QUEUES WITH TIME-DEPENDENT ARRIVAL RATES
III — A MILD RUSH HOUR

G. F. NEWELL, Institute of Transportation and Traffic Engineering,
University of California, Berkeley

Abstract
The arrival rate of customers to a service facility is assumed to have the form
\( \lambda(t) = \lambda(0) - \beta t^2 \) for some constant \( \beta \). Diffusion approximations show that
for \( \lambda(0) \) sufficiently close to the service rate \( \mu \), the mean queue length at time 0
is proportional to \( \beta^{-1/5} \). A dimensionless form of the diffusion equation is
evaluated numerically from which queue lengths can be evaluated as a function
of time for all \( \lambda(0) \) and \( \beta \). Particular attention is given to those situations in
which neither deterministic queueing theory nor equilibrium stochastic
queueing theory apply.

1. Introduction

In I and II [1, 2] we considered the evolution of a queue during a rush hour
in which the arrival rate \( \lambda(t) \) exceeded the service rate \( \mu \) by so much and/or for
so long a time that the deterministic queueing models gave a correct first approx-
imation to the queue behavior. We also saw that if the rush hour was too weak,
then the deterministic models would fail, and so would the analysis described
in Parts I and II. The failure, however, is not necessarily a failure of the appli-
cability of diffusion approximations, rather it is failure of the methods for de-
termining solutions of the diffusion equations.

Here we shall reexamine the solution of the diffusion equations under somewhat
different conditions corresponding to a mild rush hour. We will be concerned
mostly with the queue behavior near such time when \( \lambda(t) \) goes through a maximum.
For a small range of time near this maximum it is reasonable to assume that
\( \lambda(t) \) is a quadratic function of \( t \). For the present analysis it is convenient to choose
\( t = 0 \) to be the time when \( \lambda(t) \) is maximum (unlike the choice of time origin
in Parts I and II), i.e.,

\[
\begin{align*}
\lambda(t) &= \lambda(0) - \beta t^2, \\
a(t) &= \mu - \lambda(t) = a(0) + \beta t^2,
\end{align*}
\]

Received 20 November 1967.
for some constants \(\lambda(0)\) and \(\beta, \beta > 0\). Also unlike the conditions in I and II, we do not necessarily restrict \(a(0)\) to be negative. A negative \(a(0)\) means that the system is oversaturated at the peak arrival rate, \(\lambda(0)\). Since we will be concerned only with a short (in some sense) range of \(t\), we shall also assume that the coefficient \(b(t)\), I(3)*, is independent of time

\[
b(t) = b = (I_\lambda + I_u)\mu.
\]

The diffusion equation I(2) now takes the form

\[
\frac{\partial F(x, t)}{\partial t} = \left(\alpha(0) + \beta t^2\right) \frac{\partial F(x, t)}{\partial x} + \frac{b}{2} \frac{\partial^2 F(x, t)}{\partial x^2}.
\]

Again it is convenient to rescale the length and time coordinates. Let

\[
t^\dagger = t/T^\dagger, \quad x^\dagger = x/L^\dagger.
\]

But this time we will choose \(T^\dagger\) and \(L^\dagger\) so that the coefficients of \(t^2\) and the second derivative become 1 and 1/2 respectively, i.e.,

\[
\frac{\beta T^\dagger^3}{L^\dagger^2} = 1 \quad \text{and} \quad \frac{b T^\dagger}{L^\dagger^2} = 1
\]

or

\[
L^\dagger = \frac{b^{3/5}}{\beta^{1/5}} \quad \text{and} \quad T^\dagger = \frac{b^{1/5}}{\beta^{2/5}}.
\]

Equation (4) then becomes

\[
\frac{\partial F}{\partial t^\dagger} = \left(\epsilon + t^\dagger^2\right) \frac{\partial F}{\partial x^\dagger} + \frac{1}{2} \frac{\partial^2 F}{\partial x^\dagger^2}
\]

with

\[
\epsilon = \frac{\alpha(0) T^\dagger}{L^\dagger} = \left[\mu - \lambda(0)\right] \beta^{-1/5} b^{-2/5}.
\]

The form of (3a), which involves only one parameter \(\epsilon\) along with numbers of order 1, suggests immediately that the qualitative nature of the solution of (3a) will depend upon whether \(|\epsilon| = O(1)\) or \(|\epsilon| \gg 1\). If \(\epsilon\) is negative and \(|\epsilon| \gg 1\), the queue is highly oversaturated at \(t^\dagger = 0\), in which case the theory described in Parts I and II should apply. If \(\epsilon\) is positive and \(|\epsilon| \gg 1\), the queue is highly undersaturated, in which case the queue distribution should stay close to the prevailing equilibrium distribution.

