Queueing Syst (2009) 63: 253–280
DOI 10.1007/s11134-009-9149-2

Erlang loss bounds for OT–ICU systems

N.M. van Dijk · N. Kortbeek

Received: 22 December 2008 / Revised: 30 October 2009 / Published online: 26 November 2009

Abstract In hospitals, patients can be rejected at both the operating theater (OT) and the intensive care unit (ICU) due to limited ICU capacity. The corresponding ICU rejection probability is an important service factor for hospitals. Rejection of an ICU request may lead to health deterioration for patients, and for hospitals to costly actions and a loss of precious capacity when an operation is canceled.

There is no simple expression available for this ICU rejection probability that takes the interaction with the OT into account. With c the ICU capacity (number of ICU beds), this paper proves and numerically illustrates a lower bound by an $M|G|c|c$ system and an upper bound by an $M|G|c\text{-}1|c\text{-}1$ system, hence by simple Erlang loss expressions.

The result is based on a product form modification for a special OT–ICU tandem formulation and proved by a technically complicated Markov reward comparison approach. The upper bound result is of particular practical interest for dimensioning an ICU to secure a prespecified service quality. The numerical results include a case study.

Keywords Tandem queues · Markov reward approach · Health services · Capacity planning · Intensive care units · Operating rooms

Mathematics Subject Classification (2000) 60J27 · 60K25 · 90B22

N.M. van Dijk · N. Kortbeek (✉)
Operations Research Group, Department of Economics and Business, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands
e-mail: n.kortbeek@utwente.nl

N.M. van Dijk
e-mail: n.m.vandijk@uva.nl

1 Introduction

1.1 Motivation

Intensive care units (ICUs) and operating theaters (OTs) are critical components within hospitals. Patients are admitted to an ICU for intensive care (i.e., monitoring and artificial ventilation), because their vital functions are compromised and their lives are in danger. They may require an ICU bed directly, or for postoperative care after an invasive operation. But the ICU can become congested due to limited capacity (the finite number of beds) and as a result requests for ICU beds have to be rejected.

Because the OTs and ICUs are among their most expensive resources, hospitals aim to keep them highly utilized. The drawback of a higher ICU occupancy level is deterioration of accessibility. Thus the size of an ICU needs to be dimensioned carefully so as to secure a sufficient service level within budget.

For patients who need direct admission to the ICU (without having first undergone surgery, mostly emergency patients), the consequences of a rejection are obvious for both the patients and the hospital. For patients this may lead to further, possibly life-threatening, delays. For the hospital (or the public health care system) this rejection may result in additional costs because ad hoc solutions have to be found, such as transferring a patient who might be less critical to another location.

For patients who need to be admitted to an ICU for postoperative care, the intensive care is often strictly required in conjunction with the operation. Therefore, if no ICU bed is available, it is common practice not to admit such patients to the OT. In these cases, the operation is either postponed or canceled, even if the patient is already present. OT-patients account for a substantial percentage of all ICU patients (roughly 40% in the case study).

For these patients the consequences are obvious for both the patient and the hospital. Because surgeries that require postoperative ICU care are generally serious, cancellation may pose a severe health risk or have a major emotional impact. For the hospital in its turn, the cancellation may lead to an unutilized operating room and thus a loss of resource capacity, as most of the time it is not possible to start another operation immediately. Unfortunately, OT and ICU planning generally take place independently of each other, as surgeons have to schedule elective surgeries well in advance under the assumption that an ICU bed will be available.

The availability of ICU beds is thus a highly important factor, one that reflects the service quality of the hospital. Nevertheless, rough aggregate figures in a recent report for the Dutch Ministry of Health [9] indicated that the rejection percentages are still substantial, in the order of 10 to 15% (also see Sect. 5). This paper will therefore focus on this ICU rejection probability, both for patients who require postoperative care and for patients who enter the ICU directly.

1.2 Literature

Various authors have already argued that the ICU can be seen as a multi-server queue ([8, 14, 15, 17, 19, 23]). More precisely, different standard queueing systems are proposed with both infinite ([14, 23]) and finite capacities ([8, 17, 19]). Some papers

have paid particular attention to the non-exponential character of ICU sojourn times, and for example phase type distributions are fitted ([6, 8, 17, 23, 42]). Remarkably, though, none of these report on the effect of such precise fitting of the sojourn times. Most notably, in [19], experimental data for the ICU rejection probability were shown to be reasonably approximated by the loss probability of an $M|G|c|c$. The approximations show both under- (the majority) and overestimation. A survey of some of these models can be found in [42]. This reference also includes a double ICU with overflow. Recently, in [17], the overflow aspect (which is of interest in itself), was elegantly addressed by applying a so-called equivalent random method, which is well known in the queueing and telecommunication literature. The objective of [17] was to determine the total number of ICU beds required for a particular region. The objective of determining the number of beds has also been addressed in [5], but was based on deterministic scheduling and data analysis, and thus without queueing and rejection phenomena. Reference [20] is of a similar deterministic nature and uses a multi-mode job-shop model with blocking. Nevertheless, these results do not contain:

- Formal justification for why these standard-type one-dimensional queueing approximations are accurate (other than by the rough experimental data in [17] and [19] and simulation support in [17] and [14]).
- Inclusion of the OT and its admission rejection protocol at the OT (in [20], the coupling of the OT and the ICU is mentioned as being highly important for its "job-shop" scheduling). For one thing, one might argue that, in practice, OT arrivals are partly scheduled rather than random. In addition, the ICU sojourn times will generally be shorter for OT patients. Last but not least, it remains uncertain to which extent a standard (one-dimensional) queueing system is sufficiently representative for capturing the interaction between the OT and the ICU.
- Secure bounds for estimating the rejection probability from below or above, or a discussion as to whether the results give under- or overestimation.

1.3 A complication and objectives

There is one additional complication, both in practice and throughout this study. Even with the admission rejection protocol at the OT, patients requiring ICU care after an operation may still find that there is no bed available for one of two reasons:

- An emergency patient might arrive who requires priority over an OT patient with a reserved bed (due to an extremely critical condition, the impossibility of a transfer, or legislation).
- An unexpected complication during an operation could mean that the patient requires an ICU bed. Because this would not have been known in advance, no admission rejection would have taken place and no ICU bed would have been reserved.

For a realistic analysis of the rejection probability, this complication will have to be incorporated into an OT–ICU model.

Therefore, the main objectives of this paper are:

- To justify the Erlang loss expression as a simple analytical approximation for the ICU rejection probability for both patients who need direct admission and those who are admitted through the OT, taking the OT interaction into account.
- To establish a strict lower and upper bound for this probability.

1.4 Results and outline

In Sect. 2, we argue, and numerically support with the results from a case study, that the $M|G|c|c$ queue is a reasonable basic queueing model for approximating the ICU rejection probability. In addition, a coupled OT–ICU tandem queueing system is presented. This is our system of interest, and is referred to as the original system.

In Sect. 3, the $M|G|c|c$-approximation is formally justified by an analytical product form result for a slightly modified OT–ICU tandem queue. This product form result implies that the OT and ICU may indeed be analyzed separately. Though strongly related to results in the queueing network literature, these product form results can be regarded as new in their specific combined form (as is specified in more detail in Sect. 3).

In Sect. 4, the product form result from Sect. 3 is used to prove that the ICU rejection probability for the original OT–ICU system can be bounded from below and above by $M|G|c|c$ and $M|G|c\text{-}1|c\text{-}1$ queues. This proof relies upon application of an analytical Markov reward comparison approach, adopted from the literature, and technical verifications (bounding) of so-called bias-terms. These verifications and the comparison results are themselves of theoretical interest.

In Sect. 5, simulation is used to show these bounds on a numerical basis, and the practical usefulness of the results for dimensioning an ICU is shown for a case study. As the assumption of Poisson rather than scheduled arrivals at the OT will lead to a conservative rejection probability, the $M|G|c\text{-}1|c\text{-}1$ upper bound in particular seems to be of practical interest for securing a guaranteed rejection probability for real-life OT–ICU practices.

2 Model formulation

In this section we first present some motivational case data. Then, in line with the literature, we briefly discuss three basic queueing models. Next, a number of properties are defined to argue a finite tandem queueing model for the OT–ICU system. Finally, the practical usefulness of the Erlang loss system as an approximation is argued numerically. In Sects. 3 through 5 this usefulness will be supported by formal proofs and bounds.

2.1 Patient groups and a case study

Patient groups The ICU inflow consists of emergency patients (the majority) and elective patients, which can be further subdivided into various patient groups. However, as we are particularly interested in the effect of the limited ICU capacity and its interaction with the OT, we make a distinction between patients that need to visit the ICU after having undergone an operation and patients who enter the ICU directly without having had an operation. These patients will be referred to respectively as:

- OT (or type 1) patients
- Direct (or type 2) patients

The distinction between type 1 and type 2 patients will also be kept explicit at the ICU itself for two reasons: first, to make explicit the influence of direct patients on the OT and ICU patients, and second, because their average sojourn times in the ICU are significantly different.

Case study Data for a case study were collected over a one year period. The case study was done in the Groot Ziekengasthuis (GZG) in 's-Hertogenbosch, the Netherlands, and included data for the year 2005. It showed that a substantial portion of the patients admitted to the ICU were OT patients. The percentages of type 1 and type 2 patients were 39% and 61% respectively. The overall average sojourn time spent in the ICU was 5.2 days, roughly 4.0 days for type 1 patients and 6.0 days for type 2 patients. The average duration of an operation for type 1 patients was 4.0 hours. Other numbers for the case study were:

- OT capacity (number of operating rooms): 8,
- ICU capacity (number of beds): 12,
- ICU offered load: 85%.

Remark 1

1. The offered load of 85% is used rather fictitiously as a rough estimate. In practice there are only measurements on occupancy level, which exclude the rejected patients from the offered load. Measurements on rejections are available only occasionally, because attempts are usually made to find ad hoc solutions such as pre-discharges or transfers. These actions are not always registered, and certainly not as rejections.
2. Clearly, the number of operating rooms in use is not always fixed.

2.2 Basic queueing models

Poisson arrivals It is assumed that patients arrive at the ICU according to a Poisson flow process. For emergency patients the Poisson assumption seems highly justified by the fact that these patients arrive independently and "at random" (except for occasional accidents involving more than one person). For elective patients one might state that the Poisson process is not an appropriate assumption because arrivals are planned. But, seeing that OT planning is generally done without taking the availability of ICU capacities into account, it can be reasonably argued that these patients arrive independently at the ICU as well. A similar argument can be found in [17]: "However, a surgeon is not aware of the occupation of the ICU when planning operations. As only a fraction of 5% of operated patients require intensive care after the operation, the assumption of Poisson arrivals is reasonable." Furthermore, last-minute changes are frequently made to the OT schedule.

