Approximating the Performance of a “Last Mile” Transportation System

Hai Wang and Amedeo Odoni

Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA 02139

{haiwang@mit.edu, arodoni@mit.edu}

Abstract: The Last Mile Problem (LMP) refers to the provision of travel service from the nearest public transportation node to a home or office. We study the supply side of this problem in a stochastic setting, with batch demands resulting from the arrival of groups of passengers who request last-mile service at urban rail stations or bus stops. Closed-form approximations are derived for the performance of Last Mile Transportation Systems as a function of the fundamental design parameters of such systems. An initial set of results is obtained for the case in which a fleet of vehicles of unit capacity provides the Last Mile service and each delivery route consists of a simple round-trip between the rail station or bus stop and a single passenger’s destination. These results are then extended to the general case in which the capacity of a vehicle is a small number (up to 20). It is shown through comparisons with simulation results that the approximations perform consistently well for a broad and realistic range of input values and conditions. These expressions can therefore be used for the preliminary planning and design of Last Mile Transportation Systems, especially for determining approximately resource requirements, such as the number of vehicles/servers needed to achieve some pre-specified level of service, as measured by the expected waiting time until a passenger is picked up from the station or delivered to her destination.
Keywords: Last mile problem; queueing; batch demands; waiting time approximation; cyclic assignment, vehicle routing.

1. Background and Literature Survey

The Last Mile Problem (LMP) refers to the provision of travel service from a public transportation node to home or workplace (“last mile”) or vice versa (“first mile”). This public transportation node could be the nearest rapid transit rail station or a stop of a scheduled bus line. The unavailability of this type of service is one of the main deterrents to the use of public transport in urban areas, especially for certain demographic groups, such as schoolchildren, seniors and people with certain physical handicaps. Currently, the default solutions to the LMP are walking, taking a taxi, or driving a private vehicle.

A conceptual Last Mile Transportation System (LMTS) is described schematically in Figure 1, which shows an urban area surrounding a public-transit rail station, where trains arrive and discharge passengers. The passengers’ final destinations (homes, offices and workplaces) are distributed in the area. A fleet of vehicles transports these passengers to their eventual destinations and empty vehicles return to the station to pick up waiting passengers or newly arriving ones. We describe the setting in more detail in Section 2.

Figure 1: Schematic of a Last Mile Transportation System (LMTS)
Many issues must be addressed when designing and operating a LMTS. On the supply side, it is essential to deal with difficult questions concerning the stochastic aspects of the system. The demand side requires an understanding and estimation of the potential LMTS loads as a function of demographic characteristics, nature of trip, level of service, cost, etc. The focus of this paper is solely on the supply side: given a probabilistic description of demand, design a LMTS that operates under dynamic and stochastic conditions according to certain guidelines and satisfies a set of Level of Service (LOS) requirements. This implies specifying such system characteristics as vehicle fleet size, service frequency, vehicle dispatching strategies, vehicle routing strategies, monitoring and control of operations, etc.

Addressing these questions is difficult analytically, as the planning and management of a LMTS generally involves such complications as: stochastic travel times; batch arrivals of prospective passengers; partitioning of demands among vehicles; routing of the vehicles; queueing issues; and, obviously, numerous considerations concerning staffing and economic sustainability. With the exception of staffing and economic issues, these complications are addressed in this paper in a static setting.

An extensive literature in this general area has generated various models for a number of application contexts related to the LMP with early papers dating back to the 1960s. We mention here only a few that are among the most influential in the field and especially relevant to the approach we have adopted. The Dynamic Traveling Repairman Problem (DTRP) was introduced in two papers by Bertsimas and Van Ryzin. They consider the DTRP in the cases of a single-vehicle “fleet” (1991) and of multiple vehicles (1993). The Dynamic Pick-up and Delivery Problem (DPDP) was studied by Swihart and Papastavrou
(1999), who derived bounds on the performance of several DPDP variants for light and heavy traffic. The Car Pooling Problem (CPP), introduced by Baldacci et al. (2004), also has features similar to the LMP – or, more exactly, to the First Mile Problem. The paper presents both exact and heuristic methods for solving the CPP based on integer programming formulations. Last mile supply chains are studied by Boyer et al. (2009). They examine the effects of two factors, customer density and delivery window length, on the overall efficiency of “last mile routes” for package deliveries. Balcik et al. (2008) consider a different vehicle-based last mile distribution system, this one geared to the needs of humanitarian relief chains. They propose a mixed integer programming model to determine delivery schedules for vehicles and to allocate resources equitably in the face of certain types of constraints. Personal rapid transit (PRT), which refers to a variety of transportation systems with characteristics similar in some ways to the last mile transportation system studied in this paper, has attracted significant attention in recent years. Papers considering PRT systems from a range of perspectives (all different from those presented here) include those by Anderson (1998), Bly and Teychenne (2005), Lees-Miller et al. (2009, 2010), Berger et al. (2011), and Mueller and Sgouridis (2011). Finally, a large number of papers have dealt with the Dial-a-Ride Problem (DARP) and related variations – see, e.g., Jaw et al. (1986), Lei et al. (2012). A good critical review of the DARP literature by Cordeau and Laporte (2007) underlines, among other points, the fact that this body of work does not address well some of the queueing aspects of the subject systems – a deficiency that this paper tries to remedy.

Similarities also exist between the LMP, as studied here, and various queueing, dispatching, routing, and resource allocation problems arising in entirely different
contexts, such as the design of manufacturing systems, the operation of elevator banks, and the scheduling of school-bus systems. The major difference between the LMP and these more “traditional” problems is that, in the LMP, passengers arrive in (possibly large) batches, not singly. Moreover, the size of these batches is a random variable. Queueing systems with batch arrivals are notoriously difficult analytically. A further complication is that the “service times” of passengers are determined by the length (or the duration) of the routes traveled by each vehicle. Thus, in designing a LMMS, it is necessary to consider simultaneously the problems of: allocating passengers among vehicles; routing the vehicles and estimating the lengths of the routes; and computing the queueing performance characteristics of the system.

The main body of this paper is organized as follows. In Section 2, we describe in detail the version of the LMP that we are studying and discuss the associated fundamental assumptions. Section 3 outlines our overall approach: we begin by deriving a set of queueing results by considering a fleet of vehicles with capacity for a single passenger \((c = 1)\) and then extend the analysis by allowing the vehicle capacity to be arbitrary and by incorporating the resulting travel time estimates into the queueing expressions derived for the \(c = 1\) case. Section 4 presents our analysis and results for the unit-capacity case. We derive an upper bound and an approximate expression for the performance of a LMMS as a function of its design parameters and then show through a set of simulation experiments that the resulting estimates approximate well the observed waiting times. Section 5 examines the general capacity case \((c > 1)\) by, first, proposing approximate analytical expressions for the expected value and the variance of the travel times associated with fleets consisting of vehicles with general capacity, and then applying
these expressions to the queueing approximation derived in Section 4. The results again compare well with those obtained from simulations. Section 6 contains a summary and concluding remarks.

