Walter Böhm

Lattice path counting and the theory of queues

Paper

Original Citation:

Böhm, Walter
(2008)
Lattice path counting and the theory of queues.

This version is available at: https://epub.wu.ac.at/1086/
Available in ePub\textsuperscript{WU}: November 2008

ePub\textsuperscript{WU}, the institutional repository of the WU Vienna University of Economics and Business, is provided by the University Library and the IT-Services. The aim is to enable open access to the scholarly output of the WU.
Lattice Path Counting and the Theory of Queues

Walter Böhm

Department of Statistics and Mathematics
Wirtschaftsuniversität Wien

Research Report Series

Report 74
October 2008

http://statmath.wu-wien.ac.at/
Lattice Path Counting And The Theory Of Queues

Walter Böhm

Department of Statistics and Mathematics, Wirtschaftsuniversität Wien, A-1091 Vienna, Austria

Abstract

In this paper we will show how recent advances in the combinatorics of lattice paths can be applied to solve interesting and nontrivial problems in the theory of queues. The problems we discuss range from classical ones like $M^a/M^b/1$ systems to open tandem systems with and without global blocking and to queueing models that are related to random walks in a quarter plane like the Flatto-Hahn model or systems with preemptive priorities.

Key words: lattice paths, multidimensional paths, paths in a quarter plane, Markovian queues, transient analysis.

1 Introduction

It is well known that combinatorial methods provide very elegant and powerful tools of analysis in the theory of queues, see e.g. Takács (1967) or Jain, Mohanty and Böhm (2006). To demonstrate that this is still true is the major purpose of the present paper. In particular we will show how recent advances in the combinatorics of lattice paths can be applied to solve rather hard problems in queueing theory. We start in section 2 by providing a technical device, Champernowne’s Construction, for translating statements about sample paths of Markovian queueing processes to statements about lattice paths. The major advantage of this method is that counting paths in the presence of reflecting barriers, as is typical of almost all Markovian queueing systems, is completely avoided and replaced by the task of counting paths restricted by absorbing bar-
riers which is much easier\footnote{It should be noted that the same purpose could be achieved by another approach, viz. dual processes. For details see Krinik et al. (2005).}. The effectiveness of this device is demonstrated by the transient analysis of the classical $M/M/1$ model. In section 3 we are going considerably beyond simple $M/M/1$ and discuss models with arrivals or service in batches of fixed size. There we present a new formula for the transient distribution based on a combinatorial result due to Niederhausen (1980). In section 4 we introduce a method based on the factorization of formal Laurent series to analyze a queueing model in which customers arrive in batches of random size. Section 5 is devoted to the analysis of systems in which customers arrive and depart in batches of fixed size. The most general counting result in this context yields a determinant. However, we will indicate also that there are alternatives, e.g., the kernel method. As an example we give a complete busy period analysis of a $M^3/M^2/1$ model. Section 6 is devoted to multidimensional paths and their application to queues. We start with the analysis of a queueing system with heterogeneous arrivals. The major topic of this section is Gessel and Zeilberger’s (1992) generalization of the reflection principle. Based on that result we provide a transient analysis of open and closed tandem systems. The formula we proved for the closed tandem system is apparently new. In the last section we embark briefly on the problem of counting paths in a quarter plane. Several interesting queueing models may be formulated within this framework, as we demonstrate by several examples. In particular we give a busy period analysis of the Flatto-Hahn model based on a result due to Bousquet-Mélou (2005). This formula has been derived by the kernel method. In our last example we show how the kernel method can be used to analyze a queueing model with preemptive priorities.

2 Champernowne’s Construction

In his famous paper (Champernowne (1956)) Champernowne provided for the first time a combinatorial solution of the classical $M/M/1$ queueing model. This model is particularly appealing because of its structural simplicity:

- There is a single server, service times are i.i.d. exponential random variables with mean $1/\mu$.
- Customers arrive according to a Poisson process with rate $\lambda$.
- The system can hold an infinite number of customers.

These assumptions guarantee that the process $Q_t$, the number of customers in the system at time $t$, is a Markov process in continuous time, moreover, it is the simplest nontrivial example of a birth-death process with constant birthrate $\lambda$ and constant death rate $\mu$. Basically Champernowne’s idea was
to associate with \( Q_t \) a basic process \( X_t \). This process \( X_t \), sometimes called randomized random walk (Feller (1971, pp. 479)), is Markovian with jumps of magnitude +1 occurring with rate \( \lambda \) and jumps of magnitude −1 occurring with rate \( \mu \). The essential difference between the processes \( Q_t \) and \( X_t \) is that \( Q_t \) is restricted by an impenetrable barrier at zero, thus its state space is \( \mathbb{Z}_0^+ \), whereas \( X_t \) is free to move on the set of all integers \( \mathbb{Z} \).

But what is the connection between \( Q_t \) and \( X_t \)? Champernowne realized that \( Q_t \) and \( X_t \) are linked together by the fundamental relationship

\[
Q_t = \max [X_t - \sigma_t, Q_0 + X_t],
\]

where \( Q_0 \) is the number of customers waiting at time zero when the service station starts up and \( \sigma_t \) is the smallest value the process \( X_t \) attained during the time interval \((0, t)\), i.e.

\[
\sigma_t = \inf_{0 \leq s \leq t} X_s.
\]

Equation (1) is indeed fundamental and holds for a much more general class of stochastic processes than we need here, \( X_t \) might be any separable infinitely divisible process which can be represented as the difference of two infinitely divisible nonnegative processes, as has been proved by Gani and Pyke (1960), but see also Prabhu (1965, pp. 249). In fact, in our case

\[
X_t = A_t - D_t,
\]

where \( A_t \) is a Poisson process with rate \( \lambda \), the arrival process, and \( D_t \) is Poisson with rate \( \mu \), the service process. Equation (1) tells us also that the processes \( Q_t \) and \( X_t \) behave identically as long as the queueing system does not become empty, i.e. \( Q_0 + X_t > 0 \).

The next step is to consider the processes \( X_t \) and \( Q_t \) just immediately after those time instances where jumps occur. Suppose jumps occur at times \( \tau_1, \tau_2, \ldots \) and let \( X_n = X_{\tau_n} \) and \( Q_n = Q_{\tau_n} \). The sample paths of these embedded processes may now be represented by lattice paths in the plane. \( X_n \) has step set \{1, 1\}, \{1, −1\}\} whereas the step set of \( Q_n \) is given by \{(1, 1), (1, 0), (1, −1)\}. Note that for \( X_n \) jumps of size +1 have probability \( \lambda/(\lambda+\mu) \), those of size −1 have probability \( \mu/(\lambda+\mu) \). Figure 1 shows graphically the connection between \( Q_n \) and \( X_n \).

