Analysis and Design of Vehicle Platooning Operations on Mixed-Traffic Highways

Li Jin, *Member, IEEE*, Mladen Čičić, Karl H. Johansson, *Fellow, IEEE*, and Saurabh Amin, *Member, IEEE*

***Abstract*—Platooning of connected and autonomous vehicles (CAVs) has a significant potential for throughput improvement. However, the interaction between CAVs and non-CAVs may limit the practically attainable improvement due to platooning. To better understand and address this limitation, we introduce a new fluid model of mixed-autonomy traffic flow and use this model to analyze and design platoon coordination strategies. We propose a tandem-link fluid model that considers randomly arriving platoons sharing highway capacity with non-CAVs. We derive verifiable conditions for stability of the fluid model by analyzing an underlying M/D/1 queuing process and establishing a Foster–Lyapunov drift condition for the fluid model. These stability conditions enable a quantitative analysis of highway throughput under various scenarios. The model is useful for designing platoon coordination strategies that maximize throughput and minimize delay. Such coordination strategies are provably optimal in the fluid model and are practically relevant. We also validate our results using standard macroscopic (cell transmission model) and microscopic (simulation for urban mobility) simulation models.**

***Index Terms*—Fluid model, piecewise-deterministic Markov processes, traffic control, vehicle platooning.**

I. Introduction

PLATOONING of connected and autonomous vehicles (CAVs) has the potential for significant throughput improvement [1], [2]. The idea of automatically regulating strings of vehicles is well-known [3], [4], and several experimental studies in real-world traffic conditions have been conducted in the past decades [5]–[8]. Important progress has been made in vehicle-level control design [6], [7], [9] and system-level simulations [10]. However, we still lack both link- or network-level models for evaluating the impact of platooning on highway traffic and coordinating the movement of platoons on mixed-traffic highways. In particular, although platooned CAVs have smaller intervehicle spacing, uncoordinated and randomly arriving CAV platoons may act as "moving obstacles" and result in recurrent local congestion, especially at bottlenecks (see Fig. 1(a) for illustration). Appropriate interplatoon coordination, such as regulating the headways between platoons and managing the platoon sizes, is essential for realizing the full benefits of platooning in mixed-traffic conditions.

In this article, we consider a platoon coordination problem over a generic highway section as shown in Fig. 1(a). The highway section has a downstream on-ramp, which forms a geometric bottleneck. If more traffic is arriving at the bottleneck than the bottleneck can discharge, then a queue would be growing there. The bottleneck may impede free flow of traffic, and consequently uncoordinated platoons may start inducing congestion (via local queuing), thereby impacting the upstream off-ramp. One of the objectives of platoon coordination is to ensure that the potential throughput gain due to platooning is not limited by such local congestion. Interplatoon coordination can be achieved by regulating the times at which platoons arrive

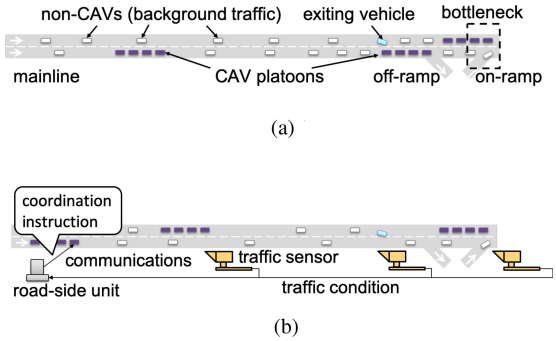


Fig. 1. Traffic scenarios without (top) and with (bottom) interplatoon coordination. (a) Without interplatoon coordination, the exiting traffic can be blocked by local congestion induced by platoons. (b) Interplatoon coordination can help mitigate the "movingobstacle" effect and maintain free flow.

Manuscript received February 18, 2020; revised August 28, 2020; accepted October 9, 2020. Date of publication October 30, 2020; date of current version September 27, 2021. This work was supported in part by the NYU Tandon School of Engineering, in part by the C2SMART University Transportation Center, in part by the US National Science Foundation CAREER Award CNS-1453126, in part by the US Air Force Office of Scientific Research, in part by the EU's Horizon 2020 Marie Skłodowska-Curie 674875, in part by the FFI VINNOVA 2014-06200, in part by the Swedish Research Council, in part by the Swedish Foundation for Strategic Research, and in part by the Knut and Alice Wallenberg Foundation. Recommended by Associate Editor C. Mahulea. *(Corresponding author: Li Jin.)*

Li Jin is with the UM Joint Institute and the School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China and with the Tandon School of Engineering, New York University, Brooklyn, NY 11201 USA (e-mail: lijin@nyu.edu).

Mladen Čičić and Karl H. Johansson are with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, 17171 Stockholm, Sweden (e-mail: cicic@kth.se; kallej@kth.se).

Saurabh Amin is with the Laboratory for Information and Decision Systems and the Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: amins@mit.edu).

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TAC.2020.3034871

TABLE I
MODELS FOR TRAFFIC WITH PLATOONING

Scale	Model	Control actions	References
Vehicle	Vehicle dynamic, car following	Throttle, break, steering	[14], [15] [11], [16]
Intersection	Discrete queuing	Signal timing, CAV coordination	[17], [18]
Segment	Partial differential equation (PDE)	Speed regulation, platoon merge/split, lane management	[19], [20]
Link	Fluid	Headway regulation, platoon merge/split,	this paper
Network	Static	Routing	[21], [22]

at the bottleneck by specifying the average or reference speed of platoons over the highway section (called *headway regulation*) and maintaining desirable platoon sizes through split or merge maneuvers (called *size management*). Implementation of such coordination strategies can be facilitated by traffic sensors that collect real-time traffic information and road-side units that send coordination instructions to each platoon [see Fig. 1(b)]; such capabilities are already available in modern transportation systems [11].

To model the interaction between platoons and background traffic, we present a multiclass tandem-link fluid model. Fluid models are standard for highway bottleneck analysis [12, Ch. 2] and allow tractable analysis [13]. Our model specifically captures two important features of platooning operations. First, platoons travel in a clustered manner (i.e., with small between-vehicle spacings and relatively large interplatoon headways), whereas non-CAVs do not follow such a configuration. In our model, random platoon arrivals are modeled as Poisson jumps in the traffic queue at the bottleneck. This model captures the inherent randomness in platooning operations [7]. Second, platoons share highway capacity with the background traffic and may act as "moving obstacles"; our model captures this interaction by constraining the sum of discharge rates of platoons and background traffic at the bottleneck with an overall capacity.

For comparison, Table I lists various models that researchers have studied for traffic flow with CAV platoons under various contexts.Fluid models are particularly suitable for link-level analysis and control design due to the following advantages. Compared with discrete queuing models, fluid models do not track individual vehicles; instead, only the aggregate flow is required. This considerably simplifies the state space and the system dynamics. Compared with partial differential equation (PDE) models, fluid models entail smaller computational requirements and enable tractable analysis. Compared with static models, fluid models allow real-time control design rather than long-term decisions such as day-to-day traffic assignment.

Two major differences are also worth noting here. First, the behavior of discrete queuing models (e.g., M/M/1) is closely related to their fluid counterparts [23]. As opposed to fully discrete queuing models, our model considers platoon as discrete Poisson arrivals and non-CAVs as a continuous inflow. This is motivated by practical situations where the arrival rate of CAV platoons (less than 5 per minute) is much lower than that of non-CAVs (50–100 per minute). Second, fluid models share some common features with PDE models [24] or their discretization [e.g., cell transmission model (CTM)][25]. In contrast to PDE models, the fluid model retains the queuing delay due to demand–capacity imbalance, but does not capture the evolution of congestion waves. Still, both the traffic fluid model and classical traffic flow models are based on conservation laws. Some authors showed that these two types of models lead to equivalent results in traffic network optimization [26]. This article also demonstrates the consistency between the fluid model and more detailed models via simulation (see Section V-B).

Using the fluid model, we study the throughput of the highway section with uncoordinated platoons.[1] We utilize an M/D/1 queuing characterization of the fluid model and establish a Foster–Lyapunov drift condition for stability of the fluid model. This leads to an easy-to-check sufficient condition for bounded queues of the uncontrolled system (see Theorem 1), which relies on the general stability/convergence theory of Markov processes [28]. We also derive explicit lower and upper bounds for throughput in the uncoordinated scenario (see Theorem 2). These results also contribute to the literature on fluid queuing systems [13], [29]–[32].

We also design a class of platoon coordination strategies that realize the full potential of platooning for throughput improvement. The control actions that we consider include regulating interplatoon headways and splitting platoons into shorter platoons. In terms of the fluid model, these regulation strategies are formulated as control laws regulating the arrival process of platoons with the knowledge of the arrival times of previous platoons. In practice, such knowledge can be obtained by tracking the movement of existing platoons on the highway section via vehicle-to-infrastructure (V2I) communications [27]. We prove that a set of control laws (see Theorem 3) stabilize the system in a fairly strong sense (bounded moment-generating function and exponential convergence to steady-state distribution) if and only if the total inflow is less than the capacity. Thus, they are optimal in the sense of throughput maximization and delay minimization. Intuitively, these control laws coordinate the movement of platoons so that they arrive at the bottleneck with relatively evenly distributed headways; thus, queuing delay is absorbed en route, congestion at the bottleneck does not build up, and spillback is eliminated.

Note that the control actions considered in this article are related to, but different from, a class of longitudinal/lateral cooperative adaptive cruise control (CACC) capabilities. We focus on link-level decision variables including the reference speed or the average speed that a traffic manager would recommend a platoon to take over a highway section and the decision whether to maintain or split a platoon over a highway section, both of which are concerned with a typical scale of 10 km or 10 min. In the context of vehicle-level CACC, however, the real-time speed is dynamically regulated to maintain string stability [14]

[1] In this article, "coordination" refers to the link-level coordination of multiple platoons. We acknowledge that "coordination" is also used to refer to the CACC applied to individual vehicles within one platoon [27].

or to mitigate local stop-and-go behavior [16], [33], [34], which involves much finer space and time resolutions. The objective of CACC design is to optimize microscopic driving behavior and regulate congestion waves, whereas our control objective is to reduce the congestion due to randomness in platoon arrivals and develops coordination strategies that work effectively in the presence of such randomness.

Finally, we discuss how the fluid model-based results can be translated to implementable actions for actual CAV platoons and validate the proposed coordination strategies in standard simulation environments. We implement the proposed strategies in a macroscopic (CTM) [35] and a microscopic (simulation for urban mobility, SUMO) [36] simulation models. Results indicate that the proposed strategy effectively and consistently improves travel times in both simulation models. In spite of multiple simplifications that our modeling approach makes, simulation results suggest that the theoretically optimal headway regulation strategy attains more than 80% of the improvement attained by the simulation-optimal strategies.

The main contributions of this article include the following.

1) A novel fluid model highway with randomly arriving platoons and constant inflow of background traffic that captures the queuing and throughput loss due to interaction between various traffic classes.
2) A set of easily checkable conditions for stability of the fluid model derived by combining ideas from queuing theory (mainly M/D/1 process) and the theory of stability of continuous-time, continuous-state Markov processes. These conditions enable quantitative analysis of throughput of mixed-autonomy highways.
3) A set of platooning strategies that regulate movement of platoons to attain maximum throughput as well as minimum delay.
4) Validation of the fluid model-based analysis and design results via simulation of macroscopic and microscopic models.

The rest of this article is organized as follows. Section II introduces the fluid model. Section III presents throughput analysis based on stability conditions. Section IV discusses a class of optimal control strategies. Section V presents the implementation and validation in simulation environments. Section VI summarizes the main results and mentions several directions for the future work.

II. Modeling and Problem Definition

In this section, we introduce our stochastic fluid model for highway bottlenecks with mixed traffic consisting of both CAVs and non-CAVs (see Section II-A) and formally define the problems that we study in the rest of this article (see Section II-B).

A. Stochastic Fluid Queuing Model

Consider a highway section with a downstream bottleneck and an off-ramp, as illustrated in Fig. 2(a). We model the highway as a tandem-link fluid queuing system in Fig. 2(b). Link 1 has a *mainline capacity* F (veh/h) and an *off-ramp capacity* R (veh/h). Link 2 has an on-ramp with capacity R, which

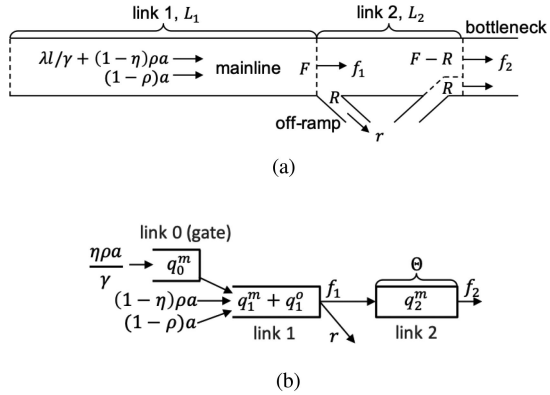


Fig. 2. Fluid queuing system for a highway section with CAVs and non-CAVs. (a) Two-link highway section and main parameters. (b) Fluid model for a system in (a).

creates a bottleneck with capacity F–R.[2] To model the effect of coordinated platooning operations [see Fig. 1(b)], we introduce a "virtual link" 0 upstream to link 1, which we refer as the *gate*. In our model, this gate can temporarily hold platoons and control their rate of release in order to regulate the downstream traffic flow. Furthermore, the storage space in link 2 (i.e., the maximum queue length that it can admit before the traffic spills over to link 1) is finite, modeled as a *buffer* with space Θ (veh). On the other hand, links 0 and 1 are assumed to have an infinite buffer space. However, the model can be extended to other typical road configurations [26], which allows extension of our approach to more general settings. Note that our model does not account for details such as car following and lane changing.