2. Problem Description and Assumptions

We now describe in more detail the LMP scenario of Figure 1. The Last Mile Transportation System (LMTS) operates as follows: let STA be the transit rail station served by the LMTS. Consider a passenger, PAX, who boards a train at any station ("ORIGIN") for the purpose of traveling to STA and will then board a LMTS vehicle for transport to her home. PAX is required to provide advance notice to LMTS of her impending arrival at STA. The time interval between the advance notice and the actual arrival of PAX at STA is of the order of several minutes (e.g., at least 5 or 10 minutes) to give the LMTS system sufficient time to plan the service of PAX. In practical terms, the advance notice could be generated by PAX in a number of alternative ways. For example, PAX could use a smart-phone when she arrives at ORIGIN or when she enters her train to STA; or, she could tap a smart card on a special-purpose screen, as she is entering ORIGIN or while aboard the train. The resulting message to the LMTS includes the expected time of arrival of PAX at STA (easy to predict, once the passenger is at the ORIGIN station or aboard a train) and her ultimate destination, e.g., her home address. If the great majority of LMTS users are subscribers whose home addresses are pre-registered, then the only information that PAX will have to provide will be an identification number or code.

Once the information about PAX is received the LMTS will assign PAX to one of the vehicles of the LMTS fleet, plan the route of that vehicle so it includes a visit to the
ultimate destination of PAX, estimate the departure time of the vehicle from STA, and notify PAX accordingly. PAX will receive a message (on her smart-phone or by tapping her card on a screen when she arrives at STA) that indicates the vehicle she has been assigned to and the planned departure time of the vehicle from STA (e.g., “please board Vehicle 123 which will depart from STA at 4:26 PM”). Once all the passengers assigned to a vehicle are on board, the vehicle will execute a delivery route, visiting the destination of each of the passengers and will then return to STA to pick up the passengers for its next delivery tour.

The LMTS described above possesses the generic system features that we are most interested in: arrivals of passengers in “batches” (groups) at STA; clustering of passengers into subgroups for assignment to a fleet of vehicles; routing of the vehicles to deliver the passengers on board; and a requirement for fast computation of waiting times and other performance parameters so that, for example, passengers can be notified in a timely way of the departure time of the vehicle they have been assigned to. Actual implementations would probably involve some simpler variants of the above features.

Given the service region’s geometry, passenger demand, the spatial distribution of the passenger destinations, and the number, capacity and travel speed of the LMTS vehicles, examples of performance metrics that we wish to compute include: the average waiting time until boarding a vehicle, the average riding time of passengers, the average waiting time until delivery, the minimum number of vehicles we need to reach stable operation, vehicle productivity and workload, and eventually (but not in this paper) the general cost of operating the system and the service vs. cost trade-offs involved.
We now identify briefly the specifics of the model considered. With reference to Figure 1, we make the following assumptions: (i) headways, $h$, between arrivals of trains at the station (and discharges of passengers) are constant; (ii) passengers are discharged in batches after each train’s arrival; (iii) the batch size is a general random variable, $N$, with known expectation $E(N) = n$ and variance $Var(N) = \sigma_N^2$; (iv) all passengers arriving in a single batch request service essentially simultaneously; (v) given the size of any particular batch, $N = N_0$, the destinations of the $N_0$ passengers in the batch are distributed identically, uniformly and independently in a service region; (vi) the service region is convex and compact with known dimensions; (vii) the delivery fleet (or pick-up fleet, in the case of “First Mile” service) consists of $m$ vehicles, each with integer capacity, $c$.

We believe that these assumptions are sufficiently general for approximating, to a first order, the characteristics of many potential variations of LMMS. Note that our model includes the most difficult, from the analytical point of view, features that one might encounter in an LMMS: batch arrivals, stochastic demand, stochastic service times, and the presence of queueing phenomena interfaced with clustering and routing problems.

3. Description of Overall Approach

Sections 4 and 5 of the paper describe in detail our analysis and results. In this section we provide a brief description of the overall approach we have followed to provide perspective for these detailed sections. We have adopted a viewpoint under which the LMMS is regarded as a spatially distributed queueing system. In line, with typical queueing terminology, we shall refer henceforth to passengers as “customers”. The $m$ parallel servers (the vehicle fleet) serve customers in groups of $c$ or smaller, where $c$ is
the capacity of each vehicle. The service time for each group is equal to the travel time associated with a vehicle tour that begins at the station/depot, visits each of the \( c \) (or fewer) customer destinations and returns to the station/depot to pick up a new group.

Because queueing systems with batch arrivals (like the arrivals of customers at STA) and batch services (like the service of groups of customers by each vehicle) are difficult to analyze, we resort to a two-step approach. In Step 1, we assume that \( c = 1 \), i.e., that the delivery vehicles have unit capacity. Thus, service times consist simply of the duration of a round-trip between STA and one customer’s destination (Figure 2), with the destination being randomly and uniformly distributed within the service area per our assumption (v) in Section 2. In this way, we obtain a \( D^N/G/m/\infty \) system in queueing theory notation, where \( D^N \) indicates batch arrivals at constant (“Deterministic”) intervals with the number of arriving customers in each batch described by random variable \( N \); \( G \) denotes the fact that the distribution of service times (i.e., the duration of the round trips between STA and customer destinations) is “general”; and \( m \) and \( \infty \) indicate, respectively, the number of service vehicles and the fact that no \textit{a priori} limit is placed on the number of customers waiting for pickup at STA.

![Diagram of customer destinations and vehicles routes](image)

Figure 2: Customer destinations and vehicles routes of the Unit-Capacity, Multi-Vehicle LMP
As no closed-form expressions are available for the fundamental quantities that describe the performance of a $D^N/G/m/\infty$ system, we then attempt to obtain expressions for similar queueing systems, which are more tractable mathematically. Through a series of simplifications, we derive (i) an upper bound and (ii) an approximate expression for the mean waiting time associated with $D^N/G/m/\infty$ queues. We then carry out an extensive set of simple simulation experiments and conclude that these expressions provide good estimates of the performance of the system (with $c = 1$) under a broad range of system design parameters.