The point now is that by (1) the distribution of \( Q_t \) can be expressed in terms of the probability that the process \( m + X_t, m > 0 \) does does not touch or cross
the boundary at zero. This can be seen as follows:

\[
P(Q_t < k | Q_0 = m) = P(X_t - \sigma_t < k, m + X_t < k) \\
= \sum_{\nu > m} P(X_t - \sigma_t < k, X_t = k - \nu) \\
= \sum_{\nu > m} P(\nu + \sigma_t > 0, \nu + X_t = k) \\
= \sum_{\nu > m} P(\nu + X_t = k, \nu + X_s > 0, 0 \leq s \leq t) \\
= \sum_{\nu > m} P_{\nu,k}^0(t) \quad \text{(say)}.
\]

Observe that

\[
P(\nu + X_t = k, \nu + X_s > 0, 0 \leq s \leq t) = P(Q_t = k, Q_s > 0, 0 \leq s \leq t | Q_0 = \nu).
\]

At this point combinatorics of lattice paths comes into play. The generic term in the summation (2) is the conditional probability that a path of \(X_t\) moves from initial altitude \(\nu\) to terminal altitude \(k\) without touching the boundary at zero, given there are \(a\) steps of size +1 and \(b\) steps of size −1. By the reflection principle the number of such paths is

\[
\binom{a+b}{a} - \binom{a+b}{a+\nu}.
\]

Any such path has probability

\[
\left( \frac{\lambda}{\lambda + \mu} \right)^a \left( \frac{\mu}{\lambda + \mu} \right)^b e^{-(\lambda+\mu)t} \frac{((\lambda + \mu)t)^{a+b}}{(a+b)!}.
\]

Hence

\[
P(\nu + X_t = k, \nu + X_s > 0, 0 \leq s \leq t) =
\]

\[
= \sum_{\nu + a - b = k} \left[ \binom{a+b}{a} - \binom{a+b}{a+\nu} \right] e^{-(\lambda+\mu)t} \frac{\lambda^a \mu^b t^{a+b}}{(a+b)!}
\]

\[
= e^{-(\lambda+\mu)t} \rho^{\frac{k-\nu}{2}} [I_{k-\nu}(2t\sqrt{\lambda\nu}) - I_{k+\nu}(2t\sqrt{\lambda\nu})],
\]

where \(\rho = \lambda/\mu\) denotes the traffic intensity and \(I_n(z)\) denotes the modified Bessel functions defined by

\[
I_n(z) = \sum_{k \geq 0} \frac{(z/2)^{n+2k}}{k!(n+k)!}.
\]

Inserting (4) into (2) yields finally one of the classical representations of the transient distribution of the queue length in an \(M/M/1\) system:

\[
P(Q_t < k | Q_0 = m) = e^{-(\lambda+\mu)t} \sum_{\nu > m} \rho^{\frac{k-\nu}{2}} [I_{k-\nu}(2t\sqrt{\lambda\nu}) - I_{k+\nu}(2t\sqrt{\lambda\nu})].
\]

(5)
For some alternative formulas see e.g. Jain, Mohanty and Böhm (2006, pp 39).

\[ X_n - \sigma_n \]

Fig. 1. The processes \( X_n \) and \( Q_n \) with \( Q_0 = 0 \)

### 3 Going beyond \( M/M/1 \) - Bulk Arrivals and Bulk Service

The reflection principle for lattice paths together with Champernowne’s method outlined in the previous section allows an elementary and elegant derivation of a result which otherwise would have required the use of generating functions and Laplace transforms. Of course there is nothing to say against integral transforms. But the point is that Champernowne’s method exploits the simple probabilistic structure of the model by transforming the problem into one of counting weighted lattice paths thereby providing a transparent and constructive analysis. This has to be seen in contrast to alternative approaches like the Difference-Equation technique (Bailey (1954), Conolly (1958)) or the Spectral Method (Ledermann and Reuter (1954)). For an overview see e.g. Medhi (1991, pp. 114) or Jain, Mohanty and Böhm (2006, chapter 2).

Clearly, we cannot expect such a simple technique of solution in more complicated queueing models, in particular non-Markovian systems. In such cases generally (though not always) our odds are bad, but there are good chances that methods based on the combinatorics of paths might be applied successfully to queueing models much more general than simple \( M/M/1 \). So the question is: how far can we go?

In order to give an impression of the amazing answer to this question, let us
begin by allowing customers to arrive in bulks or batches.

**Bulk Arrivals.** In the simplest case customers arrive in groups of constant size $r$ with rate $\lambda$. All other assumptions of the model (see the beginning of section 2) remain unchanged. This model, in standard notation $M^r/M/1$, is particularly interesting for two reasons. Firstly, it requires the solution of an interesting path counting problem, and secondly, the model $M^r/M/1$ is actually equivalent to the system $M/E_r/1$, i.e., a single server system in which service times have an Erlangian distribution, thus we may mimic in this way a queueing system which is no longer Markovian.

We can apply the ideas of Champernowne’s construction again. With the process $Q_t$, we associate the basic process $X_t$ which is now a randomized random walk having jumps of size $+r$ with rate $\lambda$ and jumps of size $-1$ with rate $\mu$. The representation (1) still holds and therefore also

$$P(Q_t < k \mid Q_0 = m) = \sum_{\nu > m} P(\nu + X_t = k, \nu + X_s > 0, 0 \leq s \leq t)$$

$$= \sum_{\nu > m} P_{\nu,k}^0(t). \quad (6)$$

Calculation of $P_{\nu,k}^0(t)$ requires the determination of the number of paths going from initial altitude $m > 0$ to terminal altitude $k > 0$, such that the paths do not touch or cross the boundary at zero. Let us denote this number by $N_{m,k}(a)$, where $a$ is the number of arriving batches of size $r$. This time, however, the step set of the paths is different. Now paths have steps in the set $S = \{(1, r), (1, -1)\}$. Let

$$A_{m,k}(a) = \# \text{ of unrestricted paths from altitude } m \text{ to altitude } k$$

$$B_{m,k}(a) = \# \text{ of paths from altitude } m \text{ to altitude } k, \text{ which touch or cross the boundary at zero}$$

Clearly the following decomposition holds:

$$N_{m,k}(a) = A_{m,k}(a) - B_{m,k}(a). \quad (7)$$

Now $A_{m,k}(a)$ is easy to find. Since paths move from altitude $m$ to $k$ and have $a$ steps up, there must be $m - k + ar$ down-steps, hence

$$A_{m,k}(a) = \binom{m - k + ar + 1}{a}. \quad (8)$$

The second part, determination of $B_{m,k}(a)$, is not that easy, since for such paths the reflection principle will not work. Looking at Figure 2 we can see, that paths which touch the boundary can be split into two parts: a path segment starting at altitude $m$ which touches the boundary for the first time after a certain number of steps, followed by a segment which starts at altitude
0 and moves to terminal altitude $k$. The latter segment is an arbitrary path without any restriction. The first segment may be counted by a generalization

$$m \quad m + i(r + 1) \binom{m + i(r + 1)}{i}.$$

The second segment having $j$ up-steps must have $jr - k$ down-steps and it is counted by the binomial coefficient

$$\binom{j(r + 1) - k}{j}.$$

By forming the convolution of these two expressions we obtain

$$B_{m,k}(a) = \sum_{\ell \geq 0} \frac{m}{m + \ell(r + 1)} \binom{m + \ell(r + 1)}{\ell} \binom{(a - \ell)(r + 1) - k}{a - \ell}.$$

And this finally determines (7) together with (8). In this formula and also in subsequent ones we use the following convention regarding binomial coefficients: $\binom{x}{y}$ is defined to be zero when $x < 0$.

