1-2020

We are on the way: Analysis of on-demand ride-hailing systems

Guiyun FENG
Singapore Management University, gyfeng@smu.edu.sg

Guangwen KONG

Zizhuo WANG

Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0000-0000 | EISSN 0000-0000 | 00 | 0000 | 0001

We Are on the Way: Analysis of On-Demand Ride-Hailing Systems

Guiyun Feng
Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, fengx421@umn.edu

Guangwen Kong
Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, gkong@umn.edu

Zizhuo Wang
Department of Industrial and Systems Engineering, University of Minnesota, Minneapolis, MN 55455, zwang@umn.edu

Recently, there has been a rapid rise of on-demand ride-hailing platforms, such as Uber and Didi, which allow passengers with smartphones to submit trip requests and match them to drivers based on their locations and drivers' availability. This increased demand has raised questions about how such a new matching mechanism will affect the efficiency of the transportation system, in particular, whether it will help reduce passengers' average waiting time compared to traditional street-hailing systems.

In this paper, we address this question by building a stylized model of a circular road and comparing the average waiting times of passengers under various matching mechanisms. After identifying key tradeoffs between different mechanisms, we find that surprisingly, the on-demand matching mechanism could result in higher or lower efficiency than the traditional street-hailing mechanism, depending on the parameters of the system. To overcome the disadvantage of both systems, we further propose adding response caps to the on-demand hailing mechanism and develop a heuristic method to calculate a near-optimal cap. We also test our model using more complex road networks to show that our key observations still exist.

Keywords: on-demand ride-hailing; queueing systems; matching mechanism

1. Introduction

In the past few years, with the growth of information technology, we have witnessed the rapid rise of on-demand ride-hailing platforms such as Uber, Lyft and Didi. Thanks to these platforms, passengers nowadays can request rides using their smartphones instead of hailing taxis on the streets, which used to be the norm in big metropolitan areas. Ride-hailing platforms help connect passengers with drivers (private drivers and traditional taxi drivers) in real time. The growth of these platforms has been astonishing: Uber completed its two-billionth trip in July 2016 (Hawkins 2016b), Didi was responsible for 1.43 billion trips in 2015 alone (Hawkins 2016a), and the growth continues to accelerate.

Despite the rapid growth, policy makers worldwide continue to debate about whether such platforms should be encouraged, and if so, to what extent. Prominent in the debate are concerns about safety, privacy and liability of the drivers and the platform (Rogers 2015). Yet there is also the central question of whether these platforms always increase convenience for passengers and the efficiency of the transportation system. In particular, policy makers struggle to understand how much increased efficiency, if any, such platforms bring to the transportation system; whether the platform can reduce passengers' average waiting times when they need a ride, and if the platform has negative effects on the transportation system. Understanding the answers to these questions would enable policy makers to make more informed decisions when enacting regulations for this emerging service paradigm.

Unfortunately, the answers to these questions are not obvious. Even though it is generally believed that such platforms may help people obtain transportation more efficiently, it is also realized that a platform may decrease the efficiency of transportation during rush hour when the supply easily matches the demand even without the use of ride-sharing platforms. In particular, in a high-traffic situation, it is likely that a driver heading to pick up a passenger who requested service from a platform would encounter another passenger en route, thus increasing the otherwise shorter pickup time and resulting in a loss of efficiency. For example, Steinberg (2012) points out that "People waiting on Manhattan's Fifth Avenue in the middle of rush hour don't need an app to tell cabs to come to the area." An example of action taken in response to such situations is the Shanghai government's 2014 ban on the use of on-demand riding platforms during peak traffic hours (from 7:30 AM to 9:30 AM and from 6:30 PM to 8:30 PM). Although the ban was lifted, the question remains regarding when these platforms tend to increase or decrease the efficiency of the system.

This research attempts to shed light on this question from an operations point of view. Specifically, we consider a stylized model in which taxis[1] drive on a circular road, and we study the performance of different demand-matching mechanisms under the proposed model. Although highly stylized, this model captures the main tradeoffs between the street-hailing system and the on-demand hailing system and thus provides useful insights. In particular, for the street-hailing system, we consider a corresponding "no-call mechanism," in which no platform matches passengers with taxis, and arriving passengers are picked up by the first available taxi passing by. For the on-demand hailing system, we consider a corresponding "call mechanism," in which waiting passengers

[1] In this paper, we use "taxis" to refer to the cars that are used to transport passengers. The term may include traditional taxis, as well as private cars or even self-driving cars that are used to provide transportation services.

are matched to the nearest idle taxi whenever there are taxis available. We are interested in comparing the average waiting time of passengers using different mechanisms. To this end, we first pinpoint the advantages and disadvantages of each mechanism. The advantage of the call mechanism lies in its ability to tell the driver where (which direction) to go to pick up the next customer. However, taxis using the call mechanism may suffer from the possibility of forgoing future better matching opportunities when accepting an incoming request. Upon understanding these tradeoffs, we illustrate the conditions under which one mechanism is more efficient than the other by using an approximate $M/M/k$ queue. Somewhat surprisingly, we show that the call mechanism is more efficient (has lower average waiting time) when the traffic intensity is low or high, while the no-call mechanism could be more efficient when the traffic intensity is in the middle range and when the density of taxis distributed in the transportation system is low. This result, partly confirming what people previously thought, provides new insights into when the call mechanism should be encouraged and when it should not be.

Based on the tradeoffs between the two mechanisms, we further propose a modified call mechanism that adds a distance cap such that a passenger is matched to a taxi only when the distance is below the cap. We show that adding such a cap would preserve most of the benefit of the call mechanism, while greatly diminishing the disadvantage, thus achieving a superior overall performance. We further propose a heuristic to calculate a near-optimal cap that is shown to provide a very good policy compared to the optimal selection.

Finally, we also test our conclusions using more complex grid networks. We find that although the observed phenomenon could be less pronounced, it nevertheless remains. Therefore, we believe that our results could be instrumental for decision makers to understand the tradeoffs of the new service paradigm.

The remainder of the paper is organized as follows: In Section 2, we review the literature that is related to this work. In Section 3, we introduce our models. In Section 4, we consider a simpler one-direction system to illustrate main insights. Then we extend our discussion to a more realistic two-direction system in Section 5. In Section 6, we propose to add a distance cap to the call mechanism and study the optimal selection of the cap. In Section 7, we conduct numerical experiments on more complex road networks. We conclude the paper in Section 8.

2. Literature Review

The emergence and popularity of the sharing economy in recent years have raised many interesting research questions and attracted significant academic interest. The research conducted in this area can be categorized into strategic and operational levels. In the following, we review literature in both categories.

On the strategic level, several recent works examine how the emergence of the sharing economy changes the way people behave and subsequently its impact on the economy. For example, Fraiberger and Sundararajan (2015) use aggregate data from the U.S. automobile industry to show that a shift from ownership to collaborative consumption may lead to less usage, lower used-good prices and a higher consumer surplus. Benjaafar et al. (2015) consider an equilibrium model that endogenizes market friction and provide analytical results showing that product usage and ownership may increase with the presence of a sharing platform. Jiang and Tian (2016) examine the impact of a product-sharing platform on the manufacturer's profit and consumers' surplus. They find that when product quality is exogenous, the manufacturer and consumers are better off when the marginal production cost is high, but both are worse off otherwise. However, when the product quality is endogenous, the consumer surplus is always lower in the presence of a sharing platform. In comparison to the research focusing on product sharing, our study considers operational decisions in matching supply and demand of transportation service via on-demand hailing platforms and studies the impact of different matching mechanisms. In addition, most papers consider consumer surplus that depends on product price and quality. Instead, we focus on whether an on-demand hailing platform can reduce average passenger waiting time.

Another growing stream of papers focus on the operational decision-making problems faced by the on-demand service platforms. One problem that has been considered extensively is capacity management via dynamic pricing. In particular, on-demand service platforms have enabled service providers to choose their own flexible work schedules, presenting new challenges to managing the service capacity. To address this problem, analysis often focuses on how the platform can adjust agent payment and service price to efficiently allocate capacity. For example, Gurvich et al. (2015) employ a newsvendor model to study the capacity management problem in sharing marketplaces where workers have the flexibility to choose their own work schedules. Cachon et al. (2015) consider several contractual forms ranging from fixed price/compensation to surge pricing under a two-period framework and find that providers and consumers are generally better off with surge pricing. Riquelme et al. (2015) model a ride-sharing platform as a queue with customers' arrival and the drivers' work hours depending on the real-time dynamic service price, and they show that the platform cannot significantly increase its revenue by using a dynamic pricing policy based on a threshold number of drivers. Taylor (2016) examines how two defining features of an on-demand service platform – congestion-driven delay disutility and agent independence – affect the platform's optimal per-service price and wage. Tang et al. (2016) use the steady-state equilibrium to characterize the optimal price, wage and payout ratio that maximize the profit of the platform, where an $M/M/1$ queuing model is used to get the approximated waiting time for passengers. In our paper, we do not focus on the pricing and capacity decisions. Instead, we assume the demand and the supply are exogenously given and focus on comparing different matching mechanisms under various utilization levels.

In addition to capacity management, another important operational problem faced by platforms is how to efficiently match service providers with customers. Several recent works have taken on this task. Allon et al. (2012) explore the role of the platforms in facilitating information gathering, operational efficiency and communication by considering three different market models employed by platforms and then characterize the corresponding market outcomes. Cullen and Farronato (2014) study the problem of balancing highly variable demand and supply for a frictional matching market and calibrate their model by using data from TaskRabbit. Anderson et al. (2015) investigate timely exchanges for agents in a barter marketplace, and their results show that a greedy policy that attempts to match upon each arrival is approximately optimal (minimizes average waiting time) among a large class of policies including batching policies. Baccara et al. (2015) consider two-sided matching with vertically different preferences over agents. They show that the optimal mechanism always matches congruent pairs immediately and holds on to a stock of incongruent pairs up to a certain threshold, and a centralized market is more appealing than the corresponding decentralized one. Akbarpour et al. (2016) study different dynamic matching strategies in network markets where agents arrive stochastically and stay for a random period of time before leaving the system if not matched. They show that waiting to thicken the market is highly valuable if the central planner knows the agents' departure time, otherwise, a greedy local algorithm is close to optimal. Hu and Zhou (2016) model dynamic matching between the demand and supply of heterogeneous types in a periodic-review fashion. They provide sufficient conditions on matching rewards such that the optimal matching policy follows a priority hierarchy among possible matching pairs. Similar to these works, the on-demand matching problem studied in our work is also a two-sided dynamic matching problem faced by a centralized platform; in particular, we can model the preference level (matching quality) for any taxi-passenger pair by their distance. However, it would be challenging to use the methods in these papers to evaluate the efficiency of the system. This is because a key differentiating factor in our paper is that since taxis keep moving, the matching scores are constantly evolving even when no new arrival occurs. We obtain insights that are particular to this dynamic.

In addition to the operations management literature, the dynamic matching problem described above has also been studied in the transportation literature under the taxi fleet dispatch context. For example, Meyer and Wolfe (1961) provide approximate stationary analysis for different dispatch systems, providing explicit expressions for operational decisions. McLeod (1972) develop performance models for the taxi system based on queueing theory and show the important tradeoffs between passenger waiting time, fleet size and taxi productivity. Bailey and Clark (1992) study the structure and behavior of an urban taxi system, where the average waiting time and taxi utilization are presented for simulation models under different dispatch and idle-time strategies. In this paper, we focus on comparing two matching strategies: street-hailing and on-demand hailing, and we obtain insights about under what circumstances each strategy has better performance. We also propose potential ways to eliminate the disadvantage of both strategies and thus improve the efficiency of the system.

3. Models

We consider a stylized transportation system on a circular road with perimeter R. There are k taxis in the system. Passengers arrive at the system according to a Poisson process with rate λ, and their arrival locations are uniformly distributed on the circle. Each arriving passenger requests a service with duration d, where d follows an exponential distribution with mean $1/\mu$.[2] We assume that all taxis have a constant speed of one; i.e., they travel one unit of distance on the circle in a unit of time.

In this paper, we consider two different systems in terms of the direction of travel: a one-direction system and a two-direction system. In the one-direction system, only a clockwise direction of travel is allowed. That is, all taxis travel clockwise, and so are the passengers' requests (i.e., if a passenger requests a service with duration d, then she is requesting to be transported in a clockwise direction for a distance of d). In contrast, in the two-direction system, taxis and passengers are allowed to travel in both directions. More precisely, in the two-direction system, the initial direction of each taxi, the requested service direction of each passenger, as well as the travel direction of a taxi after completing a service, are all randomly chosen, with a 50% chance being clockwise and a 50% chance being counterclockwise.

The focus of this paper is on comparing two matching mechanisms. The first mechanism is referred to as the *call mechanism*, which is used to model the matching process made by the

[2] In our model, it is possible for d to be greater than R. However, our model can be easily modified to one in which the requested durations of the passengers follow a bounded distribution.

on-demand ride-hailing platforms. In the call mechanism, when a passenger arrives, a platform immediately matches the passenger to the nearest taxi if there is a taxi available; otherwise, arrived passengers will wait until the next taxi becomes available, at which time the taxi will be matched to its nearest passenger (the remaining passengers, if any, would keep waiting for the next taxi to become available). Once a match between a taxi and a passenger is made, the taxi is committed to serving that passenger next, and the passenger waits for the matched taxi to come. In the one-direction system, the distance between a passenger and a taxi is defined as the distance for a taxi to reach a passenger when driving clockwise; while in the two-direction system, the distance is defined as the shorter distance between driving clockwise and counterclockwise. The other mechanism is the *no-call mechanism*, which is used to model the matching process of traditional street hailing. In the no-call mechanism, no matching occurs, and a passenger is picked up by the first available taxi passing by.

To summarize, we consider four settings in this paper: one-direction call/no-call systems and two-direction call/no-call systems. The features of the four settings are listed in Table 1.

Table 1 Features of call/no-call mechanisms in one-direction/two-direction systems

Transportation systems	*1-d no-call*	*1-d call*	*2-d no-call*	*2-d call*
Driving direction	Clockwise	Clockwise	Both	Both
Initial direction/ Request direction/ Direction upon finish	Clockwise	Clockwise	Random	Random
Matching mechanism	No matching	Matched to nearest clockwise	No matching	Matched to nearest between two directions

The goal of this work is to study and compare the efficiency of these systems, particularly call versus no-call systems. We use the average waiting time of passengers as the performance measure, which is defined as the average time interval between a passenger's arrival and being picked up. By Little's Law, the average waiting time of passengers is proportional to the average number of passengers waiting in the system. Thus, the average waiting time can measure the passenger's satisfaction and the congestion of the system.

4. One-Direction Call/No-Call Systems

In this section, we first study the one-direction systems. Through studying the one-direction systems, we illustrate the tradeoffs between the call and no-call mechanisms, which will be helpful for studying the more complicated two-direction systems in Section 5. The approach of our study is as follows: We first perform numerical experiments on these systems, which lead to some key observations. Then, we provide intuitive explanations for those observations. Finally, we verify the observations with theoretical analysis using a novel approximation approach.

4.1. Numerical Experiments

In the following, we perform numerical experiments about the one-direction systems. In each numerical experiment, we fix the service rate μ, the number of taxis k, and the road length R, and we adjust the arrival rate λ to see how the average waiting time of passengers changes with λ. Note that this effectively changes the utilization level $\rho \doteq \lambda/k\mu$, which is an indicator of the congestion level of the system. In addition, without loss of generality, we fix $\mu = 0.1$ throughout our study.[3] Furthermore, we vary the values of k and R to see how these parameters affect the relation between the average waiting time and the utilization level ρ.