The prototype of the cases \(\epsilon = O(1)\) is the case \(\epsilon = 0\) in which the arrival rate increases to a maximum exactly equal to the service rate, and then decreases

---

* Equations from Parts I or II will be designated by I(·) or II(·).
Queues with time-dependent arrival rates. III: A mild rush hour

again. One would expect from the form of (3a) that the queue lengths in this case will be of order $x^\dagger = 1$ over times $t^\dagger$ of order 1, or, in the original units, the queue will be of order $L^\dagger$ for $t$ of order $T^\dagger$. The actual queue will be proportional to $\beta^{-1/5}$. This result is the analogue of the $\alpha^{-1/3}$ dependence of the queue length in I (for which there was no intuitively obvious explanation either). Certainly for $\beta \to \infty$ the arrival rate reaches a maximum and returns so quickly that there is no time for a queue to form. The queue length should, therefore, be zero. But if $\beta \to 0$, the arrival rate stays arbitrarily close to saturation for a very long time and the queue will try to reach the equilibrium condition which gives infinite queues.

To establish the connection with the results of Parts I and II we consider first the queue behavior for $\epsilon > 0$, $\eta > 1$, and the conditions for an approximate equilibrium distribution. We will then focus our attention on transition conditions where $t^\dagger$, $x^\dagger$ and $\epsilon$ are all of order 1.

2. Highly oversaturated queues

In Section 4 of II, we considered in some detail the example in which $a(t)$ was a quadratic function of $t$. In II we chose the time origin as the first time $a(t)$ vanished whereas here we prefer to measure time from the time $a(t)$ has a maximum. In terms of the time $t$ of II, $a(t)$ had the form $-t + t^2/\gamma$, which had a maximum at $t = \gamma/2$. In the present notation

$$a(t) = -(t + \gamma/2) + (t + \gamma/2)^2/\gamma = -(\gamma/4) + t^2/\gamma, \quad b(t) = 1,$$

i.e., $a(0) = -\gamma/4$ and $\beta = 1/\gamma$. These parameters in turn give

$$L^\dagger = \gamma^{1/5}, \quad T^\dagger = \gamma^{2/5}, \quad \epsilon = -\gamma^{6/5}/4.$$

The condition for the validity of the results in I and II was $\gamma > 1$ (actually the theory was reasonably correct for $\gamma \gtrsim 3$). Thus we would expect solutions of (3a) to map into solutions of II if $\epsilon \lesssim -1$, or conversely, the solutions of II to map into solutions of (3a).

To find the solution of (3a) in terms of the parameter $\epsilon$ for $\epsilon \lesssim -1$, from the results of II, we first evaluate $\gamma$

$$\gamma = (-4\epsilon)^{5/6}.$$

For this $\gamma$, the solutions from II must now be translated in time $t \to t - \gamma/2$ and then rescaled by dividing the lengths and times of II by

$$L^\dagger = (-4\epsilon)^{1/6} \quad \text{and} \quad T^\dagger = (-4\epsilon)^{1/3}.$$

The description of qualitative features of Figure 2 of II translates into the present notation as follows. The queue reaches saturation at $t = -(-\epsilon)^{1/2}$, (instead of 0), at which time the queue length is of order $L^{-1} = (-4\epsilon)^{-1/6}$, (instead
of $1).$ The transition through saturation takes place with a time of order $T^{-1} = (-4\varepsilon)^{-1/3}$, (instead of $1).$ The mean queue reaches its maximum value at $t = +(-\varepsilon)^{1/2}$, (instead of $\gamma$) and has a value of order $(-4\varepsilon)^{3/2}$, (instead of $\gamma^2$). The standard deviation of the queue length at $t = 0$ is of order $(-4\varepsilon)^{1/4}$, (instead of $\gamma$). The final transition back to equilibrium occurs near $t = 2(-\varepsilon)^{1/2}$ over a time interval, the duration of which is of order $(-4\varepsilon)^{-3/4}$. The equilibrium queue length after this transition is of order $(-4\varepsilon)^{-1}$. There are seven different powers of $\varepsilon$ involved (there were only six powers of $\gamma$ described in II but the units of time and length were chosen so that the queue length and the transition time had the same order, namely $1$).

