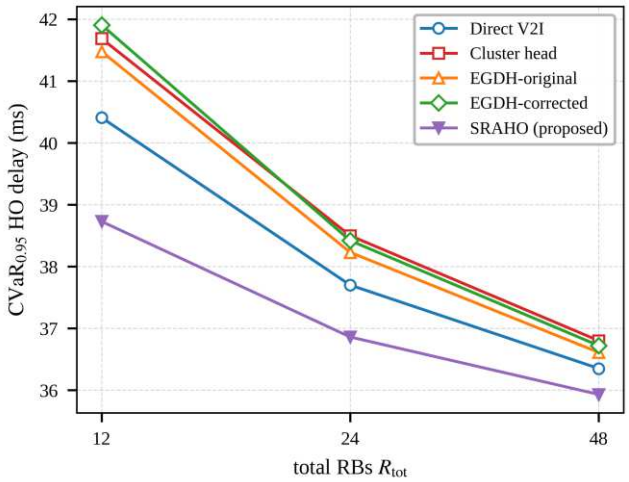


Fig. 10. $\mathrm{CVaR}_{0.95}$ HO delay versus total RBs in the cell-edge scenarios (mean of five seeds).

Fig. 11 and Table IV compare all methods over the eight scenarios. SRAHO attains the lowest $\mathrm{CVaR}_{0.95}$ in every scenario, and paired per-seed differences against Direct V2I have 95% Student-t confidence intervals that exclude zero (for example, 1.679 ± 0.039 ms at 12 RBs and 8.835 ± 3.366 ms under interference). The gain peaks under interference (14.72%), where the relay baselines deteriorate sharply: a relayed packet traverses two interfered hops, which by Corollary 1 erodes the geometric advantage of relaying, and it additionally loads the helper's multiclass queue (28), whose waiting time scales with the aggregate second moment $\sum_i \lambda_i B_i^2$. No run violated the RB or power constraints.

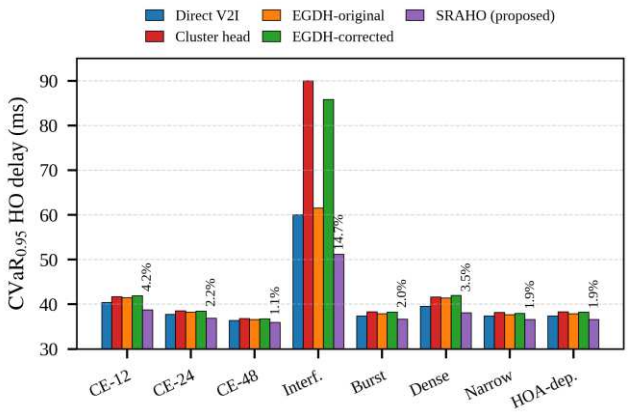


Fig. 11. $CVaR_{0.95}$ HO delay for all methods over the eight scenarios. Numbers above the SRAHO bars give the relative reduction with respect to Direct V2I.

TABLE IV: $CVaR_{0.95}$ HANDOFF DELAY (MS), MEAN OF FIVE SEEDS

Scenario	Direct	Cluster	EGDH-o	EGDH-c	SRAHO	Gain (%)
Cell-edge, 12 RB	40.41	41.69	41.48	41.91	**38.73**	4.16
Cell-edge, 24 RB	37.70	38.50	38.23	38.42	**36.86**	2.22
Cell-edge, 48 RB	36.35	36.80	36.61	36.72	**35.93**	1.14
Interference	60.01	89.95	61.49	85.81	**51.18**	14.72
Burst load	37.36	38.30	37.90	38.24	**36.63**	1.96
Dense noisy	39.52	41.64	41.39	41.95	**38.12**	3.55
Narrow overlap	37.35	38.15	37.65	37.95	**36.62**	1.95
HOA departure	37.35	38.29	37.89	38.23	**36.62**	1.95

C. Reliability and Tail-Delay Behaviour

Fig. 12 shows the empirical CDF of the per-step mean delay under interference. The relay baselines exhibit a heavy right tail, whereas SRAHO shifts the entire distribution to the left and also shortens its upper tail relative to Direct V2I, in line with the mean–tail coupling of Theorem 1. The success probability under interference is 0.983 for SRAHO versus 0.955 for Direct V2I, and under HOA departure it is 1.000 versus 0.972–0.976 for the relay baselines, reflecting the survival penalty of Proposition 8.

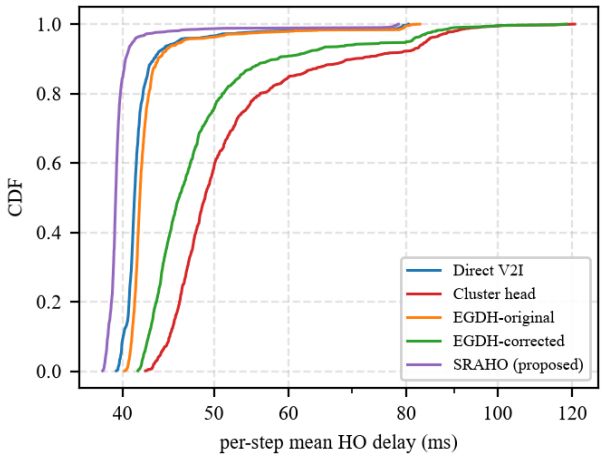


Fig. 12. Empirical CDF of the per-step mean HO delay under interference (all seeds and steps).

D. Ablation Study

Fig. 13 removes one component at a time. Semantic adaptation alone reduces the payload from 100 608 to 69 406 bits per step (31.0%) and the perception proxy from 1.00 to approximately 0.80. In the three non-interference scenarios, semantic-only and full SRAHO differ by at most 0.03 ms, so the gain there is attributable to payload adaptation, as anticipated by Proposition 4. The CVaR-greedy stage contributes principally under interference, where it removes a further 3.1 ms (from 54.27 to 51.18 ms) relative to semantic adaptation alone and 3.9 ms relative to the projection variant. The two components are thus complementary: compression shrinks $\mathbb{E}[S_i^2]$ uniformly, whereas risk-aware allocation redistributes RBs toward the vehicles populating the tail set $\mathcal{T}(\mathbf{r})$.

