Volume 38, Issue 5, September 2010

ELSEVIER

Operations Research Letters

Operations Research Letters 38 (2010) 432–435

Contents lists available at ScienceDirect

Operations Research Letters

journal homepage: www.elsevier.com/locate/orl

A stochastic model for a vehicle in a dial-a-ride system

E. Hyytiä [*], S. Aalto, A. Penttinen, R. Sulonen

Helsinki University of Technology, TKK, Finland

ARTICLE INFO

Article history:
Received 14 January 2010
Accepted 7 May 2010
Available online 24 May 2010

Keywords:
Dial-a-ride problem
Erlang's loss system
Interruptions

ABSTRACT

We consider a stochastic model for a system which can serve n customers concurrently, and each accepted and departing customer generates a service interruption. The proposed model describes a single vehicle in a dial-a-ride transport system and is closely related to Erlang's loss system. We give closed-form expressions for the blocking probability, the acceptance rate, and the mean sojourn time, which are all shown to be insensitive with respect to the forms of the distributions defining the workload and interruption durations.

1. Introduction

In a dynamic dial-a-ride problem (DARP) [2,1], the aim is to provide a transport service that satisfies the dynamically arriving trip requests by controlling a fleet of N vehicles, each having n customer seats. Upon receiving a new trip request, the system immediately assigns the request to some vehicle, the route of which is modified accordingly to include the new pick-up and drop-off locations, and an acknowledgement about the trip is sent back to the customer. One important consequence is that the delivery of the other customers may get delayed when a route changes due to a new customer. In general, one can identify two somewhat conflicting goals: (i) the system's effort (e.g., driven kilometers) and (ii) the customers' service level (e.g., mean travel time). The fundamental decision is to juggle between the interests of the system, the new customer, the old customers, and the future customers. A policy maneuvering the fleet sets a balance between these two objectives. Controlling the fleet is an interesting and very challenging problem itself, but outside the scope of this paper.

In a dial-a-ride system, vehicles follow constantly changing routes while picking up and dropping off customers. In a high-demand situation with a large number of vehicles, only trips along some meaningful route are assigned to each vehicle. Conversely, these trips implicitly define a vehicle's route, which becomes self-evident. In fact, an often used heuristic rule simply *inserts* a new trip to the existing route without changing the relative order of the existing stops [1], thus reducing the complexity considerably.

In this paper, we propose an elementary model for a single vehicle in a dial-a-ride system. The model captures several aspects of the real system. Firstly, the vehicle has a finite capacity, and thus during a congestion customers may have to be rejected. Secondly, each stop delays the travel of the other customers sharing the same vehicle. Consequently, the model allows us to gauge, for example, the length of pick-up/drop-off detours that should be taken.

2. Model

We study a demand-responsive transport system, where each vehicle, having n seats and a velocity v, picks up and drops off customers along its route. Customers are distributed on the road according to a Poisson process with density λ^* per unit distance [1/m], and defining $\lambda = \lambda^* \cdot v$ gives the offered customer rate per unit time [1/s]. Trip distances, denoted by the X_i^*, are drawn from an arbitrary distribution, $X_i^* \sim X^*$, so that $X_i = X_i^* v$ denotes a "direct workload" in time [s]. The random variable S denotes the total time to pick-up/drop-off a customer, consisting of the driving time and the actual stop time (see Fig. 1). The total additional work (in time) a customer creates, denoted by D, is thus $S_i + S_j$, i.e., $\mathrm{E}[D] = 2\,\mathrm{E}[S]$. We further assume that a boarding customer experiences half of the corresponding interruption period, and a departing customer, similarly, leaves the system in the middle of the departure interruption. Thus, on average, each customer experiences a delay of $\mathrm{E}[S]$ due to his own interruptions. During each detour, i.e., interruption, the other customers get no service in terms of progress towards their destination. Neither can new customers arrive for the same reason.