We start with the one-direction call system. The results are shown in Figure 1. From Figure 1, we have the following observation.[4]

Observation 1. *In the one-direction call system, the average waiting time is not always monotonically increasing in ρ. In particular, it increases in ρ when ρ is small or large. However, it could decrease in ρ when ρ is medium. Moreover, such non-monotonicity is more pronounced when R is large.*

To understand the intuition behind the non-monotonicity observed in Figure 1, we decompose the waiting time for one passenger into two components: *response time* and *en route time*. Here, the response time is defined as the time between a passenger's arrival and when the passenger's ride request is responded to by the nearest taxi, and the en route time is defined as the time between the request response and the passenger pickup. In particular, when a passenger arrives, if there are

[3] In our models, it is easy to see that a system with parameters (λ, μ, k, R) is equivalent to a system with parameters $(c\lambda, c\mu, k, R/c)$ except that the time scale is multiplied by $1/c$, i.e., one unit of time in the original system becomes $1/c$ unit of time in the new system. Therefore, we can fix one parameter among λ, μ and R in our study without loss of generality. In our paper, we choose to fix μ.

[4] In all our simulation results, each average waiting time is based on the average of 500 replications of sample paths. In each sample path, we run the system from an idle state. Then we use the average waiting time between the t_0-th and t_1-th passengers as the average waiting time in that sample path. Here, t_0 can be viewed as having a warm-up period until the system enters a steady state. The value of t_0 is determined by a commonly used graphical method, see Welch (1983). In particular, let Y_j denote the average waiting time of passenger j over 500 sample paths. Then we take $w = 20$ and let $\bar{Y}_j = (1/w)\sum_{i=j-w+1}^{j} Y_i$ be the average waiting time between the $(j-w+1)$-th and the j-th passengers. We plot $\bar{Y}_j$ and observe visually when $\bar{Y}_j$ becomes stable. Then we choose a t_0 that is large enough so that $\bar{Y}_j$ has entered a steady state. We further choose $t_1 = t_0 + 6000$.

Also, we note that the observations in this section and Section 5 are obtained from extensive numerical experiments, and the figures we choose to show are representative of the numerical results. As we will show in Sections 4.2 and 5.2, these observations are generally valid in an approximation system.

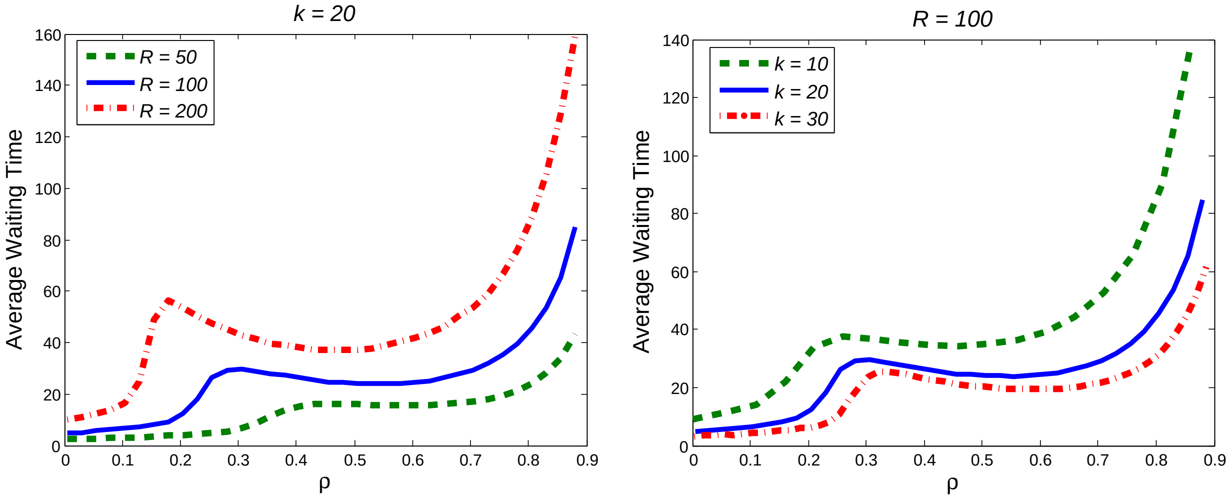

Figure 1 **Average waiting time of the one-direction call system under different R and k.**

idle taxis in the system, then the passenger's ride request would be responded to immediately by the nearest taxi. In this case, the response time is zero. Otherwise, the passenger has to wait until a taxi becomes available to which the passenger is the nearest.

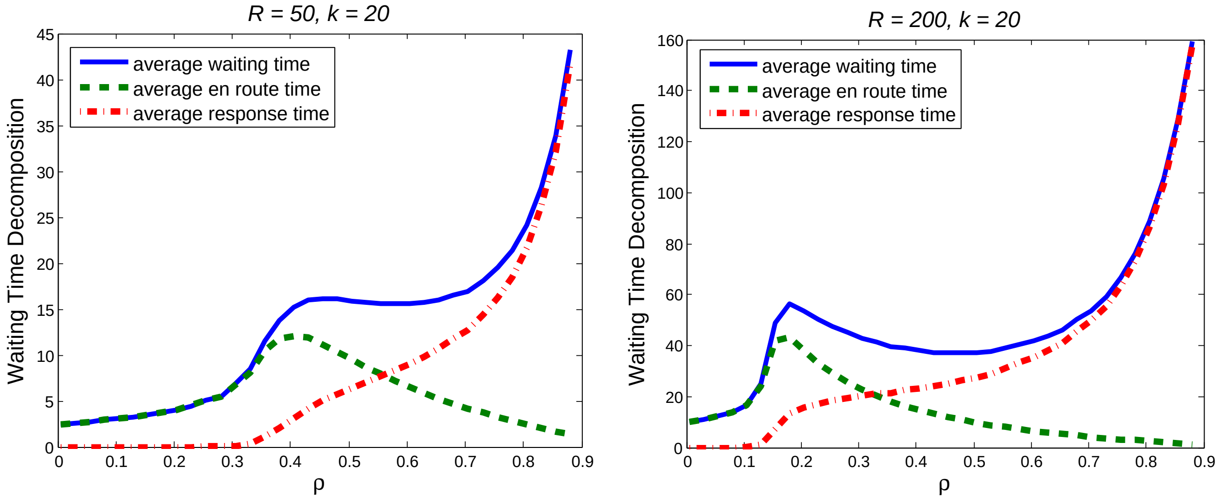

Figure 2 **Decomposition of waiting time into response time and en route time for one-direction call system.**

Figure 2 shows the decomposition of the average waiting time into the average response time and the average en route time when $R = 50$, $k = 20$, and $R = 200$, $k = 20$. From Figure 2, we can see that the average response time is monotone in ρ while the average en route time is not. It is apparent that the non-monotonicity of the average en route time leads to the non-monotonicity of the average waiting time. To understand the non-monotonicity of the average en route time, consider the case when ρ is small. In such a case, increasing ρ leads to fewer idle taxis on average, which means that the average distance between an arriving passenger and the nearest idle taxi is longer, and thus, the en route time increases with ρ in this range. When ρ is high, a larger ρ implies that, on average, more passengers are waiting for service in the system. Therefore, the average distance between the taxi and the nearest passenger decreases in ρ; thus the en route time decreases in ρ. When ρ is in the middle range (near where the en route time peaks), a passenger arriving at the system is likely to find him/herself among the few waiting passengers, and at the same time, not more than one taxi is available. In this case, the average distance between the passenger and the taxi is the largest, which explains the peak of the en route time. Moreover, this effect is more significant when R is large, because the en route time will play a more significant role in determining the total waiting time in that case. Therefore, the non-monotonicity is more pronounced when R is large.

Next we consider the one-direction no-call system. We perform similar numerical experiments, and the results are shown in Figure 3. From Figure 3, we make the following observation.

Observation 2. *In the one-direction no-call system, the average waiting time is monotonically increasing in ρ.*

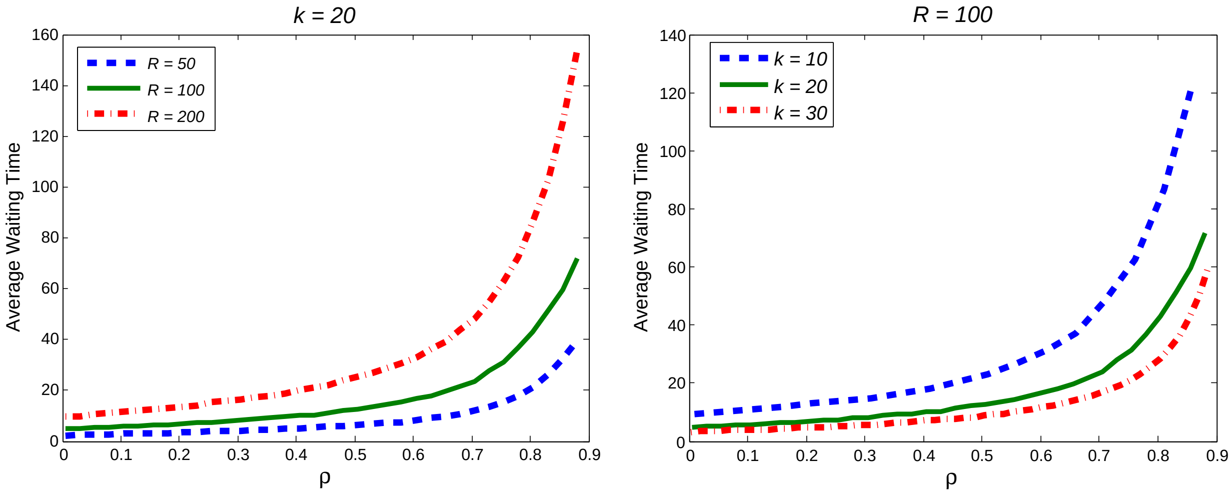

Figure 3 Average waiting time of the one-direction no-call system under different R and k.

The intuition behind the results shown in Figure 3 is as follows: In the no-call system, there are no exact counterparts of the response time and the en route time as in the call system. However, similar to the en route time in the call system, for each finally matched passenger-taxi pair, we define a measure named *the virtual en route time*, which is the time between the passenger and the taxi becoming available and the pickup moment. The virtual en route time has similar features as the en route time in the call system but also has some key differences. When ρ is small, the virtual en route time also increases in ρ since there will be fewer idle taxis as ρ increases. When ρ is large, the virtual en route time also decreases in ρ since there will be more passengers in the system as ρ increases and the average distance traveled between the time the taxi becomes available and encounters a passenger is shorter. However, the fluctuation of the virtual en route time is much smaller than that of the en route time in the call system. This is because in the no-call system, when the distance between a taxi and its nearest passenger is large, there is a high chance that the taxi will encounter another passenger en route and end up picking up the nearer passenger (since taxis are not committed to passengers). As a result, the virtual en route time in the no-call system is much less than the en route time in the call system. As we can see from Figure 4, the virtual en route time is much smoother than the en route time in Figure 2, which explains the monotonicity observed in Figure 3.

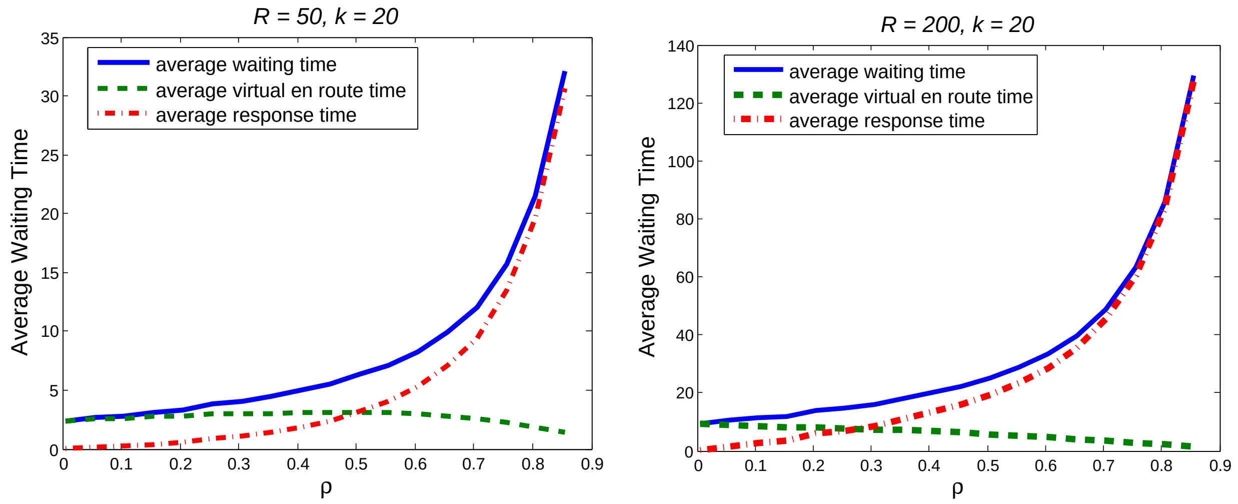

Figure 4 **Decomposition of waiting time into response time and virtual en route time for one-direction no-call system.**

Next, we compare the one-direction call and no-call systems. The numerical results are shown in Figure 5. We have the following observation.

Observation 3. *The average waiting time in the one-direction no-call system is smaller than that in the one-direction call system.*

To explain this observation, we note that in the one-direction case, the no-call system always expedites the next pickup compared to the call system. More specifically, in the call system, the

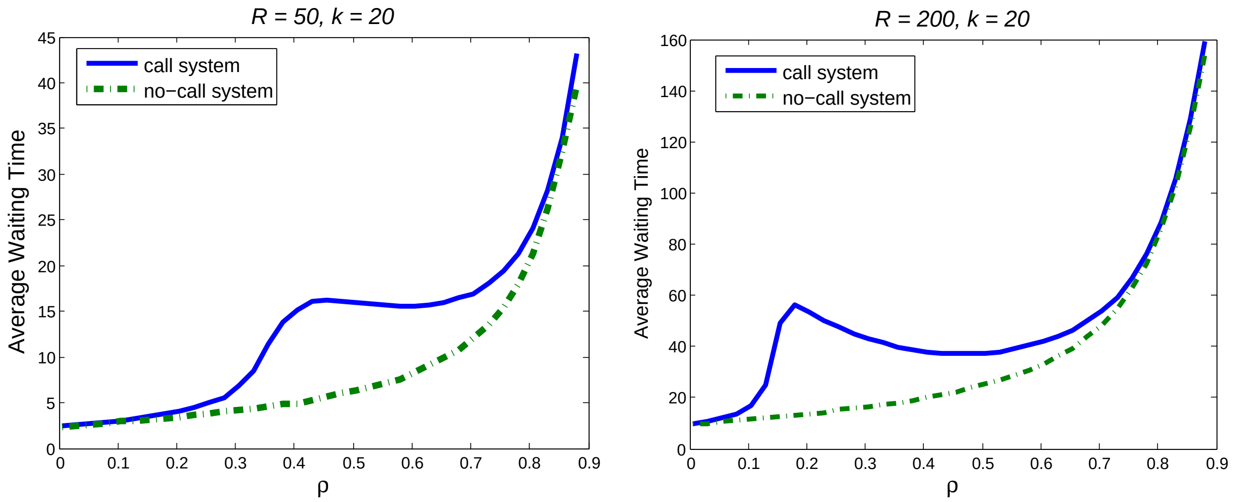

Figure 5 Comparison of the one-direction call and no-call systems.

nearest passenger and taxi are matched as soon as available; in contrast, in the no-call system, a passenger-taxi match is not established until the taxi passes by a waiting passenger, which provides the possibility of more efficient matches. In particular, the advantage of the no-call system can be demonstrated with two possible scenarios. For the current nearest pair of passenger and taxi: 1) the passenger might be picked up by another taxi that becomes available later and has a shorter distance to the passenger upon its availability; 2) the taxi may encounter a new arriving passenger with a shorter distance to the taxi. In either scenario, the no-call system shortens the next pickup time by postponing the decision-making time and thus achieves shorter average waiting time.

In addition, the difference between the call and no-call systems is most significant around the point when the call system has the highest average en route time. We note that when ρ approaches zero, the average waiting times for both systems are the same. This is because when ρ approaches zero, all k taxis are available when a passenger arrives, and few new passengers would arrive before the next pickup. Therefore, call and no-call systems would provide the same match and thus have the same waiting time. We also note that when ρ approaches one, the average waiting times in the two systems converge and go to infinity. This is because when ρ is large, the number of waiting passengers in both systems is large, and thus, the en route time is negligible compared to the response time. Therefore, the difference between the call and no-call systems is small.

