AN APPROACH TO TIME-DEPENDENT MODELLING OF QUEUES IN MULTIPLE LANES WITH TURNING MOVEMENTS

Nicholas Taylor

PhD Research Student (part-time)

UCL

Abstract and Introduction

Urban road networks are dominated by junctions with turning movements. Well-established methods exist to calculate capacities of these movements from geometry and volume information, and well-established analytical queue models exist to estimate queue sizes according to demand, capacity and queuing process. However, there is currently no standard analytical, as opposed to empirical or micro-simulation, method to account for interaction of movements sharing a lane or using adjacent lanes.

This paper looks at these questions starting with the basic observation that if a junction approach is divided into two lanes, with fully shared service, the queue size in each lane should be on average half that of the queue on a single lane with the same overall demand and capacity. But in the Pollaczek-Khinchin formula, provided the process is fixed, queue size depends only on the ratio of demand to capacity, making the total queue apparently double that on the individual lanes.

While it can be argued that service sharing is an idealisation or even unrealistic, it may still serve as an approximation where there is some interaction of service between neighbouring lanes. It is shown that modification of process statistics on individual lanes is insufficient. Pointers are laid to possible lines of further study, for example by extending the M/M/c or G/G/c multi-server queue models. However, these models seem unable to address all features of the problem.

In practical applications, a balance has to be struck between theoretical rigour, accuracy, flexibility and efficiency. The paper develops a simple generalised analytical model and computational method to predict queue sizes in the general case of a two lane approach with turning movements, taking account of lane sharing by different movements, and time-dependence. Modelled and event-based micro-simulation results are compared, showing that model accuracy is comparable with the inherent uncertainty of the outputs.

1 The basic problem with two lanes

Consider a road junction approach where two lanes are available, and suppose that arrivals can choose between lanes, either randomly or according to some principle, and that there can be some kind of sharing of service at the exit, as in Figure 1.

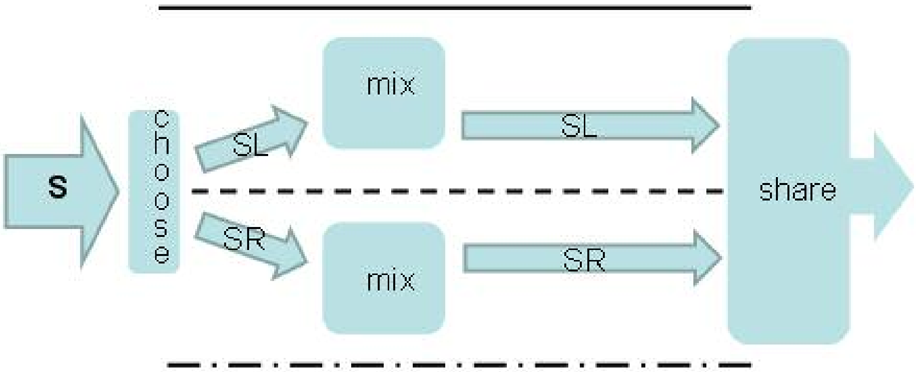


Figure 1. Two lane approach with possibility of lane choice and service sharing

If there is no change of lane after arrival and no sharing of capacity, then in effect there are two independent queues. In the steady state, queue size depends only on the traffic intensity, not the absolute volumes. If the lanes have equal capacity $\mu/2$, arrival rates are divided γq left and $(1-\gamma)q$ right, the overall traffic intensity is $\rho=q/\mu$, and the process is M/M/1, ('Markov/Markov/single server') then the total steady-state queue is:

$$L_e(\phi) = \frac{2\gamma\rho}{1-2\gamma\rho} + \frac{2(1-\gamma)\rho}{1-2(1-\lambda)\rho} = \frac{2\rho(1-2\gamma\rho(1-\gamma))}{(1-2\gamma\rho)(1-2(1-\gamma)\rho)} \qquad (1)$$

The queue is maximised when γ is 0 or 1, when it can become oversaturated even if $\rho<1$. It is minimised when $\gamma=0.5$, being twice what it would be in a single lane, namely:

$$L_e = \frac{2\rho}{1-\rho} \qquad \text{(symmetrical separate lanes)} \qquad (2)$$

Neither of these scenarios is realistic. If either lane can be used freely, we expect arrivals to choose what they perceive to be the shorter queue[1]. If there is no formal lane separation there can be free movement between the lanes, but in practice this is of little avail unless it increases throughput, so it is represented more usefully as a free choice between lane servers provided they are available. This makes the system equivalent to a single lane, so:

$$L_e = \frac{\rho}{1-\rho} \qquad \text{(merged lanes/service or single lane)} \qquad (3)$$

By symmetry, the queue in each lane must be half this, yet as both the arrivals and the nominal capacity in each lane are halved, the traffic intensity in each lane is still ρ. Therefore, either the inference is false or the statistics of the queue process in each lane are no longer M/M/1.

A third possibility is that this idealised situation simply cannot occur, but this is a weak argument against addressing something theoretically, since theory itself ought to forbid it! If two lanes ultimately feed into a single stream then merging will result in some turbulence. This might for example be the situation where a multiple-lane stop line at a roundabout feeds into an unmarked circulating section. In practice it may be difficult to observe and describe such movements, making it difficult to verify any model except in aggregate terms. Nevertheless one can imagine a range of possibilities from no sharing of service to perfect sharing, and suppose that any actual case can be approximated by a theoretical one in that range.

