Joint Beamforming and Resource Allocation for Latency Minimization in Vehicular ISCC Systems: A Two-Timescale Approach

Kai Dong*, Lei Wang*, Sergiy A. Vorobyov*, Yanghai Yu†, Tarik Taleb‡, and Zhu Han§
*Department of Information and Communications Engineering, Aalto University, Espoo, Finland.
†National Space Science Center, Chinese Academy of Sciences, Beijing, China.
‡Faculty of Electrical Engineering and Information Technology, Ruhr University Bochum, Bochum, Germany.
§Department of Electrical and Computer Engineering, University of Houston, Houston, USA.

***Abstract*—Vehicular integrated sensing, communication, and computation (ISCC) systems demand low-latency joint optimization of beamforming and resource allocation. Conventional single-timescale approaches optimize beamforming and resource allocation simultaneously, incurring prohibitive computational complexity for real-time vehicular systems. The heterogeneous channel dynamics in high-mobility environments offer a natural two-timescale structure: beamforming needs to be frequently updated to track fast-varying small-scale fading for maintaining directional gain, while resource allocation decisions depend on slowly evolving large-scale statistics. Inspired by this intrinsic structure, we propose a two-timescale framework for minimizing the maximum completion latency, jointly optimizing beamforming, resource allocation, and computational task offloading subject to sensing quality constraints. Specifically, at the fast timescale, an alternating optimization updates the communication/sensing beamformer and receive combiner. At the slow timescale, logarithmic substitution and successive convex approximation (SCA) handle non-convex fractional constraints based on a grid-based statistical channel state information dictionary. Simulations show that the proposed scheme closely approaches the latency performance of the single-timescale benchmark while significantly outperforming other baselines, with reduced resource-allocation update overhead that enables practical deployment on resource-constrained scenarios.**

***Index Terms*—Integrated sensing, communication, and computation, two-timescale optimization, latency minimization, resource allocation, beamforming.**

I. Introduction

The integration of integrated sensing and communication (ISAC) with multi-access edge computing (MEC) facilitates a fundamental paradigm shift toward integrated sensing, communication, and computation (ISCC) supporting for latency-critical vehicular applications such as cooperative perception and autonomous navigation [1]. By exploiting the intrinsic coupling among sensing, communication, and computation over a shared spectrum and hardware platform, ISCC enables more efficient system design than traditional decoupled approaches [2]. In vehicular ISCC systems, the overall performance is governed by the joint design of beamforming and resource allocation, where these three functionalities inherently compete for the same limited system budgets.

Latency-aware ISCC system design has been attracting increasing attention. The authors of [3] proposed a two-tier binary offloading framework for latency-aware resource allocation in cell-free massive multiple-input and multiple-output (MIMO) systems to balance distributed and centralized computation. The work in [4] minimized the total sensing task execution latency by jointly optimizing beamforming and binary offloading variables subject to power and computational capacity constraints. In [1], beamforming and resource allocation were jointly optimized to minimize the maximum sensing task completion latency in an ISCC-aided vehicular network, assuming a fixed offloading strategy for raw sensing data. In [5], a multi-agent reinforcement learning approach was proposed to minimize the long-term average service latency of vehicular user equipments (VUEs) via joint binary offloading, power, bandwidth, and computational resource optimization. Also, a three-tier inference framework was proposed in [6] to jointly optimize deep neural network partitioning, ISAC beamforming, and computational resource allocation for minimizing sensing task inference latency across all ISAC devices.

However, the above works uniformly adopt a single-timescale paradigm where beamforming and resource allocation variables are jointly updated at every instantaneous channel state information (CSI) acquisition interval. In high-mobility vehicular environments, the wireless channel exhibits heterogeneous dynamics: small-scale fading fluctuates at the millisecond level [7], while large-scale components (e.g., path loss and shadowing) vary much more slowly [8]. Forcing all variables to be recomputed at each coherence interval imposes unnecessary signaling and computational overhead, since resource allocation decisions remain near-optimal over tens to hundreds of milliseconds [9]. Although two-timescale frameworks have been explored in metasurface-aided systems [10] and cooperative ISAC networks [11], these works neither

The work of Tarik Taleb is supported in part by the Federal Ministry of Research, Technology, and Space (BMFTR), Germany, through the Project 6GEM+ under Grant 16KIS2411; and in part by the European Union's Horizon Europe research and innovation programme under the 6G-Path project (Grant No. 101139172).

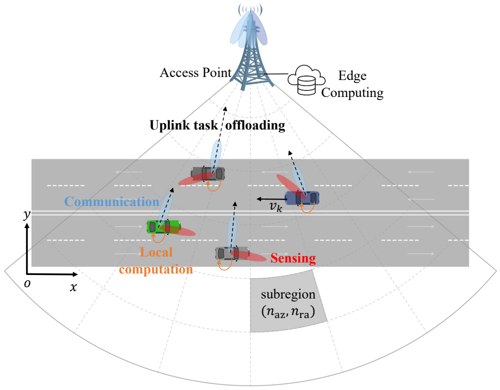


Fig. 1: AP-assisted Vehicular ISCC system.

account for sensing link quality constraints nor address the coupling between task offloading and computational resource allocation under vehicular mobility in ISCC systems.

To address these limitations, we propose a two-timescale approach for vehicular ISCC latency minimization. Specifically, the main contributions of this paper are: *(i)* We formulate a latency minimization problem for computational task in vehicular ISCC systems. Inspired by the heterogeneous channel dynamics in timescale, the optimization problem is then decoupled into fast-timescale beamforming based on instantaneous CSI and slow-timescale resource allocation based on a statistical CSI dictionary constructed offline; *(ii)* A hierarchical double-loop algorithm is then developed to handle the non-convexity of the formulated problem. The fast-timescale inner loop employs alternating optimization (AO) framework for beamforming updates, while slow-timescale outer loop applies successive convex approximation (SCA) approach convert non-convex fractional constraints into tractable convex subproblems; and *(iii)* Simulation results demonstrate our proposed two-timescale scheme closely approaches the latency performance of the single-timescale benchmark with significantly reduced resource-allocation update frequency and an order-of-magnitude reduction in computational complexity, while significantly outperforming all the other baselines.

The rest of this paper is organized as follows. Section II presents the system model and the formulation of the latency minimization problem. Section III develops the proposed two-timescale algorithm. Simulation results and conclusions are provided in Sections IV and V, respectively.