The maximum expected queue at $t = +(-\varepsilon)^{1/2}$ is actually given approximately as

$$\frac{1}{6}(-4\varepsilon)^{3/2} + (0.95)(-4\varepsilon)^{-1/6}.$$

### 3. Undersaturated conditions

We saw in I that a queue should stay close to equilibrium if the coefficient $a(t)$ is positive and changes by only a small fraction of itself in a relaxation time $T_0 = b(t)/a^2(t)$, i.e.,

$$\frac{|a(t + o(T_0)) - a(t)|}{a(t)} \ll 1.$$

(In I this condition was expressed in terms of derivatives of $a(t)$, but we should be a little more careful here because $da(t)/dt$ vanishes at $t = 0$). As applied to (3a) this condition is equivalent to

$$\frac{2|t|}{(\varepsilon + t^2)^3} + \frac{1}{(\varepsilon + t^2)^5} \ll 1, \quad \varepsilon + t^2 > 0.$$

We are not concerned here with the case $\varepsilon$ negative and $|\varepsilon| \gg 1$ which was treated previously in II. If $-1 \leq \varepsilon < 0$, (10) is satisfied only if $|t| \gg 1$. In fact it should suffice for $(\varepsilon + t^2)$ to be larger than about 4. For $-(-\varepsilon)^{1/2} < t < (-\varepsilon)^{1/2}$, the equilibrium condition will certainly fail because the system is oversaturated. For $\varepsilon = 0$, (10) is equivalent to $|t|^{-5} \ll 1$ and the equilibrium condition should become quite effective by the time $|t|$ exceeds about 2.

For $\varepsilon > 0$, (10) will hold even at $t = 0$ if $\varepsilon^{-5} \ll 1$, i.e., for $\varepsilon \gtrsim 2$. But $t = 0$ is not the most critical time. The first term of (10) has a maximum as a function of $t$ when $t^2 = \varepsilon/5$ where the left hand side of (10) becomes approximately $\frac{1}{6}\varepsilon^{-3/2}$. For slightly larger $\varepsilon$, $\varepsilon \gtrsim 3$; this however is also small.

The equilibrium distribution is

$$F(x,t) = 1 - \exp[-2(\varepsilon + t^2)x].$$
and has a mean

$$E\{X(t)\} \sim \frac{1}{2(\epsilon + t^2)}.$$  

This is a simple symmetric function of $t$ with a maximum at $t = 0$. It falls off within a time of order $(\epsilon^{1/2})$. The maximum queue is $1/(2\epsilon)$.

4. Numerical calculations, $\epsilon = 0$

Equation (3a) was solved numerically by the same type of procedure as used in Part I, i.e., we replaced the differential equation by a finite difference equation, which could also be interpreted as the exact equation for a hypothetical queueing process in discrete time or of a random walk. The only difference is that Equation I (27) is replaced by

$$p_j = p_0 - \beta j^2/2, \quad 1 - p_j = 1 - p_0 + \beta j^2/2,$$

so that Equations I (28) are the discrete analogue of

$$\frac{\partial F_j(k)}{\partial j} = \left[1 - 2p_0 + \beta j^2\right] \frac{\partial F_j(k)}{\partial k} + \frac{1}{2} \frac{\partial^2 F_j(k)}{\partial k^2},$$

i.e., Equation (3) with $a(0) = 1 - 2p_0$ and $b = 1$.

Equation (14) has two parameters $1 - 2p_0$ and $\beta$. In order that queues be large compared with the integer scale of the difference equation and vary slowly relative to the integer time scale, both $1 - 2p_0$ and $\beta$ should be small compared with 1. To test the convergence of the difference equation solution to a limit solution, we first set $1 - 2p_0 = 0$, ($\epsilon = 0$) and solved Equations I (20) and (13) for $\beta = 10^{-2}$, $10^{-4}$, and $10^{-6}$. These values of $\beta$ are roughly comparable with the sequence of $\alpha$ values in I. The value of $p_j$ goes from $1/2$ to 0 as $|j|$ varies from 0 to $10^2$, $10^3$ respectively.

Figure 1 shows the mean queue length in units of $L^i$ versus time in units of $T^i$, analogous to Figure 2 of Part I, for $\beta = 10^{-2}$, $10^{-4}$, and $10^{-6}$. For $\beta = 10^{-2}$, the range of $t^i$ runs only to $|t^i| = 1.2$ because $p_j$ becomes zero by then. As was the case for $\alpha = 10^{-1}$ in Part I, the curve $\beta = 10^{-2}$ is surprisingly close to the others even though the postulates for the diffusion approximation are not valid for $\beta = 10^{-2}$. For the curve $\beta = 10^{-4}$, the range of $|t^i|$ runs only to about 2.5 before $p_j$ becomes 0 at $|j| = 100$. The postulate that the variance per step of the random walk, $4p_j(1 - p_j)$, remains nearly constant is not satisfied in this case either over the range of $t^i$ shown in Figure 1.

