Efficiency or Fairness? Carpooling Design for Online Ride-hailing Platform in Transport Hubs at Midnight

Chenbei Lu, Jiaman Wu
Institute for Interdisciplinary Information Sciences
Tsinghua University
Beijing, China
{lcb20,wjm18}@mails.tsinghua.edu.cn

Chenye Wu*
School of Science and Engineering,
The Chinese University of Hong Kong, Shenzhen
Shenzhen, Guangdong, China
chenyewu@yeah.net

Yongli Qin, Qun Li
Nan Ma
Didi Chuxing
Beijing, China
{dsqinyongli,liquntracy}@didiglobal.com
mandymanan@didiglobal.com

ABSTRACT
The online ride-hailing platform has revolutionized urban transport. However, there is much room for improvement. We consider meeting the demand for an online ride-hailing at transport hubs late at night, when the public transport system stops its operations. Passengers arriving late at night face a long wait before service. We launch the ride-hailing model in the theoretical framework of queueing and introduce the arrival and the service processes. To improve the efficiency of the ride-hailing platform, as well as to maintain fairness between different types of passengers, we study three variations of carpool service policies. We then provide practical guidelines on the trade-off between efficiency and fairness to assist the online platform designers. Specifically, we derive the analytical trade-off bounds with the passenger parameters. Furthermore, we suggest that these bounds can be good performance estimators for the empirical trade-off when only limited passenger information is available. This analysis motivates us to design the optimal service rate for the entire platform. Finally, we conduct numerical studies based on field data retrieved from Didi Chuxing, highlighting the remarkable performance of our proposed method in terms of improving the quality of online ride-hailing service.

CCS CONCEPTS
• Applied computing → Transportation; Decision analysis.

KEYWORDS
Carpool, Queueing Theory, Performance Evaluation, Optimization

1 INTRODUCTION
Sharing economy has profoundly altered our life styles in many ways. As a result, ride-hailing and ride-sharing services are reshaping the urban transport by providing convenient and personalized ways for passengers to commute and travel. This related market grew 21.7% from 2016 to 2019 on an annual basis in China [19].

The online transport service is a valuable addition to the current public transport system, which often requires upfront planning, huge capital investment, and limited flexibility during operation. The combined online and public transport system generally works efficiently. However, serious issues can arise when one of the forms is missing. According to the Beijing Transport Annual Report, public transport supplies 89.60% of traffic demands [16]. As a result, passengers arriving at transport hubs late at night may face only a few transport options, such as the online ride-hailing. This leads to extremely long waiting time before travelling. According to the Didi platform, passengers who arrived at the Beijing South railway station within the first half-hour after the subway shut down had to wait for about 20 minutes. In addition, when large numbers of vehicles flood the transport hub, it often becomes very congested, increasing passengers’ waiting times for the service.

Specifically, this work considers a theoretical integrated ride-hailing design for busy transport hubs without public transport. To relieve the peak demand, one realistic means is to conduct carpooling. We analyze the trade-off between its efficiency (i.e., the mean waiting time) and fairness (i.e., the differential effects for passengers with diversified carpool preferences). In particular, we seek to understand the fundamental limit of each service policy in terms of negotiating the efficiency and fairness.

1.1 Related Works
The boom in ride-hailing online platforms has attracted significant recent academic attention, much of which focuses on improving platform efficiency. Empirical studies mainly use learning approaches. Methods varying from reinforcement learning [21, 27] to multi-graph convolution network [10] are adopted to enable the effective operation of ride-hailing platform. Although sound in practice, they provide limited insight to the structure of the service process. In contrast to this line of research, we provide theoretical guidelines to improve the efficiency of the system.

From a theoretical perspective, much effort is devoted to modeling and understanding the structure of the ride-hailing process,
which is crucial to improve the platform efficiency. In this context, Sadowsky et al. use the regression discontinuity design to examine the impact of Uber and Lyft’s entry into public transport in [20]. Afeche et al. propose an integrated model to characterize the system equilibrium for ride-hailing in [1]. Henao et al. further analyze the impacts of the ride-hailing on the total vehicle miles traveled [13] and vehicle parking [12]. We further the literature by considering the carpooling process, and characterize its influence to the whole ride-hailing process.

Carpooling is a key way to improve the efficiency of ride-hailing in practice. Teal et al. provide the initial insights into the different types of carpool passengers and describe how and why passengers share rides in [22]. Much has followed this seminal work, including designing a real-time carpool system [23], considering the recommendation system for carpool services [28], assessing the potential benefit of carpooling in terms of meeting more requests, and reducing city’s traffic flows [25]. To further improve the platform efficiency, solution concepts range from optimization [6, 9, 14, 17, 18, 24, 26] to learning [2]. In contrast to most literature, which focuses on improving system efficiency through empirical carpool service policy design, we seek to provide theoretical insights on carpooling policy design and analyze the crucial trade-off between efficiency and fairness.

1.2 Our Contributions
In seek of improving system efficiency through empirical policy design, our main contribution is threefold:

- We conduct theoretical analysis on the carpool service policy design. Specifically, we analyze the impact of carpool service policies on the mean passenger waiting time and the length of the busy period.
- We introduce empirical metrics to enable future carpool service policy design. More precisely, we study the theoretical tension between fairness and efficiency for different carpool service policies. We offer analytical trade-off bounds between efficiency and fairness, which can serve as empirical performance estimations, when limited passenger information is available.
- We investigate the carpool service policy design customized for transport hubs. A special challenge for the transport hub analysis is intra-hub congestion, which is not a major focus in general carpooling models. We are motivated by existing models, such as the dynamic model [5], the non-equilibrium traffic model [29], and the integrated traffic congestion model [4]. We enhance these models by incorporating the interaction between congestion and ride-hailing. Furthermore, most existing congestion models are intractable or require strong assumptions to derive the analytical solution. Our proposed model can analyze the expected performance of the carpool in a very general setting.

The rest of the paper is organized as follows: Section 2 introduces the ride-hailing model without public transport at late night transport hubs. Section 3 evaluates two key performance metrics for the classical settings. Section 4 investigates the tension between efficiency and fairness for different carpool service policies. Section 5 models the congestion process and analyzes its theoretical properties. Section 6 numerically verifies the usefulness of the analytical trade-off bounds using field ride-hailing data, and demonstrates the designed service rate is able to optimize the quality of ride-hailing service. The concluding remarks are delivered in Section 7. Due to page limit, we defer all the necessary proof sketches in the Appendix.

2 RIDE-HAILING MODEL
In this section, we introduce the ride-hailing model for transport hubs without public transport, which consists of the arrival process and the service process. Then, we present a carefully designed multi-stage approximation of the arrival process to facilitate the subsequent theoretical analysis of the service process.

2.1 Arrival Process
Late at night, public transport near the transport hubs often stops working, making ride-hailing the only viable option for passengers. This situation sets up the arrival process: in the first stage, arrival is very intense but remains for a short period (although it is late at night, trains and flights are not scheduled to arrive too late at the hub); in the second stage, the arrival is very stable and its rate is kept at a very low level. These features are highlighted by field data at the Beijing South Railway Station (the orange curve in Figure 1).

