Explicit solutions for queueing problems

Citation for published version (APA):

Document status and date:
Published: 01/01/1997

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Please check the document version of this publication:
• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher’s website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal.

If the publication is distributed under the terms of Article 25fa of the Dutch Copyright Act, indicated by the “Taverne” license above, please follow below link for the End User Agreement:
www.tue.nl/taverne

Take down policy
If you believe that this document breaches copyright please contact us at:
openaccess@tue.nl
providing details and we will investigate your claim.
Explicit solutions for queueing problems

Ivo Adan
Eindhoven University of Technology
Department of Mathematics and Computer Science
Box 513, 5600 MB - Eindhoven
the Netherlands

April 14, 1997

1 Introduction

Many queueing problems can be modeled as random walks on a multi-dimensional grid. A time-dependent analysis of such models appears to be possible only in rare cases, and even then the solutions are quite complicated. Therefore, in this paper attention is focussed upon the equilibrium behavior of these models, rather than upon their time-dependent behavior.

The equilibrium distribution of a random walk on a grid is the solution of a set of equilibrium equations. These equations can be viewed as difference equations. In the theory of differential equations, the continuous analogue of difference equations, a well-known solution approach is separation of variables (see e.g. Garabedian [11]). This method attempts to solve differential equations by constructing sums of product-form solutions. It seems natural to investigate whether it is also possible to solve equilibrium equations by sums of product-form solutions. And if so, under which conditions are such solutions feasible, and which techniques can be used to find such solutions? In the attempts to find sums of product forms as solution, three main directions may be distinguished, namely:

1. Exactly one product form as solution;

2. A finite sum of product forms as solution;

3. A countably infinite sum of product forms as solution.

The first direction is the oldest one. It has also been studied most systematically, see e.g. Jackson [14], Gordon and Newell [12], Baskett et al. [5], Kelly [15], Lavenberg and Reiser [17] and Van Dijk [9]. In [5] a practical characterization is given of queueing networks with a product-form solution. The conditions under which the solution is a single product form are rather severe, but, most strikingly, they do not depend on the dimension of the state
space. This feature makes this product-form approach very important, since it is virtually the only more general approach for really complex queueing systems.

The other directions are newer and may be viewed as a generalization of the first one. The results in these directions are less systematic than the first one. This paper aims at reviewing some of the particular techniques required for the construction of sums of product form solutions.

An important application in the second direction is the multi-server queue with Erlang (or Cox) arrivals and services. Queueing problems of this type can be described as a random walk on a multi-dimensional grid which is unbounded in only one direction. In section 2 we will show that the queue length distribution can be written as a sum of products. To find this solution we use a direct approach which is based on separation of variables. The third direction is based on the compensation approach. This approach has been developed for one rather general class of multi-dimensional random walks (unbounded in each direction) and for several related special cases. An important example in this direction is the shortest queue problem. In section 3 we will sketch the analysis of this problem to demonstrate the basic ideas of the compensation approach.

2 Finite sum of products: System with Erlang servers

In this section we study a system with \( c \) parallel identical servers and a common queue. The service times are Erlang-\( r \) distributed with mean \( r/\mu \). This means that a service has to go through up to \( r \) exponential stages, each with mean \( 1/\mu \). The service discipline is first-come first-served. Jobs arrive according to a Poisson stream with rate \( \lambda \). This system can be modeled as a continuous-time Markov process with states \( \mathbf{n} = (n_0, n_1, \ldots, n_c) \), where \( n_0 \) is the number of waiting jobs and \( n_i \) is the number of remaining service stages for server \( i \), \( i = 1, \ldots, c \). So \( n_0 \) ranges from 0 to \( \infty \) and \( n_i, i = 1, \ldots, c \), from 0 to \( r \), where \( n_i = 0 \) is only possible if \( n_0 = 0 \). Note that completion of a service stage at server \( i \) leads to a departure if \( n_i = 1 \). The flow diagram for \( c = 1 \) is depicted in figure 1.