2 Role of coefficients of variation

The Pollaczek-Khinchin (P-K) steady-state mean queue formula (eg Medhi 2003), as modified by Heydecker (unpublished, personal communication) is given by equation (4):

$$L_e = I_b\rho + \frac{(I_a-1)\rho}{2(1-\rho)} + \frac{(1+c_b^2)\rho^2}{2(1-\rho)} \qquad \text{where} \qquad (4)$$

I_a is the dispersion (mean/variance) of arrivals as introduced by Heydecker
I_b is a parameter reflecting the unit-in-service as identified by Kimber and Hollis (1979)
c_b is the coefficient of variation (mean/s.d.) of the service time.

Kimber, Summersgill and Burrow (1986) suggested that the P-K formula might be generalised by replacing the coefficient $(1+c_b^2)$ by $(c_a^2+c_b^2)$, where c_a is the coefficient of variation of the arrival headway time. However, this intuitive proposal is inconsistent with the formal derivation involving I_a, indeed the coefficient $(1+c_b^2)$ emerges naturally in the derivation of (4) with or without including dispersion in the arrivals.

[1] A degree of misperception can be represented by adding to the actual queue size a random amount proportional to the square-root of queue size. This effectively treats the space occupied by units as a Poisson variable. In practice, it is the split ratio which matters not the method used to achieve it.

It might be expected that where arrivals choose the shorter of two queues, this could affect the dispersion of arrivals in each queue separately. In practice the effect would be small compared to random selection, and with symmetrical service random selection by arrivals would be expected to produce similar average results, so we can assume I_a=1, eliminating the second term in (4) and returning it to the more familiar form. To accommodate the effect of sharing service there are now only the parameters I_b and c_b. For a normal M/M/1 queue, c_b=1. Halving the coefficient of the last term in (4) would require c_b to be zero, which actually represents uniform service. It is not obvious that the unit-in-service coefficient, which represents the unavoidable average time needed to service each unit, can meaningfully be halved either. If both of these *could* be achieved then the mean queue would be halved, as required by perfect sharing, but this cannot be extended to more than two lanes. So manipulating the coefficients in (4) is insufficient – the formula can accommodate only the statistics of individual servers not correlations between servers.

3 Multi-channel process

Standard queuing theory for multiple servers centres on the M/M/*c* or G/G/*c* processes (G='General'), which represent something like a supermarket with *c* checkout channels. The focus tends to be on service times when some servers are idle and queues, if any, are short. Arrivals choose an idle server if one is available and otherwise join a queue at random, and there is no interaction between the servers or queues. In transport, however, the focus is more on a few heavily loaded servers and interaction between queues can occur, so the standard model may have limited application. However, it is instructive to see where it leads.

If ρ is defined to be the overall traffic intensity of the system, treating the individual checkout capacities as additive, the steady-state queue (Medhi 2003 with modified notation[2]) is:

$$L_e^{(c)} = c\rho + \frac{\rho C(c,\rho)}{1-\rho} \qquad \text{where} \qquad C(c,\rho) = \frac{(c\rho)^c p_0}{c!(1-\rho)} \tag{5}$$

$$\text{and} \quad p_0 = \left[\sum_{i=0}^{c-1} \frac{(c\rho)^i}{i!} + \frac{(c\rho)^c}{c!(1-\rho)}\right]^{-1} \qquad p_i = \frac{c}{\min(i,c)}\rho p_{i-1} \; (i{>}0) \tag{6}$$

Here p_0 is the probability that the system is empty, and the coefficient of variation of service is assumed to be 1 for all the servers. For *c*=1, we recover the usual formula (3), noting in this case that p_0=1-ρ. For *c*=2, we get:

$$p_0^{(2)} = \frac{1-\rho}{1+\rho} \qquad \text{and} \qquad L_e^{(2)} = \frac{2\rho}{1-\rho^2} \tag{7}$$

This can be considered to account for selection of channel by arrivals, but clearly does not account for service sharing since the 'checkouts' are independent. However, it is easy to show that the queue size (7) lies somewhere between (3) and (2).

Queue size probabilities have been calculated using (6-7) for a two-server system, using a Markov process based on recurrence relations for two lanes with shared service, and assuming each lane queue is M/M/1 with half the total queue. These are graphed in Figure 2 and some invariants given in Table 1. The graphs show clearly the difference between the two-server (broken line) and M/M/1 (nearby solid line) distributions, and the similarity between the distributions on individual lanes regardless of how generated. The whole system is constrained to look the same overall whatever goes on 'inside the box', but the lesson is that unless the arrival process is very peculiar it is the service we need to concentrate on, and that at least in the symmetrical case it is possible to get away with treating the individual lane processes as M/M/1.

[2] Number of channels *c* and function *C*() are unconnected with the coefficients c_b and *C* in P-K.

