Resilient Operations of Smart Highways: Platooning, Ramp Metering, and Incident Management

by

Li Jin

B.Eng., Shanghai Jiao Tong University (2011)
M.S., Purdue University (2012)

Submitted to the Department of Civil and Environmental Engineering in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Transportation

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2018

© Massachusetts Institute of Technology 2018. All rights reserved.

Signature redacted

Author ....
Department of Civil and Environmental Engineering
August 8, 2018

Signature redacted

Certified by.............. ....
Saurabh Amin
Robert N. Noyce Career Development Associate Professor
of Civil and Environmental Engineering
Thesis Supervisor

Signature redacted

Accepted by ....
Heidi Nepf
Donald and Martha Harleman Professor
of Civil and Environmental Engineering
Chair, Graduate Program Committee

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
OCT 11 2018
LIBRARIES
ARCHIVES

MITLibraries

77 Massachusetts Avenue
Cambridge, MA 02139
http://libraries.mit.edu/ask

DISCLAIMER NOTICE

Due to the condition of the original material, there are unavoidable flaws in this reproduction. We have made every effort possible to provide you with the best copy available.

Thank you.

The images contained in this document are of the best quality available.

Resilient Operations of Smart Highways: Platooning, Ramp Metering, and Incident Management

by

Li Jin

Submitted to the Department of Civil and Environmental Engineering
on August 1, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Transportation

Abstract

Highway systems have witnessed a significant modernization in recent years due to the deployment of traffic sensing and control capabilities. In addition, the ongoing developments in connected and autonomous vehicle technology are poised to enable advanced capabilities such as platooning and vehicle-to-infrastructure communications. On one hand, these advancements offer new opportunities for improving the operational efficiency of highway systems. On the other hand, most highway system operators still face significant challenges in ensuring adequate performance under disruptions such as incidents and other capacity-reducing events, as well as demand fluctuations. Furthermore, the inherent vulnerabilities of cyber-physical components in smart highway systems are prone to exploitation by adversarial agents, who can introduce strategic disruptions. Thus, ensuring the resiliency of highway operations is a principal concern of system operators.

In this thesis, we contribute to the above-mentioned challenge by developing a system-theoretic approach for maintaining resilient highway operations under a broad range of disruptions, modeled as stochastic perturbations in highway capacity or traffic demand. In particular, we focus on three types of highway operations: vehicle platooning, ramp metering, and capacity-aware routing/demand management. Our approach relies on (i) modeling partially automated traffic flow dynamics under disruptions as stochastically switching dynamical systems, (ii) analyzing their long-time properties (stability and/or convergence), and (iii) designing traffic control schemes that improve system throughput with stability guarantees. We demonstrate the application of our approach to several realistic situations ranging from capacity perturbations at incident hotspots to moving bottlenecks created by heavy-duty vehicles to stochastic arrivals/progression of autonomous vehicle platoons.

To model traffic flow dynamics under disruptions, we extend classical macroscopic traffic flow/queuing models by combining them with Markovian switches in flow/queuing dynamics that capture the stochasticity in occurrence/clearance of disruptions. Specifically, we propose two models: Piecewise-Deterministic Queuing (PDQ) model, and Stochastic Switching Cell Transmission Model (SS-CTM). The

PDQ model is the most basic model that captures the dynamic evolution of a traffic queue upstream of a highway bottleneck under perturbations in capacity or demand. We use this model to analyze link-level capacity management schemes and design capacity-aware routing schemes for parallel-route highway systems. The SS-CTM captures the spatial propagation of a disturbance created by capacity perturbations, and is useful for identifying the congestion bottlenecks induced by these perturbations. We adopt this model to analyze the impact of perturbations on the on-ramp queues and highway throughput as well as to design new ramp control schemes with improved performance guarantees.

Our results on the stability analysis of PDQ and SS-CTM utilize more general results on the stability of continuous-time Markov processes. We refine them for the purpose of evaluating the boundedness of traffic queues upstream of highway bottlenecks and on the ramps. Our key contribution is a computationally tractable approach for verifying the classical Foster-Lyapunov drift condition over a finite subset of states, which happen to be the vertices of an invariant set for the stochastic traffic dynamics. This requires us to exploit the long-time properties of the PDQ and SS-CTM—in particular, the cooperativity of traffic flow dynamics and ergodicity of Markov chain that models disruptions. Our analysis approach enables us to estimate how performance metrics such as throughput and travel time change with location and intensity (rate) of disruptions. We also extend our results to the problem of designing traffic control schemes that improve system throughput under perturbations, while maintaining stable traffic queues. This leads us to identify somewhat surprising ways to prioritize and route traffic on real-world highway systems, and relate them to important operational capabilities such as lane control on automated highways, speed regulation of platoons, incident-aware routing, and stabilization of on-ramp queues.

Finally, we also consider the modeling and impact evaluation of security disruptions. We report an initial game-theoretic model that captures an emerging security concern in multi-priority highway systems. The model is relevant to study the incentives of strategic misbehavior by individual vehicles who can exploit the security vulnerabilities in vehicle-to-infrastructure communications and impact the highway operations. We also discuss strategic response to cyber-physical attacks on smart highway infrastructure for timely recovery of compromised traffic links.

Thesis Supervisor: Saurabh Amin
Title: Robert N. Noyce Career Development Associate Professor
of Civil and Environmental Engineering

Acknowledgments

First, I would like to express my deepest gratitude to my Ph.D. advisor Saurabh Amin. This work is impossible without his guidance and support. During the past five years, he has devoted an enormous amount of time and effort to the supervision of my research and help of my graduate life. I am extremely lucky to have him as my Ph.D. advisor.

I sincerely appreciate the other members of my thesis committee. Profs. Nigel H. M. Wilson, Demosthenis Teneketzis, and Hamsa Balakrishnan have provided valuable feedback for this thesis work. Their supervision in the past two years has been helpful in terms of both the extension of the general scope of my research and the refinement of specific results. I enjoyed my interaction with them, and it is a privilege for me to have these three outstanding researchers and mentors on my thesis committee.

I am grateful to a number of great mentors and valuable collaborators. In particular, I am very thankful to Prof. Patrick Jaillet (research collaborator as well as teaching assistant mentor) and Mrs. Manxi Wu (lab mate) at MIT, Prof. Karl H. Johansson and Mr. Mladen Čičić from the KTH Royal Institute of Technology, Dr. Alexander A. Kurzhanskiy from the University of California, Berkeley, Prof. Dengfeng Sun (master's advisor) from Purdue University, and Dr. Falk Hante and Prof. Martin Gugat from the University of Erlangen-Nuremburg. The collaboration and/or discussion with them significantly helped me deepen understanding and increase knowledge on the topics involved in this thesis.

This thesis work also benefited from my interaction with other researchers from MIT and elsewhere. The other lab mates of mine, including Mr. Jeffrey Liu, Mr. Devendra Shelar, Mr. Mathieu Dahan, Mr. Hao-Yu Derek Chang, and Major Andrew Lee, provided valuable feedback on my thesis work. Ms. Julia Romanski (MIT ORC) provided insightful comments on the paper based on which Chapter 2 of this thesis is written. The discussion with Prof. Pravin Varaiya (Berkeley EECS) was very helpful as well.

In addition, I appreciate the support from the faculty and staff in the Interdepartmental Transportation Program at MIT. Besides the three Transportation-affiliated thesis committee members (Profs. Wilson, Balakrishnan, and Amin), Prof. Moshe Ben-Akiva, as my first-year academic advisor, also provided valuable suggestions for my coursework and research. I would also like to thank Ms. Kiley Clapper, Ms. Eunice Kim, Ms. Roberta Pizzinato, and Mr. Max Martelli for helping me through the program.

Finally, I would like to thank my parents for their support. Without that support, I could not have completed the PhD program.

The work reported in this thesis was supported in part by NSF Grant No. 1239054 CPS Frontiers (FORCES), MIT CEE Graduate Fellowship, NSF CAREER Award CNS 1453126, and the Singapore MIT Alliance for Research and Technology (SMART) Center for Future Mobility (FM). Their support is gratefully acknowledged.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis proposes a system-theoretic framework for macroscopic modeling, analysis, and design of resiliency-improving control strategies for smart highway systems. The concept of the smart highway includes controlling congested highway sections with sensor-actuator systems and integration of connected and autonomous vehicles (CAVs). We focus on a class of stochastic perturbations resulting from (i) unreliable capacity of individual highway sections (traffic incidents) and (ii) dynamic bottlenecks resulting from the integration of CAV platoons with normal traffic. In addition, we discuss the modeling of security failures in the context of smart highway systems. We propose a novel modeling framework that captures the macroscopic traffic dynamics under these perturbations, and develop new analysis and control design tools to improve the system performance in terms of throughput and travel time (queue length). While most of this thesis deals with perturbations that are non-strategic, we also consider simple models of interaction between strategic users/malicious adversaries and system operators. The technical contributions presented in this thesis are grounded in the theory of stochastic switching systems, control of continuous-time Markov processes, and game-theoretic models of queuing systems.

In this introductory chapter, we present the technological features of smart highway systems, discuss the main challenges in ensuring operational resiliency of this class of systems, and highlight our main contributions. In Section 1.1, we review the background of smart highway systems and summarize key developments in this area.

In particular, we focus on highway operations with integration of CAV platoons and dynamic, feedback control strategies for ramp metering and demand management. In Section 1.2, we review existing approaches and point out several practical challenges that need to be addressed for extensive deployment of these technologies. We argue that resolving these challenges entails development of operational tools to ensure efficiency in nominal operations, robustness against random perturbations (faults and non-strategic disturbances), and survivability under security failures (adversarial attacks on cyber/physical infrastructure). In Section 1.3, we discuss how the models and analysis/design tools developed in this thesis contribute towards addressing these challenges. Finally, in Section 1.4, we provide an outline of the subsequent chapters.

1.1 Overview of smart highway systems

To motivate our research, we recap the control systems architecture of smart highways proposed by Varaiya [98]. The author introduced automatic highway control and connected and autonomous vehicles as two key components of smart highway systems; see Fig. 1-1. The primary difference between conventional and smart highway systems is the level of information exchange. The former involves mostly static and one-way information exchange, in particular, via conventional road-side signals. On the other hand, an essential feature of the latter is extensive real-time information

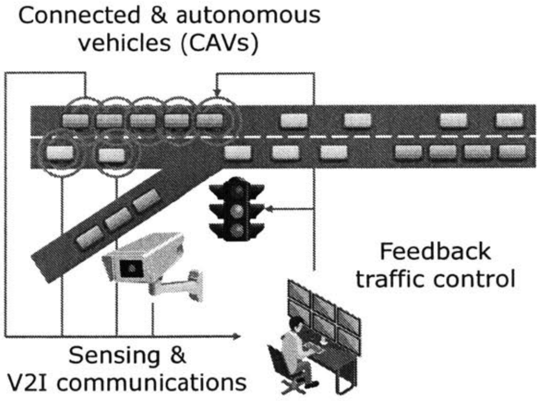

Connected & autonomous vehicles (CAVs)
Feedback traffic control
Sensing & V2I communications

Figure 1-1: A smart highway system with feedback traffic control and vehicle platooning.

exchange between the vehicles and the infrastructure (V2I) as well as between vehicles (V2V). Indeed, in the last 25 years since the article [98] was published, we have witnessed continued progresses in information, communications, and control technology for smart highways, and we are currently in the midst of another revolution in autonomous vehicle technology. Next, let us briefly discuss the relevant aspects of these technologies.

1.1.1 Automatic control of highway traffic

Broadly speaking, the control capabilities for smart highways contribute to demand management (routing) and/or capacity allocation (ramp metering). Both capabilities are intended to improve operational efficiency, including resolving congestion at bottlenecks and improving throughput. Note that the time scale of operation that we typically consider ranges from minutes to hours—the questions of how the deployment of control capabilities is related to objectives of transportation planners is beyond the scope of this thesis.

Demand management

Demand management refers to mechanisms that influence travelers' choices of routes to improve system-wide performance. Route guidance and tolling are two typical demand management mechanisms in the context of smart highways.

Route guidance is enabled by broadcasting real-time traffic information to drivers via roadside message boards, GPS-enabled navigation tools, or V2I communications. Technologically, this mechanism relies on traffic measurement and communications capabilities, which have improved considerably during the past several decades due to the increased deployment of roadside sensors (loop inductors, video cameras, etc.) and use of smart phones [58]. The intended objective is to inform travelers about traffic congestion and/or traffic incidents, and encourage travelers to take alternative routes to avoid congestion. Without this information, travelers can only learn about traffic conditions from their day-to-day experience and thus often make inefficient decisions.

A typical assumption that is made in studying the impact of route guidance is that every traveler rationally chooses the fastest route available to him/her, resulting in an outcome governed by the so-called user equilibrium or Wardrop equilibrium [102].

Another classical mechanism for demand management is congestion pricing or tolling, where tolls are charged on congested roads to incentivize travelers to choose alternative routes or times. A typical objective in the design of tolling schemes is to internalize the congestion externality, i.e. to move the user equilibrium under tolling close to the system optimum. Indeed, a system optimum traffic assignment plan can result in heterogeneous travel times on various routes; thus a system optimum can entail some travelers taking a slower route, even if a faster route is available. Such a route choice strategy cannot be justified when every traveler selfishly plans his/her route using a route guidance tool. However, under a well designed tolling scheme, some travelers may select the slower route, even if the faster route is available. Tolling has been demonstrated to be effective in terms of improving system-wide performance [90]. Although tolling involves the economic interaction between travelers and system operators and thus requires micro-economic modeling; the system-level performance metrics (congestion level, travel time) are identical to the ones used for evaluating the effectiveness of tolling schemes.

One of the main goals of this thesis is to develop operational control tools for improving performance of individual highway sections and simple (parallel-route) networks under a class of stochastic perturbations. We do not pursue the secondary question of how network-level route choice strategies would alter when these control tools are deployed. Nevertheless, we identify demand patterns (i.e. spatial distribution of traffic queues) for congested highway sections that result in effective utilization of available traffic capacity. Again, the issue of how such demand patterns can be implemented at the network level using route guidance or tolling schemes is not considered.

Capacity allocation

Capacity allocation refers to the allocation of limited road capacity to various sources of demand. For the purpose of achieving operational resiliency, we mainly focus on ramp metering as a control mechanism for capacity allocation.

Ramp metering involves restricting the rate at which on-ramps discharge traffic into the mainline of a highway. This operation came to the attention of transportation researchers as early as the 1960s [6, 73, 103]. Since vehicle speed decreases as traffic density exceeds a certain threshold (i.e. the critical density) [38], a traffic manager can restrict the inflow from an on-ramp to maintain optimal speed on the mainline of a highway. In practice, this restriction is imposed by either temporarily closing the on-ramp or regulating the on-ramp discharge rate via a traffic signal [80].

Modern control technology enables traffic managers to regulate the on-ramp discharge rates in response to real-time traffic conditions. A feedback ramp controller can be distributed or centralized. A distributed ramp controller is responsive only to the local traffic conditions. A typical distributed ramp controller is ALINEA proposed by Papageorgiou et al. [79]. This controller requires the measurement of local traffic density, and is in fact a proportional-integral (PI) controller. On the other hand, a coordinated ramp controller simultaneously regulates multiple on-ramps. Such a controller needs to be designed using model-predictive control design approach, and is capable of achieving better performance than distributed controllers. However, implementation of coordinated ramp controllers heavily depends on accurate modeling and calibration, and reliable communication between sensors and actuators. Currently, these capabilities may not be always achievable in practice.

We view ramp metering as a control mechanism that helps the traffic manager allocate limited highway capacity between the traffic on the mainline and the traffic from the on-ramps, under a class of perturbations. When demand temporarily exceeds capacity, congestion inevitably occurs at some location(s). In the absence of ramp metering, the congestion first arises on the mainline before eventually propagating to the on-ramps. With ramp metering, on the contrary, traffic congestion can be partly shifted to the on-ramps, and thus the mainline capacity can be effectively utilized. In Chapters 4–5, we argue that a ramp control strategy can be used for assigning relative priorities between mainline and on-ramp traffic to limit the effects of stochastic perturbations on system-wide performance.

1.1.2 Connected and autonomous vehicles

Safe and efficient integration of connected and autonomous vehicles (CAVs) is another key aspect of smart highway systems. CAV technology continues to improve with the ultimate objective of eliminating the inefficiency and unreliability of human driving [98]. With regard to efficiency, human drivers are prone to irregular driving behavior, including unnecessary acceleration and deceleration, which introduce perturbations in traffic flow as well as increase fuel consumption [52]. With regard to safety, 45%–75% of traffic accidents are due to human error [104]. In addition, traffic rule violations by inattentive or aggressive drivers are also a significant contributor to traffic accidents [109].

Next, we briefly comment on technological features of autonomous vehicles and vehicle platooning.

Autonomous vehicles (AV)

Self-driving involves two main tasks, i.e. sensing/perceiving the environment and controlling the vehicle's movement. Take as an example an autonomous vehicle built by a group of researchers from MIT [66]. The control system of this vehicle has a classical architecture, which consists of three components: (i) sensing, (ii) perception, and (iii) planning and control (P&C).

A typical sensing component installed on an AV includes a range of sensors such as camera, radar, and lidar, which collect information from the surroundings. The authors of [66] also considered the Global Positioning System (GPS) connectivity for vehicle tracking and localization. In addition to the onboard sensors, an AV can also collect information from other vehicles and from the infrastructure via V2V/V2I communications. Based on the information collected by the sensing component, the perception component maps the dynamically evolving environment, including the lanes, other vehicles, obstacles, etc. Then, the P&C component decides the path and maneuvers (car-following, lane changing, overtaking, etc.), and commands the actuators (steer, throttle, break, etc.) to realize the planned movement.

Vehicle platooning

Vehicle platooning refers to the cooperative movement of a group of autonomous vehicles, or a platoon. Current platooning technology typically relies on a human-driven vehicle acting as the platoon leader, while the following vehicles are driven by computers [2]. In the future, this technology is expected to evolve to a stage where the lead vehicle can also be computer-driven [98]. The movement of the following vehicles is governed by the adaptive cruise control (ACC) system. In a platoon, the ACC system of a following vehicle collects information of the lead vehicle via radar, lidar, and wireless communications, and regulates the movement of the vehicle [2]. Currently, a following vehicle is driven by ACC only if it is part of an active platoon; otherwise, it is controlled by a human driver.

Although the concept of automatically regulating a platoon of vehicles already existed in the 1960s [67], it is only over the last few years that extensive experimental studies in real-world traffic conditions have been conducted [2, 76, 96]. Currently, experiments of platooning are conducted mainly on highways, where the environment is simpler than urban streets. With the rapid advancements in vehicle platooning technology [86], it seems plausible that semi-automated highway systems will be practically viable soon [41].

In the context of traffic operations, platooning of vehicles can be considered as an effective way to improve traffic throughput [15, 69, 93] and reduce environmental externalities [2, 12, 96]. Platooning can improve throughput by reducing the inter-vehicle spacing. Since the following vehicles in a platoon are driven by computers, they can react faster to the movement of the lead vehicle, and thus travel with a smaller distance from the front vehicle. Consequently, vehicles on a highway can travel at a fast speed even with shorter separation. In addition, the reduced spacing leads to a reduction in the air resistance experienced by the following vehicles, which improves fuel efficiency.

1.2 System resiliency: previous work and challenges

Resiliency is a major concern for design and operation of smart highways. In this thesis, we define resiliency as a "super-attribute" comprising three aspects (i) efficiency under nominal operational conditions, (ii) robustness against random perturbations, and (iii) survivability under security failures; see Fig. 1-2. We will focus on the first two aspects in the main body of this thesis (Chapters 3–5), and briefly discuss on the third in the ongoing work (Section 6.2). Next, we review previous work on the operational resiliency of highway systems, and summarize the main challenges that define the focus of our research.

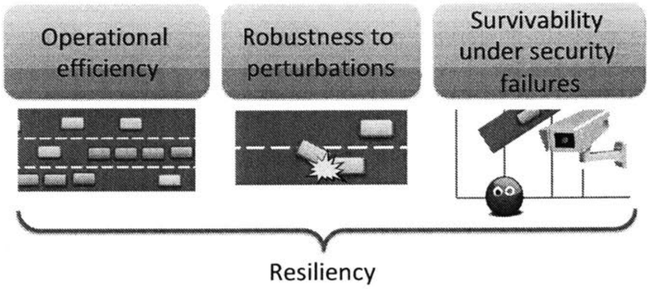

Operational efficiency
Robustness to perturbations
Survivability under security failures
Resiliency

Figure 1-2: Three aspects of resiliency.

1.2.1 Efficient operation under nominal conditions

Currently, the main challenge in development of a system-level (macroscopic) model-based control design framework for improving the operational resiliency of smart highways is the lack of realistic and tractable models that capture specific features of CAVs and their interaction with human-driven vehicles. In our research, we consider a class of macroscopic traffic flow models as they are well-suited for system-level design of control strategies with performance guarantees. Indeed, vehicle-level (microscopic) models are also relevant; however, they are outside the scope of this thesis.

The traditional macroscopic models for highway traffic flow that have been used by the transportation community can be classified into the following three classes:

1. *Stochastic queuing models* consider vehicles as "customers" and a highway section as a "server". The inter-arrival times as well as the service times of vehicles are random variables. This class of models focuses on the delay due to randomness in the movement of individual vehicles [77]. The field of queuing theory is very well developed (see standard textbooks such as [33]); typical stochastic queuing models are tractable and enable analytical characterization of the steady-state distribution of the traffic queue. This class of models are usually applicable to study performance of urban intersections [72, 84] as well as highway sections [8].

2. *Fluid queuing models* consider the aggregate flow of vehicles rather than tracking individual vehicles. This class of models mainly considers the delay due to fluctuations in demand and/or capacity, rather than randomness of the movement of individual vehicles. Fluid queuing models are well tractable, even if the demand/capacity fluctuations are stochastic. However, they do not account for the spatial distribution of traffic (unless a network of fluid queuing links is considered). This class of models is also applicable to study performance of urban intersections [100] and highway bottlenecks [77].

3. *Partial differential equation (PDE)-based models* consider traffic flow as a dynamic fluid with a particular flow-density relation, called the *fundamental diagram.* The basic idea of the fundamental diagram is that vehicles move slower when traffic density increases [38]. In the 1950s, Lighthill and Whitham [68] and Richards [85] developed a fluid model for highway traffic flows, called the LWR model. This model captures important features of highway traffic including the flow-density relation and the propagation of congestion waves. However, since this model is governed by a system of partial differential equations, its cal-

ibration and use for control design is more complicated than the previous two classes of models. A significant development in this direction is the introduction of Daganzo's cell transmission model (CTM, [24]), which is essentially a first-order discrete Godunov approximation of the LWR model [36, 37]. Numerous researchers have adopted the CTM for performance evaluation and control design for highway traffic [22, 37, 40, 70, 99].

The control problem in the non-autonomous setting has been considered by transportation researchers for decades, both theoretically and experimentally. In a 1965 paper, Athol [6] considered a simple control problem for a highway section with several on-ramps. The author [6] qualitatively studied both distributed and coordinated ramp control, and identified the importance of models for ramp control design. More recently, relying on the advancements offered by sensing and control technologies, Papageorgiou et al. [79] proposed a practical ramp controller, called ALINEA, which provably stabilizes the traffic flow at isolated highway merges. Gomes et al. [37] further synthesized the design of coordinated metering of multiple on-ramps to improve highway throughput.

The existing literature focuses on maintaining free flow on the highway mainline. Consequently, in high demand situations, ramp metering can lead to queuing at on-ramps, which can be highly costly if the resulting congestion propagates to the upstream highways/arterials. This problem was pointed out by May [73] as early as 1964. Unfortunately, to the best of our knowledge, limited results are available to systematically address the issue of on-ramp queues. In this thesis, we build on the network version of the CTM proposed by Daganzo [25] to explicitly consider the impact of on-ramp queuing in ramp control design for highway systems.

It is not yet clear how well the conventional models apply to (or rather how they should be modified to) account for traffic flow with CAVs. A very recent survey [15] summarizes the state-of-the-art of this area and provides a comprehensive literature review. According to Calvert et al. [16], some simulation-based studies and theoretical analyses have been proposed; however, very limited justification is available regarding their relevance to empirical evidence from field experiments. Talebpour and Mahmassani [94] conducted a simulation-based analysis of the influence of CAVs on highway performance (travel time and throughput). Lioris et al. [69] studied the throughput of intersections with flows of fully automated vehicles traveling in platoons. However, a systematic modeling approach that considers the interaction between CAVs and human-driven vehicles is still an outstanding issue and serves as the first motivating challenge behind our research.

1.2.2 Robustness against random perturbations

Notably, most of the ramp metering approaches developed in the literature assume nominal operating conditions, i.e. they do not provide congestion mitigation guarantees under perturbations due to random capacity variations, incidents, or stochastic traffic arrivals. However, most highway systems are routinely subject to capacity perturbations, for example, crashes, road blockages, and other capacity-reducing incidents [50, 55, 61, 89]. In practice, these events can introduce significant congestion in highways [62, 88], and control strategies designed for nominal operations are typically ineffective in resolving such congestion. This serves as the second challenge for our research.

Previous work on stochastic models for incidents has focused on two broad classes of problems. The first class is prediction and detection of incidents. One of the first contributions in this direction is by Willsky et al. [105]. The authors proposed a macroscopic approach to the detection of highway accidents using sensor data. More recently, Lee et al. [65] identified several precursors that can help predict the likelihood of accidents. The second class is estimation of the impact of accidents using historical data, and design of control schemes for capacity-reducing incidents. Khattak et al. [54] developed a stochastic queuing model that estimates the consequences of accidents. Recent work by Miller and Gupta [75] used a classification model to assess the severity and induced delay due to reported accidents. In [60], Kurzhanskiy explored practical control schemes for several accident scenarios in California highways. Yet, to the best of our knowledge, the available literature does not present a systematic approach that incorporates the randomness of accidents into highway traffic dynamics.

In terms of modeling the impact of perturbations, there has been previous work on stochastic extensions of the CTM. Sumalee et al. [92] proposed a CTM with parameters subject to random noise; however, their model is largely motivated by the intrinsic randomness in demand and in the flow-density relation, but does not explicitly model capacity-reducing events. Zhong et al. [111] consider a closely related model in which traffic flow dynamics randomly switch between free-flow and congested modes. Using that model, the authors of [111] studied an optimal control problem over a finite time horizon. However, these models do not capture the dynamic propagation of incident-induced congestion (spillback), which is a critical mechanism affecting the traffic dynamics in highways.

In terms of control design under uncertain link or node capacities, previous work either assumes a static (but uncertain) capacity model, or considers time-varying capacities [5, 20, 82]. In the former approach, the actual capacity is assumed to lie in a known set [20], or is realized according to a given probability distribution [35]. Such models are useful for evaluating the system's performance against worst-case disturbances. The latter approach is motivated by situations where the capacity is inherently dynamic. These models can enable more accurate assessment of system performance in comparison to static models. In contrast to the above two classes, we consider the control design in situations where the capacity can be modeled as a Markovian process [8, 47, 82].

1.2.3 Survivability in the face of security failures

The growing deployment of cyber-physical components, including sensors (induction loops, video camera), actuators (adaptive traffic signals, navigation tools such as Google and WAZE), and vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communications in road traffic networks raises concerns for security failures. A cyber-physical component can be compromised via either illegal attack/intrusion or injection of malicious data, which leads to physical impacts on highway performance/safety.

It has been practically demonstrated by various research groups that cyber-physical components in smart highway systems are vulnerable to security failures:

1. On the vehicle side, Kosher et al. [56] demonstrated that a broad set of safety-critical systems in a modern car can be compromised via hacking the electronic control units. Spaar [91] identified the security vulnerability of the BMW ConnectedDrive system, through which a BMW car can be remotely unlocked. DARPA researchers also managed to hack into a Chevrolet car via its OnStar system [14].

2. On the infrastructure side, Ghena et al. [34] demonstrated a remote intrusion into a traffic signal controller via the communications network connecting the traffic signals. Petit and Shladover [83] also pointed out vulnerabilities in the V2V/V2I communications, which could compromise either individual vehicles, or a highway system collectively. In fact, hacking of traffic controllers has occurred in reality and caused significant economic loss [11].

Because of the large number of cyber-physical components involved in smart highway systems, it is neither technologically nor economically feasible to protect every component from security failures. Therefore, we should design the system such that

1. the most critical components are protected or inspected, and

2. the system architecture does not allow a local security failure to extensively propagate through the system.

Besides ensuring adequate level of investment and oversight in deployment of security solutions, a major challenge in achieving resiliency to security attacks is the lack of models that allow control engineers and traffic system operators to assess the negative impact of plausible security failures on system performance. This serves as the third motivating challenge behind our research. Prior work in this direction includes the paper by Laszka et al. [64], who consider a dynamic traffic flow network model for evaluating the impact (in terms of travel delay) of compromised intersections. The authors of [64] also provided a rather complete review of previous work following this approach. Como et al. [20, 21] considered the problem of evaluating the resiliency of the traffic transmission over a single-origin-single-destination network against worst-case physical disturbances (disruptions). Instead of travel delay, they considered throughput loss as the metric for link disruption. In another approach [4, 87], game-theoretic models have been used to study the strategic interaction between malicious or fraudulent users and system operators. Importantly, Babaioff et al. [7] considered the equilibria of a class of congestion games involving strategic travelers. However, in general, game-theoretic models that specifically consider strategic attack models for transportation systems (and in particular smart highways) are still in their infancy.

1.3 Our contributions

In this thesis, we develop a system-theoretic approach for modeling, analysis, and control of smart highways subject to perturbations resulting from capacity-reducing events (random incidents) and/or heterogeneous traffic classes (CAVs vs. normal traffic).

Our approach enables analysis of traffic flow dynamics under a broad class of stochastic demand/supply fluctuations, and can be used for designing control strategies that limit the impact of these fluctuations on system performance. We model such stochastic fluctuations as a finite-state Markov process and study their impact on the dynamical evolution of the traffic state (vehicle densities and on-ramp queues). A key finding is that stochastic fluctuations can lead to traffic bottlenecks and congestion that do not exist with the nominal or average settings; therefore, highway operations designed for the nominal/average setting can be inefficient or ineffective in practice. We use this model to design capacity allocation schemes that guarantee the stability of traffic queues and improve network-wide throughput. These schemes suggest new and somewhat surprising ways to prioritize and route traffic flows on real-world highway systems, and motivate several applications including management of traffic incidents in highways and lane control of connected vehicles in mixed traffic conditions. Furthermore, we consider a specific security failure that is relevant given the existing vulnerabilities in technologies supporting V2I-based highway operations.

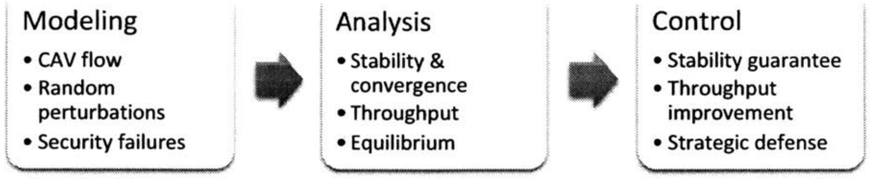


Figure 1-3: Summary of the contributions.

We now highlight our specific contributions in modeling, analysis, and control, as summarized by Fig. 1-3. The main body of this thesis focuses on operational efficiency and robustness against random perturbations; modeling and analysis of security failures as well as strategic defense are presented in Chapter 6 as ongoing work.

1.3.1 Modeling

Fig. 1-4 illustrates a hierarchical control system for smart highways proposed in [98]. In this thesis, we focus on the macroscopic layers. That is, we are concerned with the aggregate behavior of flows of vehicles, rather than the movement of individual vehicles. This modeling approach enables us to apply well-known tools in the theory of Markov processes (especially queuing theory) and control theory to analysis/design in the macroscopic layers, and to derive useful insights for highway operations.

Our approach builds on two classical macroscopic traffic flow models, viz. fluid queuing model (FQM) and cell transmission model (CTM), as indicated in Fig. 1-4. Under stochastic demand and/or stochastic capacity perturbations—which we model as a Markov process—these models become the piecewise-deterministic queuing (PDQ, see [46]) model and stochastic switching cell transmission model (SS-CTM, see [45]), respectively. In fact, both these models belong to the more general class of piecewise-determinstic Markov processes (PDMP, see [9, 29]).

In general, one can view the PDQ model as an abstraction of the SS-CTM. The PDQ model captures the queuing due to demand/capacity perturbations, but accounts for neither the spatial distribution of traffic queues nor the relation between

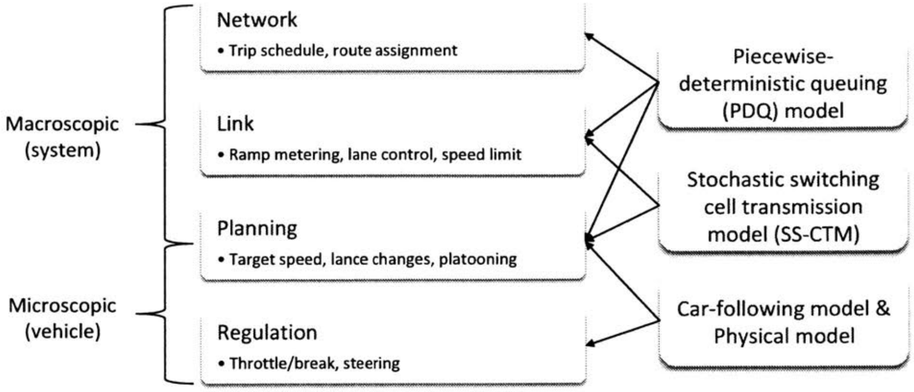

Macroscopic (system)
Microscopic (vehicle)
Network
• Trip schedule, route assignment
Link
• Ramp metering, lane control, speed limit
Planning
• Target speed, lance changes, platooning
Regulation
• Throttle/break, steering
Piecewise-deterministic queuing (PDQ) model
Stochastic switching cell transmission model (SS-CTM)
Car-following model & Physical model

Figure 1-4: Control layers and models involved in smart highway systems. This thesis focuses on the macroscopic layers/models.

flow rate and queue length (or traffic density). On the other hand, the SS-CTM not only tracks the total queuing delay, but also where the queue is located (on which mainline section or at which on-ramp); furthermore, the SS-CTM explicitly accounts for the flow-density relation of highway traffic. However, the PDQ model is more tractable than the SS-CTM.

Next, we summarize the key features of these models.

Piecewise-deterministic queuing (PDQ) model

A key feature of the PDQ model for highway sections is that the traffic queue at a bottleneck is always discharged at the capacity (which follows from the fluid queuing model [77]). In addition, demand and/or capacity fluctuations vary stochastically in this class of models. We show that the PDQ model can be used to capture the effects of stochastic arrivals of vehicle platoons at a highway bottleneck, as well as their macroscopic interaction with normal traffic. Furthermore, we also demonstrate that the PDQ model can serve as a representative model for the design of stabilizing routing strategies under stochastically fluctuating route capacities (which result from recurring incidents). These applications demonstrate that, for network-level analysis, PDQ models can serve as reasonable abstractions for highway systems facing a broad class of Markovian perturbations.

PDQ for platooning operations: We consider a highway section with both conventional vehicles and platoons of CAVs. While previous work provides a good foundation to study platooning in specific scenarios [30, 51], it does not naturally lead to a tractable approach for designing efficient operational strategies under mixed traffic conditions. We focus on the macroscopic interaction between platoons of connected vehicles and ordinary vehicles, and show that the analytical tractability of PDQ model leads to practically relevant insights on the design of platoon operations.

The PDQ model for platooning captures the following features of CAV platoons. First, vehicle platoons can act as temporary bottlenecks for other vehicles. We use a two-class fluid queuing model to capture the sharing of highway capacity between vehicle platoons and the background (normal) traffic. Second, the headways between platoons and the lengths of platoons are subject to random variations. We use a Markov process to capture this randomness. Third, vehicles within a platoon have smaller spacing compared to ordinary vehicles. We scale down the queuing effect due to vehicle platoons according to a pre-defined inter-vehicle spacing ratio for the two traffic classes. Note, however, that the model does not account for (i) the impact of speed difference between platoons and background traffic, (ii) the formation/split of platoons, (iii) the microscopic (vehicle-level) interaction between platoons and background traffic. Inclusion of these features is part of our ongoing work, see Section 6.2.1.

PDQ for incident management:

The PDQ model can also be used to represent a network of parallel routes facing stochastic capacity perturbations. This model conveniently captures the dynamic effects of capacity-reducing traffic incidents which are known to have random occurrence/clearance rates. The parallel-route PDQ network model allows us to study the following questions:

1. How to model the traffic delay induced by random incidents?

2. How do incident characteristics, including occurrence rate, duration, and capacity reduction, affect the induced delay?

3. How to route traffic in response to incidents?

Since we focus on the behavior of aggregate traffic flows, fluid queueing models are better suited to our objectives than conventional queueing models (e.g. $M/M/1$) [82, 77]. Single server fluid queueing systems with stochastically switching saturation rates have been studied previously; see [5, 18, 57]. This line of work focuses on the analysis of the stationary distribution of queue length under a fixed inflow or an open-loop control policy. Some results are also available on feedback-controlled fluid queueing systems with stochastic capacities [82, 108]. However, to the best of our knowledge, stability of parallel-link fluid queueing systems with uncertain capacities has not previously been considered.

In our parallel-route PDQ network model, the saturation rates of the links switch between a finite set of values, or modes, according to a Markov chain, while the evolution of queue lengths between mode switches is deterministic. An advantage of this model is that it can be easily calibrated using commonly available traffic data [48]. Furthermore, since the capacity and the queue lengths can be obtained using modern sensing technologies, this model can be used to design capacity-sensitive control policies.

Stochastic switching cell transmission model (SS-CTM)

The SS-CTM combines a stochastic switching capacity model with the classical CTM. This model has a continuous state that describes the traffic state (traffic density on the mainline and queues at the on-ramps), and a discrete state that captures change in dynamics due to accidents. For given parameters (calibrated offline) and inputs, the model is capable of simulating the evolution of traffic under a range of recurrent capacity-reducing events such as incidents and moving bottlenecks (slow vehicles or CAV platoons) through the highway. This model can be used by traffic controllers to evaluate the impact of randomly occurring events that affect highway capacity. The model is also useful for designing control strategies that account for the nature of incident occurrence/clearance events. While our main focus is on modeling incidents, our modeling approach can be extended to other capacity-reducing events, e.g. reduction in capacity due to blockage or road surface damage.

The SS-CTM captures capacity-reducing events as switches of highway capacity between a finite set of values (*modes*); the switches are governed by a continuous-time finite-state Markov chain. The mode transitions can represent abrupt changes in traffic dynamics including: (i) occurrence of primary incidents [1, 50], (ii) occurrence of secondary incidents [54, 81], and (iii) clearance of incidents [47, 89]. In a given mode, the traffic density in each section (the continuous state of the model) evolves according to the CTM. The mode transitions essentially govern the build-up and release of traffic queues in the system.

We note that true capacity fluctuation may be more complicated than implied by the finite-state Markov model [44, 48]. However, this model is adequate to study the build-up and release of traffic queues due to stochastic capacity, and also enables a tractable analysis of long-time properties of the traffic queues. In a related work [48] (not included in this thesis), we showed that calibration of this model is simple, and that it satisfactorily captures the stochastic variation of capacity in practical situations; see [8, 53, 111] for related models.

1.3.2 Analysis

In this thesis, we develop tools for analyzing performance of smart highway systems, based on either the PDQ model or the SS-CTM. We consider travel time and throughput the key performance metrics. Specifically, our analysis focuses on the following practical questions:

1. Under what conditions do the traffic queues induced by demand/capacity perturbations remain bounded?

2. What is the maximum rate at which a smart highway system can discharge traffic without inducing unbounded queues?

The first question is relevant for estimating the average travel time, since travel delay is directly related to traffic queue length. The second question is relevant for throughput analysis. These two metrics are standard for nominal performance evaluation of highway systems [61, 99], and useful for estimating the efficiency loss due to random perturbations [8, 77] or security failures [64].

In the context of the PDQ model or the SS-CTM, boundedness of traffic queues means that the stochastic process governing the evolution of traffic (in particular, build-up and release of traffic queues in either model) is stable. We are interested in deriving intuitive and verifiable stability criteria for both models. Specifically, we consider boundedness of either the time-average moment generating function (MGF) or the time-average expected value of the total number of vehicles in our analysis. In the PDQ, this means that the queues on every link are bounded. In the SS-CTM, this means that the queue at every on-ramp is bounded. In addition, for the more tractable PDQ model, we also consider convergence, a notion related to stability, which means that the traffic queues converges to a unique steady-state probability distribution. Our notion of stability and convergence is similar to that considered by Dai and Meyn [28].

The main results of our analysis include necessary and sufficient conditions for stability of both models, viz. the PDQ and the SS-CTM. To establish these results, we build on the theory of stability of continuous-time Markov processes [9, 33, 74]. Note, however, that the application of standard stability results to our setting is not straightforward due to non-linearity of the PDQ/CTM dynamics. In this thesis, we exploit the properties of the mode transition process and the PDQ/CTM dynamics to develop a computationally tractable approach to characterizing the set of stabilizing inflow vectors—specifically, an over- and an under-approximation of this set. We now introduce the main features of our stability conditions.

For the SS-CTM model, the necessary condition essentially state that the on-ramp queues are stable only if for each highway section, the incoming traffic flow does not exceed the time-average of the "spillback-adjusted" capacity. The notion of spillback-adjusted capacity essentially captures the effect of downstream congestion (i.e. spillback) resulting from capacity fluctuations on the traffic discharging ability of each cell. To the best of our knowledge, this notion has not been reported previously in the literature. The computation of spillback-adjusted capacity builds on the construction of an invariant set that is also globally attracting. Using this construction, we show that the capacity fluctuation can result in an unbounded traffic queue even when the inflow in each cell is less than the average capacity; this property essentially results from the effect of traffic spillback. The necessary condition also provides a way to identify an over-approximation of the set of stabilizing inflow vectors (i.e. the inflows that satisfy the necessary condition).

To establish sufficient conditions for stability, we consider a well-known approach formalized by Meyn and Tweedie [74]. Application of this approach to our model is challenging, since it requires verification of the Foster-Lyapunov drift condition, i.e. that a Lyapunov function for the stochastic process is decreasing in expectation everywhere over the state space. Computationally, verifying this condition involves checking a set of non-linear inequalities everywhere over the invariant set (which we explicitly construct for the SS-CTM). To resolve this issue, we construct a switched Lyapunov function that captures both the queuing process as well as the demand/capacity perturbations as "mode transitions". We also utilize properties of the SS-CTM dynamics to show that the drift condition holds everywhere over the invariant set if it holds over the finite set of vertices of the invariant sets that we construct. Thus, to establish stability, we only need to verify the drift condition at finitely many states. Consequently, standard computational tools [71] can be used to check whether or not the PDQ model/SS-CTM is stable.

Overall, these results enable a systematic analysis of the congestion induced by stochastic demand/capacity. We also discuss the tightness of our results, i.e., the gap between the necessary and the sufficient condition. In addition, we present illustrative examples to study the impact of capacity fluctuation (including its frequency, intensity, and spatial correlation) on throughput.

For the PDQ model, similar necessary/sufficient conditions for stability can be derived, using similar tools. Furthermore, we can argue for the convergence of the probability distribution of the traffic queues. That is, in addition to being bounded, the queues on a single link or on a network of parallel links converge to a unique steady-state distribution (in the sense of total variation), or an invariant probability measure. The sufficient condition for convergence has two requirements: (i) the model is stable, i.e. the traffic queues are bounded on average; (ii) all queues eventually vanish in a "nominal" mode. Condition (i) essentially ensures the existence of an invariant probability measure, and Condition (ii) ensures the uniqueness of the invariant measure. This condition also provides an exponential convergence rate towards the invariant measure. The sufficient conditions for convergence can be verified in a more straightforward manner in the case when the PDQ system has two modes. Furthermore, under a mode-responsive control policy the sufficient condition and the necessary coincide for a two-mode PDQ.

1.3.3 Control design

Varaiya [98] proposed a four-layer hierarchy for the control system of smart highways, including network, link, planning, and regulation. We adapt Varaiya's layering for our resilient control problem and focus on the system-level layers, i.e. network and link layers. In these layers, limited results on resiliency-improving control, either theoretical or experimental, are available, and a systematic framework that supports control design is lacking. We focus on the control design problem for the network layer and the link layer of smart highway systems under a class of Markovian demand/capacity perturbations. Control design under integration of CAVs is part of our ongoing work and beyond the scope of this thesis.

The network layer refers to the assignment of given traffic demand over a set of alternative routes. The network layer involves network-wide, long-term (order of hours/days) decisions, while the link layer involves local, short-term (order of minutes) decisions. At the network layer, routing policies (network layer) are designed based on an abstraction of metered highway segments. We use the more tractable PDQ model in this layer.

The link layer refers to the operation of a single highway link, possibly with on-ramps and off-ramps. At the link layer, ramp metering policies (segment layer) are designed in a distributed manner, independent of metering policies for other links, and independent of network-wide routing. We use the more detailed SS-CTM in this layer. This thesis focuses on the analysis in each individual layer. The synthesis of the two layers is part of our ongoing work.

Specifically, we focus on:

1. Network layer: routing over a network of parallel highways facing time-varying capacity perturbations.

2. Link layer: demand management and ramp control for a single highway with multiple on-ramps and off-ramps and facing spatially and temporally evolving capacity perturbations.

In this thesis, we deal with control design under stochastic perturbations in each layer separately. The synthesis of multiple layers with a consistency guarantee is indeed an important issue and is part of our ongoing work. Nevertheless, since we use analogous models and a common design approach over both layers, we expect the synthesis to be achievable.

