A Queueing Model and Analysis for Autonomous Vehicles on Highways

Neda Mirzaeian, Soo-Haeng Cho, Alan Scheller-Wolf
Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213
mneda@andrew.cmu.edu, soohaeng@cmu.edu, awolf@andrew.cmu.edu

November 2018

We investigate the effects of autonomous vehicles (AVs) on highway congestion. AVs have the potential to significantly reduce highway congestion, since these vehicles are able to maintain smaller inter-vehicle gaps and travel together in larger platoons (or batches) than human-driven vehicles (HVs). Various policies have been proposed to regulate AVs on highways, yet no in-depth comparison of these policies exists. To address this shortcoming, we develop a queueing model for a multi-lane highway, and analyze two policies for a mixed fleet of HVs and AVs: the designated-lane policy (“D policy”) under which one lane is designated to AVs, and the integrated policy (“I policy”) under which AVs travel together with HVs in all lanes. Our analysis connects the service rate of the queueing system to congestion on the highway, as well as inter-vehicle gaps using a Markovian arrival process (MAP). We measure the performance of these policies using mean travel time and throughput as metrics. We show that although the I policy performs at least as well as a benchmark case with no AVs, the D policy outperforms the benchmark only when the highway is heavily congested and AVs constitute the majority of vehicles; in such a case this policy may outperform the I policy as well. These findings caution against recent industry and government proposals that the D policy should be employed at the beginning of the mass appearance of AVs. Finally, we calibrate our model to data; our numerical analysis shows that for highly congested highways, a moderate number of AVs can make substantial improvement (e.g., 22% AVs can improve throughput by 30%), and when all vehicles are AVs, throughput can be increased over 400%.

Key words: Autonomous vehicles, Platooning, Queues with state-dependent service rates, Smart city operations, Transportation

1. Introduction

The autonomous vehicle (AV) industry is growing enormously fast. The very first experiment with AVs dates back to the 1920s (Maurer et al. 2016), but the first true AV was invented in the mid 1980s by Carnegie Mellon University’s Navlab and ALV projects (Kanade et al. 1986). Today all major automobile manufacturers, as well as many research centers, are running experiments with AVs, and several manufacturers have announced long-term plans to mass-produce AVs (Muoio 2017). In fact, we already see AVs operating on the roads of several major cities in the U.S. and several other countries. For example, Google’s self-driving vehicles have driven more than seven million miles and are on the streets of six states in the U.S. (Waymo 2018). More than half of the largest U.S. cities are preparing for autonomous vehicles in their long-range transportation plans (National League of Cities 2018). Industry experts predict that by 2025, fully autonomous vehicles will arrive on highways (The automated driving community 2018), and by 2035, one out of every four vehicles on the road will be an AV (Bierstedt et al. 2014). KPMG predicts that AVs are capable of increasing the capacity of highways by 500%, without building new roads (Albright
et al. 2015). But, despite all these claims, very little is actually known regarding the post-AV era, for example how exactly AVs will affect traffic flow is an open question.

This paper focuses on how AVs will affect the congestion of highways, which are the arteries of nations. Congestion has always been a critical issue for urban planners, particularly due to its effect on commute time. There are about 120 million workers in the U.S. (Statista 2017), and according to the U.S. Census Bureau, every worker in the United States spends, on average, 50 minutes each day to commute from home to work and back, and at least 60% of these trips happen with vehicles. On average 5.5% of this commute time is spent in congestion (INRIX 2017), and highways carry about 24% of total travel (Federal Highway Administration 2011); as a result, roughly speaking, in the U.S. alone, 48 million minutes of commute time is wasted due to congestion on highways every workday. AVs have a great potential to reduce such waste, should they be utilized properly. But will AVs mitigate congestion in reality? The answer to this question is not simple, and depends on several factors, such as the adoption rate of AVs, highway characteristics, and policies and regulations.

In this paper, we develop a queueing model for a multi-lane highway, and analyze two policies for a mixed fleet of human-driven vehicles (HVs) and AVs: the designated-lane policy ("D policy" in short) under which one lane is designated to AVs, and the integrated policy ("I policy" in short) under which AVs travel together with HVs in all lanes. The Departments of Transportation in Colorado, Wisconsin and Washington are considering designated lanes for AVs (Aguilar 2018). Bierstedt et al. (2014) predict a different evolution. They claim that before 2025 there will be very few AVs on highways and their effect on highway traffic flow will be negligible. They predict by 2025 there will be enough AVs on highways such that designating one separate lane to them will become reasonable. By 2030, after having AVs on designated lanes of highways for 5 years, HVs and AVs can be integrated on all lanes of highways.

Our queueing model captures the potential benefits of AVs by explicitly modeling platoons and headway. Observing traffic on a highway, one notices that HVs usually move in platoons (batches). Within each platoon vehicles follow one another keeping small intraplatoon headway - the time gap (in seconds) between two vehicles. The headway between two consecutive platoons, interplatoon headway, is typically significantly greater than the intraplatoon headway. Thus, the overall headway between vehicles depends on the size of platoons and the intraplatoon and interplatoon headways. Moreover, it is different for AVs and HVs: According to several field experiments (e.g., Bergenhem et al. (2012), Amoozadeh et al. (2015), and Zhao and Sun (2013)), AVs are capable of forming

1 Should large enough numbers of AVs enter the highway, one could assign multiple AV lanes. We consider this case in Appendix E.5, while focusing our analysis on designating one lane to AVs.
larger platoons than HVs, and the intraplatoon headway tends to be smaller for AVs than HVs. These two benefits arise because AVs can communicate with each other, and also move and brake more smoothly than HVs. As a result, there is a common belief that AVs will increase the capacity of highways, without constructing new lanes or increasing legal speed limits.

Considering these potential benefits of AVs, we analyze the effect of AVs on highway congestion under the D policy and the I policy. Specifically, we answer the following questions:

(1) When is it optimal to use the D policy over the I policy?

(2) How much will AVs improve highway traffic flow under each of these two policies?

To answer these questions we model traffic flow on a highway segment as an $M/G_n/c/c$ queueing system. Based on this model, we compare the mean travel time of a single vehicle as well as the throughput of the highway under each of the two policies, and also against a benchmark case in which all the vehicles are HVs. In our queueing model, vehicles arrive individually to the highway segment, and the service time of a vehicle is defined as the amount of time it takes to travel the segment. This travel time depends on congestion – the number of vehicles ($n$) that are simultaneously using the highway segment (i.e., the state of the queueing system) – so it is state-dependent. The queue capacity $c$ of the highway is defined as the number of vehicles that it can accommodate at saturation, i.e., when the traffic forms a jam. We use a Markov Arrival Process (MAP) to capture the platooning process; within a semi-renewal framework, a MAP enables us to describe the intraplatoon and interplatoon headways as well as the size of platoons. A platoon consists of only HVs, either AVs or HVs, or a mix of AVs and HVs in the benchmark case, under the D policy, and under the I policy, respectively. The difference in the mix of vehicle fleets leads to different service rates under different policies.

We calibrate our models to data, and evaluate our policies both analytically and numerically. Our analysis shows that, in terms of mean travel time, while the I policy improves the performance of the highway over the benchmark case, the D policy outperforms the benchmark case only when the highway is congested and AVs constitute a significant proportion of the vehicles. This calls into question the industry proposals in Bierstedt et al. (2014), as well as the policies being considered in Colorado, Wisconsin and Washington.2

In terms of throughput, we show that the performance of the D and I policies depends on the arrival rate to the highway and the AV proportion. For a lightly loaded highway, the throughput under the D policy is at most as high as that of the benchmark case, while the I policy increases the throughput over the benchmark case negligibly (by less than 5%). For a highly loaded highway,

---

2 This paper focuses on the potential effects of employing the D and I policies on congestion. We do not consider other possible benefits (e.g., fuel consumption, environments, safety), and concerns (e.g., confusion of human drivers under the I policy) of AVs.
both policies are capable of increasing the throughput over the benchmark case: Under the D policy and the I policy, a 30% increase in throughput is achievable when the AV proportion is 0.24 and 0.22, respectively. This implies that for highly congested highways, a moderate number of AVs can make substantial improvement. Finally, when all vehicles are AVs, the throughput of one lane of the highway is increased by 437% over the benchmark case. Thus, the bold prediction of KPMG that the platooning of AVs could increase throughput by 500% (Albright et al. 2015) is not entirely unrealistic.

We provide suggestions to policy makers about when and under what conditions each of the D and I policies should be employed. Based on our analysis, if the mean travel time of vehicles is of more importance, the I policy should be employed. If a policy maker bases a decision primarily on throughput in a congested highway, then for moderate values of the AV proportion, the D policy is recommended, otherwise, the I policy should be used. Specially, in our calibrated model, only when the AV proportion is between 0.25 and 0.55 does the D policy result in a higher throughput than the I policy in a congested highway. However, under the I policy, the mean travel time is lower than that under the D policy for all values of the AV proportion.

The rest of this paper is organized as follows. In §2, we review the related literature. Two policies with AVs as well as a benchmark case for highway traffic flow are presented in §3. In §4 we calibrate the model to data, and in §5, we compare the policies with AVs with the benchmark case. Finally, §6 provides our conclusion.

2. Related Literature

Our work is related to three streams of research: smart city operations, highway traffic flow modeling and platooning of vehicles, and autonomous vehicles.

This paper contributes to the expanding literature on smart city operations. Ride-sharing and electric vehicles (EVs) are both well-studied areas under this body of literature. Ride-sharing literature covers a broad range of topics from equilibrium models of peer-to-peer vehicle sharing (Benjaafar et al. 2018) to use of ride-sharing in last-mile delivery (Qi et al. 2018). Prior work on EVs includes business models for mass adoption of EVs (Lim et al. 2014) and infrastructure planning for EVs (Mak et al. 2013). In addition, the intersection of ride-sharing and EVs is investigated by He et al. (2017). For a comprehensive review of ride-sharing and electric vehicles, we refer readers to He et al. (2018) and Pelletier et al. (2016), respectively.

To model highway traffic flow, a variety of queueing models have been used. Bell (1980) studies vehicle queueing at traffic signaled junctions by using simulation. Kuwahara and Newell (1987)

---

3 Another stream of research investigates the movement of two consecutive vehicles at a microscopic level; see Brackstone and McDonald (1999) for a review of car-following models, in which the variables of the model represent microscopic properties, e.g., the position of a single vehicle.
use a non-stationary queueing network to model traffic flow into a core city. Heidemann (1996) models an uninterrupted traffic flow as the stationary queueing models $M/M/1$ and $M/G/1$. Jain and Smith (1997) use an $M/G_n/c/c$ model in which state $n$ is defined as the number of vehicles simultaneously driving on the highway. In their model, state-dependent service rates account for the effect of congestion on the speed of vehicles, but vehicles never form platoons. Vandaele et al. (2000) study the transient behavior of $M/M/1$, $M/G/1$ and $GI/G/1$ queues with or without state-dependent service times in traffic modeling. The validity of an $M/G/c$ queueing model is shown by Van Woensel and Vandaele (2006) and Van Woensel et al. (2006). Van Woensel and Vandaele (2007) provide a review of different queueing models for traffic on highways. In this paper, we follow Jain and Smith (1997) by using an $M/G_n/c/c$ queueing model, augmented with an incorporation of platooning.

Platoon formation in the case of interrupted traffic flow (e.g., when vehicular motion is interrupted by stoppages such as traffic lights) is studied in Dunne (1967), Lehoczky (1972) and Daganzo (1994). Neuts and Chakravarthy (1981) discuss a continuous-time MAP for platoon formation on highways. A MAP was first introduced by Neuts (1979) as a versatile Markovian point process; he describes it as an extension of a Poisson process. Later Lucantoni (1991) provides a more convenient notation for the MAP. Alfa and Neuts (1995) show that the MAP is a valid model for platoon arrivals to a highway, but they do not provide a queueing model to examine traffic flow; Alfa (1995) uses the MAP to model traffic flow at signalized intersections. Breuer and Alfa (2005) present a procedure for estimating the parameters of the MAP. In this paper, we use the MAP to model the formation of platoons as vehicles drive on highways. We then use the headway between vehicles derived from the MAP, to calculate the state-dependent service rates of the $M/G_n/c/c$ queueing system, coupling the two models.

Research on autonomous vehicles is nascent, but growing fast. This literature covers a broad range of topics related to AVs; those studies that investigate the effects of AVs on traffic flow are particularly related to our paper. Qom et al. (2016) and Talebpour et al. (2017) conduct simulation studies investigating the effect of designating a lane to AVs on throughput. Chen et al. (2017) develop an analytical model to show that segregating AVs and HVs leads to a smaller improvement in the capacity of the highway than mixing them. Different from our paper, they assume that the number of AVs entering the highway is fixed, and that the headway between vehicles is deterministic, thus ignoring the effect of congestion on traffic flow. For a mixed fleet of AVs and HVs, several papers have studied the effects of adding some preliminary autonomous features to vehicles on throughput. For example, adaptive cruise control (ACC) and cooperative ACC (CACC) have been studied in multiple papers by California Partners for Advanced Transportation Technology
(PATH), including Shladover et al. (2012) and Liu et al. (2018). Shladover et al. (2012) is the first work that uses experimental data to study the effects of ACC and CACC on traffic. They find that ACC does not have a remarkable impact on highway capacity; however, if a moderate to high percentage of vehicles adopt CACC, a significant increase in capacity is expected. In contrast, Stern et al. (2018) show through a field experiment that congestion can be eradicated by only a few AVs (1 out of 21 vehicles or about 5%). There exist several other experimental and simulation analyses of mixed traffic; see Liu et al. (2018) and references therein. In our paper, as in Shladover et al. (2012), AVs are equipped with CACC. We find that although a small proportion of AVs can have a substantial effect on highway performance, the result obtained by Stern et al. (2018) is overly optimistic: to fully eradicate congestion, a substantial number of AVs are typically needed.

Ghiasi et al. (2017) is the closest work to our paper. They model the platoon structure of a mixed fleet of AVs and HVs driving on a one-lane highway segment, using a Markov chain. In their model, the arrival process to the highway is a vehicle stream of a fixed length, where any number of consecutive AVs in this stream can form a platoon, but no HV can be a part of a platoon. They show that, when the mean headway between an HV and an AV is lower than that between two HVs and higher than that between two AVs, the throughput of the highway increases with the AV proportion. Our model is more general than Ghiasi et al. (2017) in several respects. First, the arrival of vehicles to the highway follows a stochastic process, so the number of vehicles on the highway is not necessarily fixed. Second, the headway between two vehicles in our model not only depends on their vehicle types (i.e., HV-HV, HV-AV, AV-HV, or AV-AV), but also on the number of vehicles simultaneously present on the highway (i.e., state of our queueing system). As a result, the speed of a vehicle on the highway is impacted by all other vehicles. Third, we consider a multi-lane highway, which incorporates the effects of lane changing in the state-dependent speed of vehicles. Our data indicate that, not surprisingly, the speed of vehicles on a 3-lane highway is higher than that on a 2-lane highway due to higher chance of lane changing. In addition, our multi-lane traffic model enables us to compare the performance of the D and I policies. Finally, our richer model yields different results than Ghiasi et al. (2017): for example, our result indicates that an integrated fleet of AVs and HVs can improve the throughput of the highway under more general conditions than those indicated by Ghiasi et al. (2017).

In summary, our paper presents the first queueing model for a multi-lane highway with AVs. Our model captures several realistic features of highways, such as stochastic headway between vehicles, state-dependent speed of vehicles, stochastic arrival of vehicles to the highway, and mixed platoons of AVs and HVs. Whereas prior papers, except Chen et al. (2017), focus on either the D policy or I policy, we compare these two policies to provide a guideline for policy makers. Although prior
studies measure the impact of AVs by throughput, we show that a policy which results in a higher throughput does not necessarily have a lower mean travel time.

3. Model

In §3.1 we present a general model for highway traffic flow. In §3.2, we adapt this model to a benchmark and two policies with AVs. Table A1 in the Appendix summarizes our notation.

3.1. The Highway Traffic Flow Model

To model traffic flow on a highway, we consider an $M/G/c/c$ queueing system with state-dependent service times. This queueing system is also known as an $M/G_n/c/c$ queue, where $n$ is the state of the system, and it is defined as the number of vehicles simultaneously driving on the highway. In our model, vehicles arrive individually to the highway according to a Poisson process with rate $\lambda$, and they form platoons while traveling on the highway (i.e., while receiving service). We use a MAP to make a connection between platooning and the service rate of the queueing system. Although different vehicle types (e.g., trucks, sedans, SUVs, etc.) travel on a highway, we assume for simplicity that all vehicle types are identical. In §3.1.1 we first describe the queueing system, and in §3.1.2 we explain how we use our MAP to model the formation of platoons. In §3.1.3 the impact of platooning on service time is illustrated.

3.1.1. The Queueing System

We restrict our attention to a segment of length $L$ in a highway with $N$ lanes. The capacity per mile of each lane is called jam density $J$, and it is defined as the maximum number of vehicles per mile per lane of the highway; Once $J$ is reached flow comes to a jam, at which point vehicles travel at minimal speed. The capacity of the entire queueing system is $c = J \times L \times N$. We assume a vehicle that finds $c$ other vehicles on the segment upon its arrival turns away, possibly taking an alternative route. This is the standard assumption in prior literature, but it also reflects today’s reality that drivers may take alternative local routes suggested by navigation systems or apps (e.g., Google Maps) when highways are extremely congested.

Service is defined as the travel time of a single vehicle from the beginning of the highway segment to the end of the highway segment. The speed of a vehicle, and hence its travel time, depends on the number of vehicles present on the highway (i.e., the state of the queueing system): a vehicle travels freely on the highway in the absence of other vehicles, but as the highway becomes more crowded, a vehicle tends to drive at a lower speed. We use the state-dependent speed as a measure of a service rate. Let $V_n$ be the mean speed when there are $n$ vehicles on the highway. We assume
$V_n$ is decreasing in $n$, and $V_n = 0$ for $n \geq c + 1$. When only a single vehicle drives on the highway, the mean speed $V_1$ is called the free-flow speed.\footnote{Ideally, the free-flow speed should be equal to the speed limit. Yet, our data indicate that the free-flow speed is close to, but not equal to the speed limit. This is also observed in prior literature.}

As a vehicle enters the highway, it immediately occupies a server and starts receiving service (i.e., there is no waiting time). As a result, the number of servers is equal to the maximum possible number of vehicles that are traveling on the highway, i.e., the capacity $c$ of the highway.

### 3.1.2. The Platooning Process

To capture the platooning effect, we use a MAP.\footnote{Since we analyze a steady-state queueing system, platooning is modeled also in steady-state.} A MAP is defined by two $m \times m$ matrices $C^1$ and $C^0$. The matrix $C^0$ (resp., $C^1$) is associated with the rate of transitions to non-absorbing states (resp., the absorbing state). The matrix $C = C^1 + C^0$ is the irreducible generator matrix of the MAP. The steady state distribution of the MAP, $\tilde{\pi}$, satisfies $\tilde{\pi}C = 0$ and $\tilde{\pi}1 = 1$, where $1$ is a vector of ones. The mean of a MAP is calculated as

$$h_n = \frac{1}{\tilde{\pi}_n C^1_n 1} \quad \text{(in unit of time)},$$

where $C^1_n$ and $\tilde{\pi}_n$ represent the transition rate matrix to the absorbing state, and the steady state distribution of the MAP, when there are $n$ vehicles on the highway, respectively. The value $h_n$ represents the mean headway (i.e., the mean time gap between any two consecutive vehicles on the highway) when there are $n$ vehicles on the highway.