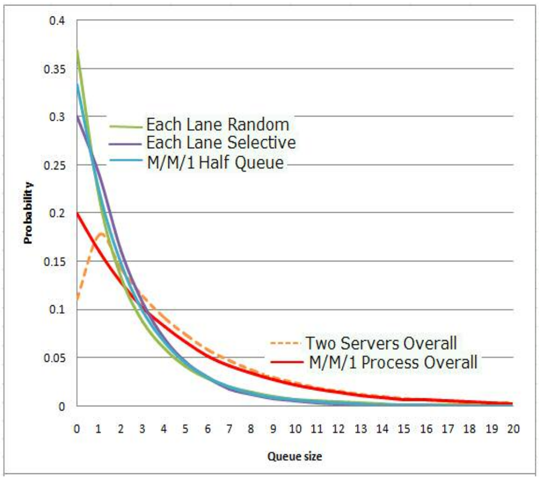
Table 1. Characteristic values for the cases in Figure 2

Queue Size *i*	Two Servers Overall	M/M/1 Process Overall	Each Lane Random	Each Lane Selective	M/M/1 Half Queue
p_0	0.111	0.200	0.369	0.300	0.333
L_e	4.444	4.000	2.000	2.000	2.000

4 Practical concerns

At this point it is useful to be reminded of some things:

1. As pointed out above, multi-server theory tends to be concerned mainly with a large number of lightly-loaded servers, several of which may be idle, whereas traffic modelling is mostly concerned with a smaller number of service processes which tend to be busy and may be heavily loaded.
2. Smooth probability distributions conceal the fact that real queues are highly variable; even after thousands of events simulated distributions are very ragged.
3. Equilibrium or steady state is a practical fiction, and the range of situations of interest in which equilibrium applies is quite limited, as Figure 3 shows.
4. Practical traffic modelling needs to consider the presence of turning flows, which standard queuing theory does not deal with.
5. It would be advantageous if time-dependent multi-lane and turning cases could be handled using existing efficient closed-form time-dependent queuing models.



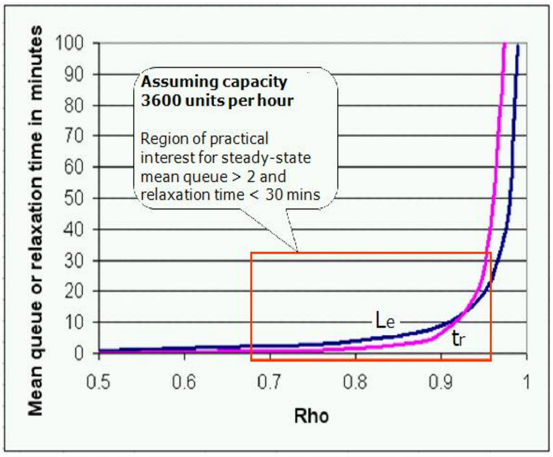


Figure 2 (left). Graphs of queue size distributions for some two-lane cases with ρ=0.8

Figure 3 (right). M/M/1 equilibrium queue and relaxation time showing limited useful region

5 The effect of service sharing on utilisation

In Figure 1, if arrivals choose a lane randomly and there is a common service process then the total queue must be the same as for a single lane. This is an admittedly artificial model but it may give some insight into real situations where there is some interaction between the lanes. Further, if the queue in each lane is supposed to be approximately M/M/1, how can it be expressed in the form of (4)? It has already been shown that dispersion in arrivals is not an issue and the coefficient of variation of service cannot be used. In Table 2 the states of the lane queues are divided into empty and non-empty. The utilisation of each lane is the proportion of time for which it is busy.

Table 2. Grouping of elementary probabilities for two lanes with shared service

	Lane 2 queue = 0	**Lane 2 queue > 0**
Lane 1 queue = 0	P_0	
	p_{00}	$\Sigma_1^\infty p_{0j} = P_0 - p_{00}$
Lane 1 queue > 0	$\Sigma_1^\infty p_{i0} = P_0 - p_{00}$	
	1-P_0	

If each lane can use all the capacity available to the other when it is empty then the effective factor of gain in utilisation for each lane is:

$$f_u^{(2)} = \frac{1 - P_0 + P_0 - p_{00}}{1 - P_0} = \frac{1 - p_{00}}{1 - P_0} \tag{8}$$

Since during some random intervals the throughput capacity on a lane is doubled, the distribution of service headways is no longer perfectly Exponential. In terms of *time*, the proportion of total time needed to service the queue in the lane is reduced, so the effective utilisation in terms of time is reduced. A similar calculation can be performed for three lanes:

Table 3. Utilisation construction for three lanes with shared service

Case	**Absolute probability**	**Utilisation factor**
All lanes zero	$p_{000} = 1 - \rho$	0
Both other lanes zero	$\Sigma p_{00i} = P_{00} - p_{000}$	3 (contributes additional 2x)
One other lane zero	$P_0 - 2(\mathrm{P}_{00} - p_{000}) - p_{000}$	2 (contributes additional 1x)
Neither other lane zero	*remainder*	1
Lane 1 non-zero	$1\text{-}\Sigma p_{0ij} = 1\text{-}P_0$	Includes three cases above

The effective utilisation is then:

$$f_u^{(n)} = \frac{1 - P_0 + \left(P_0 - 2\left(P_{00} - p_{000}\right) - p_{000}\right) + 2\left(P_{00} - p_{000}\right)}{1 - P_0} = \frac{1 - p_{000}}{1 - P_0} \tag{9}$$

For any number of lanes *n*, the probability of the total queue being zero $p_{0\ldots0}$ is the same, namely 1-ρ where ρ is the traffic intensity for the whole system. Strictly, under time-dependence, ρ should be replaced by the overall utilisation at the stop line *x*, but for the moment stay with the steady state. To calculate $f_u^{(n)}$ we need an estimate of P_0, the probability that an individual lane is empty. Defining first a 'transfer probability' $P_x^{(n)}$ by:

$$P_x^{(n)} = \frac{P_0 - p_{0\ldots0}}{(n-1)(1 - P_0)} \tag{10}$$

the factor $f_u^{(n)}$ can be developed as the Binomial expansion (11), whose components represent the separate contributions to one lane's utilisation when: no other lane is empty, one other lane is empty, two are empty etc, up to all other lanes empty. Each contribution has a weighting factor one more than the number of empty lanes. For example, one empty lane doubles the utilisation available to the current lane with a certain probability, two empty lanes triple it etc, and each case can occur in a combinatorial number of ways:

$$f_u^{(n)} = \left(1 - P_x\right)^{n-1} + 2\binom{n-1}{1}P_x\left(1 - P_x\right)^{n-2} + \ldots + nP_x^{n-1} \tag{11}$$