For $\beta = 10^{-6}$, however, $p_j \sim 0.35$ and $4p_j(1 - p_j) \sim 0.9$ at $|j| = 550$; the variance stays close to 1 over most of the range $|t^i| < 2$. Despite the apparently slow convergence as $\beta$ decreases to $10^{-6}$, the curve for $\beta = 10^{-6}$ should be rather close to the limit curve for $\beta = 0$. This is further indicated...
by the fact that, as $\beta$ decreases, the successive curves appear to shift upward with a nearly constant displacement. But for $\beta \to 0$ we should expect the mean to stay below the broken line curve for the mean of the equilibrium distribution, $\frac{1}{2}t^{\dagger -2}$, for $t^{\dagger} < 0$, and approach the equilibrium mean from above as $t^{\dagger} \to +\infty$. The curve labeled $\beta = 0$ was drawn by simply displacing the curve $\beta = 10^{-6}$ upward by about 0.03. The curve $\beta = 0$ now approaches the equilibrium curve very rapidly at both $t^{\dagger} \sim -2$ and $t^{\dagger} \sim +2.5$. Certainly Figure 1 shows that the scaling of coordinates by $T^{\dagger}$ and $L^{\dagger}$ is very effective in producing similar curves over a wide range of $\beta$.

The shape of the curves in Figure 1 confirms that as $t^{\dagger}$ increases toward $-2$, the mean queue is trying to stay close to the mean of the prevailing equilibrium distribution. As $t^{\dagger}$ increases further, the equilibrium mean is increasing too fast; the actual queue cannot adjust to it quickly enough. At time 0, when the arrival rate is equal to the service rate, the equilibrium mean is infinite, but

$$E\{X(0)\} \sim 0.75L^{\dagger}$$

and still increasing. Before the queue can increase much further, the arrival rate has started to decrease again. The mean queue reaches a maximum value of

$$E\{X(t)\} \sim 0.92L^{\dagger} \text{ at } t \sim 0.75T^{\dagger}.$$  

By this time the mean of the equilibrium distribution has decreased, and the actual mean has about caught up with it (although the actual queue distribution is not necessarily of the same shape as the equilibrium distribution). Still later the actual queue tries to catch an equilibrium distribution, the mean of which is decreasing too fast. The actual mean finally converges to the equilibrium mean at $t^{\dagger} \sim 2.5$. 

---

Figure 1
Figure 2 shows similar curves for the variance of the queue in units of $U^2$ versus $t^\dagger$. As was true also in Figure 3 of I, the variance is much less sensitive to $\beta$ than the mean. We did not even try to extrapolate to $\beta = 0$. These curves for the variance have a shape similar to those for the mean. For $t^\dagger \sim -2$ or $t^\dagger \sim +2.5$, they are very close to the equilibrium variance, $t^\dagger -4/4$. At $t = 0$, the variance is about $0.34U^2$ and rises to a maximum of about $0.58U^2$ at $t \sim 0.95T^\dagger$. 

![Figure 2](image-url)

![Figure 3](image-url)
As one might expect, the maximum variance occurs later than the maximum mean. As the queue was growing, the distribution was developing a long tail. Having done so, it takes a long time to pull it down.

Figure 3 shows a few of the queue distributions evaluated for $\beta = 10^{-6}$. Unlike the distributions of Figure 4 in Part I which showed tendencies toward a normal distribution, these smoothed distributions all show a maximum density at queue length zero (actually $-1$ because of the way the step function was smoothed). The interesting feature of these curves is the relative shapes as the mean is increasing or decreasing. The curve for $t^* = 1.4$, for example, has a smaller mean than for $t^* = -0.2$, but a much longer tail. As the arrival rate is increasing, the tail of the distribution lags behind what the equilibrium would predict, but when the arrival rate starts to decrease again, the tail is slow to be dissipated.