**Remark.** Formula (9) has a long and interesting history, see Mohanty (1979, pp. 7) for details. There are several ways to prove this formula, we will see one proof based on generating functions and Lagrange inversion below.

An alternative formula for $N_{m,k}(a)$ has been given by Niederhausen (1980):

$$N_{m,k}(a) = \sum_{\ell \geq 0} (-1)^\ell \frac{m}{m + (r + 1)(a - \ell)} \binom{m + (r + 1)(a - \ell)}{a - \ell} \binom{k - r\ell - 1}{\ell}.$$

$$7$$
Actually, from a numerical point of view (10) seems to be superior to $A_{m,k}(a) - B_{m,k}(a)$.

Now we are in a position to assign probabilities to steps, in particular we assign $\lambda/(\lambda + \mu)$ to up-steps and $\mu/(\lambda + \mu)$ to down-steps. Upon summation of all possible values of $a$, we obtain

$$P^0_{m,k}(t) = e^{-(\lambda+\mu)t} \sum_{a \geq 0} N_{m,k}(a) \frac{\lambda^a \mu^{m-ar-k} t^{m+a(r+1)-k}}{(m+a(r+1)-k)!},$$  \hspace{1cm} (11)

and this in turn determines the general time-dependent distribution of $Q_t$ via (6).

**Bulk Service.** In this model we assume that customers arrive one after the other with rate $\lambda$ and are served in batches of constant size $r \geq 1$ with mean service time $1/\mu$. Service times and interarrival times are still exponentially distributed so that the Markov property is retained. However, one additional assumption is needed. If at any time $t$ the number of customers in the system is $< r$ and a new service is to start, then the server does not wait until there is a complete batch of size $r$, instead he takes whatever number of customers is in the system at that time. This queueing system, denoted by $M/M^r/1$ can be shown to be equivalent to the model $E_r/M/1$, thus again we can mimic a non-Markovian queueing system (see Jain, Mohanty, Böhm (2006, pp. 84)).

Also in this model Champernowne’s representation (1) of the process $Q_t$ in terms of a basic process $X_t$ holds and again $P(Q_t < k|Q_0 = m) = \sum_{a>m} P^0_{\nu,k}(t)$. This time the lattice paths of the embedded process $X_n$ have step set $S = \{(1,1),(1,-r)\}$. Thus we are once more faced with the problem of finding the number of lattice paths avoiding a boundary at zero. At a first sight this counting problem seems to be more difficult than it was the case of batch arrivals, because (see Figure (3.a)) path may cross the boundary but not at a lattice point and may have some overshoot. But this difficulty is easily resolved by reversing the paths, as is shown in the Figure (3.b). Two observations are immediate: the embedded process $Q_n$ exhibits a somewhat complicated behavior near the boundary whereas the reversed paths of the embedded random walk behave exactly like paths of a $M/M^r/1$ system. Thus the whole machinery we developed for $M/M^r/1$ works after a proper reparameterization again. Hence all that has to be done is to exchange $m \leftrightarrow k$ and $\lambda \leftrightarrow \mu$ in formula (11). We omit the details here, for more information see Böhm and Mohanty (1994) and Böhm (1993) where these models are dealt with in discrete time and the corresponding continuous time formulas are derived by a limiting procedure.
In many applications, in particular when modeling production systems, the assumption that the batches arriving or being served have a constant size is not very realistic. It is more plausible that batches have a size which may vary randomly. In this section we will deal with the model $M^B/M/1$. Customers are served one after the other with rate $\mu$ and arrive in batches having size $1 \leq q \leq r$, for some fixed maximum batch size $r$. It is assumed that batches of size $q$ arrive with rate $\lambda_q$, thus the probability that an arrival consists of a group of size $q$ equals $\lambda_q/(\mu + \sum_{i=1}^r \lambda_i)$.

In order to find $P(Q_t < k|Q_0 = m)$ we may still apply Champernowne’s construction (2), thus again we need to determine the probabilities $P_{\nu,k}^0(t)$, where the basic process $X_t$ can be represented by lattice paths having step set

$$S = \{(1,-1), (1,1), (1,2), \ldots, (1,r)\}.$$

In contrast to section 3, it no longer makes sense to separate the counting problem and the assignment of probabilities, since we would have to keep track of all possible batch sizes that arrived during the interval $(0,t)$. It is much more convenient to use generating functions.

Let us define a step generating function $P(y, z)$ by

$$P(y, z) = \frac{\mu z}{y} + \lambda_1 yz + \lambda_2 y^2 z + \ldots + \lambda_r y^r z$$

$$= z T(y) \quad \text{(say).} \quad (12)$$

Indeed, this is still possible by means of a multidimensional generalization of the ballot theorem, see Mohanty (1979, p 25, Exercise 10).
Then \([z^n y^k] P^n(y, z)\) yields the weight of all paths having \(n\) steps that lead from initial altitude 0 to terminal altitude \(k\). Let us denote these weight functions by \(w_k(n)\):

\[
w_k(n) = [y^k] \left( \frac{z}{y} + \lambda_1 y + \ldots + \lambda_r y^r \right)^n = [y^k] T(y)^n.
\]

It will turn out that these functions carry all the information we need to find the transient characteristics of this queueing system. They can be calculated easily by recursion. Since the residue of the derivative of \(T^n(y)\) is equal to zero, we have the identity

\[
\text{Res} \left( \frac{d}{dy} T^n(y) \right) = \text{Res} \left( \frac{-k}{y^{k+1}} T^n(y) + \frac{n}{y^k} T^{n-1}(y) \left[ \sum_{\ell=1}^{r} \ell \lambda_\ell y^{\ell-1} - \frac{\mu}{y^2} \right] \right) = 0
\]

By comparing coefficients we obtain the recurrence formula

\[
k w_k(n) = n \sum_{\ell=1}^{r} \ell \lambda_\ell w_{k-\ell}(n-1) - n \mu w_{k+1}(n-1), \quad -n \leq k \leq nr \tag{13}
\]

with boundary conditions

\[
w_{-n}(n) = \mu^n, \quad w_0(0) = 1 \quad \text{and} \quad w_k(n) = 0, \text{ if } k < -n \text{ or } k > nr
\]

Let us define for paths with \(n\) steps:

\[
A_{m,k}(n) = \text{weight of unrestricted paths from altitude } m \text{ to altitude } k \quad \text{and} \quad B_{m,k}(n) = \text{weight of paths from altitude } m \text{ to altitude } k \text{ which touch or cross the boundary at zero}
\]

Clearly, \(A_{m,k}(n) = w_{k-m}(n)\). To find \(B_{m,k}(n)\) we apply a beautiful theorem on the factorization of Laurent series due to Gessel (1980). Gessel has shown that the generating function of paths starting at initial altitude zero and terminating with a first passage at a lower boundary, such paths are also called minus-paths (see Goulden and Jackson (1983, pp. 314)), is given by

\[
\Phi(y, z) = \frac{1}{1 - \xi(z)/y}, \quad \xi = yzT(\xi).
\]