Basic queueing models As a first-order approximation for evaluating the ICU rejection probability, it seems quite natural and plausible to simply regard the ICU in isolation. More specifically, as a "standard" multi-server queueing system, ignoring the link between the OT and ICU. Although several models have been suggested,

they have not been compared in the literature ([8, 14, 15, 17, 19, 23]). In [37] we examine and argue which of these queueing models for the ICU in isolation seems most realistic for estimating the rejection probability due to a limited ICU capacity (with c the number of ICU beds):

- an $M|G|c|\infty$ queue,
- an $M|G|\infty|\infty$ queue,
- an $M|G|c|c$ queue.

In [37] we argued, and numerically illustrated, that the $M|G|c|c$ system (in line with the literature) seems to be by far the best simple queueing approximation of the ICU-rejection probability. But the question remains as to whether this assertion also holds true and can be formally justified when the OT is included. This question will be the main focus of Sect. 3.

2.3 The OT–ICU tandem model

A more extensive model was formulated to study the ICU-rejection probability at a more realistic level. This model, which incorporates the OT and its interaction with the ICU, is presented below and will be referred to as the *original model* throughout this paper.

For both OT patients and direct patients, we are interested in the probability of all ICU beds being occupied, when such a bed is requested (i.e., the ICU rejection probability). Initially, one might regard the ICU in isolation, as described in Sect. 2.2 and numerically illustrated in [37]. This has some modeling discrepancies.

First of all, an OT patient who requires an ICU bed after surgery cannot simply be rejected if no such bed is available. Therefore, it is common practice not even to admit such patients to the OT if no ICU bed is available. The rejection of OT patients thus takes place at the OT rather than at the ICU. Clearly, under a Poisson arrival assumption of OT patients, the PASTA (Poisson arrivals see time averages) property might be recalled. Due to the interaction between the OT and ICU, it no longer seems justified to regard the ICU as a standard loss queue for an OT patient whose operation has been completed and who is kept "on hold" in the recovery room until an ICU bed becomes available.

Furthermore, even under a Poisson assumption for the arrival process at the OT, the flow of OT patients from the OT into the ICU is no longer Poisson. Thus, a model that contains both the OT and ICU is required so as to capture the interaction between these two. A tandem model is therefore proposed (see Fig. 1).

Modeling assumptions We use the following modeling assumptions. In [37] each of these assumptions has been argued and justified by simulation as being quite reasonable for practical modeling. The exponentiality assumptions (4) and (5) in particular have been shown to be reasonable, because the ICU rejection probability (not the total delay, of course) appears to be "nearly" insensitive.

(1) Patients who do not require an ICU bed are not included.
(2) A Poisson arrival rate λ_1 of OT patients (type 1) at the OT.
(3) A Poisson arrival rate λ_2 of direct patients (type 2) at the ICU.

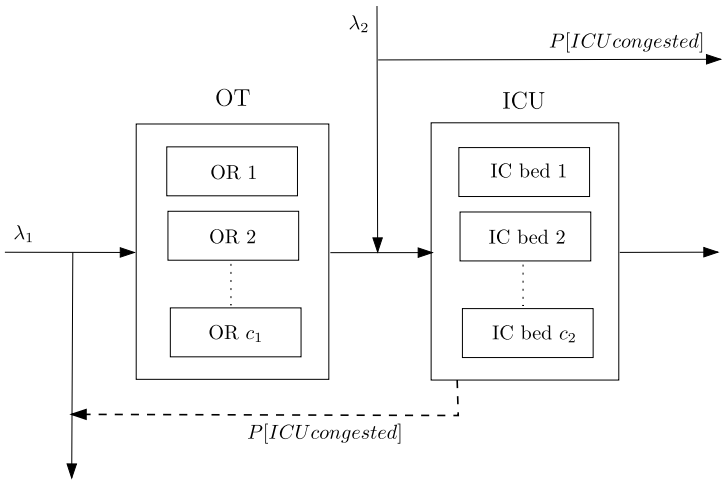

Fig. 1 Tandem model OT–ICU

(4) An exponential service time for surgery at the OT with parameter μ_1.
(5) An exponential sojourn time at the ICU with parameter μ_{21} for type 1 patients and μ_{22} for type 2 patients. Let $\tau_i = (\mu_{2i})^{-1}$.
(6) The OT has c_1 identical operating rooms with an infinite waiting facility; the ICU has a limited capacity for a maximum of c_2 patients and no waiting facility.

In addition to these assumptions, blocking protocol properties were defined that uniquely specify the dynamics of the system, so as to model realistic practice in the best possible manner. The "realistic" assumption (8) is an essential complicating factor for our analysis. The following blocking protocol is defined in the original system when all of the ICU beds are occupied:

(7) Type 1 patients are rejected upon arrival at the OT.
Type 2 patients are rejected upon arrival at the ICU.
(8) An ongoing operation is continued. Upon completion of an operation, the patient is kept in the OR (recovery room) and that operating room is suspended; that is, when the OR (or recovery room) is occupied, no new patient is brought in for surgery until an ICU bed is available again and the recovery room is idled.
When all ICU beds become occupied due to a completed operation, it is still possible to start a new operation in the OR that has become available for a patient who has already been accepted and is waiting.

No analytical solution For the original system described above, no analytical solution seems to have yet been reported for the joint steady-state probability of the occupancy levels at the OT and ICU, or, in relation to this for performance measures such as the ICU rejection probability. Therefore, in the next section we will first argue and consider a slightly modified description, in which only assumption (8) is modified. This modification leads to an analytic solution and a justification of an $M|G|c|c$ model.

Remark 2 ('New-operation' protocol) Alternatively, in (8) we could have assumed that no new operation would be started when completing an operation would lead to a congested ICU (even though an operating room is available). Both procedures appear to be used in practice depending on the hospital (and the urgency of the situation). The present protocol as contained in (8) seems more common, as it fills up the available capacities and thus progresses the processes as much as is physically feasible.

With reference to Remarks 3, 9, and 12 below, for both protocols the eventuating Result 1 can be shown to remain identical.

Remark 3 (Reservation of ICU bed) Note that our description does not assume that an accepted type 1 patient has a reserved ICU bed. Although in practice, such reservations may be made upon starting an operation when the need for postoperative intensive care is known in advance. However, these reservations can (or might have to) be violated in specific critical situations (see also Sect. 1.3).

2.4 Comparison of original model and $M|G|c|c$

The question arises as to what extent an $M|G|c|c$ can provide a good estimate of the ICU rejection probability of the original model.

As shown in Table 1, simulation results for the OT–ICU tandem system of interest as described in Sect. 2.3 seem to strongly support the accuracy of an $M|G|c|c$-approximation (with $c = c_2$). But despite this numerical evidence, so far the $M|G|c|c$-approximation still has the following three shortcomings:

(1) It has no formal justification that takes the OT interaction into account.
(2) It is not clear whether the $M|G|c|c$-approximation yields a lower or upper bound for the rejection probability. In fact, the experimental results in [19] are inconclusive. Simulation results (see Table 1) consistently show it to be a lower bound.
(3) Also, with regard to the partial presence of scheduled arrivals in practice, an upper bound for the ICU blocking probability of the system of interest would clearly be of more practical interest so as to secure a guaranteed service level.

The remainder of this paper will therefore be involved with each of these aspects, more specifically:

- Aspect 1 in Sect. 3.
- Aspects 2 and 3 in Sect. 4 on theoretical basis.
- Aspects 2 and 3 in Sect. 5 on numerical basis.

Table 1 The rejection probability of the $M|G|c|c$ queue and the simulated original model

Offered load ICU $\left(\frac{\lambda_1\tau_1+\lambda_2\tau_2}{c_2}\right)$	$M\|G\|c\|c$	Original model
0.70	0.06332	0.06388
0.75	0.08309	0.08349
0.80	0.10465	0.10534
0.85	0.12744	0.12797
0.90	0.15097	0.15170
0.95	0.17479	0.17616
1.00	0.19857	0.19928

3 Product form OT–ICU modification

As mentioned in Sect. 2.3, there is no known analytical solution for the original OT–ICU system of interest. However, an Erlang loss system seems to approximate the ICU rejection probability reasonably well, more precisely, an $M|G|c|c$ queue with $c = c_2$ the number of ICU beds. (In this section we consistently use c_2.)

Let us first provide some formal support for this approximation by slightly modifying the original system. Under this modification the OT–ICU system is shown to exhibit a product form solution. Furthermore, based on this modification, the lower and upper bounds for the ICU rejection probability are concluded and numerically illustrated in Sects. 4 and 5.

The modified OT–ICU tandem system The modified OT–ICU system is identical to the original OT–ICU system as described in Sect. 2.3 under the assumptions (1)–(6) as well as the realistic assumption (7) for the admission rejection at the OT. It only differs in assumption (8) for its blocking protocol at the OT when the ICU is congested. The following artificial modification of (8) is made:

(8′) When the ICU becomes congested, ongoing operations are immediately interrupted and no new patient is brought in for surgery. The operations are resumed as soon as the ICU is no longer congested.

This modification is meant for purely analytical purposes, as will appear below, and not for realistic modeling. Its influence is expected to be small, as operation durations are in general substantially shorter than the ICU sojourn times.

For the purpose of the insight, self-containment and its novel interest in and of itself (see Remark 5 below), in this section we will briefly study this modified OT–ICU tandem system for exponential sojourn times at the ICU and the simple blocking structure as used herein. An extension to more complex blocking structures and general ICU times can be found in [37].

The system defined by assumptions (1)–(7) and (8′) can be analyzed by studying the steady-state behavior of the corresponding continuous-time Markov chain. The following result can then be obtained.

