A tandem queueing network with feedback admission control

A Tandem Queueing Network with Feedback Admission Control

Lasse Leskelä[1] and Jacques Resing[2]

[1] Helsinki University of Technology
P.O. Box 1100, FI-02015 TKK, Finland
lasse.leskela@iki.fi
[2] Eindhoven University of Technology
P.O. Box 513, 5600 MB Eindhoven, The Netherlands
j.a.c.resing@tue.nl

Abstract. Admission control in queueing networks is often based on partial information on the network state. This paper studies how the lack of state information affects performance by considering a simple model for admission control. The model is analyzed by studying a related censored process that has a matrix-geometric steady-state distribution. Numerical results show how partial information may cause some performance characteristics in queueing networks to be nonmonotone with respect to service rates.

1 Introduction

Admisson control can be used to avoid congestion in heavily loaded queueing networks. In many queueing networks of practical interest, the admission controller may not be able to fully monitor the state of all queues in the network. This might happen for example in models of Internet-based services that rely on the infrastructure of several network operators. As a consequence, the admission decisions must be based on partial information on the network state. The goal of this paper is to study how the lack of state information affects the performance in this type of queueing models.

To deal with the question mathematically, we will restrict our attention to simple two-station tandem queues with unlimited buffers. Probably the simplest admission control mechanism based on partial information is the one where the admission decisions are based on the number of customers at the first server only, so that arriving customers are rejected whenever the length of the first queue exceeds a certain threshold. The steady-state decay rate of the number of customers in the second queue in this model was recently studied by Kroese, Scheinhardt, and Taylor [6]. Another well-studied class of tandem queues where the control is based on partial state information are the models where the first server stops processing when the second queue becomes too long [3,4,5,7].

In this paper we look at a different type of system where the control is based on the state of the second queue, but in such a way that only the admissions to the network, not the operation of the first server, can be controlled. More precisely,

T. Chahed and B. Tuffin (Eds.): NET-COOP 2007, LNCS 4465, pp. 129–137, 2007.

we consider a two-station tandem queueing network, where the interarrival times at station 1 and the service times in stations 1 and 2 are independent and exponentially distributed with parameters λ, μ_1, and μ_2, respectively. We denote the number of customers in station i by X_i, and assume that the admission controller accepts new customers to the system if and only if $X_2 \leq K$, see Figure 1. The stability of this system was recently studied by Leskelä [8], who showed that the queue length process $X = (X_1, X_2)$ is positive recurrent if and only if the triple (λ, μ_1, μ_2) satisfies the relation

$$\lambda \left(1 - (\mu_1/\mu_2)^{K+1}\right) < \mu_1. \tag{1}$$

Here, we will focus on determining the steady-state distribution of X. Observe that, in contrast with the other aforementioned control models, in this system both queues can grow arbitrarily big.

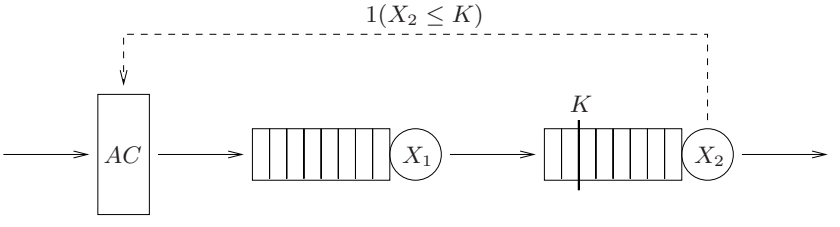

Fig. 1. The admission control mechanism

The rest of the paper is organized as follows. In Section 2, the problem of finding the steady-state distribution of the system is reduced using censoring to a simpler problem with a matrix-geometric structure. Section 3 presents an efficient numerical approach for performance evaluation of the model, with examples that illustrate how partial information affects the system. Finally, Section 4 concludes the paper.

2 Queue Length Analysis

We will assume from now on that (1) holds, so that the continuous-time Markov process $X = (X_1, X_2)$ is positive recurrent. In the sequel we will first study a modification of X restricted to periods of time during which $X_2 \leq K$. This censored process has a generator with a special block structure of the so-called $G/M/1$ type [9], which will be exploited to evaluate its steady-state distribution. Afterwards, we will find the steady-state probabilities of the process X itself.