Expanding the geometric series and applying the Lagrange inversion theorem,
we obtain the expansion
\[
\Phi(y, z) = 1 + \sum_{i \geq 0} \frac{1}{y^i} \sum_{n \geq i} \frac{z^n}{n} [v^{n-i}] i(\mu + \lambda_1 v^2 + \ldots + \lambda_r v^{r+1})^n
\]
\[
= 1 + \sum_{i \geq 0} \sum_{j \geq 0} \frac{1}{y^i} \frac{z^{i+j}}{i+j} [v^j](\mu + \lambda_1 v^2 + \ldots + \lambda_r v^{r+1})^n
\]
\[
= 1 + \sum_{i \geq 0} \sum_{j \geq 0} \frac{1}{y^i} \frac{z^{i+j}}{i+j} [w_{-i}(i+j)].
\]

It follows that the weight of the set of paths starting at altitude \(m > 0\) and terminating after \(n\) steps with a first passage at zero is given by
\[
[y^{-m} z^n] P(y, z) = \frac{m}{n} w_{-m}(n). \tag{14}
\]

Since after the first passage the paths move without any further restriction from altitude zero to terminal altitude \(k\), we have
\[
B_{m,k}(n) = \sum_{\ell \geq 1} \frac{m}{\ell} w_{-m}(\ell) w_k(n-\ell). \tag{15}
\]

Passing from weights to probabilities is now easy, we have only to multiply 
\(A_{m,k}(n) - B_{m,k}(n)\) by \(e^{-(\mu + \sum_{i=1}^r \lambda_i) t n / n!}\) and sum over \(n\) to obtain finally
\[
P_{m,k}^0(t) = e^{-(\mu + \sum_{i=1}^r \lambda_i) t} \left[ \sum_{n \geq 0} \frac{w_{k-m}(n) t^n}{n!} - m \sum_{n \geq 0} \frac{t^n}{n!} \sum_{\ell=1}^n \frac{1}{\ell} w_{-m}(\ell) w_k(n-\ell) \right], \tag{16}
\]

and from this formula we obtain immediately the general transient distribution 
\(P(Q_t < k | Q_0 = m)\) by one more summation.

**Remark.** Formula (14) is indeed a special case of L. Takács’ *ballot theorem*. Takács has derived it under much more general conditions than we needed here and has devoted a complete book to this theorem (Takács (1967)).

### 5 Systems with batches and paths with general boundaries

In the last sections dealing with queueing systems involving batches the counting problem could always be reduced to one which was solvable essentially by Lagrange inversion. Unfortunately the situation becomes very different and much more complicated if we allow that customers arrive and depart in batches. To be concrete let us assume that customers arrive in batches of constant size \(a > 1\) with rate \(\lambda\) and depart in batches of constant size \(d > 1\).
with rate $\mu$. Interarrival times and service times are independent exponentially distributed, so that the queueing process $Q_t$ is still Markovian.

Champernowne’s construction (1) and (2) is still applicable but the associated randomized random walk is now more difficult. The process $X_t$ has step set $S = \{(1, -d), (1, a)\}$ with associated probabilities $\mu/(\lambda + \mu)$ and $\lambda/(\lambda + \mu)$.

In order to obtain a formula for
\[
P_{m,k}^0(t) = P(m + X_t = k, m + X_s > 0, 0 \leq s \leq t),
\]
we will use another lattice path representation of $X_n$, the process embedded in $X_t$, which is more convenient. The lattice paths representing $X_n$ now have step set $S^* = \{(1, 0), (0, 1)\}$ where a $(1, 0)$-step encodes an arrival of a batch and a $(0, 1)$-step encodes a batch departure. Such paths are sometimes called underdiagonal. Suppose that there have been $A_t = n$ arrivals and $D_t = \ell$ departures, then the paths representing the event
\[
\{m + X_t = k, m + X_s > 0, 0 \leq s \leq t, A_t = n, D_t = \ell\}
\]
lead from the origin to the point $(n, \ell)$ and do not touch or cross the line $dy = ax + m$. And clearly $m + an - \ell d = k$. Now define critical points $(i, b_i), i = 1, \ldots, n$, by the requirement that the altitude $y_i$ of the $i$-th horizontal step of the path satisfies $y_i \leq b_i, i = 1, \ldots, n$. In our example shown in Figure 4 we have $n = \ell = 7$ and
\[
b = (0, 2, 3, 5, 6, 7, 7).
\]
In general, for the boundary \( dy = ax + m \) these critical altitudes are given by

\[
b_i = \min \left( \ell, \left\lceil \frac{a(i-1) + m}{d} \right\rceil - 1 \right), \quad i = 1, \ldots n
\]  

(17)

We require the number of paths \( N_{m,k}(n, b) \) from \((0, 0)\) to the point \((n, \ell)\) such that \( k = m + an - d\ell \) and for which in each point \((i, y_i)\) we have \( y_i \leq b_i \). It is well known that this number is given by the determinant

\[
N_{m,k}(n, b) = \det_{1 \leq i, j \leq n} \left( \frac{b_i + 1}{j - i + 1} \right).
\]

(18)

By assigning probabilities to the \( n + \ell \) steps we get finally

\[
P(m + X_t = k, m + X_s > 0, 0 \leq s \leq t) = e^{-(\lambda + \mu) t} \sum_{m+an-d\ell=k}^{n \geq 0} N_{m,k}(n, b) \frac{\lambda^n \mu^\ell t^{n+\ell}}{(n+\ell)!}
\]

(19)

And again, as in (2), summation over \( m \) yields the general transient distribution of \( Q_t \). Unfortunately the computational work to evaluate (19) is considerable since the size the determinant increases with \( n \). The determinant formula (18) has been found for the first time by Kreweras (1965) in case of two boundaries, the one-boundary case has been solved by Narayana (1979), see also Mohanty (1979, chapter 2).

Since determinants are not very handy objects, it would be nice to have an alternative solution. In fact such an alternative exists, it is based on generating functions and interestingly in this technique Lagrange inversion enters again the stage, though through the back door. Two quite different approaches based on generating functions are due to Banderier and Flajolet (2002) and Sato (1989). Still the results are rather difficult to use. Banderier and Flajolet’s technique uses bivariate generating functions and the kernel method, whereas Sato’s method uses matrix valued generating functions. Since these methods are very interesting we comment briefly on a very special case of the \( Ma/Md/1 \) model, the case \( M^3/M^2/1 \), i.e. \( a = 3 \) and \( d = 2 \), the simplest nontrivial case of this model. We are interested in finding the density of the length \( T \) of a busy period initiated by a single customer. This queueing model is in fact a variant of Duchon’s Club problem (Duchon (2000)): There is a night club which upon opening has a single guest. Now from time to time new guests arrive, always in groups of three, and also from time to time guests leave the club in pairs for reasons we can only speculate about. The club closes when there are no more guests. Assume that the times between arrivals are

\[\text{It should me mentioned that there is yet another approach based on probabilistic counting due to Gessel (1986). This method extends a technique originally due to Dwass (1967).}\]
exponentially distributed with mean $1/\lambda$ and the times between successive departures have an exponential distribution with mean $1/\mu$.

