View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by DSpace@MIT

Modeling the Impact of Vehicle Platooning on Highway Congestion: A Fluid Queuing Approach

Li Jin
Massachusetts Institute of Technology
jnl@mit.edu

Mladen Čičić
KTH Royal Institute of Technology
cicic@kth.se

Saurabh Amin
Massachusetts Institute of Technology
amins@mit.edu

Karl H. Johansson
KTH Royal Institute of Technology
kallej@kth.se

ABSTRACT

Vehicle platooning is a promising technology that can lead to significant fuel savings and emission reduction. However, the macroscopic impact of vehicle platoons on highway traffic is not yet well understood. In this article, we propose a new fluid queuing model to study the macroscopic interaction between randomly arriving vehicle platoons and the background traffic at highway bottlenecks. This model, viewed as a stochastic switched system, is analyzed for two practically relevant priority rules: proportional (or mixed) and segmented priority. We provide intuitive stability conditions, and obtain bounds on the long-run average length and variance of queues for both priority rules. We use these results to study how platoon-induced congestion varies with the fraction of platooned vehicles, and their characteristics such as intra-platoon spacing and arrival rate. Our analysis reveals a basic tradeoff between congestion induced by the randomness of platoon arrivals, and efficiency gain due to a tighter intra-platoon spacing. This naturally leads to conditions under which the proportional priority is preferred over segmented priority. Somewhat surprisingly, our analytical results are in agreement with the simulation results based on a more sophisticated two-class cell transmission model.

Keywords

Connected and autonomous vehicles, vehicle platooning, smart highways, fluid queuing model, stochastic switched systems.

1. INTRODUCTION

Platooning of connected vehicles is considered as an effective way of improving traffic throughput [5, 19, 27] and reducing environmental externalities [1, 4, 28]. Although the idea of automatically regulating a string of vehicles is well-known [6, 18], it is only over the last few years that extensive experimental studies in real-world traffic conditions have been conducted [1, 21, 28]. With the rapid advancements in vehicle platooning technology [24], it seems plausible that semi-automated highway systems will be practically viable soon [12]. However, we still lack a realistic and tractable model that captures the macroscopic impact of platooning operations on highway congestion. We posit that a major challenge in integration of vehicle platoons into existing highway systems is our limited understanding of how vehicle platoons interact with (and impact) highway traffic.

In this article, we propose a new fluid queuing model that captures the macroscopic impact of platooning operations on highway congestion. This model belongs to a class of stochastic switched models or piecewise-determinstic Markov processes (PDMP). Our work is complementary to the two lines of existing literature: first, on modeling and control of microscopic platoon behavior [1, 23, 26]; and second, on partial differential equation models of interaction between slowly moving (large) vehicles on background traffic [10, 15]. While the previous work provides a good foundation to study platooning in specific scenarios, it does not naturally lead to a tractable way to design efficient network-level operations. Our model captures the macroscopic interaction between platoons of connected vehicles and ordinary vehicles, and permits a tractable analysis that can lead to practical insights on the design of platoon operations. We mainly focus on three questions:

1. How to model the sharing of highway capacity between vehicle platoons and the background traffic?
2. How do the key parameters of vehicle platoons, including penetration rate, platoon length, and vehicle spacing within a platoon, affect highway performance?
3. How to evaluate the strategies for allocating road capacity between ordinary vehicles and platoons?

Our model (Section 2) captures the following important features of vehicle platoons. First, vehicle platoons can act as temporary bottlenecks for other vehicles. We use a two-class fluid queuing model to capture the sharing of highway capacity between vehicle platoons and the background traffic. Second, the headways between platoons and the lengths of platoons are subject to random variations. We use a Markov process to capture such randomness. Third, vehicles within a platoon have smaller spacing compared to ordinary vehicles. We scale down the queuing effect due to vehicle platoons according to a pre-defined inter-vehicle spacing ratio for the two traffic classes. Note that our model does not account for (i) the impact of speed difference between platoons and background traffic, (ii) the formation/split of platoons, (iii) the microscopic (vehicle-level) interaction between platoons and background traffic.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
HSCC '18: 21st International Conference on Hybrid Systems: Computation and Control (part of CPS Week), April 11–13, 2018, Porto, Portugal
Copyright 2018 ACM X-XXXXX-XX-X/XX/XX ...$15.00.

Importantly, we view the fluid queuing model as a reduced-order model of the more sophisticated (and rather well-studied) cell-transmission model (CTM, see Section 2.3). Although our model does not capture the spatial propagation of congestion (which CTM does), we find that the estimates of traffic queues for a single highway bottleneck – as obtained by our model – are in general agreement with the simulation results obtained from the CTM (Section 4.1). Thus, an important aspect of our work is the simplicity and analytical tractability of the fluid queuing model for study of platooning operations at individual highway bottlenecks.

Our stability analysis (Section 3) focuses on the queuing resulting the interaction between two classes of traffic. We first provide an intuitive stability result based on the theory of convergence of stochastic fluid queuing systems [14, 20]. We consider the traffic queue to be stable if the time-average of its moment generating function is bounded. Then, based on known results regarding the steady-state distribution of stochastic fluid queuing systems [16], we derive analytical bounds for the average and variance of the queue lengths under proportional priority, and the exact queue lengths under segmented priority.

We also consider the impact of key parameters of vehicle platoons on traffic queue (Section 4). Main insights include: (i) increase of the fraction of connected vehicles typically reduce congestion; however, if the highway is in free flow without platooning, then introduction of platooning may induce congestion due the randomness in platoon arrivals; (ii) short platoons lead to less congestion than long platoons; (iii) prioritizing platoons over background traffic does not necessarily reduces congestion.

2. TRAFFIC MODELS WITH PLATOONS

In this section, we introduce two models for highway traffic with vehicle platooning at highway bottlenecks. We first introduce a stochastic two-class fluid queuing model (FQM), and then an analogous stochastic two-class cell transmission model (CTM).

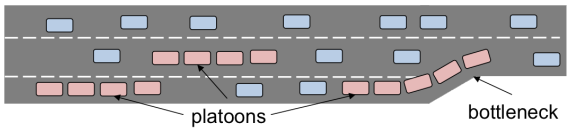


Figure 1: A highway bottleneck.

We focus on the most basic setting of a highway bottleneck with both vehicle platoons and ordinary vehicles (Figure 1). When a platoon is passing through the bottleneck, for a period of time, one lane is occupied by the platoon and not available to the background traffic. Thus, queuing happens upstream to the bottleneck.

2.1 Stochastic platoon arrival process

Let us model the randomness in the arrival process at the highway bottleneck; as we show subsequently, this model is simple enough to be integrated with the FQM and the CTM, both of which account for the interaction between the two traffic classes (although in different ways). The first class is the background traffic, with a constant inflow rate $a > 0$. The second class is the connected vehicles (platoons), with a stochastic, time-varying inflow rate $B(t)$. The unit of traffic flows is vehicles per hour (veh/hr).

We assume that (i) the inter-platoon headways are i.i.d. and exponentially distributed with the average $1/\lambda$, and (ii) the number of vehicles in platoons are also i.i.d. and exponentially distributed with the average $v/(\mu h)$, where v is the *free-flow speed* and h is the *intra-platoon spacing*. These assumptions are motivated by the inherent uncertainty in the formation, split, and movement of platoons [17]. Specifically, exponential distribution is commonly used model of randomness in vehicle headways [13]. In addition, for our purposes, the random platoon lengths can be also modeled as exponentially distributed random variables. With these assumptions, we use a two-state Markov process to model the arrival of platoons. Thus, $\{B(t); t \geq 0\}$ is a continuous-time, two-state Markov process with state space $\mathcal{B} := \{0, v/h\}$; see Figure 2 for an illustration.

