DISSERTATION

Advanced Markov Chain Techniques in Queueing Networks

Ausgeführt zum Zwecke der Erlangung des akademischen Grades
eines Doktors der technischen Wissenschaften

unter der Leitung von

o. Univ.–Prof. Dr.–Ing. Harmen R. van As
Institut für Breitbandkommunikation
Technische Universität Wien

und

ao. Univ.–Prof. Dr. Karl Grill
Institut für Statistik und Wahrscheinlichkeitstheorie
Technische Universität Wien

eingereicht an der

Technischen Universität Wien
Fakultät für Elektrotechnik und Informationstechnik

von

Dipl.–Ing. Markus Peter Sommereder
Matrikelnummer 0225618

Wien, 5. Februar 2009

Kurzfassung

Die Modellierung mit Markov-Ketten ist ein elegantes und äußerst effizientes Verfahren zur Untersuchung von Warteschlangensystemen: Es ist für eine große Klasse von Warteschlangensystemen geeignet, die zugrunde liegende Theorie ist nicht allzu schwierig zu erlernen, und man kann damit viele verschiedene Informationen zum untersuchten Warteschlangensystem erhalten.

Derzeit werden Markov-Ketten hauptsächlich verwendet, um die transienten oder stationären Zustandswahrscheinlichkeiten eines Warteschlangensystems zu ermitteln. Mithilfe dieser Zustandswahrscheinlichkeiten können beispielsweise die Anzahl der Anforderungen im System oder die Auslastung der Bedieneinheiten berechnet werden. Manchmal wird auch die Durchflusszeit durch ein Warteschlangensystem mit Markov-Ketten bestimmt oder es wird ermittelt, wieviel Zeit vergeht, bis ein bestimmter Zustand erreicht wird. Damit werden die Möglichkeiten der Modellierung mit Markov-Ketten jedoch nicht ausgeschöpft.

In dieser Arbeit werden fortgeschrittene Techniken der Modellierung von Warteschlangensystemen mit zeitkontinuierlichen Markov-Ketten gezeigt. Wir zeigen Techniken zur Analyse von Leerlauf- und Arbeitsphasen der Bedieneinheiten von Warteschlangensystemen (Länge der Leerlaufphase, Länge der Arbeitsphase, Anzahl der während einer Arbeitsphase bedienten Anforderungen), des Ausgangsstroms von Warteschlangensystemen mit einer Bedieneinheit (Zwischenereigniszeiten des Ausgangsstroms) und des Überlauf-Verkehrs von Warteschlangensystemen (Zeit zwischen zwei Abweisungen, Anzahl der erfolgreichen Ankünfte zwischen zwei Abweisungen). Weiters wird gezeigt, wie Markov-Ketten verwendet werden können, um die Überlagerung und die Aufteilung von Verkehrsströmen zu untersuchen (Zwischenereigniszeiten).

Die gezeigten Techniken werden anhand zahlreicher Beispiele erläutert. Für die praktische Anwendung wichtige verwandte Themen, wie die Annäherung von gegebenen Verteilungen durch Phasenverteilungen und die Auswirkung statistischer Abhängigkeiten innerhalb von Verkehrsströmen, werden ebenfalls besprochen.

Abstract

Modelling with Markov chains is a very powerful and efficient technique for the investigation of queueing systems: it is suitable for a broad class of queueing systems, the underlying theory is relatively easy to understand, and many different characteristics of queueing systems can be explored. Traditionally, Markov chains are used to calculate the transient or stationary system state probabilities of queueing systems, from which characteristics such as the number of customers in the system and the server utilisation can be obtained. Sometimes the flow time through a queueing system is determined using Markov chains, or the time that is needed to reach a certain state. However, with these applications the capabilities of Markov chain modelling are not fully utilised.

In this work, more advanced ways of modelling queueing systems with continuous-time Markov chains are presented. We show techniques to analyse the idle and the busy period of queueing systems (length of the idle period, length of the busy period, number of customers served during the busy period), the departure stream of single-server queueing systems (interdeparture times), and the overflow traffic of queueing systems (interoverflow times, number of successful arrivals between two overflows). Moreover, we show how Markov chains can be used to analyse the superposition and the decomposition of traffic streams (interevent times).

The techniques are explained with many examples. Related issues, which are important for the practical application, such as the approximation of given distributions by phase-type distributions and the effects of statistical interdependence within traffic streams, are also discussed.

Contents

Part I.

Introduction

1. Introduction

Modelling with Markov chains is a very powerful and efficient technique for the investigation of queueing systems: it is suitable for a broad class of queueing systems, the underlying theory is relatively easy to understand, and many different characteristics of queueing systems can be explored.

Unfortunately, the full potential of modelling queueing systems with Markov chains is seldom utilised. Mostly only the system state probabilities are calculated and then simple characteristics such as the number of customers in the system and the server utilisation are obtained.

The aim of this work is to show more sophisticated ways of modelling queueing systems with continuous-time Markov chains, whereby, in particular, we focus on the investigation of traffic streams. We present Markov chain techniques for the analysis of

- the idle and the busy period of queueing systems,
- the departure stream of queueing systems,
- the overflow stream of queueing systems,
- the superposition of streams, and
- the decomposition of streams.

An application of the techniques presented is network decomposition. Network decomposition means that a network, which is too complex to be analysed as a whole, is broken up into several subsystems, which are small enough to be analysable. These subsystems are then investigated individually, whereby the input stream into a subsystem consists – depending on the network topology – of the output streams of one or more other subsystems (Figure 1.1).[1]

The Markov chain techniques presented in this work cover all tasks needed to analyse a queueing network using network decomposition.

[1] Since, in most cases, the traffic streams between the subsystems cannot be described exactly and have to be approximated, e.g., by renewal processes that match in the first k moments, the network decomposition method does not yield exact results. However, by varying the size of the subsystems and the precision with which the traffic streams are described, the accuracy of the results can be controlled.

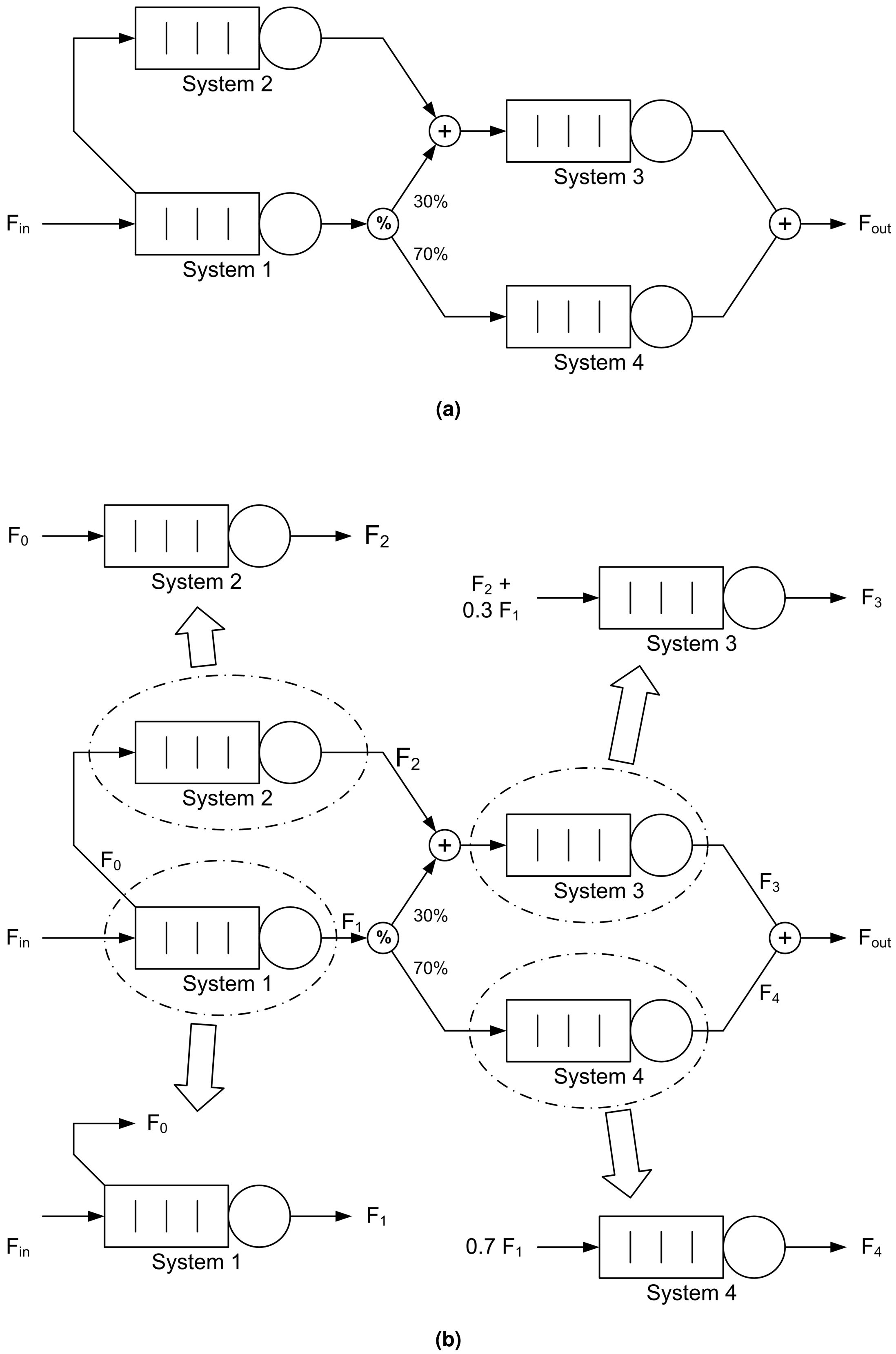


Figure 1.1.: Network decomposition. The queueing network shown in (a) is broken up into subsystems (b), which are analysed individually. Since the input of some subsystems depends on the output of other subsystems, we must choose an appropriate order, e.g., system 1 – system 4 – system 2 – system 3.

Our work is organised as follows:

Part I, “Introduction”, gives an introduction to modelling with Markov chains.

After an overview of the mathematical theory of Markov chains in Chapter 2, Chapter 3 reviews well-known techniques currently used to model queueing systems with Markov chains, such as determining the system state and the flow time through a queueing system. It is also shown how general probability distributions can be approximated by phase type distributions. In Chapter 4, we briefly discuss some topics related to the content of this work.

In part II, “Advanced Markov Chain Techniques”, the advanced Markov chain techniques are presented. For the sake of clarity, we chose simple examples to explain the techniques. Of course, it is possible to apply the techniques to queueing systems of arbitrary complexity. For some of the problems described other techniques rather than modelling with Markov chains might be more appropriate. However, even in such cases it can be useful to have, in addition, a Markov chain model so that the solutions obtained can be verified.

In Chapter 5, we analyse the idle and the busy period of queueing systems: the length of the idle period, the length of the busy period, and the number of customers served during a busy period.

These results are needed in Chapter 6, where the departure stream of single-server queueing systems is considered.

In Chapter 7, we investigate characteristics of overflow traffic. We determine the blocking probability, the interoverflow time, and the number of successful arrivals between two overflows.

Finally, Chapters 8 and 9 deal with the manipulation of traffic streams: In Chapter 8, we consider the superposition of streams, and in Chapter 9, we analyse the decomposition (splitting) of streams.

2. Markov chains

2.1. Stochastic processes

Stochastic processes

A **stochastic process** (or **random process**) $X_t,\ t \in \mathcal{T}$ is a family of random variables defined on a given probability space.

The ordered set $\mathcal{T}$ is called the **index set**. $\mathcal{T}$ is usually interpreted as time; so X_t denotes the value of the stochastic process at time t. When the index set is countable (e.g., $\mathcal{T} = \mathbb{N}$), the process is said to be a discrete-time process, otherwise (e.g., when $\mathcal{T} = \mathbb{R}_0^+$) it is said to be a continuous-time stochastic process.

The values that can be assumed by the random variables X_t are called the **states** of the stochastic process. The **state space** is the set of all states. The state space can be finite, countably infinite or uncountably infinite.

Markov processes

A Markov process (named after the Russian mathematician Andrej A. Markov, 1856-1922) is a stochastic process, whose further evolution depends only on its current state, and not on the past history, that is, when the current state has been reached and how it has been reached:

$$\begin{aligned}\mathrm{P}\left\{X_t \leq x \mid X_{t_0} = x_0, X_{t_1} = x_1, \ldots, X_{t_n} = x_n\right\} = \\ \mathrm{P}\left\{X_t \leq x \mid X_{t_n} = x_n\right\} \quad \text{for all } t_0 < t_1 < \cdots < t_n < t \end{aligned} \tag{2.1}$$

This important property is called **Markov property** or **memoryless property**.

Markov processes are much easier to analyse than general random processes. In many cases, it is possible to convert general random processes into Markov processes.

If the behaviour of a Markov process is independent of the absolute time, that is, if

$$\mathrm{P}\left\{X_t = j \mid X_0 = i\right\} = \mathrm{P}\left\{X_{t+s} = j \mid X_s = i\right\} \tag{2.2}$$

holds for all $s \in \mathcal{T}$, the Markov process is said to be **homogeneous**, otherwise it is said to be **nonhomogeneous**. In the following, we consider only homogeneous Markov processes.

Markov chains

Markov processes with a countable state space are called **Markov chains**. Usually the state space of Markov chains is set to the natural numbers $\mathbb{N}$ or to a subset of $\mathbb{N}$.

In this work, we number the states of the Markov chains with the positive integers 1,2,3, ... For the sake of readability, in addition, we give a unique name to all states, such as "Idle", "2", "3/1/a". These names help to understand what the respective state means. We use these two notations as needed. To distinguish the number of a state from its name, we enclose the names in angle brackets. For example, $X_t = 2$ means that the Markov chain is in the state with number 2, whereas $X_t = \langle 2 \rangle$ means that the Markov chain is in the state with the name "2".

In the following, we discuss some important properties of discrete-time Markov chains (DTMC) and continuous-time Markov chains (CTMC).

2.2. Discrete-time Markov chains

2.2.1. Transition probabilities

Discrete-time Markov chains can change their state at discrete points in time. The probability for such a change is determined by the transition probabilities.

The transition probability p_{ij} is the probability that the Markov chain will be in state j at the time $n+1$, given that it is in state i at time n:

$$p_{ij} = \mathrm{P}\left\{X_{n+1} = j \mid X_n = i\right\} \tag{2.3}$$

The transition probabilities p_{ij} can be written as a matrix $\mathcal{P}$:

$$\mathcal{P} = (p_{ij}) = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots \\ p_{2,1} & p_{2,2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{2.4}$$

This matrix is called the **transition probability matrix** for the discrete-time Markov chain.

2.2.2. Multi-step transition probabilities

The probability $p_{ij}^{(m)}$ that the Markov chain is in state j at time $n+m$ $(m \geq 2)$, given that it is in state i at time n,

$$p_{ij}^{(m)} = \mathrm{P}\left\{X_{n+m} = j \mid X_n = i\right\} \tag{2.5}$$

can be calculated as follows: In order to get in m steps from state i to state j, first an intermediate state k must be reached in l $(0 < l < m)$ steps (probability $p_{ik}^{(l)}$). Then there must be a transition from this intermediate state k to state j in $m-l$ steps (probability $p_{kj}^{(m-l)}$). For the intermediate states all states of the Markov chain are possible. Therefore, we have

$$p_{ij}^{(1)} = p_{ij} \tag{2.6}$$

$$p_{ij}^{(m)} = \sum_{\text{all } k} p_{ik}^{(l)}\, p_{kj}^{(m-l)}, \quad \text{for } 0 < l < m \tag{2.7}$$

or in matrix notation

$$\mathcal{P}^{(1)} = \mathcal{P} \tag{2.8}$$

$$\mathcal{P}^{(m)} = \mathcal{P}^{(l)}\,\mathcal{P}^{(m-l)}, \quad \text{for } 0 < l < m \tag{2.9}$$

From this it follows (with $l = 1$)

$$\mathcal{P}^{(m)} = \mathcal{P}\,\mathcal{P}^{(m-1)} = \mathcal{P}\,\mathcal{P}\,\mathcal{P}^{(m-2)} = \cdots = \mathcal{P}^{m} \tag{2.10}$$

2.2.3. Representation by state diagrams

Discrete-time Markov chains can be graphically represented by state diagrams: The states of the Markov chains are represented by circles, and the possible transitions between the states are represented by arcs, which are annotated with the transition probabilities.

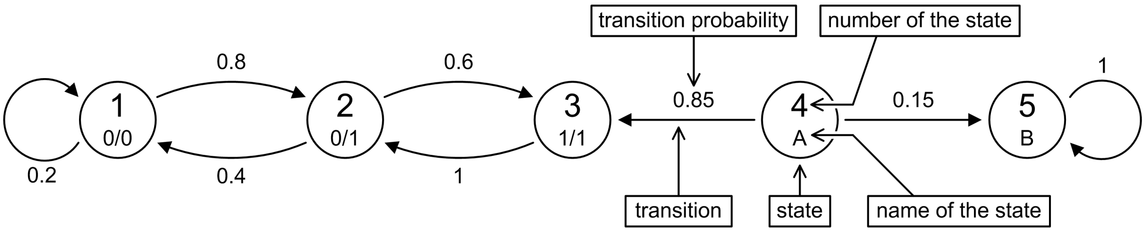


Figure 2.1.: State diagram of a discrete-time Markov chain.

Figure 2.1 shows the state diagram for a discrete-time Markov chain with the transition probabilities

$$\mathcal{P} = \begin{pmatrix} 0.2 & 0.8 & 0 & 0 & 0 \\ 0.4 & 0 & 0.6 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0.85 & 0 & 0.15 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

2.2.4. Classification of states

A state j is **accessible** from a state i ($i \rightsquigarrow j$), if j can be reached from i in an arbitrary number of steps:

$$i \rightsquigarrow j \Leftrightarrow \exists\, m \geq 0 : p_{ij}^{(m)} > 0 \tag{2.11}$$

A set of states is said to be **closed** if no state outside the set can be reached from a state within the set. If a closed set consists of only one state, this state is said to be **absorbing.**

Two states i and j **communicate** ($i \leftrightsquigarrow j$), if they are accessible to each other:

$$i \leftrightsquigarrow j \Leftrightarrow i \rightsquigarrow j \wedge j \rightsquigarrow i \tag{2.12}$$

The relation of communication is an equivalence relation ($i \leftrightsquigarrow i, i \leftrightsquigarrow j \Leftrightarrow j \leftrightsquigarrow i$, and $i \leftrightsquigarrow j \wedge j \leftrightsquigarrow k \Rightarrow i \leftrightsquigarrow k$ for all i, j, k). The associated equivalence classes are called **communication classes**.

A Markov chain is said to be **irreducible** if all states communicate with each other, that is, there is only one communication class.

The states of a communication class that is not closed are said to be **transient**. Transient states are taken only a finite number of times (because once the communication class of such a state is left, it cannot be reached again).

In the Markov chain shown in Figure 2.1, there are three communication classes: $\{1, 2, 3\}$, $\{4\}$ and $\{5\}$. The sets $\{1, 2, 3\}$ and $\{5\}$ are closed, state 5 is absorbing. State 4 is transient.

States which are not transient are said to be **recurrent**. When a recurrent state is taken once, it is taken infinitely often.

With $f_i^{(n)}$ we denote the probability that the Markov chain will return to state i in exactly n steps after leaving it. The probability that the Markov chain ever returns to state i (that is, in an arbitrary number of steps), is

$$f_i = \sum_{m=1}^{\infty} f_i^{(m)} \tag{2.13}$$

For a recurrent state i this probability is $f_i = 1$, for a transient state j it is $f_j < 1$.

The expected number of steps between two consecutive sojourns in a recurrent state i (the **mean recurrence time**) is

$$M_i = \sum_{n=1}^{\infty} n\, f_i^{(n)} \tag{2.14}$$

When the mean recurrence time is finite ($M_i < \infty$), state i is said to be **positive-recurrent**, otherwise **null-recurrent**. Null-recurrent states can occur only in Markov chains with an infinite state space.

The states of an irreducible Markov chain are all either positive-recurrent, null-recurrent or transient.

The states of a finite, irreducible Markov chain are all positive-recurrent.

A state j is said to be **periodic** with period p, when, after the state has been left, a return to the state is possible only in a number of steps, which is a multiple of p. That is, p is the greatest common divisor or all numbers r, for which $p_{jj}^{(r)} > 0$. A state with period $p = 1$ is said to be **aperiodic**.

If a state of an irreducible Markov chain is aperiodic, then all other states of the Markov chains are aperiodic, too.

A discrete-time Markov chain is said to be **ergodic**, if it is irreducible, and all states of the Markov chain are aperiodic and positive-recurrent.

2.2.5. State probabilities

The probability that the Markov chain is in state i at time n (state probability) is $\pi_i(n)$,

$$\pi_i(n) = \mathrm{P}\left\{X_n = i\right\} \tag{2.15}$$

The probability distribution of a Markov chain (the state probabilities of all states) can be written as a vector,

$$\pi(n) = (\pi_0(n),\ \pi_1(n),\ \dots) \tag{2.16}$$

If the initial probability distribution $\pi(0)$ is known, the state probabilities can be calculated with

$$\pi_i(n) = \sum_{\text{all } k} p_{ki}^{(n)}\ \pi_k(0) \tag{2.17}$$

or in matrix notation

$$\pi(n) = \pi(0)\ \mathcal{P}^{(n)} = \pi(0)\ \mathcal{P}^n \tag{2.18}$$

A **stationary probability distribution** $\pi = (\pi_1,\ \pi_2,\ \dots)$ is a distribution where

$$\pi \cdot \mathcal{P} = \pi \tag{2.19}$$

holds. That is, when the Markov chain has reached a stationary probability distribution, it will be retained forever.

If a certain initial probability distribution $\pi(0)$ is given and the limit $\tilde{\pi} = \lim\limits_{n\to\infty} \pi(n)$ exists, then $\tilde{\pi}$ is called **limiting probability distribution**.

For the existence and uniqueness of the two probability distributions, we have:

- In an aperiodic Markov chain, there exists the limiting probability distribution for every initial probability distribution. In a periodic Markov chain, this is not the case, which can easily be seen when we look at a Markov chain with $\mathcal{P} = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$.
- In an aperiodic and irreducible Markov chain, there exists a unique limiting probability distribution, which is independent of the initial probability distribution. In a reducible Markov chain, this is not the case. For example, in a Markov chain with $\mathcal{P} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ every initial probability distribution is its own limiting distribution.
- In an irreducible and positive-recurrent Markov chain, there is a unique stationary probability distribution. In a reducible Markov chain, this is not the case. For example, in a Markov chain with $\mathcal{P} = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ every probability distribution is stationary.
- In an aperiodic, irreducible, and positive-recurrent (that is, ergodic) Markov chain, there is a unique limiting probability distribution, which equals the unique stationary probability distribution. This probability distribution can be calculated by solving equation (2.19) under the side condition $\sum_i \pi_i = 1$. Alternatively π can be obtained from the relation

$$\lim_{n\to\infty} \mathcal{P}^n = \left(\pi,\ \pi,\ \dots\right)^T \tag{2.20}$$

- In an aperiodic, irreducible, but not positive-recurrent Markov chain (such Markov chains have an infinite state space), there is no stationary probability distribution.

2.2.6. Sojourn times

The time R_i that a Markov chain spends in state i is called **sojourn time**. Because of the Markov property at any point in time the remaining sojourn time of a state is independent of the time already spent in the state. Therefore, the sojourn times are geometrically distributed,

$$\mathrm{P}\left\{R_i = k\right\} = p_{ii}^{k-1} \sum_{j \neq i} p_{ij} \tag{2.21}$$

The mean sojourn time is

$$\mathrm{E}\left(R_i\right) = \frac{1}{1 - p_{ii}} \tag{2.22}$$

and its variance is

$$\mathrm{Var}\left(R_i\right) = \frac{p_{ii}}{(1 - p_{ii})^2} \tag{2.23}$$

2.3. Continuous-time Markov chains

2.3.1. Transition rates

Continuous-time Markov chains can change their state at each point in time. Therefore, the transition probabilities depend on the considered interval:

The **transition probability** $p_{ij}(\tau)$ $(\tau \geq 0)$ is the probability that the Markov chain is in state j at time $t + \tau$, given that it is in state i at time t:

$$p_{ij}(\tau) = \mathrm{P}\left\{X_{t+\tau} = j \mid X_t = i\right\} \tag{2.24}$$

The smaller the interval τ is, the smaller the probability for a transition to another state is. For $\tau = 0$ we have

$$p_{ij}(0) = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases} \tag{2.25}$$

As the interval τ becomes larger, the probability for a transition increases, whereby also multiple transitions (that is, the Markov chain changes from state i to some intermediate state k, and then from state k to state j) can occur and have to be taken into account. In order to describe only single transitions, τ must be chosen as small as possible.

The **transition rate** q_{ij} is defined as the derivation of the transition probability $p_{ij}(\tau)$ at time $\tau = 0$ (see Figure 2.2):

$$q_{ij}(\tau) = p'_{ij}(0) = \lim_{\tau \to 0} \frac{p_{ij}(\tau) - p_{ij}(0)}{\tau - 0} \tag{2.26}$$

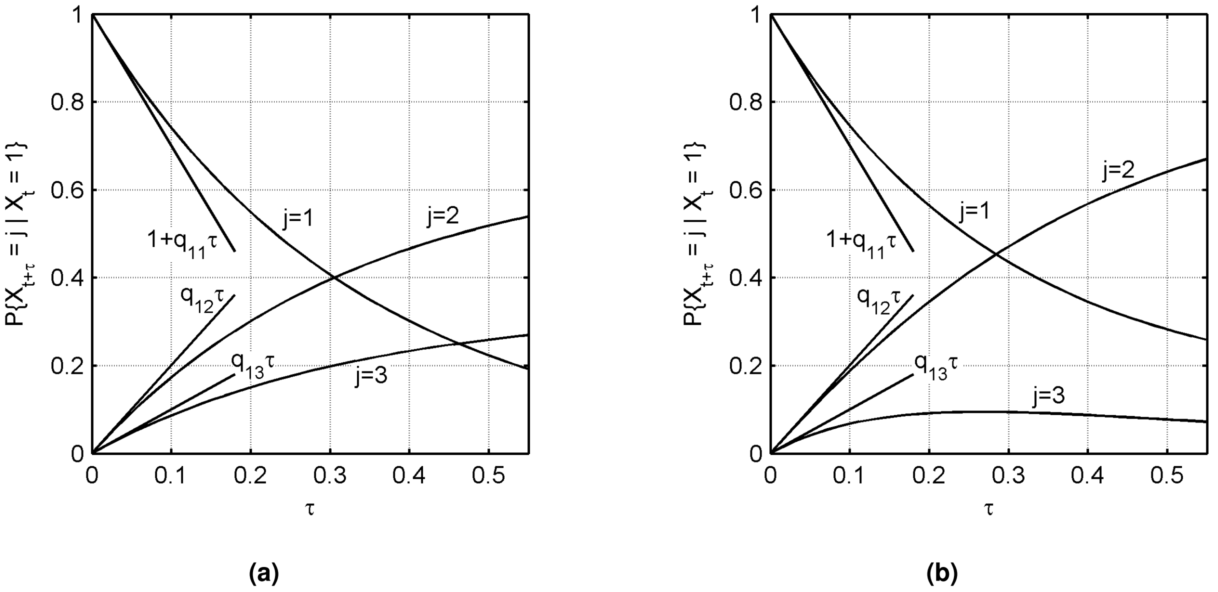


Figure 2.2.: The transition rate is defined as the derivation of the transition probability $p_{ij}(\tau)$ at $\tau = 0$. (a) $\mathcal{Q} = \begin{pmatrix} -3 & 2 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. (b) $\mathcal{Q} = \begin{pmatrix} -3 & 2 & 1 \\ 0.5 & -0.5 & 0 \\ 0 & 5 & -5 \end{pmatrix}$.

For $j \neq i$ we have

$$q_{ij}(\tau) = \lim_{\tau \to 0} \frac{p_{ij}(\tau)}{\tau} \tag{2.27}$$

For $j = i$ we have

$$q_{ii}(\tau) = \lim_{\tau \to 0} \frac{p_{ii}(\tau) - 1}{\tau} \tag{2.28}$$

With $p_{ii}(\tau) = 1 - \sum_{j \neq i} p_{ij}(\tau)$ follows

$$q_{ii}(\tau) = \lim_{\tau \to 0} \frac{1 - \sum_{j \neq i} p_{ij}(\tau) - 1}{\tau} = -\sum_{j \neq i} j \lim_{\tau \to 0} \frac{p_{ij}(\tau)}{\tau} \tag{2.29}$$

$$q_{ii}(\tau) = -\sum_{j \neq i} q_{ij}(\tau) \tag{2.30}$$

The transition rates can also be written as a matrix,

$$\mathcal{Q} = (q_{ij}) = \begin{pmatrix} q_{1,1} & q_{1,2} & \cdots \\ q_{2,1} & q_{2,2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{2.31}$$

This matrix is called the **transition rate matrix** or **infinitesimal generator matrix** for the continuous-time Markov chain.

2.3.2. Representation by state diagrams

Continuous-time Markov chains can be graphically represented by state diagrams: The states are represented by circles, and the possible transitions between states are represented by arcs, which are annotated with the transition rates.

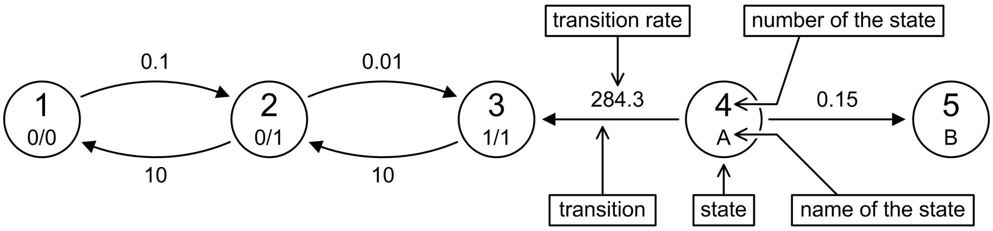


Figure 2.3.: State diagram of a continuous-time Markov chain.

Figure 2.3 shows the state diagram for a continuous-time Markov chain with the transition rates

$$\mathcal{Q} = \begin{pmatrix} -0.1 & 0.1 & 0 & 0 & 0 \\ 10 & -10.01 & 0.01 & 0 & 0 \\ 0 & 10 & -10 & 0 & 0 \\ 0 & 0 & 284.3 & -284.45 & 0.15 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

2.3.3. Chapman-Kolmogorov equations

The transition probability between two states (also considering multiple transitions) can be calculated with the aid of the following consideration:

Let the Markov chain be in state i at time t. If the Markov chain is in state j at time $t + \tau + \alpha$, then it must have been at time $t + \tau$ in some intermediate state k. The probability for this is $p_{ik}(\tau)$. Then there must have been a transition from state k to state j in α time units. The probability for this is $p_{kj}(\alpha)$. Since all states of the Markov chain (including i and j) can be intermediate states, we have

$$p_{ij}(\tau + \alpha) = \sum_k p_{ik}(\tau)\, p_{kj}(\alpha) \tag{2.32}$$

Because of $\tau + \alpha = \alpha + \tau$ we also have

$$p_{ij}(\tau + \alpha) = \sum_k p_{ik}(\alpha)\, p_{kj}(\tau) \tag{2.33}$$

From this it follows that

$$\begin{aligned} p_{ij}(\tau + \alpha) - p_{ij}(\tau) &= \sum_{k \neq j} p_{ik}(\tau)\, p_{kj}(\alpha) + p_{ij}(\tau)\, p_{jj}(\alpha) - p_{ij}(\tau) \\ &= \sum_{k \neq j} p_{ik}(\tau)\, p_{kj}(\alpha) + p_{ij}(\tau)\ \left(p_{jj}(\alpha) - 1\right) \end{aligned}$$

and

$$\begin{aligned} p_{ij}(\tau+\alpha) - p_{ij}(\tau) &= \sum_{k\neq i} p_{ik}(\alpha)\, p_{kj}(\tau) + p_{ii}(\alpha)\, p_{ij}(\tau) - p_{ij}(\tau) \\ &= \sum_{k\neq i} p_{ik}(\alpha)\, p_{kj}(\tau) + p_{ij}(\tau)\ \ (p_{ii}(\alpha) - 1) \end{aligned}$$

With equation (2.25) follows

$$p_{ij}(\tau+\alpha) - p_{ij}(\tau) = \sum_{k\neq j} p_{ik}(\tau)\ \ (p_{kj}(\alpha) - p_{kj}(0)) + p_{ij}(\tau)\ \ (p_{jj}(\alpha) - p_{jj}(0)) \tag{2.34}$$

and

$$p_{ij}(\tau+\alpha) - p_{ij}(\tau) = \sum_{k\neq i} p_{kj}(\tau)\ \ (p_{ik}(\alpha) - p_{ik}(0)) + p_{ij}(\tau)\ \ (p_{ii}(\alpha) - p_{ii}(0)) \tag{2.35}$$

Dividing both sides by α and taking the limit $\alpha \to 0$ yields

$$\begin{aligned} &\lim_{\alpha\to 0} \underbrace{\frac{p_{ij}(\tau+\alpha) - p_{ij}(\tau)}{\alpha}}_{\to p'_{ij}(\tau)} = \\ &\qquad \lim_{\alpha\to 0} \sum_{k\neq j} p_{ik}(\tau)\, \underbrace{\frac{p_{kj}(\alpha) - p_{kj}(0)}{\alpha}}_{\to p'_{kj}(0)=q_{kj}} + \lim_{\alpha\to 0} p_{ij}(\tau)\, \underbrace{\frac{p_{jj}(\alpha) - p_{jj}(0)}{\alpha}}_{\to p'_{jj}(0)=q_{jj}} \end{aligned} \tag{2.36}$$

and

$$\begin{aligned} &\lim_{\alpha\to 0} \underbrace{\frac{p_{ij}(\tau+\alpha) - p_{ij}(\tau)}{\alpha}}_{\to p'_{ij}(\tau)} = \\ &\qquad \lim_{\alpha\to 0} \sum_{k\neq i} p_{kj}(\tau)\, \underbrace{\frac{p_{ik}(\alpha) - p_{ik}(0)}{\alpha}}_{\to p'_{ik}(0)=q_{ik}} + \lim_{\alpha\to 0} p_{ij}(\tau)\, \underbrace{\frac{p_{ii}(\alpha) - p_{ii}(0)}{\alpha}}_{\to p'_{ii}(0)=q_{ii}} \end{aligned} \tag{2.37}$$

The first equation is called the **Chapman-Kolmogorov forward equation**:

$$\boxed{p'_{ij}(\tau) = \sum_{k} p_{ik}(\tau)\ q_{kj}} \tag{2.38}$$

The second equation is the **Chapman-Kolmogorov backward equation**:

$$\boxed{p'_{ij}(\tau) = \sum_{k} q_{ik}\ p_{kj}(\tau)} \tag{2.39}$$

2.3.4. Classification of states

States of a countinuous-time Markov chain are classified similarly to states of discrete-time Markov chains:

A state j is **accessible** from state i ($i \rightsquigarrow j$), if j can be reached from i in arbitrary time:

$$i \rightsquigarrow j \Leftrightarrow \exists\, t : p_{ij}(t) > 0 \tag{2.40}$$

A set of states is said to be **closed** if no state outside the set can be reached from a state within the set. When a closed set consists of only one state, that state is said to be **absorbing**.

Two states i and j **communicate** ($i \leftrightsquigarrow j$), if both are accessible to each other:

$$i \leftrightsquigarrow j \Leftrightarrow i \rightsquigarrow j \wedge j \rightsquigarrow i \tag{2.41}$$

A **communication class** is a set which contains all communicating states.

A Markov chain is said to be **irreducible** if all states communicate with each other, that is, there is only one communication class.

The states of a communication class that is not closed are said to be **transient**. Transient states are taken only a finite number of times (because once the communication class of such a state is left, it cannot be reached again).

States that are not transient are said to be **recurrent**. When a recurrent state is taken once, it is taken infinitely often.

The expected time between two consecutive sojourns in a recurrent state i is called **mean recurrence time**. When the mean recurrence time is finite, state i is said to be **positive-recurrent**, otherwise **null-recurrent**.

A continuous-time Markov chain is said to be **ergodic** if it is irreducible and all states are positive-recurrent.

2.3.5. State probabilities

The probability $\pi_j(\tau)$ that the Markov chain is in state j at time τ can be determined by means of the Chapman-Kolmogorov forward equation (2.38).

We have

$$\pi_j(\tau) = \sum_i \pi_i(0)\, p_{ij}(\tau) \tag{2.42}$$

Differentiation of both sides yields:

$$\pi_j'(\tau) = \sum_i \pi_i(0)\; p_{ij}'(\tau) = \sum_i \pi_i(0) \sum_k p_{ik}(\tau)\; q_{kj} = \sum_k \underbrace{\sum_i \pi_i(0)\; p_{ik}(\tau)}_{\pi_k(\tau)}\; q_{kj} \tag{2.43}$$

$$\boxed{\pi_j'(\tau) = \sum_k \pi_k(\tau)\; q_{kj}} \tag{2.44}$$

Using $\pi(\tau) = (\pi_1(\tau),\; \pi_2(\tau),\; \ldots,\; \pi_N(\tau))$, this equation can be written in matrix form:

$$\boxed{\pi'(\tau) = \pi(\tau)\; \mathcal{Q}} \tag{2.45}$$

A probability distribution $\pi = (\pi_1,\; \pi_2,\; \ldots)$ is called **stationary probability distribution**, if

$$\pi' = \pi\; \mathcal{Q} = 0 \tag{2.46}$$

holds. This means that if a Markov chain has reached a stationary probability distribution, the change in the probability distribution is 0 and the probability distribution will be retained forever.

If an initial probability distribution $\pi(0)$ is given and the limit $\tilde{\pi} = \lim\limits_{t\to\infty} \pi(t)$ exists, $\tilde{\pi}$ is called **limiting probability distribution**.

For the existence and the uniqueness of these two probability distributions, the following holds:

- In an irreducible Markov chain, there is a unique limiting probability distribution, which is independent of the initial probability distribution.
- In an irreducible and finite Markov chain, there is unique stationary probability distribution. This probability distribution can be obtained by solving equation (2.46) under the side condition $\sum_i \pi_i = 1$.

2.3.6. Sojourn times

The time R_i that a Markov chain spends in a state i is called **sojourn time**. Because of the memory-less property of Markov chains, the sojourn times are exponentially distributed (with parameter $-q_{ii}$):

$$\mathrm{P}\left\{R_i \le t\right\} = 1 - \mathrm{e}^{q_{ii} t} \tag{2.47}$$

The mean of the sojourn time is

$$\mathrm{E}\left(R_i\right) = -\frac{1}{q_{ii}} \tag{2.48}$$

and the variance is

$$\mathrm{Var}\left(R_i\right) = \frac{1}{{q_{ii}}^2} \tag{2.49}$$

Sometimes we are interested in the probability that the Markov chain is at a certain point in time in a state that does not belong to a subset $\mathcal{H}$ of the state space. This probability can be calculated with the Chapman-Kolmogorov backward equation:

Let $\varphi_i(\tau)$ be the probability that the Markov chain is in a state that is not contained in $\mathcal{H}$ after τ time units, given that it is in state i at time t:

$$\varphi_i(\tau) = P\{X_{t+\tau} \notin \mathcal{H} \mid X_t = i\} \tag{2.50}$$

Then we have

$$\varphi_i(\tau) = \sum_{j \notin \mathcal{H}} p_{ij}(\tau) \tag{2.51}$$

Differentiation of both sides yields

$$\varphi_i'(\tau) = \sum_{j \notin \mathcal{H}} p_{ij}'(\tau) = \sum_{j \notin \mathcal{H}} \sum_k q_{ik}\, p_{kj}(\tau) = \sum_k q_{ik} \underbrace{\sum_{j \notin \mathcal{H}} p_{kj}(\tau)}_{\varphi_k(\tau)} \tag{2.52}$$

$$\boxed{\varphi_i'(\tau) = \sum_k q_{ik}\, \varphi_k(\tau)} \tag{2.53}$$

Using $\varphi(\tau) = (\varphi_1(\tau),\ \varphi_2(\tau),\ \ldots,\ \varphi_N(\tau))^T$, this equation can be written in matrix form:

$$\boxed{\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau)} \tag{2.54}$$

For the initial conditions $\varphi_i(0)$, we have

$$\varphi_i(0) = 1 \quad \forall i \notin \mathcal{H} \qquad\qquad \varphi_i(0) = 0 \quad \forall i \in \mathcal{H} \tag{2.55}$$

2.3.7. Embedded Markov chain

If we consider only the transitions between the states of a continuous-time Markov chain and not the time spent in the states, we obtain a discrete-time Markov chain, the so-called **embedded Markov chain**. The embedded Markov chain has the same state space as the continuous-time Markov chain.

For the calculation of the transition probabilities s_{ij} of the embedded Markov chain, we first show that Y_i is the smallest of the exponentially distributed random variables $Y_1, Y_2, \ldots, Y_n$ (with respective rates $\lambda_1, \lambda_2, \ldots, \lambda_n$) with probability $\lambda_i / (\lambda_1 + \lambda_2 + \cdots + \lambda_n)$:

$$\begin{aligned}P\Big\{Y_i = \min_j Y_j\Big\} &= \mathrm{P}\Big\{Y_i \le \overbrace{\min_{j\neq i} Y_j}^{=:Y_M}\Big\} = \mathrm{P}\left\{Y_i \le Y_M\right\} = \\ &= \int_0^\infty \mathrm{P}\left\{Y_i \le Y_M \mid Y_i = x\right\} \mathrm{P}\left\{Y_i = x\right\} \mathrm{d}x = \\ &= \int_0^\infty \mathrm{P}\left\{x \le Y_M\right\} \mathrm{P}\left\{Y_i = x\right\} \mathrm{d}x = \\ &= \int_0^\infty \mathrm{e}^{-\sum_{j\neq i} \lambda_j x} \lambda_i \mathrm{e}^{-\lambda_i x} \mathrm{d}x = \\ &= \lambda_i \int_0^\infty \mathrm{e}^{-\sum_j \lambda_j x} \mathrm{d}x = \frac{\lambda_i}{\sum_j \lambda_j}\end{aligned} \tag{2.56}$$

Now let $\mathcal{N}_i$ be the states of the continuous-time Markov chain that can be directly reached from state i,

$$\mathcal{N}_i = \{j \mid q_{ij} > 0\}$$

and let T_{ij}, $j \in \mathcal{N}_i$, be the time the continuous-time Markov chain would need to go from state i to state j if we would allow no other transitions than that from i to j. Then the T_{ij} are exponentially distributed random variables with respective rates q_{ij}. The probability that T_{ik}, $k \in \mathcal{N}_i$, is the smallest of these random variables – and thus the next transition will be to state k – is (according to Equation 2.56) $q_{ik}/\sum_{j\in\mathcal{N}_i} q_{ij} = -q_{ik}/q_{ii}$

Therefore, the transition probabilities s_{ij} of the embedded Markov chain are

$$s_{ij} = \begin{cases} -\frac{q_{ij}}{q_{ii}} = \frac{q_{ij}}{\sum_{k\neq i} q_{ik}} & \text{for } i \neq j \\ 0 & \text{for } i = j \end{cases} \tag{2.57}$$

or in matrix form

$$\mathcal{S} = (s_{ij}) = \mathcal{I} - (\operatorname{diag}(\mathcal{Q}))^{-1} \mathcal{Q} \tag{2.58}$$

The embedded Markov chain can be used to analyse the underlying continuous-time Markov chain. For example, the stationary state probabilities π_i of the continuous-time Markov chain can be obtained by multiplying the stationary state probabilities of the embedded Markov chain ϵ_i by the sojourn time in each state:

$$\pi_i = C \cdot \epsilon_i \frac{1}{-q_{ii}} \tag{2.59}$$

where C is chosen such that $\sum_i \pi_i = 1$ holds.

3. Modelling of queueing systems with Markov chains

In general, modelling of queueing systems with Markov chains is done in three steps:

First, a Markov chain that deals with the problem under consideration is constructed. Attention should be paid that the Markov chain contains all the important properties of the queueing system. However, at the same time it should not contain too many details, because otherwise the effort for the calculation of the properties of the Markov chain is unnecessarily great.

Then the properties of the Markov chain (such as transient and stationary state probabilities) are calculated. This is done by means of the formulas shown in Chapter 2. Mathematical procedures that can be used for this task are described in [Heath 1997], [Press et al. 1992], [Stewart 1994] and [Watkins 2002]. Finally, the quantities of interest for the real system can be calculated from the properties of the Markov chain.

In this chapter, we briefly show some well-known techniques that are used to model queueing systems with Markov chains: In Section 3.1, we show how the system state is modelled, in Section 3.2, we explain how the flow time through a queueing system is calculated. In Section 3.3, we describe how to determine the time needed until a certain state is reached or until a certain event occurs for the first time.

Due to the memory-less property of Markov chains, only such processes can be modelled directly whose interevent times are exponentially distributed. How general processes can be modelled is described in Section 3.4.

Finally, in Section 3.5, we show how traffic streams can be modelled.

3.1. System state

For most problems it is necessary to model the system state. The Markov chain for the system state, we call it $\mathcal{M}_S$, contains a state for each possible state of the queueing system under consideration. The transition rates between the states of the Markov chain are the respective transition rates (e.g., arrival and service rates) in the queueing system. From the state probabilities of the states of the Markov chain $\mathcal{M}_S$ we can determine the probability for the states of the queueing system. The transient state probabilities correspond to transient processes in the queueing system, and the stationary state probabilities correspond to the steady-state of the queueing system. By means of the state probabilities, we can determine quantities such as

- number of customers in the system,
- length of the waiting queue,
- utilisation of the server,
- probability that the system is full,
- mean waiting time of the customers.

3.1.1. M/M/1/S queueing system

The Markov chain for the system state of an M/M/1/S queueing system with system size $S = 3$, arrival rate λ and service rate μ is shown in Figure 3.1. Each state of this Markov chain corresponds to one of the four possible states the queueing system can be in: State 1 / $\langle$Idle$\rangle$ corresponds to an empty system, State 2 / $\langle 1 \rangle$ corresponds to a system in which there is one customer, and so on.

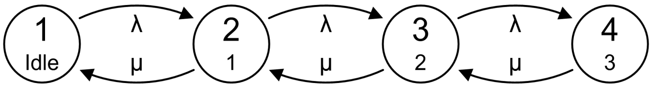


Figure 3.1.: M/M/1/S queueing system: Markov chain for the system state ($S = 3$). Meaning of the names of the states: number of customers in the system, or *Idle* if the system is empty.

The transition rate matrix of this Markov chain is

$$\mathcal{Q} = \begin{pmatrix} -\lambda & \lambda & 0 & 0 \\ \mu & -\lambda-\mu & \lambda & 0 \\ 0 & \mu & -\lambda-\mu & \lambda \\ 0 & 0 & \mu & -\mu \end{pmatrix}$$

It should be noted that in this queueing system all states of the system have a corresponding state in the Markov chain $\mathcal{M}_S$, but not all events have a corresponding transition: When the queueing system is full and a customer arrives, there is an event but no transition in the Markov chain. We call such events "silent events". The corresponding transitions – which are not contained in the Markov chain we currently use, but might be contained in another Markov chain for the queueing system under consideration – are called “hidden transitions”.

The stationary state probabilities of the Markov chain $\pi = \left(\pi_{\langle \text{Idle} \rangle}, \pi_{\langle 1 \rangle}, \pi_{\langle 2 \rangle}, \pi_{\langle 3 \rangle}\right)$ are calculated by solving the system of linear equations

$$\pi \cdot \mathcal{Q} = 0 \tag{3.1}$$

under the side condition

$$\pi_{\langle \text{Idle} \rangle} + \pi_{\langle 1 \rangle} + \pi_{\langle 2 \rangle} + \pi_{\langle 3 \rangle} = 1 \tag{3.2}$$

When we have calculated the stationary system state probabilities of the Markov chain, we can determine, among other things, the following steady-state characteristics of the queueing system:

The expected number of customers in the system is

$$\mathrm{E}(X) = \sum_{k=1}^{3} k \cdot \pi_{\langle k \rangle} \tag{3.3}$$

Figure 3.2 shows the number of customers in the system as a function of the arrival rate λ.

The probability that an arriving customer is rejected (blocking probability) is the probability that the system is full:

$$p_{\text{blocking}} = \pi_4 \tag{3.4}$$

Figure 3.3 shows the blocking probability as a function of the arrival rate λ.

The utilisation of the server ρ is the probability that it is not idle,

$$\rho = 1 - \pi_{\langle \text{Idle} \rangle} = \pi_{\langle 1 \rangle} + \pi_{\langle 2 \rangle} + \pi_{\langle 3 \rangle} \tag{3.5}$$

The transient state probabilities of the Markov chain $\pi(t) = \big(\pi_{\langle \text{Idle} \rangle}(t), \pi_{\langle 1 \rangle}(t), \pi_{\langle 2 \rangle}(t), \pi_{\langle 3 \rangle}(t)\big)$ are calculated by solving the system of first order ordinary differential equations

$$\pi'(t) = \pi(t) \cdot \mathcal{Q} \tag{3.6}$$

The value for $\pi(0)$ is chosen according to the initial state of the queueing system.

For example, if we want to investigate the transient behaviour of a queueing system that is empty at the beginning, we would have $\pi(0) = (1, 0, 0, 0)$.

The probability that at time t there are n customers in the system is

$$\mathrm{P}\left\{X(t) = n\right\} = \begin{cases} \pi_{\langle \text{Idle} \rangle}(t) & \text{for } n = 0 \\ \pi_{\langle n \rangle}(t) & \text{for } n \geq 1 \end{cases} \tag{3.7}$$

The expected number of customers in the system at time t is

$$\mathrm{E}(X)(t) = \sum_{k=1}^{3} k \cdot \pi_{\langle k \rangle}(t) \tag{3.8}$$

Some results are shown in Figures 3.4 and 3.5.

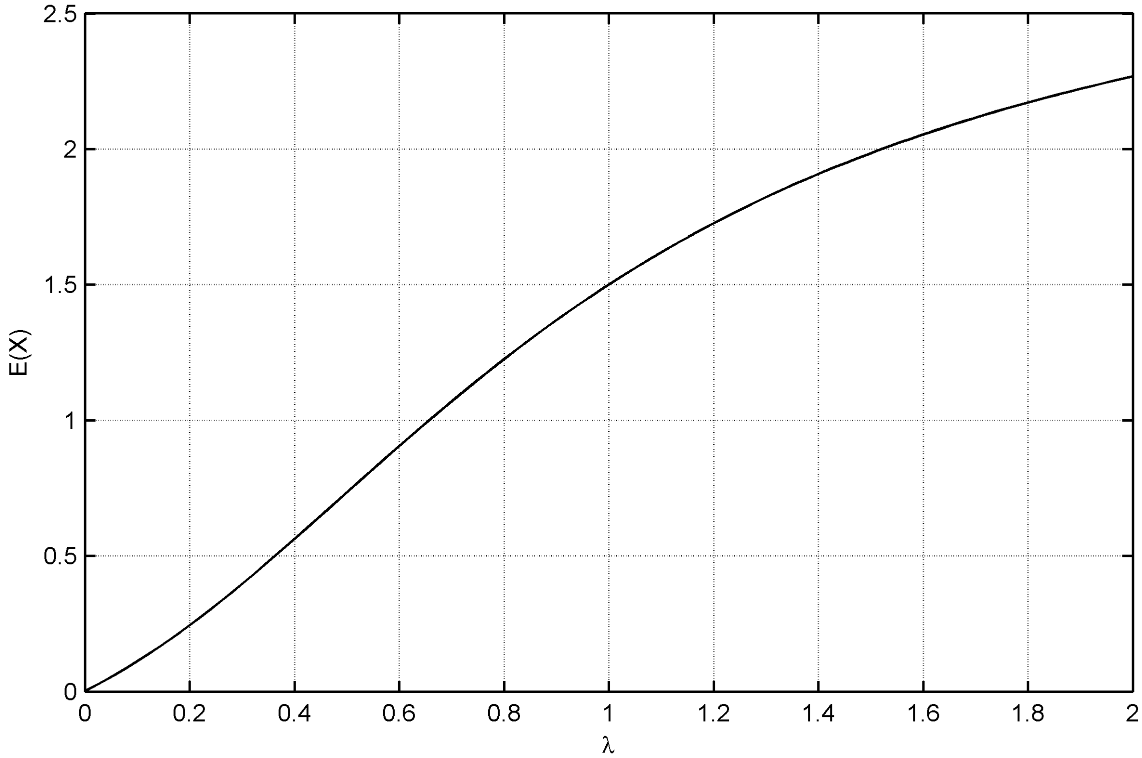


Figure 3.2.: M/M/1/S queueing system with $S = 3$, arrival rate λ and service rate $\mu = 1$: expected number of customers in the system in the steady state.

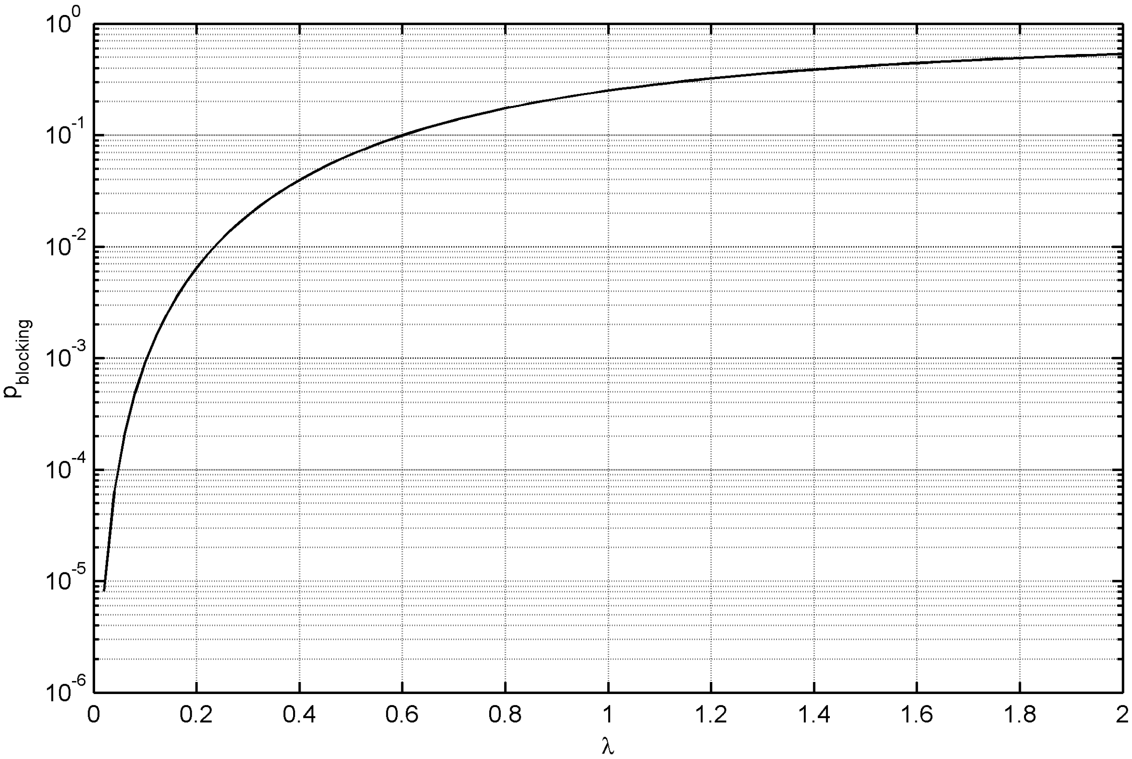


Figure 3.3.: M/M/1/S queueing system with $S = 3$, arrival rate λ and service rate $\mu = 1$: blocking probability in the steady state.

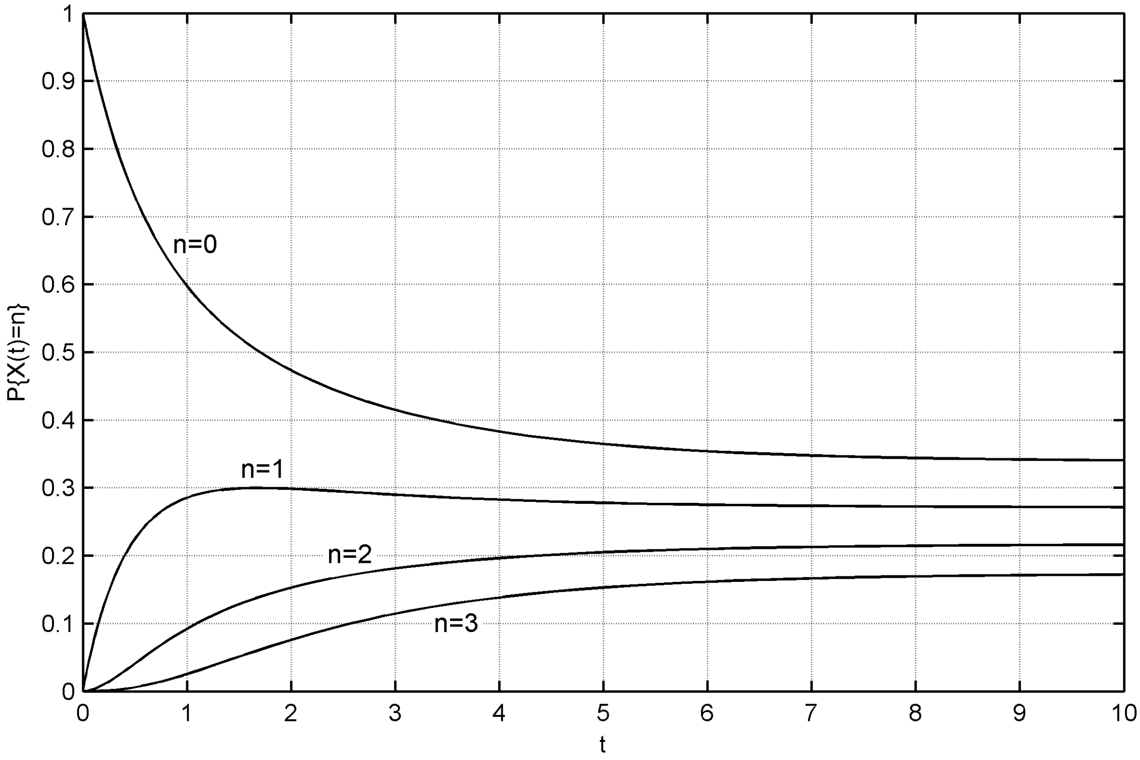


Figure 3.4.: Transient behaviour of an M/M/1/S queueing system: probability that there are n customers in the system at time t, given that the system is empty at time 0. $S = 3$, arrival rate $\lambda = 0.8$, service rate $\mu = 1$.

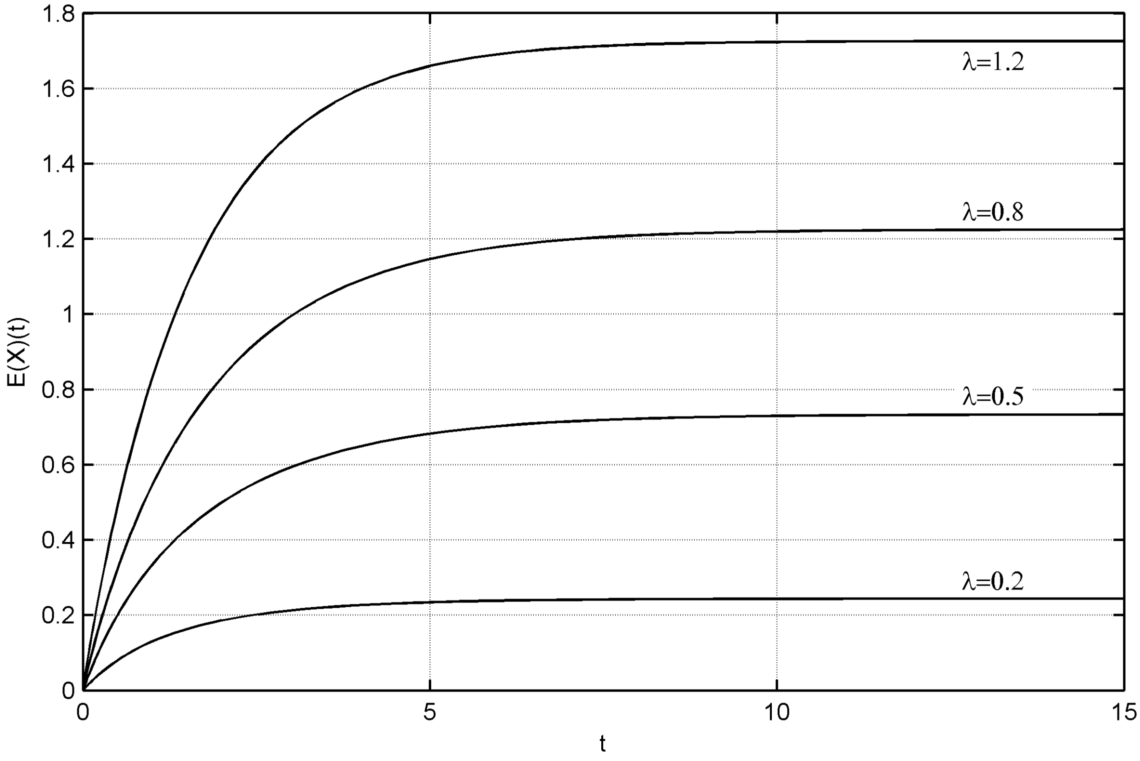


Figure 3.5.: Transient behaviour of an M/M/1/S queueing system: expected number of customers in the system at time t, given that the system is empty at time 0. $S = 3$, arrival rate λ, service rate $\mu = 1$.

Closed-form solution for the stationary state probabilities

A closed-form solution for the stationary state probabilities can be obtained by using the so-called balance equations. The balance equations state that, in steady state, the rate at which a state is left equals the rate at which it is entered:

$$\underbrace{\pi_i \sum_{j \neq i} q_{ij}}_{\text{rate at which state } i \text{ is left}} \quad = \quad \underbrace{\sum_{j \neq i} \pi_j q_{ji}}_{\text{rate at which state } i \text{ is entered}} \qquad \forall\ i \tag{3.9}$$

Therefore, in the Markov chain for the M/M/1/S queueing system we have

$$\begin{aligned}
\pi_{\langle \text{Idle} \rangle}\lambda = \pi_{\langle 1 \rangle}\mu \quad &\Rightarrow \quad \pi_{\langle 1 \rangle} = \pi_{\langle \text{Idle} \rangle}\frac{\lambda}{\mu} \\
\pi_{\langle 1 \rangle}\lambda = \pi_{\langle 2 \rangle}\mu \quad &\Rightarrow \quad \pi_{\langle 2 \rangle} = \pi_{\langle 1 \rangle}\frac{\lambda}{\mu} = \pi_{\langle \text{Idle} \rangle}\left(\frac{\lambda}{\mu}\right)^2 \\
&\dots \\
\pi_{S-1}\lambda = \pi_{\langle S \rangle}\mu \quad &\Rightarrow \quad \pi_{\langle S \rangle} = \pi_{\langle S-1 \rangle}\frac{\lambda}{\mu} = \pi_{\langle \text{Idle} \rangle}\left(\frac{\lambda}{\mu}\right)^S
\end{aligned} \tag{3.10}$$

To simplify the formulas, we change the name of state 1 from $\langle \text{Idle} \rangle$ to $\langle 0 \rangle$. So we have

$$\pi_{\langle n \rangle} = \pi_{\langle 0 \rangle}\left(\frac{\lambda}{\mu}\right)^n \qquad n = 1, \dots, S \tag{3.11}$$

Using the fact that the sum of all probabilities must equal 1, we can write

$$\begin{aligned}
1 &= \sum_{k=0}^{S} \pi_{\langle k \rangle} = \pi_{\langle 0 \rangle} + \pi_{\langle 1 \rangle} + \dots + \pi_{\langle S \rangle} = \\
&= \pi_{\langle 0 \rangle} + \pi_{\langle 0 \rangle}\frac{\lambda}{\mu} + \pi_{\langle 0 \rangle}\left(\frac{\lambda}{\mu}\right)^2 + \dots + \pi_{\langle 0 \rangle}\left(\frac{\lambda}{\mu}\right)^S = \pi_{\langle 0 \rangle}\sum_{k=0}^{S}\left(\frac{\lambda}{\mu}\right)^k = \\
&= \pi_{\langle 0 \rangle}\left(\frac{1}{1-\frac{\lambda}{\mu}} - \frac{\left(\frac{\lambda}{\mu}\right)^{S+1}}{1-\frac{\lambda}{\mu}}\right) = \pi_{\langle 0 \rangle}\left(\frac{1-\left(\frac{\lambda}{\mu}\right)^{S+1}}{1-\frac{\lambda}{\mu}}\right)
\end{aligned} \tag{3.12}$$

and

$$\pi_{\langle 0 \rangle} = \frac{1-\frac{\lambda}{\mu}}{1-\left(\frac{\lambda}{\mu}\right)^{S+1}} \tag{3.13}$$

In the special case $\lambda = \mu$ we obtain

$$\begin{aligned}
1 &= \sum_{k=0}^{S} \pi_{\langle k \rangle} = \pi_{\langle 0 \rangle} + \pi_{\langle 1 \rangle} + \dots + \pi_{\langle S \rangle} = \\
&= \pi_{\langle 0 \rangle} + \pi_{\langle 0 \rangle}1 + \pi_{\langle 0 \rangle}1^2 + \dots + \pi_{\langle 0 \rangle}1^S = \pi_{\langle 0 \rangle}(S+1)
\end{aligned} \tag{3.14}$$

$$\pi_{\langle 0 \rangle} = \frac{1}{S+1} \tag{3.15}$$

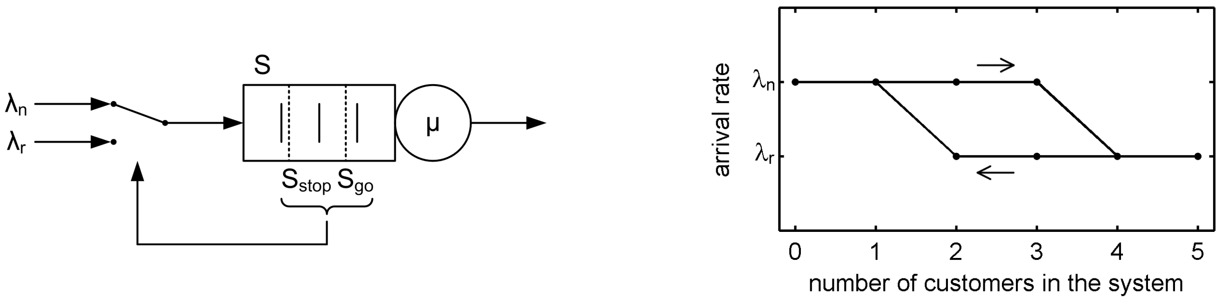
Therefore, we have

$$\pi_{\langle 0\rangle} = \begin{cases} \frac{1-\lambda/\mu}{1-(\lambda/\mu)^{S+1}} & \text{for } \lambda \neq \mu \\ \frac{1}{S+1} & \text{for } \lambda = \mu \end{cases} \tag{3.16}$$

$$\pi_{\langle n\rangle} = \pi_{\langle 0\rangle} \left(\frac{\lambda}{\mu}\right)^n \qquad n = 1, \ldots, S \tag{3.17}$$

3.1.2. M/M/1/S queueing system with controlled arrival rate

In this queueing system (Figure 3.6a), the arrival rate is controlled depending on the number of customers in the system. The normal arrival rate is λ_n. If the number of customers in the queueing system reaches S_stop, the arrival rate is reduced to λ_r, in order to decrease the probability that the system will become full. When the number of customers in the system reaches S_go, the arrival rate is switched back to its normal rate λ_n (see Figure 3.6b). The service rate is always μ.



(a)

(b)

Figure 3.6.: M/M/1/S queueing system with controlled arrival rate. (a) Overview, (b) arrival rate for $S = 5, S_\text{stop} = 4, S_\text{go} = 1$

The Markov chain for the system state of this queueing system (with $S = 5$, $S_\text{stop} = 4$ and $S_\text{go} = 1$) is shown in Figure 3.7.

The transition rate matrix of this Markov chain is

$$\mathcal{Q} = \begin{pmatrix} -\lambda_\mathrm{n} & \lambda_\mathrm{n} & 0 & 0 & 0 & 0 & 0 & 0 \\ \mu & -\lambda_\mathrm{n}-\mu & \lambda_\mathrm{n} & 0 & 0 & 0 & 0 & 0 \\ 0 & \mu & -\lambda_\mathrm{n}-\mu & \lambda_\mathrm{n} & 0 & 0 & 0 & 0 \\ 0 & 0 & \mu & -\lambda_\mathrm{n}-\mu & 0 & 0 & \lambda_\mathrm{n} & 0 \\ 0 & \mu & 0 & 0 & -\lambda_\mathrm{r}-\mu & \lambda_\mathrm{r} & 0 & 0 \\ 0 & 0 & 0 & 0 & \mu & -\lambda_\mathrm{r}-\mu & \lambda_\mathrm{r} & 0 \\ 0 & 0 & 0 & 0 & 0 & \mu & -\lambda_\mathrm{r}-\mu & \lambda_\mathrm{r} \\ 0 & 0 & 0 & 0 & 0 & 0 & \mu & -\mu \end{pmatrix}$$

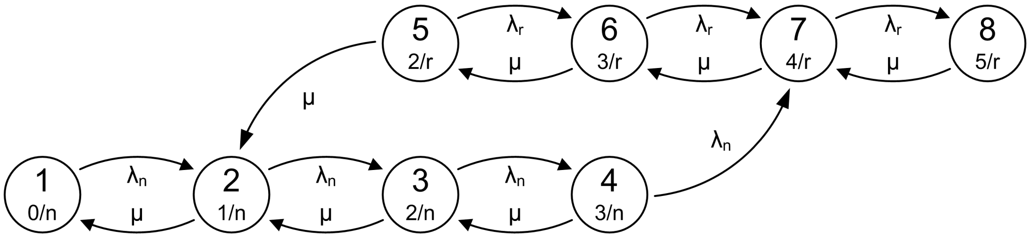


Figure 3.7.: M/M/1/S queueing system with controlled arrival rate. $S = 5, S_{\text{stop}} = 4, S_{\text{go}} = 1$. Meaning of the names of the states: number of customers in the system / "n" for normal arrival rate or "r" for reduced arrival rate.

After calculating the stationary state probabilities of this Markov chain, we can obtain some characteristics of the queueing system:

The expected number of customers in the system is

$$\mathrm{E}(X) = 1 \cdot \pi_{\langle 1/\mathrm{n}\rangle} + 2 \cdot \left(\pi_{\langle 2/\mathrm{n}\rangle} + \pi_{\langle 2/\mathrm{r}\rangle}\right) + \\ 3 \cdot \left(\pi_{\langle 3/\mathrm{n}\rangle} + \pi_{\langle 3/\mathrm{r}\rangle}\right) + 4 \cdot \pi_{\langle 4/\mathrm{r}\rangle} + 5 \cdot \pi_{\langle 5/\mathrm{r}\rangle} \tag{3.18}$$

or generally

$$\mathrm{E}(X) = \sum_{k=0}^{S_{\text{stop}}-1} k \cdot \pi_{\langle k/\mathrm{n}\rangle} + \sum_{k=S_{\text{go}}+1}^{S} k \cdot \pi_{\langle k/\mathrm{r}\rangle} \tag{3.19}$$

The probability that the arrival rate is reduced is

$$p_{\text{reduced}} = \pi_{\langle 2/\mathrm{r}\rangle} + \pi_{\langle 3/\mathrm{r}\rangle} + \pi_{\langle 4/\mathrm{r}\rangle} + \pi_{\langle 5/\mathrm{r}\rangle} \tag{3.20}$$

or generally

$$p_{\text{reduced}} = \sum_{k=S_{\text{go}}+1}^{S} \pi_{\langle k/\mathrm{r}\rangle} \tag{3.21}$$

The utilisation of the server is

$$\rho = 1 - \pi_{\langle 0/\mathrm{n}\rangle} \tag{3.22}$$

3.2. Calculation of the flow time

In general, it is not possible to determine the time Φ that a customer needs to flow through a queueing system based only on the system state. Instead, the fate of a so-called test customer is considered.[1] The test customer enters the queueing system at an arbitrary point of time and begins its flow process. The flow process ends when the test customer has again left the system. The length of the interval between entry and departure of the test customer is the flow time. Only such test customers are considered that are accepted by the queueing system when they arrive. Test customers who are rejected (for example, because the system is full) are ignored.

In most cases, the fate of the test customer depends on the system state as well as on its evolution within the service discipline of the system (in the following referred to as service process). Therefore, the Markov chain for the flow process $\mathcal{M}_\Phi$ is a combination of the Markov chain for the system state and the Markov chain that describes the service process.

In $\mathcal{M}_\Phi$ three kinds of states can be distinguished:

- States in which the flow process can start. These states can be taken immediately after the arrival of the test customer. The set of these states is denoted by $\mathcal{M}_1$.
- States in which the flow process cannot start. These states can be reached only indirectly. The set of these states is denoted by $\mathcal{M}_2$.
- States in which the test customer has left the queueing system. The set of these states is denoted by $\mathcal{H}$. These states are absorbing, because when such a state is reached, the flow process has ended.

Depending on in which state the queueing system is when the test customer arrives, and therefore in which state $\mathcal{M}_\Phi$ is when the flow process begins, the flow time will be longer or shorter.

Let $\varphi_i(\cdot)$ be the complementary cumulative distribution function of the time the Markov chain needs to go from state i to one of the absorbing states $\mathcal{H}$. This time equals the flow time given that $\mathcal{M}_\Phi$ is in state i when the flow process begins,

$$\varphi_i(\tau) = \mathrm{P}\left\{\Phi > \tau \mid \text{starting state } i\right\} \tag{3.23}$$

According to Equations 2.54 and 2.55, we calculate $\varphi_i(\cdot)$ with

$$\varphi_i(0) = \begin{cases} 1 & \text{for } i \in \mathcal{M}_1 \cup \mathcal{M}_2 \\ 0 & \text{for } i \in \mathcal{H} \end{cases} \tag{3.24}$$

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{3.25}$$

Now the cumulative distribution function of the total flow time is

$$F_\Phi(\tau) = 1 - \sum_i \sigma_i^\Phi \, \varphi_i(\tau) \tag{3.26}$$

[1] cf. [Kühn 1972]

where σ_i^Φ is the probability that the Markov chain $\mathcal{M}_\Phi$ is in state i when the flow process begins.

The mean flow time is obtained with[2]

$$\mathrm{E}\left(\Phi\right) = \int\limits_{\tau=0}^{\infty} F_\Phi^C(\tau)\,\mathrm{d}\tau \tag{3.27}$$

(where $F_\Phi^C(\cdot)$ is the complementary cumulative distribution function of Φ) or with Little's law.

Little's law states that the average number of customers in a system $\mathrm{E}(X)$ is equal to the average arrival rate λ multiplied by the average time $\mathrm{E}\left(\Phi\right)$ that a customer spends in the system:

$$\mathrm{E}(X) = \lambda \cdot \mathrm{E}\left(\Phi\right) \tag{3.28}$$

Therefore, we have

$$\mathrm{E}(\Phi) = \mathrm{E}\left(X\right) \cdot \lambda^{-1} \tag{3.29}$$

3.2.1. M/M/1/S queueing system

First, we consider a simple M/M/1/S queueing system with $S = 5$, arrival rate λ, service rate μ and service discipline FIFO (first in – first out).

In this queueing system the system state does not affect the flow process, except for the determination of the probabilities of the starting states. The evolution within the service discipline, and thus the evolution within the flow process, is determined by the number of customers that are in the queueing system when the test customer arrives (and, therefore, will be served before the test customer).

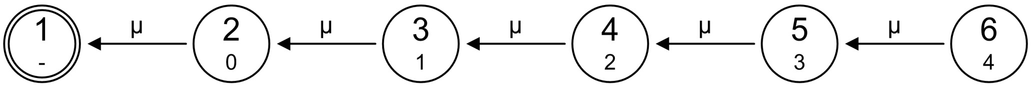


Figure 3.8.: M/M/1/S queueing system: Markov chain for the flow process (=service process) for $S = 5$. Meaning of the names of the states: number of customers which will be served before the test customer.

The Markov chain for the flow process is shown in Figure 3.8. The flow process is in the absorbing state 1 when the test customer has left the queueing system. The other states

[2] Let $f_\Phi(\cdot)$ be the probability density function of Φ. Then we have:

$$\int\limits_{x=0}^{\infty} F_\Phi^C(x)\,\mathrm{d}x = \int\limits_{x=0}^{\infty}\int\limits_{y=x}^{\infty} f_\Phi(y)\,\mathrm{d}y\,\mathrm{d}x = \int\limits_{y=0}^{\infty}\int\limits_{x=0}^{y} f_\Phi(y)\,\mathrm{d}x\,\mathrm{d}y = \int\limits_{y=0}^{\infty} f_\Phi(y) \underbrace{\int\limits_{x=0}^{y} \mathrm{d}x}_{=y}\,\mathrm{d}y = \mathrm{E}\left(\Phi\right)$$

are possible starting states for the flow process. For example, the flow process starts in state 2, if the test customer finds an empty system upon arrival. The test customer is served immediately, and at the service rate μ the flow process reaches state 1. The flow process starts in state 3, if the test customer finds a system in which there is one customer. This customer leaves the system at the service rate μ (the flow process is then in state 2), whereupon the test customer is served. The service of the test customer finishes at rate μ, afterwards the test customer leaves the system (the flow process is then in state 1).

The probabilities for the different starting states depend on the system state probabilities $\pi(0)$ at the moment when the test customer arrives:

$$\begin{aligned}
\mathrm{P}\{\text{starting state 2}\} &= \sigma_2^{\Phi} = \pi_{\langle \text{Idle} \rangle}(0)\,\frac{1}{1-\pi_{\langle 5 \rangle}(0)}\\
\mathrm{P}\{\text{starting state 3}\} &= \sigma_3^{\Phi} = \pi_{\langle 1 \rangle}(0)\,\frac{1}{1-\pi_{\langle 5 \rangle}(0)}\\
\mathrm{P}\{\text{starting state 4}\} &= \sigma_4^{\Phi} = \pi_{\langle 2 \rangle}(0)\,\frac{1}{1-\pi_{\langle 5 \rangle}(0)}\\
\mathrm{P}\{\text{starting state 5}\} &= \sigma_5^{\Phi} = \pi_{\langle 3 \rangle}(0)\,\frac{1}{1-\pi_{\langle 5 \rangle}(0)}\\
\mathrm{P}\{\text{starting state 6}\} &= \sigma_6^{\Phi} = \pi_{\langle 4 \rangle}(0)\,\frac{1}{1-\pi_{\langle 5 \rangle}(0)}
\end{aligned}$$

The factor $1/\left(1-\pi_{\langle 5 \rangle}(0)\right)$ occurs because the test customer can begin its flow process only if the queueing system is not full when it arrives.[3]

If we look at the Markov chain in Figure 3.8, we see that the flow time of a test customer that finds n customers in the system upon arrival is the convolution of $n+1$ exponentially distributed service times.

Therefore, we can express the flow time through an M/M/1/S queueing system as weighted sum of Erlang distributions (see Section 3.4.1):

$$\begin{aligned}
\Phi \sim\ & \sigma_2^F \operatorname{Exp}(\mu) + \\
& \sigma_3^F \operatorname{HypoExp}(\mu,\mu) + \\
& \sigma_4^F \operatorname{HypoExp}(\mu,\mu,\mu) + \\
& \sigma_5^F \operatorname{HypoExp}(\mu,\mu,\mu,\mu) + \\
& \sigma_6^F \operatorname{HypoExp}(\mu,\mu,\mu,\mu,\mu)
\end{aligned} \tag{3.30}$$

If we want to use Equation 3.25 to calculate the flow time, we would solve

$$\begin{pmatrix} \varphi_1'(t)\\ \varphi_2'(t)\\ \varphi_3'(t)\\ \varphi_4'(t)\\ \varphi_5'(t)\\ \varphi_6'(t) \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0\\ \mu & -\mu & 0 & 0 & 0 & 0\\ 0 & \mu & -\mu & 0 & 0 & 0\\ 0 & 0 & \mu & -\mu & 0 & 0\\ 0 & 0 & 0 & \mu & -\mu & 0\\ 0 & 0 & 0 & 0 & \mu & -\mu \end{pmatrix} \cdot \begin{pmatrix} \varphi_1(t)\\ \varphi_2(t)\\ \varphi_3(t)\\ \varphi_4(t)\\ \varphi_5(t)\\ \varphi_6(t) \end{pmatrix} \tag{3.31}$$

[3] $\mathrm{P}\{\text{system in state } \langle i \rangle\ (0 \le i \le 4) \mid \text{system not full}\} = \frac{\mathrm{P}\{\text{system in state } \langle i \rangle \wedge \text{system not full}\}}{\mathrm{P}\{\text{system not full}\}} = \frac{\pi_i}{1-\pi_{\langle 5 \rangle}}$

or

$$\begin{aligned}
&\varphi_1'(t) = 0\\
&\varphi_2'(t) = \varphi_1(t)\,\mu - \varphi_2(t)\,\mu = -\mu\,\varphi_2(t)\\
&\varphi_3'(t) = \varphi_2(t)\,\mu - \varphi_3(t)\,\mu = \mu\,(\varphi_2(t) - \varphi_3(t))\\
&\varphi_4'(t) = \varphi_3(t)\,\mu - \varphi_4(t)\,\mu = \mu\,(\varphi_3(t) - \varphi_4(t))\\
&\varphi_5'(t) = \varphi_4(t)\,\mu - \varphi_5(t)\,\mu = \mu\,(\varphi_4(t) - \varphi_5(t))\\
&\varphi_6'(t) = \varphi_5(t)\,\mu - \varphi_6(t)\,\mu = \mu\,(\varphi_5(t) - \varphi_6(t))
\end{aligned} \tag{3.32}$$

with the initial conditions

$$\varphi_i(0) = \begin{cases} 0 & \text{for } i = 1\\ 1 & \text{otherwise} \end{cases} \tag{3.33}$$

The solution is

$$\begin{aligned}
&\varphi_1(t) = 0\\
&\varphi_2(t) = \mathrm{e}^{-\mu\,t}\\
&\varphi_3(t) = \mathrm{e}^{-\mu\,t}\,(1 + \mu\,t)\\
&\varphi_4(t) = \mathrm{e}^{-\mu\,t}\left(1 + \mu\,t + \frac{\mu^2\,t^2}{2}\right)\\
&\varphi_5(t) = \mathrm{e}^{-\mu\,t}\left(1 + \mu\,t + \frac{\mu^2\,t^2}{2} + \frac{\mu^3\,t^3}{6}\right)\\
&\varphi_6(t) = \mathrm{e}^{-\mu\,t}\left(1 + \mu\,t + \frac{\mu^2\,t^2}{2} + \frac{\mu^3\,t^3}{6} + \frac{\mu^4\,t^4}{24}\right)
\end{aligned} \tag{3.34}$$

As expected, for $k \geq 2$, $\varphi_k(\cdot)$ is the complementary cumulative distribution function of a $(k-1)$-stage Erlang distribution (Equation 3.55).

The cumulative distribution function of the total flow time is

$$\begin{aligned}
F_\Phi(\tau) &= 1 - \sum_i \varphi_i(\tau)\sigma_i^\Phi\\
&= 1 - \big(\varphi_2(\tau)\,\pi_{\langle\text{Idle}\rangle}(0) + \varphi_3(\tau)\,\pi_{\langle 1\rangle}(0) + \varphi_4(\tau)\,\pi_{\langle 2\rangle}(0) +\\
&\qquad \varphi_5(\tau)\,\pi_{\langle 3\rangle}(0) + \varphi_6(\tau)\,\pi_{\langle 4\rangle}(0)\big) \,/\, \big(1 - \pi_{\langle 5\rangle}(0)\big)
\end{aligned} \tag{3.35}$$

3.2.2. M/M/1/S queueing system with controlled service rate

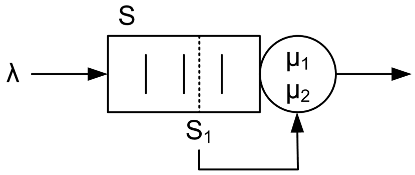


Figure 3.9.: M/M/1/S queueing system with controlled service rate.

In this queueing system (Figure 3.9), the service rate depends on the number of customers in the system: The normal service rate is μ_1. If the length of the queue exceeds S_1, the service rate is increased to μ_2. This is done in order to decrease the probability that the system will become full. The service rate is decreased to μ_1, if the length of the queue reaches or falls below S_1. The arrival rate is constant λ, and the service discipline is FIFO.

Now the fate of the test customer is also influenced by following customers: If after the test customer has arrived, there are so many arrivals of further customers that there are always more than S_1 customers in the system, then the test customer and all customers that are served before the test customer are served at the increased service rate μ_2, so that the flow time decreases. On the other hand, if for a long time no customers arrive at the queueing system after the test customer has, the test customer and up to $S_1 - 1$ customers that are served before the test customer are served only at the normal service rate μ_1, which leads to a higher flow time.

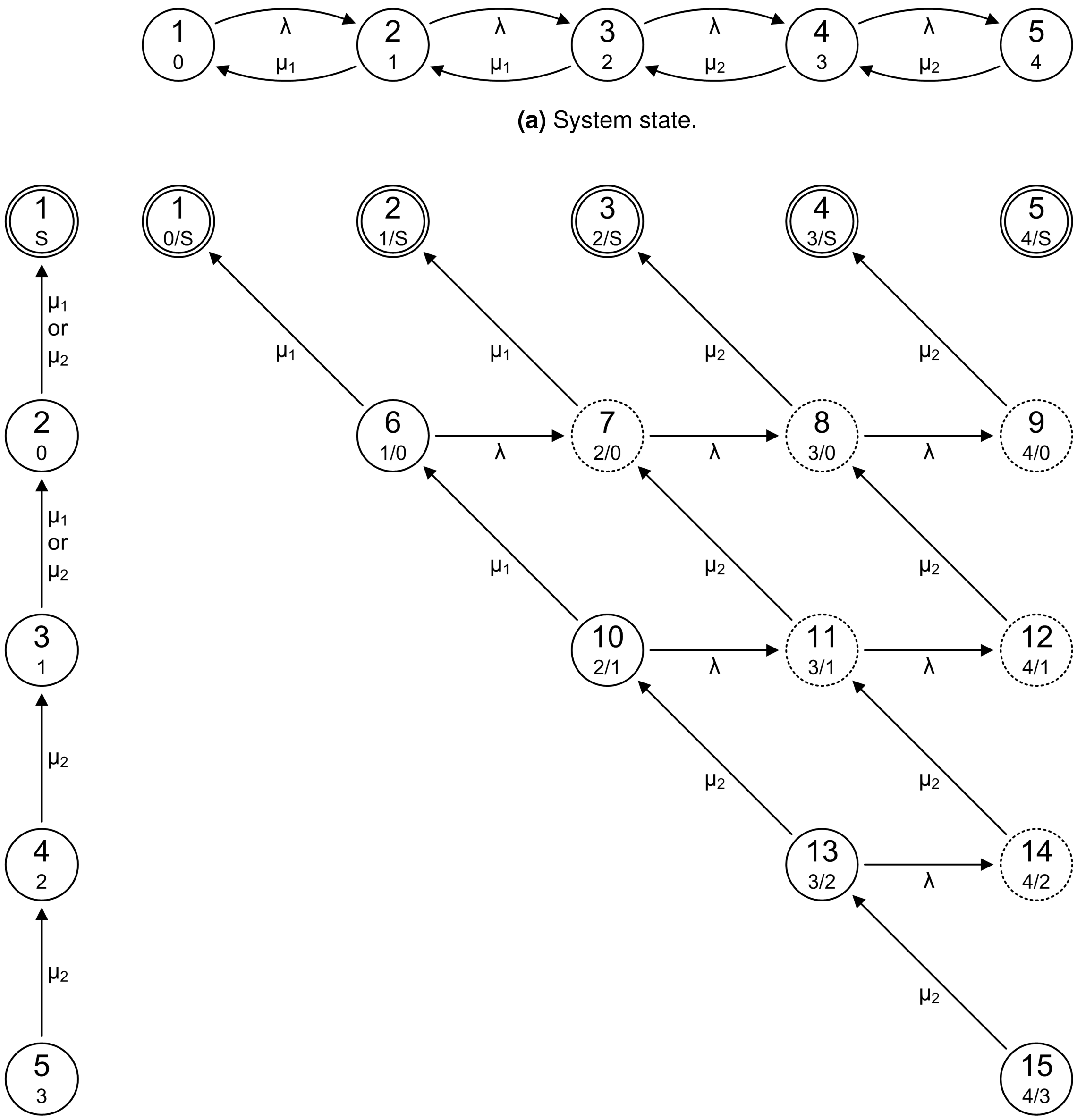
The Markov chain for the flow process is shown in Figure 3.10c. It can be constructed from the Markov chain for the system state (Figure 3.10a) and the Markov chain for the service process (Figure 3.10b).

The probabilities for the different starting states depend on the system state probabilities $\pi(0)$ at the moment of the arrival of the test customer:

$$
\begin{aligned}
\mathrm{P}\left\{\text{starting state } \langle 1/0\rangle\right\} &= \sigma^{\Phi}_{\langle 1/0\rangle} = \pi_{\langle 0\rangle}(0)\,\frac{1}{1-\pi_{\langle 4\rangle}(0)}\\
\mathrm{P}\left\{\text{starting state } \langle 2/1\rangle\right\} &= \sigma^{\Phi}_{\langle 2/1\rangle} = \pi_{\langle 1\rangle}(0)\,\frac{1}{1-\pi_{\langle 4\rangle}(0)}\\
\mathrm{P}\left\{\text{starting state } \langle 3/2\rangle\right\} &= \sigma^{\Phi}_{\langle 3/2\rangle} = \pi_{\langle 2\rangle}(0)\,\frac{1}{1-\pi_{\langle 4\rangle}(0)}\\
\mathrm{P}\left\{\text{starting state } \langle 4/3\rangle\right\} &= \sigma^{\Phi}_{\langle 4/3\rangle} = \pi_{\langle 3\rangle}(0)\,\frac{1}{1-\pi_{\langle 4\rangle}(0)}
\end{aligned}
$$

After calculating $\varphi_i(\cdot)$ with

$$
\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \qquad \varphi_i(0) = \begin{cases} 0 & \text{for } 1 \le i \le 5 \\ 1 & \text{for } 6 \le i \le 15 \end{cases} \tag{3.36}
$$



(a) System state.

(b) Service process.

(c) Flow process. Meaning of the names of the states: number of customers in the system / number of customers which are served before the test customer or “S”, if the test customer has been served. States which can be reached only indirectly (set $\mathcal{M}_2$) are painted with dashed lines, absorbing states (set $\mathcal{H}$) are painted with double lines.

Figure 3.10.: M/M/1/S queueing system with controlled service rate: Markov chains for system state, service process and flow process. ($S = 4, S_1 = 2$.)

we can calculate the cumulative distribution function of the total flow time with

$$\begin{aligned} F_\Phi(\tau) &= 1 - \sum_i \varphi_i(\tau)\ \sigma_i^\Phi \\ &= 1 - \big(\varphi_{\langle 1/0\rangle}(\tau)\ \pi_{\langle 0\rangle}(0) + \varphi_{\langle 2/1\rangle}(\tau)\ \pi_{\langle 1\rangle}(0) + \\ &\qquad \varphi_{\langle 3/2\rangle}(\tau)\ \pi_{\langle 2\rangle}(0) + \varphi_{\langle 4/3\rangle}(\tau)\ \pi_{\langle 3\rangle}(0)\big)\ /\ \big(1 - \pi_{\langle 4\rangle}(0)\big) \end{aligned} \tag{3.37}$$

We could also express the flow time as a weighted sum of hypoexponential random variables. However, since we have to consider all possible paths from the starting states to the absorbing states, this approach is not suitable here:

$$\begin{aligned} \Phi \sim\ & \sigma^\Phi_{\langle 1/0\rangle}\ \frac{\mu_1}{\lambda+\mu_1}\ \mathrm{Exp}\left(\mu_1\right) + \\ & \sigma^\Phi_{\langle 1/0\rangle}\ \frac{\lambda}{\lambda+\mu_1}\ \frac{\mu_1}{\lambda+\mu_1}\ \mathrm{HypoExp}\left(\lambda, \mu_1\right) + \\ & \sigma^\Phi_{\langle 1/0\rangle}\ \frac{\lambda}{\lambda+\mu_1}\ \frac{\lambda}{\lambda+\mu_1}\ \frac{\mu_2}{\lambda+\mu_2}\ \mathrm{HypoExp}\left(\lambda, \lambda, \mu_2\right) + \\ & \sigma^\Phi_{\langle 1/0\rangle}\ \frac{\lambda}{\lambda+\mu_1}\ \frac{\lambda}{\lambda+\mu_1}\ \frac{\lambda}{\lambda+\mu_2}\ \mathrm{HypoExp}\left(\lambda, \lambda, \lambda, \mu_2\right) + \\ & \sigma^\Phi_{\langle 2/1\rangle}\ \frac{\mu_1}{\lambda+\mu_1}\ \frac{\mu_1}{\lambda+\mu_1}\ \mathrm{HypoExp}\left(\mu_1, \mu_1\right) + \\ & \sigma^\Phi_{\langle 2/1\rangle}\ \frac{\mu_1}{\lambda+\mu_1}\ \frac{\lambda}{\lambda+\mu_1}\ \frac{\mu_1}{\lambda+\mu_1}\ \mathrm{HypoExp}\left(\mu_1, \lambda, \mu_1\right) + \\ & \ldots \end{aligned} \tag{3.38}$$

Figure 3.11 shows the flow time depending on the system state at the moment when the test customer arrives.

The impact of following customers on the flow time of the test customer can be seen in Figure 3.12: we consider an M/M/1/S queueing system with controlled service rate in which the arrival rate changes immediately after the test customer has arrived. If the arrival rate decreases, the flow time of the test customer increases, and vice versa.

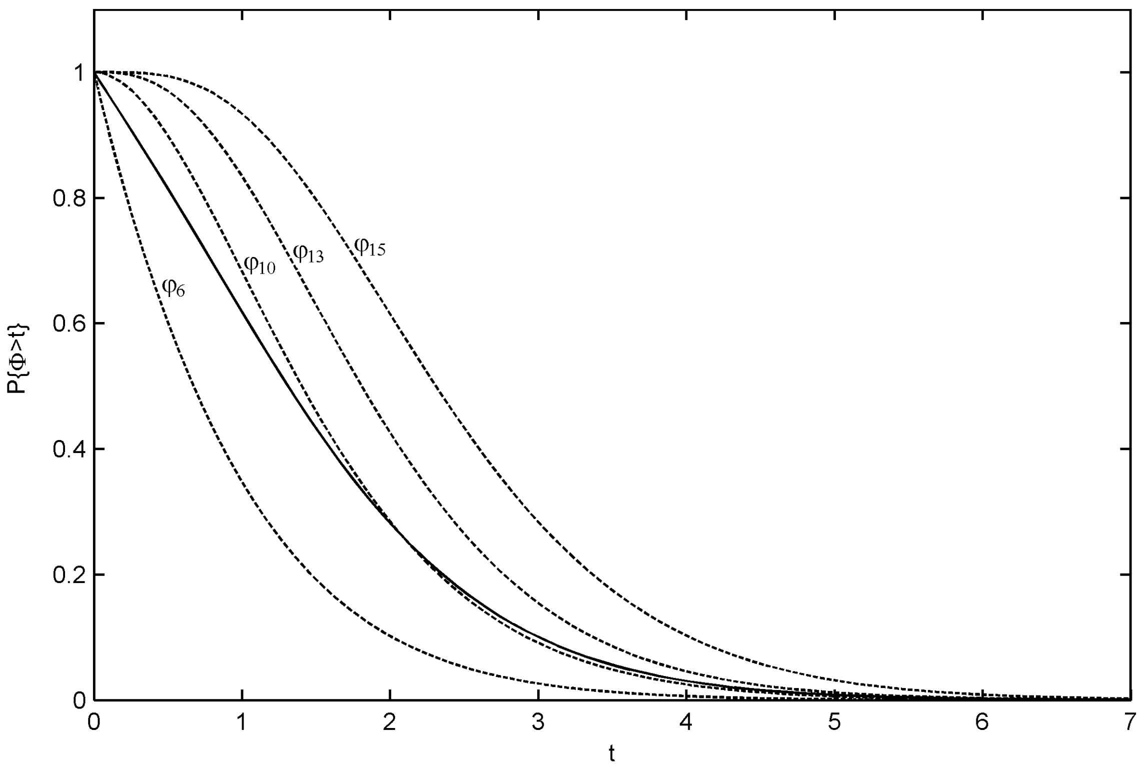


Figure 3.11.: M/M/1/S queueing system with controlled service rate: flow time depending on the system state when the test customer arrives (dashed lines), total flow time (solid line). $S = 3, \lambda = 0.8,\ \mu_1 = 1,\ \mu_2 = 2.$

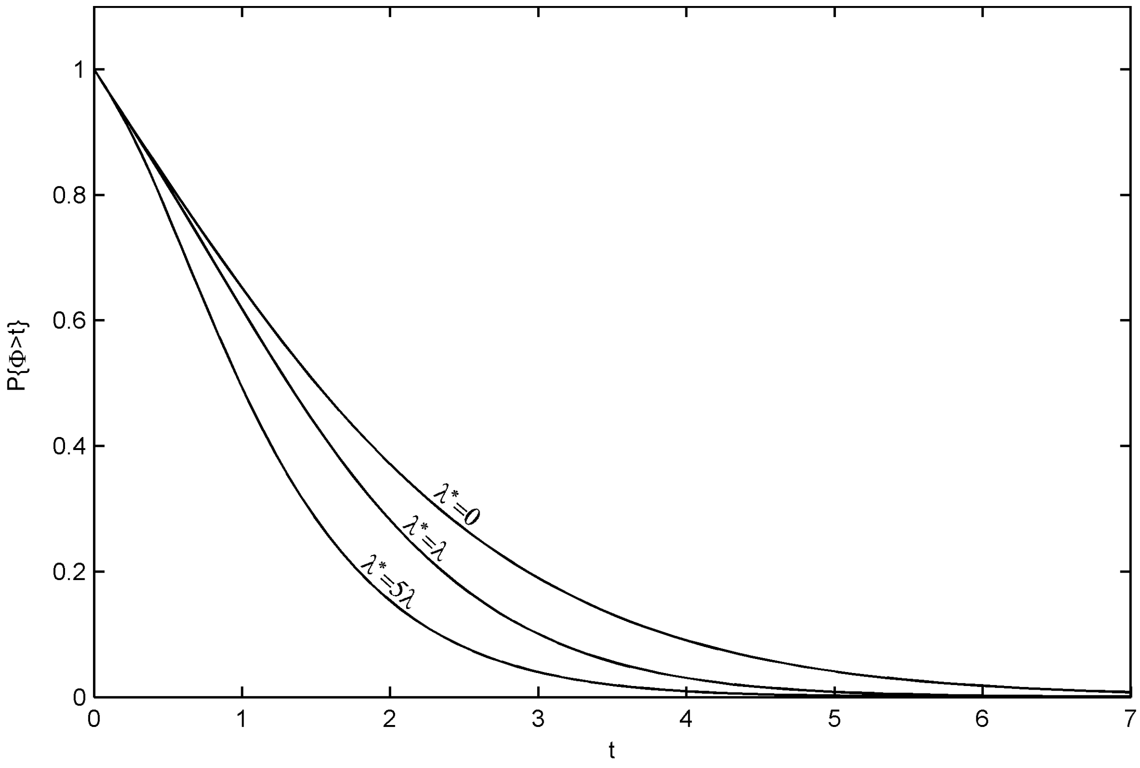


Figure 3.12.: M/M/1/S queueing system with controlled service rate: impact of following customers on the flow time Φ of the test customer. The arrival rate is $\lambda = 0.8$ until the test customer arrives. Then it changes to λ^*. The service rates are $\mu_1 = 1,\ \mu_2 = 2.\ S = 3.$

3.3. Calculation of the time until a certain state is reached

A problem similar to the calculation of the flow time through a queueing system is that we know the queueing system is in state j now and we want to know how long it will take until it is, for the first time, in some state $h \in \mathcal{H}$, or until, for the first time, a certain event takes place.

This time can be calculated as follows: If we are interested in the time until a state of the set $\mathcal{H}$ is reached, we remove all transitions originating in a state $h \in \mathcal{H}$, that is, we make the states in $\mathcal{H}$ absorbing. If we are interested in the first occurrence of an event, we add a new state $\langle \mathrm{R} \rangle$ – which will constitute the set $\mathcal{H}$ – to the Markov chain for the system state and redirect all transitions that correspond to the event under consideration to this new state.

Then we calculate the complementary cumulative distribution function $\varphi_i(\cdot)$ of the time that the Markov chain needs to go from state i to a state in $\mathcal{H}$ with

$$\varphi_i(0) = \begin{cases} 1 & i \notin \mathcal{H} \\ 0 & i \in \mathcal{H} \end{cases} \tag{3.39}$$

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{3.40}$$

The time T we are looking for is

$$\mathrm{P}\left(T > t\right) = \varphi_j(t) \tag{3.41}$$

The expected value of T is

$$\mathrm{E}(T) = \int\limits_{t=0}^{\infty} \varphi_j(t)\,\mathrm{d}t \tag{3.42}$$

3.3.1. M/M/1/S queueing system

We are interested in how long an M/M/1/S queueing system ($S = 3$, arrival rate λ, service rate μ), which is empty at time $t = 0$, is able to serve all arriving customers. That is, we want to know how long it takes until, for the first time, an arriving customer is blocked because the system is full.

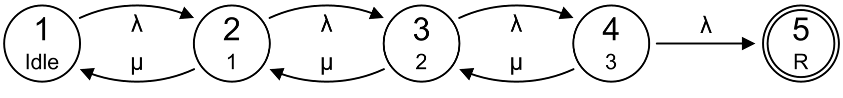


Figure 3.13.: M/M/1/S queueing system: Markov chain for the system state, extended by the state $\langle \mathrm{R} \rangle$, which is taken, when for the first time a customer is rejected.

The needed Markov chain is shown in Figure 3.13. It is the Markov chain for the system state of the M/M/1/S queueing system, extended by the state $\langle \mathrm{R} \rangle$. This state is taken when the system is full and a customer arrives.

Now we set

$$\varphi_i(0) = \begin{cases} 1 & i \in \{1,2,3,4\} \\ 0 & i = 5 \end{cases} \tag{3.43}$$

and calculate

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{3.44}$$

We start with an empty system (state 1), so the solution is

$$\mathrm{P}\left\{\text{no rejection until time } t\right\} = \varphi_1(t) \tag{3.45}$$

The expected time $\mathrm{E}(T)$ to the first rejection is

$$\mathrm{E}(T) = \int\limits_{t=0}^{\infty} \varphi_1(t)\ \mathrm{d}t \tag{3.46}$$

Some results are shown in Figure 3.14.