II. System Model and Problem Formulation

As illustrated in Fig. 1, we consider a roadside access point (AP)-assisted vehicular ISCC system, where K VUEs move within the AP's coverage sector defined by azimuth range $[-\pi/3, \pi/3]$ and maximum radius r_{ap}. Each VUE $k \in \mathcal{K} = \{1, \cdots, K\}$ moves along the x-axis with a desired velocity governed by the intelligent driver model (IDM) [12], which adjusts the velocity based on relative velocity and spacing to the preceding vehicle. Each VUE k executes a computational task of Q_k bits (e.g., multi-modal sensing data) while simultaneously performing a sensing task towards azimuth ψ_k, with adaptive task offloading adopted to accelerate data processing. The AP is equipped with a uniform linear array (ULA) of N_{a} antenna elements, and each VUE has two ULAs of $N_{\mathrm{u}}^{\mathrm{t}}$ and $N_{\mathrm{u}}^{\mathrm{r}}$ antenna elements for signal transmission and reception, respectively.

A. Signal Model

The transmitted communication-sensing signal $\mathbf{x}_k$ at VUE k is[1]

$$\mathbf{x}_k = \sqrt{p_k^{\mathrm{c}}}\mathbf{f}_k s_k^{\mathrm{c}} + \sqrt{p_k^{\mathrm{s}}}\mathbf{u}_k s_k^{\mathrm{s}}, \tag{1}$$

where s_k^{c} and s_k^{s} are the zero-mean unit-power data and sensing symbols, assumed mutually uncorrelated; $\mathbf{f}_k \in \mathbb{C}^{N_{\mathrm{u}}^{\mathrm{t}}\times 1}$ and $\mathbf{u}_k \in \mathbb{C}^{N_{\mathrm{u}}^{\mathrm{t}}\times 1}$ are the unit-norm communication and sensing beamforming vectors, i.e., $\|\mathbf{f}_k\| = \|\mathbf{u}_k\| = 1$; p_k^{c} and p_k^{s} are the allocated communication and sensing power, subject to a maximum transmit power budget $P_{\max}$, i.e., $p_k^{\mathrm{c}} + p_k^{\mathrm{s}} \leq P_{\max}$.

Under orthogonal frequency-division multiple access transmissions, the received signal at the AP from VUE k is[2]

$$\begin{aligned} y_k^{\mathrm{ap}} &= \mathbf{w}_k^H \mathbf{H}_k \mathbf{x}_k + \mathbf{w}_k^H \mathbf{n}_k \\ &= \underbrace{\sqrt{p_k^{\mathrm{c}}}\mathbf{w}_k^H \mathbf{H}_k \mathbf{f}_k s_k^{\mathrm{c}}}_{\text{Communication signal}} + \underbrace{\sqrt{p_k^{\mathrm{s}}}\mathbf{w}_k^H \mathbf{H}_k \mathbf{u}_k s_k^{\mathrm{s}}}_{\text{Interference from sensing signal}} + \underbrace{\mathbf{w}_k^H \mathbf{n}_k}_{\text{Noise}}, \end{aligned} \tag{2}$$

where $\mathbf{H}_k \in \mathbb{C}^{N_{\mathrm{a}}\times N_{\mathrm{u}}^{\mathrm{t}}}$ is the VUE-AP MIMO channel, $\mathbf{w}_k \in \mathbb{C}^{N_{\mathrm{a}}\times 1}$ is the AP receive combiner, and $\mathbf{n}_k \sim \mathcal{CN}(\mathbf{0}, \sigma_{\mathrm{n}}^2 \mathbf{I}_{N_{\mathrm{a}}})$ with σ_{n}^2 being the noise variance. The communication signal-to-interference-plus-noise ratio (SINR) is then given by

$$\gamma_k^{\mathrm{c}} = \frac{p_k^{\mathrm{c}} |\mathbf{w}_k^H \mathbf{H}_k \mathbf{f}_k|^2}{p_k^{\mathrm{s}} |\mathbf{w}_k^H \mathbf{H}_k \mathbf{u}_k|^2 + \sigma_{\mathrm{n}}^2 \|\mathbf{w}_k\|^2}, \tag{3}$$

and the achievable rate is expressed as

$$R_k = \alpha_k B \log_2(1 + \gamma_k^{\mathrm{c}}), \tag{4}$$

where B is the total bandwidth and $\alpha_k \in (0, 1]$ is the bandwidth fraction allocated to VUE k with $\sum_k \alpha_k = 1$. The sensing beampattern gain towards ψ_k is given by

$$g(\psi_k) = \mathbf{a}^H(\psi_k)(p_k^{\mathrm{c}} \mathbf{f}_k \mathbf{f}_k^H + p_k^{\mathrm{s}} \mathbf{u}_k \mathbf{u}_k^H)\mathbf{a}(\psi_k), \tag{5}$$

where $\mathbf{a}(\psi_k)$ is the ULA transmit steering vector.

B. Latency Model

Under a partial offloading policy, the computational task of VUE k is split into two parallel segments: a fraction $\beta_k \in [0, 1]$ of the Q_k bits is offloaded to the AP for edge computing, while the remaining $(1 - \beta_k)Q_k$ bits are processed locally. The local computation latency at VUE k is expressed as

$$t_k^{\mathrm{loc}} = \frac{(1 - \beta_k) C Q_k}{f_k^{\mathrm{v}}}, \tag{6}$$

where C is the number of CPU cycles per bit, and f_k^{v} is the local CPU frequency of VUE k, subject to the maximum

[1]The time index is omitted throughout the paper for the brevity.

