# Dimensioning On-Demand Vehicle Sharing Systems

Saif Benjaafar[†], Shining Wu[‡], Hanlin Liu[§], Einar Bjarki Gunnarsson[†]

[†]Department of Industrial & Systems Engineering, University of Minnesota, Twin Cities, USA

[‡]Department of Logistics & Maritime Studies, Hong Kong Polytechnic University, Hong Kong

[§]Division of Information Systems and Management Engineering, College of Business,
Southern University of Science and Technology, Shenzhen, China

saif@umn.edu, sn.wu@polyu.edu.hk, liuhl@sustech.edu.cn, gunna042@umn.edu

**Abstract**

We consider the problem of optimal fleet sizing in a vehicle sharing system. Vehicles are available for short-term rental and are accessible from multiple locations. A vehicle rented at one location can be returned to any other location. The size of the fleet must account not only for the nominal load and for the randomness in demand and rental duration but also for the randomness in the number of vehicles that are available at each location due to vehicle *roaming* (vehicles not returning to the same location from which they were picked up). We model the dynamics of the system using a closed queueing network and obtain explicit and closed form lower and upper bounds on the optimal number of vehicles (the minimum number of vehicles needed to meet a target service level). Specifically, we show that starting with any pair of lower and upper bounds, we can always obtain another pair of lower and upper bounds with gaps between the lower and upper bounds that are independent of demand and bounded by a function that depends only on the prescribed service level. We show that the generated bounds are asymptotically exact under several regimes. We use features of the bounds to construct a simple and closed form approximation that we show to be always within the generated lower and upper bounds and is exact under the asymptotic regimes considered. Extensive numerical experiments show that the approximate and exact values are nearly indistinguishable for a wide range of parameter values. The approximation is highly interpretable with buffer capacity expressed in terms of three explicit terms that can be interpreted as follows: (1) standard buffer capacity that is protection against randomness in demand and rental times, (2) buffer capacity that is protection against vehicle roaming, and (3) a correction term. Our analysis reveals important differences between the optimal sizing of standard queueing systems (where servers always return to the same queue upon service completion) and that of systems where servers, upon service completion, randomly join any one of the queues in the system. We show that the additional capacity needed to buffer against vehicle roaming can be substantial even in systems with vanishingly small demand.

**Keywords:** On-demand vehicle sharing systems; closed queueing networks; capacity optimization; bounds and approximations

# 1 Introduction

We consider a vehicle sharing system with a fixed number of vehicles distributed across multiple locations. The vehicles are available for short term rental and are accessible from any location and can be returned to any other location. In other words, the system allows for *one-way* trips, as opposed to one that requires *round trips* in which a vehicle picked up at a location always returns to the same location. Demand at each location is random and assumed to arise continuously over time according to a Poisson process. Rental duration is also random and dependent on the origin and destination of the trip and so are the locations at which vehicles are returned at the end of the trip.

Vehicle sharing systems with features similar to the ones described above are increasingly common and include one-way car, bike, and scooter sharing systems. Features that are common to these systems include *on-demand access* (vehicles are accessed without prior reservation and without requiring customers to divulge trip duration or destination), *multi-locations* (vehicles can be accessed at multiple locations that are spatially distributed), and *one-way service* (vehicles can be dropped off at locations that are different from those at which they were picked up). These features, along with the associated randomness in demand, rental duration, and returns, make the design (e.g., the scale and scope of these systems) and operation (e.g., dynamic vehicle allocation) of these systems challenging in some cases.

In this paper, we address the problem of how to optimally dimension a vehicle sharing system with the above features. Specifically, for the class of vehicle sharing systems we consider, we are interested in determining the minimum number of vehicles that can guarantee a specified service level, where the service level refers to a threshold on the probability that a customer who seeks to rent a vehicle at a location would find one available (the problem can also be viewed as one of minimizing the number of vehicles subject to a service level constraint).

We model the dynamics of a vehicle sharing system using a closed queueing network (where the items moving through the network correspond to vehicles) with two types of queues: pick-up queues modeled as single server queues and corresponding to locations and transit queues modeled as infinite server queues and corresponding to vehicles in transit. We first consider *balanced* systems (systems where each location is as popular an origin as it is a destination). Using mean value analysis, we show how it is possible to obtain a recursive relationship between the service level of a system with $K$ vehicles and the service level of a system with $K - 1$ vehicles. This recursive relationship (and its analogue for unbalanced systems) can be used to efficiently compute the optimal number of vehicles (i.e., the minimum number of vehicles that meets the service level constraint).

We use this recursive relationship to obtain explicit and closed form lower and upper bounds on the optimal number of vehicles. Specifically, we show that starting with any pair of lower and upper bounds, we can always obtain another pair of lower and upper bounds. In the case where the procedure is initialized with a particular (and well specified) pair of bounds, the bounds can be made tighter with additional iterations. In all cases, we show that the gap between the lower and upper bounds is finite and bounded by $\frac{1}{1-\alpha}$, where $\alpha$ is the service level (note that the bound on the gap is independent of the number of locations and the demand level).

We show that the bounds generated using our procedure are exact (i.e., converge to the exact value of the optimal number of vehicles) under several asymptotic regimes, including when the demand approaches 0 or infinity, the number of locations approaches infinity, and the service level approaches 1. We use features of the bounds to construct a simple and closed form approximation that we show to be always within the generated lower and upper bounds and is exact under the asymptotic regimes considered. Extensive numerical experiments show that the approximate and exact values for the optimal number of vehicles are nearly indistinguishable for a wide range of parameter values[1].

The approximation reveals the following important insights into factors that affect the optimal number of vehicles, or optimal capacity.

- Optimal capacity can be expressed as the sum of two terms: nominal load (number of vehicles needed to handle the induced load on the system) and buffer capacity.
- Buffer capacity consists of three terms that can be interpreted as follows: *standard* buffer capacity that is protection against randomness in demand and service (rental) times, buffer capacity that is protection against vehicle *roaming* (vehicles not returning to the same location from which they were originally and the resulting randomness in the service capacity available at each location), and a correction term. We attribute the second term to vehicle roaming since it reduces to zero when vehicles do not roam (i.e., return to the same location upon trip completion). Because this additional capacity can also buffer against randomness from other sources, the overall need for buffer capacity diminishes and there is a need for the correction term.
- The buffer capacity we attribute to protection against vehicle roaming is given by $(N-1)\frac{\alpha}{1-\alpha}$, which is increasing in the number of locations $N$ and the service level $\alpha$. This buffer capacity is independent of the demand level in contrast to the standard buffer capacity which increases with demand, as higher demand translates into a higher load.
- The fact that the term $(N-1)\frac{\alpha}{1-\alpha}$ is independent of demand means that even for vanishingly small demand (or service times), buffer capacity can still be substantial.
- The above insights reveal fundamental differences between queueing systems where “servers” (i.e., the resources involved in fulfilling the demand from customers) always return to the same queue upon completing service and queueing systems where servers roam (i.e, systems where, upon completing

[1] Our results apply to the important special case of a system with a single location, which corresponds to the widely studied Erlang loss system. We show that our approximation, specialized to this case, performs well against approximations from the literature.

service, a server may join another queue), with the additional capacity resulting from vehicle roaming being substantial.

We extend our analysis to unbalanced systems (i.e., systems where some locations are more popular as a destination than they are as an origin, or vice-versa). We discuss how an unbalanced system can be balanced via vehicle repositioning and show how our results for balanced systems, including the approximation for the optimal number of vehicles, can be adapted for unbalanced systems under repositioning.

The results in the paper highlight the fact that one-way vehicle sharing systems (and more generally queueing systems where servers may roam) may be substantially more expensive to operate than systems that require round trips. This implies that the economic viability of the one-way model crucially depends on the additional revenue a service provider expects from the increased convenience of one-way service.

The rest of the paper is organized as follows. In Section 2, we review related literature. In Section 3, we describe the model. In Section 4, we describe our procedure for generating bounds. In Section 5, we describe the asymptotic results and the approximation. In Section 6, we draw managerial insights. In Section 7, we discuss the case of unbalanced systems. In Section 8, we offer concluding comments.

## 2 Related Literature

Although the literature on the design, planning, and operation of on-demand vehicle sharing systems is extensive and growing (see for example recent reviews in Freund et al. (2019), He et al. (2019), and Benjaafar and Hu (2020)), papers that consider optimal fleet sizing, while accounting for the underlying queueing dynamics, are relatively few. One of the earliest papers to consider a problem similar to ours is George and Xia (2011). They use a closed queueing network model, as we do, to model systems dynamics and develop exact and approximate solution algorithms to determine optimal fleet size where the objective is to maximize system profit, but do not provide, as we do, explicit expressions (exact or approximate) for the minimal fleet size. He et al. (2017) also consider a similar problem and a similar closed queueing network formulation to ours. They use the fixed population mean (FPM) approach (see Whitt (2002)) to approximate the closed queueing network with one that is open. They derive an approximation for the minimum fleet size needed to meet a specified service level. The same approximation is used in Bellos et al. (2017) in a different context for a system with a single location. The approximation in He et al. (2017) and Bellos et al. (2017) corresponds to one of the upper bounds we identify (see expression (11) in this paper). As we show in the paper, this bound is not asymptotically exact under the various regimes we consider.

Other papers that consider the optimal fleet sizing while accounting for the queueing dynamics include Hu and Liu (2016), Zhang et al. (2019) and Li et al. (2019). These papers primarily rely on algorithms and numerical procedures to determine the minimal fleet size. There are papers that rely on queueing models to consider other design and operation aspects of vehicle sharing systems. For example, Banerjee et al. (2017)

characterize the steady state distribution of a vehicle sharing system with price-dependent demand flows and develop efficient algorithms for trip pricing and vehicle rebalancing. Braverman et al. (2019) establish a fluid approximation of a large-scale ridesharing system and derive a fluid-based optimal empty vehicle routing policy.

For the class of closed queueing networks implied by vehicle sharing systems, George et al. (2012) derive the exact-order asymptotic growth rate of system throughput as the number of items (vehicles) increases and Banerjee et al. (2017) provide a lower bound on the service level as the number of vehicles increases to infinity in a balanced system that is induced by the optimal pricing under an elevated flow relaxation. Although these results can be applied to obtain the optimal number of vehicles in an asymptotic sense, they do not lead to closed-form approximations. The analysis and the approximation we propose lead to an exact-order asymptotics that is consistent with that of George et al. (2012) and tighter than that of Banerjee et al. (2017). Waserhole and Jost (2016) consider a setting similar to ours but assume that trips are instantaneous (i.e., trip durations are zero). They obtain a relationship between service level and the number of vehicles that is a special case of the relationship we obtain (see equation (8)) for when $1/\mu = 0$.

There is significant literature that focuses on approximating the normalizing constant of the steady-state distribution in closed queueing networks (see for example Kogan and Birman (1992), Kogan (1992), and Hofri and Kogan (1994) who consider a class of systems consisting of a single infinite-server queue and many single-server queues with application in computer networks). Though they simplify computations, these approximations do not typically yield simple closed form expressions for performance measures of interest.

There is a large body of literature that considers the problem of optimal capacity in the context of a queueing system with a single location (i.e., a system with multiple servers and a single queue). The problem is often referred as the optimal staffing problem in reference to the staffing of call centers, an important application; see for example Gans et al. (2003) and Whitt (2007). An important result from this literature is the so-called square root staffing rule, whereby the number of servers is set equal to $a + \beta\sqrt{a}$ where $a$ is the nominal load and $\beta$ is a function of the service level (the probability that a customer does not need to wait for service). The result arises naturally under a normal approximation of the number of customers in the system and in heavy traffic under appropriate scaling. In particular, Halfin and Whitt (1981) show that taking an $M/M/n$ queueing system to heavy traffic by scaling the number of servers as $a + \beta\sqrt{a}$, the probability that a customer does not wait for service is guaranteed to be strictly between zero and one, the so-called Halfin-Whitt regime. The follow on literature on this topic is extensive and we refer the reader to reviews by Gans et al. (2003), Whitt (2007), Mandelbaum and Zeltyn (2009), Dai and He (2014) and Ward (2012). In our case, we consider a system with a network feature where servers are routed probabilistically to different queues upon service completion. We show that the introduction of this feature adds a new component to buffer capacity that is independent of demand and increasing in the number of locations. In fact, in the limiting case of an infinitely large number of locations, buffer capacity ceases to depend on demand altogether.

A special case of the setting we consider in this paper is the well-studied Erlang loss system (an $M/M/n/n$ queueing system). Although the literature on Erlang loss systems is extensive (see for example Jagerman (1974), Cooper (1981), Harel (1988), Janssen et al. (2008), Adelman (2008)), literature on the optimal sizing of these systems is relatively limited. There is significant literature that studies, for a given number of servers, the blocking probability (the probability that an arriving customer finds all servers busy, also known as the Erlang loss formula). This includes literature that offers various bounds and approximation; see for example, Janssen et al. (2008), and Adelman (2008) and the references therein. However, the inverse problem (the problem of determining the number of servers needed to guarantee a certain threshold on the blocking probability) is less studied. In this paper, we show that the results we obtain for a general network can be specialized for the case of an Erlang loss system. In particular, we show that our approximation, specialized for a single location problem, performs well relative to approximations considered among the best in the literature such as those in Berezner et al. (1998) and Harel (2010).

Finally, we note that there is emerging literature that considers the issues of optimal service capacity in the context of ridesharing platforms (platforms that connect passengers with independent drivers, such as Uber and Lyft). In this setting, service capacity is determined indirectly via the choice of wages the platform pays the drivers; see for example Cachon et al. (2017), Taylor (2018), Benjaafar et al. (2020) and the references therein. Papers that consider the spatial features of ridehailing include Castillo et al. (2018), Afeche et al. (2018), Bimpikis et al. (2019) and Besbes et al. (2020). Besbes et al. (2019) study the problem of optimal service capacity when the ridehailing system is modeled as a single multi-server queue with a state-dependent service rate (the state dependency is needed to account for the customer pick up time portion of total service time). They show that, under heavy traffic, a square root staffing approach is not sufficient to achieve a Halfin-Whitt-like regime and that instead buffer capacity that is proportional to the nominal load to the power of 2/3 is needed to account for pick up time.

# 3 Model Description and Preliminaries

Consider a vehicle sharing system where $K$ vehicles are available for short-term rental. Vehicles can be picked up and dropped off at one of $N$ locations. Customers arrive continuously over time at each location according to a Poisson process with arrival rate $\lambda_i$ at location $i$, where $i = 1, \ldots, N$. A vehicle picked up at location $i$ is returned to location $j$ with probability $p_{ij}$, where $\sum_{j=1}^{N} p_{ij} = 1$ for all $i$. The rental duration for a vehicle picked up at location $i$ and returned to location $j$ follows a distribution that has rational Laplace transform and a mean $\frac{1}{\mu_{ij}}$.[2] A customer who arrives at a location and finds no vehicles available at that location immediately leaves the system without renting. We consider a balanced system where $\lambda_i = \sum_j \lambda_j p_{ji}$.

[2] The family of distributions that have rational Laplace transform is dense in all non-negative distributions (Botta et al., 1987). Exponential, hypoexponential, hyperexponential, mixed generalized Erlangs, and generalized hyperexponential distributions all belong to this family.

This condition implies that each location is as popular an origin as it is a destination. A special case of a balanced system is a symmetric system where $\lambda_i = \lambda_j$ for any $i \neq j$ and $p_{ij} = p_{ij'} = \frac{1}{N}$ for any pair $j$ and $j'$. Note that balance arises naturally in many vehicle sharing systems because of the rebalancing of vehicles that is typically carried out by the service provider; see Section 7 for further discussion. In Section 7, we provide analysis for unblanced systems.

The system as described above can be viewed as a closed queueing network where the items moving through the network correspond to vehicles. In particular, each location can be viewed as a single server queue with service times corresponding to the customer inter-arrival times at that location. We refer to such queues as *pickup* queues. A vehicle that is picked up at location $i$ with intended destination $j$ can be viewed as entering an infinite-server queue with service times corresponding to the travel times between location $i$ and $j$. We refer to such queues as *transit* queues. Note that a pick up queue is associated with each location $i$ for $i = 1, \dots, N$ and a transit queue $(i, j)$ is associated with each pair of locations $i$ and $j$ for which $p_{ij} > 0$. A vehicle that completes service at transit queue $(i, j)$ joins pick-up queue $j$. Without loss of generality, we assume that the routing matrix specified by the probabilities $p_{ij}$ is irreducible (i.e., a vehicle at any location $i$ can reach any other location in finitely many steps with positive probability). The network, as specified above, is an instance of a BCMP network (Baskett et al., 1975)[3].

Our objective is to characterize the relationship between service level and the number of vehicles in the system which, in turn, would allow us to determine the minimum number of vehicles needed to achieve a specified service level. To do so, we can proceed in at least one of two ways. The first involves characterizing the probability distribution of system states while the other does not. We describe the first approach next. The state space can be specified by the number of vehicles at each pick up and transit queue. Let $X_i$ denote the number of vehicles in pick up queue $i$ (this corresponds to the number of idle vehicles in location $i$) and $Y_{ij}$ denote the number of vehicles in transit queue $(i, j)$ (this corresponds to the number of rented vehicles from location $i$ destined to location $j$) for $i, j \in V$ where $V = \{1, \dots, N\}$. The state space can then be defined as $S = \left\{(\boldsymbol{X}, \boldsymbol{Y}) \,\middle|\, \sum_{i \in V} X_i + \sum_{i,j \in V} Y_{ij} = K\right\}$, where $\boldsymbol{X}$ is the vector with components $X_i$ and $\boldsymbol{Y}$ is the matrix with component $Y_{ij}$ for $i, j \in V$. A BCMP network is known to have a product form for the steady state probability distribution over the system states (Baskett et al., 1975). In particular,

$$\Pr(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}) = C \prod_{i \in V} (\frac{\gamma_i}{\lambda_i})^{x_i} \prod_{i,j} (\frac{\gamma_{ij}}{\mu_{ij}})^{y_{ij}} \frac{1}{y_{ij}!} \tag{1}$$

[3] A queueing network is called a BCMP network if (i) it has a finite number of queues (locations in our case) and a finite number of classes of items, (ii) the routing among the queues is governed by fixed transition probabilities, and (iii) the arrival processes and service disciplines are of the types specified in §2.1 of Baskett et al. (1975). Since our network is closed (i.e., no external arrivals) and the queues have either one server with exponential service times or infinite servers with service time distributions having rational Laplace transform, this condition (iii) is satisfied.

for all $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{S}$, where $C$ is a normalization constant that satisfies

$$\sum_{(\boldsymbol{x},\boldsymbol{y})\in S} \Pr(\boldsymbol{X} = \boldsymbol{x}, \boldsymbol{Y} = \boldsymbol{y}) = C \sum_{(\boldsymbol{x},\boldsymbol{y})\in S} \prod_{i\in V} (\frac{\nu_i}{\lambda_i})^{x_i} \prod_{i,j} (\frac{\nu_{ij}}{\mu_{ij}})^{y_{ij}} \frac{1}{y_{ij}!} = 1, \tag{2}$$

$\nu_i$ is the steady state *throughput* of pick-up queue $i$ (the average number of vehicles successfully rented at location $i$ per unit time), and $\nu_{ij}$ is the steady state throughput of transit queue $(i, j)$ (the average number of vehicles picked up at location $i$ and returned to location $j$ per unit time).

The throughput rates can be computed as follows. Noting that, in steady state, the rate at which vehicles are returned to location $i$ is equal to the rate at which they are picked up at location $i$ leads to the following set of balance equations

$$\sum_{j=1}^{N} \nu_j p_{ji} = \nu_i, \tag{3}$$

for all $i = 1, \ldots, N$. Moreover, $\nu_{ij} = \nu_i p_{ij}$. Because the routing matrix $[p_{ij}]$ is irreducible, the $N$ balance equations in (3) allow us to solve for the throughput rates $\nu_i$'s up to a scalar multiple. This is sufficient for computing the probability distribution since this scalar multiple can be subsumed in the normalization constant in (2).

Let $\Lambda := \sum\limits_{i\in V} \lambda_i$ denote the total arrival rate to the network, $\nu := \sum\limits_{i\in V} \nu_i$ the total throughput rate, and

$$\alpha_i := \frac{\nu_i}{\lambda_i}$$

the long run fraction of customers who find an available vehicle (also the probability in steady state that a customer finds an available vehicle upon arrival at location $i$). We refer to $\alpha_i$ as the service level at location $i$. Recall that for a balanced network, $\lambda_i = \sum_j \lambda_j p_{ji}$ for all $i = 1, 2, \cdots, N$. This condition makes the $\lambda_i$'s a solution to (3), implying that the throughput rates have the form $\nu_i = \alpha\lambda_i$ for all $i$, where $\alpha$ is a scalar multiple that can be computed using the normalization equation (2). Noting that the service level at location $i$ is given by $\alpha_i = \frac{\nu_i}{\lambda_i}$, we have $\alpha_i = \alpha$ for all $i$. That is, perhaps consistent with intuition, in a balanced network, the service level is the same at all locations.

Having obtained the probability distribution of the systems states, various system performance metrics can in principle be calculated. However, the computational effort involved can be significant. In particular, solving for the normalization constant $C$ is challenging since the number of states $\binom{K+N+N^2-1}{K}$ increases exponentially in $N$ and $K$. Moreover, the fact that performance measures of interest are not in closed form limits our ability to carry out further analysis. Therefore, in what follows, we resort to a different approach, *mean value analysis*, which allows us to bypass the need to compute the steady state distribution in order to compute throughput rate, the measure that is of primary interest for our analysis in this paper.

Mean value analysis relies on the *random observer property* (Reiser and Lavenberg, 1980), which states that an arrival to any queue in a BCMP network with $K$ items (vehicles in our case) observes the stationary distribution of an identical network with $K-1$ items (vehicles). As we describe below, this property can be exploited to derive a relationship between the throughput of a system with $K$ vehicles and the throughput of a system with $K-1$ vehicles. This relationship can be used to recursively compute the throughput for a system with $K$ vehicles.