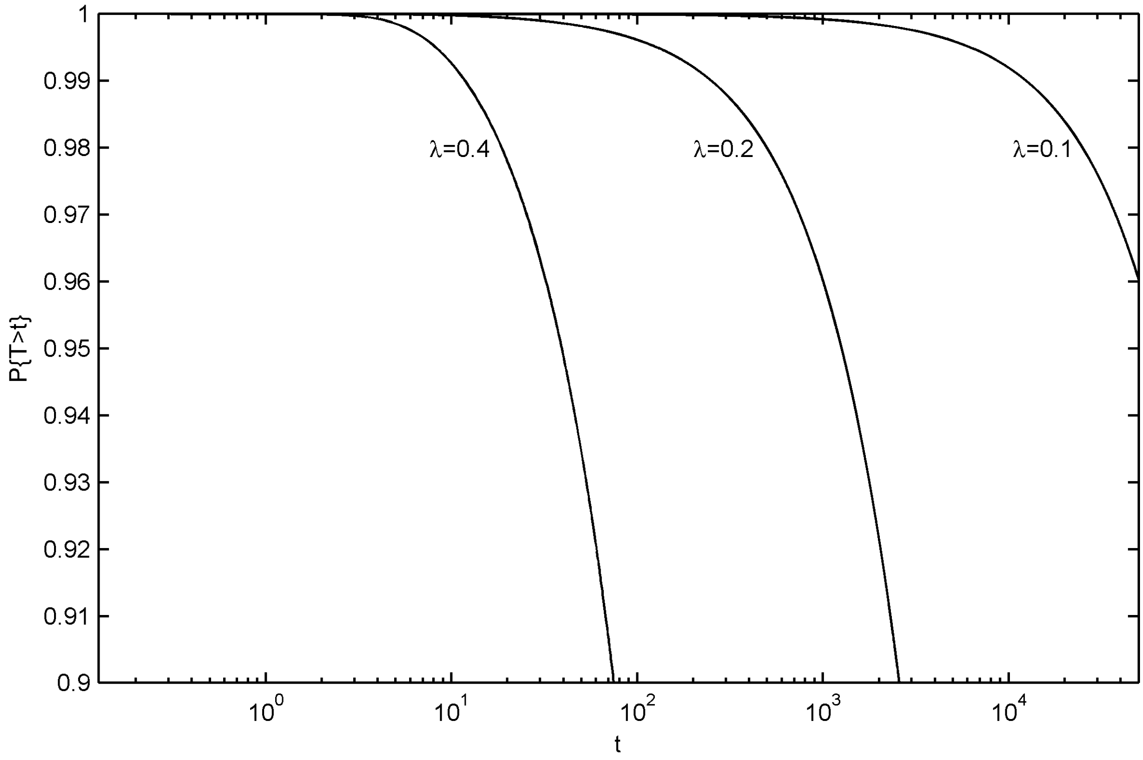


Figure 3.14.: M/M/1/S queueing system: probability that at the first rejection of an arriving customer takes place after time t. $S = 5, \mu = 1$.

3.4. Modelling general processes

In Markov chains the time between two consecutive transitions (that is, the sojourn time in a state) is exponentially distributed. Hence, in Markov chains only such processes can be modelled directly, in which the interevent times are likewise exponentially distributed (so-called Poisson processes).

Nevertheless, the interevent times of general processes can often be approximated by phase-type distributions. Phase-type distributions are probability distributions that originate from a combination of several exponential distributions and therefore can be generated in a Markov chain.

The approximation of the actual interevent times can be done with arbitrary accuracy. However, in general a better approximation requires more exponential distributions, which increases the size of the resulting Markov chains.

In this section, some important phase-type distributions are presented. Moreover, we show algorithms with which the parameters of phase-type distributions can be determined so that they can be used to approximate a given distribution.

3.4.1. Hypoexponential distribution

Let $X_1, \dots, X_k$, be independent exponential random variables with respective rates $\mu_1, \dots, \mu_k$. The random variable

$$Y = \sum_{j=1}^{k} X_j \tag{3.47}$$

is said to be a hypoexponential random variable with parameters $\mu_1, \dots, \mu_k$,

$$Y \sim \mathrm{HypoExp}(\mu_1, \dots, \mu_k) \tag{3.48}$$

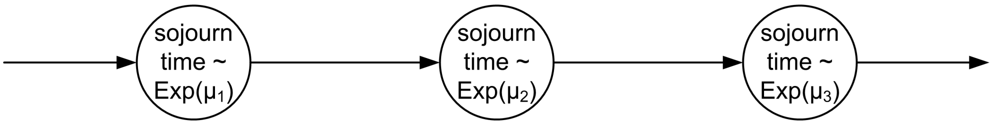


Figure 3.15.: Hypoexponential distribution (3 stages).

Hypoexponential random variables arise when a customer passes through several stages with exponentially distributed sojourn times, as shown in Figure 3.15. The flow time through such a network has a hypoexponential distribution.

A hypoexponential random variable X with parameters $\mu_1, \dots, \mu_k$ (where we assume that $\mu_i \neq \mu_j$ for $i \neq j$) has the probability density function

$$f_X(t) = \sum_{i=1}^{k} \left(\prod_{j=1, j\neq i}^{k} \frac{\mu_j}{\mu_j - \mu_i} \right) \mu_i \, \mathrm{e}^{-\mu_i t}, \quad t \geq 0 \tag{3.49}$$

and the cumulative distribution function

$$F_X(t) = \sum_{i=1}^{k} \left(\prod_{j=1, j\neq i}^{k} \frac{\mu_j}{\mu_j - \mu_i} \right) \left(1 - e^{-\mu_i t}\right), \quad t \geq 0 \tag{3.50}$$

Mean and variance are

$$\mathrm{E}(X) = \sum_{i=1}^{k} \frac{1}{\mu_i} \tag{3.51}$$

$$\mathrm{Var}(X) = \sum_{i=1}^{k} \frac{1}{\mu_i^2} \tag{3.52}$$

The coefficient of variation of hypoexponential random variables is ≤ 1:

$$c_X = \frac{\sqrt{\mathrm{Var}(X)}}{\mathrm{E}(X)} = \sqrt{\frac{\sum\limits_{i=1}^{k} \frac{1}{\mu_i^2}}{\sum\limits_{i=1}^{k} \frac{1}{\mu_i^2} + 2 \sum\limits_{i<j} \frac{1}{\mu_i \mu_j}}} \leq 1 \tag{3.53}$$

Figure 3.16 shows the probability density function of two hypoexponential distributions.

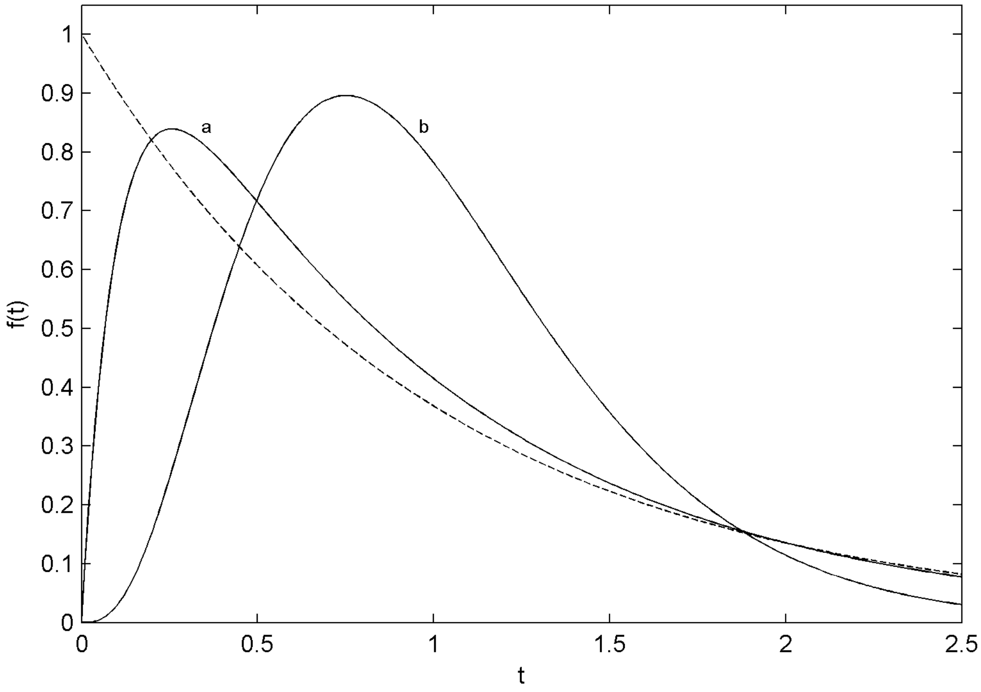


Figure 3.16.: Probability density functions of two hypoexponential random variables. (a) $\mu_1 = 1.12$, $\mu_2 = 9.41$ $(c_X = 0.9)$, (b) $\mu_1 = \mu_2 = \mu_3 = \mu_4 = 4$ $(c_X = 0.5)$. Dashed lines: exponential random variable. All three random variables have mean 1.

Erlang distribution

An important special case of the hypoexponential distribution is the Erlang distribution. An Erlang distribution is a hypoexponential distribution with $\mu_1 = \mu_2 = \cdots = \mu_k$.

An Erlang random variable has the probability density function

$$f_X(t) = \frac{\mu_1^k}{(k-1)!}\, t^{k-1}\, e^{-\mu_1 t}, \quad t \geq 0 \tag{3.54}$$

and the cumulative distribution function

$$F_X(t) = 1 - e^{-\mu_1 t} \sum_{j=0}^{k-1} \frac{(\mu_1 t)^j}{j!}, \quad t \geq 0 \tag{3.55}$$

Mean and variance are

$$\mathrm{E}(X) = \frac{k}{\mu_1} \tag{3.56}$$

$$\mathrm{Var}(X) = \frac{k}{\mu_1^2} \tag{3.57}$$

The coefficient of variation is

$$c_X = \frac{1}{\sqrt{k}} \leq 1 \tag{3.58}$$

As shown in Figure 3.17, with a growing number of stages the Erlang distribution approaches a normal distribution, whereby the coefficient of variation becomes smaller and smaller. Hence, Erlang distributions with a large number of stages can be used as an approximation of deterministic distributions. However, because of the huge number of required stages this should be done only in special cases.

Modelling hypoexponentially distributed transition times in Markov chains

Assume we have a system with states A, B and C (Figure 3.18a). The system spends an exponentially distributed time (parameter λ) in state A. Then it goes to state B, in which it spends a hypoexponentially distributed time (parameters μ_1, μ_2, μ_3), before it goes to state C. We model such a situation by thinking of state B as a network as shown in Figure 3.15. That is, we represent state B by three Markov chain states $\langle B/1\rangle, \langle B/2\rangle$ and $\langle B/3\rangle$ (Figure 3.18b). The transition of the system into state B corresponds to the transition by which the network is entered. The transition of the system out of state B corresponds to the transition by which the network is left.

In Figure 3.19, it is shown how the system state of a GI/M/1/S queueing system with hypoexponentially distributed interarrival times (a Hypo/M/1/S queueing system) is modelled.

A state of the Markov chain is defined by the number of customers in the queueing system and the current stage of the arrival process. Before an arrival really takes place,

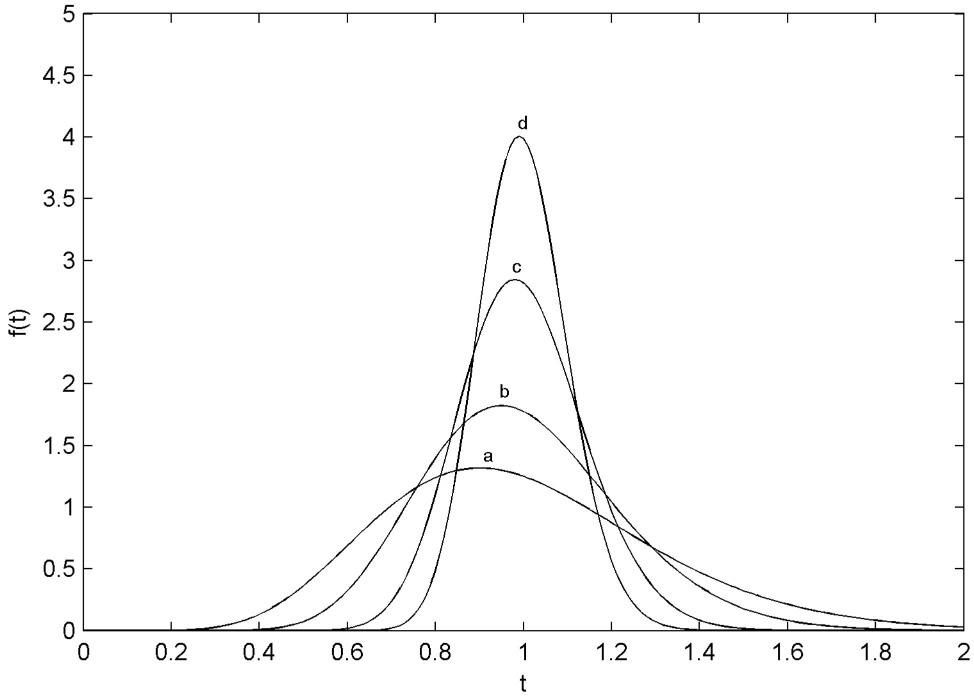


Figure 3.17.: Probability density functions of some Erlang random variables. (a) 10 stages, $\mu_1 = 10$ $(c_X = 0.32)$, (b) 20 stages, $\mu_1 = 20$ $(c_X = 0.22)$, (c) 50 stages, $\mu_1 = 50$ $(c_X = 0.14)$, (d) 100 stages, $\mu_1 = 100$ $(c_X = 0.1)$. All four random variables have mean 1.

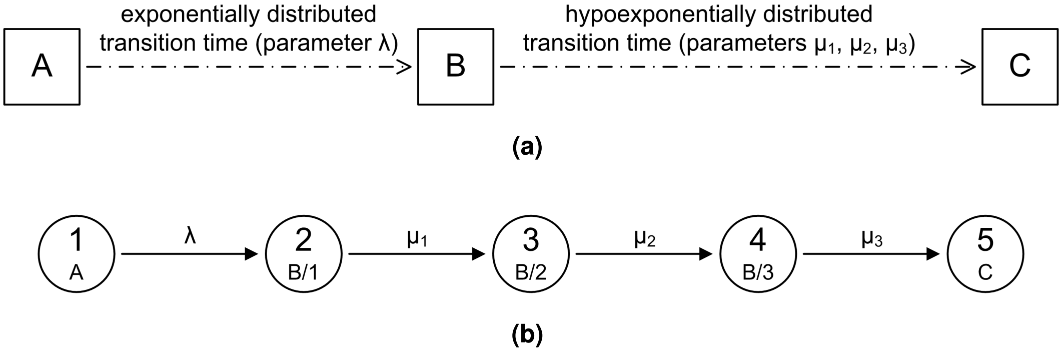

exponentially distributed transition time (parameter λ)
hypoexponentially distributed transition time (parameters μ1, μ2, μ3)
A
B
C
(a)
1
A
λ
2
B/1
μ1
3
B/2
μ2
4
B/3
μ3
5
C
(b)

Figure 3.18.: Modelling a hypoexponentially distributed transition time in a Markov chain.

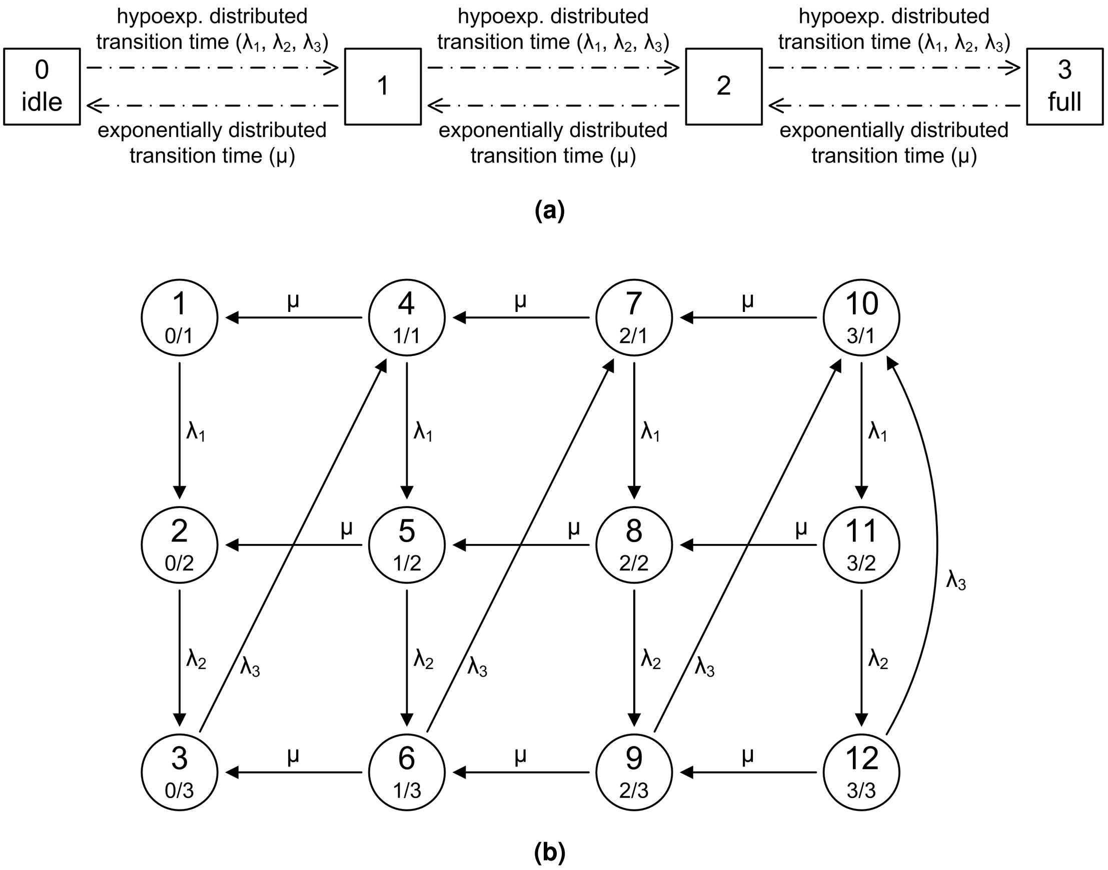


Figure 3.19.: Modelling the system state of a Hypo/M/1/S queueing system. (a) State transition diagram, (b) Markov chain. Meaning of the names of the states: number of customers in the system / state of the hypoexponential distribution of the arrival process.

several states of the Markov chain (namely those that represent the stages of the hypoexponential distribution) have to be passed. Services take place in one step and can occur – independent of the arrival process – always when there is a customer in the system.

A state of the queueing system is represented by several states of the Markov chain. For example, the state *Idle* of the queueing system is represented by the states $\langle 0/1\rangle, \langle 0/2\rangle$ and $\langle 0/3\rangle$ of the Markov chain. To calculate the probability of a state of the queueing system, the state probabilities of all associated states of the Markov chain must be added.

For example, we have for the probability p_{idle} that the system is idle

$$p_{\text{idle}} = \sum_{p=1}^{3} \pi_{\langle 0/p\rangle} \tag{3.59}$$

and for the number of customers in the system X

$$\mathrm{E}(X) = \sum_{n=0}^{3}\sum_{p=1}^{3} n\pi_{\langle n/p\rangle} \tag{3.60}$$

Figure 3.20 shows the Markov chain for the system state of an M/G/1/S queueing system with hypoexponentially distributed service times (an M/Hypo/1/S queueing system).

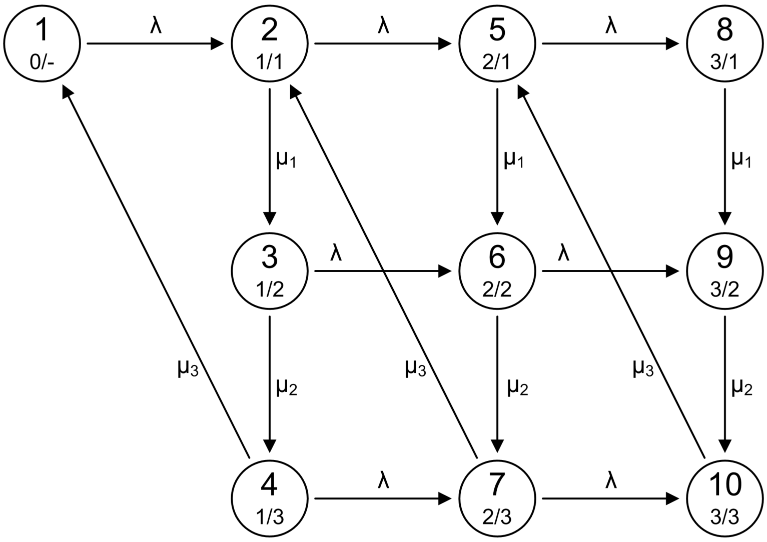


Figure 3.20.: Markov chain for the system state of an M/Hypo/1/S queueing system. Meaning of the names of the states: number of customers in the system / state of the hypoexponential distribution of the service process.

3.4.2. Hyperexponential distribution

Let $X_1, \ldots, X_k$ be independent exponential random variables with respective rates $\mu_1, \ldots, \mu_k$, and let D be a discrete random variable with $\mathrm{P}\{D = d\} = \alpha_d$, where $\sum_{d=1}^{k} \alpha_d = 1$ holds. The random variable

$$Y = X_D \tag{3.61}$$

is said to be a hyperexponential random variable with parameters $\mu_1, \alpha_1, \ldots, \mu_k, \alpha_k$,

$$Y \sim \mathrm{HyperExp}(\mu_1, \alpha_1, \ldots, \mu_k, \alpha_k) \tag{3.62}$$

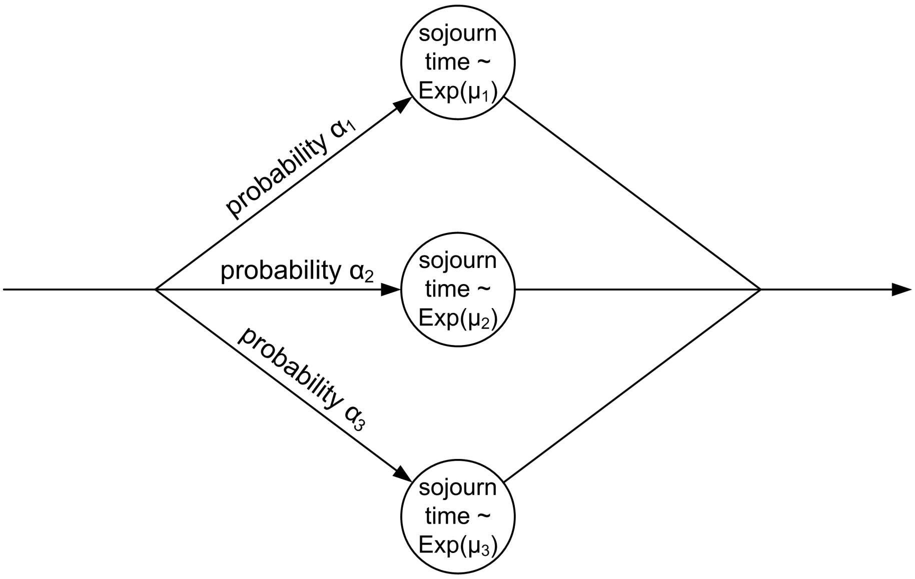


Figure 3.21.: Hyperexponential distribution (3 stages).

Hyperexponential random variables arise when a customer passes through a randomly chosen stage of a group of in parallel switched stages with exponentially distributed sojourn times, as shown in Figure 3.21. The flow time through such a network has a hyperexponential distribution.

A hyperexponential random variable with the parameters $\mu_1, \alpha_1, \ldots, \mu_k, \alpha_k$ has the probability density function

$$f_X(t) = \sum_{j=1}^{k} \alpha_j \; \mu_j \; e^{-\mu_j t}, \quad t \geq 0 \tag{3.63}$$

and the cumulative distribution function

$$F_X(t) = \sum_{j=1}^{k} \alpha_j \left(1 - e^{-\mu_j t}\right), \quad t \geq 0 \tag{3.64}$$

Mean and variance are

$$\mathrm{E}(X) = \sum_{j=1}^{k} \frac{\alpha_j}{\mu_j} \tag{3.65}$$

$$\mathrm{Var}(X) = 2\sum_{j=1}^{k} \frac{\alpha_j}{\mu_j^2} - \mathrm{E}(X)^2 \tag{3.66}$$

The coefficient of variation of hyperexponential distributions is ≥ 1:

$$c_X = \sqrt{2\,\frac{1}{\mathrm{E}(X)^2}\sum_{j=1}^{k} \frac{\alpha_j}{\mu_j^2} - 1} \geq 1 \tag{3.67}$$

In Figure 3.22, the probability density function of two hyperexponential random variables is shown.

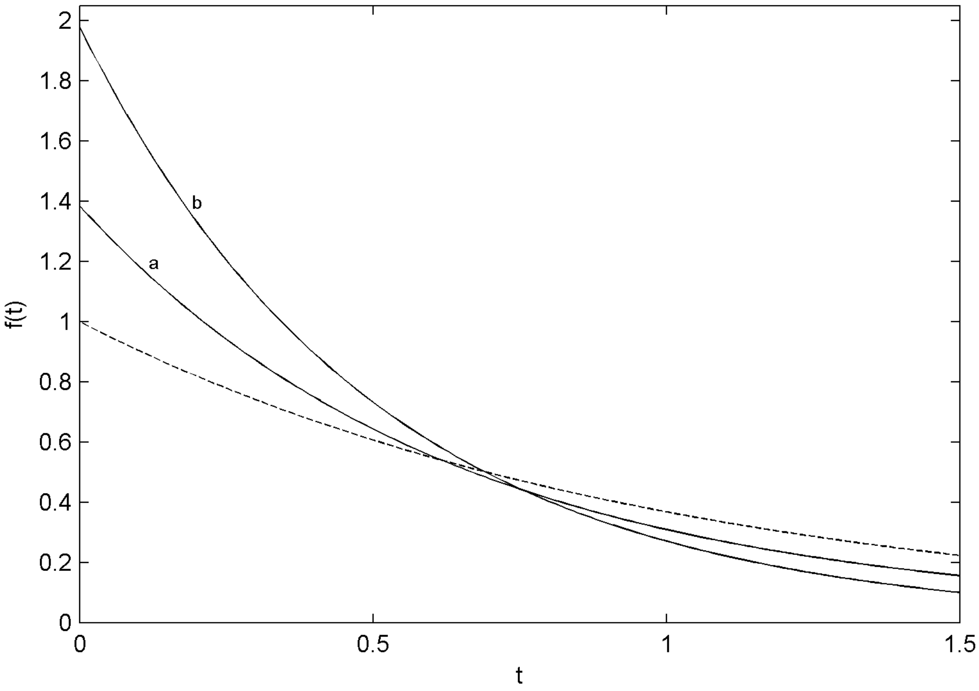


Figure 3.22.: Probability density functions of two hyperexponential random variables. (a) $\mu_1 = 0.38$, $\mu_2 = 1.62$, $\alpha_1 = 0.19$, $\alpha_2 = 0.81$ $(c_X = 1.5)$. (b) $\mu_1 = 0.01$, $\mu_2 = 1.99$, $\alpha_1 = 0.005$, $\alpha_2 = 0.995$ $(c_X = 10)$. Dashed line: exponential random variable. All three random variables have mean 1.

Modelling hyperexponentially distributed transition times in Markov chains

Assume we have a system with states A, B and C (Figure 3.23a). The system spends an exponentially distributed time (parameter λ) in state A. Then it goes to state B, in which it spends a hyperexponentially distributed time (parameters $\mu_1, \alpha_1, \mu_2, \alpha_2, \mu_3, \alpha_3$),

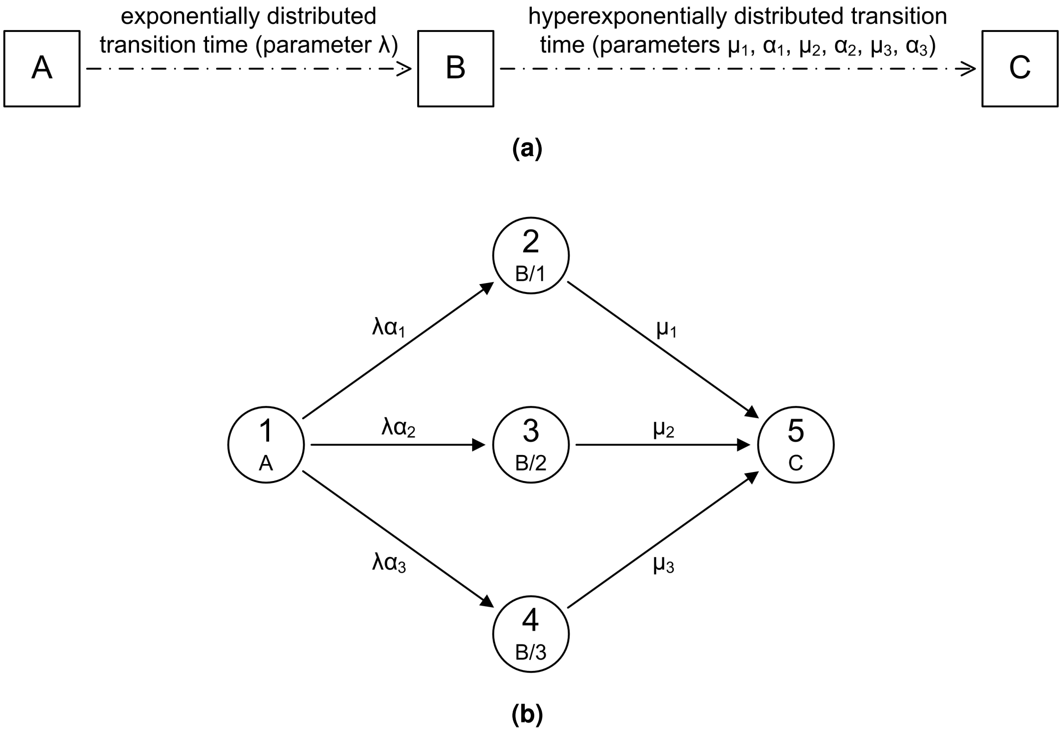

exponentially distributed transition time (parameter λ)
hyperexponentially distributed transition time (parameters μ1, α1, μ2, α2, μ3, α3)
A
B
C
(a)
1 A
2 B/1
3 B/2
4 B/3
5 C
λα1
λα2
λα3
μ1
μ2
μ3
(b)

Figure 3.23.: Modelling a hyperexponentially distributed transition time in a Markov chain.

before it goes to state C. We model such a situation by thinking of state B as a network as shown in Figure 3.21. That is, we represent state B by three Markov chain states $\langle B/1\rangle, \langle B/2\rangle$ and $\langle B/3\rangle$ (Figure 3.23b). The transition of the system into state B corresponds to the transitions by which the network is entered. The transition of the system out of state B corresponds to the transitions by which the network is left.

Figure 3.24 shows how the system state of a GI/M/1/S queueing system with hyperexponentially distributed interarrival times (a Hyper/M/1/S queueing system) is modelled.

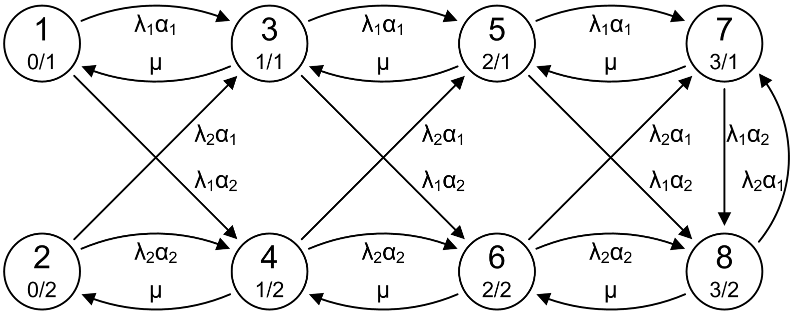


Figure 3.24.: Markov chain for the system state of a Hyper/M/1/S queueing system. Meaning of the names of the states: number of customers in the system / state of the hyperexponential distribution of the arrival process.

A state of the Markov chain is defined by the number of customers in the queueing system and the current stage of the arrival process. Arrivals take place at different rates, depending on the stage of the arrival process. After each arrival, the rate for the next arrival is chosen. Services take place in one step and can occur – independent of the arrival process – always when there is a customer in the system.

A state of the queueing system is represented by several states of the Markov chain. For example, the state *Idle* of the queueing system is represented by the states $\langle 0/1 \rangle$ and $\langle 0/2 \rangle$ of the Markov chain. To calculate the probability of a state of the queueing system, the state probabilities of all associated states of the Markov chain must be added.

For example, we have for the probability p_{idle} that the system is idle

$$p_{\text{idle}} = \pi_{\langle 0/1 \rangle} + \pi_{\langle 0/2 \rangle} \tag{3.68}$$

and for the number of customers in the system X

$$\mathrm{E}(X) = \sum_{n=0}^{3} n \left(\pi_{\langle n/1 \rangle} + \pi_{\langle n/2 \rangle} \right) \tag{3.69}$$

Figure 3.25 shows the Markov chain for the system state of an M/G/1/S queueing system with hyperexponentially distributed service times (an M/Hyper/1/S queueing system).

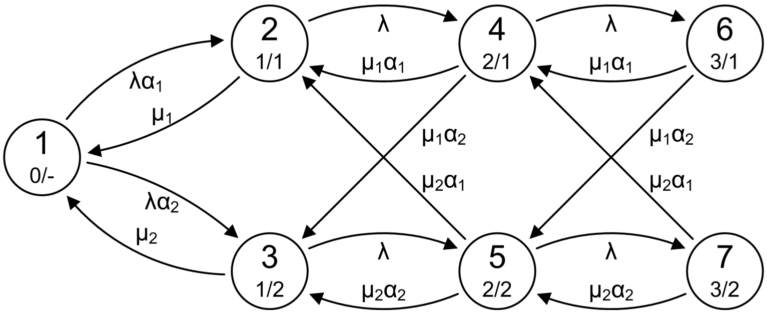


Figure 3.25.: Markov chain for the system state of an M/Hyper/1/S queueing system. Meaning of the names of the states: number of customers in the system / state of the hyperexponential distribution of the service process.

3.4.3. Coxian distribution

Let $X_1, \ldots, X_k$ be independent exponential random variables with respective rates $\mu_1, \ldots, \mu_k$, and let D be a discrete random variable with $\mathrm{P}\{D = d\} = P_d$, where $\sum_{d=1}^{k} P_d = 1$ holds. The random variable

$$Y = \sum_{j=1}^{D} X_j \tag{3.70}$$

is said to be a Coxian random variable with parameters $\mu_1, \alpha_1, \ldots, \mu_{k-1}, \alpha_{k-1}, \mu_k$,

$$Y \sim \mathrm{Cox}(\mu_1, \alpha_1, \ldots, \mu_{k-1}, \alpha_{k-1}, \mu_k) \tag{3.71}$$

where

$$P_d = (1 - \alpha_d) \sum_{i=1}^{d-1} \alpha_i \tag{3.72}$$

Coxian random variables arise when a customer passes through a sequence of stages with exponentially distributed sojourn times, whereby after each stage i it enters the next stage with probability α_i or leaves the network (with probability $1 - \alpha_i$). The flow time through such a network has a Coxian distribution.

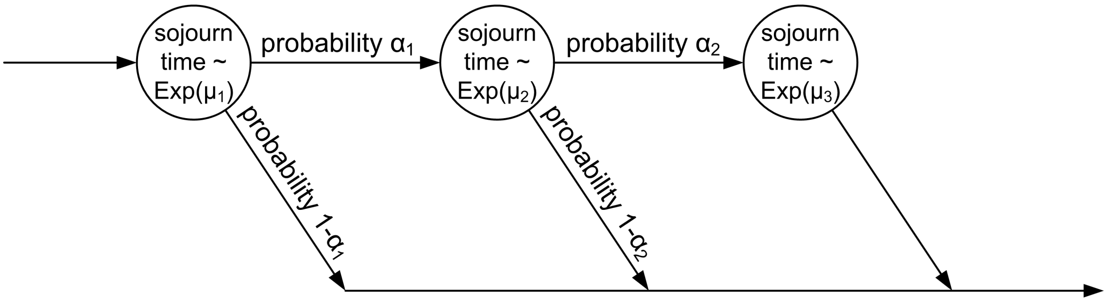


Figure 3.26.: Coxian distribution (3 stages).

For a 2-stage Coxian distribution we have

$$\mathrm{E}(X) = \frac{1}{\mu_1} + \frac{\alpha_1}{\mu_2} \tag{3.73}$$

$$c_X^2 = 1 - \frac{2\alpha_1\mu_1\left(\mu_2 - \mu_1(1 - \alpha_1)\right)}{(\mu_2 + \alpha_1\mu_1)^2} \tag{3.74}$$

With Coxian distributions, many probability distributions can be approximated. Figure 3.27 shows two examples.

In general, Markov modelling is done best with Coxian distributions. The reason is that with the same structure of the Markov chain it is possible to model interevent times with coefficients of variation smaller than, equal to, and greater than 1. If hypoexponential, hyperexponential, and exponential distributions are used for modelling, one

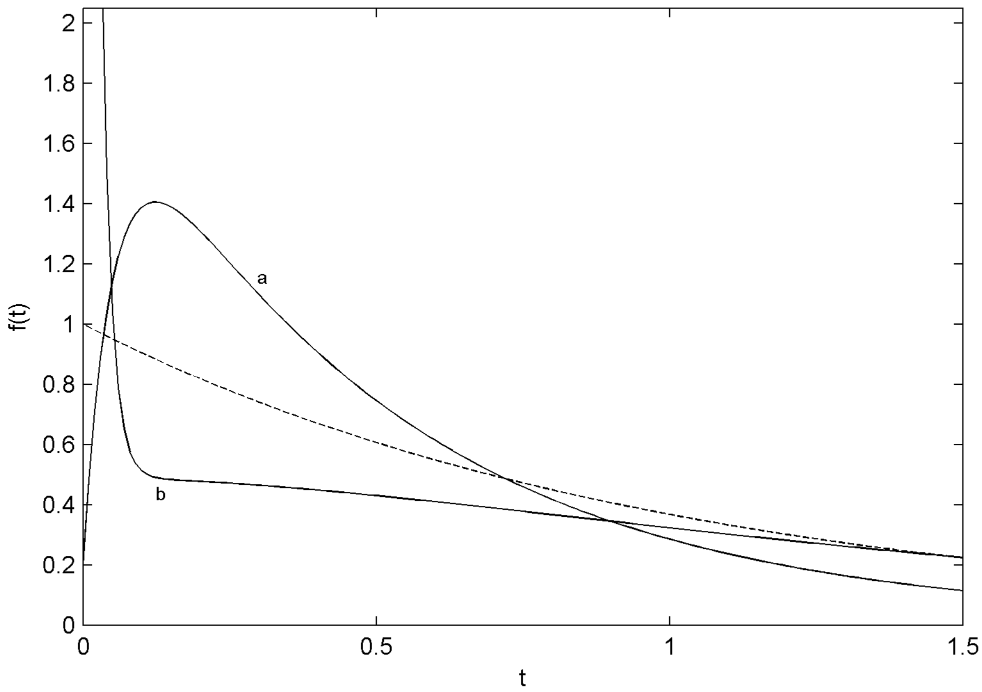


Figure 3.27.: Probability density functions of two Coxian random variables. (a) $\mu_1 = 1.99$, $\mu_2 = 19.9$, $\mu_3 = 0.199$, $\alpha_1 = 0.9$, $\alpha_2 = 0.1$. (b) $\mu_1 = 61$, $\mu_2 = 1.22$, $\mu_3 = 1.22$, $\alpha_1 = 0.8$, $\alpha_2 = 0.5$. Dashed line: exponential random variable. All three random variables have mean 1.

needs several different Markov chains to cover all possible combinations of coefficients of variation.[4]

Because a Coxian distribution can also be an exponential distribution,[5] it is often possible to test the correctness of Markov chains containing Coxian distributions: We set the parameters of the Coxian distribution to values that result in an exponential distribution and compare the achieved results with those of a model where the same distribution is assumed to be exponential (e.g., we could test the Markov chain for a Cox/M/1/S queueing system by comparing it to an M/M/1/S queueing system). For such queueing systems, the Markov chains are easier, and sometimes even closed-form solutions exist. With hypoexponential and hyperexponential distributions this procedure is not possible – these distributions always have a coefficient of variation that is strictly greater and smaller than 1, respectively.

Further information on the Coxian distribution can be found in [Augustin 1982].

[4] If we want to model a GI/G/1/S queueing system with arbitrary coefficients of variation for both arrival and service process, we would need $|\{\mathrm{Hypo}, \mathrm{M}, \mathrm{Hyper}\} \times \{\mathrm{Hypo}, \mathrm{M}, \mathrm{Hyper}\}| = 9$ different Markov chains.

[5] This can be achieved by setting $\mu_k = r$ and $\mu_i = \mu_k/(1 - \alpha_i), i = 1 \ldots k - 1$, where r is the desired rate of the exponential distribution and k is the number of stages of the Coxian distribution.

Coxian* distribution

Let $X_1, \ldots, X_k$ be independent exponential random variables with respective rates $\mu_1, \ldots, \mu_k$, and let D be a discrete random variable with $\mathrm{P}\{D = d\} = P_d$, where $\sum_{d=0}^{k} P_d = 1$ holds. We will call the random variable

$$Y = \sum_{j=0}^{D} X_j \tag{3.75}$$

a Coxian* random variable with parameters $p, \mu_1, \alpha_1, \ldots, \mu_{k-1}, \alpha_{k-1}, \mu_k$,

$$Y \sim \text{Cox*}(p, \mu_1, \alpha_1, \ldots, \mu_{k-1}, \alpha_{k-1}, \mu_k) \tag{3.76}$$

where

$$P_0 = 1 - p \tag{3.77}$$

$$P_d = p(1 - \alpha_d) \sum_{i=1}^{d-1} \alpha_i \quad d = 1 \ldots k \tag{3.78}$$

Coxian* random variables arise when a customer traverses through a network shown in Figure 3.26 with probability p and bypasses the network with probability $1 - p$. The flow time through such a network (Figure 3.28) has a Coxian* distribution.

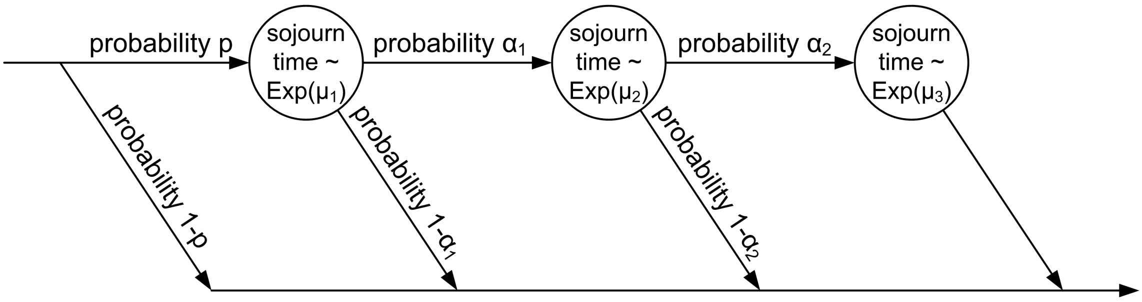


Figure 3.28.: Coxian distribution with bypass (3 stages).

Coxian* random variables equal 0 with a positive probability. This means that if they are used to describe interevent times, batches of events can occur.

It is possible to use Coxian* distributions with Markov modelling, but since we have to take batches of events into account, the resulting Markov chains are much more complex than if we use, for example, a normal Coxian distribution (see Figures 3.32 and 3.33).

Although in this work we do not consider arrival streams or service processes where batches of events can occur, we describe how some of the techniques that we present can be used in combination with Coxian* distributions. The reason is that one of the most important algorithms to find a phase type distribution with given moments (see Section 3.4.4) is based on this distribution.

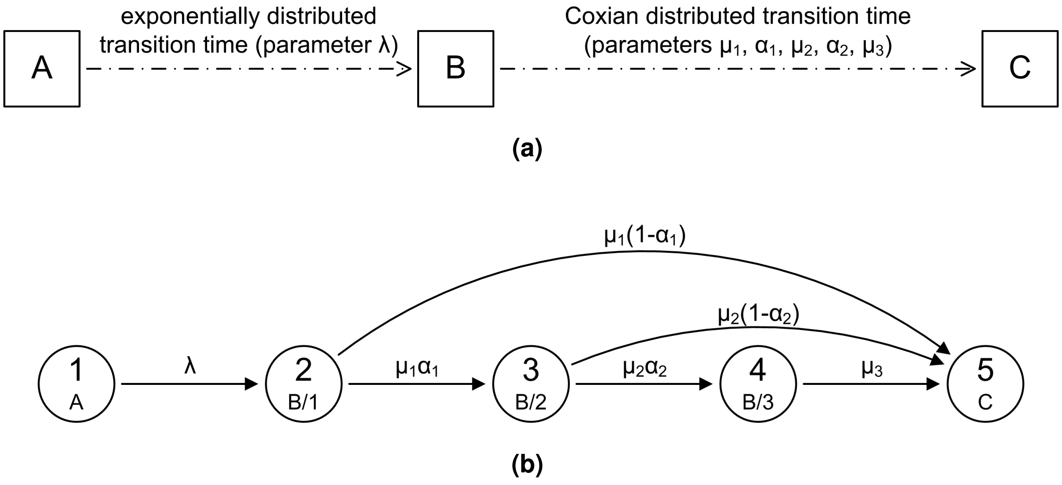

exponentially distributed
transition time (parameter λ)
Coxian distributed transition time
(parameters μ1, α1, μ2, α2, μ3)
A
B
C
(a)
μ1(1-α1)
μ2(1-α2)
λ
μ1α1
μ2α2
μ3
1 A
2 B/1
3 B/2
4 B/3
5 C
(b)

Figure 3.29.: Modelling a Coxian distributed transition time in a Markov chain.

Modelling Coxian distributed transition times in Markov chains

Assume we have a system with states A, B and C (Figure 3.18a). The system spends an exponentially distributed time (parameter λ) in state A. Then it goes to state B, in which it spends a Coxian distributed time (parameters $\mu_1, \alpha_1, \mu_2, \alpha_2, \mu_3$), before it goes to state C. We model such a situation by thinking of state B as a network as shown in Figure 3.26. That is, we represent state B by three Markov chain states $\langle B/1\rangle, \langle B/2\rangle$ and $\langle B/3\rangle$ (Figure 3.29b). The transition of the system into state B corresponds to the transition by which the network is entered. The transition of the system out of state B corresponds to the transitions by which the network is left.

Figure 3.30 shows how the system state of a GI/M/1/S queueing system with Coxian distributed interarrival times (a Cox/M/1/S queueing system) is modelled.

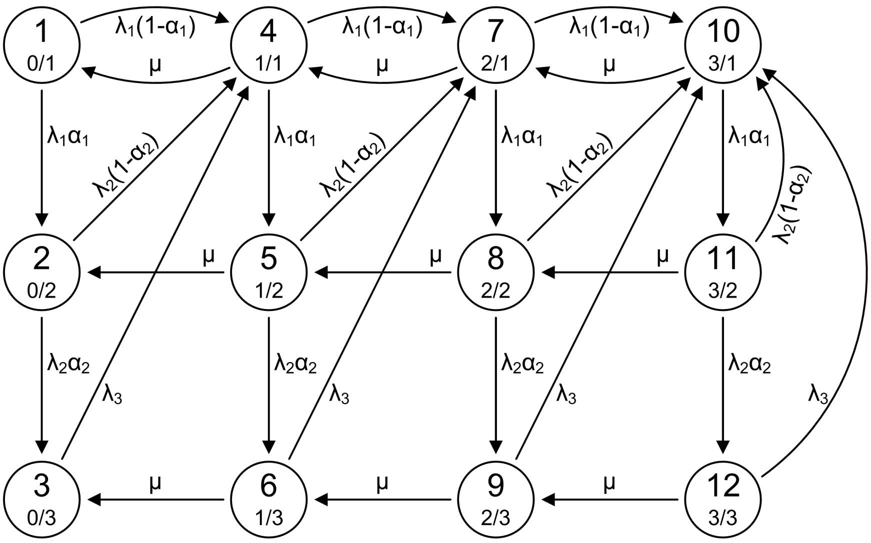


Figure 3.30.: Markov chain for the system state of a Cox/M/1/S queueing system. Meaning of the names of the states: number of customers in the system / state of the Coxian distribution of the arrival process.

A state of the Markov chain is defined by the number of customers in the queueing system and the current stage of the arrival process. Before an arrival really takes place, one or more states of the Markov chain have to be passed. Services take place in one step and can occur – independent of the arrival process – always when there is a customer in the system.

A state of the queueing system is represented by several states of the Markov chain. For example, the state *Idle* of the queueing system is represented by the states $\langle 0/1\rangle, \langle 0/2\rangle$ and $\langle 0/3\rangle$ of the Markov chain. To calculate the probability of a state of the queueing system, the state probabilities of all associated states of the Markov chain must be added.

For example, we have for the probability p_{idle} that the system is idle

$$p_{\text{idle}} = \sum_{p=1}^{3} \pi_{\langle 0/p\rangle} \tag{3.79}$$

and for the number of customers in the system X

$$\mathrm{E}(X) = \sum_{n=0}^{3} \sum_{p=1}^{3} n\pi_{\langle n/p\rangle} \tag{3.80}$$

Figure 3.31 shows the Markov chain for the system state of an M/G/1/S queueing system with Coxian distributed service times (an M/Cox/1/S queueing system).

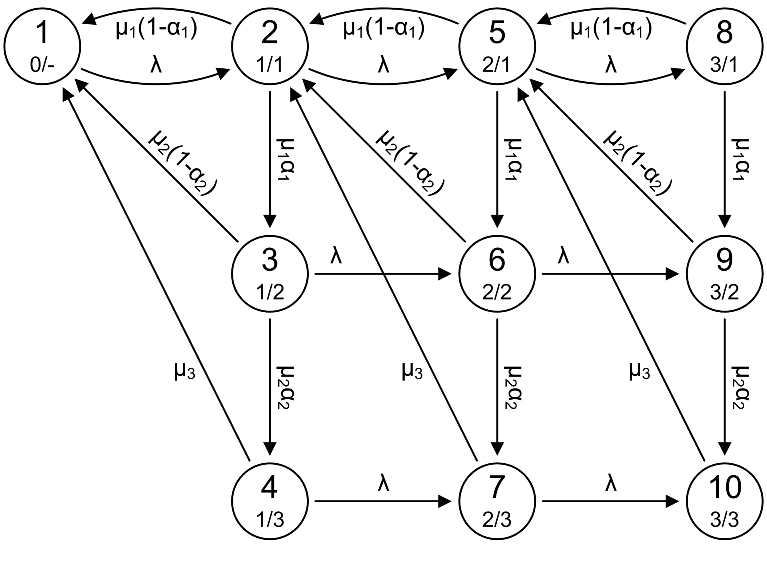


Figure 3.31.: Markov chain for the system state of an M/Cox/1/S queueing system. Meaning of the names of the states: number of customers in the system / state of the Coxian distribution of the service process.

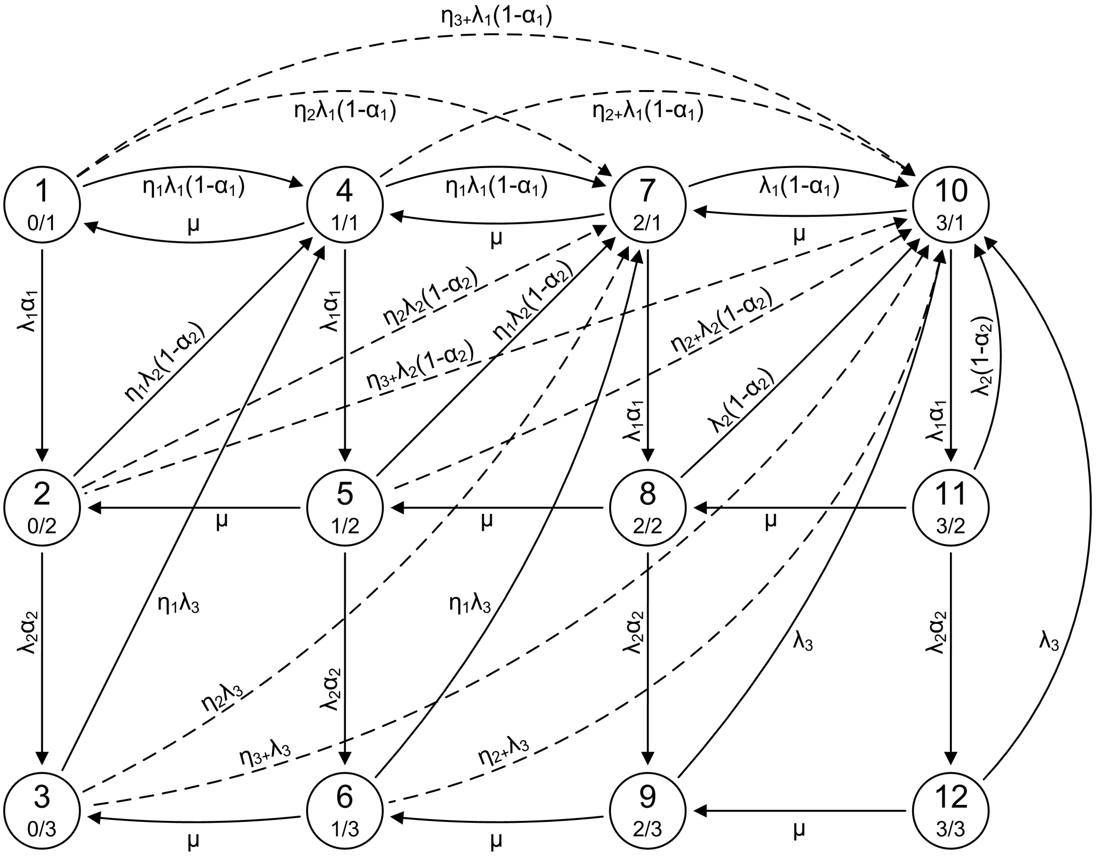


Figure 3.32.: Markov chain for the system state of a Cox*/M/1/S queueing system. Meaning of the names of the states: number of customers in the system / state of the Coxian* distribution of the arrival process. $\eta_n = (1-p)^{n-1}p$ is the probability that a batch of n customers arrives. $\eta_{n+} = (1-p)^{n-1}$ is the probability that a batch of n or more customers arrives. The transitions represented by dashed lines are needed to take batch arrivals into account.

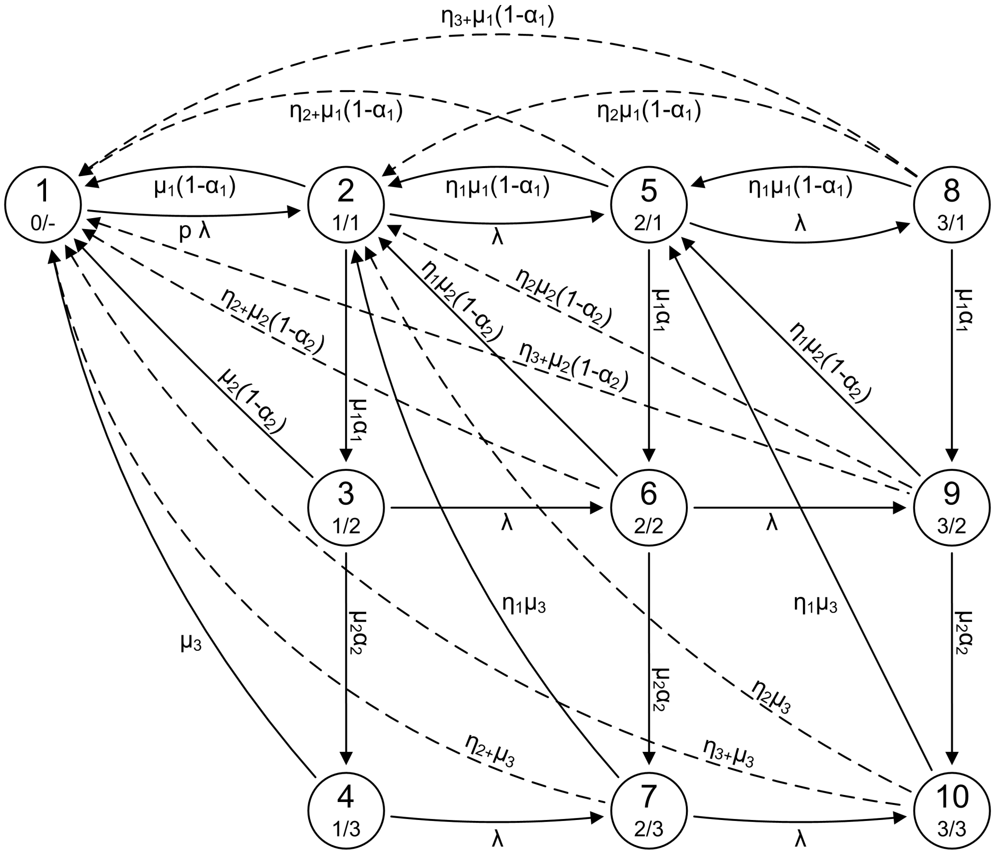

$\eta_{3+}\mu_1(1\text{-}\alpha_1)$
$\eta_{2+}\mu_1(1\text{-}\alpha_1)$
$\eta_2\mu_1(1\text{-}\alpha_1)$
1
0/-
$\mu_1(1\text{-}\alpha_1)$
p λ
2
1/1
$\eta_1\mu_1(1\text{-}\alpha_1)$
λ
5
2/1
$\eta_1\mu_1(1\text{-}\alpha_1)$
λ
8
3/1
$\eta_1\mu_2(1\text{-}\alpha_2)$
$\eta_2\mu_2(1\text{-}\alpha_2)$
$\eta_{2+}\mu_2(1\text{-}\alpha_2)$
$\eta_{3+}\mu_2(1\text{-}\alpha_2)$
$\mu_1\alpha_1$
$\mu_1\alpha_1$
$\mu_1\alpha_1$
$\eta_1\mu_2(1\text{-}\alpha_2)$
$\mu_2(1\text{-}\alpha_2)$
3
1/2
λ
6
2/2
λ
9
3/2
$\mu_2\alpha_2$
μ_3
$\eta_1\mu_3$
$\eta_1\mu_3$
$\mu_2\alpha_2$
$\mu_2\alpha_2$
$\eta_2\mu_3$
$\eta_{2+}\mu_3$
$\eta_{3+}\mu_3$
4
1/3
λ
7
2/3
λ
10
3/3

Figure 3.33.: Markov chain for the system state of an M/Cox*/1/S queueing system. Meaning of the names of the states: number of customers in the system / state of the Coxian* distribution of the service process. $\eta_n = (1-p)^{n-1}p$ is the probability that a batch of n customers is served. $\eta_{n+} = (1-p)^{n-1}$ is the probability that a batch of n or more customers is served. The transitions represented by dashed lines are needed to take batch services into account.

3.4.4. Approximation of a given distribution by a phase-type distribution

To approximate a given general probability distribution, usually a phase-type distribution whose first moments match those of the given distribution is used. For many applications, it is sufficient to consider only the first two moments (mean and variance). In this case, a suitable phase-type distribution can be found easily. However, for some applications two moments are not sufficient. To achieve a match in more than two moments, complex phase-type distributions (e.g., multistage Coxian distributions) are needed. Determining the parameters of such distributions is often very difficult.

Another approach to approximate a given probability distribution is to find a phase-type distribution whose cumulative distribution function has a curve shape which is similar to that of the given distribution. Whether it is better to match the moments or the curve shape depends on the application.

In the following, we present some important approximation algorithms. Unless otherwise mentioned, we use r and c to denote the rate and the coefficient of variation, respectively, of the given distribution.

Two moments, hypoexponential + hyperexponential distribution

The easiest method is to approximate a given distribution by finding a phase-type distribution that matches the first two moments. If the coefficient of variation is greater than 1, we choose a hyperexponential distribution, if the coefficient of variation is smaller than 1, we choose a hypoexponential distribution.

The hyperexponential distribution has two stages and its parameters are calculated from Equations 3.65 and 3.66 as follows:

$$\alpha_1 = \frac{1}{2}\left(1 - \sqrt{\frac{c^2-1}{c^2+1}}\right) \tag{3.81}$$

$$\alpha_2 = 1 - \alpha_1 \tag{3.82}$$

$$\mu_1 = 2\alpha_1 r \tag{3.83}$$

$$\mu_2 = 2\alpha_2 r \tag{3.84}$$

If $1/\sqrt{2} \leq c < 1$, we use a 2-stage hypoexponential distribution whose parameters are calculated from Equations 3.51 and 3.52:

$$\mu_2 = \frac{r\left(1+\sqrt{2c^2-1}\right)}{1-c^2} \tag{3.85}$$

$$\mu_1 = \frac{\mu_2 r}{\mu_2 - r} \tag{3.86}$$

If $c < 1/\sqrt{2}$, we need three or more stages whose rates cannot be determined so easily. In this case, the hypoexponential distribution can be constructed as follows:

The coefficient of variation of an n-stage Erlang distribution – which has the lowest variability among all hypoexponential distributions with the same number of stages –

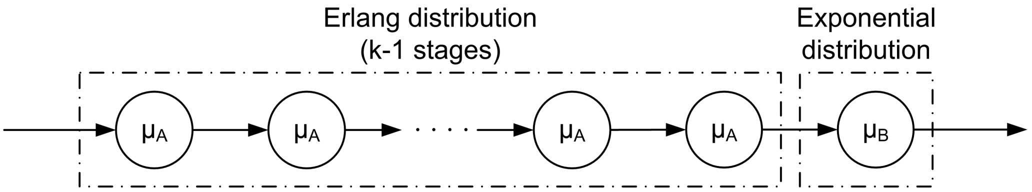


Figure 3.34.: Hypoexponential distribution as a sum of an Erlang and an exponential distribution.

is $1/\sqrt{n}$. That means we can achieve the given coefficient of variation by setting the number k of used stages to

$$k = \left\lceil \frac{1}{c^2} \right\rceil \tag{3.87}$$

Now we use a combination of a $k-1$-stage Erlang distribution A and an exponential distribution B as shown in Figure 3.34. The Erlang distribution produces a coefficient of variation of $c_A = 1/\sqrt{k-1} > c$, and we determine the rate of the exponential distribution so that the total coefficient of variation is decreased to c.

The procedure consists of two steps. In the first step, we find a hypoexponential distribution (consisting of the Erlang distribution $A*$ with mean 1 and the exponential distribution $B*$) that has coefficient of variation c but a rate that may differ from r. In the second step, we adjust the rates of the stages of the hypoexponential distribution so that its rate becomes r. This approach is possible since changing the rates of the stages of a hypoexponential distribution while keeping the ratio of the rates fixed does not change the coefficient of variation.

Mean and variance of the Erlang distribution $A*$ are

$$\mathrm{E}(A*) = 1 \tag{3.88}$$

$$\mathrm{Var}(A*) = (c_A\,\mathrm{E}(A*))^2 = c_A^2 = \frac{1}{k-1} \tag{3.89}$$

Now we have

$$c = \frac{\sqrt{\mathrm{Var}(A*) + \mathrm{Var}(B*)}}{\mathrm{E}(A*) + \mathrm{E}(B*)} = \frac{\sqrt{\mathrm{Var}(A*) + (\mathrm{E}(B*))^2}}{1 + \mathrm{E}(B*)} \tag{3.90}$$

From this we can calculate $\mathrm{E}(B*)$ with

$$\mathrm{E}(B*) = \frac{c^2 \pm \sqrt{\mathrm{Var}(A*)(c^2-1) + c^2}}{1-c^2} = \frac{c^2 \pm \sqrt{\frac{c^2k-1}{k-1}}}{1-c^2} \tag{3.91}$$

Unless $c^2k - 1 = 0$, in which case we need a pure Erlang distribution for the approximation ($c = 1/\sqrt{k} \Rightarrow c^2k = 1$), we get two results for $\mathrm{E}(B*)$.

In principle both results can be used, but if $\mathrm{E}(B*)$ is very small or very large this results in extremely unbalanced rates which can lead to numerical instabilities.[6] Therefore, we should use $\mathrm{E}(B*)^{(1)}$ if $\left|\log \mathrm{E}(B*)^{(1)}\right| \leq \left|\log \mathrm{E}(B*)^{(2)}\right|$ and $\mathrm{E}(B*)^{(2)}$ otherwise.

We adjust the total rate of the hypoexponential distribution:

$$\frac{1}{r} = \mathrm{E}(A)\,(1 + \mathrm{E}(B*)) \tag{3.92}$$

$$\mathrm{E}(A) = \frac{1}{r((1 + \mathrm{E}(B*))} \tag{3.93}$$

$$\mathrm{E}(B) = \mathrm{E}(A)\,\mathrm{E}(B*) \tag{3.94}$$

The rates of the stages of the hypoexponential distribution are

$$\mu_1, \ldots, \mu_{k-1} = \frac{k-1}{\mathrm{E}(A)} \tag{3.95}$$

$$\mu_k = \frac{1}{\mathrm{E}(B)} \tag{3.96}$$

Another method is to use a combination of two Erlang distributions A and B with a similar number of stages (k_A and k_B, respectively), as shown in Figure 3.35

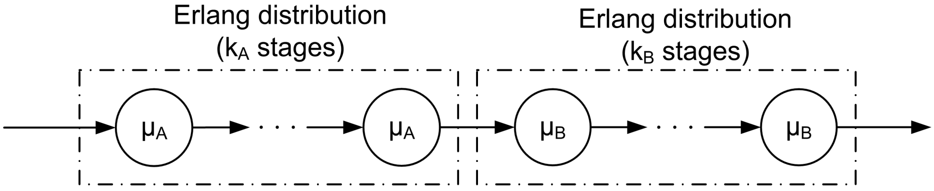


Figure 3.35.: Hypoexponential distribution as a sum of two Erlang distributions.

The number of stages for the distributions is

$$k = \left\lceil \frac{1}{c^2} \right\rceil \tag{3.97}$$

$$k_A = \left\lceil \frac{k}{2} \right\rceil \tag{3.98}$$

$$k_B = k - k_A \tag{3.99}$$

We set the mean of distribution $A*$ to 1 and calculate which mean distribution $B*$ must have to achieve the requested coefficient of variation c:

[6] An example: When $r = 1$ and $c = 0.499$ we have $\mathrm{E}(B*)^{(1)} = 0.00201$ and $\mathrm{E}(B*)^{(2)} = 0.661$. If we use $\mathrm{E}(B*)^{(1)}$, the resulting rates are $\mu = (4.01, 4.01, 4.01, 4.01, 498)$, otherwise we have $\mu = (6.64, 6.64, 6.64, 6.64, 2.51)$. Such extreme cases occur when c is only slightly below the coefficient of variation of a pure Erlang distribution.

$$c = \frac{\sqrt{\operatorname{Var}(A*) + \operatorname{Var}(B*)}}{\operatorname{E}(A*) + \operatorname{E}(B*)} = \frac{\sqrt{\frac{1}{k_A} + \frac{(\operatorname{E}(B*))^2}{k_B}}}{1 + \operatorname{E}(B*)} \tag{3.100}$$

$$\operatorname{E}(B*) = \frac{k_A k_B c^2 \pm \sqrt{k_A k_B (c^2(k_A + k_B) - 1)}}{k_A(1 - c^2 k_B)} \tag{3.101}$$

Again, in most cases there are two solutions, both of which can be used and – in contrast to the previously shown method – lead to acceptable results.[7]

Adjusting the total rate of the hypoexponential distribution:

$$\operatorname{E}(A) = \frac{1}{r((1 + \operatorname{E}(B*))} \tag{3.102}$$

$$\operatorname{E}(B) = \operatorname{E}(A)\operatorname{E}(B*) \tag{3.103}$$

The rates of the stages of the hypoexponential distribution are

$$\mu_A = \frac{k_A}{\operatorname{E}(A)} \tag{3.104}$$

$$\mu_B = \frac{k_B}{\operatorname{E}(B)} \tag{3.105}$$

Sauer/Chandy: two moments, hyperexponential + generalised Erlang distribution

In [Sauer 1975], C. H. Sauer and K. M. Chandy use a generalised Erlang distribution if $c < 1$ and a two-branch hyperexponential distributions if $c > 1$.

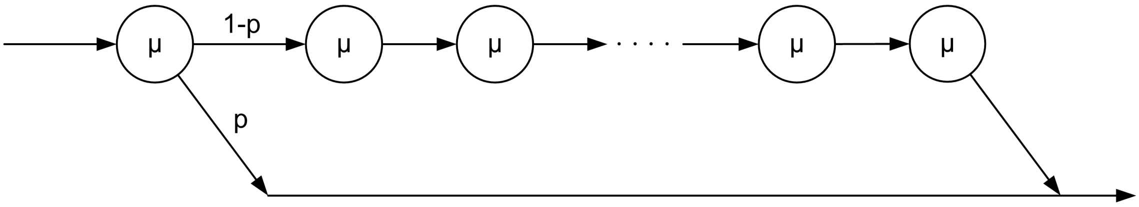


Figure 3.36.: Generalised Erlang distribution.

The generalised Erlang (GE) distribution is a modification of a standard Erlang distribution with k stages where a customer, after passing through the first stage, leaves the system with probability p, or continues through the remaining $k - 1$ stages with probability $1 - p$ (see Figure 3.36). By varying p between 1 to 0, the coefficient of variation can be set to a value in the range from 1 to $1/\sqrt{k}$.

[7] When $r = 1$ and $c = 0.499$ we have $\operatorname{E}(B*)^{(1)} = 0.187$ and $\operatorname{E}(B*)^{(2)} = 1.797$. If we use $\operatorname{E}(B*)^{(1)}$, the resulting rates are $\mu = (3.56, 3.56, 3.56, 12.7, 12.7)$, otherwise we have $\mu = (8.39, 8.39, 8.39, 3.11, 3.11)$.

The parameters are calculated as follows:

$$k = \left\lceil \frac{1}{c^2} \right\rceil \tag{3.106}$$

$$p = \frac{2kc^2 + k - 2 - \sqrt{k^2 + 4 - 4kc^2}}{2(c^2+1)(k-1)} \tag{3.107}$$

$$\mu = r(k - p(k-1)) \tag{3.108}$$

The parameters of the hyperexponential distribution are calculated with

$$\alpha_1 = \frac{c^2 + 1 - \sqrt{c^4 - 1}}{2(c^2+1)} \tag{3.109}$$

$$\alpha_2 = 1 - \alpha_1 \tag{3.110}$$

$$\lambda_1 = 2r\alpha_1 \tag{3.111}$$

$$\lambda_2 = 2r\alpha_2 \tag{3.112}$$

Figures 3.37 and 3.38 show the Markov chains for the system state of a GE/M/1/S queueing system and an M/GE/1/S queueing system.

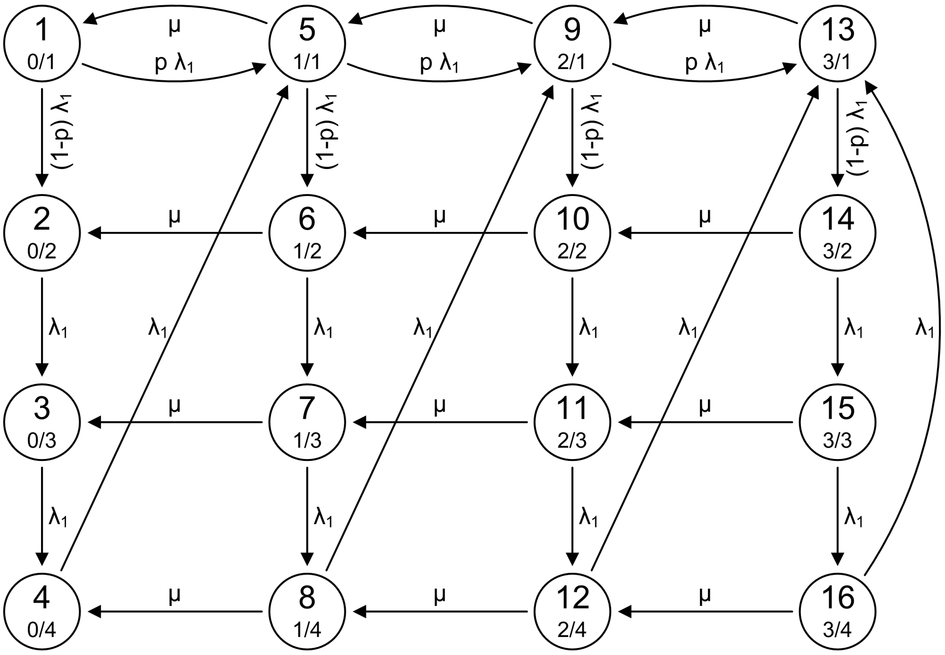


Figure 3.37.: Markov chain for the system state of a GE/M/1/S queueing system. Meaning of the names of the states: number of customers in the system / state of the GE distribution of the arrival process.

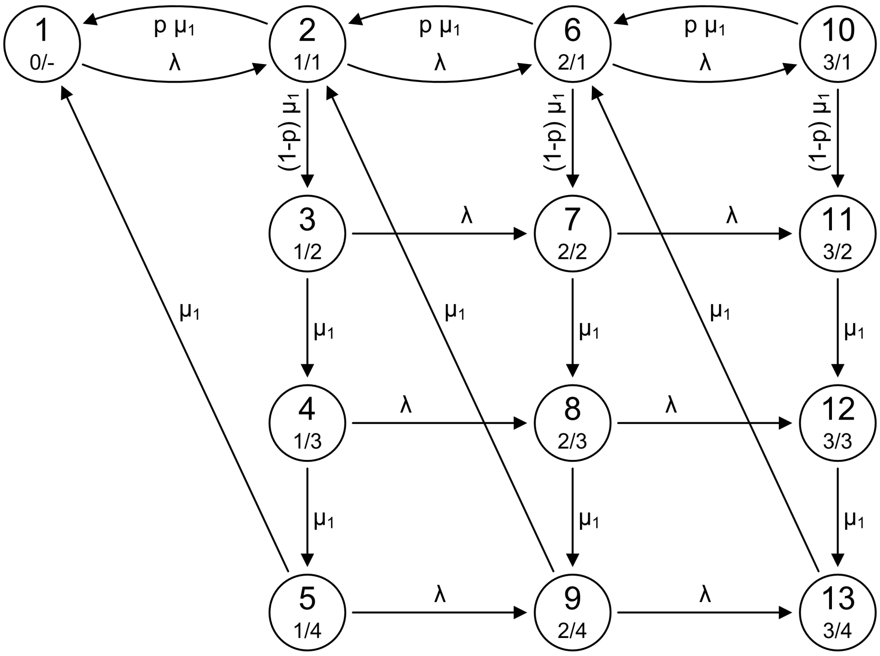


Figure 3.38.: Markov chain for the system state of an M/GE/1/S queueing system. Meaning of the names of the states: number of customers in the system / state of the GE distribution of the service process.

Marie: two moments, 2-stage Coxian distribution

Marie [Marie 1980] uses a 2-stage Coxian distribution, with which coefficients of variation $\geq 1/\sqrt{2}$ can be achieved.

The parameters of the Coxian distribution are calculated as follows:

$$\mu_1 = 2r \tag{3.113}$$

$$\mu_2 = \frac{r}{c^2} \tag{3.114}$$

$$\alpha_1 = \frac{1}{2c^2} \tag{3.115}$$

For coefficients of variation smaller than $1/\sqrt{2}$ he also suggests using a generalised Erlang distribution.

Osogami/Marchol-Balter: three moments, EC distribution

T. Osogami and M. Harchol-Balter present in [Osogami 2005] an algorithm that calculates the parameters of a so-called “EC distribution” (a combination of an Erlang distribution Y and a 2-stage Coxian distribution X, see Figure 3.39), which matches 3 moments of a given distribution.

The algorithm is fast, finds a solution with a nearly minimal number of phases, and works for almost all non-negative input distributions. A big disadvantage is that the EC distribution is in fact a Coxian* distribution, which is not very suitable for the use in Markov chains.

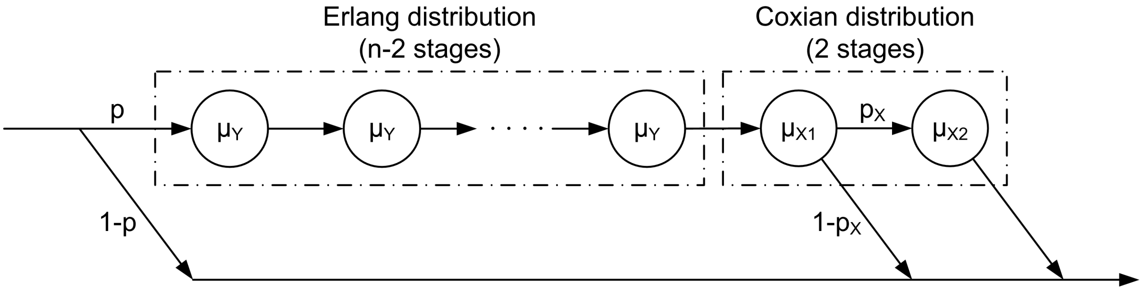


Figure 3.39.: EC distribution.

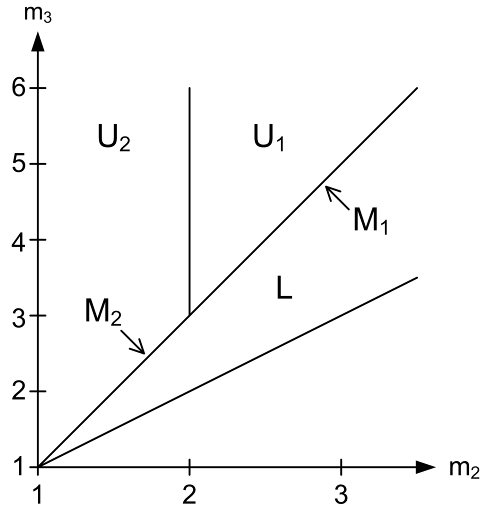


Figure 3.40.: Classification of distributions.

The parameters of the EC distribution are calculated as follows:

First it is determined to which of the following sets the distribution G that we want to approximate belongs (Figure 3.40):

$$U_1 = \left\{F : m_2^F > 2 \text{ and } m_3^F > 2m_2^F - 1\right\} \tag{3.116}$$

$$U_2 = \left\{F : 1 < m_2^F \leq 2 \text{ and } m_3^F > 2m_2^F - 1\right\} \tag{3.117}$$

$$M_1 = \left\{F : m_2^F \geq 2 \text{ and } m_3^F = 2m_2^F - 1\right\} \tag{3.118}$$

$$M_2 = \left\{F : 1 < m_2^F < 2 \text{ and } m_3^F = 2m_2^F - 1\right\} \tag{3.119}$$

$$L = \left\{F : m_2^F > 1 \text{ and } m_2^F < m_3^F < 2m_2^F - 1\right\} \tag{3.120}$$

m_2^F and m_3^F refer to the second and third normalised moments of the random variable F, which are defined as

$$m_2^F = \frac{\mathrm{E}(F^2)}{(\mathrm{E}(F))^2} = c_F^2 + 1 \tag{3.121}$$

$$m_3^F = \frac{\mathrm{E}(F^3)}{\mathrm{E}(F)\ \mathrm{E}(F^2)} = \gamma_F \sqrt{m_2^F} \tag{3.122}$$

(γ_F is the skewness of F.)

Depending on to which set G belongs, we need different structures of the EC distribution.

(i) If $G \in U_1 \cup M_1$, then a 2-stage Coxian distribution suffices to match the first three moments of G. Let Z be the EC distribution, then we have

$$Z = X \tag{3.123}$$

$$\Rightarrow p = 1, n = 2 \tag{3.124}$$

The parameters of the Coxian distribution are calculated as follows:

$$\mu_{X1} = \frac{u + \sqrt{u^2 - 4v}}{2\,\mathrm{E}(G)} \tag{3.125}$$

$$\mu_{X2} = \frac{u - \sqrt{u^2 - 4v}}{2\,\mathrm{E}(G)} \tag{3.126}$$

$$p_X = \frac{\mu_{X2}\,\mathrm{E}(G)(\mu_{X1}\,\mathrm{E}(G) - 1)}{\mu_{X1}\,\mathrm{E}(G)} \tag{3.127}$$

where

$$u = \frac{6 - 2m_3^G}{3m_2^G - 2m_3^G} \tag{3.128}$$

$$v = \frac{12 - 6m_2^G}{m_2^G(3m_2^G - 2m_3^G)} \tag{3.129}$$

(ii) If $G \in U_2 \cup M_2$, we need an Erlang distribution and a 2-stage Coxian distribution:

$$p = 1 \tag{3.130}$$

$$Z = \sum_{k=1}^{n-2} Y_k + X \tag{3.131}$$

The necessary number of stages is

$$n = \left\lfloor \frac{m_2^G}{m_2^G - 1} + 1 \right\rfloor \tag{3.132}$$

The Coxian distribution has the moments

$$m_2^X = \frac{(n-3)m_2^G - (n-2)}{(n-2)m_2^G - (n-1)} \tag{3.133}$$

$$m_3^X = \frac{\gamma m_3^G - \beta}{m_2^X} \tag{3.134}$$

$$\mathrm{E}(X) = \frac{\mathrm{E}(G)}{(n-2)m_2^X - (n-3)} \tag{3.135}$$

where

$$\beta = (n-2)(m_2^X - 1)(n(n-1)(m_2^X)^2 - n(2n-5)m_2^X + (n-1)(n-3)) \tag{3.136}$$
$$\gamma = \left((n-1)m_2^X - (n-2)\right)\left((n-2)m_2^X - (n-3)\right)^2 \tag{3.137}$$

The parameters of the Coxian distributed are calculated according to case (i), whereby the moments of X (Equations 3.133 - 3.135) are used instead of the moments of G.

(iii) If $G \in L$, we need a probability mass at 0.

$$Z = \begin{cases} W & \text{with probability } p \\ 0 & \text{with probability } 1-p \end{cases} \tag{3.138}$$

where

$$p = \frac{1}{2m_2^G - m_3^G} \tag{3.139}$$

The moments of the distribution W are

$$m_2^W = p\, m_2^G \tag{3.140}$$
$$m_3^W = p\, m_3^G \tag{3.141}$$
$$\mathrm{E}(W) = \frac{\mathrm{E}(G)}{p} \tag{3.142}$$

Depending on to which distribution set W belongs (Equations 3.116 - 3.120), we need either a Coxian distribution or a Coxian/Erlang combination for the generation of W: If $W \in M_1$, the parameters of W are calculated according to case (i), otherwise (we then have $W \in M_2$) the parameters are calculated according to case (ii). In both cases, the moments of W (Equations 3.140 - 3.142) are used instead of the moments of G.

Whitt: three moments, hyperexponential distribution

W. Whitt shows in [Whitt 1982] how the parameters of a 2-stage hyperexponential distribution can be determined so that the distribution matches 3 given moments.

$$\lambda_{1,2} = \frac{6M_1 y}{(x + 1.5y^2 + 3{M_1}^2 y) \pm \sqrt{(x + 1.5y^2 - 3{M_1}^2 y)^2 + 18{M_1}^2 y^3}} \tag{3.143}$$
$$p_1 = \frac{M_1 - \frac{1}{\lambda_2}}{\frac{1}{\lambda_1} - \frac{1}{\lambda_2}} \tag{3.144}$$
$$p_2 = 1 - p_1 \tag{3.145}$$

where M_i is the ith moment of the given distribution and

$$x = M_1 M_3 - 1.5{M_2}^2 \tag{3.146}$$
$$y = M_2 - 2{M_1}^2 \tag{3.147}$$

distribution	m_1	m_2	m_3	m_4	m_5
given distribution	1	1.475	1.929	2.375	2.817
exponential	1	2	3	4	5
hypoexponential	1	1.475	2.016	2.604	3.220
generalised Erlang	1	1.475	1.885	2.280	2.669

Table 3.1.: Normalised moments of the given distribution and the distributions used for the approximation.

If M_3 is not sufficiently high, no 2-stage hyperexponential distribution with the given moments exists. In this case, Whitt suggests that one could use a hyperexponential distribution with a third moment slightly above $3{M_2}^2/(2M_1)$. (In a 2-stage hyperexponential distribution, $M_3 M_1 \geq 3/2 \cdot {M_2}^2$ holds.)

Curve fitting algorithms

A. Feldmann and W. Whitt present in [Feldmann 1998] a method with which long-tail distributions can be approximated by hyperexponential distributions. Y. Sasaki describes in [Sasaki 2001] how a Coxian distribution can be used to approximate both short and long-tail distributions.

In [Sommereder 2008], it is shown how an evolutionary algorithm can be used to find a Coxian distribution that has a similar cumulative distribution function as a given distribution. This algorithm can easily be adapted so that a Coxian distribution with certain moments (instead of a certain shape of the cumulative distribution function) is found.

Comparison of approximation algorithms

Which approximation method should be used depends strongly on the actual application. It is not possible to say that a certain approximation method is superior to the others in all cases.

In the following, we demonstrate how some of the shown approximation algorithms perform when used to approximate given distributions that are used as distributions for the interarrival times of a GI/M/1/S queueing system and as service time distributions of an M/G/1/S queueing system.

The first distribution[8] we want to approximate has a coefficient of variation smaller than 1. We approximate it by

- an exponential distribution, which matches in the first moment,
- a 3-stage hypoexponential distribution, which is constructed as shown on page 68 and matches in the first two moments, and
- a 3-stage generalised Erlang distribution, which matches in the first two moments.

Figure 3.41 shows the probability density function of the given distribution and the approximations, Table 3.1 shows the higher moments of the distributions.

[8] It is a Coxian distribution with parameters $\lambda_1 = \lambda_2 = 2.235$, $\lambda_3 = 2.98$, $\alpha_1 = 0.95$, $\alpha_2 = 0.4$.

Figures 3.42 and 3.43 show the number of customers and the probability that the server is idle in a GI/M/1/S (M/G/1/S) queueing system when, for the distribution of the interarrival times (service times), the given distribution and the approximation distributions are used.

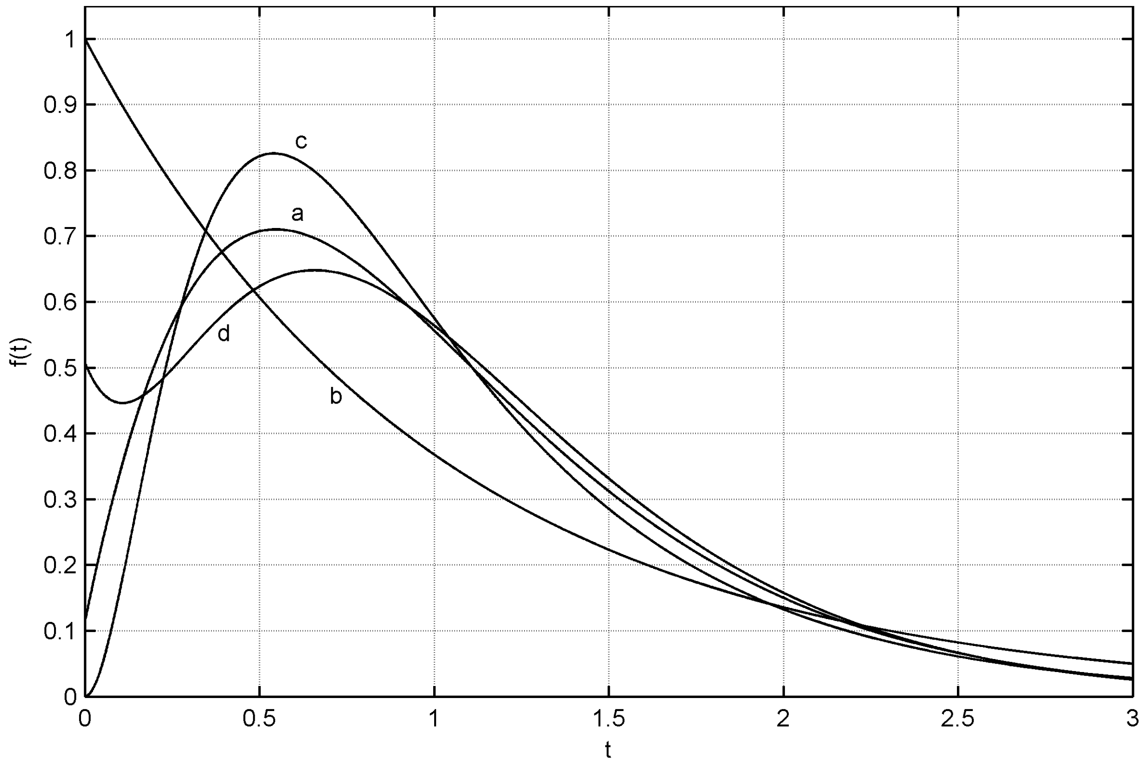


Figure 3.41.: Approximation of a given distribution (a) by (b) an exponential distribution, (c) a hypoexponential distribution and (d) a generalised Erlang distribution. The figure shows the probability density function of the distributions.

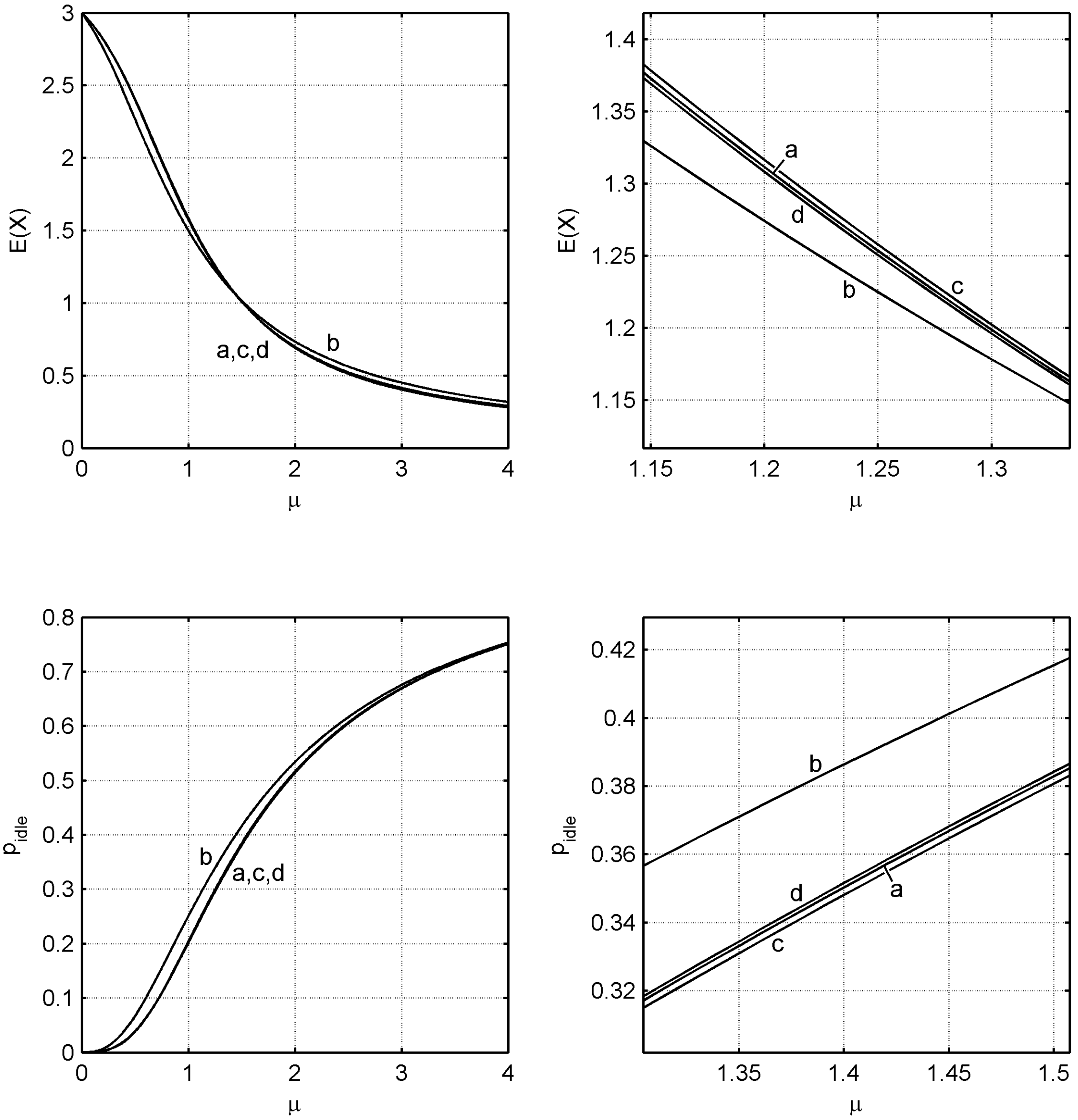


Figure 3.42.: Number of customers in the system and probability that the server is idle in a GI/M/1/S queueing system where the interarrival times are distributed according to (a) the given distribution and (b) the exponential distribution, (c) the hypoexponential distribution and (d) the generalised Erlang distribution that are used to approximate the given distribution. $S = 3$, the service rate is μ.

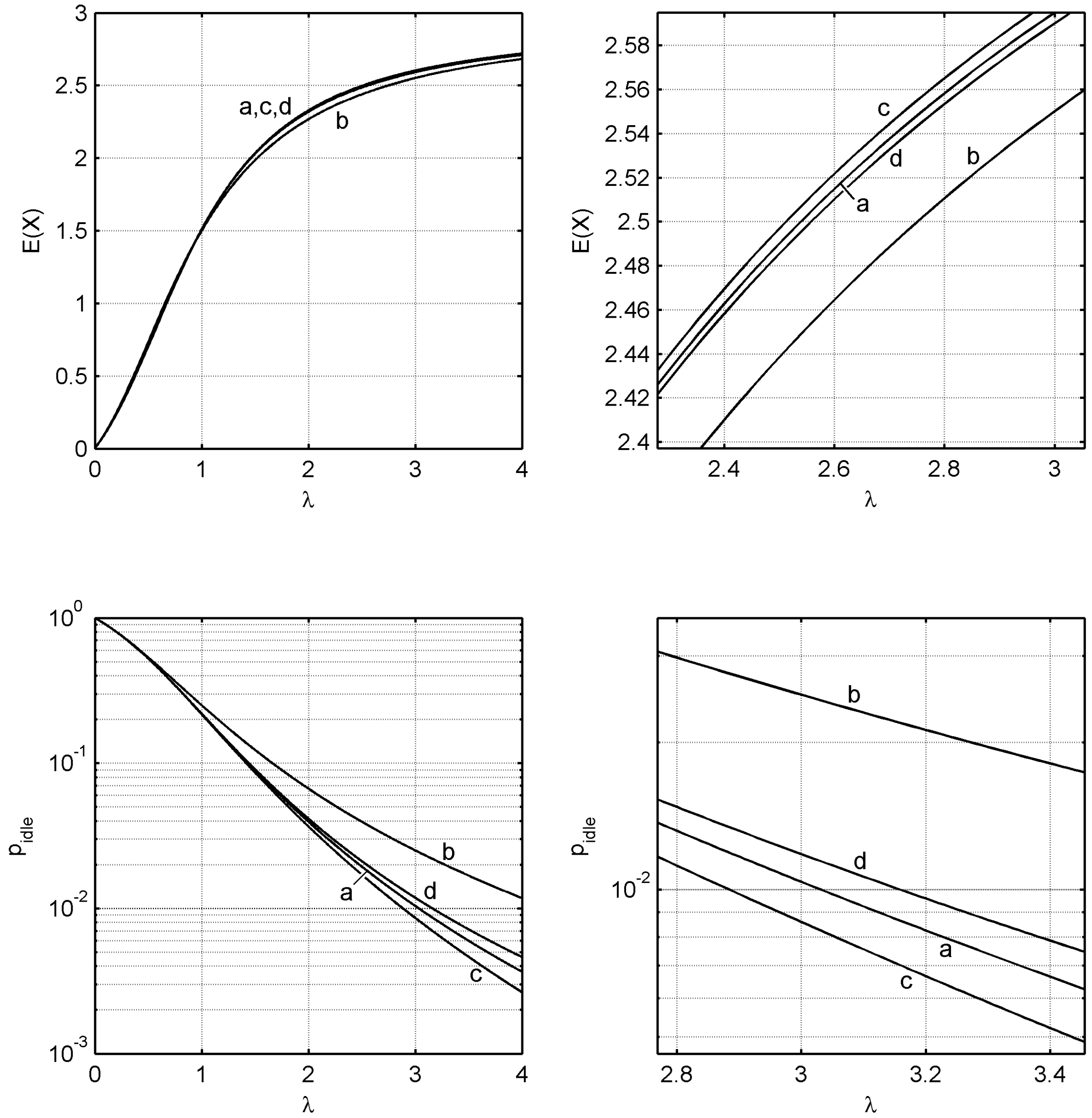


Figure 3.43.: Number of customers in the system and probability that the server is idle in an M/G/1/S queueing system where the service times are distributed according to (a) the given distribution and (b) the exponential distribution, (c) the hypoexponential distribution and (d) the generalised Erlang distribution that are used to approximate the given distribution. $S = 3$, the arrival rate is λ.

The second distribution[9] we want to approximate has a coefficient of variation greater than 1. We approximate it by

- an exponential distribution, which matches in the first moment,
- a 2-stage hyperexponential distribution, which is constructed as shown on page 66 and matches in the first two moments,
- a 2-stage hyperexponential distribution, which is constructed according to Whitt's formula and matches in the first three moments, and
- an EC-distribution, which matches in the first three moments.

Figure 3.44 shows the probability density function of the given distribution and the approximations, Table 3.2 shows the higher moments of the distributions.

Figures 3.45 and 3.46 show the number of customers and the probability that the server is idle in a GI/M/1/S (M/G/1/S) queueing system when, for the distribution of the interarrival times (service times), the given distribution and the approximation distributions are used.

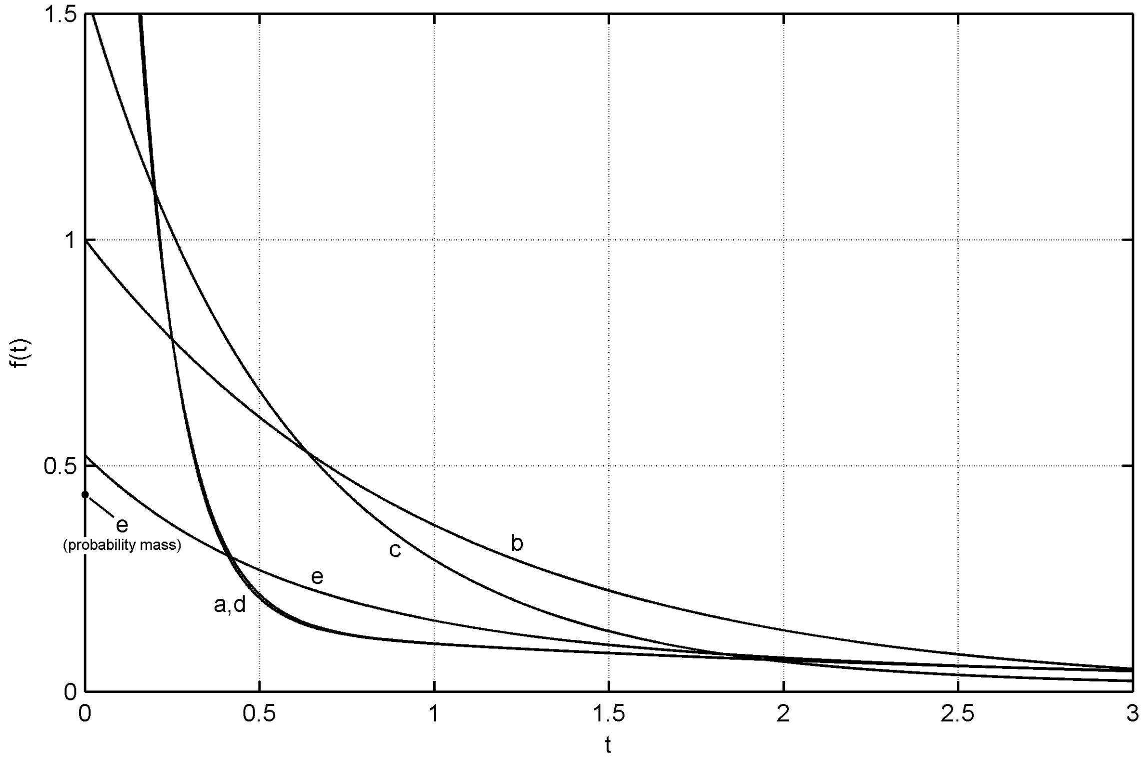


Figure 3.44.: Approximation of a given distribution (a) by (b) an exponential distribution, (c) a 2-stage hyperexponential distribution (2 moments match), (d) a 2-stage hyperexponential distribution (3 moments match), (e) an EC distribution. The figure shows the probability density function of the distributions.

[9] It is a Coxian distribution with parameters $\lambda_1 = 70.2247$, $\lambda_2 = 7.9824$, $\lambda_3 = 0.4122$, $\alpha_1 = 0.9$, $\alpha_2 = 0.4$.

distribution	m_1	m_2	m_3	m_4	m_5
given distribution	1	4.51	7.247	9.699	12.13
exponential	1	2	3	4	5
hyperexponential	1	4.51	10.53	15.47	19.63
hyperexponential (Whitt)	1	4.51	7.247	9.697	12.12
EC	1	4.51	7.247	9.82	12.32

Table 3.2.: Normalised moments of the given distribution and the distributions used for the approximation.

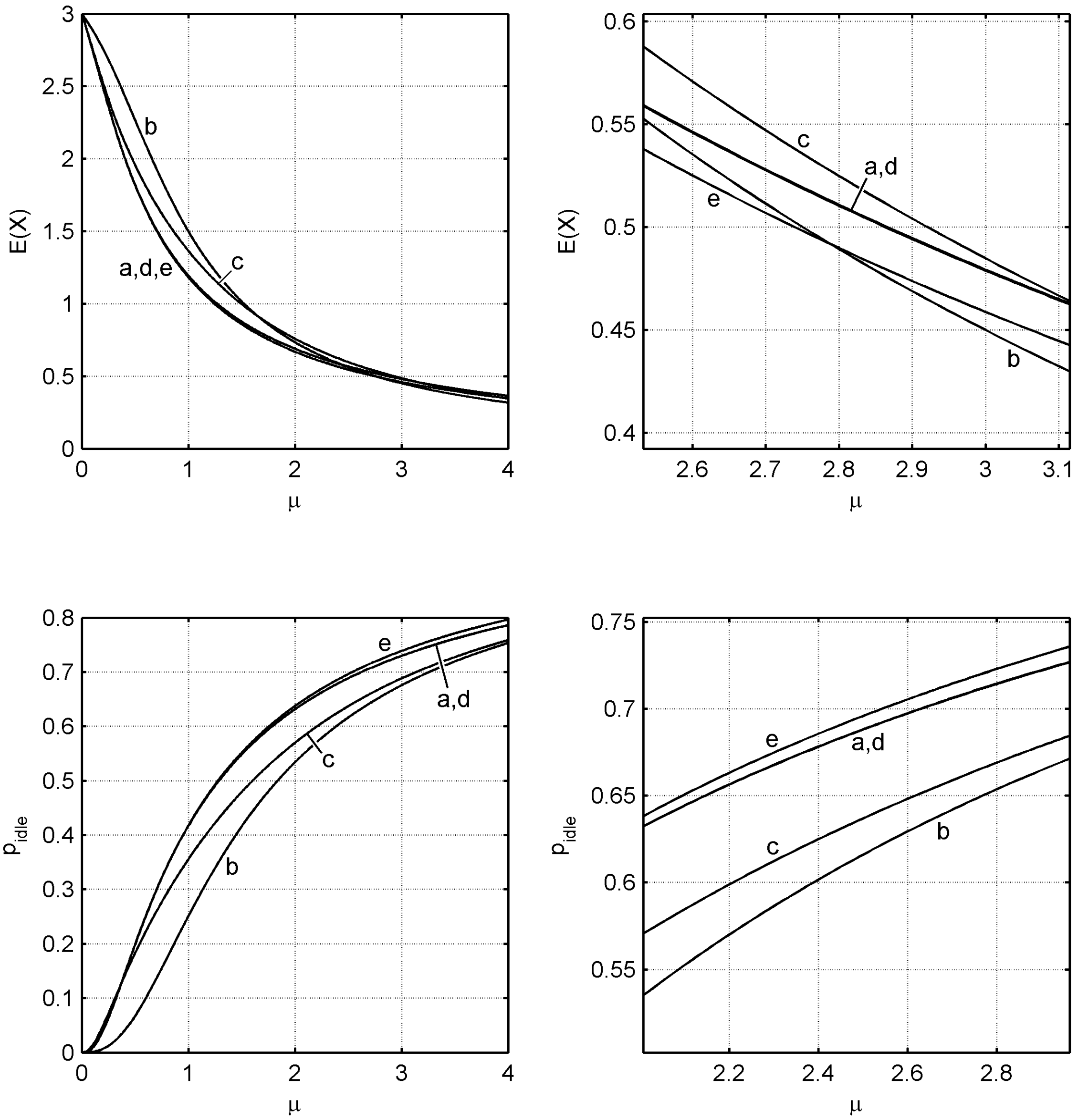


Figure 3.45.: Number of customers and probability that the server is idle in a GI/M/1/S queueing system where the interarrival times are distributed according to (a) the given distribution and (b) the exponential distribution, (c) the first hyperexponential distribution (2 moments match), (d) the second hyperexponential distribution (3 moments match) and (d) the EC distribution that are used to approximate the given distribution. $S = 3$, the service rate is μ.