[2]Sensing signal reflections at the AP are assumed negligible due to severe two-way propagation loss at new radio (NR) frequency bands [13].

available computational capacity F_k^{v}, i.e., $0 \leq f_k^{\mathrm{v}} \leq F_k^{\mathrm{v}}$. The offloading latency comprises transmission and edge processing delays, which can be expressed as

$$t_k^{\mathrm{off}} = \frac{\beta_k Q_k}{R_k} + \frac{\beta_k C Q_k}{f_k^{\mathrm{ap}}}, \tag{7}$$

where f_k^{ap} is the CPU frequency allocated to VUE k at the AP, subject to $\sum_k f_k^{\mathrm{ap}} \leq F^{\mathrm{ap}}$. Taking the parallel execution of local computing and task offloading into account, the overall completion latency for the k-th VUE is defined as the bottleneck between the two processes, which is formulated as[3]

$$T_k = \max\{t_k^{\mathrm{loc}},\, t_k^{\mathrm{off}}\}. \tag{8}$$

C. Optimization Problem Formulation

In vehicular ISCC systems, the stringent latency requirements of delay-critical applications necessitate efficient utilization of the limited communication and computational resources. To guarantee fairness among all VUEs, the minmax task execution latency is adopted as the optimization objective. Defining $\mathcal{B}_k = \{\alpha_k, \beta_k, p_k^{\mathrm{c}}, p_k^{\mathrm{s}}, f_k^{\mathrm{ap}}, f_k^{\mathrm{u}}, \mathbf{f}_k, \mathbf{u}_k, \mathbf{w}_k\}$, the optimization problem is formulated as

$$\begin{aligned}
\min_{\mathcal{B}_k} \quad & \max\ T_k && \text{(9a)}\\
\text{s.t.} \quad & g(\psi_k) \geq \Gamma^{\mathrm{s}}, \forall k, && \text{(9b)}\\
& 0 \leq f_k^{\mathrm{u}} \leq F_k^{\mathrm{v}},\ \forall k, && \text{(9c)}\\
& \sum_{k=1}^{K} f_k^{\mathrm{ap}} \leq F^{\mathrm{ap}},\ f_k^{\mathrm{ap}} \geq 0,\ \forall k, && \text{(9d)}\\
& \sum_{k=1}^{K} \alpha_k = 1,\ 0 \leq \alpha_k \leq 1\ \forall k, && \text{(9e)}\\
& 0 \leq \beta_k \leq 1,\ \forall k, && \text{(9f)}\\
& p_k^{\mathrm{c}} + p_k^{\mathrm{s}} \leq P_{\max}, \forall k, && \text{(9g)}\\
& \|\mathbf{f}_k\| = 1, \|\mathbf{u}_k\| = 1, \|\mathbf{w}_k\| = 1,\ \forall k, && \text{(9h)}
\end{aligned}$$

where (9b) ensures the minimum sensing beampattern gain Γ^{s} required for reliable target detection; (9c)-(9d) impose local and AP computational capacity limits; (9e)-(9f) constrain bandwidth and offloading ratios; (9g) enforces the total transmit power budget at each VUE; and (9h) restricts beamforming vectors to unit-norm for power scaling.

By introducing an auxiliary variable t for handling the non-smooth objective, the above optimization problem is equivalently reformulated as

$$\begin{aligned}
\min_{\mathcal{B}_k, t} \quad & t && \text{(10a)}\\
\text{s.t.} \quad & t_k^{\mathrm{loc}} \leq t,\ \forall k, && \text{(10b)}\\
& t_k^{\mathrm{off}} \leq t,\ \forall k, && \text{(10c)}\\
& \text{(9b)} - \text{(9h)}, && \text{(10d)}
\end{aligned}$$

This reformulated problem remains non-convex due to: *(i)* the bilinear coupling of β_k and f_k^{ap} in (7); *(ii)* the fractional SINR structure in (3) making the achievable data rate R_k non-concave; and *(iii)* the non-convex unit-norm constraints in (9h). A two-timescale optimization framework is proposed in the next section to tackle these challenges.

[3]The latency associated with CSI estimation and signaling exchange between the AP and VUEs is omitted here for the brevity.

III. Proposed Two-Timescale Solution

Problem (10) involves variables operating at distinct timescales. Specifically, the beamforming variables $\mathcal{B}_k^{\mathrm{fast}} = \{\mathbf{f}_k, \mathbf{u}_k, \mathbf{w}_k\}$ need to adapt to instantaneous CSI at the fast timescale Δt_{fast} (millisecond-level), while the resource allocation variables $\mathcal{B}_k^{\mathrm{slow}} = \{\alpha_k, \beta_k, f_k^{\mathrm{ap}}, f_k^{\mathrm{v}}, p_k^{\mathrm{c}}, p_k^{\mathrm{s}}\}$ are governed by slowly varying large-scale statistics at the slow timescale Δt_{slow}. Exploiting this separation, problem (10) can be decoupled into two subproblems solved via a hierarchical AO framework, as detailed below. To capture location-dependent channel statistics, the AP's coverage area is partitioned into a two-dimentional (2D) grid: the azimuth range $[-\pi/3, \pi/3]$ is divided into N_{az} sectors and the radial coverage $(0, r_{\mathrm{ap}}]$ into N_{ra} range cells. For each subregion $(n_{\mathrm{az}}, n_{\mathrm{ra}})$, the instantaneous uplink channel is modeled as a Rician fading model [11]. The statistical channel for each subregion is obtained via Monte Carlo (MC) simulations as $\tilde{\mathbf{H}}_{n_{\mathrm{az}}, n_{\mathrm{ra}}} = \mathbb{E}[\mathbf{H}_{n_{\mathrm{az}}, n_{\mathrm{ra}}}]$, and stored in the dictionary $\boldsymbol{\mathcal{H}}_{\mathrm{sta}}$ with $[\boldsymbol{\mathcal{H}}_{\mathrm{sta}}]_{n_{\mathrm{az}}, n_{\mathrm{ra}}} = \tilde{\mathbf{H}}_{n_{\mathrm{az}}, n_{\mathrm{ra}}}$.

A. Beamforming Update in Fast-Timescale

With $\mathcal{B}_k^{\mathrm{slow}}$ fixed at each fast timescale interval, the fast-timescale subproblem optimizes $\mathcal{B}_k^{\mathrm{fast}} = \{\mathbf{f}_k, \mathbf{u}_k, \mathbf{w}_k\}$ and auxiliary variable t:

$$\begin{aligned}
\min_{\mathcal{B}_k^{\mathrm{fast}}, t} \quad & t && \text{(11a)}\\
\text{s.t.} \quad & \text{(9b)}, \text{(9g)} - \text{(9h)}, \text{(10b)} - \text{(10c)}. && \text{(11b)}
\end{aligned}$$

This optimization problem remains non-convex due to the coupling among beamforming vectors, and can be solved via an AO framework as detailed below.