![Flow diagram for the model with \( c = 1 \)](image)

Figure 1: Flow diagram for the model with \( c = 1 \)

Our aim is to determine the equilibrium probabilities \( p(\mathbf{n}) \). Once these probabilities are known we can compute performance characteristics such as, for example, the mean queue length and the mean waiting time. We will show that \( p(\mathbf{n}) \) can be expressed as a finite sum
of products of powers. To do so we first introduce some notations and formulate the equilibrium equations for $p(n)$ that will be relevant to the analysis. Let $e_i = (0, \ldots, 0, 1, 0, \ldots, 0)$ have $c+1$ components, with the one at the same place as $n_i$ in $n$ and let $\delta(n)$ be 1 if $n = r$ and 0 otherwise. By equating the rate out of and the rate into state $n$ we obtain

$$p(n)(\lambda + c\mu) = p(n - e_0)\lambda + \sum_{i=1}^c p(n + e_i)\mu(1 - \delta(n_i)) + \sum_{i=1}^c p(n + e_0 - (r - 1)e_i)\mu\delta(n_i),$$

which is valid for all states $n$ with $n_0 \geq 1$. The equations (1) form the inner conditions, the equations in states with $n_0 = 0$ form the boundary conditions. The precise form of the boundary conditions is not relevant to the analysis, and therefore it is omitted.

The approach to solve the equilibrium equations will be based on separation of variables, an elementary approach for the solution of partial differential equations. Below we demonstrate this approach for a simple problem of conduction of heat (cf. Carslaw and Jaeger [8]). Then we will show how the same approach can be used to determine the probabilities $p(n)$.

### An analogue: Conduction of heat in a thin rod

Consider the following problem (see also figure 2):

$$u_{xx} - u_t = 0, \quad 0 < x < 1, t > 0, \quad (2)$$

$$u(0, t) = u_x(1, t) = 0, \quad t \geq 0, \quad (3)$$

$$u(x, 0) = v(x), \quad 0 \leq x \leq 1. \quad (4)$$

The function $u(x, t)$ can be interpreted as the temperature in a thin rod along the interval $0 \leq x \leq 1$. The end $x = 0$ is maintained at zero temperature, while the end $x = 1$ is isolated (no flow of heat). The initial temperature at $t = 0$ is given by $v(x)$. This problem may be solved by first constructing solutions of the form

$$u(x, t) = X(x)T(t), \quad (5)$$

satisfying (2) and boundary conditions (3). By linearity of (2)–(3) any linear combination of solutions in this set satisfies these equations. The next step is to construct a linear combination which also satisfies initial condition (4). Substitution of (5) in (2)–(3) yields

$$\frac{X''(x)}{X(x)} = \frac{T'(t)}{T(t)} = \text{constant} = \lambda, \quad X(0) = X'(1) = 0,$$

where $\lambda$ is the separation constant. Only for the values $\lambda = \lambda_j = -(j + 1/2)^2\pi^2$ with $j = 0, 1, \ldots$ these equations have a nontrivial solution, namely

$$u_j(x, t) = \sin(x\sqrt{-\lambda_j})e^{\lambda_j t}.$$  

Then the solution of (2)–(4) can be written as a linear combination of functions in this set,

$$u(x, t) = \sum_{j=0}^\infty c_j u_j(x, t),$$

3
where the coefficients $c_j$ follow from initial condition (4).

We now return to the queueing problem and try to use the same approach as above. In the first step we construct solutions of the form

$$ p(n) = F_0(n_0)F_1(n_1)\ldots F_c(n_c) $$

(6)

satisfying the equations (1). In the second step we use these functions to construct a linear combination also satisfying the conditions for $n_0 = 0$. Insertion of (6) in (1) yields

$$ \lambda + c\mu = \frac{F_0(n_0 - 1)}{F_0(n_0)} \lambda + \sum_{i=1}^{c} \frac{F_i(n_i + 1)}{F_i(n_i)} \mu (1 - \delta(n_i)) + \sum_{i=1}^{c} \frac{F_0(n_0 + 1)}{F_0(n_0)} \frac{F_i(1)}{F_i(r)} \mu \delta(n_i). $$

