Real-time Operations Management for Emerging Mobility Systems

by

Mojtaba Abdolmaleki

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Civil Engineering)
in the University of Michigan
2023

Doctoral Committee in alphabetical order:

Professor Neda Masoud, Co-Chair
Professor Yafeng Yin, Co-Chair
Professor Xiuli Chao
Professor Jon Lee
Professor Mark Van Oyen

Mojtaba Abdolmaleki

mojtabaa@umich.edu

ORCID iD: 0000-0002-6337-1939

DEDICATION

To my beloved wife who stood by my side through all the ups and downs of my Ph.D.

To my parents - the reason I am who I am today. They have shown so much faith in me and provided endless encouragement.

To the Women, Life, Freedom revolution.

ACKNOWLEDGMENTS

I want to thank everyone who offered support and gave me a hand in this chapter of my life and in preparing this dissertation. Nevertheless, I need more than a rusty memory and the limited space of this page to let me acknowledge everyone. Therefore, the names may be missing herein; nevertheless: I am forever grateful to everyone who guided me to choose the University of Michigan, Ann Arbor, about five years ago and to everyone who has been/has become a part of my life since.

Yafeng and Neda: thank you for the patient guidance, for supporting my education, and for supporting me in moving toward my passion and dreams; It was a pleasure working with both of you.

Xiuli: thank you for all the Sundays of patiently teaching me what I needed to learn, especially in the last year of my Ph.D. Special thanks to you and Tara for putting in extra effort to help me when I needed the help the most.

Mark: Thank you for listening to my concerns and offering support and guidance when I needed advice the most.

Jon: Thank you for teaching me and letting me borrow your book (Matroid Theory) forever.

My committee: thank you again for your insightful comments and suggestions on my dissertation and for supporting and encouraging me. I have learned much from you, your courses, and your work, and it was my pleasure and honor to have you on my dissertation committee.

My family and friends: thank you for bearing with me, sacrificing for me, supporting me, and cheering me up in all these years of hard work. None of these achievements would have been possible without your love and support.

Coffee: my friend, thank you for your support and service to each and every Ph.D. student.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF ACRONYMS

OD Origin-Destination

FCFS First Come, First Served

SRS Square Root Safety

CRP Complete Resource Pooling

LCP Linear Complementarity Problem

LP Linear Programming

QD Quality-Driven

DTMC Discrete Time Markov Chain

CTMC Continuous Time Markov Chain

TSC Traffic Signal Control

GCC Graph Coloring Control

PTAS Polynomial-Time Approximation Scheme

ABSTRACT

Real-time operations management of emerging mobility systems requires designing and developing efficient algorithms to support fast decision-making with provable performance guarantees using minimal resources. This dissertation discusses the challenges of operating ride-hailing systems and centrally-controlled intersections as two examples of large-scale operations in emerging mobility systems and presents policies to improve the system's performance.

The first chapter describes our proposed policy for the operations management of ride-hailing systems. With the development of shared mobility (e.g., ride-hailing systems such as Uber and Lyft), there has been a growing interest in pricing and empty vehicle relocation to maximize system performance. The impatience of passengers during waiting is an important feature of such systems, but it has been neglected in most studies due to the complexities it introduces. In this study, we develop a provably near-optimal dynamic pricing and empty vehicle relocation policy for a ride-hailing system with limited passenger patience. We model the ride-hailing system as a network of double-ended queues. To derive a near-optimal control policy, we first establish a fluid limit of the network in a large market regime where the demand for rides and vehicle supply is large, and show that the fluid-based optimal solution provides an upper bound for the performance of the original ride-hailing system among all dynamic policies. Then, we develop a specific dynamic policy based on the fluid solution for the original problem and investigate the performance of this policy. Among the results, we answer two open questions raised in the literature: (i) the performance of our dynamic policy converges to that of the true optimal value exponentially fast in time when the market size is large [Braverman et al., 2019], and (ii) the passenger loss of our policy decreases to zero exponentially fast when demand size increases [Banerjee et al., 2018a]. Further, we show that our dynamic policy can balance supply utilization and customer waiting times under the Square Root Safety (SRS) rule. The effectiveness of the proposed policy is demonstrated through a numerical experiment using empirical data from DiDi Chuxing.

The second chapter investigates the potential of vehicle motion control to improve intersection control. As the bottleneck of transportation networks, the safety and efficiency of transportation networks depend on intersection control. While intersections take up a small portion of the road segments, more than 50% percent of fatal and injury crashes (combined) occur at or near intersections. Intersection control aims to coordinate conflicting traffic movements under safety and kinodynamic constraints in real-time to maximize intersection throughput or minimize traffic delay. This study aims to optimally control intersections to maximize throughput in the presence of trajectory control. These problems have existed as cornerstones of traffic control for over a century. In the current practice, the task is accomplished by heuristic right-of-way allocation rules or principles. For example, traffic signal control has been the primary means of controlling critical intersections with high traffic demand for almost a century. We leverage the graph coloring techniques to devise a control approach that provides an approximation algorithm to the intersection control problem. For an intersection with a sufficiently large footprint, we prove that the proposed algorithm provides a fully polynomial-time approximation scheme for the throughput maximization problem.

CHAPTER 1

Dynamic Joint Pricing and Empty Relocation Policies For Ride-hailing Systems

With the proliferation of shared mobility (e.g., ride-hailing systems such as Uber and Lyft), there has been an increasing interest in developing optimal policies to maximize the system performance (e.g., revenue, throughput, social welfare) of ridesharing systems. Compared with other resource allocation problems, the modeling and optimization of ride-hailing systems are challenging due to two significant complexities. First, as ride-hailing systems are two-sided markets consisting of drivers and passengers, it is crucial for the model to capture the endogenous system dynamics on each side as well as their interactions. Second, there are spatial and temporal supply externalities that need to be captured in the ride-hailing system, e.g., fulfilling demand in a region can generate supply in its destination in a future time. Furthermore, the passenger waiting experience is important and should be incorporated in the search for optimal solution. Indeed, the platform has to consider the passenger's waiting time and limited patience, as a passenger would leave the system for other alternatives when the waiting time becomes long.

In a ride-hailing network, there is a finite number of vehicles, and passengers arrive randomly and request to move from an origin to a destination. The passenger demand is price-sensitive such that an increase in the platform's announced price for an origin-destination (OD) pair lowers their willingness to ride. The platform matches an arriving passenger with an available driver in the same region upon arrival; and if there is no available driver, the arriving passengers form a queue to be served, typically on a first-come-first-served (FCFS) basis. The passengers waiting in a queue renege the system when their patience runs out. A critical inefficiency hindering system performance in ride-hailing systems is the imbalance generated by the origin-destination demand pattern, which is widely referred to as "OD unbalancedness" [Bimpikis et al., 2019]. In that, the number of vehicles that enter some regions is higher than the demand leaving them, resulting in the accumulation of supply units in the absence of a proper rebalancing strategy. To balance supply with demand, pricing is a commonly used lever [Bimpikis et al., 2019]. However, pricing alone cannot resolve the OD unbalancedness. Rebalancing strategies such as empty vehicle relocation have to be employed to optimize system performance. We illustrate the necessity of a proper empty relocation policy in the presence of pricing schemes in a simple two-region example below.

A two-region example. The following two-region network is often used to illustrate supply-demand imbalance (see e.g., Braverman et al. 2019). Consider the time horizon of morning rush hour with region 1 representing the residential district and region 2 representing the business district. The demand for rides from region 1 to region 2 is price sensitive such that with the price increase, demand for rides decreases. Suppose the demand from region 2 to region 1 is negligible during the morning rush hour, even when the price is very low. In this scenario, a pricing-only policy is ineffective since a low price in region 1 leads to an accumulation of supply in region 2, and a high price leads to no rides in the network and an accumulation of supply in region 1. Therefore, having a joint pricing-empty relocation policy is inevitable to optimize the system performance.

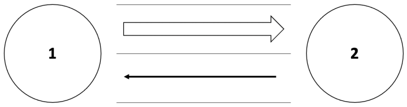


Figure 1.1: Two region city

Because of its practical importance, many research works have been done on optimizing ride-hailing systems (see our literature review in Section 2). For example, Braverman et al. [2019] considers an empty vehicle relocation problem with no pricing decision. The authors develop a static empty vehicle relocation policy and establish its asymptotic optimality. They observe a long convergence time to the optimal performance, and list the development of rigorous time-varying (dynamic) policies as interesting future research to resolve the convergence issue. Banerjee et al. [2018a] considers a dynamic matching problem (with no pricing or empty vehicle relocation) and investigates the asymptotic passenger loss rate. Assuming *"demand arrival rates satisfying an approximate balance condition"*, the authors obtain an exponentially small passenger loss rate when the market size becomes large. They further show that their balance condition is necessary for achieving the exponentially small passenger loss. In the presence of passenger waiting but no passenger reneging, Besbes et al. [2021a] analyzes a stylized model and under a balanced OD demand assumption, the authors find the minimum required fleet size to balance the server utilization and wait times. They show that for their model the square root safety (SRS) rule does not lead to such a balance in spatial systems, and they leave the study on the effects of passenger reneging and OD unbalancedness on the required excess capacity for future research.

In this paper, we close these gaps through a systematic study of a dynamic joint pricing-empty relocation policy in a ride-hailing network. We address the following research questions: (1) How much faster can dynamic policies converge to their equilibrium state than static policies? (2) Can we find policies with tight optimality gaps in finite time and finite supply? (3) Can dynamic policies reduce the expected passenger loss compared to static policies? (4) what is the minimum excess supply to reach a quality-driven (QD) regime in the presence of passenger queues? To address these questions, we formulate the problem as a double-ended queueing network with passenger impatience and develop a dynamic policy based on a simple solution derived from a fluid model. We show that (i) the system performance of the proposed policy converges exponentially fast in time to the true optimal value when the market size is large; (ii) the passenger loss decreases to zero exponentially fast when demand size increases under general OD demand patterns (so no balancing conditions are required); and (iii) with our empty vehicle relocation policy, a near SRS rule achieves the minimum required excess capacity to balance supply utilization and customer waiting times.

Since the ride-hailing problem gives rise to a complex queueing network stochastic process, to develop an efficient control policy we first study a fluid-based model that describes the equilibrium state under a given policy. Then, by conducting a queueing analysis of the ride-hailing system that properly models passengers' patience, we establish a process level convergence of the scaled queue length process to the equilibrium state. We formulate the problem of optimizing the dynamic joint pricing-empty relocation policy as a concave maximization problem, and prove that the resulting policy is optimal in the set of all dynamic joint pricing-empty relocation policies. Lastly, we prove that under our optimized policy, expected passenger loss decreases exponentially fast as the market size tends to infinity.

Major Contributions. At a high level, the major contributions of this paper can be summarized as follows:

(a) Exponentially fast convergence in time.

Banerjee et al. [2021] considers a single-node problem with the objective of maximizing throughput, and the authors show that the gap between the asymptotic utility rate of optimal dynamic policy and the optimal objective value of the fluid-based optimization cannot be less than $\Omega(N^{-1/2})$. In this paper, we consider a general ride-hailing network with an arbitrary utility function (including e.g., revenue, throughput, social welfare). We develop a fluid model for the problem and show that the optimal objective value for the fluid problem is an upper bound to the asymptotic utility rate of all feasible dynamic policies. We propose a dynamic policy for the ride-hailing problem based on the fluid solution and prove that the utility rate for our proposed policy converges to any small neighborhood of $O(N^{-1/2+\epsilon})$ (for any arbitrarily small $\epsilon > 0$) of the true optimal value exponentially fast over time.

The theoretical result of exponentially fast convergence has significant practical implications. It has been observed in the analysis of Braverman et al. [2019] that, while their static policies are asymptotically optimal, the convergence in time is relatively slow; and in practice, the study horizon is shorter than static policies' required convergence time. For example, in the case study of Braverman et al. [2019], the rush hour duration is approximately 2 hours, while the convergence time is approximately 10 hours, as noted in Braverman et al. [2019]:

> *"The Didi dataset we use suggests it is reasonable to assume constant parameters over time windows 1-2 hours in length; see Figure 4. However, numerical experiments suggest that for certain choices of parameters and initial conditions, convergence of our system to equilibrium can occur on timescales on the order of 10 hours. With the rate at which parameters vary, and the slow convergence to equilibrium, the system never really reaches steady-state."*

We close this gap by presenting the first dynamic policy that converges to its equilibrium at an exponentially fast rate in time. Practically, our proposed class of policies is effective even for short time horizon, as observed in our numerical study using the DiDi dataset in Section 1.4.

(b) Exponentially small passenger loss. While there has been an extensive investigation on control policies that reduce the expected passenger loss, the general problem has remained unsolved. Banerjee et al. [2018a] studies this problem under *"an approximate balance condition"* on the demand arrival rates similar to the complete resource pooling (CRP) condition in the queueing literature.

Under the CRP condition, Banerjee et al. [2018a] proposes a set of dynamic policies for the resource pooling problem that reduces the passenger loss exponentially fast in the market size. They further demonstrate that the demand balance condition is necessary for their setting. Additionally, they prove that the best reduction in passenger loss with static policies would be polynomial in market size. We close this gap. Specifically, we develop the first utility-optimized policy and show that it simultaneously decreases the expected passenger loss exponentially fast in the market size without imposing any supply demand condition.

(c) Square Root Safety (SRS) capacity rule for QD regimes. The research literature on waiting probability for passengers has studied the so-called quality-driven (QD) and efficiency-driven (ED) regimes. In the former, the probability of waiting for a passenger is vanishingly small, while in the latter, the probability of waiting for passengers approaches one, as the system scales.

In the presence of passenger waiting but no passenger reneging, Besbes et al. [2021a] shows that the excess supply required for achieving a QD regime is $\Omega(N^{2/3})$, and the authors propose two future research directions: (i) "*Analyzing the case when customers are impatient and might abandon the system if not served after some time is a natural extension",* and (ii) "*Another interesting extension is to study how the results in this study can be generalized to cases where origin-destination demand patterns generate imbalances in the system. In this case, the additional workload stemming from pickups might be even larger. How would this impact capacity sizing?"*

This paper tries to answer these questions in its different setting through optimal empty vehicle relocation in a double-ended queueing network. Our result shows that a simple dynamic relocation policy achieves a QD regime with an excess supply of $\Omega(N^{1/2+\varepsilon})$ (ε can be arbitrarily small), or it (almost) follows the SRS rule, when passengers have finite patience.

Methodological contributions. This paper also makes a number of methodological contributions. Among others, it establishes the convergence of a family of dynamical systems and the ergodicity for a family of infinite-state Markov chains. These contributions are summarized as follows:

(i) Network-level control and passenger queues. Previous studies on ride-hailing systems under limited patience behaviors (completely impatient/patient) have recognized the challenge of incorporating passengers' patience time. When coupled with ride-hailing controls like empty relocation policies, further complexity rises [Besbes et al., 2021a]. In this paper, we formulate a framework to model passengers' patience using double-ended queues and extend this contribution by proposing a set of dynamic policies to optimize the system performance.

(ii) Unconditional process-level convergence. We develop a fluid model that describes the equilibrium state under the given policy, and establish process-level convergence of the scaled queue process to the equilibrium state. Specifically, we prove an $O(N^{-1/2})$ convergence rate for the scaled queue length process. We believe this is the first unconditional convergence result for the joint pricing and empty vehicle relocation (for empty vehicle routing with no pricing decisions, this has been established in Braverman et al. 2019). Furthermore, this paper is the first to conduct a rigorous analysis of the passenger queueing process under general demand patterns.

(iii) A family of LCP-based Lyapunov functions. We introduce a Lyapunov function approach to establish the convergence of the fluid model based on the solution to a *linear complementary problem.* To the best of our knowledge, we are the first to utilize such an approach to establishing the stability of dynamical systems. Moreover, due to its unique feature, this approach is a valuable toolbox for studying networks of double-ended queues in other settings/applications. In this paper we prove the ergodicity result of the resulting CTMC by adopting a Foster-Lyapunov approach.

(iv) Network-level analysis of double-ended queues. We are the first to generalize the analysis and control of double-ended queues to a network setting. This is a challenging task partly because establishing the convergence results for the dynamical system that describes the evolution of the fluid model is relatively challenging; see Liu [2019] for analysis of single-node double-ended queue.

Organization. The remainder of this paper is organized as follows. In Section 1.1, we discuss the relevant literature. Section 1.2 presents the problem formulation and our proposed dynamic policy. In Section 1.3, we present our main results on the performance of our optimized fluid-based dynamic policy and discuss their managerial insights. In Section 1.4, we report the results of several numerical experiments on a real dataset. In Section 1.5, we outline the proof of the main theoretical results. We conclude the paper with a discussion about future research in Section 2.7. Finally, all the technical proofs are presented in the Appendix. Throughout the paper, for any positive integer n, we denote by D^n the set of all Cadlag functions $x : \mathbb{R}^+ \to \mathbb{R}^n$, right continuous functions with left limit. Also for a function $f : \mathbb{R} \to \mathbb{R}^n$, we use $\dot{f}(t) : t \in \mathbb{R}$ to denote the derivative of f at t when it exists. For given positive functions $f(t)$ and $g(t)$, the notation $f(t) = O(g(t))$ means that there exists some constant c such that $f(t) \leq cg(t)$ for all $t \geq 0$, while $f(g) = o(g(t))$ means $\lim_{t\to\infty} f(t)/g(t) = 0$. For any real number x, we denote $x^+ = \max\{x, 0\}$ and $x^- = \max\{-x, 0\}$. Finally, for any $x(t) = (x_1(t), \ldots, x_n(t)) \in D^n$ with $0 \leq t \leq T$, the infinite norm is denoted by

$$\|x\|_T = \max_{1\leq i\leq n} \sup_{0\leq t\leq T} |x_i(t)|.$$

1.1 Literature Review

The closest works related to ours are Braverman et al. [2019], Banerjee et al. [2021], and Banerjee et al. [2018a]. We first review these several papers in relationship with this paper and then discuss the main streams of research related to our work. In the review below, we use N to denote market size.

In Braverman et al. [2019], the authors study optimization of empty vehicle relocation policies in a setting where a passenger immediately leaves the station when no driver is available. There is no pricing decisions. They propose a family of static fluid-based policies to optimize the system-wide performance and design a fluid-based convex optimization problem to optimize the policy. They further show that the optimal static empty relocation policy asymptotically outperforms all dynamic empty relocation policies when the market size grows. They observe that static policies demonstrate a poor convergence rate in time, so they propose a set of dynamic heuristics to address this issue and leave a rigorous analysis of dynamic policies and their potential in improving the convergence rate in time to future research. Also, assuming that passengers do not wait, Banerjee et al. [2021] considers the problem of joint pricing-empty relocation optimization to maximize the system-wide utility at equilibrium. They prove that their optimized static policy attains an $O(N^{-1/2})$ optimality gap in the infinite horizon. Compared with these two papers, our problem involves joint pricing and empty vehicle relocation, and passengers can wait but with finite patience time, leading to double-ended queues. Our performance guarantees are valid for a finite time and finite supply, and our Lyapunov approach shows that our proposed dynamic pricing-empty relocation policy attains an $O(N^{-1/2+\epsilon} + e^{-\alpha_M T})$ optimality gap, where T represents the time, ϵ is an arbitrary small positive number and α_M is a positive constant independent of N and T.

Banerjee et al. [2018a] considers a dynamic matching problem with no pricing or empty vehicle relocation, and the objective is to minimize the asymptotic passenger loss. In particular, upon arrival of a passenger demand at a node, the decision maker can choose the empty vehicle from an adjacent node to instantaneously serve the demand. This is similar to empty vehicle relocation with zero traveling time from a neighboring region. Similar to Braverman et al. [2019], Banerjee et al. [2021] consider a setting where passengers immediately leave the station when no driver is available. They prove that no static policy results in passenger loss less than $\Omega(N^{-2})$. Then, under the condition of "complete resource pooling" (CRP), they propose a dynamic policy and show that the passenger loss under the proposed policy drops to zero exponentially fast in the market size N. The authors further prove that the CRP condition is necessary for achieving the exponentially fast passenger loss and point out that the empirical data from Manhattan violates the CRP condition. In contrast with Banerjee et al. [2018a], our pricing and empty vehicle relocation relaxes zero traveling time assumption between neighboring regions for empty vehicle relocation and drops the CRP condition. Nonetheless, we prove that the passenger loss under our policy decreases to zero exponentially fast. Another key difference is that this paper analyzes the loss of finite time horizon system performance. This is an extension to and more general than the infinite time horizon loss considered in Banerjee et al. [2018a].

Besbes et al. [2021a] investigates capacity planning for an asymptotically optimal control policy (nearest neighbor) in a stylized setting that approximates the network dynamics by considering passengers arriving with origin/destinations uniformly drawn from a uniform distribution in a two-dimensional square. Assuming the passengers never leave the system, they obtain the minimum excess supply required for achieving the QD regime is $\Omega(N^{2/3})$. They leave the question of modeling the passenger reneging and the effect of OD unbalancedness for future research. In comparison, our network-level double-ended queueing analysis enables us to consider the passenger reneging and model network dynamics. Also, we provide the minimum excess supply required for achieving a QD regime under our proposed pricing-empty relocation policy. Specifically, we show that by enabling empty relocation, we can achieve the QD regime with an excess of supply $\Omega(N^{1/2})$.

Closed queueing network control in ride-hailing systems. There is a young but growing literature on the applications of closed queueing network control in ride-hailing systems. We categorize the literature based on their modeling approach into two main streams:

The first stream approximates the endogenous system dynamics by aggregated exogenous Poisson processes. In this stream of research, the queueing analysis is built on the fixed population mean idea introduced by Whitt [1984] where they cut each link in the closed queueing network and assume, in aggregate, the individual driver decisions yield a Poisson arrival to each region. For example, Brooks et al. [2013] utilizes this approach to derive policies for matching debris removal vehicles to routes following natural disasters. However, it only provides performance guarantees in restricted settings. Özkan and Ward [2020] exploits aggregated approximation to devise continuous linear programming-based policy for the joint pricing and matching decisions. The second stream considers system performance when supply and demand, referred to as the market size, simultaneously grow. These studies propose asymptotically optimal policies derived from a set of fluid optimization problems and provide performance guarantees. Moreover, they prove that the optimality gap for the proposed fluid-based policies tightens as the market size increases. The performance guarantees in initial investigations in this stream are asymptotic both in time and market size [George and Xia, 2011, Iglesias et al., 2019, Zhang and Pavone, 2016, Waserhole and Jost, 2016, Iglesias et al., 2019, Braverman et al., 2019]. Banerjee et al. [2021] and Benjaafar and Shen [2022] investigate the problem of optimizing pricing and empty relocation policies and provide performance guarantees that are asymptotic in time for a finite-size market. Kanoria and Qian [2019] investigates the pricing and assignment decision optimization problem to maximize system-wide utility when pickup and service time are instantaneous. They propose a family of Mirror Backpressure control policies that provide performance guarantees when the demand pattern changes smoothly in time, independent of information on the exact demand pattern. They require a strong connectivity assumption on the demand pattern and demonstrate that this constraint is less restrictive than CRP. Our extensions are mainly motivated by the second stream of research. The introduction of passenger queues with abandonment enables us to develop a dynamic policy that converges exponentially fast to its equilibrium and the passenger loss decreases to zero exponentially fast with market size.

Double ended queueing models. A separate line of research on stylized models in ride-hailing systems aims to formulate stylized queueing models that relate drivers' closed queueing network and passengers' open queueing network by approximating endogenous system dynamics such as pick-up time and travel time via aggregated exponential distributions. Besbes et al. [2021a] studies higher-level strategic capacity sizing. Wang et al. [2019] and Castillo et al. [2017], assuming fixed pricing, study admission control based on a pickup-time threshold in a two-sided model with an open rider-side queue and a closed driver-side queue that captures space in reduced form. Compared to the double-ended queueing models, our network-level double-ended queueing analysis enables us to study the performance of fine-grained optimized policies that outperforms asymptotically optimal coarse-grained policies like the nearest neighbor.

Fluid-based models. Another stream considers the fluid-based models where supply units are infinitesimally small. Compared to the first approach, these models assume a different approach in modeling the driver repositioning. Bimpikis et al. [2019] studies pricing under steady-state conditions in a network where drivers behave in equilibrium and decide whether and when to provide service and reposition. They are the first to identify the "balancedness" property of the demand pattern and investigate the potential benefits of employing spatial price discrimination to maximize the system-wide utility. Afeche et al. [2018] studies the demand admission controls and drivers' repositioning in a two-location network without pricing and indicates that the value of the controls is significant when capacity is moderate, and demand is imbalanced. Ma et al. [2021] proposes an optimal and incentive-aligned spatio-temporal pricing mechanism for the finite-horizon problem with complete information. Özkan [2020] studies asymptotically optimal policies for assigning customers to nearby drivers in an open and time-varying system.

Empirical studies. Papers in this stream use ride-hailing data or taxi data. For example, using Uber data, Chen and Sheldon [2016] demonstrates that surge pricing persuades drivers to work longer and consequently increases efficiency; Hall et al. [2017] finds that the driver supply is highly elastic to wage and underlying fare changes. Yan et al. [2020] review operational matching and dynamic pricing techniques and discuss a dynamic waiting mechanism inspired by Uber. Using NYC taxi data, Buchholz [2015] and Ata et al. [2019] analyze the dynamic spatial equilibrium with strategic taxi drivers and study the effects of matching and spatial pricing on performance. Buchholz [2022] indicates that matching technology can improve performance significantly even under optimized pricing, which supports the value of studying the impact of operational controls.

Other related literature on ride-hailing. In addition to the above literature, extensive literature focuses on specific controls on ride-hailing systems such as pricing, matching, empty relocation, and their joint controls. Interested reader is referred to, e.g., [Balseiro et al., 2021, Besbes et al., 2021b, Guda and Subramanian, 2019, Banerjee et al., 2015, Hu and Zhou, 2022, Özkan, 2020, Gurvich and Ward, 2015].

1.2 Problem Formulation

The ride-hailing system under consideration consists of $S > 0$ regions, and each region is called a node. There is a fixed number of vehicles in the system, which we denote by N. We model the system as a network of infinite-server and double-ended queues. The vehicles traveling or empty relocating between two nodes are modeled as an infinite-server queue, with travel time being the service time. The vehicles or passengers waiting in a node are modeled as a double-ended queue, i.e., the passenger and vehicle queues do not coexist simultaneously[1]. For convenience we denote the set of double ended queues by $\mathcal{S} = \{1, 2, \ldots, S\}$, and denote the set of infinite server queues, or the links connecting nodes, by $\mathcal{I}$. We use $i \in \mathcal{S}$ to denote node, and use $ij \in \mathcal{I}$ or $(i, j) \in \mathcal{I}$, denote the connecting links.

It is worthy noting that, viewing the system from the standpoint of vehicles, vehicles circulate through the network neither enter nor exit the system, similar to that in a closed queueing network. On the other hand, when viewing from the standpoint of passengers, the passengers in each queue

[1] When an empty vehicle is matched with a passenger, the time it takes to pick up the passenger is modeled as part of the traveling time between the passenger's origin and destination.

leave the system after being served or reach the limit of their waiting patience, so they constitute an opening queueing network with customer reneging.

Passenger demand. Passenger demand for O-D pair (i, j) occurs according to a Poisson process with a price-sensitive rate. Upon the arrival of a passenger at node i, if there is an empty vehicle available in the node, we match them immediately. If no empty vehicle is available, arriving passengers form a queue in that node and are served according to the FCFS rule (that is, when an empty vehicle becomes available at the node, it serves the customer at the head of the passenger queue). The waiting passengers have limited patience, and a passenger waiting at node i reneges the system after an exponentially distributed time with rate θ_i if still not yet matched. We assume that $c_j^W(\$/hr)$ is the per-hour *waiting cost* for a passenger in node j. Similar to passengers, when vehicles arrive at node i with no passenger waiting, they also form a queue and are matched to the arriving passengers on an FCFS basis.

Traveling time. We assume that the traveling times from node i to node j, for both occupied and empty vehicles, are random with a general probability distribution; we denote the mean traveling time by $1/\mu_{ij}$, $ij \in \mathcal{I}$.

State of system. To keep track of the system dynamics, we define the system state as

$$X(t) = (X_i(t), X_{ij}(t), Z_{ij}(t) : \ i \in \mathcal{S}, (i, j) \in \mathcal{I}),$$

where $X_{ij}(t) \geq 0$ and $Z_{ij}(t) \geq 0$ denote the number of full vehicles and number of empty vehicles traveling from node i to node j at time t, and if $X_i(t) \geq 0$ then it denotes the number of drivers waiting in node i at time t, if $X_i(t) \leq 0$ then $-X_i(t)$ is the number of passengers waiting in node i at time t. Note that under such a definition, the resulting stochastic process is Markovian only when the traveling times are exponentially distributed (otherwise it can be augmented to become Markovian). Let the state space be denoted by $\mathcal{X}$. **Pricing.** Let c_{ij} denote the announced origin-destination price from node i to node j. When the price is c_{ij}, the passenger demand rate is $\lambda_{ij}(c_{ij})$ which is assumed to be strictly decreasing. Thus, there is a one-to-one correspondence between the platform's announced price and the demand rate for each pair of nodes; we denote its inverse function by $c_{ij}(\lambda_{ij})$. With some abuse of notation, we interchangeably use $c_{ij} = c_{ij}(\lambda_{ij})$ and $\lambda_{ij} = \lambda_{ij}(c_{ij})$, and $\boldsymbol{c} = (c_{ij})_{\mathcal{S}\times\mathcal{S}}$ and $\boldsymbol{\lambda} = (\lambda_{ij})_{\mathcal{S}\times\mathcal{S}}$. All these decisions can be time and state dependent.

Empty vehicle relocation. We model empty vehicle relocation by two sets of parameters: a time-to-relocate vector $\boldsymbol{\gamma} = (\gamma_i; \ i \in \mathcal{S})$, and a relocation probability matrix $\boldsymbol{Q} = (q_{ij})_{\mathcal{S}\times\mathcal{S}}$. Specifically, we assume that after a vehicle dwells in node i for exponentially distributed amount of time with mean $1/\gamma_i$, if still not matched with a passenger yet, it will be relocated to node j with probability $q_{ij}, j \in \mathcal{S}$. Here, both $\boldsymbol{\gamma}$ and $\boldsymbol{Q}$ are decisions[2] to be made and they can be time and state dependent. Clearly, $\sum_{j \in \mathcal{S}} q_{ij} = 1, i \in \mathcal{S}$, and q_{ii} denote the probability for the vehicle to continue to wait at node i for arriving passengers. There is a vehicle relocation cost, defined by $c^V_{ij}(\$/trip)$, that is the per-trip cost for relocating an empty vehicle from node i to node j.

Performance measures. The system measure we consider includes a reward component associated with the trips the platform serves and a cost element associated with empty vehicle relocation and passenger waiting. we denote the per-ride reward function by $I_{ij} : \mathbb{R} \to \mathbb{R}$, which corresponds to the reward obtained from a customer taking a ride from i to j. Considering the bijection that relates the price and the rate of passenger requests for each pair of nodes, we use $I_{ij}(c_{ij})$ and $I_{ij}(\lambda_{ij})$ interchangeably to refer to the reward per ride.

Three important canonical reward functions often considered in the literature are

i) Throughput: $I_{ij}(\lambda_{ij}) = 1, ij \in \mathcal{I}$

ii) Revenue: $I_{ij}(\lambda_{ij}) = c_{ij}(\lambda_{ij}), ij \in \mathcal{I}$

iii) Social Welfare: $I_{ij}(\lambda_{ij}) = \mathbb{E}[\lambda_{ij}(\tilde{c}_{ij})\tilde{c}_{ij}|\tilde{c}_{ij} \geq c_{ij}(\lambda_{ij})], ij \in \mathcal{I}$

The throughout rate for a ride is always 1 regardless of its price, and it is concerned with the number of trips served. The revenue rate associated with a ride is the price the rider pays $c_{ij}(\lambda_{ij})$. For social welfare, we follow the definition of Banerjee et al. [2021] and Benjaafar and Shen [2022], where $\mathbb{E}[\lambda_{ij}(\tilde{c}_{ij})\tilde{c}_{ij}|\tilde{c}_{ij} \geq c_{ij}(\lambda_{ij})]$ represents the average trip valuation for the passengers willing to pay the platform announced price $c_{ij}(\lambda_{ij})$. Following the ride-hailing literature (see e.g., Banerjee et al. 2021, Bimpikis et al. 2019, Kanoria and Qian 2019), we assume the average revenue function $\lambda_{ij} I_{ij}(\lambda_{ij})$ is concave in λ_{ij}.

To compute the objective function, we let $K_{ji}(t, T)$ denote the number of passengers' request from node j to node i by time t and completed by time T. Also, we let $J_{ji}(t)$ denote the number of empty vehicles relocated from node j to node i by time t. Then, the objective function that we intend to maximize is

$$U(T, N) = \frac{1}{N}\mathbb{E}\left[\sum_{ji \in \mathcal{I}} \int_0^T I_{ji}(\lambda_{ji}(t)) dK_{ji}(t, T) - \sum_{ji \in \mathcal{I}} \int_0^T c^V_{ji} dJ_{ji}(t) - \sum_{j \in \mathcal{S}} \int_0^T c^W_j X^-_j(t) dt\right]. \tag{1.1}$$

The first term on the right hand side is the reward received from the served trips by time T, the second term represents the total empty vehicle relocation cost, while the last is total passenger waiting cost. Clearly, (1.1) is the expected total net-utility value up to time T per vehicle (e.g., total profit if I represents revenue).

[2] An alternative but equivalent definition for relocation decision is $v_{ij} = \gamma_i q_{ij}$, i.e., each empty vehicle in node i is relocated to j after exponentially distributed amount of time with mean $1/v_{ij}$ if still not yet matched.

Objective. Our goal is to find the pricing $\boldsymbol{c}$, the relocation rates rates $\boldsymbol{\gamma}$, and the relocation probabilities matrix $\boldsymbol{Q}$, to maximize the objective function (1.1).

1.3 Main Results

The optimization problem as presented in the previous section is too complex to solve. Indeed, finding the exact optimal policy for the problem is challenging even in its very special cases. For example, Braverman et al. [2019] considered a network with two nodes and fixed static pricing, and no passenger reneging, and they demonstrated that optimizing the relocation policy with $N = 2000$ vehicles is not solvable to optimality with a commercial solver. Our approach in this paper is to first analyze a simplified version of the problem and derive a policy for the simplified problem. Then, we develop a solution for the original problem based on that of the simplified problem. Finally, we prove that the solution for the original problem performs very well and near-optimality. This simplified model is the fluid limit of the original network when the market size is large.

1.3.1 Fluid optimization and construction of dynamic policy

We analyze the problem with a large market size N. To that end, we define the scaled queue length process as queue lengths divided by N. From now on, to emphasize the dependency on the market size, whenever necessary we use a superscript (N) to refer to the system when the number of vehicles in system is N and demand arrival rates are $N\lambda_{ji}(c_{ji})$, $ji \in \mathcal{I}$.

Given a joint pricing-empty relocation policy, we first present a set of constraints that explicitly characterizes the equilibrium for the scaled queue length process. Denote by $f_j, j \in \mathcal{S}$, as the equilibrium portion of vehicles waiting in the queue in node j, and by $f_{ji}, e_{ji}, (j,i) \in \mathcal{I}$ the number of full and empty vehicles traveling from node j to i. Let β_j denote the average passenger queue length at node j. Then, the constraints that characterize the fluid model are

$$\sum_{kj\in\mathcal{I}}\mu_{kj}f_{kj}+\sum_{kj\in\mathcal{I}}\mu_{kj}e_{kj}=\sum_{ji\in\mathcal{I}}\mu_{ji}f_{ji}+\sum_{ji\in\mathcal{I}}\mu_{ji}e_{ji};\quad \forall j\in\mathcal{S}\qquad (\textit{vehicle flow balance})\qquad (1.2a)$$

$$\mu_{ji}f_{ji}\le\lambda_{ji};\quad \forall ji\in\mathcal{I}\qquad (\textit{passenger flow})\qquad (1.2b)$$

$$\mu_{ji}f_{ji}\sum_{k\in\mathcal{S}}\lambda_{jk}=\lambda_{ji}\sum_{k\in\mathcal{S}}\mu_{jk}f_{jk};\quad \forall j\in\mathcal{S}\qquad (\textit{rate balance})\qquad (1.2c)$$

$$\Big(\sum_{ji\in\mathcal{I}}\lambda_{ji}-\sum_{ji\in\mathcal{I}}\mu_{ji}f_{ji}\Big)f_j=0;\quad \forall j\in\mathcal{S}\qquad (\textit{complementary slackness})\qquad (1.2d)$$

$$\beta_j=(\sum_{ji\in\mathcal{I}}\lambda_{ji}-\sum_{ji\in\mathcal{I}}\mu_{ji}f_{ji})/\theta_j;\quad \forall j\in\mathcal{S}\qquad (\textit{Little's law})\qquad (1.2e)$$

$$\sum_{ji\in\mathcal{I}}f_{ji}+\sum_{ji\in\mathcal{I}}e_{ji}+\sum_{j\in\mathcal{S}}f_j=1\qquad (\textit{unit mass})\qquad (1.2f)$$

$$f_{ji},e_{ji},f_j,\lambda_{ji},\beta_j\ge 0\qquad (\textit{non-negativity})\qquad (1.2g)$$

Constraint (1.2a) ensures the vehicle flow balance for each node $j\in\mathcal{S}$. This is because the inflow rates for full vehicles and empty vehicles from node k to node j equals $\mu_{kj}f_{kj}$ and $\mu_{kj}e_{kj}$, respectively. Constraint (1.2b) guarantees that full vehicle departure rates do not exceed the passenger request rate from node j to i. Constraint (1.2c) ensures that the probability a full vehicle departing node i for destination j is equal to the probability a passenger requesting a ride in node i for destination j. The complementarity constraint (1.2d) implies that passenger and driver queues do not coexist in a node. The constraint (1.2e) represents Little's law for passenger queue in node i, i.e., the passenger queue length β_j equals the product of their average waiting time $1/\theta_j$ and the passenger reneging rate $\sum_{i,ji\in\mathcal{I}}(\lambda_{ji}-\mu_{ji}f_{ji})$. Finally, constraint (1.2f) is the unit mass constraint, and (1.2g) is the standard non-negativity constraint. A solution that satisfies constraints (1.2a) to (1.2g), denoted by $\mathbf{f}=(\boldsymbol{\lambda},\boldsymbol{f},\boldsymbol{e},\boldsymbol{\beta})$ is called a fluid solution.

We formulate the fluid optimization problem as maximizing objective function (1.1) subject to the constraints that characterize the fluid scaled queue lengths at equilibrium, which in equilibrium can be expressed as

$$\max_{\boldsymbol{\lambda},\boldsymbol{f},\boldsymbol{e},\boldsymbol{\beta}}\quad \sum_{ji\in\mathcal{I}}\mu_{ji}f_{ji}I_{ij}(\lambda_{ji})-\sum_{ji\in\mathcal{I}}c^V_{ji}\mu_{ji}e_{ji}-\sum_{j\in\mathcal{S}}c^W_j\beta_j\qquad (1.3)$$

$$\text{s.t.}\quad 1.2a-1.2g\qquad (1.4)$$

To show the equivalence between (1.1) and (1.3) in equilibrium, we first note that Little's law applied to the infinite-server queues implies $\lim_{T\to\infty}K_{ji}(T)/T=\mu_{ji}f_{ji}$ and $\lim_{T\to\infty}J_{ji}(T)/T=$

$\mu_{ji}e_{ji}$. For the equality of third terms in (1.1) and (1.3), it suffices to note that e_{ji} represents the equilibrium scaled queue length process for Z_{ji}, and β_j is the equilibrium scaled queue length process for X_j^-.

The optimization problem (1.3) and (1.4) is not a convex optimization problem as the objective function is not concave, and the constraints do not give rise to a convex set because of the complementarity slackness condition. We next show that the problem can be converted to a convex optimization problem. In the following lemma, we first show that some conditions are always satisfied by the optimal solution of (1.3) and (1.4), thus we can modify the objective function as well as the constraints using them. Its proof, as well as all other omitted proofs, are given in the Appendix.

Lemma 1. *Any optimal solution* $\mathrm{f} = (\boldsymbol{\lambda}, \boldsymbol{f}, \boldsymbol{e}, \boldsymbol{\beta})$ *to the fluid optimization problem (1.3) and (1.4) satisfies* $\mu_{ji}f_{ji} = \lambda_{ji}$ *for all* $ji \in \mathcal{I}$.

Before discussing the implications of Lemma 1 on the optimization problem, we first shed some lights on why this result is expected. When passenger arrival rate is higher than full vehicle departure rate, we can slightly increase the price to lower the passenger arrival rate, serving the same number of customers with a higher objective function value. Thus at optimum, arrival rate and departure rate balance. Also note that, Lemma 1 implies $\beta_j = 0$ for $j \in \mathcal{S}$, which shows that, under an optimal solution, the number of customers waiting in the fluid model is zero in every node.

To simplify the optimization problem, we consider (1.3) and (1.4) within the subset of feasible region that satisfies $\mu_{ji}f_{ji} = \lambda_{ji}$. Then, the optimization problem becomes

$$\begin{aligned} \max_{\boldsymbol{\lambda}, \boldsymbol{f}, \boldsymbol{e}} \quad & \sum_{ji \in \mathcal{I}} \lambda_{ji} I_{ji}(\lambda_{ji}) - \sum_{ji \in \mathcal{I}} c_{ji}^V \mu_{ji} e_{ji} \\ \text{s.t.} \quad & \sum_{kj \in \mathcal{I}} \lambda_{kj} + \sum_{kj \in \mathcal{I}} \mu_{kj} e_{kj} = \sum_{ji \in \mathcal{I}} \lambda_{ji} + \sum_{ji \in \mathcal{I}} \mu_{ji} e_{ji}; \quad \forall j \in \mathcal{S} \\ & \sum_{ji \in \mathcal{I}} \lambda_{ji}/\mu_{ji} + \sum_{ji \in \mathcal{I}} e_{ji} + \sum_{i \in \mathcal{S}} f_i = 1, \\ & e_{ji}, f_j, \lambda_{ji} \geq 0 \end{aligned}$$

This optimization problem only has decision variables λ_{ji}, e_{ji} for $ji \in \mathcal{I}$, and f_i for $i \in \mathcal{S}$. Further, by our assumption, the objective function is a separately concave function of λ_{ji} and linear in e_{ji}. Hence it is a convex optimization problem that can be efficiently solved.

After solving the convex optimization problem with optimal solution λ_{ji}, e_{ji} for $ji \in \mathcal{I}$, we set $f_{ji} = \lambda_{ji}/\mu_{ji}$ for all $ji \in \mathcal{I}$, and $\beta_i = 0$ for all $i \in \mathcal{S}$ to obtain a feasible solution for problem (1.3) and (1.4), which is also the optimal solution for the fluid model.

Construction of the fluid-based dynamic policy. After obtaining the optimal solution for problem (1.3) and (1.4), we are ready to construct a dynamic pricing and empty vehicle relocation policy for our original problem (1.1).

As the first step to finding a dynamic policy, we set the price to $c_{ji}(\lambda_{ji})$, which leads to arrival rate of λ_{ji} for the passenger requests from node j to node i. We define

$$q_{ji} = \frac{\mu_{ji} e_{ji}}{\sum_{k,\ jk \in \mathcal{I}} \mu_{jk} e_{jk}}, \quad ji \in \mathcal{I} \tag{1.5}$$

when the denominator is not zero, and $q_{ji} = 0$ otherwise. This gives the empty relocation policy $\boldsymbol{Q}$. Lastly, let

$$\gamma_i = \frac{\sum_{k,\ ik \in \mathcal{I}} \mu_{ik} e_{ik}}{f_i}, \qquad i \in \mathcal{S} \tag{1.6}$$

be the dwell time parameter at node i before relocating the empty vehicles.

Remark 1. *It is clear from (1.6) that the policy requires $f_i > 0$ for all i. We point out that this is not necessary. In the case $f_i = 0$, we can introduce a network transformation to make the policy feasible and all the results in the paper remain to hold. For details the reader is referred to Section A.2 in the Appendix.*

Why is this pricing-empty vehicle relocation policy dynamic? Our proposed policy is dynamic because the relocation rate of empty vehicles in a node depends on the number of empty vehicles in the node. To see that, note that the rate at which we relate an empty vehicle at node i is γ_i. This means that when there are n_i empty vehicles in the node i at time t, the total rate of relocating an empty vehicle to other nodes is $n_i\gamma_i$, which is linear in the number of empty vehicles waiting.

The key question is how well does this dynamic policy perform in terms of system measure. This is answered in the following subsection.

1.3.2 Theoretical performance guarantees

We presented a fluid model in the previous subsection, but have not established its formal connection with the original problem. In this subsection, we first show that the model is indeed the fluid limit of the original problem when the market size becomes large.

Let $(\boldsymbol{\lambda}, \boldsymbol{f}, \boldsymbol{e}, \boldsymbol{\beta})$ be a feasible solution for the fluid optimization problem (1.2) and let $(\boldsymbol{Q}, \boldsymbol{\gamma}, \boldsymbol{c}(\boldsymbol{\lambda}))$ be its associated dynamic policy described in the previous subsection. We first show that, under the policy $(\boldsymbol{Q}, \boldsymbol{\gamma}, \boldsymbol{c}(\boldsymbol{\lambda}))$, the scaled queue length process $X^{(N)}(T)/N$ of the original stochastic system converges to the fluid solution $(\boldsymbol{\lambda}, \boldsymbol{f}, \boldsymbol{e}, \boldsymbol{\beta})$ as we scale the market size N. Additionally, we will unveil the relationship between the convergence speed, time, and the market size.

We make the following assumption, which we impose throughout the paper.

Assumption 1. A matrix $\tilde{\boldsymbol{Q}}$ of dimension $|\mathcal{S}|$ with its j-th row being the j-th row of either matrix $\boldsymbol{\lambda}$ or matrix $\boldsymbol{Q}$ is irreducible.

Note that when all the rows of $\tilde{\boldsymbol{Q}}$ are that of matrix $\boldsymbol{\lambda}$, then Assumption 1 implies that the demand from each node should be able to reach any other node, though that may need to go through other nodes. On the other hand, when all the rows of $\tilde{\boldsymbol{Q}}$ are that of $\boldsymbol{Q}$, then it means that empty vehicle allocation from any node is reachable to any other node as well, though that may need many transitions. These are plausible assumptions. We remark that Assumption 1 is weaker than those in Braverman et al. [2019], where the authors assume that all entries of $\boldsymbol{\lambda}$ and $\boldsymbol{Q}$ are strictly positive, though they conjecture (in their Remark 2) that their condition can be relaxed.

Theorem 1. ***(Convergence to fluid model)*** *Let* $\mathbf{f} = (f_j, f_{ji}, e_{ji}, \lambda_{ji}; j \in \mathcal{S}, ji \in \mathcal{I})$ *be a feasible solution to the fluid optimization problem (1.3) and (1.4). Consider the associated dynamic policy* $(\boldsymbol{Q}, \boldsymbol{\gamma}, \boldsymbol{c}(\boldsymbol{\lambda}))$ *described in Section 1.3.1. There exists a unique* $x = (x_{ji}(t), z_{ji}(t), x_j(t); ji \in \mathcal{I}, j \in \mathcal{S})$ *representing the fluid-based dynamical system approximation, such that for any* $t \geq 0$,

$$\lim_{N\to\infty} N^{\frac{1}{2}-\epsilon}\, \mathbb{E}\left[\left|\frac{X_{ji}^{(N)}(t)}{N} - x_{ji}(t)\right|\right] = 0, \qquad \forall ji \in \mathcal{I} \tag{1.7a}$$

$$\lim_{N\to\infty} N^{\frac{1}{2}-\epsilon}\, \mathbb{E}\left[\left|\frac{Z_{ji}^{(N)}(t)}{N} - z_{ji}(t)\right|\right] = 0, \qquad \forall ji \in \mathcal{I} \tag{1.7b}$$

$$\lim_{N\to\infty} N^{\frac{1}{2}-\epsilon}\, \mathbb{E}\left[\left|\frac{X_{j}^{(N)}(t)}{N} - x_{j}(t)\right|\right] = 0, \qquad \forall j \in \mathcal{S} \tag{1.7c}$$

Moreover, there exists constant $\alpha_M > 0$*, such that for any time* $t > 0$,

$$\sum_{j\in\mathcal{S}} |x_j^+(t) - f_j| + \sum_{ji\in\mathcal{I}} |x_{ji}(t) - f_{ji}| + \sum_{ji\in\mathcal{I}} |z_{ji}(t) - e_{ji}| = O(e^{-\alpha_M t})$$

$$\sum_{j\in\mathcal{S}} |x_j^-(t) - \beta_j| = O(t^{-1}) \tag{1.8}$$

Remark 2. *To the best of our knowledge, Theorem 1 provides the first fluid limit result for a network of double-ended queues. Limiting result for single-node doubled-ended queue has been reported in Liu et al. [2015], Liu [2019].*

We will refer to the process $x(t) = (x_{ji}(t), z_{ji}(t), x_j(t); ji \in \mathcal{I}, j \in \mathcal{S})$ as the transient fluid solution of the original network, while $\mathbf{f} = (f_j, f_{ji}, e_{ji}, \lambda_{ji}, \beta_j; j \in \mathcal{S}, ji \in \mathcal{I})$ is the steady state limit of the fluid process. Theorem 1 indicates that the scaled queueing length process of the original ride-hailing network at any time t converges to the transient fluid solution at the rate of nearly

$O(N^{-1/2})$; the transient fluid process of vehicle queueing length processes at each node and each traveling link, as well as the the rate of lost passengers, converge their steady state exponentially fast, while transient fluid process of the number of passengers waiting for vehicles converges to steady state at rate $O(1/T)$. Putting all these results together, Theorem 1 confirms that under the set of dynamic policies considered in this paper, the scaled queueing length process converges to the fluid steady state solution $\mathbf{f} = (f_j, f_{ji}, e_{ji}, \beta_j)$ at their respective rate.

Remark 3. *The numerical results in Braverman et al. [2019] show that their static policies converge in time to their fluid limit but very slowly. For example, the numerical results of a case study in Braverman et al. [2019] demonstrates that while the peak hour is roughly* 2 *hours, it might take up to* 10 *hours for a static policy to converge to its fluid limit. They state that*

"it would be very interesting to be able to say something rigorous about time-varying policies. Along this line, studying the transient control problem would also be of interest (given the long time it takes the fluid to converge to equilibrium from certain initial conditions). Even studying the fluid transient control problem is non-trivial, because the fluid model is a non-linear dynamical system."

This paper proposes the first dynamic policy that reaches an optimality gap of $O(1/N^{1/2-\epsilon})$ *exponentially fast in time. The numerical experiments in Section 1.4 demonstrate that it also converges very fast numerically.*

One important question is, how well does our fluid-based policy perform compared to other dynamic policies? And how does the performance of the policy change/improve when the market size and time vary? These questions are answered in the following theorem.

Theorem 2. ***(Optimality gap)*** *Consider the dynamic policy derived from the optimization solution of fluid problem (1.3) and (1.4),* $(\boldsymbol{Q}^*, \boldsymbol{\lambda}^*, \boldsymbol{\gamma}^*)$. *Let* U^* *be the optimal objective value to the fluid optimization problem (1.3) and (1.4).*

a) For any dynamic policy that imposes a single recurrent class on the system states $\mathcal{X}$*, its long run performance is upper bounded by* U^**, i.e.,*

$$\limsup_{T\to\infty} \frac{1}{T} U(T, N) \le U^*.$$

Furthermore, the long run performance of our proposed policy, which we denote by $U_{\boldsymbol{Q}^*,\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*}(T, N)$*, achieve* U^**, i.e.,*

$$\limsup_{T,N\to\infty} \frac{1}{T} U_{\boldsymbol{Q}^*,\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*}(T, N) = U^*.$$

b) There exists a constant $\alpha_S > 0$, such that for any positive constant $\delta > 0$, it holds that

$$\frac{\partial U_{\boldsymbol{Q}^*,\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*}(T,N)}{\partial T} \geq \Big[1 - O\Big(N^{-1/2+\delta} + e^{-\alpha_S T}\Big)\Big] U^*. \tag{1.9}$$

Therefore, we have

$$\frac{U_{\boldsymbol{Q}^*,\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*}(T,N)}{T} \geq \Big[1 - O\Big(N^{-1/2+\delta} + \frac{1}{T}\Big)\Big] U^*.$$

Part (a) of Theorem 2 states that, in terms of long run utility rate, no dynamic policy can outperform that of the optimal fluid problem (1.2), while our proposed policy achieves just that. For part (b), since the left hand side represents the utility rate at time T for a system with market size N, the result claims that the utility rate converges to an $O(N^{-1/2+\delta})$ neighborhood of the optimal long run utility rate exponentially fast in time for arbitrarily given small $\delta > 0$. Theorem 2 presents the first performance guarantee result for a finite-time and finite-supply system. Banerjee et al. [2021] considers the infinite time horizon problem and shows that the best asymptotic optimality gap of infinite horizon long-run average for any dynamic policy is $\Omega(N^{-1/2})$. Our results extend the literature on performance bound to finite time by establishing that our dynamic policy achieves the (any small neighborhood of) long-run average rate exponentially fast over time.

Recall from Lemma 1 that the limiting passenger queue associated with an optimal fluid-based policy at each station is zero. Thus, Theorem 1 suggests that, under an optimal policy, the probability of having a passenger queue at each node converges to 0 as the market size grows. This result motivates us to consider another assessment measure that how fast the passenger loss converges to zero as the market size grows. This criterion has been studied in Banerjee et al. [2018a] in which the authors search for policies that minimize the asymptotic expected passenger loss. Since the passengers' patience times are exponentially distributed, the rate of passengers reneging from the queue at station j is $\theta_j X_j^{(N)-}(t)$ at each time $t \geq 0$. Therefore, the expected passenger loss in a time interval $[t_1, t_2]$ can be computed as

$$L(t_1, t_2) = \mathbb{E}\Big[\sum_{j \in \mathcal{S}} \int_{t_1}^{t_2} \theta_j X_j^{(N)-}(t) dt\Big].$$

The passenger loss under our policy turns out to be very small, as the next result illustrates.

Theorem 3. ***(Passenger loss rate)*** *Consider the optimal solution $(\boldsymbol{\lambda}, \boldsymbol{f}, \boldsymbol{e}, \boldsymbol{\beta})$ to fluid optimization problem (1.3) and (1.4) and its associated dynamic policy $(\boldsymbol{Q}, \boldsymbol{\gamma}, \boldsymbol{c}(\boldsymbol{\lambda}))$. The expected passenger loss decreases exponentially with the market size N. That is, there exist constants α_L and t_0, such that for any fixed $T > t_0$,*

$$L(t_0, T) = O(e^{-\alpha_L N}). \tag{1.10}$$

Theorem 3 shows that the dynamic policy derived from fluid model (1.2) not only maximizes the objective function (revenue, social welfare, throughput), it also exhibits the important property that, under the policy the passenger loss goes to zero exponentially fast when market size increases. The most relevant study of this property is Banerjee et al. [2018b], where the authors minimize the asymptotic expected passenger loss by proposing a dynamic ride-pooling mechanism. Assuming a demand pattern condition called complete resource pooling (CRP) is satified and immediate relocation between neighboring nodes, they prove that the asymptotic expected passenger loss per unit of time decreases to zero exponentially fast when the market size grows. Furthermore, they explain the importance of relaxing the CRP condition as follows:

"We leave the cases where demand is time-varying or violates CRP for future research. Our numerical study in Section 6.4 regarding transient performance may be seen as a first step towards the time-varying case, and for a setting where CRP is violated, we conjecture that a back-pressure type policy that generalizes SMW could work well."

Our result in Theorem 3 addresses the question left open in Banerjee et al. [2018b] by extending that to general demand and finite time. In fact, we can show that the exponential passenger loss can be achieved by the dynamic policy derived from any feasible solution of the fluid optimization problem (1.3) and (1.4), as long as the condition $f_j > 0$ is satisfied for all j.