Lemma 1 *Let $(n_1; m_1, m_2)$ denote that there are n_1 patients at the OT and m_i patients at the ICU of type i $(i = 1, 2)$. For the modified OT–ICU system, with*

$$F_1(n_1) = \prod_{k=1}^{n_1} \big[f_1(k)\big]^{-1} = \begin{cases} [n_1!]^{-1}, & n_1 \le c_1, \\ [c_1! c_1^{(n_1 - c_1)}]^{-1}, & n_1 > c_1 \end{cases} \tag{1}$$

a normalizing constant α and $m_1 + m_2 \le c_2$, we have for the steady-state probabilities $\pi(n_1; m_1, m_2)$:

$$\pi(n_1; m_1, m_2) = \alpha\, F_1(n_1) \left(\frac{\lambda_1}{\mu_1}\right)^{n_1} \frac{1}{m_1!} (\lambda_1 \tau_1)^{m_1} \frac{1}{m_2!} (\lambda_2 \tau_2)^{m_2}. \tag{2}$$

Proof It is sufficient to verify the global balance equation (3) for any state $(n_1; m_1, m_2)$. As argued below, it will be convenient to order the detailed in- and outrates as:

$$\left\{\begin{array}{ll}
\pi(n_1; m_1, m_2)\min\{n_1, c_1\}\mu_1 1_{(n_1>0)} 1_{(m_1+m_2<c_2)}+ & (3.1)\\
\pi(n_1; m_1, m_2)m_1\mu_{21} 1_{(m_1>0)}+ & (3.2)\\
\pi(n_1; m_1, m_2)m_2\mu_{22} 1_{(m_2>0)}+ & (3.3)\\
\pi(n_1; m_1, m_2)\lambda_1 1_{(m_1+m_2<c_2)}+ & (3.4)\\
\pi(n_1; m_1, m_2)\lambda_2 1_{(m_1+m_2<c_2)} & (3.5)
\end{array}\right\}$$
$$= \qquad (3)$$
$$\left\{\begin{array}{ll}
\pi(n_1-1; m_1, m_2)\lambda 1_{(n_1>0)} 1_{(m_1+m_2<c_2)}+ & (3.1)'\\
\pi(n_1+1; m_1-1, m_2)\min\{n_1+1, c_1\}\mu_1 1_{(m_1>0)}+ & (3.2)'\\
\pi(n_1; m_1, m_2-1)\lambda_2 1_{(m_2>0)}+ & (3.3)'\\
\pi(n_1; m_1+1, m_2)(m_1+1)\mu_{21} 1_{(m_1+m_2<c_2)}+ & (3.4)'\\
\pi(n_1; m_1, m_2+1)(m_2+1)\mu_{22} 1_{(m_1+m_2<c_2)} & (3.5)'
\end{array}\right\}$$

(It is noted that some of the indicator notation and implicit assumptions overlap. Nevertheless, the indicators are used to keep the 'boundary aspects' explicit, as used below).

The global balance equation (3) is ordered as if it can be decomposed into five local balances $(3.i) = (3.i)'$, $i = 1, \ldots, 5$. To this end, first observe that the local balances $(3.1) = (3.1)'$, $(3.4) = (3.4)'$ and $(3.5) = (3.5)'$ immediately follow in states for which $m_1 + m_2 = c_2$, as both the left-hand side and the right-hand side are then equal to 0. Each of $(3.i) = (3.i)'$ $(i = 1, \ldots, 5)$ can then be verified directly by substituting (2). □

Expression (2) can be rewritten in an expression for the steady-state distribution of n_1 patients at the OT and $m = m_1 + m_2$ at the ICU.

Lemma 2 *With*

$$\tau = \left[\frac{\lambda_1}{\lambda_1+\lambda_2}\right]\tau_1 + \left[\frac{\lambda_2}{\lambda_1+\lambda_2}\right]\tau_2 \qquad (4)$$

the mean sojourn time at the ICU, normalizing constant α, *and* $\lambda = (\lambda_1 + \lambda_2)$:

$$\pi(n; m) = \alpha F(n_1)\left(\frac{\lambda_1}{\mu_1}\right)^{n_1}\frac{1}{m!}(\lambda\tau)^m, \quad n_1 \geq 0 \text{ and } 0 \leq m \leq c_2 \qquad (5)$$

which factorizes in the steady-state distributions for:

- *an* $M|M|c_1|\infty$ *queue with arrival rate* λ_1 *and service rate* μ_1
- *an* $M|G|c_2|c_2$ *queue with arrival rate* λ *and mean service time* τ.

The ICU rejection probability for patients of both type 1 (*at the OT*) *and type* 2 (*at the ICU*) *is thus determined by the Erlang loss expression, with* $c = c_2$ *servers.*

Proof The proof is directly seen by summing over all possible values of m_1 and m_2 with $m = m_1 + m_2$ and noting that also the normalizing constant can be factorized in

$\alpha = \alpha_1 \cdot \alpha_2$, with α_1 and α_2 the normalizing constants for two separate queues as if in isolation. □

Remark 4 (Insensitivity) The product form expression (5) can already be regarded as a first illustration of insensitivity in that it only depends on the mean arrival rate and the mean sojourn time at the ICU, and not on the specific values for each patient type separately.

Remark 5 (Literature) Though related to results in the literature, as will be specified in more detail below, even in this exponential case expression (2) (and its consequence (5)) can be regarded as "new" because it includes and combines

- a non-reversible routing with blocking and
- multiple job-types.

More precisely, the classical product form papers [2, 4] and [3] do include multiple job-types but without any form of service or arrival blocking. Product form results for queueing networks with finite capacity constraints (and thus blocking) are also well known, among which by the famous book [13] and in relation to this, by [7, 16] or [21]. But the blocking results in these references rely upon the restrictive condition of a reversible routing. This condition is necessarily violated by the serial routing of a tandem system. For the situation of a single job-class, the present product form can directly be concluded from [10] or [34]. (In fact, in the latter references and [38], arrivals are artificially blocked when the second service station is congested to provide an easily computable performance bound for finite tandem queues. For the present OT–ICU system, in contrast, this blocking is natural.) The framework in [11] also allows multiple job-types and non-reversible blocking. But instead it requires a notion of balance for each job separately (job-local-balance). This notion is necessarily violated by "first-come, first-served" disciplines as in the first service station for the OT. Product forms for networks with both first-come, first-served (and thus non-insensitive) and insensitive (e.g., processor sharing type) disciplines are also known (e.g., [3]), but again, without blocking. No product form result seems to have been reported yet that combines these aspects, as in the present case in its specific form.

4 Lower and upper Erlang loss bounds

The product-form stationary distribution (5) factorizes into two terms that are the same as we would observe if the OT and ICU queues were fed independently by Poisson arrival processes with rates λ_1 and $\lambda_1 + \lambda_2$ respectively. For the stationary measures, it is therefore justified to consider the ICU in isolation as in Sect. 2.2. Of course, the whole queueing processes at the OT and ICU are clearly not stochastically independent.

It was also concluded that the $M|G|c|c$ system even provides an exact expression, although under the modification that OT service is interrupted when the ICU becomes congested. In the original system, continuing the OT service may put a patient in a

"waiting" position. Because this patient will immediately re-saturate the ICU when an ICU bed becomes available, we can expect a lower bound. In contrast, in the modified system, an OT service has to be completed or a direct arrival has to take place before the ICU can get congested again. Thus, the original system keeps the ICU more congested than the modified system. Intuitively this would imply that the $M|G|c|c$ computation provides a lower bound for the rejection probability. A more formal support for this intuitive statement would be of some practical interest so as to be sure that a certain capacity is "insufficient." For dimensioning purposes, though, it would be of even more practical interest if we could also provide an upper bound, so as to secure a sufficiently small rejection probability.

Conversely, it should therefore be noted that the modified OT–ICU tandem system only differs from the original OT–ICU tandem system for a patient who is undergoing surgery when the ICU becomes congested. One may expect that the effect of continuing OT service while the ICU is congested can be bounded as if one ICU bed is permanently reserved for a patient who is being kept "on hold." In turn, this negative effect (enlarging the rejection probability) cannot be more than if this patient were to permanently occupy one of the ICU beds, or if one ICU bed were to be permanently reserved for such a patient. By this modified system we expect to obtain an upper bound by an $M|G|c\text{-}1|c\text{-}1$ system.

Intuitively, the original system thus seems to be bounded from above and below by an $M|G|c\text{-}1|c\text{-}1$ and $M|G|c|c$ system. One might wonder whether these bounds can be proven formally, especially since in reality the arrival process of OT patients is partially scheduled rather than purely random. As a consequence, the rejection probability can actually be expected to be smaller than that for the original system in Sect. 3. Thus, a secure upper bound for this original model is of considerable interest for dimensioning the ICU in such a way that the "real" rejection probability meets a guaranteed norm. This section provides formal support for these bounds in the specific exponential case (see Remark 10 on the non-exponential case).

Result 1 (Erlang loss bounds) *With*

- $\mathbf{R}$ *the ICU rejection probability upon arrival at the OT for a type* 1 *patient and at the ICU for a type* 2 *patient for the original OT–ICU system as described in Sect.* 3,
- $\mathbf{B}(M|M|c|c)$ *the loss probability of an Erlang loss system with* c *servers with arrival rate* $\lambda = \lambda_1 + \lambda_2$ *and mean service time* τ *as by* (4),

we have

$$\mathbf{B}(M|M|c|c) \leq \mathbf{R} \leq \mathbf{B}(M|M|c\text{-}1|c\text{-}1). \tag{6}$$

Proof By (5) we have: $\mathbf{B}(M|M|c|c)$ is given by the Erlang loss expression as based on the tandem queueing system as described in Sect. 3; that is, with the OT suspended when the ICU is saturated. We can thus compare the original tandem queue in which the service (operation) at an OT is first completed before the OT is stopped with the tandem queue in which the OT is stopped when $m_1 + m_2 = c$. We will use the following abbreviations to refer to these different tandem queues:

- $O(TQ)$ for the original tandem queue.
- $S(TQ)(c)$ for the tandem queue as in Sect. 3 under the stop protocol with $c_2 = c$ the number of ICU beds.

Remark 6 (Sample path proof) At first glance, a proof based on sample path comparison and weak coupling, in line with the intuitive arguments given above and the literature on stochastic monotonicity (e.g., [1, 12, 18, 25–30, 32, 41]), might seem appropriate. But for finite tandem queues, as in [40] or [39], one can also give counterintuitive examples by which intuitive and monotonicity results are violated. More precisely, as shown in [36], Sect. 2.2, monotonicity results for finite tandem queues can still be proven but not in the direction required for showing an ordering of loss probabilities as in this article. In addition, the sample paths will not be free of overtaking, because multiple servers are involved and a distinction has been made between type 1 and type 2 patients. Patients would thus have to be interchanged stochastically. Finally, such a proof would also have to apply for both exponential and non-exponential sojourn times. The technical details can therefore be expected to be highly complicated and have not as yet been established.

Analytical proof by Markov Reward Approach An analytical proof therefore follows here, one that is based on a Markov Reward Approach (first used in [38] and surveyed in [35]). To this end, with reference to Remark 10 below, we will limit the proof to the exponential case. For presentational convenience and without loss of generality, we also assume there is only one OT (see Remark 8).

Clearly, for the $S(TQ)(c)$ and $S(TQ)(c-1)$ it suffices to keep track of the state $(n_1; m_1, m_2)$. For the $O(TQ)$, in contrast, we need to keep track of the status of the patient who was already in the OT or went into OT service when the ICU became saturated. To this end, with n_1, m_1, m_2 the numbers as in Sect. 3, let the state be described by:

$(n_1; m_1, m_2)$ for $m_1 + m_2 < c$

$(n_1, \theta; m_1, m_2)$ for $m_1 + m_2 = c$ with

$\theta = 0$ when the OT server is suspended;

$\theta = 1$ when there is one patient at the OT server who is continuing its OT service;

$\theta = 2$ when there is one patient who has completed its OT service and is waiting at the OT server for an ICU server to become available with the OT server suspended.