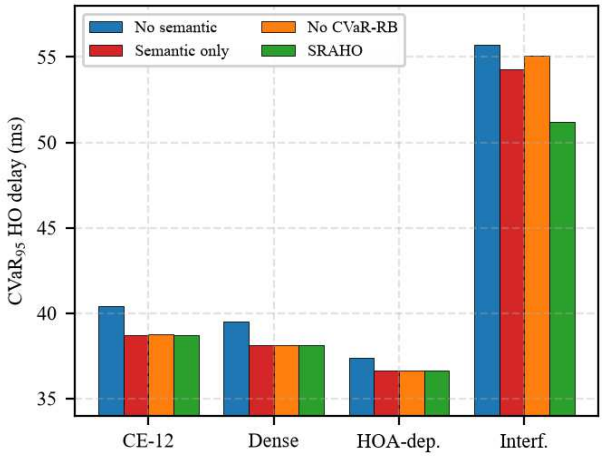


Fig. 13. Ablation of SRAHO: no semantic adaptation, semantic adaptation only, greedy stage removed (proportional projection), and full SRAHO.

E. Sensitivity Analysis

Fig. 14 varies the packet arrival rate in the interference scenario (three seeds of 100 steps). The delay of every method grows with the arrival rate, since the waiting time in (26) scales

as $(1-\rho_i)^{-1}$. SRAHO remains lowest: from 1 to 6 packets/s its $\mathrm{CVaR}_{0.95}$ rises from 51.71 to 56.38 ms, against 58.52 to 63.85 ms for Direct V2I, a gap of 6.81 ms at 1 packet/s and 7.47 ms at 6 packets/s (11.6% and 11.7%). The relay baselines are markedly higher (Cluster-head 75.12 to 87.75 ms, EGDH-corrected 72.80 to 83.77 ms), and their success probabilities remain near 0.80 and 0.86.

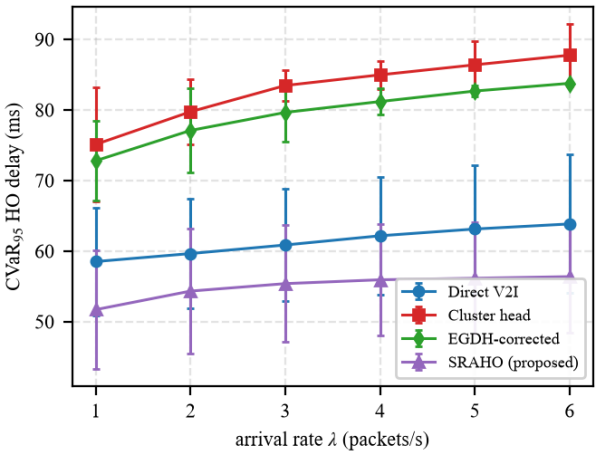


Fig. 14. $\mathrm{CVaR}_{0.95}$ HO delay versus packet arrival rate in the interference scenario (mean ± standard deviation over three seeds of 100 steps).

Fig. 15 varies the number of vehicles with 24 RBs. SRAHO has the lowest $\mathrm{CVaR}_{0.95}$ for every N, with reductions of 7.9%, 8.9%, 20.0%, 12.3%, and 12.3% relative to Direct V2I at $N = 8$, 12, 16, 20, and 24. The curve is not monotone in N because the vehicle set is redrawn for each N and only three seeds are used; we therefore refrain from inferring a trend. The clearer effect concerns reliability: at $N = 24$, where each vehicle receives a single RB and the spare capacity $R_{\mathrm{tot}} - \sum_i c_i$ in (35) is exhausted, the success probability of SRAHO is 0.943, against 0.769 for Direct V2I, 0.446 for Cluster-head, and 0.355 for EGDH-corrected.

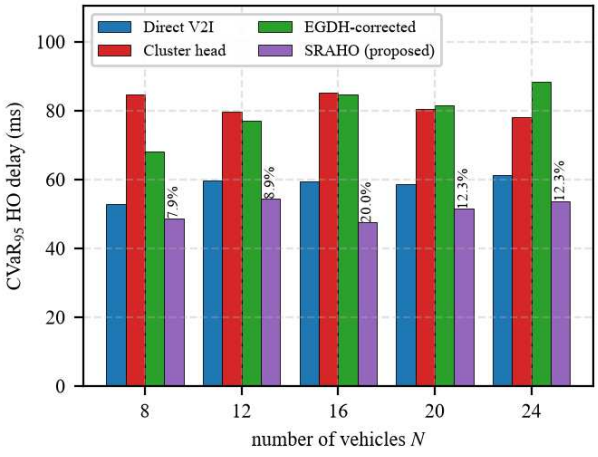


Fig. 15. $\mathrm{CVaR}_{0.95}$ HO delay versus number of vehicles in the interference scenario ($R_{\mathrm{tot}} = 24$).

Fig. 16 sweeps the perception floor. Because the semantic levels are discrete, the achieved perception is 0.573, 0.775, 0.798, 0.999, and 1.000 for floors of 0.50, 0.60, 0.72, 0.85, and 1.00, respectively, an instance of the quantisation effect bounded in (23). The $\mathrm{CVaR}_{0.95}$ delay of SRAHO increases from 48.87 ms at the lowest floor to 58.04 ms at full perception, an increase of 9.17 ms (18.8%) that constitutes the price of retaining more of the scene. SRAHO remains below semantic-only at every floor, by 3.52, 1.94, 1.99, 1.61, and 1.61 ms, which isolates the contribution of helper selection and CVaR-greedy allocation. At full perception, where Direct V2I and semantic-only coincide at 59.65 ms, SRAHO attains 58.04 ms, a reduction of 2.7%.