Step 2 examines the general case, in which service times are equal to the duration of delivery tours consisting of $c(> 1)$ or fewer delivery stops, as shown in Figure 3. To apply to the general capacity case the queueing expressions that were derived in Step 1, we need to partition the customers, design near-optimal tours/routes for the vehicles, and compute the approximate expectation and variance of the vehicle tour length. We accomplish this by using arguments from geometrical probability and from the literature on the Traveling Salesman Problem and Vehicle Routing Problem. We then use these expressions, along with the queueing-based approximation derived in Step 1, to complete the process of estimating the performance of the LMTS for the general case of arbitrary fleet size and arbitrary vehicle capacity. Finally, we compare again our approximate estimates to the results of a series of simulations over a broad range of input values.
4. The Unit-Capacity, Multi-Vehicle LMP

In this section we present the analysis of the case described in Section 3 as Step 1, in which \( c = 1 \), and \( m \) is an arbitrary positive integer. As already indicated (Figure 2), the length of the vehicle trips in this case is equal to two times the distance between the rail station and a customer’s destination. If we postulate constant and unit travel speed, the expressions for travel times are identical with those derived for travel distances.

The basic notation is summarized as follows:

\[
\begin{align*}
h & = \text{the constant headway between arrivals of trains (and discharges of customers) at the station STA;} \\
N & = \text{a random variable denoting the number of LMTS customers ("batch size") discharged after the arrival of a train at STA, with the sizes of successive batches being mutually independent, and } E(N) = n \text{ and } Var(N) = \sigma_N^2 \text{ denoting, respectively, the expectation and variance of } N; \\
\lambda_a & = \text{the arrival rate of customer batches (} = 1/h); \\
\end{align*}
\]
\[ \sigma_a^2 = \text{the variance of the inter-arrival time of customer batches} \quad (\sigma_a^2 = 0 \text{ for constant headways}); \]

\[ S = \text{a random variable denoting the service time of a random LMTS customer with} \]

\[ E(S) = s \text{ and variance } Var(S) = \sigma_s^2. \]

Note that the successive service times of any given vehicle in the fleet are independent and identically distributed. The traffic load (or utilization ratio) is given by \( \rho = ns/hm \), since \( m/s \) is the service rate of the LMTS, while \( n/h \) is the rate of customer arrivals per unit of time.

We are particularly interested in the expected waiting time, \( W_q \), of LMTS customers until they board one of the \( m \) vehicles to be transported to their eventual destination. Determining this expected waiting time as a function of the LMTS design parameters is a critical step toward developing the means to design LMTS satisfying certain level-of-service requirements.

### 4.1 General Upper Bound and Approximation

We begin by obtaining a general upper bound and approximate expression for \( W_q \) in the original Unit-Capacity, Multi-Vehicle \( D^N/G/m/\infty \) model. To do this, for each train’s arrival, we pre-assign the discharged customers to different vehicles and then construct a corresponding single-server queueing model \( D^{N_s}/G/1/\infty \) for each vehicle, where \( N_s \) is the random variable indicating the number of customers from any single train assigned to the same vehicle. Each customer can be served only by the vehicle to which she has been pre-assigned.
With such an assignment policy, service inefficiencies will exist since a customer is required to wait for his or her assigned vehicle, even when other vehicles may be available. Thus, the average waiting time in this case will be larger than the average waiting time in the original model. The customer flow is shown schematically in Figure 4.

![Customer flow diagram](image)

**Figure 4:** Customer flow in the pre-assignment policy

The $D^{N_S}/G/1/\infty$ model is still difficult to work with. To obtain approximate expressions for $W_q$, we decompose the problem into two parts. First, the $N_S$ customers in a batch who are assigned to the same vehicle are treated as a single “macro-customer” $P$. If we only consider the “macro-customer”, this reduces the $D^{N_S}/G/1/\infty$ model to the more tractable $D/G/1/\infty$ model and allows us to obtain an approximation for $W_{q_1}$, the expected waiting time until the first customer in $P$ receives service.

Let $T$ be the service time of the “macro-customer”, $T = \sum_{i=1}^{N_S} S_i$, where $N_S$ depends on the assignment policy and $S_1, S_2, ..., S_N$ are the service times of the real customers, which are mutually independent and identically distributed. Note that $N_S$ is a random variable. Therefore,
\[ E(T) = \sum_{i=1}^{N_S} E(S_i) = E(N_S)s, \quad \text{Var}(T) = E(N_S)\sigma_s^2 + s^2\text{Var}(N_S), \]

\[ C_t^2 = \frac{E(N_S)\sigma_s^2 + s^2\text{Var}(N_S)}{E^2(N_S)s^2} \]

Additionally, \( \sigma_a^2 = 0 \) because of constant “macro-customer” inter-arrival times, \( \lambda_a = 1/h, \rho = E(T)/h = E(N_S)s/h. \) According to Kingman (1961), Kingman (1962), and Ott (1987), an upper bound for \( W_{q1} \), the expected waiting time of \( D/G/1/\infty \) queue is:

\[ W_{q1} = W_q(D/G/1/\infty) \leq \frac{\lambda_a(\sigma_a^2 + \sigma_f^2)}{2(1 - \rho)} = \frac{1}{h}(0 + \text{Var}(T)) = \frac{E(N_S)\sigma_s^2 + s^2\text{Var}(N_S)}{2(h - E(N_S)s)} \quad (1) \]

According to Kraemer et al. (1976), an approximation of \( W_{q1} \) is provided by:

\[ W_{q1} = W_q(D/G/1/\infty) \approx \frac{\text{Var}(T)}{2(h - E(T))} \cdot \exp \left[ -\frac{2(h - E(T))E(T)}{3\text{Var}(T)} \right] = \frac{E(N_S)\sigma_s^2 + s^2\text{Var}(N_S)}{2(h - E(N_S)s)} \cdot \exp \left[ -\frac{2(h - E(N_S)s)E(N_S)s}{3E(N_S)\sigma_s^2 + 3s^2\text{Var}(N_S)} \right] \quad (2) \]