1) Update of Transmit Beamforming $\mathbf{f}_k$*:* The offloading latency constraint (10c) is first converted into an equivalent SINR requirement $\gamma_k^{\mathrm{c}} \geq 2^{\frac{\beta_k Q_k}{\alpha_k B \Delta t_k}} - 1$, where $\Delta t_k = t - \frac{\beta_k C Q_k}{f_k^{\mathrm{ap}}}$. With fixed $\{p_k^{\mathrm{c}}, p_k^{\mathrm{s}}, \mathbf{w}_k, \mathbf{u}_k, t\}$, $\mathbf{f}_k$ is updated by solving

$$\begin{aligned}
\max_{\mathbf{f}_k} \quad & |\mathbf{w}_k^H \mathbf{H}_k \mathbf{f}_k|^2 && \text{(12a)}\\
\text{s.t.} \quad & \|\mathbf{f}_k\| = 1,\ p_k^{\mathrm{c}} |\mathbf{a}^H(\psi_k) \mathbf{f}_k|^2 \geq \tilde{\Gamma}_k^{\mathrm{s}}, \forall k, && \text{(12b)}
\end{aligned}$$

where $\tilde{\Gamma}_k^{\mathrm{s}} \triangleq \Gamma^{\mathrm{s}} - p_k^{\mathrm{s}} |\mathbf{a}^H(\psi_k) \mathbf{u}_k|^2$. By defining $\mathbf{F}_k \triangleq \mathbf{f}_k \mathbf{f}_k^H$ and applying semidefinite relaxation (SDR) [14], the above problem (12) is relaxed to

$$\begin{aligned}
\max_{\mathbf{F}_k} \ & \mathrm{Tr}(\mathbf{H}_k^{\mathrm{c}} \mathbf{F}_k) && \text{(13a)}\\
\text{s.t.} \ & \mathrm{Tr}(\mathbf{F}_k) = 1,\ \mathbf{F}_k \succeq \mathbf{0},\ \mathrm{Tr}(\mathbf{A}_k \mathbf{F}_k) \geq \tilde{\Gamma}_k^{\mathrm{s}} / p_k^{\mathrm{c}}, \forall k, && \text{(13b)}
\end{aligned}$$

where $\mathbf{H}_k^{\mathrm{c}} = \mathbf{H}_k^H \mathbf{w}_k \mathbf{w}_k^H \mathbf{H}_k$ and $\mathbf{A}_k = \mathbf{a}(\psi_k) \mathbf{a}^H(\psi_k)$. The relaxed problem is convex and can be solved efficiently. Since (13a) admits a rank-one optimal solution [14], $\mathbf{f}_k^\star$ is recovered via eigenvalue decomposition (EVD) as $\mathbf{f}_k^\star = \sqrt{\lambda_{\mathbf{F}_k^\star}} \mathbf{v}_{\mathbf{F}_k^\star}$, where $\lambda_{\mathbf{F}_k^\star}$ and $\mathbf{v}_{\mathbf{F}_k^\star}$ denote the dominant eigenvalue and its corresponding eigenvector of $\mathbf{F}_k^\star$, respectively.

2) Update of Sensing Beamforming $\mathbf{u}_k$: Given fixed $\{\mathbf{f}_k^\star, \mathbf{w}_k, p_k^{\mathrm{c}}, p_k^{\mathrm{s}}, t\}$, $\mathbf{u}_k$ is optimized to minimize interference leakage while satisfying the sensing constraint. By defining $\mathbf{U}_k \triangleq \mathbf{u}_k\mathbf{u}_k^H$ and applying the same SDR procedure, the relaxed subproblem is

$$\min_{\mathbf{U}_k} \ \mathrm{Tr}(\mathbf{H}_k^{\mathrm{c}}\mathbf{U}_k) \tag{14a}$$

$$\text{s.t. } \mathrm{Tr}(\mathbf{U}_k) = 1, \ \mathbf{U}_k \succeq \mathbf{0}, \ \mathrm{Tr}(\mathbf{A}_k\mathbf{U}_k) \geq \bar{\Gamma}_k^{\mathrm{s}}/p_k^{\mathrm{s}}, \forall k, \tag{14b}$$

where $\bar{\Gamma}_k^{\mathrm{s}} \triangleq \Gamma^{\mathrm{s}} - p_k^{\mathrm{c}}|\mathbf{a}^H(\psi_k)\mathbf{f}_k^\star|^2$. The sensing beamformer $\mathbf{u}_k^\star$ is recovered from $\mathbf{U}_k^\star$ via EVD following the same procedure.

3) Update of Receive Combiner $\mathbf{w}_k$: With fixed $\mathbf{f}_k^\star$ and $\mathbf{u}_k^\star$, the optimal MMSE receive combiner is [15]: $\mathbf{w}_k^* = \frac{\mathbf{R}_k^{-1}\mathbf{H}_k\mathbf{f}_k^\star}{\|\mathbf{R}_k^{-1}\mathbf{H}_k\mathbf{f}_k^\star\|}$, where $\mathbf{R}_k \triangleq p_k^{\mathrm{s}}\mathbf{H}_k\mathbf{u}_k^\star\mathbf{u}_k^{*H}\mathbf{H}_k^H + \sigma_{\mathrm{n}}^2\mathbf{I}_{N_a}$.

B. Slow-Timescale Resource Allocation Update

Given the VUE k located at position $(\tilde{r}_k, \tilde{\psi}_k)$, the subregion index $(\tilde{n}_{\mathrm{az}}, \tilde{n}_{\mathrm{ra}})$ is determined by

$$\begin{cases} \tilde{n}_{\mathrm{az}} = \min\left\{ \left\lfloor \frac{\tilde{\psi}_k + \frac{\pi}{3}}{\frac{2\pi}{3N_{\mathrm{az}}}} \right\rfloor + 1, N_{\mathrm{az}} \right\}, \forall \tilde{n}_{\mathrm{az}} \in \{1, \cdots, N_{\mathrm{az}}\}, \\ \tilde{n}_{\mathrm{ra}} = \min\left\{ \left\lfloor \frac{\tilde{r}_k}{r_{\mathrm{ap}}/N_{\mathrm{ra}}} \right\rfloor + 1, \ N_{\mathrm{ra}} \right\}, \ \forall \tilde{n}_{\mathrm{ra}} = \{1, \cdots, N_{\mathrm{ra}}\}. \end{cases}$$