This observation motivates us to approximate the two stages of the arrival process with a short peak period followed by a longer valley period. Each period has a stable demand, denoted by $\tilde{d}$, $d$ respectively. We assume that the entire arrival process is divided into $N_e$ segments. The first time interval (i.e., $t = 1$) starts when the public transport stops service, for example, 11:00pm in our configuration. The optimal approximation can be decided by solving the following optimization problem:

$$\min \sum_{i=1}^{k} (\tilde{d} - d_i^{true})^2 + \sum_{i=k+1}^{N_e} (d - d_i^{true})^2$$

s.t. \quad 1 \leq k < N_e,

$$i \in \mathbb{Z},$$

where decision variables are the division time $k$, peak demand $\tilde{d}$ and valley demand $d_i$. The parameter $d_i^{true}$ denotes the actual demand during period $i$. Figure 1 exemplifies the optimal approximation (the bar graphs), corresponding to the orange curve.
2.2 Service Process

The typical service process consists of three stages: in the first stage, passengers virtually queue for ride-hailing and wait to be matched with a driver. In the second stage, the matched driver picks up the passenger. Finally, the driver takes the passenger to the destination. In our work, we seek to improve the quality of passenger service at transport hub. Consequently, we focus only on the first two stages. This is particularly true in our setting, as after leaving the transport hub, traffic conditions are often free of congestion.

Then, we model and analyze the first two stages, respectively. Based on these analysis, we use various carpool service policies to improve the passenger ride-hailing experience.

3 QUEUEING FOR SERVICES: THE CLASSICS

This section studies the queueing process for passengers after arrival. Specifically, we evaluate two key performance metrics for the queueing process: the length of busy period and passenger’s mean waiting time. We use two periods to approximate the arrival process. In the subsequent analysis, we focus only on the second period, i.e., the long valley period. This is due to the small impact of the short peak in the calculation of the two performance metrics. After the intense but short peak period, we consider the following scenario: there are still A passengers in the queue waiting for service; and new passengers keep joining the queue following the Poisson Process with the arrival rate of A. We assume that the intervals between the arrivals of two neighboring vehicles, denoted by random variable S, follow an identical predefined distribution f_S(s). Based on such a configuration, we can theoretically characterize the two performance metrics of interest.

3.1 Busy Period Length

Busy period is an important metric in queueing theory. It describes the time it takes for the queue to empty for the first time since its current state. A busy period can well characterize the impact of the short but intense peak on the service process.

Denote the length of busy period with A initial passengers by B(A). Then, it holds that

\[ B(A) = T_A + B(N_A), \]  

where \( T_A \) denotes the time to serve the initial A passengers; \( N_A \) denotes the number of arrivals during \( T_A \); and \( B(N_A) \) denotes the busy period caused by \( N_A \) passengers. Eq. (2) illustrates the recursive relationship between busy periods.

We can verify that the length of the busy period \( B(A) \) is additive in A. In addition, standard Laplace transformation and moment analysis techniques provide the following fact.

**Fact 3.1.** The mean and variance of \( B(A) \) are as follows:

\[ \mathbb{E}[B(A)] = \frac{A\mathbb{E}[S]}{1 - \rho}, \]  
\[ \text{Var}[B(A)] = A\mathbb{E}[S] - (1 - \rho)\mathbb{E}[S], \]  

where \( \rho \) denotes the post-peak load rate:

\[ \rho = \frac{\lambda\mathbb{E}[S]}{\lambda\mathbb{E}[S]}, \]  

**Remark:** This result can be found in classic textbooks, e.g., Chapter 27 in [11]. It indicates that the busy period length increases linearly in A. The greater the post-peak load rate \( \rho \), the longer the busy period length. This fact helps us to examine the mean waiting time for passengers being served during the busy period.

3.2 Mean Waiting Time

After deriving the busy period length, it is possible to study the waiting time, denoted by \( T_Q(A) \), for passengers being served during the busy period under a given carpool service policy. Specifically, we adopt the most common First-Come-First-Serve (FCFS) policy and derive both the upper and lower bounds for the mean waiting time.

**Theorem 3.2.** The mean waiting time \( \mathbb{E}[T_Q(A)] \) is upper bounded by \( T_u \), and lower bounded by \( T_l \), where

\[ T_u = \frac{A + (1 - \rho)^2}{2(1 - \rho)^2} \mathbb{E}[S], \]  
\[ T_l = \frac{(1 - \rho)(A + 1)\mathbb{E}[S]}{2}. \]  

**Remark:** This theorem coincides with our intuition: the mean waiting time is linear in A. When \( \rho = 0 \), it represents the case where there are no new passengers arriving; and the bounds are tight. In practice, the post-peak load rate \( \rho \) is usually very small (less than 0.1), which makes the bound practically tight. We use these practical insights to analyze the subsequent mean waiting time for various carpool service policies.

4 THE POWER OF CARPOOLING

While FCFS is a widely adopted policy, much can be improved. This section goes beyond FCFS and investigates the effects of various carpool service policies on the ride-hailing platform efficiency.

Specifically, we consider an offline approximation. We assume that, when matching the carpooling, the platform knows all the necessary information. While this seems like a strong assumption, advanced machine learning techniques are able to predict passenger demand very well. Furthermore, due to the long waiting time of most passengers, the platform can often successfully match passengers before departing the transport hub. This transforms the stochastic queueing model into tractable static analysis.

4.1 Carpool Service Policies

Carpooling, while sacrificing some comfort, often allows for quick service at low cost. Suppose there are \( N \) arrivals during the busy period of interest, we divide the passengers into two groups. Group \( A \) is composed of those who successfully participate in the carpooling. Denote the total number of passengers in Group A by \( B(N) \). All other passengers belong to Group B. They may not want to join in carpooling or simply cannot find the carpooling partners. Clearly, the total number of passengers in Group B is \( (1 - \beta)N \). The service rate for matching is again \( \mu = (\mathbb{E}[S])^{-1} \), which is totally determined by the vehicles’ arrival process. We want to emphasize that, throughout the paper, we only consider the carpooling between passengers departing from the transport hub. Specifically, we compare three variants of carpooling, which are inspired by the real dispatch strategies of Didi platform:
• FCFS: We can modify the FCFS policy for carpooling as the benchmark. During dispatch, we always respond to the passenger request at the front of the queue. If this passenger want to board the carpool and there are suitable partners in the line, the first one (ones) will attend together with the head passenger. Thus, all served passengers will be labeled as Group A. Otherwise, the head passenger will be directly served and be labeled as Group B.
• Carpooling First (CF): the CF policy prioritizes meeting the demands of passengers who successfully carpool. This policy can be described recursively. Consider matching a carpool request for the head passenger without any label. If he wants to join in a carpool and there are suitable partners in the line, the first one (ones) will attend and be served together with the head passenger, they will be labeled as Group A. Otherwise, the head passenger will be labeled as Group B. Only till there are all Group B passengers in the line, they will be served FCFS.

• Average Arrival (AA): AA is a policy based on the average arrival time of passengers. The labeling procedure is identical to the CF policy. The only difference is that AA will serve passengers in the order of the average arrival time (that is, the average arrival time for Group A, and the algorithmic average arrival time for matched passengers in Group A).

### 4.2 2-passenger Carpooling Analysis

We then compare the performance of the three carpooling variants to the 2-passenger carpooling and then extend the conclusion to more general carpooling configurations. Denote \( \mathcal{P} = \{CF, FCFS, AA\} \) the policy set. For Group \( i, i \in \{A, B, aaq (i.e., A \cup B)\} \), define \( T_i^p \) the waiting time under policy \( p \in \mathcal{P} \).