Remark 7

- At first glance, one might wonder why the states $(n_1, 0; m_1, m_2)$ and $(n_1, 1; m_1, m_2)$ with $m_1 + m_2 = c$ have to be distinguished. Note that a state with $\theta = 2$ can only be reached out of a state with $\theta = 1$ but not out of a state with $\theta = 0$ as specified below in (7).
- Note here that the situation $\theta = 0$ arises when a bed in the ICU idles and is immediately occupied by a patient who has completed its OT service and has been waiting at the OT server, which is out of a state with $\theta = 2$, as also specified in (7).

- Furthermore, also note that when congestion at the ICU ends (i.e., when $m_1 + m_2$ becomes less than c) and when there is no patient who has already completed its OT service and is waiting at the OT server (i.e., when $m_1 + m_2 < c$ remains), the OT server will become operative (will become available and take a new patient into service when $n_1 > 0$). The state $(n_1; m_1, m_2)$ with $m_1 + m_2 < c$ can thus always be implicitly regarded as with $\theta = 1$.

Remark 8 In the case of multiple operating rooms we would need a similar triple specification $\theta_i = 0, 1, 2$ for each operating room i. This would substantially increase the already heavy notation and verifications later on, without further insights. The proof is therefore restricted to a single operating room.

Let q denote the corresponding transition rates for the $O(TQ)$. More precisely, that is:

$$\begin{array}{ll}
q[(n_1; m_1, m_2), (n_1 + 1; m_1, m_2)] = \lambda_1, & m_1 + m_2 < c, \\
q[(n_1; m_1, m_2), (n_1; m_1, m_2 + 1)] = \lambda_2, & m_1 + m_2 + 1 < c, \\
q[(n_1; m_1, m_2), (n_1, 1; m_1 + 1, c)] = \lambda_2, & m_1 + m_2 + 1 = c, \\
q[(n_1; m_1, m_2), (n_1 - 1; m_1 + 1, m_2)] = \mu_1, & n_1 > 0, m_1 + m_2 + 1 < c, \\
q[(n_1; m_1, m_2), (n_1 - 1, 1; m_1 + 1, m_2)] = \mu_1, & n_1 > 1, m_1 + m_2 + 1 = c, \\
q[(n_1; m_1, m_2), (n_1 - 1, 0; m_1 + 1, m_2)] = \mu_1, & n_1 = 1, m_1 + m_2 + 1 = c, \\
q[(n_1, 1; m_1, m_2), (n_1, 2; m_1, m_2)] = \mu_1, & n_1 > 0, m_1 + m_2 = c, \\
q[(n_1; m_1, m_2), (n_1; m_1 - 1, m_2)] = m_1\mu_{21}, & m_1 + m_2 < c, \\
q[(n_1; m_1, m_2), (n_1; m_1, m_2 - 1)] = m_2\mu_{22}, & m_1 + m_2 < c, \\
q[(n_1, \theta; m_1, m_2), (n_1; m_1 - 1, m_2)] = m_1\mu_{21}, & m_1 + m_2 = c, \theta = 0, 1, \\
q[(n_1, \theta; m_1, m_2), (n_1; m_1, m_2 - 1)] = m_2\mu_{22}, & m_1 + m_2 = c, \theta = 0, 1, \\
q[(n_1, 2; m_1, m_2), (n_1 - 1, 0; m_1, m_2)] = m_1\mu_{21}, & m_1 + m_2 = c, \theta = 2, \\
q[(n_1, 2; m_1, m_2), (n_1 - 1, 0; m_1 + 1, m_2 - 1)] = m_2\mu_{22}, & m_1 + m_2 = c, \theta = 2.
\end{array} \tag{7}$$

When convenient, such as for unification, we also use the identifications:

$$\begin{cases} (n_1; m_1, m_2) = (n_1, 1; m_1, m_2) = (n_1, \theta; m_1, m_2) & \text{for } m_1 + m_2 < c, \\ (0; m_1, m_2) \ = (0, 1; m_1, m_2) & \text{for } m_1 + m_2 = c. \end{cases} \tag{8}$$

The state space becomes:

$$S = \big\{(n_1, \theta; m_1, m_2) \mid \theta = 1 \text{ for } m_1 + m_2 < c \text{ and } \theta = 0, 1, 2 \text{ for } m_1 + m_2 = c\big\}. \tag{9}$$

In order to compare the $O(TQ)$ with the $S(TQ)$ systems, we first introduce a discrete-time cumulative reward structure. (The probability transition matrix P and the function $\mathbf{V}^k$, as will be defined below, are needed for Result 1 and Lemma 3 below and its proof.) Let h be a sufficiently small positive number such that

$$h \le [\lambda_1 + \lambda_2 + \mu_1 + c\mu_{21} + c\mu_{22}]^{-1} \tag{10}$$

and with the transition rates q as given by (7) and define the one-step transition matrix P at the same state space S of the $O(TQ)$ by

$$P\big[(n_1,\theta;m_1,m_2);(n_1,\theta;m_1,m_2)'\big] = \begin{cases} hq[(n_1,\theta;m_1,m_2);(n_1,\theta;m_1,m_2)'] \\ \quad \text{for } (n_1,\theta;m_1,m_2)' \neq (n_1,\theta;m_1,m_2), \\ 1-h\sum_{(n_1,\theta;m_1,m_2)'} q[(n_1,\theta;m_1,m_2);(n_1,\theta;m_1,m_2)'] \\ \qquad \text{for } (n_1,\theta;m_1,m_2)' = (n_1,\theta;m_1,m_2). \end{cases} \tag{11}$$

Now for a given reward rate function r at S, define the expected cumulative reward functions V^k at S, for $k = 0, 1, 2, \ldots$, by

$$\begin{cases} \mathbf{V}^0(n_1,\theta;m_1,m_2) = 0 \\ \mathbf{V}^{k+1}(n_1,\theta;m_1,m_2) = hr(n_1,\theta;m_1,m_2) \\ \quad + \sum_{(n_1,\theta;m_1,m_2)'} P[(n_1,\theta;m_1,m_2);(n_1,\theta;m_1,m_2)'] \\ \quad \times \mathbf{V}^k(n_1,\theta;m_1,m_2)' \qquad \text{for any } (n_1,\theta;m_1,m_2). \end{cases} \tag{12}$$

Then by standard stochastic dynamic programming (e.g., see [22, 31]), $\mathbf{V}^k(n_1,\theta;m_1,m_2)$ represents the expected cumulative reward over k steps for the discrete-time Markov chain with one-step transition matrix P for each step and one-step reward function hr, as by (an expected) step length of time h. Furthermore, by virtue of the well-known uniformization method (e.g., [31]),

$$\lim_{k\to\infty} k^{-1}h^{-1}\mathbf{V}^k(n_1,\theta;m_1,m_2) = G, \tag{13}$$

where G represents the expected average reward per time unit of the continuous-time Markov chain at S with transition rates q and reward rate r. More specifically, our measure of interest, the rejection probability $\mathbf{R}$, is obtained by

$$\begin{cases} r(n_1,\theta;m_1,m_2) = 1_{(m_1+m_2=c)}, \\ G = \mathbf{R}. \end{cases} \tag{14}$$

The following result (Lemma 3) is now adopted from [35], Theorem 2.1, for comparing the expected average reward G and $\bar{G}$ of two continuous-time ergodic Markov chains with transition rates $q[i,j]$ and $\bar{q}[i,j]$ respectively for a transition from a state i into another state $j \neq i$, reward rates r and $\bar{r}$ and abstract state spaces S and $\bar{S}$, where it is essential that one state space covers the other, say:

$$\bar{S} \subseteq S. \tag{15}$$

Lemma 3

$$G \geq (\leq)\ \bar{G} \tag{16}$$

if for all k and $i \in \bar{S}$:

$$[r-\bar{r}](i) + \sum_{j\in S}\big[q[i,j]-\bar{q}[i,j]\big]\big[\mathbf{V}^k(j)-\mathbf{V}^k(i)\big] \geq (\leq)\ 0. \tag{17}$$

(*As will be essential below, note that we only need to include states $j \in S$ for which $q(i,j) > 0$ with $i \in \bar{S}$.*)

We apply Lemma 3 for both inequalities in (6), the lower and upper bounds. For both applications the transition rates q and state space S correspond to those for the original system $O(TQ)$ as specified by (9). The functions $\mathbf{V}^k$ and reward rate r in (17) are hereby specified by (11)–(13) and $G = \mathbf{R}$.

Lower bound in (6) Here, let the $\bar{q}$-system in Lemma 3 correspond to the $S(TQ)(c)$ as described in Sect. 3 and note that the corresponding state space $\bar{S}$ satisfies (15) if written as:

$$\bar{S} = \big\{(n_1, \theta; m_1, m_2) \,|\, \theta = 1 \text{ for } m_1 + m_2 < c \text{ and } \theta = 0 \text{ for } m_1 + m_2 = c\big\}. \quad (18)$$

(In fact, as the configuration $(n_1; m_1, m_2)$ in this case uniquely determines whether the OT server is operative ($\theta = 1$) or not ($\theta = 0$), in Sect. 3 we simply limited the notation to: $(n_1; m_1, m_2)$.) Furthermore, by choosing $\bar{r}(n_1, \theta; m_1, m_2) = r(n_1, \theta; m_1, m_2)$ as by (14), the average value $\bar{G}$ represents the rejection probability in the $S(TQ)(c)$ system. Hence, by Sect. 3 (Lemma 2):

$$\bar{G} = \mathbf{B}(M|G|c|c). \quad (19)$$

Hence, it suffices to verify (17) with $\geq$ sign for any $(n_1, \theta; m_1, m_2) \in \bar{S}$. By comparing the system behavior of the $O(TQ)$ and $S(TQ)(c)$, we directly conclude that:

$$\bar{q}\big[(n_1, \theta; m_1, m_2); (n_1, \theta; m_1, m_2)'\big] = q\big[(n_1, \theta; m_1, m_2); (n_1, \theta; m_1, m_2)'\big]$$

for any

$$\begin{cases} (n_1, \theta; m_1, m_2) \in \bar{S} & \text{with } m_1 + m_2 + 1 < c \quad \text{(hence with } \theta = 1\text{)}, \\ (n_1, \theta; m_1, m_2) \in \bar{S} & \text{with } m_1 + m_2 = c \quad\;\; \text{(hence only with } \theta = 0\text{)} \end{cases} \quad (20)$$

and any $(n_1, \theta; m_1, m_2)'$. Since also $\bar{r} = r$, condition (17) is thus trivially satisfied for all states $(n_1, \theta; m_1, m_2) \in \bar{S}$ with $m_1 + m_2 + 1 \neq c$.