Network layer

In the network layer, we consider a PDQ model with parallel links (see Fig. 1-5) that accounts for stochastically varying capacities of individual highway links, and investigate its stability under a class of feedback control policies. Since we focus on the behavior of aggregate traffic flows, fluid queueing models are better suited to our objectives than stochastic queueing models (e.g. $M/M/1$) [77, 82]. In addition, for routing policy design, the PDQ model is more tractable and insightful than the more sophisticated CTM.

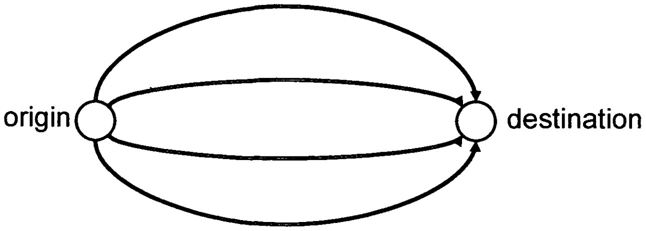


Figure 1-5: A network of parallel PDQ links.

We mainly study the stability and convergence of the PDQ network under a class of feedback routing policies. Our results are based on two assumptions: (i) the mode transition process is ergodic, and (ii) the feedback control policy is bounded and continuous in the queue lengths, and also satisfies a monotonicity condition to ensure that more traffic is routed through links with smaller queues.

Under these assumptions, and based on the analysis results, we derive necessary and sufficient conditions for stability as well as convergence. The necessary condition is that, for every link, a suitably defined lower bound on the time-average inflow does not exceed the corresponding link's time-average saturation rate. The sufficiency result requires a lower bound on the discharge rate of the system in individual modes verify a bilinear matrix inequality (BMI). The sufficient conditions for stability can be verified in a more straightforward manner in the case when the PDQ system has two modes. Furthermore, under a mode-responsive control policy the necessary and sufficient conditions coincide for a two-mode PDQ.

Link layer

In the link layer, we consider the improvement of the throughput of a highway segment (see Fig. 1-6) via joint demand management and capacity allocation. Based on previous results in the nominal setting [23, 37], we study stabilization of on-ramp queues and improvement of highway throughput under stochastic capacity perturbations, using the SS-CTM.

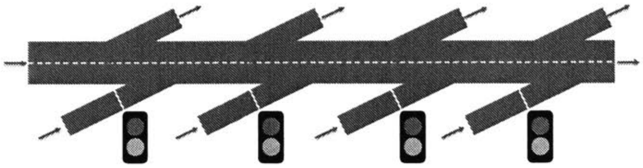


Figure 1-6: A highway with multiple on-ramps and off-ramps.

The decision variables (control inputs) include (i) the amount of demand that is accepted at each entrance, called the *inflow*, and (ii) the priority of each on-ramp with respect to the mainline, called the *priority rule*. For a highway with a given demand pattern, we are interested in maximizing the throughput (in terms of vehicle-miles traveled), while keeping the on-ramp queues bounded; i.e.

$$\begin{aligned} \max \quad & \text{throughput} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(P0)} \\ s.t. \quad & \text{every on-ramp queue is bounded on average.} \\ & \text{constraints on control input.} \end{aligned}$$

We develop a systematic approach to designing the optimal operations. The main contributions include:

1. An easily checkable sufficient condition for boundedness of the on-ramp queues, which enables us to simplify the boundedness constraint in (P0). The stability condition that we derive is bilinear or linear (depending on the complexity of the stability condition) in the decision variables.

2. A mixed integer program formulation of the throughput-maximizing problem under a broad class of stochastic capacity perturbations. Using the sufficient condition for stability, the max-throughput problem is formulated as a mixed integer bilinear/linear program (MIBLP/MILP).

3. Characterization of throughput-improving priority rules. Although analytical solution to the max-throughput problem is in general not easy, we are able to characterize the structure of a class of optimal traffic control configurations. We find that, to improve throughput, an on-ramp should be prioritized if it has a smaller capacity-demand margin than the mainline.

1.4 Thesis outline

The rest of this thesis is organized as follows. Chapter 2 introduces a piecewise-deterministic queueing (PDQ) model to study the stability of traffic queues in parallel-link transportation systems facing stochastic capacity fluctuations. In Chapter 3, we extend the PDQ model to study the macroscopic interaction between randomly arriving vehicle platoons and the background traffic at highway bottlenecks. Chapter 4 introduces the SS-CTM for highway traffic dynamics under stochastic capacity-reducing incidents, and provides insights for highway incident management by analyzing long-time (stability) properties of the proposed model. In Chapter 5, we consider highway control under the influence of stochastic capacity perturbations, based on the analysis presented in Chapter 4. In Chapter 6, we conclude the thesis by summarizing the contributions and introducing several future directions. Importantly, we present preliminary results on modeling strategic misbehavior in V2I-based highway operations, including a novel signaling game formulation and some practical insights. We also introduce our ongoing work on strategic response to adversarial cyber-physical attacks on smart highway systems.

Chapter 2

Incident-Aware Routing over Parallel Highways

Capacity fluctuations in transportation systems can cause significant efficiency losses to the system operators [62]. In practice, these fluctuations can be frequent and also hard to predict [48, 82]. Thus, traffic control strategies that assume fixed (or nominal) link capacities may fail to limit the inefficiencies resulting from capacity fluctuations, especially when their intensity and/or frequency is non-negligible.

In this chapter, we introduce a piecewise-deterministic queueing (PDQ) model to study the stability of traffic queues in parallel-link transportation systems facing stochastic capacity fluctuations. The saturation rate (capacity) of the PDQ model switches between a finite set of modes according to a Markov chain, and link inflows are controlled by a state-feedback policy. A PDQ system is stable only if a lower bound on the time-average link inflows does not exceed the corresponding time-average saturation rate. Furthermore, a PDQ system is stable if the following two conditions hold: the nominal mode's saturation rate is high enough that all queues vanish in this mode, and a bilinear matrix inequality (BMI) involving an underestimate of the discharge rates of the PDQ in individual modes is feasible.

In Section 2.1, we introduce the parallel-link fluid queueing model and the class of routing policies that we consider. Our analysis involves two assumptions: (i) the mode transition process is ergodic, and (ii) the feedback control policy is bounded and continuous in the queue lengths, and also satisfies a monotonicity condition to ensure that more traffic is routed through links with smaller queues. Under these assumptions, in Section 2.2, we derive a necessary condition (Theorem 2.1) and a sufficient condition (Theorem 2.2) for stability. In Section 2.3, we further show that the sufficient conditions for stability can be verified in a more straightforward manner in the case when the PDQ system has two modes (Proposition 1). Furthermore, under a mode-responsive control policy the necessary and sufficient conditions for a two-mode PDQ coincide (Proposition 2). Finally, in Section 2.4, we illustrate some applications of our results for designing stabilizing traffic routing policies in parallel-link networks with stochastic capacity fluctuations.

2.1 Piecewise-deterministic queueing system

Consider the PDQ system in Figure 2-1 (left). A constant *demand* $A \geq 0$ of traffic arrives at the system and is allocated to n parallel servers. The *inflow* vector $F(t) = [F_1(t), \dots, F_n(t)]^T \in \mathbb{R}^n_{\geq 0}$ is such that $\sum_{k=1}^n F_k(t) = A$ for all $t \geq 0$. Traffic can be temporarily stored in queueing buffers and discharged downstream. We denote the vector of *queue lengths* by $Q(t) = [Q_1(t), \dots, Q_n(t)]^T$. Let $U(t) = [U_1(t), \dots, U_n(t)]^T$ denote the vector of stochastic *saturation rates*, where $U_k(t)$ is the maximum rate at which the k-th server can release traffic at time t.

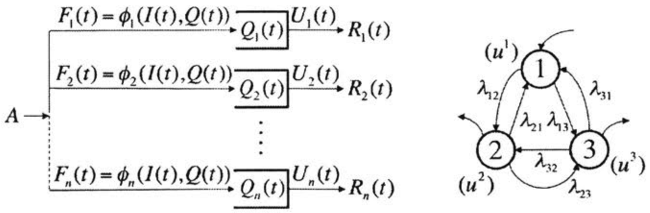


Figure 2-1: Illustration of a PDQ system with n parallel servers (left) and the mode transition process (right).

For the k-th server, if $Q_k(t) = 0$ and $F_k(t) \leq U_k(t)$, the *discharge rate* $R_k(t)$, i.e. the rate at which traffic departs from the system through the k-th server, is given by $R_k(t) = F_k(t)$; otherwise $R_k(t) = U_k(t)$. We assume infinite buffer sizes; i.e. $Q(t)$ can take any value in the set $\mathcal{Q} := \mathbb{R}^n_{\geq 0}$. This assumption enables us to account for all traffic arriving at the system and not just the traffic that is ultimately discharged by the system.

2.1.1 Markovian capacity model

In our model, the saturation rates of the n servers stochastically switch between a finite set of values. To model this switching process, we introduce the set of ***modes*** $\mathcal{I}$ of the PDQ system and let $m = |\mathcal{I}|$. We denote the mode of the PDQ system at time t by $I(t)$. Each mode $i \in \mathcal{I}$ is associated with a fixed saturation rate, denoted by $u^i = [u^i_1, \dots, u^i_n]^T$, which is distinct for each mode. The evolution of $I(t)$ is governed by a finite-state Markov process with state space $\mathcal{I}$ and constant transition rates $\{\lambda_{ij}; i, j \in \mathcal{I}\}$. We assume that $\lambda_{ii} = 0$ for all $i \in \mathcal{I}$. Note that this is without loss of generality, since self-transitions do not change the saturation rate; thus, including them will not affect the PDQ dynamics. Let

$$\nu_i := \sum_{j \in \mathcal{I}} \lambda_{ij}, \tag{2.1}$$

which is the rate at which the system leaves mode i. Given a fixed initial mode $I_0 \in \mathcal{I}$ at $t = T_0 := 0$, let $\{T_z; z = 1, 2, \dots\}$ be the *epochs* at which the mode transitions occur. Let I_{z-1} be the mode during $[T_{z-1}, T_z)$ and $S_z := T_z - T_{z-1}$. Then, S_z follows an exponential distribution with the cumulative distribution function (CDF):

$$\mathsf{F}_{S_z}(s) = 1 - e^{-\nu_{I_{z-1}} s}, \quad z = 1, 2 \dots \tag{2.2}$$

One can capture the transition rates in the $m \times m$ matrix:

$$\Lambda := \begin{bmatrix} -\nu_1 & \lambda_{12} & \dots & \lambda_{1m} \\ \lambda_{21} & -\nu_2 & \dots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \dots & -\nu_m \end{bmatrix}. \tag{2.3}$$

We assume the following about the mode transition process:

Assumption 1. *The Markov process $\{I(t); t \geq 0\}$ is* ergodic.

This assumption ensures that the process $\{I(t); t \geq 0\}$ converges to a unique steady-state distribution, i.e. a row vector $\mathsf{p} = [\mathsf{p}_1, \dots, \mathsf{p}_m]$ satisfying the following:

$$\mathsf{p}\Lambda = 0, \ |\mathsf{p}| = 1, \ \mathsf{p} \geq 0, \tag{2.4}$$

where $|\cdot|$ is the 1-norm.

2.1.2 Stochastic queuing dynamics

We consider that the demand A is distributed across the n servers according to a state-feedback *routing policy*, which we denote as $\phi : \mathcal{I} \times \mathcal{Q} \to \mathbb{R}^n_{\geq 0}$. A routing policy is *admissible* if $|\phi(i,q)| = A$ for all $(i,q) \in \mathcal{I} \times \mathcal{Q}$. For a given routing policy ϕ, the vector of discharge rates $R(t)$ is specified by the vector-valued function $r^\phi : \mathcal{I} \times \mathcal{Q} \to \mathbb{R}^n_{\geq 0}$ with following components:

$$r^\phi_k(i,q) := \begin{cases} \phi_k(i,q), & q = 0, \ \phi_k(i,q) \leq u^i_k, \\ u^i_k, & \text{o.w.} \end{cases}$$
$$k \in \{1, \dots, n\}; \tag{2.5}$$

i.e., for each $t \geq 0$, we have $R_k(t) = r^\phi_k(I(t), Q(t)) \leq U_k(t)$.

Let us define a vector field $D^\phi : \mathcal{I} \times \mathcal{Q} \to \mathbb{R}^n$ as follows:

$$D^\phi(i,q) := \phi(i,q) - r^\phi(i,q). \tag{2.6}$$

Then, the evolution of the *hybrid state* $(I(t), Q(t))$ of the PDQ system is specified by the following dynamics:

$$I(0) = i, \ Q(0) = q, \quad (i,q) \in \mathcal{I} \times \mathcal{Q}, \tag{2.7a}$$

$$\Pr\{I(t+\delta) = j' | I(t) = j\} = \lambda_{jj'}\delta + o(\delta), \ j' \neq j, \tag{2.7b}$$

$$\frac{dQ(t)}{dt} = D^{\phi}\Big(I(t), Q(t)\Big), \tag{2.7c}$$

where δ is an infinitesimal time increment. Henceforth, we consider routing policies that satisfy the following assumption:

Assumption 2. *The routing policy $\phi(i,q) = [\phi_1(i,q), \dots, \phi_n(i,q)]^T$ is bounded and continuous in q. Furthermore, for $k \in \{1, \dots, n\}$, ϕ_k is non-increasing in q_k, and non-decreasing in q_h for $h \neq k$.*

The assumption of boundedness and continuity ensures that the Markov process $\{(I(t), Q(t)); t \geq 0\}$ is right continuous with left limits (RCLL, or *càdlàg*) [29]. Furthermore, since $Q(t)$ is not reset after mode transitions, $Q(t)$ is necessarily continuous in t. With the RCLL property, following [29, Theorem 5.5], the *infinitesimal generator $\mathcal{L}^{\phi}$* of a PDQ with an admissible routing policy ϕ satisfying Assumption 2 is given by

$$\begin{aligned}\mathcal{L}^{\phi} g(i,q) &= \big(D^{\phi}(i,q)\big)^T \nabla_q g(i,q) \\ &\quad + \sum_{j \in \mathcal{I}} \lambda_{ij}\Big(g(j,q) - g(i,q)\Big), \quad (i,q) \in \mathcal{I} \times \mathcal{Q},\end{aligned} \tag{2.8}$$

where g is any function on $\mathcal{I} \times \mathcal{Q}$ smooth in the continuous argument.

The assumption of monotonicity of controlled inflows with respect to queue lengths is practically relevant: more traffic is allocated to servers with smaller queues. In addition, this assumption ensures the existence of the following limits:

$$\varphi^i_{kh} := \lim_{q_h \to \infty} \phi_k(i, q_h \mathbf{e}_h), \quad h, k \in \{1, \dots, n\},\ i \in \mathcal{I}, \tag{2.9}$$

where $\mathbf{e}_h$ is the n-dimensional vector such that the h-th element is 1 and the others are 0. Particularly, the monotonicity of ϕ also implies that $\phi_k(i,q) \geq \varphi^i_{kk}$ for all $k \in \{1, \dots, n\}$ and all $(i,q) \in \mathcal{I} \times \mathcal{Q}$.

Many practically relevant routing policies satisfy Assumption 2. Examples include:

1. *Mode-responsive* routing policy:

$$\phi_k^{\mathrm{mod}}(i) = \sum_{j\in\mathcal{I}} \mathbf{1}_{\{j=i\}}\psi_k^j, \quad k \in \{1,\dots,n\}, \tag{2.10}$$

   where $\mathbf{1}_{\{\cdot\}}$ is the indicator function, and $\psi_k^i \geq 0$ for $k \in \{1,\dots,n\}$ and $i \in \mathcal{I}$. This policy can be viewed as simple re-direction of traffic during disruptions.

2. *Piecewise-affine* routing policy:

$$\phi_k^{\mathrm{pwa}}(q) = \min\Big\{A, \Big(\theta_k - \alpha_{kk}q_k + \sum_{h\neq k}\alpha_{kh}q_h\Big)_+\Big\}, \\ k \in \{1,\dots,n\}, \tag{2.11}$$

   where $\alpha_{kh} \geq 0$ for all $k, h \in \{1,\dots,n\}$ and $(\cdot)_+$ indicates the positive part. This policy is an example of a *queue-responsive* traffic control policy. Note that θ_k can be interpreted as the "nominal" inflow sent to each server when no queue exists throughout the system, and the linear terms $\alpha_{kh}q_h$ as adjustment to these inflows that accounts for the queue lengths.

3. *Logit* routing policy:

$$\phi_k^{\mathrm{log}}(q) = \frac{A\exp(\gamma_k - \beta_k q_k)}{\sum_{h=1}^n \exp(\gamma_h - \beta_h q_h)}, \; k \in \{1,\dots,n\}, \tag{2.12}$$

   where $\beta_k \geq 0$ for $k \in \{1,\dots,n\}$. This is a classical model of travelers' route choice. One can interpret β_k as sensitivity parameter that reflects travelers' preference to the queue length in the k-th server, and γ_k the parameter governing travelers' preference when every server has a zero queue.

Note that the computation of the limiting inflows φ_{kh}^i is rather straightforward for the above-mentioned routing policies (see Section 2.4).

Next, we introduce the notion of stability. The *transition kernel* [74] of a PDQ at time $t \geq 0$ is a map P_t from $\mathcal{I} \times \mathcal{Q}$ to the set of probability measures on $\mathcal{I} \times \mathcal{Q}$. Essentially, for an initial condition $(i, q) \in \mathcal{I} \times \mathcal{Q}$ and a measurable set $\mathcal{E} \subseteq \mathcal{I} \times \mathcal{Q}$, we have

$$P_t((i,q);\mathcal{E}) = \Pr\{(I(t),Q(t)) \in \mathcal{E} | I(0) = i, Q(0) = q\}.$$

One can also consider P_t as an operator acting on probability measures μ on $\mathcal{I} \times \mathcal{Q}$ via

$$\mu P_t(\mathcal{E}) = \int_{\mathcal{I}\times\mathcal{Q}} P_t((i,q);\mathcal{E})d\mu. \tag{2.13}$$

An *invariant probability measure* [74] of a PDQ system with routing policy ϕ is a probability measure μ_ϕ such that

$$\mu_\phi P_t = \mu_\phi, \quad \forall t \geq 0.$$

Definition 2.1 (Stability [9, 19]). *The PDQ system with routing policy ϕ is* stable *if there exists a probability measure μ_ϕ on $\mathcal{I} \times \mathcal{Q}$ such that, for each initial condition $(i,q) \in \mathcal{I} \times \mathcal{Q}$,*

$$\lim_{t\to\infty} \| P_t((i,q);\cdot) - \mu_\phi(\cdot) \|_{\mathrm{TV}} = 0, \ \forall (i,q) \in \mathcal{I} \times \mathcal{Q}, \tag{2.14}$$

where $\|\cdot\|_{\mathrm{TV}}$ is the total variation distance. Furthermore, the PDQ system is exponentially stable *if it is stable and there exist constants $B > 0$ and $c > 0$, and a norm-like function*[1] *$W : \mathcal{I} \times \mathcal{Q} \to [1,\infty)$ such that, for any $(i,q) \in \mathcal{I} \times \mathcal{Q}$,*

$$\| P_t((i,q);\cdot) - \mu_\phi(\cdot) \|_{\mathrm{TV}} \leq BW(i,q)e^{-ct}, \ \forall t \geq 0. \tag{2.15}$$

Finally, the PDQ system is said to be *unstable* if (2.14) does not hold.

[1] Following [74], W is norm-like if $W(i,q) \to \infty$ as $\|q\| \to \infty$ for $i \in \mathcal{I}$.

2.2 Stability of feedback-controlled PDQs

In this section, we study the stability of controlled PDQ systems. The main results are Theorem 2.1 (a necessary condition for stability) and Theorem 2.2 (a sufficient condition for stability).

2.2.1 Necessary condition for stability

Theorem 2.1. *Suppose that a PDQ system with n parallel servers is subject to a total demand $A \in \mathbb{R}_{\geq 0}$ and is controlled by an admissible policy ϕ. If the PDQ system is stable, then*

$$\sum_{i \in \mathcal{I}} \mathsf{p}_i \varphi^i_{kk} \leq \sum_{i \in \mathcal{I}} \mathsf{p}_i u^i_k, \; k \in \{1, \dots, n\}, \tag{2.16}$$

where p_i are given by (2.4) *and φ^i_{kk} are given by* (2.9).

Proof. Suppose that the PDQ system is stable.

For each server $k \in \{1, \dots, n\}$ and for each initial condition $(i, q) \in \mathcal{I} \times \mathcal{Q}$, we obtain from (2.6) and (2.7c) that, for all $t \geq 0$,

$$Q_k(t) = \int_0^t \Big(\phi_k(I(s), Q(s)) - r^\phi_k(I(s), Q(s)) \Big)\, ds + q_k.$$

Since $\lim_{t \to \infty} q_k/t = 0$, we have

$$\begin{aligned} 0 &= \lim_{t \to \infty} \frac{1}{t} \left(\int_0^t \Big(\phi_k(I(s), Q(s)) - r^\phi_k(I(s), Q(s)) \Big)\, ds + q_k - Q_k(t) \right) \\ &= \lim_{t \to \infty} \frac{1}{t} \left(\int_0^t \Big(\phi_k(I(s), Q(s)) - r^\phi_k(I(s), Q(s)) \Big) ds - Q_k(t) \right). \end{aligned} \tag{2.17}$$

Since the k-th queue is stable, for each initial condition $(i, q) \in \mathcal{I} \times \mathcal{Q}$, $\Pr\{\lim_{t \to \infty} Q(t) = \infty\} = 0$ (i.e. *non-evanescence*, see [74, pp. 524] for details), and we have $\lim_{t \to \infty} Q_k(t)/t =$

0 a.s. Hence, we can rewrite (2.17) as

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t \Big(\phi_k(I(s), Q(s)) - r_k^{\phi}(I(s), Q(s))\Big) ds = 0, \ a.s.$$

Now we can make two observations. First, by Assumption 2 (monotonicity), we have

$$\begin{aligned} \phi_k(I(s), Q(s)) &\geq \phi_k(I(s), Q_k(s)\mathbf{e}_k) \\ &\geq \lim_{q_k\to\infty} \phi_k(I(s), q_k\mathbf{e}_k) = \varphi_{kk}^{I(s)}, \ \forall s \geq 0. \end{aligned}$$

Secondly, recall that (2.5) implies $r_k^{\phi}(I(s), Q(s)) \leq U_k(s)$ for $s \geq 0$. Thus, we have

$$\begin{aligned} 0 &= \lim_{t\to\infty} \frac{1}{t} \int_0^t \Big(\phi_k(I(s), Q(s)) - r_k^{\phi}(I(s), Q(s))\Big) ds \\ &\geq \lim_{t\to\infty} \frac{1}{t} \int_0^t \Big(\varphi_{kk}^{I(s)} - U_k(s)\Big) ds. \end{aligned} \tag{2.18}$$

In addition, for every $i \in \mathcal{I}$, let $M_i(t)$ be the amount of time that the PDQ system is in mode i up to time t, i.e.:

$$M_i(t) = \int_0^t \mathbf{1}_{\{I(s)=i\}} ds.$$

Then, under Assumption 1, we have

$$\lim_{t\to\infty} \frac{M_i(t)}{t} = \mathsf{p}_i, \ a.s. \ \forall i \in \mathcal{I}.$$

Hence,

$$\begin{aligned} \lim_{t\to\infty} \frac{1}{t} \int_0^t \varphi_{kk}^{I(s)} ds &= \lim_{t\to\infty} \frac{1}{t} \int_0^t \Big(\sum_{i\in\mathcal{I}} \mathbf{1}_{\{I(s)=i\}} \varphi_{kk}^i\Big) ds \\ = \lim_{t\to\infty} \sum_{i\in\mathcal{I}} \frac{M_i(t)}{t} \varphi_{kk}^i &= \sum_{i\in\mathcal{I}} \mathsf{p}_i \varphi_{kk}^i, \ a.s. \end{aligned} \tag{2.19}$$

Similarly, we can obtain

$$\lim_{t\to\infty} \frac{1}{t} \int_0^t U_k(s)ds = \sum_{i\in\mathcal{I}} \mathsf{p}_i u_k^i, \ a.s. \tag{2.20}$$

Combining (2.18)–(2.20), we obtain (2.16). □

Theorem 2.1 provides a way of identifying unstable control policies. As argued in the proof, φ_{kk}^i is in fact the lower bound for $\phi_k(i,q)$ for all $q \in \mathcal{Q}$. Hence, Theorem 2.1 essentially states that if the PDQ system is stable, then the (time-average) lower bound of the inflow does not exceed the average saturation rate.

2.2.2 Sufficient condition for stability

To introduce our next result, we define $R_{\min} = [R_{\min}^1, \dots, R_{\min}^m]^T$ as follows:

$$R_{\min}^i = \min_k \left(u_k^i + \sum_{h:h\neq k} \min\{u_h^i, \varphi_{hk}^i\} \right), \quad i \in \mathcal{I}. \tag{2.21}$$

One can interpret $R_{\min}^i$ as a lower bound on the total discharge rate of the n servers in mode i when at least one of the n servers has a non-zero queue. Our next result uses $R_{\min}^i$ to provide a sufficient condition for the stability of feedback-controlled PDQ systems.

Theorem 2.2. *Suppose that a PDQ system of n parallel servers is subject to a total demand $A \in \mathbb{R}_{\geq 0}$ and is controlled by an admissible policy ϕ. Let the elements of the vector $R_{\min}$ be as defined in (2.21). Then, the PDQ system is stable if*

$$\exists i^* \in \mathcal{I},\ \forall k \in \{1, \dots, n\}, \quad \phi_k(i^*, 0) < u_k^{i^*}, \tag{2.22}$$

and if

$$\exists a = [a^1, \dots, a^m]^T \in \mathbb{R}_{>0}^m,\ \exists b > 0,$$
$$\Big(\mathrm{diag}(A\mathbf{e} - R_{\min})b + \Lambda\Big) a \leq -\mathbf{e}, \tag{2.23}$$

where $\mathbf{e}$ *is the* m*-dimensional vector of 1's. Furthermore, under the above conditions, there exists a positive constant* $c = \min_{i \in \mathcal{I}} 1/(2a^i)$ *such that, for some* $B > 0$,

$$\|P_t((i,q);\cdot) - \mu_\phi(\cdot)\|_{\mathrm{TV}} \leq B\left(a^i e^{b|q|} + 1\right)e^{-ct}, \qquad \forall (i,q) \in \mathcal{I} \times \mathcal{Q},\ \forall t \geq 0, \tag{2.24}$$

where μ_ϕ *is the unique invariant probability measure.*

The proof of Theorem 2.2 is based on a more general result [74, Theorem 6.1], which we recall here in the setting of PDQ systems. To conclude stability of the PDQ system, [74, Theorem 6.1] requires that the following two conditions hold:

(A) For any two initial conditions $(i,q), (j,\ell) \in \mathcal{I} \times \mathcal{Q}$, there exist $\delta > 0$ and $T > 0$ such that

$$\|P_T((i,q);\cdot) - P_T((j,\ell);\cdot)\|_{\mathrm{TV}} \leq 1 - \delta. \tag{2.25}$$

(B) There exist a norm-like function $V : \mathcal{I} \times \mathcal{Q} \to \mathbb{R}_{\geq 0}$ (called the *Lyapunov function*) and constants $c > 0$ and $d < \infty$ such that

$$\mathcal{L}V(i,q) \leq -cV(i,q) + d, \quad \forall (i,q) \in \mathcal{I} \times \mathcal{Q}. \tag{2.26}$$

Condition (A) is required for the uniqueness of the invariant probability measure [28]. Condition (B) is usually referred to as the *drift condition*, which essentially ensures the existence of invariant probability measures [74, Theorem 4.5].

We are now ready to prove the theorem:

Proof of Theorem 2.2. Suppose that (2.22) and (2.23) hold. We verify condition (A) (resp. (B)) using (2.22) (resp. (2.23)).

Condition (A):

Consider any initial condition $(i_0, q_0) \in \mathcal{I} \times \mathcal{Q}$.

First, Assumption 1 ensures that the Markov process $\{I(t), Q(t); t \geq 0\}$ recurrently visits the mode i^*. That is, for any $X_1 > 0$, there exists $\sigma > 0$ such that

$$\Pr\{I(X_1) = i^* | I(0) = i_0, Q(0) = q_0\} = \sigma. \tag{2.27}$$

Furthermore, we can obtain from (2.7c) that

$$\begin{aligned} |Q(X_1)| &= \left| q_0 + \int_0^{X_1} \Big(A - \sum_{k=1}^{n} r_k^{\phi}(I(s), Q(s)) \Big) ds \right| \\ &\leq |q_0| + \int_0^{X_1} \left| A - \sum_{k=1}^{n} r_k^{\phi}(I(s), Q(s)) \right| ds \\ &\leq |q_0| + AX_1. \end{aligned} \tag{2.28}$$

Secondly, in mode i^*, the vector of queue length $Q(t)$ necessarily converges to $q^* = 0$. To see this, consider mode i^* and any $q \in \mathcal{Q}$. For each $k \in \{1, \dots, n\}$ such that $q_k = 0$, by Assumption 2, we have $\phi_k(i^*, q) \geq \phi_k(i^*, 0)$, and thus

$$\begin{aligned} r_k^{\phi}(i^*, q) &= \min\{u_k^{i^*}, \phi_k(i^*, q)\} \\ &\geq \min\{u_k^{i^*}, \phi_k(i^*, 0)\} = r_k^{\phi}(i^*, 0). \end{aligned} \tag{2.29}$$

Therefore, for each $q \in \mathcal{Q} \backslash \{0\}$, we have

$$\begin{aligned} \sum_{k=1}^{n} D_k^{\phi}(i^*, q) &\overset{(2.6)}{=} A - \sum_{k=1}^{n} r_k^{\phi}(i^*, q) \\ &= A - \sum_{k: q_k > 0} u_k^{i^*} - \sum_{k: q_k = 0} r_k^{\phi}(i^*, q) \\ &\overset{(2.29)}{\leq} A - \sum_{k: q_k > 0} u_k^{i^*} - \sum_{k: q_k = 0} r_k^{\phi}(i^*, 0) \\ &\leq A - \min_{k: q_k > 0} \Big(u_k^{i^*} + \sum_{\substack{h: q_h > 0 \\ h \neq k}} r_h^{\phi}(i^*, 0) \Big) - \sum_{k: q_k = 0} r_k^{\phi}(i^*, 0) \\ &\leq A - \min_{k \in \{1, \dots, n\}} \Big(u_k^{i^*} + \sum_{h \neq k} r_h^{\phi}(i^*, 0) \Big) \end{aligned}$$

$$\overset{(2.22)}{<} A - \sum_{k=1}^{n} r_k^{\phi}(i^*, 0) \overset{(2.5)(2.22)}{=} 0. \tag{2.30}$$

One can see from (2.28) and (2.30) that there exists

$$X_2 = \frac{|q_0|+AX_1}{A - \min_k \left(u_k^{i^*} + \sum_{h\neq k} r_h^{\phi}(i^*, q)\right)}$$

such that $Q(X_1 + X_2) = 0$ if $I(t) = i^*$ for all $t \in [X_1, X_2 + X_2)$. Note that

$$\Pr\{I(t) = i^*; t \in [X_1, X_1 + X_2)|I(X_1) = i^*\} = e^{-\nu_{i^*}X_2}.$$

Thus, we have

$$\Pr\{Q(X_1 + X_2) = 0|I(0) = i_0, Q(0) = q_0\} \geq \sigma e^{-\nu_{i^*}X_2} > 0,$$

where σ satisfies (2.27). Hence, we have

$$P_{X_1+X_2}((i_0, q_0), \{(i^*, 0)\}) \geq \sigma e^{-\nu_{i^*}X_2}.$$

Then, for any $T \geq X_1 + X_2$, we have

$$P_T((i_0, q_0), \{(i^*, 0)\}) \geq \sigma e^{-\nu_{i^*}(T-X_1)}.$$

Thus, for arbitrary initial conditions (i, q) and (j, ℓ), there exist $\sigma' > 0$, $X_1' > 0$, $X_2' > 0$, and $T' > 0$ such that

$$P_{T'}((i, q), \{(i^*, 0)\}) \geq \sigma' e^{-\nu_{i^*}(T'-X_2')},$$
$$P_{T'}((j, \ell), \{(i^*, 0)\}) \geq \sigma' e^{-\nu_{i^*}(T'-X_2')},$$

which verifies (2.25) with $T = T'$ and $\delta = \sigma' e^{-\nu_{i^*}(T'-X_2')}$.

Condition (B):

Consider the Lyapunov function

$$V(i,q) = a^i e^{b|q|}, \quad (i,q) \in \mathcal{I} \times \mathcal{Q}, \tag{2.31}$$

where $a^1, \dots, a^m$, and b are positive constants.

For each server k, by the definition of φ^i_{kh} (2.9), there necessarily exists $L_k < \infty$ such that, for all $h \neq k$,

$$\min\left\{u^i_h, \phi_h(i, L_k \mathbf{e}_k)\right\} \geq \min\left\{u^i_h, \varphi^i_{hk}\right\} - \frac{1}{2nb \max_{j\in\mathcal{I}} a^j}. \tag{2.32}$$

Let $L = [L_1, \dots, L_n]^T$. We claim that the constants

$$c := \frac{1}{2\max_{j\in\mathcal{I}} a^j}, \tag{2.33a}$$

$$d := \max_{i\in\mathcal{I}} |\mathcal{L}V(i,L) + cV(i,L)|, \tag{2.33b}$$

verify the drift condition (2.26). Let us prove this claim.

Plugging the Lyapunov function defined in (2.31) into the expression of the infinitesimal generator (2.8), we obtain

$$\mathcal{L}V(i,q) = \left(\sum_{k=1}^{n} \left(\phi_k(i,q) - r^\phi_k(i,q)\right) a^i b \right. \\ \left. + \sum_{j\in\mathcal{I}} \lambda_{ij}(a^j - a^i) \right) e^{b|q|}. \tag{2.34}$$

Then, to check (2.26), we need to consider two cases:

Case I: $q \in \{\zeta \in \mathcal{Q} : 0 \leq \zeta \leq L\}$. Since each such q's are bounded, $V(i,q)$ is also bounded. Hence, we can verify in a rather straightforward manner that, with c and d given by (2.33), $\mathcal{L}V \leq -cV + d$ for all $i \in \mathcal{I}$ and $0 \leq q \leq L$.

Case II: $q \in \mathcal{Q}\backslash\{\zeta \in \mathcal{Q} : 0 \leq \zeta \leq L\}$. For each such q, there necessarily exists a server k_1 such that $q_{k_1} > L_{k_1}$. For the k_1-th server, since $q_{k_1} > L_{k_1} \geq 0$, we have

$$r^\phi_{k_1}(i,q) = u^i_{k_1}, \; \forall i \in \mathcal{I}. \tag{2.35}$$

For the other servers, i.e. for each $h \neq k_1$, we have

$$\begin{aligned} r_h^{\phi}(i,q) &\overset{(2.5)}{=} \min\left\{u_h^i, \phi_h(i,q)\right\} \\ &\geq \min\{u_h^i, \phi_h(i, q_{k_1}\mathbf{e}_{k_1})\} \geq \min\{u_h^i, \phi_h(i, L_{k_1}\mathbf{e}_{k_1})\} && (2.36\text{a}) \\ &\overset{(2.32)}{\geq} \min\{u_h^i, \varphi_h^{k_1}(i)\} - \frac{1}{2nb\max_{j\in\mathcal{I}} a^j}, \ \forall i \in \mathcal{I}, && (2.36\text{b}) \end{aligned}$$

where (2.36a) results from Assumption 2 (monotonicity). Combining (2.35) and (2.36b), we can write

$$\begin{aligned} \sum_{h=1}^{n} r_h^{\phi}(i,q) &\geq u_{k_1}^i + \sum_{h:h\neq k_1} \min\left\{u_h^i, \varphi_h^{k_1}(i)\right\} - \frac{1}{2b\max_{j\in\mathcal{I}} a^j} \\ &\overset{(2.21)}{\geq} R_{\min}^i - \frac{1}{2b\max_{j\in\mathcal{I}} a^j}. && (2.37) \end{aligned}$$

Then,

$$\begin{aligned} &\sum_{k=1}^{n} \left(\phi_k(i,q) - r_k^{\phi}(i,q)\right) a^i b + \sum_{j\in\mathcal{I}} \lambda_{ij}(a^j - a^i) \\ &\overset{(2.37)}{\leq} \left(A - R_{\min}^i + \frac{1}{2b\max_{j\in\mathcal{I}} a^j}\right) a^i b + \sum_{j\in\mathcal{I}} \lambda_{ij}(a^j - a^i) \\ &\overset{(2.23)}{\leq} -1 + \frac{1}{2} = -\frac{1}{2}. \end{aligned}$$

Finally,

$$\mathcal{L}V(i,q) \overset{(2.34)}{\leq} -\frac{1}{2}e^{b|q|} \overset{(2.33\text{a})}{\leq} -ca^i e^{b|q|} \overset{(2.31)}{=} -cV.$$

Hence, (2.26) holds for all $i \in \mathcal{I}$, all $q \in \mathcal{Q}\backslash\{q : 0 \leq \zeta \leq L\}$, and all $d \geq 0$.

Thus, we have verified that the drift condition (2.26) holds for all $(i,q) \in \mathcal{I} \times \mathcal{Q}$.

Finally, note that we have verified conditions (A) and (B) for the controlled PDQ system. Thus, we obtain from [74, Theorem 6.1] that the PDQ system is exponentially stable.

□

The condition (2.22) states that there exists a mode i^* in which every queue decreases to zero. Practically, one can interpret i^* as a "nominal" or "normal" mode in which the saturation rates are sufficiently high and satisfy (2.22). This condition leads to Condition (A).

The condition (2.23) essentially imposes a lower bound on the total discharged flow from the n servers, which is characterized by $R^i_{\min}$. This condition leads to Condition (B). To verify this condition, one needs to determine whether BMI (2.23) admits positive solutions for $a_1, \dots, a_m$ and b. This can be done using the known computational methods to solve BMIs (see e.g. [97, 71]).

Remark 2.1. *Using the exponential Lyapunov function (31), one can also apply [74, Theorem 4.3] to obtain that, under (23), for each initial condition* $(i, q) \in \mathcal{I} \times \mathcal{Q}$, *we have*

$$\limsup_{t\to\infty} \frac{1}{t} \int_0^t \mathsf{E}[e^{|Q(s)|}]ds < \infty.$$

That is, moments of the queue lengths are bounded.

2.3 Two-mode systems

If the system has only two modes, solutions for b and a can be constructed in a more straightforward manner, which motivates the next result.

Proposition 2.1. *A PDQ system of n parallel servers with two modes* $\{1, 2\}$ *and with an admissible control policy ϕ is stable if*

$$\exists i^* \in \{1,2\},\ \phi_k(i^*, 0) < u_k^{i^*},\ k \in \{1,2\}, \tag{2.38}$$

and if

$$A < \mathsf{p}_1 R^1_{\min} + \mathsf{p}_2 R^2_{\min}, \tag{2.39}$$

where $R^i_{\min}$ is defined in (2.21).

Proof. First, let us define the following quantities

$$D_{\min} = \min\left\{A - R^1_{\min}, A - R^2_{\min}\right\}, \tag{2.40a}$$

$$D_{\max} = \max\left\{A - R^1_{\min}, A - R^2_{\min}\right\}, \tag{2.40b}$$

$$\overline{D} = A - (\mathsf{p}_1 R^1_{\min} + \mathsf{p}_2 R^2_{\min}), \tag{2.40c}$$

$$i_{\min} = \begin{cases} 1, & \text{if } D_{\min} = A - R^1_{\min}, \\ 2, & \text{o.w.} \end{cases} \tag{2.40d}$$

$$i_{\max} = \begin{cases} 2, & \text{if } D_{\min} = A - R^1_{\min}, \\ 1, & \text{o.w.} \end{cases} \tag{2.40e}$$

$$\lambda_{\min} = \begin{cases} \lambda_{12}, & \text{if } D_{\min} = A - R^1_{\min}, \\ \lambda_{21}, & \text{o.w.} \end{cases} \tag{2.40f}$$

$$\lambda_{\max} = \begin{cases} \lambda_{21}, & \text{if } D_{\min} = A - R^1_{\min}, \\ \lambda_{12}, & \text{o.w.} \end{cases} \tag{2.40g}$$

Under (2.39), we explicitly construct constants $a^{i_{\min}}$, $a^{i_{\max}}$, and b satisfying the BMI (2.23). Condition (2.39) implies

$$A - \mathsf{p}_1 R^1_{\min} - \mathsf{p}_2 R^2_{\min} = \mathsf{p}_{i_{\min}} D_{\min} + \mathsf{p}_{i_{\max}} D_{\max} < 0 \tag{2.41}$$

Since $D_{\min} \leq D_{\max}$, (2.41) implies that $D_{\min} < 0$. Thus, we only need to consider two cases:

In the case that $D_{\min} < 0$, $D_{\max} \leq 0$, we can select an arbitrary $a^{i_{\min}} > \max_i\{1/\lambda_i\}$ and let

$$a^{i_{\max}} = 2a^{i_{\min}}, \; b = \frac{\lambda_{\min} a^{i_{\min}} + 1}{-D_{\min} a^{i_{\min}}}. \tag{2.42}$$

It is not hard to see that $a^{i_{\min}}$, $a^{i_{\max}}$, and b are positive and satisfy the BMI (2.23).

In the case that $D_{\text{min}} < 0$, $D_{\text{max}} > 0$, we let

$$b = \frac{(\lambda_{12} + \lambda_{21})\overline{D}}{2D_{\text{min}}D_{\text{max}}}, \tag{2.43a}$$

$$a^{i_{\text{min}}} = \frac{-D_{\text{max}}b + \lambda_{12} + \lambda_{21}}{\det[\operatorname{diag}(A\mathbf{e} - R_{\text{min}})b + \Lambda]}, \tag{2.43b}$$

$$a^{i_{\text{max}}} = \frac{-D_{\text{min}}b + \lambda_{12} + \lambda_{21}}{\det[\operatorname{diag}(A\mathbf{e} - R_{\text{min}})b + \Lambda]}. \tag{2.43c}$$

Now, we show that these constants are positive. First, note that (2.39) implies $\overline{D} < 0$. Then, since $D_{\text{min}} < 0$ and $D_{\text{max}} > 0$, and since $\overline{D} < 0$, b is positive. Secondly, to see that $a^{i_{\text{min}}} > 0$, note that

$$\begin{aligned}
&\operatorname{diag}(A\mathbf{e} - R_{\text{min}})b + \Lambda \\
&= \begin{bmatrix} b(A - R^1_{\text{min}}) - \lambda_{12} & \lambda_{12} \\ \lambda_{21} & b(A - R^2_{\text{min}}) - \lambda_{21} \end{bmatrix},
\end{aligned}$$

and

$$\begin{aligned}
&\det[\operatorname{diag}(A\mathbf{e} - R_{\text{min}})b + \Lambda] \\
&= b^2 \left(A - R^1_{\text{min}}\right)\left(A - R^2_{\text{min}}\right) \\
&\qquad - \lambda_{12}b\left(A - R^2_{\text{min}}\right) - \lambda_{21}b\left(A - R^1_{\text{min}}\right) \\
&= b^2 \left(A - R^1_{\text{min}}\right)\left(A - R^2_{\text{min}}\right) - b(\lambda_{12} + \lambda_{21})\overline{D} \\
&= b^2 D_{\text{min}}D_{\text{max}} - b(\lambda_{12} + \lambda_{21})\overline{D}.
\end{aligned}$$

Again, since $D_{\text{min}} < 0$ and $D_{\text{max}} > 0$, one can check that the b given in (2.43a) ensures that $\det[\operatorname{diag}(A\mathbf{e} - R_{\text{min}})b + \Lambda] > 0$. In addition, note that

$$\begin{aligned}
b &= \frac{(\lambda_{12} + \lambda_{21})\overline{D}}{2D_{\text{min}}D_{\text{max}}} \\
&= \frac{\lambda_{12} + \lambda_{21}}{D_{\text{max}}}\left(\frac{-\mathsf{p}_{i_{\text{min}}}D_{\text{min}} - \mathsf{p}_{i_{\text{max}}}D_{\text{max}}}{-2D_{\text{min}}}\right) \\
&< \frac{\lambda_{12} + \lambda_{21}}{D_{\text{max}}}\left(\frac{-\mathsf{p}_{i_{\text{min}}}D_{\text{min}} - \mathsf{p}_{i_{\text{max}}}D_{\text{min}}}{-2D_{\text{min}}}\right)
\end{aligned}$$

$$= \frac{\lambda_{12} + \lambda_{21}}{2D_{\max}} < \frac{\lambda_{12} + \lambda_{21}}{D_{\max}},$$

which, along with $D_{\max} > 0$, implies $a^{i_{\min}} > 0$. Finally, since $D_{\min} < 0$, $a^{i_{\max}}$ is also positive.

From (2.40d) and (2.40e), we know that

$$a^1 = \begin{cases} a^{i_{\min}}, & \text{if } D_{\min} = A - R^1_{\min}, \\ a^{i_{\max}}, & \text{o.w.} \end{cases}$$
$$a^2 = \begin{cases} a^{i_{\max}}, & \text{if } D_{\min} = A - R^1_{\min}, \\ a^{i_{\min}}, & \text{o.w.} \end{cases}$$

Let $a = [a^1, a^2]^T$. Then, one can check that a and b satisfy

$$[\text{diag}(A\mathbf{e} - R_{\min})b + \Lambda]a = -\mathbf{e},$$

and thus satisfy the BMI (2.23).

In addition, (2.38) is analogous to (2.22). Thus, we can conclude from Theorem 2.2 that the two-mode PDQ system is stable.

□

In comparison to Theorem 2.2, Proposition 2.1 provides a simpler criterion (2.39) for stability of PDQ systems with two modes, since it does not involve solving a BMI.