**Theorem 4.1.** The mean waiting time under the three policies can be analytically derived as follows:

\[
\begin{align*}
\mathbb{E}[T_{FA}^{FCFS}] &= \frac{6 + 4N - \beta N}{12\mu}, \\
\mathbb{E}[T_{FB}^{FCFS}] &= \frac{3 + 3N - \beta N}{6\mu}, \\
\mathbb{E}[T_{FA}^{FCFS}] &= \frac{6 + 6N - 4\beta N + \beta^2 N}{12\mu}, \\
\mathbb{E}[T_{FB}^{FCFS}] &= \frac{2 + \beta N}{4\mu}, \\
\mathbb{E}[T_{FA}^{AA}] &= \mathbb{E}[T_{FA}^{CF}] = \mathbb{E}[T_{FA}^{AAC}] = \frac{2 + 2N - \beta N}{4\mu}.
\end{align*}
\]

**Remark:** The most interesting observation from Theorem 4.1 is the understanding of how the proportion of carpool passengers \( \beta \) affects the waiting time in different policies. Specifically, we have:

\[
\begin{align*}
\mathbb{E}[T_{FA}^{CF}] &\leq \mathbb{E}[T_{FA}^{AAC}] \leq \mathbb{E}[T_{FA}^{AA}], \\
\mathbb{E}[T_{FA}^{AAC}] &\leq \mathbb{E}[T_{FA}^{FCFS}] \leq \mathbb{E}[T_{FA}^{CF}], \\
\mathbb{E}[T_{FA}^{AAC}] &\leq \mathbb{E}[T_{FA}^{CF}] \leq \mathbb{E}[T_{FA}^{CF}], \quad (8) \\
\mathbb{E}[T_{FA}^{AAC}] &\leq \mathbb{E}[T_{FA}^{FCFS}] \leq \mathbb{E}[T_{FA}^{AAC}], \quad (9) \\
\mathbb{E}[T_{FA}^{AAC}] &\leq \mathbb{E}[T_{FA}^{CF}] \leq \mathbb{E}[T_{FA}^{AAC}], \quad (10)
\end{align*}
\]

This partial order relationship indicates that the CF policy achieves the optimal mean waiting time for Group A and all passengers, because carpooling can carry more passengers each time compared to no-carpooling. However, this policy brings unfairness: Group B will have to endure the longest waiting time. The AA policy focuses more on the fairness between different groups: the two groups share the same mean waiting time. This policy also achieves the optimal mean waiting time for Group B. In contrast, the overall mean waiting time is the longest under AA policy. The FCFS policy strikes the best balance between fairness and efficiency, i.e., it achieves the second shortest waiting time for both groups. This may explain why FCFS policy is widely used.

#### 4.3 Metrics Design: Efficiency versus Fairness

We propose a weighted metric to trade-off efficiency and fairness:

\[
M(\beta, k) = M_{ef} + k M_{fair}.
\]

where \( k \) is the trade-off coefficient; \( M_{ef} \) and \( M_{fair} \) denote the efficiency and fairness metric, respectively. We adopt the mean waiting time to assess the efficiency and mean absolute bias of the waiting time among different carpooling groups among passengers to assess fairness, i.e.,

\[
M_{ef} = \mathbb{E}[T_{avg}], \quad (12)
\]

\[
M_{fair} = \frac{1}{N} \mathbb{E} \left[ \sum_{i \in \{A,B\}} |T_i^p - \mathbb{E}[T_{avg}]| \right]. \quad (13)
\]

Note that the set \( \{A,B\} \) can be easily extended for general carpooling analysis by replacing it with \( \{1,...,M\} \), which is defined in the next subsection. It is important to note that, the fairness metric \( M_{fair} \) actually measures unfairness. Consequently, the larger the \( M_{fair} \), the more unfair the policy \( p \).

The optimal policy \( p(\beta, k) \) with respect to carpooling parameter \( \beta \) and trade-off coefficient \( k \) can be obtained as follows:

\[
p(\beta, k) = \arg \min_{p \in \mathcal{P}} M(\beta, k). \quad (14)
\]

Specifically,

\[
p(\beta, k) = \begin{cases} 
  CF, & k < 0.5(2 - \beta)^{-1}, \\
  any policy, & k = 0.5(2 - \beta)^{-1}, \\
  AA, & k > 0.5(2 - \beta)^{-1}.
\end{cases} \quad (15)
\]

**Remark:** This suggests that with knowledge of \( \beta \) and \( k, CF \) or AA might be better alternative than FCFS. For the ride-hailing platform, choosing CF may attract more passengers with greater preference for efficiency, while choosing AA may attract those with better preference for fairness. This is not a universal conclusion. In practice, different passengers have heterogeneous preferences regarding efficiency and fairness, and therefore \( k \) may vary greatly. Nonetheless, under any trade-off coefficient \( k, FCFS \) is at least the second best policy and can be a safe choice.

#### 4.4 General Carpooling Analysis

We can now extend our 2-passenger carpooling analysis to general carpool service. Denote the percentages of passengers who are interested in joining in 1-passenger carpooling (i.e., no carpooling), 2-passenger carpooling, \( ... , M \)-passenger carpooling by \( \beta_1, \beta_2, ... , \beta_M \). Clearly, it is required that \( \sum_{i=1}^{M} \beta_i = 1 \).

We then define the carpooling rate vector as follows:

\[
\beta = [\beta_2, \beta_3, ..., \beta_M]^T. \quad (16)
\]
Note that, we use $\beta_1$ as the slack variable to ensure the total rate is 1. Consequently, we do not include $\beta_1$ in $\beta$. In general carpooling settings, the carpool service policies in Section 4.1 only need to be changed slightly. Specifically, passengers will be grouped into more carpooling groups (i.e., 1, 2, ..., $M$-passenger carpooling groups) following the same rule in Section 4.1. Carpooling groups with more passengers have higher priority in $CF$.

We also assume that the locations of passengers in the queue are independent of their types. For $k$-passenger carpooling group, $k = 1, ..., M$, define $\tau_k^p$ the waiting time under policy $p \in \mathcal{P}$. Define $\tau_{avg}$ the overall waiting time under policy $p \in \mathcal{P}$. We can then derive the following result for general carpooling settings:

**Theorem 4.2.** The mean waiting time under three policies with general carpooling types can be analytically derived as follows:

\[
\mathbb{E}[\tau_{FCFS}^k] = \frac{1}{2}\mu + \left(\frac{1}{1 + k} - \frac{a^k}{\mu} \right) N, \quad (17)
\]

\[
\mathbb{E}[\tau_{FCS}^k] = \frac{1}{2}\mu + \left(\frac{1}{2} - a^k \right) N, \quad (18)
\]

\[
\mathbb{E}[\tau_{CF}^k] = \frac{1}{2}\mu + b^k \beta N, \quad (19)
\]

\[
\mathbb{E}[\tau_{AA}^k] = \mathbb{E}[\tau_{FCS}^k] = \frac{1}{2}\mu + \left(\frac{1}{2} - b^k \right) N, \quad (20)
\]

where

- $a^k = [a^k_i]_{i=1}^{M-1}, a^k_1 = \frac{1}{1+k} - \frac{1}{\mu},$
- $a = [a_i]_{i=1}^{M-1}, a_i = \frac{1}{1+k};$
- $A = [A_{ij}]_{i=1}^{M-1} \times \{M-1\}, A_{ij} = 1 + \frac{2}{\mu_{ijk}} = \frac{4}{\mu_{ij}};$
- $b^k = [b^k_i]_{i=1}^{M-1}, b^k_1 = \frac{1}{\mu_{ijk}},$
- $b_i = [b_{ij}]_{i=1}^{M-1} \times \{M-1\}, b_{ij} = \min(i,j) \max(i,j) + 1.$