We encode the paths of the embedded random walk by lattice paths with step set $\mathcal{S} = \{(1,-2),(1,3)\}$. Figure 5 shows typical paths. Using the method of

![Fig. 5. Busy period with even and odd number of arrivals](image)

Banderier and Flajolet (2002) it can be shown that the number of paths from $(0,1)$ to $(5n/2,1)$ not touching the boundary at zero, $n$ even, is given by

$$N_0(n) = \sum_{\ell=0}^{n} (-1)^{\ell} \frac{1}{1 + 5\ell} \left( \frac{1+5\ell}{2} \right) \frac{1}{1 + 5(n-\ell)} \left( \frac{1+5(n-\ell)}{2(n-\ell)} \right),$$

whereas for odd $n$ the number of paths not touching the boundary and going from $(0,1)$ to $((5n-1)/2,2)$ is given by the simple formula

$$N_1(n) = \frac{2}{1 + 5n} \left( \frac{1+5n}{2n} \right).$$

Now for $n$ even there are $3n/2$ departures and for $n$ odd we have $(3n-1)/2$ departures. Any of these paths is terminated by a final departure. Thus assigning probabilities we obtain for the density $f(t)$ of $T$ the following beautiful formula:

$$f(t) = e^{-(\lambda+\mu)t} \left[ \sum_{n \geq 0} N_0(n) \frac{\lambda^n \mu^{n-1} (5n/2)!}{(5n/2)!} + \sum_{n \geq 0} N_1(n) \frac{\lambda^n \mu^{n-1} (5n-1)/2)!}{(5n-1)/2)!} \right] \mu$$

$$= e^{-(\lambda+\mu)t} \sum_{\ell \geq 0} \lambda^{2\ell} \mu^{3\ell} 5^{5\ell} \left[ \frac{\mu N_0(2\ell)}{(5\ell)!} + \frac{\lambda \mu^2 t^2}{(2\ell + 1)(3\ell + 2)!} \right].$$

### 6 Queues and higher dimensional paths

Many queuing models lead quite naturally to combinatorial problems where lattice paths in higher dimensional spaces arise. This is particularly true when
we consider networks of queues. Here we have arrival streams at various nodes in the network as well as departure streams. The paths of the embedded random walks now become lattice paths with several dimensions which must satisfy certain constraints. We may also think of queueing systems with some inherent heterogeneity, e.g. different types of customers, different types of servers etc. In this section we want to give just an overview of some typical problems which have been successfully solved by combinatorial methods.

In Mohanty (1979, p. 25) we find the following formula which generalizes the classical ballot theorem: Let $E_m$ be a hyperplane of the form

$$E_m : x_0 = m + \sum_{i=1}^{r} b_i x_i,$$

where the $b_i$ are nonnegative integers. Then the number of simple lattice paths, i.e. paths having $n_i, 0 \leq i \leq r$ steps of unit length parallel to the coordinate axes which do not touch or cross $E_m$ except at the end, is given by

$$Q(m, \mathbf{n}) = \frac{m}{n} \binom{n}{\mathbf{n}} = \frac{m}{n} \frac{n!}{n_0! \cdots n_r!}, \quad n = \sum_{i=0}^{r} n_i, \quad \mathbf{n} = (n_0, \ldots, n_r). \quad (23)$$

Consider now the following queueing system: there is a single server providing service with exponentially distributed service times having mean $1/\mu$. Customers arrive from $r$ different sources in batches of size $b_i, i = 1, \ldots, r$, inter-arrival times of customers from source $i$ having an exponential distribution with mean $1/\lambda_i$.

A typical example is a communication channel which transmits data packets of unit size. Various sources, e.g., terminals, are connected to the channel and are sending data packets of different size and with different intensity to the channel for further transmission (Mitrani (1987, p. 75)). Due to congestion a queue will form in front of the transmitter and thus the time dependent distribution of the queue length becomes an issue. However, due to the heterogeneous nature of the arrival process, the joint distribution of the number of arrivals from the various sources and the number of completed services may be more informative. For instance, data packets may be charged with a certain rate $\pi_i$, depending on the source. If there is a cost of $\gamma$ for processing a data packet, then the revenue of the transmitter may be calculated. This model has been discussed in more detail by Böhm and Mohanty (1994).

Let $A_t(i)$ and $D_t$ denote the number of arrivals from source $i$ and the number of departures in $(0, t)$ and define $T_m = \inf\{t > 0 : m + \sum_{i=1}^{r} b_i A_t(i) - D_t = 0\}$. 

15
Let \( A(t) = (A_i(t), i = 1, \ldots, r) \). Then by (23) we obtain immediately

\[
P(A(t) = n, D_t = n_0, T_m > 0) = \left[ \binom{n}{n} - \frac{m}{n} \sum_{m+n = 0+\sum_{i=1}^r n_i} \binom{n}{c} \binom{a}{c} \right] \times \]

\[
\times e^{-\mu \sum \lambda_i} \prod \lambda_i^{n_i} \mu^{n_0} t^n, \tag{24}
\]

where \( c = \sum_{i=1}^r c_i, a = \sum_{i=1}^r a_i \). Similarly we get for the density of \( T_m \):

\[
f_m(t) = \frac{m}{t} e^{-\mu \sum \lambda_i} \sum_{n \geq 0}^{n_0 + \sum \lambda_i} \sum_{n_0 + \sum n_i = n} \frac{\prod \lambda_i^{n_i} \mu^{n_0} t^n}{n_0! n_i!}. \tag{25}
\]

Let us now return to the simple \( M/M/1 \) model. As we have seen, it is the associated random walk constrained to stay positive, which essentially determines the distribution of the queue length in this model. Let us have a closer look at this model to find out what is going on during a time period when the server is continuously busy. Again \( T_m \) denotes the length of this time period assuming that there are \( m \) customers waiting initially, and let \( A_t \) and \( D_t \) denote the arrival and the departure process both being Poisson with rates \( \lambda \) and \( \mu \), respectively. At the lowest level these two processes behave quite independently up to the time where the busy period ends, see Figure 6. Obviously, given the processes \( A_t \) and \( D_t \) start with an initial distance \( m \), the busy period terminates when the paths of \( A_t \) and \( D_t \) touch for the first time, i.e. the processes are coincident, they occupy the same state at the same time. In section 2 we represented the processes \( A_t \) and \( D_t \) simultaneously in a single lattice path by encoding a jump of \( A_t \) by a rise and a jump of \( D_t \) by a fall. And we solved our queueing problem by counting paths not touching or crossing a boundary at zero. Equivalently we could encode these processes by a simple lattice path having step set \( S = \{(1,0), (0,1)\} \), such that a jump of \( A_t \) corresponds to a \((1,0)\)-step and a jump of \( D_t \) to a \((0,1)\)-step. Then we have to count the

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig6.png}
\caption{The cases \( T_m > t \) and \( T_m = s < t \)}
\end{figure}
number of paths from \((m, 0)\) to a point \((a, b)\) staying strictly below the line \(y = x\), i.e., for each point \((x, y)\) on the path there holds \(x > y\).