2.1 Censoring

The behavior of the process $X = (X_1, X_2)$ during periods of time when $X_2 \leq K$ can be studied by censoring the parts of the sample path where X does not belong

to the set $S^- = \{(n,k) \in \mathbb{Z}_+^2 : k \le K\}$. The censored process $Y = (Y_1, Y_2)$ is defined by

$$Y_i(t) = X_i(\gamma(t)), \quad t \ge 0,$$

where

$$\gamma(t) = \inf\{\tau \ge 0 : \int_0^\tau 1_{\{X_2(s) \le K\}}\, ds > t\}.$$

It follows from the strong Markov property [10, Section III.21] that Y is a Markov process on S^-.

To conveniently describe the infinitesimal generator of Y, we will employ the following notation for $(K+1)$-dimensional square matrices. Denote by I the identity matrix, while T_L and T_R will stand for the left and right shift matrices given by $(T_L)_{i,j} = \delta_{i-1,j}$ and $(T_R)_{i,j} = \delta_{i+1,j}$ for $0 \le i, j \le K$ where $\delta_{i,j}$ denotes the Kronecker delta. Further, denote the projection matrices onto 0-th and K-th coordinate by U_0 and U_K, that is, $(U_0)_{i,j} = \delta_{i,0}\delta_{j,0}$ and $(U_K)_{i,j} = \delta_{i,K}\delta_{j,K}$. Ordering the states in S^- lexicographically, the generator of Y can be written in the form

$$Q = \begin{pmatrix} B_0 & A_0 & 0 & 0 & \cdots \\ B_1 & A_1 & A_0 & 0 & \cdots \\ B_2 & A_2 & A_1 & A_0 & \cdots \\ B_3 & A_3 & A_2 & A_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{pmatrix},$$

where the matrices A_n and B_n are given by

$$\begin{aligned} A_0 &= \lambda I, \\ A_1 &= \mu_2 T_L - (\lambda + \mu_1 + \mu_2) I + \mu_2 U_0, \\ A_2 &= \mu_1 (T_R + q_1 U_K), \\ A_{n+1} &= \mu_1 q_n U_K, \quad n \ge 2, \end{aligned}$$

and

$$\begin{aligned} B_0 &= \mu_2 T_L - (\lambda + \mu_2) I + \mu_2 U_0, \\ B_1 &= \mu_1 (T_R + U_K), \\ B_{n+1} &= \mu_1 (1 - q_1 - \cdots - q_n) U_K, \quad n \ge 1. \end{aligned}$$

The numbers q_n represent the probabilities that, if the process X leaves the set S^- in some state $(m+n, K)$, it enters S^- again in state (m, K), where $m \ge 1$. It is not hard to check that q_n is equal to the probability that a random walk on the integers starting at state 0 and with probabilities $\mu_1/(\mu_1+\mu_2)$ and $\mu_2/(\mu_1+\mu_2)$ of going to the right and left, respectively, reaches state -1 for the first time in exactly $2n-1$ steps. It is well-known [2] that this quantity equals

$$q_n = C_{n-1} \left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{n-1} \left(\frac{\mu_2}{\mu_1+\mu_2}\right)^n,$$

where $C_n = \frac{1}{n+1}\binom{2n}{n}$ are the Catalan numbers.

For $k = 0, \ldots, K$, denote by e_k the k-th basis vector of the $(K+1)$-dimensional euclidean space, and let $e = \sum_{k=0}^{K} e_k$. By convention, all vectors are treated as row vectors. One can check that Neuts' mean drift condition [9, Formula (1.7.11)] for the stability of Y is equivalent to (1). Furthermore, the steady-state probabilities of the censored process are given [9] in the matrix-geometric form

$$\mathrm{P}(Y = (n,k)) = x_0 R^n e_k^T, \quad (n,k) \in S^-, \tag{2}$$

where the matrix R is the unique minimal non-negative solution of

$$\sum_{n=0}^{\infty} R^n A_n = 0, \tag{3}$$

and x_0 is the unique positive row vector satisfying

$$x_0 \sum_{n=0}^{\infty} R^n B_n = 0, \quad \text{and} \quad x_0 (I-R)^{-1} e^T = 1. \tag{4}$$