Specifically, the following order inequalities hold:

\[
\mathbb{E}[\tau_{CF}^k] \leq \mathbb{E}[\tau_{FCFS}^k] \leq \mathbb{E}[\tau_{AA}^k], \quad (22)
\]

\[
\mathbb{E}[\tau_{AA}^k] \leq \mathbb{E}[\tau_{FCS}^k] \leq \mathbb{E}[\tau_{CF}^k], \quad (23)
\]

\[
\mathbb{E}[\tau_{CF}^k] \leq \mathbb{E}[\tau_{FCS}^k] \leq \mathbb{E}[\tau_{AA}^k]. \quad (24)
\]

We can then examine the marginal benefits of having each type of passenger as follows:

\[
\frac{\partial \mathbb{E}[\tau_{AA}^k]}{\partial b} = -c N \mu, \quad (25)
\]

\[
\frac{\partial \mathbb{E}[\tau_{FCS}^k]}{\partial b} = (-b + B) N \mu, \quad (26)
\]

\[
\frac{\partial \mathbb{E}[\tau_{AA}^k]}{\partial \beta} = (-a + A) N \mu \quad (27)
\]

**Remark:** These results have several important indications. First, it indicates that passengers with greater tolerance for carpooling will bring more marginal benefits in terms of reduced mean waiting time. A larger $k$ will bring at most twice the marginal benefit of the 2-passenger carpooling for both $AA$ and $CF$, and will bring at most three times the marginal benefit for $FCFS$. When the carpooling willingness is low, $CF$ enjoys the greatest marginal benefit from increasing the carpooling rate, while $AA$ receives the least. Furthermore, with the increase in the carpooling rate, the marginal benefits of $CF$ and $AA$ decrease. In general, the higher the carpooling rate, the better the performance of $CF$ and $FCFS$ policy.

**Theorem 4.3.** For service policies $CF$ and $FCFS$ with total passengers $N$ and maximal carpooling rate $k$, the maximal ratios between $M_{f_{air}}$ and $M_{f_{CF}}$, denoted by $R_{CF}$ and $R_{FCFS}$, satisfy the following relationship:

\[
R_{CF} = \sup_{0 \leq x \leq 1} \frac{2N(K(1-x) + x) - x}{K + N(2x - x^2 + K(1-x)^2)}, \quad (28)
\]

\[
R_{FCFS} = \sup_{0 \leq x \leq 1} \frac{2N(K(1-x) + x) - x}{K + N(2x - x^2 + K(1-x)^2)}, \quad (29)
\]

\[
R_{CF} \geq R_{FCFS}. \quad (30)
\]

**Remark:** This theorem shows the fundamental bounds between $M_{f_{air}}$ and $M_{f_{CF}}$ for $CF$ and $FCFS$. The closed form expressions for $R_{CF}$ and $R_{FCFS}$ can be derived using Mathematica [15]. However, exact expressions are tedious and offer no insights, and thus we choose not to include them in the paper. We will demonstrate that these bounds are also close to the true trade-off performance for the policies in practice using field data.

## 5 PICKING UP THE PASSENGERS

Intra-hub congestion is another major component that contributes to the long waiting time in ride-hailing. We analyze the dynamics of traffic congestion in the second stage of the service process.

Denote $A(t)$ and $D(t)$ the total numbers of vehicles arriving and departing up to time $t$ respectively. By definition, $A(0) = 0$. Denote the duration of the delay by $T(t)$. Thus $A(t) = 0$. $T(t) = T(t+T(t))$, which shows that arrivals will depart with a delay of $T(t)$.

We use $N(t)$ to denote the number of vehicles in the transport hub at time $t$, i.e., $N(t) = A(t) - D(t)$. This model is the "proper model" proposed by [3], which is intractable. We make the following simplification for theoretical insights. We assume that the delay would be linearly influenced by the number of current vehicles:

\[
T(t) = M + aN(t), \quad (31)
\]

where $M$ is a random variable denoting congestion-free waiting time; $aN(t)$ denotes the additional waiting time due to congestion, and $a$ is a coefficient that represents the influence of the total number of vehicles during congestion. This parameter can be related to the properties of the transport hub, i.e., the number of lanes.

**Remark:** This simplified model is able to provide practical insights. First, distances from vehicles to transport hubs can vary greatly, which makes the congestion-free waiting time stochastic. Second, when congestion is relatively severe, the vehicles will leave one by one, and the extra waiting time is linearly related to the number of vehicles ahead of the vehicle of interest.

### 5.1 Prototype Linear System

To obtain the first cutting intuition, we consider a simplified prototype system. We make two strong assumptions for simplification:
Assumption 5.1. Vehicles arrive at a constant rate \( \mu: A(t) = \mu t \).

Assumption 5.2. \( M \) is deterministic.

Thus, the simplified system describes the mean dynamics of the traffic condition, in line with the notion of mean field analysis [7]. This system can be characterized as follows:

\[
D((1 + \alpha \mu) t + M - \alpha D(t)) = \mu, \quad (32) \\
D(t) = 0, \quad \forall t \in [0, M]. \quad (33)
\]

By taking the derivative of both sides of Eq. (32), we can obtain the following differential equation:

\[
D'((1 + \alpha \mu) t + M - \alpha D(t)) \cdot (1 + \alpha \mu - \alpha D'(t)) = \mu. \quad (34)
\]

Theorem 5.3. The differential equation in Eq. (34) has a unique solution \( D(t) \), which can be characterized as follows:

- If \( \alpha \mu = 1 \):
  \[
  D(t) = \frac{n(n - 1)M}{2} + \frac{\mu n + 1}{n + 1} \left( t - \frac{n(n + 1)M}{2} \right),
  \]
  \[\text{where } t \in \left[ \frac{n(n + 2)M}{2}, \frac{(n + 2)(n + 1)M}{2} \right], n \geq 0.\]
- If \( \alpha \mu \neq 1 \):
  \[
  D(t) = \mu n - \frac{a_{n-1} - a_n}{a_{n+1} - a_n} (t - a_n),
  \]
  \[\text{where } t \in [a_n, a_{n+1}], n \geq 0, \text{ and } a_n = \frac{[(\alpha \mu)n^2 - (\alpha \mu)(n + 1) + n]M}{(\alpha \mu - 1)^2}. \quad (37)
  \]

Remark: It indicates that \( D(t) \) is piecewise linear. We can further characterize its key properties.