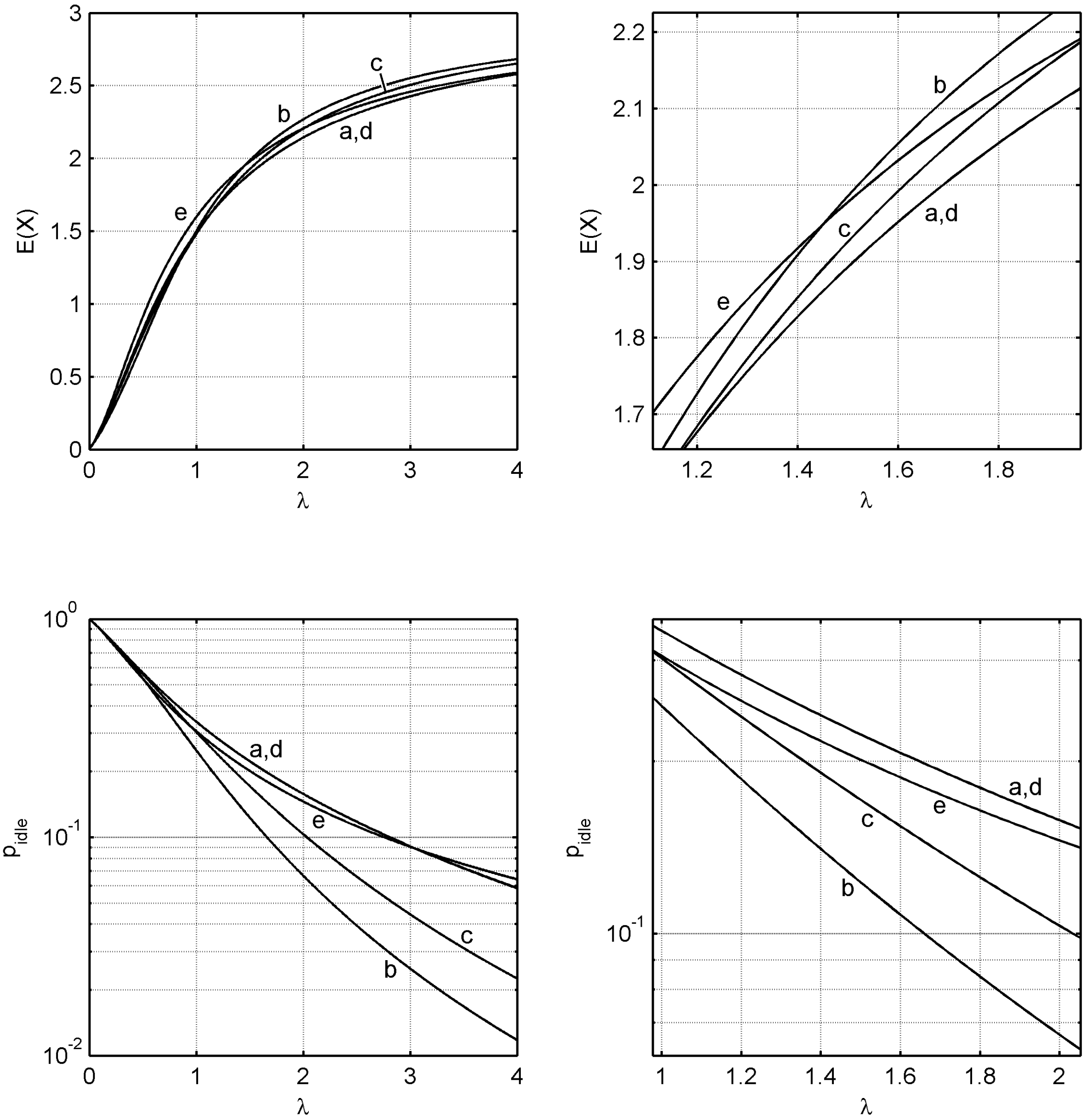


Figure 3.46.: Number of customers and probability that the server is idle in an M/G/1/S queueing system where the service times are distributed according to (a) the given distribution and (b) the exponential distribution, (c) the first hyperexponential distribution (2 moments match), (d) the second hyperexponential distribution (3 moments match) and (d) the EC distribution that are used to approximate the given distribution. $S = 3$, the arrival rate is λ.

3.5. Modelling of traffic streams

The counting process $N(t) = \max\{0,\, j : T_j \leq t\}$ of a point process (traffic stream) $T = \langle T_1, T_2, \dots \rangle$ can be thought of as the number of customers in a queueing system with an infinite queue and no servers whose arrival stream is the point process under consideration. Therefore, we can model the state of a counting process of a traffic stream with phase-type distributed interevent times with the techniques shown in the previous section.

Figure 3.47a shows the state transition diagram of the counting process of a traffic stream with Coxian distributed interevent times. By applying the rules shown in Figure 3.29, we obtain the Markov chain $\mathcal{M}_C$ for the state of the counting process (Figure 3.47b).

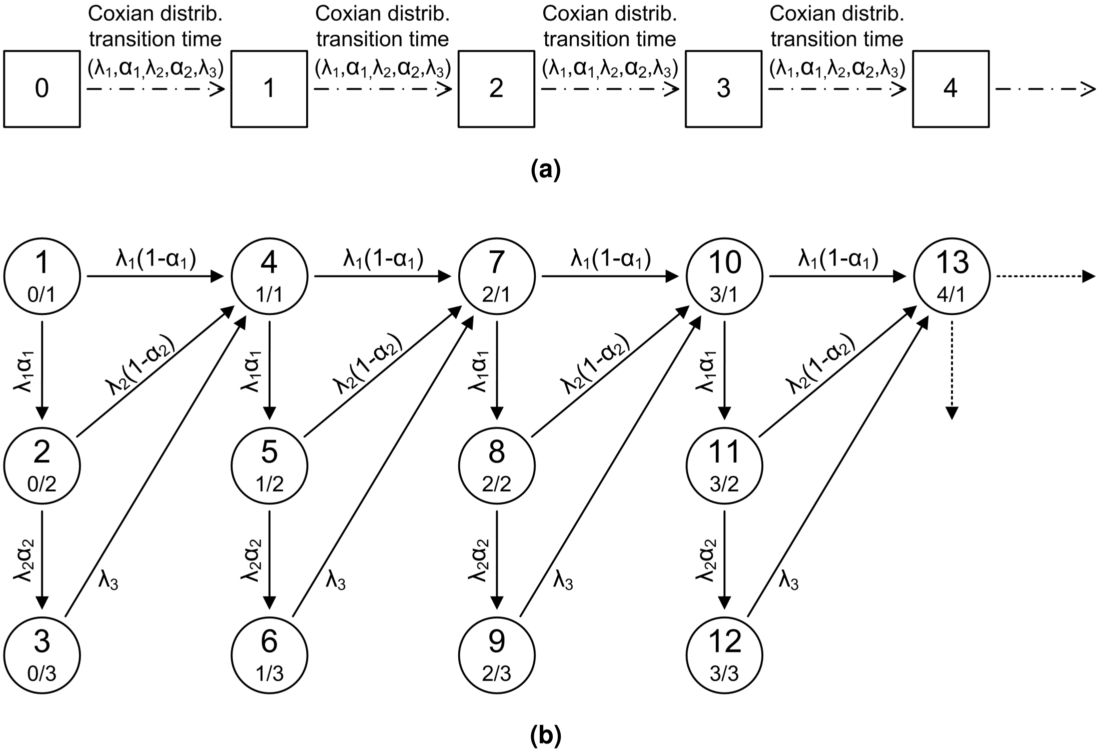
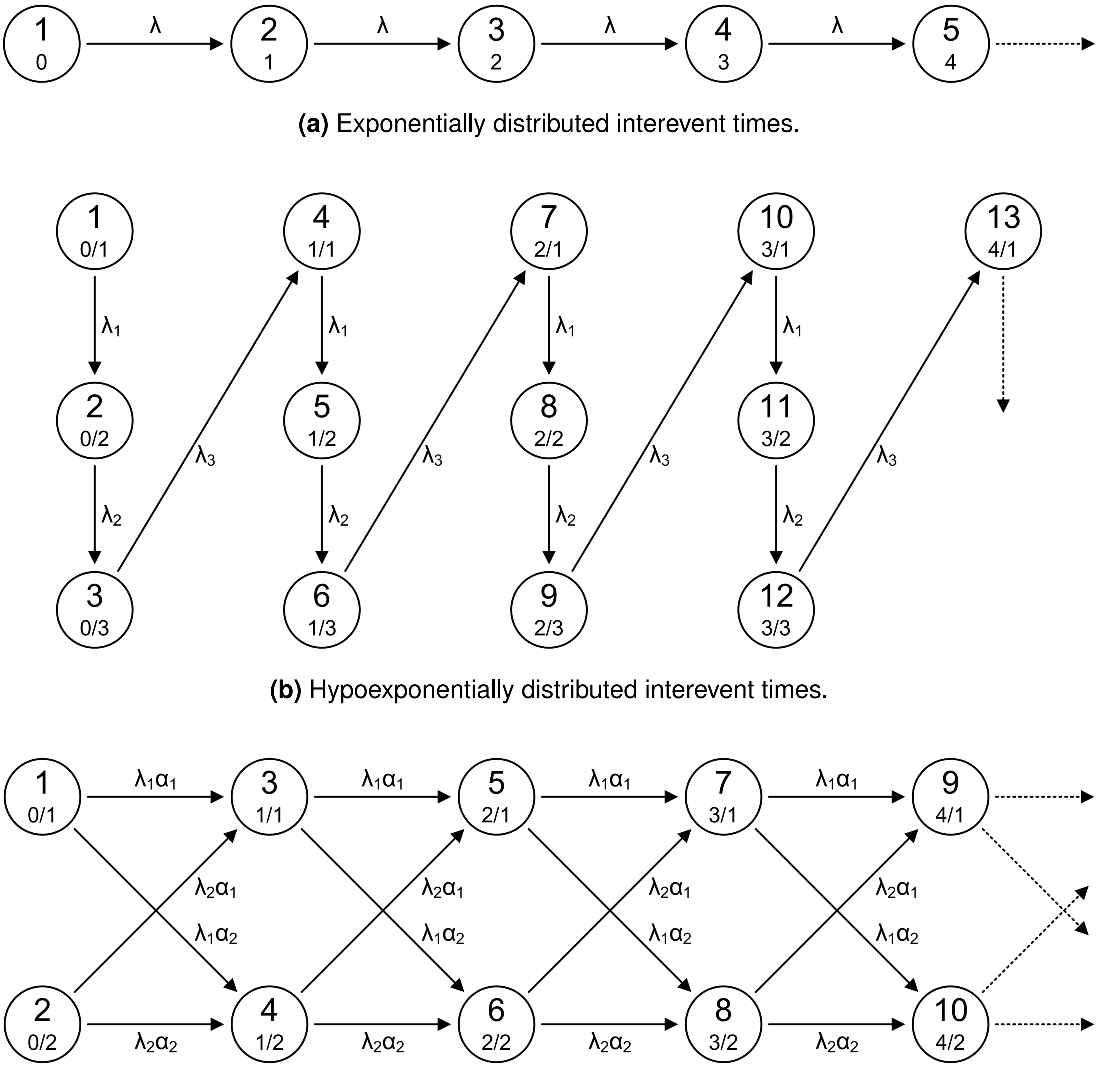


Figure 3.47.: Counting process of a point process with Coxian distributed interevent times. Meaning of the names of the states: value of the counting process / state of the Coxian distribution.

The Markov chains for the state of the counting process of traffic streams with exponentially, hypoexponentially and hyperexponentially distributed interevent times are shown in Figure 3.48.



(a) Exponentially distributed interevent times.

(b) Hypoexponentially distributed interevent times.

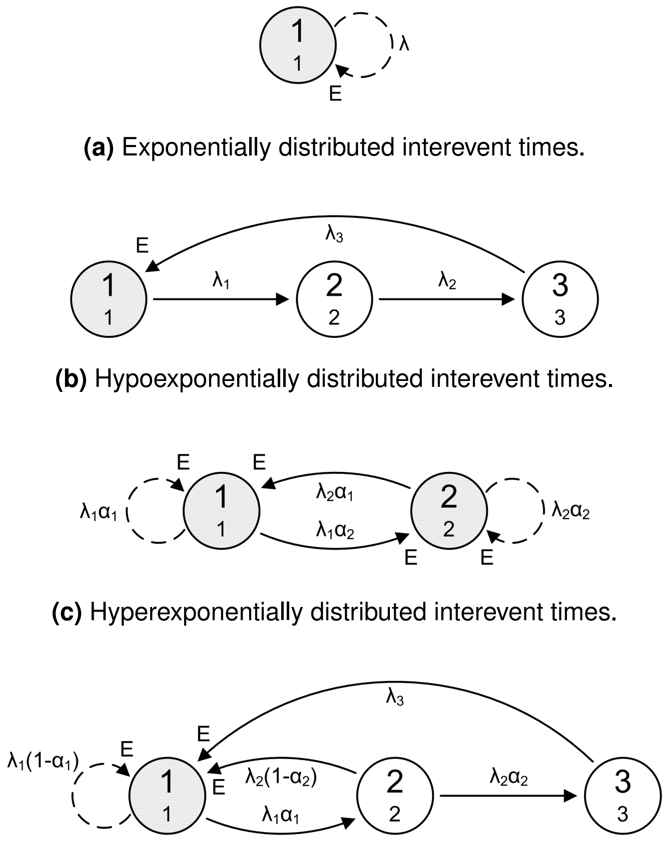
(c) Hyperexponentially distributed interevent times.

Figure 3.48.: Markov chains for the state of the counting process of traffic streams.

In Chapters 8 and 9, where we deal with the manipulation of traffic streams, we are interested only in the state of the point process, and not in the state of the counting process.

To model the state of the point process, we remove the part that describes the value of the counting process in the names of the states of the Markov chain for the state of the counting process $\mathcal{M}_C$. If we do this, many states will have the same name. These states describe the same state of the point process and are therefore combined into one single state.

Figure 3.49 shows the Markov chains for the state of point processes with exponentially, hypoexponentially, hyperexponentially and Coxian distributed interevent times.



(a) Exponentially distributed interevent times.

(b) Hypoexponentially distributed interevent times.

(c) Hyperexponentially distributed interevent times.

(d) Coxian distributed interevent times.

Figure 3.49.: Markov chains for the system state of streams. Hidden transitions are painted with dashed lines. States in which the Markov chain is after an event has taken place are shaded grey. Transitions annotated with “E” create events.

4. Related topics

In this chapter, we give an overview of some topics related to the content of this work. In Sections 4.1 to 4.3, we discuss performance evaluation techniques based on Markov chains: the embedded Markov chain method, the matrix geometric method, and matrix analytic methods. In Section 4.4, we present the regenerative method, which is not based on Markov chains, but contains interesting probabilistic arguments, which might be combined with the techniques shown in this work.

4.1. Embedded Markov chain method

When the service times or the interarrival times of a queueing system are not exponentially (or phase-type) distributed, but have a general distribution, the state of the system cannot be modelled as a Markov chain any more. The reason is that the further evolution of the state of the queueing system does not depend only on the current state, but also on the time that has elapsed since the last arrival or the time when the current service began.

However, it is often possible to find quantities of the queueing system that do constitute a Markov chain. By means of this *embedded Markov chain*, many characteristics of the underlying queueing system can be determined.

Let us consider an M/G/1 queueing system.[1] The interarrival times are independent and exponentially distributed with rate λ, and the service times are independent and have a general distribution with cumulative distribution function $F_S(\cdot)$ and rate μ.

Assume $t_0 = 0 < t_1 < t_2 < \dots$ denote the times at which services are completed, and let the sequence $\langle Y_n \rangle_{n \geq 0}$ be the number of customers in the queueing system immediately after the service completions that take place at times t_n (Figure 4.1). Then $\langle Y_n \rangle$ constitutes a discrete-time Markov chain. (Since the interarrival times are exponentially distributed and independent of the system state, Y_{j+1} depends only on Y_j.)

Let the random variable k be the number of customers that arrive during a service and let k_n be the probability that there are n arrivals during a service period. Then we can write[2]

$$k_n = \mathrm{P}\left\{k = n\right\} = \int\limits_{t=0}^{\infty} \mathrm{P}\left\{k = n \mid S = t\right\} \mathrm{d}F_S(t) = \int\limits_{t=0}^{\infty} \mathrm{e}^{-\lambda t} \frac{(\lambda t)^n}{n!} \, \mathrm{d}F_S(t) \tag{4.1}$$

[1] cf. [Allen 1978]

[2] The Stiltjes integral $\int_{t=0}^{\infty} g(t) \mathrm{d}F_X(t)$ is evaluated as $\int_{t=0}^{\infty} g(t) f_X(t) \, \mathrm{d}t$ or as $\sum_i g(t_i) p_X(t_i)$, depending upon whether the random variable X with distribution function $F_X(\cdot)$ is continuous with probability density function $f_X(\cdot)$ or discrete with probability mass function $p_X(\cdot)$.

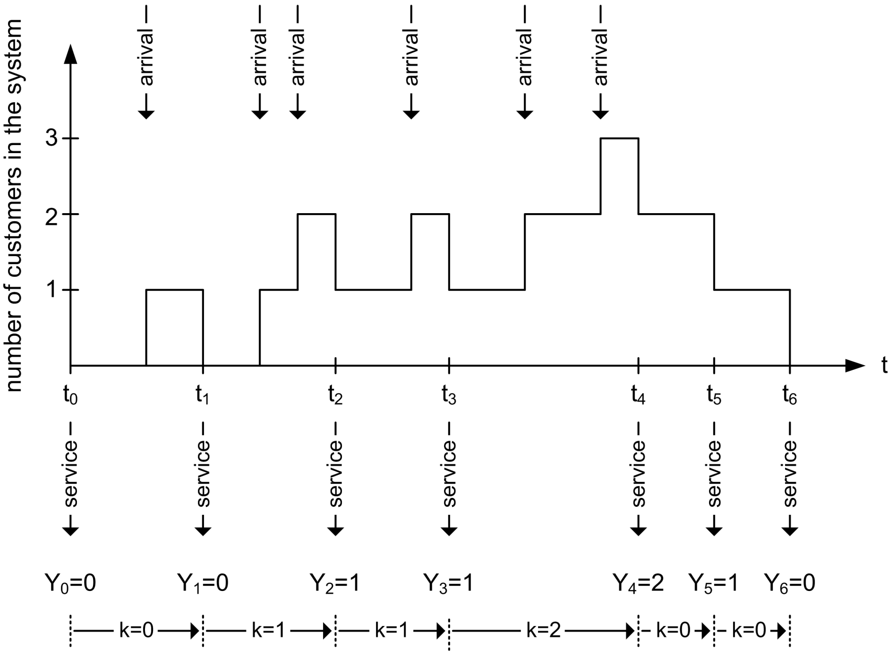

number of customers in the system
1
2
3
t
arrival
t_0 t_1 t_2 t_3 t_4 t_5 t_6
service
$Y_0=0$ $Y_1=0$ $Y_2=1$ $Y_3=1$ $Y_4=2$ $Y_5=1$ $Y_6=0$
k=0 k=1 k=1 k=2 k=0 k=0

Figure 4.1.: Embedded Markov chain method. The sequence of the number of customers in the system immediately after service completions constitute a discrete-time Markov chain.

Now we have

$$Y_{n+1} = \begin{cases} Y_n - 1 + k & \text{if } Y_n \geq 1 \\ k & \text{if } Y_n = 0 \end{cases} \tag{4.2}$$

and therefore the transition probabilities of the embedded Markov chain are

$$\mathcal{P} = \begin{pmatrix} k_0 & k_1 & k_2 & k_3 & k_4 & \cdots \\ k_0 & k_1 & k_2 & k_3 & k_4 & \cdots \\ 0 & k_0 & k_1 & k_2 & k_3 & \cdots \\ 0 & 0 & k_0 & k_1 & k_2 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{4.3}$$

It can be shown[3] that in an M/G/1 queueing system the steady-state probability that after a service completion there are n customers in the system equals the steady-state probability $\pi_{\langle n \rangle}$ that there are n customers in the system at an arbitrary moment.

So we have

$$\pi_{\langle n \rangle} = \pi_{\langle 0 \rangle} k_n + \sum_{j=1}^{n+1} \pi_{\langle j \rangle} k_{n-j+1} \qquad n \geq 0 \tag{4.4}$$

[3] cf. [Gross/Harris 1974] and [Kleinrock 1975]

To calculate $\pi_{\langle 0\rangle}$, we first note that

$$\begin{aligned}
\mathrm{E}\left(k\right) = \sum_{n=0}^{\infty} n k_n &= \int\limits_{t=0}^{\infty} \mathrm{e}^{-\lambda t} \sum_{n=0}^{\infty} \frac{n(\lambda t)^n}{n!} \mathrm{d}F_S(t) \\
&= \int\limits_{t=0}^{\infty} \mathrm{e}^{-\lambda t} (\lambda t) \sum_{n=0}^{\infty} \frac{(\lambda t)^n}{n!} \mathrm{d}F_S(t) \\
&= \lambda \int\limits_{t=0}^{\infty} t \, \mathrm{d}F_S(t) = \lambda \, \mathrm{E}\left(S\right) = \frac{\lambda}{\mu}
\end{aligned} \tag{4.5}$$

Now we define the generating functions of $\langle \pi_{\langle n\rangle} \rangle_{n \geq 0}$ and $\langle k_0, k_1, k_2, \ldots \rangle$ by

$$\Pi(z) = \sum_{n=0}^{\infty} \pi_{\langle n\rangle} z^n \tag{4.6}$$

and

$$K(z) = \sum_{n=0}^{\infty} k_n z^n \tag{4.7}$$

Multiplying Equation 4.4 by z^n yields

$$\pi_{\langle n\rangle} z^n = \pi_{\langle 0\rangle} k_n z^n + \frac{1}{z} \sum_{j=0}^{n+1} \pi_{\langle j\rangle} k_{n-j+1} z^{n+1} - \frac{\pi_{\langle 0\rangle} k_{n+1} z^{n+1}}{z} \qquad i \geq 0 \tag{4.8}$$

and

$$\sum_{n=0}^{\infty} \pi_{\langle n\rangle} z^n = \Pi(z) = \pi_{\langle 0\rangle} K(z) + \frac{1}{z}\left(K(z)\Pi(z) - \pi_0 k_0\right) - \frac{\pi_{\langle 0\rangle}}{z}\left(K(z) - k_0\right) \tag{4.9}$$

So we have

$$\Pi(z) = \frac{\pi_{\langle 0\rangle}(1-z)K(z)}{K(z) - z} \tag{4.10}$$

With

$$\sum_{n=0}^{\infty} \pi_{\langle n\rangle} = \Pi(1) = 1 \tag{4.11}$$

$$\sum_{n=0}^{\infty} k_n = K(1) = 1 \tag{4.12}$$

and

$$K'(1) = \mathrm{E}\left(k\right) = \frac{\lambda}{\mu} \tag{4.13}$$

we get

$$\lim_{z\to 1}\Pi(z) = \lim_{z\to 1}\frac{\pi_{\langle 0\rangle}\left((1-z)K'(z)-K(z)\right)}{K'(z)-1} = \\ \frac{\pi_{\langle 0\rangle}K(1)}{1-\frac{\lambda}{\mu}} = \frac{\pi_{\langle 0\rangle}}{1-\frac{\lambda}{\mu}} = \Pi(1) = 1 \quad (4.14)$$

or

$$\pi_{\langle 0\rangle} = 1-\frac{\lambda}{\mu} \quad (4.15)$$

Therefore, the stationary state probabilities of an M/G/1 queueing system are

$$\pi_{\langle 0\rangle} = 1-\frac{\lambda}{\mu} \quad (4.16)$$

$$\pi_{\langle n+1\rangle} = \left(\pi_{\langle n\rangle} - \pi_{\langle 0\rangle}k_n + \sum_{j=1}^{n}\pi_{\langle j\rangle}k_{n-j+1}\right)\frac{1}{k_0} \qquad n \geq 0 \quad (4.17)$$

with

$$k_n = \int\limits_{t=0}^{\infty} \mathrm{e}^{-\lambda t}\frac{(\lambda t)^n}{n!}\,\mathrm{d}F_S(t) \quad (4.18)$$

4.2. Matrix geometric method

The matrix geometric method, which was developed by M. Neuts,[4] exploits regularities in the structure of Markov chains for the calculation of the stationary state probabilities.

To illustrate how it works, we show how to determine the stationary state probabilities of an M/Hypo/1 queueing system using the matrix geometric method.[5]

Let us first consider an M/M/1 queueing system. If the system is empty, customers arrive at rate λ^*, otherwise at rate λ. The service rate is always μ. Figure 4.2 shows the Markov chain for the system state of this queueing system.

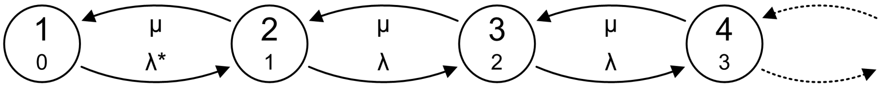


Figure 4.2.: M/M/1 queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system.

The transition rate matrix of this Markov chain is

$$\mathcal{Q} = \begin{pmatrix} -\lambda^* & \lambda^* & 0 & 0 & 0 & \cdots \\ \mu & -(\lambda+\mu) & \lambda & 0 & 0 & \cdots \\ 0 & \mu & -(\lambda+\mu) & \lambda & 0 & \cdots \\ 0 & 0 & \mu & -(\lambda+\mu) & \lambda & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{4.19}$$

We see that columns 4, 5, ... have the same structure as column 3, with the only difference that they are shifted down by $j-3$ steps, where j is the number of the column. We call equations arising from these columns the *repeating portion* of the process. The associated states ($\langle 2\rangle, \langle 3\rangle, \ldots$) are called *repeating states*. The remaining equations are called the *boundary portion* of the process, the associated states ($\langle 0\rangle$ and $\langle 1\rangle$) are called *boundary states*.

To calculate the stationary state probabilities $\pi = (\pi_1, \pi_2, \ldots)$, we need to find a solution to

$$\pi \cdot \mathcal{Q} = 0 \tag{4.20}$$

under the side condition

$$\sum_i \pi_i = 1 \tag{4.21}$$

[4] [Neuts 1981], [Neuts 1989]

[5] This example is taken from [Nelson 1991].

This is done with the following approach: We assume[6] that it is possible to express the state probability of the repeating states π_j as a function of π_{j-1}. This is reasonable, because the probability of each state depends only on the probabilities of its direct neighbours. Moreover, we assume that the function has the form

$$\pi_j = \pi_{j-1} \cdot \rho \qquad j \geq 3 \tag{4.22}$$

with an unknown constant ρ.

Therefore, the state probabilities $\pi_j, j \geq 3$ are assumed to have the form

$$\pi_j = \pi_2 \cdot \rho^{j-2} \qquad j \geq 3 \tag{4.23}$$

To determine ρ, we combine Equation 4.22 with the balance equation for the repeating portion of the process,

$$\pi_j(\lambda + \mu) = \pi_{j-1}\lambda + \pi_{j+1}\mu \qquad j \geq 3 \tag{4.24}$$

This leads to

$$\pi_{j-1}\rho(\lambda + \mu) = \pi_{j-1}\lambda + \pi_{j-1}\rho^2\mu \tag{4.25}$$

and

$$\lambda - \rho(\lambda + \mu) + \rho^2\mu = 0 \tag{4.26}$$

The two solutions of this quadratic equation are

$$\rho_1 = \frac{\lambda}{\mu} \qquad \rho_2 = 1 \tag{4.27}$$

ρ_2 would require $\pi_j = 0$ for all $j \geq 2$ to satisfy Equation 4.21, therefore

$$\rho = \rho_1 = \frac{\lambda}{\mu} \tag{4.28}$$

The equations for the boundary portion of the process are

$$-\pi_1\lambda^* + \pi_2\mu = 0 \tag{4.29}$$
$$\pi_1\lambda^* - \pi_2(\lambda + \mu) + \pi_3\mu = 0 \tag{4.30}$$

or in matrix form

$$(\pi_1, \pi_2)\begin{pmatrix} -\lambda^* & \lambda^* \\ \mu & -\mu \end{pmatrix} = 0 \tag{4.31}$$

[6] Of course, we know that this is true because in steady state $\pi_{j-1}\lambda = \pi_j\mu$ must hold. However, we "guess" the form of the solution because we will need to do so when we use the matrix geometric method with the M/Hypo/1 queueing system.

Together with Equation 4.21, which states that

$$1 = \pi_1 + \pi_2 \sum_{j=0}^{\infty} \rho^j = \pi_1 + \pi_2 \frac{1}{1-\rho} \tag{4.32}$$

we get the solution

$$\pi_1 = \frac{\mu - \lambda}{\mu - \lambda + \lambda^*} \tag{4.33}$$

$$\pi_2 = \frac{\lambda^*(\mu - \lambda}{\mu(\mu - \lambda + \lambda^*)} \tag{4.34}$$

Finally we have to prove that the assumption in Equation 4.22 is correct. This is done by showing that the solution (Equations 4.23, 4.33 and 4.34), which must be unique, satisfies Equations 4.20 and 4.21.

Now we analyse the M/Hypo/1 queueing system.

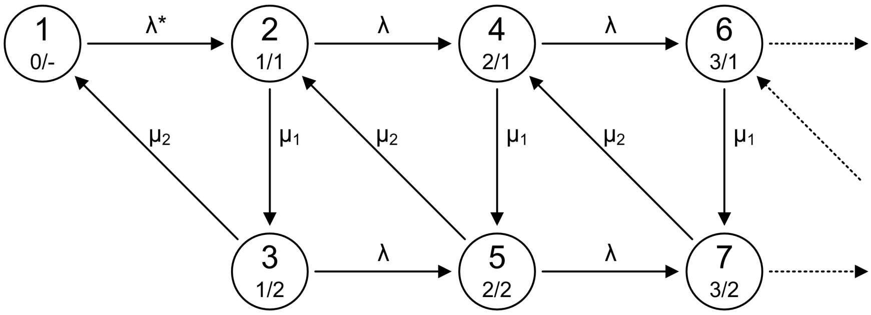


Figure 4.3.: M/Hypo/1 queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the service process.

Figure 4.3 shows the Markov chain for the system state. The transition rate matrix of this Markov chain is

$$\mathcal{Q} = \begin{pmatrix} -\lambda^* & \lambda^* & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\ 0 & -a_1 & \mu_1 & \lambda & 0 & 0 & 0 & 0 & 0 & \cdots \\ \mu_2 & 0 & -a_2 & 0 & \lambda & 0 & 0 & 0 & 0 & \cdots \\ 0 & 0 & 0 & -a_1 & \mu_1 & \lambda & 0 & 0 & 0 & \cdots \\ 0 & \mu_2 & 0 & 0 & -a_2 & 0 & \lambda & 0 & 0 & \cdots \\ 0 & 0 & 0 & 0 & 0 & -a_1 & \mu_1 & \lambda & 0 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{4.35}$$

where $a_1 = \lambda + \mu_1$ and $a_2 = \lambda + \mu_2$.

If we group the elements of $\mathcal{Q}$ into blocks,

$$\mathcal{Q} = \left(\begin{array}{ccc|cc|cc|cc|c}
-\lambda^* & \lambda^* & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdots \\
0 & -a_1 & \mu_1 & \lambda & 0 & 0 & 0 & 0 & 0 & \cdots \\
\mu_2 & 0 & -a_2 & 0 & \lambda & 0 & 0 & 0 & 0 & \cdots \\
\hline
0 & 0 & 0 & -a_1 & \mu_1 & \lambda & 0 & 0 & 0 & \cdots \\
0 & \mu_2 & 0 & 0 & -a_2 & 0 & \lambda & 0 & 0 & \cdots \\
\hline
0 & 0 & 0 & 0 & 0 & -a_1 & \mu_1 & \lambda & 0 & \cdots \\
0 & 0 & 0 & \mu_2 & 0 & 0 & -a_2 & 0 & \lambda & \cdots \\
\hline
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}\right) \tag{4.36}$$

and set

$$\begin{aligned}
\mathcal{L}^* &= \begin{pmatrix} -\lambda^* & \lambda^* & 0 \\ 0 & -a_1 & \mu_1 \\ \mu_2 & 0 & -a_2 \end{pmatrix} & \mathcal{F}^* &= \begin{pmatrix} 0 & 0 \\ \lambda & 0 \\ 0 & \lambda \end{pmatrix} \\
\mathcal{B}^* &= \begin{pmatrix} 0 & 0 & 0 \\ 0 & \mu_2 & 0 \end{pmatrix} \\
\mathcal{L} &= \begin{pmatrix} -a_1 & \mu_1 \\ 0 & -a_2 \end{pmatrix} & \mathcal{F} &= \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix} \\
\mathcal{B} &= \begin{pmatrix} 0 & 0 \\ \mu_2 & 0 \end{pmatrix}
\end{aligned} \tag{4.37}$$

(The letters “L”, “B” and “F” relate to the type of the transitions – “local”, “backward” and “forward”.)

we get a new matrix

$$\mathcal{Q} = \begin{pmatrix}
\mathcal{L}^* & \mathcal{F}^* & 0 & 0 & 0 & \cdots \\
\mathcal{B}^* & \mathcal{L} & \mathcal{F} & 0 & 0 & \cdots \\
0 & \mathcal{B} & \mathcal{L} & \mathcal{F} & 0 & \cdots \\
0 & 0 & \mathcal{B} & \mathcal{L} & \mathcal{F} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{pmatrix} \tag{4.38}$$

which has the same structure as the matrix in Equation 4.19. The only difference is that scalar entries have been replaced with matrix entries. Therefore, we will try to calculate the stationary state probabilities in the same manner as in the M/M/1 example, except that we deal with vectors and matrices instead of scalars.

Again we call equations that arise from columns 3, 4, . . . of $\mathcal{Q}$ the *repeating portion* of the process and equations arising from columns 1 and 2 the *boundary portion*. States 1, . . . , 5 are called *boundary states*, states 6, 7, . . . are are called *repeating states*.

We assume that it is possible to express the probabilities of repeating states $\left(\pi_{\langle j,1\rangle}, \pi_{\langle j,2\rangle}\right)$ as a function of $\left(\pi_{\langle j-1,1\rangle}, \pi_{\langle j-1,2\rangle}\right)$ and we assume that the function has the form

$$\left(\pi_{\langle j,1\rangle}, \pi_{\langle j,2\rangle}\right) = \left(\pi_{\langle j-1,1\rangle}, \pi_{\langle j-1,2\rangle}\right)\mathcal{R} \qquad j \geq 3 \tag{4.39}$$

with an unknown constant matrix $\mathcal{R}$.

Therefore, the state probabilities $\left(\pi_{\langle j,1\rangle}, \pi_{\langle j,2\rangle}\right), j \geq 3$ are assumed to have the form

$$\left(\pi_{\langle j,1\rangle}, \pi_{\langle j,2\rangle}\right) = \left(\pi_{\langle 2,1\rangle}, \pi_{\langle 2,2\rangle}\right) \mathcal{R}^{j-2} \qquad j \geq 3 \tag{4.40}$$

To determine $\mathcal{R}$, we combine Equation 4.39 with the balance equation for the repeating portion of the process,

$$\left(\pi_{\langle j-1,1\rangle}, \pi_{\langle j-1,2\rangle}\right) \mathcal{F} + \left(\pi_{\langle j,1\rangle}, \pi_{\langle j,2\rangle}\right) \mathcal{L} + \left(\pi_{\langle j+1,1\rangle}, \pi_{\langle j+1,2\rangle}\right) \mathcal{B} = 0 \qquad j \geq 3 \tag{4.41}$$

This leads to

$$\begin{aligned}\left(\pi_{\langle j-1,1\rangle}, \pi_{\langle j-1,2\rangle}\right) \mathcal{F} + \left(\pi_{\langle j-1,1\rangle}, \pi_{\langle j-1,2\rangle}\right) \mathcal{R}\mathcal{L} + \\ \left(\pi_{\langle j-1,1\rangle}, \pi_{\langle j-1,2\rangle}\right) \mathcal{R}^2\mathcal{B} = 0 \qquad j \geq 3\end{aligned} \tag{4.42}$$

and

$$\mathcal{F} + \mathcal{R}\mathcal{L} + \mathcal{R}^2\mathcal{B} = 0 \tag{4.43}$$

As in the M/M/1 example, there can be more than one solution to this quadratic equation. We need to find the minimal non-negative solution.

There are a number of algorithms for calculating $\mathcal{R}$. For example, we could set

$$\mathcal{R}^{(0)} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix} \tag{4.44}$$

and repeat the iteration

$$\mathcal{R}^{(n+1)} = -\left(\mathcal{F} + {\mathcal{R}^{(n)}}^2\mathcal{B}\right)\mathcal{L}^{-1} \tag{4.45}$$

until $|\mathcal{R}^{(n+1)} - \mathcal{R}^{(n)}|$ becomes sufficiently small.

The equations for the boundary portion of the process are

$$\left(\pi_{\langle 0/-\rangle}, \pi_{\langle 1,1\rangle}, \pi_{\langle 1,2\rangle}\right) \mathcal{L}^* + \left(\pi_{\langle 2,1\rangle}, \pi_{\langle 2,2\rangle}\right) \mathcal{B}^* = 0 \tag{4.46}$$

$$\left(\pi_{\langle 0/-\rangle}, \pi_{\langle 1,1\rangle}, \pi_{\langle 1,2\rangle}\right) \mathcal{F}^* + \left(\pi_{\langle 2,1\rangle}, \pi_{\langle 2,2\rangle}\right) \mathcal{L} + \left(\pi_{\langle 3,1\rangle}, \pi_{\langle 3,2\rangle}\right) \mathcal{B} = 0 \tag{4.47}$$

or in matrix form

$$\left(\left(\pi_{\langle 0/-\rangle}, \pi_{\langle 1,1\rangle}, \pi_{\langle 1,2\rangle}\right), \left(\pi_{\langle 2,1\rangle}, \pi_{\langle 2,2\rangle}\right)\right) \begin{pmatrix} \mathcal{L}^* & \mathcal{F}^* \\ \mathcal{B}^* & \mathcal{L} + \mathcal{R}\mathcal{B} \end{pmatrix} \tag{4.48}$$

Together with Equation 4.21, which states that

$$1 = \pi_{\langle 0/-\rangle} + \pi_{\langle 1,1\rangle} + \pi_{\langle 1,2\rangle} + \left(\pi_{\langle 2,1\rangle}, \pi_{\langle 2,2\rangle}\right) \sum_{j=0}^{\infty} \mathcal{R}^j \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{4.49}$$

$$= \pi_{\langle 0/-\rangle} + \pi_{\langle 1,1\rangle} + \pi_{\langle 1,2\rangle} + \left(\pi_{\langle 2,1\rangle}, \pi_{\langle 2,2\rangle}\right) (\mathcal{I} - \mathcal{R})^{-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \tag{4.50}$$

we can find a solution for $\left(\pi_{\langle 0/-\rangle}, \pi_{\langle 1,1\rangle}, \pi_{\langle 1,2\rangle}\right)$ and $\left(\pi_{\langle 2,1\rangle}, \pi_{\langle 2,2\rangle}\right)$.

Finally, we have to prove that the assumption in Equation 4.39 is correct. This is done by showing that the solution satisfies Equations 4.20 and 4.21.

The matrix geometric method can be used to determine the stationary state probabilities of all quasi-birth-death processes (QBD), such as the state of PH/PH/n queueing systems. In general, the transition rate matrices of QBD processes have the form

$$\mathcal{Q} = \begin{pmatrix} \mathcal{L}^* & \mathcal{F}^* & 0 & 0 & 0 & \cdots \\ \mathcal{B}^* & \mathcal{L} & \mathcal{F} & 0 & 0 & \cdots \\ 0 & \mathcal{B} & \mathcal{L} & \mathcal{F} & 0 & \cdots \\ 0 & 0 & \mathcal{B} & \mathcal{L} & \mathcal{F} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{4.51}$$

Moreover, GI/M/1-type Markov chains can be analysed. GI/M/1-type Markov chains are Markov chains which are a matrix generalisation of the embedded Markov chains of GI/M/1 queueing systems. These Markov chains have transition rate matrices of lower block Hessenberg form,

$$\mathcal{Q} = \begin{pmatrix} \mathcal{L}^* & \mathcal{F}^* & 0 & 0 & 0 & \cdots \\ \mathcal{B}_1^* & \mathcal{L} & \mathcal{F} & 0 & 0 & \cdots \\ \mathcal{B}_2^* & \mathcal{B}_1 & \mathcal{L} & \mathcal{F} & 0 & \cdots \\ \mathcal{B}_3^* & \mathcal{B}_2 & \mathcal{B}_1 & \mathcal{L} & \mathcal{F} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{4.52}$$

and the matrix $\mathcal{R}$ is a solution to

$$\mathcal{F} + \mathcal{R}\mathcal{L} + \sum_{k=1}^{\infty} \mathcal{R}^{k+1}\mathcal{B}_k = 0 \tag{4.53}$$

4.3. Matrix analytic methods

For M/G/1-type Markov chains (Markov chains which are a matrix generalisation of the embedded Markov chains of M/G/1 queueing systems), there is no geometric relation between the probability vectors. Therefore, they cannot be analysed using the matrix geometric method. Instead, matrix analytic techniques are needed.

As an example for an M/G/1-type Markov chain, let us consider a BMAP/Cox/1 queueing system.[7] In this system, the interarrival times are exponentially distributed with rate 2λ, and customers arrive in batches of size n with probability $1/2^n$. The service times are Coxian distributed with parameters (μ_1, α_1, μ_2). Figure 4.4 shows the Marov chain for the system state.

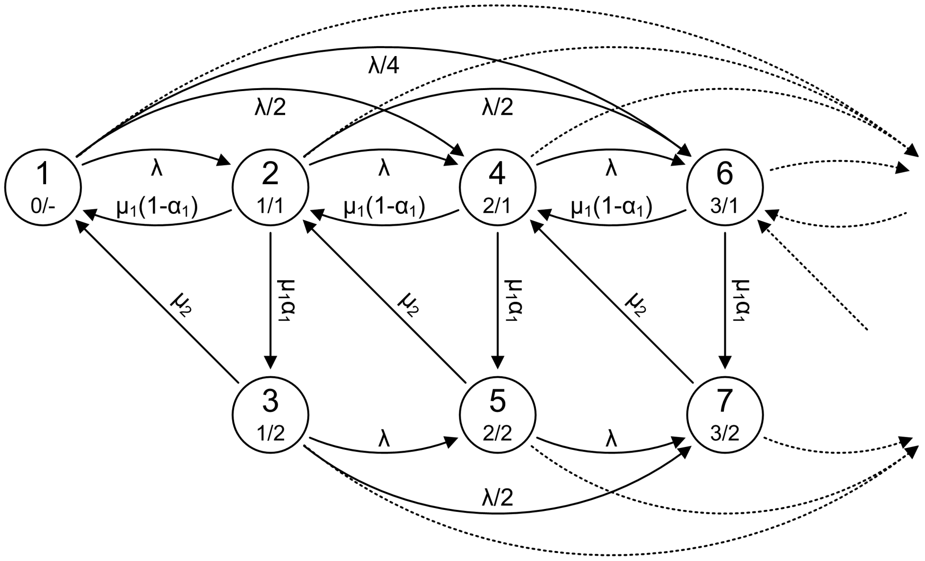


Figure 4.4.: Markov chain for the system state of a BMAP/Cox/1 queueing system.

The transition rate matrix is

$$\mathcal{Q} = \left(\begin{array}{c|cc|cc|cc|c} -2\lambda & \lambda & 0 & \lambda/2 & 0 & \lambda/4 & 0 & \cdots \\ \hline \nu & -a_1 & \mu_1\alpha_1 & \lambda & 0 & \lambda/2 & 0 & \cdots \\ \mu_2 & 0 & -a_2 & 0 & \lambda & 0 & \lambda/2 & \cdots \\ \hline 0 & \nu & 0 & -a_1 & \mu_1\alpha_1 & \lambda & 0 & \cdots \\ 0 & \mu_2 & 0 & 0 & -a_2 & 0 & \lambda & \cdots \\ \hline 0 & 0 & 0 & \nu & 0 & -a_1 & \mu_1\alpha_1 & \cdots \\ 0 & 0 & 0 & \mu_2 & 0 & 0 & -a_2 & \cdots \\ \hline \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{array} \right) \qquad (4.54)$$

with $\nu = \mu_1(1-\alpha_1)$, $a_1 = -2\lambda - \mu_1$ and $a_2 = -2\lambda - \mu_2$.

[7] A MAP (Markovian Arrival Process) is an arrival process where arrivals are associated with transitions in an underlying Markov chain. Examples for MAPs are Poisson process, Markov-modulated Poisson process (MMPP) and the point processes discussed in Section 3.5. A BMAP (Batch Markovian Arrival Process) is a MAP where customers can arrive in batches.

If we group the elements of $\mathcal{Q}$ into blocks, and substitute

$$\mathcal{L}^* = \begin{pmatrix} -2\lambda \end{pmatrix} \qquad \mathcal{F}_i^* = \begin{pmatrix} \lambda/2^{i-1} & 0 \end{pmatrix} \tag{4.55}$$

$$\mathcal{B}^* = \begin{pmatrix} \mu_1(1-\alpha_1) \\ \mu_2 \end{pmatrix} \tag{4.56}$$

$$\mathcal{L} = \begin{pmatrix} -2\lambda - \mu_1 & \mu_1\alpha_1 \\ 0 & -2\lambda - \mu_2 \end{pmatrix} \qquad \mathcal{F}_i = \begin{pmatrix} \lambda/2^{i-1} & 0 \\ 0 & \lambda/2^{i-1} \end{pmatrix} \tag{4.57}$$

$$\mathcal{B} = \begin{pmatrix} \mu_1(1-\alpha_1) & 0 \\ \mu_2 & 0 \end{pmatrix} \tag{4.58}$$

we get a matrix in upper block Hessenberg form:

$$\mathcal{Q} = \begin{pmatrix} \mathcal{L}^* & \mathcal{F}_1^* & \mathcal{F}_2^* & \mathcal{F}_3^* & \cdots \\ \mathcal{B}^* & \mathcal{L} & \mathcal{F}_1 & \mathcal{F}_2 & \cdots \\ 0 & \mathcal{B} & \mathcal{L} & \mathcal{F}_1 & \cdots \\ 0 & 0 & \mathcal{B} & \mathcal{L} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{4.59}$$

We will not discuss how a solution can be found, but it should be mentioned that in general an important step is to calculate an auxiliary matrix $\mathcal{G}$, which is the solution to

$$\mathcal{B} + \mathcal{L}\mathcal{G} + \sum_{i=1}^{\infty} \mathcal{F}_i \mathcal{G}^{i+1} = 0 \tag{4.60}$$

Further information about matrix analytic methods can be found in [Riska/Smirni 2002] (which also contains the solution to the BMAP/Cox/1 queueing system) and in [Lucantoni 1993] and [Neuts 1989].

4.4. Regenerative method

The regenerative method, introduced in [Hordijk/Tijms 1976], considers regeneration cycles of the the queueing process. Within a regeneration cycle, quantities, which are strongly related to the stationary state probabilities, are described in two different ways using sophisticated probabilistic arguments. By combining the two resulting sets of equations, the quantities and thus the stationary state probabilities can be determined.

We demonstrate the regenerative method by calculating the stationary state probabilities of an M/G/1 queueing system.[8] Customers arrive according to a Poisson process with rate λ, the service times S have a general distribution with rate μ and cumulative distribution function $F_S(t)$.

As a regeneration cycle, we can use a busy cycle consisting of the idle period and the busy period (Figure 4.5).

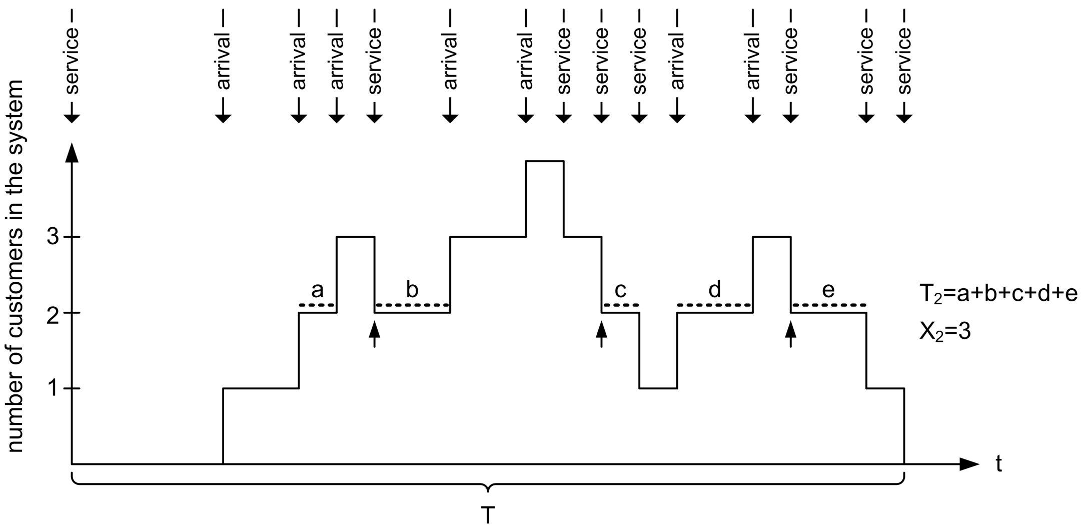


Figure 4.5.: Number of customers in an M/G/1 queueing system during a busy cycle.

First, we define the following variables:

T ... length of the busy cycle. We assume that the busy cycle begins at time 0.

T_n ... amount of time in $[0, T]$ that there are n customers in the system. Due to the Poisson arrivals, the length of the idle period T_0 is exponentially distributed with rate λ, therefore $\mathrm{E}(T_0) = 1/\lambda$.

ξ ... number of customers served during the busy cycle.

X_j ... number of services in a busy cycle after which there are n customers in the system. There is always only one service after which there are no customers in the system (this is the last service in the busy cycle), therefore $X_0 = 1$.

[8] cf. [van Hoorn 1983] and [Tijms 1986]

Our goal is to find two sets of equations that describe the relation between $\mathrm{E}(T_n)$ and $\mathrm{E}(X_n)$. By combining these two sets of equations, we can then determine $\mathrm{E}(T_n)$ and $\mathrm{E}(X_n)$.

Let state $\langle n\rangle$ be the state when there are n customers in the queueing system.

For the first set of relations, we consider transitions between state $\langle n\rangle$ and state $\langle n+1\rangle$.

A transition from state $\langle n\rangle$ to state $\langle n+1\rangle$ occurs at rate λ when there are n customers in the system, so we have on average $\lambda\,\mathrm{E}(T_n)$ transitions from state $\langle n\rangle$ to state $\langle n+1\rangle$ during a busy cycle.

A transition from state $\langle n+1\rangle$ to state $\langle n\rangle$ takes place when there are $n+1$ customers in the system and a service is finished, after which there are n customers in the system. The expected number of such services in a busy cycle is $\mathrm{E}(X_n)$.

Since both at the beginning and at the end of the busy cycle the queueing system is empty and we do not have batch arrivals or services, the number of transitions from state $\langle n\rangle$ to state $\langle n+1\rangle$ must equal the number of transitions from state $\langle n+1\rangle$ to state $\langle n\rangle$. Therefore, we have

$$\lambda\,\mathrm{E}(T_n) = \mathrm{E}(X_n) \qquad n \geq 0 \tag{4.61}$$

To derive the second set of equations, we first note that the number of services at whose beginning there are $j \geq 2$ customers in the system equals X_n, the number of services after whose completion there are j customers in the system. Since at the beginning of the first service in the busy cycle there is one customer in the system and this service does not follow another service, the number of services at whose beginning there is one customer in the system is $X_1 + 1$.

Now assume that there are j customers in the system when a service begins. While this service is in progress, there will be j or more customers in the system. That is during the service the system can spend time in states $\langle j\rangle$, $\langle j+1\rangle$, $\langle j+2\rangle$, …

Let $A_{jn}(j, n \geq 1)$ be the amount of time during which there are n customers in the system until the current service is completed, given that there are j customers in the system when the current service has begun (Figure 4.6).

For the calculation of A_{jn}, we define the indicator function $\chi_{jn}(t)$. Given that at the beginning of the service under consideration there were j customers in the system, $\chi_{jn}(t) = 1$ if at time t the service is still in progress and there are exactly n customers in the system. Otherwise, $\chi_{jn}(t) = 0$.

$$\begin{aligned} \mathrm{P}\{\chi_{jn}(t)=1\} = \mathrm{E}(\chi_{jn}(t)) = \mathrm{P}\{S > t, n-j \text{ arrivals until time } t\} = \\ \mathrm{P}\{S>t\}\,\mathrm{P}\{n-j \text{ arrivals until time } t\} = (1-F_S(t))\,\mathrm{e}^{-\lambda t}\frac{(\lambda t)^{n-j}}{(n-j)!} \end{aligned} \tag{4.62}$$

Now we have

$$A_{jn} = \int\limits_{t=0}^{\infty} \mathrm{E}(\chi_{jn}(t))\,\mathrm{d}t = \int\limits_{t=0}^{\infty} (1-F_S(t))\,\mathrm{e}^{-\lambda t}\frac{(\lambda t)^{n-j}}{(n-j)!}\mathrm{d}t \tag{4.63}$$

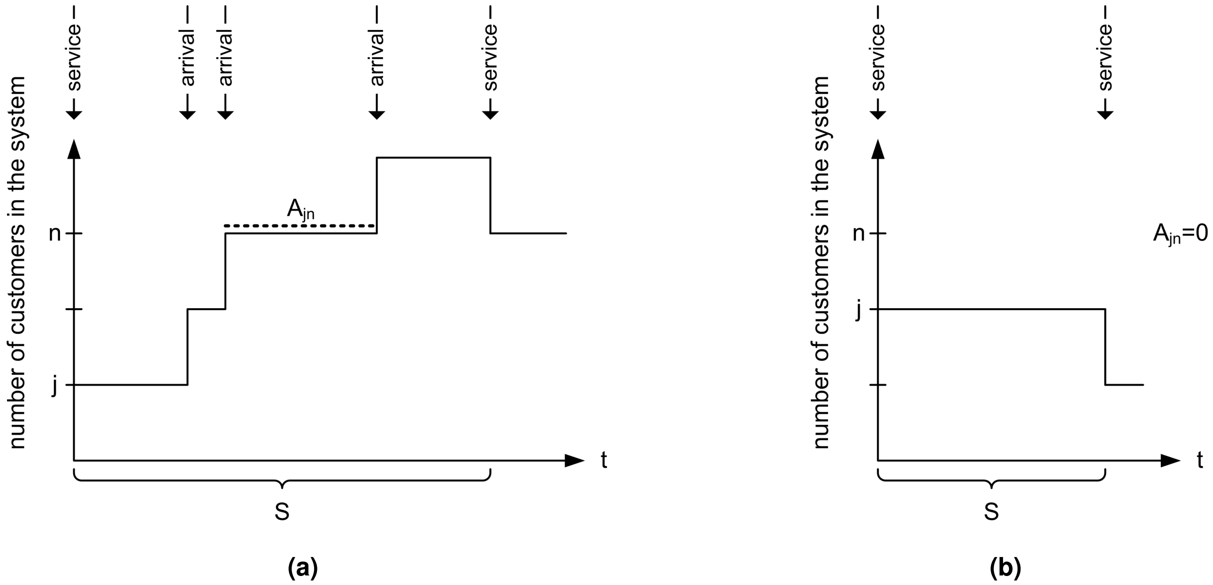


Figure 4.6.: Number of customers in an M/G/1 queueing system during a service. When the service begins, there are j customers in the system. (a) During the service there are arrivals, (b) during the service there are no arrivals.

Considering only services starting with j customers in the system, the time T_n is a compound random variable consisting of A_{jn} and of the number of services X_j after which there are j customers in the system.

Therefore, we have

$$\mathrm{E}\left(T_n\right) = A_{1,n} + \sum_{j=1}^{n} A_{jn}\,\mathrm{E}\left(X_j\right) \qquad n \geq 1 \tag{4.64}$$

From Equations 4.61 and 4.64 follows the solution

$$\mathrm{E}\left(T_0\right) = \frac{1}{\lambda} \tag{4.65}$$

$$\mathrm{E}\left(X_0\right) = 1 \tag{4.66}$$

$$\mathrm{E}\left(T_n\right) = \frac{A_{1n} + \sum_{j=1}^{n-1} A_{jn}\,\mathrm{E}\left(X_j\right)}{1 - \lambda A_{nn}} \qquad n \geq 1 \tag{4.67}$$

$$\mathrm{E}\left(X_n\right) = \lambda\,\mathrm{E}\left(T_n\right) \qquad n \geq 1 \tag{4.68}$$

It remains to calculate the length of the busy cycle T.

The length of the busy period $T - T_0$ is a compound random variable consisting of the number of customers served during the busy period and the service time for each customer. Therefore, we have

$$\mathrm{E}\left(T\right) - \mathrm{E}\left(T_0\right) = \mathrm{E}\left(\xi\right)\frac{1}{\mu} \tag{4.69}$$

Moreover, the number of customers served during a busy cycle equals the number of customers arriving during the busy cycle,

$$\mathrm{E}\left(\xi\right) = \mathrm{E}\left(T\right) \cdot \lambda \tag{4.70}$$

From these two equations, we can express the length of the idle period as

$$\mathrm{E}\left(T_0\right) = \mathrm{E}\left(T\right)\left(1 - \frac{\lambda}{\mu}\right) \tag{4.71}$$

Together with $\mathrm{E}\left(T_0\right) = 1/\lambda$ we get for the length of the busy cycle

$$\mathrm{E}\left(T\right) = \frac{1}{\lambda\left(1 - \frac{\lambda}{\mu}\right)} \tag{4.72}$$

Now we can calculate the steady-state probability $\pi_{\langle n\rangle}$ that there are n customers in the queueing system. Since $\pi_{\langle n\rangle}$ equals the long-run proportion of time that there are n customers in the system and since a regeneration cycle has the same statistical characteristics as the whole process, we have

$$\pi_{\langle n\rangle} = \frac{\mathrm{E}\left(T_n\right)}{\mathrm{E}\left(T\right)} \tag{4.73}$$

Part II.

Advanced Markov chain techniques

5. Idle and busy period

In this chapter we show how the idle and the busy period of queueing systems (Figure 5.1) can be analysed using Markov chains.

In Sections 5.1 and 5.2, we calculate the length of the idle and the busy period. In Section 5.3, we determine the number of customers that are served during the busy period.

To simplify the explanations, we restrict ourselves to single-server queueing systems. However, the techniques shown work just as well with multi-server queueing systems.

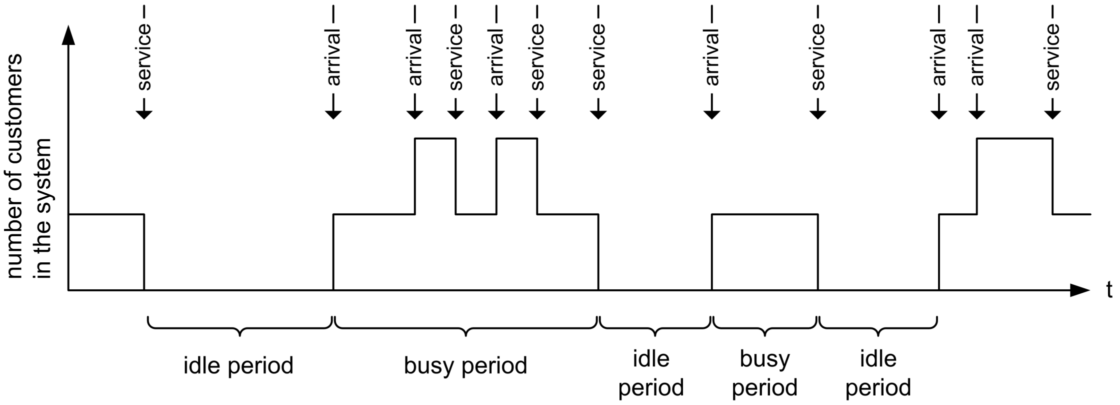


Figure 5.1.: Idle and busy period of a single-server queueing system.

In the following discussion, we call states of the Markov chain for the system state that correspond to states of the queueing system where the server is idle *idle states*, and states of the Markov chain for the system state corresponding to states of the queueing system where the server is busy *busy states*. The set of the idle states is denoted with $\mathcal{I}$, the set of the busy states is denoted with $\mathcal{B}$.

5.1. Length of the idle period

The idle period is the time frame in which there are no customers in the queueing systems and, therefore, the server is idle. The idle period begins when the last customer leaves the system, and it ends, when the next customer arrives.

To calculate the length I of the idle period we identify (by means of the Markov chain for the system state $\mathcal{M}_S$) the states in which the Markov chain can be when the idle period begins (that is, the states in which the Markov chain can be after the last customer has left the system). The probability σ_j^I that the Markov chain is in state j when the idle period begins is the ratio of the rate at which state j is entered from busy states to the rate at which all idle states are entered from busy states:

$$\sigma_j^I = \frac{\sum\limits_{b\in\mathcal{B}} \pi_b q_{bj}}{\sum\limits_{b\in\mathcal{B}} \pi_b \sum\limits_{i\in\mathcal{I}} q_{bi}} \qquad j \in \mathcal{I} \tag{5.1}$$

Next we create a new Markov chain $\mathcal{M}_I$ by removing all transitions of $\mathcal{M}_S$ that do not originate in an idle state. That means, in $\mathcal{M}_I$, all busy states are absorbing states. With this Markov chain, we calculate $\varphi_i(\cdot)$, the complementary cumulative distribution function of the remaining length of the idle period (that is, the time needed to reach a busy state) given that the Markov chain is in state i:

$$\varphi_i(0) = \begin{cases} 1 & \text{if } i \text{ is an idle state} \\ 0 & \text{if } i \text{ is a busy state} \end{cases} \tag{5.2}$$

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{5.3}$$

The cumulative distribution function of the length of the idle period is then

$$\mathrm{P}(I \le t) = 1 - \sum_j \sigma_j^I \cdot \varphi_j(t) \tag{5.4}$$

Often it is possible to express the length of the idle period as a weighted sum of hypoexponential random variables[1]:

$$I \sim \sigma_1^I \,\mathrm{HypoExp}(\cdots) + \sigma_2^I \,\mathrm{HypoExp}(\cdots) + \dots \tag{5.5}$$

This approach should be preferred, because since we can use closed-form solutions (instead of numerical approximation algorithms), the achieved solution will be more accurate.

5.1.1. M/M/1/S queueing system

Figure 5.2a shows the Markov chain for the system state of an M/M/1/S queueing system with $S = 3$.

[1] In this work we will use the following notation: Let $X \sim \mathrm{A}(a_1, a_2, \dots)$ (the random variable X is distributed according to the distribution A with parameters $a_1, a_2, \dots$) and $Y \sim \mathrm{B}(b_1, b_2, \dots)$. Then we define: $Z = c_1 X + c_2 Y \Leftrightarrow Z \sim c_1 \,\mathrm{A}(a_1, a_2, \dots) + c_2 \,\mathrm{B}(b_1, b_2, \dots) \Leftrightarrow f_Z(\tau) = c_1 f_X(\tau) + c_2 f_Y(\tau)$, where $f_i(\tau)$ are the probability density functions of the random variables X, Y and Z.

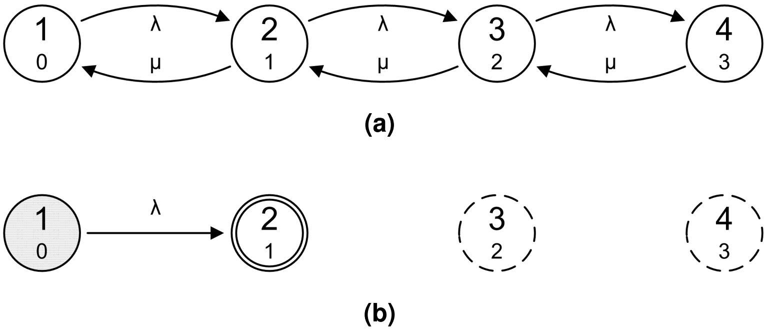


Figure 5.2.: M/M/1/S queueing system ($S = 3$). (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the idle period. Meaning of the names of the states: number of customers in the system.

The only idle state is state $\langle 0 \rangle$, therefore,

$$\sigma^I_{\langle 0 \rangle} = 1 \tag{5.6}$$

The Markov chain for the calculation of the length of the idle period is shown in Figure 5.2b. All transitions of $\mathcal{M}_S$ have been removed, except for the transition $\langle 0 \rangle \rightarrow \langle 1 \rangle$, which originates in an idle state.

Since the idle period always begins in state $\langle 0 \rangle$ and it needs only one transition with rate λ to reach a busy state, the idle period is exponentially distributed with mean $1/\lambda$:

$$I \sim \mathrm{Exp}(\lambda) \tag{5.7}$$

$$\mathrm{E}(I) = \frac{1}{\lambda} \tag{5.8}$$

5.1.2. Hypo/M/1/S queueing system

The Markov chain for a Hypo/M/1/S queueing system is shown in Figure 5.3a. Now we have two states that represent an empty system: state $\langle 0/1 \rangle$ and state $\langle 0/2 \rangle$. The Markov chain can be in either of them when the idle period begins. If it was in state $\langle 1/1 \rangle$ before the last service finished (probability $\pi_{\langle 1/1 \rangle} / \left(\pi_{\langle 1/1 \rangle} + \pi_{\langle 1/2 \rangle}\right)$), it is in state $\langle 0/1 \rangle$ when the idle period begins. If it was in state $\langle 1/2 \rangle$ (probability $\pi_{\langle 1/2 \rangle} / \left(\pi_{\langle 1/1 \rangle} + \pi_{\langle 1/2 \rangle}\right)$), it is in state $\langle 0/2 \rangle$ when the idle period begins. Therefore, we have

$$\sigma^I_{\langle 0/1 \rangle} = \frac{\pi_{\langle 1/1 \rangle}}{\pi_{\langle 1/1 \rangle} + \pi_{\langle 1/2 \rangle}} \tag{5.9}$$

$$\sigma^I_{\langle 0/2 \rangle} = \frac{\pi_{\langle 1/2 \rangle}}{\pi_{\langle 1/1 \rangle} + \pi_{\langle 1/2 \rangle}} \tag{5.10}$$

The Markov chain for the calculation of the length of the idle period is shown in Figure 5.3b. We see that if the Markov chain is in state $\langle 0/1 \rangle$ when the idle period begins, it needs two transitions (with rates λ_1 and λ_2) to reach a busy state. That means, in this case, the idle period is hypoexponentially distributed. If the Markov chain is in state $\langle 0/2 \rangle$ when the idle period begins, it needs only one transition (with rate λ_2) to reach a

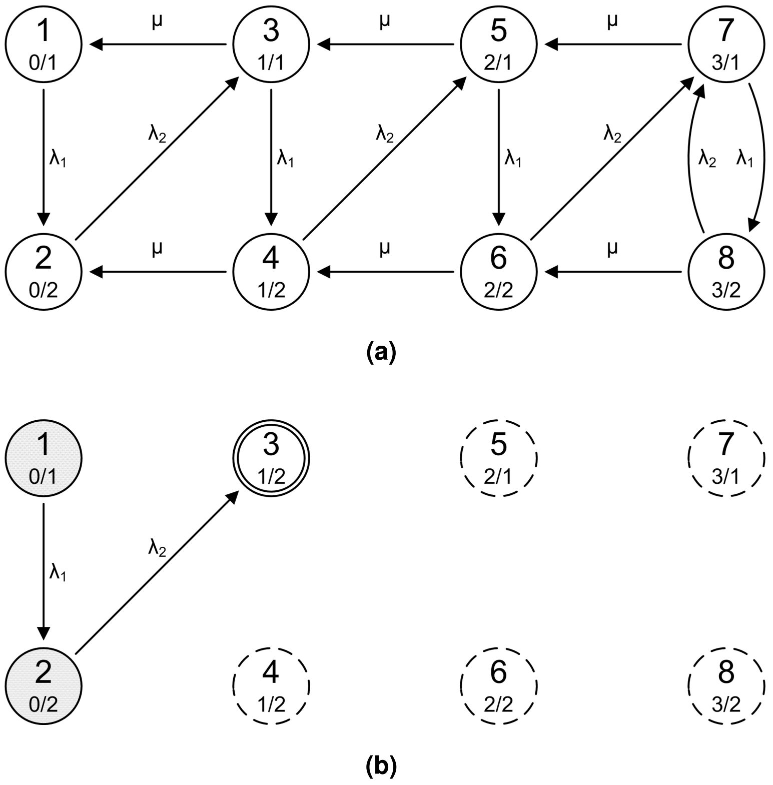


Figure 5.3.: Hypo/M/1/S queueing system $(S = 3)$. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the idle period. Meaning of the names of the states: number of customers in the system / state of the arrival process.

busy state, which means the idle period would be exponentially distributed. Therefore, we can express the length of the idle period as a weighted sum of an exponential and a hypoexponential distribution:

$$I \sim \frac{\pi_{\langle 1/1\rangle}}{\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}} \operatorname{HypoExp}(\lambda_1, \lambda_2) + \frac{\pi_{\langle 1/2\rangle}}{\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}} \operatorname{Exp}(\lambda_2) \tag{5.11}$$

The mean length of the idle period is

$$\mathrm{E}(I) = \frac{\pi_{\langle 1/1\rangle}}{\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}} \frac{1}{\lambda_1} + \frac{1}{\lambda_2} \tag{5.12}$$

5.1.3. Hyper/M/1/S queueing system

Figure 5.4a shows the Markov chain for the system state of a Hyper/M/1/S queueing system. Again there are two states in which the Markov chain can be when the idle period begins: state $\langle 0/1\rangle$ and state $\langle 0/2\rangle$. If the Markov chain was in state $\langle 1/1\rangle$ before the last customer left the system (probability $\pi_{\langle 1/1\rangle} / \left(\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}\right)$), it is state $\langle 0/1\rangle$ when the idle period begins. If it was in state $\langle 1/2\rangle$ (probability $1 - \pi_{\langle 1/1\rangle} / \left(\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}\right) = \pi_{\langle 1/2\rangle} / \left(\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}\right)$), it is in state $\langle 0/2\rangle$ when the idle period begins. Therefore, we have

$$\sigma^I_{\langle 0/1\rangle} = \frac{\pi_{\langle 1/1\rangle}}{\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}} \tag{5.13}$$

$$\sigma^I_{\langle 0/2\rangle} = \frac{\pi_{\langle 1/2\rangle}}{\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}} \tag{5.14}$$

The Markov chain for the calculation of the length of the idle period in a Hyper/M/1/S queueing system is shown in Figure 5.4b. Since $\alpha_1 + \alpha_2$ equals 1, the length of the idle period, given that the Markov chain is in state 1 or in state 2 when the idle period begins, is exponentially distributed.

The length of the idle period is the weighted sum of these two exponential distributions:

$$I \sim \frac{\pi_{\langle 1/1\rangle}}{\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}} \operatorname{Exp}(\lambda_1) + \frac{\pi_{\langle 1/2\rangle}}{\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}} \operatorname{Exp}(\lambda_2) = \operatorname{HyperExp}\left(\lambda_1, \frac{\pi_{\langle 1/1\rangle}}{\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}}, \lambda_2, \frac{\pi_{\langle 1/2\rangle}}{\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}}\right) \tag{5.15}$$

Its mean is

$$\mathrm{E}(I) = \frac{\pi_{\langle 1/1\rangle}}{\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}} \frac{1}{\lambda_1} + \frac{\pi_{\langle 1/2\rangle}}{\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}} \frac{1}{\lambda_2} \tag{5.16}$$

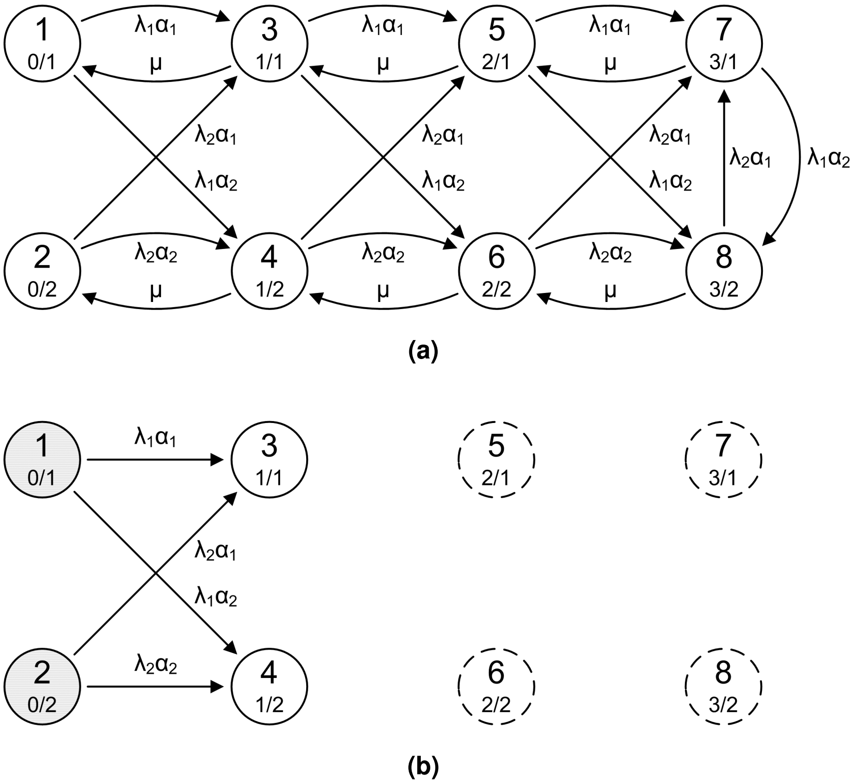


Figure 5.4.: Hyper/M/1/S queueing system $(S = 3)$. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the idle period. Meaning of the names of the states: number of customers in the system / state of the arrival process.

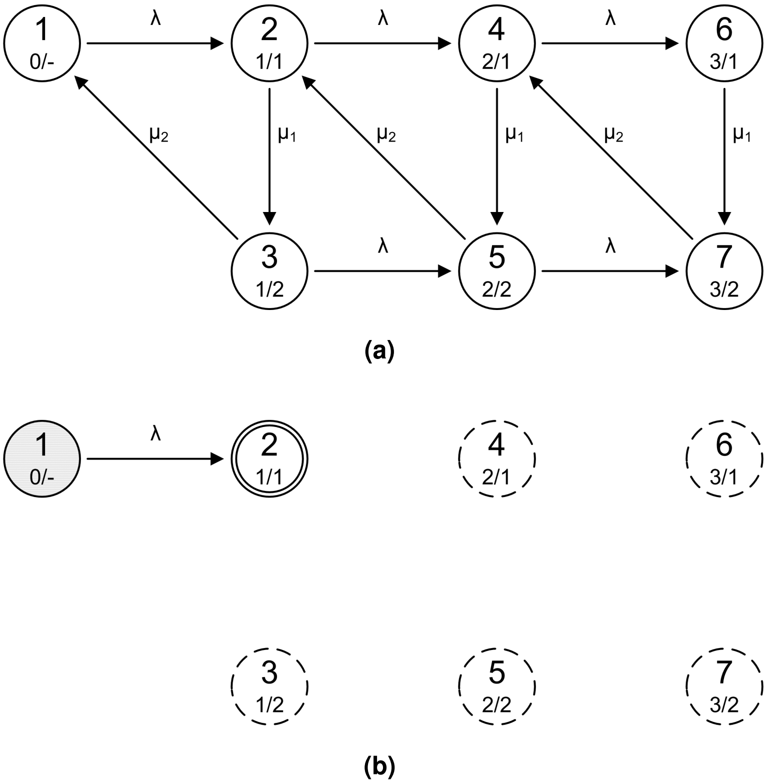


Figure 5.5.: M/Hypo/1/S queueing system $(S = 3)$. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the idle period. Meaning of the names of the states: number of customers in the system / state of the service process.

5.1.4. Other PH/PH/1/S queueing systems

M/Hypo/1/S queueing system

The Markov chains for an M/Hypo/1/S queueing system are shown in Figure 5.5.

As in the M/M/1/S queueing system, there is only one idle state (state $\langle 0 \rangle$), so the length of the idle period is exponentially distributed with mean $1/\lambda$.

$$I \sim \mathrm{Exp}(\lambda) \tag{5.17}$$

$$\mathrm{E}(I) = \frac{1}{\lambda} \tag{5.18}$$

M/Hyper/1/S queueing system

The Markov chains for an M/Hyper/1/S queueing system are shown in Figure 5.6.

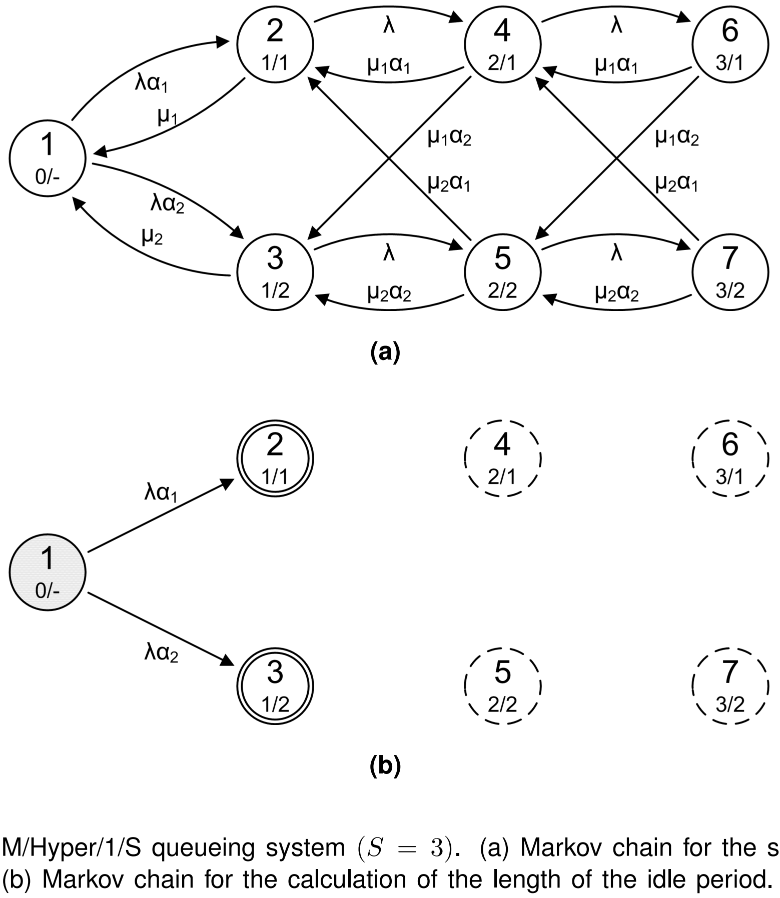


Figure 5.6.: M/Hyper/1/S queueing system ($S = 3$). (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the idle period. Meaning of the names of the states: number of customers in the system / state of the service process.

Again there is only one idle state (state $\langle 0 \rangle$), so the length of the idle period is exponentially distributed with mean $1/\lambda$:

$$I \sim \mathrm{Exp}(\lambda) \tag{5.19}$$

$$\mathrm{E}(I) = \frac{1}{\lambda} \tag{5.20}$$

Hypo/Hypo/1/S queueing system

Figure 5.7a shows the Markov chain for the system state of a Hypo/Hypo/1/S queueing system.

The Markov chain can be in the states $\langle 0/1/-\rangle$ and $\langle 0/2/-\rangle$ when the idle period begins. If it was in state $\langle 1/1/2\rangle$ before the last customer left the system (probability $\pi_{\langle 1/1/2\rangle}/\left(\pi_{\langle 1/1/2\rangle}+\pi_{\langle 1/2/2\rangle}\right)$), it is in state $\langle 0/1/-\rangle$ when the idle period begins. If it was in state $\langle 1/2/2\rangle$ (probability $\pi_{\langle 1/2/2\rangle}/\left(\pi_{\langle 1/1/2\rangle}+\pi_{\langle 1/2/2\rangle}\right)$), it is in state $\langle 0/2/-\rangle$ when the idle period begins. Therefore, we have

$$\sigma^I_{\langle 0/1/-\rangle} = \frac{\pi_{\langle 1/1/2\rangle}}{\pi_{\langle 1/1/2\rangle}+\pi_{\langle 1/2/2\rangle}} \tag{5.21}$$

$$\sigma^I_{\langle 0/2/-\rangle} = \frac{\pi_{\langle 1/2/2\rangle}}{\pi_{\langle 1/1/2\rangle}+\pi_{\langle 1/2/2\rangle}} \tag{5.22}$$

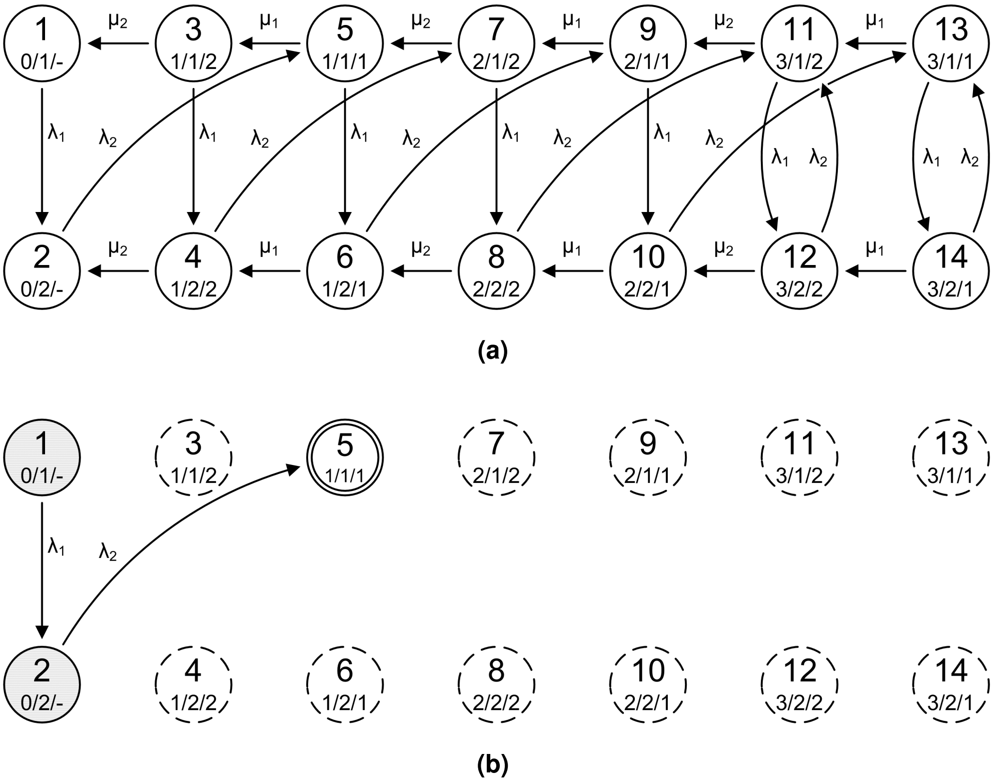


Figure 5.7.: Hypo/Hypo/1/S queueing system $(S = 3)$. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the idle period. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

The Markov chain for the calculation of the length of the idle period is shown in Figure 5.7b. If the Markov chain is in state $\langle 0/1/-\rangle$ when the idle period begins, the length of the idle period is hypoexponentially distributed. If it is in state $\langle 0/2/-\rangle$ when the idle period begins, the length of the idle period is exponentially distributed. Therefore, I has the distribution

$$I \sim \frac{\pi_{\langle 1/1/2\rangle}}{\pi_{\langle 1/1/2\rangle} + \pi_{\langle 1/2/2\rangle}} \,\mathrm{HypoExp}(\lambda_1, \lambda_2) + \frac{\pi_{\langle 1/2/2\rangle}}{\pi_{\langle 1/1/2\rangle} + \pi_{\langle 1/2/2\rangle}} \,\mathrm{Exp}(\lambda_2) \tag{5.23}$$

Its mean is

$$\mathrm{E}(I) = \frac{1}{\lambda_2} + \frac{\pi_{\langle 1/1/2\rangle}}{\pi_{\langle 1/1/2\rangle} + \pi_{\langle 1/2/2\rangle}} \frac{1}{\lambda_1} \tag{5.24}$$

Hypo/Hyper/1/S queueing system

The Markov chains for a Hypo/Hyper/1/S queueing system are shown in Figure 5.8.

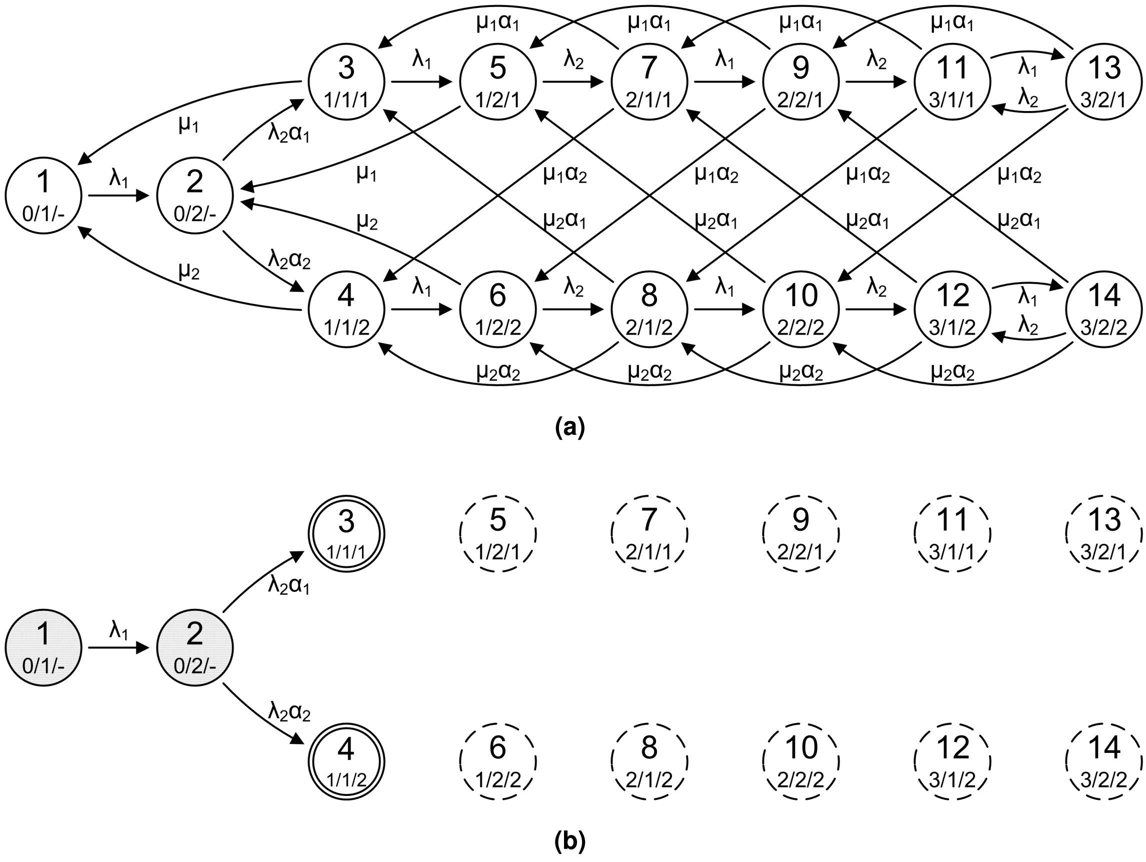


Figure 5.8.: Hypo/Hyper/1/S queueing system ($S = 3$). (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the idle period. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

There are two idle states, state $\langle 0/1/-\rangle$ and state $\langle 0/2/-\rangle$. The Markov chain can be in both of them when the idle period begins. If it was in state $\langle 1/1/1\rangle$ or in state

$\langle 1/1/2\rangle$ before the last customer left the system (probability $\left(\mu_1\pi_{\langle 1/1/1\rangle}+\mu_2\pi_{\langle 1/1/2\rangle}\right)/\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}$), it is in state $\langle 0/1/-\rangle$ afterwards. If it was in state $\langle 1/2/1\rangle$ or in state $\langle 1/2/2\rangle$ (probability $\left(\mu_1\pi_{\langle 1/2/1\rangle}+\mu_2\pi_{\langle 1/2/2\rangle}\right)/\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}$), it is in state $\langle 0/2/-\rangle$:

$$\sigma^I_{\langle 0/1/-\rangle}=\frac{\mu_1\pi_{\langle 1/1/1\rangle}+\mu_2\pi_{\langle 1/1/2\rangle}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}} \tag{5.25}$$

$$\sigma^I_{\langle 0/2/-\rangle}=\frac{\mu_1\pi_{\langle 1/2/1\rangle}+\mu_2\pi_{\langle 1/2/2\rangle}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}} \tag{5.26}$$

If the Markov chain is in state $\langle 0/1/-\rangle$ when the idle period begins, the length of the idle period is hypoexponentially distributed (because $\alpha_1+\alpha_2=1$). If it is in state $\langle 0/2/-\rangle$ when the idle period begins, the length of the idle period is exponentially distributed. Therefore, I has the distribution

$$I\sim\frac{\mu_1\pi_{\langle 1/1/1\rangle}+\mu_2\pi_{\langle 1/1/2\rangle}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}}\operatorname{HypoExp}(\lambda_1,\lambda_2)+\frac{\mu_1\pi_{\langle 1/2/1\rangle}+\mu_2\pi_{\langle 1/2/2\rangle}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}}\operatorname{Exp}(\lambda_2) \tag{5.27}$$

Its mean is

$$\mathrm{E}(I)=\frac{1}{\lambda_2}+\frac{\mu_1\pi_{\langle 1/1/1\rangle}+\mu_2\pi_{\langle 1/1/2\rangle}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}}\frac{1}{\lambda_1} \tag{5.28}$$

Hyper/Hypo/1/S queueing system

Figure 5.9a shows the Markov chain for the system state of a Hyper/Hypo/1/S queueing system.

The Markov chain can be in the states $\langle 0/1/-\rangle$ and $\langle 0/2/-\rangle$ when the idle period begins. If it was in state $\langle 1/1/2\rangle$ before the last customer left the system (probability $\pi_{\langle 1/1/2\rangle}/\left(\pi_{\langle 1/1/2\rangle}+\pi_{\langle 1/2/2\rangle}\right)$), it is in state $\langle 0/1/-\rangle$ when the idle period begins. If it was in state $\langle 1/2/2\rangle$ (probability $\pi_{\langle 1/2/2\rangle}/\left(\pi_{\langle 1/1/2\rangle}+\pi_{\langle 1/2/2\rangle}\right)$), it is in state $\langle 0/2/-\rangle$ when the idle period begins. Therefore, we have

$$\sigma^I_{\langle 0/1/-\rangle}=\frac{\pi_{\langle 1/1/2\rangle}}{\pi_{\langle 1/1/2\rangle}+\pi_{\langle 1/2/2\rangle}} \tag{5.29}$$

$$\sigma^I_{\langle 0/2/-\rangle}=\frac{\pi_{\langle 1/2/2\rangle}}{\pi_{\langle 1/1/2\rangle}+\pi_{\langle 1/2/2\rangle}} \tag{5.30}$$

The Markov chain for the calculation of the length of the idle period is shown in Figure 5.9b. If the Markov chain is in state $\langle 0/1/-\rangle$ when the idle period begins, the length of the idle period is exponentially distributed with rate λ_1. If it is in state $\langle 0/2/-\rangle$ when the idle period begins, the length of the idle period is exponentially distributed with rate λ_2. Therefore, I has the distribution

$$\begin{aligned}I&\sim\frac{\pi_{\langle 1/1/2\rangle}}{\pi_{\langle 1/1/2\rangle}+\pi_{\langle 1/2/2\rangle}}\operatorname{Exp}(\lambda_1)+\frac{\pi_{\langle 1/2/2\rangle}}{\pi_{\langle 1/1/2\rangle}+\pi_{\langle 1/2/2\rangle}}\operatorname{Exp}(\lambda_2)=\\&\operatorname{HyperExp}\left(\lambda_1,\frac{\pi_{\langle 1/1/2\rangle}}{\pi_{\langle 1/1/2\rangle}+\pi_{\langle 1/2/2\rangle}},\lambda_2,\frac{\pi_{\langle 1/2/2\rangle}}{\pi_{\langle 1/1/2\rangle}+\pi_{\langle 1/2/2\rangle}}\right)\end{aligned} \tag{5.31}$$

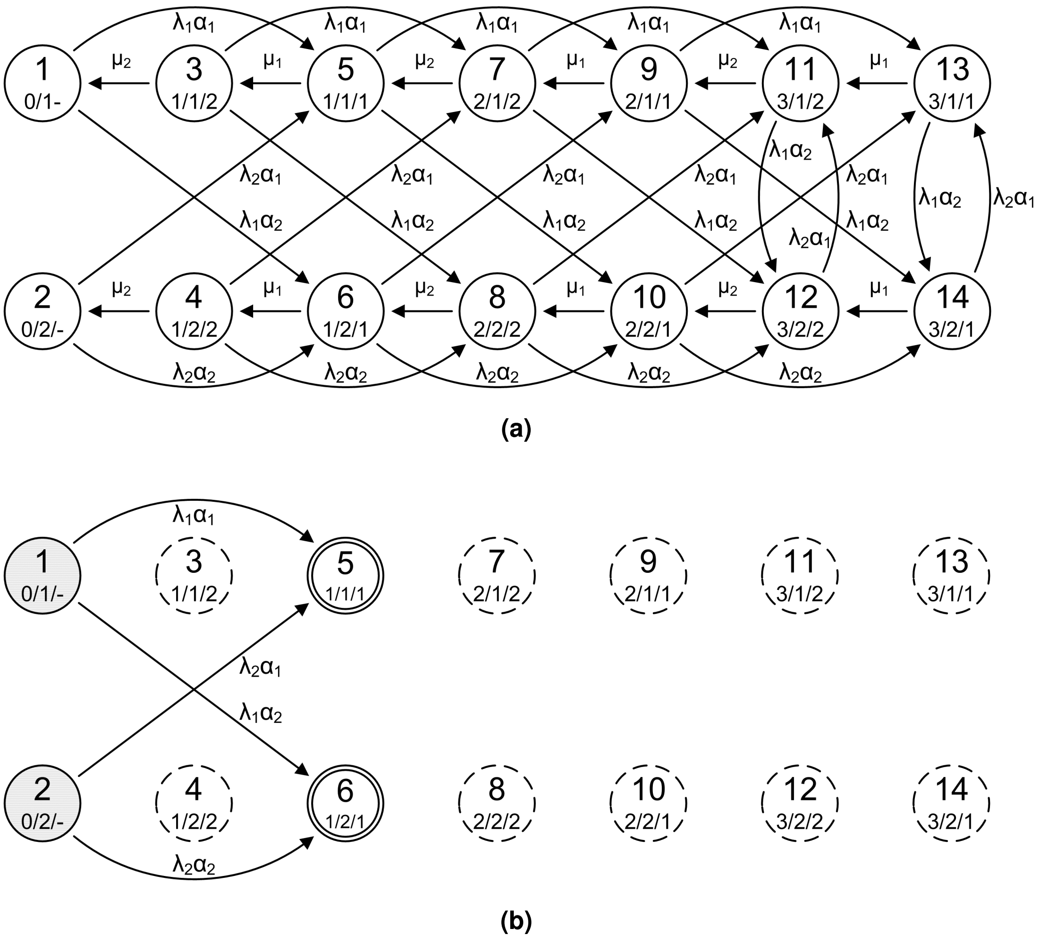


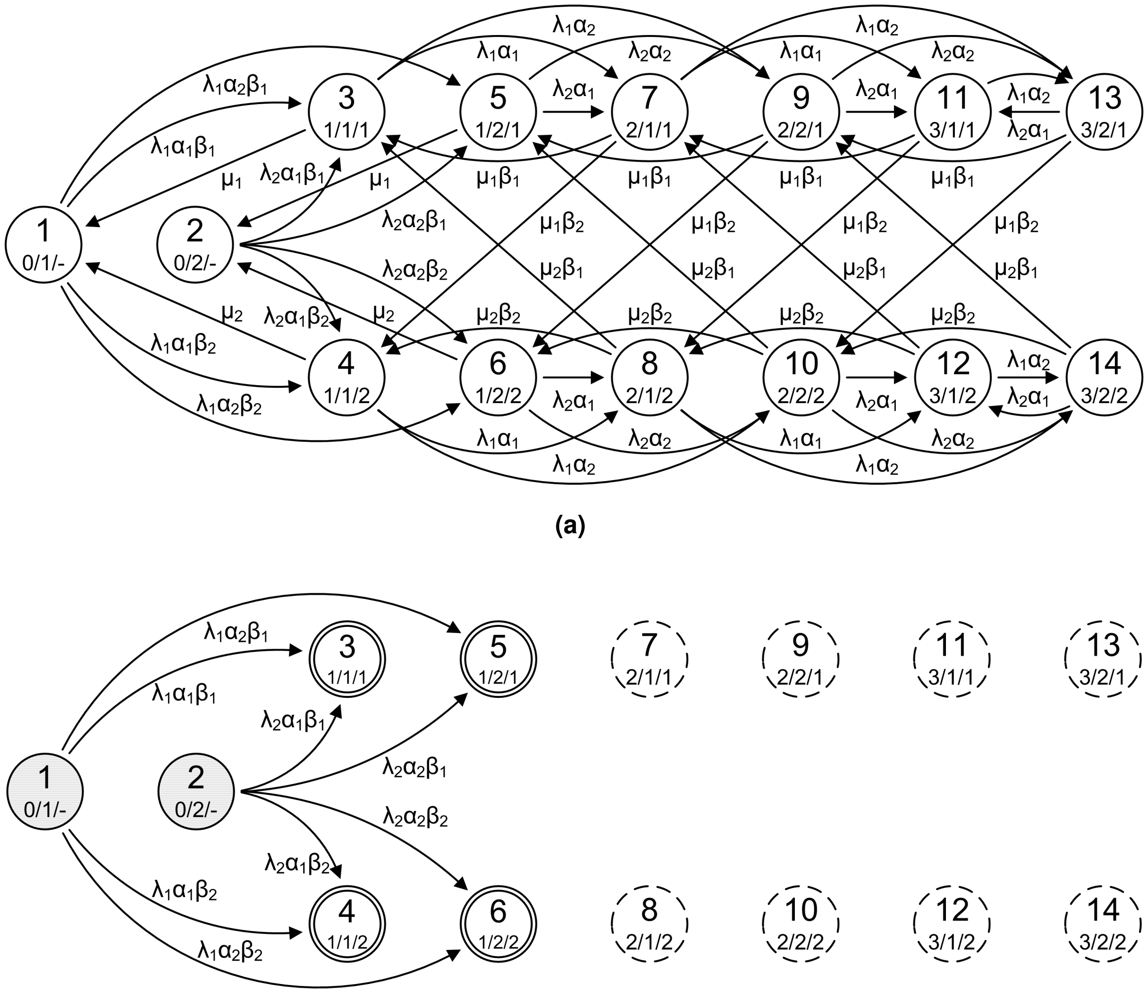
Figure 5.9.: Hyper/Hypo/1/S queueing system $(S = 3)$. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the idle period. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

Its mean is

$$\mathrm{E}(I) = \frac{\pi_{\langle 1/1/2\rangle}}{\pi_{\langle 1/1/2\rangle} + \pi_{\langle 1/2/2\rangle}} \frac{1}{\lambda_1} + \frac{\pi_{\langle 1/2/2\rangle}}{\pi_{\langle 1/1/2\rangle} + \pi_{\langle 1/2/2\rangle}} \frac{1}{\lambda_2} \tag{5.32}$$

Hyper/Hyper/1/S queueing system

The Markov chains for a Hyper/Hyper/1/S queueing system are shown in Figure 5.10.



(a)

(b)

Figure 5.10.: Hyper/Hyper/1/S queueing system ($S = 3$). (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the idle period. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

There are two idle states, state $\langle 0/1/-\rangle$ and state $\langle 0/2/-\rangle$. The Markov chain can be in both of them when the idle period begins. If it was in state $\langle 1/1/1\rangle$ or in state $\langle 1/1/2\rangle$ before the last customer left the system (probability $\left(\mu_1\pi_{\langle 1/1/1\rangle} + \mu_2\pi_{\langle 1/1/2\rangle}\right) / \sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}$), it is in state $\langle 0/1/-\rangle$ afterwards. If it was in state $\langle 1/2/1\rangle$ or in state $\langle 1/2/2\rangle$ (probability $\left(\mu_1\pi_{\langle 1/2/1\rangle} + \mu_2\pi_{\langle 1/2/2\rangle}\right)/\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}$), it is in state $\langle 0/2/-\rangle$:

$$\sigma^I_{\langle 0/1/-\rangle} = \frac{\mu_1\pi_{\langle 1/1/1\rangle} + \mu_2\pi_{\langle 1/1/2\rangle}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}} \tag{5.33}$$

$$\sigma^I_{\langle 0/2/-\rangle} = \frac{\mu_1\pi_{\langle 1/2/1\rangle} + \mu_2\pi_{\langle 1/2/2\rangle}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}} \tag{5.34}$$

If the Markov chain is in state $\langle 0/1/-\rangle$ when the idle period begins, the length of the idle period is exponentially distributed with rate λ_1 (because $\alpha_1 + \alpha_2 = \beta_1 + \beta_2 = 1$). If it is in state $\langle 0/2/-\rangle$ when the idle period begins, the length of the idle period is exponentially distributed with rate λ_2. Therefore, I has the distribution

$$\begin{aligned} I \sim \frac{\mu_1\pi_{\langle 1/1/1\rangle} + \mu_2\pi_{\langle 1/1/2\rangle}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}}\,\mathrm{Exp}(\lambda_1) + \frac{\mu_1\pi_{\langle 1/2/1\rangle} + \mu_2\pi_{\langle 1/2/2\rangle}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}}\,\mathrm{Exp}(\lambda_2) = \\ \mathrm{HyperExp}\left(\lambda_1, \frac{\mu_1\pi_{\langle 1/1/1\rangle} + \mu_2\pi_{\langle 1/1/2\rangle}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}}, \lambda_2, \frac{\mu_1\pi_{\langle 1/2/1\rangle} + \mu_2\pi_{\langle 1/2/2\rangle}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}}\right) \end{aligned} \tag{5.35}$$

Its mean is

$$\mathrm{E}(I) = \frac{\mu_1\pi_{\langle 1/1/1\rangle} + \mu_2\pi_{\langle 1/1/2\rangle}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}}\frac{1}{\lambda_1} + \frac{\mu_1\pi_{\langle 1/2/1\rangle} + \mu_2\pi_{\langle 1/2/2\rangle}}{\sum_{i=1}^{2}\sum_{j=1}^{2}\mu_j\pi_{\langle 1/i/j\rangle}}\frac{1}{\lambda_2} \tag{5.36}$$

5.2. Length of the busy period

The busy period is the time frame in which there are customers in the system, and, therefore, the server is busy. The busy period begins when the system is idle and a customer arrives. It ends when the system becomes idle again.