The statistical CSI is then retrieved as $\bar{\mathbf{H}}_k = [\boldsymbol{\mathcal{H}}_{\mathrm{sta}}]_{\tilde{n}_{\mathrm{az}}, \tilde{n}_{\mathrm{ra}}}$. The reference sensing beamformer $\bar{\mathbf{u}}_k = \mathbf{a}(\psi_k)/\|\mathbf{a}(\psi_k)\|$ is adopted to guarantee the sensing beampattern requirement without instantaneous CSI. Invoking the fast-timescale steps with $\bar{\mathbf{H}}_k$ and $\bar{\mathbf{u}}_k$ yields $\bar{\mathbf{f}}_k, \bar{\mathbf{w}}_k$, and the statistical channel gains are expressed as

$$\begin{cases} a_k = |\bar{\mathbf{w}}_k^H\bar{\mathbf{H}}_k\bar{\mathbf{f}}_k|^2, \ b_k = |\bar{\mathbf{w}}_k^H\bar{\mathbf{H}}_k\bar{\mathbf{u}}_k|^2, \\ \bar{a}_k = |\mathbf{a}^H(\psi_k)\bar{\mathbf{f}}_k|^2, \bar{b}_k = |\mathbf{a}^H(\psi_k)\bar{\mathbf{u}}_k|^2. \end{cases} \tag{15}$$

The statistical SINR and achievable rate under full bandwidth allocation for the k-th VUE are then given by

$$\bar{\gamma}_k(p_k^{\mathrm{c}}, p_k^{\mathrm{s}}) = \frac{p_k^{\mathrm{c}}a_k}{p_k^{\mathrm{s}}b_k + \sigma_{\mathrm{n}}^2}, \tag{16}$$

$$\bar{R}_k(p_k^{\mathrm{c}}, p_k^{\mathrm{s}}) = B\log_2(1 + \bar{\gamma}_k(p_k^{\mathrm{c}}, p_k^{\mathrm{s}})). \tag{17}$$

By incorporating the bandwidth allocation factor α_k, the effective achievable data rate is expressed as $\tilde{R}_k = \alpha_k\bar{R}_k(p_k^{\mathrm{c}}, p_k^{\mathrm{s}})$. The slow-timescale resource allocation subproblem is then expressed as

$$\min_{\mathcal{B}_k^{\mathrm{slow}}, t} \quad t \tag{18a}$$

$$\text{s.t.} \quad \frac{\beta_k Q_k}{\alpha_k\bar{R}_k(p_k^{\mathrm{c}}, p_k^{\mathrm{s}})} + \frac{\beta_k C Q_k}{f_k^{\mathrm{ap}}} \leq t, \forall k, \tag{18b}$$

$$p_k^{\mathrm{c}}\bar{a}_k + p_k^{\mathrm{s}}\bar{b}_k \geq \Gamma^{\mathrm{s}}, \text{(9c)} - \text{(9g)}, \text{(10b)}, \forall k. \tag{18c}$$

This optimization problem is non-convex due to the fractional coupling of β_k/α_k, $\beta_k/f_k^{\mathrm{ap}}$, and the multiplicative coupling between $\bar{R}_k(p_k^{\mathrm{c}}, p_k^{\mathrm{s}})$ and β_k. To handle this, logarithmic substitutions are introduced as $\tilde{\alpha}_k = \ln\alpha_k$, $\tilde{\beta}_k = \ln\beta_k$, $\tilde{f}_k^{\mathrm{ap}} = \ln f_k^{\mathrm{ap}}$, $\forall k$, which convert the fractional constraints into convex exponential forms.[4] The offloading latency constraint (10c) is further decomposed via auxiliary variables $t_{k,1}, t_{k,2}$ as

$$t_{k,1} + t_{k,2} \leq t, \forall k, \tag{19}$$

$$\frac{Q_k\exp(\tilde{\beta}_k - \tilde{\alpha}_k)}{\bar{R}_k(p_k^{\mathrm{c}}, p_k^{\mathrm{s}})} \leq t_{k,1}, \forall k, \tag{20}$$

$$CQ_k\exp(\tilde{\beta}_k - \tilde{f}_k^{\mathrm{ap}}) \leq t_{k,2}, \forall k. \tag{21}$$

Constraint (21) is convex, since the exponential function is convex and its argument $\tilde{\beta}_k - \tilde{f}_k^{\mathrm{ap}}$ is affine. However, (20) remains non-convex due to the coupling of $p_k^{\mathrm{c}}, p_k^{\mathrm{s}}$ in $\bar{R}_k(p_k^{\mathrm{c}}, p_k^{\mathrm{s}})$. Taking the logarithm on both sides of (20) yields

$$\ln Q_k + \tilde{\beta}_k - \tilde{\alpha}_k - \ln\bar{R}_k(p_k^{\mathrm{c}}, p_k^{\mathrm{s}}) \leq \ln t_{k,1}. \tag{22}$$

To handle the non-convex term in (22), $\ln\bar{R}_k(p_k^{\mathrm{c}}, p_k^{\mathrm{s}})$ is locally approximated by its first-order Taylor expansion around $(p_k^{\mathrm{c}(l)}, p_k^{\mathrm{s}(l)})$ as (23) (shown at the tope of the next page), where $\bar{R}_k^{(l)} = \bar{R}_k(p_k^{\mathrm{c}(l)}, p_k^{\mathrm{s}(l)})$ is achieved by substituting $p_k^{\mathrm{c}(l)}$ and $p_k^{\mathrm{s}(l)}$ into (17), and corresponding partial derivatives with respect to p_k^{c} and p_k^{s} around the point $(p_k^{\mathrm{c}(l)}, p_k^{\mathrm{s}(l)})$ at the l-th iteration are expressed in (24) at the top of the next page.

Using the local approximation in (23), the transmission latency constraint in (22) is approximated as

$$Q_k\exp\left(\tilde{\beta}_k - \tilde{\alpha}_k - \Psi_k^{(l)}(p_k^{\mathrm{c}}, p_k^{\mathrm{s}})\right) \leq t_{k,1}. \tag{25}$$

Based on the logarithmic transformations and the SCA-based linearization derived above, problem (18) is approximated as the following tractable problem

$$\min_{\substack{\tilde{\mathcal{B}}_k^{\mathrm{slow}}, t_{k,1}, \\ t_{k,2}, \ t}} \quad t \tag{26a}$$

$$\text{s.t.} \quad \sum_{k\in\mathcal{K}} \exp(\tilde{f}_k^{\mathrm{ap}}) \leq F^{\mathrm{ap}}, \forall k, \ \sum_{k\in\mathcal{K}} e^{\tilde{\alpha}_k} = 1, \tag{26b}$$

$$\text{(9c)}, \text{(9g)}, \text{(10b)}, \text{(19)}, \text{(21)}, \text{(25)}, \tag{26c}$$

where $\tilde{\mathcal{B}}_k^{\mathrm{slow}} = \{\tilde{\alpha}_k, \tilde{\beta}_k, \tilde{f}_k^{\mathrm{ap}}, f_k^{\mathrm{v}}, p_k^{\mathrm{c}}, p_k^{\mathrm{s}}\}$. This resulting convex subproblem can be solved iteratively, with solutions mapped back to real domain, e.g., $\alpha_k^* = \exp(\tilde{\alpha}_k^*)$.