Let $\frac{1}{\mu} := \sum\limits_{i,j\in V} \frac{\nu_{ij}}{\nu\mu_{ij}}$ denote the average rental time of a vehicle in the network (note that we use the fact that $\frac{\nu_{ij}}{\nu}$ is the proportion of effective rentals that originate in location $i$ and terminate in location $j$). Because, in a balanced network, $\frac{\nu_i}{\lambda_i} = \frac{\nu}{\Lambda}$ and $\nu_{ij} = \nu_i p_{ij}$, we can also express the average rental time as $\frac{1}{\mu} = \sum\limits_{i,j\in V} \frac{\lambda_i p_{ij}}{\Lambda\mu_{ij}}$. Let $\mathbb{E}[X_i(K)]$ denote the expected number of vehicles at a pick up queue $i$ and $\mathbb{E}[Y_{ij}(K)]$ denote the expected number of vehicles at a transit queue $(i,j)$ given there are $K$ vehicles in the system (let also $\nu_i(K)$, $\nu_{ij}(K)$, $\nu(K)$, and $\alpha(K)$ be similarly defined). By Little's Law,

$$\sum_{i,j\in V} \mathbb{E}[Y_{ij}(K)] = \frac{\nu(K)}{\mu}. \tag{4}$$

Because $\sum\limits_{i,j\in V} \mathbb{E}[Y_{ij}(K)] + \sum\limits_{i\in V} \mathbb{E}[X_i(K)] = K$, we have

$$\sum_{i\in V} \mathbb{E}[X_i(K)] = K - \frac{\nu(K)}{\mu}. \tag{5}$$

Noting that the expected number of vehicles at a pick up queue $i$ observed by an arriving vehicle at location $i$ is, by virtue of the random observed property, given by $\mathbb{E}[X_i(K-1)]$, the expected time the vehicle spends in that queue is given by $\frac{1+\mathbb{E}[X_i(K-1)]}{\lambda_i}$. Applying Little's Law, we obtain $\mathbb{E}[X_i(K)] = \nu_i(K)\frac{1+\mathbb{E}[X_i(K-1)]}{\lambda_i}$, and hence

$$\sum_{i\in V} \mathbb{E}[X_i(K)] = \sum_{i\in V} \left\{ \nu_i(K) \frac{1+\mathbb{E}[X_i(K-1)]}{\lambda_i} \right\} = \frac{\nu(K)}{\Lambda} \sum_i \left\{ 1 + \mathbb{E}[X_i(K-1)] \right\}. \tag{6}$$

Substituting the expression in (5) for $\sum\limits_{i\in V} \mathbb{E}[X_i(K)]$ and $\sum\limits_{i\in V} \mathbb{E}[X_i(K-1)]$ into (6) leads to $K - \frac{\nu(K)}{\mu} = \frac{\nu(K)}{\Lambda}\left[N + (K-1) - \frac{\nu(K-1)}{\mu}\right]$, which yields for $K \geq 2$,

$$\nu(K) = \frac{K\Lambda\mu}{(K+N-1)\mu + \Lambda - \nu(K-1)}. \tag{7}$$

Equation (7) provides a recursive relationship that allows us to compute the throughput for a system with $K$ vehicles knowing the throughput for a system with $K-1$ vehicles. For $K=1$, we can directly show that $\nu(1) = \frac{\Lambda\mu}{N\mu+\Lambda}$. Hence, by letting $\nu(0) := 0$, equation (7) holds for all $K \geq 1$. We are now ready to state the following important lemma.

**Lemma 1.** *For $K \geq 1$*

$$\alpha(K) = \frac{K}{(K+N-1) + \frac{\Lambda}{\mu}[1-\alpha(K-1)]}, \tag{8}$$

*where $\alpha(0) := 0$.*

The result in the lemma follows immediately from Equation (7) using the fact that $\alpha(K) := \frac{\nu(K)}{\Lambda}$. Lemma 1 provides a recursive relationship for computing the service level $\alpha(K)$ for a system with $K$ vehicles knowing the service level for a system with $K-1$ vehicles. More importantly, equation (8) allows us to calculate the minimum number of vehicles that can guarantee a specified service level. That is, (8) can be the basis for an algorithm to solve for

$$K(\alpha) := \min\{K \geq 0 : \alpha(K) \geq \alpha\},$$

where $\alpha$ is the target service level that must be satisfied. A myopic search approach is sufficient here since the service level can be shown to be monotonically increasing in $K$.

Although it is possible to efficiently compute the minimum number of vehicles to meet a specified service level using (8), (8) does not provide an explicit expression for the minimum number of vehicles. This has two shortcomings: (i) it is difficult to obtain insights into the determining factors behind optimal fleet sizing (a main objective of this paper) and (ii) it may be difficult to carry out further analysis that involves the minimal fleet size (e.g., endogenizing the service level using a profit maximization model as we do in Appendix C).

**Remark 1** Throughout the paper, we abuse notation and use $\alpha$ to denote a fixed scalar value for service level and $\alpha(K)$ to refer to service level as a function of the number of vehicles. We similarly abuse notation and use $K$ and $K(\alpha)$ to refer respectively to a fixed number of vehicles and the minimum number of vehicles needed to meet a target service level $\alpha$.

# 4 Bounds

In this section, we derive closed form bounds for the minimum number of vehicles $K(\alpha)$ needed to guarantee a specified service level $\alpha$, to which we refer as *the minimal fleet size*. We show that the gap between these bounds narrows to 1 under various asymptotic regimes (i.e., they provide approximations that are asymptotically exact). Moreover, we describe how each pair of lower and upper bounds can be used, with proper initialization, to construct another tighter pair of lower and upper bounds.

We rewrite equation (8) as follows

$$\alpha(K) = \frac{K}{(K+N-1) + \frac{\Lambda}{\mu}[1-\alpha(K)+\Delta\alpha(K)]},$$

where $\Delta\alpha(K) := \alpha(K) - \alpha(K-1)$, which, by simple algebra, yields

$$K = \frac{\Lambda}{\mu}\alpha(K) + (N-1)\frac{\alpha(K)}{1-\alpha(K)} + \frac{\Lambda}{\mu}\frac{\alpha(K)}{1-\alpha(K)}\Delta\alpha(K). \tag{9}$$

The above expression suggests that bounds on $\Delta\alpha(K)$ can be used to obtain bounds on $K(\alpha)$. In the lemma below, we obtain such bounds (the proof of Lemma 2 and all other results, unless indicated otherwise, can be found in the Appendix).

**Lemma 2.** *For all $K$, $0 < \Delta\alpha(K) < \frac{\mu}{\Lambda}$.*

The bounds on $\Delta\alpha(K)$ in Lemma 2 immediately lead to the bounds on $K(\alpha)$ described in the following proposition.

**Proposition 1.** *For $\alpha > 0$,*

$$K(\alpha) > \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} := L_0, \text{ and} \tag{10}$$

$$K(\alpha) < \frac{\Lambda}{\mu}\alpha + N\frac{\alpha}{1-\alpha} + 1 := U_0 \tag{11}$$

The bounds in Proposition 1, though simple to obtain, are surprisingly tight. This can be seen from the difference between the lower and upper bounds, given by $U_0 - L_0 = \frac{1}{1-\alpha}$. This difference is independent of the number of locations $N$ and the demand level $\Lambda$ (in particular, it does not increase with an increase in either parameter).

Next, we describe an iterative procedure that allows us to produce even tighter bounds and to lead to approximations that are asymptotically exact. In particular, given any pair of lower and upper bounds, we show that it is possible to produce another (tighter under proper initialization and up to an upper bound on the number of iterations) pair of lower and upper bounds.

We first briefly describe the idea of how we generate the iterative bounds. Note that by (8) we have

$$\Delta\alpha(K) = \frac{K}{(K+N-1) + \frac{\Lambda}{\mu}(1-\alpha(K-1))} - \frac{K-1}{(K+N-2) + \frac{\Lambda}{\mu}(1-\alpha(K-2))}$$

which we can rewrite as

$$\Delta\alpha(K) = \frac{(N-1) + \frac{\Lambda}{\mu}[1-\alpha(K-1)] + K\frac{\Lambda}{\mu}[\Delta\alpha(K-1)]}{[(K+N-1) + \frac{\Lambda}{\mu}(1-\alpha(K-1))][(K+N-2) + \frac{\Lambda}{\mu}(1-\alpha(K-2))]}. \tag{12}$$

Note that (12) provides us with a relationship between $\Delta\alpha(K)$ and $\Delta\alpha(K-1)$. By iterating on this relationship $s-1$ times, we can obtain a relationship between $\Delta\alpha(K)$ and $\Delta\alpha(K-s)$. A pair of lower and upper bounds on $\Delta\alpha(K)$ can be obtained by replacing $\Delta\alpha(K-s)$ with its lower and upper bounds. That is, each iteration yields a different pairs of bounds on $\Delta\alpha(K)$. In turn, this allows us to obtain iteratively bounds on $K(\alpha)$.

Suppose $L$ and $U$ are known lower and upper bounds for $K(\alpha)$ (i.e., $L \leq K(\alpha) \leq U$). We define quantities $\eta_s^t(L,U,\alpha)$ and $\zeta_s^t(L,U,\alpha)$, for $s = 1,2,\cdots,\lfloor L \rfloor - 1$, $t = 0,1,2,\cdots,\lfloor L \rfloor - 1$ and $L \geq 2$, that can be computed iteratively as follows:

$$\eta_s^t(L,U,\alpha) := \frac{(N-1) + \frac{\Lambda}{\mu}(1-\alpha) + (L-t)\frac{\Lambda}{\mu}\eta_{s-1}^{t+1}(L,U,\alpha)}{\left[(U+N) + \frac{\Lambda}{\mu}(1-\alpha)\right]^2}$$

$$\zeta_s^t(L,U,\alpha) := \frac{(N-1) + \frac{\Lambda}{\mu}[1-\alpha + \frac{t+1}{(L+N-t-1)+\frac{\Lambda}{\mu}(1-\alpha)}] + (U-t)\frac{\Lambda}{\mu}\zeta_{s-1}^{t+1}(L,U,\alpha)}{[(L+N-t-1) + \frac{\Lambda}{\mu}(1-\alpha)][(L+N-t-2) + \frac{\Lambda}{\mu}(1-\alpha)]},$$

where we let $\eta_0^t(L,U,\alpha) := 0$ and $\zeta_0^t(L,U,\alpha) := \frac{\mu}{\Lambda}$. That is, $\eta_s^t$ and $\zeta_s^t$ are recursively defined starting with boundary values 0 and $\frac{\mu}{\Lambda}$ (recall that these two boundary values correspond to lower and upper bounds on $\Delta\alpha(K)$ per Lemma 2).

In the following lemma, we show that $\eta_s^t(L,U,\alpha)$ and $\zeta_s^t(L,U,\alpha)$ are respectively lower and upper bounds for $\Delta\alpha(K)$.

**Lemma 3.** *If $L \leq K(\alpha) \leq U$, then $\eta_s^0(L,U,\alpha) < \Delta\alpha\big(K(\alpha)\big) < \zeta_s^0(L,U,\alpha)$ and $\eta_s^1(L,U,\alpha) < \Delta\alpha\big(K(\alpha)-1\big) < \zeta_s^1(L,U,\alpha)$ for $1 \leq s \leq L-1$.*

The idea for the proof can be seen from the similarities between (12) and the definitions of $\eta_s^t(L,U,\alpha)$ and $\zeta_s^t(L,U,\alpha)$. By iterating (12) multiple times and replacing $\Delta\alpha(K-t)$ with $\eta_s^t(L,U,\alpha)$ and $\zeta_s^t(L,U,\alpha)$, we can prove by induction that $\eta_s^0(L,U,\alpha)$ and $\zeta_s^0(L,U,\alpha)$ are lower and upper bounds of $\Delta\alpha(K(\alpha))$. With the bounds on $\Delta\alpha(K(\alpha))$ and $\Delta\alpha(K(\alpha)-1)$, we can obtain bounds on $K(\alpha)$ by (9).

**Proposition 2.** *If $L \leq K(\alpha) \leq U$, then*

$$K(\alpha) > \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\Lambda}{\mu}\frac{\alpha}{1-\alpha}\eta_s^0(L,U,\alpha), \textit{ and}$$
$$K(\alpha) < \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\Lambda}{\mu}\frac{\alpha}{1-\alpha}\zeta_s^1(L,U,\alpha) + 1$$

*for any $s \geq 1$ such that $(L+N-s-1) + \frac{\Lambda}{\mu}(1-\alpha) > 0$.*

The above proposition shows that, starting with any pair of lower and upper bounds, successive bounds can be achieved. It immediately follows that, by applying Proposition 2 to the bounds $(L_0, U_0)$ in Proposition 1, we can obtain additional pairs of lower and upper bounds. In this case, the bounds obtained are successively tighter (for the upper bound, this is subject to a maximum value on the number of iterations).

**Corollary 1.** *For $s = 1, \ldots, L_0 - 1$, $L_s < K(\alpha) < U_s$, where*

$$L_s := \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\Lambda}{\mu}\frac{\alpha}{1-\alpha}\eta_s^0(L_0, U_0, \alpha), \text{ and} \tag{13}$$

$$U_s := \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\Lambda}{\mu}\frac{\alpha}{1-\alpha}\zeta_s^1(L_0, U_0, \alpha) + 1. \tag{14}$$

*Moreover, $L_s \geq L_{s-1}$ for any $s > 0$; and $U_s \leq U_{s-1}$ for $0 < s < \frac{1}{6}\frac{N-1}{1-\alpha}$ if $N \geq 4$.*

Note that while the lower bounds always improve with each iteration, the upper bounds do as long as $s < \frac{1}{6}\frac{N-1}{1-\alpha}$ (this threshold on $s$ increases in $N$ and $\alpha$); see the Appendix for further discussion.

## 5 Asymptotic Analysis and Approximations

In this section, we examine the asymptotic performance of the lower and upper bounds $(L_s, U_s)$ identified in the previous section and consider the extent to which they can be used as a basis for constructing approximations for the optimal number of vehicles $K(\alpha)$.

The following proposition shows that the bounds in Corollary 1 converge to $K(\alpha)$ under several asymptotic regimes.

**Proposition 3.** *For any pair of bounds, $(L_s, U_s)$, where $s \geq 1$, specified in Corollary 1, the following holds.*

1. *For fixed $N$, $\lim_{\Lambda \to 0^+}(U_s - L_s) = 1$, $\lim_{\Lambda \to 0^+} L_s = (N-1)\frac{\alpha}{1-\alpha}$, and $\lim_{\Lambda \to 0^+} U_s = (N-1)\frac{\alpha}{1-\alpha} + 1$.*
2. *For fixed $N$, $\lim_{\Lambda \to \infty}(U_s - L_s) = 1 + \frac{\alpha^{s+1}}{1-\alpha}$, $\lim_{s \to \infty}\left\{\lim_{\Lambda \to \infty}(U_s - L_s)\right\} = 1$,*

$$\lim_{\Lambda \to \infty}\left\{\frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + (1-\alpha^s)\frac{\alpha}{1-\alpha} - L_s\right\} = 0,$$

   *and*

$$\lim_{\Lambda \to \infty}\left\{\frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\alpha}{1-\alpha} + 1 - U_s\right\} = 0.$$

3. *For fixed $\Lambda$, $\lim_{N \to \infty}(U_s - L_s) = 1$,*

$$\lim_{N \to \infty}\left\{\frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} - L_s\right\} = 0,$$

   *and*

$$\lim_{N \to \infty}\left\{\frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + 1 - U_s\right\} = 0.$$

4. *Let* $\Lambda = N\lambda$. *For fixed* $\lambda$, $\lim_{N\to\infty}(U_s - L_s) = 1 + \left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{(1-\alpha)}+\frac{\lambda}{\mu}}\right)^s$, $\lim_{s\to\infty}\left\{\lim_{\Lambda\to\infty}(U_s - L_s)\right\} = 1$, *and*

$$\lim_{N\to\infty}\left\{\frac{N\lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\alpha}{1-\alpha}\frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{(1-\alpha)} + \frac{\lambda}{\mu}(1-\alpha)}\left[1 - \left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{(1-\alpha)} + \frac{\lambda}{\mu}}\right)^s\right] - L_s\right\} = 0,$$

$$\lim_{N\to\infty}\left\{\frac{N\lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\alpha}{1-\alpha}\left[\frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{(1-\alpha)} + \frac{\lambda}{\mu}(1-\alpha)} + \frac{\frac{1}{1-\alpha}}{\frac{1}{(1-\alpha)} + \frac{\lambda}{\mu}(1-\alpha)}\left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{(1-\alpha)} + \frac{\lambda}{\mu}}\right)^s\right] + 1 - U_s\right\} = 0.$$

5. *For fixed* $\Lambda$ *and* $N > 1$, $\lim_{\alpha\to 1}(U_s - L_s) = 1$, $s \geq 2$,

$$\lim_{\alpha\to 1}\left\{\frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} - L_s\right\} = 0, \quad s \geq 1,$$

$$\lim_{\alpha\to 1}\left\{\frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + 1 - U_s\right\} = 0, \quad s \geq 2, \text{ and}$$

$$\lim_{\alpha\to 1}\left\{\frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\Lambda}{\mu}\frac{N}{(N-1)^2} + 1 - U_1\right\} = 0.$$

Proposition 3 shows that the lower and upper bounds become exact (i.e., converge to the exact optimal value of $K(\alpha)$ under several asymptotic regimes), including when the demand $\Lambda$ approaches 0, the number of locations $N$ approaches infinity (for fixed $\Lambda$, corresponding to a setting where an increase in $N$ corresponds to an increase in the density of pick up and drop off locations), and the service level $\alpha$ approaches 1. When demand $\Lambda$ approaches infinity (for fixed $N$), the gap between the lower and upper bound approaches $1 + \frac{\alpha^{s+1}}{1-\alpha}$ which is decreasing in $s$ and approaches 1 as $s$ approaches infinity. When the number of locations $N$ approaches infinity (for fixed $\lambda$, corresponding to an increase in the service region), the gap between the lower and upper bound approaches $1 + \left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{(1-\alpha)}+\frac{\lambda}{\mu}}\right)^s$ which is decreasing in $s$ and approaches 1 as $s$ approaches infinity. These results suggest that these bounds possess properties similar to those of the exact value for $K(\alpha)$.

## 5.1 From Bounds to an Approximation

Motivated by the bounds, we propose next a closed form approximation, denoted by $\hat{K}(\alpha)$, for $K(\alpha)$ that falls between the lower and upper bounds $L_s$ and $U_s$ and satisfies the asymptotic properties of the lower and upper bounds, and other properties of $K(\alpha)$ per Proposition 3. First note that $L_s$ and $U_s$ share the first two terms (namely, $\frac{\Lambda}{\mu}\alpha$ and $(N-1)\frac{\alpha}{1-\alpha}$). Hence, any approximation reduces to approximating the third term. This term would ideally, per the results in Proposition 3, (i) converge to 0 as $\Lambda \to 0^+$, $\alpha \to 0$, and $N \to \infty$ for fixed $\Lambda$, (ii) converge to $\frac{\alpha}{1-\alpha}$ as $\Lambda \to \infty$ for fixed $N$, and (iii) converge to $\frac{\alpha}{1-\alpha}\frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{(1-\alpha)}+\frac{\lambda}{\mu}(1-\alpha)} = \frac{\alpha}{1-\alpha}\frac{\frac{\Lambda}{N\mu}(1-\alpha)}{\frac{1}{(1-\alpha)}+\frac{\Lambda}{N\mu}(1-\alpha)}$ as $N \to \infty$ where $\Lambda = N\lambda$ for fixed $\lambda$. An expression that satisfies these properties is given by $\frac{\frac{\Lambda}{\mu}\alpha}{\frac{N}{(1-\alpha)}+\frac{\Lambda}{\mu}(1-\alpha)}$.

**Proposition 4.** *For $s = 0, \dots, L_0 - 1$ (i.e., all indices with which $(L_s, U_s)$ is well-defined in Corollary 1), $L_s < \hat{K}(\alpha) < U_s - 1$ where*

$$\hat{K}(\alpha) = \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\frac{\Lambda}{\mu}}{\frac{N}{1-\alpha} + \frac{\Lambda}{\mu}(1-\alpha)}\alpha. \tag{15}$$

Note that, by Propositions 3 and 4, (15) is exact under several asymptotic regimes, including when the demand approaches 0 or approaches infinity, the number of locations approaches infinity for a fixed $\Lambda$ or for a fixed demand density per location, and the service level $\alpha$ approaches 1.

Moreover, we can show that $\hat{K}(\alpha)$ possesses several important properties that are satisfied by the exact value $K(\alpha)$. In the following lemma, we state these properties (which are of independent interest). For clarity, we use the notation $K(\alpha, \Lambda, \mu, N)$ and $\hat{K}(\alpha, \Lambda, \mu, N)$ to indicate the dependence of $K(\alpha)$ and $\hat{K}(\alpha)$ on $\Lambda$, $\mu$, and $N$ and state compactly results regarding the impact of these parameters.

**Lemma 4.** *The following properties are satisfied by $K(\alpha, \Lambda, \mu, N)$.*

*(1) $K(\alpha, C\Lambda, C\mu, N) = K(\alpha, \Lambda, \mu, N)$ for any $C > 0$.*

*(2) $K(\alpha, \Lambda, \mu, N)$ increases in $N$.*

*(3) $\alpha(NK, N\lambda, \mu, N) \le \alpha(K, \lambda, \mu, 1)$. That is, $K(\alpha, N\lambda, \mu, N) \ge NK(\alpha, \lambda, \mu, 1)$ if $K(\alpha, \Lambda, \mu, N)$ is defined as a continuous inverse of the function $\alpha(K, \Lambda, \mu, N)$.*

*(4) $\lim_{\lambda \to \infty} \{K(\alpha, N\lambda, \mu, N) - NK(\alpha, \lambda, \mu, 1)\} = 0$ if $K(\alpha, \Lambda, \mu, N)$ is defined as a continuous inverse of the function $\alpha(K, \Lambda, \mu, N)$.*

Property 1 states $K(\alpha, \Lambda, \mu, N)$ depends on the parameters $\Lambda$ and $\mu$ only through their ratio $\frac{\Lambda}{\mu}$. Property 2 states that more vehicles are needed to guarantee the same service level if the number of locations increases even though $\Lambda$ and $\mu$ remain unchanged. Property 3 states that an $N$-location system requires more vehicles than $N$ independent single-location systems if their per location demand and service level are the same. Property 4 states that the difference between an $N$-location system and $N$ independent single-location systems vanishes as $\lambda$ becomes very large.