2.1. Erlang's loss system with interruptions

Erlang's loss system is a well-known stochastic model with n servers, no waiting places, a Poisson arrival process with rate λ, and

[*] Corresponding address: Comnet, P.O. Box 3000, 02015 TKK, Finland.
E-mail addresses: esa@netlab.tkk.fi (E. Hyytiä), samuli.aalto@tkk.fi (S. Aalto), aleksi@netlab.tkk.fi (A. Penttinen), reijo.sulonen@tkk.fi (R. Sulonen).

0167-6377/$ – see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.orl.2010.05.009

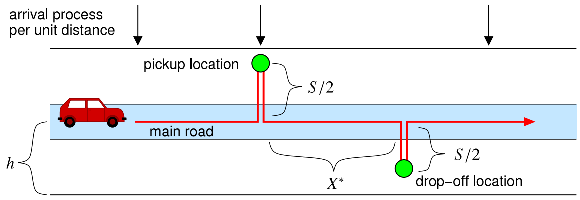

Fig. 1. Model illustrated. Pick-ups and drop-offs correspond to "perpendicular" deviations from the main route/road and the random variable S defines the additional delay experienced by the other customers.

an arbitrary service time distribution with mean $\mathrm{E}[X]$. For Erlang's loss system, the stationary probability distribution is

$$\pi_i^{(n)} = \frac{a^i/i!}{\sum_{j=0}^{n} a^j/j!}, \quad i = 0, \ldots, n \tag{1}$$

where $a = \lambda\,\mathrm{E}[X]$, and the blocking probability in particular is given by $\mathrm{Erl}(n, a) = \pi_n^{(n)}$. Note that this result holds for an arbitrary service time distribution with mean $\mathrm{E}[X]$ (insensitivity) [5,4,8]. Due to the so-called PASTA property (*Poisson arrivals see time averages*), the steady-state distribution the accepted customers see is given by $\pi_i^{(n-1)}$, i.e., arriving customers see the system as if it had one server less. Thirdly, the mean number of customers N in an ordinary Erlang loss system is given by

$$\mathrm{E}[N] = a \cdot (1 - \mathrm{Erl}(n, a)). \tag{2}$$

Here we consider a variant of this, referred to as *Erlang's loss system with interruptions*.

Def 1 (*Erlang's Loss System with Interruptions*)**.** The system has n servers, no waiting places, and accepts a new customer given a free server is available. The system is either in the active state, or interrupted. In the active state, customers arrive according to a Poisson process with rate λ [1/s], and have i.i.d. service times $X_i \sim X$ with finite mean $\mathrm{E}[X]$. Each accepted and departing customer triggers an interruption for a duration $S_i^{(\mathrm{in})}$ and $S_i^{(\mathrm{out})}$, respectively, where both are i.i.d. random variables, $S_i^{(\mathrm{in})} \sim S^{(\mathrm{in})}$ and $S_i^{(\mathrm{out})} \sim S^{(\mathrm{out})}$, having finite means $\mathrm{E}[S^{(\mathrm{in})}]$ and $\mathrm{E}[S^{(\mathrm{out})}]$. During the interruptions, no customer is served, and also the arrival rate is zero, i.e., the system is offline.

Interruptions are internal, i.e., not due to some external process, and may correspond, for example, to a reconfiguration delay or, as in our case, to a pick-up/drop-off delay. This is not the same as *vacations*, where an empty server enters into an idle state (=vacation), in which the server only occasionally checks if new customers have arrived [3].

3. Analysis

Let a denote the offered load in the active state, $a = \lambda \cdot \mathrm{E}[X]$. Neglecting the time periods corresponding to service interruptions yields an ordinary Erlang loss system. Therefore the blocking probability b in the active state is, simply, $b = \mathrm{Erl}(n, a)$, which is also the system's overall blocking probability as arrivals occur only in the active state.