The highway section is subject to a total demand $a > 0$ (veh/h). This demand comprises ρa amount of mainline demand that is discharged into link 2, and $(1–\rho)a$ amount of demand exiting through the off-ramp. Out of the ρa mainline demand, CAVs traveling in platoons amount to a fraction $\eta \in [0, 1]$, and the remaining $(1–\eta)$ fraction is comprised of non-CAV traffic. In this demand pattern, CAVs only contribute to the mainline traffic (with demand $\eta\rho a$), and the off-ramp demand $(1–\rho)a$ is entirely comprised of non-CAVs.[3] We call ρ the *mainline ratio* and η the *platooning ratio*. We refer to the total non-CAV demand $(1 - \eta\rho)a$ as *background traffic*. For CAV platoons, the mean interarrival time is much greater than minimal interarrival times, so we consider platoons as discrete arrivals. For non-CAVs, however, the mean and the minimal interarrival times are in the same order, so we consider background traffic as continuous fluid. We consider a Poisson process rather than a general renewal process, since Poisson processes are standard models for random arrivals in transportation systems [12] and ensure that the model is Markovian [37].

CAV platoons arrive according to a Poisson process of rate λ (platoons per hour), which is given by

$$\lambda := \frac{\eta\rho a}{l} \tag{1}$$

[2]For simplicity, we assume the on-ramp capacity to be identical to the off-ramp capacity.

[3]Our setup can be extended to a more general case of CAV platoons that are bound to different destinations.

where l is the number of CAVs in each platoon. Typical values of l are between 2 and 10 [7]. For ease of presentation, we consider homogeneous platoon lengths. Nonhomogeneous platoon lengths can be modeled as jumps with randomized magnitudes, and our Lyapunov function-based approach (see Appendixes 1–3) is still valid. The Poisson process captures the randomness of the arrival of platoons.

We are now ready to introduce the stochastic fluid model.

Definition 1: The stochastic fluid model is defined as the tuple $\langle \mathcal{Q}, \mathcal{U}, G, \lambda, \mathcal{V}, S \rangle$, where following the statements hold:

1) $\mathcal{Q} := \mathbb{R}^3_{\geq 0} \times [0, \Theta]$ is the state space as well as the set of initial conditions;
2) $\mathcal{U} \subseteq \mathbb{R}_{\geq 0}$ is the set of control inputs to the vector field G (called the "gate discharge");
3) $G : \mathcal{Q} \times \mathcal{U} \to \mathbb{R}^4$ is a vector field such that $(d/dt)Q(t) = G(Q(t))$;
4) $\lambda \in \mathbb{R}_{\geq 0}$ is the Poisson rate at which resets occur (called the "arrival rate");
5) $\mathcal{V} \subseteq \mathbb{R}^3$ is the set of control inputs to the reset mapping S (called the "allocation");
6) $S : \mathcal{Q} \times \mathcal{V} \to \mathcal{Q}$ is a mapping that resets the state when a platoon arrives.

In our model, queuing happens due to sudden increases in queues, which occur at rate λ and according to the reset mapping S, and interaction between queues in various links, which is captured by the vector field G. We consider the control inputs (u, v) to be determined by a *control policy* (μ, ν), i.e., $(u, v) = (\mu(q), \nu(q))$.

Definition 2 (Control policy): A control policy (μ, ν) is specified by functions $\mu : \mathcal{Q} \to \mathcal{U}$ and $\nu : \mathcal{Q} \to \mathcal{V}$.

We will describe how control policies for the stochastic fluid model can be translated to platoon coordination strategies in Section IV.

Given a control policy (μ, ν), the model's dynamics can be expressed via the *infinitesimal generator*

$$\mathscr{L} g(q) = G^T(q; \mu(q)) \nabla_q g(q) + \lambda \Big(g(S(q; \nu(q))) - g(q) \Big) \quad q \in \mathcal{Q} \tag{2}$$

where g is a differentiable function [38]. In the above, the first term on the right-hand side results from the fluid dynamics governed by the vector field G, and the second term results from the resets governed by the reset mapping S.

In summary, our model (as well as the subsequent analysis) focuses on the impact due to the following parameters:

1) total demand a;
2) platooning ratio η (or equivalently, platoon arrival rate λ);
3) platoon size l;
4) buffer size Θ.

The rest of this section is devoted to specifying the elements in the tuple $\langle \mathcal{Q}, \mathcal{U}, G, \lambda, \mathcal{V}, S \rangle$.

1) State Space $\mathcal{Q}$: We use q_0^m to denote the CAVs held in the gate, q_1^m and q_2^m to denote the queues of mainline traffic in links 1 and 2, respectively, and q_1^o to denote the queue of off-ramp traffic in link 1. The *state* of the stochastic fluid model is $q = [q_0^m \; q_1^m \; q_1^o \; q_2^m]^T \in \mathcal{Q}$. Note that q_0^m consists of only CAVs, q_1^m and q_2^m consist of both CAVs and non-CAVs, and q_1^o consists of only non-CAVs. A key characteristic of platooning is the reduced intervehicle spacing. We model this by scaling down the CAV part of q_k^m by a factor $\gamma > 1$: a platoon of l CAVs are roughly equivalent to l/γ non-CAVs in terms of the occupied road space, where γ is the ratio between the inter-non-CAV spacing H and the inter-CAV spacing h (see Fig. 3). A typical value for γ is 2 [7], [39]. Hence, q is the vector of *effective* queue lengths with the CAV part scaled down, which is in general smaller than the *nominal* queue lengths. Throughout this article, we use $Q(t) = [Q_0^m(t) \; Q_1^m(t) \; Q_1^o(t) \; Q_2^m(t)]^T$ to denote the vector of queues at time t and q to denote a particular state.

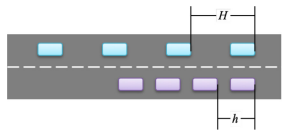


Fig. 3. H is the spacing between ordinary vehicles, whereas h is the spacing within a platoon; $\gamma = H/h$.

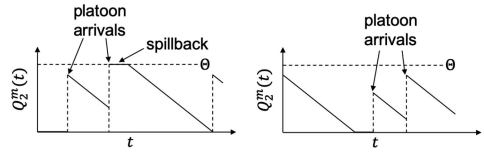


Fig. 4. Two examples of trajectories of $Q_2^m(t)$ with (left) and without spillback (right).

2) Gate Discharge $\mathcal{U}$: The rate at which the gate discharges traffic to link 1 is $u \in \mathcal{U}$. We consider u to be controlled by a *gate discharge policy* $\mu : \mathcal{Q} \to \mathbb{R}_{\geq 0}$ satisfying the following.

Assumption 1: The gate discharge policy μ satisfies the following:

i) $\mu(q)$ is nonnegative, bounded, and piecewise continuous in q;
ii) $\mu(q) = 0$ for q such that $q_0^m = 0$;
iii) $\mu(q)$ is nonincreasing in q_1^m, q_1^o, q_2^m.

In the above, (i) ensures regularity to facilitate analysis, (ii) means that the gate discharge must vanish if $q_0^m = 0$, and (iii) means that CAVs in the gate will be discharged slower if the downstream queues are longer. We use $\mathscr{U}$ to denote the set of gate discharge policies satisfying the above assumption. This assumption basically ensures that $Q(t)$ is bounded and piecewise continuous in t.

3) Vector Field G: The vector field G specifies the model's dynamics between resets. Specifically, the inflow of mainline background traffic is $(1-\eta)\rho a$, and the inflow of off-ramp background traffic is $(1-\rho)a$. If the queue in link 2 is less than the buffer size, then the queue in link 1 is discharged at the link's capacity F; otherwise, the queue propagates to link 1 and reduces the off-ramp flow (called *spillback*, see Fig. 4). The flows in Fig. 2(b) are given by

$$f_0(q; \mu(q)) := \mu(q) \tag{3a}$$

$$f_1(q;\mu(q)) := \begin{cases} \min\{(1-\eta)\rho a + \mu(q), F\} & \text{if} q_1^m = 0, q_2^m < \Theta \\ F & \text{if} q_1^m > 0, q_2^m < \Theta \\ \min\{(1-\eta)\rho a + \mu(q), F-R\} & \text{if} q_1^m = 0, q_2^m = \Theta \\ F-R & \text{if} q_1^m > 0, q_2^m = \Theta \end{cases} \tag{3b}$$

$$f_2(q;\mu(q)) := \begin{cases} \min\{f_1(q;\mu(q)), F-R\} & \text{if} q_2^m = 0 \\ F-R & \text{if} q_2^m > 0 \end{cases} \tag{3c}$$

$$r(q;\mu(q)) := \begin{cases} \min\{(1-\rho)a, F - f_1(q;\mu(q)), R\} & \text{if} q_1^o = 0, q_2^m < \Theta \\ \min\{F - f_1(q;\mu(q)), R\} & \text{if} q_1^o > 0, q_2^m < \Theta \\ \min\{(1-\rho)a, F-R & \\ \quad -f_1(q;\mu(q)), R\} & \text{if} q_1^o = 0, q_2^m = \Theta \\ \min\{F-R-f_1(q;\mu(q)), R\} & \text{if} q_1^o > 0, q_2^m = \Theta. \end{cases}$$

Equations (3b) and (3d) indicate that spillback happens if $q_2^m = \Theta$: whenever the threshold is attained, the upstream capacity is dropped. To focus on the impact of capacity, we assume that traffic queues are always discharged at the rate of F or F–R.[4] Then, the fluid dynamics is specified by the vector field $G : \mathcal{Q} \times \mathcal{U} \to \mathbb{R}^4$ defined as

$$G_0^m(q;\mu(q)) := -\mu(q) \tag{4a}$$

$$G_1^m(q;\mu(q)) := (1-\eta)\rho a - (f_1(q;\mu(q)) + r(q;\mu(q))) \tag{4b}$$

$$G_1^o(q;\mu(q)) := (1-\rho)a - r(q;\mu(q)) \tag{4c}$$

$$G_2^m(q;\mu(q)) := f_1(q;\mu(q)) - f_2(q;\mu(q)). \tag{4d}$$

To emphasize that the vector field depends on the control policy, we use the notation $G(q,\mu(q))$. Note that for an admissible gate discharge policy $\mu \in \mathscr{U}$, G is bounded and piecewise continuous in q. Since we focus on the aggregate behavior of both traffic classes, our results hold for a variety of capacity-sharing models (see, e.g., [40]). In addition, microscopic maneuvers, such as overtaking, are implicitly captured by the flow dynamics.

4) Arrival Rate λ: CAV platoons arrive according to a Poisson process of rate λ (per hour), which is given by (1). λ specifies the rate at which the continuous state $Q(t)$ is reset. Randomness of the platoon arrival process can be attributed to the process of platoon formation [41], [42].

5) Allocation $\mathcal{V}$: Platoon control is modeled by a vector $v = [v_0\ v_1\ v_2]^T \in \mathcal{V}$. Arriving platoons are allocated to each link according to v: for example, $v = [l,0,0]^T$ means that a platoon is allocated to the gate. Recall that v is determined by a mapping $\nu : \mathcal{Q} \to \mathcal{V}$. We assume the following for ν.

Assumption 2: The allocation policy ν satisfies the following:

i) $\sum_{k=0}^{2} \nu_k(q) = 0, 0 \le \nu_0(q) \le l/\gamma, -l/\gamma \le \nu_1(q) \le 0$, $\max\{-l/\gamma, q_2^m - \Theta\} \le \nu_2(q) \le 0$;

ii) $\nu_2(q) = 0$ for q such that $q_1^m > 0$, and $\nu_1(q) = \nu_2(q) = 0$ for q such that $q_0^m > 0$;

iii) $\nu_k(q)$ is nonincreasing in q_k^m and nondecreasing in q_j^m for $j \neq k$, and $\nu_k(q)$ is nonincreasing in q_1^o for $k = 1$ and nondecreasing in q_1^o for $k \neq 1$.