2.2 Steady-State Queue Lengths

Once the steady-state distribution of the censored process Y is found, we will next show how this distribution can be used to obtain the steady-state probabilities for the queue length process X. First, note that

$$\mathrm{P}(X = (n,k)) = \mathrm{P}(X_2 \le K)\ \mathrm{P}(Y = (n,k)), \quad (n,k) \in S^-. \tag{5}$$

Second, because the steady-state rate at which customers are accepted into the system must be equal to the rate of customers coming out of the system, it follows that

$$\lambda\, \mathrm{P}(X_2 \le K) = \mu_2 (1 - \mathrm{P}(X_2 = 0)). \tag{6}$$

Note that (5) implies $\mathrm{P}(X_2 = 0) = \mathrm{P}(X_2 \le K)\, \mathrm{P}(Y_2 = 0)$. Substituting this into (6) we get

$$\mathrm{P}(X_2 \le K) = \frac{\mu_2}{\lambda + \mu_2\, \mathrm{P}(Y_2 = 0)}. \tag{7}$$

Thus, (5) and (7) yield the steady-state probabilities of X for all states $(n,k) \in S^-$, and what remains is to find the corresponding quantities on $S^+ = \mathbb{Z}_+^2 \setminus S^-$.

For the states in S^+, we will first find out the probabilities $\mathrm{P}(X = (n, K+k))$ for $n, k > 0$ by inspecting the excursions X makes in S^+. Note that if X visits the state $(n, K+k)$, the time it spends there has mean $1/(\mu_1 + \mu_2)$. Furthermore, note that the steady-state rate of transitions from $(n+m, K)$ to S^+ equals $\mu_1\, \mathrm{P}(X = (n+m, K))$. Now we see by conditioning on the state in S^- from where X enters S^+ that for all $n, k > 0$,

$$\mathrm{P}(X = (n, K+k)) = \frac{\mu_1}{\mu_1 + \mu_2} \sum_{m=k}^{\infty} q_{k,m}\ \mathrm{P}(X = (n+m, K)), \tag{8}$$

where $q_{k,m}$ is the probability that X will visit state $(n, K+k)$ during an excursion in S^+ which was initiated from state $(n+m, K)$. It is not hard to see that $q_{k,m}$ does not depend on the values of n and K, and is equal to the probability that a random walk on the integers starting from state 1 at time 0 and with probabilities $\mu_1/(\mu_1+\mu_2)$ and $\mu_2/(\mu_1+\mu_2)$ of going to the right and to the left, respectively, will visit state k at time $2m-k-1$ without visiting state 0 in any time inbetween. Using the ballot theorem (see Takács [11]) one can verify that this quantity equals

$$q_{k,m} = \frac{k}{m}\binom{2m-k-1}{m-1}\left(\frac{\mu_1}{\mu_1+\mu_2}\right)^{m-1}\left(\frac{\mu_2}{\mu_1+\mu_2}\right)^{m-k}. \tag{9}$$

Finally, the probabilities $\mathrm{P}(X=(0,k))$ with $k>K$ can be found by observing that the steady-state rate of transitions out of the set $\{(n,k') : n=0, k' \geq k\}$ equals the corresponding rate into that set, so that

$$\mu_2\,\mathrm{P}(X=(0,k)) = \mu_1 \sum_{m=k-1}^{\infty} \mathrm{P}(X_1=1, X_2=m), \quad k>K. \tag{10}$$

Alternatively, the probabilities on the left-hand side of equations (8) and (10) can be recursively determined from the balance equations, starting from the ones for $X_2=K$. In this way one avoids the infinite sums appearing on the right-hand side of equations (8) and (10).

Remark 1. In the special case where $K=0$, (3) degenerates into a scalar equation. In this case one can explicitly solve the balance equations to conclude that the steady-state distribution of the system equals

$$\mathrm{P}(X_1=n, X_2=k) = \begin{cases} \dfrac{\lambda}{\lambda+\mu_2}(1-R)\left(1-\dfrac{\mu_1}{\lambda}R\right)^{k-1}, & n=0,\ k\geq 1, \\ \dfrac{\mu_2}{\lambda+\mu_2}(1-R)\,R^n\left(1-\dfrac{\mu_1}{\lambda}R\right)^{k}, & \text{otherwise,} \end{cases}$$

where

$$R = \frac{\lambda}{2\mu_1\mu_2}\left(\sqrt{(\lambda+\mu_1-\mu_2)^2+4\mu_1\mu_2}-(\lambda+\mu_1-\mu_2)\right).$$

This result has also been independently derived by Adan and Weiss [1], who studied a two-machine 3-step re-entrant line with an infinite supply of work.