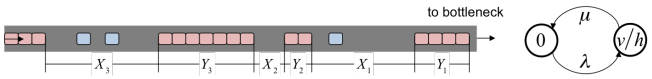


Figure 2: Platoon headway X_k and length Y_k are random (left). The arrival process of connected vehicles $B(t)$ is a two-state Markov process (right).

By standard results in Markov process (see e.g. [11]), the average inflow rate of connected vehicles is

$$\overline{B} = \lim_{t\to\infty} \frac{1}{t}\int_0^t B(\tau)d\tau = \frac{\lambda}{\lambda+\mu}\frac{v}{h}, \quad \text{a.s.} \tag{1}$$

where "a.s." means almost surely.

2.2 Fluid queuing model

The fluid queuing model is a simple model that can be used to study highway bottlenecks [22]. The essence of the FQM is to consider the highway bottleneck as a server with an infinite-sized buffer that stores the vehicles waiting for discharge. If there are vehicles waiting in the buffer, then the server discharges the vehicles at the *saturation rate*, denoted by u. The unit of u is veh/hr. If no traffic is waiting in the buffer, then the rate at which the server discharges traffic is the minimum of the saturation rate and the inflow rate.

The evolution of the traffic queue depends on the *priority rule*, i.e. how the server's saturation rate (i.e. the bottleneck's capacity) is allocated to the two traffic classes. Thanks to the simplicity of the FQM, we can consider two operational policies for capacity allocation. In the first policy, we model a highway bottleneck as a single server with the *proportional priority*; i.e., the road's capacity is allocated to a class of traffic is proportional to the fraction of this class of traffic in the aggregate traffic queue. In the second policy, we consider the case where vehicle platoons are prioritized to get discharged; we name this policy as *segmented priority*, which is motivated by the idea of dedicated lanes for connected vehicles [3].

Queuing dynamics: proportional priority

This priority rule corresponds to a highway where connected and ordinary vehicles share all lanes of the highway. This is a typical capacity allocation model for a highway that allows for mixing between connected and ordinary vehicles [29].

Figure 3 shows the two-class FQM. The (hybrid) state of the fluid queuing system is (b, q^a, q^b), where $b \in \mathcal{B}$ is the inflow of connected vehicles, $q^a \in \mathbb{R}_{\geq 0}$ is the queue of ordinary

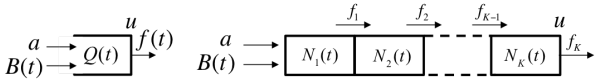


Figure 3: FQM (left) and CTM (right) under the proportional priority rule.

vehicles, and $q^b \in \mathbb{R}_{\geq 0}$ is the queue of connected vehicles. To capture the reduced intra-platoon vehicle spacing, we scale down queues of connected vehicles according to the spacing reduction enabled by platooning. More specifically, currently available platooning technology is able to reduce intra-platoon spacing to less than half of that between ordinary vehicles [1, 19]. We model this by scaling down the traffic queue and flow of connected vehicles with a coefficient (h/H). Thus, we define the *effective queue length* as

$$q = q^a + \frac{h}{H} q^b,$$

and the *effective discharge rate* as

$$f = f^a + \frac{h}{H} f^b.$$

Then, the effective discharge rate can be expressed as a function of b and q:

$$f(b,q) = \begin{cases} \min\{a + (h/H)b, u\}, & q = 0, \\ u, & q > 0. \end{cases}$$

Furthermore, the discharge rates of each class of traffic are given by

$$f^a(b, q^a, q^b) = \begin{cases} \frac{q^a}{q^a + \frac{h}{H} q^b} f(b, q^a + \frac{h}{H} q^b), & q^a + q^b > 0, \\ \min\Big\{a, \frac{a}{a + \frac{h}{H} b} u\Big\}, & q^a + q^b = 0, \end{cases} \tag{2a}$$

$$\frac{h}{H} f^b(b, q^a, q^b) = f(b, q^a + \frac{h}{H} q^b) - f^a(b, q^a, q^b). \tag{2b}$$

The above formulae essentially mean that the server's saturation rate is allocated to a class of traffic is proportional to this class's fraction in the aggregate (effective) queue. If $q^a + q^b = 0$, then the server's saturation rate is allocated according to a class's fraction in the aggregate (effective) inflow rate.

Throughout this article, we use lower-case letters (e.g. b and q) to denote the state variable, and upper-case letters (e.g. $B(t)$ and $Q(t)$) to denote the stochastic processes. Thus, the evolution of the queues $Q^a(t)$ and $Q^b(t)$ is governed by the following dynamics:

$$Q^a(0) = q^a, \ \frac{d}{dt} Q^a(t) = a - f^a(B(t), Q^a(t), Q^b(t)), \tag{3a}$$

$$Q^b(0) = q^b, \ \frac{d}{dt} Q^b(t) = B(t) - f^b(B(t), Q^a(t), Q^b(t)). \tag{3b}$$

One can check that, with the discharged rates defined in (2), $Q^a(t)$ and $Q^b(t)$ are continuous in t; thus $Q(t) = Q^a(t) + (h/H)Q^b(t)$ is also continuous in t.

We can also use the *infinitesimal generator* to represent the stochastic dynamics of the FQM. Since $\{B(t); t \geq 0\}$ is a stationary two-state Markov process and since $Q(t)$ is continuous in t, the FQM under proportional priority is right-continuous with left limits (RCLL, see [2]). Hence, by [9], the infinitesimal generator of the FQM under proportional priority can be written in operator form as follows:

$$\begin{aligned} &\mathcal{L}g(b, q^a, q^b) \\ &= \Big(a - f^a(b, q^a, q^b)\Big) \frac{\partial g}{\partial q^a} + \Big(b - f^b(b, q^a, q^b)\Big) \frac{\partial g}{\partial q^b} \\ &\quad + \mathbb{1}_{\{b=0\}} \lambda \Big(g(v/h, q^a, q^b) - g(0, q^a, q^b)\Big) \\ &\quad + \mathbb{1}_{\{b=v/h\}} \mu \Big(g(0, q^a, q^b) - g(v/h, q^a, q^b)\Big), \end{aligned} \tag{4}$$

where g is any function smooth in the continuous arguments, and $\mathbb{1}$ is the indicator function.

We say that the FQM under proportional priority is *stable* if there exists a constant $C > 0$ such that, for any initial condition $(b, q^a, q^b) \in \mathcal{B} \times \mathbb{R}^2_{\geq 0}$,

$$\limsup_{t \to \infty} \frac{1}{t} \int_0^t \mathsf{E}\Big[\exp\Big(Q^a(s) + (h/H) Q^b(s)\Big)\Big] ds \leq C. \tag{5}$$

This notion of stability is in line with that considered by Dai and Meyn for FQMs [8]. Essentially, it captures the boundedness of moments of queue lengths.

We are also interested in the steady-state joint distribution of $(B(t), Q^a(t), Q^b(t))$, called the *invariant probability measure*, denoted by π_{prop}. This measure is defined on the hybrid space $\mathcal{B} \times \mathbb{R}^2_{\geq 0}$. In general, boundedness of moments does not ensure convergence towards a unique invariant probability measure [8]. However, we will show while proving Theorem 1 that a stable FQM necessarily converges to a unique invariant probability measure.