This idea may be generalized considerably thereby enabling us to solve several nontrivial problems in the theory of networks of queues. As a first example let us consider a \(r\)-node series Jackson network (see Figure 7). This simplest form of a Jackson network consists of a single arrival stream at the first node which is Poisson with rate \(\lambda\). Customers finishing service at the first node are transferred to node 2 where they have to wait for service. The output of node 2 is input of node 3 and so on until a customer finishes service at the last node. The service times at each node are assumed to be independent exponentially distributed random variables with means \(1/\mu_i\). We shall also assume that each node has an infinite amount of space available to hold its customers, so there is no blocking. Blocking makes the analysis truly complicated, so we exclude this possibility here for simplicity. However, one special aspect of blocking will be discussed below.

To fix notation, let \(N_0(t)\) denote the arrival process at the first node and \(N_i(t), i = 1, \ldots, r\) the accumulated number of services in \((0, t)\) at node \(i\). During time periods where all servers are continuously busy, these processes are Poisson with rates \(\lambda_i, i = 1, \ldots, r\). We assume that at time zero there are \(m_i > 0\) customers already waiting at node \(i\). Let \(Q_i(t)\) be the number of customers waiting at node \(i\) at time \(t\) and define the stopping times

\[
T_i = \inf\{t > 0 : Q_i(t) = 0, Q_i(0) = m_i\} = \inf\{t > 0 : m_i + N_{i-1}(t) - N_i(t) = 0\} \quad i = 1, \ldots, r,
\]

and let \(T = \min_{1 \leq i \leq r} T_i\), i.e., \(T\) denotes the length of the interval all servers are continuously busy. Furthermore define

\[
P_0(\mathbf{m}, k; t) = P(Q(t) = k, T > t|Q(0) = \mathbf{m}) \quad (26)
\]

with \(Q(t) = (Q_1(t), \ldots, Q_r(t)), \quad \mathbf{m} = (m_1, \ldots, m_r), \quad \mathbf{k} = (k_1, \ldots, k_r)\).

For \(r = 2\) the situation is illustrated in Figure 8. The combinatorial problem can now be stated as follows: what is number of possible arrangements of jumps of the processes \(N_i(t), 0 \leq i \leq r\), such that no coincidence occurs, thus their sample paths do not touch or cross?

To solve this problem, we encode the sample paths in a single lattice path in \((r + 1)\)-dimensional space with step set \(\mathcal{S} = \{e_0, \ldots, e_r\}, e_i\) being the \(i\)-th unit
vector. Define the points \( a \) and \( b \) in \( \mathbb{R}^{r+1} \) by

\[
a_i = \sum_{\ell=i+1}^{r} m_{\ell}, \quad b_i = n + \sum_{\ell=1}^{r} m_{\ell} - \sum_{i=1}^{i} k_i, \quad i = 0, \ldots, r. \tag{27}
\]

We are looking for the number \( N_n(a, b) \) of paths from \( a \) to \( b \) such that for each point \( x \) on the path there holds

\[
x_0 > x_1 > \ldots > x_r \tag{28}
\]

Actually this requires again a version of the ballot theorem which in its weak form (replace \( > \) by \( \geq \) in the above sequence of inequalities) has been proved by Barton and Mallows (1965) and by Kreweras (1965).

In our case the required number turns out to be a determinant:

\[
N_n(a, b) = \left[ \sum_{i=0}^{r} (b_i - a_i) \right]! \det_{0\leq i,j\leq r} \left( \frac{1}{(b_i - a_j)!} \right) \tag{29}
\]

Assigning probabilities to the steps \( e_i \) in the usual way we obtain the following nice formula

\[
P_0(m, k; t) = \sum_{n\geq 0} \det_{0\leq i,j\leq r} \left( \frac{1}{(b_i - a_j)!} \right) e^{-t \sum_{i=0}^{r} \lambda_i \prod_{i=0}^{r} (\lambda_i t)^{b_i - a_i}}. \tag{30}
\]

This formula has been derived by Böhm, Jain and Mohanty (1993) using Kreweras’s approach and by Böhm and Mohanty (1997) using the Karlin-McGregor Theorem. See also Karlin and McGregor (1959).

The beautiful formula (29) can be proved in various ways, the most elegant way being the reflection principle, actually a multidimensional version thereof. The lattice paths going from \( a \) to \( b \) and satisfying (28) are constrained by the hyperplanes \( H_i : x_{i-1} - x_i = 0, i = 1, \ldots, r \). When a path touches \( H_i \) and the
initial portion of the path prior to the last contact with \( H_i \) is reflected, this results in a transposition of the \((i-1)\)-th and the \(i\)-th coordinate in \(a\). Taking care of all possible transformations of \(a\) that arise from iterated reflections on the hyperplanes \(H_i\) one obtains

\[
N_n(a, b) = \sum_{\sigma \in \mathcal{S}_{r+1}} (-1)^\sigma |\sigma(a) \rightarrow b|,
\]

which is just (29). Here \(\mathcal{S}_{r+1}\) is the group of permutations of order \(r+1\), \(\sigma(a)\) is a permutation of the coordinates in \(a\) and \(|a \rightarrow b|\) denotes the number of unconstrained paths from \(a\) to \(b\).

The foregoing discussion raises the question: how far can we go with the reflection principle? This question is not only interesting from a combinatorial point of view but also with regard to applications. A complete answer has been given by Gessel and Zeilberger in their celebrated paper on random walks in a Weyl chamber (1992).

Let us give another example of this powerful method which solves one more really difficult queueing problem in an almost elementary way. We consider a cyclic series Jackson network with \(r+1\) nodes with a fixed population of customers of size \(L\). Customers finishing service at the last node are sent back to node 1 and the process continues \textit{ad infinitum}. This cyclic tandem system which is the basic network model in the so called window flow control mechanism of communication systems has been analyzed by Baccelli, Massey and Wright (1994). It can be shown that this model is equivalent to the open tandem network we discussed earlier with global blocking at \(L\). Global blocking means, the operation of the queueing system stops whenever the number of customers in the system reaches a critical level \(L\).

As in the case of the open tandem queue discussed above we are interested in the joint distribution of the number of customers waiting at the various nodes during time intervals where all servers are continuously busy.

Define now points \(a = (a_0, \ldots, a_r)\) and \(b = (b_0, \ldots, b_r)\) by

\[
a_i = L - \sum_{\ell=1}^{i} m_\ell, \quad b_i = n + L - \sum_{\ell=1}^{i} k_\ell, \quad i = 0, \ldots, r
\]
This time we have to count paths from \( a \) to \( b \), such that for each point \( x \) on a path there holds

\[
x_0 > x_1 > \ldots > x_r > x_0 - L
\]  

(31)

Again the reflection principle works, the hyperplanes which restrict the paths are of the form \( x_{i-1} - x_i = k, k \in \mathbb{Z} \). The group generated by these reflections is affine to the symmetric group \( S_{r+1} \). The affine permutations act by permuting the coordinates of \( x \) and adding a vector \( \mathbf{v} = (v_0, \ldots, v_r) \) with \( v_0 + \ldots + v_r = 0 \). By the Gessel-Zeilberger Theorem the number of paths satisfying (31) is given by

\[
M_n(a, b) = \left[ \sum_{i=0}^{r} (b_i - a_i)! \right]! \sum_{v_0 + \ldots + v_r = 0} \det_{0 \leq i, j \leq r} \left( \frac{1}{(b_i - a_j + v_i L)!} \right). 
\]  

(32)

A variant of this formula has been obtained by Filaseta (1985).

