Queuing Networks in Healthcare Systems

Maartje E. Zonderland and Richard J. Boucherie

Abstract Over the last decades, the concept of patient flow has received an increased amount of attention. Healthcare professionals have become aware that in order to analyze the performance of a single healthcare facility, its relationship with other healthcare facilities should also be taken into account. A natural choice for analysis of networks of healthcare facilities is queuing theory. With a queuing network a fast and flexible analysis is provided that discovers bottlenecks and allows for the evaluation of alternative set-ups of the network. In this chapter we describe how queuing theory, and networks of queues in particular, can be invoked to model, study, analyze and solve healthcare problems. We describe important theoretical queuing results, give a review of the literature on the topic, discuss in detail two examples of how a healthcare problem is analyzed using a queuing network, and suggest directions for future research.

1 Introduction

With an aging population, the rising cost of new medical technologies, and a society wanting higher quality care, the demand for healthcare is increasing annually. In European countries, such as the Netherlands, healthcare expenditures consume around 10% of the GDP. In the United States this percentage is even bigger at 16% [45] (2008 data). Since the supply of healthcare is finite, policy makers have to ration care and make choices on how to distribute physical, human, and monetary resources. Such choices also have to be made at the hospital level (e.g., which pa-

Maartje E. Zonderland · Richard J. Boucherie
Stochastic Operations Research & Center for Healthcare Operations Improvement and Research, University of Twente, Postbox 217, 7500 AE Enschede, the Netherlands, e-mail: m.e.zonderland@utwente.nl · r.j.boucherie@utwente.nl

Maartje E. Zonderland
Division I, Leiden University Medical Center, Postbox 9600, 2300 RC Leiden, the Netherlands

tient groups will be treated in this hospital), and on a departmental level (e.g., who gets which available bed).

An immediate consequence of rationing resources is the evolution of queues. This brings us immediately to queuing theory, which is the mathematical theory that studies queues. The methods available in this field can support healthcare professionals in their decision making. Already in 1952, Bailey recognized that queuing theory would be of value to make a trade-off between patient waiting time and healthcare provider idle time: short waiting time means a low provider utilization rate, while low provider idle time results in long waiting times [7]. With queuing theory a balance between these two performance measures can be found. Another example is calculating the required number of beds on a nursing ward that ensures the patient rejection rate stays below a certain threshold [13]. Finally, consider an example from the operating room (OR), where a queuing model can be used to find the optimal amount of OR time to allocate to semi-urgent patients. A surplus of allocated OR time results in an empty OR (a waste of resources), while a shortage will result in elective patients that need to be canceled to accommodate the semi-urgent patients. The challenge is to find a balance [62]. The book chapter by Linda Green [29] provides an overview of queuing theory applications in healthcare.

1.1 Some General Queuing Concepts

A queue can generally be characterized by its arrival and service processes, the number of servers, and the service discipline. The arrival process is specified by a probability distribution that has an arrival rate associated with it, which is usually the mean number of patients that arrives during a time unit (e.g., minutes, hours or days). A common choice for the probabilistic arrival process is the Poisson process, in which the inter-arrival times of patients are independent and exponentially distributed.

The service process specifies the service requirements of patients, again using a probability distribution with associated service rate. A common choice is the exponential distribution, which is convenient for obtaining analytical tractable results. The number of servers in a healthcare setting may represent the number of doctors at an outpatient clinic, the number of MRI scanners at a diagnostic department, and so on. The service discipline specifies how incoming patients are served. The most common discipline is First Come First Serve (FCFS), where patients are served in order of arrival. Other examples are briefly addressed in Subsection 2.2.5. Some patients may have priority over other patients (see Subsection 2.2.6). This can be such that the service of a lower priority patient is interrupted when a higher priority patient arrives (preemptive priority), or the service of the lower priority patient is finished first (non-preemptive priority).

Typical measures for the performance of the system include the mean sojourn time, $\mathbb{E}[W]$, the mean time that a patient spends in the queue and in service. The sojourn time is a random variable as it is determined by the stochastic arrival and

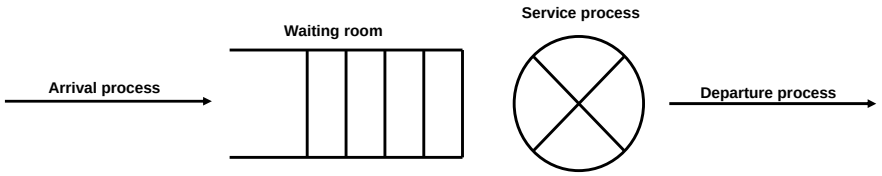

Fig. 1: A simple queue

service processes. The mean waiting time, $\mathbb{E}[W^q]$, gives the mean time a patient spends in the queue waiting for service. How $\mathbb{E}[W]$ and $\mathbb{E}[W^q]$ are calculated depends, among other things, on the choice for the arrival and service processes, and is given for several basic queues in Subsection 2.2.

Kendall's Notation

All queues in this chapter are described using the so-called Kendall notation: $A/B/s$, where **A** denotes the arrival process, **B** denotes the service process, and **s** is the number of servers. There are several extensions to this notation, see for example [56]. Clearly, there are many distinctive cases of queues:

$M/M/1$: The single-server queue with Poisson arrivals and exponential service times. The M stands for the Markovian or Memoryless property.
$M/D/1$: The single-server queue with Poisson arrivals and Deterministic service times.
$M/G/1$: The single-server queue with Poisson arrivals and General (i.e., not specified) service time distribution.

Other arrival processes may also apply: consider for example the $D/M/1$, $G/M/1$ and $G/G/1$ queue. All of the forms above also exist in the case of multiple servers ($s > 1$).

The load of the queue is defined as the mean utilization rate per server, which is the amount of work that arrives on average per time unit, divided by the amount of work the queue can handle on average per time unit. Suppose our server is a single doctor in an outpatient clinic, then the load specifies the fraction of time the doctor is working. The load, ρ, equals the amount of work brought to the system per time unit, i.e. the patient arrival rate, λ, multiplied by the mean service time per patient, $\mathbb{E}[S]$:

$$\rho = \lambda \mathbb{E}[S]. \tag{1}$$

The load is the fraction of time the server, working at unit rate, must work to handle the arriving amount of work. It is required that $\rho < 1$ (in other words, the server should work less than 100 percent of the time). If $\rho > 1$, then on average more work

arrives at the queue than can be handled, which inevitably leads to a continuously growing number of patients in the queue waiting for service, i.e., an unstable system. Only when the arrival and service processes are deterministic (i.e., the inter-arrival and service times have zero variance), the load may equal 1. The mean waiting time, $\mathbb{E}[W^q]$, increases with load ρ. As an illustration, consider a single-server queue with Poisson arrivals and general service times (the so-called $M/G/1$ queue), with mean $\mathbb{E}[S]$ and squared coefficient of variation (scv) c_S^2, which is calculated by dividing the variance by the squared mean. For this queue, the relationship between ρ and $\mathbb{E}[W^q]$ is characterized by the Pollaczek-Khinchine formula [21]:

$$\mathbb{E}[W^q] = \mathbb{E}[S]\frac{\rho}{1-\rho}\frac{1+c_S^2}{2}, \tag{2}$$

In Figure 2 the relation is shown graphically for $c_S^2 = 1$. We see that the mean wait-

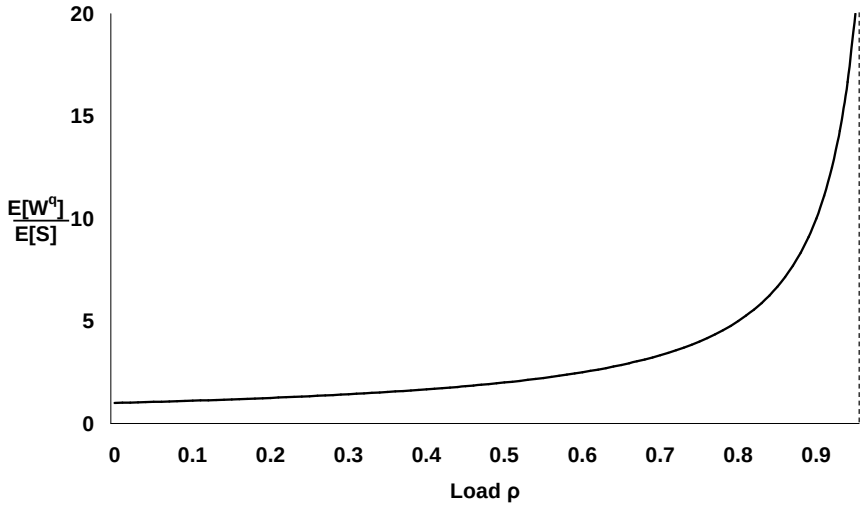

Fig. 2: The relationship between load ρ and mean waiting time $\mathbb{E}[W^q]$ for the $M/M/1$ queue with Poisson arrivals and exponential service times

ing time increases with the load. When the load is low, a small increase therein has a minimal effect on the mean waiting time. However, when the load is high, a small increase has a tremendous effect on the mean waiting time. As an illustration, increasing the load from 50% to 55% increases the waiting time by 10%, but increasing the load from 90% to 95% increases the waiting time by 100%! This explains why a minor change (for example a small increase in the number of patients) can result in a major increase in waiting times as sometimes seen in outpatient clinics. Formulas such as (2) allow for an exact and fast quantification of the relationships

between (influencable) parameters and system outcomes. Queuing theory is a very valuable tool to identify bottlenecks and to calculate the effect of removing them.

We conclude this subsection with a basic queuing network: the $M/M/1$ tandem queue. In this network we have two queues with exponential service, which are placed in series. Patients arrive at the first queue according to a Poisson process with rate λ. When the service at the first queue is completed, the patient is routed immediately to the second queue. Upon service completion at this queue, the patient leaves the system. At both queues the service discipline is FCFS, and there is an infinite waiting room (see Figure 3). It can be shown that the mean sojourn time in

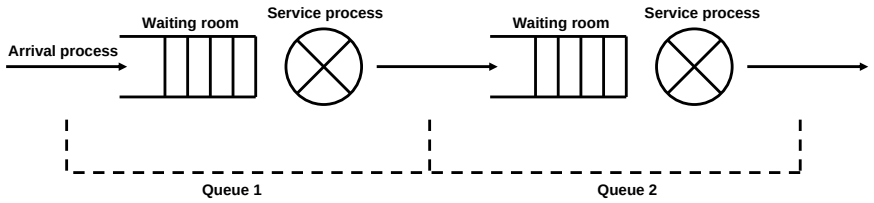

Fig. 3: The $M/M/1$ tandem queue

the entire system, $\mathbb{E}[W]$, is just the sum of the mean sojourn times in the individual queues, $\mathbb{E}[W_j]$ for queue j:

$$\mathbb{E}[W] = \mathbb{E}[W_1] + \mathbb{E}[W_2], \tag{3}$$

since the departure process from each queue has the same characteristics as its input process. This remarkable result can be generalized to larger networks of queues, as is shown in Subsection 2.3.2.

1.2 Queuing Networks in Healthcare

When patients share and use multiple resources, a queuing network usually arises. Consider, for example, a patient that visits the Orthopedic outpatient clinic and then needs to have an X-ray at Radiology; or the surgical patient who is operated in the OR, then cared for at the Intensive Care Unit (ICU) and subsequently cared for in a nursing ward. The formulation and analysis of these queuing network models is usually not straightforward. This likely explains why (discrete-event) simulation [41] is a commonly used approach to analyze healthcare problems. Simulation models are robust in terms of the setting they can represent, however they are very time consuming to develop and require a vast amount of data (-analysis). Also, the resulting model is, with a few exceptions, not generic and thus not suitable to represent other problems or organizations other than the one it was build for.

In this chapter we describe how queuing theory, and networks of queues in particular, can be invoked to study, analyze and solve healthcare problems. In Section 2 we provide an introduction to the theory of queues and queuing networks. In Section 3 we give a review of the literature on the topic, and discuss in detail two examples of how healthcare problems are analyzed using queuing networks. In the last section we suggest directions for further research. Given the numerous modeling opportunities of queuing networks, many difficult healthcare problems can, and hopefully will, be solved in the future. The literature references on applications of queuing theory in healthcare are included in the categorized ORchestra bibliography [46], provided by research institute CHOIR from the University of Twente, Enschede, the Netherlands.

2 Basic Queuing Networks

In this section we discuss several basic queuing networks. We start by introducing the Poisson process, which is a basic element in many queuing systems. We then proceed to the building blocks for the networks: the individual queues. We conclude by describing various queuing networks.