Furthermore, if a PDQ system with two modes is controlled by a mode-responsive routing policy (2.10), then we can obtain a *necessary and sufficient condition* for stability:

Proposition 2.2. *A system of n parallel servers two modes $\{1, 2\}$ and with a mode-responsive routing policy ϕ given by* (2.10) *is stable if and only if*

$$\mathsf{p}_1\psi^1_k + \mathsf{p}_2\psi^2_k < \mathsf{p}_1 u^1_k + \mathsf{p}_2 u^2_k, \ \forall k \in \{1, \dots, n\}. \tag{2.44}$$

Proof. Since the system is controlled by a mode-responsive policy, the queues in various servers do not interact. Therefore, we can consider the n servers independently. For the k-th server, consider the Lyapunov function

$$V_k(i, q_k) = a_k^i \exp(b_k q_k),\ (i, q) \in \{1, 2\} \times \mathbb{R}_{\geq 0}$$

with parameters $[a_k^1, a_k^2]^T \in \mathbb{R}_{>0}^2$ and $b_k > 0$. With this Lyapunov function, one can adapt the proof of Proposition 2.1 and conclude that the k-th server is stable if (2.44) holds.

To obtain the necessity of (2.44), first note that the k-th server is unstable if $\mathsf{p}_1\psi_k^1 + \mathsf{p}_2\psi_k^2 > \mathsf{p}_1 u_k^1 + \mathsf{p}_2 u_k^2$. Secondly, to argue that the k-th server is unstable if

$$\mathsf{p}_1\psi_k^1 + \mathsf{p}_2\psi_k^2 = \mathsf{p}_1 u_k^1 + \mathsf{p}_2 u_k^2, \tag{2.45}$$

one can first assume by contradiction the existence of an invariant probability measure μ_ϕ, and then consider $\mu_\phi(\mathcal{I} \times \{0\})$ to arrive at a contradiction to (2.45). □

In addition, for the setting of Proposition 2.2, expression for the invariant probability measure μ_ϕ has been reported in the literature [57], which makes possible analytical optimization of the routing policy.

2.4 Insights for incident-aware routing

In this section, we demonstrate how our results can provide insights for traffic flow routing under stochastic capacity fluctuations. Consider a network of two parallel servers. The total inflow is $A = 1$. Our results in Section 2.2 can be applied to obtain stability conditions of this network. We focus on the practically motivated routing policies given in (2.10)–(2.12).

2.4.1 A two-mode network

Suppose that the network has two modes $\{1, 2\}$ with symmetric transition rates $\lambda_{12} = \lambda_{21} = 1$. Thus, the steady-state probabilities are $\mathsf{p}_1 = \mathsf{p}_2 = 0.5$. The saturation rates in both modes are given as $u^1 = [1.2, 0.7]^T$ and $u^2 = [0.2, 0.7]^T$. Thus, both servers have an average saturation rate of 0.7.

Mode-responsive routing

For this two-mode system, the policy given by (2.10) can be parametrized by two constants $\psi_1^1, \psi_1^2 \in [0,1]$ (note that admissibility requires $\psi_1^i + \psi_2^i = 1$ for $i \in \{1,2\}$). By Proposition 2.2, the routing policy ϕ^{mod} is stabilizing if and only if

$$0.3 < (\psi_1^1 + \psi_1^2)/2 < 0.7, \quad \psi_1^1 \in [0,1], \psi_1^2 \in [0,1].$$

That is, the PDQ system is stable if and only if the average inflows into each server are less than their respective average saturation rate (note that $(\psi_1^1 + \psi_1^2)/2 > 0.3$ is equivalent to $(\psi_2^1 + \psi_2^2)/2 < 0.7$).

Piecewise-affine feedback routing

Consider the policy given by (2.11). Admissibility requires $\alpha_{11} = \alpha_{12}$, $\alpha_{21} = \alpha_{22}$, and $\theta_1 + \theta_2 = 1$. Hence, we denote $\alpha_1 = \alpha_{11} = \alpha_{12}$ and $\alpha_2 = \alpha_{21} = \alpha_{22}$. For $k = 1, 2$ and $i \in \{1,2\}$, the expression of the limiting inflows (2.9) are as follows:

$$\varphi_{kk}^i = \begin{cases} 0, & \text{if } \alpha_k > 0, \\ \min\{A, \theta_k\}, & \text{if } \alpha_k = 0, \end{cases}$$
$$\varphi_{kh}^i = \begin{cases} 1, & \text{if } \alpha_h > 0, \\ \min\{A, \theta_k\}, & \text{if } \alpha_h = 0, \end{cases} \quad h \neq k.$$

Table 2.1 shows the necessary condition for stability given by Theorem 2.1 and the sufficient condition for stability given by Proposition 2.1. Note that the restriction on θ_k is stronger if $\alpha_k = 0$. The intuition is that, if the routing policy is not responsive to the queue length in a server, then an appropriate selection of the nominal inflow θ_k is crucial to ensure stability. In addition, the structures of the stability conditions strongly depend on whether α_k is zero, but not on the exact magnitude of α_k. In this example, the gap between the necessary condition and the sufficient condition mainly

Table 2.1: Stability conditions (two modes, PWA routing).

α_1	α_2	Necessary condition	Sufficient condition
$=0$	$=0$	$0.3 \leq \theta_1 \leq 0.7$	$0.3 < \theta_1 < 0.7$
$=0$	>0	$\theta_1 \leq 0.7$	$0.3 < \theta_1 < 0.7$
>0	$=0$	$\theta_1 \geq 0.3$	$\theta_1 > 0.3$
>0	>0	$\theta_1 \in \mathbb{R}$	$\theta_1 > 0.3$

results from the condition (2.22), which requires $\theta_1 > 0.3$.

Logit routing

Now, consider the policy (2.12). For $k = 1, 2$, $i \in \{1, 2\}$, the limiting inflows are

$$\varphi^i_{kk} = \begin{cases} 0, & \text{if } \beta_k > 0, \\ \frac{A \exp(\gamma_k)}{\sum_{h=1}^{2} \exp(\gamma_h)}, & \text{if } \beta_k = 0, \end{cases} \tag{2.46a}$$

$$\varphi^i_{kh} = \begin{cases} A, & \text{if } \beta_h > 0, \\ \frac{A \exp(\gamma_k)}{\sum_{h=1}^{2} \exp(\gamma_h)}, & \text{if } \beta_h = 0, \end{cases} \quad h \neq k. \tag{2.46b}$$

Again, we can obtain stability conditions from Theorem 2.1 and Proposition 2.1. Table 2.2 implies that the constants γ_k have a stronger impact on stability of the PDQ system than the coefficients β_k capturing the sensitivity to queue lengths. Once again, the gap between the necessary condition and the sufficient condition results from (2.22), which requires $|\gamma_1 - \gamma_2| < \log(7/3)$.

Table 2.2: Stability conditions (two modes, logit routing).

β_1	β_2	Necessary condition	Sufficient condition
$=0$	$=0$	$\lvert\gamma_1 - \gamma_2\rvert \leq \log(7/3)$	$\lvert\gamma_1 - \gamma_2\rvert < \log(7/3)$
$=0$	>0	$\gamma_1 - \gamma_2 \leq \log(7/3)$	
>0	$=0$	$\gamma_1 - \gamma_2 \geq -\log(7/3)$	
>0	>0	$\gamma_1 \in \mathbb{R}, \gamma_2 \in \mathbb{R}$	

2.4.2 A three-mode network

Suppose that the network has three modes $\{1, 2, 3\}$ with symmetric transition rates $\lambda_{ij} = 1$ for all $i, j \in \mathcal{I}$. Thus, the steady-state probabilities are $\mathsf{p}_1 = \mathsf{p}_2 = \mathsf{p}_3 = 1/3$.

The saturation rates in the three modes are $u^1 = [1.2, 0.7]^T$, $u^2 = [0.7, 0.7]^T$, and $u^3 = [0.2, 0.7]^T$; i.e. the average saturation rates are equal to those in the two-mode case. The main difference between the analysis in this subsection and that in the previous subsection is that the sufficient conditions for stability below are obtained numerically (in terms of solving the BMI (2.23)) instead of analytically.

Mode-responsive routing

For ease of presentation, we assume that $\psi_k^2 = \psi_k^3$ for $k \in \{1, 2\}$. The limiting inflows φ_{kh}^i are given by

$$\varphi_{kh}^i = \psi_k^i, \ h \in \{1, 2\}, k \in \{1, 2\}, i \in \mathcal{I}.$$

Theorem 2.1 gives a necessary condition for stability:

$$0.3 \le 1/3\psi_1^1 + 2/3\psi_1^2 \le 0.7, \tag{2.47}$$

whose complement is the "Unstable" region in Figure 2-2. Figure 2-2 also shows a "Stable" region obtained from Theorem 2.2; the BMI (2.23) is solved using YALMIP [71]. In contrast to the two-mode case, there is an "Unknown" region between the "Stable" and "Unstable" regions, due to the gap between the necessary condition (Theorem 2.1) and the sufficient condition (Theorem 2.2).

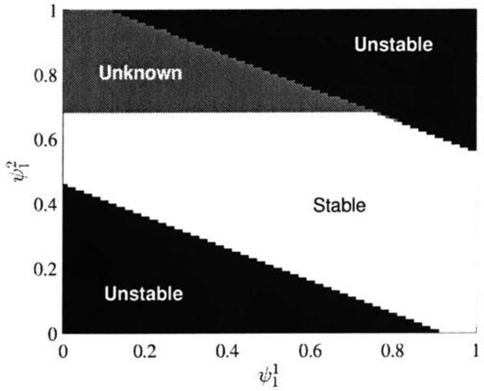


Figure 2-2: Stability of various (ψ_1^1, ψ_1^2) pairs.

Queue-responsive routing policies

For the piecewise-affine routing policy (2.11) and the logit routing policy (2.12),

Table 2.3: Stability conditions (three modes, PWA routing).

α_1	α_2	Necessary condition	Sufficient condition
$= 0$	$= 0$	$0.3 \leq \theta_1 \leq 0.7$	$0.41 \leq \theta_1 \leq 0.59$
$= 0$	> 0	$\theta_1 \leq 0.7$	$0.41 \leq \theta_1 \leq 0.59$
> 0	$= 0$	$\theta_1 \geq 0.3$	$\theta_1 \geq 0.36$
> 0	> 0	$\theta_1 \in \mathbb{R}$	$\theta_1 > 0.3$

Tables 2.3 and 2.4 show the stability conditions. In comparison to the two-mode case, the necessary conditions are unchanged, but the sufficient conditions in the three-mode case are more restrictive. This indicates that the sufficient condition becomes more restrictive as the number of modes (and thus the number of bilinear inequality constraints) increases.

2.5 Summary

In this chapter, we proposed an approach to routing policy design in networks with unreliable capacities, based on a network extension of the PDQ model introduced in the previous chapter. We model link saturation rates as piecewise-constant signals that randomly switch between finite sets of values. We derived a necessary condition (Theorem 2.1) for stability, which essentially states that the average inflow cannot exceed the average saturation rate. We also derived a sufficient condition (Theorem 2.2) based on properties of PDMPs and the Foster-Lyapunov criteria along with an argument for the uniqueness of the invariant probability measure. For bimodal PDQs, we refined the results (Propositions 2.1 and 2.2) and analyzed the impact of

Table 2.4: Stability conditions (three modes, logit routing).

β_1	β_2	Necessary condition	Sufficient condition
$= 0$	$= 0$	$\lvert\gamma_1 - \gamma_2\rvert \leq \log(7/3)$	$\lvert\gamma_1 - \gamma_2\rvert \leq \log 1.7$
$= 0$	> 0	$\gamma_1 - \gamma_2 \leq \log(7/3)$	
> 0	$= 0$	$\gamma_1 - \gamma_2 \geq -\log(7/3)$	
> 0	> 0	$\gamma_1 \in \mathbb{R}, \gamma_2 \in \mathbb{R}$	

control policies on the average queue length and the rate of convergence. Based on long-time properties of PDQs and their network extensions, we derive some useful insights for incident-aware routing policy design.

Chapter 3

Modeling Highway Traffic with Vehicle Platooning

Vehicle platooning is a promising technology that can lead to significant fuel savings and emission reduction. However, the macroscopic impact of vehicle platoons on highway traffic is not yet well understood. In this chapter, we propose a new fluid queuing model to study the macroscopic interaction between randomly arriving vehicle platoons and the background traffic at highway bottlenecks. Specifically, we focus on three questions:

1. How to model the sharing of highway capacity between vehicle platoons and the background traffic?

2. How do the key parameters of vehicle platoons, including penetration rate, platoon length, and vehicle spacing within a platoon, affect highway performance?

3. How to evaluate the strategies for allocating road capacity between ordinary vehicles and platoons?

Our analysis is based on a stochastic extension of the classical fluid queuing model, called the *piecewise-deterministic queuing (PDQ) model.* Our model (Section 3.1) captures the following important features of vehicle platoons. First, vehicle platoons can act as temporary bottlenecks for other vehicles. Second, the headways between platoons and the lengths of platoons are subject to random variations. We use a Markov process to capture such randomness. Third, vehicles within a platoon have smaller spacing compared to ordinary vehicles. Our stability analysis (Section 3.2) focuses on the queuing resulting from the interaction between the two classes of traffic. The analysis presented in this chapter are based on the following assumptions:

1. The headways between consecutive vehicle platoons are i.i.d. exponential random variables. This is a typical assumption for arrival processes with random headways [77].

2. The lengths of vehicle platoons are i.i.d. exponential random variables. Note that this assumption only applies to the fluid limit of traffic; in practice, the number of vehicles in a platoon is always an integer. Since the discrete correspondence of exponential distribution is geometric distribution, this assumption essentially means that the formation of a platoon is a Bernoulli process, which makes practical sense. In addition, this assumption ensures that the fluid queuing model is a Markov process, and thus significantly improves tractability. In reality, platoon lengths are more likely to be concentrated within a certain range (e.g. 2–10 vehicles) rather than spread from 1 to infinity. In this sense, our model overstates the variance in platoon lengths and thus overestimates platooning-induced congestion.

3. A platoon of n CAVs is equivalent to $(h/H)n$ ordinary vehicles in terms of queuing effect, where h and H are the inter-vehicle spacings between two CAVs and two ordinary vehicles, respectively. This assumption is consistent with the model proposed by [69].

Our PDQ model focuses on the aggregate congestion due to platooning. Note that the PDQ model does not account for (i) the spatial propagation of such congestion, or (ii) congestion due to microscopic interactions such as formation/split of platoons and speed difference between CAVs and ordinary vehicles. Regarding the first limitation, we demonstrated in [49] that the main insights derived from the PDQ model is consistent with those obtained from the more detailed and practical multi-class cell transmission model (CTM). We are also studying properties of tandem PDQ links to better understand the impact of propagation of platooning-induced congestion. Regarding the second limitation, part of our ongoing work is to establish the consistency (both theoretical and empirical) between microscopic CAV models and the PDQ model.

In the rest of this chapter, we first provide an intuitive stability result based on the theory of convergence of stochastic fluid queuing systems [74, 46]. We also consider the impact of key parameters of vehicle platoons on traffic queues (Section 3.3). Main insights include: (i) increasing the fraction of connected vehicles typically reduces congestion; however, if the highway is in free flow without platooning, then introduction of platooning may induce congestion due to the randomness in platoon arrivals; (ii) short platoons lead to less congestion than long platoons; (iii) prioritizing platoons over background traffic does not necessarily reduce congestion.

3.1 Traffic models with platoons

In this section, we introduce a stochastic two-class piecewise-deterministic queuing (PDQ) model for highway traffic with vehicle platooning at highway bottlenecks.

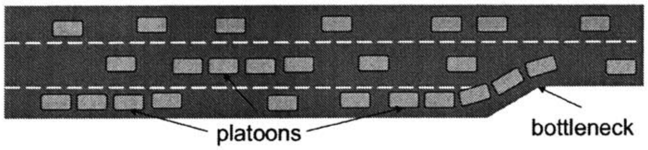


Figure 3-1: A highway bottleneck.

We focus on the most basic setting of a highway bottleneck with both vehicle platoons and ordinary vehicles (Figure 3-1). When a platoon is passing through the bottleneck, for a period of time, one lane is occupied by the platoon and not available to the background traffic. Thus, queuing happens upstream from the bottleneck.

3.1.1 Stochastic platoon arrival process

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

We assume that (i) the inter-platoon headways are i.i.d. and exponentially distributed with mean $1/\lambda$, and (ii) the numbers of vehicles in platoons are also i.i.d. and exponentially distributed with mean $v/(\mu h)$, where v is the *free-flow speed* and h is the *intra-platoon spacing.* These assumptions are motivated by the inherent uncertainty in the formation, split, and movement of platoons [63]. Specifically, the exponential distribution is commonly used to model the randomness in vehicle headways [43]. In addition, for our purposes, the random platoon lengths can be also modeled as exponentially distributed random variables. With these assumptions, we use a two-state Markov process to model the arrival of platoons. Thus, $\{B(t); t \geq 0\}$ is a continuous-time, two-state Markov process with state space $\mathcal{B} := \{0, v/h\}$; see Figure 3-2 for an illustration.

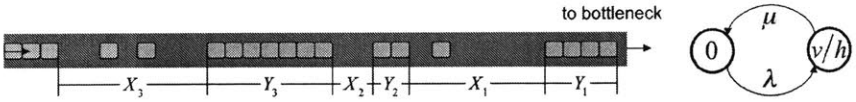


Figure 3-2: Platoon headway X_k and length Y_k are random (left). The arrival process of connected vehicles $B(t)$ is a two-state Markov process (right).

By standard results in Markov processes (see e.g. [33]), the average inflow rate of connected vehicles is

$$\overline{B} = \lim_{t \to \infty} \frac{1}{t} \int_0^t B(\tau) d\tau = \frac{\lambda}{\lambda + \mu} \frac{v}{h}, \quad \text{a.s.} \tag{3.1}$$

where "a.s." means almost surely.

3.1.2 Fluid queuing model

The fluid queuing model is a simple model that can be used to study highway bottlenecks [77]. The essence of the PDQ is to consider the highway bottleneck as a server with an infinite-sized buffer that stores the vehicles waiting for discharge. If there are vehicles waiting in the buffer, then the server discharges the vehicles at the *saturation rate*, denoted by u. The unit of u is veh/hr. If no traffic is waiting in the buffer, then the rate at which the server discharges traffic is the minimum of the saturation rate and the inflow rate.

The evolution of the traffic queue depends on the *priority rule*, i.e. how the server's saturation rate (i.e. the bottleneck's capacity) is allocated to the two traffic classes. Thanks to the simplicity of the PDQ, we can consider two operational policies for capacity allocation. In the first policy, we model a highway bottleneck as a single server with *proportional priority*; i.e., the road capacity allocated to a class of traffic is proportional to the fraction of this class of traffic in the aggregate traffic queue. In the second policy, vehicle platoons are prioritized; we name this policy *segmented priority*, which is motivated by the idea of dedicated lanes for connected vehicles [10].

Queuing dynamics: proportional priority

This priority rule corresponds to a highway where connected and ordinary vehicles share all lanes of the highway. This is a typical capacity allocation model for a highway that allows mixing between connected and ordinary vehicles [106].

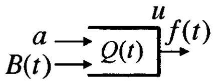


Figure 3-3: PDQ model under the proportional priority rule.

Figure 3-3 shows the two-class PDQ. The (hybrid) state of the fluid queuing system is (b, q^a, q^b), where $b \in \mathcal{B}$ is the inflow of connected vehicles, $q^a \in \mathbb{R}_{\geq 0}$ is the queue of ordinary vehicles, and $q^b \in \mathbb{R}_{\geq 0}$ is the queue of connected vehicles. To capture the reduced intra-platoon vehicle spacing, we scale down queues of connected vehicles according to the spacing reduction enabled by platooning. More specifically, currently available platooning technology is able to reduce intra-platoon spacing to less than half that between ordinary vehicles [2, 69]. We model this by scaling down the traffic queue and flow of connected vehicles with a coefficient (h/H). Thus, we define the *effective queue length* as

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

and the *effective discharge rate* as

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

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

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

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

$$f^a(b, q^a, q^b) = \begin{cases} \frac{q^a}{q^a + \frac{h}{H} q^b} f(b, q^a + \frac{h}{H} q^b), & q^a + q^b > 0, \\ \min\left\{a, \frac{a}{a + \frac{h}{H} b} u\right\}, & q^a + q^b = 0, \end{cases} \tag{3.2a}$$

$$\frac{h}{H} f^b(b, q^a, q^b) = f(b, q^a + \frac{h}{H} q^b) - f^a(b, q^a, q^b). \tag{3.2b}$$

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

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

$$Q^a(0) = q^a, \ \frac{d}{dt}Q^a(t) = a - f^a(B(t), Q^a(t), Q^b(t)), \tag{3.3a}$$

$$Q^b(0) = q^b, \ \frac{d}{dt}Q^b(t) = B(t) - f^b(B(t), Q^a(t), Q^b(t)). \tag{3.3b}$$

One can check that, with the discharged rates defined in (3.2), $Q^a(t)$ and $Q^b(t)$ are continuous in t; thus $Q(t) = Q^a(t) + (h/H)Q^b(t)$ is also continuous in t.

We can also use the *infinitesimal generator* to represent the stochastic dynamics of the PDQ. Since $\{B(t); t \geq 0\}$ is a stationary two-state Markov process and since $Q(t)$ is continuous in t, the PDQ under proportional priority is right-continuous with left limits (RCLL, see [9]). Hence, by [29], the infinitesimal generator of the PDQ under proportional priority can be written in operator form as follows:

$$\begin{aligned}
&\mathcal{L}g(b, q^a, q^b) \\
&= \Big(a - f^a(b, q^a, q^b)\Big)\frac{\partial g}{\partial q^a} + \Big(b - f^b(b, q^a, q^b)\Big)\frac{\partial g}{\partial q^b} \\
&\quad + \mathbf{1}_{\{b=0\}}\lambda\Big(g(v/h, q^a, q^b) - g(0, q^a, q^b)\Big) \\
&\quad + \mathbf{1}_{\{b=v/h\}}\mu\Big(g(0, q^a, q^b) - g(v/h, q^a, q^b)\Big),
\end{aligned} \tag{3.4}$$

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

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

$$\limsup_{t\to\infty} \frac{1}{t}\int_0^t \mathsf{E}\Big[\exp\Big(Q^a(s) + (h/H)Q^b(s)\Big)\Big]ds \leq C. \tag{3.5}$$

This notion of stability is in line with that considered by Dai and Meyn for PDQs [28]. Essentially, it captures the boundedness of moments of queue lengths.

We are also interested in the steady-state joint distribution of $(B(t), Q^a(t), Q^b(t))$, called the *invariant probability measure*, denoted by π_{prop}. This measure is defined on the hybrid space $\mathcal{B} \times \mathbb{R}^2_{\geq 0}$. In general, boundedness of moments does not ensure convergence towards a unique invariant probability measure [28]. However, we will show while proving Theorem 3.1 that a stable PDQ necessarily converges to a unique invariant probability measure.

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

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

We derive $\overline{q}_{\text{prop}}$ and σ^2_{prop} in Section 3.2. Based on properties of the effective queue length, we will also derive bounds on the actual queue length $Q^a(t)+Q^b(t)$.

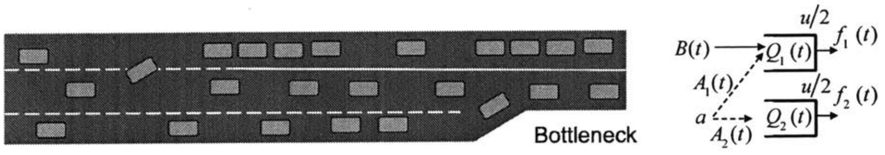
Furthermore, we define the *throughput under proportional priority*, denoted by J_{prop}, as follows:

$$J_{\text{prop}} = \sup\{a+\overline{B} : (3.5) \text{ holds}\}. \tag{3.6}$$

i.e. the supremum of the set of average aggregate arrival rates $a+\overline{B}$ such that the effective queue is stable; see (3.1) for the definition of $\overline{B}$.

Queuing dynamics: segmented priority

This priority rule is motivated by the idea of segmenting ordinary and connected vehicles and prioritizing connected vehicles in certain lanes [10]. For ease of presentation, we consider a highway bottleneck with two identical lanes; see Figure 3-4(a). Since



(a) A bottleneck with segmented priority.

(b) PDQ model under the segmented priority rule.

Figure 3-4: Relation between queue length and fraction of connected vehicles.

the total capacity of the bottleneck is u, each lane has a capacity of $u/2$. The two traffic classes travel through the bottleneck as follows. When no connected vehicles are arriving, i.e. when $B(t) = 0$, ordinary vehicles are evenly distributed over two lanes; that is, background traffic enters each lane at rate $a/2$. When $B(t) = v/h$, ordinary vehicles are restricted to one lane (server 2); the other lane (server 1) is dedicated to platoons. Note that in this setting lane changes are not allowed at the bottleneck.

Under the above priority rule, we can model the bottleneck as two parallel servers as shown in Figure 3-4(b). Let $A_k(t)$ be the rate at which the background traffic enters the k-th server. The segmented priority rule leads to the following:

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

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

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

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

The discharge rates are given by

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

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

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

$$\begin{aligned}
Q_1(0) &= q_1, \quad \frac{d}{dt}Q_1(t) = A_1(t) + \frac{h}{H}B(t) - f_1(B(t), Q(t)), \\
Q_2(0) &= q_2, \quad \frac{d}{dt}Q_2(t) = A_1(t) - f_2(B(t), Q(t)).
\end{aligned}$$

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

$$a < u, \quad v/H \leq u/2. \tag{3.7}$$

The first assumption is a trivial necessary condition for stability. The second assumption essentially ensures that vehicle platoons are always in free flow if not interacting with the background traffic. This assumption is typically satisfied by highway traffic, since the capacity of a highway lane ($u/2$ in this case) is equal to the quotient between free-flow speed v and minimal free-flow spacing H [24].

Assuming that (3.7) holds implies that the inflow to server 1 is always less than the capacity of server 1; hence $Q_1(t)$ vanishes. Therefore, we only need to consider $Q_2(t)$ for steady-state analysis. Note that server 2 is essentially a single-class fluid queuing system, since no platoons enter server 2. Hence, $Q_2(t) = Q_2^a(t)$.

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

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

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

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

We will compute $\overline{q}_{\text{seg}}$, σ^2_{seg}, in Section 3.2.

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

3.2 Stability analysis of fluid queuing model

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

3.2.1 Sufficient condition for bounded queue

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

Theorem 3.1 (Stability under proportional priority). *The two-class fluid queuing system is stable under proportional priority if*

$$a + \frac{\lambda}{\lambda + \mu}\frac{v}{H} < u. \tag{3.8}$$

Furthermore, if (3.8) *holds, then, for any initial condition* $(b, q_a, q_b) \in \mathcal{B} \times \mathbb{R}^2_{\geq 0}$, *the joint distribution of the hybrid state* $(B(t), Q_a(t), Q_b(t))$, *denoted by* $P_t(b, q_a, q_b)$, *converges to a unique probability measure* π_{prop}, *i.e.*

$$\lim_{t\to\infty} \|P_t(b, q_a, q_b) - \pi_{\text{prop}}\|_{\text{TV}} = 0, \tag{3.9}$$

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

Proof. The proof of the boundedness of moments (in the sense of (3.5)) is based on a Foster-Lyapunov-type criterion introduced by Meyn and Tweedie [74, Theorem 4.3], which we recall in our setting as follows: if there exist constants $c > 0$ and $d < \infty$, and a norm-like function[1] $V : \mathcal{B} \times \mathbb{R}^2_{\geq 0} \to \mathbb{R}$, such that

$$\mathcal{L}V(b, q_a, q_b) \leq -cV(b, q_a, q_b) + d, \quad \forall (b, q_a, q_b) \in \mathcal{B} \times \mathbb{R}^2_{\geq 0}, \tag{3.10}$$

then the PDQ model is stable in the sense of (3.5). Next, we prescribe the function V and explicitly construct the constants c and d.

Suppose that (3.8) holds. Let us consider the switched exponential Lyapunov function

$$V(b, q_a, q_b) = \begin{cases} k_0 e^{\gamma(q_a + (h/H) q_b)}, & b = 0, \\ k_1 e^{\gamma(q_a + (h/H) q_b)}, & b = v/h. \end{cases} \tag{3.11}$$

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

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

which are positive under (3.8); otherwise, we let

$$\gamma = \frac{(\lambda + \mu)(u - a - \frac{\lambda}{\lambda+\mu}\frac{v}{H})}{2(u-a)(a + \frac{v}{H} - u)}, \tag{3.12a}$$

$$k_0 = \frac{\gamma(a + \frac{v}{H} - u) + \lambda + \mu}{\gamma((\lambda+\mu)(u - a - \frac{\lambda}{\lambda+\mu}\frac{v}{H}) - \gamma(u-a)(a + \frac{v}{H} - u))}, \tag{3.12b}$$

$$k_1 = \frac{\gamma(a - u) + \lambda + \mu}{\gamma((\lambda+\mu)(u - a - \frac{\lambda}{\lambda+\mu}\frac{v}{H}) - \gamma(u-a)(a + \frac{v}{H} - u))}. \tag{3.12c}$$

which are also positive under (3.8) and $a + v/H > u$. In addition, we construct the constants c and d as follows:

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

Next, we verify (3.10) with V, c, and d as constructed above. Note that, for

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

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

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

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

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

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

Finally, since we have verified (3.10), we can apply [74, Theorem 4.3] and obtain (3.5).

To obtain (3.9), i.e. the convergence towards a unique invariant probability measure π_{prop}, note that, under (3.8), we have $a < u$. Hence, the aggregate traffic queue necessarily decreases when $B(t) = 0$. Therefore, for any initial condition, there is a strictly positive probability that $Q_a(t) = Q_b(t) = 0$ for a sufficiently large t. That is, the state $(0,0,0) \in \mathcal{B} \times \mathbb{R}^2_{\geq 0}$ can be attained with positive probability. Then, one can adapt the proof of [9, Theorem 4.6] and obtain the convergence to a unique invariant probability measure (in the sense of total variation distance). For details of this argument, we refer readers to [57, 46].

□

3.2.2 Steady-state queue length

For a stable PDQ model, we can also study the queue length:

Proposition 3.1. *For the PDQ model under proportional priority, if* (3.8) *holds, the steady-state effective queue length* $\overline{q}_{\text{prop}}$ *and variance* σ^2_{prop} *can be analytically* *expressed as follows:*

$$\overline{q}_{\text{prop}} = \begin{cases} 0, & a + \frac{v}{H} < u, \\ \frac{\lambda}{(\lambda+\mu)^2} \frac{a+\frac{v}{H}-u}{u-a-\frac{\lambda}{\lambda+\mu}\frac{v}{H}} \frac{v}{H}, & o.w. \end{cases} \tag{3.13a}$$

$$\sigma^2_{\text{prop}} = \begin{cases} 0, & a + \frac{v}{H} < u, \\ \frac{\lambda}{(\lambda+\mu)^3} \frac{(a+\frac{v}{H}-u)(u-a)}{\left(u-a-\frac{\lambda}{\lambda+\mu}\frac{v}{H}\right)^2} \frac{v}{H}, & o.w. \end{cases} \tag{3.13b}$$

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

$$\overline{q}_{\text{prop}} \leq \tilde{q} \leq \left(\frac{1}{1+\theta} + \frac{\theta}{1+\theta}\frac{H}{h}\right)\overline{q}_{\text{prop}},$$

$$\sigma^2_{\text{prop}} \leq \tilde{\sigma}^2 \leq \left(\frac{1}{1+\theta} + \frac{\theta}{1+\theta}\frac{H}{h}\right)^2 \sigma^2_{\text{prop}},$$

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

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

Lemma 3.1. *Under proportional priority, the following set*

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

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

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

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

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

This lemma can be proved by utilizing properties of the queuing dynamics (3.3). We omit the proof here due to space limitations. Figure 3-5 illustrates the basic

[2] See [9] for details regarding invariant sets for PDMPs.

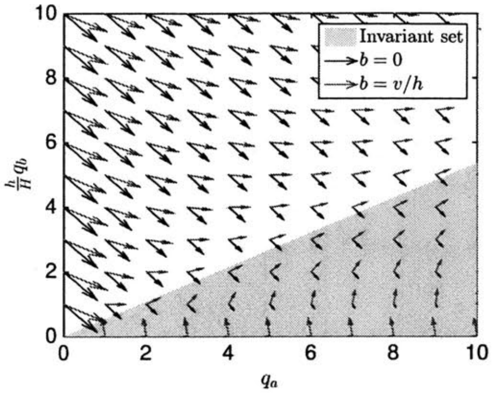


Figure 3-5: Illustration of the queuing dynamics and the invariant set $\mathcal{Q}_{\text{inv}}$ under proportional priority. The arrows represent the vectors of time-derivatives defined in (3.3) for both $b = 0$ and for $b = v/h$.

intuition behind this result. The proof entails that, for any $b \in \mathcal{B}$ and for any $[q_a, q_b]^T$ such that $[q_a, q_b]^T \notin \mathcal{Q}_{\text{inv}}$, the vector of time-derivatives of the queue lengths has a non-zero component that points to the interior of the invariant set $\mathcal{Q}_{\text{inv}}$.

Proof of Proposition 3.1. Average effective queue lengths and variance: Kulkarni gives an analytical expression for the steady-state distribution of the queue length in a single-class PDQ model that switches between a finite number of modes [57, Theorem 11.6]. In the particular setting of this proposition, the steady-state joint distribution of (b, q) can be represented as a probability density function (pdf) as follows:

$$\mathsf{f}(b, q) = \begin{cases} \mathsf{z}\delta_0 + \alpha_1 e^{-q/\beta}, & b = 0, \\ \alpha_2 e^{-q/\beta}, & b = v/H, \end{cases} \tag{3.14}$$

where

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

and δ_0 is the Dirac delta function centered at 0. Hence, we can obtain the expected value $\overline{q}_{\text{prop}}$ and variance σ^2_{prop} of the effective queue q, which are given by (3.13a) and

(3.13b), respectively.

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

Upper bounds for the actual queue length: Recall the invariant set $\mathcal{Q}_{\text{inv}}$ from Lemma 5.1. For each $(q_a, q_b) \in \mathcal{Q}_{\text{inv}}$, since $(h/H)q_b \leq \theta q_a$, we have

$$\frac{1+\theta}{\theta}\frac{h}{H}q_b \leq q_a + \frac{h}{H}q_b. \tag{3.15}$$

Then,

$$\begin{aligned} q_a + q_b &= q_a + (H/h)\frac{h}{H}q_b = q_a + \frac{h}{H}q_b + (H/h-1)\frac{h}{H}q_b \\ &\overset{(3.15)}{\leq} (q_a + \frac{h}{H}q_b) + (H/h-1)\frac{\theta}{1+\theta}(q_a + \frac{h}{H}q_b) \\ &= \left(\frac{1}{1+\theta} + \frac{\theta}{1+\theta}\frac{H}{h}\right)(q_a + \frac{h}{H}q_b). \end{aligned} \tag{3.16}$$

Since the set $\mathcal{Q}_{\text{inv}}$ is globally attracting and positively invariant, the invariant probability measure π_{prop} vanishes outside $\mathcal{Q}_{\text{inv}}$ [9]. Therefore,

$$\begin{aligned} \tilde{q} &= \int_{\mathcal{B}\times\mathbb{R}^2_{\geq 0}} (q_a + q_b)d\pi_{\text{prop}} = \int_{\mathcal{B}\times\mathcal{Q}_{\text{inv}}} (q_a + q_b)d\pi_{\text{prop}} \\ &\overset{(3.16)}{\leq} \left(\frac{1}{1+\theta} + \frac{\theta}{1+\theta}\frac{H}{h}\right)\int_{\{0,v/h\}\times\mathbb{R}^2_{\geq 0}} (q_a + \frac{h}{H}q_b)d\pi_{\text{prop}} \\ &= \left(\frac{1}{1+\theta} + \frac{\theta}{1+\theta}\frac{H}{h}\right)\overline{q}; \end{aligned}$$

the last equality results from the fact that π_{prop} gives the same average value of $(q_a + \frac{h}{H}q_b)$ as the pdf in (3.14) does. The upper bound for variance the variance σ^2_{prop} of the actual queue can be similarly obtained. □

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

Proposition 3.2 (Segmented priority). *Consider the two-class fluid queuing model* *and assume that* (3.7) *holds. Then the model is stable if*

$$\frac{\lambda+\mu/2}{\lambda+\mu}a < u/2. \tag{3.17}$$

Furthermore, under (3.17)*, the average and variance of queue length are given by*

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

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

Proof. Note that, under (3.7), the set $\{(q_1^a, q_1^b, q_2^a, q_2^b) \in \mathbb{R}_{\geq 0}^4 : q_1^a = q_1^b = q_2^b = 0\}$ is globally attracting and positively invariant under the segmented priority; i.e. $Q_2^a(t)$ could be arbitrarily large, but $Q_1^a(t)$, $Q_1^b(t)$, and $Q_2^b(t)$ necessarily vanish after sufficiently long time. Hence, we only need to consider the queue $Q_2^a(t)$. Note that Server 2 can be viewed as a single-class PDQ model. Thus, the rest of the proof is analogous to that of Theorem 3.1. □

3.3 Performance analysis of platooning operations

We are now ready to discuss how characteristics of platoons (specifically, penetration rate of connected vehicles, vehicle spacing within platoons, platoon length, and priority rule) affect traffic queue. Table 3.1 lists the nominal values considered in this section.

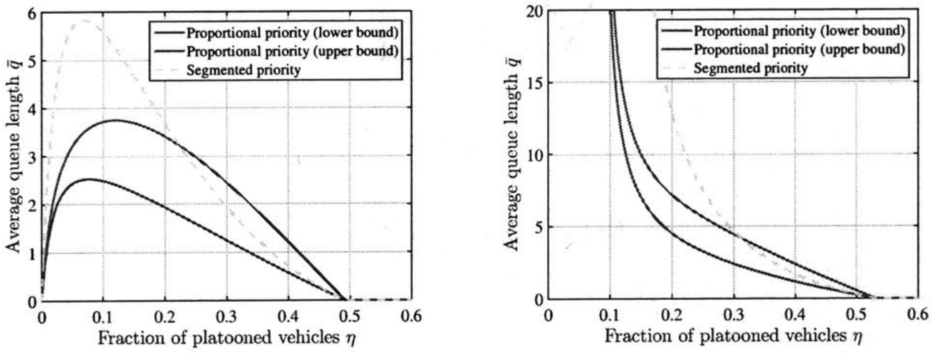
3.3.1 Fraction of platooned vehicles

The fraction of platooned vehicles can be written as

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

Table 3.1: Nominal parameters of traffic flow and platoons.

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



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

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

Figure 3-6: Impact of fraction of platooned vehicles on (actual) queue length.

where $\overline{B}$ is the average inflow of connected vehicles given by (3.1). Suppose that we fix the aggregate average demand $a+\overline{B}$, the platoon lengths μ, and the space h, and vary λ (or equivalently η). Figure 3-6 shows how the (bounds of) queue length vary with the fraction of platooned vehicles.

When the average aggregate demand $a+\overline{B}$ is smaller than the capacity u, this relation is characterized by a cap-shaped curve (Figure 3-6(a)). The points worth noting are: (i) at a low fraction, platooning increases the randomness of the arrival process, and thus increases the traffic queue, and (ii) as the fraction increases further, the gain of the reduced within-platoon spacing compensates for the increase in randomness of the arrival process. From a practical perspective, the inefficient fraction of platooned vehicles (≈ 0.1 in this example) should be avoided to limit the effect of random platoon arrivals. Furthermore, there exists a threshold η_0 beyond which no queue exists:

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

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

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

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

When the average aggregate demand is greater than the capacity (Figure 3-6(b)), the $\overline{q} - \eta$ curve has an elbow-shaped shape. In this case, note that, to ensure stability, at least a certain fraction of the total demand should be connected vehicles such that the excessive demand is compensated by the reduced spacing between platooned vehicles. This threshold, η_1, can be obtained from Theorem 1:

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

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

3.3.2 Intra-platoon spacing

Now we study the benefit of reducing the intra-platoon spacing. Current technology enables reduction of inter-vehicle spacing by 50% or more [2]. Suppose that we fix the aggregate average demand $a + \overline{B}$ and vary h. For the queue to be stable, the spacing should not exceed the following threshold:

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

Figure 3-7 shows how the queue varies with the ratio H/h when the average aggregate demand is greater than the capacity, i.e. $a + \overline{B} > u$. As expected, queue

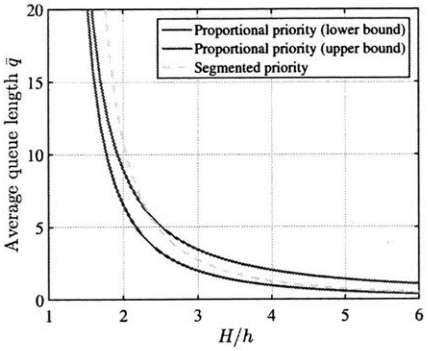


Figure 3-7: Impact of intra-platoon spacing on queue length.

length decreases as H/h increases. In addition, the curve becomes shallow as the ratio increases, implying that an excessively high ratio (more than 3 in Figure 3-7) does not bring much additional benefit. Note that high H/h ratios are not recommended for safety considerations either [2].

Arrival frequency and lengths of platoons

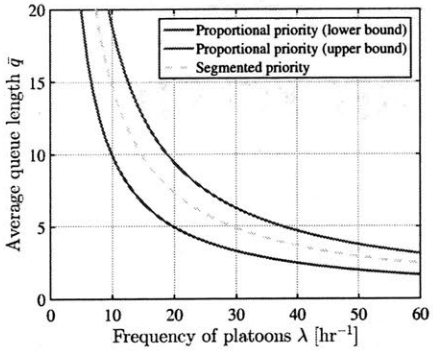


Figure 3-8: Impact of platoon arrival frequency on queue length.

Another question of practical interest is whether connected vehicles should form a large number of short platoons or a small number of long platoons. Platoon lengths affect fuel consumption and the ease of implementation [2]. Here, we focus on how average platoon length affects the traffic queue. Suppose that we fix the ratio between λ and μ, and vary λ. That is, we fix the fraction of platooned vehicles η, but vary the frequency and lengths of the platoons. Figure 3-8 shows that higher frequencies lead to smaller queues. The reason is that, as the platoons become more frequent and shorter, the probability of forming a long queue decreases. A practical interpretation in the setting illustrated in Figure 3-1 is that it is more difficult for long platoons to go through the bottleneck than short ones.

3.3.3 Priority rule

In Figures 3-6, 3-7, and 3-8, the queue lengths resulting from segmented priority are also plotted. Figure 3-6(a) implies that, with a low fraction of platooning, proportional priority leads to smaller traffic queues. This is intuitive in that prioritization of platooned vehicles under-utilizes the road's capacity if the fraction η is low. However, as the fraction increases (say greater than 0.4 in the figure), the queue length associated with segmented priority approaches the lower bound of that associated with proportional priority. In addition, Figure 3-7 implies that the relative benefit of segmenting two classes of traffic increases as the intra-platoon spacing decreases. Figure 3-8 implies that the relative benefit of segmenting does not significantly vary with the transition rates. However, in all the above-mentioned figures, the queue lengths associated with segmented priority are never below the lower bounds associated with proportional priority. Therefore, segmented priority is not guaranteed to outperform proportional priority, at least in the setting being considered here. In a broader range of settings, segmented priority may outperform proportional priority when the ratio H/h is very high, i.e. when the intra-platoon spacing is very short.

Finally, we can obtain from Theorem 3.1 that the throughput (as defined in (3.6)) under proportional priority is

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

That is, throughput increases with the fraction of connected vehicles. Similarly, we can obtain from Proposition 3.2 that the throughput under segmented priority is

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

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

$$\begin{aligned} J_{\text{prop}} &> J_{\text{seg}}, \quad \text{if } \eta > \frac{\frac{\lambda}{\lambda+\mu}}{\frac{\lambda}{\lambda+\mu}+\frac{h}{H}}; \\ J_{\text{prop}} &< J_{\text{seg}}, \quad \text{if } \eta < \frac{\frac{\lambda}{\lambda+\mu}}{\frac{\lambda}{\lambda+\mu}+\frac{h}{H}}. \end{aligned}$$

That is if the fraction of connected vehicles is high, then segmented priority leads to a smaller throughput. The intuition is that, in such a scenario, one lane (server 1 in Figure 3-4(b)) is not sufficient to serve the platoons, while the other lane (server 2) is under-utilized.

3.4 Summary

In this chapter, we proposed a two-class fluid queuing model to study the traffic congestion induced by vehicle platooning at highway bottlenecks. Using this model, we are able to evaluate the impact of parameters of vehicle platoons and the priority rule on traffic congestion and throughput. As we have argued at the beginning of this chapter, our model considers exponentially distributed platoon lengths, which is likely to be an over-approximation. Consequently, the stability analysis and control design are thus likely conservative. One way of estimating the conservativeness is to simulate the platooning-induced queues for alternative distributions of platoon lengths and compare with our results.

This work is being extended in several directions. First, to consider the impact of congestion downstream to a bottleneck, tandem PDQ models with finite buffers can be considered. Known results [57] imply that, for PDQ models with finite buffers, average platoon length affects not only queue length, but also stability. Second, our approach can be used to study control of platoons in response to local traffic conditions, such as time-varying demand of background traffic and road capacity perturbations. Of particular interest is the tradeoff between throughput gain and fuel savings. More information about our ongoing work is available in Section 6.2.