Theorems 2 and 3 present the theoretical performance guarantee of our proposed dynamic policy for the network when the arrival rates are $N\boldsymbol{\lambda}$ and the number of vehicles in system is N. Now, we address the following question: Assuming the arrival rates remain $N\boldsymbol{\lambda}$, can the same performance guarantee be achieved using few number of vehicles? To answer this question, we consider the system with passenger arrival rates $N\boldsymbol{\lambda}$ but the number of vehicles is changed to N'. With a slight abuse of notation, we let the objective function for the system be denoted by $U(T, N, N')$:

$$U(T,N,N') = \frac{1}{N}\mathbb{E}\left[\sum_{ji\in\mathcal{I}}\int_0^T I_{ji}(\lambda_{ji}(t))dK_{ji}(t,T) - \sum_{ji\in\mathcal{I}}\int_0^T c^V_{ji}dJ_{ji}(t) - \sum_{j\in\mathcal{S}}\int_0^T c^W_j X^-_j(t)dt\right].$$

We want to find the minimum number of vehicles N' to achieve the optimal objective NU^*.

The results in Theorems 2 and 3 suggest that, to achieve objective value NU^*, the number of vehicles we require is at least $N\sum_{ji\in\mathcal{I}}(f_{ji}+e_{ji})$ which is less than or equal to N. We are interested in the case that this minimum capacity requirement is strictly less than N or $\sum_{ji\in\mathcal{I}}(f_{ji}+e_{ji}) < 1$. We show that the additional number of vehicles required to reach the same level of service is $|\mathcal{S}|N^{1/2+\varepsilon}$, where ε is an arbitrarily small positive number. That is,

$$N' = N\sum_{ji\in\mathcal{I}}(f_{ji}+e_{ji}) + |\mathcal{S}|N^{1/2+\varepsilon}. \tag{1.11}$$

Note that N' is much smaller than N. For this part only, we refer to N as market size.

We present the result formally in the following Theorem.

Theorem 4. ***(Capacity planning)*** *Consider a ride-hailing network with the passenger arrival rate $N\boldsymbol{\lambda}$ and number of vehicles (1.11). Using the optimal solution of the fluid optimization problem (1.3) and (1.4), $(f_j, \beta_j, f_{ji}, e_{ji}, \lambda_{ji}, j \in \mathcal{S}, ji \in \mathcal{I})$, we define a joint pricing-empty relocation policy $(\boldsymbol{Q}, \boldsymbol{\gamma}, \boldsymbol{c}(\boldsymbol{\lambda}))$ by (1.5) and (1.6) but (1.6) is replaced by*

$$\gamma_i = N^{1/2-\varepsilon} \sum_{k,\, ik \in \mathcal{I}} \mu_{ik} e_{ik}, \qquad i \in \mathcal{S}.$$

Then, there exists constant $\alpha_S > 0$, such that for arbitrary positive $\delta > 0$ with $\delta < \epsilon$, we have

$$\frac{\partial U_{\boldsymbol{Q},\boldsymbol{\lambda},\boldsymbol{\gamma}}(T, N, N')}{\partial T} \geq \left[1 - O\Big(N^{-1/2+\delta} + e^{-\alpha_S T}\Big)\right] U^*,$$

and

$$\limsup_{T,N\to\infty} \frac{1}{T} U_{\boldsymbol{Q},\boldsymbol{\lambda},\boldsymbol{\gamma}}(T, N, N') = U^*.$$

Furthermore, under the policy $(\boldsymbol{Q}, \boldsymbol{\gamma}, \boldsymbol{c}(\boldsymbol{\lambda}))$, there exist constants $t_0, \alpha_V > 0$ such that for any time $T \geq t_0$ the probability of having passenger waiting at time T is

$$\mathbb{P}\left\{X_j(T) < 0 \, \textit{for some } j \in \mathcal{S}\right\} = O(|\mathcal{S}| \; e^{-\alpha_V N^{\varepsilon}}).$$

In essence, in Theorem 4 we modify the policy by reducing the average empty vehicle dwelling time at each station i, $1/\gamma_i$, as the market size increases. As a result, we achieve the same long run objective value using fewer number of vehicles while simultaneously having an exponentially small passenger waiting probability in terms of market size N.

This result differs from the result of Besbes et al. [2021a], where the authors use a stylized model with OD randomly drawn from a uniform distribution in a two-dimensional square to investigate the minimum vehicle supply for achieving a balance between vehicle utilization and passenger waiting time. Assuming no passenger abandonment, the authors show that the minimum excess supply required for a near-optimal policy to have passenger waiting probability converging to zero as market size N increases is of the order $\Omega(N^{2/3})$. They further raise the question on the impact on the capacity sizing rule when customers are impatient and might abandon the system if not served after some time. Theorem 4 addresses their question by showing that our dynamic policy achieves an exponentially small passenger loss probability with only $\Omega(N^{1/2+\varepsilon})$ excess supply (where ε can be arbitrarily small), which is more consistent with the staffing literature on QD regime (see e.g., Halfin and Whitt [1981]).

1.4 Numerical Experiments

In this section, we evaluate the performance of our joint pricing-empty relocation policy in optimizing the platform's objective. To ground the study, we use the data from the Di-Tech challenge held by DiDi Research Institute in 2016. The original data set containing individual trip order information from January 1, 2016, to January 21, 2016, in a city in China is no longer publicly available. Therefore, we utilize the data for our numerical experiments for the 5 PM-6 PM evening rush hour in the nine-region network extracted from the original data set by Braverman et al. [2019].

Benchmark. Our benchmark for performance comparison is the optimal static policy developed in Banerjee et al. [2021] because i) their results guarantee the system performance to be asymptotically optimal for long-run average criterion, and ii) for a given pricing scheme, Braverman et al. [2019] has shown that the static relocation policy outperforms a set of intuitive, dynamic routing heuristics. Hence, we compare our results with those of Banerjee et al. [2021] to showcase our policy's performance and discuss managerial insights.

Simulation setup. We simulate the system introduced in Section 1.2 to evaluate the performance measures based on market size and length of time horizon. In our basic setting, we consider a two-hour time horizon and $N = 2000$ vehicles. Following Braverman et al. [2019], we assume vehicle travel time follows an exponential distribution. Using the data presented in Ridesharing-Driver [2018] website, we consider 25% of the platform's announced price for the estimation of the empty relocation cost. Following Bimpikis et al. [2019], we assume passengers' willingness to pay for each pair of nodes follows an exponential distribution, and we adopt the same method as Bimpikis et al. [2019] to estimate its parameter. Also, following the study of Goldszmidt et al. [2020] on the value of time for riders on the Lyft platform, we consider 120 (CNY/h) as the passenger's waiting cost. Throughout this section, the system is initialized by assuming all cars are idle and distributing them across nodes uniformly.

1.4.1 Throughput, social welfare, and revenue

In addition to comparing the results of optimal static policy from Banerjee et al. [2021], we also present the theoretical upper bound for the performance measures in Theorem 2.

Performance change with market size. First, to ensure the irreducibility of the empty relocation matrix $\boldsymbol{Q}$, we enforce a constraint to the fluid optimization problem (1.2). The constraint requires the limiting number of empty vehicles along a closed cycle in the network is at least 1% of all vehicles. Figures 1.2-1.4 demonstrate the performance (revenue, social welfare, and throughput) of our optimized dynamic policy compared to the benchmarks. In these figures, we consider a fixed time horizon of 2 hours, and vary the market size N from 200 to 2000.

It is seen from the numerical results that, the performance measures quickly converge to the optimal value in Theorem 2 under our optimized policy in the two hour time horizon when the market size N increases. More importantly, it is clear that our (dynamic) policy significantly outperforms the benchmark solution in all the three performance measures: throughout, revenue and social welfare.

We also make an interesting observation from our numerical results: It is shown in Banerjee et al. [2021] that, under optimized static policies over infinite time horizon, the performance measures converge to the fluid-based optimal value when the market size increases. However, the speed of convergence is unclear. Our numerical results in Figures 1.2-1.4 indicates that performance of the static policy is distant from the theoretical upper bound in the 2 hour time horizon (and also much lower that that of our dynamic policy), suggesting that the speed of convergence may be rather slow. This is in line with the observations of Braverman et al. [2019] that static policies require a time scale on the order of 10 hours to reach their optimal performance. To further investigate the convergence time of our policy in comparison to the static policy, we analyze performance measure values over time in the following section.

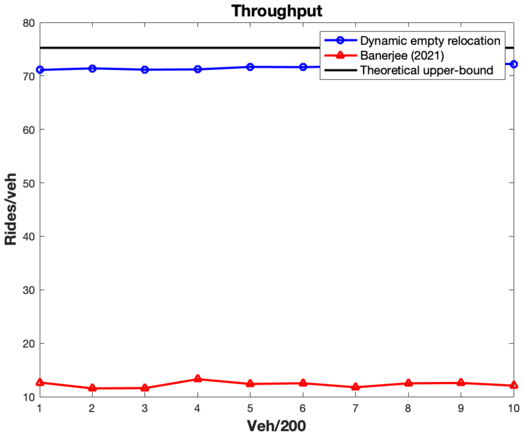


Figure 1.2: Throughput as market size grows

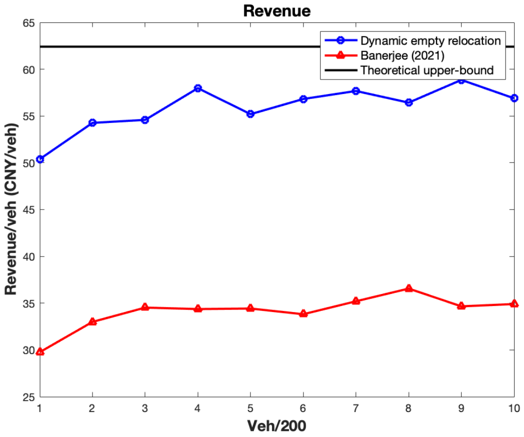


Figure 1.3: Revenue as market size grows

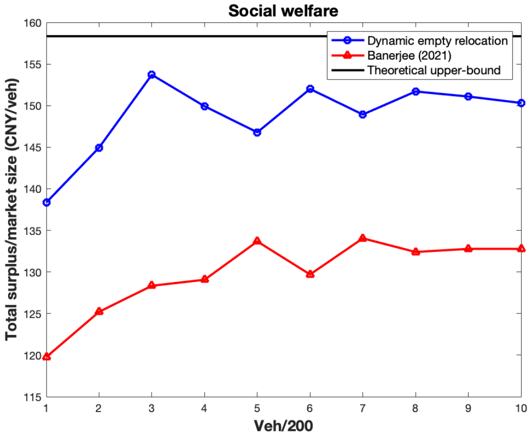


Figure 1.4: Social welfare as market size grows

Performance change with time. Next, we fix the market size N and test the performance of our dynamic policy when time horizon T varies. Again, we compare the three performance measures: Throughput, Revenue, and Social Welfare. Figures 1.5-1.7 present the results of our dynamic policy and benchmark static policies of Banerjee et al. [2021] for a fixed market size $N = 2000$. The time horizon changes from 1 hour to 10 hours. Similar to Braverman et al. [2019], we observe that convergence of optimized static policies to equilibrium occurs on timescales of around 10 hours. In contrast, our optimized dynamic policy reaches a 10% optimality gap within 1 hour for the revenue and social welfare and 3 hours for throughput.

We offer an explanation why the static policy converges slowly, but dynamic policy converges very fast. This is because in static policies the outflow of empty vehicles from a station is proportionate to the inflow of full vehicles. However, in our dynamic policy it is proportionate to the number of idle vehicles in that node. Therefore, when more empty vehicles gather at one region, they are more quickly relocated to other regions in need of them.

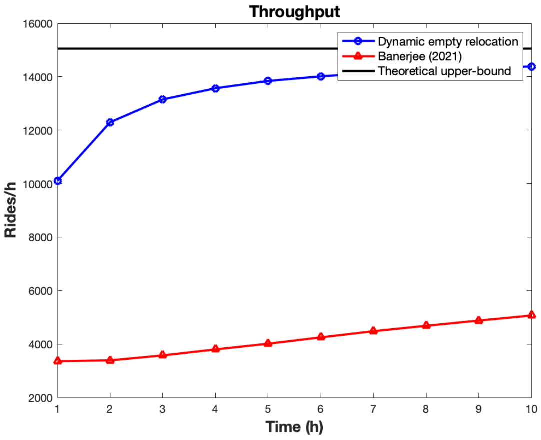


Figure 1.5: Throughput over time

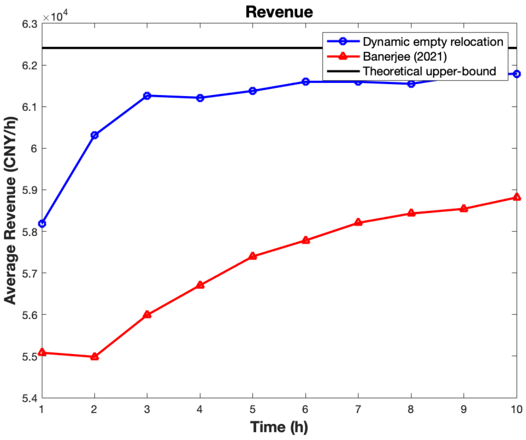


Figure 1.6: Revenue over time

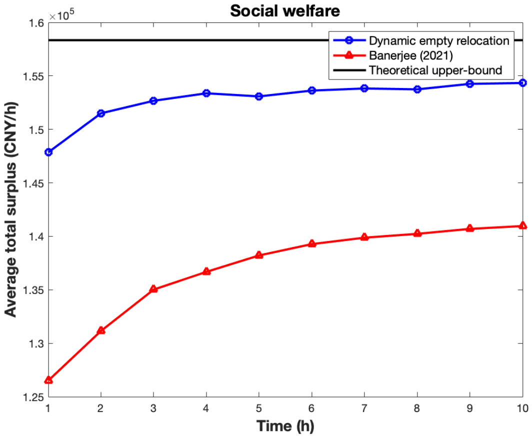


Figure 1.7: Social welfare over time

1.4.2 Passenger loss rate

We now analyze the passenger loss rate when either market size or time horizon changes. In Figure 1.8 the time horizon is fixed at 2 hours, and the market size again varies from 200 to 2000. The results demonstrate that the passenger loss indeed drops exponentially when market size increases, and in this example it reduces to nearly zero when the market sizes reaches 800. This confirms our theoretical results in Theorem 3. It is also observed from Figure 1.8 that the loss rate under optimal static policy decreases rather slowly, and it confirms the result of Banerjee et al. [2018a] that the best reduction in passenger loss for a static policy is polynomial in market size.

Figure 1.9 demonstrates that passenger loss decreases over time for both static and dynamic policies. However, we observe that passenger loss is significantly lower in dynamic policies than in static policies.

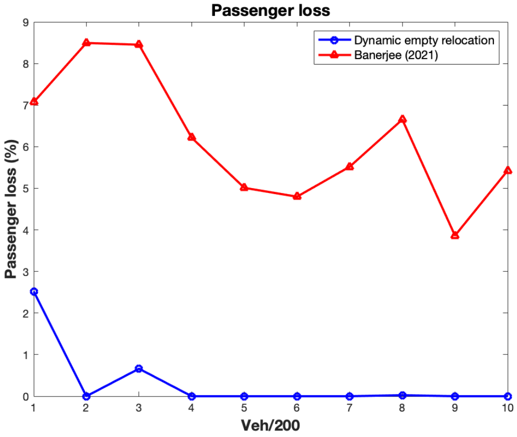


Figure 1.8: Passenger loss as market size grows

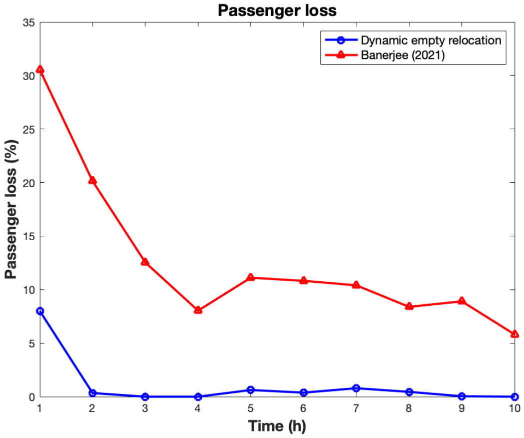


Figure 1.9: Passenger loss over time

Sensitivity to passenger patience parameter. Next, we analyze the sensitivity of passenger loss to passenger waiting patience time. In Figure 1.10, we compute the percentage of loss for three patience parameters, $\theta = 3, 6$, and 12, so the passengers are becoming increasingly impatient. The time horizon is 2 hours, and we present the results when the market size increases from 200 to 2000.

As seen from Figure 1.10, the numerical results are consistent with our intuition: Passenger loss generally decreases with an increase in their waiting patience. Furthermore, the loss decreases linearly in the logarithmic scale of market size. This confirms our theoretical results in Theorem 3.

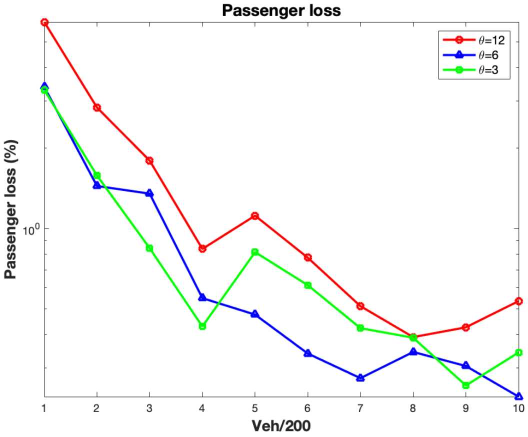


Figure 1.10: Sensitivity analysis on the patience threshold

Capacity planning. We now numerically illustrate the results of Theorem 4. After enforcing the irreducibility constraint by limiting number of empty vehicles along a closed cycle in the network to be at least 4% of all vehicles, we solve the optimization problem in (1.3) and (1.4) to obtain the parameters for the policy in Theorem 4. We consider three scenarios indicated by $\epsilon \in \{0.05, 0.1, 0.2\}$ such that in each scenario, we set the number of vehicles to

$$N' = N \sum_{ji \in \mathcal{I}} (f_{ji} + e_{ji}) + \frac{|\mathcal{S}|}{3} N^{1/2+\varepsilon}.$$

In Figures 1.11 and 1.12, the time horizon is fixed at 3 hours and the market size varies from 500 to 5000. The curves in Figures 1.11 and 1.12 present the probability of waiting and the revenue in (in Chinese yuan) per market size for the three scenarios. As shown, the numerical and theoretical results are consistent. The probability of waiting for passengers is in the order of $O(e^{-\alpha_V N^{\varepsilon}})$. Specifically, for $\epsilon = 0.05$ the probability of waiting reduces from 12% to 7%. For $\epsilon = 0.1$, it reduces from 10.5% for $N = 500$ to 3.8% for $N = 5000$; while for $\epsilon = 0.2$, it reduces from 2.2% for $N = 500$ to 0.02% for $N = 5000$. By increasing ϵ, the probability of waiting for passengers converges to zero at the corresponding exponential rate. Additionally, Figure 1.12 shows that the revenues per market size for all three scenarios are within 4% gap of the theoretical optimal value when the market range varies from 500 to 5000 vehicles.

Figure 1.13 displays the number of vehicles N' as N changes from 500 to 5000. This figure confirms that the number of vehicles in the system is substantially smaller than N in the three scenarios. For example, for $\epsilon = 0.1$ when $N = 1000$, we have $N' = 799$, and when $N = 5000$, we have $N' = 3542$. The result shows that we can achieve almost the same objective function and passenger loss probability with much few number of vehicles.

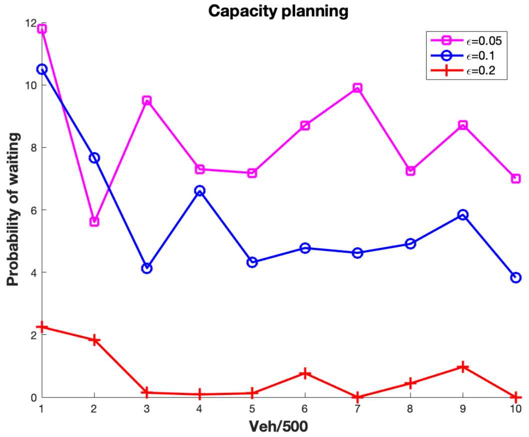


Figure 1.11: Probability of waiting

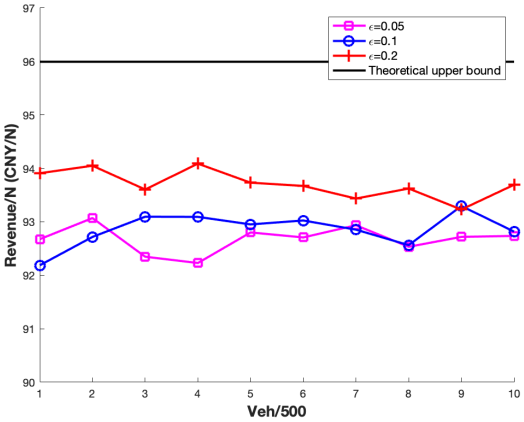


Figure 1.12: Revenue over time

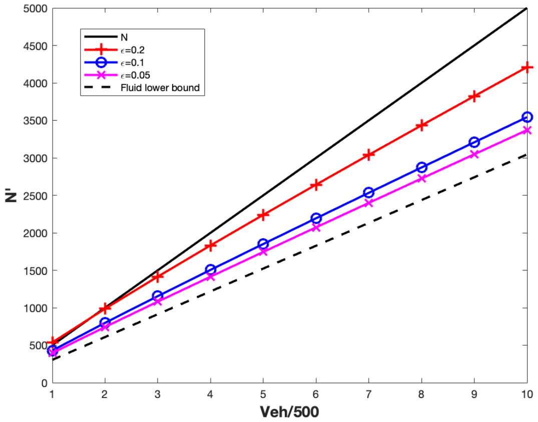


Figure 1.13: Capacity

1.5 Outline of Proofs of Main Results

In this section, we outline the main steps in proving Theorems 1 to 3. The proof for Theorem 4 follows the same procedure as those for Theorem 2 and Theorem 3, so it it not discussed here. The details of all the proofs are provided in the Appendix. We first present the system dynamics of the ride-hailing problem.

1.5.1 System dynamics

To prepare for the proof of the main results, we need to write down the system dynamics. For illustration, we assume that the travel time for vehicles moving from node j to node i are i.i.d. with exponential distribution of rate μ_{ji}. For general traveling time distributions, the typical approach is to replace the links in the original network with a sub-network of infinite and double-ended queues using that mixtures of Erlang distributions that are dense among all continuous distributions. We discuss this transformation in detail in Section A.3.

Passengers originate at node j with destination node i according to a Poisson process with rate denoted by λ_{ji}. To model the passenger arrival process, we introduce unit rate Poisson Processes $A_i = \left\{A_i(t), t \geq 0\right\}, i \in \mathcal{S}$. That means we denote the number of passenger requests with origin j and destination i until the time t by $A_i(\lambda_{ji}t)$. To compute passengers' reneging after their patience runs out, we define unit rate Poison Processes $G_i = \left\{G_i(t), t \geq 0\right\}, i \in \mathcal{S}$. Given that, calculation of the number of passengers who renege from node j by the time t is

$$G_j\left(\theta_j \int_0^t X_j^{(N)-}(s)ds\right).$$

Among the first n full vehicles dispatched from the station j, we denote the number of those with destination i by $\phi_{ji}(n), ji \in \mathcal{I}$. Note that the probability that a full vehicle departing node j for destination i is equivalent to the probability that the destination of a passenger arriving at j is i, thus

$$\mathbb{P}\left\{\phi_{ji}(n) - \phi_{ji}(n-1) = 1\right\} = \frac{\lambda_{ji}}{\sum_k \lambda_{jk}} = p_{ji}, \quad \mathbb{P}\left\{\phi_{ji}(n) - \phi_{ji}(n-1) = 0\right\} = 1 - \frac{\lambda_{ji}}{\sum_k \lambda_{jk}}.$$

Similarly, to compute the number of empty relocation from each node, we define the unit rate Poison processes $H_i = \left\{H_i(t), t \geq 0\right\}, i \in \mathcal{S}$. So, the number of empty vehicles from i to j by the time t is

$$H_j\left(\gamma_j \int_0^t X_j^{(N)+}(s)ds\right).$$

Let $\sigma_{ji}(n)$ denote the number of the first n empty vehicles that left node j to destination node $i, ji \in \mathcal{I}$, then

$$\mathbb{P}\left\{\sigma_{ji}(n) - \sigma_{ji}(n-1) = 1\right\} = q_{ji}, \qquad \mathbb{P}\left\{\sigma_{ji}(n) - \sigma_{ji}(n-1) = 0\right\} = 1 - q_{ji}.$$

Finally, let $F_{ji}(t)$ and $E_{ji}(t)$ denote unit rate Poisson processes, then the number of full and empty vehicles that enter destination node j from node i can be expressed as

$$F_{ji}\left(\mu_{ji} \int_0^t X_{ji}^{(N)}(s)ds\right), \qquad E_{ji}\left(\mu_{ji} \int_0^t Z_{ji}^{(N)}(s)ds\right).$$

Recall that in the scaled system N, the arrival rates are $N\lambda_{ji}$. We have the following system dynamics for the double-ended queue i of the network:

$$\begin{aligned} X_i^{(N)}(t) &= X_i^{(N)}(0) - A_i\left(N\lambda_i t\right) + \sum_{ji\in\mathcal{I}} F_{ji}\left(\mu_{ji} \int_0^t X_{ji}^{(N)}(s)ds\right) \\ &\quad + \sum_{ji\in\mathcal{I}} E_{ji}\left(\mu_{ji} \int_0^t Z_{ji}^{(N)}(s)ds\right) + G_i\left(\theta_i \int_0^t X_i^{(N)-}(s)ds\right) \\ &\quad - H_i\left(\gamma_i \int_0^t X_i^{(N)+}(s)ds\right), \qquad i \in \mathcal{S}, \end{aligned} \tag{1.12}$$

where $X_i^{(N)}(0)$ is the state of the system at station i at the time 0, $A_i(N\lambda_i t)$ is the passenger arrivals to station i by the time t, $\sum_{ji\in\mathcal{I}} F_{ji}\big(\mu_{ji} \int_0^t X_{ji}^{(N)}(s)ds\big)$ represents the number of completed trips of full vehicles travelling from node j to i by the time t, and $\sum_{ji\in\mathcal{I}} E_{ji}\big(\mu_{ji} \int_0^t Z_{ji}^{(N)}(s)ds\big)$ denotes the number of empty vehicles arriving to node i from node j by the time t.

Next, we present the system dynamics for the infinite server stations to model the full vehicles traveling on each link ji, and it is given by

$$\begin{aligned} X_{ji}^{(N)}(t) &= X_{ji}^{(N)}(0) - F_{ji}\bigg(\mu_{ji}\int_0^t X_{ji}^{(N)}(s)ds\bigg) \\ &\quad +\phi_{ji}\Bigg(\sum_{k,kj\in\mathcal{I}} F_{kj}\bigg(\mu_{kj}\int_0^t X_{kj}^{(N)}(s)ds\bigg) + \sum_{k,kj\in\mathcal{I}} E_{kj}\bigg(\mu_{kj}\int_0^t Z_{kj}^{(N)}(s)ds\bigg) \\ &\quad +X_j^{(N)+}(0) - X_j^{(N)+}(t) - H_j\bigg(\gamma_j\int_0^t X_j^{(N)+}(s)ds\bigg)\Bigg), \qquad ji\in\mathcal{I}, \quad (1.13) \end{aligned}$$

where $X_{ji}^{(N)}(0)$ is the initial number of full vehicles travelling from node j to i at time 0. Note that the output of term ϕ_{ji} gives the total number of full vehicles leaving their trips from j toward i up to time t, and the quantity in the open brackets of $\phi_{ji}(\cdot)$ represents the total number of departures from node j by time t. This is because, following the conservation law for vehicles, the input to ϕ_{ji} calculates the number of full vehicles departing j by the time t. Specifically, $\sum_{kj\in\mathcal{I}} F_{kj}(\mu_{kj}\int_0^t X_{kj}^{(N)}(s)ds)$ indicates the number of completed trips to node j, $\sum_{kj\in\mathcal{I}} E_{kj}(\mu_{kj}\int_0^t Z_{kj}^{(N)}(s)ds)$ denotes the number of completed empty trips to node j, $X_j^{(N)+}(0)$ represents the initial number of vehicles in node j, $X_j^{(N)+}(t)$ characterizes the number of current vehicles in node j, and $H_j(\gamma_j\int_0^t X_j^{(N)+}(s)ds)$ denotes the total number of empty vehicles reneged from node j by time t.

Lastly, we note the defining equations for the system dynamics in the infinite server stations to model the travel of empty vehicles on link ji, as follows

$$Z_{ji}^{(N)}(t) = Z_{ji}^{(N)}(0) + \sigma_{ji}\Bigg(H_j\bigg(\gamma_j\int_0^t X_j^{(N)+}(s)ds\bigg)\Bigg) - E_{ji}\bigg(\mu_{ji}\int_0^t Z_{ji}^{(N)}(s)ds\bigg), \quad ji\in\mathcal{I}. \tag{1.14}$$

Here, $Z_{ji}^{(N)}(0)$ is the initial number of full vehicles travelling from node j to i at the time 0, the second term on the right hand side is the total number of empty vehicles relocated from j to i by time t, while the third term is the total number of empty vehicles from j to i that have arrived node i by time t.

Fluid system dynamics. We now rewrite the system dynamics by separating the stochastic and deterministic terms so that each stochastic term has a mean of 0. For each stochastic term, we subtract its mean from it so all stochastic will have mean 0. For convenience, we use the notation "$\tilde{}$" to indicate the centered stochastic components:

$$\begin{aligned} &\tilde{A}_i(t) = A_i(t) - t; \qquad \tilde{E}_{ij}(t) = E_{ij}(t) - t; \qquad \tilde{F}_{ij}(t) = F_{ij}(t) - t; \\ &\tilde{G}_i(t) = G_i(t) - t; \qquad \tilde{H}_i(t) = H_i(t) - t; \\ &\tilde{\phi}_{ij}(n) = \phi_{ij}(n) - p_{ji}n; \qquad \tilde{\sigma}_{ij}(n) = \sigma_{ij}(n) - q_{ij}n. \end{aligned}$$

Then, for a double-ended queue at station i, the centered stochastic term of system dynamics (1.12) can be expressed as

$$\begin{aligned} \tilde{X}_i^{(N)}(t) &= -\tilde{A}_i\Big(N\lambda_i t\Big) + \sum_{ji\in\mathcal{I}} \tilde{F}_{ji}\Big(\mu_{ji}\int_0^t X_{ji}^{(N)}(s)ds\Big) + \sum_{ji\in\mathcal{I}} \tilde{E}_{ji}\Big(\mu_{ji}\int_0^t Z_{ji}^{(N)}(s)ds\Big) \\ &\quad +\tilde{G}_i\Big(\theta_i\int_0^t X_i^{(N)-}(s)ds\Big) - \tilde{H}_i\Big(\gamma_i\int_0^t X_i^{(N)+}(s)ds\Big), \qquad i\in\mathcal{S}. \end{aligned} \tag{1.15}$$

We similarly obtain the stochastic terms of $\tilde{X}_{ji}^{(N)}(t), \forall ji \in \mathcal{I}$ and $\tilde{Z}_{kl}^{(N)}(t), \forall kl \in \mathcal{I}$. Refer to Appendix A.4 for detailed expressions.

To proceed, we introduce the following deterministic dynamic system for give process $\tilde{x}(t)$:

$$\begin{aligned} x_i(t) &= x_i(0) + \tilde{x}_i(t) - \lambda_i t + \sum_{ji\in\mathcal{I}}\mu_{ji}\int_0^t x_{ji}(s)ds + \sum_{ji\in\mathcal{I}}\mu_{ji}\int_0^t z_{ji}(s)ds \\ &\quad +\theta_i\int_0^t x_i^-(s)ds - \gamma_i\int_0^t x_i^+(s)ds, \qquad i\in\mathcal{S} \end{aligned} \tag{1.16}$$

$$\begin{aligned} x_{ji}(t) &= x_{ji}(0) + \tilde{x}_{ji}(t) - \mu_{ji}\int_0^t x_{ji}(s)ds + p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}\int_0^t x_{kj}(s)ds \\ &\quad +p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}\int_0^t z_{kj}(s)ds + p_{ji}x_j^+(0) - p_{ji}x_j^+(t) - p_{ji}\gamma_j\int_0^t x_j^+(s)ds, \quad ji\in\mathcal{I} \end{aligned} \tag{1.17}$$

$$z_{ji}(t) = z_{ji}(0) + \tilde{z}_{ji}(t) - \mu_{ji}\int_0^t z_{ji}(s)ds + q_{ji}\gamma_j\int_0^t x_j^+(s)ds, \qquad ji\in\mathcal{I} \tag{1.18}$$

Note that $X^{(N)}/N$ satisfies these equations with $\tilde{x}(t) = \tilde{X}^{(N)}(t)/N$.

The following result provides the properties of solution to the system (1.16-1.18).

Lemma 2. *There exists a unique solution to the dynamical system (1.16-1.18), and the solution is Lipschitz continuous in the process $\tilde{x}(t)$.*

Outline of proof of Theorem 1.

We first prove that, under the policy $(\boldsymbol{Q}, \boldsymbol{\gamma}, \boldsymbol{c}(\boldsymbol{\lambda}))$, the scaled queue length process $X^{(N)}(T)/N$ converges to the fluid solution $x(t)$ that is defined as the solution of system (1.16-1.18) with $\tilde{x} = 0$ as $N \to \infty$. We make use of the Lipschitz continuity property of the solution to dynamical system (1.16-1.18) presented in Lemma 2 to bound the difference between the scaled queue length process and the centered system (1.16-1.18) as follows:

$$\mathbb{E}\Bigg[\bigg|\frac{X^{(N)}(t)}{N} - x(t)\bigg|\Bigg] \le C'\,\mathbb{E}\Bigg[\bigg\|\frac{\tilde{X}^{(N)}}{N}\bigg\|_t\Bigg] \tag{1.19}$$

for some constant C'. Then, we apply Donsker's theorem to bound the magnitude of $\tilde{X}^{(N)}/N$ (see details in Lemma 7 in the Appendix) to prove that, for an arbitrary small $\epsilon > 0$, we have

$$\lim_{N\to\infty} N^{1/2-\epsilon}\mathbb{E}\left[\left\|\frac{\tilde{X}^{(N)}}{N}\right\|_t\right] = 0.$$

We next prove that the solution to (1.16-1.18) with $\tilde{x} = 0$ converges to its equilibrium solution $(\boldsymbol{\lambda}, \boldsymbol{f}, \boldsymbol{e}, \boldsymbol{\beta})$ at the respective rates presented in (1.8). We first show that a Linear Complementarity Problem (LCP) that describes the equilibrium vehicle queue length x^+ under a given policy $(\boldsymbol{Q}, \boldsymbol{\gamma}, \boldsymbol{c}(\boldsymbol{\lambda}))$ has a unique solution with any given total supply Θ. Then, we construct a Lyapunov function using the maximum value of Θ for which the solution to the LCP is a lower bound of the vehicle queue length x^+. We iteratively bound the rate of improvement in the Lyapunov function using the convergence rate of a weakly diagonally dominant linear dynamical system. Finally we use the Cheeger inequality for the second largest eigenvalue of the linear dynamical system, that determines the convergence rate, to show that the Lyapunov function converges exponentially in time. The linear convergence of the passenger queue length follows subsequently by letting time go to infinity. The proof of these statements are technical and are laid out in Lemmas 8 to 16 in the Appendix. This concludes the proof for Theorem 1.

Remark 4. *Establishing the uniqueness of the solution to LCPs is typically a challenging task. Due to the unique structure of the coefficient matrix in our LCP, we are able to solve it in closed form and then prove its uniqueness (see Lemma 8 in the Appendix). To the best of our knowledge, this paper is the first to use LCP based approach to construct a Lyapunov function in establishing the stability of a dynamical system.*

Outline of proof of Theorem 2.

Part (a) in Theorem 2 is proved in three steps. First, in Lemma 18 in the Appendix, we use the Foster-Lyapunov Theorem to show that every dynamic policy that impose a single recurrent class on the system states leads to an infinite-state Continuous Time Markov Chain (CTMC) that is ergodic, hence it has a stationary distribution. Second, in Lemma 19 in the Appendix, we show that the expected scaled stationary queue length $\mathbb{E}[X^N(\infty)/N]$ satisfies the constraints in the fluid optimization in (1.3) and (1.4). Third, in Lemma 20 in the Appendix, we prove that the objective value associated with the long-run system performance measures is upper bounded by the objective value of the fluid optimization in (1.3) and (1.4).

The proof for part (b) of Theorem 2 is more involved. It is clear that it suffices to prove

$$\left|\frac{\partial U_{\boldsymbol{Q}^*,\boldsymbol{\lambda}^*,\boldsymbol{\gamma}^*}(T,N)}{\partial T} - U^*\right| \le O\Big(N^{-1/2+\delta} + e^{-\alpha_S T}\Big). \tag{1.20}$$

Since we fix the pricing decisions in our dynamic joint pricing empty relocation policy, the rate of objective function at the time t can be written as

$$\frac{\partial U_{\boldsymbol{Q},\boldsymbol{\lambda},\boldsymbol{\gamma}}(T,N)}{\partial T} = \frac{1}{N}\mathbb{E}\Big[\sum_{ji\in\mathcal{I}}\mu_{ji}X_{ji}^{(N)}(t)I_{ji}(\lambda_{ji}) - \sum_{ji\in\mathcal{I}}c_{ji}^{V}\mu_{ji}Z_{ji}^{(N)}(t) - \sum_{j\in\mathcal{S}}c_j^W X_j^{(N)-}(t)\Big].$$

Recall that U^* is the optimal objective value to the fluid optimization problem in (1.3) and (1.4), we have

$$U^* = \sum_{ji\in\mathcal{I}}\mu_{ji}f_{ji}^* I_{ij}(\lambda_{ji}) - \sum_{ji\in\mathcal{I}}c_{ji}^V\mu_{ji}e_{ji}^* - \sum_{j\in\mathcal{S}}c_j^W\beta_j^*,$$

here $\mathbf{f}^* = (f_j^*, f_{ji}^*, e_{ji}^*, \beta_j^*, \lambda_{ji}^*, j \in \mathcal{S}, ji \in \mathcal{I})$ is the optimal solution to the fluid optimization problem in (1.3) and (1.4). Subtracting we bound the left hand side using triangle inequality as follows

$$\begin{aligned}\Big|\frac{\partial U_{\boldsymbol{Q}^*,\boldsymbol{\lambda^*},\boldsymbol{\gamma^*}}(T,N)}{\partial T} - U^*\Big| \;&\leq\; \frac{1}{N}\sum_{ji\in\mathcal{I}}\mu_{ji}I_{ji}(\lambda_{ji})\mathbb{E}\Big[\Big|X_{ji}^{(N)}(T) - Nf_{ji}^*\Big|\Big]\\ &\quad+\frac{1}{N}\sum_{ji\in\mathcal{I}}c_{ji}^V\mu_{ji}\mathbb{E}\Big[\Big|Z_{ji}^{(N)}(T) - Ne_{ji}^*\Big|\Big] + \frac{1}{N}\sum_{j\in\mathcal{S}}c_j^W\mathbb{E}\Big[\Big|X_j^{(N)-}(T)\Big|\Big]\\ &\leq\; \frac{1}{N}\sum_{ji\in\mathcal{I}}\mu_{ji}I_{ji}(\lambda_{ji})\int_0^\infty \mathbb{P}\Big\{\Big|X_{ji}^{(N)}(T) - Nf_{ji}^*\Big| > \xi\Big\}d\xi\\ &\quad+\frac{1}{N}\sum_{ji\in\mathcal{I}}c_{ji}^V\mu_{ji}\int_0^\infty \mathbb{P}\Big\{\Big|Z_{ji}^{(N)}(T) - Ne_{ji}^*\Big| > \xi\Big\}d\xi\\ &\quad+\frac{1}{N}\sum_{j\in\mathcal{S}}c_j^W\int_0^\infty \mathbb{P}\Big\{\Big|X_{ji}^{(N)-}(T)\Big| > \xi\Big\}d\xi, \qquad (1.21)\end{aligned}$$

where the first inequality follows from $\beta_j^* = 0$ for all j, the second inequality follows from the continuous tail sum formula.

To complete the proof of Theorem 2, we need to evaluate the three tail probabilities in (1.21). We will only illustrate the first one. In Lemma 22 of the Appendix, we have done, by making use of Schilder's and Mogulskii's theorems, that there exist constant $\alpha_S > 0$ and $m > 0$ such that for $\xi \geq N^{1/2+\delta}$, we have

$$\mathbb{P}\Big\{|X^N(T) - Nf^*| > \xi + mNe^{-\alpha_S T}\Big\} = O(e^{-\alpha_U\xi^2/N}). \qquad (1.22)$$

Substituting, we obtain

$$\begin{aligned}
& \int_0^{\infty} \mathbb{P}\Big\{\Big|X_{ji}^{(N)}(T) - Nf_{ji}^*\Big| > \xi\Big\}d\xi \\
= & \int_0^{N^{1/2+\delta}+mNe^{-\alpha_S T}} \mathbb{P}\Big\{\Big|X_{ji}^{(N)}(T) - Nf_{ji}^*\Big| > \xi\Big\}d\xi \\
& + \int_{N^{1/2+\delta}+mNe^{-\alpha_S T}}^{\infty} \mathbb{P}\Big\{\Big|X_{ji}^{(N)}(T) - Nf_{ji}^*\Big| > \xi\Big\}d\xi \\
\leq & \; O(N^{1/2+\delta}) + O(Ne^{-\alpha_S T}) + \int_{N^{1/2+\delta}}^{\infty} \mathbb{P}\Big\{\Big|X_{ji}^{(N)}(T) - Nf_{ji}^*\Big| > \xi + mNe^{-\alpha_S T}\Big\}d\xi \\
\leq & \; O(N^{1/2+\delta}) + O(Ne^{-\alpha_S T}) + \int_0^{\infty} O(e^{-\alpha_U \xi^2/N})dx \\
= & \; O(N^{1/2+\delta}) + O(Ne^{-\alpha_S T}),
\end{aligned}$$

where the first inequality is obtained by replacing $P\{(\cdot)\}$ in the first term by 1, the second inequality follows from (1.22), and the last equality is obtained by including $O(N^{1/2})$ in $O(N^{1/2+\delta})$.

Similar results are obtained for the second and third integral in (1.21). Substituting these results in (1.21) we obtain (1.20). This completes the proof of Theorem 2.

Outline of proof of Theorem 3.

We want to show that there exist $\alpha_L > 0$ and $t_0 > 0$, such that for any $T \geq t_0$, (1.10) is satisfied. By Lemma 16 in the Appendix, we know that $x_i(t)$ converges to f_i, thus there exists a $t_0 > 0$ such that $|x_i(t) - f_i| < f_i/2$ for $t \geq t_0$, thus $x_i(t) \geq f_i/2$. We first upper bound the left hand side of (1.10) by

$$\mathbb{E}\Big[\sum_{j\in\mathcal{S}} \int_{t_0}^{T} \theta_j X_j^{(N)-}(t)dt\Big] \leq (T - t_0) \sum_{j\in\mathcal{S}} \theta_j \mathbb{E}\Big[\sup_{t_0\leq t\leq T} X_j^{(N)-}(t)\Big]. \tag{1.23}$$

We upper bound the expectation on the right hand side as follows

$$\begin{aligned}
\mathbb{E}\Big[\sup_{t_0\le t\le T} X_j^{(N)-}(t)\Big] &= \sum_{k=0}^{\infty}\mathbb{P}\Big\{\sup_{t_0\le t\le T} X_j^{(N)-}(t)\ge k\Big\}\\
&= \sum_{k=0}^{\infty}\mathbb{P}\Big\{\sup_{t_0\le t\le T}\Big\{-X_j^{(N)}(t)\Big\}\ge k\Big\}\\
&= \sum_{k=0}^{\infty}\mathbb{P}\Big\{\sup_{t_0\le t\le T}\Big\{-X_j^{(N)}(t)+N\frac{f_j}{2}\Big\}\ge k+N\frac{f_j}{2}\Big\}\\
&\le \sum_{k=0}^{\infty}\mathbb{P}\Big\{\sup_{t_0\le t\le T}\Big\{-X_j^{(N)}(t)+Nx_j(t)\Big\}\ge k+N\frac{f_j}{2}\Big\}\\
&\le \sum_{k=0}^{\infty}\mathbb{P}\Big\{\sup_{t_0\le t\le T}\Big|X_j^{(N)}(t)-Nx_j(t)\Big|\ge k+N\frac{f_j}{2}\Big\}\\
&\le \sum_{i=0}^{\infty}\sum_{k=iN}^{(i+1)N-1}\mathbb{P}\Big\{\sup_{t_0\le t\le T}\Big|X_j^{(N)}(t)-Nx_j(t)\Big|\ge k+N\frac{f_j}{2}\Big\}\\
&\le N\sum_{i=0}^{\infty}\mathbb{P}\Big\{\sup_{t_0\le t\le T}\Big|X_j^{(N)}(t)-Nx_j(t)\Big|\ge iN+N\frac{f_j}{2}\Big\}\\
&\le N\sum_{k=0}^{\infty}\mathbb{P}\left\{\Big\|\frac{X_j^N}{N}-x_j\Big\|_T\ge k+\frac{f_j}{2}\right\}\\
&\le N\sum_{k=0}^{\infty}\mathbb{P}\left\{\Big\|\frac{X^N}{N}-x\Big\|_T\ge k+\frac{f_j}{2}\right\}, \qquad (1.24)
\end{aligned}$$

where the first equality follows the tail sum formula for the non-negative integer valued random variable $\sup_{t_0\le t\le T} X^{(N)-}$, the second equality follows from the fact that the summand k is positive, the first inequality follows from the fact that $x_j > f_j/2$, and the fifth inequality follows from the definition of infinity norm.

By Lipschitz continuity of $x_j(t)$ (see Lemma 2), we have

$$\Big\|\frac{X^N}{N}-x\Big\|_T\le C'\Big\|\frac{\tilde{X}^N}{N}\Big\|_T,$$

Where C' is the Lipschitz constant introduced in Lemma 2. Thus

$$\mathbb{P}\left\{\Big\|\frac{X^N}{N}-x\Big\|_T\ge k+\frac{f_j}{2}\right\}\le\mathbb{P}\left\{\Big\|\frac{\tilde{X}^N}{N}\Big\|_T>\frac{k+f_j/2}{C'}\right\}.$$

In Lemma 25 in the Appendix, we will show that for a positive constant $\alpha>0$ and any positive

$\epsilon > 0$

$$\mathbb{P}\Big\{\Big\|\frac{\tilde{X}^N}{N}\Big\|_T > \epsilon\Big\} = O\big(e^{-2\alpha\epsilon^2 N}\big). \tag{1.25}$$

Setting $\epsilon = (k + f_j/2)/C'$ and $\alpha_L = \alpha(f_j/2C')^2$ in (1.25) and substituting the result in (1.24), we obtain

$$\mathbb{E}\Big[\sup_{t_0 \le t \le T} X_j^{(N)-}(t)\Big] \le N \sum_{k=0}^{\infty} O\Big(e^{-2\alpha\left(\frac{k+f_j/2}{C'}\right)^2 N}\Big) = O(e^{-\alpha_L N}), \tag{1.26}$$

where the last term follows from that $N = e^{\log N} \le e^{k+\alpha_L N}$ for some constant k. Finally, substituting (1.26) into (1.23) concludes the proof of Theorem 3.

CHAPTER 2

A Unifying Graph-Coloring Approach for Intersection Control in a Connected and Automated Vehicle Environment

2.1 Introduction

Vehicular traffic control at at-grade intersections has been extensively studied in the traffic science and engineering literature. Intersections are where conflicting traffic movements compete for the right-of-way, and thus delay or accidents will likely occur. The essence of intersection control is resource/right-of-way allocation. In general, the problem involves determining spatio-temporal trajectories for vehicles approaching the intersection to maximize throughput or minimize delay while ensuring a safe buffer between these trajectories in the space-time prism. Moreover, these spatio-temporal trajectories must respect certain kinodynamic constraints such as maximum acceleration, deceleration, or speed (these constraints can be reflected by the limits on curvature and slope of a spatio-temporal trajectory). Finding optimal intersection control belongs to the class of NP-complete problems even under various simplifying assumptions such as relaxing the acceleration/deceleration limits [Dasler and Mount, 2015].

In practice, this daunting task is accomplished by heuristic right-of-way allocation rules or principles. For example, a stop-sign control allocates the intersection's entire space-time prism to a single vehicle during its passage to ensure safety. It is designed for low-traffic intersections and thus yields low throughput and high delay when implemented at intersections with high traffic demand. Traffic signal control (TSC), on the other hand, has served as the primary means of controlling critical, high-traffic intersections for almost a century. Its right-of-way allocation principle decomposes the traffic demand into a number of movement groups (each group is a union of spatially non-conflicting movement streams; a movement stream is a group of through/turning lanes with the same entering and exiting road/approach). It then allocates the right-of-way to one group at a time while keeping all other groups waiting. Typically, this allocation is cyclic, and thus signal timing aims to optimize the length of the cycle and its allocation among movement streams (the so-called green split) [Urbanik et al., 2015].

The advancement of connected and automated vehicle (CAV) technology has motivated researchers to investigate whether the technology can be leveraged to transform intersection control. In addition to the provision of real-time vehicle location information, the CAV technology substantially decreases the required buffer in the space-time prism between two moving vehicles to ensure a safe passage. More importantly, in a fully CAV environment, the right-of-way allocation rule can be more complex, to the point that would otherwise confuse a human driver. The proposed control on this quest usually consists of two stages: a planning/scheduling algorithm puts the approaching CAVs into an order, and then a cooperative longitudinal motion controller controls the motion of CAVs for their passage of the intersection. Below we elaborate on two most relevant control approaches investigated in the literature: reservation-based schemes and rhythmic control. For a more comprehensive review, see, e.g., Rios-Torres and Malikopoulos [2016].

Initiated by Dresner and Stone [2008], the reservation-based scheme requires vehicles approaching an intersection to send a reservation request to an intersection manager, which subsequently decides whether it is safe to accommodate the request. If positive, the manager accepts the reservation and provides a motion control plan that satisfies the safety and kinodynamic constraints. Otherwise, the intersection manager rejects the request and responds with a counter offer. Fundamentally, this scheme performs a grid decomposition of the space-time prism of the intersection into reservation tiles. Then, it finds a safe passage for a new vehicle by ensuring the associated tiles for the movement of this vehicle do not overlap with those reserved by previously confirmed requests. The reservation-based scheme inherently follows the first-come-first-served principle to process the requests. Several follow-up studies aim to improve upon this principle by batching the requests and optimizing the sequence in which the intersection manager processes them [Levin et al., 2016, Levin and Rey, 2017]. However, it has been pointed out that finding an optimal sequence is not tractable, especially for high traffic regimes. That being said, the reservation-based scheme has been demonstrated to effectively reduce intersection delay as compared to TSC in low traffic regimes.

Recently, Chen et al. [2021] proposed a novel rhythmic control scheme that redesigns the intersection layout and performs a pre-decomposition of the space-time prism into a union of spatio-temporal trajectories. Then, vehicles approaching the intersection would traverse through the first unassigned spatio-temporal trajectory. This scheme has been shown to substantially outperform TSC reasonably balanced traffic demand patterns using simulation. A balanced traffic demand pattern is one in which the per-lane demand for each movement stream, turning or through, is the same. However, the scheme shows an inferior performance when dealing with unbalanced demand patterns. Moreover, it imposes several restrictive assumptions on vehicles' specifications and intersection layout to facilitate the pre-decomposition of the space-time prism. These restrictions include homogeneous vehicle dimensions and increased lane spacing, among others.

2.1.1 Contributions

Given the NP-hardness of the intersection control problem, a control with a guaranteed approximation factor would be desirable. Unfortunately, to the best of our knowledge, there is no known control scheme that can guarantee a near-optimal throughput for a generic traffic demand pattern. This paper aims to fill this void by proposing an approximation algorithm that leverages graph coloring techniques. We hereinafter refer to this algorithm as graph coloring control (GCC). GCC is unifying because it includes as a special case the traditional signal control for manually driven vehicles, several reservation-based schemes, and the rhythmic control for automated vehicles. Consider the two main objectives of an efficient intersection control, namely maximizing throughput and minimizing delay. For the former, given a sufficiently large footprint for the intersection and a generic demand pattern, GCC provides a polynomial-time approximation scheme for the throughput maximization problem. For the latter, with stationary, admissible vehicle arrivals, the delay that each vehicle experiences is less than a constant value that linearly depends on the number of movement streams and the inverse of the approximation algorithm's precision factor. Moreover, this constant value is independent of the number of lanes within each movement stream and the intersection's demand pattern.

2.1.2 Organization

Section 2.2 states the problem and the study assumptions. Section 2.3 presents the approximation algorithm and its performance guarantees for a simplified version of the intersection control problem. Section 2.4 generalizes the algorithm and its performance guarantees to the most general case of the problem. Section 2.5 discusses the relationships between GCC and other controls. Section 2.6 presents simulation results to demonstrate the performance of GCC in various scenarios. Lastly, Section 2.7 summarizes our GCC algorithm and generalizes it to solve a subclass of job-shop scheduling problems with sequence-dependent setup times.

2.2 Problem Statement

Consider a generic intersection layout. Let L represent the set of all left-turn, through, and right-turn lanes. If the intersection includes merged, shared, or diverged lanes, we consider them as separate lanes in L. Let N be the set of conflict points representing the crossing locations of these lanes. The nodes in N divide each lane $l \in L$ into a set of consecutive directed segments, A_l. Let

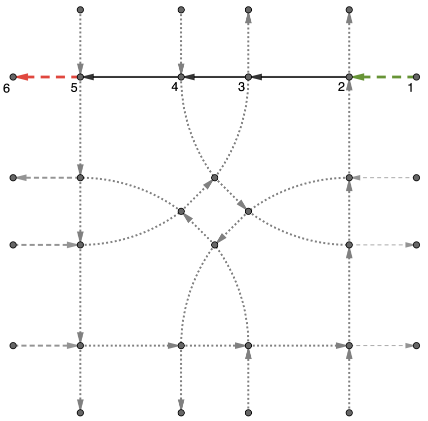


Figure 2.1: Representation of a lane and its segments inside an intersection

us denote by $A = \cup_l A_l$ the union of these segments and indicate by $G = (N, A)$ the underlying digraph of the intersection. Also, we call the first and last segment of each lane the entrance and exit segment of that lane, respectively. We refer to other segments as intermediate. Figure 2.1 demonstrates an intersection with eight lanes, where the through lane l_1 is highlighted. The set of its intermediate segments $\{(2,3),(3,4),(4,5)\}$ are shown by solid lines, and its entrance and exit segments $\{(1,2),(5,6)\}$ are shown by dashed lines. Dotted lines show other segments.

Now consider a set of CAVs of different sizes crossing the intersection. The intersection manager assigns an entrance time and an acceleration/deceleration profile that satisfies safety and kinodynamic constraints for each vehicle approaching the intersection to maximize the throughput of the intersection. For a meaningful representation of the intersection throughput, we adopt the concept of reserve capacity, defined as the greatest common multiplier of the existing traffic demand pattern that can be accommodated subject to lane capacity and other constraints [Allsop, 1972]. The reserve capacity has also been utilized to measure the capacity for signal-controlled traffic networks [Wong and Yang, 1997]. Finding the reserve capacity for general networks with multi-commodity flows is known as the maximum concurrent flow problem [Shahrokhi and Matula, 1990].