**Proposition 5.** *The approximation $\hat{K}(\alpha, \Lambda, \mu, N)$ satisfies properties 1-4 in Lemma 4.*

To assess the performance of the approximation $\hat{K}(\alpha, \Lambda, \mu, N)$, we conducted extensive numerical experiments to compare $K(\alpha, \Lambda, \mu, N)$ and $\hat{K}(\alpha, \Lambda, \mu, N)$. In particular, we varied $N = 2, \cdots, 100$, $\Lambda = 1i$, $i = 1, \cdots, 1000$, and $\alpha = 0.03j$, $j = 1, \cdots, 33$ for fixed $\mu = 1$. This constituted $3,267,000$ cases. The range of the parameter values associated with these cases is chosen so that the starting value for each parameter is equal or nearly equal to the smallest possible value and the largest value yields results that are consistent with those obtained under the relevant asymptotic regimes (i.e., values that are large enough to lead to results similar to those obtained when these values reach their limit). For each case, we recorded the difference between $K(\alpha)$ and $\lceil\hat{K}\rceil$, where the notation $\lceil\cdot\rceil$ refers to the integer ceiling of the argument, and the relative difference $\frac{K(\alpha)-\lceil\hat{K}\rceil}{K(\alpha)}$. We find that the value of the difference for all the cases considered to be always no smaller than 0 and less than 5 with a mean of 0.015 and the relative difference to be less than 33% with a mean of 0.056‰. The difference is observed to be larger when the following conditions simultaneously hold: $N$ is small, $\alpha$ is large, and $\Lambda$ is large but not too large (since, from Proposition 3, case 2, we know that the approximation is exact when $\Lambda \to \infty$ for given $N$ and $\alpha$). The relative difference is observed to be larger when $K(\alpha)$ is small (i.e., when both $N$ and $\Lambda$, or $\alpha$ are small). Representative results are shown in Table 1, illustrating how remarkably accurate the approximation is.

| Parameters | | $K(\alpha)$ | $\lceil\hat{K}\rceil$ | $\lceil L_0\rceil$ | $\lfloor U_0\rfloor$ | $K(\alpha)-\lceil\hat{K}\rceil$ | $\frac{K(\alpha)-\lceil\hat{K}\rceil}{K(\alpha)}$ |
|---|---|---|---|---|---|---|---|
| | $\Lambda$ | | | | | | |
| | 1 | 28 | 28 | 28 | 37 | 0 | 0.00% |
| $\alpha = 0.9$ | 10 | 37 | 37 | 37 | 46 | 0 | 0.00% |
| $N = 4$ | 100 | 120 | 119 | 117 | 123 | 1 | 0.83% |
| | 200 | 211 | 210 | 207 | 214 | 1 | 0.47% |
| | 1000 | 934 | 934 | 927 | 935 | 0 | 0.00% |
| | $N$ | | | | | | |
| | 2 | 39 | 38 | 36 | 46 | 1 | 2.56% |
| $\alpha = 0.9$ | 4 | 55 | 55 | 55 | 64 | 0 | 0.00% |
| $\Lambda = 30$ | 8 | 91 | 91 | 91 | 100 | 0 | 0.00% |
| | 16 | 163 | 163 | 163 | 172 | 0 | 0.00% |
| | 32 | 307 | 307 | 307 | 316 | 0 | 0.00% |
| | $N$ | | | | | | |
| | 2 | 48 | 47 | 45 | 55 | 1 | 2.08% |
| $\alpha = 0.9$ | 4 | 101 | 101 | 99 | 109 | 0 | 0.00% |
| $\Lambda = 20N$ | 8 | 209 | 209 | 207 | 217 | 0 | 0.00% |
| | 16 | 425 | 425 | 423 | 433 | 0 | 0.00% |
| | 32 | 857 | 857 | 855 | 865 | 0 | 0.00% |
| | $\alpha$ | | | | | | |
| | 0.03 | 2 | 2 | 2 | 2 | 0 | 0.00% |
| $N = 4$ | 0.30 | 14 | 14 | 14 | 14 | 0 | 0.00% |
| $\Lambda = 40$ | 0.60 | 30 | 30 | 29 | 31 | 0 | 0.00% |
| | 0.90 | 65 | 64 | 64 | 73 | 1 | 1.54% |
| | 0.99 | 337 | 337 | 337 | 436 | 0 | 0.00% |

Table 1: Representative numerical results

## 5.2 Case of a Single Location

The case of a single location ($N = 1$) with exponentially distributed rental times corresponds to the well studied Erlang loss system, with vehicles returning to the same location upon completing service. Hence, it is worthwhile (though the single location is not the focus of this paper) to compare the approximation in (15) to the approximations obtained in the literature that consider Erlang loss systems. In Appendix D.2, we list notable results for the approximation of the inverse *Erlang Loss formula*, which include (5), (6) and (21) in Berezner et al. (1998), and (35), (36), (39), and (40) in Harel (2010). For ease of reference, we rewrite these approximations in the appendix using our notation. We add the initial "B" to equation numbers when we refer to expressions from Berezner et al. (1998), and the initial "H" when we refer to expressions from Harel (2010). Specifically, the maximum of (B.6) and (B.21) (referred to as (B.621) in our paper) and (H.36) are lower bounds for $K(\alpha, \Lambda, \mu, 1)$, and (B.5) and (H.35) are upper bounds. To the best of our knowledge, (H.39) and (H.40) are among the best performing approximations found in the literature. Note that the upper bound (B.5) and lower bound (B.6) are just special cases of our simple bounds $U_0$ and $L_0$ for $N = 1$. Except for these two, the other approximations, especially (B.21), do not have a simple form, which makes them less tractable (e.g., if used as part of a larger model) and more difficult to interpret (see the discussion in Section 6).

To further assess the effectiveness of our approximation, we carried out extensive numerical experiments to compare its performance to that of (39) and (40) in Harel (2010). Fixing $\mu = 1$, we varied $\Lambda$ ($\Lambda = 1i$, $i = 1, \cdots, 1000$) and $\alpha$ ($\alpha = 0.03j$, $j = 1, \cdots, 33$). For the cases tested, the mean absolute gaps between $K(\alpha, \Lambda, \mu, 1)$ and $\hat{K}(\alpha, \Lambda, \mu, 1)$, (B.621), (B.5), (H.39), (H.40) are 1.51, 2.55, 3.10, 2.06, and 0.64, respectively. That is, the approximation error of approximation (15) is among the best ones for Erlang loss systems. Interested readers are referred to Appendix D.2 for further numerical comparisons using the same examples considered by Berezner et al. (1998) and Harel (2010). Hence, a secondary contribution of the paper is to the literature on Erlang loss systems.

A potential limitation of (15), as an approximation for the case of $N = 1$, is that it does not satisfy the result in regime 5 ($\alpha \to 1$) of Proposition 3. That is, when $N = 1$, $\hat{K}(\alpha, \Lambda, \mu, 1)$ does not approach infinity as $\alpha \to 1$ (note that $\hat{K}(\alpha, \Lambda, \mu, 1)$ still satisfies all of the other results of Proposition 3). However, as implied by the following proposition, this issue arises only when $\alpha$ is nearly 1.

**Proposition 6.** *When* $N = 1$, $\lim_{K \to \infty} \left\{[1 - \alpha(K)](K!)(\frac{\Lambda}{\mu})^{-K}\right\} \in (0, 1)$. *That is,* $\frac{1}{1-\alpha(K)}$ *grows to infinity in the order of* $O\big((K!)(\frac{\Lambda}{\mu})^{-K}\big)$ *as* $K \to \infty$. *In contrast, when* $N > 1$, $\lim_{K \to \infty} \left\{[1 - \alpha(K)]\frac{K}{N-1}\right\} = 1$, *i.e.,* $\frac{1}{1-\alpha(K)}$ *grows to infinity in the order of* $O(K)$ *as* $K \to \infty$.[4]

[4] Note that the exact-order asymptotics we show in the proposition holds not only for the true $K(\alpha)$ but also for the approximation $\hat{K}(\alpha)$. The second statement (for $N > 1$) of the proposition provides an exact-order asymptotics that is tighter than the lower bound result in Lemma 20 of Banerjee et al. (2017) who show that the service level is at least $1 - O(\frac{1}{\sqrt{K}})$ as $K \to \infty$. Moreover, it is consistent with the asymptotic approximation, $\alpha(K) \approx \left(1 - \frac{1}{K}\right)^{N-1}$, implied by (14) in George et al. (2012) (a Taylor expansion of their result yields $\alpha(K) = 1 - \frac{N-1}{K} + o(\frac{1}{K})$ as $K \to \infty$).

Note that $(K!)(\frac{\Lambda}{\mu})^{-K}$ increases with $K$ at a very fast rate (faster than any exponential function), and $\frac{K}{N-1}$ linearly increases in $K$. Because $K(\alpha, \Lambda, \mu, N)$ is in fact the inverse function of $\alpha(K)$, the above proposition implies that the order in which $K(\alpha, \Lambda, \mu, N)$ grows to infinity as $\alpha \to 1$ is fundamentally different for $N = 1$ and $N > 1$. When $N > 1$, $\lim_{\alpha \to 1} K(\alpha, \Lambda, \mu, N)$ grows to infinity in the order of $O(\frac{1}{1-\alpha})$. When $N = 1$, $\lim_{\alpha \to 1} K(\alpha, \Lambda, \mu, 1)$ grows to infinity at a very slow rate. This rate is so negligible that it is smaller than that of any power series of order $O((1-\alpha)^{-\delta})$, $\delta > 0$. In Appendix D.1, we describe a correction term that could be added if necessary so that the resulting approximation for the optimal fleet size approaches infinity as $\alpha \to 1$.

# 6 Insights

In order to obtain insights into the determinants of minimal fleet size, it is useful to consider the benchmark case of a network where vehicles always return to the location from which they were picked up. That is, there is no vehicle roaming and the network can be viewed as consisting of $N$ independent queues. To isolate the impact of roaming[5], consider the case of a symmetric system, where the demand and service rates at each location are given by $\lambda_i = \frac{\Lambda}{N}$ and $\mu_{ij} = \mu$ for all $i, j = 1, \cdots, N$. The problem in this case decomposes into $N$ independent subsystems, with each subsystem consisting of a single location, with the optimal fleet size given by:

$$N\hat{K}(\alpha, \frac{\Lambda}{N}, \mu, 1) = \frac{\Lambda}{\mu}\alpha + B_0, \tag{16}$$

where $B_0 = \frac{N\frac{\Lambda}{\mu}}{\frac{N}{1-\alpha} + \frac{\Lambda}{\mu}(1-\alpha)}\alpha$. The above expression consists of two terms: the nominal load $\frac{\Lambda}{\mu}\alpha$ and additional buffer capacity $B_0$. This additional capacity is protection against the randomness in demand and service times.

Contrast equation (16) with equation (15) for the original system with vehicle roaming which can be rewritten as follows:

$$\hat{K}(\alpha, \Lambda, \mu, N) = \frac{\Lambda}{\mu}\alpha + B_0 + (N-1)\frac{\alpha}{1-\alpha} - B_0(1 - \frac{1}{N}). \tag{17}$$

The following observations can be made.

- Buffer capacity in a system with roaming can be viewed as consisting of three terms which can be interpreted as follows: *standard* buffer capacity that is protection against randomness in demand and service times, $B_0$, buffer capacity that is protection against roaming, $(N-1)\frac{\alpha}{1-\alpha}$, and a *correction* term, $-B_0(1 - \frac{1}{N})$ (we refer to this term as a "correction" term because it is negative).
- We attribute the second term to vehicle roaming since it reduces to zero when $N = 1$ (more about this below). However, because this additional capacity can also buffer against randomness from demand

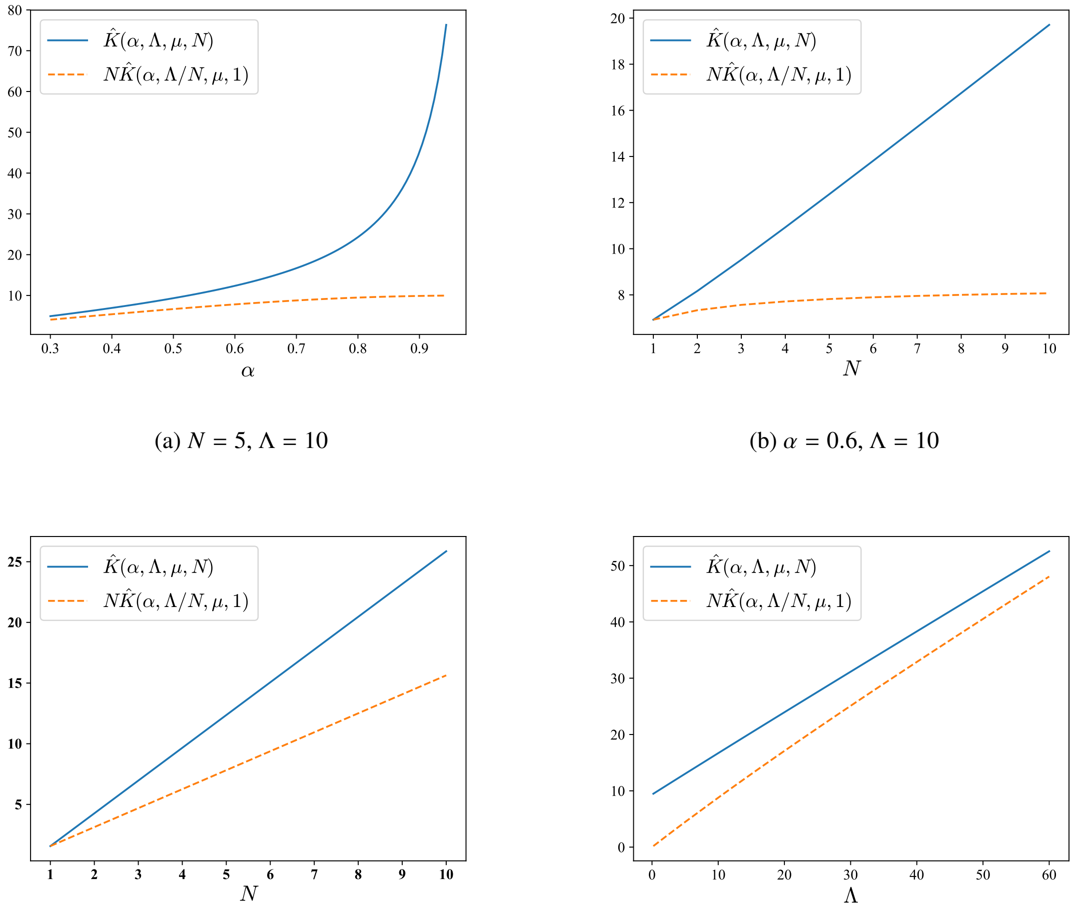
[5] Formally, we use the term a "vehicle network with roaming" to refer to an $N$-location network with a routing probability matrix with components $p_{ij} \geq 0$, $i, j \in \{1, \cdots, N\}$ that is irreducible.

and service times, the overall need for buffer capacity diminishes and there is a need for the correction term $-B_0(1-\frac{1}{N})$.

- The magnitude of the correction term (i.e., $B_0(1-\frac{1}{N})$) increases with the number of locations. In the limit, as $N \longrightarrow \infty$, the correction term approaches $B_0$. That is, the standard buffer capacity is no longer needed, with the buffer capacity $(N-1)\frac{\alpha}{1-\alpha}$ sufficient to protect against both vehicle roaming and randomness in demand and service times. The optimal fleet size then reduces to $\frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha}$.
- The standard buffer capacity term $B_0$ increases in the ratio of demand rate to service rate, $\frac{\Lambda}{\mu}$, and the service level, $\alpha$, and increases in the number of locations, $N$. The buffer capacity term, $(N-1)\frac{\alpha}{1-\alpha}$, is independent of the demand and service rates and increases in the service level and the number of locations.
- The fact that the term $(N-1)\frac{\alpha}{1-\alpha}$ is independent of demand means that even for vanishingly small demand (or service times), buffer capacity can be substantial (for example, a system with 2 locations and a service level of 0.95, the minimal fleet size is 20 vehicles when $\Lambda \longrightarrow 0$; contrast that with the minimal fleet size of 2 for the benchmark case where vehicles always return to the location from which they originated).
- A system without roaming requires fewer vehicles than a system with roaming, with the difference given by $\Delta_{\hat{K}} = (N-1)(\frac{\alpha}{1-\alpha} - \frac{B_0}{N})$. This difference increases in $N$ and $\alpha$, with $\Delta_{\hat{K}} \longrightarrow \infty$ as $N \longrightarrow \infty$ or $\alpha \longrightarrow 1$, and decreases in $\frac{\Lambda}{\mu}$, with $\Delta_{\hat{K}} \longrightarrow (N-1)\frac{\alpha}{1-\alpha}$ as $\frac{\Lambda}{\mu} \longrightarrow 0$ and $\Delta_{\hat{K}} \longrightarrow 0$ as $\frac{\Lambda}{\mu} \longrightarrow \infty$; see Figure 1 for a numerical illustration of these effects.
- The impact of an increase in the service level is different in the two systems. In a system with roaming, the minimal fleet size increases at a rate that is of order $O(\frac{1}{1-\alpha})$. In contrast, in a system without roaming, the minimal fleet size increases at a rate of an order that is smaller than any power series $O((1-\alpha)^{-\delta})$, $\delta > 0$, per Proposition 6 in Section 5.2. Figure 1a provides a numerical illustration of this result.

An intuitive explanation for why vehicle roaming requires additional capacity buffering and why the buffer capacity term $(N-1)\frac{\alpha}{1-\alpha}$ may be attributed to this roaming feature is that "roaming," even for balanced systems, creates short term unbalances in the distribution of vehicles across locations, requiring the additional buffering. Put a different way, vehicle roaming produces randomness in the service capacity available at each location. This feature is present in any irreducible closed queueing network (i.e., a network where the probability routing matrix $[p_{ij}]$ is irreducible). This intuition is supported by considering the following three examples.

**Example 1:** Consider a system where vehicles picked up at one location are always returned to the same location (i.e., $p_{ii} = 1$ for $i = 1, \ldots, N$) – the benchmark case we have considered so far. In this case, vehicles



(a) $N = 5, \Lambda = 10$

(b) $\alpha = 0.6, \Lambda = 10$

(c) $\alpha = 0.6, \Lambda = 2N$

(d) $\alpha = 0.7, N = 5$

Figure 1: Minimal Fleet Size With and Without Vehicle Roaming ($\mu = 1$)

do not roam, the routing matrix is not irreducible, and the service capacity available at each location is constant. In this case, and per equation (17), the term $(N-1)\frac{\alpha}{1-\alpha}$ is no longer part of the buffer capacity.

**Example 2:** Consider a system where vehicles picked up at location $i$ are always returned to location $i+1$ for $i = 1,\ldots,N-1$) and vehicles picked up at location $N$ are returned to location 1 (i.e., $p_{i,i+1} = 1$ for $i = 1,\ldots,N-1$ and $p_{N,1} = 1$). That is, one could view pick up locations arranged on a circular road. In such a network, there is no randomness in the routing of vehicles. However, vehicles still roam, the routing matrix is irreducible, and the service capacity available at each location is random. This system, assuming the demand arrival rate is the same at each location, satisfies the requirement of a balanced and irreducible network. Hence, buffer capacity, per equation (15), contains the term $(N-1)\frac{\alpha}{1-\alpha}$.

**Example 3(a):** Consider a balanced system where trips durations are zero (i.e. $\frac{1}{\mu} = 0$), per the setup described in Waserhole and Jost (2016). In such a network, there is no randomness in trip durations.

However, vehicles still roam, the routing matrix is irreducible, and the service capacity available at each location is random. Hence, buffer capacity, per equation (15) contains the term $(N-1)\frac{\alpha}{1-\alpha}$.

**Example 3(b):** Consider the same setup as in Example 3(a), except that the inter-arrival times of customers at each location is now deterministic and $p_{ij} = \frac{1}{N}$. In other words, there is no randomness in demand. However, vehicles still roam, the routing matrix is irreducible, and the service capacity available at each location is random. In this case, even though the demand process at the different locations is not Poisson, we can prove that the optimal buffer size is the same as that of Example 3(a) and contains the term $(N-1)\frac{\alpha}{1-\alpha}$.

Example 1 illustrates that absent the roaming feature, the term $(N-1)\frac{\alpha}{1-\alpha}$ disappears. Examples 2, 3(a), and 3(b) illustrate, respectively, that this term is present even if there is no randomness in vehicles routing and even if there no uncertainty in service times and customer inter-arrival times.

# 7 Unbalanced Networks

In this section (and the accompanying appendix), we discuss how our analysis and results can be extended to the case of unbalanced networks where the condition $\lambda_i = \sum_j \lambda_j p_{ji}$ may not hold for all locations. In an unbalanced network, the problem of determining a fleet size that guarantees a specified service level at each location may not have a feasible solution (George and Xia, 2011). That is, even with an infinitely large number of vehicles, it may not be possible to achieve a target service level (if this target is sufficiently high) at each location. In particular, per Proposition 7 in the Appendix, the service levels at some locations (so-called non-bottleneck locations) are bounded by specific thresholds. The average numbers of vehicles at these locations and in transit are also bounded by finite fixed thresholds (no matter how large is the total number of vehicles) and so are the associated throughputs.

In view of these challenges, it is common practice for operators of vehicle sharing systems to periodically reposition vehicles to reduce the degree of unbalance. In what follows, we show how we can account for such vehicle repositioning and how vehicle repositioning impacts fleet sizing. We also consider the problem of optimal vehicle repositioning when repositioning is costly. We adopt the approach of He et al. (2017) which models the process of vehicle repositioning as one where the requests for vehicle repositioning are issued continuously over time to move a vehicle from one location to another. If a vehicle is present at that location, then the move is carried out. Otherwise, the request is ignored. The process by which repositioning requests are sent to a location $i$ is modeled as a Poisson process with rate $\psi_i \geq 0$. Vehicles are repositioned from location $i$ to location $j$ according to transition probabilities $q_{ij}$ with repositioning time having a mean $\frac{1}{\theta_{ij}}$. In other words, repositioning requests function as "virtual" demand for rentals at a particular location and are governed by similar dynamics as those of actual rentals. Treating vehicles repositioning as an uncontrolled process is of course an approximation of how repositioning may occur in practice. However, for the purpose of making decision over long time scales (e.g., deciding on the size of the fleet) such an approximation can be reasonable (see He et al. (2017) for further discussion and justifications).