![Figure 4](image)

5. **Numerical calculations, $\varepsilon \neq 0$**

Having established that for $\beta = 10^{-6}$, the solution of the finite difference equations with $p_0 = 1/2$ gives a reasonable approximation to the solution of the differential equation (3a) for $\varepsilon = 0$, we now keep $\beta$ at $10^{-6}$ and vary $p_0$ in order to investigate how the solution of (3a) depends upon $\varepsilon$. 
Figure 4 shows the mean queue length versus time for $\varepsilon = -0.8, -0.5, 0, +1,$ and $+2$ as determined from the numerical solution of the finite difference equations with $\beta = 10^{-6}$. The curve $\varepsilon = 0$ is identical to that of Figure 1 for $\beta = 10^{-6}$. The scales of queue length and time are shown both in the original integer units of the finite difference equation and in the “non-dimensional” units of $L_t^\prime$ and $T_t^\prime$, the conversion factor from one to the other being independent of $\varepsilon$.

On the three curves for $\varepsilon \leq 0$ ($\varepsilon = 0$, $\varepsilon = -0.5$ and $\varepsilon = -0.8$), the times $t = \pm (\varepsilon)^{1/2}$ where the arrival rate and service rates are equal, $a(t) = 0$, are marked by the vertical crosses on these curves. For $\varepsilon = 0$, the maximum queue occurs well after the time $t = (-\varepsilon)^{1/2} = 0$, but as $\varepsilon$ becomes more negative, the maximum queue occurs nearer the time $t = (-\varepsilon)^{1/2}$, where the deterministic theory or the theory of II predicts the maximum should occur.

An attempt was made to evaluate the mean queue for $\varepsilon = -0.8$ also by the methods of II. These methods consist of first calculating the mean queue through the first transition near $t = (-\varepsilon)^{1/2}$ according to the procedure of I, then for later times applying a constant correction $(0.95)(-4\varepsilon)^{-1/6}$ as in (9) to the deterministic queueing theory; and finally using an equation like II(31) to give the effects of reflection off the barrier at $x = 0$ in the final transition to equilibrium.

The results of the first two steps are shown by the broken line curve. This is obviously quite accurate until $t^\prime \sim 1$ where the final corrections for reflection start to show some influence. We cannot apply II (31) directly, however, because the part of the probability mass which has hit the barrier is put into the prevailing equilibrium distribution to obtain II (31). If some probability has hit the barrier already before the queue has become undersaturated, one obviously cannot put it into the non-existent equilibrium distribution, one must put it into a distribution with a finite mean, perhaps comparable with the equilibrium distribution associated with some typical time during the final transition.

Although there is some uncertainty as to how one should correct for this reflection off the boundary, the value of the mean queue is not very sensitive to what one does. In the early stages very little probability is reflected anyway and it could not have acquired a very large mean. By the time a considerable probability has been reflected, the equilibrium distribution is well-defined and the difficulty disappears. With any sensible corrections, the theory of Part II continues to agree to within 5% or so with the solid curve for $\varepsilon = -0.8$, also for $t^\prime > 1$. Since the theory of Part II works reasonably well for $\varepsilon = -0.8$, it obviously should work even better for $\varepsilon < -0.8$.

For positive $\varepsilon$, we see from Figure 4 that as $\varepsilon$ goes from 0 to 1 to 2, the curves for mean queue versus time become more nearly symmetric about $t^\prime = 0$. The broken line curve for $\varepsilon = 1$ is the equilibrium mean associated with the finite difference equations (the geometric distribution rather than exponential distribution of the differential equation). A corresponding equilibrium curve for $\varepsilon = 2$...
would be so close to the solid line curve as to hardly show on a graph of this scale. Thus the equilibrium theory should apply very well for \( \varepsilon \geq 2 \).

Figure 5 shows corresponding curves for the variance. These curves are of similar shape to those of Figure 4 but more asymmetric. For \( \varepsilon < 0 \), the variance will continue to grow after the mean has reached its peak. It will not start to decrease until there is a significant reflection off the barrier.

For \( \varepsilon = -0.8 \), the theory of Part II does not work as well for calculation of the variance as for the mean. It only gives rather crude estimates (within about 20\%). For the theory of Part II to apply here one should have \( \varepsilon \) less than about \(-1\). One can also see from the curve for \( \varepsilon = +2 \) that as \( \varepsilon \) increases the variance is slower to converge to a symmetric shape than the mean of Figure 4.