Given an intersection layout and its underlying traffic demand pattern, we attempt to devise an intersection control to maximize the reserve capacity of the intersection. To mathematically define this problem, we discretize the study horizon, say, one hour, into time intervals of length δ_t to form a set $T = \{t_1, t_2, \ldots, t_r\}$, where $t_i \in T$ represents the i^{th} ordered time interval. The intersection manager assigns an entrance time and a pre-specified acceleration/deceleration profile upon each vehicle's arrival. To model each vehicle's acceleration/deceleration profile, we discretize each segment $a \in A_l$ by introducing additional tracking points. The vehicle moves forward to the next tracking point along its lane during each time interval. The intersection manager must ensure the existence of an acceptable longitudinal motion control that can materialize the discretized movement plan. Note that the segment discretization is vehicle-specific and may differ between vehicles passing over the same segment. More specifically, the longitudinal motion control must satisfy kinodynamic constraints such as respecting the maximum acceleration, deceleration, and speed. Moreover, to ensure a safe passage, define a following headway as the minimum safe tip-to-tail time difference between the passage of two vehicles from the same lane over the same conflict point. We assume this time difference is the same for all pairs of vehicles traveling on the same lane and denote it by h_f. Also, define a crossing headway as the minimum safe tip-to-tail time difference between the passage of two vehicles from two different lanes over the same conflict point. Although the crossing headway may differ between different pairs of vehicles, we denote by h_c its maximum for all pairs of vehicles in the system. Considering vehicle dynamics and control, we know that these two headways are not necessarily equal and h_f is likely less than h_c.

We now formulate a linear program to find an upper bound on the reserve capacity of the intersection. The idea is that the right-of-way at each conflict point of the intersection can be allocated to at most one vehicle at any time during the study period. As vehicles in the system may have different lengths, we denote by U_l the sum of lengths of all vehicles in the demand for lane l during the study horizon, and by d_l the demand for lane l in the number of vehicles during the study horizon. To compute an upper bound on the reserve capacity of the intersection, we define Problem 2.1 as follows:

$$\max \quad \alpha \tag{2.1a}$$

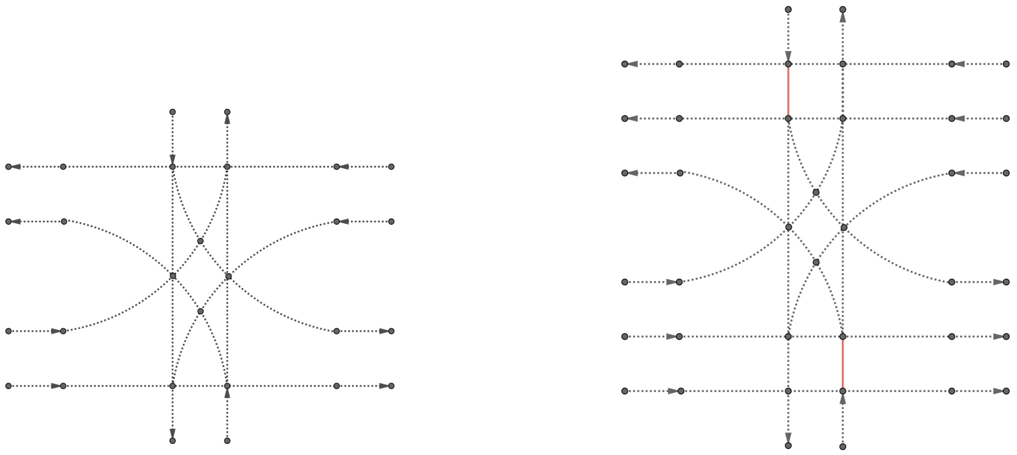
$$\text{s.t.} \quad H \geq \alpha \sum_{l:e=(j,i)\in l} (\frac{U_l}{v_{\max}} + d_l h_f); \qquad \forall i \in N \tag{2.1b}$$

where objective 2.1a is to maximize the reserve capacity. H is the length of the study horizon and $\frac{U_l}{v_{\max}} + d_l h_f$ is the minimum time required for a set of d_l vehicles with total length U_l to pass through a conflict point with headway h_f. For constraint 2.1b, note that the right-of-way of conflict point i is allocated to the lanes crossing it during the study time H. For a control to have a reserve capacity of α, it must allocate the right-of-way of conflict point i to each lane l crossing it for at least $\alpha(\sum_{l:e=(j,i)\in l} \frac{U_l}{v_{\max}} + d_l h_f)$ time.

Lastly, unless explicitly specified, we focus on simple intersections defined as follows:

Definition 1. *In a simple intersection, starting from each intermediate segment $a \in A$, for all lanes passing the segment a, there is a unique way to exit the intersection. An intersection is complicated otherwise.*

Figure 2.2 presents examples of simple and complicated intersections. In the complicated intersection on the right, the segments, depicted in solid red lines, are shared by lanes that exit the



(a) A simple intersection layout

(b) A complicated intersection layout

Figure 2.2: At-grade road intersection layouts

intersection from different outbound approaches.

Note that an intersection with merged lanes, like the merged left-turn and through lanes, can be a simple intersection. The key difference between simple and complicated intersections is to consider the intermediate segments in the underlying digraph of the intersection. Additionally, if a left-turn lane is merged with a through lane, the intersection is still simple as long as the two lanes do not diverge along an intermediate segment. We further note that complicated intersections are not common in practice. They typically appear when a one-lane street intersects with a multi-lane street (the case in Figure 2.2(b)). Although, our method can be modified to accommodate complicated intersection, it cannot achieve the same performance guarantees as the ones we present for simple intersections.

2.3 GCC for Dimensionless Vehicles

To facilitate the presentation of our scheme, we start with a simplified setting where vehicles are dimensionless, and both the following and crossing headways are set equal to the time discretization unit, δ_t.

2.3.1 Motivating Example and Intuition

In this section, we provide a motivating example to illustrate how graph coloring can be leveraged to avoid collision when routing spatially conflicting traffic movements. Consider a directed gird network $\hat{G}$ (Figure 2.3a) with six set of vehicles traveling from their origin to their destination

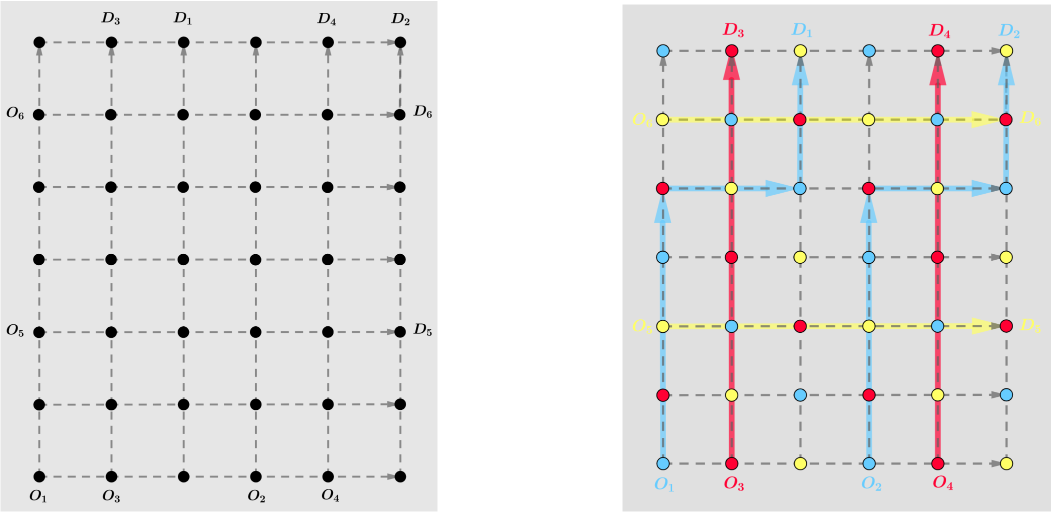
(O_i, D_i), $i = 1, 2, ..., 6$. We now route these vehicles to traverse the network without collision, i.e., ensuring that no two vehicles occupy the same node at the same time. We first decompose these six sets of vehicles into three movement groups while assigning a path for each set of vehicles. The decomposition is shown in Figure 2.3b, where we assign the same color to the paths belonging to the same movement group.

We subsequently assign three colors to the vertices of $\hat{G}$ with color 1 illustrated by cyan, color 2 by red, and color 3 by yellow (Figure 2.3b). Note that the colors are assigned to nodes of the digraph such that each edge is directed from a node with color $r \equiv \pmod 3$ to a node with color $r + 1 \equiv \pmod 3$. Additionally, the start nodes O_1 and O_2 posses color 1, start nodes O_3 and O_4 posses color 2, and start nodes O_5 and O_6 posses color 3. We discretize the time horizon into time steps $t \in \{t_1, t_2, \ldots, t_r\}$. Then, we release one vehicle at each of the start nodes at the beginning of the j-th time interval if and only if $j \equiv 0 \pmod 3$. Then, vehicles in each set are required to follow their associated path from their start to end node moving one node forward during each time interval. Consequently, we can observe that at the start of any time interval t, all vehicles with endpoints (O_1, D_1) and (O_2, D_2) are in nodes of color $1 + t \equiv \pmod 3$, all vehicles with endpoints (O_3, D_3) and (O_4, D_4) are in nodes of color $2 + t \equiv \pmod 3$ and all vehicles with endpoints (O_5, D_5) and (O_6, D_6) are in nodes of color $t \equiv \pmod 3$.

Note that there are two types of potential collisions. The first type is between two vehicles assigned to the same movement group, and the other is between two vehicles assigned to different movement groups. We implicitly avoid the first type of collision by enforcing vehicles in each of the six groups to follow their associated paths that are spatially non-conflicting. To prevent the second type of collisions, observe that at each time step $t \in \{t_1, t_2, \ldots, t_r\}$, the vehicles that occupy nodes of the same color are from the same movement group. So there is no collision between vehicles from different movement groups. As a result, no two vehicles will be at the same node at the same time.

In this example, to route spatially conflicting traffic movements, we first decompose them into three movement groups; each group contains movements that follow spatially non-conflicting paths. Afterward, we color the underlying digraph of the network with three colors and leverage the graph coloring property of the network to allocate the right-of-way of each node to one movement group during each time step. Finally, we allow the vehicles from each movement group to enter an entrance node at time steps that right-of-way of the entrance node is allocated to their associated group. Then, these vehicles can continue their movement along their associated path moving forward one node during each time step.

In the same spirit, our GCC algorithm for intersection control follows these three steps, as illustrated in Figure 2.4. At the first step, GCC decomposes the traffic demand to be served into groups of non-conflicting movements. This step appears similar to what traditional traffic signal



(a) The grid network

(b) colored grid network

Figure 2.3: Collision-free example

control or TSC does, but the key difference is that in TSC, the movement along each lane, e.g., a left-turn movement, needs to be entirely contained in at least one of these groups. To ensure each lane is covered the same number of times with the movement groups, this entirety inclusion constraint may increase the number of required movement groups to serve the intersection demand. As we will show later, the number of groups in the decomposition phase of GCC is directly related to the throughput associated with each movement group. In particular, if we decompose the traffic demand into k groups and implement GCC on a k-colorable digraph, we can admit traffic from each group for $\frac{1}{k}$ fraction of the study horizon. Section 2.3.2 discusses how demand decomposition is implemented in a generic directed graph.

At the second step, we exploit graph coloring techniques to allocate the right-of-way at conflict points of the intersection. In particular, note that a key feature in the motivating example in Section 2.3.1 is that the vertices in the underlying digraph of the intersection can be colored with three colors such that for each node of color $i \equiv (\bmod\ 3)$ all outgoing edges are connected to nodes of color $i+1 \equiv (\bmod\ 3)$. We exploit this feature to allocate the right-of-way at each node with color i at time step $t \equiv (\bmod\ 3)$ to the movement group $i-t \equiv (\bmod\ 3)$. Section 2.3.3 generalizes this idea into periodic k-colorable digraphs and discusses how to modify the right-of-way allocation in this general case.

At the third step, GCC assigns the movement of each vehicle along each segment of its lane to one of the decomposition movement groups. Doing so, some vehicles might switch between movement groups while crossing the intersection. Therefore, it is essential to ensure that vehicles do not block each others' movement when awaiting to switch to another movement group.

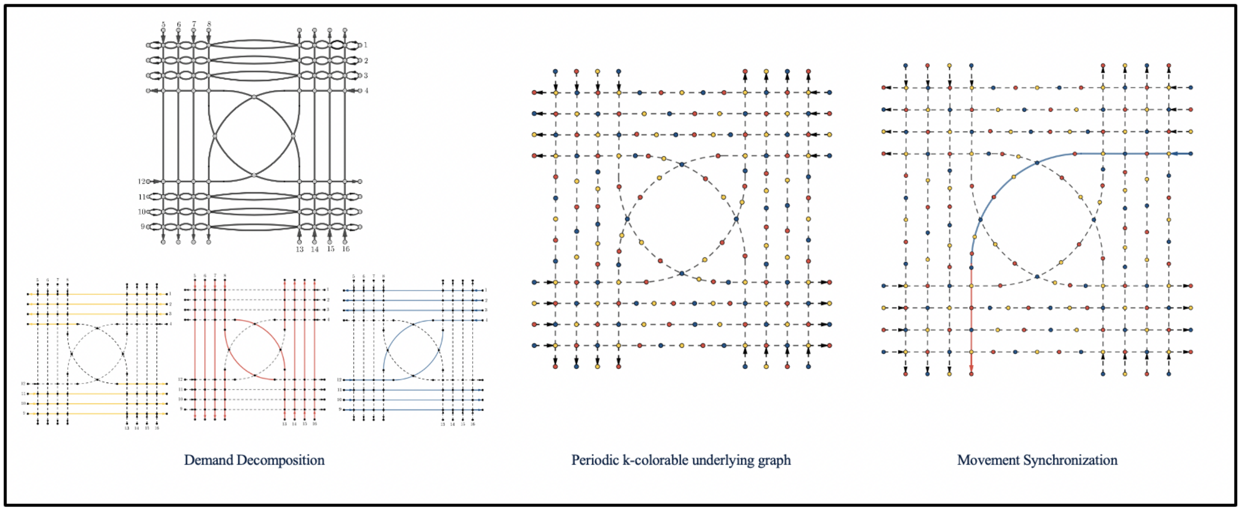


Figure 2.4: Overview of the GCC algorithm

Note that vehicles in each movement group can continue their movement along a path entirely contained in their movement group moving forward one node during each time step and the cyclic right-of-way allocation can ensure collision avoidance. However, as a result of breaking the entirety inclusion constraint, the vehicles traveling along some of the lanes must switch between the movement groups. Section 2.3.4 illustrates the details of this procedure.

2.3.2 Step 1: Demand Decomposition

As aforementioned, the demand decomposition component in GCC deliberately removes the entirety inclusion constraint by breaking each through or turn lane into its segments. Then, it finds a set of non-conflicting movement groups that can cover the demand for these segments. The intent of this key difference is to reduce the number of decomposition groups, thereby improving the intersection's throughput. To be specific, GCC considers an optimal solution, α^*, to Problem 2.1 and breaks the right-of-way allocation to each lane, i.e., $\alpha^*(\frac{U_l}{v_{\max}} + d_l h_f)$, into segment allocated right-of-way $d_a = \sum_{l:a\in l} \alpha^*(\frac{U_l}{v_{\max}} + d_l h_f) \, : \, a \in A$. Here, the summation is over the set of lanes that include segment a. Then, GCC discretizes the segment allocated right-of-way to exploit the graph decomposition techniques that can find a set of non-conflicting movement groups covering the demand for those segments. To do so, denote by Δ the maximum number of lanes that pass through a node $i \in N$. Then, GCC discretizes the segment allocated right-of-way into multiples of $H/(k - \Delta + 1)$, i.e., for each segment $a \in A$, we replace segment a with $\sum_{l:a\in l} \left\lceil \frac{\alpha^*(\frac{U_l}{v_{\max}} + d_l h_f)(k-\Delta+1)}{H} \right\rceil$ parallel edges between the endpoints of a in the original digraph G. This yields a new directed multigraph $\tilde{G}$ such that the number of parallel edges between the endpoints of a segment is an index of demand for that segment. Subsequently, GCC decomposes $\tilde{G}$ into a union of non-conflicting movement groups. Note that a non-conflicting movement group is a union of mutually exclusive directed paths and directed cycles, which are defined as directed $\{1, 2\}$-factors [Chiba and Yamashita, 2018] in the graph theory literature. Thus, we hereinafter refer to non-conflicting movement groups by directed $\{1, 2\}$-factors.

Definition 2. *[Chiba and Yamashita, 2018] A directed $\{1, 2\}$-factor D_i in a directed graph is a directed subdigraph formed by the union of directed cycles $C = \bigcup C_i$ and directed paths $P = \bigcup P_i$ that do not include common nodes.*

Definition 3. *[Chiba and Yamashita, 2018] A directed 2-factor of a digraph is a spanning subdigraph in which every component is a directed cycle.*

To decompose the directed multigraph $\tilde{G}$ into a union of directed $\{1, 2\}$-factors, we utilize the 2-factor theorem introduced by Lovász and Plummer [2009]:

Theorem 5. *[Lovász and Plummer, 2009] Let G be a directed graph whose nodes have in-degree and out-degree equal to k. Then the edges of G can be partitioned into k edge-disjoint 2-factors.*

Theorem 5 implies that for regular digraphs $\tilde{G}$, a digraph whose nodes have the same in-degree and out-degree, the number of subflow groups required to cover edges in $\tilde{G}$ equals the maximum in/out-degree of $\tilde{G}$. Therefore, we find an upper bound on the degree of a node $i \in \tilde{G}$ in Lemma 3.

Lemma 3. *The maximum in-degree and out-degree for the digraph G^* is at most k.*

Proof. Proof. See Appendix A.15.

□

To use Theorem 5 on directed multigraph $\tilde{G}$, we are required to add edges to $\tilde{G}$ to construct a directed multigraph G^* whose nodes have in-degree and out-degree equal to k. This can be done in an iterative procedure where we add a virtual directed edge from a node v with out-degree less than k to a node u, possibly the same as v, with in-degree less than k. As the total in-degree and total out-degree of a digraph are equal and as the in/out degree of all vertices are less than or equal to k, there is a node with in-degree less than k if and only if there is a node with out-degree less than k. Finally, we utilize Theorem 5 to decompose the edges in G^* into k 2-factors. Removing the edges in $G^* \setminus \tilde{G}$ we obtain a $\{1, 2\}$-factor decomposition of $\tilde{G}$. To be self-contained, we provide a brief description of the algorithm adopted from Lovász and Plummer [2009] in Appendix A.16. For brevity, hereinafter, we refer to the edges $\tilde{e} \in \tilde{A}$ resulted from applying the decomposition algorithm on segment $e \in A$ as parallel edge copies of e. Also, to avoid confusion, we refer to elements of A as segments and to elements of $\tilde{A}$ as edges.

2.3.3 Step 2: Making Underlying Digraph Periodic k-colorable

To generalize the graph coloring feature exploited in the motivating example in Section 2.3.1, we define a periodic k-colorable digraph as follows:

Definition 4. *A digraph is periodic k-colorable if its vertices can be colored with k colors such that for each node with color $i \equiv \pmod{k}$ all outgoing edges are connected to nodes with color $i+1 \equiv \pmod{k}$.*

Trivially, not all digraphs are periodic k-colorable. However, Theorem 6 shows how to add tracking points to any generic network to construct a modified periodic k-colorable digraph.

Theorem 6. *Consider a multigraph G, and assume there exist the same number of initial tracking points along any two parallel edges. We can add additional tracking points to transform G into a periodic k-colorable digraph. Moreover, the numbers of tracking points along parallel edges remain equal.*

Proof. Proof. See Appendix A.17.

□

Next, we discuss how to allocate the right-of-way at each node to directed $\{1,2\}$-factors crossing that node in a periodic k-colorable digraph. For each edge $\tilde{a} \in \tilde{A}$, denote by $z^1_{\tilde{a}}$ the number of tracking points required to traverse it with the maximum allowable speed $v_{\max}$, this value can be obtained by dividing the time it takes to traverse $\tilde{a}$ with maximum allowable speed by the length of the time intervals δ_t. Add $z^1_{\tilde{a}}$ tracking points along each edge $\tilde{a} \in \tilde{A}$ and apply the algorithm presented in Theorem 6 to obtain the periodic k-colorable digraph $\bar{G}$. Decompose $\bar{G}$ into k directed $\{1,2\}$-factors $D_p : p \in \{1,2,\dots,k\}$. Consider a set of vehicles whose routes belong to one of these directed $\{1,2\}$-factors. We allocate the right-of-way of a node $v \in \bar{G}$ possessing color $i \equiv (\bmod\ k)$ to vehicles travelling along D_p only in time intervals $t \equiv i-p(\bmod\ k)$. As a result, we ensure at any time step $t \equiv \pmod{k}$ all vehicles traveling along a path contained in D_p are at nodes possessing color $i \equiv t+p\pmod{k}$. Hence, we can make sure each node i cannot be occupied by vehicles from more than one directed $\{1,2\}$-factor D_p. Doing so, we associated a set of time intervals to directed $\{1,2\}$-factors that pass through each node. In other words, we allocate the right-of-way at each node as a resource to directed $\{1,2\}$-factors crossing that node.

2.3.4 Step 3: Movement Synchronization

This last step of GCC is to devise a scheduling algorithm that synchronizes the movement of vehicles through the intersection to ensure a collision-free route assignment while obtaining a throughput for the intersection close to the upper bound obtained in Problem 2.1.

The graph coloring control, upon arrival of each vehicle at the intersection, considers its associated lane. Then, it assigns the movement of that vehicle along each segment of the associated lane to a directed $\{1,2\}$-factor. Afterward, the vehicle crosses each node, including its entrance node to the intersection, according to the right-of-way allocation presented in Section 2.3.3. In particular, it crosses each node during a time interval that the right-of-way is allocated to the directed $\{1,2\}$-factor containing the movement on the segment directed out from that node.

Note that if the scheduling algorithm assigns the movement of a vehicle along each segment of its associated lane to the same directed $\{1,2\}$-factor, the vehicle can move forward one tracking point during each time step and the scheduling algorithm ensures collision avoidance. To conclude that, we should consider two cases. First, the right-of-way allocation in section 2.3.3 ensures that vehicles from different $\{1,2\}$-factors do not collide. Second, each directed $\{1,2\}$-factor resulted from the decomposition phase does not intersect itself, so vehicles travelling along a $\{1,2\}$-factor do not collide.

Furthermore, recall from Section 2.3.2 that we break the demand for each lane into demand for its segments. Thus, we cannot make sure each of the final $\{1,2\}$-factors contains a parallel edge copy of all segments for a lane. As such, the vehicles crossing the intersection might switch between different $\{1,2\}$-factors. Consider a vehicle moving along lane l aims to switch from an edge $\tilde{a} \in \tilde{A}$ to an edge $\tilde{b} \in \tilde{A}$ at a node v with color i. Also, $\tilde{a}$ belongs to the directed $\{1,2\}$-factor D_p and $\tilde{b}$ belongs to the directed $\{1,2\}$-factor D_q. If we add $p - q \equiv \pmod{k}$ virtual tracking points on the last edge on D_p, $\tilde{a}$, prior to transition to D_q, we can ensure this vehicle will cross the transition node during a time interval t that satisfies $t \equiv i - p + (p - q) \equiv i - q (\mod k)$. Thus, this vehicle can start its movement on D_q upon completion of its switch from D_p to D_q. The right-of-way allocation presented in Section 2.3.3 ensures that if vehicles move forward one tracking point during each time step no two vehicle from different $\{1,2\}$-factors appear at a node of the digraph during the same time interval. The only remaining caveat is to ensure these vehicles will not block each other's paths while waiting to switch between directed $\{1,2\}$-factors. To address this caveat, we first restate the problem of assigning the vehicle movement along each segment of its lane to a directed $\{1,2\}$-factor as an instance of the general matching problem.

To mathematically approach this problem, for each edge $\tilde{a} \in \tilde{A}$ recall $z^1_{\tilde{a}}$ is the number of tracking points required to traverse it with the maximum allowable speed $v_{\max}$. Also, denote by $z^2_{\tilde{a}}$ the number of additional points on $\tilde{a}$ we obtain from implementing Theorem 6 to make the directed multigraph $\tilde{G}$ periodic k-colorable. For each entrance/intermediate edge $\tilde{e} \in \tilde{A}$ denote by $A^+(\tilde{e})$ the set of edges in $\tilde{A}$ that follow $\tilde{e}$ along any lane l. As we consider simple intersections, all edges in the set $A^+(\tilde{e})$ are parallel edge copies of the same segment $b \in A$.

To find a feasible switching scheme, we define a new graph $\check{G} = (\check{V}, \check{A})$ where each intermediate edge $\tilde{a} \in \tilde{A}$ is replaced by two vertices $v^+(\tilde{e}), v^-(\tilde{e}) \in \check{V}$, each entrance edge is replaced by one vertex $v^+(\tilde{e})$ and each exit edge is replaced by one vertex $v^-(\tilde{e})$. Also, there is a directed edge $\check{a} \in \check{A}$ from each node $v^+(\tilde{e}) \in \check{V}$ to all nodes $v^-(\tilde{b}) \in \check{V}$ corresponding to edges $\tilde{b} \in A^+(\tilde{e})$.

Furthermore, for each edge $\check{a} \in \check{A}$ between nodes corresponding to a pair of consecutive edges $\tilde{e}, \tilde{f} \in \tilde{A}$, define $w'_{\tilde{e},\tilde{f}} \equiv z^2_{\tilde{e}} + p - q(\mod\ k)$, and denote by $w_{\tilde{e},\tilde{f}} = w'^{(1+\epsilon)}_{\tilde{e},\tilde{f}}$ the weight of $\check{a}$. Here, $\tilde{e}$ and $\tilde{f}$ belong to the $\{1, 2\}$-factors p and q, respectively. Also, $\epsilon > 0$ is a sufficiently small positive number. Finally, we implement standard algorithms, such as the blossom algorithm by Edmonds [1965a,b], to find a minimum weight perfect matching in $\check{G}$.

Lemma 4. *The graph $\check{G}$ contains at least one perfect matching.*

Proof. Proof. See Appendix A.20.

□

Note that each edge $(v^+(\tilde{e}), v^-(\tilde{f}))$ in a perfect matching in graph $\check{G}$ corresponds to a switch between the $\{1, 2\}$-factor containing $\tilde{e}$ to the $\{1, 2\}$-factor containing $\tilde{f}$. Therefore, each intermediate edge in $\tilde{G}$ is matched with an incoming edge as well as an outgoing edge in $\tilde{G}$. Also, each entrance and exit edge is matched with an outgoing and incoming edge in $\tilde{G}$, respectively. Following GCC, vehicles from each lane form a queue before entering the intersection. Then, assuming we are at time step t, we determine the entrance time t_1 of the vehicle in front of a queue behind start segment $a \in A$ to intersection according to the right-of-way allocation in Section 2.3.3. In particular, the vehicle enters a node with color i at the minimum time step $t_1 \geq t$ that satisfies $t_1 \equiv (i - p)(\mod\ k)$ for a $\{1, 2\}$-factor D_p that contains a parallel edge copy of segment a, i.e, $\tilde{a}$. Then the vehicle travels along its lane by moving forward one tracking point in $\tilde{G}$ at each time interval. When the vehicle traversed all $z^1_{\tilde{e}}$ tracking points on an edge $\tilde{e}$ it will switch to an edge $\tilde{f}$ such that the edges $v^+(\tilde{e})$ and $v^-(\tilde{f}))$ are matched in the minimum weight perfect matching obtained. Furthermore, to switch from $\tilde{e}$ to $\tilde{f}$ we should add $w_{\tilde{e},\tilde{f}}$ additional tracking points on e. As the intersection is a simple intersection, all lanes that contain a section $a \in A$ will coincide with the unique way to exit the intersection. Therefore, vehicles do not deviate from their associated lane when switching between edges. For readers convenience, we provide a brief description of the process of finding a feasible switching scheme, presented in this section, as Algorithm 3 presented in Appendix A.24. Finally, Lemma 5 proves that following this procedure no vehicle blocks another vehicles' path while waiting to switch between segments.

Lemma 5. *If we follow the switching scheme presented in Section 2.3.4, we can ensure no vehicle blocks another vehicle's movement.*

Proof. Proof. See Appendix A.18.

□

For readers convenience, we provide a brief description of dimensionless GCC in Algorithm 1.

Algorithm 1 Dimensionless Graph Coloring Control

```
1. Solve Problem 2.1 in intersection underlying digraph $G$ to find maximum reserve capacity $\alpha^*$.
2. Replace each edge $e \in G$ by $\sum_{l:a\in l} \left\lceil \frac{\alpha^*(\frac{U_l}{v_{\max}}+d_l h_f)(k-\Delta+1)}{H} \right\rceil$ parallel edges to find multigraph $\tilde{G}$.
3. Decompose $\tilde{G}$ into a union of $k$ mutually exclusive $\{1,2\}$-factors, $D_1, D_2, \ldots, D_k$.
4. Add $z_{\tilde{e}}^2$ tracking points on each edge $\tilde{e} \in \tilde{G}$ to modify $\tilde{G}$ into a periodic $k$-colorable digraph.
5. Allocate the right-of-way of a node $v \in \bar{G}$ possessing color $i \equiv (\bmod\ k)$ to a vehicle travelling along $D_p$ only in time intervals $t \equiv i - p(\bmod\ k)$.
6. Construct the weighted graph $\check{G}$. Then, obtain a feasible switching scheme as a minimum weight perfect matching $\check{G}$.
7. Find entrance time to intersection for a vehicle according to the right-of-way allocation rule.
8. The vehicle travels along its lane by moving forward one tracking point in $\tilde{G}$ during each time interval. When traversed all $z_{\tilde{e}}^1$ tracking points on $\tilde{e}$ it will switch to an edge $\tilde{f}$ such that $v^+(\tilde{e})$ and $v^-(\tilde{f})$ are matched in $\check{G}$ (To switch from $\tilde{e}$ to $\tilde{f}$ we should add $w_{\tilde{e},\tilde{f}}$ tracking points on $\tilde{e}$).
```

Theorem 7. *Algorithm 1 obtains an objective value of at least $\frac{k-\Delta+1}{k}\alpha^*$ where α^* is the optimal value of the optimization problem 2.1.*

Proof. Proof. See Appendix A.19. □

For an illustrative example of GCC for dimensionless vehicles we refer readers to Appendix A.21. Also, we refer the readers to the supporting materials (2.7.2) for a video demonstrating the operations of dimensionless graph coloring control. Note that, at each moment, the color of a vehicle indicates the $\{1,2\}$-factor associated with the movement of the vehicle on $\tilde{G}$.

2.4 Generalizing GCC for Practical Implementation

This section generalizes GCC by allowing arbitrary vehicle length and different following and crossing headways. We also discuss the issues regarding the intersection space limitation for accommodating additional tracking points.

As mentioned in Section 2.1, the following headway is expected to be shorter than the crossing headway, and the upper bound obtained in Problem 2.1 can be achieved under the condition that

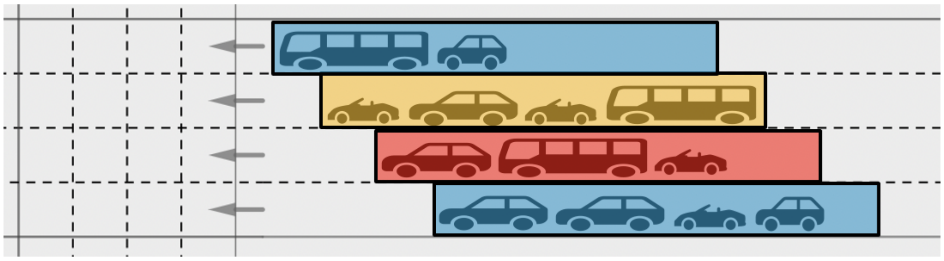


Figure 2.5: four traffic units moving in parallel lanes

all consecutive vehicles must maintain the following headway h_f. Thus, it is intuitive that in order to improve the objective value we may want that most of the consecutive vehicles passing over the same node belong to the same lane, forming a platoon of vehicles. Furthermore, because vehicles in the system may have different lengths, it is desirable to find a way to normalize the demand for each lane to facilitate the operations of a proposed control. Thus, GCC introduces the concept of traffic unit. The idea is to decompose the demand for each lane into units of traffic with the same time length, u, i.e., the difference between the time that the tip of a traffic unit passes over a node and the time its tail passes the point is u time steps of length δ_t. Then, within each traffic unit, the vehicles' tip-to-tail headway equals the minimum following headway h_f. Figure 2.5 demonstrates four traffic units of the same length moving in parallel lanes. As demonstrated in Figure 2.5, in real-time operations it may happen that a traffic unit is not fully utilized, either due to the low demand magnitude or the configuration of consecutive vehicle's lengths.

To address the issue of arbitrary vehicle length and different headway for following and crossing vehicles, we find a collision-free routing scheme for a set of traffic units consisting of a leading vehicle followed by a number of following vehicles that together can fit into a traffic unit of length $u - \delta$ time steps of length δ_t. Here, δ represents the difference between the following and crossing headways in time steps, $\frac{h_c - h_f}{\delta_t}$. This difference accounts for the potential offset in curvilinear position between two vehicles in merging/crossing paths. Thus, the safe following headway is implicitly satisfied since it is inherent to the definition of traffic units. Moreover, the crossing headway is satisfied since we discard headway δ time steps from the tail of each traffic unit. Below, we provide a motivating example to demonstrate the basic idea that ensures a minimum safe distance between leading vehicles released from each lane.

2.4.1 Motivating Example

Recall the network $\hat{G}$ and six groups of vehicles in the motivating example in Section 2.3.1. Also, assume all vehicles in the system move forward to the next tracking point along their associated paths. Consider a set of traffic units with the same time length equal to two time steps. The

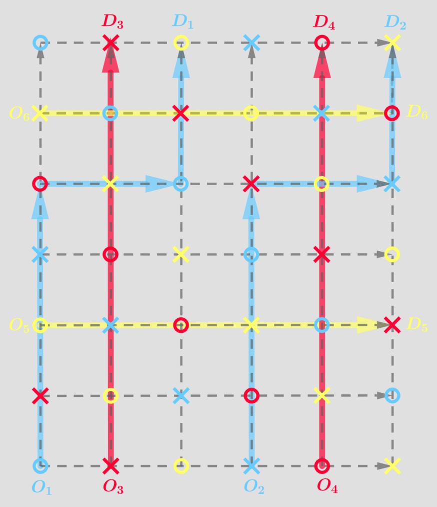


Figure 2.6: Minimum distance example

basic idea is to synchronize the movement of the leading vehicles in each traffic unit such that two main criteria are satisfied: first, no two vehicles occupy the same node in $\hat{G}$; second, the distance between any two nodes occupied by leading vehicles is at least the length of a traffic unit. As such, it takes at least two time steps for each leading vehicle to be at the location of another leading vehicle. Thus, the traffic units are non overlapping. To achieve this aim, we modify the routing assignment presented in Section 2.3.1 to ensure that the distance between any two leading vehicles is a multiple of the traffic unit length.

To illustrate the new approach, we modify the graph coloring presented in Section 2.3.1 to become a 6-coloring such that each edge is directed from a node with color $r \equiv \pmod{6}$ to a node with color $r + 1 \equiv \pmod{6}$. Figure 2.6 demonstrates such coloring by separating each of the three colors presented in Figure 2.3b into two groups demonstrated by circles and crosses. Note that the distance between any two circles and any two crosses is an even number.

Furthermore, we release one vehicle at a circle start node O_i at any time interval $t = 3k, k \in \mathbb{N}$ if and only if $t \equiv 0 \pmod{2}$. Also, we release one vehicle at a cross start node O_i at any time interval $t = 3k, k \in \mathbb{N}$ if and only if $t \equiv 1 \pmod{2}$. As a result, at any time step t, either all vehicles in the system are on circle nodes or cross nodes. This means the distance between any two vehicles in the system at any time is an even number. Additionally, from the analysis in the motivating example in Section 2.4.1, we know that these vehicles do not collide. This is because we release a subset of vehicles routed in the motivating example in Section 2.3.1. In conclusion, we can ensure all vehicles in the system maintain a minimum safe distance of two nodes.

In this second example, to accommodate traffic units, we modify some of the steps presented in Section 2.3.1. First, similar to dimensionless GCC, we decompose the intersection demand into three non-conflicting movement groups. In the next step, unlike the dimensionless GCC, we color the underlying digraph of the network with six colors and leverage the graph coloring property of the network to allocate the right-of-way at each node to one traffic unit leader during each time step such that the traffic unit leaders maintain a minimum distance of two. Finally, Similar to dimensionless GCC, we allow the vehicles from each movement group to enter an entrance node at time steps that right-of-way at the entrance node is allocated to their associated group. Then, these vehicles can continue their movement along their associated path moving forward one node during each time step.

In sum, the idea in generalizing GCC is to follow the dimensionless GCC to decompose the traffic demand into k groups of non-conflicting movements. Then, modify the underlying digraph so that it admits a periodic ku-coloring. Finally, GCC exploits the periodic ku-coloring of the directed network to allocates the right-of-way at each node to tip of traffic units. In particular, GCC ensures the tip of every two traffic unit has a distance that is a non-zero multiple of u. Section 2.4.2 discusses these modifications in further details that enable the dimensionless GCC to accommodate the traffic units.

Note that introducing the concept of traffic units makes the control more robust in response to the online variations in safety measures, such as an increase in the minimum safety gap between vehicles caused by adverse weather and environmental conditions. After realizing any changes to the safety gap, we can recompute the number of allowable vehicles within each traffic unit that maintain the minimum safe distance. The control remains valid as long as traffic unit time length fits into the design length.

2.4.2 General GCC Movement Synchronization

Here, we discuss how to utilize the idea presented in the motivating example in Section 2.4.1 to generalize the design in Section 2.3. As most steps are similar to GCC for dimensionless vehicles, we focus on the differences between the two methods and omit the repetitive steps. To mathematically formalize this idea, we first follow GCC for dimensionless vehicles to decompose the traffic demand into a union of k separate $\{1, 2\}$-factors. Then, we implement Theorem 6 on directed multigraph $\tilde{G}$ resulted from the decomposition phase to make it a periodic ku-colorable digraph and denote by $z^2_{\tilde{e}}$ the number of additional points on edge $\tilde{e}$ obtained during this procedure. To regulate the right-of-way of nodes, we allocate the right-of-way of a node $v \in \bar{G}$ possessing color $i \equiv (\mod \ ku)$ to a traffic unit travelling along D_p only at time intervals $t \equiv i - pu (\mod \ ku)$. To find an acceptable switching scheme, we follow the idea presented in the dimensionless GCC with a minor modification. For a switch between two consecutive edges $\tilde{e} \in \tilde{A}$ and $\tilde{f} \in A^+(\tilde{e})$, we define $w'_{\tilde{e},\tilde{f}} \equiv z^2_{\tilde{e}} + pu - qu (\mod \ ku)$, and denote by $w_{\tilde{e},\tilde{f}} = w'^{(1+\epsilon)}_{\tilde{e},\tilde{f}}$ the weight of this switch. Then, we follow the dimensionless GCC to implement Edmonds' blossom algorithm to find a minimum weight perfect matching on the modified graph $\check{G}$.

Similar to dimensionless GCC, traffic units from each lane form a queue before entering the intersection. Then, assuming we are at time step t, we determine the entrance time t_1 of the tip of a traffic unit in front of a queue behind start segment $a \in A$ to intersection. In particular, the tip of the traffic unit enters a node with color i at the minimum time step $t_1 \geq t$ that satisfies $t_1 \equiv (i - pu)(\mathrm{mod}\ \ ku)$ for a $\{1, 2\}$-factor D_p that contains a parallel edge copy of segment a, i.e, $\tilde{a}$. Then, the tip of traffic units travels along its lane by moving forward one tracking point in $\tilde{G}$ during each time interval. When a traffic unit traversed all $z^1_{\tilde{e}}$ tracking points on an edge $\tilde{e}$ it switches to an edge $\tilde{f}$ such that the edges $v^+(\tilde{e})$ and $v^-(\tilde{f})$ are matched in the minimum weight perfect matching obtained. Furthermore, to switch from $\tilde{e}$ to $\tilde{f}$ we should add $w_{\tilde{e},\tilde{f}}$ additional tracking points on e. As the intersection is a simple intersection, all lanes that contain a section $a \in A$ will coincide with the unique way to exit the intersection. Therefore, the tip of traffic units do not deviate from their associated lane when switching between edges. As the routing scheme for the tip of traffic units is a subset of routes in a dimensionless GCC with ku colors, the result of Lemma 5 still holds for the tip of traffic units. Therefore, no traffic unit blocks another traffic unit while waiting to switch between segments. Finally, we can accompany the leading vehicle in each traffic unit with a set of following vehicles that together can fit into a traffic unit of length $u - \delta$. Note that at each time interval t the tip of all traffic units occupy a node with color $i \equiv t(\mathrm{mod}\ \ u)$. Thus, they maintain a distance that is a multiple of u. For the readers convenience, we summarize the general GCC algorithm for traffic units as follows:

Algorithm 2 General Graph Coloring Control

1. Steps (1-3) are the same as Algorithm 1

4. Add $z^2_{\tilde{e}}$ tracking points on each edge $\tilde{e} \in \tilde{G}$ to transfer $\tilde{G}$ into a periodic ku-colorable digraph.

5. Allocate the right-of-way of a node $v \in \bar{G}$ possessing color $i \equiv (\mathrm{mod}\ \ ku)$ to a traffic unit travelling along D_p only in time intervals $t \equiv i - pu(\mathrm{mod}\ \ ku)$.

6. Construct the weighted graph $\check{G}$. Then, obtain a feasible switching scheme as a minimum weight perfect matching in $\check{G}$.

7. Find entrance time of tip of a traffic unit to intersection by the right-of-way allocation rule.

8. The tip of traffic unit travels along its lane moving forward one tracking point in $\tilde{G}$ during each time interval. When traversed all $z^1_{\tilde{e}}$ tracking points on $\tilde{e}$, it switches to $\tilde{f}$ such that $v^+(\tilde{e})$ and $v^-(\tilde{f}))$ are matched in $\check{G}$(To switch from $\tilde{e}$ to $\tilde{f}$, we must add $w_{\tilde{e},\tilde{f}}$ tracking points on $\tilde{e}$) .

10. Within each traffic unit, a leading vehicle is followed by a set of following vehicles with total time length less than $u - \delta$.

Note that the spacing between the added tracking points within each section is not necessarily homogeneous. Finding the optimal spacing between these tracking points and an associated acceleration/deceleration profile that respects the kinodynamic constraints is an optimal control problem, which is out of the scope of this paper [Feng et al., 2018, Chen et al., 2021]. However, given an intersection with a sufficient footprint, we present a feasible acceleration/deceleration profile in Appendix A.22. We conclude this section by establishing the optimality guarantee of the general GCC algorithm in Theorem 8 and Lemma 5.

Theorem 8. *If we implement Algorithm 2 on an intersection with a generic demand pattern and layout, we can obtain a reserve capacity of at least* $\frac{u-\delta-u_m}{u}\frac{k-\Delta+1}{k}\alpha^*$ *where* α^* *is the optimal value of the optimization problem 2.1 and* u_m *is the maximum tip-to-tail time length of a vehicle under maximum speed,* $v_{\max}$*, in the system.*

Proof. Proof. See Appendix A.23.

□

Remark 5. *If* $\Delta = 1$*, Algorithm 2 provides a PTAS for the throughput maximization problem at a generic intersection with sufficiently large footprint.*

Proof. Proof. Set $\Delta = 1$ and choose $u \geq \frac{\delta+u_m}{\epsilon}$ to obtain the reserve capacity of the intersection as $\frac{u-\delta-u_m}{u}\frac{k-\Delta+1}{k}\alpha^*(1-\epsilon) = (1-\epsilon)^2\alpha^*$.

□

2.4.3 Restricted Regions and Maximum Delay

As mentioned previously, given an intersection with a sufficiently large footprint, there exists a feasible acceleration/deceleration profile to accommodate GCC, presented in Appendix A.22. However, to attain such acceleration/deceleration profile we might be required to add tracking points along some edges beyond their physical capacity. Moreover, it is preferred that the spacing between parallel lanes be minimized in the vicinity of entrance and exit sections of lanes to enhance safety and facilitation of the motion control of vehicles approaching/exiting the intersection. While the space limitation may affect the main results presented in the previous sections, due to the unique features of road intersections, we present a few remedies that can address the space limitation issues for most of the real-world instances. A common location for edges with limited capacity is in the intersection area of two sets of parallel lanes belonging to two movement streams. We call these regions restricted regions. Figure 2.7 demonstrates four restricted regions inside an intersection.

In what follows, we first demonstrate how to modify GCC to avoid the addition of any tracking points along the edges in restricted regions. Then, we present a few heuristic remedies to address

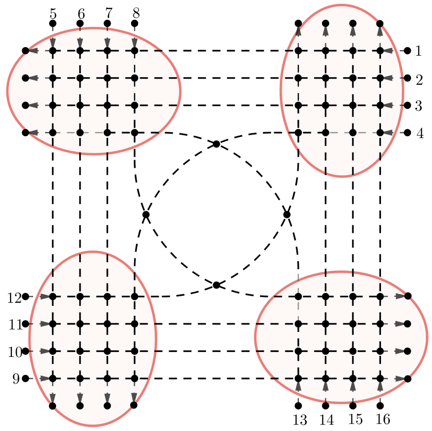


Figure 2.7: Demonstration of restricted zones inside an intersection

the space limitation problem outside the restricted regions. Lastly, we provide an upper bound on the maximum delay incurred by vehicles crossing the intersection while following GCC design.

To avoid the addition of tracking points in restricted regions, we modify Algorithm 2. Recall from Section 2.4.2 that we add $w'_{\tilde{e},\tilde{f}} \equiv pu - qu + z^2_{\tilde{e}} (\bmod \; ku)$ tracking point along an edge $\tilde{e}$ if we switch from $\tilde{e}$ to a segment $\tilde{f}$ in the matching step of Algorithm 2. Thus, it suffices to modify Algorithm 2 so that both $z^2_{\tilde{e}}$ and $qu - pu$ equal to zero.

Lemma 6. *We can modify the algorithm in Theorem 6 such that $z^2_{\tilde{e}} = 0$ for all edges $\tilde{e}$ in the intersection of two movement streams.*

Proof. Proof. See Appendix A.26.

□

To satisfy $qu - pu = 0$, we must modify the matching step in Algorithm 2. The idea is to ensure the segments of a single lane within each restricted region are not included in different $\{1,2\}$-factors, and so $p = q$. Note that as each restricted region results from the intersection of two movement streams, the conflict points in the intersection of these lanes are entirely contained in these two movement streams. Denote by L_1 and L_2 the set of lanes in the first and second movement stream, respectively. Define $x = \max_{l\in L_1} \left\lceil \frac{\alpha^*(\frac{U_l}{v_{\max}}+d_l h_f)(k-\Delta+1)}{H} \right\rceil$ and $y = \max_{l\in L_2} \left\lceil \frac{\alpha^*(\frac{U_l}{v_{\max}}+d_l h_f)(k-\Delta+1)}{H} \right\rceil$. Also, note that as each node i in the restricted region satisfies constraint 2.1b and any two lanes in L_1 and L_2 intersect, $x + y \leq k$. Therefore, we can remove the edges connected to nodes inside the restricted regions for all k $\{1,2\}$-factors and replace those edges with x copies of L_1 and y copies of L_2 inside the restricted region. We briefly elaborate on the detailed procedure for modifying the decomposition algorithm to respect the restricted regions in Algorithm 4 presented in Appendix A.25.

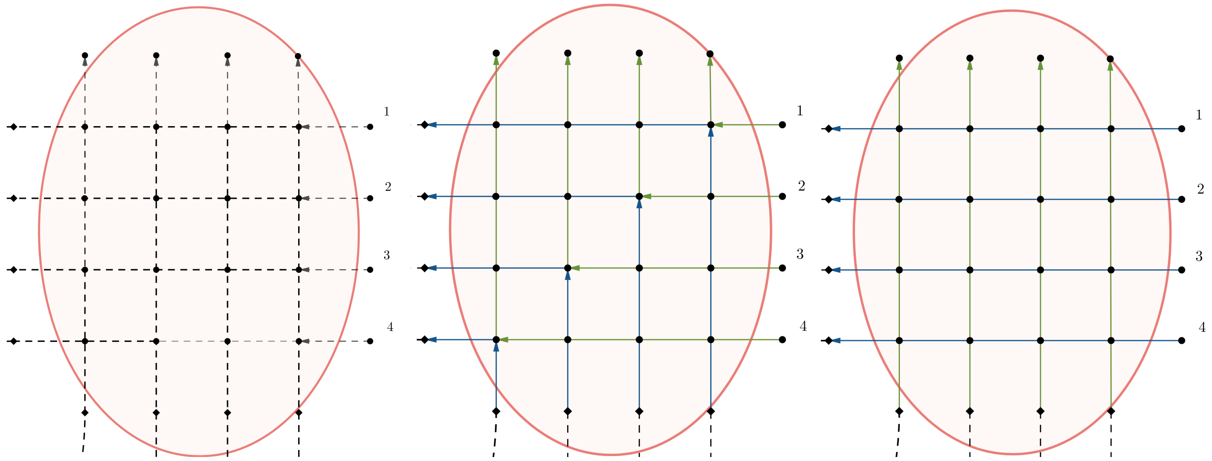


Figure 2.8: Step 1

Figure 2.9: Step 2

Figure 2.10: Steps 3 and 4

Figure 2.11: Demonstration of the steps in Algorithm 4

A heuristic remedy to address the space limitation issue is to release an inscribed traffic unit with length u' within the design traffic unit with length u. Therefore, we can add/remove a few tracking points along each lane in the GCC design. In particular, the addition/removal of a tracking point along a lane moves the inscribed traffic unit one time step backward/forward within the design traffic unit. Therefore, as long as the number of additions/removals along each lane does not exceed $u - u'$, the design remains valid.

Another greedy heuristic approach to address this issue is to avoid releasing one traffic unit at all entrance times associated with a lane during each cycle of length ku. In this case, we can still utilize GCC while regulating right-of-way allocation to avoid any collision. In particular, for some lanes we can release a traffic unit every two cycles at the associated entrance times for them.

In the last part of this section, Theorem 9 finds an upper bound on the delay each vehicle experiences while crossing the intersection by analyzing the number of tracking points added along each lane.

Theorem 9. *The required additional tracking points along each lane to accommodate an optimal GCC linearly depends on the number of movement streams.*

Proof. Proof. See Appendix A.31.

□

To realize the implication of Theorem 9, we call a stationary vehicle arrival admissible if and only if the demand for each lane l during the study horizon H does not exceed the maximum number of traffic units admitted by GCC, calculated in Theorem 8. Proposition 1 presents an upper bound on the delay each vehicle experiences under GCC when the vehicle arrival is admissible.

Proposition 1. *our scheme ensures that under stationary, admissible vehicle arrivals, the delay each vehicle experiences is less than a constant that linearly depends on the number of movement streams and the inverse of the approximation algorithm's precision factor.*

Proof. Proof. The proof follows directly from Theorem 9 and the fact that no queues will be formed under an stationary admissible vehicle arrival.

□

We refer interested readers to Appendix A.27 for an Illustrative Example of the general GCC.

2.5 Relationships Between GCC and Other Controls

GCC provides a general framework that allows us to derive a unifying insight on several controls proposed for intersection management. In this section, we present the traditional signal control or TSC as a special case of GCC and compare the performance of optimal GCC and TSC. We also present rhythmic control as a special case of GCC with parameters $k = 2$ and $l = 1$. Moreover, we present several reservation based-schemes as special cases of GCC. For brevity, we include the detailed discussions on rhythmic control and reservation-based schemes in Appendix A.30.

In Theorem 10, we show that for any given TSC, there exists a GCC with the intersection capacity at least as high as an intersection capacity of the given TSC. The idea in Theorem 10 is to replace the demand decomposition phase of GCC with that of a signalized intersection and find the parameters in Algorithm 2, k and u, so that the resulted GCC can ensure to replicate a similar performance as the considered TSC.

Theorem 10. *For any given TSC, we can modify the demand decomposition phase in Algorithm 2 to construct a corresponding solution to GCC control with intersection reserve capacity at least as high as that of the given TSC.*

Proof. Proof. See Appendix A.28.

□

Next, we compare the throughput of an optimized GCC and an optimized TSC, consider an R-way intersection where the number of lanes for each movement stream is proportionate to its associated demand. Theorem 11 quantifies the improvement we obtain from implementing an optimal GCC over an optimal TSC.

Theorem 11. *For an R-way intersection with sufficient footprint to implement GCC, under a balanced demand pattern, an optimal GCC multiplies the throughput of the intersection by the factor $\frac{R}{2}$ over an optimal TSC.*

Proof. Proof. See Appendix A.29.

□

Now, we showcase the improvement for a generic demand pattern in a 4-way intersection. As the right-turn lanes do not directly intersect with other movement streams, we focus on left turn and through movement streams. First, we note that we can group the intersection lanes into 8 through and left-turn movement streams. Without loss of generality, we assume the demand for all lanes belonging to the same movement stream is the same since they share the same in- and out-bound approach. Note that at each stage of the signal cycle, we may allocate right-of-way to at most two of the eight streams. Denote by M the set of movement streams, and by M_1 the set of pairs of non-intersecting movement streams.

Then, neglecting the loss time between signal stages, we can find the maximum reserve capacity of a TSC by solving the following linear programming:

$$\begin{aligned} \text{Max} \quad & \alpha && && \text{(2.2a)} \\ \text{s.t.} \quad & \sum_{i,j \in M_1} (H\eta_{ij}) \mathbf{1}_{i,j,l} \geq \alpha\left(\frac{U_l}{v_{\max}} + d_l h_f\right); && \forall l \in M && \text{(2.2b)} \\ & \sum_{i,j \in M_1} \eta_{ij} = 1; && && \text{(2.2c)} \end{aligned}$$

where η_{ij} represents the fraction of a cycle that right-of-way allocation is dedicated to a stage with streams i and j. Also, $\mathbf{1}_{i,j,l}$ equals one if $l \in \{i, j\}$; i.e., stream l is either stream i or j. The constraint set 2.2b ensures the served demand from each stream l is at most the sum of admitted demand in stages that include this stream. The constraint set 2.2c ensures the cycle length equals the sum of time duration allocated to the set of stages.

Although the number of lanes within each movement stream is usually designed proportionate to the demand for that stream, the problem's day-to-day dynamic nature prevents us from achieving this 100% improvement [Yin, 2008]. Indeed, if the average demand per lane varies within 50% to 150% of a target demand per lane, we can utilize Theorem 8 and the solution to Problem 2.2 to compute the improvement from TSC to GCC if the footprint of intersection is sufficiently large. The results are shown in Figure 2.12.

2.6 Simulation Comparison

This section presents simulation results to demonstrate the performance of GCC in various scenarios. The benchmarks include TSC and RC. We consider a 4-way intersection with three through

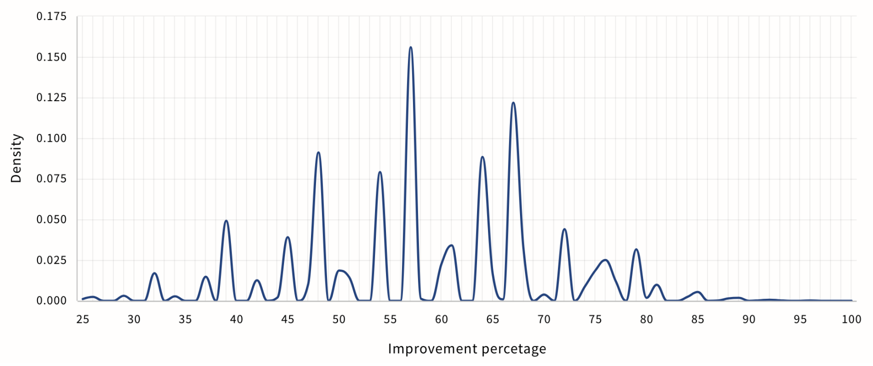


Figure 2.12: Reserve capacity improvement of GCC in comparison with TSC

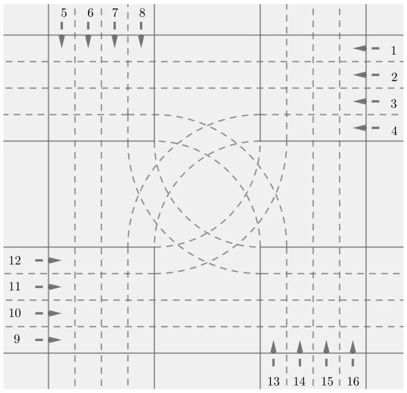


Figure 2.13: GCC Intersection Layout

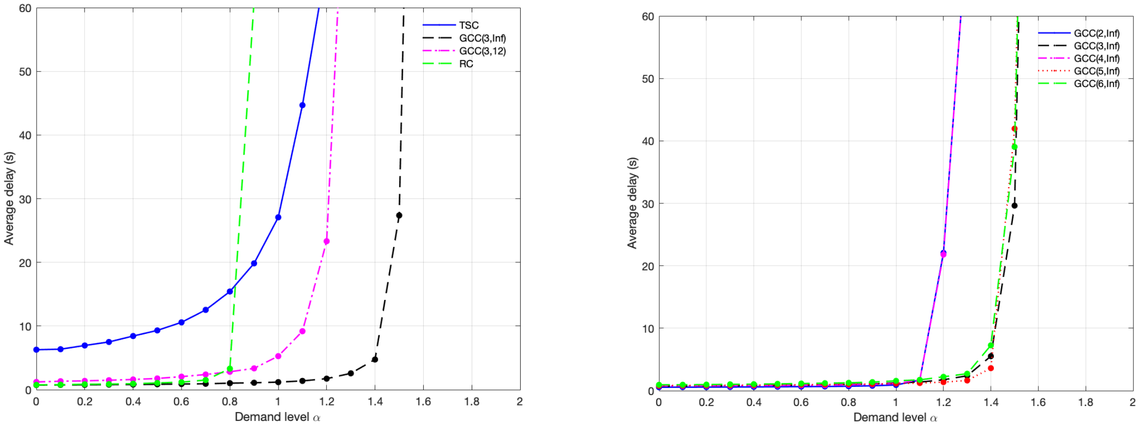
lanes and two left turn lanes for each approach adopted from Chen et al. [2021], under a demand pattern αd where α is the demand level and d is the demand vector as follows:

$$d = [2600, 1400, 1400, 1400, 400, 400, 400, 400]$$

Here, the first and the last four elements represent the per-lane demand for the four through approaches and the four left-turn approaches, respectively.