Lemma 1. *The rate of accepted customers (throughput), denoted by λ_a, is given by*

$$\lambda_a = \frac{1 - \mathrm{Erl}(n, a)}{1/\lambda + \mathrm{E}[D](1 - \mathrm{Erl}(n, a))},$$

where $\mathrm{E}[D] = \mathrm{E}[S_i^{(\mathrm{in})}] + \mathrm{E}[S_i^{(\mathrm{out})}]$. (3)

Proof. A system without interruptions is an ordinary Erlang loss system. Moments when such a system becomes empty are regeneration points [7]. Let $\tilde{A}(t)$ denote the number of accepted customers during $(0, t)$, $\tilde{N}_a$ the number of accepted customers during a cycle, and $\tilde{T}_a$ the duration of the cycle in Erlang's loss system. For the rate of accepted customers in Erlang's loss system (i.e., throughput), it holds that

$$\tilde{\lambda}_a := \lim_{t\to\infty} \frac{\tilde{A}(t)}{t} = \frac{\mathrm{E}[\tilde{N}_a]}{\mathrm{E}[\tilde{T}_a]} = \lambda(1 - \mathrm{Erl}(n, a)). \tag{4}$$

The modified system with interruptions behaves identically during the active periods. Therefore, it also constitutes a regenerative process with the same number of accepted customers per cycle, $\tilde{N}_a = N_a$. Moreover, $T_a = \tilde{T}_a + \sum_{i=1}^{N_a}(S_i^{(\mathrm{in})} + S_i^{(\mathrm{out})})$, which gives

$$\mathrm{E}[T_a] = \mathrm{E}[\tilde{T}_a] + \mathrm{E}[N_a]\,(\mathrm{E}[S^{(\mathrm{in})}] + S^{(\mathrm{out})}) = \mathrm{E}[\tilde{T}_a] + \mathrm{E}[N_a]\,\mathrm{E}[D],$$

where the expectation of the random sum is due to Wald's equation [8,7]. Finally, the rate of accepted customers in the modified system with service interruptions is

$$\lambda_a = \frac{\mathrm{E}[N_a]}{\mathrm{E}[T_a]} = \frac{\mathrm{E}[\tilde{N}_a]}{\mathrm{E}[\tilde{T}_a] + \mathrm{E}[\tilde{N}_a] \cdot \mathrm{E}[D]} = \frac{1}{1/\tilde{\lambda}_a + \mathrm{E}[D]},$$

and substituting (4) into the above yields (3). □

Corollary 2. *The blocking probability of a system in which customers arrive independently of the system's state, but are categorically rejected during an interruption, is given by*

$$b^* = 1 - \frac{\lambda_a}{\lambda} = \frac{\mathrm{Erl}(n, a) + \lambda \cdot \mathrm{E}[D] \cdot (1 - \mathrm{Erl}(n, a))}{1 + \lambda \cdot \mathrm{E}[D] \cdot (1 - \mathrm{Erl}(n, a))}.$$

That is, b^* is the blocking probability for a system where, instead of switching off the arrival process, customers arriving during interruptions are categorically rejected.

Another interesting quantity is the sojourn time of customers, denoted by T, and in particular, its mean value $\mathrm{E}[T]$. Note that, without interruptions, $\mathrm{E}[T] = \mathrm{E}[X]$, but here interruptions cause an additional delay that customers may experience. Assuming that the service times X_i are i.i.d. and that they obey an exponential distribution, one can utilize the memoryless property of Markovian systems, write a set of linear equations, and solve $\mathrm{E}[T]$. However, it turns out that a more general result exists.

Before proceeding any further, we must define when a customer actually enters and leaves the system: at the start, at the end, or at the middle of the corresponding interruption? These definitions make no difference to the system's behavior in the active state, but merely define a base line for the mean sojourn time. Thus, without loss of generality, we first assume that *the sojourn time does not include the customer's own interruptions*.

Lemma 3. *The mean sojourn time in the system is given by*

$$\mathrm{E}[T] = \mathrm{E}[X] + a \cdot (1 - \mathrm{Erl}(n - 1, a)) \cdot \mathrm{E}[D]. \tag{5}$$

Consequently, $\mathrm{E}[T]$ depends only on the mean values, $\mathrm{E}[X]$, $\mathrm{E}[S^{(\mathrm{in})}]$ and $\mathrm{E}[S^{(\mathrm{out})}]$, i.e., the *mean sojourn time is insensitive* with respect to the forms of the distributions defining the amount of work and the interruption times. Note that, letting $\mathrm{E}[N']$ denote the mean number of customers in Erlang's loss system with $n - 1$ servers, that is $\mathrm{E}[N'] = a(1 - \mathrm{Erl}(n - 1, a))$, then (5) can also be written as

$$\mathrm{E}[T] = \mathrm{E}[X] + \mathrm{E}[N'] \cdot \mathrm{E}[D].$$

The last term corresponds to the mean additional delay caused by the other customers.