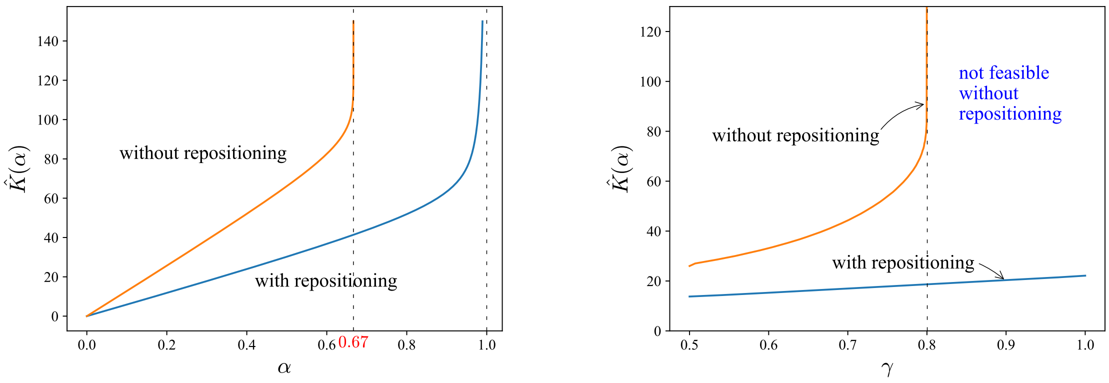
Given the above specification, a vehicle repositioning policy can be fully characterized by the vector of rates $\{\psi_i \geq 0 \mid i \in V\}$ and the matrix of routing probabilities $\{q_{ij} \geq 0 \mid i, j \in V, \sum_j q_{ij} = 1\}$. The network, in the presence of repositioning[6], is balanced if

$$\lambda_i + \psi_i = \sum_j (\lambda_j p_{ji} + \psi_j q_{ji}), \quad \text{for all } i \in V. \tag{18}$$

Let $\Psi := \sum_{i \in V} \psi_i$ denote the total repositioning rate, and $\frac{1}{\theta} := \sum_{i,j \in V} \frac{\psi_i q_{ij}}{\Psi \theta_{ij}}$ denote the average duration of repositioning trips. Then, the total demand rate for rentals (actual+virtual) is given by $\Lambda + \Psi$ and the average trip duration (averaged over both actual rentals and repositioning trips) is given by $\frac{\Lambda}{\Lambda+\Psi}\frac{1}{\mu} + \frac{\Psi}{\Lambda+\Psi}\frac{1}{\theta}$. Substituting $\Lambda + \Psi$ and $\frac{\Lambda}{\Lambda+\Psi}\frac{1}{\mu} + \frac{\Psi}{\Lambda+\Psi}\frac{1}{\theta}$ for $\Lambda$ and $\frac{1}{\mu}$ in (15), we obtain the following modified version of the approximation of the minimal fleet size:

$$\hat{K}(\alpha, \Lambda, \mu, \Psi, \theta, N) = (\frac{\Lambda}{\mu} + \frac{\Psi}{\theta})\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\frac{\Lambda}{\mu} + \frac{\Psi}{\theta}}{\frac{N}{1-\alpha} + (\frac{\Lambda}{\mu} + \frac{\Psi}{\theta})(1-\alpha)}\alpha. \tag{19}$$

In Figure 2, we illustrate, for an example with two locations, the impact of repositioning on the minimal number of vehicles required to guarantee a service level at least $\alpha$ at every location. This impact is particularly significant when $\gamma$ or $\alpha$ is large. Noting that $\gamma$ is the probability that a trip terminates at location 1 and $1-\gamma$ is the probability that it terminates at location 2, the parameter $\gamma$ measures how unbalanced the system is. The larger the $\gamma$ for $\gamma > 0.5$, the larger the degree of unbalance.



(a) $\gamma = 0.6$

(b) $\alpha = 0.25$

Figure 2: Minimal Fleet Size With and Without Repositioning (for a Two-location Unbalanced Network with $\lambda_1 = \lambda_2 = 50$, $\mu_{ij} = 1$, $\theta_{ij} = 1.5\mu_{ij}$, $p_{11} = p_{21} = \gamma$, and $p_{12} = p_{22} = 1 - \gamma$)

Note that there are multiple ways network balancing can be achieved. In setting where repositioning is

[6] Note that the network continues to be irreducible.

costly and the cost depends on the origin and destination of the respositioned vehicle, the following problem can be used to determine optimal repositioning:

$$\min_{\psi_i, q_{ij}} \left\{ \alpha c_{ij} \sum_{i,j \in V} \frac{\psi_i q_{ij}}{\theta_{ij}} \right\},$$

subject to (18), where $c_{ij}$ refers to the cost per unit of time of repositioning a vehicle from location $i$ to location $j$. The problem can be easily solved by recognizing that it can be transformed into a linear programing problem (e.g., reformulating the problem in terms of variables $z_{ij}$, where $z_{ij} = \psi_i q_{ij}$ and using the fact that $\sum_j z_{ij} = \psi_i$).

## 8 Concluding Remarks

The setting we consider can be viewed as an instance of a class of service systems where servers do not necessarily return to their original queue upon service completion and may instead join other queues. It can be viewed as an instance of an even broader class of service systems where the number of servers at each queue is random (or varies according to some specified logic). In our case, this randomness is determined by the demand process and the travel patterns of customers. In other settings, this randomness may arise for different reasons, such as external shocks that affect server availability or servers being independent agents who decide when to work and for how long (or who act strategically by deciding, based on the state of the system, on which queue to join). The impact of this randomness on system performance appears to be less well understood than the impact of other types of randomness, and hence is worthy of further study. There are other ways to generalize the setting we consider, including allowing for limits on the number of vehicles at each location.

## Acknowledgement

The authors thank the department editor, the associate editor, and the three anonymous reviewers for their valuable comments and suggestions. The research of Saif Benjaafar is in part funded by the United States National Science Foundation [Award SCC-1831140]. The research of Shining Wu is supported in part by the Hong Kong Research Grants Council [Grants No. 25504517 and 15507919]. The research of H. Liu is in part supported by National Science Foundation of China [Grants 71731009, 72061127002] and Guangdong Basic and Applied Basic Research Foundation [Grant 2020A1515111131].

# References

Adelman, D. (2008). A simple algebraic approximation to the erlang loss system. *Operations Research Letters 36*(4), 484 – 491.

Afeche, P., Z. Liu, and C. Maglaras (2018). Ride-hailing networks with strategic drivers: The impact of platform control capabilities on performance. Working paper, University of Toronto, Toronto.

Banerjee, S., D. Freund, and T. Lykouris (2017). Pricing and optimization in shared vehicle systems: An approximation framework. Working paper, Cornell University, Ithaca, NY.

Baskett, F., K. M. Chandy, R. R. Muntz, and F. G. Palacios (1975). Open, closed, and mixed networks of queues with different classes of customers. *Journal of the ACM 22*(2), 248–260.

Bellos, I., M. Ferguson, and L. B. Toktay (2017). The car sharing economy: Interaction of business model choice and product line design. *Manufacturing & Service Operations Management 19*(2), 185–201.

Benjaafar, S., J.-Y. Ding, G. Kong, and T. Taylor (2020). Labor welfare in on-demand service platforms. *Manufacturing & Service Operations Management*, Forthcoming.

Benjaafar, S. and M. Hu (2020). Operations management in the age of the sharing economy: What is old and what is new? *Manufacturing & Service Operations Management 22*(1), 93–101.

Berezner, S. A., A. E. Krzesinski, and P. G. Taylor (1998). On the inverse of erlang's function. *Journal of Applied Probability 35*(1), 246–252.

Besbes, O., F. Castro, and I. Lobel (2019). Spatial capacity planning. Working paper, Columbia University, New York.

Besbes, O., F. Castro, and I. Lobel (2020). Surge pricing and its spatial supply response. *Management Science*, Forthcoming.

Bimpikis, K., O. Candogan, and D. Saban (2019). Spatial pricing in ride-sharing networks. *Operations Research 67*(3), 744–769.

Botta, R. F., C. M. Harris, and W. G. Marchal (1987). Characterizations of generalized hyperexponential distribution functions. *Communications in Statistics. Stochastic Models 3*(1), 115–148.

Braverman, A., J. G. Dai, X. Liu, and L. Ying (2019). Empty-car routing in ridesharing systems. *Operations Research 67*(5), 1437–1452.

Cachon, G. P., K. M. Daniels, and R. Lobel (2017). The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing & Service Operations Management 19*(3), 368–384.

Castillo, J. C., D. T. Knoepfle, and E. G. Weyl (2018). Surge pricing solves the wild goose chase. Working paper, Stanford University, Stanford, CA.

Cooper, R. B. (1981). *Introduction to queueing theory*. North Holland.

Dai, J. G. and S. He (2014). Queues in service systems: Customer abandonment and diffusion approximations. In *INFORMS TutORials in Operations Research*, Chapter 3, pp. 36–59.

Freund, D., S. G. Henderson, and D. B. Shmoys (2019). Bike sharing. In M. Hu (Ed.), *Sharing Economy: Making Supply Meet Demand*, Chapter 18, pp. 435–459. Cham, Switzerland: Springer.

Gans, N., G. Koole, and A. Mandelbaum (2003). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing & Service Operations Management 5*(2), 79–141.

George, D. K. and C. H. Xia (2011). Fleet-sizing and service availability for a vehicle rental system via closed queueing networks. *European Journal of Operational Research 211*(1), 198 – 207.

George, D. K., C. H. Xia, and M. S. Squillante (2012). Exact-order asymptotic analysis for closed queueing networks. *Journal of Applied Probability 49*(2), 503–520.

Halfin, S. and W. Whitt (1981). Heavy-traffic limits for queues with many exponential servers. *Operations Research 29*(3), 567–588.

Harel, A. (1988). Sharp bounds and simple approximations for the erlang delay and loss formulas. *Management Science 34*(8), 959–972.

Harel, A. (2010). Sharp and simple bounds for the erlang delay and loss formulae. *Queueing Systems 64*(2), 119–143.

He, L., H.-Y. Mak, and Y. Rong (2019). Operations management of vehicle sharing systems. In M. Hu (Ed.), *Sharing Economy: Making Supply Meet Demand*, Chapter 19, pp. 461–484. Cham, Switzerland: Springer.

He, L., H.-Y. Mak, Y. Rong, and Z.-J. M. Shen (2017). Service region design for urban electric vehicle sharing systems. *Manufacturing & Service Operations Management 19*(2), 309–327.

Hofri, M. and Y. Kogan (1994). Asymptotic analysis of product-form distributions related to large interconnection networks. *Theoretical Computer Science 125*(1), 61 – 90.

Hu, L. and Y. Liu (2016). Joint design of parking capacities and fleet size for one-way station-based carsharing systems with road congestion constraints. *Transportation Research Part B: Methodological 93*, 268 – 299.

Jagerman, D. L. (1974). Some properties of the erlang loss function. *Bell System Technical Journal 53*(3), 525–551.

Janssen, A. J. E. M., J. S. H. van Leeuwaarden, and B. Zwart (2008). Gaussian expansions and bounds for the poisson distribution applied to the erlang b formula. *Advances in Applied Probability 40*(1), 122–143.

Kogan, Y. (1992). Another approach to asymptotic expansions for large closed queueing networks. *Operations Research Letters 11*(5), 317 – 321.

Kogan, Y. and A. Birman (1992). Asymptotic analysis of closed queueing networks with bottlenecks. In *Performance of Distributed Systems and Integrated Communication Networks*, IFIP Transactions C: Communication Systems, pp. 265 – 280. Amsterdam: North-Holland.

Li, S., Q. Luo, and R. Hampshire (2019). Optimizing large on-demand transportation systems. Working paper, Northwestern University, Evanston.

Mandelbaum, A. and S. Zeltyn (2009). Staffing many-server queues with impatient customers: Constraint satisfaction in call centers. *Operations Research 57*(5), 1189–1205.

Reiser, M. and S. S. Lavenberg (1980). Mean-value analysis of closed multichain queuing networks. *Journal of ACM 27*(2), 313–322.

Taylor, T. A. (2018). On-demand service platforms. *Manufacturing & Service Operations Management 20*(4), 704–720.

Ward, A. R. (2012). Asymptotic analysis of queueing systems with reneging: A survey of results for FIFO, single class models. *Surveys in Operations Research and Management Science 17*(1), 1–14.

Waserhole, A. and V. Jost (2016). Pricing in vehicle sharing systems: optimization in queuing networks with product forms. *EURO Journal on Transportation and Logistics 5*(3), 293–320.

Whitt, W. (2002). *Stochastic-process limits: an introduction to stochastic-process limits and their application to queues*. Springer-Verlag New York.

Whitt, W. (2007). What you should know about queueing models to set staffing requirements in service systems. *Naval Research Logistics 54*(5), 476–484.

Zhang, R., F. Rossi, and M. Pavone (2019). Analysis, control, and evaluation of mobility-on-demand systems: A queueing-theoretical approach. *IEEE Transactions on Control of Network Systems 6*(1), 115–126.

# Online Appendices: Dimensioning On-Demand Vehicle Sharing Systems

Saif Benjaafar[†], Shining Wu[‡], Hanlin Liu[§], Einar Bjarki Gunnarsson[†]

[†]Department of Industrial & Systems Engineering, University of Minnesota, Twin Cities, USA

[‡]Department of Logistics & Maritime Studies, Hong Kong Polytechnic University, Hong Kong

[§]Division of Information Systems and Management Engineering, College of Business,
Southern University of Science and Technology, Shenzhen, China

saif@umn.edu, sn.wu@polyu.edu.hk, liuhl@sustech.edu.cn, gunna042@umn.edu

# Appendix A Proofs

In the section, we provide proofs for our main results.

To prove Lemma 2, it is sufficient to prove the following more general lemma.

**Lemma 5.** *(i)* $\Delta\alpha(K) \leq \frac{\mu}{N\mu+\Lambda} < \frac{\mu}{\Lambda}$ *holds for all $K$, and the first inequality is strict for $K > 1$.*

*(ii)* $\Delta\alpha(K) \leq \frac{\mu}{(K+N-1)\mu+\Lambda[1-\alpha(K-1)]}$ *holds for all $K$ and is strict for $K > 1$.*

*(iii)* $0 < \Delta\alpha(K) < \Delta\alpha(K-1)$. *That is, $\alpha(K)$ is increasing concave in $K$.*

*(iv)* $\Delta\alpha(K) > \frac{(N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}{\left\{(K+N)+\frac{\Lambda}{\mu}[1-\alpha(K)]\right\}^2}$.

**Proof of Lemma 5.** First, note that from (8), we can obtain

$$\Delta\alpha(K) = \frac{(N-1) + \frac{\Lambda}{\mu}[1-\alpha(K-1)] + K\frac{\Lambda}{\mu}[\Delta\alpha(K-1)]}{[(K+N-1) + \frac{\Lambda}{\mu}(1-\alpha(K-1))][(K+N-2) + \frac{\Lambda}{\mu}(1-\alpha(K-2))]},$$

which also corresponds to equation (12) in the main text. We will use this result in the course of this proof.

We prove the results (i)–(iv) in sequence.

(i) We prove $\Delta\alpha(K) \leq \frac{\mu}{N\mu+\Lambda} < \frac{\mu}{\Lambda}$ by induction. Note that the inequalities are shown to be strict for $K > 1$ in our proof.

- When $K = 1$, $\Delta\alpha(1) = \alpha(1) - \alpha(0) = \frac{\mu}{N\mu+\Lambda} < \frac{\mu}{\Lambda}$.
- Suppose that $\Delta\alpha(k) \leq \frac{\mu}{N\mu+\Lambda} < \frac{\mu}{\Lambda}$ holds for $k \leq K-1$, we then prove that the inequality strictly holds for $k = K$.

  By plugging $\alpha(K-1) - \alpha(K-2) < \frac{\mu}{\Lambda}$ into (12), we have

  $$\alpha(K) - \alpha(K-1) < \frac{1}{(K+N-2) + \frac{\Lambda}{\mu}(1-\alpha(K-2))} \leq \frac{\mu}{N\mu+\Lambda} < \frac{\mu}{\Lambda},$$

  where the second inequality holds because $\alpha(k) \leq \frac{k\mu}{\Lambda}$ for any $k \geq 0$.

(ii) Second, note that the inequality $\Delta\alpha(K) < \frac{\mu}{\Lambda}$ guarantees that $(K+N-1) + \frac{\Lambda}{\mu}[1-\alpha(K-1)]$ increases in $K$. Therefore,

$$\begin{aligned}
\alpha(K) - \alpha(K-1) &= \frac{K}{(K+N-1) + \frac{\Lambda}{\mu}(1-\alpha(K-1))} - \frac{K-1}{(K+N-2) + \frac{\Lambda}{\mu}(1-\alpha(K-2))} \\
&< \frac{K}{(K+N-1)\mu + \frac{\Lambda}{\mu}(1-\alpha(K-1))} - \frac{K-1}{(K+N-1) + \frac{\Lambda}{\mu}(1-\alpha(K-1))} \\
&= \frac{\mu}{(K+N-1)\mu + \Lambda(1-\alpha(K-1))},
\end{aligned}$$

for $K \geq 2$ where both $\alpha(K)$ and $\alpha(K-1)$ can be expressed by the recursive equation (8). Furthermore, it is obvious that the above inequality also holds for $K = 1$ since $\alpha(0) = 0$.

(iii) Next, we prove result (iii) by induction. On the one hand, by (12) and the fact that $\alpha(1) > \alpha(0)$, a simple induction proves $\alpha(K) > \alpha(K-1)$, $\forall K$. That is, $\Delta\alpha(K) > 0$. On the other hand, it is easy to verify $\Delta\alpha(2) < \Delta\alpha(1)$. Assume that $\Delta\alpha(k) < \Delta\alpha(k-1)$ holds for $k \leq K-1$, we then prove $\Delta\alpha(K) < \Delta\alpha(K-1)$. By (12), we know that $\Delta\alpha(K) < \Delta\alpha(K-1)$ holds if and only if

$$(N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)] < \Delta\alpha(K-1)\left\{\left[(K+N-1)+\frac{\Lambda}{\mu}(1-\alpha(K-1))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]-K\frac{\Lambda}{\mu}\right\}.$$

Note that

$$\begin{aligned}&\left[(K+N-1)+\frac{\Lambda}{\mu}(1-\alpha(K-1))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]-K\frac{\Lambda}{\mu}\\=&\left[(K+N-3)+\frac{\Lambda}{\mu}(1-\alpha(K-3))+2-\frac{\Lambda}{\mu}(\alpha(K-1)-\alpha(K-3))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]-K\frac{\Lambda}{\mu}.\end{aligned}$$

Thus, the condition is satisfied if

$$\begin{aligned}&\Delta\alpha(K-1)\\>&\frac{(N-1)+\frac{\Lambda}{\mu}(1-\alpha(K-1))+\Delta\alpha(K-1)\left\{K\frac{\Lambda}{\mu}-\left[2-\frac{\Lambda}{\mu}(\alpha(K-1)-\alpha(K-3))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]\right\}}{\left[(K+N-3)+\frac{\Lambda}{\mu}(1-\alpha(K-3))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]}\\=&\frac{(N-1)+\frac{\Lambda}{\mu}(1-\alpha(K-2))+\Delta\alpha(K-1)\left\{(K-1)\frac{\Lambda}{\mu}-\left[2-\frac{\Lambda}{\mu}(\alpha(K-1)-\alpha(K-3))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]\right\}}{\left[(K+N-3)+\frac{\Lambda}{\mu}(1-\alpha(K-3))\right]\left[(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))\right]},\end{aligned}$$

which holds according to (12) because $\Delta\alpha(K-2) > \Delta\alpha(K-1)$ by the induction assumption and that $[2-\frac{\Lambda}{\mu}(\alpha(K-1)-\alpha(K-3))] > 0$ by the result (i) in this proposition.

(iv) Lastly, according to the result (i), $(K+N)\mu + \Lambda[1-\alpha(K)]$ increases in $K$ because $\Delta\alpha(K) \leq \frac{\mu}{\Lambda}$. Therefore, both $\alpha(K)$ and $(K+N)\mu + \Lambda[1-\alpha(K)]$ increase in $K$, and hence equation (12) yields

$$\begin{aligned}\alpha(K)-\alpha(K-1) &> \frac{(N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}{[(K+N-1)+\frac{\Lambda}{\mu}(1-\alpha(K-1))][(K+N-2)+\frac{\Lambda}{\mu}(1-\alpha(K-2))]}\\&> \frac{(N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}{[(K+N)+\frac{\Lambda}{\mu}(1-\alpha(K))][(K+N)+\frac{\Lambda}{\mu}(1-\alpha(K))]}.\end{aligned}$$

□

**Proof of Proposition 1.** From (8), we have

$$K = \frac{\Lambda}{\mu}\alpha(K)+(N-1)\frac{\alpha(K)}{1-\alpha(K)}+\frac{\Lambda}{\mu}\frac{\alpha(K)}{1-\alpha(K)}\Delta\alpha(K)$$

Because $K(\alpha)$ is the smallest number that satisfies $\alpha(K) \geq \alpha$, we know $\alpha\big(K(\alpha)-1\big) < \alpha \leq \alpha\big(K(\alpha)\big)$. Therefore,

$$\begin{aligned} K(\alpha) &= \frac{\Lambda}{\mu}\alpha\big(K(\alpha)\big) + (N-1)\frac{\alpha\big(K(\alpha)\big)}{1-\alpha\big(K(\alpha)\big)} + \frac{\Lambda}{\mu}\frac{\alpha\big(K(\alpha)\big)}{1-\alpha\big(K(\alpha)\big)}\Delta\alpha\big(K(\alpha)\big) \\ &\geq \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\Lambda}{\mu}\frac{\alpha}{1-\alpha}\Delta\alpha\big(K(\alpha)\big), \qquad (20) \\ K(\alpha)-1 &= \frac{\Lambda}{\mu}\alpha\big(K(\alpha)-1\big) + (N-1)\frac{\alpha\big(K(\alpha)-1\big)}{1-\alpha\big(K(\alpha)-1\big)} + \frac{\Lambda}{\mu}\frac{\alpha\big(K(\alpha)-1\big)}{1-\alpha\big(K(\alpha)-1\big)}\Delta\alpha\big(K(\alpha)-1\big) \\ &< \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\Lambda}{\mu}\frac{\alpha}{1-\alpha}\Delta\alpha\big(K(\alpha)-1\big). \qquad (21) \end{aligned}$$

First, it is easy to see

$$K(\alpha) \geq \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha},$$

since $\Delta\alpha(K) > 0$. The inequality is strict if $K(\alpha) \geq 1$, i.e., $\alpha > 0$.

Second, by substituting $\Delta\alpha(K) < \frac{\mu}{\Lambda}$ into the right hand side of (21), we obtain

$$K(\alpha) < \frac{\Lambda}{\mu}\alpha + N\frac{\alpha}{1-\alpha} + 1.$$

□

Note that equation (12) also describes a recursive relationship between $\Delta\alpha(K)$ and $\Delta\alpha(K-1)$. A series of bounds on $\Delta\alpha(K)$ can be obtained if we expand (12) for multiple times and bound $\Delta\alpha(K-s)$ according to Lemma 5. We introduce the following result as a corollary of Lemma 5.