To model platooning on a highway using a MAP, one needs to specify the distributions of the following three elements: (1) the size of each platoon, (2) the time gap between two consecutive vehicles traveling in the same platoon (“intraplatoon headway”), and (3) the time gap between the last vehicle of one platoon and the first vehicle of the following platoon (“interplatoon headway”).

We model the size of a platoon using a discrete phase type (PH-type) distribution of order $l$. A distribution on $1, 2, \cdots, l$ is a discrete phase-type distribution if it is the distribution of the first passage time to the absorbing state of a Markov chain with $l$ states, such that state $l$ is absorbing and the rest of the states are transient. Any discrete distribution can be written as a PH-type distribution. This distribution is represented by $(\delta^0, G^0)$. The vector $\delta^0$ corresponds to the probability of starting at the non-absorbing states $1, 2, \cdots, l - 1$. Similarly, $\delta$ is the probability of starting at the absorbing state $l$. The vector $(\delta, \delta^0)$ represents the initial distribution of states, where $\delta + \delta^0 1 = 1$. The $(l - 1) \times (l - 1)$ matrix $G^0$ is the probability matrix associated with non-absorbing transitions (transitions among non-absorbing states). Analogously, $G$ corresponds to transitions to the absorbing state, satisfying $G^0 1 + G = 1$, where $1$ is a vector of ones. The transition probability
matrix of this discrete-time Markov chain (DTMC) is then \( M = [G^0 G] \). We present the example of a geometric distribution in §3.2, while presenting another example of a uniform distribution on \( 1, 2, \cdots, l \) in Appendix B.

We model the intraplatoon headway, when there are \( n \) vehicles in the system, using a continuous PH-type distribution of order \( m_1 \) represented by \((\alpha^0_n, Q^0_n)\), having mean \( 1/\xi_n \). The vector \( \alpha^0_n(1) \) (resp., \( \alpha_n(1) \)) demonstrates the initial distribution of the non-absorbing states (resp., the absorbing state). The matrices \( Q^0_n \) and \( Q_n \) represent a non-absorbing transition rate matrix and an absorbing transition rate matrix, respectively.

Lastly, we also model the interplatoon headway, when there are \( n \) vehicles in the system, using a continuous PH-type distribution of order \( m_2 \) represented by \((\alpha^0_n, Q^0_n)\), and with mean \( 1/\psi_n \).

Having these three elements specified, the platooning process of a single lane can be characterized as a MAP with the following matrices:

\[
\begin{align*}
C^0_n &= \begin{bmatrix} Q^0_n(2) & 0 \\ 0 & I \otimes Q^0_n(1) \end{bmatrix}, \quad C^1_n = \begin{bmatrix} \delta Q^0_n(2)\alpha^0_n(2) & \delta \otimes Q_n(2)\alpha^0_n(1) \\ G \otimes Q_n(1)\alpha^0_n(2) & G \otimes Q_n(1)\alpha^0_n(1) \end{bmatrix}.
\end{align*}
\] (2)

The size of these matrices is \( m \times m \), where \( m = m_2 + m_1(l-1) \). See Example 2 in Appendix B.

For platoon arrivals to a highway with \( N > 1 \) lanes, we consider a MAP with matrices \((C^0_n, C^1_n)\), and assume that a vehicle joins one of the lanes with probability \( 1/N \). The platoon formation process on a specific lane itself is a MAP with matrices \((C^0_n + (1 - \frac{1}{N})C^1_n, \frac{1}{N}C^1_n)\). This is called the random thinning of a MAP (see Proposition 2.2.3 in He (2014) for more details).

### 3.1.3. The Effect of Platooning on a Service Rate

Platooning of vehicles affects the service rate of the queueing system through mean headway. We now derive the relationship between platooning, headway and service rate (vehicle speed).

A highway traffic stream is characterized by three factors: speed, density and flow. Speed or velocity \( V_n \) is in miles per unit time, traffic density \( k = \frac{n}{NL} \) is defined as the number of vehicles per unit distance, and traffic flow \( q \) is defined as the rate (in vehicles per unit time) at which vehicles travel through some designated roadway point. These three measures are related according to \( q = V_n k = \frac{nV_n}{NL} \), and the mean headway \( h_n \) is equal to the inverse of flow \( q \). Thus,

\[
\frac{1}{h_n} = \frac{nV_n}{NL}.
\] (3)

For each \( n \), once we compute \( h_n \) from (1) (which depends on platoon characteristics such as platoon size, and interplatoon and intraplatoon headways), we can derive the speed \( V_n \) from (3).
3.2. Models with a Specific Fleet Composition

We adapt the model described in §3.1 to a benchmark case and two policies with AVs. In §3.2.1, we specify the benchmark case which depicts the current situation where all vehicles on highways are HVs. In §3.2.2 we present a model that assigns a specific lane to AVs, the designated-lane (D) policy. In §3.2.3 we develop a model in which AVs and HVs are allowed to use any lanes, the integrated (I) policy. Let $p$ denote the proportion of AVs in the latter two models. In the rest of this paper, we use superscripts $B, I, D, DA$ and $DH$ that represent the benchmark case, the I policy, the D policy, and the AV queue and the HV queue of the D policy, respectively.

3.2.1. The Benchmark Case

To characterize the platooning process for the benchmark case, we specify the platoon size, the intraplatoon headway, and the interplatoon headway, respectively. The platoon size follows a geometric distribution with mean $1/\delta^B$. We can represent this distribution as a discrete PH-type distribution of order $l = 2$, where $G^0 = \delta^0 = 1 - \delta$. The intraplatoon headway follows an exponential distribution with mean $1/\xi^B_n$, and restating this distribution as a continuous PH-type distribution of order $m_1 = 1$, we let $Q^0_n(1) = -Q_n(1) = -\xi^B$ and $\alpha^0_n(1) = 1 - \alpha_n(1) = 1$. Similarly, the interplatoon headway follows an exponential distribution with mean $1/\psi^B_n$, so $Q_n^0(2) = -\psi^B$ and $\alpha^0_n(2) = 1$.

By forming the matrices $C^0_n$ and $C^1_n$ based on equation (2), the mean headway of vehicles in the benchmark case is determined by (1) as $h^B_n = \delta^B/\psi^B_n + (1 - \delta^B)/\xi^B_n$. Thus, according to (3), $V^B_n$ can be represented as a function of the platoon characteristics as follows:

$$V^B_n = \frac{NL}{n} \frac{\xi^B_n \psi^B_n}{\delta^B \xi^B_n + (1 - \delta^B) \psi^B_n}. \quad (4)$$

3.2.2. D Policy

In this model, a vehicle entering a highway segment is an AV with probability $p$, and this vehicle must use the designated lane. HVs are allowed to use all lanes except the one designated to AVs. As a result, we can consider two independent queueing systems: an AV queueing system and an HV queueing system.

The HV queue is similar to that of the benchmark case, except it has one fewer lane. The capacity of this queue is equal to $JL(N - 1)$, and the arrival rate is $(1 - p)\lambda$. As in the benchmark case, the platoon size follows a geometric distribution with mean $1/\delta^{DH}$, the intraplatoon headway follows an exponential distribution with mean $1/\xi^{DH}_n$, and the interplatoon headway follows an exponential distribution with mean $1/\psi^{DH}_n$. These parameters may be different from those in the benchmark case, since the mean speed in an $N$-lane highway is usually higher than the mean speed in an $(N-1)$-lane highway.
The AV queue is also modeled as an $M/G_n/c/c$ queueing system described in §3.1. In this system, the number of lanes is one, and the capacity is equal to $JL$. According to the Poisson splitting, the arrival rate is $p\lambda$. In this queue, the platoon size follows a geometric distribution with mean $1/\delta^{DA}$. The intraplatoon and interplatoon headways follow exponential distributions with mean $1/\xi^{DA}_n$ and $1/\psi^{DA}_n$, respectively.

### 3.2.3. I policy

We consider an $M/G_n/c/c$ queueing system with arrival rate $\lambda$, where a proportion $p$ of the vehicles are AVs. Under the I policy, a mixed fleet of AVs and HVs form platoons. Following Liu et al. (2018), we assume that HVs are equipped with Vehicle Awareness Devices (VAD) that enable them to jointly form platoons with AVs. We characterize the platooning process as follows. First, the platoon size follows a geometric distribution with mean $\delta^I$. Second, the intraplatoon headway follows a hyperexponential distribution with mean $1/\xi^I_n$; the intraplatoon headway depends on the types of vehicles following each other. For example, if an AV follows another AV, since they can communicate with each other and also move/brake more smoothly than HVs, the intraplatoon headway between them is lower than the intraplatoon headway between two HVs. There exist four possible pairs of vehicles: AV-AV (denoted by $i = IAA$), AV-HV ($i = IAH$), HV-AV ($i = IHA$), and HV-HV ($i = IHH$); where X-Y means vehicle X is followed by vehicle Y. Since AVs constitute a proportion $p$ of vehicles, the probability of observing each of those four pairs is $p^2$, $p(1-p)$, $(1-p)p$, and $(1-p)^2$, respectively. Assuming the intraplatoon headway follows an exponential distribution with rate $\xi^i_n$ for the pair $i \in \{IAA, IAH, IHA, IHH\}$, the overall intraplatoon headway follows a hyperexponential distribution. By linearity of expectation, the mean intraplatoon headway is calculated as

\[
1/\xi^I_n = p^2/\xi^I_{nAA} + p(1-p)/\xi^I_{nAH} + (1-p)p/\xi^I_{nHA} + (1-p)^2/\xi^I_{nHH} \text{ for } n = 1, 2, \cdots, c. \tag{5}
\]

Lastly, similar to the intraplatoon headway, the interplatoon headway also follows a hyperexponential distribution with parameters $\psi^i_n$, where $i \in \{IAA, IAH, IHA, IHH\}$. The mean interplatoon headway is as follows

\[
1/\psi^I_n = p^2/\psi^I_{nAA} + p(1-p)/\psi^I_{nAH} + (1-p)p/\psi^I_{nHA} + (1-p)^2/\psi^I_{nHH} \text{ for } n = 1, 2, \cdots, c. \tag{6}
\]

Having the platoon characteristics specified, similar to the benchmark case, we are able to calculate the mean headway $h^I_n$, and then represent $V^I_n$ as a function of $h^I_n$.

### 4. Model Calibration

In this section we calibrate our queueing model to data. The arrival rate to a highway can differ depending on the highway location, but the speed of vehicles primarily depends on the number of
vehicles currently driving on the highway and the headway between them, so we focus on estimating the steady-state speed of vehicles in our models. Without loss of generality, we assume $L$ is equal to one mile. Then, we estimate $V_n$ for $n = 1, 2, \cdots, c$, where $c = NJL = NJ$ is the capacity of the highway segment with $N$ lanes and a jam density of $J$. We use $J = 185$, since $J$ is typically between 185 and 250 (Holtzman and Goodman (2012) and Wang et al. (2010)). Thus, the capacities of the benchmark case and I policy with $N = 3$ are equal to 555, the capacity of the HV queue in D policy with $N = 2$ is 370, and the capacity of the AV queue in D policy with $N = 1$ is 185.

In the rest of this section, we first use the data collected from a highway in Arizona to estimate $V_n$ in the benchmark case. Then, we discuss the speed estimation for each of the HV queue and the AV queue under the D policy. For the estimation of $V_n$ in the HV queue, we also use the data from Arizona. For the AV queue we propose a procedure to estimate $V_n$ using several parameters that reflect an AV’s driving performance; since a designated lane for AVs has yet to be implemented in reality, $V_n$ cannot be directly estimated from data. Lastly, $V_n$ is estimated under the I policy using a procedure similar to the AV queue under the D policy.

4.1. The benchmark case
We estimate the state-dependent speed $V_n^B$ based on the data from Arizona Department of Transportation. Our data include about 10,000 instances of 5-minute average volume and speed of vehicles collected from the segments of Interstate 10 (I-10) with three lanes ($N = 3$) in January 2017. Figure 1(a) shows a scatter plot of this data as well as its fitted curve given as follows:

$$V_n^B = 70e^{-\frac{n^2}{21.049}} + 4.7 \text{ (miles/hour)} \text{ for } n = 1, 2, \cdots, 555. \quad (7)$$

Appendix D.1 provides more information about how we estimate this curve.

4.2. D policy
As mentioned in §3.2.3, under this policy, the highway is divided into two queues: (1) the HV queue with two lanes, and (2) the AV queue with one lane.

4.2.1. The HV queue This queue is similar to that in the benchmark case, except it has one fewer lane. Using about 16,000 data points collected from state route 101 in Arizona with two lanes\(^7\) in January 2017, we estimate the state-dependent speed of the HV queue $V_n^{DH}$ (see Figure 1(b)), which is expressed as follows:

$$V_n^{DH} = 66e^{-\frac{n^{3.4}}{21.999}} + 2 \text{ (miles/hour)} \text{ for } n = 1, 2, \cdots, 370. \quad (8)$$

\(^6\) Considering several different values of $J$, we observe that our results continue to hold.

\(^7\) All the segments of I-10 that we use in our benchmark case have three lanes. Thus, we use data from the 2-lane state route 101 which has the same speed limit as the 3-lane segments of I-10.
By comparing Figures 1(a) and 1(b), we observe that $V_{DH}^n < V_B^n$ for any given $n$. Intuitively, when the same number of vehicles travel on two different segments, the number of vehicles per lane is lower for the segment with more lanes, and vehicles are able to drive faster on this segment.

4.2.2. The AV queue We use a MAP to estimate the state-dependent speed of the AV queue, $V_n^{DA}$. Equation (3) relates $V_n^{DA}$ to the mean headway $h_n^{DA}$ of vehicles derived from the MAP. In this queue $N$ is equal to one and $L$ is one mile, so $V_n^{DA} = \frac{1}{nh_n^{DA}}$, where $h_n^{DA}$ is derived from equation (1); it is a function of mean platoon size, mean intraplatoon headway, and mean interplatoon headway. Therefore, to estimate $V_n^{DA}$, we need to estimate these three parameters.

First, we specify the mean platoon size. All the previous papers that consider platooning of AVs assume a fixed platoon size: for example, Amoozadeh et al. (2015) and Liu et al. (2018) consider platoons of size 10, and Zhao and Sun (2013) consider platoons of size 6. We take into account randomness in platoon sizes by using a geometric distribution with the same mean value as in Amoozadeh et al. (2015) and Liu et al. (2018), i.e., $1/\delta^{DA} = 10$ vehicles. See Appendix D.2 for more details about the estimation of a platoon size distribution.

Second, following Amoozadeh et al. (2015), Vander Werf et al. (2002), and Zhao and Sun (2013), we set the mean intraplatoon headway of AVs equal to 0.55 seconds, i.e., $1/\xi_n^{DA} = 0.55$ seconds for all $n = 1, 2, \cdots, 185$, because the intraplatoon headway of AVs does not depend on congestion or speed.\(^8\) (Siciliano and Khatib 2016).

\(^8\) This property also holds for the intraplatooon headway of HVs. According to Virginia DMV (2016) and Tientrakool et al. (2011), the intraplatooon headway for HVs reflects the reaction time of human drivers. This reaction time depends neither on speed nor on the congestion caused by a high number of vehicles in a highway segment. The interplatooon headway of both AVs and HVs, however, depends on $n$. This headway is what vehicles maintain to avoid instant crashes, so it depends on how fast they are moving and how congested the highway is.
Figure 2  Comparison of the state-dependent speeds: the benchmark case vs. the D policy.

Note: The capacities of the AV queue, the HV queue, and the benchmark case are different (185, 370, and 555, respectively), and the speed in each of these queues is shown as a function of the number of vehicles per lane.

Lastly, we estimate the mean interplatoon headway of AVs. According to Guzzella and Kiencke (1995) and Bergenhem et al. (2012), the interplatoon headway of AVs is set equal to the safe stopping time to avoid chain-reaction crashes. In other words, the interplatoon headway is set such that if one platoon of AVs crashes, the following platoon has enough time to stop before hitting the crashed platoon. We estimate this value by using data from National Highway Traffic Safety Administration (NHTSA 2015) as follows (see Appendix D.3 for more details):

$$1/\psi_n^{DA} = 3,600(0.001 - 0.006/V_n^{DA}) \text{ (seconds) for } n = 1, 2, \cdots, 185.$$  

Having characterized the platooning process, we are able to calculate the mean headway between vehicles. For a MAP with geometric(δ^{DA}) platoon size, exp(ξ^{DA}) intraplatoon headway, and exp(ψ^{DA}) interplatoon headway, by equation (1) we get

$$h_n^{DA} = \frac{\delta^{DA}}{\psi^{DA} n} + \frac{(1-\delta^{DA})}{\psi^{DA} n}$$

(see Appendix B for details); substituting the values of δ^{DA}, ξ^{DA}, and ψ^{DA} into

$$V_n^{DA} = \frac{1}{nh_n^{DA}}$$

we obtain after simplifications

$$V_n^{DA} = \min(74.7, \frac{3,600(0.001 - 0.006/V_n^{DA})}{0.855n})$$

where 74.7 miles per hour is the free flow speed under the D policy. Figure 2 illustrates $V_n^{DA}$ and compares it with $V_n^B$ and $V_n^{DH}$. We can observe that the speed of AVs is always higher than HVs. For example, when $n/N = 185$, HVs move slowly at about 4 miles per hour in the benchmark case, but AVs have a smooth flow of platoons moving at 25 miles per hour. Based on our analysis, we assume that the jam speed in an N-lane highway, $V_{NC}^B$, is higher than that in an (N−1)-lane highway, $V_{(N−1)c}^{DH}$, in the rest of this paper.

The maximum speed of AVs must be capped, because otherwise the speed keeps increasing as n decreases. For ease of comparison, we set the free flow speed of the highway under the D and I policies equal to that of the benchmark case (which is obtained by setting $n = 1$ in (7)).

Figure 2 also compares $V_n^B$ and $V_n^{DH}$. The speed pattern of a 3-lane highway ($V_n^B$) differs slightly from that of a 2-lane highway ($V_n^{DH}$). When the number of vehicles per lane is very low (i.e., $n/N \leq 15$), vehicles drive at the
4.3. I policy

Similar to the AV queue under the D policy, we estimate the state-dependent speed $V_I^t$ by employing the mean headway of vehicles derived from a MAP. For any value of $p \in [0, 1]$, we characterize the MAP by first specifying the mean platoon size, and then jointly describing the mean intraplatoon headway and the mean interplatoon headway under this policy.