In the above, (i) means that ν only distributes but does not create traffic, (ii) results from the "first-come-first-served" principle: a platoon cannot be allocated to link k if there is a nonzero queue in link $k-1$, and (iii) means that more traffic is allocated to a link with a shorter queue. We use $\mathscr{V}$ to denote the set of gate discharge policies satisfying the above assumption.

6) Reset Mapping S: Arrivals of platoons lead to sudden increases in the state $Q(t)$, and the reset mapping is given by

$$S_0^m(q;v) := q_0^m + v_0 \tag{5a}$$

$$S_1^m(q;v) := q_1^m + (q_2^m + l/\gamma - \Theta)_+ + v_1 \tag{5b}$$

$$S_1^o(q;v) := q_1^o \tag{5c}$$

$$S_2^m(q;v) := \min\{\Theta, q_2^m + l/\gamma\} + v_2. \tag{5d}$$

In particular, $S(q;0)$ represents the reset mapping if no control is applied (i.e., platoons are not coordinated).[5] For $v = 0$, no platoons will be allocated to the gate upon arrival; instead, every platoon will be allocated to link 2 unless $Q_2^m(t)$ attains the buffer size Θ. If a platoon arrives at time t, then the state is reset according to

$$Q(t) = S\left(Q(t_-); v\right)$$

where $Q(t_-)$ is the vector of queues immediately before the arrival.

B. Problem Definition

The main questions that we study are as follows.

1) Given model parameters and a control policy, how to determine whether the queues are bounded (in expectation) and compute or estimate the model's throughput?
2) How to design the control policy to ensure bounded queues and improve throughput and travel time?

To study the above questions, we introduce the following definitions.

First, following [23], we define stability as follows.

Definition 3 (Stability): The stochastic fluid model is stable if there exists $Z < \infty$ such that for each initial condition $q \in \mathcal{Q}$

$$\limsup_{t\to\infty} \frac{1}{t} \int_{s=0}^{t} \mathrm{E}\left[|Q(s)|\right] ds \le Z. \tag{6}$$

Practically, stability means that expected queue size is bounded, and hence the probability of long queues is small.

Second, given a control policy (μ,ν), the stochastic fluid model typically admits an *invariant set*, which is defined as follows:

Definition 4 (Invariant set): Given a control policy (μ,ν), a compact set $\mathcal{M}_{\mu,\nu} \subseteq \mathcal{Q}$ is an invariant set if the following statements hold:

i) $\lim_{t\to\infty} \Pr\{Q(t) \in \mathcal{M}_{\mu,\nu} | Q(0) = q\} = 1, \quad \forall q \in \mathcal{Q}$;

ii) $Q(t) \in \mathcal{M}_{\mu,\nu}$, $\forall Q(0) = q \in \mathcal{M}_{\mu,\nu}$.

[4] Simulation results show that this simplified flow model is largely consistent with more sophisticated models such as the CTM (see Section V-B1).

[5] If $v = 0$, then no CAVs will ever be allocated to the gate; hence the gate discharge has no impact.

The interpretation of an invariant set is that (i) for each initial condition, the process $\{Q(t); t \geq 0\}$ enters the set $\mathcal{M}_{\mu,\nu}$ almost surely (a.s.), and (ii) if the process $\{Q(t); t \geq 0\}$ starts within $\mathcal{M}_{\mu,\nu}$, then it never leaves $\mathcal{M}_{\mu,\nu}$. Since stability is defined for an infinite time horizon in (6), we can focus on the model's evolution over $\mathcal{M}_{\mu,\nu}$ rather than over $\mathcal{Q}$; this simplifies the analysis. Also note that $\mathcal{M}_{\mu,\nu}$ depends on (μ, ν) and is thus typically determined based on characteristics of (μ, ν).

The theoretical tool that we use to establish stability is the *Foster–Lyapunov criterion*, which is a sufficient condition for (6):

Foster–Lyapunov criterion [43]: Consider a Markov process with an invariant set $\mathcal{Y}$ and infinitesimal generator $\mathscr{A}$. If there exist a Lyapunov function $W : \mathcal{Y} \to \mathbb{R}_{\geq 0}$ and constants $c > 0$, $d < \infty$ satisfying

$$\mathscr{A}W(y) \leq -cg(y) + d \quad \forall y \in \mathcal{Y} \tag{7}$$

then for each initial condition $y \in \mathcal{Y}$

$$\limsup_{t\to\infty} \frac{1}{t} \int_{\tau=0}^{t} \mathrm{E}[g(Y(t))]d\tau \leq d/c. \tag{8}$$

In the above, (7) is called the "drift condition" [43]. Verifying the drift condition is in general challenging, since it requires finding an effective Lyapunov function, which is not straightforward for a nonlinear system as our stochastic fluid model, and checking the inequality (7) over a possibly unbounded set $\mathcal{Y}$, which involves a nonconvex optimization. In Section III, we argue how we address these challenges.

Third, with the notion of stability, we define the stochastic fluid model's *throughput* as follows.

Definition 5 (Throughput): Given a control policy $(\mu, \nu) \in \mathscr{U} \times \mathscr{V}$, the throughput of the stochastic fluid model is

$$\bar{a}_{\mu,\nu} := \sup a \quad \text{s.t. fluid model is stable under } (\mu, \nu) \in \mathscr{U} \times \mathscr{V}.$$

where $\bar{a}_{\mu,\nu}$ is defined as supremum rather than maximum, since the stability constraint may lead to strict inequalities. As indicated in the above definition, the key to throughput analysis is to develop stability conditions for the stochastic fluid model, which we discuss in Section III. We have the following preliminary result for throughput.

Lemma 1 (Nominal throughput): For any control policy $(\mu, \nu) \in \mathscr{U} \times \mathscr{V}$, throughput $\bar{a}_{\mu,\nu}$ of the stochastic fluid model is upper bounded by

$$\bar{a}_{\mu,\nu} \leq a^* := \min\left\{\frac{R}{1-\rho}, \frac{F-R}{(\eta/\gamma+1-\eta)\rho}\right\}. \tag{9}$$

We call a^* as defined in (9) the *nominal throughput*. To interpret the expression for a^*, note that the first (resp. second) term in $\min\{\cdot\}$ results from the capacity constraint of the off-ramp (resp. the mainline bottleneck). Importantly, we will show that the nominal throughput cannot always be attained due to the interaction between CAV and non-CAV traffic and due to spillback of traffic queues.

Finally, for control design, we consider the following formulation:

$$\text{(P)} \quad \begin{aligned} &\text{decision:} && (\mu, \nu) \in \mathscr{U} \times \mathscr{V} \\ &\text{objective:} && \max \quad \bar{a}_{\mu,\nu} \\ &&& \text{s.t.} \quad \text{stability}. \end{aligned}$$

That is, the first objective is stabilization, and the second objective is throughput maximization or queue minimization. Note that the objective value of (P) is upper bounded by a^* in Lemma 1.

III. Stability and Throughput

In this section, we study the stability and throughput of the stochastic fluid model, which provides insights about the efficiency of the mixed-traffic highway.

The first main result gives a criterion to check the stability of the stochastic fluid model under a given control policy (μ, ν).

Theorem 1 (Stability criterion): Suppose that the stochastic fluid model admits an invariant set $\mathcal{M}_{\mu,\nu} \subseteq \mathcal{Q}$. The stochastic fluid model is stable if

$$a < a^*, \quad \text{and} \tag{10a}$$

$$\begin{aligned} \max_{\substack{\xi\in\mathcal{M}_{\mu,\nu}:\\ \xi_0^m=\xi_1^m=\xi_1^o=0}} & \frac{\xi_2^m}{\Theta}\left((1-\eta)\rho a - (F-R)\right) \\ & + \frac{\eta\rho a}{l}\Bigg(S_0^m(\xi;\nu(\xi)) + S_1^m(\xi;\nu(\xi)) \\ & + \frac{1}{2\Theta}\left(S_2^m(\xi;\nu(\xi))\right)^2 - \frac{1}{2\Theta}(\xi_2^m)^2\Bigg) \\ < & \frac{R-(1-\rho)a}{(1-\rho)a}\left((F-R) - (\eta/\gamma+1-\eta)\rho a\right). \end{aligned} \tag{10b}$$

One can interpret the stability criterion as follows. Equation (10a) results from the nominal upper bound in Lemma 1. Equation (10b) essentially results from the interaction between the mainline and the off-ramp traffic. Equation (10b) also captures the influence of the control policy (μ, ν). Although the complexity of the maximization on the left-hand side of (10b) depends on the control policy (μ, ν), the decision variable ξ can only vary in the direction of ξ_2^m; all the other components of ξ must be zero. Hence, the maximization involves essentially only one decision variable ξ_2^m, which takes values from a compact interval $[0, \Theta]$, and is thus not hard to solve numerically. This is a significant refinement of the Foster–Lyapunov criterion, which, in its general form, does not give a ready-to-use stability criterion for our stochastic fluid model. Theorem 1 can be used for throughput analysis by finding the largest demand a that satisfies (10a)–(10b). Since Theorem 1 is a sufficient condition for stability, it leads to a lower bound for throughput.

We prove Theorem 1 by considering a quadratic Lyapunov function and establishing the Foster–Lyapunov criterion [28] for the queuing process. The main technique is to relate the fluid queuing process to an underlying M/D/1 process, which we discuss below. The detailed proof is in Appendix 1.

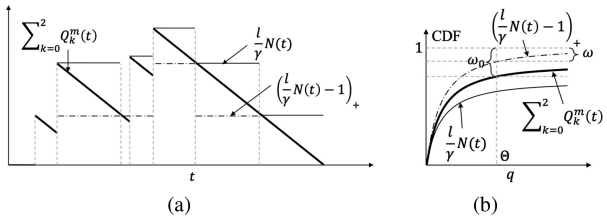


Fig. 5. Relation between M/D/1 process $N(t)$ and fluid process $\sum_{k=0}^{2} Q_k^m(t)$. (a) Trajectories. (b) Steady-state CDFs.

In particular, if no control is applied, then (10b) can be manually solved, and explicit lower and upper bounds for throughput can be derived. To see this, consider the process

$$N(t) := \left\lceil \frac{\gamma}{l} \left(\sum_{k=0}^{2} Q_k^m(t) \right) \right\rceil \quad t \geq 0 \tag{11}$$

which satisfies the following.

Lemma 2: The process $\{N(t); t \geq 0\}$ is an M/D/1 process with arrival rate $\lambda = \eta\rho a/l$ and service time $s := \frac{l}{\gamma(F-R-(1-\eta)\rho a))}$. Furthermore

$$((l/\gamma)N(t)-1)_+ \leq \sum_{k=0}^{2} Q_k^m(t) \leq (l/\gamma)N(t) \quad \forall t \geq 0. \tag{12}$$

That is, we can use $N(t)$ to bound $\sum_{k=0}^{2} Q_k^m(t)$ [see Fig. 5(a)]. The steady-state probabilities π_n of the M/D/1 process are given by a standard result in queuing theory [44]

$$\pi_n = \begin{cases} 1-\lambda s & n=0 \\ (1-\lambda s)(e^{\lambda s}-1) & n=1 \\ (1-\lambda s)\left(e^{n\lambda s} + \sum_{k=1}^{n-1} e^{k\lambda s}(-1)^{n-k} \right. & \\ \quad \left. \times \left[\frac{(k\lambda s)^{n-k}}{(n-k)!} + \frac{(k\lambda s)^{n-k-1}}{(n-k-1)!} \right] \right) & n \geq 2. \end{cases} \tag{13}$$

Using the above result, we can obtain a lower bound

$$\omega = 1 - \sum_{n=0}^{\lceil \gamma\Theta/l \rceil} \pi_n \tag{14}$$

for the actual fraction of time ω_0 that the stochastic fluid model experiences spillback [see Fig. 5(b)].

With the above arguments, we can state the second main result of this section as follows (the proof is in Appendix 2).

Theorem 2 (Throughput without control): Suppose that the stochastic fluid model is not controlled, i.e., $\mu(q) = 0$ and $\nu(q) = 0$ for all $q \in \mathcal{Q}$. Then, the throughput $\bar{a}$ of the model is bounded by

$$\min \left\{ \frac{F-R}{\rho(\eta/\gamma+1-\eta)}, \frac{R}{1-\rho+\frac{1}{2}\left(\sqrt{\zeta^2+\frac{2\rho R l}{\gamma\Theta(F-R)}}-\zeta\right)} \right\} \leq \bar{a} \leq \min \left\{ \frac{F-R}{\rho(\eta/\gamma+1-\eta)}, \frac{(1-\omega)R}{(1-\rho)} \right\} \tag{15}$$

where

$$\zeta = (1-\rho) - \rho(\eta/\gamma+1-\eta)\frac{R}{F-R} \tag{16}$$

TABLE II
NOMINAL MODEL PARAMETER VALUES

Quantity	Notation	Nominal value
mainline capacity	F	4500 veh/hr
ramp capacity	R	1500 veh/hr
platooning ratio	η	0.2
platoon scaling coefficient	γ	2
platoon size	l	5 veh
mainline ratio	ρ	0.75
buffer size	Θ	50 veh

and ω is given by (14).