Considering this equation for $n_i < r$ (so $\delta(n_i) = 0$) leads to the conclusion that

$$ \frac{F_0(n_0 - 1)}{F_0(n_0)} = \text{constant} = \frac{1}{\alpha_0}, \quad n_0 \geq 1, $$

$$ \frac{F_i(n_i + 1)}{F_i(n_i)} = \text{constant} = \alpha_i, \quad n_i = 1, \ldots, r - 1, \ i = 1, \ldots, c. $$

So the functions $F_i(n_i)$ are powers of $\alpha_i$. Hence the solutions (6) are of the form

$$ p(n) = \alpha_0^{n_0} \ldots \alpha_c^{n_c}, $$

where the (separation) constants $\alpha_i$ have to satisfy

$$ \lambda + c\mu = \frac{\lambda}{\alpha_0} + \sum_{i=1}^{c} \alpha_i \mu + \sum_{i=1}^{c} \frac{\alpha_0}{\alpha_i} - 1 \alpha_i \mu \delta(n_i). $$

(7)

To satisfy this relation for all $n_i$ the coefficients of the functions $\delta(n_i)$ must be zero, so we obtain that $\alpha_0 = \alpha_1^r = \ldots = \alpha_c^r$, or equivalently,

$$ \alpha_0 = \alpha_1^r, \quad \alpha_2 = x_2 \alpha_1, \ldots, \alpha_c = x_c \alpha_1, $$

(8)

with $x_2^r = \ldots = x_c^r = 1$. Substitution of (8) in (7) yields the following equation for $\alpha_1$:

$$ \lambda + c\mu = \frac{\lambda}{\alpha_1^r} + \alpha_1 \mu + \sum_{i=2}^{c} x_i \alpha_1 \mu. $$

(9)
The condition that the sum of all probabilities \( p(n) \) is equal to 1 implies that only products with \( |\alpha_0| < 1 \), or equivalently, \( |\alpha_1| < 1 \) are useful. For each feasible choice of \( x_i \) it can be shown that equation (9) has exactly \( r \) roots \( \alpha_1 \) with \( |\alpha_1| < 1 \) provided the utilization condition \( \lambda r / \mu < c \) holds. So we find \( r \) products satisfying (1). We label these products \( \alpha_0^{n_0} \cdots \alpha_c^{n_c}, j = 1, \ldots, r \). This concludes the first step. In the second step we express \( p(n) \) as a linear combination of the products in this set. The number of unknown coefficients in this linear combination is sufficient to also satisfy the equilibrium equations for states \( n \) with \( n_0 = 0 \). Below we summarize our findings. Rigorous proofs may be found in [4].

**Theorem 2.1** Provided \( \lambda r / \mu < c \), there exist coefficients \( c_j \) such that

\[
p(n) = \sum_{j=1}^{r} c_j \alpha_0^{n_0} \cdots \alpha_c^{n_c}
\]

for all \( n \) with \( n_i \geq 1, i = 1, \ldots, c \).

**Remark 2.2** (Extensions) The approach demonstrated in this section also works in case of Erlang distributed interarrival times, see [4]. A similar technique, based on separation of variables, has been applied by Bertsimas [7, 6] to solve the \( E_k/C_2/s \) and \( C_k/C_r/s \) system, respectively.

**Remark 2.3** (Matrix-geometric approach) The equilibrium distribution can also be represented in a matrix-geometric form (see e.g. chapter 3 in Neuts [19]). There is a close relation between Theorem 2.1 and this representation. Namely, the factors \( \alpha_0^{n_0} \cdots \alpha_c^{n_c} \) are the eigenvalues of the rate matrix and the products \( \alpha_0^{n_0} \cdots \alpha_c^{n_c} \) are the associated eigenvectors.