2.1 The Poisson Process

As mentioned in Subsection 1.1, the Poisson process is commonly used to model the arrival of customers to a queue, and in general to model independent arrivals from a large population. As an example, consider patient arrivals at a hospital emergency department (ED). They originate from a large population (the demographic area surrounding the hospital) and usually arrive independently. The probability that an arbitrary person has an urgent medical problem is very small. Then it can be shown that the arrival process tends to a Poisson process [13].

The Poisson process is common in real world processes and has many interesting and for analysis very useful properties. For example, the number of ticks a Geiger counter records is a Poisson process. This example also indicates that merging or splitting Poisson processes independently results in Poisson processes, as this corresponds to joining two lumps of radioactive material or breaking one lump into parts. Or, for the population example, ED arrivals from a population subgroup (men, women, children, ...) are also Poisson.

For a Poisson process, the time between two successive arrivals is exponentially distributed [57]. A very important property of the exponential distribution is that it is memoryless: the probability that the inter-arrival time exceeds $u+t$ time units, given that it already has exceeded u time units, equals the probability that the inter-arrival time exceeds t time units. Mathematically, a random variable X that has an exponential distribution satisfies:

$$\mathbb{P}(X > u+t|X > u) = \mathbb{P}(X > t), \quad \forall u,t \geq 0. \tag{4}$$

We may also rephrase this property as: what happens in the future is independent of what happened in the past. Because of this Markovian or memoryless property, the complexity of analyzing systems with this property significantly reduces, as we show in the subsequent subsections.

2.2 Basic Queues

We introduce the most commonly used queues: single and multi-server queues with Poisson arrivals and exponential or general service times. Unless mentioned otherwise, we consider the FCFS service discipline and queues with infinite capacity for waiting patients.

2.2.1 The $M/M/1$ Queue

In an $M/M/1$ queue, patients arrive according to a Poisson process with rate λ and exponentially distributed service requirement with mean service time $\mathbb{E}[S]$. The service rate per unit time is $\mu = \frac{1}{\mathbb{E}[S]}$, the number of patients that would be completed per time unit when the system would continuously be serving patients. As denoted in Section 1.1, the load of the queue is $\rho = \lambda\mathbb{E}[S]$, where it is required that $\rho < 1$, that is, the amount of work brought into the queue should be less than the rate of the server. The number of patients present in the queue at time t, i.e., those waiting in line and in service, is obtained from Markov chain analysis.

Let $N(t)$ record the number of patients in the system at time t. Then $N = (N(t), t \geq 0)$ is a Markov chain with state space $\mathbb{N}_0 = \{0,1,2,\ldots\}$, arrival rate λ, which is the rate at which a transition occurs from a state with n patients to a state with $n+1$ patients, and departure rate μ from state n to state $n-1$. We are interested

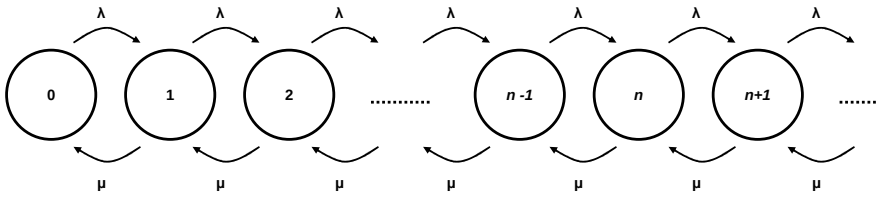

Fig. 4: Transition rates in the $M/M/1$ queue

in the probability P_n that at an arbitrary point in time in statistical equilibrium the

system contains n patients[1]:

$$P_n = \lim_{t\to\infty} \mathbb{P}(N(t) = n). \tag{5}$$

The probability P_n also reflects the fraction of time that the system contains n patients. The total probability may be seen as an amount of fluid of total volume 1 that is distributed over the states of the Markov chain and flows from state to state according to the transition rates (for the $M/M/1$ queue the arrival and departure rates). The system is in statistical equilibrium when these flows out of state n balance the flows into state n for each state n, $n = 0,1,2,\ldots$ (see Figure 4). Mathematically, this is expressed as:

$$\begin{aligned} \lambda P_0 &= \mu P_1, \\ (\lambda+\mu)P_1 &= \lambda P_0 + \mu P_2, \\ (\lambda+\mu)P_2 &= \lambda P_1 + \mu P_3, \\ &\vdots \end{aligned} \tag{6}$$

and in general:

$$\begin{aligned} \lambda P_0 &= \mu P_1, \\ (\lambda+\mu)P_n &= \lambda P_{n-1} + \mu P_{n+1} \quad \text{for} \quad n > 0. \end{aligned} \tag{7}$$

Since P_n is a probability, the summation of all probabilities P_n, $n = 0,1,\ldots$, should equal unity:

$$\sum_{n=0}^{\infty} P_n = 1. \tag{8}$$

Using equation (7) and this additional property, we derive the queue length distribution P_n:

$$\begin{aligned} P_0 &= 1-\rho, \\ P_n &= (1-\rho)\rho^n \quad \text{for} \quad n > 0. \end{aligned} \tag{9}$$

Note that P_0, also called the normalization constant, denotes the probability that there are zero patients present, but also the fraction of time the queue is empty. Further, ρ is the probability there are one or more patients present, and the fraction

[1] We consider the system in statistical equilibrium only, as is customary in queuing theory. For the $M/M/1$ queue, relaxation or convergence to equilibrium usually occurs fast. See [28] for a discussion on the validity of equilibrium analysis.

of time the queue is busy.

The PASTA Property

In a queuing system with Poisson arrivals, the probability that an arriving patient finds n patients in the queue is equal to the fraction of time the queue contains n patients. This property is referred to as PASTA, or Poisson Arrivals See Time Averages [57].

Usually, queuing systems with non-Poisson arrival processes do not conform to this property. For example, consider the $D/D/1$ queue with deterministic inter-arrival and service times. Time is equally distributed in slots of length one, and the service time is half a slot. Suppose that at the start of each time slot a patient arrives (so the inter-arrival time is one slot). Then the queue is empty upon arrival for all patients, while half of the time the queue contains one patient.

The mean number of patients in the queue, $\mathbb{E}[L]$, including those in service, is given by:

$$\mathbb{E}[L] = \sum_{n=0}^{\infty} nP_n = \frac{\rho}{1-\rho}. \tag{10}$$

Since ρ is the mean utilization rate of the server, the mean number of patients waiting, $\mathbb{E}[L^q]$, equals:

$$\mathbb{E}[L^q] = \frac{\rho}{1-\rho} - \rho = \frac{\rho^2}{1-\rho}. \tag{11}$$

Using a basic result in queuing theory, known as Little's Law, the relationship between the mean number of patients in the queue, $\mathbb{E}[L]$, and the mean sojourn time, $\mathbb{E}[W]$, can be explicitly quantified as follows [43]:

$$\mathbb{E}[L] = \lambda \mathbb{E}[W]. \tag{12}$$

This also holds for the relationship between the mean number of patients waiting for service, $\mathbb{E}[L^q]$, and the mean waiting time in the queue, $\mathbb{E}[W^q]$:

$$\mathbb{E}[L^q] = \lambda \mathbb{E}[W^q]. \tag{13}$$

Note that the equilibrium distribution and performance measures are characterized by the single parameter ρ and can be calculated in a straightforward manner. As we will see in the subsequent subsections, this is more involved for more complicated

queuing systems.

Little's Law

The simple relationship $\mathbb{E}[L] = \lambda \mathbb{E}[W]$, presented in 1961 by J.D.C. Little [43], is known as Little's Law. It relates the mean number of patients in the queue, $\mathbb{E}[L]$, the average arrival rate, λ, and the mean time the patient spends in the queue, $\mathbb{E}[W]$.

A common intuitive reasoning for obtaining Little's Law is the following. Suppose patients pay 1 Euro for each time unit they spend in the queue. On average, the queue receives $\mathbb{E}[L]$ Euro per time unit, since there are on average $\mathbb{E}[L]$ patients present in the queue. Alternatively, if each patient would pay upon entering the queue for its entire time spent in the queue, a patient would on average have to pay $\mathbb{E}[W]$ to finance the entire stay. Since each time unit on average λ patients enter the queue, the amount received by the queue per time unit then equals $\lambda \mathbb{E}[W]$. Both methods of payment must result in the same benefit for the queue, thus $\mathbb{E}[L] = \lambda \mathbb{E}[W]$. The formal proof actually follows the lines of this reasoning. It is remarkable that Little's Law requires only mild assumptions on the system in equilibrium, and is valid irrespective of the number of servers, distribution of the arrival and service processes, queuing and service order. Thus Little's Law applies to many types of queues.

2.2.2 The $M/M/s$ Queue

The $M/M/s$ queue is the multi-server variant of the $M/M/1$ queue. Patients arrive with rate λ, each patient is served by one server and a patient waits in queue when all servers are occupied. There are s servers so that the maximum service rate of the queue is $s\mu$, where μ is the service rate of the individual servers. If the number of patients in the queue, n, is less than the number of servers, s, the service rate equals $n\mu$ (see the transition rate diagram in Figure 5). Again it is required that the amount

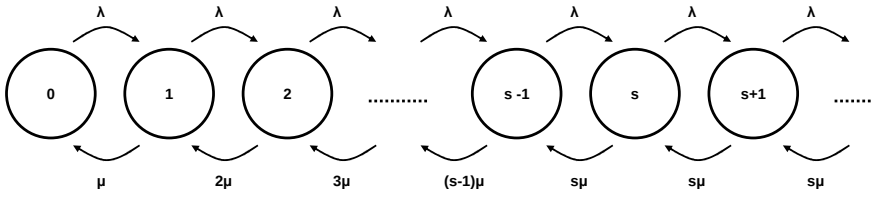

Fig. 5: Transition rates in the $M/M/s$ queue

of work that arrives per time unit (ρ) is less than the maximum service rate, i.e.,

$\rho = \lambda \mathbb{E}[S] < s$. The equilibrium distribution is obtained from:

$$\begin{array}{rll} \lambda P_0 = \mu P_1, & & \\ (\lambda + n\mu)P_n = \lambda P_{n-1} + (n+1)\mu P_{n+1} & \text{for} & n < s, \\ (\lambda + s\mu)P_n = \lambda P_{n-1} + s\mu P_{n+1} & \text{for} & n \geq s. \end{array} \tag{14}$$

Thus

$$P_n = \frac{\rho^n}{m(n)} P_0, \tag{15}$$

where

$$m(n) = \begin{cases} n! & \text{for} \quad 0 \leq n < s, \\ s^{n-s} s! & \text{for} \quad n \geq s. \end{cases} \tag{16}$$

Invoking the normalization condition (8), we obtain:

$$P_0 = \left(\sum_{n=0}^{s-1} \frac{\rho^n}{n!} + \frac{\rho^s}{s!} \frac{s}{s-\rho} \right)^{-1}. \tag{17}$$

For $s = 1$, equations (15)–(17) reduce to the queue length distribution for the $M/M/1$ queue (9). The probability P_s deserves special attention; this is the fraction of time all servers are occupied, and because of the PASTA property, also the fraction of arriving patients that find all servers occupied. Thus the probability that a patient will be served immediately upon arrival is $1 - \sum_{n=s}^{\infty} P_n = \sum_{n=0}^{s-1} P_n$, and the probability that a patient has to wait is $\sum_{n=s}^{\infty} P_n$. The latter probability can be calculated using the Erlang-C formula [31]:

$$P_{s^+} = \mathbb{P}(n \geq s) = \frac{\rho^s}{s!} \frac{s}{s-\rho} P_0. \tag{18}$$

There are several Erlang-C calculators available online to compute P_{s^+}, see e.g. [26] and [54]. The mean number of patients waiting for service is:

$$\mathbb{E}[L^q] = \sum_{n=s+1}^{\infty} (n-s) P_n = \frac{\rho}{s-\rho} P_{s^+}. \tag{19}$$

By applying Little's Law we find the mean waiting time:

$$\mathbb{E}[W^q] = \frac{\mathbb{E}[L^q]}{\lambda}. \tag{20}$$

The mean sojourn time is then obtained by adding the mean service time to the mean waiting time:

$$\mathbb{E}[W] = \mathbb{E}[S] + \mathbb{E}[W^q]. \tag{21}$$

The mean number of patients in the queue can be calculated by adding the mean number of patients in service, ρ, to the mean number of patients waiting [31]:

$$\mathbb{E}[L] = \rho + \mathbb{E}[L^q]. \tag{22}$$

2.2.3 The $M/M/s/s$ Queue

The $M/M/s/s$ queue, or Erlang loss queue, is different from the $M/M/s$ queue in that it has no waiting capacity. Thus when all servers are occupied, patients are blocked and lost (i.e., they leave and do not come back). This type of queue is very useful when modeling healthcare systems with limited capacity, where patients are routed to another facility when all capacity is in use. Examples are nursing wards and the ICU. Figure 6 gives the transition rates for this queue. We obtain:

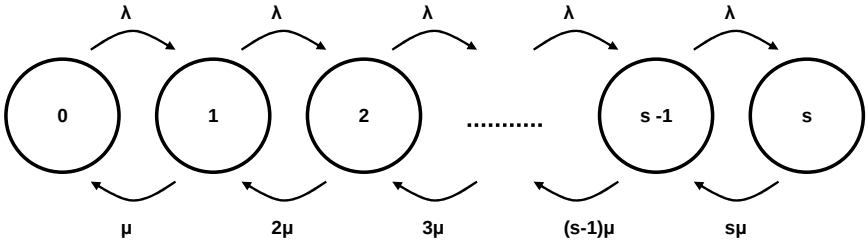

Fig. 6: Transition rates in the $M/M/s/s$ queue

$$\begin{aligned} \lambda P_0 &= \mu P_1 \\ (\lambda + n\mu)P_n &= \lambda P_{n-1} + (n+1)\mu P_{n+1} \quad \text{for} \quad 0 < n < s \\ \lambda P_{s-1} &= s\mu P_s, \end{aligned} \tag{23}$$

with solution:

$$P_n = \frac{\rho^n/n!}{\sum_{i=0}^{s} \rho^i/i!} \quad \text{for} \quad 0 \leq n \leq s, \tag{24}$$

where $\rho = \lambda \mathbb{E}[S]$. Surprisingly, (24) also holds for general service times (the $M/G/s/s$ queue) and is thus insensitive to the service time distribution [31]. The probability that all servers are occupied, is often called the blocking probability, and is given by:

$$P_s = \frac{\rho^s/s!}{\sum_{i=0}^{s} \rho^i/i!}. \tag{25}$$

Formula (25) is often referred to as the Erlang loss formula, or Erlang-B [31]. For large s, the direct calculation of P_s by using (25) often introduces numerical problems. The following stable recursion exists where these problems are avoided [60].