With π_{prop}, the steady-state average $\overline{q}_{\text{prop}}$ and variance σ^2_{prop} of the effective queue lengths can be obtained as follows:

$$\overline{q}_{\text{prop}} = \int_{\mathcal{B} \times \mathbb{R}^2_{\geq 0}} q d\pi_{\text{prop}},$$

$$\sigma^2_{\text{prop}} = \int_{\mathcal{B} \times \mathbb{R}^2_{\geq 0}} (q - \overline{q}_{\text{prop}})^2 d\pi_{\text{prop}}.$$

We are able to derive $\overline{q}_{\text{prop}}$ and σ^2_{prop} in Section 3. Based on properties of the effective queue length, we will also derive bounds on the actual queue length $Q^a(t) + Q^b(t)$.

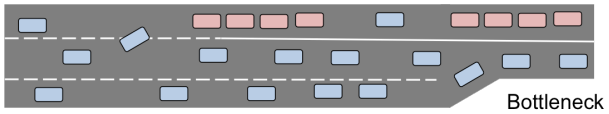
Furthermore, we define the *throughput under proportional priority*, denoted by J_{prop}, as follows:

$$J_{\text{prop}} = \sup\{a + \overline{B} : (5) \text{ holds}\}. \tag{6}$$

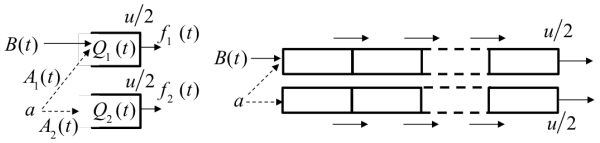
i.e. the supremum of the set of average aggregate arrival rates $a + \overline{B}$ such that the effective queue is stable; see (1) for the definition of $\overline{B}$.

Queuing dynamics: segmented priority

This priority rule is motivated by the idea of segmenting ordinary and connected vehicles and prioritizing connected vehicles in certain lanes [3]. For ease of presentation, we consider a highway bottleneck with two identical lanes; see Figure 4(a). Since the total capacity of the bottleneck is u, each lane has a capacity of $u/2$. The two traffic classes travel through the bottleneck as follows. When no connected vehicles are arriving, i.e. when $B(t) = 0$, ordinary vehicles are evenly distributed over two lanes; that is, background traffic enters each lane at rate $a/2$. When $B(t) = v/h$, ordinary vehicles are restricted to one lane (server 2); the other lane (server 1) is dedicated to platoons. Note that in this setting lane changes are not allowed at the bottleneck.



(a) A bottleneck with segmented priority.



(b) FQM (left) and CTM (right) under the segmented priority rule.

Figure 4: Relation between queue length and fraction of connected vehicles.

Under the above priority rule, we can model the bottleneck as two parallel servers as shown in Figure 4(b). Let $A_k(t)$ be the rate at which the background traffic enters the k-th server. The segmented priority rule leads to the following:

$$A_1(t) = \begin{cases} 0, & B(t) > 0, \\ a/2, & B(t) = 0, \end{cases}$$

$$A_2(t) = \begin{cases} a, & B(t) > 0, \\ a/2, & B(t) = 0, \end{cases}$$

Let q_k^a (resp. q_k^b) be the traffic queue of ordinary vehicles (resp. connected vehicles) in the k-th server. The effective queue lengths are

$$q_k = q_k^a + (h/H)q_k^b, \quad k = 1, 2.$$

The discharge rates are given by

$$f_1(b,q) = \begin{cases} \min\{a/2, u/2\}, & q = 0, b = 0, \\ \min\{b, u/2\}, & q = 0, b > 0, \\ u/2, & q > 0. \end{cases}$$

$$f_2(b,q) = \begin{cases} \min\{a/2, u/2\}, & q = 0, b = 0, \\ \min\{a, u/2\}, & q = 0, b > 0, \\ u/2, & q > 0. \end{cases}$$

Then, the dynamics of the effective queues can be written as follows:

$$Q_1(0) = q_1, \quad \frac{d}{dt}Q_1(t) = A_1(t) + \frac{h}{H}B(t) - f_1(B(t), Q(t)),$$

$$Q_2(0) = q_2, \quad \frac{d}{dt}Q_2(t) = A_1(t) - f_2(B(t), Q(t)).$$

For the above two-server system, we assume the following:

$$a < u, \quad v/H \le u/2. \tag{7}$$

The first assumption is a trivial necessary condition for stability. The second assumption essentially ensures that vehicle platoons are always in free flow if not interacting with the background traffic. This assumption is typically satisfied by highway traffic, since the capacity of a highway lane ($u/2$ in this case) is equal to the quotient between free-flow speed v and minimal free-flow spacing H [7].

Assuming that (7) holds implies that the inflow to server 1 is always less than the capacity of server 1; hence $Q_1(t)$ vanishes. Therefore, we only need to consider $Q_2(t)$ for steady-state analysis. Note that server 2 is essentially a single-class fluid queuing system, since no platoons enter server 2. Hence, $Q_2(t) = Q_2^a(t)$.

We say that the FQM under segmented priority is *stable* if there exists $C > 0$ such that, for any initial condition $(b, q_1^a, q_1^b, q_2^a, q_2^b) \in \mathcal{B} \times \mathbb{R}^4_{\ge 0}$,

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t \mathsf{E}\Big[\exp\Big(Q_2(s)\Big)\Big]ds \le C.$$

If the system is stable, there exists an invariant probability measure π_{seg} on $\mathcal{B} \times \mathbb{R}^4_{\ge 0}$, and the steady-state average $\overline{q}_{\text{seg}}$ and variance σ^2_{seg} of queue lengths can be obtained as follows:

$$\overline{q}_{\text{seg}} = \int_{\mathcal{B}\times\mathbb{R}^4_{\ge 0}} q_2 d\pi_{\text{seg}},$$

$$\sigma^2_{\text{seg}} = \int_{\mathcal{B}\times\mathbb{R}^4_{\ge 0}} (q_2 - \overline{q}_{\text{seg}})^2 d\pi_{\text{seg}}.$$

We will compute $\overline{q}_{\text{seg}}$, σ^2_{seg}, in Section 3.

Furthermore, we define the *throughput under segmented priority*, denoted by J_{seg}, as the supremum of the set of average aggregate demand $\bar{a} = \frac{\lambda+\mu/2}{\lambda+\mu}a$ such that the system is stable.

2.3 Cell transmission model

We now integrate our model of stochastic arrival process of vehicle platoons with the (classical) cell transmission model (CTM), which enables us to capture the spatial distribution of the two traffic classes within a highway bottleneck.

Before specifying the CTM, we briefly discuss on the relation between the FQM and the CTM. The two models are broadly consistent for the purpose of modeling aggregate congestion effect resulting from mismatch between traffic demand and available capacity. For example, Shen et al. [25] showed that the optimal routing policy for certain CTM networks can also be obtained from their FQM counterparts. In this article, we assume correspondence between the key parameters of the two models, including free-flow speed and bottleneck capacity; see Figures 3 and 4(b).

However, the two models also have important differences. First, the FQM always discharges the stored queue at the maximum saturation rate, while the discharged flow of the CTM may be smaller than the capacity if the traffic density at the bottleneck is less than the critical density. Therefore, the FQM gives a smaller estimate of traffic congestion than the CTM. Second, the CTM captures the spatial distribution of congestion and its propagation over various sections of the highway, while the FQM only considers the aggregate traffic queue.

In this article, we simulate the steady-state traffic volumes upstream to the bottleneck for the two-class CTM and compare their qualitative behavior with that of the analytical estimates (or bounds) of queue lengths obtained from the FQM (Section 4.1). This enables us to compare the aggregate impact of stochastic platoon arrivals on the build-up of traffic queues across the two models, without having to handle the spatial propagation of congestion.