Using the bounds in Theorem 2, we can analyze the model's throughput without control. Fig. 6(a)–(c) summarizes our results for throughput analysis with the nominal parameters in Table II. Importantly, due to the interaction between multiple traffic classes and due to lack of coordination between platoons, the nominal throughput given by Lemma 1 is not attained. Specific discussions about the figures are as follows.

1) Fig. 6(a) shows that higher fraction of platooning lead to higher throughput, which is consistent with previous results [17]. The nominal throughput is not attained due to interaction between CAVs and non-CAVs.
2) Fig. 6(b) shows that platooning does improve throughput, but overly long platoons may result in lower throughput. The reason is that longer platoons have stronger impact on local traffic and cause larger local congestion. For this example, the empirically optimal platoon size is four CAVs.
3) As shown in Fig. 6(c), when the buffer size Θ is small (e.g., less than 20), spillback occurs frequently, and thus an obvious throughput drop (with respect to the nominal throughput) is observed. As Θ approaches infinity, spillback hardly occurs, and both bounds approach the nominal throughput given by Lemma 1.

IV. CONTROL DESIGN

In this section, we study a set of control policies that attain the nominal throughput given by Lemma 1.

One can indeed use Theorem 1 as the stability constraint and solve (P). However, since Theorem 1 is a sufficient condition, it in general leads to a suboptimal solution. The main result of this section is a sufficient condition for optimality of a given control policy.

Theorem 3 (Optimality criterion): Suppose that a control policy $(\mu, \nu) \in \mathscr{U} \times \mathscr{V}$ satisfies:

$$\mu(q) = 0 \quad \forall q : q_1^m > 0 \text{ or } q_2^m = \Theta \tag{17a}$$

$$\mu(q) \leq F - ((1-\eta)\rho + (1-\rho))a \quad \forall q \in \mathcal{Q} \tag{17b}$$

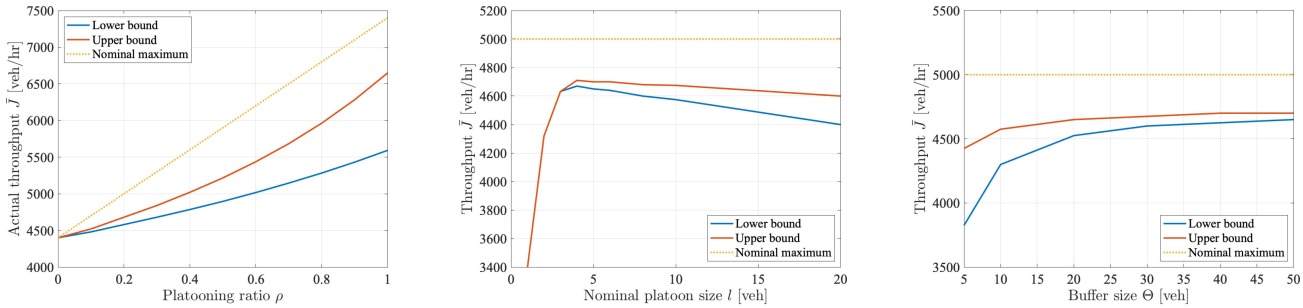
$$\mu(q) + (1-\eta)\rho a \geq F - R \quad \forall q : q_0^m > 0 \text{ and } q_2^m = 0 \tag{17c}$$

$$S_1^m(q; \nu(q)) = 0 \quad \forall q \in \mathcal{Q} \tag{17d}$$

$$S_2^m(q; \nu(q)) < \Theta \quad \forall q \in \mathcal{Q}. \tag{17e}$$

Then, (μ, ν) stabilizes the stochastic fluid model if and only if

$$a < a^* \tag{18}$$



(a) (b) (c)

Fig. 6. Relation between throughput and model parameters. Bounds are computed using Theorem 2. "Nominal maximum" refers to a^* in Lemma 1. In (a) [resp. (b)], $\eta = 0$ (resp. $l = 1$) means no platooning. (a) Throughput versus fraction. (b) Throughput versus platoon size. (c) Throughput versus buffer size.

where a^* is the nominal throughput given by (9). Furthermore, under (18), the control (μ, ν) minimizes the total queue length $|Q(t)|$ among all admissible controls at any time $t \geq 0$, and the time-average queuing delay converges as follows:

$$\lim_{t\to\infty} \frac{1}{t}\int_{\tau=0}^{t} |Q(\tau)|d\tau \overset{a.s.}{=} \bar{Q} := \frac{\eta\rho a l}{2\gamma^2(F-R-(1-\eta)\rho a)} \times \left(\frac{\eta\rho a}{\gamma(F-R-(\eta/\gamma+1-\eta)\rho a)}+1\right). \tag{19}$$

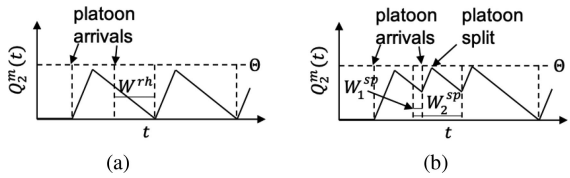
Note that Theorem 3 addresses not only stabilization, but also throughput maximization and queue minimization. The optimality criterion (17a)–(17d) can be interpreted as follows. Equations (17a), (17b), and (17d) ensure no queuing in link 1 (i.e., $Q_1^m(t) = 0$). Equations (17a) and (17e) ensure no spillback at link 2 (i.e., $Q_2^m(t) < \Theta$). Equation (17c) ensures that as long as there is a nonzero queue in the gate (i.e., $Q_0^m(t) > 0$), link 2 must be discharging traffic at its capacity F–R.

We use $\mathscr{U}^* \times \mathscr{V}^*$ to denote the set of control policies satisfying (17a)–(17e). Since $(\mu, \nu) \in \mathscr{U}^* \times \mathscr{V}^*$ stabilizes the stochastic fluid model if and only if the demand is less than the nominal throughput, (μ, ν) maximizes the throughput. In addition, each control in $\mathscr{U}^* \times \mathscr{V}^*$ not only maximizes throughput, but also minimizes queuing delay in a sample pathwise manner. Furthermore, an analytical expression for the mean queuing delay is obtained. The proof of Theorem 3 is in Appendix 3.

In the rest of this section, we discuss two concrete control policies with practical interpretations, viz., headway regulation and platoon size management. Here, we focus on their formulation in the stochastic fluid model. In Section V-A, we discuss and demonstrate how they can be implemented in practice (e.g., speed).

A. Headway Regulation

Under this strategy, if a large number of platoons arrive within a short time period, some platoons will be allocated to the gate so that their arrival at the bottleneck is postponed to avoid cumulative congestion at the bottleneck. In the stochastic fluid model, the headway regulation strategy can be formulated as a



(a) (b)

Fig. 7. Illustration of how regulating headway $(\mu^{\text{hr}}, \nu^{\text{hr}})$ and splitting platoons $(\mu^{\text{sm}}, \nu^{\text{sm}})$ avoid spillback (see Fig. 4 for the uncontrolled case). (a) Regulating headway. (b) Splitting platoon.

control policy (μ, ν) such that

$$\mu^{\text{hr}}(q) := \begin{cases} \alpha \text{ if } q_0^m > 0, q_2^m = 0 \\ \alpha \text{ if } \gamma q_0^m/l \notin \{0,1,2,\dots,\} \\ 0 \text{ o.w.} \end{cases} \tag{20a}$$

$$\nu^{\text{hr}}(q) := \begin{bmatrix} l/\gamma \\ q_1^m - S_1(q;0) \\ q_2^m - S_2(q;0) \end{bmatrix} \tag{20b}$$

where α is the saturation flow rate of CAVs given by

$$\alpha = v_0/h \tag{21}$$

where v_0 is the nominal speed of the highway section, and h is the between-vehicle spacing within platoons. One can check that $(\mu^{hr}, \nu^{hr}) \in \mathscr{U}^* \times \mathscr{V}^*$.

Fig. 7(a) illustrates the idea of this control policy: if two platoons enter the highway with a short interarrival time, the following platoon is decelerated so that its arrival at the bottleneck is postponed by W^{hr} amount of time. Consequently, the platoons arrive at the bottleneck with sufficient headway in between. As illustrated in Fig. 7(a), $(\mu^{\text{hr}}, \nu^{\text{hr}})$ essentially regulates times at which platoons arrive at the bottleneck so that congestion does not build up or spill back from the bottleneck, and the off-ramp traffic is not blocked. To compute W^{hr}, suppose that a platoon enters the highway at time t and the state immediately before the arrival is $Q(t_-)$, then W^{hr} is the solution to the deterministic equation

$$\int_{s=t}^{W^{\text{hr}}} \mu^{\text{hr}}(Q(s))ds = Q_0^m(t_-). \tag{22}$$

Note that W^{hr} is independent of any platoon arrivals after t.

B. Platoon Size Management

If the highway is congested, long platoons will be disadvantageous at bottlenecks due to their sizes. Consequently, the operator can instruct platoons to split into shorter platoons to mitigate local congestion. The decision variable is whether to split or maintain a platoon as it enters the highway. In the stochastic fluid model, splitting a platoon can be modeled by a control policy $(\mu^{\text{sm}}, \nu^{\text{sm}})$ defined as follows:

$$\mu^{\text{sm}}(q) := \begin{cases} \alpha \text{ if } q_0^m > 0, q_2^a \le \Theta - \frac{l((\eta/\gamma+1-\eta)\rho a-(F-R))}{2\gamma\eta\rho a} \\ \alpha \text{ if } (\gamma q_0^m/l) \notin \{0, \frac{1}{2}, 1, \frac{3}{2}, \dots, \} \\ 0 \text{ o.w.} \end{cases} \tag{23}$$

$$\nu^{\text{sm}}(q) := \begin{bmatrix} l/\gamma \\ q_1^m - S_1(q;0) \\ q_2^m - S_2(q;0) \end{bmatrix}. \tag{24}$$

One can check that $(\mu^{\text{sm}}, \nu^{\text{sm}}) \in \mathscr{U}^* \times \mathscr{V}^*$. As shown in Fig. 7(b), μ^{sm} opens the gate only if link 2 has sufficient space to accept at least half a platoon. Furthermore, if q_2^a is close to the buffer size Θ, ν^{sm} will split a platoon into two short platoons and allocate the two short platoons to links 0 and 2 to avoid spillback. In practice, suppose a platoon enters the highway at time t and let T be the solution to

$$\int_{s=t}^{T} f_2(Q(s))ds = Q_0^m(t_-) + Q_1^m(t_-) + \min\left\{\frac{l}{2\gamma}, \left(\Theta - \frac{l}{2\gamma} - Q_2^m(t_-)\right)_+\right\} \tag{25}$$

where $Q_k^m(t_-)$ is the queue size in the gate immediately before the platoon arrives. Then, the delays W_1^{sm} and W_2^{sm} indicated in Fig. 7(b) are given by

$$W_1^{\text{sm}} = T,\ W_2^{\text{sm}} = T + \frac{l}{2\gamma(F-R-(1-\eta)\rho a)}. \tag{26}$$

V. Implementation, Simulation, and Discussion

In this section, we translate the control laws discussed in the previous section to platoon coordination instructions that can be implemented in practice (see Section V-A). We also validate the optimality of the headway regulation strategy via two standard simulation environments, viz., CTM and SUMO (see Section V-B).

A. Implementation of Proposed Control Policies

We now discuss how the two platoon coordination strategies presented in Section IV, viz., headway regulation and size management, can be translated to implementable instructions for platoons. Fig. 8 illustrates the implementation. These strategies are enabled by modern V2I communications technologies [45]. We do not explicitly consider lower level control actions such as longitudinal and lateral control; instead, we assume that the platoons are equipped with adequate lower level controllers that can implement the instructions from the operator.

1) Headway Regulation: To regulate headways, the operator sends a recommended speed to each platoon when it enters the highway, and no more instructions need to be sent to this platoon [see Fig. 8(a)]. The decision variable is the average speed for each platoon over the highway section, or, equivalently, the time at which a platoon is scheduled to arrive at the bottleneck. In practice, let L_1 and L_2 be the lengths of links 1 and 2, respectively, as in Fig. 2(a), and let v_0 be the nominal speed for the highway section. The recommended speed for an incoming platoon is

$$v^{\text{hr}} = \frac{L_1 + L_2}{\frac{L_1+L_2}{v_0} + W^{\text{hr}}} \tag{27}$$

where W^{hr} is given by (22).