Recursion for Erlang-B

Step 1.
Set $X_0 = 1$.

Step 2.
For $j = 1, \ldots, s$ compute

$$X_j = 1 + \frac{jX_{j-1}}{\rho}. \tag{26}$$

Step 3.
The blocking probability P_s is given by

$$P_s = \frac{1}{X_s}. \tag{27}$$

Another option is to use one of the Erlang-B calculators available online, see e.g. [48] and [54]. The performance measures are given by:

$$\mathbb{E}[L] = \rho\,(1 - P_s), \quad \mathbb{E}[W] = \mathbb{E}[S]. \tag{28}$$

As we have seen in this subsection, the computation of the blocking probabilities can be quite involved. The infinite server, or $M/M/\infty$ queue, is often used to approximate the $M/M/s/s$ queue for a large number of servers. In this queue, upon arrival each patient obtains his own server. The queue length has a Poisson distribution with parameter ρ, where $\rho = \lambda\mathbb{E}[S]$, and is thus given by

$$\begin{aligned} P_n^\infty &= \frac{\rho^n}{n!}P_0, \quad \text{where} \\ P_0^\infty &= e^{-\rho}. \end{aligned} \tag{29}$$

The blocking probability for the system with s servers is approximated by [52]:

$$P_s \approx \sum_{n \geq s}^{\infty} P_n^\infty. \tag{30}$$

2.2.4 Queues with General Arrival and/or Service Processes

For the $M/M/s$ queue a single parameter suffices to calculate the queue length distribution and related performance measures. However, assuming exponentiality of the distributions involved in a queuing process is not always a valid choice. When the coefficient of variation is not close to 1 (the value for the exponential distribution) other probability distributions should be used to obtain reliable outcomes,

since the variance of the inter-arrival and service times has strong influence on the performance measures.

Results for non-exponential systems are scarce and are often characterized via the scv, c^2. In general, when the scv increases, the variability in the related queuing system also increases. In this subsection we will focus on results for mean waiting times. Additional results are given in the books [31], [52] and [57]. The software package QtsPlus that accompanies [31] supports the calculation of many relevant performance measures, is free available online [49] and implemented in MS Excel, but also has an open source variant.

For the $M/G/1$ queue the Laplace-Stieltjes transform for the waiting time distribution is known. From this result, we obtain the Pollaczek-Khinchine formula [21] that characterizes the waiting time in the single-server queue with Poisson arrivals and general service times:

$$\mathbb{E}[W^q] = \mathbb{E}[S]\frac{\rho}{1-\rho}\frac{1+c_S^2}{2}, \tag{31}$$

where c_S^2 denotes the scv of the service time. The mean sojourn time for the $G/M/1$ queue is:

$$\mathbb{E}[W] = \frac{\mathbb{E}[S]}{1-\sigma}, \tag{32}$$

where σ is the unique root in the range $0 < \sigma < 1$ of the following equation:

$$\sigma = \bar{A}(\mu - \mu\sigma), \tag{33}$$

with $\bar{A}$ the Laplace-Stieltjes transform of the inter-arrival time and $\mu = \frac{1}{\mathbb{E}[S]}$ [57]. For the $G/G/1$ queue the following approximation solution is often used [52]:

$$\mathbb{E}[W^q] \approx \mathbb{E}[S]\frac{\rho}{1-\rho}\frac{c_A^2+c_S^2}{2}, \tag{34}$$

where c_A^2 denotes the scv of the arrival process. This result includes the $G/M/1$ queue and is exact for the $M/G/1$ queue.

It is hard to determine the exact effect of using the exponential distribution to represent a non-exponential process. As a rule of thumb, we suggest that as long as the actual variance is below that of the exponential distribution, then the exponential distribution provides a conservative estimate. In other words, the calculated expectations of the queue length and waiting times will over-estimate the actual values. Such a conservative estimate is for instance useful when a strategic decision that does not involve a lot of detail needs to be made.

For the mean waiting time in the $G/G/s$ queue the following approximation is very useful [31]:

$$\mathbb{E}[W^q] \approx \mathbb{E}[W_{q(M/M/s)}]\frac{c_A^2+c_S^2}{2}, \tag{35}$$

where $\mathbb{E}[W_{q(M/M/s)}]$ denotes the mean waiting time in the $M/M/s$ queue with identical λ and μ. In [31] lower and upper bounds on $\mathbb{E}[W^q]$ are also provided. Using the results for $\mathbb{E}[W^q]$, Little's Law can be applied to determine the mean number of patients in the queues mentioned in this subsection.

2.2.5 Service Disciplines

So far, we only discussed the FCFS service discipline. Other options are Processor Sharing (PS) and Last Come First Serve (LCFS). We will elaborate on queuing networks with these kind of queues in Subsection 2.4.2.

In the processor sharing service discipline, all arriving patients are immediately served, thus there is no queuing. A single server is shared equally among patients, where each patient class may have its own service requirement. For the $M/M/1-PS$ queue the queue length distribution, P_n, is identical to that of the $M/M/1-FCFS$ queue (9). Intuitively, this can be explained as follows. The server works at rate μ, and when there are n patients in the queue, an individual patient is served with rate $\frac{\mu}{n}$. However, since n patients are served simultaneously, the overall completion rate is still μ ($\frac{\mu}{n} \cdot n = \mu$). Since the patient arrival rate equals λ, the flow in and out of the queue is identical to that of the $M/M/1-FCFS$ queue.

The $M/M/1-LCFS$ queue with preemptive resume can be seen as a stack, for instance of patient files, where a single server (the doctor) works on the top item of the stack. Whenever a new item is added, the server immediately starts working on this item. However, when the server returns to the previous item, it resumes service (i.e., the queue is work conserving). The queue length distribution is again given by (9), where the same argument holds as for the $M/M/1-PS$ queue.

2.2.6 Miscellaneous Queuing Results

In this subsection we briefly mention a couple other queuing results. Some of the results that can be obtained for $G/G/1$ queues are exact, but do not transfer to queuing networks. In particular, the equilibrium distribution at arrival instants in the $G/M/1$ queue is:

$$P_n = (1-\sigma)\sigma^n, \tag{36}$$

where σ is defined as in (33).

The equilibrium distribution of the $M/M/1$ queue and the $G/M/1$ queue at arrival epochs have a geometric form. At arbitrary epochs, the equilibrium distribution for the $M/G/1$ and $G/M/1$ queues is not available in an amenable form. These distributions, however, can be obtained using the theory of matrix geometric queues. To this end, we introduce the class of so-called phase type distributions [40]. A distribution is of phase-type if it can be represented as a continuous time Markov chain on the phases such that the chain remains in a phase during an exponential time and jumps from phase to phase according to transition probabilities, see [40] for details.

It is interesting to observe that each probability distribution that attains positive values, only, can be approximated arbitrarily closely by a phase-type distribution. Using phase-type distribution for respectively the service time and inter-arrival time distribution, the equilibrium distributions for the $M/Ph_r/1$ and $Ph_s/M/1$ queues are available in closed form. For these queues, the state description requires the number of patients *n and* the phase of the service or inter-arrival times r resp. s. The equilibrium distribution is obtained in closed form:

$$P_n = P_0 R^n, \quad n = 0,1,2,\ldots, \tag{37}$$

where P_0 and P_n are r resp. s vectors over the phases of the service or inter-arrival times and R is an $r \times r$ or $s \times s$ matrix over these phases. The result generalizes to the $Ph_r/Ph_s/1$ queue where P_0 and P_n become rs vectors recording the joint phases of inter-arrival and service times. Although the form (37) is geometric, obtaining the matrix R is quite involved and goes beyond the scope of this chapter, see [39] for details. We specifically mention this queue since phase-type distributions are common in healthcare. For example the length of stay in geriatric care has been modeled using phase-type distributions [24].

Instead of joining the queue, patients may be impatient and leave the queue before service. When this happens upon arrival, it is called balking. When patients leave after waiting some time, it is referred to as reneging. In the $M/M/s/s$ queue it is assumed that patients who are blocked are lost to the system. When blocked and/or impatient patients return to the queue after some time, we have a retrial queue [31].

In this subsection we have considered only queues with a single class of patients. When more than one patient class arrives at the queue, and classes have priority over one another, we have a priority queue [57]. In the case of preemptive priority, the service of the low priority patient is interrupted immediately when a higher prioritized patient arrives. Afterwards, the service of the low priority patient is resumed (work conserving) or may have to start allover again (work is lost). In the case of non-preemptive priority, a patient that is already in service is completed first.

Vacation queues are a generalization of the $M/G/1$ queue, where the server may take a vacation (i.e., becomes idle for a certain amount of time), also when there are patients in the queue [57]. A generalization of the vacation queue is the polling model, where a single server visits multiple queues [51]. In this chapter we restrict our focus to networks of queues with continuous availability.

2.3 Networks of Exponential Queues

Now that we have defined the building blocks, we can proceed to queuing networks. We start with networks of exponential queues with either a single or multiple servers.

2.3.1 Tandem Networks

Consider a tandem network of J queues that are placed in series. All queues have infinite waiting room, a single-server, and the service requirement at queue j, $j = 1,\ldots,J$, has an exponential distribution with mean service time $\mathbb{E}[S_j]$. Patients arrive at queue 1 according to a Poisson process with rate λ. Upon service completion at queue j the patient routes to queue $j+1$, $j = 1,\ldots,J-1$, and finally departs from queue J.

From Burke's theorem [14] it follows that the departure process of a queue with Poisson arrivals and exponential service times, is again a Poisson process with the same rate as the arrival process, and that departures from queue 1 before time t_0 are independent of the queue length of queue 1 at time t_0. This fundamental result indicates that the queue length at time t_0 in queue 1 and queue 2 are statistically independent. Hence, for the tandem queue of Figure 3,

$$P(n_1,n_2) = \mathbb{P}(N_1 = n_1, N_2 = n_2) = (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}, \quad n_1,n_2 \geq 0, \tag{38}$$

where $\rho_1 = \lambda\mathbb{E}[S_1]$, $\rho_2 = \lambda\mathbb{E}[S_2]$, and N_j is the random queue length at queue j in equilibrium. Continuing this argument, for a tandem network of J queues, we obtain the so-called product-form solution [52]:

$$P(n_1,\ldots,n_J) = \prod_{j=1}^{J}(1-\rho_j)\rho_j^{n_j}, \tag{39}$$

where $\rho_j = \lambda\mathbb{E}[S_j]$. This elegant result leads us to Open Jackson Networks with general patient routing.