We first estimate the mean platoon size as a function of $p$. Since AVs are able to move/brake more smoothly and react faster than HVs, they are more likely to form long platoons. As before, we assume that the platoon size follows a geometric distribution with mean $1/\delta_I^t$. When all vehicles are AVs (i.e., $p = 1$), as in the AV queue under the D policy, we use $1/\delta_I^t = 10$ vehicles. On the other hand, when all vehicles are HVs (i.e., $p = 0$), we use $1/\delta_I^t = 1.5$ vehicles, which is obtained from the data used for the benchmark case. For any $p \in (0, 1)$, we assume $1/\delta_I^t = \frac{3}{2-1.7p}$, so that the mean size of platoons increases with $p$.\footnote{Note that $\delta$ is the probability that a vehicle forms a new platoon instead of joining the very last one. This probability is equal to $1/1.5 = 2/3$ for an HV, and $1/10$ for an AV. Thus, under the I policy, when the proportion of AVs is $p$, the probability of forming a new platoon is $\delta_I^t = \frac{2}{3}(1-p) + \frac{1}{10}p = \frac{2-1.7p}{3-1.7p}$, and $1/\delta_I^t = \frac{3}{2-1.7p}$.}

Next we proceed to describe the mean intraplatoon and interplatoon headways. Recall that a pair of consecutive vehicles can be of four different types: AV-AV, HV-AV, HV-HV, and AV-HV. For each of these types, we estimate the mean intraplatoon headway and the state-dependent mean interplatoon headway separately.

- **AV-AV:** Similar to the AV queue in the D policy, we set the mean intraplatoon headway $1/\xi_{IAA}^t = 0.55$ seconds, and estimate the mean interplatoon headway $1/\psi_{IAA}^t$ from equations (1), (3) and (9) as follows: $1/\psi_{IAA}^t = \frac{3.600N-6n(1-\delta_I^t)/\xi_{IAA}^t}{1.000N+6n\delta_I^t}$. Substituting $N = 3$ lanes, $1/\delta_I^t = 10$ vehicles, and $1/\xi_{IAA}^t = 0.55$ seconds, $1/\psi_{IAA}^t = \frac{10.800-2.97n}{3.000+0.6n}$ seconds for $n = 1, 2, \cdots, 555$.\footnote{Note that $\delta$ is the probability that a vehicle forms a new platoon instead of joining the very last one. This probability is equal to $1/1.5 = 2/3$ for an HV, and $1/10$ for an AV. Thus, under the I policy, when the proportion of AVs is $p$, the probability of forming a new platoon is $\delta_I^t = \frac{2}{3}(1-p) + \frac{1}{10}p = \frac{2-1.7p}{3-1.7p}$, and $1/\delta_I^t = \frac{3}{2-1.7p}$.}

- **HV-AV:** Following Zhao and Sun (2013), we set the mean intraplatoon headway $1/\xi_{IHA}^t = 1.4$ seconds: Due to lack of communication between AVs and HVs, an AV maintains a longer intraplatoon headway from an HV than another AV. Similar to the AV-AV pair, we can then estimate the interplatoon headway of this pair as $\psi_{IHA}^t = \frac{10.800-7.56n}{3.000+0.6n}$ seconds for $n = 1, 2, \cdots, 555$.

- **HV-HV:** Tientrakool et al. (2011) state that on average two HVs maintain a longer intraplatoon headway from an HV than another AV. Similar to the AV-AV pair, we can then estimate the interplatoon headway of this pair as $\psi_{IHH}^t = 1.1$ seconds for $n = 1, 2, \cdots, 555$. Note that this value is lower than that of an HV-AV pair. The mean interplatoon headway of HVs is derived free-flow speed, which is higher for a highway with more lanes. At moderately low values of $n/N$ (i.e., $15 < n/N < 55$), the highways are not congested and vehicles are able to still drive fast, so there is not much difference between 3-lane and 2-lane highways. As the highways become more congested (i.e., $n/N \geq 55$), speed decreases on both highways. In this case, if one lane is moving slowly, some of the vehicles on this lane may switch to the other lanes, and maintain a higher speed. The chance of lane-changing is higher for a 3-lane highway than a 2-lane one. Thus, vehicles drive faster on average at a congested highway with more lanes. For further discussion, see Appendix E.2.
by rearranging equation (4) as $1/\psi_n^{IHH} = \frac{\xi_n^{IHH}N^L e^{-(1-\delta_n^{IHH})N^B}}{nV_n^B \xi_n^{IHH}}$ seconds.\footnote{Note that this depends on how human drivers determine their speed based on the characteristics of the highway such as the number of lanes ($N$) and the length of the highway segment ($L$), whereas that of AVs is determined by the safe stopping time.} Substituting $N = 3$ lanes, $L = 1$ mile, $1/\delta_n^{IHH} = 1.5$ vehicles, $1/\xi_n^{IHH} = 1.1$ seconds, and $V_n^B = 70e^{-\frac{n^2}{22000} + 4.7}$ (miles/hour), we obtain

$$1/\psi_n^{IHH} = \frac{10,800 - 0.55n(46.67e^{-\frac{n^2}{22000} + 3.13})}{n(46.67e^{-\frac{n^2}{22000} + 3.13})} \text{ (seconds)}.$$ 

- AV-HV: Following the literature (e.g., Zhao and Sun (2013)), we assume that an HV maintains its intraplatoon and interplatoon headways independent of the type of vehicle in front, so that $1/\xi_n^{IAH} = 1/\xi_n^{IHH} = 1.1$ seconds, and $1/\psi_n^{IAH} = 1/\psi_n^{IHH}$.

As mentioned in §3.2.3, the intraplatoon headway of vehicles in the I policy follows a hyperexponential distribution. Substituting the values of the mean intraplatoon headways for each of the vehicle pairs into equation (5), we obtain the mean intraplatoon headway as follows:

$$1/\xi_n^I = 0.55p^2 + 1.4p(1 - p) + 1.1(1 - p) \text{ (seconds) for } n = 1, 2, \cdots, 555, \text{ and } p \in [0, 1].$$

Similarly, by substituting the values of mean interplatoon headways for each of the vehicle pairs into equation (6), after some simplifications, we have the following for $n = 1, 2, \cdots, 555, \text{ and } p \in [0, 1]$:

$$1/\psi_n^I = p \frac{(10,800 - 7.56n) + 4.59np}{3,000 + 0.6n} + (1 - p) \frac{10,800 - 0.55n(46.67e^{-\frac{n^2}{22000} + 3.13})}{n(46.67e^{-\frac{n^2}{22000} + 3.13})} \text{ (seconds).}$$

Finally, substituting $\delta^I, \xi_n^I$, and $\psi_n^I$ into $V_n^I(p) = \frac{N}{n\xi_n^I} = \frac{N\xi_n^I\psi_n^I}{(1-\delta^I)\psi_n^I}$ (miles/hour), we have the following for $n = 1, 2, \cdots, 555, \text{ and } p \in [0, 1]$:

$$V_n^I(p) = \min\{47.4, \frac{10,800N}{n(1 + 1.7p)} + \frac{10,800(1 - p)}{n(46.67e^{-\frac{n^2}{22000} + 3.13})} - 0.55(1 - p) + [0.55p^2 + 1.4p(1 - p) + 1.1(1 - p)]\}.$$
5. Analysis

In §5.1 and §5.2, we analyze our queueing model under the D policy and the I policy, respectively. In §5.3 we compare the performances of these policies. All proofs are presented in Appendix C.

For these comparisons, we use two quality of service (QoS) measures: throughput ($\theta$) and mean travel time ($W$). To compute these measures, we use the steady-state distribution of an $M/G_n/c/c$ queueing model, $\pi_n$, which is expressed as follows (Cheah and Smith 1994):

$$\pi_n = \left( \frac{(\lambda L)^n}{n!V_n \cdots V_1} \right) \pi_0,$$

where $\pi_0 = (1 + \sum_{n=1}^{c} \frac{(\lambda L)^n}{n!V_n \cdots V_1})^{-1}$. \hspace{1cm} (13)

For a queueing model with finite capacity, throughput is the rate at which vehicles exit the highway, also equal to the effective arrival rate. In our model, a proportion $\pi_c$ of vehicles find $c$ other vehicles on the highway upon their arrival, and turn away. Thus, the effective arrival rate is $\theta = \lambda (1 - \pi_c)$.

Throughput is an important measure to urban designers. Mean travel time of a single vehicle (also known as mean response time) is obtained by Little’s law as follows:

$$W = \left( \sum_{n=1}^{c} n\pi_n \right)/\theta. \hspace{1cm} (14)$$

This measure mostly concerns individual users. Continuing from the previous sections, we use superscript $i \in \{B, DH, DA, D, I\}$ to $\pi_n^i$, $\theta^i$, and $W^i$ to indicate different policies and queues.

In our analysis, a highway is considered to be highly (heavily) loaded, if $\lambda \geq \lambda_{jam}^i$, $i \in \{B, DA, DH, I\}$, where $\lambda_{jam}^i$ is the smallest $\lambda \leq \max_n nV_n$ such that, for a highway with capacity $c$, $\pi_c^i \gg \pi_{[\eta]}^i$ for all $[\eta]$ that satisfy $\eta V_{[\eta]} = \lambda$; see Appendix E.4 for more details about $\eta$. In this case, the arrival rate to the highway is so high that the highway tends to be full. We call $V_{Nc}^B$, $V_{(N-1)c}^{DH}$, $V_c^{DA}$, and $V_{Nc}(p)$ jam speeds, which are the speed of vehicles when a highway with $N$ lanes each
having $c$ capacity is full, in the benchmark, the HV queue of the D policy, the AV queue of the D policy, and the I policy with AVs constituting a proportion $p$ of arrivals, respectively. In addition, we call the throughput of a jammed highway jam throughput, which is equal to the product of the capacity of the system and the jam speed. For example, in the benchmark case, the rate at which a single vehicle leaves the highway is $\mu_{NcV}^c = V_{NcV}^c$, hence the total rate at which vehicles exit the highway, the jam throughput, is $NcV_{NcV}^c$ vehicles per hour.

Our analysis focuses primarily on high values of $\lambda$, because, in general, a policy-maker is more concerned about improving the performance of a highway when it is heavily loaded than when it is lightly loaded. We complement our analytical results with a numerical study that also examines lighter loads.

### 5.1. D Policy

Under the D policy the highway segment is split into two separate queueing systems, as opposed to the benchmark case in which all lanes of the highway are pooled in one queueing system. In general, a pooled server is more efficient than multiple servers of the same total capacity. However, in our setting, although the service capacity is divided into two queues, the service rate in the AV queue is higher than that in the HV queue (see §4.2). Thus, there is a trade-off between pooling the servers and increasing the service rate by designating one lane to AVs. As a result, the pooled server in the benchmark case can be inferior to the split servers under the D policy. The next proposition, which holds for any values of the model parameters described in §3, presents the condition under which each of these factors outweighs the other in terms of $W$ and $\theta$.\(^{13}\)

#### Proposition 1

a) There exist $\lambda^{(D,W)}$ and $\lambda^{(D,W)}$ such that for $\lambda \leq \lambda^{(D,W)}$, $W^D(p) \geq W^B$ for $p \in [0,1]$, and for $\lambda \geq \lambda^{(D,W)}$, $W^D(p) \leq W^B$ if and only if $p \geq p^{(D,W)} = \frac{V_{c}^{DA}V_{Nc}^B - V_{c}^{DA}N_{c}^{DH}(N-1)c}{V_{c}^{DA}V_{Nc}^B - V_{c}^{DA}V_{Nc}^{DH}(N-1)c}$.

b) $\theta^D(p) \geq \theta^B$ if and only if $p^{\pi_{p}^{DA}} \leq N_{c}^{V}c_{V}^{DA} + (1-p)\pi_{p}^{DH}(N-1)c \leq \pi_{NcV}^B$. This can be further simplified for $\lambda \geq \lambda^{(D,\theta)} \equiv \min_{\pi_{p}^{DA}} \max_{\pi_{p}^{DH}(N-1)c} \left\{ \lambda^{DA}_{jam}/p, \lambda^{DH}_{jam}/(1-p) \right\}$, when $\theta^{DH}$ and $\theta^{DA}$ are increasing in $\lambda$: $\theta^D(p) \geq \theta^B$ if and only if $\min_{\pi_{p}^{DA}} \geq \theta^{(D,\theta)} \leq p \leq p^{(D,\theta)}$, where $p^{(D,\theta)} = \left\{ \frac{N_{c}^{V}c_{V}^{DA} + (N-1)c_{V}^{DH}(N-1)c}{\lambda^{[1-\pi_{p}^{DA}(\pi^{DH}(N-1)c)]}} \right\}^{+}$ and $1 - p^{(D,\theta)} = \left\{ \frac{N_{c}^{V}c_{V}^{DA} - e_{V}^{DA}}{\lambda^{[1-\pi_{p}^{DA}(\pi^{DH}(N-1)c)]}} \right\}^{+}$.

---

\(^{13}\) For thresholds of $\lambda$ and $p$, we use superscripts $(i,j)$ for $i \in \{D,I,DI\}$ and $j \in \{W,\theta,W\theta\}$, where $D$, $I$, and $DI$ represent the D policy, the I policy, and comparison between D and I policies, respectively; and $W$, $\theta$, and $W\theta$ indicate thresholds for $W$, $\theta$, and comparison between $W$ and $\theta$, respectively. We also use underscore and overscore to indicate smaller and larger thresholds, respectively. Table A2 summarizes the notation used in this section.

\(^{14}\) For the AV queue, $\theta^{DA}$ is increasing in $\lambda$, if $n^{V^{DA}}$ is increasing in $n$ (see Lemma 2 in the appendix). This assumption guarantees that $V_{c}^{DA}$ decreases in $n$ no faster than linearly. Figure 16(a) in the appendix shows that this property holds for our calibrated model. Similarly, in Lemma 2, we derive the condition for $\theta^{DH}$ to be increasing in $\lambda$. This condition also holds for our calibrated model.
Proposition 1(a) indicates that adding AVs and designating a lane to them decreases $W$ compared to the benchmark case only when both the arrival rate $\lambda$ and the AV proportion $p$ are high. This result cautions against adopting this policy too early, despite recent industry proposals. For example, Bierstedt et al. (2014) propose that at the beginning of the mass appearance of AVs (i.e., when $p$ is relatively low), one lane of highways should be designated for AVs. Likewise, the Departments of Transportation in Colorado, Wisconsin and Washington are considering designated lanes for AVs (Aguilar 2018). When the highway is lightly loaded with $\lambda \leq \bar{\lambda}^{(D,W)}$, designating a lane to AVs may not be beneficial, in this case the D policy potentially slows down vehicles in $(N - 1)$ lanes of the highway segment, while increasing the speed of vehicles in the AV lane only moderately. Even for a heavily loaded highway with $\lambda \geq \bar{\lambda}^{(D,W)}$, dedicating a lane to AVs does not necessarily reduce the mean travel time, unless AVs constitute at least a proportion $p^{(D,W)}$ of vehicles. This threshold depends on the jam speed of vehicles in each queue as well as the characteristics of the highway such as the number of lanes and the capacity of each lane.$^{15}$

For illustration, consider two different values of $\lambda$ in the calibrated model in §4.2: the mean arrival rate of 2,217 vehicles per hour and the maximum arrival rate of 11,342 vehicles per hour to I-10 with three lanes in 2016. The former is considered to be a light load, and the latter is a heavy load, because $\Delta^{(D,W)} = \bar{\lambda}^{(D,W)} = 2,510$ vehicles per hour. When the highway is lightly loaded with $\lambda = 2,217$, we observe that the D policy can increase $W$ (see Appendix E.1). This value of $\lambda$ is so

$^{15}$Although in all our numerical expriments $\Delta^{(D,W)} = \bar{\lambda}^{(D,W)}$, it is theoretically possible that $\Delta^{(D,W)} < \bar{\lambda}^{(D,W)}$, as shown in the proof of Proposition 1(a). When $\Delta^{(D,W)} < \lambda < \bar{\lambda}^{(D,W)}$, unlike the highly loaded case, $W^{D}(p)$ depends on $V_{n}^{DH}$ and $V_{n}^{DA}$ for all values of $n$, and $W^{D}(p)$ can be lower or higher than $W^{B}$. 

\[ \text{Figure 4} \] QoS measures for the D policy when $\lambda = 11,342$ vehicles per hour: (a) mean travel time, and (b) throughput.

Note: “overall” indicates the overall performance of AVs and HVs under the D policy.
low that there is no congestion in the benchmark case, and vehicles are able to drive freely on the highway. However, since the D policy divides the highway into two queueing systems, when \( p \) is low, the HV queue becomes congested. In this case, the large \( W \) of this queue increases the overall \( W \) of the D policy beyond \( W \) of the benchmark case.

At \( \lambda = 11,342 \) vehicles per hour, the highway is heavily loaded.\(^{16}\) For this value of \( \lambda \), if AVs constitute at least \( p^{(D,W)} = 0.63 \) of vehicles, dedicating one out of three lanes to AVs leads to a lower overall mean travel time, \( W^D \), than the benchmark case (see Figure 4(a)). The overall mean travel time under the D policy, \( W^D = pW^{DA} + (1 - p)W^{DH} \), is a weighted average of mean travel time of the AV queue, \( W^{DA} \), and that of the HV queue, \( W^{DH} \). To understand the effect of \( p \) on \( W^D \), we discuss its effect on \( W^{DH} \) and \( W^{DA} \). First, when \( p \leq 0.93 \), the reduction in capacity of the HV queue leads to congestion in this queue, so \( W^{DH} \) is substantially larger than \( W^B = 13 \) minutes. When \( p \geq 0.93 \), the arrival rate to the HV queue becomes so low that HVs do not need to reduce their speed due to congestion. Appendix E.4 provides more details about \( W^{DH} \), and its phase-change-type behavior at \( p = 0.93 \). Next, as \( p \) increases, \( W^{DA} \) increases from 0.8 minutes to 2.3 minutes; both the magnitude and change of \( W^{DA} \) are negligible compared to \( W^{DH} \). Finally, increasing \( p \) raises the weight of \( W^{DA} \), and reduces the weight of \( W^{DH} \). Taken together, as \( p \) increases, \( W^D \) decreases as shown in Figure 4(a). Since \( W^{DH} \) is significantly higher than \( W^B \), more than 63% AVs are needed to bring \( W^D \) below \( W^B \). To have a tangible reduction in \( W^D \), even more AVs are needed; for example, to get 50% reduction in \( W^B \), 80% of vehicles need to be autonomous.

Proposition 1(b) provides conditions under which adding AVs to the highway under the D policy improves the throughput over the benchmark case. The proposition first presents the general condition (i.e., \( p\pi^{DA}_c + (1 - p)\pi^{(N-1)c}_{(N-1)c} \leq \pi^{B}_{Nc} \)) that holds for any \( \lambda \). This condition indicates that having AVs on the highway under the D policy improves the throughput, if and only if the highway blocks fewer vehicles under this policy than in the benchmark case. We can further simplify the condition in terms of the proportion of AVs (\( p \)) for a highly loaded highway with \( \lambda \geq \lambda^{(D,\theta)} \).\(^{17}\) In this case, there exists a unique interval of \( p \) (i.e., \( \underline{p} \leq p \leq \bar{p} \)) in which the D policy improves throughput over the benchmark case. When \( p < \underline{p} \), both the benchmark case and the HV queue are heavily loaded, but the AV lane may still flow relatively freely. If the difference between the jam throughput of the benchmark case and that of the HV queue under the D policy (i.e., \( N\pi^{B}_{Nc} - (N - 1)\pi^{DH}_{(N-1)c} \)) is larger than the throughput of the AV queue (i.e., \( p\lambda(1 - \pi^{DA}_c) \)),

\(^{16}\) According to Varaiya (2005), a highway has an ideal throughput of 2,000 vehicles per hour per lane, so a highway with an arrival rate higher than 2,000 \( \times \) \( N \) vehicles per hour is considered to be highly loaded.