Theorem 5.4. The asymptotic increasing rates of \( T(t) \), \( N(t) \) are as follows:

- If \( \alpha \mu < 1 \), \( T(t) \) and \( N(t) \) are bounded and converge linearly:
  \[
  \lim_{t \to \infty} T(t) = \frac{M}{1 - \alpha \mu}, \quad \lim_{t \to \infty} N(t) = \frac{\mu M}{1 - \alpha \mu}.
  \]
- If \( \alpha \mu = 1 \), \( T(t) \) and \( N(t) \) are asymptotically increasing at the rate of \( O(\sqrt{t}) \):
  \[
  \lim_{t \to \infty} \frac{T(t)}{\sqrt{t}} = \frac{M + \alpha N(t)}{\sqrt{t}} = \frac{\sqrt{2M}}{\alpha},
  \]
  \[
  \lim_{t \to \infty} \frac{N(t)}{\sqrt{t}} = \frac{\sqrt{2M}}{\alpha}, \quad \lim_{t \to \infty} \frac{N(t)}{A(t)} = 1.
  \]
- If \( \alpha \mu > 1 \), \( T(t) \) and \( N(t) \) are asymptotically increasing at the rate of \( O(t) \):
  \[
  \lim_{t \to \infty} \frac{T(t)}{t} = \frac{M + \alpha N(t)}{t} = \alpha \mu - 1,
  \]
  \[
  \lim_{t \to \infty} \frac{N(t)}{t} = \frac{\alpha \mu - 1}{t}, \quad \lim_{t \to \infty} \frac{N(t)}{A(t)} = \frac{\alpha \mu - 1}{\alpha \mu}.
  \]

Remark: This theorem is very intuitive and interesting: \( \alpha \mu \) plays the vital role. When the arrival rate \( \mu \) is lower than the threshold \( \alpha^{-1} \), congestion is fully controlled: as time goes on, the delay and the number of congested vehicles will both be bounded. They are also proportional to the congestion-free delay \( M \).

When \( \mu > \alpha^{-1} \), the arrival rate is greater than the service capacity of the transport hub. As a result, the hub become very congested.

Another interesting observation is that, the congestion process can be compared to the standard queueing process, where the parameter \( \alpha^{-1} \) plays the same role as the service rate.

5.2 Relaxation: Stochastic Arrival

We first relax Assumption 5.1 allowing stochastic arrival. Denote the arrival rate at time \( t \) by \( \mu(t) \). Thus, the total arrival to \( t \), denoted by \( A(t) \), has an integral form:

\[
A(t) = \int_0^t \mu(x)dx. \quad (38)
\]

The dynamics of the system can be described as follows:

\[
D(t + M + \alpha(A(t) - D(t))) = A(t). \quad (39)
\]

Following the same routine in analyzing the prototype system, we make basic observations as follows:

Corollary 5.5. Consider two arrival processes \( A_1(t) \) and \( A_2(t) \). Their corresponding departure processes are \( D_1(t) \) and \( D_2(t) \). If \( A_1(t) \geq A_2(t), \forall t \geq 0, \text{ then } D_1(t) \geq D_2(t), \forall t \geq 0. \]

Remark: This corollary is intuitive: a faster arrival will speed up the departure process.

Corollary 5.6. With stochastic arrival, \( D(t) \) is continuous and its growth rate is bounded by \( \alpha^{-1} \).

Remark: This corollary indicates that there is a fundamental limit on the growth rate of \( D(t) \), i.e., \( \alpha^{-1} \).

Theorem 5.7. If there exists a constant \( c, A(t) \) satisfies \( A(t) - \mu t \leq c, \forall t \geq 0, \) when \( \mu < \frac{1}{\alpha} \), the mean waiting time converges:

\[
\lim_{t \to \infty} \frac{\int_0^t T(x)dx}{t} = \frac{M}{1 - \alpha \mu}. \quad (40)
\]

Remark: This result has the same form as Theorem 5.4. This indicates that, for an asymptotically stationary arrival process, the departure process will also be asymptotically stationary.

5.3 Relaxation: Stochastic Departure

Now, we relax Assumption 5.2 even further, allowing \( M \) to be a random variable. Denote the pdf, cdf of \( M \) by \( f_M(x), F_M(x) \). The system now satisfies the following dynamics:

\[
G(t) = t + \alpha(A(t) - D(t)),
\]

\[
D(t) = \int_0^{G^{-1}(t)} \mu(x)F_M(G^{-1}(t) - x)dx. \quad (41)
\]

Such dynamics can be characterized by differential equation:

\[
D'(t) = \frac{\partial}{\partial t} \int_0^{G^{-1}(t)} \mu(x)f_M(G^{-1}(t) - x)dx. \quad (43)
\]

Corollary 5.8. Consider two arrival processes \( A_1(t) \) and \( A_2(t) \). Their corresponding departure processes are \( D_1(t) \) and \( D_2(t) \). If \( \mu_1(t) \geq \mu_2(t), \forall t \geq 0, \text{ then } D_1(t) \geq D_2(t), \forall t \geq 0. \]

Remark: This corollary is similar to Corollary 5.5 but requires a stronger condition. Specifically, a higher arrival rate \( \mu(t) \) across \( t \) is required to get a faster departure.

We can make two more observations similar to those in Section 5.2. They indicate that the structure of the problem remains the same, allowing for stochastic arrival and departure, simultaneously.
Corollary 5.9. With stochastic arrival and stochastic departure, $D(t)$ is continuous and its growth rate is bounded by $\alpha^{-1}$.

Theorem 5.10. If there exists a constant $c$, $A(t)$ satisfies $|A(t) - \mu t| \leq c, \forall t \geq 0$, when $\mu < \frac{1}{T}$, the mean waiting time converges:

$$\lim_{t\to\infty} \int_0^t T(x)dx = \mathbb{E}[M].$$

\section{Optimal Service Rate Design}

The service rate $\mu$ decided by the platform will influence the waiting time in both the queueing stage and picking-up stage. Consider there are $N$ passengers, we seek to derive the optimal service rate $\mu^*$ to minimize the total waiting time:

$$T_{\text{total}}(\mu) = \frac{NQ(p, \beta)}{\mu} + \frac{\mathbb{E}[M]}{1 - \alpha \mu}.$$  \hfill (45)

where

$$Q(p, \beta) = \lim_{N \to \infty} \frac{\mu^N}{N} \mathbb{E}[T_{\text{avg}}]^p.$$  \hfill (46)

Recall the expressions of $\mathbb{E}[T_{\text{avg}}]$ in Theorem 4.2, $Q(p, \beta)$ is a constant decided purely by policy $p$ and carpooling rate vector $\beta$. Under the condition $\alpha \mu < 1$, the optimal $\mu^*$ and optimal total waiting time $T_{\text{total}}^*$ satisfy:

$$\mu^* = \left(\alpha + \sqrt{\frac{\alpha M}{NQ(p, \beta)}}\right)^{-1}.$$  \hfill (47)

$$T_{\text{total}}^* = \mathbb{E}[M] + \alpha NQ(p, \beta)^p + 2\sqrt{\alpha M NQ(p, \beta)}.$$  \hfill (48)

Remark: When the congestion coefficient $\alpha$, $N$ or $\mathbb{E}[M]$ increases, the corresponding service rate needs to be lower, although the total waiting time will inevitably increase almost linearly. When the queueing factor $Q(p, \beta)$ becomes larger, the optimal service rate and the total waiting time will increase simultaneously.

\section{Numerical Analysis}

In this section, we examine the performance of our proposed methods using field data and measure how different carpool service policies and carpooling rates improve service quality. Further, we systematically estimate the transport hub congestion-related parameters to enable the pick up stage analysis. Finally, we demonstrate the performance of the optimal service rate design. Throughout the numerical study, we use field ride-hailing data in Beijing South Railway Station retrieved from Didi Chuxing platform, including more than 1 million desensitized driver trajectory records with geographical coordinates and time stamps.

\subsection{Queueing Evaluation: Benchmark}

We model the queueing process and calculate the length of busy period and the mean waiting time for each day of the year 2019 with different service rates. Figure 2 plots the distribution of busy period length and mean waiting time. We can infer from the figure that, with the increase in the service rate $\mu$, the number of days with shorter busy period and shorter mean waiting time both increase notably. This is very intuitive.