For presentation of two-class CTM, we view the highway bottleneck introduced earlier as a segment consisting of K *cells*, and for the sake of simplicity, we assume that the cells are homogeneous and have unit lengths. We now present the dynamics of the CTM.

Under the proportional priority, the state of the highway is the vector of *ordinary vehicles' density* $n^a = [n_1^a, \dots n_K^a]^T$ and the vector of *connected vehicles' density* $n^b = [n_1^b, \dots n_K^b]^T$; see Figure 3. Given density vectors n^a and n^b, the *aggregate flow* out of each cell is given by

$$f_k(n^a, n^b) = \min\left\{ v\left(n_k^a + \frac{h}{H} n_k^b\right), U, w\left(\bar{n} - \left(n_{k+1}^a + \frac{h}{H} n_{k+1}^b\right)\right)\right\}, \quad k = 1, \dots, K-1,$$

$$f_K(n^a, n^b) = \min\left\{ v\left(n_K^a + \frac{h}{H} n_K^b\right), u\right\},$$

where v, h, H are the same as defined for the FQM, U is the *capacity* of Cells 1 through $K-1$, u is the capacity of Cell k, i.e. the bottleneck, w is the *congestion wave speed*, and $\bar{n}$ is the *jam density*. Here u is equal to the saturation rate of the FQM, and is assumed to be less than U. This flow-density relation follows from the classical triangular/trapezoidal fundamental diagram for highway traffic [7].

The aggregate flow is proportionally distributed to both traffic classes as follows:

$$f_k^a(n^a, n^b) = \begin{cases} \frac{n_k^a}{n_k^a + \frac{h}{H} n_k^b} f_k(n^a, n^b), & n_a + \frac{h}{H} n_b > 0, \\ 0, & \text{o.w.}, \end{cases}$$

$$f_k^b(n^a, n^b) = \begin{cases} \frac{n_k^b}{n_k^a + \frac{h}{H} n_k^b} f_k(n^a, n^b), & n_a + \frac{h}{H} n_b > 0, \\ 0, & \text{o.w.}, \end{cases} \quad k = 1, \dots, K.$$

Then, for any initial condition n^a, n^b the dynamics of the CTM is specified by

$$\begin{aligned} \frac{dN_1^a(t)}{dt} &= a - f_1^a(N^a(t), N^b(t)), \\ \frac{dN_1^b(t)}{dt} &= B(t) - f_1^b(N^a(t), N^b(t)), \\ \frac{dN_k^a(t)}{dt} &= f_{k-1}^a(N^a(t), N^b(t)) - f_k^a(N^a(t), N^b(t)), \\ \frac{dN_k^b(t)}{dt} &= f_{k-1}^b(N^a(t), N^b(t)) - f_k^b(N^a(t), N^b(t)), \quad k = 2, \dots, K. \end{aligned}$$

Note that the CTM under proportional priority is similar to the model introduced by Wright et al. [29], except for the stochastic arrival process $B(t)$.

Analogous to the FQM under segmented priority, the CTM for highway under segmented priority can be defined by considering independent, parallel CTMs; see Figure 4(b). That is, the top lane can be viewed as a two-class CTM, and the bottom lane can be viewed as a single-class CTM. The presentation of this model is similar to the CTM under proportional priority, and is omitted for the sake of brevity.

3. STABILITY ANALYSIS OF FLUID QUEUING MODEL

In this section, we study the stability of the FQM under two priority rules and characterize the effective and actual queue lengths under the two priority rules.

Our first result states that the FQM is stable under proportional priority if the average aggregate inflow rate is strictly less than the server's saturation rate:

Theorem 1 (Stability under proportional priority)**.** *The two-class fluid queuing system is stable under proportional priority if*

$$a + \frac{\lambda}{\lambda + \mu} \frac{v}{H} < u. \tag{8}$$

Furthermore, if (8) *holds, then, for any initial condition* $(b, q_a, q_b) \in \mathcal{B} \times \mathbb{R}_{\geq 0}^2$*, the joint distribution of the hybrid state* $(B(t), Q_a(t), Q_b(t))$*, denoted by* $P_t(b, q_a, q_b)$*, converges to a unique probability measure* π_{prop}*, i.e.*

$$\lim_{t \to \infty} \|P_t(b, q_a, q_b) - \pi_{\text{prop}}\|_{\text{TV}} = 0, \tag{9}$$

where $\|\cdot\|_{\text{TV}}$ *is the total variation distance.*

Proof. The proof of the boundedness of moments (in the sense of (5)) is based on a Foster-Lyapunov-type criterion introduced by Meyn and Tweedie [20, Theorem 4.3], which we recall in our setting as follows: if there exist constants $c > 0$ and $d < \infty$, and a norm-like function[1] $V : \mathcal{B} \times \mathbb{R}_{\geq 0}^2 \to \mathbb{R}$, such that

$$\mathcal{L}V(b, q_a, q_b) \leq -cV(b, q_a, q_b) + d, \quad \forall (b, q_a, q_b) \in \mathcal{B} \times \mathbb{R}_{\geq 0}^2, \tag{10}$$

then the FQM is stable in the sense of (5). Next, we prescribe the function V and explicitly construct the constants c and d.

Suppose that (8) holds. Let us consider the switched exponential Lyapunov function

$$V(b, q_a, q_b) = \begin{cases} k_0 e^{\gamma(q_a + (h/H) q_b)}, & b = 0, \\ k_1 e^{\gamma(q_a + (h/H) q_b)}, & b = v/h. \end{cases} \tag{11}$$

The parameters γ, k_0, and k_1 are constructed as follows. If $a + v/H \leq u$, we let

$$k_0 = 2\max\{1/\lambda, 1/\mu\},\ k_1 = 2k_0,\ \gamma = \frac{\lambda k_0 + 1}{(u-a)k_0},$$

which are positive under (8); otherwise, we let

$$\gamma = \frac{(\lambda + \mu)(u - a - \frac{\lambda}{\lambda+\mu}\frac{v}{H})}{2(u-a)(a + \frac{v}{H} - u)}, \tag{12a}$$

$$k_0 = \frac{\gamma(a + \frac{v}{H} - u) + \lambda + \mu}{\gamma((\lambda+\mu)(u - a - \frac{\lambda}{\lambda+\mu}\frac{v}{H}) - \gamma(u-a)(a + \frac{v}{H} - u))}, \tag{12b}$$

$$k_1 = \frac{\gamma(a - u) + \lambda + \mu}{\gamma((\lambda+\mu)(u - a - \frac{\lambda}{\lambda+\mu}\frac{v}{H}) - \gamma(u-a)(a + \frac{v}{H} - u))}. \tag{12c}$$

which are also positive under (8) and $a + v/H > u$. In addition, we construct the constants c and d as follows:

$$c = \frac{1}{2\gamma k_1},\ d = \max_{b \in \mathcal{B}} |\mathcal{L}V(b, 0, 0) + cV(b, 0, 0)|.$$

Next, we verify (10) with V, c, and d as constructed above. Note that, for $b = 0, q_a + q_b = 0$, we have

$$\begin{aligned} \mathcal{L}V(0,0,0) &\leq |\mathcal{L}V(0,0,0)| \\ &\leq \max_{b \in \mathcal{B}} |\mathcal{L}V(b,0,0) + cV(b,0,0)| - cV(0,0,0) \\ &= -cV(0,0,0) + d; \end{aligned}$$

[1] That is, for each $b \in \mathcal{B}$, $V \to \infty$ if $q_a \to \infty$ or $q_b \to \infty$.

for $b = 0, q_a + q_b > 0$, we have

$$\begin{aligned}\mathcal{L}V &= k_0(a-u)\gamma e^{\gamma(q_a+(h/H)q_b)} + \lambda(k_0-k_1)e^{\gamma(q_a+(h/H)q_b)} \\ &= \Big(k_0\gamma(a-u) + \lambda(k_0-k_1)\Big)e^{\gamma(q_a+(h/H)q_b)} \\ &\le -e^{\gamma(q_a+(h/H)q_b)} \le -cV \le -cV + d;\end{aligned}$$

similarly, one can show that $\mathcal{L}V \le -cV + d$ for $b = v/h$ and $(q_a, q_b) \in \mathbb{R}^2_{\ge 0}$.