In a second step, we then compute the additional expected waiting time, \( W_{q2} \), until each of the individual customers in macro-customer \( P \) receives service, following the service to the “macro-customer”. For the \( i th \) customer in \( P \), we consider the additional expected waiting time due to being preceded by \( i - 1 \) other customers in \( P \). If the macro-customer consists of \( k \) customers and \( k \geq 1 \), the customer in the \( i th \) position suffers the expected additional total waiting time \( W_{q,i \text{th}} = \sum_{j=1}^{i-1} s_j = (i - 1)s \), where \( s_j \) is the expected service time of the \( jth \) customer served before the \( ith \) customer. Let \( W_{q,k \text{ customers}} \) denote the expected total additional waiting time of the \( k \) customers:

\[ W_{q,k \text{ customers}} = \frac{\sum_{i=1}^{k} W_{q,i \text{th}}}{k} = \frac{\sum_{i=1}^{k} (i - 1)s}{k} = \frac{(k - 1)s}{2}, \quad k \geq 1 \]
If \( k = 0 \), no customers are served, so that \( W_{q,0} = 0 \). According to the Law of Total Expectation, the expected additional waiting time of a customer is then given by:

\[
W_{q2} = \frac{\sum_{k=0}^{\infty} P(k) W_{q,k \text{ customers}} k}{\sum_{k=0}^{\infty} P(k) k} = \frac{s\text{Var}(N_S) + sE^2(N_S) - sE(N_S)}{2E(N_S)} \tag{3}
\]

Thus the upper bound we seek is:

\[
W_q = W_{q1} + W_{q2} \leq \frac{E(N_S)\sigma_S^2 + s^2\text{Var}(N_S)}{2(h - E(N_S)s)} + \frac{s\text{Var}(N_S) + sE^2(N_S) - sE(N_S)}{2E(N_S)} \tag{4}
\]

The approximation, using (2) and \( W_{q2} \), is:

\[
W_q = W_{q1} + W_{q2} \approx \frac{E(N_S)\sigma_S^2 + s^2\text{Var}(N_S)}{2(h - E(N_S)s)} \cdot \exp \left[ -\frac{2(h - E(N_S)s)E(N_S)s}{3E(N_S)\sigma_S^2 + 3s^2\text{Var}(N_S)} \right] + \frac{s\text{Var}(N_S) + sE^2(N_S) - sE(N_S)}{2E(N_S)} \tag{5}
\]

Expression (4) and (5) are valid under general assumptions about the probability density functions of the batch size, \( N \), and the service times, \( S \). Moreover, (4) and (5) have been derived without considering how exactly customers are assigned to vehicles.

We next analyze one particular reasonable policy for customer assignment to vehicles. The policy will provide a modified \( D^{N_S}/G/1/\infty \) model with \( E(N_S) \) and \( \text{Var}(N_S) \), leading to corresponding expressions for \( W_{q1} \) and \( W_{q2} \), and, ultimately, to an upper bound and an approximation for \( W_q \).

### 4.2 Cyclic Assignment Policy

One possible policy for allocating customers to vehicles is to assign customers in cyclic order to the vehicles: the first customer in the batch is assigned to Vehicle 1, the second
to Vehicle 2, ..., the \((m + 1)\)th to Vehicle 1 again, and so forth. No jockeying of customers, after being assigned to vehicles, is allowed. Figure 5 illustrates this policy, which requires assigning an “identification number” to each vehicle to distinguish among them.

We have a total of \(m\) vehicles, labeled as “Vehicle 1”, “Vehicle 2”,..., “Vehicle \(m\)”. Let \(N_i\) be the random variable indicating the number of customers assigned to “Vehicle \(i\)” after the arrival of a particular train, with the assignment process upon arrival of each train being independent of the arrival process upon arrival of any other train. When one train arrives, we order the \(m\) vehicles in sequence: the vehicle that receives customers first is called “1st server”, the vehicle that receives customers second is called “2nd server”, etc. Let \(X_i\) be the random variable indicating the number of customers assigned to the “\(i\) th server” after the arrival of a particular train. Then,

\[
X_1 = \left\lfloor \frac{N + m - 1}{m} \right\rfloor, X_2 = \left\lfloor \frac{N + m - 2}{m} \right\rfloor, \ldots, X_{m-1} = \left\lfloor \frac{N + 1}{m} \right\rfloor, X_m = \left\lfloor \frac{N}{m} \right\rfloor
\]

\(N = X_1 + X_2 + \cdots + X_{m-1} + X_m\)
If we order the vehicles randomly, the probability that Vehicle \( i \) will become the \( j \)th server for some train is \( 1/m \). The modified model can be considered as \( D^{N_i}/G/1/\infty \). Since \( N_1, N_2, ..., N_m \) are identically distributed, all \( D^{N_i}/G/1/\infty \) models can be viewed as identical \( D^{N_i}/G/1/\infty \) models although \( N_1, N_2, ..., N_m \) are not necessarily independent.

Recalling that \( N \) is the random variable indicating the total number of customers coming from one train, let \( N = Km + R \), where \( K = \lfloor N/m \rfloor \), and \( R \) is the remainder after division of \( N \) by \( m \). We can therefore express \( N \) as a 2-dimensional random vector, \((K, R)\).

\[
X_i = \begin{cases} 
K + 1, & 1 \leq i \leq R; \\
K, & R + 1 \leq i \leq m;
\end{cases}
\]

\[
E(N_2|K, R)) = \frac{1}{m} [E(X_1|K, R)) + E(X_2|K, R)) + ... + E(X_m|K, R)) = \frac{N}{m};
\]

\[
Var(N_2|K, R)) = P(N_2 = K + 1)(K + 1 - E(N_2|K, R))^2 + P(N_2 = K)(K - E(N_2|K, R))^2
\]

\[
= \frac{R}{m} \cdot (K + 1 - \frac{Km + R}{m})^2 + \frac{m - R}{m} \cdot (K - \frac{Km + R}{m})^2 = \frac{Rm - R^2}{m^2}.
\]

\[
E(N_2) = E(E(N_2|K, R))) = E(\frac{N}{m}) = \frac{n}{m}
\]

\[
Var(N_2) = E(Var(N_2|K, R))) + Var(E(N_2|K, R))) = E\left(\frac{Rm - R^2}{m^2}\right) + Var\left(\frac{N}{m}\right)
\]

\[
= \frac{E(Rm - R^2)}{m^2} + \frac{\sigma^2_N}{m^2}
\]

Since \( R < m \), it is also true that \( Rm - R^2 \leq m^2/4 \), and

\[
Var(N_2) \leq \frac{1}{4} + \frac{\sigma^2_N}{m^2} = \frac{4\sigma^2_N + m^2}{4m^2}
\]