2.3.2 Open Jackson Networks

We now consider a network consisting of J single-server queues. The external arrival process at queue j, $j = 1,\ldots,J$, is Poisson distributed with rate γ_j, $\gamma_j \geq 0\ \forall j$. Each queue j has an exponentially distributed service requirement with mean service time $\mathbb{E}[S_j]$. Patients are routed from queue i to queue j with state independent routing probability r_{ij}, $0 \leq r_{ij} \leq 1$, i.e., a fraction r_{ij} of patients served at queue i routes to queue j. The parameter r_{i0} denotes the fraction of patients leaving the network at queue i. The total arrival rate λ_j at queue j is given by:

$$\lambda_j = \gamma_j + \sum_{i=1}^{J}\lambda_i r_{ij}, \quad j = 1,\ldots,J, \tag{40}$$

and is composed of the arrivals to queue j from outside and inside the network. A queuing network with these characteristics is called an Open Jackson Network, named after James R. Jackson who first studied its properties in 1957 [32]. In Figure 7 an example of an Open Jackson Network is given. According to Jackson's

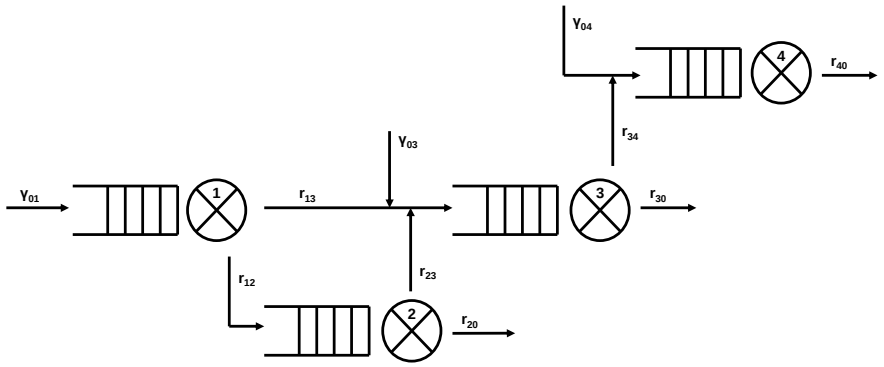

Fig. 7: An example of an Open Jackson Network with four queues and patient routing from queues 1→2, 1→3, 2→3, and 3→4. External arrivals occur at queue 1, 3, and 4; departures occur at queue 2, 3, and 4

Theorem [32], the product-form solution for this type of network is given by:

$$P(n_1,\ldots,n_J) = \prod_{j=1}^{J}(1-\rho_j)\rho_j^{n_j}, \quad n_j \geq 0, \quad j = 1,\ldots,J, \tag{41}$$

where $\rho_j = \lambda_j \mathbb{E}[S_j]$.

The Power of Jackson's Theorem

From Jackson's theorem it follows that per queue only a single parameter, ρ_j, is required for the calculation of $P(n_1,\ldots,n_J)$. Consequently, only J parameters are required to analyze the entire network! This result is surprising since usually many parameters are required to characterize a probability distribution.

Since the queues in the network act as if they are independent $M/M/1$ queues, the performance measures are easy to compute:

$$\mathbb{E}[L_j] = \frac{\rho_j}{1-\rho_j}, \quad \mathbb{E}[W_j] = \frac{\mathbb{E}[L_j]}{\lambda_j}. \tag{42}$$

The mean sojourn time for an arbitrary patient can be calculated using Little's Law:

$$\mathbb{E}[W] = \frac{\sum_{j=1}^{J}\mathbb{E}[L_j]}{\sum_{j=1}^{J}\gamma_j}. \tag{43}$$

Note that this is not equal to $\sum_{j=1}^{J}\mathbb{E}[W_j]$, since patients may not visit all queues in the network or visit some queues several times. Jackson's result can be extended to

the multi-server case. We obtain:

$$P(n_1,\ldots,n_J) = \prod_{j=1}^{J} \frac{\rho_j^{n_j}}{m(n_j)} P_{0j}, \tag{44}$$

where $\rho_j = \lambda_j \mathbb{E}[S_j]$,

$$m(n_j) = \begin{cases} n_j! & \text{for} \quad 0 \le n_j < s_j, \\ s_j^{n_j - s_j} s_j! & \text{for} \quad n_j \ge s_j, \end{cases} \tag{45}$$

and $s_j \geq 1$ for $j = 1,\ldots,J$. The normalization constant P_{0j} is given by

$$P_{0j} = \left(\sum_{n_j=0}^{s_j-1} \frac{\rho_j^{n_j}}{n_j!} + \frac{\rho_j^{s_j}}{s_j!} \frac{s_j}{s_j - \rho_j} \right)^{-1}. \tag{46}$$

2.3.3 Closed Jackson Networks

A Jackson Network where the external arrival rates $\gamma_j = 0 \ \forall j$ and the departure probabilities $r_{i0} = 0 \ \forall i$, is a called a Gordon-Newell or Closed Jackson Network, since patients do not enter or leave (see Figure 8). The finite number N of patients

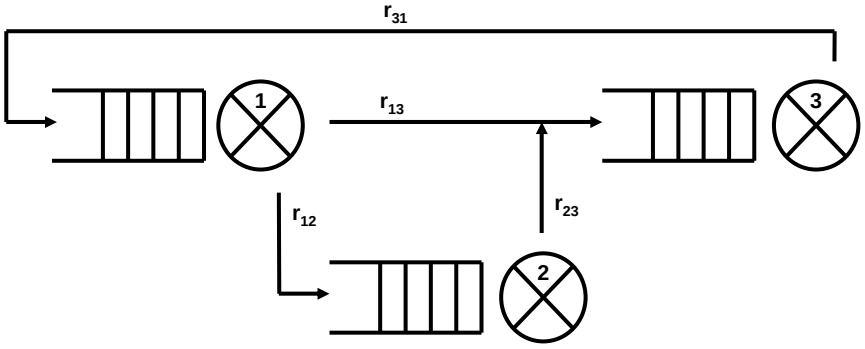

Fig. 8: An example of a Closed Jackson Network with three queues and patient routing from queues 1→2, 1→3, 2→3, and 3→1

that is present in the network is continuously routed among J queues according to the state independent routing probabilities r_{ij}. For the single-server case we obtain a product-form solution [27]:

$$P(n_1,\ldots,n_J) = B(N)^{-1}\prod_{j=1}^{J}\rho_j^{n_j}, \tag{47}$$

where $\sum_{j=1}^{J} n_j = N$. In this formula $B(N)$ is called the normalization constant. In the open network variant, the expression $\prod_{j=1}^{J}(1-\rho_j)$ is actually the normalization constant and easy to compute. In the closed network variant, $B(N)$ is given by:

$$B(N) = \sum_{\sum_{j=1}^{J} n_j = N}\prod_{j=1}^{J}\rho_j^{n_j}. \tag{48}$$

Calculating $B(N)$ can be quite cumbersome, even for small N. Buzen's algorithm [16] is very helpful in this case and works as follows.

Buzen's Algorithm

Step 1.
Define

$$G_j(k), \quad \text{where} \quad j = 0,\ldots,J \quad \text{and} \quad k = 0,\ldots,N, \tag{49}$$

with initial values

$$G_1(k) = \rho_1^k, \quad G_j(0) = 1. \tag{50}$$

Step 2.
Recursively compute

$$G_j(k) = G_{j-1}(k) + \rho_j G_j(k-1). \tag{51}$$

Step 3.
The normalization constant is given by:

$$B(N) = G_J(N). \tag{52}$$

Buzen's algorithm can also be used to compute other performance measures of interest. The marginal probability that n_j patients are present at queue j is given by:

$$P(n_j) = B(N)^{-1}\rho_j^{n_j}\left(G_J(N-n_j) - \rho_j G_J(N-n_j-1)\right). \tag{53}$$

The mean number of patients present at queue j is given by:

$$\mathbb{E}[L_j] = \sum_{n_j=1}^{N}\rho_j^{n_j}B(N)^{-1}G_J(N-n_j). \tag{54}$$

The Closed Jackson Network can also be extended to the multi-server case. The product-form solution is then given by:

$$P(n_1,\ldots,n_J) = B(N)^{-1}\prod_{j=1}^{J}\frac{\rho_j^{n_j}}{m(n_j)}, \tag{55}$$

where $\sum_{j=1}^{J} n_j = N$, $m(n_j)$ is given by (45), and

$$B(N) = \sum_{\sum_{j=1}^{J} n_j = N}\prod_{j=1}^{J}\frac{\rho_j^{n_j}}{m(n_j)}. \tag{56}$$

For the multi-server case $B(N)$ can also be calculated using Buzen's algorithm.

In a closed single-server Jackson network the mean waiting time and mean number of patients at queue j can be calculated without evaluating $B(N)$ [31]. This algorithmic approach is called Mean Value Analysis (MVA). We present the basic algorithm, but MVA has been extended to many other queuing systems, see [2].

MVA Algorithm

Step 1.
Set $\lambda_1 = 1$ and solve the traffic equations:

$$\lambda_j = \sum_{i=1}^{J}\lambda_i r_{ij}, \quad j = 1,\ldots,J. \tag{57}$$

Step 2.
Define $L_j(0) = 0$ for $j = 1,\ldots,J$.

Step 3.
For $n = 1,\ldots,N$, calculate

$$\begin{aligned} W_j(n) &= (1+L_j(n-1))\,\mathbb{E}[S_j], \quad j = 1,\ldots,J,\\ \nu_1(n) &= \frac{n}{\sum_{j=1}^{J}\lambda_j W_j(n)},\\ \nu_j(n) &= \nu_1(n)\lambda_j \quad j = 2,\ldots,J,\\ L_j(n) &= \nu_j(n)W_j(n), \quad j = 1,\ldots,J. \end{aligned} \tag{58}$$

Step 4.
The mean waiting time at queue j is given by:

$$\mathbb{E}[W_j] = W_j(N). \tag{59}$$

The mean number of patients at queue j is given by:

$$\mathbb{E}[L_j] = L_j(N). \tag{60}$$

2.4 Networks of Queues with General Arrival and/or Service Processes

As said, the few exact results that exist for general queues cannot be transferred to general queuing networks. However, many of the approximation results are. In this subsection we describe three types of networks that have an exact solution for the queue length distribution, namely networks with fixed routing, BCMP networks, and loss networks. We conclude with the Queuing Network Analyzer (QNA). This is a generalization of MVA for networks of $G/G/s$ queues.

2.4.1 Networks with Fixed Routing

All of the queuing networks we have discussed so far employ Markovian routing. This means that after departure, patients are routed to other queues or leave the network with a certain probability. This excludes fixed routes in which patients follow a prescribed path.

Consider a network in which each patient class has its own route. The route of patient class k, $k = 1,\ldots,K$, is given by the sequence of queues to visit before leaving the system [34]:

$$r(k,1), r(k,2), \ldots, r(k,H(k)). \tag{61}$$

So in stage h, $h = 1,\ldots,H(k)$, patient class k visits queue $r(k,h)$. Note that one queue may appear multiple times in the route. Using this notation enables to include patients that visit the same queue multiple times, but have a different destination depending on the times the queue has been visited. An example route for a patient class could be $3 \to 2 \to 3 \to 4$, where queue 2 is visited after the patient departs from queue 3 the first time, and queue 4 is visited after the patient departs from queue 2 the second time. This type of queuing network can be seen as a set of intertwined tandem networks (Subsection 2.3.1). Each patient class is routed through its own tandem network of queues, and different patient classes may meet each other at one of the queues.

Let γ_k denote the arrival rate of patient class k. As a consequence of fixed routes, the arrival rate of patient class k at stage h to queue $r(k,h)$ equals the arrival rate of the patient class to the network. In order to be able to determine how many patients of class k being in stage h of their route, are present at queue j, we have to record the position in the queue for each individual patient. We introduce some additional notation. Let $k_j(\ell)$ denote the class of the patient that holds position ℓ in queue j, and let $h_j(\ell)$ denote the stage the patient is currently in. Then $c_j(\ell) = (k_j(\ell), h_j(\ell))$ gives the type of this patient. Since a patient may visit one queue several times, his type potentially gives more information than his class. The state of queue j is given by the vector $c_j = (c_j(1),\ldots,c_j(n_j))$, and $C = (c_1,\ldots,c_J)$ gives the state of the queuing network. Now if we define the parameter $\alpha_j(k,h)$ as follows:

$$\alpha_j(k,h) = \begin{cases} \nu_k & \text{if} \quad r(k,h) \equiv j, \\ 0 & \text{otherwise}, \end{cases} \tag{62}$$

where ν_j is given by $\lambda_j \mathbb{E}[S_j]$, and a_j is the load of queue j:

$$a_j = \sum_{k=1}^{K} \sum_{h=1}^{H(k)} \alpha_j(k,h), \tag{63}$$

then the marginal queue length distribution of the number of patients of class k, $k = 1,\ldots,K$, present at queue j, is given by:

$$P_j(c_j) = B_j^{-1} \prod_{\ell=1}^{n_j} \alpha_j(k_j(\ell), h_j(\ell)), \quad \text{where}$$
$$B_j = \sum_{n=0}^{\infty} a_j^n. \tag{64}$$

The queue length distribution for the entire queuing network is then given by:

$$P(C) = \prod_{j=1}^{J} P_j(c_j). \tag{65}$$

The queue length distribution of the number of patients at the queues in the network is given by:

$$P(n_1,\ldots,n_J) = \prod_{j=1}^{J} (1-\nu_j)\nu_j^{n_j}. \tag{66}$$

Note that this result does not discriminate among patient classes. Even though the notation required can be quite cumbersome, networks with fixed routing introduce substantial modeling flexibility.