\(^{17}\) For lightly loaded highways with \( \lambda < \lambda^{(D,\theta)} \), consider two cases. First, when \( \lambda \) is so low that \( \pi^{B}_{Nc} = 0 \), then \( \theta^D(p) \) is at most as high as \( \theta^B \). The D policy may even reduce throughput, because \( \theta \) in the benchmark case already reaches the maximum possible \( \theta \) (= \( \lambda \)). Second, when \( \pi^{B}_{Nc} > 0 \), \( \theta^D(p) \) is not necessarily concave, so multiple intervals of \( p \) in which \( \theta^D(p) \geq \theta^B \) may exist.
then the AV queue’s throughput cannot offset this difference and \( \theta \) under the D policy is lower than that of the benchmark case. Similarly, when \( p > \bar{p}^{(D, \theta)} \), the HV queue is not highly loaded, and the throughput of this queue is lower than the difference between the jam throughput of the benchmark case and that of the AV queue. Thus, for a highly loaded highway, in order for the D policy to increase \( \theta \) over the benchmark case, \( p \) should be moderate, i.e., \( p^{(D, \theta)} \leq p \leq \bar{p}^{(D, \theta)} \), to balance the load between the HV queue and the AV queue.

To illustrate, we consider the same two values of \( \lambda \), one below and one above \( \lambda^{(D, \theta)} = 2,510 \) vehicles per hour. At \( \lambda = 2,217 \) vehicles per hour, the benchmark case has enough capacity for all the vehicles that enter the highway, and no vehicle is blocked. Thus, assigning a lane to AVs may even reduce \( \theta \) when \( p \) is low, and employing this policy is not beneficial (see Appendix E.1). However, when the highway is heavily loaded, the D policy is able to improve \( \theta \). As Figure 4(b) shows, at \( \lambda = 11,342 \) vehicles per hour, the benchmark case has reached its maximum jam throughput (\( 555V_{555}/3 = 2590/3 = 863 \) vehicles per hour per lane), but the AV lane of the D policy has not; the maximum throughput of this lane of the D policy is about five times the throughput of a lane in the benchmark case (i.e., \( 185V_{185}^{DA} = 4,675 \) vs. 863 vehicles per hour per lane). When \( p \geq p^{(D, \theta)} = 0.17 \), the increase in the throughput of the lane dedicated to AVs is so significant that it compensates for the decrease in the throughput of the HV lanes. As a result, for \( 0.43 \leq p \leq 0.93 \) the overall throughput of the highway segment becomes about twice that of the benchmark case (i.e., 5,396 vs. 2,590 vehicles per hour). This is same as the ratio of the jam throughput under the D policy to that of the benchmark case (i.e., \( \frac{V_{e}^{DA}\lambda+(N-1)V_{(N-1)c}^{DH}}{NV_{c}} = 2.09 \)). As \( p \) increases further from 0.93, the HV queue is no longer highly-loaded, and \( \theta \) of this queue decreases as its arrival rate, \( (1-p)\lambda \), decreases, and \( \theta \) of the AV queue cannot increase further. Eventually, as \( p \) approaches 1, the overall \( \theta \) under this policy converges to the jam throughput of the AV lane (i.e., \( cV_{e}^{DA} \)). Since the jam throughput of the designated lane alone is higher than that of the benchmark case (see Figure 4(b)), when \( p \geq p^{(D, \theta)} \), \( \theta^{D}(p) \) does not go below \( \theta^{B} \), hence \( \bar{p}^{(D, \theta)} = 1.18 \).

Our results reveal that the throughput of the D policy depends crucially on the arrival rate to the highway; unfortunately the role of the arrival rate has been neglected in the AV literature. For example, in the simulation study performed by Liu et al. (2018), at a fixed value of \( \lambda \), one lane is dedicated to AVs when \( p = 0.4 \), resulting in about 24% improvement in \( \theta \). However, Liu et al. (2018) is silent on whether the benchmark case without AVs has reached its maximum throughput at this value of \( \lambda \), and how throughput \( \theta \) would change with different values of \( \lambda \).

\[ ^{18} \text{It is possible for the threshold } \bar{p}^{(D, \theta)} \text{ to be less than one, i.e., } \bar{p}^{(D, \theta)} < 1. \text{ In this case, the number of lanes, } N, \text{ is so high that the throughput of the designated AV lane under the D policy is lower than the throughput of } N \text{ lanes in the benchmark case.} \]
Parts (a) and (b) of Proposition 1 together show that the effect of employing the D policy on $\theta$ is not always the same as that on $W$. Corollary 1 presents the condition that determines which metric improves first, by comparing the thresholds $p(D,W)$, $p(D,\theta)$, and $\tilde{p}(D,\theta)$.

**Corollary 1.** (a) $p(D,W) \geq p(D,\theta)$, if $\lambda \geq \max\{\bar{\lambda}(D,W), \bar{\lambda}(D,\theta), \lambda(D,\theta)\}$, where

$$\bar{\lambda}(D,W) = \frac{V^B_c(V^DA - V^{DH}_{(N-1)c})[NcV^B_c - (N-1)cV^{DH}_{(N-1)c}]}{V^D_c[V^B_c - V^{DH}_{(N-1)c}][1 - \pi^D_c(p(D,\theta))]}.\]

(b) $p(D,W) \leq \tilde{p}(D,\theta)$, if $\lambda \geq \max\{\bar{\lambda}(D,W), \bar{\lambda}(D,\theta), \lambda(D,\theta)\}$, where

$$\bar{\lambda}(D,\theta) = \frac{V^B_c(V^DA - V^{DH}_{(N-1)c})[NcV^B_c - cV^D_c]}{\pi^{DH}_{(N-1)c}(1 - \tilde{p}(D,\theta))V^B_c[V^DA - V^{DH}_{(N-1)c}] + V^{DH}_{(N-1)c}[V^DA - V^B_c]].\]

Corollary 1(a) indicates that if the arrival rate to the highway is high with $\lambda \geq \max\{\bar{\lambda}(D,W), \bar{\lambda}(D,\theta), \lambda(D,\theta)\}$, then the D policy improves $\theta$ over the benchmark case before (i.e., at lower $p$) improving $W$, i.e., $p(D,W) \geq p(D,\theta)$. The intuition behind this result is that the overall $W$ of this policy is the weighted average of $W^{DH}$ in the HV queue and $W^DA$ in the AV queue, while the overall $\theta$ is the sum of $\theta$’s in these two queues. Consequently, when $(1-p)$ (i.e., the weight of $W^{DH}$ in $W$) is high, the low performance of the HV queue has a more significant impact on $W$ than $\theta$. Corollary 1(b) states that if the arrival rate to the highway is high with $\lambda \geq \max\{\bar{\lambda}(D,W), \bar{\lambda}(D,\theta), \lambda(D,\theta)\}$, there exists an interval of $p$, i.e., $\max\{p(D,W), p(D,\theta)\} \leq p \leq \tilde{p}(D,\theta)$, in which the D policy simultaneously improves $\theta$ and $W$ over the benchmark case. In our calibrated model, $\bar{\lambda}(D,W) = 2.510$, $\bar{\lambda}(D,\theta) = 2.992$, and $\lambda(D,W) = 0$ vehicles per hour, and $\pi^{DA}_c(p(D,\theta)) = \pi^{DH}_{(N-1)c}(1 - \tilde{p}(D,\theta)) = 0$; Figure 4 illustrates the case where $\lambda = 11,342 \geq \max\{\lambda(D,W), \lambda(D,\theta)\}$, hence $p(D,\theta) = 0.17 \leq p(D,W) = 0.63 \leq \tilde{p}(D,\theta) = 1$.

In a nutshell, the performance of the D policy depends significantly on an arrival rate and a proportion of AVs. Although this policy has the potential to reduce mean travel time as well as throughput, this requires that a substantial proportion of vehicles must be AVs, even in a congested highway with a high arrival rate.

### 5.2. I policy

This section analyzes the effect of AVs on highway congestion under the I policy, and compares its performance with that of the benchmark case. We first present Proposition 2 that compares this policy with the benchmark case for highly loaded highways.

**Proposition 2.** a) For any given $p \in [0,1]$, there exists $\lambda(I,W) \geq 0$ such that for $\lambda \geq \lambda(I,W)$, the I policy has a lower mean travel time $W$ than the benchmark case, if and only if $V^I_c(p) \geq V^B_N$, or equivalently

$$(1 - \delta^I)(\frac{2}{\psi^IH_N} - \frac{1}{\psi^IA_N} - \frac{1}{\psi^HA_N}) + (1 - \delta^I)(\frac{2}{\xi^IH_N} - \frac{1}{\xi^IA_N} - \frac{1}{\xi^HA_N}) \geq (\frac{2}{\psi^IH_N} - \frac{1}{\psi^IA_N} - \frac{1}{\psi^HA_N} \xi^IH_N \xi^IA_N \xi^HA_N) - (\frac{2}{\xi^IH_N} - \frac{1}{\xi^IA_N} - \frac{1}{\xi^HA_N} \xi^IH_N \xi^IA_N \xi^HA_N)\]

(15)
b) For any given $p \in [0, 1]$, there exists $\lambda(I, \theta) \geq 0$ such that for $\lambda \geq \lambda(I, \theta)$, the I policy has a higher throughput $\theta$ than the benchmark case, if and only if (15) holds.

Proposition 2 states that, when the highway is highly loaded, the I policy outperforms the benchmark case in terms of both $W$ and $\theta$, if and only if the jam speed of vehicles (state-dependent speed $V_{Nc}^{I}$ when $n = Nc$) is higher than that of the benchmark case (i.e., $V_{Nc}^{I}(p) \geq V_{Nc}^{B}$) for any given $p$. We illustrate this result using the calibrated model for $\lambda = 2,217$ and 11,342 vehicles per hour as in §5.1. We compute $\lambda(I,W) = \lambda(I,\theta) = 2,594$ vehicles per hour, so 2,217 and 11,342 vehicles per hour are considered as light load and high load, respectively. For light traffic, our numerical analysis in Appendix E.1 shows that AVs do improve the performance of the highway under the I policy, but the amount of improvement is not substantial because the benchmark model already performs quite well under light traffic. For example, when $\lambda = 2,217$, $W$ is improved by 5%, and $\theta$ is hardly improved; see Table A3 in Appendix E.1 for further discussion. At $\lambda = 11,342$ vehicles per hour, Figure 5(a) illustrates that $W$ is lower under the I policy than that of the benchmark case for all values of $p$. This happens because, as shown in Figure 3(a), the jam speed of vehicles under the I policy, $V_{Nc}^{I}(p)$, is increasing in $p$ for $p \in [0, 1]$, where $p = 0$ corresponds to the benchmark case. When $p < 0.95$, $\lambda = 11,342$ vehicles per hour is a high load (i.e., for these values of $p$, the jam service rate, $NcV_{Nc}^{I}(p)$, is lower than $\lambda$), so vehicles drive at the jam speed, $V_{Nc}^{I}(p)$. In this case, since $V_{Nc}^{I}(p)$ is increasing in $p$, $W = l/V_{Nc}^{I}(p)$ decreases in $p$. When $p \geq 0.95$, AVs are so prevalent that vehicles drive at the free-flow speed of the highway, and $W^{I}(p)$ is minimal. Figure 5(b) shows that $\theta$ under the I policy is strictly increasing for $p \leq 0.95$. For $p \geq 0.95$, $\theta$ is equal to $\lambda = 11,342$ vehicles per hour.
vehicles per hour, and it cannot grow any further, because vehicles are traveling at the free-flow speed. In Appendix E.4, we offer further discussion about a sharp decrease in $W$ at $p = 0.95$.

Since our model captures various characteristics of a mixed traffic flow, our results offer deeper insights than previous studies. Liu et al. (2018) and Bierstedt et al. (2014) state that in order for the I policy to improve the performance of the benchmark case (in terms of $\theta$) by about 30%, the AV proportion should be substantial - 60% and 75%, respectively. However, we observe that the performance of this policy crucially depends on $\lambda$: Whereas the I policy does not have a significant impact (about 5%) on $W$ or $\theta$ at $\lambda = 2,217$ vehicles per hour, 50% AVs halve the mean travel time and double the throughput at $\lambda = 11,342$ vehicles per hour (see Figure 5). This discrepancy may stem from the fact that these two studies focus on the role of AVs in reducing the mean intraplatoon headway. However, as described in §4.3, the weighted mean intraplatoon headway is not necessarily decreasing in $p$, and the magnitude of its change is not significant in any case. It turns out that the capability of AVs to reduce the weighted mean interplatoon headway is the primary driver for the speed increase under the I policy.

Ghiasi et al. (2017) consider the I policy for a one-lane highway. They show that $\theta$ increases in $p$, if and only if $1/\xi_{IAA} \leq 1/\psi_n^{IAH} + 1/\psi_n^{IH\lambda} \leq 1/\psi_n^{IH\lambda}$, where $n$ is a fixed number. In contrast, for our calibrated model in §4.3, although there exist values of $n$ such that the condition in Ghiasi et al. (2017) does not hold (e.g., at $\bar{n} = 100$, $3.28 + 2.81 \not\leq 2.81$ seconds), $\theta$ is still increasing in $p$. This happens because the number of vehicles on a highway is not a fixed value, and speed of vehicles (as well as the mean headway) changes with $n$ even if $p$ is fixed. Thus, by considering only one instance of $n$, Ghiasi et al. (2017) do not fully capture the effect of congestion on speed.

### 5.3. Comparison of D Policy and I Policy

Building on the analyses in §5.1 and §5.2, we compare the performance of the D policy and the I policy. Under the premise of Proposition 1, Proposition 3 presents conditions under which the I policy outperforms the benchmark case in terms of $W$ and $\theta$, and vice versa.

**Proposition 3.** Suppose $\lambda \geq \lambda_{jam}(1)$, and $V_{N_c}(p)$ is concave everywhere or convex everywhere, and it is increasing in $p$.\(^\text{19}\) Then:

a) When $p \to 0$, $W^I(p) \to W^B \leq W^D(p)$, and $\theta^I(p) \to \theta^B \geq \theta^D(p)$.

\(^{19}\)This condition of $V$ can be expressed in terms of parameters as follows: condition (15) at $n = N_c$ and either $\partial V^I / \partial p + \delta \partial (\psi_n^I \lambda_{jam}) / \partial p \leq 0$ or $\partial V^I / \partial p + \delta \partial (\psi_n^I \lambda_{jam}) / \partial p \geq 0$ for all $p \in [0,1]$. These assumption are not restrictive; as Figure 3(b) shows, these conditions are satisfied in our calibrated model for all the heavily loaded queues. Note that the first assumption about (15) requires $V^I$ to be increasing in $p$ only at $n = N_c$. When $V_{N_c}(p)$ is increasing in $p$ and $\lambda \geq \lambda_{jam}(1)$, all queues are highly loaded, because $V_{N_c}(1) = V^D$ is the maximum jam speed among all queues for all values of $p$. In addition, $\lambda_{jam}(1)$ is at least as high as both $\lambda^{(I,W)}$ and $\lambda^{(I,\theta)}$. This means that the I policy outperforms the benchmark case even before it becomes jammed.
b) When \( p \to 1 \), \( W^I(p) \to W^D(p) \leq W^B \), \( \theta^I(p) \geq \theta^B \geq \theta^D(p) \) if \( N \geq \frac{V_{cD}}{V_{Nc}} \), and \( \theta^I(p) \geq \theta^D(p) \geq \theta^B \) if \( N \leq \frac{V_{cD}}{V_{Nc}} \).

c) When \( p \neq 0 \) or 1,
   - \( W^I(p) \leq W^B \leq W^D(p) \) if \( p \leq p^{(D,W)} \), \( W^I(p) \leq W^D(p) \leq W^B \) if \( p^{(D,W)} \leq p \leq p^{(DI,W)} \), and \( W^D(p) \leq W^I(p) \leq W^B \) if \( p \geq p^{(DI,W)} \), where \( p^{(DI,W)} \) is the smallest \( p \) such that \( \frac{V_{cD} - V_{DH}(N_{cV})}{V_{cD} - V_{(N-1)c}} \leq V_{Nc}(p) \left( 1 - \frac{V_{cD}}{V_{(N-1)c}} \right) \).
   - For \( \lambda \geq \max(\lambda_{jam}(1), \lambda^{(D,g)}) \), \( \theta^D(p) \leq \theta^B \leq \theta^I(p) \) if \( p \leq \bar{p}^{(D,g)} \) or \( p \geq \bar{p}^{(D,g)} \), \( \theta^B \leq \theta^D(p) \leq \theta^I(p) \) if \( \bar{p}^{(D,g)} \leq p \leq \bar{p}^{(D,g)} \) or \( \bar{p}^{(D,g)} \leq p \leq \bar{p}^{(D,g)} \), where \( \bar{p}^{(D,g)} \) is the smallest \( p \) such that \( NcV^I_{Nc}(p) - p\lambda(1 - \pi_{D,A}) = (N - 1)cV^{DH}_{(N-1)c} \), and \( \bar{p}^{(D,g)} \) either satisfies \( NcV^I_{Nc}(\bar{p}^{(D,g)}) - (1 - \bar{p}^{(D,g)})\lambda(1 - \pi^{DH}_{(N-1)c}) = cV^{DA}_{c} \), or it is the largest \( p \) such that \( NcV^I_{Nc}(p) - p\lambda(1 - \pi_{D,A}) = (N - 1)cV^{DH}_{(N-1)c} \).

Proposition 3 characterizes ranges of \( p \) where the D policy should be employed over the I policy, and vice versa. When \( p \) converges to 0 or 1, almost all vehicles are HVs or AVs, respectively. In this case, Proposition 3(a)-(b) state that the I policy outperforms the D policy as well as the benchmark case. Under the D policy, when \( p \) approaches 0 (resp., 1), the arrival rate to the AV queue (resp., HV queue) approaches zero, so the highway does not utilize its full capacity. When \( p \) approaches 1, the jam throughput of the benchmark case is higher than that under the D policy, only if the highway has a sufficient number of lanes (i.e., \( N \geq \frac{V_{cD}}{V_{Nc}} \)), such that they collectively produce a higher throughput than one fast lane of AVs. In contrast, when \( p = 0 \), the I policy is equivalent to the benchmark case, and when \( p = 1 \), this policy performs better than the benchmark case, since it replaces all the HVs with fast AVs.

When \( 0 < p < 1 \), Proposition 3(c) specifies intervals of \( p \) where the D policy performs better than the I policy in terms of \( W \) and \( \theta \). We first discuss \( W \), and then \( \theta \). As discussed in §5.1 and §5.2, unlike the D policy that decreases \( W \) over the benchmark case only when \( p \geq p^{(D,W)} \), the I policy has a lower \( W \) than the benchmark case for any value of \( p \), due to the premise in Proposition 3 that the jam speed of vehicles is increasing in \( p \) under the I policy. As a result, when \( p \leq p^{(D,W)} \), if \( W \) is the only decision metric, the I policy should be chosen over the D policy. When \( p^{(D,W)} \leq p \leq p^{(DI,W)} \), although the D policy has a lower \( W \) than the benchmark case, the I policy still results in the lowest \( W \) and should be employed. In this case, the \( W \) of HVs under the D policy is so high that it increases the overall \( W \) of this policy, whereas the I policy enables these vehicles to travel faster by mixing them with fast moving AVs. When \( p \geq p^{(DI,W)} \), the highway has a lower \( W \) under the D policy than under the I policy. In this case, the arrival rate to the HV queue is so low that this queue flows freely. Thus, the overall \( W \) under the D policy, which is the
weighted average of $W^{DH}$ and $W^{DA}$, is determined primarily by the jam speed of vehicles on the heavily loaded AV lane. Under the I policy, the highway is also heavily loaded, but the jam speed of the mixture of vehicles is lower than the jam speed of the AV queue under the D policy.