In practice, the number of customers \( N \) from each batch will typically be much larger than the number of vehicles \( m \), and the remainder \( R \) will tend to be uniformly distributed in \( \{0, 1, ..., m - 1\} \). Then,
By substituting the bound and approximation of \(E(N_s)\) and \(Var(N_s)\) into (4) and (5), respectively, the model corresponding to the cyclic assignment policy finally leads to the following upper bound and approximation for the case of a General service time distribution:

\[
W_q \leq \frac{4mn^2(a_N^2 + s^2) - 4n^3s^2 + 4hms(a_N^2 + n^2) + hm^3s - 4hm^2ns}{8mn(hm - ns)}
\]  
\(6\)

\[
W_q \approx \frac{6mn^2a_N^2 + 6s^2a_N^2 + m^2s^2 - s^2}{12m(hm - ns)} \cdot \exp \left[ - \frac{4(hm - ns)ns}{6mn^2a_N^2 + 6s^2a_N^2 + m^2s^2 - s^2} \right] + \frac{(6a_N^2 + m^2 + 6n^2 - 6mn - 1)s}{12mn}
\]  
\(7\)

Assuming the service area is a \(b \times b\) square with the train station located at the square’s center, the travel metric is right angle, and the travel speed is constant throughout the service region and equal to 1, and for Poisson batch sizes, the bound (6) becomes:

\[
W_q \leq \frac{14b^2mn^2 + 12b^2hn^2 - 12b^2n^3 + 12b^2hn - 12bhn - 3bmn}{24mn(hm - bn)}
\]  
\(8\)

The approximation for this special case is:

\[
W_q \approx \frac{(m + 6)b^2n + b^2m^2 - b^2}{12m(hm - bn)} \cdot \exp \left[ - \frac{4(hm - bn)n}{bmn + 6bn + bm^2 - b} \right] + \frac{(m^2 + 6n^2 + 6n - 6mn + 1)b}{12mn}
\]  
\(9\)

A different method can provide another approximation for the general case. The waiting time is decomposed into two parts: \(W_{q3}\), the waiting time until the first passenger in a batch receives service; and \(W_{q4}\), the waiting time until the following individual
customers in that batch receive service. We treat all the customers from one arrival batch as a single “macro-customer” \( P' \) and do not pre-assign them to vehicles. This reduces the \( D^N/G/m/\infty \) model to the \( D/G/m/\infty \) model and allows us to obtain an approximation for \( W_{q3} \) using approximations of the \( G/G/m/\infty \) model, such as those of Köllerström (1974) and Whitt (1993). The performance of approximations obtained in this way was found to perform worse than approximation (9), except in a few cases when both approximations perform poorly.

### 4.3 Numerical Experiments for the Unit-Capacity, Multi-Vehicle LMP

To assess the performance of the expressions obtained in Sections 4.1 and 4.2 under a broad range of conditions, a simple simulation of the Unit-Capacity, Multi-Vehicle LMP was carried out with a program written in java. We consider a square service region with geometry \( b/v = 2.5 \) min = 150 sec, headway \( h = 10 \) min = 600 sec, and Poisson-distributed batch sizes of \( n = 20, 40, 60, 80 \). We selected these parameters so that the system would make sense physically.

![Simulation Results](image)

Figure 6: Simulation results, bounds and approximations of average waiting time when \( n = 20 \)
Figure 7: Simulation results, bounds and approximations of average waiting time when $n = 40$

Figure 8: Simulation results, bounds and approximations of average waiting time when $n = 60$

Figure 9: Simulation results, bounds and approximations of average waiting time when $n = 80$
Figures 6 – 9 plot the simulation results and our estimates for the average waiting time per customer $W_q$ (in seconds) against the utilization ratio $\rho = sn/hm$. Since the simulated system has Poisson customer batch size and a square service region, the upper bound (expression (8)), and the approximation (expression (9)) from Sections 4.2 are applicable and considered here. For each demand intensity $n$, the utilization ratio $\rho$ takes on a set of discrete values because the number of vehicles, $m$, is integer. We have plotted the points with utilization ratio less than 0.9, above which the system is highly unstable and the average waiting time is too long to be accepted practically.

It can be seen that (8) is a consistently reliable upper bound for $W_q$, while (9) provides a very good approximation for the entire range of parameter values for which the LMTS remains stable. In a practical system, it would be desirable to achieve values of 1 to 5 minutes, for the average waiting time until customers board a vehicle. Note from Figures 6 – 9 that for this range of values (60 to 300 seconds) the difference between the approximation and the simulation results stays small in absolute or percentage terms. For example, when $n = 20$ (Figure 6), this difference never exceeds the greater of 15 seconds or 12% for values of $W_q$ between 1 and 4 minutes.

We have also performed simulation experiments with rectangular and diamond-shaped service regions and with discontinuities in the travel medium, such as an impenetrable barrier to travel. For these environments we have derived expressions for $W_q$, analogous to (8) and (9), based on (6) and (7) – see Wang (2012). These experiments led to the conclusion that the analytical upper bound and approximation continue to perform well under a wide range of conditions.
5. General-Capacity, Multi-Vehicle LMP: Approximations

In this section we shall generalize the results of Section 4 by considering the General-Capacity, Multi-Vehicle LMP, in which the vehicle capacity, $c$, and the number of vehicles, $m$, are arbitrary positive integers. The vehicles will now travel along more complicated routes than in the $c = 1$ case to deliver customers to their destinations. In practice, one would expect the vehicle capacity to be smaller than that of a regular bus – typically a number between 3, for service provided by taxi-like vehicles, and 20, for large vans.

As explained in Section 3, the General-Capacity, Multi-Vehicle LMTS will be viewed as a spatially distributed queueing system in which the service times are equal to the amount of time it takes to complete a customer delivery tour and return to the train station (Figure 3). After each batch of arrivals, the customers must be partitioned into clusters and assigned to vehicles according to their destinations and the vehicles must then be routed with the objective of obtaining a shortest total travel distance – which translates into shortest service times and smallest overall queueing. This means that the estimation of model parameters, such as the expected value and the variance of service times, is now far more complicated than when $c = 1$.

5.1 Adjustment of the Queueing Model

We first need to make some adjustments to the principal expression (7) that we have derived from our queueing model. For the General ($c > 1$) Capacity case, General distribution of customer batch size and General service times, the approximation for the waiting time until boarding a vehicle is given by:
The expression (10) is exactly the same as (7), except \( S \) is substituted by \( S_E \), the travel time to serve \( c \) customers, and \( N \) by \( N_E \), the random variable indicating the number of tours formed following the arrival of a batch of customers.