\section{Conclusion}

Then we expand the distribution of passengers’ waiting time for representative days. Consider a standard service rate $\mu = 500$ (vehicles/hour). Figure 3 illustrates 4 representative days’ waiting time distributions from peak to 2:00am (+1). They are New Year’s Eve, Lunar New Year’s Eve, Labor Day, and New Year’s Day. We can observe that, despite fewer passengers arriving on New Year’s Eve and Lunar New Year’s Eve, almost all orders were responded within 5 minutes. And the busy period ends fast. However, on Labor Day, a significant number of passengers will wait for more than 30 minutes. It is interesting to note the peak in the waiting time around 0 min: it indicates the busy period ends before 2:00am (+1). And for New Year’s Day, almost all passengers have to wait more than 50 minutes, which means the peak effects last extremely long, and the current service rate provided by the platform is not enough.

\subsection{Queueing Evaluation: Carpooling}

We seek to evaluate the mean waiting time of different groups when the maximum carpooled passenger is 2. Figure 4 visualizes the mean waiting time from January 2 to January 9 at different carpooling rates and under different carpool service policies. We set $p$ to be 2000 (vehicles/hour).

It is obvious that, for the carpool group, the $CF$ minimizes the mean waiting time while $AA$ maximizes it, and with the increased carpooling rate $\beta$, the differences in the mean waiting time between the different policies are decreasing. However, for no-carpooling group, the differences become more significant. Overall, the mean
waiting time for carpooling and no-carpooling groups under AA and FCFS decreases, but increases under CF.

For general carpooling cases, we allow up to 8-passenger carpooling to illustrate the properties of multiple-passenger carpooling. Specifically, we fix all the other elements in the carpooling rate vector $\beta$ to be 0.05, and let the fifth element evolve linearly from 0.05 to 0.55. We simulate the FCFS, AA and CF service process on the January 1 ride-hailing data for 1000 times, and plot the mean waiting time in Figure 5.

We can observe that, under FCFS policy, the average waiting time of all types of carpooling groups is decreasing in relation to the increase in the carpooling rate. Under AA policy, the mean waiting time for different groups falls at the same rate. And under the CF policy, only the 1-passenger carpooling group’s mean waiting time is decreasing, the 2 to 6-passenger carpooling group’s mean waiting time is increasing, while the 7 and 8-passenger carpooling groups’ mean waiting time remains unchanged. This result is consistent with the theoretical analysis, dictated by Theorem 4.3.

### 6.3 Fairness versus Effectiveness

In this section, we investigate the difference in efficiency and fairness among FCFS, AA and CF policies with the evolution of the carpooling rate vector $\beta$. Specifically, based on a fixed service rate $\mu$, we visualize the efficiency metric $M_{ef}$ and the fairness metric $M_{fair}$ using passenger demand data through the year 2019, for different carpool service policy $p$, and random carpooling rate $\beta$.

Figure 6 indicates that the orange dots (AA) are located on the $M_{ef}$ axis. This indicates that AA achieves the desired fairness among the different carpooling groups. The green dots (FCFS) are mostly to the upper left of the blue dots (CF): this indicates that FCFS policy tends to be fairer, albeit with a longer mean waiting time. Also, in Figure 6, we find the green dots and blue dots above the two lines (solid green and blue lines), which are the trade-off bounds between fairness and efficiency of FCFS and CF in any carpooling vector, which is described in Theorem 4.3.

We can also verify that the linear regression of the dots (purple and red dotted lines) are close to the bounds. Therefore, the bounds can serve as good performance estimators for the real ratios between fairness and efficiency with limited knowledge of the carpooling rate vector $\beta$. In practice, such estimates are valuable for the platform to design more customized products for heterogeneous passengers.

We then assess the relative performance of different carpool service policies as the proportion of carpooling rate increases. Consider a process from no passengers willing to carpool to all passengers willing to 4-passenger carpooling. During the process, we control the mean waiting time of the AA policy to be the same (by adjusting the service rate $\mu$). The relative effects of the other two policies on effectiveness and efficiency are shown in Figure 7 (a).

It is surprising that the curves corresponding to the two carpool service policies display a closed necklace. This means that, with the increasing carpooling rate, the change in relative efficiency can be divided into two stages: in the first stage, the unfairness increases, but the relative mean waiting time decreases; and, in the second stage, unfairness diminishes as the mean waiting time grows. Furthermore, comparing these two stages, under the same fairness, CF and FCFS will have a better relative mean waiting time in the second stage. That is, the efficiency of these two service policies will be more prominent when the carpooling rate is higher.

According to Figure 7 (b), we find that all the green dots are located exactly at the 40% quantile of the line segment connecting the corresponding blue and orange dots. This confirms Eq. (15) that FCFS is not the best with any trade-off coefficient $k$, and its overall performance is always between the other two service policies.

This observation is not true for all processes in the evolution of the carpooling rate. In Figure 7 (c), we build a process that $\beta$ linearly evolves from $[0,0,0] \rightarrow [0.05,0.5]$. The curves of CF and FCFS are not smooth: The trend changes drastically after a certain breakpoint, which is marked by a red star. In Figure 7 (d), each green dot is below the line connecting the corresponding blue dot.
6.4 Picking-up Stage Evaluation

In this section, we estimate the congestion-related parameters, i.e., the distribution of the random variable $M$ and the service delay ratio $\alpha$. Specifically, they satisfy:

$$E[T(t)] = E[M] + \alpha N(t). \quad (49)$$

We use linear regression to estimate $E[M]$ and $\alpha$ using the Didi Chuxing ride-hailing data, including the data from vehicles departing from Beijing South Railway Station from 11:00 pm to 01:00 am (+1) for the whole 2019 year.

The regression suggests that $\hat{\alpha}$ is 0.033 (min/passenger). This means that for every extra vehicle at the railway station, congestion time increases by 0.033 minute on average. It also indicates that the maximal capacity of the transportation hub is 1818 vehicles/hour. The data also suggest that $E[M]$ is 10.69 minutes.

We then investigate how the service rate will influence the length of the picking-up stage and the entire service process. We simulate the queuing stage and the picking-up stage for different service rates under the FCFS policy (the policy currently implemented in Didi Chuxing) using ride-hailing demand (with real carpooling rate) from the whole 2019 year. Figure 8 illustrates the duration of the queuing stage, the picking-up stage and the sum of both stages. Shaded areas indicate 25\% percentile to 75\% percentile.

It also suggests that the duration of queuing process decreases with $\mu$, and the duration of congestion process increases with $\mu$. There is an optimal dispatch rate $\mu^*$ to minimize the duration of the entire process as the green line is convex. The optimal theoretical decision (the green star) fits well.

Next we illustrate geographically how the dispatch rate $\mu$ will impact the total waiting time. The demand for ride-hailing in July, 2019 is visualized in Figure 9. We can observe that, the ride-hailing demand to destinations within the central city is strong. Clearly, it matches the residential characteristics of Beijing.

We analyze the daily demand for ride-hailing and carpooling data. We calculate the optimal daily dispatch rate $\mu^*$ according to Eq. (47), assuming that the dispatch process is Poisson. In addition, we set up two control groups with the daily dispatch rates of $0.5\mu^*$ and $2\mu^*$. We simulate the service process on the collected ride-hailing demand data. Figure 10 compares the geographical distribution of total waiting time.