Finally, since we have verified (10), we can apply [20, Theorem 4.3] and obtain (5).

To obtain (9), i.e. the convergence towards a unique invariant probability measure π_{prop}, note that, under (8), we have $a < u$. Hence, the aggregate traffic queue necessarily decreases when $B(t) = 0$. Therefore, for any initial condition, there is a strictly positive probability that $Q_a(t) = Q_b(t) = 0$ for a sufficiently large t. That is, the state $(0, 0, 0) \in \mathcal{B} \times \mathbb{R}^2_{\ge 0}$ can be attained with positive probability. Then, one can adapt the proof of [2, Theorem 4.6] and obtain the convergence to a unique invariant probability measure (in the sense of total variation distance). For details of this argument, we refer the readers to [14, 16]. □

In fact, for a stable FQM, we can also study the queue length:

Proposition 1. *For the FQM under proportional priority, if (8) holds, the steady-state effective queue length $\overline{q}_{\text{prop}}$ and variance σ^2_{prop} can be analytically expressed as follows:*

$$\overline{q}_{\text{prop}} = \begin{cases} 0, & a + \frac{v}{H} < u, \\ \frac{\lambda}{(\lambda+\mu)^2}\frac{a+\frac{v}{H}-u}{u-a-\frac{\lambda}{\lambda+\mu}\frac{v}{H}}\frac{v}{H}, & o.w. \end{cases} \tag{13a}$$

$$\sigma^2_{\text{prop}} = \begin{cases} 0, & a + \frac{v}{H} < u, \\ \frac{\lambda}{(\lambda+\mu)^3}\frac{(a+\frac{v}{H}-u)(u-a)}{\left(u-a-\frac{\lambda}{\lambda+\mu}\frac{v}{H}\right)^2}\frac{v}{H}, & o.w. \end{cases} \tag{13b}$$

Furthermore, the steady-state actual queue length $\tilde{q} = \overline{q}_{a,\text{prop}} + \overline{q}_{b,\text{prop}}$ and its variance $\tilde{\sigma}^2$ satisfy

$$\begin{aligned}\overline{q}_{\text{prop}} \le \tilde{q} &\le \left(\frac{1}{1+\theta} + \frac{\theta}{1+\theta}\frac{H}{h}\right)\overline{q}_{\text{prop}}, \\ \sigma^2_{\text{prop}} \le \tilde{\sigma}^2 &\le \left(\frac{1}{1+\theta} + \frac{\theta}{1+\theta}\frac{H}{h}\right)^2\sigma^2_{\text{prop}},\end{aligned}$$

where $\theta = \frac{v}{Ha}$.

The derivation of the above result is based on the following lemma:

Lemma 1. *Under proportional priority, the following set*

$$\mathcal{Q}_{\text{inv}} := \left\{[q_a, q_b]^T \in \mathbb{R}^2_{\ge 0} : \frac{h}{H}q_b \le \theta q_a\right\},$$

is globally attracting, i.e., for any initial condition $(b, q_a, q_b) \in \mathcal{B} \times \mathbb{R}^2_{\ge 0}$,

$$\lim_{t\to\infty} \inf_{\substack{[\xi_a,\xi_b]^T \\ \in \mathcal{Q}_{\text{inv}}}} \left\|[Q_a(t), Q_b(t)]^T - [\xi_a, \xi_b]^T\right\|_2 = 0,$$

and positively invariant[2], i.e., for any initial condition $(b, q_a, q_b) \in \mathcal{B} \times \mathcal{Q}_{\text{inv}}$,

$$[Q_a(t), Q_b(t)]^T \in \mathcal{Q}_{\text{inv}}, \quad \forall t \ge 0.$$

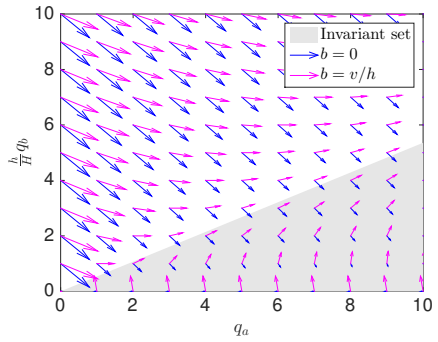


Figure 5: Illustration of the queuing dynamics and the invariant set $\mathcal{Q}_{\text{inv}}$ under proportional priority. The arrows represent the vectors of time-derivatives defined in (3) for both $b = 0$ and for $b = v/h$.

This lemma can be proved by utilizing properties of the queuing dynamics (3). We omit the proof here due to space limitations. Figure 5 illustrates the basic intuition behind this result. The proof entails that, for any $b \in \mathcal{B}$ and for any $[q_a, q_b]^T$ such that $[q_a, q_b]^T \notin \mathcal{Q}_{\text{inv}}$, the vector of time-derivatives of the queue lengths has a non-zero component that points to the interior of the invariant set $\mathcal{Q}_{\text{inv}}$.

Proof of Proposition 1. Average effective queue lengths and variance: Kulkarni gives an analytical expression for the steady-state distribution of the queue length in a single-class FQM that switches between a finite number of modes [16, Theorem 11.6]. In the particular setting of this proposition, the steady-state joint distribution of (b, q) can be represented as a probability density function (pdf) as follows:

$$\mathsf{f}(b, q) = \begin{cases} \mathsf{z}\delta_0 + \alpha_1 e^{-q/\beta}, & b = 0, \\ \alpha_2 e^{-q/\beta}, & b = v/H, \end{cases} \tag{14}$$

where

$$\begin{aligned}&\mathsf{z} = \frac{1}{\lambda+\mu}\left(\mu - \lambda\frac{a + v/H - u}{u-a}\right), \alpha_1 = \frac{\lambda\mathsf{z}}{u-a}, \\ &\alpha_2 = \frac{\lambda\mathsf{z}}{a+v/H-u},\ \beta = \left(\frac{\mu}{a+v/H-u} - \frac{\lambda}{u-a}\right)^{-1},\end{aligned}$$

and δ_0 is the Dirac delta function centered at 0. Hence, we can obtain the expected value $\overline{q}_{\text{prop}}$ and variance σ^2_{prop} of the effective queue q, which are given by (13a) and (13b), respectively.

Lower bounds for the actual queue length: Since the actual queue length $(q_a + q_b)$ is no less than the effective queue length q, $\overline{q}_{\text{prop}}$ and σ^2_{prop} are straightforward lower bounds for the expected value $\tilde{q}$ and variance $\tilde{\sigma}^2$ of the actual queue.