### 3 Infinite sum of products: Shortest queue problem

The system we consider in this section consists of two parallel and identical servers, each with its own queue (see figure 3(a)). The service times are exponentially distributed with mean 1. Jobs arrive according to a Poisson stream with rate \( 2 \rho \) where \( 0 < \rho < 1 \). On arrival a job joins the shortest queue, and if queues have equal length, joins either queue with probability \( 1/2 \). The state of the system can be described by the pair \((m, n)\) where \( m \) is the length of the shortest queue and \( n \) is the difference between the longest and shortest queue. The flow diagram is shown in figure 3(b). Let \( p(m, n) \) be the equilibrium probability for state \((m, n)\). The object in this section is the determination of \( p(m, n) \).

The equilibrium equations are given below. In these equations we have eliminated the probabilities \( p(m, 0) \) from (11) and (13) by substituting (14)–(15). This is done to simplify the presentation. The analysis can now be restricted to the probabilities \( p(m, n) \) with \( n > 0 \).
satisfying (10)–(13). The equations (14)–(15) may be treated as definition for \( p(m, 0) \).

\[
\begin{align*}
    p(m, n)(2\rho + 1) & = p(m - 1, n + 1)2\rho + p(m, n + 1) + p(m + 1, n - 1), \\
    m > 0, n > 1, \\
    (10) \\
    p(m, 1)(2\rho + 1) & = p(m - 1, 2)2\rho + p(m, 2) \\
    & + (p(m, 1)2\rho + p(m + 1, 1)) \frac{1}{\rho + 1} \\
    & + (p(m - 1, 1)2\rho + p(m, 1)) \frac{\rho}{\rho + 1}, \\
    m > 0, \\
    (11) \\
    p(0, n)(2\rho + 1) & = p(0, n + 1) + p(1, n - 1), \\
    n > 1, \\
    (12) \\
    p(0, 1)(2\rho + 1) & = p(0, 2) + (p(0, 1)2\rho + p(1, 1)) \frac{1}{\rho + 1} + p(0, 1), \\
    (13) \\
    p(m, 0)(2\rho + 1) & = p(m - 1, 1)2\rho + p(m, 1), \\
    m > 0, \\
    (14) \\
    p(0, 0)2\rho & = p(0, 1). \\
    (15)
\end{align*}
\]

The usual approaches to solve the equilibrium equations are based on generating functions (see e.g. [16, 10]). In this section we present an approach which directly tries to solve the equations. The idea is similar to the one in the previous section. We first try to find a set of products \( \alpha^m \beta^n \) satisfying the inner conditions (10). Then we use the products in this set to construct a linear combination which also satisfies the boundary conditions (11)–(13). The first part is easy. Substituting \( \alpha^m \beta^n \) into (10) and then dividing by common factors yield a quadratic equation for \( \alpha \) and \( \beta \) (see figure 4).

**Lemma 3.1** The product \( \alpha^m \beta^n \) satisfies (10) if

\[
\alpha \beta^2(\rho + 1) = \beta^22\rho + \alpha \beta^2 + \alpha^2. 
\]

The problem we are now facing is different from the one in the previous section. There we found a finite set of products satisfying the inner conditions and these products were all used to construct the solution. Now we have a continuum of products satisfying (10).
How do we select the appropriate products from this set? The selection is based on a compensation idea (which explains the name of the approach). This idea has an interesting analogue in electrostatics, where it is known as the method of images.

**An analogue: Potential problem of conducting spheres**

Consider two non-intersecting conducting spheres, whose centers are $A$ and $B$, their radii $a$ and $b$ and their potentials $\Phi_a$ and 0, respectively. Suppose that their distance of centers is $c$ (see figure 5). Below we show how the potential $\Phi$ outside the spheres can be found by the method of images (see e.g. Maxwell [18]).