C. Overall Framework and Computational Complexity

The overall framework is summarized in **Algorithm 1** that operates in two stages. *Offline stage:* the statistical CSI dictionary $\boldsymbol{\mathcal{H}}_{\mathrm{sta}}$ is pre-built via MC simulations over $N_{\mathrm{az}} \times N_{\mathrm{ra}}$ subregions. *Online stage:* slow-timescale resource allocation is triggered every Δt_{slow} to update $\mathcal{B}_k^{\mathrm{slow}}$ based on VUE positions, while fast-timescale beamforming update runs every Δt_{fast} to refine $\mathcal{B}_k^{\mathrm{fast}}$ using instantaneous CSI. The per-slot complexity of fast-timescale beamforming update is $\mathrm{Comp}_{\mathrm{fast}} = \mathcal{O}\big(LK\left((N_{\mathrm{u}}^{\mathrm{t}})^{3.5}\log(1/\xi) + N_{\mathrm{a}}^3\right)\big)$, and that of slow-timescale resource allocation is $\mathrm{Comp}_{\mathrm{slow}} = \mathcal{O}\big(\bar{L}K^3\log(1/\bar{\xi})\big)$. The average per-slot complexity of the proposed two-timescale framework is thus $\mathcal{O}\big(\mathrm{Comp}_{\mathrm{fast}} + \frac{\Delta t_{\mathrm{fast}}}{\Delta t_{\mathrm{slow}}}\mathrm{Comp}_{\mathrm{slow}}\big)$. Since $\Delta t_{\mathrm{slow}} \gg \Delta t_{\mathrm{fast}}$ in practice, thereby reducing the average computational complexity compared with the single-timescale baseline of $\mathcal{O}(\mathrm{Comp}_{\mathrm{fast}} +$

[4]The log substitutions are applied to positive variables. Zero-valued boundary points are approximated by a sufficiently small positive constant numerically.

$$\ln \bar{R}_k(p_k^{\mathrm{c}}, p_k^{\mathrm{s}}) \leq \ln \bar{R}_k^{(l)} + \frac{1}{\bar{R}_k^{(l)}} \frac{\partial \bar{R}_k}{\partial p_k^{\mathrm{c}}}\bigg|_{p_k^{\mathrm{c}(l)}} (p_k^{\mathrm{c}} - p_k^{\mathrm{c}(l)}) + \frac{1}{\bar{R}_k^{(l)}} \frac{\partial \bar{R}_k}{\partial p_k^{\mathrm{s}}}\bigg|_{p_k^{\mathrm{s}(l)}} (p_k^{\mathrm{s}} - p_k^{\mathrm{s}(l)}) \approx \Psi_k^{(l)}(p_k^{\mathrm{c}}, p_k^{\mathrm{s}}). \tag{23}$$

$$\frac{\partial \bar{R}_k}{\partial p_k^{\mathrm{c}}}\bigg|_{p_k^{\mathrm{c}(l)}} = \frac{Ba_k/\ln 2}{(1+\bar{\gamma}_k^{(l)}(p_k^{\mathrm{c}}, p_k^{\mathrm{s}}))(p_k^{\mathrm{s}(l)} b_k + \sigma_{\mathrm{n}}^2)}, \quad \frac{\partial \bar{R}_k}{\partial p_k^{\mathrm{s}}}\bigg|_{p_k^{\mathrm{s}(l)}} = \frac{-Bp_k^{\mathrm{c}(l)} a_k b_k/\ln 2}{(1+\bar{\gamma}_k^{(l)}(p_k^{\mathrm{c}}, p_k^{\mathrm{s}}))(p_k^{\mathrm{s}(l)} b_k + \sigma_{\mathrm{n}}^2)^2}. \tag{24}$$

Algorithm 1: Proposed Two-Timescale Optimization Framework

Input: VUE positions, Q_k, C, F^{ap}, F_k^{v}, $\boldsymbol{\mathcal{H}}_{\mathrm{sta}}$, ψ_k, $\forall k$, update periods Δt_{fast}, Δt_{slow}.
Output: $\mathcal{B}_k^{\mathrm{fast}*}$, $\mathcal{B}_k^{\mathrm{slow}*}$, and t^*.

Initialize feasible $\mathcal{B}_k^{\mathrm{slow}}$, $\mathcal{B}_k^{\mathrm{fast}}$;
for each time slot $n = 1, 2, \ldots$ **do**
  if $n \bmod \Delta t_{\mathrm{fast}} == 0$ **then**
    Estimate instantaneous CSI $\mathbf{H}_k$, $\forall k$;
    while $|t^{(\ell)} - t^{(\ell-1)}| \geq \epsilon$ and $\ell \leq L$ **do**
      $\ell \leftarrow \ell + 1$;
      for $k = 1$ **to** K **do**
        Solve (12), recover $\mathbf{f}_k^{(\ell)}$ via EVD;
        Solve (14), recover $\mathbf{u}_k^{(\ell)}$ via EVD;
        Update $\mathbf{w}_k^{(\ell)}$ via MMSE;
      $t^{(\ell)} \leftarrow \max_k \left\{\max\left\{t_k^{\mathrm{loc}},\ t_k^{\mathrm{off}(\ell)}\right\}\right\}$;
    Update $\mathcal{B}_k^{\mathrm{fast}} \leftarrow \mathcal{B}_k^{\mathrm{fast}*}$, $t^* \leftarrow t^{(\ell)}$;
  if $n \bmod \Delta t_{\mathrm{slow}} == 0$ **then**
    for $k = 1$ **to** K **do**
      Retrieve $\bar{\mathbf{H}}_k$ from $\boldsymbol{\mathcal{H}}_{\mathrm{sta}}$, construct $\bar{\mathbf{u}}_k$;
      Obtain $\bar{\mathbf{f}}_k$, $\bar{\mathbf{w}}_k$ via fast-timescale steps with $\bar{\mathbf{H}}_k$, $\bar{\mathbf{u}}_k$;
      Compute $a_k, b_k, \bar{a}_k, \bar{b}_k$ via (15);
    while $|t^{(l)} - t^{(l-1)}| \geq \epsilon$ and $l \leq \bar{L}$ **do**
      $l \leftarrow l + 1$;
      Compute $\Psi_k^{(l)}$ via Taylor expansion of $\ln \bar{R}_k$ (23);
      Solve $(\mathbf{P3.1})$, recover $\alpha_k^{(l)}, \beta_k^{(l)}, f_k^{\mathrm{ap}(l)}$;
    Update $\mathcal{B}_k^{\mathrm{slow}} \leftarrow \mathcal{B}_k^{\mathrm{slow}*}$;
return $\mathcal{B}_k^{\mathrm{fast}*}$, $\mathcal{B}_k^{\mathrm{slow}*}$, t^*.