Proposition 3(c) also compares $\theta$ between the two policies. For highly loaded highways, $\theta$ is increasing in $p$ under the I policy, and this policy outperforms the benchmark case for any $p$ (see §5.2). When $p \leq \underline{p}^{(D,\theta)}$, if $\theta$ is the only decision metric that is considered, the I policy should also be chosen over the D policy, because the D policy leads to a lower $\theta$ than the benchmark case. In addition, when $\underline{p}^{(D,\theta)} \leq p \leq \overline{p}^{(D,\theta)}$, even though the D policy has a higher $\theta$ than the benchmark case, $\theta$ under the I policy is still the highest. Only when $\overline{p}^{(D,\theta)} \leq p \leq \underline{p}^{(D,\theta)}$ does the D policy outperform the I policy in terms of $\theta$. In this case, the arrival rate to the AV queue of the D policy is so high that this queue works at its jam $\theta$ which is much higher than the jam $\theta$ of one lane of only HVs. The high $\theta$ of this fast AV lane makes up for the low $\theta$ of the HV lanes, and increases the overall $\theta$ under the D policy. In contrast, under the I policy, the state-dependent speed is balanced between fast AVs and slow HVs, so $\theta$ under this policy is not as high as that under the D policy. When $p \geq \overline{p}^{(D,\theta)}$, the I policy has a higher $\theta$ than the D policy: AVs, that constitute the majority of vehicles, can run on all lanes under the I policy, whereas they can run only on one designated lane under the D policy.

We illustrate Proposition 3 using the calibrated model in §4 for highly loaded highways with $\lambda = 11,342$ vehicles per hour. As Figure 6(a) shows, the I policy leads to a lower $W$ than the D policy for all values of $p$, except when $p$ is between $\underline{p}^{(D,I,W)} = 0.93$ and 0.94. (Note that when $p > 0.94$, since $V^{I}_N(p)$ is increasing in $p$, the condition in Proposition 3 is violated, and $\lambda = 11,342$ vehicles per hour is not a jam load for this highway.) Figure 6(b) compares $\theta$ under the I policy with that under the D policy. As stated in Proposition 3, when $p \in [0.25, 0.55]$ (where $\underline{p}^{(D,I,\theta)} = 0.25$ and $\overline{p}^{(D,I,\theta)} = 0.55$), the D policy leads to a higher throughput $\theta$ than the I policy. When the highway is lightly loaded with $\lambda = 2,217$ vehicles per hour, in terms of both $W$ and $\theta$, the I policy performs at least as well as the benchmark case, while the D policy can be inferior to the benchmark case. In this case, integrating AVs and HVs improves the performance of the highway more than assigning one lane to AVs for any value of $p \in [0, 1]$ (see Appendix E.1).

Proposition 3(c) shows, interestingly, that on some interval of $p$, either policy outperforms the other in terms of $W$ or $\theta$. Corollary 2 specifies this interval.

**Corollary 2.** $\underline{p}^{(D,I,\theta)} \leq \min\{\underline{p}^{(D,I,W)}, \overline{p}^{(D,I,\theta)}\}$.

Corollary 2 implies that there exists an interval of $p$, $[\underline{p}^{(D,I,\theta)}, \min\{\underline{p}^{(D,I,W)}, \overline{p}^{(D,I,\theta)}\}]$ (e.g., $[0.25, 0.55]$ in Figure 6(b)), such that, in terms of throughput, the D policy performs better than
the I policy, but in terms of mean travel time, it is worse. The driver of this trade-off is the fact that, under the D policy, the highway is divided into a fast queueing system for AVs and a slow one for HVs. When \( p \) is small, the HV queue is still heavily loaded, so the mean travel time of these vehicles is much longer than that of the benchmark case (i.e., \( \frac{W_{DH}}{W_B} = \frac{V_B}{(N-1)c} \geq 1 \)). The throughput of the HV queue is reduced compared to the benchmark case, but the amount of reduction in \( \theta \) is smaller than the amount of increase in \( W \) (i.e., \( \frac{\theta_B}{\theta_I} = \frac{V_B}{V_{NC}(N)} \leq W_{DH}W_B \)). Thus, when \( p \in [\bar{p}^{(D,\theta)}, \min\{\bar{p}^{(D,I,W)}, \bar{p}^{(D,\theta)}\}] \), the throughput derived from the balanced speed of vehicles (\( V_{NC}(p) \)) under the I policy is lower than the high throughput the D policy achieves by utilizing the fast AV lane.

When \( p \) is high, it seems intuitive to increase the number of designated lanes to AVs under the D policy. Numerically analyzing the D policy with two lanes designated to AVs, we observe that although this policy increases \( \theta \), it still performs poorly in terms of \( W \). See Appendix E.5 for further discussion.

6. Policy Recommendation and Conclusion

In this paper we investigate the effects of autonomous vehicles on highway traffic flow under two policies: the D and the I policies. We model the traffic flow as a queueing system, and calibrate it to data. Then, we analyze each of these policies, as well as the benchmark case, and provide recommendations about when each of these policies should be employed.
We use two metrics to measure the performance of a highway: the mean travel time $W$ of a single vehicle, and the throughput $\theta$ of the highway. The former concerns users of the highway, while the latter is important for urban designers. Unfortunately, these two metrics are not always aligned: a high utilization of the highway does not guarantee a short travel time, and vice versa. Thus, it is crucial for policy makers to take both of these metrics into account.

The performance of different policies depends crucially on an arrival rate to the highway, $\lambda$, which has been overlooked in the AV literature to date, as well as a proportion of AVs, $p$. Since AVs are primarily intended to alleviate congested highways, we focus on high values of $\lambda$. For each metric, we recommend policies in three different regions of $p$: low, moderate, and high. Our analysis indicates:

1) When $p$ is low, the I policy is recommended. This is intuitive because the number of AVs on the highway is so low that it is not worth designating one lane to them. However, this policy recommendation is in contrast with the suggestion provided by Bierstedt et al. (2014) to employ the D policy at the beginning of the mass appearance of AVs. This is because in their simulation analysis as well as in most other studies mentioned in §2, the effect of AVs on interplatoon headway is not considered; furthermore it is not clear if the highway they study is congested. Thus, the Departments of Transportations in several states, including Colorado, Wisconsin and Washington, should recognize that their plan to designate a lane to AVs (Aguilar 2018) could lead to significantly more congestion than the I policy (although it may have other benefits, such as helping human drivers become acquainted with the new era of AVs, and incentivizing adoption).

2) In the moderate $p$ region, a policy maker could consider adopting either policy, depending on which metric he or she cares about more. If a lower travel time ($W$) for vehicles is of more importance, the I policy is the solution; but if improving the overall utilization ($\theta$) of the highway is of high priority, the D policy should be employed. In terms of $W$, under the I policy, the state-dependent speed of vehicles is jointly determined by AVs and HVs, and it is faster (slower) than the speed when there are only HVs (AVs) on the highway, so this policy has a lower $W$ than the benchmark case where only HVs are present. Under the D policy, HVs experience a significantly long $W$. Since in this case HVs still constitute a significant proportion of vehicles, the overall $W$ of the highway becomes longer than that in the benchmark case, and hence also that under the I policy. In terms of $\theta$, under the D policy, the fast AV lane has a very high throughput, thereby improving the overall $\theta$ of the highway substantially. Under the I policy, although $\theta$ improves over the benchmark case, this improvement is not as significant.

3) When $p$ is high, the I policy again performs better than the D policy on both metrics. This is because AVs, which constitute the majority of vehicles, are allowed to use only one lane of the
highway under the D policy, but they can drive on any lane under the I policy. In this case, the D policy can still improve the performance of the highway over the benchmark case, but since it is outperformed by the I policy, we recommend the I policy. Whereas Bierstedt et al. (2014) predict that, under the I policy, even high values of the AV proportion will not have a significant impact on throughput, (e.g., 60% AVs increase throughput by only 30%), we observe that 60% AVs increase throughput by about 130%. Moreover, in line with KPMG’s prediction that AVs could increase the capacity of highways by 500%, our calibrated model shows that, when all vehicles are autonomous, the I policy increases the throughput of the highway by 437% on a congested highway.

Our paper is the first to model and analyze a multi-lane highway using a queueing system coupled with a MAP. There are several interesting avenues to expand this research. For example, one may explore endogenizing the adoption rate of AVs based on quality of traffic flow on highways and regular roads: If having AVs helps reduce congestion, more people would be interested in trading their conventional vehicle for an autonomous one. As a result, the adoption rate of AVs may be determined depending on various factors that include the amount of improvement in congestion on regular roads and highways, the price of AVs (which is unknown yet), possibly additional trips due to the autonomous feature (which gives free time to drivers), and so on. In addition, it would be interesting to study the effect of AVs on highway crash rates. On the one hand, AVs are more capable of preventing crashes than HVs, due to their low reaction time. On the other hand, platoons of AVs are longer and more dense than platoons of HVs, so an accident between two AVs could propagate through the entire platoon and affect a longer string of vehicles. Thus, it is not immediately clear if AVs will improve highway safety.

References


Transportation Research Record: Journal of the Transportation Research Board (2324):63–70.


Stern RE, Cui S, Delle Monache ML, Bhadani R, Bunting M, Churchill M, Hamilton N, Pohlmann H, 


Appendix A: Summary of Notation

Throughout this paper, for model parameters, we use a superscript $i \in \{B, D, DH, DA, I\}$ and a subscript $n \in \{1, 2, \cdots , N_c\}$, where B, D, DH, DA, and I represent the benchmark case, the D policy, the HV queue of the D policy, the AV queue of the D policy, and the I policy, respectively, and $n$ is the number of vehicles on highway. Moreover, for thresholds of $\lambda$ and $p$, we use superscripts $(i,j)$ for $i \in \{D, I, DI\}$ and $j \in \{W, \theta, W\theta\}$, where D, I, and DI represent the D policy, the I policy, and comparison between D and I policies, respectively; and $W$, $\theta$, and $W\theta$ indicate thresholds for $W$, $\theta$, and comparison between $W$ and $\theta$, respectively. We also use underscore and overscore to indicate smaller and larger thresholds, respectively.

<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
<th>Notation</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\lambda$</td>
<td>Arrival rate to the highway</td>
<td>$N$</td>
<td>Number of lanes</td>
</tr>
<tr>
<td>$L$</td>
<td>Length of the highway segment</td>
<td>$J$</td>
<td>Jam density of the highway</td>
</tr>
<tr>
<td>$c$</td>
<td>Highway capacity</td>
<td>$C$</td>
<td>Irreducible generator matrix of the MAP</td>
</tr>
<tr>
<td>$C^0$</td>
<td>Transition rate matrix to the non-absorbing states of the MAP</td>
<td>$C^1$</td>
<td>Transition rate matrix to the absorbing state of the MAP</td>
</tr>
<tr>
<td>$\delta^0$</td>
<td>Initial distribution of the non-absorbing states of the platoon size distribution</td>
<td>$\delta$</td>
<td>Initial distribution of the absorbing state of the platoon size distribution</td>
</tr>
<tr>
<td>$G^0$</td>
<td>Probability transition matrix associated with the non-absorbing states of the platoon size distribution</td>
<td>$G$</td>
<td>Probability transition matrix associated with the absorbing state of the platoon size distribution</td>
</tr>
<tr>
<td>$\delta^i$</td>
<td>Parameter of the platoon size distribution (or equivalently the reciprocal of the mean platoon size) in the model $i \in {B, DA, DH, I}$</td>
<td>$M$</td>
<td>Transition probability matrix of the platoon size distribution</td>
</tr>
<tr>
<td>$\alpha_n^0 (1)$</td>
<td>Initial distribution of the non-absorbing states of the state-dependent intraplatoon headway distribution</td>
<td>$\alpha_n^1 (1)$</td>
<td>Initial distribution of the absorbing state of the state-dependent intraplatoon headway distribution</td>
</tr>
<tr>
<td>$Q_n^0 (1)$</td>
<td>Transition rate matrix associated with the non-absorbing states of the state-dependent intraplatoon headway distribution</td>
<td>$Q_n^1 (1)$</td>
<td>Transition rate matrix associated with the absorbing state of the state-dependent intraplatoon headway distribution</td>
</tr>
<tr>
<td>$\alpha_n^0 (2)$</td>
<td>Initial distribution of the non-absorbing states of the state-dependent interplatoon headway distribution</td>
<td>$\alpha_n^1 (2)$</td>
<td>Initial distribution of the absorbing state of the state-dependent interplatoon headway distribution</td>
</tr>
<tr>
<td>$Q_n^0 (2)$</td>
<td>Transition rate matrix associated with the non-absorbing states of the state-dependent interplatoon headway distribution</td>
<td>$Q_n^1 (2)$</td>
<td>Transition rate matrix associated with the absorbing state of the state-dependent interplatoon headway distribution</td>
</tr>
<tr>
<td>$1/\xi_n^i$</td>
<td>Mean intraplatoon headway when there are $n$ vehicles on the highway in the model $i \in {B, DA, DH, IA, IR, IHA, IAH, I}$</td>
<td>$1/\psi_n^i$</td>
<td>Mean interplatoon headway when there are $n$ vehicles on the highway in the model $i \in {B, DA, DH, IA, IR, IHA, IAH, I}$</td>
</tr>
<tr>
<td>$\mu_n^i$</td>
<td>Service rate of a single vehicle when there are $n$ vehicles on the highway in the model $i \in {B, DA, DH, I}$</td>
<td>$V_n^i (p)$</td>
<td>Speed of a single vehicle when there are $n$ vehicles on the highway in the model $i \in {B, DA, DH, I}$, and the AV proportion is $p$</td>
</tr>
<tr>
<td>$k$</td>
<td>Highway density</td>
<td>$q$</td>
<td>Highway flow</td>
</tr>
<tr>
<td>$h_n$</td>
<td>Mean headway when there are $n$ vehicles on the highway</td>
<td>$d_n$</td>
<td>Stopping distance when there are $n$ vehicles on the highway</td>
</tr>
<tr>
<td>$\pi$</td>
<td>Steady-state distribution of the MAP</td>
<td>$\pi$</td>
<td>Steady-state distribution of an $M/G_n/c/c$ queueing model</td>
</tr>
<tr>
<td>$p$</td>
<td>Proportion of AVs</td>
<td>$\pi_c$</td>
<td>Blocking probability</td>
</tr>
<tr>
<td>$W$</td>
<td>Mean travel time of vehicles</td>
<td>$\theta$</td>
<td>Throughput of the highway</td>
</tr>
</tbody>
</table>

Table A1: Table of Notation
<table>
<thead>
<tr>
<th>Notation</th>
<th>Definition</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{\lambda}^{(D,W)}$</td>
<td>The arrival rate threshold before which the D policy has a higher $W$ than the benchmark case.</td>
<td>Proposition 1</td>
</tr>
<tr>
<td>$\lambda^{(D,W)}$</td>
<td>The arrival rate threshold after which the D policy may have a lower $W$ than the benchmark case.</td>
<td>Proposition 1</td>
</tr>
<tr>
<td>$\lambda^{(D,\theta)}$</td>
<td>The arrival rate threshold after which the D policy may have a higher $\theta$ than the benchmark case.</td>
<td>Proposition 1</td>
</tr>
<tr>
<td>$\bar{\lambda}^{(D,W,\theta)}$</td>
<td>The arrival rate threshold after which there exists an interval of $p$ such that the D policy tends to increase $\theta$, but it does not decrease $W$ over those of the benchmark case.</td>
<td>Corollary 1</td>
</tr>
<tr>
<td>$\underline{\lambda}^{(D,W)}$</td>
<td>The arrival rate threshold after which there exists an interval of $p$ such that the D policy tends to increase $\theta$ and decrease $W$ over those of the benchmark case.</td>
<td>Corollary 1</td>
</tr>
<tr>
<td>$\lambda^{(I,W)}$</td>
<td>The arrival rate threshold after which the I policy may have a lower $W$ than the benchmark case.</td>
<td>Proposition 2</td>
</tr>
<tr>
<td>$\lambda^{(I,\theta)}$</td>
<td>The arrival rate threshold after which the I policy may have a higher $\theta$ than the benchmark case.</td>
<td>Proposition 2</td>
</tr>
<tr>
<td>$p^{(D,W)}$</td>
<td>The AV proportion threshold after which the D policy has a lower $W$ than the benchmark case for a highly loaded highway.</td>
<td>Proposition 1</td>
</tr>
<tr>
<td>$p^{(D,\theta)} \leq p \leq \bar{p}^{(D,\theta)}$</td>
<td>The AV proportion interval in which the D policy has a higher $\theta$ than the benchmark case for a highly loaded highway.</td>
<td>Proposition 1</td>
</tr>
<tr>
<td>$p^{(DI,W)}$</td>
<td>The AV proportion threshold after which the D policy has a lower $W$ than the I policy for a highly loaded highway.</td>
<td>Proposition 3</td>
</tr>
<tr>
<td>$p^{(DI,\theta)} \leq p \leq \bar{p}^{(DI,\theta)}$</td>
<td>The AV proportion interval in which the D policy has a higher $\theta$ than the I policy for a highly loaded highway.</td>
<td>Proposition 3</td>
</tr>
</tbody>
</table>

Table A2  Table of Thresholds

Appendix B: Examples

Example 1. (Discrete PH-type distribution) A uniform distribution on $1, 2, \cdots, l$ can be represented as:

$$G^0 = \begin{bmatrix}
0 & \frac{1}{l-1} & 0 & \cdots & 0 \\
0 & 0 & \frac{1}{l-2} & \cdots & 0 \\
& \ddots & & & \\
0 & 0 & 0 & \frac{1}{2} & 0 \\
0 & 0 & 0 & 0 & 0
\end{bmatrix},$$

and $\delta^0 = [\frac{l-1}{l}, 0, \cdots, 0]$. Note that for this Markov chain the first passage time to the absorbing state $l$ happens with probability $\frac{1}{l}$ in $i$ steps, where $i = 1, 2, \cdots, l$. The Markov chain associated with this uniform distribution is illustrated in Figure 7. In this Markov chain, transition to $l$ happens in one step, only if we start at $l$, and this happens with probability $\frac{1}{l}$. If we start at state one and then go to $l$, it happens in two steps with the probability $\frac{1}{l-1} \times \frac{1}{l-1} = \frac{1}{l}$. Repeating this process, one can observe that transition to the absorbing state happens in any $i \in \{1, 2, \cdots, l\}$ steps with the same probability $\frac{1}{l}$, representing a uniform distribution on $1, 2, \cdots, l$. Note that uniform distribution is only a special case of the PH-type distributions, and other distributions have different $\delta^0$ and $G^0$.