This simplifies naturally to:

$$f_u^{(n)} = \left(1 - P_X\right) + nP_x \equiv 1 + (n-1)P_x \tag{12}$$

Even though most of the terms in the Binomial formulation cancel out, it has value both as explanation and because the contributions are mutually independent, thus avoiding the need for a statistical representation of correlation between lanes.

Knowing from symmetry that the mean steady-state queue in each of *n* lanes must be:

$$L_e^{(n)} = \frac{\rho}{n(1-\rho)} \tag{13}$$

and treating each lane queue as quasi-M/M/1, this implies the *effective* ρ in each lane is:

$$\rho^{/} = \frac{L_e^{(n)}}{L_e^{(n)}+1} = \frac{\rho}{\rho+n(1-\rho)} = \frac{\rho}{f_u^{(n)}} \tag{14}$$

where the last equality is numerical because of the assumed symmetry. If, again based on quasi-M/M/1, we suppose that approximately:

$$P_0 \approx 1-\rho^{/} \tag{15}$$

equations (12) and (14) lead to:

$$P_x \approx 1-\rho \tag{16}$$

So the transfer contribution from an empty lane is not only independent of the other lanes but reflects the nominal probability of the lane being empty as if it were isolated. This is very useful because it can be calculated by conventional methods. The linearity of (12) and (16) suggests that if only a proportion ϕ of an empty lane's capacity is sharable then P_x can simply be reduced by this factor. This might be the case if effective sharing arises from weaving either between lanes or after departure.

Considering two lanes, on a time basis we can write the probabilities that when one lane is ready for service the other lane is empty or otherwise not ready, $P^{(t)}{}_{x0}$, or its complement where the other lane is ready for service, $P^{(t)}{}_{xx}$:

$$P_{x0}^{(t)} = \phi\left(\frac{P_0 - p_{00}}{1-P_0}\right), \qquad P_{xx}^{(t)} = \frac{1-(1+\phi)P_0+\phi p_{00}}{1-P_0} \tag{17}$$

These are not the same as the probabilities that a unit will be served, since more can be served on one lane when the other lane not being served, hence on a service basis:

$$P_{x0}^{(s)} = \frac{2p_{x0}^{(t)}}{1+p_{xo}^{(t)}} = \frac{2\phi(P_0-p_{00})}{1-(1-\phi)P_0-\phi p_{00}},\ P_{xx}^{(s)} = \frac{1-(1+\phi)P_0+\phi p_{00}}{1-(1-\phi)P_0-\phi p_{00}} \tag{18}$$

The significance of (18) is that these probabilities can be measured directly in microscopic simulations by counting the numbers of units served under different conditions, allowing models to be tested, and where applicable the factor ϕ calibrated. Of course, we have assumed symmetry between the lanes. If this does not apply then the formulae will be more complicated. For straight movements on two lanes, using (12) and (14), the effective ρs are:

$$\rho'_{SL} = \frac{\rho_{SL}}{1+\phi(1-\rho_{SR})}, \qquad \rho'_{SR} = \frac{\rho_{SR}}{1+\phi(1-\rho_{SL})} \tag{19}$$

This allows service sharing to be included when the arrivals and possibly the basic service are not necessarily symmetrical between the lanes.

6 Space sharing in a single lane

If a lane carries two independent movements, we need to be sure that the queue model is internally consistent. This can be called the 'red queue, blue queue' problem, meaning that if units in the system are arbitrarily painted in two colours it should not affect the result.

Suppose that the total arrivals are *q*, the capacity is μ, the 'red' and 'blue' component arrival rates are q_r and q_b, and u_r, u_b are the corresponding utilisations relative to the combined capacity. Since each 'colour' takes capacity away from the other, the partial capacities are:

$$\mu_r^* = \mu(1-u_b), \quad \mu_b^* = \mu(1-u_r) \quad \text{where} \quad u_r = q_r/\mu\,, \ u_b = q_b/\mu \qquad (20)$$

Define the effective degrees of saturation simply as follows:

$$x_r^* = q_r/\mu_r^* = q_r/(\mu - q_b), \qquad x_b^* = q_b/\mu_b^* = q_b/(\mu - q_r) \qquad (21)$$

The mean steady-state queues according to the standard M/M/1 formula are:

$$\begin{aligned} L_r^* &= x_r^*/(1-x_r^*) = q_r/(\mu - q_r - q_b), \\ L_b^* &= x_b^*/(1-x_b^*) = q_b/(\mu - q_r - q_b) \end{aligned} \qquad (22)$$

Since these queues are independent the total queue is just their sum:

$$L^* = (q_r + q_b)/(\mu - q_r - q_b) \qquad (23)$$

But since the total volume $q = q_r + q_b$, this is the same as:

$$L^* = q/(\mu - q) = \rho/(1-\rho) \equiv L_e(\rho) \qquad (24)$$

Thus the result is independent of the 'red/blue' split, showing that the model is internally consistent. This will not necessarily work for queue processes other than M/M/1, if one queue affects the headway distributions of the other.