$\mathrm{Comp}_{\mathrm{slow}}$) per fast-timescale interval. The corresponding relative reduction in average computational complexity is $\left(1 - \frac{\Delta t_{\mathrm{fast}}}{\Delta t_{\mathrm{slow}}}\right) \frac{\mathrm{Comp}_{\mathrm{slow}}}{\mathrm{Comp}_{\mathrm{fast}} + \mathrm{Comp}_{\mathrm{slow}}}$.

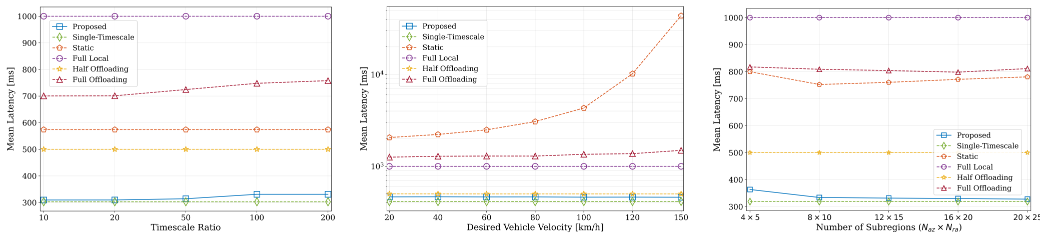
IV. Numerical Simulations

The default key simulation parameters are listed in Table I. The statistical CSI dictionary $\boldsymbol{\mathcal{H}}_{\mathrm{sta}}$ is constructed via 200 offline MC runs per subregion, and system latency is averaged over 100 independent realizations. Each VUE mobility follows the IDM with a desired velocity [12]. Five benchmarks are considered in this work: *(i)* **Single-Timescale**: Both beamforming and resource allocation updated at every fast timescale using instantaneous CSI; *(ii)* **Static**: Beamforming and resource allocation fixed at initial values (lowest-overhead baseline), without any dynamic updating; *(iii)* **Full Local**: All tasks processed locally ($\beta_k = 0$, $\forall k$); *(iv)* **Half Offloading**: Fixed $\beta_k = 0.5$, $\forall k$, with slow-timescale resource allocation and fast-timescale beamforming updates; and *(v)* **Full Offloading**: All tasks offloaded ($\beta_k = 1$, $\forall k$), with the same update structure as Half Offloading.

TABLE I: Simulation Parameters

Parameter	Value
Carrier frequency f_{c}	28 GHz
Antenna elements N_{a}, $N_{\mathrm{u}}^{\mathrm{t}}$, $N_{\mathrm{u}}^{\mathrm{r}}$	32, 16, 16
AP's coverage radius r_{ap}	300 m
Number of lanes M	4
Bandwidth B	20 MHz
Max transmit power $P_{\max}$ and noise power σ_{n}^2	20 dBm, -85 dBm
Number of VUEs K	4
Computational task size $Q_k, \forall k$	10^6 bits
VUE computational capacity $C_k^{\mathrm{u}}, \forall k$	1 GHz
AP computational capacity C^{ap}	10 GHz
CPU cycles per bit C	1,000 cycles/bit
Sensing SINR threshold $\gamma_{\mathrm{th}}^{\mathrm{s}}$	5 dB
Azimuth and range grids $(N_{\mathrm{az}}, N_{\mathrm{ra}})$	(12, 15)
Per slot length, $\Delta t_{\mathrm{fast}}, \Delta t_{\mathrm{slow}}$	1 ms, 5 ms, 100 ms
Desired velocity $v_k, \forall k$	60 km/h
Maximum number of AO iterations $L, \bar{L}$	20, 20
Convergence tolerance of the AO framework ϵ	10^{-4}

The mean latency are evaluated in Fig. 2 varying with timescale ratio $\Delta t_{\mathrm{slow}}/\Delta t_{\mathrm{fast}}$, desired VUE velocity, and grid size $N_{\mathrm{az}} \times N_{\mathrm{ra}}$ of the statistical CSI dictionary, respectively. Fig. 2a shows the mean latency versus timescale ratio within $[10, 200]$. The proposed scheme closely approaches the Single-Timescale benchmark (approximately 300 ms) at ratios up to 50, with a gap of less than 10 ms, confirming that the location-based statistical CSI dictionary provides sufficient channel information at moderate update intervals. As the ratio increases to 200, the proposed scheme experiences a slight latency increase to approximately 330 ms, remaining within a gap of 8.9% compared with the single-timescale benchmark and outperforming the Static benchmark by more than 43.6%. The Full Local and Static schemes remain insensitive to the timescale ratio due to their fixed strategies, while Full Offloading exhibits a gradual latency increase as resource allocation becomes less frequent. In addition, under a communication resource-constrained condition such as $B = 10$ MHz, the system mean latency (on a logarithmic scale) is evaluated in Fig. 2b varying desired vehicle velocities from 20 to 150 km/h. The proposed scheme and Single-Timescale baseline remain nearly flat across all velocities, demonstrating its strong



(a) (b) (c)

Fig. 2: Mean latency varying with: (a) timescale ratio $\Delta t_{\mathrm{slow}}/\Delta t_{\mathrm{fast}}$; (b) VUE's desired velocity; (c) grid size $N_{\mathrm{az}} \times N_{\mathrm{ra}}$.