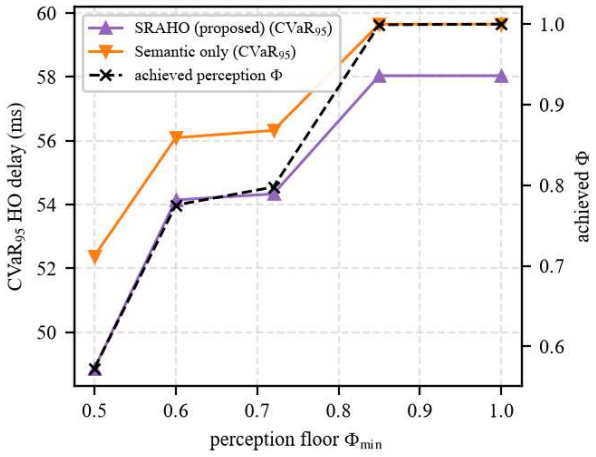


Fig. 16. $\mathrm{CVaR}_{0.95}$ HO delay and achieved perception versus perception floor $\Phi_{\min}$ in the interference scenario.

F. Model-Based Validation of the Allocation Stage

The comparisons above evaluate SRAHO end to end. To isolate the RB-allocation stage from semantic selection and helper choice, and to answer three questions that the end-to-end results do not settle, Algorithm 1 is also evaluated in a reference Python/NumPy implementation of the analytical model of Sections II–IV. Unless stated otherwise, every vehicle uses the direct V2I path, carries the level-2 payload of 5784 bits (the per-vehicle average reported in Section V-D), and contributes the grid delay samples $D_{i,m} = \min\{D_{\mathrm{fix}} + B_i a_i(\xi_m)/r_i + W_i, \kappa D_{\max}\}$, $m = 1, \ldots, M$, with W_i from (26) and all remaining parameters taken from Table III.

1) CVaR-driven versus mean-driven allocation: Table V compares the CVaR-greedy rule (54) with a mean-delay allocator that applies the same marginal rule to the pooled sample mean; since that objective is separable, Theorem 2 guarantees that the mean-delay allocator is exactly optimal for its own objective, so the comparison isolates the choice of objective. Both allocators are computed on the same fading grid, and the resulting allocations are evaluated out of sample on 4000 independent Rayleigh fades per vehicle, conditioned on $h \geq \xi_1$, and averaged over 100 random vehicle drops. In the noise-limited cell-edge scenarios, the two allocations nearly coincide because the delay distribution is too concentrated for the mean and the tail to disagree. Under interference, the CVaR

allocator lowers the 99th-percentile delay from 42.07 to 41.28 ms, the $\mathrm{CVaR}_{0.95}$ from 40.91 to 40.34 ms, and the deadline-violation probability $\mathbb{P}(D_i > D_{\max})$ from 0.27% to 0.19% (a 29% reduction), at the cost of a 0.18-ms (0.5%) increase in mean delay; with 24 vehicles the violation probability falls from 0.90% to 0.78%. CVaR-driven allocation therefore deliberately trades a small increase in the average for a proportionally larger reduction of rare deadline misses, a trade-off that a mean-delay allocator cannot make by construction.

TABLE V: MEAN-DELAY VERSUS $\mathrm{CVaR}_{0.95}$-DRIVEN RB ALLOCATION (DELAYS IN MS, VIOLATION IN %)

Scenario	Alloc.	Mean	P95	P99	$\mathrm{CVaR}_{0.95}$	$\mathbb{P}(D > D_{\max})$
Cell-edge, 24 RB	Mean	35.83	36.09	36.28	36.20	0.00
	CVaR	35.83	36.08	36.26	36.19	0.00
Cell-edge, 48 RB	Mean	35.41	35.53	35.61	35.58	0.00
	CVaR	35.41	35.51	35.58	35.55	0.00
Interf., 48 RB	Mean	36.51	37.91	42.07	40.91	0.27
	CVaR	36.69	37.99	41.28	40.34	0.19
Interf., $N = 24$	Mean	38.07	40.85	49.02	46.78	0.90
	CVaR	38.20	40.93	48.03	45.98	0.78

2) Optimality gap of the greedy allocator: Theorem 2 guarantees exactness only for separable objectives, and Proposition 9 bounds the gap a posteriori. To measure the gap directly, Table VI compares (54) with exhaustive enumeration of every positive integer allocation on 200 random instances per budget, with $N = 4$ vehicles located uniformly between 300 and 1800 m from the gNB, interference $I = 2 \times 10^{-12}$ W, and payloads drawn from the four semantic levels. Over $R_{\mathrm{tot}} \in \{6,8,\dots,16\}$, the greedy allocation attained the exhaustive optimum in 1198 of the 1200 instances, and the relative gap in the remaining two was 0.12% and 0.09%. The number of objective evaluations of (54) grows linearly as $(R_{\mathrm{tot}} - N)(N + 1) + 1$, whereas the number of allocations enumerated by exhaustive search grows combinatorially, reaching 455 against 61 at $R_{\mathrm{tot}} = 16$.

TABLE VI: CVAR-GREEDY ALLOCATION VERSUS EXHAUSTIVE SEARCH ($N = 4$, 200 INSTANCES PER BUDGET)

R_{tot}	Exhaustive allocations	Greedy evaluations	Optimal instances	Max. gap (%)
6	10	11	200/200	0.00
8	35	21	200/200	0.00
10	84	31	199/200	0.12
12	165	41	200/200	0.00
14	286	51	199/200	0.09
16	455	61	200/200	0.00

3) Execution time: Table VII reports the wall-clock time of each stage of Algorithm 1, including relay-path evaluation for the three HOAs, measured on one core of an Intel Xeon processor at 2.1 GHz (Python 3.12, NumPy 2.4) and averaged over 100 decision slots per configuration. Semantic selection and HOA scoring take less than 0.1 ms, path evaluation grows linearly with N and stays below 0.6 ms, and the CVaR-greedy stage accounts for 94–98% of the total, consistent with (57). For the default configuration of Table III ($N = 12$, $R_{\mathrm{tot}} = 48$), a complete decision takes 8.85 ms on average (95th percentile 9.07 ms). The largest configuration ($N = 24$, $R_{\mathrm{tot}} = 96$) requires 41.81 ms on average (95th percentile 44.84 ms), which remains below the 50-ms handoff deadline even in this interpreted implementation; the order-statistic update described after (57) or a compiled implementation would reduce these times further.