**Corollary 2.** *Let $\tilde{\eta}_0(K) := 0$ and $\tilde{\zeta}_0(K) := \frac{\mu}{(K+N-1)\mu+\Lambda(1-\alpha(K-1))}$ for all $K \geq 1$. Define $\tilde{\eta}_s(K)$ and $\tilde{\zeta}_s(K)$ iteratively for $s = 1, 2, \cdots, K-1$, where $K \geq 2$ as follows.*

$$\begin{aligned} \tilde{\eta}_s(K) &= \frac{(N-1) + \frac{\Lambda}{\mu}[1-\alpha(K-1)] + K\frac{\Lambda}{\mu}\tilde{\eta}_{s-1}(K-1)}{\{(K+N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]\}\{(K+N-2)+\frac{\Lambda}{\mu}[1-\alpha(K-2)]\}}, \\ \tilde{\zeta}_s(K) &= \frac{(N-1) + \frac{\Lambda}{\mu}[1-\alpha(K-1)] + K\frac{\Lambda}{\mu}\tilde{\zeta}_{s-1}(K-1)}{\{(K+N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]\}\{(K+N-2)+\frac{\Lambda}{\mu}[1-\alpha(K-2)]\}}. \end{aligned}$$

*Then, $\tilde{\eta}_s(K) < \Delta\alpha(K) \leq \tilde{\zeta}_s(K)$ for all $0 \leq s \leq K-1$. Furthermore, $\tilde{\eta}_s(K) > \tilde{\eta}_{s-1}(K)$ and $\tilde{\zeta}_s(K) < \tilde{\zeta}_{s-1}(K)$.*

**Proof of Corollary 2.** By iterating equation (12) multiple times on its right hand side, we are able to express $\Delta\alpha(K)$ by $\Delta\alpha(K-s)$, where the last iteration has terms $(N-1) + \frac{\Lambda}{\mu}[1-\alpha(K-s)] + (K-s+1)\frac{\Lambda}{\mu}\Delta\alpha(K-s)$ in its numerator. Noting that $(K-s)\Delta\alpha(K-s) \leq \alpha(K-s)$ by result (iii) of Lemma 5, we know $\frac{\Lambda}{\mu}[1-\alpha(K-s)] + (K-s+1)\frac{\Lambda}{\mu}\Delta\alpha(K-s) \leq \frac{\Lambda}{\mu} + \frac{\Lambda}{\mu}\Delta\alpha(K-s) \leq \frac{\Lambda}{\mu} + \tilde{\zeta}_0(K-s)$. By replacing the term $(K-s+1)\frac{\Lambda}{\mu}\Delta\alpha(K-s)$ with 0 on the right hand side of the iteration, we have $\Delta\alpha(K) > \tilde{\eta}_s(K)$. By replacing the term $\frac{\Delta}{\mu}[1-\alpha(K-s)]+(K-s+1)\frac{\Delta}{\mu}\Delta\alpha(K-s)$ with $\frac{\Delta}{\mu}+\tilde{\zeta}_0(K-s)$ on the right hand side of the iteration, we have $\Delta\alpha(K) \le \tilde{\zeta}_s(K)$.

According to Lemma 5, $\tilde{\eta}_1(K) > 0 = \tilde{\eta}_0(K)$ and

$$\begin{aligned}\tilde{\zeta}_1(K) &= \frac{(N-1)+\frac{\Delta}{\mu}[1-\alpha(K-1)]}{\{(K+N-1)+\frac{\Delta}{\mu}[1-\alpha(K-1)]\}\{(K+N-2)+\frac{\Delta}{\mu}[1-\alpha(K-2)]\}}\\ &< \frac{1}{\{(K+N-1)+\frac{\Delta}{\mu}[1-\alpha(K-1)]\}} = \tilde{\zeta}_0(K)\end{aligned}$$

for $K \ge 2$. By the definition of $\tilde{\eta}_s(K)$ and $\tilde{\zeta}_s(K)$, we can easily prove $\tilde{\eta}_s(K) > \tilde{\eta}_{s-1}(K)$ and $\tilde{\zeta}_s(K) < \tilde{\zeta}_{s-1}(K)$ by induction. □

Although $\tilde{\eta}_s(K)$ and $\tilde{\zeta}_s(K)$ bound $\Delta\alpha(K)$ in general, they are not useful bounds for $\Delta\alpha\big(K(\alpha)\big)$ because the value of $K(\alpha)$ is unknown. Thus, we replace the value $K$ in the definitions of $\tilde{\eta}_s(K)$ and $\tilde{\zeta}_s(K)$ with its upper bound $U$ and lower bound $L$ to yield $\eta_s^t(L,U,\alpha)$ and $\zeta_s^t(L,U,\alpha)$, whose expressions do not involve $K$ and hence are now practical bounds for $\Delta\alpha(K)$. Recall that the definitions of $\eta_s^t(L,U,\alpha)$ and $\zeta_s^t(L,U,\alpha)$ are presented in the main text. Here we prove the following proposition.

**Lemma 6.** *If $L \le K(\alpha) \le U$, then $\tilde{\eta}_s\big(K(\alpha)-t\big) \ge \eta_s^t(L,U,\alpha)$ and $\tilde{\zeta}_s\big(K(\alpha)-t\big) \le \zeta_s^t(L,U,\alpha)$, for all $t \ge 0$ and $0 \le s \le L-1$.*

**Proof of Lemma 6.** Note that $\eta_s^t(L,U,\alpha)$ and $\zeta_s^t(L,U,\alpha)$ are defined in similar ways as $\tilde{\eta}_s(K)$ and $\tilde{\zeta}_s(K)$ in Corollary 2 with minor differences in the numerators and denominators. While the denominator of $\tilde{\eta}_s(K)$ is updated according to $K$, that of $\eta_s^t(L,U,\alpha)$ depends only on $U$ and $\alpha$.

By the fact that $L \le K(\alpha) \le U$ and $\alpha\big(K(\alpha)-1\big) < \alpha \le \alpha\big(K(\alpha)\big)$, we prove the results by induction for $s = 0,1,2,\cdots,L-1$.

- First, $s = 0$. We know that $\tilde{\eta}_0\big(K(\alpha)-t\big) = \eta_0^t(L,U,\alpha) = 0$ and $\tilde{\zeta}_0\big(K(\alpha)-t\big) \le \zeta_0^t(L,U,\alpha)$, $\forall t \ge 0$ by result (i) of Lemma 5.

- Suppose that we have proved the result for indices up to $s-1$. Then, by the definition of $\tilde{\eta}_s$, we have

$$\begin{aligned}
\tilde{\eta}_s(K(\alpha)-t) =& \frac{(N-1)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-1\big)]+\big(K(\alpha)-t\big)\frac{\Delta}{\mu}\tilde{\eta}_{s-1}(K(\alpha)-t-1)}{\Big\{(K(\alpha)+N-t-1)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-1\big)]\Big\}\Big\{(K(\alpha)+N-t-2)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-2\big)]\Big\}} \\
>& \frac{(N-1)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-1\big)]+\big(K(\alpha)-t\big)\frac{\Delta}{\mu}\tilde{\eta}_{s-1}(K(\alpha)-t-1)}{\Big\{(K(\alpha)+N)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)\big)]\Big\}^2} \\
>& \frac{(N-1)+\frac{\Delta}{\mu}(1-\alpha)+(L-t)\frac{\Delta}{\mu}\tilde{\eta}_{s-1}(K(\alpha)-t-1)}{\Big[(U+N)+\frac{\Delta}{\mu}(1-\alpha)\Big]^2} \\
\geq& \frac{(N-1)+\frac{\Delta}{\mu}(1-\alpha)+(L-t)\frac{\Delta}{\mu}\eta_{s-1}^{t+1}(L,U,\alpha)}{\Big[(U+N)+\frac{\Delta}{\mu}(1-\alpha)\Big]^2} = \eta_s^t(L,U,\alpha),
\end{aligned}$$

where the first inequality holds because $K+\frac{\Delta}{\mu}[1-\alpha(K)]$ increases in $K$ (by result (i) of Lemma 5), the second inequality follows from the fact that $L \leq K(\alpha) \leq U$, $\alpha\big(K(\alpha)-1\big) < \alpha \leq \alpha\big(K(\alpha)\big)$, and the last inequality holds by the induction hypothesis.

On the other hand, by the definition of $\tilde{\zeta}_s$,

$$\begin{aligned}
\tilde{\zeta}_s(K(\alpha)-t) =& \frac{(N-1)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-1\big)]+\big(K(\alpha)-t\big)\frac{\Delta}{\mu}\tilde{\zeta}_{s-1}(K(\alpha)-t-1)}{\Big\{(K(\alpha)+N-t-1)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-1\big)]\Big\}\Big\{(K(\alpha)+N-t-2)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-2\big)]\Big\}} \\
<& \frac{(N-1)+\frac{\Delta}{\mu}\Big\{1-\alpha\big(K(\alpha)\big)+\frac{t+1}{K(\alpha)+N-t-1+\frac{\Delta}{\mu}\big[1-\alpha(K(\alpha)-t-1)\big]}\Big\}+\big(K(\alpha)-t\big)\frac{\Delta}{\mu}\tilde{\zeta}_{s-1}(K(\alpha)-t-1)}{\Big\{(K(\alpha)+N-t-1)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-1\big)]\Big\}\Big\{(K(\alpha)+N-t-2)+\frac{\Delta}{\mu}[1-\alpha\big(K(\alpha)-t-2\big)]\Big\}} \\
<& \frac{(N-1)+\frac{\Delta}{\mu}\big\{1-\alpha+\frac{t+1}{L+N-t-1+\frac{\Delta}{\mu}(1-\alpha)}\big\}+\big(U-t\big)\frac{\Delta}{\mu}\tilde{\zeta}_{s-1}(K(\alpha)-t-1)}{\Big\{(L+N-t-1)+\frac{\Delta}{\mu}(1-\alpha)\Big\}\Big\{(L+N-t-2)+\frac{\Delta}{\mu}(1-\alpha)\Big\}} \\
\leq& \frac{(N-1)+\frac{\Delta}{\mu}\big\{1-\alpha+\frac{t+1}{L+N-t-1+\frac{\Delta}{\mu}(1-\alpha)}\big\}+\big(U-t\big)\frac{\Delta}{\mu}\zeta_{s-1}^{t+1}(L,U,\alpha)}{\Big\{(L+N-t-1)+\frac{\Delta}{\mu}(1-\alpha)\Big\}\Big\{(L+N-t-2)+\frac{\Delta}{\mu}(1-\alpha)\Big\}} \\
=& \zeta_s^t(L,U,\alpha),
\end{aligned}$$

where the second inequality follows from the fact that $L \leq K(\alpha) \leq U$, $\alpha\big(K(\alpha)-1\big) < \alpha \leq \alpha\big(K(\alpha)\big)$ and the last inequality holds by the induction hypothesis. The first inequality holds because

$$\begin{aligned}
1-\alpha\big(K-t-1\big) = 1-\alpha(K)+\alpha(K)-\alpha(K-t-1) &\leq 1-\alpha(K)+(t+1)\Delta\alpha(K-t) \\
&\leq 1-\alpha(K)+(t+1)\frac{1}{K-t+N-1+\frac{\Delta}{\mu}[1-\alpha(K-t-1)]}
\end{aligned}$$

from results (ii) and (iii) of Lemma 5.

□

**Proof of Lemma 3.** The results follow directly from Corollary 2 and Lemma 6. □

**Proof of Proposition 2.** The results follow directly from Lemma 3 and inequalities (20) and (21). □

**Proof of Corollary 1.** By applying Proposition 2 to the bounds $(L_0, U_0)$ established in Proposition 1, we have $L_s < K(\alpha) < U_s$.

To prove $L_s \geq L_{s-1}$ and $U_s \leq U_{s-1}$, it is sufficient to show that $\eta_s^t(L_0, U_0, \alpha) > \eta_{s-1}^t(L_0, U_0, \alpha)$ and $\zeta_s^t(L_0, U_0, \alpha) > \zeta_{s-1}^t(L_0, U_0, \alpha)$ for $s = 0, 1, \cdots$.

For the monotonicity result of $\eta_s^t$, we can prove for any pair of positive bounds as argument of $\eta_s^t(L, U, \alpha)$. It is obvious that $\eta_1^t(L, U, \alpha) > 0 = \eta_0^t(L, U, \alpha)$, $\forall t$. According to the definition of $\eta_s^t(L, U, \alpha)$, we know $\eta_s^t(L, U, \alpha) > \eta_{s-1}^t(L, U, \alpha)$ for all $0 \leq s \leq L - 1$, which is an analogue to the result $\tilde{\eta}_s(K) > \tilde{\eta}_{s-1}(K)$ in Corollary 2.

For the monotonicity result of $\zeta_s^t$, we only prove the case with $(L_0, U_0)$ as argument and for certain $s$'s. Again, according to the definition of $\zeta_s^t(L, U, \alpha)$, it is sufficient to prove $\zeta_1^t(L_0, U_0, \alpha) < \zeta_0^t(L_0, U_0, \alpha)$. We then examine the condition that can guarantee this inequality. Note that

$$\zeta_1^t(L, U, \alpha) = \frac{(N-1) + \frac{\Lambda}{\mu}[1 - \alpha + \frac{t+1}{(L_0+N-t-1)+\frac{\Lambda}{\mu}(1-\alpha)}] + (U_0 - t)\frac{\Lambda}{\mu}\zeta_0^{t+1}(L_0, U_0, \alpha)}{[(L_0 + N - t - 1) + \frac{\Lambda}{\mu}(1 - \alpha)][(L_0 + N - t - 2) + \frac{\Lambda}{\mu}(1 - \alpha)]} < \frac{\mu}{\Lambda} = \zeta_0^t(K, L, \alpha)$$

$$\Leftrightarrow \frac{\Lambda}{\mu}(U_0 - L_0) + \frac{(\frac{\Lambda}{\mu})^2(t+1)}{(L_0 + N - t - 1) + \frac{\Lambda}{\mu}(1 - \alpha)} < [(L_0 + N - t - 1) + \frac{\Lambda}{\mu}(1 - \alpha)][(L_0 + N - t - 2) + \frac{\Lambda}{\mu}(1 - \alpha) - \frac{\Lambda}{\mu}].$$

By substituting the values of $L_0$ and $U_0$ into the above inequality, we know that the condition

$$\frac{\Lambda}{\mu}\left[\frac{1}{1-\alpha} + 1 + \frac{\frac{\Lambda}{\mu}(t+1)}{\frac{\Lambda}{\mu} + \frac{N-1}{1-\alpha} - t}\right] < \left(\frac{\Lambda}{\mu} + \frac{N-1}{1-\alpha} - t\right)\left(\frac{N-1}{1-\alpha} - t - 1\right)$$

is needed to guarantee $\zeta_1^t(L_0, U_0, \alpha) < \zeta_0^t(L_0, U_0, \alpha)$. If $t + 1 \leq \frac{1}{6}\frac{N-1}{1-\alpha}$, simple algebra shows that the left hand side is no greater than $\frac{\Lambda}{\mu}\frac{2+\frac{N-1}{6}}{1-\alpha} - \frac{\Lambda}{\mu}\frac{\alpha}{1-\alpha}$ and the right hand side is greater than $\frac{\Lambda}{\mu}\frac{5}{6}\frac{N-1}{1-\alpha} + (\frac{5}{6}\frac{N-1}{1-\alpha})^2$. When $N \geq 4$, the inequality holds regardless of the value of $\frac{\Lambda}{\mu}$ and $\alpha$ because $2 + \frac{N-1}{6} \leq \frac{5}{6}(N-1)$. Noting that $s + t$ is constant in the iteration, the condition $t + 1 \leq \frac{1}{6}\frac{N-1}{1-\alpha}$ requires $s + 1 \leq \frac{1}{6}\frac{N-1}{1-\alpha}$ as well. Therefore, $\zeta_s^t(L_0, U_0, \alpha) < \zeta_{s-1}^t(L_0, U_0, \alpha)$ and hence $U_s < U_{s-1}$ if $s < \frac{1}{6}\frac{N-1}{1-\alpha}$ and $N \geq 4$. □

As one can see, $\eta_s^t$ and $\zeta_s^t$ are defined using recursive equations that are similar to (12) for $\Delta\alpha(K - t)$'s, $t = 0, 1, \cdots, K$ except that the unknown variables in (12) are replaced by their known bounds, e.g, $K(\alpha)$ by $L$ or $U$, $\alpha(K)$ by $\alpha$, $\alpha(K - t)$ by $\alpha$ or $\alpha - \frac{t}{L+N-t+\frac{\Lambda}{\mu}(1-\alpha)}$, etc. On the one hand, the expressions of $\eta_s^t$ and $\zeta_s^t$ better resemble the recursive formulation of the actual $\Delta\alpha(K)$, leading to better approximations, as more iterations are applied. On the other hand, the replacement of the unknown actual variables with their bounds introduces additional errors with each iteration, which may lead to a deterioration in the quality of the generated bounds. In particular, such relaxation errors could be significant for the upper bounds $\zeta_s^t$'s because one of the approximate term, $\frac{(t+1)\Lambda/\mu}{[L+N-t-1+\frac{\Lambda}{\mu}(1-\alpha)]^2[L+N-t-2+\frac{\Lambda}{\mu}(1-\alpha)]}$, is increasing convex in $t$ (i.e., larger and larger relaxation errors are introduced with each iteration). We observe that the net effect is that the upper bounds $\zeta_s^1(L,U,\alpha)$'s may deteriorate with additional iterations when many iterations have already been applied and the bound is already very close to the actual $\Delta\alpha\big(K(\alpha)\big)$. This does not happen to the lower bounds because the unknown variables in their expressions are replaced by the parameters that remain constant over iterations. Thus, the $\eta_s^0(L,U,\alpha)$'s always increase (improve) over iterations. An illustration of this effect is presented in Figure 3.

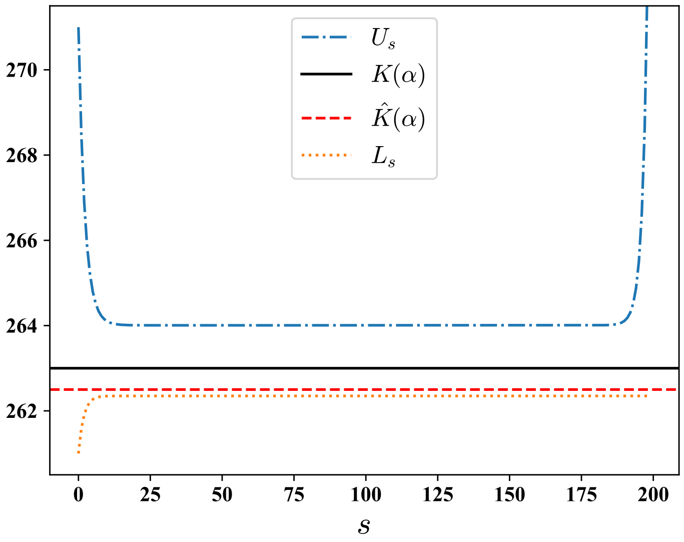


Figure 3: The lower and upper bounds as a function of the number of iterations ($\alpha = 0.9$, $\Lambda = 200$, $\mu = 1$, and $N = 10$)

**Proof of Proposition 3.** According to the definition of $L_s$ and $U_s$, it is sufficient to prove the following results.

1. For fixed $N$ and any finite $s, t \geq 1$, $\lim\limits_{\Lambda\to 0^+} \frac{\Lambda}{\mu}\eta_s^0(L_0, U_0, \alpha) = \lim\limits_{\Lambda\to 0^+} \frac{\Lambda}{\mu}\zeta_s^1(L_0, U_0, \alpha) = 0$.
2. For fixed $N$ and any finite $s, t \geq 0$, $\lim\limits_{\Lambda\to\infty} \frac{\Lambda}{\mu}\eta_s^t(L_0, U_0, \alpha) = 1 - \alpha^s$, and $\lim\limits_{\Lambda\to\infty} \frac{\Lambda}{\mu}\zeta_s^t(L_0, U_0, \alpha) = 1$.
3. For fixed $\Lambda$ and any finite $s, t \geq 1$, $\lim\limits_{N\to\infty} \frac{\Lambda}{\mu}\eta_s^t(L_0, U_0, \alpha) = \lim\limits_{\Lambda\to 0^+} \frac{\Lambda}{\mu}\zeta_s^t(L_0, U_0, \alpha) = 0$.

4. For fixed $\lambda = \frac{\Lambda}{N}$ and any finite $s, t \geq 0$,

$$\lim_{N\to\infty} \frac{\Lambda}{\mu}\eta_s^t(L_0, U_0, \alpha) = \frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}(1-\alpha)}\left[1 - \left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}}\right)^s\right], \text{ and}$$
$$\lim_{N\to\infty} \frac{\Lambda}{\mu}\zeta_s^t(L_0, U_0, \alpha) = \frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}(1-\alpha)} + \frac{\frac{1}{1-\alpha}}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}(1-\alpha)}\left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}}\right)^s.$$

5. For fixed $\Lambda > 0$, $N > 1$, and any finite $s, t \geq 2$, $\lim\limits_{\alpha\to 1} \frac{\Lambda}{\mu}\eta_s^t(L_0, U_0, \alpha)\frac{\alpha}{1-\alpha} = \lim\limits_{\alpha\to 1} \frac{\Lambda}{\mu}\zeta_s^t(L_0, U_0, \alpha)\frac{\alpha}{1-\alpha} = 0$. For $s = 1$, $\lim\limits_{\alpha\to 1} \frac{\Lambda}{\mu}\eta_1^t(L_0, U_0, \alpha)\frac{\alpha}{1-\alpha} = 0$ and $\lim\limits_{\alpha\to 1} \frac{\Lambda}{\mu}\zeta_1^t(L_0, U_0, \alpha)\frac{\alpha}{1-\alpha} = \frac{\Lambda}{\mu}\frac{N}{(N-1)^2}$.