Upper bounds for the actual queue length: Recall the invariant set $\mathcal{Q}_{\text{inv}}$ from Lemma 1. For each $(q_a, q_b) \in \mathcal{Q}_{\text{inv}}$, since $(h/H)q_b \le \theta q_a$, we have

$$\frac{1+\theta}{\theta}\frac{h}{H}q_b \le q_a + \frac{h}{H}q_b. \tag{15}$$

Then,

$$q_a + q_b = q_a + (H/h)\frac{h}{H}q_b = q_a + \frac{h}{H}q_b + (H/h - 1)\frac{h}{H}q_b$$

[2]See [2] for details regarding invariant sets for PDMPs.

$$\overset{(15)}{\leq} (q_a + \frac{h}{H} q_b) + (H/h - 1)\frac{\theta}{1+\theta}(q_a + \frac{h}{H} q_b)$$
$$= \left(\frac{1}{1+\theta} + \frac{\theta}{1+\theta}\frac{H}{h}\right)(q_a + \frac{h}{H} q_b). \quad (16)$$

Since the set $\mathcal{Q}_{\text{inv}}$ is globally attracting and positively invariant, the invariant probability measure π_{prop} vanishes outside $\mathcal{Q}_{\text{inv}}$ [2]. Therefore,

$$\tilde{q} = \int_{\mathcal{B}\times\mathbb{R}^2_{\geq 0}} (q_a + q_b) d\pi_{\text{prop}} = \int_{\mathcal{B}\times\mathcal{Q}_{\text{inv}}} (q_a + q_b) d\pi_{\text{prop}}$$
$$\overset{(16)}{\leq} \left(\frac{1}{1+\theta} + \frac{\theta}{1+\theta}\frac{H}{h}\right) \int_{\{0,v/h\}\times\mathbb{R}^2_{\geq 0}} (q_a + \frac{h}{H} q_b) d\pi_{\text{prop}}$$
$$= \left(\frac{1}{1+\theta} + \frac{\theta}{1+\theta}\frac{H}{h}\right)\overline{q};$$

the last equality results from the fact that π_{prop} gives the same average value of $(q_a + \frac{h}{H} q_b)$ as the pdf in (14) does. The upper bound for variance the variance σ^2_{prop} of the actual queue can be similarly obtained. □

An analogous result regarding the stability and queue length of the FQM under segmented priority can be derived:

Proposition 2 (Segmented priority)**.** *Consider the two-class fluid queuing model and assume that* (7) *holds. Then the model is stable if*

$$\frac{\lambda + \mu/2}{\lambda + \mu} a < u/2. \quad (17)$$

Furthermore, under (17)*, the average and variance of queue length are given by*

$$\overline{q}_{\text{seg}} = \begin{cases} 0, & a < u/2, \\ \frac{\lambda}{(\lambda+\mu)^2} \frac{(a-u/2)a/2}{u/2 - \frac{\lambda+\mu/2}{\lambda+\mu} a}, & o.w. \end{cases}$$
$$\sigma^2_{\text{seg}} = \begin{cases} 0, & a < u/2, \\ \frac{\lambda}{(\lambda+\mu)^2} \frac{(a-u/2)(u/2-a/2)a/2}{\left(u/2 - \frac{\lambda+\mu/2}{\lambda+\mu} a\right)^2}, & o.w. \end{cases}$$

Proof. Note that, under (7), the set $\{(q_1^a, q_1^b, q_2^a, q_2^b) \in \mathbb{R}^4_{\geq 0} : q_1^a = q_1^b = q_2^b = 0\}$ is globally attracting and positively invariant under the segmented priority; i.e. $Q_2^a(t)$ could be arbitrarily large, but $Q_1^a(t)$, $Q_1^b(t)$, and $Q_2^b(t)$ necessarily vanish after sufficiently long time. Hence, we only need to consider the queue $Q_2^a(t)$. Note that the server 2 can be viewed as a single-class FQM. Thus, the rest of the proof is analogous to that of Theorem 1. □

4. PLATOONING OPERATIONS

We are now ready to discuss how characteristics of platoons (specifically, penetration rate of connected vehicles, vehicle spacing within platoons, platoon length, and priority rule) affect traffic queue. Table 1 lists the nominal values considered in this section.

Fraction of platooned vehicles

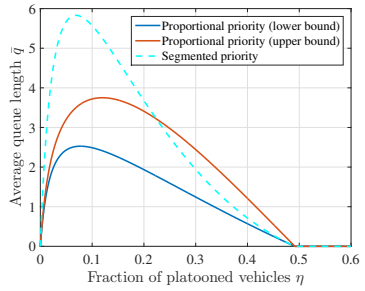
The fraction of platooned vehicles can be written as

$$\eta = \frac{\overline{B}}{a + \overline{B}} = \frac{\frac{\lambda}{\lambda+\mu}\frac{v}{h}}{a + \frac{\lambda}{\lambda+\mu}\frac{v}{h}},$$

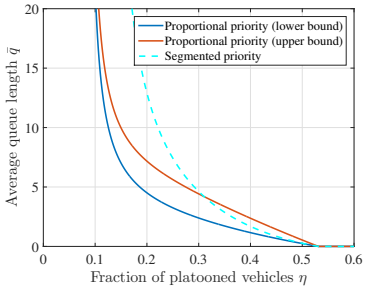
Table 1: Nominal parameters of traffic flow and platoons.

Name	Symbol	Value	unit
Cell length	l	1	mi
Free-flow speed	v	60	mi/hr
Congestion wave speed	w	20	mi/hr
Jam density (per lane)	$\overline{\rho}$	100	veh/mi
Capacity (per lane)	u	1500	veh/hr
Average aggregate demand	$a + \overline{B}$	3600	veh/hr
Spacing ratio	h/H	1/3	N/A
Penetration rate of platooned vehicles	η	0.4375	N/A
Platoon arrival rate	λ	30	hr^{-1}

where $\overline{B}$ is the average inflow of connected vehicles given by (1). Suppose that we fix the aggregate average demand $a + \overline{B}$, the platoon lengths μ, and the space h, and vary λ (or equivalently η). Figure 6 shows how the (bounds of) queue length vary with fraction of platooned vehicles. When the



(a) $a + \overline{B} < u$.



(b) $a + \overline{B} > u$.

Figure 6: Impact of fraction of platooned vehicles on (actual) queue length.

average aggregate demand $a + \overline{B}$ is smaller than the capacity u, this relation is characterized by a cap-shaped curve (Figure 6(a)). The points worth noting are: (i) at a low fraction, platooning increases the randomness of the arrival process, and thus increases the traffic queue, and (ii) as the fraction increases further, the gain of the reduced within-platoon spacing compensates for the increase in randomness of the arrival process. From a practical perspective, the inefficient fraction of platooned vehicles (≈ 0.1 in this example) should be avoided to limit the effect of random platoon arrivals. Furthermore, there exists a threshold η_0 beyond which no queue exists:

$$\eta_0 = 1 - \frac{u - v/H}{a + \overline{B}}.$$

To see this, note that, for $\eta > \eta_0$, we have

$$a + \frac{h}{H}B(t) \le a + \frac{v}{H} = (1-\eta)(a+\overline{B}) + \frac{v}{H} < u,$$

and thus the queue never grows. Hence, if the fraction of connected vehicles is greater than η_0, then the traffic on the highway can maintain free flow even with a high density, thanks to the reduced spacing between platooned vehicles.

When the average aggregate demand is greater than the capacity (Figure 6(b)), the $\overline{q} - \eta$ curve has an elbow-shaped shape. In this case, note that, to ensure stability, at least a certain fraction of the total demand should be connected vehicles such that the excessive demand is compensated by the reduced spacing between platooned vehicles. This threshold, η_1, can be obtained from Theorem 1:

$$\eta_1 = \frac{(a+\overline{B}-u)_+}{(a+\overline{B})(1-h/H)}.$$

Beyond this threshold, the queue length decreases with the fraction of platooned vehicles.