Proof. The fraction of time the system is interrupted is $\lambda_a\,\mathrm{E}[D]$. Outside interruptions, one obtains an ordinary Erlang loss system, and the probability of *active state* i (active and i customers in the system) is obtained with aid of (1), $(1 - \lambda_a \cdot \mathrm{E}[D]) \cdot \pi_i^{(n)}$. Both interruption states occur with rate λ_a. An accepted customer

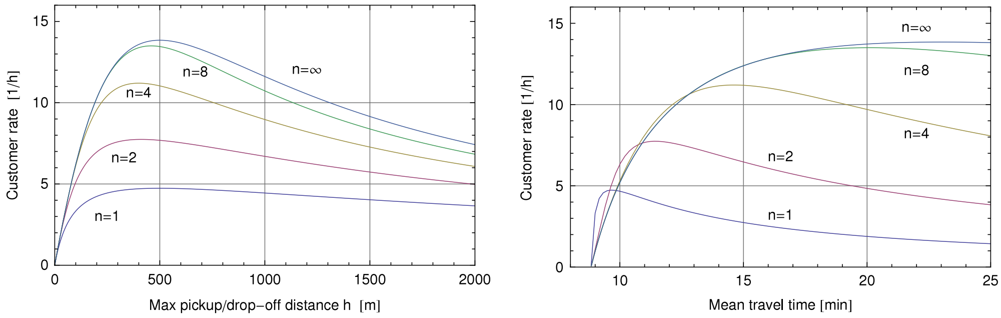

Fig. 2. Left: customer rate λ_a (throughput) as a function of range parameter h for $n = 1, 2, 4, 8, \infty$ seats per vehicle. Right: Trade-off between the customer rate and the customer's mean travel time $\mathrm{E}[T^*]$ obtained by increasing parameter h from $h = 0$ m corresponding to the common point at the bottom left corner.

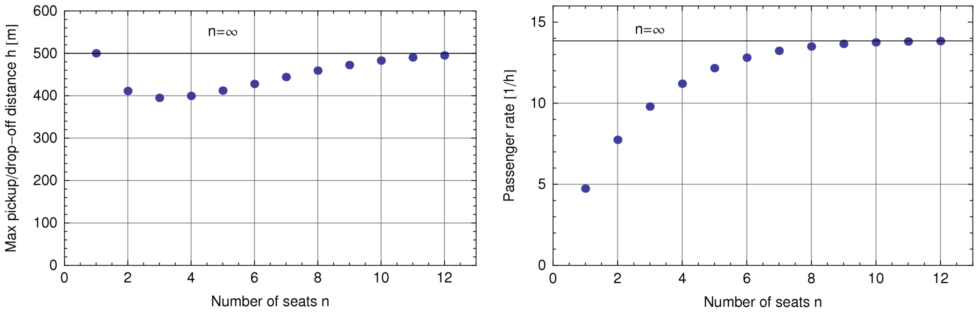

Fig. 3. Left: optimal pick-up/drop-off distance h for different numbers of seats n. Right: resulting customer rate (throughput) with optimal h as a function of n. The constant lines correspond to cases $n = \infty$.