2.4.2 BCMP Networks

If each queue j in a network of J queues is one of the following types:

1. $M/M/s-FCFS$
2. $M/G/1-PS$
3. $M/G/1-LCFS$ preemptive resume
4. $M/G/\infty$,

an exact solution exists and the network is a BCMP network. It is named after the authors Baskett, Chandy, Muntz and Palacios, who described it in 1975 [8]. The network may be open or closed with multiple patient classes, and employ Markovian or fixed routing. In the case of an open network, the external arrival rates to the

queues are Poisson. For notational convenience, we give the product-form solution for a BCMP network with Markovian routing and a single patient class. In this case the queue length distribution is given by:

$$P(n_1,\ldots,n_J) = B(N)\prod_{j=1}^{J} P_j(n_j), \tag{67}$$

where $B(N)$ is the normalization constant such that $\sum_N P(n_1,\ldots,n_J)=1$, and $P_j(n_j)$ is the equilibrium distribution for queue j, $j=1,\ldots,J$. If queue j is of type 1:

$$\begin{aligned} P_j(n_j) &= \frac{\rho_j^{n_j}}{m(n_j)}P_j(0), \quad \text{where} \\ m(n_j) &= \begin{cases} n_j! & \text{for} \quad 0 \le n_j < s_j, \\ s_j^{n_j-s_j}s_j! & \text{for} \quad n_j \ge s_j, \end{cases} \quad \text{and} \\ P_j(0) &= \left(\sum_{n_j=0}^{s_j-1}\frac{\rho_j^n}{n_j!} + \frac{\rho_j^s}{s_j!}\frac{s_j}{s_j-\rho_j}\right)^{-1}. \end{aligned} \tag{68}$$

If queue j is of type 2 or 3:

$$\begin{aligned} P_j(n_j) &= \rho_j^{n_j}P_j(0), \quad \text{where} \\ P_j(0) &= 1-\rho_j. \end{aligned} \tag{69}$$

If queue j is of type 4:

$$\begin{aligned} P_j(n_j) &= \frac{\rho_j^{n_j}}{n_j!}P_j(0), \quad \text{where} \\ P_j(0) &= e^{-\rho_j}. \end{aligned} \tag{70}$$

Note that the four queue types include the service disciplines we discussed in Subsection 2.2.5. For BCMP networks the queue length distributions for these service disciplines are insensitive to the service requirement distribution, that is, only the mean service times are required to obtain the equilibrium distribution (67).

2.4.3 Loss Networks

A loss network is the multi-dimensional generalization of the Erlang loss queue (Subsection 2.2.3). In a loss network, patients simultaneously claim at least one server in at least one queue. When upon arrival at the network one of the designated queues is full, the patient is blocked and lost. Note that this kind of queuing network shows an analogy with some hospital processes. For instance, a patient that needs to be admitted to the ICU after surgery, will not be operated on when there is no

ICU bed available. Thus the patient simultaneously claims an operating room and an ICU bed. If either one is not available, the surgery will not commence.

For a loss network handling K patient classes, the queue length distribution of the number of patients of class k, $k = 1,\ldots,K$, is given by [35, 59]:

$$\begin{aligned} P(n_1,\ldots,n_K) &= B(S)^{-1}\prod_{k=1}^{K}\frac{\rho_k^{n_k}}{n_k!}, \quad \text{where} \quad n \in S(S), \\ S(S) &= \{n \in \mathbb{N}_0, \sum_{k=1}^{K} A_{jk}n_k \leq s_j\}, \\ B(S) &= \sum_{n\in S(S)}\prod_{k=1}^{K}\frac{\rho_k^{n_k}}{n_k!}, \quad \rho_k = \lambda_k \mathbb{E}[S_k], \end{aligned} \tag{71}$$

with λ_k the arrival rate to the network of patients of class k, $\mathbb{E}[S_k]$ the mean sojourn time in the network, s_j the number of servers at queue j and A_{jk} the number of servers a patient of class k claims at queue j. Loss networks are insensitive to the sojourn time distribution. Various algorithms and approximations exist to obtain blocking probabilities [35, 59].

2.4.4 The Queuing Network Analyzer

Despite the fact that many real world problems do not exhibit exponential service times, open Jackson networks have been used in numerous applications, often with good results. However, to analyze networks of general queues, the Queuing Network Analyzer (QNA) is a better alternative. The QNA was developed in 1983 by Ward Whitt [55] for approximate analysis of open networks of $G/G/s$ queues with FCFS service discipline. There are several variations on the QNA, also known as reduction or decomposition methods (see [15]). In this subsection we summarize the basic QNA algorithm.

QNA Algorithm

Step 1.

Calculate the aggregate arrival rates at queue j, λ_j:

$$\lambda_j = \gamma_j + \sum_{i=1}^{J}\lambda_i r_{ij}. \tag{72}$$

Step 2.

Calculate the load of a server at queue j, ϕ_j:

$$\phi_j = \frac{\lambda_j \mathbb{E}[S_j]}{s_j}. \tag{73}$$

Step 3.
Calculate the flow from queue i to queue j, λ_{ij}:

$$\lambda_{ij} = \lambda_i r_{ij}, \tag{74}$$

and the fraction of arrivals at queue j that come from queue i, q_{ij}:

$$q_{0j} = \frac{\gamma_j}{\lambda_j}, \quad q_{ij} = \frac{\lambda_{ij}}{\lambda_j}, \tag{75}$$

where q_{0j} denotes the fraction of external arrivals to queue j.

Step 4.
Calculate the scv for the arrival process at queue j, $c_{A,j}^2$:

$$\begin{aligned} c_{A,j}^2 &= a_j + \sum_{i=1}^{J} c_{A,i}^2 b_{ij}, \quad \text{with} \\ a_j &= 1 + w_j \left[(q_{0j} c_{0j}^2 - 1) + \sum_{i=1}^{J} q_{ij}((1-r_{ij}) + r_{ij}\phi_i^2 x_i) \right], \end{aligned} \tag{76}$$

where c_{0j}^2 is the scv of the external arrival process at queue j, and

$$x_i = 1 + \frac{1}{\sqrt{m_i}}(\max(c_{S,i}^2, \frac{1}{5}) - 1), \tag{77}$$

with $c_{S,i}^2$ the scv of the service process at queue i. We have

$$\begin{aligned} b_{ij} &= w_j q_{ij} r_{ij}(1-\phi_i^2), \quad w_j = \left[(1+4(1-\phi_j)^2(\eta_j - 1)\right]^{-1}, \quad \text{and} \\ \eta_j &= \left[\sum_{i=0}^{J} q_{ij}^2 \right]^{-1}. \end{aligned} \tag{78}$$

Step 5.
The mean waiting time at queue j, $\mathbb{E}[W_j]$, is given by

$$\mathbb{E}[W_j] = \mathbb{E}[W_{M/M/s}] \frac{c_{A,j}^2 + c_{S,j}^2}{2}. \tag{79}$$

The calculations involved with the QNA are usually straightforward and can be done by hand. However, when the parameters need to be changed often, we suggest using a spreadsheet program such as MS Excel. QtsPlus [49] also supports the analysis of general queuing networks. Even though the QNA has proved to be very useful, other approximation methods give better results when the network is highly congested (see [15] for further reference).

2.5 State of the Art in Networks of Queues

Queuing theory traces back to Erlang's historical work for telephony networks in 1909 [12]. The simplicity and fundamental flavour of Erlang's famous expressions, such as his loss formula for an incoming call in a circuit switched system to be lost, see Subsection 2.2.3, has remained intriguing, and has motivated the development of results with similar elegance and expression power for various systems modeling congestion and competition over resources.

A second milestone was the step of queuing theory into queuing networks as motivated by the product form results for manufacturing systems in the nineteen fifties obtained by Jackson [32]. These results revealed that the queue lengths at nodes of a network, where customers route among the nodes upon service completion in equilibrium can be regarded as independent random variables, that is, the equilibrium distribution of the network of nodes factorizes over (is a product of) the marginal equilibrium distributions of the individual nodes as if in isolation, see Subsection 2.3.2. These networks are nowadays referred to as Jackson networks.

A third milestone was inspired by the rapid development of computer systems and brought the attention for service disciplines such as the Processor Sharing discipline introduced by Kleinrock in 1967 [36]. More complicated multi-server nodes and service disciplines such as First Come First Served, Last Come First Served and Processor Sharing, and their mixing within a network have led to a surge in theoretical developments and a wide applicability of queuing theory, see Subsection 2.4.2.

Queuing networks have obtained their place in both theory and practice. New technological developments such as Internet and wireless communications, but also advancements in existing applications such as manufacturing and production systems, public transportation, and logistics, have triggered many theoretical and practical results. The questions arising in health care will no doubt again lead to a surge in the development of queuing theoretical results and applications, a fourth milestone in queuing theory.

Queuing network theory has focused on both the analysis of complex nodes, and the interaction between nodes in networks. Many textbooks and handbooks include or are devoted to queuing theory. Basic level textbooks include [50, 56], and more advanced handbooks are [31, 36, 37, 44, 52, 57]. The state of the art in the mathematical theory for queuing networks is described in the handbook [11]. Topics treated include:

- A general theory for product form equilibrium distributions far beyond those for Jackson and BCMP networks.
- Monotonicity and comparison results that allow analytical bounds on performance measures for networks that slightly deviate from Jackson or BCMP type networks.
- Fluid and diffusion limits that aim at analyzing networks in the regimes dominated by the mean or the variances of the underlying processes such as service times and inter arrival times.

- Computational results that are far more general than the queuing network analyzer of Subsection 2.4.4.

In the last chapter an application of networks of queues in healthcare is presented, indicating that many available theoretical results for networks of queues are waiting to be disclosed for application in healthcare.

3 Examples of Healthcare Applications

As we have seen in the previous section, for some queuing networks that consist of only exponential queues analytical solutions are available. When either the arrival or service process is non-exponential, approximation methods are usually required. In this section we provide several references to healthcare examples that involve queuing networks, and discuss two examples in detail. For examples that involve single queues, we refer to [29].

Generally speaking, three types of healthcare networks have been studied using queuing network topologies. We distinguish between networks of healthcare facilities, networks of departments within a facility, and networks of healthcare providers within a department (see Figure 9). Using this network classification, and the dis-

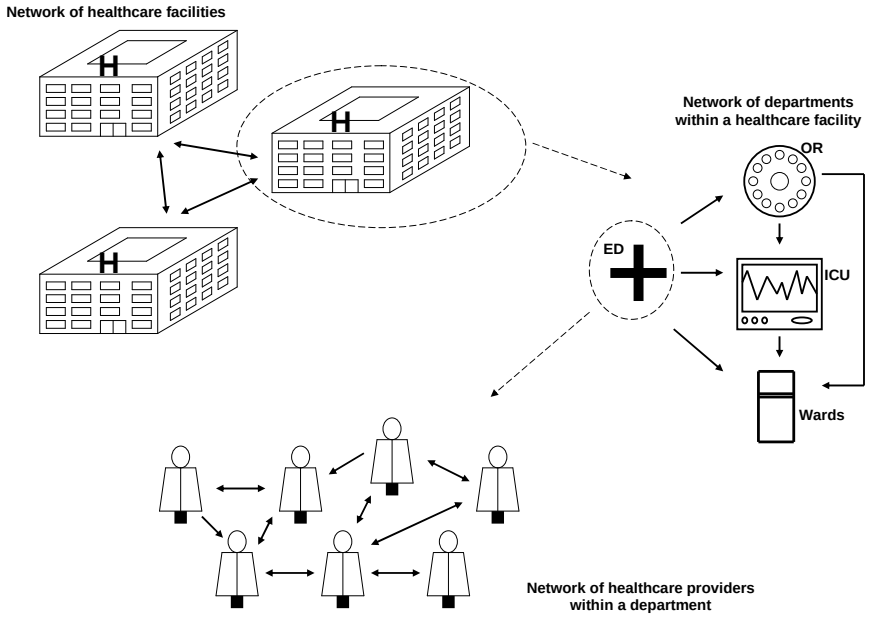

Fig. 9: Different types of networks in healthcare

tinction among exponential and general networks, the references provided in this section can be categorized as presented in Table 1.