To calculate the length B of the busy period, we identify (by means of the Markov chain for the system state $\mathcal{M}_S$) the states in which the Markov chain can be when the busy period begins (that is, the states in which the Markov chain can be after a customer enters an empty system). The probability σ_j^B that the Markov chain is in state j when the busy period begins is the ratio of the rate at which state j is entered from idle states to the rate at which all busy states are entered from idle states:

$$\sigma_j^B = \frac{\sum\limits_{i\in\mathcal{I}} \pi_i q_{ij}}{\sum\limits_{i\in\mathcal{I}} \pi_i \sum\limits_{b\in\mathcal{B}} q_{ib}} \qquad j \in \mathcal{B} \tag{5.37}$$

Next we create a new Markov chain $\mathcal{M}_B$ by removing all transitions of $\mathcal{M}_S$ which do not originate in a busy state. That means, in $\mathcal{M}_B$, all idle states are absorbing states. With this Markov chain, we calculate $\varphi_i(\cdot)$, the complementary cumulative distribution function of the remaining length of the busy period (that is, the time needed to reach an idle state) given that the Markov chain is in state i:

$$\varphi_i(0) = \begin{cases} 1 & \text{if } i \text{ is a busy state} \\ 0 & \text{if } i \text{ is an idle state} \end{cases} \tag{5.38}$$

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{5.39}$$

The cumulative distribution function of the length of the busy period is then

$$\mathrm{P}(B \le t) = 1 - \sum_j \sigma_j^B \cdot \varphi_j(t) \tag{5.40}$$

The mean length of the busy period can be calculated with

$$\mathrm{E}(B) = \int\limits_{t=0}^{\infty} \sum_j \sigma_j^B \varphi_j(t) \mathrm{d}t \tag{5.41}$$

If we know the utilisation of the server and the mean length of the idle period, we can also use these quantities to calculate the mean length of the busy period: Since the steady-state probability that the server is busy equals the long-run ratio of busy periods to the total time,

$$\mathrm{P}\left\{\text{Server is busy}\right\} = \rho = \frac{\mathrm{E}(B)}{\mathrm{E}(B) + \mathrm{E}(I)} \tag{5.42}$$

we have

$$\mathrm{E}(B) = \mathrm{E}(I) \cdot \frac{\rho}{1-\rho} \tag{5.43}$$

Another approach to determine the mean length of the busy period is the following:

Suppose the Markov chain is in state i and let $L(i)$ be the expected time the Markov chain needs to go from state i to an idle state. $L(i)$ depends on the sojourn time in state i and the state that is taken next. If the next transition is to state j, the expected time until an idle state is reached is the mean sojourn time in state i plus the time the Markov chain needs to go from state j to an idle state. If we denote the mean sojourn time in state i by $S(i)$, we have

$$L(i) \mid \text{next transition is to state } j = S(i) + L(j) \tag{5.44}$$

The probability that the next state is j is $-q_{ij}/q_{ii}$. The mean sojourn time in state i is $S(i) = -1/q_{ii}$. Summing over all possible transitions yields

$$L(i) = \begin{cases} S(i) + \sum\limits_{j \neq i} L(j) \frac{q_{ij}}{-q_{ii}} & i \text{ is no idle state} \\ 0 & i \text{ is an idle state} \end{cases} \tag{5.45}$$

Now the expected length of the busy period is

$$\mathrm{E}(B) = \sum_{b \in \mathcal{B}} L(b) \sigma_b^B \tag{5.46}$$

5.2.1. M/M/1/S queueing system

The Markov chain for the system state of an M/M/1/S queueing system is shown in Figure 5.11a. The only idle state is state $\langle 0 \rangle$, all other states are busy states. When the system is empty and a customer arrives, the Markov chain is in state $\langle 1 \rangle$:

$$\sigma_{\langle 1 \rangle}^B = 1 \tag{5.47}$$

Therefore, we have to calculate the time it takes to go from state $\langle 1 \rangle$ to the idle state $\langle 0 \rangle$. This is done with the Markov chain $\mathcal{M}_B$, which is shown in Figure 5.11b.

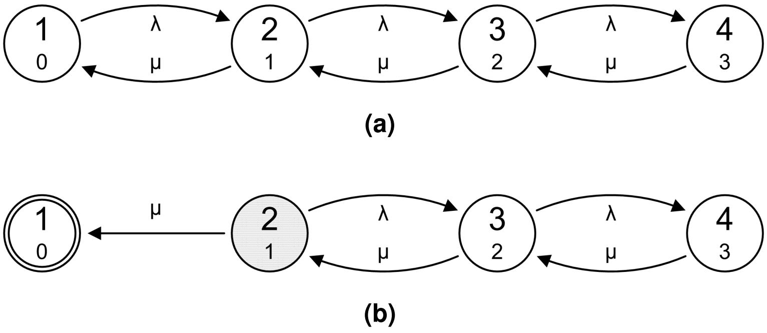


Figure 5.11.: M/M/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the busy period. Meaning of the names of the states: number of customers in the system.

We have

$$\varphi'(t) = \mathcal{Q} \cdot \varphi(t) \qquad\qquad \varphi_i(0) = \begin{cases} 0 & i = \langle 0 \rangle \\ 1 & i \neq \langle 0 \rangle \end{cases} \tag{5.48}$$

The cumulative distribution function of the length of the busy period is

$$\mathrm{P}(B \leq t) = 1 - \varphi_2(t) \tag{5.49}$$

The expected length of the busy period is

$$\mathrm{E}(B) = \int\limits_{t=0}^{\infty} \varphi_2(t) \mathrm{d}t \tag{5.50}$$

If we use the result that the mean length of the idle period is

$$\mathrm{E}(I) = \frac{1}{\lambda} \tag{5.51}$$

we can calculate the mean length of the busy period with

$$\mathrm{E}(B) = \mathrm{E}(I) \cdot \frac{\rho}{1-\rho} = \frac{1 - \pi_{\langle 0 \rangle}}{\lambda \pi_{\langle 0 \rangle}} \tag{5.52}$$

Using the closed-form solution

$$\pi_{\langle 0 \rangle} = \begin{cases} \frac{1-\lambda/\mu}{1-(\lambda/\mu)^{S+1}} & \lambda \neq \mu \\ \frac{1}{S+1} & \lambda = \mu \end{cases} \tag{5.53}$$

we obtain

$$\mathrm{E}(B) = \begin{cases} \frac{\lambda/\mu-(\lambda/\mu)^{S+1}}{\lambda(1-\lambda/\mu)} & \lambda \neq \mu \\ \frac{S}{\lambda} & \lambda = \mu \end{cases} \tag{5.54}$$

Alternative method for the calculation of the mean of the length of the busy period:

The mean sojourn times in states 2, 3 and 4 are

$$\begin{aligned} S(2) &= \frac{1}{-q_{2,2}} = \frac{1}{\lambda + \mu} \\ S(3) &= \frac{1}{-q_{3,3}} = \frac{1}{\lambda + \mu} \\ S(4) &= \frac{1}{-q_{4,4}} = \frac{1}{\mu} \end{aligned} \tag{5.55}$$

The remaining length of the busy period given that the Markov chain is in state i, $i = 2 \dots 4$ is

$$\begin{aligned} L(2) &= S(2) + \frac{\lambda}{\lambda+\mu} L(3) \\ L(3) &= S(3) + \frac{\lambda}{\lambda+\mu} L(4) + \frac{\mu}{\lambda+\mu} L(2) \\ L(4) &= S(4) + L(3) \end{aligned} \tag{5.56}$$

Solving this system of equations yields

$$\begin{aligned} L(2) &= \frac{\mu^2 + \lambda\mu + \lambda^2}{\mu^3} \\ L(3) &= \frac{2\mu^2 + 2\lambda\mu + \lambda^2}{\mu^3} \\ L(4) &= \frac{3\mu^2 + 2\lambda\mu + \lambda^2}{\mu^3} \end{aligned} \tag{5.57}$$

Since the Markov chain is in state 2 when the busy period begins, the expected length of the busy period is

$$\mathrm{E}(B) = L(2) = \frac{\mu^2 + \lambda\mu + \lambda^2}{\mu^3} \tag{5.58}$$

Figure 5.12 shows the length of the idle and the busy period in an M/M/1/S queueing system as a function of the arrival rate.

5.2.2. Hypo/M/1/S queueing system

Figure 5.13a shows the Markov chain for the system state of a Hypo/M/1/S queueing system. States $\langle 0/1 \rangle$ and $\langle 0/2 \rangle$ are idle states, the other states are busy states. When the busy period begins, the Markov chain is in state $\langle 1/1 \rangle$.

$$\sigma^B_{\langle 1/1 \rangle} = 1 \tag{5.59}$$

Therefore, we have to calculate the time which it takes to go from state $\langle 1/1 \rangle$ to one of the idle states $\langle 0/1 \rangle$ and $\langle 0/2 \rangle$. This is done with the Markov chain $\mathcal{M}_B$, which is shown in Figure 5.13b.

We have

$$\varphi'(t) = \mathcal{Q} \cdot \varphi(t) \qquad\qquad \varphi_{\langle i,j \rangle}(0) = \begin{cases} 0 & i = 0 \\ 1 & i \neq 0 \end{cases} \tag{5.60}$$

The cumulative distribution function of the length of the busy period is

$$\mathrm{P}(B \leq t) = 1 - \varphi_{\langle 1/1 \rangle}(t) \tag{5.61}$$

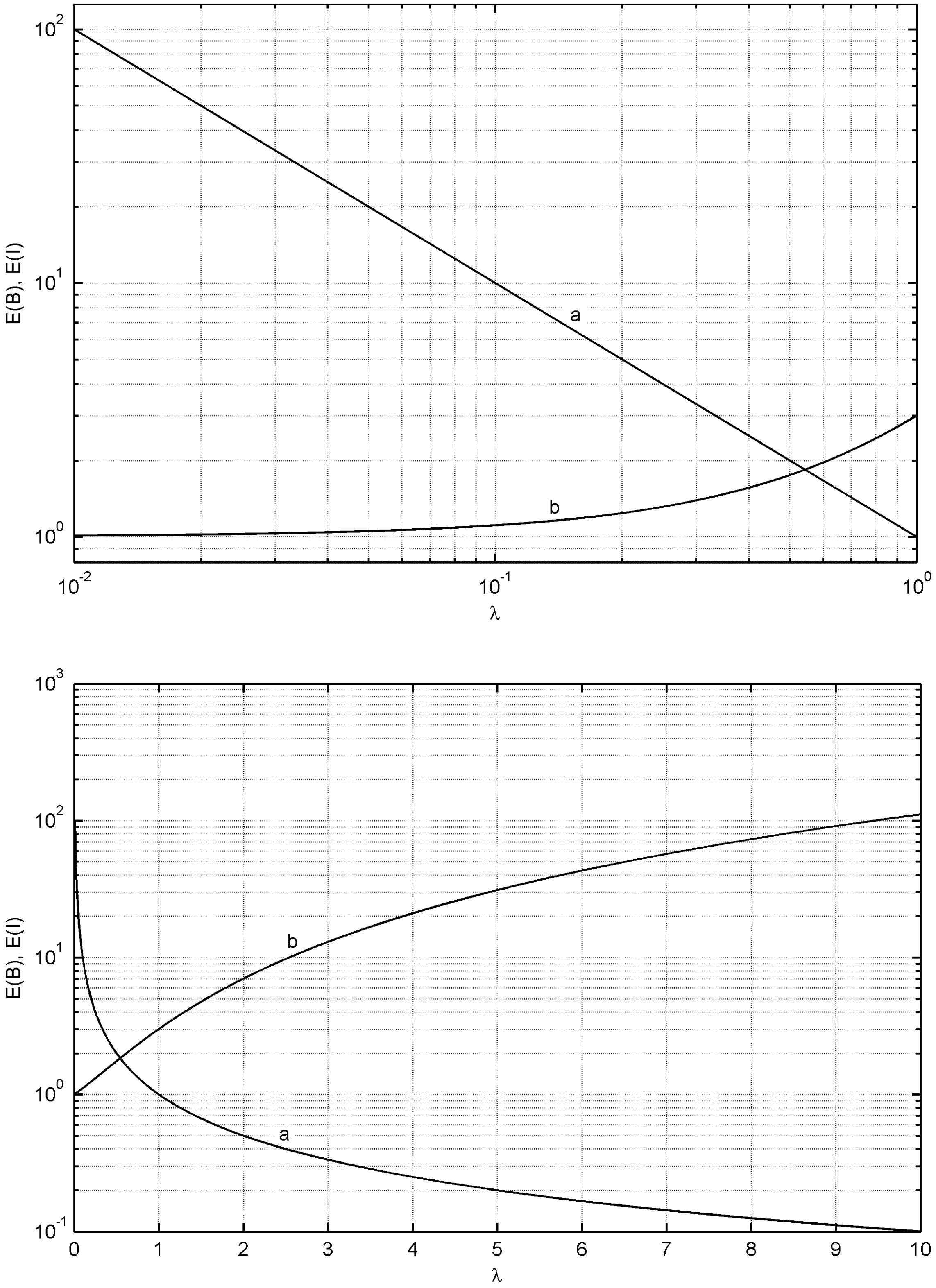


Figure 5.12.: M/M/1/S queueing system. Mean length of (a) idle period and (b) busy period. $S = 3, \mu = 1$.

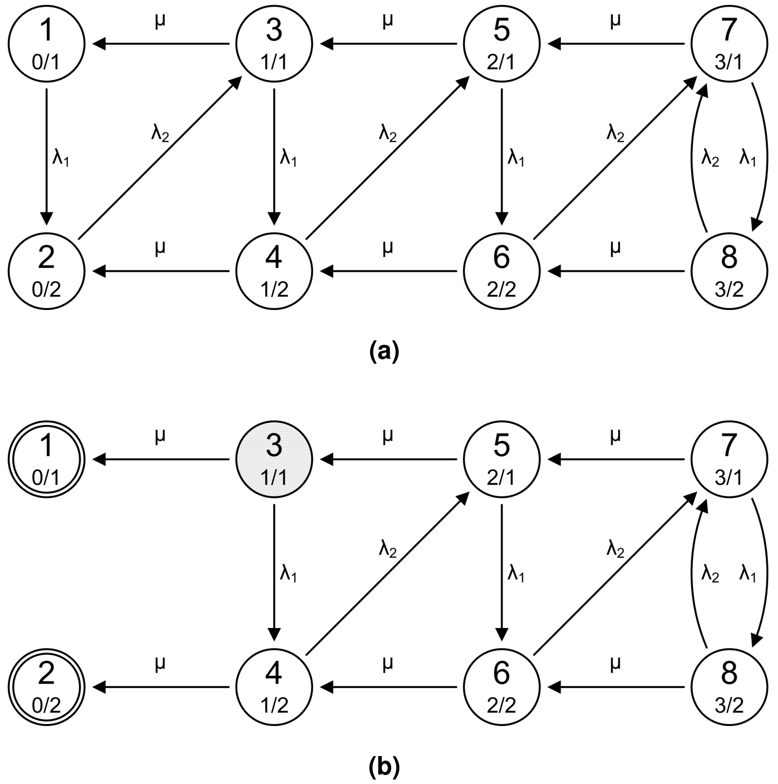


Figure 5.13.: Hypo/M/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the busy period. Meaning of the names of the states: number of customers in the system / state of the arrival process.

The expected length of the busy period is

$$\mathrm{E}(B) = \int\limits_{t=0}^{\infty} \varphi_{\langle 1/1 \rangle} \mathrm{d}t \tag{5.62}$$

Alternative method for the calculation of the mean of the length of the busy period:

The mean sojourn times in the busy states are

$$\begin{aligned}
S(3) &= \frac{1}{-q_{3,3}} = \frac{1}{\lambda_1 + \mu} \\
S(4) &= \frac{1}{-q_{4,4}} = \frac{1}{\lambda_2 + \mu} \\
S(5) &= \frac{1}{-q_{5,5}} = \frac{1}{\lambda_1 + \mu} \\
S(6) &= \frac{1}{-q_{6,6}} = \frac{1}{\lambda_2 + \mu} \\
S(7) &= \frac{1}{-q_{7,7}} = \frac{1}{\lambda_1 + \mu} \\
S(8) &= \frac{1}{-q_{8,8}} = \frac{1}{\lambda_2 + \mu}
\end{aligned} \tag{5.63}$$

The remaining length of the busy period given that the Markov chain is in state i, $i = 3 \ldots 8$ is

$$\begin{aligned}
L(3) &= S(3) + \frac{\lambda_1}{\lambda_1+\mu}L(4)\\
L(4) &= S(4) + \frac{\lambda_2}{\lambda_2+\mu}L(5)\\
L(5) &= S(5) + \frac{\lambda_1}{\lambda_1+\mu}L(6) + \frac{\mu}{\lambda_1+\mu}L(3)\\
L(6) &= S(6) + \frac{\lambda_2}{\lambda_2+\mu}L(7) + \frac{\mu}{\lambda_2+\mu}L(4)\\
L(7) &= S(7) + \frac{\lambda_1}{\lambda_1+\mu}L(8) + \frac{\mu}{\lambda_1+\mu}L(5)\\
L(8) &= S(8) + \frac{\lambda_2}{\lambda_2+\mu}L(7) + \frac{\mu}{\lambda_2+\mu}L(6)
\end{aligned} \tag{5.64}$$

Solving this system of equations yields

$$\begin{aligned}
L(3) = \; &\big(\mu^5 + (3\lambda_2 + 3\lambda_1)\,\mu^4 + \left(3{\lambda_1}^2 + 5\lambda_2\lambda_1 + 3{\lambda_2}^2\right)\mu^3 +\\
&(\lambda_2+\lambda_1)^3\,\mu^2 + \lambda_2\lambda_1\,(\lambda_2+\lambda_1)^2\,\mu + {\lambda_2}^2{\lambda_1}^2\,(\lambda_2+\lambda_1)\big)\,/\alpha
\end{aligned} \tag{5.65}$$

$$\begin{aligned}
L(4) = \; &\big(\mu^5 + (4\lambda_2 + 3\lambda_1)\,\mu^4 + \left(3{\lambda_1}^2 + 7\lambda_2\lambda_1 + 5{\lambda_2}^2\right)\mu^3 +\\
&(2\lambda_2+\lambda_1)\,(\lambda_2+\lambda_1)^2\,\mu^2 +\\
&\lambda_2\lambda_1\,(2\lambda_2+\lambda_1)\,(\lambda_2+\lambda_1)\,\mu + {\lambda_2}^2{\lambda_1}^2\,(\lambda_2+\lambda_1)\big)\,/\alpha
\end{aligned} \tag{5.66}$$

$$\begin{aligned}
L(5) = \; &\big(2\mu^5 + (6\lambda_2 + 6\lambda_1)\,\mu^4 + \left(6{\lambda_1}^2 + 11\lambda_2\lambda_1 + 6{\lambda_2}^2\right)\mu^3 +\\
&(2\lambda_2+\lambda_1)\,(\lambda_2+\lambda_1)\,(2\lambda_1+\lambda_2)\,\mu^2 +\\
&2\lambda_2\lambda_1\,(\lambda_2+\lambda_1)^2\,\mu + {\lambda_2}^2{\lambda_1}^2\,(\lambda_2+\lambda_1)\big)\,/\alpha
\end{aligned} \tag{5.67}$$

$$\begin{aligned}
L(6) = \; &\big(2\mu^5 + (8\lambda_2 + 6\lambda_1)\,\mu^4 + \left(6{\lambda_1}^2 + 14\lambda_2\lambda_1 + 9{\lambda_2}^2\right)\mu^3 +\\
&\left(2{\lambda_1}^3 + 8{\lambda_1}^2\lambda_2 + 3{\lambda_2}^3 + 9{\lambda_2}^2\lambda_1\right)\mu^2 +\\
&2\lambda_2\lambda_1\,(\lambda_2+\lambda_1)^2\,\mu + {\lambda_2}^2{\lambda_1}^2\,(\lambda_2+\lambda_1)\big)\,/\alpha
\end{aligned} \tag{5.68}$$

$$\begin{aligned}
L(7) = \; &\big(3\mu^5 + (9\lambda_2 + 9\lambda_1)\,\mu^4 + \left(9{\lambda_2}^2 + 17\lambda_2\lambda_1 + 9{\lambda_1}^2\right)\mu^3 +\\
&3\,(\lambda_2+\lambda_1)^3\,\mu^2 + 2\lambda_2\lambda_1\,(\lambda_2+\lambda_1)^2\,\mu + {\lambda_2}^2{\lambda_1}^2\,(\lambda_2+\lambda_1)\big)\,/\alpha
\end{aligned} \tag{5.69}$$

$$\begin{aligned}
L(8) = \; &\big(3\mu^5 + (11\lambda_2 + 9\lambda_1)\,\mu^4 + \left(10{\lambda_2}^2 + 18\lambda_2\lambda_1 + 9{\lambda_1}^2\right)\mu^3 +\\
&3\,(\lambda_2+\lambda_{\mathit{1}})^3\,\mu^2 + 2\lambda_2\lambda_1\,(\lambda_2+\lambda_1)^2\,\mu + {\lambda_2}^2{\lambda_1}^2\,(\lambda_2+\lambda_1)\big)\,/\alpha
\end{aligned} \tag{5.70}$$

where

$$\alpha = (\lambda_1 + \lambda_2 + \mu)(\lambda_1^2 + \lambda_2^2 + \mu^2 + 2\mu(\lambda_1 + \lambda_2))\mu^3 \tag{5.71}$$

When the busy period begins, the Markov chain is in state 3, therefore,

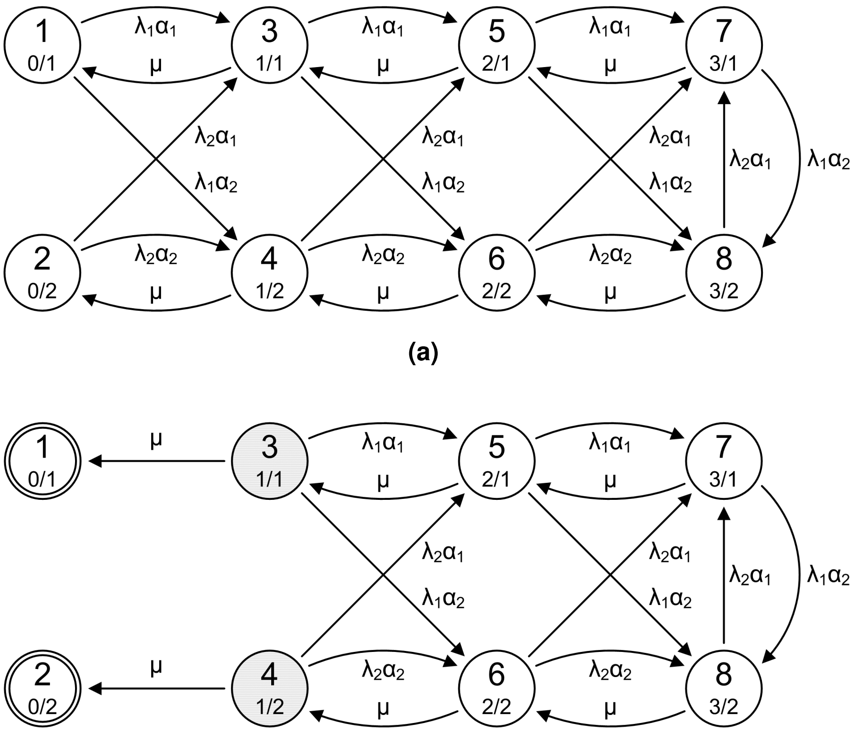
$$\mathrm{E}(B) = L(3) \tag{5.72}$$

5.2.3. Hyper/M/1/S queueing system

The Markov chain for the system state of a Hyper/M/1/S queueing system is shown in Figure 5.14a. In this system, the states $\langle 0/1\rangle$ and $\langle 0/2\rangle$ are idle states, all other states are busy states. When the system is empty and a customer arrives, the Markov chain is in state $\langle 1/1\rangle$ with probability α_1 and in state $\langle 1/2\rangle$ with probability α_2:

$$\sigma^B_{\langle 1/1\rangle} = \alpha_1 \tag{5.73}$$

$$\sigma^B_{\langle 1/2\rangle} = \alpha_2 \tag{5.74}$$



(a)

(b)

Figure 5.14.: Hyper/M/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the busy period. Meaning of the names of the states: number of customers in the system / state of the arrival process.

The Markov chain for the calculation of the length of the busy period is shown in Figure 5.14b.

We have

$$\varphi'(t) = \mathcal{Q} \cdot \varphi(t) \qquad\qquad \varphi_{\langle i/j\rangle}(0) = \begin{cases} 0 & i = 0 \\ 1 & i > 0 \end{cases} \tag{5.75}$$

If the Markov chain is in state $\langle 1/1\rangle$ when the busy period begins, the complementary cumulative distribution function of the length of the busy period is $\varphi_{\langle 1/1\rangle}(\cdot)$, otherwise it is $\varphi_{\langle 1/2\rangle}(\cdot)$. Therefore, the cumulative distribution function of the length of the busy period is

$$\mathrm{P}(B \le t) = 1 - \alpha_1\varphi_{\langle 1/1\rangle}(t) - \alpha_2\varphi_{\langle 1/2\rangle}(t) \tag{5.76}$$

The expected length of the busy period is

$$\mathrm{E}(B) = \int\limits_{t=0}^{\infty} \alpha_1\varphi_{\langle 1/1\rangle}(t) + \alpha_2\varphi_{\langle 1/2\rangle}(t)\ \mathrm{d}t \tag{5.77}$$

If we use the result that the mean length of the idle period is

$$\mathrm{E}(I) = \frac{\pi_{\langle 1/1\rangle}}{\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}}\frac{1}{\lambda_1} + \frac{\pi_{\langle 1/2\rangle}}{\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}}\frac{1}{\lambda_2} \tag{5.78}$$

we can calculate the mean length of the busy period with

$$\mathrm{E}(B) = \mathrm{E}(I) \cdot \frac{\rho}{1-\rho} = \frac{\left(\lambda_1\pi_{\langle 1/2\rangle} + \lambda_2\pi_{\langle 1/1\rangle}\right)\left(1 - \pi_{\langle 0/1\rangle} - \pi_{\langle 0/2\rangle}\right)}{\lambda_1\lambda_2\left(\pi_{\langle 1/1\rangle} + \pi_{\langle 1/2\rangle}\right)\left(\pi_{\langle 0/1\rangle} + \pi_{\langle 0/2\rangle}\right)} \tag{5.79}$$

5.2.4. Other PH/PH/1/S queueing systems

M/Hypo/1/S queueing system

Figure 5.15a shows the Markov chain for the system state of an M/Hypo/1/S queueing system. State $\langle 0/-\rangle$ is the only idle state. When the busy period begins, the Markov chain is in state $\langle 1/1\rangle$:

$$\sigma^B_{\langle 1/1\rangle} = 1 \tag{5.80}$$

Therefore, we have to calculate the time it takes to go from state $\langle 1/1\rangle$ to the idle state $\langle 0/-\rangle$. This is done with the Markov chain $\mathcal{M}_B$, which is shown in Figure 5.15b.

We have

$$\varphi'(t) = \mathcal{Q}\cdot\varphi(t) \qquad\qquad \varphi_i(0) = \begin{cases} 0 & i = \langle 0/-\rangle \\ 1 & i \ne \langle 0/-\rangle \end{cases} \tag{5.81}$$

The cumulative distribution function of the length of the busy period is

$$\mathrm{P}(B \le t) = 1 - \varphi_{\langle 1/1\rangle}(t) \tag{5.82}$$

The expected length of the busy period is

$$\mathrm{E}(B) = \int\limits_{t=0}^{\infty} \varphi_{\langle 1/1\rangle}\mathrm{d}t \tag{5.83}$$

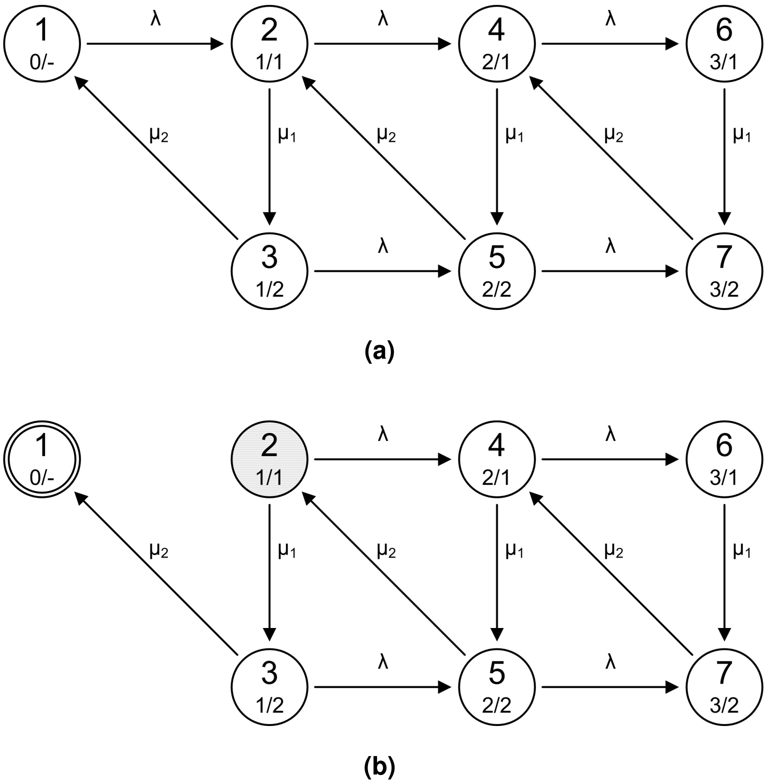


Figure 5.15.: M/Hypo/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the busy period. Meaning of the names of the states: number of customers in the system / state of the service process.

M/Hyper/1/S queueing system

The Markov chain for the system state of an M/Hyper/1/S queueing system is shown in Figure 5.16a. In this system, the state $\langle 0/-\rangle$ is the only idle state. When the system is empty and a customer arrives, the Markov chain is in state $\langle 1/1\rangle$ with probability α_1 and in state $\langle 1/2\rangle$ with probability α_2:

$$\sigma^B_{\langle 1/1\rangle} = \alpha_1 \tag{5.84}$$

$$\sigma^B_{\langle 1/2\rangle} = \alpha_2 \tag{5.85}$$

The Markov chain for the calculation of the length of the busy period is shown in Figure 5.16b.

We have

$$\varphi'(t) = \mathcal{Q} \cdot \varphi(t) \qquad\qquad \varphi_{\langle i/j\rangle}(0) = \begin{cases} 0 & i = 0 \\ 1 & i > 0 \end{cases} \tag{5.86}$$

If the Markov chain is in state $\langle 1/1\rangle$ when the busy period begins, the complementary cumulative distribution function of the length of the busy period is $\varphi_{\langle 1/1\rangle}(t)$, otherwise it is $\varphi_{\langle 1/2\rangle}(t)$. Therefore, the cumulative distribution function of the length of the busy period is

$$\mathrm{P}(B \le t) = 1 - \alpha_1\varphi_{\langle 1/1\rangle}(t) - \alpha_2\varphi_{\langle 1/2\rangle}(t) \tag{5.87}$$

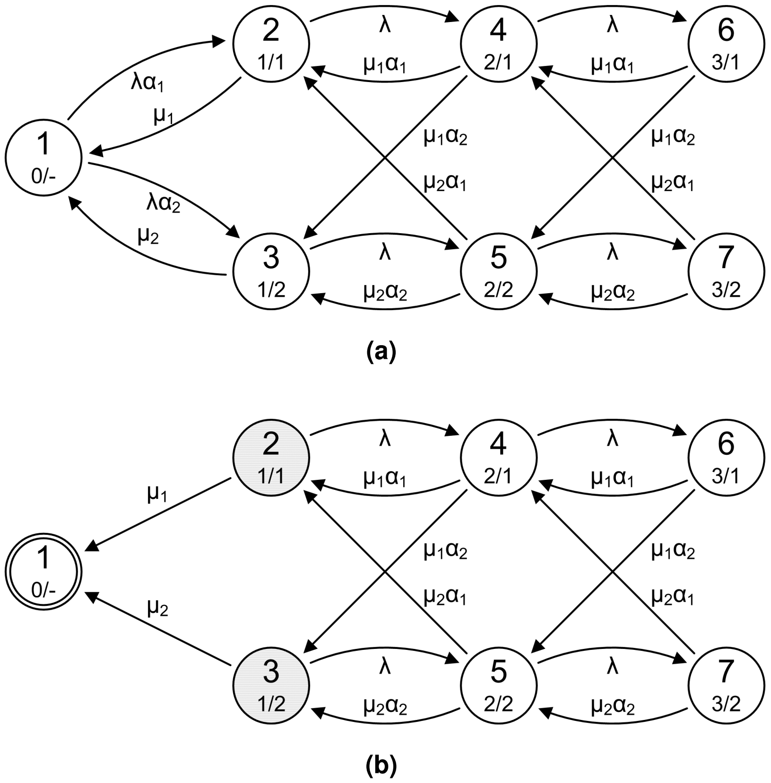


Figure 5.16.: M/Hyper/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the busy period. Meaning of the names of the states: number of customers in the system / state of the service process.

The expected length of the busy period is

$$\mathrm{E}(B) = \int\limits_{t=0}^{\infty} \alpha_1 \varphi_{\langle 1/1 \rangle}(t) + \alpha_2 \varphi_{\langle 1/2 \rangle}(t) \; \mathrm{d}t \tag{5.88}$$

Hypo/Hypo/1/S queueing system

Figure 5.17a shows the Markov chain for the system state of a Hypo/Hypo/1/S queueing system. The states $\langle 0/1/- \rangle$ and $\langle 0/2/- \rangle$ are idle states, all other states are busy states. When the busy period begins, the Markov chain is in state $\langle 1/1/1 \rangle$:

$$\sigma^B_{\langle 1/1/1 \rangle} = 1 \tag{5.89}$$

Therefore, we have to calculate the time it takes to go from state $\langle 1/1/1 \rangle$ to one of the idle states. This is done with the Markov chain $\mathcal{M}_B$, which is shown in Figure 5.17b.

We have

$$\varphi'(t) = \mathcal{Q} \cdot \varphi(t) \qquad\qquad \varphi_{\langle i/j/k \rangle}(0) = \begin{cases} 0 & i = 0 \\ 1 & i > 0 \end{cases} \tag{5.90}$$

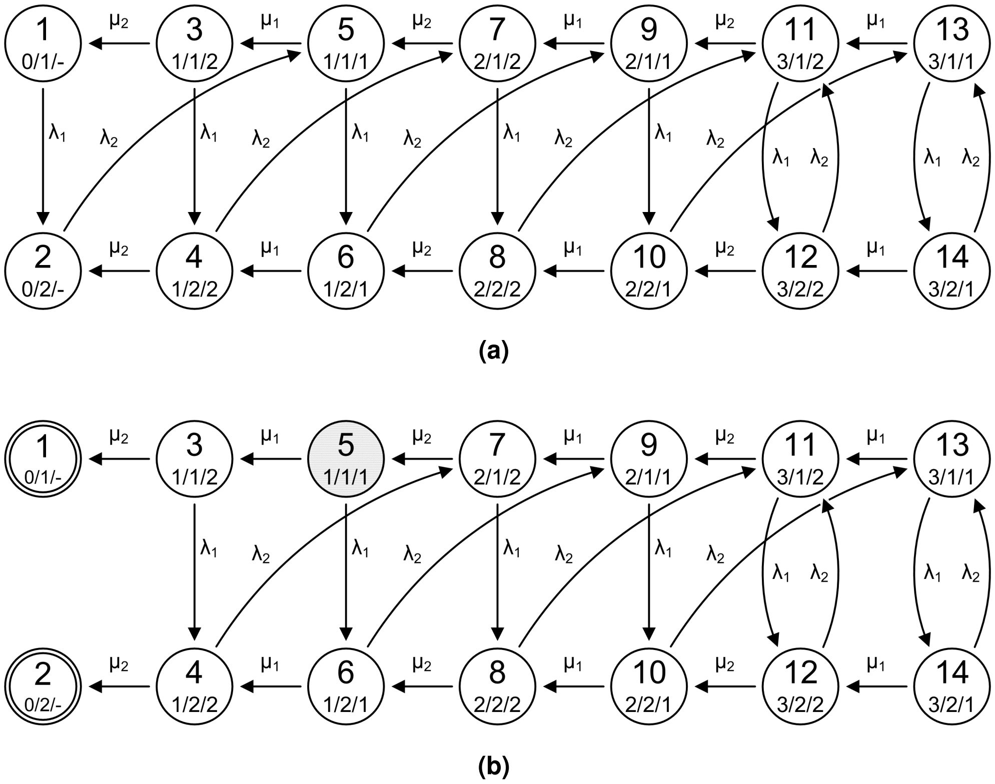


Figure 5.17.: Hypo/Hypo/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the busy period. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

The cumulative distribution function of the length of the busy period is

$$\mathrm{P}(B \leq t) = 1 - \varphi_{\langle 1/1/1 \rangle}(t) \tag{5.91}$$

The expected length of the busy period is

$$\mathrm{E}(B) = \int\limits_{t=0}^{\infty} \varphi_{\langle 1/1/1 \rangle} \mathrm{d}t \tag{5.92}$$

Hypo/Hyper/1/S queueing system

Figure 5.18a shows the Markov chain for the system state of a Hypo/Hyper/1/S queueing system. The states $\langle 0/1/- \rangle$ and $\langle 0/2/- \rangle$ are idle state, all other states are busy states. When the busy period begins, the Markov chain is in state $\langle 1/1/1 \rangle$ or in state $\langle 1/1/2 \rangle$:

$$\sigma^B_{\langle 1/1/1 \rangle} = \alpha_1 \tag{5.93}$$
$$\sigma^B_{\langle 1/1/2 \rangle} = \alpha_2 \tag{5.94}$$

The Markov chain $\mathcal{M}_B$ is shown in Figure 5.18b.

We have

$$\varphi'(t) = \mathcal{Q} \cdot \varphi(t) \qquad\qquad \varphi_{\langle i/j/k \rangle}(0) = \begin{cases} 0 & i = 0 \\ 1 & i > 0 \end{cases} \tag{5.95}$$

The cumulative distribution function of the length of the busy period is

$$\mathrm{P}(B \leq t) = 1 - \alpha_1 \varphi_{\langle 1/1/1 \rangle}(t) - \alpha_2 \varphi_{\langle 1/1/2 \rangle}(t) \tag{5.96}$$

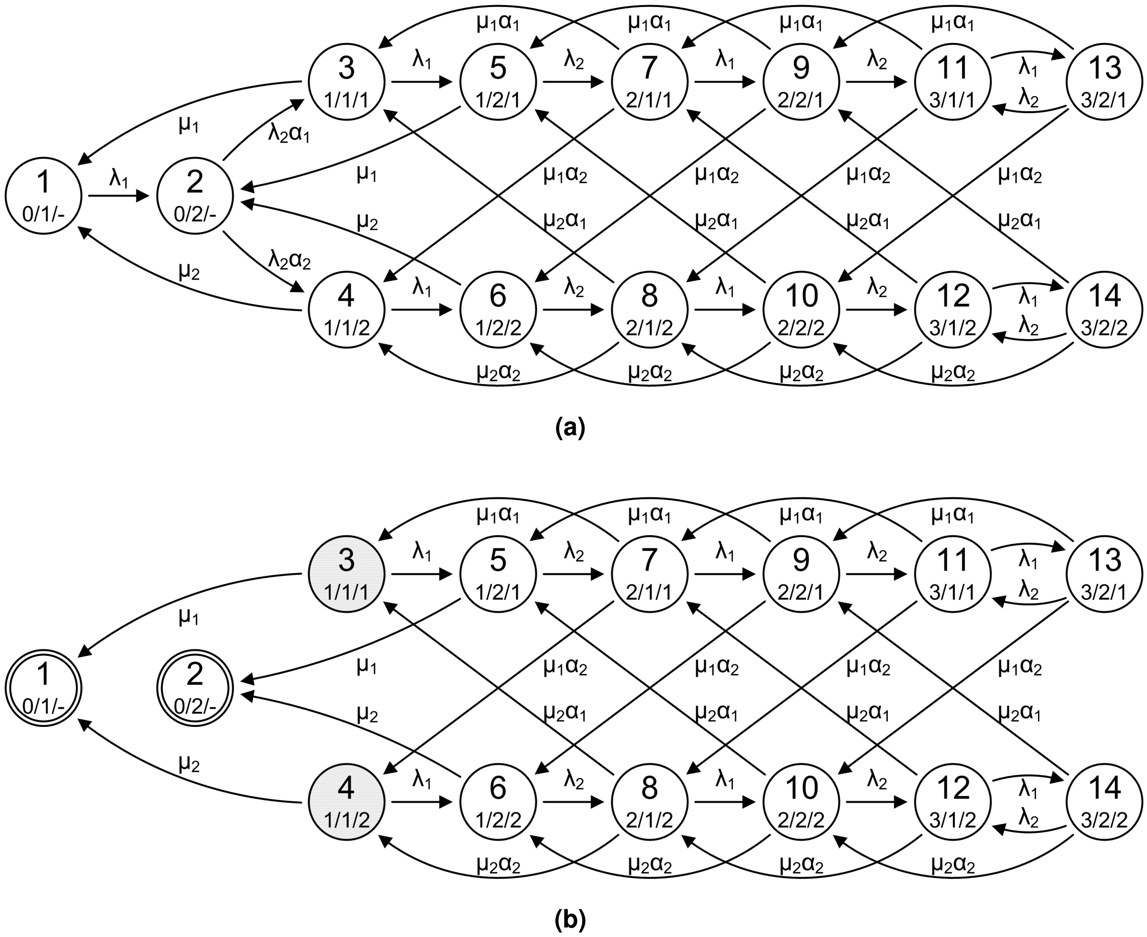
The expected length of the busy period is

$$\mathrm{E}(B) = \int\limits_{t=0}^{\infty} \alpha_1 \varphi_{\langle 1/1/1 \rangle}(t) + \alpha_2 \varphi_{\langle 1/1/2 \rangle}(t) \; \mathrm{d}t \tag{5.97}$$

Hyper/Hypo/1/S queueing system

The Markov chain for the system state of a Hyper/Hypo/1/S queueing system is shown in Figure 5.19a. In this system, the states $\langle 0/1/- \rangle$ and $\langle 0/2/- \rangle$ are idle states, all other states are busy states. When the system is empty and a customer arrives, the Markov chain is in state $\langle 1/1/1 \rangle$ with probability α_1 and in state $\langle 1/2/1 \rangle$ with probability α_2:

$$\sigma^B_{\langle 1/1/1 \rangle} = \alpha_1 \tag{5.98}$$
$$\sigma^B_{\langle 1/2/1 \rangle} = \alpha_2 \tag{5.99}$$



(a)

(b)

Figure 5.18.: Hypo/Hyper/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the busy period. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

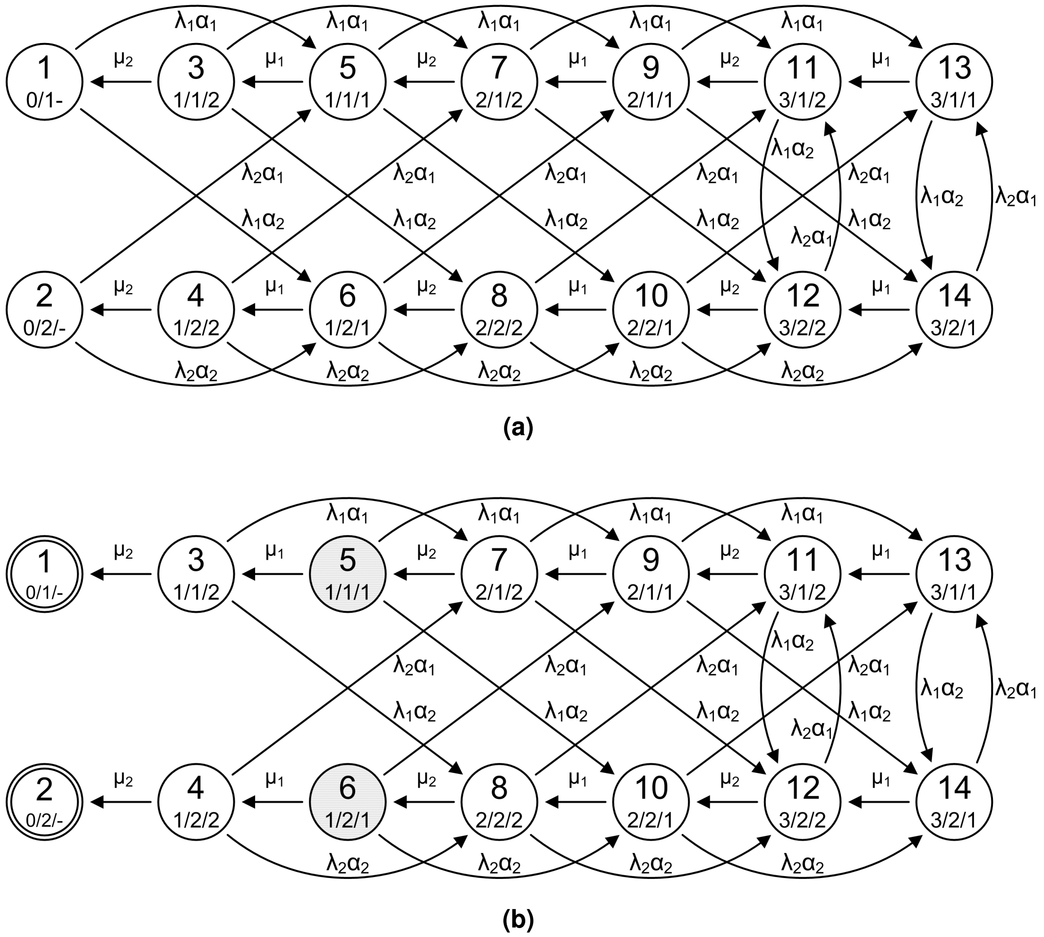


Figure 5.19.: Hyper/Hypo/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the busy period. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

The Markov chain for the calculation of the length of the busy period is shown in Figure 5.19b.

We have

$$\varphi'(t) = \mathcal{Q} \cdot \varphi(t) \qquad \varphi_{\langle i/j/k \rangle}(0) = \begin{cases} 0 & i = 0 \\ 1 & i > 0 \end{cases} \tag{5.100}$$

If the Markov chain is in state $\langle 1/1/1 \rangle$ when the busy period begins, the complementary cumulative distribution function of the length of the busy period is $\varphi_{\langle 1/1/1 \rangle}(t)$, otherwise it is $\varphi_{\langle 1/2/1 \rangle}(t)$. Therefore, the cumulative distribution function of the length of the busy period is

$$\mathrm{P}(B \leq t) = 1 - \alpha_1 \varphi_{\langle 1/1/1 \rangle}(t) - \alpha_2 \varphi_{\langle 1/2/1 \rangle}(t) \tag{5.101}$$

The expected length of the busy period is

$$\mathrm{E}(B) = \int\limits_{t=0}^{\infty} \alpha_1 \varphi_{\langle 1/1/1 \rangle}(t) + \alpha_2 \varphi_{\langle 1/2/1 \rangle}(t) \, \mathrm{d}t \tag{5.102}$$

Hyper/Hyper/1/S queueing system

Figure 5.20a shows the Markov chain for the system state of a Hyper/Hyper/1/S queueing system. The states $\langle 0/1/- \rangle$ and $\langle 0/2/- \rangle$ are idle states, all other states are busy states. When the busy period begins, the Markov chain is in one of the states $\langle 1/1/1 \rangle$, $\langle 1/1/2 \rangle$, $\langle 1/2/1 \rangle$ and $\langle 1/2/2 \rangle$:

$$\sigma^B_{\langle 1/1/1 \rangle} = \alpha_1 \beta_1 \tag{5.103}$$
$$\sigma^B_{\langle 1/1/2 \rangle} = \alpha_1 \beta_2 \tag{5.104}$$
$$\sigma^B_{\langle 1/2/1 \rangle} = \alpha_2 \beta_1 \tag{5.105}$$
$$\sigma^B_{\langle 1/2/2 \rangle} = \alpha_2 \beta_2 \tag{5.106}$$

The Markov chain for the calculation of the length of the busy period is shown in Figure 5.20b.

We have

$$\varphi'(t) = \mathcal{Q} \cdot \varphi(t) \qquad \varphi_{\langle i/j/k \rangle}(0) = \begin{cases} 0 & i = 0 \\ 1 & i > 0 \end{cases} \tag{5.107}$$

The cumulative distribution function of the length of the busy period is

$$\begin{aligned} \mathrm{P}(B \leq t) = 1 - \alpha_1 \beta_1 \varphi_{\langle 1/1/1 \rangle}(t) - \alpha_1 \beta_2 \varphi_{\langle 1/1/2 \rangle}(t) \\ - \alpha_2 \beta_1 \varphi_{\langle 1/2/1 \rangle}(t) - \alpha_2 \beta_2 \varphi_{\langle 1/2/2 \rangle}(t) \end{aligned} \tag{5.108}$$

The expected length of the busy period is

$$
\begin{aligned}
\mathrm{E}(B) = \int\limits_{t=0}^{\infty} & \alpha_1\beta_1\varphi_{\langle 1/1/1\rangle}(t) + \alpha_1\beta_2\varphi_{\langle 1/1/2\rangle}(t) \\
& + \alpha_2\beta_1\varphi_{\langle 1/2/1\rangle}(t) + \alpha_2\beta_2\varphi_{\langle 1/2/2\rangle}(t)\ \mathrm{d}t \qquad (5.109)
\end{aligned}
$$

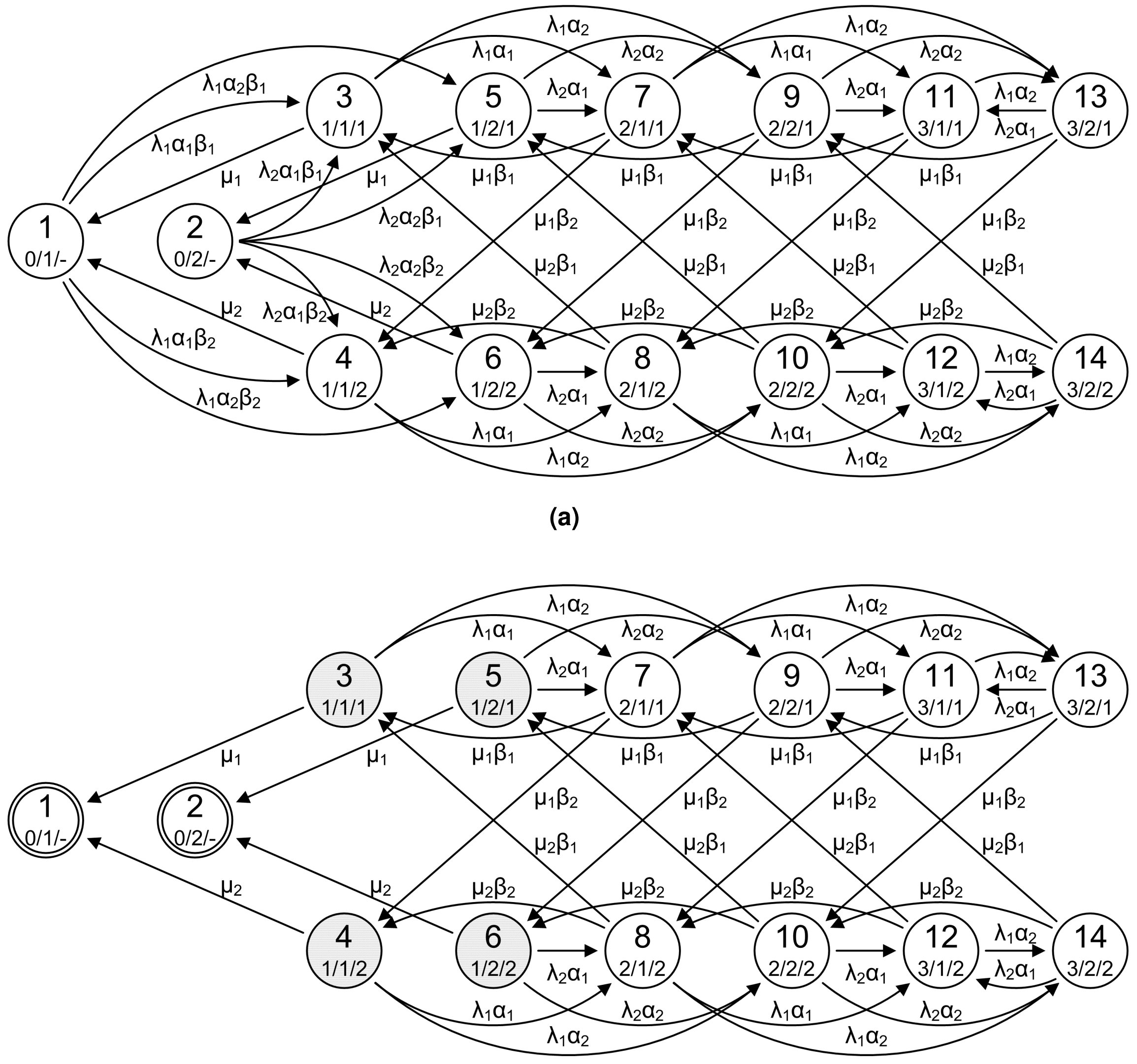


Figure 5.20.: Hyper/Hyper/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the length of the busy period. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

5.3. Number of customers served during the busy period

The length of the busy period is a compound random variable consisting of the number of customers served during the busy period ξ and the service times S_i of the single customers:

$$B = \sum_{i=1}^{\xi} S_i \tag{5.110}$$

and

$$\mathrm{E}(B) = \mathrm{E}(S_i) \cdot \mathrm{E}(\xi) \tag{5.111}$$

Therefore, if we have a constant service rate μ and know the mean length of the busy period $\mathrm{E}(B)$, we can calculate the mean number of customers being served during a busy period with

$$\mathrm{E}\left(\xi\right) = \mathrm{E}(B) \cdot \mu \tag{5.112}$$

The probability distribution of ξ is determined by extending the Markov chain $\mathcal{M}_B$ we used in the previous section for the calculation of the length of the busy period with a counting Markov chain $\mathcal{M}_C$, which counts services.

We start the observation at the moment when the busy period begins. Since there has not been a service yet, the new Markov chain $\mathcal{M}_\xi$ is in a state $\langle b/0\rangle, b \in \mathcal{B}$. The probabilities $\sigma^{\xi}_{\langle b/0\rangle}$ that it is in state $\langle b/0\rangle$ are

$$\sigma^{\xi}_{\langle b/0\rangle} = \sigma^{B}_{b} \tag{5.113}$$

As time goes by and $\mathcal{M}_\xi$ evolves, the counting Markov chain $\mathcal{M}_C$ increases its value whenever there is a service, until eventually $\mathcal{M}_\xi$ reaches an absorbing state $\langle i/n\rangle, i \in \mathcal{I}, n \in \mathbb{N}^+$. This means that the system has become idle and we are interested in the number of counted services.

So we have

$$\pi_{\langle b/0\rangle}(0) = \sigma^{B}_{b} \qquad b \in \mathcal{B} \tag{5.114}$$

$$\pi'(\tau) = \pi(\tau) \cdot \mathcal{Q} \tag{5.115}$$

If we had an infinite counting Markov chain $\mathcal{M}_C$, the probability that there were n services during the busy period would be

$$\mathrm{P}\left(\xi = n\right) = \sum_{i \in \mathcal{I}} \lim_{t \to \infty} \pi_{\langle i/n\rangle}(t) \tag{5.116}$$

However, since we do not concern ourselves with infinite Markov chains in this work, $\mathcal{M}_C$ is finite. Assume it counts from 0 to N. In this case we have

$$\mathrm{P}\left(\xi = n\right) = \sum_{i \in \mathcal{I}} \lim_{t \to \infty} \pi_{\langle i/n\rangle}(t) \qquad n < N \tag{5.117}$$

$$\mathrm{P}\left(\xi \geq N\right) = \sum_{i \in \mathcal{I}} \lim_{t \to \infty} \pi_{\langle i/N\rangle}(t) \tag{5.118}$$

In practice the computation of $\pi(t)$ can be stopped when the Markov chain is with very high probability in a state $\langle i/\cdot\rangle, i \in \mathcal{I}$, that is, $\sum_{n=0}^{N} \sum_{i\in\mathcal{I}} \pi_{\langle i/n\rangle}(t) \approx 1$.

It should be noted that it is not possible to obtain the limiting probability distribution $\lim_{t\to\infty} \pi(t)$, which is also a stationary probability distribution, by solving the system of linear equations $\pi \cdot \mathcal{Q} = 0$ or by using other techniques for the calculation of unique stationary state probabilities. The reason is that the Markov chain $\mathcal{M}_\xi$ has absorbing states, and therefore the stationary probability distribution is not unique, but depends on the initial state.

We can also use the following method to calculate the number of customers served during the busy period:

Suppose the Markov chain $\mathcal{M}_B$ is in state i and let $C(i)$ be the expected number of remaining services in the current busy period given that the Markov chain is in state i. $C(i)$ depends on the state taken next. If the next transition is to state j (probability $-q_{ij}/q_{ii}$) and corresponds to a service, then $C(i) = C(j) + 1$. If the transition to j does not correspond to a service, then $C(i) = C(j)$. For idle states we have $C(i) = 0$.

$$C(i) = \begin{cases} \sum\limits_{j\neq i} \frac{q_{ij}}{-q_{ii}}\ \left(C(j) + s(i,j)\right) & i \notin \mathcal{I} \\ 0 & i \in \mathcal{I} \end{cases} \tag{5.119}$$

where

$$s(i,j) = \begin{cases} 1 & \text{transition } i \to j \text{ corresponds to a service} \\ 0 & \text{otherwise} \end{cases} \tag{5.120}$$

Now the mean number of customers served during the busy period is

$$\mathrm{E}(\xi) = \sum_{b\in\mathcal{B}} C(b)\sigma_b^B \tag{5.121}$$

Let further be $T(i,n)$ the probability that there will be n services until an idle state is reached, given that the Markov chain is in state i. $T(i,n)$ depends on the next transition. If the next transition leads to state j (probability $-q_{ij}/q_{ii}$) and corresponds to a service, then $T(i,n) = T(j, n-1)$. If the transition to j does not correspond to a service, we have $T(i,n) = T(j,n)$:

$$T(i,n) = \begin{cases} \sum\limits_{j\neq i} \frac{q_{ij}}{-q_{ii}} T(j, n - s(i,j)) & i \notin \mathcal{I}, n \geq 0 \\ 0 & i \in \mathcal{I}, n > 0 \\ 1 & i \in \mathcal{I}, n = 0 \\ 0 & n < 0 \end{cases} \tag{5.122}$$

where

$$s(i,j) = \begin{cases} 1 & \text{transition } i \to j \text{ corresponds to a service} \\ 0 & \text{otherwise} \end{cases} \tag{5.123}$$

The probability that there are n customers served during the busy period is

$$\mathrm{P}(\xi = n) = \sum_{b \in \mathcal{B}} T(b, n)\sigma_b^B \tag{5.124}$$

5.3.1. M/M/1/S queueing system

The construction of the Markov chain $\mathcal{M}_\xi$ for an M/M/1/S queueing system is shown in Figure 5.21. Figure 5.21a shows the Markov chain $\mathcal{M}_B$ used for the calculation of the length of the busy period. This Markov chain is combined with the counting Markov chain $\mathcal{M}_C$ shown in Figure 5.21b, which increases its value when a service takes place.

When the busy period begins, $\mathcal{M}_B$ is in state $\langle 1 \rangle$ and $\mathcal{M}_C$ is in state $\langle 0 \rangle$ (because by then there were no services in the current busy period). Therefore, $\mathcal{M}_\xi$ is in state $\langle 1/0 \rangle$:

$$\pi_i(0) = \begin{cases} 1 & i = \langle 1/0 \rangle \\ 0 & \text{otherwise} \end{cases} \tag{5.125}$$

Now we calculate the state probabilities with

$$\pi'(\tau) = \pi(\tau) \cdot \mathcal{Q} \tag{5.126}$$

until, for example,

$$1 - \sum_{n=0}^{N} \pi_{\langle 0/n \rangle}(\tau) < 10^{-6} \tag{5.127}$$

The probability distribution of ξ is

$$\mathrm{P}(\xi = n) = \lim_{t \to \infty} \pi_{\langle 0/n \rangle}(t) \qquad n < N \tag{5.128}$$

$$\mathrm{P}(\xi \geq N) = \lim_{t \to \infty} \pi_{\langle 0/N \rangle}(t) \tag{5.129}$$

or

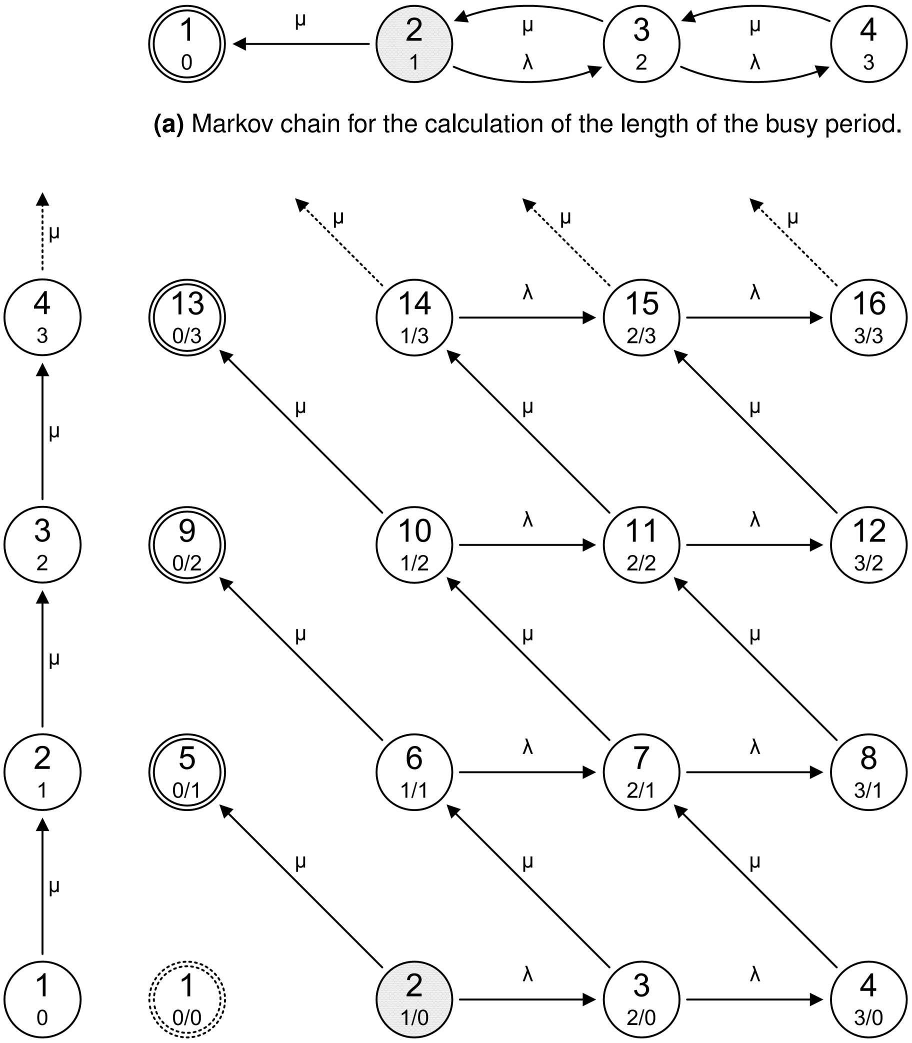
$$\mathrm{P}(\xi > n) = \sum_{i=n}^{N} \lim_{t \to \infty} \pi_{\langle 0/n \rangle}(t) \qquad n \leq N \tag{5.130}$$

Given the mean length of the busy period we can calculate $\mathrm{E}(\xi)$ with

$$\mathrm{E}(\xi) = \mathrm{E}(B)\,\mu \tag{5.131}$$

By using the closed-form solution for the mean length of the busy period (Equation 5.54), we obtain

$$\mathrm{E}(\xi) = \begin{cases} \frac{1-(\lambda/\mu)^S}{1-\lambda/\mu} & \lambda \neq \mu \\ S & \lambda = \mu \end{cases} \tag{5.132}$$



(a) Markov chain for the calculation of the length of the busy period.

(b) Counting Markov chain.

(c) Markov chain for the calculation of the probability distribution of the number of customers served during a busy period. Meaning of the names of the states: number of customers in the system / number of counted services. Absorbing states are painted with double lines, states in which the system is when the busy period begins are shaded grey.

Figure 5.21.: M/M/1/S queueing system ($S = 3$): calculation of the number of customers served during a busy period

Alternative method for the calculation of the number of customers served during the busy period:

$$\begin{aligned}
C(2) &= \frac{\mu}{\lambda+\mu}(1+\underbrace{C(1)}_{0}) + \frac{\lambda}{\lambda+\mu}C(3) \\
C(3) &= \frac{\mu}{\lambda+\mu}(1+C(2)) + \frac{\lambda}{\lambda+\mu}C(4) \\
C(4) &= 1+C(3)
\end{aligned} \tag{5.133}$$

Solving this system of equations yields

$$\begin{aligned}
C(2) &= \frac{\mu^2+\lambda\mu+\lambda^2}{\mu^2} \\
C(3) &= \frac{2\mu^2+2\lambda\mu+\lambda^2}{\mu^2} \\
C(4) &= \frac{3\mu^2+2\lambda\mu+\lambda^2}{\mu^2}
\end{aligned} \tag{5.134}$$

The Markov chain is in state 2 when the busy period begins, therefore,

$$\mathrm{E}\left(\xi\right) = C(2) = \frac{\mu^2+\lambda\mu+\lambda^2}{\mu^2} = 1+\frac{\lambda}{\mu}+\frac{\lambda^2}{\mu^2} \tag{5.135}$$

Moreover, $T(2,n)$ is the probability that n customers are served during the busy period. We show the calculation for $n = 1 \dots 3$.

$n = 1:$

$$\begin{aligned}
T(2,1) &= \underbrace{T(1,0)}_{1}\frac{\mu}{\lambda+\mu} + T(3,1)\frac{\lambda}{\lambda+\mu} \\
T(3,1) &= T(2,0)\frac{\mu}{\lambda+\mu} + T(4,1)\frac{\lambda}{\lambda+\mu} \\
T(4,1) &= T(3,0) \\
T(3,0) &= \underbrace{T(2,-1)}_{0}\frac{\mu}{\lambda+\mu} + T(4,0)\frac{\lambda}{\lambda+\mu} \\
T(4,0) &= \underbrace{T(3,-1)}_{0} \\
T(4,0) &= 0 \\
T(3,0) &= 0 \\
T(4,1) &= 0 \\
T(2,0) &= \underbrace{T(1,-1)}_{0}\frac{\mu}{\lambda+\mu} + \underbrace{T(3,0)}_{0}\frac{\lambda}{\lambda+\mu} \\
T(2,0) &= 0 \\
T(3,1) &= 0 \\
T(2,1) &= \frac{\mu}{\lambda+\mu}
\end{aligned} \tag{5.136}$$

$n = 2:$

$$\begin{aligned}
T(2,2) &= \underbrace{T(1,1)}_{0}\frac{\mu}{\lambda+\mu} + T(3,2)\frac{\lambda}{\lambda+\mu} \\
T(3,2) &= \underbrace{T(2,1)}_{\mu/(\lambda+\mu)}\frac{\mu}{\lambda+\mu} + T(4,2)\frac{\lambda}{\lambda+\mu} \\
T(4,2) &= \underbrace{T(3,1)}_{0} \\
T(4,2) &= 0 \\
T(3,2) &= \frac{\mu^2}{(\lambda+\mu)^2} \\
T(2,2) &= \frac{\lambda\mu^2}{(\lambda+\mu)^3}
\end{aligned} \tag{5.137}$$

$n = 3:$

$$\begin{aligned}
T(2,3) &= \underbrace{T(1,2)}_{0}\frac{\mu}{\lambda+\mu} + T(3,3)\frac{\lambda}{\lambda+\mu} \\
T(3,3) &= \underbrace{T(2,2)}_{\lambda\mu^2/(\lambda+\mu)^3}\frac{\mu}{\lambda+\mu} + T(4,3)\frac{\lambda}{\lambda+\mu} \\
T(4,3) &= \underbrace{T(3,2)}_{\mu^2/(\lambda+\mu)^2} \\
T(4,3) &= \frac{\mu^2}{(\lambda+\mu)^2} \\
T(3,3) &= \frac{\lambda\mu^2(\lambda+2\mu)}{(\lambda+\mu)^4} \\
T(2,3) &= \frac{\lambda^2\mu^2(\lambda+2\mu)}{(\lambda+\mu)^5}
\end{aligned} \tag{5.138}$$

So we have

$$\begin{aligned}
\mathrm{P}(\xi = 1) &= \frac{\mu}{\lambda+\mu} \\
\mathrm{P}(\xi = 2) &= \frac{\lambda\mu^2}{(\lambda+\mu)^3} \\
\mathrm{P}(\xi = 3) &= \frac{\lambda^2\mu^2(\lambda+2\mu)}{(\lambda+\mu)^5}
\end{aligned} \tag{5.139}$$

We also can obtain a general solution:

$$\mathrm{P}(\xi = 1) = \frac{\mu}{\lambda+\mu} \tag{5.140}$$

$$\mathrm{P}(\xi = n) = \frac{\lambda^{n-1}\mu^2(\lambda+2\mu)^{n-2}}{(\lambda+\mu)^{2n-1}} \qquad n \geq 2 \tag{5.141}$$

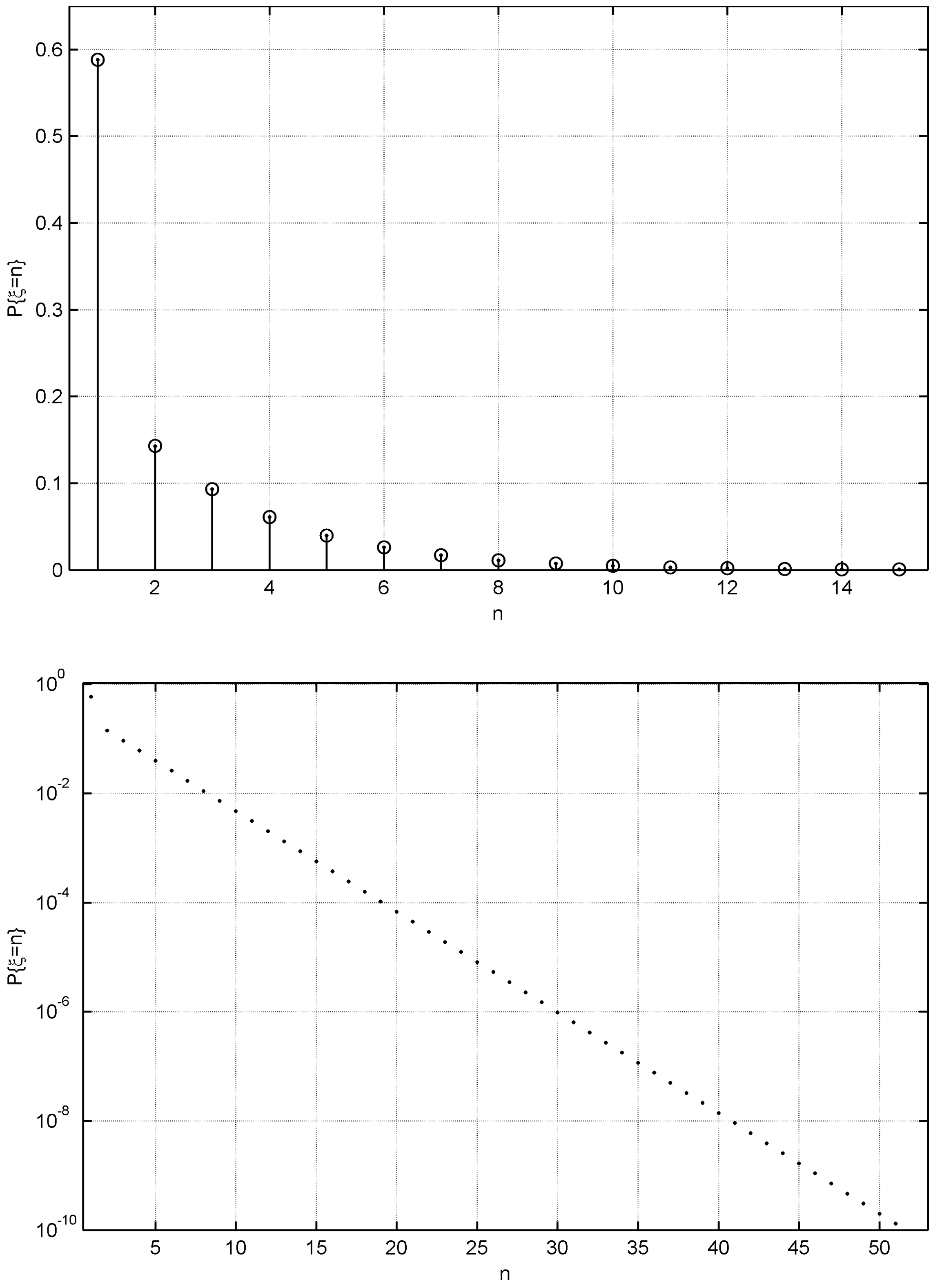


Figure 5.22.: M/M/1/S queueing system: number of customers served during the busy period. $S = 3, \lambda = 0.7, \mu = 1$.
Dots: calculation, circles: simulation.

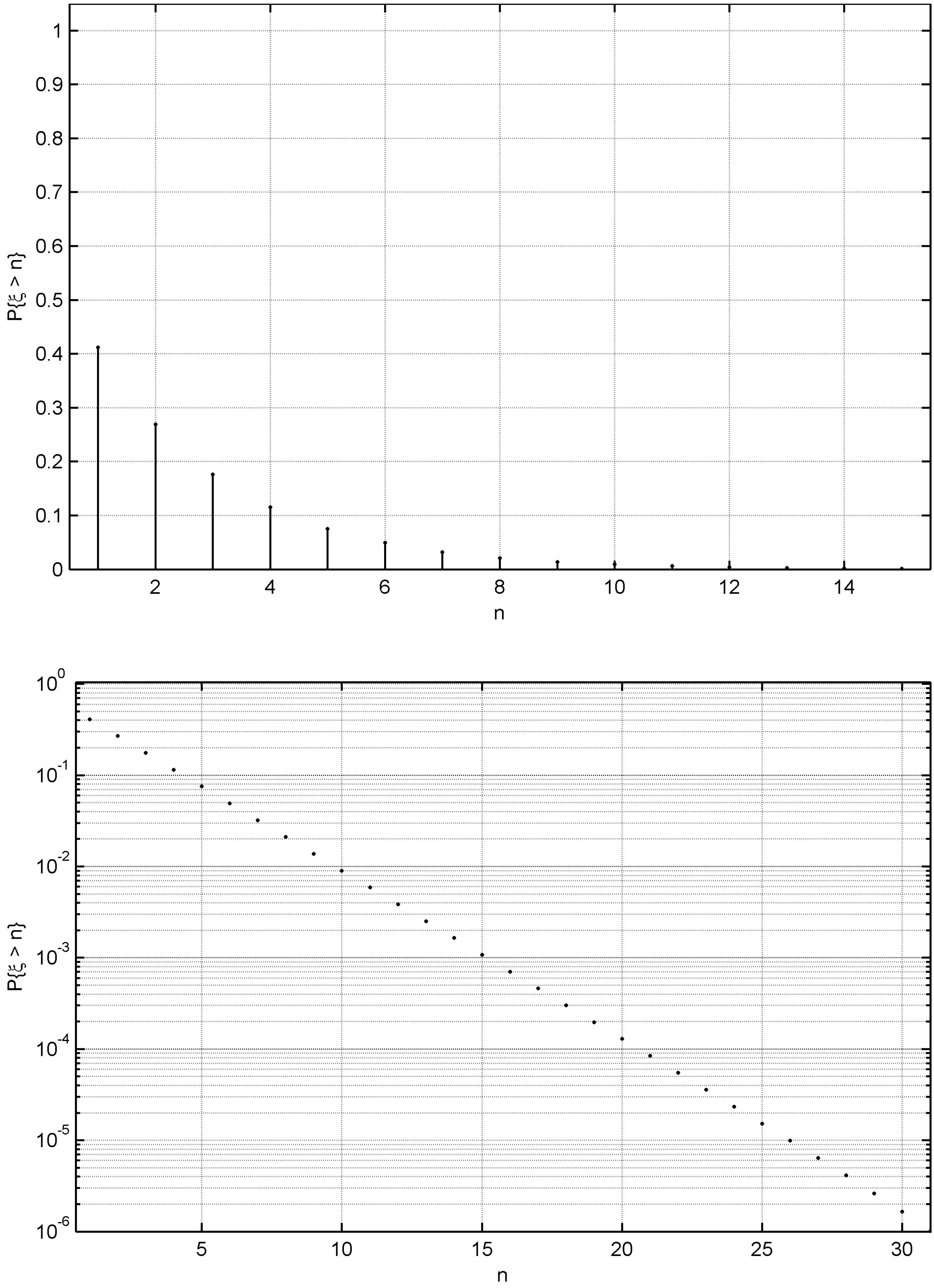


Figure 5.23.: M/M/1/S queueing system: number of customers served during the busy period. $S = 3, \lambda = 0.7, \mu = 1.$

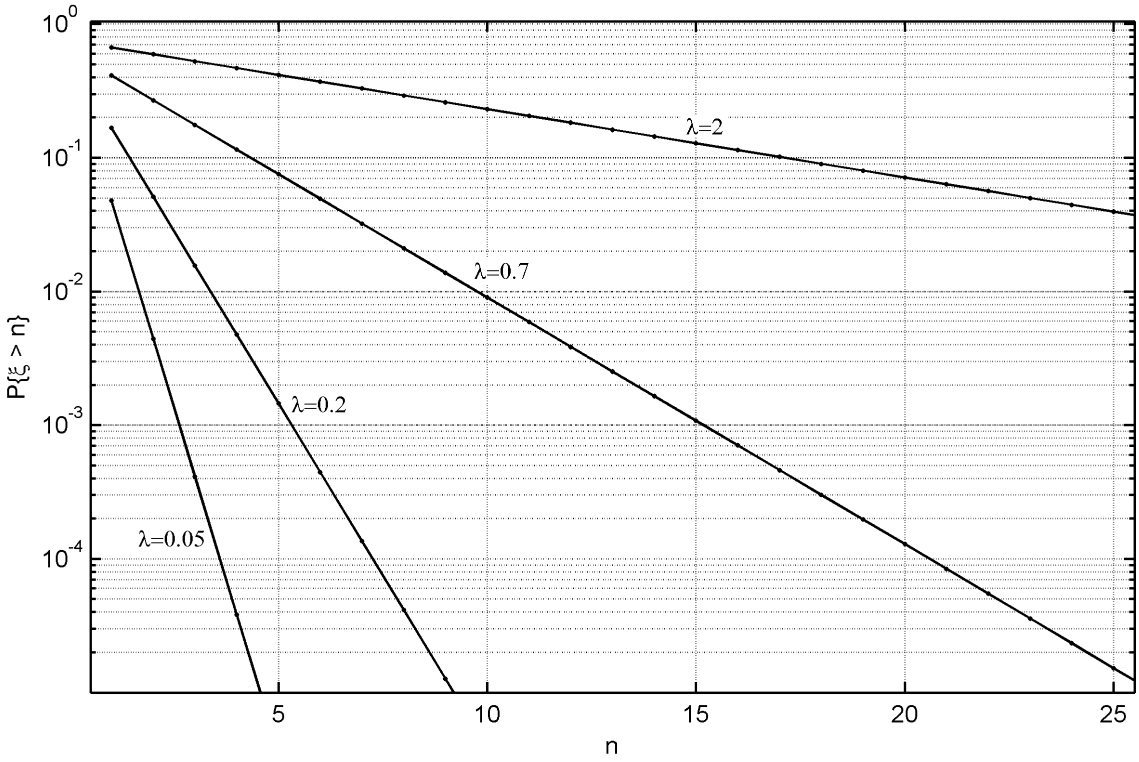


Figure 5.24.: M/M/1/S queueing system: number of customers served during the busy period for various arrival rates. $S = 3, \mu = 1$.

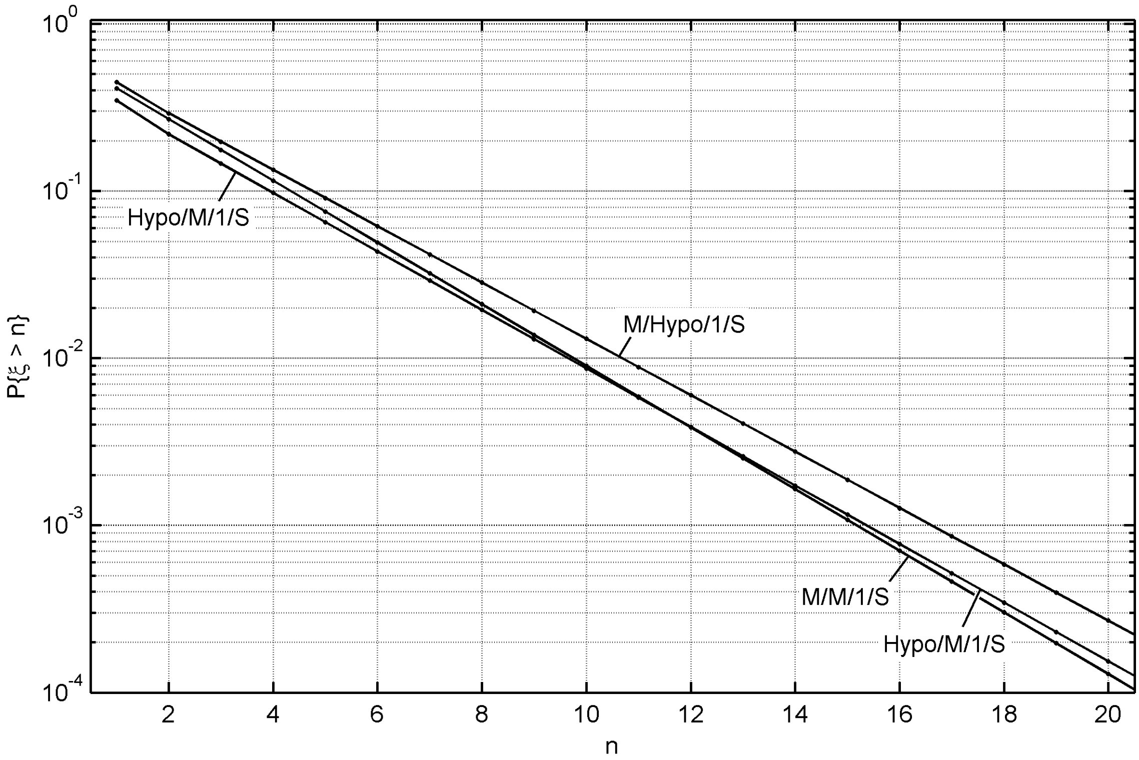


Figure 5.25.: Number of customers served during the busy period in an M/M/1/S queueing system, a Hypo/M/1/S queueing system ($c_A = 0.75$) and an M/Hypo/1/S queueing system ($c_S = 0.75$). All: $S = 3, \lambda = 1, \mu = 1$.

An alternative approach to model an M/M/1/S queueing system

In Chapter 6, we have a situation where it would be helpful to be able to describe the state of a queueing system by the number of remaining services in the current busy period. So we could try to model an M/M/1/S queueing system with the Markov chain shown in Figure 5.26.

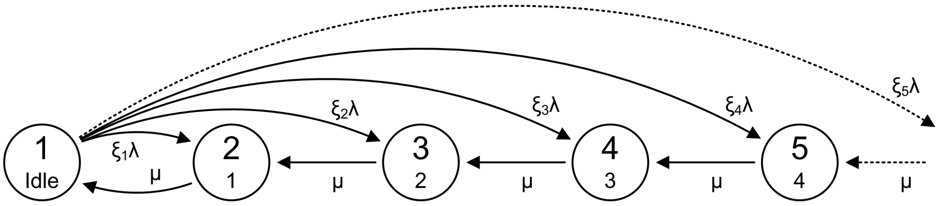


Figure 5.26.: An alternative approach to model an M/M/1/S queueing system. This approach does not work! Meaning of the names of the states: remaining number of customers served in the current busy period, or "Idle" if the queueing system is empty. $\xi_i = \mathrm{P}\,(\xi = i)$.

Unfortunately, this is not possible. Although the service times are independent, the service times within a busy period are not independent. For example, if $S = 3$, it is very unlikely that we would have 8 consecutive short service times within the same busy period: Unless we have a few short interarrival times at the same time, the busy period ends after around 3 short service times because the system becomes empty. That is, the probability that within a busy period we have 8 short service times is the probability that we have 8 short service times and that we have, in addition, a few short interarrival times.

Another reason is that the transition rate between state *"n remaining services in the current busy period"* and state *"$n-1$ remaining services in the current busy period"* depends on n: The last service in a busy period finishes at rate $\lambda + \mu$ instead of μ, because a service time can be the last one only if it is smaller than the next interarrival time.[2] The second to last service is not that short, but also shorter than an average service time. (If it were longer, it would be likely that customers arrive, which would prevent it from being the second to last by prolonging the busy period.) The first service time in the busy period tends to be longer than average, and so on.

Figures 5.27 and 5.28 show the actual transition rates and the coefficient of variation of the sojourn times in the states of the process we tried to model with the Markov chain shown in Figure 5.26. We see that the rates and coefficients of variation depend not only on the state, but also on ξ.

[2] The smallest of independent exponential random variables $X_1, X_2, \ldots, X_n$ with respective rates $\lambda_1, \lambda_2, \ldots, \lambda_n$ is also an exponential random variable, whose rate is the sum of the λ_js: $\mathrm{P}\left\{\min_j X_j \geq x\right\} = \mathrm{P}\left\{X_j \geq x\right\} \quad \forall j = \prod_j \mathrm{P}\left\{X_j \geq x\right\} = \prod_j \mathrm{e}^{-\lambda_j x} = \mathrm{e}^{-\sum_j \lambda_j x}$. The last expression is the complementary cumulative distribution function for an exponential random variable with rate $\sum_j \lambda_j$.

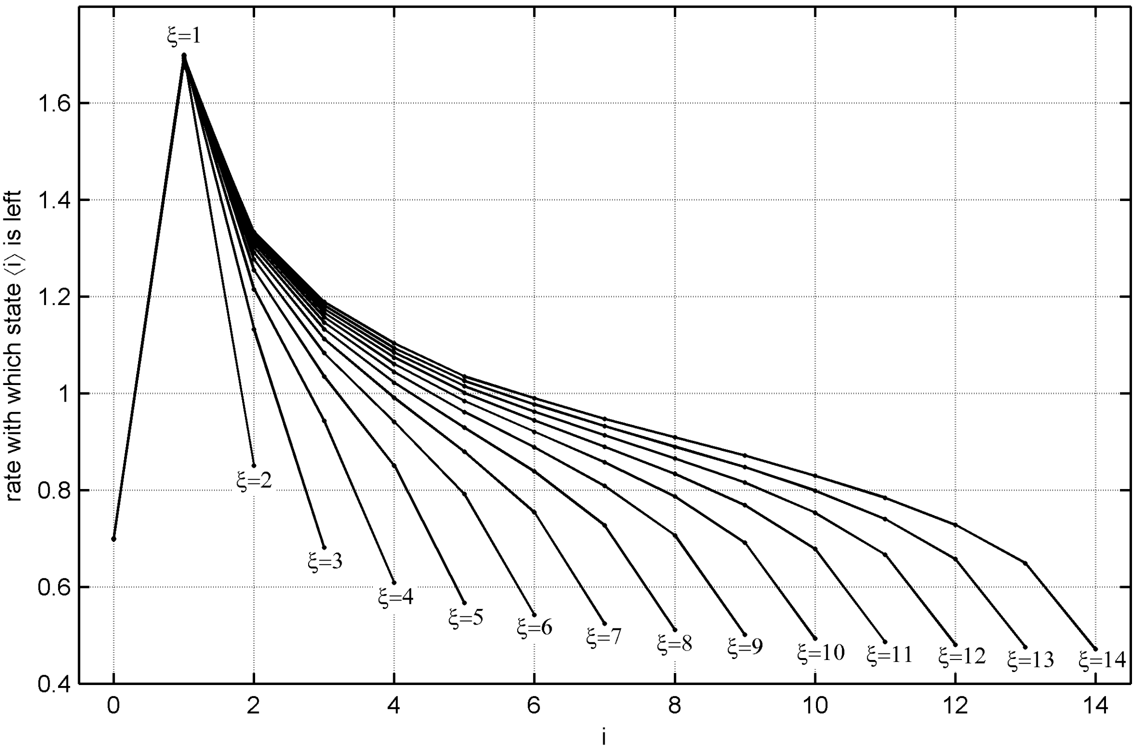


Figure 5.27.: Actual transition rates in the process we tried to describe by the Markov chain shown in Figure 5.26. $S = 3, \lambda = 0.7, \mu = 1$. Simulation study.

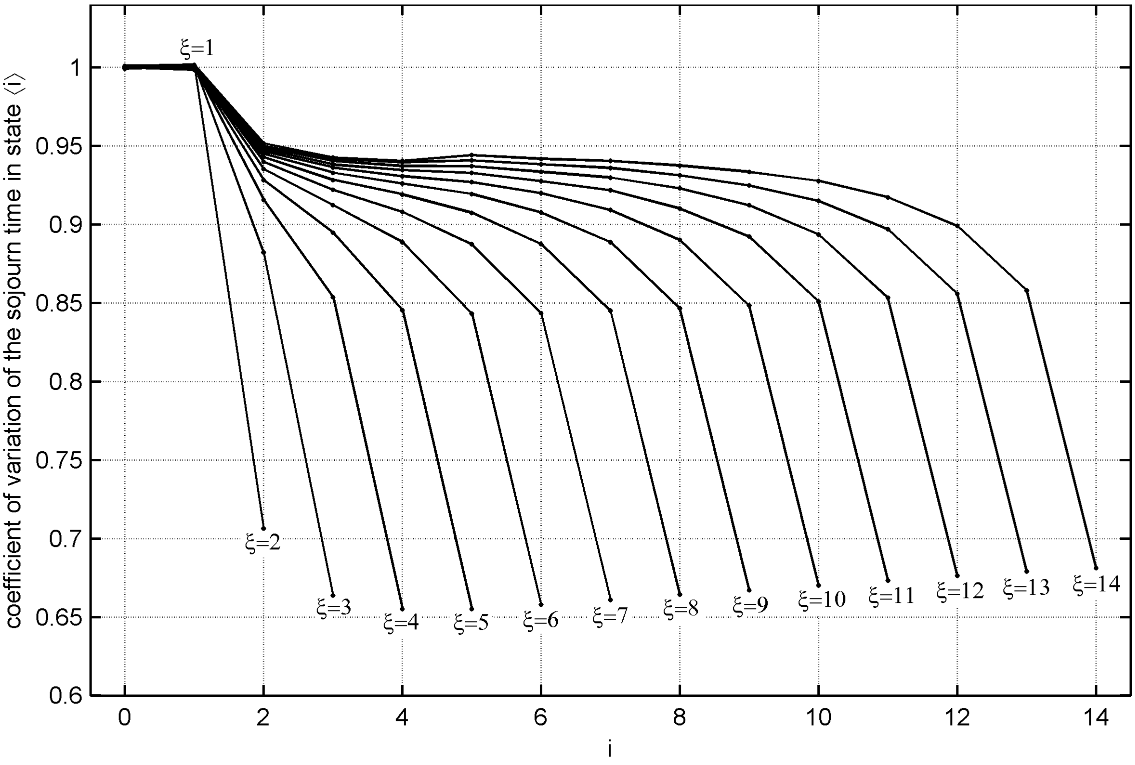


Figure 5.28.: Actual coefficient of variation of the sojourn times in the states of the process we tried to describe by the Markov chain shown in Figure 5.26. $S = 3, \lambda = 0.7, \mu = 1$. Simulation study.

5.3.2. Hypo/M/1/S queueing system

Figure 5.29 shows the Markov chain for the calculation of the probability distribution of the number of customers served during a busy period in a Hypo/M/1/S queueing system.

When the busy period begins, there is one customer in the system, the arrival process is in its initial state and there have not yet been any services in the current busy period. Therefore, the Markov chain is in state $\langle 1/1/0 \rangle$:

$$\pi_i(0) = \begin{cases} 1 & i = \langle 1/1/0 \rangle \\ 0 & \text{otherwise} \end{cases} \tag{5.142}$$

Now we calculate the state probabilities with

$$\pi'(\tau) = \pi(\tau) \cdot \mathcal{Q} \tag{5.143}$$

until, for example,

$$1 - \sum_{n=0}^{N} \left(\pi_{\langle 0/1/n \rangle}(\tau) + \pi_{\langle 0/2/n \rangle}(\tau) \right) < 10^{-6} \tag{5.144}$$

The probability distribution of ξ is

$$\mathrm{P}\,(\xi = n) = \lim_{t \to \infty} \pi_{\langle 0/1/n \rangle}(t) + \pi_{\langle 0/2/n \rangle}(t) \qquad n < N \tag{5.145}$$

$$\mathrm{P}\,(\xi \geq N) = \lim_{t \to \infty} \pi_{\langle 0/1/N \rangle}(t) + \pi_{\langle 0/2/N \rangle}(t) \tag{5.146}$$

or

$$\mathrm{P}\,(\xi \geq n) = \sum_{i=n}^{N} \lim_{t \to \infty} \pi_{\langle 0/1/n \rangle}(t) + \pi_{\langle 0/2/n \rangle}(t) \qquad n \leq N \tag{5.147}$$

5.3.3. M/Hypo/1/S queueing system

The Markov chain for the calculation of the probability distribution of the number of customers served during a busy period of an M/Hypo/1/S queueing system is shown in Figure 5.30.

When the system is idle and a customer arrives, there is one customer in the system, the service process is in its initial state and there have not yet been any services in the current busy period. Therefore, the Markov chain is in state $\langle 1/1/0 \rangle$:

$$\pi_i(0) = \begin{cases} 1 & i = \langle 1/1/0 \rangle \\ 0 & \text{otherwise} \end{cases} \tag{5.148}$$

Now we calculate the state probabilities with

$$\pi'(\tau) = \pi(\tau) \cdot \mathcal{Q} \tag{5.149}$$

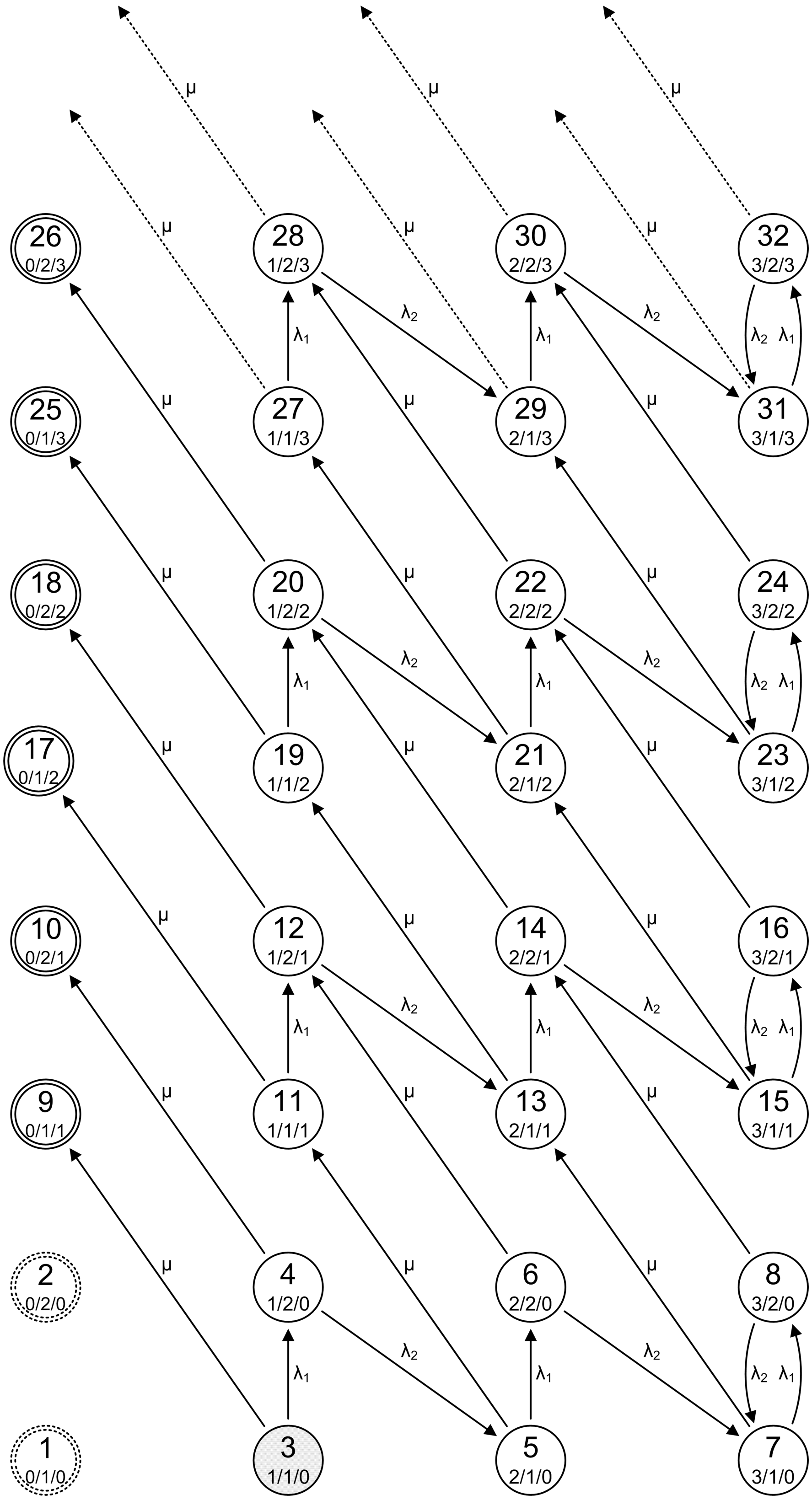


Figure 5.29.: Hypo/M/1/S queueing system: calculation of the number of customers served during a busy period. Meaning of the names of the states: number of customers in the system / state of the arrival process / number of counted services.

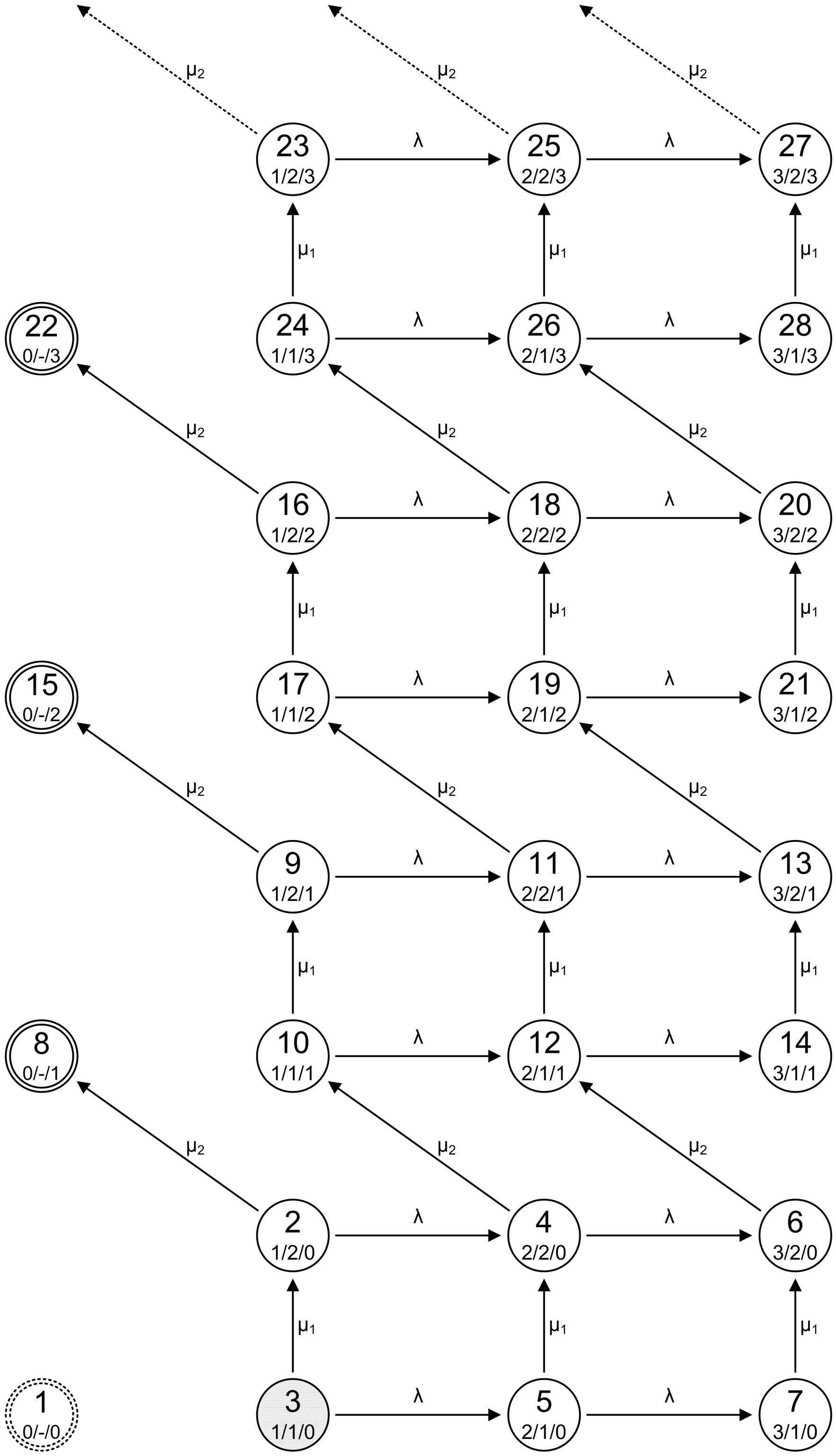


Figure 5.30.: M/Hypo/1/S queueing system: calculation of the number of customers served during a busy period. Meaning of the names of the states: number of customers in the system / state of the service process / number of counted services.

until, for example,

$$1 - \sum_{n=0}^{N} \pi_{\langle 0/-/n \rangle}(\tau) < 10^{-6} \tag{5.150}$$

The probability distribution of ξ is

$$\mathrm{P}\left(\xi = n\right) = \lim_{t\to\infty} \pi_{\langle 0/-/n \rangle}(t) \qquad n < N \tag{5.151}$$

$$\mathrm{P}\left(\xi \geq N\right) = \lim_{t\to\infty} \pi_{\langle 0/-/N \rangle}(t) \tag{5.152}$$

or

$$\mathrm{P}\left(\xi \geq n\right) = \sum_{i=n}^{N} \lim_{t\to\infty} \pi_{\langle 0/-/n \rangle}(t) \qquad n \leq N \tag{5.153}$$

6. Departure stream

This chapter deals with the departure stream of queueing systems. The departure stream is the stream created by those customers who have been served in a queueing system and leave it.

The interdeparture times of single-server queueing systems can be calculated easily if we know the length of the idle period and the number of customers served during the busy period. We show how to do this in Section 6.1.

In most cases, the interdeparture times of a queueing system are not independent. In Section 6.2 we shall briefly discuss this topic.

Finally, we show how to model a small queueing network where the departure stream of a queueing system constitutes the arrival stream of another queueing system. We first model the queueing network by using one single Markov chain for both systems (Section 6.3), and then we show how the network can be analysed using network decomposition (Section 6.4).