For $m_1 + m_2 + 1 = c$ (hence with $\theta = 1$), a difference in the dynamics of the system arises as the OT server continues to work in the $O(TQ)$ while it stops in the $S(TQ)(c)$ system when the ICU gets congested, so that in state $(n_1; m_1, m_2)$ with $m_1 + m_2 + 1 = c$,

$$\mu_1 1_{(n_1>0)} = \begin{cases} q[(n_1, 1; m_1, m_2); (n_1 - 1, 1; m_1 + 1, m_2)], \\ \bar{q}[(n_1, 1; m_1, m_2); (n_1 - 1, 0; m_1 + 1, m_2)] \end{cases} \quad (21)$$

and

$$\lambda_2 = \begin{cases} q[(n_1, 1; m_1, m_2); (n_1, 1; m_1, m_2 + 1)], \\ \bar{q}[(n_1, 1; m_1, m_2); (n_1, 0; m_1, m_2 + 1)]. \end{cases} \quad (22)$$

As a consequence, for $(n_1, \theta; m_1, m_2) \in \bar{S}$ with $m_1 + m_2 + 1 = c$, and using again that $\bar{r} \equiv r$, the left-hand side of (17) becomes:

$$
\begin{aligned}
&\sum_{(n_1,\theta;m_1,m_2)'} \Big[q\big[(n_1,\theta;m_1,m_2);(n_1,\theta;m_1,m_2)'\big] \\
&\qquad - \bar{q}\big[(n_1,\theta;m_1,m_2);(n_1,\theta;m_1,m_2)'\big]\Big] \\
&\qquad \times \big[\mathbf{V}^k\big((n_1,\theta;m_1,m_2)'\big) - \mathbf{V}^k(n_1,\theta;m_1,m_2)\big] \\
&= 1_{(m_1+m_2+1=c)}\mu_1 1_{(n_1>0)} \\
&\qquad \times \big[\mathbf{V}^k(n_1-1,1;m_1+1,m_2) - \mathbf{V}^k(n_1-1,0;m_1+1,m_2)\big] \\
&\qquad + 1_{(m_1+m_2+1=c)}\lambda_2\big[\mathbf{V}^k(n_1,1;m_1,m_2+1) - \mathbf{V}^k(n_1,0;m_1,m_2+1)\big].
\end{aligned}
\tag{23}
$$

By Lemma 4 below, the right-hand side of equality (23) can be estimated from below by 0. Hence, by Lemma 3: $G \geq \bar{G}$, which by (14) and (19) completes the proof of the lower bound inequality in (6).

Upper bound in (6) Now let the $\bar{q}$-system in Lemma 3 correspond to the $S(TQ)(c-1)$ system as also described in Sect. 3, but with $c-1$ rather than c ICU beds. The corresponding state space $\bar{S}$ could then be given by (18) with c replaced by $c-1$. However, in that case, the inclusion (15) would no longer be satisfied as $\theta \neq 0$ for $m_1+m_2=c-1$ in S. Therefore, note that due to the exponential assumption for the OT server, we can also describe $S(TQ)(c-1)$ as in Sect. 3 except that the OT server always continues to operate (i.e., $\theta = 1$) but that a type 1 patient who completes its OT service has to undergo a new OT service when the ICU is congested, that is, when $m_1+m_2=c-1$. According to this description the transition rates remain equal to those for the stop protocol in Sect. 3, so that the product form result still applies. (In fact, more general equivalencies of stop and recirculate protocols can be proven under product form conditions, as shown in [33].) Now clearly, (15) is satisfied by

$$\bar{S} = \big\{(n_1;m_1,m_2) \mid m_1+m_2 \leq c-1\big\} \subseteq S.$$

Furthermore, by choosing $\bar{r}(n_1;m_1,m_2) = 1_{(m_1+m_2=c-1)}$, the average reward per time unit $\bar{G}$ represents the rejection rate of type 1 patients in the $S(TQ)(c-1)$ system. Hence, by Sect. 3 (Lemma 2):

$$\bar{G} = \mathbf{B}(M|G|c-1|c-1). \tag{24}$$

Hence, it suffices to verify condition (17) with $\leq$ sign for any $(n_1;m_1,m_2) \in \bar{S}$. As the OT server is also not suspended in the $O(TQ)$ when $m_1+m_2=c-1$, we can now directly conclude that for $(n_1;m_1,m_2) \in \bar{S}$ with $m_1+m_2 < c-1$:

$$
\begin{aligned}
&\bar{r}(n_1;m_1,m_2) = r(n_1;m_1,m_2) \quad \text{and} \\
&\bar{q}\big[(n_1;m_1,m_2);(n_1;m_1,m_2)'\big] = q\big[(n_1;m_1,m_2);(n_1;m_1,m_2)'\big].
\end{aligned}
\tag{25}
$$

For $m_1+m_2=c-1$, however, a difference arises in not only the transition state rates but also the reward rates, because in the $S(TQ)(c-1)$ system, a rejection reward

is incurred. More precisely, for $m_1 + m_2 = c - 1$:

$$\begin{aligned}
&r(n_1; m_1, m_2) = 0,\\
&\bar{r}(n_1; m_1, m_2) = 1,\\
&q\big[(n_1; m_1, m_2); (n_1 + 1; m_1, m_2)\big] = \lambda_1,\\
&\bar{q}\big[(n_1; m_1, m_2); (n_1 + 1; m_1, m_2)\big] = 0,\\
&q\big[(n_1; m_1, m_2); (n_1, 1; m_1, m_2 + 1)\big] = \lambda_2,\\
&\bar{q}\big[(n_1; m_1, m_2); (n_1, 1; m_1, m_2 + 1)\big] = 0,\\
&q\big[(n_1; m_1, m_2); (n_1 - 1, 1; m_1 + 1, m_2)\big] = \mu_1 1_{(n_1>0)},\\
&\bar{q}\big[(n_1; m_1, m_2); (n_1 - 1, 1; m_1 + 1, m_2)\big] = 0
\end{aligned} \tag{26}$$

so that the left-hand side of (17) becomes:

$$\begin{aligned}
&\big[r(n_1; m_1, m_2) - \bar{r}(n_1; m_1, m_2)\big]\\
&\quad + \sum_{(n_1;m_1,m_2)'} \big[q\big[(n_1; m_1, m_2); (n_1; m_1, m_2)'\big] - \bar{q}\big[(n_1; m_1, m_2); (n_1; m_1, m_2)'\big]\big]\\
&\quad \times \big[\mathbf{V}^k((n_1; m_1, m_2)') - \mathbf{V}^k(n_1; m_1, m_2)\big]\\
&= -1_{(m_1+m_2=c-1)}\\
&\quad + 1_{(m_1+m_2=c-1)}\lambda_1\big[\mathbf{V}^k(n_1 + 1; m_1, m_2) - \mathbf{V}^k(n_1; m_1, m_2)\big]\\
&\quad + 1_{(m_1+m_2=c-1)}\lambda_2\big[\mathbf{V}^k(n_1, 1; m_1, m_2 + 1) - \mathbf{V}^k(n_1; m_1, m_2)\big]\\
&\quad + 1_{(m_1+m_2=c-1)}\mu_1 1_{(n_1>0)}\big[\mathbf{V}^k(n_1 - 1, 1; m_1 + 1, m_2) - \mathbf{V}^k(n_1; m_1, m_2)\big].
\end{aligned} \tag{27}$$

In Lemma 4 below, each of the three difference terms in $\mathbf{V}^k$ is non-negative but still estimated from above by $[\lambda_1 + \lambda_2 + \mu_1]^{-1}$. As a consequence, the right-hand side of (27) can still be estimated from above by 0. Hence, by Lemma 3: $G \le \bar{G}$. By (14) and (24) this completes the proof of the upper bound inequality in (6). □

Lemma 4 *With $Q = [\lambda_1 + \lambda_2 + \mu_1]^{-1}$, for all states $(n_1, \theta; m_1, m_2)$ is such that both states in the difference term below are contained in S and all $t \ge 0$.*

With $\theta = 1$ for $m_1 + m_2 < c$ and $\theta = 0, 1, 2$ for $m_1 + m_2 = c$:

$$0 \le \mathbf{V}^t(n_1 + 1, \theta; m_1, m_2) - \mathbf{V}^t(n_1, \theta; m_1, m_2) \le Q. \tag{28}$$

With $\theta = 1$ for $m_1 + m_2 + 1 < c$ and $\theta = 0, 1, 2$ for $m_1 + m_2 + 1 = c$:

$$0 \le \mathbf{V}^t(n_1, \theta; m_1 + 1, m_2) - \mathbf{V}^t(n_1, 1; m_1, m_2) \le Q, \tag{29}$$

$$0 \le \mathbf{V}^t(n_1, \theta; m_1, m_2 + 1) - \mathbf{V}^t(n_1, \theta; m_1, m_2) \le Q, \tag{30}$$

$$0 \le \mathbf{V}^t(n_1 - 1, \theta; m_1 + 1, m_2) - \mathbf{V}^t(n_1, 1; m_1, m_2) \le Q. \tag{31}$$

With $m_1 + m_2 = c$:

$$0 \le \mathbf{V}^t(n_1, 2; m_1, m_2) - \mathbf{V}^t(n_1, 1; m_1, m_2) \le Q, \tag{32}$$

$$0 \le \mathbf{V}^t(n_1, 1; m_1, m_2) - \mathbf{V}^t(n_1, 0; m_1, m_2) \le Q, \tag{33}$$

$$0 \le \mathbf{V}^t(1, 2; m_1, m_2) - \mathbf{V}^t(0, 1; m_1, m_2) \le Q. \tag{34}$$

Proof See the Appendix. □

Remark 9 Under the alternative assumption instead of (8), as in Remark 2, that no new operation will be started when the ICU has become congested due to the completion of an operation, the fifth term in (7) will become:

$$q\big[(n_1; m_1, m_2), (n_1 - 1, 0; m_1 + 1, m_2)\big] = \mu_1, \quad m_1 + m_2 + 1 = c$$

The right-hand side of (23) then reduces to

$$1_{(m_1+m_2+1=c)}\mu_1 1_{(n_1>0)}\big[\mathbf{V}^k(n_1 - 1, 1; m_1 + 1, m_2) - \mathbf{V}^k(n_1 - 1, 0; m_1 + 1, m_2)\big].$$

In other words, the second term has vanished. Nevertheless, as shown in the proof of Lemma 4 in the Appendix, the same inequalities in Lemma 4 will still be required, due to the interdependencies of the difference terms in (28)–(34) in Lemma 4.