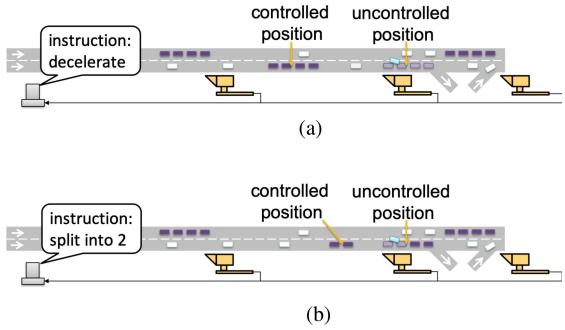


Fig. 8. Two practical platoon coordination strategies. (a) Headway regulation. (b) Platoon size management.

2) Size Management: To manage the size of platoons, when a platoon enters the highway, the coordination strategy first predicts the traffic condition if the platoon arrives at the bottleneck without any intervention; then, if congestion is predicted at the bottleneck, the strategy will split the platoon into shorter platoons [see Fig. 8(b)]. The fluid model can be used to make such predictions. In practice, suppose a platoon enters the highway at time t and let T be the solution to (25). Specifically, the coordination decision is made in the following two steps.

1) *Whether to split*. The platoon is instructed to split if $Q_2^m(T) \ge \Theta - \frac{l}{2\gamma}$ and not to split otherwise; if the platoon is splitting, then the separation (i.e., headway) between the short platoons will be $l/(2\gamma(F-R-(1-\eta)\rho a))$.
2) *When to arrive at bottleneck*. If the platoon is not splitting, then it will travel at the nominal speed and no intervention will be needed. If the platoons are splitting into two short platoons, the leading short platoon will travel at the speed

$$v_1^{\text{sm}} = \frac{L_1 + L_2}{\frac{L_1+L_2}{v_0} + W_1^{\text{sm}}}$$

and the following short platoon will travel at the speed

$$v_2^{\text{sm}} = \frac{L_1 + L_2}{\frac{L_1+L_2}{v_0} + W_2^{\text{sm}}}$$

where W_1^{sm} and W_2^{sm} are given by (26).

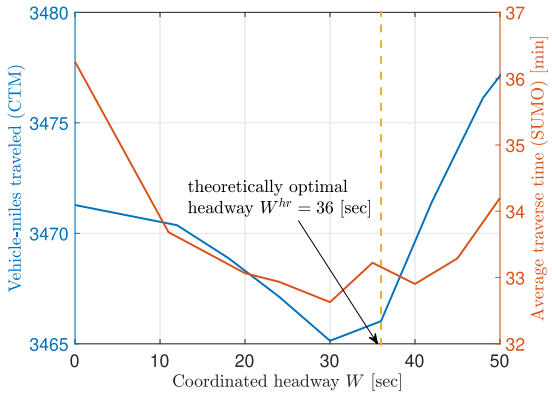
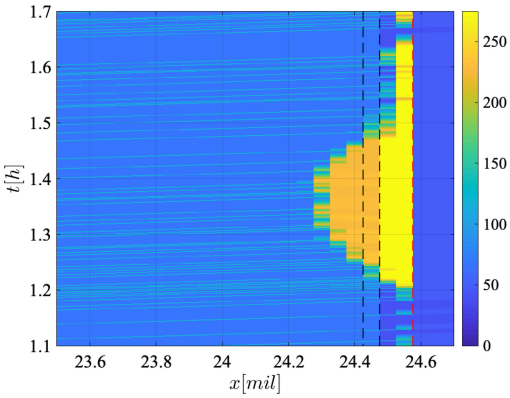
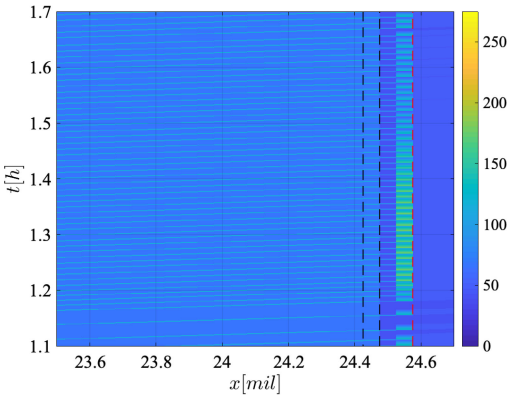


Fig. 9. Simulated performance metrics for various values of minimal interplatoon headways.



(a) Without coordination.



(b) With coordination.

Fig. 10. Traffic density contour plots for CTM simulation. The vertical red (resp. black) dashed line indicates the location of the bottleneck (resp. off-ramp). Color indicates traffic density in veh/mi. (a) Without coordination. (b) With coordination.

B. Simulation-Based Validation

The purpose of the simulations is to show that the optimal headway regulation strategy designed using the fluid model-based approach is consistent with the simulation-optimal values. We use two standard simulation models, viz., the CTM) [35] and the SUMO [36]. The CTM is a macroscopic traffic flow model, which evolves according to the conservation law and the flow-density relation (also called "fundamental diagram" by transportation researchers [35]). The CTM accounts for the spatial distribution of traffic and the detailed flow-density relation, which are not captured by the fluid model. The SUMO is a microscopic simulation model, which evolves according to vehicle-following and lane-changing behavior models for individual drivers. Such microscopic details are captured by neither the CTM nor the fluid model.

Fig. 9 shows the simulation results with parameters in Table II. The mainline demand is 2500 veh/h, and the off-ramp demand is 1400 veh/h. The theoretical optimal headway W^{hr} (36 s) is close to the simulation-obtained value (30 s). As expected, the theoretical optimum W^{hr} is greater than the simulated one. The main reason is that in the fluid model decelerating a platoon only affects the platoon itself and does not directly impact the neighboring traffic. In both the CTM and the SUMO, however, a decelerated platoon will induce local congestion. Hence, both CTM and SUMO prefer less deceleration than the fluid model.

Next, we provide more details and discussion of the simulations.

1) Macroscopic Simulation (CTM): We consider the CTM [35] for the highway section in Fig. 2(a). In particular, we consider CAVs and non-CAVs as multiple traffic classes in the CTM. More details of the multiclass CTM is available in [46].

The parameters of the macroscopic model are chosen in accordance with the ones given in Table II, i.e., the capacity of the bottleneck will be set to F–R, and platoon and buffer lengths chosen appropriately. Note that in order for platoons to be properly represented in this framework, we need the spatial and temporal discretization steps to be fairly short, with the platoon length spanning at least two cells. In this article, the physical platoon length is taken to be 0.1 mi, or 10 cells.

Fig. 10(a) shows the simulated traffic evolution without interplatoon coordination using color-coded traffic density. The location of the bottleneck is shown in dashed red line, and the location of the off-ramp is outlined in dashed black line. The streaks of brighter color represent the increased traffic density near the moving platoons. The congestion from the bottleneck propagates upstream and the off-ramp cell becomes congested approximately 1 h into the simulation. Because of the congestion, the off-ramp is partially blocked, preventing vehicles from exiting the highway and further degrading the traffic situation. Although the total demand a is lower than the capacity of the bottleneck, randomness of the platoon arrivals may still create local congestion that disrupt the traffic flow. Such disruptions produce congestion at the bottleneck and block the off-ramp (i.e., the bright-color area in the interval indicated by black dashed lines near distance $x = 24.4$ mil), as shown in Fig. 10(a) from $t \approx 1.25$ h to $t \approx 1.5$ h.

We apply the recommended speed (27) to coordinate platoons in the CTM simulation. Fig. 10(b) shows the traffic evolution under headway regulation of platoons. The simulation run considers the same situation as in the uncoordinated case [see Fig. 10(a)]; whereas in the uncoordinated case, the congestion from the bottleneck blocked the off-ramp, in the coordinated

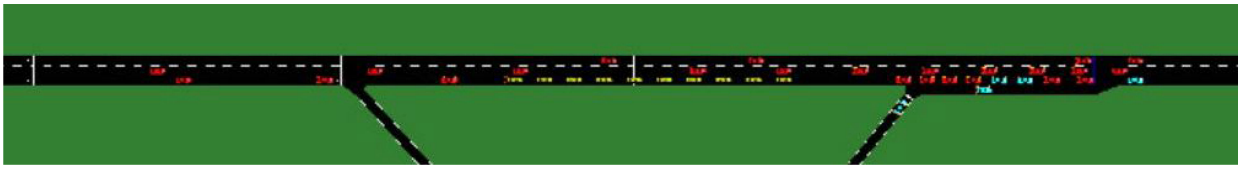


Fig. 11. Microsimulation environment. Red, cyan, and yellow vehicles represent mainline traffic, on-ramp/off-ramp traffic, and CAV platoons, respectively.

case, we are able to spread the arrival of platoons more evenly, thus avoiding causing spillback. Fig. 9 further illustrates how close the theoretical optimal strategy (36 s) is closed to the simulated one (30 s).

2) Microscopic Simulation (SUMO): We also implement the headway regulation strategy introduced in Section V-A in a microsimulation model (see Fig. 11). We use the traffic control interface (TraCI) to customize the simulation and realize the functions required for this specific experiment. The TraCI features 13 individual modules varying from simulation, vehicle type, vehicle. We use Python to code the route and runner files. Some variables that we control in particular include platoon speed, total simulation time, platooning ratio, platoon length, and platooning state. We also customized the lane-changing function to prevent platoons from breaking apart at the bottleneck. The coordination instructions are realized by the runner script.

The simulation results are shown in Fig. 9. The simulations lead to a simulation-optimal interplatoon headway (30 s), which is close to the value W^{hr} given by the fluid model and computed via (22) (36 s). A prominent pattern to be noted in Fig. 9 is that as the coordinated headway increases, the travel time in SUMO grows more slowly than the vehicle hours traveled (VHT) in the CTM. A reasonable explanation is that the CTM that we simulated assumes the first-come-first-serve principle: if a platoon is decelerated, then all the traffic behind will be simultaneously decelerated. However, overtaking is allowed in SUMO, which is implemented according to SUMO's internal overtaking algorithm. Consequently, the impact of decelerating platoons is less significant in SUMO than in the CTM.

C. Further Discussion

The stochastic fluid model that we consider focuses on the following:

1) the capacity sharing between platoons and background traffic;
2) the throughput gain due to reduced intervehicle spacing in platoons;
3) the impact of congestion propagation (spillback).

From a practical perspective, our model is based on the following simplifications.

1) The interaction between platoons and background traffic only occurs at the link boundaries. This is actually a characteristic of any queuing model. Consequently, our model does not account for interactions occurring over a distance, such as the impact of speed difference between platoons and background traffic.

TABLE III
IMPROVEMENT ATTAINED BY VARIOUS COORDINATION STRATEGIES

Strategy	VHT, CTM	Improvement (VHT)	Traverse time, SUMO [min]	Improvement [min]
No coordination	3471	0	36.26	0
Theoretically optimal	3466	5.3	33.27	2.99
Simulation optimal	3465	6.2	32.58	3.68

2) At the interface between links 1 and 2, mainline traffic (i.e., demands a and b) are prioritized for discharging. This is reflected in the definition of the off-ramp flow (3). Alternative models for discharging exist [47], which can be incorporated in our model as well.
3) Each link's capacity is independent of speed difference between CAVs and non-CAVs. In practice, headway regulation modifies CAVs' speeds. Consequently, highway capacity is dependent on the speed difference between CAVs and non-CAVs as well as the traffic mixture, i.e., the percentage of CAVs. In our model, the impact of heterogeneous speed can be modeled as state-dependent link capacities.

Note that the CTM and SUMO simulations that we conducted are not restricted by the above simplifications. The simulation results indicate that the fluid-model approach is adequate in spite of the above simplifications. As Table III shows, compared to the baseline scenario without headway regulation, the theoretically optimal strategy attains 85% (resp. 81%) of the improvement attained by the simulation optimum in CTM (resp. SUMO).

VI. CONCLUSION

In this article, we develop a stochastic fluid model for analysis and design of platoon coordination strategies. The model focuses on the interaction between CAV platoons and non-CAVs, the impact of key platooning parameters (including platooning ratio and platoon size), and road geometry (buffer space). Based on the theory of Markov processes and queuing properties of the model, we derive theoretical bounds for the throughput of the model and identify a set of coordination strategies that maximize throughput as well as minimize delay. We discuss how such strategies can be implemented in practice and validate the fluid model-based results using standard macroscopic and microscopic simulation environments. Our results are useful for link-level coordination of CAV platoons.

This article can be extended in the following directions:

1) evaluation of macroscopic impact due to various vehicle-level controllers;
2) extension to multiple-section highways;
3) integration with network-level scheduling and routing of CAVs.