Intra-platoon spacing

Now we study the benefit of reducing the intra-platoon spacing. Current technology enables reduction of inter-vehicle spacing by 50% or more [1]. Suppose that we fix the aggregate average demand $a + \overline{B}$ and vary h. For the queue to be stable, the spacing should not exceed the following threshold:

$$h_1 < \frac{u - \eta(a+\overline{B})}{(1-\eta)(a+\overline{B})}H.$$

Figure 7 shows how the queue varies with the ratio H/h when the average aggregate demand is greater than the capacity, i.e. $a + \overline{B} > u$. As expected, queue length decreases

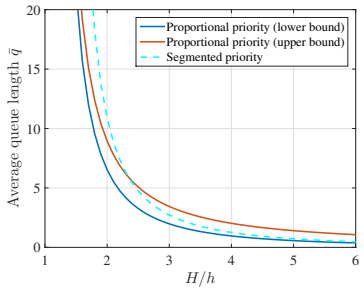


Figure 7: Impact of intra-platoon spacing on queue length.

as H/h increases. In addition, the curve becomes shallow as the ratio increases, implying that an excessively high ratio (more than 3 in Figure 7) does not bring much benefit. Note that high H/h ratios are not recommended for safety consideration either [1].

Arrival frequency and lengths of platoons

Another question of practical interest is whether connected vehicles should form a large number of short platoons or a small number of long platoons. Platoon lengths affect fuel consumption and the ease of implementation [1]. Here, we focus on how average platoon length affect the traffic queue. Suppose that we fix the ratio between λ and μ, and vary λ. That is, we fix the fraction of platooned vehicles η, but

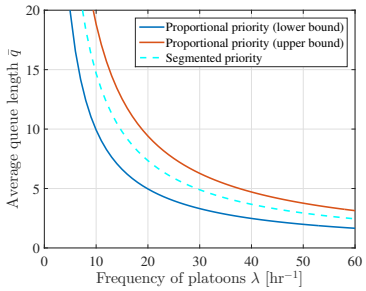


Figure 8: Impact of platoon arrival frequency on queue length.

vary the frequency and lengths of the platoons. Figure 8 shows that a high frequency leads to a smaller queue. The reason is that, as the platoons become more frequent and shorter, the probability of forming a long queue decreases. A practical interpretation in the setting illustrated in Figure 1 is that it is more difficult for long platoons to go through the bottleneck than short ones.

Priority rule

In Figures 6, 7, and 8, the queue lengths resulting from segmented priority are also plotted. Figure 6(a) implies that, with a low fraction of platooning, proportional priority leads to smaller traffic queues. This is intuitive in that prioritization of platooned vehicles under-utilizes the road's capacity if the fraction η is low. However, as the fraction increases (say greater than 0.4 in the figure), the queue length associated with segmented priority approaches the lower bound of that associated with proportional priority. In addition, Figure 7 implies that the relative benefit of segmenting two classes of traffic increases as the intra-platoon spacing decreases. Figure 8 implies that the relative benefit of segmenting does not significantly vary with the transition rates. However, in all the above-mentioned figures, the queue lengths associated with segmented priority are never below the lower bounds associated with proportional priority. Therefore, segmented priority is not guaranteed to outperform the proportional priority, at least in the setting being considered here. In a broader range of settings, segmented priority may outperform proportional priority when the ratio H/h is very high, i.e. when the intra-platoon spacing is very short.

Finally, we can obtain from Theorem 1 that the throughput (as defined in (6)) under proportional priority is

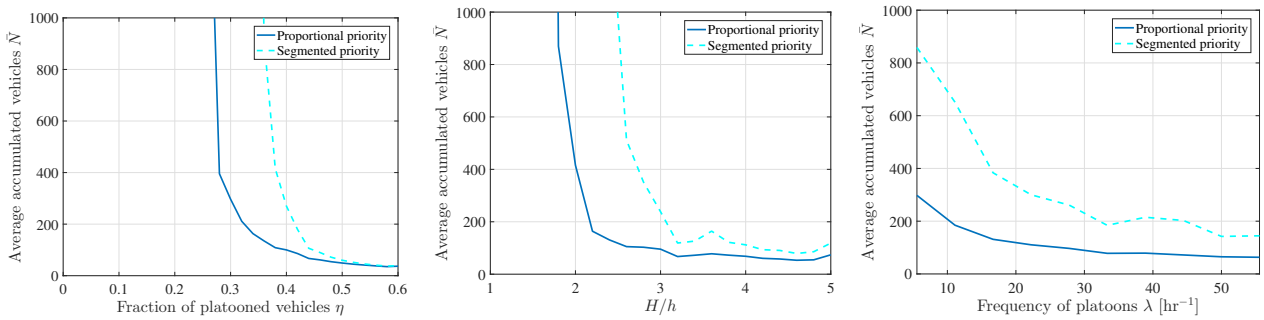
$$J_{\text{prop}} = \frac{u}{1-\eta+(h/H)\eta}.$$

That is, throughput increases with the fraction of connected vehicles. Similarly, we can obtain from Proposition 2 that the throughput under segmented priority is

$$J_{\text{seg}} = \min\left\{\frac{u}{1-\eta}, \frac{2H}{h\eta}u, \frac{\lambda+\mu}{(1-\eta)(2\lambda+\mu)}u\right\}$$

for $0 < \eta < 1$. One can show that

$$J_{\text{prop}} > J_{\text{seg}}, \quad \text{if } \eta > \frac{\frac{\lambda}{\lambda+\mu}}{\frac{\lambda}{\lambda+\mu}+\frac{h}{H}};$$



(a) Varying fraction of connected vehicles.

(b) Varying spacing ratio.

(c) Varying transition rates.

Figure 9: Relation between traffic queue in CTM and platooning characteristics ($a + \overline{B} > u$).

$$J_{\text{prop}} < J_{\text{seg}}, \quad \text{if } \eta < \frac{\frac{\lambda}{\lambda+\mu}}{\frac{\lambda}{\lambda+\mu} + \frac{h}{H}}.$$

That is if the fraction of connected vehicles is high, then the segmented priority leads to a smaller throughput. The intuition is that, in such a scenario, one lane (server 1 in Figure 4(b)) is not sufficient to serve the platoons, while the other lane (server 2) is under-utilized.

4.1 Comparison with cell transition model

In this subsection, we demonstrate via simulation that the insights obtained using the FQM largely extend to more detailed CTM as described in Section 2.3. The simulation model consists of $K = 10$ two-lane cells, and its parameters are given in Table 1.

In order to compare the behavior of the CTM and the FQM, we need to define an analogue notion to queues based on traffic flows or densities. One simple choice is to use the total number of vehicles accumulated per cell $\overline{N}$ averaged over a long time. This accumulation is due to congestion induced by stochastic vehicle platoon arrival, and it can be calculated by averaging the difference between the inflow to the first cell and the outflow from the last cell. With discretized simulation time steps, we have

$$\overline{N} = \sum_{\tau=1}^{T} \frac{a + B(\tau\Delta t) - f_K(\tau\Delta t)}{TK}, \tag{18}$$

over the duration $T\Delta t$ of the simulation.

Note that the quantity defined above does not immediately correspond to the traffic queue in FQM. The reason is that, in the CTM, vehicles need a nonzero time to go through the cells, and thus the traffic densities in the CTM are not zero even when the entire highway is in free flow. However, since the evolution of queue lengths (3) is also governed by the difference between inflow and outflow, $\overline{N}$ is the natural choice for comparison against $\bar{q}$, and we indeed see from Figures 6 to 8 that there is a qualitative agreement.

The results of the simulations with different fractions of connected vehicles η are shown on Figure 9(a). The simulated average number of accumulated vehicles decrease with η, similarly to the plot of analytical expression of average queue lengths, as shown on Figure 6(b). In case of segmented priority, the number of accumulated vehicles is larger due to lower lane utilization. In this case, the congestion solely results from the accumulation of the ordinary vehicles, and connected vehicles can pass through freely.