Remark 10 With simulation, the rejection probability appears to be rather insensitive ([37]) and to be well within the Erlang loss bounds (Sect. 5). The simulation experiments also seem to consistently support the validity of the bounds for the non-exponential case ([37]). Because the Erlang loss probabilities are insensitive, these bounds can thus be conjectured to be insensitive as well. To provide a formal proof for the non-exponential case as well, a proof can be thought of along the lines presented here and by using phase-type distributions, as in [32]. However, the notation and technical details will then be substantially more complicated. These details have not as yet been worked out and as such no formal support has been provided.

5 Numerical results

This section contains numerical support for the results obtained in the previous sections. First, in Sect. 5.1, the bounds of Result 1 are supported. Next, a dimensioning application for the case study of Sect. 2.1 is presented.

5.1 Bounds

To support the bounding in Result 1, Table 2 shows numerical results and the comparison with simulation for the original OT–ICU system. The situations are within the range of realistic figures, like recently reported by the Dutch Ministry of Health. This report states that roughly 10% of patients requesting an ICU bed are rejected outright, 4% are admitted elsewhere and 3% are admitted because of a pre-discharge or a transfer. Furthermore, in a recent Dutch study [24], an occupancy level of 75% is suggested as the norm. Typically, the number of ICU beds is in the order of:

Table 2 $M|G|c|c$ and $M|G|c\text{-}1|c\text{-}1$ bounds and simulated rejection probability

Offered load	Number of beds	$M\|G\|c\|c$	Original model	$M\|G\|c\text{-}1\|c\text{-}1$
80%	10	0.12166	0.12256	0.17314
	20	0.06441	0.06556	0.08606
	30	0.04012	0.04175	0.05225
	40	0.02684	0.02865	0.03447
90%	10	0.16796	0.16925	0.22430
	20	0.10921	0.11051	0.13623
	30	0.08188	0.08476	0.09909
	40	0.06537	0.06870	0.07771

- 10–20 With an offered load of roughly 60–80% for small and medium size hospitals.
- 20–40 With an offered load of roughly 80–90% for larger hospitals.

The simulation results consistently support the lower and upper bounds. Particularly for situations with smaller rejection probabilities, say in the order of 5–10%, as is more natural in larger hospitals that generally have higher occupancy levels, the bounds might even be regarded as being reasonably accurate (in absolute sense). The results are useful for practical purposes such as to guarantee a sufficiently small rejection percentage by the upper bound. But the results might still be of practical interest for smaller hospitals (which generally have larger rejection probabilities) in providing a realistic order of magnitude.

(In Table 2, the offered load applies to the original and $M|G|c|c$ model. For the $M|G|c\text{-}1|c\text{-}1$ model, the effective offered load is a factor $\frac{c}{c-1}$ larger, as the bounds are to be calculated with the same arrival and service parameters and the $M|G|c\text{-}1|c\text{-}1$ has one server less. The occupancies are determined by varying λ_1. The operating and ICU sojourn times are taken from the case study, i.e., Sect. 2.1.)

5.2 Application: case study

To provide an example of the application, a numerical example is shown for the case study presented in Sect. 2.1. For the case study data, the $M|G|c|c$ and $M|G|c\text{-}1|c\text{-}1$ computations (bounds) for varying offered load (by varying arrival rate λ_1) lead to the results shown in Table 3.

For the 85% offered load as measured in practice, the results lead to an upper bound of 0.172 and lower bound of 0.127 (with simulation result 0.128). Unfortunately, it is infeasible to make a direct comparison with actual practice, as only occasional measurements were available for rejected OT and ICU requests. The limited availability of measurements seems to be a general problem within hospital practice (see also Sect. 2.1).

Nevertheless, this high figure was in line with the general perception of the "rejection" frequency when taking into account all occasions in which ad hoc solutions had to be organized. The results of Table 3 can be used by the hospital to obtain insight

Table 3 $M|G|c|c$ and $M|G|c\text{-}1|c\text{-}1$ bounds and simulated rejection probability for the case study

Offered load	$M\|G\|c\|c$	Original model	$M\|G\|c\text{-}1\|c\text{-}1$
0.60	0.03127	0.03135	0.05380
0.65	0.04589	0.04634	0.07399
0.70	0.06332	0.06408	0.09657
0.75	0.08309	0.08383	0.12082
0.80	0.10465	0.10525	0.14610
0.85	0.12744	0.12782	0.17183
0.90	0.15097	0.15155	0.19757

Table 4 Dimensioning for required rejection probability

Number of beds	Rejection probability
10	0.282
12	0.172
14	0.091
16	0.041
18	0.015
19	0.008
20	0.004

into the influence of considerable changes in arrival rates or ICU sojourn times. With regard to achieving an acceptable rejection figure, questions can be encountered as to how much the mean ICU sojourn time should be reduced or how much the arrival rate could grow given the current ICU capacity.

As a most directly conceivable application of the secure $M|G|c\text{-}1|c\text{-}1$ upper bound computation, given the current offered load, the required number of ICU beds could be computed to guarantee a specified rejection probability R (see Table 4), such as:

- 16 beds for a maximum of 5%.
- 19 beds for a maximum of 1%.

6 Evaluation

The rejection probability for ICU beds, due to the limited ICU capacity is a factor of considerable interest within hospitals. It affects both emergency patients who require intensive care directly and monitoring and patients who have undergone a severe operation. Although it is a factor that involves high capacity costs, it is also one that may put lives at risk on a daily basis.

In practice, rejection figures for different types of patients may not be readily available, because attempts are made to fine ad hoc solutions such as transfer to another hospital, temporary placement in a medium or regular care facility, or pre-discharge of a less critical patient. The support this paper provides is twofold:

- In a practical way: by the Erlang loss expression for approximating the ICU rejection probability at an OT and ICU.

- In a theoretical and practical way: by justifying this expression as an upper (as well as lower) bound so as to secure a sufficiently low rejection percentage.

As such, it can be regarded as a present-day tribute to Erlang's pioneering.

Appendix A: Proof of Lemma 4

Proof The proofs for the difference terms (28)–(31) essentially go similarly to those in [38] (for the single server case) or [40] (for the multi-server case), except that in these references θ does not appear (or only takes the value $\theta = 1$ as under the stop protocol, that there are no direct arrivals at the ICU ($\lambda_2 = 0$) and that there is no distinction between type 1 and type 2 jobs (patients) at the ICU (station 2)). However, as the rejection and transition mechanism at the OT (station 1) are only influenced by the total number of jobs (patients) at station 2: $m_1 + m_2$, for fixed $\theta = 1, \theta = 2$ or $\theta = 0$ the steps for the proofs of (28)–(31) are primarily notationally more complex but essentially similar to those in these references. Only changes of the extra component θ need to be included. The technical details of these steps for each of the difference terms in (28)–(31) with $\theta = 1, 2, 0$ would be (too) lengthy and cumbersome. Nevertheless:

- for the sake of self-containment,
- to show how the recurrence relations can be derived,
- to illustrate how these three extensions (these θ-values, as well as direct arrivals at station 2 and type 1 and type 2 distinctions) can be included, and
- to show how the difference terms are related,

we will present the steps and proof for (28) in detail while leaving those for (29)–(30) to the reader. Those for (32), (33), and (34) will also be given in detail, as these types of difference terms are new. The proofs all go by induction in t. Clearly, (28)–(34) hold for $t = 0$ as $\mathbf{V}^0(\cdot) \equiv 0$. Assume that (28)–(34) hold for $t = k$. Let

$$\Delta^1 \mathbf{V}^t(n_1, \theta; m_1, m_2) = \mathbf{V}^t(n_1 + 1, \theta; m_1, m_2) - \mathbf{V}^t(n_1, \theta; m_1, m_2). \tag{35}$$

Inequalities (28)–(31) We need to verify (28) for $t = k + 1$. To this end, consider a state $(n_1, \theta; m_1, m_2)$ with $\theta = 1$. Then, by the Markov reward (or dynamic programming) relation (12) with (11) and (7) substituted, we find

$$\begin{aligned}
&\mathbf{V}^{k+1}(n_1, 1; m_1, m_2) \\
&\quad = h 1_{(m_1+m_2=c)} \\
&\qquad + h\lambda_1 1_{(m_1+m_2<c)} \mathbf{V}^k(n_1 + 1, 1; m_1, m_2) \\
&\qquad + h\lambda_2 1_{(m_1+m_2<c)} \mathbf{V}^k(n_1, 1; m_1, m_2 + 1)
\end{aligned}$$

$$
\begin{aligned}
&+ h\mu_1 1_{(m_1+m_2<c)} 1_{(n_1>0)} \mathbf{V}^k(n_1-1,1;m_1+1,m_2)\\
&+ h\mu_1 1_{(m_1+m_2=c)} 1_{(n_1>0)} \mathbf{V}^k(n_1,2;m_1,m_2)\\
&+ hm_1\mu_{21} 1_{(m_1>0)} \mathbf{V}^k(n_1,1;m_1-1,m_2)\\
&+ hm_2\mu_{22} 1_{(m_2>0)} \mathbf{V}^k(n_1,1;m_1,m_2-1)\\
&+ [1 - h\lambda_1 1_{(m_1+m_2<c)} - h\lambda_2 1_{(m_1+m_2<c)} - h\mu_1 1_{(n_1>0)} - hm_1\mu_{21} - hm_2\mu_{22}]\\
&\times \mathbf{V}^k(n_1,1;m_1,m_2) \qquad (36)
\end{aligned}
$$

and similarly

$$
\begin{aligned}
&\mathbf{V}^{k+1}(n_1+1,1;m_1,m_2)\\
&= h1_{(m_1+m_2=c)}\\
&\quad + h\lambda_1 1_{(m_1+m_2<c)} \mathbf{V}^k(n_1+2,1;m_1,m_2)\\
&\quad + h\lambda_2 1_{(m_1+m_2<c)} \mathbf{V}^k(n_1+1,1;m_1,m_2+1)\\
&\quad + h\mu_1 1_{(m_1+m_2<c)} \mathbf{V}^k(n_1,1;m_1+1,m_2)\\
&\quad + h\mu_1 1_{(m_1+m_2=c)} \mathbf{V}^k(n_1+1,2;m_1,m_2)\\
&\quad + hm_1\mu_{21} 1_{(m_1>0)} \mathbf{V}^k(n_1+1,1;m_1-1,m_2)\\
&\quad + hm_2\mu_{22} 1_{(m_2>0)} \mathbf{V}^k(n_1+1,1;m_1,m_2-1)\\
&\quad + [1 - h\lambda_1 1_{(m_1+m_2<c)} - h\lambda_2 1_{(m_1+m_2<c)} - h\mu_1 - hm_1\mu_{21} - hm_2\mu_{22}]\\
&\quad \times \mathbf{V}^k(n_1+1,1;m_1,m_2). \qquad (37)
\end{aligned}
$$