6.1. Interdeparture times

Under the assumption of state-independent service rates, the interdeparture times of a single-server queueing system consist of two types of random variables: The first interdeparture time $D^{(1)}$ in a busy cycle (that is, a cycle consisting of the idle period and the busy period) is the sum of the length of the idle period I and the first service time S. The remaining interdeparture times $D^{(2)}$ equal the service times (see Figure 6.1).

$$D^{(1)} = I + S \tag{6.1}$$

$$D^{(2)} = S \tag{6.2}$$

If we know the number of customers served during a busy period ξ, we can calculate the average interdeparture time of the sequence

$$\underbrace{I+S, S, S}_{\xi \text{ departures}},\ \underbrace{I+S}_{\xi \text{ departures}},\ \underbrace{I+S, S, S, S, S, S}_{\xi \text{ departures}},\ \underbrace{I+S, S}_{\xi \text{ departures}}, \ldots$$

with

$$D = \frac{(I+S) + (\xi - 1)\, S}{\xi} \tag{6.3}$$

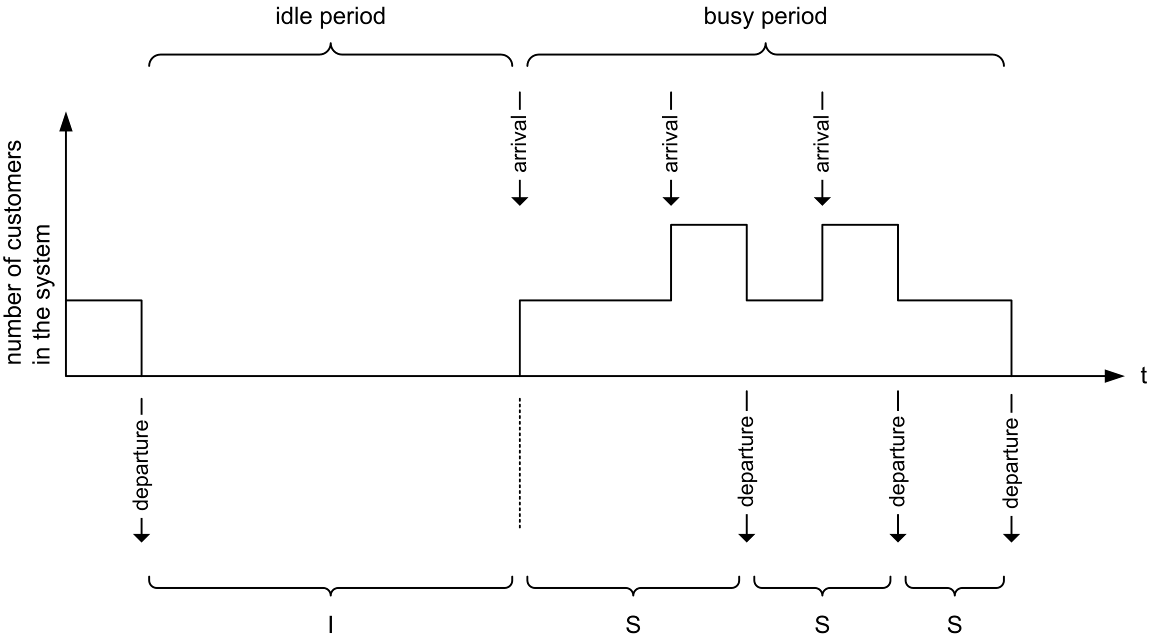


Figure 6.1.: Departures of a single-server queueing system during a cycle consisting of the idle period and the busy period.

The probability density function of the interdeparture times is

$$f_D(t) = \frac{(f_I(t) * f_S(t)) + (\xi - 1)\, f_S(t)}{\xi} \tag{6.4}$$

Often $I + S$ is expressible as a weighted sum of hypoexponential random variables:

$$I + S \sim \sigma^I_{i_1} \,\mathrm{HypoExp}(\cdots) + \sigma^I_{i_2} \,\mathrm{HypoExp}(\cdots) + \ldots \tag{6.5}$$

If this is not possible, the distribution of $I+S$ should not be calculated by first calculating I and S and then convolving the probability density functions,

$$f_{I+S}(t) = \int\limits_{-\infty}^{\infty} f_S(\tau) f_I(t - \tau) \mathrm{d}\tau \tag{6.6}$$

It is more accurate to calculate the complementary cumulative distribution function of $I + S$ by means of the Markov chain $\mathcal{M}_I$ we used for the calculation of the length of the busy period: We add states and transitions that describe the first service of a customer to all states in which the Markov chain can be when the busy period begins. Then we calculate the time needed to go from the states in which the Markov chain is when the idle period begins to reach the busy period *and* to traverse the added transitions that describe the first service.

The mean of the interdeparture times can be calculated from $f_D(t)$ with

$$\mathrm{E}(D) = \int\limits_{0}^{\infty} t\ f_D(t)\ \mathrm{d}t \tag{6.7}$$

or from the service rate μ and the utilisation ρ with

$$\mathrm{E}(D) = \frac{1}{\mu\,\rho} \tag{6.8}$$

Another method, which also can be used if the service rates are state-dependent, is to calculate the departure rate from the stationary state probabilities and the rate at which each state produces departures:

$$\frac{1}{\mathrm{E}(D)} = \sum_{b\in\mathcal{B}} \sum_{j\neq b} h_{bj}\pi_b \tag{6.9}$$

where

$$h_{bj} = \begin{cases} q_{bj} & \text{if } b \to j \text{ corresponds to a departure} \\ 0 & \text{otherwise} \end{cases} \tag{6.10}$$

6.1.1. M/M/1/S queueing system

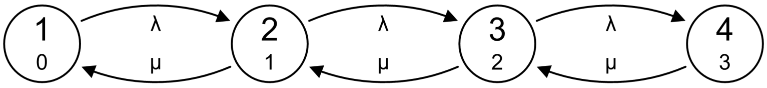


Figure 6.2.: M/M/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system.

As can be seen from Figure 6.2, the length of the idle period of an M/M/1/S queueing system is exponentially distributed with rate λ. The service times are exponentially distributed with rate μ:

$$I \sim \mathrm{Exp}(\lambda) \tag{6.11}$$
$$S \sim \mathrm{Exp}(\mu) \tag{6.12}$$

Therefore, we have

$$D = \frac{(I+S) + (\xi-1)\,S}{\xi} \sim \frac{\mathrm{HypoExp}(\lambda,\mu) + (\xi-1)\,\mathrm{Exp}(\mu)}{\xi} \tag{6.13}$$

Using the closed-form solution for $\mathrm{E}(\xi)$ (Equation 5.132) we obtain

for $\lambda \neq \mu$:

$$\begin{aligned}
f_D(t) &= \frac{\frac{\lambda\mu\left(\mathrm{e}^{-\lambda t}-\mathrm{e}^{-\mu t}\right)}{\mu-\lambda} + (\xi-1)\mu\mathrm{e}^{-\mu t}}{\xi} = \\
&= \frac{\frac{\lambda\mu\left(\mathrm{e}^{-\lambda t}-\mathrm{e}^{-\mu t}\right)}{\mu-\lambda} + \left(\frac{1-(\lambda/\mu)^S}{1-\lambda/\mu} - 1\right)\mu\mathrm{e}^{-\mu t}}{\frac{1-(\lambda/\mu)^S}{1-\lambda/\mu}} = \\
&= \cdots = \frac{\mathrm{e}^{-\mu t}\left(\frac{\lambda}{\mu}\right)^S \mu - \lambda\mathrm{e}^{-\lambda t}}{\left(\frac{\lambda}{\mu}\right)^S - 1}
\end{aligned} \tag{6.14}$$

$$F_D(t) = \cdots = \frac{1-\mathrm{e}^{-\lambda t} - \left(\frac{\lambda}{\mu}\right)^S + \mathrm{e}^{-\mu t}\left(\frac{\lambda}{\mu}\right)^S}{1-\left(\frac{\lambda}{\mu}\right)^S} \tag{6.15}$$

$$\mathrm{E}(D) = \int_{t=0}^{\infty} t\, f_D(t)\,\mathrm{d}t = \cdots = \frac{\lambda^{S+1}-\mu^{S+1}}{\lambda\mu(\lambda^S-\mu^S)} \tag{6.16}$$

$$\begin{aligned}
\mathrm{Var}(D) &= \int_{t=0}^{\infty} (t-\mathrm{E}(D))^2\, f_D(t)\,\mathrm{d}t = \cdots = \\
&= \frac{\mu^{2S+2} - 2\lambda^{S+2}\mu^S - 2\mu^{S+2}\lambda^S + 2\lambda^{S+1}\mu^{S+1} + \lambda^{2S+2}}{\lambda^2\mu^2\left(\mu^{2S} - 2\mu^S\lambda^S + \lambda^{2S}\right)}
\end{aligned} \tag{6.17}$$

for $\lambda = \mu$:

$$f_D(t) = \frac{-\mu + \mu^2 t + \mu\, S}{\mathrm{e}^{\mu\, t} S} \tag{6.18}$$

$$F_D(t) = \frac{S\mathrm{e}^{\mu t} - S - \mu\, t}{S\mathrm{e}^{\mu t}} \tag{6.19}$$

$$\mathrm{E}(D) = \frac{S+1}{\mu S} \tag{6.20}$$

$$\mathrm{Var}(D) = \frac{S^2+2S-1}{S^2\mu^2} \tag{6.21}$$

6.1.2. Hypo/M/1/S queueing system

Figure 6.3 shows the Markov chain for the system state of a Hypo/M/1/S queueing system. We see that the idle period is hypoexponentially distributed (with parameters λ_1 and λ_2 if the Markov chain is in state $\langle 0/1\rangle$ when the idle period begins, and it is exponentially distributed (with parameter λ_2) if the Markov chain is in state $\langle 0/2\rangle$ when the idle period begins. The service times are exponentially distributed with rate μ. Therefore, the first interdeparture time is

$$D^{(1)} \sim \sigma^I_{\langle 0/1\rangle}\,\mathrm{HypoExp}(\lambda_1,\lambda_2,\mu) + \sigma^I_{\langle 0/2\rangle}\,\mathrm{HypoExp}(\lambda_2,\mu) \tag{6.22}$$

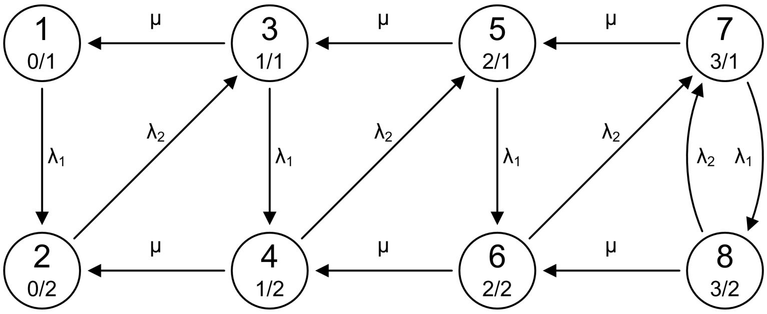


Figure 6.3.: Hypo/M/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the arrival process.

The following interdeparture times are

$$D^{(2)} \sim \mathrm{Exp}(\mu) \tag{6.23}$$

For the average interdeparture times, we obtain

$$D \sim \frac{\sigma^I_{\langle 0/1\rangle} \,\mathrm{HypoExp}(\lambda_1, \lambda_2, \mu) + \sigma^I_{\langle 0/2\rangle} \,\mathrm{HypoExp}(\lambda_2, \mu)}{\xi} + \frac{(\xi - 1)\,\mathrm{Exp}(\mu)}{\xi} \tag{6.24}$$

6.1.3. Hyper/M/1/S queueing system

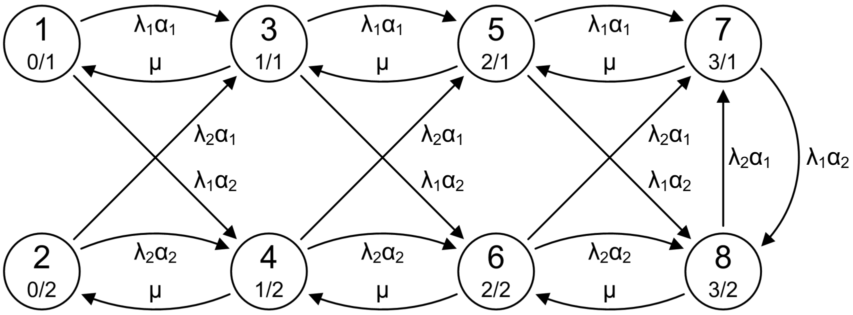


Figure 6.4.: Hyper/M/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the arrival process.

Figure 6.4 shows the Markov chain for the system state of a Hyper/M/1/S queueing system. If the Markov chain is in state $\langle 0/1\rangle$ when the idle period begins, the idle period is exponentially distributed with rate λ_1 (because $\alpha_1 + \alpha_2 = 1$). If the Markov chain is in state $\langle 0/2\rangle$ when the idle period begins, the idle period is exponentially distributed with rate λ_2. Therefore, the first interdeparture time is

$$D^{(1)} \sim \sigma^I_{\langle 0/1\rangle} \,\mathrm{HypoExp}(\lambda_1, \mu) + \sigma^I_{\langle 0/2\rangle} \,\mathrm{HypoExp}(\lambda_2, \mu) \tag{6.25}$$

The following interdeparture times are

$$D^{(2)} \sim \mathrm{Exp}(\mu) \tag{6.26}$$

For the average interdeparture times we obtain

$$D \sim \frac{\sigma^I_{\langle 0/1\rangle}\,\mathrm{HypoExp}(\lambda_1,\mu) + \sigma^I_{\langle 0/2\rangle}\,\mathrm{HypoExp}(\lambda_2,\mu) + (\xi - 1)\,\mathrm{Exp}(\mu)}{\xi} \tag{6.27}$$

6.1.4. Other PH/PH/1/S queueing systems

M/Hypo/1/S queueing system

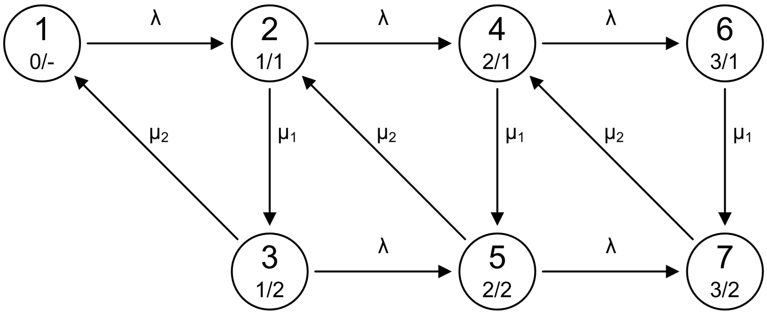


Figure 6.5.: M/Hypo/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the service process.

As can be seen from Figure 6.5, the length of the idle period of an M/Hypo/1/S queueing system is exponentially distributed with rate λ. The service times are hypoexponentially distributed with parameters μ_1 and μ_2:

$$I \sim \mathrm{Exp}(\lambda) \tag{6.28}$$
$$S \sim \mathrm{HypoExp}(\mu_1,\mu_2) \tag{6.29}$$

Therefore, we have

$$D \sim \frac{\mathrm{HypoExp}(\lambda,\mu_1,\mu_2) + (\xi - 1)\,\mathrm{HypoExp}(\mu_1,\mu_2)}{\xi} \tag{6.30}$$

M/Hyper/1/S queueing system

Figure 6.6 shows the Markov chain for the system state of an M/Hyper/1/S queueing system. The length of the idle period is exponentially distributed with rate λ. The service times are hyperexponentially distributed with parameters $\mu_1, \alpha_1, \mu_2, \alpha_2$. That is, the first service time is with probability α_1 exponentially distributed with rate μ_1, and with probability α_2 it is exponentially distributed with rate μ_2. Therefore, the first interdeparture time is

$$D^{(1)} \sim \alpha_1\,\mathrm{HypoExp}(\lambda,\mu_1) + \alpha_2\,\mathrm{HypoExp}(\lambda,\mu_2) \tag{6.31}$$

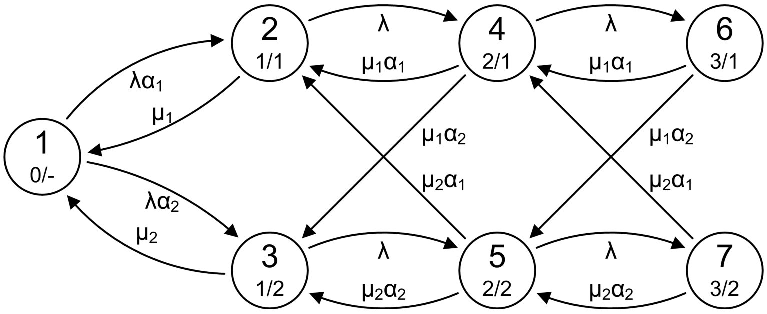


Figure 6.6.: M/Hyper/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the service process.

and the following interdeparture times are

$$D^{(2)} \sim \text{HyperExp}(\mu_1, \alpha_1, \mu_2, \alpha_2) \tag{6.32}$$

For the average interdeparture times, we obtain

$$\begin{aligned} D \sim \frac{\alpha_1 \,\text{HypoExp}(\lambda, \mu_1) + \alpha_2 \,\text{HypoExp}(\lambda, \mu_2)}{\xi} + \\ \frac{(\xi - 1)\,\text{HyperExp}(\mu_1, \alpha_1, \mu_2, \alpha_2)}{\xi} \end{aligned} \tag{6.33}$$

If we are only interested in the mean of the interdeparture times, we can calculate the departure rate with

$$\frac{1}{\text{E}(D)} = \sum_{n=1}^{3} \pi_{\langle n/1 \rangle} \mu_1 + \sum_{n=1}^{3} \pi_{\langle n/2 \rangle} \mu_2 \tag{6.34}$$

Hyper/Hyper/1/S queueing system

Figure 6.7 shows the Markov chain for the system state of a Hyper/Hyper/1/S queueing system.

If the Markov chain is in state $\langle 0/1/- \rangle$ when the idle period begins, the length of the idle period is exponentially distributed with rate λ_1. If the Markov chain is in state $\langle 0/2/- \rangle$ when the idle period begins, the length of the idle period is exponentially distributed with rate λ_2. The service times are hyperexponentially distributed with parameters $\mu_1, \beta_1, \mu_2, \beta_2$. That is, the first service time is with probability β_1 exponentially distributed with rate μ_1, and with probability β_2 it is exponentially distributed with rate μ_2. Therefore, the first interdeparture time is

$$\begin{aligned} D^{(1)} \sim\ & \sigma^I_{\langle 0/1/- \rangle} \beta_1 \,\text{HypoExp}(\lambda_1, \mu_1) + \\ & \sigma^I_{\langle 0/1/- \rangle} \beta_2 \,\text{HypoExp}(\lambda_1, \mu_2) + \\ & \sigma^I_{\langle 0/2/- \rangle} \beta_1 \,\text{HypoExp}(\lambda_2, \mu_1) + \\ & \sigma^I_{\langle 0/2/- \rangle} \beta_2 \,\text{HypoExp}(\lambda_2, \mu_2) \end{aligned} \tag{6.35}$$

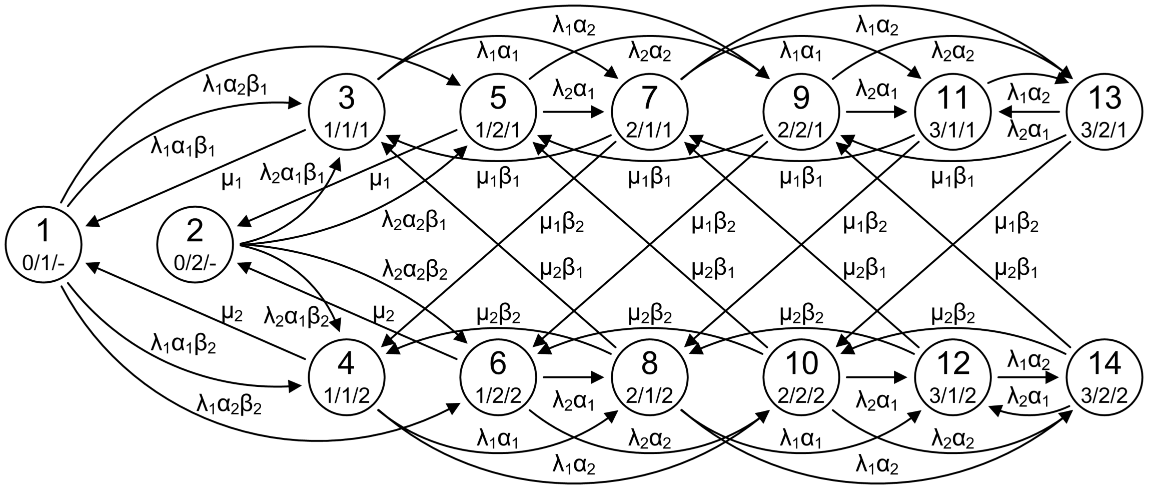


Figure 6.7.: Hyper/Hyper/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

and the following interdeparture times are

$$D^{(2)} \sim \text{HyperExp}(\mu_1, \beta_1, \mu_2, \beta_2) \tag{6.36}$$

The average interdeparture times are

$$\begin{aligned} D \sim (&\sigma^I_{\langle 0/1/-\rangle}\beta_1 \,\text{HypoExp}(\lambda_1, \mu_1) + \sigma^I_{\langle 0/1/-\rangle}\beta_2 \,\text{HypoExp}(\lambda_1, \mu_2) + \\ &\sigma^I_{\langle 0/2/-\rangle}\beta_1 \,\text{HypoExp}(\lambda_2, \mu_1) + \sigma^I_{\langle 0/2/-\rangle}\beta_2 \,\text{HypoExp}(\lambda_2, \mu_2) + \\ &(\xi - 1) \,\text{HyperExp}(\mu_1, \beta_1, \mu_2, \beta_2)) \,/\xi \end{aligned} \tag{6.37}$$

Hyper/Hypo/1/S queueing system

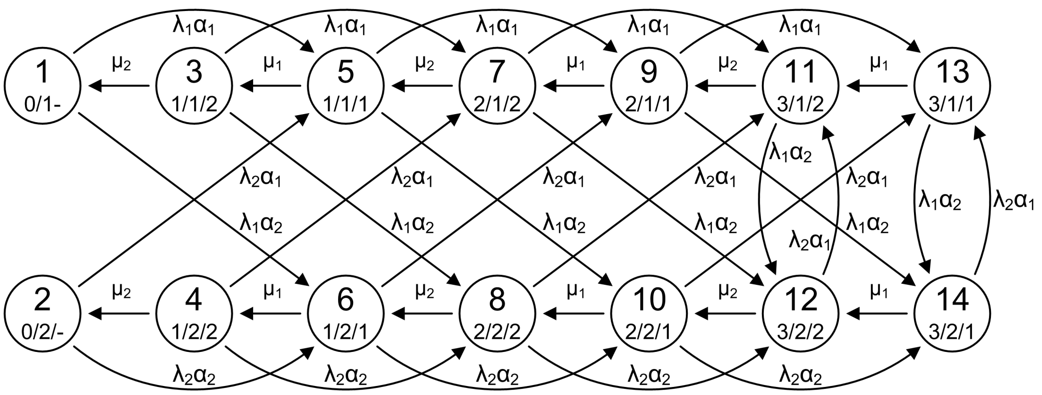


Figure 6.8.: Hyper/Hypo/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

Figure 6.8 shows the Markov chain for the system state of a Hyper/Hypo/1/S queueing system.

If the Markov chain is in state $\langle 0/1/-\rangle$ when the idle period begins, the length of the idle period is exponentially distributed with rate λ_1. If the Markov chain is in state $\langle 0/2/-\rangle$ when the idle period begins, the length of the idle period is exponentially distributed with rate λ_2. The service times are hypoexponentially distributed with parameters μ_1 and μ_2. Therefore, the first interdeparture time is

$$D^{(1)} \sim \ \sigma^I_{\langle 0/1/-\rangle} \operatorname{HypoExp}(\lambda_1, \mu_1, \mu_2) + \sigma^I_{\langle 0/2/-\rangle} \operatorname{HypoExp}(\lambda_2, \mu_1, \mu_2) \tag{6.38}$$

and the following interdeparture times are

$$D^{(2)} \sim \operatorname{HypoExp}(\mu_1, \mu_2) \tag{6.39}$$

The average interdeparture times are

$$\begin{aligned} D \sim \ & \frac{\sigma^I_{\langle 0/1/-\rangle} \operatorname{HypoExp}(\lambda_1, \mu_1, \mu_2) + \sigma^I_{\langle 0/2/-\rangle} \operatorname{HypoExp}(\lambda_2, \mu_1, \mu_2)}{\xi} + \\ & \frac{(\xi - 1) \operatorname{HypoExp}(\mu_1, \mu_2)}{\xi} \end{aligned} \tag{6.40}$$

If we are only interested in the mean of the interdeparture times, we can calculate the departure rate with

$$\frac{1}{\mathrm{E}(D)} = \sum_{n=1}^{3} \left(\pi_{\langle n/1/2\rangle} + \pi_{\langle n/2/2\rangle}\right) \mu_2 \tag{6.41}$$

Hypo/Hyper/1/S queueing system

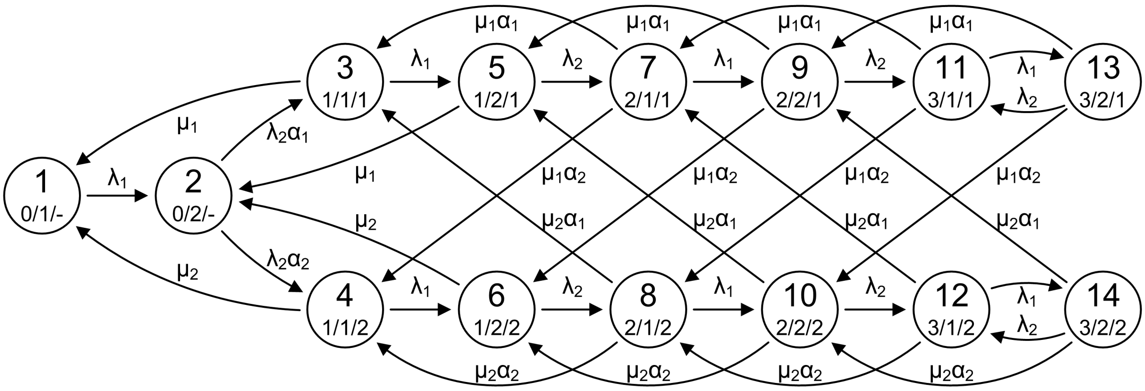


Figure 6.9.: Hypo/Hyper/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

Figure 6.9 shows the Markov chain for the system state of a Hypo/Hyper/1/S queueing system. If the Markov chain is in state $\langle 0/1/-\rangle$ when the idle period begins, the length of the idle period is hypoexponentially distributed (with parameters λ_1 and λ_2). If the Markov chain is in state $\langle 0/2/-\rangle$ when the idle period begins, it is exponentially distributed (with parameter λ_2). The service times are hyperexponentially distributed with parameters $\mu_1, \alpha_1, \mu_2, \alpha_2$. That is, the first service time is with probability α_1 exponentially distributed with rate μ_1, and with probability α_2 it is exponentially distributed with rate μ_2. Therefore, the first interdeparture time is

$$\begin{aligned} D^{(1)} \sim\ & \sigma^I_{\langle 0/1/-\rangle} \alpha_1 \operatorname{HypoExp}(\lambda_1, \lambda_2, \mu_1) + \\ & \sigma^I_{\langle 0/1/-\rangle} \alpha_2 \operatorname{HypoExp}(\lambda_1, \lambda_2, \mu_2) + \\ & \sigma^I_{\langle 0/2/-\rangle} \alpha_1 \operatorname{HypoExp}(\lambda_2, \mu_1) + \\ & \sigma^I_{\langle 0/2/-\rangle} \alpha_2 \operatorname{HypoExp}(\lambda_2, \mu_2) \end{aligned} \tag{6.42}$$

and the following interdeparture times are

$$D^{(2)} \sim \operatorname{HyperExp}(\mu_1, \alpha_1, \mu_2, \alpha_2) \tag{6.43}$$

The average interdeparture times are

$$\begin{aligned} D \sim\ & \left(\sigma^I_{\langle 0/1/-\rangle} \alpha_1 \operatorname{HypoExp}(\lambda_1, \lambda_2, \mu_1) + \sigma^I_{\langle 0/1/-\rangle} \alpha_2 \operatorname{HypoExp}(\lambda_1, \lambda_2, \mu_2) + \right. \\ & \sigma^I_{\langle 0/2/-\rangle} \alpha_1 \operatorname{HypoExp}(\lambda_2, \mu_1) + \sigma^I_{\langle 0/2/-\rangle} \alpha_2 \operatorname{HypoExp}(\lambda_2, \mu_2) + \\ & \left. (\xi - 1) \operatorname{HyperExp}(\mu_1, \alpha_1, \mu_2, \alpha_2) \right) / \xi \end{aligned} \tag{6.44}$$

Hypo/Hypo/1/S queueing system

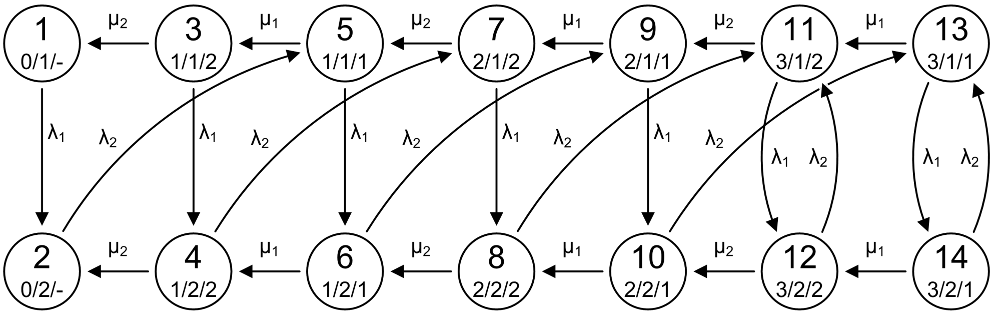


Figure 6.10.: Hypo/Hypo/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

Figure 6.9 shows the Markov chain for the system state of a Hypo/Hypo/1/S queueing system. If the Markov chain is in state $\langle 0/1/-\rangle$ when the idle period begins, the length of the idle period is hypoexponentially distributed (with parameters λ_1 and λ_2). If the Markov chain is in state $\langle 0/2/-\rangle$ when the idle period begins, it is exponentially distributed (with parameter λ_2). The service times are hypoexponentially distributed with parameters μ_1 and μ_2. Therefore, the first interdeparture time is

$$D^{(1)} \sim\ \sigma^I_{\langle 0/1/-\rangle} \operatorname{HypoExp}(\lambda_1, \lambda_2, \mu_1, \mu_2) + \sigma^I_{\langle 0/2/-\rangle} \operatorname{HypoExp}(\lambda_2, \mu_1, \mu_2) \tag{6.45}$$

and the following interdeparture times are

$$D^{(2)} \sim \operatorname{HypoExp}(\mu_1, \mu_2) \tag{6.46}$$

The average interdeparture times are

$$D \sim \frac{\sigma^I_{\langle 0/1/-\rangle}\,\mathrm{HypoExp}(\lambda_1,\lambda_2,\mu_1,\mu_2)+\sigma^I_{\langle 0/2/-\rangle}\,\mathrm{HypoExp}(\lambda_2,\mu_1,\mu_2)}{\xi}+ \frac{(\xi-1)\,\mathrm{HypoExp}(\mu_1,\mu_2)}{\xi} \tag{6.47}$$

Figures 6.11, 6.12 and 6.13 show the first two moments of the departure streams of the queueing systems discussed.

We see that when the arrival rate is low, the interdeparture times are mainly of the type $I+S$, so the statistical characteristics of the arrival process dominate. As the arrival rate increases, there are more and more interdeparture times of type S, therefore the statistical characteristics of the service process dominate.

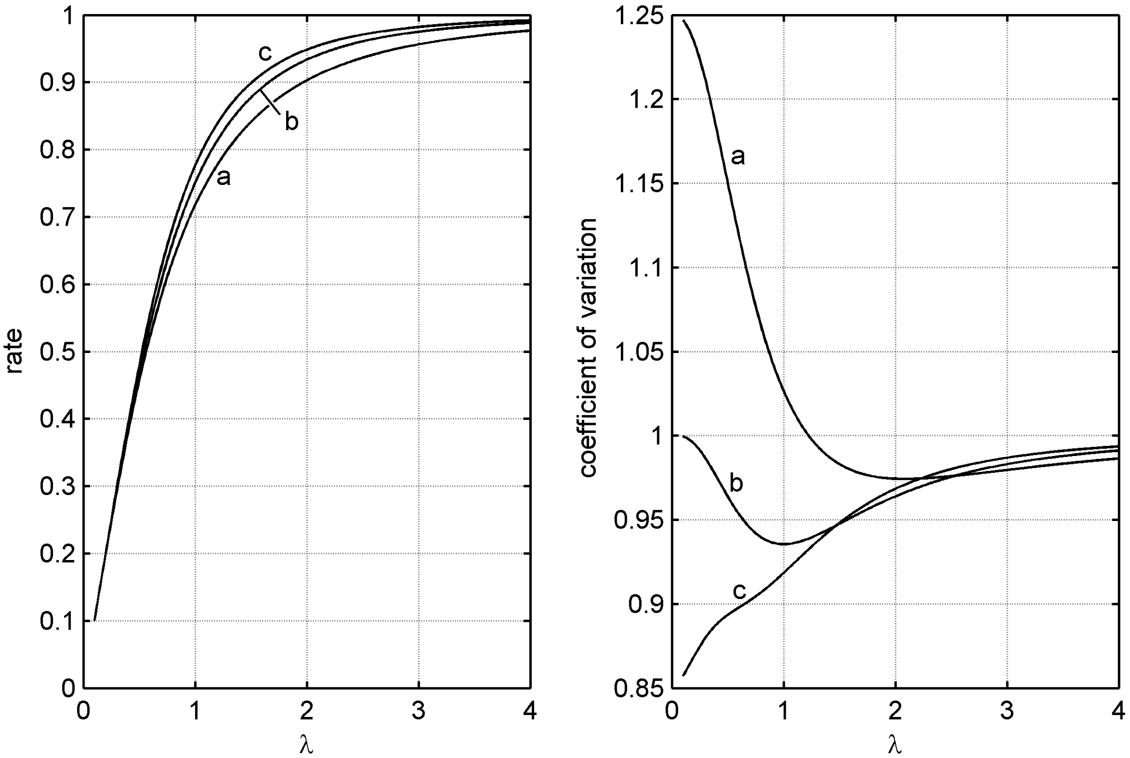


Figure 6.11.: Rate and coefficient of variation of the interdeparture times of PH/M/1/S queueing systems with arrival rate λ, service rate $\mu = 1$ and $S = 3$. (a) Hyper/M/1/S queueing system ($c_A = 1.25$), (b) M/M/1/S queueing system, (c) Hypo/M/1/S queueing system ($c_A = 0.85$).

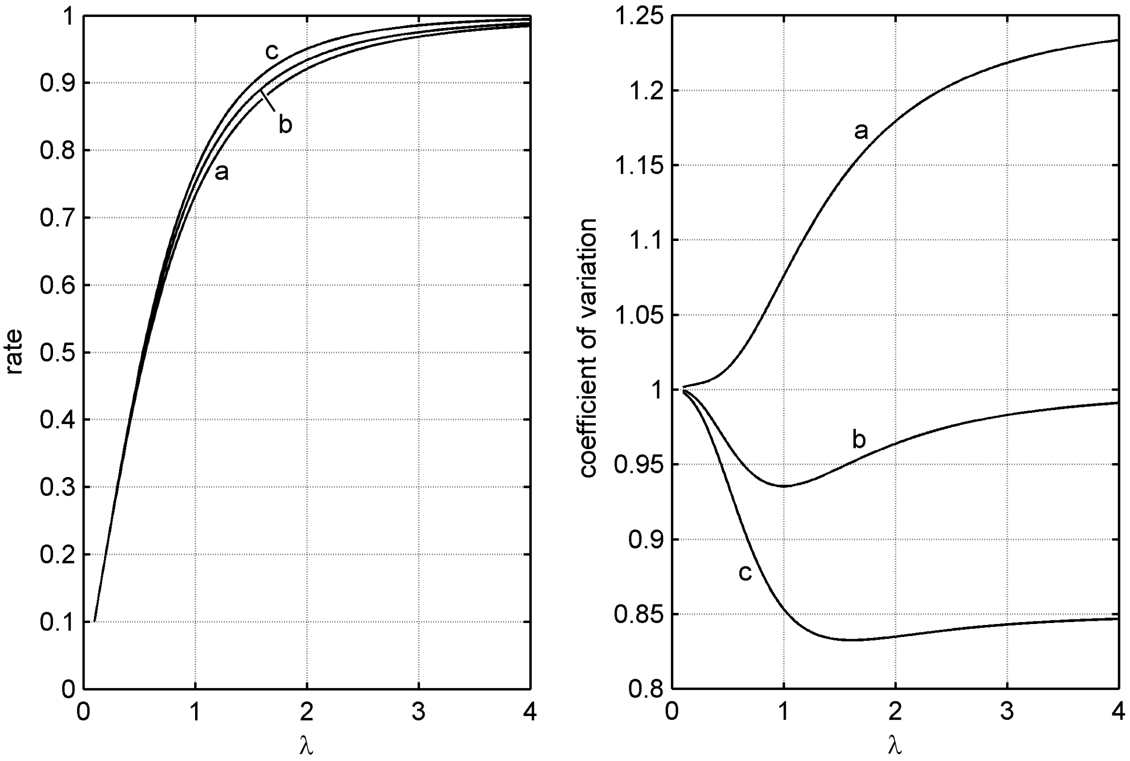


Figure 6.12.: Rate and coefficient of variation of the interdeparture times of M/PH/1/S queueing systems with arrival rate λ, service rate $\mu = 1$ and $S = 3$. (a) M/Hyper/1/S queueing system $(c_S = 1.25)$, (b) M/M/1/S queueing system, (c) M/Hypo/1/S queueing system $(c_S = 0.85)$.

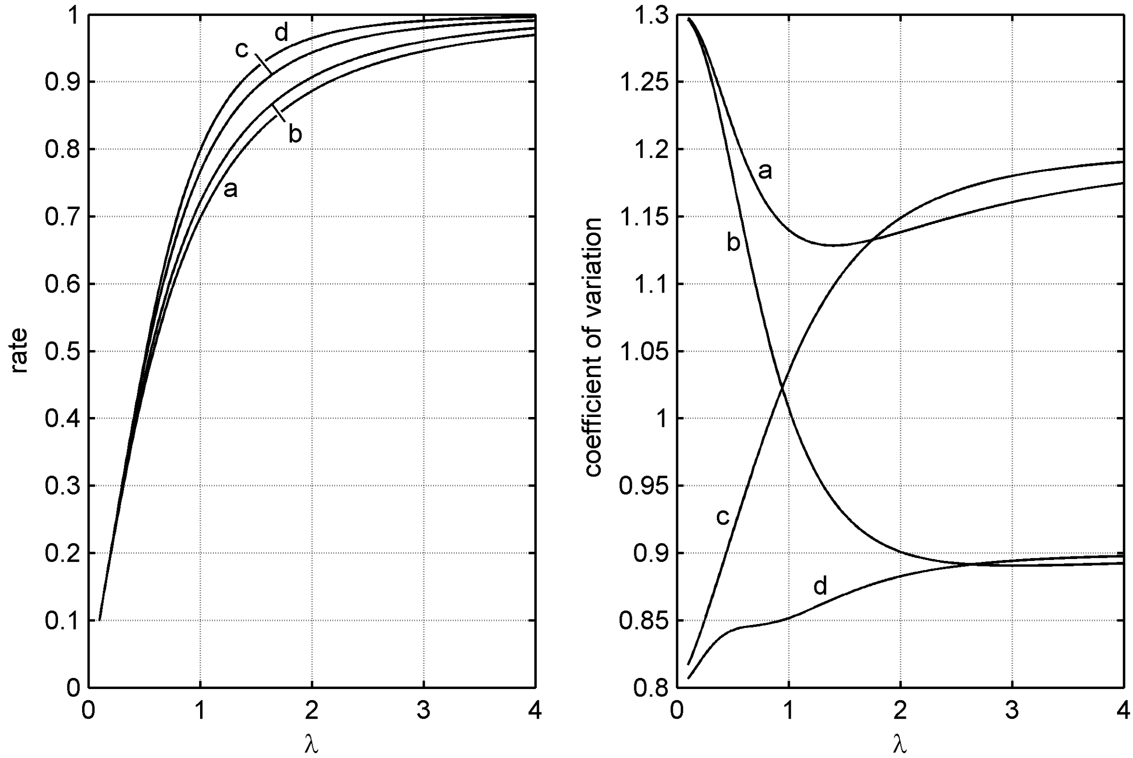


Figure 6.13.: Rate and coefficient of variation of the interdeparture times of PH/PH/1/S queueing systems with arrival rate λ, service rate $\mu = 1$ and $S = 3$. (a) Hyper/Hyper/1/S queueing system $(c_A = 1.3, c_S = 1.2)$, (b) Hyper/Hypo/1/S queueing system $(c_A = 1.3, c_S = 0.9)$, (c) Hypo/Hyper/1/S queueing system $(c_A = 0.8, c_S = 1.2)$, (d) Hypo/Hypo/1/S queueing system $(c_A = 0.8, c_S = 0.9)$.

6.2. Dependencies between the interdeparture times

Even if both the interarrival times and the service times of a queueing system are independent and identically distributed, this does, in general, not hold for the interdeparture times.[1]

The reasons are that the interdeparture times consist of two types of random variables (in a busy cycle we have one interevent time of length $I + S$ and $\xi - 1$ interevent times of length S) and that the service times within a busy period are not independent, as was discussed in Section 5.3.1: For example, the last service time in a busy period has rate $\lambda + \mu$, and the finite system size limits the number of consecutive short services within a busy period.

An important measure for the interdependence in a sequence of random variables is the autocorrelation, which is discussed in Section 8.2. In brief, the autocorrelation of a sequence of random variables $\langle X_1, X_2, \dots \rangle$ is a measure of the extent to which X_{i+d} depends on X_i. The closer the autocorrelation is to 1 or -1, the stronger the interdependence is. Figure 6.14 shows the autocorrelation of the interdeparture times of an M/M/1/S queueing system for various system sizes.

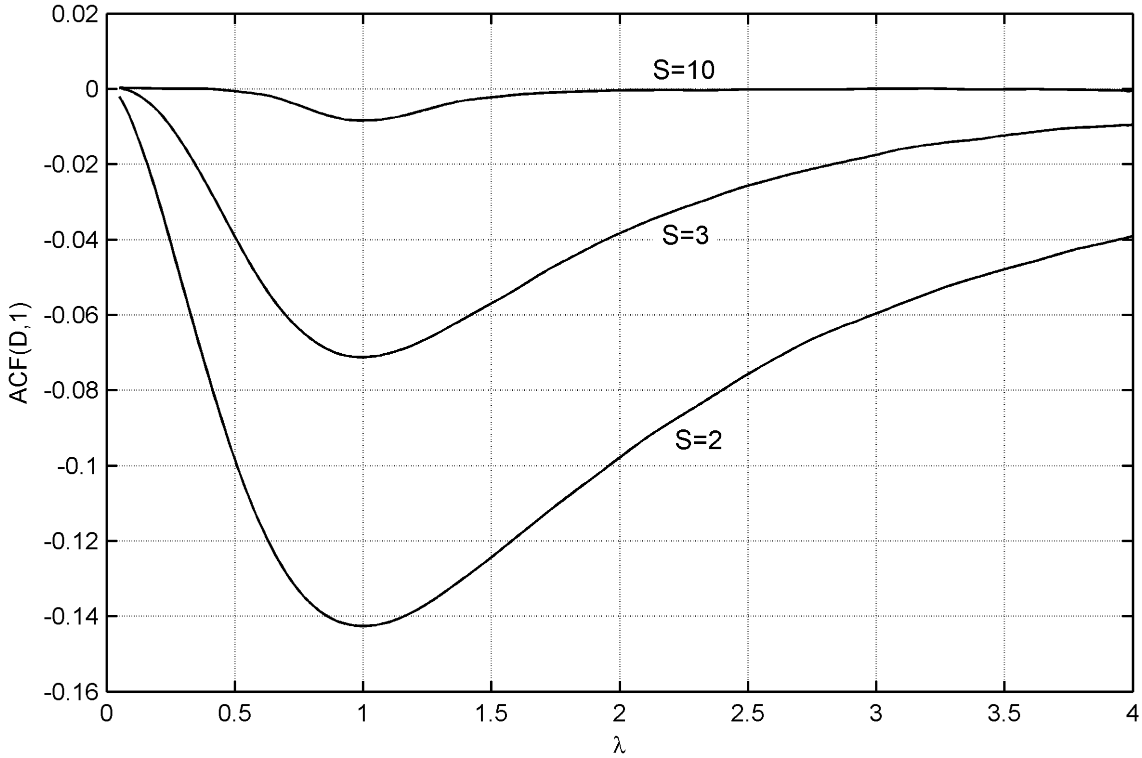


Figure 6.14.: M/M/1/S queueing system: Autocorrelation of the interdeparture times. Arrival rate λ, service rate $\mu = 1$. Simulation study.

[1] An important case where the interdeparture times are i.i.d. is the M/M/1 queueing system. The steady-state departure stream of an M/M/1 queueing system is a Poisson process whose rate equals the arrival rate (see [Burke 1966]).

6.2.1. Effect of the dependencies

For an estimation of the effect of the dependencies between the interdeparture times, we consider two queueing systems in tandem and compare the number of customers in the downstream system for two different arrival streams: The first stream is the departure stream of the upstream queueing system (Figure 6.15a). The second stream has the same interevent times as the first stream, but these interevent times have been shuffled (Figure 6.15b), so that they are not interdependent any more.

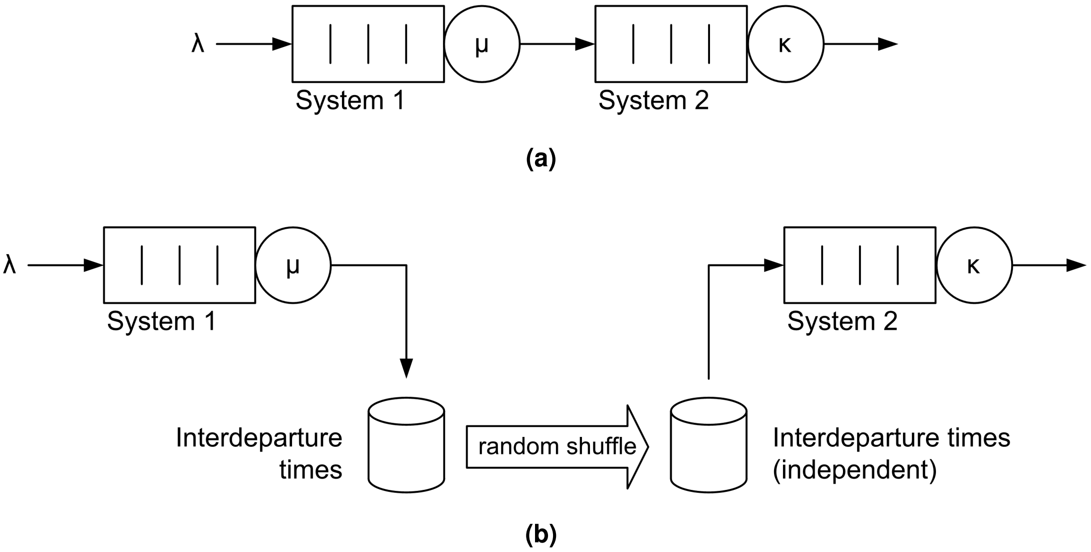


Figure 6.15.: Estimation of the effect of the dependencies between the interdeparture times. (a) The interarrival times at system 2 are the real interdeparture times of system 1 (including dependencies). (b) The interarrival times at system 2 are the interdeparture times of system 1 without dependencies.

The results are shown in Figure 6.16. When the arrival rate of the upstream system is very low or very high, there is little difference between the results. In the former case, the interdeparture times are mainly of type $I + S$ and are, therefore, independent. In the latter case, the interdeparture times are mainly of type S. They are not completely independent, but since we have long busy periods and a high probability that the system is full, the interdependence is very weak.

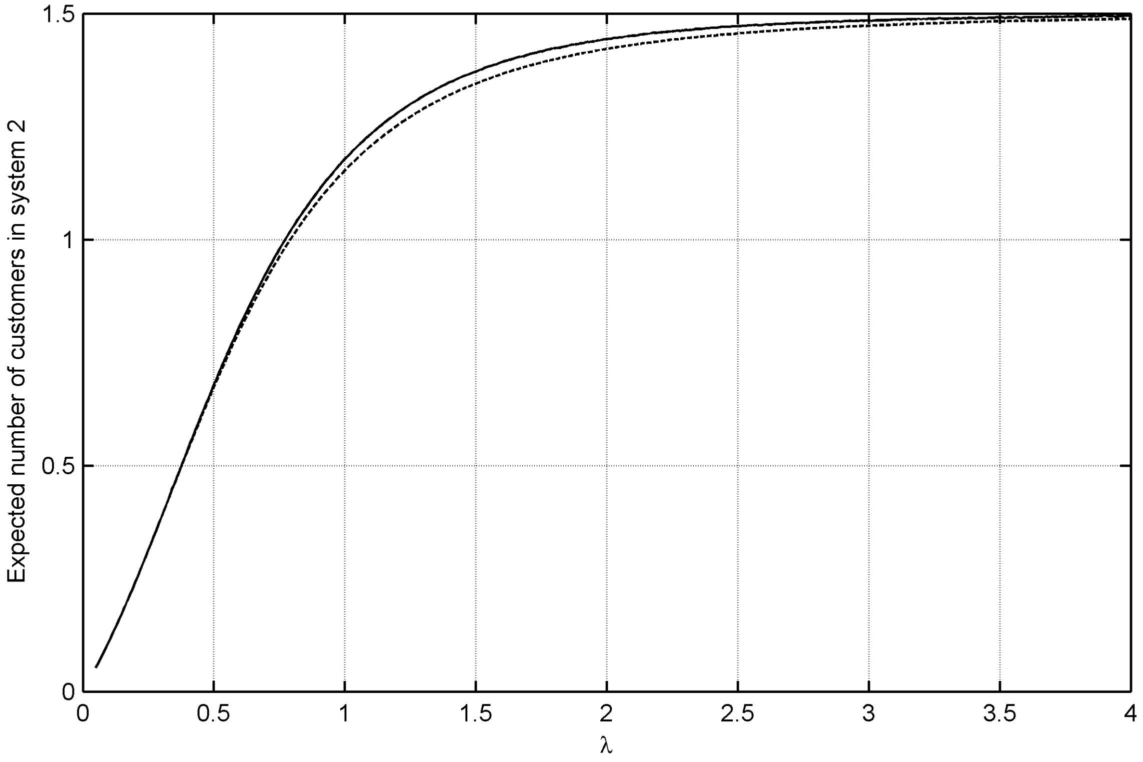


Figure 6.16.: Two queueing systems in tandem: number of customers in the downstream system. Solid line: real interarrival times, dashed line: interarrival times without dependencies.

6.3. Modelling a tandem system

We consider the queueing network shown in Figure 6.17. The network consists of two GI/M/1/S queueing systems. Customers arrive according to a Poisson process at system 1. When they have been served in system 1, they are forwarded to system 2. Such a network is called a *(two-station) tandem system*. If system 2 is full when a customer arrives, the customer is lost.

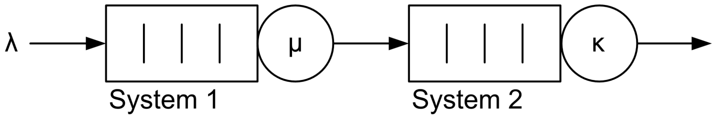


Figure 6.17.: Tandem system.

We are interested in the number of customers in system 2 and in the interdeparture times D_2 of system 2. (The number of customers in system 1 and the interdeparture times of system 1 are independent of the fact that served customers are forwarded to another system, so they are calculated as shown in Sections 3.1 and 6.1.)

The Markov chain for the system state of the network is shown in Figure 6.18.

Arrivals at system 1 (rate λ) increase the number of customers in system 1. Services in system 1 (rate μ) decrease the number of customers in system 1 and increase the number of customers in system 2. Services in system 2 (rate κ) decrease the number of customers in system 2.

For the calculation of the number of customers in the queueing systems, we calculate the stationary system state probabilities π by solving the system of linear equations

$$\pi \cdot \mathcal{Q} = 0 \tag{6.48}$$

$$\sum_i \pi_i = 1 \tag{6.49}$$

The number of customers in the first system X_1 is

$$\mathrm{P}\left\{X_1 = i\right\} = \sum_{k=0}^{3} \pi_{\langle i/k\rangle} \tag{6.50}$$

$$\mathrm{E}(X_1) = \sum_{i=1}^{3} i \sum_{k=0}^{3} \pi_{\langle i/k\rangle} \tag{6.51}$$

The number of customers in the second system X_2 is

$$\mathrm{P}\left\{X_2 = i\right\} = \sum_{k=0}^{3} \pi_{\langle k/i\rangle} \tag{6.52}$$

$$\mathrm{E}(X_2) = \sum_{i=1}^{3} i \sum_{k=0}^{3} \pi_{\langle k/i\rangle} \tag{6.53}$$

The length of the idle period of the second system I_2 is calculated with the Markov chain shown in Figure 6.19.

The Markov chain can be in states $\langle 0/0\rangle$, $\langle 1/0\rangle$, $\langle 2/0\rangle$ and $\langle 3/0\rangle$ when the idle period begins, depending on the state in which the network was before the last customer in system 2 was served. The probabilities σ_i^I that the Markov chain is in state $i, i = 1 \ldots 4$, when the idle period begins are

$$\sigma_{\langle 0/0\rangle}^I = \sigma_1^I = \pi_{\langle 0/1\rangle} \left(\sum_{k=0}^{3} \pi_{\langle k/1\rangle} \right)^{-1} = \pi_5 \left(\sum_{k=5}^{8} \pi_k \right)^{-1} \tag{6.54}$$

$$\sigma_{\langle 1/0\rangle}^I = \sigma_2^I = \pi_{\langle 1/1\rangle} \left(\sum_{k=0}^{3} \pi_{\langle k/1\rangle} \right)^{-1} = \pi_6 \left(\sum_{k=5}^{8} \pi_k \right)^{-1} \tag{6.55}$$

$$\sigma_{\langle 2/0\rangle}^I = \sigma_3^I = \pi_{\langle 2/1\rangle} \left(\sum_{k=0}^{3} \pi_{\langle k/1\rangle} \right)^{-1} = \pi_7 \left(\sum_{k=5}^{8} \pi_k \right)^{-1} \tag{6.56}$$

$$\sigma_{\langle 3/0\rangle}^I = \sigma_4^I = \pi_{\langle 3/1\rangle} \left(\sum_{k=0}^{3} \pi_{\langle k/1\rangle} \right)^{-1} = \pi_8 \left(\sum_{k=5}^{8} \pi_k \right)^{-1} \tag{6.57}$$

The idle period ends when a customer arrives at system 2, that is, if one of the states $\langle 0/1\rangle$, $\langle 1/1\rangle$ or $\langle 2/1\rangle$ is reached.

The length of the idle period is

$$\begin{aligned} I_2 \sim \sigma_{\langle 0/0\rangle}^I \Big(& \frac{\mu}{\lambda+\mu} \operatorname{HypoExp}(\lambda,\mu) + \frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu} \operatorname{HypoExp}(\lambda,\lambda,\mu) + \\ & \frac{\lambda}{\lambda+\mu}\frac{\lambda}{\lambda+\mu} \operatorname{HypoExp}(\lambda,\lambda,\lambda,\mu)\Big) + \\ \sigma_{\langle 1/0\rangle}^I \Big(& \frac{\mu}{\lambda+\mu} \operatorname{Exp}(\mu) + \frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu} \operatorname{HypoExp}(\lambda,\mu) + \\ & \frac{\lambda}{\lambda+\mu}\frac{\lambda}{\lambda+\mu} \operatorname{HypoExp}(\lambda,\lambda,\mu)\Big) + \\ \sigma_{\langle 2/0\rangle}^I \Big(& \frac{\mu}{\lambda+\mu} \operatorname{Exp}(\mu) + \frac{\lambda}{\lambda+\mu} \operatorname{HypoExp}(\lambda,\mu)\Big) + \\ \sigma_{\langle 3/0\rangle}^I & \operatorname{Exp}(\mu) \end{aligned} \tag{6.58}$$

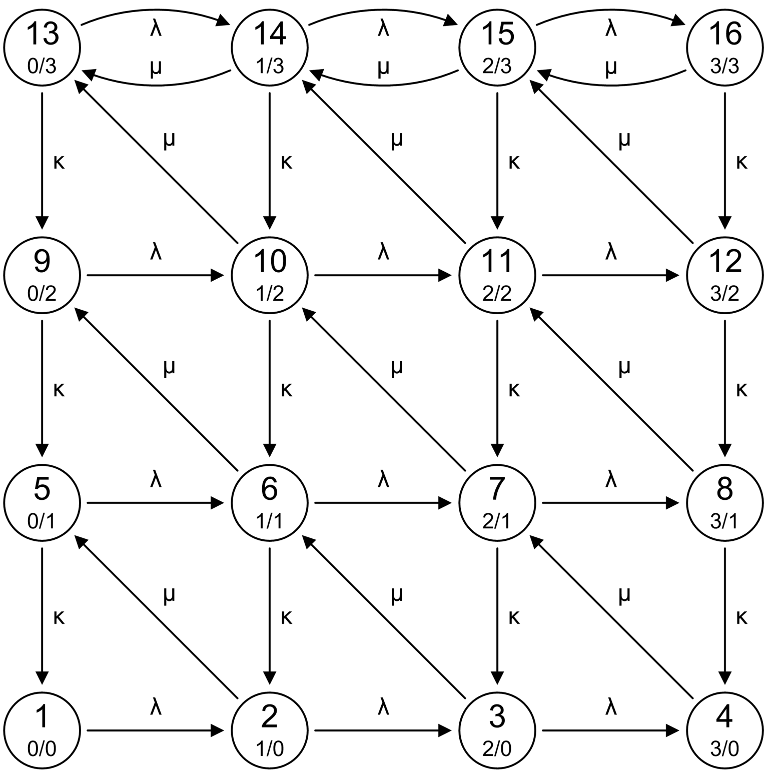


Figure 6.18.: Tandem system: Markov chain for the system state. Meaning of the names of the states: number of customers in system 1 / number of customers in system 2.

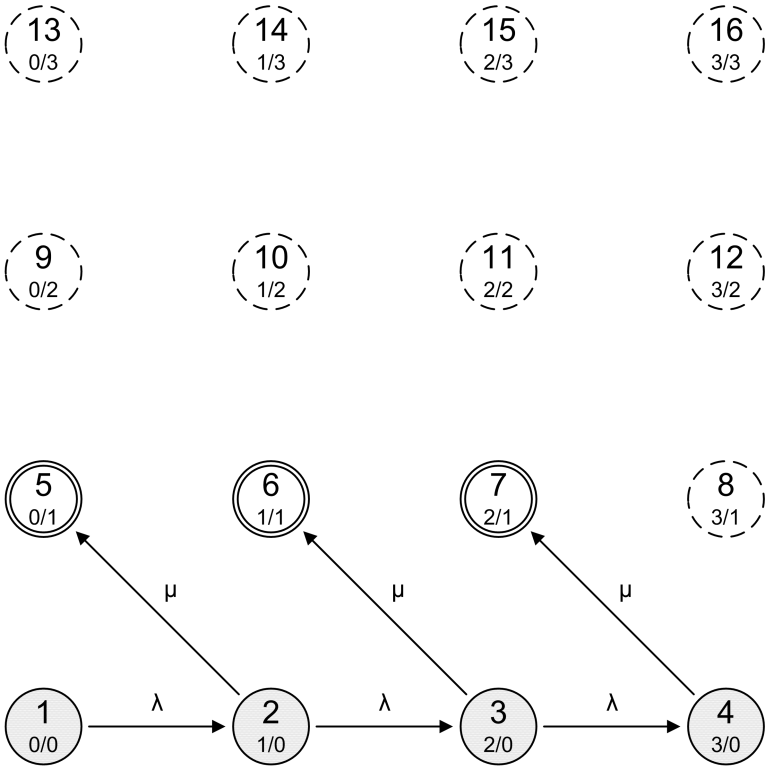


Figure 6.19.: Tandem system: Markov chain for the calculation of the length of the idle period. Meaning of the names of the states: number of customers in system 1 / number of customers in system 2.

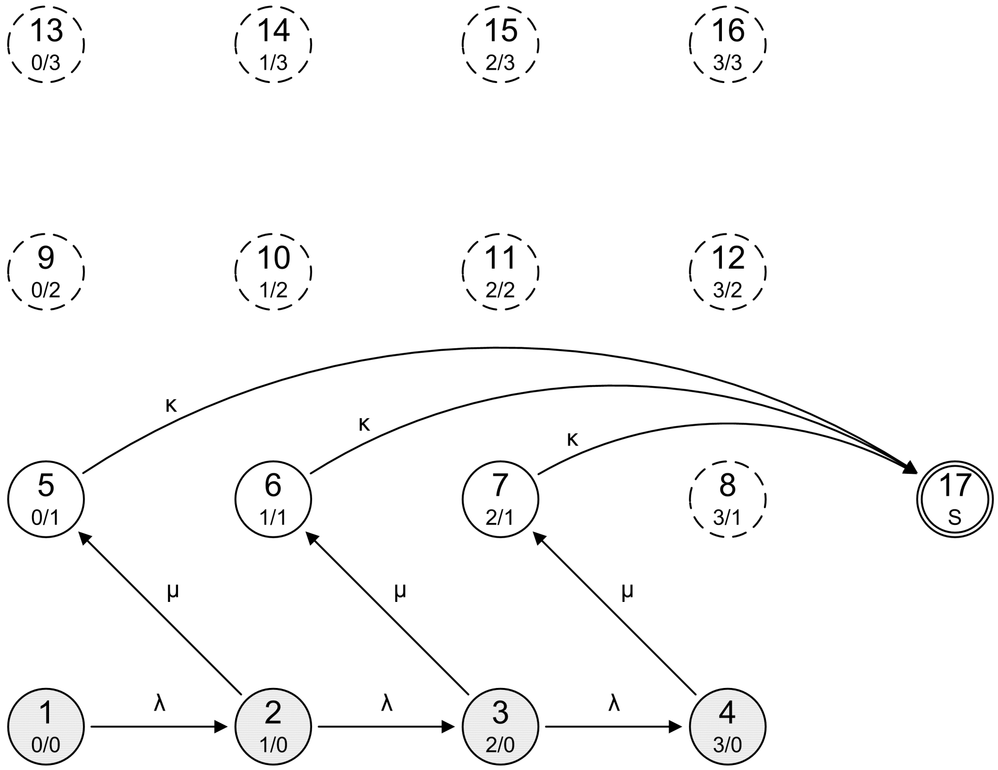


Figure 6.20.: Tandem system: Markov chain for the calculation of the length of the first interdeparture time. Meaning of the names of the states: number of customers in system 1 / number of customers in system 2.

The length of the first interdeparture time $D_2^{(1)}$ is the sum of length of the idle period and the first service,

$$\begin{aligned} D_2^{(1)} \sim \sigma^I_{\langle 0/0\rangle}\Big(&\frac{\mu}{\lambda+\mu}\operatorname{HypoExp}(\lambda,\mu,\kappa) + \frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}\operatorname{HypoExp}(\lambda,\lambda,\mu,\kappa) + \\ &\frac{\lambda}{\lambda+\mu}\frac{\lambda}{\lambda+\mu}\operatorname{HypoExp}(\lambda,\lambda,\lambda,\mu,\kappa)\Big) + \\ &\sigma^I_{\langle 1/0\rangle}\Big(\frac{\mu}{\lambda+\mu}\operatorname{HypoExp}(\mu,\kappa) + \frac{\lambda}{\lambda+\mu}\frac{\mu}{\lambda+\mu}\operatorname{HypoExp}(\lambda,\mu,\kappa) + \\ &\frac{\lambda}{\lambda+\mu}\frac{\lambda}{\lambda+\mu}\operatorname{HypoExp}(\lambda,\lambda,\mu,\kappa)\Big) + \\ &\sigma^I_{\langle 2/0\rangle}\Big(\frac{\mu}{\lambda+\mu}\operatorname{HypoExp}(\mu,\kappa) + \frac{\lambda}{\lambda+\mu}\operatorname{HypoExp}(\lambda,\mu,\kappa)\Big) + \\ &\sigma^I_{\langle 3/0\rangle}\operatorname{HypoExp}(\mu,\kappa) \end{aligned} \tag{6.59}$$

Another way of determining $D_2^{(1)}$ is to calculate the time the Markov chain shown in Figure 6.20 needs to go from states 1...4 to state 17:

$$\varphi_k(0) = \begin{cases} 1 & 1 \le k \le 7 \\ 0 & \text{otherwise} \end{cases} \tag{6.60}$$

$$\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \tag{6.61}$$

$$\begin{aligned} \mathrm{P}\left\{D_2^{(1)} > t\right\} = \sigma^I_{\langle 0/0\rangle}\varphi_{\langle 0/0\rangle}(t) + \sigma^I_{\langle 1/0\rangle}\varphi_{\langle 1/0\rangle}(t) + \\ \sigma^I_{\langle 2/0\rangle}\varphi_{\langle 2/0\rangle}(t) + \sigma^I_{\langle 3/0\rangle}\varphi_{\langle 3/0\rangle}(t) \end{aligned} \tag{6.62}$$

The length of the busy period is calculated with the Markov chain shown in Figure 6.21.

The busy period begins when system 2 is empty (states $\langle 1/0\rangle, \langle 2/0\rangle, \langle 3/0\rangle$) and in system 1 a customer is served (states $\langle 0/1\rangle, \langle 1/1\rangle, \langle 2/1\rangle$). The probability σ_i^B that the Markov chain is in state i when the busy period begins depends on the stationary system state probabilities $\pi_{\langle 1/0\rangle}$, $\pi_{\langle 2/0\rangle}$ and $\pi_{\langle 3/0\rangle}$:

$$\sigma^B_{\langle 0/1\rangle} = \sigma^B_5 = \pi_{\langle 1/0\rangle}\left(\sum_{k=1}^{3}\pi_{\langle k/0\rangle}\right)^{-1} = \pi_2\left(\sum_{k=2}^{4}\pi_k\right)^{-1} \tag{6.63}$$

$$\sigma^B_{\langle 1/1\rangle} = \sigma^B_6 = \pi_{\langle 2/0\rangle}\left(\sum_{k=1}^{3}\pi_{\langle k/0\rangle}\right)^{-1} = \pi_3\left(\sum_{k=2}^{4}\pi_k\right)^{-1} \tag{6.64}$$

$$\sigma^B_{\langle 2/1\rangle} = \sigma^B_7 = \pi_{\langle 3/0\rangle}\left(\sum_{k=1}^{3}\pi_{\langle k/0\rangle}\right)^{-1} = \pi_4\left(\sum_{k=2}^{4}\pi_k\right)^{-1} \tag{6.65}$$

The busy period ends when system 2 becomes idle again (states $\langle k/0\rangle, k = 1\ldots 4$).

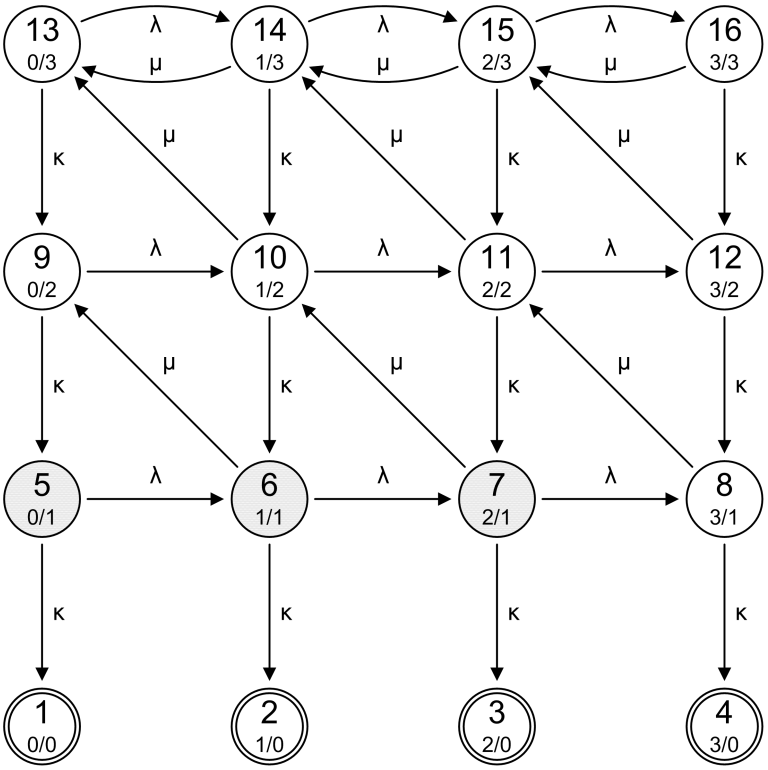


Figure 6.21.: Tandem system: Markov chain for the calculation of the length of the busy period. Meaning of the names of the states: number of customers in system 1 / number of customers in system 2.

Therefore, the length B_2 of the busy period of system 2 is calculated as follows:

$$\varphi_{\langle i/j\rangle}(0) = \begin{cases} 1 & j \geq 1 \\ 0 & j = 0 \end{cases} \tag{6.66}$$

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{6.67}$$

and

$$\mathrm{P}\left\{B_2 > t\right\} = F^C_{B_2}(t) = \sigma^B_{\langle 0/1\rangle}\varphi_{\langle 0/1\rangle}(t) + \sigma^B_{\langle 1/1\rangle}\varphi_{\langle 1/1\rangle}(t) + \sigma^B_{\langle 2/1\rangle}\varphi_{\langle 2/1\rangle}(t) \tag{6.68}$$

$$\mathrm{E}(B_2) = \int_0^\infty F^C_{B_2}(t)\mathrm{d}t \tag{6.69}$$

The number of customers served during a busy period ξ is

$$\xi = \kappa \cdot \mathrm{E}(B_2) \tag{6.70}$$

Finally, the interdeparture time D_2 of the system 2 is

$$D_2 \sim \frac{D_2^{(1)} + (\xi - 1)\,\mathrm{Exp}(\kappa)}{\xi} \tag{6.71}$$

6.4. Modelling a tandem system using network decomposition

Again, we consider the queueing network shown in Figure 6.17. But now we use network decomposition for our analysis: we break up the network into subsystems, and analyse these subsystems individually.

It should be noted that – due to the simplicity and the small size of the involved queueing systems – in our example the Markov chain that describes the state of the network is quite small. The Markov chains we need for doing network decomposition have a similar size or are even bigger, so that the advantage of network decomposition might not be seen easily. However, as the complexity and size of the involved queueing systems increases, the size of the Markov chain for the state of the network increases much faster than the size of the Markov chains needed for the network decomposition (see Figure 6.22).

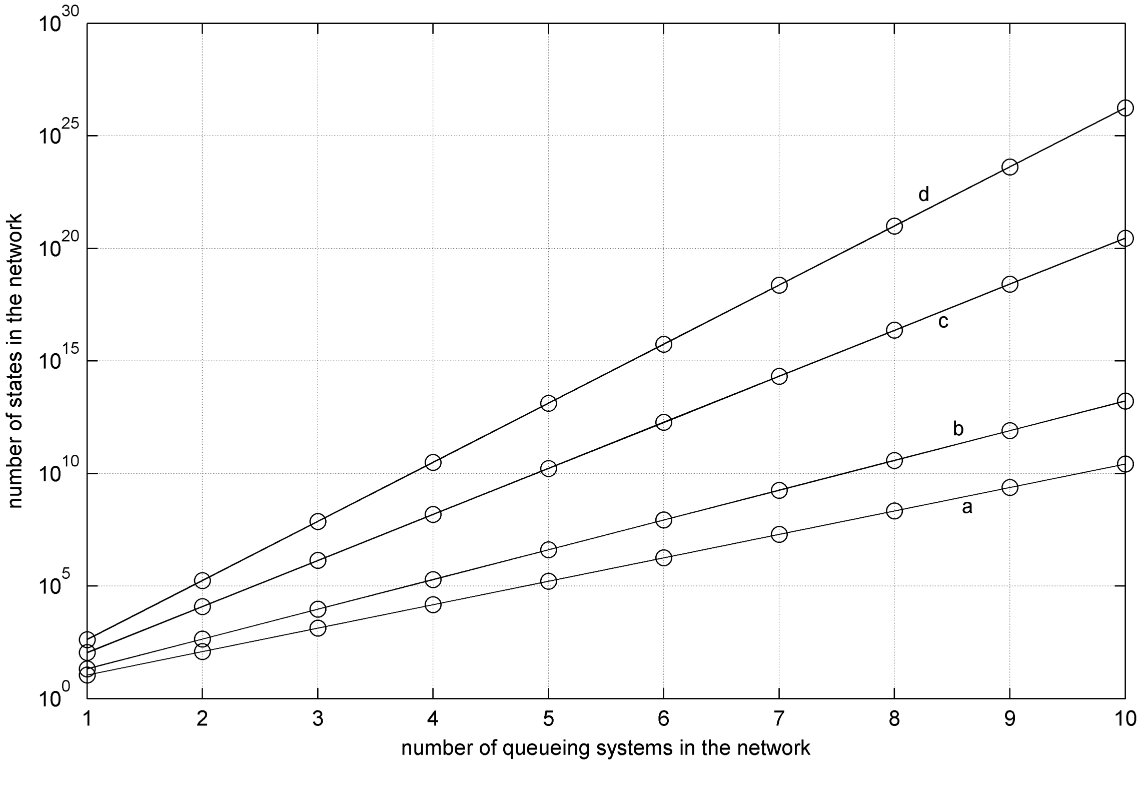
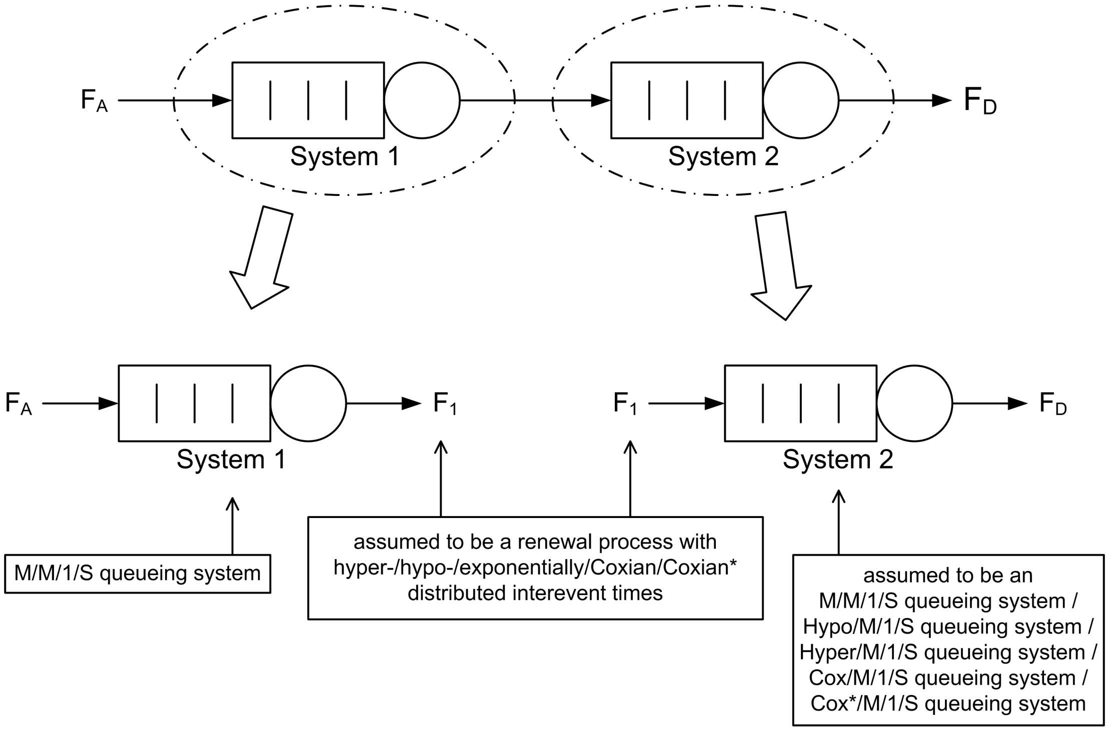


Figure 6.22.: The number of states in a queueing network grows exponentially with the number of queueing systems in the network. Here: M/M/1/S queueing systems with (a) $S = 10$, (b) $S = 20$, (c) $S = 10$, two classes of customers, (d) $S = 20$, two classes of customers.

Our first approach (Figure 6.23) is to assume that the stream between system 1 and system 2 has independent and identically distributed interevent times.



F_A
System 1
System 2
F_D
F_A
F_1
System 1
F_1
System 2
F_D
M/M/1/S queueing system
assumed to be a renewal process with hyper-/hypo-/exponentially/Coxian/Coxian* distributed interevent times
assumed to be an M/M/1/S queueing system / Hypo/M/1/S queueing system / Hyper/M/1/S queueing system / Cox/M/1/S queueing system / Cox*/M/1/S queueing system

Figure 6.23.: Two GI/M/1/S queueing systems in tandem: network decomposition.

The first step is to determine the probability distribution of the interdeparture times D_1 of the first queueing system. This system is an M/M/1/S queueing system, so we can use the techniques shown in Section 6.1.

Now we approximate this distribution by a phase-type distribution (cf. Section 3.4). For the approximation, we use a phase-type distribution whose first k moments match the first k moments of the distribution of D_1, where $k = 1, 2, 3$ and 5:

number of moments	*phase-type distribution*
1	exponential
2	hypoexponential ($c < 1$) exponential ($c = 1$) hyperexponential ($c > 1$)
3	2- or 3-stage EC distribution (cf. page 71)
5	3-stage Coxian

For example, if we match 5 moments, we determine the parameters of a 3-stage Coxian distribution such that its first 5 moments match the first 5 moments of the distribution of the interdeparture times of system 1.

Then we assume system 2 to be a PH/M/1/S queueing system, whereby the distribution of the interarrival times is the previously determined phase-type distribution:

number of moments	*type of system 2*
1	M/M/1/S
2	Hypo/M/1/S ($c < 1$) M/M/1/S ($c = 1$) Hyper/M/1/S ($c > 1$)
3	Cox*/M/1/S
5	Cox/M/1/S

The expected number of customers and the interdeparture times of these PH/M/1/S queueing systems can be determined using the techniques shown in Sections 3.1 and 6.1.

Figure 6.24 shows the results. It should be noted that, since we ignore the interdependence between the interdeparture times, the best achievable result is indicated by the dashed line.

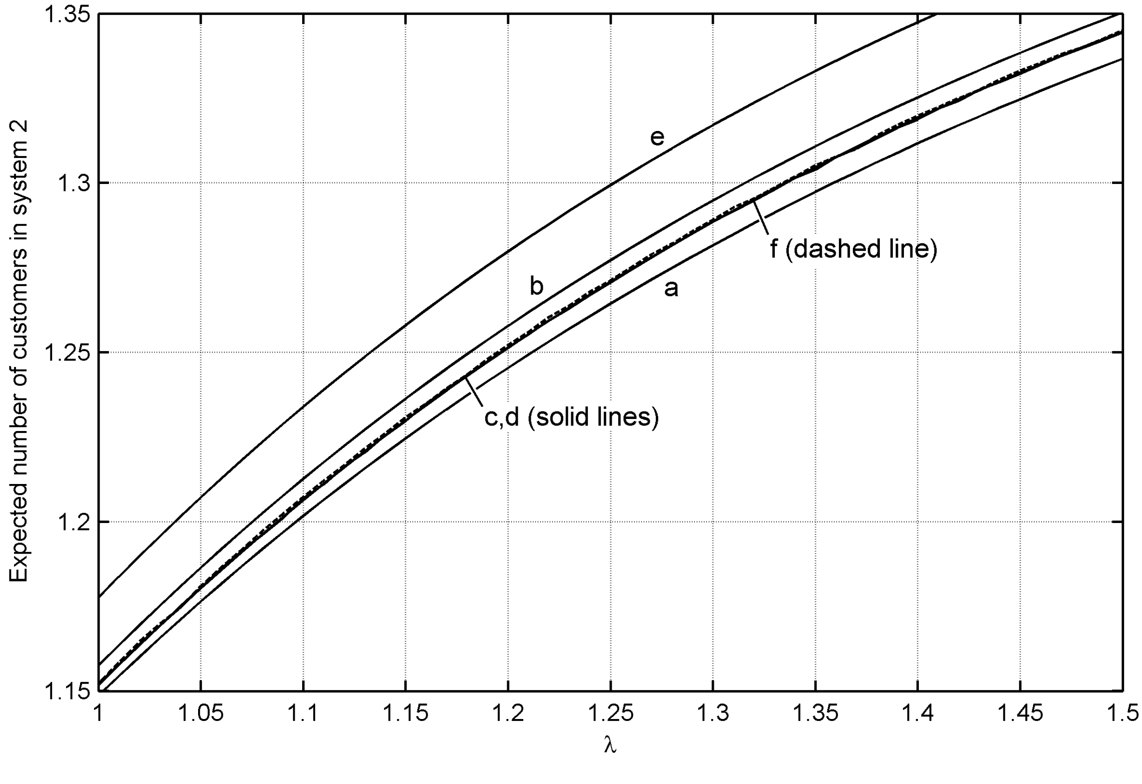


Figure 6.24.: Two queueing systems in tandem: number of customers in the downstream system. The arrival rate in the upstream system (system 1) is λ, the service rates are $\mu = \kappa = 1$. $S = 3$. Network decomposition: (a) 1 moment, (b) 2 moments, (c) 3 moments, (d) 5 moments. (e) Exact results, (f) the interarrival times of the second system are the interdeparture times of the first system without interdependencies.

For our second approach (Figure 6.25), we use our knowledge of the structure of the departure stream: In a busy cycle the first interdeparture time is of type $I + S$, the following $\xi - 1$ interdeparture times are of type S.

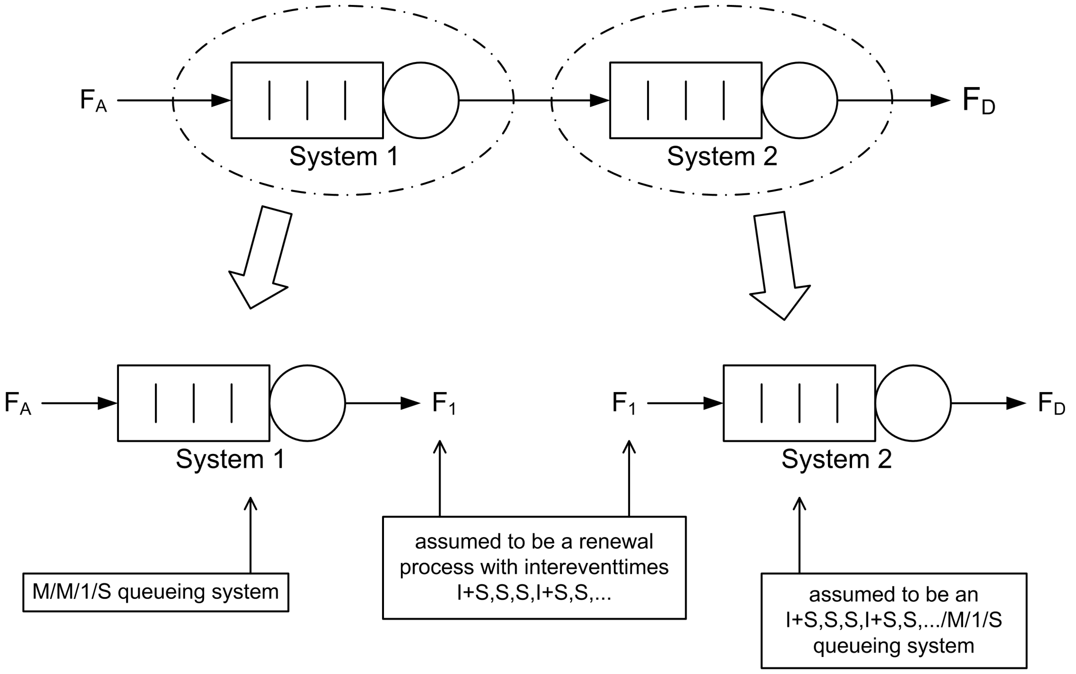

F_A
System 1
System 2
F_D
F_A
System 1
F_1
F_1
System 2
F_D
M/M/1/S queueing system
assumed to be a renewal process with intereventtimes I+S,S,S,I+S,S,...
assumed to be an I+S,S,S,I+S,S,.../M/1/S queueing system

Figure 6.25.: Two GI/M/1/S queueing systems in tandem: network decomposition.

The Markov chain for the system state of system 2 is shown in Figure 6.27. (Figures 6.28 and 6.29 show two other Markov chains that could be used.)

If system 1 is empty (states 1,2,3 and 4) and a customer arrives (rate λ), we chose the length of the following busy period in system 1 according to the distribution of ξ. These transitions do not affect the number of customers in system 2. They are part of the interdeparture times of type $I + S$. With each service in system 1, the number of customers in system 2 increases and the number of remaining services during the busy period of system 1 decreases. Depending on the chosen length of the busy period, we have a certain number of arrivals at system 2. Then system 1 is assumed to be idle again, and the next transition is the arrival of a customer at system 1.

One might expect this approach to achieve good results. However, as discussed in Section 5.3.1, it is not possible to model the state of a queueing system by describing the remaining number of customers to be served in the current busy period. Therefore, the approach fails. The results, which are shown in Figure 6.26, are not better than those we obtain when we approximate the arrival process of system 2 by a renewal process whose interevent times match only the first moment of the interdeparture times of system 1 (Figure 6.24 - line a).

Of course, we could use an even more complicated Markov chain in order to take the interdependencies between the interdeparture times of system 1 into account. But with each additional detail we describe in the Markov chain, the Markov chain becomes bigger. If we construct a Markov chain that includes a complete description of the structure of the departure stream, this Markov chain is as big the Markov chain that describes the system state for both systems (Figure 6.18). This means, when we use network decomposition, we have to find a trade-off between the accuracy and the size of the used Markov chains.

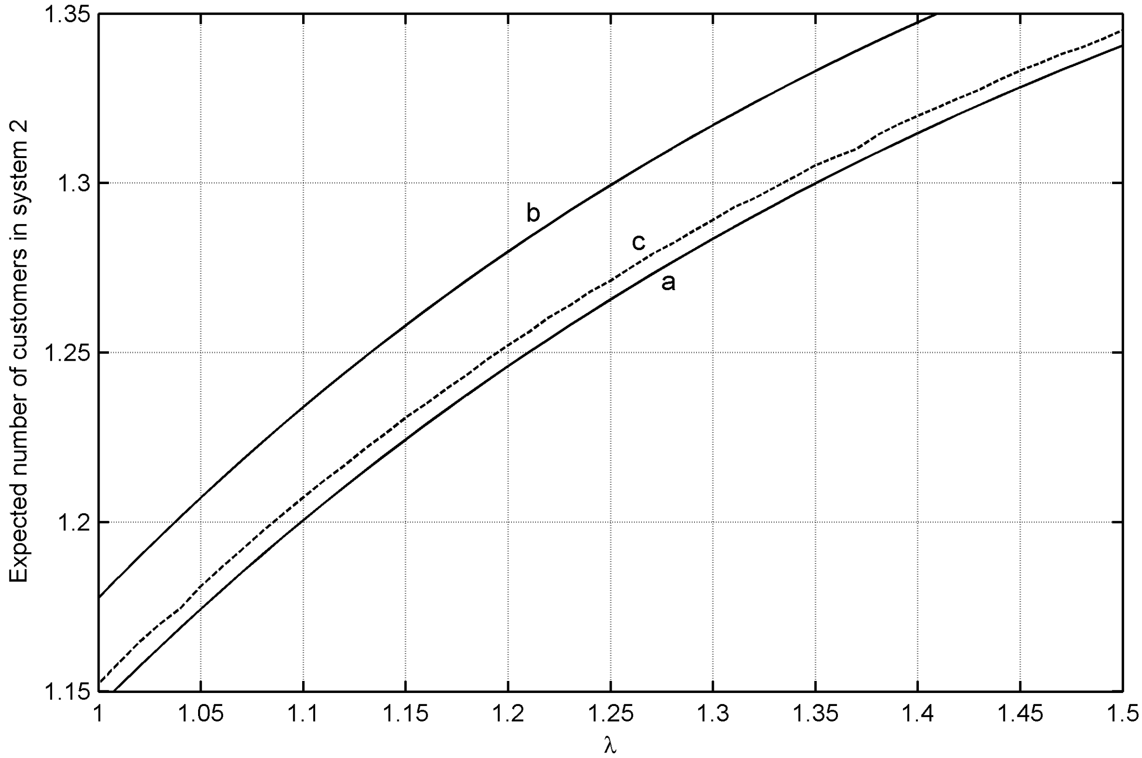


Figure 6.26.: Two queueing systems in tandem: number of customers in the downstream system. The arrival rate in the upstream system (system 1) is λ, the service rates are $\mu_1 = \mu_2 = 1$. $S = 3$. (a) network decomposition using the structure I+S,S,S,... (b) exact results, (c) the interarrival times of the second system are the interdeparture times of the first system without interdependencies.

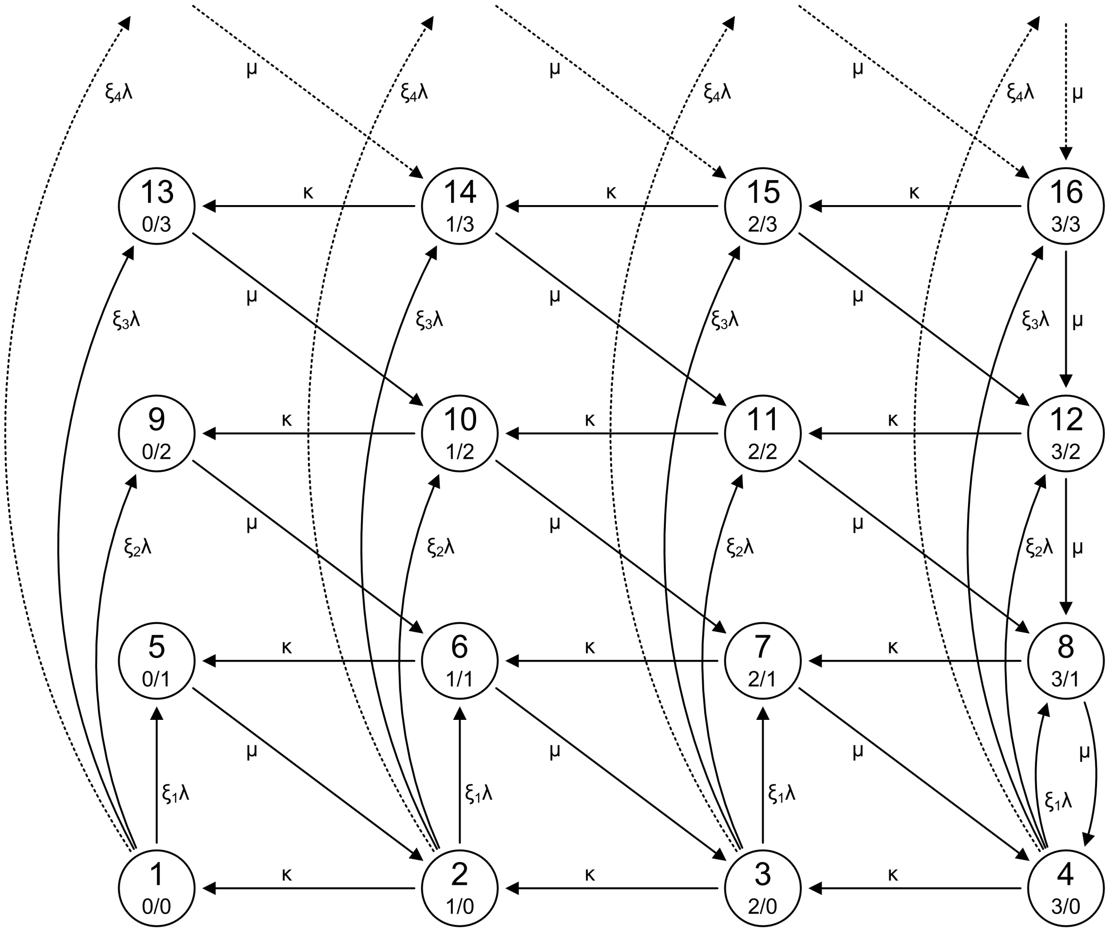


Figure 6.27.: Two GI/M/1/S queueing systems in tandem: Markov chain for the system state of system 2. Meaning of the names of the states: number of customers in system 2 / remaining number of customers which are served during the current busy period of system 1 or 0, if system 1 is empty. λ... arrival rate at system 1, μ ... service rate of system 1, κ ... service rate of system 2.

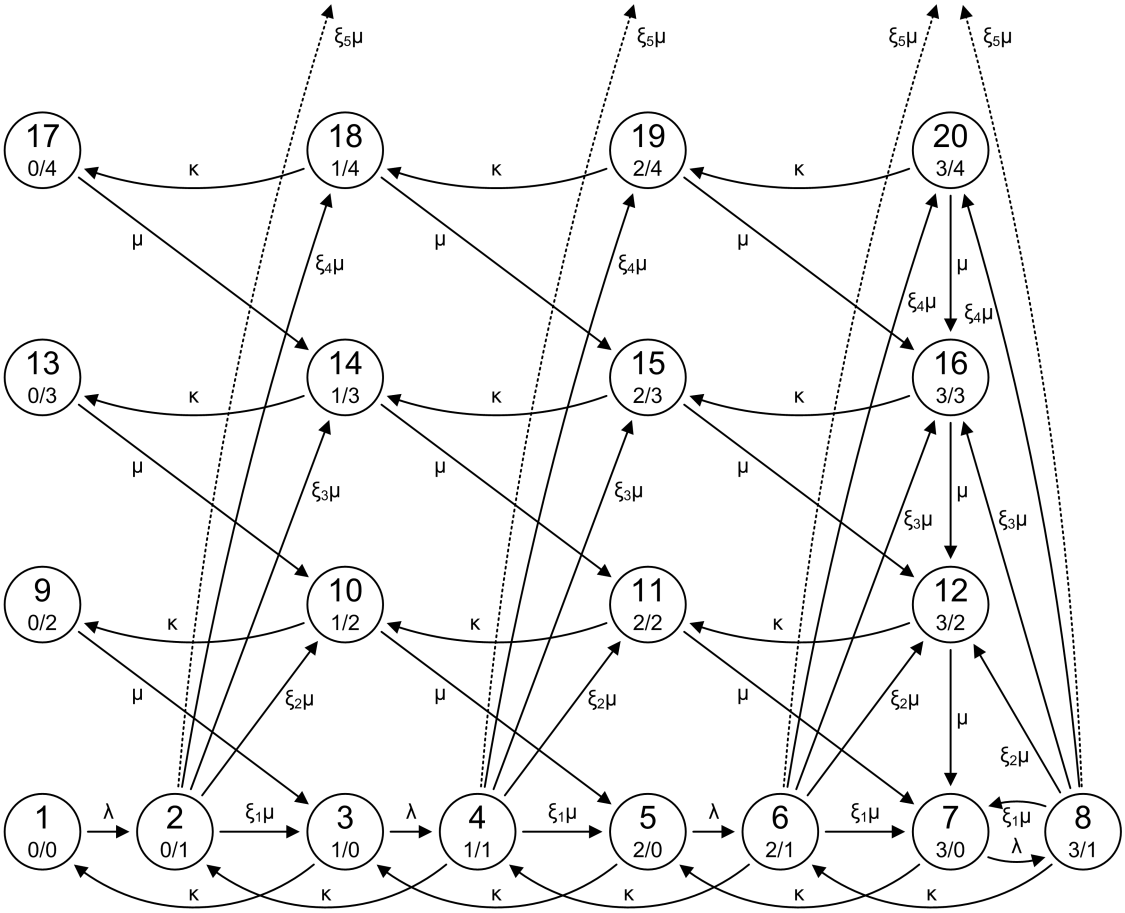


Figure 6.28.: Two GI/M/1/S queueing systems in tandem: Markov chain for the system state of system 2. Meaning of the names of the states: number of customers in system 2 / remaining number of customers which are served during the current busy period of system 1 or 0, if system 1 is empty. λ... arrival rate at system 1, μ ... service rate of system 1, κ ... service rate of system 2.

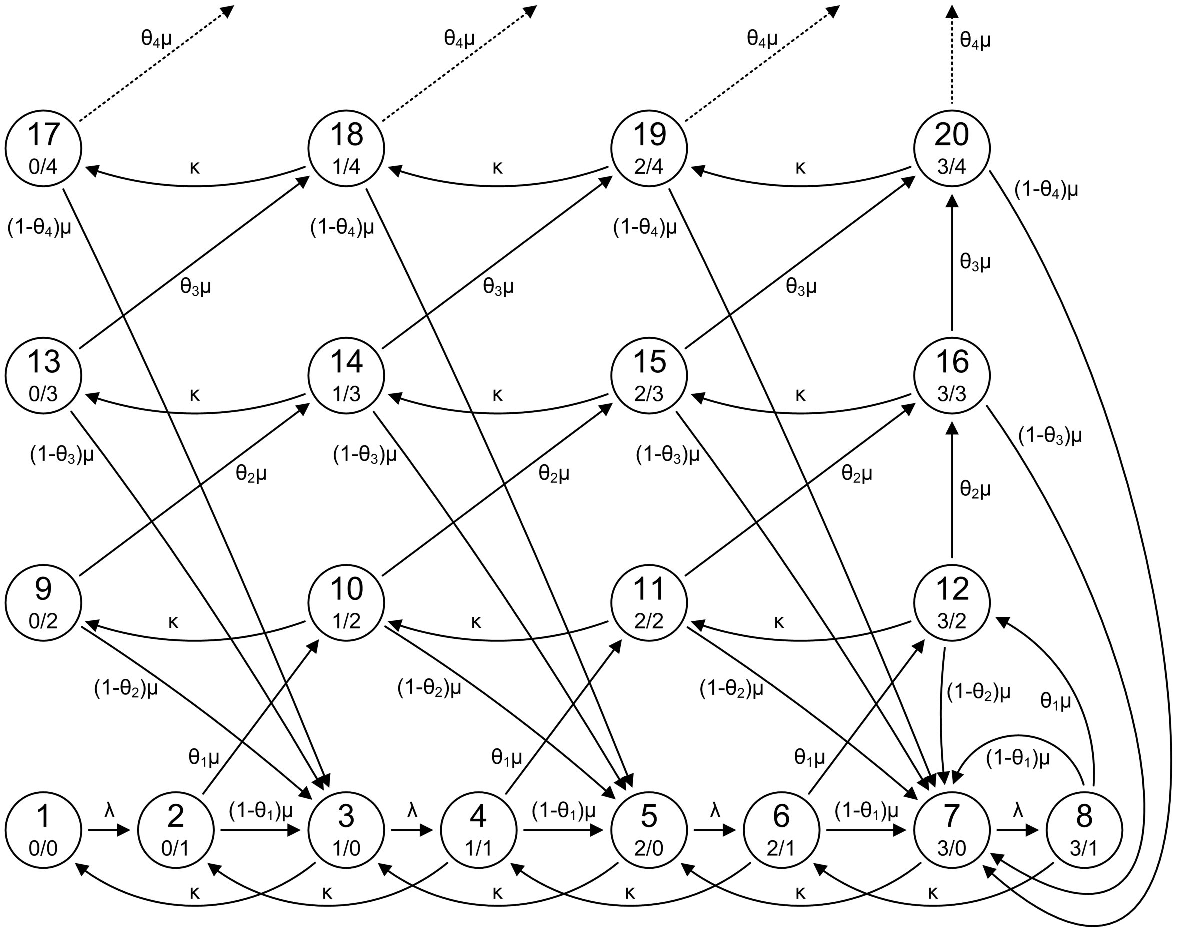


Figure 6.29.: Two GI/M/1/S queueing systems in tandem: Markov chain for the system state of system 2. Meaning of the names of the states: number of customers in system 2 / minimum number of customers which are served during the current busy period of system 1 or 0, if system 1 is empty. λ... arrival rate at system 1, μ ... service rate of system 1, κ ... service rate of system 2. $\theta_i = \mathrm{P}(\xi > i \,|\, \xi \geq i)$.

7. Overflow traffic

Customers that are prevented from entering a queueing system (e.g., because the queueing system is full) are either discarded or redirected to other queueing systems. In the latter case, these customers constitute the so-called overflow traffic, with which we deal in this chapter.

In Sections 7.1 and 7.3, we determine the interoverflow time and the number of successful arrivals between two consecutive overflows. In Section 7.2, we calculate the blocking probability of queueing systems. In Section 7.4, we show how a small network containing an overflow stream can be modelled. In Section 7.5, we show how such a network can by analysed using network decomposition.

7.1. Probability distribution of the interoverflow times

For the calculation of the probability distribution of the interoverflow times R we use two Markov chains. The first one is the Markov chain for the system state, $\mathcal{M}_S$. The second one, $\mathcal{M}_R$, is a modification of the Markov chain for the system state, which contains an additional absorbing state $\langle \mathrm{R} \rangle$ *("Customer has been rejected")*. This state is reached whenever a customer is blocked. That means we have to redirect all transitions (including hidden transitions) that correspond to a blocking of a customer to state $\langle \mathrm{R} \rangle$.

The first step is to determine (with the aid of the Markov chain for the system state) the probabilities σ_i^R that the queueing system is in state i immediately after a rejection has occurred.

The second step is to calculate (in $\mathcal{M}_R$) for all states i with $\sigma_i^R > 0$ the time R_i that the Markov chain needs to go from state i to state $\langle \mathrm{R} \rangle$. Let $\varphi_i(\cdot)$ be the complementary cumulative distribution function of this time. According to Equations 2.54 and 2.55, we have

$$\begin{aligned}\varphi_i(0) &= \begin{cases} 1 & i \neq \langle \mathrm{R} \rangle \\ 0 & i = \langle \mathrm{R} \rangle \end{cases} \\ \varphi'(\tau) &= \mathcal{Q} \cdot \varphi(\tau)\end{aligned} \tag{7.1}$$

Now we know the length of the interval between two consecutive rejections of customers given that the system is in state i after a rejection, and we know the probability of this happening. We can calculate the cumulative distribution function $F_R(t)$ of the interoverflow times with

$$F_R(\tau) = 1 - \sum_i \sigma_i^R \varphi_i(\tau) \tag{7.2}$$

The mean of the interblocking times can be calculated from $F_R(t)$ with

$$\mathrm{E}(R) = \int\limits_0^\infty (1 - F_R(\tau))\,\mathrm{d}\tau = \int\limits_0^\infty \sum_i \sigma_i^R \varphi_i(\tau)\mathrm{d}\tau \tag{7.3}$$

Another method is to calculate the overflow rate from the stationary state probabilities of $\mathcal{M}_S$ and the rate at which each state produces overflows:

$$\frac{1}{\mathrm{E}(R)} = \sum_i \pi_i \left(g_i + \sum_{j \neq i} h_{ij} \right) \tag{7.4}$$

where

$$h_{ij} = \begin{cases} q_{ij} & \text{if } i \to j \text{ corresponds to an overflow} \\ 0 & \text{otherwise} \end{cases} \tag{7.5}$$

and g_i is the rate at which overflows are produced while the system remains in state i (silent events). If we use the Markov chain $\mathcal{M}_R$, which we used for the calculation of the distribution of the interoverflow time, we can write

$$\frac{1}{\mathrm{E}(R)} = \sum_i \pi_i q_{i,\langle\mathrm{R}\rangle} \tag{7.6}$$

7.1.1. M/M/1/S queueing system

Figure 7.1a shows the Markov chain for the system state of an M/M/1/S queueing system with $S = 3$. An overflow takes place when the system is full and a customer arrives. The blocking of the customer does not change the state of the queueing system. Therefore, immediately after the occurrence of an overflow the system is in state $\langle 3 \rangle$, and we have $\sigma^R_{\langle 3 \rangle} = 1$.

The second Markov chain needed, $\mathcal{M}_R$, is shown in Figure 7.1b. The hidden transition that describes the arrival of customers that are blocked now leads from state $\langle 3 \rangle$ to the new state $\langle \mathrm{R} \rangle$.