Finally Figure 6 shows some curves for the maximum of the expected queue length, \( \max_t E\{X(t)\} \), versus \( \varepsilon \). The curve labeled deterministic theory is the first term of (9), \((-4\varepsilon)^{3/2}/6\). The curve labeled \( \varepsilon < 0, |\varepsilon| \gg 1 \) is Equation (9). The curve \( \varepsilon > 0, \varepsilon \gg 1 \) is \( 1/(2\varepsilon) \) for the equilibrium distribution. The five circled points are calculated values for \( \varepsilon = -0.8, -0.5, 0, +1, \) and \(+2\). These values are slightly higher (by an amount 0.031) than those shown in Figure 4. They
have been corrected for the difference between calculations for $\beta = 10^{-6}$ and $\beta = 0$. As explained below, the difference between the means for $\beta = 10^{-6}$ and $\beta = 0$ seems to be essentially the difference between the means of a geometric and an exponential distribution. Although this correction is not very important for $\varepsilon < 0$, where the means are large, it does make some difference as $\varepsilon$ becomes positive and the queues relatively short. This figure shows that even with only five values of $\varepsilon$, one can easily interpolate between the two asymptotes for $\varepsilon < 0$ and $\varepsilon > 0$ and show that the queue behavior is a smooth function of $\varepsilon$ over the range of $\varepsilon = O(1)$.

6. Accuracy of calculations

The numerical calculations described here and in Part I were all done in a total of about 8 hours of computer time on a slow computer (IBM 1620) with no special algorithms to achieve high accuracy or rapid convergence for $\alpha \to 0$ or $\beta \to 0$. If the object were to obtain accurate solutions of the diffusion equation
and tabulate the solutions as a function of $\varepsilon$, we could easily reduce the errors by a factor of 10, 100 or even 1,000 by using better difference-equation approximations to the differential equation and/or faster computers, but it is not obvious that this would be of much practical value. The crude calculations described here probably furnish more insight into the accuracy of diffusion approximations and their applications than more accurate calculations.

In Figure 1 we saw that the curves of mean queue versus time for different $\beta$ were nearly the same except for a uniform displacement. We also saw in Figure 3 that although the shape of the queue distributions varied with time, they were all reasonably similar in shape to an exponential or geometric distribution. The equilibrium distribution $I_2$ (29) corresponding to a given value of $p$ has a mean

$$\frac{p}{1 - 2p} = \frac{1}{2(1 - 2p)} - \frac{1}{2},$$

whereas the mean for the exponential equilibrium distribution of the corresponding differential equation is $[2(1 - 2p)]^{-1}$. They differ by 1/2, independent of $p$. This is a difference of 1/2 in the difference equation units of queue, a difference of $1/2(Lt)$ in the “non-dimension” units.

In Figure 1, the difference between the curves for $\beta > 0$ and $\beta = 0$ is almost exactly $1/(2L) = \beta^{1/5}/2$ uniformly in $t^*$ (even for the crude calculation with $\beta = 10^{-2}$). Although the distributions are not geometric for all $t^*$, they apparently are sufficiently close to geometric as to yield a nearly uniform displacement as $\beta$ changes. This correction of $\beta^{1/5}/2 = 0.031$ was used to construct Figure 6 where we wanted to show the smooth transition to the equilibrium mean of an exponential distribution as $\varepsilon$ became large.

If we wished to obtain more accurate estimates of the limiting queue distribution for $\beta \to 0$, we obviously should not simply repeat the calculations described here for smaller and smaller values of $\beta$, because the convergence to a limit apparently goes as $\beta^{1/5}$. A reduction of $\beta$ by a factor of $10^{-5}$ would reduce the error in the mean by a factor of only $10^{-1}$, but increase the computation time beyond reasonable limits (by a factor of perhaps $10^{10}$). To achieve high accuracy we must use more efficient numerical schemes to solve the differential equation.

In Figure 2 of Part I, we had an analogous situation. We see that as $\alpha \to 0$, the mean queue has a slow convergence for $t^* \sim -2$ but a very rapid convergence for $t^* \sim +2$. For $t^* \sim -2$ the queue distribution is close to equilibrium and so the difference between the means for $\alpha = 0$ and $\alpha > 0$ should be approximately $1/2$ in the original units, or $\alpha^{1/3}/2$ in the nondimensional units of Figure 2 of Part I. For $t^* > 0$, however, the distribution starts to approach a normal distribution that moves away from the boundary $x^* = 0$. It seems quite clear that if the goal is to obtain better approximations to the limit distribution, we could achieve this to a large extent merely by modifying the reflecting condition on
the difference equation, i.e., change the transition probabilities from the state 
j = 0 (it appears that one should make the random walk go to state 1 with prob-
ability 1 if it starts at state 0). Again, one cannot achieve much improvement
in accuracy merely by repeating the calculations of I for smaller \( \alpha \) because the
error decreases too slowly as \( \alpha \to 0 \).