If the spheres did not influence each other ($c = \infty$), then the potential $\Phi$ is that of point charge $\alpha_0 = a\Phi_a$ located at $A$. However, since $c$ is finite, the potential does not vanish on sphere $B$. Therefore we place inside sphere $B$ a new point charge $\beta_0$ at distance $\alpha_0$ from $B$ on the ray $AB$, and choose $\beta_0$ and $\alpha_0$ such that the sum of the potentials of the charges $\alpha_0$ and $\beta_0$ vanishes on sphere $B$. Note that the charges must be placed inside the spheres, since their potentials must be solutions to the Laplace equation outside the spheres. But, by adding charge $\beta_0$ we alter the potential on sphere $A$. To keep that potential unaltered we again place inside sphere $A$ a point charge $\alpha_1$ at distance $d_1$ from $A$ on the ray $AB$, and choose $\alpha_1$ and $d_1$ such that the potential of $\alpha_1$ and $\beta_0$ vanishes on sphere $A$. In doing so we altered the potential on sphere $B$, and so on. We keep on adding point charges inside sphere $A$ and $B$ to alternately satisfy the boundary conditions on the two spheres. This results in an infinite sequence of point charges. The value of $\Phi$ outside the two spheres is given by the sum of the potentials of these charges.
We will now use the same approach as above to find the probabilities $p(m, n)$. The products $\alpha^m \beta^n$ satisfying (10) play the role of the point charges inside the spheres. The conditions (11)–(13) on the horizontal and vertical boundary act as the boundary conditions on the two spheres.

The starting solution, which is correct far away from the vertical boundary, is given by

$$p(m, n) = \alpha_0^m \beta_0^n$$

where $\alpha_0 = \rho^2$, $\beta_0 = \rho^2/(2 + \rho)$. This solution satisfies the inner conditions (10) and the horizontal boundary conditions (11). But it violates the vertical boundary conditions (12). Therefore we add a product $c_1 \alpha_1^m \beta_1^n$ to (17) and choose $c_1$, $\alpha_1$ and $\beta_1$ with $\alpha_1$, $\beta_1$ satisfying (16) such that the sum

$$p(m, n) = \alpha_0^m \beta_0^n + c_1 \alpha_1^m \beta_1^n$$

satisfies (12). But the new term violates the horizontal boundary conditions (11). So we add again a product $c_2 \alpha_2^m \beta_2^n$ and so on. We keep on adding products, each one satisfying (10), so as to alternately satisfy the two boundary conditions. This results in an infinite sum of products. The sum is a formal solution of the equilibrium equations. What remains is the proof of convergence. This can be found in [1, 3]. The conclusion is formulated below.

**Theorem 3.2** There exist products $\alpha_i^m \beta_i^n$ satisfying (10) and coefficients $c_i$ such that the equilibrium probabilities $p(m, n)$ can be expressed as

$$p(m, n) = \sum_{i=0}^{\infty} c_i \alpha_i^m \beta_i^n, \quad m \geq 0, n > 0.$$  

**Remark 3.3** (Explicit determination of $\alpha_i$, $\beta_i$, $c_i$) The $\alpha_i$, $\beta_i$, $c_i$ mentioned in Theorem 3.2 can be solved explicitly, see [3]. Hence Theorem 3.2 provides an explicit characterization of $p(m, n)$. And based on the expression for $p(m, n)$ similar expressions may be derived for performance characteristics, such as the mean waiting time and mean queue lengths.

**Remark 3.4** (General result) Above we developed an approach to solve the shortest queue problem. But what is the scope of this approach? In [2, 3] it has been applied to a class of two dimensional Markov processes on the lattice in the positive quadrant of $\mathbb{R}^2$. For the processes in this class the transition rates are restricted to neighboring states and they are constant in the interior points and also constant on each of the axes (see figure 6). It appears that the equilibrium probabilities $p(m, n)$ can be expressed as an infinite sum of products, which can be found by the compensation approach, provided

$$q_{0,1} = q_{1,1} = q_{1,0} = 0. \quad (18)$$

So there may be no transitions in the interior points to the North, North-East and East. This result can be extended to processes of dimension 3 or higher (see Van Houtum [13]).
Figure 6: Flow diagram for a Markov process with constant rates and transitions restricted to neighboring states. $q_{i,j}$ is the transition rate from $(m, n)$ to $(m + i, n + j)$ with $m, n > 0$ and a similar notation is used for the transition rates on each of the axes.

References