Chapter 4

Performance Analysis of Highways Facing Perturbations

Freeway traffic networks are prone to capacity disruptions, for example, crashes, road blockage, and other capacity-reducing incidents [50, 55, 61, 89]. In practice, these events can introduce significant congestion in freeways [62, 88]. To design traffic control strategies that reduce the congestion and throughput loss resulting from such disruptions, we need to systematically analyze traffic dynamics under stochastic capacity fluctuations. This chapter introduces a stochastic switching model of freeway traffic dynamics under capacity perturbations, and studies its stability (in the sense of bounded traffic queue) under fixed inflows.

Perturbations (incidents) on a multi-cell freeway are modeled by reduction in capacity at the affected freeway sections, which occur and clear according to a Markov chain. We develop conditions under which the traffic queue induced by stochastic incidents is bounded. A necessary condition is that the demand must not exceed the time-average capacity adjusted for spillback. A sufficient condition, in the form of a set of bilinear inequalities, is also established by constructing a Lyapunov function and applying the classical Foster-Lyapunov drift condition. Both conditions can be easily verified for realistic instances of the stochastic incident model. Our analysis relies on the construction of a globally attracting invariant set, and exploits the properties of the traffic flow dynamics. We use our results to analyze the impact of stochastic capacity fluctuation (frequency, intensity, and spatial correlation) on the throughput of a freeway segment.

In our SS-CTM (formally defined in Section 4.1), the capacity of a freeway section switches between a finite set of values (*modes*); the switches are governed by a continuous-time finite-state Markov chain. The main results of this chapter (Theorems 4.1 and 4.2; presented in Section 4.2) include a necessary condition and a sufficient condition for stability of the SS-CTM with fixed inflows and an ergodic mode transition process. Proofs of the main results are provided in Section 4.3. In Section 4.4, we provide examples to show how Theorems 4.1 and 4.2 can be used to characterize the set of stabilizing inflow vectors.

4.1 Stochastic Switching Cell Transmission Model

In this section, we define the stochastic switching cell transmission model (SS-CTM). To develop this model, we introduce a Markovian capacity model, and combine it with the classical CTM [24]. We also introduce key definitions that are needed for our subsequent analysis.

4.1.1 Markovian capacity model

Consider a freeway consisting of K *cells*, as shown in Figure 4-1. The *capacity* (or saturation rate, in vehicles per hour, or veh/hr) of the k-th cell at time $t \geq 0$ is denoted by $F_k(t)$. Let $F(t) = [F_1(t), \dots, F_K(t)]^T$ denote the vector of cell capacities at time t. One can interpret $F_k(t)$ as the maximum rate at which cell k can discharge traffic to the downstream cell at time t.

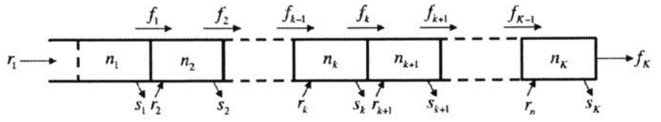


Figure 4-1: SS-CTM with K cells. Cell 1 includes an infinite-sized buffer to accommodate the upstream queue.

To model stochastic capacity disruptions, we assume that $F(t)$ is a finite-state

Markov process. Specifically, let $\mathcal{I}$ be a finite set of *modes* of the freeway and let $m = |\mathcal{I}|$. Each mode $i \in \mathcal{I}$ is associated with a vector of cell capacities $F^i = [F^i_1, \dots, F^i_K]^T$. We define

$$F^{\min}_k = \min_{i \in \mathcal{I}} F^i_k, \quad F^{\max}_k = \max_{i \in \mathcal{I}} F^i_k, \tag{4.1}$$

and refer to $F^{\max}_k$ as the *normal* (maximum) *capacity* of cell k. For ease of presentation, we assume an identical normal capacity for all cells throughout the chapter, i.e. $F^{\max}_k = F^{\max}$ for $k = 1, 2, \dots, K$.

In our model, a mode represents a particular configuration of capacities at various locations (cells). We say that the freeway is in the *normal mode* if the maximum capacity is available at every cell. We model an incident in cell k by introducing a mode i such that $F^i_k < \mathsf{F}$ and $F^i_h = \mathsf{F}$ for $h \neq k$. Transition from the normal mode to mode i can be viewed as occurrence of an incident in cell k; similarly, the transition from i to the normal mode can be viewed as clearance of the incident. Furthermore, we call the k-th cell an *incident hotspot* if $F^{\min}_k < F^{\max}$.

Note that the Markovian capacity model can be used to represent more complex situations. For example, two modes can be associated with incidents in the same cell(s), but with different values of capacities, reflecting the difference in incident intensities (e.g. minor and major). Furthermore, the occurrence of secondary (or induced) incidents [54] can be modeled as a transition from a mode with an incident in a single cell to a mode with incidents in multiple cells.

Throughout this chapter, we use i to denote elements of $\mathcal{I}$, and use $I(t)$ to denote the stochastic mode of the freeway at time t. The mode $I(t)$ switches according to a continuous-time, finite-state Markov chain defined over the set $\mathcal{I}$ with (time-invariant) transition rates $\{\lambda_{ij}; i, j \in \mathcal{I}\}$. We assume that $\lambda_{ii} = 0$ for each $i \in \mathcal{I}$; note that this is without loss of generality, as inclusion of self-transitions would not affect the traffic flow dynamics. Let $\nu_i = \sum_{j \in \mathcal{I}} \lambda_{ij}$ and define the *transition matrix* as follows:

$$\Lambda = \begin{bmatrix} -\nu_1 & \lambda_{12} & \dots & \lambda_{1m} \\ \lambda_{21} & -\nu_2 & \dots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{m1} & \lambda_{m2} & \dots & -\nu_m \end{bmatrix}. \tag{4.2}$$

We assume the following for the mode switching process:

Assumption 3. *The finite-state Markov process* $\{I(t); t \geq 0\}$ *is* ergodic.

This assumption ensures that the dwell times in each mode are finite almost surely (a.s.). Under this assumption, the process $\{I(t); t \geq 0\}$ admits a unique steady-state probability distribution $\mathsf{p} = [\mathsf{p}_1, \dots, \mathsf{p}_m]$ satisfying:

$$\mathsf{p}\Lambda = 0, \ |\mathsf{p}| = 1, \ \mathsf{p} \geq 0, \tag{4.3}$$

where $|\cdot|$ indicates the 1-norm of (row or column) vectors [33].

4.1.2 Traffic flow under stochastic capacities

The formal definition of the SS-CTM is as follows:

Definition 4.1. *The SS-CTM is a tuple* $\langle \mathcal{I}, \mathcal{N}, \mathcal{R}, \Lambda, G \rangle$*, where*

- $\mathcal{I}$ *is a finite set of modes (discrete state space) with* $|\mathcal{I}| = m$*,*

- $\mathcal{N} = [0, \infty) \times [0, n_k^{\max}]^{K-1}$ *is the set of traffic densities (continuous state space),*

- $\mathcal{R} \subseteq \mathbb{R}^K_{\geq 0}$ *is the set of inflow vectors,*

- $\Lambda \in \mathbb{R}^{m \times m}$ *is the transition rate matrix governing the mode transitions, and*

- $G : \mathcal{I} \times \mathcal{N} \times \mathcal{R} \to \mathbb{R}^K$ *is the vector field governing the continuous dynamics.*

We have defined $\mathcal{I}$ and Λ, and now introduce $\mathcal{N}$, $\mathcal{R}$, and G.

Let $N_k(t)$ denote the *traffic density* (in vehicles per mile, veh/mi) in the k-th cell at time t, as shown in Figure 4-1. Traffic density $N_k(t)$ is non-negative and upper bounded by $n_k^{\max}$, the k-th cell's *jam density.* The K-dimensional vector $N(t) = [N_1(t), N_2(t), \dots, N_K(t)]^T \in \mathcal{N}$ represents the stochastic continuous state of the SS-CTM.

For ease of presentation, we assume that each cell k has the unit length of 1 mi. Furthermore, each cell k has a free-flow speed α_k, a congestion-wave speed β_k, a jam density $n_k^{\max}$, and a normal capacity F_k. The unit of α and β is miles per hour (mi/hr). We define the *critical density* of vehicles as

$$n_k^{\text{crit}} = \frac{\mathsf{F}_k}{\alpha_k}. \tag{4.4}$$

The *sending flows* S_k and the *receiving flows* R_k can be written as follows:

$$S_k(i, n_k) = \min\left\{\alpha_k n_k, F_k^i\right\}, \ k = 1, 2, \dots, K, \tag{4.5a}$$

$$R_k(n_k) = \beta_k(n_k^{\max} - n_k), \ k = 2, 3, \dots, K. \tag{4.5b}$$

Thus, S_k is the traffic flow that cell k can discharge downstream and R_k is the traffic flow from upstream that cell k can accommodate. The receiving flow of cell 1 will be discussed later in this subsection. Following [24], we assume that,

$$\forall k \in \{1, 2, \dots, K\}, \ \max_{i \in I} S_k(i, n_k^{\text{crit}}) \leq R_k(n_k^{\text{crit}}),$$

which, along with (4.4) and (4.5), implies that,

$$\mathsf{F}_k \leq \frac{\alpha_k \beta_k}{\alpha_k + \beta_k} n_k^{\max}. \tag{4.6}$$

Let $r = [r_1, r_2, \dots, r_K]^T \in \mathcal{R} = \mathbb{R}_{\geq 0}^K$ denote the *inflow vector* to the freeway; the unit of r_k is veh/hr. Throughout this chapter, we assume that the freeway is subject to a fixed (i.e. time-invariant) inflow vector. Importantly, we also make the standard assumption that, for each cell k, the on-ramp flow r_k is prioritized over the sending inflow S_{k-1} from the upstream cell [37]. Under this priority rule, the flow (in veh/hr) from cell k to cell $k+1$, denoted by f_k, is given by the flow function (i.e., the so-called *fundamental diagram*):

$$f_0 = 0, \tag{4.7a}$$

$$\begin{aligned} f_k(i, n_k, n_{k+1}, r_{k+1}) = \min\{&\rho_k S_k(i, n_k), \\ &(R_{k+1}(n_{k+1}) - r_{k+1})_+\}, \; k = 1, 2, \ldots, K-1, \end{aligned} \tag{4.7b}$$

$$f_K(i, n_K) = \rho_K S_K(i, n_K). \tag{4.7c}$$

where $(\cdot)_+$ stands for the positive part and $\rho_k = f_k/(f_k + s_k) \in (0, 1]$ denotes the fixed *mainline ratio*, i.e. the fraction of traffic from cell k entering cell $k+1$. The *off-ramp flow* s_k from cell k is given by $s_k(t) = (1/\rho_k - 1)f_k(t)$ for $k = 1, 2, \ldots, K$.

Let $f(i, n, r)$ denote the K-dimensional vector of flows. For notational convenience, we denote $S_k(t) = S_k(I(t), N_k(t))$, $R_k(t) = R_k(N_k(t))$, and $f(t) = f(I(t), N(t), r)$. We say that cell k is experiencing *spillback* at time t if $\rho_k S_k(t) > R_{k+1}(t) - r_{k+1}$, i.e. if the sending flow from cell k exceeds the receiving flow of cell $k+1$.

Due to spillback, there might be traffic queues at the entrances (on-ramps) to the freeway. We track the queue upstream to cell 1 by assuming that cell 1 has a buffer with infinite space to admit this queue, i.e. $n_1^{\max} = \infty$ (see Figure 4-1). However, we do not consider the on-ramp queues (i.e. queues at on-ramps to cells $2, 3, \ldots, K$). Note that not including the on-ramp queues to cells 2 through K does not affect our stability analysis of the upstream queue, since our priority rule implies that the boundedness of the upstream queue is a sufficient condition for the boundedness of the on-ramp queues. Hence, we denote the *continuous state space* of the SS-CTM as $\mathcal{N} = [0, \infty) \times \prod_{k=2}^{K}[0, n_k^{\max}]$.

By mass conservation, traffic density in each cell evolves as follows [37]:

$$\dot{N}_k(t) = f_{k-1}(t) + r_k - f_k(t)/\rho_k, \; k = 1, 2, \ldots, K. \tag{4.8}$$

From (4.7) and (4.8), we can define the vector field $G : \mathcal{I} \times \mathcal{N} \times \mathcal{R} \to \mathbb{R}^n$ governing the continuous state of the SS-CTM as follows:

$$\begin{aligned}
G_1(i,n,r) &= r_1 - f_1(i,n_1,n_2,r_2)/\rho_1, && \text{(4.9a)}\\
G_k(i,n,r) &= f_{k-1}(i,n_{k-1},n_k,r_k) + r_k \\
&\quad - f_k(i,n_k,n_{k+1},r_{k+1})/\rho_k,\ k = 2,3,\dots,K-1, && \text{(4.9b)}\\
G_K(i,n,r) &= f_{K-1}(i,n_{K-1},n_K,r_K) + r_k - f_K(i,n_K)/\rho_K. && \text{(4.9c)}
\end{aligned}$$

Note that the vector field G is bounded and continuous in n. In a given mode $i \in \mathcal{I}$, the *integral curve* starting from $n \in \mathcal{N}$, denoted by $\phi_t^i(n) = [\phi_t^i(n)_1, \dots, \phi_t^i(n)_K]^T$, can be expressed as follows:

$$\phi_t^i(n) = n + \int_{\tau=0}^{t} G\Big(i, \phi_\tau^i(n), r\Big) d\tau. \tag{4.10}$$

The *hybrid state* of the SS-CTM is $(I(t), N(t))$ at time t, and the *hybrid state space* is $\mathcal{I} \times \mathcal{N}$. The evolution of the discrete (resp. continuous) state is governed by the finite-state Markov process with transition matrix Λ (resp. the vector field G). For an initial condition $(i,n) \in \mathcal{I} \times \mathcal{N}$, the stochastic process $\{(I(t), N(t)); t \geq 0\}$ is given by

$$\begin{aligned}
N(t) &= n + \int_{\tau=0}^{t} G\Big(I(\tau), N(\tau), r\Big) d\tau, && \text{(4.11a)}\\
\Pr\Big\{I(t+\delta) = j | I(t) = i\Big\} &= \lambda_{ij}\delta + \mathrm{o}(\delta),\ j \neq i. && \text{(4.11b)}
\end{aligned}$$

4.1.3 Additional definitions

For an SS-CTM with a given inflow vector $r \in \mathcal{R}$, the total number of vehicles $|N(t)|$ at time t is given by

$$|N(t)| = \sum_{k=1}^{K} N_k(t). \tag{4.12}$$

We say that the SS-CTM is *stable* if the moment generating function (MGF) of $N(t)$ is bounded on average; i.e., for some $\rho \in \mathbb{R}^K_{>0}$ and some $C > 0$, and for each initial condition $(i,n) \in \mathcal{I} \times \mathcal{N}$,

$$\limsup_{t\to\infty} \frac{1}{t}\int_{\tau=0}^{t} \mathsf{E}\left[\exp\left(\rho^T N(\tau)\right)\right] d\tau \leq C. \tag{4.13}$$

Since a bounded MGF implies a bounded p-th moment for all $p \in \mathbb{Z}_{>0}$, our notion of stability is in line with the notion of bounded moments considered by Dai and Meyn [28]. Recall that, in our model, $N_2(t), \ldots, N_K(t)$ are always upper-bounded by the jam density $n_k^{\max}$; therefore, $N(t)$ is bounded if and only if $N_1(t)$ is bounded.

An alternative notion of stability that is used in the analysis of queueing systems [28, 57, 46] and PDMPs [9, 19] is the convergence of the traffic queue towards a unique invariant probability measure. This notion of stability is equivalent to boundedness of the traffic queue in many simple settings (e.g. for $M/M/1$ queues [33] or fluid queueing systems with stochastic service rates [57, 46]). However, convergence to a unique invariant probability measure does not always guarantee bounded moments of the traffic queue, which is of practical significance for freeway traffic management. Therefore, in this chapter, we consider boundedness of the upstream queue as the stability notion of interest.

A major issue in analyzing the stability of the SS-CTM is to ensure (4.13) for all initial conditions $(i,n) \in \mathcal{I} \times \mathcal{N}$. We address this issue by constructing a positive invariant set that is also globally attracting and positively invariant [9], i.e. a set $\widetilde{\mathcal{N}} \subseteq \mathcal{N}$ such that

$$\text{(Invariant)} \quad \forall (i,n) \in \mathcal{I} \times \widetilde{\mathcal{N}},\ \forall t \geq 0,\ \phi_t^i(n) \in \widetilde{\mathcal{N}}; \tag{4.14a}$$

$$\text{(Attracting)} \quad \forall (i,n) \in \mathcal{I} \times \mathcal{N},\ I(0) = i, N(0) = n,$$

$$\forall \epsilon > 0,\ \exists T \geq 0,\ \forall t \geq T,\ \min_{\nu \in \widetilde{\mathcal{N}}} \|N(t) - \nu\|_2 \leq \epsilon. \tag{4.14b}$$

For convenience, we henceforth refer to any set satisfying (4.14a) and (4.14b) simply as an *invariant set.* Construction of an invariant set considerably simplifies the proofs of our main results (Theorems 4.1 and 4.2), since, to capture long-time behavior of SS-CTM, we only need to consider initial conditions in $\widetilde{\mathcal{N}}$ rather than in $\mathcal{N}$.

Before proceeding further, we introduce two properties of the SS-CTM. First, the natural filtration $\mathcal{F}_t$ of the SS-CTM is the σ-algebra generated by $\{(I(s), N(s)); s \leq t\}$ for all $t \geq 0$ [29]. Since the realizations of the continuous state are always continuous in time, and since the transition rates λ_{ij} are finite and constant, $\mathcal{F}_t$ is *right continuous with left limits* (RCLL, or *càdlàg* [9]). Second, by [9, Proposition 2.1], thanks to the RCLL property, the *infinitesimal generator* of the SS-CTM with a fixed inflow $r \in \mathcal{R}$ can be written as an operator $\mathcal{L}$ as follows:

$$\mathcal{L}g(i,n) = G^T(i,n,r)\nabla_n g(i,n) + \sum_{j\in\mathcal{I}} \lambda_{ij}\Big(g(j,n) - g(i,n)\Big), \qquad \forall (i,n) \in \mathcal{I} \times \mathcal{N}, \tag{4.15}$$

where $g : \mathcal{I} \times \mathcal{N} \to \mathbb{R}$ is a function smooth in the second argument, and $\nabla_n g(i,n)$ is the gradient of g with respect to n.[1] We utilize the expression of the infinitesimal generator in our stability analysis (while applying the Foster-Lyapunov drift condition [74] in Appendix C).

4.2 Stability of SS-CTM

In this section, we present our results and demonstrate their application via a simple example. The proofs for these results are available in Section 4.3.

4.2.1 Main results

Our results include a necessary condition (Theorem 4.1) and a sufficient condition (Theorem 4.2) for the stability of the SS-CTM under fixed inflows. Both of these conditions rely on the construction of a "rectangular" invariant set of the following form:

$$\widetilde{\mathcal{N}} = [\underline{n}_1, \infty) \times \prod_{k=2}^{K} [\underline{n}_k, \overline{n}_k]\,. \tag{4.16}$$

[1] We consider $\nabla_n g$ as a column vector.

Proposition 4.1. *For an SS-CTM with an inflow vector* $r \in \mathcal{R}$*, the set* $\widetilde{\mathcal{N}}$ *of the form in* (4.16) *is an invariant set in the sense of* (4.14) *with the interval boundaries specified as follows:*

$$\underline{n}_1 = \min\left\{\frac{r_1}{\alpha_1}, \frac{\mathsf{F}_1}{\alpha_1}\right\}, \tag{4.17a}$$

$$\underline{n}_k = \min\left\{\rho_{k-1}\underline{n}_{k-1} + \frac{r_k}{\alpha_k}, \frac{\rho_{k-1}F_{k-1}^{\min} + r_k}{\alpha_k}, \frac{\mathsf{F}_k}{\alpha_k}\right\}, \qquad k = 2, 3, \ldots, K, \tag{4.17b}$$

$$\overline{n}_K = \begin{cases} \dfrac{\rho_{K-1}\mathsf{F}_K + r_K}{\alpha_K}, & \textit{if } \rho_{K-1}\mathsf{F}_K + r_K \le F_K^{\min}, \\ n_K^{\max} - \dfrac{F_K^{\min}}{\beta_K}, & o.w. \end{cases} \tag{4.17c}$$

$$\overline{n}_k = \begin{cases} \dfrac{\beta_{k-1}\mathsf{F}_k + r_k}{\alpha_k}, & \textit{if } \beta_{k-1}\mathsf{F}_k + r_k \\ & \quad \le \min\left\{F_k^{\min}, \dfrac{(R_{k+1}(\overline{n}_{k+1}) - r_{k+1})_+}{\rho_k}\right\}, \\ n^{\max} - \dfrac{1}{\beta_k}\min\left\{F_k^{\min}, \dfrac{(R_{k+1}(\overline{n}_{k+1}) - r_{k+1})_+}{\rho_k}\right\}, & \\ & o.w., \end{cases} \qquad k = K-1, K-2, \ldots, 2, \tag{4.17d}$$

where S_k *and* R_k *are given by* (4.5).

The set $\widetilde{\mathcal{N}}$ is constructed by considering the properties of the sending and receiving flows (4.5). Specifically, for each cell k, the lower boundary $\underline{n}_k$ can be viewed as the limiting density when the flow f_{k-1} from upstream is at its minimum and when the flow f_k discharged to downstream is not constrained by the $(k+1)$-th cell's receiving flow. Thus, for each k and each $n \in \widetilde{\mathcal{N}}$ such that $n_k = \underline{n}_k$, we have $G_k(i, n, r) \ge 0$ in each mode $i \in \mathcal{I}$; i.e. the vector field points in the direction of non-decreasing cell density. Similarly, on the upper boundary of $\widetilde{\mathcal{N}}$, the vector field points in the direction of non-increasing cell density; i.e. for each $k \ge 2$ and each $n \in \widetilde{\mathcal{N}}$ such that $n_k = \overline{n}_k$, we have $G_k(i, n, r) \le 0$ in each mode $i \in \mathcal{I}$.

We choose this specific form of invariant set (i.e. Cartesian product of intervals) because of its simple representation [13]. Note that, for a given $r \in \mathcal{R}$, Proposition 4.1

only provides one such construction; indeed, other rectangular sets satisfying (4.14) exist. Importantly, this particular construction leads to intuitive and practically relevant conditions (Theorems 4.1 and 4.2) that can be used to identify sets of stabilizing and unstabilizing inflow vectors. In fact, the sharpness of our stability conditions is directly related to the properties of the invariant set; please refer to Proposition 4.2 at the end of this subsection.

Before introducing the necessary condition for stability, we need to define a new notion of *spillback-adjusted capacities*: for each $i \in \mathcal{I}$,

$$\widetilde{F}^i_k(\underline{n}, r) = \min\left\{F^i_k,\ \frac{1}{\beta_k}\left(R_{k+1}\left(\underline{n}_{k+1}\right) - r_{k+1}\right)_+\right\},$$

$$k = 1, 2, \ldots, K-1, \tag{4.18a}$$

$$\widetilde{F}^i_K(\underline{n}, r) = F^i_K, \tag{4.18b}$$

where $\underline{n}_k$'s are given by (4.17a) and (4.17b). Recalling (4.5) and (4.7) and noting that R_k is non-increasing in n_k, one can see that

$$\forall (i, n) \in \mathcal{I} \times \widetilde{\mathcal{N}},\ \forall r \in \mathcal{R},\ \forall t \geq 0,$$
$$f_k(t) \leq \rho_k \min\left\{vN_k(t), \widetilde{F}^{I(t)}_k\right\},\ k = 1, 2 \ldots, K. \tag{4.19}$$

Thus, $\widetilde{F}_k$ can be interpreted as the capacity adjusted for the receiving flow admissible by the downstream the $(k+1)$-th cell, and hence the name "spillback-adjusted". By considering $\widetilde{F}_k$, we do not need to explicitly involve the receiving flow in our necessary condition for stability.

In addition, we define the following parameters:

$$\rho^k_k = 1, \quad k = 1, \ldots, K, \tag{4.20a}$$

$$\rho^{k_2}_{k_1} = \prod_{h=k_1}^{k_2-1} \rho_h, \quad 1 \leq k_1 \leq k_2 - 1,\ k_2 = 2, \ldots, K. \tag{4.20b}$$

Note that $\rho^{k_2}_{k_1}$ can be viewed as the fraction of the inflow r_{k_1} that is routed to cell k_2.

Thus, for each k, we can view $\sum_{h=1}^{k}\rho_h^k r_h$ as the total *nominal flow* through cell k.

Then, we have the following result:

Theorem 4.1 (Necessary condition). *Consider an SS-CTM with an inflow vector* $r \in \mathcal{R}$. *Let* p_i *be the solution to* (4.3), $\widetilde{F}_k^i(\underline{n}, r)$ *be as defined in* (4.18), *and* $\rho_{k_1}^{k_2}$ *be as defined in* (4.20). *If the SS-CTM is stable in the sense of* (4.13), *then,*

$$\sum_{h=1}^{k}\rho_h^k r_h \leq \sum_{i\in\mathcal{I}}\mathsf{p}_i\widetilde{F}_k^i\left(\underline{n}, r\right), \quad k = 1, 2, \ldots, K. \tag{4.21}$$

The left-hand side of (4.21) is the nominal flow through cell k. The right-hand side of (4.21) can be viewed as the long-time average of the spillback-adjusted capacity. Thus, Theorem 4.1 necessitates that, for the SS-CTM to be stable, the nominal flow cannot exceed the average spillback-adjusted capacity. This result provides a simple criterion to check for the instability of SS-CTM for a given inflow vector $r \in \mathcal{R}$: if $\sum_{h=1}^{k}\rho_h^k r_h > \sum_{i\in\mathcal{I}}\mathsf{p}_i\widetilde{F}_k^i(\underline{n}, r)$ for some k, then the system is unstable.

An important implication of Theorem 4.1 is that the SS-CTM may be unstable even if, for each cell, the nominal flow is strictly less than the average capacity of the respective cell, i.e.

$$\sum_{h=1}^{k}\rho_h^k r_h < \sum_{i\in\mathcal{I}}\mathsf{p}_i F_k^i, \quad k = 1, 2, \ldots, K. \tag{4.22}$$

To see this, one can note that (4.22) does not guarantee (4.21), unless $\widetilde{F}_k^i(\underline{n}, r) = F_k^i$ for all k and all i, which holds only when the inflows are sufficiently small. In summary, our necessary condition imposes a restriction on the inflow vector that captures the joint effect of capacity fluctuation and spillback. Note that Theorem 4.1 only involves the steady state probabilities p_i but not the elements of of Λ directly.

To develop the sufficient condition, let us limit our attention to the set of inflow vectors satisfying (4.22). For each r satisfying (4.22), we define the vectors $\gamma =$

$[\gamma_1, \dots, \gamma_K]^T$ and $\Gamma = [\Gamma_1, \dots, \Gamma_K]^T$ as follows:

$$\gamma_k = \frac{\sum_{i\in\mathcal{I}} \mathsf{p}_i F_k^i}{\sum_{i\in\mathcal{I}} \mathsf{p}_i F_k^i - \sum_{h=1}^{K} \rho_h^k r_h}, \quad k = 1, 2, \dots, K, \tag{4.23a}$$

$$\Gamma_K = \gamma_K, \tag{4.23b}$$

$$\Gamma_k = \rho_k(\Gamma_{k+1} + \gamma_k), \quad k = K-1, K-2, \dots, 1. \tag{4.23c}$$

In our sufficient condition, we consider the sum of inflows weighted by Γ_k:

$$\mathscr{R}(r) = \Gamma^T r. \tag{4.24}$$

Essentially, Γ_k can be viewed as a weight assigned to the inflow or the traffic density in the k-th cell with the following properties: (i) upstream cells have higher weights; (ii) a cell's weight increases with the cell's inflow-capacity ratio.

In addition, we define the following sets

$$\Theta = \left\{n \in \mathcal{N} : n_1 = n^{\text{crit}}, n_k \in \{\underline{n}_k, \overline{n}_k\}, k = 2, 3, \dots, K\right\}, \tag{4.25a}$$

$$\hat{\Theta} = \left\{n \in \mathcal{N} : n_1 = \underline{n}_1, n_k \in \{\underline{n}_k, \overline{n}_k\}, k = 2, 3, \dots, K\right\}, \tag{4.25b}$$

where $\underline{n}_k$ and $\overline{n}_k$ are given by (4.17). Note that Θ and $\hat{\Theta}$ both have cardinality of 2^{K-1}, where K is the number of cells. Furthermore, let

$$\mathscr{F}_i(\underline{n}, \overline{n}, r) = \min_{n\in\Theta} \gamma^T f(i, n, r), \ i \in \mathcal{I}, \tag{4.26a}$$

$$\hat{\mathscr{F}}_i(\underline{n}, \overline{n}, r) = \min_{n\in\hat{\Theta}} \gamma^T f(i, n, r), \ i \in \mathcal{I}, \tag{4.26b}$$

where $f(i, n, r)$ is given by (4.7). Although (4.26) involves evaluating minima of $\gamma^T f$ over discrete sets. We note that, for typical freeway lengths (in the order of 10 cells), $\mathscr{F}_i(\underline{n}, \overline{n}, r)$ and $\hat{\mathscr{F}}_i(\underline{n}, \overline{n}, r)$ can be obtained by simple enumeration. As we will show in Appendix C, $\mathscr{F}_i(\underline{n}, \overline{n}, r)$ and $\hat{\mathscr{F}}_i(\underline{n}, \overline{n}, r)$ can be viewed as lower bounds on the weighted sum of the discharged flows f_k in mode i.

Then, we have the following result:

Theorem 4.2 (Sufficient condition). *Consider an SS-CTM with an inflow vector $r \in \mathcal{R}$ satisfying* (4.22). *Let $\underline{n}$ and $\overline{n}$ be as defined in* (4.17), *$\mathscr{R}(r)$ as defined in* (4.24), *and $\mathscr{F}_i(\underline{n}, \overline{n}, r)$ and $\hat{\mathscr{F}}_i(\underline{n}, \overline{n}, r)$ as defined in* (4.26).

If there exist positive constants a_1, $a_2, \ldots a_m$ and b such that

$$\forall i \in \mathcal{I},\ a_i b\Big(\mathscr{R}(r) - \mathscr{F}_i(\underline{n}, \overline{n}, r)\Big) + \sum_{j \in \mathcal{I}} \lambda_{ij}(a_j - a_i) \leq -1, \tag{4.27}$$

then, by defining

$$c = \frac{1}{\max_i a_i}, \tag{4.28a}$$

$$\begin{aligned} d &\\ &= \max_{i \in \mathcal{I}} \left| a_i b \left(\mathscr{R}(r) - \hat{\mathscr{F}}_i(\underline{n}, \overline{n}, r)\right) + \sum_{i \in \mathcal{I}} \lambda_{ij}(a_j - a_i) + a_i c \right| \\ &\quad \times \exp\Big(b(\Gamma_1 n_1^{\mathrm{crit}} + \Gamma_2 \overline{n}_2 + \cdots + \Gamma_K \overline{n}_K)\Big), \end{aligned} \tag{4.28b}$$

we obtain that, for each initial condition $(i, n) \in \mathcal{I} \times \mathcal{N}$,

$$\limsup_{t \to \infty} \frac{1}{t} \int_{\tau=0}^{t} \mathsf{E}\left[\exp\left(b\Gamma^T N(\tau)\right)\right] d\tau \leq \frac{d}{c \min_i a_i}.. \tag{4.29}$$

The bilinear inequalities (4.27) essentially restrict the weighted inflow $\mathscr{R}$, and thus restrict the inflow vector r to ensure stability. The first term on the left-hand side of (4.27) captures the difference between the (weighted) inflow and the (weighted) discharged flows; the second term captures the effect of stochastic mode transitions. Note that, unlike Theorem 4.1, Theorem 4.2 explicitly involves the elements of Λ.

Theorem 4.2 is derived based on an approach introduced by Meyn and Tweedie [74]. For readers' convenience, we state the relevant result [74, Theorem 4.3] as follows. Recall that the SS-CTM is RCLL and its infinitesimal generator $\mathcal{L}$ is given by (4.15). Suppose that there exists a norm-like[2] function $V : \mathcal{I} \times \widetilde{\mathcal{N}} \to \mathbb{R}_{\geq 0}$ (called

[2] The function $V : \mathcal{I} \times \widetilde{\mathcal{N}} \to [0, \infty)$ is norm like if $\lim_{n \to \infty} V(i, n) = \infty$ for all $i \in \mathcal{I}$.

the Lyapunov function) such that, for some $c > 0$ and $d < \infty$,

$$\mathcal{L}V(i,n) \leq -cV(i,n) + d, \quad \forall (i,n) \in \mathcal{I} \times \widetilde{\mathcal{N}}. \tag{4.30}$$

The above condition is usually referred to as the *drift condition* [74]. Under this condition, for any initial condition $(i,n) \in \mathcal{I} \times \widetilde{\mathcal{N}}$,

$$\limsup_{t\to\infty} \frac{1}{t} \int_{\tau=0}^{t} \mathcal{E}\Big[V(I(\tau), N(\tau)) | I(0) = i, N(0) = n\Big] d\tau \leq d/c. \tag{4.31}$$

We consider the Lyapunov function $V : \mathcal{I} \times \widetilde{\mathcal{N}} \to \mathbb{R}_{\geq 0}$ defined as follows:

$$V(i,n) = a_i \exp\left(b\Gamma^T n\right), \tag{4.32}$$

where $a = [a_1, \ldots, a_m]^T$ and b are strictly positive constants (to be determined) and Γ is defined in (4.23). The switched Lyapunov function captures the effect of both the mode (via the coefficient a_i) and the traffic density (via the exponential term $\exp(b\Gamma^T n)$). Intuitively, V decreases when the freeway switches to a mode with larger capacities or when traffic is discharged from upstream cells to downstream cells. The exponential form is in line with our notion of stability (4.13) and facilitates verification of the drift condition (4.30).

The main challenge of verifying (4.30) is to show that

$$d \geq \max_{(i,n) \in \mathcal{I} \times \widetilde{\mathcal{N}}} \mathcal{L}V(i,n) + cV(i,n). \tag{4.33}$$

Since the maximization problem in the right-hand side of (4.33) is rather complex, the drift condition is not easy to verify and is thus far from checkable in its original form. To address this challenge, we utilize properties of CTM dynamics to show that (4.30) can be established by minimizing a concave function (see (4.51) in Appendix C) over the rectangular set $\widetilde{\mathcal{N}}$, where an optimal solution must lie at one of the vertices of $\widetilde{\mathcal{N}}$.

Finally, we note that, in general, there exists a gap between the necessary condition

(Theorem 4.1) and the sufficient condition (Theorem 4.2). Since both these results rely on the invariant set $\widetilde{\mathcal{N}}$, the gap depends on the construction of the invariant set. Indeed, both results also apply to other invariant sets that can be expressed of the form in (4.16). The following result addresses how the construction of the invariant set affects the gap:

Proposition 4.2. *Consider two invariant sets*

$$\widetilde{\mathcal{N}} = [\underline{n}_1, \infty) \times \prod_{k=2}^{K} [\underline{n}_k, \overline{n}_k], \; \widetilde{\mathcal{N}}' = [\underline{n}_1', \infty) \times \prod_{k=2}^{K} [\underline{n}_k', \overline{n}_k']$$

such that $\widetilde{\mathcal{N}} \subseteq \widetilde{\mathcal{N}}'$, i.e. $\underline{n} \geq \underline{n}'$ and $\overline{n} \leq \overline{n}'$. For a given $r \in \mathcal{R}$,

(i) if $\widetilde{F}_k^i(\underline{n}, r)$ satisfies (4.21), so does $\widetilde{F}_k^i(\underline{n}', r)$;

(ii) if $\mathscr{F}_i(\underline{n}', \overline{n}', r)$ allows positive solutions for $a_1, \ldots, a_m$, b to (4.27), so does $\mathscr{F}_i(\underline{n}, \overline{n}, r)$.

Proposition 4.2 implies that a smaller invariant set leads to sharper stability conditions (i.e. a smaller gap between the necessary condition and the sufficient condition). Indeed, the invariant set given by Proposition 4.1 is in some cases the smallest invariant set of the form in (4.16).

4.3 Proofs of Main Results

4.3.1 Proof of Proposition 4.1

Invariant

To show the invariance of $\widetilde{\mathcal{N}}$, we demonstrate that the vector field G points towards the interior of $\widetilde{\mathcal{N}}$ everywhere on the boundary of $\widetilde{\mathcal{N}}$. That is, for each $n \in \widetilde{\mathcal{N}}$ such that $n_k = \underline{n}_k$ (resp. $n_k = \overline{n}_k$) for some $k \in \{1, \ldots, K\}$ (resp. $k \in \{2, \ldots, K\}$), we have $G_k(i, n, r) \geq 0$ (resp. $G_k(i, n, r) \leq 0$) for all $i \in \mathcal{I}$.

a) We first study the directionality of the vector field on the lower boundaries. Consider a given $r \in \mathcal{R}$.

a.1) For each $n \in \widetilde{\mathcal{N}}$ such that $n_1 = \underline{n}_1$, we have

$$\begin{aligned} G_1(i,n,r) &\overset{(4.9a)}{=} r_1 - f_1(i,\underline{n}_1,n_2,r_2)/\rho_1 \\ &\overset{(4.7b)}{=} r_1 - \min\left\{\alpha_1\underline{n}_1, F_1^i, (R_2(n_2)-r_2)_+/\rho_2\right\} \\ &\ge r_1 - \alpha_1\underline{n}_1 \overset{(4.17a)}{\ge} r_1 - \alpha_1\frac{r_1}{\alpha_1} = 0, \quad \forall i \in \mathcal{I}. \end{aligned} \tag{4.34}$$

a.2) For each $n \in \widetilde{\mathcal{N}}$ such that $n_k = \underline{n}_k$ for some $k \in \{2,\dots,K-1\}$, we need to show that $G_k \ge 0$.

First, note that

$$\begin{aligned} & f_{k-1}\left(i, n_{k-1}, \underline{n}_k, r_k\right) \\ & \overset{(4.7b)}{=} \min\left\{\rho_{k-1}\alpha_{k-1}n_{k-1}, \rho_{k-1}F_{k-1}^i, R_k(\underline{n}_k) - r_k\right\} \\ & \ge \min\left\{\rho_{k-1}\alpha_{k-1}\underline{n}_{k-1}, \rho_{k-1}F_{k-1}^i, R_k(\underline{n}_k) - r_k\right\} \\ & \overset{(4.5b)(4.17b)}{\ge} \min\left\{\alpha_k\underline{n}_k - r_k, \beta_k(n_k^{\max} - \underline{n}_k) - r_k\right\}. \end{aligned} \tag{4.35}$$

Since $\underline{n}_k \overset{(4.17b)}{\le} \mathsf{F}_k/\alpha_k$, we can obtain from (4.6) that

$$\alpha_k\underline{n}_k \le \beta_k\left(n_k^{\max} - \underline{n}_k\right).$$

Plugging the above into (4.35), we obtain

$$f_{k-1}\left(i, n_{k-1}, \underline{n}_k, r_k\right) \ge \alpha_k\underline{n}_k - r_k. \tag{4.36}$$

Next, note that

$$\begin{aligned} & f_k\left(i, \underline{n}_k, n_{k+1}, r_{k+1}\right)/\rho_k \\ & \overset{(4.7b)}{=} \min\left\{\alpha_k\underline{n}_k, F_k^i, \frac{\beta_{k+1}\left(n_{k+1}^{\max} - n_{k+1}\right) - r_{k+1}}{\rho_k}\right\} \le \alpha_k\underline{n}_k. \end{aligned} \tag{4.37}$$

Hence, we have

$$
\begin{aligned}
& G_k(i,n,k) \\
& = f_{k-1}\left(i, n_{k-1}, \underline{n}_k, r_k\right) + r_k - f_k\left(i, \underline{n}_k, n_{k+1}, r_{k+1}\right)/\rho_k \\
& \overset{(4.36)(4.37)}{\geq} \quad \alpha_k \underline{n}_k - r_k + r_k - \alpha_k \underline{n}_k = 0, \quad \forall i \in \mathcal{I}.
\end{aligned} \tag{4.38}
$$

a.3) The proof for $k = K$ is analogous.

b) Next, we study the directionality of the vector field G on the upper boundaries. Again, consider a given $r \in \mathcal{R}$.

b.1) For each $n \in \widetilde{\mathcal{N}}$ such that $n_K = \overline{n}_K$, we need to consider the two subcases in (4.17c):

If $\rho_{K-1}\mathsf{F}_K + r_K \leq F_K^{\min}$, then we have

$$
\begin{aligned}
& f_{K-1}(i, n_{K-1}, \overline{n}_K, r_K) \\
& \overset{(4.7b)}{=} \min\{\rho_{K-1}\alpha_{K-1}n_{K-1}, \rho_{K-1}F_{K-1}^i, \rho_K(n_k^{\max} - \overline{n}_K) - r_K\} \\
& \leq \rho_{K-1}F_{K-1}^i \leq \rho_{K-1}\mathsf{F}, & (4.39a) \\
& \frac{f_K(i, \overline{n}_K)}{\rho_K} \overset{(4.7c)}{=} \min\{\alpha_K \overline{n}_K, F_K^i\} \overset{(4.17c)}{=} \rho_{K-1}\mathsf{F} + r_K. & (4.39b)
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
G_K(i,n,r) & \overset{(4.9c)}{=} f_{K-1}(i, n_{K-1}, \overline{n}_K, r_K) + r_K - \frac{f_K(i_K, \overline{n}_K)}{\rho_K} \\
& \overset{(4.39)}{\leq} \rho_{K-1}\mathsf{F}_{K-1} + r_K - (\rho_{K-1}\mathsf{F}_{K-1} + r_K) = 0, \quad \forall i \in \mathcal{I}.
\end{aligned}
$$

Otherwise, we have

$$
\begin{aligned}
& f_{K-1}(i, n_{K-1}, \overline{n}_K, r_K) \overset{(4.7b)}{\leq} \beta_k(n_k^{\max} - \overline{n}_K) - r_K \\
& \overset{(4.17c)}{=} F_K^{\min} - r_K, & (4.40a) \\
& f_K(i_K, \overline{n}_K)/\rho_k \overset{(4.7c)(4.17c)}{=} \min\{\alpha_K(n_k^{\max} - F_K^{\min}/\rho_k), F_K^i\} \\
& \overset{(4.6)}{=} F_K^i \geq F_K^{\min}. & (4.40b)
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
G_K(i,n,r) &\overset{(4.9c)}{=} f_{K-1}(i,n_{K-1},\overline{n}_K,r_K)+r_K-\frac{f_K(i_K,\overline{n}_K)}{\rho_K} \\
&\overset{(4.40)}{\le} (F_K^{\min}-r_K)+r_K-F_K^{\min}=0, \quad \forall i\in\mathcal{I}.
\end{aligned}
$$

b.2) For $n\in\widetilde{\mathcal{N}}$ such that $n_k=\overline{n}_k$ for some $k\in\{1,\dots,K-1\}$, we again need to consider both cases indicated in (4.17d):

If $\rho_{k-1}\mathsf{F}_{k-1}+r_k\le\min\{F_k^{\min},\frac{R_{k+1}(\overline{n}_{k+1})-r_{k+1}}{\rho_k}\}$, then we have

$$
f_{k-1}(i,n_{k-1},\overline{n}_k,r_k)\le\rho_{k-1}F_{k-1}^i\le\rho_{k-1}\mathsf{F}_{k-1}, \tag{4.41a}
$$

$$
\begin{aligned}
&f_k(i,\overline{n}_k,n_{k+1},r_{k+1})/\rho_k \\
&\overset{(4.7b)}{=}\min\left\{\alpha_k\overline{n}_k,F_k^i,\frac{\beta_{k+1}(n_{k+1}^{\max}-n_{k+1})-r_{k+1}}{\rho_k}\right\} \\
&\ge\min\left\{\alpha_k\overline{n}_k,F_k^i,\frac{\beta_{k+1}(n_{k+1}^{\max}-\overline{n}_{k+1})-r_{k+1}}{\rho_k}\right\} \\
&\overset{(4.17d)}{=}\min\left\{\rho_{k-1}\mathsf{F}_{k-1}+r_k,F_k^i,\frac{\beta_{k+1}(n_{k+1}^{\max}-\overline{n}_{k+1})-r_{k+1}}{\rho_k}\right\} \\
&=\rho_{k-1}\mathsf{F}_{k-1}+r_k.
\end{aligned} \tag{4.41b}
$$

Thus, we have

$$
\begin{aligned}
&G_k(i,n,r) \\
&=f_{k-1}(i_{k-1},n_{k-1},\overline{n}_k,r_k)+r_k-f_k(i,\overline{n}_k,n_{k+1},r_{k+1})/\rho_k \\
&\overset{(4.41)}{\le}\rho_{k-1}\mathsf{F}_{k-1}+r_k-(\rho_{k-1}\mathsf{F}_{k-1}+r_k)=0, \quad \forall i\in\mathcal{I}.
\end{aligned}
$$

Otherwise, we have

$$
\begin{aligned}
&f_{k-1}(i,n_{k-1},\overline{n}_k,r_k)\le\beta_k(n_k^{\max}-\overline{n}_k)-r_k \\
&\overset{(4.17d)}{=}\min\left\{F_k^{\min},\frac{\beta_{k+1}(n_{k+1}^{\max}-\overline{n}_{k+1})-r_{k+1}}{\rho_k}\right\}-r_k,
\end{aligned} \tag{4.42a}
$$

$$
f_k(i,\overline{n}_k,n_{k+1},r_{k+1})/\rho_k
$$

$$
\begin{aligned}
&= \min\left\{\alpha_k\overline{n}_k, F_k^i, \frac{\beta_{k+1}(n_{k+1}^{\max} - n_{k+1}) - r_{k+1}}{\rho_k}\right\}\\
&\geq \min\left\{\alpha_k\overline{n}_k, F_k^i, \frac{\beta_{k+1}(n_{k+1}^{\max} - \overline{n}_{k+1}) - r_{k+1}}{\rho_k}\right\}\\
&\overset{(4.17d)}{=} \min\left\{F_k^i, \frac{\beta_{k+1}(n_{k+1}^{\max} - \overline{n}_{k+1}) - r_{k+1}}{\rho_k}\right\}\\
&\geq \min\left\{F_k^{\min}, \frac{\beta_{k+1}(n_{k+1}^{\max} - \overline{n}_{k+1}) - r_{k+1}}{\rho_k}\right\}. \qquad (4.42b)
\end{aligned}
$$

Thus, we have

$$
\begin{aligned}
G_k(i,n,r) &= f_{k-1}(i, n_{k-1}, \overline{n}_k, r_k) + r_k\\
&\quad - f_k(i_k, \overline{n}_k, n_{k+1}, r_{k+1})/\rho_k \overset{(4.42)}{\leq} 0, \quad \forall i \in \mathcal{I}.
\end{aligned}
$$

Combining cases a) and b), we obtain that $\widetilde{\mathcal{N}}$ is invariant.