Varying spacing ratio gives similar results (Figure 9(b)) to the ones shown in Figure 7. Note that since the road space that a vehicle takes consists of the length of the vehicle and the headway it keeps, and platooning can only reduce the headway, the maximum realistic spacing ratio will be limited by both physical and safety reasons.

The results of simulations with different transition rates λ are shown on Figure 9(c). The transition rate effectively determines the average platoon length, with low λ corresponding to few very long platoons, and high λ with many short platoons. The number of accumulated vehicles decreases with increasing λ, since shorter queues are discharged faster; this trend agrees with Figure 8.

Thus, our computational study shows that the FQM is fairly consistent with the CTM in terms of estimating the impact of platooned vehicles on traffic conditions.

5. CONCLUDING REMARKS

In this article, we propose a two-class fluid queuing model to study the traffic congestion induced by vehicle platooning at highway bottlenecks. Using this model, we are able to evaluate the impact of parameters of vehicle platoons and the priority rule on traffic congestion and throughput. This work can be extended in several directions. First, to consider the impact of congestion downstream to a bottleneck, tandem FQMs with finite buffers can be considered. Known results [16] imply that, for FQMs with finite buffers, average platoon length affects not only queue length, but also stability. Second, our approach can be used to study control of platoons in response to local traffic conditions, such as time-varying demand of background traffic and road capacity perturbations. Of particular interest is the tradeoff between throughput gain and fuel savings.

ACKNOWLEDGMENTS

This work was supported by NSF CNS-1239054 CPS Frontiers, NSF CAREER Award CNS-1453126, the Future Urban Mobility project under the Singapore NRF, European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 674875, VINNOVA within the FFI program under contract 2014-06200, the Swedish Research Council, Knut and Alice Wallenberg Foundation and the Swedish Foundation for Strategic Research. We deeply appreciate the feedback from the

anonymous reviewers.

6. REFERENCES

[1] A. Alam, B. Besselink, V. Turri, J. Martensson, and K. H. Johansson. Heavy-duty vehicle platooning for sustainable freight transportation: A cooperative method to enhance safety and efficiency. *IEEE Control Systems*, 35(6):34–56, 2015.

[2] 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*, volume 51, pages 1040–1075. Institut Henri Poincaré, 2015.

[3] C. Bergenhem, S. Shladover, E. Coelingh, C. Englund, and S. Tsugawa. Overview of platooning systems. In *Proceedings of the 19th ITS World Congress, Oct 22-26, Vienna, Austria (2012)*, 2012.

[4] B. Besselink, V. Turri, S. van de Hoef, K.-Y. Liang, A. Alam, J. Mårtensson, and K. H. Johansson. Cyber-physical control of road freight transport. *Proceedings of IEEE*, 104(5):1128–1141, 2016.

[5] S. Calvert, H. Mahmassani, J.-N. Meier, P. Varaiya, S. Hamdar, D. Chen, X. Li, A. Talebpour, and S. P. Mattingly. Traffic flow of connected and automated vehicles: Challenges and opportunities. In *Road Vehicle Automation 4*, pages 235–245. Springer, 2018.

[6] S. Coogan and M. Arcak. A dissipativity approach to safety verification for interconnected systems. *IEEE Transactions on Automatic Control*, 60(6):1722–1727, 2015.

[7] C. F. Daganzo. The cell transmission model: A dynamic representation of highway traffic consistent with the hydrodynamic theory. *Transportation Research Part B: Methodological*, 28(4):269–287, 1994.

[8] J. G. Dai and S. P. Meyn. Stability and convergence of moments for multiclass queueing networks via fluid limit models. *IEEE Transactions on Automatic Control*, 40(11):1889–1904, 1995.

[9] M. H. A. Davis. Piecewise-deterministic Markov processes: A general class of non-diffusion stochastic models. *Journal of the Royal Statistical Society. Series B. Methodological*, 46(3):353–388, 1984.

[10] M. L. Delle Monache and P. Goatin. Scalar conservation laws with moving constraints arising in traffic flow modeling: an existence result. *Journal of Differential Equations*, 257:4015–4029, 2014.

[11] R. G. Gallager. *Stochastic Processes: Theory for Applications*. Cambridge University Press, 2013.

[12] R. Horowitz and P. Varaiya. Control design of an automated highway system. *Proceedings of the IEEE*, 88(7):913–925, 2000.

[13] S. E. Jabari and H. X. Liu. A stochastic model of traffic flow: Theoretical foundations. *Transportation Research Part B: Methodological*, 46(1):156–174, 2012.

[14] L. Jin and S. Amin. Stability of fluid queueing systems with parallel servers and stochastic capacities. *IEEE Transactions on Automatic Control*, to appear.

[15] A. Keimer, N. Laurent-Brouty, F. Farokhi, H. Signargout, V. Cvetkovic, A. M. Bayen, and K. H. Johansson. Integration of information patterns in the modeling and design of mobility management services. Technical report, arXiv:1707.07371, 2017.

[16] V. G. Kulkarni. Fluid models for single buffer systems. *Frontiers in queueing: Models and applications in science and engineering*, 321:338, 1997.

[17] J. Larson, K.-Y. Liang, and K. H. Johansson. A distributed framework for coordinated heavy-duty vehicle platooning. *Intelligent Transportation Systems, IEEE Transactions on*, 16(1):419–429, 2015.

[18] W. Levine and M. Athans. On the optimal error regulation of a string of moving vehicles. *IEEE Transactions on Automatic Control*, 11(3):355–361, 1966.

[19] J. Lioris, R. Pedarsani, F. Y. Tascikaraoglu, and P. Varaiya. Platoons of connected vehicles can double throughput in urban roads. *Transportation Research Part C: Emerging Technologies*, 77:292–305, 2017.

[20] S. P. Meyn and R. L. Tweedie. Stability of Markovian processes III: Foster-Lyapunov criteria for continuous-time processes. *Advances in Applied Probability*, pages 518–548, 1993.

[21] 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 Transactions on Vehicular Technology*, 59(9):4268–4279, 2010.

[22] G. F. Newell. *Applications of Queueing Theory*, volume 4. Springer Science & Business Media, 2013.

[23] J. Ploeg, N. van de Wouw, and H. Nijmeijer. L_p string stability of cascaded systems: Application to vehicle platooning. *IEEE Transactions on Control Systems Technology*, 22(2):786–793, 2014.

[24] R. Rijkswaterstaat, the Ministry of Infrastructure, and t. N. the Environment. European truck platooning challenge 2016—lessons learnt. www.eutruckplatooning.com, 2016.

[25] W. Shen and H. Zhang. System optimal dynamic traffic assignment: Properties and solution procedures in the case of a many-to-one network. *Transportation Research Part B: Methodological*, 65:1–17, 2014.

[26] D. Swaroop, J. K. Hedrick, and S. B. Choi. Direct adaptive longitudinal control of vehicle platoons. *IEEE Transactions on Vehicular Technology*, 50(1):150–161, 2001.

[27] A. Talebpour and H. S. Mahmassani. Influence of connected and autonomous vehicles on traffic flow stability and throughput. *Transportation Research Part C: Emerging Technologies*, 71:143–163, 2016.

[28] S. Tsugawa, S. Jeschke, and S. E. Shladover. A review of truck platooning projects for energy savings. *IEEE Transactions on Intelligent Vehicles*, 1(1):68–77, 2016.

[29] M. Wright, G. Gomes, R. Horowitz, and A. A. Kurzhanskiy. A new model for multi-commodity macroscopic modeling of complex traffic networks. *arXiv preprint arXiv:1509.04995*, 2015.