3 Performance Analysis

3.1 Throughput and Sojourn Time

We will analyse the steady-state performance of the system in terms of the throughput θ, measured as the number of customers served per unit time, and the mean sojourn time $\mathrm{E}(D)$ of accepted customers. To evaluate θ and $\mathrm{E}(D)$,

one could use formulas (5) – (10) derived in Section 2.2. However, this approach is computationally not very appealing, because it involves multiple infinite summations over the state space. As an alternative, the following theorem shows how θ and $\mathrm{E}(D)$ can be calculated directly in terms of the steady-state distribution of the censored process Y.

Theorem 1. *Assume that (1) holds. Then the steady-state throughput θ and mean sojourn time $\mathrm{E}(D)$ are given in terms of the steady-state distribution of the censored process Y by*

$$\theta = \left(\frac{1}{\lambda} \mathrm{P}(Y_2 = 0) + \frac{1}{\mu_2} \right)^{-1}, \tag{11}$$

and

$$\mathrm{E}(D) = \frac{1}{\lambda} \mathrm{E}(Y_1 1_{\{Y_2=0\}}) + \frac{1}{\mu_2} \mathrm{E}(Y_1 + Y_2 + 1). \tag{12}$$

Proof. Because $\theta = \lambda \mathrm{P}(X_2 \le K)$, the validity of (11) follows immediately from (7). To prove the second claim, let us consider the level transitions for the total amount of customers $X_1 + X_2$. Under stability, the rate of events where the value of $X_1 + X_2$ changes from n to $n+1$ is given by $\lambda \mathrm{P}(X_1 + X_2 = n, X_2 \le K)$, while the corresponding rate backwards from $n+1$ to n equals $\mu_2 \mathrm{P}(X_1 + X_2 = n+1, X_2 > 0)$. Thus,

$$\lambda \mathrm{P}(X_1 + X_2 = n, X_2 \le K) = \mu_2 \mathrm{P}(X_1 + X_2 = n+1) - \mu_2 \mathrm{P}(X_1 = n+1, X_2 = 0) \tag{13}$$

for all $n \ge 0$. Multiplying both sides of (13) by $n+1$ and then summing over n we see that

$$\lambda \mathrm{E}(\, (X_1 + X_2 + 1) 1_{\{X_2 \le K\}} \,) = \mu_2 \mathrm{E}(X_1 + X_2) - \mu_2 \mathrm{E}(X_1 1_{\{X_2 = 0\}}). \tag{14}$$

Because $\mathrm{E}(D) = \theta^{-1} \mathrm{E}(X_1 + X_2)$ by Little's law, the validity of (12) now follows from solving (14) for $\mathrm{E}(X_1 + X_2)$ and using $\theta = \lambda \mathrm{P}(X_2 \le K)$.

3.2 Long-Term Behavior of the Unstable System

To better understand how the choice of parameters affects the performance of the system, it is also interesting to see what happens in the unstable parameter region. Recall from (1) that instability of the system implies $\mu_1 < \mu_2$, so that the rate at which work is fed into the second server is strictly less than its service capacity. Thus, intuition suggests that only the first queue will grow to infinity. The next theorem verifies the validity of these heuristics.

Theorem 2. *Assume $\lambda(1 - (\mu_1/\mu_2)^{K+1}) > \mu_1$. Then the process X started from an arbitrary initial state satisfies as $t \to \infty$,*

$$\begin{aligned} X_1(t) &\to \infty \quad \textit{almost surely}, \\ X_2(t) &\to Z \quad \textit{in distribution}, \end{aligned}$$

where Z is a geometrically distributed random variable with parameter μ_1/μ_2.