sees the system's state according to $\pi_i^{(n-1)}$, $i = 0, \ldots, n-1$, as in Erlang's system. Similarly, the system's state after a departure obeys the same distribution. Thus, excluding the present customer, the probability of *arrival interruption (AI) state i* is given by $\lambda_a \cdot \mathrm{E}[S^{(\mathrm{in})}] \cdot \pi_i^{(n-1)}$, and *departure interruption (DI) state i*, similarly, $\lambda_a \cdot \mathrm{E}[S^{(\mathrm{out})}] \cdot \pi_i^{(n-1)}$. For the mean number of customers, on condition that the system is in the arrival interruption state, one obtains

$$\mathrm{E}[N \mid \mathrm{AI}] = \sum_{i=0}^{n-1} i \cdot \pi_i^{(n-1)} = a\,(1 - \mathrm{Erl}(n-1, a)),$$

as according to (2) we have $\sum_{i=0}^{n} i\,\pi_i = a(1 - \mathrm{Erl}(n, a))$. Similarly, $\mathrm{E}[N \mid \mathrm{DI}] = a\,(1 - \mathrm{Erl}(n-1, a))$. Thus, the mean number of customers in our system is given by

$$\begin{aligned}\mathrm{E}[N] &= (1 - \lambda_a \mathrm{E}[D])\,\mathrm{E}[N \mid \text{active}] + \lambda_a\,\mathrm{E}[S^{(\mathrm{in})}]\,\mathrm{E}[N \mid \mathrm{AI}] \\ &\quad + \lambda_a\,\mathrm{E}[S^{(\mathrm{out})}]\,\mathrm{E}[N \mid \mathrm{DI}] \\ &= (1 - \lambda_a \mathrm{E}[D]) \cdot a(1 - \mathrm{Erl}(n, a)) \\ &\quad + \lambda_a \mathrm{E}[D] \cdot a(1 - \mathrm{Erl}(n-1, a)) \\ &= a(1 - \mathrm{Erl}(n, a)) + \lambda_a \cdot a \cdot (\mathrm{Erl}(n, a) \\ &\quad - \mathrm{Erl}(n-1, a)) \cdot \mathrm{E}[D].\end{aligned}$$

Little's result [6], together with the arrival rate λ_a, then gives

$$\begin{aligned}\mathrm{E}[T] &= \mathrm{E}[N]/\lambda_a = a(1 - \mathrm{Erl}(n, a))/\lambda_a \\ &\quad + a \cdot (\mathrm{Erl}(n, a) - \mathrm{Erl}(n-1, a)) \cdot \mathrm{E}[D].\end{aligned}$$

For the first term, we have $a \cdot (1 - \mathrm{Erl}(n, a))/\lambda_a = \mathrm{E}[X] + a \cdot (1 - \mathrm{Erl}(n, a)) \cdot \mathrm{E}[D]$, and consequently, two terms cancel each other, yielding (5). □

Both an upper and a lower bound for $\mathrm{E}[T]$ follow immediately. Firstly,

$$\mathrm{E}[X] + a\,\mathrm{E}[D] \geq \mathrm{E}[T] \geq \mathrm{E}[X],$$

where the first inequality follows from $0 \leq \mathrm{Erl}(n-1, a) \leq 1$, and the second is obtained when $a \to 0$ and there is at most one customer in the system at a time. Similarly, when $a \to \infty$, the system is always full, and each departure is immediately followed by a new customer taking up the free place, $\lim_{a\to\infty} \mathrm{E}[T] = \mathrm{E}[X] + \mathrm{E}[D] \lim_{a\to\infty} a(1 - \mathrm{Erl}(n-1, a))$, where the limit obviously converges to $n - 1$ (see (2)). Thus,

$$\mathrm{E}[T] \leq \mathrm{E}[X] + (n-1)\,\mathrm{E}[D].$$

4. Numerical example

Let us next briefly return back to the original model for a single vehicle in a dial-a-ride system (see Fig. 1), where the pick-up and drop-off delays obey the same distribution S, and $\mathrm{E}[D] = 2\,\mathrm{E}[S]$. Moreover, each customer experiences half of his own interruptions, i.e., on average time $\mathrm{E}[S]$. Let $\mathrm{E}[T^*]$ denote the customer's mean sojourn time including his/her own interruptions, $\mathrm{E}[T^*] = \mathrm{E}[T] + \mathrm{E}[S]$. Substituting this into (5) gives

$$\mathrm{E}[T^*] = \mathrm{E}[X] + \mathrm{E}[S] + 2\,a \cdot (1 - \mathrm{Erl}(n-1, a)) \cdot \mathrm{E}[S].$$