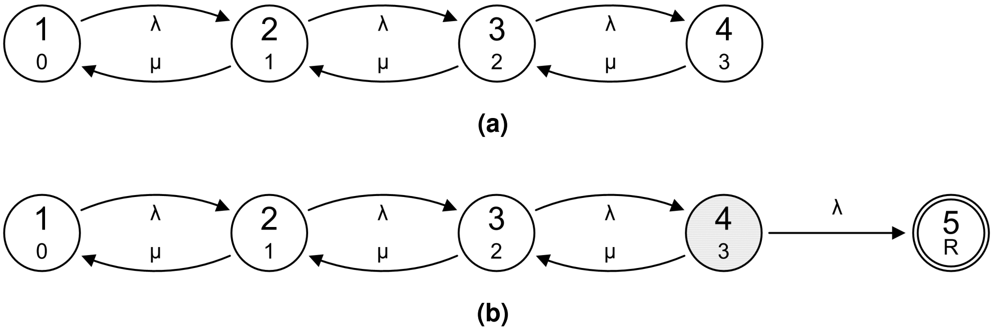


Figure 7.1.: Overflow stream of an M/M/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the time to the next rejection ($\mathcal{M}_R$). Meaning of the names of the states: number of customers in the system.

Now we use Equation 7.1 to calculate the complementary cumulative distribution function of the time $R_{\langle 3\rangle}$, which the Markov chain needs to go from state $\langle 3\rangle$ to state $\langle \mathrm{R}\rangle$.

$$\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \qquad\qquad \varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{R}\rangle \\ 0 & i = \langle \mathrm{R}\rangle \end{cases} \tag{7.7}$$

Since state $\langle 3\rangle$ is the only state in which the system can be after an overflow has occurred, $R_{\langle 3\rangle}$ is the interoverflow time of this queueing system.

$$F_R(t) = 1 - \varphi_{\langle 3\rangle}(t) \tag{7.8}$$

The mean of the interblocking times is

$$\mathrm{E}(R) = \int\limits_0^\infty \varphi_{\langle 3\rangle}(t)\mathrm{d}t \tag{7.9}$$

or

$$\mathrm{E}(R) = \frac{1}{\pi_{\langle 3\rangle}\cdot\lambda} \tag{7.10}$$

because when the system is in state $\langle 3\rangle$, it produces overflows with the arrival rate λ. Using the closed-form solution for the stationary system state probabilities (Equations 3.16 and 3.17),

$$\pi_0 = \begin{cases} \frac{1-\lambda/\mu}{1-(\lambda/\mu)^{S+1}} & \text{for } \lambda \neq \mu \\ \frac{1}{S+1} & \text{for } \lambda = \mu \end{cases} \tag{7.11}$$

$$\pi_n = \pi_0 \left(\frac{\lambda}{\mu}\right)^n \qquad n = 1,\ldots,S \tag{7.12}$$

we obtain

$$\mathrm{E}(R) = \begin{cases} \frac{1-(\lambda/\mu)^{S+1}}{(1-\lambda/\mu)(\lambda/\mu)^S\lambda} & \text{for } \lambda \neq \mu \\ \frac{S+1}{\lambda} & \text{for } \lambda = \mu \end{cases} \tag{7.13}$$

7.1.2. M/M/1/S queueing system with RED

Random early detection (RED) is a queue management algorithm, where customers that arrive at the queueing system are accepted with a probability that depends on the queue length. If the system is empty, all customers are accepted. With increasing queue length also the probability for rejecting a customer increases. If the system is full, all customers are rejected.

The Markov chain for the system state of an M/M/1/S queueing system with RED is shown in Figure 7.2a. When the system is in state $\langle 0 \rangle$, all arriving customers are accepted. When the system is in state $\langle 1 \rangle$, arriving customers are accepted with probability $1 - d_1$, and with probability d_1 they are rejected. When the system is in state $\langle 2 \rangle$, arriving customers are accepted with probability $1 - d_2$, and with probability d_2 they are rejected. When the system is in state $\langle 3 \rangle$, all arriving customers are rejected.

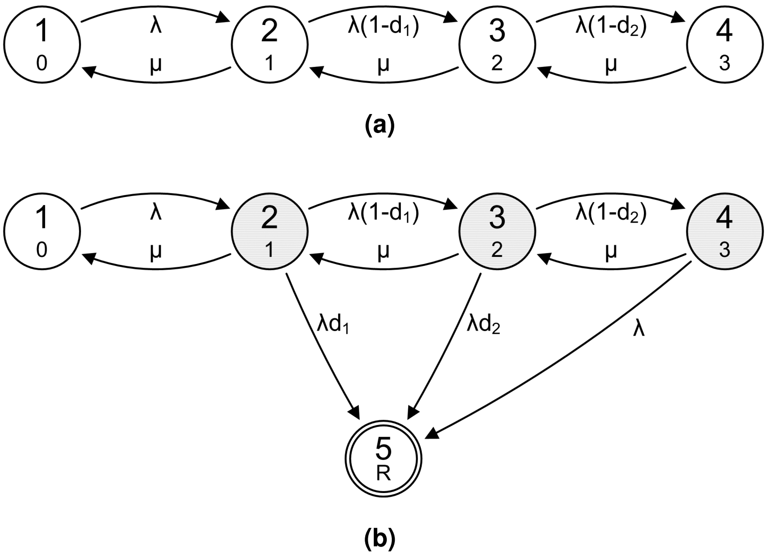


Figure 7.2.: Overflow stream of an M/M/1/S queueing system with RED. (a) Markov chain for the system state, (b) Markov chain for the calculation of the time to the next rejection ($\mathcal{M}_R$). Meaning of the names of the states: number of customers in the system.

Since the states $\langle 1 \rangle$ and $\langle 2 \rangle$ produce rejections at rate $\lambda\pi_{\langle 1 \rangle}d_1$ and $\lambda\pi_{\langle 2 \rangle}d_2$, respectively, and state $\langle 3 \rangle$ produces rejections at rate $\lambda\pi_{\langle 3 \rangle}$, the total rate at which rejections are produced is $\sum_{i=1}^{2} \lambda\pi_{\langle i \rangle}d_i + \lambda\pi_{\langle 3 \rangle}$. The probability that the system is in state $\langle k \rangle$ when a customer is rejected (and therefore will be in that state immediately after the rejection), is

$$\sigma^R_{\langle k \rangle} = \frac{\pi_{\langle k \rangle}d_k}{\sum\limits_{i=1}^{S-1} \pi_{\langle i \rangle}d_i + \pi_{\langle S \rangle}} \qquad \text{for } k = 1 \ldots S-1 \tag{7.14}$$

$$\sigma^R_{\langle S \rangle} = \frac{\pi_{\langle S \rangle}}{\sum\limits_{i=1}^{S-1} \pi_{\langle i \rangle}d_i + \pi_{\langle S \rangle}} \tag{7.15}$$

The Markov chain we use for the calculation of the time to the next rejection, $\mathcal{M}_R$, is shown in Figure 7.2b. The hidden transitions that correspond to rejections now lead to state $\langle \mathrm{R} \rangle$. We calculate the complementary cumulative distribution function $\varphi_{\langle k \rangle}(\cdot),\ k = 1 \dots S$, of the time that will pass until the next blocking occurs, given that the system is in state $\langle k \rangle$ with

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \qquad\qquad \varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{R} \rangle \\ 0 & i = \langle \mathrm{R} \rangle \end{cases} \tag{7.16}$$

The cumulative distribution function of the interblocking time is

$$F_R(\tau) = 1 - \sum_i \sigma_i^R \varphi_i(\tau) \tag{7.17}$$

A comparison of the overflow traffic of an M/M/1/S queueing system with and without RED is shown in Figure 7.3.

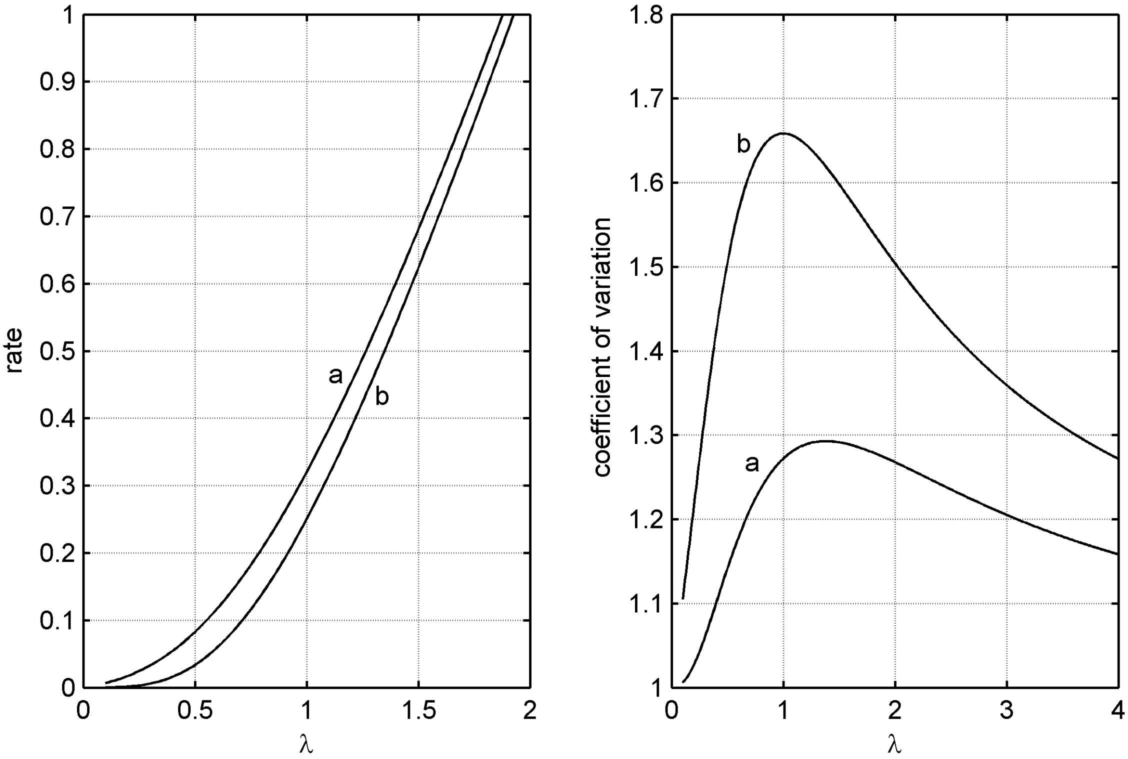


Figure 7.3.: Overflow stream of an M/M/1/S queueing system (a) with $(d_0 = 0.05,\ d_1 = 0.2,\ d_2 = 0.45)$ and (b) without RED. $S = 3$, arrival rate λ, service rate $\mu = 1$.

7.1.3. Hypo/M/1/S queueing system

Figure 7.4a shows the Markov chain for the system state of a Hypo/M/1/S queueing system with $S = 3$.

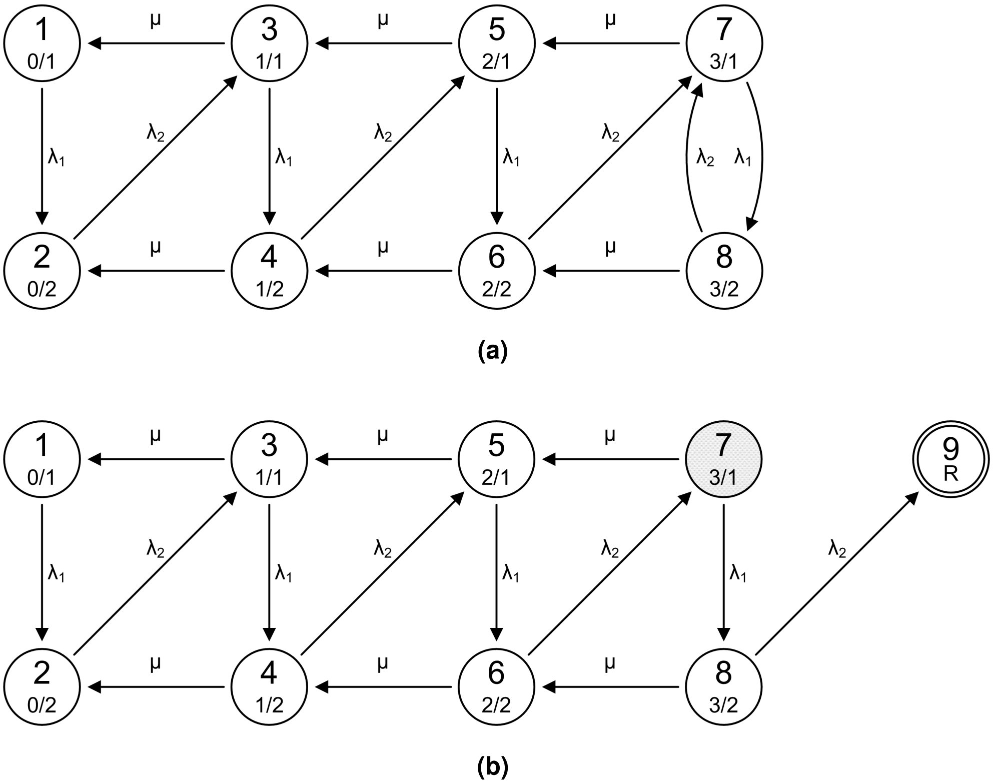


Figure 7.4.: Overflow stream of a Hypo/M/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the time to the next rejection. Meaning of the names of the states: number of customers in the system / state of the arrival process.

An overflow takes place when the system is full (states $\langle 3/\cdot\rangle$) and a customer arrives (transitions $\langle\cdot/2\rangle \to \langle\cdot/1\rangle$). Therefore, transition $\langle 3/2\rangle \to \langle 3/1\rangle$ corresponds to a rejection. After the rejection, the system is in state $\langle 3/1\rangle$, so we have

$$\sigma^{R}_{\langle 3/1\rangle} = 1 \tag{7.18}$$

Figure 7.4b shows the extended Markov chain. The transition from state $\langle 3/2\rangle$ to state $\langle 3/1\rangle$ now leads to state $\langle \mathrm{R}\rangle$

The cumulative distribution function of the interblocking times is calculated with

$$\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \qquad\qquad \varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{R}\rangle \\ 0 & i = \langle \mathrm{R}\rangle \end{cases} \tag{7.19}$$

and

$$F_R(\tau) = 1 - \varphi_{\langle 3/1\rangle}(\tau) \tag{7.20}$$

The mean of the interblocking times is

$$\mathrm{E}(R) = \int\limits_{0}^{\infty} \varphi_{\langle 3/1 \rangle}(t)\mathrm{d}t \tag{7.21}$$

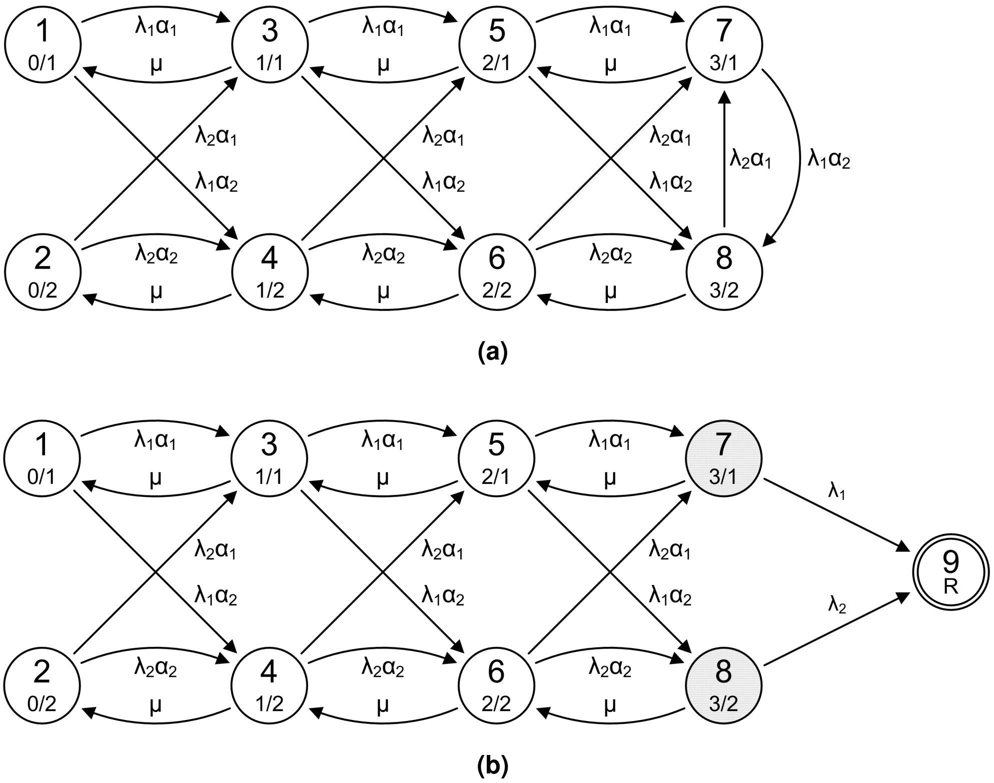
or

$$\mathrm{E}(R) = \frac{1}{\pi_{\langle 3/2 \rangle} \cdot \lambda_2} \tag{7.22}$$

because when the system is in state $\langle 3/2 \rangle$, it produces overflows at rate λ_2. It should be noted that the transition from state $\langle 3/1 \rangle$ to $\langle 3/2 \rangle$ does not produce an overflow, because it does not correspond to the actual arrival of a customer, but only to a change of the state of the arrival process.

7.1.4. Hyper/M/1/S queueing system

Figure 7.5a shows the Markov chain for the system state of a Hyper/M/1/S queueing system with $S = 3$.



(a)

(b)

Figure 7.5.: Overflow stream of a Hyper/M/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the time to the next rejection. Meaning of the names of the states: number of customers in the system / state of the arrival process.

A rejection takes place when the system is full (states $\langle 3/1\rangle$ and $\langle 3/2\rangle$) and a customer arrives. When the system is in state $\langle 3/1\rangle$, customers arrive at rate λ_1, when it is in state $\langle 3/2\rangle$, customers arrive at rate λ_2.

Independent of the state in which the Markov chain was before the rejection occurred, after a blocking it is state $\langle 3/1\rangle$ with probability α_1 and in state $\langle 3/2\rangle$ with probability α_2:

$$\sigma^R_{\langle 3/1\rangle} = \alpha_1 \tag{7.23}$$
$$\sigma^R_{\langle 3/2\rangle} = \alpha_2 \tag{7.24}$$

Therefore, the cumulative distribution function of the interblocking times is calculated with

$$\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \qquad\qquad \varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{R}\rangle \\ 0 & i = \langle \mathrm{R}\rangle \end{cases} \tag{7.25}$$

and

$$F_R(\tau) = 1 - \alpha_1\varphi_{\langle 3/1\rangle}(\tau) - \alpha_2\varphi_{\langle 3/2\rangle}(\tau) \tag{7.26}$$

The mean of the interblocking times is

$$\mathrm{E}(R) = \int\limits_0^\infty \big(\alpha_1\varphi_{\langle 3/1\rangle}(t) + \alpha_2\varphi_{\langle 3/2\rangle}(t)\big)\,\mathrm{d}t \tag{7.27}$$

or

$$\mathrm{E}(R) = \frac{1}{\pi_{\langle 3/1\rangle}\cdot\lambda_1 + \pi_{\langle 3/2\rangle}\cdot\lambda_2} \tag{7.28}$$

7.1.5. Other PH/PH/1/S queueing systems

M/Hypo/1/S queueing system

Figure 7.6a shows the Markov chain for the system state of an M/Hypo/1/S queueing system with $S = 3$. There is no transition that corresponds to the rejection of a customer, so after a rejection, the Markov chain is in the same state as it was before the rejection occurred. If it was in state $\langle 3/1\rangle$ (probability $\pi_{\langle 3/1\rangle}/\big(\pi_{\langle 3/1\rangle} + \pi_{\langle 3/2\rangle}\big)$), it is in state $\langle 3/1\rangle$, if it was in state $\langle 3/2\rangle$ (probability $\pi_{\langle 3/2\rangle}/\big(\pi_{\langle 3/1\rangle} + \pi_{\langle 3/2\rangle}\big)$), it is in state $\langle 3/2\rangle$. So we have

$$\sigma^R_{\langle 3/1\rangle} = \frac{\pi_{\langle 3/1\rangle}}{\pi_{\langle 3/1\rangle} + \pi_{\langle 3/2\rangle}} \tag{7.29}$$
$$\sigma^R_{\langle 3/2\rangle} = \frac{\pi_{\langle 3/2\rangle}}{\pi_{\langle 3/1\rangle} + \pi_{\langle 3/2\rangle}} \tag{7.30}$$

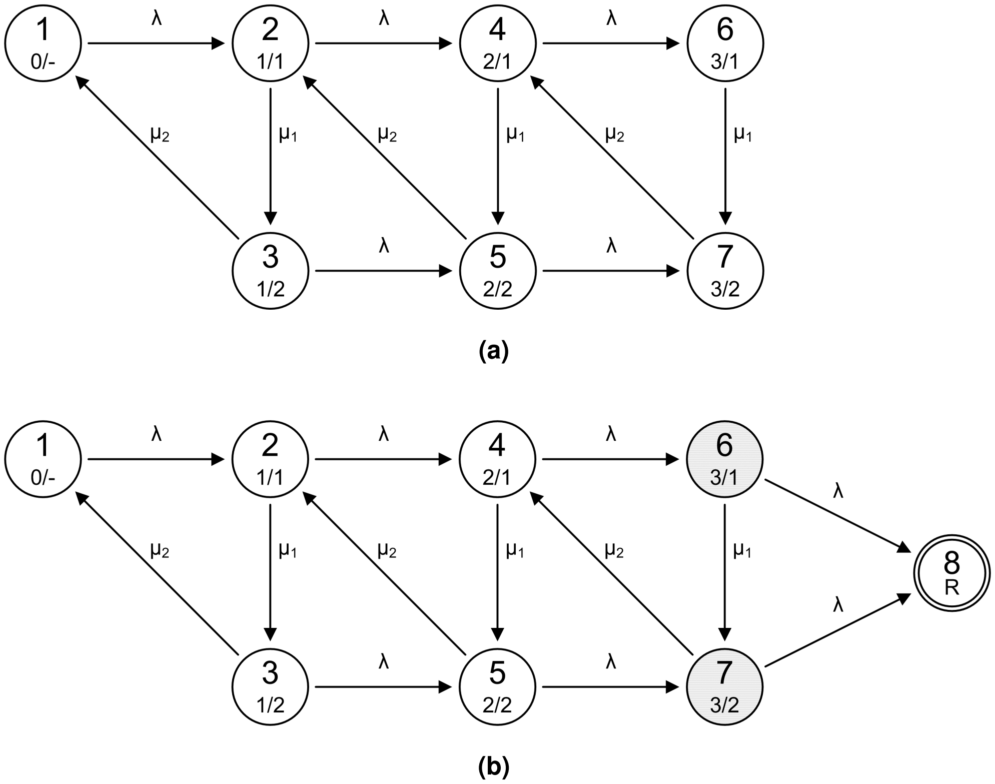


Figure 7.6.: Overflow stream of an M/Hypo/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the time to the next rejection. Meaning of the names of the states: number of customers in the system / state of the service process.

The second Markov chain needed, $\mathcal{M}_R$, is shown in Figure 7.6b. The hidden transitions that correspond to the arrival of customers that are blocked now lead from states $\langle 3/1 \rangle$ and $\langle 3/2 \rangle$ to the new state $\langle \mathrm{R} \rangle$.

Now the cumulative distribution function of the interblocking times is calculated with

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \qquad \varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{R} \rangle \\ 0 & i = \langle \mathrm{R} \rangle \end{cases} \tag{7.31}$$

and

$$F_R(\tau) = 1 - \frac{\pi_{\langle 3/1 \rangle}}{\pi_{\langle 3/1 \rangle} + \pi_{\langle 3/2 \rangle}} \varphi_{\langle 3/1 \rangle}(\tau) - \frac{\pi_{\langle 3/2 \rangle}}{\pi_{\langle 3/1 \rangle} + \pi_{\langle 3/2 \rangle}} \varphi_{\langle 3/2 \rangle}(\tau) \tag{7.32}$$

M/Hyper/1/S queueing system

The Markov chains for an M/Hyper/1/S queueing system with $S = 3$ are shown in Figure 7.7.

After a rejection, the Markov chain is in the same state as it was before the blocking occurred. This is state $\langle 3/1 \rangle$ (with probability $\pi_{\langle 3/1 \rangle} / \left(\pi_{\langle 3/1 \rangle} + \pi_{\langle 3/2 \rangle} \right)$) or state $\langle 3/2 \rangle$

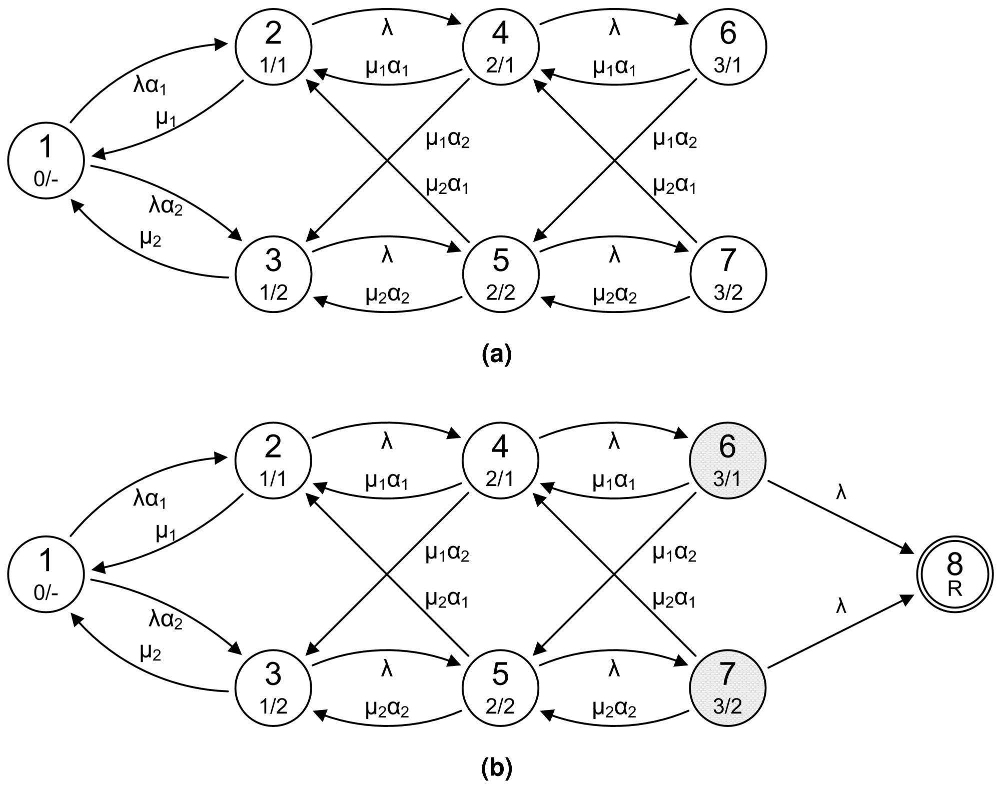


Figure 7.7.: Overflow stream of an M/Hyper/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the time to the next rejection. Meaning of the names of the states: number of customers in the system / state of the service process.

(with probability $\pi_{\langle 3/2\rangle}/\left(\pi_{\langle 3/1\rangle}+\pi_{\langle 3/2\rangle}\right)$). Therefore, we have

$$\sigma^R_{\langle 3/1\rangle} = \frac{\pi_{\langle 3/1\rangle}}{\pi_{\langle 3/1\rangle}+\pi_{\langle 3/2\rangle}} \tag{7.33}$$

$$\sigma^R_{\langle 3/2\rangle} = \frac{\pi_{\langle 3/2\rangle}}{\pi_{\langle 3/1\rangle}+\pi_{\langle 3/2\rangle}} \tag{7.34}$$

The cumulative distribution function of the interblocking times is calculated with

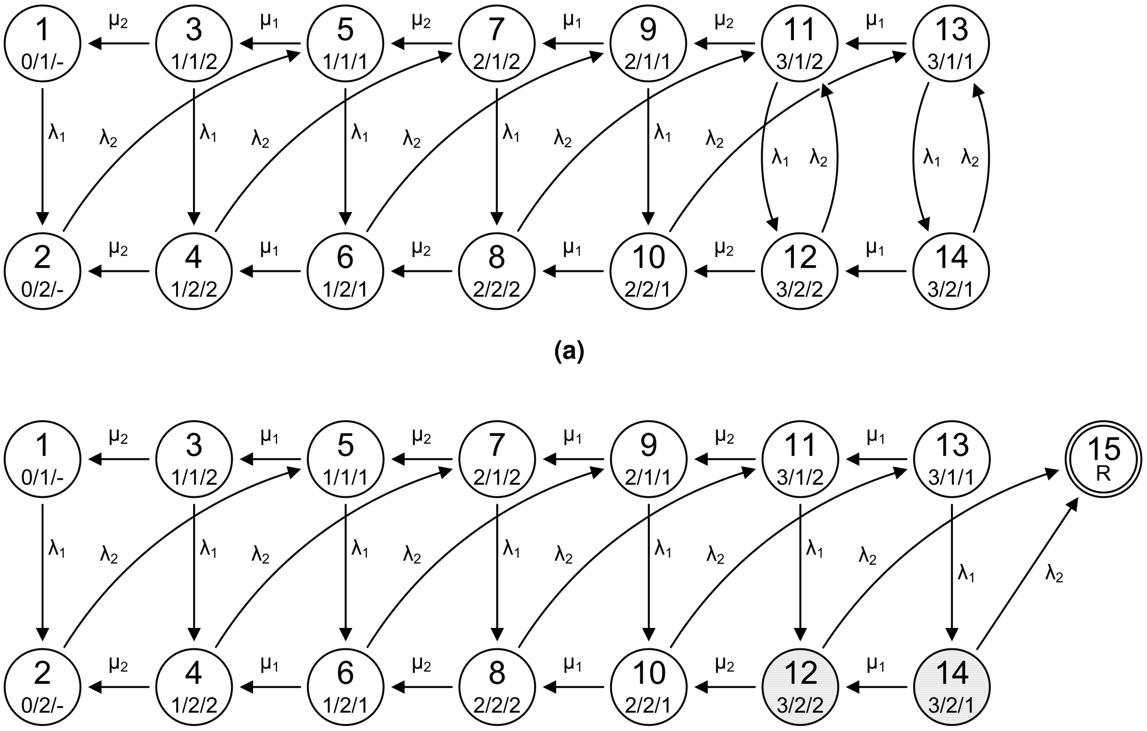
$$\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \qquad\qquad \varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{R}\rangle \\ 0 & i = \langle \mathrm{R}\rangle \end{cases} \tag{7.35}$$

and

$$F_R(\tau) = 1 - \frac{\pi_{\langle 3/1\rangle}}{\pi_{\langle 3/1\rangle}+\pi_{\langle 3/2\rangle}}\varphi_{\langle 3/1\rangle}(\tau) - \frac{\pi_{\langle 3/2\rangle}}{\pi_{\langle 3/1\rangle}+\pi_{\langle 3/2\rangle}}\varphi_{\langle 3/2\rangle}(\tau) \tag{7.36}$$

Hypo/Hypo/1/S queueing system

The Markov chains for a Hypo/Hypo/1/S queueing system with $S = 3$ are shown in Figure 7.8.



(a)

(b)

Figure 7.8.: Overflow stream of a Hypo/Hypo/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the time to the next rejection. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

In this system there are four states that represent a full system: $\langle 3/1/1\rangle, \langle 3/1/2\rangle$, $\langle 3/2/1\rangle$ and $\langle 3/2/2\rangle$. Transitions $\langle \cdot/2/\cdot\rangle \rightarrow \langle \cdot/1/\cdot\rangle$ correspond to arrivals of customers. Therefore, the transitions $\langle 3/2/1\rangle \rightarrow \langle 3/1/1\rangle$ and $\langle 3/2/2\rangle \rightarrow \langle 3/1/2\rangle$ correspond to rejections of customers.

Before a rejection occurred, the system must have been in state $\langle 3/2/1\rangle$ or in state $\langle 3/2/2\rangle$. If it was in state $\langle 3/2/1\rangle$ (probability $\pi_{\langle 3/2/1\rangle}/\left(\pi_{\langle 3/2/1\rangle}+\pi_{\langle 3/2/2\rangle}\right)$), it is in state $\langle 3/1/1\rangle$ afterwards. If it was in state $\langle 3/2/2\rangle$ (probability $\pi_{\langle 3/2/2\rangle}/\left(\pi_{\langle 3/2/1\rangle}+\pi_{\langle 3/2/2\rangle}\right)$), it is in state $\langle 3/1/2\rangle$ afterwards. Therefore, we have

$$\sigma^R_{\langle 3/1/1\rangle} = \frac{\pi_{\langle 3/2/1\rangle}}{\pi_{\langle 3/2/1\rangle}+\pi_{\langle 3/2/2\rangle}} \tag{7.37}$$

$$\sigma^R_{\langle 3/1/2\rangle} = \frac{\pi_{\langle 3/2/2\rangle}}{\pi_{\langle 3/2/1\rangle}+\pi_{\langle 3/2/2\rangle}} \tag{7.38}$$

The cumulative distribution function of the interblocking times is calculated with

$$\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \qquad\qquad \varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{R}\rangle \\ 0 & i = \langle \mathrm{R}\rangle \end{cases} \tag{7.39}$$

and

$$F_R(\tau) = 1 - \frac{\pi_{\langle 3/2/1\rangle}}{\pi_{\langle 3/2/1\rangle}+\pi_{\langle 3/2/2\rangle}}\varphi_{\langle 3/1/1\rangle}(\tau) - \frac{\pi_{\langle 3/2/2\rangle}}{\pi_{\langle 3/2/1\rangle}+\pi_{\langle 3/2/2\rangle}}\varphi_{\langle 3/1/2\rangle}(\tau) \tag{7.40}$$

Hypo/Hyper/1/S queueing system

The Markov chains for a Hypo/Hyper/1/S queueing system with $S = 3$ are shown in Figure 7.9.

In this system, there are four states that represent a full system: $\langle 3/1/1\rangle, \langle 3/1/2\rangle$, $\langle 3/2/1\rangle$ and $\langle 3/2/2\rangle$. Transitions $\langle \cdot/2/\cdot\rangle \rightarrow \langle \cdot/1/\cdot\rangle$ correspond to arrivals of customers. Therefore, the transitions $\langle 3/2/1\rangle \rightarrow \langle 3/1/1\rangle$ and $\langle 3/2/2\rangle \rightarrow \langle 3/1/2\rangle$ correspond to rejections of customers.

Before a rejection occurred, the system was in state $\langle 3/2/1\rangle$ or in state $\langle 3/2/2\rangle$. If it was in state $\langle 3/2/1\rangle$ (probability $\pi_{\langle 3/2/1\rangle}/\left(\pi_{\langle 3/2/1\rangle}+\pi_{\langle 3/2/2\rangle}\right)$), it is in state $\langle 3/1/1\rangle$ afterwards. If it was in state $\langle 3/2/2\rangle$ (probability $\pi_{\langle 3/2/2\rangle}/\left(\pi_{\langle 3/2/1\rangle}+\pi_{\langle 3/2/2\rangle}\right)$), it is in state $\langle 3/1/2\rangle$ afterwards. Therefore, we have

$$\sigma^R_{\langle 3/1/1\rangle} = \frac{\pi_{\langle 3/2/1\rangle}}{\pi_{\langle 3/2/1\rangle}+\pi_{\langle 3/2/2\rangle}} \tag{7.41}$$

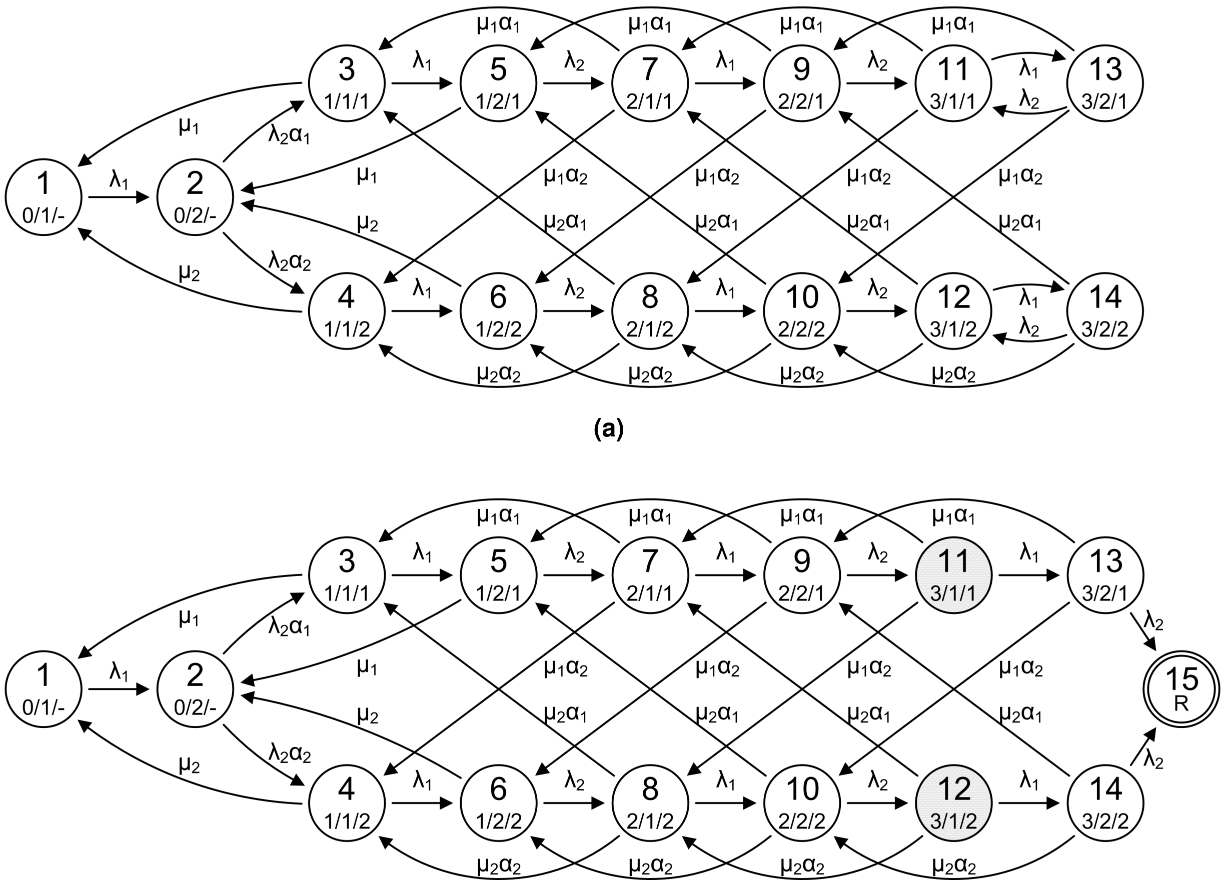
$$\sigma^R_{\langle 3/1/2\rangle} = \frac{\pi_{\langle 3/2/2\rangle}}{\pi_{\langle 3/2/1\rangle}+\pi_{\langle 3/2/2\rangle}} \tag{7.42}$$

The cumulative distribution function of the interblocking times is calculated with

$$\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \qquad\qquad \varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{R}\rangle \\ 0 & i = \langle \mathrm{R}\rangle \end{cases} \tag{7.43}$$

and

$$F_R(\tau) = 1 - \frac{\pi_{\langle 3/2/1\rangle}}{\pi_{\langle 3/2/1\rangle}+\pi_{\langle 3/2/2\rangle}}\varphi_{\langle 3/1/1\rangle}(\tau) - \frac{\pi_{\langle 3/2/2\rangle}}{\pi_{\langle 3/2/1\rangle}+\pi_{\langle 3/2/2\rangle}}\varphi_{\langle 3/1/2\rangle}(\tau) \tag{7.44}$$



(a)

(b)

Figure 7.9.: Overflow stream of a Hypo/Hyper/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the time to the next rejection. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

Hyper/Hypo/1/S queueing system

The Markov chains for a Hyper/Hypo/1/S queueing system with $S = 3$ are shown in Figure 7.10.

There are four states that represent a full system: $\langle 3/1/1\rangle, \langle 3/1/2\rangle, \langle 3/2/1\rangle$ and $\langle 3/2/2\rangle$. In all of these states, rejections can take place.

After a rejection, the Markov chain is in state $\langle 3/1/1\rangle$ or $\langle 3/1/2\rangle$ with probability α_1 and in state $\langle 3/2/1\rangle$ or $\langle 3/2/2\rangle$ with probability α_2.

If the system was in state $\langle 3/1/1\rangle$ or $\langle 3/2/1\rangle$ before the rejection (probability $\big(\lambda_1\pi_{\langle 3/1/1\rangle}+ \lambda_2\pi_{\langle 3/2/1\rangle}\big) \,/\, \big(\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}\big)$), it is in state $\langle 3/1/1\rangle$ or $\langle 3/2/1\rangle$ afterwards. If it was in state $\langle 3/1/2\rangle$ or $\langle 3/2/2\rangle$ before the rejection (probability $\big(\lambda_1\pi_{\langle 3/1/2\rangle}+ \lambda_2\pi_{\langle 3/2/2\rangle}\big) \,/\, \big(\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_1\pi_{\langle 3/1/2\rangle}+ \lambda_2\pi_{\langle 3/2/1\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}\big)$), it is in states $\langle 3/1/2\rangle$ or $\langle 3/2/2\rangle$ afterwards. Therefore, we have

$$\sigma^R_{\langle 3/1/1\rangle} = \alpha_1 \frac{\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle}}{\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}} \tag{7.45}$$

$$\sigma^R_{\langle 3/1/2\rangle} = \alpha_1 \frac{\lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}}{\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}} \tag{7.46}$$

$$\sigma^R_{\langle 3/2/1\rangle} = \alpha_2 \frac{\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle}}{\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}} \tag{7.47}$$

$$\sigma^R_{\langle 3/2/2\rangle} = \alpha_2 \frac{\lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}}{\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}} \tag{7.48}$$

The cumulative distribution function of the interblocking times is calculated with

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \qquad\qquad \varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{R}\rangle \\ 0 & i = \langle \mathrm{R}\rangle \end{cases} \tag{7.49}$$

and

$$F_R(\tau) = 1 - \sum_i \sigma^R_i \varphi_i(\tau) \tag{7.50}$$

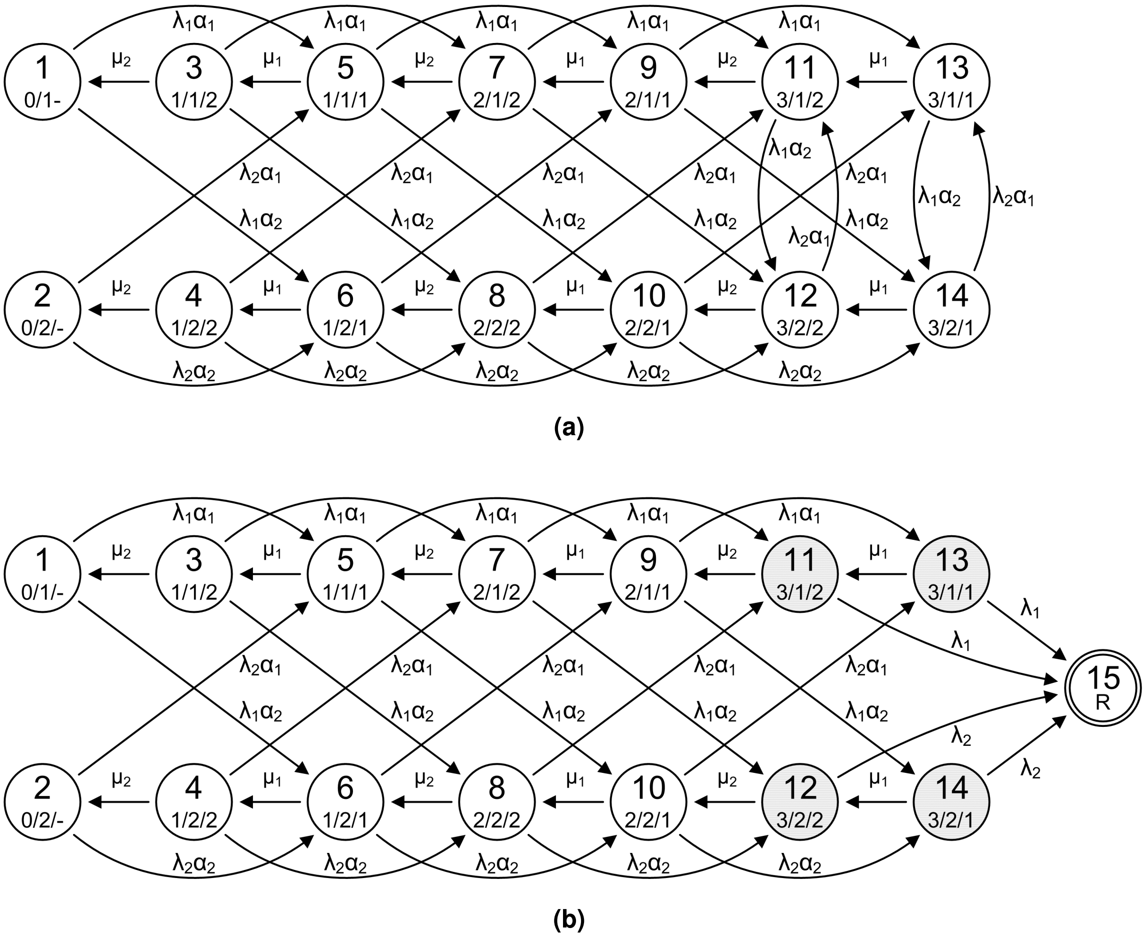
Hyper/Hyper/1/S queueing system

The Markov chains for a Hyper/Hyper/1/S queueing system with $S = 3$ are shown in Figure 7.11.

There are four states that represent a full system: $\langle 3/1/1\rangle, \langle 3/1/2\rangle, \langle 3/2/1\rangle$ and $\langle 3/2/2\rangle$. In all of these states, rejections can take place.

After a rejection, the Markov chain is in state $\langle 3/1/1\rangle$ or $\langle 3/1/2\rangle$ with probability α_1 and in state $\langle 3/2/1\rangle$ or $\langle 3/2/2\rangle$ with probability α_2.

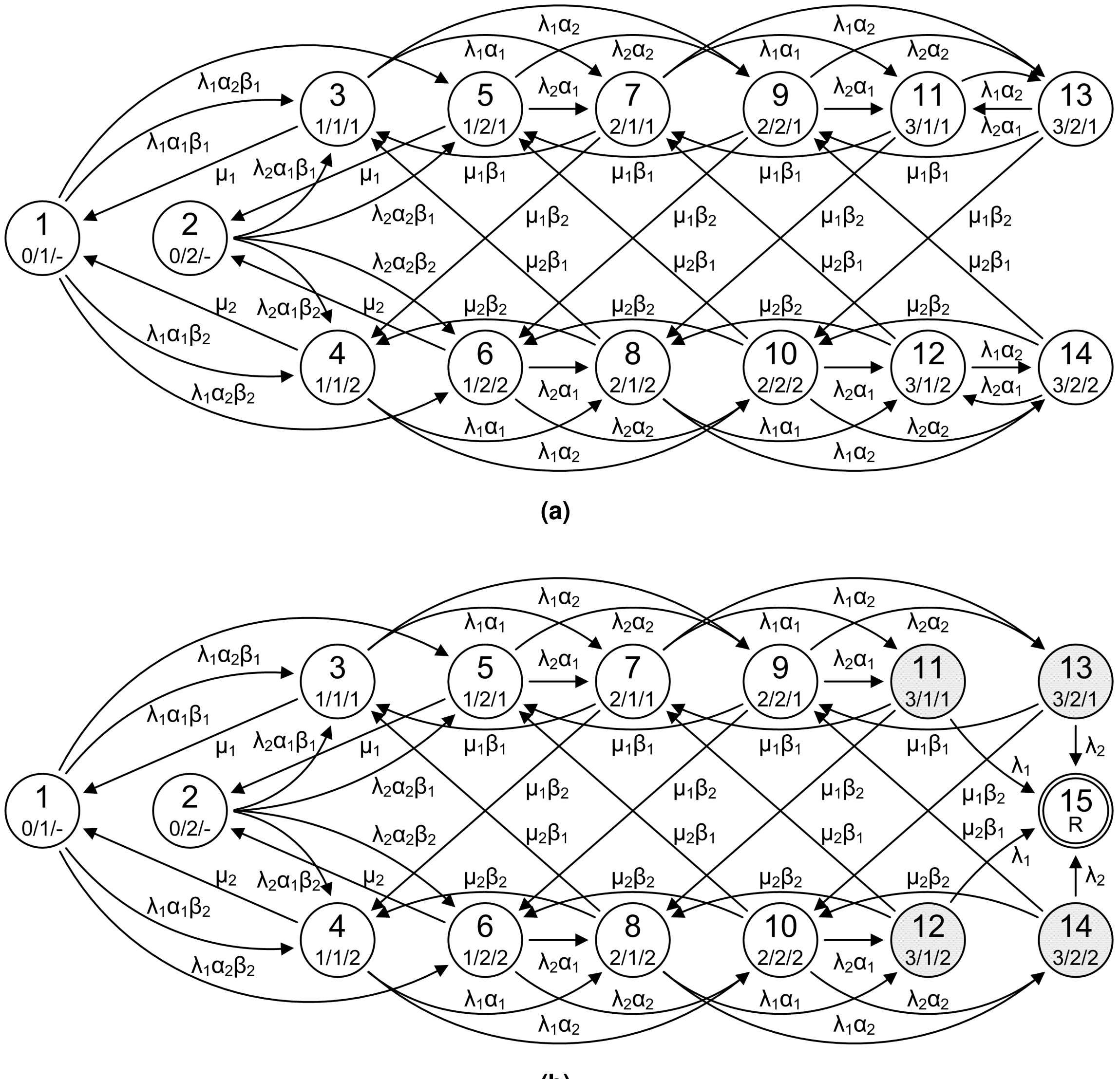
If the system was in state $\langle 3/1/1\rangle$ or $\langle 3/2/1\rangle$ before the rejection (probability $\big(\lambda_1\pi_{\langle 3/1/1\rangle}+ \lambda_2\pi_{\langle 3/2/1\rangle}\big) \,/\, \big(\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}\big)$), it is in state $\langle 3/1/1\rangle$ or $\langle 3/2/1\rangle$ afterwards. If it was in state $\langle 3/1/2\rangle$ or $\langle 3/2/2\rangle$ before the rejection (probability $\big(\lambda_1\pi_{\langle 3/1/2\rangle}+$



(a)

(b)

Figure 7.10.: Overflow stream of a Hyper/Hypo/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the time to the next rejection. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.



(a)

(b)

Figure 7.11.: Overflow stream of a Hyper/Hyper/1/S queueing system. (a) Markov chain for the system state, (b) Markov chain for the calculation of the time to the next rejection. Meaning of the names of the states: number of customers in the system / state of the arrival process / state of the service process.

$\lambda_2\pi_{\langle 3/2/2\rangle}\big) \,/\, \big(\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}\big)\big)$, it is in states $\langle 3/1/2\rangle$ or $\langle 3/2/2\rangle$ afterwards. Therefore, we have

$$\sigma^R_{\langle 3/1/1\rangle} = \alpha_1 \frac{\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle}}{\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}} \tag{7.51}$$

$$\sigma^R_{\langle 3/1/2\rangle} = \alpha_1 \frac{\lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}}{\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}} \tag{7.52}$$

$$\sigma^R_{\langle 3/2/1\rangle} = \alpha_2 \frac{\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle}}{\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}} \tag{7.53}$$

$$\sigma^R_{\langle 3/2/2\rangle} = \alpha_2 \frac{\lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}}{\lambda_1\pi_{\langle 3/1/1\rangle} + \lambda_1\pi_{\langle 3/1/2\rangle} + \lambda_2\pi_{\langle 3/2/1\rangle} + \lambda_2\pi_{\langle 3/2/2\rangle}} \tag{7.54}$$

The cumulative distribution function of the interblocking times is calculated with

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \qquad \varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{R} \rangle \\ 0 & i = \langle \mathrm{R} \rangle \end{cases} \tag{7.55}$$

and

$$F_R(\tau) = 1 - \sum_i \sigma_i^R \varphi_i(\tau) \tag{7.56}$$

Figures 7.12, 7.13 and 7.14 show the first two moments of the interoverflow times of the discussed queueing systems.

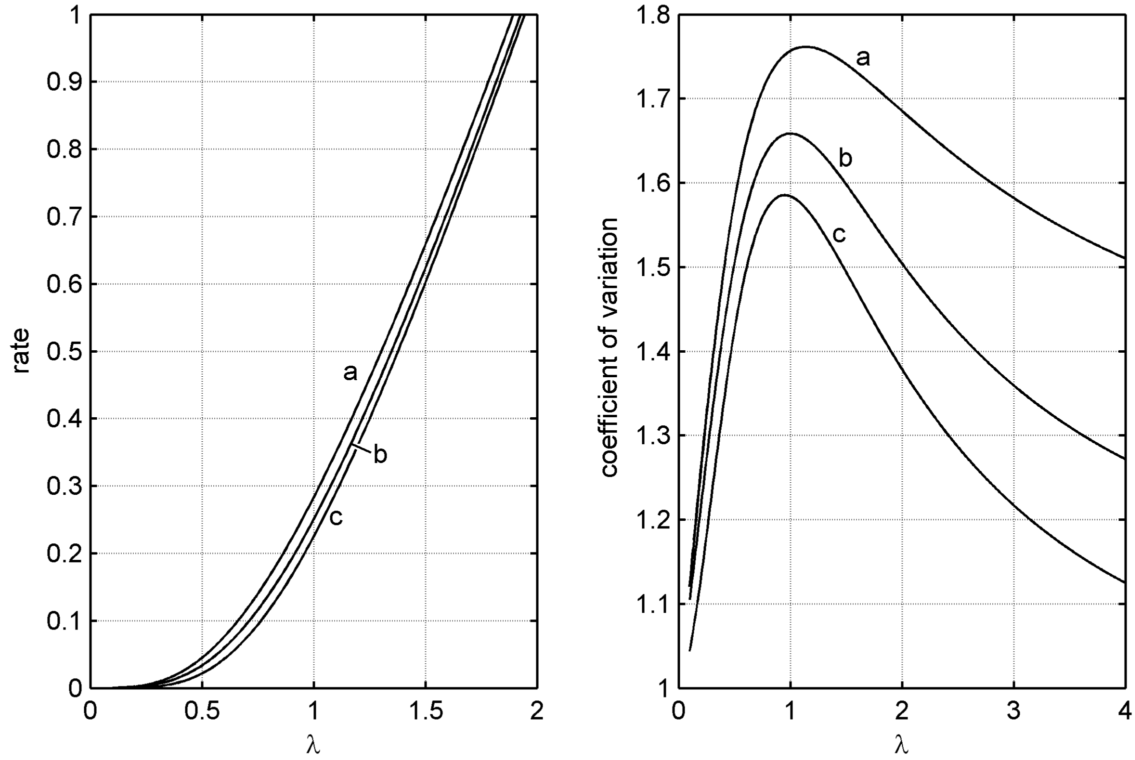


Figure 7.12.: Interoverflow times of PH/M/1/S queueing systems with arrival rate λ, service rate 1 and $S = 3$. (a) Hyper/M/1/S queueing system $(c_A = 1.25)$, (b) M/M/1/S queueing system, (c) Hypo/M/1/S queueing system $(c_A = 0.85)$.

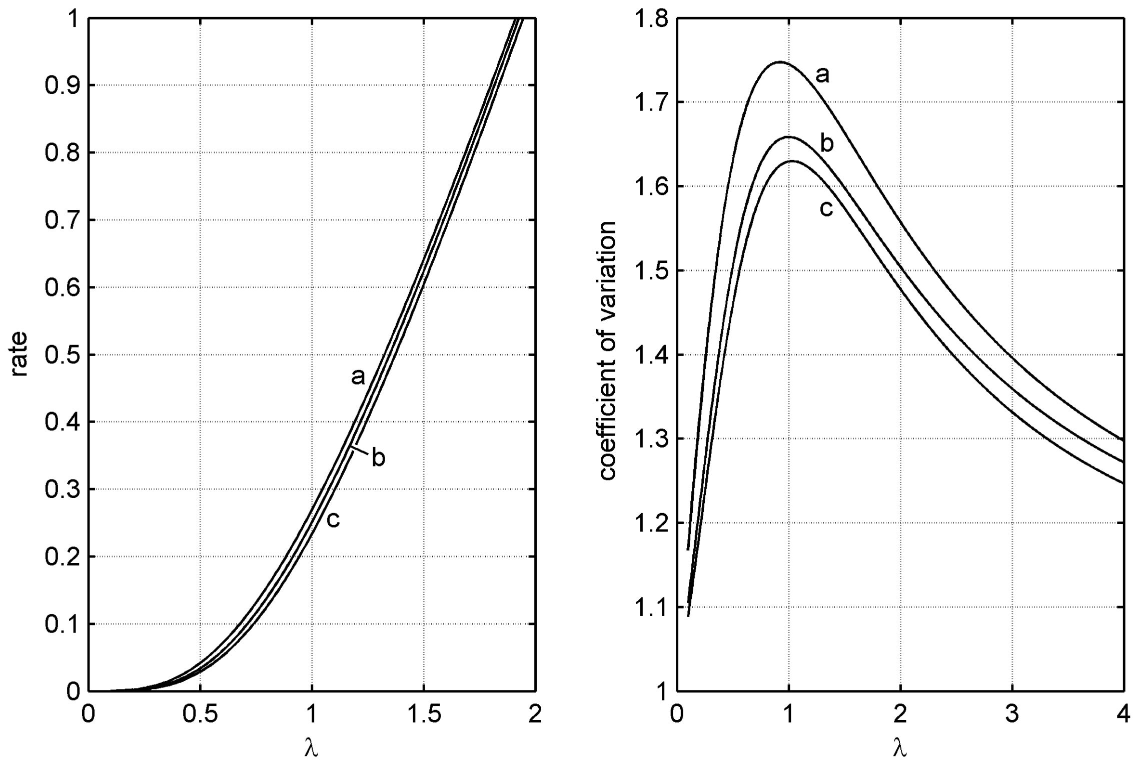


Figure 7.13.: Interoverflow times of M/PH/1/S queueing systems with arrival rate λ, service rate 1 and $S = 3$. (a) M/Hyper/1/S queueing system $(c_S = 1.25)$, (b) M/M/1/S queueing system, (c) M/Hypo/1/S queueing system $(c_S = 0.85)$.

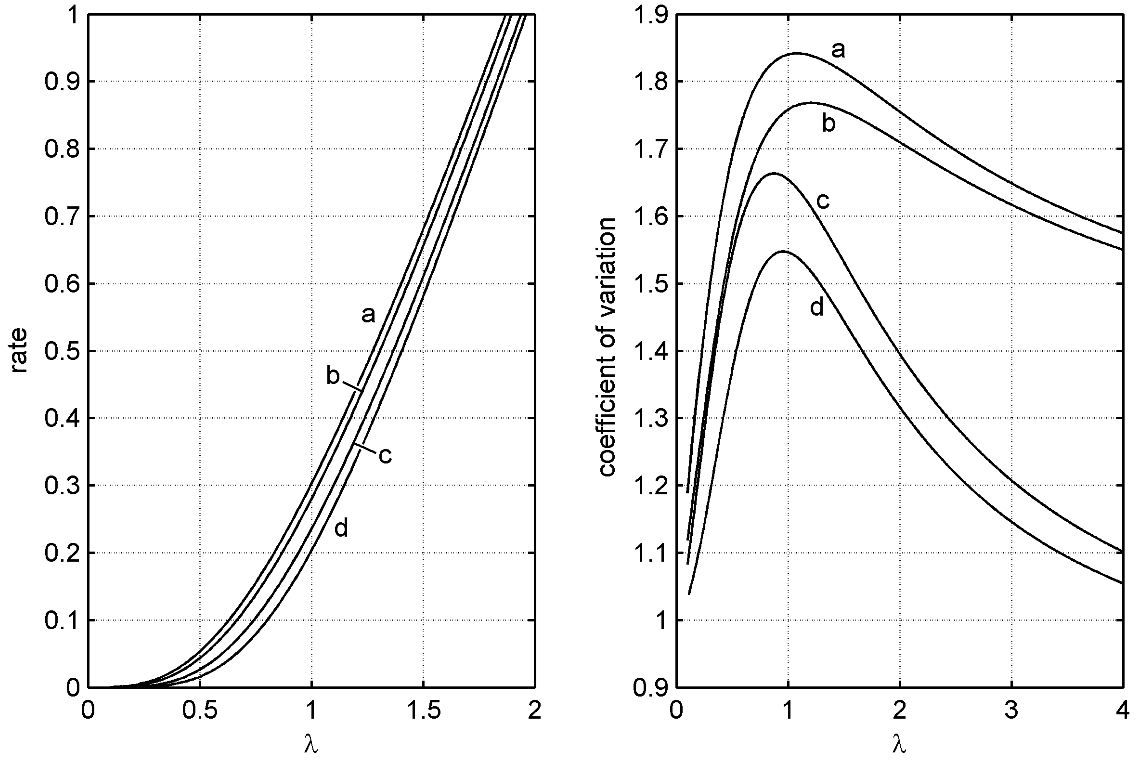


Figure 7.14.: Interoverflow times of PH/PH/1/S queueing systems with arrival rate λ, service rate 1 and $S = 3$. (a) Hyper/Hyper/1/S queueing system $(c_A = 1.3, c_S = 1.2)$, (b) Hyper/Hypo/1/S queueing system $(c_A = 1.3, c_S = 0.9)$, (c) Hypo/Hyper/1/S queueing system $(c_A = 0.8, c_S = 1.2)$, (d) Hypo/Hypo/1/S queueing system $(c_A = 0.8, c_S = 0.9)$.

7.2. Blocking probability

The blocking probability is the ratio of arriving customers that are rejected by the queueing system to all arriving customers. This probability is not the probability that the system rejects a customer that arrives at an arbitrary point in time, but the probability that a customer of the actual arrival stream is rejected.

To determine the blocking probability, we first need to calculate the total arrival rate λ_{total}:

$$\lambda_{\text{total}} = \sum_i \pi_i \left(g_i + \sum_{j \neq i} h_{ij} \right) \tag{7.57}$$

where

$$h_{ij} = \begin{cases} q_{ij} & \text{if } i \rightarrow j \text{ corresponds to an arrival} \\ 0 & \text{otherwise} \end{cases} \tag{7.58}$$

and g_i is the rate at which arrivals take place while the system remains in state i (silent events).

If the arrival rate is not state-dependent, λ_{total} is the arrival rate.

The rate at which rejections occur was calculated in the previous section (Equation 7.4):

$$\lambda_{\text{rejected}} = \sum_i \pi_i \left(g_i + \sum_{j \neq i} h_{ij} \right) \tag{7.59}$$

where

$$h_{ij} = \begin{cases} q_{ij} & \text{if } i \rightarrow j \text{ corresponds to an overflow} \\ 0 & \text{otherwise} \end{cases} \tag{7.60}$$

and g_i is the rate at which overflows are produced while the system remains in state i.

Now we have

$$p_{\text{blocking}} = \frac{\lambda_{\text{rejected}}}{\lambda_{\text{total}}} \tag{7.61}$$

7.2.1. M/M/1/S queueing system

In an M/M/1/S queueing system (Figure 7.15), the blocking probability equals the probability that the system is full (Figure 7.16):

$$p_{\text{blocking}} = p_{\text{full}} = \pi_{\langle S \rangle} \tag{7.62}$$

The reason is that when we have Poisson arrivals, the probability that a customer arrives is the same for each point in time; therefore the probability that an arriving customer finds a full system equals the steady-state probability that the system is full.

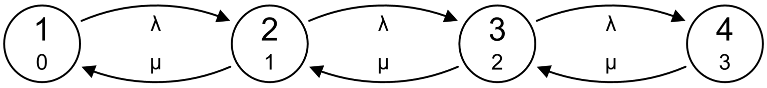


Figure 7.15.: M/M/1/S queueing system: Markov chain for the system state.

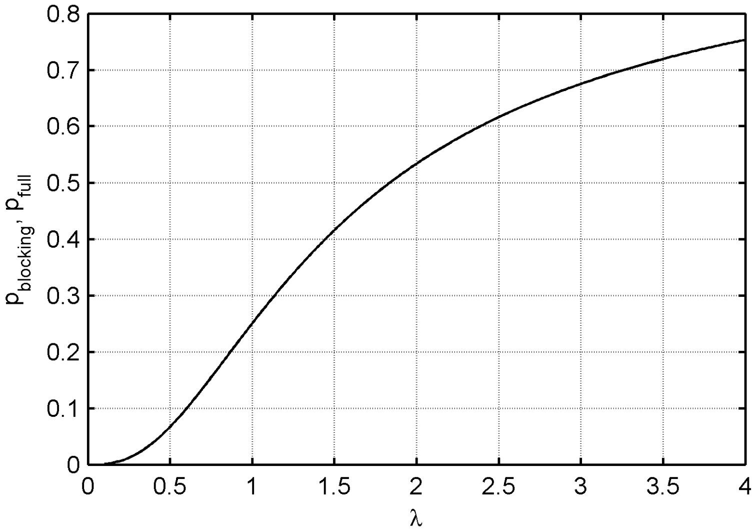


Figure 7.16.: M/M/1/S queueing system: the blocking probability equals the probability that the system is full. $S = 3$, arrival rate λ, service rate $\mu = 1$.

7.2.2. M/M/1/S queueing system with controlled arrival rate

Figure 7.17 shows the Markov chain for the system state of an M/M/1/S queueing system with controlled arrival rate (cf. Section 3.1). When the system is in a state $\langle\cdot/\mathrm{n}\rangle$, customers arrive at the normal arrival rate λ_n. When the system is in a state $\langle\cdot/\mathrm{r}\rangle$, customers arrive at the reduced arrival rate λ_r.

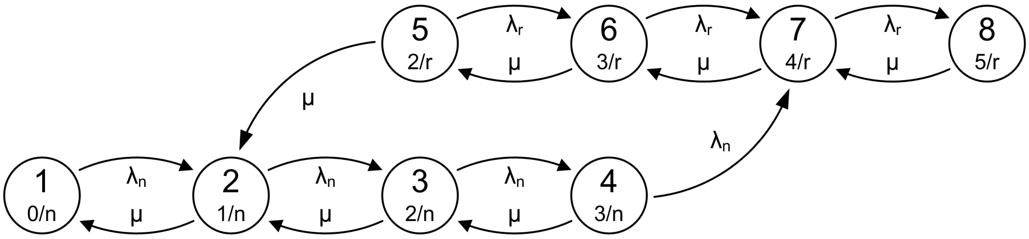


Figure 7.17.: M/M/1/S queueing system with controlled arrival rate: Markov chain for the system state. $S = 5, S_\mathsf{stop} = 4, S_\mathsf{go} = 2$. Meaning of the names of the states: number of customers in the system / “n” for normal arrival rate or “r” for reduced arrival rate.

Therefore, the total arrival rate is

$$\lambda_\text{total} = \sum_{k=0}^{3} \pi_{\langle k/\mathrm{n}\rangle}\lambda_\mathrm{n} + \sum_{k=2}^{5} \pi_{\langle k/\mathrm{r}\rangle}\lambda_\mathrm{r} \tag{7.63}$$

Rejections take place when the system is in state $\langle 5/\mathrm{r}\rangle$. The arrival rate in this state is λ_r, so the rejection rate is

$$\lambda_\text{rejected} = \pi_{\langle 5/\mathrm{r}\rangle}\lambda_\mathrm{r} \tag{7.64}$$

The blocking probability is

$$p_\text{blocking} = \frac{\pi_{\langle 5/\mathrm{r}\rangle}\lambda_\mathrm{r}}{\sum\limits_{k=0}^{3} \pi_{\langle k/\mathrm{n}\rangle}\lambda_\mathrm{n} + \sum\limits_{k=2}^{5} \pi_{\langle k/\mathrm{r}\rangle}\lambda_\mathrm{r}} \tag{7.65}$$

7.2.3. Hypo/M/1/S queueing system

Figure 7.18 shows the Markov chain for the system state of a Hypo/M/1/S queueing system with $S = 3$. The total arrival rate in this queueing system is

$$\lambda_{\text{total}} = \lambda = \lambda_2 \sum_{k=0}^{3} \pi_{\langle k/2 \rangle} \tag{7.66}$$

The rejection rate is

$$\lambda_{\text{rejected}} = \lambda_2 \pi_{\langle 3/2 \rangle} \tag{7.67}$$

Therefore, the blocking probability is

$$p_{\text{blocking}} = \frac{\pi_{\langle 3/2 \rangle}}{\sum_{k=0}^{3} \pi_{\langle k/2 \rangle}} \tag{7.68}$$

whereas the probability that the system is full is

$$p_{\text{full}} = \pi_{\langle 3/1 \rangle} + \pi_{\langle 3/2 \rangle} \tag{7.69}$$

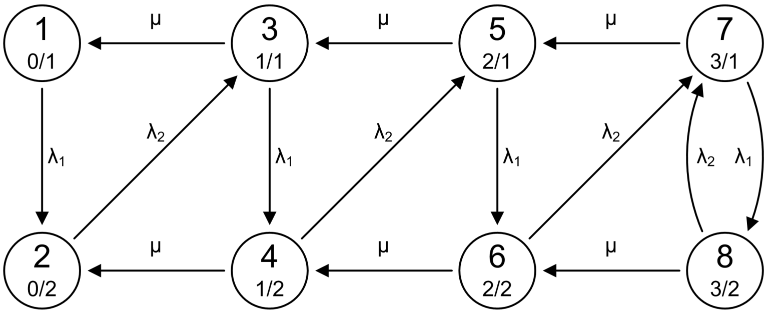


Figure 7.18.: Hypo/M/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the arrival process.

As can be seen in Figure 7.20, the blocking probability in a Hypo/M/1/S queueing system is lower than the probability that the system is full. The reason is that the probability that the interarrival time is small is lower than if we had Poisson arrivals, so that it is more likely that the server finishes a service between two arrivals. In other words, the customers tend to arrive at “advantageous” points in time, where the probability of finding a full system is smaller than it would be if they arrived at arbitrary points in time.

7.2.4. Hyper/M/1/S queueing system

Figure 7.19 shows the Markov chain for the system state of a Hyper/M/1/S queueing system with $S = 3$. In this queueing system, the blocking probability is calculated as follows: The total arrival rate is

$$\lambda_{\text{total}} = \lambda = \lambda_1 \sum_{k=0}^{3} \pi_{\langle k/1\rangle} + \lambda_2 \sum_{k=0}^{3} \pi_{\langle k/2\rangle} \tag{7.70}$$

The rejection rate is

$$\lambda_{\text{rejected}} = \lambda_1 \pi_{\langle 3/1\rangle} + \lambda_2 \pi_{\langle 3/2\rangle} \tag{7.71}$$

Therefore, the blocking probability is

$$p_{\text{blocking}} = \frac{\lambda_1 \pi_{\langle 3/1\rangle} + \lambda_2 \pi_{\langle 3/2\rangle}}{\lambda_1 \sum\limits_{k=0}^{3} \pi_{\langle k/1\rangle} + \lambda_2 \sum\limits_{k=0}^{3} \pi_{\langle k/2\rangle}} \tag{7.72}$$

whereas the probability that the system is full is

$$p_{\text{full}} = \pi_{\langle 3/1\rangle} + \pi_{\langle 3/2\rangle} \tag{7.73}$$

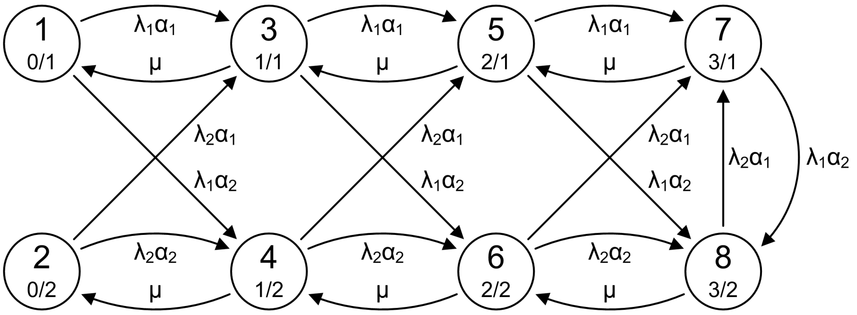


Figure 7.19.: Hyper/M/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the arrival process.

As can be seen in Figure 7.21, the blocking probability in a Hyper/M/1/S queueing system is higher than the probability that the system is full. The reason is that the probability that the interarrival time is small is higher than if we had Poisson arrivals, so that it is less likely that the server finishes a service between two arrivals. In other words, the customers tend to arrive at "disadvantageous" points in time, where the probability of finding a full system is higher than it would be if they arrived at arbitrary points in time.

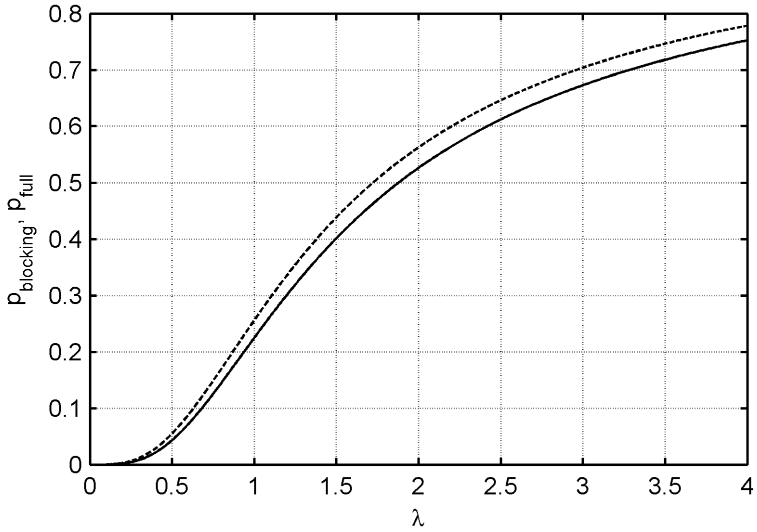


Figure 7.20.: Hypo/M/1/S queueing system: blocking probability (solid line) and probability that the system is full (dashed line)

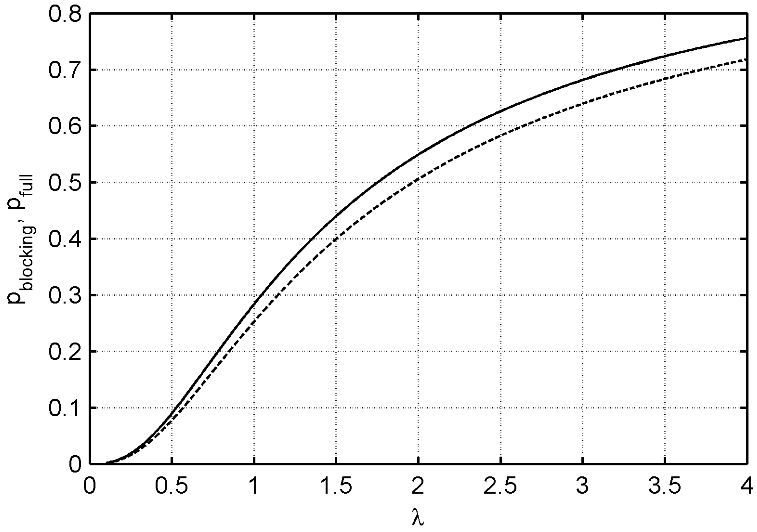


Figure 7.21.: Hyper/M/1/S queueing system: blocking probability (solid line) and probability that the system is full (dashed line)

7.3. Successful arrivals between two consecutive overflows

The probability distribution of the number of successful arrivals between two consecutive rejections ζ is determined by extending the Markov chain $\mathcal{M}_R$ we used in the previous section for the calculation of the time to the next rejection with a Markov chain $\mathcal{M}_C$ that counts successful arrivals.

We start the observation at the moment when a rejection has taken place. Since there has not been a successful arrival yet, the new Markov chain $\mathcal{M}_\zeta$ is in a state $\langle i/0 \rangle$, where $\langle i \rangle$ is a state of $\mathcal{M}_R$. The probabilities $\sigma^\zeta_{\langle i/0 \rangle}$ that it is in state $\langle i/0 \rangle$ are

$$\sigma^\zeta_{\langle i/0 \rangle} = \sigma^R_{\langle i \rangle} \tag{7.74}$$

As time goes by and $\mathcal{M}_\zeta$ evolves, the counting Markov chain $\mathcal{M}_C$ increases its value when there are successful arrivals, until eventually $\mathcal{M}_\zeta$ reaches an absorbing state $\langle \mathrm{R}/n \rangle, n \in \mathbb{N}$. This means that another rejection has taken place, and we are interested in the number of counted successful arrivals at this moment.

So we have

$$\pi_{\langle i/0 \rangle}(0) = \sigma^R_{\langle i \rangle} \tag{7.75}$$

$$\pi'(\tau) = \pi(\tau) \cdot \mathcal{Q} \tag{7.76}$$

If we had an infinite Markov chain $\mathcal{M}_C$, the probability that there are n successful arrivals between two consecutive rejections would be

$$\mathrm{P}\left(\zeta = n\right) = \lim_{t \to \infty} \pi_{\langle \mathrm{R}/n \rangle}(t) \tag{7.77}$$

However, since we do not deal with infinite Markov chains, $\mathcal{M}_C$ is finite. Assume it counts from 0 to N. In this case, we have

$$\mathrm{P}\left(\zeta = n\right) = \lim_{t \to \infty} \pi_{\langle \mathrm{R}/n \rangle}(t) \qquad n < N \tag{7.78}$$

$$\mathrm{P}\left(\zeta \geq N\right) = \lim_{t \to \infty} \pi_{\langle \mathrm{R}/N \rangle}(t) \tag{7.79}$$

In practice, the computation of $\pi(t)$ can be stopped when the Markov chain is with very high probability in a state $\langle \mathrm{R}/n \rangle$, that is, when $\sum_{n=0}^{N} \pi_{\langle \mathrm{R}/n \rangle}(t) \approx 1$.

Unless N is very large, in many cases the probability $\mathrm{P}\left(\zeta \geq N\right)$ has a non-negligible value (cf. Table 7.1). Therefore, we should avoid calculating the expected number of successful arrivals based on the obtained probabilities.

If the arrival rate is constant, we should instead calculate the expected number of customers that arrive during an interblocking time B and subtract 1 for the last (blocked) arrival:

$$\mathrm{E}\left(\zeta\right) = B\lambda - 1 \tag{7.80}$$

7.3.1. M/M/1/S queueing system

Figure 7.22 shows the construction of the Markov chain for the calculation of the number of successful arrivals between two overflows (Figure 7.22c) from the Markov chain for the calculation of the time to the next rejection (Figure 7.22a) and a counting Markov chain (Figure 7.22b).

After a rejection, the system is full and there have been no successful arrivals since the last rejection, so the Markov chain $\mathcal{M}_\zeta$ is in state $\langle 3/0\rangle$:

$$\pi_i(0) = \begin{cases} 1 & i = \langle 3/0\rangle \\ 0 & \text{otherwise} \end{cases} \tag{7.81}$$

Now we calculate the state probabilities with

$$\pi'(\tau) = \pi(\tau) \cdot \mathcal{Q} \tag{7.82}$$

until, for example,

$$1 - \sum_{n=0}^{N} \pi_{\langle \mathrm{R}/n\rangle}(\tau) < 10^{-6} \tag{7.83}$$

The probability distribution of ζ is

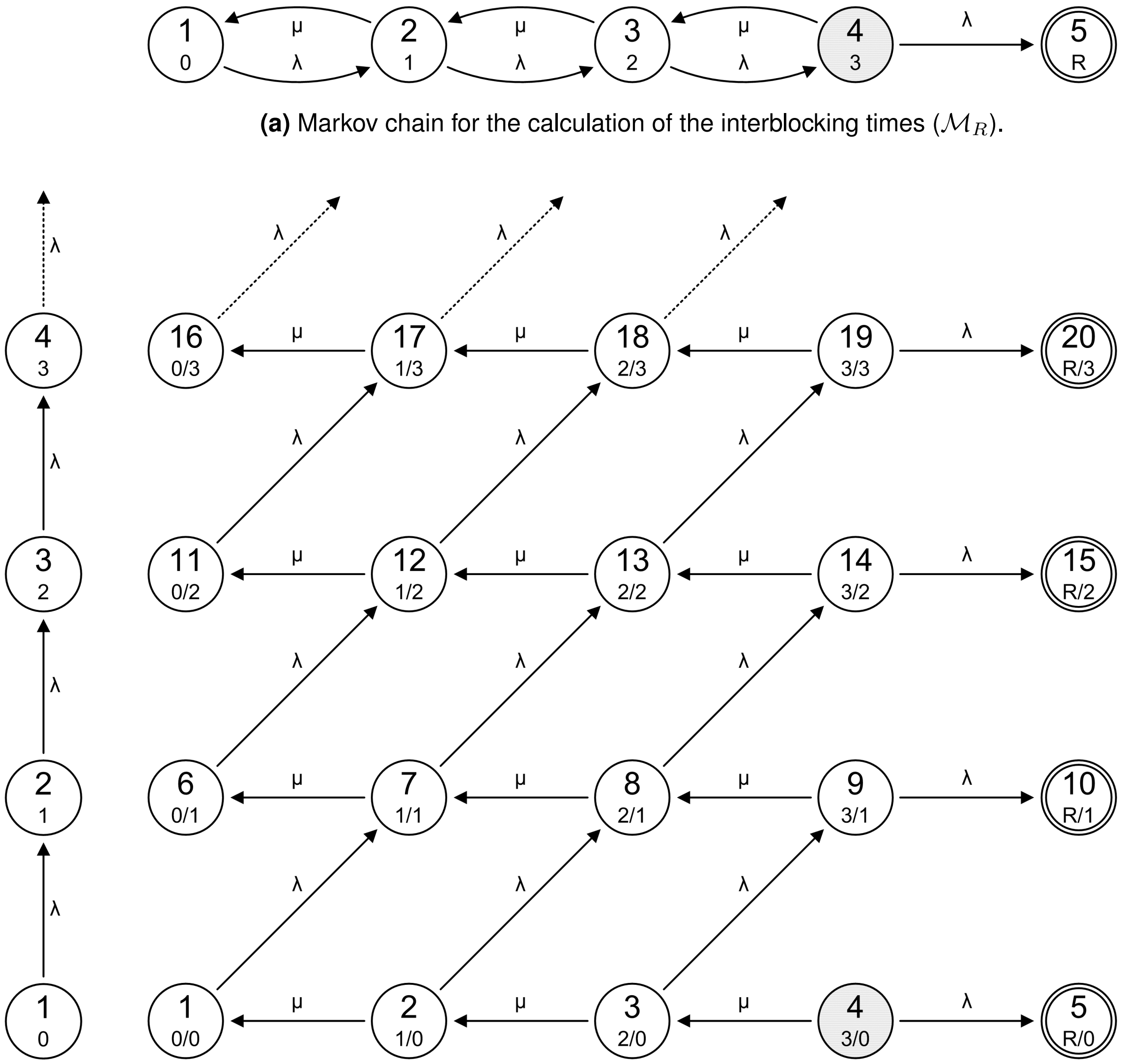
$$\mathrm{P}\left(\zeta = n\right) = \lim_{t\to\infty} \pi_{\langle \mathrm{R}/n\rangle}(t) \quad n < N \tag{7.84}$$

$$\mathrm{P}\left(\zeta \geq N\right) = \lim_{t\to\infty} \pi_{\langle \mathrm{R}/N\rangle}(t) \tag{7.85}$$

or

$$\mathrm{P}\left(\zeta \geq n\right) = \sum_{i=n}^{N} \lim_{t\to\infty} \pi_{\langle \mathrm{R}/n\rangle}(t) \quad n \leq N \tag{7.86}$$

Figures 7.23 to 7.26 show some results.



(a) Markov chain for the calculation of the interblocking times ($\mathcal{M}_R$).

(b) Counting Markov chain ($\mathcal{M}_C$).

(c) Markov chain for the calculation of the number of successful arrivals between two consecutive overflows ($\mathcal{M}_\varsigma$). Meaning of the names of the states: number of customers in the system or "R" if the next rejection has taken place / number of counted successful arrivals. Absorbing states are painted with double lines, states in which the system is after an overflow are shaded grey.

Figure 7.22.: M/M/1/S queueing system: calculation of the number of successful arrivals between two consecutive overflows

$\lambda = 1$

N	$\mathrm{P}(\zeta > N)$	$\mathrm{E}(\zeta)$
10	0.0876	1.53
20	0.0180	2.52
50	$1.55 \cdot 10^{-4}$	2.99
100	$5.67 \cdot 10^{-8}$	3
500	≈ 0	3
1000	≈ 0	3
∞	0	3

$\lambda = 0.5$

N	$\mathrm{P}(\zeta > N)$	$\mathrm{E}(\zeta)$
10	0.390	1.20
20	0.249	3.33
50	0.0650	9.27
100	0.00691	13.2
500	≈ 0	14
1000	≈ 0	14
∞	0	14

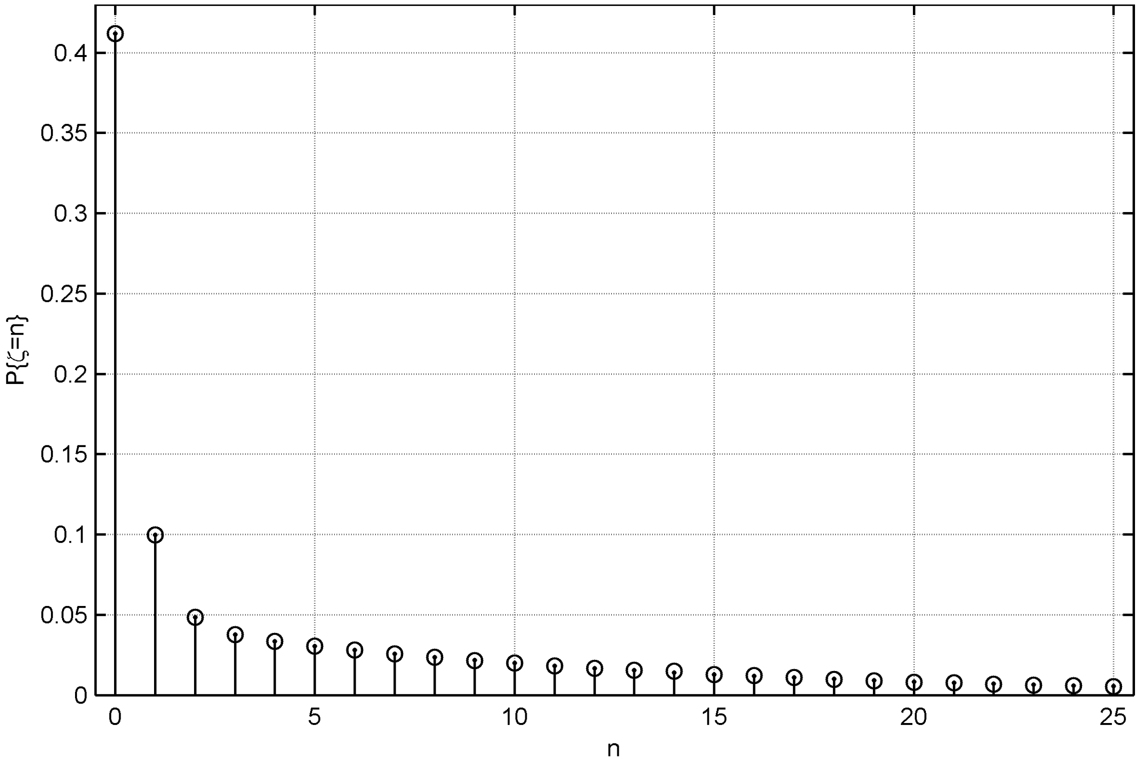
$\lambda = 0.2$

N	$\mathrm{P}(\zeta > N)$	$\mathrm{E}(\zeta)$
10	0.770	0.251
20	0.731	0.860
50	0.624	4.60
100	0.480	15.3
500	0.0587	114
1000	0.00425	150
∞	0	155

$\lambda = 0.1$

N	$\mathrm{P}(\zeta > N)$	$\mathrm{E}(\zeta)$
10	0.894	0.0480
20	0.887	0.160
50	0.866	0.918
100	0.831	3.51
500	0.601	70.4
1000	0.4	217
∞	0	1110

Table 7.1.: Successful arrivals between two overflows in an M/M/1/S queueing system with $S = 3$, arrival rate λ and service rate $\mu = 1$. The calculation of the mean is done based on the known probability distribution of ζ (which depends on N).



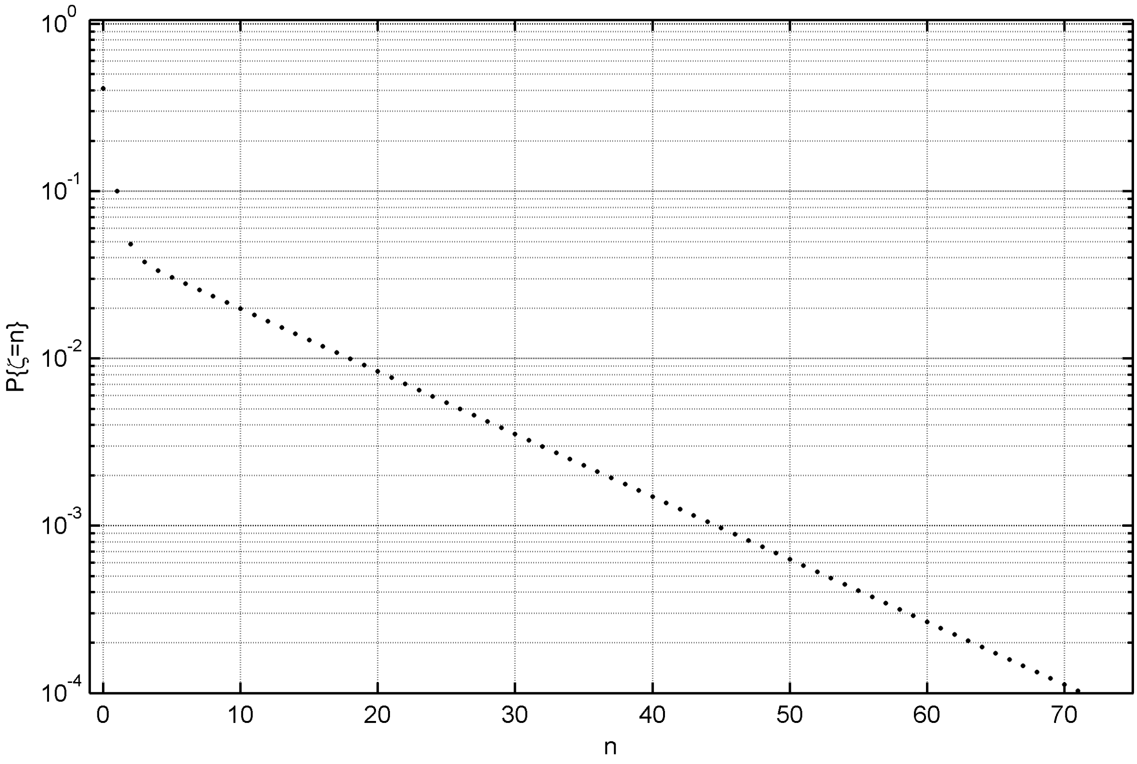
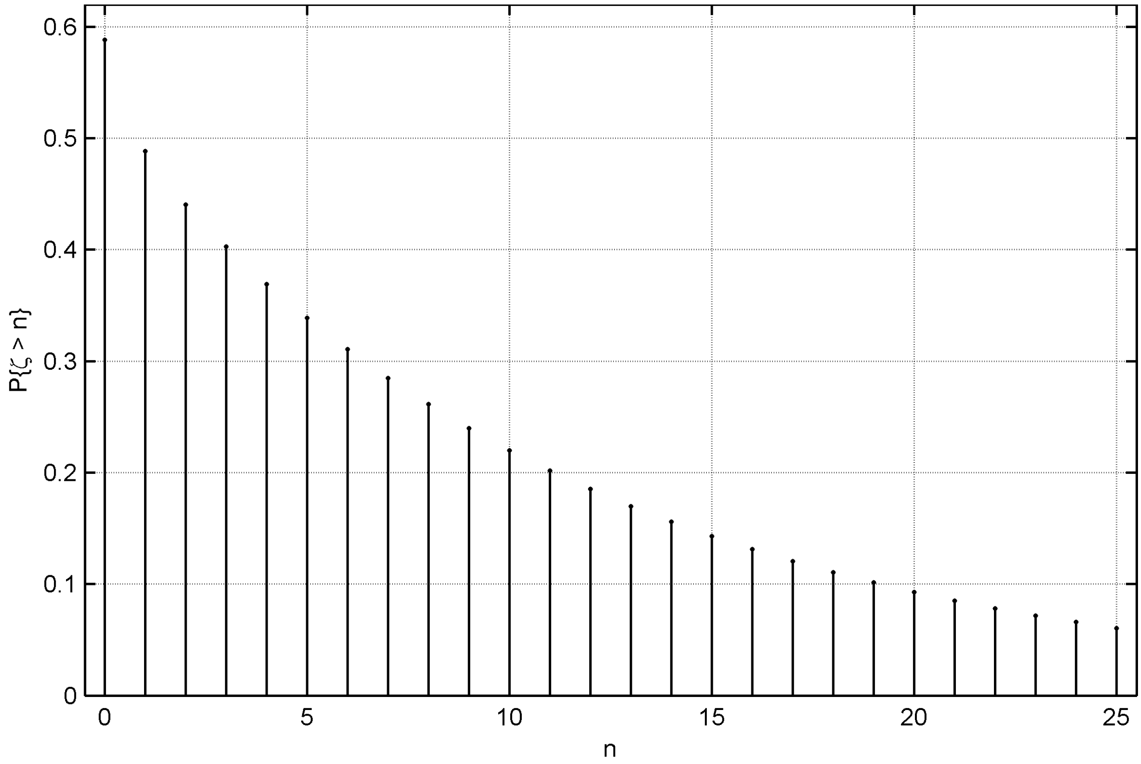


Figure 7.23.: M/M/1/S queueing system: number of successful arrivals between two consecutive overflows. $S = 3$, arrival rate $\lambda = 0.7$, service rate $\mu = 1$.
Dots: calculation, circles: simulation.



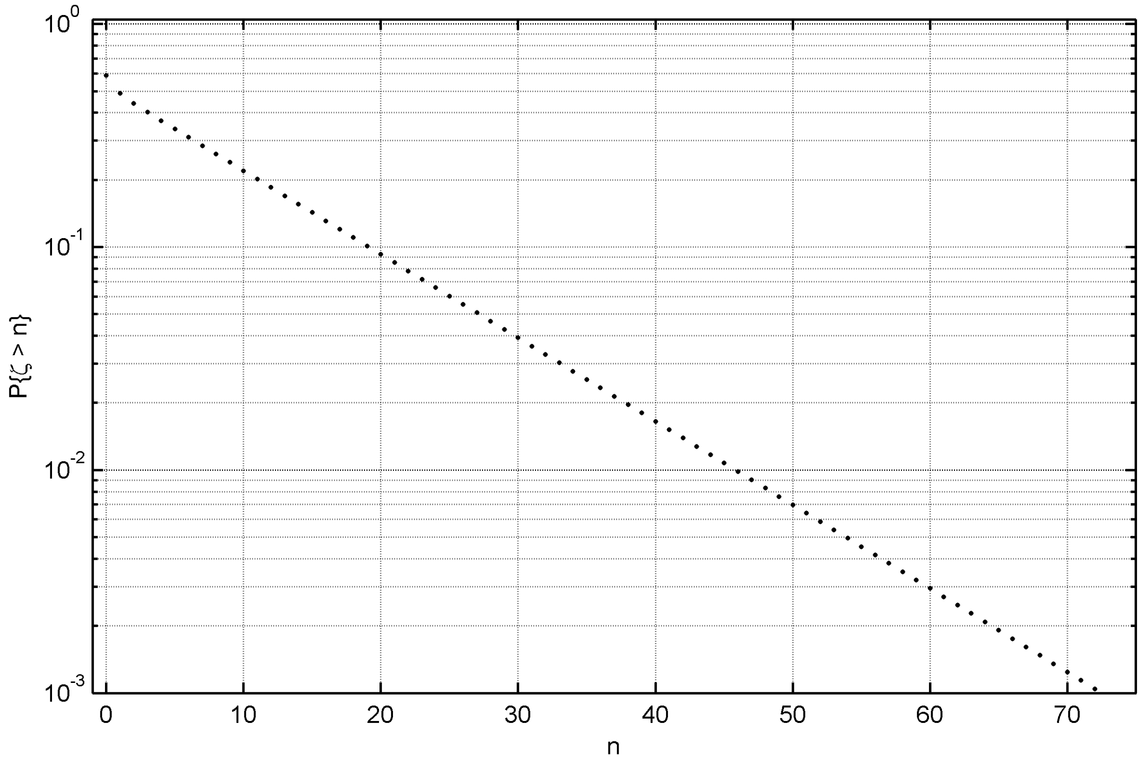


Figure 7.24.: M/M/1/S queueing system: number of successful arrivals between two consecutive overflows. $S = 3$, arrival rate $\lambda = 0.7$, service rate $\mu = 1$.

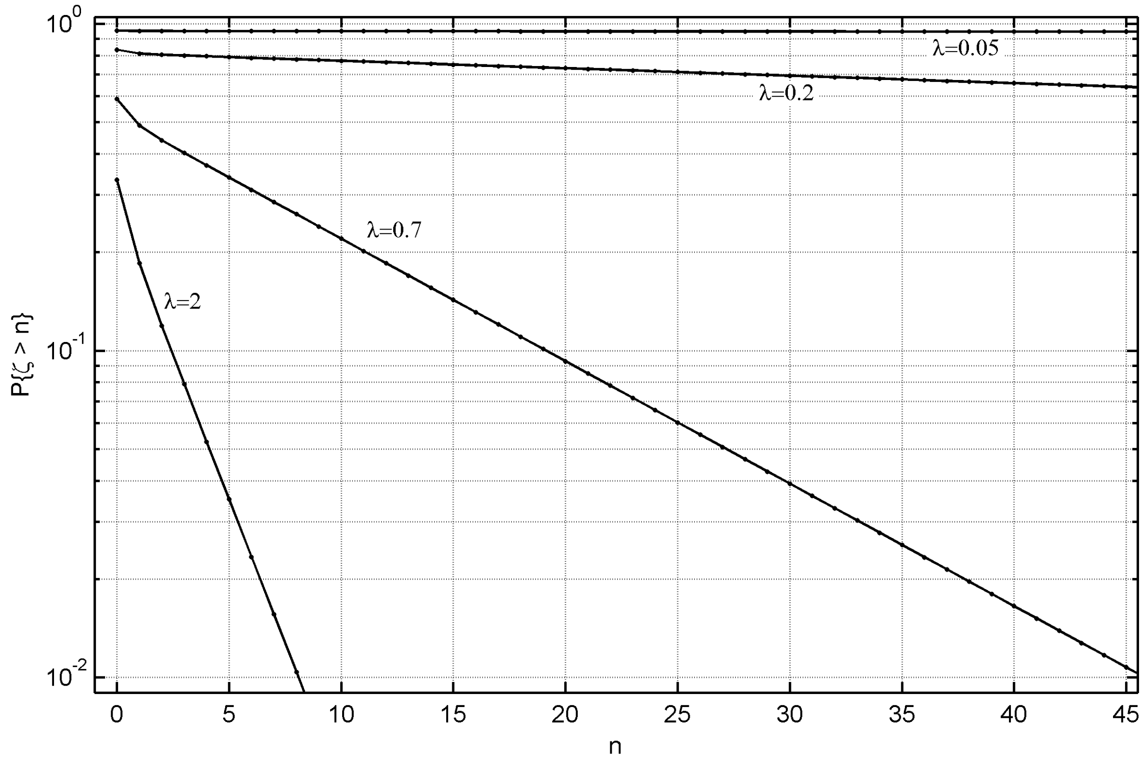


Figure 7.25.: M/M/1/S queueing system: number of successful arrivals between two consecutive overflows for various arrival rates. $S = 3$, arrival rate λ, service rate $\mu = 1$.

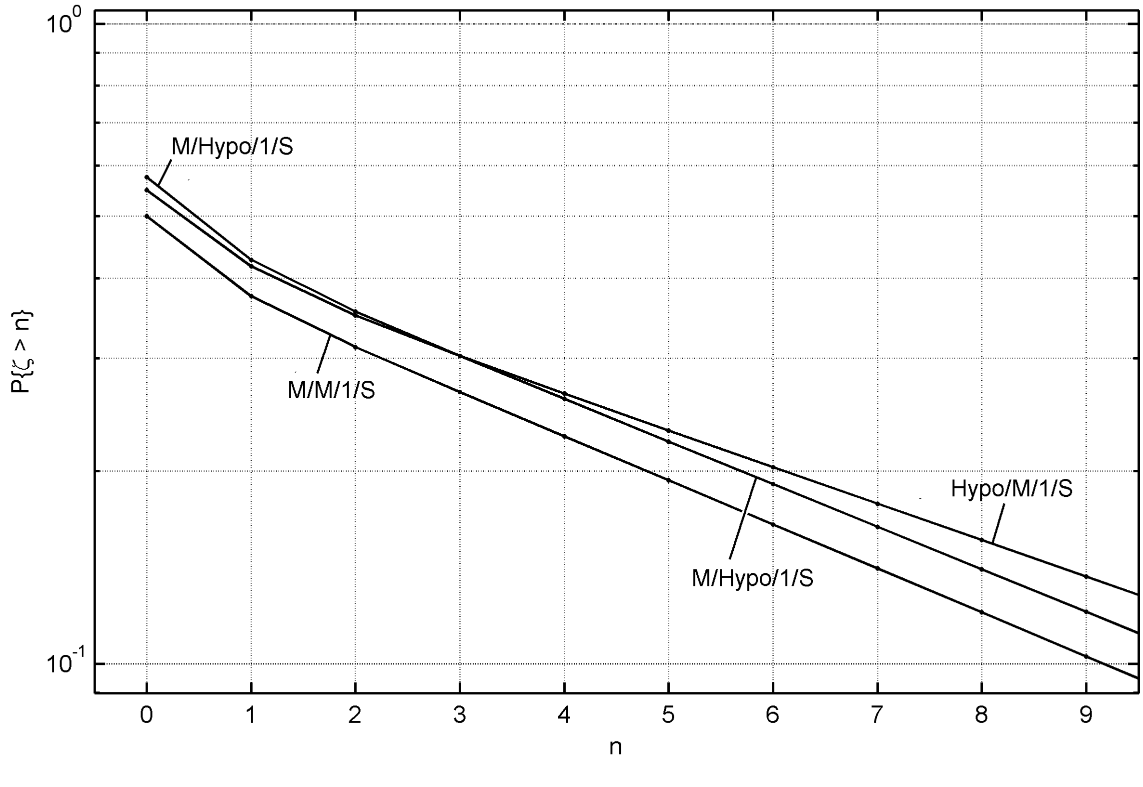


Figure 7.26.: Number of successful arrivals between two consecutive overflows in an M/M/1/S queueing system, a Hypo/M/1/S queueing system ($c_A = 0.75$) and an M/Hypo/1/S queueing system ($c_S = 0.75$). All: $S = 3$, arrival rate $\lambda = 1$, service rate $\mu = 1$.

7.3.2. Hypo/M/1/S queueing system

Figure 7.27 shows the Markov chain for the calculation of the number of successful arrivals between two rejections in a Hypo/M/1/S queueing system.