Table 1: Categorization of references

	Exponential networks	General networks
Network of healthcare facilities	[5],[6],[10],[38],[42]	[1]
Network of departments within a facility	[18],[19],[47]	[22]
Network of healthcare providers within a department	-	[3],[20],[33],[61]

3.1 Applications of Exponential Networks

Modeling a healthcare network with exponential queues gives a lot of insight into the structural behavior, such as bottlenecks. The modeling power of these networks is most when many of the details on patient behavior are not yet specified, but randomness is an essential part of the behavior of the system, i.e., at the strategical level of allocation of capacity, facilities and resources.

3.1.1 Facility Location and Bed Blocking Problems

One of the earliest developments in this area is given in [10], where a network of $M/M/s/s$ queues is combined with an algorithm to determine the optimal location of burn care facilities in the state of New York. The resulting system of equations can be solved, but due to computational difficulties only for a small number of facilities and beds. This type of network is further studied in [47]. The latter paper involves an example where patients are routed through a network of operative and post-operative units (such as the OR, ICU and nursing wards), and may experience bed-blocking when the next unit on the route operates at full capacity. Also in this model the numerical computations remain problematic when there are numerous units and beds. The relationship between the OR and bed availability on the ICU is further studied in [23], where the authors use a loss network to determine the blocking probability for surgical patients caused by a lack of ICU beds. The bed blocking problem is also considered in [38], where the flow of psychiatric patients within a network of healthcare facilities is considered. A relatively simple steady-state analysis results in a product-form solution. The capacity planning problem for neonatal units (how many cots to place at each care unit) is analyzed in [5] using a loss network model. The implementation of the solution is described in [6].

3.1.2 Patient Flow

Modeling patient flow has received limited attention [53]. Patient flow between different hospital departments is studied in two papers by the same author. In [18] the patient flow from the ED to the ICU and nursing wards is studied using an open Jackson Network. The same methodology is used in [19] to analyze flow of obstetric patients. Patient flow within a care facility is studied from another perspective in [17] and [58]. In these papers, different phases in the care trajectory of a patient are considered. While in [17] a closed queuing network is used, in [58] the model is extended to a semi-open queuing network with a capacity constraint (the maximum number of patients that can be admitted).

3.1.3 Clinical Capacity Problem

Patients with renal failure are considered in [42]. These patients either receive dialysis at a clinic, or when their condition worsens, (temporarily) hospitalized. A multi-class open queuing network with two queues (the clinic and the hospital respectively) is used to determine the clinic's capacity and the maximum number of patients to be admitted into the clinic, given that patients do not use clinic resources when they are hospitalized.

3.2 Applications of General Networks

When a higher level of detail is required, for example when networks of healthcare providers withing a department are studied, models with general queues are of more value.

3.2.1 Organization of Acute Care

The organization of acute care is studied in [20] and [33]. In [20] an ED is modeled with a multi-class open network of $M/G/s$ queues. The main purpose of this model is to determine the required ED capacity needed to achieve service targets such as waiting time and overflow probabilities. In [33] the same kind of network is used to model an urgent care center (UCC), which is basically an outpatient clinic that delivers ambulatory urgent care to relieve pressure from the ED. The main goal of this model is to determine whether parallelization of tasks in the patient's care trajectory has a positive effect on the patient's length of stay at the UCC.

3.2.2 Other Applications

In [22] hospital departments and their interdepartmental relationships are modeled as a network with $G/G/s$ queues. Analysis of the network gives relevant information such as utilization rates and mean waiting times for each queue, and also allows for exploring the impact of service interruptions, aggregating patient flows, and determining the optimal number of patients in a clinic session. Another application is the recent outbreaks of viruses, such as the H1N1 influenza virus, which call for a rapid response of the authorities. In [1] the authors show how a queuing network can help to plan emergency mass dispensing and vaccination clinics. In [3] an outpatient clinic is studied using the Queuing Network Analyzer. The paper provides a nice example of how a queuing network can be of added value when performing bottleneck analysis.

3.3 Example I: Distribution of Patient Classes over Nursing Wards

This example is based on a project carried out by the authors at Leiden University Medical Center (LUMC), one of the eight university hospitals in the Netherlands. The LUMC admits 20,500 inpatients per year and has 14 wards with a total of 390 beds (2009 data).

3.3.1 The Problem

LUMC management wanted to study the distribution of patient classes over the nursing wards and the related bed requirements. We supported them by developing a loss network model that allows for an exact calculation of the fraction of patients that are blocked because the ward is full, and the mean utilization rate per ward. Of course, in practice arrival and service processes at the wards are very complex; arrivals are not homogeneously distributed over the day; patients are not always blocked when the ward is full (e.g. an extra temporary bed is created), and so on. However, for the purpose of this project, this model was a sufficient and fitting tool.

3.3.2 The Model

Figure 10 gives a simple representation of the nursing ward loss network. Patients enter the wards via the ED, the ICU, another hospital, or come from (a nursing) home. Ultimately patients leave the ward again to go home, to another hospital, or sometimes, unfortunately, die. Each patient has an attending physician from specialty i, $i = 1,\ldots,I$. We assume that patients are routed to the ward of their attending physician. Patients come in three classes k: elective short-stay patients ($k = es$),

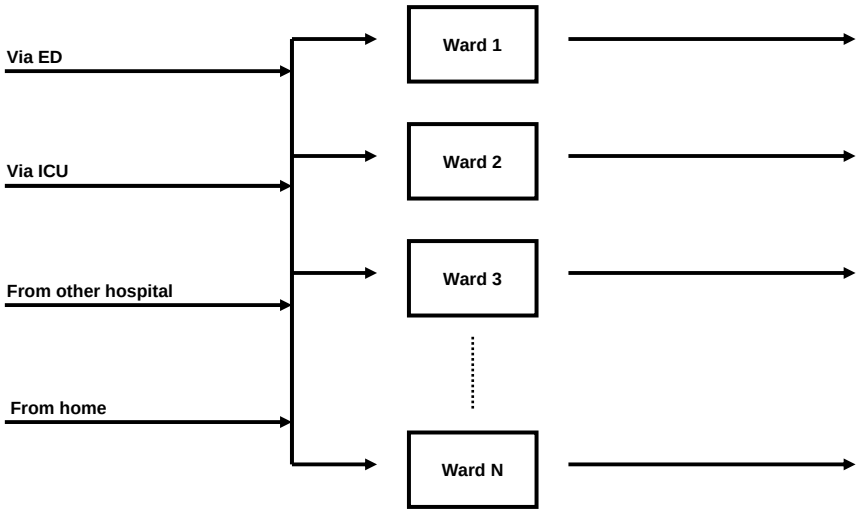

Fig. 10: Nursing ward loss network

elective long-stay patients ($k = el$), and urgent patients ($k = u$), and have mean sojourn time $\mathbb{E}[S_{ik}]$. They originate from one of the four sources m: ED ($m = ed$), ICU ($m = ic$), another hospital ($m = oh$), or home ($m = ho$). Patients are admitted to one of the wards j, $j = 1,\ldots,J$, with routing probability $P_{ik,m,j}$, where $P_{ik,m,j} \in \{0,1\}$ and $\sum_j P_{ik,m,j} = 1\ \forall i,k,m$ (all patients should be admitted to a ward). Each ward has c_j physical beds, of which s_j are staffed and can be used to admit patients. It may occur that a ward has more physical than staffed beds, so $s_j \leq c_j$. If all staffed beds at the designated ward are full, the patient is blocked and not admitted to the ward (patients will not be admitted at another ward). The mean sojourn time, $\mathbb{E}[S_j]$, and arrival rate, λ_j, at ward j are calculated using the fraction of patients that is routed to ward j:

$$\begin{aligned}\lambda_j &= \sum_{i=1}^{I} \sum_{k=\{es,el,u\}} \sum_{m=\{ed,ic,oh,ho\}} \lambda_{ik,m} P_{ik,m,j},\\ \mathbb{E}[S_j] &= \sum_{i=1}^{I} \sum_{k=\{es,el,u\}} \sum_{m=\{ed,ic,oh,ho\}} \mathbb{E}[S_{ij}] P_{ik,m,j}.\end{aligned} \tag{80}$$

We assume that the departure rates from the sources m are Poisson; thus, the ward arrival rates are also Poisson. The problem we study is how the hospital should distribute the patient groups ik over the wards j. Depending on the number of staffed beds, each ward can offer a certain amount of care. The hospital should choose whether it wants to focus on achieving a blocking probability which is below a

certain value, or a mean utilization rate which is above a threshold[2]. An additional benefit of a distribution that optimally uses the ward capacities is that it might be possible to close one or more wards. Since we consider each ward j as a separate entity, the blocking probability, $\mathbb{P}_{s_j}$, is given by

$$\mathbb{P}_{s_j} = \frac{(\lambda_j \mathbb{E}[S_j])^{s_j} / s_j!}{\sum_{l=0}^{s_j} \frac{(\lambda_j \mathbb{E}[S_j])^l}{l!}}. \tag{81}$$

The utilization rate of the beds at ward j, ϕ_j, is given by

$$\phi_j = \frac{\left(1 - \mathbb{P}_{s_j}\right) \lambda_j \mathbb{E}[S_j]}{s_j}. \tag{82}$$

To attain the desired value of either $\mathbb{P}_{s_j}$ or ϕ_j, one can calculate the required value of $\lambda_j \mathbb{E}[S_j]$. This can be done by hand or by using spreadsheet software such as MS Excel. An easier option is to use one of the Erlang-B calculators available online (see e.g. [48]). By amending the routing probabilities $P_{ik,m,j}$, it is possible to evaluate all kinds of patient class distributions over the wards.

During the project, we developed a practical extension to the model. We observed it was hard for hospital management to obtain a 'gut feeling' for which patient classes could be combined at a ward. We therefore printed a large map of the hospital with the locations of the wards. For each ward we printed the maximum value of $\lambda_j \mathbb{E}[S_j]$ (which depends on s_j). We also made cards that for each patient class ik had the value of $\sum_m \lambda_{ik,m} \mathbb{E}[S_{ik}]$ printed on it. Hospital management could put the cards with patient classes on the locations on the map, and explore the effect of combining various patient classes. This example shows that queuing techniques can also provide online decision support.

Using the theory of loss networks (Subsection 2.4.3), we can further improve the performance of the wards. Patient groups are still routed to a dedicated ward, but nursing staff can be shared among wards. This way, the previously unstaffed physical beds can be used as well, resulting in a lower blocking probability and a higher utilization rate. Consider for example a simple system with two wards. Ward 1 has $c_1 = 5$ physical beds, $s_1 = 3$ staffed beds, and arrival rate $\lambda_1 = 2$ patients per day. Ward 2 has $c_2 = 5$ physical beds, $s_2 = 4$ staffed beds, and arrival rate $\lambda_2 = 3$ patients per day. At both wards the mean sojourn time equals one day. If the wards would operate separately as in the example above, both wards would have a blocking probability of 21% and an utilization rate of 53% resp. 60%.

If the two wards would share nursing staff, we can formulate this example as a loss network:

[2] Many hospitals aim for a mean utilization of 85% and a blocking probability below 5% at the same time. This is only possible when the ward has a large (around 50) number of beds [13].

$$P(n_1,n_2) = B(S)^{-1}\frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!},$$
$$n_1 \le c_1, \quad n_2 \le c_2, \quad n_1+n_2 \le s_1+s_2, \quad \text{and}$$
$$B(S) = \sum_{n_1,n_2}\frac{\rho_1^{n_1}}{n_1!}\frac{\rho_2^{n_2}}{n_2!}, \tag{83}$$

where n_1, n_2 denotes the number of patients present at ward 1 resp. 2. We see that in total still at most $s_1+s_2=7$ patients could be present at the same time. However, now at ward 1 at most $c_1=5$ instead of $s_1=3$ patients can be admitted, and at ward 2 at most $c_2=5$ instead of $s_2=4$ patients can be admitted, as long as the total number of patients does not exceed 7. The blocking probability then decreases to 16%, while the utilization rate per staffed bed at the wards increases to 56% resp. 63%.

3.4 Example II: Redesign of a Preanesthesia Evaluation Clinic

This example is based on [61].

3.4.1 The Problem

We consider a preanesthesia evaluation clinic (PAC). At this clinic, which is operated by the department of Anesthesiology, patients are screened before undergoing elective surgery. In the last decades most hospitals have organized this screening in an outpatient setting. Not only will a well-performed screening reduce the surgical risk for the patient, but also it reduces the number of canceled surgeries due to the physical condition of the patient [25]. Initially, the screening process at the PAC was organized as follows. Four anesthesia care providers performed the actual screening, supported by a secretary and two clinic assistants. The screening consisted of several separate medical and administrative tasks. The majority of patients (70%) would be screened directly after their consultation at the surgeon's outpatient clinic. This direct (walk-in) screening would only be possible for non-complex patients with ASA I&II classification [4], patients with a more severe health status (ASA III&IV classification) would receive an appointment, since additional medical information and a longer consultation time was required. An increased staff workload, resulting from the introduction of an electronic patient data management system, led to lower job satisfaction, work stress and prolonged patient waiting times. Although 90% of the annual 6,000 PAC patients were eligible for walk-in, one third of these patients were seen on appointment basis, due to an overcrowded waiting room when they first presented themselves at the PAC.