At the limit when $n \to \infty$, we have

$$\lim_{n\to\infty} \lambda_a = (1/\lambda + 2\,\mathrm{E}[S])^{-1}, \quad \text{and}$$

$$\lim_{n\to\infty} \mathrm{E}[T^*] = (2\lambda\,\mathrm{E}[S] + 1) \cdot \mathrm{E}[X] + \mathrm{E}[S].$$

Let ρ denote the customer density per unit area $[1/\mathrm{km}^2]$, corresponding to the customers assigned to this particular vehicle, i.e., here we assume a constant customer density along the main road. Then let h denote the maximum pick-up/drop-off distance from the main road (see Fig. 1), so that the mean distance is $h/2$. With these,

$$\lambda = 2v\rho \cdot h \quad \text{and} \quad \mathrm{E}[S] = h/v + \mathrm{E}[t_{\mathrm{stop}}].$$

Consequently, asymptotically we have, for example, for the pick-up rate λ_a,

$$\begin{array}{lll} \text{Initially,} & \lambda_a \approx \lambda = 2v\rho \cdot h, & \text{when } h \to 0. \\ \text{For tail,} & \lambda_a \approx 1/\mathrm{E}[D] = v/(2h), & \text{when } h \to \infty. \end{array}$$

Let us further assume a velocity of $v = 10$ m/s, customer density of $\rho = 1/\mathrm{km}^2$ (Poisson point process), a mean stop time of $\mathrm{E}[t_{\mathrm{stop}}] = 30$ s, and a mean direct distance of $\mathrm{E}[X^*] = 5$ km. Fig. 2 (left) illustrates the resulting customer pick-up rate λ_a for

$n = 1, 2, 4, 8$ and infinite number of seats (per vehicle) as a function of h. Initially, the pick-up rate increases as more customers become available until too long detours start to hinder it. Fig. 2 (right) illustrates the trade-off between the system's interest (high pick-up rate) and the increased travel time the customers experience. Each curve is obtained by varying parameter h from zero. Obviously, the right-hand tails of the curves after their particular turning points correspond to adverse parameter values where decreasing h improves both the pick-up rate and the mean travel time. Additionally, we notice that initially the mean travel time increases much more slowly than the customers per hour rate, which means that if the customers are even a bit flexible then the vehicle's efficiency can be increased considerably, allowing a lower cost per trip.

Fig. 3 illustrates the performance when h is chosen so that the customer rate is maximized (system's gain). First, we can observe that the optimal range h does not vary much as a function of the number of seats n. Secondly, in this setting, with $n = 8$ seats the performance is already rather close to that of a vehicle with an infinite capacity.

Acknowledgements

This work was conducted in the Metropol project that is supported by the Finnish Funding Agency for Technology and Innovation, Finnish Ministry of Transport and Communications, Helsinki Region Transport and City of Helsinki.

References

[1] G. Berbeglia, J.-F. Cordeau, G. Laporte, Dynamic pickup and delivery problems, European Journal of Operational Research 202 (1) (2010) 8–15.
[2] J.-F. Cordeau, G. Laporte, The dial-a-ride problem: models and algorithms, Annals of Operations Research 153 (1) (2007) 29–46.
[3] B.T. Doshi, Queueing systems with vacations—a survey, Queueing Systems 1 (1) (1986) 29–66.
[4] D. Gross, C.M. Harris, Fundamentals of Queueing Theory, 3rd ed., 1998.
[5] L. Kleinrock, Queueing Systems, Volume I: Theory, Wiley Interscience, 1975.
[6] J.D.C. Little, A proof of the queueing formula $L = \lambda W$, Operations Research 9 (3) (1961) 383–387.
[7] S.M. Ross, Applied Probability Models with Optimization Applications, 1970.
[8] S.M. Ross, Introduction to Probability Models, 7th ed., Academic Press, 2000.