Above we have made important observations about the one-direction systems and provided intuitive explanations. In the next subsection, we propose an approximation scheme and provide theoretical justifications for those observations. As we shall see from the theoretical analysis, the observations we made are likely to be generally true in such systems.

4.2. Approximation Scheme

In this subsection, we propose an approximation scheme for the one-direction system studied above. The approximation scheme is useful in several ways. First, it provides an efficient way to obtain performance measures for such systems, rather than performing potentially burdensome simulations. Second, it captures the key characteristics of different matching mechanisms and thus allows us to obtain meaningful insights. As we will show, under the approximation scheme, we can prove several insights we observed in the actual system. Moreover, the approximation scheme allows extension to more complex settings, such as other road setups or matching mechanisms, thus offering a powerful tool for analyzing such problems.

To establish the approximation scheme, we note that the transportation system shares some similarities with a queueing system. In particular, taxis can be viewed as servers, while passengers waiting for a ride can be viewed as customers waiting for service. The arrival rate of passengers and the service rate of taxis are also similar to the counterparts in a queueing system. However, unlike a queueing system where customers are served according to certain priority rules (e.g., first-come first-served), in the transportation system, the locations of taxis and passengers play a key role in determining the sequence of service. In addition, the waiting time for a passenger is not only impacted by waiting for busy taxis to become available, but also affected by the time en route.

To incorporate these features, we define the *extended service time* in the transportation system, which is the sum of the actual service time and the en route time. In our model, the service time is exponentially distributed with mean $1/\mu$. However, the explicit distribution of the en route time is hard to obtain. To address this issue, in the approximation scheme, we first compute an approximate expected en route time for the next service, which we denote by $\bar{t}_e$. We note that $\bar{t}_e$ is dependent on the number of waiting passengers and available taxis in the system. Then we approximate the extended service time by assuming that it follows an exponential distribution with mean $1/\mu + \bar{t}_e$. Finally, with this approximation for the extended service time, we approximate the transportation system by an $M/M/k$ queue, in which arrivals form a single queue and are governed by a Poisson process, and there are k servers with state-dependent service rates as previously described. In the following, we specify the $M/M/k$ approximations for the one-direction call/no-call systems.

For the one-direction call system, let n denote the total number of passengers in the system (waiting for service or being served), and $\lambda_n^{(1c)}$ and $\mu_n^{(1c)}$ denote the arrival rate and the service rate when there are n passengers in the system, respectively. By definition, $\lambda_n^{(1c)} = \lambda$ for all n. In the following, we calculate the average en route time in order to obtain $\mu_n^{(1c)}$.

In the one-direction call system, when $n<k$, the n-th passenger (the most recent arrival) would be matched to one of the $k-n+1$ idle taxis. In the approximation scheme, we assume the $k-n+1$ idle taxis are uniformly located on the road, labeled by $1,2,\ldots,k-n+1$. Let d_i denote the clockwise distance between the i-th idle taxi and the arriving passenger. We let $t_e^{(1c)}$ represent the en route time. Then $t_e^{(1c)} = \min\limits_{i=1,\ldots,k-n+1} d_i$ is the shortest distance between the taxis and the passenger (since the speed of taxi is one, $t_e^{(1c)}$ is also the pickup time). We have

$$\mathbb{P}(t_e^{(1c)} \geq x) = \prod_{i=1}^{k-n+1} \mathbb{P}(d_i \geq x) = \left(\frac{R-x}{R}\right)^{k-n+1},$$

and thus, the expected en route time $\bar{t}_e^{(1c)} = \mathbb{E}[t_e^{(1c)}] = \int_0^R \mathbb{P}(t_e^{(1c)} \geq x)dx = \frac{R}{k-n+2}$.

When $n>k$, there are $n-k$ passengers in the system waiting to be matched. When a taxi becomes available, it would be immediately matched to the nearest passenger. In the approximation scheme, we assume that the waiting passengers are uniformly located on the road; therefore, by a similar argument as in the $n<k$ case, we have $\bar{t}_e^{(1c)} = \frac{R}{n-k+1}$. When $n=k$, we approximate the en route time by $R/2$, which is the average distance between a taxi and a passenger of random locations. To summarize, we approximate the one-direction call system by an $M/M/k$ queue with state-dependent arrival rate $\lambda_n^{(1c)} = \lambda$ and service rate

$$\mu_n^{(1c)} = \begin{cases} \frac{n}{1/\mu + R/(k-n+2)}, & \text{if } n \leq k, \\ \frac{k}{1/\mu + R/(n-k+1)}, & \text{if } n > k. \end{cases} \tag{1}$$

Figure 6 shows the comparison between the average waiting time under the approximation system versus the true system. As we can see, the approximation scheme provides a good approximation for the average waiting time of the true system. In particular, it retains the important features of the original system.[5]

Next, using the approximation system given by (1), we prove the properties described in Observation 1. In the following, let $\mathcal{W}^{(1c)}(\lambda,\mu,k,R)$ be the average waiting time under the approximation scheme given parameters λ, μ, k and R. We have the following result:

THEOREM 1. *Consider the approximated one-direction call system given by (1).*

1. $\mathcal{W}^{(1c)}(\lambda,\mu,k,R) < \infty$ *if and only if* $0 \leq \lambda < k\mu$.
2. *For any given μ, k and R,* $\left.\frac{\partial \mathcal{W}^{(1c)}(\lambda,\mu,k,R)}{\partial \lambda}\right|_{\lambda=0} \geq 0$ *and* $\underline{\lim}_{\lambda \to k\mu-} \frac{\partial \mathcal{W}^{(1c)}(\lambda,\mu,k,R)}{\partial \lambda} > 0$.

[5] In most cases, the approximation system gives an underestimate of the average waiting time compared to the true system, especially when the utilization is high. This may be because we "reduced" some variance in the system by assuming the idling taxis and the waiting passengers are uniformly located on the road and the extended service time follows an exponential distribution (the true extended service time tends to have a longer tail). Such reduction in variance in our modeling leads to the smaller estimation of the average waiting time when the utilization is high.

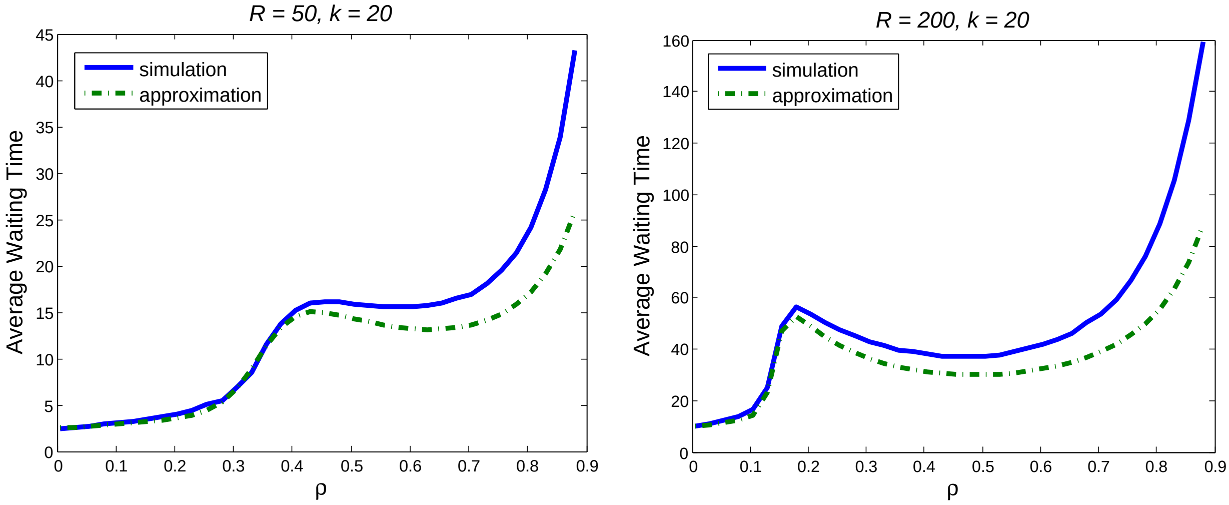

Figure 6 **Performance of the approximation scheme for the one-direction call system.**

3. *For any given* μ *and* $k \geq 2$*, there exists a constant* $R^*(\mu, k)$ *such that when* $R > R^*(\mu, k)$*, there exists* $0 \leq \lambda(\mu, k, R) < k\mu$ *such that* $\frac{\partial \mathcal{W}^{(1c)}(\lambda, \mu, k, R)}{\partial \lambda}\Big|_{\lambda = \lambda(\mu, k, R)} < 0$.

The proofs of Theorem 1 and all subsequent theorems are relegated to the Appendix. Theorem 1 provides strong support for Observation 1 under the approximation scheme. Note that the monotonicity property in Theorem 1 is equivalent to the monotonicity property in Observation 1. In particular, the second part of Theorem 1 shows that when λ is very small (close to 0) or very large (close to the high-traffic regime where $\lambda = k\mu$), the average waiting time is increasing in λ. Moreover, the third part of Theorem 1 states that given any k, as long as the road length R is large enough, there will be a range of λ such that the average waiting time is decreasing in λ. These results are all consistent with the findings in Observation 1. Therefore, with the help of the approximation scheme, we are able to verify that the interesting findings we obtained in the simulation are generally true in such systems.

Next, we propose an approximation scheme for the one-direction no-call system. We apply the same idea as in the call system and approximate the system with a state-dependent $M/M/k$ queue. Let n denote the number of passengers in the system and $\lambda_n^{(1n)}$ and $\mu_n^{(1n)}$ denote the arrival and service rates when there are n passengers in the system. Apparently, the arrival rate $\lambda_n^{(1n)} = \lambda$. We now approximate the average en route time $\bar{t}_e^{(1n)}$ when there are n passengers in the system, in order to obtain the service rate $\mu_n^{(1n)}$.

When $n > k$, there are more passengers than the total number of taxis. Though the number of taxis in service could be smaller than k, we simplify the analysis by assuming that all k taxis are in service and $n - k$ passengers are waiting for service. Let $t_e^{(1n)}$ be the en route time, which is the pickup time for the next available taxi in this case. The event $t_e^{(1n)} \geq x$ implies that no passenger is located within clockwise distance x of the taxi, and the taxi does not encounter any new passenger arriving within time period x. Here we ignore the possibility of another taxi becoming available during this interval and picking up a passenger before this taxi. Let l_i denote the clockwise distance between the i-th waiting passenger and the taxi. Based on the discussions above, we have

$$\mathbb{P}(t_e^{(1n)} \geq x) = \mathbb{P}\left(\min_{1\leq i\leq n-k} l_i \geq x\right)\mathbb{P}(A),$$

where A denotes the event that the taxi encounters no new passenger arriving within time period x (note that the event A depends only on future arrivals and is independent of the event that $\min_{1\leq i\leq n-k} l_i \geq x$).

Similar to the analysis for the call system, it is easy to see that $P(\min\limits_{1\leq i\leq n-k} l_i \geq x) = (\frac{R-x}{R})^{n-k}$. Now we compute the probability of event A. We note that the new passenger arrivals within time period x follow a Poisson process with arrival rate λ. For an arrival at time y, $y \leq x$, to be encountered by the taxi before time x, the passenger must be located between y and x, which is of length $x-y$. By the assumption of the uniformly distributed arrival location, the probability that an arrival could be in a region of length $x-y$ is $\frac{x-y}{R}$; therefore, the arrivals that would be encountered by the taxi within time x follow a Poisson process with non-homogenous arrival rate $\lambda_y = \lambda\frac{x-y}{R}$ at any time $y \leq x$. Consequently, the total number of arrivals encountered by the taxi within time x, which we denote by N, follows a Poisson distribution with mean $\int_0^x \lambda\frac{x-y}{R}dy = \frac{\lambda x^2}{2R}$. Therefore,

$$\mathbb{P}(A) = \mathbb{P}(N=0) = \exp\left(-\frac{\lambda x^2}{2R}\right),$$

and thus, the expected en route time $\bar{t}_e^{(1n)} = \mathbb{E}[t_e^{(1n)}] = \int_0^R \mathbb{P}\left(t_e^{(1n)} \geq x\right)dx = \int_0^R \left(\frac{R-x}{R}\right)^{n-k}\exp\left(-\frac{\lambda x^2}{2R}\right)dx.$

Now we analyze the case when $n \leq k$. We assume that $n-1$ taxis are already in service, the n-th passenger would be picked up by one of the $k-n+1$ taxis that are idle and that the locations of the idle taxis are independent and uniformly distributed on the circular road. We approximate the waiting time for a new arrival passenger, which is also the en route time $t_e^{(1n)}$ for the taxi picking this passenger up. If we further assume no other passenger arrives before this passenger is picked up, then the event $t_e^{(1n)} \geq x$ is equivalent to no idle taxis being within the x clockwise distance from the passenger and no busy taxi freed up within the $x-y$ clockwise distance for any time point y satisfying $y \leq x$. Let B denote the latter event (no busy taxi frees up within $x-y$ clockwise distance for any time point y satisfying $y \leq x$), and let d_i denote the clockwise distance between the i-th idle taxi and the passenger. By similar arguments as in the $n > k$ case, we have

$$\mathbb{P}(t_e^{(1n)} \geq x) = \mathbb{P}\left(\min_{1\leq i\leq k-n+1} d_i \geq x\right)\mathbb{P}(B) = \left(\frac{R-x}{R}\right)^{k-n+1}\exp\left(-\frac{(n-1)\mu x^2}{2R}\right),$$

and $\bar{t}_e^{(1n)} = \mathbb{E}[t_e^{(1n)}] = \int_0^R \mathbb{P}(t_e^{(1n)} \geq x)dx = \int_0^R (\frac{R-x}{R})^{k-n+1} \exp(-\frac{(n-1)\mu x^2}{2R})dx$.

Thus, the one-direction no-call system can be approximated by an $M/M/k$ queue with arrival rate $\lambda_n^{(1n)} = \lambda$ and state-dependent service rate

$$\mu_n^{(1n)} = \begin{cases} \frac{n}{1/\mu + \int_0^R \left(\frac{R-x}{R}\right)^{k-n+1} \exp\left(-\frac{(n-1)\mu x^2}{2R}\right)dx}, & n \leq k, \\ \frac{k}{1/\mu + \int_0^R \left(\frac{R-x}{R}\right)^{n-k} \exp\left(-\frac{\lambda x^2}{2R}\right)dx}, & n > k. \end{cases} \tag{2}$$

Figure 7 shows numerical experiments that compare the approximated average waiting time using this scheme versus the true average waiting time of the one-direction no-call system. One can see that similar as the one-direction call system, the approximation scheme performs quite well except for large values of ρ.

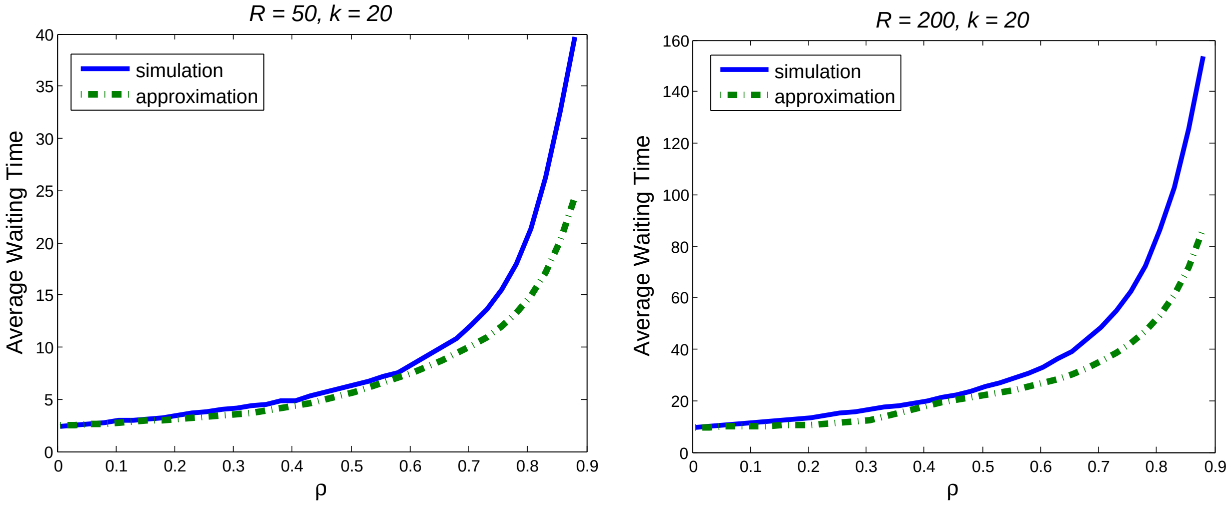

Figure 7 Performance of the approximation scheme for the one-direction no-call system.