What if the 'colours' *are* distinguishable, in particular their capacities differ? Suppose these capacities are μ_r and μ_b, being the capacities which each 'colour' would experience if the other were absent. How should capacities be averaged? At first sight it seems that the harmonic mean ought to be used, as it is related to the mean service time, but this neglects the relative proportions in the flow, which are reflected in the respective utilisations.

Remembering that utilisation is the proportion of time that service takes place on a movement, and that the movements are mutually exclusive, at any point in a given time period a unit can be using either the 'red channel' or the 'blue channel', or no service is taking place. The combined utilisation is therefore the *sum* of the component utilisations. It is reasonable to define the combined capacity as the total throughput divided by the sum of the component utilisations. Incidentally, the same result is got if the service time on each movement is weighted by throughput. Thus:

$$\bar{\mu} = \frac{u_r\mu_r + u_b\mu_b}{u_r + u_b} \qquad (25)$$

Contrary to what might appear at first sight, the average capacity in equation (25) is not dominated by the larger component capacity. Remembering that the sum of the utilisations is constrained to be less than 1, if the 'red' capacity approaches zero while the 'blue' capacity stays finite, then if there is any 'red' demand its utilisation will approach 1 and the 'blue' utilisation is therefore forced towards zero. The combined capacity then approaches the 'red' capacity, that is zero as expected and as would also be the case if the combined capacity *were* the harmonic mean of the component capacities.

7 Deterministic model of two lanes with turning movements

Figure 4 extends Figure 1 to include turning movements[3]. Left turners always use the left lane and right turners the right lane. Straight movers may use either lane according to some choice mechanism, such as selecting the queue which appears shorter. Using (19) and (20) we can calculate effective capacities for all the movements, treating straight-left-lane and straight-right-lane as separate. To allow for later introduction of time-dependence we now work in terms of utilisations, the proportions of time spent serving a movement.

[3] The figure shows left-driving convention but that does not affect the mathematics.

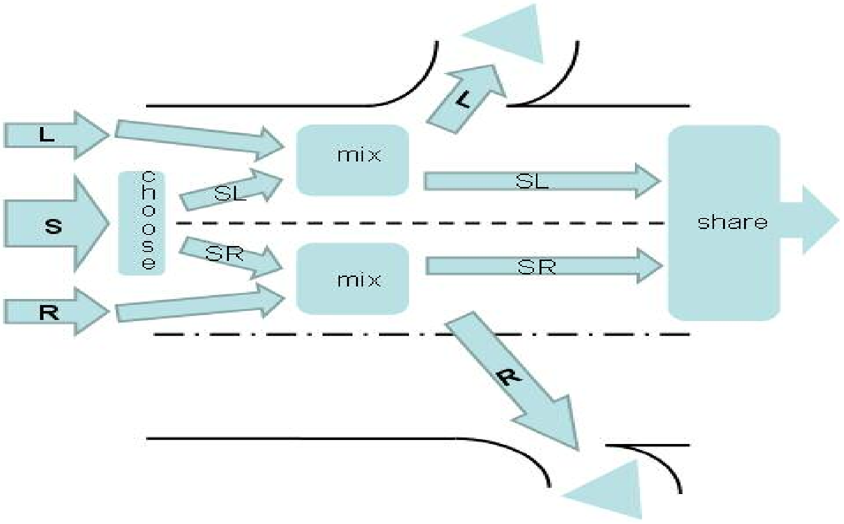


Figure 4. Two lanes with turning movements and service sharing

The effective capacities can be expressed in different ways in terms of absolute or effective utilisations, assuming that each lane claims half the total straight-movement capacity μ_S. The factor ϕ would add nothing to the argument but can be built into in any actual calculations.

$$\mu_L^* = \mu_L\left(1 - u_{SL}^{/}\right) = \mu_L\left(\frac{2 - u_{SL} - u_{SR}}{2 - u_{SR}}\right)$$

$$\mu_{SL}^* = .5\mu_S\left(1 - u_L\right)\frac{u_{SL}}{u_{SL}^{/}} = .5\mu_S\left(1 - u_L\right)\left(2 - u_{SR}\right) = \frac{\mu_S\left(1 - u_L\right)\left(1 - u_{SR}^{/}\right)}{\left(1 - u_{SL}^{/}u_{SR}^{/}\right)}$$

$$\mu_{SR}^* = .5\mu_S\left(1 - u_R\right)\frac{u_{SR}}{u_{SR}^{/}} = .5\mu_S\left(1 - u_R\right)\left(2 - u_{SL}\right) = \frac{\mu_S\left(1 - u_R\right)\left(1 - u_{SL}^{/}\right)}{\left(1 - u_{SL}^{/}u_{SR}^{/}\right)}$$

$$\mu_R^* = \mu_R\left(\frac{2 - u_{SL} - u_{SR}}{2 - u_{SL}}\right) = \mu_R\left(1 - u_{SR}^{/}\right) \qquad (26)$$

Then assuming quasi-M/M/1 processes, the component queues (27) can be calculated from the effective traffic intensities of the component services. Finally, in accordance with (23) and (24), the component queues in each lane can be added to get the total queues.

$$\rho_i^* = \frac{q_i}{\mu_i^*}, \qquad L_i = \frac{\rho_i^*}{1 - \rho_i^*} \qquad \text{if undersaturated in the steady state} \quad (27)$$