TABLE VII: EXECUTION TIME PER DECISION SLOT (MS), MEAN OF 100 SLOTS

N	R_{tot}	Sem.	HOA	Path	Greedy	Total	Total P95
4	48	0.03	0.04	0.11	3.06	3.24	3.44
4	96	0.04	0.04	0.11	6.36	6.54	6.93
8	48	0.05	0.05	0.21	5.93	6.23	6.64
8	96	0.05	0.04	0.22	12.54	12.85	13.34
12	48	0.04	0.04	0.29	8.47	8.85	9.07
12	96	0.05	0.05	0.30	19.60	20.00	20.76
16	48	0.05	0.05	0.39	10.70	11.19	11.74
16	96	0.05	0.05	0.39	26.52	27.02	28.16
20	48	0.05	0.05	0.46	12.68	13.24	13.75
20	96	0.06	0.05	0.48	32.97	33.56	34.40
24	48	0.05	0.05	0.52	13.97	14.60	15.25
24	96	0.07	0.06	0.57	41.10	41.81	44.84

G. Limitations and Practical Scope

The present evidence is simulation-based. The communication results rely on the analytical Rayleigh-fading and M/G/1 models of Sections II and III rather than a full 3GPP system-level channel implementation [27] or measured mobility traces. The semantic score is a retained-priority proxy rather than a detection metric; the reported payload reduction, obtained at a retained priority mass near 0.80, should therefore not be read as preserved object-detection accuracy. The relay queue relies on a store-and-forward independence approximation, and the reversed fading grid yields an optimistic lower approximation of the two-hop CVaR (Remark 1). The model-based validation of Section V-F covers direct paths only and relies on a reference implementation of the analytical model rather than on the full simulator; the greedy optimality gap was measured for four vehicles, and execution times were measured on a general-purpose processor rather than on in-vehicle or edge hardware.

VI. CONCLUSION AND FUTURE WORK

This paper developed SRAHO, a semantic and risk-aware handoff-orchestration framework for cooperative perception in highway V2X networks. The framework treats semantic payload size, direct-versus-HOA path selection, and RB allocation as coupled decisions. The analytical development links mobility persistence, Rayleigh-fading service statistics, M/G/1 queueing, and a CVaR handoff-delay objective. Semantic selection chooses the minimum payload satisfying a retained-priority floor; mobility-aware screening excludes short-lived V2V helpers; and marginal CVaR reduction guides the discrete RB allocation.

In the synthetic evaluation, SRAHO achieves the lowest $\mathrm{CVaR}_{0.95}$ delay among all evaluated methods in all eight scenarios. The largest reduction relative to Direct V2I occurs under interference, where the $\mathrm{CVaR}_{0.95}$ decreases from 60.01 ms to 51.18 ms (14.72%) and the handoff success probability rises from 0.955 to 0.983. The ablation study shows that semantic adaptation reduces the transmitted payload by 31.0%, while the CVaR-greedy stage delivers additional benefit in the interference-limited regime, confirming that semantic payload control and tail-risk-aware allocation are complementary rather than redundant.

Future work will evaluate SRAHO with 3GPP-compliant V2X channel models [27], measured highway mobility traces, and an end-to-end cooperative-perception detector on datasets such as OPV2V [3], so that semantic retention can be linked to task metrics such as mean average precision [4], [9]. The model-based validation of Section V-F will also be extended to relay paths, larger networks, and the full simulator, and the controller will be profiled on embedded in-vehicle hardware. Further directions include replacing the grid-induced truncation with the physically motivated outage cutoff $\varepsilon = \gamma_{\mathrm{th}}/\bar{\gamma}$, stability-aware greedy initialisation (Remark 2), interference-coupled power control formulated as a game, and hardware-in-the-loop emulation to validate controller run time and signalling overhead.

APPENDIX: PROOFS

Proof of Lemma 1

Since $a_u - a_v \sim \mathcal{N}(0, 2\sigma_a^2)$, $\Delta_{u,v}(\tau) \sim \mathcal{N}\big(m_{u,v}(\tau), \varsigma_\tau^2\big)$ with $\varsigma_\tau = 1/2\,\tau^2\sqrt{2}\sigma_a = \tau^2\sigma_a/\sqrt{2}$, which yields (3). For (a), $|\Delta_{u,v}(\tau)| > d_{\mathrm{V2V}}$ implies $|\Delta_{u,v}(\tau) - m_{u,v}(\tau)| > d_{\mathrm{V2V}} - |m_{u,v}(\tau)| \geq z_{\mathrm{m}}\tau^2\sigma_a = \sqrt{2}z_{\mathrm{m}}\varsigma_\tau$ under $\mathcal{A}_{u,v}$, so the probability is at most $\mathbb{P}\big(|Z| > \sqrt{2}z_{\mathrm{m}}\big) = 2Q\big(\sqrt{2}z_{\mathrm{m}}\big)$ with $Z \sim \mathcal{N}(0,1)$. For (b), $\sup_{t\leq\tau}|\Delta_{u,v}(t) - m_{u,v}(t)| = 1/2\,|a_u - a_v|\tau^2$ is controlled by a single Gaussian variable, while $|m_{u,v}(t)|$ is the modulus of an affine function and is maximised at $t \in \{0, \tau\}$; the argument of (a) then applies uniformly in t.