In the following, let $\mathcal{W}^{(1n)}(\lambda, \mu, k, R)$ be the average waiting time under the approximated one-direction no-call system. We have the following theorem:

Theorem 2. *For any μ, k, there exists $R^*(\mu, k)$ such that when $R < R^*(\mu, k)$, $\mathcal{W}^{(1n)}(\lambda, \mu, k, R)$ is increasing in λ. Furthermore, for any λ, μ, k and R, $\mathcal{W}^{(1n)}(\lambda, \mu, k, R) \leq \mathcal{W}^{(1c)}(\lambda, \mu, k, R)$.*

Theorem 2 provides support for Observations 2 and 3 under the approximation scheme. By Theorem 2, in the approximated one-direction no-call system with a short road length, the waiting time is always increasing in λ, which is consistent with Observation 2.[6] Furthermore, under the approximation scheme, the average waiting time of the one-direction no-call system is always

[6] Unfortunately, we are not able to prove this statement for an arbitrary value of R.

smaller than that of the one-direction call system, which verifies Observation 3. Thus, by using the approximation scheme, we are able to provide justifications for most of the observations about the one-direction system.

5. Two-Direction Call/No-Call Systems

In this section, we study the two-direction call/no-call systems. As in Section 4, we first perform numerical experiments on the systems and summarize our observations. Then, we use an approximation scheme to obtain theoretical analysis.

5.1. Numerical Experiments

We start with the call system. As before, we study the relation between the average waiting time and the utilization level ρ under different values of k and R. The results are shown in Figure 8. From Figure 8, we make the following observation.

Observation 4. *In the two-direction call system, the average waiting time is not always monotonically increasing in ρ. In particular, it increases in ρ when ρ is small or large. However, it could decrease in ρ when ρ is medium. Moreover, such non-monotonicity is more pronounced when R is large and is less pronounced than that in the corresponding one-direction system.*

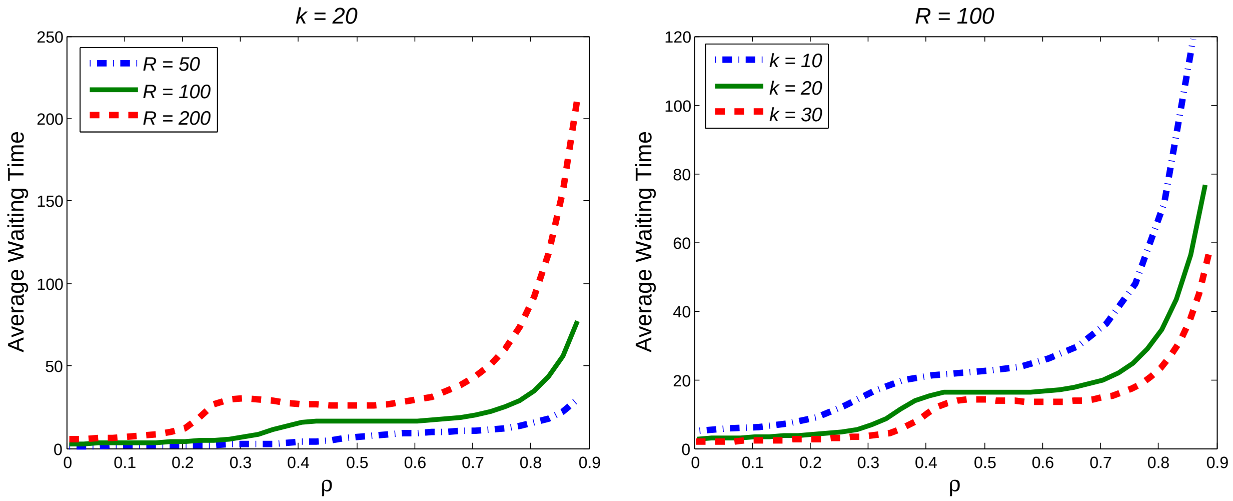

Figure 8 **Average waiting time of the two-direction call system under different R and k.**

Observation 4 is similar to Observation 1 in the one-direction call system. Recall that to explain Observation 1, we decompose the average waiting time into the response time and the en route time. Here we do the same, and the result is shown in Figure 9. In Figure 9, we can see that the same phenomenon as in the one-direction system still exists in the two-direction system, which explains the non-monotonicity of the waiting time in ρ. However, as can be seen by comparing Figure 1 and Figure 8, the non-monotonicity in the two-direction call system is less significant than that in the one-direction call system. This is because in the two-direction system, the taxis can be matched to passengers in both directions; thus, the matching distance is, in general, shorter. This implies that the en route time plays a less dominant role in determining the waiting time in the two-direction systems. To further illustrate this, we compare the average en route time in one-direction versus two-direction call systems in Figure 10. As one can see from Figure 10, the non-monotonicity of the en route time is much less significant in the two-direction system.

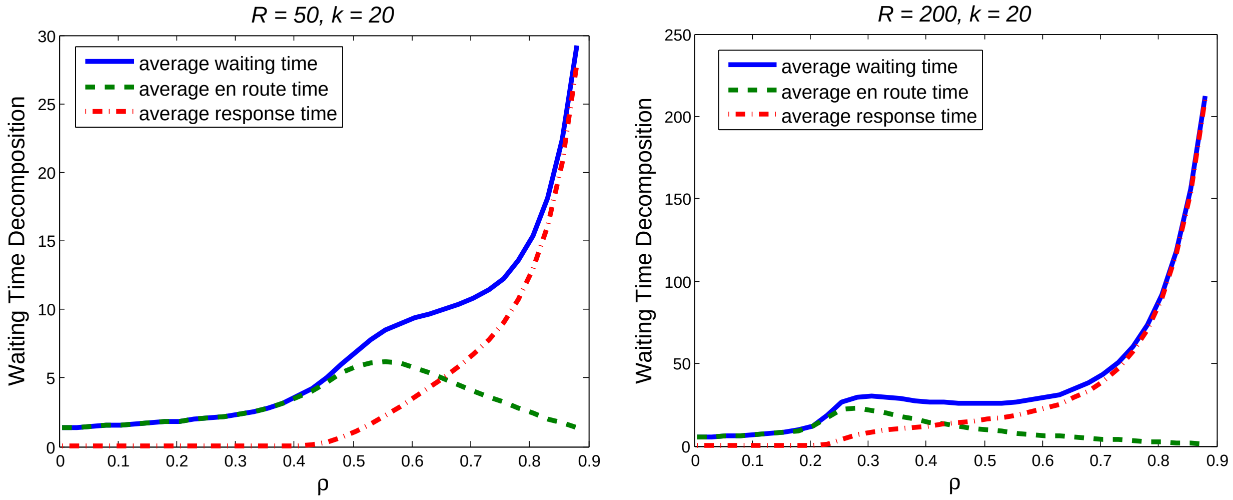

Figure 9 **Decomposition of the average waiting time for the two-direction call system.**

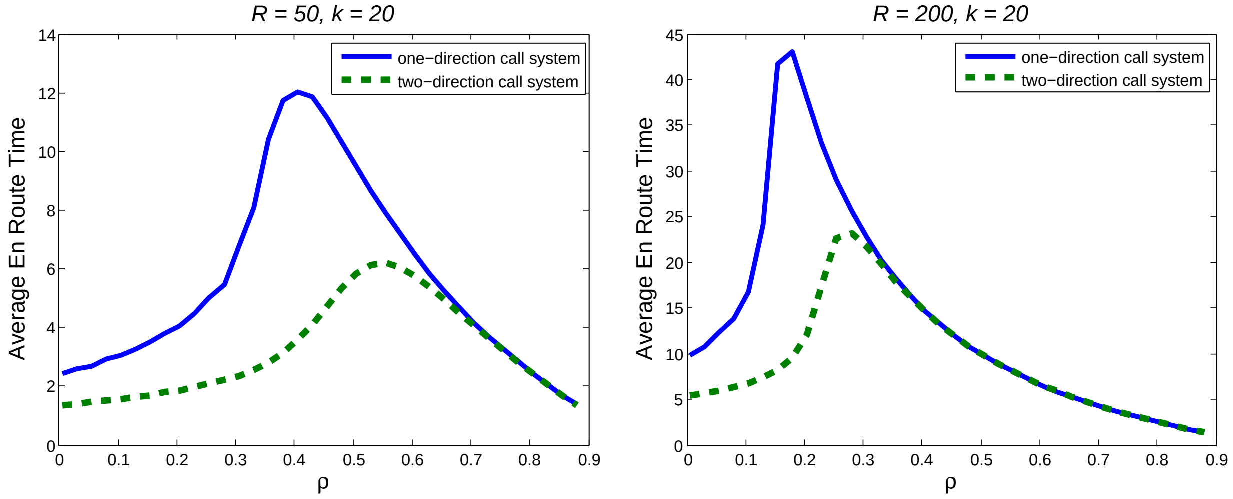

Figure 10 **Comparison of the average en route time of the one- and two-direction call systems.**

For the two-direction no-call system, we have exactly the same result as in the one-direction system, which we summarize in the following observation.

Observation 5. *The average waiting time in the two-direction no-call system is monotonically increasing in ρ.*

5.2. Approximation Scheme

In this section, we construct an approximation scheme for the two-direction systems. Using the approximation scheme, we provide justifications for Observations 4 and 5 and show parallel results as in Theorems 1 and 2.

We apply the idea introduced in Section 4.2, that is, to approximate the system by an $M/M/k$ queue. In particular, for the two-direction no-call system, since the matching mechanism is exactly the same as in the one-direction no-call system (except that taxis now are allowed to travel in both directions), the adopted approximation scheme for the two-direction no-call system is identical to that of the one-direction no-call system. That is, the approximation scheme for the two-direction no-call system is an $M/M/k$ queue with arrival rate λ and state-dependent service rates as in (2), i.e., $\mu_n^{(2n)} = \mu_n^{(1n)}$. And we have the following theorem:

THEOREM 3. *Let $\mathcal{W}^{(2n)}(\lambda,\mu,k,R)$ be the average waiting time in the approximated two-direction no-call system. For any μ, k, there exists $R(\mu,k)$ such that when $R < R(\mu,k)$, $\mathcal{W}^{(2n)}(\lambda,\mu,k,R)$ is increasing in λ.*

For the two-direction call system, we use the similar idea as in the one-direction call system. However, when computing the average en route time $\bar{t}_e^{(2c)}$, we should consider the distance for both directions. More specifically, let $t_e^{(2c)}$ be the en route time and d_i be the distance (in the two-direction sense) between the i-th taxi and a passenger in the two-direction call system. Then, when $n < k$, we have

$$\mathbb{P}(t_e^{(2c)} \geq x) = \prod_{i=1}^{k-n+1} \mathbb{P}(d_i \geq x) = \left(\frac{R-2x}{R}\right)^{k-n+1} \quad \text{and } \bar{t}_e^{(2c)} = \mathbb{E}[t_e^{(2c)}] = \int_0^{R/2} \mathbb{P}(t_e^{(2c)} \geq x)dx = \frac{R}{2(k-n+2)}.$$

Similarly, when $n > k$, we approximate the average en route time $\bar{t}_e^{(2c)}$ by $\frac{R}{2(n-k+1)}$ and when $n = k$, we approximate the average en route time by $R/4$ (which is the average distance between two uniformly random points on the circle, where the distance is in the two-direction sense). Therefore, we can approximate the two-direction call system by an $M/M/k$ queue with state-dependent arrival rate $\lambda_n^{(2c)} = \lambda$ and service rate

$$\mu_n^{(2c)} = \begin{cases} \frac{n}{1/\mu + \frac{R}{2(k-n+2)}}, \text{ if } n \leq k, \\ \frac{k}{1/\mu + \frac{R}{2(n-k+1)}}, \text{ if } n > k. \end{cases} \tag{3}$$

We have the following theorem for the approximated two-direction call system:

Theorem 4. *Consider the approximated two-direction call system given by (3). Let $\mathcal{W}^{(2c)}(\lambda,\mu,k,R)$ be the average waiting time given parameters λ, μ, k and R. Then we have:*

1. $\mathcal{W}^{(2c)}(\lambda,\mu,k,R) < \infty$ *if and only if* $0 \le \lambda < k\mu$.
2. *For any given μ, k and R,* $\frac{\partial \mathcal{W}^{(2c)}(\lambda,\mu,k,R)}{\partial \lambda}\Big|_{\lambda=0} \ge 0$ *and* $\underline{\lim}_{\lambda\to k\mu-} \frac{\partial \mathcal{W}^{(2c)}(\lambda,\mu,k,R)}{\partial \lambda} > 0$.
3. *For any given μ and $k \ge 2$, there exists a constant $R^*(\mu,k)$ such that when $R > R^*(\mu,k)$, there exists $0 \le \lambda(\mu,k,R) < k\mu$ such that* $\frac{\partial \mathcal{W}^{(2c)}(\lambda,\mu,k,R)}{\partial \lambda}\Big|_{\lambda=\lambda(\mu,k,R)} < 0$.

Figures 11 and 12 show comparisons between the approximated systems and the actual systems in the two-direction case. As one can see from Figures 11 and 12, the approximation scheme is also quite accurate in the two-direction systems, especially for small and medium values of ρ.

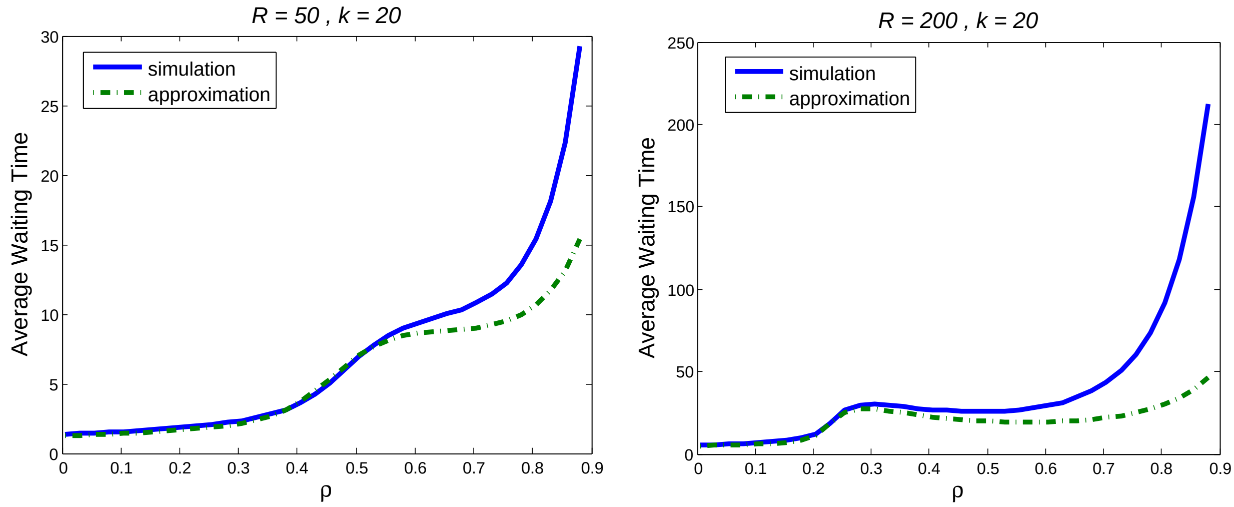

Figure 11 **Performance of the approximation scheme for the two-direction call systems.**

5.3. Comparison between Two-Direction Call and No-Call Systems

In this section, we compare the waiting time in the two-direction call system with that in the two-direction no-call system. We first obtain some numerical results in Figure 13. As shown in Figure 13, the average waiting time in the two-direction call system could be larger than that in the two-direction no-call system when the utilization level ρ is medium, especially when R is large. Otherwise, the average waiting time in the two-direction call system is smaller than that in the two-direction no-call system.