8 Reminder of the sheared queue model

Time-dependent queue solutions are needed in what follows. As described by Kimber and Hollis (1979) and elsewhere, and interpreted by Taylor (2003), shearing transforms the Pollaczek-Khinchin steady-state mean queue formula to be asymptotic to the deterministic queue formula, creating a time-dependent model which handles the transition between under- and oversaturation seamlessly. In effect, it treats the dynamic queue as quasi-static with the degree of saturation or utilisation at the stop line replacing the traffic intensity, to which it is no longer equal. In its simplest form, the sheared model can be written as (28), where L_0 is the initial queue and C now represents the randomness coefficient ($1+c_b^2$):

$$L(t) \equiv L_0 + (\rho - x)\frac{\mu t}{a} = I_b x + \frac{Cx^2}{1 - x} \qquad (a=1 \text{ for final, } a=2 \text{ for average}) \quad (28)$$

In the sheared model with constant parameters the average queue over time is approximately equal to the instantaneous queue at half time (Kimber and Hollis 1979). Hence the factor *a* is 1 if the final queue at time *t* is required, and 2 if the average queue over [0,*t*] is required. Average values are needed to adjust average utilisations for feasibility (see later) and also if the chosen benchmark is relatively stable time-averaged rather than highly variable final simulated queue values. The solution for the degree of saturation *x* is:

$$x(t)=\frac{g-\sqrt{g^{2}-4fh}}{2f}\quad (f\neq 0)\ ,\qquad x(t)=\frac{h}{g}\quad (f=0 \text{ and } g\neq 0)\ \text{ where}\qquad (29)$$

$$f=\frac{\mu t}{a}-\left(C-I_{b}\right),\quad g=L_{0}+I_{b}+\left(\rho+1\right)\frac{\mu t}{a},\quad h=L_{0}+\rho\frac{\mu t}{a}\qquad (30)$$

For M/M/1, where $I_b = C = 1$, and L_0=0, the queue can be calculated in three steps:

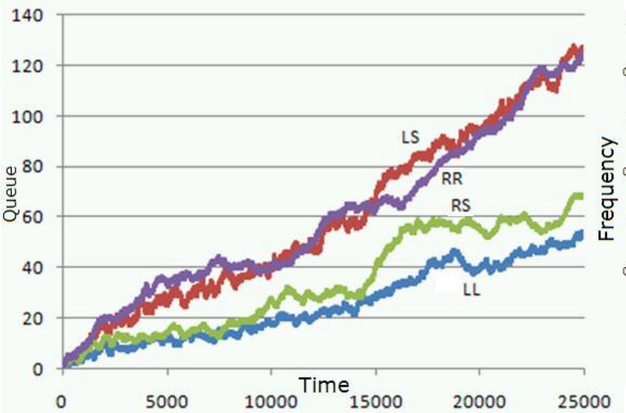
$$\left[g/2f\right]=\tfrac{1}{2}\left(1+\rho+\frac{a}{\mu t}\right),\quad x=\left[g/2f\right]-\sqrt{\left[g/2f\right]^{2}-\rho}\ ,\quad L=\frac{x}{1-x}\qquad (31)$$

9 Including time-dependence in the multilane model

Time dependence is important not just where service is oversaturated, but also in what may be termed '*de facto* oversaturation' where interference between movements reduces effective capacity below demand, and even below saturation if the traffic intensity is high enough to make relaxation time long compared to the modelled time period (see Figure 3).

An illustration is given by Figure 5 where a demand of 1100 units/hour has been imposed on a set up with total capacity summed over all movements of 1750 units/hour ('Case 110B'). This has been generated by averaging the results of nine event-based simulations each running to an extended simulated duration of about 7 hours. Individual simulation runs exhibit much greater variability, as shown by Figure 7, although Figure 6 shows that the statistics of the Exponential event generator are accurate. In Figure 7 and the early growth part of Figure 5, there is an apparent slowing of the queue growth rates, but this turns out to be illusory. Undersaturated cases have the same ragged appearance but without the secular growth trend, in fact the variability is maximised around saturation (Taylor 2005).

In time-dependent cases, the traffic intensities of arrivals (demands) no longer equal the utilisations at the stop lines. This leads to complications which appear soluble only by an iterative approach, particularly as it becomes impractical to solve for the proportion λ of straight-moving arrivals which should use each lane according to whatever policy is adopted. Indeed, if component queues fail to grow in step this proportion could vary with time. However, for the moment we wish only to show it is possible to arrive at a solution.



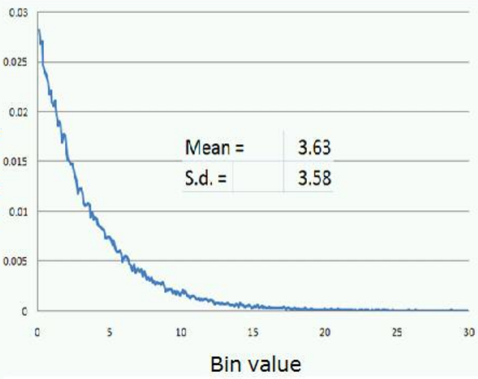


Figure 5 (left). Event-simulated component queues with *de facto* oversaturation

Figure 6 (right). Verification of Exponential random generator in simulation (100,000 events)