These results can be proven by induction for $s = 1, 2, \cdots$. Noting that

$$L_0 - t = \frac{\Lambda}{\mu}\alpha + \frac{(N-1)\alpha}{1-\alpha} - t, \qquad L_0 + N - t - 1 + \frac{\Lambda}{\mu}(1-\alpha) = \frac{\Lambda}{\mu} + \frac{N-1}{1-\alpha} - t,$$
$$U_0 - t = \frac{\Lambda}{\mu}\alpha + \frac{N\alpha}{1-\alpha} + 1 - t, \text{ and} \qquad U_0 + N + \frac{\Lambda}{\mu}(1-\alpha) = \frac{\Lambda}{\mu} + \frac{N}{1-\alpha} + 1,$$

we know that (1) as $\Lambda \longrightarrow 0$, $\lim\limits_{\Lambda\to 0} L_0 = 0$, $\lim\limits_{\Lambda\to 0} U_0 = 1$; (2) as $\Lambda \longrightarrow \infty$,

$$L_0 - t = \frac{\Lambda}{\mu}\alpha + O(1), \qquad L_0 + N - t - 1 + \frac{\Lambda}{\mu}(1-\alpha) = \frac{\Lambda}{\mu} + O(1),$$
$$U_0 - t = \frac{\Lambda}{\mu}\alpha + O(1), \text{ and} \qquad U_0 + N + \frac{\Lambda}{\mu}(1-\alpha) = \frac{\Lambda}{\mu} + O(1);$$

(3) as $N \longrightarrow \infty$,

$$L_0 - t = \frac{N\alpha}{1-\alpha} + O(1), \qquad L_0 + N - t - 1 + \frac{\Lambda}{\mu}(1-\alpha) = \frac{N}{1-\alpha} + O(1),$$
$$U_0 - t = \frac{N\alpha}{1-\alpha} + O(1), \text{ and} \qquad U_0 + N + \frac{\Lambda}{\mu}(1-\alpha) = \frac{N}{1-\alpha} + O(1);$$

(4) as $N \longrightarrow \infty$ and $\Lambda = N\lambda$,

$$L_0 - t = N\left(\frac{\lambda}{\mu} + \frac{1}{1-\alpha}\right)\alpha + O(1), \qquad L_0 + N - t - 1 + \frac{\Lambda}{\mu}(1-\alpha) = N\left(\frac{\lambda}{\mu} + \frac{1}{1-\alpha}\right) + O(1),$$
$$U_0 - t = N\left(\frac{\lambda}{\mu} + \frac{1}{1-\alpha}\right)\alpha + O(1), \text{ and} \qquad U_0 + N + \frac{\Lambda}{\mu}(1-\alpha) = N\left(\frac{\lambda}{\mu} + \frac{1}{1-\alpha}\right) + O(1);$$

and (5) as $\alpha \longrightarrow 1$,

$$L_0 - t = \frac{(N-1)\alpha}{1-\alpha} + O(1) = \frac{N-1}{1-\alpha} + O(1), \qquad L_0 + N - t - 1 + \frac{\Lambda}{\mu}(1-\alpha) = \frac{N-1}{1-\alpha} + O(1),$$
$$U_0 - t = \frac{N\alpha}{1-\alpha} + O(1) = \frac{N}{1-\alpha} + O(1), \text{ and} \qquad U_0 + N + \frac{\Lambda}{\mu}(1-\alpha) = \frac{N}{1-\alpha} + O(1).$$

The proofs for regimes 1, 2, 3, and 5 can be completed using induction by applying the above limits to the definition of $\eta_s^t$ and $\zeta_s^t$. Here, we only provide the detailed steps for regime 4, the case that requires most involved expressions.

- For $s = 0$, $\frac{\Lambda}{\mu}\eta_0^t(L_0, U_0, \alpha) = 0$ and $\frac{\Lambda}{\mu}\zeta_0^t(L_0, U_0, \alpha) = 1$, which satisfy the induction hypothesis for all $t$.
- Suppose that the results hold for subscript $0, 1, \cdots, s-1$ for all $t$, then

$$\begin{aligned}
\lim_{N\to\infty} \frac{\Lambda}{\mu}\eta_s^t(L_0, U_0, \alpha) &= \frac{\frac{\lambda}{\mu}\left\{1 + \frac{\lambda}{\mu}(1-\alpha) + (\frac{\lambda}{\mu} + \frac{1}{1-\alpha})\alpha \frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}(1-\alpha)}\left[1 - \left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{1-\alpha}+\frac{\lambda}{\mu}}\right)^{s-1}\right]\right\}}{(\frac{\lambda}{\mu} + \frac{1}{1-\alpha})^2} \\
&= \frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{1-\alpha} + \frac{\lambda}{\mu}(1-\alpha)}\left[1 - \left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{1-\alpha}+\frac{\lambda}{\mu}}\right)^{s}\right], \\
\lim_{N\to\infty} \frac{\Lambda}{\mu}\zeta_s^t(L_0, U_0, \alpha) &= \frac{\frac{\lambda}{\mu}\left\{1 + \frac{\lambda}{\mu}(1-\alpha) + (\frac{\lambda}{\mu} + \frac{1}{1-\alpha})\alpha \left[\frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{(1-\alpha)} + \frac{\lambda}{\mu}(1-\alpha)} + \frac{\frac{1}{1-\alpha}}{\frac{1}{(1-\alpha)} + \frac{\lambda}{\mu}(1-\alpha)}\left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{(1-\alpha)}+\frac{\lambda}{\mu}}\right)^{s-1}\right]\right\}}{(\frac{\lambda}{\mu} + \frac{1}{1-\alpha})^2} \\
&= \frac{\frac{\lambda}{\mu}(1-\alpha)}{\frac{1}{(1-\alpha)} + \frac{\lambda}{\mu}(1-\alpha)} + \frac{\frac{1}{1-\alpha}}{\frac{1}{(1-\alpha)} + \frac{\lambda}{\mu}(1-\alpha)}\left(\frac{\frac{\lambda}{\mu}\alpha}{\frac{1}{(1-\alpha)}+\frac{\lambda}{\mu}}\right)^{s}.
\end{aligned}$$

□

**Proof of Proposition 4.** We prove the result for all $s = 0, \ldots, L_0 - 1$ where $(L_s, U_s)$ is well-defined in Corollary 1. We only consider the non-trivial case where $L_0 \geq 1$, i.e., $(\frac{\Lambda}{\mu} + \frac{N-1}{1-\alpha})\alpha \geq 1$. By comparing the expressions of (13), (14), and (15), it suffices to prove that

$$\eta_s^t(L_0, U_0, \alpha) < \frac{1}{\frac{N}{(1-\alpha)^2} + \frac{\Lambda}{\mu}} < \zeta_s^{t+1}(L_0, U_0, \alpha) \tag{22}$$

holds for all $s \geq 0$ and $t \geq 0$. We prove by induction on progressing index $s$ as follows.

- For $s = 0$, $\eta_0^t(L_0, U_0, \alpha) = 0 < \left[\frac{N}{(1-\alpha)^2} + \frac{\Lambda}{\mu}\right]^{-1} < \frac{\mu}{\Lambda} = \eta_0^{t+1}(L_0, U_0, \alpha)$, which satisfies the induction hypothesis.
- Suppose that (22) holds for subscripts $0, 1, \cdots, s-1$ for all $t$, we then examine whether it holds for subscript $s$. We know by definition that

$$\begin{aligned}
\eta_s^0(L_0, U_0, \alpha) &= \frac{(N-1) + \frac{\Lambda}{\mu}(1-\alpha) + L_0\frac{\Lambda}{\mu}\eta_{s-1}^1(L_0, U_0, \alpha)}{(\frac{N}{1-\alpha} + 1 + \frac{\Lambda}{\mu})^2} \quad \text{and} \\
\zeta_s^1(L_0, U_0, \alpha) &= \frac{(N-1) + \frac{\Lambda}{\mu}[1 - \alpha + \frac{2}{(L_0+N-2)+\frac{\Lambda}{\mu}(1-\alpha)}] + (U_0 - 1)\frac{\Lambda}{\mu}\zeta_{s-1}^2(L_0, U_0, \alpha)}{(\frac{N-1}{1-\alpha} - 1 + \frac{\Lambda}{\mu}) \cdot (\frac{N-1}{1-\alpha} - 2 + \frac{\Lambda}{\mu})}.
\end{aligned}$$

By simple algebra, the former is smaller than $\left[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\right]^{-1}$ if

$$\left[L_0\frac{\Lambda}{\mu}\eta^1_{s-1}(L_0,U_0,\alpha)-1\right]\big[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\big]<\big(\frac{\Lambda}{\mu}+\frac{N}{1-\alpha}\big)\big(\frac{\Lambda}{\mu}\alpha+2\big)+1;$$

the latter is greater than $\left[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\right]^{-1}$ if

$$(U_0-1)\frac{\Lambda}{\mu}\zeta^2_{s-1}(L_0,U_0,\alpha)\big[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\big]+\frac{2\frac{\Lambda}{\mu}\big[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\big]}{\frac{\Lambda}{\mu}+\frac{N-1}{1-\alpha}-1}>\big(\frac{\Lambda}{\mu}+\frac{N-1}{1-\alpha}\big)\frac{\Lambda}{\mu}\alpha-\big(\frac{\Lambda}{\mu}+\frac{N-1}{1-\alpha}\big)(3+\frac{1}{\alpha})+2.$$

Both of these conditions hold because

- $L_0\frac{\Lambda}{\mu}\eta^1_{s-1}(L_0,U_0,\alpha)\big[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\big]<L_0\frac{\Lambda}{\mu}=\big(\frac{\Lambda}{\mu}+\frac{N-1}{1-\alpha}\big)\frac{\Lambda}{\mu}\alpha$ by the induction hypothesis for subscript $s-1$.
- $(U_0-1)\frac{\Lambda}{\mu}\zeta^2_{s-1}(L_0,U_0,\alpha)\big[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\big]>(U_0-1)\frac{\Lambda}{\mu}=\big(\frac{\Lambda}{\mu}+\frac{N}{1-\alpha}\big)\frac{\Lambda}{\mu}\alpha$ by the induction hypothesis for subscript $s-1$.
- $-\big(\frac{\Lambda}{\mu}+\frac{N-1}{1-\alpha}\big)(3+\frac{1}{\alpha})+2<-\frac{1}{\alpha}(3+\frac{1}{\alpha})+2<0$ because $(\frac{\Lambda}{\mu}+\frac{N-1}{1-\alpha})\alpha\geq 1$ in the non-trivial case.

Therefore, we obtain $\eta^0_s(L_0,U_0,\alpha)<\left[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\right]^{-1}<\zeta^1_s(L_0,U_0,\alpha)$. Since $\eta^t_s(L_0,U_0,\alpha)$ decreases in $t$ and $\zeta^t_s(L_0,U_0,\alpha)$ increases in $t$, we further obtain $\eta^t_s(L_0,U_0,\alpha)<\left[\frac{N}{(1-\alpha)^2}+\frac{\Lambda}{\mu}\right]^{-1}<\zeta^{t+1}_s(L_0,U_0,\alpha)$ for all $t\geq 0$. The induction is completed.

□

**Proof of Proposition 5.** The proof is trivial according to (15). □

**Proof of Proposition 6.** By the monotonicity of $\alpha(K)$ and (8), we know that $\lim\limits_{K\to\infty}\alpha(K)$ exists and equals to 1. By (8), we know that

$$1-\alpha(K)=\frac{(N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}{(K+N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}. \tag{23}$$

<u>If $N>1$</u>, then

$$\lim_{K\to\infty}\left\{[1-\alpha(K)]\frac{K}{N-1}\right\}=\lim_{K\to\infty}\left\{\frac{(N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}{N-1}\frac{K}{(K+N-1)+\frac{\Lambda}{\mu}[1-\alpha(K-1)]}\right\}=1.$$

$\underline{\text{If } N = 1}$, by (23), we know

$$\frac{1}{1-\alpha(K)} = 1 + \frac{K(\frac{\Lambda}{\mu})^{-1}}{1-\alpha(K-1)}, \text{ for all } K, \text{ and} \tag{24}$$

$$\frac{1}{1-\alpha(K)} + \frac{\frac{\Lambda}{\mu}}{K-\frac{\Lambda}{\mu}} = K(\frac{\Lambda}{\mu})^{-1}\left(\frac{1}{1-\alpha(K-1)} + \frac{\frac{\Lambda}{\mu}}{K-\frac{\Lambda}{\mu}}\right), \text{ for all } K \neq \tfrac{\Lambda}{\mu}. \tag{25}$$

On the one hand, let $\underline{\omega}(K) := K(\frac{\Lambda}{\mu})^{-1}\underline{\omega}(K-1)$ be recursively defined with $\underline{\omega}(1) := \frac{1}{1-\alpha(1)} = (\frac{\Lambda}{\mu})^{-1}(1+\frac{\Lambda}{\mu})$. Then, $\underline{\omega}(K) = (K!)(\frac{\Lambda}{\mu})^{-K}(1+\frac{\Lambda}{\mu})$, and $\frac{1}{1-\alpha(K)} \geq \underline{\omega}(K)$ for all $K \geq 1$ by comparing (24) and the definition of $\underline{\omega}(K)$ via induction. This implies that $\frac{1}{1-\alpha(K)} \geq (K!)(\frac{\Lambda}{\mu})^{-K}(1+\frac{\Lambda}{\mu})$, and hence

$$\limsup_{K\to\infty}\left\{[1-\alpha(K)](K!)(\frac{\Lambda}{\mu})^{-K}\right\} \leq \frac{1}{1+\frac{\Lambda}{\mu}} < 1. \tag{26}$$

On the other hand, for any $k > \frac{\Lambda}{\mu}$, let $\bar{\omega}^k(K) := K(\frac{\Lambda}{\mu})^{-1}\bar{\omega}^k(K-1)$ be recursively defined for all $K \geq k$, where $\bar{\omega}^k(k-1) := \frac{1}{1-\alpha(k-1)} + \frac{\frac{\Lambda}{\mu}}{k-\frac{\Lambda}{\mu}}$. Then, $\bar{\omega}^k(K) = (K!)(\frac{\Lambda}{\mu})^{-K} \cdot \frac{\frac{1}{1-\alpha(k-1)}+\frac{\frac{\Lambda}{\mu}}{k-\frac{\Lambda}{\mu}}}{(k-1)!(\frac{\Lambda}{\mu})^{-(k-1)}}$, and $\bar{\omega}^k(K) \geq \frac{1}{1-\alpha(K)} + \frac{\frac{\Lambda}{\mu}}{K-\frac{\Lambda}{\mu}}$ for all $K \geq k$ by comparing (25) and the definition of $\bar{\omega}^k(K)$ via induction. This implies that

$$\liminf_{K\to\infty}\left\{[1-\alpha(K)](K!)(\frac{\Lambda}{\mu})^{-K}\right\} \geq \liminf_{K\to\infty}\frac{(k-1)!(\frac{\Lambda}{\mu})^{-(k-1)}}{\frac{1}{1-\alpha(k-1)}+\frac{\frac{\Lambda}{\mu}}{k-\frac{\Lambda}{\mu}}}\left[1+\frac{\frac{\Lambda}{\mu}(1-\alpha(K))}{K-\frac{\Lambda}{\mu}}\right] = \frac{(k-1)!(\frac{\Lambda}{\mu})^{-(k-1)}}{\frac{1}{1-\alpha(k-1)}+\frac{\frac{\Lambda}{\mu}}{k-\frac{\Lambda}{\mu}}} > 0. \tag{27}$$

Let $k \to \infty$, we obtain

$$\liminf_{K\to\infty}\left\{[1-\alpha(K)](K!)(\frac{\Lambda}{\mu})^{-K}\right\} \geq \limsup_{k\to\infty}\frac{(k-1)!(\frac{\Lambda}{\mu})^{-(k-1)}}{\frac{1}{1-\alpha(k-1)}+\frac{\frac{\Lambda}{\mu}}{k-\frac{\Lambda}{\mu}}} \geq \frac{\limsup_{k\to\infty}\left\{[1-\alpha(k-1)](k-1)!(\frac{\Lambda}{\mu})^{-(k-1)}\right\}}{\liminf_{k\to\infty}\left\{1+\frac{\frac{\Lambda}{\mu}[1-\alpha(k-1)]}{k-\frac{\Lambda}{\mu}}\right\}}$$
$$= \limsup_{k\to\infty}\left\{[1-\alpha(k)](k!)(\frac{\Lambda}{\mu})^{-k}\right\},$$

which suggests that

$$\liminf_{K\to\infty}\left\{[1-\alpha(K)](K!)(\frac{\Lambda}{\mu})^{-K}\right\} = \limsup_{K\to\infty}\left\{[1-\alpha(K)](K!)(\frac{\Lambda}{\mu})^{-K}\right\}$$

and $\lim_{K\to\infty}\left\{[1-\alpha(K)](K!)(\frac{\Lambda}{\mu})^{-K}\right\}$ exits. Its value lies in $(0,1)$ due to (26) and (27). □

**Proof of Lemma 4.** The proofs of properties 1 and 2 are straightforward. Property 1 obviously holds because the recursive equation (8) depends only on the ratio of $\Lambda$ to $\mu$. From (8), we can also easily prove, by induction, that $\alpha(K,\Lambda,\mu,N)$ decreases in $N$, which immediately yields property 2.

If (8) is extended to be appropriately defined for real-valued $K$, then $K(\alpha)$ can be re-defined as the continuous inverse of $\alpha(K)$ such that $K(\alpha) := \{K : \alpha(K) = \alpha\}$. In this case, (9) holds for real-valued $K$, which yields

$$K(\alpha) = \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\Lambda}{\mu}\frac{\alpha}{1-\alpha}\Delta\alpha\big(K(\alpha)\big).$$

By Lemma 3 and the proof of result 2 of Proposition 3, we obtain

$$1 - \alpha^s = \lim_{\Lambda\to\infty}\frac{\Lambda}{\mu}\eta_s^0(L_0, U_0, \alpha) < \frac{\Lambda}{\mu}\Delta\alpha\big(K(\alpha)\big) < \lim_{\Lambda\to\infty}\frac{\Lambda}{\mu}\zeta_s^1(L_0, U_0, \alpha) = 1,$$

which, by letting $s \to \infty$ as $\Lambda \to \infty$, proves that $\lim_{\Lambda\to\infty}\{K(\alpha, \Lambda, \mu, N) - (\frac{\Lambda}{\mu}\alpha + N\frac{\alpha}{1-\alpha})\} = 0$. This result is sufficient to guarantee property 4.

The proof of property 3 is more involved. Note that the recursive equation (8) depends on the $\lambda_i$'s only via $\Lambda$ and the $\mu_{ij}$'s only via $\mu$. Therefore, balanced systems with the same total arrival rate and overall average rental time have the same performance even though they may have different $\lambda_i$'s, $\mu_{ij}$'s, and rental time distributions. This implies that it is sufficient to prove that property 3 holds for a symmetric system where each region has the same demand rate $\lambda$ and each customer has the same exponentially distributed rental time with mean $\frac{1}{\mu}$. We may also further restrict the routing matrix $\boldsymbol{P}$ as follows:

$$p_{ij} = \begin{cases} 1, & \text{if } j = i+1, \\ 1, & \text{if } i = N, j = 1, \\ 0, & \text{otherwise,} \end{cases}$$

which implies that a vehicle picked up at location $i$ is dropped off at location $i+1$ when $i < N$ and vehicles picked up at location $N$ are dropped off at location 1. A graphic representation of this *cyclic* network is shown in Figure 4. We use the notation $X_i$ and $Y_i$ to denote the number of vehicles at location $i$'s pick-up queue and transit queue respectively. In this network, the throughput of all the queues are the same and we use the notation $f_N(K)$ to denote the throughput rate per location when the cyclic network has $K$ vehicles and $N$ locations. This throughput is a concave function of $K$ for any $N$ (a result we use below) since the total throughput of a closed Jackson queueing network is nondecreasing concave in the number of items in the network (Shanthikumar and Yao (1988)).

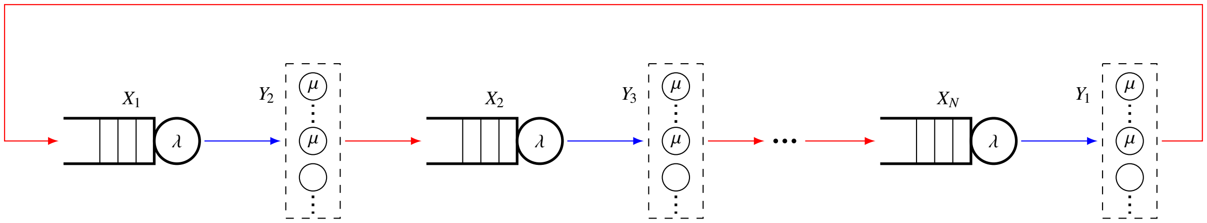


Figure 4: A cyclic queueing network representation of a one-way vehicle sharing system

Next, we describe a network aggregation procedure due to Chandy et al. (1975) that we will deploy in our proof. Consider a closed queuing network with exponential service queues. A subnetwork $\sigma$ is a subset of queues in the network such that items enter this subset through only one starting point and exit it through only one endpoint. For example, the transit queue 2 and pick-up queue 2 in Figure 4 constitute a subnetwork, while the transit queues 2 and 3 do not. A reduced network with $\sigma$ "shorted" is a modification to the original network in which the service times of all the servers in the subnetwork $\sigma$ are set to zero. Let $T(K)$ be the throughput rate passing the endpoint of the shorted subnetwork when there are a total number of $K$ items in the reduced network. Then, we construct an equivalent network with a composite $\sigma^c$ by replacing all the queues in the original network, except those in the subnetwork $\sigma$, by a single composite queue which has a state dependent service rate $T(k)$ when there are $k$ items in its queue. That is, the equivalent network consists of the queues in the original subnetwork $\sigma$ and a single composite queue.