Proof of Proposition 1

For a nonnegative variable, $\mathbb{E}[X] = \int_0^\infty \mathbb{P}\,(X > t)dt$. With $X = \ln(1 + \bar{\gamma}h)$, $\mathbb{P}(X > t) = \exp(-(e^t - 1)/\bar{\gamma})$, and the substitution $s = e^t$ gives

$$\mathbb{E}[X] = \int_1^\infty \frac{e^{-(s-1)/\bar{\gamma}}}{s}\,ds = e^{1/\bar{\gamma}}E_1(1/\bar{\gamma}), \tag{59}$$

which is (8) after division by ln2. Likewise, $\mathbb{E}[X^2] = \int_0^\infty 2\,t\,\mathbb{P}(X > t)dt$, and the same substitution yields (9). For $\bar{\gamma} \to \infty$, the expansion $E_1(x) = -\gamma_{\mathrm{E}} - \ln x + x + O(x^2)$ [28] with $x = 1/\bar{\gamma}$ and $e^x = 1 + O(x)$ give the first line of (10). For $\bar{\gamma} \to 0$, $\ln(1 + y) = y - y^2/2 + y^3/3 - \cdots$ and $\mathbb{E}[h^n] = n!$ give the second line. Finally, $X = \ln\bar{\gamma} + \ln h + \ln\big(1 + 1/(\bar{\gamma}h)\big)$; the last term vanishes pointwise and is dominated by $\ln(1 + 1/h)$, which is square integrable, so $\mathrm{Var}[X] \to \mathrm{Var}[\ln h] = \pi^2/6$, the variance of the Gumbel-type law of $\ln h$.

Proof of (12) and Corollary 1

Conditioning on the interference, $\mathbb{P}\big(X \geq \gamma_{\mathrm{th}}\big(N_0 + \sum_j Y_j\big) \mid \mathbf{Y}\big) = \exp\big(-\gamma_{\mathrm{th}}\big(N_0 + \sum_j Y_j\big)/\bar{X}\big)$. Averaging with the Laplace transform $\mathbb{E}\big[e^{-sY_j}\big] = \big(1 + s\Lambda_j\big)^{-1}$ and independence [29] yields (12). For (14), $e^{-x} \geq 1 - x$ and $(1 + y)^{-1} \geq 1 - y$ together with $1 - \prod_k(1 - c_k) \leq \sum_k c_k$ for $c_k \in [0,1]$ give the bound; the two-hop union bound follows identically. For (b), with common deterministic interference, $\mathbb{P}_{u,v} = \exp\big(-\kappa_0 d_{u,v}^\alpha/P_u\big)$ with $\kappa_0 = \gamma_{\mathrm{th}}(N_0 + I)/G_0$, so $\mathbb{P}_{i,j}\mathbb{P}_{j,b} > \mathbb{P}_{i,b}$ is equivalent to (15).

Proof of Proposition 2

Since $\ln(1 + x) \leq x$, $a \geq c_{\bar{\gamma}}/h$ and $\mathbb{E}[1/h] = \int_0^\infty h^{-1}\,e^{-h}dh = \infty$. Since $\ln(1 + x) \geq x/(1 + x)$, $a \leq c_{\bar{\gamma}}(1/h + \bar{\gamma})$. Integrating both bounds against e^{-h} over $[\varepsilon, \infty)$ gives (16). Squaring and using

$$\int_\varepsilon^\infty \frac{e^{-h}}{h^2}\,dh = \frac{e^{-\varepsilon}}{\varepsilon} - E_1(\varepsilon) = \Psi(\varepsilon), \tag{60}$$

obtained by integration by parts, gives (17). The asymptotics (18) follow from $E_1(\varepsilon) = \ln\varepsilon^{-1} - \gamma_{\mathrm{E}} + O(\varepsilon)$ and $\Psi(\varepsilon) = \varepsilon^{-1} + O(\ln\varepsilon^{-1})$.

Proof of Proposition 3

(a) All regions have identical bit cost c_{reg}, so a subset of cardinality K costs $B_{\mathrm{hdr}} + Kc_{\mathrm{reg}}$ bits, and the maximal mass achievable with K regions is $\sum_{k\leq K}\pi_i^{(k)}$ by the rearrangement argument. The minimal feasible cardinality is therefore $K_i^\star$. (b) The increment $\Phi_i(K + 1) - \Phi_i(K) = \pi_i^{(K+1)}/\Pi_i^{\mathrm{tot}}$ is nonnegative and nonincreasing in K by the ordering. (c) Level ℓ is feasible if and only if $K_i(\ell) \geq K_i^\star$. Infeasibility of $\ell_i^\star - 1$ gives $K_i(\ell_i^\star - 1) \leq K_i^\star - 1$, hence $K_i(\ell_i^\star) - K_i^\star \leq K_i(\ell_i^\star) - K_i(\ell_i^\star - 1) - 1$. Since $B_i(\ell)$ is increasing in ℓ, the smallest feasible level minimises the payload.

Proof of Lemma 2

The identity $\frac{1}{r(r-c)} = \frac{1}{c}\left(\frac{1}{r-c} - \frac{1}{r}\right)$ and $\lambda_i B_i^2\mu_2/(2c_i) = B_i\mu_2/(2\mu_1)$ give (31). Differentiating,

$$\frac{\partial D_i}{\partial r_i} = -\frac{B_i\mu_1}{r_i^2} - \frac{B_i\mu_2}{2\mu_1}\left(\frac{1}{(r_i - c_i)^2} - \frac{1}{r_i^2}\right), \tag{61}$$

and $\frac{1}{(r-c)^2} - \frac{1}{r^2} = \frac{c(2r-c)}{r^2(r-c)^2}$ recovers (32); both terms are negative because $0 < r_i - c_i < r_i$. Differentiating once more gives (33), positive for the same reason. The limits follow directly from (61). Monotonicity in B_i holds because each term of (26)–(27) increases with B_i, and $\min\{\cdot, \kappa D_{\max}\}$ preserves monotonicity. On a relay path, T_i^{rel} is a sum of a convex function of r_i and a convex function of r_j, hence jointly convex.