Now, in order to compare (36) and (37) pairwise by transition, (36) and (37) are first slightly rewritten as follows. In (36) we artificially add as well as subtract an extra term:

$$h\mu_1 1_{(n_1=0)} \mathbf{V}^k(n_1,1;m_1,m_2)$$

while in (37) we rewrite the coefficient $h\mu_1$ as

$$h\mu_1 = h\mu_1 1_{(n_1>0)} + h\mu_1 1_{(n_1=0)}.$$

Then by subtracting (36) from (37) and collecting terms transition-wise we obtain:

$$
\begin{aligned}
&\Delta^1\mathbf{V}^{k+1}(n_1,1;m_1,m_2)\\
&= h\lambda_1 1_{(m_1+m_2<c)} \Delta^1\mathbf{V}^k(n_1+1,1;m_1,m_2)\\
&\quad + h\lambda_2 1_{(m_1+m_2<c)} \Delta^1\mathbf{V}^k(n_1,1;m_1,m_2+1)\\
&\quad + h\mu_1 1_{(m_1+m_2<c)} 1_{(n_1>0)} \Delta^1\mathbf{V}^k(n_1-1,1;m_1+1,m_2)\\
&\quad + h\mu_1 1_{(m_1+m_2<c)} 1_{(n_1=0)} \big[\mathbf{V}^k(0,1;m_1+1,m_2) - \mathbf{V}^k(0,1;m_1,m_2)\big]\\
&\quad + h\mu_1 1_{(m_1+m_2=c)} 1_{(n_1>0)} \big[\mathbf{V}^k(n_1+1,2;m_1,m_2) - \mathbf{V}^k(n_1,2;m_1,m_2)\big]
\end{aligned}
$$

$$
\begin{aligned}
&+ h\mu_1 1_{(m_1+m_2=c)} 1_{(n_1=0)} \big[\mathbf{V}^k(1,2;m_1,m_2) - \mathbf{V}^k(0,1;m_1,m_2)\big] \\
&+ hm_1\mu_{21} 1_{(m_1>0)} \Delta^1 \mathbf{V}^k(n_1,1;m_1-1,m_2) \\
&+ hm_2\mu_{22} 1_{(m_2>0)} \Delta^1 \mathbf{V}^k(n_1,1;m_1,m_2-1) \\
&+ [1 - h\lambda_1 1_{(m_1+m_2<c)} - h\lambda_2 1_{(m_1+m_2<c)} - h\mu_1 - hm_1\mu_{21} - hm_2\mu_{22}] \\
&\times \Delta^1 \mathbf{V}^k(n_1,1;m_1,m_2). \qquad (38)
\end{aligned}
$$

Now note that the fourth $\mathbf{V}^k$-difference term between brackets on the right-hand side of (38) corresponds to the difference in inequality (29) for $\theta = 1$, the fifth to $\Delta^1 \mathbf{V}^k(n_1,\theta;m_1,m_2)$ and thus inequality (28) for $\theta = 2$, and the sixth to inequality (34). As a consequence, by substituting the lower bound 0 from the induction hypothesis (28), (29), and (34) for $t = k$, the right-hand side of (38) is estimated from below by 0, that is $\Delta^1 \mathbf{V}^t(n_1,1;m_1,m_2) \geq 0$ for $t = k+1$. Similarly, by substituting the upper bounds Q from (28), (29) and (34) for $t = k$, and noting that all coefficients sum up to 1, we also verify $\Delta^1 \mathbf{V}^t(n_1,1;m_1,m_2) \leq Q$ for $t = k+1$. This proves (28) with $\theta = 1$ for $t = k+1$.

Following steps similar to those in (36), (37), and (38), for $\theta = 2$ and thus necessarily $m_1 + m_2 = c$, we find:

$$
\begin{aligned}
&\Delta^1 \mathbf{V}^{k+1}(n_1,2;m_1,m_2) \\
&\quad = hm_1\mu_{21} 1_{(m_1>0)} 1_{(n_1>0)} \Delta^1 \mathbf{V}^k(n_1-1,0;m_1,m_2) \\
&\qquad + hm_2\mu_{22} 1_{(m_2>0)} 1_{(n_1>0)} \Delta^1 \mathbf{V}^k(n_1-1,0;m_1+1,m_2-1) \\
&\qquad + hm_1\mu_{21} 1_{(m_1>0)} 1_{(n_1=0)} \Delta^1 \big[\mathbf{V}^k(0,1;m_1,m_2) - \mathbf{V}^k(0,1;m_1-1,m_2)\big] \\
&\qquad + hm_2\mu_{22} 1_{(m_2>0)} 1_{(n_1=0)} \Delta^1 \big[\mathbf{V}^k(0,1;m_1+1,m_2-1) - \mathbf{V}^k(0,1;m_1,m_2-1)\big] \\
&\qquad + \big[1 - (hm_1\mu_{21} + hm_2\mu_{22})(1_{(n_1>0)} + 1_{(n_1=0)})\big] \Delta^1 \mathbf{V}^k(n_1,2;m_1,m_2) \qquad (39)
\end{aligned}
$$

Hence, by substituting the induction hypotheses (28) and (29) again for $t = k$, we also verify (28) for $t = k+1$ and $\theta = 2$. Similarly, with $\theta = 0$ and thus necessarily again $m_1 + m_2 = c$,

$$
\begin{aligned}
&\Delta^1 \mathbf{V}^{k+1}(n_1,0;m_1,m_2) \\
&\quad = hm_1\mu_{21} 1_{(m_1>0)} 1_{(n_1>0)} \Delta^1 \mathbf{V}^k(n_1,1;m_1-1,m_2) \\
&\qquad + hm_2\mu_{22} 1_{(m_2>0)} 1_{(n_1>0)} \Delta^1 \mathbf{V}^k(n_1,1;m_1,m_2-1) \\
&\qquad + [1 - hm_1\mu_{21} - hm_2\mu_{22}] \Delta^1 \mathbf{V}^k(n_1,0;m_1,m_2) \qquad (40)
\end{aligned}
$$

by which (28) is also verified, by using the induction hypothesis, for $t = k+1$ and $\theta = 0$. Accordingly, we have thus proven the induction step for the inequalities (28). The induction steps for inequalities (29)–(31) follow along similar lines and are left to the reader. □

Remark 11 Similarly to [38] and [40], the value $Q = \lambda_1 + \lambda_2 + \mu_1$ follows from the inductive verification of difference terms as in (29)–(31), as the throughput rate for

station 2. (E.g., in [38] in the single-server case a slightly sharper bound $Q = 1/\mu_2$ was concluded.)

Inequalities (32)–(34) With $m_1 + m_2 = c$ and $n_1 > 0$, as in (36), (37) and (38), we derive for $t = k + 1$

$$\begin{aligned}
&\big[\mathbf{V}^{k+1}(n_1, 2; m_1, m_2) - \mathbf{V}^{k+1}(n_1, 1; m_1, m_2)\big] \\
&\quad = h\mu_1\big[\mathbf{V}^k(n_1, 2; m_1, m_2) - \mathbf{V}^k(n_1, 2; m_1, m_2)\big] \\
&\qquad + hm_1\mu_{21}1_{(n_1>0)}1_{(m_1>0)}\big[\mathbf{V}^k(n_1 - 1, 0; m_1, m_2) - \mathbf{V}^k(n_1, 1; m_1 - 1, m_2)\big] \\
&\qquad + hm_2\mu_{22}1_{(n_1>0)}1_{(m_2>0)}\big[\mathbf{V}^k(n_1 - 1, 0; m_1 + 1, m_2 - 1) \\
&\qquad - \mathbf{V}^k(n_1, 1; m_1, m_2 - 1)\big] \\
&\qquad + [1 - h\mu_1 - hm_1\mu_{21}1_{(n_1>0)}1_{(m_1>0)} - hm_2\mu_{22}1_{(n_1>0)}1_{(m_2>0)}] \\
&\qquad \times \big[\mathbf{V}^k(n_1, 2; m_1, m_2) - \mathbf{V}^k(n_1, 1; m_1, m_2)\big]. \qquad (41)
\end{aligned}$$

Here, the first $\mathbf{V}^k$-difference term on the right-hand side is indeed equal to 0 but left in for clarity of its derivation. The second and third correspond to those in inequality (31) for $\theta = 0$. Hence, the induction hypotheses (31) for $t = k$ and $\theta = 0$ and (32) for $t = k$ can be applied to also conclude (32) for $t = k + 1$.

Similarly, with $m_1 + m_2 = c$ and $n_1 > 0$, for $t = k + 1$

$$\begin{aligned}
&\big[\mathbf{V}^{k+1}(n_1, 1; m_1, m_2) - \mathbf{V}^{k+1}(n_1, 0; m_1, m_2)\big] \\
&\quad = h\mu_1\big[\mathbf{V}^k(n_1, 2; m_1, m_2) - \mathbf{V}^k(n_1, 0; m_1, m_2)\big] \\
&\qquad + hm_1\mu_{21}1_{(n_1>0)}1_{(m_1>0)}\big[\mathbf{V}^k(n_1 - 1, 0; m_1, m_2) - \mathbf{V}^k(n_1, 1; m_1 - 1, m_2)\big] \\
&\qquad + hm_2\mu_{22}1_{(n_1>0)}1_{(m_2>0)}\big[\mathbf{V}^k(n_1 - 1, 0; m_1 + 1, m_2 - 1) \\
&\qquad - \mathbf{V}^k(n_1, 1; m_1, m_2 - 1)\big] \\
&\qquad + [1 - h\mu_1 - hm_1\mu_{21}1_{(n_1>0)}1_{(m_1>0)} - hm_2\mu_{22}1_{(n_1>0)}1_{(m_2>0)}] \\
&\qquad \times \big[\mathbf{V}^k(n_1, 1; m_1, m_2) - \mathbf{V}^k(n_1, 0; m_1, m_2)\big]. \qquad (42)
\end{aligned}$$

So again, the induction step for (33) and $t = k + 1$ could be concluded by substituting (31) for $\theta = 0$ and $t = k$ in the second and third difference term on the right-hand side of (42) and (33) for $t = k$ in the fourth, provided also

$$0 \le \mathbf{V}^k(n_1, 2; m_1, m_2) - \mathbf{V}^k(n_1, 0; m_1, m_2) \le Q. \qquad (43)$$