To understand this result, we pinpoint the disadvantage of both systems and study when these disadvantages are most significant. For the call system, its disadvantage can be illustrated by the

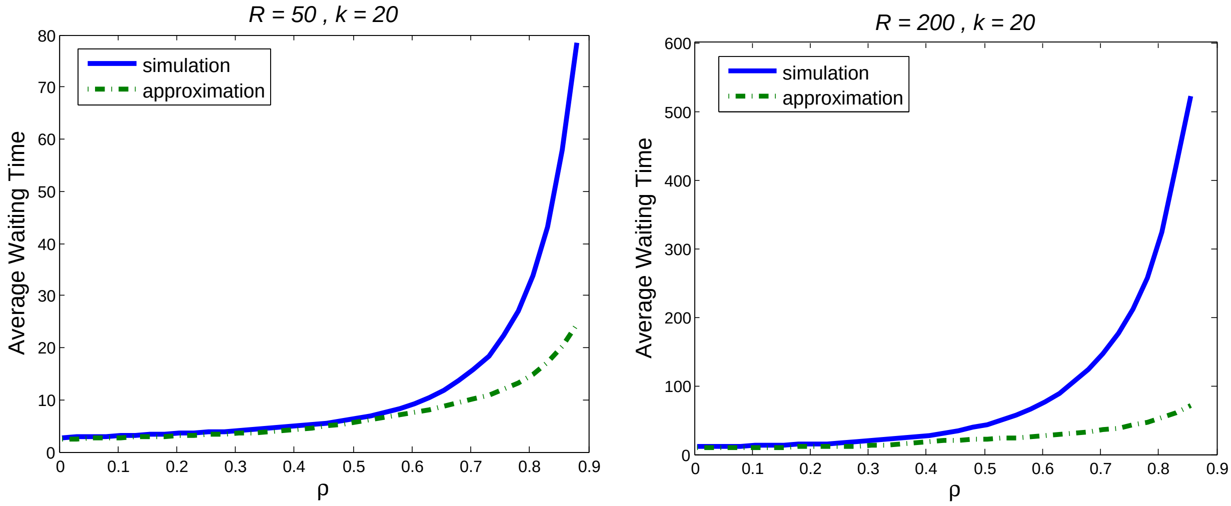

Figure 12 **Performance of the approximation scheme for the two-direction no-call systems.**

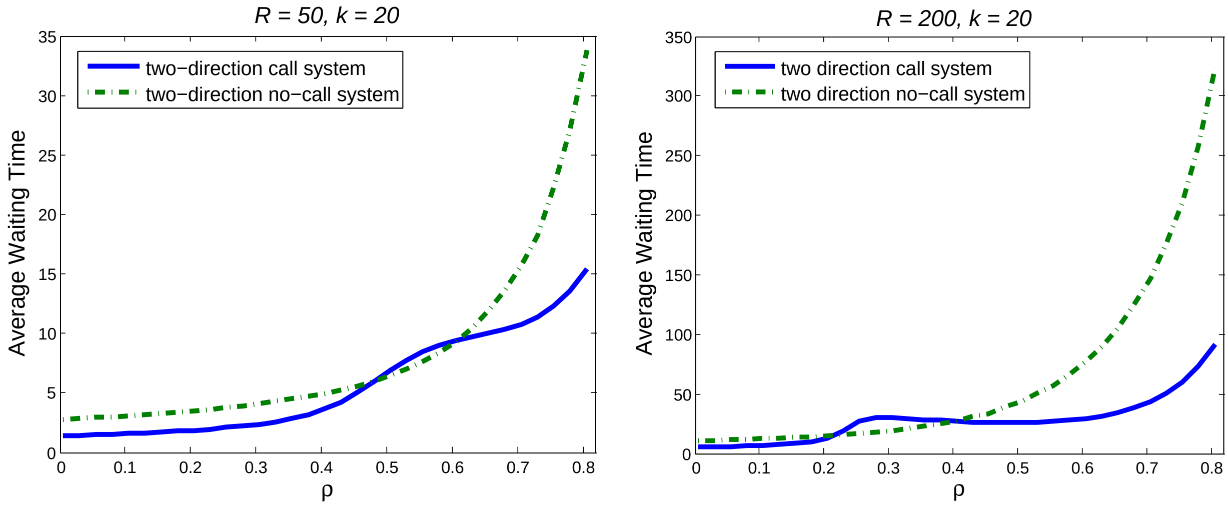

Figure 13 **Comparison of the two-direction call and no-call systems.**

following scenario: Suppose there is one waiting passenger and one idle taxi in the system, and their distance is large. Due to the call mechanism, the passenger will be immediately matched to the taxi, and the pickup time will be equal to their distance. However, such a matching is unlikely to be the most efficient one since the taxi may well encounter a new arrival passenger while it is en route to the matched passenger, or a taxi much closer to the passenger may become available later. In either case, the system would have been more efficient had the matching not happened. In particular, such an undesirable scenario is more likely to occur when the road is long (thus, the distance could be long compared to the service rate) and when the numbers of unserved passengers and idle taxis are small in the system. The latter occurs when the utilization ρ is in the middle range. We further note that a no-call system would not have such inefficiency since the matching would not have been established and the taxi (passenger, respectively) would be free to pick up new passengers (be picked up by other taxis, respectively).

For the no-call system, the disadvantage can be illustrated by the following scenario: Again, suppose there is only one passenger and one taxi in the system. This time, they are very close to each other. However, the taxi is driving away from this passenger. In the no-call system, there is no way to establish this would-be-efficient matching. Yet in the call system, the match would be established, and the passenger would be picked up immediately.

6. A Capped Matching Mechanism

In the last two sections, we investigate the characteristics of the call and no-call mechanisms and pinpoint their advantages and disadvantages. In the following, we focus on the two-direction systems as they are closer to reality. Recall that in the two-direction call system, the taxis have information about which driving direction is more efficient, but by matching to a passenger immediately, there is a possibility of missing a later-arrival passenger with a shorter pickup distance. This can be avoided in the two-direction no-call system. However, taxis in the two-direction no-call system lack the information about the location of the nearest passenger and thus could miss a nearby passenger by driving in the other direction.

In this section, we propose a modified mechanism that exploits the advantages and mitigates the disadvantages of the call and no-call mechanisms. In particular, we consider a call mechanism with a response cap. Specifically, the mechanism runs like the call mechanism except that a taxi and a passenger will not be matched unless their distance is within a designated cap. We call this mechanism the *capped matching mechanism* and the corresponding transportation system the *capped system*. The focus of this section is on studying the effect of adding such a cap on the average waiting time, as well as proposing a heuristic method to calculate a near-optimal cap.[7]

To start, we perform numerical tests of the capped matching mechanism under different caps c (from 0 to $R/2$) and find the cap that leads to the smallest average waiting time. Note that $c=0$ corresponds to the no-call mechanism and $c=R/2$ (the maximum distance between a taxi and a passenger in a circular road) corresponds to the call mechanism. The performance of the capped system with the optimal cap is shown in Figure 14.

[7] In theory, the cap could be chosen according to the state of the system. In the extreme case, the cap can be chosen such that the decision of whether to match a passenger and a taxi can depend on the entire history of the system. Such more complicated controls could improve the performance of the system. However, they will be very complicated to calculate or implement. Therefore, in this section, we focus on the case where the cap is a fixed constant. We will see that even such a relatively simple added lever could improve the efficiency of the system significantly.

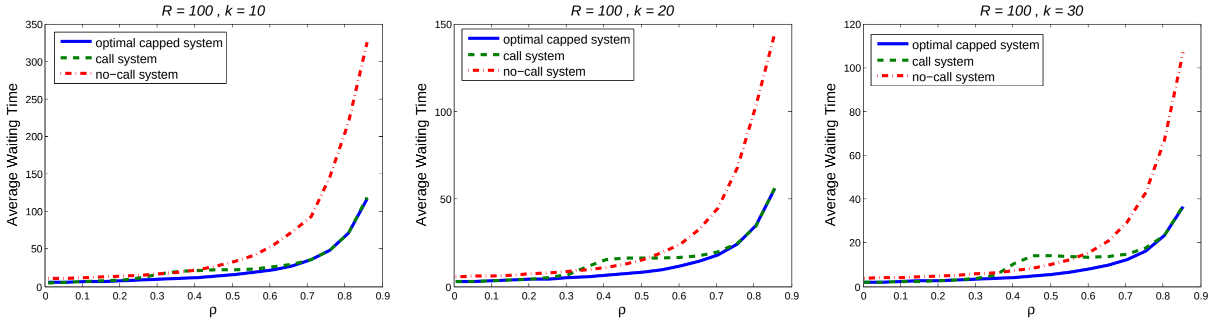

Figure 14 **Comparison of the two-direction optimal capped system, no-call system and call system.**

From Figure 14, we can see that the capped system with the optimal cap outperforms the call and no-call systems under different parameter settings. In particular, using an optimal cap could significantly reduce the average waiting time, and the reduction is more significant when there are more taxis in the system.

Having observed the potential value of adding a cap, a natural question is how to obtain a good cap without having to go through extensive simulations. In the following, we propose a heuristic method that provides a cap with good performance.

To find a good cap such that we should match a taxi and a passenger if their distance is within the cap and should not otherwise, we consider the following question: How close do a taxi and a passenger need to be so that matching them would be a *good* matching? To answer this question, we think from the taxi's point of view. Recall that if we establish a match between a taxi and a passenger, then the en route time for this taxi will be equal to the distance between the taxi and the passenger. Now we approximate the expected time this taxi has to drive in order to encounter a passenger if we do not establish the match. If the expected time is shorter than the distance from the current passenger, then intuitively the current match is not efficient, in which case we should not match them. Otherwise, matching will result in a shorter en route time than not matching, in which case we should match them.

Now it remains to calculate the expected time the taxi has to drive in order to encounter a passenger. To this end, we first compute the probability that a new passenger arrival will be encountered by this taxi within time x. Remember that new passengers arrive according to a Poisson process with rate λ. Similar to the discussions in Section 4.2, for an arrival at time y, $y \leq x$, to be encountered by the taxi before time x, the passenger must be located between y and x, which is of length $x-y$. By the assumption of the uniformly distributed arriving location, the probability that an arrival will be in a region of length $x-y$ is $\frac{x-y}{R}$; therefore, the arrivals that would be encountered by the taxi within time x follow a Poisson process with a non-homogenous arrival rate

$\lambda_y = \lambda \frac{x-y}{R}$ at any time $y \le x$. Consequently, the total number of arrivals encountered by the taxi within time x, which we denote by N, follows a Poisson distribution with mean $\int_0^x \lambda \frac{x-y}{R} dy = \frac{\lambda x^2}{2R}$. Therefore, the probability that it takes longer than x for a taxi to encounter a passenger equals

$$\mathbb{P}(N=0) = \exp\left(-\frac{\lambda x^2}{2R}\right),$$

and the expected time for a taxi to encounter a passenger is

$$\int_0^\infty \exp\left(-\frac{\lambda x^2}{2R}\right) dx = \sqrt{\frac{\pi R}{2\lambda}}.$$

Thus, based on the above idea, we choose a heuristic cap c^* to be $c^* = \min\left\{R/2, \sqrt{\frac{\pi R}{2\lambda}}\right\}$. Here, we add the term $R/2$ because for any $c^* \ge R/2$, the mechanism will be equivalent to a call mechanism.

Next, we test the performance of our proposed heuristic cap. We define the extra average waiting time as the waiting time difference divided by the waiting time in the optimal capped system. Figure 15 shows the extra average waiting time (in percentage) of the heuristic capped system, call and no-call systems compared to the optimal capped system.

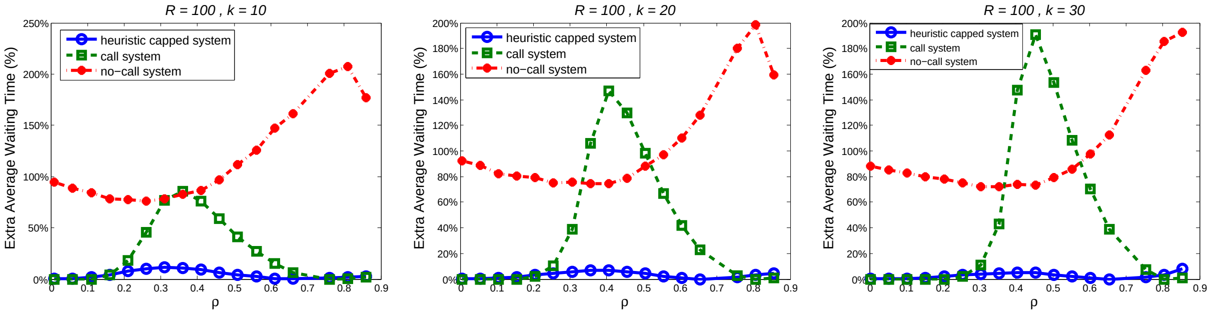

Figure 15 Extra average waiting time (%) of call and no-call systems compared to the optimal capped system.

From Figure 15, we can see that the average waiting time in the no-call/call system could be triple the average waiting time in the optimal capped system in the worst case. However, the heuristic cap performs quite well, with no more than 12% extra average waiting time versus using the optimal cap. Therefore the heuristic capped system has much better performance than both the call and no-call systems. And such a mechanism is also very easy to implement in practice.

7. Two-Dimensional Grid Network

In this section, we extend our numerical study to a more realistic setting: a two-dimensional grid network. The two-dimensional grid network provides a closer model for city road networks and has been used extensively in transportation-related studies (see, e.g., Fawaz and Newell 1976).

In particular, we consider a rectangular region that is divided into equally sized sub-squares by vertical and horizontal roads. The sub-squares may be viewed as blocks in the city. Mathematically, an $m \times n$ two-dimensional grid network is comprised of roads of $\{x = k, 1 \leq y \leq n\}$, $k = 1, \ldots, m$ and $\{y = l, 1 \leq x \leq m\}$, $l = 1, \ldots, n$. In the numerical experiments, we simulate the call and no-call systems in the two-dimensional grid network to see if the characteristics of these systems in the circular road setting still hold.

In the simulation, we assume that the passengers' arrival still follows a Poisson process and they arrive only at intersection points of the road, i.e., passengers arrive only at points (x, y) where $1 \leq x \leq m$ and $1 \leq y \leq n$ are both integers. This is mainly for the simplicity of analysis, yet it is also plausible in practice. Taxis drive on the road, and at each intersection point, they choose either to keep the current direction or turn left or right (each option is available only when it is eligible) in a uniformly random manner. Note that under this setup, a taxi would not drive back and forth between two adjacent intersection points, which is reasonable and intuitively more efficient in practice. We assume taxis drive one block per unit of time, and the system is updated at each integer point of time. In the no-call system, at the end of each time epoch, if an available taxi encounters a passenger at an intersection, then a match between this taxi and the passenger will be established immediately. In the call system, at the end of each time epoch, if there are available taxis in the system, then the passengers who arrive within the past unit of time will be matched one by one to the taxis in a first-come-first-serve manner. In either system, regardless of the pickup point, the passenger's destination is uniformly distributed over all intersection points in the grid. For the service rate, we first calculate the average distance (1-norm distance) between any two uniformly random points on the grid and choose μ to be the reciprocal of the average distance. By simple calculation, we have $\mu = 3mn/(m(n^2 - 1) + n(m^2 - 1))$.

In the following, we perform numerical experiments using grid networks of different sizes and with different numbers of taxis. As shown in Figure 16, the main features that we have observed in the circular road setting still hold. In particular, the no-call mechanism could be more efficient than the call mechanism when the utilization is in the middle range. Otherwise, the call mechanism is more efficient. Moreover, we observe that in the grid network, the call mechanism tends to be more efficient when there are few taxis in the system or when the grid is large. This is because if the grid is large (with many roads in the grid), then the call mechanism would be more advantageous in helping taxis find passengers. Imagine a network with numerous roads; in that case, it is hard for a taxi to find a passenger without using a call mechanism. Similarly, when there are very few taxis, it is hard for them to find passengers without using the call mechanism too. Thus, the call mechanism is more efficient in those settings.