**Lemma 7** (Theorems 1 & 2 of Chandy et al. (1975))**.** *The behavior of $\sigma$ in the equivalent network is identical to that in the original network, i.e., they have the same queue length and queue time distributions.*

For our cyclic network with $2N$ queues and $NK$ items (see Figure 4), we propose the following aggregation procedure, consistent with the procedure described above, to construct an equivalent network with $N$ indentical queues and $NK$ items. Specifically, we sequentially aggregate, starting with location 1, the pickup and transit queues into a single queue (see Figure 5 for an illustration). The resulting equivalent network (with the same throughput per location) has $N$ identical queues, with each queue having a queue-length dependent service rate function $f_1(\cdot)$. Let $Z_i, i = 1, 2, \cdots, N$ denote the length of each queue in the equivalent network. By definition, the total throughput of the equivalent network can be derived as

$$\sum_{Z_1+Z_2+\cdots+Z_N=NK} [P(Z_1, Z_2, \cdots, Z_N) \sum_{i=1}^{N} f_1(Z_i)],$$

which equals the total throughput rate of all pick-up queues in the original network, yielding

$$\begin{aligned} Nf_N(NK) =& \sum_{Z_1+Z_2+\cdots+Z_N=NK} [P(Z_1, Z_2, \cdots, Z_N) \sum_{i=1}^{N} f_1(Z_i)] \\ \leq& \sum_{Z_1+Z_2+\cdots+Z_N=NK} [P(Z_1, Z_2, \cdots, Z_N) N f_1(K)] \\ =& N f_1(K), \end{aligned}$$

where the inequality holds by the concavity of $f_1(\cdot)$ and Jensen's inequality. Noting $\alpha(NK, N\lambda, \mu, N) = \frac{f_N(NK)}{\lambda}$ and $\alpha(K, \lambda, \mu, 1) = \frac{f_1(K)}{\lambda}$, we have $\alpha(NK, N\lambda, \mu, N) \leq \alpha(K, \lambda, \mu, 1)$. □

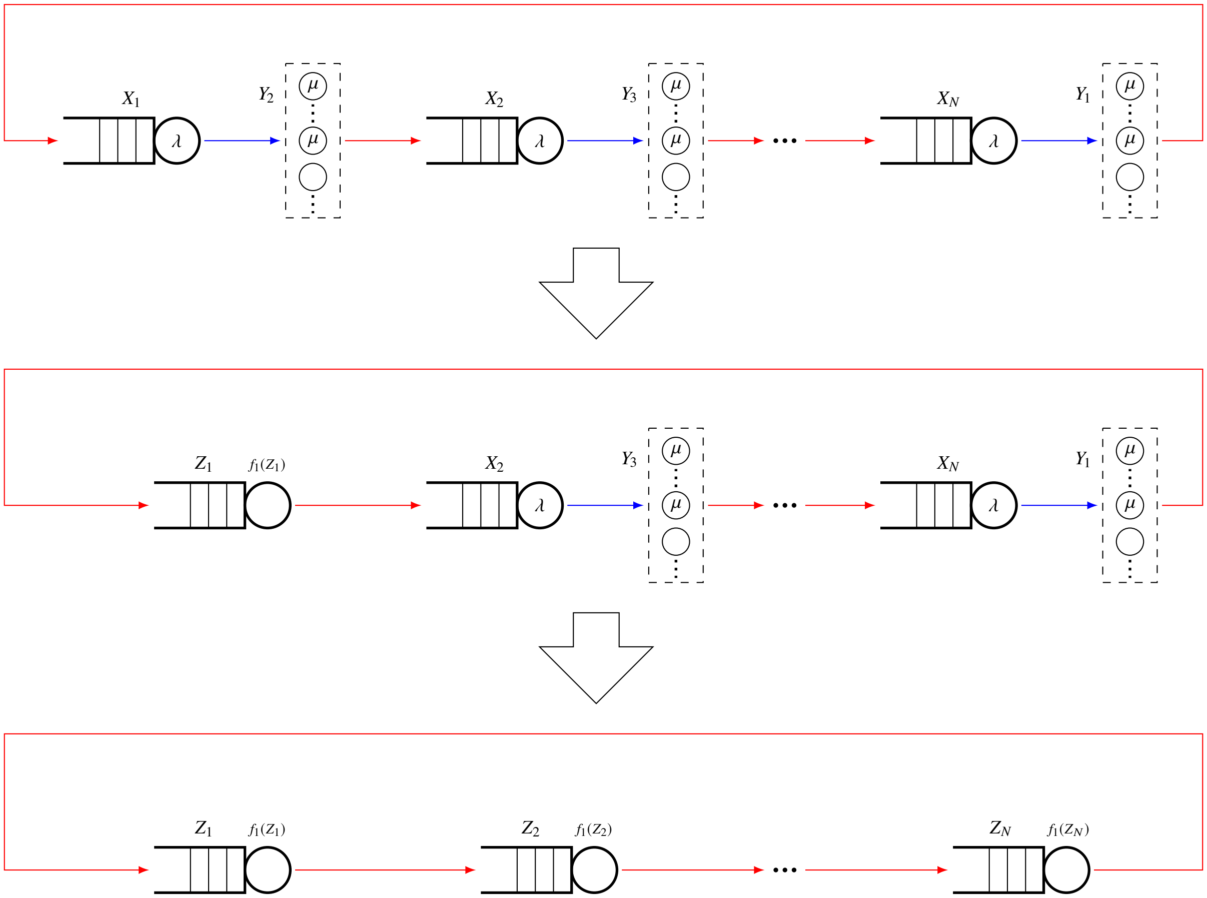


Figure 5: A graphical illustration of the aggregation procedure

# Appendix B Dimensioning Unbalanced Networks

In this section, we derive recursive equations for unbalanced networks that are analogous to (8) for balanced networks. These equations allow us to efficiently compute the average system performance and determine the optimal fleet size, though not yielding closed-form expressions. We discuss how in an unbalanced network, the problem of determining a fleet size that guarantees a specified service level at each location may not have a feasible solution. That is, even with an infinitely large number of vehicles, it may not be possible to achieve a target service level (if this target is sufficiently high) at each location; these results are due to George and Xia (2011) and we refer the interested reader to their paper for more details.

First, note that the network continues to be a BCMP Network. Therefore, the stationary distribution of system states can still be obtained via (1)–(3) and used to compute various performance measures, including throughput at each transit and pickup queue, total system throughput, and service level at each location. Note that because $\lambda_i = \sum_j \lambda_j p_{ji}$ does not hold for all $i$, the service level induced in steady state by a given number of vehicles is no longer the same at different locations. Hence, the dimensioning problem becomes one of finding the smallest number of vehicles that guarantees that the smallest service level is greater or equal than the specified target service level. As in the case of a balanced network, this approach requires significant computational effort and lacks interpretability.

Mean value analysis can alternatively be used. This requires some modifications from the balanced system case which we describe next. Let $r_{ij}$ be the proportion of effective rentals that originate in location $i$ and terminate in location $j$ when there is only a single vehicle in the system (i.e., $K = 1$) and let $r_i = \sum_{j \in V} r_{ij}$. From the point of view of this vehicle, its transitions between locations, regardless of the lengths of stay, are governed by the transition probabilities $p_{ij}$. Therefore, the $r_i$'s are the steady-state probabilities of a discrete time Markov Chain with transition matrix with elements $\{p_{ij}\}$ (which can be easily computed by substituting $\nu_i$ with $r_i$ in (3) and requiring $\sum_{i \in V} r_i = 1$), and $r_{ij} = r_i p_{ij}$. Noting that the above argument does not involve the lengths of stay which depend on the demand rates and the number of vehicles, it applies to each individual vehicle even when there are multiple vehicles in the system. Therefore, the proportion of effective rentals that originate in location $i$ and terminate in location $j$ in a system with $K$ vehicles always equals $r_{ij}$ regardless of $K$, i.e., $\frac{\nu_{ij}(K)}{\nu(K)} = r_{ij}$. From the balance equation of the above-mentioned single-vehicle Markov Chain, we know that $r_i = \sum_{j \in V} r_{ij} = \sum_{j \in V} r_{ji}$. That is, $r_i$ is the proportion of effective rentals that originate in location $i$ and also the proportion of effective rentals that terminate in location $i$.

Noting that equations (4) and (5) also hold for unbalanced networks and that the first equality of (6) remains valid leads to the following modified version of (6)

$$\mathbb{E}[X_i(K)] = \nu_i(K)\frac{1 + \mathbb{E}[X_i(K-1)]}{\lambda_i} = \frac{r_i}{\lambda_i}\big(1 + \mathbb{E}[X_i(K-1)]\big)\nu(K), \tag{28}$$

and

$$\sum_i \mathbb{E}[X_i(K)] = \nu(K)\sum_i \Big\{\frac{r_i}{\lambda_i}(1 + \mathbb{E}[X_i(K-1)])\Big\}. \tag{29}$$

Substituting (5) into (29) leads to

$$\nu(K) = \frac{K\lambda\mu}{\lambda + \mu + \lambda\mu \sum_{i \in V}\big(\frac{r_i}{\lambda_i}\mathbb{E}[X_i(K-1)]\big)}, \tag{30}$$

where $\lambda := (\sum_{i \in V} \frac{r_i}{\lambda_i})^{-1}$ and $\mu = \big(\sum_{i,j \in V} \frac{r_{ij}}{\mu_{ij}}\big)^{-1}$. By letting $\mathbb{E}[X_i(0)] = 0$, equations (30) and (28) hold for $K \geq 1$, and can be used recursively to compute the throughput rate $\nu(K)$ and to do so efficiently (note that, as with a balanced network, the computational effort does not depend on the state space). Having obtained $\nu(K)$, other performance measures can be derived, including $\nu_i(K) = r_i \nu(K)$, $\nu_{ij}(K) = r_{ij}\nu(K)$, and $\alpha_i(K) = \frac{\nu_i(K)}{\lambda_i}$ ($\alpha_i(K)$ can also be rewritten as $\alpha_i(K) = \frac{r_i}{\lambda_i}\nu(K)$, where $\frac{r_i}{\lambda_i}$ is independent of $K$), which implies that $\alpha_i(K) > \alpha_j(K)$ if an only if $\frac{r_i}{\lambda_i} > \frac{r_j}{\lambda_j}$.

Let $B = \{i \in V : \frac{r_i}{\lambda_i} \geq \frac{r_j}{\lambda_j}, \forall j \in V\}$ denote the set of locations with the largest ratio $\frac{r_i}{\lambda_i}$ (because the system is unbalanced, not all the ratios can be equal and the set $B$ is a proper subset of $V$). Let $\alpha_i(\infty) := \lim_{K \to \infty} \alpha_i(K)$ and define similarly $\nu_i(\infty)$, $\mathbb{E}[X_i(\infty)]$, and $\mathbb{E}[Y_{ij}(\infty)]$. The following proposition describes several useful properties of unbalanced systems, including asymptotic results as $K$ becomes large.

**Proposition 7** (Reproduction of Theorems 1 of George and Xia (2011) with Minor Extension)**.** [7]

*(i).* $\alpha_i(K)$, $\nu_i(K)$, $\mathbb{E}[X_i(K)]$, *and* $\mathbb{E}[Y_{ij}(K)]$ *increase in* $K$*, for all* $i, j \in V$.

*(ii).* $\alpha_i(\infty) = 1$ *and* $\alpha_j(\infty) = \frac{r_j \lambda_i}{\lambda_j r_i} < 1$*, for all* $i \in B,\ j \in V \setminus B$.

*(iii).* $\nu_i(\infty) = \lambda_i$ *and* $\nu_j(\infty) = \frac{r_j}{r_i}\lambda_i < \lambda_j$*, for all* $i \in B,\ j \in V \setminus B$.

*(iv).* $\mathbb{E}[X_i(\infty)] = \infty$ *and* $\mathbb{E}[X_j(\infty)] = \frac{\alpha_j(\infty)}{1-\alpha_j(\infty)} < \infty$*, for all* $i \in B,\ j \in V \setminus B$.

*(v).* $\mathbb{E}[Y_{ij}(\infty)] = \frac{\lambda_i p_{ij}}{\mu_{ij}}\alpha_i(\infty)$*, for all* $i, j \in V$.

**Proof of Proposition 7.** We prove results (i)–(v) in sequence.

(i). It is easy to prove by the coupling technique that $\nu(K)$, $\mathbb{E}[X_i(K)]$, and $\mathbb{E}[Y_{ij}(K)]$ increase in $K$. Because $\nu_i(K) = r_i \nu(K)$ and $\alpha_i(K) = \frac{\nu_i(K)}{\lambda_i}$, they increase in $K$ as well. Since $\alpha_i(K)$ is bounded above by 1, the limit $\lim\limits_{K\to\infty} \alpha_i(K)$ exists.

(ii). By the expression $\alpha_i(K) = \frac{r_i}{\lambda_i}\nu(K)$, we know that $\frac{\alpha_j(K)}{\alpha_i(K)} = \frac{r_j\lambda_i}{\lambda_j r_i}$ and $\alpha_j(\infty) = \frac{r_j\lambda_i}{\lambda_j r_i}\alpha_i(\infty)$, $\forall i, j \in V,\ K \geq 1$. Because $\nu(K)$ is bounded above by $\Lambda$, we know by (30) that $\lim\limits_{K\to\infty} \sum\limits_{i\in V} \left(\frac{r_i}{\lambda_i}\mathbb{E}[X_i(K-1)]\right) = \infty$. There must exist at least one $i' \in V$ such that $\lim\limits_{K\to\infty} \mathbb{E}[X_{i'}(K-1)] = \infty$. Noting that equation (28) can be written as

$$\mathbb{E}[X_i(K)] = \alpha_i(K)\big(1 + \mathbb{E}[X_i(K-1)]\big), \tag{31}$$

we must have $\alpha_{i'}(\infty) = 1$ since it will lead to a contradiction otherwise. By the equation $\alpha_i(\infty) = \frac{r_i\lambda_{i'}}{\lambda_i r_{i'}}\alpha_{i'}(\infty)$, we know that this $i'$ must be in $B$. Otherwise, $\alpha_i(\infty)$ is greater than 1 for $i \in B$. Thus, we have $\alpha_i(\infty) = 1$, and hence $\alpha_j(\infty) = \frac{r_j\lambda_i}{\lambda_j r_i} < 1$, $\forall i \in B,\ j \in V \setminus B$.

(iii). This result follows directly from result (ii) and the expression $\nu_i(K) = \alpha_i(K)\lambda_i$.

(iv). We prove by contradiction. Since $\mathbb{E}[X_i(K)]$ increases in $K$, it either grows unboundedly or converges to a finite value as $K \to \infty$. If $\lim\limits_{K\to\infty} \mathbb{E}[X_i(K)] = C < \infty$ for some $i \in B$, then by letting $K \to \infty$ on both sides of (31) we have $C = 1 + C$, leading to a contradiction. Thus, $\lim\limits_{K\to\infty} \mathbb{E}[X_i(K)] = \infty$ for any $i \in B$.

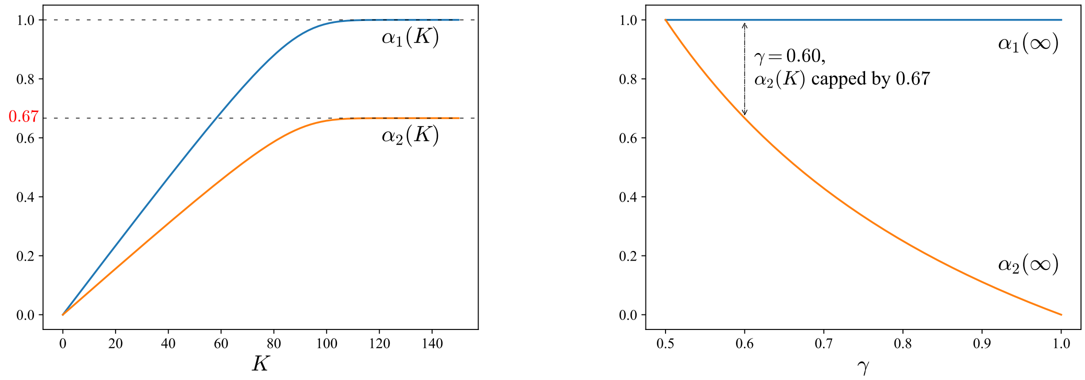
Because $\lim\limits_{K\to\infty} \alpha_j(K) < 1$ for $j \in V \setminus B$ and $\mathbb{E}[X_j(0)] = 0$, it is also straightforward to see from (31) that $\lim\limits_{K\to\infty} \mathbb{E}[X_j(K)] < \infty$. By letting $K \to \infty$ on both sides of (31), we have $\lim\limits_{K\to\infty} \mathbb{E}[X_j(K)] = \frac{\alpha_j(\infty)}{1-\alpha_j(\infty)} < \infty$, $\forall j \in V \setminus B$.

(v). By Little's Law, $\mathbb{E}[Y_{ij}(K)] = \frac{\nu_{ij}(K)}{\mu_{ij}} = \frac{\nu_i(K)p_{ij}}{\mu_{ij}} = \frac{\alpha_i(K)\lambda_i p_{ij}}{\mu_{ij}}$.

[7] The proposition recasts Theorem 1 of George and Xia (2011) using our notation and it extends by including $\nu_i(\infty)$ and $\mathbb{E}[X_i(\infty)]$.

□

Proposition 7 shows while some location(s) can have arbitrarily high service level as the total number of vehicles increases, the service levels at other locations are bounded by specific thresholds. That is, it is impossible to achieve service levels above these thresholds even with an infinite number of vehicles. The average numbers of vehicles at locations in $V \setminus B$ and in transit are also bounded by finite fixed thresholds (no matter how large is the total number of vehicles) and so are the associated throughputs. These results are illustrated for an example with two locations in Figure 6. Figure 6 shows how even modest differences in the relative popularity of locations (either as origins or destinations) can lead to significant differences in the achievable service levels at these locations.



(a) $\gamma = 0.6$

(b) $K = \infty$

Figure 6: A two-location unbalanced network with $\lambda_1 = \lambda_2 = 50$, $\mu_{ij} = 1$, $p_{11} = p_{21} = \gamma$, and $p_{12} = p_{22} = 1 - \gamma$.

The intuition for the results is as follows. Note that in an unbalanced network, locations may be more popular as an origin or as a destination. The popularity of a location as a destination relative to it being an origin is indicated by the ratio $\frac{r_i}{\lambda_i}$, which is referred to as the relative utilization of the location by George and Xia (2011). In the long run, vehicles accumulate at locations that are most popular as destination (relative to their popularity as origin), i.e., with the greatest $\frac{r_i}{\lambda_i}$. These locations are referred to as bottleneck locations of the closed network. In this case, the supply of vehicles the non-bottleneck locations receive depends on the bottleneck throughput rates, which are bounded by the demand rates at the bottleneck locations regardless of the fleet size. Therefore, no matter how large the fleet size is and how many vehicles are initially provisioned to non-bottleneck locations, a large (majority) number of the vehicles will later accumulate and stay idle at the bottleneck locations, guaranteeing an arbitrarily high service level there and at the same time making the service levels at non-bottleneck locations bounded by specific thresholds in the long run.

## Appendix C An Application: Optimizing the Service Level

In this section, we briefly illustrate how the minimal fleet size approximation in (15) can be embedded in an optimization problem. In particular, we illustrate how the service level can be endogenized by letting it to be a decision that the service provider makes. We do not intend this to be a full treatment of the problem, but simply an illustration of the usefulness of an approximation in supporting operational decision making and in obtaining additional managerial insights[8].

Let $r$ denote the price the service provider charges for each rental per unit time the vehicle is rented. Let also $c$ denote the cost of a vehicle per unit of time (this cost may include the amortized purchase cost of the vehicle as well as its operating cost). Assume $c < r$ (otherwise offering the service would not be profitable) and assume that the network is balanced. The service provider's profit maximization problem can be stated as follows:

$$\begin{aligned}\max{}_{\alpha}\pi(\alpha) &= \max{}_{\alpha}\left\{r\Lambda\frac{1}{\mu}\alpha - c\hat{K}(\alpha,\Lambda,\mu,N)\right\}\\ &= \max{}_{\alpha}\left\{\frac{r\Lambda\alpha}{\mu} - c\left[\frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\Lambda}{\frac{N\mu}{1-\alpha}+\Lambda(1-\alpha)}\alpha\right]\right\}. \qquad (32)\end{aligned}$$

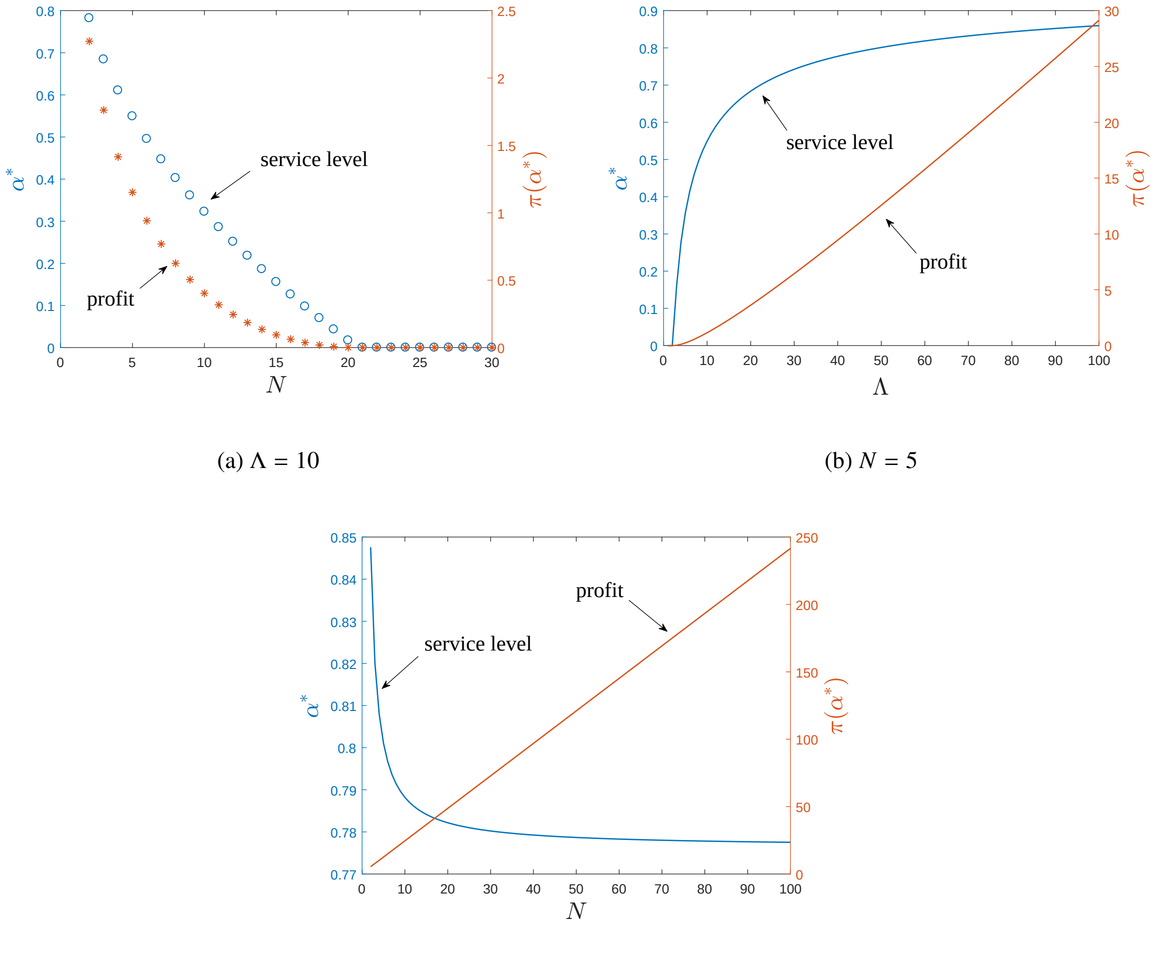
**Proposition 8.** *$\pi(\alpha)$ is concave in $\alpha$. The optimal service level, $\alpha^*(N,\Lambda)$, is unique and has the following properties:*

1. *$\alpha^*(N,\Lambda) > 0$ if and only if $c \le \frac{\frac{r\Lambda}{\mu}}{\frac{\Lambda}{\mu}+N-1+\frac{\Lambda}{N\mu+\Lambda}}$;*
2. *for fixed $\Lambda$, $\alpha^*(N,\Lambda)$ decreases in $N$ and $\lim\limits_{N\to\infty}\alpha^*(N,\Lambda) = 0$;*
3. *for fixed $N$, $\alpha^*(N,\Lambda)$ increases in $\Lambda$ and $\lim\limits_{\Lambda\to\infty}\alpha^*(N,\Lambda) = 1$; and*
4. *for $\frac{\Lambda}{N} \equiv \lambda$, $\alpha^*(N,\Lambda)$ decreases in $N$. If $\lambda < \frac{c\mu}{r-c}$, then $\alpha^*(N,\Lambda) > 0$ holds only for $N \le \frac{c\mu^2}{(\mu+\lambda)(r-c)(\frac{c\mu}{r-c}-\lambda)}$; otherwise, $\alpha^*(N,\Lambda) > 0$ holds for any $N$ and $\lim\limits_{N\to\infty}\alpha^*(N,\Lambda) = 1-\sqrt{\frac{c\mu}{(r-c)\lambda}}$.*

The proposition shows that the problem is concave and, hence, admits a unique solution. Property 1 in the proposition provides a necessary and sufficient condition for the service provider to realize a positive profit. Property 2 shows that, all else being equal, there is a tradeoff between location density (number of locations) and service level. Property 3 shows that, perhaps surprisingly, increased demand does not lead to a deterioration in the service level but instead to an increase (because of the pooling effect, the service provider makes additional investments in vehicles resulting in a higher service level). Property 4 shows that there is a tradeoff between the size of the service region (where an increase in demand requires an increase in the number of locations) and service level. If the demand density (average demand per location) is sufficiently

[8] It is possible to consider other problems, including determining location density (the optimal number of locations for fixed demand), sizing of the service region (optimal demand level), and service pricing (optimal price to charge).

large, then the firm is always profitable no matter how large the service region is; otherwise, the firm is profitable only when the number of locations is sufficiently small. The underlying reason why the limit exists is that, as $N \longrightarrow \infty$, the correction term $B_0(1-\frac{1}{N})$ approaches the standard buffer $B_0$ and the minimal fleet size reduces to $\frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha}$. That is, buffer capacity $(N-1)\frac{\alpha}{1-\alpha}$ is sufficient to protect against both vehicle roaming and randomness in demand and service times (see also the discussion in Section 6). These results are illustrated graphically in Figure 7.