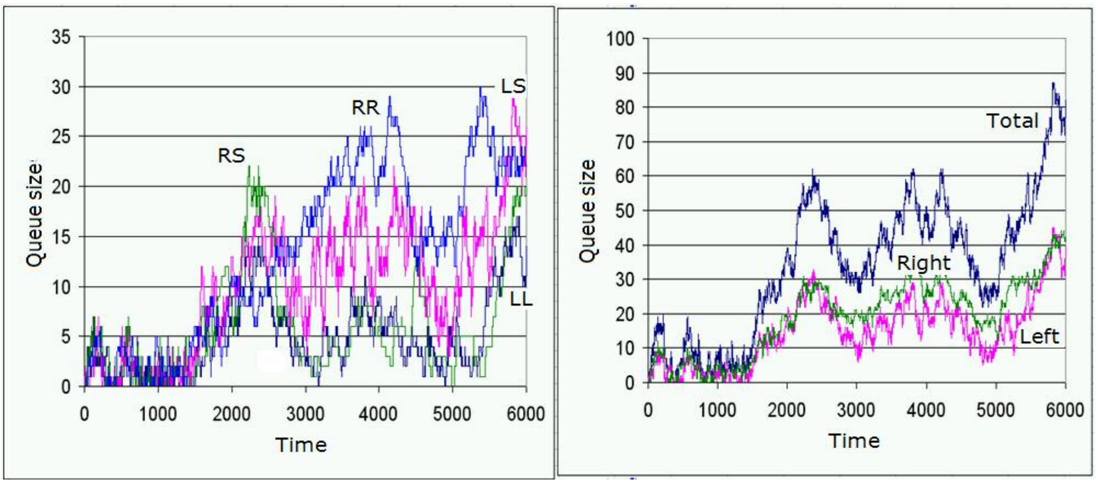


Figure 7. Results of a single event-simulation run for the above case showing variability

The algorithm adopted is the following:

1. Given the arrival rates on each movement, or total arrivals and turning proportions, and the basic capacities for each movement, calculate 'raw' average utilisations (degrees of saturation) on each movement using the sheared queue method. If steady state is assured, x=ρ, so this step can be omitted.

2. Calculate initial values of the effective utilisations on movements affected by shared service. This affects only the SL and SR movements in Figure 4.

3. If the sum of the effective utilisations on a lane exceeds 1 then they are infeasible and must be adjusted. This can be done by factoring the utilisations so their sum does not exceed the single utilisation value derived from the total demand, an average capacity on the lane calculated from the basic capacities and raw effective utilisations according to equation (25), and the average utilisation in the lane estimated using the sheared model with a=2.

4. Using the adjusted effective utilisations, calculate the effective capacities and traffic intensities according to equations (26) and (27), and use the sheared model to calculate either the time-averaged or final component queues, as required.

Although this algorithm needs up to three evaluations of the sheared queue formula rather than just one, the closed form queue model (31) is quite efficient to evaluate. Experiment indicates that up to five iterations of the whole algorithm may be needed to get a left-right split factor γ which equalises the lane queues. This is insensitive to the starting estimate which therefore might as well be 50:50. A descent direction calculated using derivatives of queue size with respect to γ can be used to adjust the factor.

9 Results and discussion

The model and algorithm described have been implemented in a spreadsheet and compared with event simulation. The test cases, defined in Table 4, are not intended to reflect typical realistic situations. In Table 4 and Figure 8 the model has been provided with the outturn simulated arrival rates and capacities rather than the specified values. This is arguably a fair test as results can be very sensitive to the inputs. 'Reconstruction' is a simplified application of the model based on outturn statistics such as equations (18), and therefore gives an indication of how well the model and event simulation correspond at a basic level without, for example, re-adjusting the straight-movement split factor γ.

In Figure 9, the model is working with the specified data, which can be a few percent different from those actually simulated. There is some difference between the modelled queue components and event simulation, and in particular the model appears to overestimate turning queues. Figure 10 plots the standard deviations of the component queues between the nine simulation runs with different random seeds, giving an idea of their inherent uncertainty which is around 30% of the average queue values and therefore of similar order to the model-simulation difference, suggesting the latter may be irreducible.

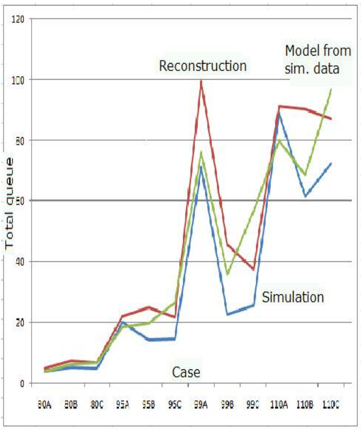
The algorithm given above has a somewhat circular nature, but solving directly for feasible, mutually consistent component utilisations under time-dependence appears to be intractable, and it is felt that a stepwise analytical approach is more transparent than numerical solution.

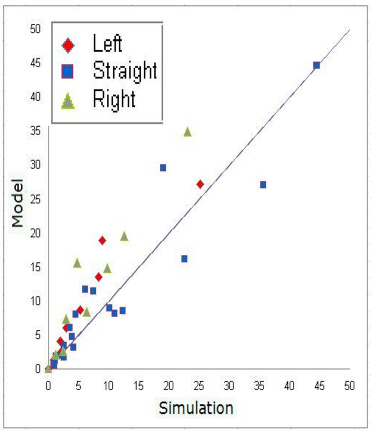
Table 4. Model results using specified arrival and capacity rates, with iterated balance

Movement	Left	Straight	Right
Capacities	500	1000	250
Sub-Case	**Turning proportions**		
A	0	1	0
B	0.2	0.6	0.2
C	0.35	0.5	0.15

‘Reconstruction’ is a simplified application of the model based on the outturn statistics from simulation. In the results modelled from simulated data, only the simulated arrival and capacity values are used, and the left-right balance is adjusted by hand.