Attracting

To show that the set $\widetilde{\mathcal{N}}$ is attracting, we define

$$
\begin{aligned}
\underline{\mathcal{N}}_k &= \{n \in \mathcal{N} : n_k \geq \underline{n}_k\}, \; k = 1, 2, \ldots, n,\\
\overline{\mathcal{N}}_k &= \{n \in \mathcal{N} : n_k \leq \overline{n}_k\}, \; k = 2, 3, \ldots, n.
\end{aligned}
$$

Thus, we have $\widetilde{\mathcal{N}} = (\cap_{k=1}^K \underline{\mathcal{N}}_k) \cap (\cap_{k=2}^K \overline{\mathcal{N}}_k)$. Consider a given $r \in \mathcal{R}$.

a) First, we show by induction that the set $\cap_{k=1}^K \underline{\mathcal{N}}_k$ is attracting.

a.1) For any $\epsilon > 0$, consider $\mathcal{B}(\underline{\mathcal{N}}_1, \epsilon)$, i.e. the neighborhood of $\underline{\mathcal{N}}_1$ such that $\min_{\varrho \in \underline{\mathcal{N}}_1} \|n - \varrho\|_2 \leq \epsilon$. Without loss of generality, we consider $0 < \epsilon < \min_{k:\underline{n}_k > 0} \underline{n}_k$. If $\underline{n}_1 > 0$, for any $n \in \mathcal{B}^c(\underline{\mathcal{N}}_1, \epsilon)$ (complement of $\mathcal{B}(\underline{\mathcal{N}}_1, \epsilon)$), we have $n_1 < \underline{n}_1 - \epsilon$. Then, we obtain from (4.5a) and (4.7) that

$$
\begin{aligned}
G_1(i,n,r) &= r_1 - f_1(i, n_1, n_2, r_2)/\rho_1 \geq r_1 - \alpha_1 n_1\\
&\geq r_1 - \alpha_1(\underline{n}_1 - \epsilon) \geq \alpha_1\epsilon > 0, \quad \forall i \in \mathcal{I}.
\end{aligned}
$$

Therefore, for any initial condition $(i,n) \in \mathcal{I} \times \mathcal{B}^c(\underline{\mathcal{N}}_1, \epsilon)$, there exists $T = \underline{n}_1/(\alpha_1\epsilon)$

such that $N(t) \in \mathcal{B}(\underline{\mathcal{N}}_1, \epsilon)$ for all $t \geq T$. Hence, the set $\underline{\mathcal{N}}_1$ is attracting in the sense of (4.14b).

If $\underline{n}_1 = 0$, the proof is trivial.

a.2) Now, suppose that the set $(\cap_{h=1}^{k} \underline{\mathcal{N}}_h)$ is attracting. If $\underline{n}_{k+1} = 0$, for any $\epsilon > 0$, consider the neighborhood $\mathcal{B}(\cap_{h=1}^{k+1} \underline{\mathcal{N}}_h, \epsilon)$. For each $n \in (\cap_{h=1}^{k} \underline{\mathcal{N}}_h) \cap \mathcal{B}^{c}(\cap_{h=1}^{k+1} \underline{\mathcal{N}}_h, \epsilon)$, we have $n_1 \geq \underline{n}_1, \ldots, n_k \geq \underline{n}_k, n_{k+1} < \underline{n}_{k+1} - \epsilon$. Then, we obtain from (4.5a) and (4.7) that

$$\begin{aligned}
G_{k+1}(i,n,r) &= f_k(i, n_k, n_{k+1}, r_{k+1}) + r_{k+1} \\
&\quad - f_{k+1}(i, n_{k+1}, n_{k+2}, r_{k+2})/\rho_{k+1} \\
&\overset{(4.7b)}{\geq} f_k\left(i, \underline{n}_k, n_{k+1}, r_{k+1}\right) + r_{k+1} \\
&\quad - f_{k+1}(i, n_{k+1}, n_{k+2}, r_{k+2})/\rho_{k+1} \\
&\overset{(4.17b)}{\geq} \left(\alpha_{k+1}\underline{n}_{k+1} - r_{k+1}\right) + r_{k+1} - \alpha_{k+1}\left(\underline{n}_{k+1} - \epsilon\right) \\
&\geq \alpha_{k+1}\epsilon > 0, \quad \forall i \in \mathcal{I}.
\end{aligned}$$

Therefore, for any initial condition $(i, n) \in \mathcal{I} \times (\cap_{h=1}^{k} \underline{\mathcal{N}}_h) \cap \mathcal{B}^{c}(\cap_{h=1}^{k+1} \underline{\mathcal{N}}_h, \epsilon)$, there exists $T = \underline{n}_{k+1}/(\alpha_{k+1}\epsilon)$ such that $N(t) \in \mathcal{B}(\cap_{h=1}^{k+1} \underline{\mathcal{N}}_h, \epsilon)$ for all $t \geq T$. Recall that, by the inductive hypothesis, $\cap_{h=1}^{k} \underline{\mathcal{N}}_h$ is (globally) attracting. Hence, the set $\cap_{h=1}^{k+1} \underline{\mathcal{N}}_h$ is attracting.

If $\underline{n}_{k+1} = 0$, the proof is trivial.

In conclusion, $\cap_{k=1}^{K} \underline{\mathcal{N}}_k$ is attracting.

b) The proof for $\cap_{k=2}^{K} \overline{\mathcal{N}}_k$ being attracting is analogous.

4.3.2 Proof of Theorem 4.1

Suppose that the SS-CTM with a given inflow vector r is stable in the sense of (4.13). To establish the necessary condition, we can limit our attention to a particular initial condition in the invariant set, i.e. $N(0) = n \in \widetilde{\mathcal{N}}$. The proof consists of two steps. In Step 1), we show that the average flow is equal to the nominal flow. In Step 2), we show that the average flow is less than or equal to the average spillback-adjusted capacity.

1) Integrating (4.8), we obtain that, for $t \geq 0$,

$$N_k(t) = \int_{\tau=0}^{t} \Big(f_{k-1}(\tau) + r_k - f_k(\tau)/\rho_k \Big) d\tau + n_k,$$
$$k = 1, 2 \ldots, n.$$

Since $\lim_{t\to\infty} n_k/t = 0$ for $k = 1, 2, \ldots, n$, we can write

$$\begin{aligned} 0 &= \lim_{t\to\infty} \frac{1}{t} \left(\int_{\tau=0}^{t} \Big(f_{k-1}(\tau) + r_k - f_k(\tau)/\rho_k \Big) d\tau \right. \\ &\qquad \left. + n_k - N_k(t) \right) \\ &= \lim_{t\to\infty} \frac{1}{t} \left(\int_{\tau=0}^{t} \Big(f_{k-1}(\tau) + r_k - f_k(\tau)/\rho_k \Big) d\tau - N_k(t) \right). \end{aligned}$$

Since the MGF of $|N(t)|$ is bounded on average, we have $\Pr\{\lim_{t\to\infty} N_k(t) = \infty\} = 0$ for $k = 1, 2, \ldots, n$. Thus, we have

$$\lim_{t\to\infty} \frac{1}{t} \int_{\tau=0}^{t} \Big(f_{k-1}(\tau) + r_k - \frac{f_k(\tau)}{\rho_k} \Big) d\tau = 0, \quad a.s. \tag{4.43}$$

For $k = 1$, since $f_0 = 0$ by definition, we have

$$\lim_{t\to\infty} \frac{1}{t} \int_{\tau=0}^{t} \Big(r_1 - f_1(\tau)/\rho_1 \Big) d\tau = 0, \quad a.s. \tag{4.44}$$

which implies that

$$\lim_{t\to\infty} \frac{1}{t} \int_{\tau=0}^{t} f_1(\tau) d\tau = \rho_1 r_1. \quad a.s.$$

To proceed by induction, we assume that, for some $k \geq 1$, we have

$$\lim_{t\to\infty} \frac{1}{t} \int_{\tau=0}^{t} f_k(\tau) d\tau = \rho_k \sum_{h=1}^{k} \rho_h^k r_h, \quad a.s. \tag{4.45}$$

Then, we can obtain from (4.43) that

$$\begin{aligned}
&\lim_{t\to\infty}\frac{1}{t}\int_{\tau=0}^{t} f_{k+1}(\tau)d\tau \\
&= \rho_{k+1}\left(\lim_{t\to\infty}\frac{1}{t}\int_{\tau=0}^{t} f_k(\tau)d\tau + r_{k+1}\right) \\
&\overset{(4.45)}{=} \rho_{k+1}\left(\rho_k\sum_{h=1}^{k}\rho_h^k r_h + r_{k+1}\right) = \rho_{k+1}\sum_{h=1}^{k+1}\rho_h^{k+1}r_h,\ a.s.
\end{aligned}$$

Hence, we conclude that

$$\begin{aligned}
\lim_{t\to\infty}\frac{1}{t}\int_{\tau=0}^{t} f_k(\tau)d\tau = \rho_k\sum_{h=1}^{k}\rho_h^k r_h,\ a.s. \\
k = 1, 2, \dots, n,
\end{aligned}\tag{4.46}$$

which means that the average flow is equal to the nominal flow.

2) For every $i \in \mathcal{I}$, let $T_i(t)$ be the amount of time that the SS-CTM is in mode i up to time t, i.e.

$$T_i(t) = \int_{\tau=0}^{t} \mathbf{1}_{\{I(\tau)=i\}}d\tau. \tag{4.47}$$

Then, recalling Assumption 3 and using [33, Theorem 7.2.6], we obtain

$$\lim_{t\to\infty}\frac{T_i(t)}{t} = \mathsf{p}_i, \quad a.s. \tag{4.48}$$

In addition, since $\widetilde{\mathcal{N}}$ is invariant, we know from (4.14a) that $n(\tau) \in \widetilde{\mathcal{N}}$ for all $\tau \geq 0$. Then, we have, for $k = 1, 2, \dots, n$,

$$\begin{aligned}
&\lim_{t\to\infty}\frac{1}{t}\int_{\tau=0}^{t} f_k(\tau)d\tau \overset{(4.19)}{\leq} \lim_{t\to\infty}\frac{1}{t}\int_{\tau=0}^{t} \rho_k\widetilde{F}_k^{I(\tau)}(\underline{n}, r)d\tau \\
&\overset{(4.47)}{=} \rho_k\lim_{t\to\infty}\sum_{i\in\mathcal{I}}\frac{T_i(t)\widetilde{F}_k^i(\underline{n}, r)}{t} \overset{(4.48)}{=} \rho_k\sum_{i\in\mathcal{I}}\mathsf{p}_i\widetilde{F}_k^i(\underline{n}, r). \quad a.s.
\end{aligned}\tag{4.49}$$

Combining (4.46) and (4.49), we obtain (4.21).

4.3.3 Proof of Theorem 4.2

Recall from Section 4.2.1 that, to obtain stability, we need to show (4.33). Let us proceed with the following expression:

$$\begin{aligned}
&\max_{n\in\widetilde{\mathcal{N}}} \mathcal{L}V(i,n) + cV(i,n) \\
&\overset{(4.15)(4.32)}{=} \max_{n\in\widetilde{\mathcal{N}}} \left(a_i b \sum_{k=1}^{K} \Gamma_k (f_{k-1} + r_k - f_k/\rho_k) \right. \\
&\qquad\qquad \left. + \sum_{j=1}^{m} \lambda_{ij}(a_j - a_i) + a_i c \right) \frac{V}{a_i} \\
&\overset{(4.23)(4.24)}{=} \left(a_i b \Big(\mathscr{R}(r) - \min_{n\in\widetilde{\mathcal{N}}} \gamma^T f(i,n,r) \Big) \right. \\
&\qquad\qquad \left. + \sum_{j=1}^{m} \lambda_{ij}(a_j - a_i) + a_i c \right) \frac{V}{a_i}, \ \forall i \in \mathcal{I},
\end{aligned} \tag{4.50}$$

where c and d are given by (4.28). The key to evaluate (4.50) is to compute $\min_{n\in\widetilde{\mathcal{N}}} \gamma^T f(i,n,r)$. To do this, we define

$$\begin{aligned}
\widetilde{\mathcal{N}}_1 &:= \left[n^{\text{crit}}, \infty\right) \times \prod_{k=2}^{K} \left[\underline{n}_k, \overline{n}_k\right], \\
\widetilde{\mathcal{N}}_2 &:= \left[\underline{n}_1, n^{\text{crit}}\right] \times \prod_{k=2}^{K} \left[\underline{n}_k, \overline{n}_k\right]
\end{aligned}$$

and consider two cases:

1) $n \in \widetilde{\mathcal{N}}_1$. For each $r \in \mathcal{R}$ and each $i \in \mathcal{I}$, consider

$$\min_{n\in\widetilde{\mathcal{N}}_1} \quad \gamma^T f(i,n,r). \tag{4.51}$$

We claim that an optimal solution of (4.51) lies on one of the vertices of $\widetilde{\mathcal{N}}_1$, i.e. for all $i \in \mathcal{I}$,

$$\min_{n\in\widetilde{\mathcal{N}}_1} \gamma^T f(i,n,r) = \min_{n\in\Theta} \gamma^T f(i,n,r) \overset{(4.26a)}{=} \mathscr{F}_i(\underline{n}, \overline{n}, r), \tag{4.52}$$

where Θ is defined in (4.25a). We will prove this claim at the end of this subsection.

Then, for each $i \in \mathcal{I}$, we have

$$\begin{aligned}
&\max_{n\in\widetilde{\mathcal{N}}_1} \mathcal{L}V(i,n) + cV(i,n) \\
&\overset{(4.50)(4.52)}{\leq} \left(a_i b\left(\mathscr{R}(r) - \mathscr{F}_i(\underline{n},\overline{n},r)\right) + \sum_{j=1}^{m} \lambda_{ij}(a_j - a_i) + a_i c \right) \\
&\quad \times \exp\left(\Gamma^T n\right) \\
&\overset{(4.27)}{\leq} (-1 + a_i c)\exp\left(\Gamma^T n\right) \overset{(4.28a)}{\leq} 0 \overset{(4.28b)}{\leq} d.
\end{aligned}$$

2) $n \in \widetilde{\mathcal{N}}_2$. This case is straightforward, since $\widetilde{\mathcal{N}}_2$ is bounded. We claim that, for all $i \in \mathcal{I}$,

$$\begin{aligned}
\min_{n\in\widetilde{\mathcal{N}}_2} \gamma^T f(i,n,r) &= \min_{n\in\hat{\Theta}} \gamma^T f(i,n,r) \\
&\overset{(4.26b)}{=} \hat{\mathscr{F}}_i(\underline{n},\overline{n},r),
\end{aligned} \tag{4.53}$$

where $\hat{\Theta}$ is as defined in (4.25b); again, we will prove this claim at the end of this subsection. Then, we have

$$\begin{aligned}
&\max_{n\in\widetilde{\mathcal{N}}_2} \mathcal{L}V + cV \\
&\leq \left| a_i b\left(\mathscr{R}(r) - \hat{\mathscr{F}}_i(\underline{n},\overline{n},r)\right) + \sum_{j=1}^{m} \lambda_{ij}(a_j - a_i) + a_i c \right| \\
&\quad \times \frac{\max_{n\in\widetilde{\mathcal{N}}_2} V}{a_i} \overset{(4.28b)}{\leq} d,\ \forall i \in \mathcal{I}.
\end{aligned}$$

Since $\widetilde{\mathcal{N}} = \widetilde{\mathcal{N}}_1 \cup \widetilde{\mathcal{N}}_2$, combining cases a) and b), we obtain (4.33) and thus the drift condition (4.30). Using [74, Theorem 4.3], we obtain (4.29) from the drift condition.

The rest of this subsection is devoted to the derivation of (4.52) and (4.53):

To show (4.52), it suffices to argue that, for a given $i \in \mathcal{I}$ and a given $r \in \mathcal{R}$, and for each $k \in \{1, \dots, K\}$, $\gamma^T f(i,n,r)$ is concave in n_k. For each $i \in \mathcal{I}$, let us consider the following quantity:

$$H_k^i(n) = \begin{cases} \gamma_1 f_1(i, n_1, n_2, r_2), & k = 1, \\ \gamma_{k-1} f_{k-1}(i, n_{k-1}, n_k, r_k) & \\ \quad + \gamma_k f_k(i, n_k, n_{k+1}, r_{k+1}), & 2 \leq k \leq K-1, \\ \gamma_{K-1} f_{K-1}(i, n_{K-1}, n_K, r_K) & \\ \quad + \gamma_K f_K(i, n_K), & k = K. \end{cases}$$

Then, for each $k \in \{1, \ldots, K\}$, we have

$$\gamma^T f(i, n, r) = H_k^i(n) + M_k^i(n_1, \ldots, n_{k-1}, n_{k+1}, \ldots, n_K),$$

where M_k^i is a term independent of n_k. Hence, to show that $\gamma^T f(i, n, r)$ is concave in n_k, we only need to show that $H_k(n)$ is concave in n_k.

We need to consider the following subcases of k:

a.1): For each $k \in \{2, \ldots, K-1\}$, we have

$$\begin{aligned} & H_k(n) \\ & = \gamma_{k-1} \min \left\{ \rho_{k-1} \alpha_{k-1} n_{k-1}, \beta_k F_{k-1}^i, \beta_{k-1}(n_k^{\max} - n_k) - r_k \right\} \\ & \quad + \gamma_k \min \left\{ \rho_k \alpha_k n_k, \rho_k F_k^i, \beta_{k+1}(n_{k+1}^{\max} - n_{k+1}) - r_{k+1} \right\}. \end{aligned}$$

In the above, the first term on the right side (corresponding to $\gamma_{k-1} f_{k-1}$) is non-increasing in n_k, while the second term (corresponding to $\gamma_k f_k$) is non-decreasing in n_k; both terms are piecewise affine in n_k with exactly one intersecting point (see Figure 4-2).

Note that the intersecting points n_k^* and n_k^{**} of the piecewise-linear functions $\gamma_{k-1} f_{k-1}$ and $\gamma_k f_k$, respectively, satisfy the following:

$$\begin{aligned} n_k^* &= n_k^{\max} - \frac{1}{w}(\min\{\rho_{k-1} \alpha n_{k-1}, \rho_{k-1} F_{k-1}^i\} - r_k) \\ &\geq n_k^{\max} - \frac{\rho_{k-1} F_{k-1}^i}{\beta_{k-1}} \geq n_k^{\max} - \frac{\mathsf{F}_{k-1}}{\beta_{k-1}} = n^{\text{crit}}, \end{aligned}$$

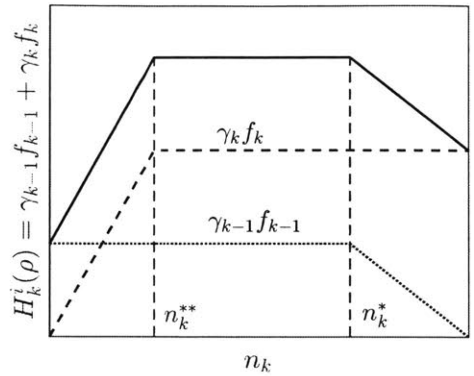


Figure 4-2: The function $H_k(n)$ is concave in n_k.

$$\begin{aligned} n_k^{**} &= \frac{1}{\rho_k \alpha_k} \min\{\rho_k F_k^i, \beta_{k+1}(n_{k+1}^{\max} - n_{k+1}) - r_{k+1}\} \\ &\leq \frac{F_k^i}{\alpha_k} \leq \frac{\mathsf{F}_k}{\alpha_k} = n^{\text{crit}}, \end{aligned}$$

Hence, we have $n_k^* \geq n_k^{**}$. Thus, we conclude that $H_k(n)$ is concave in n_k; see Figure 4-2.

a.2): For $k = 1$ and $k = K$, the expression of H_k is simpler and thus the derivation is straightforward.

The proof of (4.53) is analogous.

4.3.4 Proof of Proposition 4.2

We show the two conclusions in the statement separately:

(i): We can observe from (4.18) that $\widetilde{F}_k^i(\underline{n}, r)$ is non-increasing in $\underline{n}$. Hence, for each $\underline{n}' \leq \underline{n}$, we have

$$\widetilde{F}_k^i(r, \underline{n}') \geq \widetilde{F}_k^i(\underline{n}, r), \quad k = 1, 2, \ldots, n, \; i \in \mathcal{I}.$$

The conclusion follows from the above.

(ii): Consider $\widetilde{\mathcal{N}} \subseteq \widetilde{\mathcal{N}}'$. Define

$$\widetilde{\mathcal{N}}_1' := \left[n^{\text{crit}}, \infty\right) \times \prod_{k=2}^{K} \left[\underline{n}_k', \overline{n}_k'\right],$$
$$\widetilde{\mathcal{N}}_2' := \left[\underline{n}_1', n^{\text{crit}}\right] \times \prod_{k=2}^{K} \left[\underline{n}_k', \overline{n}_k'\right].$$

Clearly, $\widetilde{\mathcal{N}}_1 \subseteq \widetilde{\mathcal{N}}_1'$ and $\widetilde{\mathcal{N}}_2 \subseteq \widetilde{\mathcal{N}}_2'$. Then, from (4.52) and (4.53), we can obtain that

$$\begin{aligned}
&\mathscr{F}_i(\underline{n}, \overline{n}, r) = \min_{n \in \widetilde{\mathcal{N}}_1} \gamma^T f(i, n, r) \geq \min_{n' \in \widetilde{\mathcal{N}}_1'} \gamma^T f(i, n', r) \\
&= \mathscr{F}_i(r, \underline{n}', \overline{n}'), \quad \forall i \in \mathcal{I}, \\
&\mathscr{F}_i(\underline{n}, \overline{n}, r) = \min_{n \in \widetilde{\mathcal{N}}_2} \gamma^T f(i, n, r) \geq \min_{n' \in \widetilde{\mathcal{N}}_2'} \gamma^T f(i, n', r) \\
&= \mathscr{F}_i(r, \underline{n}', \overline{n}'), \quad \forall i \in \mathcal{I}.
\end{aligned}$$

The conclusion follows from the above and the fact that $a_1, a_2, \ldots, a_m$ and b in (4.27) are positive.

4.4 Some Practical Insights

In this section, we derive some practical insights for freeway traffic management under capacity fluctuations. Specifically, we use our results to (i) identify the set of stable inflow vectors for a given capacity model, (ii) analyze the impact due to frequency and intensity of capacity fluctuation on throughput, and (iii) study the effect of correlation between capacity fluctuation at various locations.

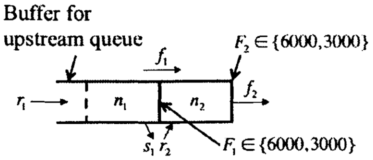
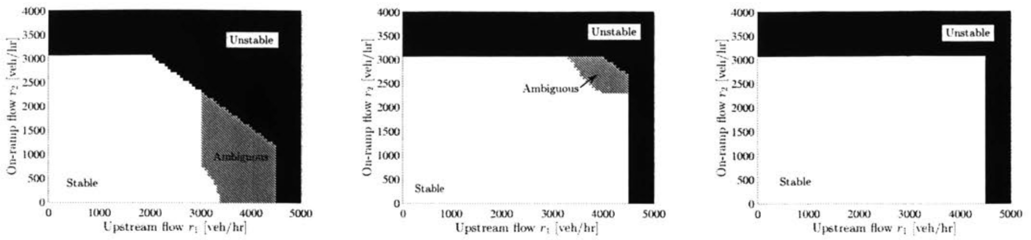


Figure 4-3: A two-cell freeway with two incident hotspots.



(a) Baseline model.

(b) Variant 1.

(c) Variant 2.

Figure 4-4: Stability of the two-cell freeway with various inflow vectors $r = [r_1, r_2]^T$ determined by Theorems 4.1 and 4.2.

A two-cell system as shown in Figure 4-3 is sufficient for our purpose. The parameters for the (baseline) capacity model is as follows:

$$\mathcal{I} = \{1, 2, 3, 4\}. \tag{4.54a}$$

$$F^1 = [6000, 6000]^T,\ F^2 = [3000, 6000]^T, \tag{4.54b}$$

$$F^3 = [6000, 3000]^T,\ F^4 = [3000, 3000]^T, \tag{4.54c}$$

$$\Lambda = \begin{bmatrix} -2 & 1 & 1 & 0 \\ 1 & -2 & 0 & 1 \\ 1 & 0 & -2 & 1 \\ 0 & 1 & 1 & -2 \end{bmatrix}. \tag{4.54d}$$

Note that the transition rate matrix defined above implies that the capacity fluctuations at both cells are mutually independent. We will discuss the impact of correlation between cell capacities in Section 4.4.3.

4.4.1 Set of stabilizing inflow vectors

For an inflow vector $r = [r_1, r_2]^T \in \mathbb{R}^2_{\geq 0}$, we know that r is unstable if it does not satisfy the necessary condition (Theorem 4.1), and that r is stable if it satisfies the sufficient condition (Theorem 4.2). For practicality, we also assume that the on-ramp has a fixed saturation rate of 3000 veh/hr. Thus, the on-ramp inflow r_2 cannot exceed 3000 veh/hr if the freeway is stable.

Applying our stability conditions to various inflow vectors $[r_1, r_2]^T$, we obtain

Figure 4-4(a). In this figure, the r_1-r_2 plane is partitioned into three regimes: The "Unstable" regime (in black) depicts the set of inflow vectors violating the necessary condition. Thus, any inflow vector in this regime leads to an infinite traffic queue. We denote the complement of this regime as $\mathcal{R}_1$. The "Stable" regime (in white) depicts the set of inflow vectors satisfying sufficient condition. In this example, we solve the bilinear inequalities (4.27) using YALMIP, a MATLAB-based optimization tool [71]. By Theorem 4.2, the inflow vectors in this regime lead to a traffic queue bounded in the sense of (4.13). We denote this regime as $\mathcal{R}_2$.

Notice that there is a gap, labeled as "Ambiguous", between the "Stable" and the "Unstable" regimes. This regime shows the gap between the necessary condition and the sufficient condition; for inflow vectors in this regime, our stability conditions do not provide a conclusive answer.

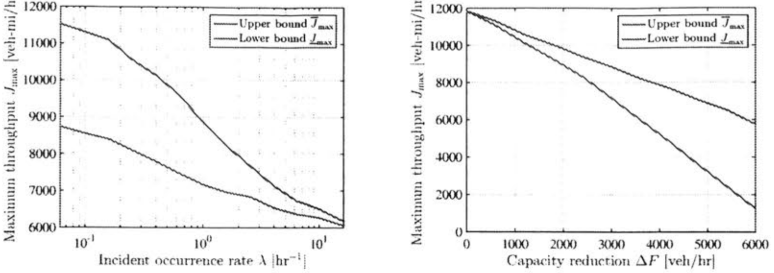
These results can be used to calculate stabilizing inflows that lead to maximum throughput, which partially motivates the results to be presented in the next chapter.

4.4.2 Impact of capacity fluctuation on throughput

Now we estimate the maximum throughput that can be achieved under a class of capacity models parameterized by ΔF and λ as follows:

$$
\begin{aligned}
&\mathcal{I} = \{1,2,3,4\}.\\
&F^1 = [6000, 6000]^T,\ F^2 = [6000 - \Delta F, 6000]^T,\\
&F^3 = [6000, 6000 - \Delta F]^T,\\
&F^4 = [6000 - \Delta F, 6000 - \Delta F]^T,\\
&\Lambda = \begin{bmatrix} -2\lambda & \lambda & \lambda & 0 \\ 1 & -(1+\lambda) & 0 & \lambda \\ 1 & 0 & -(1+\lambda) & \lambda \\ 0 & 1 & 1 & -2 \end{bmatrix}.
\end{aligned}
$$

By varying the parameters ΔF and λ, we can study the effect the intensity and the frequency of capacity fluctuations on the maximum throughput of SS-CTM.



(a) Throughput decreases as incident frequency increases, with ΔF fixed at 3000 veh/hr.

(b) Throughput decreases as incident intensity increases, with λ fixed at 1 per hour.

Figure 4-5: Relation between maximum throughput and incident frequency/intensity. The upper (resp. lower) bounds result from Theorem 4.1 (resp. Theorem 4.2).

For various $(\lambda, \Delta F)$ pairs, we numerically determine $\overline{J}_{\max}$ and $\underline{J}_{\max}$. Figure 4-5(a) shows that, with ΔF fixed at 3000 veh/hr, $\overline{J}_{\max}$ and $\underline{J}_{\max}$ decreases as λ increases. This is intuitive: more frequent capacity disruptions leads to lower throughput. Figure 4-5(b) shows that, with λ fixed at 1 per hr, $\overline{J}_{\max}$ and $\underline{J}_{\max}$ decreases as ΔF increases. This is also intuitive: larger capacity reduction leads to lower throughput.

Note that the throughput tends to be more sensitive to ΔF than to λ. Indeed, as shown in Figure 4-5(a), if λ is doubled from 1 (the baseline) to 2 per hr, the upper (resp. lower) bound is reduced by 7% (resp. 3%). However, if ΔF is doubled from 3000 (the baseline) to 6000 veh/hr, we can observe from Figure 4-5(b) that the upper (resp. lower) bound is reduced by 35% (resp. 82%).

The gap between $\overline{J}_{\max}$ and $\underline{J}_{\max}$ in Figures 4-5(a) and 4-5(b) result from the gap between the necessary condition and the sufficient condition for stability.

4.4.3 Impact of correlated capacity fluctuation

So far we have assumed that the cell capacities in the two-cell freeway are independent. Now we consider the case where the cells' capacities are correlated. We consider two extreme cases as follows.

Case 1: Suppose that the capacities of both cells are identical for all $t \geq 0$. In other words, the freeway has two modes $\{1, 2\}$ associated with $F^1 = [6000, 6000]^T$

and $F^2 = [3000, 3000]^T$. In this case, the transition matrix is

$$\Lambda_{\text{Case 1}} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

By implementing Theorems 4.1 and 4.2, we obtain the following bounds for the maximum throughput:

$$7485 \leq J_{\max} \leq 8910[\text{veh-mi/hr}].$$

Case 2: Suppose that the freeway always has exactly one incident in either cell. In other words, the freeway has two modes $\{1, 2\}$ associated with $F^1 = [6000, 3000]^T$ and $F^2 = [6000, 3000]^T$. In this case, the transition matrix is

$$\Lambda_{\text{Case 2}} = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}.$$

By implementing Theorems 4.1 and 4.2, we obtain the following bounds for the maximum throughput:

$$6720 \leq J_{\max} \leq 8910[\text{veh-mi/hr}].$$

In conclusion, even with the same average cell capacities (4500 veh/hr for both cells), the freeway's maximum throughput estimated by the sufficient condition could vary (in the order of 10%) due to spatial correlation. Therefore, compared to traditional approaches that assume independent cell capacities (e.g. [92]), our approach is able to capture the effect of spatial correlation and possibly achieve better throughput. In a related work [48], we have found that the extent of spatial correlation can be quite significant in practice.

We finally note that, in some situations, the transition rate matrix Λ may not be easy to calibrate. Analysis of the SS-CTM with partially known transition rates is indeed a practically relevant question, but is beyond the scope of this paper. We refer readers to [110] for more information on stochastic switching systems with partially known transition probabilities.

4.5 Summary

In this chapter, we developed (i) a stochastic switching cell transmission model for traffic dynamics in incident-prone freeways and (ii) easily checkable stability conditions for the SS-CTM (Theorems 4.1 and 4.2). A sufficient condition for stability is that the inflow does not exceed the spillback-adjusted capacity. A necessary condition for stability is that a set of bilinear inequalities, which is derived from the Foster-Lyapunov drift condition, admit positive solutions. Both conditions build on a construction of a globally attracting and invariant set of the SS-CTM (Proposition 4.1). Using these results, we derive new insights for the impact of capacity fluctuation on the upstream traffic queue length and the attainable throughput.

The results in this chapter motivated additional work in several directions. First, we developed a calibration approach for the SS-CTM , and constructed an SS-CTM for a section of the US Route 101 using real data [48]. Second, recent results on network traffic flow models [46, 22, 20] makes possible the extension of our model and method to the network setting, and to feedback-controlled systems, which naturally leads to the topic in the next chapter.

Chapter 5

Control Design for Highways Facing Perturbations

In the previous chapter, we analyzed the performance of highways facing stochastic capacity perturbations. In this chapter, we further consider the control design problem. We use a variation of the SS-CTM introduced in the previous chapter with on-ramp buffers to capture the evolution of traffic. For control design, we formulate a throughput-maximizing problem: the decision variables are the accepted demands at on-ramps (demand management) and priority of on-ramp traffic with respect to mainline (capacity allocation), and the constraint is the boundedness of the on-ramp queues. Using the stability theory of Markov processes, we derive necessary and sufficient conditions for bounded queues. We show that our stability conditions make the max-throughput problem a mixed integer bilinear or linear (depending whether the parameters of the Lyapunov function is solved or constructed) program. We also show that the throughput-maximizing control scheme prioritizes an on-ramp only if the capacity-demand margin of the on-ramp is smaller than that of the mainline, i.e. the "margin criterion."

The rest of this chapter is organized as follows. In Section 5.1, we introduce the stochastic traffic flow model and the max-throughput problem. In Section 5.2, we develop sufficient conditions for stable on-ramp queues. In Section 5.3, we present the MIBPL/MILP formulation and characterize optimal solutions of the max-throughput problem (the margin criterion). In Section 5.4, we present the results of the SR123/I210 simulation.

5.1 SS-CTM with control input

Now we consider a version of SS-CTM that is similar to that in the previous chapter, but with two important differences. First, in addition to the upstream queue, we now explicitly track the queues at the on-ramps $2, 3, \ldots, K$ as well. Second, we include control inputs, including the accepted inflow at each on-ramp and the priority of each on-ramp (with respect to the mainline), into the model and the subsequent argument.

Specifically, we consider a highway segment modeled as a compartmental system of K mainline *cells* with on-ramp *buffers* and off-ramp exits, as shown in Fig. 5-1. This model is based on [37, 45] with the exception this model also explicitly tracks the on-ramp queues. We call buffer 1 the *upstream buffer*, and buffers 2 through K the *on-ramp buffers*. The upstream buffer has a saturation rate $\mathsf{R}_1 = \mathsf{F}_1$ (veh/hr), same as the capacity of cell 1. The other on-ramp buffers have *saturation rates* R_k (veh/hr). The other model parameters are defined in the same way as in the previous chapter.

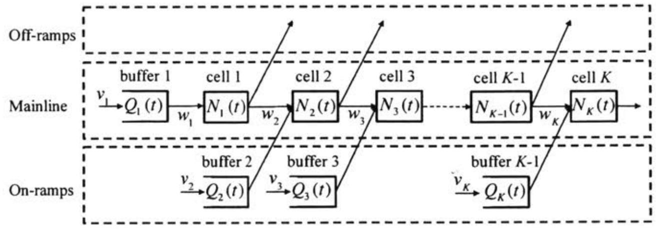


Figure 5-1: A highway with K mainline cells and K on-ramp buffers.

The *continuous state* of the highway is $x = (q, n)$, where $q = [q_1\ q_2\ \cdots\ q_K]^T$ is the lengths of the on-ramp queues and $n = [n_1\ n_2\ \cdots\ n_K]^T$ is the traffic densities in the mainline cells. For simplicity, we assume that every buffer has an infinite size; thus, the set of queue lengths is $\mathcal{Q} = \mathbb{R}^K_{\geq 0}$. Since the traffic density in each cell is non-negative and upper bounded by the jam density $n_k^{\max}$, the set of traffic densities is $\mathcal{N} := [0, n_k^{\max}]^K$. By including on-ramp queues as state variables, we can explicitly track performance (throughput) loss induced by capacity perturbations.

In our model, a *control input* is described by $u = (v, w)$, where $v = [v_1\ v_2\ \cdots\ v_K]^T$ denotes the vector of *inflows* (i.e. the demands that are admitted into the mainline) at the on-ramps, and $w = [w_1\ w_2\ \cdots\ w_K]^T$ denotes the vector of *priorities* assigned to the on-ramp traffic flows (with respect to the mainline traffic flow). Thus, in our model, each cell-buffer pair has two control inputs. The first is $v_k \in [0, d_k]$, the inflow into the mainline from on-ramp k, where d_k is the demand at the kth on-ramp (see Fig. 5-1). From a practical viewpoint, $v_k = d_k$ means that all the demand at on-ramp k is admitted, and $v_k < d_k$ means that only a fraction of the demand is admitted. We assume that any non-admitted demand is permanently rejected from the system and not redistributed to other locations. We make this conservative assumption to focus on how individual on-ramp buffers are affected by stochastic capacity perturbations. Indeed, our results can be extended to the case of interacting on-ramp buffers which can exchange traffic demand.[1] Also, we denote $d = [d_1\ d_2\ \cdots\ d_K]^T$.

The second control input w_k denotes the priority of inflow from buffer k with respect to the mainline. Specifically, $w_k = 1$ (resp. $w_k = 0$) means that the inflow from the kth on-ramp (resp. mainline) is prioritized over the flow from the mainline (resp. kth on-ramp). From a practical viewpoint, prioritizing the kth on-ramp (i.e. $w_k = 1$) means that the flow from the kth on-ramp is not metered and is given full priority over the mainline flow, and mainline priority (i.e. $w_k = 0$) means that the kth on-ramp is metered to give priority to the mainline flow. One can also view assigning priority as a way of allocating highway capacity between mainline and on-ramp, or distributing congestion on the mainline or queues at the on-ramps, as suggested by Varaiya [99].

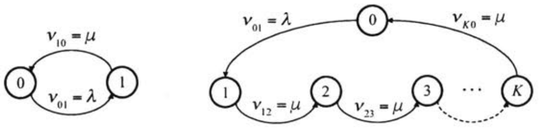
[1] For example, instead of $v_k \in [0, d_k]$ for each k, we can impose the constraint $0 \leq \sum_{k=1}^{K} v_k \leq \sum_{k=1}^{K} d_k]$.

5.1.1 Stochastic capacity model

We again consider the class of Markovian capacity perturbations defined in the previous chapter. That is, the cell capacities stochastically vary over time according to a finite-state Markov process over a set of *modes* denoted by $\mathcal{I}$. The inter-mode transition rates are $\{\nu_{ij}; i, j \in \mathcal{I}\}$. Every mode i is associated with a vector of cell capacities $F(i) = [F_1(i) \ \cdots \ F_k(i)]^T$. For ease of presentation, we assume that $F_k(i) \in \{\mathsf{F}_k, \mathsf{F}_k - \Delta_k\}$ for each k and for all i; that is, the capacity of the kth cell can only switch between two values F_k and $\mathsf{F}_k - \Delta_k$, where $\Delta_k \geq 0$ characterizes the intensity of capacity perturbation. We assume that the Markov chain governing the the mode transition process $\{I(t); t \geq 0\}$ is ergodic and associated with a unique (row) vector of steady-state probabilities $\mathsf{p} = [\mathsf{p}_0 \ \mathsf{p}_1 \ \cdots, \mathsf{p}_m]$ such that

$$\sum_{j \in \mathcal{I}} \nu_{ij} \mathsf{p}_i = \sum_{j \in \mathcal{I}} \nu_{ji} \mathsf{p}_j \quad \forall i \in \mathcal{I}, \quad |\mathsf{p}|= 1, \quad \mathsf{p} \geq 0. \tag{5.1}$$

The main results (Theorem 5.1 and Proposition 5.1) apply to this general class of capacity perturbations. In addition, we also derive particular results for two specific types of capacity perturbations of practical interest, which we introduce below.



(a) Stationary hotspot.
(b) Moving bottleneck.

Figure 5-2: Markov chains for the stochastic capacity models.

Stationary hotspot: Consider the setting where the Kth cell represents a highway section that recurrently experience capacity-reducing incidents. Suppose that cells 1 through $K-1$ have constant capacities, and only $\mathsf{F}_K(t)$ is stochastic. In this case, we call cell K as the *stationary hotspot*. When there is an incident in the Kth cell, its capacity is reduced from the nominal value F_K to $\mathsf{F}_K - \Delta_K$; for ease of presentation, we assume that every incident leads to the same amount of capacity reduction Δ_K. To model the occurrence and clearance of incidents, we consider that the highway stochastically switches between a *nominal mode* "0" and a *perturbed mode* "1." The *set of modes* is $\mathcal{I} = \{0, 1\}$, and the cell capacities are given by:

$$F_k(i) = \mathsf{F}_k \quad i = 0, 1, \ k = 1, \ldots, K-1,$$
$$F_K(i) = \begin{cases} \mathsf{F}_K & i = 0, \\ \mathsf{F}_K - \Delta_K & i = 1, \end{cases}$$

If the stationary hotspot faces incidents occurring at a rate λ [hr^{-1}] and clearing at rate μ [hr^{-1}], one can represent these transitions as switches of the mode $I(t)$ from 0 to 1 (resp. 1 to 0) at rate λ (resp. μ).

Moving bottlenecks: This perturbation model is relevant to highway sections that face recurrent congestion due to the presence of slow vehicles [26] or, in a proposed scenario, heavy-duty vehicle platoons [49]. These slow vehicles or vehicle platoons act as randomly moving bottlenecks for the regular traffic. In our model, we represent the initiation and movement of these bottlenecks by introducing mode switches between a *nominal mode* "0" and a set of *perturbed modes* "1", "2," ..., "K", where mode "k" means that the moving bottleneck is in cell k. Then, the *set of modes* is $\mathcal{I} = \{0, 1, \ldots, K\}$, and the mode-specific cell capacities are given by

$$F_k(i) = \begin{cases} \mathsf{F}_k - \Delta_k, & i = k, \\ \mathsf{F}_k, & \text{o.w.} \end{cases} \qquad k = 1, \ldots, K.$$

Following standard analysis approach [77], the randomness in the arrival of moving bottlenecks can be approximated as a time-homogeneous Markovian arrival process with rate λ. In addition, to account for the randomness in the movement of bottlenecks through the highway, we assume that the time that a moving bottleneck spends in a cell is an exponentially distributed random variable with mean $1/\mu$. Furthermore, for simplicity, we assume that, for each time t, at most one moving bottleneck can be present in the highway. Under these assumptions, the mode $I(t)$ evolves according to a Markov chain with transitions illustrated in Figure 5-2(b): as a moving bottleneck enters and moves through the highway, the mode switches from "0" to "1" to "2", and so on; as it leaves the highway, the mode switches from mode "K" back to "0."

5.1.2 Traffic flow model

We now describe the stochastic switching cell transmission model (SS-CTM, see [45]) as shown in Figure 5-1. We use $(i, x) = (i, q, n)$ to denote state variables and $(I(t), X(t)) = (I(t), Q(t), N(t))$ to denote (hyrbid) stochastic process. Following this convention, we let $Q_k(t)$ be the queue length in the kth buffer and $N_k(t)$ denote the traffic density in the kth cell at time t. Let $Q(t) = [Q_1(t), \dots, Q_K(t)]^T \in \mathcal{Q}$ and $N(t) = [N_1(t), \dots, N_K(t)]^T \in \mathcal{N}$, where $\mathcal{Q} = [0, \infty]^K$ and $\mathcal{N} = \prod_{k=1}^{K}[0, n_k^{\max}]$. For a fixed control input $u = (v, w) \in [0, d] \times \{0, 1\}^K$, the stochastic dynamics of the mode $I(t)$, the on-ramp queues $Q(t)$, and traffic densities $N(t)$ can be written as follows:

$$\Pr\Big\{I(t+\delta) = j | I(t) = i, I(s), s < t\Big\} = \nu_{ij}\delta + \mathrm{o}(\delta) \quad i, j \in \mathcal{I} : j \neq i, \tag{5.2a}$$

$$\dot{Q}(t) = G(I(t), X(t), u), \tag{5.2b}$$

$$\dot{N}(t) = H(I(t), X(t), u), \tag{5.2c}$$

where $G : \mathcal{I} \times (\mathcal{Q} \times \mathcal{N}) \times ([0, d] \times \{0, 1\}^K) \to \mathbb{R}^K$ and $H : \mathcal{I} \times (\mathcal{Q} \times \mathcal{N}) \times ([0, d] \times \{0, 1\}^K) \to \mathbb{R}^K$ are vector fields governing the dynamics of on-ramp queues and cell traffic densities to be developed below.