Note that in (10) we have used the notation \( W_{q,Board} \) for the expected waiting time until a customer will board a vehicle, while in (7) we used the notation \( W_q \) for the same quantity. This is because we also want to introduce here another quantity, \( W_{Riding} \), which is defined as the expected time a customer will spend riding on the vehicle before being delivered to her destination. Considering the riding component of the trip, the total expected time from the instant a customer arrives at the rail station until she is delivered at her destination is given by

\[
W_{Delivered} = W_{q,Board} + W_{Riding}
\]  

(11)

The expected riding time of the \( i \)th delivered customer in a tour with \( c \) customer deliveries is approximated by \( i \times E(S_E)/(c + 1) \) and the expected riding time of a random customer is \( E(S_E)/2 \).

5.2 Approximating the Expected Value of Customer Service Times

We turn next to the task of evaluating the performance of the general expressions (10) and (11). To do this, expressions must be developed for all terms involving \( S_E \) and \( N_E \). In
this subsection and the next two, we propose a set of such approximate expressions for the case in which the destinations of the customers are uniformly and independently distributed within a square $b \times b$ district, assuming Euclidean travel. An entirely different operating environment is examined in Section 5.5.

We start with the critical quantity $E(S_E)$, the expected travel time to deliver $c$ customers. Consider the situation in which $j$ customers are to be delivered by vehicles with capacity $c$ each within the district of interest in a minimum total amount of travel time. Vehicles must return to their origin (the train station). This is a classical Vehicle Routing Problem (VRP).

Eilon et al. (1971) proposed an empirical formula for $E(TVRT_{j,c})$, the total length of vehicle routing tours when a total of $j$ customers are delivered using vehicles of capacity $c$, but tested it for only up to $j = 70$ and $c = 10$. Daganzo (1984) provided another simple and intuitive analytical approximation:

$$E(TVRT_{j,c}) \approx \frac{2rj}{c} + 0.57\sqrt{ja} \quad (12)$$

where $r$ is the average distance between the customers and the depot and $A$ is the area of service region. For a $b \times b$ square region, uniformly distributed customer destinations, and the depot located at the center of the region, $r = 0.382b$ and expression (12) becomes:

$$E(TVRT_{j,c}) \approx 0.764\frac{j}{c}b + 0.57\sqrt{jb} \quad (13)$$

The expectation of a single route length of the vehicle routing tours $VRT_{j,c}$ can then be approximated as:

$$E(VRT_{j,c}) \approx \frac{E(TVRT_{j,c})}{j/c} \approx 0.764b + 0.57\frac{c}{\sqrt{j}}b \quad (14)$$
To assess the accuracy of (14), we took advantage of the fact that good heuristics exist for the VRP. Specifically, we simulated hundreds of thousands of instances of LMSTS train arrivals and associated customer destinations. To create the clusters and routes we applied two widely used VRP heuristics, the Sweep algorithm (coupled with a TSP heuristic) and the Clark-Wright algorithm. According to Cordeau et al. (2002), these fast and simple heuristics provided an average deviation of 6.71% and 7.09%, respectively, from the best solutions obtained on CMT benchmark instances. We solved all the simulated instances we generated using each of the two heuristics separately and, for each instance, we chose the better of the two solutions. Figures 10 shows the best solutions obtained for two examples, both with \( j = 40 \) customers but in one case with \( c = 10 \) and in the other with \( c = 4 \). As might be expected, the Sweep algorithm generated the solution shown on the left and Clark-Wright the one shown on the right. For a broad range of vehicle capacities, \( c \) (2 to 20), and number of routes \( j/c \) (2 to 20), the average error of (14), in absolute value terms, was of the order of 2%. Table 1 shows part of this assessment.
Table 1: Error of expression (14) compared to results of simulation

<table>
<thead>
<tr>
<th>$j/c$</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>5.90%</td>
<td>2.58%</td>
<td>1.48%</td>
<td>0.94%</td>
<td>0.52%</td>
<td>0.29%</td>
<td>-0.12%</td>
<td>-0.49%</td>
</tr>
<tr>
<td>5</td>
<td>6.02%</td>
<td>3.24%</td>
<td>2.40%</td>
<td>1.56%</td>
<td>0.79%</td>
<td>0.77%</td>
<td>0.68%</td>
<td>0.54%</td>
</tr>
<tr>
<td>6</td>
<td>5.68%</td>
<td>2.39%</td>
<td>1.97%</td>
<td>1.75%</td>
<td>1.40%</td>
<td>1.06%</td>
<td>0.82%</td>
<td>0.97%</td>
</tr>
<tr>
<td>7</td>
<td>5.04%</td>
<td>2.99%</td>
<td>2.14%</td>
<td>1.38%</td>
<td>1.31%</td>
<td>1.24%</td>
<td>1.24%</td>
<td>1.47%</td>
</tr>
<tr>
<td>8</td>
<td>4.29%</td>
<td>2.88%</td>
<td>1.86%</td>
<td>1.41%</td>
<td>1.25%</td>
<td>1.24%</td>
<td>1.41%</td>
<td>1.30%</td>
</tr>
<tr>
<td>9</td>
<td>5.20%</td>
<td>2.94%</td>
<td>1.79%</td>
<td>1.33%</td>
<td>1.29%</td>
<td>1.20%</td>
<td>1.34%</td>
<td>1.28%</td>
</tr>
<tr>
<td>10</td>
<td>4.39%</td>
<td>2.44%</td>
<td>1.58%</td>
<td>0.97%</td>
<td>0.92%</td>
<td>0.77%</td>
<td>0.87%</td>
<td>1.03%</td>
</tr>
<tr>
<td>11</td>
<td>4.36%</td>
<td>2.54%</td>
<td>1.73%</td>
<td>0.92%</td>
<td>0.90%</td>
<td>0.84%</td>
<td>0.90%</td>
<td>1.27%</td>
</tr>
<tr>
<td>12</td>
<td>4.09%</td>
<td>2.27%</td>
<td>1.48%</td>
<td>1.11%</td>
<td>0.84%</td>
<td>0.64%</td>
<td>0.64%</td>
<td>0.92%</td>
</tr>
<tr>
<td>13</td>
<td>4.18%</td>
<td>2.20%</td>
<td>1.55%</td>
<td>1.08%</td>
<td>0.64%</td>
<td>0.43%</td>
<td>0.53%</td>
<td>0.51%</td>
</tr>
<tr>
<td>14</td>
<td>4.05%</td>
<td>2.30%</td>
<td>1.29%</td>
<td>0.85%</td>
<td>0.75%</td>
<td>0.25%</td>
<td>0.38%</td>
<td>0.49%</td>
</tr>
<tr>
<td>15</td>
<td>4.12%</td>
<td>2.38%</td>
<td>1.58%</td>
<td>0.81%</td>
<td>0.65%</td>
<td>0.28%</td>
<td>0.27%</td>
<td>0.30%</td>
</tr>
</tbody>
</table>