Case	Demand	Event Simulation				Reconstruction		Modelled from sim. data		
	q	Hours	γ	L_L	L_R	L_L	L_R	L_L	L_R	γ
80A	800	19.571	0.555	2.1	2	2.9	1.9	2	2	0.5
80B	800	19.561	0.708	2	3.4	2.5	5	3.2	3.2	0.853
80C	800	19.543	0.487	2.7	2.4	3.9	2.7	3.4	3.4	0.395
95A	950	18.506	0.512	10.1	10.1	12.3	9.6	9.2	9.2	0.5
95B	950	18.49	0.757	5.4	8.9	6.8	18.1	9.9	10	0.856
95C	950	18.49	0.437	7.8	6.9	12.6	8.9	13.2	13.3	0.396
99A	990	18.174	0.502	35.6	35.6	52.1	47.2	38	38	0.5
99B	990	18.164	0.788	9	13.8	12.1	33.5	18	18.1	0.865
99C	990	18.203	0.421	13.4	12.2	23.3	14.2	28.2	28.4	0.381
110A	1100	1.726	0.504	44.4	44.5	47.8	43.3	40	40	0.5
110B	1100	1.732	0.803	27.3	34.1	28.4	61.8	34.2	34.3	0.827
110C	1100	1.69	0.395	37.4	35.2	45.8	41.1	48.3	48.4	0.371





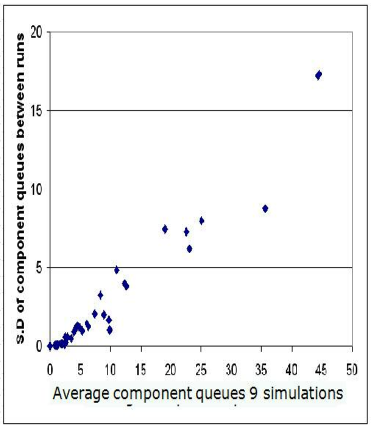


Figure 8 (left). Total queues in test cases simulated and modelled from simulation data

Figure 9 (middle). Modelled from spec. versus simulated component queues in test cases

Figure 10 (right). Standard deviations of simulated component queues between runs

Some questions arise about realism. First, of course, there is the idea of service sharing itself. The degree of sharing ϕ can have a dramatic effect on queue sizes, and the model implementation allows it to be specified. What its value ought to be can be a matter for further theoretical or empirical study.

Second, it has been found in some microscopic simulations that even if straight-moving arrivals select the shorter queue at their moment of arrival, the lane queues may not end up being equal. The longer queue tends to occur on the lane where the turning flow is most dominant (H Gibson, personal communication). A possible explanation for this is that since lane selection affects only straight-movers, as a queue fluctuates they will tend to join it only when it is short, thereby tending to oppose downward fluctuations in queue size. Turners however, having no choice, will not be put off when the queue is long, thereby tending to preserve upward fluctuations. The net result will be to increase the queue on average. This effect may be masked by an algorithm which actively aims to equalise average lane queues.

10 Conclusion

Some theoretical and calculated results have been presented to describe approximately idealised queuing cases involving multiple lanes and turning movements, both in steady state and with time-dependence. In the process it has been shown that the Pollaczek-Khinchin formula cannot accommodate non-independence of lane service processes through its randomness coefficient.

Given the inherent variability road traffic it is seldom possible to make exact predictions and variability may diminish the value of exactitude anyway (except in the somewhat artificial case of scheme benefit comparison). What can be said to be essential is to achieve enough structural conformity to reality that predictions are of the right magnitude, and respond in the right direction to changes in inputs, to enable valid comparisons, conclusions and decisions.

It is believed that the method described embodies some structural validity. Further research could be done in several areas: the effect of service processes which may not be Exponentially distributed or intrinsically independent; the possibility of extending the M/M/c model (or G/G/c if processes depart significantly from Exponential); the prevalence of service sharing in reality; the validity of the symmetrical treatment of turners and non-turners; and actual lane choice behaviour and asymmetric influences on it.

11 Acknowledgments

This work is intended to contribute to the author’s studies towards PhD under the supervision of Prof. Ben Heydecker of University College London. The author is grateful for the support of Dr Alan Stevens, TRL Transportation Chief Scientist, and also for observations by Dr Helen Gibson, Graham Burtenshaw and Ben Morris of TRL who have developed their own micro-simulations of two lane queuing cases. The views and assertions are entirely the author’s and not those of UCL or TRL.

12 References

Kimber R M and Hollis E M (1979). Traffic queues and delays at road junctions. *TRL Report LR 909.* Transport Research Laboratory, Crowthorne House.

Kimber R M, Summersgill I and Burrow I (1986). Delay processes at unsignalised junctions: the interrelation between geometric and queueing delay. *Transportation Research B*, 20B(**6**).

Medhi J (2003). *Stochastic models in queueing theory.* Elsevier Academic Press.

Taylor N B (2003). The CONTRAM dynamic traffic assignment model. *Networks and Spatial Economics 3: (2003) 297-322,* Kluwer Academic Publishers.

Taylor N B (2005). Variance and accuracy of the sheared queue model. *Proc. IMA Conference on Mathematics in Transport.* Institute of Mathematics and its Applications, 7-9 September 2005.