In Figure 10 (b) with the optimal dispatch rate $\mu^*$, we can observe that the total waiting time of more than half passengers is less than 18 minutes. For most passengers, their waiting time is bounded by 26 minutes. Compared with Figure 10 (b), Figure 10 (a) illustrates that, the total waiting time for most passengers increases to more than 20 minutes due to the increase in queuing time. And in Figure 10 (c), most passengers also have to endure long waiting times. The proportion of passengers with waiting time longer than 30 minutes increases significantly, which undermines the efficiency of ride-hailing services. It also infers that severe congestion occurs at the railway station under the excessive service rate.

7 CONCLUSION

This paper examines the tension between efficiency and fairness for various carpool service policies. We use theoretic queueing analysis and numerical studies to highlight guidelines for selecting the most
appropriately service policy. Specifically, we offer theoretical trade-off bounds for these policies, which can serve as a performance estimation for platform designers to better understand the passengers. In addition, we offer the optimal service rate with practical verification. Our analytical framework can provide guidance to general carpooling design, and be easily applied to other scenarios sharing similar characteristics.

There are many interesting research directions that can be addressed in the future, for example, how to design the optimal time-varying service to dynamically control the platform. The main difficulties come from the complexity of the dynamics and the uncertainty in future demands. Other interesting directions include how to design a data-driven service policy with guaranteed performance, which might be solved by model predictive control [8].

REFERENCES


APPENDIX

7.1 Proof for Theorem 3.2

The mean waiting time for all passengers served during the busy period initiated with A passengers is:

\[
\mathbb{E}[T_Q(A)] = \frac{\sum_{i=0}^{\infty} P(k=i)(A+i)\mathbb{E}[T_Q(A)|N_A = i]}{\sum_{i=0}^{\infty} P(k=i)(A+i)},
\]

(50)

where \( P(k=i) \) indicates the probability that there are \( i \) passengers arriving during the whole busy period.

First, we describe the event \( e \) associated with \( \mathbb{E}[T_Q(A)|N_A = i] \):

\[
e = \{ T^u = \{ t_1^u, \ldots, t_{N_f+N_A}^u \}, T^d = \{ t_1^d, \ldots, t_{N_f}^d, T_d, \ldots, T_d \} \}.
\]

(51)

where \( N_f \) denotes the number of initial passengers; \( N_A \) denotes the number of passengers who arrive during the whole busy period. \( T^u \) denotes the set of \( N_f+N_A \) passengers' arrival times; \( T^d \) denotes the set of corresponding departure times.

For any event \( e \) defined in Eq. (51), denote \( T_d \) the corresponding busy period ending time. Thus, we can further define two mappings \( F^u, F^d \) as follows:

\[
F^u(e) = \{ T^u = \{ t_1^u, \ldots, t_{N_f+N_A}^u \}, T^d = \{ t_1^d, \ldots, t_{N_f}^d, T_d, \ldots, T_d \} \}.
\]

\[
F^d(e) = \{ T^u = \{ t_1^u, \ldots, t_{N_f}^u \}, T^d = \{ t_1^d, \ldots, t_{N_f+N_A}^d \} \}.
\]

Denote \( T(e) \) as the mean waiting time of event \( e \). The two mappings help us establish the derived bounds.

\[
T(F^d(e)) \leq T(e) \leq T(F^u(e)).
\]

(52)

This leads to the upper and lower bounds for \( \mathbb{E}[T_Q(A)|N_A = i] \):

\[
\mathbb{E}[T_Q(A)|N_A = i] \leq \sum_{e:N_f=N_A=i} f(e)T(F^u(e))de,
\]

(53)

\[
\mathbb{E}[T_Q(A)|N_A = i] \geq \sum_{e:N_f=N_A=i} f(e)T(F^d(e))de.
\]

(54)

Injecting Eq. (53) and (54) back to Eq. (50) allows us to construct the upper bound \( T^u \) and the lower bound \( T^d \) for \( \mathbb{E}[T_Q(A)] \).

7.2 Proof for Theorem 4.1

According to Little’s Law, the mean waiting time for each passenger will be decided by the mean number of passengers ahead of him. Following this logic, we derive mean waiting time for each policy.

Take \( \mathbb{E}[T_{AFCFS}^A] \) as an example, the following equation holds:

\[
\mathbb{E}[T_{AFCFS}^A] = \mathbb{E}[N_A] + \mathbb{E}[N_B] + 1 - \frac{1}{\mu},
\]

(55)

where \( N_A, N_B \) respectively denote the mean number of pairs in Group A, Group B ahead of the passenger of interest. Applying standard probabilistic method, the follow equations hold:

\[
\mathbb{E}[N_A] = \frac{\beta N}{4} - \frac{1}{2},
\]

(56)

\[
\mathbb{E}[N_B] = \frac{(1-\beta)N}{3}.
\]

(57)

Hence, \( \mathbb{E}[T_{AFCFS}^A] \) can be derived immediately. Following the same routine, we can derive all the results.

7.3 Proof Sketch for Theorem 4.2

We follow the same routine in the proof of Theorem 4.1 to derive these analytical results. Note that, the only difference is that when analyzing the performance of FCFS policy, we should determine the priority of carpooling pairs with the help of the following fact.

FACT 7.1. Consider two carpooling pairs, one with \( m-passerenger \) carpooling, another with \( n-passerenger \) carpooling. All these passengers are randomly distributed in the queue. The probability \( P_{mn} \) that \( m-passerenger \) pair is dispatched prior to \( n-passerenger \) pair is \( \frac{m}{m+n} \).

7.4 Proof Sketch for Theorem 4.3

By maximizing \( R_{CF} \) towards \( \beta \), we can derive the optimal carpooling rate \( \beta \) which only contains no-carpooling and \( M \)-passenger carpooling. Then the original problem is transformed into a univariate optimization problem. Standard mathematical manipulation leads to the optimal \( R_{CF} \).

\( R_{FCFS} \) can be derived in the same vein. The last result \( R_{CF} \geq R_{FCFS} \) can obtained through comparison.

7.5 Proof for Theorem 5.3

As \( D(0) = A(0) = N(0) = 0 \), we know \( T(0) = M \). Hence, \( D(0 + T(0)) = D(M) = A(0) = 0 \). Substituting them into Eq. (32) yields \( D((2+\alpha \mu)M) = \mu M \).

Following this iterative approach, we can derive a series of fixed points, where

\[
D(a_{n+1}) = \frac{n(n+1)M}{2}, \quad n \geq 0, \quad \alpha \mu = 1,
\]

(58)

\[
a_n = \frac{((\alpha \mu)^{n+1} - (\alpha \mu)(n+1) + n)M}{(\alpha \mu - 1)^2}, \quad n \geq 0, \quad \alpha \mu = 1.
\]

(60)

We can show that \( D(x) \) is linear within \( [a_n, a_{n+1}] \) \( n \geq 0 \) by mathematical induction. It turns out that the piece-wise linear function is the unique solution to the original differential equation.

7.6 Proof for Corollary 5.5

For any solution pair \( (t_1) \) and its corresponding solution \( D_1(t) \), we consider a generative action \( \mathcal{A}_{down} \) for arriving process \( A_2(t) \), such that \( A_2(t) = \mathcal{A}_{down}(A_1(t), x) \):

- If \( \lim_{t \to x^-} A_1(x) - A_1(t) = 0 \):

\[
A_2(t) = \begin{cases}
A_1(x), & t \in [x, x + \epsilon), \\
A_1(t), & t \notin [x, x + \epsilon).
\end{cases}
\]

(61)

- If \( \lim_{t \to x^-} A_1(x) - A_1(t) = \Delta(x) > 0 \):

\[
A_2(t) = \begin{cases}
A_1(x) - \gamma \Delta(x), & t \in [x, x + \epsilon), \\
A_1(t), & t \notin [x, x + \epsilon).
\end{cases}
\]

(62)

where \( \epsilon < M \) and \( 0 < \gamma < 1 \).