3.4.2 The Model

To identify bottlenecks in the PAC's operations, the clinic was modeled as a multi-class open queuing network (see Figure 11). There were three patient classes: children, adults eligible for direct (walk-in) screening, and adults requiring an appointment because of their (more severe) health status. The PAC queuing network has three separate (connected) queues, where the employees act as servers. Patients only enter the PAC through the secretary queue, but may leave the system at any queue. The PAC queuing network was analyzed using a decomposition method, based on the QNA. This method consists of three steps. We first summarize the method and then provide a detailed description of the model with the corresponding formulas.

First, the multi-class network is reduced to a single class network. This is done by aggregating all patient flows that enter a queue. Then the workload ρ is calculated for each queue. This already gives significant and valuable information; recall that ρ is a measure for the fraction of time employees are busy. In the next step, the single class open queuing network is analyzed, where the mean contact time and scv of the joint arrival and service processes at the three queues are deduced. In the final step the mean waiting time per queue is calculated, using the variables that were derived in step 1 and 2.

In the initial analysis of the PAC queuing network, it was found that the secretary and anesthesia care providers functioned as bottlenecks. Consequently, several alternatives were formulated together with clinic staff, in order to remove these bottlenecks. All alternatives were evaluated using the queuing network model, resulting in one alternative that outperformed the others. In this alternative, several tasks were redistributed and the patient arrival process was amended such that the arrivals were

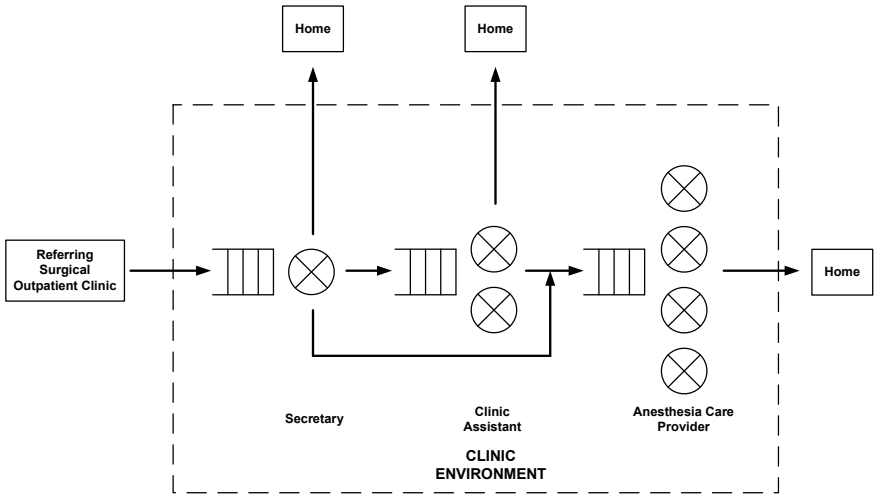

Fig. 11: Queuing network representation of PAC

spread more equally over the day. In the year following the implementation of the alternative clinic design, patient arrivals increased (unexpectedly) by 16%. In the old situation, this would likely have resulted in even longer patient waiting times (recall Figure 2). However, the mean patient length of stay at the PAC did not increase significantly, and more patients (81%) were offered the direct screening.

Detailed Description of the Decomposition Method

The PAC queuing network consists of three queues. The secretary queue is a single-server queue whereas the clinic assistant and anesthesia care providers are represented by multi-server queues. Patients enter the queuing network via the secretary queue and depart the system from any of the queues. Furthermore, if upon arrival at a queue an employee is available patients are served immediately; otherwise they join the queue and are treated on first come first serve basis. We use an approximate decomposition method [9] that is based on the QNA to analyze the model. The model we will present here is more involved than the initial QNA formulation as given in Subsection 2.4.4. Practical situations can usually not be directly translated into an existing model. Instead, the theory has to be amended and extended to represent reality. We will describe in detail the changes we have made to the QNA algorithm.

First we introduce some notation. There are k distinct patient classes, where $k = 1$ are patients deferred to an appointment by the secretary, $k = 2$ adults with ASA I or II, $k = 3$ adults with ASA III or IV, and $k = 4$ are children. To evaluate the alternative clinic design, we also introduce $k = \{5,6,7\}$ to represent patients (adults with ASA I or II, adults with ASA III or IV, and children, respectively) who return for their appointment. We have j queues, $j = 1,2,3$, representing the secretary, clinic assistant and anesthesia care provider.

Step 1.
The aggregated arrival rates at queue j are:

$$\lambda_1 = \sum_{k=1+d}^{4+3d} \gamma_k, \quad \lambda_2 = \sum_{k=2}^{3} \gamma_k, \quad \lambda_3 = \sum_{k=2}^{4} (1 - d a_k)\gamma_k + d \sum_{k=5}^{7} \gamma_k, \tag{84}$$

where γ_k is the arrival rate of patient class k at queue 1, and a_k is the fraction of patients of class k who are deferred to an appointment in the alternative clinic design. Since the indices $k = \{5,6,7\}$ only exist when the alternative clinic design is evaluated, we introduce the binary variable d, which equals 1 if the alternative clinic design is evaluated and 0 otherwise.

Step 2.
The load per patient class per server for queue 1,2, and 3 is:

$$\begin{aligned}
\phi_{1,k} &= \gamma_k \mathbb{E}[S_{k,1}] \tfrac{1}{e_1 s_1} && \text{for} \quad k = \{1+d,\ldots,4+3d\},\\
\phi_{2,k} &= \gamma_k \mathbb{E}[S_{k,2}] \tfrac{1}{s_2} && \text{for} \quad k = \{2,3\},\\
\phi_{3,k} &= \gamma_k \mathbb{E}[S_{k,3}] \tfrac{1}{e_3 s_3} + d(1-a_k)\gamma_k \mathbb{E}[S_{k,3}] \tfrac{1}{e_3 s_3} && \text{for} \quad k = \{2,\ldots,4+3d\},
\end{aligned} \tag{85}$$

where $\mathbb{E}[S_{k,j}]$ is the mean service time for patient class k at queue j. Since the secretary is often consulted by other patients and co-workers while handling a patient at the reception desk, an effective capacity e_1, $0 < e_1 \leq 1$, is taken into account when calculating the mean time a patient spends at this queue. The anesthesia care provider is often disturbed, but not while treating patients and therefore the effective capacity, e_3, $0 < e_3 \leq 1$, is only used in calculating the load. These effective capacities are calculated by using direct observations and interviews with the employees. The number of servers (i.e. employees) at queue j equals s_j. Adding the load over all patient classes gives the aggregated load per server of queue j, $j = 1,2,3$:

$$\phi_1 = \sum_{k=1+d}^{4+3d} \phi_{1,k}, \quad \phi_2 = \sum_{k=2}^{3} \phi_{2,k}, \quad \phi_3 = \sum_{k=2}^{4+3d} \phi_{3,k}. \tag{86}$$

For stability it is required that $\phi_j < 1$ for all queues j.

Step 3.
The flow from queue 1 to queue 2 or 3 and from queue 2 to queue 3 is given by:

$$\lambda_{1,2} = \sum_{k=2}^{3} \frac{(1-da_k)\gamma_k}{\lambda_1}, \quad \lambda_{1,3} = \frac{\sum_{k=4}^{4+3d}(1-da_k)\gamma_k}{\lambda_1}, \quad \lambda_{2,3} = \sum_{k=2}^{3} \frac{(1-da_k)\gamma_k}{\lambda_2} \tag{87}$$

The fraction of arrivals at queue 3 that come from queue 1 or 2 is given by (note that $q_{1,2} = 1$):

$$q_{1,3} = \frac{\sum_{k=4}^{4+3d}(1-da_k)\gamma_4}{\lambda_3}, \quad q_{2,3} = \sum_{k=2}^{3} \frac{(1-da_k)\gamma_k}{\lambda_3}. \tag{88}$$

Step 4.
The arrival process at queue 1 has scv, $c_{A,1}^2$:

$$c_{A,1}^2 = w_1 \sum_{k=1+d}^{4+3d} Q_{k,1} c_{A,k,1}^2 + 1 - w_1, \tag{89}$$

where $c_{A,k,1}^2$ is the scv of the arrival process of patient class k at queue 1, and

$$w_1 = \left(1 + 4(1-\phi_1)^2(\eta_1 - 1)\right)^{-1}, \quad \eta_1 = \frac{\lambda_1^2}{\sum_{k=1+d}^{4+3d} \gamma_k^2}, \quad Q_{k,1} = \frac{\gamma_k}{\lambda_1}. \tag{90}$$

The mean service time, $\mathbb{E}[S_1]$ and scv at queue 1, $c_{S,1}^2$, are:

$$\mathbb{E}[S_1] = \frac{\sum_{k=1+d}^{4+3d} \gamma_k \mathbb{E}[S_{k,1}]}{\lambda_1}, \quad c_{S,1}^2 = \frac{\sum_{k=1+d}^{4+3d} \gamma_k \mathbb{E}^2[S_{k,1}](c_{S,k,1}^2+1)}{\lambda_1 \mathbb{E}^2[S_1]} - 1, \quad (91)$$

where $c_{S,k,j}^2$ is the scv of the service time for patient class k at queue j. The arrival process at queue 2 has scv, $c_{A,2}^2$:

$$c_{A,2}^2 = \lambda_{1,2} c_{D,1}^2 + 1 - \lambda_{1,2}, \quad (92)$$

where $c_{D,1}^2$ is the scv of the departure process at queue 1. Queue 2 has mean service time, $\mathbb{E}[S_2]$, and scv, $c_{S,2}^2$:

$$\mathbb{E}[S_2] = \frac{\sum_{k=2}^{3} \gamma_k \mathbb{E}[S_{k,2}]}{\lambda_2}, \quad c_{S,2}^2 = \frac{\sum_{k=2}^{3} \gamma_k \mathbb{E}^2[S_{k,2}](c_{S,k,2}^2+1)}{\lambda_2 \mathbb{E}^2[S_2]} - 1. \quad (93)$$

The arrival process at queue 3 has scv, $c_{A,3}^2$:

$$\begin{aligned}
c_{A,3}^2 &= w_3(q_{2,3}c_{2,3}^2 + q_{1,3}c_{1,3}^2) + 1 - w_3, \quad \text{with} \\
w_3 &= \left(1 + 4(1-\phi_3)^2(\eta_3 - 1)\right)^{-1}, \quad \eta_3 = \left(q_{2,3}^2 + q_{1,3}^2\right)^{-1}, \\
c_{1,3}^2 &= \lambda_{1,3} c_{D,1}^2 + 1 - \lambda_{1,3}, \quad c_{2,3}^2 = (1-d)c_{D,2}^2 + d(\lambda_{2,3} c_{D,2}^2 + 1 - \lambda_{2,3}), \\
c_{D,2}^2 &= 1 + (1-\phi_2^2)(c_{A,2}^2 - 1) + \frac{\phi_2^2}{\sqrt{s_2}}(c_{S,2}^2 - 1), \quad (94)
\end{aligned}$$

where $c_{2,3}^2$ is the scv of the patient flow from queue 2 to queue 3, $c_{1,3}^2$ the scv of the patient flow from queue 1 to queue 3, and $c_{D,2}^2$ is the scv of the departure process at queue 2. Queue 3 has mean service time, $\mathbb{E}[S_3]$, and scv, $c_{S,3}^2$:

$$\begin{aligned}
\mathbb{E}[S_3] &= \frac{\sum_{k=2}^{4}(1-da_k)\gamma_k \mathbb{E}[S_{k,3}]}{\lambda_3} + d\sum_{k=5}^{7} \gamma_k \mathbb{E}[S_{k,3}], \\
c_{S,3}^2 &= \frac{\sum_{k=2}^{4}(1-da_k)\gamma_k \mathbb{E}^2[S_{k,3}](c_{S,k,3}^2+1) + \sum_{k=5}^{7} \gamma_k \mathbb{E}^2[S_{k,3}](c_{S,k,3}^2+1)}{\lambda_3 \mathbb{E}^2[S_3]} - 1. \quad (95)
\end{aligned}$$

Step 5.
We are interested in the waiting times for patients per queue and the load per employee at each queue. The latter is given by the aggregated load derived in step 1, while the mean waiting times are obtained by using the scv and mean service time calculated in step 2. The mean waiting time, $\mathbb{E}[W_j^q]$, is equal for all patient classes.

$$\mathbb{E}[W_1^q] = \frac{c_{A,1}^2 + c_{S,1}^2}{2} \frac{\phi_1}{1-\phi_1} \frac{\mathbb{E}[S_1]}{e_1},$$
$$\mathbb{E}[W_j^q] = \frac{c_{A,j}^2 + c_{S,j}^2}{2} \mathbb{E}[W_{j(M/M/s)}^q], \quad \text{where}$$
$$\mathbb{E}[W_{j(M/M/s)}^q] = G_j^{-1} \frac{(s_j\phi_j)^{s_j}}{s_j!} \frac{\mathbb{E}[S_j]}{s_j(1-\phi_j)^2},$$
$$G_j = \sum_{n=0}^{s_j-1} \frac{(s_j\phi_j)^n}{n!} + \frac{(s_j\phi_j)^{s_j}}{(1-\phi_j)s_j!} \quad \text{for} \quad j = 2,3. \tag{96}$$

Patient length of stay for each patient class can now be calculated by adding the mean waiting and length of stay of all care queues the patient calls at on his visit to the PAC.