APPENDIX 1
PROOF OF THEOREM 1

The stability criterion is obtained by showing that the Lyapunov function

$$V(q) := \frac{1}{2}(q_0^m + q_1^m + q_2^m)^2 + \left(k\left(q_0^m + q_1^m + \frac{1}{2\Theta}(q_2^m)^2\right) + q_1^o\right)q_1^o \tag{28}$$

satisfies the Foster–Lyapunov criterion if (10a)–(10b) hold. Specifically, we show that there exist $k > 0$, $c > 0$, and $d < \infty$ such that

$$\mathscr{L}V(q) \le -c|q| + d \quad \forall q \in \mathcal{M} \tag{29}$$

which implies stability via the Foster–Lyapunov criterion. To proceed, we decompose V into

$$V^m(q) := \frac{1}{2}(q_0^m + q_1^m + q_2^m)^2 \tag{30a}$$

$$V^o(q) := \left(k\left(q_1^m + \frac{1}{2\Theta}(q_2^m)^2\right) + q_1^o\right)q_1^o \tag{30b}$$

and show that there exists $c^m, c^o > 0$ and $d^m, d^o < \infty$ such that for all $q \in \mathcal{M}_{\mu,\nu}$

$$\mathscr{L}V^m(q) \le -c^m(q_0^m + q_1^m + q_2^m) + d^m \tag{31a}$$

$$\mathscr{L}V^o(q) \le -c^o q_1^o + d^o. \tag{31b}$$

Note that the above implies (29) and hence stability. The rest of this section is devoted to the proof of (31a)–(31b).

1) Proof of (31a): For q such that $q_2^m = 0$, we have

$$\mathscr{L}V^m(q) = \frac{1}{2}(\lambda l/\gamma)^2 = \frac{\eta\rho a}{2\gamma}.$$

For q such that $q_2^m > 0$, we have

$$\begin{aligned}\mathscr{L}V^m(q) &= ((1-\eta)\rho a + \lambda l/\gamma - (F-R))(q_0^m + q_1^m + q_2^m)\\ &= ((\eta/\gamma + 1 - \eta)\rho a - (F-R))(q_0^m +{}_1^m + q_2^m).\end{aligned}$$

Hence, we have

$$c^m = (F-R) - (\eta/\gamma + 1 - \eta)\rho a \overset{(10a)}{>} 0$$

$$d^m = \frac{\eta\rho a}{2\gamma} < \infty$$

that satisfy (31a).

2) Proof of (31b): For q such that $q_1^o = 0$, we have

$$\mathscr{L}V^o(q) = 0. \tag{32}$$

For q such that $q_1^o > 0$, we need to consider the following cases:

1) $q_0^m = q_1^m = 0$. In this case, we require $c^o > 0$ such that

$$\begin{aligned}&\mathscr{L}V^o(q)\\ &= \Big(k(q_2^m/\Theta)((1-\eta)\rho a - (F-R)) + \lambda k\,(S_0^m(q;\nu(q))\\ &\quad + S_1^m(q;\nu(q)) + \frac{1}{2\Theta}(S_2^m(q;\nu(q)))^2 - \frac{1}{2\Theta}(q_2^m)^2\Big)\\ &\quad + (1-\rho)a - R\Big)q_1^o\\ &\le -c^o q_1^o \quad \forall q : q_1^m = 0, q_1^o > 0.\end{aligned} \tag{33}$$

By Assumption 2, we have

$$\begin{aligned}&S_0^m(q;\nu(q)) + S_1^m(q;\nu(q)) + \frac{1}{2\Theta}(S_2^m(q;\nu(q)))^2\\ &- \frac{1}{2\Theta}(q_2^m)^2 \le S_0^m(q;\nu(q)) + S_1^m(q;\nu(q))\\ &+ \frac{1}{2\Theta}(S_2^m(q;\nu(q)))^2 - \frac{1}{2\Theta}(q_2^m)^2\Big|_{q_1^o=0} \quad \forall q \in \mathcal{Q}.\end{aligned}$$

Hence, we require $k > 0$ such that

$$\begin{aligned}&(q_2^m/\Theta)((1-\eta)\rho a - (F-R)) + \lambda\,(S_0^m(q;\nu(q))\\ &+ S_1^m(q;\nu(q)) + \frac{1}{2\Theta}(S_2^m(q;\nu(q)))^2 - \frac{1}{2\Theta}(q_2^m)^2\Big|_{q_1^o=0}\Big)\\ &< \frac{R-(1-\rho)a}{k} \quad \forall q : q_0^m = q_1^m = 0, q_1^o > 0.\end{aligned} \tag{34}$$

2) $q_0^m > 0$ or $q_1^m > 0$, $q_2^m < \Theta$. In this case, we require $c^0 > 0$ such that

$$\begin{aligned}&\mathscr{L}V^o(q) = (k\,((1-\eta)\rho a - F + (q_2^m/\Theta)(F\\ &\quad -(F-R))) + k\lambda(l/\gamma)) + (1-\rho)a - R\Big)q_1^o\\ &= \left(k\left((\frac{\eta}{\gamma}+1-\eta)\rho a - (F-R)\right) + (1-\rho)a - R\right)q_1^o\\ &\le -c_1^{oq} o \quad \forall q : q_0^m > 0 \text{ or } q_1^m > 0, q_1^o < \Theta.\end{aligned} \tag{35}$$

For all $k > 0$, the existence of such c^o is ensured by (10a).

3) $q_0^{m>0}$ or $q_1^{m>0,}$ $q_2^m = \Theta$. In this case

$$\mathscr{L}V^o(q) = \Big(k\,((\eta/\gamma + 1 - \eta)\rho a - (F-R)) + (1-\rho)a\Big)q_1^o \le -c^o q_1^o. \tag{36}$$

We require $k > 0$ such that

$$k\,((F-R) - (\eta/\gamma + 1 - \eta)\rho a) > (1-\rho)a. \tag{37}$$

Note that (10b) ensures the existence of $k > 0$ that simultaneously satisfies (34) and (37). Hence, there exists $k > 0$ and $c^o > 0$ such that satisfying (33), (35), and (36), which, together with (32), imply (31b).

APPENDIX 2
PROOF OF THEOREM 2

1) Lower Bound: The lower bound is the minimum of the following two terms:

$$\underline{a}_1 := \frac{F-R}{\rho(\eta/\gamma+1-\eta)}$$

$$\underline{a}_2 := \frac{R}{1-\rho+\frac{1}{2}\left(\sqrt{\zeta^2+\frac{2\rho Rl}{\gamma\Theta(F-R)}}-\zeta\right)}$$

where ζ is given by (16). We prove the lower bound by applying Theorem 1, i.e., verifying that (10a)-(10b) hold if $a < \min\{\underline{a}_1, \underline{a}_2\}$.

1) Since $2\rho Rl/(\gamma\Theta(F–R)) > 0$, we have $\underline{a}_2 \le R/(1-\rho)$. Hence, $a < \underline{a}_2$ and $a < \underline{a}_1$ ensure that (10a) holds.
2) For $\xi \in \mathcal{Q}$ such that $\xi_0^m = \xi_1^m = 0$

$$\begin{aligned}&\frac{\xi_2^m}{\Theta}\left((1-\eta)\rho a-(F-R)\right)+\frac{\eta\rho a}{l}\Big(S_0^m(\xi;0)\\&\quad+S_1^m(\xi;0)+\frac{1}{2\Theta}\left(S_2^m(\xi;0)\right)^2-\frac{1}{2\Theta}(\xi_2^m)^2\Big)\\&=\left(\frac{\xi_2^m}{\Theta}((1-\eta)\rho a-(F-R)\right)\\&\quad+\frac{\lambda}{2\Theta}(2\xi_2^m l/\gamma+(l/\gamma)^2).\end{aligned}$$

Since $a < \underline{a}_1$ and since $\xi_2^m \le \Theta$, the above implies that

$$\begin{aligned}&\max_{\substack{\xi\in\mathcal{Q}:\xi_0^m\\=\xi_1^m=\xi_1^o=0}}\frac{\xi_2^m}{\Theta}\left((1-\eta)\rho a-(F-R)\right)+\frac{\eta\rho a}{l}\Big(S_0^m(\xi;0)\\&\quad+S_1^m(\xi;0)+\frac{1}{2\Theta}\left(S_2^m(\xi;0)\right)^2-\frac{1}{2\Theta}(\xi_2^m)^2\Big)\\&\le\left(\frac{\lambda k}{2\Theta}(2\Theta l/\gamma+(l/\gamma)^2)\right).\end{aligned}$$

With the above, one can verify that if $a < \underline{a}_2$, then (10b) holds.

2) Upper Bound: We prove the upper bound by showing that if $\{Q(t); t > 0\}$ is stable, then $a \le \min\{\bar{a}_1, \bar{a}_2\}$, where

$$\bar{a}_1 := \frac{F-R}{\rho(\eta/\gamma+1-\eta)}$$

$$\bar{a}_2 := \frac{(1-\omega)R}{1-\rho}.$$

1) $a \le \bar{a}_1$ can be obtained from the nominal throughput given by Lemma 1.
2) To show $a \le a_2$, note that when $a < \bar{a}_1$, the M/D/1 process $\{N(t); t > 0\}$ is stable and admits a steady-state distribution $\{\pi_n; n = 0, 1, \dots,\}$ defined in (13). Hence, there exists ω_0 and ω such that

$$\lim_{t\to\infty}\frac{1}{t}\int_{\tau=0}^{t}\mathbb{I}_{Q_1^m>0,Q_2^m=\Theta}d\tau=\omega_0 \quad \text{a.s.}$$

$$\lim_{t\to\infty}\frac{1}{t}\int_{\tau=0}^{t}\mathbb{I}_{N(t)\ge/ceil\gamma\Theta/l}d\tau=\omega \quad \text{a.s.}$$

where ω is in fact given by (14).
Next, consider the set $\mathcal{M}_{0,0} \subset \mathcal{Q}$ defined by

$$\mathcal{M}_{0,0}=((\{0\}\times[0,\Theta])\cup((0,\infty)\times\{\Theta\}))\times[0,\infty). \tag{38}$$

One can show that $\mathcal{M}_{0,0}$ is an invariant set. Hence, for each initial condition $q \in \mathcal{M}_{\mu,\nu}$, we have $Q(t) \in \mathcal{M}_{\mu,\nu}$ for all $t > 0$. Thus, we have $Q_2^m(t) = \Theta$ if $Q_1^m(t) > 0$ for sufficiently large t. Hence, if $Q_1^m(t) + Q_2^m(t) > \Theta$, i.e., if $Q_1^m(t) > 0$ and $Q_2^m(t) = \Theta$, then $N(t) \ge /ceil\gamma\Theta/l$. Therefore, we have $\omega_0 \ge \omega$. Finally, note that if $|Q(t)|$ is bounded, then

$$\begin{aligned}(1-\rho)a&\le\lim_{t\to\infty}\frac{1}{t}\int_{\tau=0}^{t}r(Q(\tau);0)d\tau\\&=\lim_{t\to\infty}\frac{1}{t}\left(\int_{\tau:Q_1^m(\tau)+Q_1^m(\tau)\le\Theta}^{t}r(Q(\tau);0)d\tau\right.\\&\quad\left.+\int_{\tau:Q_1^m(\tau)+Q_1^m(\tau)>\Theta}^{t}r(Q(\tau);0)d\tau\right)\\&\overset{a.s.}{=}(1-\omega_0)R\le(1-\omega)R=\bar{a}_2.\end{aligned}$$

APPENDIX 3
PROOF OF THEOREM 3

1) Stability: The necessity of (18) results from Lemma 1. To show the sufficiency, one can indeed use Theorem 1 to show that the model is stabilized by $(\mu, \nu) \in \mathscr{U}^* \times \mathscr{V}^*$ if $a < a^*$ in the sense of a bounded 1-norm. In this section, we use an alternative Lyapunov function

$$\tilde{V}(q)=e^{\beta|q|},\quad q\in\mathcal{Q} \tag{39}$$

and obtain a stronger stability

$$\limsup_{t\to\infty}\frac{1}{t}\int_{\tau=0}^{t}\mathrm{E}[e^{\beta|Q(\tau)|}]d\tau\le Z. \tag{40}$$

That is, the state is bounded in the moment-generating function (MGF).

To proceed, define

$$\begin{aligned}\zeta_{\mu,\nu}:=\max\Big\{&\max_{q\in\mathcal{Q}}S_2^m(q;\nu(q)),\ \sup\{\zeta\ge0:\\&(\forall q:q_2^m=\zeta)\ \mu(q)\ge F-R-(1-\eta)\rho a\}\Big\}.\end{aligned} \tag{41}$$

We know from (17a)–(17d) that $\zeta_{\mu,\nu} < \Theta$ for all $(\mu, \nu) \in \mathscr{U}^* \times \mathscr{V}^*$. Then, consider the set

$$\mathcal{M}_{\mu,\nu}=[0,\infty)\times\{0\}^2\times[0,\zeta_{\mu,\nu}]. \tag{42}$$

Lemma 3: The set $\mathcal{M}_{\mu,\nu}$ as defined in (42) is an invariant set.
Proof.