While we implement TSC and RC on the 20-lane intersection, we integrate the demand for the two left-turn lanes into a single left-turn lane to shrink the required footprint for the GCC layout into 16 lanes in GCC implementation. In this new design, demonstrated in Figure 2.13, the required footprint is less than the footprint of a typical intersection designed based on the AASHTO Green Book [Hancock and Wright, 2013]. We refer the readers to appendix A.32 for a comprehensive description of the simulation settings.

Denote by $GCC(k, u)$ a GCC with parameters k and u. Also, denote by $GCC(k, Inf)$ a GCC with parameter k and $u = 20$ implemented on a sufficiently large intersection that can accommo-



(a) Comparison between RC, TSC, GCC with space limitation and GCC with no space limitation

(b) Sensitivity Analysis for GCC on the parameter k

Figure 2.14: Sensitivity analysis results

date all additional tracking points. First, we present the results of GCC design comparison with the other controls in Figure 2.14a and then we perform sensitivity analysis on the number of colors used in GCC. We observe that performance of GCC does not monotonically increase with the number of colors for some demand patterns, while the worst case performance guarantee for the GCC provided by Theorem 7 improves with the number of colors k. We refer the readers to the supporting materials (2.7.2) for a video demonstrating the operations of the graph coloring control and traffic signal control. Note that at each moment the color of a vehicle indicates the $\{1, 2\}$-factor associated with the movement of its traffic unit on $\tilde{G}$.

2.7 Concluding Remarks

2.7.1 Summary

In this paper, we have investigated the problem of designing an efficient intersection control in a fully CAV environment that respects safety and kinodynamic constraints. The main three results of the paper are as follows: 1) for an intersection with a sufficiently large footprint, we prove that our algorithm provides a PTAS for the throughput maximization problem; 2) with stationary admissible vehicle arrivals, our scheme ensures the delay each vehicle experiences is less than a constant value that linearly depends on the number of movement streams and the inverse of the approximation algorithm's precision factor; and 3) we prove in an R-way intersection where the number of lanes for each movement stream is proportionate to its associated demand, our scheme multiplies the reserve capacity of the intersection by the factor $\frac{R}{2}$ over an optimal TSC. Furthermore, the proposed algorithm is robust in handling a set of vehicles with heterogeneous dimensions and addressing the online variations in safety measures.

The results presented in this paper suggest that an optimal intersection layout design may not follow the traditional way of designing at-grade intersections. For example, the opposite left turns may not be required to avoid each other to maximize the intersection throughput. This suggests that the intersection layout design must be adapted to the CAV environment. The layout design is a strongly complicated problem that we aim to investigate in our future research. Additionally, we show that GCC achieves the reserve capacity of our headway-based LP relaxation. While we can formulate a similar LP relaxation to define the reserve capacity of a network with multiple intersections, when applying GCC to network control, the resilience aspects remains a challenging question we aim to address in our future research.

2.7.2 GCC for General Job Shop Problems with Family Dependent Setup Times

It is worth mentioning that the application of GCC is not limited to intersection control, which bears similarity with mobile robot scheduling [Lozano-Perez, 1983, Siméon et al., 2002, Altché et al., 2016], automated guided vehicle traffic control [Evers and Koppers, 1996, Lombard et al., 2016] and constant delay lattice train schedules [De Carufel et al., 2021]. In fact, Algorithm 2 provides a PTAS to solve a more general job-shop problem. Here, we discuss the connection between these two problems and how we can formulate the problem of finding an optimal traffic control as an instance of job-shop scheduling.

In a general job-shop scheduling, we are given n jobs $J_1, J_2, \ldots, J_n$ of varying processing times, which need to be scheduled on m machines with varying processing power, while trying to optimize an objective function, e.g., the total length of the schedule, i.e., the time for all the jobs processed). In the specific variant known as job-shop scheduling, each job consists of a set of operations $O_1, O_2, ..., O_n$ that need to be processed in a specific order. Each operation has a specific machine that it needs to be processed on, and only one operation in a job can be processed at a given time.

Algorithm 2 can be applied to a more general scheduling problem. Using the scheduling well-known three field notation [Graham et al., 1979], the general job shop problem we are interested in is $Jm|fmls, s, p_j|C_{max}$. In this instance of the job shop problem, the n jobs belong to F different job families. Jobs from the same family can be processed on a machine one after another without requiring any setup time in between. However, if the machine switches over from one family to another, say from family g to family h, then a setup time equal to s is required. The objective C_{max} to minimize is to the completion time of the last job to leave the system. According to the survey study Allahverdi et al. [2008], most of the proposed solution methodologies for sequence-dependent machine-scheduling problems are based on heuristics- such as tabu search and simulated annealing. That being said, even if we relax the sequence dependent setup time constraints, most of the proposed exact algorithms for different variants of the job shop problem are mixed-integer linear programming formulations [Lamorgese and Mannino, 2015, Lamorgese et al., 2016, Lamorgese and Mannino, 2019]. Although these formulations provide a high quality solution for several applications, they are inherently unscalable when faced a high traffic application such as designing an intersection control. This urges the need for investigating more scalable approaches.

Now, we discuss the connection between the two problems. First note that, the passage of vehicles over conflict points of an intersection can be viewed as the jobs in the job shop problem. The families of jobs are the vehicles crossing the intersection using the same lane. The processing time of a job is denoted as the time length of the vehicle with addition of the following headway h_f. Besides, the difference between following and crossing headway is associated with the setup time required to switch from one family to another one. Also, note that in the intersection control problem the vehicles from the same lane while passing over the same segment between two conflict points have to respect the first come first served principle. However, since in the general job-shop instance $Jm|fmls, s, p_j|C_{max}$ the queues between consecutive machines are not required to respect the first come first served principle, we do not have to restrict our approach to the job shop environment satisfying the simple intersection assumption stated in definition 1.

Assuming GCC can provide a θ-approximation, $\theta \leq 1$, for maximizing the reserve capacity problem, we can modify GCC to provides a $\frac{1+\epsilon}{\theta}$-approximation for minimizing the makespan problem. To do so, set the time horizon length in Problem 2.1 as $H = H_{opt}$ where H_{opt} is defined as the optimal makespan and realize that α^*, the optimal solution to Problem 2.1, satisfies $\alpha^* \geq 1$. Now, repeat the optimal GCC for $\left\lceil \frac{H_{opt}(1+\epsilon)}{ku} \right\rceil$ cycles of length ku. Theorem 8 yields that the reserve capacity of the intersection is at least $\frac{\alpha^*}{(1+\epsilon)}$. Therefore, during the time $[0, \left\lceil \frac{H_{opt}(1+\epsilon)}{ku} \right\rceil ku]$ we admit at least $\frac{\alpha^*(1+\epsilon)}{(1+\epsilon)} \geq 1$ multiples of demand while following GCC. This concludes the proof.

Supporting Materials

- This illustrative video demonstrates the operations of dimensionless GCC.
- This illustrative video demonstrates the operations of general GCC, as well as the queue

process comparison with TSC.

Appendix A

Appendix

In this Appendix, we provide the detailed proofs for all the results presented in the main body of the paper. For clarity, some proofs that are purely algebraic are postponed to Section A.13.

A.1 Proof of Lemma 1

Consider an optimal solution to the optimization problem in (1.3) and (1.4). First note that if for one $ji \in \mathcal{I}$ we have $\lambda_{ji} = \mu_{ji} f_{ji}$, then equations (1.2b) and (1.2c) ensure that for all $ji \in \mathcal{I}$ we have $\lambda_{ji} = \mu_{ji} f_{ji}$. Therefore, it suffices to consider the case $\lambda_{ji} < \mu_{ji} f_{ji}, \forall ji \in \mathcal{I}$. Set

$$\epsilon_j = \min_{i, ji \in \mathcal{I}} \frac{\mu_{ji} f_{ji} - \lambda_{ji}}{\lambda_{ji}}.$$

Then, substituting $\lambda'_{ji} = \lambda_{ji}(1 - \epsilon_j)$ we obtain a new solution to the optimization problem in (1.3) and (1.4). Note that according to (1.2e), β_j is decreased and the new solution has a higher objective value.

A.2 General case with some with or all $f_i = 0$

We prove that the problem with some or all $f_j = 0$, that is discussed in Remark 1, can be converted to another network problem with all f_j values positive.

First, if for at least one station $j \in \mathcal{S}$ we have $f_j > 0$. Then, we can adjust the solution to the optimization problem (1.3) and (1.4) without decreasing its objective value by substituting

$$f'_j = \frac{\sum_{j \in \mathcal{S}} f_j}{|\mathcal{S}|}, \quad \forall j \in \mathcal{S}.$$

Therefore, to conclude the discussion on the feasibility of our policy it is sufficient to consider the case where $f_j = 0, \forall j \in \mathcal{S}$. In the following, we define a secondary system and show that how the system dynamics in the original system correspond to the system dynamics in a related secondary system with finite $\gamma_j, \forall j \in \mathcal{S}$. In particular, for this transformation we allow the vehicles to match with the upcoming passengers waiting at their destination when they are close to it.

Next, consider each link $ji \in \mathcal{I}$ and denote by $\tilde{t}_{ji}$ the random variable that represents its travel time. In Lemma 33 we prove that there exists an exponential random variable $\hat{t} \sim \exp \tau$ such that for each link $ji \in \mathcal{I}$, there exists a non-negative random variable $\tilde{t'}_{ji}$ that satisfies

$$\tilde{t}_{ji} = \tilde{t'}_{ji} + \hat{t}.$$

Then, divide each trip along a link $ji \in \mathcal{I}$ into two parts such that the first part is distributed according to $\tilde{t'}_{ji}$ and the second part is distributed according to $\hat{t}$. Note that the second part for all trips has the same distribution.

Now, let the vehicles in the base system travelling on a link $ji \in \mathcal{I}$ to match with upcoming passengers at station i when they are within $\hat{t}$ time to arrive at i. Consider a secondary system with stations $\mathcal{S}', \mathcal{I}'_1, \mathcal{I}'_2$. For each station $i \in \mathcal{S}$ we have an associated station in the secondary system that, with abuse of notation, we show by $i' \in \mathcal{S}'$. Also, for each station $ji \in \mathcal{I}$ we have two associated station in the secondary system that, with abuse of notation, we show by $j_1^{'}i_1^{'} \in \mathcal{I}'_1$ and $j_2^{'}i_2^{'} \in \mathcal{I}'_2$. Here, $j' \in \mathcal{S}'$ in the secondary system represents a vehicle in the base system that completed the first part of its trip toward station j and is currently in the second part of its trip. Also, $j_1^{'}i_1^{'}$ in the secondary system demonstrates a full vehicle in the base system assigned to a passenger at j' with destination i' and is either in the second part of its trip toward j' or arrived at j' and started the first part of its travel toward i'. Also, $j_2^{'}i_2^{'}$ in the secondary system demonstrates an empty vehicle in the base system that is relocating in the first part of its travel toward i'. Therefore, dwell times at stations are $\hat{t}$. The travel times for infinite servers $j_1^{'}i_1^{'} \in \mathcal{I}'_1$ are $\tilde{t}_{ji}$ and travel times for infinite servers $j_2^{'}i_2^{'} \in \mathcal{I}'_2$ are $\tilde{t'}_{ji}$.

The two system are related in the following sense: a vehicle at station $i' \in \mathcal{S}'$ in the secondary system represents a vehicle in the base system that completed the first part of its trip toward station i and is currently in the second part of its trip. An empty vehicle at station $j_2'i_2' \in \mathcal{I}'_2$ in the secondary system represents a vehicle in the base system that is empty relocating in the first part of its travel toward i'. Also, a full vehicle at station $j_1'i_1' \in \mathcal{I}'_1$ in the secondary system represents a vehicle in the base system that is assigned to a passenger at j' with destination i' and either in the second part of its trip toward j' or arrived at j' and started its travel toward i'. Similarly we can describe the associated system dynamics for the two systems. In particular, for each vehicle currently at station $i' \in \mathcal{S}'$ two cases may happen. If it will match with an upcoming passenger within $\hat{t}$ time it takes its associated vehicle to arrive at i, then it picks up the passenger from i' and travels out of i'. Otherwise, the vehicle departs i immediately after arriving. Therefore, the system dynamics in the base system when we allow the vehicles to match with upcoming passengers within $\hat{t}$ time to arrive at their destination corresponds to a secondary system with $\gamma_j = \tau, \forall j \in \mathcal{S}'$. As $\boldsymbol{\gamma} > 0$, the results of Theorems 1, 2, and 3 are valid for this secondary system.

A.3 Relaxing the exponential travel times

To model the travel times drawn from general distributions, we first note that mixtures of Erlang distributions are dense among all distributions. Next, consider a link $ji \in \mathcal{I}$ in the original network, and denote its original travel time distribution by $\tilde{t}_{ji}$. As mixtures of Erlang distribution yields are dense, there exists a set $(k_l, \rho_l); l \in L$ such that

$$\left\|\tilde{t}_{ji} - \sum_{l \in L} Erlang(k_l, \rho_l)\right\| \leq \frac{\epsilon}{2}$$

We state the proof for the case where for all $l \in L$ we have $k_l \equiv 0(mod2)$. Then, we show how to extend the proof to the case where the values k_l are odd numbers. To do so, replace each link $ji \in \mathcal{I}$ in the original network with L parallel paths $ji(l); l \in L$. Then along each path $ji(l)$ we present $k_l/2$ double-ended queueing stations $ji(l,r); 1 \leq r \leq k_l/2$. such that for all stations we have $\lambda_{ji(l,r)} = 0$ and $\gamma_{ji(l,r)} = \rho_l$. Moreover, each two consecutive stations $ji(l,r)$ and $ji(l,r+1)$ are connected with an infinite server link with travel time $exp(\rho_l)$. Finally, to extend the results to the case where we have $k_l \equiv 1(mod2)$, we perform a similar procedure for $k_l + 1$ with the exception that we set the travel time on the last infinite server link on each path that is connected to i by $\exp(\epsilon/2)$. Finally, note that Assumptions 1 remains valid under this transformation. In particular, this transformation replaces the links in the associated directed graph of $\tilde{\boldsymbol{Q}}$ by a directed sub network. Thus, under this transformation $\tilde{\boldsymbol{Q}}$ remains irreducible.

A.4 Rewriting system dynamic equations

For all infinite servers that model full vehicle trips from node j to node i, define

$$\begin{aligned}
\tilde{X}_{ji}^{(N)}(t) \quad = \quad & -\tilde{F}_{ji}(\mu_{ji}\int_0^t X_{ji}^{(N)}(s)ds) + \tilde{\phi}_{ji}((\sum_{kj \in \mathcal{I}} F_{kj}(\mu_{kj}\int_0^t X_{kj}^{(N)}(s)ds)) + (\sum_{kj \in \mathcal{I}} E_{kj}(\mu_{kj}\int_0^t Z_{kj}^{(N)}(s)ds)) \\
& + X_j^{(N)+}(0) - X_j^{(N)+}(t) - H_j(\gamma_j \int_0^t X_j^{(N)+}(s)ds)) + p_{ji}(\sum_{kj \in \mathcal{I}} \tilde{F}_{kj}(\mu_{kj}\int_0^t X_{kj}^{(N)}(s)ds)) \\
& + p_{ji}(\sum_{kj \in \mathcal{I}} \tilde{E}_{kj}(\mu_{kj}\int_0^t Z_{kj}^{(N)}(s)ds)) - p_{ji}\tilde{H}_j(\gamma_j \int_0^t X_j^{(N)+}(s)ds), \qquad ji \in \mathcal{I}
\end{aligned}$$

And all infinite servers that model empty vehicle trips from node j to node i, let

$$\begin{aligned}\tilde{Z}_{ji}^{(N)}(t) &= -\tilde{E}_{ji}(\mu_{ji}\int_0^t Z_{ji}^{(N)}(s)ds) + \tilde{\sigma}_{ji}(H_j(\gamma_j\int_0^t X_j^{(N)+}(s)ds)) \\ &\quad +q_{ji}\tilde{H}_j(\gamma_j\int_0^t X_j^{(N)+}(s)ds), \qquad ji \in \mathcal{I}\end{aligned}$$

Then, we can rewrite the set of equations that define system dynamics by differentiating the stochastic and deterministic terms for single servers as follows

$$\begin{aligned}\frac{X_i^{(N)}(t)}{N} &= \frac{X_i^{(N)}(0)}{N} + \frac{\tilde{X}_i^{(N)}(t)}{N} - (\lambda_i t) + \sum_{ji\in\mathcal{I}}\mu_{ji}\int_0^t \frac{X_{ji}^{(N)}(s)}{N}ds + \sum_{ji\in\mathcal{I}}\mu_{ji}\int_0^t \frac{Z_{ji}^{(N)}(s)}{N}ds \\ &\quad +\theta_i\int_0^t \frac{X_i^{(N)-}(s)}{N}ds - \gamma_i\int_0^t \frac{X_i^{(N)+}(s)}{N}ds, \qquad i \in \mathcal{S} \quad (A.1) \\ \frac{X_{ji}^{(N)}(t)}{N} &= \frac{X_{ji}^{(N)}(0)}{N} + \frac{\tilde{X}_{ji}^{(N)}(t)}{N} - \mu_{ji}\int_0^t \frac{X_{ji}^{(N)}(s)}{N}ds \\ &\quad +p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}\int_0^t \frac{X_{kj}^{(N)}(s)}{N}ds + p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}\int_0^t \frac{Z_{kj}^{(N)}(s)}{N}ds \\ &\quad +p_{ji}\frac{X_j^{(N)+}(0)}{N} - p_{ji}\frac{X_j^{(N)+}(t)}{N} - p_{ji}\gamma_j\int_0^t \frac{X_j^{(N)+}(s)}{N}ds, \qquad ji \in \mathcal{I} \quad (A.2) \\ \frac{Z_{ji}^{(N)}(t)}{N} &= \frac{Z_{ji}^{(N)}(0)}{N} + \frac{\tilde{Z}_{ji}^{(N)}(t)}{N} - \mu_{ji}\int_0^t \frac{Z_{ji}^{(N)}(s)}{N}ds + q_{ji}\gamma_j\int_0^t \frac{X_j^{(N)+}(s)}{N}ds, \qquad ji \in \mathcal{I} \quad (A.3)\end{aligned}$$

A.5 Proof of Lemma 2

In this section, we first rewrite the system dynamic equations in matrix format and then establish existence, uniqueness, and Lipschitz continuity of a fluid limit to the resulting dynamical system for any given initial system state.

To start, we rewrite the system of equations (1.16-1.18) in matrix format by defining three matrices J_1, J_2, and J_3 as square matrices with size $|\mathcal{S}| + 2|\mathcal{I}|$. To facilitate the illustration we use $\mathcal{S}$, $\mathcal{I}_1$ and $\mathcal{I}_2$ to refer to the state of the system at the stations, the full vehicles traveling on infinite servers, and empty vehicles traveling on infinite servers, respectively. To define each matrix, set $i \in \mathcal{S}$, $ji \in \mathcal{I}_1$ and $kl \in \mathcal{I}_2$ to indicate the rows of these matrices associated to the single servers, infinite servers representing full vehicles travel and infinite servers representing empty vehicles travel, respectively. In particular, we define the elements of matrix J_1 as follows

$$J_1(i,v) = \begin{cases} -\gamma_i, & v = i \in \mathcal{S} \\ \mu_{ji}, & v = ji \in \mathcal{I}_1 \\ \mu_{ji}, & v = ji \in \mathcal{I}_2 \\ 0, & otherwise \end{cases} \quad J_1(ji,v) = \begin{cases} -\mu_{ji}, & j \neq i \& v = ji \\ p_{ii}\mu_{ii} - \mu_{ii}, & j = i \& v = ii \in \mathcal{I}_1 \\ p_{ji}\mu_{mj}, & v = mj \neq ji \in \mathcal{I}_1 \\ p_{ji}\mu_{mj}, & v = mj \in \mathcal{I}_2 \\ 0, & otherwise \end{cases}$$

$$J_1(kl,v) = \begin{cases} q_{kl}\gamma_k, & v = k \in \mathcal{S} \\ -\mu_{kl}, & v = kl \in \mathcal{I}_2 \\ 0, & otherwise \end{cases}$$

Next, we define matrix J_2 as follows

$$J_2(i,v) = \begin{cases} \theta_i - \gamma_i, & v = i \in \mathcal{S} \\ q_{kl}\gamma_k, & v = kl \in \mathcal{I}_2 \\ 0, & otherwise \end{cases} \quad J_2(ji,v) = 0 \quad J_2(kl,v) = 0$$

Lastly, we define the matrix J_3 as follows

$$J_3(i,v) = 0 \quad J_3(ji,v) = \begin{cases} p_{ji}, & v = ji \\ 0, & otherwise \end{cases} \quad J_3(kl,v) = 0$$

Besides, define the time dependent vector $q(t)$ with size $|\mathcal{S}| + |\mathcal{I}_1| + |\mathcal{I}_2|$ as follows

$$q_i(t) = x_i(0) - \lambda_i t \quad q_{ji}(t) = x_{ji}(0) + p_{ji}x_j^+(0) \quad q_{kl}(t) = x_{kl}(0)$$

Therefore, we can rewrite the dynamical system (1.16-1.18) in matrix format as follows

$$x(t) = q(t) + J_1 \int_0^t x(s)ds + J_2 \int_0^t x^-(s)ds - J_3 x^+(t) \tag{A.4}$$

Now, define the vector of variables $y(t) = x(t) + J_3 x^+(t)$, and note that $y^-(t) = x^-(t)$. Set J_4 as follows

$$J_4(i,v) = 0 \quad J_4(ji,v) = \begin{cases} \frac{1}{1+p_{ji}}, & v = ji \\ 0, & otherwise \end{cases} \quad J_4(kl,v) = 0$$

Next, observe that $x(t) = J_4 y^+(t) + y^-(t) = J_4 y(t) + (I - J_4)y^-(t)$. Then, we can rewrite (A.4) as follows

$$y(t) = q(t) + J_1 J_4 \int_0^t y(s)ds + (J_1(I - J_4) + J_2) \int_0^t y^-(s)ds$$

Therefore, we can define matrices J_5 and J_6 to write the system (A.4) as follows

$$y(t) = q(t) + J_5 \int_0^t y(s)ds + J_6 \int_0^t y^-(s)ds \tag{A.5}$$

Therefore, it suffices to prove existence uniqueness and liptictiz continuity for $y(t)$ the solution to the system of (A.5). First, we prove that there exists a solution to the dynamical system (A.5). Consider $y^0(t) = 0$, then define $y^{n+1}(t) = q(t)+J_5 \int_0^t y^n(s)ds+J_6 \int_0^t y^{n-}(s)(s)ds$. Next, compute the consecutive differences $\left\|y^{n+1} - y^n\right\|_t$ as follows

$$\begin{aligned}
\left\|y^{n+1} - y^n\right\|_t &= J_5 \int_0^t (y^{n+1} - y^n)(s_1)ds_1 + J_6 \int_0^t (y^{n+1-} - y^{n-})(s_1)ds_1 \\
&\leq |J_5| \int_0^t \left\|y^n - y^{n-1}\right\|_{s_1} ds_1 + |J_6| \int_0^t \left\|y^{n+1-} - y^{n-}\right\|_{s_1} ds_1
\end{aligned}$$

As we have $\left\|y^{n+1-} - y^{n-}\right\|_{s_1} \leq \left\|y^n - y^{n-1}\right\|_{s_1}$. If we define $C = \|J_5\| + \|J_6\|$ we have

$$\begin{aligned}
\left\|y^{n+1} - y^n\right\|_t &\leq C \int_0^t \left\|y^n - y^{n-1}\right\|_{s_1} ds_1 \\
&\leq C^2 \int_0^t \int_0^{s_1} \left\|y^{n-1} - y^{n-2}\right\|_{s_2} ds_2 ds_1 \\
&\leq C^n \int_0^t \int_0^{s_1} \dots \int_0^{s_{n-1}} \left\|y^1 - y^0\right\|_{s_n} ds_n \dots ds_1 \\
&\leq C^n \int_0^t \int_0^{s_1} \dots \int_0^{s_{n-1}} \left\|y^1 - y^0\right\|_t ds_n \dots ds_1 \\
&\leq \frac{(C \left\|y^1 - y^0\right\|_t)^n}{n!}
\end{aligned}$$

Thus, for any epsilon there exist a large n such that $\left\|y^{n+1} - y^n\right\|_t \leq \epsilon$. This shows that the sequence of functions $\{y^n(t)\}_{n=1}^\infty$ is a Cauchy sequence in the space of Cadlag functions. As the space of Cadlag functions is complete we have there exist a limit for the sequence $\{y^n(t)\}_{n=1}^\infty$ that is a solution to the dynamical system (13).

To prove uniqueness, we assume there exist two solutions $y(t)$ and $\tilde{y}(t)$ for the same input $q(t)$. Then using the same recursive argument we have

$$\begin{aligned}
\|y - \tilde{y}\|_t &\leq |J_5| \int_0^t \|y - \tilde{y}\|_{s_1} ds_1 + |J_6| \int_0^t \left\|y^- - \tilde{y}^-\right\|_{s_1} ds_1 \\
&\leq C \int_0^t \|y - \tilde{y}\|_{s_1} ds_1 \\
&\leq \frac{(C \|y - \tilde{y}\|_t)^n}{n!}
\end{aligned}$$

Hence, for any epsilon there exists a large n such that $\|y-\tilde{y}\|_t \leq \epsilon$. As such $\tilde{y}(t) = y(t), \forall t \in R^+$. Lipschitz continuity can be established by considering two inputs $q(t)$ and $\tilde{q}(t)$ and their corresponding solutions $y(t)$ and $\tilde{y}(t)$ and applying the same recursive argument as the ones for existence and uniqueness.

$$\begin{aligned}
\|y-\tilde{y}\|_t &\leq \|q-\tilde{q}\|_t + |J_5| \int_0^t \|y-\tilde{y}\|_{s_1} ds_1 + |J_6| \int_0^t \left\|y^- - \tilde{y}^-\right\|_{s_1} ds_1 \\
&\leq \|q-\tilde{q}\|_t + C \int_0^t \|y-\tilde{y}\|_{s_1} ds_1 \\
&\leq \|q-\tilde{q}\|_t \sum_{i=0}^{n-1} (\frac{(Ct)^i}{i!}) + \frac{(Ct\,\|y-\tilde{y}\|_t)^n}{n!}
\end{aligned}$$

Choosing n large enough, we obtain $\|y-\tilde{y}\|_t \leq e^{Ct}\,\|q-\tilde{q}\|_t$.

A.6 Proof of Lemma 7

Lemma 7. *Consider the system dynamics presented in Theorem 1, Then for any* $t \geq 0$ *we have*

$$\lim_{N\to\infty} N^{1/2-\epsilon} \mathbb{E}\left[\left|\frac{X^{(N)}(t)}{N} - x(t)\right|\right] = 0.$$

Proof. According to Lemma 2, we bound the difference between the solution to the dynamical system and the scaled queue length process as follows

$$\lim_{N\to\infty} \mathbb{E}\Big[\Big\|\frac{X_i^N}{N} - x_i\Big\|_T\Big] \leq \lim_{N\to\infty} e^{CT} \mathbb{E}\Big[\Big\|\frac{\tilde{X}^N}{N}\Big\|_T\Big]$$

Hence, it suffices to prove that for some $\epsilon_1 > 0$ we have

$$\lim_{N\to\infty} \mathbb{E}\Big[\Big\|\frac{\tilde{X}^N}{N}\Big\|_T\Big] = O(N^{-1/2+\epsilon_1}) \tag{A.6}$$

Considering the linearity of expectation, to prove the lemma statement it suffices to prove the following three statements

- $\lim_{N\to\infty} \mathbb{E}[\|\frac{\tilde{X}_i^N}{N}\|_T] = O(N^{-1/2+\epsilon_1});$ $\qquad \forall i \in \mathcal{S}$
- $\lim_{N\to\infty} \mathbb{E}[\|\frac{\tilde{X}_{ji}^N}{N}\|_T] = O(N^{-1/2+\epsilon_1});$ $\qquad \forall ji \in \mathcal{I}$
- $\lim_{N\to\infty} \mathbb{E}[\|\frac{\tilde{Z}_{kl}^N}{N}\|_T] = O(N^{-1/2+\epsilon_1});$ $\qquad \forall kl \in \mathcal{I}$

First, for all single servers i the linearity of expectation yields

$$\begin{aligned}\mathbb{E}\Big[\|\tilde{X}_i^N\|_T\Big] \;&\leq\; \mathbb{E}\Big[\sup_{0\leq t\leq T}\Big|\tilde{A}_i(N\lambda_i t)\Big|\Big] + \sum_{ji\in\mathcal{I}}\mathbb{E}\Big[\sup_{0\leq t\leq T}\Big|\tilde{F}_{ji}(\mu_{ji}\int_0^t X_{ji}^{(N)}(s)ds)\Big|\Big]\\ &+\sum_{ji\in\mathcal{I}}\mathbb{E}\Big[\sup_{0\leq t\leq T}\Big|\tilde{E}_{ji}(\mu_{ji}\int_0^t Z_{ji}^{(N)}(s)ds)\Big|\Big] + \mathbb{E}\Big[\sup_{0\leq t\leq T}\Big|\tilde{H}_i(\gamma_i\int_0^t X_i^{(N)+}(s)ds)\Big|\Big]\\ &+\mathbb{E}\Big[\sup_{0\leq t\leq T}\Big|\tilde{G}_i(\theta_i\int_0^t X_i^{(N)-}(s)ds)\Big|\Big]\end{aligned}$$

Next, we note that $X_j^{(N)+}$, $X_{ji}^{(N)}$, and $Z_{ji}^{(N)}$ are positive and their sum,

$$\sum_{j\in\mathcal{S}} X_j^{(N)+} + \sum_{ji\in\mathcal{I}} X_{ji}^{(N)} + \sum_{ji\in\mathcal{I}} Z_{ji}^{(N)} = N$$

, is invariant over time and equals the market size N. Therefore we can upper bound the above equation as follows:

$$\begin{aligned}\mathbb{E}\Big[\|\tilde{X}_i^N\|_T\Big] \;&\leq\; \mathbb{E}\Big[\sup_{0\leq t\leq T}\Big|\tilde{A}_i(N\lambda_i t)\Big|\Big] + \sum_{ji\in\mathcal{I}}\mathbb{E}\Big[\sup_{0\leq t\leq T}\Big|\tilde{F}_{ji}(\mu_{ji}Nt)\Big|\Big]\\ &\sum_{ji\in\mathcal{I}}\mathbb{E}\Big[\sup_{0\leq t\leq T}\Big|\tilde{E}_{ji}(\mu_{ji}Nt)\Big|\Big] + \mathbb{E}\Big[\sup_{0\leq t\leq T}\Big|\tilde{H}_i(\gamma_i Nt)\Big|\Big]\\ &\mathbb{E}\Big[\sup_{0\leq t\leq T}\Big|\tilde{G}_i(\theta_i\int_0^t X_i^{(N)-}(s)ds)\Big|\Big]\end{aligned}$$

Applying Exercise II.1.16 in Revuz and Yor [2013] to the first four terms yields

$$\mathbb{E}\big[\big\|\tilde{X}_i^N\big\|_T\big] \leq O\Big(\frac{e(1+\mathbb{E}[\tilde{A}_i(N\lambda_i T)log^+(\tilde{A}_i(N\lambda_i T))])}{e}\Big) + \mathbb{E}\Big[\sup_{0\leq t\leq T}\Big|\tilde{G}_i\Big(\theta_i\int_0^t X_i^{(N)-}(s)ds\Big)\Big|\Big]$$

Noting that $X\log X = O(X^{1+\epsilon})$, and $\mathbb{E}[\tilde{A}_i(N\lambda_i T)^{1+\epsilon}] = O(N^{1/2+\epsilon_1})$. We can further upper bound the right hand side as follows

$$\mathbb{E}\big[\big\|\tilde{X}_i^N\big\|_T\big] \leq O(N^{1/2+\epsilon_1}) + \mathbb{E}\Big[\sup_{0\leq t\leq T}\Big|\tilde{G}_i\Big(\theta_i\int_0^t X_i^{(N)-}(s)ds\Big)\Big|\Big]$$

Next, we can upper bound the last term as follows

$$\mathbb{E}\big[\big\|\tilde{X}_i^N\big\|_T\big] \leq O(N^{1/2+\epsilon_1}) + \mathbb{E}\Big[\Big|\tilde{G}_i\Big(\theta_i T\|\tilde{X}_i^N\|_T\Big)^{1+\epsilon}\Big|\Big]$$

Now we can expand the right hand side using the Law of the unconscious statistician (LOTUS) as follows

$$\mathbb{E}[|\tilde{G}_i(\theta_i T\|\tilde{X}_i^N\|_T)^{1+\epsilon}|] = \int_{-\infty}^{\infty} \mathbb{E}[|\tilde{G}_i(s)|^{1+\epsilon}] \quad d\mathbb{P}\left\{\theta_i T\|\tilde{X}_i^N\|_T = s\right\}$$

Therefore,

$$\begin{aligned}\mathbb{E}[|\tilde{G}_i(\theta_i T\|\tilde{X}_i^N\|_T)^{1+\epsilon}|] &= \int_{-\infty}^{\infty} O(s^{1/2+\epsilon}) \quad d\mathbb{P}\left\{\theta_i T\|\tilde{X}_i^N\|_T = s\right\} \\ &\leq O(\mathbb{E}[\|\tilde{X}_i^{(N)}\|_T^{1/2+\epsilon_1}])\end{aligned}$$

where the first inequality also folows from Exercise II.1.16 in Revuz and Yor [2013]. Therefore,

$$\mathbb{E}[\|\tilde{X}_i^N\|_T] = O(N^{1/2+\epsilon_1})$$

Similarly, for infinite servers $ji \in \mathcal{I}$ we write

$$\begin{aligned}& \mathbb{E}[\|\tilde{X}_{ji}^{(N)}\|_T] \\ \leq\ & \mathbb{E}[\sup_{0\leq t\leq T} \|\tilde{F}_{ji}\Big(\mu_{ji}\int_0^t X_{ji}^{(N)}(s)ds\Big)\|] \\ & + \sum_{kj\in\mathcal{I}} \mathbb{E}\Bigg(p_{ji}\sup_{0\leq t\leq T}\Big|\tilde{F}_{kj}\Big(\mu_{kj}\int_0^t X_{kj}^{(N)}(s)ds\Big)\Big|\Bigg) \\ & + \sum_{kj\in\mathcal{I}} \mathbb{E}\Bigg(p_{ji}\sup_{0\leq t\leq T}\Big|\tilde{E}_{kj}\Big(\mu_{kj}\int_0^t Z_{kj}^{(N)}(s)ds\Big)\Big|\Bigg) \\ & +\mathbb{E}\Bigg(\sup_{0\leq t\leq T}\Big|\tilde{\phi}_{ji}\Big(\textstyle\sum_{kj\in\mathcal{I}} F_{kj}\big(\mu_{kj}\int_0^t X_{kj}^{(N)}(s)ds\big) + \sum_{kj\in\mathcal{I}} E_{kj}\big(\mu_{kj}\int_0^t Z_{kj}^{(N)}(s)ds\big) \\ & . + X_j^{(N)+}(0) - X_j^{(N)+}(t) - H_j\big(\gamma_j\int_0^t X_j^{(N)+}(s)ds\big)\Big)\Big|\Bigg)\end{aligned}$$

Similar to the case for single servers we can upper bound the first three terms by $O(N^{1/2+\epsilon_1})$. To upper bound the last term define

$$T_5 = \sum_{kj\in\mathcal{I}}\sup_{0\leq t\leq T} F_{kj}(\mu_{kj}Nt) + \sum_{kj\in\mathcal{I}}\sup_{0\leq t\leq T} E_{kj}(\mu_{kj}Nt) + N - \sup_{0\leq t\leq T} H_j(\gamma_j Nt)$$

Then, we upper bound the last term as $\mathbb{E}\Big(\sup_{0\leq s\leq T_5}|\tilde{\phi}_{ji}(s)|\Big)$. For $0 < \epsilon < \epsilon_1$, this can be further upper bounded by $\mathbb{E}\Big(\|\tilde{\phi}_{ji}\|_{T_5}^{1+\epsilon}\Big)$ Now, we can expand this expression using LOTUS as follows

$$\int_{-\infty}^{\infty} \mathbb{E}\Big(|\tilde{\phi}_{ji}(s)|^{1+\epsilon}\Big) \quad d\mathbb{P}\{T_5 = s\}$$

Thus,

$$\int_{-\infty}^{\infty} s^{1/2+\epsilon} \quad d\mathbb{P}\left\{T_5 = s\right\} ds \;\; = \;\; \mathbb{E}[T_5^{1/2+\epsilon}] \leq O(N^{1/2+\epsilon_1})$$

where the last inequality follows from Exercise II.1.16 in Revuz and Yor [2013]. Therefore similar to single server queues we conclude that

$$\mathbb{E}\big(\|\tilde{X}_{ji}^{(N)}\|_T\big) \;= O(N^{1/2+\epsilon_1})$$

Lastly, for all infinite servers that model empty vehicle trips from node j to node i set

$$\begin{aligned}\mathbb{E}\Big(\|\tilde{Z}_{ji}^{(N)}\|_T\Big) \;\; & \leq \;\; \mathbb{E}\Big(\sup_{0\leq t\leq T}\Big|\tilde{E}_{ji}(\mu_{ji}\int_0^t Z_{ji}^{(N)}(s)ds)\Big|\Big) \\ & \quad +\mathbb{E}\Big(\sup_{0\leq t\leq T}\Big|\tilde{\sigma}_{ji}(H_j(\gamma_j\int_0^t X_j^{(N)+}(s)ds))\Big|\Big) \\ & \quad +\mathbb{E}\Big(\sup_{0\leq t\leq T}\Big|q_{ji}\tilde{H}_j(\gamma_j\int_0^t X_j^{(N)+}(s)ds)\Big|\Big)\end{aligned}$$

Following the same procedure for single servers and infinite servers, it is straight forward to show that for arbitrarily small $\epsilon > 0$ all three terms are bounded by $O(N^{1/2+\epsilon_1})$ This concludes the proof. □

A.7 Proof of Theorem 1

To prepare for the proof of Lemma 16, we first identify the limit for the dynamical system (1.16-1.18). Define a Lyapunov function based on the solution to a Linear Complimentary Problem

(LCP) as follows

$$\sum_i \mu_{ji} u_{ji} + \sum_i \mu_{ji} v_{ji} = \sum_k \mu_{kj} u_{kj} + \sum_k \mu_{kj} v_{kj} \qquad \forall j \in \mathcal{S} \tag{A.7a}$$

$$u_{ji}\mu_{ji} = p_{ji} \sum_k \mu_{jk} u_{jk}, \qquad \forall i, j \in R \tag{A.7b}$$

$$\sum_k \mu_{jk} u_{jk} \leq \lambda_j, \qquad \forall i, j \in R \tag{A.7c}$$

$$(\lambda_j - \sum_k \mu_{jk} u_{jk}) u_j = 0, \qquad \forall i, j \in R \tag{A.7d}$$

$$\mu_{ji} v_{ji} = q_{ji} \gamma_j u_j, \qquad \forall i, j \in R \tag{A.7e}$$

$$\sum_{ji \in \mathcal{I}} u_{ji} + \sum_{ji \in \mathcal{I}} v_{ji} + \sum_{j \in \mathcal{S}} u_j = \Theta \tag{A.7f}$$

$$u_{ji}, v_{ji}, u_j \geq 0, \qquad \forall j, i \in R \tag{A.7g}$$

It is worth noting that the system of equations (A.7) is equivalent to the system of equations (1.2) except that we substitute the unit mass equation with a $\Theta \in [0, 1]$ and replace λ_{ji}/λ_j, where $\lambda_j = \sum_k \lambda_{jk}$, by p_{ji} in LCP (A.7). Intuitively, this corresponds to equilibrium under the same control when the total vehicle flow in the system is reduced from unity 1 to Θ.

In Lemma 8, we will prove that for any given $\Theta > 0$, there exists a unique solution $u(\Theta) = (u_{ji}(\Theta), v_{ji}(\Theta), u_j(\Theta))$ to the system (A.7), and $u_{ji}(\Theta), ji \in \mathcal{I}, v_{ji}(\Theta), ji \in \mathcal{I}, u_j(\Theta), j \in \mathcal{S}$ are continuous piece-wise linear and increasing functions. Moreover, the number of pieces of each function is at most $|\mathcal{S}|$.

A.8 Uniqueness of the solution to the LCP problem

In this section, we establish the uniqueness of the solution for the LCP problem. The result is forally stated below.

Lemma 8. *For $\Theta > 0$, the system (A.7) has a unique solution $u(\Theta) = (u_{ji}(\Theta), v_{ji}(\Theta), u_j(\Theta))$ such that*

- *The unique functions $u_{ji}(\Theta), ji \in \mathcal{I}, v_{ji}(\Theta), ji \in \mathcal{I}, u_j(\Theta), j \in \mathcal{S}$ are continuous piece-wise linear and increasing and number of pieces of each function is at most $|\mathcal{S}|$.*
- *The right limit slope for the functions $u_j(\Theta), j \in \mathcal{S}$ is strictly positive unless $u_j(\Theta) = 0$.*
- *The right limit slope for the functions $v_{ji}(\Theta), ji \in \mathcal{I}$ is strictly positive unless $v_{ji}(\Theta) = 0$.*
- *The right limit slope for the functions $u_{ji}(\Theta), ji \in \mathcal{I}$ is strictly positive unless $u_{ji}(\Theta) \in \{0, u_{ji}(1)\}$.*

- *The minimum nonzero right limit slope for the functions $u_{ji}(\Theta), ji \in \mathcal{I}, v_{ji}(\Theta), ji \in \mathcal{I}, u_j(\Theta), j \in \mathcal{S}$ is ι_1 such that*

$$\iota_1 \geq \frac{1}{|\mathcal{I}| + |\mathcal{S}|} \min \left(\frac{\min_{ji,lm,\lambda_{ji}>0} \lambda_{ji}\mu_{lm}}{\max_{k,lm} \lambda_k \mu_{lm}}, \frac{\min_{ji,k,\lambda_{ji}>0} \lambda_{ji}\gamma_k}{\max_{k,lm} \lambda_k \mu_{lm}}, \right. \\ \left. \frac{\min_{ji,lm,q_{ji}>0} q_{ji}\mu_{lm}}{\max_{lm} \mu_{lm}}, \frac{\min_{ji,k,q_{ji}>0} q_{ji}\gamma_k}{\max_{lm} \mu_{lm}}, \frac{\min_{lm} \mu_{lm}}{\max_j \gamma_j}, 1 \right)$$

- *The maximum nonzero right limit slope for the functions $u_{ji}(\Theta), ji \in \mathcal{I}, v_{ji}(\Theta), ji \in \mathcal{I}, u_j(\Theta), j \in \mathcal{S}$ is ι_2 such that $\iota_2 \leq 1$.*

Definition 5. *Set ι_1 and ι_2 as the minimum nonzero and maximum slopes of the piecewise linear functions $u_{ji}(\Theta), ji \in \mathcal{I}, v_{ji}(\Theta), ji \in \mathcal{I}, u_j(\Theta), j \in \mathcal{S}$.*

Lyapunov Function. Define the Lyapunov function $L(x)$ for a state $x = (x_{ji}(t), z_{ji}(t), x_j(t); ji \in \mathcal{I}, j \in \mathcal{S})$ as follows

$$L(x) = \sup_{\Theta, u(\Theta) \leq x^+(t)} \Theta,$$

where, with some abuse of notation, we use "$\leq$" to also denote component-wise inequality. Let $\Theta^{(i)} = \min(1, \arg\sup_\Theta \{u_i(\Theta) \leq 0\}), i \in \mathcal{S}$. This is the maximum value for Θ such that the mass of vehicles at station i is 0 obtained from solving the LCP (A.7). Without loss of generality, we assume Θ^i is non-decreasing in i, and let $\Theta^{|S|+1} = 1$. To prove the convergence of the solution to dynamical system (1.16-1.18) to the solution of the fluid equilibrium (1.2), we first present two basic properties of the function L in the following lemma.

Lemma 9. *The Lyapunov function $L(x(.))$ satisfies the following monotonicity and drift conditions:*

- *Monotonicity: $L(x(t))$ is increasing in t.*

- *Drift condition: If the unique solution to the dynamical system (1.16-1.18) with initial condition $x(0)$ satisfies $\Theta^i \leq L(x(t)) < \Theta^{i+1}$ and $x_j(t) \geq 0, \forall j \in \mathcal{S}, u_j(\Theta^{i+1}) > 0$, then there exists constants $\iota > 0$ and T_2 such that*

$$L(x(t')) \geq \Theta^{j+1} - (\Theta^{j+1} - L(x(t)))(1 - \iota/4), \forall t' \geq t + T_2.$$

Also, $\iota \geq \iota_1$ is at least greater than the minimum slope of functions $u_{ji}, v_{ji}, ji \in \mathcal{I}$.

Informally speaking, the drift condition ensures the Lyapunov function is continually increasing, and as long as the Lyapunov function at the current state lies between two of the breakpoints in the LCP (A.7), it reaches the next breakpoint in an exponentially short time. Justifying the improvement results when the system state is close to a breakpoint requires a detailed algebraic derivation that we discuss in the proof for Lemma 16.

A.9 Stability of the dynamical system

In this section we prove the stability of the deterministic dynamic system (1.16-1.18). The proof if quite long and involved, hence we divide that into a sequence of lemmas. First, we establish a number of properties for the system (1.16-1.18) in Lemmas 10 to 14. The proof of these lemmas are purely algebraic hence they are provided in Sections A.13 to A.14.4.

Lemma 10. *For any given $t \geq 0$ the solution to the dynamical system (1.16-1.18) starting from initial condition $x_{ji}(0) \geq 0, z_{ji}(0) \geq 0, \forall ji \in \mathcal{I}$ satisfies the following conditions:*

a) The following invariant equation is satisfied:

$$\sum_{ji \in \mathcal{I}} x_{ji}(t) + \sum_{ji \in \mathcal{I}} z_{ji}(t) + \sum_{i \in \mathcal{S}} x_i^+(t) = \sum_{ji \in \mathcal{I}} x_{ji}(0) + \sum_{ji \in \mathcal{I}} z_{ji}(0) + \sum_{i \in \mathcal{S}} x_i^+(0)$$

b) The solution to dynamic system for any $t \geq 0$ satisfies

$$x_{ji}(t) \geq 0, \forall ji \in \mathcal{I}, z_{ji}(t) \geq 0, \forall ji \in \mathcal{I}$$

c) The functions $x_j^+(t)$, $x_{ji}(t)$ and $z_{ji}(t)$ are lipschitz continuous.

d) For the positive constant $C_2 = \max\{x_i^-(0), \lambda_i/\theta_i\}$ we have $x_i^-(t) \geq -C_2, \forall t \geq 0$

Lemma 11. *Decompose the set of single servers $\mathcal{S}$ into two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\lambda_i = 0, \forall i \in \mathcal{S}_1$ and $\lambda_i > 0, \forall i \in \mathcal{S}_2$. Then, for $i \in \mathcal{S}_1$ and the constant C_2 defined in Lemma 10 we have*

a) If $x_i(0) \geq 0$, Then, $x_i(t) \geq 0, \forall t \geq 0$

b) If $x_i(0) < 0$, Then, $x_i(t) \geq -\min(|x_i(0)|, C_2)e^{-\theta_i t}, \forall t \geq 0$

Definition 6. *We denote by $\Psi(x,t); t \geq 0$ the unique solution $x(t); t \geq 0$ to the dynamical system (1.16-1.18) with initial state $x(0)$.*

Lemma 12. *$L(x(t))$ is an increasing function. Moreover, if $L(x(t)) \geq \Theta^i$, then we have $x_i(t') \geq -\min(x_i(0), C_2)e^{-\theta_i(t'-t)}, \forall t' \geq t$. Here, C_2 is a constant defined in Lemma 10.*

Proposition 2. *Assume $x(t) = \Psi(x(0), t)$, the unique solution to the dynamical system (1.16-1.18), satisfies $\Theta^i \leq L(x(t))$. For an arbitrarily small $\delta > 0$ t, There exists a time $T_1 \geq 1$ such that*

$$x_i(t') \geq -\delta, \forall t' \geq t + T_1$$

Proof. Setting $T_1 = \frac{1}{\theta_j} \log(\frac{C_2}{\delta})$ in Lemma 12 leads to the desired result. □

Lemma 13. *Decompose the set of single servers $\mathcal{S}$ into two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\lambda_i = 0, \forall i \in \mathcal{S}_1$ and $\lambda_i > 0, \forall i \in \mathcal{S}_2$. Consider two solutions to the dynamical system (1.16-1.18) with different initial conditions:*

i) $x(t) = (x_i(t), x_{ji}(t), z_{kl}(t), i \in \mathcal{S}, ji \in \mathcal{I}, kl \in \mathcal{I})$, with initial condition $x(0)$.

ii) $y(t) = (y_i(t), y_{ji}(t), w_{kl}(t), i \in \mathcal{S}, ji \in \mathcal{I}, kl \in \mathcal{I})$, with initial condition $y(0)$.

Such that $\sum_{ji \in \mathcal{I}} y_{ji}(t) + \sum_{ji \in \mathcal{I}} w_{ji}(t) + \sum_{i \in \mathcal{S}} y_i^+(t) \leq \min_{j \in \mathcal{S}_2} \lambda_j/2$, and $y^+(0) < x^+(0)$. Further, assume that $y_i(t) \leq 0, \forall i \in \mathcal{S}_2; \forall t \geq 0$ and $x_i(t), y_i(t) \geq 0, \forall i \in \mathcal{S}_1; \forall t \geq 0$. Then, $x^+(t) \geq y^+(t), \forall t \geq 0$.

Lemma 14. *Decompose the set of single servers $\mathcal{S}$ into two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\lambda_i = 0, \forall i \in \mathcal{S}_1$ and $\lambda_i > 0, \forall i \in \mathcal{S}_2$. Starting from initial solution with conditions*

$$\max_{ji} \mu_{ji} \Big(\sum_{ji \in \mathcal{I}} x_{ji}(0) + \sum_{ji \in \mathcal{I}} z_{ji}(0) + \sum_{i \in \mathcal{S}} x_i^+(0) \Big) < \frac{1}{2} \lambda_{min}$$

and $x_i(0) = 0, \forall i \in \mathcal{S}$, the dynamical system (1.16-1.18) is exponentially stable, and the limit is the unique solution to LCP (A.7) by setting

$$\Theta = \sum_{ji \in \mathcal{I}} x_{ji}(0) + \sum_{ji \in \mathcal{I}} z_{ji}(0) + \sum_{i \in \mathcal{S}} x_i^+(0).$$

Besides the exponential rate of convergence is at least $\min^2(p_{\min}, q_{\min})$.

Now we are ready to prove Lemma 9.

Proof of Lemma 9. Set $\Theta = L(x(t))$. Therefore, $x^+(t) \geq u(\Theta)$ and Lemma 10 implies that

$$\sum_{ji \in \mathcal{I}} (x_{ji}(t) - u_{ji}(\Theta)) + \sum_{ji \in \mathcal{I}} (z_{ji}(t) - v_{ji}(\Theta)) + \sum_{i \in \mathcal{S}} (x_i^+(t) - u_i(\Theta))) \geq \Theta^i - \Theta$$

Define

$$\lambda_{\min} = \min_{\substack{i \in S: \\ \lambda_i \neq \sum_{j, ji \in \mathcal{I}} u_{ji}(\Theta)}} (\lambda_i - \sum_{j, ji \in \mathcal{I}} u_{ji}(\Theta)).$$

As such, there exists a system state $y(t)$ such that $u(\Theta) \leq y^+(t) \leq x^+(t)$, and $y(t) \leq x(t)$ and

$$\sum_{ji\in\mathcal{I}}(y_{ji}(t) - u_{ji}(\Theta)) + \sum_{ji\in\mathcal{I}}(w_{ji}(t) - v_{ji}(\Theta)) + \sum_{i\in\mathcal{S}}(y_i^+(t) - u_i(\Theta))) = \frac{\lambda_{\min}}{2}$$

Define the initial states $\tilde{x}(0) = x(t) - u(\Theta)$ and $\tilde{y}(0) = y(t) - u(\Theta)$. Then, consider $\tilde{x}(t) = (\tilde{x}_i, \tilde{x}_{ji}, \tilde{z}_{kl})$ the unique solution to the following system of equations

$$\begin{aligned}
\tilde{x}_i(t) &= \tilde{x}_i(0) - (\lambda_i - \lambda_i')t + \sum_{ji\in\mathcal{I}}\mu_{ji}\int_0^t \tilde{x}_{ji}(s)ds + \sum_{ji\in\mathcal{I}}\mu_{ji}\int_0^t \tilde{z}_{ji}(s)ds \\
&\quad +\theta_i\int_0^t \tilde{x}_i^-(s)ds - \gamma_i\int_0^t \tilde{x}_i^+(s)ds, \qquad i\in\mathcal{S} \\
\tilde{x}_{ji}(t) &= \tilde{x}_{ji}(0) - \mu_{ji}\int_0^t \tilde{x}_{ji}(s)ds \\
&\quad +p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}\int_0^t \tilde{x}_{kj}(s)ds + p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}\int_0^t \tilde{z}_{kj}(s)ds \\
&\quad +p_{ji}y_j^+(0) - p_{ji}\tilde{x}_j^+(t) - p_{ji}\gamma_j\int_0^t \tilde{x}_j^+(s)ds, \qquad ji\in\mathcal{I} \\
\tilde{z}_{ji}(t) &= \tilde{z}_{ji}(0) - \mu_{ji}\int_0^t \tilde{z}_{ji}(s)ds + q_{ji}\gamma_j\int_0^t \tilde{x}_j^+(s)ds, \qquad ji\in\mathcal{I}
\end{aligned}$$

Similarly, define the dynamical system that uniquely determines $\tilde{y}$ by substituting the elements of $\tilde{x}$ by elements of $\tilde{y}$ in above equations. Note that in the proof for Lemma 12 we prove that $\tilde{x}(t'-t) = x(t') - u(\Theta)$, $\tilde{y}(t'-t) = y(t') - u(\Theta)\forall t' \geq t$, $\tilde{u}(\Theta') = u(\Theta' + \Theta) - u(\Theta)$, and $\tilde{L}(\tilde{x}(t')) = L(x(t')) - \Theta$.