Clearly, \( A_2(t) \leq A_1(t), Vt \). Denoting \( D_2(t) \) the corresponding departure process of \( A_2(t) \), we now prove \( D_2(t) \leq D_1(t), Vt \).
It is clear that for \( t \in [0, x + M + \alpha(A_1(x) - D_1(x)), D_2(t) = D_1(t). \) (63)

The key is to show that \( D_2(t) \leq D_1(t) \) for \( t \in [x + M + \alpha(A_1(x) - D_1(x)), x + c + \epsilon + M + \alpha(A_1(x + \epsilon) - D_1(x + \epsilon))]. \)

By definition of \( A_2(t) \), we can indicate that \( D_2(t) \) is 2-segment piecewise linear over this range. The derivatives of the two segments are 0 and \( \epsilon^{-1} \), respectively. \( D_2(t) = D_1(t) \) holds at the two endpoints of the range. Together, they indicate \( D_1(t) \leq D_1(t) \) over this range.

Define the following recursion:

\[
a_{n+1} = a_n + M + \alpha(A_1(a_n) - D_1(a_n)), \quad n \geq 0, \quad b_{n+1} = b_n + M + \alpha(A_1(b_n) - D_1(b_n)), \quad n \geq 0,\]

where \( a_0 = x, b_0 = x + \epsilon. \)

We prove the following equations by mathematical induction:

\[
D_2(t) = D_1(t), \quad t \notin [a_i, b_i], \forall i, \]

\[
D_2(t) \leq D_1(t), \quad t \in [a_i, b_i], \forall i, \]

which concludes the proof for \( D_2(t) \leq D_1(t) \).

Similarly, we propose a symmetric action \( \mathcal{A}_{up} \) for arriving process \( A_3(t) = \mathcal{A}_{up}(A_1(t), x): \)

- If \( \lim \rightarrow (x + \epsilon) - A_1(x + \epsilon) - A_1(t) = 0: \)
  \[
  A_3(t) = \begin{cases} 
  A_1(x + \epsilon), & t \in [x, x + \epsilon), \\
  A_1(t), & t \notin [x, x + \epsilon). 
  \end{cases} \] \hspace{1cm} (68)

- If \( \lim \rightarrow (x + \epsilon) - A_1(x + \epsilon) - A_1(t) = \Delta(x + \epsilon) > 0: \)
  \[
  A_3(t) = \begin{cases} 
  A_1(x) + \gamma \Delta(x + \epsilon), & t \in [x, x + \epsilon), \\
  A_1(t), & t \notin [x, x + \epsilon), \end{cases} \] \hspace{1cm} (69)

where \( \epsilon < M \) and \( 0 < \gamma < 1 \).

Following the same routine, we can prove \( A_1(t) \leq A_3(t) \) and \( D_1(t) \leq D_3(t). \)

Now consider any two arriving processes \( A_{up}(t), A_{down}(t) \), \( A_{up}(t) \geq A_{down}(t) \). It is apparent that we can repeatedly conduct action \( \mathcal{A}_{down}(t) \) on \( A_{up}(t) \), and conduct action \( \mathcal{A}_{up}(t) \) on \( A_{down}(t) \) to make the two processes identical. Hence, we know

\[
D_{up}(t) \geq D_{down}(t). \]

7.7 Proof for Corollary 5.6

Consider the arrivals during \([x, x + \delta]\), the total waiting time at time \( x \) and \( x + \delta \) satisfies:

\[
T(x) = M + \alpha(A_1(x) - D_1(x)). \]

Hence the difference of the corresponding departure time is:

\[
\Delta = T(x + \delta) + x + \delta - T(x) - x \]

\[
= \alpha(A_1(x) + \Delta - A_1(x)) - \alpha(D_1(x + \delta) - D_1(x)) + \delta. \]

Let \( \delta \rightarrow 0^+ \), the following equation holds:

\[
D'(x + T(x)) = \frac{A'(x)}{1 + \alpha(A'(x) - D'(x))}. \]

According to Corollary 5.5, we know \( D'(x + T(x)) \) increases monotonically in \( A'(x) \). Hence \( 1 - \alpha D'(x) \geq 0 \). By the continuity of \( x \) we can obtain the desired result.

7.8 Proof Sketch for Theorem 5.7

For any \( A(t) \) and its solution \( D(t) \), we consider the following recursion:

\[
b_n = A(a_n), \forall n \geq 0, \quad a_{n+1} = D^{-1}(b_n), \forall n \geq 0. \]

According to \( \lim \rightarrow \infty |A(t) - \mu t| < \epsilon \) and Eq. (39), it holds:

\[
\lim \rightarrow \infty a_n = \mu, \forall n \geq 0, \quad a_{n+1} - a_n = M + \alpha(b_n - b_{n-1}), \forall n \geq 1. \]

The elimination summation of Eq. (78) satisfies:

\[
a_{n+1} - a_n = nM + \alpha(b_n - b_0). \]

We can verify that \( \lim \rightarrow \infty \frac{a_{n+1} - a_n}{n} = 0 \). Combining Eq. (75) and Eq. (79) yields that

\[
\lim \rightarrow \infty \frac{a_n}{n} = \frac{M}{1 - \alpha \mu}. \]

In fact, Eq. (80) provides the sample mean of \( T(t) \), as we can arbitrarily select \( a_0 \). Further mathematical manipulation yields the conclusion.

7.9 Proof Sketch for Corollary 5.9

From Eq. (43), we can infer that \( C'(t) > 0. \) And according to Corollary 5.8, the longest \( D(t) \) will be achieved when \( \mu(t) \) achieves maximal. Hence we can prove this lemma following the same routine as the proof for Corollary 5.6.

7.10 Proof Sketch for Theorem 5.10

We can prove this theorem in the same vein of the proof for Theorem 5.7.

Specifically, again we construct a recursive formula:

\[
a_{n+1} = a_n + \alpha(A(a_n) - D(a_n)) + \tilde{M}, \quad a_{n+1} = b_n + \alpha(A(b_n) - D(b_n)), \]

where \( \tilde{M} \) is the upper bound of \( M, a_1 \in [0, M] \). We can infer that \( a_n \geq n\tilde{M}, b_n \geq a_n \).

Furthermore, \( D(a_{n+1}) \) can be represented by:

\[
D(a_{n+1}) = A(a_n) + \int_{a_n}^{b_n} F_M(a_{n+1} - B(x)) \mu(x) dx. \]

Summing Eq. (81) over \( n \) yields

\[
\frac{a_{n+1} - a_1}{n} = \frac{\alpha \sum_{i=1}^{n} (A(a_i) - D(a_i))}{n} + \tilde{M}. \]

Combining \( \lim \rightarrow \infty |A(t) - \mu t| < \epsilon \) and Eq. (42) with Eqs. (83), (84), standard mathematical manipulation indicates the key results:

\[
\lim \rightarrow \infty \frac{A(a_n)}{D(a_n)} = 1, \quad \lim \rightarrow \infty \frac{a_n}{n} = \frac{\mathbb{E}[M]}{1 - \alpha \mu}. \]

These key results immediately lead to the conclusion.