robustness to mobility. In contrast, the Static scheme suffers from exponential latency increase without dynamic update of beamforming and resource allocation. Specifically, it exceeds 4×10^4 ms at 150 km/h as its fixed beamforming and resource allocation fail to track the rapidly varying channel conditions under high mobility. Full Offloading baseline shows a gradual increase, while Full Local and Half Offloading ones remain relatively stable but at significantly higher latency levels compared to our proposed scheme. These results validate that the proposed two-timescale framework effectively decouples the sensitivity to vehicular mobility through adaptive beamforming at the fast timescale, while maintaining low overhead via slow-timescale resource allocation. Moreover, Fig. 2c shows that the proposed scheme exhibits a monotonic latency decrease from approximately 366.9 ms to approximately 333.6 ms as the grid size increases from 4×5 to 12×15, after which latency maintains stable and closely approaches the Single-Timescale benchmark (around 318.8 ms). This validates that a moderate grid resolution of 12×15 provides sufficiently accurate statistical CSI for effective resource allocation without incurring excessive offline overhead. Even at the coarsest grid of 4×5, the proposed scheme still remarkably outperforms the other fixed offloading and static strategies, demonstrating the robustness of the proposed framework across different grid resolutions.

V. Conclusion

This paper proposed a two-timescale framework for joint beamforming and resource allocation in vehicular ISCC systems to minimize peak task execution latency. The fast-timescale loop updates beamforming via AO-based SDR using instantaneous CSI, while the slow-timescale loop jointly optimizes bandwidth, power, offloading ratio, and computational resources via log-domain SCA over a pre-established statistical CSI dictionary. Simulation results demonstrate that the proposed scheme closely approaches (within a gap of 8.9%) the single-timescale benchmark while achieving an order-of-magnitude reduction in per-slot computational overhead, validating its potential deployment for high-mobility vehicular ISCC systems.

References

[1] P. Liu, X. Wang, Z. Fei, Y. Wu, J. Xu, and A. Nallanathan, "Latency minimization oriented radio and computation resource allocations for 6G V2X networks with ISCC," *IEEE Trans. Commun.*, vol. 73, no. 12, pp. 15 851–15 865, Dec. 2025.

[2] K. Dong, S. A. Vorobyov, H. Yu, and T. Taleb, "Beamforming design for integrated sensing, over-the-air computation, and communication in internet of robotic things," *IEEE Internet Things J.*, vol. 11, no. 20, pp. 32 478–32 489, Oct. 2024.

[3] Q. Peng, Q. Luo, Z. Chu, Z. Lin, M. Elkashlan, P. Xiao, G. K. Karagiannidis, and C. Masouros, "Latency-aware resource allocation for integrated communications, computation, and sensing in cell-free mMIMO systems," *IEEE Trans. Wireless Commun.*, vol. 25, pp. 11 128–11 142, Feb. 2026.

[4] P. Liu, Z. Fei, X. Wang, J. Huang, J. Hu, and J. A. Zhang, "Joint offloading and beamforming design in integrating sensing, communication, and computing systems: A distributed approach," *IEEE Trans. Commun.*, vol. 73, no. 7, pp. 4697–4712, Jul. 2025.

[5] B. Hu, W. Zhang, Y. Gao, J. Du, and X. Chu, "Multiagent deep deterministic policy gradient-based computation offloading and resource allocation for ISAC-aided 6G V2X networks," *IEEE Internet Things J.*, vol. 11, no. 20, pp. 33 890–33 902, Oct. 2024.

[6] P. Liu, Z. Fei, X. Wang, X. Li, W. Yuan, Y. Li, C. Hu, and D. Niyato, "Toward intelligent edge sensing for ISCC network: Joint multi-tier DNN partitioning and beamforming design," *IEEE Trans. Wireless Commun.*, vol. 25, pp. 6774–6789, Nov. 2025.

[7] S. Hur, T. Kim, D. J. Love, J. V. Krogmeier, T. A. Thomas, and A. Ghosh, "Millimeter wave beamforming for wireless backhaul and access in small cell networks," *IEEE transactions on communications*, vol. 61, no. 10, pp. 4391–4403, Oct. 2013.

[8] K. Dong, M. Mizmizi, D. Tagliaferri, and U. Spagnolini, "Vehicular blockage modelling and performance analysis for mmwave V2V communications," in *Proc. IEEE Int. Conf. Commun.*, Seoul, Republic of Korea, May 2022, pp. 3604–3609.

[9] H. Yang, K. Zheng, L. Zhao, and L. Hanzo, "Twin-timescale radio resource management for ultra-reliable and low-latency vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 1, pp. 1023–1036, Jan. 2020.

[10] J. Yuan, H. Zhou, and Y.-C. Liang, "A two-timescale resource allocation scheme for RIS-aided cognitive radio systems," *IEEE Trans. Cognit. Commun. Networking*, vol. 12, pp. 115–127, Mar. 2025.

[11] Z. Ren, C. Pan, H. Ren, D. Wang, L. Xu, and J. Wang, "Two-timescale design for AP mode selection and power allocation of cooperative ISAC networks," *IEEE Trans. Wireless Commun.*, vol. 25, pp. 815–831, Jul. 2025.

[12] M. Treiber, A. Hennecke, and D. Helbing, "Congested traffic states in empirical observations and microscopic simulations," *Phys. Rev. E*, vol. 62, no. 2, pp. 1805–1824, Aug. 2000.

[13] X. Li, F. Liu, Z. Zhou, G. Zhu, S. Wang, K. Huang, and Y. Gong, "Integrated sensing, communication, and computation over-the-air: MIMO beamforming design," *IEEE Trans. Wireless Commun.*, vol. 22, no. 8, pp. 5383–5398, Aug. 2023.

[14] Y. Huang, H. Fu, S. A. Vorobyov, and Z.-Q. Luo, "Robust adaptive beamforming via worst-case SINR maximization with nonconvex uncertainty sets," *IEEE Trans. Signal Process.*, vol. 71, pp. 218–232, Jan. 2023.

[15] S. A. Vorobyov, A. B. Gershman, and Z.-Q. Luo, "Robust adaptive beamforming using worst-case performance optimization: A solution to the signal mismatch problem," *IEEE Trans. Signal Process.*, vol. 51, no. 2, pp. 313–324, Feb. 2003.