It immediately follows upon assigning probabilities in the usual way that the joint distribution of the number of customers waiting in front of the various servers, all of them being continuously busy in \((0, t)\), is given by:

\[
P_0^L(m, k; t) = \sum_{n \geq 0} \sum_{v_0 + \ldots + v_r = 0} \det_{0 \leq i, j \leq r} \left( \frac{1}{(b_i - a_j + v_i L)!} \right) e^{-t \sum_{i=0}^{r} \lambda_i} \prod_{i=0}^{r}(\lambda_i t)^{b_i - a_i},
\]  

(33)

a result which otherwise could be obtained only by rather intricate analytical methods (Baccelli, Massey and Wright (1994)).
In this last section we consider two-dimensional paths that are restricted to a quarter plane, a very interesting area of research with many new results. Many of these results have a strong relation to the theory of queues either from the application’s side or that research has been stimulated by various queueing problems.

Let us give a few examples first.

Consider the open tandem queue we considered earlier with two nodes in series. The corresponding paths have step set \( S = \{(1,0),(-1,1),(0,-1)\} \) and during time periods where both servers are continuously busy the paths are bound to the quarter plane \( \mathbb{Z}_+ \times \mathbb{Z}_+ \). Arrivals from outside the system at the first node are mapped to \((1,0)\)-steps, when a job finishes service at node 1 and joins the queue at node 2 this corresponds to a \((-1,1)\)-step, a departure from node 2 gives rise to a \((0,-1)\)-step, see Figure 11 for an illustration.

The joint distribution \( P_0(m_1, m_2; k_1, k_2; t) \) has been given already in (30) where it has been derived by the reflection principle. However, the general distribution is far more difficult because the boundaries become now reflecting barriers and they modify the transition probabilities of the paths in a particular way. No elegant combinatorial proof exists, although there is a result by Baccelli and Massey (1987) which has been derived by operator analytic methods.

Another queueing model with a fascinating combinatorial background has been discussed by Flatto and Hahn (1984). Customers are arriving from an external source and upon their arrival they simultaneously place two demands which are handled by two service facilities. A typical application is the maintenance of a large database: here temporal deviations may occur between the number of records in the database and the number of items present in the stock. These inconsistencies are due to the fact that updating of the system
occurs at times different from those times when records are added or deleted from the stock. If deviations are due to removals of records only then the system may be modeled as two parallel queues with the same arrival process. A generalization of this model which included two more external arrival streams has been analyzed by Wright (1992). The sample paths of the Flatto-Hahn model may be mapped to lattice paths in the non-negative quarter plane with step set \( S = \{(-1,0),(0,-1),(1,1)\} \) where the \((1,1)\)-step represents an incoming job and the other steps represent services by server 1 or 2, respectively, see Figure 12 (left). The analysis of this queueing model is pretty difficult not only in finite time but also regarding its limiting properties. Still it is still possible to decompose this two-dimensional problem into a three dimensional one (see Figure 12, right), as we have done in the tandem case: there is a Poisson process \( X_0(t) \) representing arrivals and two service processes \( X_1(t) \) and \( X_2(t) \). Transitions during time intervals where both servers are continuously busy correspond now to transitions of three independent Poisson processes \( X_0(t), X_1(t) \) and \( X_2(t) \) starting at altitudes determined by the initial conditions, such that for all \( 0 < s < t \) there holds \( X_0(s) > \max\{X_1(s),X_2(s)\} \). This is equivalent to a three candidate ballot problem which has been discussed by Kreweras (1965) and Niederhausen (1978). A proof of Kreweras’s result by means of a probabilistic counting method has been given by Gessel (1986). Recently Bousquet-Méloù (2005) has presented a thorough analysis of this problem based on the kernel method. Among other things she has given a constructive proof of the formula for the number of paths having \( 3n \) steps that start at \((1,1)\) and return to \((1,1)\) without touching the \(x\)-axis or the \(y\)-axis. This number is given by the remarkable formula

\[
a_{3n} = \frac{4^n}{(n+1)(2n+1)} \binom{3n}{n}
\]

Her technique of analysis is the kernel method. Actually this method has been well established in queueing theory for a rather long time (see e.g. Bailey (1954)). Before we demonstrate the kernel method in a little bit more detail by means of another example let us state one more result obtained by Bousquet-Méloù. The number of paths starting at the origin and terminating after \( 3n+2i \)
steps on the $x$-axis is given by

$$a_{i,0}(3n + 2i) = \frac{4^n(2i + 1)}{(n + i + 1)(2n + 2i + 1)} \binom{2i}{i} \binom{3n + 2i}{n}$$

(35)

This formula gives us immediately an expression for the density of a busy period of server 2, given $Q_1(0) = Q_2(0) = 1$. By shifting the origin one unit to the left and one unit down we get

$$P(t < T_2 \leq t + dt, T_1 > t, Q_1(t) = i) =
\begin{align*}
  &\sum_{n \geq 0} a_{i-1,0}(3n + 2i - 2)e^{-(\lambda + \mu_1 + \mu_2)t} \frac{\lambda^{n+i-1}\mu_1^n\mu_2^{n+i}t^{3n+2i-2}}{(3n + 2i - 2)!} dt \\
  &= \frac{e^{-(\lambda + \mu_1 + \mu_2)t}}{t} \sum_{n \geq 0} \frac{\lambda^{n+i-1}\mu_1^n\mu_2^{n+i}t^{3n+2i-1}}{(n+i)(3n + 2i - 1)} \frac{2i - 1}{n!(2n + 2i - 1)!} \binom{2i - 2}{i - 1} dt,
\end{align*}$$

$T_1$ and $T_2$ denoting the lengths of the busy periods at server 1 and server 2, respectively. By a symmetry argument we get in a similar fashion the density of $T_1$.

It seems to be rather difficult to obtain formulas like (34) or (35) by purely combinatorial arguments. As far as results are available they have mostly been found by analytic methods, see e.g. the monograph by Fayolle, Iasnogorodski and Malyshev (1999) for an excellent treatment of random walks in a quarter plane.

As a final example let us consider a queueing system with preemptive priorities. A first analysis of this model in a Markovian environment has been carried out by Heathcote (1959). There are two types of customers: Type-I customers arrive with rate $\lambda_1$ and are served with rate $\mu$, these have priority over type-II customers. The latter arrive with rate $\lambda_2$ and are served with the same rate $\mu$ as type-I customers. However, when a type-II customer is in service at time $t$ and a type-I customer arrives, then the service of the type-II customer is interrupted, service of the customer with priority starts immediately and all low priority customers have to wait until there are no more type-I customers in the system.