Example 2. (MAP) Suppose the platoon size follows a geometric distribution with parameters $\delta = G$ and $\delta^0 = G^0 = 1 - \delta$. Assuming the interplatoon and intraplatoon headways follow $\text{exp}(\psi)$ and $\text{exp}(\xi)$,
Finally, respectively, we have

\[ C^0 = \begin{bmatrix} -\psi & 0 \\ 0 & -\xi \end{bmatrix} \quad \text{and} \quad C^1 = \begin{bmatrix} -(1 - \delta)\psi & (1 - \delta)\psi \\ \delta \xi & -\delta \xi \end{bmatrix}. \]

The generator matrix of the MAP is then

\[ C = C^0 + C^1 = \begin{bmatrix} -(1 - \delta)\psi & (1 - \delta)\psi \\ \delta \xi & -\delta \xi \end{bmatrix}. \]

The stationary probability vector of this MAP, \( \tilde{\pi} \), is given by

\[ \tilde{\pi}C = 0, \quad \text{and} \quad \tilde{\pi} = \frac{(1 - \delta)\psi}{\delta \xi + (1 - \delta)\psi}, \]

is the arrival rate of the MAP.

**Appendix C: Proofs**

**Lemma 1.** For an \( M/G_c/c \) queueing model, the mean travel time \( W \) is increasing in arrival rate \( \lambda \geq 0 \).

**Proof.** By (13) and (14), we can state the mean travel time as follows:

\[
W = \sum_{n=0}^{c} \frac{n\pi_n}{\lambda(1 - \pi_c)} = \frac{\sum_{n=1}^{c} \left\{ \frac{\lambda^n/n!V_n\cdots V_1}{\lambda[1 - \lambda/n!(V_n\cdots V_1)]/[1 + \sum_{i=1}^{c} \lambda^i/(i!V_i\cdots V_1)]} \right\}}{\lambda[1 + \sum_{n=1}^{c} \lambda^n/(n!V_n\cdots V_1)]/[1 + \sum_{i=1}^{c} \lambda^i/(i!V_i\cdots V_1)]}
\]

\[
= \frac{\sum_{n=1}^{c} \left\{ \frac{\lambda^n/[n!V_{n+1}V_n\cdots V_1]}{1 + \sum_{n=1}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]} \frac{\lambda^{n-1}/[(n-1)!V_n\cdots V_1]}{1 + \sum_{n=1}^{c} \lambda^{n-1}/[(n-1)!V_n\cdots V_1]} \right\}}{1 + \sum_{n=1}^{c} \lambda^n/[n!V_n\cdots V_1]}
\]

\[
= \frac{\sum_{n=1}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]}{1 + \sum_{n=0}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]^2}
\]

Taking the derivative with respect to \( \lambda \), we get the following:

\[
W' = \frac{\partial W}{\partial \lambda} = \frac{\sum_{n=1}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]}{1 + \sum_{n=0}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]^2}
\]

\[
= \frac{\sum_{n=0}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]}{1 + \sum_{n=0}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]^2}
\]

\[
\cdot \left\{ \frac{\sum_{n=1}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]}{1 + \sum_{n=0}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]^2} \right\}
\]

\[
\cdot \left\{ \frac{\sum_{n=1}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]}{1 + \sum_{n=0}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]^2} \right\}
\]

\[
\cdot \left\{ \frac{\sum_{n=1}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]}{1 + \sum_{n=0}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]^2} \right\}
\]

\[
\cdot \left\{ \frac{\sum_{n=1}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]}{1 + \sum_{n=0}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]^2} \right\}
\]

\[
\cdot \left\{ \frac{\sum_{n=1}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]}{1 + \sum_{n=0}^{c} \lambda^n/[n!V_{n+1}V_n\cdots V_1]^2} \right\}
\]
Let \( \alpha_0 = 1 \) and \( \alpha_n = \lambda^n/[n!V_n \cdots V_1] \) for \( n = 1, 2, \cdots, c \), then we have:

\[
W' = \frac{\left[ \sum_{n=0}^{c-2} \alpha_n/(V_{n+2}V_{n+1}) \right] (1 + \sum_{n=1}^{c-1} \alpha_n) - \left( \sum_{n=0}^{c-2} \alpha_n/V_{n+1} \right) \left( \sum_{n=0}^{c-1} \alpha_n/V_{n+1} \right)}{(1 + \sum_{n=1}^{c-1} \alpha_n)^2}.
\]

Using induction, we show \( W \) is increasing in \( \lambda \) for \( c = 1, 2, 3, \cdots \). For \( c = 1 \), \( W = 1/V_1 \) is weakly increasing in \( \lambda \) (it is independent of \( \lambda \)). As the induction hypothesis, we assume \( W' \geq 0 \) when the capacity of the system is \( c \). Since the denominator of \( W' \) is positive, we have:

\[
\left[ \sum_{n=0}^{c-2} \alpha_n/(V_{n+2}V_{n+1}) \right] (1 + \sum_{n=1}^{c-1} \alpha_n) - \left( \sum_{n=0}^{c-2} \alpha_n/V_{n+1} \right) \left( \sum_{n=0}^{c-1} \alpha_n/V_{n+1} \right) \geq 0.
\]

Assume the capacity of the system is \( c + 1 \), then \( W' \) is as follows:

\[
W' = \frac{\left[ \sum_{n=0}^{c-2} \alpha_n/(V_{n+2}V_{n+1}) \right] (1 + \sum_{n=1}^{c-1} \alpha_n) - \left( \sum_{n=0}^{c-2} \alpha_n/V_{n+1} \right) \left( \sum_{n=0}^{c-1} \alpha_n/V_{n+1} \right)}{(1 + \sum_{n=1}^{c-1} \alpha_n)^2}.
\]

\[
= \frac{\left[ \sum_{n=0}^{c-2} \alpha_n/(V_{n+2}V_{n+1}) \right] (1 + \sum_{n=1}^{c-1} \alpha_n) - \left( \sum_{n=0}^{c-2} \alpha_n/V_{n+1} \right) \left( \sum_{n=0}^{c-1} \alpha_n/V_{n+1} \right) + \left\{ \left[ \sum_{n=0}^{c-2} \alpha_n/(V_{n+2}V_{n+1}) \right] (1 + \sum_{n=1}^{c-1} \alpha_n) + \left( \sum_{n=0}^{c-2} \alpha_n/V_{n+1} \right) \left( \sum_{n=0}^{c-1} \alpha_n/V_{n+1} \right) \right\}}{(1 + \sum_{n=1}^{c-1} \alpha_n)^2}.
\]

Since \( \frac{1}{n+1} - \frac{1}{c} \geq 0 \), \( \frac{1}{n+1} - \frac{1}{c} > 0 \), and by the induction hypothesis, the numerator of \( W' \) is positive. The denominator is also positive. Thus, \( W' \geq 0 \), when the capacity of the system is \( c + 1 \).

By the principle of mathematical induction \( W \) is increasing in \( \lambda \geq 0 \). ■

**Lemma 2.** For an \( M/G_n/c/c \) queueing model, the throughput \( \theta \) is an increasing function of the arrival rate \( \lambda \), if at least one of the following conditions holds:

\[
\lambda \geq \frac{V_1(c + 1 - 1/\pi_c) - c\pi_c(V_1 - V_c)}{1 - \pi_c} \quad \text{(17)}
\]

or

\[
nV_n \text{ is increasing in } n. \quad \text{(18)}
\]

**proof.** We first prove the sufficiency of condition (17) as follows.
By 13, \( \theta \) for an \( M/G_c/c \) queueing model can be expressed as:

\[
\theta = \lambda (1 - \pi_c) = \lambda \left[ 1 - \frac{\lambda^e/(cV_e \cdots V_1)}{1 + \sum_{n=1}^c \lambda^n/(n!V_n \cdots V_1)} \right] = \lambda \left[ 1 + \sum_{n=1}^{c-1} \frac{\lambda^n/(n!V_n \cdots V_1)}{1 + \sum_{n=1}^c \lambda^n/(n!V_n \cdots V_1)} \right].
\]

Letting \( \alpha_0 = 1 \) and \( \alpha_n = \lambda^n/[n!V_n \cdots V_1] \) for \( n = 1, 2, \ldots, c \), we get \( \theta = \frac{\lambda + \sum_{n=1}^{c-1} \alpha_n}{1 + \sum_{n=1}^c \alpha_n} \), and \( \pi_c = \alpha_c/(1 + \sum_{n=1}^c \alpha_n) \).

The derivative of \( \theta \) with respect to \( \lambda \) is as follows:

\[
\theta' = \frac{[1 + \sum_{n=1}^{c-1} (n+1)\alpha_n] (1 + \sum_{n=1}^c \alpha_n) - (1 + \sum_{n=1}^c \alpha_n) (\sum_{n=1}^c n\alpha_n)}{(1 + \sum_{n=1}^c \alpha_n)^2}
\]

\[
= \frac{(1 + \sum_{n=1}^{c-1} \alpha_n) (1 + \sum_{n=1}^c \alpha_n) + \sum_{n=1}^{c-1} n\alpha_n (1 + \sum_{n=1}^c \alpha_n) - (1 + \sum_{n=1}^c \alpha_n) (\sum_{n=1}^c n\alpha_n)}{(1 + \sum_{n=1}^c \alpha_n)^2}
\]

\[
= \frac{[\alpha_c (-1 + 1/\pi_c)] (\alpha_c/\pi_c) + \alpha_c \sum_{n=1}^{c-1} n\alpha_n - \alpha_c [\alpha_c (-1 + 1/\pi_c)]}{(\alpha_c/\pi_c)^2}
\]

\[
= \frac{\alpha_c [-\alpha_c + 1/\pi_c + 1/\pi_c^2] + \sum_{n=1}^{c-1} n\alpha_n - \alpha_c (-c + c/\pi_c)}{(\alpha_c/\pi_c)^2}
\]

\[
= \frac{\alpha_c [c - (c+1)/\pi_c + 1/\pi_c^2] + \sum_{n=1}^{c-1} n\alpha_n}{\alpha_c/\pi_c^2}.
\]

In the numerator of the above expression, \( c - (c+1)/\pi_c + 1/\pi_c^2 \) is non-negative, if \( 0 \leq \pi_c \leq \frac{1}{c} \) or \( \pi_c = 1 \). For \( \frac{1}{c} < \pi_c < 1 \), we restate the numerator as follows:

\[
\alpha_c [c - (c+1)/\pi_c + 1/\pi_c^2] + \sum_{n=1}^{c-1} n\alpha_n = \alpha_c [c - (c+1)/\pi_c + 1/\pi_c^2] + \sum_{n=1}^{c-1} \frac{\lambda}{V_n} \alpha_{n-1}
\]

\[
= \alpha_c [c - (c+1)/\pi_c + 1/\pi_c^2] + \sum_{n=0}^{c-2} \frac{\lambda}{V_{n+1}} \alpha_n
\]

\[
\geq \alpha_c [c - (c+1)/\pi_c + 1/\pi_c^2] + \sum_{n=0}^{c-2} \frac{\lambda}{V_1} \alpha_{n+1}
\]

\[
= \alpha_c [c - (c+1)/\pi_c + 1/\pi_c^2] + \frac{\lambda}{V_1} \left[ (\sum_{n=0}^{c} \alpha_n) - \alpha_c - \alpha_{c-1} \right]
\]

\[
= \alpha_c [c - (c+1)/\pi_c + 1/\pi_c^2] + \frac{\alpha_c}{\pi_c} - \alpha_c - \frac{cV_c}{\lambda} \alpha_c
\]

\[
= \alpha_c [c - (c+1)/\pi_c + 1/\pi_c^2] + \frac{\lambda}{V_1} \frac{1}{\pi_c} - 1 - \frac{cV_c}{\lambda}.
\]
The above expression is positive if \( \lambda \geq \frac{V_1(c + 1 - 1/\pi_c) - c\pi_c(V_1 - V_c)}{1 - \pi_c} \). Thus, \( \theta' \geq 0 \), when \( \lambda \geq V_1(c + 1 - 1/\pi_c) - c\pi_c(V_1 - V_c) \).

Next, we use induction to prove the sufficiency of condition (18) for \( c = 1, 2, 3, \ldots \). For \( c = 1 \), \( \pi_0 = 1 - \pi_1 = \frac{V_1}{cV_c} \), so \( \theta = \frac{\lambda_0}{\lambda V_1} \) and \( \theta' = \frac{\lambda^2}{(\lambda + V_1)^2} \geq 0 \). Thus \( \theta \) is increasing in \( \lambda \). Suppose \( \theta' \geq 0 \) when the capacity of the system is \( c \) and \( cV_c \geq (c - 1)V_{c - 1} \). Since in (19) the denominator of \( \theta' \) is positive, we have:

\[
[1 + \sum_{n=1}^{c-1} (n + 1)\alpha_n](1 + \sum_{n=1}^{c} \alpha_n) - (1 + \sum_{n=1}^{c-1} \alpha_n)(\sum_{n=1}^{c} n\alpha_n) \geq 0.
\]

Assume the capacity of the system is \( c + 1 \), then \( \theta' \) is as follows:

\[
\theta' = \frac{[1 + \sum_{n=1}^{c} (n + 1)\alpha_n](1 + \sum_{n=1}^{c} \alpha_n) - (1 + \sum_{n=1}^{c-1} \alpha_n)(\sum_{n=1}^{c} n\alpha_n)}{(1 + \sum_{n=1}^{c} \alpha_n)^2}.
\]

Due to the induction hypothesis, the first fraction is positive. To show that the second fraction is positive, it suffices to show that each \( \alpha_c\pi_{n+1} = -\alpha_{c+1}\pi_n \) for \( n = 0, 1, \ldots, c - 1 \) is positive, since the procedure is similar, we show this only for \( n = c - 1 \).

\[
\alpha_c\lambda_c - \alpha_{c+1}\lambda_{c-1} = \left(\frac{\lambda_c}{cV_c} \cdots \frac{\lambda_c}{V_1}\right) - \left(\frac{\lambda_{c+1}}{(c+1)!V_{c+1} \cdots V_1}\right)\left[\frac{\lambda_{c-1}}{(c-1)!V_{c-1} \cdots V_1}\right] = \frac{\lambda^c}{(cV_c)^c} \left(\frac{1}{cV_c} - \frac{1}{(c + 1)V_{c+1}}\right).
\]

This expression is positive, if \( cV_c \leq (c + 1)V_{c+1} \).

By the principle of mathematical induction \( \theta \) is increasing in \( \lambda \), if \( n\pi_n \) is increasing in \( n \). ■

**Proof of Proposition 1**  a) By (16), for an \( M/G_n/c/c \) queueing system, the mean travel time is calculated as follows:

\[
W = \frac{\sum_{n=0}^{c} \pi_n}{\lambda(1 - \pi_c)} = \frac{\sum_{n=0}^{c-1} \lambda^n/[n!V_{n+1} \cdots V_1]}{1 + \sum_{n=1}^{c-1} \lambda^n/[n!V_n \cdots V_1]}.
\]
In terms of our queueing models, we have \( W_0^B = 1/V_1^B \), and \( W_0^D(p) = pW_0^{DA} + (1 - p)W_0^{DH} = p/V_1^{DA} + (1 - p)/V_1^{DH} \). Since, \( V_1^{DA} = V_1^B \) and \( V_1^{DH} \leq V_1^B \), \( W_0^D(p) \geq W_0^B \). Moreover, according to Lemma 1, \( W^B \) and \( W^D(p) \) are both increasing in \( \lambda \). Thus, for every \( p \) there exists a \( \Delta^{(D,W)}(p) \) such that for \( \lambda \leq \Delta^{(D,W)}(p) \), \( W^D(p) \geq W^B \). Let \( \Delta^{(D,W)} = \min_p \Delta^{(D,W)}(p) \), then when \( \lambda \leq \Delta^{(D,W)} \), \( W^D(p) \geq W^B \) for \( p \in [0, 1] \).

Taking the limit of \( W \) as the arrival rate \( \lambda \) approaches \( \infty \), we get the following expression:

\[
W_\infty = \lim_{\lambda \to \infty} W = \lim_{\lambda \to \infty} \frac{\sum_{n=0}^{c-1} \lambda^n / [n!V_n V_1 \cdots V_1]}{1 + \sum_{n=0}^{c-1} \lambda^n / [n!V_n \cdots V_1]} = 1/V_c.
\]

In terms of our queueing models, we have \( W_\infty^B = 1/V_N^B \), and \( W_\infty^D(p) = pW_\infty^{DA} + (1 - p)W_\infty^{DH} = p/V_1^{DA} + (1 - p)/V_1^{DH} \). As a result, \( W_\infty^D(p) \leq W_\infty^B \) if and only if \( p \geq p^{(D,W)} = \frac{\max_{N_c} V_1^{DA} - V_1^{DA}V_1^{DH}(N_1-c)}{V_1^{DA}V_N^B - V_1^{DA}V_1^{DH}(N_1-c)} \) (since \( V_1^{DH} \leq V_1^{DA} \leq 0 \) the direction of inequality changes). Moreover, according to Lemma 1, \( W^B \) and \( W^D(p) \) are both increasing in \( \lambda \). Thus, for a given \( p \geq p^{(D,W)} \), there exists a \( \bar{\lambda}^{(D,W)}(p) \) such that \( \lambda \geq \bar{\lambda}^{(D,W)}(p) \), \( W_\infty^D(p) \leq W^B \). Let \( \bar{\lambda}^{(D,W)} = \max_p \bar{\lambda}^{(D,W)}(p) \), then when \( \lambda \geq \bar{\lambda}^{(D,W)} \), \( W_\infty^D(p) \geq W^B \) for \( p \geq p^{(D,W)} \). It is immediately clear that \( \bar{\lambda}^{(D,W)} \leq \lambda^{(D,W)} \).

b) For an \( M/G_\infty/c/c \) queueing system, the throughput \( \theta \) is equal to \( \lambda (1 - \pi_e) \); in terms of our queueing models we have: \( \theta^B = \lambda (1 - \pi_{N_c}^B) \) and \( \theta^D = \theta^{DA} + \theta^{DH} = p\lambda (1 - \pi_{N_c}^{DA}) + (1 - p)\lambda (1 - \pi_{N_c}^{DH}) \). Thus, \( \theta^D(p) \geq \theta^B \) if and only if \( p\pi_{N_c}^{DA} + (1 - p)\pi_{N_c}^{DH} \leq \pi_{N_c}^B \). Taking the limit of \( \theta \) for any \( M/G_\infty/c/c \) queue as the arrival rate \( \lambda \) approaches \( \infty \), we get the following:

\[
\theta_\infty = \lim_{\lambda \to \infty} \theta = \lim_{\lambda \to \infty} \lambda (1 - \pi_e) = \lambda \left[ 1 - \frac{\lambda^e/(c!V_c \cdots V_1)}{1 + \sum_{n=1}^{\infty} \lambda^n/[n!V_n \cdots V_1]} \right] = \lambda \left[ 1 + \sum_{n=1}^{\infty} \frac{\lambda^n/[n!V_n \cdots V_1]}{1 + \sum_{n=1}^{\infty} \lambda^n/[n!V_n \cdots V_1]} \right] = \lambda \left[ \frac{\lambda^n/[n!V_n \cdots V_1]}{1 + \sum_{n=1}^{\infty} \lambda^n/[n!V_n \cdots V_1]} \right] = \lim_{\lambda \to \infty} \left[ \frac{\lambda^e/(c!V_c \cdots V_1)}{\lambda^e/(c!V_c \cdots V_1)} \right] = cV_e.
\]