4 Challenges and Directions for Future Research

In the last decade the number of healthcare problems that have been studied using a queuing network approach has increased tremendously. Except for [3] and [10], all of the references included in Section 3 were published in the years 2000-2010. In this final section we point out a few directions for future research. We distinguish between mathematical challenges: healthcare problems for which appropriate queuing network models have not yet been developed, and healthcare challenges: healthcare problems which have not been studied yet, but could be studied with the queuing techniques available in literature.

4.1 Mathematical Challenges

The mathematical challenges mainly lie in the modeling aspect. One example is the development of models for networks of care providers who perform several tasks in parallel, in sequence, and sometimes even in a mixed form. Polling models [51] could be of interest here. Also, clinics where patients have to (re-) visit specific care providers in a network of care queues still involve modeling complications. However, re-visiting occurs often in reality (consider for example the complex network of multiple care providers in ED treatment).

The application of time inhomogeneous models that capture the time-dependent arrival patterns of patients has attained only limited attention, see for example [30]. Introducing time inhomogeneity in queuing networks is a tremendous challenge. Related is the development of computationally efficient methods that explicitly take into account opening hours of healthcare facilities.

4.2 Healthcare Challenges

Healthcare professionals in a couple of fields are familiar with the possibilities of mathematical decision support techniques in general and queuing theory in particular. As we have seen in Section 3, modeling networks of healthcare facilities, departments and care providers has received some attention. However, capturing the complex relationships between hospital departments has proved to be quite involved. The relationship studied is usually that with a downstream department [53], while that with upstream departments is not considered, even though it can be of significant influence.

Our aging population requires more and more care, which has to be delivered with limited resources. Rationing care and the consequences thereof has therefore become an important research topic. Decisions regarding which patient class will be offered what type of care are inevitable. The influence of these decisions on other patient classes, regarding accessibility and other important matters, should be studied in detail. Moreover, the dimensioning of healthcare facilities, not only in the number of beds required, but also regarding care that will be offered to certain patient classes only, will become increasingly important.

This chapter has provided a thorough theoretical background on networks of queues and examples of how networks of queues may be used to model, analyze and solve health care problems. In that respect, often, the theory has to be amended or extended. We are confident that this contribution has made health care professionals increasingly aware of the possibilities and opportunities queueing networks have to offer to tackle the challenges they are facing, now and in the future.

5 References

1. Aaby K, Herrmann JW, Jordan CS, Treadwell M, Wood K (2006) Montgomery county's public health service uses operations research to plan emergency mass dispensing and vaccination clinics. Interfaces 36(6):569-579
2. Adan I, van der Wal J (2011) Mean Values Techniques. In: Boucherie and van Dijk (eds) Queueing networks: a fundamental approach. Springer, New York, NY, USA
3. Albin SL, Barrett J, Ito D, Mueller JE (1990) A queueing network analysis of a health center. Queueing Systems 7:51-61
4. American Society of Anesthesiologists, www.asahq.org/For-Members/Clinical-Information/ASA-Physical-Status-Classification-System.aspx, retrieved April 20, 2011
5. Asaduzzaman Md, Chaussalet TJ, Robertson NJ (2010) A loss network model with overflow for capacity planning of a neonatal unit. Annals of Operations Research 178:67-76
6. Asaduzzaman Md, Chaussalet TJ, Adeyemi S, et al (2010) Towards effective capacity planning in a perinatal network centre. Archives of Disease in Childhood Fetal Neonatal 95:F283-287

7. Bailey NTJ (1952) A study of queues and appointment systems in hospital out-patient departments, with special reference to waiting-times. Journal of the Royal Statistical Society. Series B-Statistical Methodology 14(2):185-199
8. Baskett F, Chandy KM, Muntz RR, Palacios FG (1975) Open, closed, and mixed networks of queues with different classes of customers. Journal of the Association for Computing Machinery 22(2):248-260
9. Bitran GR, Morabito R (1996) Survey open queueing networks: optimization and performance evaluation models for discrete manufacturing systems. Production and Operations Management 5(2):163-193
10. Blair EL, Lawrence CE (1981) A queueing network approach to health care planning with an application to burn care in New York state. Socio-Economic Planning Sciences 15(5):207-216
11. Boucherie RJ, van Dijk NM (2011) Queueing networks: a fundamental approach. Springer, New York, NY, USA
12. Brockmeyer E, Halström HL, Jensen A (1948) The life and works of A.K. Erlang. Translations of the Danish Academy of Technical Sciences
13. de Bruin AM, Bekker R, van Zanten L, Koole GM (2010) Dimensioning hospital wards using the Erlang loss model. Annals of Operations Research 178(1):23-43
14. Burke PJ (1956) The output of a queuing system. Operations Research 4(6):699-704
15. Buzacott JA, Shanthikumar JG (1993) Stochastic models of manufacturing systems. Prentice Hall, Englewood Cliffs, NJ, USA
16. Buzen JP (1973) Computational algorithms for closed queueing networks with exponential servers. Communications of the ACM 16(9):527-531
17. Chaussalet TJ, Xie H, Millard P (2006) A closed queueing network approach to the analysis of patient flow in health care systems. Methods of Information in Medicine 45(5):492-497
18. Cochran JK, Bharti A (2006) A multi-stage stochastic methodology for whole hospital bed planning under peak loading. International Journal of Industrial and Systems Engineering 1(1-2):8-36
19. Cochran JK, Bharti A (2006) Stochastic bed balancing of an obstetrics hospital. Health Care Management Science 9(1):31-45
20. Cochran JK, Roche KT (2008) A multi-class queueing network analysis methodology for improving hospital emergency department performance. Computers & Operations Research 36(5):1497-1512
21. Cohen JW (1982) The single server queue. 8th ed. North-Holland Publishing Company, Amsterdam, the Netherlands
22. Creemers S, Lambrecht M (2011) Modeling a hospital queueing network. In: Boucherie and van Dijk (eds) Queueing networks: a fundamental approach. Springer, New York, NY, USA
23. van Dijk NM, Kortbeek N (2009) Erlang loss bounds for OT–ICU systems. Queueing systems 63(1):253-280
24. Fackrell M (2009) Modelling healthcare systems with phase-type distributions. Health Care Management Science 12(1):11-26
25. Ferschl MB, Tung A, Sweitzer B, Huo D, Glick DB (2005) Preoperative clinic visits reduce operating room cancellations and delays. Anesthesiology 103(4):855-9
26. Free University, Department of Mathematics, Erlang-C Calculator, www.few.vu.nl/˜koole/ccmath/ErlangC/
27. Gordon WJ, Newell GF (1967) Closed queuing systems with exponential servers. Operations Research 15(2):254-265
28. Green LV, Kolesar PJ, Soares J (2001) Improving the SIPP approach for staffing service systems that have cyclic demands. Operations Research 49(4):549-564
29. Green LV (2006) Queueing analysis in healthcare. In: Hall (ed) Patient flow: reducing delay in healthcare delivery. Springer, New York, NY, USA
30. Green LV, Soares J, Giglio JF, Green RA (2006) Using queueing theory to increase the effectiveness of emergency department provider staffing. Academic Emergency Medicine 13(1):61-68
31. Gross D, Shortle JF, Harris CM (2008) Fundamentals of queueing theory. 4th ed. John Wiley & Sons, Hoboken, NJ, USA

32. Jackson JR (1957) Networks of waiting lines. Operations Research 5(4):518-521
33. Jiang L, Giachetti RE (2007) A queueing network model to analyze the impact of parallelization of care on patient cycle time. Health Care Management Science 11(3):248-261
34. Kelly FP (1979) Reversibility and stochastic networks. Available online via www.statslab.cam.ac.uk/˜frank/rsn.html
35. Kelly FP (1991) Loss networks. The Annals of Applied Probability 1(3):319-378
36. Kleinrock L (1967) Queuing systems: theory, vol. 1. John Wiley & Sons, New York, NY, USA
37. Kleinrock L (1976) Queueing systems: computer applications, vol. 2. John Wiley & Sons, New York, NY, USA
38. Koizumi N, Kuno E, Smith TE (2005) Modeling patient flows using a queueing network with blocking. Health Care Management Science 8(1):49-60
39. Latouche G, Ramaswami V (1999) Introduction to Matrix Analytic Methods in Stochastic Modeling. American Statistical Association and the Society for Industrial and Applied Mathematics, USA
40. Latouche G, Taylor P (2002) Matrix-analytic methods: theory and applications. Proceedings of the fourth international conference: Adelaide, Australia. Imperial College Press, London, UK
41. Law AM, Kelton WD (1991) Simulation modeling and analysis. McGraw-Hill New York, NY, USA
42. Lee DKK, Zenios SA (2009) Optimal capacity overbooking for the regular treatment of chronic conditions. Operations Research 57(4):852-865
43. Little JDC (1961) A proof for the queuing formula $L = \lambda W$. Operations Research 9(3):383-387
44. Nelson RD (1995) Probability, stochastic processes, and queueing theory: the mathematics of computer performance modelling. Springer, New York, NY, USA
45. Organisation for Economic Co-operation and Development www.oecd.org, retrieved on April 19, 2011.
46. Research Institute CHOIR, University of Twente, Enschede, the Netherlands (2011) ORchestra bibliography. www.utwente.nl/choir/en/orchestra/
47. Osorio C, Bierlaire M (2009) An analytic finite capacity queueing network model capturing the propagation of congestion and blocking. European Journal of Operational Research 196(3):996-1007
48. Patient Flow Improvement Center Amsterdam, Erlang-B calculator www.vumc.nl/afdelingen/pica/Software/erlang_b/
49. QtsPlus Software, ftp://ftp.wiley.com/public/sci_tech_med/queuing_theory/
50. Taha HA (1997) Operations research: an introduction. Prentice Hall, Englewood Cliffs, NJ, USA
51. Takagi H (2000) Analysis and application of polling models. In: Haring G, Lindemann C, Reiser M (eds) Performance Evaluation: Origins and Directions. Lecture Notes in Computer Science 1769, Springer Verlag, Berlin, Germany
52. Tijms HC (2003) A first course in stochastic models. John Wiley & Sons, Chichester, UK
53. Vanberkel PT, Boucherie RJ, Hans EW, Hurink JL, Litvak N (2010) A survey of health care models that encompass multiple departments. International Journal of Health Management and Information 1(1):37-69
54. Westbay Online Traffic Calculators, www.erlang.com/calculator
55. Whitt W (1983) The queueing network analyzer. The Bell System Technical Journal 62(9):2779-2815
56. Winston WL (1994) Operations research: applications and algorithms. 3th ed. Duxbury Press, Belmont, CA, USA
57. Wolff RW (1989) Stochastic modeling and the theory of queues. Prentice Hall, Englewood Cliffs, NJ, USA
58. Xie H, Chaussalet T, Rees M (2007) A semi-open queueing network approach to the analysis of patient flow in healthcare systems. IEEE Proceedings Twentieth IEEE International Symposium on Computer-Based Medical Systems: 719-724

59. Zachary S, Ziedins I (2011) Loss networks. In: Boucherie and van Dijk (eds) Queueing networks: a fundamental approach. Springer, New York, NY, USA
60. Zeng G (2003) Two common properties of the Erlang-B function, Erlang-C function, and Engset blocking function. Mathematical and Computer Modelling 37(12-13):1287-1296
61. Zonderland ME, Boer F, Boucherie RJ, de Roode A, van Kleef JW (2009) Redesign of a university hospital preanesthesia evaluation clinic using a queuing theory approach. Anesthesia and Analgesia 109(5):1612-1621
62. Zonderland ME, Boucherie RJ, Litvak N, Vleggeert-Lankamp LCAM (2010) Planning and scheduling of semi-urgent surgeries. Health Care Management Science 13(3):256-267