For each cell k, the sending flow S_k and the receiving flow T_k can be written as follows:

$$S_k(i, x) := \min\{\alpha_k n_k, F_k(i)\} \quad k = 1, 2, \dots, K, \tag{5.3a}$$

$$T_k(x) := \beta_k(n_k^{\max} - n_k) \quad k = 1, 2, \dots, K, \tag{5.3b}$$

where S_k is the traffic flow that cell k can discharge downstream and T_k is the traffic flow from upstream that cell k can accept. Following [24], we assume that

$$\mathsf{F}_k \leq \frac{\alpha_k \beta_k}{\alpha_k + \beta_k} n_k^{\max}. \tag{5.4}$$

For $k = 1, \dots, K-1$, let $\rho_k \in (0,1]$ denote the fixed *mainline ratio*, i.e. the fraction of traffic from cell k entering cell $k+1$; the remaining traffic flow leaves the highway at the kth off-ramp. Since the mainline ends at cell K, we have $\rho_K = 0$. In addition, we define

$$\rho_k^k := 1 \quad k = 1, 2 \dots, K, \tag{5.5a}$$

$$\rho_{k_1}^{k_2} := \prod_{h=k_1}^{k_2-1} \rho_h \quad 1 \le k_1 \le k_2 - 1,\ k_2 = 2, 3, \dots, K. \tag{5.5b}$$

Note that $\rho_{k_1}^{k_2}$ can be viewed as the fraction of the flow out of cell k_1 that eventually go through cell k_2.

We assume that every on-ramp isa fluid queueing system with an infinite buffer size [46]. That is, the sending flow from the kth buffer is given by

$$D_k(x, u) = \begin{cases} \min\{v_k, \mathsf{R}_k\} & q_k = 0, \\ \mathsf{R}_k & \text{o.w.} \end{cases} \qquad k = 1, \dots, K.$$

The *flow* r_k (resp. F_k) discharged by the kth buffer (resp. cell) is defined as follows:

$$r_k(i, x, u) := \min\left\{D_k(x, u), \Big(T_k(x) - (1 - w_k)S_k(i, x)\Big)_+\right\} \quad k = 1, 2, \dots, K, \tag{5.6a}$$

$$f_k(i, x, u) = \min\left\{S_k(i, x), \Big(T_{k+1}(x) - w_k D_{k+1}(i, x)\Big)_+\right\} \quad k = 1, 2, \dots, K-1, \tag{5.6b}$$

$$f_k(i, x, u) = S_K(i, x). \tag{5.6c}$$

where $(\cdot)_+$ stands for the positive part. In the above, we make the standard assumption that the Kth cell is not constrained from downstream [37]. One can see from (5.6a)–(5.6c) that the priority w_k determines whether the available receiving flow T_k is first allocated to the on-ramp ($w_k = 1$) or to the mainline $w_k = 0$.

We say that cell k is experiencing *spillback* at time t if $S_k(t) < \mathsf{F}_k(t)$, i.e. if the sending flow from cell k is strictly less than the actual flow. We say a buffer (resp. cell) k to be *congested* at time t if $Q_k(t) > 0$ (resp. $N_k(t) \geq F_k/\alpha_k$, where F_k/α_k is the critical density [37]). Then, we say that buffer k is a bottleneck at time t if buffer k is congested but cell k is not, and that cell k is a bottleneck at time t if cell k is congested but cell $k+1$ is not.

Remark 5.1. *Our notion of bottlenecks can be viewed as a time-varying extension of the static notion of bottlenecks considered in [37]. The authors of [37] consider a setting with constant demand and constant capacities, and thus the highway converges to a particular congestion pattern. In our model, a particular congestion pattern can recurrently occur and terminate because of the occurrence and clearance of capacity perturbations.*

Then, the vector fields G and H in (5.2b)-(5.2c) follow from mass conservation:

$$G_k(i,x,u) := v_k - r_k(i,x,u) \quad k = 1,\dots,K, \tag{5.7a}$$

$$H_1(i,x,u) := r_1(i,x,u) - f_1(i,x,u), \tag{5.7b}$$

$$H_k(i,x,u) := \rho_{k-1} f_{k-1}(i,x,u) + r_k(i,x,u) - f_k(i,x,u) \quad k = 2,\dots,K. \tag{5.7c}$$

5.1.3 Control design problem

We now introduce two definitions that are required for formulating the max-throughput problem, viz. stability and throughput.

For a given control input $u = (v,w)$, we say that the SS-CTM is *stable* if the limiting time-average on-ramp queues are bounded; i.e., there exists $Z < \infty$ such that for each initial condition $(i,q,n) \in \mathcal{I} \times \mathcal{Q} \times \mathcal{N}$,

$$\limsup_{t\to\infty} \frac{1}{t}\int_{s=0}^{t} \mathsf{E}[|Q(s)|]ds \leq Z. \tag{5.8}$$

This definition follows the notion of stability considered by Dai and Meyn in the context of fluid queuing systems [27].

We say that a demand vector d is *feasible* if there exists a stabilizing control input $u = (v, w)$ such that $v = d$, and *infeasible* otherwise.

Instead, our stability analysis in Section 5.2 focuses on deriving a sufficient condition in the form

$$C(u, \theta) \leq 0 \tag{5.9}$$

where C is a vector-valued function. Note that C may contain auxiliary variables θ (taking value in some set Θ) that do not show up in the objective function. Thus, the set of u satisfying $C(u, \theta) \leq 0$ is a subset of stabilizing control inputs.

Next, for a given control input $u = (v, w)$, *throughput* is defined as the time-average off-ramp flows discharged by the highway:

$$J := \lim_{t \to \infty} \frac{1}{t} \int_{s=0}^{t} \sum_{k=1}^{K} (1 - \rho_k) \mathsf{F}_k\Big(I(s), Q(s), N(s)\Big) ds.$$

Direct computation of the above limit involves integration of traffic flows, which evolve according to non-linear stochastic dynamics, and is thus not easy. However, we note that, if the on-ramp queues are stable in the sense of (5.8), then, by mass conservation, the time-average flow out of a cell becomes equal to the time-average flow into the cell almost surely (a.s.):

$$\lim_{t \to \infty} \frac{1}{t} \int_{s=0}^{t} f_k\Big(I(s), Q(s), N(s)\Big) ds = \sum_{h=1}^{k} \rho_h^k v_h \quad a.s. \quad k = 1, 2, \ldots, K.$$

Thus, we can rewrite the throughput as

$$J \overset{a.s.}{=} \sum_{k=1}^{K} \sum_{h=k}^{K} (1 - \rho_k) \rho_h^K v_h = \sum_{k=1}^{K} \left(\sum_{h=k}^{K} (1 - \rho_h) \rho_k^h \right) v_k = \sum_{k=1}^{K} v_k. \tag{5.10}$$

We can now rewrite the max-throughput problem (P0) as follows:

$$\max \quad \sum_{k=1}^{K} v_k \tag{P}$$

$$
\begin{aligned}
s.t.\quad & C(u,\theta) \leq 0, \\
& u \in [0,d] \times \{0,1\}^K,\ \theta \in \Theta. \qquad (5.11)
\end{aligned}
$$

Finally, we say that u is *feasible w.r.t.* (P) if u is a feasible solution to (P).

5.2 Sufficient condition for stability

In this section, we develop a sufficient condition for the boundedness of on-ramp queues that can be expressed as (5.9). To state the main result (Theorem 5.1), for each control input $u = (v, w)$, we define a finite set of states $\mathcal{V}(u) \subset \mathcal{Q} \times \mathcal{N}$:

$$
\mathcal{V}(u) := \bigcup_{q \in \prod_{k=1}^{K}\{\underline{q}_k, \bar{q}_k\}} \left(\{q\} \times \prod_{k=1}^{K} \{\underline{n}_k(q), \overline{n}_k\} \right) \qquad (5.12)
$$

where

$$
\underline{q}_k := \begin{cases} \infty & \text{if } v_k > \mathsf{R}_k \\ 0 & \text{o.w.} \end{cases} \qquad k = 1, \ldots, K, \qquad (5.13a)
$$

$$
\underline{n}_1(q) := \begin{cases} v_1/\alpha_1 & \text{if } q_1 = 0 \\ \mathsf{F}_1/\alpha_1, & \text{o.w.} \end{cases} \qquad (5.13b)
$$

$$
\underline{n}_k(q) := \begin{cases} \min\{\rho_{k-1}\underline{n}_{k-1} + v_k/\alpha_k, \mathsf{F}_k/\alpha_k\} & \text{if } q_k = 0 \\ \min\{\rho_{k-1}\underline{n}_{k-1} + \mathsf{R}_k/\alpha_k, \mathsf{F}_k/\alpha_k\} & \text{o.w.} \end{cases} \qquad k = 2, \ldots, K, \qquad (5.13c)
$$

$$
\overline{n}_K := n_K^{\max} - \min_i F_K(i)/\beta_K, \qquad (5.13d)
$$

$$
\begin{aligned}
\overline{n}_k &:= \min\left\{ \frac{\mathsf{F}_k}{\alpha_k}, n_k^{\max} - \frac{\min_i F_k(i)}{\beta_k}, n_k^{\max} - \frac{\beta_{k+1}(n_{k+1}^{\max} - \overline{n}_{k+1}) - v_{k+1}w_{k+1}}{\rho_k\beta_k} \right\} \\
& k = K-1, K-2, \ldots, 1, \qquad (5.13e)
\end{aligned}
$$

$$
\begin{aligned}
\bar{q}_k &:= \begin{cases} 0 & \text{if } v_k \leq \mathsf{R}_k \text{ and } v_k \leq \beta_k(n_k^{\max} - \bar{n}_k) - (1 - w_k)\rho_{k-1}\min\{\alpha_{k-1}\bar{n}_{k-1}, \mathsf{F}_k\} \\ \infty & \text{o.w.} \end{cases} \\
& k = 1, \ldots, K. \qquad (5.13f)
\end{aligned}
$$

Note that $\underline{q}_k$, $\bar{q}_k$, $\underline{n}_k$, $\bar{n}_k$ all depend on the control input u; for notational convenience we do not explicitly write them as functions of u. We will elaborate on the interpretation of $\mathcal{V}$ after stating Theorem 5.1. Furthermore, define

$$\mathcal{V}_0(u) = \{(q, n) \in \mathcal{V}(u) : q = 0\}, \tag{5.14a}$$

$$\mathcal{V}_1(u) = \{(q, n) \in \mathcal{V}(u) : q \neq 0\}. \tag{5.14b}$$

In addition, let us define e to be a K-dimensional vector of 1's, and define a constant matrix $D = (d_{kh}) \in \mathbb{R}^{K \times K}$ such that

$$d_{kh} = \begin{cases} 1 & \text{if } k = h = 1 \text{ or } h = k - 1, \\ 0 & \text{o.w.} \end{cases} \tag{5.15}$$

Then, we can state the sufficient condition for stability as follows.

Theorem 5.1 (Stability condition). *Consider a K-cell SS-CTM with a set of modes $\mathcal{I}$ with a demand vector $d \in \mathbb{R}^K_{\geq 0}$. For a given control input $u = (v, w) \in [0, d] \times \{0, 1\}^K$, the SS-CTM is stable in the sense of (5.8) if there exist a symmetric $K \times K$ matrix A satisfying*

$$a_{k,h} \geq a_{k+1,h} \quad k = 1, \ldots, K-1, \; h = 1, \ldots, K, \quad a_{K,K} > 0 \tag{5.16}$$

and a set of K-dimensional vectors $\{b^{(i)}; i \in \mathcal{I}\}$, which jointly verify the following set of inequalities linear in A and $b^{(i)}$:

$$A\Big(DG(i, x, u) + H(i, x, u)\Big) + \sum_{j \in \mathcal{I}} \nu_{ij} \left(b^{(j)} - b^{(i)}\right) \leq -e \quad \forall (i, x) \in \mathcal{I} \times \mathcal{V}_1(u). \tag{5.17}$$

For a given control input u, the vector fields G and H are fixed, and thus the inequalities (5.17) are linear in A and $b^{(i)}$. In addition, the cardinality of the set $\mathcal{I} \times \mathcal{V}_1$ is upper-bounded by $2^{2K}m$. That is, checking (5.17) entails checking no more than $2^{2K}m$ inequalities. Furthermore, for practical instances of the SS-CTM, the cardinality of $\mathcal{I} \times \mathcal{V}_1$ is typically smaller than this upper bound, and therefore the model is still tractable as we show in the subsequent Examples 5.2 and 5.3. As we will show in the next section, (5.17) is essentially a system of either linear or bilinear inequalities (depending on whether A is variable or not) and can be solved using existing computational tools [97]. In this chapter, solutions of inequalities are obtained using YALMIP [71].

Theorem 5.1 specializes a more general result on the stability of continuous-time Markov processes [74]. To conclude stability, the generic result in [74] (to be recalled in Section 5.2.2 as the "Foster-Lyapunov criterion") requires that a "drift condition" is verified everywhere over the hybrid state space $\mathcal{I} \times (\mathcal{Q} \times \mathcal{N})$. Essentially, the drift condition involves checking that the time derivative of an appropriately chosen Lyapunov function is negative in expectation for those states far away from the "origin", i.e. the states (q, n) such that $q_k = 0$ for each k. To prove Theorem 5.1, we choose a switched quadratic Lyapunov function $V : \mathcal{I} \times (\mathcal{Q} \times \mathcal{N}) \to \mathbb{R}$ that is specifically tailored to the SS-CTM:

$$V(i, x) := \frac{1}{2}(Dq + n)^T A(Dq + n) + (b^{(i)})^T (Dq + n). \tag{5.18}$$

We require the matrix A in the above to satisfy (5.16) to ensure that

(i) V is norm-like, i.e. $\lim_{\|x\|\to\infty} V(i, x) = \infty$ for each $i \in \mathcal{I}$, and

(ii) V decreases as traffic moves downstream through the highway, i.e. for each $i \in \mathcal{I}$, each $k \in \{1, \ldots, K-1\}$, each $\delta > 0$, and each $(q, n), (q', n') \in \mathcal{Q} \times \mathcal{N}$ such that

$$\begin{aligned} &q = q', \\ &n_k = n'_k + \delta,\ n_{k+1} = n'_{k+1} - \delta,\ n_h = n'_h\ \forall h \notin \{k, k+1\} \end{aligned}$$

we have $V(i, q, n) > V(i, q', n')$.

Recall from Fig. 5-1 that both buffer k and cell $(k-1)$ are immediately upstream of cell k; thus, the traffic out of buffer k and the traffic out of cell $k-1$ merge with each other and cannot be further distinguished in our model. This feature of the SS-CTM is captured by the Lyapunov function which equally penalizes q_k and n_{k-1} for $k = 2, \dots, K$ thanks to the structure of D. Also note that the vector $b^{(i)}$ depends on the mode i while the matrix A does not; thus, the second terms in (5.18) captures the impact of mode transitions. Finally, we do not need to restrict the range of $b^{(i)}$, since only the differences between them is involved in (5.17); one can always set $b^{(i)} \geq 0$ to ensure that $V \geq 0$ (if necessary).

By choosing the Lyapunov function as defined in (5.18), we can conclude that verifying the drift condition reduces to checking the feasibility of the system of linear inequalities (5.17). More importantly, we only require checking feasibility of this system for the finite number of states in the set $\mathcal{V}$. Note that straightforward use of the generic result in [74] for the SS-CTM would require checking the feasibility of (5.17) everywhere over $\mathcal{Q} \times \mathcal{N}$, which essentially requires maximizing the left-hand side of (5.17) over $\mathcal{Q} \times \mathcal{N}$; this maximization is a non-linear and non-convex optimization problem. Theorem 5.1 addresses this challenge by exploiting the cooperative property (in the sense of [39]) of the SS-CTM dynamics in each mode. Using to this property, we are able to find a globally attracting and positively invariant set of the continuous state $x = (q, n)$ as follows:

$$\mathcal{M}(u) := \bigcup_{q \in \prod_{k=1}^{K} [\underline{q}_k, \bar{q}_k]} \left(\{q\} \times \prod_{k=1}^{K} [\underline{n}_k(q), \overline{n}_k] \right). \tag{5.19}$$

Note that $\mathcal{M}(u)$ only involves the mode-specific capacities $F(i)$ (see (5.13b)–(5.13f)), but is independent of the transition rates ν_{ij}.

One can see from (5.18) that $\mathcal{V}$ is the (finite) set of vertices of $\mathcal{M}$. We also say $\mathcal{V}$ to be a set of *critical states*, since they represent typical congestion patterns that can recurrently happen due to capacity perturbations. For example, if $\underline{n}_k < \mathsf{F}_k/\alpha_k \leq \bar{n}_k$, then we can conclude that cell k recurrently switches between congestion and free flow. By looking at the critical states, we can identify the bottlenecks that capacity perturbations can recurrently induce. We emphasize here that a bottleneck is not always a location of capacity perturbation. Due to spillback, even after a perturbation clears, traffic can still be stuck in an upstream cell/buffer, which we call an *induced bottleneck.* The critical states clearly show where the induced bottlenecks are. Example 5.1 illustrates this point.

Remark 5.2. *$\mathcal{M}(u)$ is a generalization of the invariant set proposed in [45], which does not consider the impact of on-ramp queues $q_2, \dots, q_K$.*

To prove Theorem 5.1, we show that $\mathcal{M}$ is globally attracting and positively invariant (Section 5.2.1) and then argue that verification of the drift condition reduces to finding a feasible solution to (5.17) (Section 5.2.2).

5.2.1 Invariant set

Following [9], a set $\mathcal{A} \subseteq \mathcal{Q} \times \mathcal{N}$ is a globally attracting and positively invariant set for the continuous state (q, n) if

$$\begin{aligned}
&\text{(Attracting)} \quad \forall (I(0), Q(0), N(0)) \in \mathcal{I} \times \mathcal{Q} \times \mathcal{N},\ \forall \epsilon > 0,\ \exists T \geq 0,\ \forall t \geq T, \\
&\qquad\qquad \min_{(q,n)\in\mathcal{A}} \|(Q(t), N(t)) - (q, n)\|_2 \leq \epsilon, \qquad \text{(5.20a)} \\
&\text{(Invariant)} \quad \forall (I(0), Q(0), N(0)) \in \mathcal{I} \times \mathcal{A},\ \forall t \geq 0,\ (Q(t), N(t)) \in \mathcal{A}. \qquad \text{(5.20b)}
\end{aligned}$$

Intuitively, $\mathcal{A}$ is a set of states such that, for all initial conditions $(i, q, n) \in \mathcal{I} \times \mathcal{Q} \times \mathcal{N}$, the process $(I(t), Q(t), N(t))$ eventually enters and does not leave $\mathcal{A}$. For convenience, we henceforth refer to any set satisfying (5.20a)–(5.20b) simply as an *invariant set.*

Lemma 5.1. *The set $\mathcal{M}(u)$ as defined in* (5.19) *is globally attracting and positively invariant.*

With this result, for stability analysis, we can restrict our attention to the evolution of the continuous states over the invariant set $\mathcal{M}$ instead of the entire continuous state space $\mathcal{Q} \times \mathcal{N}$.

Proof of Lemma 5.1. Consider a given control input $u = (v, w)$. One can adapt the proof of [45, Proposition 1] and show that the set $\hat{\mathcal{M}} = [0, \infty]^K \times \prod_{k=1}^{K} [\underline{n}_k(0), \overline{n}_k]$ is globally attracting and positively invariant. Note that $\mathcal{M} \subseteq \hat{\mathcal{M}}$. In this proof, we refine $\hat{\mathcal{M}}$ and eventually obtain the invariance of $\mathcal{M}$. The refinement is done in three steps:

1. $[0,\infty]^K \times \prod_{k=1}^K [\underline{n}_k(0), \bar{n}_k]$ to $\prod_{k=1}^K [\underline{q}_k, \infty] \times \prod_{k=1}^K [\underline{n}_k(0), \bar{n}_k]$,
2. $\prod_{k=1}^K [\underline{q}_k, \infty] \times \prod_{k=1}^K [\underline{n}_k(0), \bar{n}_k]$ to $\prod_{k=1}^K [\underline{q}_k, \bar{q}_k] \times \prod_{k=1}^K [\underline{n}_k(0), \bar{n}_k]$,
3. $\prod_{k=1}^K [\underline{q}_k, \bar{q}_k] \times \prod_{k=1}^K [\underline{n}_k(0), \bar{n}_k]$ to $\bigcup_{\prod_{k=1}^K [\underline{q}_k, \bar{q}_k]} \{q\} \times \prod_{k=1}^K [\underline{n}_k(q), \bar{n}_k]$

Step (1). For the k-th buffer, if $v_k > R_k$ (and thus $\underline{q}_k = \infty$ according to (5.13a)), then for any initial condition $(i, x, u) \in \mathcal{I} \times \hat{\mathcal{M}}$, we have $\lim_{t\to\infty} Q_k(t) = \infty$. Hence, the following set

$$[0,\infty]^{k-1} \times \{\infty\} \times [0,\infty]^{K-k} \times \prod_{h=1}^K [\underline{n}_h(0), \bar{n}_h]$$

is attracting and invariant. Repeating the above argument for each k, we conclude that the following set

$$\hat{\mathcal{M}}_1 = \prod_{k=1}^K [\underline{q}_k, \infty] \times \prod_{k=1}^K [\underline{n}_k(0), \bar{n}_k]$$

is attracting and invariant.

Step (2). For the k-th buffer, if $v_k < R_k$ and $v_k < \beta_k(n_k^{\max} - \bar{n}_k) - (1 - w_k)\rho_{k-1} \min\{\alpha_{k-1}\bar{n}_{k-1}, F_{k-1}\}$ (and thus $\bar{q}_k = 0$), we have

$$\begin{aligned}
G_k(i,x,u) &= v_k - r(i,x,u) = v_k - \min\{\mathbf{1}_{\{q=0\}} v_k + \mathbf{1}_{\{q>0\}} R_k, \beta_k(n_k^{\max} - n_k) - \rho_{k-1} f_{k-1}(i,x,u)\} \\
&\le v_k - \min\{\mathbf{1}_{\{q=0\}} v_k + \mathbf{1}_{\{q>0\}} R_k, \beta_k(n_k^{\max} - \bar{n}_k) \\
&\quad - (1 - w_k)\rho_{k-1} \min\{\alpha_{k-1}\bar{n}_{k-1}, F_{k-1}\}\} \quad \text{if } n \in \prod_{h=1}^K [\underline{n}_h(0), \bar{n}_h] \\
&< 0 \quad \text{if } q_k > 0.
\end{aligned}$$

Hence, the set

$$[0,\infty]^{k-1} \times \{0\} \times [0,\infty]^{K-k} \times \prod_{k=1}^K [\underline{n}_k(0), \bar{n}_k]$$

is attracting and invariant. Repeating the above argument for each k, we conclude that the following set

$$\hat{\mathcal{M}}_2 = \prod_{k=1}^{K} [0, \bar{q}_k] \times \prod_{k=1}^{K} [\underline{n}_k(0), \bar{n}_k]$$

is attracting and invariant. Then, we conclude that $\hat{\mathcal{M}}_1 \cap \hat{\mathcal{M}}_2$ is attracting and invariant.

Step (3). We complete this step by showing (i) $\mathcal{M}$ is attracting for each initial condition $(q, n) \in (\hat{\mathcal{M}}_1 \cap \hat{\mathcal{M}}_2)$, and (ii) $\mathcal{M}$ is invariant.

Step (3i). First, we consider the case that $\bar{q}_1 = \infty$. For each $q \in (\hat{\mathcal{M}}_1 \cap \hat{\mathcal{M}}_2)$ such that $q_1 > 0$, we obtain from (5.13b) that

$$\underline{n}_1(q) = \mathsf{R}_1 / \alpha_1.$$

Then, for each $i \in \mathcal{I}$ and each $(q, n) \in \{(\xi, \zeta) \in (\hat{\mathcal{M}}_1 \cap \hat{\mathcal{M}}_2) : \xi_1 > 0, \zeta_1 < \underline{n}_1(\xi)\}$, we have

$$\begin{aligned} H_1(i, x, u) &= \mathsf{R}_1 - \min\left\{\alpha_1 n_1, \mathsf{F}_1(i), \frac{\beta_2(n_2^{\max} - n_2) - r_2(q, n)}{\rho_1}\right\} \geq \mathsf{R}_1 - \alpha_1 n_1 \\ &> \mathsf{R}_1 - \alpha_1 \underline{n}_1(q) = 0. \end{aligned}$$

That is, for each $(q, n) \in \{(\xi, \zeta) \in (\hat{\mathcal{M}}_1 \cap \hat{\mathcal{M}}_2) : \xi_1 > 0, \zeta_1 < \underline{n}_1(\xi)\}$, the vector field H has a strictly positive component pointing to the set

$$[\underline{n}_1(q), \bar{n}_1] \times \prod_{k=2}^{K} [\underline{n}_k(0), \bar{n}_k].$$

Therefore, the set

$$\mathcal{M}_1^1 = \left(\bigcup_{\substack{q \in (\hat{\mathcal{M}}_1 \cap \hat{\mathcal{M}}_2): \\ q_1 = 0}} \left(\{q\} \times \prod_{k=1}^{K} [\underline{n}_k(0), \bar{n}_k] \right) \right) \cup \left(\bigcup_{\substack{q \in (\hat{\mathcal{M}}_1 \cap \hat{\mathcal{M}}_2): \\ q_1 > 0}} \left(\{q\} \times [\underline{n}_1(q), \bar{n}_1] \times \prod_{k=2}^{K} [\underline{n}_k(0), \bar{n}_k] \right) \right).$$

is attracting, where the subscript $e_1 = [1\ 0\ \cdots\ 0]^T$.

Next, for each $i \in \mathcal{I}$ and each $(q,n) \in \{(\xi,\zeta) \in \mathcal{M}_1^1 : \xi_1 > 0, \zeta_2 < \underline{n}_2(\xi)\}$, we have

$$\begin{aligned} H_2(i,x,u) &= \rho_1 f_1(i,x,u) + r_2(i,x,u) - f_2(i,x,u) \\ &= \min\left\{\rho_1 \min\{\alpha_1 n_1, \mathsf{F}_1(i)\} + r_2(q,n), \beta_2(n_2^{\max} - n_2)\right\} \\ &\quad - \min\left\{\alpha_2 n_2, \mathsf{F}_2(i), \frac{\beta_3(n_3^{\max} - n_3) - r_3(i,x,u)}{\rho_2}\right\} \\ &\geq \rho_1\alpha_1\underline{n}_1(q) + v_2 - \alpha_2 n_2 > \rho_1 v \underline{n}_1(q) + v_2 - \alpha_2\underline{n}_2(q) \overset{(5.13b)(5.13c)}{=} 0. \end{aligned}$$

That is, for each $i \in \mathcal{I}$ and each $(q,n) \in \{(\xi,\zeta) \in \mathcal{M}_1^1 : \xi_1 > 0, \zeta_2 < \underline{n}_2(\xi)\}$, the vector field H has a strictly positive component pointing to the set

$$\prod_{k=1}^{2}[\underline{n}_k(q), \overline{n}_1] \times \prod_{k=3}^{K}[\underline{n}_k(0), \overline{n}_k].$$

Therefore, the set

$$\begin{aligned} \mathcal{M}_1^2 = &\left(\bigcup_{\substack{q\in(\hat{\mathcal{M}}_1\cap\hat{\mathcal{M}}_2):\\ q_1=0}} \left(\{q\} \times \prod_{k=1}^{K}[\underline{n}_k(0), \overline{n}_k]\right)\right) \\ &\cup \left(\bigcup_{\substack{q\in(\hat{\mathcal{M}}_1\cap\hat{\mathcal{M}}_2):\\ q_1>0}} \left(\{q\} \times \prod_{k=1}^{2}[\underline{n}_1(q), \overline{n}_1] \times \prod_{k=3}^{K}[\underline{n}_k(0), \overline{n}_k]\right)\right). \end{aligned}$$

is attracting.

Similarly, for each $i \in \mathcal{I}$ and each $(q,n) \in \mathcal{M}_1^h$, we can show that H has a strictly positive component pointing to the set $\prod_{k=1}^{h+1}[\underline{n}_k(q), \overline{n}_k] \times \prod_{k=k_1+2}^{K}[\underline{n}_k(0), \overline{n}_k]$, and thus the set

$$\mathcal{M}_1^h = \left(\bigcup_{\substack{q\in(\hat{\mathcal{M}}_1\cap\hat{\mathcal{M}}_2):\\ q_1=0}} \left(\{q\} \times \prod_{k=1}^{K}[\underline{n}_k(0), \overline{n}_k]\right)\right)$$

$$\cup \left(\bigcup_{\substack{q \in (\hat{\mathcal{M}}_1 \cap \hat{\mathcal{M}}_2): \\ q_1 > 0}} \left(\{q\} \times \prod_{k=1}^{h} [\underline{n}_1(q), \overline{n}_1] \times \prod_{k=h+1}^{K} [\underline{n}_k(0), \overline{n}_k] \right) \right).$$

is attracting. Repeating this argument, we obtain that the set

$$\mathcal{M}_1^K = \left(\bigcup_{\substack{q \in (\hat{\mathcal{M}}_1 \cap \hat{\mathcal{M}}_2): \\ q_1 = 0}} \left(\{q\} \times \prod_{k=1}^{K} [\underline{n}_k(0), \overline{n}_k] \right) \right) \cup \left(\bigcup_{\substack{q \in (\hat{\mathcal{M}}_1 \cap \hat{\mathcal{M}}_2): \\ q_1 > 0}} \left(\{q\} \times \prod_{k=1}^{K} [\underline{n}_1(q), \overline{n}_1] \right) \right).$$

is attracting.

By analogous arguments, we can show that, for every h such that $\bar{q}_h = \infty$, the set

$$\mathcal{M}_h^K = \left(\bigcup_{\substack{q \in (\hat{\mathcal{M}}_1 \cap \hat{\mathcal{M}}_2): \\ q_1 = 0}} \left(\{q\} \times \prod_{k=1}^{K} [\underline{n}_k(0), \overline{n}_k] \right) \right) \cup \left(\bigcup_{\substack{q \in (\hat{\mathcal{M}}_1 \cap \hat{\mathcal{M}}_2): \\ q_h > 0}} \left(\{q\} \times \prod_{k=1}^{K} [\underline{n}_1(q), \overline{n}_1] \right) \right).$$

is attracting.

Since $\mathcal{M} = \cap_{k=1}^{K} \mathcal{M}_k^K$, we conclude that $\mathcal{M}$ is also attracting.

Step (3ii): One can prove the invariance of $\mathcal{M}$ by adapting the proof of [45, Proposition 12] and considering the direction of the vector field H on the boundary of $\mathcal{M}$. ■

Note that the invariant set $\mathcal{M}$ may or may not be bounded: $\mathcal{M}$ is bounded if and only if $\bar{q}_k = 0$ for all k. A bounded invariant set immediately imply stability of the SS-CTM. Thus, we only need to focus on the case where $\mathcal{M}$ is unbounded, i.e. $\bar{q}_k = \infty$ for some k; see (5.13f). If the invariant set $\mathcal{M}$ is unbounded, then we need to show that the drift condition holds everywhere over the invariant set $\mathcal{M}$.

Next, we interpret the invariant set $\mathcal{M}$ as well as the critical states $\mathcal{V}$ via a practical example.

Example 5.1. *In this example, we consider the South Mountain Avenue on-ramp to I210 East-bound in Monrovia, California, USA as shown in Fig. 5-15. Suppose that the mainline is subject to capacity perturbation that randomly occur and terminates.*

The average time interval between two incidents is 1 hour, and the average duration of an incident is 0.5 hour. The mainline nominal capacity is 9850 veh/hr, the capacity reduction is 1970 veh/hr, and the on-ramp capacity is 3000 veh/hr. Fig. 5-4 shows the SS-CTM for this highway section. Calibrated parameters are listed in Tab. 5.1; see [42] for details about the simulation model. For ease of presentation, we only illustrate the invariant set and critical states in the case without metering (i.e. $w_2 = 1$*); the case of* $w_2 = 0$ *can be similarly illustrated.*

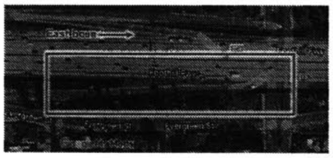


Figure 5-3: Traffic control for a stationary hotspot on I210.

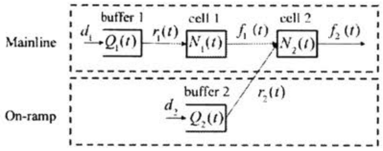


Figure 5-4: The SS-CTM for the highway section in Fig. 5-12.

Table 5.1: Parameters of the SS-CTM shown Fig. 5-4.

	α_k	F_k, R_k	Δ_k	$n_k^{\max}$
Cell 1	108 km/hr	7360 veh/hr	1840 veh/hr	494 veh/km
Cell 2	111 km/hr	9850 veh/hr	1970 veh/hr	661 veh/km
Buffer 1	N/A	7360 veh/hr	N/A	N/A
Buffer 2	N/A	3000 veh/hr	N/A	N/A

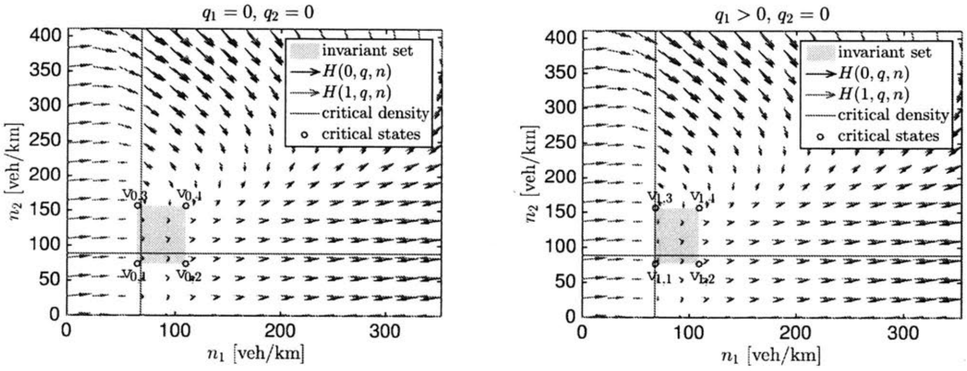
Suppose that cell 2 is a stationary hotspot for perturbation, while cell 1 has a constant capacity. That is, the two-cell model has two modes $\mathcal{I} = \{0, 1\}$*, and the mode-specific cell capacities are*

$$F(0) = [7360\ 9850]^T, \quad F(1) = [7360\ 7880]^T.$$

Furthermore, we assume the mode transition rates to be

$$\lambda = 1[hr^{-1}], \quad \mu = 1[hr^{-1}],$$

where λ (resp. μ) is the occurrence (resp. clearance) rate of capacity perturbations; see Fig. 5-2(a). The control input $v = [7000\ 1000]^T$, $w = [1\ 1]^T$ is used in this example; we will show in Example 5.2 that this control input is stabilizing.



(a) $\mathcal{M}_0 = \{[0\ 0]^T\} \times \prod_{k=1}^{2} [\underline{n}_k(0), \bar{n}_k]$

(b) $\mathcal{M}_1 = ((0, \infty] \times \{0\}) \times \prod_{k=1}^{2} [\underline{n}_k([\infty, 0]^T), \bar{n}_k]$

Figure 5-5: The vector field, the invariant set, and the critical states of the two-cell model with a stationary hotspot.

One can obtain from (5.13a) and (5.13f) that $\underline{q}_1 = \underline{q}_2 = \bar{q}_2 = 0$, while $\bar{q}_1 = \infty$. Fig. 5-5 shows the vector field $H(i, x, u)$ (governing the evolution of traffic densities n) as well as the set $[\underline{n}_1, \bar{n}_1] \times [\underline{n}_2, \bar{n}_2]$ on the two-dimensional plane $\mathcal{N} = [0, n_1^{\max}] \times [0, n_2^{\max}]$.[2] *Hence, the invariant set is*

$$\mathcal{M}(u) = \left(\{[0\ 0]^T\} \times \prod_{k=1}^{2} [\underline{n}_k([0\ 0]^T), \bar{n}_k]\right) \cup \left(\Big((0, \infty] \times \{0\}\Big) \times \prod_{k=1}^{2} [\underline{n}_k([\infty\ 0]^T), \bar{n}_k]\right),$$

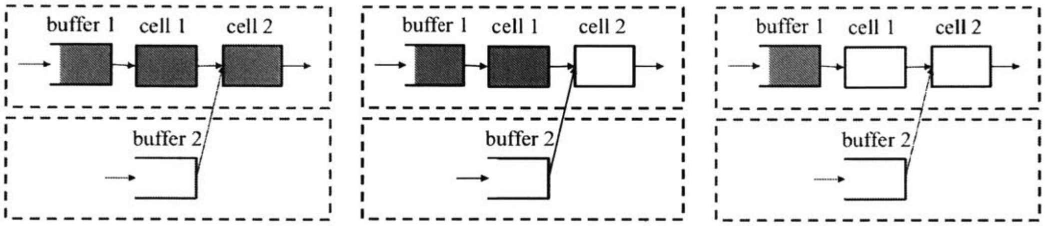
and the sets of critical states are

$$\mathcal{V}_0(u) = \{[0\ 0]^T\} \times \prod_{k=1}^{2} \{\underline{n}_k([0\ 0]^T), \bar{n}_k\} = \{\mathsf{v}_{0,1}, \mathsf{v}_{0,2}, \mathsf{v}_{0,3}, \mathsf{v}_{0,4}\},$$

[2] Visualization of the vector field $G(i, x, u)$ over the plane $\mathcal{Q} = [0, \infty]^2$ is not easy, since G also depends on the specific value of n.

$$\mathcal{V}_1(u) = \{[\infty\ 0]^T\} \times \prod_{k=1}^{2}\{\underline{n}_k([\infty\ 0]^T), \bar{n}_k\} = \{\mathsf{v}_{1,1}, \mathsf{v}_{1,2}, \mathsf{v}_{1,3}, \mathsf{v}_{1,4}\},$$

where $\mathsf{v}_{0,1}$–$\mathsf{v}_{1,4}$ are indicated in Fig. 5-5. This invariant set $\mathcal{M}$ is unbounded. To see this, note that the total inflow v_1+v_2 is greater than the perturbed capacity $(\mathsf{F}_2-\Delta_2)$; hence, if the capacity perturbation lasts sufficiently long, which can always happen with a positive probability (as long as $\mu > 0$), then the total number of vehicles $|Q(t)|+|N(t)|$ in the system can grow arbitrarily large.



(a) Possible $(\mathsf{v}_{1,3}, \mathsf{v}_{1,4})$. (b) Possible $(\mathsf{v}_{1,1}, \mathsf{v}_{1,2})$. (c) Impossible.

Figure 5-6: Congestion patterns associated with the critical states indicated in Fig. 5-5(b). Shaded cells/buffers are congested.

The critical states imply the following. First, cell 2 may be recurrently congested, which means that cell 2 is a bottleneck (see Fig. 5-6(a)); this situation is captured by the trivial states in Fig. 5-5(a) and non-trivial critical states $\mathsf{v}_{1,3}$ and $\mathsf{v}_{1,4}$ in Fig. 5-5(b). Second, it recurrently happens that cell 1 is congested but cell 2 is not; that is, although cell 1 does not directly experience capacity perturbations, this cell is an induced bottleneck. This situation is captured by the critical state $\mathsf{v}_{1,1}$ and $\mathsf{v}_{1,2}$ in Fig. 5-5(b) and illustrated in Fig. 5-6(b). Third, if buffer 1 is congested, cell 1 is necessarily congested. That is, buffer 1 is never a bottleneck, and the situation in Fig. 5-6(c) does not happen.

5.2.2 Verification of drift condition

Now we verify the drift condition for the Lyapunov function V defined in (5.18) for the SS-CTM dynamics and utilize the Foster-Lyapunov criterion to obtain stability. To formally state the Foster-Lyapunov criterion, we recall that the evolution of the process $(I(t), X(t)) = (I(t), Q(t), N(t))$ is captured by the *infinitesimal generator* of the traffic flow dynamics [45]. Since $I(t)$ is a Markov chain and since $Q(t)$ and $N(t)$ are always continuous in t, this process is *right-continuous with left limits (RCLL).* Hence, by [9, Proposition 2.1], for a given control input $u = (v, w)$, the infinitesimal generator can be written as an operator $\mathcal{L}$ as follows:

$$\mathcal{L}V(i,x) = G^T(i,x,u)\nabla_q V(i,x) + H^T(i,x,u)\nabla_n V(i,x) + \sum_{j\in\mathcal{I}} \nu_{ij}\Big(V(j,x) - V(i,x)\Big)$$
$$(i,x) \in \mathcal{I} \times (\mathcal{Q} \times \mathcal{N}), \tag{5.21}$$

where $\nabla_q V$ (resp. $\nabla_n V$) is the gradient of V with respect to q (resp. n). With the above definition, we can state the generic result [74, Theorem 4.3] as follows:

Theorem (Foster Lyapunov criterion [74]). *For an RCLL Markov process with state ξ and invariant set Ξ, if there exist a norm-like function $V : \Xi \to \mathbb{R}_{\geq 0}$, a function $g : \Xi \to \mathbb{R}_{\geq 0}$, and constants $\epsilon > 0$ and $Z < \infty$ such that*

$$(\textit{Drift condition}) \quad \mathcal{L}V(\xi) \leq -\epsilon g(\xi) + Z \quad \forall \xi \in \Xi, \tag{5.22}$$

then, for each initial condition $\xi \in \Xi$,

$$\limsup_{t\to\infty} \frac{1}{t} \int_{s=0}^{t} \mathcal{E}[g(\xi(s))]ds \leq Z/\epsilon. \tag{5.23}$$

It turns out that we can conclude that the SS-CTM is stable if the drift condition (5.22) can be verified over the set of critical states $\mathcal{V}$ instead of the invariant set $\mathcal{M}$. The key is to show that $\mathcal{V}$ contains the states where the expected time derivative of the Lyapunov function attains its maximum, i.e. where the rate of traffic discharge is minimal. Verifying the drift condition is straightforward over $\mathcal{V}_0$, but requires more work over $\mathcal{V}_1$.

Our task is to show that (5.17) is sufficient for verifying the drift condition (5.22) with $g(i,x) = |q|$. Suppose that there exist a matrix A and vectors $b^{(i)}$ that satisfy the conditions in the statement of Theorem 5.1. Consider the invariant set $\mathcal{M}$ as defined in (5.19). By (5.2b)–(5.2c) and (5.21), for a given control input $u = (v, w)$, we have

$$\begin{aligned}\mathcal{L}V(i,x) &= (DG+H)^T A(Dq+n) + (b^{(i)})^T(DG+H) + \sum_{i\in\mathcal{I}} \nu_{ij}\left(b^{(j)} - b^{(i)}\right)^T (Dq+n) \\ &= \left((DG+H)^T A + \sum_{i\in\mathcal{I}} \nu_{ij}\left(b^{(j)} - b^{(i)}\right)^T\right)(Dq+n) + (b^{(i)})^T(DG+H).\end{aligned} \tag{5.24}$$

Since G_k and H_k are bounded, we can define a constant

$$W := \left(\max_{i,k}|b_k^{(i)}|\right) \sum_{k=1}^{K} (\mathsf{R}_k + \rho_k \mathsf{F}_k)$$

and obtain that

$$\begin{aligned}(b^{(i)})^T(DG(i,x,u) + H(i,x,u)) &\overset{(5.15)}{\leq} \max_{i,x}|b_k^{(i)}|\sum_{k=1}^{K}(G_k + H_k) \\ &\overset{(5.7\text{a})-(5.7\text{c})}{\leq} \max_{i,x}|b_k^{(i)}|\sum_{k=1}^{K}(v_k - r_k(x,u) + \rho_{k-1} f_{k-1}(i,x,u) + r_k(x,u) - f_k(i,x,u) \\ &\overset{(5.6\text{a})-(5.6\text{c})}{\leq} \left(\max_{i,k}|b_k^{(i)}|\right)\sum_{k=1}^{K}(\mathsf{R}_k + \rho_k \mathsf{F}_k) = W \quad \forall (i,x) \in \mathcal{I} \times (\mathcal{Q} \times \mathcal{N}).\end{aligned} \tag{5.25}$$

Here, we utilize a technical result;

Lemma 5.2. *If* (5.17) *holds, then* $\underline{q}_K = 0$ *for* $q = 1, \ldots, K$.

Proof. Suppose that (5.17) holds for a given control input $u = (v, w)$. Then, the invariant set is $\mathcal{M}(u)$ and the set of vertices is $\mathcal{V}(u)$. Assume by contradiction that there exists k such that $\underline{q}_K = \infty$. By (5.13a), this means that $v_k \geq R_k$. Consider the state $x' = (q', n') = ([\underline{q}_1 \ \cdots \ \underline{q}_K]^T, [\underline{n}_1 \ \cdots \ \underline{n}_K]^T)$. Since $\underline{q}_K = \infty$, $(q', n') \in \mathcal{V}_1$. By the proof of Lemma 5.1, $G_h(i, x', u) \geq 0$ and $H_h(i, x', u) \geq 0$ for $h = 1, \ldots, K$ and for each $i \in \mathcal{I}$. Then, for each $h \in \{1, \ldots, K\}$, letting $i' = \arg\min_j b_h^{(j)}$, we can expand the hth row of (5.17) as follows:

$$\begin{aligned} &a_{h,1}(G_1(i',x',u) + G_2(i',x',u) + H_1(i',x',u)) + \sum_{\ell=2}^{K} a_{h,\ell}(G_{\ell+1}(i',x',u) + H_\ell(i',x',u)) \\ &\quad + \sum_{j\in\mathcal{I}} \nu_{i'j}(b_h^{(j)} - b_h^{(i')}) \\ &\quad \geq 0 + \sum_{j\in\mathcal{I}} \nu_{i'j}(b_h^{(j)} - b_h^{(i')}) \geq \sum_{j\in\mathcal{I}} \nu_{i'j}(b_h^{(i')} - b_h^{(i')}) = 0, \end{aligned}$$

which contradicts (5.17). ■

Thus, we can partition $\mathcal{M}$ into two subsets:

$$\begin{aligned} \mathcal{M}_0(u) &:= \{(q,n) \in \mathcal{M}(u) : q = 0\}, \\ \mathcal{M}_1(u) &:= \{(q,n) \in \mathcal{M}(u) : q \neq 0\}. \end{aligned}$$

Note that $\mathcal{V}_0$ and $\mathcal{V}_1$ as defined in (5.14a) and (5.14b), respectively are the vertices of $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively.