After a rejection, the system is full, there have been no successful arrivals since the last rejection, and the arrival process is in its initial state. Therefore, the Markov chain is in state $\langle 3/1/0 \rangle$:

$$\pi_i(0) = \begin{cases} 1 & i = \langle 3/1/0 \rangle \\ 0 & \text{otherwise} \end{cases} \tag{7.87}$$

Now we calculate the state probabilities with

$$\pi'(\tau) = \pi(\tau) \cdot \mathcal{Q} \tag{7.88}$$

until, for example,

$$1 - \sum_{n=0}^{N} \pi_{\langle \mathrm{R}/n \rangle}(\tau) < 10^{-6} \tag{7.89}$$

The probability distribution of ζ is

$$\mathrm{P}(\zeta = n) = \lim_{t \to \infty} \pi_{\langle \mathrm{R}/n \rangle}(t) \qquad n < N \tag{7.90}$$

$$\mathrm{P}(\zeta \geq N) = \lim_{t \to \infty} \pi_{\langle \mathrm{R}/N \rangle}(t) \tag{7.91}$$

or

$$\mathrm{P}(\zeta \geq n) = \sum_{i=n}^{N} \lim_{t \to \infty} \pi_{\langle \mathrm{R}/n \rangle}(t) \qquad n \leq N \tag{7.92}$$

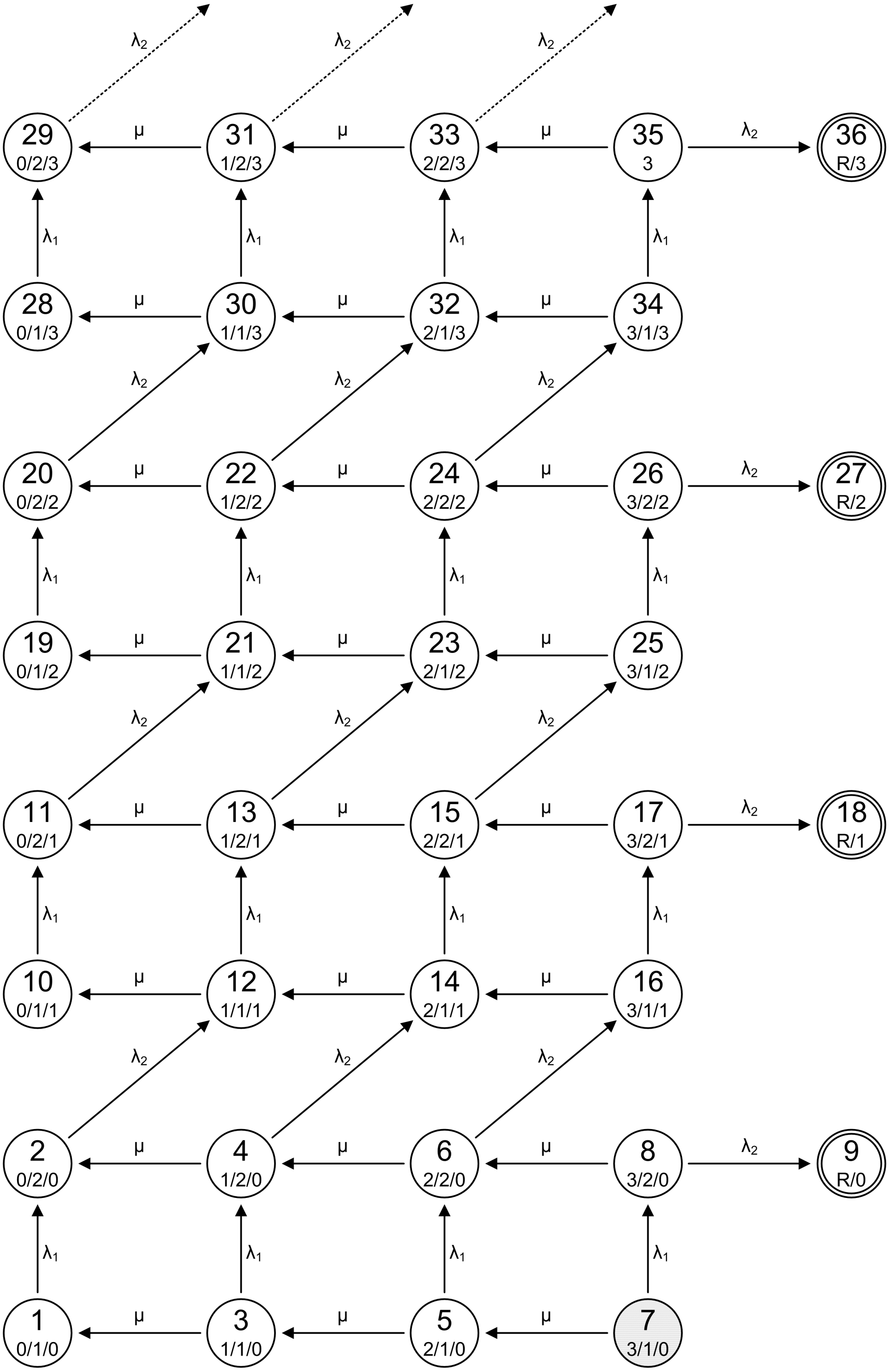


Figure 7.27.: Hypo/M/1/S queueing system: calculation of the number of successful arrivals between two consecutive overflows. Meaning of the names of the states: $\langle i/j/k \rangle$: number of customers in the system / state of the arrival process / number of counted successful arrivals; $\langle \mathrm{R}/k \rangle$: the next overflow has taken place / number of counted successful arrivals.

7.3.3. M/Hypo/1/S queueing system

Figure 7.28 shows the Markov chain for the calculation of the number of successful arrivals between two rejections in an M/Hypo/1/S queueing system.

After a rejection, the system is full and there have been no successful arrivals since the last rejection. The service process is not affected by rejections, so after a rejection the Markov chain is in state $\langle 3/1/0\rangle$ or in state $\langle 3/2/0\rangle$.

$$\pi_{\langle 3/1/0\rangle}(0) = \sigma^R_{\langle 3/1\rangle} = \frac{\pi_{\langle 3/2\rangle}}{\pi_{\langle 3/1\rangle} + \pi_{\langle 3/2\rangle}} \tag{7.93}$$

$$\pi_{\langle 3/2/0\rangle}(0) = \sigma^R_{\langle 3/2\rangle} = \frac{\pi_{\langle 3/2\rangle}}{\pi_{\langle 3/1\rangle} + \pi_{\langle 3/2\rangle}} \tag{7.94}$$

Now we calculate the state probabilities with

$$\pi'(\tau) = \pi(\tau) \cdot \mathcal{Q} \tag{7.95}$$

until, for example,

$$1 - \sum_{n=0}^{N} \pi_{\langle \mathrm{R}/n\rangle}(\tau) < 10^{-6} \tag{7.96}$$

The probability distribution of ζ is

$$\mathrm{P}\left(\zeta = n\right) = \lim_{t\to\infty} \pi_{\langle \mathrm{R}/n\rangle}(t) \quad n < N \tag{7.97}$$

$$\mathrm{P}\left(\zeta \geq N\right) = \lim_{t\to\infty} \pi_{\langle \mathrm{R}/N\rangle}(t) \tag{7.98}$$

or

$$\mathrm{P}\left(\zeta \geq n\right) = \sum_{i=n}^{N} \lim_{t\to\infty} \pi_{\langle \mathrm{R}/n\rangle}(t) \quad n \leq N \tag{7.99}$$

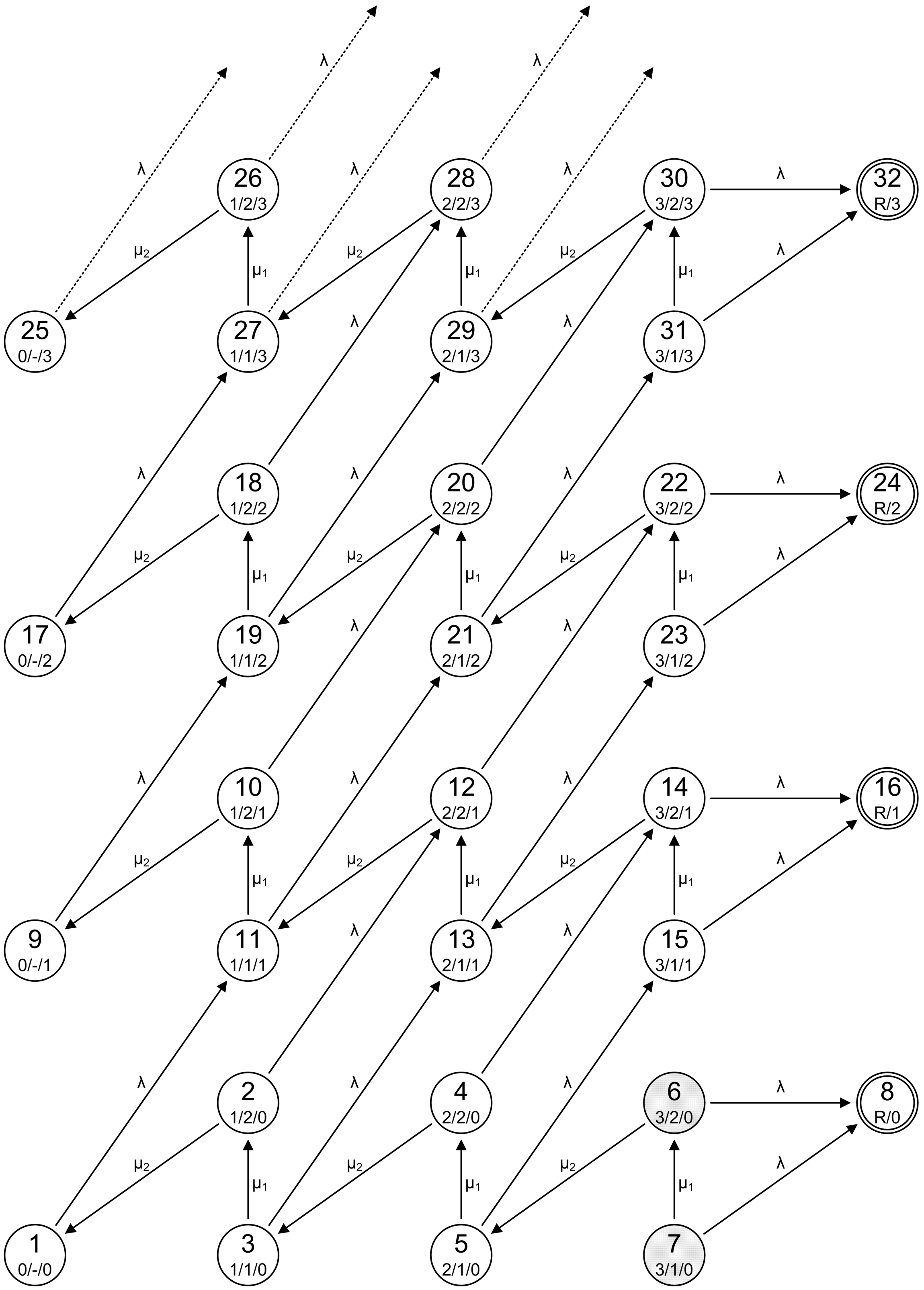


Figure 7.28.: M/Hypo/1/S queueing system: calculation of the number of successful arrivals between two consecutive overflows. Meaning of the names of the states: $\langle i/j/k \rangle$: number of customers in the system / state of the service process / number of counted successful arrivals; $\langle \mathrm{R}/k \rangle$: the next overflow has taken place / number of counted successful arrivals.

7.4. Modelling an overflow tandem system

We consider the queueing network shown in Figure 7.29. The network consists of two GI/M/1/S queueing systems. Customers arrive according to a Poisson process at system 1. If system 1 is full, they are redirected to system 2. If system 2 is full when such a customer arrives, the customer is discarded.

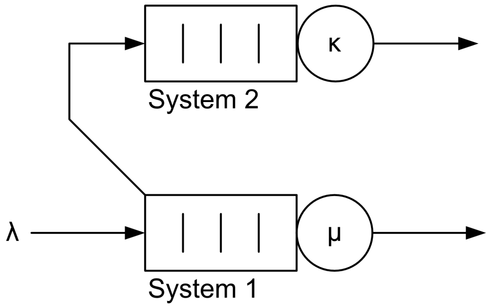


Figure 7.29.: Overflow tandem system.

We are interested in the number of customers X_2 in system 2 and in the interdeparture times D_2 of system 2. (The number of customers in system 1 and the interdeparture times of system 1 are independent of the fact that rejected customers are redirected to another system, so they are calculated as shown in Sections 3.1 and 6.1.)

The Markov chain for the system state of the network is shown in Figure 7.30.

Arrivals at system 1 (rate λ) increase the number of customers in system 1. Services in system 1 (rate μ) decrease the number of customers in system 1. If system 1 is full when a customer arrives (states $\langle 3/\cdot \rangle$), the number of customers in system 2 is increased. Services in system 2 (rate κ) decrease the number of customers in system 2.

For the calculation of the number of customers in the queueing systems, we calculate the stationary system state probabilities π by solving the system of linear equations

$$\pi \cdot \mathcal{Q} = 0 \tag{7.100}$$

$$\sum_i \pi_i = 1 \tag{7.101}$$

The number of customers in the first system X_1 is

$$\mathrm{P}\{X_1 = i\} = \sum_{k=0}^{3} \pi_{\langle i/k \rangle} \tag{7.102}$$

$$\mathrm{E}(X_1) = \sum_{i=1}^{3} i \sum_{k=0}^{3} \pi_{\langle i/k \rangle} \tag{7.103}$$

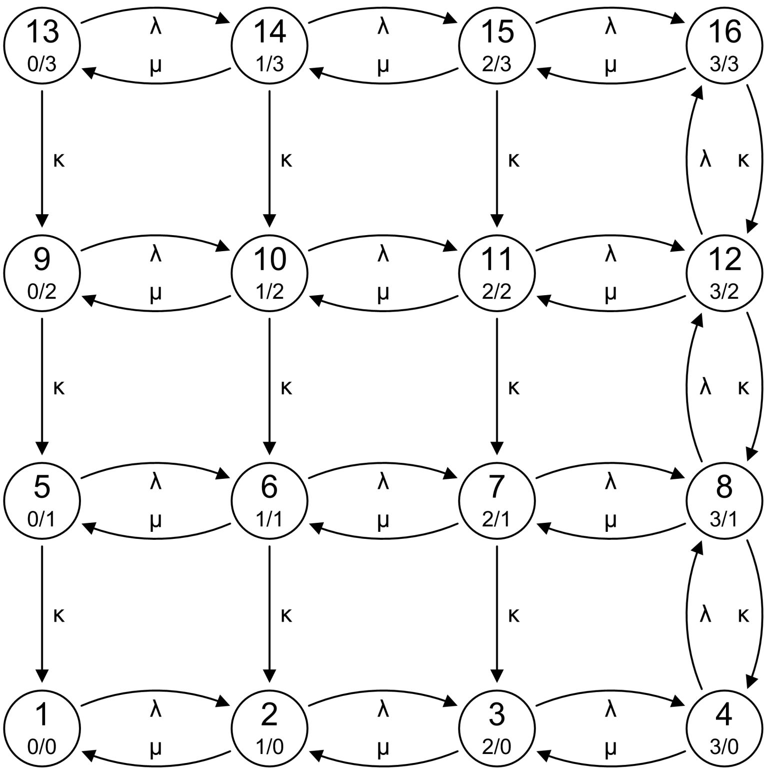


Figure 7.30.: Overflow tandem: Markov chain for the system state. Meaning of the names of the states: number of customers in system 1 / number of customers in system 2.

The number of customers in the second system X_2 is

$$\mathrm{P}\left\{X_2 = i\right\} = \sum_{k=0}^{3} \pi_{\langle k/i\rangle} \tag{7.104}$$

$$\mathrm{E}(X_2) = \sum_{i=1}^{3} i \sum_{k=0}^{3} \pi_{\langle k/i\rangle} \tag{7.105}$$

As can be seen in Figure 7.31, we cannot express the length of the idle period of system 2 as a sum of hypoexponentially distributed random variables. Therefore, the length of the first interdeparture time in a busy cycle in the second system $D_2^{(1)}$ is calculated with the Markov chain shown in Figure 7.32. This Markov chain is constructed from the Markov chain $\mathcal{M}_I$ for the calculation of the length of the idle period (Figure 7.31) by adding a transition with rate κ – that corresponds to the first service of a customer in system 2 – originating in the state in which the Markov chain is when the busy period begins (transition $\langle 3/1\rangle \rightarrow \langle \mathrm{S}\rangle$).

Now we calculate the time needed to go from the states in which the Markov chain can be when the idle period begins to reach the busy period (state $\langle 3/1\rangle$) and to traverse the added transition $\langle 3/1\rangle \rightarrow \langle \mathrm{S}\rangle$.

The Markov chain can be in state 1, 2, 3 or 4, when the idle period begins, depending on the state in which it was before the last customer in system 2 was served. The probabilities σ_i^I that the Markov chain is in state $i, i = 1 \dots 4$, when the idle period begins are

$$\sigma_{\langle 0/0\rangle}^I = \sigma_1^I = \pi_{\langle 0/1\rangle} \left(\sum_{k=0}^{3} \pi_{\langle k/1\rangle}\right)^{-1} = \pi_5 \left(\sum_{k=5}^{8} \pi_k\right)^{-1} \tag{7.106}$$

$$\sigma_{\langle 1/0\rangle}^I = \sigma_2^I = \pi_{\langle 1/1\rangle} \left(\sum_{k=0}^{3} \pi_{\langle k/1\rangle}\right)^{-1} = \pi_6 \left(\sum_{k=5}^{8} \pi_k\right)^{-1} \tag{7.107}$$

$$\sigma_{\langle 2/0\rangle}^I = \sigma_3^I = \pi_{\langle 2/1\rangle} \left(\sum_{k=0}^{3} \pi_{\langle k/1\rangle}\right)^{-1} = \pi_7 \left(\sum_{k=5}^{8} \pi_k\right)^{-1} \tag{7.108}$$

$$\sigma_{\langle 3/0\rangle}^I = \sigma_4^I = \pi_{\langle 3/1\rangle} \left(\sum_{k=0}^{3} \pi_{\langle k/1\rangle}\right)^{-1} = \pi_8 \left(\sum_{k=5}^{8} \pi_k\right)^{-1} \tag{7.109}$$

The time the Markov chain needs to go from states $1 \dots 4$ to state $\langle \mathrm{S}\rangle$ is calculated with

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \qquad\qquad \varphi_k(0) = \begin{cases} 0 & k = \langle \mathrm{S}\rangle \\ 1 & \text{otherwise} \end{cases} \tag{7.110}$$

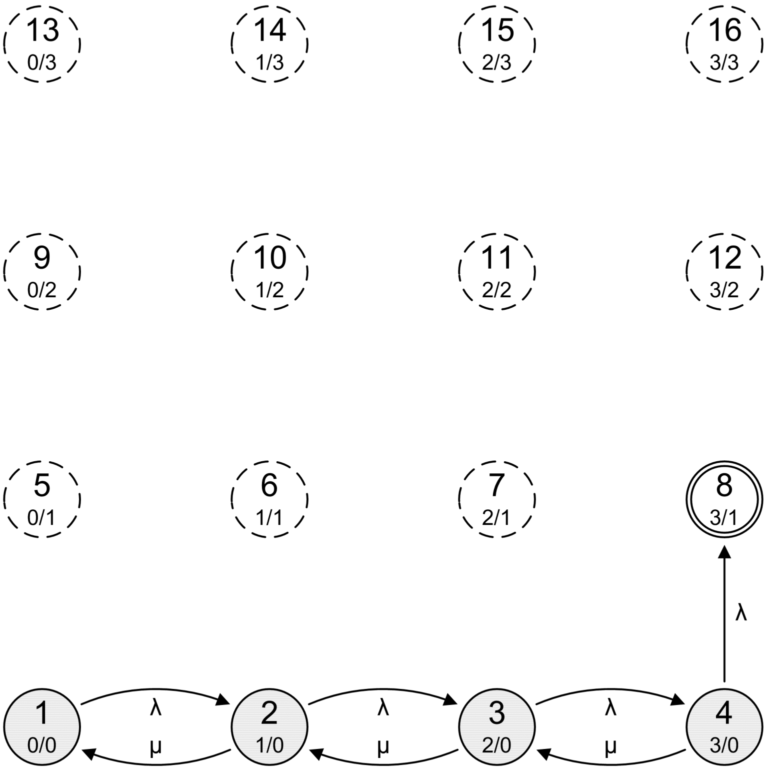


Figure 7.31.: Overflow tandem system: Markov chain for the calculation of the length of the idle period. Meaning of the names of the states: number of customers in system 1 / number of customers in system 2.

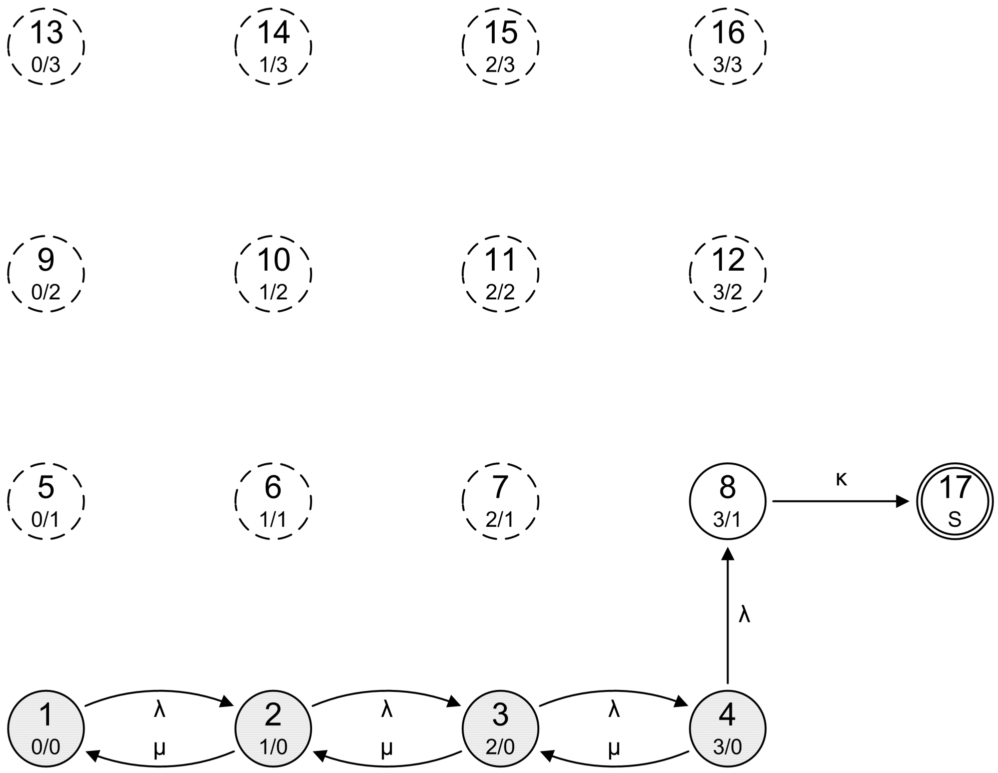


Figure 7.32.: Overflow tandem system: Markov chain for the calculation of the length of the first interdeparture time in a busy cycle of system 2. Meaning of the names of the states: number of customers in system 1 / number of customers in system 2.

Now $D_2^{(1)}$ is

$$\mathrm{P}\left\{D_2^{(1)} > t\right\} = \sigma^I_{\langle 0/0\rangle}\varphi_{\langle 0/0\rangle}(t) + \sigma^I_{\langle 1/0\rangle}\varphi_{\langle 1/0\rangle}(t) + \sigma^I_{\langle 2/0\rangle}\varphi_{\langle 2/0\rangle}(t) + \sigma^I_{\langle 3/0\rangle}\varphi_{\langle 3/0\rangle}(t) \quad (7.111)$$

$$\mathrm{E}(D_2^{(1)}) = \int\limits_0^\infty \mathrm{P}\left\{D_2^{(1)} > t\right\} \mathrm{d}t \quad (7.112)$$

The length of the busy period of system 2 is calculated with the Markov chain shown in Figure 7.33.

The busy period begins if system 2 is empty, system 1 is full (state $\langle 3/0\rangle$) and a customer arrives at system 1 (state $\langle 3/1\rangle$). It ends when system 2 becomes idle again (states $\langle \cdot/0\rangle$).

Therefore, the length B_2 of the busy period of system 2 is calculated as follows:

$$\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \qquad \varphi_{\langle i/j\rangle}(0) = \begin{cases} 1 & j \geq 1 \\ 0 & j = 0 \end{cases} \quad (7.113)$$

and

$$\mathrm{P}\left\{B_2 > t\right\} = \varphi_{\langle 3/1\rangle}(t) \quad (7.114)$$

$$\mathrm{E}(B_2) = \int\limits_0^\infty \mathrm{P}\left\{B_2 > t\right\} \mathrm{d}t \quad (7.115)$$

The number of customers served during a busy period is

$$\xi = \kappa \cdot \mathrm{E}(B_2) \quad (7.116)$$

Finally, the interdeparture time of system 2 is

$$D_2 \sim \frac{D_2^{(1)} + (\xi - 1)\,\mathrm{Exp}(\kappa)}{\xi} \quad (7.117)$$

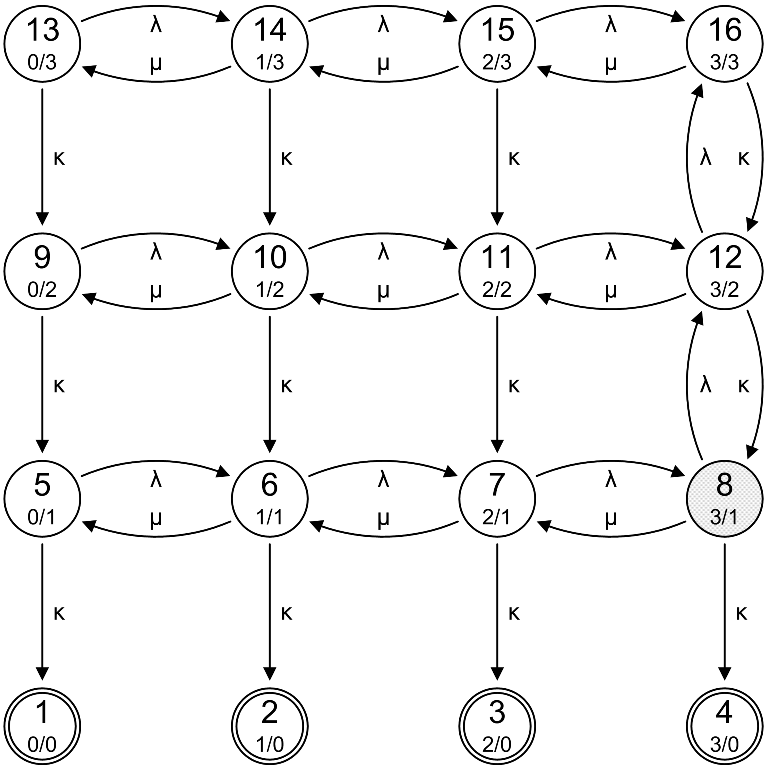


Figure 7.33.: Overflow tandem system: Markov chain for the calculation of the length of the busy period. Meaning of the names of the states: number of customers in system 1 / number of customers in system 2.

7.5. Modelling an overflow tandem system using network decomposition

As can be seen in Figure 7.34, the overflow stream of an M/M/1/S queueing system consitutes an Interrupted Poisson Process[1] (IPP): When the system is full, arriving customers (which arrive according to a Poisson process) are rejected, so the IPP is in state *"on"*. When the system is not full, arriving customers are not rejected, so the IPP is in state *"off"*.

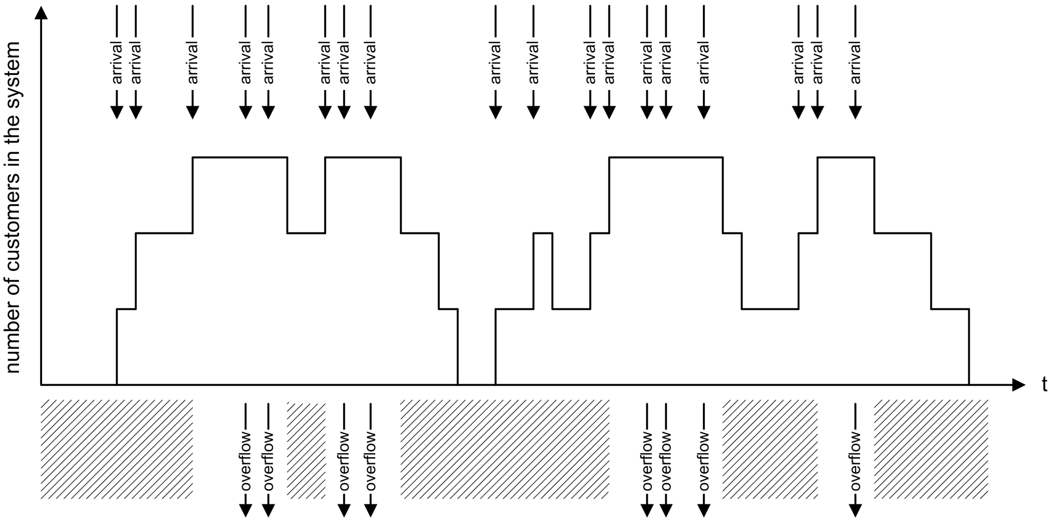


Figure 7.34.: The overflow stream of an M/M/1/S queueing system constitutes an IPP. When the system is full $(S = 3)$, the IPP is in state *"on"*, otherwise (shaded areas) it is in state *"off"*.

Therefore, if we want to use network decomposition to analyse the network shown in Figure 7.29, we could use the following approach (Figure 7.36):

We assume the overflow stream to be an IPP with exponentially distributed *"on"* and *"off"* times. The state of such an IPP can be modelled with the Markov chain shown in Figure 7.35.

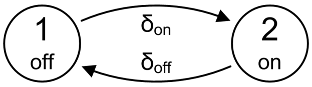


Figure 7.35.: Markov chain for the state of an IPP.

[1] An IPP is a process that can be in two states: *"off"* and *"on"*. When the process is in state *"off"*, it does not create any events. When the process is in state *"on"*, it behaves like a Poisson process.

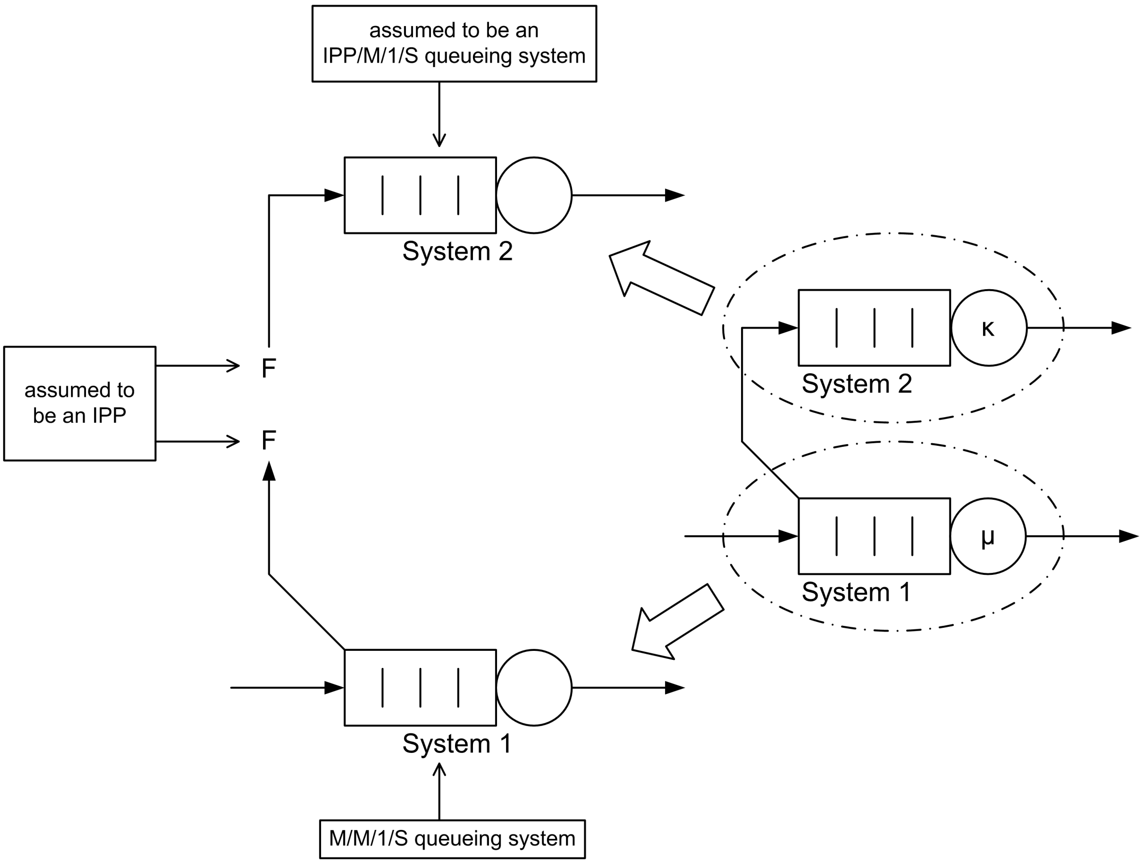


Figure 7.36.: Overflow tandem: network decomposition.

The rate δ_{on} at which the IPP goes from state *"off"* to state *"on"* equals the rate at which system 1 goes from states $\langle 0 \rangle$, $\langle 1 \rangle$ and $\langle 2 \rangle$ (where customer are not redirected to system 2) to state $\langle 3 \rangle$ (where customers are redirected to system 2), given that it is in state $\langle 0 \rangle$, state $\langle 1 \rangle$ or state $\langle 2 \rangle$ (see Figure 7.37):

$$\delta_{\text{on}} = \frac{\pi_{\langle 2 \rangle}}{\pi_{\langle 0 \rangle} + \pi_{\langle 1 \rangle} + \pi_{\langle 2 \rangle}} \lambda \tag{7.118}$$

The rate δ_{off} at which the IPP goes from state *"on"* to state *"off"* equals the rate at which system 1 goes from state $\langle 3 \rangle$ to states $\langle 0 \rangle$, $\langle 1 \rangle$ and $\langle 2 \rangle$, given that it is in state $\langle 3 \rangle$:

$$\delta_{\text{off}} = \mu \tag{7.119}$$

Now system 2 can be modelled with the Markov chain shown in Figure 7.38. When the Markov chain is in a state $\langle \cdot/\text{on} \rangle$, system 1 is full and customers are redirected to system 2. In this case, the system behaves like an M/M/1/S queueing system. When the Markov chain is in a state $\langle \cdot/\text{off} \rangle$, system 1 is not full, therefore, there are no arrivals at system 2, but only services.

Figure 7.39 shows some results.

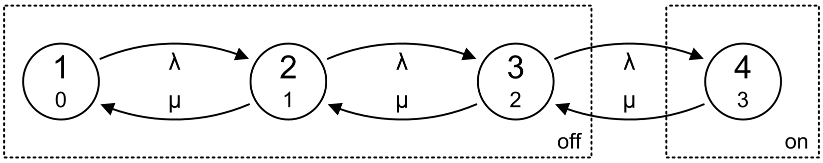


Figure 7.37.: Overflow tandem: Markov chain for the state of system 1. Meaning of the names of the states: number of customers in the system. Dashed boxes: state of the IPP of the overflow stream.

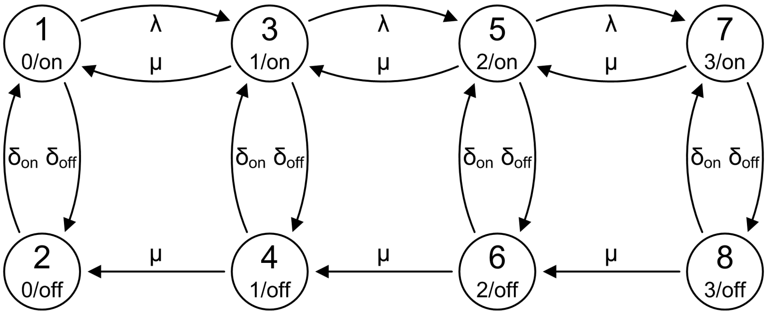


Figure 7.38.: Overflow tandem: Markov chain for the state of system 2 (IPP/M/1/S queueing system). Meaning of the names of the states: number of customers in the system / state of the IPP.

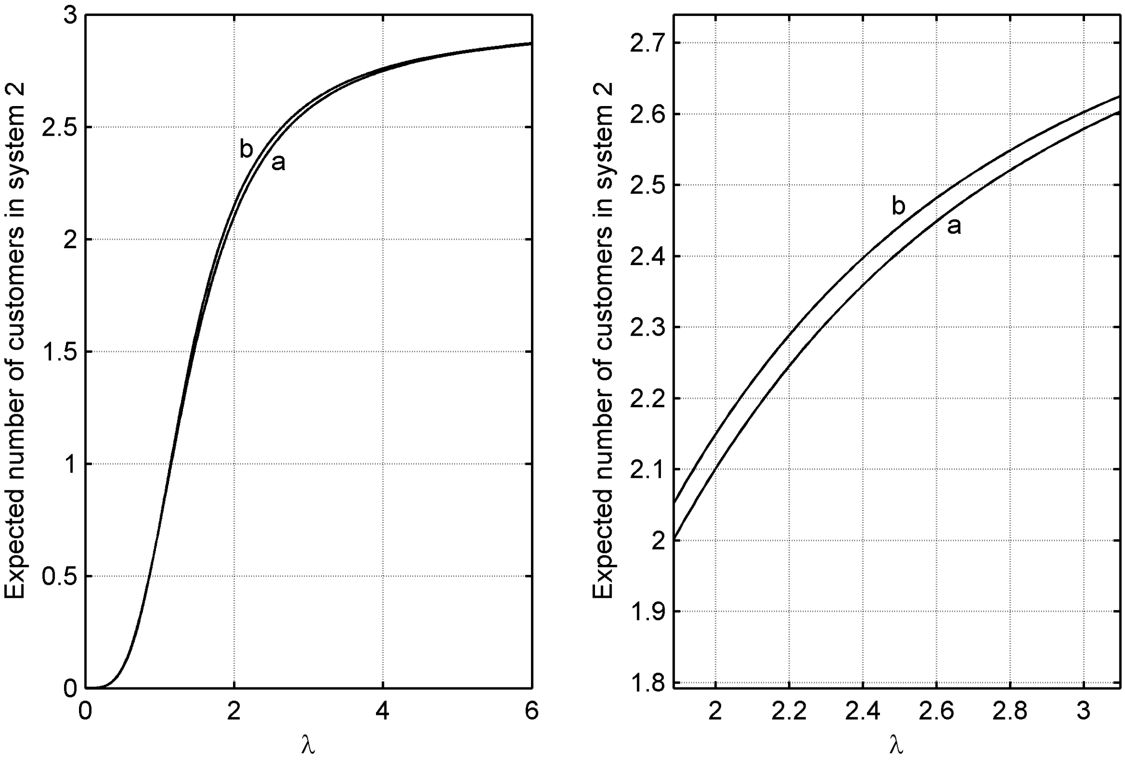


Figure 7.39.: Overflow tandem: number of customers in system 2. (a) exact results, (b) network decomposition (IPP). $S = 3$, the service rate of the system 1 is $\mu = 1$, the service rate of system 2 is $\kappa = 0.5$.

8. Superposition of traffic streams

In this chapter, we deal with the problem of superposing traffic streams (Figure 8.1).

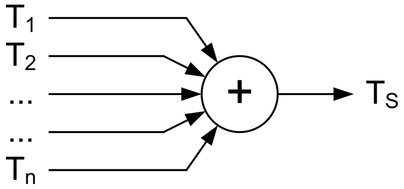


Figure 8.1.: Superposition of traffic streams.

Given n independent traffic streams $T_i = \langle T_{i1}, T_{i2}, \dots \rangle$ with independent and (within the same stream) identically distributed interevent times X_{ij} (that is, they constitute stationary renewal processes), the superposition T_S of these streams is defined as the union of the events of the component streams (Figure 8.2a),

$$T_S = \bigcup_{i=1}^{n} T_i \tag{8.1}$$

If we consider the associated counting processes $N_i(t) = \max\{0,\ j : T_{ij} \leq t\}$, the superposition N_S is the sum of the component counting processes (Figure 8.2b),

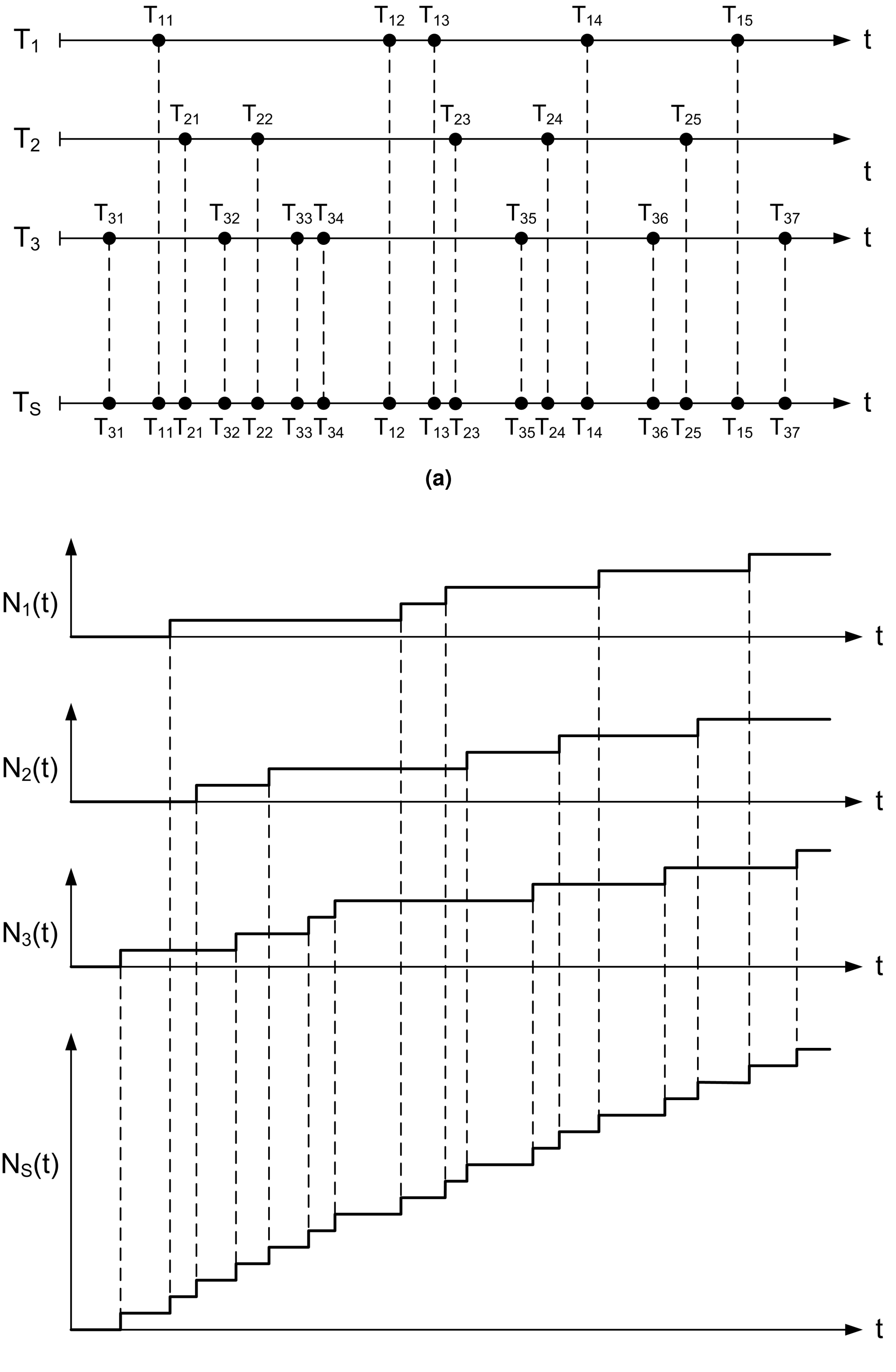
$$N_S(t) = \sum_{i=1}^{n} N_i(t) \tag{8.2}$$

We know the statistical characteristics of the component streams, and we are interested in the statistical characteristics of the resulting stream.

In Section 8.1, we show how the interevent times X_S of the superposition can be calculated.

Unless all component processes are Poisson processes, the superposition is not a renewal process any more. (This becomes obvious if we consider the superposition of two deterministic processes (Figure 8.3). In this case, each interevent time depends on its predecessor. If the current interevent time is δ, the next interevent time will be $1 - \delta$, and vice versa.) Therefore, in Section 8.2 we take a brief look at the dependencies between the interevent times.

In Section 8.3, we show how queueing systems with several input streams, which are superposed in the system, can be modelled.



(a)

(b)

Figure 8.2.: Superposition of 3 point processes. (a) Superposition as the union of the events of the component processes, (b) superposition as the sum of the component counting processes.

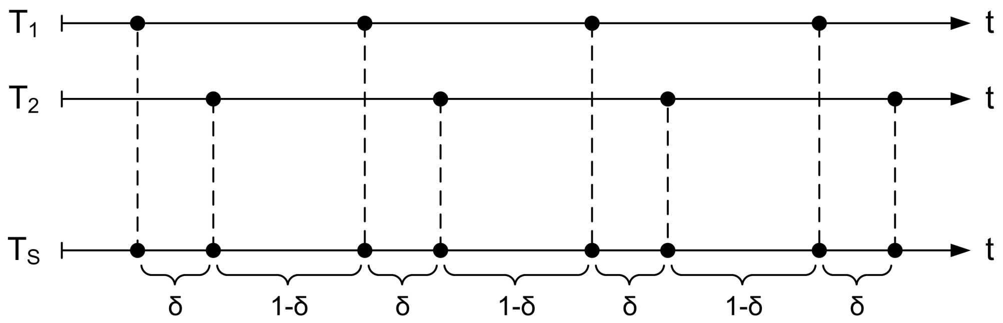


Figure 8.3.: Superposition of two deterministic processes.

To simplify matters, we mostly consider the superposition of two traffic streams. All techniques shown are extensible to an arbitrary number of streams. However, it is sufficient to be able to superpose two traffic streams, because the superposition of more than two streams can be achieved by combining several superpositions of two streams, as shown in Figure 8.4.

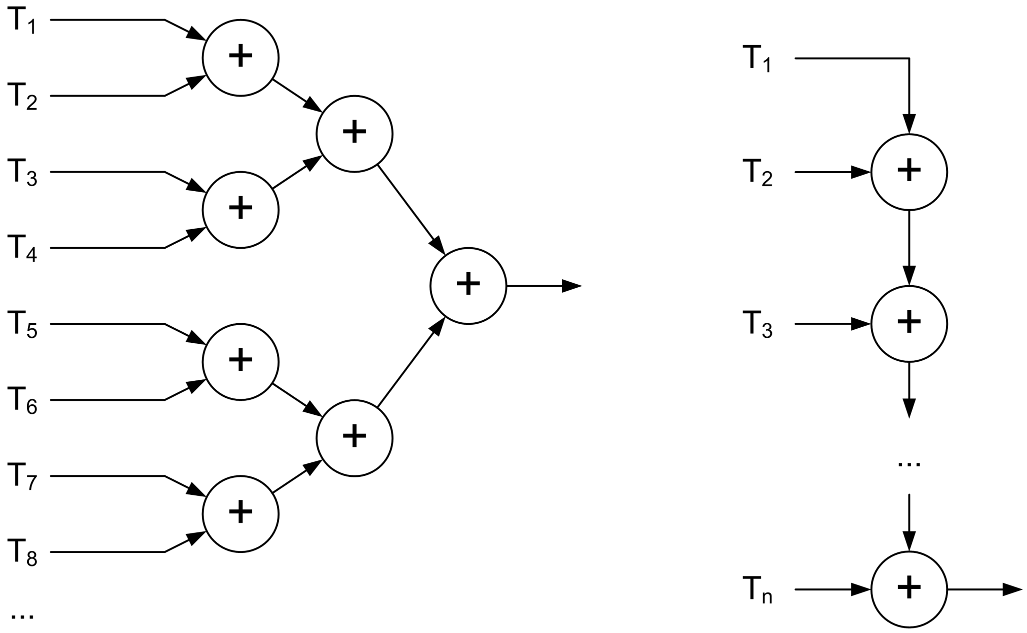


Figure 8.4.: Superposition of more than two traffic streams.

8.1. Interevent times

For the calculation of the probability distribution of the interevent times, we first combine the Markov chains that describe the state of the component processes (cf. Section 3.5) into a new Markov chain $\mathcal{M}_A$, which describes the state of the component processes as a whole. Transitions in the Markov chains for the state of the component processes that correspond to events keep this property in $\mathcal{M}_A$. That is, whenever there is an event in a component process, there is an event in the superposition, too.

Now we identify the probability σ_i^A that the Markov chain $\mathcal{M}_A$ is in state i after an event has occurred. This probability depends on the stationary state probabilities π immediately before the event and the probability that a certain stream has caused the event.

The probability P_j that an observed event has been created by component stream j equals the ratio of the rate of this stream to the total rate. If the rates of the component streams are $\lambda_1, \lambda_2, \ldots, \lambda_n$, we have

$$P_j = \frac{\lambda_j}{\sum\limits_{i=1}^{n} \lambda_i} \tag{8.3}$$

When stream i created an event, its phase-type distribution is in the initial state, whereas the states of the phase-type distributions of all other streams are unchanged.

Finally, we extend the Markov chain $\mathcal{M}_A$ by a state $\langle \mathrm{E} \rangle$, which is reached when an event occurs. All transitions that correspond to events are redirected to this new state. We calculate the complementary cumulative distribution function $\varphi_i(\cdot)$ of the time that the Markov chain needs to go from state i to state $\langle \mathrm{E} \rangle$ with

$$\varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{E} \rangle \\ 0 & i = \langle \mathrm{E} \rangle \end{cases} \tag{8.4}$$

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{8.5}$$

Now, the cumulative distribution function $F_A(\cdot)$ of the interevent time is

$$F_A(t) = 1 - \sum_i \sigma_i^A \varphi_i(t) \tag{8.6}$$

In the following examples, we use the following notation:

λ . . . rate of the first process
μ . . . rate of the second process

For hypoexponentially distributed interevent times:
λ_i . . . rates of the distribution of the first process
μ_i . . . rates of the distribution of the second process

For hyperexponentially or Coxian distributed interevent times:
λ_i, α_i . . . rates and probabilities of the distribution of the first process
μ_i, β_i . . . rates and probabilities of the distribution of the second process

8.1.1. Superposition of two streams with hypoexponentially distributed interevent times (Hypo+Hypo)

The construction of the Markov chain for the state of all streams ($\mathcal{M}_A$) is shown in Figure 8.5.

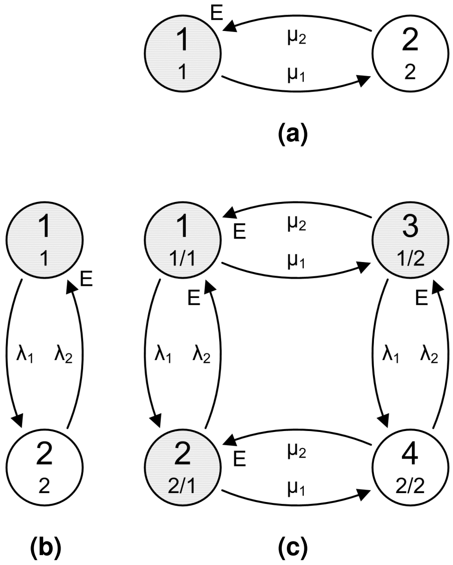


Figure 8.5.: Superposition of two streams with hypoexponentially distributed interevent times: Markov chains for the state of the component streams. (a) Second stream, (b) first stream, (c) both streams ($\mathcal{M}_A$).

The probability that an event in the superposition was caused by the first stream is $\lambda / (\lambda + \mu)$. In this case, the Markov chain is in states $\langle 1/1 \rangle$ or $\langle 1/2 \rangle$ immediately after the event. If it was in state $\langle 2/1 \rangle$ before the event occurred (the probability for this is $\pi_{\langle 2/1 \rangle} / \left(\pi_{\langle 2/1 \rangle} + \pi_{\langle 2/2 \rangle}\right)$), it is in state $\langle 1/1 \rangle$, if it was in state $\langle 2/2 \rangle$ (probability $\pi_{\langle 2/2 \rangle} / \left(\pi_{\langle 2/1 \rangle} + \pi_{\langle 2/2 \rangle}\right)$, it is in state $\langle 1/2 \rangle$.

The probability that an event in the superposition was caused by the second stream is $\mu / (\lambda + \mu)$. In this case, the Markov chain is in states $\langle 1/1 \rangle$ or $\langle 2/1 \rangle$ immediately after the event. If it was in state $\langle 1/2\rangle$ before the event occurred (the probability for this is $\pi_{\langle 1/2\rangle}/\left(\pi_{\langle 1/2\rangle}+\pi_{\langle 2/2\rangle}\right)$), it is in state $\langle 1/1\rangle$, if it was in state $\langle 2/2\rangle$ (probability $\pi_{\langle 2/2\rangle}/\left(\pi_{\langle 1/2\rangle}+\pi_{\langle 2/2\rangle}\right)$, it is in state $\langle 2/1\rangle$.

Therefore, we have

$$\begin{aligned}
\sigma^A_{\langle 1/1\rangle} &= \frac{\lambda}{\lambda+\mu}\frac{\pi_{\langle 2/1\rangle}}{\pi_{\langle 2/1\rangle}+\pi_{\langle 2/2\rangle}} + \frac{\mu}{\lambda+\mu}\frac{\pi_{\langle 1/2\rangle}}{\pi_{\langle 1/2\rangle}+\pi_{\langle 2/2\rangle}}\\
\sigma^A_{\langle 2/1\rangle} &= \frac{\mu}{\lambda+\mu}\frac{\pi_{\langle 2/2\rangle}}{\pi_{\langle 1/2\rangle}+\pi_{\langle 2/2\rangle}}\\
\sigma^A_{\langle 1/2\rangle} &= \frac{\lambda}{\lambda+\mu}\frac{\pi_{\langle 2/2\rangle}}{\pi_{\langle 2/1\rangle}+\pi_{\langle 2/2\rangle}}\\
\sigma^A_{\langle 2/2\rangle} &= 0
\end{aligned} \tag{8.7}$$

Adding the new state $\langle \mathrm{E}\rangle$ to the Markov chain for the state of all streams and redirecting all transitions that correspond to events to this state leads to the Markov chain shown in Figure 8.6.

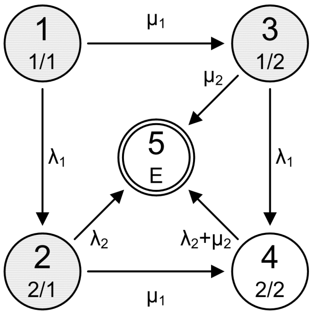


Figure 8.6.: Superposition of two streams with hypoexponentially distributed interevent times: Markov chain for the calculation of the time to the next event.

With this Markov chain, we calculate the complementary cumulative distribution function $\varphi_i(\cdot)$ of the time to the next event given that the Markov chain is in state i:

$$\varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{E}\rangle \\ 0 & i = \langle \mathrm{E}\rangle \end{cases} \tag{8.8}$$

$$\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \tag{8.9}$$

Now we calculate the complementary distribution function $F_A(\cdot)$ of the interevent times of the superposition with

$$F_A(t) = 1 - \sum_i \sigma^A_i\ \varphi_i(t) \tag{8.10}$$

8.1.2. Superposition of two streams with hyperexponentially distributed interevent times (Hyper+Hyper)

The construction of the Markov chain for the state of all streams ($\mathcal{M}_A$) is shown in Figure 8.7.

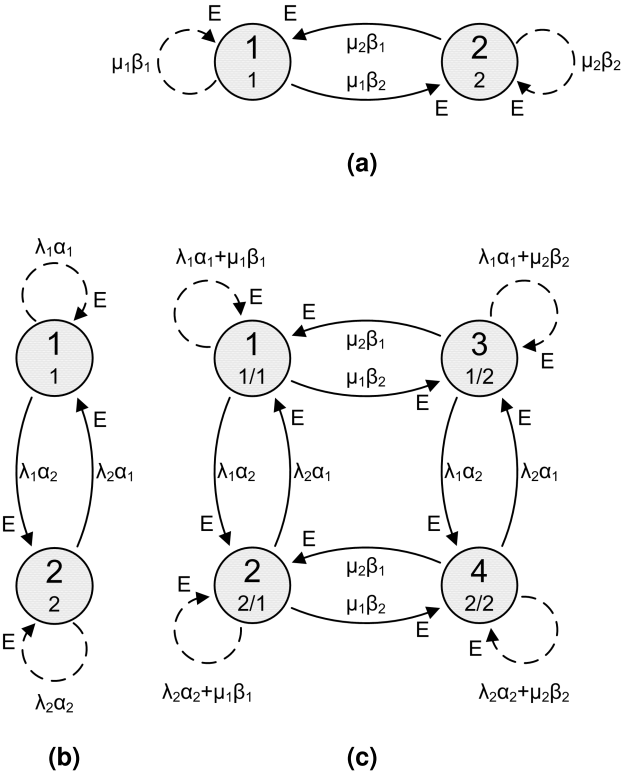


Figure 8.7.: Superposition of two streams with hyperexponentially distributed interevent times: Markov chain for the state of the component streams. (a) Second stream, (b) first stream, (c) both streams ($\mathcal{M}_A$). Hidden transitions are indicated with dashed lines.

The probability that an event in the superposition was caused by the first stream is $\lambda / (\lambda + \mu)$. In this case, immediately after the event the Markov chain is in states $\langle 1/1 \rangle$ or $\langle 1/2 \rangle$ with probability α_1 or in states $\langle 2/1 \rangle$ or $\langle 2/2 \rangle$ with probability α_2. If it was in state $\langle 1/1 \rangle$ or $\langle 2/1 \rangle$ before the event, it is in state $\langle 1/1 \rangle$ or $\langle 2/1 \rangle$ afterwards. If it was in state $\langle 1/2 \rangle$ or $\langle 2/2 \rangle$ before the event, it is in state $\langle 1/2 \rangle$ or $\langle 2/2 \rangle$ afterwards.

The probability that an event in the superposition was caused by the second stream is $\mu / (\lambda + \mu)$. In this case, immediately after the event the Markov chain is in states $\langle 1/1 \rangle$ or $\langle 2/1 \rangle$ with probability β_1 or in states $\langle 1/2 \rangle$ or $\langle 2/2 \rangle$ with probability β_2. If it was in state $\langle 1/1 \rangle$ or $\langle 1/2 \rangle$ before the event, it is in state $\langle 1/1 \rangle$ or $\langle 1/2 \rangle$ afterwards. If it was in state $\langle 2/1 \rangle$ or $\langle 2/2 \rangle$ before the event, it is in state $\langle 2/1 \rangle$ or $\langle 2/2 \rangle$ afterwards.

Therefore, we have

$$\begin{aligned}
\sigma^A_{\langle 1/1\rangle} &= \frac{\lambda}{\lambda+\mu}(\pi_{\langle 1/1\rangle}+\pi_{\langle 2/1\rangle})\alpha_1 + \frac{\mu}{\lambda+\mu}(\pi_{\langle 1/1\rangle}+\pi_{\langle 1/2\rangle})\beta_1\\
\sigma^A_{\langle 2/1\rangle} &= \frac{\lambda}{\lambda+\mu}(\pi_{\langle 1/1\rangle}+\pi_{\langle 2/1\rangle})\alpha_2 + \frac{\mu}{\lambda+\mu}(\pi_{\langle 2/1\rangle}+\pi_{\langle 2/2\rangle})\beta_1\\
\sigma^A_{\langle 1/2\rangle} &= \frac{\lambda}{\lambda+\mu}(\pi_{\langle 1/2\rangle}+\pi_{\langle 2/2\rangle})\alpha_1 + \frac{\mu}{\lambda+\mu}(\pi_{\langle 1/1\rangle}+\pi_{\langle 1/2\rangle})\beta_2\\
\sigma^A_{\langle 2/2\rangle} &= \frac{\lambda}{\lambda+\mu}(\pi_{\langle 1/2\rangle}+\pi_{\langle 2/2\rangle})\alpha_2 + \frac{\mu}{\lambda+\mu}(\pi_{\langle 2/1\rangle}+\pi_{\langle 2/2\rangle})\beta_2
\end{aligned} \tag{8.11}$$

Adding the new state $\langle \mathrm{E}\rangle$ to the Markov chain $\mathcal{M}_A$ and redirecting all transitions that correspond to events to this state leads to the Markov chain shown in Figure 8.7.

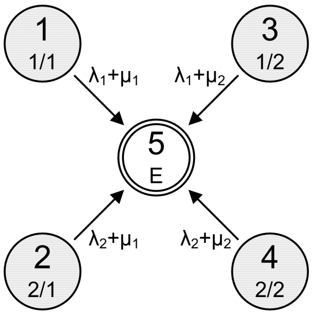


Figure 8.8.: Superposition of two streams with hyperexponentially distributed interevent times: Markov chain for the calculation of the time to the next event.

With this Markov chain, we calculate the complementary cumulative distribution function $\varphi_i(\cdot)$ of the time to the next event given that the Markov chain is in state i:

$$\varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{E}\rangle \\ 0 & i = \langle \mathrm{E}\rangle \end{cases} \tag{8.12}$$

$$\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \tag{8.13}$$

Now we calculate the complementary distribution function $F_A(\cdot)$ of the interevent times with

$$F_A(t) = 1 - \sum_i \sigma^A_i\ \varphi_i(t) \tag{8.14}$$

Results can be seen in Figure 8.9.

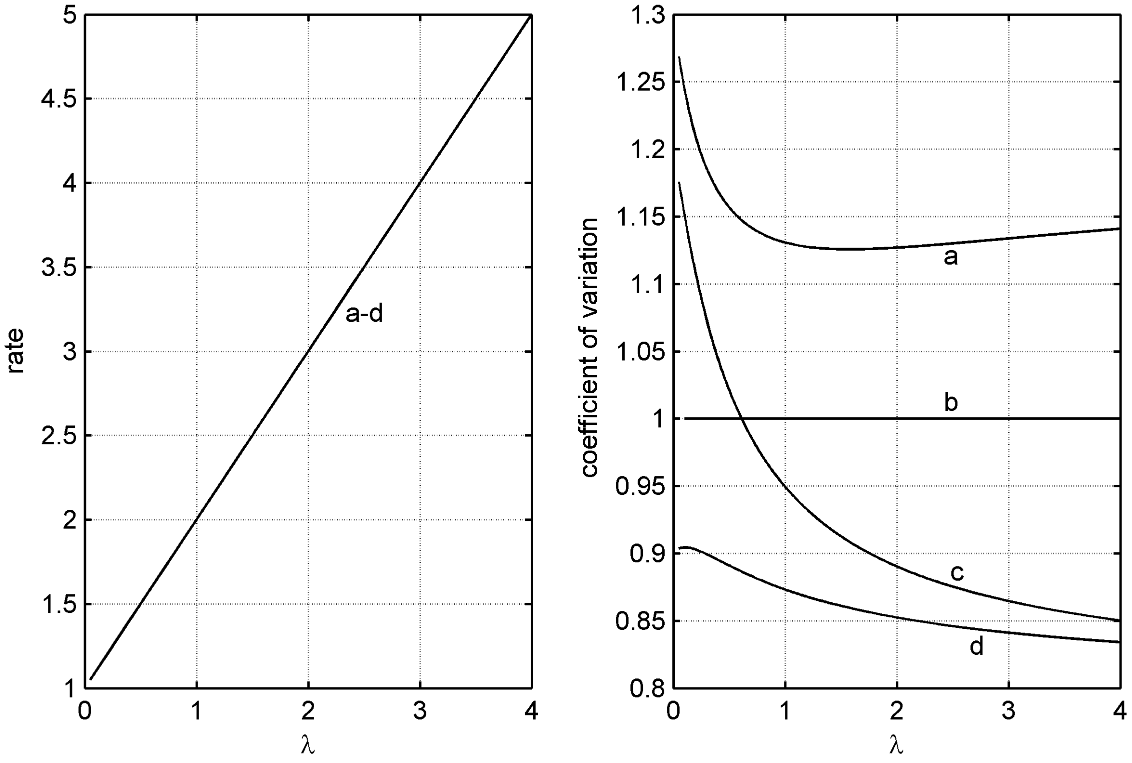


Figure 8.9.: Superposition of two traffic streams. (a) Hyper+Hyper: $r_1 = \lambda$, $c_1 = 1.2$, $r_2 = 1$, $c_2 = 1.3$ (b) M+M: $r_1 = \lambda$, $c_1 = 1$, $r_2 = 1$, $c_2 = 1$ (c) Hypo+Hyper: $r_1 = \lambda$, $c_1 = 0.8$, $r_2 = 1$, $c_2 = 1.1$ (d) Hypo+Hypo: $r_1 = \lambda$, $c_1 = 0.8$, $r_2 = 1$, $c_2 = 0.9$

8.1.3. Superposition of two streams with Coxian distributed interevent times (Cox+Cox)

The Markov chain for the state of the streams is shown in Figure 8.10.

The probability that an event in the superposition was caused by the first stream is $\lambda / (\lambda + \mu)$. In this case, immediately after the event the Markov chain is in state $\langle 1/1 \rangle$, state $\langle 1/2 \rangle$ or state $\langle 1/3 \rangle$. It is in state $\langle 1/x \rangle$ if it was in a state $\langle \cdot /x \rangle$ before the event occurred. The probability for this is $\sum_{i=1}^{3} \pi_{\langle i/x \rangle}$.

The probability that an event in the superposition was caused by the second stream is $\mu / (\lambda + \mu)$. In this case, immediately after the event the Markov chain is in state $\langle 1/1 \rangle$, state $\langle 2/1 \rangle$ or state $\langle 3/1 \rangle$. It is in state $\langle x/1 \rangle$ if it was in a state $\langle x/ \cdot \rangle$ before the event occurred. The probability for this is $\sum_{i=1}^{3} \pi_{\langle x/i \rangle}$.

Therefore, we have

$$\begin{aligned}
\sigma^A_{\langle 1/1\rangle} &= \frac{\lambda}{\lambda+\mu}(\pi_{\langle 1/1\rangle}+\pi_{\langle 2/1\rangle}+\pi_{\langle 3/1\rangle}) + \frac{\mu}{\lambda+\mu}(\pi_{\langle 1/1\rangle}+\pi_{\langle 1/2\rangle}+\pi_{\langle 1/3\rangle})\\
\sigma^A_{\langle 1/2\rangle} &= \frac{\lambda}{\lambda+\mu}(\pi_{\langle 1/2\rangle}+\pi_{\langle 2/2\rangle}+\pi_{\langle 3/2\rangle})\\
\sigma^A_{\langle 1/3\rangle} &= \frac{\lambda}{\lambda+\mu}(\pi_{\langle 1/3\rangle}+\pi_{\langle 2/3\rangle}+\pi_{\langle 3/3\rangle})\\
\sigma^A_{\langle 2/1\rangle} &= \frac{\mu}{\lambda+\mu}(\pi_{\langle 2/1\rangle}+\pi_{\langle 2/2\rangle}+\pi_{\langle 2/3\rangle})\\
\sigma^A_{\langle 3/1\rangle} &= \frac{\mu}{\lambda+\mu}(\pi_{\langle 3/1\rangle}+\pi_{\langle 3/2\rangle}+\pi_{\langle 3/3\rangle})
\end{aligned} \tag{8.15}$$

Adding the new state $\langle \mathrm{E}\rangle$ to the Markov chain $\mathcal{M}_A$ and redirecting all transitions that correspond to events to this state leads to the Markov chain shown in Figure 8.11.

With this Markov chain, we calculate the complementary cumulative distribution function $\varphi_i(\cdot)$ of the time to the next event given that the Markov chain is in state i:

$$\varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{E}\rangle \\ 0 & i = \langle \mathrm{E}\rangle \end{cases} \tag{8.16}$$

$$\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \tag{8.17}$$

Now we calculate the complementary distribution function $F_A(\cdot)$ of the interevent times with

$$F_A(t) = 1 - \sum_{i,j} \sigma^A_{\langle i/j\rangle}\ \varphi_{\langle i/j\rangle}(t) \tag{8.18}$$

Coxian distribution with bypass (Cox*+Cox*)

Let $1-p_1$ and $1-p_2$ be the bypass probabilities of the two Coxian* distributions, that is, the probabilities that the interevent times are zero.

The probability that in the superposition the interevent time is zero depends on the probability that an event is caused by a certain process and on the bypass probability of this process. If S is the interevent time of the superposition, we have

$$\mathrm{P}\{S=0\} = \frac{\lambda}{\lambda+\mu}(1-p_1) + \frac{\mu}{\lambda+\mu}(1-p_2) \tag{8.19}$$

The time between batches of events is calculated as described in the Cox+Cox case, whereby we neglect the probability mass at 0 (that is, we use the pure Coxian part of the Coxian* distribution). If F_Z is the cumulative distribution function of the thereby resulting interbatch time, we have

$$\mathrm{P}\{S\le t\} = \mathrm{P}\{S=0\} + F_Z(t)\left(\frac{\lambda}{\lambda+\mu}p_1 + \frac{\mu}{\lambda+\mu}p_2\right) \tag{8.20}$$

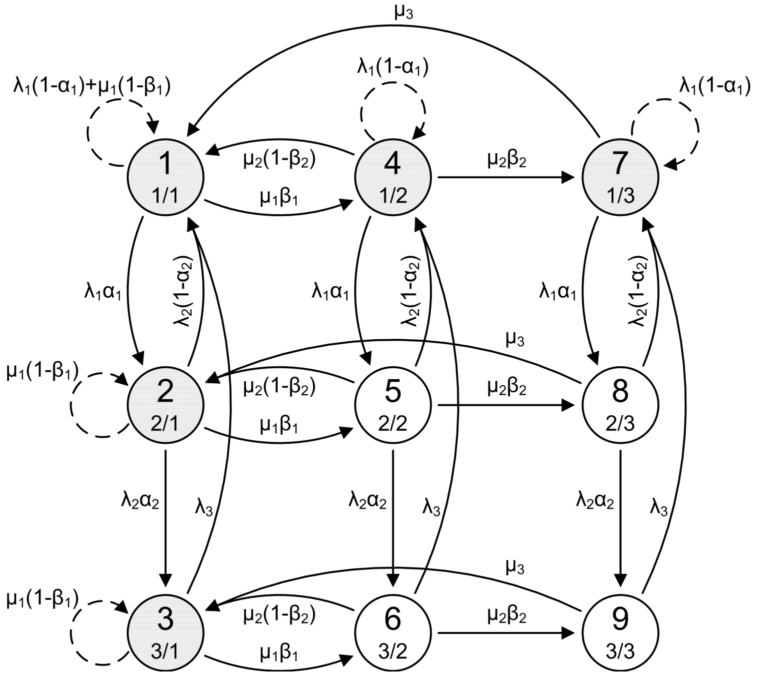


Figure 8.10.: Superposition of two streams with Coxian distributed interevent times: Markov chain for the state of the component streams. Hidden transitions are indicated with dashed lines.

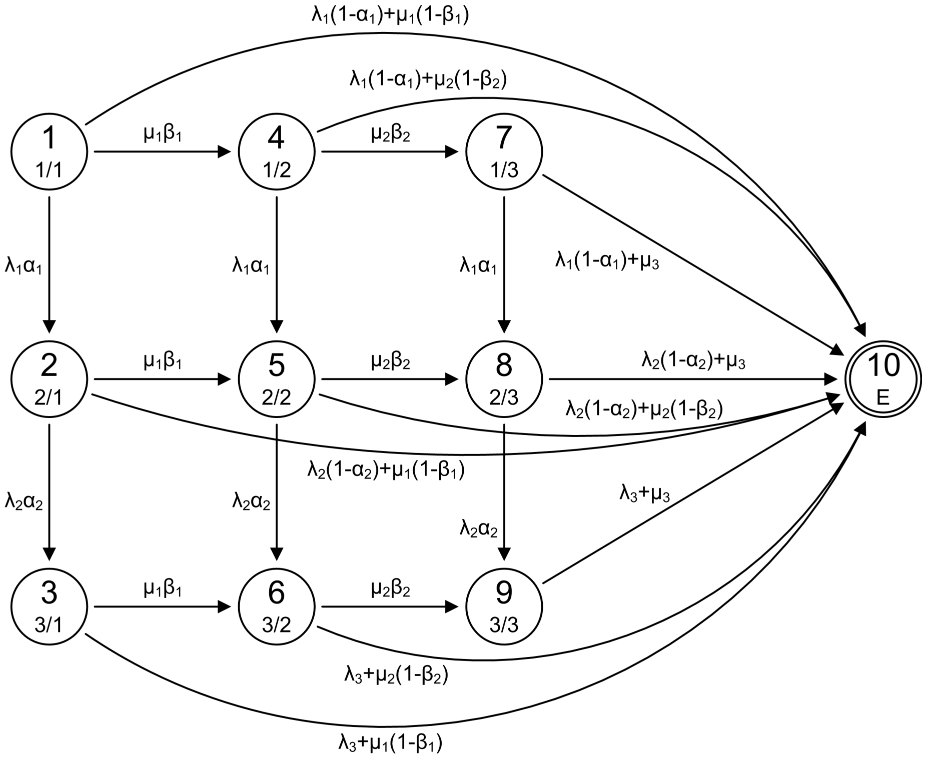


Figure 8.11.: Superposition of two streams with Coxian distributed interevent times: Markov chain for the calculation of the time to the next event.

8.1.4. Further examples for the superposition of two streams with PH-distributed interevent times

M+M

The needed Markov chains are shown in Figure 8.12.

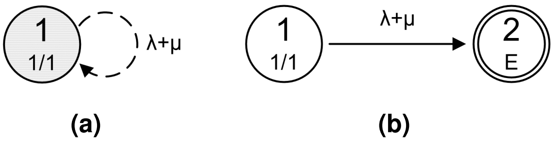


Figure 8.12.: Superposition of two streams with exponentially distributed interevent times: (a) Markov chain $\mathcal{M}_A$ for the state of the streams, (b) Markov chain for the calculation of the time to the next event.

Since both streams are memoryless, $\mathcal{M}_A$ consists of only one state. No matter which stream has caused an event, the Markov chain is always in this state:

$$\sigma^A_{\langle 1/1\rangle} = 1 \tag{8.21}$$

The complementary cumulative distribution function of the time to the next event is calculated with

$$\varphi_i(0) = \begin{cases} 1 & i \neq \langle \mathrm{E}\rangle \\ 0 & i = \langle \mathrm{E}\rangle \end{cases} \tag{8.22}$$

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{8.23}$$

which has the solution

$$\varphi_{\langle 1/1\rangle}(t) = \mathrm{e}^{-(\lambda+\mu)} \tag{8.24}$$

$$\varphi_{\langle \mathrm{E}\rangle}(t) = 0 \tag{8.25}$$

The complementary distribution function of the interevent times is

$$F_A(t) = 1 - \sigma^A_{\langle 1/1\rangle}\, \varphi_{\langle 1/1\rangle}(t) = 1 - \mathrm{e}^{-(\lambda+\mu)} \tag{8.26}$$

which is the complementary distributon function of an exponentially distributed random variable. This means a superposition of streams with exponentially distributed interevent times has exponentially distributed interevent times, too.

In the following examples, we only show the needed Markov chains and the determination of σ^A_i. The calculation of the interevent times is always done according to Equations 8.4 – 8.6.

M+Hypo

The needed Markov chains are shown in Figure 8.13. If the first process (the process with the exponentially distributed interevent times) causes an event, the state of the Markov chain for the state of the streams does not change. If the second process (the process with the hypoexponentially distributed interevent times) causes an event, the Markov chain is in state $\langle 1/1 \rangle$ after the event. Therefore, we have

$$\begin{aligned} \sigma^A_{\langle 1/1 \rangle} &= \frac{\lambda}{\lambda+\mu}\pi_{\langle 1/1 \rangle} + \frac{\mu}{\lambda+\mu} \\ \sigma^A_{\langle 1/2 \rangle} &= \frac{\lambda}{\lambda+\mu}\pi_{\langle 1/2 \rangle} \end{aligned} \tag{8.27}$$

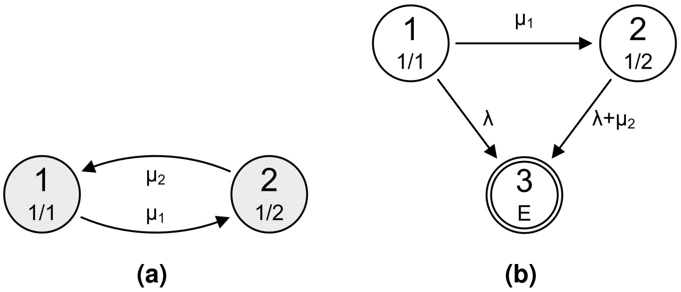


Figure 8.13.: Superposition of a stream with exponentially distributed interevent times and a stream with hypoexponentially distributed interevent times: (a) Markov chain for the state of the streams, (b) Markov chain for the calculation of the time to the next event.

M+Hyper

The needed Markov chains are shown in Figure 8.14. If the first process (the process with the exponentially distributed interevent times) causes an event, the state of the Markov chain for the state of the streams does not change. If the second process (the process with the hyperexponentially distributed interevent times) causes an event, after the event the Markov chain is in state $\langle 1/1 \rangle$ with probability β_1 and in state $\langle 1/2 \rangle$ with probability β_2. Therefore, we have

$$\begin{aligned} \sigma^A_{\langle 1/1 \rangle} &= \frac{\lambda}{\lambda+\mu}\pi_{\langle 1/1 \rangle} + \frac{\mu}{\lambda+\mu}\beta_1 \\ \sigma^A_{\langle 1/2 \rangle} &= \frac{\lambda}{\lambda+\mu}\pi_{\langle 1/2 \rangle} + \frac{\mu}{\lambda+\mu}\beta_2 \end{aligned} \tag{8.28}$$

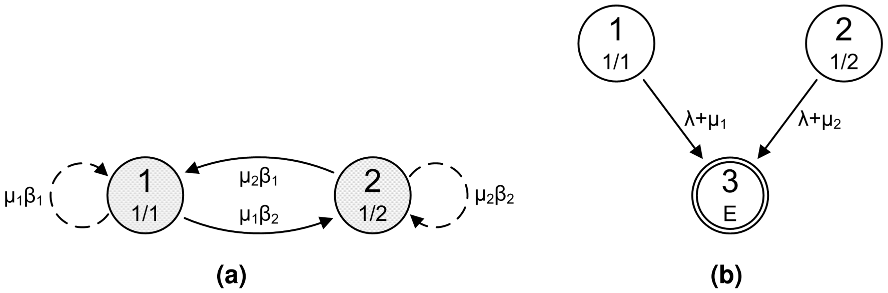


Figure 8.14.: Superposition of a stream with exponentially distributed interevent times and a stream with hyperexponentially distributed interevent times: (a) Markov chain for the state of the streams, (b) Markov chain for the calculation of the time to the next event.

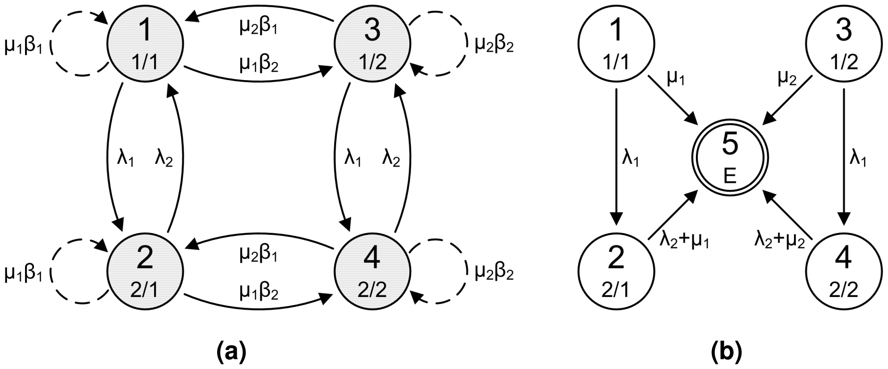


Figure 8.15.: Superposition of a stream with hypoexponentially distributed interevent times and a stream with hyperexponentially distributed interevent times: (a) Markov chain for the state of the streams, (b) Markov chain for the calculation of the time to the next event.

Hypo+Hyper

The needed Markov chains are shown in Figure 8.15.

If the first process (the process with the hypoexponentially distributed interevent times) causes an event, the Markov chain for the state of the streams is in state $\langle 1/1\rangle$ or in state $\langle 1/2\rangle$. If it was in state $\langle 2/1\rangle$ before the event occurred, it is in state $\langle 1/1\rangle$ afterwards, otherwise it is in state $\langle 1/2\rangle$. If the second process (the process with the hyperexponentially distributed interevent times) causes an event, the Markov chain is in a state $\langle \cdot/1\rangle$ with probability β_1 and in a state $\langle \cdot/2\rangle$ with probability β_2. If it was in a state $\langle 1/\cdot\rangle$ before the event occurred, it is in a state $\langle 1/\cdot\rangle$ afterwards; the same holds for states $\langle 2/\cdot\rangle$. Therefore, we have

$$\begin{aligned}
\sigma^A_{\langle 1/1\rangle} &= \frac{\lambda}{\lambda+\mu}\frac{\pi_{\langle 2/1\rangle}}{\pi_{\langle 2/1\rangle}+\pi_{\langle 2/2\rangle}} + \frac{\mu}{\lambda+\mu}(\pi_{\langle 1/1\rangle}+\pi_{\langle 1/2\rangle})\beta_1 \\
\sigma^A_{\langle 2/1\rangle} &= \frac{\mu}{\lambda+\mu}(\pi_{\langle 2/1\rangle}+\pi_{\langle 2/2\rangle})\beta_1 \\
\sigma^A_{\langle 1/2\rangle} &= \frac{\lambda}{\lambda+\mu}\frac{\pi_{\langle 2/2\rangle}}{\pi_{\langle 2/1\rangle}+\pi_{\langle 2/2\rangle}} + \frac{\mu}{\lambda+\mu}(\pi_{\langle 1/1\rangle}+\pi_{\langle 1/2\rangle})\beta_2 \\
\sigma^A_{\langle 2/2\rangle} &= \frac{\mu}{\lambda+\mu}(\pi_{\langle 2/1\rangle}+\pi_{\langle 2/2\rangle})\beta_2
\end{aligned} \tag{8.29}$$

8.1.5. Superposition of two overflow streams

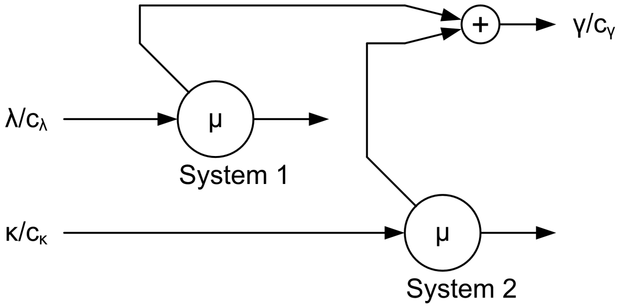


Figure 8.16.: Superposition of two overflow streams.

Finally we present a more demanding example: We consider a network consisting of two Hypo/M/1/1 queueing systems (i.e., queueing systems consisting only of a single server) whose overflow streams are superposed (Figure 8.16). We are interested in the probability distribution of the interevent times of the resulting stream.

First, we create the Markov chains for the system state for the two Hypo/M/1/1 queueing systems (Figure 8.17) and the Markov chain for the system state of the network by combining the Markov chains for the system state of the single queueing systems (Figure 8.18).

Let π be the stationary system state probabilities of the Markov chain for the system state of the queueing network. Then we can calculate the overflow rates r_i for queueing system i with

$$r_1 = \lambda_2 \sum \pi_{\langle 12\cdot\cdot\rangle} = \lambda_2 \left(\pi_4 + \pi_8 + \pi_{12} + \pi_{16}\right) \tag{8.30}$$

$$r_2 = \kappa_2 \sum \pi_{\langle\cdot\cdot 12\rangle} = \kappa_2 \left(\pi_{13} + \pi_{14} + \pi_{15} + \pi_{16}\right) \tag{8.31}$$

(See Chapter 7 for details.)

The probability that an overflow was produced by a certain system equals the ratio of its overflow rate to the total overflow rate. Therefore, we have for the probabilities p_i that system i produced an overflow

$$p_1 = \frac{r_1}{r_1 + r_2} \tag{8.32}$$

$$p_2 = \frac{r_2}{r_1 + r_2} \tag{8.33}$$

Let σ_i^R be the probabilities that the network is in state i after an overflow. After an overflow occurred, the network is in a state $\langle 11\cdot\cdot\rangle$ (if system 1 produced the overflow) or in a state $\langle\cdot\cdot 11\rangle$ (if system 2 produced the overflow). If system 1 produced the overflow, it is in state $\langle 11xy\rangle$ if it was in state $\langle 12xy\rangle$ before. If system 2 produced the

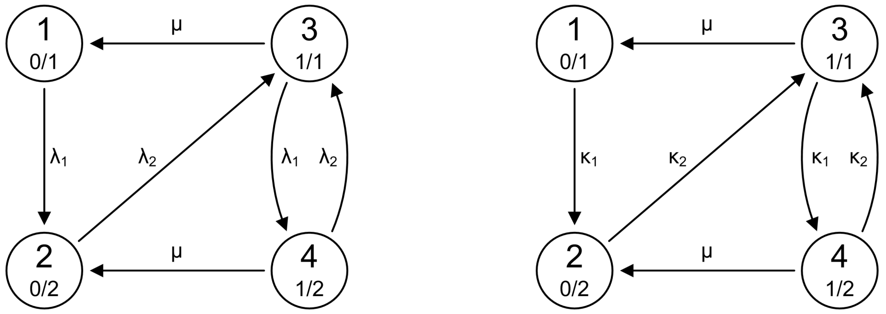


Figure 8.17.: Superposition of two overflow streams: Markov chains for the system state of the single queueing systems.

overflow, it is in state $\langle xy11\rangle$, if it was in state $\langle xy12\rangle$ before. Therefore, we have

$$\sigma^R_{\langle 1,1,0,1\rangle} = \sigma^R_3 = p_1 \frac{\pi_4}{\pi_4+\pi_8+\pi_{12}+\pi_{16}} \tag{8.34}$$
$$\sigma^R_{\langle 1,1,0,2\rangle} = \sigma^R_7 = p_1 \frac{\pi_8}{\pi_4+\pi_8+\pi_{12}+\pi_{16}} \tag{8.35}$$
$$\sigma^R_{\langle 1,1,1,2\rangle} = \sigma^R_{15} = p_1 \frac{\pi_{16}}{\pi_4+\pi_8+\pi_{12}+\pi_{16}} \tag{8.36}$$
$$\sigma^R_{\langle 0,1,1,1\rangle} = \sigma^R_9 = p_2 \frac{\pi_{13}}{\pi_{13}+\pi_{14}+\pi_{15}+\pi_{16}} \tag{8.37}$$
$$\sigma^R_{\langle 0,2,1,1\rangle} = \sigma^R_{10} = p_2 \frac{\pi_{14}}{\pi_{13}+\pi_{14}+\pi_{15}+\pi_{16}} \tag{8.38}$$
$$\sigma^R_{\langle 1,2,1,1\rangle} = \sigma^R_{12} = p_2 \frac{\pi_{16}}{\pi_{13}+\pi_{14}+\pi_{15}+\pi_{16}} \tag{8.39}$$
$$\sigma^R_{\langle 1,1,1,1\rangle} = \sigma^R_{11} = p_1 \frac{\pi_{12}}{\pi_4+\pi_8+\pi_{12}+\pi_{16}} + p_2 \frac{\pi_{15}}{\pi_{13}+\pi_{14}+\pi_{15}+\pi_{16}} \tag{8.40}$$

Now we extend the Markov chain for the state of the network with a new state $\langle \mathrm{E}\rangle$, which represents the occurrence of an overflow (Figure 8.19). All transitions that correspond to overflows now lead to this state.

With this Markov chain, we calculate the complementary cumulative distribution function φ_i of the time to the next overflow given that the network is in state i:

$$\varphi_1(0) = \cdots = \varphi_{16}(0) = 1, \varphi_{\langle \mathrm{E}\rangle}(0) = 0 \tag{8.41}$$
$$\varphi'(t) = \mathcal{Q} \cdot \varphi(t) \tag{8.42}$$

Finally, we calculate the complementary cumulative distribution function of the interevent times $F^C_A(\cdot)$ of the superposition with

$$F^C_A(t) = \sigma^R_3\varphi_3(t) + \sigma^R_7\varphi_7(t) + \sigma^R_9\varphi_9(t) + \sigma^R_{10}\varphi_{10}(t) + \\ \sigma^R_{11}\varphi_{11}(t) + \sigma^R_{12}\varphi_{12}(t) + \sigma^R_{15}\varphi_{15}(t) \tag{8.43}$$

Some results are shown in Figure 8.20.

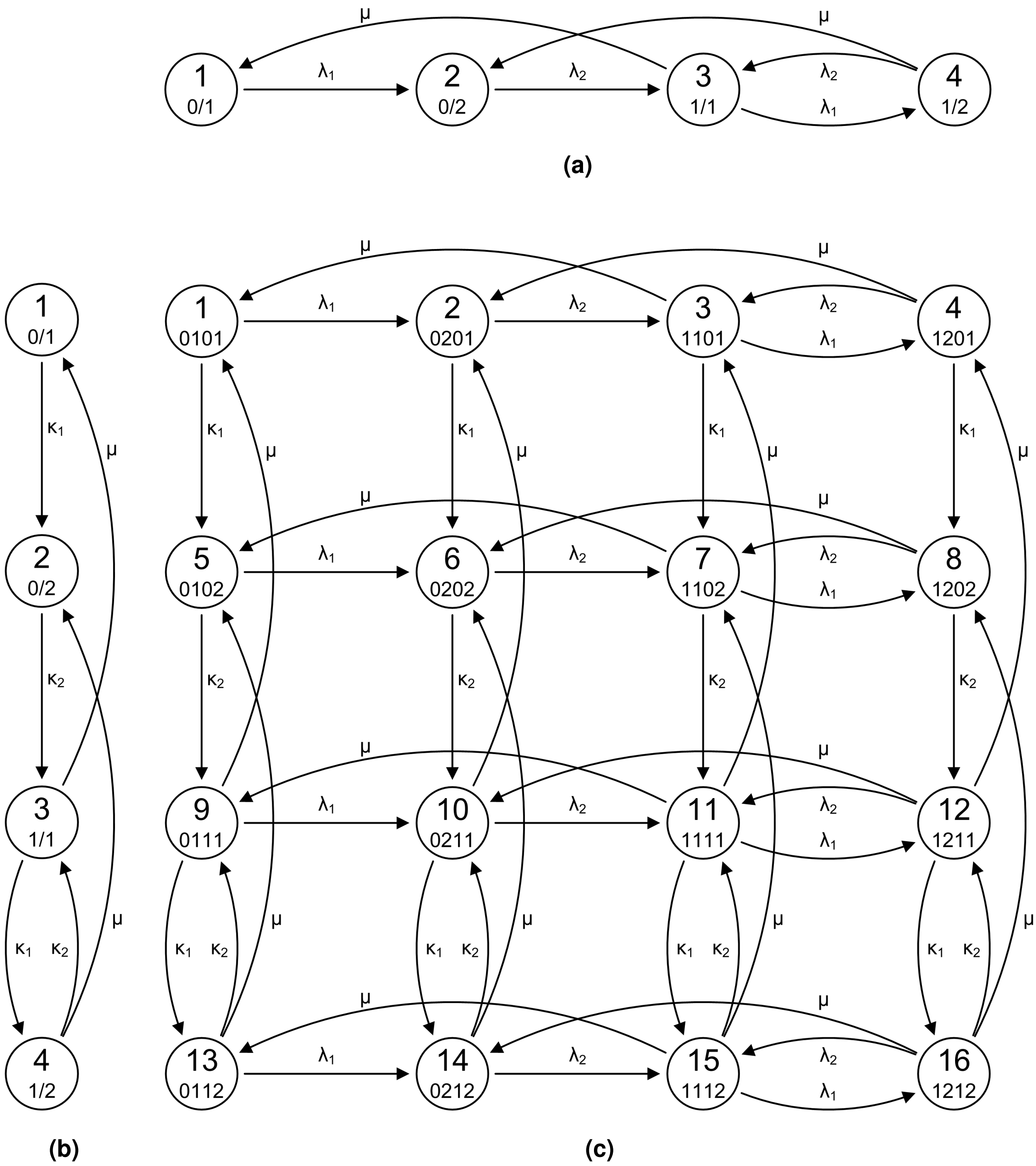


Figure 8.18.: Superposition of two overflow streams: Markov chains for the system state of (a) the first queueing system, (b) the second queueing system and (c) the network. Meaning of the names of the states: (a), (b) number of customers in the system / state of the arrival process, (c) number of customers in system 1 / state of the first arrival process / number of customers in system 2 / state of the second arrival process.

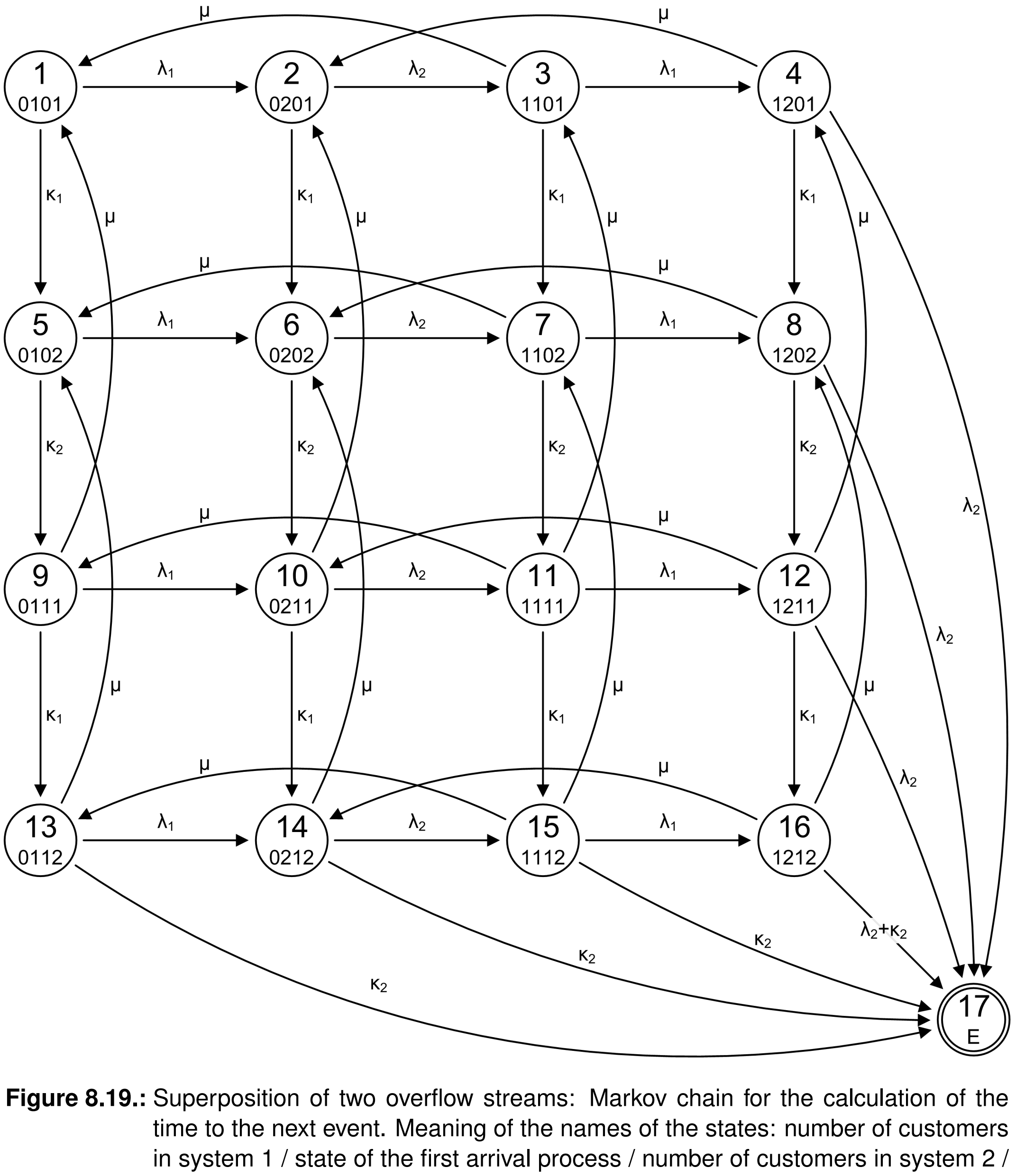


Figure 8.19.: Superposition of two overflow streams: Markov chain for the calculation of the time to the next event. Meaning of the names of the states: number of customers in system 1 / state of the first arrival process / number of customers in system 2 / state of the second arrival process.

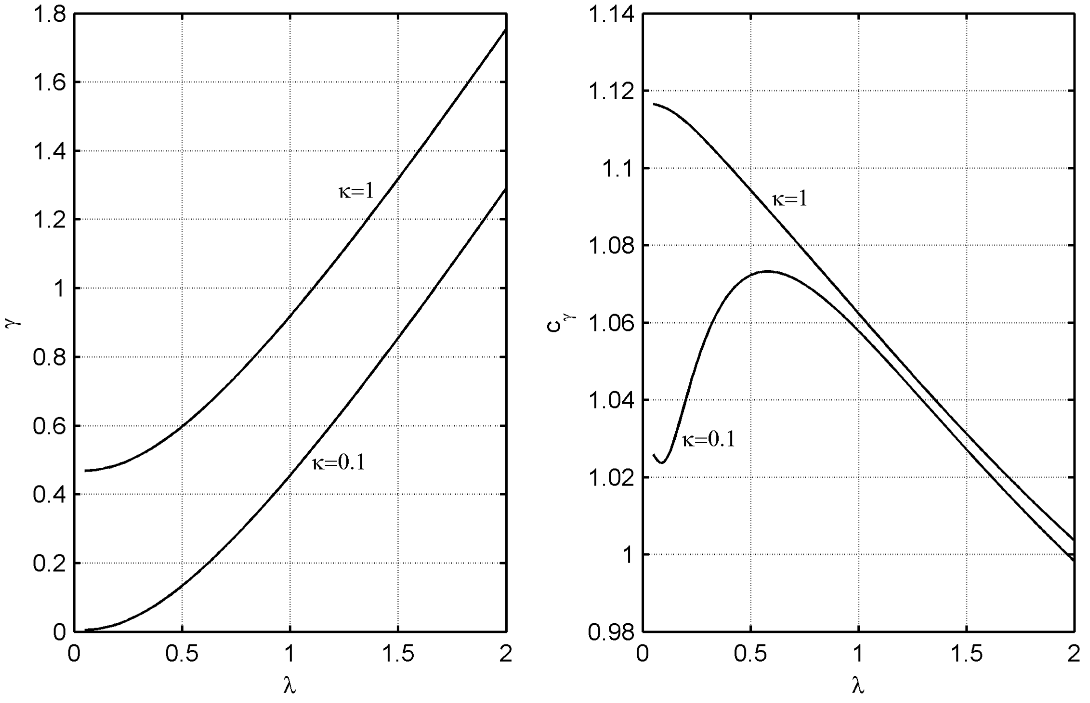


Figure 8.20.: Superposition of two overflow streams: rate and coefficient of variation. $c_\lambda = 0.75$, $c_\kappa = 0.85$, $\mu = 1$.

8.1.6. Closed-form solution

In [Kühn 1979], it is shown how a closed-form solution for the distribution of the interevent times of a superposition of two point processes can be obtained:

The forward recurrence time $T_{Vj}(t_0)$ is the interval between an arbitrary instant t_0 and the next event of the component process S_j. If process S_j is stationary, $T_{Vj}(t_0)$ is independent of t_0, and its probability density function $V_j'(t)$ is given by

$$V_j'(t) = \lambda_j F_j^C(t) \tag{8.44}$$

where λ_j is the rate of process S_j and $F_j^C(t)$ is the complementary cumulative distribution function of the interevent times of process S_j.

From this follows

$$\mathrm{P}\left\{T_{Vj} > t\right\} = \int\limits_{u=t}^{\infty} \lambda_j F_j^C(u) \mathrm{d}u \tag{8.45}$$

The time T_V to the next event in the superposition is greater than t if it is in both component processes greater than t:

$$\begin{aligned} V^C(t) = \mathrm{P}\left\{T_V > t\right\} = \\ \mathrm{P}\left\{T_{V1} > t\right\} \mathrm{P}\left\{T_{V2} > t\right\} = \int\limits_{u=t}^{\infty} \lambda_1 F_1^C(u) \mathrm{d}u \int\limits_{u=t}^{\infty} \lambda_2 F_2^C(u) \mathrm{d}u \end{aligned} \tag{8.46}$$

Assuming the renewal property and using the relation

$$V'(t) = \lambda F^C(t) \tag{8.47}$$

leads to

$$F^C(t) = \frac{V'(t)}{\lambda} = \frac{1}{\lambda_1 + \lambda_2} \frac{\mathrm{d}}{\mathrm{d}t} \left(\int\limits_{u=t}^{\infty} \lambda_1 F_1^C(u) \mathrm{d}u \int\limits_{u=t}^{\infty} \lambda_2 F_2^C(u) \mathrm{d}u \right) \tag{8.48}$$

$$F(t) = 1 - \frac{\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \left(F_1^C(t) \int\limits_{u=t}^{\infty} F_2^C(u) \mathrm{d}u + F_2^C(t) \int\limits_{u=t}^{\infty} F_1^C(u) \mathrm{d}u \right) \tag{8.49}$$

It is not possible to give the moments of the interevent times T,

$$\mathrm{E}\left(T^k\right) = \int\limits_{0-}^{\infty} t^k F'(t) \mathrm{d}t = -\frac{1}{\lambda} \int\limits_{0-}^{\infty} t^k V''(t) \mathrm{d}t \tag{8.50}$$

as a function of the moments of the component processes (except for $k = 1$). Therefore, if only the moments and not the distribution of the interevent times of the component processes are known, one has to approximate these distributions by distributions with a known distribution function.

Kühn uses, for the approximation, distributions that match two moments of the given distribution. For coefficients of variation greater than 1, he uses a 2-stage hyperexponential distribution, for coefficients of variation smaller than 1, he uses the sum of a deterministic and an exponential distribution. The sum of a deterministic and an exponential distribution is more complicated to deal with than a hypoexponential distribution, but it has the advantage that all coefficients of variation $c \in (0, 1)$ can be obtained.

Some results are shown in Figure 8.21.

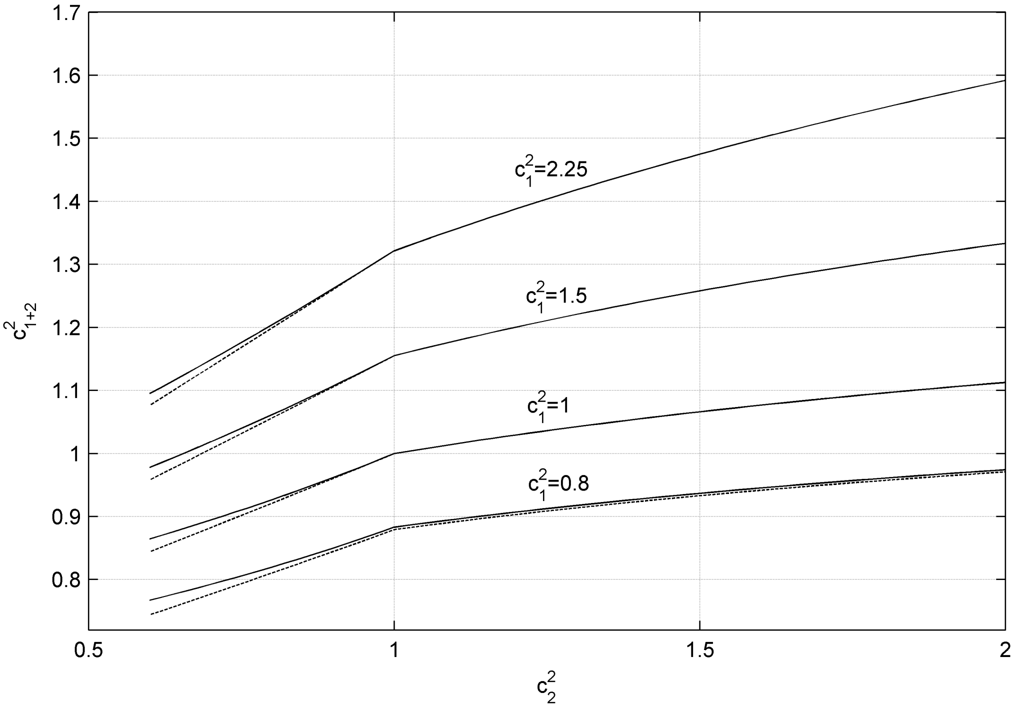


Figure 8.21.: Coefficient of variation of the superposition of two processes. For $c > 1$ hyperexponentially distributed interevent times are assumed, for $c < 1$ hypoexponentially distributed interevent times (solid line) and interevent times that are the sum of a deterministic and an exponential distribution (dashed line) are assumed.