Proof. Let $N_\lambda, N_{\mu_1}, N_{\mu_2}$ be independent Poisson processes with rates λ, μ_1 and μ_2, respectively. Then X can be represented as the unique solution of

$$X_1(t) = X_1(0) + \int_{(0,t]} 1_{\{X_2(s)\le K\}}\, N_\lambda(ds) - \int_{(0,t]} 1_{\{X_1(s)>0\}}\, N_{\mu_1}(ds),$$
$$X_2(t) = X_2(0) + \int_{(0,t]} 1_{\{X_1(s)>0\}}\, N_{\mu_1}(ds) - \int_{(0,t]} 1_{\{X_2(s)>0\}}\, N_{\mu_2}(ds). \qquad (15)$$

Let $\tilde{X}_2(t)$ be the solution of

$$\tilde{X}_2(t) = X_2(0) + N_{\mu_1}(t) - \int_{(0,t]} 1_{\{\tilde{X}_2(s)>0\}}\, N_{\mu_2}(ds). \qquad (16)$$

Then a pathwise comparison of (15) and (16) shows that $\tilde{X}_2(t) \ge X_2(t)$ for all t almost surely. This implies that $X_1(t) \ge U(t)$ for all t a.s., where

$$U(t) = X_1(0) + \int_{(0,t]} 1_{\{\tilde{X}_2(s)\le K\}}\, N_\lambda(ds) - N_{\mu_1}(t).$$

Note that $\tilde{X}_2$ equals the number of customers in a stable $M/M/1$ queue with arrival rate μ_1 and mean service time $1/\mu_2$. Thus, $\tilde{X}_2(t) \to Z$ in distribution, where Z is geometric with parameter μ_1/μ_2. Since N_λ is independent of $\tilde{X}_2$, it is not hard to see that $\lim_{t\to\infty} \frac{1}{t} \int_{(0,t]} 1_{\{\tilde{X}_2(s)\le K\}}\, N_\lambda(ds) = \lambda \,\mathrm{P}(Z \le K)$ a.s. Now using $\lim_{t\to\infty} \frac{1}{t} N_{\mu_1}(t) = \mu_1$ a.s., we see that with probability one,

$$\lim_{t\to\infty} U(t)/t = \lambda(1 - (\mu_1/\mu_2)^{K+1}) - \mu_1.$$

Since the above limit is strictly positive, $U(t) \to \infty$ and thus $X_1(t) \to \infty$ almost surely.

To verify that $X_2(t) \to Z$ in distribution, it is enough to show that X_2 and $\tilde{X}_2$ will couple in finite time. Let $T_0 = \sup\{t : X_1(t) = 0\}$. Since $X_1(t) \to \infty$, T_0 is a.s. finite. Define $T_1 = \inf\{t \ge T_0 : \tilde{X}_2(t) = 0\}$. Since $\tilde{X}_2$ represents the state of a stable $M/M/1$ queue, T_1 is finite a.s. Further, since X_2 is dominated by $\tilde{X}_2$, we see that $X_2(T_1) = \tilde{X}_2(T_1) = 0$. Since the pathwise dynamics of X_2 and $\tilde{X}_2$ coincide for $t \ge T_1$, we conclude that $X_2(t) = \tilde{X}_2(t)$ for all $t \ge T_1$.

3.3 Numerical Results

The steady-state distribution of the censored process $Y = (Y_1, Y_2)$ can be numerically calculated by first solving the matrix R from equation (3) using the method of successive substitutions [9], and then solving the vector x_0 from (4). The steady-state throughput θ and mean sojourn time $\mathrm{E}(D)$ are then found by combining the expressions of Theorem 1 with formula (2).

Figure 2 illustrates numerically computed contours of the throughput θ for varying μ_1 and μ_2, where $\lambda = 1$ and $K = 5$. The thick curve in the middle indicates the boundary of the stability region, so that the queue length process

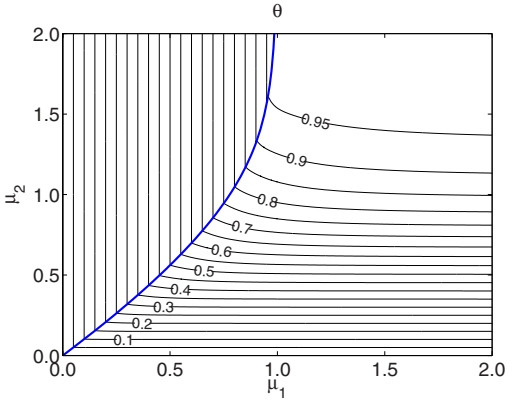

Fig. 2. Contours of θ as a function of μ_1 and μ_2 with $\lambda = 1$ and $K = 5$