Moreover, from Lemma 14, we know that

$$\tilde{L}(\tilde{y}(t')) - \frac{\lambda_{\min}}{2} \geq \lambda_{\min}e^{-\min(p_{\min},q_{\min})^2(t'-t)}.$$

Therefore,

$$L(y(t')) \geq \Theta + \frac{\lambda_{\min}}{2} - \lambda_{\min}e^{-\min(p_{\min},q_{\min})^2(t'-t)}.$$

Also, from Lemma 13 we know that $x^+(t') \geq y^+(t')$, we have,

$$L(x(t')) \geq \Theta + \frac{\lambda_{\min}}{2} - \lambda_{\min}e^{-\min(p_{\min},q_{\min})^2(t'-t)}$$

Now, set $T_2 = 2/\min^2(p_{\min}, q_{\min})$ to conclude that

$$L(x(t')) \geq \Theta + \frac{\lambda_{\min}}{4}, \forall t' \geq t + T_2.$$

Lastly, note that

$$\lambda_{\min} \geq \lambda_i - \sum_{j,ji\in\mathcal{I}} u_{ji}(\Theta) = \sum_{j,ji\in\mathcal{I}} (u_{ji}(\Theta^j) - u_{ji}(\Theta)) \geq \iota|\Theta^j - \Theta|.$$

This concludes the proof. Also, note that ι is at least greater than the minimum slope of functions $u_{ji}, v_{ji}, \forall ji \in \mathcal{I}$.

Lemma 15. *Starting from an initial solution $|x_0| \leq M$, there exists a positive constants t_4, α_M such that the solution to the dynamical system (1.16-1.18) satisfies*

$$\Big|x^+(t+t_4) - f\Big| \leq \frac{\iota_2}{\iota_1}\Big|x^+(0) - f\Big|e^{-\alpha_M t}$$

Also,

$$\Big|L(x(t+t_4)) - 1\Big| \leq \Big|L(x(0)) - 1\Big|e^{-\alpha_M t}$$

Proof. Consider δ such that

$$\delta \leq \iota_1^2 \min_i(\Theta^{i+1} - \Theta^i)/8e^{CT_2}. \tag{A.8}$$

Set the precision parameter in Proposition 2 as $\delta/2$ to obtain $T_1 = \frac{1}{\min_j \theta_j}\log(\frac{C_2}{\delta/2})$. Consider

$$\delta_1 = \frac{\delta^{(C/\min_j\theta_j)+1}}{2^{(C/\min_j\theta_j)+1}C_2^{C/\min_j\theta_j}}. \tag{A.9}$$

Note that

$$T_1 = \frac{1}{\min_j \theta_j}\log(\frac{C_2}{\delta/2}) = (C/\min_j \theta_j^2)log(1/\delta_1) + O(1) \tag{A.10}$$

More over, consider the minimum integer d_i such that

$$\Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1 e^{CT_2} \leq \iota_1^2/8(1-\iota_1/8)^{d_i-2}(\Theta^{i+1} - \Theta^i) \tag{A.11}$$

Define $d = \min_{i:0\leq i\leq|\mathcal{S}|} d_i$. Taking logarithm from both sides in (A.11) yields

$$d\log\frac{1}{1-\iota_1/8} = \log(1/\delta_1) + O(1)$$

Therefore,

$$d = \frac{\log(1/\delta_1)}{\log\frac{1}{1-\iota_1/8}} + O(1) \tag{A.12}$$

Also, note that if for $k \in \mathcal{I}$ we have $u_k(\Theta^i) = u_k(\Theta^{i+1}) > 0$, then Lemma 8 yields $u_k(\Theta^i) =$

$u_k(\Theta^{|S|+1})$. Next, consider the set $\{t^{(i,j)} : 0 \leq i \leq |\mathcal{S}|, 0 \leq j \leq d_i\}$ such that

$$\begin{aligned} t^{(i,0)} &= \sum_{k=0}^{i}(T_1 + (d_k - 1)T_2) \\ t^{(i,j)} &= t^{(i,0)} + T_1 + (j-1)T_2; \quad 1 \leq j \leq d-1 \end{aligned}$$

To prove the lemma, we first use induction on $i \in [0, |\mathcal{S}| + 1]$ to show that for $t \geq t^{(i,0)}$ we have

$$x^+(t) + \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1\mathbf{1} \geq u(\Theta^i)$$

For the induction base, $i = 0$, note that

$$x^+(0) + \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|}\delta_1 \geq u(\Theta^0) = \mathbf{0}.$$

Assume the induction hypothesis is correct for i, in the following three steps we prove that the induction hypothesis is correct for $i+1$

Step 1. From induction hypothesis we have for $t \geq t^{(i,0)}$

$$x^+(t) + \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1\mathbf{1} \geq u(\Theta^i). \tag{A.13}$$

In this step, for each l such that $u_l(\Theta^{i+1}) > 0$ and for $t \geq t^{(i,1)}$ we prove that

$$\begin{aligned} x_l(t) &\geq \psi(x^{(i,0)} + \delta_1\mathbf{1}, T_1)_l - \delta_1 e^{CT_1} \\ &\geq -\delta/2 - \delta/2 \\ &\geq -\delta \end{aligned} \tag{A.14}$$

where, for $y(t) = (y_{ji}(t), w_{ji}(t), y_i(t))$ the unique solution to the dynamical system (1.16-1.18), with abuse of notation we denote $\psi(y,t)_l = y_l(t)$. To proceed, note that for $t \geq t^{(i,1)}$ we have $t - T_1 \geq t^{i,0}$. Then, Lemma 2 yields

$$|\psi(x(t-T_1), T_1) - \psi(x(t-T_1) + \delta_1\mathbf{1}, T_1)| \leq \delta_1 e^{CT_1}$$

Next, Lemma 10 and Proposition 2 yield for each $l \in \mathcal{S}$ such that $u_l(\Theta^{i+1}) > 0$ and for $t \geq t^{(i,1)}$,

$$(\psi(x(t-T_1) + \delta_1\mathbf{1}, T_1))_l \geq -\delta/2$$

Combining with (A.9), for each l such that $u_l(\Theta^{i+1}) > 0$ and (A.14) satified.

Step 2. From induction hypothesis and Step 1 we have for $t \geq t^{(i,1)}$,

$$x^{+}(t) + \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1\mathbf{1} \geq u(\Theta^i),$$

and (A.14) holds for all $l \in \mathcal{S}$ such that $u_l(\Theta^{i+1}) > 0$. In this step we prove that $\forall l \in \mathcal{S}, u_l(\Theta^{i+1}) > 0$ and for $t \geq t^{(i,2)}$,

$$x_l(t) \geq 0 \tag{A.15}$$

To proceed, for $t \geq t^{(i,2)}$ Lemma 2 yields

$$|\psi(x(t-T_2), T_2) - \psi(x(t-T_2) + \delta\mathbf{1}, T_2)| \leq \delta e^{CT_2}$$

Next, Lemma 9 yields

$$L(\psi(x(t-T_2) + \delta\mathbf{1}, T_2)) \geq \Theta^{i+1} - (1 - \iota_1/4)(\Theta^{i+1} - \Theta^i)$$

Therefore, for all $k \in \mathcal{S}$ such that $0 < u_k(\Theta^{i+1})$ and for $t \geq t^{(i,2)}$we have

$$\begin{aligned} x_k(t) &\geq (\psi(x(t-T_2) + \delta\mathbf{1}, T_2))_k - \delta e^{CT_2} \\ &\geq u_k(\Theta^{j+1} - (\Theta^{j+1} - \Theta^j)(1 - \iota_1/4)) - \delta e^{CT_2} \end{aligned}$$

As the minimum nonzero slope in the piece wise linear functions $u(.)$ is ι_1 we can lower bound the right hand side as follows

$$\begin{aligned} & u_k(\Theta^{j+1} - (\Theta^{j+1} - \Theta^j)(1 - \iota_1/8)) + \iota_1^2(\Theta^{j+1} - \Theta^j)/8 - \delta e^{CT_2} \\ \geq\ & u_k(\Theta^{j+1} - (\Theta^{j+1} - \Theta^j)(1 - \iota_1/8)) \end{aligned}$$

Therefore, it follows from (A.8) that $\forall l \in \mathcal{S}, u_l(\Theta^{i+1}) > 0$ and (A.15) holds.

Step 3. In this step we complete the induction proof by showing that for $t \geq t^{(i+1,0)}$,

$$x^{+}(t) + \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i-1}\delta_1\mathbf{1} \geq u(\Theta^{i+1}) \tag{A.16}$$

To proceed, note that the maximum slope in the piece-wise linear functions u_k is ι_2. Therefore

$$u(\Theta^{i+1} - (1 - \iota_1/8)^{d_i - 2}(\Theta^{i+1} - \Theta^i))\mathbf{1} \geq u(\Theta^{i+1}) - \iota_2(1 - \iota_1/8)^{d_i - 2}(\Theta^{i+1} - \Theta^i)\mathbf{1}$$

Thus, to prove (A.16), it is sufficient to prove for $t \geq t^{(i+1,0)}$,

$$x(t) + (\frac{\iota_1}{e^{CT_2}\iota_2})^{|\mathcal{S}|-i-1}\delta_1\mathbf{1} - \iota_2(1-\iota_1/8)^{d_i-2}(\Theta^{i+1}-\Theta^i)\mathbf{1} \geq u(\Theta^{i+1} - (1-\iota_1/8)^{d_i-2}(\Theta^{i+1}-\Theta^i))$$

Also, note that (A.11) yields

$$\begin{aligned} & \left(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\right)^{|\mathcal{S}|-i-1}\delta_1 - \iota_2(1-\iota_1/8)^{d_i-2}(\Theta^{i+1}-\Theta^i) \\ \geq & \left(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\right)^{|\mathcal{S}|-i-1}\delta_1 - \frac{8\iota_2}{\iota_1^2}(\frac{16\iota_1^2}{e^{CT_2}\iota_2})^{|\mathcal{S}|-i+2}\delta_1 e^{CT_2} \geq \left(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\right)^{|\mathcal{S}|-i}\delta_1 \end{aligned}$$

Thus, to prove (A.16) it suffices to prove for each $2 \leq j \leq d_i$ and $t \geq t^{(i,j)}$,

$$x(t) + \left(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\right)^{|\mathcal{S}|-i}\delta_1\mathbf{1} \geq u\Big(\Theta^{i+1} - (1-\iota_1/8)^{j-2}(\Theta^{i+1}-\Theta^i)\Big) \tag{A.17}$$

Note that in Step 2 we have proved (A.17) holds for $j = 2$. Next, we prove (A.17) in an iterative argument. To proceed, assume (A.17) holds for j and we want to prove it holds for $j+1 \leq d_i$. Lemma 9 yields

$$L\Big(\psi\Big(x(t-T_2) + \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1\mathbf{1}, T_2\Big)\Big) \geq \Theta^{i+1} - (1-\iota_1/4)(1-\iota_1/8)^{j-2}(\Theta^{i+1}-\Theta^i)$$

Next, the Lipschitz continuity of the function $\psi(.)$ at $x(t-T_2)$ for $t \geq t^{(i,j)}$ yields

$$\Big|\psi(x(t-T_2), T_2) - \psi\Big(x(t-T_2) + \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1\mathbf{1}, T_2\Big)\Big| \leq \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1 e^{CT_2}$$

For all $k \in \mathcal{S} \cup \mathcal{I}$ such that $0 < u_k(\Theta^i) < u_k(\Theta^{i+1})$ we have

$$\begin{aligned} x_k(t) & \geq \psi\Big(x(t-T_2) + \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta\mathbf{1}, T_2\Big)_k - \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1 e^{CT_2} \\ & \geq u_k(\Theta^{i+1} - (1-\iota_1/4)(1-\iota_1/8)^{j-2}(\Theta^{i+1}-\Theta^i)) - \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1 e^{CT_2} \end{aligned}$$

Similarly, For all $k \in \mathcal{I}$ such that $0 < v_k(\Theta^i) < v_k(\Theta^{i+1})$ we have

$$\begin{aligned} z_k(t) & \geq \psi\Big(x(t-T_2) + \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta\mathbf{1}, T_2\Big)_k - \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1 e^{CT_2} \\ & \geq v_k(\Theta^{i+1} - (1-\iota_1/4)(1-\iota_1/8)^{j-2}(\Theta^{i+1}-\Theta^i)) - \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1 e^{CT_2} \end{aligned}$$

As the minimum nonzero slope in the piece-wise linear functions u_k and v_k is ι_1, we can lower bound the right-hand sides as follows (to avoid repetition we only present the lower bounds for the functions u)

$$\begin{aligned}
& u_k(\Theta^{i+1} - (1-\iota_1/8)^{j-1}(\Theta^{i+1} - \Theta^i)) + \iota_1^2(1-\iota_1/8)^{j-2}(\Theta^{i+1} - \Theta^i)/8 \\
& -\Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1 e^{CT_2} \\
\geq\ & u_k(\Theta^{i+1} - (1-\iota_1/8)^{j-1}(\Theta^{i+1} - \Theta^i))
\end{aligned}$$

The last inequality follows from (A.11). Therefore,

$$x(t) + \Big(\frac{16\iota_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1\mathbf{1} \geq u(\Theta^{i+1} - (1-\iota_1/8)^{j-1}(\Theta^{i+1} - \Theta^i))$$

This concludes the proof for the three steps of the induction proof.

Next, note that the induction result yields, for $t \geq t^{(|\mathcal{S}|,d-1)}$,

$$x(t) + \delta_1\mathbf{1} \quad\geq\quad u(\Theta^{|\mathcal{S}|+1}) = u(1)$$

Furthermore, note that

$$\begin{aligned}
t^{(|\mathcal{S}|,d-1)} &\leq (|\mathcal{S}|+1)(T_1 + (d-1)T_2) \\
&\leq |\mathcal{S}|\big((C/\min_j \theta_j^2) + \frac{1}{\log\frac{1}{1-\iota_1/8}}\big)\log(1/\delta_1) + O(1)
\end{aligned}$$

Therefore, we conclude there exists a constant $T' > 0$ such that for $T \geq T'$

$$\begin{aligned}
& \Big| x^+\Big(|\mathcal{S}|[(C/\min_j \theta_j^2) + \frac{1}{\log\frac{1}{1-\iota_1/8}}]\log(1/\delta_1) + T\Big) \\
& \qquad\qquad . - u\Big(\sum_{ji\in\mathcal{I}_1} x_{ji}(0) + \sum_{ji\in\mathcal{I}_2} z_{ji}(0) + \sum_{i\in\mathcal{S}} x_i^+(0)\Big)\Big| \\
\leq\ & \delta_1 = e^{-log(1/\delta_1)}
\end{aligned}$$

Also, note that if we set

$$\alpha_M^{'} = \Big(|\mathcal{S}|[(C/\min_j \theta_j^2) + \frac{1}{\log\frac{1}{1-\iota_1/8}}]\log(1/\delta_1)\Big)^{-1}$$

Then, as δ_1 can be chosen arbitrarily small, we conclude that

$$\Big| x^+(t+T') - u(\sum_{ji\in\mathcal{I}} x_{ji}(0) + \sum_{ji\in\mathcal{I}} z_{ji}(0) + \sum_{i\in\mathcal{S}} x_i^+(0))\Big| \leq e^{-\alpha_M^{'}t}$$

Now, consider $t_4 > T' + \frac{\log(\Theta^{|\mathcal{S}|+1} - \Theta^{|\mathcal{S}|})}{\alpha_M^{'}}$ to realize that

$$\max(L(x^+(t), \Theta^{|\mathcal{S}|}) < L(x^+(t+t_4))$$

Therefore, as $\Theta^{|\mathcal{S}|+1} = 1$, and by lemma 9 we conclude

$$1 - L(x(t_4 + kT_2)) \leq \Big(1 - L(x(t+t_4))\Big)(1 - \iota_1/4)^k$$

The increasing property of the Lyapunov function L yields

$$1 - L(x(t+t_4)) \leq \Big(1 - L(x(0))\Big)e^{\frac{t}{T_2}\log(1-\iota_1/4)}$$

As the minimum slope of the functions $u(.)$ is ι_1 we conclude

$$|x^+(t+t_4) - f| \leq \frac{1}{\iota_1}\Big|1 - L(x(0))\Big|e^{\frac{t}{T_2}\log(1-\iota_1/4)} \leq \frac{\iota_2}{\iota_1}\Big|x^+(0) - f\Big|e^{\frac{t}{T_2}\log(1-\iota_1/4)}.$$

Now, we set

$$\alpha_M = -\frac{\log(1-\iota_1/4)}{T_2} = -\frac{\log(1-\iota_1/4)\min_{ji,j}\frac{\lambda_{ji}}{\lambda_j}}{2}$$

to conclude the proof. □

Lemma 16. *There exists constants $\iota_1, \iota_2, t_4 > 0$ and $\alpha_M > 0$, such that starting from an initial solution $|x_0| \leq M$ the solution to the dynamical system (1.16-1.18) satisfies*

a) $|x^+(t+t_4) - f| \leq \frac{\iota_2}{\iota_1}|x^+(0) - f|e^{-\alpha_M t}$

b) $|x^-(t) - \beta| = O(t^{-1})$

Proof. Note that part a) is proved in Lemma 15. Next, to prove part b), note that if $f_i > 0$, then (1.2d) and (1.2e) yield $\beta_i = 0$. Therefore, Lemma 11 concludes the proof for this case.

Next, we consider the case $f_i \leq 0$. In this case, (1.2e) yields $\theta_i\beta_i = -\lambda_i + \sum_{ji}\mu_{ji}f_{ji}$. Thus, to conclude the Lemma statement it is sufficient to prove the following two equations

$$\theta_i x_i(t) \geq -\lambda_i + \sum_{ji\in\mathcal{I}}\mu_{ji}f_{ji} - O(t^{-1}). \tag{A.18}$$

$$\theta_i x_i(t) \leq -\lambda_i + \sum_{ji\in\mathcal{I}}\mu_{ji}f_{ji} + O(t^{-1}). \tag{A.19}$$

Also, note that Lemma 10 part d) yields

$$\frac{|x_i(t)|}{t} \leq \frac{C_2}{t}.$$

Thus, (1.16) and Lemma 15 yield $\forall i \in \mathcal{S}$

$$\frac{x_i(t)}{t} \leq -\lambda_i + \sum_{ji \in \mathcal{I}} \mu_{ji} f_{ji} + \theta_i \frac{\int_0^t x_i^-(s)ds}{t} + C_2 \frac{\iota_2}{\iota_1 t} e^{-\alpha_M t}\Big(2 \sum_{ji} \mu_{ji} + 1\Big) \tag{A.20}$$

$$\frac{x_i(t)}{t} \geq -\lambda_i + \sum_{ji \in \mathcal{I}} \mu_{ji} f_{ji} + \theta_i \frac{\int_0^t x_i^-(s)ds}{t} - C_2 \frac{\iota_2}{\iota_1 t} e^{-\alpha_M t}\Big(2 \sum_{ji} \mu_{ji} + 1\Big). \tag{A.21}$$

Next, Consider

$$t_0 = \max\bigg(1, \frac{1}{\alpha_M} \log\Big(\frac{\iota_1 t}{C_2 \iota_2 (2 \sum_{ji} \mu_{ji} + 1)}\Big)\bigg).$$

Note that $t_0 = O(\log(t))$. From (A.21) we have

$$\theta_i \frac{\int_{t_0}^t x_i^-(s)ds}{t} \leq \lambda_i - \sum_{ji \in \mathcal{I}} \mu_{ji} f_{ji} + t^{-1}.$$

So, there exists $s \leq t$ such that $s \geq t_0$ and

$$\theta_i x_i^-(s) \leq \lambda_i - \sum_{ji \in \mathcal{I}} \mu_{ji} f_{ji} + 2t^{-1}.$$

Rearranging the terms yields

$$\theta_i x_i(s) \geq -\lambda_i + \sum_{ji \in \mathcal{I}} \mu_{ji} f_{ji} - 2t^{-1}.$$

Define t_2 as follows:

$$t_2 = \inf\Big\{t \geq s | \theta_i x_i(t) \leq -\lambda_i + \sum_{ji \in \mathcal{I}} \mu_{ji} f_{ji} - 4t^{-1}\Big\}$$

If $t_2 = \infty$ we obtain (A.18). Otherwise, define t_1 as follows:

$$t_1 = \sup\Big\{t \leq t_2 | \theta_i x_i(t) \geq -\lambda_i + \sum_{ji \in \mathcal{I}} \mu_{ji} f_{ji} - 2t^{-1}\Big\}. \tag{A.22}$$

Note that $t_1 \neq -\infty$ because s satisfies (A.22). As a result,

$$\begin{aligned}
-t^{-1} &\geq x_i(t_2) - x_i(t_1) \\
&= \int_{t_1}^{t_2} \dot{x}_i(s) ds \\
&= \int_{t_1}^{t_2} \Big(- \lambda_i + \sum_{ji \in \mathcal{I}} \mu_{ji} x_{ji}(s) + \sum_{ji \in \mathcal{I}} \mu_{ji} z_{ji}(s) + \theta_i x_i^-(s) - \gamma_i x_i^+(s) \Big) ds \\
&\geq \int_{t_1}^{t_2} (1/t) ds \\
&> 0
\end{aligned}$$

where, the first and second inequality follow from the definition of t_1 and t_2 and the fact that t_2 is finite. We proved the third equation as part of the proof of Lemma 11 (details are presented in (A.68)). The fourth inequality follows from Lemma 15 and definition of t_1 and t_2. This contradiction concludes $t_2 = \infty$. As such, (A.18) is satisfied. For $\beta_i < 0$, a completely similar procedure yields (A.19). Also, for $\beta_i = 0$ (A.18) together with $x_i^-(t) \geq 0$ concludes the proof. Thus,

$$|x_i^-(t) - \beta_i| = O(t^{-1}).$$

□

A.10 Lemmas used in the proof of Theorem 2

To start note that for each state dependent policy $(\boldsymbol{\lambda}, \boldsymbol{Q}, \boldsymbol{\gamma})$, we write the system dynamics to define a continuous time Markov chain on the system state space as follows:

Table A.1: Markov Chain Transition Rates

Description	Transition	Rate
Passenger arrives to node j and joins passenger queue	$X_j - 1$	$\sum_{i,ji\in\mathcal{I}} \lambda_{ji}\mathbf{1}(X_j^-)$
Passenger arrives to node j and matches a waiting driver in node j and travels to node i	$X_i - 1$ $X_{ik} + 1$	$\lambda_{ji}\mathbf{1}(X_j^+)$
Passenger reneges from the passenger queue in node i	$X_i + 1$	$\theta_i X_i^-$
Vehicle reneges empty from the vehicle queue in node i toward node k	$X_i - 1$ $Z_{ik} + 1$	$\gamma_i X_i^+ q_{ik}$
Full vehicle drops off a passenger at node i, originally picked up from node j. Then, it picks up a passenger from node i and departs toward destination k.	$X_{ji} - 1$ $X_{ik} + 1$ $X_i + 1$	$\mu_{ji}X_{ji}(\frac{\lambda_{ik}}{\sum_l \lambda_{il}})\mathbf{1}(X_i^-)$
Full vehicle drops off a passenger at node i, originally picked up from node j. Then, it joins the vehicle queue at node j.	$X_{ji} - 1$ $X_i + 1$	$\mu_{ji}X_{ji}\mathbf{1}(X_i^+)$
Empty vehicle drops off a passenger at node i, originally picked up from node j. Then, it picks up a passenger from node i with destination k	$Z_{ji} - 1$ $X_{ik} + 1$ $X_i + 1$	$\mu_{ji}Z_{ji}(\frac{\lambda_{ik}}{\sum_l \lambda_{il}})\mathbf{1}(X_i^-)$
Empty vehicle drops off a passenger at node i, originally picked up from node j. Then, it joins the vehicle queue at node j.	$Z_{ji} - 1$ $X_i + 1$	$\mu_{ji}Z_{ji}\mathbf{1}(X_i^+)$

Also, recall from Section 1.3 that we can consider (1.3) and (1.4) within the subset of feasible region that satisfies $\mu_{ji}f_{ji} = \lambda_{ji}$ to rewrite the optimization problem as follows:

$$\max_{\boldsymbol{\lambda},\boldsymbol{f},\boldsymbol{e}} \quad \sum_{ji\in\mathcal{I}} \mu_{ji}f_{ji}I_{ji}(\mu_{ji}f_{ji}) - \sum_{ji\in\mathcal{I}} c^V_{ji}\mu_{ji}e_{ji} \tag{A.23}$$

$$\text{s.t.} \quad \sum_{kj\in\mathcal{I}} \mu_{kj}f_{kj} + \sum_{kj\in\mathcal{I}} \mu_{kj}e_{kj} = \sum_{ji\in\mathcal{I}} \mu_{ji}f_{ji} + \sum_{ji\in\mathcal{I}} \mu_{ji}e_{ji}; \quad \forall j \in \mathcal{S} \tag{A.24}$$

$$\sum_{ji\in\mathcal{I}} f_{ji} + \sum_{ji\in\mathcal{I}} e_{ji} + \sum_{i\in\mathcal{S}} f_i = 1, \tag{A.25}$$

$$e_{ji}, f_j, f_{ji} \geq 0 \tag{A.26}$$

The key in proving the ergodicity of the CTMC is applying Foster-Lyapunov theorem (see e.g., Tweedie 1975), which is stated as follows.

Lemma 17. *(Foster-Lyapunov Theorem for CTMCs) Consider an irreducible CTMC on a countable state space $\mathcal{X}$ having a transition rate matrix R with elements r_{XY} for pairs $X, Y \in \mathcal{X}$.*

Then,

(i) The Markov chain is positive recurrent if there exists a Lyapunov function $V : \mathcal{X} \to \mathbb{R}$ and a partition of $\mathcal{X}$ into two subsets including a finite subset $\mathcal{X}_1$ and an infinite subset $\mathcal{X}_2$ that satisfy the following two conditions:

$$\sum_{j \in \mathcal{X}} r_{XY} V(Y) \leq -1 \qquad \forall X \in \mathcal{X}_1$$

$$\sum_{j \in \mathcal{X}} r_{XY} V(Y) < \infty \qquad \forall X \in \mathcal{X}_2$$

$$\sum_{X \in \mathcal{X}} \boldsymbol{1}(V(X) < M) \leq \infty \qquad \forall M \in \mathbb{R}$$

(ii) The embedded discrete time Markov chain (E-DTMC) associated with the original CTMC is positive recurrent if there exists a Lyapunov function $V' : \mathcal{X} \to \mathbb{R}$ and a partition of $\mathcal{X}$ into two subsets including a finite subset $\mathcal{X}_1$ and an infinite subset $\mathcal{X}_2$ that satisfy the following two conditions:

$$\sum_{Y \in \mathcal{X}} r_{XY} V'(Y) \leq r_{XX} \qquad \forall X \in \mathcal{X}_1$$

$$\sum_{Y \in \mathcal{X}} \frac{r_{XY}}{r_{XX}} V'(Y) < \infty \qquad \forall X \in \mathcal{X}_2$$

Lemma 18. *Under any dynamic joint pricing and empty relocation policy, the resulting CTMC and its embedded DTMC are positive recurrent.*

Proof. We first define the value function

$$V(X) = |\mathcal{S}| \max_{i \in \mathcal{S}} X_i^- + \sum_{j \in \mathcal{S}} \mathbf{1}_{X_j^- = \max_{i \in \mathcal{S}} X_i^-}$$

Then, we set $\mathcal{X}_2$ to be the states X such that

$$\max_{i \in \mathcal{S}} X_i^- \leq \frac{2}{\min_i \theta_i} |2\mathcal{I} + 2\mathcal{S}| N(\max_{ji \in \mathcal{I}} \mu_{ji} + \max_i \lambda_i + \max_i \gamma_i)|\mathcal{S}| + 2$$

Next, consider a state $X \in \mathcal{X}_1$. Note that the number of states Y with a nonzero transition rate from X to Y is bounded by $|2\mathcal{I} + 2\mathcal{S}|$. Then, consider a state Y with a nonzero transition from state X to Y, $r_{XY} > 0$. Consider the following three cases:

- $V(Y) > V(X)$: in this case we have $r_{XY} \leq N(\max_{ji \in \mathcal{I}} \mu_{ji} + \max_i \lambda_i + \max_i \gamma_i)$

- $V(Y) < V(X)$: in this case we have $r_{XY} \geq (\min_i \theta_i)\Big(\frac{V(X)}{|\mathcal{S}|} - 1\Big)$

Thus we have

$$\begin{aligned}\sum_{Y \in \mathcal{X}} r_{XY} V(Y) &= \sum_{\substack{Y \in \mathcal{X}: \\ V(Y) > V(X)}} r_{XY}(V(Y) - V(X)) + \sum_{\substack{Y \in \mathcal{X}: \\ V(Y) = V(X)}} 0 \\ &\quad + \sum_{\substack{Y \in \mathcal{X}: \\ V(Y) < V(X)}} r_{XY}(V(Y) - V(X)) \\ &\leq |2\mathcal{I} + 2\mathcal{S}| N(\max_{ji \in \mathcal{I}} \mu_{ji} + \max_i \lambda_i + \max_i \gamma_i)|\mathcal{S}| - (\min_i \theta_i)\Big(\frac{V(X)}{|\mathcal{S}|} - 1\Big) \\ &\leq -\frac{1}{2}(\min_i \theta_i)\Big(\frac{V(X)}{|\mathcal{S}|} - 1\Big) \leq -1\end{aligned}$$

Thus, Lemma 17 implies that the continuous time markov chain is ergodic and therefore, denote by $\pi : \mathcal{X} \to \mathbb{R}$ the stationary probability distribution of the continuous time markov chain. To prove the ergodicity of the E-DTMC set

$$V'(X) = 4|2\mathcal{I} + 2\mathcal{S}|V(X)|S|\frac{\max_i \theta_i}{\min_i \theta_i}$$

Next, note that for each $X \in \mathcal{X}_1$ we have

$$|r_{XX}| \leq |2\mathcal{I} + 2\mathcal{S}|(\max_i \theta_i)(\max_i X_i^-)$$

Then similar to the proof scheme for ergodicity of the CTMC we have

$$\sum_{Y \in \mathcal{X}} r_{XY} V(Y) \leq -\frac{1}{2}(\min_i \theta_i)\Big(\frac{V(X)}{|\mathcal{S}|} - 1\Big) \leq -r_{XX}.$$

□

Lemma 19. *The expected steady-state scaled queue length* $\mathbb{E}[X(\infty)/N]$ *satisfies all the constraints of fluid optimization problem (A.23-A.26).*

Proof. To prove this lemma we take the same approach as Braverman et al. [2019] by applying Proposition 3 of Glynn and Zeevi [2008]. Specifically, if a CTMC with rate matrix R imposses a stationary distribution π and moreover if Rg is π integrable, then $\pi R g = 0$. Lemma 18 proves exsistence of stationary distribution π and π integrablity of Rg for any bounded test function g. Now, we prove that the CTMC satisfies the constraints in the fluid optimization problem (A.23-A.26). First, to prove (A.24), we consider the test function $g(X) = X_i^+(X) + \sum_{ik} X_{ik}(X) +$

$\sum_{ik} Z_{ik}(X)$. Doing so, $\pi R g = 0$ yields

$$\begin{aligned}
& \sum_X \pi(X) R g(X) \\
= \;& \sum_{X,Y} \pi(X) r_{X,Y} g(Y) \\
= \;& \sum_X \sum_{\substack{Y:r_{XY}>0,\\ g(Y)=g(X)-1}} \pi(X) r_{XY}(-1) + \sum_X \sum_{\substack{Y:r_{XY}>0\\ g(Y)=g(X)+1}} \pi(X) r_{XY}(1) \\
= \;& \sum_X \pi(X)\Big(\sum_{ik} \mu_{ik} X_{ik}(X) + \sum_{ik} \mu_{ik} Z_{ik}(X)\Big)(-1) \\
& + \sum_X \pi(X)\Big(\sum_{ji} \mu_{ji} X_{ji}(X) + \sum_{ji} \mu_{ji} Z_{ji}(X)\Big)(1) \\
= \;& \sum_{X,Y} \pi(X)\Big(- \sum_{ik} \mu_{ik} X_{ik}(X) - \sum_{ik} \mu_{ik} Z_{ik}(X) + \sum_{ji} \mu_{ji} X_{ji}(X) + \sum_{ji} \mu_{ji} Z_{ji}(X)\Big) \\
= \;& 0
\end{aligned}$$

The last equality gives (A.24). Second, it is straightforward to see that constraint (A.25) is satisfied, since the total number of vehicles is invariant according to the transitions demonstrated in Table A.1. Lastly, the (A.26) is satisfies since the transitions demonstrated in Table A.1 yield $X_{ji}(T), Z_{ji}(T), X_{ji}^{+}(T) \geq 0$. □

Lemma 20. *The long-run average utility rate of any dynamic policy is upper bounded by the optimal objective value of the fluid optimization in (A.23-A.26).*

Proof. Consider a Dynamic policy $(\boldsymbol{\lambda}, \boldsymbol{Q}, \boldsymbol{\gamma})$. Denote by $\lambda_{ji}^{(s)}$ the rate of passenger arrivals for OD pair (j,i), when system state is $s \in \mathcal{X}$. We refer to a passenger who enters the system in state s as a type s passenger and the system reward when serving a type s passenger with OD pair (j,i) is $I(\lambda_{ji}^{(s)})$. To prove the lemma statement, it suffices to prove that

$$\begin{aligned}
& \limsup_{T\to\infty} \frac{1}{NT} \mathbb{E}\left[\sum_{ji\in\mathcal{I}} \int_0^T I_{ji}(\lambda_{ji}(t)) dK_{ji}(t,T) - \sum_{ji\in\mathcal{I}} \int_0^T c_{ji}^V dJ_{ji}(t) - \sum_{j\in\mathcal{S}} \int_0^T c_j^W X_j^-(t) dt \right] \\
\leq \;& \sum_{ji\in\mathcal{I}} \mu_{ji} \mathbb{E}[X_{ji}(\infty)/N] I_{ji}(\mu_{ji}\mathbb{E}[X_{ji}(\infty)/N]) - \sum_{ji\in\mathcal{I}} c_{ji}^V \mu_{ji} \mathbb{E}[Z_{ji}(\infty)/N] \qquad \text{(A.29)}
\end{aligned}$$

We prove (A.29) in three steps. In the first step, define

$$\zeta_{ji}^{(s,T)} = \frac{\textit{\#completed trips of type } s \textit{ from } j \textit{ to } i \textit{ by time } T}{N \lambda_{ji}^{(s)} T}$$

Note that we can rewrite the reward portion of the objective function as follows:

$$\limsup_{T\to\infty} \frac{1}{NT}\mathbb{E}\left[\sum_{ji\in\mathcal{I}}\int_0^T I_{ji}(\lambda_{ji}(t))dK_{ji}(t,T)\right] \quad = \quad \limsup_{T\to\infty}\mathbb{E}[\sum_{ji\in\mathcal{I}}\sum_{s\in\mathcal{X}}\zeta_{ji}^{(s,T)}\lambda_{ji}^{(s)}I_{ji}(\lambda_{ji}^{s})] \tag{A.30}$$

Next, we upper bound the number of type s completed trips with the number of type s requests made by time T, and apply uniform law of large numbers to conclude

$$\limsup_{T\to\infty}\zeta_{ji}^{(s,T)} \le \lim_{T\to\infty}\frac{N\pi_s\lambda_{ji}^{(s)}T}{N\lambda_{ji}^{(s)}T} = \pi_s \tag{A.31}$$

Furthermore, note that

$$\limsup_{T\to\infty}\sum_{s\in\mathcal{X}}\zeta_{ji}^{(s,T)}\lambda_{ji}^{(s)} = \mu_{ji}\mathbb{E}[X_{ji}(\infty)/N] \tag{A.32}$$

where the first term is the total number of completed trips from node j to node i by time T divided by NT. As the first term represents the average arrival rate of full vehicles from j to i divided by N, the equality follows from proposition 1.1 in Little [1961]. As the functions $\lambda_{ji}I_{ji}(\lambda_{ji})$ are concave in λ_{ji}, Jensen's inequality yields

$$\sum_{ji\in\mathcal{I}}\sum_{s\in\mathcal{X}}\zeta_{ji}^{(s,T)}\lambda_{ji}^{(s)}I_{ji}(\lambda_{ji}^{s}) \quad \le \quad \sum_{ji\in\mathcal{I}}\Big(\sum_{s\in\mathcal{X}}\zeta_{ji}^{(s,T)}\Big)\Big(\sum_{s\in\mathcal{X}}\zeta_{ji}^{(s,T)}\lambda_{ji}^{(s)}I_{ji}\Big(\frac{\sum_{s\in\mathcal{X}}\zeta_{ji}^{(s,T)}\lambda_{ji}^{(s)}}{\sum_{s\in\mathcal{X}}\zeta_{ji}^{(s,T)}}\Big)\Big)$$

Combining with (A.31), $\sum_{s\in\mathcal{X}}\pi_s = 1$, and the fact that $I_{ji}(\lambda_{ji})$ are non-increasing in λ_{ji} yields

$$\limsup_{T\to\infty}\sum_{ji\in\mathcal{I}}\sum_{s\in\mathcal{X}}\zeta_{ji}^{(s,T)}\lambda_{ji}^{(s)}I_{ji}(\lambda_{ji}^{(s)}) \le \limsup_{T\to\infty}\sum_{ji\in\mathcal{I}}(\sum_{s\in\mathcal{X}}\zeta_{ji}^{(s,T)}\lambda_{ji}^{(s)})I_{ji}(\sum_{s\in\mathcal{X}}\zeta_{ji}^{(s,T)}\lambda_{ji}^{(s)})$$

Note that the non-increasing property follows from the concave reward assumption, see e.g., Banerjee et al. [2021]. Combining with (A.32), we upper bound the right hand side in (A.30) as follows:

$$\sum_{ji\in\mathcal{I}}\mu_{ji}\mathbb{E}[X_{ji}(\infty)/N]I_{ji}(\mu_{ji}\mathbb{E}[X_{ji}(\infty)/N])$$

In the second step, to compute the empty relocation cost in the objective function, we refer to a vehicle who completes its empty relocation in state s as a type s empty relocation. Now, define

$$\zeta_{ji}^{\prime(s,T)} = \frac{\#\textit{empty relocation of type } s \textit{ by time } T}{N\mu_{ji}T}$$

Next, we lower bound the empty relocation portion of the objective function as follows:

$$\limsup_{T\to\infty} \frac{1}{NT}\mathbb{E}\left[-\sum_{ji\in\mathcal{I}}\int_0^T c_{ji}^V dJ_{ji}(T)\right] = \limsup_{T\to\infty}\mathbb{E}\Big[-\sum_{ji\in\mathcal{I}}\zeta_{ji}^{\prime(s,T)}\mu_{ji}c_{ji}^V\Big] \tag{A.33}$$

Also, we apply proposition 1.1 in Little [1961] to conclude

$$\liminf_{T\to\infty}\zeta_{ji}^{\prime(s,T)} = \lim_{T\to\infty}\frac{\pi_s\mu_{ji}Z_{ji}(s)T}{N\mu_{ji}T} = \pi_s\frac{Z_{ji}(s)}{N} \tag{A.34}$$

Then, we upper bound the right hand side in (A.33) as follows:

$$\limsup_{T\to\infty}\mathbb{E}\Big[-\sum_{ji\in\mathcal{I}}\zeta_{ji}^{\prime(s,T)}\mu_{ji}c_{ji}^V\Big] = \sum_{ji\in\mathcal{I}}\mu_{ji}\frac{\mathbb{E}[Z_{ji}(\infty)]}{N}c_{ji}^V$$

In the third step, note that for the third part of the objective we have

$$\limsup_{T\to\infty}\frac{1}{NT}\mathbb{E}[-\sum_{j\in\mathcal{S}}\int_0^T c_j^W X_j^-(t)dt] \leq 0$$

This concludes the proof. □

Lemma 21. *There exists a positive constant $\alpha' > 0$ such that for optimal solution to optimization problem in (1.3) and (1.4) we have*

$$\mathbb{P}\left\{\Big|\frac{X^N(T)}{N} - f^*\Big| < M\right\} \leq 1 - O(e^{-\alpha' N}),$$

where $M = 4\max\Big(C_2, 1\Big)\Big(2|\mathcal{I}| + |\mathcal{S}|\Big)$ and $C_2 = \max\{x_i^-(0), \lambda_i/\theta_i\}$ as defined in Lemma 10.

Proof. First note that from Lemma 10 we have

$$\sum_{ji}\frac{X_{ji}(T)}{N} + \frac{Z_{ji}(T)}{N} + \sum_{j\in\mathcal{S}}\frac{X_j^+(T)}{N} = 1$$

Thus,

$$\sum_{ji}\Big|\frac{X_{ji}(T)}{N} - f_{ji}^*\Big| + \sum_{ji}\Big|\frac{Z_{ji}(T)}{N} - e_{ji}^*\Big| < 2|\mathcal{I}|$$

Also note that

$$\mathbb{P}\left\{\frac{X_j^{N-}(T)}{N} - \beta_j^* < M/2\right\} \leq \mathbb{P}\left\{\frac{X_j^{N-}(T)}{N} < M/2\right\}$$

To prove the lemma, it is sufficient to prove that

$$\mathbb{P}\left\{\frac{X_j^{N-}(T)}{N} < \frac{M}{2|\mathcal{S}|}\right\} \leq 1 - O(e^{-\alpha' N}) \tag{A.35}$$

To prove the last statement, for each station $i \in \mathcal{S}$ we consider a parallel infinite server queue, $M/M/\infty$, with exponential service time with mean $1/\theta_i$. To relate the two systems, whenever a passenger arrives at a station i in the original system with an exponential patience time with mean $1/\theta_i$, a job arrives in the infinite server queue in the parallel system with mean service time $1/\theta_i$. Also, when ever a passenger is served with a vehicle in the original system we mark the corresponding job in the parallel system. Clearly, the number of unmarked jobs in the parallel system equals the passenger queue length at station i in the original system. Also, the number of unmarked jobs in the parallel system is upper bounded by the total number of jobs in the parallel system. Finally, to derive (A.35) we note that when the system is initialized the maximum passenger queue length is at most $M/(4|\mathcal{S}|)$. Therefore, we can upper bound the number of jobs in the infinite server queue by $M/(4|\mathcal{S}|)$ plus the number of jobs arrive after time $t = 0$. Thus, we use the transient probability of infinite server queues to derive (A.35) as follows

$$\begin{aligned}
\mathbb{P}\left\{\frac{X_j^{N-}(T)}{N} < \frac{M}{2|\mathcal{S}|}\right\} &\leq 1 - \mathbb{P}\left\{\frac{X_j^{N-}(T)}{N} > \frac{M}{4|\mathcal{S}|} \;\Big|\; \frac{X_j^{N-}(0)}{N} = 0\right\} \\
&\leq 1 - \sum_{i=\frac{MN}{4|\mathcal{S}|}}^{\infty} e^{-(N\lambda_j/\theta_j)(1-e^{\theta_j t})} \frac{\left(\frac{N\lambda_j}{\theta_j}(1-e^{-\theta_j t})\right)^i}{i!} \\
&\leq 1 - O(e^{-\alpha' N})
\end{aligned}$$

where $\alpha > 0$ is a constant. This concludes the proof. □

Lemma 22. *There exists positive constants $\alpha_S, \alpha_U, M > 0$ such that we have*

$$\mathbb{P}\Big\{\Big|\frac{X^N(T)}{N} - f^*\Big| > 4N^{-1/2+\delta} + Me^{-\alpha_S T}\Big\} = O(e^{-\alpha_U N^\delta}),$$

where $M = 4\max\Big(C_2, 1\Big)\Big(2|\mathcal{I}| + |\mathcal{S}|\Big)$ and $C_2 = \max\{x_i^-(0), \lambda_i/\theta_i\}$ as defined in Lemma 21.

Proof. As we discussed in Section A.2, the problem with some or all $f_j^* = 0$, can be converted to another network problem with all f_j^* values positive. Hence, it is sufficient to prove the lemma statement for the case $f_j^* > 0, \forall j \in \mathcal{S}$. Set

$$T' = t_4 + \alpha_M^{-1} \log\Big(\frac{\iota_2}{\iota_1} \max\big(1, \big|\frac{X^N(0)}{N} - f^*\big|\big)\Big)$$

Consider the constant M as defined in Lemma 21. We first consider the case $T \geq T' \log(MN^{1/2-\delta})$. From Lemma 21, there exists a positive constant α' such that

$$\mathbb{P}\Big\{\Big|\frac{X^N(T - T'\log(MN^{1/2-\delta}))}{N} - f^*\Big| < M\Big\} \leq 1 - O(e^{-\alpha' N}). \tag{A.36}$$

For $k \leq \log(MN^{1/2-\delta})$, define

$$X^{N,k} = X^N(T + kT' - T'\log(MN^{1/2-\delta}))$$

Setting the time reference at kT' and $\alpha_U = \min(\alpha e^{-2CT'}, \alpha')$,

$$\begin{aligned}\mathbb{P}\Big\{\Big|\frac{X^{N,k+1}}{N} - \Psi\Big(\frac{X^{N,k}}{N}, T'\Big)\Big| > N^{-1/2+\delta}\Big\} &\leq \mathbb{P}\Big\{\Big|\frac{\tilde{X}^N(T)}{N}\Big|_{(T')} > \frac{N^{-1/2+\delta}}{2e^{CT'}}\Big\} \\ &= O(e^{-\alpha_U N^{2\delta}}).\end{aligned}$$

where the last inequality follows from Lemma 25. Also, from Lemma 16 we know that

$$\Big|\Psi\Big(\frac{X^{N,k}}{N}, T'\Big) - f^*\Big| \leq \Big|\frac{X^{N,k}}{N} - f^*\Big| e^{-1}.$$

As a result,

$$\mathbb{P}\Big\{\Big|\frac{X^{N,k+1}}{N} - f^*\Big| \leq N^{-1/2+\delta} + \Big|\frac{X^{N,k+1}}{N} - f^*\Big| e^{-1}\Big\} = 1 - O(e^{-\alpha_U N^{2\delta}})$$

Iteratively applying this bound yields

$$\begin{aligned}&\mathbb{P}\Bigg\{\Big|\frac{X^{N,\log(MN^{1/2-\delta})}}{N} - f^*\Big| \leq N^{-1/2+\delta}\Big(\sum_{l=0}^{\log MN^{1/2+\delta}} e^{-l}\Big) + \frac{\Big|\frac{X^{N,0}}{N} - f^*\Big|}{MN^{1/2-\delta}}\Bigg\} \\ \leq\ & 1 - \log(MN^{1/2-\delta}) O(e^{-\alpha_U N^{2\delta}})\end{aligned}$$

Combining with (A.36) yields

$$\mathbb{P}\Bigg\{\Big|\frac{X^{N,\log(MN^{1/2-\delta})}}{N} - f^*\Big| \leq 3N^{-1/2+\delta}\Bigg\} = 1 - \log(MN^{1/2-\delta}) O(e^{-\alpha_U N^{2\delta}})$$

As our choice of $\delta > 0$ can be arbitrarily small we conclude

$$\mathbb{P}\Big\{|\frac{X^N(T)}{N} - f^*| > 3N^{-1/2+\delta}\Big\} = O(e^{-\alpha_U N^{\delta}}).$$

Next, consider the case $T \leq T' \log(MN^{1/2-\delta})$ similarly for $0 \leq k \leq \lfloor T/T' \rfloor$, define

$$X^{N,k} = X^N(kT' + \{T/T'\}T')$$

where $\{T/T'\}$ denotes the fractional part of $\{T/T'\}$. Applying the same iterative procedure we presented for the previous case we conclude

$$\begin{aligned}&\mathbb{P}\Big\{\Big|\frac{X^N(T)}{N} - f^*\Big| \leq N^{-1/2+\delta}\Big(\sum_{l=0}^{\lfloor T/T' \rfloor} e^{-l}\Big) + \frac{\Big|\frac{X^{N,0}}{N} - f^*\Big|}{e^{\lfloor T/T' \rfloor}}\Big\}\\ \leq\ & 1 - (\lfloor T/T' \rfloor)O(e^{-\alpha_U N^{2\delta}})\\ \leq\ & 1 - \log(MN^{1/2-\delta})O(e^{-\alpha_U N^{2\delta}})\end{aligned}$$

As our choice of $\delta > 0$ can be arbitrarily small, setting $\alpha_S = 1/T'$ yields

$$\mathbb{P}\Big\{|\frac{X^N(T)}{N} - f^*| > 4N^{-1/2+\delta} + Me^{-\alpha_S T}\Big\} = O(e^{-\alpha_U N^{\delta}}).$$

This conclude the proof. □

A.11 Lemmas used in the proof of Theorem 3

First, we apply moderate deviation theory (Theorem 3.7.1 of Dembo and Zeitouni 2009) and Schilder's theorem for Poisson processes (Exercise 5.2.12 Dembo and Zeitouni 2009) to obtain the following result.

Lemma 23. *(Moderate deviation analysis for Poisson processes) For a unit rate Poisson process $A(t)$, set $\tilde{A} = U(t) - t$. So, for $-1/2 < \vartheta < 0$ we have*

$$\mathbb{P}\Big\{\frac{\|\tilde{A}\|_T}{T} > \epsilon T^{\vartheta}\Big\} = e^{-TI_1(\epsilon T^{\vartheta})} = O(e^{-1/2\epsilon^2 T^{1+2\vartheta}})$$

Here, $I_1(.)$ is a rate function.

Then we apply moderate deviation theory and Mogulskii's theorem on Bernoulli processes (Theorem 5.1.2 in Dembo and Zeitouni 2009) to obtain the following result.

Lemma 24. *(Moderate deviation analysis for Bernoulli trial) For a unit Bernouli trial ϕ_{ji} with success rate p_{ji}, set $\tilde{\phi}_{ji} = \phi_{ji}(N) - p_{ji}N$. Then, for $-1/2 < \vartheta < 0$ we have*

$$\mathbb{P}\Big\{ \sup_{1 \le n \le N} \Big| \frac{\tilde{\phi}_{ji}(n)}{N} \Big| > \epsilon N^{\vartheta} \Big\} = e^{-NI_2(\epsilon N^{\vartheta})} = O(e^{-\frac{1}{2}(p_{ji}-p_{ji}^2)^{-1}\epsilon^2 N^{1+2\vartheta}})$$

Here, $I_2(.)$ is a rate function.

The following is the key result used in the proof of Theorem 3.