The rest of this proof has two steps for verifying the drift condition over $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively:

Step 1: For each $(q,n) \in \mathcal{M}_0$, since $\mathcal{M}_0$ is a bounded set, there exists $Z_0 < \infty$ such that

$$\left((DG+H)^T A + \sum_{j\in\mathcal{I}} \nu_{ij}\left(b^{(j)} - b^{(i)}\right)^T\right)(Dq+n) \leq Z_0 \quad \forall (i,x) \in \mathcal{I}\times\mathcal{M}_0.$$

Hence, we can obtain from (5.25) and the above that

$$\mathcal{L}V(i,x) \leq Z_0 + W = -|q| + Z_0 + W \quad \forall (i,x) \in \mathcal{I}\times\mathcal{M}_0. \tag{5.26}$$

Step 2: For each $(q, n) \in \mathcal{M}_1$ and for $h = 1, \ldots, K$, note that

$$(DG + H)^T \begin{bmatrix} a_{h,1} \\ a_{h,2} \\ \vdots \\ a_{h,K} \end{bmatrix} = a_{h,1}v_1 + \sum_{k=2}^{K} a_{h,k-1}v_k - \sum_{k=1}^{K-1} \Big((a_{h,k} - \rho_k a_{h,k+1})f_k(i, x, u)$$

$$+ (a_{h,k} - a_{h,k+1})r_{k+1}(x, u)\Big) - a_{h,K}f_K(i, x, u)$$

$$\leq a_{h,1}v_1 + \sum_{k=2}^{K} a_{h,k-1}v_k - \sum_{k=1}^{K-1}(a_{h,k} - a_{h,k+1})\Big(\rho_k f_k(i, x, u) + r_{k+1}(x, u)\Big)$$

$$- a_{h,K}f_K(i, x, u).$$

Note that (5.16) ensures that $a_{h,k} - a_{h,k+1} \geq 0$ for $k = 1, \ldots, K-1$. For each $q \neq 0$, $\rho_{k-1}f_{k-1} + r_k$ is concave in n over the box $\prod_{k=1}^{K}[\underline{n}_k(q), \overline{n}_k]$ (see [45, Proof of Theorem 2] for details). Therefore, $\rho_{k-1}f_{k-1} + r_k$ attains its minimum at one of the vertices of the box, i.e.

$$\min_{n \in \prod_{k=1}^{K}[\underline{n}_k(q), \overline{n}_k]} \sum_{k=1}^{K-1}(a_{h,k} - a_{h,k+1})\Big(\rho_k f_k(i, x, u) + r_{k+1}(x, u)\Big) + a_{h,K}f_K(i, x, u)$$

$$= \min_{n \in \prod_{k=1}^{K}\{\underline{n}_k(q), \overline{n}_k\}} \sum_{k=1}^{K-1}(a_{h,k} - a_{h,k+1})\Big(\rho_k f_k(i, x, u) + r_{k+1}(x, u)\Big) + a_{h,K}f_K(i, x, u).$$

Then, we have

$$a_{k,1}v_1 + \sum_{k=2}^{K} a_{k,k-1}v_k - \sum_{k=1}^{K-1} \Big((a_{k,k} - \rho_k a_{k,k+1})f_k(i, q, n) + (a_{k,k} - a_{k,k+1})r_{k+1}(i, q, n)\Big)$$

$$- a_{k,K}f_K(i, q, n) + \sum_{j \in \mathcal{I}} \nu_{ij}(b_k^{(j)} - b_k^{(i)})$$

$$\leq a_{k,1}v_1 + \sum_{k=2}^{K} a_{k,k-1}v_k - \min_{(q,n) \in \mathcal{V}_1} \left(\sum_{k=1}^{K-1}(a_{k,k} - a_{k,k+1})(\rho_k f_k(i, x) + r_{k+1}(i, x)) + a_{k,K}f_k(i, x)\right)$$

$$+ \sum_{j \in \mathcal{I}} \nu_{ij}(b_k^{(j)} - b_k^{(i)}) \leq -1 \quad \forall (i, x) \in \mathcal{I} \times \mathcal{M}_1.$$

Thus, we can obtain from the above together with (5.24) and (5.25) that

$$\mathcal{L}V \leq -e^T(Dq+n)+W \quad \forall(i,x) \in \mathcal{I} \times \mathcal{M}_1.$$

Recalling the definition of D in (5.15), we have

$$\mathcal{L}V \leq -|q|+W \quad \forall(i,x) \in \mathcal{I} \times \mathcal{M}_1. \tag{5.27}$$

Finally, let $Z := Z_0 + W$, we can obtain from (5.26) and (5.27) that

$$\mathcal{L}V(i,x) \leq -|q|+Z \quad \forall(i,x) \in \mathcal{I} \times \mathcal{M},$$

which verifies the drift condition. This finishes the proof of Theorem 5.1.

5.2.3 Application of Theorem 5.1

Now we apply Theorem 5.1 to two particular scenarios of capacity perturbations, viz. stationary hotspots and moving bottlenecks as introduced in Section 5.1.1, and also evaluate the "sharpness" of Theorem 5.1, i.e. how large is the gap between Theorem 5.1 and a necessary condition for stability of the SS-CTM.

Following [45, Theorem 1], a necessary condition for stability of the SS-CTM is

$$\sum_{h=1}^{k} \rho_h^k v_h \leq \sum_{i \in \mathcal{I}} \mathsf{p}_i \min\left\{F_k(i), \frac{1}{\rho_k}\left(\beta_k(n_k^{\max} - \min_q \underline{n}_{k+1}(q)) - v_{k+1}w_{k+1}\right)\right\}$$

$$k = 1, \ldots, K-1, \tag{5.28a}$$

$$\sum_{h=1}^{K} \rho_h^K v_h \leq \sum_{i \in \mathcal{I}} \mathsf{p}_i F_K(i), \tag{5.28b}$$

$$v_k \leq \mathsf{R}_k \quad k = 2, \ldots, K. \tag{5.28c}$$

The necessary condition is based on the principle that if the on-ramp queues are bounded, then the inflows $v_1, \ldots, v_K$ do not exceed the cell/buffer capacity. In the setting of SS-CTM, this principle implies that for each cell (resp. buffer), the time-average inflow into the cell (resp. buffer) is no greater than the cell's (resp. buffer's) time-average capacity. Furthermore, (5.28a) is refined with respect to the simple average capacity, i.e. $\bar{F}_k = \sum_{i\in\mathcal{I}} \mathsf{p}_i F_k(i)$, to capture the impact of the receiving flow of the downstream cell.

In general, there is a gap between the sufficient condition (Theorem 5.1) and the necessary condition (5.28a)–(5.28c). For those control inputs lying in this gap, our stability conditions do not give a conclusive answer to whether the SS-CTM is stable or unstable. Factors affecting the size of the gap include (i) the particular form of the Lyapunov function V, (ii) the tightness of the invariant set $\mathcal{M}$, and (iii) the model parameters. Here, we only discuss about (i); some discussion on (ii) and (iii) is available in [45]. Particularly, we focus on how the structure of A affects the size of the gap, or the sharpness.

When applying Theorem 5.1, one can consider A in (5.17) as an unknown to be solved. For a given control input u, this involves solving a system of linear inequalities. Alternatively, A can also be explicitly constructed using insights about the SS-CTM dynamics. Although this practice may affect the sharpness of the sufficient condition, this will make (5.17) easier to check, and thus make the max-throughput problem (P) easier to solve. A simple construction is

$$a_{k,h} = \gamma \quad k = 1, \dots, K, \; h = 1, \dots, K, \tag{5.29}$$

where γ is any constant in $[1, \infty)$. That is, the Lyapunov function V as defined in (5.18) penalizes the on-ramp queues and mainline traffic densities equally. Consequently, the Lyapunov function depends only the 1-norms of q and n (i.e. total number of vehicles in the system) but not the spatial distribution of traffic over various cells/buffers. This construction is suitable for a highway with a single or a dominating bottleneck (see Example 5.2). An alternative construction is

$$a_{k,h} = \gamma k h \quad k = 1, \dots, K, \; h = 1, \dots, K, \tag{5.30}$$

where γ is any constant in $[1, \infty)$. This construction accounts for the spatial distribution of traffic: the Lyapunov function will strictly decrease as traffic is discharged downstream. Such a matrix A is suitable for a highway with multiple bottlenecks (see Example 5.3). Next, we illustrate the stability conditions resulting from both variable and constructed A matrices.

Example 5.2 (stability, stationary). *Consider the two-cell highway with a stationary hotspot, as described in Example 5.1. For ease of presentation, we only consider on-ramp priority, i.e.* $w = [1\ 1]^T$*, and only vary* $v = [v_1\ v_2]^T$ *over the set* $[0, \infty)^2$ *(assuming infinite demands). To implement Theorem 5.1 for this highway, we need to verify* (5.17) *over the set* $\mathcal{V}_1$*. That is, if there exist* A*,* $b^{(0)}$*, and* $b^{(1)}$ *satisfying these inequalities, then the on-ramp queues are stable.*

We consider three candidate A *matrices, viz. a matrix of unknowns to be determined, and those given by* (5.29) *and* (5.30)*. For a given control input* $u = (v, w)$*, if* A *is also given, then we only need to solve* (5.17) *for* $b^{(0)}$ *and* $b^{(1)}$*. Otherwise, we can obtain* A*,* $b^{(0)}$*, and* $b^{(1)}$ *by solving the linear inequalities* (5.17) *(note that* u *is fixed for now). In addition, we can obtain a set of destabilizing inputs by checking the necessary condition* (5.28a)–(5.28c)*: if a control input* u *does not satisfy these inequalities, then it is destabilizing.*

The results are illustrated in Fig. 5-7 and the nomenclature of various regions is in Tab. 5.4. Specifically, the union of regions 1-4 is the set of inflows $v = [v_1\ v_2]$ *that satisfy Theorem 5.1 (with a variable* A*) and are thus stabilizing. In addition, every* $[v_1\ v_2]^T$ *in region 6 violates the necessary condition* (5.28a)–(5.28c)*, and is thus destabilizing. Finally, there is a gap (region 5) between the sufficient condition and the necessary condition; for control inputs in that region, our stability conditions do not give a conclusive answer.*

Furthermore, every $[v_1\ v_2]^T$ *in region 3 verifies the stability criterion* (5.17) *together with a matrix* A *such that* $a_{k,h} = \gamma kh$*, but does not verify* (5.17) *if* $a_{k,h} = \gamma$*. In addition, in region 4, only a variable* A *matrix is able to verify* (5.17)*. Therefore, compared with a variable* A*, a matrix* A *such that* $a_{k,h} = \gamma$ *only increase the unknown region by a small size (union of regions 3 and 4). Hence,* $a_{k,h} = \gamma$ *is an adequate approximation to a variable* A *for this example.*

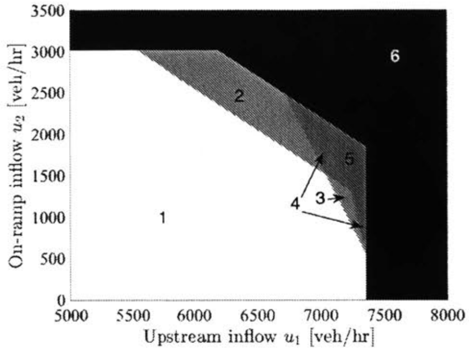


Figure 5-7: Stability of the SS-CTM with a stationary hotspot and with various control inputs $[v_1\ v_2]^T$. The white and light gray regions are the stabilizing inputs, while the black region is the destabilizing inputs. The dark gray region represents the gap between the necessary condition and the sufficient condition.

Table 5.2: Stability of various regions in Fig. 5-7.

Region	variable $a_{k,h}$	$a_{k,h} = \gamma$	$a_{k,h} = \gamma kh$
1	Stable	Stable	Stable
2	Stable	Stable	Unknown
3	Stable	Unknown	Stable
4	Stable	Unknown	Unknown
5	Unknown	Unknown	Unknown
6	Unstable	Unstable	Unstable

Example 5.3 (stability, moving). *Consider again the two-cell highway as described in Example 5.1. Now, we consider moving bottlenecks instead of a stationary hotspot. That is, the two-cell model has three modes $\mathcal{I} = \{0, 1, 2\}$, and, according to Tab. 5.1, the mode-specific cell capacities are*

$$F(0) = [7360\ 9850]^T, \quad F(1) = [5520\ 9850]^T, \quad F(2) = [7360\ 7880]^T.$$

Furthermore, we assume the mode transition rates to be

$$\lambda = 12[hr^{-1}], \quad \mu = 12[hr^{-1}],$$

which means that on average moving bottlenecks arrive at the highway every five minutes, and the average time that a moving bottleneck spends in a cell is five minutes;

see Fig. 5-2(b). Again, for various control inputs $[v_1\ v_2]^T$ *we apply Theorem 5.1 and* (5.28a)–(5.28c) *to check stability.*

The results are illustrated in Fig. 5-8 and the nomenclature of various regions is in Tab. 5.5. Specifically, the union of regions 1–3 consist of the set of inflows $v = [v_1\ v_2]$ *that satisfy Theorem 5.1 (with a variable A) and are thus stabilizing. In addition, every* $[v_1\ v_2]^T$ *in region 5 violates the necessary condition* (5.28a)–(5.28c)*, and is thus destabilizing. Finally, there is a gap (region 4) between the sufficient condition and the necessary condition.*

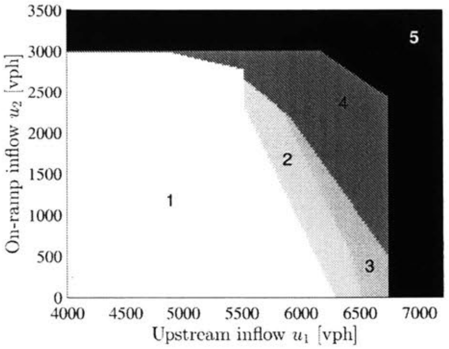


Figure 5-8: Stability of the SS-CTM with moving bottlenecks and with various control inputs $[v_1\ v_2]^T$. The white and light gray regions are the stabilizing inputs, while the black region is the destabilizing inputs. The dark gray region represents the gap between the necessary condition and the sufficient condition.

Table 5.3: Stability of various regions in Fig. 5-8.

Region	variable $a_{k,h}$	$a_{k,h} = \gamma$	$a_{k,h} = \gamma kh$
1	Stable	Stable	Stable
2	Stable	Stable	Unknown
3	Stable	Unknown	Unknown
4	Unknown	Unknown	Unknown
5	Unstable	Unstable	Unstable

Furthermore, every $[v_1\ v_2]^T$ *in region 2 verifies the stability criterion* (5.17) *together with a matrix A such that* $a_{k,h} = \gamma kh$*, but does not verify* (5.17) *if* $a_{k,h} = \gamma$*. In addition, in region 3, only a variable A matrix is able to verify* (5.17)*. Therefore, compared with a variable A, a matrix A such that* $a_{k,h} = \gamma kh$ *only increase the* *unknown region by a fairly small size (region 3). Hence, $a_{k,h} = \gamma kh$ is an adequate approximation to a variable A for this example.*

5.3 Formulation and analysis of max-throughput problem

By Theorem 5.1, we can formulate the max-throughput problem as follows:

$$\begin{aligned}
\max \quad & J = \sum_{k=1}^{K} v_k && \text{(P1)}\\
s.t. \quad & A\Big(G(i,x,u) + DH(i,x,u)\Big) + \sum_{j\in\mathcal{I}} \nu_{ij}\left(b^{(j)} - b^{(i)}\right) \le -e \quad \forall (i,x) \in \mathcal{I} \times \mathcal{V}_1(u), && \text{(5.31a)}\\
& u \in [0,d] \times \{0,1\}^K. && \text{(5.31b)}
\end{aligned}$$

The decision variables of (P1) are $u = (v, w)$, A, and $b^{(i)}$. Since G and H as well as the critical states $\mathcal{V}(u)$ are non-linear in u, even with a given matrix A, (5.31a) is non-linear in the u. In this section, we address this challenge by reformulating (P1) such that it is

(i) a mixed integer bilinear program (MIBLP), with a linear objective function and bilinear constraints, if A is variable, and

(ii) a mixed integer linear program (MILP) if A is given.

In addition, we derive some insights about the structure of optimal solutions to the max-throughput problem.

5.3.1 Reformulation

The main techniques involved in the reformulation include (i) substitution of G and H with a new set of variables,

$$\left\{\tilde{f}^{(i)}_{k,y,z},\ \tilde{r}^{(i)}_{k,y,z};\ k=1,\dots,K,\ i\in\mathcal{I},\ y\in\{0,1\}^K\backslash\{0\}^K,\ z\in\{0,1\}^K\right\},$$

and (ii) eliminate the cross-product term $v_k w_k$ (as appears in (5.13e)) using the big-M method. One can interpret $\tilde{f}^{(i)}_{k,y,z}$ and $\tilde{r}^{(i)}_{k,y,z}$ as the cell/buffer flow evaluated at the critical states. One can approximately interpret y and z as indices for the critical states: $y_k=0$ (resp. $y_k=1$) corresponds to $q_k=0$ (resp. $q_k=\infty$ or $q>0$),[3] and $z_k=0$ (resp. $z_k=1$) corresponds to $n_k=\underline{n}_k(q)$ (resp. $n_k=\bar{n}_k$). We reformulate (P1) as follows:

Proposition 5.1. (P1) *can be reformulated as the following mixed-integer program:*

$$\max \quad J=\sum_{k=1}^{K} v_k \tag{P2}$$

$$\begin{aligned}
s.t.\quad & \forall i\in\mathcal{I},\ \forall y\in\{0,1\}^K\backslash\{0\}^K,\ \forall z\in\{0,1\}^K,\\
& a_{h,1}v_1+\sum_{k=2}^{K}a_{h,k-1}v_k-\sum_{k=1}^{K-1}(a_{h,k}-a_{h,k+1})\left(\rho_k\tilde{f}^{(i)}_{k,y,z}+\tilde{r}^{(i)}_{k+1,y,z}\right)-a_{h,K}\tilde{f}^{(i)}_{k,y,z}\\
& \quad +\sum_{j\in\mathcal{I}}\nu_{ij}(b_h^{(j)}-b_h^{(i)})\le -1+\sum_{k=1}^{K}M_1 y_k\xi_k \quad h=1,\dots,K, && (5.32a)\\
& \tilde{r}^{(i)}_{k,y,z}\le v_k \quad \textit{if } y_k=0,\ k=1,\dots,K, && (5.32b)\\
& \tilde{r}^{(i)}_{k,y,z}\le \mathsf{R}_k \quad k=1,\dots,K, && (5.32c)\\
& \tilde{f}^{(i)}_{k,y,z}\le \rho_h^k\left(\min_i F_h(i)\right)+\sum_{\ell=h+1}^{k-1}\rho_\ell^k v_\ell+\tilde{r}^{(i)}_{k,y,z}\\
& \quad \textit{if } z_k=0,\ h=1,\dots,k-1, k=1,\dots,K, && (5.32d)\\
& \tilde{f}^{(i)}_{k,y,z}\le \sum_{\ell=1}^{k-1}\rho_\ell^k v_\ell+\tilde{r}^{(i)}_{k,y,z} \quad \textit{if } z_k=0,\ k=1,\dots,K, && (5.32e)
\end{aligned}$$

[3] Note that $y_k=0$ (resp. $y_k=1$) does not necessarily correspond to $q=\underline{q}_k$ (resp. $q=\bar{q}_k$), since sometimes $\underline{q}_k=\infty$ and $\bar{q}_k=0$.

$$\tilde{f}^{(i)}_{k,y,z} \leq F_k(i) \quad k = 1, \ldots, K, \tag{5.32f}$$

$$\tilde{f}^{(i)}_{k,y,z} \leq \frac{\tilde{f}^{(i)}_{k+1,y,z} - \tilde{r}^{(i)}_{k+1,y,z}}{\rho_k} \quad \textit{if } z_{k+1} = 1,\ k = 1, \ldots, K-1, \tag{5.32g}$$

$$v_k \leq \tilde{f}^{(i)}_{k,y,z} - \rho_{k-1}\tilde{f}^{(i)}_{k-1,y,z} + M_2 w_k + M_2(1-\xi_k) \quad k = 2, \ldots, K, \tag{5.32h}$$

$$v_k \leq \tilde{f}^{(i)}_{k,y,z} + M_2(1-w_k) + M_2(1-\xi_k) \quad k = 2, \ldots, K, \tag{5.32i}$$

$$a_{k,h} \geq a_{k+1,h} \quad k = 1, \ldots, K-1, \quad a_{K,h} \geq 1 \quad h = 1, \ldots, K, \tag{5.32j}$$

$$v \in [0, d],\ w \in \{0,1\}^K, \xi \in \{0,1\}^K. \tag{5.32k}$$

In summary, the decision variables in (P2) are v_k, w_k, $\tilde{f}^{(i)}_{k,y,z}$, $\tilde{r}^{(i)}_{k,y,z}$, ξ_k, $a_{k,h}$, and $b^{(i)}_k$. Note that y_k, z_k are not decision variables; instead, they are only notations intended for a compact representation of multiple inequalities. If A is variable, then (P2) is an integer problem with a linear objective function and bilinear constraints. If A is given, then (P2) is a mixed-integer linear program (MILP).

Next, we interpret the constraints (5.32a)–(5.32k). Constraints (5.32a)–(5.32i) are imposed for each $i \in \mathcal{I}$, each $y \in \{0,1\}^K \backslash \{0\}$, and each $z \in \{0,1\}^K$. (5.32a) is a reformulation of (5.31a), where the vector fields G and H are replaced by critical-state flows $\tilde{f}^{(i)}_{k,y,z}$ and $\tilde{r}^{(i)}_{k,y,z}$, and the matrix product is expanded for every row of A. The right-hand side of (5.32a) includes a big-M term, which means that this constraint is active[4] if and only if $y_k w_k = 1$ for some k. The auxiliary binary variable ξ_k results from (5.13f), which we will elaborate on as we interpret (5.32h)–(5.32i).

(5.32b)–(5.32c) result from (5.6a), i.e. the definition of the buffer-discharged flow.

(5.32d)–(5.32g) results from (5.6c) and (5.13b)–(5.13e), i.e. the expression for the flows and for the boundaries of the invariant set $\mathcal{M}$. Specifically, (5.32d)–(5.32e) result from $\underline{n}_k$; recall that $z_k = 0$ corresponds to $n = \underline{n}_k$. (5.32f) is the capacity constraint. (5.32g) is associated with the receiving flow constraint, which is only active if $z_{k+1} = 1$, i.e. if $n_{k+1} = \bar{n}_{k+1}$.

(5.32h)–(5.32i) are associated with the expression (5.13f) for $\bar{q}_k$. The big-M terms associated with ξ_k replace the cross-product term $v_k w_k$ in (5.13e) (and carried over to

[4] By "active", we mean that the constraint is imposed (instead of that the constraint is binding); by "inactive", we mean that the constraint is essentially a dummy one that does not affect the optimal solution under any circumstances.

(5.13f)). The auxiliary variable ξ_k serves the following purpose. If both (5.32h) and (5.32i) hold with $\xi_k = 1$, one can obtain from (5.13f) that $\bar{q}_k = 0$, and thus $q_k = 0$ for every $(q, n) \in \mathcal{V}$; therefore, we do not need to verify the drift condition at those states where $q_k > 0$ (i.e. $y_k = 1$), and ξ_k will "inactivate" (5.32a) for those y such that $y_k = 1$.

(5.32j) results from (5.16); this constraint is not needed if A is given. Recall that $b^{(i)}$ do not need to be constrained (see Section 5.2).

(5.32k) indicates the set of admissible control inputs $u = (v, w)$ and the auxiliary decision variables ξ_k. The auxiliary decision variables $\tilde{f}^{(i)}_{k,y,z}$ and $\tilde{r}^{(i)}_{k,y,z}$ are naturally constrained by (5.32b)–(5.32i) and do not need explicit range constraints.

In addition, for the big-M constraints, if A is variable, then we can use

$$M_1 = 2,$$
$$M_2 = \max_k \{\mathsf{F}_k + \mathsf{R}_k\}.$$

If A is given, then we can use the following values:

$$M_1 = K a_{1,1} \max_k \mathsf{F}_k + 1,$$
$$M_2 = \max_k \{\mathsf{F}_k + \mathsf{R}_k\}.$$

5.3.2 Structure of optimal solution

With a fixed A matrix, we are able to characterize the optimal solutions (under particular assumptions) and thus derive useful insights for highway control. In this subsection, we study how the structure of optimal control inputs is jointly influenced by the demand and the capacity. Because of the coupling between these two factors, analysis of a general K-cell highway is neither tractable nor insightful. However, we are able to derive useful insights by studying a two-cell highway section (as in Fig. 5-9) with either a stationary hotspot or moving bottlenecks. Similar two-cell models are commonly considered for ramp metering design and throughput analysis in the literature [59, 79, 93]. For ease of presentation, we do not consider the impact of off-ramps and assume $\rho_1 = 1$. Furthermore, we assume that $\mathsf{R}_1 = \mathsf{F}_1$.

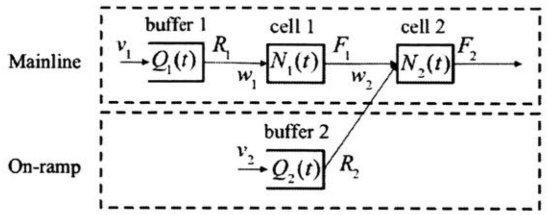


Figure 5-9: Two-cell highway with a single on-ramp.

Specifically, we show that the structure of the optimal control strongly depends on whether the given demand is feasible or infeasible. First, for a feasible demand, the optimal control follows the *margin criterion*: an on-ramp should be prioritized over the mainline (i.e. not metered) if the on-ramp has a smaller capacity-to-demand margin than the mainline. Second, for an infeasible demand, the optimal control input prioritizes the mainline (resp. the on-ramp) if the on-ramp (resp. the mainline) has a sufficiently large capacity-to-inflow margin. The optimal control inputs in the above two scenarios are consistent, in that it is always optimal to place the queue at the location where it will be discharged fast. For ease of presentation, our subsequent theoretical analysis only considers either feasible or infinite demand, and excludes the case of finite but infeasible demand. However, we will consider finite but infeasible demand in the numerical examples.

Stationary hotspot

Consider the two-cell highway with a stationary hotspot. Let $a_{k,h} = \gamma$; recall from Example 5.2 that this construction of A is suitable for this setting. Then, we can set $b_1^{(i)} = b_2^{(i)}$ without loss of generality, and formulate the max-throughput problem as follows:

$$\max \quad J = v_1 + v_2 \tag{P2.1}$$

$$s.t. \quad \forall i \in \{0,1\},\ \forall y \in \{0,1\}^2 \backslash \{0\}^2,\ \forall z \in \{0,1\}^2,$$

$$v_1 + v_2 - \tilde{f}_{2,y,z}^{(i)} + \sum_{j \in \mathcal{I}} \nu_{ij}(b_1^{(j)} - b_1^{(i)}) \le -\delta + \sum_{k=1}^{2} M_1 y_k \xi_k, \tag{5.33a}$$

$$\tilde{f}^{(i)}_{2,y,z} \le v_1 + \mathsf{R}_2 \quad \text{if } y_1 = 0, y_2 = 1, z_2 = 0, \tag{5.33b}$$

$$\tilde{f}^{(i)}_{2,y,z} \le \mathsf{F}_1 + v_2 \quad \text{if } y_1 = 1, y_2 = 0, z_2 = 0, \tag{5.33c}$$

$$\tilde{f}^{(i)}_{2,y,z} \le \mathsf{F}_2 - \Delta_2 \mathbb{1}_{\{i=1\}}, \tag{5.33d}$$

$$v_1 \le \mathsf{F}_2 - \Delta_2 - v_2 + M_2(1 - w_2) + M_2(1 - \xi_1), \tag{5.33e}$$

$$v_1 \le \mathsf{F}_2 - \Delta_2 + M_2 w_2 + M_2(1 - \xi_1), \tag{5.33f}$$

$$v_2 \le \mathsf{F}_2 - \Delta_2 + M_2(1 - w_2) + M_2(1 - \xi_2), \tag{5.33g}$$

$$v_2 \le \mathsf{F}_2 - \Delta_2 - v_1 + M_2 w_2 + M_2(1 - \xi_2), \tag{5.33h}$$

$$v \in [0, d],\ w \in \{0,1\}^2, \xi \in \{0,1\}^2, \tag{5.33i}$$

where $\delta = 1/\gamma$; the selection of δ or γ depends on the required numerical precision.

If the demand d is feasible in the formulation (P2.1), i.e. if there exist $v = d$ and $w \in \{0,1\}^2$ satisfying (5.33a)–(5.33i), we have the following result:

Proposition 5.2 (Feasible demand, stationary). *Consider a two-cell highway with a stationary hotspot. Suppose that $\rho_1 = 1$ and that the demand vector $d = [d_1\ d_2]^T$ is feasible under the formulation* (P2.1). *Then,*

1. $([d_1\ d_2]^T, [1\ 1]^T)$ *is an optimal solution to* (P2.1) *if*

$$\mathsf{F}_1 - d_1 \ge \mathsf{R}_2 - d_2. \tag{5.34}$$

2. $([d_1\ d_2]^T, [1\ 0]^T)$ *is an optimal solution to* (P2.1) *if*

$$\mathsf{F}_1 - d_1 \le \mathsf{R}_2 - d_2. \tag{5.35}$$

Proposition 5.2 gives a criterion for whether to meter an on-ramp or not. Specifically, if the on-ramp has a capacity-to-demand margin $(\mathsf{R}_2 - d_2)$ that is larger than the capacity-to-demand margin $(\mathsf{F}_1 - d_1)$ of the mainline, then the on-ramp is supposed to be metered. As an intuitive interpretation, if buffer 1 has a larger margin than cell 1, then cell 1 is the bottleneck that restricts the discharge of traffic congestion; in other words, the on-ramp can discharge traffic queue faster than the mainline does. Consequently, prioritizing the traffic from the bottleneck improves throughput.

Proof of Proposition 5.2. Consider a given $u = (v, w) \in [0, d] \times \{0, 1\}^2$. To obtain the conclusion, we only need to show the following:

(i) if $(v, w) = ([d_1\ d_2]^T, [1\ 0]^T)$ satisfies (5.33a)–(5.33i) and if (5.34) holds, then $(v, w') = ([d_1\ d_2]^T, [1\ 1]^T)$ also satisfies (5.33a)–(5.33i);

(ii) if $(v, w') = ([d_1\ d_2]^T, [1\ 1]^T)$ satisfies (5.33a)–(5.33i) and if (5.35) holds, then $(v, w) = ([d_1\ d_2]^T, [1\ 0]^T)$ also satisfies (5.33a)–(5.33i).

Here we only present the proof of part (i); part (ii) can be analogously proved. Suppose that $(v, w) = ([d_1\ d_2]^T, [1\ 0]^T)$ satisfies (5.33a)–(5.33i) and that (5.34) holds. We now show that this particular solution also satisfies (5.33a)–(5.33i).

If $v_1 + v_2 \le \mathsf{F}_2 - \Delta_2$, then $\bar{q}_1 = \bar{q}_2 = 0$ and $\mathcal{M}$ is bounded regardless of w_2, and the proof is straightforward.

If $v_1 + v_2 > \mathsf{F}_2 - \Delta_2$ and if $v_1 \le \mathsf{F}_2 - \Delta_2$, then, with $w_2 = 0$, (5.33a)–(5.33i) are equivalent to the following:

$$d_1 + d_2 - \mathsf{p}_0(d_1 + \mathsf{R}_2) - \mathsf{p}_1 \min\{d_1 + \mathsf{R}_2, \mathsf{F}_2 - \Delta_2) \le -\delta. \tag{5.36}$$

With $w_2 = 1$, (5.33a)–(5.33i) are implied by the following:

$$d_1 + d_2 - \mathsf{p}_0 \min\{\mathsf{F}_1 + d_2, d_1 + \mathsf{R}_2\} - \mathsf{p}_1 \min\{\mathsf{F}_1 + d_2, d_1 + \mathsf{R}_2, \mathsf{F}_2 - \Delta_2) \le -\delta. \tag{5.37}$$

If (5.34) holds, then (5.36) implies (5.37). Hence, $(v, w') = ([d_1\ d_2]^T, [1\ 1]^T)$ also satisfies (5.33a)–(5.33i).

If $v_1 + v_2 > \mathsf{F}_2 - \Delta_2$ and if $v_1 > \mathsf{F}_2 - \Delta_2$, then $\bar{q}_1 = \bar{q}_2 = \infty$ with either $w_2 = 0$ or $w_2' = 1$; thus, (5.33a)–(5.33i) are again equivalent to (5.37). Hence, $(v, w') = ([d_1\ d_2]^T, [1\ 1]^T)$ also satisfies (5.33a)–(5.33i). ■

If the demand is infinite, we have the following result:

Proposition 5.3 (Infinite demand, stationary). *Consider a two-cell highway with a stationary hotspot. Suppose that the demands are infinite, i.e.* $d_1 = d_2 = \infty$. *Furthermore, assume that* $\rho_1 = 1$ *and* $\mathsf{F}_1 = \mathsf{R}_1$. *Let* $\bar{F}_2 = \mathsf{p}_0\mathsf{F}_2 + \mathsf{p}_1(\mathsf{F}_2 - \Delta_2)$. *Then,*

1. *Every* (v, w) *in the set* $\mathscr{V}_1^* \times \{[1\ 0]^T\}$, *where*

$$\mathscr{V}_1^* = \Big\{v \in [0, d] : v_1 + v_2 = \bar{F}_2,\ \mathsf{R}_2 - v_2 \geq \mathsf{F}_2 - \bar{F}_2,\ v_1 < \min\{\mathsf{F}_1, \mathsf{F}_2 - \Delta_2\}\Big\} \tag{5.38}$$

is an optimal solution to (P2.1).

2. *Every* (v, w) *in the set* $\mathscr{V}_2^* \times \{[1\ 1]^T\}$, *where*

$$\mathscr{V}_2^* = \Big\{v \in [0, d] : v_1 + v_2 = \bar{F}_2,\ \mathsf{F}_1 - v_1 \geq \mathsf{F}_2 - \bar{F}_2,\ v_2 < \min\{\mathsf{R}_2, \mathsf{F}_2 - \Delta_2\}\Big\} \tag{5.39}$$

is an optimal solution to (P2.1).

In summary, for infinite demand, if the on-ramp has a sufficiently large capacity-to-inflow margin (in the sense that $\mathsf{R}_2 - v_2 \geq \mathsf{F}_2 - \bar{F}_2$), then it is optimal to prioritize the mainline. The intuition is that in such a case, even if mainline priority leads to queues at the on-ramp, the queues can be discharged quickly. Similarly, if the mainline has a sufficiently large capacity-to-inflow margin (in the sense that $\mathsf{F}_1 - v_1 \geq \mathsf{F}_2 - \bar{F}_2$), then it is optimal to prioritize the on-ramp. This is consistent with the logic of the margin criterion characterized by Proposition 5.2. Note that this logic can be extended to the case where the demand is finite but infeasible with straightforward modification.

Proof of Proposition 5.3. We only present the proof of part (i); part (ii) can be analogously proved. On the one hand, consider a $v^* \in \mathscr{V}_1^*$ and $w^* = [1\ 0]^T$. Then, we have

$$\begin{aligned} & v_1^* + v_2^* - \mathsf{p}_0 \min\{v_1^* + \mathsf{R}_2, \mathsf{F}_2\} - \mathsf{p}_1\{v_1^* + \mathsf{R}_2, \mathsf{F}_2 - \Delta_2\} \\ & \overset{(5.38)}{\leq} \bar{F}_2 - \mathsf{p}_0\mathsf{F}_2 - \mathsf{p}_1(\mathsf{F}_2 - \Delta_2) = 0. \end{aligned} \tag{5.40}$$

Since $v_1 < \mathsf{F}_2 - \Delta_2$, we obtain from (5.13f) that $\bar{q}_1 = 0$. Thus, (5.40) ensures (5.33a)–(5.33i) (see the proof of Proposition 5.2). Hence, (v, w) is a feasible solution to (P2.1).

On the other hand, $v_1^* + v_2^* \leq \bar{F}_2$ is a necessary condition for stability, and $\bar{F}_2$ is the maximum throughput that can be achieved (regardless of the formulation).

In conclusion, (v^*, w^*) is an optimal solution to the (P2.1). ■

Remark 5.3. *Proposition 5.3 gives at least one optimal solution, i.e.* $\mathscr{V}_1^* \cup \mathscr{V}_2^* \neq \emptyset$, *if and only if (i)* $\mathsf{F}_1 + \mathsf{R}_2 \geq \mathsf{F}_2$ *and (ii)* $\Delta_2 \leq \min\{\mathsf{F}_1, \mathsf{R}_2\}$. *Note that (i) and (ii) typically hold for highway on-ramps [42].*

Propositions 5.2 and 5.3 can be mechanically extended to a general K-cell highway. However, the extension is notationally heavy and less insightful.

Example 5.4 (I210-EB, stationary). *Recall the two-cell highway with a stationary hotspot as described in Example 5.2. This example illustrates some insights from Propositions 5.2 and 5.3, and visualizes the structure of the max-throughput problem* (P2.1). *Figure 5-10 shows the feasible set of* (P2.1)*; the meaning of each region is listed in Tab. 5.4. Note that the feasible set is almost an exact one, since the gap (region 4) between the stable regions and the unstable region is very small.*

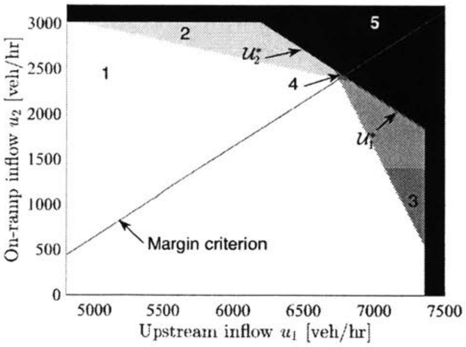


Figure 5-10: Stability of various control schemes (v, w). The margin criterion and a set of optimal solutions (green/red line segments) are also indicated.

To obtain optimal solutions to (P2.1)*, we first need to determine whether the given demand vector* $d = [d_1\ d_2]^T$ *is feasible or not. If the demand vector d falls in the union of regions 1–3, we know that it is feasible, and the corresponding control input can be*

Table 5.4: Stability of various regions in Figure 5-10.

Regime	$w_2 = 0$	$w_2 = 1$	$w_2 = v_2^*$
1	Stable	Stable	Stable
2	Unknown	Stable	Stable
3	Stable	Unknown	Stable
4	Unknown	Unknown	Unknown
5	Unstable	Unstable	Unstable

determined using the margin criterion given by Proposition 5.2: if $v^* = d = [d_1\ d_2]^T$ *is below (resp. above) the line of margin criterion, then* $w^* = [1\ 0]^T$ *(resp.* $w^* = [1\ 1]^T$*). If the demand d is infinite, then a (sub)set of optimal solutions is given by*

$$\left(\mathscr{V}_1^* \times \{[1\ 0]^T\}\right) \cup \left(\mathscr{V}_2^* \times \{[1\ 1]^T\}\right)$$

where $\mathscr{V}_1^*$ *and* $\mathscr{V}_2^*$ *are given by Proposition 5.3 and illustrated in Fig. 5-10.*

Moving bottlenecks

Consider the two-cell highway section as shown in Fig. 5-9. Suppose that moving bottlenecks randomly arrives at and moves through the highway section. That is, the highway has three modes $\{0, 1, 2\}$, and the mode-specific capacities are

$$F(0) = [\mathsf{F}_1\ \mathsf{F}_2]^T,\ F(1) = [\mathsf{F}_1 - \Delta_1\ \mathsf{F}_2]^T,\ F(2) = [\mathsf{F}_1\ \mathsf{F}_2 - \Delta_2]^T,$$

and the inter-mode transition rates are

$$\nu_{01} = \lambda,\ \nu_{12} = \mu,\ \nu_{20} = \mu.$$

Let $a_{k,h} = \gamma k h$ and $b_1^{(i)} = 2b_2^{(i)}$; recall from Section 5.3 that this construction of A is suitable for moving bottlenecks. Then formulation for the max-throughput problem is as follows:

$$\max \quad J = v_1 + v_2 \tag{P2.2}$$

s.t. $\forall i \in \{0,1,2\},\ \forall y \in \{0,1\}^2 \backslash \{0\}^2,\ \forall z \in \{0,1\}^2,$

$$2v_1 + 2v_2 - \tilde{f}^{(i)}_{1,y,z} - \tilde{r}^{(i)}_{2,y,z} - \tilde{f}^{(i)}_{2,y,z} + \sum_{j \in \mathcal{I}} \nu_{ij}(b^{(j)}_1 - b^{(i)}_1) \le -\delta + \sum_{k=1}^{2} M_1 y_k w_k, \tag{5.41a}$$

$$\tilde{r}^{(i)}_{k,y,z} \le v_k \quad \text{if } y_k = 0,\ k = 1,2, \tag{5.41b}$$

$$\tilde{r}^{(i)}_{1,y,z} \le \mathsf{R}_1, \quad \tilde{r}^{(i)}_{2,y,z} \le \mathsf{R}_2, \tag{5.41c}$$

$$\tilde{f}^{(i)}_{1,y,z} \le \tilde{r}^{(i)}_{1,y,z} \quad \text{if } y_1 = 0, \tag{5.41d}$$

$$\tilde{f}^{(i)}_{1,y,z} + \tilde{r}^{(i)}_{2,y,z} \le \tilde{f}^{(i)}_{2,y,z} \quad \text{if } z_2 = 1,$$

$$\tilde{f}^{(i)}_{2,y,z} \le \rho_1(\mathsf{F}_1 - \Delta_1) + \tilde{r}^{(i)}_{2,y,z} \quad \text{if } z_2 = 0, \tag{5.41e}$$

$$\tilde{f}^{(i)}_{2,y,z} \le \rho_1 v_1 + \tilde{r}^{(i)}_{2,y,z} \quad \text{if } z_2 = 0, \tag{5.41f}$$

$$\tilde{f}^{(i)}_{1,y,z} \le \mathsf{F}_1 - \Delta_1 \mathbf{1}_{\{i=1\}},\ \tilde{f}^{(i)}_{2,y,z} \le \mathsf{F}_2 - \Delta_2 \mathbf{1}_{\{i=2\}}, \tag{5.41g}$$

$$\tilde{f}^{(i)}_{1,y,z} \le \frac{\tilde{f}^{(i)}_{2,y,z} - \tilde{r}^{(i)}_{2,y,z}}{\rho_k} \quad \text{if } y_2 = 1, \tag{5.41h}$$

$$v_2 \le \tilde{f}^{(i)}_{2,y,z} - \rho_1 \tilde{f}^{(i)}_{1,y,z} + M_2(1 - w_2) + M_2(1 - \xi_2), \tag{5.41i}$$

$$v_2 \le \tilde{r}^{(i)}_{2,y,z} + M_2 w_2 + M_2(1 - \xi_2), \tag{5.41j}$$

$$v \in [0,d],\ w \in \{0,1\}^2,\ \xi \in \{0,1\}^2. \tag{5.41k}$$

For a feasible demand, we again have the “margin criterion” for traffic control under moving bottlenecks:

Proposition 5.4 (Feasible demand, moving). *Consider a highway of two-cells with moving bottlenecks. Suppose that (i) $\rho_1 = 1$ and (ii) the demand vector $d = [d_1\ d_2]^T$ is feasible in the formulation* (P2.2).

1. *An optimal solution is $v^* = [d_1\ d_2]^T$, $w^* = [1\ 1]^T$ if*

$$\mathsf{R}_2 - d_2 \le \mathsf{F}_1 - \Delta_1 - d_1. \tag{5.42}$$

2. *An optimal solution is $v^* = [d_1\ d_2]^T$, $w^* = [1\ 0]^T$ if*

$$\mathsf{R}_2 - d_2 \ge \mathsf{F}_1 - d_1. \tag{5.43}$$

The above result essentially states that, if the kth on-ramp has a capacity-to-demand margin $(\mathsf{R}_2 - d_2)$ smaller than that of the mainline $(\mathsf{F}_1 - \Delta_1 - d_1)$ under the influence of the moving bottleneck, then it should be metered; if the on-ramp has a capacity-to-demand margin $(\mathsf{R}_2 - d_2)$ larger than that of the mainline $(\mathsf{F}_1 - d_1)$ even without the influence of the moving bottleneck, then it should not be metered. If the margins do not fall in the above regions, this result does not provide conclusive characterization of the structure of the optimal ramp metering plan. However, we can still obtain the optimal solution by solving the MILP (P2).

The proof of Proposition 5.4 is similar to that of Proposition 5.2.

For infinite demand, we again analytically compute an optimal solution. Due to the complexity of the SS-CTM dynamics under moving bottlenecks, a complete characterization of the optimal solution involves too many cases and is thus tedious. For ease of presentation, we only consider a practically relevant case, where $\mathsf{F}_1 + \mathsf{R}_2 \geq \mathsf{F}_2$, and $\mathsf{F}_1 - \Delta_1 + \mathsf{R}_2 \leq \mathsf{F}_2 - \Delta_2/2$; both inequalities hold if $\mathsf{F}_1 + \mathsf{R}_2$ is slightly greater than F_2, which is typically true for a highway merge. (Note that the subsequent result can be easily extended to the other cases.) Then, we have the following result:

Proposition 5.5 (Infinite demand, moving). *Consider a two-cell highway with moving bottlenecks. Suppose that (i)* $d_1 = d_2 = \infty$, *(ii)* $\rho_1 = 1$, *(iii)* $\mathsf{R}_1 = \mathsf{F}_1$, *(iv)* $\mathsf{F}_1 + \mathsf{R}_2 \geq \mathsf{F}_2$, *and (v)* $\mathsf{F}_1 - \Delta_1 + \mathsf{R}_2 \leq \mathsf{F}_2 - \Delta_2/2$. *Then, an optimal solution to* (P2.2) *is*

$$v^* = \begin{bmatrix} \mathsf{p}_0(\mathsf{F}_2 - \Delta_2/2) + \mathsf{p}_1(\mathsf{F}_1 - \Delta_1 + \mathsf{R}_2) + \mathsf{p}_2(\mathsf{F}_2 - \Delta_2) - \mathsf{R}_2 \\ \mathsf{R}_2 \end{bmatrix}, \quad w^* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \tag{5.44}$$

where $\mathsf{p}_0, \mathsf{p}_1, \mathsf{p}_2$ *are the steady-state probabilities given by* (5.1).