In the queueing model, the service time, $S_E$, is the time that a vehicle takes to traverse a tour and deliver a group of $c$ customers. Estimates of the length of a VRT can be converted into time units, using information about the speed of travel in the region of interest. To simplify this conversion, we shall continue to assume here that travel speed is constant and equal to 1 throughout the region. Considering that the size, $j$, of a customer batch is sampled from the General distribution of a random variable $N$, we finally have:

$$E(S_E) \approx \frac{\sum_{j=0}^{\infty} P(N = j) \cdot \frac{j}{c} \cdot E(VRT_j,c)}{\sum_{j=0}^{\infty} P(N = j) \cdot \frac{j}{c}} = \frac{\sum_{j=0}^{\infty} P(N = j) \cdot j \cdot E(VRT_j,c)}{E(N)}$$

$$\approx \frac{\sum_{j=0}^{\infty} P(N = j) \cdot j \cdot \left(0.764b + 0.57 \frac{c}{\sqrt{j}} b\right)}{E(N)} = \frac{0.57cE(\sqrt{N})b + 0.764E(N)b}{E(N)}$$

$$= \frac{0.57cE(\sqrt{N})}{E(N)}b + 0.764b$$  (15)
5.3 Approximating the Variance of Customer Service Times

Our simulation experiments indicated that the coefficient of variation of $VRT_{j,c}$ is small, typically of the order of $0.1 - 0.25$ for most combinations of $c$ and $j/c$ values in the range of $c = 3 - 20$ and $j/c = 4 - 20$. Thus, the standard deviation of $VRT_{j,c}$ is small compared to its expected value. In addition, it is well known from queueing theory (and has been confirmed by the simulation experiments in the specific context of this paper) that the variance of the service times has only a secondary impact on expected waiting times because it does not affect the utilization ratio, $\rho$, of a queueing system. For these reasons, we can use simple approximations for the variance (or for the coefficient of variation) of $VRT_{j,c}$ without affecting by much the quality of the approximations obtained for the expected waiting time and expected time to delivery of (10) and (11).

Beardwood et al. (1959) proposed a famous asymptotic expression for the expectation of the length of a Traveling Salesman Tour (TST) with $k$ independent, uniformly distributed points in a square of area $A$ with Euclidean travel,

$$E(TST)_k \approx \beta_{1,k} \sqrt{kA}$$

(16)

For very large values of $k$, the best available estimate seems to be $\beta_{1,\infty} \approx 0.7124$ (Johnson 1996). Gremlich et al. (2004) have suggested that, the variance of the length of a TST with a large number of points can be approximated as,

$$Var(TST)_{\infty} \approx \beta_{2,\infty} A$$

(17)

where $\beta_{2,\infty} \approx 0.1385$. Let $C$ denote a coefficient of variation. Then, using (16) and (17), we shall use

$$C_{TST,k} = \frac{\sqrt{Var(TST)_k}}{E(TST)_k} \approx \frac{\sqrt{Var(TST)_{\infty}}}{E(TST)_{\infty}} \approx \frac{\beta_{2,\infty}}{\sqrt{k\beta_{1,\infty}^2}}$$

(18)
as an approximate expression for the coefficient of variation of the length of a TST.

For the case at hand, in which the number of points visited by vehicles with capacity \( c \) is \( c + 1 \), we shall use the further approximation

\[
C_{VRT,j,c} \approx C_{TST,c+1} \approx \frac{\beta_{2,\infty}}{\sqrt{(c + 1)\beta_{1,\infty}^2}}
\]  

(19)

on the premise that the variability of a VRT and a TST that visit the same number of points should be similar. This leads to

\[
Var(VRT_{j,c}) = E^2(VRT_{j,c}) \cdot C_{VRT,j,c}^2 \approx \frac{\beta_{2,\infty}}{(c + 1)\beta_{1,\infty}^2} E^2(VRT_{j,c})
\]  

(20)

and finally, for speed of travel equal to 1, to

\[
Var(S_E) \approx \frac{\beta_{2,\infty}}{(c + 1)\beta_{1,\infty}^2} E^2(S_E)
\]  

(21)

We tested the accuracy of (21) against the results of our simulations. Despite the rough nature of approximations (18) – (19) on which (21) is based, the observed average errors were of the order of only 30% for a broad range of values of the vehicle capacity \( c \) and the number of routes \( j/c \). As the next subsection indicates this is very adequate due to the limited impact of \( Var(S_E) \) on the value of the expected waiting time.

5.4 Simulation and Comparisons for the General-Capacity, Multi-Vehicle LMP

Expressions (10) and (11) will now be tested for the case in which the size of customer batches has a Poisson distribution with intensity \( n \). All the other assumptions (independent and uniform locations of customer destinations, Euclidean travel, square district with size \( b \), travel speed equal to 1) are the same as above.

Under the Poisson assumption \( E(\sqrt{N}) \approx \sqrt{n} \) in (15) and therefore,
\[
E(S_E) \approx \frac{0.57c\sqrt{n}}{n} b + 0.764b = \frac{0.57c}{\sqrt{n}} b + 0.764b
\]  

(22)

The variance of \( S_E \) is given in (21), while the various other terms of (10) and (11) take on the following values:

\[
E(N_E) \approx \frac{E(N)}{c} = \frac{n}{c} \]  

(23)

\[
VAR(N_E) \approx Sd^2 \left( \frac{N}{c} \right) \approx \left[ \sqrt{Var \left( \frac{N}{c} \right)} \right]^2 \approx \left[ \frac{\sqrt{n}}{c} \right]^2
\]  

(24)

\[
E(N_E^2) \approx \left[ \frac{\sqrt{n}}{c} \right]^2 + \left( \frac{n}{c} \right)^2
\]  

(25)

A simulation of a General-Capacity, Multi-Vehicle LMTS was performed with a program written in java. We consider a square service district with geometry \( a/v_x = b/v_y = 2.5 \text{ min} = 150 \text{ sec} \), headways between train arrivals of \( h = 10 \text{ min} = 600 \text{ sec} \), vehicle capacity \( c = 3-20 \) and customer arrivals with batch sizes described by a Poisson distribution with \( n = 40, 80 \) and 120. These parameters were selected so that the system would make sense physically. As before, vehicle tours were generated by using the two well-known vehicle routing heuristics, the Sweep algorithm and the Clark-Wright algorithm. Specifically, the simulation generated sets of points, uniformly and independently distributed in a \( b \times b \) square, and vehicle tours through these points were drawn using the better of the two solutions (shortest total length of the delivery tours).