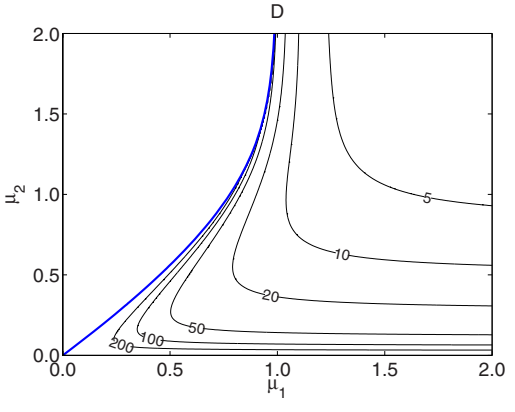

Fig. 3. Contours of $\mathrm{E}(D)$ as a function of μ_1 and μ_2 with $\lambda = 1$ and $K = 5$

$X = (X_1, X_2)$ is not positive recurrent for values of μ_1 and μ_2 located to the left of the thick curve. In the unstable region, the value of throughput does not dependent on μ_2, and is in fact equal to μ_1. This reflects the fact that $X_1(t) \to \infty$ almost surely when the system is unstable (Theorem 2), so in the long run the system serves customers at the bottleneck rate μ_1.

In Figure 3, where the corresponding contours for the mean sojourn time $\mathrm{E}(D)$ are plotted, we see that increasing the service rate for queue 2 may either increase or decrease $\mathrm{E}(D)$. In particular, if $\mu_1 < 1$, then making μ_2 large enough will eventually drive the system unstable and $\mathrm{E}(D)$ becomes infinite.

4 Conclusion

We analyzed a tandem queue with admission control based on partial information on the network state. We showed that, although the two-dimensional state

space of the system is infinite in both coordinate directions, the steady-state distribution can still be analyzed using matrix-analytic methods. The approach used in Section 2 extends to a wider class of models. For example, by making suitable modifications to the matrices A_n and B_n in Section 2.1, we can model situations where during congestion the input traffic is gradually thinned by randomly rejecting a certain proportion of the newly arriving customers. The approach of the paper can still be used as long as there exists a certain maximum threshold so that all newly arriving customers are rejected whenever the length of queue 2 exceeds this maximum threshold. A different modification to A_n and B_n allows to replace the second queue in the network by a delay node of the $M/M/\infty$ type. An interesting direction for future research is to extend this approach to networks with more than two queues.

Acknowledgments. The work presented in this article was done while both authors were visiting the Mittag–Leffler Institute.

References

1. Adan, I. J. B. F. and Weiss, G.: Analysis of a simple Markovian re-entrant line with infinite supply of work under the LBFS policy. Queueing Systems **54** (2006) 169–183.
2. Feller, W.: An Introduction to Probability Theory and Its Applications, volume I. Wiley, third edition (1966).
3. van Foreest, N. D., Mandjes, M., van Ommeren, J. C. W., and Scheinhardt, W. R. W.: A tandem queue with server slow-down and blocking. Stochastic Models **21** (2005) 695–724.
4. Grassmann, W. K. and Drekic, S.: An analytical solution for a tandem queue with blocking. Queueing Systems **36** (2000) 221–235.
5. Konheim, A. G. and Reiser, M.: A queueing model with finite waiting room and blocking. Journal of the ACM **23** (1976) 328–341.
6. Kroese, D. P., Scheinhardt, W. R. W., and Taylor, P. G.: Spectral properties of the tandem Jackson network, seen as a quasi-birth-and-death process. Ann. Appl. Probab. **14** (2004) 2057–2089.
7. Latouche, G. and Neuts, M. F.: Efficient algorithmic solutions to exponential tandem queues with blocking. SIAM J. Algebra. Discr. **1** (1980) 93–106.
8. Leskelä, L.: Stabilization of an overloaded queueing network using measurement-based admission control. J. Appl. Probab. **43** (2006) 231–244.
9. Neuts, M. F.: Matrix-Geometric Solutions in Stochastic Models. John Hopkins University Press (1981).
10. Rogers, L. C. G. and Williams, D.: Diffusions, Markov Processes, and Martingales, volume I. Wiley, second edition (1994).
11. Takács, L.: Combinatorial Methods in the Theory of Stochastic Processes. Wiley (1967).