From the point of view of applications it is quite relevant that the solution
of the differential equation is quite similar to the solution of the finite difference
equation even for \( \alpha = 10^{-1} \) in I or \( \beta = 10^{-2} \) in III, but that the rate at which
the two solutions approach each other as \( \alpha \) or \( \beta \to 0 \) is very slow. The finite
difference equations used here could have been the exact equations for some
hypothetical queueing system, so also could some other set of difference equations
with different transition probabilities from the state \( j = 0 \). We would expect,
however, the same qualitative behavior if we had approximated the differential
equation by equations of some more realistic queueing systems in continuous
time. We would expect the differential equation to approximate the solution
to a much more general sort of queueing system to within an accuracy of about
one customer, as compared with a mean queue length measured in units of \( L \)
or \( L' \).

In typical applications, the values of \( \alpha \) or \( \beta \) could be smaller than the smallest
values \( \alpha = 2 \times 10^{-3} \) or \( \beta = 10^{-6} \) used here, but not by more than one or two
powers of 10. With this slow rate of convergence a few extra powers of 10
will not make much difference, however. We expect the accuracy of the
calculations described here to be typical, in order of magnitude at least, of what
one would encounter in applications. The diffusion approximations should almost
always give good rough estimates (within 10 or 20\%) of the queue size for moder-
ately large queues, but hardly ever results of high precision (less than 1\%).
There does not, therefore, seem to be much motivation here for determining
the solution of the diffusion equation to high accuracy.

7. The maximum queue

In II we considered the distribution function \( G(z) \) for the maximum queue \( \operatorname{max},
X(t) \). The upper bound II (15) for \( G(z) \) was derived by neglecting the possibility that
the maximum occurs before some time \( t \) or that the maximum occurs after a
reflection from the barrier at a time later than \( t \). This bound is valid regardless
of whether or not the rush hour is a severe one in which the time \( t_1 \) of II is
larger than \( T = T_1 \). In II this condition simply guaranteed that the bound was
a close one, so that II (15) could be used to obtain good numerical estimates
of \( G(z) \).

If we translate II (13) and II (15) into the present notation, we must change
the time coordinate, \( t \to t - e^{t/2} \). The special form of \( a(t) \) in (1) is symmetric
about \( t = 0 \), which implies that Equation II (19) for \( G^{**} \) is identical to the equation for \( F(x, t) \) except that \( t \) is changed to \(-t\). Thus II (13), II (15) become

\[
G(z) \leq G_t(z) = \int d_x F(x, t) F(z - x, -t)
\]

for all \( t \) and, in particular,

\[
G(z) \leq G_0(z) = \int d_x F(x, 0) F(z - x, 0).
\]

Thus

\[
(15) \quad E\{\max X(t)\} \leq 2E\{X(0)\},
\]

for any \( \varepsilon \).

In accordance with the conclusions of II, this bound should be close if \( \varepsilon < 0 \) \( |\varepsilon| > 1 \), (actually they should be fairly close even for \( \varepsilon \leq -1 \)). From the nature of the postulates used to derive the bound, one might even expect the estimate to be qualitatively correct (within 20\% or so) even for \( |\varepsilon| \leq 1 \).

For \( \varepsilon < 0 \), \( |\varepsilon| \gg 1 \)

\[
2E\{X(0)\} = (-4\varepsilon)^{3/2}/6 + 2(0.95)(-4\varepsilon)^{-1/6}
\]

i.e., the expected maximum \( E\{\max X(t)\} \) differs from the deterministic theory estimate, \((-4\varepsilon)^{3/2}/6\), by just twice that for the maximum expected queue.

For \( \varepsilon = -0.8, -0.5 \), etc., the values of \( 2E\{X(0)\} \) can be read from Figure 4. The broken line of Figure 6 shows this lower bound for \( E\{\max X(t)\} \).

The relative sizes of \( \max E\{X(t)\} \) and \( E\{\max X(t)\} \) vary considerably as \( \varepsilon \) goes from large negative values to large positive values. For large negative values we have seen that the difference between these two is not only fractionally small, it is even small compared with the standard deviation of \( X(t) \) near \( t \sim |\varepsilon|^{1/2} \). This is true because \( \text{Var} X(t) \) represents the cumulative uncertainty in the queue which is built up from all fluctuations over the time \( t \sim -|\varepsilon|^{1/2} \) to \( +|\varepsilon|^{1/2} \).

For any realization of \( X(t) \) the difference between \( \max X(t) \) and \( X(|\varepsilon|^{1/2}) \) is, however, determined by the amount of fluctuation that can occur within a relatively short time around \( t = |\varepsilon|^{1/2} \). Specifically, this difference is of order \( |\varepsilon|^{-1/6} \) whereas the maximum queue is of order \( |\varepsilon|^{3/2} \) and the standard deviation is of order \( |\varepsilon|^{1/4} \).