1) For all $q \in \mathcal{M}_{\mu,\nu}$, we have

$$\begin{aligned}G_1^m(q;\mu(q))&=\mu(q)+((1-\eta)\rho+(1-\rho))a\\&\quad-\min\left\{((1-\eta)\rho+(1-\rho))\,a+\mu(q),F\right\}\\&\overset{(17b)}{=}0.\end{aligned}$$

Hence, $\{0\}$ is invariant for q_1^m. In addition, for $q \in \mathcal{M}_{\mu,\nu}$ such that $q_2^m = \zeta$, we have

$$\begin{aligned} G_2^m(q;\mu(q)) &= f_1(q;\mu(q)) - f_2(q;\mu(q)) \\ &\overset{(41)}{<} (\mu(q) + (1-\eta)\rho a) \\ &\quad - (\mu(q) + (1-\eta)\rho a) < 0. \end{aligned}$$

Hence, $[0, \zeta_{\mu,\nu}]$ is invariant for q_2^m.

2) By (17d), since $S_1^m(q;\nu(q)) = 0$ for all $q \in \mathcal{Q}$, $\{0\}$ is invariant for q_1^m. By (41), $S_2^m(q;\nu(q)) \le \zeta$ for all $q \in \mathcal{Q}$. Hence, $[0, \zeta]$ is invariant for q_2^m.□

In the rest of this proof, we show that there exists $\beta > 0$, $c > 0$, and $\tilde{d} < \infty$ verifying the drift condition

$$\mathcal{L}\tilde{V}(q) \le -ce^{\beta|q|} + \tilde{d} \quad \forall q \in \mathcal{M}_{\mu,\nu} \tag{43}$$

which leads to (40) by the Foster–Lyapunov criterion.

We partition the invariant set $\mathcal{M}_{\mu,\nu}$ into two subsets

$$\begin{aligned} \mathcal{M}_{\mu,\nu}^0 &= \{0\}^3 \times [0, \zeta_{\mu,\nu}] \\ \mathcal{M}_{\mu,\nu}^1 &= (0,\infty) \times \{0\}^2 \times [0, \zeta_{\mu,\nu}] \end{aligned}$$

For $q \in \mathcal{M}_{\mu,\nu}^0$, we have

$$\begin{aligned} \mathcal{L}\tilde{V}(q) &= \nabla_q e^{\beta|q|} G(q;\mu(q)) + \lambda(e^{\beta|S(q;\nu(q))|} - e^{\beta|q|}) \\ &= (\beta((1-\eta)\rho a - \tilde{f}_2(q;\mu(q))) + \lambda(e^{\beta l/\gamma} - 1))e^{\beta|q|} \\ &\le \lambda(e^{\beta l/\gamma} - 1)e^{\beta l/\gamma} = \tilde{d}^* \end{aligned}$$

which also defines $\tilde{d}^*$. For $q \in \mathcal{M}_{\mu,\nu}^1$, we have

$$\begin{aligned} \mathcal{L}\tilde{V}(q) &= (\beta((1-\eta)\rho a - (F-R)) + \lambda(e^{\beta l/\gamma} - 1))e^{\beta|q|} \\ &= \phi(\beta)e^{\beta|q|} \end{aligned}$$

where the definition of the function ϕ is clear. Since

$$\phi(0) = 0, \ \frac{d}{d\beta}\phi(\beta)\Big|_{\beta=0} = (1-\eta)\rho a - (F-R) + \eta\rho a/\gamma$$

there exists $\beta^* > 0$ such that $\phi(\beta^*) < 0$ if $a < a^*$.

In conclusion, there exist $\beta = \beta^*$, $c = \phi(\beta^*)$, and $\tilde{d} = \tilde{d}^*$ that verify (43). Then, by the Foster–Lyapunov stability criterion, we conclude (40).

2) Queue Minimization: Next, we use a sample path-based method to show that any $(\mu,\nu) \in \mathscr{U}^* \times \mathscr{V}^*$ minimizes the total queue size $|Q(t)|$ for all t over all control policies $(\mu,\nu) \in \mathscr{U} \times \mathscr{V}$.

Let $\{M(t); t > 0\}$ be the counting process of platoon arrivals. For a given sample path $\{m(t); t > 0\}$ of the counting process and a given initial condition $q \in \mathcal{Q}$, let $\{q(t); t > 0\}$ and $\{\psi(t); t > 0\}$ be the corresponding trajectories under a control policy $(\mu^*,\nu^*) \in \mathscr{U}^* \times \mathscr{V}^*$ and under a control policy $(\mu,\nu) \in \mathscr{U} \times \mathscr{V}$, respectively. To show the optimality of (μ^*,ν^*), it suffices to show $|q(t)| \le |\psi(t)|$ for any $t \ge 0$. Without loss of generality, we consider zero initial condition. We prove this by contradiction as follows.

Assume by contradiction that there exists $(\mu,\nu) \in \mathscr{U} \times \mathscr{V}$ such that

$$\exists t_1 > 0, \ |q(t_1)| > |\psi(t_1)|. \tag{44}$$

Between resets, the continuity of $q(t)$ and $\psi(t)$ follows from Assumption 1. Therefore, there must exist a "crossing time" $t_0 \in [0, t)$ such that

$$|q(t_0)| = |\psi(t_0)|, \ \frac{d}{dt}|q(t)| > \frac{d}{dt}|\psi(t)|. \tag{45}$$

Note that the "crossing" must happen between resets. To see this, recall that Assumption 2 ensures that if $|q(t_-)| = |\psi(t_-)|$ and if a reset occurs at time t, then

$$|q(t)| = |\psi(t)| = |q(t_-)| + l/\gamma.$$

Since the system admits the invariant set $\mathcal{M}_{\mu^*,\nu^*}$ as given in (42) under (μ^*,ν^*), we have $q_1^m(t) = q_1^o(t) = 0$ for all $t \ge 0$. Hence, a necessary condition for (45) is that

$$q_0^m(t_0) + q_2^m(t_0) \ge \psi_0^m(t_0) + \psi_1^m(t_0) + \psi_2^m(t_0) \tag{46a}$$

$$\begin{aligned} &G_0^m(q(t_0);\mu^*) + G_2^m(q(t_0);\mu^*) \\ &> G_0^m(\psi(t_0);\mu) + G_1^m(\psi(t_0);\mu) + G_2^m(\psi(t_0);\mu)). \end{aligned} \tag{46b}$$

However, one can obtain from (4a)–(4d) that if $q_0^m(t_0) + q_2^m(t_0) \ge \psi_0^m(t_0) + \psi_1^m(t_0) + \psi_2^m(t_0) = 0$, then

$$\begin{aligned} &G_0^m(q(t_0);\mu^*) + G_2^m(q(t_0);\mu^*) \\ &= G_0^m(\psi(t_0);\mu) + G_1^m(\psi(t_0);\mu) + G_2^m(\psi(t_0);\mu)). \end{aligned} \tag{47}$$

If $q_0^m(t_0) + q_2^m(t_0) \ge \psi_0^m(t_0) + \psi_1^m(t_0) + \psi_2^m(t_0) > 0$, then

$$G_0^m(q(t_0);\mu^*) + G_2^m(q(t_0);\mu^*) = (1-\eta)\rho a - (F-R) \tag{48a}$$

$$\begin{aligned} &G_0^m(\psi(t_0);\mu) + G_1^m(\psi(t_0);\mu) + G_2^m(\psi(t_0);\mu)) \\ &= (1-\eta)\rho a - r(\psi(t_0)) \ge (1-\eta)\rho a - (F-R). \end{aligned} \tag{48b}$$

Since both (47) and (48a)-(48b) contradict with (46b), we conclude that (μ,ν) cannot achieve (44). That is, if $|q(t_1)| = |\psi(t_1)|$, then $|q(t)|$ cannot increase faster (or decrease slower) than $|\psi(t)|$ at time $t = t_1$. This proves the optimality of (μ^*,ν^*).

3) Mean Queue Size: To show that the 1-norm of the state converges as in (19), we first need to show that the process $\{Q(t); t \ge 0\}$ is ergodic. Formally, let $P_t(q)$ be the distribution of $Q(t)$ given the initial condition $Q(0) = q$ for $q \in \mathcal{Q}$. Then, there exists a unique probability measure P^* on $\mathcal{Q}$ such that

$$\lim_{t\to\infty} \|P_t(q) - P^*\|_{\mathrm{TV}} = 0 \tag{49}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total-variation distance between two probability measures [37]. Ergodicity ensures convergence of the time average toward the expected value, if the expected value exists.

Convergence. The Foster–Lyapunov criterion ensures the existence of an invariant measure P^* [43, Th. 4.5]. We only need to further show that the invariant measure is unique. This can be shown via the "coupling" condition given as follows.

Coupling Condition [43]: Let $q, q' \in \mathcal{Q}$ be two initial conditions and $Q(t)$ and $Q'(t)$ be the trajectories starting therefrom. Then, there exists $\delta > 0$ and $T < \infty$ such that

$$\Pr\{Q(T) = Q'(t) | Q(0) = q, Q'(0) = q'\} = \delta. \tag{50}$$

To show that the stochastic fluid model controlled by $(\mu,\nu) \in \mathscr{U}^* \times \mathscr{V}^*$ satisfies the coupling condition, note that

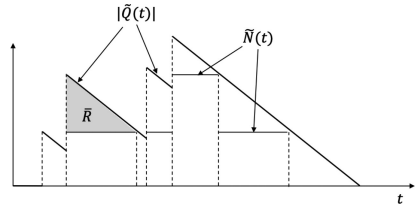


Fig. 12. Controlled fluid process $Q(t)$ envelops an M/D/1 process $\tilde{N}(t)$.

for an arbitrary initial condition $q \in \mathcal{Q}$, there exists $T = (\sum_{k=1}^{3} q_k^m)/(F-R) + q_1^o/R$ such that

$$\Pr\{Q(T) = 0 | Q(0) = q\} \geq e^{-\lambda T} > 0.$$

Then, by [28, Th. 6.1], the above, together with (43), implies convergence in the sense of (49).

Remark 1: In fact, the above argument ensures exponentially convergent [43] in the sense that there exist a constant $\kappa > 0$ and a finite-valued function $U : \tilde{\mathcal{Q}} \to \mathbb{R}_{\geq 0}$ such that

$$\|P_t(q) - P^*\|_{\mathrm{TV}} \leq U(q) e^{-\kappa t} \quad \forall t \geq 0.$$

Queuing Delay: The evolution of the total queue length $|Q(t)|$ can be viewed as the superposition of two subprocesses (see Fig. 12).

1) The first process is an M/D/1 process $\{N(t); t \geq 0\}$ defined by (11). By the Pollazcek–Khinchin formula [48, p. 248] and the Little's theorem [48, Th. 5.5.9], the mean number of waiting jobs (excluding the one being served) of this process is

$$\bar{N} = \frac{\lambda^2 s^2}{2(1-\lambda s)} = \frac{\eta^2 \rho^2 a^2/(2\gamma^2)}{(F - R - (1-\eta)\rho a)(F - R - (\eta/\gamma + 1 - \eta)\rho a)}.$$

2) The second process is the "services" each job experiences. As Fig. 12 shows, the cumulative queuing delay during services is given by

$$\bar{R} = \frac{l^2}{2\gamma^2(F - R - (1-\eta)\rho a)}.$$

Thus, the total fluid queue length is

$$\bar{Q} = \frac{l}{\gamma}\bar{N} + \frac{\eta \rho a}{l}\bar{R}$$

which leads to (19).

Acknowledgment

X. Xiong and T. Wang at NYU Tandon School of Engineering contributed to the simulation. The authors greatly appreciate the feedback from the anonymous reviewers and from the Associate Editor.