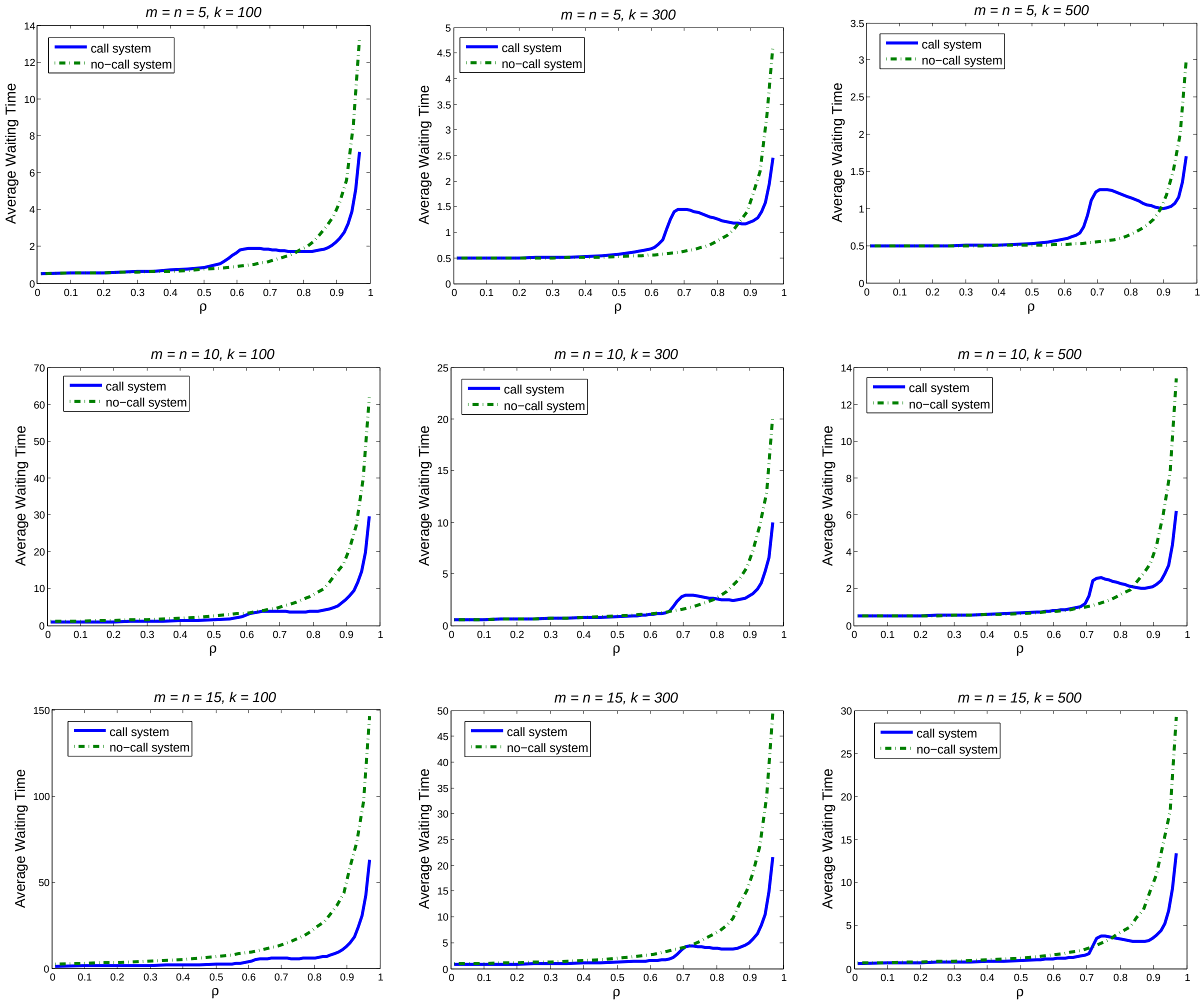

Figure 16 Numerical results for grid networks.

8. Conclusion

The emergence of on-demand ride-hailing systems over the past few years has greatly simplified the process of requesting transportation services. Although the on-demand platform has potential advantages in directing drivers to passengers in a timely manner, the platform also has its own challenges: A driver who commits to serve a passenger may miss the chance to serve another incoming passenger who is a shorter distance away. In this paper, we examine the average waiting time of on-demand hailing and traditional street-hailing systems using a stylized model and $M/M/k$ queueing approximations.

In the one-direction system, the average waiting time under the no-call system always increases with system utilization as expected. However, the average waiting time under the call system can decrease and then increase due to the non-monotonicity of the average en route time in the call system. When the system utilization is very high (with many waiting passengers) or very low (with many idle drivers), the average en route time is low, whereas the average en route time reaches the highest when the system utilization is medium, i.e., the system is balanced. In addition, the average waiting time under the call system is always higher than that in the no-call system. This is because a call system may have the disadvantage of forgoing possible future matching opportunities when accepting an incoming request, while a no-call system circumvents it by postponing the "matching" between a driver and a passenger as late as possible. That is, no matching is established until a driver is passing a passenger.

In the two-direction system, the average waiting time under a no-call system still increases with system utilization. In addition, albeit less pronounced, the non-monotonicity of the average waiting time under a call system is still observed. However, unlike in the one-direction system, a call system can be better than a no-call system in the two-direction system despite the disadvantage of forgoing possible future matching opportunities because a call system informs the driver of the shortest distance to pick up a passenger. It is especially true when the system utilization is either very low or very high. When the system utilization is medium, a no-call system may still perform better than a call system in reducing the average waiting time.

Based on the understanding of the two different matching mechanisms in one and two directions, we sought to address the matching inefficiency that arises with the on-demand hailing platform by proposing a distance cap in responding to requests from passengers. The distance cap, on one hand, reserves the advantages of the on-demand hailing system, while on the other hand helps in limiting the possibility of serving an incoming passenger. We propose a heuristic way to calculate the distance cap that is not only implementable but also effectively reduces average waiting time in the on-demand hailing system. In a more complex grid network, we show that properties observed in the circular road system still exist. A no-call system may perform better than a call system when there are more taxis or when the network is less complicated.

In sum, our analysis shows that the on-demand hailing system is more efficient than the traditional street-hailing system in some circumstances while it is less efficient in others. Awareness of advantages and disadvantages of the on-demand hailing system and street hailing would assist policy makers decide when and where to adopt on-demand hailing or street hailing. In particular, we show that street hailing may have its own value in operations and should be reserved under certain conditions. This research may provide justification for the New York City government to allow a certain number of taxis dedicated to street hailing through Street Hail Livery (SHL), and evidence that it may not be unreasonable for the Shanghai government to ban on-demand hailing apps during certain times of the day from an operations aspect. Our paper provides guidance for evaluating the efficiency of each system. In addition, to take advantage of both systems, we propose a simple yet effective solution: using a distance cap to avoid the inefficiency in matching through an on-demand platform, and thus significantly reduce the average waiting time. Our work will inspire a few interesting future research directions, such as the accuracy of the approximation scheme, the extension of the theoretical results to more complex road networks, and the validation of our findings using real data.

References

Akbarpour, M., S. Li, O. G. Shayan. 2016. Thickness and information in dynamic matching markets. Working Paper.

Allon, G., A. Bassamboo, E. B. Çil. 2012. Large-scale service marketplaces: The role of the moderating firm. *Management Science* **58**(10) 1854–1872.

Anderson, R., I. Ashlagi, D. Gamarnik, Y. Kanoria. 2015. A dynamic model of barter exchange. *In SODA'15: Proceedings of the 26th Annual ACM-SIAM Symposium on Discrete Algorithms*. 1925–1933.

Baccara, M., S. Lee, L. Yariv. 2015. Optimal dynamic matching. Working Paper.

Bailey, W., T. Clark. 1992. Taxi management and route control: A systems study and simulation experiment. *Proceedings of the 24th Conference on Winter Simulation*. 1217–1222.

Benjaafar, S., G. Kong, X. Li, C. Courcoubetis. 2015. Peer-to-peer product sharing: Implications for ownership, usage and social welfare in the sharing economy. Working Paper.

Cachon, G. P., K. M. Daniels, R. Lobel. 2015. The role of surge pricing on a service platform with self-scheduling capacity. Working Paper.

Cullen, Z., C. Farronato. 2014. Outsourcing tasks online: Matching supply and demand on peer-to-peer internet platforms. Working Paper.

Fawaz, M. Y., G. F. Newell. 1976. Optimal spacings for a rectangular grid transportation network: A hierarchy structure. *Transportation Research* **10**(2) 111–119.

Fraiberger, S. P., A. Sundararajan. 2015. Peer-to-peer rental markets in the sharing economy. Working Paper.

Gurvich, I., M. Lariviere, A. Moreno. 2015. Operations in the on-demand economy: Staffing services with self-scheduling capacity. Working Paper.

Hawkins, A. J. 2016a. It took Uber five years to get to a billion rides, and its Chinese rival just did it in one. *The Verge* .

Hawkins, A. J. 2016b. Uber just completed its two-billionth trip @Verge. `http://www.theverge.com/2016/7/18/12211710/uber-two-billion-trip-announced-kalanick-china-didi`. Accessed: 2017-04-21.

Hu, M., Y. Zhou. 2016. Dynamic type matching. Working Paper.

Jiang, B., L. Tian. 2016. Collaborative consumption: Strategic and economic implications of product sharing. *Management Science.* Forthcoming.

McLeod, M. G. 1972. The operation and performance of a taxi fleet. Master's thesis, Massachusetts Institute of Technology.

Meyer, R., H. Wolfe. 1961. The organization and operation of a taxi fleet. *Naval Research Logistics Quarterly* **8**(2) 137–150.

Riquelme, C., S. Banerjee, R. Johari. 2015. Pricing in ride-share platforms: A queueing-theoretic approach. Working Paper.

Rogers, B. 2015. The social costs of Uber. *University of Chicago Law Review Dialogue* **82**.

Steinberg, J. 2012. Smartphone taxi e-hail apps: New convenience or potential deathtrap. *The Forbes* .

Tang, C. S., J. Bai, K. C. So, X. Chen, H. Wang. 2016. Coordinating supply and demand on an on-demand platform: Price, wage, and payout ratio. Working Paper.

Taylor, T. 2016. On-demand service platforms. Working Paper.

Welch, P. 1983. The statistical analysis of simulation results. *The Computer Performance Modeling Handbook*. Academic Press, 268–328.

Wood, D. 1992. *The Computation of Polylogarithms*. University of Kent at Canterbury.

ONLINE APPENDIX

Lemma 1. *Consider an $M/M/k$ queue with arrival rate λ and state-dependent service rate*

$$\mu_n = \begin{cases} \frac{n(k-n+2)}{R}, \text{ if } n \leq k, \\ \frac{k(n-k+1)}{R}, \text{ if } n > k. \end{cases}$$

Let $\overline{\mathcal{W}}(\lambda, k, R)$ be the average waiting time. Then for any $R > 0$ and $k \geq 2$, there exists an arrival rate $\lambda^*(k,R)$ such that $\frac{\partial \overline{\mathcal{W}}(\lambda,k,R)}{\partial \lambda}\Big|_{\lambda=\lambda^*(k,R)} < 0$.

Proof of Lemma 1. By standard queueing theory, $\overline{\mathcal{W}}(\lambda,k,R) = \frac{\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}}{1+\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}}$. For $\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}$, we have $\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} = \sum_{i=1}^{k-1} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} + \sum_{i=k}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}$, where

$$\begin{aligned}
\sum_{i=k}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} &= \frac{\lambda^k}{\mu_1\mu_2\cdots\mu_k}\left(1+\sum_{i=k+1}^{\infty}\frac{\lambda^{i-k}}{\mu_{k+1}\cdots\mu_i}\right) \\
&= \frac{\lambda^k}{\mu_1\mu_2\cdots\mu_k}\left(1+\sum_{i=k+1}^{\infty}\frac{(\lambda R/k)^{i-k}}{(i-k+1)!}\right) \\
&= \frac{\lambda^k}{\mu_1\mu_2\cdots\mu_k}\frac{k}{\lambda R}\left((\lambda R/k)+\sum_{i=k+1}^{\infty}\frac{(\lambda R/k)^{i-k+1}}{(i-k+1)!}\right) \\
&= \frac{\lambda^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}}\sum_{i=1}^{\infty}\frac{(\lambda R/k)^i}{i!} = \frac{\lambda^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}}(e^{\lambda R/k}-1).
\end{aligned}$$

Similarly, for $\sum_{i=1}^{\infty} i\frac{\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}$, we have $\sum_{i=1}^{\infty} i\frac{\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i} = \sum_{i=1}^{k-1} i\frac{\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i} + \sum_{i=k}^{\infty} i\frac{\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}$, where

$$\begin{aligned}
\sum_{i=k}^{\infty} i\frac{\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i} &= \frac{\lambda^{k-1}}{\mu_1\mu_2\cdots\mu_k}\left(k+\sum_{i=k+1}^{\infty} i\frac{\lambda^{i-k}}{\mu_{k+1}\cdots\mu_i}\right) \\
&= \frac{\lambda^{k-1}}{\mu_1\mu_2\cdots\mu_k}\frac{k}{\lambda R}\left(k\frac{\lambda R}{k}+\sum_{i=k+1}^{\infty} i\frac{(\lambda R/k)^{i-k+1}}{(i-k+1)!}\right) \\
&= \frac{\lambda^{k-2}}{2\mu_1\mu_2\cdots\mu_{k-1}}\sum_{i=k}^{\infty} i\frac{(\lambda R/k)^{i-k+1}}{(i-k+1)!} \\
&= \frac{\lambda^{k-2}}{2\mu_1\mu_2\cdots\mu_{k-1}}\left((k-1)\sum_{i=k}^{\infty}\frac{(\lambda R/k)^{i-k+1}}{(i-k+1)!}+\sum_{j=1}^{\infty}\frac{j(\lambda R/k)^j}{j!}\right) \\
&= \frac{\lambda^{k-2}}{2\mu_1\mu_2\cdots\mu_{k-1}}\left[(k-1)(e^{\lambda R/k}-1)+\frac{\lambda R}{k}e^{\lambda R/k}\right].
\end{aligned}$$

Therefore

$$\overline{\mathcal{W}}(\lambda,k,R) = \frac{\sum_{i=1}^{k-1} i\frac{\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i} + \frac{\lambda^{k-2}}{2\mu_1\mu_2\cdots\mu_{k-1}}[(k-1)(e^{\lambda R/k}-1)+\frac{\lambda R}{k}e^{\lambda R/k}]}{1+\sum_{i=1}^{k-1}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}+\frac{\lambda^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}}(e^{\lambda R/k}-1)}.$$

For any given R and $k \geq 2$, we note that $\overline{\mathcal{W}}\big|_{\lambda=0} = R/(k+1)$ and $\overline{\mathcal{W}}\big|_{\lambda\to\infty} = R/k$. Also we claim that for any given R and $k \geq 2$, there exists $\tilde{\lambda}$ such that $\overline{\mathcal{W}}\big|_{\lambda=\tilde{\lambda}} > R/k$. To see this, when $k = 2$, we can simplify the expression of $\overline{\mathcal{W}}$ and get that $\overline{\mathcal{W}} = \frac{1+(1+\lambda R/2)\exp{(\lambda R/2)}}{\lambda+6/R+\lambda\exp{(\lambda R/2)}}$. By taking $\tilde{\lambda}$ satisfying $\exp{(\tilde{\lambda}R/2)} > 2+\tilde{\lambda}R/2$, we have $\overline{\mathcal{W}}\big|_{\lambda=\tilde{\lambda}} > R/2$. When $k>2$, we have

$$\overline{\mathcal{W}} > \frac{\frac{\lambda^{k-2}}{2\mu_1\mu_2\cdots\mu_{k-1}}[(k-1)(e^{\lambda R/k}-1)+\frac{\lambda R}{k}e^{\lambda R/k}]}{1+\sum_{i=1}^{k-1}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}+\frac{\lambda^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}}(e^{\lambda R/k}-1)} > \frac{\frac{\lambda^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}}(k-1+e^{\lambda R/k})\frac{R}{k}}{1+\sum_{i=1}^{k-1}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}+\frac{\lambda^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}}(e^{\lambda R/k}-1)}.$$