(a) $\Lambda = 10$

(b) $N = 5$

(c) $\Lambda/N = 10$

Figure 7: The optimal service level ($c = 0.2$, $r = 0.6$, $\mu = 1$)

**Proof of Proposition 8.** Since

$$\pi''(\alpha) = -2c\frac{(2N-1)(\frac{\Lambda}{\mu})^2(1-\alpha)^2 + 2(N-1)\frac{N^2}{(1-\alpha)^2} + N\frac{\Lambda}{\mu}(4N-3-3\alpha)}{[\frac{N}{1-\alpha} + \frac{\Lambda}{\mu}(1-\alpha)]^2(1-\alpha)^3} < 0,$$

$\pi(\alpha)$ is concave in $\alpha$ and the profit maximization problem has a unique solution $\alpha^*(N,\Lambda)$. Furthermore, the first order condition

$$\pi'(\alpha^*) = r\frac{\Lambda}{\mu} - c\Big[\frac{\Lambda}{\mu} + (N-1)\frac{1}{(1-\alpha^*)^2} + \frac{(\frac{\Lambda}{\mu})^2 + \frac{N\Lambda(1-2\alpha^*)}{\mu(1-\alpha^*)^2}}{[\frac{N}{1-\alpha^*} + \frac{\Lambda}{\mu}(1-\alpha^*)]^2}\Big] = 0 \tag{33}$$

is satisfied if $\alpha^*(N,\Lambda)$ is in $(0,1)$. Next, we prove properties 1–4.

(1). Noting that the concavity of $\pi(\alpha)$ and that $\pi(0) = 0$, the solution $\alpha^*(N,\Lambda)$ is non-zero if and only if

$$\pi'(0) = r\frac{\Lambda}{\mu} - c[\frac{\Lambda}{\mu} + (N-1) + \frac{\Lambda}{N\mu + \Lambda}] > 0. \tag{34}$$

(2). Since

$$\frac{\partial^2\pi}{\partial\alpha\partial N}(\alpha) = -c\frac{N^3 + (\frac{\Lambda}{\mu})^3(1-\alpha)^6 + (\frac{\Lambda}{\mu})^2(1-\alpha)^4(3N-1-2\alpha) + N(\frac{\Lambda}{\mu})(1-\alpha)^2(3N-1+2\alpha)}{(1-\alpha)^2[N + \frac{\Lambda}{\mu}(1-\alpha)^2]^3} \leq 0,$$

the profit function is submodular in $(\alpha, N)$ and hence the optimal service level $\alpha^*(N,\Lambda)$ decreases in $N$. Moreover, we know by property 1 that $\alpha^*(N,\Lambda) = 0$ when $N$ is sufficiently large.

(3). Note that

$$\frac{\partial^2\pi}{\partial\alpha\partial\Lambda}(\alpha) = r - c - c\frac{N\frac{\Lambda}{\mu}\frac{1+2\alpha}{1-\alpha} + N^2\frac{1-2\alpha}{(1-\alpha)^3}}{[\frac{N}{1-\alpha} + \frac{\Lambda}{\mu}(1-\alpha)]^3}.$$

By property 1, $\alpha^*(N,\Lambda) > 0$ and hence the first order condition (33) holds at $\alpha^*(N,\Lambda)$ when $\Lambda$ is sufficiently large. In this case, by substituting (33) into the above expression, we know

$$\begin{aligned}\frac{\partial^2\pi}{\partial\alpha\partial\Lambda}(\alpha^*) =& \frac{c\mu}{\Lambda(1-\alpha^*)^2}\frac{(\frac{\Lambda}{\mu})^2N(3N-2-4\alpha^*)(1-\alpha^*) + N^3(N-1)\frac{1}{(1-\alpha^*)^2} + (\frac{\Lambda}{\mu})^3N(1-\alpha^*)^3 + 3\frac{\Lambda}{\mu}N^2(N-1)\frac{1}{1-\alpha^*}}{[\frac{N}{1-\alpha^*} + \frac{\Lambda}{\mu}(1-\alpha^*)]^3}\\ \geq& 0,\end{aligned}$$

which implies that $\alpha^*(N,\Lambda)$ increases in $\Lambda$ and $\lim_{\Lambda\to\infty}\alpha^*(N,\Lambda)$ exists by the monotone convergence theorem. If $\bar{\alpha} = \lim_{\Lambda\to\infty}\alpha^*(N,\Lambda) < 1$, then (33) fails when $\Lambda$ is sufficiently large, which causes a contradiction. Therefore, $\bar{\alpha} = \lim_{\Lambda\to\infty}\alpha^*(N,\Lambda) = 1$.

(4). When $\frac{\Lambda}{N} = \lambda$ is held fixed, the first order condition (33) and the partial derivative $\frac{\partial^2\pi}{\partial\alpha\partial N}(\alpha; N, \lambda)$ can be rewritten as

$$\pi'(\alpha^*; N, \lambda) = r\frac{N\lambda}{\mu} - c\Big[\frac{N\lambda}{\mu} + (N-1)\frac{1}{(1-\alpha^*)^2} + \frac{(\frac{\lambda}{\mu})^2 + \frac{\lambda(1-2\alpha^*)}{\mu(1-\alpha^*)^2}}{[\frac{1}{1-\alpha^*} + \frac{\lambda}{\mu}(1-\alpha^*)]^2}\Big] = 0 \tag{35}$$

and

$$\frac{\partial^2 \pi}{\partial \alpha \partial N}(\alpha; N, \lambda) = r\frac{\lambda}{\mu} - c\frac{\lambda}{\mu} - c\frac{1}{(1-\alpha)^2}.$$

By substituting (35) into $\frac{\partial^2 \pi}{\partial \alpha \partial N}(\alpha; N, \lambda)$, we know

$$\frac{\partial^2 \pi}{\partial \alpha \partial N}(\alpha^*; N, \lambda) = -\frac{c}{N}\frac{\frac{\lambda}{\mu}(1+2\alpha) + \frac{1}{(1-\alpha)^2}}{[\frac{1}{1-\alpha} + \frac{\lambda}{\mu}(1-\alpha)]^2(1-\alpha)^2} \leq 0.$$

Therefore, $\alpha^*(N, \Lambda)$ decreases in $N$ when $\frac{\Lambda}{N}$ is held fixed.

Given $\frac{\Lambda}{N} = \lambda$, the firm earns a positive profit if and only if (by rewriting property 1)

$$N(r\frac{\lambda}{\mu} - c\frac{\lambda}{\mu} - c) > -\frac{c\mu}{\mu + \lambda},$$

whick holds for $N \leq \frac{c\mu^2}{(\mu+\lambda)(r-c)(\frac{c\mu}{r-c}-\lambda)}$ if $\lambda < \frac{c\mu}{r-c}$ and for all $N > 0$ otherwise. In the latter case, $\lim_{N\to\infty} \alpha^*(N, \Lambda; \frac{\Lambda}{N} = \lambda)$ exists by the monotone convergence theorem. Let $N \to \infty$ in equality (35), we obtain $\lim_{N\to\infty} \alpha^*(N, N\lambda) = 1 - \sqrt{\frac{c\mu}{(r-c)\lambda}}$.

□

# Appendix D Additional Discussion of the Single Location Case

## D.1 A Correction Term

Recall that per Proposition 6 our approximation has a minor shortcoming when $N = 1$ since it does not approach infinity as $\alpha \longrightarrow 1$. Thus, a question that naturally arises is whether we should add a correction term (a fourth term) to our approximation (15) for $N = 1$ to ensure that the approximation is consistent with the fact that the minimal fleet size grows to infinity as $\alpha$ approaches 1. If we do so, this correction term would be noticeable only when $N = 1$ and $\alpha$ is nearly 1. Note that such a term cannot be expressed in algebraic form because it grows slower than $\frac{1}{(1-\alpha)^\delta}$ for any $\delta > 0$ as $\alpha \longrightarrow 1$ per Proposition 6. In most practical cases, adding this correction term would not be necessary. Extensive numerical experiments confirm that the correction would be small for $\alpha \leq 0.99$ no matter how large $\Lambda$ is. The largest value for the difference $|K(\alpha, \Lambda, \mu, 1) - \hat{K}(\alpha, \Lambda, \mu, 1)|$ is 31.97, which is observed to occur when $\frac{\Lambda}{\mu} = 1826$ and $\alpha = 0.99$. In this case, $K(\alpha, \Lambda, \mu, 1)$ is very large so that the percentage error is rather negligible. If we were to restrict our attention to $\alpha \leq 0.95$, the largest gap reduces to 6.90.

It is important to note that most of the existing approximations in the literature share a similar limitation. In particular, neither (H.39) nor (H.40) increases to infinity as $\alpha \longrightarrow 1$ (i.e., even the best approximations in the literature suffer from this relatively minor drawback).

If a correction term must be included, we propose adding the following term:

$$\kappa(\alpha, \Lambda, \mu, N) = \ln(1 + \frac{\Lambda}{\mu}) \ln(1 + \alpha) \ln\Big(\frac{\frac{N}{(1-\alpha)^2} + \frac{\Lambda}{\mu}}{\frac{N-1}{(1-\alpha)^2} + \frac{\Lambda}{\mu}}\Big),$$

resulting in a modified expression for $\hat{K}(\alpha, \Lambda, \mu, N)$ given by:

$$\hat{K}(\alpha, \Lambda, \mu, N) = \frac{\Lambda}{\mu}\alpha + (N-1)\frac{\alpha}{1-\alpha} + \frac{\frac{\Lambda}{\mu}\alpha}{\frac{N}{1-\alpha} + \frac{\Lambda}{\mu}(1-\alpha)} + \kappa(\alpha, \Lambda, \mu, N). \tag{36}$$

Note that $\kappa(\alpha, \Lambda, \mu, N)$ satisfies the property that it grows at a smaller rate than that of any power series $O((1-\alpha)^{-\delta})$ for $\delta > 0$. It also satisfies the property $\kappa(\alpha, \Lambda, \mu, N) \longrightarrow 0$ under the five regimes considered in Proposition 3 and $\kappa(\alpha, \Lambda, \mu, 1) \longrightarrow \infty$ as $\alpha \longrightarrow 1$.

For the same numerical example considered in Section 5.2 (i.e., $N = 1$, $\mu = 1$, $\Lambda = 1i$, $i = 1, \cdots, 1000$, and $\alpha = 0.03j$, $j = 1, \cdots, 33$), we find that our modified approximation (36) (with a mean absolute error of 0.86 and a relative error of 1.02%) is comparable to the best one, (H.40) (with a mean absolute error of 0.64 and a relative error of 1.15%), from the literature.

## D.2 Comparisons with the Approximations of the Inverse Erlang loss formula in Berezner et al. (1998) and Harel (2010)

In this section, we list notable approximations of the inverse of the Erlang Loss formula derived by Berezner et al. (1998) and Harel (2010). For ease of reference, we rewrite these approximations using our notation, and add an initial "B" to equation numbers when we refer to expressions from Berezner et al. (1998) and an initial "H" when we refer to expressions from Harel (2010). Using the same numerical examples used by Berezner et al. (1998) and Harel (2010), we compare the performance of our approximations against theirs.

*Bounds from Berezner et al. (1998):*

$$K(\alpha,\Lambda,\mu,1) < \frac{\Lambda}{\mu}\alpha + \frac{1}{1-\alpha}, \tag{B.5}$$

$$K(\alpha,\Lambda,\mu,1) > \frac{\Lambda}{\mu}\alpha, \tag{B.6}$$

$$K(\alpha,\Lambda,\mu,1) > \frac{\Lambda}{\mu}\alpha + (\frac{1}{1-\alpha} - 1) - \frac{3\mu}{\Lambda(1-\alpha)^3} - \alpha^{\frac{\Lambda}{\mu}\alpha}\left(\frac{2}{1-\alpha} + \frac{\Lambda}{\mu}\alpha\right), \textit{ and} \tag{B.21}$$

$$K(\alpha,\Lambda,\mu,1) > \max\{(\text{B.6}),(\text{B.21})\}. \tag{B.621}$$

*Bounds from Harel (2010):*

$$K(\alpha,\Lambda,\mu,1) < \left(\frac{\Lambda}{2\mu} + \frac{1}{2(1-\alpha)}\sqrt{(\frac{\Lambda}{\mu})^2(1-\alpha)^2 + 4\frac{\Lambda}{\mu}(1-\alpha)}\right)\alpha, \quad K \geq 2, \Lambda > 0, \tag{H.35}$$

$$K(\alpha,\Lambda,\mu,1) > \frac{\Lambda}{\mu} - \frac{1}{2} - \frac{3\Lambda}{2\mu}(1-\alpha) + \frac{\sqrt{4\frac{\Lambda}{\mu} + \left[\frac{\Lambda}{\mu}(1-\alpha) - 1\right]^2}}{2}, \quad K \geq 1, \Lambda > 0, \tag{H.36}$$

$$K(\alpha,\Lambda,\mu,1) \approx \frac{\Lambda}{\mu}\alpha\frac{2 + \frac{\Lambda}{\mu}(1-\alpha)}{1 + \frac{\Lambda}{\mu}(1-\alpha)}, \textit{ and} \tag{H.39}$$

$$K(\alpha,\Lambda,\mu,1) \approx \frac{\Lambda}{\mu} - 2\frac{\Lambda}{\mu}(1-\alpha) - 1 + \sqrt{(\frac{\Lambda}{\mu})^2(1-\alpha)^2 + 2\frac{\Lambda}{\mu} + 1}. \tag{H.40}$$

We first compare our approximations (15) and (36) against those in Berezner et al. (1998) using the same numerical example considered in their Table 1. Since Berezner et al. (1998) prove that their bounds are strict, they use a strict ceiling of (B.621) and a strict floor of (B.5) as the lower and upper bounds for $K(\alpha)$, respectively. The results in Table 2 show that our approximations (15) and (36) consistently perform better than (B.621) and (B.5).

Next, we provide comparisons against those in Harel (2010) using the same numerical example considered in their Tables 2, 3, and 4. Noting that Harel (2010) treats $K$ as a real value $K$, we follow their setting and keep decimal parts in our approximations. One can see that all the approximations (ours and theirs) perform

| $\alpha$ | $\frac{\Lambda}{\mu}$ | (B.621) | $K(\alpha)$ | (B.5) | $\lceil\hat{K}\rceil$ | $\lceil(36)\rceil$ |
|---|---|---|---|---|---|---|
| 0.99 | 1000 | 991 | 1029 | 1089 | 999 | 1011 |
| | 10000 | 9901 | 9970 | 9999 | 9950 | 9954 |
| | 100000 | 99070 | 99092 | 99099 | 99090 | 99091 |
| | 1000000 | 990097 | 990099 | 990099 | 990099 | 990099 |
| | 10000000 | 9900099 | 9900099 | 9900099 | 9900099 | 9900099 |
| 0.999 | 1000 | 1000 | 1072 | 1998 | 1000 | 1034 |
| | 10000 | 9991 | 10170 | 10989 | 10000 | 10030 |
| | 100000 | 99901 | 100293 | 100899 | 99991 | 100010 |
| | 1000000 | 999001 | 999697 | 999999 | 999500 | 999507 |
| | 10000000 | 9990700 | 9990925 | 9990999 | 9990909 | 9990910 |

Table 2: Comparisons with the approximations in Table 1 of Berezner et al. (1998)

well when $\alpha$ is not too close to 1. When $\alpha$ is small, all the approximations produce exact values. Significant gaps are observed for the approximations when $\alpha = 0.99$ and for some when $\alpha = 0.9$. We further look into the case of $\alpha = 0.99$ for different values of $\frac{\Lambda}{\mu}$. First, (H.35) significantly overestimates the exact value for $\alpha = 0.99$. When $\frac{\Lambda}{\mu} = 10$ and $\alpha = 0.99$, (H.39) performs the best. When $\frac{\Lambda}{\mu} = 100$ and $\alpha = 0.99$, our approximation (36) performs the best. When $\frac{\Lambda}{\mu} = 1000$ and $\alpha = 0.99$, (H.40) performs the best.

| $\alpha$ | (H.35) | Exact | (H.40) | (H.39) | (H.36) | $\hat{K}$ | (36) |
|---|---|---|---|---|---|---|---|
| 0.99 | 36.65 | 17.44 | 13.38 | 18.90 | 12.54 | 10.00 | 21.40 |
| 0.90 | 14.56 | 12.53 | 11.69 | 13.50 | 11.16 | 9.82 | 13.51 |
| 0.80 | 10.93 | 10.27 | 10.00 | 10.67 | 9.70 | 9.14 | 10.91 |
| 0.70 | 8.85 | 8.58 | 8.48 | 8.75 | 8.32 | 8.11 | 9.06 |
| 0.60 | 7.24 | 7.12 | 7.08 | 7.20 | 7.00 | 6.92 | 7.47 |
| 0.50 | 5.85 | 5.80 | 5.78 | 5.83 | 5.74 | 5.71 | 6.04 |
| 0.40 | 4.58 | 4.56 | 4.55 | 4.57 | 4.53 | 4.52 | 4.72 |
| 0.30 | 3.38 | 3.37 | 3.37 | 3.38 | 3.36 | 3.36 | 3.47 |
| 0.20 | 2.22 | 2.22 | 2.22 | 2.22 | 2.22 | 2.22 | 2.28 |
| 0.10 | 1.10 | 1.10 | 1.10 | 1.10 | 1.10 | 1.10 | 1.13 |
| 0.01 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 | 0.11 |

Table 3: Comparisons with the results in Table 2 of Harel (2010) for $\frac{\Lambda}{\mu} = 10$

Hence, none of the approximations consistently outperforms the others when $\alpha = 0.99$. In general, (H.40) and (36) provide the best performance. While (H.40) and (15) appear to underestimate the exact values, (36) overestimates by a very small amount except for the case of $\alpha = 0.99$. In summary, all the

| $\alpha$ | (H.35) | Exact | (H.40) | (H.39) | (H.36) | $\hat{K}$ | (36) |
|---|---|---|---|---|---|---|---|
| 0.99 | 160.19 | 116.88 | 111.21 | 148.50 | 108.00 | 99.98 | 114.64 |
| 0.90 | 98.24 | 96.25 | 96.35 | 98.18 | 95.47 | 94.50 | 96.55 |
| 0.80 | 83.82 | 83.42 | 83.52 | 83.81 | 83.29 | 83.20 | 83.81 |
| 0.70 | 72.26 | 72.14 | 72.18 | 72.26 | 72.11 | 72.10 | 72.36 |
| 0.60 | 61.46 | 61.42 | 61.44 | 61.46 | 61.41 | 61.41 | 61.54 |
| 0.50 | 50.98 | 50.96 | 50.97 | 50.98 | 50.96 | 50.96 | 51.03 |
| 0.40 | 40.66 | 40.65 | 40.65 | 40.66 | 40.65 | 40.65 | 40.69 |
| 0.30 | 30.42 | 30.42 | 30.42 | 30.42 | 30.42 | 30.42 | 30.44 |
| 0.20 | 20.25 | 20.25 | 20.25 | 20.25 | 20.25 | 20.25 | 20.26 |
| 0.10 | 10.11 | 10.11 | 10.11 | 10.11 | 10.11 | 10.11 | 10.12 |
| 0.01 | 1.01 | 1.01 | 1.01 | 1.01 | 1.01 | 1.01 | 1.01 |

Table 4: Comparisons with the results in Table 3 of Harel (2010) for $\frac{\Lambda}{\mu} = 100$

| $\alpha$ | (H.35) | Exact | (H.40) | (H.39) | (H.36) | $\hat{K}$ | (36) |
|---|---|---|---|---|---|---|---|
| 0.99 | 1080.69 | 1028.85 | 1024.84 | 1080.00 | 1016.44 | 999.00 | 1010.40 |
| 0.90 | 908.91 | 908.32 | 908.55 | 908.91 | 908.24 | 908.18 | 908.60 |
| 0.80 | 803.98 | 803.91 | 803.94 | 803.98 | 803.90 | 803.90 | 804.00 |
| 0.70 | 702.33 | 702.31 | 702.32 | 702.33 | 702.31 | 702.31 | 702.35 |
| 0.60 | 601.50 | 601.49 | 601.49 | 601.50 | 601.49 | 601.49 | 601.51 |
| 0.50 | 501.00 | 501.00 | 501.00 | 501.00 | 501.00 | 501.00 | 501.01 |
| 0.40 | 400.67 | 400.66 | 400.67 | 400.67 | 400.66 | 400.66 | 400.67 |
| 0.30 | 300.43 | 300.43 | 300.43 | 300.43 | 300.43 | 300.43 | 300.43 |
| 0.20 | 200.25 | 200.25 | 200.25 | 200.25 | 200.25 | 200.25 | 200.25 |
| 0.10 | 100.11 | 100.11 | 100.11 | 100.11 | 100.11 | 100.11 | 100.11 |
| 0.01 | 10.01 | 10.01 | 10.01 | 10.01 | 10.01 | 10.01 | 10.01 |

Table 5: Comparisons with the results in Table 4 of Harel (2010) for $\frac{\Lambda}{\mu} = 1000$

approximations in Harel (2010) and our approximations (15) and (36) perform well as long as $\alpha$ is not too close to 1.

## References for Appendices

Chandy, K., U. Herzong, and L. Woo (1975). Parametric analysis of queueing networks. *IBM Journal of Research and Development 19*, 36.

Shanthikumar, J. G. and D. D. Yao (1988). Second-order properties of the throughput of a closed queueing network. *Mathematics of Operations Research 13*(3), 524–534.