To conclude the lower bound 0 in (43) we can simply write

$$\begin{aligned}
&\big[\mathbf{V}^k(n_1, 2; m_1, m_2) - \mathbf{V}^k(n_1, 0; m_1, m_2)\big] \\
&\quad = \big[\mathbf{V}^k(n_1, 2; m_1, m_2) - \mathbf{V}^k(n_1, 1; m_1, m_2)\big] \\
&\qquad + \big[\mathbf{V}^k(n_1, 1; m_1, m_2) - \mathbf{V}^k(n_1, 0; m_1, m_2)\big] \qquad (44)
\end{aligned}$$

and use the lower bound 0 from (32) and (33) for $t = k$ (as by induction hypothesis). To conclude the upper bound Q in (43) we can also write

$$\begin{aligned}&\big[\mathbf{V}^k(n_1, 2; m_1, m_2) - \mathbf{V}^k(n_1, 0; m_1, m_2)\big]\\&\quad= \big[\mathbf{V}^k(n_1, 2; m_1, m_2) - \mathbf{V}^k(n_1, 1; m_1 - 1, m_2)\big]\\&\qquad+ \big[\mathbf{V}^k(n_1, 1; m_1 - 1, m_2) - \mathbf{V}^k(n_1, 0; m_1, m_2)\big]\end{aligned} \tag{45}$$

provided $m_1 > 0$. Hence, by noting that the second difference on the right-hand side of (45) will be non-positive as by (29) for $t = k$, and using the upper bound Q from (29) for $t = k$ for the first difference on the right-hand side, we have also proven (43) and thus (33) for $t = k + 1$. (When $m_1 = 0$ and thus $m_2 > 0$ since $m_1 + m_2 = c > 0$, we can rewrite (45) to $m_2 - 1$ and use (30) for $t = k$.)

Finally, again with $m_1 + m_2 = c$, to prove (34) we write

$$\begin{aligned}&\big[\mathbf{V}^{k+1}(1, 2; m_1, m_2) - \mathbf{V}^{k+1}(0, 1; m_1, m_2)\big]\\&\quad= hm_1\mu_{21}1_{(m_1>0)}\big[\mathbf{V}^k(0, 1; m_1, m_2) - \mathbf{V}^k(0, 1; m_1 - 1, m_2)\big]\\&\qquad+ hm_2\mu_{22}1_{(m_2>0)}\big[\mathbf{V}^k(0, 1; m_1 + 1, m_2 - 1) - \mathbf{V}^k(0, 1; m_1, m_2 - 1)\big]\\&\qquad+ [1 - hm_1\mu_{21}1_{(m_1>0)} - hm_2\mu_{22}1_{(m_2>0)}]\big[\mathbf{V}^k(1, 2; m_1, m_2)\\&\qquad- \mathbf{V}^k(0, 1; m_1, m_2)\big].\end{aligned} \tag{46}$$

Substituting the induction hypothesis for $t = k$ from (29) for $\theta = 1$ and from (34) in (46) then also proves (34) for $t = k + 1$. We have thus proven all inequalities (28)–(34) for $t = k + 1$. The induction completes the proof of Lemma 4.

Remark 12 (New-operation protocol) Again, with reference to Remark 2, instead of assuming (8), Lemma 4 remains to be valid under the alternative new operation protocol, under which no new operation will be started when the ICU becomes congested due to a completed operation. However, its technical details and the expressions as in (36)–(46) are slightly different, as a separate term comes in when $m_1 + m_2 + 1 = c$. Because the steps are identical and the expressions very similar, the details are left to the reader for this particular case.

References

1. Adan, I.J.B.F., Van der Wal, J.: Monotonicity of the throughput of a closed queueing network in the number of jobs. Oper. Res. **37**, 935–957 (1989)
2. Baskett, F., Chandy, M., Muntz, R., Palacios, J.: Open, closed and mixed networks of queues with different classes of customers. J. ACM **22**, 248–260 (1975)
3. Chandy, K.M., Martin, A.J.: A characterization of product-form queueing networks. J. ACM **30**(2), 286–299 (1983)
4. Chandy, K.M., Howard, J.H., Towsley, D.F.: Product form and local balance in queueing networks. J. ACM **24**, 250–263 (1977)
5. Costa, A.X., Ridley, S.A., Shahani, A.K., Harper, P.R., De Senna, V., Nielsen, M.S.: Mathematical modelling and simulation for planning critical care capacity. Anaesthesia **58**, 320–327 (2003)

6. Faddy, M.J., McClean, S.I.: Analyzing data on lengths of stay of hospital patients using phase-type distributions. Stoch. Models Bus. Ind. **15**, 311–317 (1999)
7. Foschini, G.J., Gopinath, B.: Sharing memory optimally. IEEE Trans. Commun. **31**, 352–359 (1983)
8. Griffiths, J.D., Price-Lloyd, N., Smithies, M., Williams, J.: A queueing model of activities in an intensive care unit. IMA J. Manag. Math. **17**, 277–288 (2006)
9. Hautvast, J.L.A., Bakker, J., Boekema-Bakker, N., Faber, J.A.J., Grobbee, D.E., Schrijvers, A.J.P.: Plaats in de herberg, een studie naar determinanten van opname- en ontslagproblemen in IC-afdelingen in Nederland. Julius Centrum voor Huisartsgeneeskunde en Patientgebonden Onderzoek, Utrecht (in Dutch) (2001)
10. Hordijk, A., Van Dijk, N.M.: Networks of queues with blocking. In: Kylstra, F.J. (ed.) Performance, vol. 81, pp. 51–65. North-Holland, Amsterdam (1981)
11. Hordijk, A., Van Dijk, N.M.: Networks of queues, Part I: Job-local-balance and the adjoint process. Part II: General routing and service characteristics. In: Lecture Notes in Control and Information Sciences, vol. 60, pp. 158–205. Springer, Berlin (1983)
12. Keilson, J., Kester, A.: Monotone matrices and monotone Markov processes. Stoch. Process. Appl. **5**, 231–245 (1977)
13. Kelly, F.P.: Reversibility and Stochastic Networks. Wiley, New York (1979)
14. Kim, S.C., Horowitz, I., Young, K.K., Buckley, T.A.: Analysis of capacity management of the intensive care unit in a hospital. Eur. J. Oper. Res. **115**, 36–46 (1999)
15. Kim, S.C., Horowitz, I., Young, K.K., Buckley, T.A.: Flexible bed allocation and performance in the intensive care unit. J. Oper. Manag. **18**, 427–443 (2000)
16. Lam, S.S.: Queueing networks with population size constraints. IBM J. Res. Dev. 370–378 (1977)
17. Litvak, N., Rijsbergen, M., Boucherie, R.J., Houdenhoven, M.: Managing the overflow of intensive care patients. Eur. J. Oper. Res. **14**, 998–1010 (2008)
18. Massey, W.A.: Stochastic orderings for Markov processes on partially ordered spaces. Math. Oper. Res. **12**, 350–367 (1987)
19. McManus, M.L., Long, M.C., Copper, A., Litvak, E.: Queuing theory accurately models the need for critical care. Anesthesiology **100**, 1271–1276 (2004)
20. Pham, D.N., Klinkert, A.: Surgical case scheduling as a generalized job shop scheduling problem. Eur. J. Oper. Res. **185**, 1011–1025 (2008)
21. Pittel, B.: Closed exponential networks of queues with saturation. The Jackson-type stationary distribution and its asymptotic analysis. Math. Oper. Res. **4**, 357–378 (1979)
22. Puterman, M.L.: Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, New York (2005)
23. Ridge, J.C., Jones, S.K., Nielsen, M.S., Shahani, A.K.: Capacity planning for intensive care units. Eur. J. Oper. Res. **105**, 346–355 (1998)
24. Scheffer, G.J. e.a.: Richtlijn: Organisatie en werkwijze op intensive care-afdelingen voor volwassenen in Nederland, Van Zuiden (In Dutch) (2006)
25. Shanthikumar, J.G., Yao, D.D.: The effect of increasing service rates in closed queueing network. J. Appl. Probab. **23**, 474–483 (1987)
26. Shanthikumar, J.G., Yao, D.D.: Stochastic monotonicity of the queue lengths in closed queueing networks. Oper. Res. **35**, 583–588 (1987)
27. Sonderman, D.: Comparing multi-server queues with finite waiting rooms, II: Different number of servers. Adv. Appl. Probab. **11**, 448–455 (1979)
28. Stoyan, D.: Bounds and approximations in queueing through monotonicity and continuity. Oper. Res. **25**, 851–863 (1977)
29. Stoyan, D.: Comparison Methods for Queues and Other Stochastic Models. Wiley, New York (1983)
30. Szekli, R.: Stochastic Ordering and Dependence in Applied Probability. Lecture Notes in Statist., vol. 97. Springer, New York (1995)
31. Tijms, H.C.: Stochastic Models: An Algorithmic Approach. Wiley, New York (1994)
32. Van Dijk, N.M.: A formal proof for the insensitivity of simple bounds for multi-server non-exponential tandem queues based on monotonicity results. Stoch. Process. Appl. **27**, 261–277 (1988)
33. Van Dijk, N.M.: Stop = Repeat for non-exponential stochastic networks with blocking. J. Appl. Probab. **28**, 159–173 (1991)
34. Van Dijk, N.M.: Queueing Networks and Product Forms: A Systems Approach. Wiley, Chichester (1993)
35. Van Dijk, N.M.: Bounds and error bounds for queueing networks. Ann. Oper. Res. **79**, 295–319 (1998)

36. Van Dijk, N.M.: Error bounds and comparison results: the Markov reward approach for queueing networks. Research Report, Department of Economics and Business, University of Amsterdam (2008)
37. Van Dijk, N.M., Kortbeek, N.: Approximation and bounds for the ICU-rejection probability in OT–ICU tandem systems. Research Report, Department of Economics and Business, University of Amsterdam (2007)
38. Van Dijk, N.M., Lamond, B.F.: Bounds for the call congestion of finite single-server exponential tandem queues. Oper. Res. **36**, 470–477 (1988)
39. Van Dijk, N.M., Taylor, P.G.: Strong stochastic bounds for the stationary distribution of a class of multicomponent performability models. Oper. Res. **46**, 665–674 (1998)
40. Van Dijk, N.M., Van der Wal, J.: Simple bounds and monotonicity results for finite multi-server exponential tandem queues. Queueing Syst. **4**, 1–16 (1989)
41. Van Dijk, N.M., Tsoucas, P., Walrand, J.: Simple bounds and monotonicity of the call congestion of infinite multiserver delay systems. Probab. Eng. Inf. Sci. **2**, 129–138 (1988)
42. Vohra, S., Dutta, G., Gush, D.K.: Capacity management of Intensive Care Units in a multi-specialty hospital in India. IIMA Working Papers, available at http://ideas.repec.org/p/iim/iimawp/2006-07-04.html (2006)