The above result essentially implies that on-ramp priority is more efficient in this particular setting. In addition, the on-ramp inflow is maximized ($v_2 = \mathsf{R}_2$). The reason is that the capacity of buffer 2 is more reliable than that of cell 1, which is subject to perturbations. The proof of Proposition 5.5 is similar to that of Proposition 5.3.

Example 5.5 (Moving bottlenecks on I210-EB). *This example provides some insights about the max-throughput problem* (P2.2). *Consider again the two-cell example presented in Section 5.3. For each $u = (v, w)$, we can verify the stability by applying the sufficient condition, Theorem 5.1 and the necessary condition* (5.28a)–(5.28c). *Figure 5-11 shows the results; the meaning of each region is listed in Tab. 5.5.*

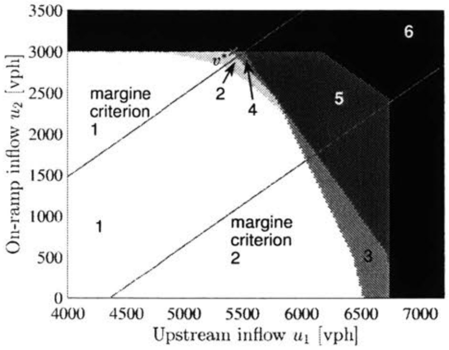


Figure 5-11: Stability of various control inputs. Margin criterion 1 (resp. 2) is specified by (5.42) (resp. (5.43)). v^* is the optimal solution in the case of infinite demand given by Proposition 5.5.

Table 5.5: Stability of various regions in Figure 5-11.

Region	Ramp metered $w_2 = 0$	Ramp not metered $w_2 = 0$	Margin criterion w_2^*
1	Stable	Stable	Stable
2	Unknown	Stable	Stable/unknown
3	Stable	Unknwon	Stable
4	Unknown	Unknown	Unknown
5	Unstable	Unstable	Unstable

To obtain the optimal solution to (P2.2), *we first need to determine whether the given demand vector $d = [d_1\ d_2]^T$ is feasible or not. If the d falls in the union of regions 1–3, we know that it is feasible, and the corresponding control input can be determined using the margin criteria given by Proposition 5.4: if $d = [d_1\ d_2]^T$ is below (resp. above) the line of margin criterion 1 (resp. 2), then $w^* = [1\ 0]^T$ (resp. $w^* = [1\ 1]^T$). If the demand vector falls between the two criteria, our results do not give an analytical characterization; however, one can obtain optimal solutions by solving the MILP* (P2.2). *If the demand d is infinite, then a particular optimal* *solution is given by Proposition 5.5 and illustrated in Fig. 5-11.*

5.4 Case study: a full-day simulation of SR-134 East/ I-210 East

In this section, we consider a 33.2-km stretch of SR-134 East/ I-210 East in Los Angeles County shown in Figure 5-12, as a test case for the margin criterion for ramp control. This freeway stretch consists of 6.3 km of SR-134 East from postmile 9.46 to postmile 13.36 and 26.9 km of I-210 East from postmile 25 to postmile 41.7. There are 28 on-ramps and 25 off-ramps. We consider the on-ramp flows measured on a day. Particularly, we focus on two scenarios of capacity perturbations, viz. a stationary perturbation hotspot and moving bottlenecks.

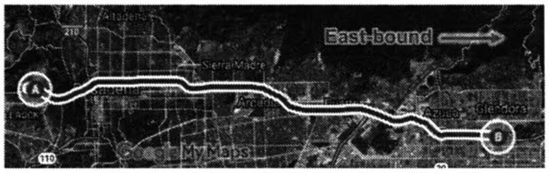
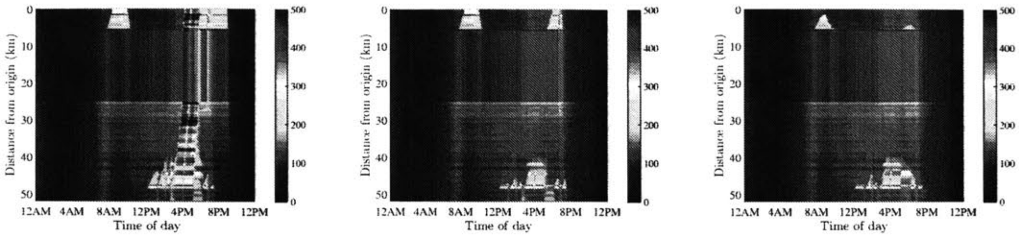


Figure 5-12: The segment of SR-134 East/ I-210 East studied in this section.

The model was built using PeMS data [101] for the corresponding segments of the SR-134 East and I-210 East for Monday, October 13, 2014. This was one of the days when most vehicle detectors on mainline, on-ramps, and off-ramps of SR-134 East and I-210 East were intact, and hence the PeMS data are reliable. Traffic flow parameters were calibrated using PeMS data following the methodology [31]. The simulations were conducted by Dr. Alexander A. Kurzhanskiy from UC Berkeley. More information about the simulation tool is available in [42].

5.4.1 Stationary perturbation hotspot

We consider a stationary perturbation hotspot near North Azusa Avenue. The capacity of the hotspot switches between 100% (mode 0) and 75% (mode 1) of its nominal



(a) On-ramp priority. (b) Mainline priority. (c) Margin criterion.

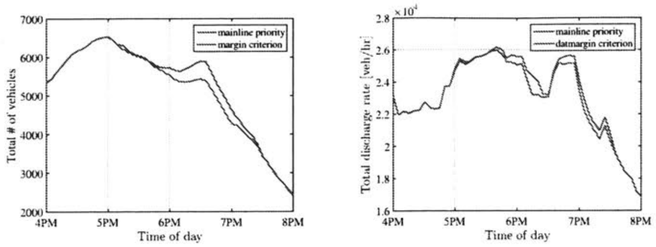
Figure 5-13: Traffic density contour for the highway with a stationary hotspot.

capacity, and spends equal time in both modes. That is, the time average capacity is 87.5% of the nominal capacity.

We run three simulations in this setting. First, every on-ramp is prioritized over the mainline, i.e. $y_k = 1$ for every on-ramp k. This is considered as the baseline, where no ramp is metered. Second, the mainline is prioritized over every on-ramp, i.e. $y_k = 0$ for every k. This can be viewed as an "aggressive" ramp metering strategy. Third, the on-ramp priorities are determined according to the margin criterion given by Proposition 5.2. For the three simulations, we track the traffic evolution over time. The resulting traffic density contour plot is show in Figures 5-13(a)–5-13(c).

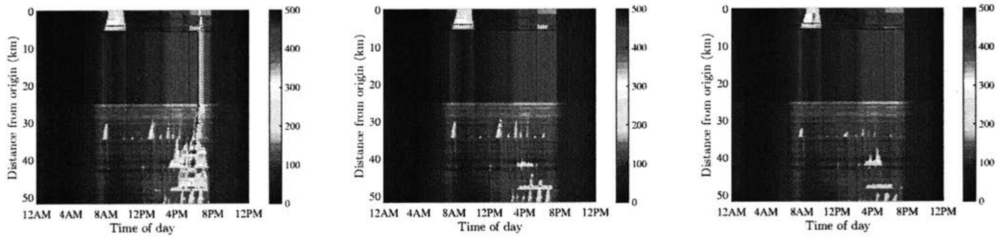
For all three control configurations, we compute the traffic delay with respect to free-flow travel time. Compared to the baseline, mainline priority reduces delay by 62%. Compared to mainline priority, the margin criterion further reduces delay by 3%. More importantly, the margin criterion eliminates several very long on-ramp queues that mainline priority induces by prioritizing particular on-ramps at those on-ramps during certain hours: the longest on-ramp queue upstream from the stationary hotspot is reduced by 39%.

Surprisingly, margin criterion leads to a significantly smaller upstream queue at the origin. The reason is that mainline priority under-utilizes highway capacity and thus discharges congestion more slowly. Figures 5-14(a) and 5-14(b) clearly illustrate how the margin criterion improves discharge rate during the evening peak hour.



(a) Total number of vehicles. (b) Total discharge rate.

Figure 5-14: Margin criterion accelerates discharge of traffic during peak hours.



(a) On-ramp priority. (b) Mainline priority. (c) Margin criterion.

Figure 5-15: Traffic density contour for the highway with moving bottlenecks.

5.4.2 Moving bottlenecks

We now consider moving bottlenecks randomly arriving at the highway. We consider an arrival rate of moving bottlenecks of 12 per hour. When the moving bottleneck is in a cell, then the cell's capacity is reduced by the capacity of one lane. The expected time μ_k that a moving bottleneck spends in a cell is given by

$$\mu_k = \frac{l_k}{v_k},$$

where l_k is the length of the cell and v_k is the free-flow speed.

Once again, we consider the control configurations in the previous subsection, viz. on-ramp priority, mainline priority, and margin criterion. The resulting traffic density contour plots are given by Figures 5-15(a)–5-15(c). Compared to the baseline, mainline priority reduces delay by 16%. Compared to mainline priority, the margin criterion further reduces delay by 25%.

5.5 Summary

In this chapter, we considered the maximization of the throughput of a perturbation-prone highway section, under the constraint that every on-ramp should remain bounded on average. We developed a sufficient condition (Theorem 5.1) for bounded on-ramp queues, which is based on the construction of a switched quadratic Lyapunov function and verification of the Foster-Lyapunov criterion. Furthermore, we formulate the max-throughput problem as either an MILP or an MIBLP, depending on whether the parameters of the Lyapunov function are given as constants or solved as unknowns. Under the MILP formulation, we characterized the structure of the optimal solutions to the max-throughput problem, which we summarized as the "margin criterion": traffic queue should be placed on a (mainline or on-ramp) link with a larger capacity-to-demand margin.

Chapter 6

Conclusions and Ongoing Work

6.1 Summary of this thesis

In this thesis, we have considered analysis and control of highway systems subject to capacity disruptions and heterogeneous demand, which are important concerns in many practical settings.

In Chapter 1, we posed the resiliency question for smart highway systems, i.e. efficiency in the nominal setting, robustness against random perturbations, and survivability under security failures. We have discussed about the first two aspects in thesis, and will mention our ongoing work in the third aspect in the next section. We argued that the main challenge is the lack of models for smart highway systems and for reliability/security failures.

In Chapter 2, we considered the routing problem over a network of parallel PDQ links. We particularly focused on the feedback-controlled stability of the system. We showed that a necessary condition for stability is that a lower bound on the time-average link inflows does not exceed the corresponding time-average saturation rate. In addition, we showed that a sufficient condition for stability is that (i) the nominal mode's saturation rate is high enough that all queues vanish in this mode (i) and a bilinear matrix inequality (BMI) involving an underestimate of the discharge rates of the PDQ in individual modes is feasible. Furthermore, under the sufficient condition, the state of the network converges to a unique invariant probability measure.

In Chapter 3, we developed a piecewise-deterministic queuing (PDQ) model for vehicle platoons. This model captures the interaction between platoons of CAVs and the background traffic in terms sharing the highway's capacity. We show that randomness in the arrival process of platoons can induce congestion on both platoons and the background traffic. Our PDQ model also allows analytical characterize the platooning-induced queue in terms of parameters of the highway and of the platooning operations. To further validate the model, one can either conduct micro-simulation or run field experiments to estimate how well the model captures the link between key platooning parameters and highway performance.

In Chapter 4, we developed the stochastic switching cell transmission model (SS-CTM) for highway sections subject to random perturbations. The main difference between the SS-CTM and the classical CTM is that cell capacities in the SS-CTM are stochastically varying according to a Markov chain. We develop a sufficient condition, in the form of a set of bilinear inequalities, for the boundedness of the upstream traffic queue. This sufficient condition is also established by constructing a Lyapunov function and applying the classical Foster-Lyapunov drift condition. The proof involves the construction of a globally attracting invariant set, and utilizes the properties of the traffic flow dynamics to show that, instead of verifying the drift condition everywhere over the continuous state space, it suffices to verify it over a finite set of states. We also use our results to analyze the impact of stochastic capacity fluctuation (frequency, intensity, and spatial correlation) on the throughput of a freeway segment.

In Chapter 5, we build on the SS-CTM and considered the control design problem. We studied the scenario where on-ramp demand can be managed and on-ramps can be either metered or not. We posed an optimization problem, where the objective is to maximize the throughput and the constraint is to ensure bounded on-ramp queues. The main result is (i) a sufficient condition for the controlled SS-CTM, and (ii) a MIBLP/MILP formulation of the max-throughput problem. Under particular assumptions, we also characterized the structure of the optimal control input. Particularly, we showed that if traffic queue is inevitable, it should be placed on a location (either on the mainline or on an on-ramp) with a larger capacity-to-demand margin, which we called the "margin criterion".

6.2 Ongoing work

The work presented in this thesis is being extended in several directions. First, the PDQ model for platooning shown in Chapter 3 is a natural basis for control design. Specifically, we are studying regulating the speed of vehicle platoons to alleviate the efficiency loss at highway bottlenecks. Second, we are synthesizing game theoretic models with traffic models to investigate the design of protection and inspection schemes for critical smart highway components. Third, we are modeling the impact (in terms of traffic delay and throughput loss) due to cyber-physical attacks on smart highway systems, and developing effective response strategies for timely recovery.

6.2.1 Speed control of vehicle platoons

In this thesis, we have considered the interaction between vehicle platoons and highway traffic from two perspectives (see Chapters 3 and 5). Following this work, we are considering the regulation of speed of vehicle platoons to alleviate the congestion and throughput loss due to platooning at highway bottlenecks.

We consider a single highway split as shown in Fig. 6-1. CAVs only travel on the mainline, while the background traffic have a fixed split between the mainline and the off-ramp. That is, there are three traffic classes:

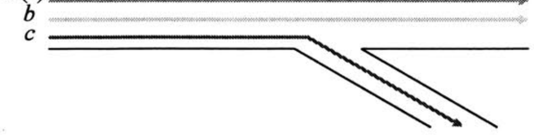
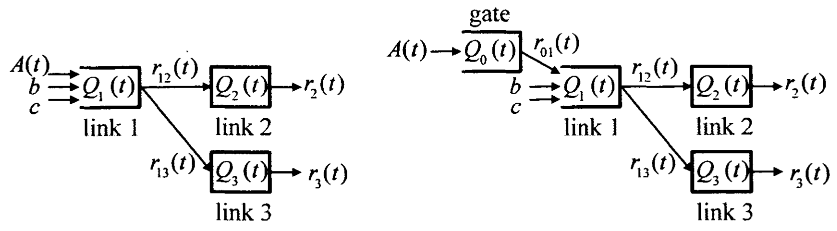


Figure 6-1: A highway section with an off-ramp.

1. Class a: CAVs, which travel in platoons on the mainline. The inflow rate of this traffic class is a Markovian process $A(t)$ (similar to that defined in Section 3.1.1).



(a) Tandem links without control. (b) Tandem links with control.

Figure 6-2: The multi-class, tandem PDQ models for the highway section in Fig. 6-1.

Specifically, $A(t)$ switches between two values a and 0 according to a Markov chain, with transition rates λ (0 to a) and μ (a to 0).

2. Class b: background traffic remaining on the mainline, with a constant inflow rate b.

3. Class c: background traffic leaving the highway via the off-ramp, with a constant inflow rate c.

We model the highway section in Fig. 6-1 as a tandem fluid queuing system, as shown in Fig. 6-2(a). links 1 and 2 are on the mainline of a highway, and link 3 is a downstream arterial road connected to the mainline via an on-ramp. The model that we consider is a stochastic hybrid system. The state of the model is (b, q^a, q^b, q^c), where b is the arrival rates of CAVs, $q^a = [q_1^a\ q_2^a]^T \in \mathbb{R}^2_{\geq 0}$ is the vectors of queue lengths of CAVs, $q^b = [q_1^b\ q_2^b]^T \in \mathbb{R}^2_{\geq 0}$ is the queues of mainline traffic, and $q^c = [q_1^c\ q_3^c]^T \in \mathbb{R}^2_{\geq 0}$ is the vector of off-ramp traffic. Furthermore, we define q_k to be the total queue length in link k, i.e.

$$
\begin{aligned}
q_1 &= q_1^a + q_1^b + q_1^c, \\
q_2 &= q_2^a + q_2^b, \\
q_3 &= q_3^c.
\end{aligned}
$$

We regulate the speed of each individual platoons so that their arrival times at the bottleneck (end of link 2) do not conflict. In terms of the fluid queuing model, this control can be considered as applying a “gate” that regulates the arrival process of vehicle platoons, as shown in Fig. 6-2(b). Practically, the decision variable of the control design problem is the time of arrival of each platoon at link 1. In the PDQ model, this is equivalent to controlling the discharge rate r_{01} from link 0 (the "gate") to link 1. That is, r_{01} is specified by a function (control law) $\phi : (a, q^a, q^b, q^c) \mapsto r_{01}$.

The controlled system is stable if there exists a finite constant C such that for each initial condition

$$\limsup_{t\to\infty} \frac{1}{t} \int_{s=0}^{t} \mathsf{E}\left[\sum_{k=0}^{3} Q_k(s)\right] ds \leq C.$$

Throughput is defined as

$$J = \lim_{t\to\infty} \frac{1}{t} \int_{s=0}^{t} (r_2(s) + r_3(s))ds$$

if this limit exists. For a stable system, we have

$$J = \left(\lim_{t\to\infty} \frac{1}{t} \int_{s=0}^{t} A(s)ds\right) + b + c$$

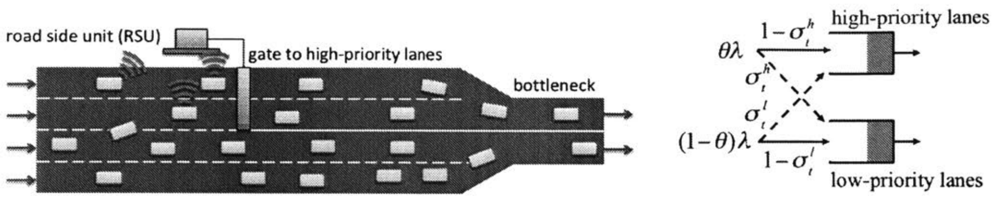
Hence, the control design problem can be formulated as an optimization problem:

$$\begin{aligned} \max_{\phi} \quad & J \\ s.t. \quad & q^a, q^b, q^c \text{ are bounded on average} \\ & \phi \text{ satisfies practical constraints} \end{aligned}$$

To convert the above formulation into a solvable optimization problem, we first need to develop a sufficient condition for the boundedness of the queues, and then identify the set of practically admissible control laws.

6.2.2 Inspection of misbehavior in V2I-based operations

Vehicle-to-Infrastructure (V2I) communications are increasingly supporting highway operations such as electronic toll collection, carpooling, and vehicle platooning. In this



(a) Highway configuration.

(b) Abstract queuing model.

Figure 6-3: A highway with V2I-based lane management operations.

work, we are studying the incentives of strategic misbehavior by individual vehicles who can exploit the security vulnerabilities in V2I communications and impact the highway operations. We consider a V2I-enabled highway segment facing two classes of vehicles (agent populations), each with authorized access to one server (subset of lanes).

Next, we demonstrate via a particular example in the context of smart highways how the models and approaches that we developed in the previous chapters can be integrated into security failure analysis.

We consider a simple model of lane management operations on a highway section equipped with vehicle-to-infrastructure (V2I) communications capability. Suppose that the highway system faces a fixed traffic demand λ comprised of two types of agent populations: a high priority type, denoted h, and a low priority type, denoted l; see Fig. 6-3(a). The fraction of type h agents is $\theta \in (0, 1)$, and the fraction of type l agents is $1 - \theta$. There are two sets of lanes on the highway, H and L, which we model as two parallel servers; see Fig. 6-3(b). In the absence of misbehavior, server H only serves type h agents, and server L only serves type l agents. The highway is equipped with a road-side unit (RSU), which collects messages from incoming agents to monitor the traffic and grants access to each incoming agent based on the received message (i.e., reported type).

Given any fraction of type h travelers, θ, the *travel cost* (or queueing delay) on the H (resp. L) server is denoted as c_{H}^{θ} (resp. c_{L}^{θ}). In general, c_{H}^{θ} (resp. c_{L}^{θ}) increases with the aggregate demand of agents using the server H (resp. L). In our setting, to reduce his/her travel cost, an agent may misreport its type to the RSU, which we consider as *misbehavior*; i.e. a misbehaving type l agent reports itself as having authorization for the sever H; similarly for a misbehaving type h agent. We use σ_l^t (resp. σ_h^t) to denote the fraction of l (resp. h) agents that misbehave. Consider a generic misbehavior strategy $\sigma^t = (\sigma_h^t, \sigma_l^t)$. Since the fraction of misbehaving agents impacts the demand of agents using each server, we can use the notations $c_{\mathrm{H}}^{\theta}(\sigma^t)$ (resp. $c_{\mathrm{L}}^{\theta}(\sigma^t)$) to denote the cost of server H (resp. L) under the strategy σ^t.

Characterization of the equilibria (in the sense of Perfect Bayesian Equilibria, or PBEs; see [32]) of this signaling game is available in [107]. Fig. 6-4 shows the regime plot of an example taken from [107]. In regime A, since the misbehavior cost is high, no traveler would have the incentive to misbehave. Consequently, the SO does not need to inspect. In regime B_1, travelers misbehave with a strictly positive probability; however, the SO chooses not to inspect due to high inspection cost. In regime B_2, travelers misbehave with a strictly positive probability, and the SO inspects any traveler with a strictly positive probability.

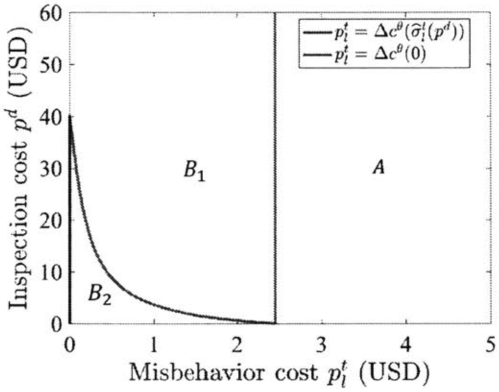


Figure 6-4: PBE regimes. Analytical expressions for the boundaries, i.e. the function Δc^{θ}, is provided in [107].

6.2.3 Response to adversarial cyber-physical attacks

We have considered random capacity perturbations in Chapters 2–5, which mainly account for "physical" capacity-reducing events. However, operations of smart highway systems rely on ubiquitous sensing and actuation capabilities, mobile and embedded computing with smartphones, and deep penetration of wireless communications networks. A significant drawback of this Information Technology (IT) modernization is lowered security of transportation systems, caused by the exposure to IT insecurities. That is, security failures of the cyber components may also lead to capacity perturbations.

Several real incidents have confirmed that transportation NCS [11, 78] and more generally, supervisory control and data acquisitions (SCADA) systems for other critical infrastructures [3, 17, 95] are subject to significant security risks. These can be broadly classfied as IT-based accidents, non-targeted attacks, and targeted attacks. IT-based accidents are caused due to unintended technology failures. The non-targeted attacks are similar to the incidents that any network-connected computer may suffer. For example, In 2003, the Sobig virus infected the CSX train control computer, shutting down the train signaling systems in the US East Coast for 4 to 6 hours. Targeted attacks could be the most damaging class of attacks because they are tailored specifically to inflict maximum damage to NCS. For example, hacking incidents have been reported in the Toronto subway system (where the traveler information was reprogrammed) and the Moscow subway system (where a hacker transferred revenue from the ticketing system).

In response to this emerging challenge, we are trying to model and analyze the interaction between malicious attackers and system operators in this setting. We consider a setting where the attacker can randomize the location and timing of attacks, and the system operator can proactively allocate resources for recovery processes and adaptively react to detected attacks. Since such recovery processes are of less urgent nature than policing and rescuing, throughput is the primary concern (objective function) for the system operator.

Our approach focuses on the three most critical aspects in this problem: adversarial disruptions, response operations, and evolution of traffic queues. Specifically, we are exploring the following questions:

1. For a given response strategy, what is the structure of optimal attack?

2. What are key variables/parameters that effect the network stability (and margin to instability)?
3. How to obtain the optimal response strategy under given attacker constraints and budget?

Preliminary results imply that the attacker does not necessarily focus on the link with the highest load. Instead, the optimal attacking strategy depends on not only the load, but the recovery time, and the travel time between multiple sites.

6.3 Final remarks

In this thesis, we contribute to a system-theoretic approach to a modeling and design framework providing efficiency and resiliency guarantees under a broad class of perturbations. We focus on the operations (vehicle platooning, dynamic routing, and ramp metering) under a broader set of random perturbations including traffic incidents, effects of heterogeneous traffic flow (moving bottlenecks), and demand fluctuations. In addition, we consider the modeling and impact evaluation of security disruptions. We believe that our results are relevant for design of smart highway systems with resiliency guarantees and provide basis for future research on this topic.

Particularly, we emphasize that an important future work is to validate or refine the modeling and design approaches introduced in this thesis through analysis of real traffic data and even lab/field experiments. The key is to identify peculiar characteristics of CAVs and their impact on the aggregate traffic flow, especially under random perturbations and/or security failures. Also of importance is to evaluate the practicality of critical modeling assumptions (e.g. Markovian capacity perturbations) and how sensitive the subsequent analysis/control design is with respect to them.

Bibliography

[1] Mohamed A Abdel-Aty and A Essam Radwan. Modeling traffic accident occurrence and involvement. *Accident Analysis & Prevention*, 32(5):633–642, 2000.

[2] Assad Alam, Bart Besselink, Valerio Turri, Jonas Martensson, and Karl H Johansson. Heavy-duty vehicle platooning for sustainable freight transportation: A cooperative method to enhance safety and efficiency. *IEEE Control Systems*, 35(6):34–56, 2015.

[3] Saurabh Amin, Xavier Litrico, S Shankar Sastry, and Alexandre M Bayen. Stealthy deception attacks on water scada systems. In *Proceedings of the 13th ACM international conference on Hybrid systems: computation and control*, pages 161–170. ACM, 2010.

[4] Saurabh Amin, Galina A Schwartz, and S Shankar Sastry. Security of interdependent and identical networked control systems. *Automatica*, 49(1):186–192, 2013.

[5] David Anick, Debasis Mitra, and Man Mohan Sondhi. Stochastic theory of a data-handling system with multiple sources. *The Bell System Technical Journal*, 61(8):1871–1894, 1982.

[6] Patrick Athol. Interdependence of certain operational characteristics within a moving traffic stream. Technical Report HS-006 579, 1965.

[7] Moshe Babaioff, Robert Kleinberg, and Christos H Papadimitriou. Congestion games with malicious players. In *Proceedings of the 8th ACM conference on Electronic commerce*, pages 103–112. ACM, 2007.

[8] Melike Baykal-Gürsoy, Weihua Xiao, and Kaan Özbay. Modeling traffic flow interrupted by incidents. *European Journal of Operational Research*, 195(1):127–138, 2009.

[9] Michel Benaïm, Stéphane Le Borgne, Florent Malrieu, and Pierre-André Zitt. Qualitative properties of certain piecewise deterministic Markov processes. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 51, pages 1040–1075. Institut Henri Poincaré, 2015.

[10] Carl Bergenhem, Steven Shladover, Erik Coelingh, Christoffer Englund, and Sadayuki Tsugawa. Overview of platooning systems. In *Proceedings of the 19th ITS World Congress, Oct 22-26, Vienna, Austria (2012)*, 2012.

[11] Sharon Bernstein and Andrew Blankstein. Key signals targeted, officials say. *Los Angeles Times.*

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

[13] Franco Blanchini. Survey paper: Set invariance in control. *Automatica (Journal of IFAC)*, 35(11):1747–1767, 1999.

[14] Gabriel Brindusescu. DARPA hacked a chevy impala through its onstar system. *autoevolution.*

[15] Simeon Calvert, Hani Mahmassani, Jan-Niklas Meier, Pravin Varaiya, Samer Hamdar, Danjue Chen, Xiaopeng Li, Alireza Talebpour, and Stephen P Mattingly. Traffic flow of connected and automated vehicles: Challenges and opportunities. In *Road Vehicle Automation 4*, pages 235–245. Springer, 2018.

[16] Simeon C Calvert, A Soekroella, IR Wilmink, and B v Arem. Considering knowledge gaps for automated driving in conventional traffic. In *Proceedings of the Fourth International Conference on Advances in Civil, Structural and Environmental Engineering?ACSEE 2016*, 2016.

[17] Alvaro A Cárdenas, Saurabh Amin, Zong-Syun Lin, Yu-Lun Huang, Chi-Yen Huang, and Shankar Sastry. Attacks against process control systems: risk assessment, detection, and response. In *Proceedings of the 6th ACM symposium on information, computer and communications security*, pages 355–366. ACM, 2011.

[18] Hong Chen and David D Yao. A fluid model for systems with random disruptions. *Operations Research*, 40(3-supplement-2):S239–S247, 1992.

[19] Bertrand Cloez, Martin Hairer, et al. Exponential ergodicity for Markov processes with random switching. *Bernoulli*, 21(1):505–536, 2015.

[20] Giacomo Como, Ketan Savla, Daron Acemoglu, Munther A Dahleh, and Emilio Frazzoli. Robust distributed routing in dynamical networks Part I: Locally responsive policies and weak resilience. *Automatic Control, IEEE Transactions on*, 58(2):317–332, 2013.

[21] Giacomo Como, Ketan Savla, Daron Acemoglu, Munther A Dahleh, and Emilio Frazzoli. Robust distributed routing in dynamical networks Part II: Strong resilience, equilibrium selection and cascaded failures. *Automatic Control, IEEE Transactions on*, 58(2):333–348, 2013.

[22] Samuel Coogan and Murat Arcak. A compartmental model for traffic networks and its dynamical behavior. *IEEE Transactions on Automatic Control*, 60(10):2698–2703, 2015.

[23] Samuel Coogan and Murat Arcak. Stability of traffic flow networks with a polytree topology. *Automatica*, 66:246–253, 2016.

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

[25] Carlos F Daganzo. The cell transmission model, part II: Network traffic. *Transportation Research Part B: Methodological*, 29(2):79–93, 1995.

[26] Carlos F Daganzo and Jorge A Laval. On the numerical treatment of moving bottlenecks. *Transportation Research Part B: Methodological*, 39(1):31–46, 2005.

[27] Jim G Dai. On positive Harris recurrence of multiclass queueing networks: a unified approach via fluid limit models. *The Annals of Applied Probability*, pages 49–77, 1995.

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

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

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

[31] Gunes Dervisoglu, Gabriel Gomes, Jaimyoung Kwon, Roberto Horowitz, and Pravin Varaiya. Automatic calibration of the fundamental diagram and empirical observations on capacity. In *Transportation Research Board 88th Annual Meeting*, number 09-3159, 2009.

[32] Drew Fudenberg and Jean Tirole. Perfect bayesian equilibrium and sequential equilibrium. *journal of Economic Theory*, 53(2):236–260, 1991.

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

[34] Branden Ghena, William Beyer, Allen Hillaker, Jonathan Pevarnek, and J Alex Halderman. Green lights forever: Analyzing the security of traffic infrastructure. *WOOT*, 14:7–7, 2014.

[35] Gregory D Glockner and George L Nemhauser. A dynamic network flow problem with uncertain arc capacities: Formulation and problem structure. *Operations Research*, 48(2):233–242, 2000.

[36] Sergei Konstantinovich Godunov. A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. *Matematicheskii Sbornik*, 89(3):271–306, 1959.

[37] Gabriel Gomes, Roberto Horowitz, Alexandr A Kurzhanskiy, Pravin Varaiya, and Jaimyoung Kwon. Behavior of the cell transmission model and effectiveness of ramp metering. *Transportation Research Part C: Emerging Technologies*, 16(4):485–513, 2008.

[38] BD Greenshields, Ws Channing, Hh Miller, et al. A study of traffic capacity. In *Highway research board proceedings*, volume 1935. National Research Council (USA), Highway Research Board, 1935.

[39] Morris W Hirsch. Systems of differential equations that are competitive or cooperative ii: Convergence almost everywhere. *SIAM Journal on Mathematical Analysis*, 16(3):423–439, 1985.

[40] Serge P Hoogendoorn and Piet HL Bovy. State-of-the-art of vehicular traffic flow modelling. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*, 215(4):283–303, 2001.

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

[42] Roberto Horowitz, Alexander A. Kurzhanskiy, and Matthew Wright. HOT lane simulation tools. Technical report, Institute of Transportation Studies, University of California, Berkeley, CA, 2016.

[43] Saif Eddin Jabari and Henry X Liu. A stochastic model of traffic flow: Theoretical foundations. *Transportation Research Part B: Methodological*, 46(1):156–174, 2012.

[44] Anxi Jia, Billy Williams, and Nagui Rouphail. Identification and calibration of site-specific stochastic freeway breakdown and queue discharge. *Transportation Research Record: Journal of the Transportation Research Board*, (2188):148–155, 2010.

[45] Li Jin and Saurabh Amin. Analysis of a stochastic switching model of freeway traffic incidents. *IEEE Transactions on Automatic Control.* to appear.

[46] Li Jin and Saurabh Amin. Stability of fluid queueing systems with parallel servers and stochastic capacities. *IEEE Transactions on Automatic Control.* to appear.

[47] Li Jin and Saurabh Amin. A piecewise-deterministic Markov model of freeway accidents. In *Decision and Control (CDC), 2014 IEEE 53rd Annual Conference on*. IEEE, 2014.

[48] Li Jin and Saurabh Amin. Calibration of a macroscopic traffic flow model with stochastic saturation rates. In *Transportation Research Board 96th Annual Meeting*, 2017.

[49] Li Jin, Mladen Čičić, Saurabh Amin, and Karl H Johansson. Modeling impact of vehicle platooning on highway congestion: A fluid queuing approach. In *Proceedings of the 21st International Conference on Hybrid Systems: Computation and Control*, New York, NY, USA, 2018. ACM.

[50] Bryan Jones, Lester Janssen, and Fred Mannering. Analysis of the frequency and duration of freeway accidents in Seattle. *Accident Analysis & Prevention*, 23(4):239–255, 1991.

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

[52] Boris S Kerner and Peter Konhäuser. Cluster effect in initially homogeneous traffic flow. *Physical Review E*, 48(4):R2335, 1993.

[53] Jeffrey P Kharoufeh and Natarajan Gautam. Deriving link travel-time distributions via stochastic speed processes. *Transportation Science*, 38(1):97–106, 2004.

[54] A Khattak, X Wang, and H Zhang. Incident management integration tool: dynamically predicting incident durations, secondary incident occurrence and incident delays. *IET Intelligent Transport Systems*, 6(2):204–214, 2012.

[55] Victor L Knoop. *Road Incidents and Network Dynamics: Effects on driving behaviour and traffic congestion.* PhD thesis, Technische Universiteit Delft, 2009.

[56] Karl Koscher, Alexei Czeskis, Franziska Roesner, Shwetak Patel, Tadayoshi Kohno, Stephen Checkoway, Damon McCoy, Brian Kantor, Danny Anderson, Hovav Shacham, et al. Experimental security analysis of a modern automobile. In *Security and Privacy (SP), 2010 IEEE Symposium on*, pages 447–462. IEEE, 2010.

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

[58] Alexander A Kurzhanskiy and Pravin Varaiya. Traffic management: An outlook. *Economics of transportation*, 4(3):135–146, 2015.

[59] Alexandr A Kurzhanskiy. Set-valued estimation of freeway traffic density. In *Control in Transportation Systems*, pages 271–277, 2009.

[60] Alexandr A Kurzhanskiy. Online traffic simulation service for highway incident management. Technical Report SHRP 2 L-15(C), Relteq Systems, Inc., Albany, CA, February 2013.

[61] Alexandr A Kurzhanskiy and Pravin Varaiya. Active traffic management on road networks: A macroscopic approach. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1928):4607–4626, 2010.

[62] Jaimyoung Kwon, Michael Mauch, and Pravin Varaiya. Components of congestion: Delay from incidents, special events, lane closures, weather, potential ramp metering gain, and excess demand. *Transportation Research Record: Journal of the Transportation Research Board*, 1959(1):84–91, 2006.

[63] Jeffrey Larson, Kuo-Yun Liang, and Karl H Johansson. A distributed framework for coordinated heavy-duty vehicle platooning. *Intelligent Transportation Systems, IEEE Transactions on*, 16(1):419–429, 2015.

[64] Aron Laszka, Bradley Potteiger, Yevgeniy Vorobeychik, Saurabh Amin, and Xenofon Koutsoukos. Vulnerability of transportation networks to traffic-signal tampering. In *Cyber-Physical Systems (ICCPS), 2016 ACM/IEEE 7th International Conference on*, pages 1–10. IEEE, 2016.

[65] Chris Lee, Bruce Hellinga, and Frank Saccomanno. Real-time crash prediction model for application to crash prevention in freeway traffic. *Transportation Research Record: Journal of the Transportation Research Board*, 1840(1):67–77, 2003.

[66] John Leonard, Jonathan How, Seth Teller, Mitch Berger, Stefan Campbell, Gaston Fiore, Luke Fletcher, Emilio Frazzoli, Albert Huang, Sertac Karaman, et al. A perception-driven autonomous urban vehicle. *Journal of Field Robotics*, 25(10):727–774, 2008.

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

[68] Michael James Lighthill and Gerald Beresford Whitham. On kinematic waves ii. a theory of traffic flow on long crowded roads. *Proc. R. Soc. Lond. A*, 229(1178):317–345, 1955.

[69] Jennie Lioris, Ramtin Pedarsani, Fatma Yildiz Tascikaraoglu, and Pravin Varaiya. Platoons of connected vehicles can double throughput in urban roads. *Transportation Research Part C: Emerging Technologies*, 77:292–305, 2017.

[70] Hong K Lo and Wai Yuen Szeto. A cell-based variational inequality formulation of the dynamic user optimal assignment problem. *Transportation Research Part B: Methodological*, 36(5):421–443, 2002.

[71] Johan Löfberg. YALMIP: A toolbox for modeling and optimization in MATLAB. In *Computer Aided Control Systems Design, 2004 IEEE International Symposium on*, pages 284–289. IEEE, 2004.

[72] Samer M Madanat, Michael J Cassidy, and Mu-Han Wang. Probabilistic delay model at stop-controlled intersection. *Journal of transportation engineering*, 120(1):21–36, 1994.

[73] Adolf D May Jr. Experimentation with manual and automatic ramp control. *Highway research record*, (59), 1964.

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

[75] Mahalia Miller and Chetan Gupta. Mining traffic incidents to forecast impact. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, pages 33–40, Beijing, China, August 2012. ACM.

[76] G. J. L. Naus, R. P. A. Vugts, J. Ploeg, M. J. G. van de Molengraft, and M. Steinbuch. String-stable cacc design and experimental validation: A frequency-domain approach. *IEEE Transactions on Vehicular Technology*, 59(9):4268–4279, 2010.

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

[78] National Research Council (US). Committee on Freight Transportation Information Systems Security. *Cybersecurity of Freight Information Systems: A Scoping Study*. Number 274. Transportation Research Board, 2003.

[79] Markos Papageorgiou, Habib Hadj-Salem, and Jean-Marc Blosseville. ALINEA: A local feedback control law for on-ramp metering. *Transportation Research Record*, (1320), 1991.

[80] Markos Papageorgiou and Apostolos Kotsialos. Freeway ramp metering: An overview. *IEEE transactions on intelligent transportation systems*, 3(4):271–281, 2002.

[81] Hyoshin Park, Ali Shafahi, and Ali Haghani. A stochastic emergency response location model considering secondary incidents on freeways. *IEEE Transactions on Intelligent Transportation Systems*, (99), 2016.

[82] Michael D Peterson, Dimitris J Bertsimas, and Amedeo R Odoni. Models and algorithms for transient queueing congestion at airports. *Management Science*, 41(8):1279–1295, 1995.

[83] Jonathan Petit and Steven E Shladover. Potential cyberattacks on automated vehicles. *IEEE Trans. Intelligent Transportation Systems*, 16(2):546–556, 2015.

[84] Carolina Osorio Pisano. *Mitigating network congestion: analytical models, optimization methods and their applications*. PhD thesis, Verlag nicht ermittelbar, 2010.

[85] Paul I Richards. Shock waves on the highway. *Operations research*, 4(1):42–51, 1956.

[86] RDW Rijkswaterstaat, the Ministry of Infrastructure, and the Netherlands the Environment. European truck platooning challenge 2016—lessons learnt. www.eutruckplatooning.com, 2016.

[87] Henrik Sandberg, Saurabh Amin, and Karl Henrik Johansson. Cyberphysical security in networked control systems: An introduction to the issue. *IEEE Control Systems*, 35(1):20–23, 2015.

[88] David Schrank, Bill Eisele, and Tim Lomax. TTI's 2012 urban mobility report. *Proceedings of the 2012 annual urban mobility report. Texas A&M Transportation Institute, Texas, USA*, 2012.

[89] Alexander Skabardonis, Karl F Petty, Robert L Bertini, Pravin P Varaiya, Hisham Noeimi, and Daniel Rydzewski. I-880 field experiment: Analysis of incident data. *Transportation Research Record: Journal of the Transportation Research Board*, 1603(1):72–79, 1997.

[90] Kenneth A Small and Jia Yan. The value of "value pricing" of roads: Second-best pricing and product differentiation. *Journal of Urban Economics*, 49(2):310–336, 2001.

[91] Dieter Spaar. Auto, öffne dich. *Sicherheitslücken bei BMWs ConnectedDrive. C*, 5:15, 2015. In German.

[92] A Sumalee, RX Zhong, TL Pan, and WY Szeto. Stochastic cell transmission model (SCTM): A stochastic dynamic traffic model for traffic state surveillance and assignment. *Transportation Research Part B: Methodological*, 45(3):507–533, 2011.

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

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

[95] André Teixeira, Saurabh Amin, Henrik Sandberg, Karl Henrik Johansson, and Shankar S Sastry. Cyber security analysis of state estimators in electric power systems. In *49th IEEE Conference on Decision and Control (CDC). Atlanta, GA. DEC 15-17, 2010*, pages 5991–5998, 2010.

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

[97] Jeremy G VanAntwerp and Richard D Braatz. A tutorial on linear and bilinear matrix inequalities. *Journal of process control*, 10(4):363–385, 2000.

[98] Pravin Varaiya. Smart cars on smart roads: Problems of control. *IEEE Transactions on Automatic Control*, 38(2):195–207, 1993.

[99] Pravin Varaiya. Congestion, ramp metering and tolls. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 366(1872):1921–1930, 2008.

[100] Pravin Varaiya. Max pressure control of a network of signalized intersections. *Transportation Research Part C: Emerging Technologies*, 36:177–195, 2013.

[101] Pravin Pratap Varaiya. *Freeway Performance Measurement System (PeMS), PeMS 9.0.* California PATH Program, Institute of Transportation Studies, University of California at Berkeley, 2009.

[102] John Glen Wardrop. Some theoretical aspects of road traffic research. In *Inst Civil Engineers Proc London/UK/*, volume 1, pages 325–378, 1952.

[103] Joseph A Wattleworth. Peak-period analysis and control of a freeway system. Technical report, Texas Transportation Institute, 1965.

[104] Walter W Wierwille, RJ Hanowski, JM Hankey, CA Kieliszewski, Suzanne E Lee, A Medina, AS Keisler, and TA Dingus. Identification and evaluation of driver errors: Overview and recommendations. Technical report, 2002.

[105] AS Willsky, PK Houpt, SB Gershwin, AL Kurkjian, CS Greene, and EY Chow. Detection of incidents on freeways. In *Decision and Control including the 17th Symposium on Adaptive Processes, 1978 IEEE Conference on*, volume 17, pages 1037–1041. IEEE, 1978.

[106] Matthew Wright, Gabriel Gomes, Roberto Horowitz, and Alex A Kurzhanskiy. A new model for multi-commodity macroscopic modeling of complex traffic networks. *arXiv preprint arXiv:1509.04995*, 2015.

[107] Manxi Wu, Li Jin, Saurabh Amin, and Patrick Jaillet. Signaling game-based misbehavior inspection in v2i-enabled highway operations. *arXiv preprint arXiv:1803.08415*, 2018.

[108] Haining Yu and Christos G Cassandras. Perturbation analysis of feedback-controlled stochastic flow systems. *IEEE Transactions on Automatic Control*, 49(8):1317–1332, 2004.

[109] Guangnan Zhang, Kelvin KW Yau, and Guanghan Chen. Risk factors associated with traffic violations and accident severity in china. *Accident Analysis & Prevention*, 59:18–25, 2013.

[110] Lixian Zhang, El-Kébir Boukas, and James Lam. Analysis and synthesis of markov jump linear systems with time-varying delays and partially known transition probabilities. *IEEE Transactions on Automatic Control*, 53(10):2458–2464, 2008.

[111] RX Zhong, A Sumalee, TL Pan, and William HK Lam. Optimal and robust strategies for freeway traffic management under demand and supply uncertainties: an overview and general theory. *Transportmetrica A: Transport Science*, 10(10):849–877, 2014.