Taking $\tilde{\lambda}$ satisfying $\frac{k\tilde{\lambda}^{k-1}}{2\mu_1\mu_2\cdots\mu_{k-1}} > 1+\sum_{i=1}^{k-1}\frac{\tilde{\lambda}^i}{\mu_1\mu_2\cdots\mu_i}$ (note such $\tilde{\lambda}$ must exist since the left-hand side is greater than the right-hand side when $\tilde{\lambda}$ is large), we obtain $\overline{\mathcal{W}}\big|_{\lambda=\tilde{\lambda}} > R/k$.
Thus for any given R and $k\geq 2$, $\overline{\mathcal{W}}\big|_{\lambda=0} = R/(k+1)$ and $\overline{\mathcal{W}}\big|_{\lambda\to\infty} = R/k$, and there exists $\tilde{\lambda}$ such that $\overline{\mathcal{W}}\big|_{\lambda=\tilde{\lambda}} > R/k$. Also it is easy to see that $\overline{\mathcal{W}}$ is continuously differentiable in λ on $\lambda>0$. Therefore there exists $\lambda^*(k,R)$ such that $\frac{\partial\overline{\mathcal{W}}(\lambda,k,R)}{\partial\lambda}\Big|_{\lambda=\lambda^*(k,R)} < 0$. □

Lemma 2. *For any given $R>0$ and $k\geq 2$, there exist constants $\lambda^*(k,R)$ and $\mu^*(\lambda^*(k,R),k,R)$ such that $\frac{\partial\mathcal{W}(\lambda,\mu,k,R)}{\partial\lambda}\Big|_{\lambda=\lambda^*(k,R)} < 0$ for any $\mu > \mu^*(\lambda^*(k,R),k,R)$, where $\mathcal{W}(\lambda,\mu,k,R) = \frac{\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}}{1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}} - 1/\mu$, with $\mu_i = \mu_i^{(1c)}$ as defined in (**??**).*

Proof of Lemma 2. Take derivative of $\mathcal{W}(\lambda,\mu,k,R)$ with respect to λ, we have that

$$\frac{\partial\mathcal{W}(\lambda,\mu,k,R)}{\partial\lambda} = \frac{\sum_{i=2}^{\infty}\frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i}(1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}) - (\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i})^2}{(1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i})^2}.$$

Let $a_i(\mu) = \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}$, $b_i(\mu) = \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}$ and $c_i(\mu) = \frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i}$. We first show that for any μ,k,R and $0\leq\lambda<k\mu$, $\sum_{i=1}^{\infty}a_i(\mu)$, $\sum_{i=1}^{\infty}b_i(\mu)$ and $\sum_{i=2}^{\infty}c_i(\mu)$ are all finite. To do so, define $\epsilon = \frac{1}{2}(k/\lambda - 1/\mu) > 0$ and $N = \min\{i : 0\leq\frac{R}{(i-k+1)}\leq\epsilon\}$. We have

$$\begin{aligned}\sum_{i=1}^{\infty}a_i(\mu) &= \sum_{i=1}^{N}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} + \frac{\lambda^N}{\mu_1\mu_2\cdots\mu_N}\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_{N+1}\mu_{N+2}\cdots\mu_{N+i}}\\ &< \sum_{i=1}^{N}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} + \frac{\lambda^N}{\mu_1\mu_2\cdots\mu_N}\sum_{i=1}^{\infty}\frac{\lambda^i(1/\mu+\epsilon)^i}{k^i}\\ &< \sum_{i=1}^{N}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i} + \frac{\lambda^N}{\mu_1\mu_2\cdots\mu_N}\frac{k\mu+\lambda}{k\mu-\lambda} < \infty,\end{aligned}$$

where the first inequality is because $\mu_l > \frac{k}{1/\mu+\epsilon}$ for any $l\geq N+1$. Similarly, we can also prove $\sum_{i=1}^{\infty}b_i < \infty$, and $\sum_{i=2}^{\infty}c_i < \infty$ when $0\leq\lambda<k\mu$.
With the finiteness of $\sum_{i=1}^{\infty}a_i$, $\sum_{i=1}^{\infty}b_i$ and $\sum_{i=2}^{\infty}c_i$, $\frac{\partial\mathcal{W}(\lambda,\mu,k,R)}{\partial\lambda}$ is finite as well. We also note that $a_i(\mu), b_i(\mu)$ and $c_i(\mu)$ are all continuous and decreasing in μ. Therefore by the monotone convergence theorem,

$$\lim_{\mu\to\infty}\sum_{i=1}^{\infty}a_i(\mu) = \sum_{i=1}^{\infty}\lim_{\mu\to\infty}a_i(\mu),\ \lim_{\mu\to\infty}\sum_{i=1}^{\infty}b_i(\mu) = \sum_{i=1}^{\infty}\lim_{\mu\to\infty}b_i(\mu),\ \text{and}\ \lim_{\mu\to\infty}\sum_{i=2}^{\infty}c_i(\mu) = \sum_{i=2}^{\infty}\lim_{\mu\to\infty}c_i(\mu).$$

We thus have that for any given k, R and $0 \leq \lambda < k\mu$,

$$\lim_{\mu\to\infty}\frac{\partial\mathcal{W}}{\partial\lambda} = \frac{\lim_{\mu\to\infty}\sum_{i=2}^{\infty}\frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i}(1+\lim_{\mu\to\infty}\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}) - \lim_{\mu\to\infty}(\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i})^2}{\lim_{\mu\to\infty}(1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i})^2} \tag{1}$$

$$= \frac{\sum_{i=2}^{\infty}\frac{i(i-1)\lambda^{i-2}}{\lim_{\mu\to\infty}(\mu_1\mu_2\cdots\mu_i)}(1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\lim_{\mu\to\infty}(\mu_1\mu_2\cdots\mu_i)}) - (\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\lim_{\mu\to\infty}(\mu_1\mu_2\cdots\mu_i)})^2}{(1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\lim_{\mu\to\infty}(\mu_1\mu_2\cdots\mu_i)})^2} = \frac{\partial\overline{\mathcal{W}}}{\partial\lambda}, \tag{2}$$

where $\overline{\mathcal{W}}$ is defined in Lemma 1. Here the first equation holds since all limits exist and are finite, and the limit for the denominator is not zero.

By Lemma 1, for any given R and k, there exists $\lambda^*(k,R)$ such that $\frac{\partial\overline{\mathcal{W}}}{\partial\lambda}\Big|_{\lambda=\lambda^*(k,R)} < 0$. Thus, by (2), there exists μ^* such that when $\mu > \mu^*$, $\frac{\partial\mathcal{W}}{\partial\lambda}\big|_{\lambda=\lambda^*(k,R)} < 0$. Thus the lemma is proved. □

Proof of Theorem 1. In the following, we will prove the three parts of Theorem 1 separately. For the ease of notation, we omit the superscripts in μ_is, λ_is and $\mathcal{W}(\cdot)$ as the meanings are clear from the context.

Part 1. For the approximated one-direction call system, the expected waiting time $\mathcal{W}(\lambda,\mu,k,R)$ satisfies that $\mathcal{W} = \frac{\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}}{1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}} - 1/\mu$, where μ_is are defined in (**??**). When $\lambda \geq k\mu$, since $\mu_i \leq k\mu$ for all i, it is easy to see that $\mathcal{W}(\lambda,\mu,k,R) = \infty$. Therefore it remains to prove that when $\lambda < k\mu$, $\mathcal{W}(\lambda,\mu,k,R) < \infty$. It is equivalent to show $\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}$ and $\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}$ are both finite when $0 \leq \lambda < k\mu$. Note that in the proof of Lemma 2 we have proved the finiteness of these two terms. Therefore, $\mathcal{W}(\lambda,\mu,k,R) < \infty$ if and only if $0 \leq \lambda < k\mu$ and thus the first part is proved.

Part 2. We first show that $\frac{\partial\mathcal{W}}{\partial\lambda}\big|_{\lambda=0} \geq 0$. To see that, we take derivative of $\mathcal{W}$ with respect to λ, we have

$$\frac{\partial\mathcal{W}}{\partial\lambda} = \frac{\sum_{i=2}^{\infty}\frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i}(1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}) - (\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i})^2}{(1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i})^2}.$$

We have that

$$\frac{\partial\mathcal{W}}{\partial\lambda}\bigg|_{\lambda=0} = \frac{2}{\mu_1\mu_2} - \frac{1}{\mu_1^2} = \begin{cases} \frac{1}{\mu_1}(\frac{1}{\mu}+\frac{R}{2}), & \text{if } k=1, \\ \frac{R}{6\mu_1}, & \text{if } k=2, \\ \frac{1}{\mu_1}(\frac{R}{k}-\frac{R}{k+1}), & \text{if } k>2. \end{cases}$$

Therefore, $\frac{\partial\mathcal{W}}{\partial\lambda}\big|_{\lambda=0} \geq 0$.

Next, we prove that $\underline{\lim}_{\lambda\to k\mu-}\frac{\partial\mathcal{W}(\lambda,\mu,k,R)}{\partial\lambda} > 0$. It suffices to show that

$$\underline{\lim}_{\lambda\to k\mu-}\left\{\sum_{i=2}^{\infty}\frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i}\left(1+\sum_{i=1}^{\infty}\frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}\right) - \left(\sum_{i=1}^{\infty}\frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}\right)^2\right\} > 0. \tag{3}$$

Let $0 < \epsilon < 1$ be a constant such that $(1-\epsilon)^2(\mu R+2) - (1+\epsilon)^2(\mu R+1) > \frac{1}{2}$, and $M = \left\lceil \max\left\{\frac{\mu R-k+1}{\sqrt[\mu R]{\frac{1}{1-\epsilon}}-1}, \frac{k-2}{1-\sqrt[\mu R]{\frac{1}{1+\epsilon}}}\right\}\right\rceil + 1$. Here we use $\lceil x\rceil$ to denote the smallest integer that is greater than x. In the following, for the ease of notation, we assume μR is an integer. The case when μR is not an integer can be proved in a very similar way, however, the notation would be much messier. For $i \geq \mu R + M + 1$, by the definition of μ_n, we have

$$\frac{\lambda^i}{\mu_1 \cdots \mu_i} = \left(\frac{\lambda}{k\mu}\right)^i \frac{k^k}{k!} \prod_{\ell=1}^{k} \left(1 + \frac{\mu R}{k - \ell + 2}\right) \prod_{\ell=k+1}^{i} \left(1 + \frac{\mu R}{\ell - k + 1}\right) = C_1 \left(\frac{\lambda}{k\mu}\right)^i \prod_{\ell=M+1}^{i} \left(1 + \frac{\mu R}{\ell - k + 1}\right)$$

where $C_1 = \frac{k^k}{k!} \prod_{\ell=1}^{k} (1 + \frac{\mu R}{k-\ell+2}) \prod_{\ell=k+1}^{M} (1 + \frac{\mu R}{\ell-k+1})$ is a constant. We have

$$\begin{aligned}\prod_{\ell=M+1}^{i} \left(1 + \frac{\mu R}{\ell - k + 1}\right) &= \frac{(\mu R + M + 2 - k)(\mu R + M + 3 - k) \cdots (\mu R + i + 1 - k)}{(M + 2 - k)(M + 3 - k) \cdots (i + 1 - k)} \\ &= \frac{(i + 2 - k) \cdots (i + 1 - k + \mu R)}{(M + 2 - k) \cdots (M + 1 - k + \mu R)} \in \left(\left(\frac{i + 1 - k + \mu R}{M + 1 - k + \mu R}\right)^{\mu R}, \left(\frac{i + 2 - k}{M + 2 - k}\right)^{\mu R} \right)\end{aligned}$$

where in the second equality $i \geq \mu R + M + 1$ and $\mu R \in \mathbb{Z}_+$ are applied. By the definition of M, we have that

$$\prod_{\ell=M+1}^{i} \left(1 + \frac{\mu R}{\ell - k + 1}\right) \in \left\{ (1 - \epsilon) \left(\frac{i}{M}\right)^{\mu R}, (1 + \epsilon) \left(\frac{i}{M}\right)^{\mu R} \right\}.$$

Therefore, for $i \geq \mu R + M + 1$,

$$\frac{\lambda^i}{\mu_1 \cdots \mu_i} \in \left\{ C_1 (1 - \epsilon) \left(\frac{\lambda}{k\mu}\right)^i \left(\frac{i}{M}\right)^{\mu R}, C_1 (1 + \epsilon) \left(\frac{\lambda}{k\mu}\right)^i \left(\frac{i}{M}\right)^{\mu R} \right\}.$$

Similarly, we have

$$\frac{i\lambda^{i-1}}{\mu_1 \cdots \mu_i} \in \left\{ \frac{C_1 (1 - \epsilon) i}{k\mu} \left(\frac{\lambda}{k\mu}\right)^{i-1} \left(\frac{i}{M}\right)^{\mu R}, \frac{C_1 (1 + \epsilon) i}{k\mu} \left(\frac{\lambda}{k\mu}\right)^{i-1} \left(\frac{i}{M}\right)^{\mu R} \right\}$$

and

$$\frac{i(i-1)\lambda^{i-2}}{\mu_1 \cdots \mu_i} \in \left\{ \frac{C_1 (1 - \epsilon) i (i - 1)}{(k\mu)^2} \left(\frac{\lambda}{k\mu}\right)^{i-2} \left(\frac{i}{M}\right)^{\mu R}, \frac{C_1 (1 + \epsilon) i (i - 1)}{(k\mu)^2} \left(\frac{\lambda}{k\mu}\right)^{i-2} \left(\frac{i}{M}\right)^{\mu R} \right\}.$$

Therefore, we have

$$\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1 \cdots \mu_i} = \sum_{i=1}^{M} \frac{\lambda^i}{\mu_1 \cdots \mu_i} + \sum_{i=M+1}^{\infty} \frac{\lambda^i}{\mu_1 \cdots \mu_i} \geq C_2 + \frac{C_1 (1 - \epsilon)}{M^{\mu R}} \sum_{i=1}^{\infty} \left(\frac{\lambda}{k\mu}\right)^i i^{\mu R}$$

where $C_2 = -MC_1(1 - \epsilon) \leq \sum_{i=1}^{M} \frac{\lambda^i}{\mu_1 \cdots \mu_i} - \frac{C_1(1-\epsilon)}{M^{\mu R}} \sum_{i=1}^{M} \left(\frac{\lambda}{k\mu}\right)^i i^{\mu R}$ is a constant. Similarly,

$$\sum_{i=2}^{\infty} \frac{i(i-1)\lambda^{i-2}}{\mu_1 \cdots \mu_i} \geq C_2' + \frac{C_1 (1 - \epsilon)}{(k\mu)^2 M^{\mu R}} \sum_{i=2}^{\infty} \left(\frac{\lambda}{k\mu}\right)^{i-2} (i - 1) i^{\mu R + 1},$$

where $C_2' = -C_1 M^3 (1 - \epsilon)/(k\mu)^2 \leq \sum_{i=2}^{M} \frac{i(i-1)\lambda^{i-2}}{\mu_1 \cdots \mu_i} - \frac{C_1(1-\epsilon)}{(k\mu)^2 M^{\mu R}} \sum_{i=2}^{M} (\frac{\lambda}{k\mu})^{i-2} (i - 1) i^{\mu R + 1}$ and

$$\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1 \cdots \mu_i} \leq C_2'' + \frac{C_1 (1 - \epsilon)}{k\mu M^{\mu R}} \sum_{i=1}^{\infty} \left(\frac{\lambda}{k\mu}\right)^{i-1} i^{\mu R + 1},$$

where $C_2'' = M^2 k^M \geq \sum_{i=1}^{M} \frac{i\lambda^{i-1}}{\mu_1 \cdots \mu_i} - \frac{C_1(1-\epsilon)}{k\mu M^{\mu R}} \sum_{i=1}^{M} (\frac{\lambda}{k\mu})^{i-1} i^{\mu R+1}$.