Proof of Proposition 4

With $c_i = \lambda_i B_i\mu_1$, (26) gives $W_i(B_i) = \lambda_i B_i^2\mu_2/\big(2r_i(r_i - c_i)\big)$ and $W_i(\vartheta B_i) = \lambda_i\vartheta^2 B_i^2\mu_2/\big(2r_i(r_i - \vartheta c_i)\big)$; the ratio is (34), which is at most ϑ^2 because $r_i - \vartheta c_i \geq r_i - c_i$. The service term is linear in B_i. Hence $D_i(\vartheta B_i) - D_{\mathrm{fix}} \leq \vartheta\mathbb{E}[S_i] + \vartheta^2 W_i \leq \vartheta(D_i(B_i) - D_{\mathrm{fix}})$ and applying the monotone map $\min\{\cdot, \kappa D_{\max}\}$ to the left-hand side can only decrease it.

Proof of Lemma 3

By (6), $\bar{\gamma}_{i,\cdot}$ is increasing in P_i when I is fixed, and $a = 1/\big(B_{\mathrm{RB}}\log_2(1 + \bar{\gamma}h)\big)$ is pointwise decreasing in $\bar{\gamma}$ for every $h > 0$. Hence every grid moment $\mu_k^{i,\cdot}$ decreases in P_i, and so do ρ_i, W_i, T_i, and D_i by Lemma 2; $\mathbb{P}_{i,\cdot}$ increases because $\bar{X}_{i,\cdot}$ increases.

Proof of Proposition 5

The objective is a sum of strictly convex functions (Lemma 2) over a convex set, so any KKT point is the unique global minimiser. Stationarity of the Lagrangian $\sum_i D_i(r_i) + \nu(\sum_i r_i - R_{\mathrm{tot}})$ gives $-\partial D_i/\partial r_i = \nu$. Each map $-\partial D_i/\partial r_i$ is continuous and strictly decreasing from $+\infty$ to 0 on (c_i, ∞), so its inverse $r_i(\nu)$ is well defined, continuous, and strictly decreasing, with $r_i(\nu) \to c_i$ as $\nu \to \infty$ and $r_i(\nu) \to \infty$ as $\nu \to 0$. The function $\nu \mapsto \sum_i r_i(\nu)$ therefore attains R_{tot} at a unique $\nu^\star$ if and only if $\sum_i c_i < R_{\mathrm{tot}}$. When the waiting term dominates, $-\partial D_i/\partial r_i \simeq B_i\mu_2/(2\mu_1(r_i - c_i)^2)$, so $r_i - c_i \simeq \omega_i/\sqrt{2\nu}$; enforcing the budget eliminates ν and gives (35).

Proof of Theorem 1

The function φ_i is convex, $\varphi_i(0) = 0$, $\varphi_i'(0) = \rho_i - 1 < 0$, and $\varphi_i(\theta) \to \infty$ because $S_i > 0$ almost surely; hence a unique positive root exists. The M/G/1 waiting time coincides in law with the ruin probability of the associated Cramér–Lundberg process, and Lundberg's inequality [21] yields the tail bound. Set $\eta_0 = \ln\big(1/(1-\beta)\big)/\theta_i^\star$ in (40):

$$\begin{aligned}\mathrm{CVaR}_\beta(W_i) &\le \eta_0 + \frac{1}{1-\beta}\int_{\eta_0}^{\infty} \mathbb{P}(W_i > x)\,dx \\ &\le \eta_0 + \frac{e^{-\theta_i^\star \eta_0}}{(1-\beta)\theta_i^\star} = \frac{1 + \ln\frac{1}{1-\beta}}{\theta_i^\star}.\end{aligned} \tag{62}$$

For the lower bound on $\theta_i^\star$, $e^x \le 1 + x + x^2/2\, e^x$ for $x \ge 0$ and $S_i \le s_i^{\max}$ give

$$\varphi_i(\theta) \le \theta\big(\rho_i - 1 + 1/2\,\lambda_i\theta\mathbb{E}[S_i^2]e^{\theta s_i^{\max}}\big). \tag{63}$$

At $\theta = \underline{\theta}_i$ one has $e^{\underline{\theta}_i s_i^{\max}} \le e$ and $1/2\,\lambda_i\underline{\theta}_i\mathbb{E}[S_i^2]e \le 1 - \rho_i$, so $\varphi_i\big(\underline{\theta}_i\big) \le 0$. Since φ_i is convex with $\varphi_i(0) = 0$ and $\varphi_i < 0$ exactly on $(0, \theta_i^\star)$, $\underline{\theta}_i \le \theta_i^\star$. Substituting $1/\underline{\theta}_i = \max\{s_i^{\max}, e\mathbb{E}[W_i]\}$, with $\mathbb{E}[W_i] = \lambda_i\mathbb{E}[S_i^2]/\big(2(1-\rho_i)\big)$, gives (38). Corollary 2 follows from the definition of $\widetilde{D}_i$, translation equivariance, subadditivity, and monotonicity of CVaR [15].

Proof of Proposition 6

(a) By (42), $\hat{C}_\beta$ is a pointwise maximum of linear functions with nonnegative coefficients, hence convex, nondecreasing, and positively homogeneous. (b) The composition of a convex nondecreasing function with convex functions is convex [30]; (44) is the subgradient of the active linear piece. (c) Almost-sure convergence follows from the uniform strong law applied to the sample-average approximation of (40) [31], since the top-k estimator differs from it by $O(1/n)$. Under the stated regularity the minimiser $\eta^\star = \mathrm{VaR}_\beta(D)$ is unique, and the central limit theorem for sample-average optimal values [31] gives the asymptotic variance $\mathrm{Var}[\eta^\star + (D - \eta^\star)^+/(1-\beta)]$, which equals the expression in (45).