For \( \varepsilon \sim 0 \), \( E\{\max X(t)\} \) is larger than \( \max E\{X(t)\} \) by 50\% or more. We now have the situation that for any given time \( t \), the queue could be very small, but if it is small, this does not imply that the queue realization is uniformly small. The queue may vanish at any time but bounce up again.

For large positive \( \varepsilon \), this trend continues further. The queue is undersaturated at all times and any queue realization is likely to vanish many times over the
“rush hour”. Each realization will have many peaks of a magnitude comparable with \( \max_t E\{X(t)\} \), the longest of which is likely to be considerably larger than the mean queue at any time \( t \).

It is possible to obtain limit distributions for \( G(z) \) by methods analogous to those for stationary diffusion processes [3]. The convergence to a simple limit distribution, however, is very slow as \( \varepsilon \to \infty \) and the limit would be useful only for extremely large \( \varepsilon \). It is also possible to obtain analytic approximations for moderate \( \varepsilon \), but they would involve solutions of transcendental equations and generally be quite awkward to evaluate.

The details of this will not be described here. They could be of some mathematical interest, but of questionable practical value. One can show, for example, that for \( \varepsilon \to +\infty \), the distribution \( G(z) \), appropriately rescaled, has the double exponential form characteristic of extreme values with

\[
E\{\max_t X(t)\} \sim \frac{\log \varepsilon}{4\varepsilon} \left[ 1 + O\left( \frac{\log \log \varepsilon}{\log \varepsilon} \right) \right].
\]

This formula is of little use for numerical calculations because \( \log \log \varepsilon/\log \varepsilon \) goes to zero at such a slow rate for \( \varepsilon \to \infty \), but it shows, at least, that for large \( \varepsilon \), the maximum queue is larger than the maximum mean by a significant factor (of order \( \log \varepsilon \)).

The most efficient way to obtain the distribution \( G(z) \) for moderate \( \varepsilon \) is probably by simulation. This has not been done.

8. Typical applications

To give some indication of the order of magnitude of some of the parameters in typical applications, suppose we choose as a unit of time, the time it would require the \( b^2 t^2 \) term in the arrival rate to become comparable with \( \mu \) (if the quadratic form of \( a(t) \) were extrapolated). Thus for \( |t| = 1 \) in the chosen time units \( b^2 t^2 \sim \mu \), i.e., \( b \sim \mu \). In many applications a “rush hour” lasts about an hour, in which case the unit of time is of this order.

We will also take a variance to mean ratio comparable with 1 so that \( b \sim \mu \). From this we have

\[
\nu^* = b^{2/5} \mu^{1/5} = \mu^{2/5}, \quad T^* = \mu^{-1/5}
\]

and

\[
\varepsilon = \mu^{2/5}(1 - \rho), \quad \rho \equiv \mu/\lambda(0),
\]

\( \rho \) being the peak traffic intensity and \( \mu \) the service rate in the chosen time units.

The theory described in Part III is primarily concerned with queue behavior for \( |\varepsilon| \lesssim 1 \), i.e., \( |1 - \rho| \lesssim \mu^{-2/5} \), for which neither deterministic queueing theory nor equilibrium theory applies.
Typical values of $\mu$ cover a wide range of possible values. A typical large highway toll facility could serve of the order of $10^4$ cars per hour. If the rush hour lasts about an hour, this is the time unit and $\mu \sim 10^4$. The condition $|\varepsilon| \lesssim 1$ would imply $|1 - \rho| \lesssim 10^{-3/5}$,

$$0.97 \lesssim \rho \lesssim 1.03,$$

which is a rather narrow range not likely to be realized very often. On the other hand, an airport runway, a grocery store counter, or a bank teller, etc., are likely to have service times of the order of a minute, and a rush hour of about an hour; thus a $\mu$ of order $10^2$. The condition $|\varepsilon| \lesssim 1$ is equivalent to $|1 - \rho| \lesssim 10^{-4/5}$ or

$$0.85 \lesssim \rho \lesssim 1.15.$$

This is quite likely to exist in practice. For $\mu \sim 10^2$, the peak queue length is of order $L^* \sim 10^{4/5} \sim 6$ over a time of order $T \sim 10^{-2/5} \sim 0.4$ (hours).

9. Acknowledgment

The numerical calculations described here were skillfully handled by Frederick L. Collins.

References