In the following, we define

$$f_0(p) = \sum_{i=1}^{\infty} p^i i^{\mu R}, \quad f_1(p) = \sum_{i=1}^{\infty} p^{i-1} i^{\mu R+1} \quad \text{and} \quad f_2(p) = \sum_{i=2}^{\infty} p^{i-2}(i-1) i^{\mu R+1}.$$

Using these functions, the left-hand side of (3) can be written as:

$$\begin{aligned} &\textstyle\sum_{i=2}^{\infty} \frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i} \left(1 + \sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}\right) - \left(\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}\right)^2 \\ \geq &\left(C_2' + \tfrac{C_1(1-\epsilon)}{(k\mu)^2 M^{\mu R}} f_2\left(\tfrac{\lambda}{k\mu}\right)\right)\left(C_2 + \tfrac{C_1(1-\epsilon)}{M^{\mu R}} f_0\left(\tfrac{\lambda}{k\mu}\right)\right) - \left(C_2'' + \tfrac{C_1(1+\epsilon)}{k\mu M^{\mu R}} f_1\left(\tfrac{\lambda}{k\mu}\right)\right)^2. \end{aligned} \tag{4}$$

In the following, we use the notation of polylogarithm, in which $\mathrm{Li}_s(z) = \sum_{i=1}^{\infty} \frac{z^i}{i^s}$. Then we have

$$f_0(p) = \mathrm{Li}_{-\mu R}(p), \quad f_1(p) = \frac{1}{p}\mathrm{Li}_{-\mu R-1}(p), \quad \text{and} \quad f_2(p) = \frac{1}{p^2}(\mathrm{Li}_{-\mu R-2}(p) - \mathrm{Li}_{-\mu R-1}(p)).$$

By the limiting behavior of the polylogarithm (Wood 1992), we have that

$$\lim_{p\to 1^-} f_0(p) = \frac{\Gamma(1+\mu R)}{(-\log p)^{\mu R+1}}, \quad \lim_{p\to 1^-} f_1(p) = \frac{\Gamma(2+\mu R)}{(-\log p)^{\mu R+2}}, \quad \text{and} \lim_{p\to 1^-} f_2(p) = \frac{\Gamma(3+\mu R)}{(-\log p)^{\mu R+3}} - \frac{\Gamma(2+\mu R)}{(-\log p)^{\mu R+2}}$$

where $\Gamma(\cdot)$ is the Gamma function. Therefore,

$$\begin{aligned} &\lim\nolimits_{p\to 1^-} (1-\epsilon)^2 f_0(p) f_2(p) - (1+\epsilon)^2 f_1(p)^2 \\ &= ((1-\epsilon)^2(\mu R)!(\mu R+2)! - (1+\epsilon)^2(\mu R+1)!^2)(-\log p)^{-2\mu R-4} - (1-\epsilon)^2(\mu R)!(\mu R+1)!(-\log p)^{-2\mu R-3} \\ &= (\mu R)!(\mu R+1)![(1-\epsilon)^2(\mu R+2) - (1+\epsilon)^2(\mu R+1)](-\log p)^{-2\mu R-4} - (1-\epsilon)^2(\mu R)!(\mu R+1)!(-\log p)^{-2\mu R-3} \\ &\geq (\mu R)!(\mu R+1)!\left(\frac{1}{2}(-\log p)^{-2\mu R-4} - (1-\epsilon)^2(-\log p)^{-2\mu R-3}\right) \\ &= \mathcal{O}(-\log p)^{-2\mu R-4} > 0 \end{aligned}$$

where the inequality is because of the definition of ϵ. Based on these results, for equation (4), we further have that

$$\begin{aligned} &\underline{\lim}_{\lambda\to k\mu-} \left(\textstyle\sum_{i=2}^{\infty} \frac{i(i-1)\lambda^{i-2}}{\mu_1\mu_2\cdots\mu_i}\right)\left(1 + \sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}\right) - \left(\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}\right)^2 \\ \geq &\underline{\lim}_{\lambda\to k\mu-} \left(C_2'' + \tfrac{C_1(1-\epsilon)}{(k\mu)^2 M^{\mu R}} f_2\left(\tfrac{\lambda}{k\mu}\right)\right)\left(C_2 + \tfrac{C_1(1-\epsilon)}{M^{\mu R}} f_0\left(\tfrac{\lambda}{k\mu}\right)\right) - \left(C_2' + \tfrac{C_1(1+\epsilon)}{k\mu M^{\mu R}} f_1\left(\tfrac{\lambda}{k\mu}\right)\right)^2 = \mathcal{O}(-\log(\tfrac{\lambda}{k\mu}))^{-2\mu R-4}. \end{aligned} \tag{5}$$

The last step is because the other terms are smaller in order than $\mathcal{O}(-\log(\frac{\lambda}{k\mu}))^{-2\mu R-4}$. Therefore, $\underline{\lim}_{\lambda\to k\mu-} \frac{\partial \mathcal{W}(\lambda,\mu,k,R)}{\partial\lambda} > 0$. And part 2 is proved.

Part 3. By standard queueing theory, $\mathcal{W}(\lambda,\mu,k,R) = \frac{\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}}{1+\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}} - 1/\mu$. Now we consider a new approximated one-direction call system with arrival rate $R\lambda$, service rate $R\mu$ and road length 1. Note that the new system is equivalent to the original system after scaling the time in the original system by a factor of $1/R$ (one unit of time in the original system corresponds to $1/R$ units of time in the new system). Therefore the corresponding expected waiting time $\mathcal{W}(R\lambda, R\mu, k, 1)$ satisfies that $\mathcal{W}(R\lambda, R\mu, k, 1) = \frac{1}{R}\mathcal{W}(\lambda, \mu, k, R)$. Define $\tilde{\lambda} = R\lambda, \tilde{\mu} = R\mu, \tilde{k} = k, \tilde{R} = 1$, we have

$$\frac{\partial \mathcal{W}(\lambda, \mu, k, R)}{\partial \lambda} = R\frac{\partial \mathcal{W}(R\lambda, R\mu, k, 1)}{\partial \lambda} = R^2 \frac{\partial \mathcal{W}(\tilde{\lambda}, \tilde{\mu}, \tilde{k}, \tilde{R})}{\partial \tilde{\lambda}}.$$

To prove that for any given μ and $k \geq 2$, there exists a constant $R^*(\mu, k)$ such that when $R > R^*(\mu, k)$, there exists $0 \leq \lambda(\mu, k, R) < k\mu$ such that $\frac{\partial \mathcal{W}(\lambda,\mu,k,R)}{\partial \lambda}\Big|_{\lambda=\lambda(\mu,k,R)} < 0$, it is equivalent to show that for the new approximated system, for any given μ and k, there exists a constant $R^*(\mu, k)$ such that when $R > R^*(\mu, k)$, there exists $0 \leq \lambda(\mu, k, R) < k\mu$ such that $\frac{\partial \mathcal{W}(\tilde{\lambda},\tilde{\mu},\tilde{k},\tilde{R})}{\partial \tilde{\lambda}}\Big|_{\tilde{\lambda}=R\lambda(\mu,k,R)} < 0$.
By applying the result in Lemma 2, for the new approximated system, there exists an arrival rate $0 \leq \tilde{\lambda}^*(\tilde{k}, \tilde{R}) < \tilde{k}\tilde{\mu}$ and $\tilde{\mu}^*(\tilde{\lambda}^*, \tilde{k}, \tilde{R})$ such that $\frac{\partial \mathcal{W}(\tilde{\lambda},\tilde{\mu},\tilde{k},\tilde{R})}{\partial \tilde{\lambda}}\Big|_{\tilde{\lambda}=\tilde{\lambda}^*(\tilde{k},\tilde{R})} < 0$ for any $\tilde{\mu} > \mu^*(\tilde{\lambda}, \tilde{k}, \tilde{R})$. Since $\tilde{\mu} = R\mu$, $\tilde{k} = k$ and $\tilde{R} = 1$, equivalently, there exists an arrival rate $0 \leq \tilde{\lambda}^*(k, 1) < kR\mu$ such that $\frac{\partial \mathcal{W}(\tilde{\lambda},\tilde{\mu},\tilde{k},\tilde{R})}{\partial \tilde{\lambda}}\Big|_{\tilde{\lambda}=\tilde{\lambda}^*(k,1)} < 0$ when $R\mu > \tilde{\mu}^*(\tilde{\lambda}^*(k, 1), k, 1)$. Note that here $\tilde{\lambda}^*(k, 1)$ only depends on k. For given k and μ, when choosing $R > R^*(\mu, k) = \max\left\{\frac{\tilde{\mu}^*(\tilde{\lambda}^*(k,1),k,1)}{\mu}, \frac{\tilde{\lambda}^*(k,1)}{k\mu}\right\}$, inequalities $\tilde{\lambda}^*(k, 1) < kR\mu$ and $R\mu > \tilde{\mu}^*(\tilde{\lambda}^*, k, 1)$ are both satisfied. If we further choose $\lambda(\mu, k, R)$ by $\lambda(\mu, k, R) = \tilde{\lambda}^*(k, 1)/R$, combining with $R > R^*(\mu, k)$, we have $\frac{\partial \mathcal{W}(\tilde{\lambda},\tilde{\mu},\tilde{k},\tilde{R})}{\partial \tilde{\lambda}}\Big|_{\tilde{\lambda}=R\lambda(\mu,k,R)} < 0$. Thus the theorem is proved. □

Proof of Theorem 2. We start with the first part of the theorem. Note that

$$\mu_n^{(1n)} = \begin{cases} \frac{n}{1/\mu + \int_0^R \left(\frac{R-x}{R}\right)^{k-n+1} \exp\left(-\frac{(n-1)\mu x^2}{2R}\right) dx} = \frac{n}{1/\mu + \int_0^1 R(1-t)^{k-n+1} \exp\left(-\frac{(n-1)\mu R t^2}{2}\right) dt}, & n \leq k, \\ \frac{k}{1/\mu + \int_0^R \left(\frac{R-x}{R}\right)^{n-k} \exp\left(-\frac{\lambda x^2}{2R}\right) dx} = \frac{k}{1/\mu + \int_0^1 R(1-t)^{n-k} \exp\left(-\frac{\lambda R t^2}{2}\right) dt}, & n > k. \end{cases}$$

We have for any $R \geq 0$,

$$\min\{n, k\}\mu \geq \mu_n^{(1n)} \geq \begin{cases} \frac{n}{1/\mu + \int_0^1 R(1-t)^{k-n+1} dt}, & 0 \leq n \leq k, \\ \frac{k}{1/\mu + \int_0^1 R(1-t)^{n-k} dt}, & n > k. \end{cases}$$

Thus, $\lim_{R \to 0} \mu_n^{(1n)} = \min\{n, k\}\mu$. That is, when R approaches zero, the one-direction no-call approximation approaches a standard $M/M/k$ queue with arrival rate λ and service rate $\min\{n, k\}\mu$, for which the average waiting time increases with arrival rate λ.
For waiting time, we have $\mathcal{W}^{(1n)} = \frac{\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1^{(1n)}\mu_2^{(1n)}\cdots\mu_i^{(1n)}}}{1+\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1^{(1n)}\mu_2^{(1n)}\cdots\mu_i^{(1n)}}} - 1/\mu$. By the bounded convergence theorem, we have $\lim_{R \to 0} \frac{\partial \mathcal{W}^{(1n)}}{\partial \lambda} = \frac{\partial \lim_{R \to 0} \mathcal{W}^{(1n)}}{\partial \lambda} = \frac{\partial \mathcal{W}_{M/M/k}}{\partial \lambda} > 0$. Since $\mathcal{W}^{(1n)}$ is differentiable in λ, when R is small enough, we must have $\frac{\partial \mathcal{W}^{(1n)}}{\partial \lambda} > 0$.
Now we prove the second part of the theorem. It is easy to see that $\mu_n^{(1n)} > \mu_n^{(1c)}$ for all n (because $\mu_n^{(1c)}$ is just $\mu_n^{(1n)}$ without the exponential part in the denominator). That is, the approximated

one-direction no-call system has a higher service rate compared to the approximated one-direction call system for any system state. Now we show that this implies the average waiting time in the no-call system is shorter. To show that, we show that in a state-dependent $M/M/k$ queue, the average waiting time $\mathcal{W}(\lambda,\mu)$ is decreasing in μ_m for any m, i.e., $\frac{\partial \mathcal{W}}{\partial \mu_m} < 0$ for any m and μ_m. We have

$$\mathcal{W} = \frac{\sum_{i=1}^{\infty} \frac{i\lambda^{i-1}}{\mu_1\mu_2\cdots\mu_i}}{1+\sum_{i=1}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}} - 1/\mu = \frac{1}{\lambda}\frac{c_1+\sum_{i=m}^{\infty} \frac{i\lambda^i}{\mu_1\mu_2\cdots\mu_i}}{c_2+\sum_{i=m}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}} - 1/\mu,$$

where $c_1 = \sum_{i=1}^{m-1} \frac{i\lambda^i}{\mu_1\mu_2\cdots\mu_i}$ and $c_2 = 1+\sum_{i=1}^{m-1} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}$.

Define $f = \frac{c_1+\sum_{i=m}^{\infty} \frac{i\lambda^i}{\mu_1\mu_2\cdots\mu_i}}{c_2+\sum_{i=m}^{\infty} \frac{\lambda^i}{\mu_1\mu_2\cdots\mu_i}} = \frac{c_1+\sum_{i=m}^{\infty} \frac{a_i}{\mu_m}}{c_2+\sum_{i=m}^{\infty} \frac{b_i}{\mu_m}}$, where $a_i = \frac{i\lambda^i}{\mu_1\mu_2\cdots\mu_{m-1}\mu_{m+1}\cdots\mu_i}$ and $b_i = \frac{a_i}{i}$. To prove $\frac{\partial \mathcal{W}}{\partial \mu_m} < 0$, it is equivalent to prove that $\frac{\partial f}{\partial \mu_m} < 0$. We have that

$$\frac{\partial f}{\partial \mu_m} = \frac{\sum_{i=m}^{\infty} \frac{b_i}{\mu_m^2}(c_1+\sum_{i=m}^{\infty} \frac{a_i}{\mu_m}) - (\sum_{i=m}^{\infty} \frac{a_i}{\mu_m^2})(c_2+\sum_{i=m}^{\infty} \frac{b_i}{\mu_m})}{(c_2+\sum_{i=m}^{\infty} \frac{b_i}{\mu_m})^2} = \frac{c_1\sum_{i=m}^{\infty} b_i - c_2\sum_{i=m}^{\infty} a_i}{\mu_m^2(c_2+\sum_{i=m}^{\infty} \frac{b_i}{\mu_m})^2}.$$

Note that $c_1 < mc_2$ and $\sum_{i=m}^{\infty} a_i > m\sum_{i=m}^{\infty} b_i$, we thus get $c_1\sum_{i=m}^{\infty} b_i - c_2\sum_{i=m}^{\infty} a_i < 0$. As a result, $\frac{\partial f}{\partial \mu_m} < 0$ and $\frac{\partial \mathcal{W}}{\partial \mu_m} < 0$. Therefore, the average waiting time in the approximated one-direction no-call system is always smaller than that in the approximated one-direction call system, that is, $\mathcal{W}^{(1n)} < \mathcal{W}^{(1c)}$. □

Proof of Theorem 4. By redefining $R' = R/2$ and comparing $\mu_n^{(1c)}$ and $\mu_n^{(2c)}$, we note that the approximated two-direction call system is identical with the approximated one-direction call system with $R' = R/2$. That is, $\mathcal{W}^{(2c)}(\lambda,\mu,k,R) = \mathcal{W}^{(1c)}(\lambda,\mu,k,R')$. Thus the theorem follows directly from the proof of Theorem 1. □

References

Wood, D. 1992. *The Computation of Polylogarithms*. University of Kent at Canterbury.