Proof of Proposition 7

Using $u \le -\ln(1-u) \le u/(1-u)$ with $u_m = (m - 1/2)/M$ and the bounds $c_{\bar\gamma}/\xi \le a(\xi) \le c_{\bar\gamma}(1/\xi + \bar\gamma)$ from Proposition 2,

$$\sum_{m=1}^{M}\frac{1}{m-\frac{1}{2}} - 1 \le \frac{1}{M}\sum_{m=1}^{M}\frac{1}{\xi_m} \le \sum_{m=1}^{M}\frac{1}{m-\frac{1}{2}}, \tag{64}$$

and $\sum_{m=1}^{M}(m - 1/2)^{-1} = 2H_{2M} - H_M = \mathrm{S}_M$, which gives (50). For (51), $\mu_2^{(M)}/M = M^{-2}\sum_m a(\xi_m)^2$; for fixed m, $M\xi_m \to m - 1/2$ and $a(\xi_m) \sim c_{\bar\gamma}/\xi_m$, so each summand tends to $c_{\bar\gamma}^2(m-1/2)^{-2}$, while $M^{-2}a(\xi_m)^2 \le 2c_{\bar\gamma}^2((m-1/2)^{-2} + \bar\gamma^2 M^{-2})$ supplies a summable dominating sequence plus a remainder of total order M^{-1}. By Tannery's theorem the limit is $c_{\bar\gamma}^2\sum_{m\ge1}(m-1/2)^{-2} = c_{\bar\gamma}^2\pi^2/2$. Finally, (52) is the one-dimensional Koksma–Hlawka inequality with the star discrepancy $1/(2M)$ of the centred grid [23].

Proof of Proposition 8

The number of attempts $\mathcal{K}$ is geometric with success probability $1 - p_{\mathrm{dep}}$, so $\mathbb{E}[\mathcal{K}] = 1/\big(1 - p_{\mathrm{dep}}\big)$. The costs are i.i.d. with mean $\bar{c}(j)$ and $\{\mathcal{K} \ge n\}$ is independent of the nth cost, so Wald's identity gives $\mathbb{E}[\sum_{n\le\mathcal{K}} c_n] = \bar{c}(j)/\big(1 - p_{\mathrm{dep}}\big)$. The comparison with the direct path, whose denominator is one, is immediate.

Proof of Theorem 2

Write $F(\mathbf{r}) = \sum_i f_i(1) - \sum_i\sum_{t=1}^{r_i-1}\delta_i(t)$. Minimising F is equivalent to selecting $\Delta = R_{\mathrm{tot}} - N$ terms from the N sequences $\{\delta_i(t)\}_{t\ge1}$ so as to maximise their sum, subject to the selections from each sequence forming a prefix. Because each sequence is nonincreasing, the Δ largest terms overall can always be chosen to form prefixes, and (54) selects exactly these terms, one per iteration, with ties broken arbitrarily. Hence the greedy allocation attains the maximum and is globally optimal [25].

Proof of Proposition 9

On $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^N : \sum_i x_i = R_{\mathrm{tot}},\ \mathbf{x} \ge \bar{\mathbf{r}}\}$ all caps are inactive, every sample is convex (Lemma 2), and $\hat{C}_{\mathrm{sys}}$ is convex (Proposition 6). The subgradient inequality gives $\hat{C}_{\mathrm{sys}}(\mathbf{r}) \ge \hat{C}_{\mathrm{sys}}(\mathbf{r}^{\mathrm{g}}) + \mathbf{g}^\top(\mathbf{r} - \mathbf{r}^{\mathrm{g}})$ for all $\mathbf{r} \in \mathcal{X}$, hence

$$\hat{C}_{\mathrm{sys}}(\mathbf{r}^{\mathrm{g}}) - \hat{C}_{\mathrm{sys}}(\mathbf{r}) \le \mathbf{g}^\top\mathbf{r}^{\mathrm{g}} - \min_{\mathbf{x}\in\mathcal{X}}\mathbf{g}^\top\mathbf{x}. \tag{65}$$

The linear minimum over the shifted simplex $\mathcal{X}$ is attained by placing all slack on a coordinate with the smallest g_i: $\min_{\mathbf{x}\in\mathcal{X}}\mathbf{g}^\top\mathbf{x} = \mathbf{g}^\top\bar{\mathbf{r}} + (R_{\mathrm{tot}} - \sum_i \bar{r}_i)g_{\min}$. Substituting and using $\sum_i\big(r_i^{\mathrm{g}} - \bar{r}_i\big) = R_{\mathrm{tot}} - \sum_i \bar{r}_i$ yields (55).

REFERENCES

[1] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3D point clouds," in *Proc. IEEE Int. Conf. Distrib. Comput. Syst. (ICDCS)*, 2019.

[2] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2VNet: Vehicle-to-vehicle communication for joint perception and prediction," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020.

[3] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2022.

[4] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2022.

[5] S. Sharma, S. Ghosh, M. R. Bhatnagar, and B. K. Panigrahi, "Evolutionary game-based handoff agent selection: A novel delay-aware group handoff mechanism in vehicular networks," *IEEE Trans. Commun.*, vol. 74, p. 7010, 2026, doi: 10.1109/TCOMM.2026.3681660.

[6] S. Sharma, S. Ghosh, M. R. Bhatnagar, and B. K. Panigrahi, "Power efficient handoff management in hybrid V2X communication: Game-theoretic approach to resource allocation with load reduction," unpublished manuscript.

[7] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How often should one update?" in *Proc. IEEE INFOCOM*, 2012, pp. 2731–2739.

[8] R. D. Yates, Y. Sun, D. R. Brown, S. K. Kaul, E. Modiano, and S. Ulukus, "Age of information: An introduction and survey," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 5, pp. 1183–1210, May 2021.

[9] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022.

[10] D. Gündüz, Z. Qin, I. E. Aguerri, H. S. Dhillon, Z. Yang, A. Yener, K. K. Wong, and C.-B. Chae, "Beyond transmitting bits: Context, semantics, and task-oriented communications," *IEEE J. Sel. Areas Commun.*, vol. 41, no. 1, pp. 5–41, Jan. 2023.