8.2. Dependencies between the interevent times

There are many forms of dependencies within a sequence of random variables. As an example, we look at the so-called autocorrelation.

8.2.1. Autocorrelation

The autocorrelation (or serial correlation) $\mathrm{ACF}(X, d)$ of a (finite) sequence $X = \langle X_i \rangle_{1 \leq i \leq n}$ of identically distributed random variables is a measure of the extent to which X_{i+d} depends on X_i. It is defined as

$$\mathrm{ACF}(X, d) = \frac{\mathrm{acov}(X, d)}{\mathrm{Var}(X)} = \frac{\mathrm{acov}(X, d)}{\mathrm{acov}(X, 0)} \tag{8.51}$$

with the autocovariance $\mathrm{acov}(X, d)$ defined as

$$\mathrm{acov}(X, d) = \frac{1}{n-k} \sum_{k=1}^{n-k} \left(X_k - \mathrm{E}(X)\right)\left(X_{k+d} - \mathrm{E}(X)\right) \tag{8.52}$$

Sometimes, in addition, the random variables $X_{n-k+1}, \ldots, X_n$ are compared with $X_1, \ldots, X_{k-1}$. Then we have

$$\mathrm{ACF}(X, d) = \frac{\frac{1}{n} \sum\limits_{k=1}^{n} \left(X_k - \mathrm{E}(X)\right)\left(X_{(k+d) \bmod n} - \mathrm{E}(X)\right)}{\mathrm{Var}(X)} \tag{8.53}$$

Formula 8.53 can also be written as

$$\mathrm{ACF}(X, d) = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - \left(\sum X_i\right)^2} \tag{8.54}$$

where $Y_i = X_{(i+d) \bmod n}$.

The parameter d is called delay or lag.

If the random variables X_i are independent, the autocorrelation $\mathrm{ACF}(X, d)$ is close to 0.[1] If there is a strong positive (negative) correlation, the autocorrelation is close to 1 (-1).

[1] For details see [Knuth 1997].

8.2.2. Autocorrelation of the interevent times

As can be seen in Figure 8.22, the autocorrelation of the interevent times has its maximum when traffic streams with the same intensity are superposed. As the ratio of the intensities becomes smaller or greater than 1, one stream dominates the superposition, and since the interevent times of the single streams are assumed to be independent, the autocorrelation tends towards 0.

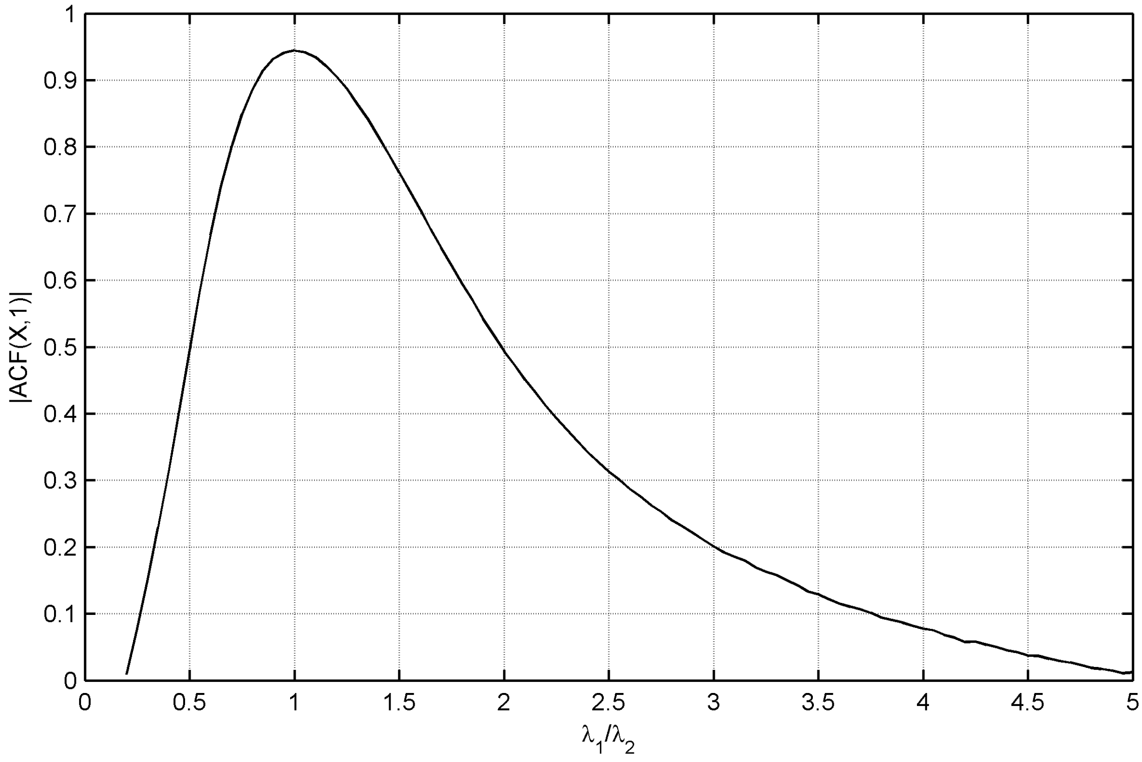
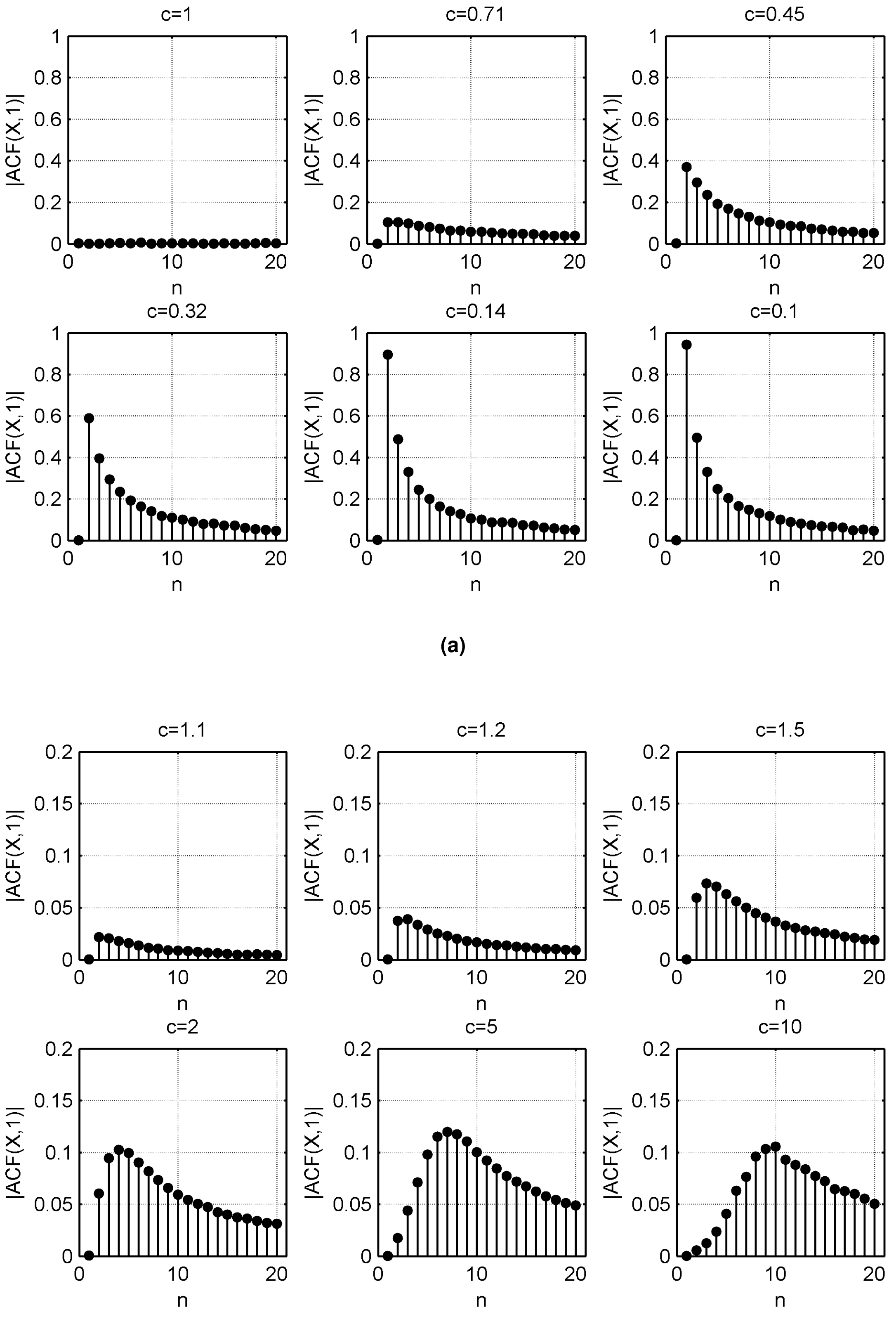


Figure 8.22.: Superposition of two traffic streams with Erlang-distributed interevent times ($c = 0.1$). The autocorrelation of the interevent times of the superposition reaches its maximum when the streams have the same intensity.

Figure 8.23 shows the autocorrelation of the interevent times of the superposition depending on the number of streams that are superposed. The superposition of streams with hyperexponentially distributed interevent times ($c > 1$) has low autocorrelation. The superposition of streams with hypoexponentially distributed interevent times ($c < 1$) has a higher autocorrelation. The smoother the stream is, the stronger the interevent times are correlated. If two streams with deterministic interevent times are superposed, the autocorrelation is 1 (for d odd) or -1 (for d even). The superposition of Poisson processes is a Poisson process and therefore has no autocorrelation.



(a)

(b)

Figure 8.23.: Superposition of n traffic streams with (a) hypoexponentially and (b) hyperexponentially distributed interevent times: autocorrelation of the interevent times. (Simulation study.)

8.2.3. Effect of dependencies

For an estimation of the effect of the dependencies between the interevent times of the superposition we compare the number of customers in a queueing system for two different arrival streams: The first stream is the superposition of two streams (Figure 8.24a). The second stream has the same interevent times as the first stream, but these interevent times have been shuffled (Figure 8.24b), so that they are not interdependent any more.

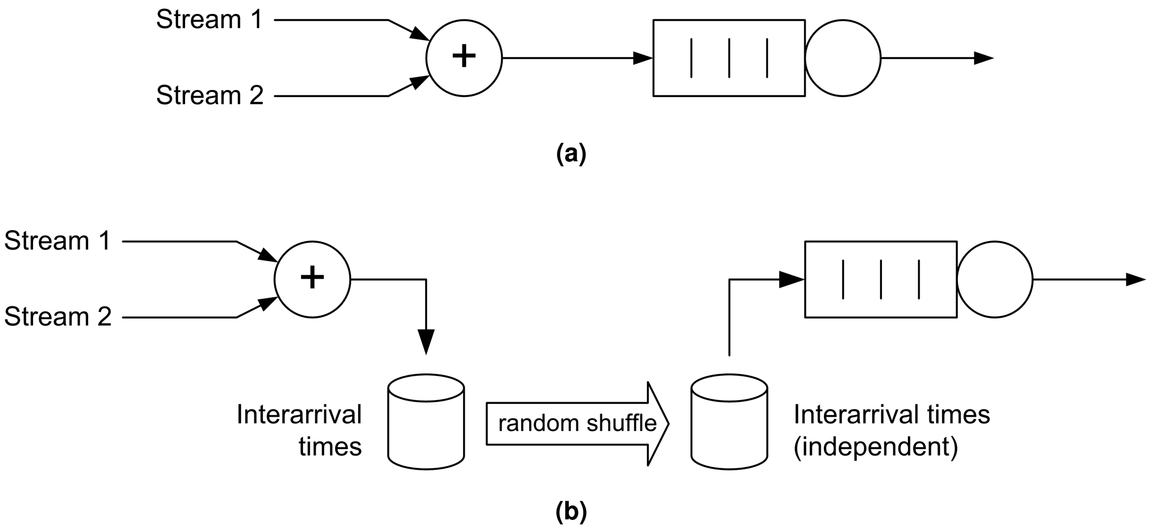


Figure 8.24.: Estimation of the effect of the dependencies between the interevent times of a superposition. (a) The interarrival times at the queueing system are the real (interdependent) interevent times. (b) The interarrival times at the queueing system are the interevent times without dependencies.

If the coefficient of variation of the streams is close to 1, there is little difference between the results (Figure 8.25 and Figure 8.26). If the coefficient of variation of the streams is close to zero, the difference is much higher (Figure 8.27).

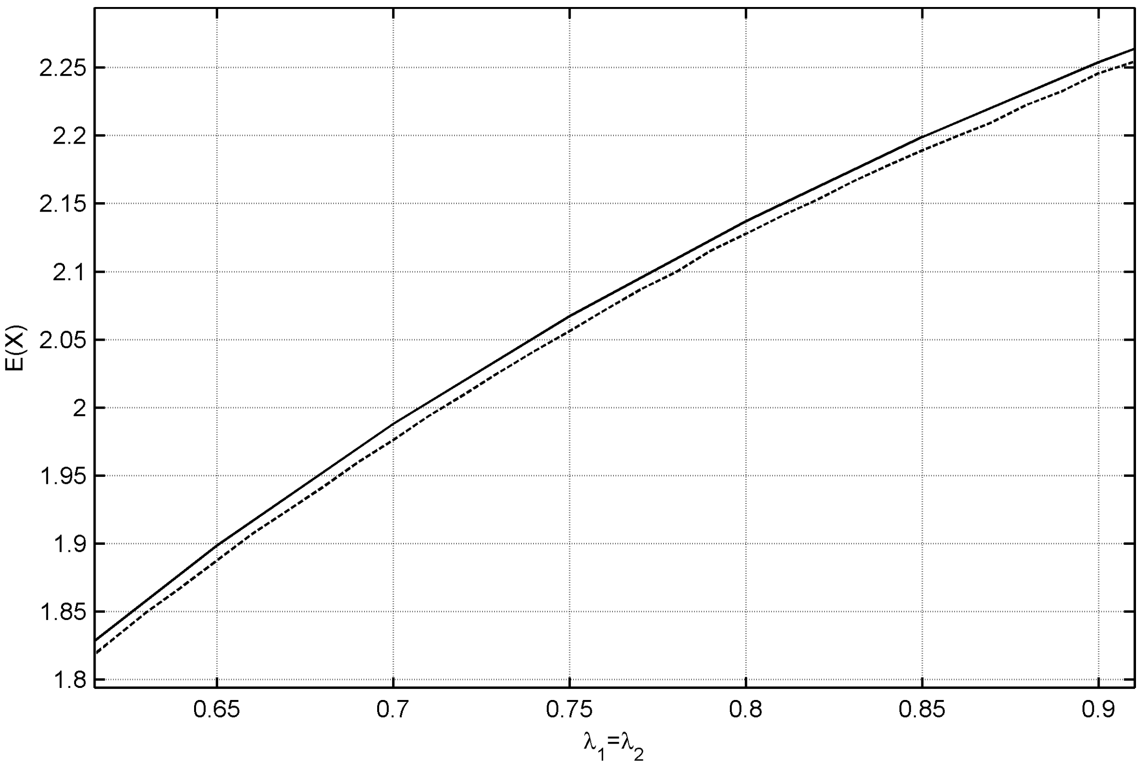


Figure 8.25.: Number of customers in a Hypo+Hypo/M/1/S queueing system. $S = 3$, $c_1 = 0.75$, $c_2 = 0.85$. Solid line: real interarrival times, dashed line: interarrival times without dependencies.

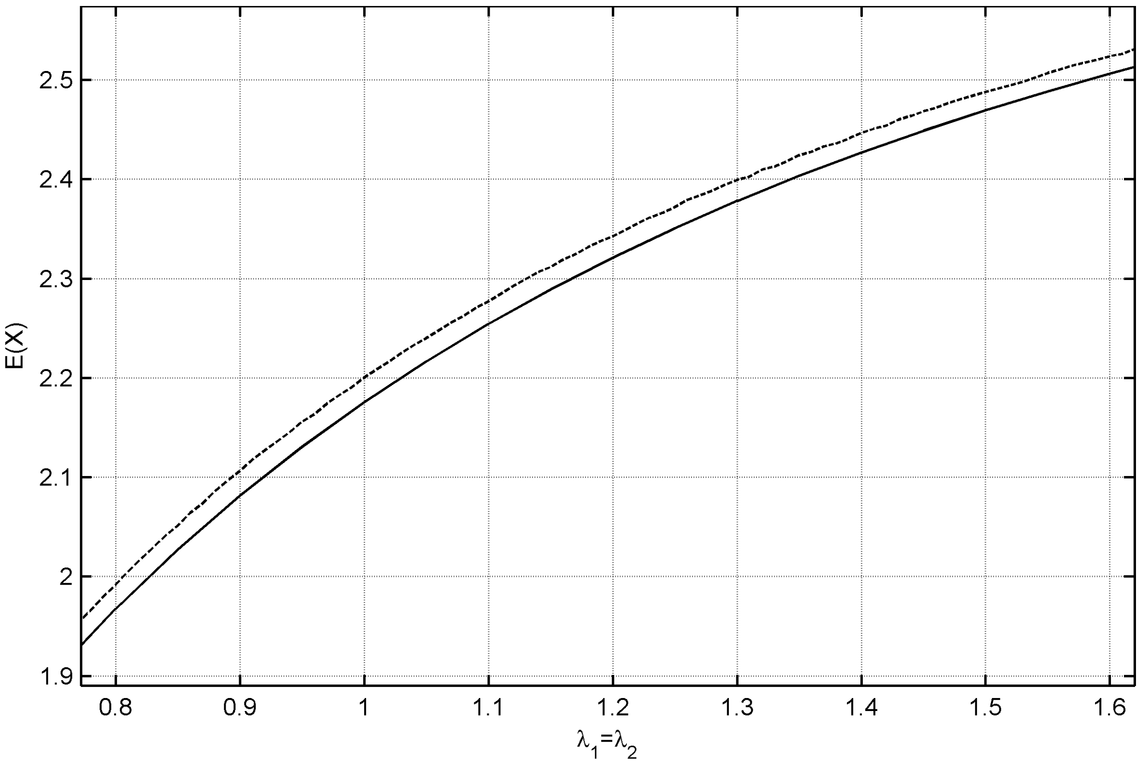


Figure 8.26.: Number of customers in a Hyper+Hyper/M/1/S queueing system. $S = 3$, $c_1 = 1.5$, $c_2 = 1.25$. Solid line: real interarrival times, dashed line: interarrival times without dependencies.

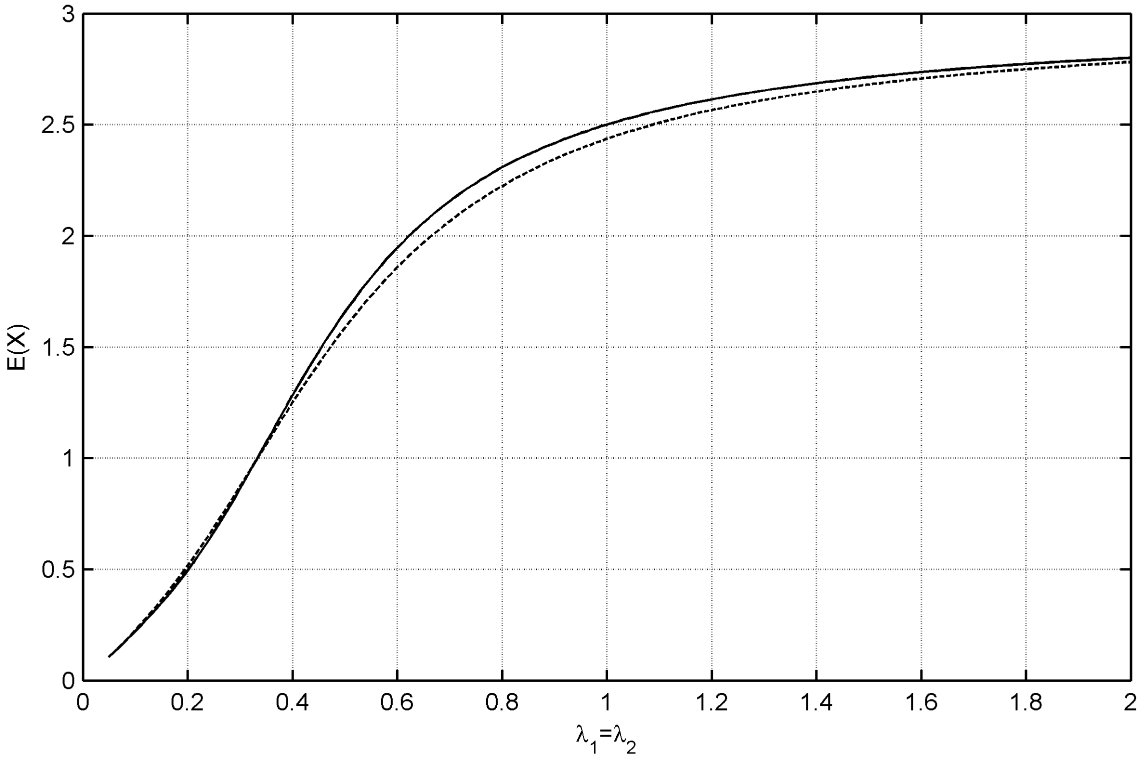


Figure 8.27.: Number of customers in a Erlang+Erlang/M/1/S queueing system. $S = 3$, $c_1 = c_2 = 0.1$. Solid line: real interarrival times, dashed line: interarrival times without dependencies.

8.3. Modelling PH+PH/M/1/S queueing systems

The superposition of traffic streams can also be included in the Markov chain for a queueing system. The Markov chain for the system state of a PH+PH/M/1/S queueing system can be constructed from the Markov chain for the system state of a PH/M/1/S queueing system by adding the states and transitions needed to describe the second arrival process.

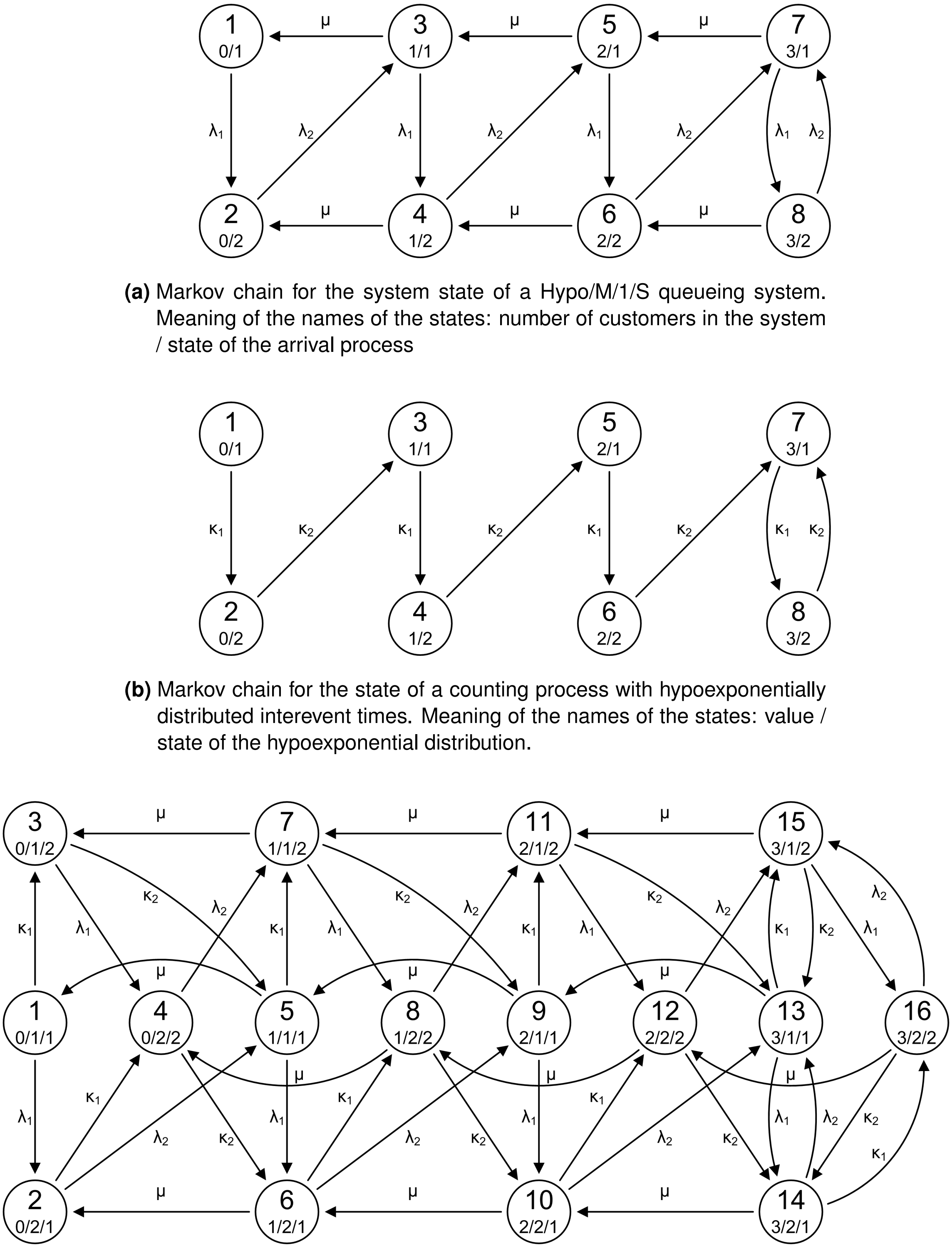
8.3.1. Hypo+Hypo/M/1/S queueing system

The construction of the Markov chain for the system state of a Hypo+Hypo/M/1/S queueing system is shown in Figure 8.28: We start with the Markov chain for the system state of a Hypo/M/1/S queueing system (Figure 8.28a) and add the second arrival process (Figure 8.28b). In the resulting Markov chain (Figure 8.28c), in each state there can be a transition of either the first or the second arrival process.

8.3.2. Other PH+PH/M/1/S queueing systems

The Markov chain for the system state of an M+M/M/1/S queueing system is shown in Figure 8.29. Since the second arrival process does not introduce additional states, the arrival rates are simply added. (This matches the result that the sum of Poisson processes is a Poisson process whose intensity is the sum of the intensities of the component processes.)

The Markov chains for the system state of queueing systems of type M+Hypo/M/1/S, M+Hyper/M/1/S, Hypo+Hyper/M/1/S and Hyper+Hyper/M/1/S are shown in Figures 8.30 to 8.33. They are constructed in the same manner as the Markov chain for the system state of the Hypo+Hypo/M/1/S queueing system, therefore, we show them without further explanation.



(a) Markov chain for the system state of a Hypo/M/1/S queueing system. Meaning of the names of the states: number of customers in the system / state of the arrival process

(b) Markov chain for the state of a counting process with hypoexponentially distributed interevent times. Meaning of the names of the states: value / state of the hypoexponential distribution.

(c) Markov chain for the system state of a Hypo+Hypo/M/1/S queueing system. Meaning of the names of the states: number of customers in the system / state of the first arrival process / state of the second arrival process.

Figure 8.28.: Hypo+Hypo/M/1/S queueing system.

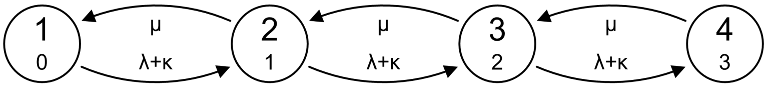


Figure 8.29.: M+M/M/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system.

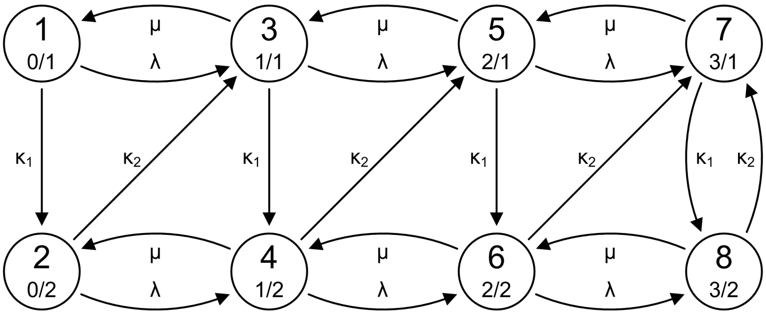


Figure 8.30.: M+Hypo/M/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the second arrival process.

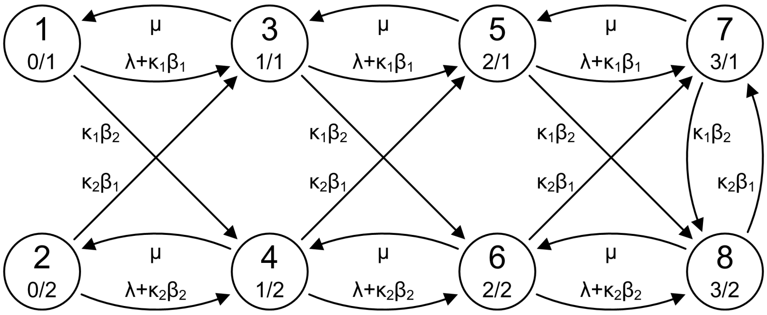


Figure 8.31.: M+Hyper/M/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the second arrival process.

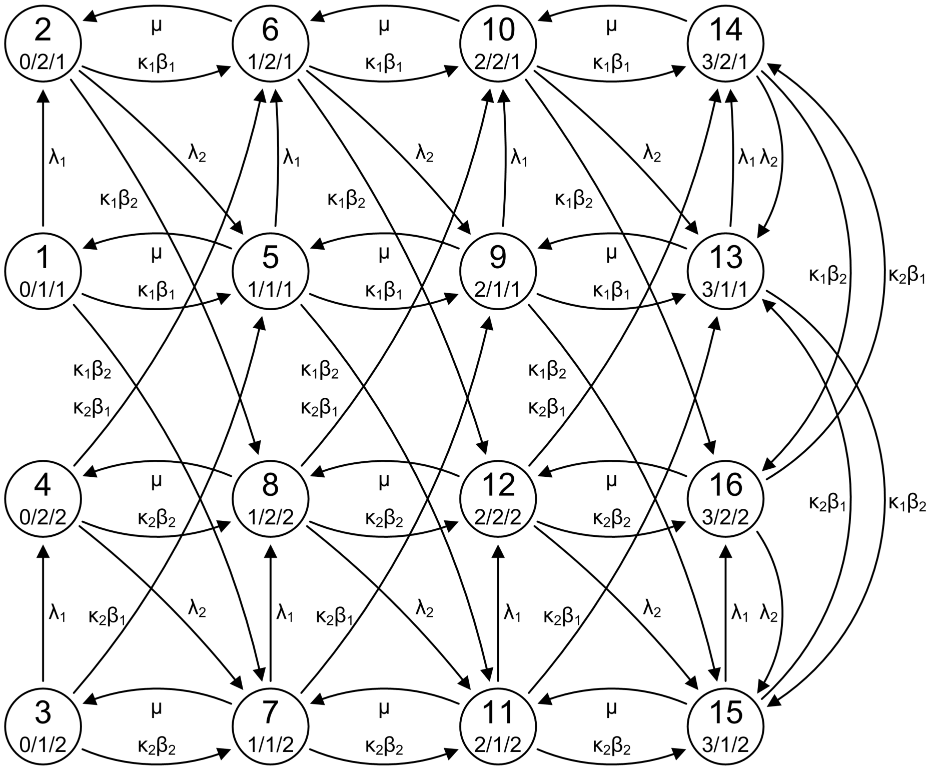


Figure 8.32.: Hypo+Hyper/M/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the first arrival process / state of the second arrival process.

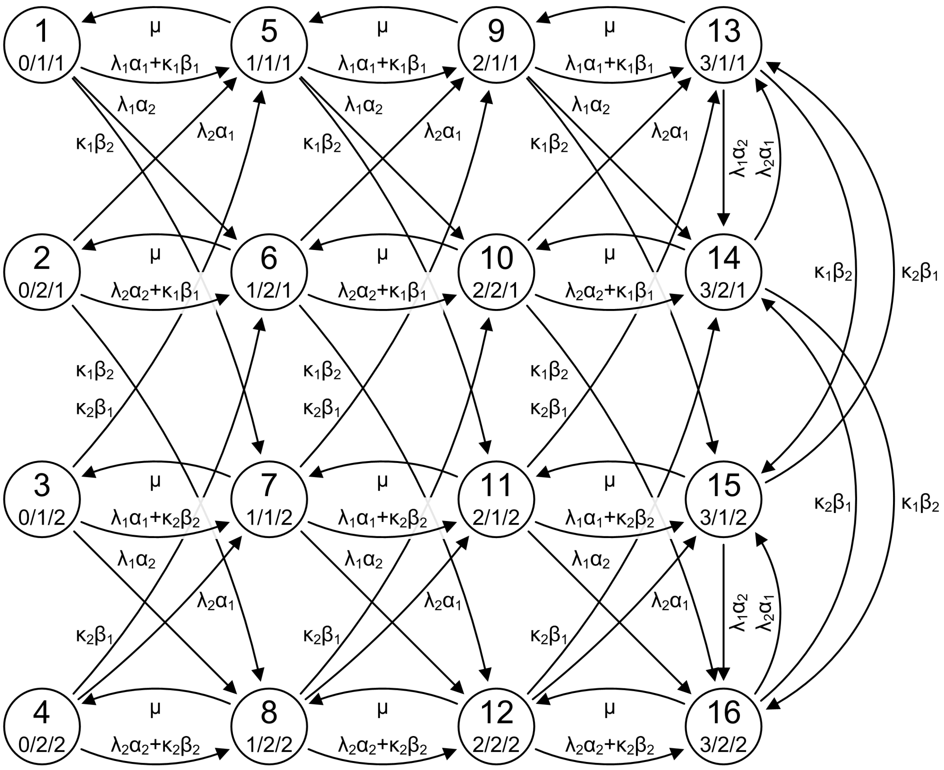


Figure 8.33.: Hyper+Hyper/M/1/S queueing system: Markov chain for the system state. Meaning of the names of the states: number of customers in the system / state of the first arrival process / state of the second arrival process.

9. Decomposition of traffic streams

The dual problem with the superposition of traffic streams is the decomposition of one traffic stream into several component streams (Figure 9.1).

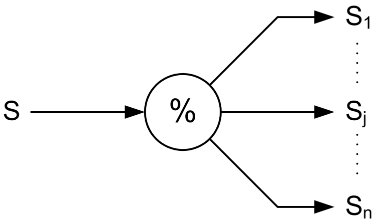


Figure 9.1.: Decomposition of a point process into n component processes.

Given a traffic stream $S = \langle T_1, T_2, \dots \rangle$ with independent and identically distributed interevent times X_j (i.e., it constitutes a stationary renewal process), a decomposition of this stream is defined as a partition of its events (customers) to n pairwise disjoint component streams $S_i = \langle T_{i,1}, T_{i,2}, \dots \rangle$, $i = 1 \dots n$,

$$S = \bigcup_i S_i = \bigcup_{i,j} T_{i,j} \tag{9.1}$$

$$S_i \cap S_j = \emptyset \quad \text{for } i \neq j \tag{9.2}$$

The decision how the customers are assigned to the component processes is made according to the routing policy. We consider the routing policies *probability based routing* and *round-robin routing*.

Probability based routing (Figure 9.2a) means that a customer is assigned to component stream S_i with probability θ_i, where the θ_i are independent of previous routing decisions.

In the case of *round-robin routing* (Figure 9.2b), customers are alternatingly assigned to the component streams:

$$T_{i,j} = T_{(j-1)n+i} \tag{9.3}$$

We know the statistical characteristics of the original stream, and we are interested in the statistical characteristics of the component streams.

The component streams constitute renewal processes, so the only task is to determine the interevent times. We do this in Section 9.1.

In Section 9.2, we show how to model queueing systems that receive only a portion of a traffic stream.

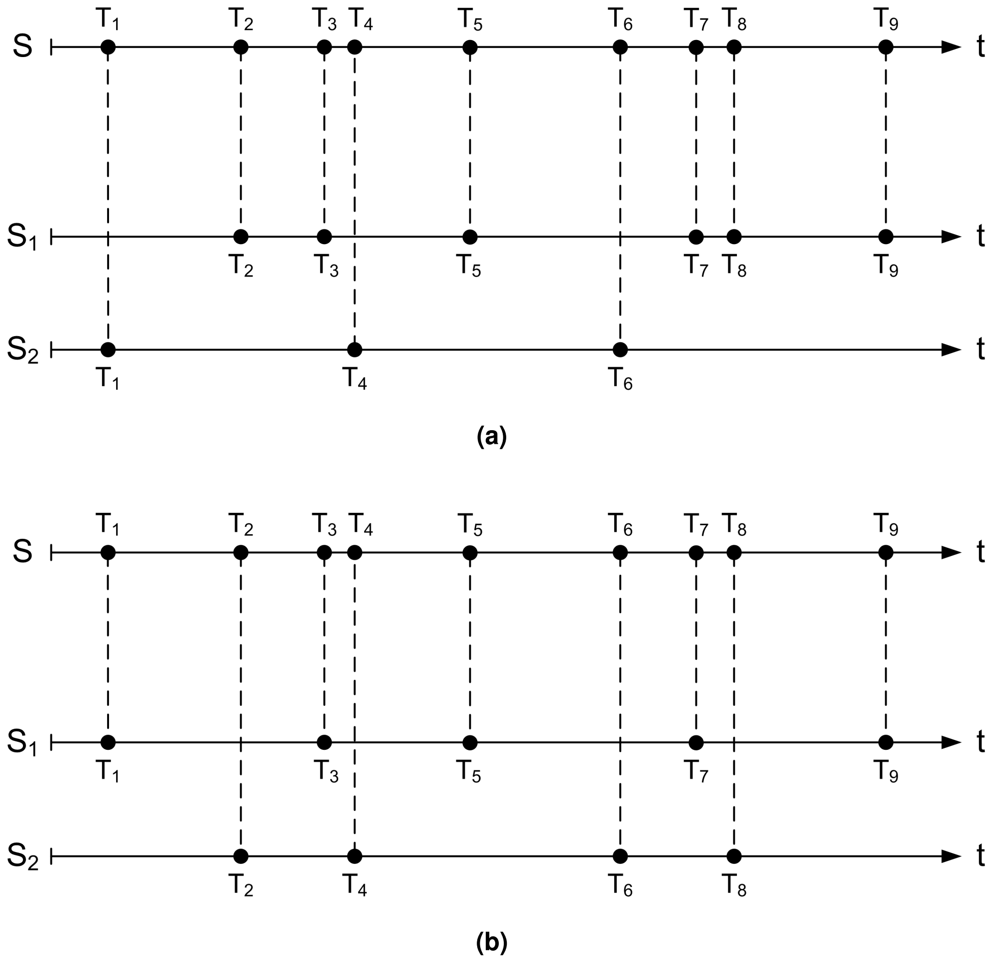


Figure 9.2.: Decomposition of a point process into two component processes. (a) Probability based, (b) round-robin based.

9.1. Interevent times

Probability based routing

The mean interevent time of component stream i is

$$\mathrm{E}\left(X_{i,j}\right) = \mathrm{E}\left(X_j\right) \cdot \frac{1}{\theta_i} \tag{9.4}$$

For the calculation of the probability distribution of the interevent times, we use the Markov chain for the state of the source stream S (cf. Section 3.5).

First, we determine the probability σ_k^D that the Markov chain is in state k after an event in S.

The next step is to extend the Markov chain by a state $\langle \mathrm{E} \rangle$, to which the Markov chain goes when there is an event in the component stream under consideration (S_i). For all transitions $k \to j$ that correspond to events in S, we set

$$q_{k,j}^{(\text{new})} = q_{k,j}^{(\text{old})} \cdot (1 - \theta_i) \tag{9.5}$$

and we add a new transition $k \to \langle \mathrm{E} \rangle$ with rate

$$q_{k,\langle \mathrm{E} \rangle} = q_{k,j}^{(\text{old})} \cdot \theta_i \tag{9.6}$$

Now we calculate the complementary cumulative distribution function $\varphi_k(\cdot)$ of the time that the Markov chain needs to go from state k to state $\langle \mathrm{E} \rangle$:

$$\varphi_k(0) = \begin{cases} 1 & k \neq \langle \mathrm{E} \rangle \\ 0 & k = \langle \mathrm{E} \rangle \end{cases} \tag{9.7}$$

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{9.8}$$

The cumulative distribution function $F_i(\cdot)$ of the interevent times of the component stream S_i is

$$F_i(t) = 1 - \sum_k \sigma_k^D \, \varphi_k(t) \tag{9.9}$$

Round-robin routing

If there are n component streams, the interevent times of all these component streams $X_{i,j}$ are the convolution of n interevent times of the source stream,

$$X_{i,j} = \underbrace{X_j + X_j + \cdots + X_j}_{n \text{ times}} \qquad i = 1 \ldots n \tag{9.10}$$

The mean interevent time of the component streams is

$$\mathrm{E}\left(X_{i,j}\right) = \mathrm{E}\left(X_j\right) \cdot n \qquad i = 1 \ldots n \tag{9.11}$$

If we want to use Markov chains for the calculation of the probability distribution of the interevent times, this is done as follows:

We start with n Markov chains $M_1 \dots M_n$ for the system state of the source stream. We identify transitions $k \to j$ that correspond to events in the source stream. Beginning with M_1, we redirect these transitions to the next Markov chain: $M_m : k \to M_m : j$ becomes $M_m : k \to M_{m+1} : j$. Transitions in the last Markov chain are redirected to Markov chain M_1: $M_n : k \to M_n : j$ becomes $M_n : k \to M_1 : j$. The resulting Markov chain describes the state of the source stream and of the round-robin scheduler. With each transition to the next partial Markov chain another component stream receives a customer. We assume that transitions from M_n to M_1 correspond to events in the component stream under consideration.

Now we determine the probabilities σ_k^D that the Markov chain is in state k after an event has occurred in the component stream. Of course, all states k with $\sigma_k^D > 0$ are states of M_1.

After redirecting all transitions from M_n to M_1 to a new state $\langle \mathrm{E} \rangle$, we calculate the complementary cumulative distribution function $\varphi_k(\cdot)$ of the time that the Markov chain needs to go from state k to state $\langle \mathrm{E} \rangle$ as shown above.

Again, the cumulative distribution function $F_i(\cdot)$ of the interevent times of the component streams is

$$F_i(t) = 1 - \sum_k \sigma_k^D \, \varphi_k(t) \tag{9.12}$$

In the following examples, we will always assume that we have round-robin based routing with two component processes.

9.1.1. Decomposition of a stream with hypoexponentially distributed interevent times

Probability based routing

Figure 9.3a shows the Markov chain for the system state of the source stream S. The transition from state $\langle 2 \rangle$ to state $\langle 1 \rangle$ corresponds to events in the source stream.

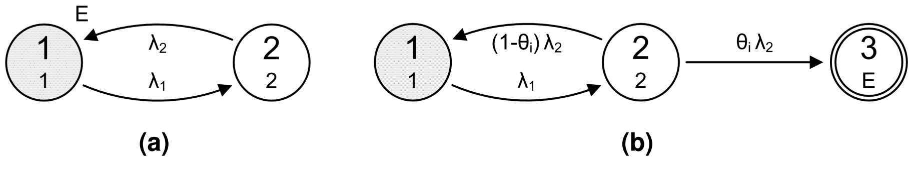


Figure 9.3.: Decomposition of a stream with hypoexponentially distributed interevent times (probability based). (a) Markov chain for the state of the stream, (b) Markov chain for the calculation of the time to the next event.

After an event has occurred, the Markov chain is in state $\langle 1 \rangle$, therefore, we have

$$\sigma^D_{\langle 1 \rangle} = 1 \tag{9.13}$$

The extended Markov chain is shown in Figure 9.3b. When the Markov chain leaves state $\langle 2 \rangle$, it goes to state $\langle \mathrm{E} \rangle$ with probability θ_i or to state $\langle 1 \rangle$ with probability $1 - \theta_i$. If it goes to state $\langle \mathrm{E} \rangle$, there is an event in the component stream under consideration S_i; otherwise there is an event in another component stream.

We calculate the complementary cumulative distribution function $\varphi_k(\cdot)$ of the time that the Markov chain needs to go from state k to state $\langle \mathrm{E} \rangle$:

$$\varphi_{\langle 1 \rangle}(0) = \varphi_{\langle 2 \rangle}(0) = 1 \tag{9.14}$$

$$\varphi_{\langle \mathrm{E} \rangle}(0) = 0 \tag{9.15}$$

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{9.16}$$

The cumulative distribution function of the interevent times of the component stream is

$$F_i(t) = 1 - \varphi_{\langle 1 \rangle}(t) \tag{9.17}$$

Round-robin routing

Figure 9.4a shows the Markov chain for the state of the source stream. The transition from state $\langle 2 \rangle$ to state $\langle 1 \rangle$ corresponds to events in the source stream. Figure 9.4b shows the Markov chain for the state of the source stream and the round-robin scheduler. In this Markov chain, transitions from state $\langle 2/2 \rangle$ to state $\langle 1/1 \rangle$ correspond to events in the component process under consideration, and transitions from state $\langle 2/1 \rangle$ to state $\langle 1/2 \rangle$ correspond to events in the other component process.

After an event in the component stream under consideration, the Markov chain is in state $\langle 1/1 \rangle$, so we have

$$\sigma^D_{\langle 1/1 \rangle} = 1 \tag{9.18}$$

In Figure 9.4c, the Markov chain for the calculation of the time to the next event is shown. The transition from state $\langle 2/2 \rangle$ to state $\langle 1/1 \rangle$ now leads to state $\langle \mathrm{E} \rangle$. With this Markov chain, we calculate the complementary cumulative distribution function $\varphi_k(\cdot)$ of the time that the Markov chain needs to go from state k to state $\langle \mathrm{E} \rangle$:

$$\varphi_k(0) = \begin{cases} 1 & k \neq \langle \mathrm{E} \rangle \\ 0 & k = \langle \mathrm{E} \rangle \end{cases} \tag{9.19}$$

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{9.20}$$

The cumulative distribution function of the interevent times of the component stream is

$$F_i(t) = 1 - \varphi_{\langle 1/1 \rangle}(t) \tag{9.21}$$

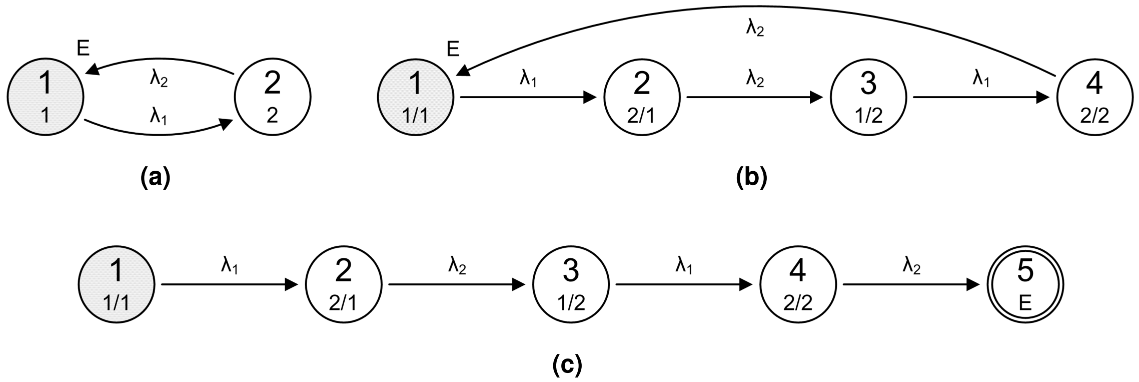


Figure 9.4.: Decomposition of a stream with hypoexponentially distributed interevent times (round-robin). (a) Markov chain for the state of the stream, (b) Markov chain for the state of the stream and the round-robin scheduler, (c) Markov chain for the calculation of the time to the next event. Meaning of the names of the states: (a) state of the stream, (b,c) state of the stream / state of the round-robin scheduler.

9.1.2. Decomposition of a stream with hyperexponentially distributed interevent times

Probability based routing

Figures 9.5a and 9.5b show the Markov chains for the system state of the source stream. All transitions correspond to events in the source stream.

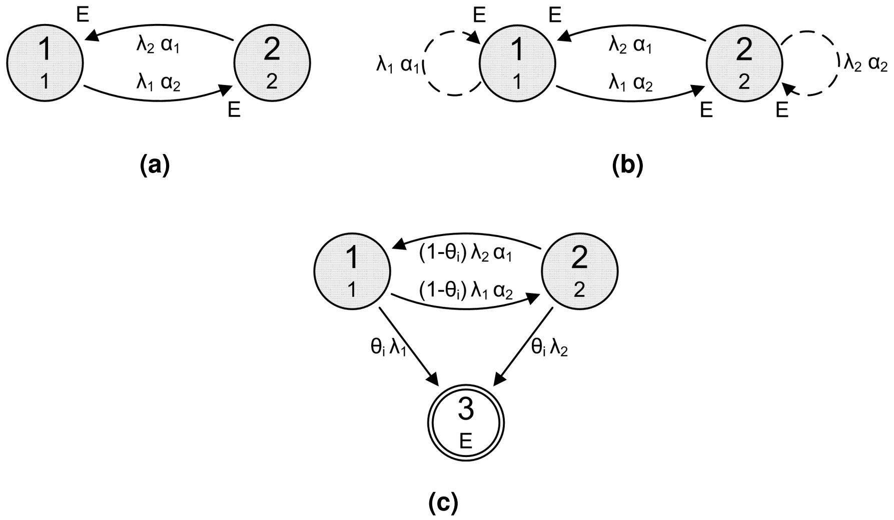


Figure 9.5.: Decomposition of a stream with hyperexponentially distributed interevent times (probability based). (a,b) Markov chains for the state of the stream, (c) Markov chain for the calculation of the time to the next event.

After an event has occurred, the Markov chain is in state $\langle 1 \rangle$ with probability α_1 and in state $\langle 2\rangle$ with probability α_2:

$$\sigma^D_{\langle 1\rangle} = \alpha_1 \tag{9.22}$$
$$\sigma^D_{\langle 2\rangle} = \alpha_2 \tag{9.23}$$

The extended Markov chain is shown in Figure 9.5c. When the Markov chain is in state $\langle 1\rangle$ and there is an event, it goes to state $\langle \mathrm{E}\rangle$ with probability θ_i. With probability $(1-\theta_i)\cdot\alpha_2$ it goes to state $\langle 2\rangle$, and with probability $(1-\theta_i)\cdot\alpha_1$ it stays in state $\langle 1\rangle$. When the Markov chain is in state $\langle 2\rangle$ and there is an event, it goes to state $\langle \mathrm{E}\rangle$ with probability θ_i. With probability $(1-\theta_i)\cdot\alpha_1$ it goes to state $\langle 1\rangle$, and with probability $(1-\theta_i)\cdot\alpha_2$ it stays in state $\langle 2\rangle$. If the Markov chain goes to state $\langle \mathrm{E}\rangle$, there is an event in the component stream.

We calculate the complementary cumulative distribution function $\varphi_k(\cdot)$ of the time that the Markov chain needs to go from state k to state $\langle \mathrm{E}\rangle$:

$$\varphi_{\langle 1\rangle}(0) = \varphi_{\langle 2\rangle}(0) = 1 \tag{9.24}$$
$$\varphi_{\langle \mathrm{E}\rangle}(0) = 0 \tag{9.25}$$
$$\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \tag{9.26}$$

The cumulative distribution function of the interevent times of the component stream is

$$F_i(t) = 1 - \alpha_1\varphi_{\langle 1\rangle}(t) - \alpha_2\varphi_{\langle 2\rangle}(t) \tag{9.27}$$

Round-robin routing

Figure 9.6a shows the Markov chain for the state of the source stream. All transitions correspond to events in the source stream. Figure 9.6b shows the Markov chain for the state of the source stream and the round-robin scheduler. In this Markov chain, transitions from states $\langle \cdot/2\rangle$ to states $\langle \cdot/1\rangle$ correspond to events in the component process under consideration, transitions from states $\langle \cdot/1\rangle$ to states $\langle \cdot/2\rangle$ correspond to events in the other component process.

After an event in the component stream, the Markov chain is in state $\langle 1/1\rangle$ with probability α_1 and in state $\langle 2/1\rangle$ with probability α_2, so we have

$$\sigma^D_{\langle 1/1\rangle} = \alpha_1 \tag{9.28}$$
$$\sigma^D_{\langle 2/1\rangle} = \alpha_2 \tag{9.29}$$

In Figure 9.6c, the Markov chain for the calculation of the time to the next event is shown. The transitions from states $\langle \cdot/2\rangle$ to states $\langle \cdot/1\rangle$ now lead to state $\langle \mathrm{E}\rangle$. With this Markov chain, we calculate the complementary cumulative distribution function $\varphi_k(\cdot)$ of the time that the Markov chain needs to go from state k to state $\langle \mathrm{E}\rangle$:

$$\varphi_k(0) = \begin{cases} 1 & k \neq \langle \mathrm{E}\rangle \\ 0 & k = \langle \mathrm{E}\rangle \end{cases} \tag{9.30}$$
$$\varphi'(\tau) = \mathcal{Q}\cdot\varphi(\tau) \tag{9.31}$$

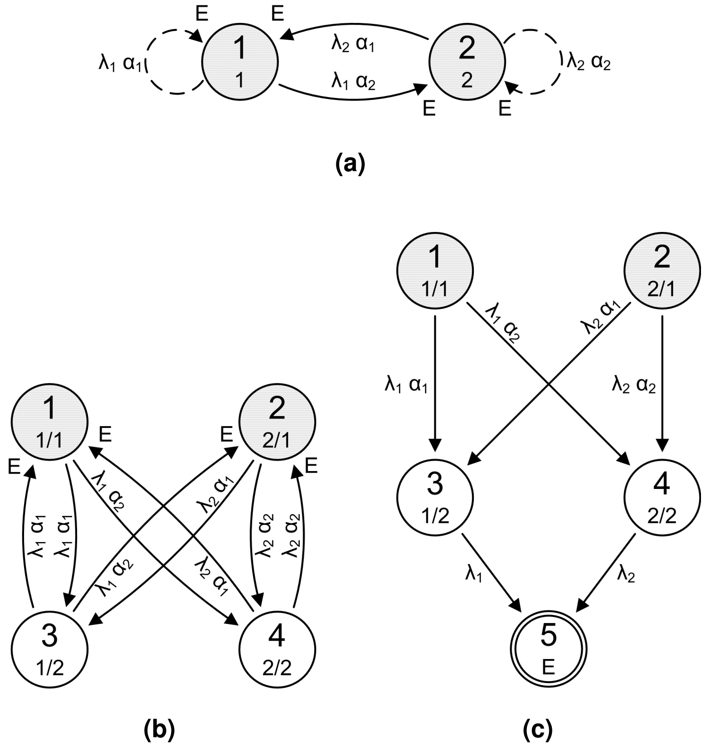


Figure 9.6.: Decomposition of a stream with hyperexponentially distributed interevent times (round-robin). (a) Markov chain for the state of the stream, (b) Markov chain for the state of the stream and the round-robin scheduler, (c) Markov chain for the calculation of the time to the next event. Meaning of the names of the states: (a) state of the stream, (b,c) state of the stream / state of the round-robin scheduler.

The cumulative distribution function of the interevent times of the component stream is

$$F_i(t) = 1 - \alpha_1 \varphi_{\langle 1/1 \rangle}(t) - \alpha_2 \varphi_{\langle 2/1 \rangle}(t) \tag{9.32}$$

9.1.3. Decomposition of a stream with exponentially distributed interevent times

Probability based routing

Figures 9.7a and 9.7b show the Markov chain for the system state of the source stream. The hidden transition corresponds to events in the source stream.

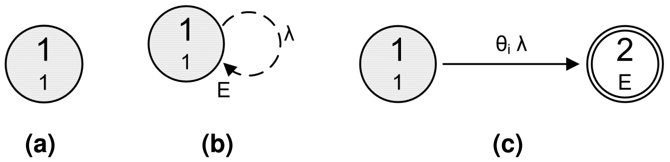


Figure 9.7.: Decomposition of a stream with exponentially distributed interevent times (probability based). (a,b) Markov chains for the state of the stream, (c) Markov chain for the calculation of the time to the next event.

After an event has occurred, the Markov chain is in state $\langle 1 \rangle$.

$$\sigma^D_{\langle 1 \rangle} = 1 \tag{9.33}$$

The extended Markov chain is shown in Figure 9.7c. When there is an event in the source stream, the Markov chain goes to state $\langle \mathrm{E} \rangle$ with probability θ_i or it stays in state $\langle 1 \rangle$ with probability $1 - \theta_i$.

The complementary cumulative distribution function of the time to the next event given that the Markov chain is in state k is calculated with

$$\varphi_k(0) = \begin{cases} 1 & k = \langle 1 \rangle \\ 0 & k = \langle \mathrm{E} \rangle \end{cases} \tag{9.34}$$

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{9.35}$$

which has the solution

$$\varphi_{\langle 1 \rangle}(t) = \mathrm{e}^{-\theta_i \lambda t} \tag{9.36}$$

$$\varphi_{\langle \mathrm{E} \rangle}(t) = 0 \tag{9.37}$$

The cumulative distribution function of the interevent times of the component stream S_i is

$$F_i(t) = 1 - \varphi_{\langle 1 \rangle}(t)\, \sigma^D_{\langle 1 \rangle} = 1 - \mathrm{e}^{-\theta_i \lambda t} \tag{9.38}$$

This means a probability based decomposition of a stream with exponentially distributed interevent times has exponentially distributed interevent times, too.

Some results are shown in Figure 9.8.

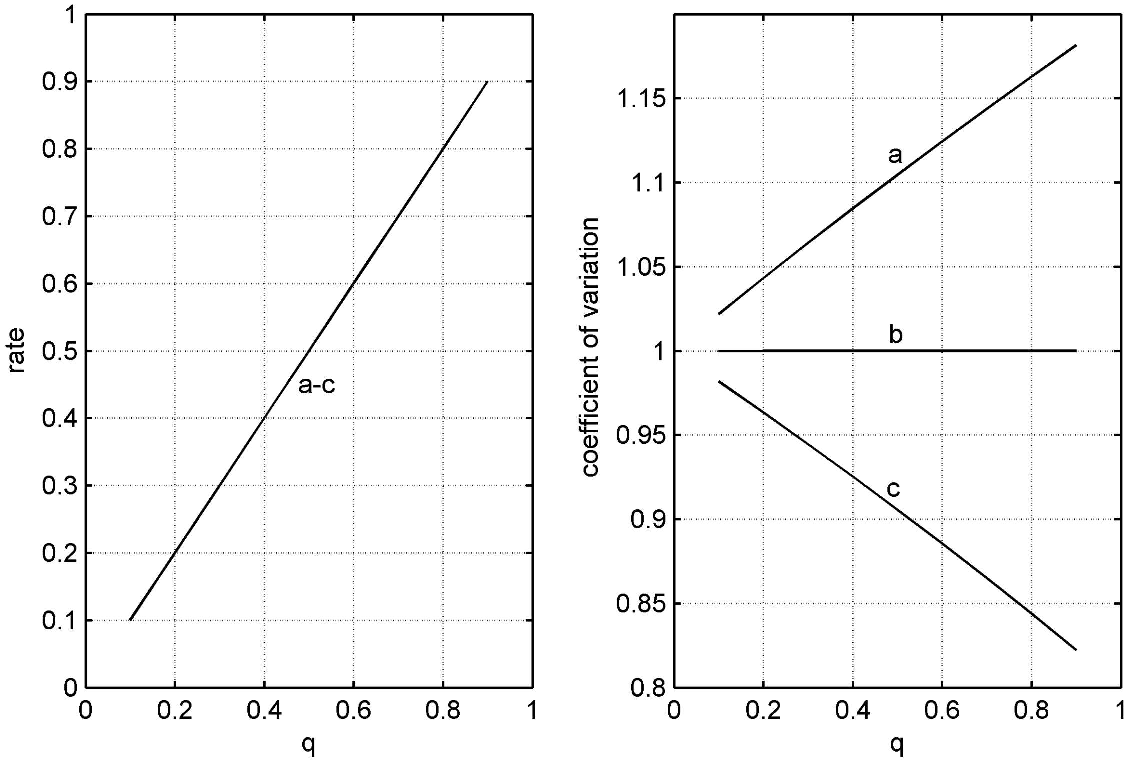


Figure 9.8.: Probability based decomposition of a traffic stream with (a) hyperexponentially ($c = 1.2$), (b) exponentially, (c) hypoexponentially ($c = 0.8$) distributed interevent times.

Round-robin routing

Figure 9.9a shows the Markov chain for the state of the source stream. The hidden transition corresponds to events in the source stream. Figure 9.9b shows the Markov chain for the state of the source stream and the round-robin scheduler. In this Markov chain, transitions from state $\langle 2\rangle$ to state $\langle 1\rangle$ correspond to events in the component process under consideration, and transitions from state $\langle 1\rangle$ to state $\langle 2\rangle$ correspond to events in the other component process.

After an event in the component stream under consideration, the Markov chain is in state $\langle 1\rangle$:

$$\sigma^{D}_{\langle 1\rangle} = 1 \tag{9.39}$$

In Figure 9.9c, the Markov chain for the calculation of the time to the next event is shown. The transition from state $\langle 2\rangle$ to state $\langle 1\rangle$ now leads to state $\langle \mathrm{E}\rangle$. With this Markov chain we calculate the complementary cumulative distribution function $\varphi_k(\cdot)$ of

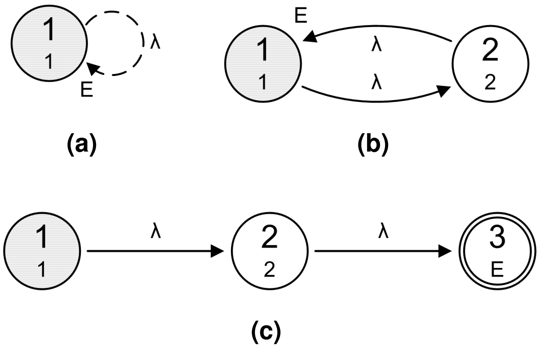


Figure 9.9.: Decomposition of a stream with exponentially distributed interevent times (round-robin). (a) Markov chain for the state of the stream, (b) Markov chain for the state of the stream and the round-robin scheduler, (c) Markov chain for the calculation of the time to the next event. Meaning of the names of the states: (a) state of the stream, (b,c) state of the round-robin scheduler.

the time that the Markov chain needs to go from state k to state $\langle\mathrm{E}\rangle$ with

$$\varphi_{\langle 1\rangle}(0) = \varphi_{\langle 2\rangle}(0) = 1 \tag{9.40}$$
$$\varphi_{\langle \mathrm{E}\rangle}(0) = 0 \tag{9.41}$$
$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{9.42}$$

which has the solution

$$\varphi_{\langle 1\rangle}(t) = (\lambda t + 1)\mathrm{e}^{-\lambda t} \tag{9.43}$$
$$\varphi_{\langle 2\rangle}(t) = \mathrm{e}^{-\lambda t} \tag{9.44}$$
$$\varphi_{\langle \mathrm{E}\rangle}(t) = 0 \tag{9.45}$$

The cumulative distribution function of the interevent times of the component stream is

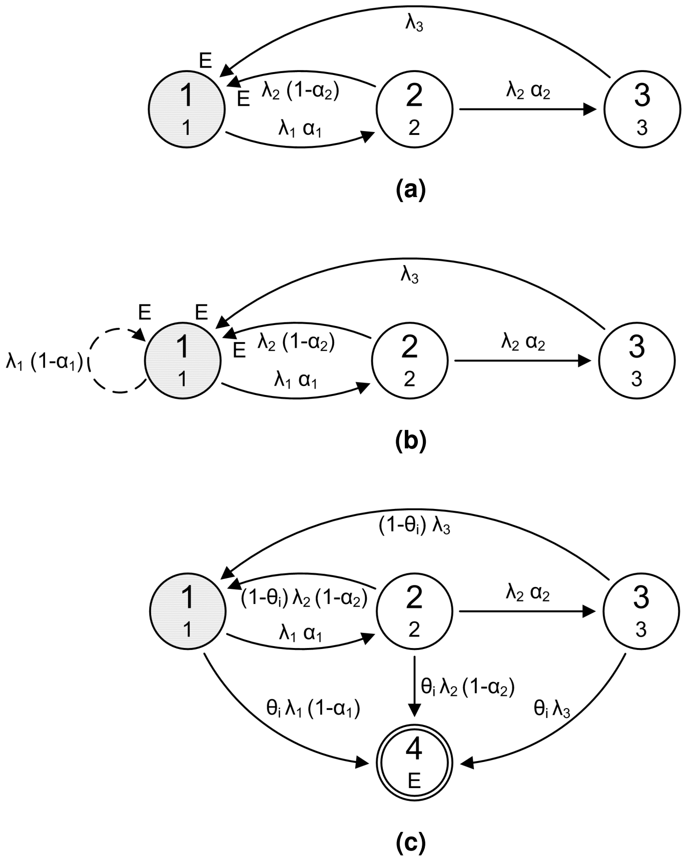
$$F_i(t) = 1 - \varphi_{\langle 1/1\rangle}(t) = 1 - (\lambda t + 1)\mathrm{e}^{-\lambda t} \tag{9.46}$$

which is, as expected, the cumulative distribution function of a 2-stage Erlang distribution.

9.1.4. Decomposition of a stream with Coxian distributed interevent times

Probability based routing

Figures 9.10a and 9.10b show the Markov chain for the system state of the source stream. All transitions to state $\langle 1 \rangle$ correspond to events in the source stream.



(a)

(b)

(c)

Figure 9.10.: Decomposition of a stream with Coxian distributed interevent times (probability based). (a,b) Markov chains for the state of the stream, (c) Markov chain for the calculation of the time to the next event.

After an event has occurred, the Markov chain is in state $\langle 1 \rangle$:

$$\sigma^D_{\langle 1 \rangle} = 1 \tag{9.47}$$

The extended Markov chain is shown in Figure 9.10c. We calculate the complementary cumulative distribution function $\varphi_k(\cdot)$ of the time that the Markov chain needs to go from state k to state $\langle \mathrm{E} \rangle$:

$$\varphi_{\langle 1 \rangle}(0) = \varphi_{\langle 2 \rangle}(0) = \varphi_{\langle 3 \rangle}(0) = 1 \tag{9.48}$$

$$\varphi_{\langle \mathrm{E} \rangle}(0) = 0 \tag{9.49}$$

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{9.50}$$

The cumulative distribution function of the interevent times of the component stream S_i is

$$F_i(t) = 1 - \varphi_{\langle 1 \rangle}(t) \tag{9.51}$$

Coxian* distribution

Probability based routing

The probability that a batch of n events occurs in the component stream under consideration when a single event or a batch of events occurs in the source stream equals the probability that there is a batch of size $m \geq n$ in the source stream (probability $(1-p)^{m-1}p$) and that exactly n of these events are routed to the component stream (probability $\theta_i^n(1-\theta_i)^{m-n}\binom{m}{n}$):

$$\mathrm{P}\left\{\text{batch of size } n\right\} = \sum_{m=n}^{\infty} (1-p)^{m-1} p\, \theta_i^n (1-\theta_i)^{m-n} \binom{m}{n} \tag{9.52}$$

The probability η that there is at least one event in the component stream after an event occurs in the source stream is the sum over all $n \geq 1$:

$$\eta = \mathrm{P}\left\{\text{batch of size} \geq 1\right\} = \sum_{n=1}^{\infty} \mathrm{P}\left\{\text{batch of size } n\right\} = \\ \sum_{n=1}^{\infty}\sum_{m=n}^{\infty} (1-p)^{m-1} p \theta_i^n (1-\theta_i)^{m-n} \binom{m}{n} = \frac{\theta_i}{p - \theta_i(p-1)} \tag{9.53}$$

If in the component stream a batch of events occurs, its expected size L is the expected size of a batch in the component stream given that exactly one event has occurred in the component stream:

$$L = \sum_{n=1}^{\infty} n \sum_{m=n}^{\infty} (1-p)^{m-1} p\, \theta_i^{n-1} (1-\theta_i)^{m-n} \binom{m-1}{n-1} = \frac{p + \theta_i - p\theta_i}{p} \tag{9.54}$$

The probability that the interevent time $X_{i,j}$ is 0 is the reciprocal of the expected batch size:

$$\mathrm{P}\left\{X_{i,j} = 0\right\} = \frac{p}{p + \theta_i - p\theta_i} \tag{9.55}$$

The Markov chain for the calculation of the interbatch times of the component stream S_i is shown in Figure 9.11. Events in the source stream (transitions $\langle 1 \rangle \to \langle 1 \rangle, \langle 2 \rangle \to \langle 1 \rangle$ and $\langle 3 \rangle \to \langle 1 \rangle$ in Figure 9.10b) cause events in the component stream (transitions $\langle 1 \rangle \to \langle \mathrm{E} \rangle, \langle 2 \rangle \to \langle \mathrm{E} \rangle$ and $\langle 3 \rangle \to \langle \mathrm{E} \rangle$) with probability η.

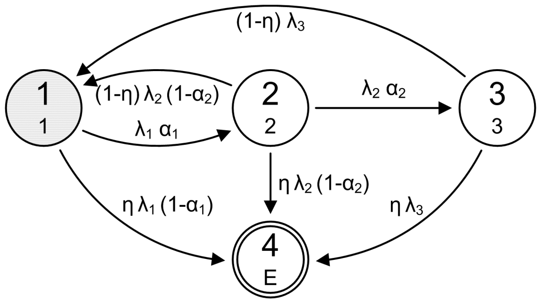


Figure 9.11.: Decomposition of a stream with Coxian* distributed interevent times: Markov chain for the calculation of the time to the next event.

The cumulative distribution function of the interbatch times Z is calculated with

$$\varphi_{\langle 1\rangle}(0) = \varphi_{\langle 2\rangle}(0) = \varphi_{\langle 3\rangle}(0) = 1 \tag{9.56}$$

$$\varphi_{\langle \mathrm{E}\rangle}(0) = 0 \tag{9.57}$$

$$\varphi'(\tau) = \mathcal{Q} \cdot \varphi(\tau) \tag{9.58}$$

and

$$F_Z(t) = 1 - \varphi_{\langle 1\rangle}(t) \tag{9.59}$$

The cumulative distribution function of the interevent times is

$$\mathrm{P}\left\{X_{i,j} \leq t\right\} = \mathrm{P}\left\{X_{i,j} = 0\right\} + F_Z(t)\left(1 - \frac{p}{p + \theta_i - p\theta_i}\right) \tag{9.60}$$

9.1.5. Closed-form solution

P. Kühn uses in [Kühn 1979] the following approach to determine a closed-form solution for the first two moments of the interevent times of a component stream (with probability based routing):

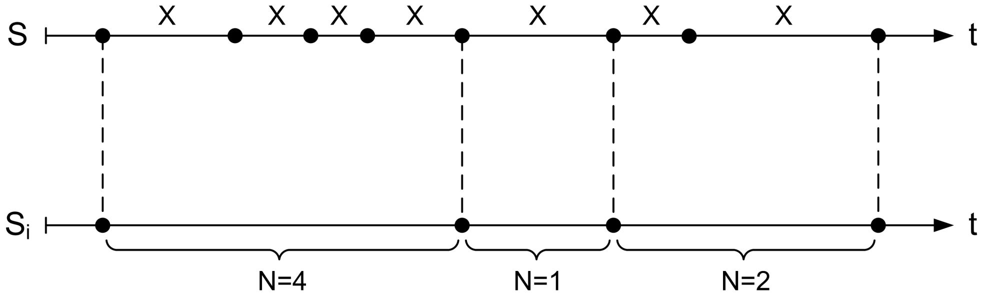


Figure 9.12.: Decomposition of a traffic stream. The interevent times of a component stream consist of N interevent times of the source stream.

As shown in Figure 9.12, an interevent time X_i in the component stream S_i consists of N interevent times X of the source stream S,

$$X_i = \sum_{n=1}^{N} X \tag{9.61}$$

The random variable N is geometrically distributed:

$$P_n = \mathrm{P}\left\{N = n\right\} = \begin{cases} 0 & \text{for } n = 0 \\ (1-\theta_i)^{n-1}\theta_i & \text{for } n \geq 1 \end{cases} \tag{9.62}$$

The sequence $\langle P_n \rangle$ has the generating function

$$G(z) = \sum_{n=0}^{\infty} P_n z^n = \frac{\theta_i z}{1-(1-\theta_i)z} \qquad |z| < 1 \tag{9.63}$$

Let

$$\Phi(s) = \int_{0-}^{\infty} \mathrm{e}^{-st} \mathrm{d}F(t) \tag{9.64}$$

and

$$\Phi_i(s) = \int_{0-}^{\infty} \mathrm{e}^{-st} \mathrm{d}F_i(t) \tag{9.65}$$

be the Laplace-Stieltjes transforms of the distribution functions of the interevent times of the source stream and the component streams, respectively.

The distribution function $F_i(t)$ can be expressed by a conditional cumulative distribution function $\mathrm{P}\{X_i < t \mid N = n\}$:

$$F_i(t) = \sum_{n=0}^{\infty} P_n \cdot \mathrm{P}\{X_i < t \mid N = n\} \tag{9.66}$$

The conditional cumulative distribution function $\mathrm{P}\{X_i < t \mid N = n\}$ is the distribution function of the sum of n independent and identically distributed random variables X that have the distribution function $F(t)$. The Laplace-Stieltjes transform of the cumulative distribution function of a sum of random variables is the product of the Laplace-Stieltjes transforms of the single cumulative distribution functions, so we have

$$\Phi_i(s) = \sum_{n=0}^{\infty} P_n \cdot (\Phi(s))^n \tag{9.67}$$

By comparing Equations 9.63 and 9.67 we find

$$\Phi_i(s) = G(\Phi(s)) = \frac{\theta_i \Phi(s)}{1 - (1-\theta_i)\Phi(s)} \tag{9.68}$$

For mean and variance of X_i we have

$$\mathrm{E}(X_i) = -\left.\frac{\mathrm{d}\Phi_i(s)}{\mathrm{d}s}\right|_{s=0} = \mathrm{E}(X)\,\mathrm{E}(N) \tag{9.69}$$

$$\mathrm{Var}(X_i) = \left.\frac{\mathrm{d}^2\Phi_i(s)}{\mathrm{d}s^2}\right|_{s=0} - (\mathrm{E}(X_i))^2 = (\mathrm{E}(X))^2\,\mathrm{Var}(N) + \mathrm{Var}(X)\,\mathrm{E}(N) \tag{9.70}$$

With

$$\mathrm{E}(N) = \frac{1}{\theta_i} \tag{9.71}$$

$$\mathrm{Var}(N) = \frac{1-\theta_i}{\theta_i^2} \tag{9.72}$$

$$\mathrm{E}(X) = \frac{1}{\lambda} \tag{9.73}$$

$$\mathrm{Var}(X) = \frac{c^2}{\lambda^2} \tag{9.74}$$

we get the final result

$$\lambda_j = \lambda \cdot \theta_i \tag{9.75}$$

$$c_i^2 = \theta_i c^2 + (1-\theta_i) \tag{9.76}$$

9.2. Modelling splitted-PH/M/1/S queueing systems

The decomposition of traffic streams can also be included in the Markov chain for a queueing system.

The Markov chain for the system state of a queueing system that receives only a portion of a traffic stream with phase-type distributed interevent times (we denote such queueing systems with %PH/. . . , e.g., %Hypo/M/1/S) can be constructed as follows:

Probability based

Let θ be the probability that a customer of the source stream is routed to the queueing system under consideration. We start with the Markov chain for the system state of the queueing system whereby we assume that the queueing system receives the whole traffic stream. Now for all transitions $i \to j$ that correspond to an arrival of a customer we set

$$q_{i,j}^{(\text{new})} = q_{i,j}^{(\text{old})} \cdot \theta \tag{9.77}$$

and we add a new transition from state i to a state k such that the arrival process is "reset" but the number of customers in the system remains unchanged. This transition corresponds to events in the source stream that are not routed to the queueing system under consideration. The rate of this new transition is

$$q_{i,k} = q_{i,j}^{(\text{old})} \cdot (1 - \theta) \tag{9.78}$$

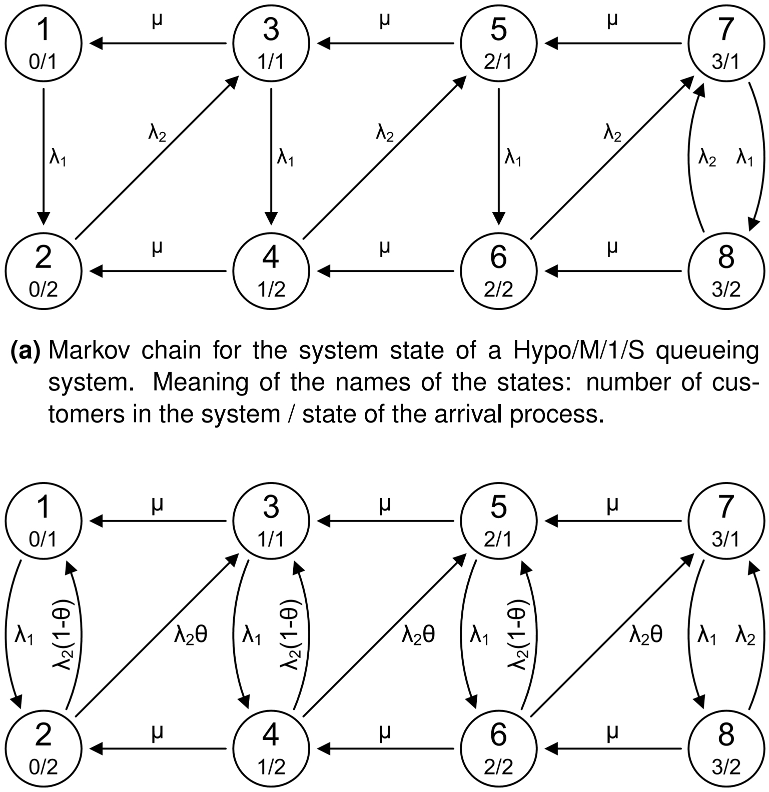
Round-robin

Again, the starting point is the Markov chain for the system state of the queueing system whereby we assume that the queueing system receives the whole traffic stream. Now we combine this Markov chain with a Markov chain that describes the state of the round-robin scheduler. If there are n component processes, this Markov chain has n states and transitions $1 \to 2, 2 \to 3, \ldots, n - 1 \to n, n \to 1$. A transition takes place whenever there is an event in the source stream. Now we modify the new Markov chain such that only when the scheduler is in state n, an event in the source stream increases the number of customers in the system.

9.2.1. Splitted-Hypo/M/1/S queueing system

Probability based routing

The construction of the Markov chain for the system state of a %Hypo/M/1/S queueing system with probability-based routing (routing probability θ) is shown in Figure 9.13. Figure 9.13a shows the Markov chain for the system state of a Hypo/M/1/S queueing system. The transitions $\langle n/2\rangle \to \langle n + 1/1\rangle$ correspond to arrivals of customers. We multiply the rates of these transitions by the routing probability θ and add new transitions $\langle n/2\rangle \to \langle n/1\rangle$ with rate $\lambda_2 \cdot (1 - \theta)$ (Figure 9.13b).



(a) Markov chain for the system state of a Hypo/M/1/S queueing system. Meaning of the names of the states: number of customers in the system / state of the arrival process.

(b) Markov chain for the system state of a %Hypo/M/1/S queueing system. Probability based routing (customers are received with probability θ).

Figure 9.13.: %Hypo/M/1/S queueing system.

Round-robin routing

The Markov chain for the system state of a %Hypo/M/1/S queueing system with round-robin routing (two component processes) is shown in Figure 9.14. Transitions that correspond to events in the source stream (all transitions with rate λ_2) change the state of the round-robin scheduler. If the scheduler is in state $n = 2$ when there is an event in the source stream, in addition the number of customers in the queueing system is increased (transitions $\langle n/2/2\rangle \rightarrow \langle n+1/1/1\rangle$).

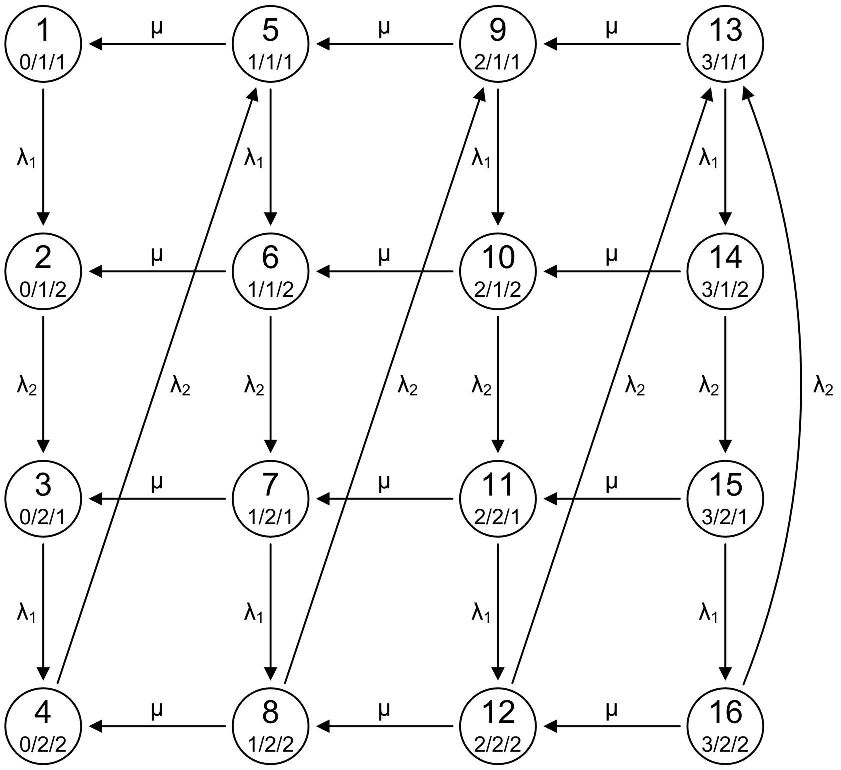
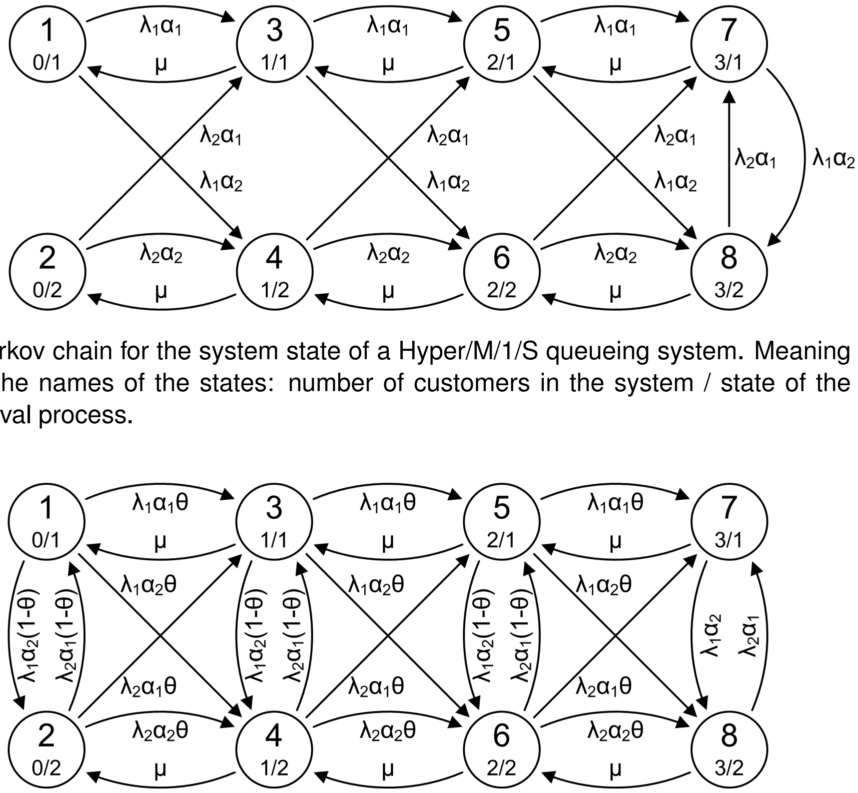


Figure 9.14.: %Hypo/M/1/S queueing system: Markov chain for the system state. Round-robin based routing (two component processes): every second customer of the arrival stream is received. Meaning of the names of the states: number of customers in the system / state of the round-robin scheduler / state of the arrival process.

9.2.2. Splitted-Hyper/M/1/S queueing system

Probability based routing

The construction of the Markov chain for the system state of a %Hyper/M/1/S queueing system with probability-based routing (routing probability θ) is shown in Figure 9.15. Figure 9.15a shows the Markov chain for the system state of a Hyper/M/1/S queueing system. The transitions $\langle n/\cdot\rangle \rightarrow \langle n+1/\cdot\rangle$ correspond to arrivals of customers. We multiply the rates of these transitions by the routing probability θ and add new transitions $\langle n/1\rangle \rightarrow \langle n/2\rangle$ with rate $\lambda_1\,\alpha_2\,(1-\theta)$ and $\langle n/2\rangle \rightarrow \langle n/1\rangle$ with rate $\lambda_2\,\alpha_1\,(1-\theta)$ (Figure 9.15b).



(a) Markov chain for the system state of a Hyper/M/1/S queueing system. Meaning of the names of the states: number of customers in the system / state of the arrival process.

(b) Markov chain for the system state of a %Hyper/M/1/S queueing system. Probability based routing (customers are received with probability θ).

Figure 9.15.: %Hyper/M/1/S queueing system.

Round-robin routing

The Markov chain for the system state of a %Hyper/M/1/S queueing system with round-robin routing (two component processes) is shown in Figure 9.16. Transitions that correspond to events in the source stream (all transitions containing λ_1 or λ_2) change the state of the round-robin scheduler. If the scheduler is in state $n = 2$ when there is an event in the source stream, in addition the number of customers in the queueing system is increased.

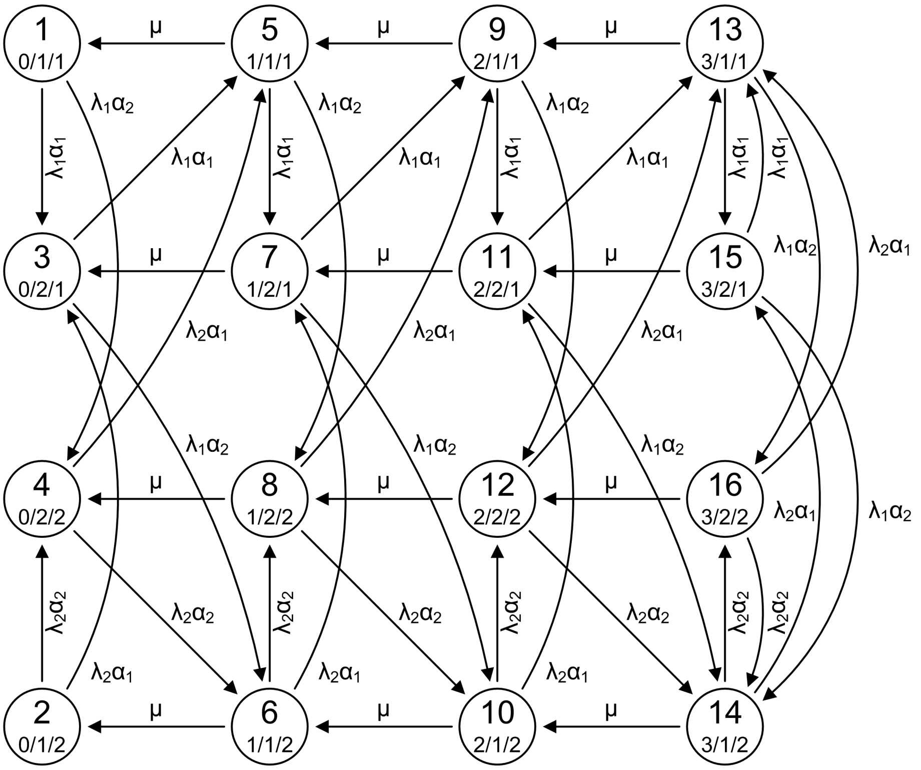


Figure 9.16.: %Hyper/M/1/S queueing system: Markov chain for the system state. Round-robin based routing (two component processes): every second customer of the arrival stream is received. Meaning of the names of the states: number of customers in the system / state of the round-robin scheduler / state of the arrival process.

Figure 9.17 shows a comparison between the expected number of customers in %PH/M/1/S queueing systems when round-robin routing or probability based routing is used.

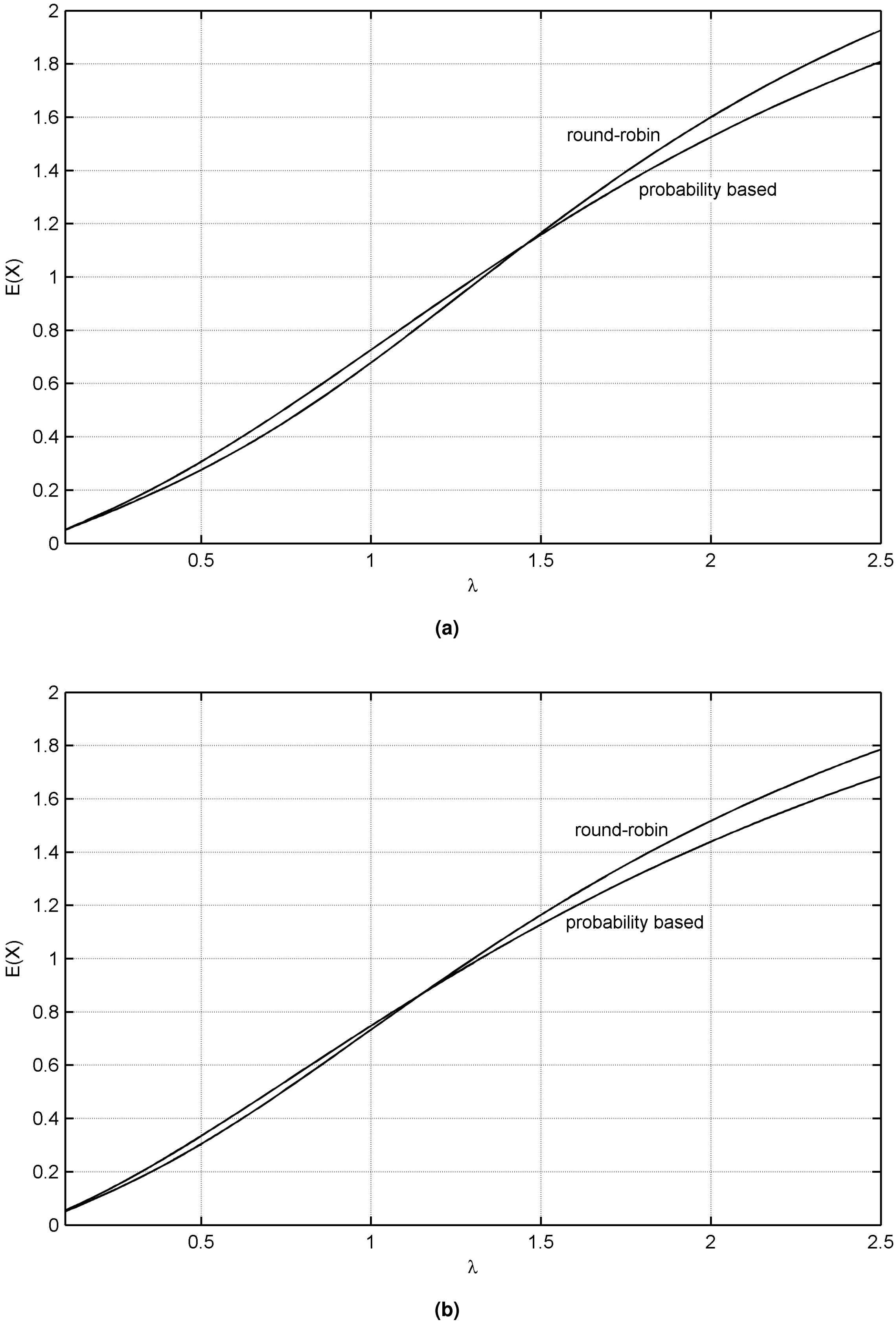


Figure 9.17.: Expected number of customers in a %PH/M/1/S queueing system $(S = 3)$ which receives 50% of the customers of a traffic stream with rate λ and (a) hypoexponentially $(c = 0.85)$, (b) hyperexponentially $(c = 1.5)$ distributed interevent times. Service rate $\mu = 1$. The routing policies are probability based routing $(\theta = 0.5)$ and round-robin routing (two component streams).

10. Summary

In this work, we show new techniques of modelling queueing systems with continuous-time Markov chains.

The first part gives an introduction to modelling of queueing systems with Markov chains. Chapter 2 reviews the theory of discrete-time and continuous-time Markov chains. In Chapter 3, traditional techniques of modelling queueing systems with Markov chains are presented: we show how to determine the system state, the flow time through a queueing system, and the time until a certain state is reached. In Chapter 4, we briefly discuss other performance evaluation techniques based on Markov chains (embedded Markov chain method, matrix geometric method, and matrix analytic methods).

In the second part of this work, the new Markov chain techniques are presented. Where possible, we give more than one method to determine the same quantity, so that one can choose the best method for the actual problem. Moreover, the techniques are presented in a form that makes it easy to implement them in a computer program.

Chapter 5 deals with the idle and the busy period of queueing systems. We show how to calculate the length of the idle period, the length of the busy period, and the number of customers served during a busy period. These quantities are used in Chapter 6, where the interdeparture times of single-server queueing systems are determined. In Chapter 7, we investigate overflow traffic, that is, traffic created by customers that are prevented from entering a queueing system and are redirected to another destination. We show how to calculate the interoverflow time and the number of successful arrivals between two consecutive rejections. Most of the techniques shown in Chapters 5 – 7 can be applied to every queueing system that can be modelled with continuous-time Markov chains, that is, no assumptions on the structure of the queueing systems are made (e.g., that arrival and service rates have to be state-independent). Finally, Chapters 8 and 9 deal with the manipulation of traffic streams. We present techniques to calculate the interevent times of both the superposition of traffic streams and the decomposition of traffic streams.

A. Mathematical symbols

c	coefficient of variation
$f_X(\cdot)$	probability density function of the random variable X
p_{ij}	transition probability from state i to state j (discrete-time Markov chain)
q_{ij}	transition rate from state i to state j (continuous-time Markov chain)
r	rate
t	time

A	interarrival time (random variable)
B	length of the busy period (random variable)
D	interdeparture time (random variable)
$D^{(1)}$	first interdeparture time in a busy cycle (random variable)
$D^{(2)}$	following interdeparture times in a busy cycle (random variable)
$F_X(\cdot)$	cumulative distribution function of the random variable X
$F_X^C(\cdot)$	complementary cumulative distribution function of the random variable X
I	length of the idle period (random variable)
R	interoverflow time (random variable)
S	service time (random variable)
X	number of customers in a queueing system (random variable)

$\mathcal{I}$	identity matrix
$\mathcal{M}_A$	Markov chain for the state of the superposition of point processes
$\mathcal{M}_B$	Markov chain for the calculation of the length of the busy period of a queueing system
$\mathcal{M}_C$	counting Markov chain

$\mathcal{M}_D$	Markov chain for the state of a point process and a component process (decomposition)
$\mathcal{M}_I$	Markov chain for the calculation of the length of the idle period of a queueing system
$\mathcal{M}_R$	Markov chain for the calculation of the length of the interblocking time of a queueing system
$\mathcal{M}_S$	Markov chain for the state of a queueing system
$\mathcal{M}_\zeta$	Markov chain for the calculation of the number of successful arrivals between two consecutive overflows
$\mathcal{M}_\xi$	Markov chain for the calculation of the number of customers served during a busy period
$\mathcal{M}_\Phi$	Markov chain for the calculation of the flow time
$\mathcal{P}$	transition probability matrix of a discrete-time Markov chain
$\mathcal{Q}$	transition rate matrix of a continuous-time Markov chain

α_i	parameters (probabilities) of hyperexponential and Coxian distributions
β_i	parameters (probabilities) of hyperexponential and Coxian distributions
ζ	number of successful arrivals between two consecutive overflows
ζ_n	probability that there are n successful arrivals between two consecutive overflows
θ	routing probability
κ	service rate
λ	arrival rate
λ_i	parameters (rates) of hypoexponential, hyperexponential, and Coxian distributions
μ	service rate
μ_i	parameters (rates) of hypoexponential, hyperexponential, and Coxian distributions
ξ	number of customers served during the busy period (random variable)
ξ_n	probability that there are n customers served during the busy period
π_i	stationary state probability of state i
$\pi_i(t)$	state probability of state i at time t

σ_i^A probability that the Markov chain $\mathcal{M}_A$ is in state i after an event has occured

σ_i^B probability that the Markov chain $\mathcal{M}_B$ is in state i when the busy period begins

σ_i^I probability that the Markov chain $\mathcal{M}_I$ is in state i when the idle period begins

σ_i^R probability that the Markov chain $\mathcal{M}_R$ is in state i after an overflow has occured

σ_i^Φ probability that the Markov chain $\mathcal{M}_\Phi$ is in state i when the flow process begins

τ time

$\varphi_i(\cdot)$ complementary cumulative distribution function of the time the Markov chain needs to reach an absorbing state given that it is in state i

Φ flow time (random variable)

$\sim$ “has the distribution”

$\mathbb{N}$ set of the natural numbers

$\mathbb{R}$ set of the real numbers

$\mathbb{R}_0^+$ set of the non-negative real numbers

Bibliography

[Allen 1978] Arnold O. Allen: *Probability, Statistics and Queueing Theory. With Computer Science Applications.* Academic Press, 1978.

[van As 1984] Harmen R. van As: *Modellierung und Analyse von Überlast-Abwehrmechanismen in Paketvermittlungsnetzen.* Universität Siegen, 1984.

[Augustin 1982] Reinhard Augustin, Klaus-Jürgen Büscher: *Characteristics of the COX-Distribution.* ACM SIGMETRICS Performance Evaluation Review (vol. 12, issue 1, pages 22-32), 1982.

[Bolch et al. 2006] Gunter Bolch, Stefan Greiner, Hermann de Meer, Kishor S. Trivedi: *Queueing Networks and Markov Chains, Second Edition.* Wiley, 2006.

[Brémaud 1999] Pierre Brémaud: *Markov Chains. Gibbs Fields, Monte Carlo Simulation, and Queues* Springer, 1999.

[Burke 1966] B. J. Burke, *The Output of a Queueing System.* Operations Research (vol. 4, pp. 699-706), 1966.

[Feldmann 1998] Anja Feldmann, Ward Whitt: *Fitting mixtures of exponentials to long-tail distributions to analyze network performance models.* Performance Evaluation (vol. 31, pages 245-279), 2001.

[Herzog/Woo/Chandy 1975] U. Herzog, L. Woo, K. M. Chandy: *Solution of Queueing Problems by a Recursive Technique.* IBM Journal of Research and Development (19, pages 295–300), 1975.

[Gross/Harris 1974] D. Gross, C.M. Harris: *Fundamentals of Queueing Theory.* Wiley, New York, 1974.

[Heath 1997] Michael T. Heath: *Scientific Computing. An Introductory Survey.* McGraw-Hill, 1997.

[van Hoorn 1983] Michiel H. van Hoorn: *Algorithms and Approximations for Queueing Systems.* Vrije Universiteit te Amsterdam, 1983.

[Hordijk/Tijms 1976] A. Hordijk, H. Tijms: *A Simple Proof of the Equivalence of the Limiting Distributions of the Continuous Time and the Embedded Process of the Queue Size in the M/G/1 Queue.* Statistica Neerlandica (volume 30, number 2, pages 97-100), 1976.

[Kleinrock 1975] Leonard Kleinrock: *Queueing Systems, Volume I: Theory.* Wiley, New York, 1975.

[Knuth 1997] Donald E. Knuth: *The Art of Computer Programming. Volume 2: Seminumerical Algorithms.* Addison-Wesley, 1997.

[Kühn 1972] Paul Kühn: *Über die Berechnung der Wartezeiten in Vermittlungs- und Rechnersystemen.* Universität Stuttgart, 1972.

[Kühn 1979] Paul Kühn: *Approximate Analysis of General Queuing Networks by Decomposition.* IEEE Transactions on Communications (vol. 27, no. 1, pages 113-126), 1972.

[Latouche/Ramaswami 1999] G. Latouche, V. Ramaswami: *Introduction to Matrix Analytic Methods in Stochastic Modeling.* ASA-SIAM Series on Statistics and Applied Probability 5, 1999.

[Lucantoni 1993] David M. Lucantoni: *The BMAP/G/1 Queue: A Tutorial.* Performance Evaluation of Computer and Communication Systems, volume 729, pp 330-358, 1993.

[Marie 1980] Raymond Marie: *Calculating Equilibrium Probabilities for* $\lambda(n)/C_k/1/N$ *queues.* Proceedings of Performance 1980 (pages 117-125), 1980.

[Nelson 1991] Randolph Nelson: *Matrix Geometric Solutions in Markov Models. A Mathematical Tutorial.* IBM Research Report RC 16777, 1991.

[Neuts 1981] Marcel F. Neuts: *Matrix-geometric solutions in stochastic models.* John Hopkins University Press, 1981.

[Neuts 1989] Marcel F. Neuts: *Structured stochastic matrices of M/G/1 type and their applications.* Marcel Dekker, New York, 1989.

[Osogami 2005] Takayuki Osogami: *Analysis of Multi-server Systems via Dimensionality Reduction of Markov Chains.* Carnegie Mellon University, 2006.

[Press et al. 1992] William H. Press, Saul A. Teukolsky, William T. Vetterling, Brian P. Flannery: *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, 1992.

[Riska/Smirni 2002] A. Riska, E. Smirni: *M/G/1-type Markov Processes: A tutorial.* Performance Evaluation of Complex Computer Systems: Techniques and Tools, LNCS vol. 2549 pp. 36-63, Springer Verlag, 2002.

[Ross 2007] Sheldon M. Ross: *Introduction to Probability Models, 9th edition.* Academic Press, 2007.

[Sasaki 2001] Yukie Sasaki, Hiroei Imai, Masahiro Tsunoyama, Ikuo Ishii: *Approximation Method for Probability Distribution Functions using Cox Distribution to Evaluate Multimedia Systems.* Dependable Computing, 2001. Proceedings. 2001 Pacific Rim International Symposium on, pp. 333-340, 2001.

[Sauer 1975] C. H. Sauer, K. M. Chandy: *Approximate Analysis of Central Server Models.* IBM Journal of Research and Development (19:301–313), 1975.

[Sommereder 2008] Markus Sommereder: *Modellierung von Warteschlangensystemen mit Markov-Ketten.* VDM Verlag Dr. Müller, 2008.

[Stewart 1994] William J. Stewart: *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, 1994.

[Tijms 1986] Henk C. Tijms: *Stochastic Modelling and Analysis: A Computational Approach.* Wiley, 1986.

[Taylor/Karlin 1994] Howard M. Taylor, Samuel Karlin: *An Introduction to Stochastic Modeling.* Academic Press, 1994.

[Watkins 2002] David S. Watkins: *Fundamentals of Matrix Computations. Second Edition.* Wiley, 2002.

[Whitt 1982] Ward Whitt: *Approximating a Point Process by a Renewal Process, I: Two Basic Methods.* Operations Research (vol. 30, no. 1, pages 125-147), 1982.