Lemma 25. *There exists a positive constant $\alpha > 0$, such that for $-1/2 < \vartheta \le 0$, we have*

$$\mathbb{P}\left\{ \Big\| \frac{\tilde{X}^N}{NT} \Big\|_T > \frac{\epsilon (NT)^{\vartheta}}{2} \right\} = O\big(e^{-\alpha\epsilon^2 (NT)^{1+2\vartheta}}\big).$$

For constant $T \ge 0$, taking $\alpha' = \alpha/2T^2$, $\vartheta = 0$ and $\epsilon' = \epsilon T$, it gives

$$\mathbb{P}\Big\{ \Big\| \frac{\tilde{X}^N}{N} \Big\|_T > \epsilon' \Big\} = O\big(e^{-2\alpha'\epsilon^2 N}\big).$$

Proof. We prove this result in two steps. First, we proves that the input to the Poisson processes that define the system dynamics are bounded. Then, we prove the probability that a Poisson process with bounded input generates an extremely large output is exponentially small. According to the law of total probability, to prove the lemma statement it suffices to prove that there exists $\alpha > 0$ such that the following three statements hold.

$$\mathbb{P}\left\{ \Big\| \frac{\tilde{X}_i^N}{NT} \Big\|_T > \frac{\epsilon (NT)^{\vartheta}}{2|\mathcal{S}| + 4|\mathcal{I}|} \right\} = O(e^{-\alpha\epsilon^2 (NT)^{1+2\vartheta}}); \qquad \forall i \in \mathcal{S}; \forall -1/2 < \vartheta \le 0$$

$$\mathbb{P}\left\{ \Big\| \frac{\tilde{X}_{ji}^N}{NT} \Big\|_T > \frac{\epsilon (NT)^{\vartheta}}{2|\mathcal{S}| + 4|\mathcal{I}|} \right\} = O(e^{-\alpha\epsilon^2 (NT)^{1+2\vartheta}}); \qquad \forall ji \in \mathcal{I}; \forall -1/2 < \vartheta \le 0$$

$$\mathbb{P}\left\{ \Big\| \frac{\tilde{Z}_{kl}^N}{NT} \Big\|_T > \frac{\epsilon (NT)^{\vartheta}}{2|\mathcal{S}| + 4|\mathcal{I}|} \right\} = O(e^{-\alpha\epsilon^2 (NT)^{1+2\vartheta}}); \qquad \forall kl \in \mathcal{I}; \forall -1/2 < \vartheta \le 0$$

First, for all single servers i the law of total probabilities yields

$$\begin{aligned}
& \mathbb{P}\left\{\left\|\frac{\tilde{X}_i^N}{NT}\right\|_T > \frac{\epsilon(NT)^{\vartheta}}{2|\mathcal{S}|+4|\mathcal{I}|}\right\} \\
\leq\ & \mathbb{P}\left\{\left\|\tilde{X}_i^N\right\|_T > \frac{\epsilon(NT)^{1+\vartheta}}{2|\mathcal{S}|+4|\mathcal{I}|}\right\} \\
\leq\ & \mathbb{P}\left\{\sup_{0\leq t\leq T}|\tilde{A}_i(N\lambda_i t)| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\right\} \\
& +\sum_{ji\in\mathcal{I}}\mathbb{P}\left\{\sup_{0\leq t\leq T}\left|\tilde{F}_{ji}(\mu_{ji}\int_0^t X_{ji}^{(N)}(s)ds)\right| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\right\} \\
& +\sum_{ji\in\mathcal{I}}\mathbb{P}\left\{\sup_{0\leq t\leq T}\left|\tilde{E}_{ji}(\mu_{ji}\int_0^t Z_{ji}^{(N)}(s)ds)\right| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\right\} \\
& +\mathbb{P}\left\{\sup_{0\leq t\leq T}\left|\tilde{H}_i(\gamma_i\int_0^t X_i^{(N)+}(s)ds)\right| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\right\} \\
& +\mathbb{P}\left\{\sup_{0\leq t\leq T}\left|\tilde{G}_i(\theta_i\int_0^t X_i^{(N)-}(s)ds)\right| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\right\}
\end{aligned}$$

where the last inequality follows from (1.15). Next, we note that $X_j^{(N)+}$, $X_{ji}^{(N)}$, and $Z_{ji}^{(N)}$ are positive and their sum,

$$\sum_{j\in\mathcal{S}}X_j^{(N)+}+\sum_{ji\in\mathcal{I}}X_{ji}^{(N)}+\sum_{ji\in\mathcal{I}}Z_{ji}^{(N)}=N$$

, is invariant over time and equals the market size N. We can upper bound the above equation as follows:

$$\begin{aligned}
& \mathbb{P}\left\{\left\|\frac{\tilde{X}_i^N}{NT}\right\|_T > \frac{\epsilon(NT)^{\vartheta}}{2|\mathcal{S}|+4|\mathcal{I}|}\right\} \\
\leq\ & \mathbb{P}\left\{\sup_{0\leq t\leq T}|\tilde{A}_i(N\lambda_i t)| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\right\} \\
& +\sum_{ji\in\mathcal{I}}\mathbb{P}\left\{\sup_{0\leq t\leq T}\left|\tilde{F}_{ji}(\mu_{ji}Nt)\right| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\right\} \\
& +\sum_{ji\in\mathcal{I}}\mathbb{P}\left\{\sup_{0\leq t\leq T}\left|\tilde{E}_{ji}(\mu_{ji}Nt)\right| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\right\} \\
& +\mathbb{P}\left\{\sup_{0\leq t\leq T}\left|\tilde{H}_i(\gamma_i Nt)\right| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\right\} \\
& +\mathbb{P}\left\{\sup_{0\leq t\leq T}\left|\tilde{G}_i(\theta_i\int_0^t X_i^{(N)-}(s)ds)\right| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\right\}
\end{aligned}$$

Define

$$\alpha_1 = \frac{1}{72(|\mathcal{S}| + 2|\mathcal{I}|)^4 \max(\max_i \lambda_i, \max_{ji} \mu_{ji}, \max_i \gamma_i)}.$$

Then, applying Lemma 23 to the first four terms yields

$$\begin{aligned} \mathbb{P}\left\{\Big\|\frac{\tilde{X}_i^N}{NT}\Big\|_T > \frac{\epsilon(NT)^{1+\vartheta}}{2|\mathcal{S}| + 4|\mathcal{I}|}\right\} \;\; &\leq \;\; O\Big(e^{-\alpha_1\epsilon^2(NT)^{1+2\vartheta}}\Big) \\ &\quad +\mathbb{P}\bigg\{\sup_{0\leq t\leq T}\Big|\tilde{G}_i(\theta_i\int_0^t X_i^{(N)-}(s)ds)\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}| + 2|\mathcal{I}|)^2}\bigg\}. \end{aligned}$$

Next, set $\chi = 2(\max_i|\theta_i| + \max_i|\lambda_i|)$ and use the law of total probability to rewrite the last term as follows

$$\mathbb{P}\bigg\{\sup_{0\leq t\leq T}\Big|\tilde{G}_i(\theta_i\int_0^t X_i^{(N)-}(s)ds)\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}| + 2|\mathcal{I}|)^2}\bigg\}$$

$$\leq \mathbb{P}\left\{\sup_{0\leq t\leq T}\Big|\tilde{G}_i(\theta_i\textstyle\int_0^t X_i^{(N)-}(s)ds)\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\Big|\theta_i\textstyle\int_0^t X_i^{(N)-}(s)ds > \chi\theta_i NT^2\right\}\mathbb{P}\left\{\theta_i\textstyle\int_0^t X_i^{(N)-}(s)ds > \chi\theta_i NT^2\right\} \quad \text{(A.37)}$$

$$+\mathbb{P}\left\{\sup_{0\leq t\leq T}\Big|\tilde{G}_i(\theta_i\textstyle\int_0^t X_i^{(N)-}(s)ds)\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\Big|\theta_i\textstyle\int_0^t X_i^{(N)-}(s)ds \leq \chi\theta_i NT^2\right\}\mathbb{P}\left\{\theta_i\textstyle\int_0^t X_i^{(N)-}(s)ds \leq \chi\theta_i NT^2\right\} \quad \text{(A.38)}$$

Next, we upper bound (A.37) by the event probability. Furthermore, we use the system dynamics (1.12) to upper bound the event probability as follows

$$\begin{aligned} &\mathbb{P}\left\{\theta_i\int_0^t X_i^{(N)-}(s)ds > \chi\theta_i NT^2\right\} \\ \leq \;\; &\mathbb{P}\left\{\theta_i\int_0^t \Big(A_i(N\lambda_i s) + H_i\Big(\gamma_i\int_0^t X_i^{(N)+}(s)ds\Big)\Big)ds > \chi\theta_i NT^2\right\} \end{aligned}$$

Noting that $X_i^{(N)+} \leq N$, we conclude

$$\begin{aligned} &\mathbb{P}\left\{\theta_i\int_0^t X_i^{(N)-}(s)ds > \chi\theta_i NT^2\right\} \\ \leq \;\; &\mathbb{P}\left\{\theta_i T\sup_{0\leq s\leq T} A_i(N\lambda_i s) + \theta_i T\sup_{0\leq s\leq T} H_i(\gamma_i Ns) > \chi\theta_i NT^2\right\} \end{aligned}$$

Define

$$\alpha_2 = \frac{\chi^2}{8\max(\max_i \lambda_i, \max_i \gamma_i)}$$

Therefore, Lemma 23 yields

$$\mathbb{P}\left\{\theta_i\int_0^t X_i^{(N)-}(s)ds > \chi\theta_i NT^2\right\} = O(e^{-\alpha_2\epsilon^2 NT})$$

On the other hand, we upper bound (A.38) by the conditional probability as follows

$$\begin{aligned}
&\mathbb{P}\Big\{\sup_{0\leq t\leq T}\Big|\tilde{G}_i(\theta_i\int_0^t X_i^{(N)-}(s)ds)\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\Big|\theta_i\int_0^t X_i^{(N)-}(s)ds \leq \chi\theta_i NT^2\Big\} \\
\leq \quad &\mathbb{P}\Big\{\sup_{0\leq s\leq \chi T^2}\Big|\tilde{G}_i(\theta_i Ns)\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\Big\}
\end{aligned}$$

Define

$$\alpha_3 = \frac{1}{72T\chi\theta_i(|\mathcal{S}|+2|\mathcal{I}|)^4}$$

Using Lemma 23 we can upper bound (A.38) by $O(e^{-\alpha_3\epsilon^2(NT)^{1+2\vartheta}})$. Also, for infinite servers $ji \in \mathcal{I}$ we write

$$\begin{aligned}
&\mathbb{P}\left\{\Big\|\frac{\tilde{X}_{ji}^{(N)}}{NT}\Big\|_T > \frac{\epsilon(NT)^{\vartheta}}{2|\mathcal{S}|+4|\mathcal{I}|}\right\} \\
\leq \;&\mathbb{P}\left\{\sup_{0\leq t\leq T}\Big|\tilde{F}_{ji}(\mu_{ji}\int_0^t X_{ji}^{(N)}(s)ds)\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\right\} \\
&+\sum_{kj\in\mathcal{I}}\mathbb{P}\left\{p_{ji}\sup_{0\leq t\leq T}\Big|\tilde{F}_{kj}(\mu_{kj}\int_0^t X_{kj}^{(N)}(s)ds))\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\right\} \\
&+\sum_{kj\in\mathcal{I}}\mathbb{P}\left\{p_{ji}\sup_{0\leq t\leq T}\Big|\tilde{E}_{kj}(\mu_{kj}\int_0^t Z_{kj}^{(N)}(s)ds)\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\right\} \\
&+\mathbb{P}\Bigg\{\sup_{0\leq t\leq T}\Big|\tilde{\phi}_{ji}((\sum_{kj\in\mathcal{I}}F_{kj}(\mu_{kj}\int_0^t X_{kj}^{(N)}(s)ds)) \\
&+(\sum_{kj\in\mathcal{I}}E_{kj}(\mu_{kj}\int_0^t Z_{kj}^{(N)}(s)ds)) + X_j^{(N)+}(0) \\
&-X_j^{(N)+}(t) - H_j(\gamma_j\int_0^t X_j^{(N)+}(s)ds))\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6|\mathcal{S}+\mathcal{I}_1+\mathcal{I}_2|^2}\Bigg\}.
\end{aligned}$$

Similar to the case for single servers we can upper bound the first three elements in the right hand side by $O(e^{-\alpha_1\epsilon^2(NT)^{1+2\vartheta}})$. Moreover, we can upper bound the last element by the following expression

$$\mathbb{P}\Big\{\sup_{0\leq t\leq T}\Big|\sum_{kj\in\mathcal{I}}F_{kj}(\mu_{kj}Nt)+\sum_{kj\in\mathcal{I}}E_{kj}(\mu_{kj}Nt)+X_j^{(N)+}(0)\Big| \geq 2NT(1+\sum_{kj\in\mathcal{I}}\mu_{kj})\Big\} \tag{A.39}$$

$$+\;\mathbb{P}\Big\{\sup_{0\leq t\leq T}\Big|\tilde{\phi}_{ji}\big(2NT(1+\sum_{kj\in\mathcal{I}}\mu_{kj})\big)\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\Big\} \tag{A.40}$$

We can upper bound (A.39) using Lemma 23 for $T \geq 1$ by $O(e^{-NT \min_{kj}(\mu_{kj})/2})$. Moreover, define

$$\alpha_4 = \frac{1}{144(1+\sum_{kj}\mu_{kj})^2(|\mathcal{S}|+2|\mathcal{I}|)^4 \max_{ji}(p_{ji}(1-p_{ji}))}$$

Applying Lemma 24, we can upper bound (A.40) as the following expression

$$\mathbb{P}\Big\{\sup_{0\leq n\leq N}\Big|\tilde{\phi}_{ji}(2n(1+T\sum_{kj\in\mathcal{I}}\mu_{kj}))\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\Big\} = O(e^{-\alpha_4\epsilon^2(NT)^{1+2\vartheta}})$$

Lastly, for all infinite servers that model empty vehicle trips from node j to node i, we have

$$\begin{aligned}
\mathbb{P}\Big\{\Big\|\frac{\tilde{Z}_{ji}^{(N)}}{NT}\Big\|_T > \frac{\epsilon(NT)^{\vartheta}}{2|\mathcal{S}|+4|\mathcal{I}|}\Big\} \;\leq\; & \mathbb{P}\Big\{\sup_{0\leq t\leq T}\Big|\tilde{E}_{ji}\Big(\mu_{ji}\int_0^t Z_{ji}^{(N)}(s)ds\Big)\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\Big\} \\
& +\mathbb{P}\Big\{\sup_{0\leq t\leq T}\Big|\tilde{\sigma}_{ji}\Big(H_j\Big(\gamma_j\int_0^t X_j^{(N)+}(s)ds\Big)\Big)\Big| \\
& \qquad . > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\Big\} \\
& +\mathbb{P}\Big\{\sup_{0\leq t\leq T}\Big|q_{ji}\tilde{H}_j\Big(\gamma_j\int_0^t X_j^{(N)+}(s)ds\Big)\Big| > \frac{\epsilon(NT)^{1+\vartheta}}{6(|\mathcal{S}|+2|\mathcal{I}|)^2}\Big\}.
\end{aligned}$$

Following the same procedure for single servers and infinite servers, the first term is upper bounded by $O(e^{-\alpha_1\epsilon^2(NT)^{1+2\vartheta}})$. Define

$$\alpha_5 = \frac{1}{144(|\mathcal{S}|+2|\mathcal{I}|)^4 \max_{ji}(q_{ji}(1-q_{ji}))}$$

The second term is upper bounded by

$$O(e^{-\frac{1}{2}NT\min_j(\gamma_j)}) + O(e^{-\alpha_4\epsilon^2(NT)^{1+2\vartheta}})$$

Also, the third term is upper bounded by $O(e^{-\alpha_1\epsilon^2(NT)^{1+2\vartheta}})$. We set

$$\alpha = \max(\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5, 1/2\min_{kj}\mu_{kj}, 1/2\min_j\gamma_j)$$

This concludes the proof. □

A.12 Proof of Theorem 4

First, we need several lemmas.

Lemma 26. *Consider the system defined in Theorem 4, and set*

$$\mu_{\max} = \max\Big(\max_{ji}\mu_{ji}, \frac{2}{\min_i \sum_{k,\ ik\in\mathcal{I}} \mu_{ik} e_{ik}}, 1\Big).$$

There exists $\alpha^{''} > 0$ such that for an arbitrary constant T_c and for any $T \le T_c$ and any $\delta \ge \epsilon$

$$\mathbb{P}\Big\{\sup_{T\le T_c}\Big|\frac{\int_0^T X_i^{(N)+}(t)dt}{T}\Big| \ge 2\mu_{\max}N^{1/2+\delta}\Big|X_i^{(N)+}(0) \le \mu_{\max}N^{1/2+\delta}\Big\} = O\big(e^{-\alpha^{''}N^{\delta+\varepsilon}}\big)$$

Proof. We first upper bound the lemma statement as follows:

$$\begin{aligned}
&\mathbb{P}\Big\{\sup_{T\le T_c}\Big|\frac{\int_0^T X_i^{(N)+}(t)dt}{T}\Big| \ge 2\mu_{\max}N^{1/2+\delta}\Big|X_i^{(N)+}(0) \le \mu_{\max}N^{1/2+\delta}\Big\} \\
\le\ &\mathbb{P}\Big\{\sup_{t\le T_c}\Big|X_i^{(N)+}(t)\Big| \ge 2\mu_{\max}N^{1/2+\delta}\Big|X_i^{(N)+}(0) \le \mu_{\max}N^{1/2+\delta}\Big\}
\end{aligned}$$

Assume there exists $T \le T_c$ such that $X_i^{(N)+}(T) \ge \mu_{\max}N^{1/2+\delta}$ and define

$$T_2 = \inf\Big\{T\Big|X_i^{(N)+}(T) \ge 2\mu_{\max}N^{1/2+\delta}, 0 \le T \le T_c\Big\}$$

Also, define

$$T_1 = \sup\Big\{T'\Big|X_i^{(N)+}(T') \le \mu_{\max}N^{1/2+\delta}, 0 \le T' \le T_2\Big\}$$

Then, we apply (1.12) to yield

$$\begin{aligned}
X_i^{(N)}(T_2) \ \le\ & X_i^{(N)}(T_1) + \sum_{ji\in\mathcal{I}} F_{ji}\Big(\mu_{ji}\int_{T_1}^{T_2} X_{ji}^{(N)}(s)ds\Big) + \sum_{ji\in\mathcal{I}} E_{ji}\Big(\mu_{ji}\int_{T_1}^{T_2} Z_{ji}^{(N)}(s)ds\Big) \\
& -H_i\Big(\gamma_i\int_{T_1}^{T_2} X_i^{(N)+}(s)ds\Big).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mu_{\max}N^{1/2+\delta} \ \le\ & \sum_{ji\in\mathcal{I}} F_{ji}\Big(\mu_{ji}\int_{T_1}^{T_2} X_{ji}^{(N)}(s)ds\Big) + \sum_{ji\in\mathcal{I}} E_{ji}\Big(\mu_{ji}\int_{T_1}^{T_2} Z_{ji}^{(N)}(s)ds\Big) \\
& -H_i\Big(\gamma_i\int_{T_1}^{T_2} X_i^{(N)+}(s)ds\Big).
\end{aligned}$$

We upper bound the right hand side by separating the stochastic and deterministic terms as follows

$$\begin{aligned}\mu_{\max} N^{1/2+\delta} \;\; \leq \;\; & \sum_{ji\in\mathcal{I}} \tilde{F}_{ji}\Big(\mu_{ji}\int_{T_1}^{T_2} X_{ji}^{(N)}(s)ds\Big) + \sum_{ji\in\mathcal{I}} \tilde{E}_{ji}\Big(\mu_{ji}\int_{T_1}^{T_2} Z_{ji}^{(N)}(s)ds\Big) \\ & - \tilde{H}_i\Big(\gamma_i\int_{T_1}^{T_2} X_i^{(N)+}(s)ds\Big) + \sum_{ji\in\mathcal{I}}\Big(\mu_{ji}\int_{T_1}^{T_2} X_{ji}^{(N)}(s)ds\Big) \\ & + \Big(\mu_{ji}\int_{T_1}^{T_2} Z_{ji}^{(N)}(s)ds\Big) - \Big(\gamma_i\int_{T_1}^{T_2} X_i^{(N)+}(s)ds\Big).\end{aligned}$$

Recall from Lemma 10 that $\sum_{ji} Z_{ji}^{(N)}(s) + X_{ji}^{(N)}(s)$ is bounded above by N. Also, by definition we have $X_i^{(N)+}(s) > 0$ for $s \in [T_1, T_2]$. Hence for large N and $\delta \geq \varepsilon$ we conclude

$$\begin{aligned}& \mu_{\max} N^{1/2+\delta} + \mu_{\max} N^{1+\delta-\epsilon}(T_2 - T_1) - \mu_{\max} N(T_2 - T_1) \\ \leq \;\; & \sum_{ji\in\mathcal{I}} \Big\|\tilde{F}_{ji}\Big\|_{\mu_{ji}\int_0^{T_c} X_{ji}^{(N)}(s)ds} + \sum_{ji\in\mathcal{I}} \Big\|\tilde{E}_{ji}\Big\|_{\mu_{ji}\int_0^{T_c} Z_{ji}^{(N)}(s)ds} + \Big\|\tilde{H}_i\Big\|_{\gamma_i\int_0^{T_2} X_i^{(N)+}(s)ds}. \quad \text{(A.41)}\end{aligned}$$

Note that the probability that a $T_2 < T_c$ exists such that $X_i^{(N)+}(T_2) \geq N^{1/2+\delta}$ is upper bounded by the probability that (A.41) holds. We refer to the event in (A.41) as B_i and upper bound its probability. First, note that Lemma 23 yields

$$\mathbb{P}\Big\{N^{1/2+\delta}/2 \;\; \leq \;\; \sum_{ji\in\mathcal{I}} \Big\|\tilde{F}_{ji}\Big\|_{\mu_{ji}N(T_c)} + \sum_{ji\in\mathcal{I}} \Big\|\tilde{E}_{ji}\Big\|_{\mu_{ji}N(T_c)}\Big\} = O(e^{-N^{2\delta}/T_c|\mathcal{I}|^2 16 \max_{ji}\mu_{ji}})$$

Next, Lemma 23 yields

$$\begin{aligned}& \mathbb{P}\Big\{\tilde{H}_i\Big(\gamma_i\int_0^{T_2} X_i^{(N)+}(s)ds\Big) > N^{1/2+\delta}/2\Big\} \\ \leq \;\; & \mathbb{P}\Big\{\tilde{H}_i\Big(\mu_{\max} T_c N^{1+\delta-\varepsilon}\Big) > N^{1/2+\delta}/2\Big\} = O\big(e^{-N^{\delta+\varepsilon}/8T_c\mu_{\max}}\big)\end{aligned}$$

Therefore, there exists constant α'' such that

$$\mathbb{P}\{B_i\} \leq O(e^{-\alpha'' N^{\delta+\varepsilon}})$$

This concludes the proof. □

Lemma 27. *Consider γ_i and $\mu_{\max}$ defined according to Theorem 4 and Lemma 26, respectively. There exists $\alpha'' > 0$ such that for an arbitrary constant T_c and for any $T \leq T_c$ and any $\delta \geq \epsilon$*

$$\mathbb{P}\Big\{X_i^{(N)+}(T_c) \geq \mu_{\max} N^{1/2+\delta}\Big| X_i^{(N)+}(0) \leq \mu_{\max} N^{1/2+\delta}\Big\} = O\big(e^{-\alpha'' N^{\delta+\varepsilon}}\big)$$

Proof. We consider two cases. First, if $X_i^{(N)+}(T) \geq \mu_{\max} N^{1/2+\delta}/2, \forall 0 \leq T \leq T_c$, then (1.12) yields

$$\begin{aligned} X_i^{(N)}(T_c) \;\; &\leq \;\; X_i^{(N)}(0) + \sum_{ji \in \mathcal{I}} F_{ji}\bigg(\mu_{ji} \int_0^{T_c} X_{ji}^{(N)}(s)ds\bigg) + \sum_{ji \in \mathcal{I}} E_{ji}\bigg(\mu_{ji} \int_0^{T_c} Z_{ji}^{(N)}(s)ds\bigg) \\ &\quad - H_i\bigg(\gamma_i \int_0^{T_c} X_i^{(N)+}(s)ds\bigg). \end{aligned}$$

repeating the same approach as we discussed in the proof for Lemma 26 it is straightforward to show that this event is bounded by $O(e^{-\alpha'' N^{\delta+\varepsilon}})$ for some $\alpha'' > 0$. In the next case, there exists $T < T_c$ such that $X_i^{(N)+}(T) \leq \mu_{\max} N^{1/2+\delta}/2$. Define the stopping time T_1 as follows

$$T_1 = \inf_{T \leq T_c} \Big\{ T : X_i^{(N)+}(T) \leq \mu_{\max} N^{1/2+\delta}/2 \Big\}$$

Then, if there exists a $T < T_c$ such that $X_i^{(N)+}(T) \leq N^{1/2+\delta}/4$ the event in the lemma statement is upper bounded by

$$\mathbb{P}\Big\{ X_i^{(N)+}(T_c) \geq \mu_{\max} N^{1/2+\delta} \Big| X_i^{(N)+}(T_1) \leq \mu_{\max} N^{1/2+\delta}/2 \Big\}$$

However, the proof for Lemma 26 also implies this probability is in the order $O(e^{-\alpha'' N^{\delta+\varepsilon}})$ for some positive α''. This concludes the proof. □

Lemma 28. *Consider γ_i and $\mu_{\max}$ defined according to Theorem 4 and Lemma 26, respectively. Then for any positive constant $T_c \geq 1$ and for $\delta \geq \varepsilon$ there exists positive constants $t_6 \leq T_c$ and α_1 such that*

$$\mathbb{P}\Big\{ X_i^{(N)}(t_6) \leq \mu_{\max} N^{1/2+\delta}, \forall i \in \mathcal{S} \Big\} = 1 - O(e^{-\alpha_1 N}), \forall i \in \mathcal{S}.$$

Proof. Seeking contradiction, assume that $\forall t_6 \leq T_c$ for some $i \in \mathcal{S}$ we have

$$X_i^{(N)}(t_6) \geq \mu_{\max} N^{1/2+\delta}, \forall i \in \mathcal{S}$$

Thus, for at least one $j \in \mathcal{S}$ we have

$$\gamma_j \int_0^{T_c} X_j^{(N)+}(s)ds \geq 2NT_c.$$

Then, Lemma 23 yields

$$\begin{aligned}
& \mathbb{P}\Big\{H_j\Big(\gamma_j\int_0^{T_c} X_j^{(N)+}(s)ds\Big) \geq \frac{5}{3}NT_c\Big\} \\
\geq \;& 1-\mathbb{P}\Big\{\tilde{H}_j\Big(\gamma_j\int_0^{T_c} X_j^{(N)+}(s)ds\Big) \geq \frac{1}{3}\gamma_j\int_0^{T_c} X_j^{(N)+}(s)ds\Big\} \\
= \;& 1-O(e^{-\frac{1}{9}NT_c}).
\end{aligned}$$

As such,

$$\begin{aligned}
& \mathbb{P}\Bigg\{\sigma_{ji}\Bigg(H_j\Big(\gamma_j\int_0^{T_c} X_j^{(N)+}(s)ds\Big)\Bigg) \geq \frac{4}{3}NT_c\Bigg\} \\
\leq \;& \mathbb{P}\Bigg\{\sigma_{ji}\Bigg(H_j\Big(\gamma_j\int_0^{T_c} X_j^{(N)+}(s)ds\Big)\Bigg) \geq \frac{4}{3}NT_c\Big| H_j\Big(\gamma_j\int_0^{T_c} X_j^{(N)+}(s)ds\Big) \geq \frac{5}{3}NT_c\Bigg\} \\
& -O(e^{-\frac{1}{9}NT_c}) \\
\leq \;& 1-\mathbb{P}\Bigg\{\tilde{\sigma}_{ji}\Bigg(H_j\Big(\gamma_j\int_0^{T_c} X_j^{(N)+}(s)ds\Big)\Bigg) \geq \frac{1}{5}H_j\Big(\gamma_j\int_0^{T_c} X_j^{(N)+}(s)\Big) \\
& \qquad .\Big| H_j\Big(\gamma_j\int_0^{T_c} X_j^{(N)+}(s)ds\Big) \geq \frac{5}{3}NT_c\Bigg\} \\
& -O(e^{-\frac{1}{9}NT_c}) = 1-O(e^{-\alpha_1 N}).
\end{aligned}$$

where α_1 is a positive constant. As a result, to the system dynamics (1.14) we have with probability at least $1-O(e^{-\alpha_1 N})$

$$\begin{aligned}
Z_{ji}^{(N)}(T_c) \;&=\; Z_{ji}^{(N)}(0)+\sigma_{ji}\Bigg(H_j\Big(\gamma_j\int_0^{T_c} X_j^{(N)+}(s)ds\Big)\Bigg) - E_{ji}\Big(\mu_{ji}\int_0^{T_c} Z_{ji}^{(N)}(s)ds\Big) \\
&\geq\; \frac{4}{3}NT_c - NT_c - \|\tilde{E}\|_{\mu_{ji}NT_c}
\end{aligned}$$

Also Lemma 23 yields

$$\mathbb{P}\Big\{\|\tilde{E}\|_{\mu_{ji}NT_c} \geq NT_c\Big\} = O(e^{-\frac{NT_c}{2\mu_{\max}}})$$

On the other hand we know that $Z_{ji}^{(N)}(T_c) \leq N$. This contradiction concludes the proof. □

Lemma 29. *Consider γ_i and $\mu_{\max}$ defined according to Theorem 4 and Lemma 26, respectively. Also, assume the initial system state satisfies $X_i^{(N)}(0) \leq \mu_{\max}N^{1/2+\delta}$, then there exists a positive* *constant $C_3 > 0$ such that for an arbitrary constant T_c we conclude*

$$\mathbb{P}\Big\{\Big|\frac{X^{(N)}(T_c)}{N} - \psi\Big(\frac{X^{(N)}(0)}{N}, T_c\Big)\Big| \geq e^{C_3 T_c}\Big\|\frac{\tilde{X}^{(N)}}{N}\Big\|_{T_c}\Big\} = O(e^{-\alpha'' N^{2\varepsilon}})$$

Proof. We adopt the results in Lemmas 26 and 27, to substitute the system dynamics equations with upper and lower bounds inequalities that hold with a high probability of at least $1-O(e^{-N^{2\varepsilon}})$. For example, from (1.14) we conclude with probability at least $1 - O(e^{-N^{2\varepsilon}})$ for any $t \leq T_c$ we have

$$Z_{ji}^{(N)}(t) \leq Z_{ji}^{(N)}(0) + \sigma_{ji}\left(H_j\left(2\mu_{\max}Nt\right)\right) - E_{ji}\left(\mu_{ji}\int_0^t Z_{ji}^{(N)}(s)ds\right), \quad ji \in \mathcal{I}.$$

$$Z_{ji}^{(N)}(t) \geq Z_{ji}^{(N)}(0) - \sigma_{ji}\left(H_j\left(2\mu_{\max}Nt\right)\right) - E_{ji}\left(\mu_{ji}\int_0^t Z_{ji}^{(N)}(s)ds\right), \quad ji \in \mathcal{I}.$$

Following the same procedure, we can rewrite (A.4) by substituting the terms that contain coefficients γ_i with two inequalities; i.e., with probability at least $1 - O(e^{-N^{2\varepsilon}})$ for any $t \leq T_c$ we have

$$x(t) \geq q(t) + J_1\int_0^t x(s)ds + J_2\int_0^t x^-(s)ds - J_3x^+(t) - 2\mu_{\max}t\mathbf{1}$$
$$x(t) \leq q(t) + J_1\int_0^t x(s)ds + J_2\int_0^t x^-(s)ds - J_3x^+(t) + 2\mu_{\max}t\mathbf{1}$$

Repeating the same iterative bounding approach we presented in Lemma 2 yields there exists C_3 such that

$$\mathbb{P}\Big\{\Big|\frac{X^{(N)}(T_c)}{N} - \psi\Big(\frac{X^{(N)}(0)}{N}, T_c\Big)\Big| \geq e^{C_3 T_c}\Big\|\frac{\tilde{X}^{(N)}}{N}\Big\|_{T_c}\Big\} = O(e^{-\alpha'' N^{2\varepsilon}}).$$

This concludes the proof of the result. □

Lemma 30. *If we define γ_i according to Theorem 4 and the initial system state satisfies $X_i^{(N)}(0) \leq \mu_{\max}N^{1/2+\delta}$, then there exists positive constant α''', such that for any constant T we have*

$$\mathbb{P}\left\{\Big\|\frac{\tilde{X}^{(N)}}{N}\Big\|_T > \frac{\epsilon_1(N)^{\vartheta}}{2}\right\} = O\big(e^{-\alpha''' \min(N^{2\varepsilon}, \epsilon_1^2 N^{1+2\vartheta})}\big); \qquad -1/2 \leq \vartheta \leq 0.$$

Proof. First note that Lemmas 26 and 27 yield

$$\mathbb{P}\Big\{\sup_{0\leq t\leq T}\gamma_i\int_0^t X_i^{(N)+}(s)ds > 2\mu_{\max}N^{1+\delta-\varepsilon}\Big\} = O(e^{-\alpha'' N^{\varepsilon+\delta}}), \forall\delta\geq\varepsilon.$$

Thus,

$$\mathbb{P}\Big\{\sup_{0\leq t\leq T}\tilde{H}\Big(\gamma_i\int_0^t X_i^{(N)+}(s)ds\Big) > \sup_{0\leq t\leq T}\tilde{H}\Big(2\mu_{\max}N^{1+\varepsilon-\varepsilon}t\Big)\Big\} = O(e^{-\alpha'' N^{2\varepsilon}}). \tag{A.42}$$

Repeating the same procedure as Lemma 25 and applying (A.42) we obtain

$$\mathbb{P}\left\{\Big\|\frac{\tilde{X}^{(N)}}{N}\Big\|_T > \frac{\epsilon_1 N^{\vartheta}}{2}\right\} = O\big(e^{-\min(\alpha'' N^{2\varepsilon},\alpha\epsilon_1^2 N^{1+2\vartheta})}\big); \qquad -1/2\leq\vartheta\leq 0.$$

This concludes the proof. □

Lemma 31. *Considering the system parameters introduced in Theorem 4, the LCP (A.7) has a unique solution $u^{(N)}(\Theta) = (u_{ji}^{(N)}(\Theta), v_{ji}^{(N)}(\Theta), u_j^{(N)}(\Theta))$ that satisfies the following properities.*

1. *The unique functions $u_{ji}^{(N)}(\Theta), ji\in\mathcal{I}, v_{ji}^{(N)}(\Theta), ji\in\mathcal{I}, u_j^{(N)}(\Theta), j\in\mathcal{S}$ are continuous piece-wise linear and increasing and number of pieces of each function is at most $|\mathcal{S}|$.*

2. *The right limit slope for the functions $u_j^{(N)}(\Theta), j\in\mathcal{S}$ is strictly positive unless $u_j^{(N)}(\Theta) = 0$.*

3. *The right limit slope for the functions $v_{ji}^{(N)}(\Theta), ji\in\mathcal{I}$ is strictly positive unless $v_{ji}^{(N)}(\Theta) = 0$.*

4. *The right limit slope for the functions $u_{ji}^{(N)}(\Theta), ji\in\mathcal{I}$ is strictly positive unless $u_{ji}^{(N)}(\Theta)\in\Big\{0, u_{ji}^{(N)}(1)\Big\}$.*

5. *The maximum value for the functions $u_j^{(N)}(\Theta), j\in\mathcal{S}$ is less than $N^{-1/2+\epsilon}\sum_{j,ji\in\mathcal{I}}\mu_{ji}e_{ji}$.*

6. *For all stations $j\in\mathcal{S}$ we have $\Theta^j\leq 1-\sum_{i,ji\in\mathcal{I}}e_{ji}$. Also, set $\Theta^{(\max)} = 1-\min_j\sum_{i,ji\in\mathcal{I}}e_{ji}$.*

7. *The minimum nonzero right limit slope for the functions $u_j^{(N)}(\Theta), j\in\mathcal{S}$ is ι_1 introduced in Lemma 8. So, there exists a positive constant $\iota_1^{'''}$ such that $\iota_1\geq\iota_1^{'''}N^{-1/2+\epsilon}$*

8. *The minimum nonzero right limit slope for the functions $u_{ji}^{(N)}(\Theta), ji\in\mathcal{I}, v_{ji}^{(N)}(\Theta), ji\in\mathcal{I}$ is a constant $\hat{\iota}_1$ independent of N such that*

$$\hat{\iota}_1\geq\frac{1}{|\mathcal{I}|+|\mathcal{S}|}\min\Big(\frac{\min_{ji,lm,\lambda_{ji}>0}\lambda_{ji}\mu_{lm}}{\max_{k,lm}\lambda_k\mu_{lm}}, \frac{\min_{ji,lm,q_{ji}>0}q_{ji}\mu_{lm}}{\max_{lm}\mu_{lm}}\Big)$$

9. *The maximum nonzero right limit slope for the functions $u_{ji}^{(N)}(\Theta), ji \in \mathcal{I}, v_{ji}^{(N)}(\Theta), ji \in \mathcal{I}$ is a constant ι_2 defined in Lemma 8.*

10. *The maximum nonzero right limit slope for the functions $u_j^{(N)}(\Theta), j \in \mathcal{S}$, is a constant $\iota_2^{'}$ independent of N.*

Proof. The proof for the first four parts of the remark are completely the same as Lemma 8. To prove part 5, note that (A.7e) yields that

$$q_{ji}N^{1/2-\epsilon}u_j^{(N)} \sum_{k,\ ki \in \mathcal{I}} \mu_{ki}e_{ki} \leq \mu_{ji}v_{ji}^{(N)}$$

This concludes the proof for part 5. To prove part 6, note that for $\Theta < \Theta^{(j)}$, (A.7e) yields $v_{ji}^{(N)}(\Theta) = 0$. However, (A.7f) yields $v_{ji}^{(N)}(1) = e_{ji}$. Noting that $v_{ji}^{(N)}(1) - v_{ji}^{(N)}(\Theta) \leq 1 - \Theta$ concludes the proof for this part. Parts 7, 8, 9, and 10 are derived by following the exact same procedure as presented in the proof for Lemma 8. □

Lemma 32. *Starting from an initial solution $|x_0| \leq M$, there exists positive constants t_4, α_M, ι_1' such that the solution to the dynamical system (1.16-1.18) satisfies*

$$|x_{ji}(t + t_4) - f_{ji}| \leq \frac{1}{\iota_1^{'}}\Big|1 - L(x(0))\Big|e^{-\alpha_M t}, \qquad \forall ji \in \mathcal{I}$$

$$|z_{ji}(t + t_4) - e_{ji}| \leq \frac{1}{\iota_1^{'}}\Big|1 - L(x(0))\Big|e^{-\alpha_M t}, \qquad \forall ji \in \mathcal{I}$$

Furthermore,

$$\Big|L(x(t + t_4)) - 1\Big| \leq \Big|L(x(0)) - 1\Big|e^{-\alpha_M t}$$

Proof. As the proof outline is very similar to the proof of Lemma 15 here we only present the key differences.

First, note that for some $t \leq \frac{1}{\min_{ji}\mu_{ji}}$ we have

$$x_j(t) \leq 3N^{-1/2+\epsilon}|S|\frac{\max_{ji}\mu_{ji}}{\min_{ji,q_{ji}>0} q_{ji}}, \qquad \forall j \in \mathcal{S}$$

Otherwise, according to (1.18) at least for one $ji \in \mathcal{I}$

$$z_{ji}\Big(\frac{1}{\min_{ji}\mu_{ji}}\Big) \;\; \geq \;\; -\mu_{ji}\Big(\frac{1}{\min_{ji}\mu_{ji}}\Big) + 3\mu_{ji}\Big(\frac{1}{\min_{ji}\mu_{ji}}\Big) \geq -1 + 3 > 1.$$

Next, note that the solution to dynamical system (1.16-1.18) satisfies the following equation for

$t' \geq t$,

$$x_i(t') \leq 3N^{-1/2+\epsilon}\big(|\mathcal{S}| + 2|\mathcal{I}|\big)\frac{\max_{ji}\mu_{ji}}{\min_{ji:q_{ji}>0} q_{ji}}, \qquad \forall i \in \mathcal{S} \tag{A.43}$$

To realize this, note that the derivative of x_i is computed as follows.

$$\dot{x}_i^+(t) = -\lambda_i + \sum_{ji\in\mathcal{I}} \mu_{ji}x_{ji}(t) + \sum_{ji\in\mathcal{I}} \mu_{ji}z_{ji}(t) + \theta_i x_i^-(t) - \gamma_i x_i^+(t) - \dot{x}_j^-(t), \qquad i \in \mathcal{S}$$

Therefore, for x_i that violates (A.43), $\dot{x}_i^+(t) < 0$. Thus, in a completely similar argument as we presented in the proof for Lemma 11 part (a), we conclude that (A.43) holds for $t' \geq t$. Using the transformation presented in Lemma 29, we conclude that there exists a positive constant $C_3 > 0$ such that

$$\big|\psi(x+\delta, t) - \psi(x,t)\big| \leq \delta e^{C_3 T_1} \tag{A.44}$$

The argument follows the same steps in the proof for Lemma 15. First, recall from Lemma 31 that $\hat{\iota}_1$, the minimum slope of functions $u_{ji}, v_{ji}, ji \in \mathcal{I}$, is a constant independent of N. Also, note that Lemma 9 holds for $\hat{\iota} = \iota$. Next, similar to Lemma 15 we introduce the parameters

$$\delta \leq \iota_1^2 \min_i(\Theta^{i+1} - \Theta^i)/(8e^{C_3T_2}). \tag{A.45}$$

Set the precision parameter in Proposition 2 as $\delta/2$ to obtain $T_1 = \frac{1}{\min_j \theta_j}\log(\frac{C_2}{\delta/2})$. Consider

$$\delta_1 = \frac{\delta^{(C_3/\min_j\theta_j)+1}}{2^{(C_3/\min_j\theta_j)+1}C_2^{C_3/\min_j\theta_j}}. \tag{A.46}$$

More over, consider the minimum integer d_i such that

$$\Big(\frac{16\hat{\iota}_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1 e^{CT_2} \leq \hat{\iota}_1^2/8(1-\hat{\iota}_1/8)^{d_i-2}(\Theta^{i+1} - \Theta^i) \tag{A.47}$$

Set $d = \min_{i,0\leq i\leq|\mathcal{S}|} d_i$. Also, we consider the set $\{t^{(i,j)}, 0 \leq i \leq |\mathcal{S}|, 0 \leq j \leq d_i\}$ such that

$$\begin{aligned} t^{(i,0)} &= \sum_{k=0}^{i}(T_1 + (d_k - 1)T_2) \\ t^{(i,j)} &= t^{(i,0)} + T_1 + (j-1)T_2; \quad 1 \leq j \leq d_i - 1 \end{aligned}$$

To prove the lemma, we first use induction on $i \in [0, |\mathcal{S}|+1]$ to show that for $t \geq t^{(i,0)}$ we have

$$x_{ji}^{+}(t) + \Big(\frac{16\hat{\iota}_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1\mathbf{1} \geq u_{ji}(\Theta^i)$$

$$z_{ji}^{+}(t) + \Big(\frac{16\hat{\iota}_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1\mathbf{1} \geq v_{ji}(\Theta^i)$$

and if $u_j(\Theta^i) > 0$ then for $t \geq t^{(i,0)}$ we have $x_j^+(t) \geq \mathbf{0}$. For the induction base, $i = 0$, note that

$$x^+(0) + \Big(\frac{16\hat{\iota}_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|}\delta_1 \geq u(\Theta^0) = \mathbf{0}.$$

Assume the induction hypothesis is correct for i, in the following three steps we prove that the induction hypothesis is correct for $i+1$. As the induction steps are very similar to Lemma 15 we only focus on differences.

Step 1. From induction hypothesis we have for $t \geq t^{(i,0)}$

$$x^+(t) + \Big(\frac{16\hat{\iota}_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1\mathbf{1} \geq u(\Theta^i).$$

In the first step we prove that for $t \geq t^{(i,1)}$, $x_l(t) \geq -\delta$ The proof for step 1 remains the same as Lemma 15.

Step 2. From induction hypothesis and Step 1 we have for $t \geq t^{(i,1)}$

$$x^+(t) + \Big(\frac{16\hat{\iota}_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i}\delta_1\mathbf{1} \geq u(\Theta^i),$$

and $x_l(t) \geq -\delta$ holds for all $l \in \mathcal{S}$ such that $u_l(\Theta^{i+1}) > 0$. In this step we prove that $\forall l \in \mathcal{S}, u_l(\Theta^{i+1}) > 0$ and for $t \geq t^{(i,3)}$

$$x_l(t) \geq 0 \tag{A.48}$$

To proceed, for $t \geq t^{(i,2)}$ Lemma 8 yields

$$|\psi(x(t-T_2), T_2) - \psi(x(t-T_2) + \delta\mathbf{1}, T_2)| \leq \delta e^{CT_2}$$

Next, Lemma 9 yields

$$L(\psi(x(t-T_2) + \delta\mathbf{1}, T_2)) \geq \Theta^{i+1} - (1 - \hat{\iota}_1/4)(\Theta^{i+1} - \Theta^i)$$

This implies that, for all $ji \in \mathcal{I}$ and for $t \geq t^{(i,2)}$, we have

$$\begin{aligned} x_{ji}(t), z_{ji}(t) &\geq \psi(x(t-T_2) + \delta\mathbf{1}, T_2)_k - \delta e^{CT_2} \\ &\geq u_k(\Theta^{j+1} - (\Theta^{j+1} - \Theta^j)(1 - \hat{\iota}_1/4)) - \delta e^{CT_2} \end{aligned}$$

As the minimum nonzero slope in the piece wise linear functions u_{ji} and v_{ji} is $\hat{\iota}_1$ we can lower bound the right hand side as follows

$$\begin{aligned} & u_k(\Theta^{j+1} - (\Theta^{j+1} - \Theta^j)(1 - \hat{\iota}_1/8)) + \hat{\iota}_1^2(\Theta^{j+1} - \Theta^j)/8 - \delta e^{CT_2} \\ \geq\ & u_k(\Theta^{j+1} - (\Theta^{j+1} - \Theta^j)(1 - \hat{\iota}_1/8)) \end{aligned}$$

This together with Assumption 1 yields for each node $l \in \mathcal{S}$ such that $0 < u_l(\Theta^{i+1})$ for $t \geq T_2$

$$-\lambda_l + \sum_{jl \in \mathcal{I}} \mu_{jl} x_{jl}(t) + \sum_{jl \in \mathcal{I}} \mu_{jl} z_{jl}(t) \geq \hat{\iota}_1(\Theta^{i+1} - \Theta^i)/8$$

Thus, for $t \geq t^{(i,2)}$ and $x_i(t) < 0$ we conclude

$$\dot{x}_i^+(t) \ = \ -\lambda_i + \sum_{ji \in \mathcal{I}} \mu_{ji} x_{ji}(t) + \sum_{ji \in \mathcal{I}} \mu_{ji} z_{ji}(t) + \theta_i x_i^-(t) - \gamma_i x_i^+(t) < \hat{\iota}_1(\Theta^{i+1} - \Theta^i)/8$$

Thus, in a completely same argument as the proof for the proof for Lemma 11 we conclude that $\forall l \in \mathcal{S}, u_l(\Theta^{i+1}) > 0$ and that (A.48) is satisfied.

Step 3. In this step we complete the induction proof by showing that for $t \geq t^{i+1,0}$ and for $kl \in \mathcal{I}$

$$x_{kl}^+(t) + \Big(\frac{16\hat{\iota}_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i-1} \delta_1 \mathbf{1} \geq u_{kl}(\Theta^{i+1})$$

$$z_{kl}^+(t) + \Big(\frac{16\hat{\iota}_1^2}{e^{CT_2}\iota_2}\Big)^{|\mathcal{S}|-i-1} \delta_1 \mathbf{1} \geq v_{kl}(\Theta^{i+1})$$

This step is completely the same as step 3 in Lemma 15 except we only need to prove the inequalities for the functions u_{ji} and v_{ji}.

The remaining of the proof is s completely the same as we presented in Lemma 15. □

Proof of Theorem 4. We first present the proof for the case that the system initial state satisfies $X_i^{(N)}(0) \leq \mu_{\max} N^{1/2+\delta}$. Then, we apply Lemma 28 to generalize the proof to the case the system initial state is arbitrary. First, consider the probability of waiting for upcoming passengers. We define

$$T' = t_4 + \max\left\{\alpha_M^{-1} \log\Big(\frac{4|\frac{X^N(0)}{N} - f^*|}{1 - \Theta^{(\max)}}\Big),\ \ \alpha_M^{-1} \log \frac{2\iota_2}{\iota_1}, \mu_{\max}\right\}$$

Setting $t_0 = T'$ in Lemma 32 yields

$$\Big|1 - L\Big(\psi\Big(\frac{X^{(N)}(0)}{N}, T'\Big)\Big)\Big| \leq \frac{1 - \Theta^{(\max)}}{4}.$$

Next, for positive constant $\epsilon_1 > 0$ let

$$\varpi = \min\left\{\alpha^{'''}(\epsilon_1 \frac{2 - 2\tfrac{1-\Theta^{(\max)}}{2}}{e^{C_3 T'}})^2, \alpha^{''}\right\}$$

From Lemma 29 and Lemma 30 we conclude

$$\begin{aligned} & \mathbb{P}\left\{\Big|\frac{X^{(N)}(T')}{N} - \psi\Big(\frac{X^{(N)}(0)}{N}, T'\Big)\Big| > \epsilon_1 N^{-1/2+\varepsilon}\right\} \\ \leq \ & \mathbb{P}\left\{\Big|\frac{\tilde{X}^{(N)}(T')}{N}\Big| > \frac{\epsilon_1 N^{-1/2+\varepsilon}}{e^{C_3 T'}}\right\} + O(e^{-\alpha^{''} N^{2\varepsilon}}) \\ \leq \ & O(e^{-\varpi N^{2\varepsilon}}). \end{aligned} \tag{A.49}$$

Also, from Lemma 32 we have

$$\psi\Big(\frac{X^{(N)}(0)}{N}, T'\Big) \geq u\Big(1 - \frac{1 - \Theta^{\max}}{4}\Big) \tag{A.50}$$

Combining (A.49) and (A.50) with parts 6, 7 and 8 of Lemma 31 yields for $ji \in \mathcal{I}$ and $j \in \mathcal{S}$,

$$\mathbb{P}\left\{\frac{X_{ji}^{(N)}(T')}{N} \leq u_{ji}\Big(1 - \frac{1 - \Theta^{(\max)}}{2}\Big)\right\} = O(e^{-\varpi N^{2\varepsilon}})$$

$$\mathbb{P}\left\{\frac{Z_{ji}^{(N)}(T')}{N} \leq v_{ji}\Big(1 - \frac{1 - \Theta^{(\max)}}{2}\Big)\right\} = O(e^{-\varpi N^{2\varepsilon}})$$

$$\mathbb{P}\left\{\frac{X_{j}^{(N)}(T')}{N} \leq u_{j}\Big(1 - \frac{1 - \Theta^{(\max)}}{2}\Big)\right\} = O(e^{-\varpi N^{2\varepsilon}})$$

Consequently,

$$\mathbb{P}\left\{L(\frac{X^{(N)}(T')}{N}) > \frac{1 - \Theta^{(\max)}}{2}\right\} = O(e^{-\varpi N^{2\varepsilon}}).$$

Iteratively applying this bound for consecutive intervals of length T' yields

$$\mathbb{P}\left\{\sup_{1\le i\le k}\left|L\Big(\frac{X^N(iT')}{N}\Big)-1\right|<\frac{1-\Theta^{(\max)}}{2}\right\}$$
$$\le(1-O(e^{-\varpi N^{2\varepsilon}}))^{\frac{T}{T'}}\le 1-O(Te^{-\varpi N^{2\varepsilon}})\le 1-O(e^{-\varpi N^{\varepsilon}}).$$

The increasing property of the Lyapunov function $L(.)$ yields for $0\le T''\le T'$ and $1\le i\le k$,

$$\mathbb{P}\left\{\left|L\Big(\frac{X^N(iT'+T'')}{N}\Big)-1\right|<\frac{1-\Theta^{(\max)}}{2}\right\}\le 1-O(e^{-\varpi N^{\varepsilon}}).$$

Therefore, parts 5, 6 and 7 of Lemma 31 imply that for $T=O(e^{\varpi N^{\varepsilon}})$,

$$\mathbb{P}\Big\{X_j^N(T)<0\Big\}\ \ \le\ \ \mathbb{P}\Big\{X_j^N(T)<N^{-1/2+\varepsilon}\sum_{j,ji\in\mathcal{I}}\mu_{ji}e_{ji}/2\Big\}=O(e^{-\varpi N^{\epsilon}})$$

Similarly the probability of having passenger waiting at time T is

$$\mathbb{P}\left\{X_j(T)<0 \text{ for some } j\in\mathcal{S}\right\}=O(|\mathcal{S}|\ e^{-\varpi N^{\epsilon}}).$$

Next, we consider optimality gap. For $k\in\mathbb{N}$, define $X^{N,k}=X^N(kT')$. From Lemma 32 we have

$$\Big|L\Big(\psi\Big(\frac{X^{N,k}}{N},T'\Big)\Big)-1\Big|\le\Big|L(\frac{X^{N,k}}{N})-1\Big|e^{-2}$$

Also, from Lemma 30 and Lemma 29, we conclude

$$\mathbb{P}\left\{\Big|\frac{X^{N,k+1}}{N}-\psi\Big(\frac{X^{N,k}}{N},T'\Big)\Big|>\epsilon_1 N^{-1/2+\varepsilon}\right\}=O(e^{-\varpi N^{2\varepsilon}}).$$

Further, parts 9 and 10 of Lemma 31 yield

$$\mathbb{P}\left\{\Big|\frac{X^{N,k+1}}{N}-f\Big|\ge\max\Big[N^{-1/2+\varepsilon},\Big|\frac{X^{N,k}}{N}-f\Big|e^{-1}\Big]\right\}=O(e^{-\varpi N^{2\varepsilon}}),\forall ji\in\mathcal{I}.$$

Thus we obtain

$$\mathbb{P}\left\{|\frac{X^{N,\log(N^{1/2-\varepsilon})}}{N}-f|\ge\max\Big[N^{-1/2+\varepsilon},\Big|\frac{X^{N,0}}{N}-f\Big|N^{-1/2+\varepsilon}\Big]\right\}=O(e^{-\varpi N^{2\varepsilon}}\log N),\forall ji\in\mathcal{I}.$$

Next, following completely the same procedure as the proof for Theorem 2, for $\delta>\varepsilon$ we conclude with the probability of at least $1 - O(e^{-\alpha_U N^{2\delta}})$ we have

$$\frac{\partial U_{\boldsymbol{Q},\boldsymbol{\lambda},\boldsymbol{\gamma}}(T, N, N')}{\partial T} \geq \Big[1 - O\Big(N^{-1/2+\delta} + e^{-\alpha_S T}\Big)\Big] U^*$$

Lastly, note that according to Lemma 28 there exists a positive $t_6 \leq T'$ such that $X_i^{(N)}(t_6) \leq \mu_{\max} N^{1/2+\delta}$. Setting the time reference at t_6 concludes the proof.

A.13 Additional proofs

The proof of the following result is a simple algebraic exercise.

Lemma 33. *We can find an exponential random variable* $\exp(\tau)$ *such that for each link* $ji \in \mathcal{I}$*, there exists a non-negative random variable* $\tilde{t'}_{ji}$ *such that*

$$\tilde{t}_{ji} = \tilde{t'}_{ji} + \exp(\tau), \mathbb{P}\{\tilde{t'}_{ji} < 0\} = 0.$$

Proof. Note that all $\tilde{t}_{ji}$ are exponentially distributed with mean μ_{ji}. Thus, their characteristic functions are

$$\varphi_{\tilde{t}_{ji}}(t) = \frac{1}{1 - i\mu_{ji} t}$$

Consider $\tau = \frac{2}{\min \mu_{ji}}$ and note that its characteristic function equals

$$\varphi_{\tilde{t}_{ji}}(t) = \frac{1}{1 - i\tau^{-1} t}$$

Now, if there exists a random variable $\tilde{t'}_{ji}$ with characteristic function

$$\varphi_{\tilde{t'}_{ji}}(t) = \frac{1 - i\tau^{-1} t}{1 - i\mu_{ji} t}.$$

Then,$\varphi_{\tilde{t'}_{ji}} \varphi_{\tilde{t}_{ji}} = \varphi_{\tilde{t}_{ji}}$ and therefore,

$$\tilde{t}_{ji} = \tilde{t'}_{ji} + \exp(\tau)$$

which concludes the proof. Now we show that $(1 - i\tau^{-1}t)/(1 - i\mu_{ji}t)$ is the characteristic function of a random variable. According to Bochner's theorem it suffices to show that the function $(1 - i\tau^{-1}t)/(1 - i\mu_{ji}t)$ is positive definite. Therefore, we must show that for any n real numbers $x_1, \ldots, x_n$ the matrix

$$A = (a_{ij})_{i,j=1}^n, a_{ij} = (1 - i\tau^{-1}t)/(1 - i\mu_{ji}t)$$

is a positive definite matrix. However, it is straightforward to conclude

$$tr(A) > 0, tr(A)^2/tr(A^2) > n - 1.$$

Therefore, according to the trace criteria for positive definiteness of Hermitian matrices we conclude the proof. □

Proof of Lemma 8.

Decompose the set of single servers $\mathcal{S}$ into two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\lambda_i = 0, \forall i \in \mathcal{S}_1$ and $\lambda_i \neq 0, \forall i \in \mathcal{S}_2$. We prove lemma statement using induction on the size of the set $\mathcal{S}_2$. Specifically, we prove that the functions $u_{ji}(\Theta), ji \in \mathcal{I}, v_{ji}(\Theta), ji \in \mathcal{I}, u_j(\Theta), j \in \mathcal{S}$ are continuous piecewise linear, increasing functions. Also, the number of pieces of each function is at most $|\mathcal{S}_2| + 1$.

a) For induction base, consider the case $|\mathcal{S}_2| = 0$. Then, inequalities (A.7c) and (A.7g) yield $u_{ji}(\Theta) = 0, \forall ji \in \mathcal{I}$. Next, we can combine equations (A.7a) and (A.7e) to conclude

$$\gamma_j u_j(\Theta) = \sum_{i, ji \in \mathcal{I}} q_{ji} \gamma_j u_j(\Theta) = \sum_{k, kj \in \mathcal{I}} q_{kj} \gamma_k u_k(\Theta) \tag{A.51}$$

Define the vector $\eta(\Theta)$ such that $\eta_j(\Theta) = \gamma_j u_j(\Theta), \forall j \in \mathcal{S}$. Next, note that (A.51) yields $\boldsymbol{Q}^T \eta(\Theta) = \eta(\Theta)$. As the matrix $\boldsymbol{Q}^T$ is an irreducible left stochastic matrix, it has a unique unit eigenvector $\eta^{(1)}$ with eigenvalue 1; i.e., $\boldsymbol{Q}^T \eta^{(1)} = \eta^{(1)}$. Hence, $\eta(\Theta) = c_\eta(\Theta) \eta^{(1)}$, for some real number $c_\eta(\Theta)$. Thus, (A.7f) yields

$$\Theta = c_\eta(\Theta) \sum_{j \in \mathcal{S}} \eta_j^{(1)} \left(\frac{1}{\gamma_j} + \sum_{i, ji \in \mathcal{I}} \frac{q_{ji}}{\mu_{ji}} \right), \quad \forall j \in \mathcal{S}$$

Therefore,

$$u_j(\Theta) = \Theta \frac{\eta_j^{(1)}}{\gamma_j \sum_{j \in \mathcal{S}} \eta_j^{(1)} (\frac{1}{\gamma_j} + \sum_{i, ji \in \mathcal{I}} \frac{q_{ji}}{\mu_{ji}})} \quad \forall j \in \mathcal{S}$$

$$v_{ji}(\Theta) = \Theta \frac{q_{ji} \eta_j^{(1)}}{\mu_{ji} \sum_{j \in \mathcal{S}} \eta_j^{(1)} (\frac{1}{\gamma_j} + \sum_{i, ji \in \mathcal{I}} \frac{q_{ji}}{\mu_{ji}})} \quad \forall ji \in \mathcal{I}$$

This concludes the proof for the induction base.

b) Then, assume that the lemma statement is valid for all LCP's of form (A.7) with size $|\mathcal{S}_2| = n_1 - 1$, we prove the statement for the case $|\mathcal{S}_2| = n_1$. Define the matrix $\tilde{\boldsymbol{Q}}$ as follows: For $ji \in \mathcal{I}$,

$$\tilde{\boldsymbol{Q}}_{ji} = \begin{cases} p_{ji}; & j \in \mathcal{S}_2 \\ q_{ji}; & j \in \mathcal{S}_1 \end{cases}$$

As $\boldsymbol{Q}$ is irreducible and $p_{ji} > 0, \forall ji \in \mathcal{I}$, $\tilde{\boldsymbol{Q}}^T$ is an irreducible left stochastic matrix and has a unique unit eigenvector $\xi^{(1)}$ with eigenvalue 1; i.e., $\xi^{(1)} = \tilde{\boldsymbol{Q}}^T \xi^{(1)}$. Let

$$W^{(1)} = \sum_{j \in \mathcal{S}} \sum_{i, ji \in \mathcal{I}} \xi_j^{(1)} \frac{\tilde{\boldsymbol{Q}}_{ji}}{\mu_{ji}} + \sum_{j \in \mathcal{S}_1} \frac{\xi_j^{(1)}}{\gamma_j}.$$

Next, set $\Theta^{(1)} = \min_{j \in \mathcal{S}_2} \frac{\lambda_j W^{(1)}}{\xi_j^{(1)}}$. Then, we prove that every solution $u(\Theta)$ to LCP (A.7) satisfies

$$\begin{aligned}
u_{ji}(\Theta) &\geq \frac{\tilde{\boldsymbol{Q}}_{ji} \xi_j^{(1)}}{W^{(1)} \mu_{ji}} \min(\Theta, \Theta^{(1)}), \quad \forall j \in \mathcal{S}_2, \forall ji \in \mathcal{I} && \text{(A.52)} \\
v_{ji}(\Theta) &\geq \frac{\tilde{\boldsymbol{Q}}_{ji} \xi_j^{(1)}}{W^{(1)} \mu_{ji}} \min(\Theta, \Theta^{(1)}), \quad \forall j \in \mathcal{S}_1, \forall ji \in \mathcal{I} && \text{(A.53)} \\
u_j(\Theta) &\geq \frac{\xi_j^{(1)}}{\gamma_j W^{(1)}} \min(\Theta, \Theta^{(1)}), \quad \forall j \in \mathcal{S}_1 && \text{(A.54)}
\end{aligned}$$

Note that (A.53) combined with (A.7e) yields (A.54). So it suffices to prove (A.52) and (A.53). Using contradiction, consider a feasible solution $u(\Theta)$ to the LCP (A.7) that violates at least one of the two inequalities.