Suppose \( \lambda \geq \lambda^{(D,\theta)} = \min_p \max \{ \lambda^{DA}/p, \lambda^{DH}/(1-p) \} \). According to Lemma 2 and our assumption in the statement of the proposition that \( \theta^{DA} \) and \( \theta^{DH} \) are increasing in \( \lambda \), \( \theta^{DA} \) and \( \theta^{DH} \) are increasing and decreasing in \( p \), respectively. Let \( p^* \) be the smallest value of the AV proportion such that \( p^* \lambda \geq \lambda_{jam}^{DA} \) and \( (1-p^*) \lambda \geq \lambda_{jam}^{DH} \). For \( p \geq p^* \) (resp., \( p \leq p^* \)), \( \theta^{DA} \) (resp., \( \theta^{DH} \)) is equal to its jam value, i.e., \( \theta^{DA} = cV_e \) (resp., \( \theta^{DH} = (N-1)cV_e \)). Since \( V_e^{DA} + (N-1)cV_e^{DH} \geq NeV_e^B \), there exists \( \bar{\theta}^{(D,\theta)} \leq p^* \), such that \( (1-p^*)\bar{\theta}^{(D,\theta)} + (N-1)cV_e^{DH} \geq NeV_e^B \), and there exists \( \bar{\theta}^{(D,\theta)} \geq p^* \), such that \( (1-p^*)\bar{\theta}^{(D,\theta)} + (N-1)cV_e^{DH} \geq NeV_e^B \).}

\[\Box\]

**Proof of Corollary 1** a) We want \( \pi^{(D,\theta)} \leq \pi^{(D,W)} \). Substituting \( \bar{\theta}^{(D,\theta)} = \frac{\max \{ \lambda^{DA}/p, \lambda^{DH}/(1-p) \} }{\lambda^{(D,W)}}, \) and \( \pi^{(D,W)} = \frac{V_e^{DA}V_e^{B} - V_e^{DA}V_e^{DH}(N_1-c)}{V_e^{DA}V_e^{B} - V_e^{DA}V_e^{DH}(N_1-c)}, \) we get \( \lambda \geq \bar{\lambda}^{(D,W)} = \frac{\max \{ \lambda^{DA}/p, \lambda^{DH}/(1-p) \} }{V_e^{DA}V_e^{B} - V_e^{DA}V_e^{DH}(N_1-c)(1-s^{DA})}, \) According to Proposition 1, \( p^{(D,W)} \) and \( \pi^{(D,\theta)} \) exist when \( \lambda \geq \bar{\lambda}^{(D,W)} \) and \( \lambda \geq \lambda^{(D,\theta)} \), respectively. \( \bar{\theta}^{(D,\theta)} \leq \pi^{(D,W)} \) if \( \lambda \geq \max \{ \bar{\lambda}^{(D,W)}, \bar{\lambda}^{(D,W)}, \lambda^{(D,\theta)} \}. \)
b) $\hat{p}^{(D,\theta)} \geq (1 - \pi_{(N-1)c}^{DH})\bar{p}^{(D,\theta)} = (1 - \pi_{(N-1)c}^{DH}) - \left(\frac{V_{Nc}^{DA}V_{Nc}^{DH} - V_{c}^{DA}V_{Nc}^{DH}}{V_{c}^{DA}V_{Nc}^{DH} - V_{c}^{DA}V_{Nc}^{DH}}\right)$, and $p^{(D,W)} = \frac{V_{c}^{DA}V_{Nc}^{DH} - V_{c}^{DA}V_{Nc}^{DH}(N-1)c}{V_{c}^{DA}V_{Nc}^{DH} - V_{c}^{DA}V_{Nc}^{DH}(N-1)c}$.

so $\hat{p}^{(D,\theta)} \geq p^{(D,W)}$ if $\lambda \geq \lambda^{(D,W)}$. According to Proposition 1, $p^{(D,W)}$ and $\hat{p}^{(D,\theta)}$ exist when $\lambda \geq \lambda^{(D,W)}$ and $\lambda \geq \lambda^{(D,\theta)}$, respectively. $\hat{p}^{(D,\theta)} \geq p^{(D,W)}$ if $\lambda \geq \max\{\lambda^{(D,W)}, \lambda^{(D,\theta)}\}$. ■

Proof of Proposition 2

a) As mentioned in Proof of Proposition 1, $W_{\infty} = 1/V_{c}$, so $W_{c}^{l}(p) = 1/V_{Nc}^{l}(p) \leq W_{\infty}^{B} = 1/V_{Nc}^{B}$ if and only if $V_{Nc}^{l}(p) \geq V_{Nc}^{B}$. Moreover, according to Lemma 1, $W_{\infty}^{B}$ and $W^{l}(p)$ are both increasing in $\lambda$. Thus, there exists a $\lambda^{1,(W)}$ such that for $\lambda \geq \lambda^{1,(W)}$, $W_{\infty}^{B}(p) \leq W_{\infty}^{B}$ if and only if $V_{Nc}^{l}(p) \geq V_{Nc}^{B}$.

Proof of (15): By (3), $V_{Nc}^{l}(p) \geq V_{Nc}^{B}$ if and only if $h_{Nc}^{l}(p) \leq h_{Nc}^{B}$. Substituting $h_{Nc}^{B}$ with $\frac{\delta B}{\xi_{Nc}} + \frac{1 - \delta B}{\xi_{Nc}}$ and $h_{Nc}^{l}(p)$ with $\delta^{l}(\frac{(1-p)^2}{\psi_{Nc}^{DA}} + \frac{p(1-p)}{\psi_{Nc}^{DA}} + \frac{p^2}{\psi_{Nc}^{DA}} + (1 - \delta^{l})(\frac{(1-p)^2}{\psi_{Nc}^{DA}} + \frac{p(1-p)}{\psi_{Nc}^{DA}} + \frac{p^2}{\psi_{Nc}^{DA}})$, we get the following:

$$\frac{\delta B}{\psi_{Nc}^{DA}} + \frac{1 - \delta B}{\xi_{Nc}} \geq \delta^{l}(\frac{(1-p)^2}{\psi_{Nc}^{DA}} + \frac{p(1-p)}{\psi_{Nc}^{DA}} + \frac{p^2}{\psi_{Nc}^{DA}} + (1 - \delta^{l})(\frac{(1-p)^2}{\psi_{Nc}^{DA}} + \frac{p(1-p)}{\psi_{Nc}^{DA}} + \frac{p^2}{\psi_{Nc}^{DA}}) = p\delta^{l}(\frac{2-p}{\psi_{Nc}^{DA}} + \frac{1-p}{\psi_{Nc}^{DA}} - \frac{p}{\psi_{Nc}^{DA}}) + p(1 - \delta^{l})(\frac{2-p}{\xi_{Nc}^{DA}} + \frac{1-p}{\xi_{Nc}^{DA}} - \frac{p}{\xi_{Nc}^{DA}}) \geq \delta^{l} \frac{\xi_{Nc}^{AA} - \xi_{Nc}^{IH}}{\xi_{Nc}^{IH} - \xi_{Nc}^{AA}} + \frac{1 - \delta^{l} - 1 + \delta^{l} - 1}{\xi_{Nc}^{IH} - \xi_{Nc}^{AA}} \cdot \frac{\xi_{Nc}^{IH} - \xi_{Nc}^{AA}}{\xi_{Nc}^{IH} - \xi_{Nc}^{AA}}$$

Canceling $p$’s out, and rearranging the right hand side, we get the following expression:

$$\delta^{l}(\frac{2-p}{\psi_{Nc}^{DA}} - \frac{1-p}{\psi_{Nc}^{DA}} - \frac{p}{\psi_{Nc}^{DA}}) + (1 - \delta^{l})(\frac{2-p}{\psi_{Nc}^{DA}} - \frac{1-p}{\psi_{Nc}^{DA}} - \frac{p}{\psi_{Nc}^{DA}}) \geq (\frac{\xi_{Nc}^{AA} - \xi_{Nc}^{IH}}{\psi_{Nc}^{DA}} + \frac{1 - \delta^{l} - 1 + \delta^{l} - 1}{\xi_{Nc}^{IH} - \xi_{Nc}^{AA}}).$$

b) As mentioned in Proof of Proposition 1, $\theta_{\infty} = \frac{c}{cV_{c}}$, so $\theta_{\infty}^{l}(p) = NcV_{Nc}^{l}(p) \geq \theta_{\infty}^{B} = NeV_{Nc}^{B}$ if and only if $V_{Nc}^{l}(p) \geq V_{Nc}^{B}$. Moreover, according to Lemma 2, $\theta_{\infty}$ and $\theta^{l}(p)$ are both increasing in $\lambda$. Thus, there exists a $\lambda^{1,\theta}$ such that for $\lambda \geq \lambda^{1,\theta}$, $\theta_{\infty}(p) \geq \theta_{\infty}^{B}$ if and only if $V_{Nc}^{l}(p) \geq V_{Nc}^{B}$. ■

Proof of Proposition 3

In this proof we assume that the conditions stated in Propositions 1 and 2 for the existence of thresholds of $p$ hold.

a) As $p$ approaches zero, $V_{c}^{l}(p) \rightarrow V_{c}^{B}$, so $W^{l}(p) \rightarrow W^{B}$ and $\theta^{l}(p) \rightarrow \theta^{B}$. Moreover, $\lim_{p \rightarrow 0} W_{c}^{DA}(p) = \lim_{p \rightarrow 0} pW_{c}^{DA} + (1-p)W_{c}^{DH} = W_{c}^{DH} = 1/V_{Nc}^{DH} = W_{c}^{B}$, and $\lim_{p \rightarrow 0} \theta^{l}(p) = \lim_{p \rightarrow 0} p\theta^{DA} + (1 - p)\theta^{DH} = \theta^{DH} = (N-1)eV_{Nc}^{DH} \leq NeV_{Nc}^{B} = \theta^{B}$.

b) As $p$ approaches one, $V_{c}^{l}(p) \rightarrow V_{c}^{DA}$, so $W^{l}(p) \rightarrow W^{B}$. Also, since $V_{c}^{l}(p)$ is increasing in $p$, $W_{c}^{DA}(p) = 1/V_{Nc}^{l}(1) \leq W_{c}^{B} = 1/V_{Nc}^{B}$. Similarly, $\theta^{l}(p) = NcV_{Nc}^{l}(1) \geq \theta_{\infty} = NeV_{Nc}^{B}$. However, $\theta^{l}(p) \rightarrow NeV_{Nc}^{B}$, $cV_{Nc}^{l}(p) = cV_{c}^{DA} = \theta^{DA} = \lim_{p \rightarrow 1} \theta^{D}$. Moreover, $\lim_{p \rightarrow 1} \theta^{D} = \theta^{DA} = cV_{c}^{DA} \geq NeV_{Nc}^{B} = \theta^{B}$ if and only if $V_{Nc}^{DA} \geq V_{c}^{DA}$. 

c) We first analyze $W$. Because we assume that $V_{Nc}^{l}(p)$ is increasing in $p$, $W^{l}(p) = 1/V_{Nc}^{l}(p) \leq 1/V_{Nc}^{l}(0) = 1/V_{Nc}^{B} = W^{B}$. For a given $p \in [0,1]$, $W^{l}(p) = \frac{p}{V_{Nc}^{DA}} + \frac{1-p}{V_{Nc}^{DA}(N-1)c} \leq 1/V_{Nc}^{c}(p) = W^{l}(p)$, if and only if
\[- \frac{V_{DA}V_{DH}^{(1)c}}{V_{DA} - V_{DH}^{(1)c}} \leq V_{Nc}(p) \left( p - \frac{V_{eDA}}{V_{DA} - V_{DH}^{(1)c}} \right);\] derived by rearranging the first inequality. Note that in this inequality $V_{eDA} - V_{DH}^{(1)c} \geq 0$, because $V_{eDA} = V_{Nc}(1) \geq V_{DH}^{B} \geq V_{DH}^{(1)c}$. The right hand side of this inequality, i.e., $V_{Nc}(p) \left( p - \frac{V_{eDA}}{V_{DA} - V_{DH}^{(1)c}} \right)$, is increasing in $p$. Thus, $W(p) \leq \tilde{W}(p)$, if and only if $p \geq p(D,W)$, where $p(D,W)$ is the smallest $p$ such that $- \frac{V_{DA}V_{DH}^{(1)c}}{V_{DA} - V_{DH}^{(1)c}} \leq V_{Nc}(p) \left( p - \frac{V_{eDA}}{V_{DA} - V_{DH}^{(1)c}} \right)$.

We need to show that $p(D,W) \geq p(D,W)$. If $p \leq p(D,W)$, $W(p) \geq W_B$, i.e., $\frac{p}{V_{DA}} + \frac{1-p}{V_{DH}^{(1)c}} \geq 1/V_{Nc} \geq 1/V_{Nc}(p) = \tilde{W}(p)$. The last inequality holds because we assume that $V_{Nc}(p)$ is increasing in $p$, and $V_n(0) = V_B$. Thus, for $p \leq p(D,W)$, $\tilde{W}(p) \leq W(p)$, and $p(D,W) \geq p(D,W)$.

Next, we analyze $\theta$ similar to the proof of Proposition 1. Three possible scenarios can occur. First, if there is no $p \in [0,1]$ that satisfies $p(1-\pi_{eDA} - (1-p)\lambda(1-\pi_{DH}^{(1)c})) = NcV_{Nc}(p)$, then $\theta(1) > \theta(p)$, and $p(D,\theta)$ and $\tilde{p}(D,\theta)$ do not exist. Second, if $p(1-\pi_{eDA} + (N-1)cV_{DH}^{(1)c}) = NcV_{Nc}(p)$ has two roots, then $p(D,\theta)$ (resp., $\tilde{p}(D,\theta)$) is the smallest (resp., largest) $p$ that satisfies this inequality. Lastly, if $p(1-\pi_{eDA} - (N-1)cV_{DH}^{(1)c}) = NcV_{Nc}(p)$ has a unique root (i.e., $\tilde{p}(D,\theta)$), then $cV_{eDA} + (1-p)\lambda(1-\pi_{DH}^{(1)c}) \geq NcV_{Nc}(p) = \theta(p)$ has a unique root ($p(D,\theta)$) as well. In this case, $p(D,\theta) \leq \tilde{p}(D,\theta)$ holds, because at $\tilde{p}(D,\theta)$ the AV queue is jammed, so this value of $p$ is higher than the value of $p$ at which the AV queue is not jammed, i.e., $p(D,\theta)$.

Since we assume that $V_{Nc}(p)$ is increasing in $p$ and it is concave (or convex) everywhere, $\theta(p)$ is also increasing in $p$, and $\theta(p) \geq \theta(p)$, if and only if $p(D,\theta) \leq p \leq \tilde{p}(D,\theta)$. Lastly, $p(D,\theta) \leq \tilde{p}(D,\theta) \leq \tilde{p}(D,\theta)$, because $V_{Nc}(p) \geq V_B$.

**Proof of Corollary 2** As we discussed in the Proof of Proposition 3, $p(D,\theta) \leq \tilde{p}(D,\theta)$. To show that $p(D,\theta) \leq p(D,W)$, we consider two cases: $p(D,W) \leq \frac{V_{eDA}}{V_{DA} - (N-1)V_{DH}^{(1)c}}$ and $p(D,W) \geq \frac{V_{eDA}}{V_{DA} + (N-1)V_{DH}^{(1)c}}$.

**Case 1** ($p(D,W) \leq \frac{V_{eDA}}{V_{DA} + (N-1)V_{DH}^{(1)c}}$): Suppose $p \leq p(D,W)$, i.e., $1/V_{Nc}(p) \leq p/V_{eDA} + (1-p)/V_{DH}^{(1)c}$.

Let $\lambda(1-\pi_{eDA}) = cV_{eDA}^{(1)c}$. By definition $p \leq \tilde{p}(D,\theta)$, if $\frac{cV_{eDA} + (N-1)cV_{DH}^{(1)c}}{Nc} \leq V_{Nc}(p)$, or equivalently $\frac{cV_{eDA} + (N-1)cV_{DH}^{(1)c}}{V_{Nc}(p)} \geq 1/V_{Nc}(p)$. Thus, if $cV_{eDA} + (N-1)cV_{DH}^{(1)c} \geq p/V_{eDA} + (1-p)/V_{DH}^{(1)c}$, then $p \leq \tilde{p}(D,\theta)$.

Rearranging this inequality we get: if $p \geq \frac{(V_{DH}^{(1)c} - V_{eDA})(V_{eDA} + (N-1)V_{DH}^{(1)c})}{V_{eDA} - V_{DH}^{(1)c}}$, then $p \leq \tilde{p}(D,\theta)$.

Note that the numerator of the right hand side of this inequality is positive, because if $V_{DH}^{(1)c} \leq V_{eDA}$,

$p \geq \frac{V_{eDA}}{V_{DA} + (N-1)V_{DH}^{(1)c}} \geq \frac{V_{eDA}}{V_{DA} - V_{DH}^{(1)c}}$, and this is a contradiction. If $V_{DH}^{(1)c} \geq V_{eDA}$, $p \geq 0 \geq \frac{V_{eDA} - V_{DH}^{(1)c}}{V_{eDA} - V_{DH}^{(1)c}}$.

**Case 2** ($p(D,W) \geq \frac{V_{eDA}}{V_{DA} + (N-1)V_{DH}^{(1)c}}$): Suppose $p \leq p(D,W)$, i.e., $1/V_{Nc}(p) \leq p/V_{eDA} + (1-p)/V_{DH}^{(1)c}$.

In this case, $1/V_{Nc}(p) \leq p/V_{eDA} + (1-p)/V_{DH}^{(1)c} \leq \frac{N}{V_{eDA} + (N-1)V_{DH}^{(1)c}}$, and only if $p \geq \frac{V_{eDA}}{V_{DA} + (N-1)V_{DH}^{(1)c}}$. In other words, if $p \geq \frac{V_{eDA}}{V_{DA} + (N-1)V_{DH}^{(1)c}}$, $\theta(p)$ is higher than the jam through-
put of the D policy, i.e., $cV_c^{DA} + (N - 1)cV_c^{DH}_{(N-1)c}$. Thus $\theta^I(p) \geq \theta^O(p)$ when $p \geq \frac{V_c^{DA}}{V_c^{DA} + (N - 1)V_c^{DH}_{(N-1)c}}$. Therefore, $p^{(DI,\theta)}_{V_c^{DA}}$, the smallest value of $p$ at which $\theta^O(p)$ becomes higher than $\theta^I(p)$, cannot be higher than $\frac{V_c^{DA}}{V_c^{DA} + (N - 1)V_c^{DH}_{(N-1)c}}$. Moreover, $p^{(DI,W)}_{V_c^{DA}} \geq \frac{V_c^{DA}}{V_c^{DA} + (N - 1)V_c^{DH}_{(N-1)c}}$, thus $p^{(DI,\theta)} \leq p^{(DI,W)}$. ■

Appendix D: Parameter Estimation

D.1. State-dependent Speed Curve

We follow the transportation literature to fit a function to our speed-volume data from Arizona. As explained in Del Castillo and Benitez (1995) and Jain and Smith (1997), there exist various functions to represent the relationship between speed and volume. Tiwari and Marsani (2014) provide a summary of these functions. We fit all different functions mentioned in Tiwari and Marsani (2014) to our data. Specifically, several linear, logarithmic, polynomial, and exponential functions were tested, and the functional form $y = ae^{-x^2/b} + c$ gave us the highest coefficient of determination which is 85%.