[11] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[12] C.-F. Liu and M. Bennis, "Ultra-reliable and low-latency vehicular transmission: An extreme value theory approach," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1292–1295, Jun. 2018.

[13] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

[14] H. Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021.

[15] P. Artzner, F. Delbaen, J.-M. Eber, and D. Heath, "Coherent measures of risk," *Math. Finance*, vol. 9, no. 3, pp. 203–228, 1999.

[16] R. T. Rockafellar and S. Uryasev, "Optimization of conditional value-at-risk," *J. Risk*, vol. 2, no. 3, pp. 21–41, 2000.

[17] R. T. Rockafellar and S. Uryasev, "Conditional value-at-risk for general loss distributions," *J. Banking Finance*, vol. 26, no. 7, pp. 1443–1471, 2002.

[18] C. Acerbi and D. Tasche, "On the coherence of expected shortfall," *J. Banking Finance*, vol. 26, no. 7, pp. 1487–1503, 2002.

[19] L. Kleinrock, *Queueing Systems, Volume I: Theory*. New York, NY, USA: Wiley, 1975.

[20] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[21] S. Asmussen, *Applied Probability and Queues*, 2nd ed. New York, NY, USA: Springer, 2003.

[22] M. H. C. Garcia, A. Molina-Galan, M. Boban, J. Gozalvez, B. Coll-Perales, T. Şahin, and A. Kousaridas, "A tutorial on 5G NR V2X communications," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 3, pp. 1972–2026, 3rd Quart. 2021.

[23] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*. Philadelphia, PA, USA: SIAM, 1992.

[24] L. Rüschendorf, *Mathematical Risk Analysis: Dependence, Risk Bounds, Optimal Allocations and Portfolios*. Berlin, Germany: Springer, 2013.

[25] B. L. Fox, "Discrete optimization via marginal analysis," *Manage. Sci.*, vol. 13, no. 3, pp. 210–216, 1966.

[26] T. Ibaraki and N. Katoh, *Resource Allocation Problems: Algorithmic Approaches*. Cambridge, MA, USA: MIT Press, 1988.

[27] 3GPP, "Study on evaluation methodology of new vehicle-to-everything (V2X) use cases for LTE and NR," 3rd Generation Partnership Project, Tech. Rep. TR 37.885, Release 15.

[28] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*. New York, NY, USA: Dover, 1972.

[29] M. Haenggi, *Stochastic Geometry for Wireless Networks*. Cambridge, U.K.: Cambridge Univ. Press, 2012.

[30] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[31] A. Shapiro, D. Dentcheva, and A. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*. Philadelphia, PA, USA: SIAM, 2009.

[32] R. M. Sohaib, O. Onireti, K. Tan, Y. Sambo, R. Swash, and M. Imran, "Meta-transfer learning-based handover optimization for V2N communication," IEEE Trans. Veh. Technol., vol. 73, no. 11, pp. 17331-17346, Nov. 2024, doi: 10.1109/TVT.2024.3431875.

[33] J. Zhang, K. Yang, H. Wang, P. Sun, and L. Song, "Efficient vehicular collaborative perception based on saptial-temporal feature compression," IEEE Trans. Veh. Technol., vol. 73, no. 11, pp. 16125-16133, Nov. 2024, doi: 10.1109/TVT.2024.3403263.

[34] Y. Zhang, Z. Fan, J. Hou, N. Chen, F. Lyu, and P. Yang, "Incentivizing point cloud-based accurate cooperative perception for connected vehicles," IEEE Trans. Veh. Technol., vol. 74, no. 4, pp. 5637-5648, Apr. 2025, doi: 10.1109/TVT.2024.3519626.

[35] B. Lu, X. Huang, Y. Wu, L. P. Qian, et al., "Joint optimization of compression, transmission and computation for cooperative perception aided intelligent vehicular networks," IEEE Trans. Veh. Technol., 2025, doi: 10.1109/TVT.2025.3528026.

[36] G. Zhang, Q. Hu, Z. Qin, Y. Cai, G. Yu, and X. Tao, "A unified multi-task semantic communication system for multimodal data," IEEE Trans. Commun., vol. 72, no. 7, pp. 4101-4116, Jul. 2024, doi: 10.1109/TCOMM.2024.3364990.

[37] L. Xia, Y. Sun, D. Niyato, D. Feng, L. Feng, and M. A. Imran, "xURLLC-aware service provisioning in vehicular networks: A semantic communication perspective," IEEE Trans. Wireless Commun., vol. 23, no. 5, pp. 4475-4488, May 2024, doi: 10.1109/TWC.2023.3319442.

[38] H. Gao, G. Yu, Y. He, and Y. Liu, "Semantic feature scheduling and rate control in multi-modal distributed network," IEEE Trans. Wireless Commun., vol. 23, no. 12, pp. 19199-19214, Dec. 2024, doi: 10.1109/TWC.2024.3480352.

[39] S. Zhang, S. Zhang, Y. Mao, L. K. Yeung, B. Clerckx, and T. Q. S. Quek, "Transformer-based channel prediction for rate-splitting multiple access-enabled vehicle-to-everything communication," IEEE Trans. Wireless Commun., vol. 23, no. 10, pp. 12717-12730, Oct. 2024, doi: 10.1109/TWC.2024.3395670.

[40] D. Kim, M. R. Castellanos, and R. W. Heath, Jr., "Joint relay selection and beam management based on deep reinforcement learning for millimeter wave vehicular communication," IEEE Trans. Veh. Technol., vol. 72, no. 10, pp. 13067–13080, Oct. 2023.

[41] K.-L. Chen, W.-Y. Chen, and R.-H. Hwang, "Optimizing resource allocation with high-reliability constraint for multicasting automotive messages in 5G NR C-V2X networks," IEEE Trans. Veh. Technol., vol. 72, no. 4, pp. 4792–4804, Apr. 2023.