Next we divide the set of infinite server stations into two subsets:

$$\mathcal{I}_1 = \Big(\bigcup_{\substack{ji \in \mathcal{I} \\ j \in \mathcal{S}_2}} ji \Big), \qquad \mathcal{I}_2 = \Big(\bigcup_{\substack{ji \in \mathcal{I} \\ j \in \mathcal{S}_1}} ji \Big).$$

Consider $\tilde{j}\tilde{i}$ such that if $\tilde{j}\tilde{i} \in \mathcal{I}_1$ then

$$\frac{u_{\tilde{j}\tilde{i}}(\Theta) W^{(1)} \mu_{\tilde{j}\tilde{i}}}{\tilde{\boldsymbol{Q}}_{\tilde{j}\tilde{i}} \xi_{\tilde{j}}^{(1)}} = \min \left(\min_{ji \in \mathcal{I}_1} \frac{u_{ji}(\Theta) W^{(1)} \mu_{ji}}{\tilde{\boldsymbol{Q}}_{ji} \xi_j^{(1)}}, \min_{ji \in \mathcal{I}_2} \frac{v_{ji}(\Theta) W^{(1)} \mu_{ji}}{\tilde{\boldsymbol{Q}}_{ji} \xi_j^{(1)}} \right)$$

and if $\tilde{j}\tilde{i} \in \mathcal{I}_2$ then

$$\frac{v_{\tilde{j}\tilde{i}}(\Theta) W^{(1)} \mu_{\tilde{j}\tilde{i}}}{\tilde{\boldsymbol{Q}}_{\tilde{j}\tilde{i}} \xi_{\tilde{j}}^{(1)}} = \min \left(\min_{ji \in \mathcal{I}_1} \frac{u_{ji}(\Theta) W^{(1)} \mu_{ji}}{\tilde{\boldsymbol{Q}}_{ji} \xi_j^{(1)}}, \min_{ji \in \mathcal{I}_2} \frac{v_{ji}(\Theta) W^{(1)} \mu_{ji}}{\tilde{\boldsymbol{Q}}_{ji} \xi_j^{(1)}} \right)$$

Also, define

$$\Theta^{\min} = \min \left(\min_{ji \in \mathcal{I}_1} \frac{u_{ji}(\Theta) W^{(1)} \mu_{ji}}{\tilde{\boldsymbol{Q}}_{ji} \xi_j^{(1)}}, \min_{ji \in \mathcal{I}_2} \frac{v_{ji}(\Theta) W^{(1)} \mu_{ji}}{\tilde{\boldsymbol{Q}}_{ji} \xi_j^{(1)}} \right).$$

Without loss of generality we assume that $\tilde{j}\tilde{i} \in \mathcal{I}_1$. As $\Theta^{\min} < \Theta^{(1)}$, equations (A.7d) and (A.7e) ensure that $v_{\tilde{j}i} = 0, \forall \tilde{j}i \in \mathcal{I}$. Therefore, we rewrite $\Theta^{\min}$ as follows

$$\begin{aligned} \frac{u_{\tilde{j}\tilde{i}}(\Theta) W^{(1)} \mu_{\tilde{j}\tilde{i}}}{\tilde{\boldsymbol{Q}}_{\tilde{j}\tilde{i}} \xi^{(1)}_{\tilde{j}}} &= \frac{W^{(1)} \sum_{i, \tilde{j}i \in \mathcal{I}} u_{\tilde{j}i}(\Theta) \mu_{\tilde{j}i}}{\xi^{(1)}_{\tilde{j}}} \\ &\geq \frac{W^{(1)} \sum_{k, k\tilde{j} \in \mathcal{I}} \Big(u_{k\tilde{j}}(\Theta) + v_{k\tilde{j}}(\Theta) \Big) \mu_{k\tilde{j}}}{\xi^{(1)}_{\tilde{j}}} \\ &= \frac{W^{(1)} \sum_{k, k\tilde{j} \in \mathcal{I}} \Big(u_{k\tilde{j}}(\Theta) + v_{k\tilde{j}}(\Theta) \Big) \mu_{k\tilde{j}}}{\sum_{k, kj \in \mathcal{I}} \tilde{\boldsymbol{Q}}_{k\tilde{j}} \xi^{(1)}_{k}}, \end{aligned}$$

where the equality follows from (A.7b), the inequality follows from (A.7a), and the second equality is due to the fact that $\xi^{(1)} = (\tilde{\boldsymbol{Q}})^T \xi^{(1)}$. Next, we can lower bound the last term as follows:

$$\begin{aligned} & \frac{W^{(1)} \sum_{k, k\tilde{j} \in \mathcal{I}} \Big(u_{k\tilde{j}}(\Theta) + v_{k\tilde{j}}(\Theta) \Big) \mu_{k\tilde{j}}}{\sum_{k, kj \in \mathcal{I}} \tilde{\boldsymbol{Q}}_{k\tilde{j}} \xi^{(1)}_{k}} \\ \geq\; & W^{(1)} \min \Big(\min_{k, k\tilde{j} \in \mathcal{I}_1} \frac{u_{k\tilde{j}}(\Theta) \mu_{k\tilde{j}}}{\tilde{\boldsymbol{Q}}_{k\tilde{j}} \xi^{(1)}_{k}}, \min_{k, k\tilde{j} \in \mathcal{I}_2} \frac{v_{k\tilde{j}}(\Theta) \mu_{k\tilde{j}}}{\tilde{\boldsymbol{Q}}_{k\tilde{j}} \xi^{(1)}_{k}} \Big) \\ \geq\; & \frac{u_{\tilde{j}\tilde{i}}(\Theta) W^{(1)} \mu_{\tilde{j}\tilde{i}}}{\tilde{\boldsymbol{Q}}_{\tilde{j}\tilde{i}} \xi^{(1)}_{\tilde{j}}} \end{aligned}$$

This means that all inequalities must hold as equality. It follows that

$$\frac{u_{k\tilde{j}}(\Theta) \mu_{k\tilde{j}}}{\tilde{\boldsymbol{Q}}_{k\tilde{j}} \xi^{(1)}_{k}} = \frac{u_{\tilde{j}\tilde{i}}(\Theta) W^{(1)} \mu_{\tilde{j}\tilde{i}}}{\tilde{\boldsymbol{Q}}_{\tilde{j}\tilde{i}} \xi^{(1)}_{\tilde{j}}} \quad \forall k, k\tilde{j} \in \mathcal{I}_1$$

Similarly, we have

$$\frac{v_{k\tilde{j}}(\Theta) \mu_{k\tilde{j}}}{\tilde{\boldsymbol{Q}}_{k\tilde{j}} \xi^{(1)}_{k}} = \frac{u_{\tilde{j}\tilde{i}}(\Theta) W^{(1)} \mu_{\tilde{j}\tilde{i}}}{\tilde{\boldsymbol{Q}}_{\tilde{j}\tilde{i}} \xi^{(1)}_{\tilde{j}}} \quad \forall k, k\tilde{j} \in \mathcal{I}_2$$

As the matrix $\tilde{\boldsymbol{Q}}$ is irreducible, iteratively applying the same procedure to all nodes $k, k\tilde{j} \in \mathcal{I}$ yields

$$\Theta^{\min} = \begin{cases} \frac{u_{ji}(\Theta) W^{(1)} \mu_{ji}}{\tilde{\boldsymbol{Q}}_{ji} \xi^{(1)}_{j}}, \forall ji \in \mathcal{I}_1 \\ \frac{v_{ji}(\Theta) W^{(1)} \mu_{ji}}{\tilde{\boldsymbol{Q}}_{ji} \xi^{(1)}_{j}}, \forall ji \in \mathcal{I}_2 \end{cases}$$

Combining this with (A.7e) yields

$$\Theta^{\min} = \frac{\gamma_j u_j(\Theta) W^{(1)}}{\xi_j^{(1)}}, \forall j \in \mathcal{S}_1$$

Also, (A.7d) yields $u_j = 0, \forall j \in \mathcal{S}_2$. Finally, (A.7f) yields

$$\Theta = \sum_{ji \in \mathcal{I}} u_{ji} + \sum_{ji \in \mathcal{I}} v_{ji} + \sum_{j \in \mathcal{S}} u_j = \sum_{ji \in \mathcal{I}} \frac{\Theta^{\min} \tilde{\boldsymbol{Q}}_{ji} \xi_j^{(1)}}{W^{(1)}} + \sum_{j \in \mathcal{S}} \frac{\Theta^{\min} \xi_j^{(1)}}{\gamma_j W^{(1)}} = \Theta^{\min} \frac{W^{(1)}}{W^{(1)}} \tag{A.55}$$

However, as one of the two equations (A.52) and (A.53) are violated, we conclude $\Theta^{\min} < \Theta$. This contradiction proves that all three equations (A.52),(A.53), and (A.54) are satisfied.

Note that (A.55) yields that for $\Theta < \Theta^{(1)}$ all three equations (A.52),(A.53), and (A.54) hold with equality, and thus the solution to LCP is unique, continuous and piecewise linear in Θ. Lastly, if $\Theta \geq \Theta^{(1)}$ we can substitute

$$\begin{aligned} u'_{ji}(\Theta) &= u_{ji}(\Theta) - \frac{\tilde{\boldsymbol{Q}}_{ji} \xi_j^{(1)}}{W^{(1)} \mu_{ji}} \Theta^{(1)}, \forall j \in \mathcal{S}_2, \forall ji \in \mathcal{I} &\text{(A.56)}\\ v'_{ji}(\Theta) &= v_{ji}(\Theta) - \frac{\tilde{\boldsymbol{Q}}_{ji} \xi_j^{(1)}}{W^{(1)} \mu_{ji}} \Theta^{(1)}, \forall j \in \mathcal{S}_1, \forall ji \in \mathcal{I} &\text{(A.57)}\\ u'_j(\Theta) &= u_j(\Theta) - \frac{\xi_j^{(1)}}{\gamma_j W^{(1)}} \Theta^{(1)}, \forall j \in \mathcal{S}_1 &\text{(A.58)} \end{aligned}$$

Observe that $u'(\Theta)$ is a solution to the following LCP

$$\begin{aligned} &\sum_i \mu_{ji} u'_{ji} + \sum_i \mu_{ji} v'_{ji} = \sum_k \mu_{kj} u'_{kj} + \sum_k \mu_{kj} v'_{kj} && \forall j \in \mathcal{S} && \text{(A.59a)}\\ &u'_{ji} \mu_{ji} = p_{ji} \sum_k \mu_{jk} u'_{jk}, && \forall i, j \in R && \text{(A.59b)}\\ &\sum_k \mu_{jk} u'_{jk} \leq \lambda'_j, && \forall i, j \in R && \text{(A.59c)}\\ &(\lambda'_j - \sum_k \mu_{jk} u'_{jk}) u_j = 0, && \forall i, j \in R && \text{(A.59d)}\\ &\mu_{ji} v'_{ji} = q_{ji} \gamma_j u_j, && \forall i, j \in R && \text{(A.59e)}\\ &\sum_{ji \in \mathcal{I}} u'_{ji} + \sum_{ji \in \mathcal{I}} v'_{ji} + \sum_{j \in \mathcal{S}} u'_j = \Theta' && && \text{(A.59f)}\\ &u'_{ji}, v'_{ji}, u'_j \geq 0, && \forall j, i \in R && \text{(A.59g)} \end{aligned}$$

Here, $\lambda'_j = \lambda_j - \min_{i:\lambda_i \neq 0} \lambda_i$. Next, note that for $j = \arg\min_{j \in \mathcal{S}_2} \lambda_j W^{(1)} / \xi_j^{(1)}$, $\lambda'_j = 0$ and

$\lambda_j \neq 0$. Therefore, the system (A.59) satisfies the induction hypothesis and this concludes the proof. Next, note that as the size of the set $\mathcal{S}_2$ is decreased by at least 1 the number of pieces in the piecewise linear functions is at most $|\mathcal{S}|+1$. Also, note that at each step we have $u_j > 0; \forall j \in \mathcal{S}_1, u_{ji}(\Theta) > 0, \forall ji \in \mathcal{I}, v_{ji}(\Theta) > 0; \forall ji \in \mathcal{I}$. Thus, the right limit slope of the functions $u_{ji}(\Theta), ji \in \mathcal{I}, v_{ji}(\Theta), ji \in \mathcal{I}, u_j(\Theta), j \in \mathcal{S}$ is strictly positive. Also note that according to (A.52), (A.53) and (A.54), we have at least one of the following equations hold:

$$\iota_1 \geq \frac{1}{|\mathcal{I}|+|\mathcal{S}|} \min\Big(\frac{\min_{ji,lm:\lambda_{ji}>0} \lambda_{ji}\mu_{lm}}{\max_{k,lm} \lambda_k \mu_{lm}}, \frac{\min_{ji,k:\lambda_{ji}>0} \lambda_{ji}\gamma_k}{\max_{k,lm} \lambda_k \mu_{lm}}\Big)$$

$$\iota_1 \geq \frac{1}{|\mathcal{I}|+|\mathcal{S}|} \min\Big(\frac{\min_{ji,lm:q_{ji}>0} q_{ji}\mu_{lm}}{\max_{lm} \mu_{lm}}, \frac{\min_{ji,k:q_{ji}>0} q_{ji}\gamma_k}{\max_{lm} \mu_{lm}}\Big)$$

$$\iota_1 \geq \frac{1}{|\mathcal{I}|+|\mathcal{S}|} \min\Big(\frac{\min_{lm} \mu_{lm}}{\max_j \gamma_j}, 1\Big)$$

Similarly we can obtain the bounds for ι_2.

A.14 Proof of lemma 10.

a) First, we sum (1.17) over all infinite servers in $\mathcal{I}$ as follows:

$$\begin{aligned}\sum_{ji\in\mathcal{I}} x_{ji}(t) &= \sum_{ji\in\mathcal{I}} x_{ji}(0) - \sum_{ji\in\mathcal{I}} \mu_{ji} \int_0^t x_{ji}(s)ds \\ &+ \sum_j \sum_{i,ji\in\mathcal{I}} p_{ji}(\sum_{kj\in\mathcal{I}} \mu_{kj} \int_0^t x_{kj}(s)ds + \sum_{kj\in\mathcal{I}} \mu_{kj} \int_0^t z_{kj}(s)ds) \\ &+ \sum_j \sum_{i,ji\in\mathcal{I}} p_{ji}(x_j^+(0) - x_j^+(t) - \gamma_j \int_0^t x_j^+(s)ds), \quad ji \in \mathcal{I}\end{aligned}$$

However, we have $\sum_{i,ji\in\mathcal{I}} p_{ji} = 1$. As such, we will have

$$\begin{aligned}\sum_{ji\in\mathcal{I}} x_{ji}(t) &= \sum_{ji\in\mathcal{I}} x_{ji}(0) - \sum_{ji\in\mathcal{I}} \mu_{ji} \int_0^t x_{ji}(s)ds \\ &+ \sum_{kj\in\mathcal{I}} \mu_{kj} \int_0^t x_{kj}(s)ds + \sum_{kj\in\mathcal{I}} \mu_{kj} \int_0^t z_{kj}(s)ds \\ &+ \sum_j (x_j^+(0) - x_j^+(t) - \gamma_j \int_0^t x_j^+(s)ds), \quad ji \in \mathcal{I} \qquad \text{(A.60)}\end{aligned}$$

Then, we sum (1.18) over all infinite servers in $\mathcal{I}$ to obtain

$$\sum_{ji\in\mathcal{I}} z_{ji}(t) = \sum_{ji\in\mathcal{I}} z_{ji}(0) - \sum_{ji\in\mathcal{I}} \mu_{ji} \int_0^t z_{ji}(s)ds + \sum_{ji\in\mathcal{I}} q_{ji}\gamma_j \int_0^t x_j^+(s)ds, \qquad ji \in \mathcal{I}$$

Following the same procedure we can rewrite this expression as

$$\sum_{ji\in\mathcal{I}} z_{ji}(t) = \sum_{ji\in\mathcal{I}} z_{ji}(0) - \sum_{ji\in\mathcal{I}} \mu_{ji} \int_0^t z_{ji}(s)ds + \sum_{j\in\mathcal{S}} \gamma_j \int_0^t x_j^+(s)ds, \qquad ji \in \mathcal{I} \tag{A.61}$$

Now, we can sum equations (A.60) and (A.61) to conclude the proof.

b) We first prove the statement for $z_{ji}(t)$. From Lemma 2 we know that there exists a unique function $z_{ji}(t)$ that satisfies (1.18). Now, define function

$$\hat{z}_{ji}(t) = e^{-\mu_{ji}t}\Big(z_{ji}(0) + \int_0^t e^{\mu_{ji}s} q_{ji}\gamma_j x_j^+(s)ds\Big)$$

and realize that it is the unique solution to the following differential equation with initial condition $\hat{z}_{ji}(0) = z_{ji}(0)$

$$\dot{\hat{z}}_{ji}(t) = -\mu_{ji}\hat{z}_{ji}(t) + q_{ji}\gamma_j x_j^+(t), \qquad ji \in \mathcal{I} \tag{A.62}$$

Moreover, we take integral from both sides of (A.62) to obtain

$$\hat{z}_{ji}(t) = z_{ji}(0) - \int_0^t \mu_{ji}\hat{z}_{ji}(s)ds + \int_0^t q_{ji}\gamma_j x_j^+(s)ds$$

Therefore, we conclude $z_{ji}(.) = \hat{z}_{ji}(.)$. Also,

$$z_{ji}(t) = e^{-\mu_{ji}t}\Big(z_{ji}(0) + \int_0^t e^{\mu_{ji}s} q_{ji}\gamma_j x_j^+(s)ds\Big) \tag{A.63}$$

As such, we conclude $z_{ji}(t) \geq 0, \forall t \geq 0$. Also if $z_{ji}(0) > \mathbf{0}$, then $z_{ji}(t) > 0, \forall t \geq 0$. Similarly, for $p_{ji} = 0$, we can prove $x_{ji} \geq 0$, because the system dynamics for x_{ji} with $p_{ji} = 0$ and z_{ji} have the same structure.

Next, note that Lemma 8 yields

$$x_{ji}(t) \geq \psi(x(0) + \delta\mathbf{1}, t)_{ji} - \delta e^{Ct}$$

For $\epsilon > 0$, set $\delta = e^{-Ct\log(1/\epsilon)}$ to realize that

$$x_{ji}(t) \geq \psi(x(0) + \delta\mathbf{1}, t)_{ji} - \epsilon$$

As our choice of $\epsilon > 0$ is arbitrary, to prove the lemma it is sufficient to prove that for arbitrarily small $\delta > 0$ we have

$$\psi(x(0) + \delta\mathbf{1}, t)_{ji} \geq 0$$

This shows that it suffices to prove the lemma for a solution $x = (x_{ji}, z_{ji}, x_j)$, such that $z_{ji} > \mathbf{0}$. As we have shown in the proof for Lemma 8, the functions $x_{ji}(.)$ and $x_j(.)$ are limits of continuous functions and hence by Uniform limit theorem are continuous. Also, we compute their derivatives as follows:

$$\begin{aligned} \dot{x}_i(t) &= \lim_{\epsilon\to 0} \frac{x_i(t+\epsilon) - x_i(t)}{\epsilon} \\ &= -\lambda_i + \sum_{ji\in\mathcal{I}} \mu_{ji} x_{ji}(t) + \sum_{ji\in\mathcal{I}} \mu_{ji} z_{ji}(t) + \theta_i x_i^-(t) - \gamma_i x_i^+(t), \quad i \in \mathcal{S} \end{aligned} \tag{A.64}$$

Similarly,

$$\dot{x}_{ji}(t) = -\mu_{ji} x_{ji}(t) + p_{ji} \sum_{kj\in\mathcal{I}} \mu_{kj} x_{kj}(t) + p_{ji} \sum_{kj\in\mathcal{I}} \mu_{kj} z_{kj}(t) - p_{ji}\dot{x}_j^+(t) - p_{ji}\gamma_j x_j^+(t), \quad ji \in \mathcal{I}$$

As the functions $x_{kj}(t)$ are continuous, we conclude their minimum, $\min x_{kj}(t)$, is also continuous. Assume for some t we have $\min x_{kj}(t) < 0$. Define

$$t_1 = \sup\left\{ t \geq 0 \middle| \min_{kj\in\mathcal{I}} x_{kj}(t') \geq 0, \forall t' \leq t \right\}.$$

The continuity of $\min x_{kj}(.)$ yields $\min_{kj} x_{kj}(t_1) = 0$. Without loss of generality, denote $ji = \arg\min_{kj} x_{kj}(t_1)$ and realize that $x_{kj}(t_1) \geq 0, \forall kj \in \mathcal{I}$. Next, continuity of function $x_{ji}(.)$ and existence of derivative for function $x_j(.)$ yields either we have $\dot{x}_j^+(t_1) = 0$ and $x_j^+(t_1) = 0$, or we have $\dot{x}_j^-(t_1) = 0$ and $x_j^-(t_1) = 0$ (Continuity of $x_{ji}(.)$ yields if $x_j(t_1) > 0$, then $\dot{x}_j^-(t_1) = 0$ and $x_j^-(t_1) = 0$. Similarly, if $x_j(t_1) < 0$, then $\dot{x}_j^+(t_1) = 0$ and $x_j^-(t_1) = 0$. Also, in case $x_j(t_1) = 0$, then the existence of derivative for $x_j(t_1)$ yields if $\dot{x}_j(t_1) \geq 0$, then $\dot{x}_j^-(t_1) = 0$, similarly, if $\dot{x}_j(t_1) \leq 0$, then $\dot{x}_j^+(t_1) = 0$.) Now, if we have $\dot{x}_j^+(t_1) = 0$ and $x_j^+(t_1) = 0$, then

$$\dot{x}_{ji}(t_1) \geq p_{ji} \sum_{kj\in\mathcal{I}} \mu_{kj} z_{kj}(t_1) > 0$$

Next, rearrange the (A.64) to obtain

$$\sum_{kj\in\mathcal{I}} \mu_{kj} x_{kj}(t) + \sum_{kj\in\mathcal{I}} \mu_{kj} z_{kj}(t) - \dot{x}_j^+(t) - \gamma_j x_j^+(t) = \lambda_j - \dot{x}_j^-(t) - \theta_j x_j^-(t) \tag{A.65}$$

Therefore, if we have $\dot{x}_j^-(t_1) = 0$ and $x_j^-(t_1) = 0$, then we can lower bound $\dot{x}_{ji}(t_1)$ applying

Assumption 1 and $p_{ji} > 0$ as follows:

$$\dot{x}_{ji}(t_1) \geq p_{ji}\lambda_j > 0$$

This shows that, for ϵ small enough we conclude that $x_{ji}(t_1 + \epsilon) > 0$. This contradicts with the definition of t_1. Thu, we conclude $t_1 = \infty$ and conclude the proof for this case.

c) As the functions $x_i^+(t)$, $x_{ji}(t)$, and $z_{ji}(t)$ are limits of continuous functions they should also be continuous. Now, we prove the functions $x_i^+(t)$ are Lipschitz continuous. Recall from the proof for part b) that the derivative of the function $x_i^+(.)$ exists at any point t and is calculated as follows

$$\dot{x}_i^+(t) = -\lambda_i + \sum_{ji\in\mathcal{I}} \mu_{ji}x_{ji}(t) + \sum_{ji\in\mathcal{I}} \mu_{ji}z_{ji}(t) + \theta_i x_i^-(t) - \gamma_i x_i^+(t) - \dot{x}_j^-(t), \quad i \in \mathcal{S} \qquad \text{(A.66)}$$

To prove the statement, it suffices to prove that $|\dot{x}_i^+(t)|$ is bounded. However, from part b) we know that if $\dot{x}_i^+(t) \neq 0$, then we have $\dot{x}_j^-(t_1) = 0$ and $x_j^-(t_1) = 0$. Moreover, we note that from parts a) and b) we know that the functions $x_i^+(t)$, $x_{ji}(t)$, and $z_{ji}(t)$ are positive and bounded. Therefore, all terms in (A.66) are bounded we conclude that $|\dot{x}_i^+(t)|$ is bounded. Next, we prove the functions $x_{ji}(t)$ are Lipschitz continuous. Recall their computed derivative from part b) equals

$$\dot{x}_{ji}(t) = -\mu_{ji}x_{ji}(t) + p_{ji}\sum_{kj\in\mathcal{I}} \mu_{kj}x_{kj}(t) + p_{ji}\sum_{kj\in\mathcal{I}} \mu_{kj}z_{kj}(t) - p_{ji}\dot{x}_j^+(t) - p_{ji}\gamma_j x_j^+(t), ji \in \mathcal{I}$$

As we shown, the derivatives $\dot{x}_j^+(t)$ are bounded. Hence, the boundedness of functions $x_i^+(t)$, $x_{ji}(t)$, and $z_{ji}(t)$ yields the derivatives $\dot{x}_{ji}(t)$ are bounded. Lastly, to prove the functions $z_{ji}(t)$ are Lipschitz continuous, we recall their computed derivative from part b) equals

$$\dot{z}_{ji}(t) = -\mu_{ji}z_{ji}(t) + q_{ji}\gamma_j x_j^+(t), ji \in \mathcal{I} \qquad \text{(A.67)}$$

However, due to boundedness of functions $x_i^+(t)$, $x_{ji}(t)$, and $z_{ji}(t)$, the derivatives $\dot{z}_{ji}(t)$ are bounded. Thus, the functions $z_{ji}(t)$ are Lipschitz continuous.

d) Set $C_2 = \max\{x_i^-(0), \lambda_i/\theta_i\}$. Assume for $t' \geq 0$ we have $x_i(t) < -C_2$. Set

$$t'' = \sup_{t\leq t'} \{t, x_i(t) \geq -C_2\}.$$

The fundamental theorem of calculus yields

$$x_i(t') - x_i(t'') = \int_{t''}^{t'} \dot{x}_i(t)dt$$

However, for values $t'' \leq t \leq t'$, we can lower bound $\dot{x}_i(t)$ using (A.64) as follows:

$$\dot{x}_i(t) \geq -\lambda_i + \theta_i C_2 \geq 0$$

This contradiction concludes the proof.

A.14.1 Proof of Lemma 11

a) First, recall from lemma 10 that the derivatives for $\dot{x}_i(t)$ can be computed as follows:

$$\begin{aligned}\dot{x}_i(t) &= -\lambda_i + \sum_{ji \in \mathcal{I}} \mu_{ji} x_{ji}(t) + \sum_{ji \in \mathcal{I}} \mu_{ji} z_{ji}(t) + \theta_i x_i^-(t) - \gamma_i x_i^+(t) \\ &= \sum_{ji \in \mathcal{I}} \mu_{ji} x_{ji}(t) + \sum_{ji \in \mathcal{I}} \mu_{ji} z_{ji}(t) + \theta_i x_i^-(t) - \gamma_i x_i^+(t) \qquad \text{(A.68)}\end{aligned}$$

Now, define

$$y_i(t) = \frac{x_i(0)}{2} - \int_0^t \gamma_i y_i(s) ds.$$

Applying Picard–Lindelöf Theorem yields $y_i(t) = x_i(0)e^{-\gamma_i}/2$. Next we show that $x_i(t) > y_i(t)$. To prove this statement, by contradiction, assume that it does not hold. Next, note that $x_i(0) > y_i(0)$. By Lipschitz continuity of $x(.)$ and $y(.)$, consider t' such that $t' = \inf\{t : x_i(t) \leq y_i(t)\}$. Thus

$$\frac{x_i(0)}{2} + \int_0^{t'} -\gamma_i y_i(t) dt = x_i(0) + \int_0^{t'} \Big(\sum_{ji \in \mathcal{I}} \mu_{ji} x_{ji}(t) + \sum_{ji \in \mathcal{I}} \mu_{ji} z_{ji}(t) + \theta_i x_i^-(t) - \gamma_i x_i^+(t) \Big) dt$$

However, by definition of t' we know that

$$-\gamma_i y_i(t) < \sum_{ji \in \mathcal{I}} \mu_{ji} x_{ji}(t) + \sum_{ji \in \mathcal{I}} \mu_{ji} z_{ji}(t) + \theta_i x_i^-(t) - \gamma_i x_i^+(t), \forall t \leq t'$$

This concludes the proof for part a).

b) Next, consider the case $x_i(0) < 0$ and by contradiction assume $x_i(t) < -e^{\gamma_i t}$. Therefore, from part (a), we conclude $x_i(t) < 0, \forall t' \leq t$ (otherwise $x_i(t) > 0$). Thus, we can compute $\dot{x}_i(t)$ as follows

$$\dot{x}_i(t) = \sum_{ji \in \mathcal{I}} \mu_{ji} x_{ji}(t) + \sum_{ji \in \mathcal{I}} \mu_{ji} z_{ji}(t) - \theta_i x_i(t)$$

Applying the fundamental theorem of calculus yields

$$x_i(t) = e^{-\theta_i t}\Big(x_i(0) + \int_0^t e^{\theta_i s}(\sum_{ji\in\mathcal{I}} \mu_{ji} x_{ji}(s) + \sum_{ji\in\mathcal{I}} \mu_{ji} z_{ji}(s))ds\Big)$$

Noting that $x_{ji} \geq 0$ and $z_{ji} \geq 0$, we conclude

$$x_i(t) \geq -\min(x_i(0), C_2)e^{-\theta_i t}(x_i(0))$$

A.14.2 Proof of lemma 12

Let $\Theta = L(x(t))$ and decompose the set $\mathcal{S}$ into two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\mathcal{S}_1 = \{i \in \mathcal{S} : L(x(t)) \geq \Theta^j\}$ and $\mathcal{S}_2 = \mathcal{S} \setminus \mathcal{S}_1$. Furthermore, set $\lambda'_i = \sum_k \mu_{ik} u_{ik}(\Theta)$ and note that (A.7d) and the continuity of the function $u(.)$ yields $\lambda_i = \lambda'_i, \forall i \in \mathcal{S}_1$. Now, as $u(\Theta)$ is a solution of (A.7), (A.7e) yields

$$\sum_{i,ji\mathcal{I}} \mu_{ji} v_{ji}(\Theta) = \gamma_j u_j(\Theta) \tag{A.69}$$

Adding λ'_i to both sides yields

$$\lambda'_i + \gamma_j u_j(\Theta) = \sum_k \mu_{ik} u_{ik}(\Theta) + \sum_{i,ji\mathcal{I}} \mu_{ji} v_{ji}(\Theta)$$

Using (A.7a) to substitute the right hand side yields

$$\begin{aligned} u_i(\Theta) = u_i(\Theta) - \lambda'_i t + \sum_{ji\in\mathcal{I}} \mu_{ji} \int_0^t u_{ji}(\Theta)ds + \sum_{ji\in\mathcal{I}} \mu_{ji} \int_0^t v_{ji}(\Theta)ds \\ +\theta_i \int_0^t u_i^-(\Theta)ds - \gamma_i \int_0^t u_i^+(\Theta)ds, \qquad i \in \mathcal{S} \end{aligned} \tag{A.70}$$

Next, we can use (A.7b) to conclude

$$u_{ji}(\Theta)\mu_{ji} = p_{ji} \sum_k \mu_{jk} u_{jk}(\Theta)$$

Combining with equations (A.69) and (A.7a) yields

$$u_{ji}(\Theta)\mu_{ji} + p_{ji}\gamma_j u_j(\Theta) = p_{ji} \sum_k \mu_{jk} u_{jk}(\Theta) + p_{ji} \sum_{i,ji\mathcal{I}} \mu_{ji} v_{ji}(\Theta)$$

Taking integration yields

$$\begin{aligned} u_{ji}(\Theta) &= u_{ji}(\Theta) - \mu_{ji}\int_0^t u_{ji}(\Theta)ds \\ &+p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}\int_0^t u_{kj}(\Theta)ds + p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}\int_0^t v_{kj}(\Theta)ds \\ &+p_{ji}u_j^+(\Theta) - p_{ji}u_j^+(\Theta) - p_{ji}\gamma_j\int_0^t u_j^+(\Theta)ds, \qquad ji\in\mathcal{I} \end{aligned} \tag{A.71}$$

Lastly, taking integration on both sides of (A.69) yields

$$v_{ji}(\Theta) = v_{ji}(\Theta) - \mu_{ji}\int_0^t v_{ji}(\Theta)ds + q_{ji}\gamma_j\int_0^t u_j^+(\Theta)ds, \qquad ji\in\mathcal{I} \tag{A.72}$$

Next, consider the modified initial state $y(0) = x(t) - u(\Theta)$. Then, consider $y(t) = (y_i, y_{ji}, w_{kl})$ the unique solution to the following system of equations

$$\begin{aligned} y_i(t) &= y_i(0) - (\lambda_i - \lambda_i')t + \sum_{ji\in\mathcal{I}}\mu_{ji}\int_0^t y_{ji}(s)ds + \sum_{ji\in\mathcal{I}}\mu_{ji}\int_0^t w_{ji}(s)ds \\ &+\theta_i\int_0^t y_i^-(s)ds - \gamma_i\int_0^t y_i^+(s)ds, \quad i\in\mathcal{S} \\ y_{ji}(t) &= y_{ji}(0) - \mu_{ji}\int_0^t y_{ji}(s)ds \\ &+p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}\int_0^t y_{kj}(s)ds + p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}\int_0^t w_{kj}(s)ds \\ &+p_{ji}y_j^+(0) - p_{ji}y_j^+(t) - p_{ji}\gamma_j\int_0^t y_j^+(s)ds, \quad ji\in\mathcal{I} \\ w_{ji}(t) &= w_{ji}(0) - \mu_{ji}\int_0^t w_{ji}(s)ds + q_{ji}\gamma_j\int_0^t y_j^+(s)ds, \quad ji\in\mathcal{I} \end{aligned}$$

Lemma 11 yields $y_i(t) \geq -\min(|y_i(0)|, C_2)e^{-\theta_i t}\forall i \in \mathcal{S}_1; \forall t \geq 0$. Now, consider $a(t) = (a_i, a_{ji}, b_{kl})$ such that $a(t') = y(t'-t) + u(\Theta); \forall t' \geq t$. We can combine the above equations with equations (A.70), (A.71), and (A.72) to conclude

$$\begin{aligned}
a_i(t) &= x_i(0) - \lambda_i t + \sum_{ji\in\mathcal{I}} \mu_{ji} \int_0^t a_{ji}(s)ds + \sum_{ji\in\mathcal{I}} \mu_{ji} \int_0^t b_{ji}(s)ds \\
&\quad +\theta_i \int_0^t a_i^-(s)ds - \gamma_i \int_0^t a_i^+(s)ds, \quad i \in \mathcal{S} \\
a_{ji}(t) &= x_{ji}(0) - \mu_{ji} \int_0^t a_{ji}(s)ds \\
&\quad +p_{ji} \sum_{kj\in\mathcal{I}} \mu_{kj} \int_0^t a_{kj}(s)ds + p_{ji} \sum_{kj\in\mathcal{I}} \mu_{kj} \int_0^t b_{kj}(s)ds \\
&\quad +p_{ji} a_j^+(0) - p_{ji} a_j^+(t) - p_{ji}\gamma_j \int_0^t a_j^+(s)ds, \quad ji \in \mathcal{I} \\
b_{ji}(t) &= z_{ji}(0) - \mu_{ji} \int_0^t b_{ji}(s)ds + q_{ji}\gamma_j \int_0^t a_j^+(s)ds, \quad ji \in \mathcal{I}
\end{aligned}$$

However, this is the unique solution to the system (1.16-1.18) with initial condition $x(0)$. Thus, $a(t') = x(t'), \forall t' \geq t$. Thus, for $t' \geq t$ we conclude

$$x_i(t') = a_i(t) = y_i(t'-t) + u_i(\Theta) \geq -\min(x_i(0), C_2)e^{-\theta_i(t'-t)} + u_i(\Theta), \quad \forall i \in \mathcal{S}_1$$

Similarly

$$x_{ji}(t') = a_{ji}(t') = y_{ji}(t'-t) + u_{ji}(\Theta) \geq u_{ji}(\Theta), \quad \forall i \in \mathcal{I}$$

$$z_{ji}(t') = b_{ji}(t') = w_{ji}(t'-t) + v_{ji}(\Theta) \geq v_{ji}(\Theta), \quad \forall i \in \mathcal{I}$$

This concludes the proof.

A.14.3 Proof of Lemma 13

We prove this lemma by contradiction. Assume the lemma statement does not hold and set

$$t_1 = \inf\left\{t \geq 0 | x^+(t_2) \geq y^+(t_2), \forall t_2, 0 \leq t_2 \leq t\right\}.$$

Note that as $y^+(0) < x^+(0)$, and the functions $x(.)$ and $y(.)$ are continuous, we conclude $t_1 > 0$. Also, note that Lemma 10 part (a) together with the assumption $x(0) > y(0)$ yields one of the following three cases hold:

i) for some $j \in \mathcal{S}$ we have $x_j^+(t_1) > y_j^+(t_1)$; ii) for some $ji \in \mathcal{I}$ we have $x_{ji}(t_1) > y_{ji}(t_1)$; iii) for some $ji \in \mathcal{I}$ we have $z_{ji}(t_1) > w_{ji}(t_1)$. In all three cases we prove that $x^+(t_1) > y^+(t_1)$, which contradicts with the choice of t_1.

To do so, for $j \in \mathcal{S}$ we call all servers $ji \in \mathcal{I}$ its consecutive servers. Also for each $ji \in \mathcal{I}$ such that $i \in \mathcal{S}_1$ we call i its consecutive server. Lastly, for each $ji \in \mathcal{I}$ such that $i \in \mathcal{S}_2$ we call all servers $ik \in \mathcal{I}$ its consecutive servers. Next, we prove that if we have $x^+_{s_1}(t_1) > y^+_{s_1}(t_1)$, then $x^+_{s_2}(t_1) > y^+_{s_2}(t_1)$, where s_2 is a consecutive server of s_1. As the matrices $\tilde{\boldsymbol{Q}}$ satisfy Assumption 1, iteratively applying this argument yields $x^+(t_1) > y^+(t_1)$. This contradicts with the choice of t_1. Noting that $x(t_1) \geq y(t_1)$, we consider the following cases:

i) $x^+_j(t_1) > y^+_j(t_1), j \in \mathcal{S}$, and $x_{ji}(t_1) = y_{ji}(t_1)$. First, note that if $j \in \mathcal{S}_1$, exploiting (A.65) yields $\dot{x}_{ji}(t) - \dot{y}_{ji}(t) = -\mu_{ji}(x_{ji}(t) - y_{ji}(t))$. Therefore, $x_{ji}(t) - y_{ji}(t) = e^{-\mu_{ji}t}(x_{ji}(0) - y_{ji}(0)) > 0$. Next, for $j \in \mathcal{S}_2$, (A.65) yields $\dot{x}_{ji} = p_{ji}\lambda_j - \mu_{ji}x_{ji}$. while, taking partial derivatives for y_{ji} yields

$$\dot{y}_{ji}(t) = -\mu_{ji}y_{ji}(t) + p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}y_{kj}(t) + p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}w_{kj}(t)$$

However, as

$$p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}y_{kj}(t) + p_{ji}\sum_{kj\in\mathcal{I}}\mu_{kj}w_{kj}(t) \leq p_{ji}\min_{i:\lambda_i\neq 0}\lambda_i/2$$

we conclude

$$\dot{y}_{ji}(t) < p_{ji}\lambda_j - \mu_{ji}y_{ji} = \dot{x}_{ji}$$

Thus for some value $\kappa > 0$ we have $\dot{x}_{ji}(t_1) - \dot{y}_{ji}(t_1) = \kappa < 0$. Therefore,

$$\lim_{\epsilon\to 0} x_{ji}(t_1 - \epsilon) - y_{ji}(t_1 - \epsilon) = \kappa\epsilon < 0$$

This shows that, for small values of $\epsilon > 0$ we have $x_{ji}(t_1 - \epsilon) < y_{ji}(t_1 - \epsilon)$, which contradicts with the choice of t_1.

ii) $x^+_j(t_1) > y^+_j(t_1), j \in \mathcal{S}$, and $z_{ji}(t_1) = w_{ji}(t_1)$. In this case, from (A.63) we have

$$z_{ji}(t) - w_{ji}(t) = e^{-\mu_{ji}t}(z_{ji}(0) - w_{ji}(0)) + \int_0^t e^{\mu_{ji}s}q_{ji}\gamma_j(x^+_j(t) - y^+_j(t)) \tag{A.73}$$

Hence $z_{ji}(t_1) - w_{ji}(t_1) > 0$ which concludes the prof of this case.

iii) ($x_{ji}(t_1) > y_{ji}(t_1)$ and $x^+_i(t_1) = y^+_i(t_1), i \in \mathcal{S}_1$) or ($z_{ji}(t_1) > w_{ji}(t_1)$ and $x^+_i(t_1) = y^+_i(t_1), i \in \mathcal{S}_1$). In this case, noting that $\lambda_i = 0, \forall i \in \mathcal{S}_1$ and $x_i(t), y_i(t) \geq 0, \forall i \in \mathcal{S}_1; \forall t \geq 0$ we can exploit the derivatives computed in Lemma 10 to conclude

$$\dot{x}_i(t) - \dot{y}_i(t) = \sum_{ji\in\mathcal{I}}\mu_{ji}(x_{ji}(t) - y_{ji}(t)) + \sum_{ji\in\mathcal{I}}\mu_{ji}(z_{ji}(t) - w_{ji}(t)) - \gamma_i(x^+_i(t) - y^+_i(t)), i \in \mathcal{S}_1$$

This concludes $\dot{x}_i(t) > \dot{y}_i(t)$. As the argument stated in part (a), this is a contradiction with the choice of t_1. d) ($x_{kj}(t_1) > y_{kj}(t_1)$, $i \in \mathcal{S}_2$ and $x_{ji}(t_1) = y_{ji}(t_1)$), or ($z_{kj}(t_1) > w_{kj}(t_1)$, $i \in \mathcal{S}_2$ and

$x_{ji}(t_1) = y_{ji}(t_1)$). If $x_i(t_1) < 0$ we have

$$\begin{aligned} & \dot{x}_{ji}(t_1) - \dot{y}_{ji}(t_1) \\ = \; & -\mu_{ji}(x_{ji}(t_1) - y_{ji}(t_1)) + p_{ji}\Big(\sum_{kj\in\mathcal{I}} \mu_{kj}(x_{kj}(t_1) - y_{kj}(t_1)) + \sum_{kj\in\mathcal{I}} \mu_{kj}(z_{kj}(t_1) - w_{kj}(t_1))\Big) \end{aligned}$$

Thus, $\dot{x}_{ji}(t_1) - \dot{y}_{ji}(t_1)$. As the argument stated in part (a), this is a contradiction with the choice of t_1. Next, if $x_i(t_1) > 0$ then from part a) we conclude

$$\dot{y}_{ji}(t) < p_{ji}\lambda_j - \mu_{ji}y_{ji} = \dot{x}_{ji}$$

As the argument stated in part (a), this is a contradiction with the choice of t_1. e)($x_{kj}(t_1) > y_{kj}(t_1)$, $i \in \mathcal{S}_2$ and $z_{ji}(t_1) = w_{ji}(t_1)$) or ($z_{kj}(t_1) > w_{kj}(t_1)$, $i \in \mathcal{S}_2$ and $z_{ji}(t_1) = w_{ji}(t_1)$). In this case, similar to part b) we conclude

$$z_{ji}(t) - w_{ji}(t) = e^{-\mu_{ji}t}(z_{ji}(0) - w_{ji}(0)) + \int_0^t e^{\mu_{ji}s} q_{ji}\gamma_j(x_j^+(t) - y_j^+(t)) \tag{A.74}$$

As a result, $z_{ji}(t_1) - w_{ji}(t_1) > 0$ and it concludes the proof of this case.

A.14.4 Proof of Lemma 14

Decompose the set of single servers $\mathcal{S}$ into two subsets $\mathcal{S}_1$ and $\mathcal{S}_2$ such that $\lambda_i = 0, \forall i \in \mathcal{S}_1$ and $\lambda_i > 0, \forall i \in \mathcal{S}_2$. First, Lemma 11 implies that $x_i(t) \geq 0, \forall t \geq 0, \forall i \in \mathcal{S}_1$. Next, we prove $x_i(t) \leq 0, \forall t \geq 0, \forall i \in \mathcal{S}_2$. Using contradiction, assume there exist $t \geq 0; x_i(t) > 0$. Define

$$t_1 = \sup\{t \geq 0; \quad \max_{i\in\mathcal{S}_2}(x_i(t_2)) \leq 0; \forall t_2 \leq t\}$$

As the functions $x_i(.)$ are continuous, there exists $i \in \mathcal{S}$ such that $x_i(t_1) = 0$ and for each $\epsilon > 0$ there exists a $\delta > 0$ such that $x_i(t_1 + \delta) > 0$. Hence, to reach contradiction it suffices to prove that $\dot{x}_i(t_1) < 0$. From Lemma 10 we know that

$$\dot{x}_i(t_1) = -\lambda_i + \sum_{ji\in\mathcal{I}} \mu_{ji}x_{ji}(t_1) + \sum_{ji\in\mathcal{I}} \mu_{ji}z_{ji}(t_1) + \theta_i x_i^-(t_1) - \gamma_i x_i^+(t_1) - \dot{x}_i^-(t_1), i \in \mathcal{S}_2$$

Setting $x_i(t_1) = 0$ yields

$$\dot{x}_i^+(t_1) \leq -\lambda_i + \mu_{ji}(\sum_{ji\in\mathcal{I}} x_{ji}(t_1) + \sum_{ji\in\mathcal{I}} z_{ji}(t_1)), i \in \mathcal{S}_2$$

Next, the lemma assumption yields

$$\begin{array}{rl} & -\lambda_i + \mu_{ji}(\sum_{ji\in\mathcal{I}} x_{ji}(t_1) + \sum_{ji\in\mathcal{I}} z_{ji}(t_1)) \\ \leq & -\lambda_i + \max_{ji} \mu_{ji}(\sum_{ji\in\mathcal{I}} x_{ji}(0) + \sum_{ji\in\mathcal{I}} z_{ji}(0) + \sum_{i\in\mathcal{S}} x_i^+(0)) < 0, i \in \mathcal{S}_2 \end{array}$$

Taking partial derivative with respect to t in equations (1.17) and (1.18) yields

$$\begin{array}{rcl} \dot{x}_i(t) & = & \sum_{ji\in\mathcal{I}} \mu_{ji} x_{ji}(t) + \sum_{ji\in\mathcal{I}} \mu_{ji} z_{ji}(t) - \gamma_i x_i(t), i \in \mathcal{S}_1 \\ \dot{x}_{ji}(t) & = & -\mu_{ji} x_{ji}(t) + p_{ji} \sum_{kj\in\mathcal{I}} \mu_{kj} x_{kj}(t) + p_{ji} \sum_{kj\in\mathcal{I}} \mu_{kj} z_{kj}(t), ji \in \mathcal{I}, j \in \mathcal{S}_2 \\ \dot{x}_{ji}(t) & = & -\mu_{ji} x_{ji}(t), ji \in \mathcal{I}, j \in \mathcal{S}_1 \\ \dot{z}_{ji}(t) & = & -\mu_{ji} z_{ji}(t), ji \in \mathcal{I}, j \in \mathcal{S}_2 \\ \dot{z}_{ji}(t) & = & -\mu_{ji} z_{ji}(t) + q_{ji}\gamma_j x_j, ji \in \mathcal{I}, j \in \mathcal{S}_1 \end{array}$$

Now, let A be a square matrix with size $|\mathcal{S}_1| + 2|\mathcal{I}|$ and refer to rows/columns associated to coefficients of x_{ji} as $\mathcal{I}_1$, and the ones associated to coefficients of z_{ji} as $\mathcal{I}_2$. Specifically, we define A as follows: (We denote each row that corresponds to an element in $\mathcal{I}_1, \mathcal{I}_2$ and $\mathcal{S}_2$ by ji, kl, and j, respectively)

$$A_{j,v} = \begin{cases} \mu_{ji} & v = ji \in \mathcal{I} \& j \in \mathcal{S}_1 \\ \mu_{ji} & v = ji \in \mathcal{I} \& j \in \mathcal{S}_1 \\ -\gamma_j & v = j \& j \in \mathcal{S}_1 \\ 0 & otherwise \end{cases} \qquad A_{ji,v} = \begin{cases} -\mu_{ji} & v = ji \\ p_{ji}\mu_{kj} & v = kj \in \mathcal{I} \& j \in \mathcal{S}_1 \\ p_{ji}\mu_{kj} & v = kj \in \mathcal{I} \& j \in \mathcal{S}_1 \\ 0 & otherwise \end{cases}$$

$$A_{kl,v} = \begin{cases} -\mu_{kl} & v = kl \\ q_{kl}\gamma_k & v = k \in \mathcal{S}_1 \\ 0 & otherwise \end{cases}$$

Thus, the dynamical system (1.16-1.18) can be transformed into the following linear dynamical system

$$\dot{x}(t) = Ax(t) \tag{A.75}$$

The sum of the elements in the column of A associated with server $ji \in \mathcal{I}$ equals

$$\sum_v A_{v,ji} = -\mu_{ji} + \mu_{ji} \sum_{l:il} p_{il} = -\mu_{kl} + \mu_{kl} = 0.$$

Similarly for servers $kl \in \mathcal{I}$ we have

$$\sum_v A_{v,kl} = -\mu_{kl} + \mu_{ji} \sum_{l:il} p_{il} = -\mu_{kl} + \mu_{kl} = 0.$$

In order to be able to provide a close form for the solution to this dynamical system we have to investigate the eigenvalues of the matrix A. By gershgorin circle theorem we have all eigenvalues of the matrix A^T belong to one of the circles

$$|z - A_{v,v}| \leq \sum_{ji \neq kl} |A_{v,v}|.$$

Which can be simplified to

$$|z + \mu_{kl}| \leq \mu_{kl}; \forall kl \in \mathcal{I}, \qquad |z + \gamma_j| \leq \gamma_j; \forall j \in \mathcal{S}.$$

So their intersection with the imaginary axis is the point 0. Moreover, that eigenvalues of a real matrix are the same as it's transpose, so the real part of the eigenvalues of the matrix A are all negative except for the ones that are exactly equal to 0. Now, we can use the fundamental theorem for Linear Systems to show that the solution x takes the form

$$x(t) = Pdiag(e^{B_j t})P^{-1}x_0. \tag{A.76}$$

Where B is the Jordan Canonical form of the matrix A. By Perron–Frobenius Theorem we know that when t goes to infinity all but one of the blocks of the matrix $diag(e^{B_j t})$ converges to 0. The only block that does not converge to zero is the one associated with the eigenvalue 0. As the matrix A has a simple eigenvalue 0, this block, $e^{B_j t} = e^0 = 1$, will remain constant over time. So we can conclude by letting t tend to infinity $x(t)$ will converge exponentially fast to $PCP^{-1}x_0$, where C is a diagonal matrix with all its diagonal elements equal to zero except for one that is equal to 1. Next, we apply the Cheeger inequality presented in Theorem 3.2 in Montenegro [2006] and also in Theorem Theorem 3 in Chung [2005] to obatain the exponential rate as follows:

$$|x - PCP^{-1}x_0| \leq e^{-\min^2(p_{\min}, q_{\min})(t'-t)}.$$

Here, $p_{\min} = \min_{ji} \frac{\lambda_{ji}}{\lambda_j}$ and $q_{\min} = \min_{ji} q_{ji}$. Now that we have proven exponential stability we just have to show that the equilibrium is unique. However, we have already proved this fact, since every equilibrium solution, x^*, for this dynamical system satisfies $\dot{x^*} = Ax^* = 0$.

A.15 Proof of Lemma 3

Without loss of generality, we prove this statement for the set of incoming edges. A similar approach for the outgoing edges can result in the bound for the outgoing edges. Consider the set of lanes that include one of the incoming edges to node n, and denote it by $L_i = \left\{l_1, l_2, ..., l_{|L_i|}\right\}$. Note that each lane includes at most one of these segments. As α^* is a solution to Problem 2.1, we conclude:

$$H \geq \alpha^* \sum_{l \in L_i} (\frac{U_l}{v_{\max}} + d_l h_f) \tag{A.77}$$

Furthermore, by the definition of the ceil function and the fact that $|L_i| \leq \Delta$, we conclude:

$$\sum_{l \in L_i} \left\lceil \frac{\alpha^* (\frac{U_l}{v_{\max}} + d_l h_f)(k - \Delta + 1)}{H} \right\rceil < \Delta + \frac{\alpha^* \sum_{l \in L_i} (\frac{U_l}{v_{\max}} + d_l h_f)(k - \Delta + 1)}{H}$$

Combining this result with equation A.77 yields:

$$\sum_{l \in L_i} \left\lceil \frac{\alpha^* (\frac{U_l}{v_{\max}} + d_l h_f)(k - \Delta + 1)}{H} \right\rceil < \Delta + k - \Delta + 1 = k + 1$$

The fact that the left-hand side is an integer concludes the proof.

A.16 Obtaining a Directed 2-factor Decomposition

In this section we prove that an algorithm adopted from Lovász and Plummer [2009] finds a directed 2-factor decomposition of $G^* = (N^*, A^*)$ where G^* is a directed k regular digraph. To do so, we construct a directed bipartite graph $H = (N_1, N_2, A_{12})$, where N_1 and N_2 are two copies of N^*, also for each edge $e = (i, j) \in A^*$ add an edge e' from the node i in the copy N_1 to the node j in the copy N_2. Thus, we observe that the edges in each perfect bipartite matching in H corresponds to a 2-factor in $\tilde{G}$. Furthermore, the fact that each k regular bipartite graph can be partitioned into a union of k disjoint perfect matchings concludes the proof [Bondy et al., 1976, West et al., 2001].

Figure A.1 demonstrates a simple digraph G and its 2-factor decomposition. Figure A.2 demonstrates the bipartite graph H corresponding to digraph G as well as two matching decompositions that cover all edges of H. To avoid confusion and redundancy, we have not included the additional edges we added to $\tilde{G}$ to make it a k-regular digraph in these two figures. We, also, did not include the entrance/exit edges in these figures to avoid confusion.