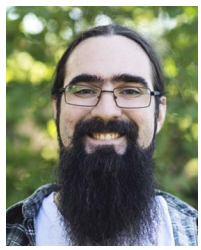
References

[1] R. Horowitz and P. Varaiya, "Control design of an automated highway system," *Proc. IEEE*, vol. 88, no. 7, pp. 913–925, Jul. 2000.
[2] T. Litman, *Autonomous Vehicle Implementation Predictions*. Victoria, BC, Canada: Victoria Transp. Policy Inst., 2017.
[3] W. Levine and M. Athans, "On the optimal error regulation of a string of moving vehicles," *IEEE Trans. Autom. Control*, vol. 11, no. 3, pp. 355–361, Jul. 1966.
[4] D. Swaroop and J. K. Hedrick, "String stability of interconnected systems," *IEEE Trans. Autom. Control*, vol. 41, no. 3, pp. 349–357, Mar. 1996.
[5] K. S. Chang, K. J. Hedrick, W.-B. Zhang, P. Varaiya, M. Tomizuka, and S. E. Shladover, "Automated highway system experiments in the path program," *J. Intell. Transp. Syst.*, vol. 1, no. 1, pp. 63–87, 1993.
[6] G. J. L. Naus, R. P. A. Vugts, J. Ploeg, M. J. G. van de Molengraft, and M. Steinbuch, "String-stable CACC design and experimental validation: A frequency-domain approach," *IEEE Trans. Veh. Technol.*, vol. 59, no. 9, pp. 4268–4279, Nov. 2010.
[7] B. Besselink *et al.*, "Cyber-physical control of road freight transport," *Proc. IEEE*, vol. 104, no. 5, pp. 1128–1141, May 2016.
[8] S. Tsugawa, S. Jeschke, and S. E. Shladover, "A review of truck platooning projects for energy savings," *IEEE Trans. Intell. Veh.*, vol. 1, no. 1, pp. 68–77, Mar. 2016.
[9] S. Coogan and M. Arcak, "A dissipativity approach to safety verification for interconnected systems," *IEEE Trans. Autom. Control*, vol. 60, no. 6, pp. 1722–1727, Jun. 2015.
[10] A. Talebpour and H. S. Mahmassani, "Influence of connected and autonomous vehicles on traffic flow stability and throughput," *Transp. Res. C, Emerg. Technol.*, vol. 71, pp. 143–163, 2016.
[11] A. Duret, M. Wang, and A. Ladino, "A hierarchical approach for splitting truck platoons near network discontinuities," *Transp. Res. B, Methodological*, vol. 132, pp. 285–302, 2020.
[12] G. F. Newell, *Applications of Queueing Theory*, vol. 4. New York, NY, USA: Springer, 2013.
[13] V. G. Kulkarni, "Fluid models for single buffer systems," in *Frontiers in Queueing: Models and Applications in Science and Engineering*, vol. 321. Boca Raton, FL, USA: CRC Press, 1997, pp. 321–338.
[14] S. E. Li *et al.*, "Dynamical modeling and distributed control of connected and automated vehicles: Challenges and opportunities," *IEEE Intell. Transp. Syst. Mag.*, vol. 9, no. 3, pp. 46–58, Fall 2017.
[15] S. W. Smith, Y. Kim, J. Guanetti, A. A. Kurzhanskiy, M. Arcak, and F. Borrelli, "Balancing safety and traffic throughput in cooperative vehicle platooning," in *Proc. 18th Eur. Control Conf.*, 2019, pp. 2197–2202.
[16] Y. Zhou, S. Ahn, M. Wang, and S. Hoogendoorn, "Stabilizing mixed vehicular platoons with connected automated vehicles: An H-infinity approach," in *Proc. 23rd Int. Symp. Transp. Traffic Theory*, 2019, pp. 441–461.
[17] J. Lioris, R. Pedarsani, F. Y. Tascikaraoglu, and P. Varaiya, "Platoons of connected vehicles can double throughput in urban roads," *Transp. Res. C, Emerg. Technol.*, vol. 77, pp. 292–305, 2017.
[18] D. Miculescu and S. Karaman, "Polling-systems-based autonomous vehicle coordination in traffic intersections with no traffic signals," *IEEE Trans. Autom. Control*, vol. 65, no. 2, pp. 680-694, Feb. 2020.
[19] A. Keimer, *et al.*, "Integration of information patterns in the modeling and design of mobility management services," 2017, *arXiv:1707.07371*.
[20] M. Čičić, L. Jin, and K. H. Johansson, "Coordinating vehicle platoons for highway bottleneck decongestion and throughput improvement," in *Proc. Transp. Res. Board 99th Annu. Meet.* (No. 20-01387), 2020.
[21] D. A. Lazar, S. Coogan, and R. Pedarsani, "Routing for traffic networks with mixed autonomy," 2018, *arXiv:1809.01283*.
[22] N. Mehr and R. Horowitz, "Can the presence of autonomous vehicles worsen the equilibrium state of traffic networks?," in *Proc. IEEE 57th Annu. Conf.* Decis. Control, 2018, pp. 1788–1793.
[23] J. G. Dai and S. P. Meyn, "Stability and convergence of moments for multiclass queueing networks via fluid limit models," *IEEE Trans. Autom. Control*, vol. 40, no. 11, pp. 1889–1904, Nov. 1995.
[24] H. Yu and M. Krstic, "Traffic congestion control for Aw–Rascle–Zhang model," *Automatica*, vol. 100, pp. 38–51, 2019.
[25] M. Čičić and K. H. Johansson, "Traffic regulation via individually controlled automated vehicles: A cell transmission model approach," in *Proc. 21st Int. Conf. Intell. Transp. Syst.*, 2018, pp. 766–771.
[26] Z. S. Qian, W. Shen, and H. Zhang, "System-optimal dynamic traffic assignment with and without queue spillback: Its path-based formulation and solution via approximate path marginal cost," *Transp. Res. B, Methodological*, vol. 46, no. 7, pp. 874–893, 2012.
[27] S. van de Hoef, K. H. Johansson, and D. V. Dimarogonas, "Fuel-efficient en route formation of truck platoons," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 1, pp. 102–112, Jan. 2018.
[28] S. P. Meyn and R. L. Tweedie, "Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes," *Adv. Appl. Probab.*, vol. 25, no. 3, pp. 518–548, 1993.
[29] D. Mitra, "Stochastic theory of a fluid model of producers and consumers coupled by a buffer," *Adv. Appl. Probab.*, vol. 20, no. 3, pp. 646–676, 1988.

[30] C. G. Cassandras, Y. Wardi, B. Melamed, G. Sun, and C. G. Panayiotou, "Perturbation analysis for online control and optimization of stochastic fluid models," *IEEE Trans. Autom. Control*, vol. 47, no. 8, pp. 1234–1248, Aug. 2002.
[31] D. P. Kroese and W. R. Scheinhardt, "Joint distributions for interacting fluid queues," *Queueing Syst.*, vol. 37, no. 1/3, pp. 99–139, 2001.
[32] L. Jin and S. Amin, "Stability of fluid queueing systems with parallel servers and stochastic capacities," *IEEE Trans. Autom. Control*, vol. 63, no. 11, pp. 3948–3955, Nov. 2018.
[33] R. E. Stern *et al.*, "Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments," *Transp. Res. C, Emerg. Technol.*, vol. 89, pp. 205–221, 2018.
[34] C. Wu, A. M. Bayen, and A. Mehta, "Stabilizing traffic with autonomous vehicles," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2018, pp. 1–7.
[35] C. F. Daganzo, "The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory," *Transp. Res. B, Methodological*, vol. 28, no. 4, pp. 269–287, 1994.
[36] D. Krajzewicz, G. Hertkorn, C. Rössel, and P. Wagner, "SUMO (simulation of urban mobility)—An open-source traffic simulation," in *Proc. 4th Middle East Symp. Simul. Model.*, 2002, pp. 183–187.
[37] M. Benaïm, S. Le Borgne, F. Malrieu, and P.-A. Zitt, "Qualitative properties of certain piecewise deterministic Markov processes," in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 51, no. 3. Paris, France: Institut Henri Poincaré, 2015, pp. 1040–1075.
[38] M. H. A. Davis, "Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models," *J. Roy. Statist. Soc.. B. Methodological*, vol. 46, no. 3, pp. 353–388, 1984.
[39] S. Santini, A. Salvi, A. S. Valente, A. Pescapé, M. Segata, and R. L. Cigno, "A consensus-based approach for platooning with intervehicular communications and its validation in realistic scenarios," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 1985–1999, Mar. 2016.
[40] L. Jin and Y. Wen, "Behavior and management of stochastic multiple-origin-destination traffic flows sharing a common link," in *Proc.IEEE 58th Conf. Decis. Control*, 2019, pp. 4109–4114.
[41] X. Xiong, E. Xiao, and L. Jin, "Analysis of a stochastic model for coordinated platooning of heavy-duty vehicles," in *Proc. IEEE Conf. Decis. Control*, 2019, pp. 3170–3175.
[42] X. Xiong, J. Sha, and L. Jin, "Optimizing coordinated vehicle platooning: An analytical approach based on stochastic dynamic programming," 2020, *arXiv:2003.13067*.
[43] S. P. Meyn and R. L. Tweedie, "Stability of Markovian processes III: Foster–Lyapunov criteria for continuous-time processes," *Adv. Appl. Probab.*, vol. 25, pp. 518–548, 1993.
[44] J. F. Shortle, J. M. Thompson, D. Gross, and C. M. Harris, *Fundamentals of Queueing Theory*, vol. 399. Hoboken, NJ, USA: Wiley, 2018.
[45] P. Papadimitratos, A. De La Fortelle, K. Evenssen, R. Brignolo, and S. Cosenza, "Vehicular communication systems: Enabling technologies, applications, and future outlook on intelligent transportation," *IEEE Commun. Mag.*, vol. 47, no. 11, pp. 84–95, Nov. 2009.
[46] M. Čičić, L. Jin, and K. H. Johansson, "Coordinating vehicle platoons for highway bottleneck decongestion and throughput improvement," 2019, *arXiv:1907.13049*.
[47] M. A. Wright, G. Gomes, R. Horowitz, and A. A. Kurzhanskiy, "On node models for high-dimensional road networks," *Transp. Res. B, Methodological*, vol. 105, pp. 212–234, 2017.
[48] R. G. Gallager, *Stochastic Processes: Theory for Applications*. Cambridge, U.K.: Cambridge Univ. Press, 2013.

Li Jin (Member, IEEE) received the B.Eng. degree in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2011, the M.S. degree in mechanical engineering from Purdue University, West Lafayette, IN, USA, in 2012, and the Ph.D. degree in transportation from the Massachusetts Institute of Technology, Cambridge, MA, USA, in 2018.

He is currently an Assistant Professor with the New York University Tandon School of Engineering, Brooklyn, NY, USA. He was also a Visiting Scholar with the University of Erlangen-Nuremberg in 2016. His background is stochastic processes, dynamic control, and optimization. His research focuses on developing resilient control algorithms for cyber–physical systems with guarantees of efficiency in nominal settings, robustness against random perturbations, and survivability under strategic disruptions. Specific applications of his work include connected and autonomous vehicles, automatic traffic control, air transportation, and gas network.



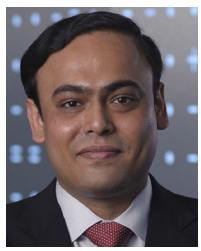
Mladen Čičić received the M.Sc. degree in automatic control from the School of Electrical Engineering, University of Belgrade, Belgrade, Serbia, in 2015. He is currently working toward the Ph.D. degree with the Department of Automatic Control, KTH Royal Institute of Technology, Stockholm, Sweden.

He is a member of Marie Skłodowska Curie oCPS ITN and an affiliated Wallenberg AI, Autonomous Systems and Software Program student. He was also a Visiting Scholar with the C2SMART University Transportation Center, New York University Tandon School of Engineering in 2019. His research interests include traffic control using mixed traffic models.

Karl H. Johansson (Fellow, IEEE) received the M.Sc. and Ph.D. degrees from Lund University, Lund, Sweden.

He is currently a Professor with the School of Electrical Engineering and Computer Science, KTH Royal Institute of Technology, Stockholm, Sweden. He has held visiting positions with UC Berkeley, Caltech, NTU, HKUST Institute of Advanced Studies, and NTNU. His research interests are in networked control systems, cyber–physical systems, and applications in transportation, energy, and automation.

Dr. Johansson is a Fellow of the Royal Swedish Academy of Engineering Sciences and an IEEE Control Systems Society Distinguished Lecturer. He is a member of the Swedish Research Council's Scientific Council for Natural Sciences and Engineering Sciences. He was on the IEEE Control Systems Society Board of Governors and the IFAC Executive Board, and is currently the Vice-President for the European Control Association Council. He was the recipient of several best paper awards and other distinctions from the IEEE, IFAC, and ACM. He has been awarded Distinguished Professor from the Swedish Research Council and Wallenberg Scholar with the Knut and Alice Wallenberg Foundation. He was also the recipient of the Future Research Leader Award from the Swedish Foundation for Strategic Research and the triennial Young Author Prize from IFAC.



Saurabh Amin (Member, IEEE) received the Ph.D. degree in systems engineering from the University of California, Berkeley, Berkeley, CA, USA, in 2011.

He is currently a Robert N. Noyce Career Development Associate Professor with the Department of Civil and Environmental Engineering and a core Faculty Member with the Laboratory of Information and Decision Systems, Massachusetts Institute of Technology, Cambridge, MA, USA. His research focuses on the design and implementation of high-confidence network control algorithms for infrastructure systems. His research interests include robust diagnostics and control problems that involve using networked systems to facilitate the monitoring and control of large-scale critical infrastructures, including transportation, water, and energy distribution systems. He also studies the effect of security attacks and random faults on the survivability of networked systems, and designs incentive-compatible control mechanisms to reduce network risks.

Dr. Amin was the recipient of NSF CAREER and Google Faculty Research awards.