D.2. HV Mean Platoon Size

To estimate the mean platoon size for HVs we use data. The data from Arizona Department of Transportation do not include the headway between vehicles, so we use another data set from the Institute for Transportation
of Iowa State University. This data set consists of more than 314,000 instances of headways between vehicles, ranging from milliseconds to hundred of seconds, and is collected in 2015 from several highways in Iowa, including I-74 and I-80. In order to distinguish between interplatoon and intraplatoon headways in this data set, we use the mixtools library in R to divide the headways into two clusters: one for the smaller headway values corresponding to the intraplatoon headways and the other for the larger headway values corresponding to the interplatoon headways. Based on the posterior distribution of the clusters, if a headway value is larger than 2.355 seconds, the probability that it belongs to the intraplatoon headway cluster is less than $10^{-7}$. Thus, when the headway between two consecutive vehicles is less than 2.355 seconds, we assume they belong to the same platoon; otherwise they are in two separate ones. Counting the number of consecutive vehicles in the same platoon, we get a sample of platoon size values. Among different discrete distributions, a geometric distribution with parameter 0.667 fits this sample well (see Figure (8)). Thus, we set the mean platoon size of HVs equal to $1/0.667 = 1.5$ vehicles.

D.3. Safe Stopping Time

To estimate the mean interplatoon headway of AVs, we use data from National Highway Traffic Safety Administration (NHTSA 2015). This dataset provides a set of $(d_n, V_n)$ pairs, where $d_n$ is the safe distance (in miles) from the vehicle in front to stop a vehicle when driving at speed $V_n$ (in miles per hour); this applies to both human-driven and autonomous vehicles. Figure 9 shows the pairs of $(d_n, V_n)$ provided by NHTSA, and the fitted line to this data. The value of coefficient of determination, $R^2$, for this curve is 97.6%. The linear regression model fitted to these data points is $d_n = 0.001 V_n - 0.006$ (miles). Noting that safe stopping time is $d_n / V_n$, we obtain the mean interplatoon headway in the AV queue, $1/\psi_n^{DA}$, in (9).

Appendix E: Additional Results

E.1. Performance of Lightly Loaded Highways with AVs

As shown in Figure 10(a), for a lightly loaded highway with $\lambda = 2,217$ vehicles per hour, when $p \leq 0.64$, the mean travel time $W$ under the D policy is higher than in the benchmark case. When $p > 0.64$, this policy works as well, but not better than the benchmark case, because at this value of $\lambda$ the average speed of vehicles on the highway in the benchmark case is already high, so that adding AVs does not improve the system further. The same argument holds for the throughput $\theta$; see Figure 10(b).

Figure 11(a) compares the mean travel time $W$ between the benchmark and those policies with AVs when $\lambda = 2,217$ vehicles per hour. Under the I policy, one can observe from Figure 11(a) that the maximum improvement in $W$ is only two seconds (0.036 minutes). Also, as shown in Figure 11(b), at this value of $\lambda$ the throughput of the benchmark case is equal to the arrival rate which is the maximum achievable throughput, and thus adding AVs does not improve the throughput.

Comparing the D policy with the I policy, Figure 11 also shows that the latter performs better than the former. However, under the I policy, AVs improve traffic flow only marginally, and therefore a choice between

---

20 We are not able to use this data in estimating the mean interplatoon and intraplatoon headways, since these parameters depend on the number of vehicles present on the highway, $n$, and the data from Iowa State University do not include $n$. 
these two policies does not have a significant impact on congestion. Table A3 shows the maximum percentage of improvement in $W$ and $\theta$ over the benchmark case, for several values of $\lambda \leq \lambda^{(D,W)} = \lambda^{(D,\theta)} = 2,510$. As this table shows, under the I policy, the increase in $\theta$ is almost zero, and $W$ decreases by at most 7%. Since, at all these values of $\lambda$, $W^B < 1$ minute, a 7% improvement in $W$ is insignificant.

In general, when $\lambda$ is low, unlike the I policy that performs at least as well as the benchmark case, the D policy may not have as good performance. In this case, HVs are capable of driving at the free-flow speed in the benchmark case, and assigning one lane to AVs slows HVs down, while AVs do not improve the speed significantly over the benchmark case.
<table>
<thead>
<tr>
<th>$\lambda$ (vehicles per hour)</th>
<th>$W_B$ (min.)</th>
<th>$\theta_B$ (vehicles per hour)</th>
<th>Max increase in $W$ (%)</th>
<th>Max increase in $\theta$ (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>500</td>
<td>0.810</td>
<td>500</td>
<td>0.29</td>
<td>0</td>
</tr>
<tr>
<td>1000</td>
<td>0.811</td>
<td>1000</td>
<td>0.99</td>
<td>0</td>
</tr>
<tr>
<td>1500</td>
<td>0.821</td>
<td>1500</td>
<td>2.13</td>
<td>0</td>
</tr>
<tr>
<td>2000</td>
<td>0.835</td>
<td>2000</td>
<td>3.76</td>
<td>0</td>
</tr>
<tr>
<td>2500</td>
<td>0.863</td>
<td>2499.97</td>
<td>6.93</td>
<td>0.001</td>
</tr>
</tbody>
</table>

Table A3  Effect of the I policy on the performance measures for lightly loaded highways

E.2. State-Dependent Speed under the D Policy

Under the D policy, the speed of AVs is higher than that of HVs for any given number of vehicles in one lane. To understand this result, recall from equation (3) that speed is a function of mean headway $h = \frac{1}{\xi} + \frac{1}{\psi}$, which is a weighted average of the mean intraplatoon headway $1/\xi$, and the mean interplatoon headway $1/\psi$, with weights determined by the mean platoon size $1/\delta$. In the following, we first compare $1/\xi$ between HVs and AVs, and then $1/\psi$.

First, we compare the weighted mean intraplatoon headways, $1/\xi$. For the benchmark case, the mean platoon size is $1/\delta_B = 1.5$ vehicles (see Appendix D.2 for details of this estimation), and the mean intraplatoon headway is $1/\xi_B = 1.1$ seconds (Tientrakool et al. 2011). Although the mean intraplatoon headway of vehicles in the AV queue (0.55 seconds) is shorter than that of HVs in the benchmark case (1.1 seconds), due to the large mean platoon size of AVs (10), the weighted mean intraplatoon headway of vehicles in the AV queue, $1/\xi_{DA} = 0.495$ seconds, is longer than that in the benchmark case, $1/\xi_B = 0.367$ seconds. These values are independent of $n$, and their difference is in the order of milliseconds.

Next, we compare the weighted mean interplatoon headways, $1/\psi$. We can derive the mean interplatoon headway of HVs by rearranging equation (4) as $1/\psi_B = \frac{\frac{1}{\zeta_B} - (1-\delta_B)^n V_B}{a V_B^\alpha (1-\delta_B)^n \xi_B}$. We plot $1/\psi_n$ and $1/\psi_n^{DA}$ in equation (9) in Figure 12(a), which shows that the mean interplatoon headway of vehicles in the AV queue is not always lower than that in the benchmark model. However, Figure 12(b) illustrates that, since the mean
size of AV platoons is much higher than that of HV platoons, the weighted mean interplatoon headway of vehicles in the AV queue ($\delta_{DA}^A/\psi_{DA}^A$) is always smaller than that in the benchmark case ($\delta_{B}^B/\psi_{B}^B$), and their minimum difference is 1.15 seconds. Finally, we combine $1 - \delta$ and $\delta/\psi$. Note that the difference in the weighted mean intraplatoon headway ($= 0.495 - 0.367$ seconds) is always smaller than the difference in the weighted mean interplatoon headway ($\geq 1.15$ seconds). Therefore, AVs maintain a lower mean headway than HVs, and the speed of vehicles in the AV queue is higher than the speed of vehicles in the benchmark case as well as in the HV queue for any $n/N$.

E.3. State-Dependent Speed under the I Policy

We observe that, for very small values of $p$ (i.e., $p < 0.08$), the speed in the benchmark case can be higher than that under the I policy. In other words, for such $p$, there exist values of $n$ such that $V_{I}^n(p) \leq V_{B}^n$. This occurs because for small $p$ and $n$ the mean intraplatoon headway and the mean interplatoon headway can be both increasing in $p$. As $p$ increases further, $V_{I}^n(p) > V_{B}^n$ for all $n = 1, 2, \ldots, 555$. As depicted in Figure 3(b), when $n$ is high ($n = 250, 350, \text{and} 450$), $V_{I}^n(p)$ increases with $p$, but for lower values of $n$ ($n = 150$), $V_{I}^n(p)$ first decreases with $p$ and then increases. As mentioned in §4.2, the mean headway $h$ is the determining factor of speed at each value of $n$, and $h$ is the weighted average of the mean intraplatoon headway $((1 - \delta)/\xi_{I}^n)$ and the mean interplatoon headway $$(\delta/\psi_{I}^n)$$. Thus, to understand this result, we first discuss each of these headways and then combine them.

First, Figure 13(a) depicts that the weighted mean intraplatoon headway, which is independent of $n$, is first increasing and then decreasing in $p$. The intraplatoon headway of the HV-HV pair (1.1 seconds) is higher than that of the AV-AV pair (0.55 seconds), and lower than that of the HV-AV pair (1.4 seconds). When the proportion of AVs ($p$) is small (high), HV-AV pairs are more (less) prevalent than AV-AV pairs; hence, as $p$ increases, the weighted mean intraplatoon headway increases (decreases).

Next, Figure 13(b) illustrates the weighted mean interplatoon headway, which is the product of $\delta/I$ and $1/\psi_{n}^I$. While $\delta/I$ is decreasing in $p$, $1/\psi_{n}^I$ is not always decreasing in $p$. According to (11), $1/\psi_{n}^I$ is the weighted
average of the mean interplatoon headway of AVs and that of HVs. There exist values of $n$ such that the mean interplatoon headway of AVs is higher than that of HVs ($n/N$ is between 20 and 60), so $1/\psi^n_I$ increases with $p$ at these values of $n$, and decreases otherwise. When $n$ is high (e.g., $n = 250, 350$, or $450$), as $p$ increases, the number of platoons decreases ($\delta_I$ decreasing in $p$), and the probability of having an AV as the leader of a platoon increases. Because the interplatoon headway maintained by an AV is set to the safe stopping time, which is lower than what a HV maintains for these values of $n$, the mean interplatoon headway $(1/\psi^n_I)$ decreases with $p$. Thus, for large values of $n$, $\delta_I/\psi^n_I$, is a decreasing function of $p$. When $n$ is low (e.g., $n = 150$), $\delta_I$ is decreasing in $p$, but $1/\psi^n_I$ can be increasing in $p$. For small values of $p$, the effect of $1/\psi^n_I$ outweighs the effect of $\delta_I$, and for higher values the opposite is true. As a result, $\delta_I/\psi^n_I$ is increasing in $p$ when $p$ is small enough (e.g., at $p = 0.01$ and $n = 150$), and it is decreasing in $p$ for higher values of $p$.

Finally, when at least one of $n$ or $p$ is high, since the rate of reduction in the weighted mean interplatoon headway is higher than the rate of increase in the weighted mean intraplatoon headway, the mean headway of vehicles is decreasing in $p$, and AVs increase the speed of vehicles by reducing the weighted mean interplatoon headway. In contrast, when $p$ and $n$ are both low, AVs decrease the speed of vehicles by forming larger platoons and increasing the mean platoon headway.

**E.4. Intuition behind the Sharp Decrease in $W_{DH}$**

The behavior of mean travel time as a function of the arrival rate depends on the steady-state speed of vehicles. Since speed depends on the number $n$ of vehicles on the highway, to calculate the steady-state speed, one needs to know the steady-state probability distribution of $n$, $\pi_n$, which depends on the arrival rate $\lambda$ and the service rate $nV_n$. Intuitively, when there are $n$ vehicles on the highway such that $nV_n \geq \lambda$ (resp., $nV_n \leq \lambda$), vehicles exit the highway faster (resp., slower) than they enter, and therefore the number of vehicles decreases (resp., increases) over time. As a result, for any given $n$, if $nV_n \neq \lambda$, the probability of having $n$ vehicles on the highway is very low. This implies that with a high probability there are $[\eta]$ vehicles on the highway in steady-state such that $\eta V_n = \lambda$ (where $\eta$ may not be unique).
In the HV queue of the D policy, as shown in Figure 4(a), the mean travel time, $W_{DH}$, decreases sharply at $p = 0.93$. This sharp decrease in $W_{DH}$ happens due to the significant difference in the steady-state speed of vehicles, $1/\sum_{n=1}^{c} \pi_n (L/V_n)$, when $p = 0.93$ and that when $p = 0.94$. Let us first consider $p = 0.93$. For this $p$, Figure 14(a) shows that, for $n < 11$ and $126 < n < 370$, $(1-p)\lambda > nV_n$, and therefore the number of vehicles tends to increase over time to the next point such that $(1-p)\lambda = nV_n$, i.e., 11 and 370, respectively. For $11 < n < 126$, $(1-p)\lambda < nV_n$, and the number of vehicles tends to decrease over time to the last point where $(1-p)\lambda = nV_n$, i.e., 11. As a result, $[\eta] \in \{11, 370\}$, and with a high probability there are either 11 or 370 vehicles on the highway in steady state. Figure 14(b) illustrates that the probability of having 370 vehicles is about 9 times the probability of having 11 vehicles. As a result, in steady state, there is a high chance that there are 370 vehicles on the highway driving at $V_{370}^{DH} = 2$ miles per hour, and the mean travel time $W_{DH}$ is primarily determined by this low speed of 2 miles per hour. Next, consider $p = 0.94$. For this $p$, Figure 15(a) displays that $[\eta]$ is equal to either 10 or 341, but the probability of having 10 vehicles is much higher than that of 341; see Figure 15(b). As a result, with a high chance vehicles drive at $V_{10}^{DH} = 68$ miles per hour, and $W_{DH}$ is primarily determined by this high speed of 68 miles per hour.

It is interesting to observe the value of $[\eta]$ at which the maximum $\pi_n$ is attained changes dramatically from $[\eta] = 370$ at $p = 0.93$ to $[\eta] = 10$ at $p = 0.94$. We can understand this better by inspecting (13) closely as follows. As $p$ increases from 0.93 to 0.94, $\lambda$ is the only parameter that is changed in $\pi_n$ given in (13). $\pi_n$ consists of two components: $\pi_0 = (1 + \sum_{n=1}^{c} \frac{(L\lambda)^n}{n!V_n-V_i})^{-1}$, and $\frac{(L\lambda)^n}{n!V_n-V_i}$. Due to the decrease in $\lambda$ from $(1-0.93)11,342 = 794$ at $p = 0.93$ to $(1-0.94)11,342 = 681$ at $p = 0.94$, $\pi_0$ is 80 times greater at $p = 0.94$ than that at $p = 0.93$. Furthermore, when $n = 10$ (resp., 370), the $\frac{(L\lambda)^n}{n!V_n-V_i}$ term of $\pi_n$ is 80 (resp., $10^{24}$) times lower at $p = 0.94$ than that at $p = 0.93$. As a result, $\pi_{10} = 0.007$ at $p = 0.93$ is significantly lower than $\pi_{10} = 0.125$ at $p = 0.94$, whereas $\pi_{370} = 0.1$ at $p = 0.93$ is significantly higher than $\pi_{370} = 1.39 \times 10^{-24}$ at $p = 0.94$.

Formation of a spontaneous jam (having an abrupt decrease in $W$) for HVs has been observed in prior literature: Bando et al. (1995) and Treiber et al. (2000) show that there exists a critical traffic density at which
the highway becomes jammed. However, analyzing AVs, we observe that this is not a universal behavior. Figure 4(a) illustrates that $W^{DA}$ increases fairly smoothly from less than one minute to 2.5 minutes. As depicted in Figure 16(a), in this case, $nV_n$ is strictly increasing in $n$, so for each value of the arrival rate $p\lambda$, $\eta$ is unique; for example, at $\lambda = 681$ vehicles per hour, $\eta = 10$, and at $\lambda = 794$ vehicles per hour, $\eta = 11$. Therefore, unlike the HV queueing system, a small increase in $p$ does not result in a substantial increase in $\eta$, and hence $\pi_\eta$.

E.5. D Policy with Two AV Lanes

We numerically analyze a designated-lane policy with two AV lanes (“D2 policy”). As mentioned in §5.3, when $p$ is high, i.e., $p \geq \bar{p}^{(D1,\theta)}$, since the D policy dedicates only one lane to the majority of vehicles, $\theta^{D}(p)$
becomes lower than $\theta^I(p)$. For high values of $p$, it seems intuitive to increase the number of designated lanes to AVs under the D policy.

Figure 17(a) illustrates mean travel time under different policies for $\lambda = 11,342$ vehicles per hour. $^{21}$ We observe that the D2 policy reduces $W$ slightly over the D policy, but the I policy is still superior to both of these policies. First, comparing $W^{D2}(p)$ with $W^B$, we observe that, similar to the D policy, $W^{D2}(p) \leq W^B$ only $p$ is higher than a threshold, i.e., $p \geq 0.58$. Second, although $W^{D2}(p)$ shows a similar pattern to $W^D(p)$, the sharp decrease in $W^{D2}(p)$ happens at a higher $p$ (i.e., $p = 0.96$) than it does in $W^D(p)$ (i.e., $p = 0.93$). As discussed in §5.1, this sharp decrease happens when the HV queue is not highly loaded anymore. Since the number of HVs lanes under the D2 policy is one fewer than that under the D policy, the HV queue under the D2 policy becomes lightly loaded at a higher $p$ than it does under the D policy. Lastly, $W^{D2}(p)$ is never lower than $W^I(p)$. This happens because the HV queue under the D2 policy stays heavily loaded unless $p$ is very high, i.e., $p \geq 0.96$, and the high $W$ of this queue negatively affects $W^{D2}(p)$.

Figure 17(b) compares throughput under different policies. In summary, for highly loaded highways, adding more lanes helps improve throughput when a proportion of AVs is moderately high, but it has a minimal impact on mean travel time. When $p$ is very low or very high (i.e., $p \leq 0.25$ or $p \geq 0.87$), the I policy performs the best; when $p$ is moderately low (i.e., $0.25 \leq p \leq 0.44$), the D policy performs the best; and when $p$ is moderately high (i.e., $0.44 \leq p \leq 0.87$), the D2 policy performs the best. This is intuitive, because the I policy always improve $\theta$ over the benchmark case, but in order for the D and D2 policies to increase $\theta$, $p$ should be moderate, so that both AV and HV queues are well utilized.

$^{21}$ When $p$ is low, we expect $W^D(p)$ to be lower than $W^{D2}(p)$, but in our numerical analysis, they are equal. Due to a lack of data for a one-lane highway, we set $V^{DH2}_o$ equal to $V^{DH}_o$, which is higher than the actual speed of vehicles on a one-lane highway with no lane-changing.