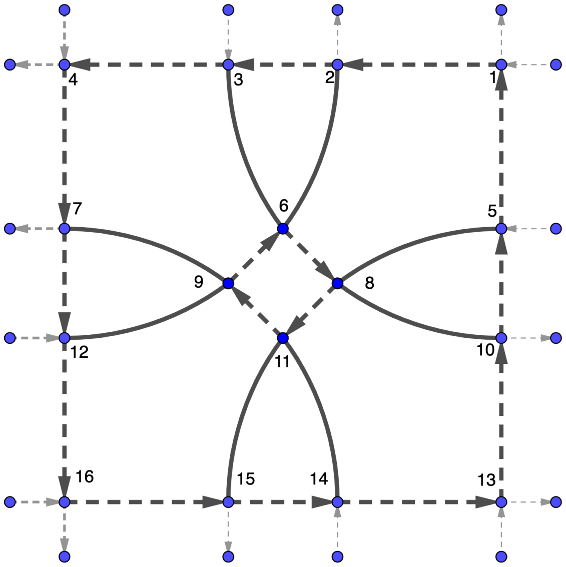


Figure A.1: The graph H and its 2-factor decomposition

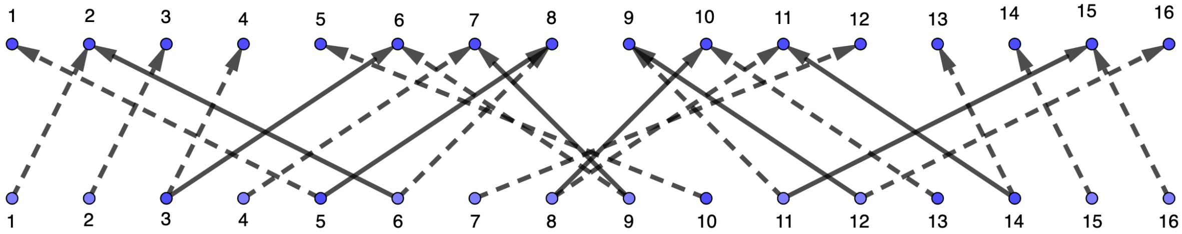


Figure A.2: The corresponding bipartite graph H

A.17 Proof of Theorem 6

Consider a spanning tree T' of the underlying undirected counterpart of G, and hang it from one of its vertices, i. We iteratively define the directed depth d'_j of node j in G. In that, the depth of root node i is zero. Then, for each node j consider its unique immediate predecessor on T' and denote it by k. Also, denote by z the number of initial tracking points on directed edges from k to j in G. If the parallel edges are directed from k to j, then we compute the directed depth $d'_j = d'_k + z + 1$. Otherwise, $d'_j = d'_k - z - 1$.

For each link from a node with depth d'_i to a node with depth d'_j add $d'_j - d'_i - 1 \equiv \pmod{k}$ tracking points on this link. Denote the resulting directed multigraph by H and assign the color c to vertices with depth $c \equiv \pmod{k}$ from the root node i.

A.18 Proof of Lemma 5

First, note that for this condition to happen both vehicles must switch between parallel edge copies of the same pair of segments in A. Assume that a vehicle switching from a copy of segment $a \in A$ to a copy of segment $b \in A$ is blocking another vehicles' movement between the copies of the same segments. Furthermore, denote by D_{p_1}, D_{p_2} the $\{1,2\}$-factors containing the segments for the first switch, and by D_{p_3}, D_{p_4} $\{1,2\}$-factors containing the segments for the second switch. Also define $w_1 \equiv z_{\tilde{e}}^2 + p_1 - p_2 \pmod{k}$, $w_2 \equiv z_{\tilde{e}}^2 + p_3 - p_4 \pmod{k}$, $w_3 \equiv z_{\tilde{e}}^2 + p_3 - p_2 \pmod{k}$ and $w_4 \equiv z_{\tilde{e}}^2 + p_1 - p_4 \pmod{k}$. As the first vehicle arrives to a sooner than the second vehicle and departs later than the second vehicle, we conclude that $w_2, w_3, w_4 < w_1$ and $w_1 + w_2 \geq w_3 + w_4$. Combining these two facts yields $w_1^{(1+\epsilon)} + w_2^{(1+\epsilon)} > w_3^{(1+\epsilon)} + w_4^{(1+\epsilon)}$. As such, we can improve the switch found by Algorithm 3 by substituting the original switches with the switches corresponding to (p_3, p_2) and (p_1, p_4) and lower the weight of the obtained perfect matching.

A.19 Proof of Theorem 7

Define $z_{\max}$ to represent the maximum number of tracking points required to traverse any lane $l \in L$ with the maximum allowable speed $v_{\max}$, this value can be obtained by dividing the time it takes to traverse any lane with maximum allowable speed by the length of the time intervals δ_t. Also note that for each edge $\tilde{a} \in \tilde{A}$ the number of additional tracking points we add while implementing the GCC algorithm is bounded above by $k-1$. Thus, in total the number of tracking points along each lane in the final GCC design is less than $z_{\max} + (|L|+1)(k-1)$. Therefore, we can ensure any vehicle entered the intersection before time step $H - z_{\max} - (|L|+1)(k-1)$ exits the intersection during the study horizon. Besides, note that as α^* is a feasible solution to Problem

2.1, we can ensure under any feasible control there is a lane l such that the number of vehicles from lane l that can cross the intersection during the time horizon $[0, H]$ does not exceed $\alpha^* d_l$.

On the other hand, we can compute the number of allocated entrance time intervals to each lane $l \in L$ during the interval $[0, H - z_{\max} - (|L|+1)(k-1)]$ as:

$$\left\lceil \frac{\alpha^*(\frac{U_l}{v_{\max}} + d_l h_f)(k - \Delta + 1)}{H} \right\rceil \left\lfloor \frac{H - z_{\max} - (|L|+1)(k-1)}{k} \right\rfloor$$

where, in the dimensionless case, $U_l = 0$ and $h_f = 1$. For the values of H that satisfy $\frac{z_{\max}-(|L|+1)(k-1)}{\epsilon} < H$, we can lower bound the number of allocated entrance time intervals to each lane $l \in L$ by $\alpha^*(\frac{U_l}{v_{\max}} + d_l h_f)\frac{k-\Delta+1}{k}(1-\epsilon) = \alpha^* d_l \frac{k-\Delta+1}{k}(1-\epsilon)$, which concludes the proof.

A.20 Proof of Lemma 4

Consider a segment $f \in A$. Denote by $A^-(f)$ the set of segments in A that are directed toward the start node of f along any lane l. As the intersection layout is simple, the sets $A^-(f)$ are mutually exclusive. As mentioned in Section 2.3.2, for each segment $e \in A^-(f)$ there are $\sum_{l:e\in l} \left\lceil \frac{\alpha^*(\frac{U_l}{v_{\max}}+d_l h_f)(k-\Delta+1)}{H} \right\rceil$ parallel edge copies in $\tilde{G}$. Moreover, segment $f \in A$ is the unique segment that follows all segments $e \in A^-(f)$ along all lanes that contain e. Thus, we conclude that segment $f \in A$ in G is also replaced by $\sum_{e\in A^-(f)} \sum_{l:e\in l} \left\lceil \frac{\alpha^*(\frac{U_l}{v_{\max}}+d_l h_f)(k-\Delta+1)}{H} \right\rceil$ parallel edges in $\tilde{G}$. Furthermore, all edges in $\check{G}$ are directed edges from a node $v^+(\tilde{e}) \in \check{V}$ to a nodes $v^-(\tilde{f}) \in \check{V}$ corresponding to edges $\tilde{f}$ that follow $\tilde{e}$ in $\tilde{G}$. Thus, $\check{G}$ is a union of complete bipartite graphs with equal cardinality for its two parts. Therefore, Hall's theorem concludes that $\check{G}$ has a perfect matching [Bondy et al., 1976, West et al., 2001].

A.21 An Illustrative Example for Dimensionless GCC

Here, we provide an illustrative example to demonstrate different steps of Algorithm 1. The example is adopted from Chen et al. [2021], which showed that TSC actually outperforms a reservation scheme with the first-come-first-served policy and rhythmic control in this example. Furthermore, as mentioned in Chen et al. [2021], the footprint of the original intersection designed for rhythmic control is significantly larger than a typical intersection designed for TSC. In our example, we have made a modification to the original intersection layout in Chen et al. [2021]. In particular, we integrate the two left-turn lanes for each approach into a single left-turn lane that serves the aggregated demand, yielding a new small intersection layout to serve the same demand. In this new design, demonstrated in Figure A.3, the required footprint is less than that of a typical intersection

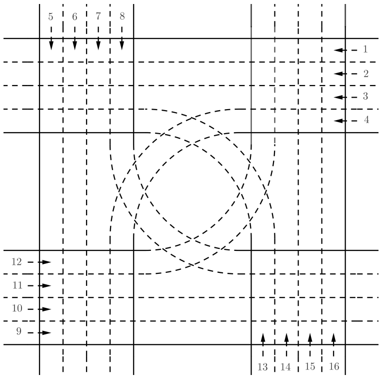


Figure A.3: Intersection Layout

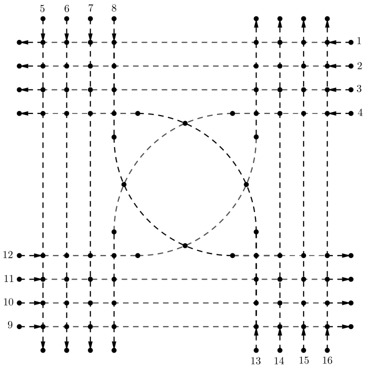


Figure A.4: Intersection Network

designed following the steps in the AASHTO Green Book [Hancock and Wright, 2013]. Note that this small footprint has resulted from two properties of GCC. First, GCC operates all lanes simultaneously. This increases the throughput of single left-turn lanes and reduces the number of required left-turn lanes. Second, GCC does not require the opposite left turns to avoid each other. As demonstrated in Figure A.3, all four left-turn lanes are crossing each other, while this does not affect the throughput of the intersection.

Chen et al. [2021] used the following demand for the intersection:

$$d = [2600, 1400, 1400, 1400, 800, 800, 800, 800]$$

In the above demand vector d, the first four elements represent the per-lane demand for the four through approaches, and the last four elements represent the per-lane demand for the four left-turn approaches. The intersection digraph is demonstrated in Figure A.4. To simplify the design process, we set $v_{max} = 10(m/s)$ and $\delta_t = 0.3(s)$ equal to the time required to move between two adjacent tracking points along a lane with maximum speed.

We implement Step 1-3 of Algorithm 1 to obtain three $\{1, 2\}$-factors as demonstrated in Figures A.5, A.6 and A.7.

Furthermore, we implement Step 4 of Algorithm 1 to obtain the periodic 3-colorable digraph counterpart of G as demonstrated in Figure A.8.

Implementing Step 5-9, we obtain the final intersection design. As each segment of the intersection belongs to a unique lane, the number of tracking points including the ones required for switching to the next edge is constant among all parallel edges of the directed multigraph $\tilde{G}$. Therefore, we depict the final intersection schedule including the tracking points as a simple directed graph in Figure A.9. Also, the entrance time of vehicles into the intersection is demonstrated in Table A.2. Note that k can be any integer number in the set $\mathbb{N}$. We refer the readers to the supporting materials (2.7.2) for a video demonstrating the operations of dimensionless graph coloring

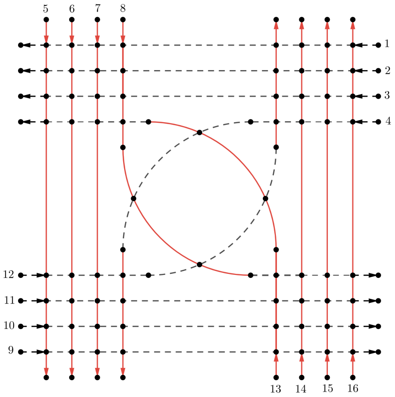


Figure A.5: 1st $\{1, 2\}$-factor

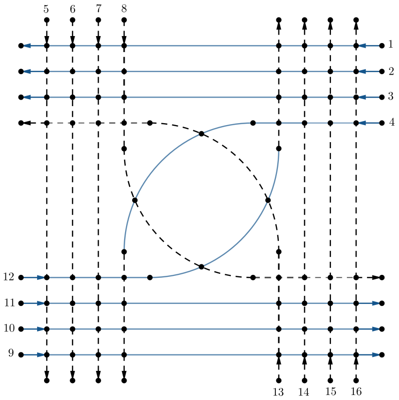


Figure A.6: 2nd $\{1, 2\}$-factor

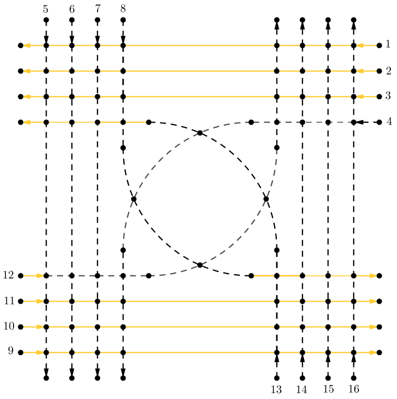


Figure A.7: 3rd $\{1, 2\}$-factor

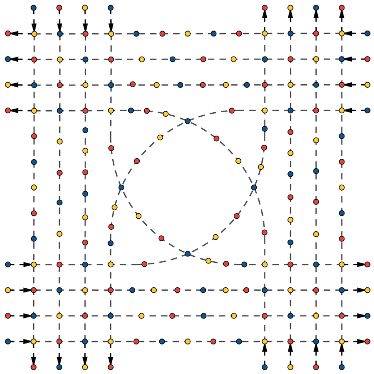


Figure A.8: The periodic 3-colorable digraph

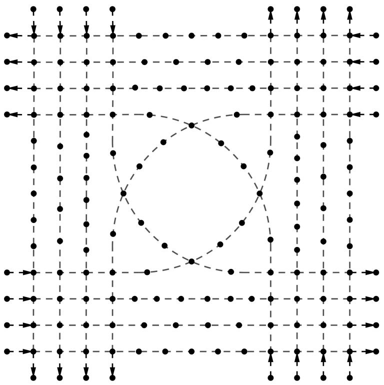


Figure A.9: The final intersection design, including the virtual tracking points

Table A.2: The set of allowable entrance times for each lane.

Lane #	Entrance time	Lane #	Entrance time
lane 1	$3(k\delta) + 2 \quad \& \quad 3(k\delta)$	lane 9	$3(k\delta) + 2 \quad \& \quad 3(k\delta)$
lane 2	$3(k\delta) + 1 \quad \& \quad 3(k\delta) + 2$	lane 10	$3(k\delta) + 1 \quad \& \quad 3(k\delta) + 2$
lane 3	$3(k\delta) \quad \& \quad 3(k\delta) + 1$	lane 11	$3(k\delta) \quad \& \quad 3(k\delta) + 1$
lane 4	$3(k\delta)$	lane 12	$3(k\delta)$
lane 5	$3(k\delta) + 1$	lane 13	$3(k\delta) + 1$
lane 6	$3(k\delta)$	lane 14	$3(k\delta) + 2$
lane 7	$3(k\delta) + 2$	lane 15	$3(k\delta)$
lane 8	$3(k\delta) + 1$	lane 16	$3(k\delta) + 1$

control. Note that, at each moment, the color of a vehicle indicates the $\{1, 2\}$-factor associated with the movement of the vehicle on $\tilde{G}$.

A.22 Finding a Feasible Acceleration/Deceleration Profile

In this section, we prove that inside an intersection with a sufficiently large footprint we can accommodate the required additional tracking points for GCC. In particular, we compute a lower bound on the number of tracking points we can accommodate along an edge $\tilde{a} \in \tilde{A}$ with length β while respecting the kinodynamic constraints, and prove its derivative with respect to β is either a positive linear function of β or a positive constant.

To obtain the lower bound we set a few simplifying assumptions. First, the vehicles crossing the intersection are supposed to enter/depart each edge $\tilde{a}$ with the maximum allowable speed, and its speed-time curve must respect the maximum acceleration/deceleration and maximum speed. Next, We make a conservative assumption that the maximum acceleration and deceleration are equal to the minimum of the two in magnitude, $|\gamma|$. Lastly, we assume the vehicles also respect a minimum speed denoted by v_{min} to ensure a minimum safe tip to tip spatial distance between two consecutive vehicles. The idea is that any two vehicles that maintain a minimum safe headway h_f and travel with a minimum speed equal to v_{min} ensure to maintain a minium distance of $v_{min}h_f$. Therefore, we can compute v_{min} as the division of the minimum safe spatial distance l_{min} by the sum of maximum time length of vehicles in the system and the minimum headway; i.e., $v_{min} = \frac{l_{min}}{u_m+h_f}$.

Next, we compute $t_{\min}$ the minimum time required for a vehicle to traverse an edge $\tilde{a}$ as $t_{\min} = \frac{\beta}{v_{max}}$. Then, we compute $t_{\max}$ the maximum time it takes a vehicle to traverse $\tilde{a}$ while respecting the kinodynamic constraints. Finally, to compute the maximum number of additional tracking points that we can accommodate along a, we divide the difference $t_{\max} - t_{\min}$ by δ_t.

To compute $t_{\max}$, set the entrance time of the vehicle to segment $\tilde{a}$ equal to zero, $t = 0$. Denote its departure time from $\tilde{a}$ by $t_{\max}$. According to the kinodynamic constraints, the speed-time curve of vehicles traversing $\tilde{a}$ must satisfy three constraints as follows:

$$v(t) \geq v_{\max} - |a_{min}|t \qquad \forall t \in [0, t_{\max}] \tag{A.78a}$$

$$v(t) \geq v_{\max} - |a_{min}|(t_{\max} - t) \qquad \forall t \in [0, t_{\max}] \tag{A.78b}$$

$$v(t) \geq v_{\min} \qquad \forall t \in [0, t_{\max}] \tag{A.78c}$$

Here, Equation A.78a ensures the vehicle enters the segment a with speed $v_{\max}$ and respects the minimum deceleration $-|\gamma|$. Equation A.78b ensures the vehicle departs the segment a with speed $v_{\max}$ and respects the maximum acceleration $|\gamma|$. Lastly, Equation A.78c ensures the speed is always greater than the minimum speed $v_{\min}$. Considering different values for the length of segment a, β, the binding constraints among the three kinodynamic constraints may vary. To compute the maximum time to traverse a, we investigate the following two cases separately:

- $l \leq \frac{v_{\max}^2 - v_{\min}^2}{a}$: In this case the constraints A.78a and A.78b are binding. As such, we compute the speed of the vehicle during the time interval $[0, \frac{t_{\max}}{2}]$ as $v_{\max} - |\gamma|t$. Similarly, we compute the speed of vehicle during the time interval $[\frac{t_{\max}}{2}, t_{\max}]$ as $v_{\max} - |\gamma|(t_{\max} - t)$. Thus, the maximum time to traverse a satisfies:

  $$\beta = \int_0^{\frac{t_{\max}}{2}} v(t)dt + \int_{\frac{t_{\max}}{2}}^{t_{\max}} v(t)dt = v_{\max}t_{\max} - |\gamma|\frac{t_{\max}^2}{4}$$

  While by definition we have $\beta = t_{\min}v_{\max}$. Combining these two together yields:

  $$v_{\max}t_{\max} - |\gamma|\frac{t_{\max}^2}{4} = t_{\min}v_{\max}$$

We rewrite this expression to obtain:

$$v_{\max}(t_{\max} - t_{\min}) = |\gamma|\frac{t_{\max}^2}{4}$$

Thus:

$$(t_{\max} - t_{\min}) = |\gamma|\frac{((t_{\max} - t_{\min}) + (t_{\min}))^2}{4v_{\max}}$$

This yields:

$$(t_{\max} - t_{\min}) = \frac{2v_{\max}^2 - \beta\gamma - 2v_{\max}\sqrt{v_{\max}^2 - \beta|\gamma|}}{|\gamma|v_{\max}}$$

Taking partial derivative with respect to β yields:

$$\begin{aligned}\frac{\partial(t_{\max} - t_{\min})}{\partial\beta} &= C|\gamma|(\frac{v_{\max}}{\sqrt{v_{\max}^2 - \beta|\gamma|}} - 1) \\ &= C|\gamma|^2|\beta|(\frac{1}{\sqrt{v_{\max}^2 - \beta|\gamma|}(v_{\max} + \sqrt{v_{\max}^2 - \beta|\gamma|})})\end{aligned}$$

We can lower bound the right-hand side to conclude:

$$C|\gamma|^2|\beta|\frac{1}{2|v_{\max}^2|} = C_1|\gamma|^2|\beta|$$

- $l \geq \frac{v_{\max}^2 - v_{\min}^2}{a}$: In this case all three constraints A.78a, A.78b, and A.78c are binding. As such, we can compute the speed of vehicle during the interval $[0, \frac{v_{\max}-v_{\min}}{|\gamma|}]$ as $v_{\max} - |a_{\max}|t$. The speed remains the constant, $v_{\min}$, during the interval $[\frac{v_{\max}-v_{\min}}{|\gamma|}, t_{\max} - \frac{v_{\max}-v_{\min}}{|\gamma|}]$. Lastly, we can compute the speed of vehicle during the interval $[t_{\max} - \frac{v_{\max}-v_{\min}}{|\gamma|}, t_{\max}]$ as $v_{\max} - |a_{\max}|(t_{\max} - t)$.

  Thus, the maximum time to traverse a satisfies:

$$\begin{aligned}\beta &= \int_0^{\frac{v_{\max}-v_{\min}}{|\gamma|}} v(t)dt + \int_{\frac{v_{\max}-v_{\min}}{|\gamma|}}^{t_{\max}-\frac{v_{\max}-v_{\min}}{|\gamma|}} v(t)dt + \int_{t_{\max}-\frac{v_{\max}-v_{\min}}{|\gamma|}}^{t_{\max}} v(t)dt = v_{\min}t_{\max} \\ &\quad + \frac{(v_{\max} - v_{\min})^2}{|\gamma|} \qquad \text{(A.79)}\end{aligned}$$

  While by definition we have $\beta = t_{\min}v_{\max}$. Combining with equation A.79 yields:

$$(t_{\max} - t_{\min}) = \frac{v_{\max} - v_{\min}}{v_{\min}}(\frac{\beta|\gamma|(v_{\max} - v_{\min})v_{\min}}{|\gamma|v_{\min}}) = C\beta$$

A.23 Proof of Theorem 8

Similar to the proof for Theorem 7, let $z_{\max}$ be the maximum number of tracking points required to traverse any lane $l \in L$ with the maximum allowable speed $v_{\max}$. Also, note that for each edge $\tilde{a} \in \tilde{A}$ the number of additional tracking points we add while implementing Algorithm 7 is bounded above by $ku - 1$. Thus, the total number of tracking points along each lane in the final GCC design is less than $z_{\max} + (|L|+1)(ku-1)$. Therefore, we can ensure any traffic unit entered the intersection before time step $H - u_m - z_{\max} - (|L|+1)(ku-1)$ exits the intersection during the study horizon. Besides, note that as α^* is an optimal solution to Problem 2.1, we can ensure under any feasible control there is a lane l such that the greatest common multiplier of the existing traffic demand that can be accommodated for lane l during the time horizon $[0, H]$ does not exceed $\alpha^* = \frac{H}{(\frac{U_l}{v_{\max}}+d_l h_f)}$.

On the other hand, we can compute the number of admitted traffic units in lane $l \in L$ during the interval $[0, H - u_m - z_{\max} - (|L|+1)(k-1)]$ as:

$$\left\lceil \frac{\alpha^*(\frac{U_l}{v_{\max}} + d_l h_f)(k - \Delta + 1)}{H} \right\rceil \left\lfloor \frac{H - u_m - z_{\max} - (|L|+1)(k-1)}{ku} \right\rfloor$$

For the values of H that satisfy $\frac{u + z_{\max} - (|L|+1)(k-1)}{\epsilon} < H$, we can lower bound the number of admitted traffic units in lane $l \in L$ by $\alpha^*(\frac{U_l}{v_{\max}} + d_l h_f)\frac{k-\Delta+1}{ku}(1-\epsilon)$. Lastly, note that due to the different lengths of consecutive vehicles entering each lane, we may not be able to utilize the $u - \delta$ length for each traffic unit. Nevertheless, if we cannot add one more vehicle to a traffic unit we can ensure that the current utilized length of the traffic unit is at least $u - u_m - \delta$. As such, under the optimal solution to Algorithm 2 we can accommodate $\frac{u-\delta-u_m}{u}\frac{k-\Delta+1}{k}(1-\epsilon)\alpha^*$ multiple of the existing traffic demand for each lane $l \in L$ during the study horizon $[0, H]$. This concludes the proof.

A.24 Algorithm 3

Algorithm 3 Finding a feasible switching scheme

1. Define a new graph $\check{G} = (\check{V}, \check{A})$ where:

   - Each entrance edge $\tilde{a} \in \tilde{A}$ is replaced with one vertex $v^+(\tilde{e}) \in \check{V}$.
   - Each intermediate edge $\tilde{a} \in \tilde{A}$ is replaced with two vertices $v^+(\tilde{e}), v^-(\tilde{e}) \in \check{V}$.
   - Each exit edge $\tilde{a} \in \tilde{A}$ is replaced with one vertex $v^-(\tilde{e}) \in \check{V}$.
   - From each node $v^+(\tilde{e}) \in \check{V}$ there is a directed edge $\check{a} \in \check{A}$ to each node $v^-(\tilde{\tilde{e}}) \in \check{V}$ corresponding to edges in $A^+(\tilde{e})$.

2. Define $w'_{\tilde{e},\tilde{f}} \equiv z_{\tilde{e}}^2 + p - q(\mathrm{mod} \;\; k)$, for each pair of consecutive edges $e, f \in \tilde{A}$. Set $w_{\tilde{e},\tilde{f}} = w'^{(1+\epsilon)}_{\tilde{e},\tilde{f}}$ as the weight of $(v^+(\tilde{e}), v^-(\tilde{f}))$.

3. Implement the blossom algorithm by Edmonds [1965a,b] to find a minimum weight perfect matching in $\check{G}$.

A.25 Algorithm 4

Algorithm 4 Modified $\{1,2\}$-factors Algorithm to respect the restricted regions

```
1. Add a virtual node along each lanes' entrance/departure edge to/from each restricted region. (Figure 2.8)
2. Implement the decomposition algorithm from section 2.3.2.
3. For each one of the $k$ $\{1,2\}$-factors, remove the edges inside the restricted regions.
for each restricted region do
    4. Denote by $L_1, L_2 \subset L$ set of lanes in the crossing movement streams. Define $x = \max_{l \in L_1} \left\lceil \frac{\alpha^*(\frac{U_l}{v_{\max}} + d_l h_f)(k-\Delta+1)}{H} \right\rceil$ and $y = \max_{l \in L_2} \left\lceil \frac{\alpha^*(\frac{U_l}{v_{\max}} + d_l h_f)(k-\Delta+1)}{H} \right\rceil$.
    5. Replace the edges connected to nodes inside the restricted region with $x$ copies of $L_1$ and $y$ copies of $L_2$
end for
```

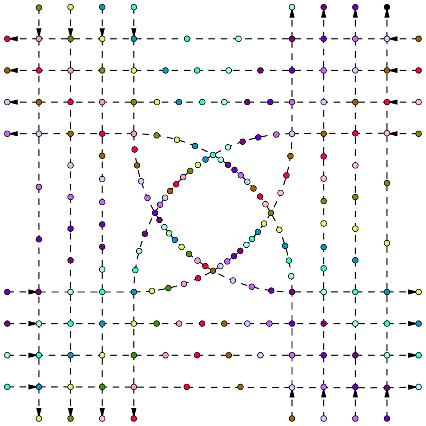


Figure A.10: The periodic 36-colorable Graph

A.26 Proof of Lemma 6

To satisfy $z_{\bar{e}}^2 = 0$, we must modify the initialization step in the algorithm presented in Theorem 6 to include all segments in the restricted regions. Consider the proof for Theorem 6, and note that the proof remains valid after substituting the spanning tree T' with any subgraph that uniquely determines the directed depth of each node. Also, note that as the segments within a restricted region are formed by the intersection of parallel lanes from two movement streams, the directed distance between any two nodes within a restricted region along any two paths is invariant. Therefore, we can initialize Algorithm 6 from a spanning connected subgraph T' that can uniquely determine the directed depth of nodes in G while including all edges in restricted regions. Then, we follow Algorithm 6 for the remaining steps to transform the directed multigraph into a directed periodic k colorable multigraph (Undirected subgraph T' can be obtained in an iterative procedure. In that, we start with the union of edges in the restricted regions. Then, we add edges to connect disconnected components until T' is connected).

A.27 An Illustrative Example for General GCC

To illustrate the implementation of the general version of GCC, we demonstrate the results of implementing Algorithm 2 on the same intersection introduced in Section A.21. As the first 3 steps of the algorithm are the same, we start our analysis from step 4. First, we set the length of a traffic unit equal to $12\delta_t$. Then, implementing step 4, we find a 36-colorable digraph as demonstrated in Figure A.10.

When we implement steps 5-10 of Algorithm 2, we realize that, due to the space limitation, the segments of the intersection cannot accommodate the additional tracking points. As this limitation

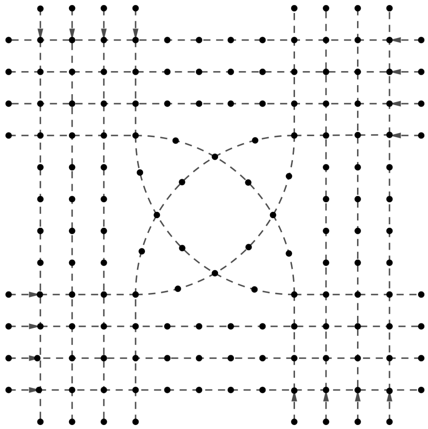


Figure A.11: The final intersection, including the virtual tracking points

is mostly visible in left-turn lanes, we follow the simple remedy presented in Section 2.4.3 to use the left-turn movements every two cycles. Doing so, within each cycle of length 36 time intervals, we allocate the right-of-way to two out of the four left-turn approaches (note that when we remove 2 left turns from each cycle the three middle segments of each left turn lane will integrate into a unified segment). Moreover, to demonstrate the flexibility of GCC, we vary the traffic unit length among lanes to adjust the number of tracking points and, therefore, avoid any longitudinal acceleration/deceleration.

Since each segment of the intersection belongs to a unique lane, number of tracking points including the ones required for switching to the next edge is the same among all parallel edges of the directed multigraph $\tilde{G}$. Therefore, we depict the final intersection schedule including the tracking points as a simple directed graph in Figure A.11. Also, the entrance time of the leading vehicles to the intersection, the length of the inscribed traffic units, u', and the number of extra tracking points to avoid acceleration/deceleration is demonstrated in Table A.3. A positive/negative tracking point modification number indicates that we added/removed that number of tracking points along the edges of $\tilde{G}$, obtained from the final GCC design, to make the number of tracking points along that lane equal to $z^1_{\tilde{e}}$. This ensures the final design does not require any longitudinal acceleration/deceleration. We refer the readers to the supporting materials (2.7.2) for a video demonstrating the operations of the graph coloring control. Note that at each moment the color of a vehicle indicates the $\{1, 2\}$-factor associated with the movement of its traffic unit on $\tilde{G}$.

A.28 Proof of Theorem 10

To present TSC as a special case of GCC, we need to adopt a few definitions from the literature [Allsop, 1972]. In particular, we consider a signalized intersection. Define a stage to be the part of a

Table A.3: The set of allowable entrance times for each lane in the general GCC.

Lane #	Entrance time	Extra tracking points	u'	Lane #	Entrance time	Extra tracking points	u'
lane 1	$36(k\delta)+23 \quad \& \quad 36(k\delta)+35$	+2	10	lane 9	$36(k\delta)+5 \quad \& \quad 36(k\delta)+17$	+2	10
lane 2	$36(k\delta)+22 \quad \& \quad 36(k\delta)+34$	0	12	lane 10	$36(k\delta)+4 \quad \& \quad 36(k\delta)+16$	0	12
lane 3	$36(k\delta)+21 \quad \& \quad 36(k\delta)+33$	-2	10	lane 11	$36(k\delta)+3 \quad \& \quad 36(k\delta)+15$	-2	10
lane 4	$72(k\delta)+20$	0	12	lane 12	$72(k\delta)+2$	0	12
lane 5	$36(k\delta)+20$	+2	10	lane 13	$72(k\delta)+35$	0	12
lane 6	$36(k\delta)+19$	0	12	lane 14	$36(k\delta)$	-2	10
lane 7	$36(k\delta)+18$	-2	10	lane 15	$36(k\delta)+1$	0	12
lane 8	$72(k\delta)+36+17$	0	12	lane 16	$36(k\delta)+2$	+2	10

signal cycle associated with one of its non-conflicting movement groups. The duration of the signal cycle is called the cycle time and is denoted by T. For each stage j, the duration of the signal cycle is divided into effective red time during which no traffic departs from any lane and effective green time during which traffic departs from lanes contained in that stage at a steady saturation rate while there is a queue. For each stage, denote by Γ_j the proportion of the signal cycle that is effectively green. Also, set $\Gamma_{\max} = \max_j \Gamma_j$. Consider a traffic signal control with stages $\{1, 2, \ldots, r\}$. Also, denote by $t_p : p \in \{1, 2, \ldots, r\}$ the time duration allocated to stage p, and set $t_{\min} = \min_p t_p$. Note that the set of lanes included in each stage p can be considered as a $\{1, 2\}$-factor for its underlying digraph. For each stage p, consider $\left\lfloor \frac{t_p}{t_{\min}\kappa} \right\rfloor$ copies of its corresponding $\{1, 2\}$-factor and note that their union forms a decomposition of the intersection demand with $k = \sum_{i=1}^{r} \left\lfloor \frac{t_i}{t_{\min}\kappa} \right\rfloor$. We can feed Algorithm 2 with this demand decomposition to convert it to a GCC. Similar to the proof for theorem 8, we conclude that the the resulting GCC can obtain a reserve capacity of at least

$$\frac{u - \delta - u_m}{u} \frac{k - \Delta + 1}{k} \min_j \frac{\left\lfloor \frac{t_j}{t_{\min}\kappa} \right\rfloor}{\left\lceil \frac{t_j}{t_{\min}\kappa} \right\rceil} \hat{\alpha}$$

Here $\hat{\alpha}$ is the reserve capacity computed in optimization problem 2.1 when considering the demand decomposition associated with the optimal TSC design. When we account for the loss time incurred due to the effective red time, we realize that the reserve capacity of the TSC is upper bounded by:

$$\hat{\alpha}\Gamma_{\max}$$

Note that $k \geq \kappa$. Therefore, we can choose u and κ sufficiently large so that:

$$\frac{u - \delta - u_m}{u} \frac{\kappa - \Delta + 1}{\kappa} \geq \Gamma_{\max}$$

it is evident that under such choice of parameters the reserve capacity of the GCC is at least as high as the considered TSC.

A.29 Proof of Theorem 11

Consider an R-way intersection, and label the R roads in a counter-clockwise ordering with labels $\equiv (\mod R)$. First, note that if the intersection footprint is sufficiently large, we can allocate paths to intersection lanes such that no more than two lanes intersect at the same conflict point. As the intersection footprint is sufficiently large, we can use Theorem 8 to prove that GCC can approach the solution to LP 2.1. As the demand pattern is balanced, the demand for all lanes is the same; i.e., following the Problem 2.1 the value $U = \frac{U_l}{v_{\max}} + d_l h_f$ is the same for all lanes of the intersection. Thus, an optimal solution to the LP 2.1 is $\alpha^* = \frac{H}{2U}$. It suffices to prove that under any signal control scheme, the reserve capacity cannot exceed $\alpha = \frac{H}{RU}$.

In traditional signal control, the approaches that have right-of-way during a phase are not intersecting each other. First, we show that the intersection requires at least R phases to accommodate the demand. To do so, we denote by A the set of proper approaches defined as the set of all movements from a line i to another line j such that $i + 1 \not\equiv j(\mod R)$. Note that the set A contains at least $R(R-2)$ approaches. Thus, it suffices to prove each phase contains at most $(R-2)$ proper approaches.

We use induction to prove this argument. The induction base for $R = 3$ and $R = 4$ is trivial. Assume the induction hypothesis is correct for all integers greater than three and less than R, we prove it must be correct for R. By contradiction, assume there exists a phase, σ, that covers at least $(R-1)$ proper approaches. Denote by (i, j) the approach covered in σ with minimum non-zero value for $j - i \equiv (\mod R)$. Define a congruence interval $[i, j]_R$ to include all integers l such that $a \equiv l - i(\mod R)$ is less than $b \equiv j - i(\mod R)$. As (i, j) has the minimum non-zero value for $j - i \equiv (\mod R)$, for each $i' \in [i, j]_R$ there is no $j' \in [i, j]_R$ such that the approach from i to j is covered in σ. Moreover, as the set of approaches covered in σ are non-intersecting, for each $i' \in [i, j]_R$ there is no $j' \notin [i, j]_R$ such that the approach from i to j is covered in σ. Thus, deleting all roads $i' \in [i, j]_R, i' \neq i, j$ does not decreases the number of proper approaches in σ. However, As the approach i, j is a proper approach we conclude the congruent interval $[i, j]_R$ includes at least 1 element excluding i, j. Thus, σ' includes at least $R - 2$ approaches which is in contradiction with the induction hypothesis.

Consider a signal control with reserve capacity is higher than $\alpha = \frac{H}{RU}$. Denote by $\sigma_1, \ldots, \sigma_{R'} : R' > R$ the phases of the considered signal control. Also, denote by $\theta_1, \ldots, \theta_{R'}$ the portion of a single cycle associated with each phase. As the intersection's reserve capacity is higher than $\alpha = \frac{H}{RU}$, the total admissible flow for the proper approaches must exceed $R(R-2)U\frac{H}{RU} = (R-2)H$. However, during each phase, σ_i, the total admissible flow for its proper approaches can be computed as $(R-2)H\theta_i$. Thus, the total admissible flow for the proper approaches can be computed as follows:

$$\sum_{i=1}^{R'}(R-2)H\theta_i = (R-2)H\sum_{i=1}^{R'}\theta_i = (R-2)H.$$

This is in contradiction with the assumption that the reserve capacity is higher than $\frac{H}{RU}$.

A.30 Relationship with Other Controls

A.30.1 Relationship with Rhythmic Control

In this section, we present rhythmic control or RC [Chen et al., 2021] at isolated intersections as a restricted special case of GCC. Chen et al. [2021] presented the RC idea from a different perspective than the graph coloring control. However, to avoid confusion, in this section we represent their idea in the GCC terminology.

To present RC, we first need to make three restrictive assumptions. First RC assumes that the intersection does not contain any conflict point resulted from intersection of more than two lanes. Second, RC assumes that the platoon of vehicles crossing the intersection contain a single vehicle; i.e., $u_m + h_c = u$. Third, the number of $\{1,2\}$-factors in the demand decomposition phase, k, equals 2. Note that when we set $k = 2$ we do not need to solve the Problem 2.1 to obtain an optimal LP solution. This is because constraint set 2.1b concludes $\left\lceil \frac{\alpha^*(\frac{U_l}{v_{\max}}+d_l h_f)(k-\Delta+1)}{H} \right\rceil = 1$ (this is the main reason the RC idea presented in Chen et al., 2021 does not include an optimization component). Therefore, the right-of-way allocation at each conflict point is alternatively assigned to the two lanes crossing it.

While the underlying control of the mentioned special case of GCC under the three aforementioned assumptions is the same as RC, there is a subtle difference between the physical implication of the two approach. The differences between the two approach is that RC modifies the length of intersection segments as well as the intersection layout to accommodate the additional tracking points, while in GCC, we adjust vehicles' acceleration/deceleration profiles to accommodate them. This difference enlarges the required footprint of RC substantially greater than that of GCC. On the other hand, GCC respects the restricted regions of the intersection and so can be applied to current intersections without incurring any cost to redesign the intersection layout.

In conclusion, RC can be considered as an alternating control that is mostly suitable for balanced demand patterns where the demand per lane is approximately the same for all intersection lanes, so the right-of-way allocation to each lane is proportionate to the demand for that lane. Increasing GCC parameter values, k and l can improve GCC performance over RC in two separate directions.

The increase in parameter k allows GCC to better respond to imbalanced demands. On the other hand, the increase in the parameter l help GCC to accommodate large difference between the following and crossing headway. Furthermore, GCC is robust in handling a set of vehicles of heterogeneous size and accommodating online variations in safety measures, such as the effect of weather and environmental conditions on the minimum safety gap between vehicles and its optimality guarantee will not be compromised when addressing these issues.

A.30.2 Relation with Reservation-based Signal-free Controls

Here, we focus on two well-known reservation-based signal-free schemes for intersection management introduced by Tachet et al. [2016], namely the fair and batch strategies. To begin, both proposed strategies consider a fixed acceleration/deceleration profile for the passage of vehicles through the same lane in the intersection. This simplifying consideration can be translated to our GCC terminology as enforcing the demand decomposition phase to respect an entirety inclusion constraint; i.e., the movement along each lane, e.g., a left-turn movement, needs to be entirely contained in at least one of these groups. Otherwise, the addition of different number of tracking points in parallel edges in the movement synchronization phase may require the vehicles using the same lane to follow different acceleration/deceleration profiles.

The fair policy, in a myopic approach, groups the non-conflicting requests and allocates the right-of-way to the groups according to a first-come-first-served policy. Therefore, fair policy can be viewed as a TSC whose stages are not cyclic and change over time. As the number of different configurations for a groups of non-conflicting requests is limited, it is straightforward to realize that the proof of Theorem 10 holds for the fair policy during a study horizon with length H. That being said, when applying GCC, we realize that there are two sources of suboptimality. First, due to its myopic nature, the fair policy might deviate from the demand decomposition that result in the optimal reserve capacity. Second, GCC allows the vehicles from conflicting lanes to enter the intersection as long as it can ensure collision avoidance. The batch strategy, intentionally, increases the minimum time allocated to each stage of the fair policy. The goal is to utilize the fact that the following headway is shorter than the crossing headway to improve the control performance. This modification is translated to GCC terminology by increasing the length of traffic units in GCC.

A.31 Proof of Lemma 9

We can apply Algorithm 4 to the intersection of any two movement streams. Doing so, we realize that the number of edges along a lane that might include additional tracking points equals the number of movement streams that intersect this lane. Besides, according to Algorithm 2, the number of additional tracking points along each segment does not exceed $ku - 1$, and we may add one more tracking point to implement algorithm 4. This concludes the proof.

A.32 Simulation Settings

We follow the setting presented in Chen et al. [2021] to consider a stationary vehicle arrival process which implies that the vehicle arrivals on each lane follow a time invariant process with a headway that obeys a shifted exponential distribution. We adopt our parameters from Chen et al. [2021]. Therefore, we set $h_f = 0.1s$, vehicle length, $L = 4.5m$, and the speed at the intersection, $v_m = 10m/s$. Also, we set the distance between parallel lanes equal to $3.6m$, and $\delta_t = 0.36(s)$. However, for the crossing headway we use a conservative design and set $h_c = 0.57(s)$. Lastly, in a traffic unit of length $12*0.36 = 4.32s$, we can accommodate at most 7 vehicles since we have $6*0.55+0.45+0.57 = 4.32s$. Additionally, we multiply the demand for each lane by a factor α that indicates the demand level. The optimal GCC design is obtained in Section A.27.

Note that GCC may exhibit a systematic delay in vehicles because of the introduction of virtual tracking points on the intersection layout. The phase transition time loss is set as two seconds for the TSC; with this loss, Webster's method [Webster, 1958] can be adopted to calculate the timing allocation and cycle length of the TSC. The minimum allowable phase timing is $g_{min} = 4s$, and the maximum allowable cycle length for the TSC is $180s$. The CPU time for all computations in this study did not exceed 1(s).

BIBLIOGRAPHY

Philipp Afeche, Zhe Liu, and Costis Maglaras. Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. *Rotman School of Management Working Paper*, (3120544):18–19, 2018.

Ali Allahverdi, Chi To Ng, TC Edwin Cheng, and Mikhail Y Kovalyov. A survey of scheduling problems with setup times or costs. *European journal of operational research*, 187(3):985–1032, 2008.

Richard E Allsop. Estimating the traffic capacity of a signalized road junction. *Transportation Research/UK/*, 6(3), 1972.

Florent Altché, Xiangjun Qian, and Arnaud de La Fortelle. Time-optimal coordination of mobile robots along specified paths. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5020–5026. IEEE, 2016.

Baris Ata, Nasser Barjesteh, and Sunil Kumar. Spatial pricing: An empirical analysis of taxi rides in new york city. Technical report, Working paper, 2019.

Santiago R Balseiro, David B Brown, and Chen Chen. Dynamic pricing of relocating resources in large networks. *Management Science*, 67(7):4075–4094, 2021.

Siddhartha Banerjee, Carlos Riquelme, and Ramesh Johari. Pricing in ride-share platforms: A queueing-theoretic approach. *Available at SSRN 2568258*, 2015.

Siddhartha Banerjee, Yash Kanoria, and Pengyu Qian. Dynamic assignment control of a closed queueing network under complete resource pooling. *arXiv preprint arXiv:1803.04959*, 2018a.

Siddhartha Banerjee, Yash Kanoria, and Pengyu Qian. State dependent control of closed queueing networks with application to ride-hailing. *arXiv preprint arXiv:1803.04959*, 2018b.

Siddhartha Banerjee, Daniel Freund, and Thodoris Lykouris. Pricing and optimization in shared vehicle systems: An approximation framework. *Operations Research*, 2021.

Saif Benjaafar and Xiaobing Shen. Pricing in on-demand (and one-way) vehicle sharing networks. *Available at SSRN 3998297*, 2022.

Omar Besbes, Francisco Castro, and Ilan Lobel. Spatial capacity planning. *Operations Research*, 2021a.

Omar Besbes, Francisco Castro, and Ilan Lobel. Surge pricing and its spatial supply response. *Management Science*, 67(3):1350–1367, 2021b.

Kostas Bimpikis, Ozan Candogan, and Daniela Saban. Spatial pricing in ride-sharing networks. *Operations Research*, 67(3):744–769, 2019.

John Adrian Bondy, Uppaluri Siva Ramachandra Murty, et al. *Graph theory with applications*, volume 290. Macmillan London, 1976.

Anton Braverman, Jim G Dai, Xin Liu, and Lei Ying. Empty-car routing in ridesharing systems. *Operations Research*, 67(5):1437–1452, 2019.

James D Brooks, Koushik Kar, and David Mendonça. Dynamic allocation of entities in closed queueing networks: An application to debris removal. In *2013 IEEE International Conference on Technologies for Homeland Security (HST)*, pages 504–510. IEEE, 2013.

Nicholas Buchholz. Spatial equilibrium, search frictions and efficient regulation in the taxi industry. *Working paper, Tech. Rep.*, 2015.

Nicholas Buchholz. Spatial equilibrium, search frictions, and dynamic efficiency in the taxi industry. *The Review of Economic Studies*, 89(2):556–591, 2022.

Juan Camilo Castillo, Dan Knoepfle, and Glen Weyl. Surge pricing solves the wild goose chase. In *Proceedings of the 2017 ACM Conference on Economics and Computation*, pages 241–242, 2017.

M Keith Chen and Michael Sheldon. Dynamic pricing in a labor market: Surge pricing and flexible work on the uber platform. *Ec*, 16:455, 2016.

Xiangdong Chen, Meng Li, Xi Lin, Yafeng Yin, and Fang He. Rhythmic control of automated traffic—part i: Concept and properties at isolated intersections. *Transportation Science*, 2021.

Shuya Chiba and Tomoki Yamashita. On directed 2-factors in digraphs and 2-factors containing perfect matchings in bipartite graphs. *SIAM Journal on Discrete Mathematics*, 32(1):394–409, 2018.

Fan Chung. Laplacians and the cheeger inequality for directed graphs. *Annals of Combinatorics*, 9(1):1–19, 2005.

Philip Dasler and David M Mount. On the complexity of an unregulated traffic crossing. In *Workshop on Algorithms and Data Structures*, pages 224–235. Springer, 2015.

Jean-Lou De Carufel, Darryl Hill, Anil Maheshwari, Sasanka Roy, and Luís Fernando Schultz Xavier da Silveira. Constant delay lattice train schedules. *arXiv preprint arXiv:2107.04657*, 2021.

Amir Dembo and Ofer Zeitouni. *Large deviations techniques and applications*, volume 38. Springer Science & Business Media, 2009.

Kurt Dresner and Peter Stone. A multiagent approach to autonomous intersection management. *Journal of artificial intelligence research*, 31:591–656, 2008.

Jack Edmonds. Maximum matching and a polyhedron with 0, 1-vertices. *Journal of research of the National Bureau of Standards B*, 69(125-130):55–56, 1965a.

Jack Edmonds. Paths, trees, and flowers. *Canadian Journal of mathematics*, 17:449–467, 1965b.

Joseph JM Evers and Stijn AJ Koppers. Automated guided vehicle traffic control at a container terminal. *Transportation Research Part A: Policy and Practice*, 30(1):21–34, 1996.

Yiheng Feng, Chunhui Yu, and Henry X Liu. Spatiotemporal intersection control in a connected and automated vehicle environment. *Transportation Research Part C: Emerging Technologies*, 89:364–383, 2018.

David K George and Cathy H Xia. Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *European journal of operational research*, 211(1):198–207, 2011.

Peter W Glynn and Assaf Zeevi. Bounding stationary expectations of markov processes. In *Markov processes and related topics: a Festschrift for Thomas G. Kurtz*, pages 195–214. Institute of Mathematical Statistics, 2008.

Ariel Goldszmidt, John A List, Robert D Metcalfe, Ian Muir, V Kerry Smith, and Jenny Wang. The value of time in the united states: Estimates from nationwide natural field experiments. Technical report, National Bureau of Economic Research, 2020.

Ronald Lewis Graham, Eugene Leighton Lawler, Jan Karel Lenstra, and AHG Rinnooy Kan. Optimization and approximation in deterministic sequencing and scheduling: a survey. In *Annals of discrete mathematics*, volume 5, pages 287–326. Elsevier, 1979.

Harish Guda and Upender Subramanian. Your uber is arriving: Managing on-demand workers through surge pricing, forecast communication, and worker incentives. *Management Science*, 65(5):1995–2014, 2019.

Itai Gurvich and Amy Ward. On the dynamic control of matching queues. *Stochastic Systems*, 4 (2):479–523, 2015.

Shlomo Halfin and Ward Whitt. Heavy-traffic limits for queues with many exponential servers. *Operations research*, 29(3):567–588, 1981.

Jonathan V Hall, John J Horton, and Daniel T Knoepfle. Labor market equilibration: Evidence from uber. *URL http://john-joseph-horton. com/papers/uber_price. pdf, working paper*, 2017.

Michael W Hancock and Bud Wright. A policy on geometric design of highways and streets. *American Association of State Highway and Transportation Officials: Washington, DC, USA*, 2013.

Ming Hu and Yun Zhou. Dynamic type matching. *Manufacturing & Service Operations Management*, 24(1):125–142, 2022.

Ramon Iglesias, Federico Rossi, Rick Zhang, and Marco Pavone. A bcmp network approach to modeling and controlling autonomous mobility-on-demand systems. *The International Journal of Robotics Research*, 38(2-3):357–374, 2019.

Yash Kanoria and Pengyu Qian. Blind dynamic resource allocation in closed networks via mirror backpressure. *arXiv preprint arXiv:1903.02764*, 2019.

Leonardo Lamorgese and Carlo Mannino. An exact decomposition approach for the real-time train dispatching problem. *Operations Research*, 63(1):48–64, 2015.

Leonardo Lamorgese and Carlo Mannino. A noncompact formulation for job-shop scheduling problems in traffic management. *Operations Research*, 67(6):1586–1609, 2019.

Leonardo Lamorgese, Carlo Mannino, and Mauro Piacentini. Optimal train dispatching by benders'-like reformulation. *Transportation Science*, 50(3):910–925, 2016.

Michael W Levin and David Rey. Conflict-point formulation of intersection control for autonomous vehicles. *Transportation Research Part C: Emerging Technologies*, 85:528–547, 2017.

Michael W Levin, Hagen Fritz, and Stephen D Boyles. On optimizing reservation-based intersection controls. *IEEE Transactions on Intelligent Transportation Systems*, 18(3):505–515, 2016.

John DC Little. A proof for the queuing formula: L= λ w. *Operations research*, 9(3):383–387, 1961.

Xin Liu. Diffusion approximations for double-ended queues with reneging in heavy traffic. *Queueing Systems*, 91(1):49–87, 2019.

Xin Liu, Qi Gong, and Vidyadhar G Kulkarni. Diffusion models for double-ended queues with renewal arrival processes. *Stochastic Systems*, 5(1):1–61, 2015.

Alexandre Lombard, Florent Perronnet, Abdeljalil Abbas-Turki, and Abdellah El Moudni. Decentralized management of intersections of automated guided vehicles. *IFAC-PapersOnLine*, 49 (12):497–502, 2016.

László Lovász and Michael D Plummer. *Matching theory*, volume 367. American Mathematical Soc., 2009.

T Lozano-Perez. Spatial planning: A configuration space approach. *IEEE Transactions on Computers*, 2(C-32):108–120, 1983.

Hongyao Ma, Fei Fang, and David C Parkes. Spatio-temporal pricing for ridesharing platforms. *Operations Research*, 2021.

Ravi Montenegro. Eigenvalues of non-reversible markov chains: their connection to mixing times, reversible markov chains, and cheeger inequalities. *arXiv preprint math/0604362*, 2006.

Erhun Özkan. Joint pricing and matching in ride-sharing systems. *European Journal of Operational Research*, 287(3):1149–1160, 2020.

Erhun Özkan and Amy R Ward. Dynamic matching for real-time ride sharing. *Stochastic Systems*, 10(1):29–70, 2020.

Daniel Revuz and Marc Yor. *Continuous martingales and Brownian motion*, volume 293. Springer Science & Business Media, 2013.

RidesharingDriver. Ridesharingdriver. *Available at* `https://www.ridesharingdriver.com/survey-data-how-much-uber-drivers-really-make-share/`, 2018. URL `https://www.ridesharingdriver.com/survey-data-how-much-uber-drivers-really-make-share/`.

Jackeline Rios-Torres and Andreas A Malikopoulos. A survey on the coordination of connected and automated vehicles at intersections and merging at highway on-ramps. *IEEE Transactions on Intelligent Transportation Systems*, 18(5):1066–1077, 2016.

Farhad Shahrokhi and David W Matula. The maximum concurrent flow problem. *Journal of the ACM (JACM)*, 37(2):318–334, 1990.

Thierry Siméon, Stéphane Leroy, and J-P Lauumond. Path coordination for multiple mobile robots: A resolution-complete algorithm. *IEEE transactions on robotics and automation*, 18(1):42–49, 2002.

Remi Tachet, Paolo Santi, Stanislav Sobolevsky, Luis Ignacio Reyes-Castro, Emilio Frazzoli, Dirk Helbing, and Carlo Ratti. Revisiting street intersections using slot-based systems. *PloS one*, 11 (3):e0149607, 2016.

Richard L Tweedie. Sufficient conditions for regularity, recurrence and ergodicity of markov processes. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 78, pages 125–136. Cambridge University Press, 1975.

Thomas Urbanik, Alison Tanaka, Bailey Lozner, Eric Lindstrom, Kevin Lee, Shaun Quayle, Scott Beaird, Shing Tsoi, Paul Ryus, Doug Gettman, et al. *Signal timing manual*, volume 1. Transportation Research Board Washington, DC, 2015.

Guangju Wang, Hailun Zhang, and Jiheng Zhang. On-demand ride-matching in a spatial model with abandonment and cancellation. *Available at SSRN 3414716*, 2019.

Ariel Waserhole and Vincent Jost. Pricing in vehicle sharing systems: Optimization in queuing networks with product forms. *EURO Journal on Transportation and Logistics*, 5(3):293–320, 2016.

FV Webster. Traffic signal settings, road research technical paper no. 39. *Road Research Laboratory*, 1958.

Douglas Brent West et al. *Introduction to graph theory*, volume 2. Prentice hall Upper Saddle River, 2001.

Ward Whitt. Open and closed models for networks of queues. *AT&T Bell Laboratories Technical Journal*, 63(9):1911–1979, 1984.

Sze Chun Wong and Hai Yang. Reserve capacity of a signal-controlled road network. *Transportation Research Part B: Methodological*, 31(5):397–402, 1997.

Chiwei Yan, Helin Zhu, Nikita Korolko, and Dawn Woodard. Dynamic pricing and matching in ride-hailing platforms. *Naval Research Logistics (NRL)*, 67(8):705–724, 2020.

Yafeng Yin. Robust optimal traffic signal timing. *Transportation Research Part B: Methodological*, 42(10):911–924, 2008.

Rick Zhang and Marco Pavone. Control of robotic mobility-on-demand systems: a queueing-theoretical perspective. *The International Journal of Robotics Research*, 35(1-3):186–203, 2016.