The embedded random walk $X_n$ of this model is a bivariate process with state space $S = \{(i,j), i, j \in \mathbb{Z}_0^+\}$, the lines $y = 0$ and $x = 0$ serve as reflecting boundaries. To demonstrate the kernel method let us study this queueing system during a busy period of type-II customers. For ease of exposition we assume the initially the system is in state $(0,1)$, i.e., there are no customers with priority in the system and only one customer of type-II having low priority. The random walk $X_n$ has four possible types of jumps (see Figure 13).
These jumps have probabilities

\[
P(X_n = (i + 1, j) | X_{n-1} = (i, j)) = \alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \mu}, \quad i, j > 0
\]

\[
P(X_n = (i - 1, j) | X_{n-1} = (i, j)) = \gamma = \frac{\mu}{\lambda_1 + \lambda_2 + \mu}
\]

\[
P(X_n = (i, j + 1) | X_{n-1} = (i, j)) = \beta = \frac{\lambda_2}{\lambda_1 + \lambda_2 + \mu}
\]

\[
P(X_n = (0, j + 1) | X_{n-1} = (0, j)) = \beta \quad j > 0
\]

\[
P(X_n = (0, j - 1) | X_{n-1} = (0, j)) = \gamma
\]

\[
P(X_n = (1, j) | X_{n-1} = (0, j)) = \alpha
\]

\[
P(X_0 = (0, 1)) = 1
\]

Now define generating functions with the proviso \(P(X_0 = (0,1)) = 1\), i.e.,

we assume that the system is initially almost empty, there is only one type-II customer, no priority customers are waiting for service:

\[
P_{a,b}(s) = P_{a,b} = \sum_{n \geq 0} s^n P(X_n = (a,b)), \quad U_a(y; s) = U_a(y) = \sum_{n \geq 1} y^n P_{a,n}(s)
\]

and

\[
W(x, y; s) = W(x, y) = \sum_{m \geq 0} U_m(y)x^m = \sum_{m \geq 0} \sum_{n \geq 1} x^m y^n P_{m,n}
\]

Setting up the difference equations for the transition probabilities we obtain the kernel equation

\[
(x - s(\alpha x^2 + \beta xy + \gamma))W(x, y) = xy + s\gamma U_0(y) \left( \frac{x}{y} - 1 \right) - xs\gamma P_{01}, \quad (36)
\]

with kernel \(x - s(\alpha x^2 + \beta xy + \gamma)\). Equation (36) contains two unknown functions, \(U_0(y)\) and \(P_{01}\). To determine these unknowns we calculate the zeroes of
the kernel. These zeroes are found to be:

\[
x_1(y) = x_1 = \frac{1}{2s\alpha} \left[ 1 - ys\beta - \sqrt{(1 - ys\beta)^2 - 4\alpha\gamma s^2} \right]
\]
\[
x_2(y) = x_2 = \frac{1}{2s\alpha} \left[ 1 - ys\beta + \sqrt{(1 - ys\beta)^2 - 4\alpha\gamma s^2} \right]
\]

For \(|y| < 1\) and \(|s| < 1\) we have \(|x_1| < 1\) and \(|x_2| > 1\). However, since \(W(x, y)\) is an analytic function of \(x\) and \(y\) in a neighborhood of \(x = 0, y = 0\), the pole of \(W(x, y)\) at \(x_1\) must be a removable singularity. Thus we have

\[
0 = x_1 y + s\gamma U_0 \left( \frac{x_1}{y} - 1 \right) - x_1 s\gamma P_{01}
\]

which yields

\[
U_0(y) = x_1 y \frac{s\gamma P_{01} - y}{s\gamma x_1 - y}, \quad (37)
\]

with one unknown function, viz. \(P_{01}\), remaining. The denominator of (37) has two roots:

\[
y_1(s) = \frac{1 - \sqrt{1 - 4\gamma s^2(\alpha + \beta)}}{2s(\alpha + \beta)} \quad y_2(s) = \frac{1 + \sqrt{1 - 4\gamma s^2(\alpha + \beta)}}{2s(\alpha + \beta)} \quad (38)
\]

The first root \(y_1\) is located inside the region \(|y| < 1\), it must cancel with the numerator yielding

\[
P_{01}(s) = \frac{1 - \sqrt{1 - 4\gamma s^2(\alpha + \beta)}}{2\gamma s^2(\alpha + \beta)} \quad (39)
\]

Let \(f_{2n+1}, n \geq 0\) denote the probability that \(X_n\) reaches the point \((0, 0)\) at step \(2n+1\) for the first time, given \(X_n\) started in \((0, 1)\). Then clearly the generating function of \(f_{2n+1}\) is given by

\[
\sum_{n \geq 0} f_{2n+1} s^{2n+1} = \gamma s P_{01} = \frac{1 - \sqrt{1 - 4\gamma s^2(\alpha + \beta)}}{2s(\alpha + \beta)} \quad (40)
\]

Expanding this generating function we obtain the nice formula

\[
f_{2n+1} = \frac{1}{n+1} \binom{2n}{n} (\alpha + \beta)^n \gamma^{n+1} \quad (41)
\]

This formula has a simple combinatorial interpretation: at any time we can have either an arrival (type-I or type II) with probability \(\alpha + \beta\) or a service
with probability $\gamma$. The Catalan number gives us the number of possible arrangements of these arrivals and services so that the system does not become completely empty.

From this formula we obtain immediately the density $f(t)$ of the busy period of low priority customers by putting

$$\alpha = \frac{\lambda_1}{\lambda_1 + \lambda_2 + \mu}, \quad \beta = \frac{\lambda_2}{\lambda_1 + \lambda_2 + \mu}, \quad \gamma = \frac{\mu}{\lambda_1 + \lambda_2 + \mu}$$

Fixing the last down-step we have

$$f(t) = \sum_{n \geq 0} \frac{1}{n+1} \binom{2n}{n} \frac{\mu^{n+1}(\lambda_1 + \lambda_2)^n}{(\lambda_1 + \lambda_2 + \mu)^{2n}} \cdot \frac{t^{2n}(\lambda_1 + \lambda_2 + \mu)^{2n}}{(2n)!} e^{-(\lambda_1+\lambda_2+\mu)t}$$

$$= \frac{1}{t} e^{-(\lambda_1+\lambda_2+\mu)t} \sum_{n \geq 0} \frac{\mu^{n+1}(\lambda_1 + \lambda_2)^n t^{2n+1}}{n!(n+1)!}$$

$$= \frac{1}{t} e^{-(\lambda_1+\lambda_2+\mu)t} \sqrt{\frac{\mu}{\lambda_1 + \lambda_2}} I_1 \left( 2t \sqrt{\mu(\lambda_1 + \lambda_2)} \right), \quad (42)$$

where $I_1$ denotes a modified Bessel function. This formula can be found (using a different method of derivation) as early as 1959 in Heathcote (1959).

The author is indebted to the anonymous referee for his helpful comments.

8 References