Figures 11 through 18 present a sample of comparisons between the simulation results and the analytical approximations of Section 5.1 for the following respective cases: \( c = 5, n = 40, 80, 120; c = 10, n = 40, 80, 120; \) and \( c = 15, n = 120; c = 20, n = 120. \)
Figure 11: Simulation and analytical results when $c = 5$ and $n = 40$

Figure 12: Simulation and analytical results when $c = 5$ and $n = 80$

Figure 13: Simulation and analytical results when $c = 5$ and $n = 120$
Figure 14: Simulation and analytical results when $c = 10$ and $n = 40$

Figure 15: Simulation and analytical results when $c = 10$ and $n = 80$

Figure 16: Simulation and analytical results when $c = 10$ and $n = 120$
Figure 17: Simulation and analytical results when $c = 15$ and $n = 120$

Figure 18: Simulation and analytical results when $c = 20$ and $n = 120$

The horizontal axis in Figures 11 – 18 shows the utilization ratio $\rho = \frac{E(S_E)E(N_E)}{hm}$, while the vertical axis shows the expected waiting time until boarding a vehicle and the expected total waiting time spent between arrival at the station and delivery at the customer’s destination. “Approximation Until Boarding” is obtained from (10) and “Approximation Until Delivery” from (11). For each combination of vehicle capacity $c$ and demand intensity $n$, the utilization ratio $\rho$ takes on a set of discrete values because the number of vehicles, $m$, is integer.
We plot in Figures 11-18 the discrete points corresponding to utilization ratios less than 0.9. At values of \( \rho \) higher than 0.9 the system is highly unstable and the average waiting time is too long to be acceptable in practice. In addition, the LMTS may take a long time to approach steady state at such high values of \( \rho \). Odoni and Roth (1983) and Newell (1971), among others, have proposed expressions for an upper bound for and an approximation to the “relaxation time” to steady state of some quite general queuing systems. We have applied these expressions to the LMTS \( D^N/G/m \) queue and found that, for the typical sets of parameter values shown in Figures 11 – 18, the upper bound for the relaxation time must be less than or in the order of 25 minutes in all cases with utilization ratios of 0.85 or lower. This is significantly shorter than the duration of the time intervals (e.g., morning rush period, or evening rush period, or midday period) during which the respective demand rates for an LMTS system can be approximated as being roughly constant. It is therefore reasonable to use the steady state approximations for all but the highest (i.e., greater than 0.9) utilization ratios.

As shown in Figures 11-18, the approximate expression (10) for the expected waiting time until boarding a vehicle performs very well for both small and large vehicles and for the broad range of customer arrival intensities \( (n = 40, 80, \text{and } 120) \) examined. The difference between the simulated average time until boarding and the analytical expression (10) is of the order of the greater of 5% or 10 seconds. Turning to the estimation of expected total time until delivery, the analytical expression (11) also works well. The difference between the analytical and simulation results is now of the order of the greater of 5% or 15 seconds.
For similar LMTS with irregular service region and non-uniform demand, the VRT length expectation can be approximated following the method described by Daganzo (1984), and the variance can be approximated following the method described in Section 5.3. Expression (10) and (11) can be then used to estimate the system performance.

5.5 Another Test

As a second test of the performance of (10) and (11), we study a last mile transportation system that operates only along two main streets, which intersect at the location of a subway station, as shown in Figure 19. The destinations of passengers alighting from the subway are uniformly distributed along the two streets up to a distance of $b = 150$ from the station. LMTS shuttles, operating in each of the four directions emanating from the station, deliver arriving passengers to their eventual destinations.

![Figure 19: Schematic LMTS around Crossroad](image)

We have obtained approximate expressions for the expectation and variance of $S_e$ in this case as follows:
We have applied these expressions to (10) and (11) and compared the resulting analytical approximations with results from simulations. For practical waiting times (1 – 5 min), the difference between the simulated average time until boarding and the analytical approximation (10) is less than the greater of 15% and 15 seconds, while the difference between the simulated average time until delivery and the analytical approximation (11) is less than 15%. The approximations thus also work well for an environment that is very different from that of Sections 5.2 – 5.4.

6. Conclusion
This paper has developed a set of fully analytical expressions to support the approximate estimation of the performance of a quite general version of a Last-Mile Transportation System (LMTS). Given a lengthy list of inputs about the system’s characteristics (headways between arrivals of trains at the station, passenger batch size from each train, number of vehicles in the service fleet, capacity of vehicles, dimensions and travel-related properties of the urban district served), the expressions we have developed estimate the expected waiting time until a passenger can board a vehicle, and the expected time between arrival at the station and delivery to the passenger’s destination. A number of simple simulation experiments suggest that these expressions approximate well the expected performance of LMTS under a broad range of conditions typical of what one may encounter in practice.

On the methodological side, the principal contribution of this research is the development of approaches for bounding and approximating the performance of a very
difficult type of queueing system involving batch arrivals and requiring the simultaneous consideration of vehicle routing, queueing issues and the use of geometrical probability arguments. On the practical side, we believe that the analytical expressions we have developed can be very useful in designing LMTS, specifically in determining resource requirements for these systems, such as how many vehicles would be necessary to achieve a specified level of service (as measured by expected time until one boards a vehicle or is delivered to one’s destination) and how many kilometers per day these vehicles would travel.

A natural extension of our analysis is to also consider the possibility of combining the “last mile” service described here with a “first mile” service, i.e., have the vehicles pick up passengers from the serviced district and transport them to the rail station. Under certain conditions, this may lead to increased efficiencies since vehicles will not be returning to the rail station empty. However, the analysis of this type of combined first- and last-mile service is significantly more complicated, if it is to be carried out at a similar level of detail as the analysis presented here for last-mile services alone. This is therefore suggested as a topic for future research.
References


Köllerström, J. (1974). Heavy traffic theory for queues with several servers. I. Journal of
Applied Probability, 544-552.

the queueing system GI/G/1. In Congressbook of the Eight International
Teletraffic Congress, pages 2351/8, Melbourne.


Lees-Miller, J. D., Hammersley, J. C., & Wilson, R. E. (2010). Theoretical maximum
capacity as benchmark for empty vehicle redistribution in personal rapid
transit. Transportation Research Record: Journal of the Transportation Research
Board, 2146(1), 76-83.

EURO Journal of Transportation and Logistics, 1-19.

transit systems: service and energy performance assessment of the Masdar City


